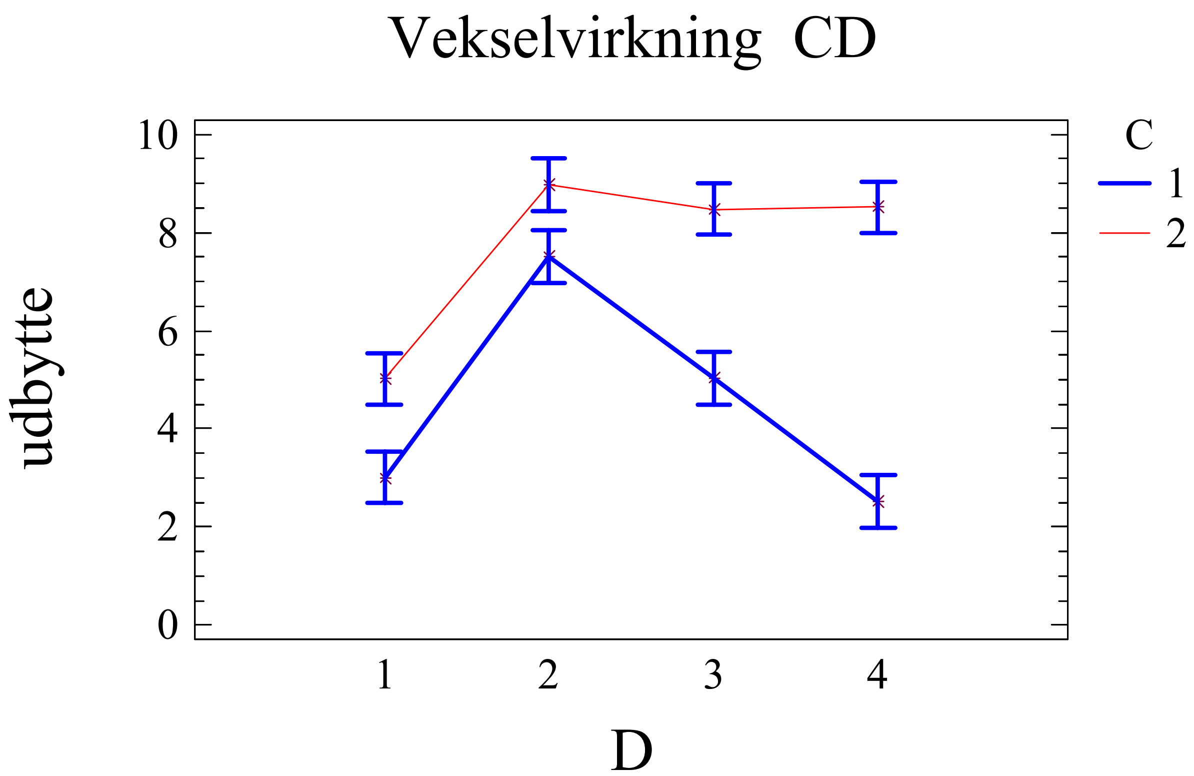
MOGENS ODDERSHEDE LARSEN

Statistisk forsøgsplanlægning

med benyttelse af Statgraphics



11 udgave 2002, DTU

FORORD

Dette notat er baseret på at de studerende både under kurset og ved evalueringen har nem adgang til at benytte statistikprogrammet Statgraphics (version 4 eller senere)
Dette er således tilfældet i det nuværende kursus i forsøgsplanlægning for diplomstuderende i kemi og levnedsmiddelingeniørstuderende på DTU.
Eksempler og øvelser er for manges vedkommende eksamensopgaver stillet ved DTU's eksaminer i forsøgsplanlægning for diplomstuderende. Sammen med notatet tænkes anvendt et hefte over tidligere eksamensopgaver. Brugen af statistikprogrammel bevirker, at undervisning og træning i "lommeregnermetoder" (eksempelvis Yates metode og ortogonale polynomier) kan undlades. Der bliver derved mere tid til model- og begrebsdannelser end tidligere.
Efter hvert større emne er der anbefalet en øvelse (en samlet oversigt kan findes i stikordsregisteret). En studerende som regner disse øvelser har været igennem nogenlunde de samme problemer, som gennemgås i notatets eksempler.

Jeg vil gerne takke L. Brøndum og J. D. Monrad for de mange gode råd de er kommet med gennem årene. Enkelte eksempler er således hentet fra deres lærebogssystem " Statistisk Forsøgsplanlægning Bind I, II, III.
En særlig tak fortjener lektor Bjarne Hellesen, som har anvendt de tidligere udgaver af notatet i undervisningen og i den forbindelse er kommet med mange konstruktive kommentarer og forslag til forbedringer samt været en stor hjælp ved korrekturlæsning. Han har endvidere udarbejdet eksempel 3.8 side 90 , eksempel 3.9 side 95, samt appendix 1.3 og 3.2. Desuden har han udarbejdet det program til konstruktion af varianskomponenter "EMSwin.exe", som udleveres i forbindelse med undervisningen.

I forhold til 10. udgave er der i kapitel 1 ændret nogle sider, foretaget enkelte rettelser i det øvrige notat, og tilsidst tilføjet nogle tabeller over t- og F-fordelingen.

11 udgave juli 2002

Mogens Oddershede Larsen

INDHOLD

5 MODELLER MED BÅDE KVANTITATIVE OG KVALITATIVE FAKTORERER

6 KOVARIANSANALYSEFORSØG

APPENDIX

1 GRUNDLÆGGENDE BEGREBER

1.1 Indledning

Et forsøg kan opspaltes i fem trin:
1) Problemformulering.
2) Udformning af forsøgsplan
3) Udførelse af forsøg
4) Analyse af forsøgsresultater
5) Konklusion
Statistisk forsøgsplanlægning drejer sig kort sagt om trinene 2, 4 og 5.

Formål med statistisk forsøgsplanlægning: At udarbejde en forsøgsplan , som med mindst mulig eksperimentel indsats og med anvendelsen af relevante statiske analysemetoder fortrinsvis varians- og regressionsanalyse giver den ønskede information .

En gylden regel er, at man skal ikke anvende mere end højst 25% af det afsatte forsøgsbudget på det første forsøg.

Endvidere skal man ud fra sin erfaring, studier af litteraturen, eventuelt ved nogle indledende forsøg på forhånd have en fornemmelse af, hvad man kan finde af betydende effekter og hvor store usikkerhederne er. Hvis man ikke har det, er man for tidligt ude.

Begreberne i forsøgsplanlægning gennemgås i det følgende på basis af nogle eksempler.

1.2. Forsøgsplan med 1 faktor

1.2.1 Indledning
I dette afsnit benyttes følgende eksempel som illustration af begreberne.

Eksempel 1.1 (én faktor). *Virkningerne af 4 tilsætningsstoffer* T_1, T_2, T_3, T_4 *på mængden af urenheder ved en kemisk proces ønskes sammenlignet. For hvert tilsætningsstof måles mængden af "uønsket stof"* 2 *gange.*
Forsøgsplanen fremgår af følgende skema:

T_1	T_2	T_3	T_4
× ×	× ×	× ×	× ×

Forsøget har 1 **faktor**: tilsætningsstof. Faktoren er på 4 **niveauer** T_1, T_2, T_3, T_4.
Hvert niveau har 2 **gentagelser.**
I dette kapitel vil vi kun betragte faktorer, der er såkaldte **"fixed" faktorer**.
Det er faktorer hvis niveauer man har valgt fordi man er specielt interesseret i netop disse. Eksempelvis har vi 4 tilsætningsstoffer, og vi er interesseret i at finde ud af hvilke af disse fire der resulterer i mindst urenhed.

Havde vi i stedet for haft et stort antal tilsætningsstoffer, og var interesseret i om tilsætningsstoffer som helhed giver anledning til en variation i resultaterne, så er tilsætningsstoffer en **"random** "faktor. Man ville så tilfældigt udtaget 4 af de mange stoffer, foretage forsøg med disse, og ud fra disse resultater have draget nogle konklusioner om tilsætningsstoffer generelt. Man vil her ikke være specielt interesseret i virkningen af en bestemt af de 4 tilfældigt udtagne stoffer.

1.2.2. Forsøg, behandlinger, forsøgsenheder og fuldstændig randomisering.

Forsøg: Et forsøg er opbygget af en række delforsøg, som hver giver en værdi (et forsøgsresultat) af den betragtede forsøgsvariabel .
I eksemplet udføres for hvert tilsætningsstof 2 delforsøg, dvs. i alt 8 delforsøg. Den betragtede forsøgsvariabel er mængden af "uønsket stof".
Hvert delforsøg i et forsøg udføres under en række forsøgsbetingelser (tilsætningsstoffer, tidspunkt, benyttet apparatur, benyttet råvareleverance, tryk, temperatur osv.).

Behandlinger: *De betingelser, der med forsæt varieres med henblik på en vurdering af deres indflydelse på forsøgsvariablen.*
I eksemplet er behandlingen : tilsætningsstof som er på 4 niveauer (4 stoffer)

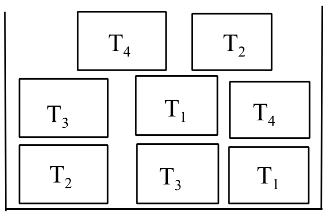
Forsøgsenheder: *Alle andre delforsøgsbetingelser* (tidspunkt benyttet apparatur, benyttet råvareleverance osv.) Der skal i eksemplet anvendes 8 forsøgsenheder svarende til de 8 delforsøg.

Randomisering. Ved en randomisering (lodtrækning) tilordnes behandlingerne til forsøgsenhederne på tilfældig måde. Herved sikres at virkningen af forsøgsenhederne fordeles tilfældigt på de forskellige behandlinger. En sammenligning mellem behandlingerne bliver derved ikke behæftet med systematiske fejl (et tilsætningsstof bliver ikke favoriseret ved at blive udsat for særlig gunstige forsøgsenheder). Herved sikrer man sig, et statistisk gyldigt forsøg.

Eksempel 1.2 (fortsættelse af eksempel 1.1) **(fuldstændig randomisering)**. *De 8 delforsøg skal indgå i den almindelige produktionsgang. når der er ledig kapacitet. Dette betyder, at man umuligt kan foretage delforsøgene på samme dag, med samme apparatur og med samme råvareleverance, hvilket ellers var ønskeligt.*
Angiv hvad der kunne tænkes at være forsøgsenheder. Forklar endvidere hvorledes en randomisering kan foretages mellem forsøgsenhederne og behandlingerne.

LØSNING:
I forsøgsenhederne må indgå hvilken dag, hvilket apparatur og hvilken råvareleverance der skal anvendes. I forsøgsenhederne indgår sandsynligvis også andre forhold udenfor vor kontrol, og som tilsammen bevirker, at selv om man udfører gentagne forsøg med samme behandling, så får vi afvigende resultater.
Lad os antage at der gælder følgende:
Mandag er det kun muligt at lave 1 forsøg, idet apparatur nr 1 og råvareleverance nr 1 kan benyttes.
Tirsdag er der kapacitet ledig til 3 forsøg:
   1 forsøg hvor apparatur nr 2 og råvareleverance 1 benyttes
   1 forsøg hvor apparatur nr 1 og råvareleverance 2 benyttes, og
   1 forsøg hvor apparatur nr 3 og råvareleverance 3 benyttes.
Onsdag kan der også laves 3 forsøg osv. (se det følgende skema).

Forsøgsenheder			Behandlinger
Dag	Apparatur	Råvareleverance	
mandag	1	1	
tirsdag	2	1	
tirsdag	1	2	
tirsdag	3	3	
onsdag	3	4	
onsdag	4	5	
onsdag	1	6	
torsdag	3	6	



Randomisering : Vi foretager nu en lodtrækning. Dette kan eksempelvis ske ved, at man har lavet 8 sedler hvor der på 2 sedler er skrevet T_1, på 2 sedler er skrevet T_2 osv. Sedlerne placeres i en dåse, f. eks. en tom kagedåse, og hver seddel kan eventuelt krølles sammen til en kugle. Sedlerne blandes ved at dåsen rystes (se figur). Hvis den første seddel der udtrækkes er T_2 så betyder det, at det delforsøg der mandag udføres med apparatur 1 og råvareleverance 1 skal anvende tilsætningsstof T_2 (se ovenstående skema). Hvis den næste seddel der udtrækkes er T_4 så betyder det, at det delforsøg der tirsdag udføres med apparatur 2 og råvareleverance 1 skal anvende tilsætningsstof T_4 osv.

Man siger, der er foretaget et **fuldstændigt randomiseret forsøg**.
E fuldstændigt randomiseret forsøg sikrer at der udføres et "statistisk gyldigt" forsøg.
Hvis vi derfor efter beregninger (ensidet variansanalyse) konkluderer, at der er forskel på tilsætningsstofferne, så er det "korrekt", idet det ville være helt tilfældigt, hvis én af tilsætningsstofferne har været begunstiget med særlig gode forsøgsenheder.

1.2.3 Ensidet variansanalyse.

Den statistiske metode til at analysere forsøget i eksempel 1.1 vil være ensidet variansanalyse. Eksemplet vil blive benyttet dels til en kort repetition af teorien bag en ensidet variansanalyse, dels til at gennemgå udskrifterne fra Statgraphics. I appendix A kan man finde hvilke menupunkter, tasttryk osv. der skal benyttes for at udføre en typisk Statgraphics beregning. I resten af notatet, vil man sædvanligvis blive henvist til dette appendix medmindre der er tale om specielle ordrer.

Eksempel 1.3 (fortsættelse af eksempel 1.2) **(ensidet variansanalyse).** *Efter at have foretaget et fuldstændigt randomiseret forsøg blev forsøgsresultaterne følgende:*

T_1	T_2	T_3	T_4
108 110	105 110	108 111	117 119

Foretag en statistisk analyse af disse resultater med henblik på at kunne angive hvilket tilsætningsstof der giver den mindste urenhed.

1.2.3.1 Modelformulering

Et forsøg med 1 faktor på *r* niveauer og med *n* gentagelser af hvert niveau kan generelt beskrives i følgende skema:

Niveauer	Observationer	Gennemsnit	Middelværdi
R_1	$x_{11}, x_{12}, \ldots, x_{1n}$	$\bar{x}_{1.}$	μ_1
R_2	$x_{21}, x_{22}, \ldots, x_{2n}$	$\bar{x}_{2.}$	μ_2
...	......	...	...
R_r	$x_{r1}, x_{r2}, \ldots, x_{rn}$	$\bar{x}_{r.}$	μ_r

Hvert af de *r* faktorniveauer har *n* gentagelser. I skemaet indgår også gennemsnittet af de *n* observationer, og middelværdien af de *r* faktorniveauer.
I eksempel 1.1 er $r = 4$ og $n = 2$.
En model for forsøget skrives $x_{ij} = \bar{\mu} + R_i + \varepsilon_{ij}$ hvor $i = 1,2,\ldots,r$ og $j = 1,2,\ldots,n$.

Her er den totale middelværdi $\bar{\mu} = \dfrac{\sum_{i=1}^{r} \mu_i}{r}$ gennemsnittet af de r niveau-middelværdier, rækkevirkningen $R_i = \mu_i - \bar{\mu}$, og ε_{ij} er "forsøgsfejlen ("støjen"), som indbefatter alle andre former for variation (stammende fra forsøgsenhederne m. m.).
En forudsætning for ved en variansanalyse at kunne teste forskellige relevante nulhypoteser. er at forsøgsfejlene ε_{ij} er uafhængige og normalfordelte variable med middelværdi 0 og varians σ^2. Variansen σ^2 antages at være den samme for alle faktorens niveauer. Man siger kort, at der er varianshomogenitet.

Ovennævnte forudsætninger medfører, at observationerne x_{ij} er uafhængige, og normalfordelte $n\left(\bar{\mu} + R_i, \sigma^2\right)$, dvs med middelværdi $\bar{\mu} + R_i$ og samme varians σ^2.

Vi vil i afsnit 1.3.7 vise, hvorledes man grafisk og ved hjælp af passende test kan undersøge om forudsætningerne er rimeligt opfyldt.

1.2.3.2 Anskuelig forklaring på den teoretiske baggrund for ensidet variansanalyse.

I eksempel 1.3 antages forudsætningerne for en variansanalyse at være tilstede .
Det er så relativt besværligt at foretage beregningerne "med lommeregner" at man sædvanligvis vil benytte et passende statistikprogram.
For at forstå den teoretiske baggrund for analysen, vil vi dog undtagelsesvis i dette eksempel starte med at foretage beregningerne "i hånden".
For at få et skøn for mængden af urenheder, udregnes gennemsnittene for hvert tilsætningsstof. Disse er angivet i nedenstående skema. Umiddelbart ud fra gennemsnit synes T_4 at adskille sig fra de tre øvrige, men hvis der er stor spredning, kan det måske blot være et tilfælde. Det er derfor naturligt at udregne spredningerne, hvilket derfor også er anført i skemaet.
Eksempel på beregning af spredning for en behandling:

$$s_{T_2} = \sqrt{\frac{(105-107.5)^2+(110-107.5)^2}{2-1}} = 3.5355$$

	T_1	T_2	T_3	T_4
Gennemsnit	109	107,5	109,5	118
Spredning	1.4142	3.5355	2,1213	1,4142

Idet hver varians er baseret på $n = 2$ målinger, har hver varians 1 frihedsgrad ($f = n - 1 = 2 - 1$).
Ifølge forudsætningen om varianshomogenitet antages de 4 varianser antages at være nogenlunde ens. Man beregner derfor et vægtet gennemsnit af disse (foretager en "pooling").

$$s_0^2 = \frac{f_1\cdot s_1^2+\ldots+f_4\cdot s_4^2}{f_1+f_2+..+f_4} = \frac{1\cdot 1.4142^2+\ldots+1\cdot s_4^2}{4} = 5.25 \text{ med } f_0 = 4 \text{ frihedsgrader.}$$

Indtastes de fire gennemsnit i en lommeregner findes $s_{\bar{x}}^2 = 22.5$.
Et gennemsnit af n tal har en varians, der er n gange mindre end variansen på den enkelte måling.
I dette tilfælde er $n = 2$. Lad $s_R^2 = 2\cdot s_{\bar{x}}^2 = 2\cdot 22.5 = 45$.

Forudsættes de 4 tilsætningsstoffer at have <u>samme</u> virkning, burde derfor $F = \frac{s_R^2}{s_0^2} \approx 1$, mens hvis de er forskellige er forholdet signifikant større end 1.

Vi får $F = \frac{s_R^2}{s_0^2} = \frac{45}{5.25} = 8.57$.

Resultaterne samles i en variansanlysetabel, hvor også $SAK = f\cdot s^2$ indgår.

Variation	SAK	f	$s^2 = \frac{SAK}{f}$	$F = \frac{s_R^2}{s_0^2}$
Tilsætningsstof	135	$r - 1 = 3$	45.0	8.57
Gentagelser	21.0	4	5.25	
Total	156.0	$n\cdot r - 1 = 7$		

Forholdet mellem de to varianser er F - fordelt F(3,4). Ved opslag i en F - tabel (tabel 2) findes $F_{0.95}(3,4) = 6.59$. Heraf ses, at sandsynligheden for at få en F - værdi på 8.57 forudsat H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$ er sand, er mindre end 5% . <u>Nulhypotesen må derfor forkastes på et signifikansniveau på 5%.</u>

I stedet for at slå op i en F-tabel kunne man have beregnet "P - værdien" = $P(F > 6.69)$. Benyttes eksempelvis Statgraphics (jævnfør appendix A, afsnit a7) fås P-værdi = 3.22%, dvs igen under 5 %.

Konklusion: , De fire tilsætningsstoffer har ikke samme virkning.

1.2.3.3 Beregning af ensidet variansanalyse ved Statgraphics.

I appendix A er beskrevet, hvilke ordrer man skal give i Statgraphics,for at kunne få de ønskede udskrifter frem. Der vil derfor i det følgende blive henvist hertil.

Data indtastes på sædvanlig måde (jævnfør appendix A afsnit a2.)
Resultatet ser således ud:

```
Stof     Urenhed
T1        108
T1        110
T2        105
T2        110
T3        108
T3        111
T4        117
T4        119
```

Husk at gemme datafilen inden man går videre.

Analyse af data

Der vælges menupunktet "Compare" (se appendix A afsnit a3)
Der fremkommer følgende variansanalysetabel med tilhørende kommentar.

```
ANOVA Table for urenhed by stof
                                Analysis of Variance
-----------------------------------------------------------------------------
Source              Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-----------------------------------------------------------------------------
Between groups               135,0      3         45,0       8,57       0,0324
Within groups                 21,0      4         5,25
-----------------------------------------------------------------------------
Total (Corr.)                156,0      7

The StatAdvisor
---------------
     The ANOVA table decomposes the variance of urenhed into two components: a
between-group component and a within-group component.
The F-ratio, which in this case equals 8,57143, is a ratio of the between-group
estimate to the within-group estimate.  Since the P-value of the F-test is less than
0,05, there is a statistically significant difference between the mean urenhed from
one level of stof
to another at the 95,0% confidence level.  To determine which means are significantly
different from which others, select Multiple Range Tests from the list of Tabular
Options.
```

Det ses, at vi får samme tabel som før, og at man her ikke behøver at slå op i en F tabel, idet P - value direkte fortæller, at nulhypotesen forkastes 1 stjernet (3,24%).
Vi har derfor samme konklusion.

Som det ses, giver "The StatAdvisor" ingen oplysninger, som ikke burde være kendt i forvejen. Den vil derfor sædvanligvis blive udeladt i det følgende.

Konfidensintervaller
Føst må det understreges, at havde vi fået en accept af nulhypotesen, havde vi konkluderet, at der ingen signifikant virkning var af de forskellige stoffer, og derfor havde vi højst beregnet et fælles gennemsnit og et fælles konfidensinterval.
Er der en signifikant forskel ved vi, at der er en reel forskel mellem det tilsætningsstof T_2 med det laveste gennemsnit og T_4 med det højeste gennemsnit. En nøjere undersøgelse kan ske ved at beregne 95% konfidensintervaller:

$\bar{x} \pm t_{0.975}(f_0) \cdot \sqrt{\frac{s_0^2}{n}}$. (t - værdien fås fra tabel 1.1 eller evt. fra Statgraphics:Appendix a afsnit a7.

I vort tilfælde fås: $\bar{x} \pm t_{0.975}(4) \cdot \sqrt{\frac{5.25}{2}} = \bar{x} \pm 2.78 \cdot 1.62 = \bar{x} \pm 4.50$.

Heraf ses, at forskellen mellem 2 gennemsnit skal være mindst $2 \cdot r = 2 \cdot 4.50 = 9.00$
Ud fra gennemsnittene ses derfor, at T_4 er signifikant større end de øvrige tilsætningsstoffer, som ikke kan adskilles.
Følgende tabel angiver konfidensintervallerne:

	T_1	T_2	T_3	T_4
Gennemsnit	109	107,5	109,5	118
95% Konfidensinter-val	[104.5 ; 113.5]	[103.0 ; 112.0]	[105.0 ; 114.0]	[113.5 ; 122.5]

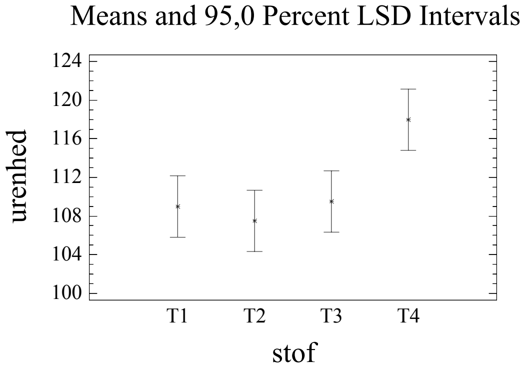
Konklusion: Man skal ikke vælge tilsætningsstof T_4, hvorimod de tre øvrige tilsætningsstoffer giver nogenlunde den samme mængde urenhed. ◆

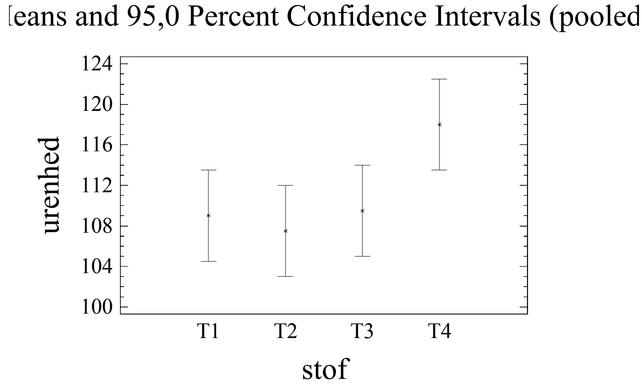
LSD = Least Signifikans Differens. Det er meget "populært" i stedet at foretage parvise sammenligninger, og på det grundlag danne konfidensintervaller.
Metoden er eksempelvis beskrevet i M. Oddershede Larsen: Videregående Statistik, afsnit 10, hvor den benyttes til at sammenligne 2 niveauer. Man beregner et konfidensinterval for differensen mellem de to middelværdier. Mens det som tidligere nævnt kræver en forskel på mindst $2 \cdot r$ mellem to gennemsnit, kræver denne metode kun en forskel på mindst $r \cdot \sqrt{2}$. I vort tilfælde altså kun $4.5 \cdot \sqrt{2} = 6.36$ mod før 9.0. Problemet ved denne metode er, at har man mange niveauer, der skal sammenlignes (10 metoder giver 45 sammenligninger) kan det let ske, at man begår en fejl af type 2 (konstatere en forskel der ikke er der). Derfor skal man være lidt forsigtig med at benytte den.
Beregning af konfidensintervaller med Statgraphics.
I appendix a3.1 vises hvorledes man får såvel LSD-intervaller som de sædvanlige konfidensinter-valler. I sidstnævnte tilfælde fås følgende udskrift:

```
Table of Means for urenhed by stof with 95,0 percent confidence intervals
--------------------------------------------------------------------------------
                                        Stnd. error
stof              Count          Mean    (pooled s)    Lower limit    Upper limit
--------------------------------------------------------------------------------
T1                    2         109,0       1,62019        104,502        113,498
T2                    2         107,5       1,62019        103,002        111,998
T3                    2         109,5       1,62019        105,002        113,998
T4                    2         118,0       1,62019        113,502        122,498
--------------------------------------------------------------------------------
Total                 8         111,0
```

Det ses, at konfidensintervallerne er de samme som vi fik ved den manuelle beregning.

De følgende plot af de to typer konfidensintervaller giver et godt overblik.





Øvelse 1.1. (ensidet variansanalyse). Regn opgave 1 i notatets opgavesamling side 235.

1.2.4. Dimensionering af forsøg.

Fejl af type I og type II: Ved enhver test kan der være to typer fejl. For bedre at forstå problemstillingen vil vi se på følgende skema.

		Beslutning	
		H_0 accepteres	H_0 forkastes
Forudsætning	H_0 er sand	Rigtig beslutning	Forkert beslutning Type I fejl
	H_0 er falsk	Forkert beslutning Type II fejl	Rigtig beslutning

Det må være et krav til en god test, at der kun er en lille sandsynlighed for at begå en fejl af type I eller type II.
I eksempel 1.1 ville en type I fejl være, hvis man konkluderer at mindst 1 tilsætningsstof afviger fra de øvrige selv om det ikke er tilfældet. Virksomheden bruger måske millionbeløb på at omlægge produktionen, og det er ganske forgæves.
En type II fejl ville være, at man ikke opdager, at et bestemt tilsætningsstof giver mindre urenhed. Dette er naturligvis uheldigt, men hvis det skyldes, at dette ikke blev opdaget fordi forskellen er ganske ringe har det muligvis ingen praktisk betydning.
Hvis en test har signifikansniveau α så vides, at $P(\text{type I fejl}) \leq \alpha$, dvs. **forkastes H_0**, så er vi rimelig sikre på, at have foretaget en korrekt beslutning.
Hvis vi **accepterer H_0** er det blot udtryk for, at vi ikke kan forkaste(svag konklusion: "H_0 frikendes på grund af bevisets stilling"). Man kan have begået en type II fejl, dvs. accepteret en falsk nulhypotese.

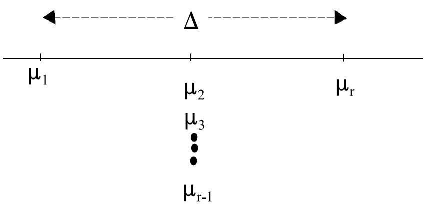
Lad os antage, at man finder, at hvis forskellen i mængden af "uønsket stof" mellem 2 tilsætningsstoffer er mindre end Δ enheder, så har det ingen praktisk interesse (Δ er bagatelgrænsen) og derfor gør det intet, hvis man ikke opdager det (begår en type II fejl). Hvis derimod forskellen er større end Δ enheder, så har det en væsentlig betydning, og sandsynligheden for at begå en type II fejl må derfor være lille, dvs. $P(\text{type II fejl}) \leq \beta$, hvor β eksempelvis kunne være 10%.

Ved en **dimensionering** af forsøget søger man at angive, hvor stor en stikprøvestørrelse *n* af hvert faktorniveau ,der mindst skal udføres, for at ovennævnte krav er opfyldt.
Udfører man det ud fra en dimensionering nødvendige antal forsøg, vil en accept af nulhypotesen nu betyde, at nok kan udbyttet være steget, men ikke så meget, at det har praktisk interesse.
Man kan dog ikke omvendt slutte, at fordi man for en forkastelse af nulhypotesen, så er forskellen større end Δ .
Et stort antal gentagelser af hvert niveau kan koste mange penge. Uden en passende planlægning kan disse penge være spildt, idet man jo risikerer at få signifikans for selv små forskelle. Disse har måske ved nærmere eftersyn ingen praktisk betydning.
En dimensionering vil altid være behæftet med nogen usikkerhed, da den bl.a. kræver, at man nogenlunde kan angive hvor stor forsøgsfejlens spredning σ ("støjen") er. Det er klart, at jo mere støj jo flere forsøg kræves for at kunne "overdøve støjen".
Da man jo endnu ikke har lavet forsøget, må man basere sin vurdering af σ på erfaringer fra tilsvarende forsøg, eller man må lave nogle få indledende forsøg for at få et estimat for den.

Det gør også en forskel hvordan niveauernes middelværdier fordeler sig. Hvis man i eksempel 1.1 tror de tre middelværdier ligger tæt sammen, mens den fjerde er placeret et stykke derfra, så giver det et andet resultat end hvis man mener de alle fordeler sig jævnt hen over et interval.
I komplicerede tilfælde, så er det bedste råd nok at simulere forsøget. Man kan ved hjælp af et EDB-program (som Statgraphics) konstruere en model, med de ønskede niveauer, med normalfordelte data som har de forventede middelværdier og den forventede spredning. For forskellige værdier af *n* kan man så statistisk vurdere resultatet, og på det grundlag finde ud af, hvad den optimale værdi for *n* er.
Til brug ved dimensionering i mindre komplicerede tilfælde er der konstrueret nogle OC - kurver. Disse findes afbildet i appendix 1.1 hvor de er hentet fra et større tabelværk.
Ud af den vandrette akse er afsat en størrelse Φ , som beregnes af formlen:

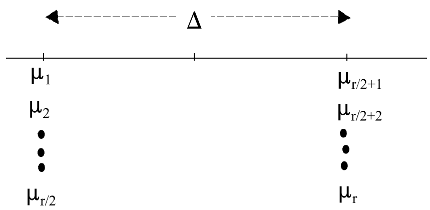
$$\Phi^2 = \frac{n\sum_{i=1}^{r} R_i^2}{r \cdot \sigma^2} \quad \text{hvor } R_i = \mu_i - \bar{\mu}$$

Hvorledes formlen skal tolkes, fremgår af følgende eksempel:



Eksempel 1.4 Illustration af beregning af Φ
Lad r være antal niveauer for den betragtede faktor.
Betragtes tilfældet på figuren hvor afstanden mellem de to yderste er Δ , og de øvrige er placeret lige midt imellem, så fås

$$\Phi^2 = \frac{n\sum_{i=1}^{r} R_i^2}{r \cdot \sigma^2} = \frac{n \cdot \left(\left(\frac{\Delta}{2}\right)^2 + \left(\frac{\Delta}{2}\right)^2\right)}{r \cdot \sigma^2} = \left(\frac{\Delta}{\sigma}\right)^2 \frac{n}{2 \cdot r}$$



Betragtes dernæst det andet ydertilfælde, hvor halvdelen af middelværdierne er placeret i det ene endepunkt, og den anden halvdel i det andet endepunkt, (der antages et lige antal niveauer), så fås

$$\Phi^2 = \frac{n\sum_{i=1}^{r} R_i^2}{r\cdot\sigma^2} = \frac{n\cdot\left(\left(\frac{\Delta}{2}\right)^2\cdot r\right)}{r\cdot\sigma^2} = \left(\frac{\Delta}{2\cdot\sigma}\right)^2\cdot n$$

Formlen kræver altså, at man har en forestilling om hvor store de enkelte virkninger er. Dette er sjældent tilfældet, så i stedet anvendes en formel, som gør, at man er på den “sikre” side, dvs. formlen giver sædvanligvis flere gentagelser end strengt nødvendigt.
Det kan vises, at de to tilfælde i eksempel 1.4 netop var ydertilfælde, og man derfor har formlerne

(1) $\Phi_{\text{flest}} \approx \frac{\Delta}{\sigma}\sqrt{\frac{n}{2\cdot r}}$ og (2) $\Phi_{\text{færrest}} \approx \frac{\Delta}{2\cdot\sigma}\sqrt{n}$.

Ved beregningerne benyttes derfor altid formel (1) ◆

Tallene ν_1 og ν_2 på kurverne er frihedsgraderne for den F - fordeling, som man anvender ved testen.

Vi illustrerer anvendelsen af formlen på problemet i eksempel 1.1.

Eksempel 1.5 (fortsættelse af eksempel 1.1) (**dimensionering**)
Lad os antage, at vi ønsker at bagatelgrænsen i forsøget er ca. $\Delta = 9$, og at variansen σ^2 =5.
Lad os endvidere dimensionere efter, at α = 5% og β = 10%.

Idet der er i alt r = 4 niveauer fås $\Phi \approx \frac{9}{\sqrt{5}}\sqrt{\frac{n}{2\cdot 4}} = \frac{9}{\sqrt{40}}\sqrt{n} = 1.423\cdot\sqrt{n}$

Startes med n = 2 fås $\Phi \approx 2.01$
Vi danner nu følgende lille variansanlysetabel:

	f	
Tilsætningsstoffer	r - 1 = 3	ν_1
Gentagelser	n· r - 1 - (r - 1) = r · (n - 1) = 7 - 3 = 4	ν_2
Total	n · r - 1 = 2 ·4 -1 = 7	

Kurvesystem V, har for α = 0.05 ikke en kurve svarende til ν_2 = 4, men det er dog klart, at β er over 35%, altså langt over 10%.
Vi laver nu en tabel for forskellige n værdier.

n	$\Phi_{\text{flest}} \approx 1.423\sqrt{n}$	ν_1	$\nu_2 = \nu_T - \nu_1 = r\cdot(n-1)$	β
2	2.01	3	4	>10
3	2.46	3	9	<10

Vi ser, at med de nævnte krav, så burde der foretages 3 gentagelser pr. niveau.

◆

Øvelse 1.2. (dimensionering). Regn opgave 2 i notatets opgavesamling side 235.

1.3. Forsøgsplan med 2 faktorer.

1.3.1 Indledning

I dette afsnit benyttes følgende eksempel som illustration af begreberne.

Eksempel 1.6 (to faktorer). *En bilfabrikant ønsker at finde ud af, hvorledes* 3 *olieblandinger* O_1, O_2, *og* O_3, *og* 2 *karburatortyper* K_1 *og* K_2 *påvirker benzinforbruget.*

Vi siger, at vi har et forsøg med 2 **faktorer:** olieblanding og karburator.
Faktoren "olieblanding" er på 3 **niveauer** O_1, O_2 , og O_3 , mens faktoren "karburator" har 2 niveauer nemlig K_1 og K_2 .
For at forsøget kan have tilstrækkelig "styrke", har man ved en dimensionering af forsøget (se senere) at der skal være mindst 4 gentagelser af hvert niveau.

1.3.2 Een faktor ad gangen

I mange forsøgsvejledninger står, at man bør kun variere en faktor ad gangen. Alle andre faktorer end den udvalgte fastholdes på et bestemt niveau.
En forsøgsplan efter disse retningslinier kunne eksempelvis være som skitseret nedenfor, hvor hvert delforsøg er markeret med et ×:

		Karburator	
		K_1	K_2
Olieblanding	O_1	× × × ×	
	O_2	× × × ×	× × × ×
	O_3	× × × ×	

I dette eksempel, hvor der kun er 2 faktorer, vælger vi først at variere olieblandingen, mens den anden faktor fastholdes.
Idet vi har valgt først at fastholde karburatoren på niveauet K_1, kan forsøget udføres således:
12 af de 16 biler, som skal anvendes, udstyres med karburator K_1, og derefter (randomiseret) får 4 af disse biler olieblanding O_1, 4 andre biler olieblanding O_2, og de sidste 4 biler olieblanding O_3.
Efter at have kørt en udvalgt strækning måles benzinforbruget.
Derefter varieres den anden faktor (her karburator), mens olieblandingen fastholdes på O_2, dvs. de sidste 4 biler udstyres med karburator K_2 og olieblanding O_2.
Igen gennemkøres den udvalgte strækning, og benzinforbruget måles.
Det er vigtigt, at **hver behandling** har **lige mange gentagelser**.
Hver af de 4 behandlinger er her gentaget 4 gange, så der er i alt 16 delforsøg.

Indtegnes for hver karburator det gennemsnitlige benzinforbrug mod olie-blandingen, fremkommer tegningen på figur 1.

Umiddelbart ses, at K_1 giver lavest benzinforbrug, og O_1 (eller O_3) skal foretrækkes.
Hvad med benzinforbruget i karburator K_2, hvis vi anvender olieblanding O_1 eller O_3?
Kan man slutte, at benzinforbruget ved olieblanding O_1 og O_3 er lavere, når man bruger karburator K_1, end når man bruger karburator K_2?
Kun, hvis man ud fra tekniske eller andre grunde mener at vide, at "karburatorkurven" for K_2 er parallel med kurven for K_1 (ingen vekselvirkning), så er forsøgsplanen anvendelig, men ikke den bedste.

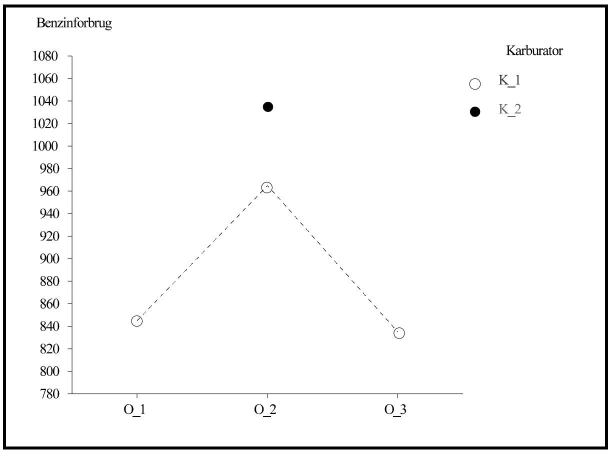


Fig. 1.

En forsøgsplan hvor der er mulighed for at undersøge, om der er vekselvirkning, og som sædvanligvis er mindre ressourcekrævende, er følgende:

1.3.3 Fuldstændig faktorstruktur

Denne plan består i, at hvert niveau af den ene faktor kombineres med ethvert niveau af den anden. Planen kan skitseres således:

		Karburator	
		K_1	K_2
Olieblanding	O_1	× ×	× ×
	O_2	× ×	× ×
	O_3	× ×	× ×

Her er hver af de 6 behandlinger gentaget 2 gange, dvs. i alt er der udført 12 delforsøg.
Endvidere er hvert niveau som ønsket gentaget mindst 4 gange.
I " en faktor ad gangen" var vi tvunget til at udføre 16 delforsøg, mens vi kun skal lave 12 delforsøg i det "fuldstændige faktorforsøg".
Vi kan altså nøjes med færre delforsøg, når vi laver et fuldstændigt faktorforsøg.
Indtegnes for hver karburator det gennemsnitlige benzinforbrug mod olie-blandingen, viser det sig, at man får figur 2.
Vi ser i modstrid med hvad vi troede ud fra "en faktor ad gangen forsøget", at kombinationen af katalysator K_2 og olieblanding O_1 giver det laveste benzinforbrug.
Det ses, at de to kurver **ikke er parallelle**, så der må være en vis **"vekselvirkning"**.
En model uden vekselvirkning (kurverne tilnærmelsesvis parallelle) siges at være **additiv.**

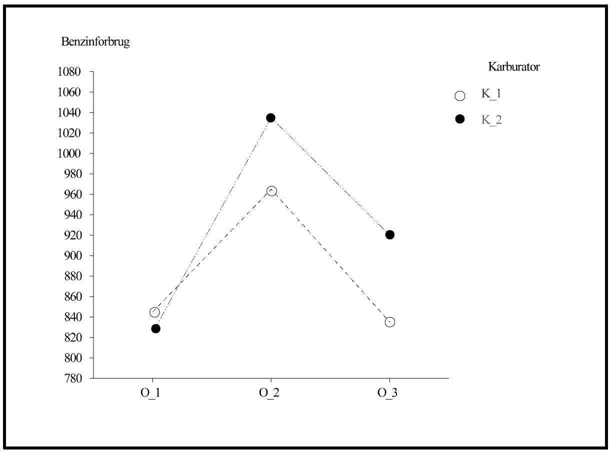


Fig. 2

1.3.4. Forsøg, behandlinger, forsøgsenheder og fuldstændig randomisering.

Begreberne blev defineret i afsnit 1.2.2. Vi benytter nu det følgende eksempel til at repetere begreberne i tilfældet med 2 faktorer.

Eksempel 1.7.(fortsættelse af eksempel 1.6) *En bilfabrikant ønsker at finde ud af, hvorledes* 3 *olieblandinger O_1, O_2, og O_3, og* 2 *karburatortyper K_1 og K_2 påvirker benzinforbruget. Forsøget planlægges som et fuldstændigt faktorforsøg idet hvert niveau skal gentages mindst 4 gange.*

1) *Angiv behandlingerne, antal gentagelser og antal forsøgsenheder.*

Et delforsøg med én bil tager 1 dag.(1 tank = 40 liter: Kører ca 15 km/l så 40 liter = 600 km, hvilket giver ca. 7 timer med 80 km/time). Anvendes forskellige biler (med tilhørende chauffør) kan det frygtes at give systematisk forskellige resultater.

Der foretages et ***fuldstændigt randomiseret forsøg*** *med samme bil og chauffør. Forsøgsenhederne bliver "dage" med tilhørende mandskab der installerer den korrekte karburator, olieblanding og vedligeholder bilen.*

2) *Angiv fordele og ulemper ved denne forsøgsplan*

3) *Angiv hvorledes randomiseringen i praksis kunne tænkes foretaget.*

LØSNING:

1) Planen kan som beskrevet i afsnit 1.3.3 skitseres således:

		Karburator	
		K_1	K_2
Olieblanding	O_1	× ×	× ×
	O_2	× ×	× ×
	O_3	× ×	× ×

Forsøget har **6 behandlinger:** O_1K_1, O_1K_2 , O_2K_1 , O_2K_2 , O_3K_1 , O_3K_2.

Hver behandling er **gentaget** 2 gange.

I alt skal der udføres 12 **delforsøg**, så der kræves 12 **forsøgsenheder**.

2) Fordele: Få delforsøg og en ret enkel forsøgsplan.
Ulemper: Forsøgsenheder måske for heterogene (stor spredning), da vejret og dermed vejforholdene kan skifte meget på 12 dage. Dette kan bevirke, at man ikke kan påvise selv ret store forskelle (stor "støj" kræver mange gentagelser for at man kan påvise en forskel)

3) **Randomisering:** Behandlingen O_1K_1 skrives på 2 sedler, behandlingen O_1K_2 skrives på 2 sedler osv. De 12 sedler placeres i en dåse. Det vedtages, at den første seddel der trækkes op af tromlen skal udføres på dag 1, den næste seddel på dag 2 osv. Resultatet blev:

O_2K_2	O_3K_1	O_3K_2	O_1K_2
O_3K_1	O_1K_2	O_2K_1	O_2K_2
O_1K_1	O_2K_1	O_1K_1	O_3K_2

dag 1	dag 2	dag 3	dag 4	dag 5	dag 6	dag 7	dag 8	dag 9	dag 10	dag 11	dag 12
O_2K_2	O_1K_2	O_3K_2	O_1K_2	O_3K_1	O_2K_1	O_1K_1	O_2K_2	O_1K_1	O_3K_2	O_2K_1	O_3K_1

1.3.5 Tosidet variansanalyse.

Den statistiske metode til at analysere et forsøg med 2 faktorer som i eksempel 1.6 kaldes tosidet variansanalyse. Som illustration af fremgangsmåden benyttes følgende eksempel:

Eksempel 1.8 (eksempel 1.6 og 1.7 fortsat).
Lad forsøgsresultaterne være følgende:

		Karburator	
		K_1	K_2
Olieblanding	O_1	830 860	810 840
	O_2	940 990	1050 1020
	O_3	855 815	930 910

Angiv hvilke kombinationer af karburator og olieblanding der giver det laveste forbrug, og giv et estimat for dette forbrug.
I undersøgelsen skal indgå en grafisk kontrol af den fundne model, og af om forudsætningerne for variansanalysen er tilfredsstillende opfyldt.

1.3.5.1 Modelformulering

Skal man mere generelt på formelform beskrive forsøget, så omdøber vi nu rækkefaktoren "olieblanding" til R og søjlefaktoren "karburator" til C (column)
I nedenstående skema er også i hver celle angivet gennemsnit angivet gennemsnit $\bar{x}$ og middelværdi $\therefore$, ligesom der er angivet de marginale gennemsnit og middelværdier.

		C : Karburator		
		C_1	C_2	
R Olieblanding	R_1	$x_{111} = 830$, $x_{112} = 860$ $\bar{x}_{11.} = 845$ μ_{11}	$x_{121} = 810$,$x_{122} = 840$ $\bar{x}_{12.} = 825$ μ_{12}	$\bar{x}_{1..} = 835$ $\mu_{1.}$
	R_2	$x_{211} = 940$, $x_{212} = 990$ $\bar{x}_{21.} = 965$ μ_{21}	$x_{221} = 1050$, $x_{222} = 1020$ $\bar{x}_{22.} = 1035$ μ_{22}	$\bar{x}_{2..} = 1000$ $\mu_{2.}$
	R_3	$x_{311} = 855$, $x_{312} = 815$ $\bar{x}_{31.} = 835$ μ_{31}	$x_{321} = 930$, $x_{322} = 910$ $\bar{x}_{32.} = 920$ μ_{32}	$\bar{x}_{3..} = 877.5$ $\mu_{3.}$
		$\bar{x}_{.1.} = 881.67$ $\mu_{.1}$	$\bar{x}_{.2.} = 926.67$ $\mu_{.2}$	$\bar{x}_{...} = 904.17$ $\mu_{..}$

Mere generelt kan et forsøg med to faktorer R og C i et fuldstændigt (r×q) faktorstruktur med n gentagelser af hver behandling skrives:

		C				
		C_1	C_2	...	C_q	
R	R_1	$x_{111}, x_{112}, \ldots x_{11n}$ $\bar{x}_{11.}$ μ_{11}	$x_{121}, x_{122}, \ldots x_{12n}$ $\bar{x}_{12.}$ μ_{12}	...	$x_{1q1}, x_{1q2}, \ldots x_{1qn}$ $\bar{x}_{1q.}$ μ_{1q}	$\bar{x}_{1..}$ $\mu_{1.}$
	R_2	$x_{211}, x_{212}, \ldots x_{21n}$ $\bar{x}_{21.}$ μ_{21}	$x_{221}, x_{222}, \ldots x_{22n}$ $\bar{x}_{22.}$ μ_{22}	...	$x_{2q1}, x_{2q2}, \ldots x_{2qn}$ $\bar{x}_{2q.}$ μ_{2q}	$\bar{x}_{2..}$ $\mu_{2.}$
	...	. . . .	....	...	...	
	R_r	$x_{r11}, x_{r12}, \ldots x_{r1n}$ $\bar{x}_{r1.}$ μ_{r1}	$x_{r21}, x_{r22}, \ldots x_{r2n}$ $\bar{x}_{r2.}$ μ_{r1}	...	$x_{rq1}, x_{rq2}, \ldots x_{rqn}$ $\bar{x}_{rq.}$ μ_{rq}	$\bar{x}_{r..}$ $\mu_{q.}$
		$\bar{x}_{.1.}$ $\mu_{.1}$	$\bar{x}_{.2.}$ $\mu_{.2}$	. . .	$\bar{x}_{.q.}$ $\mu_{.q}$	$\bar{x}_{...}$ $\mu_{..}$

En model for forsøget skrives $\tilde{x}_{ijk} = \bar{\mu} + R_i + C_j + (RC)_{ij} + \varepsilon_{ijk}$ hvor $\begin{cases} i = 1,2,\ldots,r \\ j = 1,2,\ldots,q \\ k = 1,2,\ldots,n \end{cases}$ **.**

Her er den totale middelværdi $\bar{\mu} = \dfrac{\sum_{i=1}^{r} \mu_i}{r}$ gennemsnittet af de $r \cdot q \cdot n$ niveau-middelværdier,

hovedvirkningen $R_i = \mu_{i.} - \bar{\mu}$ (også kaldet rækkevirkningen)

hovedvirkningen $C_j = \mu_{.j} - \bar{\mu}$ (også kaldet søjlevirkningen),

vekselvirkningen $(RC)_{ij} = \mu_{ij} - \bar{\mu} - R_i - C_j$ og

ε_{ij} er "forsøgsfejlen" ("støjen"), som indbefatter alle andre former for variation (stammende fra forsøgsenhederne m. m.).

Der gælder de samme variansanalyseforudsætninger som ved den ensidede variansanalyse (se evt. side 4), dvs. forsøgsfejlene ε_{ij} er uafhængige og normalfordelte variable med middelværdi 0 og konstant varians σ^2 .

Betragtes tallene i eksempel 1.8 fremgår det umiddelbart, hvordan man estimerer de forskellige virkninger. Eksempelvis er et estimat for hovedvirkningen $R_1 = \mu_1. - \bar{\mu}$ bestemt som

$\tilde{R}_1 = \bar{x}_1. - \bar{x}... = 835 - 904.17 = -69.17$

Ved variansanlysen testes først $H_0: RC = 0$ mod $H: RC \neq 0$, dvs. det testes om der er en signifikant vekselvirkning.

Er dette tilfældet søger man ud fra gennemsnit og konfidensintervaller, at bestemme de niveauer for de to faktorer som giver optimale resultater.

Er der ikke en signifikant vekselvirkning testes hypoteserne

$$H_0: \ R = 0 \quad \text{mod} \quad H: R \neq 0$$

$$H_0: \ C = 0 \quad \text{mod} \quad H: C \neq 0$$

dvs. om de 2 hovedvirkninger har en signifikant virkning.

1.3.5.2. Anskuelig forklaring på hvorledes man kan beregne vekselvirkning.

Da det kan være vanskeligt at forstå vekselvirkningsbegrebet, og hvordan man beregner det, gives her en kort forklaring.

I nedenstående skema er skitseret et forsøg med 2 faktorer R og C. R er på 3 niveauer, og C er på 4 niveauer. Der er 2 gentagelser af hver "behandling"(treatment).

	C_1	C_2	C_3	C_4
R_1	1 3	2 6	1 5	6 8
R_2	7 11	8 14	8 12	13 15
R_3	2 6	5 7	3 7	8 10

For hver af de 12 celler kan man udregne et skøn for spredningen. Hvis man forudsætter at spredningen er nogenlunde den samme i alle 12 tilfælde, kan man poole de 12 s^2 sammen til et fælles skøn s_0 for spredningen på forsøgsfejlen (støjen). Den vil have 12 frihedsgrader, da hvert enkelt s har 1 frihedsgrad.

I nedenstående skema er beregnet gennemsnit for hver celle, hver række, hver søjle og totalt.

	C_1	C_2	C_3	C_4	Gennemsnit
R_1	2	4	3	7	4
R_2	9	11	10	14	11
R_3	4	6	5	9	6
Gennemsnit	5	7	6	10	7

Tallene er konstrueret således, at vi har en helt præcis model uden vekselvirkning (Tallene i række 2 fremgår af tallene i række 1 + 7 og tallene i række 3 fremgår af tallene i række 1 + 2). Man siger, at modellen er additiv.

For en sådan model gælder helt præcist, at resultatet i celle (i, j) fås af række i's gennemsnit + søjle j's gennemsnit - totale gennemsnit.

Eksempel: I celle (2,3) står 10. Række 2's gennemsnit + søjle j's gennemsnit - totale gennemsnit = 11 + 6 - 7 = 10.

I praksis vil dette naturligvis aldrig være tilfældet på grund af den tilfældige variation (støj), men udregnes kvadratet på afvigelserne (SAK), og disse afvigelser ikke er større end hvad er rimeligt i forhold til støjen (s_0), vil vi kunne konkludere at der ikke kan konstateres nogen vekselvirkning.

1.3.5.3. Beregning af tosidet variansanlyse ved Statgraphics.

Analysen vil typisk ske i følgende rækkefølge:
a) Indtastning af data
b) Opstilling af model med tilhørende variansanalysetabel. På grundlag heraf eventuelt reducere modellen, igen opstille tabel osv., indtil man når den endelige model.
c) Grafisk kontrol af den i spørgsmål b fundne model, samt af forudsætninger for analysen.
d) Eventuelt på basis af den fundne model opstille relevante konfidensintervaller og derefter uddrage konklusion.

a) Data (fra eksempel 1.8) indtastes på sædvanlig måde (jævnfør appendix A afsnit a2.)
Resultatet ser således ud:

Karburator	oliebland	benzinforb
k1	o1	830
k1	o1	860
k1	o2	940
k1	o2	990
k1	o3	855
k1	o3	815
k2	o1	810
k2	o1	840
k2	o2	1050
k2	o2	1020
k2	o3	930
k2	o3	910

b) Opstille model med tilhørende variansanalysetabel, og eventuelt reducere model.
Vi starter med en model med 2 faktorer K (karburator) og O(olieblanding), og med mulighed for vekselvirkning mellem de to faktorer.
Der fremkommer først en variansanalysetabel uden vekselvirkning.
Vi sætter "Maximum Order Interaction" til 2 hvorved der fremkommer følgende variansanalyse-tabel:

```
Analysis of Variance for benzinforb - Type III Sums of Squares
--------------------------------------------------------------------------------
Source                 Sum of Squares     Df    Mean Square     F-Ratio    P-Value
--------------------------------------------------------------------------------
MAIN EFFECTS
 A:karburator                 6075,0      1          6075,0       10,12     0,0190
 B:oliebland                 58716,7      2         29358,3       48,93     0,0002

INTERACTIONS
 AB                           6450,0      2          3225,0        5,37     0,0460

RESIDUAL                      3600,0      6           600,0
--------------------------------------------------------------------------------
TOTAL (CORRECTED)            74841,7     11
--------------------------------------------------------------------------------
All F-ratios are based on the residual mean square error.
```

En sådan variansanalysetabel er opbygget efter følgende skema, hvor beregningen af frihedsgradstallet er angivet:

Variation	SAK (eller SS)	f	$s^2 = \frac{SAK}{f}$	$F = \frac{s_R^2}{s_0^2}$
A: Rækkefaktor		$r - 1 = 2 - 1$		
B: Søjlefaktor		$q - 1 = 3 - 1$		
AB:Vekselvirkning		$(r-1)(q-1) = 1 \cdot 2 =$		
Residual: (støj)		$r \cdot q \cdot (n-1) = 2 \cdot 3 \cdot (2-1)$		
Total		$r \cdot q \cdot n - 1$		

Et estimat for forsøgsfejlens varians fås ud for "residualen". Vi får $s_0^2 = 600$ med frihedsgradstallet $f_0 = 6$. Bidraget fra en eventuel vekselvirkning (AB) er 3225, altså væsentlig større. Forholdet mellem de to varianser er F = 5.375, og sandsynligheden for at få dette tal, hvis der ingen vekselvirkning var, er under 5% (4.60%). Anvender vi som sædvanlig et 5% signifikansniveau, så slutter vi, at der **er et svagt bevis for en vekselvirkning mellem karburator og olieblanding.**

Bemærk: Da de 2 faktorer har en virkning i form af en vekselvirkning, har det ingen mening at teste hovedvirkninger.

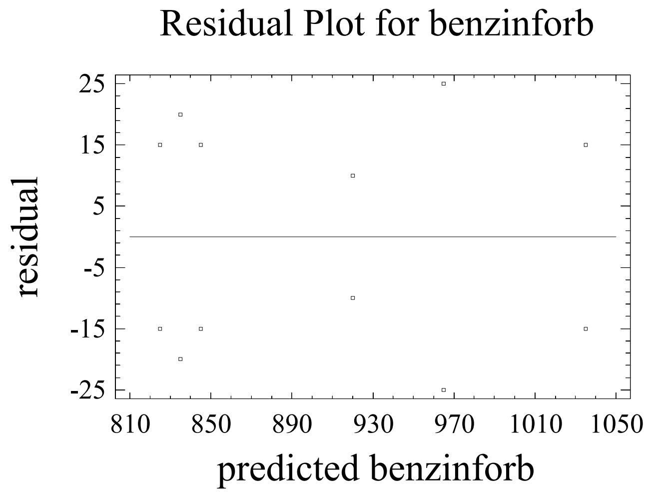
c) Grafisk kontrol af den i spørgsmål b fundne model, samt af forudsætninger for -analysen.

For at testene skal være statistisk holdbare, må man
1) være rimelig sikre på, at forsøgsresultaterne bliver rimelig beskrevet ved den fundne model,
2) kontrollere at forsøgsfejlene ε_{ij} er uafhængige og normalfordelte variable med middelværdi 0 og konstant varians σ^2.

Undersøgelser heraf udføres sædvanligvis ved på forskellig måde at studere de såkaldte residualer.
Disse er forskellen mellem de målte forsøgsresultater og de værdier som er beregnet ved anvendelse af den fundne model: $\tilde{x}_{211} = \bar{\mu} + R_2 + C_1 + (RC)_{21}$
Da modellen er den mest omfattende (da den ogaå indeholder vekselvirkningsleddet) bliver et estimat for de beregnede værdier gennemsnittene af værdierne i hver celle.
Eksempelvis er residual $r_{211} = x_{211} - \tilde{x}_{211} = 940 - 965.0$ (se tabel side 14) = - 25.0

Beregning af residual i en simplificeret model.
Antages modellen ikke at have vekselvirkning, er modellen $\tilde{x}_{211} = \bar{\mu} + R_2 + C_1$
Estimat for rækkevirkning er $\tilde{R}_2 = \bar{x}_2. - \bar{x}.. = 1000 - 904.17 = 95.83$
Estimat for søjlevirkning er $\tilde{C}_1 = \bar{x}_1. - \bar{x}.. = 881.67 - 904.17 = -22.40$
Heraf fås $\tilde{x}_{211} = 904.17 + 95.83 - 22.40 = 977.60$ og dermed residualen $r_{21} = 977.60 - 940.00 = 37.60$

1) I eksempel 1.8 var modellen den mest omfattende (da den også indeholder vekselvirkningsleddet), og man kan derfor ikke forbedre den. Imidlertid kan modellen være misvisende, hvis der findes såkaldte "outliers", dvs. af om nogle forsøgsresultater afviger mistænkeligt meget fra hvad man ville forvente ud fra modellen, og derfor måske skyldes fejlmålinger. Resultater der afviger mere end 3 σ må være under mistanke.
   Dette kan undersøges ved at tegne et såkaldt residualplot. Endvidere er det vigtigt at residualerne fordeler sig tilfældigt omkring 0, dvs. der må ikke være en tendens til at de eksempelvis mest ligger under 0-linien omkring midten og over på begge sider.
   Vælges "Residual versus Predicted" fås følgende tegning:



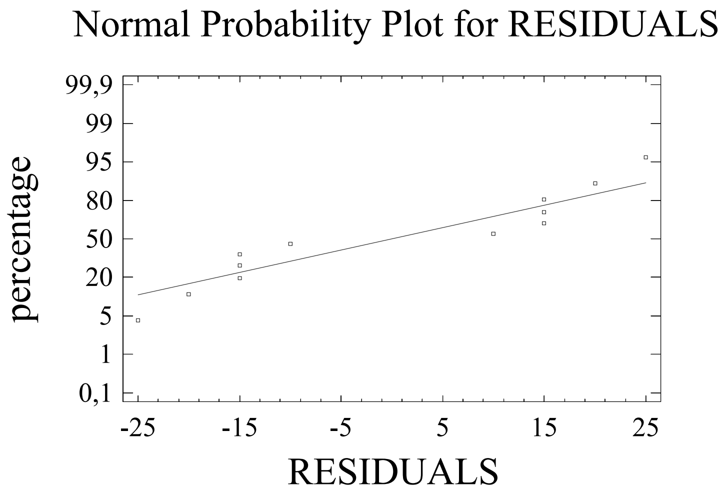
   På grund af gentagelserne er der symmetri omkring 0.
   Af variansanalysetabellen fremgår at $\sigma \approx \sqrt{600} \approx 25$. Da alle residualerne ligger indenfor $3 \cdot \sigma$ (endda indenfor σ) er der ingen outliers.

2a) De observerede resultater er værdier af **uafhængige** variable. Dette kan man sædvanligvis sikre ved en fornuftig forsøgsplan (randomisering).

2b) forsøgsfejlene ε_{ij} er normalfordelte variable med middelværdi 0 og konstant varians σ^2 .

   1) Kravet om normalfordeling undersøges grafisk ved at se på residualerne tegnet i et "normal probability plot" (manuelt kunne man have tegnet residualerne på et normalfordelingspapir). Residualerne bør så tilnærmelsesvis fordele sig omkring en ret linie. Det kræver dog et betydeligt antal målinger, for at man med rimelighed kan vurdere om normalfordelingskravet er opfyldt. Konklusionerne i variansanalyserne har dog stadig gyldighed, selv om der er nogen afvigelse fra normalitet (man siger de er robuste overfor normalitetskravet).

En grafisk test for normalitet (se appendix A) giver:



På grund af gentagelserne er der igen symmetri. Støjen synes nogenlunde normalfordelt, men antallet af målinger er for lille til, at man kan give en nogenlunde sikker vurdering (residualerne burde ligge "symmetrisk og tæt" ved en ret linie).

2) Kravet om varianshomogenitet (konstant varians σ^2) kan hvis der er gentagelser dels vurderes ved at betragte et residualplot, dels ved at foretage en mere formel test.
Da vi i eksempel 1.8 har 2 gentagelser i hver af de 6 celler, ligger residualerne to og to symmetrisk omkring 0-linien. På det på forrige side dannede residualplot ses, at afstanden mellem disse 6 punktpar er nogenlunde den samme. Man kan derfor godt tillade sig at konkludere, at der er en rimelig varianshomogenitet.
Til at foretage en mere formel testning af nulhypotesen $H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_6^2$ er der udviklet forskellige metoder. En meget benyttet metode er "Bartletts test" som imidlertid har den svaghed, at den i langt højere grad end selve variansanlysen er følsom overfor afvigelser fra normalitet. Der er derfor udviklet andre metoder, hvoraf specielt Cochrans og Levine test kan nævnes, da de ikke er så følsomme overfor normalitetskravet. Til gengæld kan Levines test ikke anvendes, hvis man kun har 2 gentagelser, hvilket jo ofte er tilfældet i praksis.
I Statgraphics foretages en test af varianserne som beskrevet i appendix A, afsnit a3.
Der fremkommer følgende udskrift med forklaring:

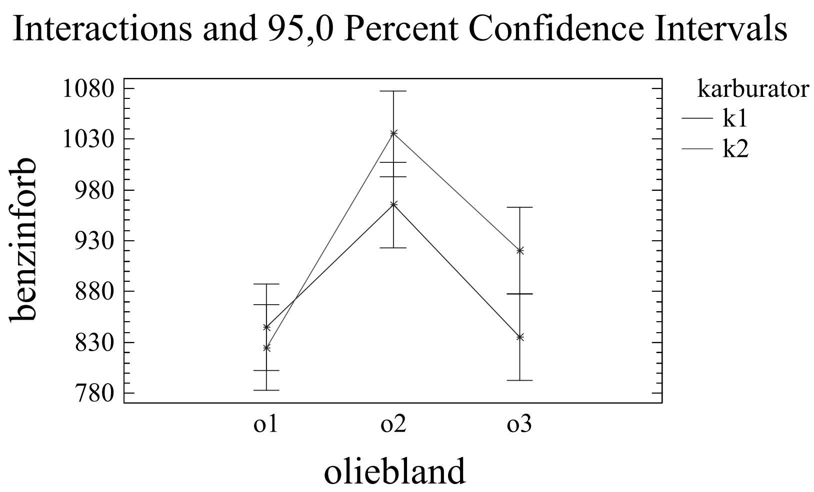
```
Variance Check
Cochran's C test: 0,347222   P-Value = 0,983143
Bartlett's test: 1,16961   P-Value = 0,984217
Hartley's test: 6,25

The StatAdvisor
---------------
The three statistics displayed in this table test the null hypothesis
that the standard deviations of benzinforb within each of the 6 levels
of behandlinger is the same.  Of particular interest are the two
P-values.  Since the smaller of the P-values is greater than or equal to
0,05, there is not a statistically significant difference amongst the
standard deviations at the 95,0% confidence level.
```

Som det ses, kan man ikke forkaste nulhypotesen, dvs ikke forkaste, at de 6 varianser er ens.
Vi vil derfor ved variansanalysen forudsætte, at der er varianshomogenitet.

d) Opstille relevante konfidensintervaller og drage konklusion
Vi ønsker nu for en given olieblanding at finde hvilken karburator, der giver det laveste benzinforbrug. Dette opnås lettest ved at beregne 95% konfidensintervaller for hver af de 6 behandlinger.
Vælges i den fundne model, "Tables of Means" fås følgende udskrift :

```
Table of Least Squares Means for benzinforb with 95,0 Percent Confidence Intervals
--------------------------------------------------------------------------------
                                           Stnd.         Lower         Upper
Level                  Count     Mean          Error         Limit         Limit
--------------------------------------------------------------------------------
GRAND MEAN             12        904,167
karburator
k1                     6         881,667       10,0          857,197       906,136
k2                     6         926,667       10,0          902,197       951,136
oliebland
o1                     4         835,0         12,2474       805,031       864,969
o2                     4         1000,0        12,2474       970,031       1029,97
o3                     4         877,5         12,2474       847,531       907,469
karburator by oliebland
k1         o1          2         845,0         17,3205       802,618       887,382
k1         o2          2         965,0         17,3205       922,618       1007,38
k1         o3          2         835,0         17,3205       792,618       877,382
k2         o1          2         825,0         17,3205       782,618       867,382
k2         o2          2         1035,0        17,3205       992,618       1077,38
k2         o3          2         920,0         17,3205       877,618       962,382
--------------------------------------------------------------------------------
```

Mere anskueligt er det nok at se på en tegning, så vi vælger at få tegnet et "interaction-plot" med afsatte konfidensintervaller:
Idet man som "second Factor" vælger den faktor med flest niveauer, fås følgende figur:



<u>Konklusion: Vi ser af tabel og figur, at man ikke bør vælge olieblanding O_3.</u>
Umidelbart giver kombinationen K_2 O_1 det laveste benzinforbrug (825), men af konfidensintervallerne ses, at der ingen signifikant forskel er mellem K_2 O_1 , K_1 O_3 og K_1 O_1.

1.3.5.4. Andre eksempler på 2-sidet variansanalyse

I det forrige afsnit blev gennemgået et eksempel med vekselvirkning og lige mange gentagelser af hver behandling. Vi vil her se på nogle eksempler hvor dette ikke er tilfældet.

1.3.5.4.1 Model uden vekselvirkning.(additiv model).

a) Eksempel 1.9 (additiv model)

Lad forsøgsresultaterne være følgende:

		Karburator	
		K_1	K_2
Olieblanding	O_1	830 860	850 840
	O_2	940 990	1050 1020
	O_3	855 815	930 910

(*I forhold til eksempel* 1.8 *er det de samme data pånær i tilfældet* O_1,K_2 *hvor tallet* 810 *er rettet til* 850).

Angiv hvilke kombinationer af karburator og olieblanding der giver det laveste forbrug, og giv et estimat for dette forbrug.

Løsning:

Med samme fremgangsmåde som før fås i dette tilfælde variansanalysetabellen :

```
Analysis of Variance for benzinforb - Type III Sums of Squares
--------------------------------------------------------------------------------
Source                Sum of Squares     Df    Mean Square    F-Ratio    P-Value
--------------------------------------------------------------------------------
MAIN EFFECTS
A:karburator                8008,33      1        8008,33      15,02     0,0082
B:oliebland                 53450,0      2        26725,0      50,11     0,0002
INTERACTIONS
AB                          4116,67      2        2058,33       3,86     0,0837
RESIDUAL                      3200,0      6        533,333
--------------------------------------------------------------------------------
TOTAL (CORRECTED)            68775,0     11
--------------------------------------------------------------------------------
```

Da *P* - value er 8.37% for vekselvirkningen AB, og dette er over vort signifikansniveau på 5%, tillader vi os i det følgende at antage, at vekselvirkningen er forsvindende.

I Statgraphics foretages nu en pooling ved at rette "Maximum order interaction" fra 2 til 1,[1] hvorved AB "pooles" ned i Residualen". Vi får følgende tabel:

```
Analysis of Variance for benzinforb - Type III Sums of Squares
--------------------------------------------------------------------------------
Source                Sum of Squares     Df    Mean Square    F-Ratio    P-Value
--------------------------------------------------------------------------------
MAIN EFFECTS
A:karburator                8008,33      1        8008,33       8,76     0,0182
B:oliebland                 53450,0      2        26725,0      29,22     0,0002

RESIDUAL                     7316,67      8        914,583
--------------------------------------------------------------------------------
TOTAL (CORRECTED)            68775,0     11
```

[1] Trykkes i stedet på "Exclude" ,fås et lille vindue med 2 kolonner ("Include" og "Exclude"), hvor man ved dobbeltklik på en vekselvirkning flytter den til den modsatte kolonne.

Vi ser nu, at forsøgsfejlens varians er 914.583, og at både karburator og olieblanding har en virkning, da P - value er under 5%.
Som før kan vi nu ved hjælp af "Means table" og eventuel "Means Plot" se, hvilken karburator og hvilken olieblanding der skal foretrækkes.

Table of Least Squares Means for benzinforb with 95,0 Percent Confidence Intervals

Level	Count	Mean	Stnd. Error	Lower Limit	Upper Limit
GRAND MEAN	12	907,5			
karburator					
k1	6	881,667	12,3463	853,196	910,137
k2	6	933,333	12,3463	904,863	961,804
oliebland					
o1	4	845,0	15,121	810,131	879,869
o2	4	1000,0	15,121	965,131	1034,87
o3	4	877,5	15,121	842,631	912,369

Konklusion: Det ses, at man skal vælge karburator K_1 og enten olieblanding O_1 eller olieblanding O_3.

◆

b) Kun én faktor har virkning

Tilfældet behandles som under tilfælde 1.3.7.2, idet også en faktor pooles ned i residualen. Dette sker ved, i nederste bjælke at vælge ikonen længst til venstre (rød "Input dialog") og derefter slette den faktor der skal pooles ned. Konfidensintervaller som ved ensidet variansanalyse.

1.3.5.4.2 Model uden gentagelser.

Havde vi kun én værdi i hver celle, kan vi ikke finde et estimat for forsøgsfejlens varians. Vi har derfor ingen mulighed for at foretage en analyse af forsøget, medmindre man på forhånd har en formodning om, at forudsætningerne for en variansanalyse er opfyldt, og der ikke er vekselvirkning mellem de to faktorer. Hvis dette er tilfældet, (kan måske sandsynliggøres grafisk), kan man teste hovedvirkningerne på sædvanlig måde.

Øvelse 1.3 (tosidet variansanalyse) Regn opgave 3 side 236.

1.3.5.4.3. Hvis forudsætninger for variansanalyse ikke er opfyldt.

1) Hvis forsøgvariablen er kontinuert, kan man foretage en passende transformation, eksempelvis ved at tage kvadratroden eller logaritmen til forsøgsresultaterne. Dette vil forminske variationen, og måske bevirke, at de nødvendige forudsætninger er opfyldt.
2) Hvis forsøgsvariablen er en tællevariabel (diskret variabel), som er Binomialfordelt transformers ved $Arc\sin\sqrt{H}$ hvor H er den relative hyppighed.
   Er den Poissonfordelt benyttes transformationen $\sqrt{x}$.
   På de transformerede tal kan man så foretage en sædvanlig variansanalyse (se nærmere herom i eksempelvis "Videregående Statistik").
3) Svigter alt andet kan udføres en rangtest (en ikke-parametrisk variansanalyse). (se nærmere herom i eksempelvis "Videregående Statistik").

1.3.5.4.4 Manglende observationer (ubalancerede forsøg).

Vi har hidtil forudsat, at forsøgsplanen var et fuldstændigt faktorforsøg med lige mange gentageser af hver behandling.vi havdeder var lige mange gentagelser af hver behandling. Dette vil i praksis ikke altid være tilfældet. Mislykkes således et delforsøg, opdager man det måske først så sent, at det ikke er muligt at foretage et nyt forsøg under de samme betingelser. Man har så et ikke-balanceret forsøg. Analysen foretages på samme måde, som hvis dataene var balancerede, men forsøget er nu mere følsomt overfor, om forudsætningen om varianshomogenitet er opfyldt.

Endvidere er det sådan, at mens der for et balanceret forsøg ingen forskel er på type I "Sum of Squares" og type III "Sum of Squares", bliver det nu vigtigt at forstå forskellen.

Til at illustrere dette, betragtes igen forsøget i eksempel 1.8, men denne gang antages, at et af forsøgene i celle O_3 K_1 er mislykket.

		Karburator	
		K_1	K_2
Olieblanding	O_1	830 860	810 840
	O_2	940 990	1050 1020
	O_3	855	930 910

Som før vælges "Multifactor Anova" og vi får nu:

```
Analysis of Variance for benzinforb - Type III Sums of Squares
--------------------------------------------------------------------------------
Source                Sum of Squares     Df    Mean Square    F-Ratio    P-Value
--------------------------------------------------------------------------------
MAIN EFFECTS
 A:oliebland                56250,0      2        28125,0      50,22     0,0005
 B:karburator               3778,57      1        3778,57       6,75     0,0484
INTERACTIONS
 AB                          4850,0      2         2425,0       4,33     0,0810
RESIDUAL                     2800,0      5          560,0
--------------------------------------------------------------------------------
TOTAL (CORRECTED)           66168,2     10
--------------------------------------------------------------------------------
All F-ratios are based on the residual mean square error.
```

Vi konkluderer, at der ikke er vekselvirkning, og eliminerer denne.

```
Analysis of Variance for benzinforb - Type III Sums of Squares
--------------------------------------------------------------------------------
Source                Sum of Squares     Df    Mean Square    F-Ratio    P-Value
--------------------------------------------------------------------------------
MAIN EFFECTS
 A:oliebland                55783,3      2        27891,7      25,52     0,0006
 B:karburator               3266,67      1        3266,67       2,99     0,1275
RESIDUAL                     7650,0      7        1092,86
--------------------------------------------------------------------------------
TOTAL (CORRECTED)           66168,2     10
--------------------------------------------------------------------------------
All F-ratios are based on the residual mean square error.
```

Bemærk, at når vekselvirkningen bortkastes så ændrer SAK for karburator og olieblanding sig også. Tilsvarende vil SAK størrelsen for olieblandingen ændre sig, når karburator bortkastes. Selvom olieblanding og karburator begge er ensstillede (hovedvirkninger) kan vi ikke på dette grundlag konkludere at olieblanding har betydning. Vi må først bortkaste karburator, og derefter se på olieblandingens *P* - værdi.
For at illustrere dette er det relevant at se på forskellen mellem type I og type III variansanalyse-tabeller.
Med cursoren på udskriften trykkes på højre musetast og "Pane Options" vælges. Derefter vælges type I. Dette giver følgende udskrift.

Analysis of Variance for benzinforb - Type I Sums of Squares

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
MAIN EFFECTS					
A:oliebland	55251,5	2	27625,8	25,28	0,0006
B:karburator	3266,67	1	3266,67	2,99	0,1275
RESIDUAL	7650,0	7	1092,86		
TOTAL (CORRECTED)	66168,2	10			

All F-ratios are based on the residual mean square error.

Havde vi i stedet valgt at indføre faktorerne i omvendt rækkefølge, havde vi fået:

Analysis of Variance for benzinforb - Type I Sums of Squares

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
MAIN EFFECTS					
A:karburator	2734,85	1	2734,85	2,50	0,1577
B:oliebland	55783,3	2	27891,7	25,52	0,0006
RESIDUAL	7650,0	7	1092,86		
TOTAL (CORRECTED)	66168,2	10			

All F-ratios are based on the residual mean square error.

The StatAdvisor

The ANOVA table decomposes the variability of benzinforb into contributions due to various factors. Since Type I sums of squares have been chosen, the contribution of each factor is measured having removed the effects of factors above it in the table. The P-values test the statistical significance of each of the factors. Since one P-value is less than 0,05, this factor has a statistically significant effect on benzinforb at the 95,0% confidence level.

Vi ser, at indføres karburator først i modellen, giver den et SAK bidrag på 2734.85, mens den giver et bidrag på 3266.67, hvis den indføres sidst. Tilsvarende gælder for olieblandinger.
Ønsker man at undersøge, om faktor A har en virkning, skal den placeres sidst, idet det er dens bidrag til at forbedre modellen, man skal vurdere, efter at alle de øvrige faktorer er indført.
Har man flere faktorer, der skal undersøges, er det lettere at anvende type III.

Dette er da også den udskrift Statgraphics automatisk giver, og som blev vist forrige side. Som det fremgår af "The StatAdvisor" enhver " SUM of Squares" (SAK) her beregnet, som om den pågældende faktor stod sidst (i overensstemmelse med, at summen af SAK'erne ikke er den totale SAK).

I appendix 1.3 har Bjarne Hellesen forklaret i detaljer, hvorledes man udregner SAK type I og type II ved benyttelse af matrixregning.
Man kan nu bortkaste den faktor, som har højeste p-value (over 5%) og så foretage en ny beregning. Man må ikke bortkaste flere faktorer på én gang, idet SAK'erne ændrer sig, når man bortkaster en faktor.

```
Analysis of Variance for benzinforb - Type III Sums of Squares
--------------------------------------------------------------------------------
Source                 Sum of Squares     Df    Mean Square    F-Ratio    P-Value
--------------------------------------------------------------------------------
MAIN EFFECTS
 A:olieblandg               55251,5      2        27625,8      20,24     0,0007

RESIDUAL                    10916,7      8        1364,58
--------------------------------------------------------------------------------
TOTAL (CORRECTED)           66168,2     10
--------------------------------------------------------------------------------
All F-ratios are based on the residual mean square error.
```

Det ses, at SAK for olieblanding har ændret sig.

Nærmere forklaring af beregninger ved ubalancerede forsøg.
Enhver variansanalyse kan udføres som en regressionsanalyse. Hvorledes dette regneteknisk kan gøres er beskrevet i "Statistisk Forsøgsplanlægning I" kapitel 14 (samt eksempel 65 side 309). Da det er den eneste måde, man kan beregne en variansanalyse med ikke balancerede data er det også denne metode der anvendes i "Multifactor Anova". En anden procedure der kan benyttes i Statgraphics, er den såkaldte "GLM procedure", som kan behandle forsøg med både kvantitative og kvalitative faktorer. Regressionsanalyse behandles i dette notat i kapitel 5 og 6.

1.3.6. Dimensionering:

I afsnit 1.2.4 er beskrevet formålet med at dimensionere og hvordan det gøres ved forsøg med 1 faktor. De samme betragtninger gælder ved forsøg med 2 faktorer. Man kan benytte de samme OC-kurver, og de samme formler.
Vi vil illustrere problemet ved følgende eksempel.
Eksempel 1.10 (fortsættelse af eksempel 1.1) **(dimensionering)**
Der ønskes en dimensionering af problemet i eksempel 1.1 idet forsøget udføres som et fuldstændigt faktorforsøg , at bagatelgrænsen i forsøget er Δ = 50 *enheder, at spredningen skønnes til at være ca.* σ = 25 *og at* α = 5% *og* β = 10%.
LØSNING:
Lad os antage, at der er *m* gentagelser i hver celle.

		Karburator	
		K_1	K_2
Olieblanding	O_1	*m*	*m*
	O_2	*m*	*m*
	O_3	*m*	*m*

Der er nu 2*m* gentagelser for hvert af de a = 3 niveauer for olieblanding, mens der er 3*m* gentagelser for hvert af de b = 2 niveauer for karburator. Det er derfor nok at beregne, hvor mange gentagelser der er nødvendige for olieblandinger, da kravene så automatisk også vil være opfyldt for karburator.
<u>Generelt skal man derfor altid dimensionere efter den faktor der har flest niveauer</u>.

Vi opstiller nu en variansanalysetabel med frihedsgrader.

Variation	f
A: Olieblanding	$a - 1 = 3 - 1 = 2$
B: Karburator	$b - 1 = 2 - 1 = 1$
AB: Vekselvirkning	$(a - 1)\cdot(b - 1) = 2$
Error (støj)	$a \cdot b \cdot (m-1) = 3 \cdot 2 \cdot (m-1) = 6 \cdot (m-1)$
Total	$a \cdot b \cdot m - 1 = 6 \cdot m - 1$

Formlen er: $\Phi_{\text{flest}} \approx \dfrac{\Delta}{\sigma}\sqrt{\dfrac{n}{2 \cdot r}}$, hvor man må være opmærksom på, at n her er antallet af gentagelser af et niveau. Idet antallet af gentagelser af niveauet er $b \cdot m$, fås derfor at $\Phi_{\text{flest}} \approx \dfrac{\Delta}{\sigma}\sqrt{\dfrac{b \cdot m}{2 \cdot a}}$. Idet $a = 3$ og $b = 2$ fås $\Phi_{\text{flest}} \approx \dfrac{\Delta}{\sigma}\sqrt{\dfrac{b \cdot m}{2 \cdot a}} = \dfrac{50}{25}\sqrt{\dfrac{2 \cdot m}{2 \cdot 3}} = 2\sqrt{\dfrac{m}{3}} = 1.155\sqrt{m}$.

Vi laver nu en tabel for forskellige m værdier.

m	$\Phi_{\text{flest}} \approx 1.155\sqrt{m}$	$\nu_1 = a - 1$	$\nu_2 = a \cdot b \cdot (m-1)$	β
2	1.63	2	6	ca 0.50
3	2.00	2	12	ca 0.20
4	2.31	2	18	ca 0.06

Vi ser, at med de nævnte krav, så burde der laves <u>4 gentagelser pr celle</u>, dvs. i alt 24 delforsøg.

◆

Øvelse 1.4. (dimensionering). Regn opgave 4 i notatets opgavesamling side 236.

1.4 Forsøgsplan med 4 faktorer

Indledning: Ved testning af modeller med mere end 1 faktor kommer man ud for at poole en række virkninger ned i residualen. Det følgende eksempel demonstrerer hvorledes dette kan gøres.

Man følger her den "normale" rækkefølge for pooling:

Først testes <u>alle</u> 3-faktorvekselvirkninger mod <u>samme</u> residual for at se dem bedømt på samme grundlag (med samme styrke). Derefter reduceres modellen, og man tester nu <u>alle</u> de 2-faktorvirkninger som ikke indgår i signifikante 3 - faktorvekselvirkninger mod den derved fremkomne residual. Efter en ny reduktion af modellen testes nu <u>alle</u> hovedvirkninger mod den nu fremkomne residual.

<u>I dette kursus følger vi ovennævnte regel</u>, men det er samtidig værd at vide, at mange anbefaler følgende regel:

Hvis frihedsgradstallet for residualen er over 10-12, så undlades at poole videre.Hvis frihedsgradstallet er mindre, så pooles endda kun, hvis P - værdien for den virkning, der står for at blive poolet er større end 0.25.

Begrundelsen herfor er naturligvis, at selv om man formelt siger, at en virkning er 0, hvis P - værdien er over signifikansniveauet, så er en accept af en nulhypotese jo ikke noget bevis herfor. Man kan derfor ved en pooling risikere, at residualen ikke er et sandt udtryk for "støjen". En sådan stor residual er derfor villedende, og kan let bevirke, at man ikke kan konstantere væsentlige forskelle.

Eksempel 1.11 (4-sidet variansanalyse med *n* = 1)

Virkningen af fire faktorer på hærdningstiden for en bestemt cementtype ønskes undersøgt.

De 4 faktorer var A: *Omrøringshastighed : 2 niveauer* A_1 , A_2
B: *Fabrikationsmetoder.* B_1 , B_2 , B_3
C: *Formalingdgrad af bestemt reaktant 2 niveauer* C_1 , C_2
D: *Råvaretype 4 niveauer* D_1 , D_2 , D_3 , D_4.

Forsøgsplanen var et fuldstændigt randomiseret forsøg med en fuldstændig faktorstruktur med 1 gentagelse af hver behandling. Følgende observationer af Y (kodede tal) fandtes:

		D_1		D_2		D_3		D_4	
		C_1	C_2	C_1	C_2	C_1	C_2	C_1	C_2
B_1	A_1	23.8	24.4	25.3	21.5	23.1	15.3	4.12	11.6
	A_2	45.8	47.7	35.5	26.1	10.0	20.0	5.30	13.2
B_2	A_1	16.0	18.9	11.6	11.8	22.7	20.9	28.9	21.1
	A_2	27.0	31.0	31.6	31.6	19.8	13.8	15.5	26.8
B_3	A_1	10.1	9.3	13.2	19.0	30.3	27.9	34.1	35.5
	A_2	17.0	27.8	24.6	25.2	17.0	27.2	21.1	34.7

Det formodes på forhånd, at kun nogle hovedvirkninger og tofaktorvirkninger er af praktisk interesse.

1) *Undersøg hvilke af de fire faktorer, der har virkninger eventuelt i form af vekselvirkninger.*
2) *Vurder grafisk i den udstrækning det er muligt ud fra residualerne i den under punkt 1 fundne model, om der er "outliers", om der er varianshomogenitet, og om "støjen" er rimeligt normalfordelt.*
3) *Angiv hvilke niveauer faktorerne skal stilles på for at give den mindst mulige størknings-tid.*(*Vink: Tegn passende interaktionsplot*).

LØSNING:

Indtastning af data

Dette sker på sædvanlig måde.

Starten af indtastningen i regnearket ses nedenfor:

```
A  B  C  D haerdetid
1  1  1  1  23.8
2  1  1  1  45.8
1  2  1  1  16.0
2  2  1  1  27.0
. . . . . . . . . . . . . . . . . .
```

1). Hvilke faktorer har virkning.

Vi starter med en model med 4 faktorer og med mulighed for op til 4-faktor vekselvirkninger (en fuldstændig model). Der opstilles den til modellen svarende variansanalysetabel:
Vi starter ved at sætte " Maximum Order Interaction til 4"
Der fremkommer følgende variansanalysetabel:

Analysis of Variance for haerdetid - Type III Sums of Squares

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
MAIN EFFECTS					
A:A	274,946	1	274,946		
B:B	22,7433	2	11,3716		
C:C	49,7761	1	49,7761		
D:D	140,396	3	46,7985		
INTERACTIONS					
AB	52,6883	2	26,3441		
AC	77,3176	1	77,3176		
AD	1009,34	3	336,448		
BC	49,7123	2	24,8561		
BD	2065,25	6	344,208		
CD	81,2751	3	27,0917		
ABC	11,2353	2	5,61763		
ABD	154,951	6	25,8252		
ACD	90,1489	3	30,0496		
BCD	61,5901	6	10,265		
ABCD	122,04	6	20,34		
RESIDUAL	0,0	0			
TOTAL (CORRECTED)	4263,41	47			

All F-ratios are based on the residual mean square error.

Det ses heraf, at forhåndsformodningen om at 3 og 4 faktorvekselvirkninger er nul synes rimelig.
Nu sættes " Maximum Order Interaction til 2", hvorved der fremkommer følgende tabel:

Analysis of Variance for haerdetid - Type III Sums of Squares

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
MAIN EFFECTS					
A:A	274,946	1	274,946	14,37	0,0009
B:B	22,7433	2	11,3716	0,59	0,5601
C:C	49,7761	1	49,7761	2,60	0,1204
D:D	140,396	3	46,7985	2,45	0,0896
INTERACTIONS					
AB	52,6883	2	26,3441	1,38	0,2723
AC	77,3176	1	77,3176	4,04	0,0563
AD	1009,34	3	336,448	17,59	0,0000
BC	49,7123	2	24,8561	1,30	0,2920
BD	2065,25	6	344,208	17,99	0,0000
CD	81,2751	3	27,0917	1,42	0,2635
RESIDUAL	439,966	23	19,1289		
TOTAL (CORRECTED)	4263,41	47			

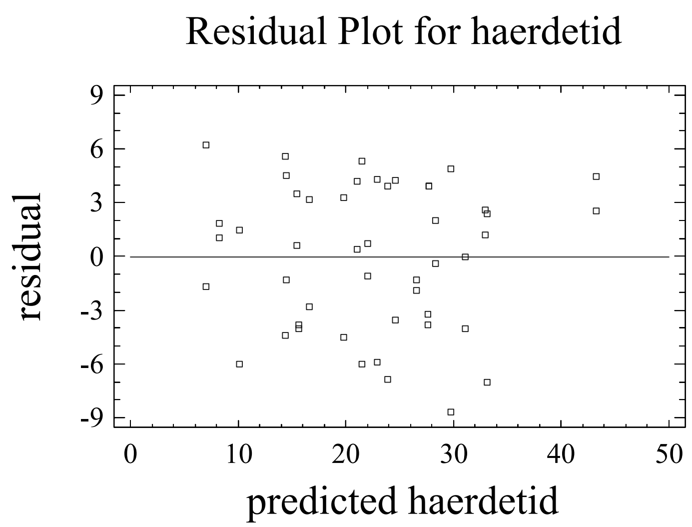
Det ses, at forsøgsfejlens varians (residual) er 19.1289. Endvidere ses, at kun vekselvirkningerne AD og BD har betydning.
Vi "excluderer" nu alle tofaktorvekselvirkningerne, der ikke har betydning (se evt. Appendix A) og får derefter følgende tabel:

```
Analysis of Variance for haerdetid - Type III Sums of Squares
--------------------------------------------------------------------------------
Source               Sum of Squares     Df    Mean Square    F-Ratio    P-Value
--------------------------------------------------------------------------------
MAIN EFFECTS
A:A                        274,946      1         274,946      12,16     0,0015
B:B                        22,7433      2         11,3716       0,50     0,6096
C:C                        49,7761      1         49,7761       2,20     0,1480
D:D                        140,396      3         46,7985       2,07     0,1245
INTERACTIONS
AD                         1009,34      3         336,448      14,88     0,0000
BD                         2065,25      6         344,208      15,22     0,0000
RESIDUAL                    700,959    31          22,6116
--------------------------------------------------------------------------------
TOTAL (CORRECTED)           4263,41    47
--------------------------------------------------------------------------------
All F-ratios are based on the residual mean square error.
```

Det ses, at faktor C ingen rolle spiller, mens A, B og D alle har betydning i form af vekselvirkning.

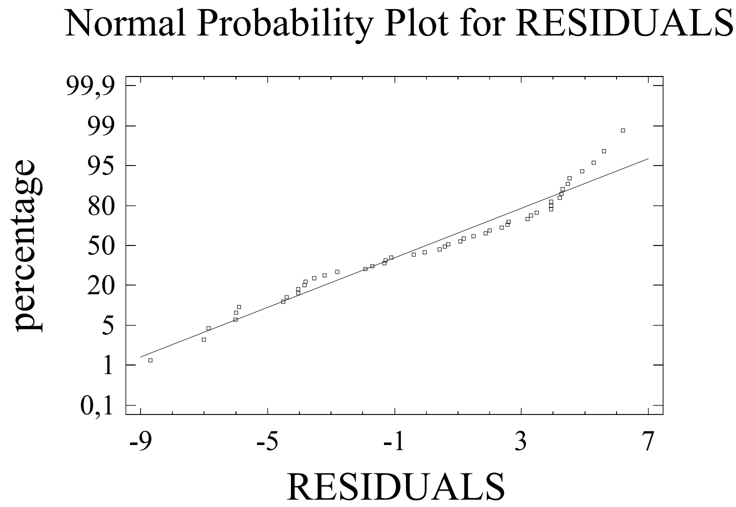
2). Grafisk vurdering af model m.m.

Af principielle hensyn slettes faktor C i modellen (det har i praksis kun meget ringe betydning, da vi jo allerede har 31 frihedsgrader i residualen).
Vælg (rød ikon = Input dialog | Slet C | OK). Dette har den lidt kedelige virkning, at der i visse udskrifter sker en omdøbning, så faktor D bliver kaldt C.
Vi må nu igen “excludere” AB.
For grafisk at vurdere gyldigheden af modellen tegnes et residualplot, “Residuals versus Predicted”.
Resultatet blev:



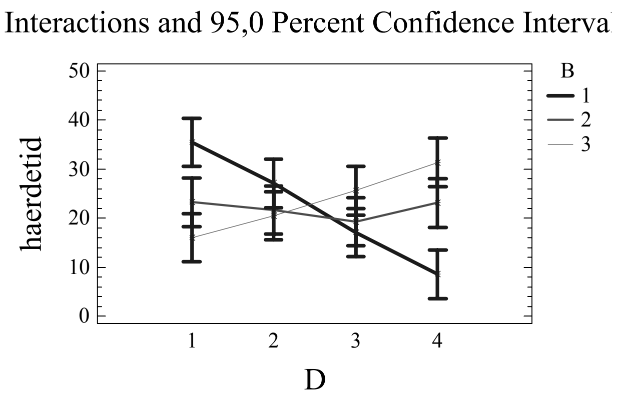
Af figuren fremgår det at residualerne ligger tilfældigt omkring 0-linien
Af variansanalysetabellen fremgår at $\sigma = \sqrt{22,6} \approx 4,8$. Da alle residualerne ligger indenfor $3 \cdot \sigma$ (endda indenfor $2 \cdot \sigma$) er der ingen outliers.
Da vi ikke har gentagelser kan vi ikke vurdere varianshomogenitet.

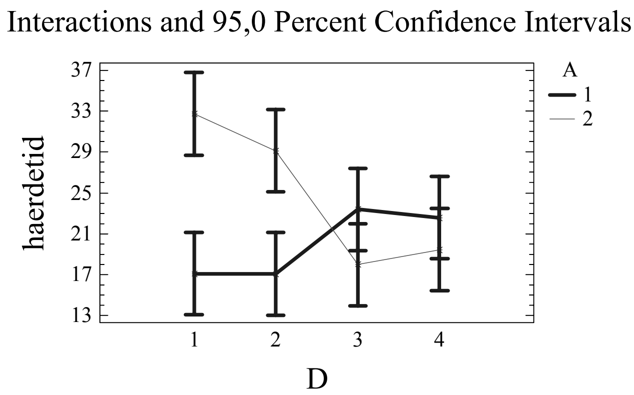
En grafisk test for normalitet, hvor vi har valgt at placere linien mere”ligeligt” ved at vælge vælge “Least Squares” giver:



Støjen synes nogenlunde normalfordelt, da residualerne ligger "symmetrisk og tæt" omkring den rette linie, selv om der dog synes at være nogen afvigelse i højre side, hvor residualerne er blevet lidt for små (punkterne ligger lidt over linien).

3). Angiv de niveauer faktorerne skal stilles på for at give den mindste størkningstid.
For at få et grafisk overblik over vekselvirkningernes betydning , indtastes modellen igen, og der vælges "Interaction Plot"
Vælg (blå ikon = Graphics Options | Interaction Plot |Cursor på figur | højre musetast | Pane options | Confidence intervals | B*D | second Factor" (da D har de fleste niveauer og vi derved får den mest overskuelige figur) | OK) .
Dette giver følgende figur: (Den sorte linie er B_1).
På samme måde fås AD plottet (den sorte linie er A_1).





Af figuren for BD ses, at vælges D på niveau 4 og B på niveau1 bliver størkningstiden mindst(kun et lille overlap med de øvrige konfidensintervaller).
Af figuren for AD ses, at et valg af D på niveau 4 gør, at man frit kan vælge A på såvel højt som lavt niveau (stort overlap på konfidensintervaller.

2) Konklusion: A (omrøringshastighed) vælges frit, B (fabrikationsmetode) på niveau 1, C (formalingsgrad) har ingen betydning og D (råvaretype) på niveau 4.
   Et estimat for den mindste størkningstid kan af tegningen ses at være ca.8.
   Ud fra resultaterne i de 4 celler, der svarer til B_1D_4 , fås ved direkte regning: $\frac{4.12 + 11.6 + 5.30 + 13.2}{4} = 8.55$

◆

Øvelse 1.5. (variansanalyse). Regn opgave 5 side 236.

1.5 Fuldstændigt randomiseret blokforsøg.

I forbindelse med planlægningen af et forsøg, kan man blive tvunget til at benytte forsøgsenheder, som er ret uensartede. Derved får den tilfældige forsøgsfejl en relativ stor spredning (stor "støj"). Dette kan bevirke, at man skal op på et urealistisk stort antal gentagelser for at kunne opnå den ønskede information. For at "dæmpe støjen" kan man inddele forsøgsenhederne i grupper (blokke), hvor de forsøgsenheder der ligger i samme blok er væsentlig mere ensartede end forsøgsenhederne i forskellige blokke. Man siger, at man har et fuldstændigt randomiseret blokforsøg, hvis hver behandling forekommer det samme antal gange (sædvanligvis netop én gang) i hver blok.
Til illustration heraf, så betragter vi igen forsøget beskrevet i eksempel 1.6 og 1.7.

Eksempel 1.12 (randomiseret blokforsøg). *En bilfabrikant ønsker at finde ud af, hvorledes* 3 *olieblandinger O_1, O_2, og O_3, og* 2 *karburatortyper K_1 og K_2 påvirker benzinforbruget. Forsøget planlægges som et fuldstændigt faktorforsøg idet hvert niveau skal gentages mindst 4 gange. Dette betyder at der skal udføres 12 delforsøg.*
Et delforsøg med én bil tager 1 dag.(1 tank = 40 liter: Kører ca 15 km/l så 40 liter = 600 km, hvilket giver ca. 7 timer med 80 km/time). Af tidsmæssige grunde kan man ikke benytte 12 *dage til forsøget. Der benyttes* 2 *biler med tilhørende chauffør, hvilket forkorter forsøgstiden til* 6 *dage.*
Da de to biler (med tilhørende chauffør) kan frygtes at give systematisk forskellige resultater, ønskes foretaget et ***randomiseret blokforsøg*** *med biler som blokke.*
1) *Angiv fordele og ulemper ved at foretage et randomiseret blokforsøg fremfor et fuldstændigt randomiseret forsøg.*
2) *Beskriv hvorledes en randomisering kunne tænkes at foregå.*
3) *Skitser udseendet af en variansanalysetabel med angivelse af frihedsgrader.*

LØSNING:
1) Fordele:Begrundelsen for ikke at foretag et fuldstændigt randomiseret forsøg er, at to biler frygtes at give så stor spredning, at selv betydelige forskelle ikke kan påvises.
   Ved blokforsøget er støjen, der skyldes eventuelle forskelle mellem biler elimineret
   Større mulighed for stabilt vejr i 6 dage end i 12 dage, hvilket også forminsker spredningen.
   Ret få delforsøg,
   Ulempe: Selv på 6 dage kan vejret skifte og give anledning til stor spredning.
2) Randomisering: To dåser mærkes henholdsvis bil1 og bil2. Behandlingen O_1K_1 skrives på 2 sedler som anbringes i hver sin dåse, behandlingen O_1K_2 skrives på 2 sedler som anbringes i hver sin dåse osv. (se figuren).
   Man trækker nu først de 6 sedler fra dåse med mærket bil 1.
   Lad den første seddel der trækkes være O_1K_2. Det betyder nu, at bil 1 skal forsynes med karburator 2 og olieblanding 1 og køre dag 1. Lad den næste seddel der trækkes være O_3K_2

. Det betyder tilsvarende at bil 1 skal forsynes med karburator 2 og olieblanding 3 og køre dag 2. Således fortsættes indtil alle 6 sedler er udtrukket

Resultatet blev:

bil 1	dag 1 O_1K_2	dag 2 O_3K_2	dag 3 O_2K_2	dag 4 O_1K_1	dag 5 O_2K_1	dag 6 O_3K_1

O_1K_1	O_1K_2	O_3K_2
O_2K_1	O_3K_1	O_2K_2

Bil 1

Derefter fortsættes med at trække sedler fra dåsen med mærket bil 2. Resultatet blev:

bil 2	dag 1 O_1K_1	dag 2 O_3K_2	dag 3 O_1K_2	dag 4 O_2K_1	dag 5 O_3K_1	dag 6 O_2K_2

O_3K_1	O_2K_1	O_2K_2
O_1K_2	O_1K_1	O_3K_2

Bil 2

3) Analyse: Tresidet variansanalyse:
Bemærk: Vi antager altid, at **blokke ikke vekselvirker med faktorerne**, idet vi forudsætter, at den ene blok (eksempelvis bil 1) bidrager med en systematisk højere resultat end den anden blok (eksempelvis at bil 1 på alle dage giver et større benzinforbrug end bil 2).

		K_1	K_2
Bil 1	O_1	-	-
	O_2	-	-
	O_3	-	-
Bil 2	O_1	-	-
	O_2	-	-
	O_3	-	-

Variansanalyse	SAK	f
Blokke (biler)		1
Olieblanding		2
Karburator		1
Olie * karburator		2
rest		5
Total		11

Bemærk: Selv om analysen viser, at blokkene mod forventning ikke kan antages at have betydning, må man **ikke poole blokkene ned**, da det svarer til, at man analyserer forsøget som om det var et fuldstændigt randomiseret forsøg. ◆

Øvelse 1.6. (planlægning) Regn opgave 6 side 238.

Eksempel 1.13 (randomiseret blokforsøg).(eksamensopgave juni 83, omskrevet)
I nedenstående tabel er anført resultaterne af et fordingsforsøg med svin. Formålet med forsøget var at undersøge, hvorvidt en ændring af vitaminindholdet i foderet gav en forskel i svinenes vægtforøgelse. Vægtforøgelsen afhænger imidlertid også af det enkelte individs genetiske egenskaber. Et fuldstændigt randomiseret forsøg vil derfor sandsynligvis kunne bevirke, at forsøgsfejlens spredning bliver så stor, at intet kan påvises (forsøget drukner i støj). Da grise fra samme kuld må forventes at være mere ensartede, vælger man at lave et randomiseret blokforsøg med kuld som blokfaktor.
Fra hvert af 4 forskellige kuld grise udtages 3 grise, der bliver fodret med hver sin af tre fodertyper A, B og C med forskelligt vitaminindhold.

Forsøgsresultaterne (vægtforøgelse i kg) var

		Fodertype		
		A	B	C
Kuld	1	7.0	14.0	8.5
	2	16.0	15.5	16.5
	3	10.5	15.0	9.5
	4	13.5	21.0	13.5

Test, om der er nogen væsentlig virkning af ændringen i foderets vitaminindhold.

LØSNING:
Analyseres svarende til en tosidet variansanalyse uden vekselvirkning (da vi forudsætter at blokke ikke vekselvirker med faktorer.
Indtastning af data

```
foder   kuld    vaegt
   A      1       7
   A      2       16
   A      3       10,5
   A      4       13,5
   B      1       14
   B      2       15,5
   B      3       15
   B      4       21
   C      1       8,5
   C      2       16,5
   C      3       9,5
   C      4       13,5
```

Vi går ind i "Compare" og danner følgende variansanalysetabel uden vekselvirkning.

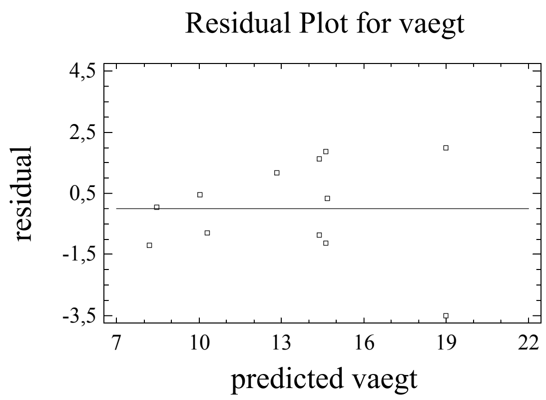
```
Analysis of Variance for vaegt - Type III Sums of Squares
--------------------------------------------------------------------------------
Source                     Sum of Squares     Df    Mean Square    F-Ratio     P-Value
--------------------------------------------------------------------------------
MAIN EFFECTS
 A:foder                           54,125      2        27,0625       5,76      0,0402
 B:kuld                           87,7292      3        29,2431       6,22      0,0285
RESIDUAL                          28,2083      6        4,70139
--------------------------------------------------------------------------------
TOTAL (CORRECTED)                 170,063     11
--------------------------------------------------------------------------------
All F-ratios are based on the residual mean square error.
```

Det ses heraf, at forsøgsfejlens varians (residual) er 4.701.

Vi ser endvidere, at der på et signifikansniveau på 5 % er signifikant forskel på fodertyperne (mindst én afviger fra de øvrige).
Vi ser endvidere, at det var fornuftigt at dele op i kuld, da der også er signifikans for kuld. Vi er imidlertid ikke interesseret i at finde ud af hvilket kuld der er det bedste, da vi jo blot har taget nogle tilfældige kuld ud.

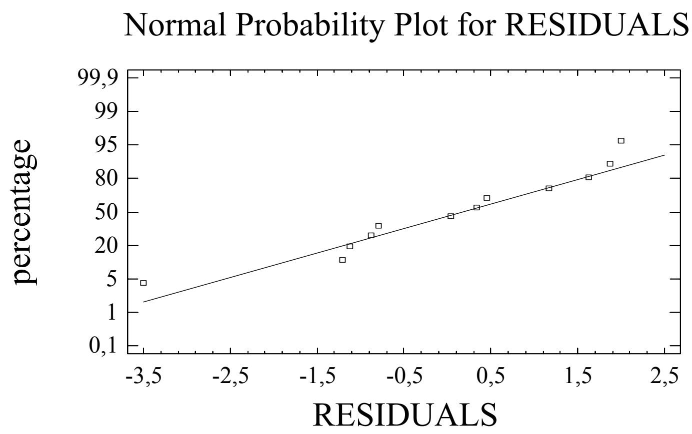
Grafisk kontrol af den fundne model, samt af forudsætninger for analysen:
Vi danner et residualplot: "Residuals versus Predicted"

Residualerne fordeler sig nogenlunde jævnt omkring linien, men der er en tendens til, at større vægt giver større afstand til linien.



Af variansanalysetabellen fremgår at $\sigma \approx \sqrt{4.71} \approx 2.16$. Da alle residualerne ligger indenfor $3 \cdot \sigma$ (endda indenfor $2 \cdot \sigma$) er der ingen outliers.

Grafisk test for normalitet :
Resultatet blev:



Støjen synes tilnærmelsesvis normalfordelt, men antallet af målinger er for lille til, at man kan give en nogenlunde sikker vurdering.

For at finde den fodertype, der giver den højeste vægtforøgelse opstilles konfidensintervaller. Bemærk: Da der er en blokvirkning vil vægtforøgelsen jo afhænge af hvilket kuld man betragter. Konfidensintervaller kan derfor kun anvendes til relative sammenligninger.

Vælges " Tables of Means" fås følgende udskrift :

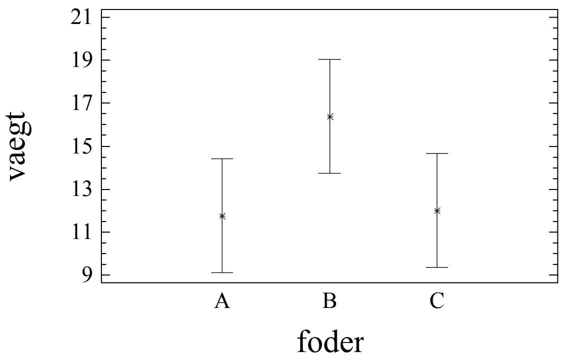
Table of Least Squares Means for vaegt with 95,0 Percent Confidence Intervals

Level	Count	Mean	Stnd. Error	Lower Limit	Upper Limit
GRAND MEAN	12	13,375			
foder					
A	4	11,75	1,08413	9,09721	14,4028
B	4	16,375	1,08413	13,7222	19,0278
C	4	12,0	1,08413	9,34721	14,6528
kuld					
1	3	9,83333	1,25185	6,77016	12,8965
2	3	16,0	1,25185	12,9368	19,0632
3	3	11,6667	1,25185	8,60349	14,7298
4	3	16,0	1,25185	12,9368	19,0632

En tegning af 95% konfidensintervallerne (ikke LSD) giver

Konfidensintervallerne viser ganske vist et svagt overlap, men da variansanalysen har vist at der er en signifikant forskel, må der gælde, at fodertype B giver den største vægtforøgelse.

Means and 95,0 Percent Confidence Intervals



Øvelse 1.7. (randomiseret blokforsøg) Regn opgave 7 side 238.

1.6. Romersk kvadratforsøg

I visse situationer, kan man som det følgende eksempel viser frygte, at der er mere end 1 betydende støjkilde. Et såkaldt romersk kvadratforsøg kan så være en mulighed for at dæmpe støjen.

Eksempel 1.14 (romersk kvadratforsøg) *En bilfabrikant ønsker at finde ud af, hvorledes* 3 *olieblandinger* O_1, O_2, *og* O_3, *og* 2 *karburatortyper* K_1 *og* K_2 *påvirker benzinforbruget.*
Forsøget planlægges som et fuldstændigt faktorforsøg idet hvert niveau skal gentages mindst 4 gange. Dette betyder at der skal udføres 12 delforsøg.
Hvert delforsøg med én bil tager 1 dag. Man frygter, at forskellige dage med deres forskellige vejrforhold kan give stor variation i resultaterne. Endvidere frygter man også, at forskellige biler (med tilhørende chauffør) frygtes give systematisk forskellige resultater.
Da der er i alt 6 forskellige behandlinger ønskes foretaget et ***romersk kvadratforsøg*** *med anvendelse af 6 biler og 6 dage som blokke.*

1) *Angiv fordele og ulemper ved at foretage et romersk kvadratforsøg fremfor et fuldstændigt randomiseret blokforsøg.*
2) *Beskriv hvorledes en randomisering kunne tænkes at foregå.*
3) *Skitser udseendet af en variansanalysetabel med angivelse af frihedsgrader.*

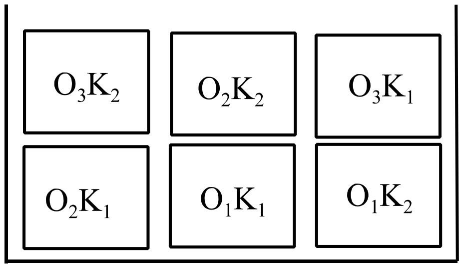
LØSNING:

1) Fordele: Støj der kan skyldes eventuelle forskelle mellem biler og mellem dage elimineres.
   Ulempe: Mange (36) delforsøg, og en ret kompliceret forsøgsplan.

2) Randomisering:
Først vælges et 6 * 6 standardkvadrat (f.eks ét, hvor bogstaverne A,B,C,D,E,F forskydes cyklisk for hver ny række)
Derefter ombyttes rækkerne "randomiseret". Dette sker ved at lægge sedler med tallene 1,2,3,4,5,6 i en dåse, og foretage lodtrækning. Trækkes først nummer 2 betyder det at række 1 i kvadratet bliver til række 2. Er resultatet 2 4 3 6 1 5 fås derfor følgende nye kvadrat:

Bil \ Dag	1	2	3	4	5	6		1	2	3	4	5	6
1	A	B	C	D	E	F	række 1 til 2	C	D	E	F	A	B
2	F	A	B	C	D	E	række 2 til 4	A	B	C	D	E	F
3	E	F	A	B	C	D	række 3 til 3	E	F	A	B	C	D
4	D	E	F	A	B	C	række 4 til 6	F	A	B	C	D	E
5	C	D	E	F	A	B	række 5 til 1	B	C	D	E	F	A
6	B	C	D	E	F	A	række 6 til 5	D	E	F	A	B	C

På samme måde randomiseres søjler. Resultat: 3 4 6 1 5 2
Endelig erstattes bogstaverne A,B,C,D,E,F med behandlinger, ved at anbringe 6 sedler med behandlingerne i en dåse, (se figuren). Trækkes først O_3K_2 betyder det, at bogstavet A erstattes af O_3K_2. Den næste udtrukne seddel erstatter bogstavet B osv.
Idet resultatet af lodtrækningen blev O_3K_2, O_3K_1 , O_1K_2 , O_2K_1 , O_2K_2, O_1K_1 fås følgende:



	Bil \ Dag	1	2	3	4	5	6		1	2	3	4	5	6
søjle 1 til 3	1	F	B	C	D	A	E	A til O_3K_2	O_1K_1	O_3K_1	O_1K_2	O_2K_1	O_3K_2	O_2K_2
søjle 2 til 4	2	D	F	A	B	E	C	B til O_3K_1	O_2K_1	O_1K_1	O_3K_2	O_3K_1	O_2K_2	O_1K_2
søjle 3 til 6	3	B	D	E	F	C	A	C til O_1K_2	O_3K_1	O_2K_1	O_2K_2	O_1K_1	O_1K_2	O_3K_2
søjle 4 til 1	4	C	E	F	A	D	B	D til O_2K_1	O_1K_2	O_2K_2	O_1K_1	O_3K_2	O_2K_1	O_3K_1
søjle 5 til 5	5	E	A	B	C	F	D	E til O_2K_2	O_2K_2	O_3K_2	O_3K_1	O_1K_2	O_1K_1	O_2K_1
søjle 6 til 2	6	A	C	D	E	B	F	F til O_1K_1	O_3K_2	O_1K_2	O_2K_1	O_2K_2	O_3K_1	O_1K_1

3) Analyse: Firesidet variansanalyse, blokke vekselvirker ikke med faktorerne.
Variansanalysetabel:

Variansanalyse	SAK	f
Blok (biler)		5
Blok(dage)		5

Olieblanding		2
Karburator		1
Olie * karburator		2
rest		20
Total		35

◆

Øvelse 1.8. (planlægning) Regn opgave 9 side 239.

Eksempel 1.15 (romersk kvadratforsøg) (eksamensopgave juni 72, omskrevet)
Der ønskes undersøgt hvorledes udbyttet af et biprodukt ved en fabrikationsproces afhænger af fabrikationsmetode og katalysatortype. Der skal vælges mellem 2 forskellige fabrikations-metoder, og 3 katalysatorerer. Det kan ikke udelukkes, at de to faktorer vekselvirker, så man vælger som forsøgsplan en fuldstændig faktorstruktur.
Lad A_i betegne fabrikationsmetode i, og lad B_j betegne katalysator j. Der er følgelig 6 behandlinger $A_1 B_1$, $A_1 B_2$, $A_1 B_3$, $A_2 B_1$, $A_2 B_2$, $A_2 B_3$.
Hvert delforsøg varer en uge, og for at forsøgene ikke skal tage for lang tid, er man nødt til at anvende flere apparater samtidig. Da det kan tænkes, at såvel fabrikationstidspunkt som apparatvalg kan påvirke udbyttet, og dermed give en for stor spredning, vælges som forsøgs-plan et romersk kvadratforsøg. Da der er 6 behandlinger bliver det et 6×6 kvadrat.
Forsøgsplan og forsøgsresultater var følgende:

		Apparater											
		1		2		3		4		5		6	
Uger	1	$A_2 B_1$	41.3	$A_2 B_2$	52.6	$A_1 B_2$	39.7	$A_2 B_3$	40.8	$A_1 B_3$	27.6	$A_1 B_1$	29.5
	2	$A_1 B_1$	9.3	$A_1 B_3$	18.7	$A_2 B_3$	19.5	$A_1 B_2$	21.1	$A_2 B_2$	31.9	$A_2 B_1$	33.2
	3	$A_2 B_3$	30.5	$A_2 B_1$	41.2	$A_1 B_3$	28.7	$A_2 B_2$	32.4	$A_1 B_1$	31.9	$A_1 B_2$	32.2
	4	$A_1 B_3$	50.7	$A_1 B_2$	48.3	$A_2 B_1$	51.6	$A_1 B_1$	40.9	$A_2 B_3$	60.4	$A_2 B_2$	49.8
	5	$A_2 B_2$	40.8	$A_2 B_3$	39.6	$A_1 B_1$	30.7	$A_2 B_1$	31.8	$A_1 B_2$	29.6	$A_1 B_3$	39.8
	6	$A_1 B_2$	42.7	$A_1 B_1$	52.4	$A_2 B_2$	62.1	$A_1 B_3$	53.6	$A_2 B_1$	63.6	$A_2 B_3$	58.7

Foretag en statistisk analyse af forsøget, idet man ønsker så lavt et udbytte som muligt.

LØSNING:
Analyseres svarende til en firesidet variansanalyse uden vekselvirkning mellem blokke og faktorer. Ordrer m.m. svarer derfor ganske til de i eksempel 1.12 anførte.
Indtastning af data sker som beskrevet i appendix A

Starten af indtastningen i regnearket ses nedenfor:
```
metode  katalysator  uge  apparat udbytte
   a2      b1         1       1    41,3
```

```
   a2      b2         1       2    52,6
   a1      b2         1       3    39,7
   a2      b3         1       4    40,8
   a1      b3         1       5    27,6
   a1      b1         1       6    29,5
   a1      b1         2       1     9,3
. . . . . . . . . . . . . . . .
```

Opstilling af variansanalysetabel.
I “Compare” vælges “Anova Tables”, hvorved fremkommer en variansanalysetabel uden vekselvirkning. Da vi ønsker at undersøge, om der er vekselvirkning, sættes “ Maximum Order Interaction til 2 “
Der fremkommer nu en kommentar: `No output due to data error.`
Dette er jo meget sandt, da der ikke er gentagelser i modellen.
Da blokkene ikke vekselvirker med faktorerne foretages følgende:
Vælg (Cursor i tabellen | højre musetast |Analysis options | Exclude) og “excluder” alle tofaktorvekselvirkningerne, pånær vekselvirkningen mellem katalysator og metode. Vi får derefter følgende tabel:

```
Analysis of Variance for udbytte - Type III Sums of Squares
--------------------------------------------------------------------------------
Source                 Sum of Squares     Df    Mean Square    F-Ratio    P-Value
--------------------------------------------------------------------------------
MAIN EFFECTS
 A:uge                        4385,54      5        877,108      29,02     0,0000
 B:apparat                    181,119      5        36,2238       1,20     0,3454
 C:katalysat                  27,8956      2        13,9478       0,46     0,6369
 D:metode                     662,204      1        662,204      21,91     0,0001

INTERACTIONS
 CD                           61,4756      2        30,7378       1,02     0,3796

RESIDUAL                      604,436     20        30,2218
--------------------------------------------------------------------------------
TOTAL (CORRECTED)             5922,67     35
--------------------------------------------------------------------------------
All F-ratios are based on the residual mean square error.
```

Det ses heraf, at der ikke er vekselvirkning mellem C og D.
Vi sætter nu “ Maximum Order Interaction til 1 “ og får derefter følgende tabel:

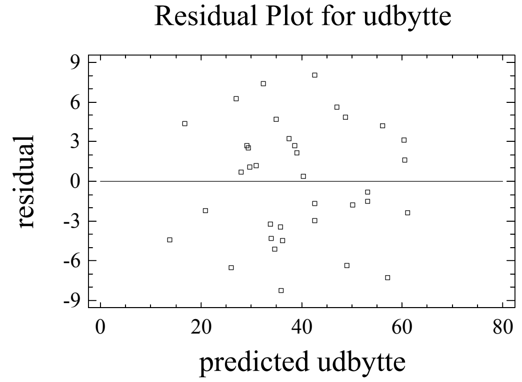
```
Analysis of Variance for udbytte - Type III Sums of Squares
--------------------------------------------------------------------------------
Source                 Sum of Squares     Df    Mean Square    F-Ratio    P-Value
--------------------------------------------------------------------------------
MAIN EFFECTS
 A:uge                        4385,54      5        877,108      28,98     0,0000
 B:apparat                    181,119      5        36,2238       1,20     0,3431
 C:katalysat                  27,8956      2        13,9478       0,46     0,6367
 D:metode                     662,204      1        662,204      21,88     0,0001
RESIDUAL                      665,911     22        30,2687
--------------------------------------------------------------------------------
TOTAL (CORRECTED)             5922,67     35
--------------------------------------------------------------------------------
All F-ratios are based on the residual mean square error.
```

Det ses, at faktor C (katalysator) ingen rolle spiller, mens metoden D har stor betydning

Det ses endvidere, at det var fornuftigt at foretage en blokopdeling efter uge, mens det var unødigt at dele op efter apparater.

Spørgsmål 2. Grafisk vurdering af model m.m.
For grafisk at vurdere gyldigheden af modellen tegnes et residualplot: “Residuals versus

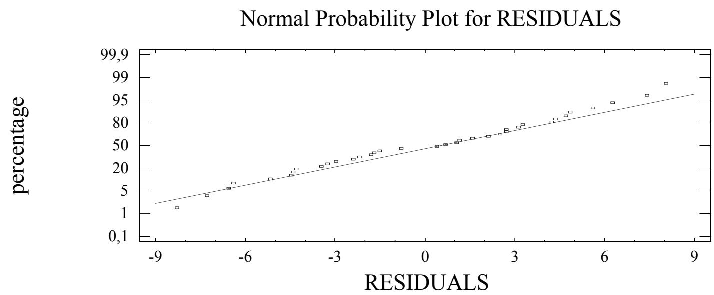
Predicted "
Resultatet blev:



<u>Af figuren fremgår det at residualerne ligger tilfældigt omkring 0-linien</u>
Af variansanalysetabellen fremgår at $\sigma = \sqrt{30.26} \approx 5{,}5$. Da alle residualerne ligger indenfor $3 \cdot \sigma$ (endda indenfor $2 \cdot \sigma$) er der <u>ingen outliers.</u>
Da vi ikke har gentagelser kan vi ikke vurdere varianshomogenitet.

Grafisk test for normalitet:

<u>Støjen synes nogenlunde normalfordelt</u>, da residualerne ligger "symmetrisk og tæt" omkring den rette linie.



For at finde den metode, der giver mindst udbytte, så vend tilbage til den endelige model, og vælger " Tables of Means"
Det giver følgende udskrift:

```
Table of Least Squares Means for udbytte
with 95,0 Percent Confidence Intervals
--------------------------------------------------------------------------------
                                              Stnd.       Lower       Upper
Level                   Count     Mean        Error       Limit       Limit
--------------------------------------------------------------------------------
GRAND MEAN              36        39,1444
apparat
1                       6         35,8833     2,19502     31,353      40,4136
2                       6         42,1333     2,19502     37,603      46,6636
```

```
3                    6         38,7167       2,19502      34,1864      43,247
4                    6         36,7667       2,19502      32,2364      41,297
5                    6         40,8333       2,19502      36,303       45,3636
6                    6         40,5333       2,19502      36,003       45,0636
uge
1                    6         38,5833       2,19502      34,053       43,1136
2                    6         22,2833       2,19502      17,753       26,8136
3                    6         32,8167       2,19502      28,2864      37,347
4                    6         50,2833       2,19502      45,753       54,8136
5                    6         35,3833       2,19502      30,853       39,9136
6                    6         55,5167       2,19502      50,9864      60,047
metode
a1                   18        34,8556       1,26729      32,24        37,4711
a2                   18        43,4333       1,26729      40,8178      46,0489
--------------------------------------------------------------------------------
```

Det ses, at metode 1 giver det laveste udbytte

Øvelse 1.9. (romersk kvadratforsøg) Regn opgave 11 side 239.

2 2^k FAKTORFORSØG

2.1 SCREENINGS-FORSØG

Har man mange faktorer der skal undersøges, må man for at begrænse antallet af delforsøg nøjes med at undersøge hver faktor på 2 niveauer "lavt" niveau og "højt" niveau. Disse niveauer forsøges valgt som "yderpunkter" med hensyn til faktorernes formodede virkning. Har man ingen formodning herom, vælges sædvanligvis to faktorniveauer, som "afviger" mest muligt.
Ved forsøget får man så afklaret, hvilke af faktorerne der har en væsentlig virkning. De sædvanligvis få faktorer, som viser sig at have en virkning, kan man så studere nøjere ved supplerende forsøg. Der gælder igen den regel, at man bør højst bruge 25% af det planlagte budget på det første forsøg. Samtidig er det fordelen ved disse screeningsforsøg, at der med forholdsvis få delforsøg kan undersøges mange faktorer. Har man derfor i den gruppe, der planlægger forsøget, en rimelig tro på at en faktor kan tænkes at spille en rolle, bør den medtages. Samtidig er det dog vigtigt, at man i gruppen søger at opstille faktorerne i en rangorden A, B, C, osv. efter den effekt, man tror de har. De faktorer man tror har stor effekt gives de første bogstaver i alfabetet.
Det har stor betydning for forsøgets succes, at man er meget omhyggelig med at sammensætte den gruppe af mennesker, som skal planlægge og udføre forsøget. Hvis det drejer sig om udvikling af et nyt produkt vil det ofte være udviklingsafdelingen der både planlægger og udfører forsøget. Hvis det drejer sig om et problem i en eksisterende produktion, som måske er koncentreret om en bestemt maskine (eller del af produktionen), er det vigtigt at gruppen har medlemmer (driftsingeniører, værkførere, opstillere), som arbejder tæt på maskinen (den del af produktionen) til daglig. Det sikrer også, at de forstår ideen i forsøget, og derved sikrer at forsøget bliver udført som planlagt.
Har man k faktorer, vil med en fuldstændig faktorstruktur antallet af behandlinger være 2^k, deraf navnet et 2^k faktorforsøg.

2.2 FORKLARING PÅ NOMENKLATUR OG FORMLER UD FRA ET TALEKSEMPEL

En særdeles bekvem notation i forbindelse med fuldstændige 2^k faktorforsøg er følgende:
(1): Alle faktorer er på lavt niveau
a: A på højt niveau, alle andre på lavt niveau.
b: B på højt niveau, alle andre på lavt niveau.
ab: A og B på højt niveau, alle andre på lavt niveau.
c: C på højt niveau, alle andre på lavt niveau.
ac: A og C på højt niveau, alle andre på lavt niveau
osv.
Nomenklaturen anvendes i en dobbelt betydning, idet eksempelvis a også betyder summen af forsøgsresultaterne med denne behandling.

Faktorerne A, B, C ... på lavt niveau benævnes A_1, B_1, C_1. . . og på højt niveau A_2, B_2, C_2 ...

Eksempel 2.1 Illustration af nomenklatur ud fra et 2^2 faktorforsøg.

Et forsøg med 2 faktorer A og B hver på 2 niveauer og med 3 gentagelser af hver behandling har resulteret i de i skemaet anførte resultater. Der er endvidere anført et sumskema og et skema over gennemsnit hvor den anvendte nomenklatur er indføjet.

Forsøgsresultater

	B_1	B_2
A_1	11 9 7	13 10 10
A_2	2 0 1	8 7 6

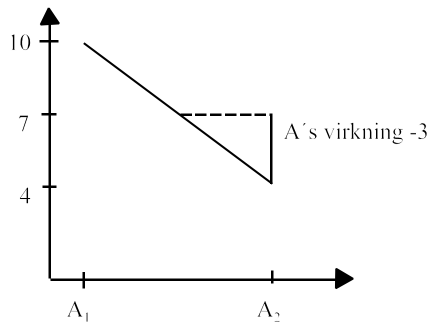
Sumskema

	B_1	B_2	SUM
A_1	(1) 27	b 33	60
A_2	a 3	ab 21	24
SUM	30	54	84

Gennemsnit

	B_1	B_2	
A_1	$\bar{x}_{11}$ 9	$\bar{x}_{12}$ 11	$\bar{x}_{1.}$ 10
A_2	$\bar{x}_{21}$ 1	$\bar{x}_{22}$ 7	$\bar{x}_{2.}$ 4
	$\bar{x}_{.1}$ 5	$\bar{x}_{.2}$ 9	$\bar{x}_{..}$ 7

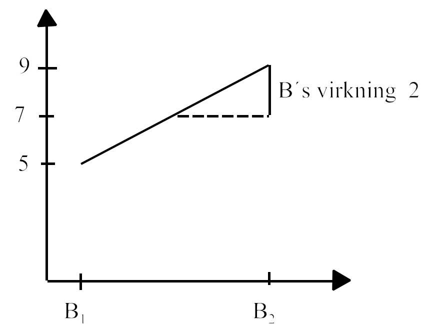
Definition af virkning: Ved A_i's virkning forstås effekten af en ændring ud fra gennemsnitsniveauet til niveauet, når A er på niveau i $(\bar{\mu}_{i.} - \bar{\mu}_{..})$.



Ved **A's hovedvirkning** forstås effekten af en ændring ud fra gennemsnitsniveauet til niveauet, når A er på højt niveau, dvs. A = A_2 (A's hovedvirkning = A's virkning på højt niveau.

Et estimat (skøn) for A's hovedvirkning er i eksempel 2.1 $\tilde{A} = 4 - 7 = -3$. (se figuren).

Bemærk: Et estimat for A skrives kort $\tilde{A}$ osv. (der sættes en "tilde" over bogstavet).



B's hovedvirkning defineres analogt (B's virkning når B er på højt niveau).

Et estimat (skøn) for B's hovedvirkning er i eksempel 2.1 $\tilde{B} = 9 - 7 = 2$. (se figuren)

Udledning af generelle formler for hovedvirkninger ud fra eksempel 2.1.

Vi har $$\tilde{A} = \frac{4-10}{2} = \frac{\frac{24-60}{2\cdot 3}}{2} = \frac{3+21-(27+33)}{2^2\cdot 3} = \frac{-(1)+a-b+ab}{2^k\cdot n}$$

Tilsvarende ses $$\tilde{B} = \frac{9-5}{2} = \frac{\frac{54-30}{2\cdot 3}}{2} = \frac{33+21-(3+27)}{2^2\cdot 3} = \frac{-(1)-a+b+ab}{2^k\cdot n}$$

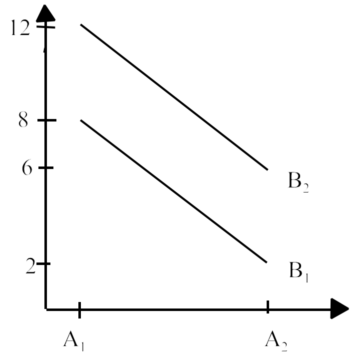
Additiv model.

Definition: En model kaldes additiv, hvis der for alle i,j gælder

$\mu_{ij} = \bar{\mu}.. + A_i + B_j = \mu_i. + \mu._j - \bar{\mu}..$ (de tilsvarende linier er parallelle)

Et "eksakt" eksempel på en additiv model, er

	B_1	B_2	
A_1	$\bar{x}_{11}$ 8	$\bar{x}_{12}$ 12	$\bar{x}_1.$ 10
A_2	$\bar{x}_{21}$ 2	$\bar{x}_{22}$ 6	$\bar{x}_2.$ 4
	$\bar{x}._1$ 5	$\bar{x}._2$ 9	$\bar{x}..$ 7



Model med vekselvirkning

En model der ikke er additiv siges at have vekselvirkning og med 2 faktorer kan modellen i så fald skrives $\mu_{ij} = \bar{\mu}.. + A_i + B_j + AB_{ij}$,

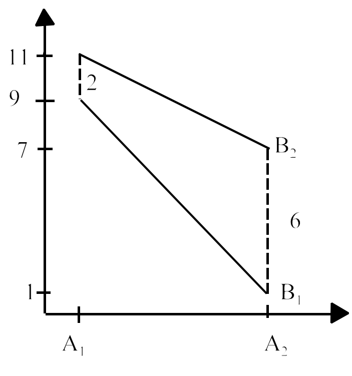
Ved AB's **vekselvirkning** forstås AB_{22} (begge på højt niveau).

Betragtes tallene fra eksempel 2.1 fås et estimat for vekselvirkningen:

$\overline{AB} = \mu_{22} - (\bar{\mu}.. + \tilde{A} + \tilde{B}) = 7 - (7 + (-3) + 2) = 1$.

Vekselvirkningen er et udtryk for hvor " lidt parallelle" linierne er, idet vi har (se figuren)

$\overline{AB} = \frac{6-2}{2^2} = 1$



Udledning af generel formel for vekselvirkning:

$$\overline{AB} = \frac{6-2}{2^2} = \frac{7-1-(11-7)}{2^2} = \frac{\frac{21-3}{3} - \frac{33-27}{3}}{2^2} = \frac{(ab-a)-(b-(-1))}{3\cdot 2^2} = \frac{(1)-a-b+ab}{2^k \cdot n}$$

Fortegnsmatrix: Et vigtigt hjælpemiddel er de såkaldte fortegnsmatricer:

	I	A	B	AB
(1)	+	−	−	+
a	+	+	−	−
b	+	−	+	−
ab	+	+	+	+

A og B søjlerne har +, hvis det til faktoren svarende lille bogstav indgår, ellers −

Fortegnet i AB-søjlen er "produktet" af tilsvarende fortegn i A og B søjle.

For at beregne A's virkning betragtes søjlen under A. Tælleren er så netop tallene i venstre kolonne med det dertil svarende fortegn.

Eksempel : $\widetilde{A} = \dfrac{-(1) + a - b + ab}{3 \cdot 2^2}$ (nævner er altid det totale antal delforsøg).

Kendes et estimat for virkningerne, kan man beregne et estimat for middelværdien f.eks. for A på højt niveau og B på lavt niveau ved i fortegnsmatricen at se på rækken ud for a.

$$\widetilde{\mu}_a = \widetilde{\mu}.. + \widetilde{A} - \widetilde{B} - \overline{AB}$$

Analogt for de andre virkninger.

Resultaterne kan generaliseres til et vilkårligt antal faktorer.

Analyse: Ved den statistiske analyse er det ofte nødvendigt at anvende følgende formler:

$$\widetilde{\mu}_a = \widetilde{\mu}.. + \widetilde{A} - \widetilde{B} - \overline{AB} \tag{1}$$

$$V(\widetilde{\mu}_a) = V(\bar{\mu}..) + V(\widetilde{A}) + V(\widetilde{B}) + V(\overline{AB}) \tag{2}$$

og $V(\widetilde{\mu}_a) = V(\bar{\mu}..) = V(\widetilde{A}) = V(\widetilde{B}) = V(\overline{AB}) = \dfrac{\sigma^2}{\text{Antal delforsøg}}$. (3)

Disse formler kan umiddelbart generaliseres til forsøg med flere faktorer.
Såfremt vi efter en testning antager, at en virkning, f. eks. B er lig med nul, sættes formelt $\widetilde{B} = 0$ og $V(\widetilde{B}) = 0$.

Hvis man ønsker at finde SAK for en virkning T, så beregnes den af formlen
SAK_T=(estimat for virkning af T) · (antal delforsøg).
Eksempelvis er $SAK_A = (-3)^2 \cdot 12 = 108$ (jævnfør den følgende variansanalysetabel).

Eksempel 2.1 (fortsat)

Det oplyses, at variansanalysetabellen i dette eksempel er:

	B_1	B_2
A_1	× ××	× ××
A_2	× ××	× ××

Variansanalyse	SAK	f	s^2	F	P-værdi
A	108	1	108	48	
B	48	1	48	21.33	
AB	12	1	12	5.33	0.0497
Gentagelser	18	8	s_o^2 =2.25		
Total		11			

Angiv på basis af tidligere beregninger samt ovenstående variansanalysetabel

1) *Hvilke faktorer der har virkning, samt hvilke niveauer disse skal have, for at man får størst mulig middelværdi. Angiv endvidere et estimat for denne middelværdi.*
2) *Angiv et konfidensinterval for den fundne maksimale middelværdi.*
3) *Undersøg om der er signifikant forskel mellem den største og den næststørste middelværdi, ved at lave et 95% konfidensinterval for differensen.*

LØSNING:

1) Da *P* - værdi er under 5% antages, at AB er forskellig fra nul.
   Faktorerne A og B har derfor en virkning i form af en vekselvirkning.
   Da vi tidligere har fundet $\widetilde{A} = -3$, $\widetilde{B} = 2$ og $\overline{AB} = 1$ må den maksimale værdi forekommer for A på lavt niveau og B på højt niveau.
   Et estimat for den maksimale værdi er
   $\widetilde{\mu}_b = \bar{x} - \widetilde{A} + \widetilde{B} - \overline{AB} = 7 - (-3) + 2 - 1 = \underline{\underline{11}}$.

2) Ifølge formlerne (1), (2) og (3) fås : $V(\widetilde{\mu}_b) = 4 \cdot \frac{\sigma^2}{2^2 \cdot 3} = \frac{\sigma^2}{3}$

   Idet $s_0^2 = 2.25$ fås $V(\widetilde{\mu}_b) = \frac{s_0^2}{3} = \frac{2.25}{3} = 0.75$. Da $f_0 = 8$ frihedsgrader er et

   95% konfidensinterval for $\widetilde{\mu}_b$: $[\widetilde{\mu}_b \pm t_{0.975}(8)\sqrt{V(\widetilde{\mu}_b)} = 11 \pm 2.31 \cdot \sqrt{0.75} = 11 \pm 2.00$

   $\underline{\underline{[9.00 ; 13.00]}}$

3) Det ses, at den næststørste værdi er $\widetilde{\mu}_{(1)}$.

   For at undersøge om der er en signifikant forskel mellem den største værdi $\widetilde{\mu}_b$ og den næststørste værdi $\widetilde{\mu}_{(1)}$ beregnes et 95% konfidensinterval for differensen $\widetilde{\mu}_b - \widetilde{\mu}_{(1)}$.

   Idet $\widetilde{\mu}_b = \bar{x} - \widetilde{A} + \widetilde{B} - \overline{AB}$ og $\widetilde{\mu}_{(1)} = \bar{x} - \widetilde{A} - \widetilde{B} + \overline{AB}$ er

   $$\widetilde{\mu}_b - \widetilde{\mu}_{(1)} = 2 \cdot \widetilde{B} - 2 \cdot \overline{AB} = 2$$

   Vi har derfor

   $$V(\widetilde{\mu}_b - \widetilde{\mu}_{(1)}) = 2^2 \cdot V(\widetilde{B}) + 2 \cdot V(\overline{AB}) = 2^2 \cdot 2 \cdot \frac{s_0^2}{2^2 \cdot 3} = 2 \cdot \frac{s_0^2}{3} = 1.50$$

95% konfidensinterval for $\tilde{\mu}_b - \tilde{\mu}_{(1)}$:

$$\tilde{\mu}_b - \tilde{\mu}_{(1)} \pm t_{0.975}(8)\sqrt{V(\tilde{\mu}_b - \tilde{\mu}_{(1)})} = 2 \pm 2.31 \cdot \sqrt{1.50} = 2 \pm 2.83 \qquad [-0.83\,;\,4.83]$$

Da konfidensintervallet indeholder 0 er der ikke signifikant forskel på den største og den næststørste værdi.

Skulle en statistisk analyse vise, at nogle af virkningerne må antages at være 0, kan de blot udelades af ovenstående udtryk. Er eksempelvis AB = 0 fås tilsvarende:

$$\tilde{\mu}_a = \bar{\mu}.. + \tilde{A} - \tilde{B} \quad \text{og} \quad \mathrm{V}(\tilde{\mu}_a) = 3 \cdot \frac{\sigma^2}{2^2 \cdot 3} = \frac{\sigma^2}{4}$$

Et 95% konfidensinterval for $\tilde{\mu}_a$: $\left[\tilde{\mu}_a - t_{0.975}(9)\frac{s_0}{\sqrt{4}}\,;\,\tilde{\mu}_a + t_{0.975}(9)\frac{s_0}{\sqrt{4}}\right]$

◆

2.3 FULDSTÆNDIGT 2^k FAKTORFORSØG

2.3.1 Fuldstændigt 2^4 faktorforsøg regnet med "Experimental Design".

I appendix A er anført de mest benyttede ordrer. Den følgende fremstilling vil derfor være baseret på, at man der kan finde de detaljerede ordrer.

Eksempel 2.2. (fuldstændigt randomiseret 2^4 faktorforsøg)

Udbyttet ved en biokemisk fabrikation kunne tænkes at afhænge af, om der blev tilsat en bestemt stabilisator (faktor A), om der blev tilsat et bestemt konserveringsmiddel (faktor B), hvilken syre der blev tilsat (faktor C) for at opnå den for fabrikationsprocessen nødvendige surhedsgrad, samt af driftstemperaturen (faktor D). Faktorniveauerne var

A: Ingen stabilisator (lavt niveau) — Stabilisator (højt niveau)
B: Intet konserveringsmiddel (lavt niveau) — Konserveringsmiddel (højt niveau)
C: Syrekoncentration (lavt niveau) — Syrekoncentration (højt niveau)
D: Temperatur (lavt niveau) — Temperatur (højt niveau)

Forsøgsplanen var et fuldstændig randomiseret 2^4 faktorforsøg uden gentagelser.

1) *Lav ved hjælp af Statgraphics en forsøgsplan med randomisering.*
Forsøgsresultaterne (% udbytte) var følgende (i randomiseringsrækkefølge):

(1) 31.54	abc 32.34	cd 6.85	bd 21.13	ab 30.85	d 24.32	a 29.06	acd 46.09
bcd 7.38	b 35.21	abcd 47.54	ad 42.20	ac 24.29	bc 22.54	c 25.94	abd 48.81

2) *Find hvilke faktorer der har virkning.*
3) *Vurder grafisk i den udstrækning det er muligt ud fra residualerne i den under punkt 2 fundne model om der er outliers, om der er varianshomogenitet, og om "støjen" er normalfordelt.*
4) *Angiv de niveauer de pågældende faktorer skal indstilles på, for at give det største middeludbytte.*
5) *Angiv et estimat og et 95% konfidensinterval for dette største middeludbytte.*

6) *Af økonomiske grunde ønskes undersøgt, om temperaturen D på lavt niveau, vil give en væsentlig ændring i udbyttet. Angiv et estimat og et 95% konfidensinterval for differensen mellem det i spørgsmål 4 fundne største middeludbytte og det største middeludbytte, der fås hvis temperaturen holdes på lavt niveau.*

LØSNING:
Spørgsmål 1 Forsøgsplan konstrueres i "Experimental Design"
Man vælger " Create Design" og sætter "No of response" = 1 og "No of experimental Factor"=4. Vi vælger at beholde Statgraphics standardnavne.
Da vi har et fuldstændigt 2^4 faktorforsøg uden blokke, vælges den øverste linie blandt valgmulighederne, og man fjerner krydset ved "Display Blocked Design" .
Vi fastholder krydset ved Randomize, da vi ønsker at rækkefølgen af forsøgene tilfældig, sådan som det bør være, hvis man selv skulle lave forsøgene (randomiseret).
Der ses nu en plan med Statgraphics standardnavne såsom -1 for lav og 1 for højt niveau.
Vælg (gule ikon = Tabular Options | Worksheet | OK)
Herved fås en arbejdsplan hvor behandlingerne er skrevet i tilfældig (randomiseret) rækkefølge.
I dette tilfælde var arbejdsplanen følgende:

run	Factor_A	Factor_B	Factor_C	Factor_D	Var_1
1	-1,0	-1,0	-1,0	-1,0	____________
2	1,0	1,0	1,0	-1,0	____________
3	-1,0	-1,0	1,0	1,0	____________
4	-1,0	1,0	-1,0	1,0	____________
5	1,0	1,0	-1,0	-1,0	____________
6	-1,0	-1,0	-1,0	1,0	____________
7	1,0	-1,0	-1,0	-1,0	____________
8	1,0	-1,0	1,0	1,0	____________
9	-1,0	1,0	1,0	1,0	____________
10	-1,0	1,0	-1,0	-1,0	____________
11	1,0	1,0	1,0	1,0	____________
12	1,0	-1,0	-1,0	1,0	____________
13	1,0	-1,0	1,0	-1,0	____________
14	-1,0	1,0	1,0	-1,0	____________
15	-1,0	-1,0	1,0	-1,0	____________
16	1,0	1,0	-1,0	1,0	____________

Forsøgene udføres (randomiseret) og indskrives på arbejdsarket.
I dette eksempel fremgår resultaterne af opgaveteksten.

Spørgsmål 2. Find de faktorer der har virkning.

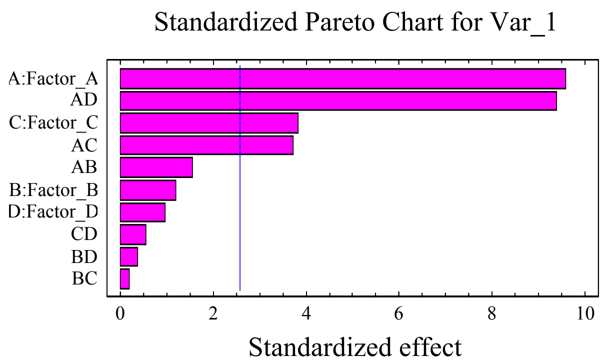
Indtastning af Data.
Vi går nu ned i bunden af skærmen og vælger et regneark “Untitled”. Vi ser at regnearkets faktorer allerede er udfyldt, så vi indtaster blot udbytterne i VAR_1.
Resultatet bliver følgende:

Block	Factor_A	Factor_B	Factor_C	Factor_D	VAR_1
1	-1,0	-1,0	-1,0	-1,0	31,54
1	1,0	1,0	1,0	-1,0	32,34
1	-1,0	-1,0	1,0	1,0	6,85
1	-1,0	1,0	-1,0	1,0	21,13
1	1,0	1,0	-1,0	-1,0	30,85
1	-1,0	-1,0	-1,0	1,0	24,32
1	1,0	-1,0	-1,0	-1,0	29,06
1	1,0	-1,0	1,0	1,0	46,09
1	-1,0	1,0	1,0	1,0	7,38
1	-1,0	1,0	-1,0	-1,0	35,21
1	1,0	1,0	1,0	1,0	47,54
1	1,0	-1,0	-1,0	1,0	42,20
1	1,0	-1,0	1,0	-1,0	24,29
1	-1,0	1,0	1,0	-1,0	22,54
1	-1,0	-1,0	1,0	-1,0	25,94
1	1,0	1,0	-1,0	1,0	48,81

Analyse af data
Vi vælger Analyze Design , og trykker på VAR_1 og derefter på ⇒ under “Data” for at fortælle, at det er denne variabel, der skal analyseres.

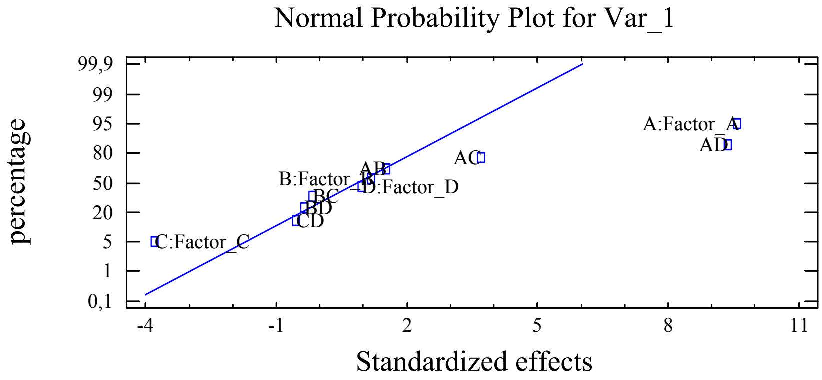
Dette resulterer i følgende tabel over virkningerne, samt et histogram over dem.

```
Analysis Summary
----------------
File name: sf2eks2-1-note.sfx

Estimated effects for Var_1
---------------------------------
average     = 29,7556  +/- 0,824493
A:Factor_A  = 15,7838  +/- 1,64899
B:Factor_B  = 1,93875  +/- 1,64899
C:Factor_C  = -6,26875 +/- 1,64899
D:Factor_D  = 1,56875  +/- 1,64899
AB          = 2,53625  +/- 1,64899
AC          = 6,10375  +/- 1,64899
AD          = 15,4563  +/- 1,64899
BC          = -0,28125 +/- 1,64899
BD          = -0,58875 +/- 1,64899
CD          = -0,88125 +/- 1,64899
---------------------------------
Standard errors are based
on total error with 5 d.f.
```



Bemærk: Statgraphics virkninger angiver **hele forskellen mellem lav og højt niveau**, mens vi i afsnit 2.2 i overensstemmelse med SF2 kun angiver den halve forskel, dvs. fra det totale gennemsnit op til højt niveau.
Grafen til højre skal læses på den måde, at de søjler der er længere end den lodrette linie har en signifikant virkning. Af figuren ses, at det gælder A, AD C og AC. Der er derfor noget der tyder på, at vekselvirkningerne AC og AD har betydning.
Dette kan man få bekræftet ved det såkaldte normalfordelingsplot.
Vælg (blå ikon = “Graphical Options | Normal Probability Plots of effects” (slet “Pareto charts”)).
Vi får så en ny figur. Med cursoren på denne tryk på højre musetast og vælg “Label effects”.

Vi får så følgende tegning:



Ideen er her den, at hvis alle virkninger var nul er den eneste grund til afvigelser herfra den tilfældige normalfordelte støj. Virkningerne burde derfor på et normalfordelingspapir ligge på en ret linie. De der afviger stærkt herfra, (over 50% fraktilen til højre for linien, under 50% fraktilen til venstre for linien) må være signifikant forskellig fra nul.
Vi ser heraf, at AD, A og C har signifikante virkninger (muligvis også AC)

Endelig kan vi som sædvanlig ved hjælp af variansanalyse teste, hvilke faktorer der har virkning. En variansanalysetabel beregnes:

Analysis of Variance for Var_1

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
A:Factor_A	996,507	1	996,507	91,62	0,0002
B:Factor_B	15,035	1	15,035	1,38	0,2926
C:Factor_C	157,189	1	157,189	14,45	0,0126
D:Factor_D	9,84391	1	9,84391	0,91	0,3851
AB	25,7303	1	25,7303	2,37	0,1846
AC	149,023	1	149,023	13,70	0,0140
AD	955,583	1	955,583	87,86	0,0002
BC	0,316406	1	0,316406	0,03	0,8713
BD	1,38651	1	1,38651	0,13	0,7356
CD	3,10641	1	3,10641	0,29	0,6160
Total error	54,383	5	10,8766		
Total (corr.)	2368,1	15			

Det ses, at vi allerede har udelukket 3-faktorvekselvirkningen.
Af *P* - value ses, at vi kan udelukke AB, BC, BD og CD. Disse pooles ned i “Total Error” (de “excluderes”).
Variansanalysetabellen ændrer sig nu til:

Analysis of Variance for Var_1

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
A:Factor_A	996,507	1	996,507	105,61	0,0000
B:Factor_B	15,035	1	15,035	1,59	0,2386
C:Factor_C	157,189	1	157,189	16,66	0,0028
D:Factor_D	9,84391	1	9,84391	1,04	0,3337
AC	149,023	1	149,023	15,79	0,0032
AD	955,583	1	955,583	101,27	0,0000
Total error	84,9226	9	9,43585		
Total (corr.)	2368,1	15			

R-squared = 96,4139 percent   R-squared (adjusted for d.f.) = 94,0232 percent

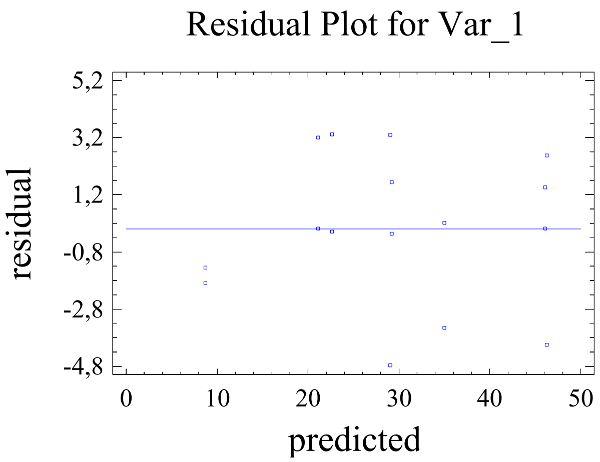
Da faktor B: konserveringsmiddel ikke indgår i de to vekselvirkninger, og har en P - value på $0.3337 > 0.05$ har konserveringsmiddel ingen virkning.
Svar på spørgsmål 2: Det ses, at faktorerne A: stabilisator, C: syre og D: temperatur har virkning i form af vekselvirkning.

Faktor B: Konserveringsmidlet pooles (excluderes) nu ned i "Error",hvorved vi får den endelige model.

Spørgsmål 3. Grafisk vurdering af model og forudsætninger

For grafisk at vurdere gyldigheden af modellen tegnes et residualplot:

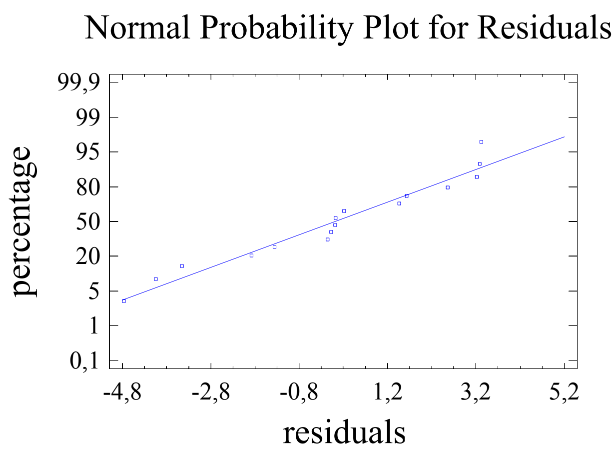
Af figuren fremgår det at residualerne ligger tilfældigt omkring 0-linien og at der ikke er nogen outliers.



Da vi har elimineret faktor B er der gentagelser.
Der synes at være ret stor forskel i varianserne, så det ville nok her være rimelig med en Bartletts test. Vi vil dog antage, at forskellene ikke er større, end at det kan forsvares at antage en rimelig varianshomogenitet.

En grafisk test for normalitet giver:

Støjen synes nogenlunde normalfordelt, da residualerne ligger "symmetrisk og tæt" omkring den rette linie.



Spørgsmål 4. Angiv de niveauer, der giver størst middeludbytte.
Umiddelbart findes niveauerne ved at vælge (gul ikon = Tabular Options | Optimization | OK).
Det giver følgende udskrift:

```
Optimize Response
-----------------
Goal: maximize Var_1    Optimum value = 46,2425
Factor                 Low               High              Optimum
-----------------------------------------------------------------------
Factor_A               -1,0              1,0               1,0
Factor_B               -1,0              1,0               -1,0
Factor_C               -1,0              1,0               -1,0
Factor_D               -1,0              1,0               1,0
```

Heraf fremgår, at A skal holdes på +1 (højt niveau) osv. Selv om vi har strøget B fra modellen angives dog, at B skal holdes på lavt niveau. Af det foregående fremgår imidlertid, at det ingen reel betydning hvilket niveau B har.
Det samme kunne gælde for nogle af de øvrige faktorer, så det er nødvendigt at foretage en nøjere undersøgelse.
For at få de relevante virkninger vælg (gul ikon = Tabular Options | Analysis Summary | OK).
Dette giver følgende tabel:

```
Analysis Summary
----------------
File name: sf2eks2-1-note.sfx
Estimated effects for Var_1
----------------------------------------------------------------------
average     = 29,7556  +/- 0,790402
A:Factor_A = 15,7838  +/- 1,5808
C:Factor_C = -6,26875 +/- 1,5808
D:Factor_D = 1,56875  +/- 1,5808
AC          = 6,10375  +/- 1,5808
AD          = 15,4563  +/- 1,5808
----------------------------------------------------------------------
Standard errors are based on total error with 10 d.f.
```

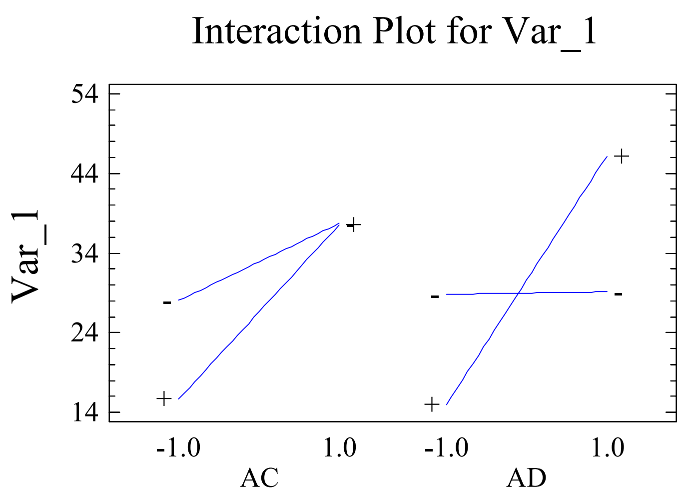
For at få det største udbytte, må derfor A holdes på +1 (højt niveau), og da A*D er positiv må også D holdes på +1 (højt niveau). Da C og A*C stort set har samme værdi, men med modsat fortegn, må det give samme resultat om man vælger syre på lavt niveau (lav koncentration) eller på højt niveau (høj koncentration).

Virkningerne er den dobbelte af de virkninger vi beskrev i afsnit 2.2. Vi har derfor, at virkningen af faktor A (stabilisator) er 7.89, af faktor C (syre) er -3.13 og af faktor D (temperatur) 0.7843. De to største tal står ud for A og A*D. For at få det største udbytte, må derfor A holdes på +1 (højt niveau), og da A*D er positiv må også D holdes på +1 (højt niveau). Da C og A*C stort set har samme værdi, men med modsat fortegn, må det give samme resultat om man vælger syre på lavt niveau (lav koncentration) eller på højt niveau (høj koncentration).

Da tolkning af vekselvirkninger kan være vanskelige kan en tegning af disse ofte være en hjælp. Vælg (blå ikon = Graphics plot |Interaction plot | OK).

<u>Tolkning af tegning</u>:
Første bogstav A i AC ses på førsteakse:til venstre -1 lavt, til højre 1 højt
Andet bogstav C i AC ses på linien: + på linie højt, - på linie lavt.

Tilsvarende tolkning af tegning for AD:
Første bogstav A niveau læses på førsteakse.
Andet bogstav D læses på linien.



Konklusion:
Af graf for AC aflæses: Stort udbytte for A= +1 og C ligegyldigt.
Af graf for AD aflæses: Stort udbytte for A= +1 og D = +

Svar på spørgsmål 4: Det er vigtigt at tilsætte stabilisator, og holde temperatur på højt niveau. Det har (hvis disse niveauer vælges) ingen betydning om der tilsættes syre med lav eller høj koncentration. Konserveringsmidlet spiller ingen rolle.

Spørgsmål 5. Beregning af estimater og konfidensintervaller.

Vælg (gul ikon = Tabular Options |Predictions | OK). Det giver følgende tabel:

```
Estimation Results for Var_1
----------------------------------------------------------------------
             Observed          Fitted  Lower 95,0% CL  Upper 95,0% CL
Row             Value           Value        for Mean        for Mean
----------------------------------------------------------------------
    1           31,54         34,9938         30,6799         39,3076
    2           32,34         29,0525         24,7386         33,3664
    3            6,85         8,73375         4,41988         13,0476
    4           21,13         21,1063         16,7924         25,4201
    5           30,85         29,2175         24,9036         33,5314
    6           24,32         21,1063         16,7924         25,4201
    7           29,06         29,2175         24,9036         33,5314
    8           46,09         46,0775         41,7636         50,3914
    9            7,38         8,73375         4,41988         13,0476
   10           35,21         34,9938         30,6799         39,3076
   11           47,54         46,0775         41,7636         50,3914
   12            42,2         46,2425         41,9286         50,5564
   13           24,29         29,0525         24,7386         33,3664
   14           22,54         22,6213         18,3074         26,9351
   15           25,94         22,6213         18,3074         26,9351
   16           48,81         46,2425         41,9286         50,5564
```

Ved at sammenligne med inddata, ses, at observationerne 8, 11, 12 og 16 alle har A og D på højt niveau, mens C der jo ingen betydning spiller, er sat på enten +1 eller -1.
Maksimalt udbytte er 46.07 (eller 46.24). Ses af "Fitted Value"
95% konfidensinterval for den optimale værdi er [41.92 ; 50.56]

Beregningerne udført ved hjælp af de i afsnit 2,2 nævnte formler.

Vor model er $\mu_{ad} = \bar{\mu} + \widetilde{A} - \widetilde{C} + \widetilde{D} - \overline{AC} + \overline{AD}$ (eller μ_{acd}) = 29.76 + 7.89 - (-3.13) + 0.78 - 3.05 + 7.72 = 46.24 dvs. et estimat for det største udbytte er 46.24% eller 46.07%

$V(\mu_{ad}) = V(\bar{\mu}..)$ + V(A) + V(C) +V(D) + V(AC) + V(AD), hvor det for hvert led gælder at dets varians

$= V(\bar{\mu}..) = \dfrac{\sigma^2}{2^4} = \dfrac{\sigma^2}{\text{antal delforsøg}}$. Vi har følgelig $V(\mu_{ad}) = 6 \cdot \dfrac{\sigma^2}{16}$.

For at få et estimat for forsøgsfejlens varians σ^2, dannes variansanalysetabellen for den endelige model (gule ikon "Tabular Options" + ANOVA)
:

```
Analysis of Variance for Var_1
-----------------------------------------------------------------------------
Source                   Sum of Squares     Df    Mean Square    F-Ratio    P-Value
-----------------------------------------------------------------------------
A:Factor_A                      996,507      1        996,507      99,69     0,0000
C:Factor_C                      157,189      1        157,189      15,73     0,0027
D:Factor_D                      9,84391      1        9,84391       0,98     0,3444
AC                              149,023      1        149,023      14,91     0,0032
AD                              955,583      1        955,583      95,60     0,0000
Total error                     99,9576     10        9,99576
-----------------------------------------------------------------------------
Total (corr.)                    2368,1     15
```

Heraf ses, at et estimat for variansen er 9.99576, med f = 10. Dermed er $V(\mu_{ad}) = 6 \cdot \dfrac{\sigma^2}{16} = 6 \cdot \dfrac{9.99576}{16} = 3.7484$.

Et 95% konfidensinterval er $\widetilde{\mu}_{ad} \pm t_{0.975}(10)\sqrt{V(\widetilde{\mu}_{ad})} = 46.24 \pm 2.23\sqrt{3.7484} = 46.24 \pm 4.32$. [41.92;50.56]

Spørgsmål 6. Estimat og konfidensinterval for differens.
Af tegningen under spørgsmål 4 ses, at hvis D holdes på lavt niveau, må A stadig holdes på højt niveau, mens C frit kan vælges på lavt eller højt niveau.
Dataene i række 2 , 5 7 og 13 opfylder dette, og det ses af udskriften i spørgsmål 4, at middeludbyttet er 29.22 for C på lavt niveau, og 29.05 for C på højt niveau.
Et konfidensinterval for differensen mellem de to udbytter kan ikke fås ud fra en Statgraphics udskrift, men ved benyttelse af de i afsnit 2.2 angivne formler:

De to modeller er: $\mu_{ad} = \bar{\mu} + \tilde{A} - \tilde{C} + \tilde{D} - \overline{AC} + \overline{AD}$

$$\mu_a = \bar{\mu} + \tilde{A} - \tilde{C} - \tilde{D} - \overline{AC} - \overline{AD}$$

Ved subtraktion fås: Diff $= \mu_{ad} - \mu_a = 2 \cdot \tilde{D} + 2 \cdot \overline{AD} = 1.569 + 15.456 = 17.025$.

$$V(Diff) = 2^2 V(D) + 2^2 V(AD) = 2^2 \cdot 2 \cdot V(D) = 8\frac{\sigma^2}{16} = \frac{9.99578}{2} = 4.9979$$

Et 95% konfidensinterval for differensen er

$$Diff \pm t_{0.975}(10)\sqrt{V(Diff)} = 17.02 \pm 2.23\sqrt{4.9979} = 17.02 \pm 4.99$$

Svar på spørgsmål 5: [12.03 ; 22.01]

Da konfidensintervallet ikke indeholder 0, er der en signifikant forskel på middeludbyttet i de to tilfælde.

◆

Øvelse 2.1 (fuldstændigt 2^k faktorforsøg) Regn opgave 12 side 240.

2.3.2. Fuldstændigt 2^3 faktorforsøg med fuldstændige blokke.

Eksempel 2.3. (fuldstændigt 2^3 faktorforsøg med blokke) (SFII eksempel 27 side 91)
Med henblik på produktionen af et bestemt stof undersøgtes om det opnåede produktionsudbytte afhang af 3 faktorer.

A:	*Ingen tilsætning af additiv (lavt niveau)*	*Tilsætning af additiv (højt niveau)*
B:	*Forbehandling (lavt niveau)*	*Efterbehandling (højt niveau)*
C:	*Fabrikationsmetode 1 (lavt niveau)*	*Fabrikationsmetode 2 (højt niveau)*

Ved produktionen benyttedes 3 forskellige fabrikationsanlæg, og det besluttedes at udføre forsøget som et randomiseret blokforsøg med fabrikationsanlæg som blokkriterium.
Forsøgsplanen var et fuldstændig randomiseret 2^3 faktorforsøg med 3 fuldstændige blokke.

Forsøgsresultaterne (% af teoretisk opnåeligt udbytte) var (i randomiseringsrækkefølge):

Blok 1	a 21	(1) 63	c 61	ac 22	ab 16	b 23	abc 37	bc 16
Blok 2	c 41	b 27	ab 2	bc 36	ac 22	(1) 37	a 35	abc 30
Blok 3	bc 20	ac 23	abc 39	(1) 43	a 33	ab 25	b 9	c 44

1) *Find hvilke faktorer der har virkning, og vurder, om det havde været nødvendigt at opdele i blokke.*

2) *Angiv de niveauer de pågældende faktorer skal indstilles på, for at give det største middeludbytte.*
3) *Forudsat blokkene ingen rolle spiller, skal man angive et estimat og et 95% konfidensinterval for dette største middeludbytte.*

LØSNING:

Spørgsmål 1. Find hvilke faktorer der har virkning

Forsøgsplan konstrueres i "Experimental Design"

Da der i eksemplet indgår en opdeling i blokke, giver det anledning til lidt andre betragtninger end i eksempel 2.1
Vi vælger " Create Design", undlader at skifte variabelnavne, vælger standardmuligheden i øverste linie, og fastholder valget ved "Display Blocked Design" .
I den næste menu med "Screening Design Options" sættes "Replicate Design" til 2 (3 blokke giver 2 ekstra udførelser), hvorved vi korrekt ser antal runs ændres til 24. Endvidere sletter vi krydset ved Randomize, da vi ønsker forsøgene skrevet i standardrækkefølge.
Der ses nu en plan med Statgraphics standardnavne såsom -1 for lav og 1 for højt niveau.

Indtastning af Data.

Vi går nu ned i bunden af skærmen og vælger et regneark "Untitled". Vi ser at regnearkets faktorer allerede er udfyldt, så vi indtaster blot udbytterne i VAR_1.
Resultatet ses nedenfor:

Block	Factor_A	Factor_B	FACTOR_C	VAR_1
1	-1,0	-1,0	-1,0	63
1	1,0	-1,0	-1,0	21
1	-1,0	1,0	-1,0	23
1	1,0	1,0	-1,0	16
1	-1,0	-1,0	1,0	61
1	1,0	-1,0	1,0	22
1	-1,0	1,0	1,0	16
1	1,0	1,0	1,0	37
2	-1,0	-1,0	-1,0	37
2	1,0	-1,0	-1,0	35
2	-1,0	1,0	-1,0	27
2	1,0	1,0	-1,0	2
2	-1,0	-1,0	1,0	41
2	1,0	-1,0	1,0	22
2	-1,0	1,0	1,0	36
2	1,0	1,0	1,0	30
3	-1,0	-1,0	-1,0	43
3	1,0	-1,0	-1,0	33
3	-1,0	1,0	-1,0	9
3	1,0	1,0	-1,0	25
3	-1,0	-1,0	1,0	44
3	1,0	-1,0	1,0	23
3	-1,0	1,0	1,0	20
3	1,0	1,0	1,0	39

Analyse af data

Vælger " Analyze Design", og VAR_1 . Dette resulterer i følgende tabel:

```
Analysis Summary
----------------
Estimated effects for Var_1
----------------------------------------------------------------------
average    = 30,2083  +/- 2,11506
A:Factor_A = -9,58333 +/- 4,23013
B:Factor_B = -13,75   +/- 4,23013
C:Factor_C = 4,75     +/- 4,23013
AB         = 12,5833  +/- 4,23013
AC         = 2,08333  +/- 4,23013
BC         = 7,91667  +/- 4,23013
block      = -2,91667 +/- 5,9823
block      = -1,41667 +/- 5,9823
----------------------------------------------------------------------
Standard errors are based on total error with 15 d.f.
```

Da vi først må teste, hvilke faktorer der har virkning, beregnes en variansanalysetabel

Analysis of Variance for Var_1

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
A:Factor_A	551,042	1	551,042	5,13	0,0387
B:Factor_B	1134,37	1	1134,37	10,57	0,0054
C:Factor_C	135,375	1	135,375	1,26	0,2791
AB	950,042	1	950,042	8,85	0,0094
AC	26,0417	1	26,0417	0,24	0,6295
BC	376,042	1	376,042	3,50	0,0809
blocks	58,5833	2	29,2917	0,27	0,7649
Total error	1610,46	15	107,364		
Total (corr.)	4841,96	23			

Af P - value ses, at vi kan udelukke AC og BC. Disse pooles (excluderes) ned i "Total Error".
Variansanalysetabellen ændrer sig nu til:

Analysis of Variance for Var_1

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
A:Factor_A	551,042	1	551,042	4,65	0,0456
B:Factor_B	1134,38	1	1134,38	9,58	0,0066
C:Factor_C	135,375	1	135,375	1,14	0,2999
AB	950,042	1	950,042	8,03	0,0115
blocks	58,5833	2	29,2917	0,25	0,7836
Total error	2012,54	17	118,385		
Total (corr.)	4841,96	23			

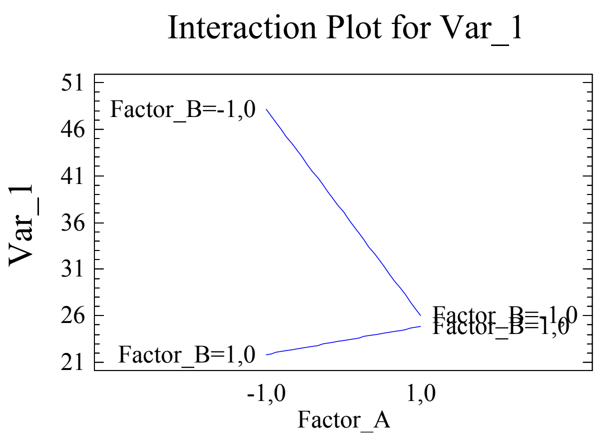
Da faktor C: Fabrikationsmetode ikke indgår i de to vekselvirkninger, og har en P-value på 0.2999 > 0.05 har fabrikationsmetoden ingen virkning.

Svar på spørgsmål 1: Det ses, at faktorerne A: additiv, B: Forbehandling har virkning i form af vekselvirkning.
Endvidere ses, at det var unødvendigt at foretage en blokopdeling da P-value er på 0.7836 > 0.05.

Spørgsmål 2. Angiv de faktorniveauer der giver størst middeludbytte .
Faktor C: Fabrikationsmetode pooles nu ned i "Error". Selvom blokke ingen virkning har er forsøget jo ikke lavet som et fuldstændigt randomiseret forsøg, og derfor må vi ikke poole blokke ned.
For at få de relevante virkninger vælges "Analysis Summary". Dette giver følgende tabel:

```
Analysis Summary
----------------
Estimated effects for Var_1
----------------------------------------------------------------------
average     = 30,2083  +/- 2,2298
A:Factor_A  = -9,58333 +/- 4,45961
B:Factor_B  = -13,75   +/- 4,45961
AB          = 12,5833  +/- 4,45961
block       = -2,91667 +/- 6,30684
block       = -1,41667 +/- 6,30684
----------------------------------------------------------------------
Standard errors are based on total error with 18 d.f.
```

Bemærk at virkningerne ikke har ændret sig fra den første model, i overensstemmelse med at disse er uafhængig af modellen.
Da virkningerne er den dobbelte af den i afsnit 2.2 angivne fås:
Virkningen af faktor A (additiv) er -4.79, af faktor B (forbehandling) er - 6.88 og af AB 6.29.



Svar på spørgsmål 2:
For at få det største udbytte, må A holdes på -1 (lavt niveau), og B på -1 (lavt niveau).
Svaret på spørgsmål 2 ses måske lettere af den figur, som fremkommer ved at vælge " Interaction plot"

Spørgsmål 3 Beregning af estimat og konfidens-interval for størst middeludbytte.
For at få et estimat for variansen vælges igen "ANOVA" og vi får følgende tabel:

Analysis of Variance for Var_1

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
A:Factor_A	551,042	1	551,042	4,62	0,0455
B:Factor_B	1134,38	1	1134,38	9,51	0,0064
AB	950,042	1	950,042	7,96	0,0113
blocks	58,5833	2	29,2917	0,25	0,7849
Total error	2147,92	18	119,329		
Total (corr.)	4841,96	23			

Det ses, at det bedste estimat for forsøgsfejlens varians er 119.3.

Et estimat for det størst mulige udbytte samt et 95% konfidensinterval fås lettest ved foretage håndregninger som beskrevet i eksempel 2.1 eller i SF II eksempel 27 side 94. Det skyldes, at Statgraphics medtager blokvirkningen i sine regninger, selvom vi klart kan se at den næppe har betydning. Elimineres blokkene, bliver forsøgsfejlens varians og dermed konfidensintervallerne forkerte.

Modellen er: $\mu_{(1)} = \bar{\mu} - \tilde{A} - \tilde{B} + \overline{AB} = 30.20 - (-4.79) - (-6.88) + 6.29 = \underline{\underline{48.16}}$

$$V(\mu_{(1)}) = V(\bar{\mu}) + V(\tilde{A}) + V(\tilde{B}) + V(\overline{AB}) = 4 \cdot \frac{\sigma^2}{2^3 \cdot 3} = \frac{\sigma^2}{6} = \frac{119.329}{6} = 19.89 .$$

Et 95% konfidensinterval for $\tilde{\mu}_{(1)}$ er

$$\tilde{\mu}_{(1)} \pm t_{0.975}(18)\sqrt{V(\tilde{\mu}_{(1)})} = 48.16 \pm 2.10\sqrt{19.89} = 48.16 \pm 9.37 \qquad \underline{\underline{[38.79 ; 57.53]}}$$

Øvelse 2.2 (fuldstændigt 2^k faktorforsøg med blokke) Regn opgave 13 side 241.

2.4 PARTIELLE 2^k FAKTORFORSØG

2.4.1 Indledning

For forsøg med en fuldstændig 2^k - faktorstruktur, stiger antallet af behandlinger i forsøget og dermed antallet af delforsøg kraftigt med antallet k af faktorer. Det betyder så også, at antallet af gentagelser af hvert faktorniveau bliver stort..
Eksempelvis vil et forsøg med 10 faktorer give 1024 delforsøg, og hver faktor, vil derfor have 512 gentagelser af hvert af de to niveauer.
Formålet med den i det følgende beskrevne partielle forsøgsstruktur er at nedsætte dette antal kraftigt. Dette opnås ved, at man udnytter, at en del af de teoretisk mulige vekselvirkninger (ikke blot alle faktorvirkninger mellem 3 eller flere faktorer, men også nogle af 2-faktorvekselvirkningerne) vides at være 0. Vi har tidligere set, at såfremt man i et forsøg med en fuldstændig faktorstruktur på forhånd kan antage, at visse vekselvirkninger er nul, så benyttes de ved variansanalysen fremkomne tilsvarende s^2 - størrelser til estimation af "støjen" (forsøgsfejlens varians). Ideen er nu, at i stedet for at benytte samtlige estimater svarende til de vekselvirkninger, der antages at være nul, til estimation af forsøgsfejlens varians, anvendes nogle af dem til estimation af virkningerne af "nye" faktorer, som indføres i en fuldstændige 2^k-te struktur.

2.4.2 Definitionsrelationer og aliasrelationer.

Lad os indledningsvist som eksempel betragte et fuldstændigt 2^k - faktorforsøg.
Fortegnsmatricen for et sådant forsøg (jævnfør side 46) er

	I	A	B	AB
(1)	+	−	−	+
a	+	+	−	−
b	+	−	+	−
ab	+	+	+	+

Ved man at vekselvirkningen AB er nul, kan man benytte fortegnssøjlen for AB til beregning af et estimat for en ny faktor C.
Faktor C kan indføres i den fuldstændige 2^2 -faktorstruktur på en måde, som vi betegner C = AB. Ved denne lader vi faktor C indgå på sit høje niveau i en behandling, når der i fortegnsmatricen for den fuldstændige 2^2 - faktorstruktur er et + i AB-søjlen, og vi lader faktor C indgå i behandlingen på sit lave niveau, når der er et minus (se følgende skema).

	I	A	B	AB	C=AB
(1)	+	−	−	+	+
a	+	+	−	−	−
b	+	−	+	−	−
ab	+	+	+	+	+

De behandlinger, der skal udføres, når vi har indført faktor C i den fuldstændige 2^2 - faktorstruktur, er en del af de behandlinger, der skal udføres i en fuldstændig 2^3 - faktorstruktur. I det foregående skema kan vi se, hvilke af den fuldstændige 2^3 - faktorstrukturs behandlinger, der skal udføres. Det fremgår, at netop halvdelen af alle behandlingerne i en fuldstændig 2^3 -faktorstruktur udføres. Den fremkomne partielle 2^3 - faktorstruktur kaldes derfor for en $\frac{1}{2}\cdot 2^3$ **eller en** 2^{3-1} **-faktorstruktur.**

Vi opskriver nu en fortegnsmatrix samtlige virkninger der indgår i den fuldstændige 2^3 - faktorstruktur.

Fortegnssøjlerne bestemmes analogt med tidligere: I A-, B- og C - søjlerne sættes + eller - eftersom det tilsvarende lille bogstav indgår i behandlingsnotationen eller ikke gør det. De øvrige søjler udfyldes efter produktregelen, hvor eksempelvis fortegnene i ABC - søjlen fås ved produkt af fortegnene i A-, B- og C-søjlerne.

Underliggende struktur	Behandlin-ger	I	A	B	AB	C=AB	AC	BC	ABC
(1)	c	+	−	−	+	+	−	−	+
a	a	+	+	−	−	−	−	+	+
b	b	+	−	+	−	−	+	−	+
ab	abc	+	+	+	+	+	+	+	+

Af fortegnsmatricen ses, at der følger en række relationer mellem fortegnssøjlerne, som symbolsk kan skrives:

A = BC
B = AC
C = AB
I = ABC

Det kan vises, at dette betyder, at vi ingen mulighed har for at estimere virkning ABC, og at de linearkombinationer, der benyttes til estimation af virkningerne A,B og C, i virkeligheden estimerer summer af virkninger: A+BC, B+AC og C+AB.

Vi benytter den terminologi, at A og BC er hinandens aliaser, B og AC er hinandens aliaser, og at C og AB er hinandens aliaser. Vi siger også, at virkningerne er sammenblandede, og at **aliasrelationerne** er A = BC, B = AC og C = AB.

Vi kalder endvidere I=ABC for den partielle faktorstrukturs **definitionsrelation**.

Vi kunne i stedet have sat C = -AB, dvs. at vi lader C-søjlen være AB - søjlen med modsat fortegn. Det giver følgende fortegnsskema

Underliggende struktur	Behandlinger	I	A	B	AB	C = - AB	AC	BC	ABC
(1)	(1)	+	−	−	+	−	+	+	−
a	ac	+	+	−	−	+	+	+	−
b	bc	+	−	+	−	+	−	+	−
ab	ab	+	+	+	+	−	−	−	−

Analogt med før ses, at der symbolsk gælder en række relationer:
A = - BC
B = - AC
C = - AB
I = -ABC
I dette tilfælde er I=-ABC den partielle faktorstrukturs definitionsrelation.

Vi kan af relationerne se, at hvis faktoren C vekselvirker med de oprindelige faktorer A og B, så kan vi ikke estimere A og B rent, da de er sammenblandede med disse vekselvirkninger. Da det naturligvis er et af formålene, så er den ovenstående forsøgsplan kun acceptabel, hvis alle 2-faktorvekselvirkninger kan antages at være 0.

Ved betragtning af fortegnssøjlerne i den tilsvarende fortegnsmatrix indses let, at der for "multiplikation" af fortegnssøjler gælder en række simple "produktregler", hvoraf vi eksempelvis med let forståelig symbolik anfører:
$A \cdot A = B \cdot B = C \cdot C = I$,
$I \cdot A = A \cdot I = A$,
$AB = A \cdot B = B \cdot A$
$ABC = A \cdot B \cdot C = B \cdot A \cdot C = C \cdot A \cdot B$ osv.

Ved benyttelse af disse produktregler fås aliasrelationerne umiddelbart udfra definitionsrelationen I=ABC:
Ved multiplikation med A: A = BC. Ved multiplikation med B: B = AC
Ved multiplikation med C: C = AB .

Omvendt kan definitionsrelationen fås ud fra en vilkårlig aliasrelation ved en multiplikation, f.eks. fås ud fra C=AB ved multiplikation med C: I=ABC.

Foranstående mønster kan generaliseres, og der kan udarbejdes en fortegnsmatrix for enhver 2^k-faktorstruktur.

Det er en erfaring, at 4-faktorvekselvirkninger i praksis er så små, at de altid kan regnes for at være 0. Det samme gælder naturligvis vekselvirkninger mellem endnu flere faktorer. Endvidere vil 3-faktorvekselvirkninger næsten altid være små sammenlignet med 2-faktorvekselvirkninger mellem de samme faktorer.
I det følgende vil vi derfor forudsætte, at kun hovedvirkninger og 2-faktorvekselvirkninger kan være forskellige fra nul.

Eksempel 2.4 (konstruktion af forsøgsplan)
Virkningerne af 7 faktorer T_1, T_2, T_3, T_4, T_5,T_6 *og* T_7 *ønskes undersøgt ved et screeningsforsøg. Man kan ikke på forhånd udelukke, at faktoren* T_6 *har 2-faktorvekselvirkninger med én eller flere af de øvrige faktorer. Alle andre vekselvirkninger kan derimod anses for udelukket.*
Opstil en forsøgsplan med færrest mulige delforsøg, og angiv herunder
a) Den underliggende fuldstændige faktorstruktur
b) Relationerne hvormed nye faktorer er indført i den underliggende struktur
c) Forsøgets behandlinger.
d) Skitser en variansanlysetabel med angivelse af frihedsgrader

LØSNING.
a) + b)
Den underliggende struktur skal altid indeholde de faktorer, som har flest mulige vekselvirkninger med andre. Dette er i dette tilfælde kun T_6 .
Da det er praktisk at benytte de sædvanlige betegnelser for faktorerne, benævnes den mest "aktive" faktor med A, og de øvrige med B, C osv.
Vi foretager følgelig følgende omdøbning: A = T_6 , B = T_1, C = T_2, D = T_3 , E = T_4, F =T_5 og G = T_7.
Der skal i alt estimeres 7 hovedvirkninger og 6 vekselvirkninger. Disse 13 virkninger kan muligvis estimeres ud fra en underliggende 2^4 faktorstruktur, da 13 < 16.
Vi prøver derfor om det er muligt at konstruere en $\frac{1}{2^3}2^7$ - faktorstruktur.
Vi opskriver samtlige hoved- og vekselvirknnger i den underliggende fuldstændige 2^4 -faktorstruktur.

	Beslaglagt virkning		Beslaglagt virkning
		D	⊗
A	⊗	AD	⊗
B	⊗	BD	
AB	⊗	ABD	
C	⊗	CD	
AC	⊗	ACD	
BC		BCD	
ABC		ABCD	

I skemaet er med ⊗ markeret de "beslaglagte" virkninger, dvs. virkninger som kan være forskellig fra 0.
Vi indfører nu eksempelvis den "nye" faktor E ved aliasrelationen E = ABCD (som jo ikke er beslaglagt)
Vi får så (ved multipication med A) at AE = BCD (ikke beslaglagt)
Vi kan så indføre faktorerne F og G ved relationerne
F = ACD (ikke beslaglagt) og G = ABD (ikke beslaglagt)
Ved multiplikation fås AF = CD (ikke beslaglagt) og AG = BD (ikke beslaglagt)

Resultatet kan samles i tabellen

	Beslaglagt virkning		Beslaglagt virkning
		D	⊗
A	⊗	AD	⊗
B	⊗	BD = AG	
AB	⊗	ABD = G	
C	⊗	CD = AF	
AC	⊗	ACD = F	
BC		BCD = AE	
ABC		ABCD = E	

Vi har dermed vist, at en $\frac{1}{2^3}2^7$ - faktorstruktur med ovenstående struktur kan benyttes til testning af de mulige hoved- og 2-faktorvekselvirkninger.

c) Strukturens behandlinger bliver

Underlig-gende struktur	A	B	C	D	E=ABC	F=ACD	G=ABD	Behandlinger
(1)	−	−	−	−	−	−	−	(1)
a	+	−	−	−	+	+	+	aefg
b	−	+	−	−	+	−	+	beg
ab	+	+	−	−	−	+	−	abf
c	−	−	+	−	+	+	−	cefg
ac	+	−	+	−	−	−	+	acg
bc	−	+	+	−	−	+	+	bcfg
abc	+	+	+	−	+	−	−	abce
d	−	−	−	+	−	+	−	df
ad	+	−	−	+	+	−	−	ade
bd	−	+	−	+	+	+	+	bdefg
abd	+	+	−	+	−	−	+	abdg
cd	−	−	+	+	+	−	+	cdeg
acd	+	−	+	+	−	+	−	acdf
bcd	−	+	+	+	−	−	−	bcd
abcd	+	+	+	+	+	+	+	abcdefg

d) En skitse til variansanlysetabel er

Variation	SAK	f	s^2	F
A		1		
B		1		
C		1		
D		1		
E		1		
F		1		
G		1		
AB		1		
AC		1		
AD		1		
AE		1		
AF		1		
AG		1		
Gentagelser		2		
Total		15		

◆

Øvelse 2.3 (partielt 2^k faktorforsøg). Regn opgave 14 side 241

2.4.3 Konstruktion af partiel forsøgsplan ved Statgraphics.

Følgende eksempel illustrerer, hvorledes man ved hjælp af Statgraphics kan konstruere en partiel forsøgsplan, der svarer til de opgivne forudsætninger.

Eksempel 2.5.(konstruktion af partiel forsøgsplan)
Virkningen af 5 faktorer A, B, C, D og E, som ikke vekselvirker, skal undersøges ved et fuldstændigt randomiseret forsøg.
Konstruer en forsøgsplan med det færrest mulige antal delforsøg.

LØSNING:
Da der kun skal undersøges 5 virkninger, vil det være naturligt at antage man kan nøjes med 8 delforsøg dvs. et 2^{5-2} faktorstruktur.
Vælg (Speciel | Experimental Design | Create Design | Screening, sæt “No of response” = 1 og “No of experimental Factor”=5 | OK | OK |OK). Vi lader Statgraphics selv vælge navne.
I den næste menu slettes krydset ved “Display Blocked Designs”, da der ikke indgår blokke.
Ved at trykke på pilen i højre side af den fremhævede linie kan man se alle muligheder. Vi ser, at det er muligt at vælge en 2^{5-2} plan med kun 8 delforsøg (kaldet “Quarter fraction”).
I den næste menu med “Screening Design Options” ses antal runs korrekt at være 8. Tryk på OK.
Der ses nu en plan med Statgraphics standardnavne såsom -1 for lav og 1 for højt niveau.
Vælg (gul ikon = Tabular Options | Alias Structure | OK).

Derved fremkommer følgende udskrift:

```
Alias Structure
Contrast  Estimates
--------  ---------
1         A+BD+CE
2         B+AD
3         C+AE
4         D+AB
5         E+AC
6         BC+DE
7         BE+CD
```

Det ses, at alle hovedvirkninger kan estimeres rent, da alle 2-faktorvekselvirkninger antages 0.
For at få en randomiseret arbejdsplan, hvor behandlingerne er skrevet i tilfældig (randomiseret) rækkefølge så vælg (gul ikon = Tabular Options | Worksheet | OK).
Derved fremkommer følgende udskrift.

run	Factor_A	Factor_B	Factor_C	Factor_D	Factor_E	Var_1
1	-1,0	1,0	1,0	-1,0	-1,0	____________
2	1,0	-1,0	1,0	-1,0	1,0	____________
3	-1,0	-1,0	-1,0	1,0	1,0	____________
4	1,0	1,0	-1,0	1,0	-1,0	____________
5	1,0	-1,0	-1,0	-1,0	-1,0	____________
6	-1,0	-1,0	1,0	1,0	-1,0	____________
7	-1,0	1,0	-1,0	-1,0	1,0	____________
8	1,0	1,0	1,0	1,0	1,0	____________

Ønsker man at se hvordan forsøgsplanen er "genereret", så
Vælg (Cursor på Worksheet | højre musetast | Analysis Options | Generators).
Vi ser så, at D = AB og E = AC
Forsøgsplanen i "worksheetet" ovenfor bliver med de sædvanlige bogstavbetegnelser:
bc, ace , de , abd , a , cd , be , abcde
Forsøgene skulle nu udføres, og analysen foretages ligesom i eksempel 2.2.

◆

Øvelse 2.4 (konstruktion af forsøgsplan ved Statgraphics) Regn opgave 15 side 241.

2.4.4 Analyse af et partielt 2^k faktorforsøg

Eksempel 2.6.(Analyse af et 2^{5-2} faktorforsøg) (SFII eksempel 32 side 115)
Virkningen af 5 faktorer A, B, C, D og E, som ikke vekselvirker, skal undersøges ved et fuldstændigt randomiseret forsøg.
Faktorerne D og E indføres ved aliasrelationerne D = -ABC og E = BC.
Forsøgsplanen (opskrevet på sædvanlig måde i standardorden) og forsøgsresultaterne er:

	I	A	B	C	D =-ABC	E= BC	behandlinger	resultat
(1)	+	−	−	−	+	+	de	7
a	+	+	−	−	−	+	ae	19
b	+	−	+	−	−	−	b	25
ab	+	+	+	−	+	−	abd	15
c	+	−	−	+	−	−	c	54
ac	+	+	−	+	+	−	acd	46
bc	+	−	+	+	+	+	bcde	12
abc	+	+	+	+	−	+	abce	31

1) *Find hvilke faktorer der har virkning*
2) *Angiv de niveauer de pågældende faktorer skal indstilles på, for at give det mindste middeludbytte, og angiv et 95% konfidensinterval for dette middeludbytte.*

LØSNING:
Spørgsmål 1. Find de faktorer der har virkning.

Forsøgsplan indføres i "Experimental Design"
Vi vælger ikke at skifte variabelnavne.
Da vi ikke selv har valgt planen, kan vi ikke forvente at finde denne som en standardmodel (den af Statgraphics angivne 2^{5-2}-plan ved vi fra eksempel 2,4 har andre aliasrelationer).
Vi må derfor vælge "User- specified design"
I den næste menu med "Screening Design Options" slettes krydset ved Randomize, da vi ønsker forsøgene skrevet i standardrækkefølge.
Der ses nu en plan med Statgraphics standardnavne såsom -1 for lav og 1 for højt niveau.

Indtastning af Data
Vi går nu ned i bunden af skærmen og vælger et regneark "Untitled". Vi bliver her nødt til at udfylde alle faktorers niveauer.

Resultatet ses her:

Block	Factor_A	Factor_B	FACTOR_C	Factor_D	FACTOR_E	VAR_1
1	-1	-1	-1	1	1	7
1	1	-1	-1	-1	1	19
1	-1	1	-1	-1	-1	25
1	1	1	-1	1	-1	15
1	-1	-1	1	-1	-1	54
1	1	-1	1	1	-1	46
1	-1	1	1	1	1	12
1	1	1	1	-1	1	31

Analyse af data (variansanalyse)
Vi vælger " Analyze Design". Indsættes response variablen fås nu følgende tabel over virkningerne

```
Analysis Summary
----------------
Estimated effects for Var_1
----------------------------------------------------------------------
average           = 26,125
A:Factor_A-DE     = 3,25
B:Factor_B+CE     = -10,75
C:Factor_C+BE     = 19,25
D:Factor_D-AE     = -12,25
E:Factor_E-AD+BC  = -17,75
AB-CD             = 1,25
AC-BD             = 2,25
----------------------------------------------------------------------
```

Da vi fik oplyst at faktorerne ikke vekselvirker, eliminerer vi alle vekselvirkningerne .
Vælg (Cursor i udskrift | højre musetast | Analysis Options | Maximum Order Effect" til 1 | OK)
Dette resulterer i følgende tabel:

```
Analysis Summary
----------------
Estimated effects for Var_1
----------------------------------------------------------------
average    = 26,125 +/- 0,910014
A:Factor_A = 3,25   +/- 1,82003
B:Factor_B = -10,75 +/- 1,82003
C:Factor_C = 19,25  +/- 1,82003
D:Factor_D = -12,25 +/- 1,82003
E:Factor_E = -17,75 +/- 1,82003
----------------------------------------------------------------
Standard errors are based on total error with 2 d.f.
```

Da vi først må teste, hvilke faktorer der har virkning, beregnes en variansanalysetabel.

Analysis of Variance for Var_1

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
A:Factor_A	21,125	1	21,125	3,19	0,2161
B:Factor_B	231,125	1	231,125	34,89	0,0275
C:Factor_C	741,125	1	741,125	111,87	0,0088
D:Factor_D	300,125	1	300,125	45,30	0,0214
E:Factor_E	630,125	1	630,125	95,11	0,0104
Total error	13,25	2	6,625		
Total (corr.)	1936,88	7			

Svar på spørgsmål 1:
B og D har en virkning, og det må antages at C og E har en stærk virkning.
Faktoren A har næppe nogen virkning, da P-value = 0.2161 > 0.05.

Spørgsmål 2. Beregning af estimat for optimal værdi samt konfidensintervaller
For at få det bedst mulige estimat for forsøgsfejlens varians pooles A ned i "Error".
Vi får følgende tabel:

Analysis of Variance for Var_1

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
B:Factor_B	231,125	1	231,125	20,17	0,0206
C:Factor_C	741,125	1	741,125	64,68	0,0040
D:Factor_D	300,125	1	300,125	26,19	0,0144
E:Factor_E	630,125	1	630,125	54,99	0,0051
Total error	34,375	3	11,4583		
Total (corr.)	1936,88	7			

Det ses at det bedste estimat for forsøgsfejlens varians er 11.46.
Svar på spørgsmål 2: Af fortegnene for virkningerne ses, at den mindste middelværdi fås for B, D og E på højt niveau og for C på lavt niveau.

$$\tilde{\mu}_{bde} = \bar{\mu} + \tilde{B} - \tilde{C} + \tilde{D} + \tilde{E} = 26.13 + \left(\frac{-10.75}{2}\right) - \left(\frac{19.25}{2}\right) + \left(\frac{-12.25}{2}\right) + \left(\frac{-17.75}{2}\right)$$

$$= \underline{\underline{-3.87}}$$

Variansen på $\tilde{\mu}_{bde}$ er $V(\tilde{\mu}_{bde}) = 5 \cdot \frac{s_0^2}{2^3} = 5 \cdot \frac{11.46}{8} = 7.16$

Et 95% konfidensinterval for $\tilde{\mu}_{bde}$ er

$$\tilde{\mu}_{bde} \pm t_{0.975}(3) \cdot \sqrt{V(\tilde{\mu}_{bde})} = -3.87 \pm 3.18 \cdot \sqrt{7.16} = -3.87 \pm 8.51 \qquad \underline{\underline{[-12.38\,;\,4.64\,]}}$$

Disse resultater kan dog også fås ved hjælp af Statgraphics.
Det ses, at behandlingen bde ikke findes blandt de oprindelige data. For at få den med tilføjes i regnearket en observation 9

Block	Factor_A	Factor_B	FACTOR_C	Factor_D	FACTOR_E	VAR_1
1	-1	-1	-1	1	1	7
1	1	-1	-1	-1	1	19
1	-1	1	-1	-1	-1	25
1	1	1	-1	1	-1	15
1	-1	-1	1	-1	-1	54
1	1	-1	1	1	-1	46
1	-1	1	1	1	1	12
1	1	1	1	-1	1	31
		1	-1	1	1	

Faktor *A* der ingen rolle spiller giver man et vilkårligt niveau.
Vælg (gul ikon = Tabular Options | Predictions | OK). Vi får følgende tabel:

Estimation Results for Var_1

Row	Observed Value	Fitted Value	Lower 95,0% CL for Mean	Upper 95,0% CL for Mean
1	7,0	6,875	-1,64152	15,3915
2	19,0	19,125	10,6085	27,6415
3	25,0	26,125	17,6085	34,6415
4	15,0	13,875	5,35848	22,3915
5	54,0	56,125	47,6085	64,6415
6	46,0	43,875	35,3585	52,3915
7	12,0	15,375	6,85848	23,8915
8	31,0	27,625	19,1085	36,1415
9		-3,875	-12,3915	4,64152

Det ses at give samme resultat.

Øvelse 2.5 (analyse ved Statgraphics af et partielt forsøg) Regn opgave 16 side 241.

2.5 SEKVENTIEL FORSØGSSTRATEGI

I en række situationer vil man selv med partielle 2^k - faktorforsøg og med rimelige forhånds-forudsætninger om hvilke vekselvirkninger der er nul få mange behandlinger og dermed mange delforsøg. Er delforsøgene dyre og budgettet begrænset, kan man prøve at formindske antallet ved at udføre færre forsøg end nødvendigt til at estimere alle virkninger, og dog håbe at nå et fornuftigt resultat. I SFII side 135-143 er metoden udførligt beskrevet. Vi vil her illustrere det ved et eksempel.

Eksempel 2.7. (sekventiel forsøgsstrategi)

Virkningen af 4 faktorer A, B, C og D ønskes undersøgt. Da vi ikke ved andet end at kun hovedvirkninger og 2-faktorvekselvirkninger kan være forskellig fra 0, er det nødvendigt at udføre et fuldstændigt 2^4 - faktorforsøg (der er i alt 10 virkninger der skal estimeres). Af økonomiske og andre grunde ønsker man at reducere dette antal.

Indføres faktoren D ved D = ABC skal man kun udføre 8 delforsøg, men til gengæld vil 2-faktorvekselvirkningerne være sammenblandede.

Idet I = ABCD ses aliasrelationerne at være

A = BCD B = ACD C = ABD D = ABC AB = CD AC = BD BC = AD

Der er derfor en risiko for, at man ikke kan konkludere noget. I så tilfælde må man udføre de resterende 8 forsøg svarende til en fuldstændig 2^4 faktorforsøg. Dette svarer til at vi havde indført D = - ABC (se nedenstående tabel).

	A	B	C	D = ABC	Behandlinger	D = - ABC	Behandlinger
(1)	−	−	−	−	(1)	+	d
a	+	−	−	+	ad	−	a
b	−	+	−	+	bd	−	b
ab	+	+	−	−	ab	+	abd
c	−	−	+	+	cd	−	c
ac	+	−	+	−	ac	+	acd
bc	−	+	+	−	bc	+	bcd
abc	+	+	+	+	abcd	−	abc

Det er en erfaring at man ofte kan få en korrekt fortolkning af forsøgsresultaterne ved at vælge den, som giver den enkleste model. (den simpleste models princip).

Man fortolker således manglende signifikanser som virkninger, der er nul, og søger den model, som giver det færreste antal hovedvirkninger og vekselvirkninger forskellig fra nul.

Dette kan bedst illustreres ved at gennemgå 7 forskellige typer testresultater, som kunne tænkes at forekomme i eksempel 2.6.

Resultaterne fremgår af den følgende tabel, hvor * betyder at der er fundet signifikans.

Virkninger	Signifikanser						
	1	2	3	4	5	6	7
A		*	*	*	*	*	
B			*	*		*	
C							
D							
AB + CD				*	*		*
AC + BD							
BC + AD						*	

Fortolkningen er i de 7 tilfælde:

Tilfælde 1: Alle faktorvirkninger er nul

Tilfælde 2: Kun A har en virkning forskellig fra 0.

Tilfælde 3: Kun A og B har virkninger forskellig fra 0.

Tilfælde 4: A og B har virkninger forskellig fra 0. A og B vekselvirker.

Tilfælde 5: A og B har virkninger forskellig fra 0. A og B vekselvirker.

Tilfælde 6: A og B har virkninger forskellig fra 0. Desuden er mindst en af 2-faktorvekselvirkningerne BC og AD forskellig fra nul. Et supplerende forsøg må udføres.

Tilfælde 7: Mindst en af 2-faktorvekselvirkningerne BC og AD er forskellig fra nul. Et supplerende forsøg må udføres.

Man ser, at kun i de 2 sidste tilfælde er det nødvendigt at udføre et supplerende forsøg. Bemærk, at man i sådanne tilfælde let får indført en blokvirkning mellem forsøgsrunde 1 og det supplerende forsøg.

Øvelse 2.6 (simpleste models princip) Regn opgave 18 side 242.

2.6 KONFUNDERET PARTIELT 2^k FAKTORFORSØG

2.6.1 Indledning

Blokke dannes som bekendt for at forminske "støjen". Eksempelvis kan det være nødvendigt f tidsmæssige grunde at benytte to ovne ved forsøget, og den ene ovn er af ældre dato, tager længere tid at varme op, og man kan derfor risikere, at forsøgsresultaterne generelt er "lavere" end et tilsvarende forsøg i den anden ovn. Hidtil har vi omhyggeligt lavet fuldstændige blokke, dvs. alle behandlinger har været repræsenteret i begge blokke. Uden en sådan balance er det klart ikke muligt at give en "fair" vurdering af behandlingerne. Imidlertid koster det jo mange forsøg, idet eksempelvis 2 blokke jo så kræver, at alle behandlinger skal udføres 2 gange. Skal antal delforsøg reduceres, er det derfor væsentligt, at have en metode til at lave lave balancerede ufuldstændige blokke. Efter nogenlunde samme teknik som ved indførelsen af faktorer i partielle forsøg, kan dette lade sig gøre. Man siger så, at man har konfunderet blokkene i strukturen.

2.6.2 Konfundering af ufuldstændige blokke

Vi vil forklare hvorledes man indfører ufuldstændige blokke i strukturen ved følgende eksempler.

Eksempel 2.8 (dannelse af ufuldstændige blokke).
Lad os betragt et forsøg med en fuldstændig 2^3 - faktorstruktur med faktorerne A,B og C. Antag, at man er nødt til at danne 2 blokke. Da man i alt har 8 delforsøg, så må hver blok kun indeholde 4 delforsøg. De 8 behandlinger kan derfor kun for halvdelens vedkommende forekomme i en blok. Man siger, at vi danner 2 ufuldstændige blokke med fire forsøgsenheder indenfor hver blok.
Dette sker på følgende måde: Betragt fortegnsmatricen for den fuldstændige 2^3 - faktorstruktur.

	Virkninger							
Behandlinger	I	A	B	AB	C	AC	BC	ABC
(1)	+	−	−	+	−	+	+	−
a	+	+	−	−	−	−	+	+
b	+	−	+	−	−	+	−	+
ab	+	+	+	+	−	−	−	−
c	+	−	−	+	+	−	−	+
ac	+	+	−	−	+	+	−	−
bc	+	−	+	−	+	−	+	−
abc	+	+	+	+	+	+	+	+

Antag, at man har én gentagelse af hver behandling. Denne antagelse vil vi gøre overalt i det følgende, medmindre andet udtrykkeligt bemærkes.
En måde at fordele de 8 behandlinger på blokkene er med tilfældig nummerering af disse, at man i blok 1 har de behandlinger, som indgår i fortegnssøjlen med $-$, mens man i blok2 sætter dem der indgår med +.

Definitionsrelation I = - ABC	Blok 1	(1)	ab	ac	bc
Definitionsrelation I = ABC	Blok 2	a	b	c	abc

Behandlingerne skal naturligvis randomiseres på forsøgsenhederne inden for hver blok.
Et estimat for vekselvirkning ABC estimerer derfor i virkeligheden summen af ABC og blokvirkningen. Med den pågældende forsøgsplan kan virkning ABC ikke estimeres særskilt, og forsøgsplanen har kun mening, hvis man på forhånd kan antage, at ABC = 0.
For de øvrige virkninger gælder, at de indenfor hver blok optræder 2 gange på højt niveau, og 2 gange på lavt niveau. Estimatet for hver af de øvrige virkninger beregnes derfor ved, at to forsøgsresultater indenfor hver blok sammenlægges og to fratrækkes.

Eksempelvis er $\tilde{A} = \dfrac{-(1) + a - b + ab - c + ac - bc + abc}{8}$

Selv om blok 1 giver 100 gange så høje resultater som blok 2, så bliver estimaterne for disse virkninger derfor ikke påvirket af blokvirkningen. Vi siger i det betragtede tilfælde, at **vekselvirkning ABC er konfunderet.**
Bemærk: 2 blokke beslaglægger én frihedsgrad, og en virkning konfunderes.

En anden måde at fordele de otte behandlinger på blokkene er:

Definitionsrelation I = AB	Blok 1	(1)	ab	c	abc
Definitionsrelation I = -AB	Blok 2	a	b	ac	bc

Man har i dette tilfælde konfunderet vekselvirkning AB.

Da man sædvanligvis med større rimelighed kan forudsætte, at ABC = 0, end at AB = 0, er det i almindelighed bedst at konfundere vekselvirkning ABC.

◆

I almindelighed gælder, at man kan konfundere en vekselvirkning V i en fuldstændig 2^k - faktorstruktur mellem 2 blokke ved at fordele behandlingerne således, at forsøgsenhederne i den ene blok får de behandlinger, der indgår med positivt fortegn i fortegnsmatricens T - søjle, og at forsøgsenhederne i den anden blok får de behandlinger, der indgår med negativt fortegn.

Det foranstående kan generaliseres. I det følgende eksempel betragtes således tilfældet med 4 blokke.

Eksempel 2.9. (dannelse af 4 ufuldstændige blokke).
Antag, at de 8 behandlinger i en fuldstændig 2^3 - faktorstruktur med faktorerne A,B og C skal fordeles på 4 blokke, hver med 2 forsøgsenheder. En måde at fordele de 8 behandlinger på blokkene er med tilfældig nummerering af disse:

Definitionsrelation I = AB = BC(=AC)	Blok 1	(1)	abc
Definitionsrelation I = AB = - BC	Blok 2	ab	c
Definitionsrelation I = -AB = BC	Blok 3	a	bc
Definitionsrelation I = - AB = - BC	Blok 4	b	ac

Forsøgsenhederne i blok 1 har fået de behandlinger, der indgår med positivt fortegn både i fortegnsmatricens AB - søjle og i dens BC - søjle.
Forsøgsenhederne i blok 2 har fået de behandlinger, der indgår med fortegnene + i AB søjlen og - BC - søjlen.
Tilsvarende er behandlingerne i blokkene 3 og 4 bestemt ved fortegnskombinationerne - +, henholdsvis - - i AB - og BC - søjlerne. Behandlingerne skal naturligvis randomiseres på forsøgsenhederne inden for hver blok.

Med den anførte fordeling af behandlingerne på blokke er det klart, at estimater for vekselvirkningerne AB og BC er sammenblandet med blokvirkningerne. Man udtryker det ved at sige, at AB og BC er blevet konfunderet. De konfunderede virkninger kan altså ikke estimeres, og forsøgsplanen har kun mening, hvis man på forhånd kan antage, at AB=BC=O.
Det ses ved hjælp af fortegnsmatricen {eller definitionsrelationerne) , at med konfunderingen af de to vekselvirkninger AB og BC er også "produktet" AC blevet konfunderet. Forsøgsplanen forudsætter altså, at også BC = 0. Estimaterne for alle andre virkninger påvirkes ikke af blokinddelingen.
Bemærk: 4 blokke beslaglægger 3 frihedsgrader, og 3 virkninger konfunderes, hvoraf de 2 kan vælges frit.

◆

Eksempel 2.10. (Konstruktion af et partielt konfunderet forsøg).
Virkningerne af 5 faktorer T_1, T_2, T_3, T_4 og T_5 ønskes undersøgt ved et screeningsforsøg. Man ved at ingen af faktorerne T_2 og T_4 vekselvirker med andre faktorer. Man er nødt til at lave et blokforsøg, med 4 forsøgsenheder i hver blok.
Opstil en forsøgsplan med færrest mulige delforsøg, og angiv herunder
a) Forsøgets behandlinger, og behandlingernes fordeling på blokke.
b) En variansanlysetabel med angivelse af frihedsgrader
LØSNING:
Den underliggende struktur skal altid indeholde de faktorer, som har flest mulige vekselvirkninger med andre. Dette er i dette tilfælde T_1, T_3 og T_5 .
Da det er praktisk at benytte de sædvanlige betegnelser for faktorerne, benævnes den mest "aktive" faktor med A, og de øvrige med B, C osv.
Vi foretager følgelig følgende omdøbning: A = T_1 , B = T_3, C = T_5, D = T_2 og E = T_4.
Der skal i alt estimeres 5 hovedvirkninger og 3 vekselvirkninger. Disse 8 virkninger kan forhåbentlig estimeres ud fra en underliggende 2^4 faktorstruktur, da 8 < 16.
Da blokkene er på 4 delforsøg, må vi så også indlægge 4 blokke. Dette koster 3 frihedsgrader, dvs i alt 8 + 4 = 12 frihedsgrader.

Vi prøver derfor om det er muligt at konstruere en konfunderet $\frac{1}{2}2^5$ - faktorstruktur.

Vi indfører nu eksempelvis den "nye" faktor D ved aliasrelationen E = - ABCD (som jo ikke er beslaglagt).Endvidere konfunderer vi ABD og CD (og dermed også ABC).
Vi opskriver den beskrevne struktur i følgende skema:

	⊗ = Beslaglagt virkning k = konfunderet virkning		⊗ = Beslaglagt virkning k = konfunderet virkning
		D	⊗
A	⊗	AD	
B	⊗	BD	
AB	⊗	ABD	k
C	⊗	CD	
AC	⊗	ACD	
BC	⊗	BCD	
ABC	k	ABCD=-E	⊗

Vi har dermed vist, at ovenstående model kan anvendes. ◆

Svarene på spørgsmålene kan ses i det følgende eksempel 2.11.

Øvelse 2.7 (konfunderet partielt forsøg) Regn opgave 19 side 243.

2.6.3 Analyse af et konfunderet partielt 2^k faktorsorsøg

Eksempel 2.11. (Analyse af et konfunderet partielt 2^{5-1} faktorsorsøg) (SFII eks. 50 side 166)
Virkningen af 5 faktorer A, B, C, D og E ønskes undersøgt. Man ved, at ingen af faktorerne D og E vekselvirker med andre faktorer. Da man kun kan udføre 4 forsøg pr apparat, indføres blokke på 4 forsøgsenheder.
I en fuldstændig 2^4 struktur med faktorerne A, B, C og D indføres E = - ABCD. Endvidere indføres blokkene ved at konfundere ABD og CD.
Forsøgsplanen (opskrevet på sædvanlig måde i standardorden) og forsøgsresultaterne er:

	A	B	C	D	E=-ABCD	behandlinger	ABD	CD	Blokke	Resultat
(1)	−	−	−	−	−	(1)	−	+	3	9
a	+	−	−	−	+	ae	+	+	1	16
b	−	+	−	−	+	be	+	+	1	11
ab	+	+	−	−	−	ab	−	+	3	13
c	−	−	+	−	+	ce	−	−	4	10
ac	+	−	+	−	−	ac	+	−	2	14
bc	−	+	+	−	−	bc	+	−	2	6
abc	+	+	+	−	+	abce	−	−	4	17
d	−	−	−	+	+	de	+	−	2	11
ad	+	−	−	+	−	ad	−	−	4	14
bd	−	+	−	+	−	bd	−	−	4	7
abd	+	+	−	+	+	abde	+	−	2	14
cd	−	−	+	+	−	cd	+	+	1	9
acd	+	−	+	+	+	acde	−	+	3	16
bcd	−	+	+	+	+	bcde	−	+	3	8
abcd	+	+	+	+	−	abcd	+	+	1	5

1) *Find hvilke faktorer der har virkning*
2) *Angiv de niveauer de pågældende faktorer skal indstilles på, for at give det største middeludbytte.*

LØSNING:

1) Forsøgsplan indføres i "Experimental Design"
   Denne gang skiftes variabelnavne til A, B osv.. Derimod fastholdes navnet VAR_1, og man fastholdes krydset ved "Display Blocked Designs", da der indgår blokke.
   Statgraphics i den nuværende version har ikke mange konfunderede planer indbygget, så man må sædvanligvis selv konstruere planen og derfor vælge "User specified design"
   Indtastning af Data: Vi går nu ned i bunden af skærmen og vælger et regneark "Untitled". Vi bliver her nødt til at udfylde alle faktorers niveauer.
   Resultatet er:

Block	A	B	C	D	E	Var_1
3	-1	-1	-1	-1	-1	9
1	1	-1	-1	-1	1	16
1	-1	1	-1	-1	1	11
3	1	1	-1	-1	-1	13
4	-1	-1	1	-1	1	10
2	1	-1	1	-1	-1	14
2	-1	1	1	-1	-1	6
4	1	1	1	-1	1	17
2	-1	-1	-1	1	1	11
4	1	-1	-1	1	-1	14
4	-1	1	-1	1	-1	7
2	1	1	-1	1	1	14
1	-1	-1	1	1	-1	9
3	1	-1	1	1	1	16
3	-1	1	1	1	1	8
1	1	1	1	1	-1	5

Analyse af data

Der vælges "Analyze Design", og efter at have valgt variabel fås følgende udskrift:

```
Analysis Summary
----------------
Not enough data to fit specified model.
```

Dette er jo sandt nok.

Vælg (Cursor i udskrift | højre musetast | Analysis Options | Exclude | dobbeltklik på AD, AE, BD, BE, CD, CE, DE | OK | OK) .

Vi får så først en tabel over virkningerne.

Vi vælger "ANOVA Tables", og får følgende variansanalysetabel:

Analysis of Variance for Var_1

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
A:A	90,25	1	90,25	11,46	0,0276
B:B	20,25	1	20,25	2,57	0,1841
C:C	6,25	1	6,25	0,79	0,4233
D:D	9,0	1	9,0	1,14	0,3453
E:E	42,25	1	42,25	5,37	0,0815
AB	1,0	1	1,0	0,13	0,7396
AC	0,0	1	0,0	0,00	1,0000
BC	4,0	1	4,0	0,51	0,5154
blocks	6,5	3	2,16667	0,28	0,8413
Total error	31,5	4	7,875		
Total (corr.)	211,0	15			

Det ses, at alle vekselvirkninger er 0.

Vi foretager nu en pooling , og får

Analysis of Variance for Var_1

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
A:A	90,25	1	90,25	17,31	0,0042
B:B	20,25	1	20,25	3,88	0,0894
C:C	6,25	1	6,25	1,20	0,3098
D:D	9,0	1	9,0	1,73	0,2303
E:E	42,25	1	42,25	8,10	0,0248
blocks	6,5	3	2,16667	0,42	0,7474
Total error	36,5	7	5,21429		
Total (corr.)	211,0	15			

Spørgsmål 1: <u>Det ses, at kun faktorerne A og E har en hovedvirkning, mens B, C og D ingen rolle spiller.</u>

2) Angiv de faktorniveauer, der giver størst middeludbytte.
B, C og D pooles ned i "Error".
Vælges "Analysis Summary" fås følgende tabel:

```
Analysis Summary
----------------
File name: C:\SGWIN\SF2\SF2EKS2_11.SFX
Estimated effects for Var_1
----------------------------------------------------------------------
average = 11,25 +/- 0,67082
A:A     = 4,75  +/- 1,34164
E:E     = 3,25  +/- 1,34164
block   = -2,0  +/- 2,32379
block   = 1,5   +/- 2,32379
block   = 0,0   +/- 2,32379
----------------------------------------------------------------------
Standard errors are based on total error with 10 d.f.
```

Heraf ses
Svar på spørgsmål 2: <u>Ønskes man den højest mulige middelværdi, må A og E være på højt niveau.</u>

Estimation af den højeste middelværdi og konfidensintervaller kan foretages (med brug af lommeregner) som i eksempel 2.3.

◆

Øvelse 2.8 (analyse af et konfunderet partielt forsøg) Regn opgave 20 side 243.

3 FIXED, RANDOM, SIDEDELT, TRIN-VIST OG MIXED FACTOR ANALYSER

3.1 INDLEDNING.

Formålet med dette notat er at lære ud fra praktiske eksempler at identificere de forskellige modeller, og ved hjælp af varianskomponentmetoden og et variansanalyseprogram at foretage de korrekte variansanalyser. De bagved liggende matematisk og statistiske beregninger og formler indgår derimod ikke.

3.2 DEFINITIONER

Fixed factor foreligger, hvis faktorens niveauer er bevidst udvalgte (kontrollerede).
Eksempel: Vi har 3 katalysatorer, hvis virkning på udbyttet af en kemisk produktion man ønsker at finde.
Faktoren "katalysator" er en "fixed factor" med 3 niveauer.

Random factor foreligger, hvis niveauerne er fremkommet som tilfældigt valgte blandt de mulige niveauer.
Eksempel: En virksomhed har et stort antal katalysatorer på lager til udskiftning i produktions-processen. Man ønsker at undersøge den variation som katalysatorerne tilfører produktionsprocessen. Forsøget udføres ved at 4 katalysatorer udtages tilfældigt fra lageret.
Faktoren "katalysator" er en "random factor" med 4 niveauer.

Sidedelt faktor (crossed factor). At faktor B er sidedelt med faktor A's niveauer, betyder, at ethvert niveau af B kombineres med ethvert niveau af A (niveauerne krydser fuldstændigt)
Eksempel: Ved en sædvanlig "sidedelt" analyse undersøges benzinforbruget ved benyttelse af 3 forskellige olieblandinger og 2 karburatorer
Ethvert niveau af faktoren karburator kombineres med ethvert niveau af faktoren olieblanding (se figur)

		Karburator	
		K_1	K_2
Olieblanding	O_1	× ×	× ×
	O_2	× ×	× ×
	O_3	× ×	× ×

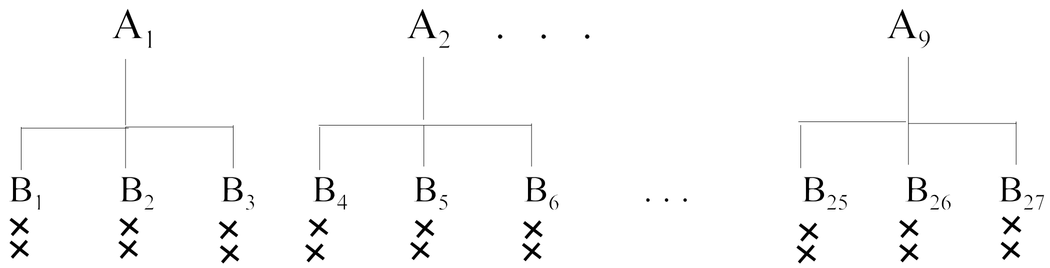
Hvis vi kun ønsker at undersøge de 3 bestemte olieblandinger og 2 karburatorer er analysen en **"fixed factor" sidedelt variansanalyse.**

Hvis man har et stort antal "ens" olieblandinger og man ønsker at undersøge den variation som olieblandingerne bevirker på benzinforbruget er analysen en **"mixed factor" sidedelt variansanalyse** med olieblanding som random factor og karburator som fixed factor.
Hvis man også udvælger de to karburatorer ud fra en stor mængde "ens" karburatorer er begge faktorer random og vi taler om en **sidedelt random factor variansanalyse.**

Trinvise niveauer. At en faktor B er undertrin (engelsk: nested factor) til faktor A's niveauer, betyder at ethvert niveau af B kun kombineres med ét niveau af A.

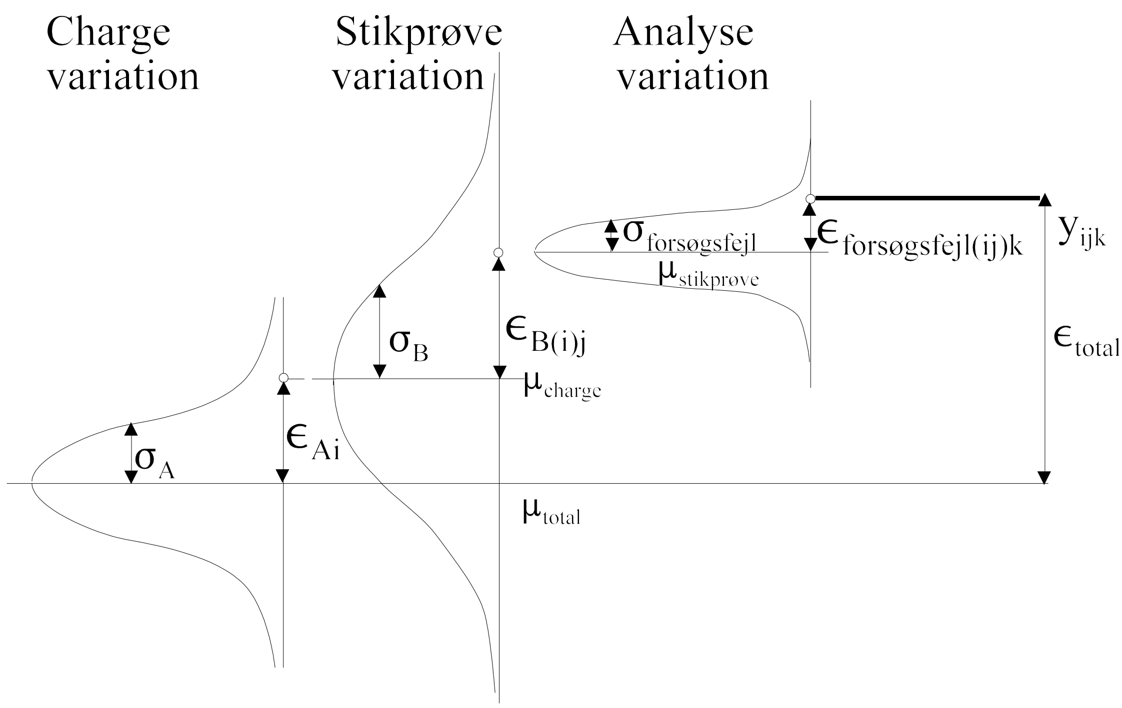
Et eksempel på en trinvis variansanalyse er følgende

Eksempel 3.1. (Trinvis variansanalyse) *I en kemisk chargeproduktion fremstilles et stof, der skal indeholde en vis procentdel vandopløselige komponenter. Driftskontrollen udføres ved, at man udtager en prøve af hver produceret charge og på denne foretager 2 uafhængige analyser af procentindholdet af vandopløseligt stof i prøven.*
De fremkomne tal udviser stor variation.
Skyldes denne variation a) *charger (fremstillingsprocessen)*
b) *prøver (måden man udtager prøverne på)*
c) *analysefejl (analysemetoden)*
Der foretages nu følgende forsøg for at få klarlagt dette.
1) 9 *charger* $A_1, A_2, \ldots, A_9$ *udtages tilfældigt.* $a = 9$
2) *Fra hver charge (bestående af* 80 *tønder) udtages* 3 *stikprøver* $b = 3$
(en tønde udtages tilfældigt og en prøve udtages
en ny tønde udtages tilfældigt og en prøve udtages , osv.)
dvs. i alt udtages 27 *stikprøver* $B_1, B_2, \ldots, B_{27}$
3) *På hver af de* $a \cdot b = 27$ *prøver foretages* 2 *uafhængige bestemmelser af % vandindhold.*
Grafisk kan forsøget fremstilles:



Det ses, at stikprøver er et undertrin (nested factor) til charger.
Begge faktorer ses at være random factors (hvilket er oftest forekommende), men kunne i andre situationer også være fixed eller en blanding af random og fixed factors.

Lad μ_{total} være middelværdien af vandindholdet og y_{ijk} være det målte vandindhold.
Den totale afvigelse $\varepsilon_{\text{total}} = y_{ijk} - \mu_{\text{total}}$ vil da være summen af 3 led
$\varepsilon_{\text{total}} = \varepsilon_{Ai} + \varepsilon_{B(i)j} + \varepsilon_{\text{forsøgsfejl}(ij)k}$
hvor de enkelte leds betydning fremgår af den følgende figur.



Ved en trinvis variansanalyse ønsker vi bl.a. at finde et skøn for σ_A , σ_B B, og σ error hvor σ error $= \sigma$ er forsøgsfejlens spredning.

3.3 SYMBOLIK

Vi vil for simpelheds skyld illustrere metoderne ved at betragte forsøg med kun 2 faktorer. For at vise hvorledes dette kan generaliseres til forsøg med flere faktorer betragtes til sidst nogle eksempler på forsøg med 3 faktorer
Lad faktorerne hedde A og B med niveauerne a og b.
En observation i en <u>sidedelt</u> variansanalyse (krydsede faktorer) kan skrives
$x_{ijk} = \mu + A_i + B_j + AB_{ij} + \varepsilon_{(ij)k}$
hvor A_i er hovedvirkningen for det i te niveau
B_j er hovedvirkningen for det j te niveau
AB_{ij} er vekselvirkningen mellem det i'te rækkeniveau og det j'te søjleniveau
og $\varepsilon_{(ij)k}$ er værdier af en normalfordelt statistisk variabel forsøgsfejlen med middelværdi 0 og varians σ^2.

En observation i en <u>trinvis</u> variansanalyse hvor B's niveauer er undertrin til A's niveauer kan skrives $x_{ijk} = \mu + A_i + B_{(i)j} + \varepsilon_{(ij)k}$
I $B_{(i)j}$ refererer parentesen til det trinvise j-niveau for fastholdt i.
I Statgraphics er symbolikken B(A), som tilsvarende tolkes at B's niveauer er undertrin (nestede) til A's.
<u>Bemærk, at ved trinvise variansanalyser har vekselvirkning mellem en faktor og dens undertrinfaktorer ingen mening.</u>

Som i det sidedelte tilfælde opstilles beregningsresultater sædvanligvis i en variansanalysetabel. I ovennævnte tilfælde vil den få følgende udseende:

Variation	f	SS	s^2	F
1.trins-faktor A	$f_A = a - 1$	SAK_A	s_A^2	
2.trins-faktor B	$f_B = a \cdot (b - 1)$	SAK_B	s_B^2	
Gentagelser	$f_0 = a \cdot b(c - 1)$	SAK_0	s_0^2	
Total	$f_{total} = a \cdot b \cdot c - 1$	SAK_{total}		

Hvorledes beregninger og analysen skal foregå vil fremgå af den følgende varianskomponentmetode. Her skal specielt bemærkes, at hvis 1. trin har a niveauer og undertrinnet B har b niveauer, så vil B have frihedsgradstallet $f_B = a \cdot (b - 1)$.
Dette følger af, at for hvert fastholdt niveau af A, er der b - 1 frihedsgrader for B.

3.4 VARIANSKOMPONENTMETODEN

Varianskomponentmetoden anvendes til at finde ud af, hvorledes testningerne skal foregå. Metoden kan anvendes på de i afsnit 3.2 nævnte modeller (fixed, random, sidedelt, trinvis, og kombinationer heraf, men altid med samme antal observationer i hver celle). Det skal bemærkes, at den ikke kan anvendes på romerske kvadratforsøg og ufuldstændige blokforsøg.

Definition af Varianskomponent.
Såfremt der i en variansanalyse forekommer

1) en fixed factor A med r niveauer defineres varianskomponenten σ_A^2 af hovedvirkningerne A_i ved

$$\sigma_A^2 = \frac{\sum_{i=1}^{r} A_i^2}{r - 1}$$

2) en random factor A med r niveauer defineres varianskomponenten σ_A^2 af hovedvirkningerne A som variansen af hovedvirkningen A_i d.v.s. $\sigma_A^2 = V(A_i)$.

Tilsvarende defineres varianskomponenter for vekselvirkninger i tilfælde, hvor begge faktorer er fixed, begge er random eller og det drejer sig om en mixed model. For nærmere detaljer henvises til L. Brøndum og J. D. Monrad: Forsøgsplanlægning I side 98.

Forsøgsfejlens varians σ^2 kaldes ofte forsøgsfejlens varianskomponent.

En nærmere beskrivelse af, hvorledes man kan konstruere varianskomponenterne kan findes i appendix 3.1.

Dette gøres efter faste regler, som det også er nemt at programmere, og derfor også kan findes i mange statistikprogammer. Dette er således tilfældet i Statgraphics's version 5.1.
Har man ikke et sådant program, så har B. Hellesen lavet såvel et Mapleprogram (kan ses i appendix 3.2) som et almindeligt Windows program (kaldet EMSwin.exe).
Vi vil i det følgende benytte det sidstnævnte program (da vi ikke i den benyttede version 4.0 af Statgraphics har denne facilitet til rådighed)
Man forstår nemmest anvendelsen af metoden ved at se på en række eksempler.

Eksempel 3.2. Trinvis variansanalyse *I eksempel* 3.1 *blev der foretaget følgende forsøg:*

1) *9 charger udtages tilfældigt* $a = 9$
2) *Fra hver charge (bestående af* 80 *tønder) udtages* 3 *stikprøver* $b = 3$
   (en tønde udtages tilfældigt og en prøve udtages
   en ny tønde udtages tilfældigt og en prøve udtages , osv.)
3) *På hver af de* $a \cdot b = 27$ *prøver foretages* 2 *uafhængige bestemmelser af % vandindhold.*

Benyt EMSwin.exe programmet til at udarbejde en varianskomponenttabel, og forklar hvorledes testningerne skal foregå.

LØSNING:
Man går ind i "Stifinder" og find programmet EMSwin.exe på sin harddisk eller diskette.
Vælg("Generate Formulas" | Input_Output | Tryk evt. på pilen til højre for at se forskellige valgmuligheder)
Da vi skriver B(A), hvis B er et undertrin for A, så skriv (eller ret en af de foreslåede modeller):
A RAN 9 B RAN 3 EPS RAN 2 mu+A+B(A)
[her er forkortheds skyld A = charge og B = stikprøver]
Tryk på "Generate Formulas", Vi får følgende udskrift på skærmen, som samtidig genereres i en tekstfil i samme mappe med navnet EMSout67.txt. Herfra kan den printes ud.
Resultatet bliver:

```
The input:
A Ran 9   B Ran 3    Eps Ran 2     mu + A + B(A)
gives the output:

  EMS(A) = 6*V(A) + 2*V(B) + V(Eps)
  EMS(B) = 2*V(B) + V(Eps)
  EMS(Eps) = V(Eps)
```

Den sidste matrix er kun beregnet af hensyn til beregningerne, og har ingen betydning ved tolkningen.
Udskriften oversættes til.

$$E(s_A^2) = 6 \cdot \sigma_A^2 + 2 \cdot \sigma_B^2 + \sigma^2$$

$$E(s_B^2) = 2 \cdot \sigma_B^2 + \sigma^2$$

$$E(s_0^2) = \sigma^2$$

Vi ser, at man kan teste B mod forsøgsfejlen EMS(EPS). (EMS = Expected Mean Square)
Hvis man ved den første testning opdager at B = 0 kan man poole B ned i "residualen" for at styrke testen, og derefter teste A mod den fremkomne "residual".
Hvis B $\neq$ 0 kan man stadig teste A mod B. Opdager man her, at A = 0, må man ikke poole A ned i "forsøgsfejlen", da EMS(A) jo netop ikke er et udtryk for denne.

◆

Eksempel 3.3. Mixed variansanalyse .
Lad os antage, at vi ønsker at finde ud hviken af 2 bestemte karburatorer der giver det laveste benzinforbrug. Da olieblandingerne kan tænkes at spille en rolle for variationen, så udvælges tilfældigt blandt mange muligheder 3 olieblandinger. Karburator er følgelig en "fixed factor" og olieblandinger er en "random factor".
Der ønskes lavet følgende mixed sidedelt faktorforsøg:

		Karburator	
		K_1	K_2
Olieblanding	O_1	× ×	× ×
	O_2	× ×	× ×
	O_3	× ×	× ×

Benyt EMSwin.exe programmet til at udarbejde en varianskomponenttabel, og forklar hvorledes testningerne skal foregå.
LØSNING:
I "Stifinder" aktiveres programmet EMSwin.exe
Vælg("Generate Formulas" | Input_Output | Tryk evt. på pilen til højre for at se forskellige valgmuligheder)
Skriv:
K Fix 2 O RAN 3 EPS RAN 2 mu + K + O + K*O
Tryk på "Generate Formulas", Vi får følgende udskrift:

```
The input:
K Fix 2   O Ran 3   Eps Ran 2    mu + K + O + K*O # Press GENERATE FORMULAS

gives the output:

   EMS(K)  = 6*f(K)  + 2*V(K*O)  + V(Eps)
   EMS(O)  = 4*V(O)  + V(Eps)
   EMS(K*O)  = 2*V(K*O)  + V(Eps)
   EMS(Eps)  = V(Eps)
```

En variansanalysetestning vil bestå i, at først dannes F størrelsen $F_{KO} = \dfrac{s_{KO}^2}{s_0^2} = \dfrac{\text{EMS(K*O)}}{\text{EMS(EPS)}}$.

Fås en accept af H_0 : K*O = 0 kan man poole og teste om der er en hovedvirkning K.

Fås forkastelse, kan man dog danne F størrelsen $F_K = \dfrac{s_K^2}{s_{KO}^2} = \dfrac{\text{EMS(K)}}{\text{EMS(K*O)}}$, og derved undersøge om nulhypotesen H_0 : K = 0 kan accepteres.

Det ses, at en testning af H_0 : O = 0 kan foretages ved $F_O = \dfrac{s_O^2}{s_0^2} = \dfrac{\text{EMS(O)}}{\text{EMS(EPS)}}$ uanset om K*O = 0. Det har ofte ikke nogen praktisk interesse i en mixed factor analyse, men kan dog vise om det er nødvendigt at tage hensyn til O's tilstedeværelse. O har som "Random faktor" mest betydning som en støjfaktor, og har derfor næppe nogen selvstændig interesse

Ved sidedelte variansanalyser er beregningerne de samme som den velkendte variansanalyse med lutter fixed factors. Derimod er testningerne som ovenfor antydet forskellig og bestemmes ud fra varianskomponentmetoden

Hvorledes man i Statgraphics foretager de nødvendige beregninger vil fremgå af de følgende eksempler.

Eksempel 3.4. Beregning af en trinvis variansanalyse
I eksempel 3.1 *og* 3.2 *blev følgende forsøg beskrevet:*

1) 9 *charger udtages tilfældigt* $a = 9$
2) *Fra hver charge (bestående af* 80 *tønder) udtages* 3 *stikprøver* $b = 3$
   (en tønde udtages tilfældigt og en prøve udtages
   en ny tønde udtages tilfældigt og en prøve udtages , osv.)
3) *På hver af de* $a \cdot b = 27$ *prøver foretages* 2 *uafhængige bestemmelser af % vandindhold.*

Resultaterne af forsøget fremgår af den følgende udskrift fra Statgraphics.

Charger	Proever	Analyser	Procent_va
1	1	1	50,3
1	1	2	49,8
1	2	1	50,1
1	2	2	49,5
1	3	1	51,1
1	3	2	49,4
2	1	1	45,8
2	1	2	45,4
2	2	1	44,4
2	2	2	44,7
2	3	1	44,7
2	3	2	44,4
3	1	1	41
3	1	2	41,4
3	2	1	42,7
3	2	2	41,6
3	3	1	43,1
3	3	2	43,3
4	1	1	48,7
4	1	2	50
4	2	1	48
4	2	2	50,4
4	3	1	47,9
4	3	2	48,5
5	1	1	48,9
5	1	2	49,4
5	2	1	48,4
5	2	2	46,8
5	3	1	46,5
5	3	2	45,4
6	1	1	47
6	1	2	46,1
6	2	1	47,4
6	2	2	47,2
6	3	1	45,1
6	3	2	47,5
7	1	1	46,3
7	1	2	45
7	2	1	44,6
7	2	2	44
7	3	1	45,6
7	3	2	44,2
8	1	1	44,9
8	1	2	42,3
8	2	1	45,1
8	2	2	43,4
8	3	1	43,3
8	3	2	41,6
9	1	1	55,7
9	1	2	55,4
9	2	1	56,3
9	2	2	56,3
9	3	1	55,1
9	3	2	55

Foretag en statistisk analyse af forsøget med henblik på at finde ud af hvilke faktorer der spiller en stor rolle for variationen.

LØSNING:
I eksempel 3.2 er varianskomponentskemaet og testrækkefølgen angivet.
I Statgraphics foretages nu beregningerne på følgende måde.
Vælg (Special | Advanced Regression | General Linear Metods (GLM)) .
En genvej for ovenstående er at vælge den lilla GLM-ikon i den næstøverste menubjælke.
Den fremkomne menu udfyldes med
Dependent variable : Procent_va
Categorical Factors : charger, proever (ikke analyser der er gentagelserne). OK
På næste menu, der hedder "GLM Model Specification" skrives under "Effects" A og B(A) (betyder at B's niveauer skal betragtes som undertrin til A's.)
Resultatet ses i følgende variansanalysetabel:

```
General Linear Models
---------------------
Number of dependent variables: 1
Number of categorical factors: 2
Number of quantitative factors: 0
Analysis of Variance for PROCENT_VA
-----------------------------------------------------------------------------
Source              Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-----------------------------------------------------------------------------
Model                      818,185     26      31,4686      42,12       0,0000
Residual                     20,17     27     0,747037
-----------------------------------------------------------------------------
Total (Corr.)              838,355     53
Type III Sums of Squares
-----------------------------------------------------------------------------------
Source                     Sum of Squares     Df   Mean Square    F-Ratio   P-Value
-----------------------------------------------------------------------------------
CHARGER                           792,881      8       99,1102     132,67    0,0000
PROEVER(CHARGER)                  25,3033     18       1,40574       1,88    0,0668
Residual                            20,17     27      0,747037
-----------------------------------------------------------------------------------
Total (corrected)                 838,355     53
```

Bemærk, at frihedsgradstallet for prøver er a (b-1) = 9 (3-1) = 18. (Til de 3 prøver der udtages af en charge svarer 2 frihedsgrader, og der var i alt 9 charger).
Resultatet viser, at prøverne er tæt ved at være signifikante (p-value = 6.68%). Imidlertid er det tydeligt, at Charger spiller en langt større rolle for variationen, så det er her man bør sætte ind på forbedringer. Man kunne poole, men det vil ikke ændre konklusionen, da frihedsgradstallet allerede er stort.
Et estimat (skøn) for variationen σ_B^2 der skyldes stikprøverne og for variationen der skyldes chargerne σ_A^2 kan fås af ligningssystemet:

```
EMS(A)  = 6*V(A) + 2*V(B) + V(Eps)
EMS(B)  = 2*V(B) + V(Eps)
EMS(Eps) = V(Eps)
```

Vi har EMS(Eps) = V(Eps) = 0.7470

$$\text{EMS(B)} = 2\cdot \text{V(B)} + \text{V(Eps)} \Leftrightarrow 1.4057 = 2\cdot \text{V(B)} + 0.7470 \Leftrightarrow \text{V(B)} = \sigma_B^2 = 0.3293$$

$$\text{EMS(A)} = 6\cdot \text{V(A)} + 2*\text{V(B)} + \text{V(Eps)} \Leftrightarrow 99.11 = 6\cdot \text{V(A)} + 2\cdot 0.3293 + 0.7470$$

$$\text{V(A)} = \sigma_A^2 = 16.28$$

I appendix 3.2 er angivet hvorledes beregningerne kan foretages med regnemaskine.
Vi ser ikke uventet, at charger giver langt den største variation, og at det derfor er her man skal sætte ind ved at prøve at homogenisere chargerne. ◆

Øvelse 3.1 (trinvis variansanalyse) Regn opgave 21 i opgavesamlingen side 243.

Til yderligere illustration af varianskomponentmetode med tilhørende variansanalyse betragtes følgende eksempel.

Eksempel 3.5 (sidedelt variansanalyse) *Det ønskes undersøgt hvorledes produktionsudbyttet ved en kemisk proces afhænger af råvareleverancerne (der er tale om løbende råvareleverancer i mindre partier) og af katalysatorer (der foreligger et stort antal "ens" katalysatorer. Forsøget udførtes ved at man udtog 5 råvarepartier og 4 katalysatorer tilfældigt. De to faktorer er "random factors" og derfor vil vi undersøge den variation som råvareleverancerne og katalysatorerne bevirker i produktionsudbyttet.Da vekselvirkning ingen praktisk betydning har når begge faktorer er random variable, vælges for at spare forsøg n = 1.*
Forsøgsresultaterne var:

		Katalysatorer B			
		B_1	B_2	B_3	B_4
Råmaterialer A	A_1	40	90	50	80
	A_2	20	80	40	50
	A_3	40	60	60	40
	A_4	50	50	60	60
	A_5	40	90	70	50

Foretag en statistisk analyse af forsøget med henblik på at finde ud af hvilke faktorer der spiller en stor rolle for variationen.

LØSNING:
Først opskrives varianskomponentskemaet:

```
The input:
A Ran 5   B Ran 4   Eps Ran 1     mu + A + B + A*B

gives the output:

  EMS(A) = 4*V(A) + V(A*B) + V(Eps)
  EMS(B) = 5*V(B) + V(A*B) + V(Eps)
  EMS(A*B) = V(A*B) + V(Eps)
  EMS(Eps) = V(Eps)
```

Vi ser, at uanset om der er vekselvirkning eller ej, kan vekselvirkningsleddet benyttes til at teste hovedvirkninger.
Da det er en sidedelte variansanalyse udføres de sædvanlige beregninger.
Ved benyttelse af Statgraphics "Multifactor Analysis" fås

```
Analysis of Variance for udbytte - Type III Sums of Squares
--------------------------------------------------------------------------------
Source                  Sum of Squares     Df    Mean Square    F-Ratio    P-Value
--------------------------------------------------------------------------------
MAIN EFFECTS
 A:raamateria                  930,0      4          232,5       1,21     0,3575
 B:katalysat                  3240,0      3         1080,0       5,61     0,0122
RESIDUAL                      2310,0     12          192,5
--------------------------------------------------------------------------------
TOTAL (CORRECTED)             6480,0     19
--------------------------------------------------------------------------------
All F-ratios are based on the residual mean square error.
```

<u>Variansanalysetabellen viser, at katalysatorerne spiller en stor rolle for variationen, mens råmaterialer ikke giver noget væsentligt bidrag.</u>

Et resultat af undersøgelsen må derfor være at man nærmere undersøger om det er muligt at "homogenisere" katalysatorerne.

Et skøn for variationen af B : σ_B^2 =V(B) kan fås af ligningssystemet:

EMS(B) = 5*V(B) + V(A*B) + V(Eps),
EMS(A*B) = V(A*B) + V(Eps)
EMS(A*B) = V(A*B) + V(Eps) = 192.5 $\Rightarrow$

EMS(B) = 5*V(B) + V(A*B) + V(Eps) $\Leftrightarrow$ 1080 = 5 * V(B) + 192.5 $\Leftrightarrow$ $\underline{\underline{V(B) = \sigma_B^2 = 177.5}}$

Analogt kan EMS(A) findes af EMS(A) = 4*V(A) + V(A*B) + V(Eps)

Vi har 232.5 = 4 * V(A) + 192.5 hvilket giver $\underline{\underline{V(A) = \sigma_A^2 = 10}}$.

3.5 SPLIT -PLOT FORSØG.

Denne type forsøg er et specialtilfælde af de i kapitel 3 angivne strukturer, men da de forekommer ofte i praksis, vil disse Split-Plot forsøg blive beskrevet mere udførligt i dette afsnit.

3.5.1. PROBLEMSTILLING

I forsøg, hvor niveauskift for den ene faktor er væsentlig besværligere end for den anden faktor kan Split-plot forsøg være en mulighed.

Eksempel 3.6 (landbrugsforsøg) *I et landbrugsforsøg ønskes undersøgt virkningen af 3 pløjemetoder* P_1, P_2, P_3 *og 5 kunstgødninger* G_1, G_2, G_3, G_4, G_5 *på udbyttet af en bestemt kornsort. Der ønskes 2 gentagelser af hver behandling, dvs. i alt 30 delforsøg.*

Forsøgsplan 1: Fuldstændigt randomiseret forsøg. En stor mark deles op i 30 markstykker (parceller) som er vor forsøgsenhed. Et fuldstændig randomiseret forsøg, hvor man tilfældigt tilordner en parcel en pløjemetode og en gødning kunne se ud som følgende skitse:

$P_2\ G_4$	$P_1\ G_2$	$P_3\ G_3$	$P_1\ G_5$	$P_1\ G_1$	$P_2\ G_3$
$P_1\ G_2$	$P_2\ G_5$	$P_1\ G_4$	$P_1\ G_3$	$P_3\ G_3$	$P_3\ G_1$
$P_3\ G_4$	$P_2\ G_1$	$P_3\ G_2$	$P_2\ G_2$	$P_2\ G_4$	$P_3\ G_1$
$P_2\ G_3$	$P_1\ G_5$	$P_3\ G_5$	$P_2\ G_5$	$P_3\ G_2$	$P_2\ G_1$
$P_3\ G_5$	$P_1\ G_3$	$P_3\ G_4$	$P_1\ G_4$	$P_2\ G_2$	$P_1\ G_1$

Analysemetoden er naturligvis en tosidet variansanalyse.

	G_1	G_2	G_3	G_4	G_5
P_1	- -	- -	- -	- -	- -
P_2	- -	- -	- -	- -	- -
P_3	- -	- -	- -	- -	- -

Variansanalyse	SAK	f
Gødning		4
Pløjning		2
Gødn x pløjn		8
Gentagelser		15
Total		29

Forsøgsplan 2: Split-Plot forsøg. Da pløjning må foregå i lange strimler er et fuldstændigt randomiseret forsøg i praksis umuligt.
I stedet deles marken op i 6 strimler (**hovedplots**), og pløjemetoderne randomiseres til hver af disse. Derefter deles hver af disse 6 strimler i 5 lige store parceller (**split-plots**), hvortil de 5 gødninger randomiseres Den resulterende split-plot forsøgsplan kunne eksempelvis se ud som vist på den følgende skitse:

Strimmel 1	Strimmel 2	Strimmel 3	Strimmel 4	Strimmel 5	Strimmel 6
$P_3\ G_3$	$P_1\ G_5$	$P_1\ G_3$	$P_2\ G_4$	$P_3\ G_3$	$P_2\ G_1$
$P_3\ G_2$	$P_1\ G_1$	$P_1\ G_4$	$P_2\ G_5$	$P_3\ G_2$	$P_2\ G_3$
$P_3\ G_5$	$P_1\ G_2$	$P_1\ G_2$	$P_2\ G_1$	$P_3\ G_5$	$P_2\ G_5$
$P_3\ G_4$	$P_1\ G_4$	$P_1\ G_5$	$P_2\ G_3$	$P_3\ G_1$	$P_2\ G_4$
$P_3\ G_1$	$P_1\ G_3$	$P_1\ G_1$	$P_2\ G_2$	$P_3\ G_4$	$P_2\ G_2$

Nedenstående skema giver en oversigt over de benyttede betegnelser:

Forsøgsenheder		Faktorer	
Hovedplot = strimler jord	6 hovedplot	Hovedplotfaktor = pløjemetoder	3 niveauer
Split-plot = parceller	5 pr. hoved-plot	Underplotfaktor = gødningsmetoder	5 niveauer

Umiddelbart må det antages, at variationen mellem hovedplot (de forskellige strimler mark) er større end mellem split plot (de enkelte parceller indenfor en markstrimmel), da markstrimlerne ligger temmelig langt fra hinanden og derfor har forskellig bonitet, mens de enkelte parceller indenfor en markstrimmel har en mere ensartet bonitet.
Det synes derfor rimeligt, at den samlede forsøgsfejl kan opdeles i to komponenter:

Forsøgsfejl 1: Variation mellem hovedplot, som anvendes til vurdering af hovedfaktors niveauer

Forsøgsfejl 2: Variation mellem split-plot indenfor hovedplot, som anvendes til vurdering af underplotfaktors niveauer

3.5.2. Opstilling af model og beregning af et Split-plot forsøg

Eksempel 3.7. (eksempel 3.6 fortsat).
I et landbrugsforsøg ønskes undersøgt virkningen af 3 pløjemetoder P_1, P_2, P_3 og 5 kunstgødninger G_1, G_2, G_3, G_4, G_5 på udbyttet af en bestemt kornsort. Forsøgsplanen er et Split-plot forsøg, og resultaterne af forsøget ses i følgende skema:

	Pløjemetoder	P1		P2		P3	
	Mark	M2	M3	M4	M6	M1	M5
Gødning	G1	29.1	41.3	22.0	25.6	40.0	35.1
	G2	31.5	42.0	21.1	27.3	45.3	39.7
	G3	36.2	44.5	25.1	23.7	40.0	34.9
	G4	28.2	36.8	20.4	22.7	42.1	38.2
	G5	39.7	36.3	31.2	29.7	50.2	44.3

LØSNING:
1) Opstilling af model:
Forsøgene kan beregningsmæssigt beskrives som et forsøg, hvor man har en blanding af en sidedelt og en trinvis variansanalyse. Markstrimlerne er undertrin af pløjemetoderne, mens gødningsmetoderne er sidedelte i forhold til såvel pløjemetoder som markstrimler. Varianskomponentmetoden er derfor velegnet at benytte ved analysen af sådanne forsøg.
Den fuldstændige model for forsøget er (skrevet med Statgraphic symbolik)

$$Y = \mu + P + M(P) + G + P*G + M(P)*G + \varepsilon$$

eller $$y_{ijkm} = \mu + P_i + M_{(i)j} + G_k + PG_{ik} + MG_{(i)jk} + \varepsilon_{(ijk)m}.$$

2) Varianskomponenttabel:
Idet P og G er fixed og M er random fås ved benyttelse af EMS-program.

```
The input:
P Fix 3  M Ran 2   G Fix 5   Eps Ran 1     mu + P + M(P) + G+P*G+M(P)*G
gives the output:
   EMS(P)  = 10*f(P)  + 5*V(M)  + V(Eps)
   EMS(M)  = 5*V(M)  + V(Eps)
   EMS(G)  = 6*f(G)  + V(M*G)  + V(Eps)
   EMS(P*G)  = 2*f(P*G)  + V(M*G)  + V(Eps)
   EMS(M*G)  = V(M*G)  + V(Eps)
   EMS(Eps)  = V(Eps)
```

Af hensyn til den efterfølgende planlægning opskrives nu en variansanalysetabel med frihedsgrader:

	f
P	$3 - 1 = 2$
M(P)	$3 \cdot (2 - 1) = 3$
G	$5 - 1 = 4$
P*G	$2 \cdot 4 = 8$
M*G	$3 \cdot 4 = 12$
Gentagelser	0
Total	$30 - 1 = 29$

3) Beregninger med benyttelse af Statgraphics.
Data indtastes på sædvanlig måde:

Ploejemetode	Markstrimmel	Goedning	Udbytte
1	2	1	29,1
1	2	2	31,5
1	2	3	36,2
1	2	4	28,2
1	2	5	39,7
1	3	1	41,3
1	3	2	42
1	3	3	44,5
1	3	4	36,8
1	3	5	36,3
2	4	1	22
2	4	2	21,1
2	4	3	25,1
2	4	4	20,4
2	4	5	31,2
2	6	1	25,6
2	6	2	27,3
2	6	3	23,7
2	6	4	22,7
2	6	5	29,7
3	1	1	40
3	1	2	45,3
3	1	3	40
3	1	4	42,1
3	1	5	50,2
3	5	1	35,1
3	5	2	39,7
3	5	3	34,9
3	5	4	38,2
3	5	5	44,3

Vælg (lilla GLM ikon) eller (Special | Advanced Regression | General Linear Methods).
Den fremkomne tabel udfyldes med
Dependent Variable : Udbytte
Chategorical Factors : ploejemetode, markstrimmel , goedning,
I det følgende vil de variable noget uhensigtsmæssigt blive omdøbt til A, B, C osv.
OK
På næste tavle der hedder "GLM Model Specifikation" skrives
A (svarer til ploejemetode)
B(A) (svarer til at markstrimmel er undertrin af ploejemetode)
C (svarer til goedning)
A*C
Derimod skrives ikke B(A)*C selvom det var det næste led i modellen. Dette skyldes, som det fremgår af søjlen over frihedsgrader i varianstabellen, at så vil residualen blive 0, da der ingen gentagelser er til bestemmelse af forsøgsfejlens varians σ^2
Vi bruger derfor det sidste led som "støjled".
Vi får nu følgende udskrift:

```
General Linear Models
---------------------
Number of dependent variables: 1
Number of categorical factors: 3
Number of quantitative factors: 0

Analysis of Variance for udbytte
-----------------------------------------------------------------------------
Source              Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-----------------------------------------------------------------------------
Model                      1878,56     17      110,503      13,41       0,0000
Residual                    98,876     12      8,23967
-----------------------------------------------------------------------------
Total (Corr.)              1977,43     29
```

```
Type III Sums of Squares
--------------------------------------------------------------------------------
Source                      Sum of Squares     Df   Mean Square    F-Ratio    P-Value
--------------------------------------------------------------------------------
ploejemetoder                       1383,9      2       691,948      83,98     0,0000
markstrimmel(ploejemetoder         204,024      3        68,008       8,25     0,0030
goedning                           186,329      4       46,5822       5,65     0,0085
ploejemetoder*goedning             104,307      8       13,0384       1,58     0,2283
Residual                            98,876     12       8,23967
--------------------------------------------------------------------------------
Total (corrected)                  1977,43     29
All F-ratios are based on the residual mean square error.

R-Squared = 94,9998 percent    R-Squared (adjusted for d.f.) = 87,9161 percent
```

Som det fremgår af varianskomponentskemaet, kan man benytte residualen til at teste vekselvirkning mellem ploejemetode og goedning.
<u>Det ses, at vekselvirkningen</u> ploejemetode · goedning <u>er 0.</u>

Vi kan nu foretage en pooling.
Vælg (gul ikon = input dialog | OK | fjern A*C | OK)
Af varianskomponenttabellen fremgår, at B(A) = markstrimmel(ploejemetode) kan benyttes til testning af ploejemetode.
Vælg (Cursor i udskrift| højre musetast | Analysis options).
I det fremkomne felt under "Factor" sættes blå streg på A. I feltet med Error Term vælges B(A). Derved ses i nederste felt "Selections" at der står A - B(A) som tegn på at A testes mod B(A))
Resultatet blev:

```
General Linear Models
---------------------
Analysis of Variance for udbytte
-----------------------------------------------------------------------------
Source                Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-----------------------------------------------------------------------------
Model                        1774,25      9      197,139      19,41       0,0000
Residual                     203,183     20      10,1592
-----------------------------------------------------------------------------
Total (Corr.)                1977,43     29

Type III Sums of Squares
--------------------------------------------------------------------------------
Source                      Sum of Squares     Df   Mean Square    F-Ratio    P-Value
--------------------------------------------------------------------------------
ploejemetoder                       1383,9      2       691,948   10,17(1)    0,0461
markstrimmel(ploejemetoder         204,024      3        68,008    6,69(0)    0,0026
goedning                           186,329      4       46,5822    4,59(0)    0,0086
Residual                           203,183     20       10,1592
--------------------------------------------------------------------------------
Total (corrected)                  1977,43     29
F-ratios are based on the following mean squares:
   (0) Residual
   (1) B(A)
```

<u>Faktor G (goedning) har en hovedvirkning (2 stjernet signifikans)</u>

<u>Der er et svagt statistisk bevis for at faktor P (ploejemetode) har en hovedvirkning</u> .

Den store variation der er konstateret fra markstrimme til markstrimmel bevirker, at vi er tæt på ikke at kunne bevise, at der er forskel på pløjemetoder.

◆

En meget hyppig fejl er, at man af praktiske grunde tvinges til at lave forsøget som et split-plot forsøg, men tror det er et fuldstændigt randomiseret forsøg som beskrevet tidligere. Analysen bliver så en almindelig tosidet variansanalyse, hvis resultat kan blive ganske misvisende.
Der gives derfor endnu et eksempel på et split-plot forsøg (udarbejdet af .Bjarne Hellesen):

Eksempel 3.8 (Split-plot, fixed og random, sidedelt[crossed] og trindelt[nested])
For at opnå den højest mulige kvalitet Y af et levnedsmiddel ønsker man at finde
*den bedste fremstillings**metode** (A1 eller A2 ?) og*
*den bedste fagforenings**gruppe** at hente medarbejdere fra (B1, B2 eller B3?).*
*Arbejdet udføres af arbejds**hold**. Der udtages tilfældigt 3 hold fra hver fagforeningsgruppe, altså 9 hold i alt (C1, C2, ..., C9: 9 "hovedplots"). Man ønsker at analysere den variation, som holdene tilfører kvaliteten. I tilfældig rækkefølge lader man hvert af de 9 hold fremstille levnedsmidlet 2 gange med hver metode (4 "split-plots" i hvert hovedplot). De målte data for kvaliteten var:*

	Gruppe B1			Gruppe B2			Gruppe B3		
Hold	C1	C2	C3	C4	C5	C6	C7	C8	C9
Metode A1	20.2 24.1	26.2 26.9	23.8 24.9	22.0 23.5	22.6 24.6	22.9 25.0	23.1 22.9	22.9 23.7	21.8 23.5
Metode A2	14.2 16.2	18.0 19.1	12.5 15.4	14.1 16.1	14.0 18.1	13.7 16.0	14.1 16.1	12.2 13.8	12.7 15.1

Modellen for forsøget er $Y = \mu + A + B + A*B + C(B) + A*C(B) + \varepsilon$

dvs. $y_{ijkm} = \mu + A_i + B_j + AB_{ij} + C_{(j)k} + AC_{i(j)k} + \varepsilon_{(ijk)m}$

a) *Udarbejd en tabel for varianskomponentsammensætningen.*
b) *Undersøg, om der er vekselvirkning mellem metoder A og hold i grupper C(B).*
c) *Undersøg, om der er vekselvirkning mellem metoder A og grupper B.*
d) *Undersøg (hvor det er relevant), om der er hovedvirkninger A, B og C(B).*
e) *Hvordan kan man i fremtiden få højest kvalitet?*
f) *Estimér varianskomponenterne* σ_C^2 *og* σ_{AC}^2.

LØSNING:
a) Ved at benytte EMS-programmet fås :

```
The input:
A Fix 2   B Fix 3   C Ran 3  Eps Ran 2     mu + A + B + A*B+C(B) +A*C(B)
gives the output:
   EMS(A) = 18*f(A) + 2*V(A*C) + V(Eps)
   EMS(B) = 12*f(B) + 4*V(C) + V(Eps)
   EMS(A*B) = 6*f(A*B) + 2*V(A*C) + V(Eps)
   EMS(C) = 4*V(C) + V(Eps)
   EMS(A*C) = 2*V(A*C) + V(Eps)
   EMS(Eps) = V(Eps)
```

b) Data for forsøget indtastes i Statgraphics:

Metoder	Grupper	Hold	Respons
1	1	1	20,2
1	1	1	24,1
1	1	2	26,2
1	1	2	26,9
1	1	3	23,8
1	1	3	24,9
1	2	4	22
1	2	4	23,5
1	2	5	22,6
1	2	5	24,6
1	2	6	22,9
1	2	6	25,0
1	3	7	23,1
1	3	7	22,9
1	3	8	22,9
1	3	8	23,7
1	3	9	21,8
1	3	9	23,5
2	1	1	14,2
2	1	1	16,2
2	1	2	18
2	1	2	19,1
2	1	3	12,5
2	1	3	15,4
2	2	4	14,1
2	2	4	16,1
2	2	5	14,0
2	2	5	18,1
2	2	6	13,7
2	2	6	16
2	3	7	14,1
2	3	7	16,1
2	3	8	12,2
2	3	8	13,8
2	3	9	12,7
2	3	9	15,1

I GLM vælges:
Dependent variable : Respons
Categorical factors : Metoder, Grupper og Hold (som Statgraphics kalder A, B og C)
GLM specifications : A , B, A*B, C(B), A*C(B) (komponenterne i vores model)
Så fås:

```
General Linear Models
---------------------
Analysis of Variance for Respons
-----------------------------------------------------------------------------
Source               Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-----------------------------------------------------------------------------
Model                        719,17     17      42,3041      18,31       0,0000
Residual                      41,59     18      2,31056
-----------------------------------------------------------------------------
Total (Corr.)                760,76     35

Type III Sums of Squares
--------------------------------------------------------------------------------
Source                        Sum of Squares     Df   Mean Square     F-Ratio    P-Value
--------------------------------------------------------------------------------
Metoder                             651,951      1       651,951      282,16     0,0000
Grupper                             16,0517      2       8,02583        3,47     0,0530
Metoder*Grupper                     1,18722      2      0,593611        0,26     0,7762
Hold(Grupper)                       39,2583      6       6,54306        2,83     0,0403
Metoder*Hold(Grupper)               10,7217      6       1,78694        0,77     0,6009
Residual                              41,59     18       2,31056
--------------------------------------------------------------------------------
Total (corrected)                    760,76     35
All F-ratios are based on the residual mean square error.
```

Af $E(s^2)$-varianskomponentsammensætningen ses, at tilstedeværelse af vekselvirkningen A*C(B) kan testes ved at udregne dens F-ratio med "Residual". Statgraphics har fundet F=0,77 svarende til en P-Value 0,6009 større end 0,05 . Så der er

ikke tegn på vekselvirkning mellem metoder og hold i grupper.

c) Den manglende vekselvirkning A*C(B) pooles ned i residualen:
Vælg ("Input dialog" og slet virkningen A*C(B) i vinduet "GLM Model Specifications".
Vi får da en ny variansanalysetabel:

Type III Sums of Squares

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Metoder	651,951	1	651,951	299,11	0,0000
Grupper	16,0517	2	8,02583	3,68	0,0403
Metoder*Grupper	1,18722	2	0,593611	0,27	0,7639
Hold(Grupper)	39,2583	6	6,54306	3,00	0,0247
Residual	52,3117	24	2,17965		
Total (corrected)	760,76	35			

All F-ratios are based on the residual mean square error.

Af $E(s^2)$-varianskomponentsammensætningen fra spørgsmål a) ses, at tilstedeværelse af vekselvirkningen A*B kan testes ved at udregne dens F-ratio med "Residual". Statgraphics har fundet F=0,27 svarende til en P-Value 0,7639 større end 0,05 . Så der er
ikke tegn på vekselvirkning mellem metoder og grupper.

d) Den manglende vekselvirkning A*B pooles ned i residualen:
Vælg ("Input dialog" og slet virkningen A*B i vinduet "GLM Model Specifications".
Vi får da en ny variansanalysetabel:

Type III Sums of Squares

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Metoder	651,951	1	651,951	316,84	0,0000
Grupper	16,0517	2	8,02583	3,90	0,0330
Hold(Grupper)	39,2583	6	6,54306	3,18	0,0178
Residual	53,4989	26	2,05765		
Total (corrected)	760,76	35			

All F-ratios are based on the residual mean square error.

Af $E(s^2)$-varianskomponentsammensætningen fra spørgsmål a ses, at tilstedeværelse af
C(B) kan testes ved dens F-ratio med "Residual". Vi finder
C(B): Hold i grupper har en virkning (P_value=0.0178<0.05, dvs. 1-stjernet signifikans).

Af $E(s^2)$-varianskomponentsammensætningen fra spørgsmål a ses, at tilstedeværelse af
A kan testes ved dens F-ratio med "Residual" (idet A*C(B) = 0)
B kan testes ved dens F-ratio med C(B)
Dette giver følgende variansanalysetabel:

Type III Sums of Squares

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Metoder	651,951	1	651,951	316,84(0)	0,0000
Grupper	16,0517	2	8,02583	1,23(1)	0,3576
Hold(Grupper)	39,2583	6	6,54306	3,18(0)	0,0178
Residual	53,4989	26	2,05765		
Total (corrected)	760,76	35			

F-ratios are based on the following mean squares:
(0) Residual
(1) C(B)

Vi finder heraf:
Metoder har en virkning (P-value =0,0000 < 0,001, dvs. 3-stjernet signifikans)
Ingen tegn på at grupper har en virkning (P-value =0,3576 > 0,05), men svarende til at vi ikke har testet B mod Residual, så pooler vi ikke B ned i residualen, fordi $E(s_B^2) = 12\sigma_B^2 + 4\sigma_C^2 + \sigma$ også indeholder en stor komponent $\sigma_C^2 \neq 0$, som ikke skal have lov til at ødelægge residualen.

e) Vælges (den gule ikon "Tabular Options" | Table of Means) fås:

Table of Least Squares Means for Respons with 95,0 Percent Confidence Intervals

Level		Count	Mean	Stnd. Error	Lower Limit	Upper Limit
GRAND MEAN		36	19,3333	0,239075	18,8419	19,8248
Metoder						
1		18	23,5889	0,338103	22,8939	24,2839
2		18	15,0778	0,338103	14,3828	15,7728
Grupper						
1		12	20,125	0,738414	18,3182	21,9318
2		12	19,3833	0,738414	17,5765	21,1902
3		12	18,4917	0,738414	16,6848	20,2985
Hold within Grupper						
1	1	4	18,675	0,717225	17,2007	20,1493
2	1	4	22,55	0,717225	21,0757	24,0243
3	1	4	19,15	0,717225	17,6757	20,6243
4	2	4	18,925	0,717225	17,4507	20,3993
5	2	4	19,825	0,717225	18,3507	21,2993
6	2	4	19,4	0,717225	17,9257	20,8743
7	3	4	19,05	0,717225	17,5757	20,5243
8	3	4	18,15	0,717225	16,6757	19,6243
9	3	4	18,275	0,717225	16,8007	19,7493

Det ses heraf (og af spørgsmål d), at det vigtigste er at vælge metode 1, da den giver langt højere kvalitet end metode 2.
Endvidere ses, at den påviste variation fra hold hovedsagelig skyldes, at hold 2 giver noget højere kvalitet end de andre de hold. Det kunne tyde på, at det næstvigtigste er at sammensætte/samarbejde et godt hold.

f) Estimation af varianskomponenter:
Da der ikke fandtes vekselvirkning mellem metoder og hold i grupper (spørgsmål b), så må vi estimere denne varianskomponent til nul: $\sigma_{AC}^2 \approx 0$.

Af $E(s^2)$ - varianskomponentsammensætningen fra spørgsmål a fås

$$E(s_C^2) = 4\sigma_C^2 + \sigma^2$$

Heri indsættes s^2-værdierne fra variansanalysetabellen, så vi får: $6.54306 \approx 4\sigma_C^2 + 2.05765$,
hvoraf $\sigma_C^2 \approx 1.121$

Øvelse 3.2 (split-plot, fixed, random, sidedelt variansanalyse) Regn opgave 22 side 244.

Tolkning af "Model koefficienterne".
Ønskes et estimat for den bedste kvalitet, vælges (den gule ikon "Tabular Options" | Model Koefficients |OK).
Det giver følgende udskrift:

95,0% confidence intervals for coefficient estimates (respons)

Parameter	Estimate	Standard Error	Lower Limit	Upper Limit	V.I.F.
CONSTANT	19,3333	0,239075	18,8419	19,8248	
metoder	4,25556	0,239075	3,76413	4,74698	1,0
grupper	0,791667	0,338103	0,0966839	1,48665	1,33333
grupper	0,05	0,338103	-0,644983	0,744983	1,33333
hold(grupper)	-1,45	0,585612	-2,65375	-0,246254	1,33333
hold(grupper)	2,425	0,585612	1,22125	3,62875	1,33333
hold(grupper)	-0,458333	0,585612	-1,66208	0,745412	1,33333
hold(grupper)	0,441667	0,585612	-0,762079	1,64541	1,33333
hold(grupper)	0,558333	0,585612	-0,645412	1,76208	1,33333
hold(grupper)	-0,341667	0,585612	-1,54541	0,862079	1,33333

For at kunne tolke en sådan udskrift, må man kende til Statgraphics indførelse af såkaldte "dummy variable" (også

kaldet "indikator variable".
Til karakterisering af de 2 metoder indføres en variabel v_A som har værdien +1 for metode 1 og -1 for metode 2.
Til karakterisering af de 3 grupper indføres 2 variable v_{B1} og v_{B2}, og til karakterisering af de 3 hold indenfor hver gruppe indføres 2 · 3 variable v_{C1}, v_{C2}, v_{C4},v_{C5}, v_{C7}, v_{C8}. Værdierne fremgår af det følgende skema.

gruppe	hold	v_{B1}	v_{B2}	v_{C1}	v_{C2}	v_{C4}	v_{C5}	v_{C7}	v_{C8}
B1	C1	1	0	1	0	0	0	0	0
B1	C2	1	0	0	1	0	0	0	0
B1	C3	1	0	-1	-1	0	0	0	0
B2	C4	0	1	0	0	1	0	0	0
B2	C5	0	1	0	0	0	1	0	0
B2	C6	0	1	0	0	-1	-1	0	0
B3	C7	-1	-1	0	0	0	0	1	0
B3	C8	-1	-1	0	0	0	0	0	1
B3	C9	-1	-1	0	0	0	0	-1	-1

Værdierne gives ved, at for en bestemt variabel, så skrives der 1 i den rubrik, som svarer til variablens niveau, der skrives -1 ud for det sidste niveau, og 0 i resten.
Den givne model bliver med "dummy variable" skrevet

$$Y = k_0 + k_1 \cdot v_a + k_2 \cdot v_{B1} + k_3 \cdot v_{B2} + k_4 \cdot v_a \cdot v_{B1} + k_5 \cdot v_a \cdot v_{B2} + k_6 \cdot v_{C1} + k_7 \cdot v_{C2} + k_8 \cdot v_{C4} + k_9 \cdot v_{C5} + k_{10} \cdot v_{C7} + k_{11} \cdot v_{C8} + k_{12} \cdot v_a \cdot v_{C1} + k_{13} \cdot v_a \cdot v_{C2} + k_{14} \cdot v_a \cdot v_{C4} + k_{15} \cdot v_a \cdot v_{C5} + k_{16} \cdot v_a \cdot v_{C7} + k_{17} \cdot v_a \cdot v_{C8}.$$

Da den fundne model ikke har vekselvirkninger A*B og A*C(B), er
$k_4 = k_5 = k_{12} = k_{13} = k_{14} = k_{15} = k_{16} = k_{17} = 0$
Modellen er derfor reduceret til

$$Y = k_0 + k_1 \cdot v_a + k_2 \cdot v_{B1} + k_3 \cdot v_{B2} + k_6 \cdot v_{C1} + k_7 \cdot v_{C2} + k_8 \cdot v_{C4} + k_9 \cdot v_{C5} + k_{10} \cdot v_{C7} + k_{11} \cdot v_{C8}.$$

Af udskriften fås konstanterne

```
Parameter                    Estimate
--------------------------------------
CONSTANT        = k0 =      19,3333
metoder         = k1 =       4,25556
grupper         = k2 =       0,791667
grupper         = k3 =       0,05
hold(grupper)   = k6 =      -1,45
hold(grupper)   = k7 =       2,425
hold(grupper)   = k8 =      -0,458333
hold(grupper)   = k9 =       0,441667
hold(grupper)   = k10 =       0,558333
hold(grupper)   = k11 =      -0,341667
--------------------------------------
```

Ønskes et estimat for den bedste kvalitet, så skal der vælges metode 1 og i gruppe B1 at vælge hold 2.
Af skemaet ses, at det sker ved i ovenstående model, at sætte $v_a = 1$, $v_{B1}=1$, $v_{C2}=1$ og resten til 0.
Vi får da: $Y_{max} = 19.333 + 4.255 + 0.79167 + 2.425 = 26.80$

3.5.3. SPLIT-PLOT FORSØG MED BLOKFAKTOR.

Et split-plot forsøg hvori der også indgår en blokfaktor er i praksis også ofte forekommende. I dette afsnit illustreres det ved to eksempler. I eksempel 3.9 indføres blokke i et til eksempel 3.6 svarende landbrugsforsøg, mens eksempel 3.10 er et typisk forsøg med at afprøve medicin på forsøgspersoner (her mus).

Eksempel 3.9 (Split-plot med blokke) (udarbejdet af Bjarne Hellesen)

Man ønsker at undersøge tre gødninger G1, G2, G3 og to pløjemetoder P1, P2. Fire bondegårde (blokke) B1, B2, B3, B4 udvælges tilfældigt. På hver gård vælges for hver pløjemetode en tilfældig strimmel mark, dvs. i alt 8 markstrimler U1, U2, ..., U8 (hovedplot-forsøgsenheder = whole plot experimental Units). Hver markstrimmel, opsplittes i 3 dele (split-plots), som kombineres tilfældigt med de tre gødninger. Altså 24 split-plots:

Pløjemetoder	Gødninger			Mark	Gård
P2	G2	G3	G1	U1	B1
P1	G1	G3	G2	U2	
P1	G2	G1	G3	U3	B2
P2	G3	G1	G2	U4	
P1	G2	G3	G1	U5	B3
P2	G2	G1	G3	U6	
P2	G3	G2	G1	U7	B4
P1	G1	G2	G3	U8	

Ovenstående er et split-plot design: De 6 behandlinger P1G1, P1G2, P1G3, P2G1, P2G2, P2G3 er ikke kombineret tilfældigt med de 6 split-plots i en blok (=gård), sådan som det ville være tilfældet i et randomiseret blokforsøg:

P1 G1	P2 G2	P1 G2	U1	B1
P2 G3	P2 G1	P1 G3	U2	
P2 G1	P1 G3	P1 G1	U3	B2
P2 G2	P1 G2	P2 G3	U4	
P2 G1	P2 G2	P1 G2	U5	B3
P1 G3	P2 G3	P1 G1	U6	
P2 G2	P1 G1	P2 G3	U7	B4
P1 G3	P1 G2	P2 G1	U8	

Fordele ved split-plot: 1) Praktisk at pløje hele markstrimler uden at skifte pløjemetode.
2) Mindre støj på testning af G og P*G.
Ulemper ved split-plot: Nogle frihedsgrader mistes, især ved test af hovedfaktoren P .

Det indses, at den mest omfattende lineære model for udbyttet Y er (led i standardrækkefølge!):

$$Y = \mu + B + P + B*P + G + B*G + P*G + B*P*G + U(B,P) + G*U(B,P) + Eps(B,P,G,U)$$

Udbytterne blev:

Gård	Metode	Mark	Gødning G1	Gødning G2	Gødning G3
B1	P1	U2	46	48	49
	P2	U1	35	39	37
B2	P1	U3	63	67	69
	P2	U4	61	65	60
B3	P1	U5	56	62	60
	P2	U6	45	47	48
B4	P1	U8	46	53	47
	P2	U7	28	29	28

a) Find varianskomponentsammensætningen, og de tilhørende frihedsgrader.
b) Undersøg, om følgende virkninger er 0:
*P*G, G, P, B*G såfremt B*P*G=0, B såfremt B*P=0.*
*c) Antag, at B*P*G=0, og benyt resultatet fra b) til at forenkle af modellen yderligere.*
Synes den forenklede model god?
d) Sammenlign gødninger. Sammenlign pløjemetoder. Sammenlign bondegårde.
e) Er forsøget bedst egnet til at sammenligne gødninger, pløjemetoder eller bondegårde?

LØSNING:

a) Benyttes EMS-program fås

```
The input:
B Ran 4  P Fix 2  G Fix 3  U Ran 1 Eps Ran1  mu + B + P + B*P + G +B*G +P*G
+ B*P*G +U(B,P) + G*U(B,P)
gives the output:
   EMS(B) = 6*V(B) + 3*V(U) + V(Eps)
   EMS(P) = 12*f(P) + 3*V(B*P) + 3*V(U) + V(Eps)
   EMS(B*P) = 3*V(B*P) + 3*V(U) + V(Eps)
   EMS(G) = 8*f(G) + 2*V(B*G) + V(G*U) + V(Eps)
   EMS(B*G) = 2*V(B*G) + V(G*U) + V(Eps)
   EMS(P*G) = 4*f(P*G) + V(B*P*G) + V(G*U) + V(Eps)
   EMS(B*P*G) = V(B*P*G) + V(G*U) + V(Eps)
   EMS(U) = 3*V(U) + V(Eps)
   EMS(G*U) = V(G*U) + V(Eps)
   EMS(Eps) = V(Eps)
```

Nu kan man tilføje de tilhørende frihedsgrader f og eventuelt skrive alt med sædvanlig symbolik:

$$
\begin{array}{llllllll}
E(s_B^2) & = & 6\cdot\sigma_B^2 & & + \; 3\cdot\sigma_U^2 & & + \; \sigma^2, & 3 \\
E(s_P^2) & = & 12\cdot\sigma_P^2 & + \; 3\cdot\sigma_{BP}^2 & + \; 3\cdot\sigma_U^2 & & + \; \sigma^2, & 1 \\
E(s_{BP}^2) & = & & 3\cdot\sigma_{BP}^2 & + \; 3\cdot\sigma_U^2 & & + \; \sigma^2, & 3\cdot 1=3 \\
\hline
E(s_G^2) & = & 8\cdot\sigma_G^2 & + \; 2\cdot\sigma_{BG}^2 & & + \; \sigma_{GU}^2 & + \; \sigma^2, & 3\text{-}1=2 \\
E(s_{BG}^2) & = & & 2\cdot\sigma_{BG}^2 & & + \; \sigma_{GU}^2 & + \; \sigma^2, & 3\cdot 2=6 \\
E(s_{PG}^2) & = & 4\cdot\sigma_{PG}^2 & + \; \sigma_{BPG}^2 & & + \; \sigma_{GU}^2 & + \; \sigma^2, & 1\cdot 2=2 \\
E(s_{BPG}^2) & = & & \sigma_{BPG}^2 & & + \; \sigma_{GU}^2 & + \; \sigma^2, & 3\cdot 1\cdot 2=6 \\
\hline
E(s_U^2) & = & & & 3\cdot\sigma_U^2 & & + \; \sigma^2, & 4\cdot 2\cdot(1\text{-}1)=0 \\
E(s_{GU}^2) & = & & & & \sigma_{GU}^2 & + \; \sigma^2, & 2\cdot 0=0 \\
E(s_0^2) & = & & & & & \sigma^2, & 0\,. \\
\hline
\end{array}
$$

b) Data indsættes i Statgraphics (U-søjlen kunne udelades, da U har 0 frihedsgrader):

B	P	G	U	Y
1	1	1	2	46
1	1	2	2	48
1	1	3	2	49
1	2	1	1	35
1	2	2	1	39
1	2	3	1	37
2	1	1	3	63
2	1	2	3	67
2	1	3	3	69
2	2	1	4	61
2	2	2	4	65
2	2	3	4	60
3	1	1	5	56
3	1	2	5	62
3	1	3	5	60
3	2	1	6	45
3	2	2	6	47
3	2	3	6	48
4	1	1	8	46
4	1	2	8	53
4	1	3	8	47
4	2	1	7	28
4	2	2	7	29
4	2	3	7	28

Vælg GLM og sæt

Dependent variable: Y

Categorical factors: B, P, G (som Statgraphics kalder A, B og C). Bemærk, at markstrimler U ikke medtages, da U har 0 frihedsgrader.

GLM specifications: A, B, A*B, C, A*C, B*C. Bemærk, at U(B P) og G*U(B P) ikke medtages, da de har 0 frihedsgrader. Bemærk, at A*B*C ikke skrives, da residualen så ville blive 0, idet der ingen gentagelser er til bestemmelse af forsøgsfejlens varians.

Så fås:

```
General Linear Models
---------------------
Analysis of Variance for Y
-----------------------------------------------------------------------------
Source               Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-----------------------------------------------------------------------------
Model                       3529,67     17      207,627      47,31       0,0001
Residual                    26,3333      6      4,38889
-----------------------------------------------------------------------------
Total (Corr.)                3556,0     23
```

Type III Sums of Squares

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
B	2398,33	3	799,444	182,15	0,0000
P	864,0	1	864,0	196,86	0,0000
B*P	195,667	3	65,2222	14,86	0,0035
G	57,0	2	28,5	6,49	0,0316
B*G	7,66667	6	1,27778	0,29	0,9206
P*G	7,0	2	3,5	0,80	0,4930
Residual	26,3333	6	4,38889		
Total (corrected)	3556,0	23			

All F-ratios are based on the residual mean square error.

R-Squared = 99,2595 percent, R-Squared (adjusted for d.f.) = 97,1613 percent

Af varianskomponentsammensætningen ses, at P*G kan testes mod residualen (=B*P*G).
P*G: P_value = 0.4930 > 0.05 ===> PG=0 , ingen vekselvirkning mellem P og G.
Vi kan nu foretage en pooling af PG . Endvidere fremgår det af varianskomponentsammensætningen, at

G kan testes mod B*G,
P kan testes mod B*P,
B*G kan testes mod B*P*G, *såfremt* B*P*G antages at være uden betydning,
B kan testes mod B*P , *såfremt* B*P antages at være uden betydning.

Dette giver følgende variansanalysetabel:

Type III Sums of Squares

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
B	2398,33	3	799,444	12,26(1)	0,0344
P	864,0	1	864,0	13,25(1)	0,0358
B*P	195,667	3	65,2222	15,65(0)	0,0010
G	57,0	2	28,5	22,30(2)	0,0017
B*G	7,66667	6	1,27778	0,31(0)	0,9164
Residual	33,3333	8	4,16667		
Total (corrected)	3556,0	23			

F-ratios are based on the following mean squares:
(0) Residual
(1) A*B
(2) A*C

G: P_value = 0.0017 < 0.01 ===> G $\neq$ 0 (2-stjernet signifikans)
P: P_value = 0.0358 < 0.05 ===> P $\neq$ 0 (1-stjernet signifikans)
B*G: P_value = 0.9164 > 0.05 ===> B*G=0, *såfremt* B*P*G antages at være uden betydning
B: P_value = 0.0344 < 0.05===> B $\neq$ 0 (1-stjernet signifikans). Skulle B*P $\neq$ 0, kan det ikke svække den opnåede konklusion "B $\neq$ 0", men det ville svække en eventuel konklusion "B=0", fordi nævneren i F-testet i middel er for stor, idet

$$\frac{\mathrm{EMS}(B)}{\mathrm{EMS}(B*P)} = \frac{6\cdot \mathrm{V}(B) + 3\cdot \mathrm{V}(U) + \mathrm{V}(EPS)}{3\cdot \mathrm{V}(B\cdot P) + 3\cdot \mathrm{V}(U) + \mathrm{V}(EPS)}.$$

Konklusion: Gødningerne giver uens middeludbytte.
Pløjemetoderne giver uens middeludbytte.
Der synes ikke at være nogen vekselvirkning mellem bondegårde og gødninger.
Det var godt at indføre bondegårde som blokke.

c) Yderligere forenkling af modellen

Antages, at 3-faktorvekselvirkningen B*P*G=0, har vi vist, at BG=0. I variansanalysetabellen vil Residual og BG da begge repræsentere split-plot-støjen V(G*U) + V(Eps), således at BG kan pooles ned i Residualen: (Input dialog | OK | slet A*C | OK).

Idet vi så skal teste G mod Residualen, fås

Type III Sums of Squares

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
B	2398,33	3	799,444	12,26(1)	0,0344
P	864,0	1	864,0	13,25(1)	0,0358
B*P	195,667	3	65,2222	22,27(0)	0,0000
G	57,0	2	28,5	9,73(0)	0,0022
Residual	41,0	14	2,92857		
Total (corrected)	3556,0	23			

```
F-ratios are based on the following mean squares:
   (0) Residual
   (1) A*B
R-Squared = 98,847 percent        R-Squared (adjusted for d.f.) = 98,1058 percent
```

Vi ser, at de tidligere konklusioner ikke ændres.

R-squared = 98,85 % > 85 % viser, at den forenklede model synes god.

d) Vælges " Table of Means" fås:

Table of Least Squares Means for Y with 95,0 Percent Confidence Intervals

Level		Count	Mean	Stnd. Error	Lower Limit	Upper Limit
GRAND MEAN		24	49,5	0,349319	48,7508	50,2492
B						
1		6	**42,3333**	3,29702	**31,8407**	**52,8259**
2		6	**64,1667**	3,29702	**53,6741**	**74,6593**
3		6	53,0	3,29702	42,5074	63,4926
4		6	**38,5**	3,29702	**28,0074**	**48,9926**
P						
1		12	**55,5**	2,33135	**48,0806**	**62,9194**
2		12	**43,5**	2,33135	**36,0806**	**50,9194**
B by P						
1	1	3	47,6667	0,988024	45,5476	49,7858
1	2	3	37,0	0,988024	34,8809	39,1191
2	1	3	66,3333	0,988024	64,2142	68,4524
2	2	3	62,0	0,988024	59,8809	64,1191
3	1	3	59,3333	0,988024	57,2142	61,4524
3	2	3	46,6667	0,988024	44,5476	48,7858
4	1	3	48,6667	0,988024	46,5476	50,7858
4	2	3	28,3333	0,988024	26,2142	30,4524
G						
1		8	**47,5**	0,605038	**46,2023**	**48,7977**
2		8	**51,25**	0,605038	**49,9523**	**52,5477**
3		8	49,75	0,605038	48,4523	51,0477

Heraf ses, at Gødning G2 giver større udbytte end G1 (intet overlap)

Pløjemetode P1 giver større udbytte end P2 (næsten intet overlap)

Bondegård B2 giver større udbytte end B1 og B4 (intet overlap)

e) Pløjemetoderne har en udbytteforskel på 12, bondegårdene på cirka 22, mens den for gødningerne kun er cirka 4. Alligevel fandt vi en meget lavere P_value for gødningerne, idet split-plot-fejl sædvanligvis er mindre end hovedplotfejl. Dvs.

Forsøgsplanen er særlig velegnet til at sammenligne gødninger.

Desuden er det praktisk at pløje lange stykker ad gangen.

Det tilsvarende randomiserede blokforsøg ville være dårligere til at sammenligne gødninger, idet støjen så også ville komme fra hovedplots. Og pløjningen ville blive upraktisk.

Da modeldannelsen ved denne type split-plot forsøg erfaringsmæssigt er vanskelig gennemgås endnu et eksempel herpå.

Eksempel 3.7. (split-plot med blok) (hentet fra SFII eksempel 75 side 243)
Ved en undersøgelse af et farmekas evne til at sænke blodkoncentrationen hos mus skulle det afgøres, om forbehandling af dyrene med 4 forskellige fodertyper igennem en måned forud havde betydning. Endvidere var man interesseret i at undersøge om en eventuel fodereffekt influerede på dosisafhængigheden. Man anvendte derfor i forsøget 2 forskellige dosiskoncentrationer af farmekaet. Da man har erfaring for at mus fra samme kuld reagerer mere ensartet end hvis man tilfældigt udtager mus til forsøget, anvendes mus af samme kuld som blokfaktor.

1) *Beskriv et split-plot forsøg, hvor der også på en skitse vises den resulterende forsøgsplan.*
2) *Opstil en model for forsøget, og angiv varianskomponentsammensætningen og de tilhørende frihedsgrader*

I opgave 26 fortsætter eksemplet, idet forsøgsresultaterne er angivet, og der ønskes en statistisk analyse af forsøget.

LØSNNG:
1) Forsøgsplan.
5 kuld på hver 4 mus udtages (tilfældigt af en større samling mus). De 4 mus i et kuld bliver (igen randomiseret) fodret med hver sin foderblanding i en måned. Hver af de 20 mus fik nu indsprøjtet en af de to doseringer af farmekaet. Hvilken det blev bestemtes ved lodtrækning. Efter en vis tid måltes den procentvise sænkning af blodsukkerkoncentrationen. Efter en passende pause (således at der ikke var tale om eftervirkning af første dosis) fik hver mus derpå den anden dosering af farmekaet og efter den samme tid som før måltes den procentvise sænkning af blodsukkerkoncentrationen.

Dette kan skitsemæssigt skrives sådan:

Foderblandinger	Dosiskoncentration		Mus	Kuld
2	D2	D1	1	1
1	D1	D2	2	
4	D1	D2	3	
3	D2	D1	4	
3	D1	D2	5	2
1	D1	D2	6	
4	D2	D1	7	
2	D1	D2	8	
2	D2	D1	9	3
1	D2	D1	10	
4	D2	D1	11	
3	D1	D2	12	
1	D1	D2	13	4
2	D2	D1	14	
4	D2	D1	15	
3	D1	D2	16	
4	D1	D2	17	5
2	D1	D2	18	
3	D2	D1	19	
1	D2	D1	20	

2) Opstilling af model.

Idet M = mus er den fuldstændige model for forsøget (med Statgraphics betegnelser)

$Y = \mu + K + F + K*F + D + K*D + F*D + K*F*D + M(K\ F) + D*M(K\ F) + \varepsilon$

eller $x_{ijklm} = \mu + K_i + F_j + KF_{ij} + D_k + KD_{ik} + FD_{jk} + KFD_{ijk} + M_{(ij)l} + DM_{(ij)kl} + \varepsilon_{(ijkl)m}$

Opstilling af varianskomponenttabel:

Benyttes EMSwin programmet fås:

```
The input:
K Ran 5  F Fix 4  D Fix 2  M Ran 1  Eps Ran 1  mu + K + F  +K*F + D + K*D +F*D +K*F*D
+M(K,F)+D*M(K,F)
gives the output:
   EMS(K) = 8*V(K) + 2*V(M) + V(Eps)
   EMS(F) = 10*f(F) + 2*V(K*F) + 2*V(M) + V(Eps)
   EMS(K*F) = 2*V(K*F) + 2*V(M) + V(Eps)
   EMS(D) = 20*f(D) + 4*V(K*D) + V(D*M) + V(Eps)
   EMS(K*D) = 4*V(K*D) + V(D*M) + V(Eps)
   EMS(F*D) = 5*f(F*D) + V(K*F*D) + V(D*M) + V(Eps)
   EMS(K*F*D) = V(K*F*D) + V(D*M) + V(Eps)
   EMS(M) = 2*V(M) + V(Eps)
   EMS(D*M) = V(D*M) + V(Eps)
   EMS(Eps) = V(Eps)
```

En tabel over frihedsgradstallet er følgende:

	f
K	$f_K = 5 - 1 = 4$
F	$f_F = 4 - 1 = 3$
KF	$f_{KF} = f_K \cdot f_F = 12$
D	$f_D = 2 - 1 = 1$
K*D	$f_{KD} = f_K \cdot f_D = 4$
F*D	$f_{FD} = f_F \cdot f_D = 3$
K*F*D	$f_{KFD} = f_K \cdot f_F \cdot f_D = 12$
M(K,F)	$f_M = 5 \cdot 4 \cdot (1 - 1) = 0$
D*M(K;F)	0
ε	0
Total	38

Af varianskomponentskemaet ses, at s^2_{KDF} kan benyttes til testning af vekselvirkningen FD mellem fodertype og dosis. Endvidere at s^2_{KD} kan benyttes til testning af dosis.
Hvis man tror, at blokfaktoren kuld ikke vekselvirker med underfaktoren dosis, kan man benytte en pooling af de to (forsøgsfejl 2) til testning af de to faktorer. Endelig ses, at s^2_{KF} (forsøgsfejl 1) kan benyttes til testning af fodertype .
Da modelkomponenterne M og DM vedrørende mus har frihedsgradstallet 0 kan de ikke indtastes i Statgraphics, men de benyttes ved testningen.

◆

Øvelse 3.3 (split-plot) Regn opgave 26 side 246.

3.6 SATTERTHWAITES METODE (skrevet af Bjarne Hellesen)

Uddrag fra en variansanalyse:

Variation	f	s^2	$E(s^2)$
. . .	. . .	. . .	. . .
. . .	. . .	. . .	. . .
C	2	400.0	$E(s_C^2) = 12\sigma_C^2 + 6\sigma_{AC}^2 + 4\sigma_{BC}^2 + \sigma^2$
AC	4	20.0	$E(s_{AC}^2) = 3\sigma_{AC}^2 + \sigma^2$
BC	10	10.0	$E(s_{BC}^2) = 4\sigma_{BC}^2 + \sigma^2$
Residual	25	5.0	$E(s_0^2) = +\sigma^2$

For at kunne teste hovedvirkning af C mangler vi en s^2-størrelse med varianssammensætningen $6\sigma_{AC}^2 + 4\sigma_{BC}^2 + \sigma^2$. Så vi **konstruerer en brugbar linearkombination**:

$$s_{lin}^2 = 2s_{AC}^2 + s_{BC}^2 - 2s_0^2 = 2 \cdot 20 + 10 - 2 \cdot 5 = 40$$

(brugbar idet $E(s_{lin}^2 = 2 \cdot (3\sigma_{AC}^2 + \sigma^2) + (4\sigma_{BC}^2 + \sigma^2) - 2\sigma^2 = 6\sigma_{AC}^2 + 4\sigma_{BC}^2 + \sigma^2)$.

Nu kan vi danne F-ratio $F_C = \frac{s_C^2}{s_{lin}^2} = \frac{400}{40} = 10.0$

og sammenligne med $F_{0.95}(f_T, f_N)$ Her er $f_T = f_C = 2$ mens vi **beregner** $f_N = f_{lin}$ **tilnærmet af formlen** $f_N = f_{lin} = \dfrac{\left(s_{lin}^2\right)^2}{\dfrac{\left(2s_{AC}^2\right)^2}{f_{AC}} + \dfrac{\left(s_{BC}^2\right)^2}{f_{BC}} + \dfrac{\left(-2s_0^2\right)^2}{f_0}} = \dfrac{40^2}{\dfrac{40^2}{4} + \dfrac{10^2}{10} + \dfrac{10^2}{25}} = 3.86$

Da 3.86 ligger mellem de hele tal 3 og 4, går vi ind i F - tabellen med disse hele tal, læser $F_{0.95}(2,3) = 9.55$, $F_{0.95}(2,4) = 6.94$ og finder ved lineær interpolation

$$F_{0.95}(2{,}3.86) \approx 9.55 + \frac{3.86 - 3}{4 - 3} \cdot (6.94 - 9.55) = 7.31 < 10.00$$

Analogt læses $F_{0.99}(2,3) = 30.8$, $F_{0.99}(2,4) = 18.0$ og ved lineær interpolation fås

$$F_{0.99}(2{,}3.86) \approx 30.8 + \frac{3.86 - 3}{4 - 3} \cdot (18.0 - 30.8) = 19.8 > 10.00$$

Vores F-test giver altså 1-stjernet signifikans , dvs. $C \neq 0$.

4 REGRESSIONSANALYSE

4.1 INDLEDNING

En regressionsanalyse har til formål at teste og estimere virkningen af én eller flere **kvantitative faktorer** , dvs. faktorer hvis niveauer er karakteriseret ved en målelig egenskab. Eksempler på kvantitative faktorer er temperatur T, tryk p og mængden af katalysatormasse ved en kemisk proces.
Vi vil i dette kapitel beskæftige os med **lineær** regressionsanalyse, hvorved forstås regressionsanalyse på en model, som er lineær i modellens parametre.

De følgende 4 ligninger er eksempler på lineære modeller, som vi i det følgende vil analysere ved hjælp af Statgraphics (x, x_1 og x_2 er de kvantitative variable).
$Y = \alpha_0 + \beta_1 x_1 + \beta_2 x_2$, hvor parametrene er α_0, β_1 og β_2 og
$Y = \alpha_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2$, hvor parametrene er $\alpha_0, \beta_1, \beta_2, \beta_3, \beta_4$ og β_5,

Mere komplicerede lineære modeller, som eksempelvis
$Y = \alpha_0 + \beta_1 x_1^2 + \beta_2 \tan(x_2) + \beta_3 x_1 x_2$, hvor parametrene er $\alpha_0, \beta_1, \beta_2$ og β_3,
er så kompliceret, at en analyse ved hjælp af Statgraphics først kan ske, efter at man har omskrevet ligningen til $Y = \alpha_0 + \beta_1 z_1 + \beta_2 z_2 + \beta_3 z_3$, hvor $z_1 = x_1^2, z_2 = \tan(x_2)$ og $z_3 = x_1 x_2$.

Modeller som ikke umiddelbart er lineære, men som en transformation kan gøres lineære, er eksempelvis
eksponentialfunktionen $Y = \alpha \cdot \mathrm{e}^{\beta \cdot x}$, som kan omformes til $\ln(Y) = \ln(\alpha) + \beta \cdot x$ som er lineær i parametrene $\alpha_0 = \ln(\alpha)$ og β, og
potensfunktionen $Y = \alpha \cdot \mathrm{x}^{\beta}$, som kan omformes til $\ln(Y) = \ln(\alpha) + \beta \cdot \ln(x)$, som er lineær i parametrene $\alpha_0 = \ln(\alpha)$ og β.

Et eksempel på en ikke lineær model, som heller ikke ved transformation kan gøres lineær, er
$Y = \alpha_0 x + \beta_1 \cdot \mathrm{e}^{\beta_2 \cdot x}$.

De grundlæggende begreber vedrørende "enkelt regressionsanalyse" forudsættes kendt[1]. En kort repetition af de væsentligste begreber kan dog findes i afsnit 4.2.1.
De grundlæggende begreber fra "enkelt regressionsanalyse" er de samme for de større modeller med mange uafhængige variable, men formlerne er naturligvis mere komplicerede, og kræver kendskab til matrixregning. De væsentligste formler er angivet i appendix 4.1, hvor der også er angivet, hvorledes man kan foretage beregningerne ved anvendelse af et program som Maple, der kan håndtere store matricer.

[1]Se evt. M. Oddershede Larsen: Videregående Statistik

4.2 REGRESSIONSANALYSE MED 1 FAKTOR

Indledning
Vi vil i dette kapitel

1) repetere de vigtigste begreber for modellen $Y = \alpha_0 + \beta_1 x$, og se på modeller, som ved en transformation kan overføres til en sådan lineær model.
3) betragte polynomiale modeller af formen $Y = \alpha_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_n x^n$.

4.2.1 Enkelt Regressionsanalyse

4.2.1.1 Regressionsanalyse med gentagelser.

Der gives i dette afsnit ved hjælp af følgende eksempel en kort repetition af de vigtigste begreber og af hvorledes man foretager en enkelt regressionsanalyse ved benyttelse af Statgraphics.

Eksempel 4.1 (enkelt regression) (F85)
Metalpladers overflader oxideres i en ovn ved 200^0 C. Med henblik på en undersøgelse af sammenhængen mellem det oxiderede lags tykkelse y (i ångstrøm) og tiden t (i minutter) foretog man følgende uafhængige målinger:

Tiden t	20	30	40	60	70	90	100	120	150	180
Tykkelse y	4.2 4.9	7.4 6.9	8.8 8.2	13.6 12.0	13.1 12.4	14.9 16.8	20.0 21.2	23.1 25.2	27.5 25.1	32.9 32.4

Det formodes på forhånd, at middelværdien af Y indenfor måleområdet [20 ; 180] *kan beskrives som et førstegradspolynomium i x*: $E(Y|x) = \alpha_0 + \beta_1 x$.

1) *Test denne formodning, og undersøg også rimeligheden heraf ved at se på et residualplot. Vurder også om der er eventuelle "outliers"*
2) *Angiv forudsætningerne for en regressionsanalyse.*
   Undersøg dernæst:
   a) ved et normalfordelingsplot om kravet om normalitet er opfyldt.
   b) dels ved et residualplot, dels ved Bartletts test om kravet om varianshomogenitet er opfyldt.

I bekræftende fald ønskes:

3) *test af om Y er uafhængig af x , dvs. af H_0*: Hældningskoefficienten $\beta_1 = 0$.
4) *angivet regressionsligningen* $Y = a + b \cdot x$ *, hvor a og b er estimater for* α_0 *og* β_1.
5) *angivet et* 95% *konfidensinterval for* β_1
6) *angivet et* 95% *konfidensinterval for middelværdien af tykkelsen y, når t* = 100 *minutter.*

LØSNING:

1) <u>Data indtastes</u> i Statgraphics på sædvanlig måde (se evtuelt appendix A):

```
x    y
20   4,2
20   4,9
30   7,4
30   6,9
40   8,8
40   8,2
...  ...
```

Regressionsanalysetabel opstilles

Vælg (Relate | Simple Regression | indsæt *t* og *y* | OK).

Der fremkommer en tegning og følgende variansanalysetabel:

```
Regression Analysis - Linear model: Y = a + b*X
-----------------------------------------------------------------------------
Dependent variable: y
Independent variable: t
-----------------------------------------------------------------------------
                               Standard          T
Parameter       Estimate         Error       Statistic        P-Value
-----------------------------------------------------------------------------
Intercept        1,65415       0,599582        2,75883         0,0129
Slope           0,172975     0,00603212        28,6756         0,0000
-----------------------------------------------------------------------------
                           Analysis of Variance
-----------------------------------------------------------------------------
Source             Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-----------------------------------------------------------------------------
Model                     1486,44      1      1486,44     822,29       0,0000
Residual                  32,5383     18      1,80768
-----------------------------------------------------------------------------
Total (Corr.)             1518,98     19

Correlation Coefficient = 0,989231      R-squared = 97,8579 percent
Standard Error of Est. = 1,3445
Correlation Coefficient = 0,99915    R-squared = 99,83 percent
```

Da vi har gentagelser ignoreres den fremkomne udskrift i første omgang.

Vælg (gul ikon = Tabular Options | Lack of Fit test | OK). Denne giver følgende tabel:

```
Analysis of Variance with Lack-of-Fit
-----------------------------------------------------------------------------
Source             Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-----------------------------------------------------------------------------
Model                     1486,44      1      1486,44     822,29       0,0000
Residual                  32,5383     18      1,80768
-----------------------------------------------------------------------------
   Lack-of-Fit            22,7283      8      2,84104       2,90       0,0591
   Pure Error                9,81     10        0,981
-----------------------------------------------------------------------------
Total (Corr.)             1518,98     19
```

Lad nulhypotesen være H_0: "Førstegradsmodellen gælder".

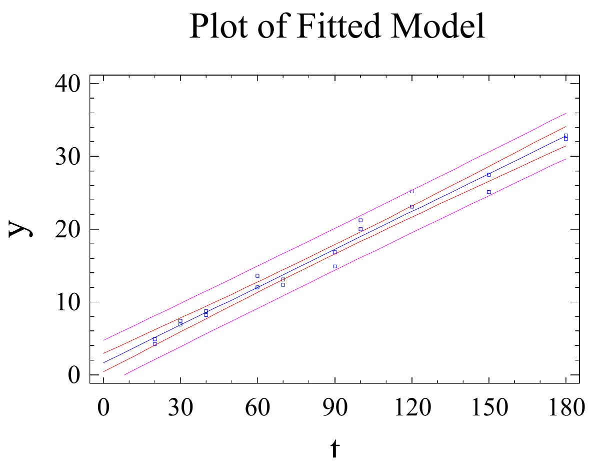
Af udskriften ses, at $F_{lack} = \frac{s_{lack}^2}{s_0^2} = 2.90$, med en P - value på 0.0591. På et signifikansniveau på 5%, ses, at <u>H_0 må accepteres, dvs. vi kan antage, at indenfor måleområdet giver førstegradsmodellen en rimelig god beskrivelse af resultaterne,</u>

Tegningen viser det samme.

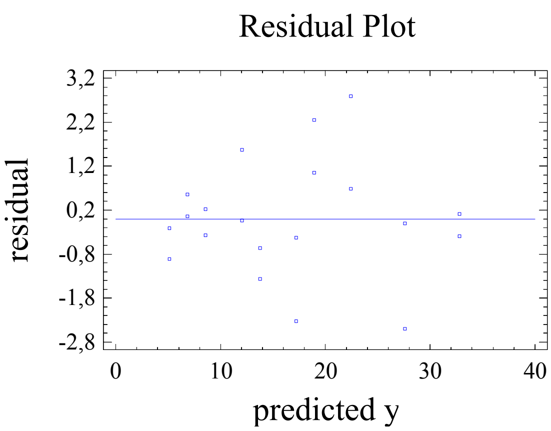
Her er der yderligere tegnet dels 95% konfidensintervallinier (det inderste sæt linier), dels de såkaldte 95% prædistinationsintervallinier.

95% konfidensintervallerne angiver indenfor hvilke intervaller "middelværdierne med 95% sikkerhed vil falde, mens prædistinationsintervallerne ("forudsigelsesintervallerne") forudsiger indenfor hvilke intervaller værdien af en kommende observation med 95% sikkerhed vil falde.

Regressionslinien går altid gennem punktet $(\bar{x}, \bar{y})$, og de nævnte intervaller er smallest her på midten og bliver bredere ud mod randen af måleområdet.



For yderligere at vurdere om den lineære model er rimelig, betragter vi et residualplot.
Vælg (blå ikon= Graphical options | Residual versus predicted | OK)
Tegningen viser de såkaldte "studentized residuals" .
De sædvanlige residualer fås af: (Cursor på tegning | højre musetast |Residuals).
Dette giver tegningen nedenfor.



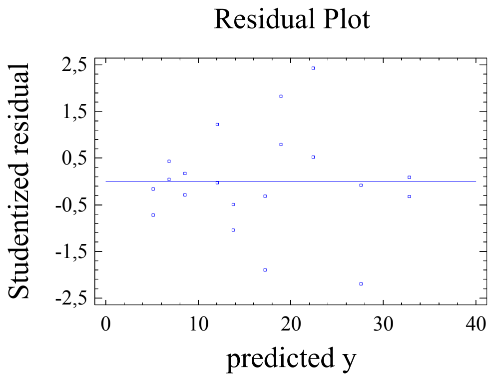
Idet det er middelværdierne der skal ligge på den rette linie, ses at netop <u>gennemsnitspunkterne fordeler sig tilfældigt omkring linien.</u>

En "outliers" er punkter, som ligger så langt fra hvad man ville forvente ud fra modellen, at det kræver en nærmere undersøgelse (det kunne være en fejlmåling),
Man har forskellige måder at undersøge det på.

a) Man kan se om der er punkter der ligger uden for predictionsgrænserne.
Kun et enkelt punkt ligger lige på grænsen, så det er ikke foruroligende.

b) Man kan også (lidt primitivt) undersøge om der er residualer, der er større end $3 \cdot \sigma$, da man ved, at 99.5 % af punkterne bør falder indenfor.
Vi har, at $3 \cdot \sigma = 3 \cdot \sqrt{0.981} = 2.97$, og af residualplottet ses igen, at kun et punkt synes at ligge lige på grænsen.

c) Idet konfidensintervallerne jo ikke er lige brede, så kan man normere residualerne (se eventuelt i appendix 4.1 herom). Dette er den såkaldte "studentized residuals" , som Statgraphics automatisk beregner:



Den kommer også med forslag hertil:
Vælg(Tabular Options | Unusual Residuals | OK). Resultatet blev

```
Unusual Residuals
--------------------------------------------------------------------------------
                                              Predicted                  Studentized
Row          X              Y                   Y         Residual        Residual
--------------------------------------------------------------------------------
16          120,0          25,2           22,4112         2,78885            2,43
18          150,0          25,1           27,6004        -2,5004            -2,20
--------------------------------------------------------------------------------
```

Statgraphics foreslår, at der er 2 punkter der er mistænkelige, idet den sætter grænsen ved 2. Det er imidlertid nok for snæver en grænse, idet man sædvanligvis først går ind, hvis "Studentized residual" er over 3.

2) Mindste kvadraters metode kan man altid udføre, så en ret linie og vurderinger ud fra en tegning kræver ingen forudsætninger.
Først når man vil udføre test, beregne konfidensintervaller osv, så må man forudsætte:

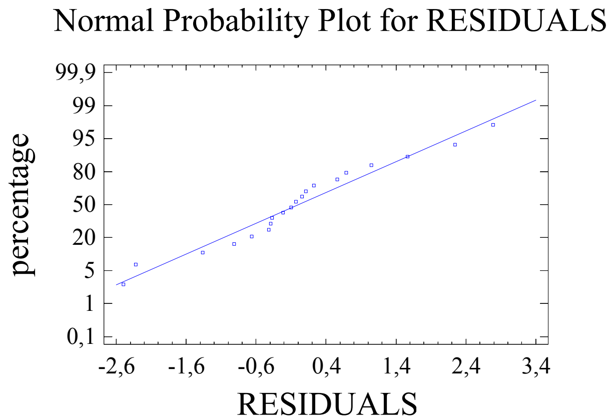
a) De enkelte observationer y_i er indbyrdes uafhængige (eksempelvis hvis der udføres flere målinger til samme tid skal de være indbyrdes uafhængige, ligesom det også skal gælde målinger til forskellige tidspunkter.
Punktet kan opfyldes ved en hensigtsmæssig forsøgsplan (eksempelvis ved at udføre et fuldstændigt randomiseret forsøg).

b) For hver værdi af x er $Y = E(Y|x) + \varepsilon = \beta_0 + \beta_1 x + \varepsilon$, hvor residualen ε er en statistisk variabel som forudsættes, at være normalfordelt med middelværdi 0 og varians σ^2.

Idet V(Y) = V(ε) skal målingerne derfor være værdier af uafhængige normalfordelte variable med samme varians.

Kravet om normalfordeling kan vurderes grafisk ved at tegne et normalfordelingsplot (se eventuelt appendix A)

Resultatet ses til højre.



Residualerne ligger ikke helt tilfældigt omkring en ret linie, og man må derfor antage, at residualerne ikke helt er normalfordelt. Da testen imidlertid er robust overfor mindre afvigelser fra normalitet slutter vi, at kravet om normalitet er tilstrækkeligt opfyldt.

Kravet om varianshomogenitet kan vurderes grafisk ud fra residualplottet. Det ses, at afstanden mellem to punkter svinger noget fra det ene par til det andet. Forskellene er dog ikke så store, at man på forhånd kan afvise at kravet om varianshomogenitet ikke er opfyldt.
Udføres Bartlets test (se evt. appendix A), fås:

```
Variance Check

Cochran's C test: 0,293578   P-Value = 0,851284
Bartlett's test: 1,84733   P-Value = 0,876251
Hartley's test: 23,04
```

Som det ses, kan man ikke forkaste, at de 10 varianser er ens, så vi vil i det følgende antage at kravet om varianshomogenitet er opfyldt.

3) Af regressionsanalysetabellen ses ud for model, at nulhypotesen H_0: $\beta = 0$ forkastes (3-stjernet), da P - value = 0.0000 < 0.001
Vi har følgelig Y afhænger af x.

4) Af udskriften ved "Estimate" og "Slope" aflæses $\tilde{\beta}_0 = 1.6542$ og $\tilde{\beta}_1 = 0.1730$.
Regressionsligningen bliver derfor $y = 1.6542 + 0.1730 \cdot x$.

5) Et konfidensinterval for $\tilde{\beta}_1$ kan kun fås i Statgraphics, ved at gå over i "Polynomial Regressionsanalyse".
Vælg (Relate | Polynomial Regression | indsæt t og y | cursor i udskrift | højre musetast | Analysis Options |sæt order til 1 | gul ikon = Tabular options | Confidence intervals | OK)
Denne giver

```
95,0% confidence intervals for coefficient estimates
----------------------------------------------------------------------------
                                       Standard
Parameter                  Estimate       Error     Lower Limit    Upper Limit
----------------------------------------------------------------------------
CONSTANT                    1,65415      0,599582       0,394468        2,91382
t                          0,172975    0,00603212       0,160302       0,185648
----------------------------------------------------------------------------
```

Et 95 % konfidensinterval er derfor [0.1603 ; 0.1856]

Benyttes lommeregner kan man ud for "slope" i udskriften under "Standard Error" findes et estimat s_β for spredningen på β_1.

$b \pm t_{0.975}(18) \cdot s_\beta = 0.172975 \pm 2.10 \cdot 0.00603212 = 0.172975 \pm 0.012667$ [0.16031 ; 0.18564]

(Formlen for konfidensintervallet er

$b \pm t_{0.975}(N-2) \cdot s_\beta$, hvor $s_\beta = \sqrt{\dfrac{s^2_{residual}}{SAK_x}}$ og $SAK_x = n \cdot (r-1) \cdot s_x^2$)

4) Det til $t = 100$ svarende 95% konfidensinterval for tykkelsen y fås ved:
Gå tilbage til "Linar Regression", vælg model som før (uden "lack of fit", da vi nu pooler de to varianser sammen)
Vælg (Gul ikon = Tabular Options | Forecast | OK | Cursor på udskrift | højre musetast | Pane Options) . Sæt i skemaet "Forecast at x" til 100 og resten stryges. Vi får da følgende tabel:

Predicted Values

		95,00%		95,00%	
	Predicted	Prediction Limits		Confidence Limits	
X	Y	Lower	Upper	Lower	Upper
100,0	18,9517	16,0518	21,8515	18,2956	19,6077

Vi får følgelig y(100) = 18.95 og 95% konfidensinterval [18,296 ; 19.608]

Benyttes lommeregner fås: Indsættes $t = 100$ i ligningen, fås $y = 18.95$.

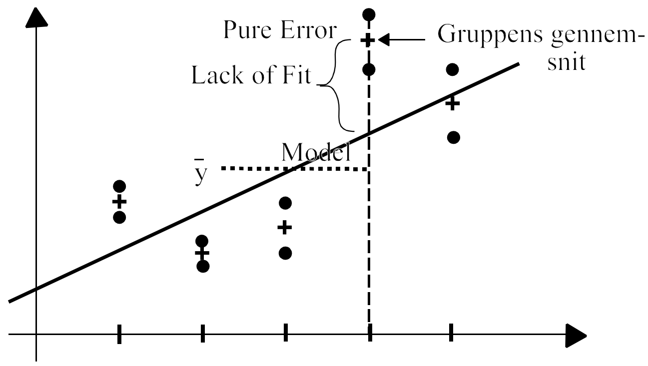
95% konfidensinterval findes af formlen:

$$\bar{\mu} \pm t_{0.975}(18)\cdot\sqrt{s^2_{residual}\cdot\left(\frac{1}{N}+\frac{(x-\bar{x})^2}{SAK_x}\right)} = \bar{\mu} \pm t_{0.975}(18)\cdot\sqrt{\frac{s^2_{residual}}{N}+(x-\bar{x})^2\,\frac{s^2_{residual}}{SAK_x}}$$

$$= 18.95 \pm 2.10\cdot\sqrt{\frac{1.80768}{20}+(100-86)^2\cdot 0.00603212^2} = 18.95 \pm 0.65577$$ dvs. [18.29 ; 19.61]

(formlerne i statistik II side 475 er anvendt).

Forklaring af de enkelte størrelser i regressionstabel



Figur 4.1 *SAK'ernes geometriske betydning*

$SAK_{\text{lack of fit}}$ = sum af kvadrater på de enkelte gruppers gennemsnits afstand fra regressionslinien.

Hvis $SAK_{\text{lack of fit}}$ er en lille størrelse ligger gennemsnitsværdierne tæt ved linien, og så giver linien en god beskrivelse af data. Frihedsgradstallet er $f_{\text{lack of fit}} = q - 2$, hvor q er antal x-værdier.
$SAK_o = SAK_{\text{pure error}}$ = sum af kvadrater på de enkelte værdiers afstand fra gruppens middelværdi.
$f_o = (n-1)\,q$, hvor n er antal gentagelser af hver x værdi.
$s_0^2 = \frac{SAK_0}{f_0}$ er variansen af den tilfældige variation.

Lad nulhypotesen være H_0: "Førstegradsmodellen gælder".

Hvis $s_{lack}^2 = \dfrac{SAK_{lack}}{f_{lack}}$ er sammenlignelig med variansen af "støjen" s_0^2, er modellen god nok.

Hvis $F_{lack} = \dfrac{s_{lack}^2}{s_0^2}$ ligger tæt ved 1, hvilket medfører at P - value er over 0.05, så accepteres modellen,

Da vi nu har to udtryk for den tilfældige variation, pooles disse til et fælles estimat for støjen $s_{residual}^2$ (kaldes i udskriften residual) med $f_{\text{residual}} = f_{\text{o}} + f_{\text{lack}}$

Den fremkomne "Residual" (i det følgende ofte kaldt residual-alle) er den residual der fremkommer hvis man tager kvadratsummen af alle punkters afstand til den bedste linie.

SAK_{model} er sum af kvadrater af liniens afstand fra det totale gennemsnit.

Den kan bruges til en vurdering af, om regressionskoefficienten (hældningskoefficienten eller slope) β er 0.

For bedre at forstå dette, kan vi også foretage en omskrivning af liniens ligning

Vi har

$$y_i = \tilde{\beta}_0 + \tilde{\beta}_1 x_i + \varepsilon_i,\ i = 1, 2, \ldots, n \qquad (1)$$

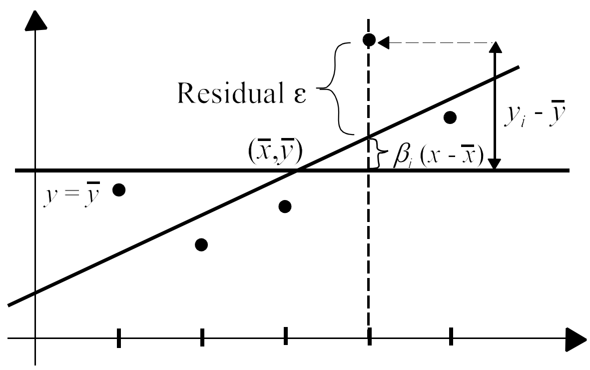
Det kan vises, at regressionslinien $Y = \tilde{\beta}_0 + \tilde{\beta}_1 x$ går gennem $(\bar{x}, \bar{y})$.

Vi har derfor : $Y = \tilde{\beta}_0 + \tilde{\beta}_1 x \Leftrightarrow Y = \bar{y} + \tilde{\beta}_1(x - \bar{x})$,

og dermed kan ligning (1) skrives $y_i = \bar{y} + \beta_1(x_i - \bar{x}) + \varepsilon_i$.

Løses denne sidste ligning med hensyn til residualen fås $\varepsilon_i = y_i - \bar{y} - \beta_i(x - \bar{x})$

På figur 4.2 er angivet de forskellige leds geometriske betydning.



Figur 4.2. *Leddenes geometriske betydning*

Sammenholdes figur 4.1 og 4.2 har man nu anskueligt :

$$y_i = \bar{y} + \beta_1(x_i - \bar{x}) + \varepsilon_i \quad \Leftrightarrow y_i = \bar{y}_i = \bar{y} + \text{Model} + \text{Residual}.$$

4.2.1.2 Regressionsanalyse uden gentagelser, transformation

I forsøgsplanlægning vil man sædvanligvis kunne planlægge forsøget, så man kan foretage flere målinger af forsøgsvariablen Y for samme værdieraf den uafhængige variabel x. Imidlertid kan der forekomme forsøg, hvor man ikke præcist kan indstille x-værdierne. Det kan eksempelvis være praktisk umuligt at indstille temperaturen præcist på en bestemt temperatur, eller man måler måske sammenhørende værdier af hårhed og brudstyrke på 25 tilfædigt udvalgte stykker aluminiums legering. I sådanne tilfælde har man intet udtryk for forsøgsfejlens variation, og kan derfor ikke lave en "lack of fit" test til at vurdere om modellen er god nok.
I sådanne tilfælde må man foretage en vurdering af modellen ved at se på en graf, lave residualplot, beregne et estimat r for "korrelationskoefficienten" eller se på "forklaringsgraden"
$$r^2 = \frac{SAK_{model}}{SAK_{total}}.$$
I det følgende eksempel illustreres hvor vigtigt det er også at se på grafer.

Eksempel 4.3 Betydning af grafer

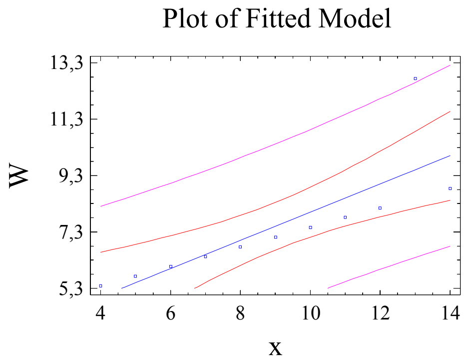
En statistiker har konstrueret følgende 3 talsæt (x,y) , (x,z), (x,w) *som giver stort set samme regressionsanalysetabel, samme korrelationskoefficient, men vidt forskellige grafer.*

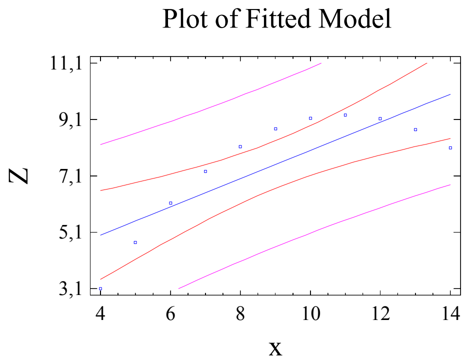
x	4	5	6	7	8	9	10	11	12	13	14
y	4.26	5.68	7.24	4.82	6.95	8.81	8.04	8.33	10.84	7.58	9.96
z	3.1	4.74	6.13	7.26	8.14	8.77	9.14	9.26	9.13	8.74	8.1
w	5.39	5.73	6.08	6.42	6.77	7.11	7.46	7.81	8.15	12.74	8.84

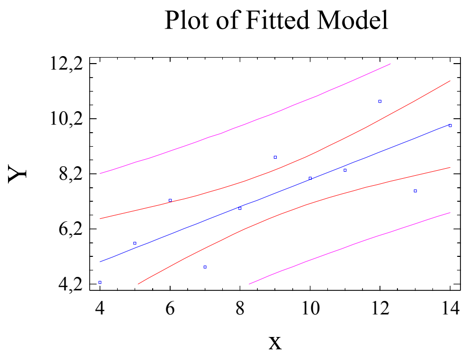
For y-dataene fås følgende regressionstabel, hvoraf man læser, at et estimat for korrelationskoefficienten $r = 0.8164$, altså ganske høj værdi, som tyder på at en lineær model er acceptabel.
For de øvrge to datasæt får man stort set samme regressionstabel og samme korrelationskoefficient (prøv selv!).

```
Simple Regression - Y vs. x
Regression Analysis - Linear model: Y = a + b*X
-----------------------------------------------------------------------------
Dependent variable: Y                  Independent variable: x
-----------------------------------------------------------------------------
                                Standard          T
Parameter        Estimate         Error       Statistic         P-Value
-----------------------------------------------------------------------------
Intercept         3,00009        1,12475        2,66735          0,0257
Slope            0,500091       0,117906        4,24146          0,0022
-----------------------------------------------------------------------------
                           Analysis of Variance
-----------------------------------------------------------------------------
Source             Sum of Squares     Df  Mean Square     F-Ratio      P-Value
-----------------------------------------------------------------------------
Model                       27,51      1        27,51       17,99       0,0022
Residual                  13,7627      9      1,52919
-----------------------------------------------------------------------------
Total (Corr.)             41,2727     10

Correlation Coefficient = 0,816421           R-squared = 66,6542 percent
Standard Error of Est. = 1,2366
```

Ser man på graferne får man et ganske andet billede.







Kun x - Y dataene synes med rimelighed at kunne beskrives ved en ret linie, mens x - Z dataene klart beskriver en krum kurve, og x - W dataene tydeligvis har en outliers, som bevirker, at den rette linie bliver helt forkert placeret i forhold til de øvrige punkter.

Eksempel 5.4. (transformation af model uden gentagelser)
Middelværdien af udbyttet Y af et kemikalium ved en bestemt kemisk reaktion afhænger af reaktionstemperaturen T. Ved et forsøg ønskes undersøgt om der indenfor det betragtede variationsområde er muligt at beskrive denne sammenhæng (eventuelt efter en transformation af data) ved et lineært udtryk i 2 parametre.
Ved forsøget fandtes følgende resultater:

T °C	10	15	20	25	30	35	40	45	50	55	60
Y gram	2.11	2.16	2.95	1.69	3.63	3.90	7.29	7.48	11.67	15.42	29.58

1) Begrund, at det næppe er rimeligt at det ønskede udtryk er af formen $Y = a \cdot Y + b$.
2) Benyt menupunktet “Comparison of Alternativ Methods” til at finde et funktionsudtryk, som på rimelig vis kan beskrive denne sammenhæng. Angiv funktionsudtrykket.

Løsning:

1) Data indtastes på sædvanlig måde.
Vælg (Relate | Simple Regression | indsæt *T* og *Y* | OK).
Der fremkommer en tegning og følgende variansanalysetabel:

Regression Analysis - Linear model: Y = a + b*X

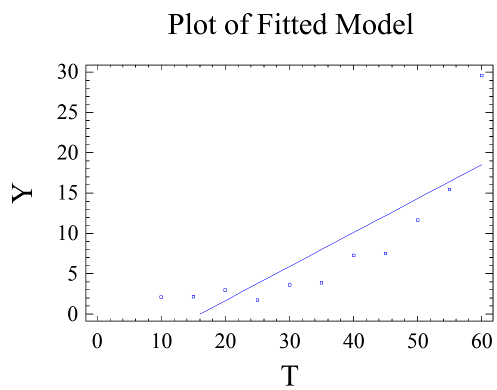
Dependent variable: Y
Independent variable: T

Parameter	Estimate	Standard Error	T Statistic	P-Value
Intercept	-6,76155	3,59201	-1,88239	0,0925
Slope	0,421473	0,0935279	4,50638	0,0015

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	488,508	1	488,508	20,31	0,0015
Residual	216,5	9	24,0556		
Total (Corr.)	705,008	10			

Correlation Coefficient = 0,832413   R-squared = 69,2911 percent



Det ses tydeligt af figuren, at en lineær model i x ikke er rimelig, da punkterne ikke fordeler sig tilfældigt omkring linien.
Forklaringsgraden på 69% er også temmelig lav.

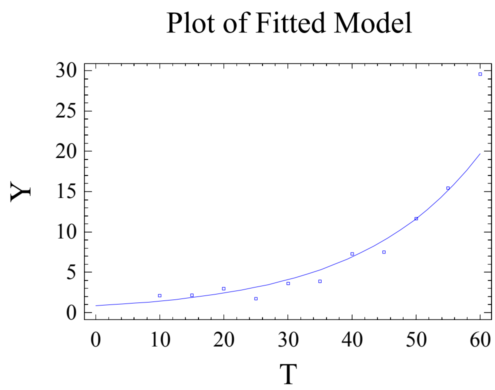
2) For at få en vurdering af hvilken funktion der kan passe bedst, så
Vælg(gul ikon = Tabular options | Comparison of Alternative Models | OK)
Det giver følgende tabel:

Comparison of Alternative Models

Model	Correlation	R-Squared
Exponential	0,9450	89,30%
Square root-Y	0,9064	82,16%
Reciprocal-Y	-0,8984	80,71%
Multiplicative	0,8573	73,50%
Linear	0,8324	69,29%
Square root-X	0,7764	60,28%
Double reciprocal	0,7511	56,42%
S-curve	-0,7196	51,79%
Logarithmic-X	0,7096	50,35%
Reciprocal-X	-0,5614	31,51%
Logistic	<no fit>	
Log probit	<no fit>	

Heraf ses, at den funktion, der tilsyneladende giver en god tilnærmelse er en exponential funktion.
Vælg (med cursor på udskrift, højre musetast | Analysis options | Exponential | OK)

Resultatet bliver følgende figur:



Punkterne synes i langt højere grad at ligge tilfældigt omkring kurven.

Udskriften viser som forventet, at forklaringsgraden er høj.

```
Regression Analysis - Exponential model: Y = exp(a + b*X)
-----------------------------------------------------------------------------
Dependent variable: Y
Independent variable: T
-----------------------------------------------------------------------------
                               Standard          T
Parameter       Estimate         Error       Statistic        P-Value
-----------------------------------------------------------------------------
Intercept      -0,168365       0,232583     -0,723894         0,4875
Slope          0,0524816     0,00605595       8,66612         0,0000
-----------------------------------------------------------------------------
                           Analysis of Variance
-----------------------------------------------------------------------------
Source                 Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-----------------------------------------------------------------------------
Model                         7,57438      1      7,57438      75,10       0,0000
Residual                     0,907695      9     0,100855
-----------------------------------------------------------------------------
Total (Corr.)                 8,48207     10

Correlation Coefficient = 0,94498          R-squared = 89,2987 percent
```

Funktionsudtrykket bliver: $y = e^{-0.168365+0.0524816T} = \underline{\underline{0.845 \cdot 1.053988^T}}$

Øvelse 4.1 (lineær regressionsanalyse, transformation) Regn opgave 28 side 247.

4.2.2 Polynomial regressionsanalyse

4.2.2.1 Indledning
Ved en polynomial regressionsanalyse er den statistiske model
$Y = \alpha_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + ... + \beta_p x^p$.
hvor den variable Y skal opfylde de sædvanlige regressionsforudsætninger.
Som det ses, er den i afsnit 4.2.1 betragtede enkelte regression et specialtilfælde. Den statistiske analyse da også meget beslægtet hermed.
En polynomial regressionsanalyse anvendes ofte, hvor man ikke kender sammenhængen mellem Y og x, men er interesseret i at bestemme et polynomium (af lavest mulig grad), der giver en tilstrækkelig god beskrivelse af Y indenfor det foreliggende variationsområde for x. Det skal dog her nævnes, at mens eksempelvis et andengradspolynomium har 3 parametre, der skal bestemmes, har de i forrige afsnit nævnte "transformerede" modeller kun 2 parametre. Ud fra et statistisk synspunkt, vil man altid foretrække den model med de færreste parametre, da de på samme datamateriale giver en sikrere bestemmelse af parametrene..
Ved planlagte forsøg vil man sædvanligvis sørge for gentagelser, da vi herved får et skøn for forsøgsfejlen, og derved kan teste om modellen er god nok. Det vil her ofte være tilstrækkeligt at benytte dobbeltbestemmelser. Har man ingen forhåndsformodninger vil en fornuftig forsøgsplan være at foretage målingerne med ækvidistante x-værdier. I det følgende eksempel 5.5 har dette været kriterierne. Har man på forhånd en formodning om, at der indenfor et bestemt delområde sker noget særligt interessant (eksempelvis nogle voldsomme ændringer i Y-værdierne) så må man her indlægge flere målepunkter.
I en række tilfælde vil data være såkaldte "drop in" data. Et eksempel herpå ses i opgave 19, hvor man måler sammenhørende værdier af vindhastighed og den jævnstrøm en vindmølle udvikler. Her er det ikke muligt at foretage nogen egentlig planlægning, da det jo vil være et rent tilfælde, hvis man eksempelvis har en gentagelse. Selv om emnet her er planlagte forsøg vil vi dog også kort i afsnit 4.2.2.3 se på, hvad man kan gøre i sådanne tilfælde.

4.2.2.2 Polynomial regressionsanalyse med gentagelser.

I dette afsnit benyttes følgende eksempel til illustration af den statistiske analyse.

Eksempel 4.2 (polynomial regression)
Man ved, at tilsættes et bestemt additiv en dunk fernis, så forkortes størkningstiden (den tid det tager for fernissen at tørre). Et forsøg udføres, for at finde hvordan størkningstiden T (i minutter) afhænger af antal gram x af additivet. Man fik følgende forsøgsresultater:

x g/l	0	1	2	3	4	5	6	7	8
T minut-ter	740 710	610 650	470 540	440 420	400 450	440 480	530 470	420 480	450 490

1) *Giv på basis af ovennævnte observationer en grafisk vurdering af, hvilket polynomium* $T = \alpha_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + ... + \beta_p x^p$ *, der indenfor måleområdet* [0 ; 8] *giver en tilfredsstillende beskrivelse af T's variation.*
2) *Test om man på basis af ovennævnte observationer kan få en accept af, at ovennævnte polynomium af p'te grad i x kan antages at kunne beskrive sammenhængen mellem T og x, og vurder også grafisk om modellen er rimelig.*

I bekræftende fald ønskes

3) *en test af, om modellen kan reduceres til et polynomium af lavere grad, samt en grafisk kontrol af om den derved fremkomne model er rimelig.*
4) *angivet regressionsligningen* $T = a_0 + b_1 x + b_2 x^2 + b_3 x^3 + ... + b_p x^p$ *hvor* $a_0, b_1, ..., b_p$ *er estimater for* $\alpha_0, \beta_1, ..., \beta_p$, *for den model, man i spørgsmål 3 har fundet frem til.*
5) *angivet et 95% konfidensinterval for regressionskoefficienten* β_p *til leddet af højest grad.*
6) *fundet et estimat for den værdi* x_m *(1 decimal) af x ,for hvilken størkningstiden T er mindst. Endvidere ønskes angivet den til* x_m *svarende estimerede middelværdi* $\tilde{T}_m$, *og et 95% konfidensinterval for* $\tilde{T}_m$.

LØSNING:

Spørgsmål 1. Grafisk vurdering af polynomiums grad.

Data indtastes i Statgraphics på sædvanlig måde.

Da vi har i alt 9 punkter (x, $\bar{T}$) , hvor $\bar{T}$ er de 9 gennemsnit, kan et 8 grads polynomium

$T = \alpha_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + ... + \beta_8 x^8$ gå eksakt gennem disse punkter.

Det vil imidlertid dels være et alt for kompliceret polynomium til de fleste praktiske formål, dels følge alle de tilfældige variationer, som vi netop ikke bør tage hensyn til.

Et grafisk overblik over punkterne og polynomiets grad fås på følgende måde:

Vælg (Relate| Polynomial Regression | indsæt x og T i den fremkomne tavle | OK)

Der vises en variansanalysetabel svarende til en andengradsmodel (det er altid startværdien for Statgraphics). Samtidig får man en figur, hvor punkterne er indtegnet. For at gøre figuren klarere fjernes konfidenslinier m.m.

Vælg (Cursor på figur | højre musetast | Pane Options | fjern markeringerne | OK).

Vi får så følgende figur:

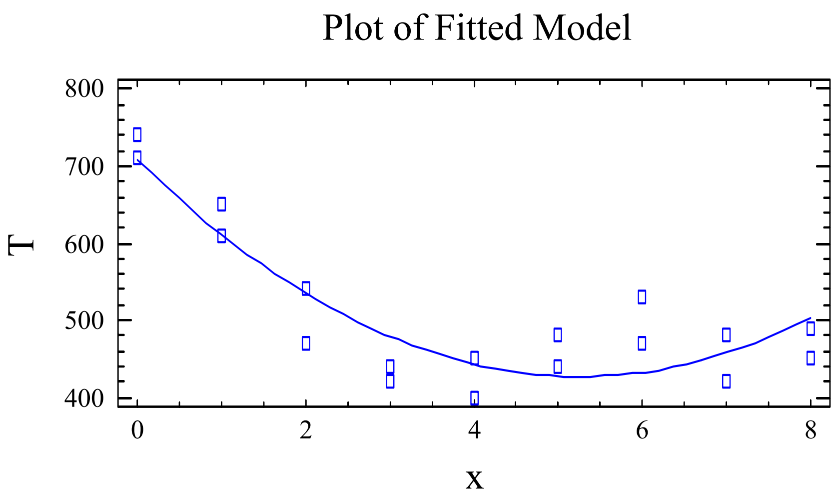


Fig 2.1 *Polynomial regression*

Punkternes gennemsnit kunne godt ligge på en parabel, hvis ikke det var for punkterne med store x-værdier. Her synes størkningstiden igen at falde, så grafen synes med tilnærmelse snarere at kunne ligge på en trediegradskurve.

Af figuren ses også, at der synes at være varianshomogenitet.

Et trediegradspolynomium må derfor antages at kunne give en tilfredsstillende beskrivelse af Y's variation.

Spørgsmål 2. Accept af trediegradsmodel.
Opstilling af regressionsanalyse tabel
For at få en trediegradsmodel:
Vælg (Cursor på udskrift | højre musetast | Analysis Options |order til 3 | OK).
Dette giver følgende variansanalysetabel.

```
Polynomial Regression Analysis
-----------------------------------------------------------------------------
Dependent variable: T
-----------------------------------------------------------------------------
                                       Standard          T
Parameter                Estimate        Error       Statistic        P-Value
-----------------------------------------------------------------------------
CONSTANT                  743,232       24,5077        30,3265         0,0000
x                        -183,255       28,2854       -6,47877         0,0000
x^2                       35,5231       8,54009        4,15957         0,0010
x^3                        -2,117      0,700505       -3,02211         0,0091
-----------------------------------------------------------------------------

                           Analysis of Variance
-----------------------------------------------------------------------------
Source              Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-----------------------------------------------------------------------------
Model                     154907,0      3      51635,7      36,91       0,0000
Residual                   19587,5     14      1399,11
-----------------------------------------------------------------------------
Total (Corr.)             174494,0     17

R-squared = 88,7747 percent
R-squared (adjusted for d.f.) = 86,3693 percent
```

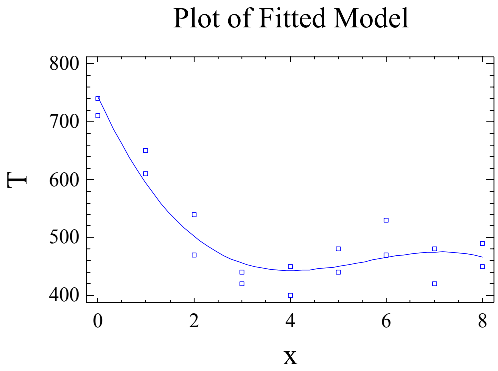
Vurdering af om trediegradsmodel kan accepteres
Da vi har gentagelser, ignoreres den fremkomne udskrift og vi foretager en "lack of fit" test.
Denne giver følgende tabel:

```
Analysis of Variance with Lack-of-Fit
-----------------------------------------------------------------------------
Source              Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-----------------------------------------------------------------------------
Model                     154907,0      3      51635,7      36,91       0,0000
Residual                   19587,5     14      1399,11
-----------------------------------------------------------------------------
   Lack-of-Fit             9237,48      5       1847,5       1,61       0,2526
   Pure Error              10350,0      9       1150,0
-----------------------------------------------------------------------------
Total (Corr.)             174494,0     17
```

Da P-value for Lack-of fit er 0,2526 > 0,05 accepteres trediegradsmodellen.
Som beskrevet under enkel regressionsanalyse, viser udskriften ved "Lack of Fit Test", at $F = \frac{s_{lack}^2}{s_0^2} = 1{,}61$ og denne værdi er ikke stor nok til at forkaste trediegradsmodellen.

Som grafisk kontrol tegnes

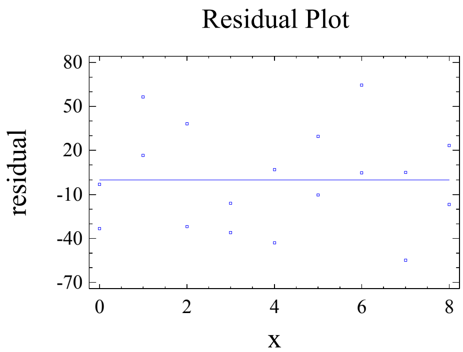
a) den fundne kurve og punkterne for at se, om modellen er rimelig.
   I Statgraphics sker det automatisk på højre side af udskriften. Den fremkomne tegning forenkles ved at slette de indtegnede linier for konfidensgrænser og prædiktionsgrænser (predictionslimits).



De 9 punkters gennemsnitsværdier synes at fordele sig tilfældigt omkring den fundne kurve

b) Man indtegner et residualplot (residualerne $r_i = \hat{y}_i - y_i$ er forskellen mellem de beregnede y-værdier $\hat{y}_i$ og de observerede værdier y_i).
Vi ser, at punkternes "gennemsnit" ligger tilfældigt omkring linien.

Endvidere synes også forudsætningen om varianshomogenitet at være opfyldt, selv om det nok kunne være rimeligt at udføre en Bartletts test.



Vi ser, at der <u>ikke er outliers</u>, da $3 \cdot \sigma = 3\sqrt{1150} = 102 > 80$

Et normalfordelingstest overlades til læseren at foretage.
<u>Samlet konklusion : Trediegradsmodellen synes at være rimelig.</u>

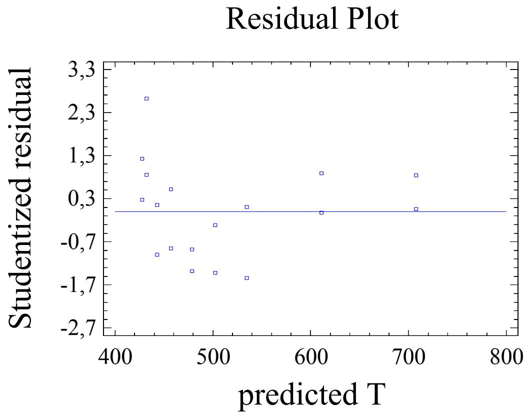
Spørgsmål 3. Reduktion af model.
Nu testes om en andengradsmodel kan anvendes .
Vælg (Cursor på udskrift | højre musetast | Analysis Options |order til 2 | OK).

Analysis of Variance with Lack-of-Fit

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	142129,0	2	71064,4	32,94	0,0000
Residual	32365,7	15	2157,71		
Lack-of-Fit	22015,7	6	3669,29	3,19	0,0579
Pure Error	10350,0	9	1150,0		
Total (Corr.)	174494,0	17			

Da P-value for Lack of fit er 0,0579 > 0,05 <u>accepteres andengradsmodellen.</u>

Vi betragter nu et residualplot:



Punkternes "gennemsnit" synes som også tidligere bemærket ikke at ligge helt tilfældigt omkring linien, så selv om punkterne ligger så tæt på linien at andengradsmodellen accepteres, så er det under ingen omstændigheder tilrådeligt, at anvende modellen ud over det angivne interval (at ekstrapolere)
Selv om en enkelt værdi har en "studentizied residual" på mere end 2, så vil vi ikke betragte det som en outliers. (skal være over 3)

Spørgsmål 4. Angiv regressionsligningen.
Vi får:

```
Polynomial Regression Analysis
-----------------------------------------------------------------------------
Dependent variable: T
-----------------------------------------------------------------------------
                                        Standard          T
Parameter               Estimate          Error       Statistic        P-Value
-----------------------------------------------------------------------------
CONSTANT                 707,667        26,6964        26,5079          0,0000
x                       -106,619        15,5615       -6,85148          0,0000
x^2                       10,119        1,87157        5,40671          0,0001
-----------------------------------------------------------------------------

                           Analysis of Variance
-----------------------------------------------------------------------------
Source             Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-----------------------------------------------------------------------------
Model                    142129,0      2      71064,4      32,94       0,0000
Residual                  32365,7     15      2157,71
-----------------------------------------------------------------------------
Total (Corr.)            174494,0     17

R-squared = 81,4517 percent    R-squared (adjusted for d.f.) = 78,9786 percent
Standard Error of Est. = 46,4512
```

Ligningen ses af udskriften at være: $\underline{\underline{\tilde{T} = 707.667 - 106.619x + 10.119x^2}}$

Spørgsmål 5. Opstilling af 95% konfidensinterval for β_2 .

Der vælges " Confidence intervals". Resultatet bliver:

95,0% confidence intervals for coefficient estimates

Parameter	Estimate	Standard Error	Lower Limit	Upper Limit
CONSTANT	707,667	26,6964	650,764	764,569
x	-106,619	15,5615	-139,788	-73,4505
x^2	10,119	1,87157	6,12988	14,1082

Et 95% konfidensinterval for β_2 er følgelig [6.13 ; 14.11].

Spørgsmål 6. Find den værdi x_m som giver den mindste størkningstid . Angiv endvidere den til x_m svarende estimerede middelværdi $\tilde{T}_m$ og et 95% konfidensinterval for $\tilde{T}_m$.

Af figur 2.1 ses, at den ønskede værdi x_m må være ca 5. En mere præcis værdi fås ved at differentiere udtrykket for T . Dette giver

$$x_m = \frac{106.619}{2 \cdot 10.119} \approx 5.27$$

I Statgraphics vælges " Forecasts " | højre musetast | Pane options | x" til 5.27 .

Vi får da følgende tabel:

Predicted Values

X	Predicted Y	95,00% Prediction Limits Lower	95,00% Prediction Limits Upper	95,00% Confidence Limits Lower	95,00% Confidence Limits Upper
5,27	426,82	322,486	531,153	393,914	459,725

Vi får følgelig $\tilde{T}_m = 426.82$ minutter og 95% konfidensinterval [393.9 ; 459.7]

Øvelse 4.2 (polynomial regressionsanalyse). Regn opgave 29 side 248.

4.2.2.3. Polynomial regressionsanalyse uden gentagelser

Har man **ingen gentagelser** kan man ikke teste modellen. Man ser så på "forklaringsgraden" R^2 eller R-squared $= \frac{SAK_{Total} - SAK_0}{SAK_{Total}} = \frac{SAK_{model}}{SAK_{Total}}$, og hvis den er over 80-85% vil man normalt godkende modellen. Man skal imidlertid være forsigtig hermed. Det er altid muligt at øge R^2 ved at addere flere led til modellen, eksempelvis vil et *n* - 1' te grads polynomium altid kunne gå gennem *n* punkter og $R^2 = 1$, men det betyder ikke nødvendigvis at denne model er den bedste. For hvert led der tilføjes mistes der en frihedsgrad i residualen, og hvis SAK for det nye led ikke giver et væsentligt bidrag kan det betyde, at den nye model er ringere end den gamle model. For at tage hensyn til dette, betragtes ofte et modificeret R^2 (R-squared (adjusted for d.f.[1])

Man bør dog i alle tilfælde se på et plot over residualerne. Disse bør ligge "tilfældigt " omkring 0. Hvis dette ikke er tilfældet, bør man være på vagt, og eventuelt gå op til en model af højere grad.

I dette afsnit benyttes igen eksempel 4.2 til illustration af den statistiske analyse, idet vi nu antager, at der kun er sket 1 måling for hver x-værdi.

Eksempel 4.4 (eksempel 4.2 uden gentagelser.)
Man ved, at tilsættes et bestemt additiv en dunk fernis, så forkortes størkningstiden (den tid det tager for fernissen at tørre). Et forsøg udføres, for at finde hvordan størkningstiden T (i minutter) afhænger af antal gram x af additivet. Man fik følgende forsøgsresultater:

x g/l	0	0,5	1	1,5	2	2,5	3	3,5	4	4,5	5	5,5	6	6,5	7	7,5	8	8,5
T min.	740	710	610	650	470	540	440	420	400	450	440	480	530	470	420	480	450	490

1) *Vurder på basis af ovennævnte observationer ud fra forklaringsgraden, hvilket polynomium*
   $T = \alpha_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \ldots + \beta_p x^p$ *af lavest mulig grad p, der indenfor måleområdet* [0 ; 8] *giver en tilfredsstillende beskrivelse af T's variation.*
2) *Giv en grafisk vurdering af, om det i 1) fundne polynomium indenfor måleområdet* [0 ; 8] *giver en tilfredsstillende beskrivelse af T's variation.*

I bekræftende fald ønskes

3) *en test af, om modellen kan reduceres til et polynomium af lavere grad, samt en grafisk kontrol af om den derved fremkomne model er rimelig.*
4) *angivet regressionsligningen* $T = a_0 + b_1 x + b_2 x^2 + b_3 x^3 + \ldots + b_p x^p$ *hvor* $a_0, b_1, \ldots, b_p$ *er estimater for* $\alpha_0, \beta_1, \ldots, \beta_p$ *, for den model, man i spørgsmål 3 har fundet frem til.*

---

[1] R^2 (adjusted) $= \frac{(n-1) \cdot R^2 + 1 - k}{n-k} = \frac{s_{total}^2 - s_0^2}{s_{total}^2}$, hvor k er antal parametre i modellen(incl konstantled).

LØSNING:
Spørgsmål 1. Vurdering ud fra forklaringsgrad om polynomiums grad.
Data indtastes i Statgraphics på sædvanlig måde.

Opstilling af regressionsanalysetabeller til beregning af forklaringsgrad.
Vi får først en variansanalysetabel svarende til en andengradsmodel (det er altid startværdien for Statgraphics).

```
Polynomial Regression Analysis
-----------------------------------------------------------------------------
Dependent variable: T
-----------------------------------------------------------------------------
                                       Standard         T
Parameter                Estimate        Error       Statistic        P-Value
-----------------------------------------------------------------------------
CONSTANT                  727,544       31,5776       23,0399          0,0000
x                        -107,975       17,2274      -6,26761          0,0000
x^2                       9,75748       1,95585       4,98886          0,0002
-----------------------------------------------------------------------------
                           Analysis of Variance
-----------------------------------------------------------------------------
Source             Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-----------------------------------------------------------------------------
Model                    137427,0      2      68713,3      27,81       0,0000
Residual                  37067,8     15      2471,18
-----------------------------------------------------------------------------
Total (Corr.)            174494,0     17
R-squared = 78,7571 percent   R-squared (adjusted for d.f.) = 75,9247 percent
```

Vi ser, at R-squared (adjusted) er 75,92%
Vi opstiller nu en trediegradsmodel, “ Order til 3"
Dette giver følgende variansanalysetabel.

```
Polynomial Regression Analysis
-----------------------------------------------------------------------------
Dependent variable: T
-----------------------------------------------------------------------------
                                       Standard         T
Parameter                Estimate        Error       Statistic        P-Value
-----------------------------------------------------------------------------
CONSTANT                  770,702        34,522       22,3249          0,0000
x                         -179,27       36,2105      -4,95077          0,0002
x^2                       31,3364       10,0704       3,11174          0,0077
x^3                      -1,69247      0,777816      -2,17592          0,0472
-----------------------------------------------------------------------------
                           Analysis of Variance
-----------------------------------------------------------------------------
Source             Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-----------------------------------------------------------------------------
Model                    146794,0      3      48931,5      24,73       0,0000
Residual                  27699,9     14      1978,57
-----------------------------------------------------------------------------
Total (Corr.)            174494,0     17
R-squared = 84,1256 percent   R-squared (adjusted for d.f.) = 80,7239 percent
```

Vi ser, at R-squared (adjusted) er steget til 80,72%.

Vi vælger nu en fjerdegradsmodel. Dette giver følgende variansanalysetabel.

```
Polynomial Regression Analysis
-----------------------------------------------------------------------------
Dependent variable: T
-----------------------------------------------------------------------------
                                       Standard          T
Parameter               Estimate         Error       Statistic        P-Value
-----------------------------------------------------------------------------
CONSTANT                 758,692       39,7228         19,0997         0,0000
x                        -141,24       68,7863        -2,05331         0,0607
x^2                      9,84872       34,3492        0,286723         0,7788
x^3                      2,31072       6,15732        0,375281         0,7135
x^4                    -0,235482       0,35917       -0,655628         0,5235
-----------------------------------------------------------------------------
                           Analysis of Variance
-----------------------------------------------------------------------------
Source             Sum of Squares     Df  Mean Square     F-Ratio      P-Value
-----------------------------------------------------------------------------
Model                   147681,0       4      36920,3       17,90       0,0000
Residual                 26813,4      13      2062,57
-----------------------------------------------------------------------------
Total (Corr.)           174494,0      17

R-squared = 84,6337 percent     R-squared (adjusted for d.f.) = 79,9056 percent
```

Vi ser, at R-squared (adjusted) nu er faldet svagt fra 80,72% til 79,90%. Heraf må sluttes, at fjerdegradsmodellen ikke har givet et væsentligt forbedret bidrag til forklaring af data. Dette stemmer også med, at *P*-value for x^4 er 0.5232 > 0.05.

Bemærk iøvrigt, at selv om alle *P*-værdier for koefficienterne er større end 0,05, kan vi ikke deraf slutte, at vi kan reducere modellen til en konstant model. Man må kun bortkaste et led af gangen.

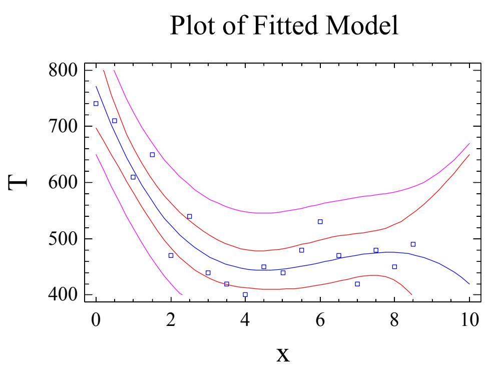
<u>Det anførte tyder på, at en trediegradsmodel er en acceptabel model.</u>

Spørgsmål 2: Grafisk kontrol af model

Man bør altid som en ekstra kontrol indtegne kurven og punkterne.

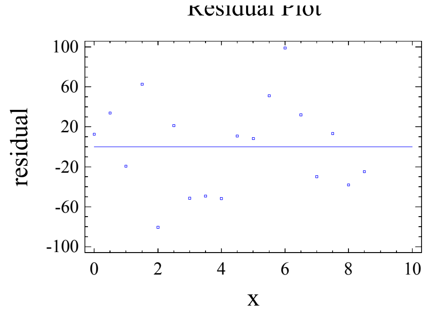
I Statgraphics sker det automatisk på højre side af udskriften. Den fremkomne tegning har indtegnede linier for konfidensgrænser og prædiktionsgrænser (predictionslimits).

De 16 punkter synes at fordele sig rimeligt omkring den fundne kurve

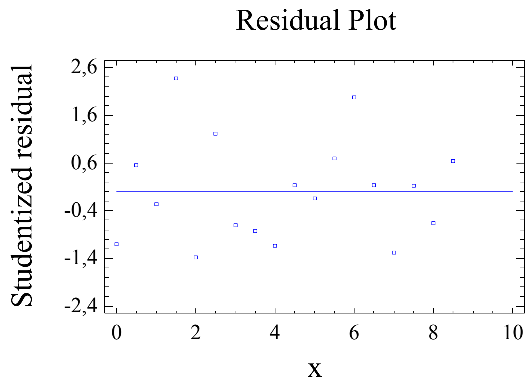


Vi laver endvidere som sædvanlig et residualplot (tegningen til højre).

Vi ser igen, at punkterne ligger tilfældigt omkring linien.

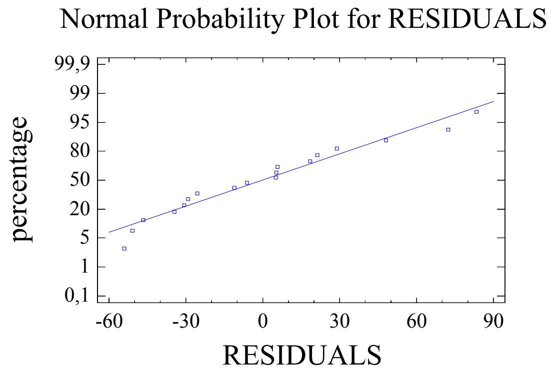


Vi danner et studentized residualplot.



Heraf ses, at der ikke er nogle "outliers", da ingen værdier er over 3

Et normalfordelingsplot udføres som beskrevet i appendix A.



Resultatet viser, at residualerne ligger rimelig tilfældigt omkring en ret linie, så forudsætningen om normalitet er opfyldt.

Samlet konklusion : Grafisk synes trediegradsmodellen at være rimelig.

Spørgsmål 3. Reduktion af model.
Da *P*-value for x^3 er 0.0472 < 0.05 er der 1 stjernet signifikans mod nulhypotesen H_0: $\beta_3 = 0$.
Både dette og de foregående betragtninger over R-squared (adjusted) gør, at det ikke er rimeligt at reducere modellen yderligere

Spørgsmål 4. Angiv regressionsligningen.
Ligningen ses af udskriften for trediegradsmodellen at være:

$$\underline{\underline{\tilde{T} = 770.70 - 179.27x + 31.336x^2 - 1.629x^3}}\,.$$

Øvelse 4.3 (regressionsanalyse uden gentagelser) Regn opgave 30 side 248.

4.3 MULTIPEL REGRESSIONSANALYSE

4.3.1 Indledning

Vi vil i dette afsnit behandle det tilfælde, hvor der indgår mere end én kvantitativ faktor. Vi vil fortrinsvis som eksempler anvende modeller med to kvantitative variable x_1, og x_2.
Vi vil først betragte modeller, hvor de variable indgår lineært som i $Y = \alpha_0 + \beta_1 x_1 + \beta_2 x_2$, hvor parametrene er α_0, β_1 og β_2 . Til sidst vil blive betragtet polynomiale modeller som $Y = \alpha_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2$, hvor parametrene er $\alpha_0, \beta_1, \beta_2, \beta_3, \beta_4$ og β_5 .
Ved forsøgsplanlægning vil man så vidt muligt basere sine forsøg på en fuldstændig faktorstruktur med lige mange gentagelser i hver celle. Dette er imidlertid ikke altid muligt, så vi vil også behandle tilfældet, hvor vi kun har en partiel struktur. I praksis har man ofte ingen gentagelser, og i sådanne tilfælde kan modellen ikke testes. Man må i sådanne tilfælde på samme måde som i afsnit 4.2.2.3 grafisk se på residualerne samt på R-Square, for at få et indtryk af om en lineær model er rimelig.
I appendix 4.1 er givet begrundelser for de væsentligste formler der benyttes i dette kapitel, ligesom der kort er skitseret hvorledes man ved hjælp af et program som Maple kan foretage beregningerne.

4.3.2 Modeller hvor de variable indgår lineært

4.3.2.1 Fuldstændig faktorstruktur (med lige mange observationer i hver celle)

Vi vil i dette afsnit benytte følgende eksempel.

Eksempel 4.5 Fuldstændig faktorstruktur. (hentet fra SF1 eksempel 36 side 184)
En fabrik fremstiller salpetersyre ved oxidering af ammoniak med luft. I løbet af processen ledes kvælstofoxider under afkøling ind i en absorptionskolonne, idet absorptionen i gennemstrømmende salpetersyre afhænger af kølevandstemperaturen x_1 (°C) og lufttemperaturen x_2 (°C) (absorptionen formindskes jo højere disse temperaturer er).
Følgende observationer af Y (kodede tal) fandtes:

		x_2			
		10	20	30	40
x_1	5	37 31	49 54	68 61	87 93
	10	50 52	65 62	84 75	94 100
	15	58 63	77 79	88 94	119 115

1) *Test om man på basis af ovennævnte observationer kan få en accept af, at sammenhængen mellem mængden Y af ikke-absorberede kvælstofoxider i et givet tidsrum og temperaturerne x_1 og x_2 (approksimativt) er lineær, dvs. på formen $Y = \alpha_0 + \beta_1 x_1 + \beta_2 x_2$.*
2) *Kontroller grafisk om modellen er rimelig.*

Det antages i det følgende, at ovenstående model gælder.

3) *Undersøg om modellen kan reduceres, dvs. om $\beta_1 = 0$ og/eller $\beta_2 = 0$.*

4) *Angiv regressionsligningen* $Y = a_0 + b_1x_1 + b_2x_2$, *hvor* a_0, b_1 *og* b_2 *er estimater for* α_0, β_1 og β_2
5) *Angiv et 95% konfidensinterval for* β_1 og β_2 .
6) *Find den til* $x_1 = 10$ og $x_2 = 20$ *svarende estimerede middelværdi* $\widetilde{Y}$ *og et 95% konfidens-interval for* $\widetilde{Y}$.

LØSNING:

Spørgsmål 1: Test af om lineær model kan accepteres

Indtastning af data. I Statgraphics indtastes tallene på sædvanlig måde og vi får

```
x1   x2   oxid
5    10   37
5    10   31
5    20   49
5    20   54
5    30   68
5    30   61
5    40   87
5    40   93
10   10   50
10   10   52
10   20   65
10   20   62
10   30   84
10   30   75
10   40   94
10   40   100
15   10   58
15   10   63
15   20   77
15   20   79
15   30   88
15   30   94
15   40   119
15   40   115
```

Opstilling af regressionsanalysetabel

Vælg ikonen "Multiple Regression" eller Vælg (Relate | Multiple Regression).
I den fremkomne tavle indsættes de variable og der fremkommer så følgende tabel.

```
Multiple Regression Analysis
-----------------------------------------------------------------------------
Dependent variable: oxid
-----------------------------------------------------------------------------
                                      Standard          T
Parameter               Estimate        Error       Statistic        P-Value
-----------------------------------------------------------------------------
CONSTANT                   3,375      2,97685         1,13375         0,2697
x1                        2,6625     0,210495         12,6488         0,0000
x2                         1,725    0,0768618         22,4429         0,0000
-----------------------------------------------------------------------------
                           Analysis of Variance
-----------------------------------------------------------------------------
Source             Sum of Squares      Df  Mean Square     F-Ratio      P-Value
-----------------------------------------------------------------------------
Model                     11762,4       2      5881,22      331,84       0,0000
Residual                  372,188      21      17,7232
-----------------------------------------------------------------------------
Total (Corr.)             12134,6      23
```

Den fremkomne "Residual" (i det følgende kaldt residual-alle) er den residual der fremkommer hvis man tager kvadratsummen af <u>alle</u> punkters afstand til den bedste plan.

Da vi har gentagelser kan vi teste modellen ved at spalte $SAK_{residual\text{-}alle}$ op i $SAK_{gentagelser}$ og en $SAK_{lack\ of\ fit}$ hvor den sidste er Kvadratsummen af afstanden fra de enkelte cellers gennemsnit til planen.
Imidlertid findes der ikke her nogen "Lack of fit" menu. Vi må derfor på én af de følgende 3 beskrevne måder finde en værdi for $SAK_{gentagelser}$.

1) Indsætte en ny søjle ("behandlinger") i dataark med tallene 1,1 2,2 3,3 ..., 12, 12 svarende til de 12 behandlinger (celler). Derefter udføres en ensidet variansanalyse med behandlinger som uafhængig variabel. Metoden kan ses i eksempel 4.6.
2) En anden mulighed ville have været, at indføre alle de højere led i modellen. Da x_1 har 3 niveauer, kan der højst optræde led indeholdende x_1 af 2-grad. Analogt kan der højst optræde led indeholdende x_2 af 3-grad. Vi får derfor
$$Y = \alpha_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2 + \beta_6 x_1^2 x_2 + \beta_7 x_1 x_2^2 + \beta_8 x_2^3 + \beta_9 x_1 x_2^3 + \beta_{10} x_1^2 x_2^2 + \beta_{11} x_1^2 x_2^3$$
Dette sidste vil (som allerede dette eksempel viser) ofte blive ganske omfattende, hvis man har mange variable.
3) Udføre en variansanalyse med maksimalt antal vekselvirkninger (den mest omfattende model), hvoraf vi kan aflæse "gentagelserne". Dette kan dog kun gøres, hvis forsøget er en fuldstændig faktorstruktur, men så er det til gengæld den hurtigste metode.

I dette tilfælde vælges derfor metode 3
Vi vælger " Compare" og på sædvanlig måde sættes "Maximum Order Interaction til 2 "
Resultatet bliver:

```
Analysis of Variance for oxid - Type III Sums of Squares
--------------------------------------------------------------------------------
Source                    Sum of Squares     Df    Mean Square    F-Ratio    P-Value
--------------------------------------------------------------------------------
MAIN EFFECTS
 A:x1                            2837,25      2        1418,63      95,37     0,0000
 B:x2                            9039,13      3        3013,04     202,56     0,0000

INTERACTIONS
 AB                                79,75      6        13,2917       0,89     0,5294

RESIDUAL                           178,5     12         14,875
--------------------------------------------------------------------------------
TOTAL (CORRECTED)                12134,6     23
--------------------------------------------------------------------------------
All F-ratios are based on the residual mean square error.
```

Heraf ses, at $SAK_0 = 178{,}5$ med f = 12 (12 celler).
Da $SAK_{residual\text{-}alle} = SAK_0 + SAK_{lack}$ fås $SAK_{lack} = 372.188 - 178.5 = 193.687$
med $f_{lack} = 21 - 12 = 9$.

Vi har følgelig at $s_{lack}^2 = \dfrac{SAK_{lack}}{f_{lack}} = \dfrac{193.687}{9} = 21.52$ og dermed $F_{lack} = \dfrac{s_{lack}^2}{s_0^2} = 1.45$.

Dette giver følgende variansanalysetabel:

Variation	SAK	f	s^2	F
Model	11762,4	2		
Lack of fit	193.687	9	21.52	1.45
Gentagelser	178.5	12	14.875	
Total	12134.5	23		

Da $F_{0.95}(9,12) = 2{,}80 > F_{lack} = 1.45$ accepteres den lineære model.

Vi kunne også have beregnet en P - værdi ved hjælp af Statgraphics (se eventuelt appendix A) Den giver udskriften:

```
Cumulative Distribution
-----------------------

Distribution: F (variance ratio)

               Lower Tail Area (<)
Variable       Dist. 1        Dist. 2        Dist. 3        Dist. 4        Dist. 5
1,45           0,742857

               Probability Density
Variable       Dist. 1        Dist. 2        Dist. 3        Dist. 4        Dist. 5
1,45           0,40805

               Upper Tail Area (>)
Variable       Dist. 1        Dist. 2        Dist. 3        Dist. 4        Dist. 5
1,45           0,257143
```

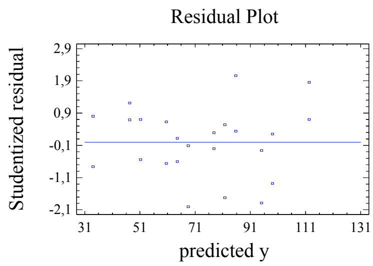
Vi har følgelig at P-værdien er 0.257, altså ogaå en accept.

Spørgsmål 2. Grafisk kontrol af model.

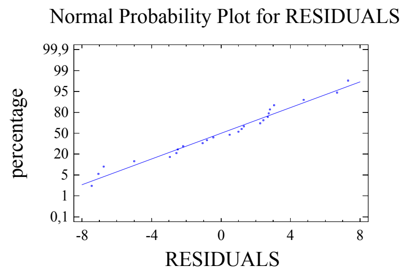
For at få et overblik over punkternes placering tegnes et residualplot.

Dette giver tegningen til højre, som viser, at punkterne ligger tilfældigt omkring linien.

Daingen af de “Studentized residuals ligger over 3 er der <u>ikke outliers</u>



Endelig dannes et normalfordelingsplot



Denne viser, at residualerne er tilnærmelsesvis normalfordelt

<u>Modellen synes at være rimelig</u>.

Spørgsmål 3 Mulig reduktion af modellen

Efter at have set, at en lineær model i 2 variable kan accepteres, går vi tilbage til udskriften med koefficienterne til førstegradsudtrykket.

Af denne ses, at nulhypotesen H_0: β_1 =0 forkastes trestjernet , da P-value ud for x1 er 0.000

og H_0: β_2 =0 forkastes trestjernet , da P-value ud for x2 er 0.000.

Dette kunne også ses hvis man på menubjælken vælger “Tabular options” og derefter menupunktet “Conditional Sums of Squares”

<u>Modellen kan følgelig ikke reduceres yderligere.</u>

Spørgsmål 4 Regresionsligningen

Af ovennævnte udskrift ses umiddelbart, at ligningen er

$$\tilde{Y} = 3.375 + 2.6625x_1 + 1.725x_2$$

Spørgsmål 5 Konfidensintervaller

Vælges “ Confidence intervals” fås en udskrift, der angiver 95% konfidensintervallerne for β_1 og β_2 .

```
 95,0% confidence intervals for coefficient estimates
-----------------------------------------------------------------------------
                                       Standard
Parameter                Estimate         Error     Lower Limit    Upper Limit
-----------------------------------------------------------------------------
CONSTANT                    3,375       2,97685         -2,8157         9,5657
x1                         2,6625      0,210495         2,22475        3,10025
x2                          1,725     0,0768618         1,56516        1,88484
-----------------------------------------------------------------------------
```

Vi har følgelig, at 95% konfidensintervallet for koefficienten β_1 til x_1 er [2.22 ; 3.10] og for koefficienten β_2 til x_2 er [1.56 ; 1.88].

Spørgsmål 6. Find den til $x_1 = 10$ og $x_2 = 20$ svarende estimerede middelværdi $\widetilde{Y}$ og et 95% konfidensinterval for $\widetilde{Y}$.

Der skrives 10 nederst i datafilen i kolonnen for x1 og 20 i kolonnen for x2.
Vælg (Tabular options | Reports | OK). Det giver følgende udskrift:

```
Regression Results for oxid
-----------------------------------------------------------------------------------
    Fitted  Stnd. Error  Lower 95,0% CL  Upper 95,0% CL  Lower 95,0% CL  Upper 95,0% CL
Row  Value   for Forecast  for Forecast   for Forecast       for Mean        for
Mean
-----------------------------------------------------------------------------------
 25  64,5        4,31386         55,5288        73,4712         62,5423
66,4577
-----------------------------------------------------------------------------------
-
```

Vi har følgelig, at $\widetilde{Y} = 64.5$ og et 95% konfidensinterval er [62.54 ; 66.46].

Øvelse 4.4 (multipel regressionsanalyse med gentagelser) Regn opgave 31 side 249.

4.3.2.2 Partiel faktorstruktur med gentagelser

I tilfælde af, at man **ikke** har en fuldstændig faktorstruktur, kan man forudsat at man har gentagelser af målingerne, stadig teste modellen.
Mens de estimerede regressionskoefficienter i modellen med den fuldstændige faktorstruktur er uafhængige af hinanden (bortkastes f.eks koefficienten til x_1 vil estimatet for regressionskoefficienten til x_2 være uændret) er dette ikke tilfældet ved en partiel struktur. Man skal derfor når man eliminerer koefficienter svarede til variable af samme grad, omhyggeligt først tage den, der har den højeste P-value, bortkaste den hvis den f.eks. er over 5%, se hvad der nu sker med de øvrige koefficienter (da de ændrer størrelse) , eventuelt bortkaste den næste koefficient osv. Selv om estimaterne for 2 regressionskoefficienter begge har p-værdier over 5% må de altså ikke bortkastes samtidigt.
Det er klart, at har man mange variable, kan det være besværligt at bortkaste i rigtig rækkefølge. I Statgraphics vil man med cursoren på en tabel, tryk på højre musetast og valg af "Analysis options" kunne få Statgraphics til at hjælpe med dette. Vælges eksempelvis "Backward selection", og F-værdi til 4, vil de variable blive bortkastet af modellen, efter F-værdiens størrelse (mindst F-værdi først) så længe denne F-værdi er under eksempelvis 4.

Eksempel 4.6 (partiel faktorstruktur)
Det formodes, at den producerede mængde Y af en given produktion er en lineær funktion af de anvendte mængder x_1 og x_2 , dvs. på formen $Y = \alpha_0 + \beta_1 x_1 + \beta_2 x_2$.
Følgende observationer foreligger:

(x_1,x_2)	(4,3)	(5,4)	(5.5)	(6,6)	(7,8)	(9,10)
y	33.0 35.7	44.4 46.9	50.0 52.1	57.4 59.1	73.0 75.0	87.1 89.3

1) *Test om man på basis af ovennævnte observationer kan få en accept af, at sammenhængen mellem den producerede mængde Y og temperaturerne x_1 og x_2 (approksimativt) er lineær.*
2) *Kontroller grafisk om modellen er rimelig.*

Det antages i det følgende, at ovenstående model gælder.

3) *Undersøg om modellen kan reduceres, dvs. om $\beta_1 = 0$ og/eller $\beta_2 = 0$.*
4) *Angiv regressionsligningen $\widetilde{Y} = a_0 + b_1 x_1 + b_2 x_2$, hvor a_0, b_1 og b_2 er estimater for α_0, β_1 og β_2*

LØSNING:
Spørgsmål 1: Test om lineær model kan accepteres.
I Statgraphic indtastes tallene på sædvanlig måde.

```
x1   x2   y
4    3    33
4    3    35,7
5    4    44,4
5    4    46,9
5    6    50
5    6    52,1
6    6    57,4
6    6    59,1
7    8    73
7    8    75
9    10   87,1
9    10   89,3
```

Vælg ikonen "Multiple Regression" og indsæt i den fremkomne tavle de variable. Dette medfører følgende udskrift.

```
Multiple Regression Analysis
-----------------------------------------------------------------------------
Dependent variable: y
-----------------------------------------------------------------------------
                                       Standard         T
Parameter               Estimate         Error       Statistic       P-Value
-----------------------------------------------------------------------------
CONSTANT                 1,46935       3,17826       0,462314         0,6548
x1                       5,65323       1,39951        4,03943         0,0029
x2                       3,76129      0,976963        3,84998         0,0039
-----------------------------------------------------------------------------
                           Analysis of Variance
-----------------------------------------------------------------------------
Source             Sum of Squares     Df  Mean Square    F-Ratio       P-Value
-----------------------------------------------------------------------------
Model                     3822,87      2      1911,44     387,61        0,0000
Residual                  44,3823      9      4,93136
-----------------------------------------------------------------------------
Total (Corr.)             3867,26     11

R-squared = 98,8524 percent    R-squared (adjusted for d.f.) = 98,5973 percent
```

Da vi har gentagelser, kan vi teste modellen ved som før at spalte $SAK_{residual\text{-}alle}$ op i $SAK_{gentagelser}=SAK_0$ og en SAK_{lack} hvor den sidste er Kvadratsummen af afstanden fra de enkelte cellers gennemsnit til planen.

Da dataene ikke kan benyttes til en tosidet variansanalyse (for spredte punkter), danner vi i stedet en ekstra søjle behandlinger, og med denne og y foretager en ensidet variansanlyse

x1	x2	y	behandlinger
4	3	33	1
4	3	35,7	1
5	4	44,4	2
5	4	46,9	2
5	6	50	3
5	6	52,1	3
6	6	57,4	4
6	6	59,1	4
7	8	73	5
7	8	75	5
9	10	87,1	6
9	10	89,3	6

Vælg (Compare | Analysis of Variance | One Way ANOVA | Udfyld tavle med y og behandlinger | OK).
Resultatet bliver

```
ANOVA Table for y by Behandlinger

                               Analysis of Variance
-----------------------------------------------------------------------------
Source             Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-----------------------------------------------------------------------------
Between groups            3852,42      5      770,483     311,52       0,0000
Within groups               14,84      6      2,47333
-----------------------------------------------------------------------------
Total (Corr.)             3867,26     11
```

Heraf ses at $SAK_0 = 14.84$ med $f_0 = 6$ (6 celler).
Da $SAK_{residual\text{-}alle} = SAK_0 + SAK_{lack}$, fås $SAK_{lack} = 44.3823 - 14.84 = 29.5423$ med $f_{lack} = 9 - 6 = 3$.

Vi har følgelig at $s^2_{lack} = \dfrac{29.5423}{3} = 9.8474$.

Dermed er $F_{lack} = \dfrac{s^2_{lack}}{s^2_0} = \dfrac{9.8474}{2.47333} = 3.9814$

Dette giver følgende variansanalysetabel:

Variation	SAK	f	s^2	F
Model	3822.87	2		
Lack of fit	29.5423	3	9.8474	3.9814
Gentagelser	14.84	6	2.47333	
Total	4485.1	11		

Da $F_{0.95}(3,6) = 4.76 > F_{lack} = 3.98$, accepteres den lineære model " $Y = \alpha_0 + \beta_1 x_1 + \beta_2 x_2$ ".
I stedet for at slå op i en F-tabel, kan man beregne P-værdien ved at benytte Statgraphics:

Vi får:

```
Cumulative Distribution
-----------------------
Distribution: F (variance ratio)
              Lower Tail Area (<)
Variable      Dist. 1       Dist. 2       Dist. 3       Dist. 4       Dist. 5
3,9874        0,920857
              Probability Density
Variable      Dist. 1       Dist. 2       Dist. 3       Dist. 4       Dist. 5
3,9874        0,03398
              Upper Tail Area (>)
Variable      Dist. 1       Dist. 2       Dist. 3       Dist. 4       Dist. 5
3,9874        0,0791428
```

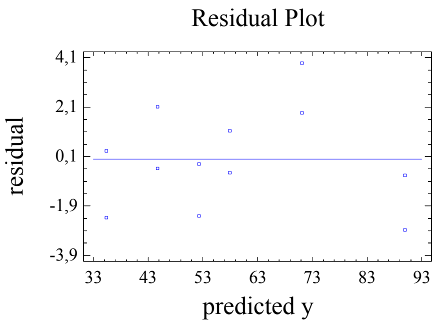
Dette giver en P-værdi på 0.079 > 0.05, og dermed igen en accept af den lineære model

Spørgsmål 2. Grafisk kontrol af model.

For at få et overblik over punkternes placering tegnes et residualplot.

Denne viser, at punkternes "gennemsnit" ligger tilfældigt omkring linien. Der er dog for få punkter til en ordentlig vurdering.

Modellen synes dog at være rimelig.



Spørgsmål 3 Mulig reduktion af modellen

Efter at have set, at en lineær model i 2 variable kan accepteres, går vi tilbage til den første udskrift med koefficienterne til førstegradsudtrykket.

Af denne ses, at nulhypotesen $H_0:\beta_1 = 0$ forkastes (2-stjernet), da *P*-value ud for x1 er 0.0029

og $H_0:\beta_2 = 0$ forkastes (2-stjernet) , da *P*-value ud for x2 er 0.039

Det ses, at vi ikke kan reducere modellen yderligere.

Spørgsmål 4 . Opstilling af regressionsligning

Ligningen bliver $Y = 1.469 + 5.653\, x_1 + 3.761\, x_2$.

Øvelse 4.5 (partiel regressionsanalyse med gentagelser) Regn opgave 32 side 250.

4.3.2.3 Partiel faktorstruktur uden gentagelser

Har man **ingen gentagelser** kan man ikke teste modellen. Man må så som beskrevet i afsnit 4.2.3.3 se på "forklaringsgraden" R^2 eller R-squared = $\frac{SAK_{Total} - SAK_0}{SAK_{Total}} = \frac{SAK_{model}}{SAK_{Total}}$, og på et modificeret R^2 (R-squared (adjusted for d.f)).
Man bør dog i alle tilfælde se på et plot over residualerne. Disse bør ligge "tilfældigt " omkring 0. Hvis dette ikke er tilfældet, bør man være på vagt, og eventuelt gå op til en model af højere grad.

Eksempel 4.7 (faktorstruktur uden gentagelser)
Det månedlige elektriske forbrug Y på en kemisk fabrik formodes at være afhængig af den gennemsnitlige udendørs temperatur x_1, antal arbejdsdage x_2 i måneden , den gennemsnitlige renhed x_3 af det fremstillede produkt og det antal tons x_4, der produceres i den pågældende måned. Det formodes, at Y er en lineær funktion af x_1, x_2, x_3 og x_4, dvs. på formen
$Y = \alpha_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$.
Følgende observationer fra det forløbne år foreligger

x_1	x_2	x_3	x_4	Y
- 4	24	91	100	240
-1	21	90	95	236
7	24	88	110	280
16	25	87	88	274
18	25	91	94	301
23	26	94	99	310
27	25	87	97	300
29	25	86	96	296
24	24	88	110	267
16	25	91	105	276
10	25	90	100	288
3	23	89	98	261

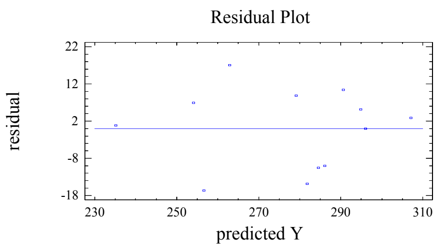
1) *Vurder ud fra forklaringsgraden og grafisk, om ovennævnte model er rimelig.*
Det antages i det følgende, at ovenstående model gælder.
2) *Undersøg om modellen kan reduceres, dvs. om nogle af koefficienterne kan antages at være* 0.
3) *Angiv regressionsligningen i den endelige model.*

Spørgsmål 1. Vurdering af om lineær model gælder.
I Statgraphics indtastes tallene på sædvanlig måde

Da der ikke er gentagelser, kan man ikke teste modellen. I stedet betragtes R-Squared og en tegning af residualerne.
Vælg ikonen "Multiple Regression" og indsæt i den fremkomne tavle de variable. Dette medfører følgende udskrift.

```
Multiple Regression Analysis
-----------------------------------------------------------------------------
Dependent variable: Y
-----------------------------------------------------------------------------
                                       Standard          T
Parameter                Estimate         Error      Statistic        P-Value
-----------------------------------------------------------------------------
CONSTANT                 -31,0531       178,358      -0,174106         0,8667
x1                        1,11251      0,567247        1,96124         0,0907
x2                        8,20832       4,64079        1,76874         0,1203
x3                        1,27901       2,05882       0,621237         0,5541
x4                      -0,212413      0,634421      -0,334814         0,7476
-----------------------------------------------------------------------------
                           Analysis of Variance
-----------------------------------------------------------------------------
Source             Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-----------------------------------------------------------------------------
Model                     4828,86      4      1207,21       6,65       0,0155
Residual                  1270,06      7      181,437
-----------------------------------------------------------------------------
Total (Corr.)             6098,92     11

R-squared = 79,1757 percent    R-squared (adjusted for d.f.) = 67,2761 percent
```

En R-squared på 79.2 % er så høj, at modellen findes rimelig.
Da R-squared (adjusted) på 67.3% er betydelig lavere, må modellen kunne reduceres.



Omstående tegning af residualerne viser, at punkterne synes at placere sig rimeligt tilfældigt omkring 0-linien,
Man kan også se, at der er ingen outliers.
Grafisk må konkluderes, at den lineære model er rimelig.

Spørgsmål 2 Mulig reduktion af modellen
Efter at have set, at en lineær model i 4 variable kan accepteres, går vi tilbage til udskriften med koefficienterne til førstegradsudtrykket.
Den β størrelse, der har størst P-værdi er β_4. Da dens P-værdi er 0.7476 > 0.05 accepteres nulhypotesen H_0: $\beta_4 = 0$.
Vi eliminerer nu x_4 : Vælg (rød ikon = Input dialog | slet x4 i tavle | OK).
Vi får :

```
Multiple Regression Analysis
-----------------------------------------------------------------------------
Dependent variable: Y
-----------------------------------------------------------------------------
                                       Standard          T
Parameter                Estimate         Error      Statistic        P-Value
-----------------------------------------------------------------------------
CONSTANT                  -48,444        160,88      -0,301119         0,7710
x1                        1,10613       0,53454         2,0693         0,0723
x2                        8,28308       4,37061        1,89518         0,0947
x3                        1,21813       1,93362       0,629976         0,5463
-----------------------------------------------------------------------------
```

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	4808,52	3	1602,84	9,94	0,0045
Residual	1290,4	8	161,3		
Total (Corr.)	6098,92	11			

R-squared = 78,8422 percent R-squared (adjusted for d.f.) = 70,908 percent

Det ses i overensstemmelse med det forventede, at R-squared (adjusted) er steget.

Den β størrelse, der nu har størst *P*-værdi er β_3. Da dens *P*-værdi er 0.5463 > 0.05 accepteres nulhypotesen H_0: $\beta_3=0$.
Vi eliminerer nu x_3 og får

Multiple Regression Analysis

Dependent variable: Y

Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	34,5475	89,1967	0,387319	0,7075
x1	0,945019	0,453396	2,08431	0,0668
x2	9,43721	3,83285	2,46219	0,0360

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	4744,5	2	2372,25	15,76	0,0011
Residual	1354,41	9	150,49		
Total (Corr.)	6098,92	11			

R-squared = 77,7926 percent R-squared (adjusted for d.f.) = 72,8576 percent

Det ses i overensstemmelse med det forventede, at R-squared (adjusted) igen er steget.

Den β størrelse, der nu har størst *P*-værdi er β_1. Da dens *P*-værdi er 0.0668 > 0.05 accepteres nulhypotesen H_0: $\beta_1 = 0$.
Imidlertid er vi nær forkastelse, så det kan være tvivlsomt om det er fornuftigt at se bort fra x_1.

Ellimineres x_1 fås

Multiple Regression Analysis

Dependent variable: Y

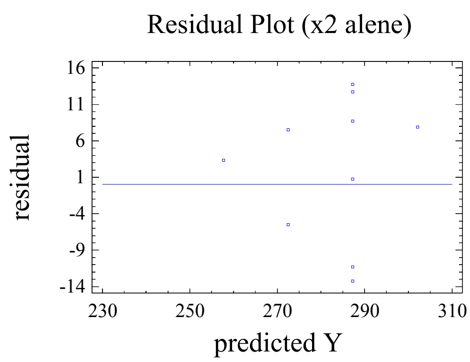
Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	-82,8036	79,9173	-1,03612	0,3246
x2	14,8036	3,27997	4,51333	0,0011

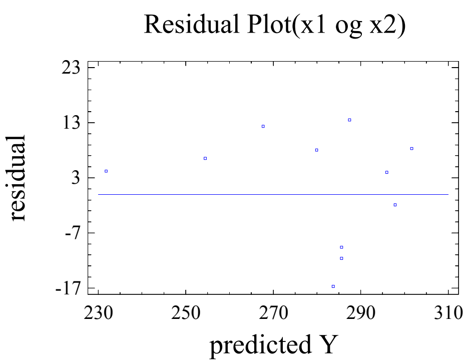
Analysis of Variance

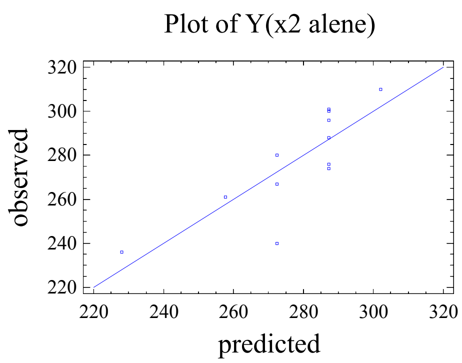
Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	4090,72	1	4090,72	20,37	0,0011
Residual	2008,2	10	200,82		
Total (Corr.)	6098,92	11			

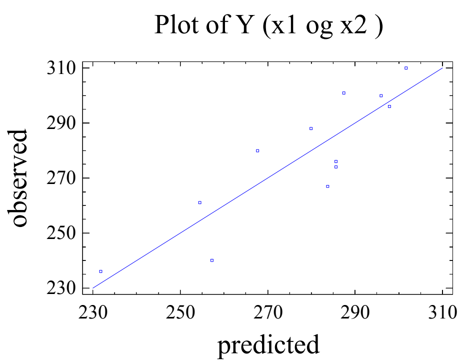
R-squared = 67,0729 percent R-squared (adjusted for d.f.) = 63,7802 percent

Denne gang er R-squared (adjusted) faldet (endda temmelig meget), hvilket kunne tyde på, at det næppe er helt fornuftigt at bortkaste x_1.
For at få et indtryk af dette tegnes ikke alene residualerne for de to tilfælde, men også de observerede værdier mod de beregnede fra modellen.









Som det ses, er det vanskeligt ud fra tegningerne at skelne mellem modellerne. Muligvis punkterne placeret lidt mere jævnt om linierne, hvis også x_1 indgår i modellen, men særlig tydeligt er det ikke.

Vi vælger derfor den simpleste model, dvs modellen med kun x_2.
Det ses, at vi ikke kan reducere modellen yderligere.

Spørgsmål 3 . Opstilling af regressionsligning

Ligningen bliver $\widetilde{Y} = -82.80 + 14.80 \cdot x_2$

Øvelse 4.6 (partiel regressionsanalyse uden gentagelser) Regn opgave 33 side 250.

4.3.3 Polynomiale modeller

4.3.3.1 Indledning

Lad der eksempelvis være givet et fuldstændigt randomiseret forsøg med følgende faktor-struktur
($Y =$ udbyttet Y af en kemisk proces,
Faktor A = Reaktionstemperatur x_1 ,Faktor B: Mængde x_2 af katalysatormasse).

		A: (temperatur x_1)		
		80	100	120
B (mængde x_2)	25	6.3 2.9	12.8 10.1	19.0 23.2
	75	5.1 6.6	8.7 14.2	39.6 39.0

Da A er på 3 niveauer, kan x_1 i modellen højst forekomme i anden potens, og da B har 2 niveauer kan x_2 højst forekomme i første potens.
En fuldstændig model af middelværdien af Y kan beskrives som et polynomium af højst 3 grad i x_1 og x_2: $Y = \alpha_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_1 x_2 + \beta_5 x_1^2 x_2$.
Bemærk, at da vi har 6 celler (gennemsnit), må antallet af parametre $\alpha_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5$ være 6.
Som det ses, er de i afsnit 5.3.2 betragtede modeller et specialtilfælde, og den statistiske analyse er da også meget beslægtet hermed.
En polynomial regressionsanalyse anvendes ofte, hvor man ikke kender sammenhængen mellem Y og de uafhængige variable, men er interesseret i at bestemme et polynomium (af lavest mulig grad), der giver en tilstrækkelig god beskrivelse af Y indenfor det foreliggende variationsområde.

4.3.3.2. Polynomial model med gentagelser

Eksempel 4.8 (3 × 4 faktorforsøg med 2 kvantitative faktorer) (SFIII eksempel 12 side 68)

Ved et fuldstændigt randomiseret forsøg med to gentagelser af hver behandling undersøgtes indflydelsen af to faktorer A (på 3 *niveauer) og B (med* 4 *niveauer) på rensningsgraden af en rensningsproces. Faktor A' s niveauer betegner* 3 *forskellige tilsætninger (*2, 4 *og* 6 *g/l) af et stof* I *og faktor B`s niveauer* 4 *forskellige tilsætninger (*6, 10, 14 *og* 18 *g/l) af et stof* II.
Forsøgsresultaterne y (% urenhed) fremgår af følgende skema:

		B: (tilsætning x_2 af stof II)			
		6	10	14	18
A (tilsætning x_1 af stof I)	2	7 8	6 4	4 3	4 6
	4	8 4	5 4	3 2	7 4
	6	8 7	4 6	3 4	6 3

Det formodes på forhånd, at middelværdien af Y kan beskrives som et polynomium af højst 2 *grad i* x_1 *og* x_2: $Y = \alpha_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2$.

1) *Vurdér om ovennævnte* 2.*gradsmodel kan antages at kunne beskrive sammenhængen mellem Y og de* 2 *uafhængige variable.*

I det følgende antages ovennævnte model at gælde.

2) *Test om en lineær model* $Y = \alpha_0 + \beta_1 x_1 + \beta_2 x_2$ *kan antages at kunne beskrive sammenhængen mellem Y og de* 2 *uafhængige variable.*
3) *Reducer om muligt modellen, og angiv regressionsligningen i den endelige model.*
4) *Bestem et estimat og et* 95% *konfidensinterval for middelurenhedsprocenten ved tilsætning af* 5 *g/l af tilsætning A og* 16 *g/l af tilsætning B*

LØSNING:

Spørgsmål 1. Vurdering af om andengradsmodel kan accepteres.

Indtastning af data

Tallene indtastes på sædvanlig måde

For at få $\text{SAK}_{\text{gentagelser}}$ kan vi udføre en tosidet variansanalyse med vekselvirkning (den mest omfattende model).

En anden mulighed ville have været, at indføre alle de højere led i modellen. Da x_1 har 3 niveauer, kan der højst optræde led indeholdende x_1 af 2-grad. Analogt kan der højst optræde led indeholdende x_2 af 3-grad. Vi får derfor

$$Y = \alpha_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2 + \beta_6 x_1^2 x_2 + \beta_7 x_1 x_2^2 + \beta_8 x_2^2 + \beta_9 x_1 x_2^3 + \beta_{10} x_1^2 x_2^2 + \beta_{11} x_1^2 x_2^3$$

Denne sidste mulighed er så omfattende, at det i denne situation er lettest, at udføre den tosidede variansanalyse.

Vi vælger Compare , Multifactor ANOVA og sætter “Maximum Order Interaction til 2".

Resultatet bliver:

```
Analysis of Variance for y - Type III Sums of Squares
--------------------------------------------------------------------------------
Source                 Sum of Squares     Df    Mean Square    F-Ratio    P-Value
--------------------------------------------------------------------------------
MAIN EFFECTS
 A:x1                            1,75      2          0,875       0,40     0,6765
 B:x2                         44,3333      3        14,7778       6,82     0,0062
INTERACTIONS
 AB                           3,91667      6       0,652778       0,30     0,9244
RESIDUAL                         26,0     12        2,16667
--------------------------------------------------------------------------------
TOTAL (CORRECTED)                76,0     23
--------------------------------------------------------------------------------
```

Heraf ses, at $\text{SAK}_0 = 26{,}0$, $s_0^2 = 2.1667$ med $f = 12$ (12 celler).

Vælg ikonen "Multiple Regression" eller Vælg (Relate | Multiple Regression).
I den fremkomne tavle under “Indenpendent Variable” indsættes x1, x2, x1^2, x2^2, x1*x2 og der fremkommer følgende tabel:

```
Multiple Regression Analysis
-----------------------------------------------------------------------------
Dependent variable: y
-----------------------------------------------------------------------------
                                      Standard          T
Parameter                Estimate       Error       Statistic        P-Value
-----------------------------------------------------------------------------
CONSTANT                     16,6      3,54966        4,67651         0,0002
x1                       -1,04375      1,25925      -0,828867         0,4180
x2                       -1,65417     0,442297       -3,73994         0,0015
x1^2                     0,140625      0,14547       0,966693         0,3465
x2^2                       0,0625    0,0171438        3,64563         0,0018
x1*x2                   -0,009375    0,0375602      -0,249599         0,8057
-----------------------------------------------------------------------------
                           Analysis of Variance
-----------------------------------------------------------------------------
Source             Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-----------------------------------------------------------------------------
Model                     43,4958      5      8,69917       4,82       0,0057
Residual                  32,5042     18      1,80579
-----------------------------------------------------------------------------
Total (Corr.)                76,0     23
```

Da $SAK_{residual\text{-}alle} = SAK_{gentagelser} + SAK_{lack}$ fås $SAK_{lack} = 32.50 - 26.0 = 6.50$
med $f_{lack} = 18 - 12 = 6$.

Vi har følgelig at $s_{lack}^2 = \frac{SAK_{lack}}{f_{lack}} = 1.08$ og dermed $F_{lack} = \frac{s_{lack}^2}{s_0^2} = 0.50$

Dette giver følgende variansanalysetabel:

Variation	SAK	f	s^2	F
Model	43.4958	5	8.69917	
Lack of fit	6.5042	6	1.08	0.50
Gentagelser	26.0	12	2.167	
Total	76.0	23		

Da $F_{lack} < 1$ accepteres den valgte andengradsmodel

Spørgsmål 2. Undersøg om modellen kan reduceres til en førstegradsmodel.

Efter at have set, at en modellen er god nok, tester vi om en lineær model er god nok, dvs om vi kan bortkaste alle andengradsled.

Vælg (rød ikon = Input dialog | slet de tre andengradsled | OK). Vi får:

```
Multiple Regression Analysis
-----------------------------------------------------------------------------
Dependent variable: renhed
-----------------------------------------------------------------------------
                                      Standard          T
Parameter                Estimate       Error       Statistic        P-Value
-----------------------------------------------------------------------------
CONSTANT                    7,425      1,28168        5,79319         0,0000
x1                       -0,03125     0,208281      -0,150038         0,8822
x2                      -0,191667    0,0760536       -2,52015         0,0199
```

```
-----------------------------------------------------------------------------
                           Analysis of Variance
-----------------------------------------------------------------------------
Source             Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-----------------------------------------------------------------------------
Model                     17,6958      2      8,84792       3,19       0,0618
Residual                  58,3042     21      2,77639
-----------------------------------------------------------------------------
Total (Corr.)                76,0     23
```

Vi ser at residualen er vokset til 58.3042 med 21 frihedsgrader. De tre bortkastede led har følgelig en samlet SAK på 58.3042 - 32.5042 = 25.80 med 21 - 18 = 3 frihedsgrader.

Vi får følgelig, at $s^2_{2.gradsled} = \frac{23.80}{3} = 8.60$ og dermed

$$F_{2.gradsled} = \frac{s^2_{2.gradsled}}{s^2_{residual-gammel}} = \frac{8.60}{1.81} = 4.75.$$

Da $F_{0.95}(3,18) = 3.16$ <u>forkastes 1. gradsmodellen.</u>

Benyttes Statgraphics fås P-værdi 0.013 < 0.05

```
Cumulative Distribution
-----------------------
              Upper Tail Area (>)
Variable      Dist. 1       Dist. 2       Dist. 3       Dist. 4       Dist. 5
4,75          0,0129841
```

Spørgsmål 3.Mulig reduktion af modellen

Vi tester nu de enkelte andengradsled.

Af tabellen på foregående side ses, at produktleddet kan bortkastes. Efter pooling opdages, at også x_1^2 leddet kan bortkastes. Vi får nu følgende tabel:

```
Multiple Regression Analysis
-----------------------------------------------------------------------------
Dependent variable: y
-----------------------------------------------------------------------------
                                          Standard          T
Parameter                  Estimate         Error       Statistic      P-Value
-----------------------------------------------------------------------------
CONSTANT                     15,175       2,30376         6,58705       0,0000
x1                         -0,03125      0,163707       -0,190889       0,8505
x2                         -1,69167      0,405431        -4,17252       0,0005
x2^2                         0,0625     0,0167083         3,74065       0,0013
-----------------------------------------------------------------------------
                           Analysis of Variance
-----------------------------------------------------------------------------
Source             Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-----------------------------------------------------------------------------
Model                     41,6958      3      13,8986       8,10       0,0010
Residual                  34,3042     20      1,71521
-----------------------------------------------------------------------------
Total (Corr.)                76,0     23
```

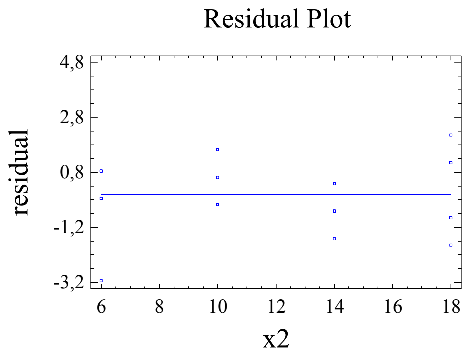
Vi kan af denne udskrift se, at også x_1 kan bortkastes.

```
Multiple Regression Analysis
-----------------------------------------------------------------------------
Dependent variable: y
-----------------------------------------------------------------------------
                                          Standard          T
Parameter                  Estimate         Error       Statistic      P-Value
-----------------------------------------------------------------------------
CONSTANT                      15,05       2,15747         6,97577       0,0000
x2                         -1,69167       0,39602        -4,27167       0,0003
x2^2                         0,0625     0,0163205         3,82954       0,0010
```

		Analysis of Variance			
Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	41,6333	2	20,8167	12,72	0,0002
Residual	34,3667	21	1,63651		
Total (Corr.)	76,0	23			

Vi kan ikke reducere modellen yderligere.

Tegnes som sædvanlig residualerne fås tegningen til højre, som intet foruroligende viser.



Vi ender følgelig i en andengradsmodel i variablen x_2.

Modellen bliver $\hat{Y} = 15.05 - 1.692 \cdot x_2 + 0.0625 \cdot x_2^2$

Spørgsmål 4. Estimat og 95% konfidensinterval.

Et skøn for middelurenhedsprocenten ved tilsætning af 5 g/l af tilsætning A og 16 g/l af tilsætning B, samt et 95% konfidensinterval herfor, fås ved at indsætte i data i sidste (25) række tallet 16 i x2 kolonne.
Vælg (gul ikon = Tabular Options | Reports | OK). Det giver følgende udskrift:

Regression Results for renhed

Row	Fitted Value	Stnd. Error for Forecast	Lower 95,0% CL for Forecast	Upper 95,0% CL for Forecast	Lower 95,0% CL for Mean	Upper 95,0% CL for Mean
25	3,98333	1,32797	1,22166	6,74501	3,24222	4,72445

Y = 3.98 og konfidensinterval [3.24; 4.72].

Øvelse 4.7 (Polynomial regressionsanalyse) Regn opgave 35 side 251.

5. MODELLER MED BÅDE KVANTITATIVE OG KVALITATIVE FAKTORER

5.1 INDLEDNING

Er alle faktorer kvalitative eller er de alle kvantitative, så benyttes de metoder og programmer (Compare, Relate osv) som er gennemgået i de tidligere kapitler. Har vi både kvalitative og kvantitative faktorer kan man naturligvis godt foretage beregningerne og analyserne som om alt var kvalitativt, men det er her mere effektivt at udnytte, at nogle af faktorerne er kvantitative.
Skal man undersøge mange faktorer, så vil man dog uanset om faktorerne er kvantitative eller kvalitative foretage et indledende 2^k screeningsforsøg som beskrevet i kapitel 2. Derved får man reduceret faktorerne til de der synes at have betydning. Man kan så efterfølgende analysere forsøget eller eventuelt lave et nyt forsøg under hensyntagen til, at nogle af de faktorer, der har virkning, er kvantitative. Vi vil derfor først give et eksempel på, hvorledes man analyserer et sådant 2^k - faktorforsøg, og derefter i afsnit 5.3 se på forsøg med faktorer på mere end 2 niveauer.

5.2. 2^k FAKTORFORSØG MED MINDST 1 KVANTITATIV FAKTOR

Efter at have foretaget et indledende screeningsforsøg som beskrevet i kapitel 2, har man fået reduceret faktorerne til de der synes at have betydning. Hvis nogle af de betydende faktorer er kvantitative, kan man så tage hensyn til det ved den videre analyse. Da hver kvantitativ faktor ved et 2^k faktorforsøg kun har 2 niveauer, bør sådanne 2^k faktorforsøg med kvantitative faktorer kun udføres, såfremt man på forhånd kan antage, at sammenhængen mellem forsøgsvariablen og hver af de kvantitative variable er lineær. Er alle faktorer kvantitative, adskiller analysen sig ellers ikke fra den der er gennemgået i kapitel 5.
Hvis vi har en blanding af kvantitative og kvalitative faktorer, bliver analysen og fortolkningerne anderledes, som det følgende eksempel viser. Da eksemplet udover de to kvantitative faktorer også en kvalitativ faktor på 2 niveauer og en (kvalitativ) blokfaktor på 5 niveauer er det velegnet til at forklare, hvorledes man i Statgraphics ved hjælp af såkaldte "dummy variable" (også kaldet "indikator variable") håndterer en blanding af kvantitative og kvalitative faktorer.

Eksempel 5.1 (2 kvantitative og en kvalitativ faktor) (hentet fra SFIII eksempel 3 side 25).
Der foreligger følgende fuldstændigt 2^3 faktorforsøg, hvor man undersøger om reaktionstemperatur T (faktor A), trykket P(faktor B) og katalysatortypen C påvirker udbyttet Y af et bestemt kemikalium. Der blev benyttet 5 apparater som kunne tænkes at have en blokvirkning.
Det vides, at udbyttet har formen $Y = \alpha_{ij} + \beta_{1i}T + \beta_{2i}P + \beta_{3i}T \cdot P$, hvor $i = 1,2$ (*katalysator*) og $j = 1,2,3,4,5$ (*apparater*).
Konstantleddet og koefficienterne kunne altså være afhængig af hvilket niveau katalysatoren C har, mens blokfaktoren (apparater) kun kan bevirke en "parallelforskydning" af ligningerne, og derfor kun påvirker konstantleddet.

Forsøgsresultaterne var:

	Faktorer			Udbytte i %				
	Temperatur T °	Tryk P i atm.	Katalysatortype	App 1	App 2	App 3	App.4	App.5
(1)	50	1	C_1	4.1	4.7	5.0	4.6	5.6
a	100	1	C_1	11.4	11.9	11.3	10.5	11.1
b	50	2	C_1	1.2	0.8	1.8	0.0	1.5
ab	100	2	C_1	8.1	8.7	8.4	9.12	9.3
c	50	1	C_2	11.0	10.1	11.2	10.7	9.3
ac	100	1	C_2	13.7	13.2	14.5	12.0	13.5
bc	50	2	C_2	7.9	8.1	7.8	7.1	6.9
abc	100	2	C_2	10.7	11.0	10.9	10.7	9.9

1) *Reducer modellen mest muligt*
2) *Kontroller grafisk den fremkomne model.*
3) *Anfør de estimerede regressionsligninger.*

LØSNING:
Spørgsmål 1. Reduktion af model.
Indtastning af data Tallene indtastes på sædvanlig måde.

```
temp  tryk  katalys  apparat  udbytte
50    1        k1       a1       4,1
50    1        k1       a2       4,7
50    1        k1       a3       5
50    1        k1       a4       4,6
50    1        k1       a5       5,6
100   1        k1       a1       11,4
100   1        k1       a2       11,9
100   1        k1       a3       11,3
100   1        k1       a4       10,5
100   1        k1       a5       11,1
50    2        k1       a1       1,2
50    2        k1       a2       0,8
50    2        k1       a3       1,8
50    2        k1       a4       0
50    2        k1       a5       1,5
100   2        k1       a1       8,1
100   2        k1       a2       8,7
100   2        k1       a3       8,4
100   2        k1       a4       9,1
100   2        k1       a5       9,3
50    1        k2       a1       11
50    1        k2       a2       10,1
50    1        k2       a3       11,2
50    1        k2       a4       10,7
50    1        k2       a5       9,3
100   1        k2       a1       13,7
100   1        k2       a2       13,2
100   1        k2       a3       14,5
100   1        k2       a4       12
100   1        k2       a5       13,5
50    2        k2       a1       7,9
50    2        k2       a2       8,1
50    2        k2       a3       7,8
50    2        k2       a4       7,1
50    2        k2       a5       6,9
100   2        k2       a1       10,7
100   2        k2       a2       11
100   2        k2       a3       10,9
100   2        k2       a4       10,7
100   2        k2       a5       9,9
```

Opstilling af model.
Vælg ikonen "GLM" eller vælg (Special | Advanced Regression | General Linear Models")
På det følgende skema sættes "Dependent variables" : udbytte
"Categorical Factors" : apparat og katalys
og "Quantitative Factors" : temp og tryk
Efter tryk på "OK" fremkommer et skema "GLM-specifications" ,hvor A= apparat, B = katalys osv. (Rækkefølgen i input bestemmer bogstavordenen).
Vi udfylder nu samtlige to og trefaktorvekselvirkninger der indbefatter katalys, temp og tryk (men ikke blokfaktoren apparat), dvs B*C, B*D, C*D, B*C*D.
Der fremkommer en udskrift med type III "Sum of Squares".

```
General Linear Models
---------------------
Number of categorical factors: 2         Number of quantitative factors: 2

Analysis of Variance for udbytte
```

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	551,116	11	50,1014	137,78	0,0000
Residual	10,182	28	0,363643		
Total (Corr.)	561,298	39			

Type III Sums of Squares

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
apparat	2,534	4	0,6335	1,74	0,1688
katalys	5,29984	1	5,29984	14,57	0,0007
temp	15,9201	1	15,9201	43,78	0,0000
tryk	16,0801	1	16,0801	44,22	0,0000
katalys*temp	1,5129	1	1,5129	4,16	0,0509
katalys*tryk	0,9025	1	0,9025	2,48	0,1264
temp*tryk	1,19025	1	1,19025	3,27	0,0812
katalys*temp*tryk	0,70225	1	0,70225	1,93	0,1756
Residual	10,182	28	0,363643		
Total (corrected)	561,298	39			

```
All F-ratios are based on the residual mean square error.
R-Squared = 98,186 percent       R-Squared (adjusted for d.f.) = 97,4733 percent
```

Da blokfaktoren apparat ikke antages at vekselvirke (bekræftes af at R-Squared (adjusted for d.f.) = 97,4733 percent) er modellen den mest omfattende, og følgelig er $s_0^2 = 0.3636$.

Type I tabel. Som omtalt i kapitel 1 beregner Type III hvor stor en ændring i SAK det giver, hvis det pågældende led var det sidst indførte i modellen. Dette er også i denne situation den mest relevante. Ønskaes den "sædvanlige" variansanalysetabel, hvor faktorerne indføres gradvist i modellen i den indtastede rækkefølge, skal man vælge type I. Forudsat faktorerne er indført i rækkefølge med hovedfaktorer, 2-faktorvekselvirkninger og 3-faktorvekselvirkninger giver Type I nedenstående tabel:

```
General Linear Models
---------------------
Analysis of Variance for udbytte
```

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	551,116	11	50,1014	137,78	0,0000
Residual	10,182	28	0,363643		
Total (Corr.)	561,298	39			

Type I Sums of Squares

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
apparat	2,534	4	0,6335	1,74	0,1688
katalys	164,43	1	164,43	452,18	0,0000
temp	252,506	1	252,506	694,38	0,0000
tryk	88,5062	1	88,5062	243,39	0,0000
katalys*temp	41,0063	1	41,0063	112,77	0,0000
katalys*tryk	0,24025	1	0,24025	0,66	0,4232
temp*tryk	1,19025	1	1,19025	3,27	0,0812
katalys*temp*tryk	0,70225	1	0,70225	1,93	0,1756
Residual	10,182	28	0,363643		
Total (corrected)	561,298	39			

Mulig reduktion af modellen

Da P-value for temp*tryk* katalys er 0.1756 > 0.05 accepteres nulhypotesen
H_0: Koefficienten til temp*tryk er uafhængig af katalysatortype
Herefter pooles dette led ned. Vælg (rød ikon = Input dialog |OK | fjern B*C*D | OK). Vi får

```
General Linear Models
---------------------
Analysis of Variance for udbytte
```

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	550,414	10	55,0414	146,65	0,0000
Residual	10,8843	29	0,375319		
Total (Corr.)	561,298	39			

Type III Sums of Squares

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
apparat	2,534	4	0,6335	1,69	0,1797
katalys	49,1643	1	49,1643	130,99	0,0000
temp	15,9201	1	15,9201	42,42	0,0000
tryk	16,0801	1	16,0801	42,84	0,0000
temp*tryk	1,19025	1	1,19025	3,17	0,0854
katalys*temp	41,0063	1	41,0063	109,26	0,0000
katalys*tryk	0,24025	1	0,24025	0,64	0,4302
Residual	10,8843	29	0,375319		
Total (corrected)	561,298	39			

All F-ratios are based on the residual mean square error.

Det ses,at vi næppe har nogen blokvirkning.
Da P-value for katalys*tryk er 0.4302 > 0.05 accepteres nulhypotesen
H_0: Koefficienten til tryk er uafhængig af katalysatortype.
Da P-value for temp*tryk er 0.0854 > 0.05 accepteres nulhypotesen
H_0: Koefficienten til temp*tryk er 0.

Nu pooles disse 2 led ned, hvorved vi får følgende udskrift:

```
General Linear Models
---------------------
Number of dependent variables: 1      Number of categorical factors: 2
Number of quantitative factors: 2
Analysis of Variance for udbytte
```

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	548,983	8	68,6229	172,74	0,0000
Residual	12,3148	31	0,39725		
Total (Corr.)	561,298	39			

Type III Sums of Squares

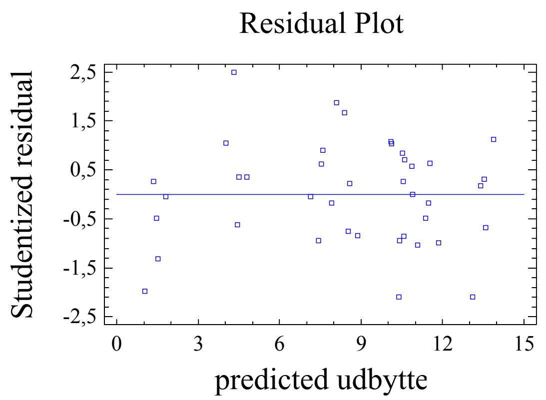
Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
apparat	2,534	4	0,6335	1,59	0,2005
katalys	102,617	1	102,617	258,32	0,0000
temp	252,506	1	252,506	635,64	0,0000
tryk	88,5062	1	88,5062	222,80	0,0000
katalys*temp	41,0063	1	41,0063	103,23	0,0000
Residual	12,3148	31	0,39725		
Total (corrected)	561,298	39			

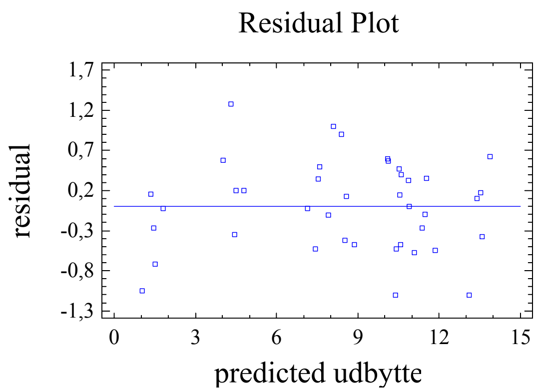
All F-ratios are based on the residual mean square error.

Vi ser, at der ikke kan bortkastes flere led.
Modellen er derfor $Y = \alpha_i + \beta_{1i} T + \beta_2 P$, hvor $i = 1,2$ (katalysator).

<u>Konstantleddet og koefficienten til temperaturen er afhængig af katalysatortypen, hvorimod koefficienten til trykket ikke er afhængig af katalysatortypen.</u>

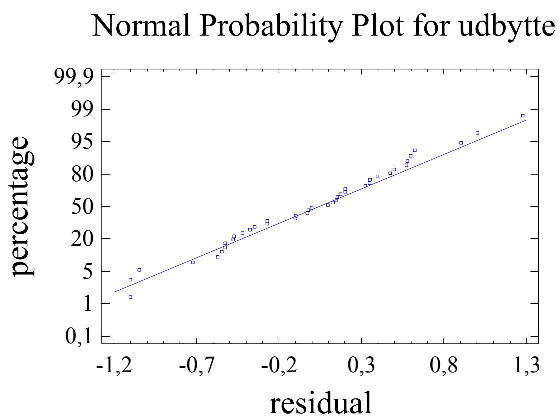
Spørgsmål 2. Kontroller grafisk modellen.
Residualplot: (Blå ikon =Graphical options |Residual plots |OK).
Vi får hermed “Studentized Residuals”
Ønskes de sædvanlige residualer:
(Cursor på figur | højre musetast |Pane Options |Vælg Residuals, Scatter plot, Predicted value).





I residualplottet synes punkterne at fordele sig rimeligt tilfældigt omkring linien, og der er ingen outliers da studentized residuals < 3 (eller $3 \cdot \sigma = 3\sqrt{0.39725} = 1.9 > 1.7$).
Residualerne synes også normalfordelte, men det ses bedre ved hjælp af et normalfordelingsplot
.

Normalfordelingsplot:(Cursor på figur | højre musetast |Normal Probability Plot |OK)



Spørgsmål 3. Opskriv regressionsligninger.
Vælg (gul ikon = Tabular Options | Model Coefficients | OK). Derved fremkommer følgende udskrift:

95,0% confidence intervals for coefficient estimates (udbytte)

Parameter	Estimate	Standard Error	Lower Limit	Upper Limit	V.I.F[1]
CONSTANT	5,4075	0,434389	4,52156	6,29344	
apparat	0,03	0,199311	-0,376499	0,436499	1,6
apparat	0,08	0,199311	-0,326499	0,486499	1,6
apparat	0,38	0,199311	-0,0264989	0,786499	1,6
apparat	-0,395	0,199311	-0,801499	0,0114989	1,6
katalys	-5,065	0,315139	-5,70773	-4,42227	10,0
temp	0,1005	0,00398623	0,09237	0,10863	1,0
tryk	-2,975	0,199311	-3,3815	-2,5685	1,0
katalys*temp	0,0405	0,00398623	0,03237	0,04863	10,0

The StatAdvisor
Included are variance inflation factors, which can be used to measure the extent to which the predictor variables are correlated amongst themselves. VIF's above 10, of which there are 0, are usually considered to indicate serious multicollinearity. Serious multicollinearity great-ly increases the estimation error of the model coefficients as compared with an orthogonal sample.

Forklaring af udskrift og dummy variable
Apparaterne spiller ingen rolle, men for at kunne forklare udskriften antager vi, at de havde betydning.
Modellen skrives som man bør gøre, når man skal tolke Statgraphics udskrift:
Y = (constant + apparat + katalys) +(const+katalys) * T + konst * P (1)
hvor den første parentes er konstantleddet, som altså afhænger af både katalysator (2 muligheder) og apparat (5 muligheder). Den næste parentes viser, at koefficienten til T afhænger af hvilken katalysator der er valgt (2 muligheder).

Ved beregningerne indfører GLM såkaldte "dummy variable", for hver "Catagorical Factors". Princippet blev forklaret i kapitel 3 i forbindelse med eksempel 3.8, men da det er vigtigt at forstå for at kunne tolke udskrifterne gentages det her.
Til karakterisering af de 2 katalysatorer indføres en variabel v_k, som har værdien +1 for katalysator 1 og værdien -1 for katalysator 2.
Til karakterisering af de 5 apparater indføres 4 variable v_{a1}, v_{a2}, v_{a3}, v_{a4}.
Apparat 1 karakteriseres ved, at v_{a1} er 1 og de øvrige nul, apparat 2 karakteriseret ved, at v_{a2} er 1 og de øvrige nul, osv, indtil vi når til det sidste apparat, hvor alle v_a'erne er -1.
Fordelen ved denne fremgangsmåde er, at hver dummy variabel får gennemsnittet 0 (over apparaterne).

[1]) VIF=Variance Inflation Factor er en tal der viser, om der er afhængighed mellem de uafhængige variable $x_1, x_2, .. x_n$. Jo stærkere lineær afhængig af x_j er af de øvrige variable, jo større er VIF(β_j). Er tallet større end 10 (se også "The StatAdvisor"), så er der tale om så stor afhængighed, og man må medtage det i sine overvejelser.

Værdierne fremgår af følgende skema:

katalys	apparat	v_k	v_{a1}	v_{a2}	v_{a3}	v_{a4}
k1	a1	1	1	0	0	0
k1	a2	1	0	1	0	0
k1	a3	1	0	0	1	0
k1	a4	1	0	0	0	1
k1	a5	1	-1	-1	-1	-1
k2	a1	-1	1	0	0	0
k2	a2	-1	0	1	0	0
k2	a3	-1	0	0	1	0
k2	a4	-1	0	0	0	1
k2	a5	-1	-1	-1	-1	-1

I Statgraphics interne formulering af modellen bliver den givne model

Y = (constant + apparat + katalys) +(const+katalys) * T + konst * P **(1)**

udtrykt ved "dummy variable" skrevet:

$$Y = k_0 + (k_k \cdot v_k + k_{a1} \cdot v_{a1} + k_{a2} \cdot v_{a2} + k_{a3} \cdot v_{a3} + k_{a4} \cdot v_{a4}) + (k_T + k_{kT} \cdot v_k) \cdot T + k_p \cdot P, \qquad (2)$$

hvor k-værdierne er modelkoefficienterne:

```
Parameter                    Estimate
-------------------------------------
CONSTANT     = k0 =          5,4075
apparat      = ka1 =          0,03
apparat      = ka2 =          0,08
apparat      = ka3 =          0,38
apparat      = ka4 =         -0,395
katalys      = kk =          -5,065
temp         = kT =           0,1005
tryk         = kP =          -2,975
katalys*temp = kkT =          0,0405
```

Ved at opskrive de 10 regressionsligninger forstås symbolikken i såvel ligning (1) som (2):

apparat 1 og katalysator 1: $\tilde{Y} = 5.4075 + 0.03 - 5.065 + (0.1005 + 0.0405)T - 2.975P$

apparat 2 og katalysator 1: $\tilde{Y} = 5.4075 + 0.08 - 5.065 + (0.1005 + 0.0405)T - 2.975P$

apparat 3 og katalysator 1: $\tilde{Y} = 5.4075 + 0.38 - 5.065 + (0.1005 + 0.0405)T - 2.975P$

apparat 4 og katalysator 1: $\tilde{Y} = 5.4075 - 0.395 - 5.065 + (0.1005 + 0.0405)T - 2.975P$

apparat 5 og katalysator 1: $\tilde{Y} = 5.4075 - 0.03 - 0.08 - 0.38 + 0.395 - 5.065 + (0.1005 + 0.0405)T - 2.975P$

apparat 1 og katalysator 2: $\tilde{Y} = 5.4075 + 0.03 + 5.065 + (0.1005 - 0.0405)T - 2.975P$

apparat 2 og katalysator 2: $\tilde{Y} = 5.4075 + 0.08 + 5.065 + (0.1005 - 0.0405)T - 2.975P$

apparat 3 og katalysator 2: $\tilde{Y} = 5.4075 + 0.38 + 5.065 + (0.1005 - 0.0405)T - 2.975P$

apparat 4 og katalysator 2: $\tilde{Y} = 5.4075 - 0.395 + 5.065 + (0.1005 - 0.0405)T - 2.975P$

apparat 5 og katalysator 2: $\tilde{Y} = 5.4075 - 0.03 - 0.08 - 0.38 + 0.395 + 5.065 + (0.1005 + 0.0405)T - 2.975P$

Bemærk: Ved apparat 5 indføres alle de forrige 4 tal med modsat fortegn

Rækkefølgen ved tolkning af udskrift.
Hvis man i inddata giver niveauerne navne (a1,a2 ..., og ikke kun tal 1,2 ...), skal variablen være af typen "Character" . I Statgraphics bestemmes rækkefølgen så af den ASCII værdi som navnet får. Dette bevirker, at rækkefølgen bliver alfabetisk. Gives apparaterne 1,2,3,4,5 eksempelvis navnene Peter, Arne, Anne, Joergen, Karsten, skal udskriften tolkes i den alfabetiske rækkefølge Anne, Arne, Joergen, Karsten, Peter. Er nogle af niveauerne karakteriseret ved tal, kan det imidlertid let give anledning til fejltagelser. Tallene -6, -3, 0,+1, 4 får rækkefølgen +1, -3,-6, 0,4, mens -6,-3,0,1,4 bliver til -3, -6,0,1,4. For at undgå sådanne fejlmuligheder bør man, hvis alle niveauerne er talstørrelser erklære variablen "numeric", hvorved man får den sædvanlige talmæssige rækkefølge.
Da det er praktisk at man konsekvent vælger at lavt niveau har den laveste interne værdi (-1) og højt niveau har den højeste værdi (+1), så kan det anbefales at hvis man ønsker at de variable skal være af typen "Character" så vælger karakterværdierne *h* (for høj) og *l* for lav. Eksempelvis hvis vi (af uforklarlige grunde) følte at katalysator k1 var den lave værdi, så burde vi have sat k1 til *l* og k2 til *h*.

Opskrivning af de fundne estimerede regressionsligninger
Da vi fandt, at apparater ingen rolle spiller vælges "gennemsnit" af apparater (de dummy variable er i gennemsnit 0)
Da vi har 2 niveauer for katalysator er dummy variable 1 og -1, så vi får

Katalysator (C_1): $\widetilde{Y} = 5.4075 + (-5.065) + (0.1005 + 0.0405)T - 2.975P$

eller $\underline{\underline{\widetilde{Y} = 0.343 + 0.141T - 2.975P}}$

Katalysator (C_2): $\widetilde{Y} = 5.4075 - (-5.065) + (0.1005 - 0.0405)T - 2.975P$

eller $\underline{\underline{\widetilde{Y} = 10.4725 + 0.060T - 2.975P}}$

Øvelse 5.1 (3 kvantitative faktorer og en blokfaktor) Regn opgave 37 side 252.

5.3. Faktorforsøg med faktorer på mere end 2 niveauer.

De følgende 3 eksempler belyser lidt forskellige problemstillinger, men følger i øvrigt stort set samme procedure som i eksempel 5.1. Ordrer i Statgraphics er derfor også de samme, og kan iøvrigt om nødvendigt findes i appendix A.

Eksempel 5.2 (1 kvalitativ og 1 kvantitativ faktor) (SFIII eksempel 13 side 76).
Ved et fuldstændigt randomiseret forsøg undersøgtes udbyttet Y af en kemisk proces, idet forsøgsfaktoren var:
Faktor A: Reaktionstemperatur x med niveauerne 80°C, 100°C *og* 120°C.
Faktor B : Katalysatortype med to niveauer (B_1) *og* (B_2)
Forsøgsresultaterne var:

		A: (temperatur x)		
		80	100	120
B (katalysatortype)	B_1	6.3 2.9	12.8 10.1	19.0 23.2
	B_2	5.1 6.6	8.7 14.2	39.6 39.0

For hver katalysatortype kan udbyttet skrives på formen
$Y = \alpha_i + \beta_{1i}x + \beta_{2i}x^2$, $i = 1, 2$ (*katalysatortype*).
Dette er den fuldstændige model.

1) *Undersøg om det er muligt at reducere modellen. Det skal således undersøges om koefficienten til* x^2 *er uafhængig af katalysatortype eventuelt om leddet helt kan bortkastes. Opskriv de estimerede regressionsligninger i den endelige model.*
2) *Giv er estimat for middelprocesudbyttet ved temperaturen* 110°C *for katalysator* (B_2), *og angiv et* 95% *konfidensinterval for dette middeludbytte.*

Spørgsmål 1. Reduktion af model.
Indtastning af data
Værdierne indtastes på sædvanlig måde:

```
x     B          udbytte
80    B1            6,3
80    B1            2,9
100   B1            12,8
100   B1            10,1
120   B1            19
120   B1            23,2
80    B2            5,1
80    B2            6,6
100   B2            8,7
100   B2            14,2
120   B2            39,6
120   B2            39
```

Opstilling af regressionstabel

I GLM sættes "Dependent variables" : udbytte
"Categorical Factors" : B
og "Quantitative Factors" : x

I "GLM specifications" hvor A= B og B = x, udfyldes med modellen (A, B, A*B, B*B, A*B*B). Der fremkommer følgende udskrift:

```
General Linear Models
---------------------
Analysis of Variance for udbytte
-----------------------------------------------------------------------------
Source              Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-----------------------------------------------------------------------------
Model                      1684,89      5      336,977      58,31       0,0001
Residual                    34,675      6      5,77917
-----------------------------------------------------------------------------
Total (Corr.)              1719,56     11

Type III Sums of Squares
------------------------------------------------------------------------------------
Source                      Sum of Squares     Df   Mean Square    F-Ratio     P-Value
------------------------------------------------------------------------------------
B                                  44,7287      1       44,7287       7,74      0,0319
x                                  66,8111      1       66,8111      11,56      0,0145
B*x                                52,3655      1       52,3655       9,06      0,0237
x*x                                104,584      1       104,584      18,10      0,0054
B*x*x                              63,0504      1       63,0504      10,91      0,0163
Residual                            34,675      6      5,77917
------------------------------------------------------------------------------------
Total (corrected)                  1719,56     11
```

Modellen skrives nu med de betegnelser, som anvendes af Statgraphics:
$Y = (\text{Constant} + \text{B}) + (\text{konst}_1 + \text{B})\, x + (\text{konst}_2 + \text{B})\, x^2$

Da P-value for B*x* x er 0.0163 < 0.05
forkastes nulhypotesen H_0: Koefficienten til x^2 er uafhængig af katalysatortype
eller kort: B*x* x: P-value = 0.0163* < 0.05 $\Longrightarrow$ β_{2i} afhænger af " i ".

Da P-value for B*x er 0.0237 < 0.05
forkastes nulhypotesen H_0: Koefficienten til x er uafhængig af katalysatortype

Modellen kan derfor ikke reduceres.

Regressionsligningerne

Koefficienterne i den endelige model (vælg "Model Coefficients", fås af følgende tabel:

```
95,0% confidence intervals for coefficient estimates (udbytte)
------------------------------------------------------------------------------------------
                                         Standard
Parameter                  Estimate         Error     Lower Limit     Upper Limit     V.I.F.
------------------------------------------------------------------------------------------
CONSTANT                    105,575       36,0799         17,2905          193,86
B                          -100,375       36,0799         -188,66        -12,0905     2703,0
x                          -2,50688      0,737294        -4,31097       -0,702776      301,0
B* x                        2,21937      0,737294        0,415276         4,02347    11588,5
x*x                       0,0156563    0,00368034      0,00665075       0,0246617      301,0
B*x*x                    -0,0121562    0,00368034      -0,0211617     -0,00315075     3265,5
------------------------------------------------------------------------------------------
```

Da vi har 2 niveauer for katalysator er dummy variable 1 og -1, og vi får

$$\text{Katalysator}(B_1)\text{:}\ \tilde{Y} = 105.575 - 100.275 + (-2.50688 + 2.21937)\cdot x + (0.015656 - 0.012156)\cdot x^2$$
$$= 5.300 - 0.2875\cdot x + 0.003500\cdot x^2\ .$$

$$\text{Katalysator}(B_2)\text{:}\ \tilde{Y} = 105.575 + 100.275 + (-2.50688 - 2.21937)\cdot x + (0.015656 + 0.012156)\cdot x^2$$
$$= 205.85 - 4.7262\cdot x + 0.027813\cdot x^2$$

Spørgsmål 2. Estimat og 95% konfidensinterval.
Et estimat for middelprocesudbyttet ved temperaturen 110° for katalysator B_2 fås ved at tilføje 110 og B_2 nederst i data.
Vælges " Report" fås følgende udskrift:

```
Regression Results for udbytte
------------------------------------------------------------------------------------------------
       Fitted      Stnd. Error  Lower 95,0% CL  Upper 95,0% CL  Lower 95,0% CL  Upper 95,0% CL
Row     Value     for Forecast    for Forecast    for Forecast        for Mean        for Mean
------------------------------------------------------------------------------------------------
13      22,5938        2,80287         15,7354         29,4521         19,0674         26,1201
------------------------------------------------------------------------------------------------
```

Vi ser, at $y = 22.59$ og konfidensinterval: [19.07 ; 26.12]

Eksempel 5.3 (1 kvalitativ og 1 kvantitativ faktor) (hentet fra SFIII eksempel 14 side 80).
Ved studiet af udbyttet ved pencillingæring undersøgtes dels virkningen af 5 sukkerarter, dels virkningen af 4 forskellige koncentrationer af sukker. Forsøgsplanen var et fuldstændig randomiseret forsøg med en fuldstændig faktorstruktur. Resultaterne var

		Sukkerkoncentration c (%)			
		2	4	6	8
Sukkerarter	Laktose	0.606	0.660	0.984	0.908
	Glukose	0.715	1.024	1.000	1.125
	Rørsukker	0.761	0.933	1.072	0.979
	Melasse	0.815	1.000	1.015	1.162
	Dextrin	0.775	0.918	1.030	1.049

Det formodes, at middeludbyttet for hver sukkerart kan beskrive sammenhængen mellem udbytte og koncentration ved et 2.grads polynomium

$Y = \alpha_i + \beta_{1i}c + \beta_{2i}c^2$, $i = 1, 2, 3, 4, 5$ (sukkerart).

1) *Benyt størrelsen af R-square til at vurdere om en sådan model er rimelig.*
2) *Søg at reducere modellen ved bl.a. at undersøge om regressionskoefficienterne afhænger af sukkerarterne.*
3) *Angiv for den endelige model regressionsligningerne.*

Spørgsmål 1. Benyt størrelsen af R-square til at vurdere om en 2.grads model er rimelig.
Indtastning af data
Værdierne indtastes på sædvanlig måde:

```
c  udbytte   sukkerart
2  0,606     laktose
4  0,66      laktose
6  0,984     laktose
8  0,908     laktose
2  0,715     glukose
4  1,024     glukose
6  1         glukose
8  1,125     glukose
2  0,761     rørsukker
4  0,933     rørsukker
6  1,072     rørsukker
8  0,979     rørsukker
2  0,815     melasse
4  1         melasse
6  1,015     melasse
8  1,162     melasse
2  0,775     dekstrin
4  0,918     dekstrin
6  1,03      dekstrin
8  1,049     dekstrin
```

Opstilling af regressionstabel

I "GLM" indsættes "Dependent variables" : udbytte
"Categorical Factors" : sukkerart
og "Quantitative Factors" : c

I "GLM specifications" hvor A= sukkerart og B = c, udfyldes med modellen (A, B, A*B, B*B, A*B*B). Der fremkommer følgende udskrift:

```
General Linear Models
---------------------
Number of dependent variables: 1
Number of categorical factors: 1
Number of quantitative factors: 1
Analysis of Variance for udbytte
-----------------------------------------------------------------------------
Source             Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-----------------------------------------------------------------------------
Model                    0,416869     14    0,0297764       3,65       0,0802
Residual                0,0407936      5   0,00815873
-----------------------------------------------------------------------------
Total (Corr.)            0,457663     19

Type III Sums of Squares
----------------------------------------------------------------------------------
Source                    Sum of Squares     Df   Mean Square    F-Ratio    P-Value
----------------------------------------------------------------------------------
sukkerart                      0,0105017      4    0,00262542       0,32     0,8527
c                              0,0650275      1     0,0650275       7,97     0,0370
sukkerart*c                   0,00599944      4    0,00149986       0,18     0,9371
c*c                            0,0274541      1     0,0274541       3,36     0,1261
sukkerart*c*c                  0,0069962      4    0,00174905       0,21     0,9196
Residual                       0,0407936      5    0,00815873
----------------------------------------------------------------------------------
Total (corrected)               0,457663     19
All F-ratios are based on the residual mean square error.
R-Squared = 91,0865 percent      R-Squared (adjusted for d.f.) = 66,1288 percent
```

Modellen kan ikke testes, da der ikke er gentagelser, men da

R-Squared = $\frac{SAK_{model}}{SAK_{total}} = \frac{0.4170}{0.4577} = 91.08\%$ er modellen ikke urimelig.

Sammenlignes R-Squared med R-Squared(adjusted for d.f.) = 66.12% som jo er væsentlig mindre, synes det åbenbart, at det må være muligt at reducere modellen. Der må være brugt en del frihedsgrader til led, som ikke har forbedret modellen væsentligt.

Spørgsmål 2. Mulig reduktion af modellen.
Modellen skrives nu med de betegnelser, som anvendes af Statgraphics:
Y = (Constant + sukkerart) + ($konst_1$ + sukkerart) c + ($konst_2$ + sukkerart) c^2

Først testes H_0: Koefficienten til c^2 er uafhængig af sukkerart
Da P-value for sukkerart*c*c er 0.9196> 0.05 accepteres nulhypotesen , dvs koefficienten er uafhængig heraf.

Modellen er nu: Y = (Constant + sukkerart) + ($konst_1$ + sukkerart) c + $konst_2$ c^2

Der foretages en pooling. (fjern A*B*B)

```
General Linear Models
---------------------
Number of dependent variables: 1
Number of categorical factors: 1
Number of quantitative factors: 1
Analysis of Variance for udbytte
```

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	0,409873	10	0,0409873	7,72	0,0026
Residual	0,0477898	9	0,00530998		
Total (Corr.)	0,457663	19			

Type III Sums of Squares

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
sukkerart	0,0303718	4	0,00759295	1,43	0,3004
c	0,0650275	1	0,0650275	12,25	0,0067
sukkerart*c	0,00680424	4	0,00170106	0,32	0,8575
c*c	0,0274541	1	0,0274541	5,17	0,0491
Residual	0,0477898	9	0,00530998		
Total (corrected)	0,457663	19			

```
R-Squared = 89,5579 percent  R-Squared (adjusted for d.f.) = 77,9555 percent
```

Nu testes H_0: Koefficienten til c^2 er nul.
Da *P*-value for c*c er 0.0491< 0.05 forkastes nulhypotesen

Nu testes H_0: Koefficienten til *c* er uafhængig af sukkerart
Da *P*-value for sukkerart*c er 0.8575> 0.05 accepteres nulhypotesen , dvs. koefficienten er uafhængig heraf.
Modellen er nu: Y = (Constant + sukkerart) + $konst_1$ c + $konst_2$ c^2.

Der foretages igen en pooling (fjern A*B), og vi får:

```
General Linear Models
---------------------
Number of dependent variables: 1         Number of categorical factors: 1
Number of quantitative factors: 1
Analysis of Variance for udbytte
-----------------------------------------------------------------------------
Source             Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-----------------------------------------------------------------------------
Model                    0,403069      6    0,0671781      16,00       0,0000
Residual                0,0545941     13   0,00419955
-----------------------------------------------------------------------------
Total (Corr.)            0,457663     19

Type III Sums of Squares
----------------------------------------------------------------------------------
Source                    Sum of Squares     Df   Mean Square    F-Ratio    P-Value
----------------------------------------------------------------------------------
sukkerart                       0,103235      4     0,0258088       6,15     0,0053
c                              0,0650275      1     0,0650275      15,48     0,0017
c*c                            0,0274541      1     0,0274541       6,54     0,0239
Residual                       0,0545941     13    0,00419955
----------------------------------------------------------------------------------
Total (corrected)               0,457663     19
All F-ratios are based on the residual mean square error.
R-Squared = 88,0711 percent  R-Squared (adjusted for d.f.) = 82,5655 percent
```

Nu testes H_0: Sukkerart har ingen virkning
Da *P*-value for sukkerart er 0.0053< 0.05 forkastes nulhypotesen .

Vi har følgelig, at de 5 regressionspolynomier kun adskiller sig fra hinanden ved at have forskellige konstantled.

Bemærk i overensstemmelse med det forventede, at R-squared er faldet lidt, men at R-squared (adjusted for d.f.) er steget ganske betydeligt.

Spørgsmål 3. Regressionsligningerne i den endelige model.
Koefficienterne i den endelige model fås ved valg af " Model Coefficients". Dette giver følgende tabel:

```
95,0% confidence intervals for coefficient estimates (udbytte)
--------------------------------------------------------------------------------------------
                                           Standard
Parameter                    Estimate         Error     Lower Limit     Upper Limit       V.I.F.
--------------------------------------------------------------------------------------------
CONSTANT                      0,48035     0,0806802        0,306051        0,654649
sukkerart                     0,01645     0,0289812      -0,0461602       0,0790602          1,6
sukkerart                     0,03945     0,0289812      -0,0231602         0,10206          1,6
sukkerart                    -0,13705     0,0289812        -0,19966      -0,0744398          1,6
sukkerart                     0,07145     0,0289812      0,00883982         0,13406          1,6
c                            0,144815     0,0368015       0,0653099         0,22432        32,25
c*c                        -0,0092625    0,00362265      -0,0170888     -0,00143623        32,25
--------------------------------------------------------------------------------------------
```

Vi har nu følgende (idet rækkefølgen af udskrift er ordnet alfabetisk):

Dekstrin: $\widetilde{Y} = 0.48035 + 0.01645 + 0.1448 \cdot c - 0.0092625 \cdot c^2$
$= 0.4968 + 0.1448 \cdot c - 0.0092625 \cdot c^2$

Glukose: $\widetilde{Y} = 0.48035 + 0.03945 + 0.1448 \cdot c - 0.0092625 \cdot c^2$
$= 0.5198 + 0.1448 \cdot c - 0.0092625 \cdot c^2$

Laktose: $\widetilde{Y} = 0.48035 - 0.13705 + 0.1448 \cdot c - 0.0092625 \cdot c^2$
$= 0.3433 + 0.1448 \cdot c - 0.0092625 \cdot c^2$

Melasse: $\widetilde{Y} = 0.48035 + 0.07145 + 0.1448 \cdot c - 0.0092625 \cdot c^2$
$\underline{\underline{= 0.5518 + 0.1448 \cdot c - 0.0092625 \cdot c^2}}$

Rørsukker: $\widetilde{Y} = 0.48035 - 0.01645 - 0.03945 + 0.13705 - 0.07145 + 0.1448 \cdot c - 0.0092625 \cdot c^2$
$\underline{\underline{= 0.4901 + 0.1448 \cdot c - 0.0092625 \cdot c^2}}$

◆

Øvelse 5.2 (2×4 faktorforsøg med 1 kvantitativ og 1 kvalitativ faktor)
Regn opgave 39 side 254.

Eksempel 5.4 (1 kvantitativ og 2 kvalitative faktorer)
En fabrik fremstiller salpetersyre ved oxidering af ammoniak med luft. I løbet af processen ledes kvælstofoxider under afkøling ind i en absorptionskolonne. Absorptionen i gennemstrømmende salpetersyre afhænger af lufttemperatur T (°C) . *Endvidere kan absorptionen tænkes at afhænge af valg af katalysator* B (3 *muligheder*) *og valg af et additiv* A (2 *muligheder*)
Følgende observationer af Y (*kodede tal*) *fandtes:*

				T = 10	T = 20	T = 30	T = 40
A	A_1	B	B_1	32 30	50 47	62 66	88 82
			B_2	48 51	60 63	73 77	96 99
			B_3	54 56	70 75	90 83	102 110
	A_2	B	B_1	32 35	49 52	64 60	86 84
			B_2	50 55	64 60	82 77	92 94
			B_3	57 53	74 77	88 86	108 105

1) *Undersøg, om sammenhængen mellem mængden Y af ikke-absorberede kvælstofoxider afhænger af de tre faktorer, idet man som udgangspunkt antager, at Y for hver af de 6 kombinationer af de* 2 *kvalitative faktorer A og B kan beskrives ved et andengradspolynomium i* T, *dvs. af formen*
$Y = \alpha_{ij} + \beta_{1ij} T + \beta_{2ij} T^2$, hvor i = 1,2 (*additiv*) , *j* = 1, 2, 3 (*katalysator*)
2) *Undersøg, om modellen kan reduceres.*
3) *Foretag en grafisk kontrol af den fremkomne model.*
4) *Opskriv ligningerne for den endelige model.*
5) *Angiv et 95% konfidensinterval for* β_2
6) *Angiv et estimat for Y i tilfældet A er på niveau* A_1 , *B er på niveau* B_2 *og T er* = 20.

LØSNING:

Spørgsmål 1. Undersøg om Y for hver af de 6 kombinationer af A og B kan beskrives ved et andengradspolynomium i T dvs. af formen $Y = \alpha_{ij} + \beta_{1ij}T + \beta_{2ij}T^2$

Indtastning af data. Tallene indtastes på sædvanlig måde.

Opstilling af regressionstabel

I "GLM" indsættes "Dependent variables" : Y
"Categorical Factors" : A og B
og "Quantitative Factors" : T

I "GLM specifications" hvor A= A, B = B og C = T udfyldes med
A, B, C, A*B, A*C, B*C, A*B*C, C*C, A*C*C, B*C*C, A*B*C*C.
Der fremkommer følgende udskrift:

```
General Linear Models
---------------------
Number of dependent variables: 1
Number of categorical factors: 2
Number of quantitative factors: 1

Analysis of Variance for Y
-----------------------------------------------------------------------------
Source            Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-----------------------------------------------------------------------------
Model                    20275,4     17      1192,67     123,57       0,0000
Residual                  289,55     30      9,65167
-----------------------------------------------------------------------------
Total (Corr.)            20565,0     47

Type III Sums of Squares
----------------------------------------------------------------------------------
Source                   Sum of Squares     Df   Mean Square    F-Ratio    P-Value
----------------------------------------------------------------------------------
A                             0,0967742      1     0,0967742       0,01     0,9209
B                               155,042      2       77,5208       8,03     0,0016
T                               225,457      1       225,457      23,36     0,0000
A*B                              7,2621      2       3,63105       0,38     0,6897
A*T                             1,64974      1       1,64974       0,17     0,6822
B*T                             32,7059      2       16,3529       1,69     0,2008
A*B*T                           9,35394      2       4,67697       0,48     0,6207
T*T                             52,0833      1       52,0833       5,40     0,0271
A*T*T                           4,08333      1       4,08333       0,42     0,5204
B*T*T                           23,2917      2       11,6458       1,21     0,3133
A*B*T*T                         10,2917      2       5,14583       0,53     0,5922
Residual                         289,55     30       9,65167
----------------------------------------------------------------------------------
Total (corrected)               20565,0     47
All F-ratios are based on the residual mean square error.

R-Squared = 98,592 percent
R-Squared (adjusted for d.f.) = 97,7942 percent
```

For at få $SAK_{gentagelser}$ kan vi udføre en tresidet variansanalyse med vekselvirkning (den mest omfattende model), hvoraf vi kan aflæse "gentagelserne".
I "Compare" sættes "Maximum Order Interaction" til 3". Resultatet bliver:

```
Analysis of Variance for Y - Type III Sums of Squares
--------------------------------------------------------------------------------
Source                  Sum of Squares     Df    Mean Square    F-Ratio    P-Value
--------------------------------------------------------------------------------
MAIN EFFECTS
 A:A                          8,33333      1        8,33333       1,01     0,3249
 B:B                          4313,63      2        2156,81     261,43     0,0000
 C:T                          15866,2      3        5288,72     641,06     0,0000
INTERACTIONS
 AB                          0,291667      2       0,145833       0,02     0,9825
 AC                           20,1667      3        6,72222       0,81     0,4983
 BC                           113,708      6        18,9514       2,30     0,0680
 ABC                          44,7083      6        7,45139       0,90     0,5089
RESIDUAL                        198,0     24           8,25
--------------------------------------------------------------------------------
TOTAL (CORRECTED)             20565,0     47
--------------------------------------------------------------------------------
```

Heraf ses, at $SAK_{gentagelser} = 198{,}0$ med f = 24 (24 celler).

Da $SAK_{residual\text{-}alle} = SAK_{gentagelser} + SAK_{lack}$ fås $SAK_{lack} = 289.55 - 198.0 = 91.55$

med $f_{lack} = 30 - 24 = 6$

Vi har følgelig at $s_{lack}^2 = \frac{SAK_{lack}}{f_{lack}} = 15.26$ og dermed $F_{lack} = \frac{s_{lack}^2}{s_0^2} = \frac{15.26}{8.25} = 1.849$

Vi kan nu danne følgende variansanalysetabel:

Variation	SAK	f	s^2	F
Model (Regres)	20275,4	17		
Lack of fit	91,55	6	15,26	1,849
Gentagelser	198.0	24	8,25	
Total	20565,0	47		

Da $F_{0.95}(6,24) = 2{,}51 > F_{lack} = 1.85$ accepteres den kvadratiske model : $Y = \alpha_{ij} + \beta_{1ij}T + \beta_{2ij}T^2$

Benyttes Statgraphic fås:

```
Cumulative Distribution
-----------------------
Distribution: F (variance ratio)
              Upper Tail Area (>)
Variable      Dist. 1       Dist. 2       Dist. 3       Dist. 4       Dist. 5
1,85          0,131541
```

dvs. P - værdi = 0.1315 > 0.05 hvilket også giver en accept af modellen.

Spørgsmål 2. Mulig reduktion af modellen

Modellen skrives nu med de betegnelser, som anvendes af Statgraphics:

Y = (Constant + A + B + A*B) + (konst$_1$ + A + B + A*B) T + (konst$_2$ +A + B + A*B) T^2

Vi går tilbage til udskriften med koefficienterne til andengradsudtrykket.

2A: Først testes nulhypotesen: H_0: Koefficienten til T^2 er uafhængig af A og B.

Modellen reduceres ved at poole de tre led tre vekselvirkninger A*T^2, B*T^2 og A*B*T^2 ned i residualen Vælg " Input dialog" og fjern A*C*C, B*C*C, A*B*C*C . Vi får så følgende tabel:

```
General Linear Models
---------------------
Number of dependent variables: 1
Number of categorical factors: 2
Number of quantitative factors: 1

Analysis of Variance for Y
-----------------------------------------------------------------------------
Source             Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-----------------------------------------------------------------------------
Model                     20237,8     12      1686,48     180,39       0,0000
Residual                  327,217     35      9,34905
-----------------------------------------------------------------------------
Total (Corr.)             20565,0     47

Type III Sums of Squares
----------------------------------------------------------------------------------
Source                        Sum of Squares     Df   Mean Square    F-Ratio    P-Value
----------------------------------------------------------------------------------
A                                    23,3472      1       23,3472       2,50     0,1230
B                                    893,861      2       446,931      47,80     0,0000
T                                    225,457      1       225,457      24,12     0,0000
A*B                                  6,02778      2       3,01389       0,32     0,7266
A*T                                  16,0167      1       16,0167       1,71     0,1991
B*T                                  60,8583      2       30,4292       3,25     0,0505
A*B*T                                7,75833      2       3,87917       0,41     0,6636
T*T                                  52,0833      1       52,0833       5,57     0,0240
Residual                             327,217     35       9,34905
----------------------------------------------------------------------------------
Total (corrected)                    20565,0     47
All F-ratios are based on the residual mean square error.

R-Squared = 98,4089 percent  R-Squared (adjusted for d.f.) = 97,8633 percent
```

Vi har nu $s_m^2 = \dfrac{327.217 - 289.55}{5} = 7.533$. Vi får $F = \dfrac{s_m^2}{s_0^2} = \dfrac{7.533}{9.6517} = 0.781$.

Da F < 1 accepteres nulhypotesen " Koefficienten til T^2 uafhængig af A og B "
Modellen skrives nu:

$$Y = (\text{Constant} + \text{A} + \text{B} + \text{A*B}) + (\text{konst}_1 + \text{A} + \text{B} + \text{A*B})\, T + (\text{konst}_2)\, T^2$$

Vi ser, at nulhypotesen: H_0: Koefficienten til T^2 er nul forkastes, da *P*-value er 0.0240< 0.05.

2B: Nu testes nulhypotesen: H_0: Koefficienten til T er uafhængig af A og B.
Modellen reduceres ved at poole de tre vekselvirkninger A*T, B*T og A*B*T ned i residualen.

Vi får så følgende tabel:

General Linear Models

Analysis of Variance for Y

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	20153,1	7	2879,02	279,62	0,0000
Residual	411,85	40	10,2963		
Total (Corr.)	20565,0	47			

Type III Sums of Squares

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
A	8,33333	1	8,33333	0,81	0,3737
B	4313,63	2	2156,81	209,48	0,0000
T	225,457	1	225,457	21,90	0,0000
A*B	0,291667	2	0,145833	0,01	0,9859
T*T	52,0833	1	52,0833	5,06	0,0301
Residual	411,85	40	10,2963		
Total (corrected)	20565,0	47			

Vi får $s_{m_1}^2 = \dfrac{411.85 - 327.217}{5} = 16.92$ og dermed $F = \dfrac{s_{m_1}^2}{s_0^2} = \dfrac{16.92}{9.3491} = 1.81$

Da $F_{0.95}(5,35) = 2.47$ (eller P - værdi = 0.1363 ifølge Statgraphics) accepteres nulhypotesen.

Modellen er nu: Y = (Constant + A + B + A*B) + (konst$_1$) T + (konst$_2$) T^2 .

2C: Vi ønsker dernæst at teste nulhypotesen H_0: Konstantleddet er uafhængigt af A og B.

Det ses umiddelbart, at vi kan se bort fra A*B og fra A, dvs. additivet spiller ingen rolle.

Vi kan nu ikke reducere modellen yderligere.

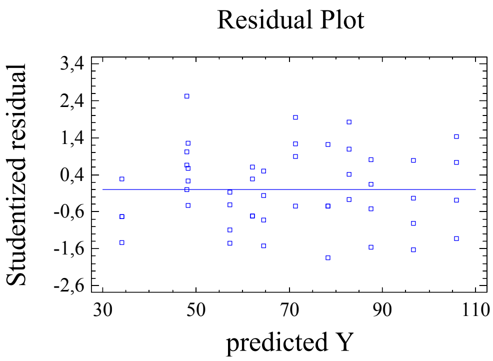
Den endelige model er : Y = (Constant + B) + (konst$_1$) T + (konst$_2$) T^2

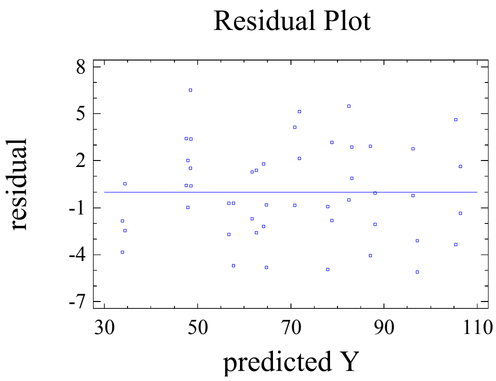
Spørgsmål 3. Foretag en grafisk kontrol af den fremkomne model.

Vi foretager en pooling af A og A*B. Vi danner nu

a) Residualplot: Studentized: Vælg "Residual plots"

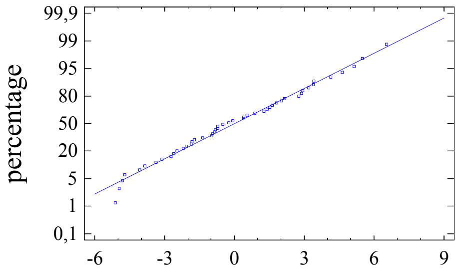
Residualer: "højre musetast | Residuals, Scatter plot, Predicted value





Det ses, at punkterne fordeler sig nogenlunde tilfældigt omkring linierne, og at der er ingen outliers, da Studentized Residuals er under 3 .

b) normalfordelingsplot:(Cursor på figur | højre musetast |Normal Probability Plot |OK)



Residualerne synes at være rimeligt normalfordelt.
Modellen synes derfor at være accepteret.

Spørgsmål 4. Regressionsligninger

Vælg " Model Coefficients". Dette giver følgende tabel:

95,0% confidence intervals for coefficient estimates (Y)

Parameter	Estimate	Standard Error	Lower Limit	Upper Limit	V.I.F.
CONSTANT	34,4167	2,51302	29,3487	39,4847	
B	-12,3125	0,638308	-13,5998	-11,0252	1,33333
B	1,5625	0,638308	0,275227	2,84977	1,33333
T	1,10083	0,229258	0,638489	1,56318	32,25
T*T	0,0104167	0,00451352	0,00131427	0,0195191	32,25

Ligningen bliver for B = B_1 :

$$\tilde{Y} = 34.42 - 12.31 + 1.10 \cdot T + 0.0104 \cdot T^2 \underline{\underline{= 22.11 + 1.10 \cdot T + 0.0104 \cdot T^2}}$$

Ligningen bliver for B = B_2 :

$$\tilde{Y} = 34.42 + 1.56 + 1.10 \cdot T + 0.0104 \cdot T^2 \underline{\underline{= 35.98 + 1.10 \cdot T + 0.0104 \cdot T^2}}$$

Ligningen bliver for B = B_3 :

$$\tilde{Y} = 34.42 + 12.3125 - 1.5625 + 1.10 \cdot T + 0.0104 \cdot T^2 \underline{\underline{= 45.17 + 1.10 \cdot T + 0.0104 \cdot T^2}}$$

Spørgsmål 5. Konfidensinterval for β_2

Af ovenstående udskrift ses, at konfidensintervallet er [0.00131 ; 0.01952].

Spørgsmål 6. Angiv et estimat for *Y* i tilfældet A er på niveau A_1 , B er på niveau B_2 og *T* er 20.

Et estimat for middelværdien af *Y* fås ved at tilføje a1, b2 og 20 nederst i data.

Vælg “ Report”. Resultatet blev følgende udskrift:

Regression Results for Y

Row	Fitted Value	Stnd. Error for Forecast	Lower 95,0% CL for Forecast	Upper 95,0% CL for Forecast	Lower 95,0% CL for Mean	Upper 95,0% CL for Mean
49	62,1625	3,261	55,5861	68,7389	60,2971	64,0279

Y er = 62.16 med et konfidensinterval på [60.30 ; 64.03]

Øvelse 5.3 (3×3×2 faktorforsøg med 2 kvantitative og 1 kvalitativ faktor)
Regn opgave 40 side 254.

6 KOVARIANSANALYSEFORSØG

6.1 INDLEDNING

For at dæmpe "støjen" dvs. nedsætte virkningen af stor heterogenitet mellem forsøgsenhederne, har vi hidtil indført blokke, som var kvalitative. I en række tilfælde, er den faktor der kunne tænkes at bevirke heterogeniteten imidlertid kvantitativ. Man kan også sige, at man forsøger at korrigere forsøgsvariablen for forskelle i størrelsen af en målelig egenskab (kaldet **kovariablen**) ved forsøgsenhederne. Der kan naturligvis også optræde flere kovariable.

6.2 ENSIDET KOVARIANSANALYSE MED 1 KVALITATIV FAKTOR

Eksempel 6.1. (ensidet kovariansanalyse)
Ved et forsøg ønskes undersøgt, om middelstyrken μ *af tre forskellige limtyper er den samme. Da man havde erfaring for, at* μ *afhænger af tykkelsen x af det benyttede limlag, og da det ikke er praktisk muligt at fastholde samme x-værdi under samtlige delforsøg, foretoges for hver limtype fire sammenhørende målinger af henholdsvis tykkelsen x og styrke y. Forsøgsresultaterne var (kodede tal):*

lim	lim 1	y	10	5	6	10
		x	5	1	3	6
	lim 2	y	9	12	5	7
		x	7	10	4	5
	lim 3	y	11	7	7	2
		x	12	8	9	4

Det antages, at den kovariable x's eventuelle virkning på y er af 1. grad, dvs. for hver limtype er $Y = \alpha_i + \beta_i x$*, hvor* $i = 1, 2, 3$ *(limtype)*
Det antages endvidere, at β_i *ikke afhænger af limtypen (de tre linier er parallelle).*

1. *Giv en vurdering af, om ovennævnte forudsætninger er rimelige.*
2. *Foretag en statistisk analyse af, om der er forskel på de tre limtypers bindingsevne.*

Bemærk: Kovariablen *x* må **ikke** afhænge af behandlingen. En kontrol af at dette ikke er tilfældet kunne ske ved at foretage en ensidet variansanalyse med *x* som afhængig variabel og lim som faktor.

LØSNING:

Spørgsmål 1. Vurdering af at den kovariables virkning på *y* er af 1.ste grad.

Indtastning af data

Tallene indtastes på sædvanlig måde:

```
lim    x    y
lim1   5    10
lim1   1    5
lim1   3    6
lim1   6    10
lim2   7    9
lim2   10   12
lim2   4    5
lim2   5    7
lim3   12   11
lim3   8    7
lim3   9    7
lim3   4    2
```

Opstilling af regressionstabel

Vælg ikonen "GLM" eller (Special | Advanced Regression | General Linear Models" (GLM)).
På det følgende skema indsættes "Dependent variables" : y
"Categorical Factors" : lim
og "Quantitative Factors" : x
Efter tryk på "OK" fremkommer et skema,"GLM specifications" hvor A= lim og B = x.
"Effects" udfyldes med modellen (A, B, A*B), da vi vil se, om forudsætningen om parallelle linier er rimelig. Der fremkommer følgende udskrift:

```
General Linear Models
---------------------
Number of dependent variables: 1
Number of categorical factors: 1
Number of quantitative factors: 1

Analysis of Variance for y
-----------------------------------------------------------------------------
Source             Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-----------------------------------------------------------------------------
Model                 90,1095         5      18,0219       38,52       0,0002
Residual              2,80715         6     0,467858
-----------------------------------------------------------------------------
Total (Corr.)              92,9167       11

Type III Sums of Squares
------------------------------------------------------------------------------
Source            Sum of Squares     Df   Mean Square    F-Ratio    P-Value
------------------------------------------------------------------------------
lim                 9,8538           2       4,9269       10,53      0,0109
x                  77,4221           1      77,4221      165,48    0,0000
lim*x               0,00854472       2     0,00427236      0,01       0,9909
Residual            2,80715          6      0,467858
-----------------------------------------------------------------------------
Total (corrected)  92,9167         11
All F-ratios are based on the residual mean square error.

R-Squared = 96,9789 percent   R-Squared (adjusted for d.f.) = 94,4612 percent
```

Vi kan ikke teste modellen, men da R-squared = 96.9 % , synes tre førstegradspolynomier at være en rimelig god model.
Da P-value for lim*x er 0.99, kan vi se bort fra dette led, dvs. de tre regressionslinier kan antages parallelle.

Spørgsmål 2. Statistisk analyse med konklusion.
Vi sletter nu lim * x fra modellen. Vælg (rød ikon = Input dialog |OK | fjern A*B | OK).
Vi får følgende udskrift:

```
General Linear Models
---------------------
Number of dependent variables: 1
Number of categorical factors: 1
Number of quantitative factors: 1

Analysis of Variance for y
-----------------------------------------------------------------------------
Source            Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-----------------------------------------------------------------------------
Model                    90,101      3      30,0337      85,33       0,0000
Residual                2,81569      8     0,351962
-----------------------------------------------------------------------------
Total (Corr.)           92,9167     11

Type III Sums of Squares
------------------------------------------------------------------------------
Source             Sum of Squares     Df   Mean Square    F-Ratio    P-Value
------------------------------------------------------------------------------
lim                      46,8971      2       23,4485      66,62     0,0000
x                        85,4343      1       85,4343     242,74     0,0000
Residual                 2,81569      8      0,351962
------------------------------------------------------------------------------
Total (corrected)        92,9167     11
All F-ratios are based on the residual mean square error.

R-Squared = 96,9697 percent    R-Squared (adjusted for d.f.) = 95,8333 percent
```

Det ses, at der er forskel på limtyperne da *P*-value er 0.000.
Endvidere ses det, at det var fornuftigt at indføre kovariablen x, da *P*-value for x også er 0.0000.

Et indtryk af hvor stor forskellen er fås påfølgende måde.
Vælg “Table of Means”. Derved fremkommer følgende udskrift:

```
Table of Least Squares Means for y with 95,0 Percent Confidence Intervals
------------------------------------------------------------------------------
Level                Count     Mean          Error       Limit       Limit
------------------------------------------------------------------------------
GRAND MEAN           12        7,58333       0,17126     7,18841     7,97826
lim
lim1                  4       10,4489       0,343509     9,65677      11,241
lim2                  4       7,87774       0,297593     7,19149     8,56399
lim3                  4       4,42336       0,332101     3,65753     5,18919
------------------------------------------------------------------------------
Means of Quantitative Factors
-----------------------------
x                   6,16667
```

Da bindingsstyrken afhænger af limtykkelsen x, må de tre limtyper sammenlignes ved samme x-værdi. I denne udskrift sker sammenligningen af bindingsstyrken for hver limtype ved en x- værdi på 6.1667 (gennemsnittet af alle x-værdier).
Det ses, at limtype 1 har den største bindingsstyrke (værdi 10.45 er den største og konfidensintervallet overlapper ikke med de øvrige).

Da limtykkelsen x er indført som støjdæmpende faktor, og man ikke kan styre tykkelsen, vil man sædvanligvis ikke være interesseret i at angive ligningerne for de tre parallelle regressionslinier. Ønsker man alligevel dette, så vælg “ Model Coefficients”. Derved fremkommer følgende udskrift:

```
95,0% confidence intervals for coefficient estimates (y)
--------------------------------------------------------------------------------------
                                      Standard
Parameter             Estimate           Error       Lower Limit     Upper Limit       V.I.F.
--------------------------------------------------------------------------------------
CONSTANT              0,696472        0,474048       -0,396688         1,78963
lim                   2,86557        0,297772        2,17891         3,55224        2,01541
lim                   0,294404       0,243375       -0,26682         0,855628       1,34631
x                     1,11679       0,0716807       0,951492         1,28208        1,60097
--------------------------------------------------------------------------------------
```

De tre ligninger bliver:

<u>lim 1:</u> $\tilde{Y} = 0.6965 + 2.8656 + 1.11679 \cdot x = \underline{\underline{3.5621 + 1.1168 \cdot x}}$

<u>lim 2:</u> $\tilde{Y} = 0.6965 + 0.2944 + 1.11679 \cdot x = \underline{\underline{0.9909 + 1.1168 \cdot x}}$

<u>lim 3:</u> $\tilde{Y} = 0.6965 - 2.8656 - 0.2944 + 1.11679 \cdot x = \underline{\underline{-2.4635 + 1.1168 \cdot x}}$

◆

6.3 TOSIDET KOVARIANSANALYSE MED 2 KVALITATIVE FAKTORER

Eksempel 6.2. (tosidet kovariansanalyse)

Ved et forsøg med en kemisk produktion ønskes undersøgt, om middeludbyttet μ *afhænger af benyttet råmateriale (faktor R) og af benyttet katalysatortype (faktor C). Delforsøgene ønskes udført ved samme konstante temperatur, men af teknisk-økonomiske grunde er dette ikke muligt. Da denne variation i temperaturen kan være en betydende støjfaktor registreres den faktiske temperatur T i hvert delforsøg. T er altså en kovariabel).*
Idet udbyttet kaldes y, fandtes følgende forsøgsresultater:

		C: (katalysatortype)			
		C_1		C_2	
		y	T	y	T
R (råmateriale)	R_1	21 19 20	4 0 1	27 20 31	8 3 7
	R_2	10 26 31	2 6 9	38 40 28	6 5 2

Det antages, at den kovariable T's eventuelle virkning på y er af 1 grad, dvs. for hver kombination af faktorerne R og C's niveauer er udbyttet af formen
$Y = \alpha_{ij} + \beta_{ij} \cdot T$ hvor *i* = 1, 2 (*råvare*) og *j* = 1,2 (*katalysator*).

Det antages endvidere, at β *ikke afhænger af faktorernes niveauer (linierne er parallelle).*
1) *Giv en vurdering af, om ovennævnte forudsætninger er rimelige.*
2) *Foretag en statisk analyse af, om udbyttet afhænger af de to faktorer.*
3) *I bekræftende fald angiv hvilke niveauer disse skal indstilles på, for at give det største udbytte.*

Spørgsmål 1.Vurdering af om forudsætninger er opfyldt.
Indtastning af data
Tallene indtastes på sædvanlig måde:

raavare	katalys	T	y
r1	c1	4	21
r1	c1	0	19
r1	c1	1	20
r1	c2	8	27
r1	c2	3	20
r1	c2	7	31
r2	c1	2	10
r2	c1	6	26
r2	c1	9	31
r2	c2	6	38
r2	c2	5	40
r2	c2	2	28

Opstilling af regressionstabel
Vælg ikonen "GLM" eller (Special | Advanced Regression | General Linear Models" (GLM)).
På det følgende skema indsættes "Dependent variables" : y
"Categorical Factors" : raavare og katalys (benævnes B og A)
og "Quantitative Factors" : T (benævnes C)
I "GLM specifications" udfyldes med modellen (A, B, C, A*B, A*C, B*C, A*B*C), da vi vil se, om forudsætningen om parallelle linier er rimelig. Der fremkommer følgende udskrift:

```
General Linear Models
---------------------
Number of dependent variables: 1
Number of categorical factors: 2
Number of quantitative factors: 1

Analysis of Variance for y
-----------------------------------------------------------------------------
Source             Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-----------------------------------------------------------------------------
Model                      756,35      7       108,05      10,65       0,0187
Residual                  40,5671      4      10,1418
-----------------------------------------------------------------------------
Total (Corr.)             796,917     11

Type III Sums of Squares
------------------------------------------------------------------------------------
Source                     Sum of Squares     Df   Mean Square    F-Ratio    P-Value
------------------------------------------------------------------------------------
raavare                           5,47367      1       5,47367       0,54     0,5033
katalys                           25,5769      1       25,5769       2,52     0,1875
T                                 193,679      1       193,679      19,10     0,0120
raavare*katalys                   62,4403      1       62,4403       6,16     0,0681
raavare*T                         38,9337      1       38,9337       3,84     0,1216
katalys*T                         3,63563      1       3,63563       0,36     0,5816
raavare*katalys*T                 6,84854      1       6,84854       0,68     0,4574
Residual                          40,5671      4       10,1418
------------------------------------------------------------------------------------
Total (corrected)                 796,917     11
All F-ratios are based on the residual mean square error.

R-Squared = 94,9095 percent
R-Squared (adjusted for d.f.) = 86,0011 percent
```

Vi kan ikke teste modellen, men da R-squared = 94.91 % ,synes førstegradspolynomier at være en rimelig god model.
Vi tester nu påstanden om at regressionslinierne kan antages parallelle, ved at poole de 3 sidste led væk. Vælg Input dialog og fjern A*C, B*C, A*B*C | OK . Vi får derved følgende udskrift:

```
General Linear Models
---------------------
Number of dependent variables: 1
Number of categorical factors: 2
Number of quantitative factors: 1

Analysis of Variance for udbytte
-----------------------------------------------------------------------------
Source             Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-----------------------------------------------------------------------------
Model                     706,744      4      176,686      13,72       0,0020
Residual                  90,1726      7      12,8818
-----------------------------------------------------------------------------
Total (Corr.)             796,917     11

Type III Sums of Squares
------------------------------------------------------------------------------------
Source                          Sum of Squares     Df   Mean Square    F-Ratio    P-Value
------------------------------------------------------------------------------------
katalys                                97,8221      1       97,8221       7,59     0,0283
raamat                                 27,6711      1       27,6711       2,15     0,1862
T                                      297,161      1       297,161      23,07     0,0020
katalys*raamat                         210,907      1       210,907      16,37     0,0049
Residual                              90,1726      7       12,8818
------------------------------------------------------------------------------------
Total (corrected)                      796,917     11
All F-ratios are based on the residual mean square error.

R-Squared = 88,6848 percent     R-Squared (adjusted for d.f.) = 82,219 percent
```

Af denne ses, at de tre bortkastede led har en samlet SAK på 90,1726 - 40,5671 = 49.61 med 7 - 4 = 3 frihedsgrader. Vi får følgelig, at $s_m^2 = \frac{49.61}{3} = 16.54$ og dermed

$F = \frac{s_m^2}{s_{residual}^2} = \frac{16.54}{10.1418} = 1.63$. Da $F_{0.95}(3,4) = 6.59$ accepteres påstanden, dvs. regressionslinierne kan antages parallelle.

Spørgsmål 2. Undersøg om udbyttet afhænger af de to faktorer.
Det ses, at der er vekselvirkning mellem raamaterialer og katalysatortype, da P-value er 0.0049. Heraf følger, at udbyttet afhænger af de to faktorer.
Endvidere ses det, at det var fornuftigt at indføre kovariablen T, da P-value for T også er 0.0020.

Spørgsmål 3. Hvilke niveauer skal vælges for at få størst udbytte.
Vælg " Table of Means" hvorved der fremkommer følgende udskrift:

```
Table of Least Squares Means for udbytte with 95,0 Percent Confidence Intervals
--------------------------------------------------------------------------------
                                          Stnd.        Lower        Upper
Level                   Count    Mean      Error        Limit        Limit
--------------------------------------------------------------------------------
GRAND MEAN              12        25,9167      1,03609      23,4667      28,3666
katalys
c1                      6         22,8943      1,50876      19,3267      26,462
c2                      6         28,939       1,50876      25,3713      32,5067
raamat
r1                      6         24,3438      1,49172      20,8164      27,8711
r2                      6         27,4896      1,49172      23,9622      31,017
katalys by raamat
c1         r1          3        26,3348      2,45633      20,5265      32,1431
c1         r2          3        19,4539      2,15716      14,353       24,5548
c2         r1          3        22,3527      2,20695      17,1341      27,5713
c2         r2          3        35,5253      2,07257      30,6244      40,4262
--------------------------------------------------------------------------------
Means of Quantitative Factors
----------------------------
T                      4,41667
```

Da udbyttet afhænger af temperaturen *T*, må de forskellige niveauer for R og C sammenlignes ved samme temperatur *T*. I denne udskrift sker sammenligningen af udbyttet ved en *T* værdi på 4.41667 (gennemsnittet af alle *T*-værdier).
Det ses, at et valg af katalysator c2 og raamateriale r2 giver det største udbytte (værdi 35.53)
Da konfidensintervallet svagt overlapper med kombinationen c1 r1, kan det dog ikke afvises, at denne kombination også kan vælges.

6.4 TOSIDET KOVARIANSANALYSE MED 2 KVALITATIVE FAKTORER OG EN BLOKFAKTOR.

Eksempel 6.3. (tosidet kovariansanalyse) (hentet fra SFIII eksempel 23 side 128).
Det i afsnit 6.3 omtalte forsøg ændres nu, idet det antages, at resultaterne hidrører fra et randomiseret blokforsøg med tre blokke, hvor øverste tal i en celle hører til blok 1, den midterste til blok2 og den nederste til blok 3. Skemaet ændrer sig følgelig til:

			C: (katalysatortype)			
		Blok	C_1		C_2	
			y	T	y	T
R (råmateriale)	R_1	b1	21	4	27	8
		b2	19	0	20	3
		b3	20	1	31	7
	R_2	b1	10	2	38	6
		b2	26	6	40	5
		b3	31	9	28	2

Med de samme forudsætninger som i afsnit 6.3 ønskes igen en kovariansanalyse af, om udbyttet afhænger af de to faktorer.

LØSNING:
Spørgsmål 1. Vurdering af om forudsætninger er opfyldt.
Indtastning af data
Tallene indtastes som i afsnit 6.3, idet der dog er tilføjet en ekstra blok kolonne.

```
raavare   katalys   T      y    blok
  r1        c1      4     21     b1
  r1        c1      0     19     b2
  r1        c1      1     20     b3
  r1        c2      8     27     b1
  r1        c2      3     20     b2
  r1        c2      7     31     b3
  r2        c1      2     10     b1
  r2        c1      6     26     b2
  r2        c1      9     31     b3
  r2        c2      6     38     b1
  r2        c2      5     40     b2
  r2        c2      2     28     b3
```

Statistisk analyse

Vælg ikonen "GLM" eller (Special | Advanced Regression | General Linear Models" (GLM)).
På det følgende skema indsættes "Dependent variables" : y
"Categorical Factors" : raavare , katalys og blok
og "Quantitative Factors" : T (benævnes D)
I "GLM specifications" har faktorerne fået betegnelser efter den rækkefølge de er indskrevet i (A for raavare, B for katalys, C for blok og D for T). Vi tilføjer A*B (vekselvirkning mellem raavare og katalys).
Vi foretager ikke her en større analyse af om linierne er parallelle, da det allerede er sket i afsnit 6.3. Der fremkommer følgende udskrift:

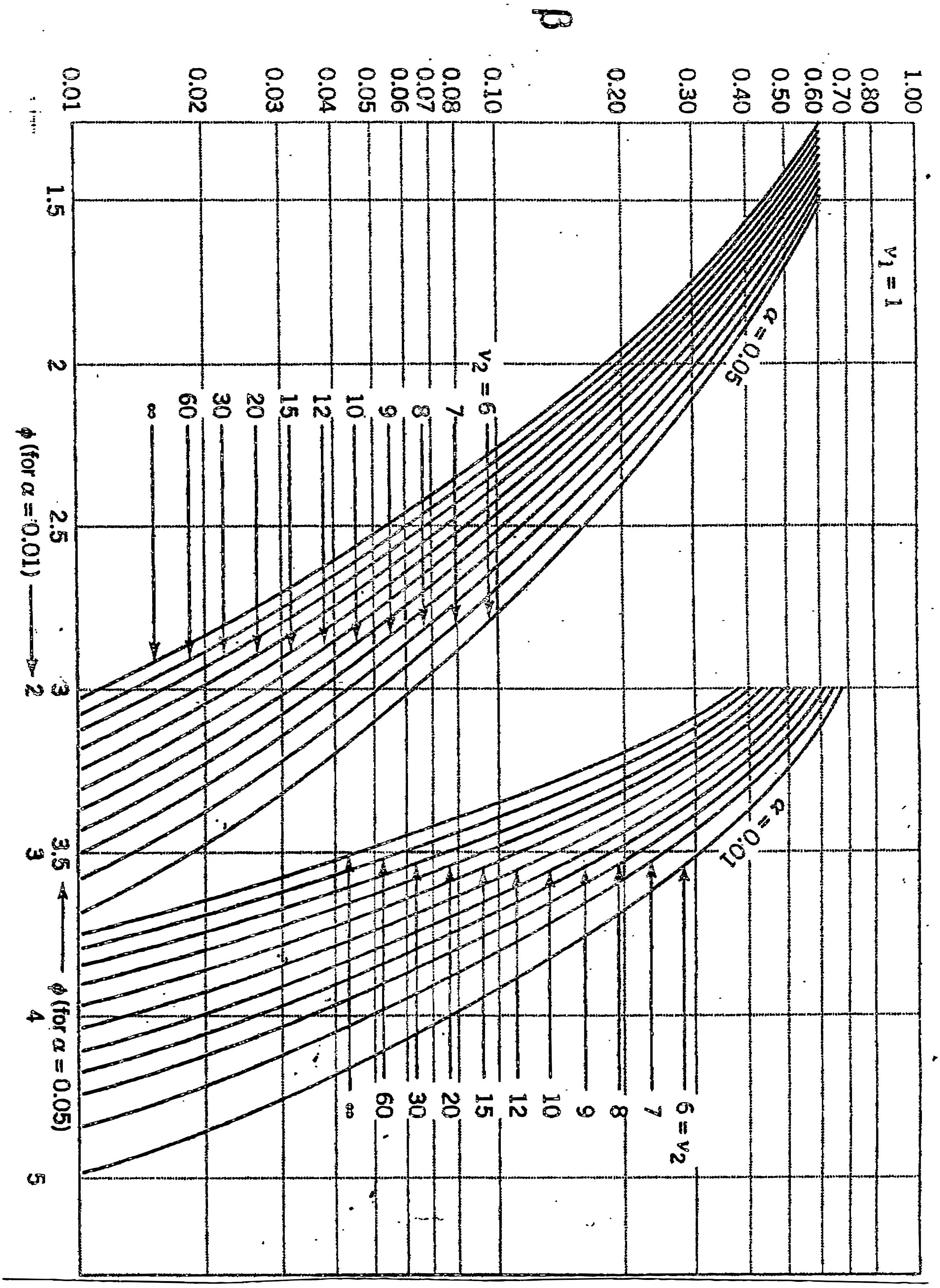
```
General Linear Models
---------------------
Number of dependent variables: 1
Number of categorical factors: 3
Number of quantitative factors: 1
Analysis of Variance for y
-----------------------------------------------------------------------------
Source                Sum of Squares     Df  Mean Square     F-Ratio      P-Value
-----------------------------------------------------------------------------
Model                        780,553      6      130,092       39,75       0,0005
Residual                     16,3633      5      3,27265
-----------------------------------------------------------------------------
Total (Corr.)                796,917     11
Type III Sums of Squares
---------------------------------------------------------------------------------
Source                         Sum of Squares     Df   Mean Square    F-Ratio     P-Value
---------------------------------------------------------------------------------
raavare                              21,6227      1        21,6227       6,61      0,0500
katalys                              82,6875      1        82,6875      25,27      0,0040
blok                                 73,8094      2        36,9047      11,28      0,0140
T                                    345,803      1        345,803     105,66      0,0001
raavare*katalys                      241,401      1        241,401      73,76      0,0004
Residual                             16,3633      5        3,27265
---------------------------------------------------------------------------------
Total (corrected)                    796,917     11
All F-ratios are based on the residual mean square error.
```

Det ses (ikke overraskende), at der er vekselvirkning mellem raamaterialer og katalysatortype, da *P*-value er 0.0004.
Endvidere ses det, at det var fornuftigt at indføre blokfaktoren , da *P*-value for blok er 0.014.
Konklusionerne er ellers som i afsnit 6.3.

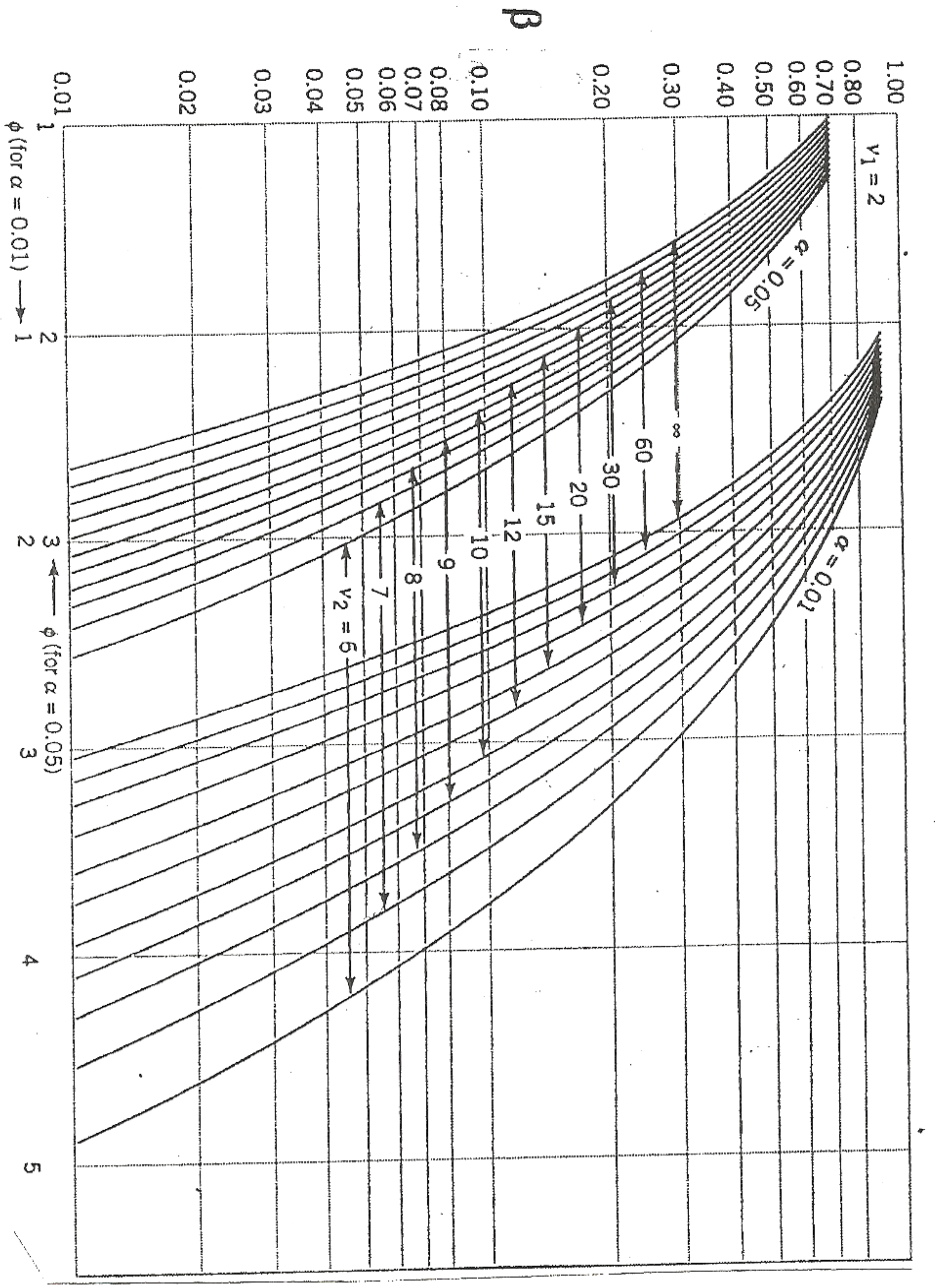
◆

Øvelse 6.1 (kovariansanalyse) Regn opgave 41 side 255.

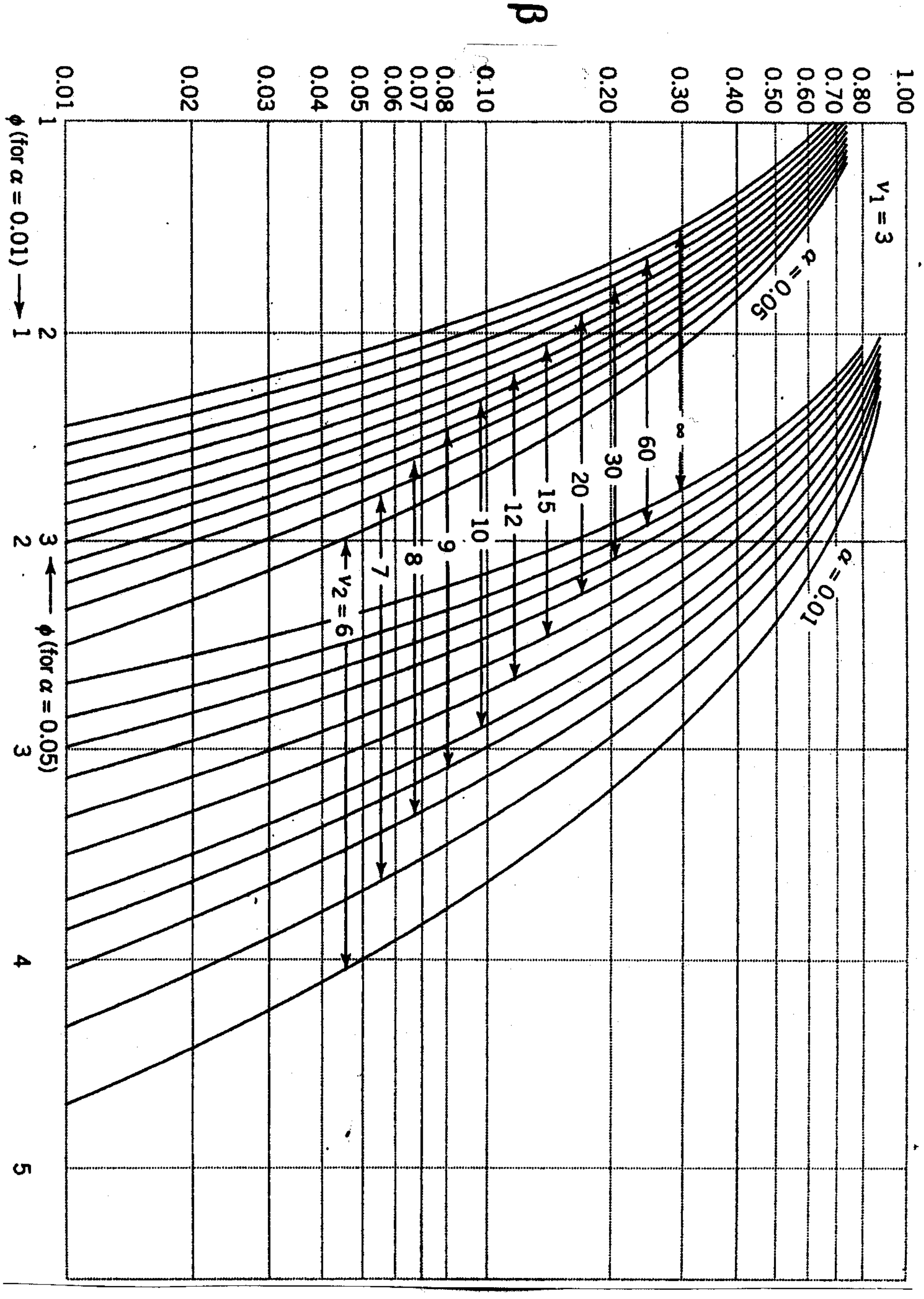
Kurve nr 1: $<_1 = 1$ er frihedsgrad for tæller og $<_2$ for nævner.



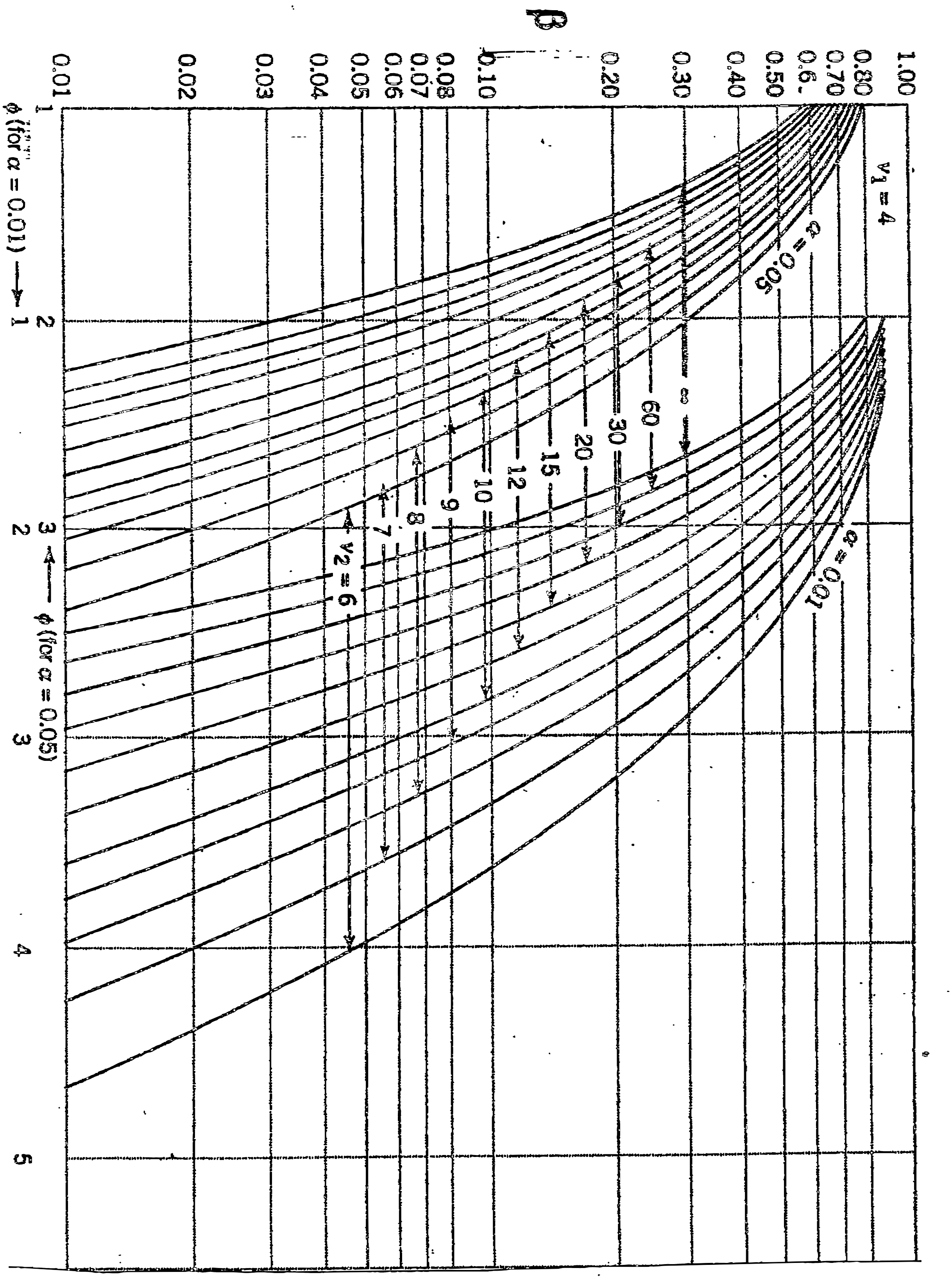
Kurve nr 2: $<_1 = 2$ er frihedsgrad for tæller og $<_2$ for nævner



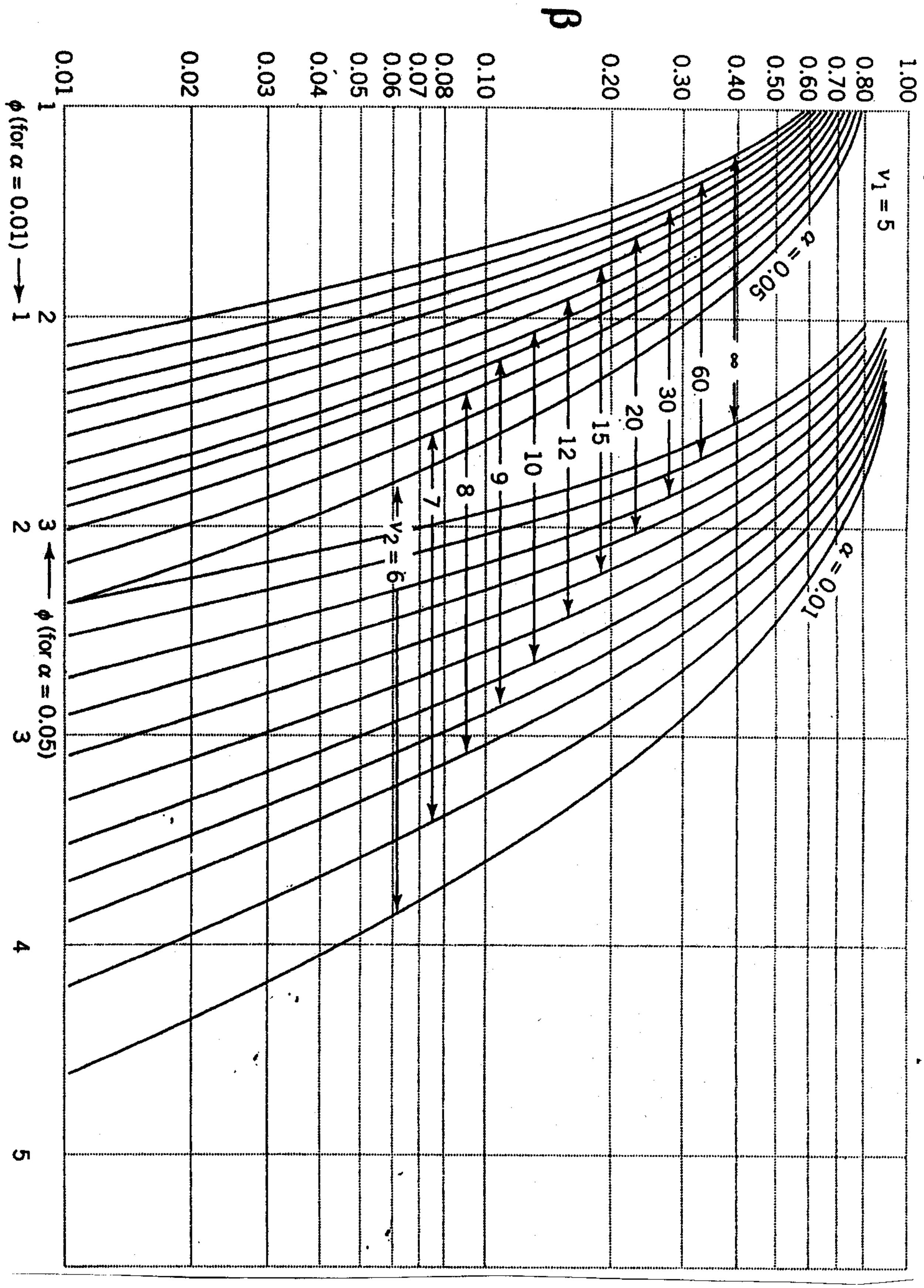
Kurve nr 3: $<_1 = 3$ **er frihedsgrad for tæller og** $<_2$ **for nævner**.



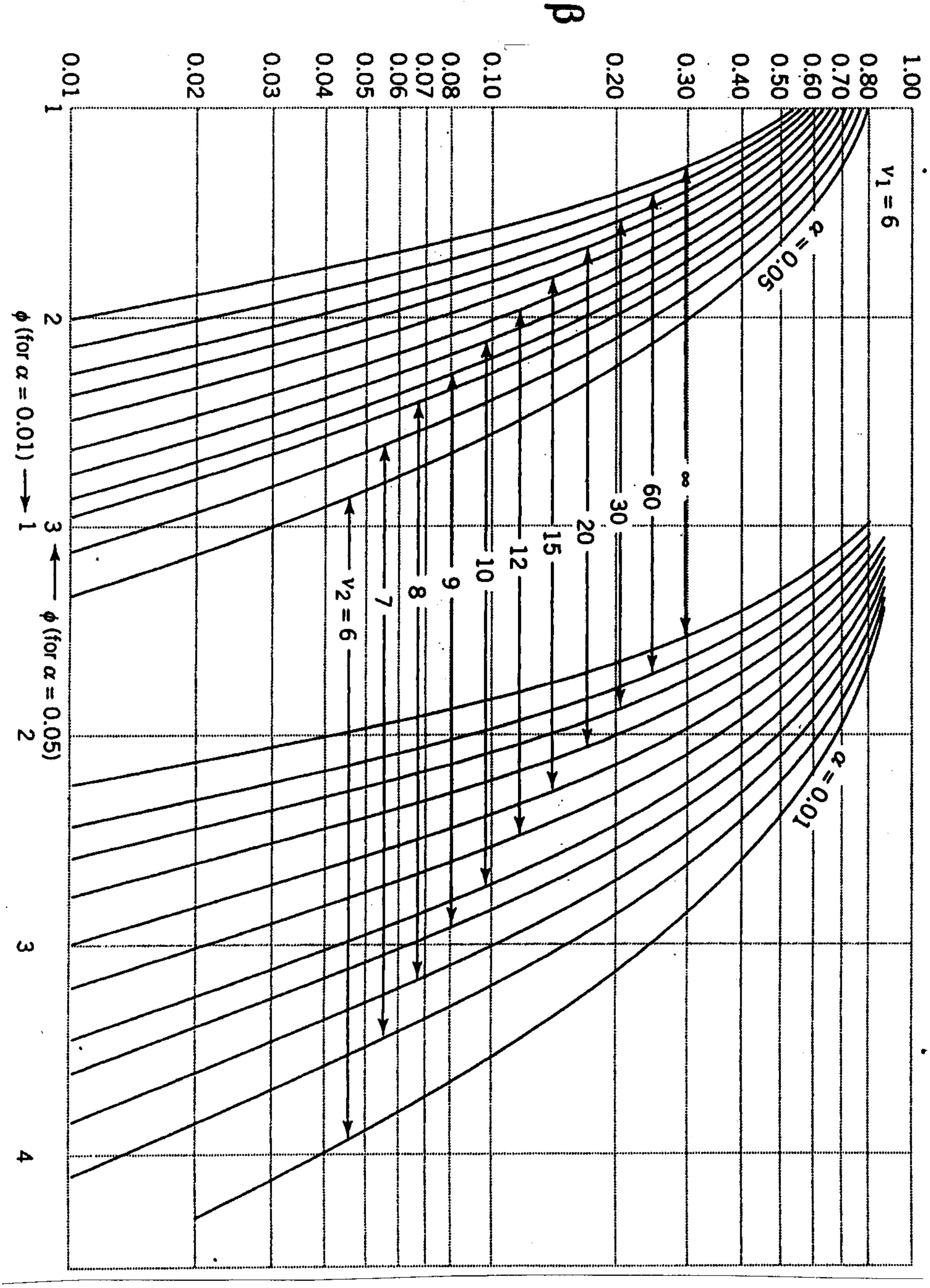
Kurve nr 4: $<_1 = 4$ **er frihedsgrad for tæller og** $<_2$ **for nævner**.



Kurve nr 5: $<_1 = 5$ **er frihedsgrad for tæller og** $<_2$ **for nævner**.

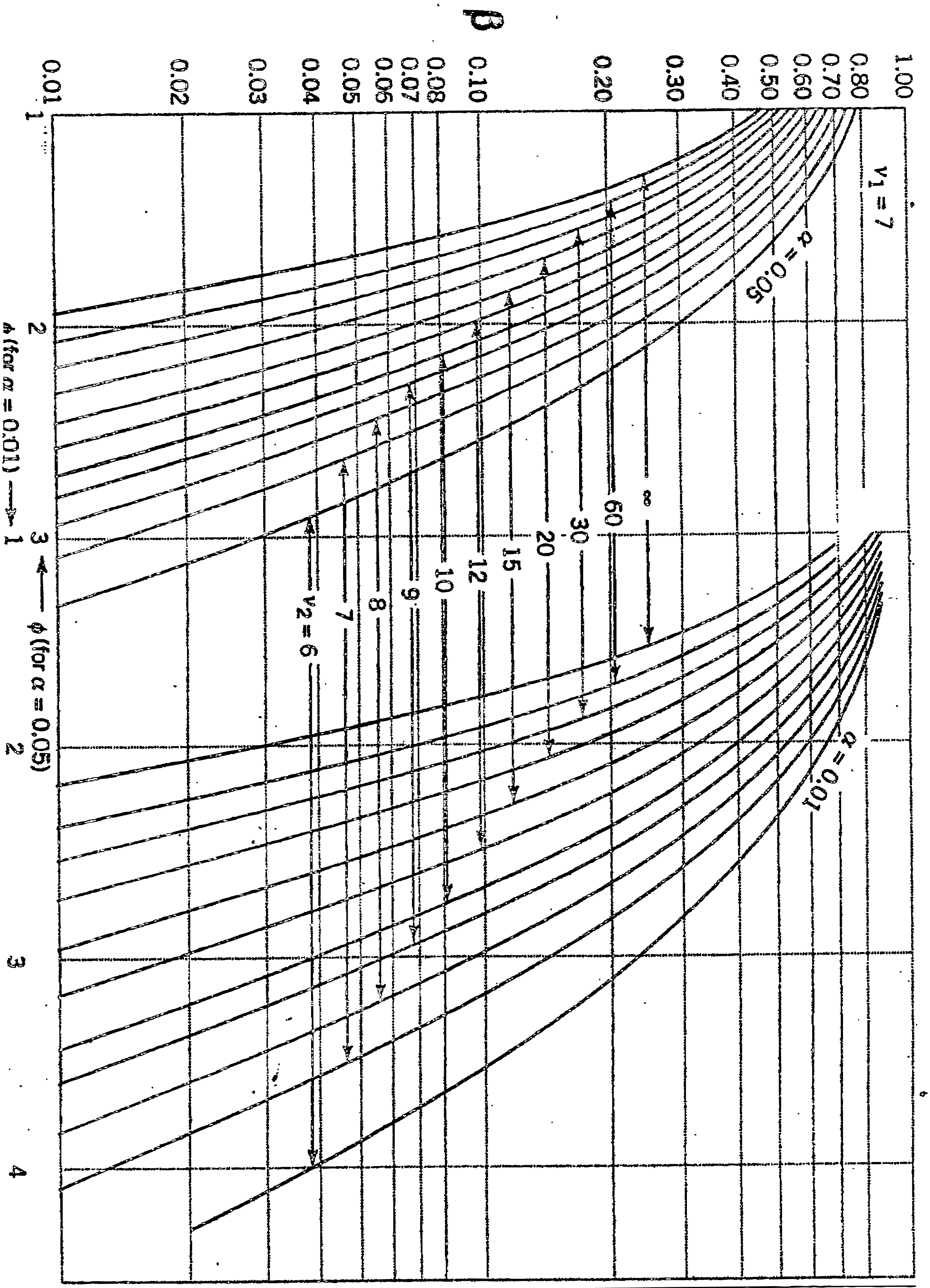


Kurve nr 6: $<_1 = 6$ er frihedsgrad for tæller og $<_2$ for nævner.

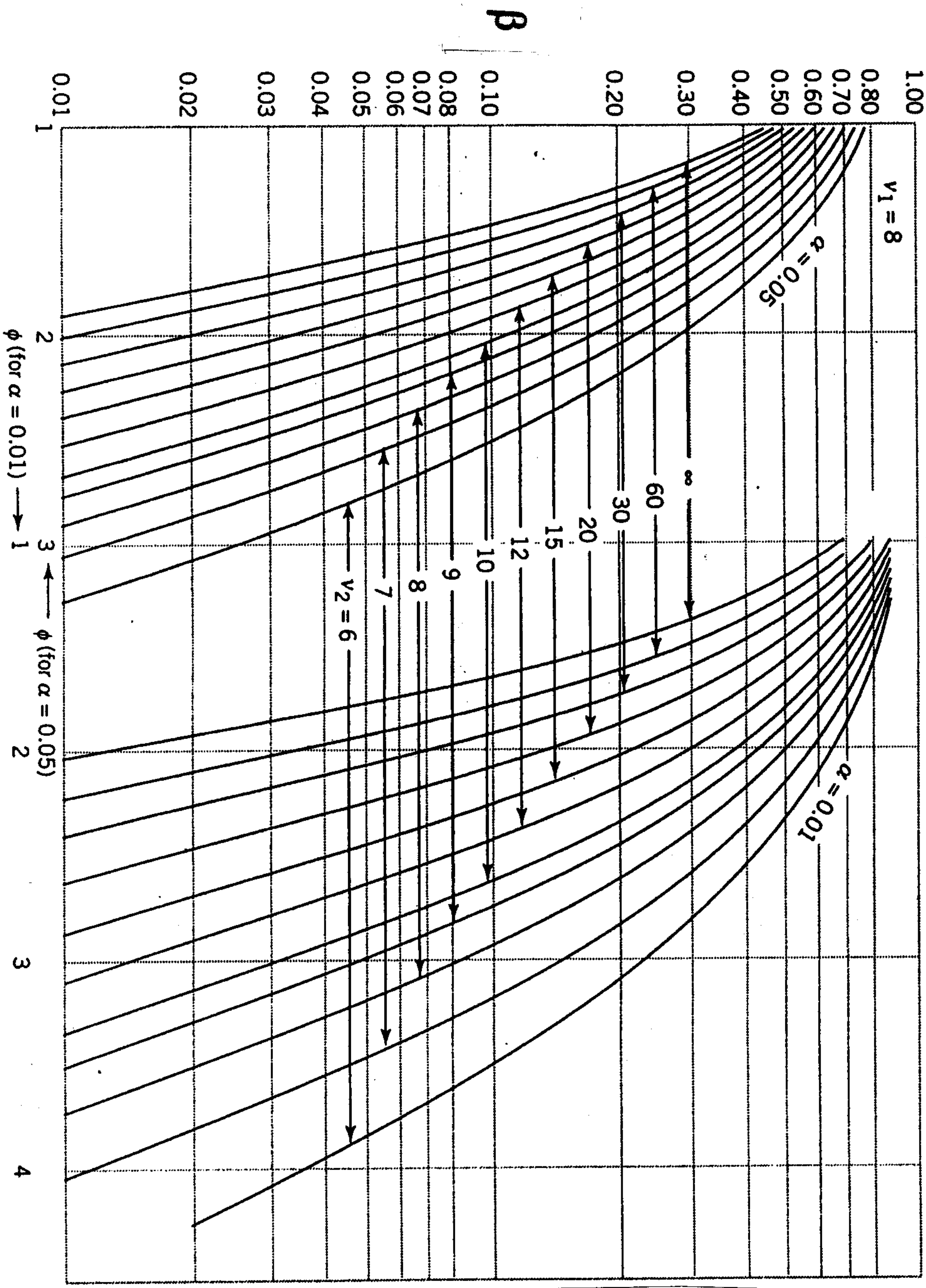


Appendix 1.1. OC-kurver for "Fixed factor" variansanalyse

Kurve nr 7: <$_1$ = 7 er frihedsgrad for tæller og <$_2$ for nævner.



Kurve nr 8: $<_1 = 8$ er frihedsgrad for tæller og $<_2$ for nævner.



APPENDIX 1.2 Beregning af tosidet variansanalyse ved brug af lommeregner

Eksempel 1.10 (tosidet variansanalyse). *En bilfabrikant ønsker at finde ud af, hvorledes* 3 *olieblandinger O_1, O_2, og O_3, og* 2 *karburatortyper K_1 og K_2 påvirker benzinforbruget.*
Lad forsøgsresultaterne være følgende:

		Karburator	
		K_1	K_2
Olieblanding	O_1	830 860	810 840
	O_2	940 990	1050 1020
	O_3	855 815	930 910

Undersøg hvilken kombination af karburator og olieblanding der giver det laveste forbrug.

a) Opstille af variansanalysetabel og uddrage konklusion

SUMSKEMA		Karburator		
		K_1	K_2	Rækkesum
Olieblanding	O_1	1690	1650	3340
	O_2	1930	2070	4000
	O_3	1670	1840	3510
	Søjlesum	5290	5560	

Antal rækker r = 3, Antal delforsøg i række =4.
Antal søjler q = 2, Antal delforsøg i søjle =6
Antal celler $r \cdot q = 6$, Antal delforsøg i celler n =2
Totalt antal = $r \cdot q \cdot n = 12$.

$$SAK_{rækker} = (r-1) \cdot \frac{s^2_{rækkesummer}}{\text{antal delforsøg i række}} = (3-1) \cdot \frac{342.68^2}{4} = 58716.67 \quad , \quad f_R = 2$$

$$SAK_{søjler} = (q-1) \cdot \frac{s^2_{søjlesummer}}{\text{antal delforsøg i søjle}} = (2-1) \cdot \frac{190.92^2}{6} = 6075.00 \quad , \quad f_C = 1$$

$$SAK_{celler} = (r \cdot q-1) \cdot \frac{s^2_{cellesummer}}{\text{antal delforsøg i celle}} = (6-1) \cdot \frac{168.804^2}{2} = 71241.66 \quad , \quad f_{celler} = 5$$

$$SAK_{vekselvirkning} = SAK_{celler} - SAK_{rækker} - SAK_{søjle} = 6450.00 \quad , \quad f_{RC} = 2$$

$$SAK_{Total} = (r \cdot q \cdot n-1) \cdot s^2_{Total} = (12-1) \cdot 82.485^2 = 74841.66 \quad , \quad f_{Total} = 11$$

Resultaterne samles i en variansanalysetabel:

Variation	f	SS	s^2	F
Olieblanding	2	58716.67	29358.33	
Karburator	1	6075.00	6075.00	
Vekselvirkning	2	6450.00	3225.00	5.38
Gentagelser	6	3600.00	600.00	
Total	11	74841.67		

Anvender vi som sædvanlig et 5% signifikansniveau, så slutter vi, at der er et svagt bevis for vekselvirkning mellem karburator og olieblanding.

b) Opstille konfidensintervaller og drage konklusion
Vi ønsker nu for en given olieblanding at finde hvilken karburator, der giver det laveste benzinforbrug.
Vi laver derfor 95% konfidensintervaller for hver af de 6 behandlinger. Idet antal gentagelser af hver celle er n=2 og forsøgsfejlens varians er $s^2_{gentagelser} = s_0^2 = 600$ med 6 frihedsgrader, fås:

$$\bar{x} \pm t_{0.975}(f_{\text{residual(poolet)}}) \frac{s_{\text{residual(poolet)}}}{\sqrt{\text{antal addender i } \bar{x}}} = \bar{x} \pm t_{0.975}(6)\frac{s_0}{\sqrt{n}} = \bar{x} \pm 2.45\frac{\sqrt{600}}{\sqrt{2}} = \bar{x} \pm 42$$

dvs. konfidensintervallerne er $[\bar{x} - 42 ;\ \bar{x} + 42]$.

I nedenstående skema er igen angivet gennemsnittene.

		Karburator	
		K_1	K_2
Olieblanding	O_1	845	825
	O_2	965	1035
	O_3	835	920

Vi får da konfidensintervallerne:

		Karburator	
		K_1	K_2
Olieblanding	O_1	[803;887]	[783;867]
	O_2	[923;1007]	[993;1077]
	O_3	[793;877]	[878;962]

Af tabellen ovenfor sluttes, at man kan enten bruge olieblanding O_1 og så er det ligegyldigt hvilken karburator man benytter, eller også skal man benytte olieblanding O_3 og Karburator K_1.

Bjarne Hellesen:

APPENDIX 1.3
HVORDAN UDREGNES SAK TYPE I OG TYPE III ?

Observationer:		Karburator	
		C_1	C_2
Olieblanding	R_1	$x_{111} = 830$, $x_{112} = 860$	$x_{121} = 810$, $x_{122} = 840$
	R_2	$x_{211} = 940$, $x_{212} = 990$	$x_{221} = 1050$, $x_{222} = 1020$
	R_3	$x_{311} = 855$	$x_{321} = 930$, $x_{322} = 910$

Antag: x_{ijm} er statistisk uafhængige observationer af statistisk uafhængige variable X_{ij} med middelværdier $E(X_{ij}) = \mu_{ij}$ og samme varians $V(X_{ij}) = \sigma^2$ (dvs. $x_{ijm} = \mu_{ij} + \overset{\text{støj}}{e_{ijm}}$, $E(e_{ijm}) = 0$, $V(e_{ijm}) = \sigma^2$).

Middelværdier:

$$\overline{\mu_{i\cdot}} \equiv \frac{\mu_{i1} + \mu_{i2}}{2}, \quad \overline{\mu_{\cdot j}} \equiv \frac{\mu_{1j} + \mu_{2j} + \mu_{3j}}{3}, \quad \overline{\mu_{\cdot\cdot}} \equiv \frac{\mu_{11} + \mu_{12} + \mu_{21} + \mu_{22} + \mu_{31} + \mu_{32}}{6}.$$

Omskrivning:

$$\begin{aligned}\mu_{ij} &= \overline{\mu_{\cdot\cdot}} + (\overline{\mu_{i\cdot}} - \overline{\mu_{\cdot\cdot}}) + (\overline{\mu_{\cdot j}} - \overline{\mu_{\cdot\cdot}}) + (\mu_{ij} - \overline{\mu_{i\cdot}} - \overline{\mu_{\cdot j}} + \overline{\mu_{\cdot\cdot}}) \\ &\equiv k + R_i + C_j + RC_{ij} \\ &\equiv \text{konstant} + \text{Rækkevirkning} + \text{Søjlevirkning} + \text{Vekselvirkning} .\end{aligned}$$

Uafhængige størrelser

Der gælder åbenbart $\sum_{i=1}^{3} R_i = 0$, $\sum_{j=1}^{2} C_j = 0$, $\sum_{j=1}^{2} RC_{ij} = 0$, $\sum_{i=1}^{3} RC_{ij} = 0$. Så der bliver kun 6 uafhængige størrelser $(k,\ R_1,\ R_2,\ C_1,\ RC_{11},\ RC_{21}) = \underline{z}$, ligesom der er 6 stk. μ_{ij} :

<u>Bevis</u>. Vi viser, at $R_3,\ C_2,\ RC_{31},\ RC_{12},\ RC_{22},\ RC_{32}$ kan elimineres.

(1) $R_1 + R_2 + R_3 = 0 \Rightarrow R_3 = -R_1 - R_2$, som bruges til at eliminere R_3

(2) $C_1 + C_2 = 0 \Rightarrow C_2 = -C_1$, som bruges til at eliminere C_2

(3) $RC_{11} + RC_{21} + RC_{31} = 0 \Rightarrow RC_{31} = -RC_{11} - RC_{21}$, som bruges til at eliminere RC_{31}

(4) $RC_{11} + RC_{12} = 0 \Rightarrow RC_{12} = -RC_{11}$, som bruges til at eliminere RC_{12}

(5) $RC_{21} + RC_{22} = 0 \Rightarrow RC_{22} = -RC_{21}$, som bruges til at eliminere RC_{22}

(6) $RC_{31} + RC_{32} = 0 \Rightarrow RC_{32} = RC_{11} + RC_{21}$, som bruges til at eliminere RC_{32}

(7) ($RC_{12} + RC_{22} + RC_{32} = 0$ giver intet nyt, da den fås af (4) + (5) + (6) - (3)). ■

Ligningssystem

Observationerne x_{ijm} kan skrives $x_{ijm} = \mu_{ij} + e_{ijm}$. Bortkastes støjen e_{ijm}, fås $\mu_{ij} \approx x_{ijm}$ eller $k + R_i + C_j + RC_{ij} \approx x_{ijm}$:

$$
\begin{array}{lllllllllll}
k & + & R_1 & + & C_1 & + & RC_{11} & \approx & x_{111} & = & 830 \\
k & + & R_1 & + & C_1 & + & RC_{11} & \approx & x_{112} & = & 860 \\
k & + & R_1 & + & C_2 & + & RC_{12} & \approx & x_{121} & = & 810 \\
k & + & R_1 & + & C_2 & + & RC_{12} & \approx & x_{122} & = & 840 \\
k & + & R_2 & + & C_1 & + & RC_{21} & \approx & x_{211} & = & 940 \\
k & + & R_2 & + & C_1 & + & RC_{21} & \approx & x_{212} & = & 990 \\
k & + & R_2 & + & C_2 & + & RC_{22} & \approx & x_{221} & = & 1050 \\
k & + & R_2 & + & C_2 & + & RC_{22} & \approx & x_{222} & = & 1020 \\
k & + & R_3 & + & C_1 & + & RC_{31} & \approx & x_{311} & = & 855 \\
k & + & R_3 & + & C_2 & + & RC_{32} & \approx & x_{321} & = & 930 \\
k & + & R_3 & + & C_2 & + & RC_{32} & \approx & x_{322} & = & 910 \quad .
\end{array}
$$

Her kan R_3, C_2, RC_{12}, RC_{22}, RC_{32}, RC_{31} elimineres (som vist ovenfor), så vi får

$$
\begin{array}{lllllllllll}
k & + & R_1 & & & + & C_1 & + & RC_{11} & & & \approx & 830 \\
k & + & R_1 & & & + & C_1 & + & RC_{11} & & & \approx & 860 \\
k & + & R_1 & & & - & C_1 & - & RC_{11} & & & \approx & 810 \\
k & + & R_1 & & & - & C_1 & - & RC_{11} & & & \approx & 840 \\
k & & & + & R_2 & + & C_1 & & & + & RC_{21} & \approx & 940 \\
k & & & + & R_2 & + & C_1 & & & + & RC_{21} & \approx & 990 \\
k & & & + & R_2 & - & C_1 & & & - & RC_{21} & \approx & 1050 \\
k & & & + & R_2 & - & C_1 & & & - & RC_{21} & \approx & 1020 \\
k & - & R_1 & - & R_2 & + & C_1 & - & RC_{11} & - & RC_{21} & \approx & 855 \\
k & - & R_1 & - & R_2 & - & C_1 & + & RC_{11} & + & RC_{21} & \approx & 930 \\
k & - & R_1 & - & R_2 & - & C_1 & + & RC_{11} & + & RC_{21} & \approx & 910 \quad .
\end{array}
$$

Dette er et overbestemt ligningssystem $A\bar{z} = \bar{b}$:

$$
\begin{pmatrix}
1 & 1 & 0 & 1 & 1 & 0 \\
1 & 1 & 0 & 1 & 1 & 0 \\
1 & 1 & 0 & -1 & -1 & 0 \\
1 & 1 & 0 & -1 & -1 & 0 \\
1 & 0 & 1 & 1 & 0 & 1 \\
1 & 0 & 1 & 1 & 0 & 1 \\
1 & 0 & 1 & -1 & 0 & -1 \\
1 & 0 & 1 & -1 & 0 & -1 \\
1 & -1 & -1 & 1 & -1 & -1 \\
1 & -1 & -1 & -1 & 1 & 1 \\
1 & -1 & -1 & -1 & 1 & 1
\end{pmatrix}
\begin{pmatrix}
k \\ R_1 \\ R_2 \\ C_1 \\ RC_{11} \\ RC_{21}
\end{pmatrix}
=
\begin{pmatrix}
830 \\ 860 \\ 810 \\ 840 \\ 940 \\ 990 \\ 1050 \\ 1020 \\ 855 \\ 930 \\ 910
\end{pmatrix} .
$$

Ligningssystemet har totalmatrix $T = (A|\overline{b})$:

$$T = \begin{array}{c} \\ \\ \end{array} \begin{array}{cccccccc} k & R_1 & R_2 & C_1 & RC_{11} & RC_{21} & & \overline{b} \end{array}$$

$$T = \begin{pmatrix} 1 & 1 & 0 & 1 & 1 & 0 & | & 830 \\ 1 & 1 & 0 & 1 & 1 & 0 & | & 860 \\ 1 & 1 & 0 & -1 & -1 & 0 & | & 810 \\ 1 & 1 & 0 & -1 & -1 & 0 & | & 840 \\ 1 & 0 & 1 & 1 & 0 & 1 & | & 940 \\ 1 & 0 & 1 & 1 & 0 & 1 & | & 990 \\ 1 & 0 & 1 & -1 & 0 & -1 & | & 1050 \\ 1 & 0 & 1 & -1 & 0 & -1 & | & 1020 \\ 1 & -1 & -1 & 1 & -1 & -1 & | & 855 \\ 1 & -1 & -1 & -1 & 1 & 1 & | & 930 \\ 1 & -1 & -1 & -1 & 1 & 1 & | & 910 \end{pmatrix} .$$

Beregning af SAK-værdi

1) Først benyttes mindste kvadraters metode (least squares) til at finde en tilnærmet løsning $\overline{z}$ til det overbestemte ligningssystem $A\overline{z} = \overline{b}$ (jævnfør "Matematik for ingeniører" bind3, afsnit 3.6).

2) Dernæst beregnes afvigelserne (residualerne) $\overline{r} = A\overline{z} - \overline{b} = (r_1, r_2, \ldots, r_{11})$. De 11 afvigelser viser, hvor godt løsningen $\overline{z}$ opfylder de 11 ligninger.

3) Til sidst findes **S**ummen af **A**fvigelsernes **K**vadrater, $SAK = r_1^2 + r_2^2 + \ldots + r_{11}^2 = \overline{r} \cdot \overline{r}$, som løsningsmetoden har gjort mindst mulig.

Benyttes den viste matrix *A*, fås $SAK_{Residual} = 2800.00$. For at finde SAK_{Total} udelades alle virkningerne *R*, *C* og *RC*. Herved slettes de tilsvarende søjler i matricen *A*, og kun den første søjle bliver tilbage. Der er nu kun én variabel k til at opfylde de 11 ligninger, så løsningen $\overline{z} = (k, 0, 0, 0, 0, 0)$ giver en meget større $SAK = 66168.18182$ $(= SAK_{Total})$. Hvis vi igen tilføjer virkningerne én ad gangen, vil vi hver gang se en formindskelse af ligningssystemets *SAK*.

For en virkning E defineres:

SAK af type I = formindskelsen af ligningssystemets *SAK*, når E tilføjes som den **næste** virkning (næste linie i variansanalysetabellen).

SAK af type III = formindskelsen af ligningssystemets *SAK*, når *E* tilføjes som den **sidste** virkning.

Variansanalysetabeller

Nu kan vi efterregne Statgraphics` SAK'er i de 6 variansanalysetabeller fra afsnit 1.2.7 , f.eks. ved at bruge matematikprogrammet Maple, hvorfra udskrift er vist senere. Vi finder følgende SAK'er.

Tabel med type III SAK'er

SAK_R = SAK(med C, RC) − SAK(med R, C, RC) = 56250.00
SAK_C = SAK(med R, RC) − SAK(med R, C, RC) = 3778.57
SAK_{RC} = SAK(med R, C) − SAK(med R, C, RC) = 4850.00
$SAK_{Residual}$ = SAK(med R, C, RC) = 2800.00
SAK_{Total} = SAK(med) = 66168.18 ($\neq$ sum af øvrige for type III)

Tabel med type III SAK'er, når RC er bortkastet

SAK_R = SAK(med C) − SAK(med R, C) = 55783.33
SAK_C = SAK(med R) − SAK(med R, C) = 3266.67
$SAK_{Residual}$ = SAK(med R, C) = 7650.00
SAK_{Total} = SAK(med) = 66168.18 ($\neq$ sum af øvrige for type III)

Tabel med type I SAK'er, når RC er bortkastet

SAK_R = SAK(med) − SAK(med R) = 55251.52
SAK_C = SAK(med R) − SAK(med R, C) = 3266.67
$SAK_{Residual}$ = SAK(med R, C) = 7650.00
SAK_{Total} = SAK(med) = 66168.18 (= sum af øvrige for type I)

Tabel med type I SAK'er, når RC er bortkastet (ny rækkefølge ændrer SAK'er)

SAK_C = SAK(med) − SAK(med C) = 2734.85
SAK_R = SAK(med C) − SAK(med R, C) = 55783.33
$SAK_{Residual}$ = SAK(med R, C) = 7650.00
SAK_{Total} = SAK(med) = 66168.18 (= sum af øvrige for type I)

Tabel med type III SAK'er, når RC er bortkastet (ny rækkefølge ændrer **ikke** SAK'er)

SAK_C = SAK(med R) − SAK(med R, C) = 3266.67
SAK_R = SAK(med C) − SAK(med R, C) = 55783.33
$SAK_{Residual}$ = SAK(med R, C) = 7650.00
SAK_{Total} = SAK(med) = 66168.18 ($\neq$ sum af øvrige for type III)

Tabel med type III SAK'er, når C og RC er bortkastet

SAK_R = SAK(med) − SAK(med R) = 55251.52
$SAK_{Residual}$ = SAK(med R) = 10916.67
SAK_{Total} = SAK(med) = 66168.18

APPENDIX 1.3.Hvordan udregnes SAK af type I og type III

```
> restart; with(linalg):   #  M A P L E :   FINDS ANOVA TABLES  FROM  OVERDETERMINED  SYSTEMS  Az=b
> SAK:= proc(A,b::array) local z,r;                         # Function "SAK" = min(sum of (residuals^2))
>       z:= leastsqrs(A,b);                                  # Best solution z=estimated effects=(k,R1,R2,C1,RC11,RC21)
>       r:= multiply(A,z) - b;                               # Residual vector r = Az-b
>       evalf(dotprod(r,r))  end:                            # Sum of squares = SAK = r^2.     End of function
> T := concat([1,1,1,1,1,1,1,1,1,1,1],[1,1,1,1,0,0,0,0,-1,-1,-1],      # Total matrix
> [0,0,0,0,1,1,1,1,-1,-1,-1],[1,1,-1,-1,1,1,-1,-1,1,-1,-1],[1,1,-1,-1,0,0,0,0,-1,1,1],
> [0,0,0,0,1,1,-1,-1,-1,1,1],[830,860,810,840,940,990,1050,1020,855,930,910]);
```

$$T := \begin{bmatrix} 1 & 1 & 0 & 1 & 1 & 0 & 830 \\ 1 & 1 & 0 & 1 & 1 & 0 & 860 \\ 1 & 1 & 0 & -1 & -1 & 0 & 810 \\ 1 & 1 & 0 & -1 & -1 & 0 & 840 \\ 1 & 0 & 1 & 1 & 0 & 1 & 940 \\ 1 & 0 & 1 & 1 & 0 & 1 & 990 \\ 1 & 0 & 1 & -1 & 0 & -1 & 1050 \\ 1 & 0 & 1 & -1 & 0 & -1 & 1020 \\ 1 & -1 & -1 & 1 & -1 & -1 & 855 \\ 1 & -1 & -1 & -1 & 1 & 1 & 930 \\ 1 & -1 & -1 & -1 & 1 & 1 & 910 \end{bmatrix}$$

```
> _k:=col(T,1):  _R:=col(T,2..3):  _C:=col(T,4):  _RC:=col(T,5..6):  b:=col(T,7):
> SAK_R    := SAK(concat(_k,    _C,_RC),b) - SAK(concat(_k,_R,_C,_RC),b);      # type III:
> SAK_C    := SAK(concat(_k,_R,    _RC),b) - SAK(concat(_k,_R,_C,_RC),b);
> SAK_RC   := SAK(concat(_k,_R,_C     ),b) - SAK(concat(_k,_R,_C,_RC),b);
> SAK_Resid:= SAK(concat(_k,_R,_C,_RC),b);                                    # SAK type III = decrease of SAK
> SAK_Total:= SAK(concat(_k          ),b);                                    #  when effect added as last one
                          SAK_R := 56250.
                          SAK_C := 3778.571429
                          SAK_RC := 4850.
                          SAK_Resid := 2800.
                       SAK_Total := 66168.18182
> SAK_R    := SAK(concat(_k,    _C),b) - SAK(concat(_k,_R,_C),b);             # RC=0, type III:
> SAK_C    := SAK(concat(_k,_R    ),b) - SAK(concat(_k,_R,_C),b);
> SAK_Resid:= SAK(concat(_k,_R,_C),b);  SAK_Total:= SAK(concat(_k),b);
                          SAK_R := 55783.33333
                          SAK_C := 3266.66667
                          SAK_Resid := 7650.
                       SAK_Total := 66168.18182
> SAK_R    := SAK(concat(_k      ),b) - SAK(concat(_k,_R),b);                 # RC=0, type I:
> SAK_C    := SAK(concat(_k,_R    ),b) - SAK(concat(_k,_R,_C),b);
> SAK_Resid:= SAK(concat(_k,_R,_C),b);                                        # SAK type I  =  decrease of SAK
> SAK_Total:= SAK(concat(_k      ),b);                                        #  when effect added as next one
                          SAK_R := 55251.51515
                          SAK_C := 3266.66667
                          SAK_Resid := 7650.
                       SAK_Total := 66168.18182
> SAK_C    := SAK(concat(_k      ),b) - SAK(concat(_k,    _C),b);             # RC=0, type I:
> SAK_R    := SAK(concat(_k,    _C),b) - SAK(concat(_k,_R,_C),b);
> SAK_Resid:= SAK(concat(_k,_R,_C),b);  SAK_Total:= SAK(concat(_k),b);
                          SAK_C := 2734.84849
                          SAK_R := 55783.33333
                          SAK_Resid := 7650.
                       SAK_Total := 66168.18182
> SAK_C    := SAK(concat(_k,_R    ),b) - SAK(concat(_k,_R,_C),b);             # RC=0, type III:
> SAK_R    := SAK(concat(_k,    _C),b) - SAK(concat(_k,_R,_C),b);
> SAK_Resid:= SAK(concat(_k,_R,_C),b);  SAK_Total:= SAK(concat(_k),b);
                          SAK_C := 3266.66667
                          SAK_R := 55783.33333
                          SAK_Resid := 7650.
                       SAK_Total := 66168.18182
> SAK_R    := SAK(concat(_k   ),b) - SAK(concat(_k,_R),b);                    # C=RC=0, type III:
> SAK_Resid:= SAK(concat(_k,_R),b);  SAK_Total:= SAK(concat(_k),b);
                          SAK_R := 55251.51515
                       SAK_Resid := 10916.66667
                       SAK_Total := 66168.18182
```

Appendix 3.1 Varianskomponentmetoden beregnet manuelt.

Symbolik
En observation i en sidedelt variansanalyse (krydsede faktorer) kan skrives
$x_{ijk} = \mu + A_i + B_j + AB_{ij} + \varepsilon_{(ij)k}$

hvor A_i er hovedvirkningen for det i te niveau
B_j er hovedvirkningen for det j te niveau
AB_{ij} er vekselvirkningen mellem det i'te rækkeniveau og det j'te søjleniveau
og $\varepsilon_{(ij)k}$ er værdier af en normalfordelt statistisk variabel forsøgsfejlen med middelværdi 0 og varians σ^2.

En observation i en trinvis variansanalyse hvor B's niveauer er undertrin til A's niveauer kan skrives $x_{ijk} = \mu + A_i + B_{(i)j} + \varepsilon_{(ij)k}$

I $B_{(i)j}$ refererer parentesen til det trinvise *j*-niveau for fastholdt *i*.

For kortheds skyld vil vi i det følgende kalde et index i parentes for et "dødt" index, og et index der ikke er i parentes for et "levende" index.

Varianskomponentmetoden.[1]
Fremgangsmåden forstås bedst, hvis den deles op i 5 punkter.
1) Der dannes følgende tabel, hvor der er en række for hvert modelkomponent og en søjle for hvert indeks. Over hvert indeks skrives om den faktor der svarer til indekset er fixed (F) eller random (R) og hvor mange niveauer faktoren har. Gentagelser betragtes altid som random. Lad antallet af gentagelser være *n*.

	R a i	F b j	R n k
A_i			
B_j			
AB_{ij}			
$\varepsilon_{(ij)k}$			

[1]En enklere mere intuitiv huskeregel (der dog kræver nogen øvelse) er følgende (formuleret af H. Spliid og B. Hellesen):
a)Enhver virkning optræder hos sig selv (virkning = hovedvirkning eller vekselvirkning)
b)Rent fixed virkning optræder ingen andre steder.
c)Rent random virkning R giver også optræden i
c1) de virkninger, som den vekselvirker med (optræder da som denne vekselvirkning),
c2) de virkninger, hvor der allerede optræder led indeholdende "overtrin" til R (optræder da som et ekstra led, der fås ved at substituere R i stedet for dens "overtrin"). [eksempelvis optræder støjen σ_0^2 overalt, da den er undertrin til alt].
Koefficient= antal observationer pr. niveau af virkning

2) I cellerne skrives 1 hvis et af rækkekomponentens "døde" indexer svarer til indexet i søjlen.

	R a i	F b j	R n k
A_i			
B_j			
AB_{ij}			
$\varepsilon_{(ij)k}$	1	1	

3) I cellerne skrives 0 hvis et af rækkekomponentens "levende" indexer svarer til indexet i en fix søjle.
   I cellerne skrives 1 hvis et af rækkekomponentens "levende" indexer svarer til indexet i en random søjle.

	R a i	F b j	R n k
A_i	1		
B_j		0	
AB_{ij}	1	0	
$\varepsilon_{(ij)k}$	1	1	1

4) I de resterende tomme celler skrives antallet af niveauer, som står over søjlen.

	R a i	F b j	R n k
A_i	1	b	n
B_j	a	0	n
AB_{ij}	1	0	n
$\varepsilon_{(ij)k}$	1	1	1

5) Beregninger på basis af skemaet.
   For en bestemt række streges alle søjler som har samme indeks som et levende index i rækkekomponenten.
   Derefter dannes i hver af de rækker, hvor **alle** den pågældende rækkekomponents indeks forekommer, produktet af de resterende tal, og dette tal multipliceres med den til rækken hørende varianskomponent.

Reglen virker ret uoverskuelig og forstås nok bedst ved at benytte den på eksemplet ovenfor.

	R a i	F b j	R n k	$E(s^2)$
A_i	1	b	n	$E(s_A^2) = b \cdot n \cdot \sigma_A^2 + \sigma^2$
B_j	a	0	n	$E(s_B^2) = a \cdot n \cdot \sigma_B^2 + 1 \cdot n \cdot \sigma_{AB}^2 + \sigma^2$
AB_{ij}	1	0	n	$E(s_{AB}^2) = n \cdot \sigma_{AB}^2 + \sigma^2$
$\varepsilon_{(ij)k}$	1	1	1	$E(s_0^2) = \sigma^2$

En variansanalysetestning i denne situation, vil bestå i, at man først danner F størrelsen

$$F_{AB} = \frac{s_{AB}^2}{s_0^2} \quad .$$

Fås en accept af H_0 : AB = 0 kan man poole og teste om der er hovedvirkninger A og B. Fås forkastelse, kan man dog danne F størrelsen $F_B = \dfrac{s_B^2}{s_{AB}^2}$, og derved undersøge om nulhypotesen H_0 : B = 0 kan accepteres.

Det ses, at en testning af H_0 : A = 0 kan foretages ved $F_A = \dfrac{s_A^2}{s_0^2}$ uanset om AB = 0, men det har ofte ikke nogen praktisk interesse i en mixed factor analyse. Den kan vise om det er nødvendigt at tage hensyn til A’s tilstedeværelse, men A har ingen selvstændig interesse, da A som random faktor jo nærmest er at betragte som en støjfaktor.

APPENDIX 3.2

B. Hellesen

Maple program til beregning af varianskomponenter.

```
> # M A P L E   F I L E :   " E M S . M W S "                    E M S   R U L E S
> # FINDS EXPECTED MEAN SQUARES FOR ANY BALANCED FACTORIAL, NESTED, OR NESTED
> # FACTORIAL EXPERIMENT.  NOT FOR PARTIALLY BALANCED ARRANGEMENTS, SUCH AS LATIN
> # SQUARES AND INCOMPLETE BLOCK DESIGNS.   (Bjarne Hellesen, IFAK-DTU, May 1998)
> # For EMS rules see D.C.Montgomery: Design and Analysis of Experiments, Wiley, 1997,
> # page 480-485   or    C.R.Hicks: Fundamental Concepts in the Design of Experiments,
> #                  Oxford University Press, 1993, page 160-164.
> restart; # First function "EMS()" will be defined, then its USAGE is shown by examples
> EMS := proc(FactorsNoise::listlist,MyModel::algebraic) # Start of function "EMS()"
>   local c,Dead,FixLL,good,i,k,LiveDead,m,mm,MoLL,MoLP,Mo1,OutL,ncols,nrows,p,q,r,rr,t;
>   q:=[op(MyModel)]:   # sum   mu+A+B(A)+C(B,A)+D+E(D)+C(B,A)*E(D)+Eps[A,B,C,D,E]-->
>                   # list [mu,A,B(A),C(B,A),D,E(D),C(B,A)*E(D),Eps[A,B,C,D,E]]
>   Mo1:=[seq(q[i],i=2..nops(q))]: # remove mu from list
>   nrows:=nops(Mo1): ncols:=nops(FactorsNoise): # Number of components, and of factors
>   MoLL:=Mo1: MoLP:=Mo1:      # Want MoLL=[[A],[B], [C] ,[D],[E], [C,E] ,   [Eps]   ]
>   Dead:=[seq([],i=1..nrows)]: # Want Dead=[ [],[A],[B,A], [],[D],[B,A,D],[A,B,C,D,E]]
>   for r to nrows do
>     if type(Mo1[r],`*`) then  # if component is a product like C(B,A)*E(D) then
>       MoLL[r]:=[op(Mo1[r])];  #   change it to a list like [C(B,A),E(D)]
>     else
>       MoLL[r]:=[Mo1[r]];
>     fi;
>     for t to nops(MoLL[r]) do
>       if type(MoLL[r,t],function) then # if atom has round brackets like C(B,A) then
>         Dead[r]:=[op(Dead[r]),op(MoLL[r,t])]; # add B,A to components Dead list
>         MoLL[r,t]:=op(0,MoLL[r,t]);         # and keep only the living factor C
>       fi;
>     od;
>     MoLP[r]:=convert(MoLL[r],`*`);  # rewrite component as a simple product like C*E
>   od:
>   m:=array(1..nrows,1..ncols):              # create matrix m
>   for r to nrows do
>     for c to ncols do
>       m[r,c]:=FactorsNoise[c,3];        # element = number of levels for c'th factor
>       if member(FactorsNoise[c,1],Dead[r]) then # if c'th factor in Dead[r] list then
>         m[r,c] := 1;                        #   change element to 1
>       fi;
>       if member(FactorsNoise[c,1],MoLL[r]) then # if c'th factor in Live[r]list then
>         if FactorsNoise[c,2]=Fix then          #   if Fixed factor then
>           m[r,c] := 0;                       #     change element to 0
>         else                               #   else
>           m[r,c] := 1;                       #     change element to 1
>         fi;
>       fi;
>     od;
>   od;
>   LiveDead:=MoLL:  # Init listlist of all factors (live or dead) in each component
>   for r to nrows do                      #for each model component
>     LiveDead[r]:=[op(MoLL[r]),op(Dead[r])]:#  concatenate live list with dead list
>   od:
>   FixLL:=LiveDead: # Init listlist of fixed factors (live or dead) in fixed comp.

>   for r to nrows do                      # for each model component
>     for c to ncols do                    # for each factor
>       if  (FactorsNoise[c,2]=Ran) and member(FactorsNoise[c,1], LiveDead[r]) then
>         FixLL[r]:=[];
>       fi;
>     od;
```

```
>   od:
>   for r to nrows do
>     mm := evalm(m);  # create working copy mm of matrix m
>     for c to ncols do
>       if member(FactorsNoise[c,1],MoLL[r]) then # if c'th factor in Live list then
>         for rr to nrows do
>           mm[rr,c]:=1;                    #    cover the column
>         od;
>       fi;
>     od;
>     OutL := [];                      # Initialize OutList to be the empty list
>     for rr from 1 to nrows do
>       good := true; #"good" will remain true if each factor from the r'th component
>       for k to nops(LiveDead[r]) do      # also occurs in the rr'th component:
>         good := good and member(LiveDead[r,k], LiveDead[rr])
>       od;
>       if good then        # if "good" is true then
>         p:=1;
>         for c to ncols do  #   find the product p of all numbers in the rr'th row
>           p:=p*mm[rr,c];
>         od;
>         if p>0 then        # if product>0 then add p*VarianceComponent to OutList:
>           if FixLL[rr]=[] then
>             OutL := [ op(OutL), p*V(MoLP[rr]) ];
>           else
>             OutL := [ op(OutL), p*f(MoLP[rr]) ];
>           fi;
>         fi;
>       fi;
>     od:
>     printf(`  EMS(%a) = %a\n`,MoLP[r],convert(OutL,`+`)); # write one EMS line
>   od; RETURN(evalm(m));
> end:   # end of function "EMS()"
>
> ##### U S A G E ########################################################################
> EMS([[A,Fix,2],[B,Ran,3],[C,Ran,2],[D,Fix,3],[E,Ran,2],[Eps,Ran,2]],   # Factors,Noise
>                mu+A+B(A)+C(B,A)+D+E(D)+C(B,A)*E(D)+Eps(A,B,C,D,E)); # The Model

> ##########################################################################################
  EMS(A)  =  72*f(A)+24*V(B)+12*V(C)+2*V(C*E)+V(Eps)
  EMS(B)  =  24*V(B)+12*V(C)+2*V(C*E)+V(Eps)
  EMS(C)  =  12*V(C)+2*V(C*E)+V(Eps)
  EMS(D)  =  48*f(D)+24*V(E)+2*V(C*E)+V(Eps)
  EMS(E)  =  24*V(E)+2*V(C*E)+V(Eps)
  EMS(C*E)  =  2*V(C*E)+V(Eps)
  EMS(Eps)  =  V(Eps)
```

$$\begin{bmatrix} 0 & 3 & 2 & 3 & 2 & 2 \\ 1 & 1 & 2 & 3 & 2 & 2 \\ 1 & 1 & 1 & 3 & 2 & 2 \\ 2 & 3 & 2 & 0 & 2 & 2 \\ 2 & 3 & 2 & 1 & 1 & 2 \\ 1 & 1 & 1 & 1 & 1 & 2 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

```
> EMS([[A,Fix,3],[B,Ran,2],[Eps,Ran,1]],mu+A+B(A) +Eps(A,B)); # Simple 1
  EMS(A)  =  2*f(A)+V(B)+V(Eps)
  EMS(B)  =  V(B)+V(Eps)
  EMS(Eps)  =  V(Eps)
```

$$\begin{bmatrix} 0 & 2 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

```
> EMS([[A,Ran,3],[B,Ran,2],[Eps,Ran,1]],mu+A+B(A) +Eps(A,B)); # Simple 2
  EMS(A) = 2*V(A)+V(B)+V(Eps)
  EMS(B) = V(B)+V(Eps)
  EMS(Eps) = V(Eps)
```

$$\begin{bmatrix} 1 & 2 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

```
> EMS([[A,Fix,3],[B,Fix,2],[Eps,Ran,1]],mu+A+B+A*B+Eps(A,B)); # Simple 3
  EMS(A) = 2*f(A)+V(Eps)
  EMS(B) = 3*f(B)+V(Eps)
  EMS(A*B) = f(A*B)+V(Eps)
  EMS(Eps) = V(Eps)
```

$$\begin{bmatrix} 0 & 2 & 1 \\ 3 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

```
> EMS([[A,Ran,3],[B,Fix,2],[Eps,Ran,1]],mu+A+B+A*B+Eps(A,B)); # Simple 4
  EMS(A) = 2*V(A)+V(Eps)
  EMS(B) = 3*f(B)+V(A*B)+V(Eps)
  EMS(A*B) = V(A*B)+V(Eps)
  EMS(Eps) = V(Eps)
```

$$\begin{bmatrix} 1 & 2 & 1 \\ 3 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

```
> EMS([[A,Ran,3],[B,Ran,2],[Eps,Ran,1]],mu+A+B+A*B+Eps(A,B)); # Simple 5
  EMS(A) = 2*V(A)+V(A*B)+V(Eps)
  EMS(B) = 3*V(B)+V(A*B)+V(Eps)
  EMS(A*B) = V(A*B)+V(Eps)
  EMS(Eps) = V(Eps)
```

$$\begin{bmatrix} 1 & 2 & 1 \\ 3 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

```
> EMS([[A,Ran,4],[B,Fix,3],[Eps,Ran,2]],mu+A+B+A*B+Eps(A,B)); # IFAK Section 3.4
  EMS(A) = 6*V(A)+V(Eps)
  EMS(B) = 8*f(B)+2*V(A*B)+V(Eps)
  EMS(A*B) = 2*V(A*B)+V(Eps)
  EMS(Eps) = V(Eps)
```

$$\begin{bmatrix} 1 & 3 & 2 \\ 4 & 0 & 2 \\ 1 & 0 & 2 \\ 1 & 1 & 1 \end{bmatrix}$$

```
> EMS([[A,Ran,9],[B,Ran,3],[Eps,Ran,2]],mu+A+B(A) +Eps(A,B)); # IFAK Example 3.2
  EMS(A) = 6*V(A)+2*V(B)+V(Eps)
  EMS(B) = 2*V(B)+V(Eps)
  EMS(Eps) = V(Eps)
```

$$\begin{bmatrix} 1 & 3 & 2 \\ 1 & 1 & 2 \\ 1 & 1 & 1 \end{bmatrix}$$

```
> EMS([[A,Ran,5],[B,Ran,4],[Eps,Ran,1]],mu+A+B+A*B+Eps(A,B)); # IFAK Example 3.3
  EMS(A) = 4*V(A)+V(A*B)+V(Eps)
  EMS(B) = 5*V(B)+V(A*B)+V(Eps)
  EMS(A*B) = V(A*B)+V(Eps)
  EMS(Eps) = V(Eps)
```

$$\begin{bmatrix} 1 & 4 & 1 \\ 5 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

```
> EMS([[A,Fix,2],[B,Fix,3],[C,Ran,3],[Eps,Ran,2]],
>        mu+A+B+A*B+C(B)+A*C(B)+Eps(A,B,C)); # IFAK Example 3.4
  EMS(A) = 18*f(A)+2*V(A*C)+V(Eps)
  EMS(B) = 12*f(B)+4*V(C)+V(Eps)
  EMS(A*B) = 6*f(A*B)+2*V(A*C)+V(Eps)
  EMS(C) = 4*V(C)+V(Eps)
  EMS(A*C) = 2*V(A*C)+V(Eps)
  EMS(Eps) = V(Eps)
```

$$\begin{bmatrix} 0 & 3 & 3 & 2 \\ 2 & 0 & 3 & 2 \\ 0 & 0 & 3 & 2 \\ 2 & 1 & 1 & 2 \\ 0 & 1 & 1 & 2 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

```
> #  Proof for the split-plot method to be used in IFAK Example 4.0
> # §U represents the whole plot experimental Units = long strips of land
> EMS([[B,Ran,4],[P,Fix,2],[G,Fix,3],[U,Ran,1],[Eps,Ran,1]],
>        mu+B+P+B*P+G+B*G+P*G+B*P*G+U(B,P)+U(B,P)*G+Eps(B,P,G,U));
  EMS(B) = 6*V(B)+3*V(U)+V(Eps)
  EMS(P) = 12*f(P)+3*V(B*P)+3*V(U)+V(Eps)
  EMS(B*P) = 3*V(B*P)+3*V(U)+V(Eps)
  EMS(G) = 8*f(G)+2*V(B*G)+V(U*G)+V(Eps)
  EMS(B*G) = 2*V(B*G)+V(U*G)+V(Eps)
  EMS(P*G) = 4*f(P*G)+V(B*P*G)+V(U*G)+V(Eps)
  EMS(B*P*G) = V(B*P*G)+V(U*G)+V(Eps)
  EMS(U) = 3*V(U)+V(Eps)
  EMS(U*G) = V(U*G)+V(Eps)
  EMS(Eps) = V(Eps)
```

$$\begin{bmatrix} 1 & 2 & 3 & 1 & 1 \\ 4 & 0 & 3 & 1 & 1 \\ 1 & 0 & 3 & 1 & 1 \\ 4 & 2 & 0 & 1 & 1 \\ 1 & 2 & 0 & 1 & 1 \\ 4 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 3 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

Appendix 3.2. Maple program til beregning af varianskomponenter.

```
> EMS([[B,Ran,4],[P,Fix,2],[G,Fix,3],[Eps,Ran,2]],
>          mu+B+P+B*P+G+B*G+P*G+B*P*G+Eps(B,P,G)); # Split-plot, IFAK Example 4.0
  EMS(B) = 12*V(B)+V(Eps)
  EMS(P) = 24*f(P)+6*V(B*P)+V(Eps)
  EMS(B*P) = 6*V(B*P)+V(Eps)
  EMS(G) = 16*f(G)+4*V(B*G)+V(Eps)
  EMS(B*G) = 4*V(B*G)+V(Eps)
  EMS(P*G) = 8*f(P*G)+2*V(B*P*G)+V(Eps)
  EMS(B*P*G) = 2*V(B*P*G)+V(Eps)
  EMS(Eps) = V(Eps)
```

$$\begin{vmatrix} 1 & 2 & 3 & 2 \\ 4 & 0 & 3 & 2 \\ 1 & 0 & 3 & 2 \\ 4 & 2 & 0 & 2 \\ 1 & 2 & 0 & 2 \\ 4 & 0 & 0 & 2 \\ 1 & 0 & 0 & 2 \\ 1 & 1 & 1 & 1 \end{vmatrix}$$

```
> EMS([[T,Fix,3],[L,Fix,4],[O,Ran,2],[Eps,Ran,1]],
>            mu+T+L+T*L+O(T)+O(T)*L+Eps(T,L,O));  # Split-plot, IFAK Section 4.5.3
  EMS(T) = 8*f(T)+4*V(O)+V(Eps)
  EMS(L) = 6*f(L)+V(O*L)+V(Eps)
  EMS(T*L) = 2*f(T*L)+V(O*L)+V(Eps)
  EMS(O) = 4*V(O)+V(Eps)
  EMS(O*L) = V(O*L)+V(Eps)
  EMS(Eps) = V(Eps)
```

$$\begin{vmatrix} 0 & 4 & 2 & 1 \\ 3 & 0 & 2 & 1 \\ 0 & 0 & 2 & 1 \\ 1 & 4 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{vmatrix}$$

```
> #  Proof for the split-plot method to be used in IFAK Section 4.6.3
> #  Mus represents the whole plot experimental units = Mouse
> EMS([[K,Ran,5],[F,Fix,4],[D,Fix,2],[Mus,Ran,1],[Eps,Ran,1]],
>    mu+K+F+K*F+D+K*D+F*D+K*F*D+Mus(K,F)+Mus(K,F)*D+Eps(K,F,D,Mus));
  EMS(K) = 8*V(K)+2*V(Mus)+V(Eps)
  EMS(F) = 10*f(F)+2*V(K*F)+2*V(Mus)+V(Eps)
  EMS(K*F) = 2*V(K*F)+2*V(Mus)+V(Eps)
  EMS(D) = 20*f(D)+4*V(K*D)+V(Mus*D)+V(Eps)
  EMS(K*D) = 4*V(K*D)+V(Mus*D)+V(Eps)
  EMS(F*D) = 5*f(F*D)+V(K*F*D)+V(Mus*D)+V(Eps)
  EMS(K*F*D) = V(K*F*D)+V(Mus*D)+V(Eps)
  EMS(Mus) = 2*V(Mus)+V(Eps)
  EMS(Mus*D) = V(Mus*D)+V(Eps)
  EMS(Eps) = V(Eps)
```

$$\begin{vmatrix} 1 & 4 & 2 & 1 & 1 \\ 5 & 0 & 2 & 1 & 1 \\ 1 & 0 & 2 & 1 & 1 \\ 5 & 4 & 0 & 1 & 1 \\ 1 & 4 & 0 & 1 & 1 \\ 5 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 2 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{vmatrix}$$

```
> EMS([[K,Ran,5],[F,Fix,4],[D,Fix,2],[Eps,Ran,1]],
```

```
>        mu+K+F+K*F+D+K*D+F*D+K*F*D+Eps(K,F,D));  # Split-plot, IFAK Section 4.6.3
  EMS(K) = 8*V(K)+V(Eps)
  EMS(F) = 10*f(F)+2*V(K*F)+V(Eps)
  EMS(K*F) = 2*V(K*F)+V(Eps)
  EMS(D) = 20*f(D)+4*V(K*D)+V(Eps)
  EMS(K*D) = 4*V(K*D)+V(Eps)
  EMS(F*D) = 5*f(F*D)+V(K*F*D)+V(Eps)
  EMS(K*F*D) = V(K*F*D)+V(Eps)
  EMS(Eps) = V(Eps)
```

$$\begin{bmatrix} 1 & 4 & 2 & 1 \\ 5 & 0 & 2 & 1 \\ 1 & 0 & 2 & 1 \\ 5 & 4 & 0 & 1 \\ 1 & 4 & 0 & 1 \\ 5 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

```
> #  General proof for the split-plot method to be used in SF2 Oev 90.
> #  The factors may be RANDOM or FIXED, so below we consider six cases.
> #  Braet represents the whole plot experimental units = Board.
> EMS([[S,Rn,2],[K,Ran,2],[L,Ran,2],[Braet,Ran,2],[Eps,Ran,1]], mu+
>  S+K+S*K+L+S*L+K*L+S*K*L+Braet(S,K)+Braet(S,K)*L+Eps(S,K,L,Braet)); # SF2 Oev90 R-R-R
  EMS(S) = 8*f(S)+4*V(S*K)+4*V(S*L)+2*V(S*K*L)+2*V(Braet)+V(Braet*L)+V(Eps)
  EMS(K) = 8*V(K)+4*V(S*K)+4*V(K*L)+2*V(S*K*L)+2*V(Braet)+V(Braet*L)+V(Eps)
  EMS(S*K) = 4*V(S*K)+2*V(S*K*L)+2*V(Braet)+V(Braet*L)+V(Eps)
  EMS(L) = 8*V(L)+4*V(S*L)+4*V(K*L)+2*V(S*K*L)+V(Braet*L)+V(Eps)
  EMS(S*L) = 4*V(S*L)+2*V(S*K*L)+V(Braet*L)+V(Eps)
  EMS(K*L) = 4*V(K*L)+2*V(S*K*L)+V(Braet*L)+V(Eps)
  EMS(S*K*L) = 2*V(S*K*L)+V(Braet*L)+V(Eps)
  EMS(Braet) = 2*V(Braet)+V(Braet*L)+V(Eps)
  EMS(Braet*L) = V(Braet*L)+V(Eps)
  EMS(Eps) = V(Eps)
```

$$\begin{bmatrix} 1 & 2 & 2 & 2 & 1 \\ 2 & 1 & 2 & 2 & 1 \\ 1 & 1 & 2 & 2 & 1 \\ 2 & 2 & 1 & 2 & 1 \\ 1 & 2 & 1 & 2 & 1 \\ 2 & 1 & 1 & 2 & 1 \\ 1 & 1 & 1 & 2 & 1 \\ 1 & 1 & 2 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

```
> EMS([[S,Ran,2],[K,Ran,2],[L,Fix,2],[Braet,Ran,2],[Eps,Ran,1]], mu+
> S+K+S*K+L+S*L+K*L+S*K*L+Braet(S,K)+Braet(S,K)*L+Eps(S,K,L,Braet)); # SF2 Oev90 R-R-F
  EMS(S) = 8*V(S)+4*V(S*K)+2*V(Braet)+V(Eps)
  EMS(K) = 8*V(K)+4*V(S*K)+2*V(Braet)+V(Eps)
  EMS(S*K) = 4*V(S*K)+2*V(Braet)+V(Eps)
  EMS(L) = 8*f(L)+4*V(S*L)+4*V(K*L)+2*V(S*K*L)+V(Braet*L)+V(Eps)
  EMS(S*L) = 4*V(S*L)+2*V(S*K*L)+V(Braet*L)+V(Eps)
  EMS(K*L) = 4*V(K*L)+2*V(S*K*L)+V(Braet*L)+V(Eps)
  EMS(S*K*L) = 2*V(S*K*L)+V(Braet*L)+V(Eps)
  EMS(Braet) = 2*V(Braet)+V(Eps)
  EMS(Braet*L) = V(Braet*L)+V(Eps)
  EMS(Eps) = V(Eps)
```

$$\begin{bmatrix} 1 & 2 & 2 & 2 & 1 \\ 2 & 1 & 2 & 2 & 1 \\ 1 & 1 & 2 & 2 & 1 \\ 2 & 2 & 0 & 2 & 1 \\ 1 & 2 & 0 & 2 & 1 \\ 2 & 1 & 0 & 2 & 1 \\ 1 & 1 & 0 & 2 & 1 \\ 1 & 1 & 2 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

```
> EMS([[S,Ran,2],[K,Fix,2],[L,Ran,2],[Braet,Ran,2],[Eps,Ran,1]],mu+
> S+K+S*K+L+S*L+K*L+S*K*L+Braet(S,K)+Braet(S,K)*L+Eps(S,K,L,Braet)); # SF2 Oev90 R-F-R
  EMS(S) = 8*V(S)+4*V(S*L)+2*V(Braet)+V(Braet*L)+V(Eps)
  EMS(K) = 8*f(K)+4*V(S*K)+4*V(K*L)+2*V(S*K*L)+2*V(Braet)+V(Braet*L)+V(Eps)
  EMS(S*K) = 4*V(S*K)+2*V(S*K*L)+2*V(Braet)+V(Braet*L)+V(Eps)
  EMS(L) = 8*V(L)+4*V(S*L)+V(Braet*L)+V(Eps)
  EMS(S*L) = 4*V(S*L)+V(Braet*L)+V(Eps)
  EMS(K*L) = 4*V(K*L)+2*V(S*K*L)+V(Braet*L)+V(Eps)
  EMS(S*K*L) = 2*V(S*K*L)+V(Braet*L)+V(Eps)
  EMS(Braet) = 2*V(Braet)+V(Braet*L)+V(Eps)
  EMS(Braet*L) = V(Braet*L)+V(Eps)
  EMS(Eps) = V(Eps)
```

$$\begin{bmatrix} 1 & 2 & 2 & 2 & 1 \\ 2 & 0 & 2 & 2 & 1 \\ 1 & 0 & 2 & 2 & 1 \\ 2 & 2 & 1 & 2 & 1 \\ 1 & 2 & 1 & 2 & 1 \\ 2 & 0 & 1 & 2 & 1 \\ 1 & 0 & 1 & 2 & 1 \\ 1 & 1 & 2 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

```
> EMS([[S,Ran,2],[K,Fix,2],[L,Fix,2],[Braet,Ran,2],[Eps,Ran,1]],mu+
> S+K+S*K+L+S*L+K*L+S*K*L+Braet(S,K)+Braet(S,K)*L+Eps(S,K,L,Braet)); # SF2 Oev90 R-F-F
  EMS(S) = 8*V(S)+2*V(Braet)+V(Eps)
  EMS(K) = 8*f(K)+4*V(S*K)+2*V(Braet)+V(Eps)
  EMS(S*K) = 4*V(S*K)+2*V(Braet)+V(Eps)
  EMS(L) = 8*f(L)+4*V(S*L)+V(Braet*L)+V(Eps)
  EMS(S*L) = 4*V(S*L)+V(Braet*L)+V(Eps)
  EMS(K*L) = 4*f(K*L)+2*V(S*K*L)+V(Braet*L)+V(Eps)
  EMS(S*K*L) = 2*V(S*K*L)+V(Braet*L)+V(Eps)
  EMS(Braet) = 2*V(Braet)+V(Eps)
  EMS(Braet*L) = V(Braet*L)+V(Eps)
  EMS(Eps) = V(Eps)
```

$$\begin{bmatrix} 1 & 2 & 2 & 2 & 1 \\ 2 & 0 & 2 & 2 & 1 \\ 1 & 0 & 2 & 2 & 1 \\ 2 & 2 & 0 & 2 & 1 \\ 1 & 2 & 0 & 2 & 1 \\ 2 & 0 & 0 & 2 & 1 \\ 1 & 0 & 0 & 2 & 1 \\ 1 & 1 & 2 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

```
> EMS([[S,Fix,2],[K,Fix,2],[L,Ran,2],[Braet,Ran,2],[Eps,Ran,1]],mu+
>  S+K+S*K+L+S*L+K*L+S*K*L+Braet(S,K)+Braet(S,K)*L+Eps(S,K,L,Braet)); # SF2 Oev90 F-F-R
  EMS(S) = 8*f(S)+4*V(S*L)+2*V(Braet)+V(Braet*L)+V(Eps)
  EMS(K) = 8*f(K)+4*V(K*L)+2*V(Braet)+V(Braet*L)+V(Eps)
  EMS(S*K) = 4*f(S*K)+2*V(S*K*L)+2*V(Braet)+V(Braet*L)+V(Eps)
  EMS(L) = 8*V(L)+V(Braet*L)+V(Eps)
  EMS(S*L) = 4*V(S*L)+V(Braet*L)+V(Eps)
  EMS(K*L) = 4*V(K*L)+V(Braet*L)+V(Eps)
  EMS(S*K*L) = 2*V(S*K*L)+V(Braet*L)+V(Eps)
  EMS(Braet) = 2*V(Braet)+V(Braet*L)+V(Eps)
  EMS(Braet*L) = V(Braet*L)+V(Eps)
  EMS(Eps) = V(Eps)
```

$$\begin{bmatrix} 0 & 2 & 2 & 2 & 1 \\ 2 & 0 & 2 & 2 & 1 \\ 0 & 0 & 2 & 2 & 1 \\ 2 & 2 & 1 & 2 & 1 \\ 0 & 2 & 1 & 2 & 1 \\ 2 & 0 & 1 & 2 & 1 \\ 0 & 0 & 1 & 2 & 1 \\ 1 & 1 & 2 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

```
> EMS([[S,Fix,2],[K,Fix,2],[L,Fix,2],[Braet,Ran,2],[Eps,Ran,1]],mu+
>  S+K+S*K+L+S*L+K*L+S*K*L+Braet(S,K)+Braet(S,K)*L+Eps(S,K,L,Braet)); # SF2 Oev90 F-F-F
  EMS(S) = 8*f(S)+2*V(Braet)+V(Eps)
  EMS(K) = 8*f(K)+2*V(Braet)+V(Eps)
  EMS(S*K) = 4*f(S*K)+2*V(Braet)+V(Eps)
  EMS(L) = 8*f(L)+V(Braet*L)+V(Eps)
  EMS(S*L) = 4*f(S*L)+V(Braet*L)+V(Eps)
  EMS(K*L) = 4*f(K*L)+V(Braet*L)+V(Eps)
  EMS(S*K*L) = 2*f(S*K*L)+V(Braet*L)+V(Eps)
  EMS(Braet) = 2*V(Braet)+V(Eps)
  EMS(Braet*L) = V(Braet*L)+V(Eps)
  EMS(Eps) = V(Eps)
```

$$\begin{bmatrix} 0 & 2 & 2 & 2 & 1 \\ 2 & 0 & 2 & 2 & 1 \\ 0 & 0 & 2 & 2 & 1 \\ 2 & 2 & 0 & 2 & 1 \\ 0 & 2 & 0 & 2 & 1 \\ 2 & 0 & 0 & 2 & 1 \\ 0 & 0 & 0 & 2 & 1 \\ 1 & 1 & 2 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

Appendix 3.2. Maple program til beregning af varianskomponenter.

```
> #  Proof for the split-plot method to be used in Hicks Example 11.1
> #  U represents the whole plot experimental units = "Oven run".
> EMS([[R,Ran,3],[T,Fix,4],[B,Fix,3],[U,Ran,1],[Eps,Ran,1]],
>   mu+R+T+R*T+B+R*B+T*B+R*T*B+U(R,T)+U(R,T)*B+Eps(R,T,B,U));
  EMS(R) = 12*V(R)+3*V(U)+V(Eps)
  EMS(T) = 9*f(T)+3*V(R*T)+3*V(U)+V(Eps)
  EMS(R*T) = 3*V(R*T)+3*V(U)+V(Eps)
  EMS(B) = 12*f(B)+4*V(R*B)+V(U*B)+V(Eps)
  EMS(R*B) = 4*V(R*B)+V(U*B)+V(Eps)
  EMS(T*B) = 3*f(T*B)+V(R*T*B)+V(U*B)+V(Eps)
  EMS(R*T*B) = V(R*T*B)+V(U*B)+V(Eps)
  EMS(U) = 3*V(U)+V(Eps)
  EMS(U*B) = V(U*B)+V(Eps)
  EMS(Eps) = V(Eps)
```

$$\begin{bmatrix} 1 & 4 & 3 & 1 & 1 \\ 3 & 0 & 3 & 1 & 1 \\ 1 & 0 & 3 & 1 & 1 \\ 3 & 4 & 0 & 1 & 1 \\ 1 & 4 & 0 & 1 & 1 \\ 3 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 3 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

```
> #  Proof for the split-plot method to be used in Montgomery Table 12.11
> #  U represents the whole plot experimental units for A_B = Block_Method
> EMS([[A,Ran,3],[B,Fix,3],[C,Fix,4],[U,Ran,1],[Eps,Ran,1]],
>           mu+A+B+A*B+C+A*C+B*C+A*B*C+U(A,B)+U(A,B)*C+Eps(A,B,C,U));
  EMS(A) = 12*V(A)+4*V(U)+V(Eps)
  EMS(B) = 12*f(B)+4*V(A*B)+4*V(U)+V(Eps)
  EMS(A*B) = 4*V(A*B)+4*V(U)+V(Eps)
  EMS(C) = 9*f(C)+3*V(A*C)+V(U*C)+V(Eps)
  EMS(A*C) = 3*V(A*C)+V(U*C)+V(Eps)
  EMS(B*C) = 3*f(B*C)+V(A*B*C)+V(U*C)+V(Eps)
  EMS(A*B*C) = V(A*B*C)+V(U*C)+V(Eps)
  EMS(U) = 4*V(U)+V(Eps)
  EMS(U*C) = V(U*C)+V(Eps)
  EMS(Eps) = V(Eps)
```

$$\begin{bmatrix} 1 & 3 & 4 & 1 & 1 \\ 3 & 0 & 4 & 1 & 1 \\ 1 & 0 & 4 & 1 & 1 \\ 3 & 3 & 0 & 1 & 1 \\ 1 & 3 & 0 & 1 & 1 \\ 3 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 4 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

```
> #  Proof for the split-plot method to be used in Hicks Example 11.2
> #  U represents the whole plot experimental units for atmospheric conditions
> EMS([[R,Ran,4],[T,Fix,3],[H,Fix,3],[S,Ran,5],[U,Ran,1],[Eps,Ran,2]],
>          mu+R+T+R*T+H+R*H+T*H+R*T*H+S+R*S+T*S+R*T*S+H*S+R*H*S+T*H*S
>                            +R*T*H*S+U(R,T,H)+U(R,T,H)*S+Eps(R,T,H,S,U));
  EMS(R) = 90*V(R)+18*V(R*S)+10*V(U)+2*V(U*S)+V(Eps)
  EMS(T) = 120*f(T)+30*V(R*T)+24*V(T*S)+6*V(R*T*S)+10*V(U)+2*V(U*S)+V(Eps)
  EMS(R*T) = 30*V(R*T)+6*V(R*T*S)+10*V(U)+2*V(U*S)+V(Eps)
  EMS(H) = 120*f(H)+30*V(R*H)+24*V(H*S)+6*V(R*H*S)+10*V(U)+2*V(U*S)+V(Eps)
  EMS(R*H) = 30*V(R*H)+6*V(R*H*S)+10*V(U)+2*V(U*S)+V(Eps)
  EMS(T*H) = 40*f(T*H)+10*V(R*T*H)+8*V(T*H*S)+2*V(R*T*H*S)+10*V(U)+2*V(U*S)+V(Eps)
  EMS(R*T*H) = 10*V(R*T*H)+2*V(R*T*H*S)+10*V(U)+2*V(U*S)+V(Eps)
  EMS(S) = 72*V(S)+18*V(R*S)+2*V(U*S)+V(Eps)
  EMS(R*S) = 18*V(R*S)+2*V(U*S)+V(Eps)
  EMS(T*S) = 24*V(T*S)+6*V(R*T*S)+2*V(U*S)+V(Eps)
  EMS(R*T*S) = 6*V(R*T*S)+2*V(U*S)+V(Eps)
  EMS(H*S) = 24*V(H*S)+6*V(R*H*S)+2*V(U*S)+V(Eps)
  EMS(R*H*S) = 6*V(R*H*S)+2*V(U*S)+V(Eps)
  EMS(T*H*S) = 8*V(T*H*S)+2*V(R*T*H*S)+2*V(U*S)+V(Eps)
  EMS(R*T*H*S) = 2*V(R*T*H*S)+2*V(U*S)+V(Eps)
  EMS(U) = 10*V(U)+2*V(U*S)+V(Eps)
  EMS(U*S) = 2*V(U*S)+V(Eps)
  EMS(Eps) = V(Eps)
```

$$
\begin{bmatrix}
1 & 3 & 3 & 5 & 1 & 2 \\
4 & 0 & 3 & 5 & 1 & 2 \\
1 & 0 & 3 & 5 & 1 & 2 \\
4 & 3 & 0 & 5 & 1 & 2 \\
1 & 3 & 0 & 5 & 1 & 2 \\
4 & 0 & 0 & 5 & 1 & 2 \\
1 & 0 & 0 & 5 & 1 & 2 \\
4 & 3 & 3 & 1 & 1 & 2 \\
1 & 3 & 3 & 1 & 1 & 2 \\
4 & 0 & 3 & 1 & 1 & 2 \\
1 & 0 & 3 & 1 & 1 & 2 \\
4 & 3 & 0 & 1 & 1 & 2 \\
1 & 3 & 0 & 1 & 1 & 2 \\
4 & 0 & 0 & 1 & 1 & 2 \\
1 & 0 & 0 & 1 & 1 & 2 \\
1 & 1 & 1 & 5 & 1 & 2 \\
1 & 1 & 1 & 1 & 1 & 2 \\
1 & 1 & 1 & 1 & 1 & 1
\end{bmatrix}
$$

Appendix 3.2. Maple program til beregning af varianskomponenter.

```
> #  Proof for the split-split-plot method to be used in Hicks Example 11.3
> #  U1 represents the whole plot experimental units for Replication_Laboratory
> #  U2 represents the split plot experimental units for Replication_Laboratory_Temp
> EMS([[R,Ran,4],[L,Fix,3],[T,Fix,3],[M,Fix,3],[U1,Ran,1],[U2,Ran,1],[Eps,Ran,1]],
>      mu+R+L+R*L+T+R*T+L*T+R*L*T+M+R*M+L*M+R*L*M+T*M+R*T*M+L*T*M+R*L*T*M+U1(R,L)
>       +U1(R,L)*T+U1(R,L)*M+U1(R,L)*T*M+U2(R,L,T)+U2(R,L,T)*M+Eps(R,L,T,M,U1,U2));
  EMS(R)  =  27*V(R)+9*V(U1)+3*V(U2)+V(Eps)
  EMS(L)  =  36*f(L)+9*V(R*L)+9*V(U1)+3*V(U2)+V(Eps)
  EMS(R*L)  =  9*V(R*L)+9*V(U1)+3*V(U2)+V(Eps)
  EMS(T)  =  36*f(T)+9*V(R*T)+3*V(U1*T)+3*V(U2)+V(Eps)
  EMS(R*T)  =  9*V(R*T)+3*V(U1*T)+3*V(U2)+V(Eps)
  EMS(L*T)  =  12*f(L*T)+3*V(R*L*T)+3*V(U1*T)+3*V(U2)+V(Eps)
  EMS(R*L*T)  =  3*V(R*L*T)+3*V(U1*T)+3*V(U2)+V(Eps)
  EMS(M)  =  36*f(M)+9*V(R*M)+3*V(U1*M)+V(U2*M)+V(Eps)
  EMS(R*M)  =  9*V(R*M)+3*V(U1*M)+V(U2*M)+V(Eps)
  EMS(L*M)  =  12*f(L*M)+3*V(R*L*M)+3*V(U1*M)+V(U2*M)+V(Eps)
  EMS(R*L*M)  =  3*V(R*L*M)+3*V(U1*M)+V(U2*M)+V(Eps)
  EMS(T*M)  =  12*f(T*M)+3*V(R*T*M)+V(U1*T*M)+V(U2*M)+V(Eps)
  EMS(R*T*M)  =  3*V(R*T*M)+V(U1*T*M)+V(U2*M)+V(Eps)
  EMS(L*T*M)  =  4*f(L*T*M)+V(R*L*T*M)+V(U1*T*M)+V(U2*M)+V(Eps)
  EMS(R*L*T*M)  =  V(R*L*T*M)+V(U1*T*M)+V(U2*M)+V(Eps)
  EMS(U1)  =  9*V(U1)+V(Eps)
  EMS(U1*T)  =  3*V(U1*T)+V(Eps)
  EMS(U1*M)  =  3*V(U1*M)+V(Eps)
  EMS(U1*T*M)  =  V(U1*T*M)+V(Eps)
  EMS(U2)  =  3*V(U2)+V(Eps)
  EMS(U2*M)  =  V(U2*M)+V(Eps)
  EMS(Eps)  =  V(Eps)
```

$$\begin{bmatrix} 1 & 3 & 3 & 3 & 1 & 1 & 1 \\ 4 & 0 & 3 & 3 & 1 & 1 & 1 \\ 1 & 0 & 3 & 3 & 1 & 1 & 1 \\ 4 & 3 & 0 & 3 & 1 & 1 & 1 \\ 1 & 3 & 0 & 3 & 1 & 1 & 1 \\ 4 & 0 & 0 & 3 & 1 & 1 & 1 \\ 1 & 0 & 0 & 3 & 1 & 1 & 1 \\ 4 & 3 & 3 & 0 & 1 & 1 & 1 \\ 1 & 3 & 3 & 0 & 1 & 1 & 1 \\ 4 & 0 & 3 & 0 & 1 & 1 & 1 \\ 1 & 0 & 3 & 0 & 1 & 1 & 1 \\ 4 & 3 & 0 & 0 & 1 & 1 & 1 \\ 1 & 3 & 0 & 0 & 1 & 1 & 1 \\ 4 & 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 1 & 3 & 3 & 1 & 1 & 1 \\ 1 & 1 & 0 & 3 & 1 & 1 & 1 \\ 1 & 1 & 3 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 3 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

```
> #  Proof for the split-split-plot method to be used in Montgomery Example 12.3
> #  U1 represents the whole plot experimental units for Day_Technician
> #  U2 represents the split plot experimental units for Day_Technician_Dose
> EMS([[Day,Ran,4],[Tech,Fix,3],[Dose,Fix,3],[Thick,Fix,4],[U1,Ran,1],[U2,Ran,1],
>     [Eps,Ran,1]],mu+Day+Tech+Day*Tech+Dose+Day*Dose+Tech*Dose+Day*Tech*Dose+Thick
>     +Day*Thick+Tech*Thick+Day*Tech*Thick+Dose*Thick+Day*Dose*Thick+Tech*Dose*Thick
>     +Day*Tech*Dose*Thick+U1(Day,Tech)+U1(Day,Tech)*Dose+U1(Day,Tech)*Thick
>     +U1(Day,Tech)*Dose*Thick+U2(Day,Tech,Dose)+U2(Day,Tech,Dose)*Thick
>     +Eps(Day,Tech,Dose,Thick,U1,U2));
  EMS(Day) = 36*V(Day)+12*V(U1)+4*V(U2)+V(Eps)
  EMS(Tech) = 48*f(Tech)+12*V(Day*Tech)+12*V(U1)+4*V(U2)+V(Eps)
  EMS(Day*Tech) = 12*V(Day*Tech)+12*V(U1)+4*V(U2)+V(Eps)
  EMS(Dose) = 48*f(Dose)+12*V(Day*Dose)+4*V(U1*Dose)+4*V(U2)+V(Eps)
  EMS(Day*Dose) = 12*V(Day*Dose)+4*V(U1*Dose)+4*V(U2)+V(Eps)
  EMS(Tech*Dose) = 16*f(Tech*Dose)+4*V(Day*Tech*Dose)+4*V(U1*Dose)+4*V(U2)+V(Eps)
  EMS(Day*Tech*Dose) = 4*V(Day*Tech*Dose)+4*V(U1*Dose)+4*V(U2)+V(Eps)
  EMS(Thick) = 36*f(Thick)+9*V(Day*Thick)+3*V(U1*Thick)+V(U2*Thick)+V(Eps)
  EMS(Day*Thick) = 9*V(Day*Thick)+3*V(U1*Thick)+V(U2*Thick)+V(Eps)
  EMS(Tech*Thick) = 12*f(Tech*Thick)+3*V(Day*Tech*Thick)+3*V(U1*Thick)+V(U2*Thick)+V(Eps)
  EMS(Day*Tech*Thick) = 3*V(Day*Tech*Thick)+3*V(U1*Thick)+V(U2*Thick)+V(Eps)
  EMS(Dose*Thick) = 12*f(Dose*Thick)+3*V(Day*Dose*Thick)+V(U1*Dose*Thick)+V(U2*Thick)+V(Eps)
  EMS(Day*Dose*Thick) = 3*V(Day*Dose*Thick)+V(U1*Dose*Thick)+V(U2*Thick)+V(Eps)
  EMS(Tech*Dose*Thick)
=4*f(Tech*Dose*Thick)+V(Day*Tech*Dose*Thick)+V(U1*Dose*Thick)+V(U2*Thick)+        V(Eps)
  EMS(Day*Tech*Dose*Thick) = V(Day*Tech*Dose*Thick)+V(U1*Dose*Thick)+V(U2*Thick)+V(Eps)
  EMS(U1) = 12*V(U1)+V(Eps)
  EMS(U1*Dose) = 4*V(U1*Dose)+V(Eps)
  EMS(U1*Thick) = 3*V(U1*Thick)+V(Eps)
  EMS(U1*Dose*Thick) = V(U1*Dose*Thick)+V(Eps)
  EMS(U2) = 4*V(U2)+V(Eps)
  EMS(U2*Thick) = V(U2*Thick)+V(Eps)
  EMS(Eps) = V(Eps)
```

$$\begin{bmatrix} 1&3&3&4&1&1&1\\ 4&0&3&4&1&1&1\\ 1&0&3&4&1&1&1\\ 4&3&0&4&1&1&1\\ 1&3&0&4&1&1&1\\ 4&0&0&4&1&1&1\\ 1&0&0&4&1&1&1\\ 4&3&3&0&1&1&1\\ 1&3&3&0&1&1&1\\ 4&0&3&0&1&1&1\\ 1&0&3&0&1&1&1\\ 4&3&0&0&1&1&1\\ 1&3&0&0&1&1&1\\ 4&0&0&0&1&1&1\\ 1&0&0&0&1&1&1\\ 1&1&3&4&1&1&1\\ 1&1&0&4&1&1&1\\ 1&1&3&0&1&1&1\\ 1&1&0&0&1&1&1\\ 1&1&1&4&1&1&1\\ 1&1&1&0&1&1&1\\ 1&1&1&1&1&1&1 \end{bmatrix}$$

```
>
```

Appendix 3.3 BEREGNING AF STØRRELSERNE I TRINVIS VARIANSANALYSE.

Da beregningerne i den trinvise variansanalyse er noget anderledes end i den sidedelte, er der i dette appendix angivet en anskuelig forklaring på de enkelte størrelser og hvorledes de kunne beregnes ved hjælp af en lommeregner.
Som eksempel er valgt det i eksempel 2 angivne forsøg.
I skemaet på næste side er dels resultaterne af forsøget angivet dels beregnet visse hjælpestørrelser som forklares nedenfor.

1) Af skemaet fås, at charge 1 , prøve 1 gav analyseresultaterne 50.3 og 49.8 med gennemsnittet $\bar{x}_B$ = 50.05
   Et skøn for σ_C^2 er derfor variansen på de to tal, som er 0.125
   Gøres dette for alle 27 stikprøver fås resultaterne i skemaet søjle 4.
   Pooles disse fås

$$v_c = \frac{0.125+0.18+...+0.005}{27} = 0.747 \quad \text{eller} \quad \sigma_C^2 \approx 0.747 \text{ med 27 frihedsgrader}$$

$$\left(\text{Generelt} \quad v_c = \frac{\sum\sum\sum (y_i - \bar{y})^2}{a \cdot b \cdot (c-1)} \quad \text{med } a \cdot b \cdot (c\text{-}1) \text{ frihedsgrader}\right)$$

2) Charge 1 gav 3 stikprøvegennemsnit på 50.05, 49.8 og 50.25.
   Et skøn for variansen mellem disse 3 gennemsnit er 0.0508 med 2 frihedsgrader.
   Gøres dette for alle 9 charger fås resultaterne i skemaet søjle 6.
   Pooles disse fås

$$v_B = \frac{0.0508+0.3675+...+0.3958}{9} = 0.7029 \quad \text{med } 9 \cdot 2 = 18 \text{ frihedsgrader}$$

   V_B er ikke et skøn for σ_B^2 alene, da hver stikprøve er repræsenteret ved et gennemsnit af c = 2 testanalyser.

   Vi har derfor at $v_B \approx \sigma_B^2 + \frac{\sigma_C^2}{2}$ (Generelt $v_B \approx \sigma_B^2 + \frac{\sigma_C^2}{c}$)

   Dette giver $0.7029 \approx \sigma_B^2 + \frac{\sigma_C^2}{2}$ eller da $\sigma_C^2 \approx 0.747$

$$\sigma_B^2 \approx 0.7029 - \frac{0.747}{2} = 0.329$$

3) Betragtes nu de 9 gennemsnit for hver charge (søjle 7) kan variansen af disse beregnes v_A = 16.518. V_A er ikke et skøn for σ_A^2 alene, da hver charge-gennemsnit består af b = 3 stikprøver, og b c = 2 · 3 testanalyser.

   Vi har derfor at $v_A \approx \sigma_A^2 + \frac{\sigma_B^2}{3} + \frac{\sigma_C^2}{2 \cdot 3}$ (Generelt $v_A \approx \sigma_A^2 + \frac{\sigma_B^2}{b} + \frac{\sigma_C^2}{b \cdot c}$)

   Dette giver $16.518 \approx \sigma_A^2 + \frac{0.329}{3} + \frac{0.747}{2 \cdot 3}$ eller

$$\sigma_A^2 \approx 16.518 - 0.1098 - 0.124 = 16.284$$

   Vi ser heraf, at langt den største variation skyldes variationen mellem chargerne, og det derfor må være her man må søge at sætte ind.

A:Charge	B:Stikprøve	C: Test	(4):v_C	(5): $\bar{x}_B$	(6):v_B	(7): $\bar{x}_A$
1	1	1: 50.3 2: 49.8	0.125	50.05	0.0508	50.033
1	2	1. 50.1 2: 49.5	0.1800	49.8		
1	3	1: 51.1 2: 49.4	1.445	50.25		
2	1	1: 45.8 2: 45.4	0.08	45.6	0.3675	44.90
2	2	1. 44.4 2: 44.7	0.045	44.55		
2	3	1: 44.7 2: 44.4	0.045	44.45		
3	1	1: 41.0 2: 41.4	0.080	41.20	1.001	42.183
3	2	1. 42.7 2: 41.6	0.605	42.15		
3	3	1: 43.1 2: 43.3	0.020	43.20		
4	1	1: 48.7 2: 50.0	0.845	49.35	0.3908	48.917
4	2	1. 48.0 2: 50.4	2.880	49.20		
4	3	1: 47.9 2: 48.5	1.180	48.20		
5	1	1: 48.9 2: 49.4	0.125	49.15	2.5608	47.566
5	2	1. 48.4 2: 46.8	1.280	47.6		
5	3	1: 46.5 2: 45.4	0.605	45.95		
6	1	1: 47.0 2: 46.1	0.405	46.55	0.2708	46.716
6	2	1: 47.4 2: 47.2	0.020	47.30		
6	3	1: 45.1 2: 47.5	2.88	46.30		
7	1	1: 46.3 2: 45.0	0.845	45.65	0.4575	44.95
7	2	1. 44.6 2: 44.0	0.180	44.30		
7	3	1: 45.6 2: 44.2	0.980	44.90		
8	1	1: 44.9 2: 42.3	3.380	43.60	0.8308	43.43
8	2	1. 45.1 2: 43.4	1.445	44.25		
8	3	1: 43.3 2: 41.6	1.445	42.25		
9	1	1: 55.7 2: 55.4	0.045	55.55	0.3958	55.63
9	2	1. 56.3 2: 56.3	0.00	56.3		
9	3	1: 55.1 2: 55.0	0.005	55.05		

Ved sammenligning med varianskomponentmetoden ses, at vi har fået samme resultat.

Statgraphics kan i den situation, hvor alle faktorer er random, direkte udregne disse varianser ved følgende fremgangsmåde:
Datafilen har samme udseende som tidligere angivet.
Nu vælges Compare, Analysis of Variance og Variance Components
I den fremkomne menu skal de nestede faktorer placeres i den korrekte trinvise rækkefølge (charger, proever, analyser).
Der fremkommer så følgende tabel:

```
Variance Components Analysis

Dependent variable: PROCENT_VA
Factors:
        CHARGER
        PROEVER
        ANALYSER

Number of complete cases: 54

Analysis of Variance for PROCENT_VA
--------------------------------------------------------------------------------
Source                Sum of Squares     Df    Mean Square   Var. Comp.  Percent
--------------------------------------------------------------------------------
TOTAL (CORRECTED)            838,355     53
--------------------------------------------------------------------------------
CHARGER                      792,881      8        99,1102      16,2841    93,80
PROEVER                      25,3033     18        1,40574     0,329352     1,90
ANALYSER                       20,17     27       0,747037     0,747037     4,30
--------------------------------------------------------------------------------


The StatAdvisor
---------------
   The analysis of variance table shown here divides the variance of
PROCENT_VA into 3 components, one for each factor.  Each factor after
the first is nested in the one above.  The goal of such an analysis is
usually to estimate the amount of variability contributed by each of
the factors, called the variance components.  In this case, the factor
contributing the most variance is CHARGER.  Its contribution
represents 93,7998% of the total variation in PROCENT_VA.
```

Resultatet passer med vore tidligere beregninger.

APPENDIX4.1 VÆSENTLIGE FORMLER I MULIPEL REGRESSIONSANALYSE

A4.1. Indledning.

De beregninger, der skal foretages i en større regressionsmodel er så omfattende, er det af tidsmæssige grunde er nødvendigt at benytte et godt edb-program som Maple, der kan foretage de nødvendige manipulationer med store matricer. Dette supplement forudsættes derfor et grundlæggende kendskab til elementær matrixregning af et omfang svarende eksempelvis til B. Hellesen og Mogens Oddershede Larsen, Matematik for ingeniører: Bind III , kapitlerne 2-4.
Alle formler vil dog blive belyst gennem eksempler, så man også på den måde får repeteret de nødvendige matrixbegreber.

I afsnit 4.2.1 blev de vigtigste begreber gennemgået med udgangspunkt i enkelt regressionsanalyse. I dette supplement vil forklaringer og eksempler tage sit udgangspunkt i dobbelt regressionsanalyse, som har den fordel, at den stadig er en overskuelig model samtidig med at man let kan generalisere herfra til modeller med flere variable.

A4.2 Udledning af normalligningssystemet til beregning af regressionskoefficienterne.

Vi vil først se på hvorledes udledningen kan ske i tilfældet dobbelt regression, og derefter generalisere.
<u>Dobbelt regression</u>
Lad der være givet 2 uafhængige variable og n observationer (x_{i1}, x_{i2}, y_i), $i = 1,2, \ldots, n$.
For overskuelighedens skyld gives observationerne sædvanligvis i form af en tabel:

x_1	x_2	y
x_{11}	x_{12}	y_1
x_{21}	x_{22}	y_2
.	.	.
.	.	.
.	.	.
x_{n1}	x_{n2}	y_n

Lad regressionsligningen være $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$, (1)
hvor $\beta_0, \beta_1, \beta_2$ er regressionskoefficienterne og ε er fejlleddet, som antages normalfordelt med middelværdi 0.
Vi benytter nu "mindste kvadraters metode" til at finde estimater for regressionskoefficienterne, dvs. vi søger den værdi af $\beta_0, \beta_1, \beta_2$, som gør summen af kvadraterne af afstandene mellem den estimerede regressionsflade og de observerede punkter mindst mulig.
Indsættes observationerne i ligning (1) fås n ligninger

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i, \qquad i = 1,2, \ldots, n.$$

til bestemmelse af de 3 ubekendte $\beta_0, \beta_1, \beta_2$.
De n fejlled ε_i antages alle at være normalfordelte variable med samme middelværdi 0 og varians σ^2.

Vi danner nu funktionen $L(\beta_0, \beta_1, \beta_2) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2})^2$

For at finde minimum for denne funktion findes de stationære punkter for funktionen L's (der differentieres med hensyn til de 3 variable $\beta_0, \beta_1, \beta_2$, og disse sættes til 0).

Benævnes de søgte estimater $\tilde{\beta}_0, \tilde{\beta}_1, \tilde{\beta}_2$ skal disse altså opfylde ligningerne

$$\frac{\partial L}{\partial \beta_0} = -2\sum_{i=1}^{n}\left(y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_{i1} - \tilde{\beta}_2 x_{i2}\right) = 0$$

$$\frac{\partial L}{\partial \beta_1} = -2\sum_{i=1}^{n}\left(y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_{i1} - \tilde{\beta}_2 x_{i2}\right)x_{i1} = 0 \qquad (2)$$

$$\frac{\partial L}{\partial \beta_2} = -2\sum_{i=1}^{n}\left(y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_{i1} - \tilde{\beta}_2 x_{i2}\right)x_{i2} = 0$$

Ligningssystemet (2) kan simplificeres til følgende **normalligningssystem.**

$$\begin{cases} n\tilde{\beta}_0 + \tilde{\beta}_1 \sum_{i=1}^{n} x_{i1} + \tilde{\beta}_2 \sum_{i=1}^{n} x_{i2} = \sum_{i=1}^{n} y_i \\ \tilde{\beta}_0 \sum_{i=0}^{n} x_{i1} + \tilde{\beta}_1 \sum_{i=1}^{n} x_{i1}^2 + \tilde{\beta}_2 \sum_{i=1}^{n} x_{i1} \cdot x_{i2} = \sum_{i=1}^{n} x_{i1} \cdot y_i \\ \tilde{\beta}_0 \sum_{i=0}^{n} x_{i2} + \tilde{\beta}_1 \sum_{i=1}^{n} x_{i2} \cdot x_{i1} + \tilde{\beta}_2 \sum_{i=1}^{n} x_{i2}^2 = \sum_{i=1}^{n} x_{i2} \cdot y_i \end{cases} \qquad (3)$$

Bemærk, at der er 3 ligninger med 3 ubekendte, som kan løses ved sædvanlig matrixregning.

I de følgende afsnit vil eksempel 5.6 blive anvendt som gennemgående eksempel.

Eksempel A4.1 (opstilling af normalligningssystem)
Lad der være givet følgende observationer:

(x_1,x_2)	(4,3)	(5,4)	(6,6)	(7,8)
y	34.4 34.7	44.4 45.5	63.2 64.2	79.9 80.6

Man (Statgraphics) fandt regressionskoefficienterne $\tilde{\beta}_0 = -1.6333, \tilde{\beta}_1 = 3.8833, \tilde{\beta}_2 = 6.88333$.
Opskriv normalligningssystemet, og eftervis, at disse regressionskoefficienter er løsninger til systemet.

LØSNING:
Idet der er 2 gentagelser, kan man ofte i formlerne nøjes med at gange med 2.

$$\begin{cases} 8\tilde{\beta}_0 + \tilde{\beta}_1 \cdot 2\cdot(4+5+6+7) + \tilde{\beta}_2 \cdot 2\cdot(3+4+6+8) = 34.4 + 34.7 + \ldots + 80.6 \\ \tilde{\beta}_0 \cdot 44 + \tilde{\beta}_1 \cdot 2\cdot(4^2+5^2+6^2+7^2) + \tilde{\beta}_2 \cdot 2\cdot(4\cdot 3+5\cdot 4+6\cdot 6+7\cdot 8) = 4\cdot 34.4 + 4\cdot 34.7 + \ldots + 7\cdot 80.6) \\ \tilde{\beta}_0 \cdot 42 + \tilde{\beta}_1 \cdot 2\cdot(3\cdot 4+4\cdot 5+6\cdot 6+8\cdot 7) + \tilde{\beta}_2 \cdot 2\cdot(3^2+4^2+6^2+8^2) = 3\cdot 34.4 + 3\cdot 34.7 + \ldots + 8\cdot 80.6 \end{cases}$$

$$\Leftrightarrow \begin{cases} 8\tilde{\beta}_0 + \tilde{\beta}_1 \cdot 44 + \tilde{\beta}_2 \cdot 42 = 446.9 \\ \tilde{\beta}_0 \cdot 44 + \tilde{\beta}_1 \cdot 252 + \tilde{\beta}_2 \cdot 248 = 2613.8 \\ \tilde{\beta}_0 \cdot 42 + \tilde{\beta}_1 \cdot 248 + \tilde{\beta}_2 \cdot 250 = 2615.3 \end{cases}$$

Ved indsættelse af $\beta_0 = -1.6333, \beta_1 = 3.8833, \beta_2 = 6.88333$ ses, at

$$\begin{cases} 8 \cdot (-1.6333) + 3.88333 \cdot 44 + 6.88333 \cdot 42 = 446.9 \\ (-1.6333) \cdot 44 + 3.88333 \cdot 126 + 6.88333 \cdot 124 = 2613.8 \\ (-1.6333) \cdot 42 + 3.88333 \cdot 124 + 6.88333 \cdot 125 = 2615.3 \end{cases}$$

dvs. normalligningssystemet har de fundne løsninger.

A4.3. Regressionsligning, normalligningssystem og estimater for regressionskoefficienter i matrixformulering

Multipel regression

Ovenstående regninger kan generaliseres på følgende måde ved benyttelse af matrixformulering.

Lad der være givet k uafhængige variable og n observationer

$(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i)$, $i = 1,2, \ldots, n$ og $n > k$.

Observationerne gives mest overskueligt i form af en tabel:

x_1	x_2	.	.	.	x_k	y
x_{11}	x_{12}	.	.	.	x_{1k}	y_1
x_{21}	x_{22}	.	.	.	x_{2k}	y_2
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
x_{n1}	x_{n2}	.	.	.	x_{nk}	y_n

Lad regressionsligningen være $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$, (1)

hvor $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ er regressionskoefficienterne og ε er fejlleddet, som antages normalfordelt med middelværdi 0.

Modellen kan i matrixnotation skrives

$$\bar{y} = X \cdot \bar{\beta} + \bar{\varepsilon}$$

$$\text{hvor } \bar{y} = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_{11} & x_{12} & . & . & . & x_{1k} \\ 1 & x_{21} & x_{22} & . & . & . & x_{2k} \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ 1 & x_{n1} & x_{n2} & . & . & . & x_{nk} \end{bmatrix}, \quad \bar{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ . \\ \beta_k \end{bmatrix} \text{ og } \bar{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ . \\ \varepsilon_n \end{bmatrix}.$$

Vi ønsker ved mindste kvadraters metode, at finde en vektor $\tilde{\beta}$, der er et estimat for vektoren $\bar{\beta}$.

Ifølge eksempelvis B. Hellesen, M. Oddershede Larsen: Bind III, kapitel 3 er løsningen af et sådant overbestemt ligningssystem $X \cdot \bar{\beta} = \bar{y}$ bestemt ved

normalligningssystemet $X^T \cdot X \cdot \tilde{\beta} = X^T \cdot \bar{y}$. (3)

Matricen $X^T \cdot X$ er en kvadratisk symmetrisk matrix, som sædvanligvis ved regressions-problemer ikke er singulær. Der eksisterer derfor en invers matrix $(X^T \cdot X)^{-1}$, hvorved løsningen til normalligningssystemet (3) bliver

$$\tilde{\beta} = (X^T \cdot X)^{-1} \cdot X^T \cdot \bar{y} \quad (4)$$

Eksempel A4.2 (matrixnotation, og beregning af estimater af regressionskoefficienter)
I eksempel A4.1 *benyttede vi følgende observationer*

(x_1,x_2)	(4,3)	(5,4)	(6,6)	(7,8)
y	34.4 34.7	44.4 45.5	63.2 64.2	79.9 80.6

1) *Beskriv problemet i matrixnotation.*
2) *Udtryk ligeledes normalligningssystemet på matrixform, og benyt Maple til en reduktion af systemet.*
3) *Angiv løsningen på matrixform, og benyt Maple til at finde regressionskoefficienterne. Kontroller løsningen ved at sammenligne med Statgraphics's løsning*
   $\tilde{\beta}_0 = -1.6333, \tilde{\beta}_1 = 3.8833, \tilde{\beta}_2 = 6.88333$.

LØSNING:

1) Regressionsligningssystemet kan skrives $\bar{y} = X \cdot \tilde{\beta}$

$$\text{hvor } \bar{y} = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ . \\ . \\ y_8 \end{bmatrix} = \begin{bmatrix} 34.4 \\ 34.7 \\ 44.4 \\ 45.5 \\ 63.2 \\ 64.2 \\ 79.9 \\ 80.6 \end{bmatrix}, \; X = \begin{bmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ 1 & x_{31} & x_{32} \\ 1 & x_{41} & x_{42} \\ 1 & x_{51} & x_{52} \\ 1 & x_{61} & x_{62} \\ 1 & x_{71} & x_{72} \\ 1 & x_{81} & x_{82} \end{bmatrix} = \begin{bmatrix} 1 & 4 & 3 \\ 1 & 4 & 3 \\ 1 & 5 & 4 \\ 1 & 5 & 4 \\ 1 & 6 & 6 \\ 1 & 6 & 6 \\ 1 & 7 & 8 \\ 1 & 7 & 8 \end{bmatrix}, \; \tilde{\beta} = \begin{bmatrix} \tilde{\beta}_0 \\ \tilde{\beta}_1 \\ \tilde{\beta}_2 \end{bmatrix}.$$

2) Normalligningssystemet er $(X^T \cdot X) \cdot \tilde{\beta} = X^T \cdot \bar{y}$

$$\Leftrightarrow \left(\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 4 & 4 & 5 & 5 & 6 & 6 & 7 & 7 \\ 3 & 3 & 4 & 4 & 6 & 6 & 8 & 8 \end{bmatrix} \cdot \begin{bmatrix} 1 & 4 & 3 \\ 1 & 4 & 3 \\ 1 & 5 & 4 \\ 1 & 5 & 4 \\ 1 & 6 & 6 \\ 1 & 6 & 6 \\ 1 & 7 & 8 \\ 1 & 7 & 8 \end{bmatrix} \right) \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 4 & 4 & 5 & 5 & 6 & 6 & 7 & 7 \\ 3 & 3 & 4 & 4 & 6 & 6 & 8 & 8 \end{bmatrix} \cdot \begin{bmatrix} 34.4 \\ 34.7 \\ 44.4 \\ 45.5 \\ 63.2 \\ 64.2 \\ 79.91 \\ 80.6 \end{bmatrix}$$

I Maple hentes algebra-pakken frem ved
\> with(linalg):
Derefter findes venstre side ved ordren:
\> U:=multiply(transpose(X),X);

$$U := \begin{bmatrix} 8 & 44 & 42 \\ 44 & 252 & 248 \\ 42 & 248 & 250 \end{bmatrix}$$

og højre side ved ordren
\> H:=multiply(transpose(X),Y);

$$H := \begin{bmatrix} 446.9 \\ 2613.8 \\ 2615.3 \end{bmatrix}$$

Dette giver normalligningssystemet: $\begin{bmatrix} 8 & 44 & 42 \\ 44 & 252 & 248 \\ 42 & 248 & 250 \end{bmatrix} \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} 446.9 \\ 2613.8 \\ 2615.3 \end{bmatrix}$

Dette ses, at det er samme normalligningssystem som i eksempel A4.1.

3) Regressionskoefficienterne findes af ligningen
$\tilde{\beta} = (X^T \cdot X)^{-1} \cdot X^T \cdot \bar{y}$

$$\Leftrightarrow \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} = \left(\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 4 & 4 & 5 & 5 & 6 & 6 & 7 & 7 \\ 3 & 3 & 4 & 4 & 6 & 6 & 8 & 8 \end{bmatrix} \cdot \begin{bmatrix} 1 & 4 & 3 \\ 1 & 4 & 3 \\ 1 & 5 & 4 \\ 1 & 5 & 4 \\ 1 & 6 & 6 \\ 1 & 6 & 6 \\ 1 & 7 & 8 \\ 1 & 7 & 8 \end{bmatrix} \right)^{-1} \cdot \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 4 & 4 & 5 & 5 & 6 & 6 & 7 & 7 \\ 3 & 3 & 4 & 4 & 6 & 6 & 8 & 8 \end{bmatrix} \cdot \begin{bmatrix} 34.4 \\ 34.7 \\ 44.4 \\ 45.5 \\ 63.2 \\ 64.2 \\ 79.91 \\ 80.6 \end{bmatrix}$$

Igen benyttes Maple til udregningerne:
Først udregnes
> invU:=inverse(U);

$$invU := \begin{bmatrix} \frac{187}{6} & -\frac{73}{6} & \frac{41}{6} \\ -\frac{73}{6} & \frac{59}{12} & -\frac{17}{6} \\ \frac{41}{6} & -\frac{17}{6} & \frac{5}{3} \end{bmatrix}$$

> B:=multiply(invU,H);

$$B := \begin{bmatrix} -1.63333 \\ 3.883330 \\ 6.883333 \end{bmatrix}$$

Det ses, at regressionskoefficienterne stemmer med de af Statgraphics fundne.

◆

A4.4 Forklaring og beregning af de øvrige størrelser i en variansanalysetabel.

Uanset hvilken statistikpakke man anvender, har udskriften fra en regressionsanalyse samme udseende som nedenstående Statgraphics-udskrift

```
Multiple Regression Analysis
-----------------------------------------------------------------------------
Dependent variable: y
-----------------------------------------------------------------------------
                                        Standard          T
Parameter               Estimate         Error        Statistic        P-Value
-----------------------------------------------------------------------------
CONSTANT                -1,63333        4,33035       -0,377183         0,7215
x1                       3,88333        1,71994        2,25783          0,0736
x2                       6,88333        1,00139        6,87379          0,0010
-----------------------------------------------------------------------------
                           Analysis of Variance
-----------------------------------------------------------------------------
Source             Sum of Squares     Df  Mean Square    F-Ratio       P-Value
-----------------------------------------------------------------------------
Model                     2457,35      2      1228,68    2042,12        0,0000
Residual                  3,00833      5     0,601667
-----------------------------------------------------------------------------
Total (Corr.)             2460,36      7
```

Vi vil i dette afsnit redegøre for hvorledes de øvrige størrelser fremkommer.

<u>Residual.</u>
Som ved den ensidede regressionsanalyse, er residualerne forskellen mellem en observeret værdi y_i og den tilsvarende værdi $\hat{y}_i$ beregnet ud fra modellen, dvs. $r_i = y_i - \hat{y}_i$.

$\text{SAK}_{\text{residual}} = \sum_{i=1}^{n} r_i^2$.

Forudsat modellen er sand, kan det vises, at størrelsen $s_{residual}^2 = \dfrac{SAK_{residual}}{n-p}$, hvor n er antal observationer, og p er antal regressionskonstanter, er et centralt estimat for forsøgsfejlen σ^2.

Sættes $\bar{r} = \begin{bmatrix} r_1 \\ r_2 \\ \cdot \\ \cdot \\ \cdot \\ r_n \end{bmatrix}$ kan vi foretage følgende omskrivning

SAK$_{\text{residual}}$= $\sum_{i=1}^{n} r_i^2 = \bar{r}^T \cdot \bar{r} = (\bar{y} - X\tilde{\beta})^T \cdot (\bar{y} - X\tilde{\beta})..$

Eksempel A4.3 (beregning af residualer og SAK$_{\text{residual}}$)
I fortsættelse af eksempel A4.2 ønsket ved hjælp af Maple beregnet residualerne og SAK$_{residual}$
.
LØSNING:
Ved benyttelse af Maple fås residualerne ved ordren
> r:=matadd(Y,multiply(X,B), 1,-1);

$$r := \begin{bmatrix} -.149989 \\ .150011 \\ -.916652 \\ .183348 \\ .233352 \\ 1.233352 \\ -.716644 \\ -.016644 \end{bmatrix}$$

og SAK$_{\text{residual}}$ ved
> SAK_residual:=multiply(transpose(r),r);
$SAK\_residual := [3.008333336]$.
◆

Testprocedure.
Den procedure, som blev udførligt forklaret i forbindelse med under ensidet regressionsanalyse kan gentages i det generelle tilfælde.
Har man gentagne observationer (som i eksempel 4.6) kan der laves en “lack - of - fit” test, til testning af modellen. Har man ikke gentagelser, ser man på forklaringsgraden
$R^2 = \frac{SAK_{model}}{SAK_{total}}$ (som gennemgået i eksempel 4.7)
Er man blevet overbevist om, at modellen er rimelig, kan man teste nulhypotesen
$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$.
Dette sker ved en F - test.
Forudsat H_0 er sand, gælder, at $\frac{SAK_{model}}{\sigma^2}$ er χ^2 -fordelt med k frihedsgrader. Ligeledes kan

vises, at $\frac{SAK_{residual}}{\sigma^2}$ er χ^2 -fordelt med n - p frihedsgrader (p = k + 1). Endelig kan vises, at $SAK_{residual}$ og $SAK_{model.}$ er uafhængige.

En test af $H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$ er derfor en sædvanlig F-test, med

$$F = \frac{\frac{SAK_{model}}{k}}{\frac{SAK_{residual}}{n-p}} = \frac{s^2_{model}}{s^2_{residual}}.$$

Beregningsteknisk kan en lommeregner let udregne SAK_{total}, og dermed som sædvanlig $SAK_{model.} = SAK_{total} - SAK_{residual}$.

SAK_{model} kan beregnes direkte af formlen $SAK_{model} = (\tilde{\beta}^T \cdot X^T) \cdot \bar{y} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}$.

Hvis H_0 forkastes ved vi, at mindst én af regressionskoefficienterne er forskellig fra nul.

Kovariansmatrix.
Det kan vises (jævnfør eksempelvis L. Brøndum, J. D. Monrad: Statistisk forsøgsplanlægning I, side 392-394, at

1) $E(\tilde{\beta}) = \beta$,dvs. estimaterne for regressionskoefficienterne er centrale estimatorer

Idet $p = k + 1$ er antal regressionskoefficienter gælder:

2) Den kvadratiske symmetriske $p \times p$ "kovariansmatrix" $\sigma^2 \cdot (X^T \cdot X)^{-1}$, med elementerne $\sigma^2 \cdot C_{ij}$, har den egenskab:
   a) diagonalelementerne er varianserne for regressionskoefficienterne, dvs. $\sigma^2 \cdot C_{ii} = V(\tilde{\beta}_i)$, $i = 0, 1, 2, \ldots k$ og
   elementerne udenfor diagonalen C_{ij} angiver kovariansen mellem $\tilde{\beta}_i$ og $\tilde{\beta}_j$, dvs.
   $\sigma^2 \cdot C_{ij} = V(\beta_i, \beta_j)$.

Estimater for disse varianser og kovarianser fås ved at erstatte σ^2 med sit estimat $\tilde{\sigma}^2$.

Standard - error og t -værdier for regressionskoefficienterne.
Kvadratroden af estimatet for variansen af $\tilde{\beta}_i$ kaldes den estimerede spredning (eller standard error) for β_i ($s(\tilde{\beta}_i) = \sqrt{\tilde{\sigma}^2 \cdot C_{ii}}$).

Testes nulhypotesen $H_0 : \beta_i = 0$ mod den alternative hypotese $H : \beta_i \neq 0$, kan dette gøres ved at beregne en værdi $t_i = \frac{\tilde{\beta}_i}{\sqrt{\tilde{\sigma}^2 \cdot C_{ii}}}$, som kan vises at være t - fordelt med n - p frihedsgrader.

Da regressionskoefficienterne ved partielle strukturer afhænger af hinanden, skal man som beskrevet i kapitel 5, bortkaste én koefficient ad gangen (dvs. ikke på én gang bortkaste alle regressionskoefficienter, der har P-værdier større end 0).

Eksempel A4.4 (beregning af standard-error og t-værdier for regressionskoefficienter.)

Eksemplet bygger på samme talmateriale som eksemplerne A4.1 - 3.
Der ønskes beregnet
a) kovariansmatricen
b) standard - error og t-værdier for regressionskoefficienterne .

LØSNING:
a) I eksempel A4.2 blev beregnet $invU = (X^T \cdot X)^{-1}$
Vi fik:

$$invU := \begin{bmatrix} \frac{187}{6} & -\frac{73}{6} & \frac{41}{6} \\ -\frac{73}{6} & \frac{59}{12} & -\frac{17}{6} \\ \frac{41}{6} & -\frac{17}{6} & \frac{5}{3} \end{bmatrix}$$

Vi får så kovariansmatricen ved at multiplicere med $\sigma^2_{residual} = 0.6016667$
> kovar:=evalm(0.60166666*invU);

$$ko\,\text{var} := \begin{bmatrix} 18.75194424 & -7.320277697 & 4.111388843 \\ -7.320277697 & 2.958194412 & -1.704722203 \\ 4.111388843 & -1.704722203 & 1.002777767 \end{bmatrix}$$

2 Vi finder "standard error" ved at tage kvadratroden af diagonalelementerne.

$$s(\tilde{\beta}_0) = \sqrt{\tilde{\sigma}^2 \cdot C_{00}} = \sqrt{18.75194424} = 4.33035\,,$$
$$s(\tilde{\beta}_1) = \sqrt{\tilde{\sigma}^2 \cdot C_{11}} = \sqrt{2.958194412} = 1.71994\,,$$
$$s(\tilde{\beta}_2) = \sqrt{\tilde{\sigma}^2 \cdot C_{22}} = \sqrt{1.002777767} = 1.00139\,.$$

De tilsvarende t-værdier er

$$t_0 = \frac{\tilde{\beta}_0}{s(\tilde{\beta}_0)} = \frac{-1.63333}{4.33035} = -0.377183$$

$$t_1 = \frac{\tilde{\beta}_1}{s(\tilde{\beta}_1)} = \frac{3.88333}{1.71994} = 2.25783$$

$$t_2 = \frac{\tilde{\beta}_2}{s(\tilde{\beta}_2)} = \frac{6.88333}{1.00139} = 6.87379$$

Det ses, at værdierne stemmer overens med Statgraphics udskriften.
◆

A4.5 Formler og beregning af konfidensintervaller.

Konfidensinterval for β_i.

Vi har fra det foregående, at middelværdien $E(\tilde{\beta}_i) = \beta_i$ og spredningen $s(\tilde{\beta}_i) = \sqrt{\tilde{\sigma}^2 \cdot C_{ii}}$. Det kan endvidere vises, at størrelsen $\frac{\tilde{\beta}_i - \beta_i}{\sqrt{\tilde{\sigma}^2 \cdot C_{ii}}}$ er *t* - fordelt med *n* - *p* frihedsgrader.

Vi kan på det grundlag opstille en formel for et $100(1-\alpha)\%$ konfidensinterval .

$$\tilde{\beta}_i - t_{1-\frac{\alpha}{2}}(n-p) \cdot \sqrt{\tilde{\sigma}^2 C_{ii}} \le \beta_i \le \tilde{\beta}_i + t_{1-\frac{\alpha}{2}}(n-p) \cdot \sqrt{\tilde{\sigma}^2 C_{ii}} .$$

Konfidensinterval for et til punktet $\bar{x}_0$ svarende værdi $\hat{y}_0$.

For lettere at kunne forklare formlerne, begynder vi med det enkle regressionspolynomium $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$, hvor $\beta_0, \beta_1, \beta_2$ er regressionskoefficienterne.

Er $\tilde{\beta}_0, \tilde{\beta}_1, \tilde{\beta}_2$ de estimerede værdier, og indsættes punktet (x_{01}, x_{02}) i ligningen, fås den dertil svarende estimerede *y* - værdi $\tilde{y}_0 = \tilde{\beta}_0 + \tilde{\beta}_1 x_{01} + \tilde{\beta}_2 x_{02}$.

Da middelværdien

$$E(\tilde{\beta}_0 + \tilde{\beta}_1 x_{01} + \tilde{\beta}_2 x_{02}) = E(\tilde{\beta}_0) + E(\tilde{\beta}_1) x_{01} + E(\tilde{\beta}_2) x_{02} = \beta_0 + \beta_1 x_{01} + \beta_2 x_{02}$$

er estimatoren central.

Ifølge reglerne for varians af en linearkombination fås

$$V(\tilde{\beta}_0 + \tilde{\beta}_1 x_{01} + \tilde{\beta}_2 x_{02}) = V(\tilde{\beta}_0) + x_{01}^2 \cdot V(\tilde{\beta}_1) + x_{02}^2 \cdot V(\tilde{\beta}_2) + 2x_{01} \cdot V(\tilde{\beta}_0 \cdot \tilde{\beta}_1) + 2x_{02} \cdot V(\tilde{\beta}_0 \cdot \tilde{\beta}_2)$$
$$+ 2x_{01} \cdot x_{02} \cdot V(\tilde{\beta}_1 \cdot \tilde{\beta}_2) .$$

Sættes $\bar{x}_0 = \begin{bmatrix} 1 \\ x_{01} \\ x_{02} \end{bmatrix}$ og idet kovariansmatricen er

$$\sigma^2 \cdot (X^T \cdot X)^{-1} = \begin{bmatrix} V(\tilde{\beta}_0) & V(\tilde{\beta}_0, \tilde{\beta}_1) & V(\tilde{\beta}_0, \tilde{\beta}_2) \\ V(\tilde{\beta}_1, \tilde{\beta}_0) & V(\tilde{\beta}_1) & V(\tilde{\beta}_1, \tilde{\beta}_2) \\ V(\tilde{\beta}_2, \tilde{\beta}_0) & V(\tilde{\beta}_2, \tilde{\beta}_1) & V(\tilde{\beta}_2) \end{bmatrix}$$ ses, at

$$V(\tilde{\beta}_0 + \tilde{\beta}_1 x_{01} + \tilde{\beta}_2 x_{02}) = \sigma^2 \cdot \begin{bmatrix} 1 & x_{01} & x_{02} \end{bmatrix} \cdot \begin{bmatrix} V(\tilde{\beta}_0) & V(\tilde{\beta}_0, \tilde{\beta}_1) & V(\tilde{\beta}_0, \tilde{\beta}_2) \\ V(\tilde{\beta}_1, \tilde{\beta}_0) & V(\tilde{\beta}_1) & V(\tilde{\beta}_1, \tilde{\beta}_2) \\ V(\tilde{\beta}_2, \tilde{\beta}_0) & V(\tilde{\beta}_2, \tilde{\beta}_1) & V(\tilde{\beta}_2) \end{bmatrix} \cdot \begin{bmatrix} 1 \\ x_{01} \\ x_{02} \end{bmatrix}$$

$$\Leftrightarrow V(\tilde{\beta}_0 + \tilde{\beta}_1 x_{01} + \tilde{\beta}_2 x_{02}) = \sigma^2 \cdot \bar{x}_0^T \cdot (X^T \cdot X)^{-1} \cdot \bar{x}_0 .$$

Konfidensintervallet bliver følgelig

$$\left[\tilde{y}_0 - t_{1-\frac{\alpha}{2}}(n-p) \cdot \sqrt{\tilde{\sigma}^2 \cdot \bar{x}_0^T \cdot (X^T \cdot X)^{-1} \cdot \bar{x}_0} ; \tilde{y}_0 + t_{1-\frac{\alpha}{2}}(n-p) \cdot \sqrt{\tilde{\sigma}^2 \cdot \bar{x}_0^T \cdot (X^T \cdot X)^{-1} \cdot \bar{x}_0} \right] .$$

Eksempel A4.5 (beregning af konfidensintervaller)
Eksemplet bygger på samme talmateriale som eksemplerne A4.1 - 4.
Der ønskes beregnet

a) *et* 95% *konfidensinterval for* $\tilde{\beta}_2$,
b) *et estimat for den y-værdi, som svarer til punktet* (5,5)*, og angivet et* 95% *konfidensinterval for denne.*

Sammenlign de fundne værdier med en Statgraphics udskrift.

LØSNING:

a) Ved at indsætte de i eksempel A4.4 fundne størrelser, fås følgende 95% konfidensinterval for $\tilde{\beta}_2$:

$$\tilde{\beta}_2 - t_{0.975}(8-3)\cdot\sqrt{\tilde{\sigma}^2 C_{ii}} \le \beta_2 \le \tilde{\beta}_2 + t_{0.975}(8-3)\cdot\sqrt{\tilde{\sigma}^2 C_{ii}}$$
$$\Leftrightarrow 6.8833 - 2.57\cdot 1.00139 \le \beta_2 \le 6.8833 + 2.57\cdot 1.00139$$
$$\Leftrightarrow \underline{\underline{4.310 \le \beta_2 \le 9.457}}$$

Den tilsvarende udskrift fra Statgraphics viser samme resultat:

```
95,0% confidence intervals for coefficient estimates
--------------------------------------------------------------------------------
                                               Standard
Parameter                   Estimate            Error      Lower Limit    Upper Limit
--------------------------------------------------------------------------------
CONSTANT                    -1,63333          4,33035        -12,7649        9,49822
x1                           3,88333          1,71994       -0,537926        8,30459
x2                           6,88333          1,00139         4,30918        9,45749
--------------------------------------------------------------------------------
```

b) Idet $\bar{x}_0 = \begin{bmatrix}1\\5\\5\end{bmatrix}$ og $B := \begin{bmatrix}\tilde{\beta}_1\\\tilde{\beta}_2\\\tilde{\beta}_3\end{bmatrix}\begin{bmatrix}-1.63333\\3.883330\\6.883333\end{bmatrix}$ fås den estimerede y-værdi

$$y_0 = \bar{x}_0^T\cdot B = \begin{bmatrix}1 & 5 & 5\end{bmatrix}\cdot\begin{bmatrix}-1.63333\\3.883330\\6.883333\end{bmatrix} = \underline{\underline{52.19999}}.$$

Resultatet kan også fås af Maple-ordrerne:

\> x0:=matrix(3,1,[1,5,5]);

$$x0 := \begin{bmatrix}1\\5\\5\end{bmatrix}$$

\> yo:=multiply(transpose(x0),B);

$$yo := \begin{bmatrix}52.199985\end{bmatrix}$$

Et 95% konfidensinterval for y_0 er bestemt ved

$$\left[\tilde{y}_0 - t_{1-\frac{\alpha}{2}}(n-p)\cdot\sqrt{\tilde{\sigma}^2\cdot\bar{x}_0^T\cdot(X^T\cdot X)^{-1}\cdot\bar{x}_0}\,;\tilde{y}_0 + t_{1-\frac{\alpha}{2}}(n-p)\cdot\sqrt{\tilde{\sigma}^2\cdot\bar{x}_0^T\cdot(X^T\cdot X)^{-1}\cdot\bar{x}_0}\right].$$

Udtrykket $\bar{x}_0^T\cdot(X^T\cdot X)^{-1}\cdot\bar{x}_0$ kan findes ved Maple-ordren

> rod:=multiply(transpose(x0),invV,x0);

$$rod := \left[\frac{3}{4}\right]$$

Vi får derfor:

$$\left[52.2 - t_{0.975}(8-3)\cdot\sqrt{0.601667\cdot 0.75};52.2 + t_{0.975}(8-3)\cdot\sqrt{0.601667\cdot 0.75}\right]$$

$$=\left[52.2 - 2.57\cdot\sqrt{0.4512};52.2 + 2.57\cdot\sqrt{0.4512}\right] = \underline{\underline{[50.47\,;53.93]}}.$$

Af nedenstående Statgraphics-udskrift ses de samme resultater.

Regression Results for y

Row	Fitted Value	Stnd. Error for Forecast	Lower 95,0% CL for Forecast	Upper 95,0% CL for Forecast	Lower 95,0% CL for Mean	Upper 95,0% CL for Mean
9	52,2	1,02612	49,5623	54,8377	50,4732	53,9268

◆

A4.6 Residual-analyse.

Plotning af residualerne $r_i = y_i - \tilde{y}_i$ spiller en vigtig rolle ved vurdering af modellen.
Dette er vist i adskillige eksempler. Mange statistikprogrammer (også Statgraphics) beregner og plotter også de såkaldte "studentized residuals", som er velegnede til at opdage såkaldte "outliers", dvs. enkelte målinger, som afviger så kraftig fra den tendens (model) som de øvrige målinger viser, at man kan forestille sig, at de enten må være fejlmålinger, eller der er en ukendt faktor, som spiller ind.
De forskellige statistikprogrammer beregner disse residuals på lidt forskellig måde, men den grundlæggende ide er, at da konfidensintervallet jo bliver bredere jo længere væk fra "centrum" af talmaterialet man er, betyder en stor residual langt fra centrum mindre end hvis den lå tæt ved centrum.
Man standardiserer derfor residualerne ved at dividere med spredningen på den enkelte residual.

Lad $h_{ii} = \bar{x}_i^T (X^T \cdot X)^{-1} \bar{x}_i$.

Vi har da, at $\tilde{\sigma}^2 \cdot h_{ii}$ indgår i udtrykket for konfidensintervallet for middelværdien af en estimeret y-værdi, og er udtryk for variansen af den y-værdi $\tilde{y}_i$, som svarer til den i'te række $\bar{x}_i^T = [1, x_{i1}, x_{i2}, ..., x_{ik}]$ i X-matricen.

Idet $y_i = \tilde{y}_i + r_i$ virker det rimeligt, at man kan vise, at

$$V(r_i) = \sigma^2 - \sigma^2 \cdot h_{ii} = \sigma^2 (1 - h_{ii}), \qquad i = 1, 2, \ldots, n.$$

Dette medfører, at formlen for den i'te"internally studentizid residuals" sr_i er

$$sr_i = \frac{r_i}{\sqrt{\tilde{\sigma}^2 \cdot (1 - h_{ii})}}, \qquad i = 1, 2, \ldots, n, \qquad \text{hvor } h_{ii} = \bar{x}_i^T (X^T \cdot X)^{-1} \bar{x}_i .$$

Ved at dividere r_i med spredningen på r_i får man "normeret" residualerne, således at egentlige "outliers" træder tydeligere frem. Værdier på over 3 bør i hvert fald give anledning til overvejelser .

Det skal bemærkes, at Statgraphics benytter samme ide, men en lidt anden formel, kaldet "the externally Studentized residual" $tr_i = sr_i\sqrt{\dfrac{n-p-1}{n-p-sr_i^2}}$

Appendix 5.1 Sammenhæng mellem virkninger og regressionskoefficienter ved 2^k faktorforsøg.

Hvis man ikke har andet end lommeregner til rådighed, er det vigtigt, at regningerne er enkle. Som det fremgår af det følgende er dette tilfældet, idet den statistiske analyse forløber på samme måde enten faktorerne er kvantitative eller kvalitative. Kun tolkningen af resultaterne er forskellig.

Lad en kvantitativ variabel x have værdien x_{lav} når den er på lavt niveau og $x_{høj}$ når den er på højt niveau, og lad $h = |x_{høj} - x_{lav}|$ være afstanden mellem de to x-værdier og lad $\bar{x}$ være deres gennemsnit.

Der indføres nu en ny variabel $z = 2\frac{x-\bar{x}}{h}$

Det ses, at z har den egenskab, at $z_{høj} = 2\frac{x_{høj}-\bar{x}}{h} = 1$ og $z_{lav} = 2\frac{x_{lav}-\bar{x}}{h} = -1$

Indføres disse normerede variable i regressionsligningerne, vises nedenfor, at de fra 2^k - faktorforsøgene kendte virkninger (jævnfør kapitel 2) = regressionskoefficienterne i disse transformerede regressionsligninger, og variansanalysetabellerne er de samme (se SF III side 18 nederst og side 20 øverst).

Eksempel: Med 3 faktorer A, B og C på 2 niveauer, bliver regressionsligningen

$$Y = A_{000} + B_{100} z_1 + B_{010} z_2 + B_{001} z_3 + B_{110} z_1 z_2 + B_{101} z_1 z_3 + B_{011} z_2 z_3 + B_{111} z_1 z_2 z_3. \quad (1)$$

hvor z_1 = - 1 for A på lavt niveau, og 1 for A på højt niveau,
z_2 = - 1 for B på lavt niveau, og 1 for B på højt niveau og
z_3 = - 1 for C på lavt niveau, og 1 for C på højt niveau .
Der gælder da:

$A_{000} = \bar{\mu}$, $B_{100} = \tilde{A}$, $B_{010} = \tilde{B}$, $B_{001} = \tilde{C}$, $B_{110} = \widetilde{AB}$,
$B_{101} = \widetilde{AC}$, $B_{011} = \widetilde{BC}$ $B_{111} = \widetilde{ABC}$.

Bevis:

k=1 <u>A er kvalitativ (lavt, højt niveau).</u>

Der gælder $\mu_{(1)} = \bar{\mu} - \tilde{A}$,

$$\mu_{(a)} = \bar{\mu} + \tilde{A} \qquad (1)$$

<u>A er kvantitativ med x på 2 niveauer.</u>

x normeres som beskrevet ovenfor:

Regressionsligningen gennem de to punkter er da $Y = A_0 + B_1 z$,hvor z = - 1 for A på lavt niveau, og z = 1 for A på højt niveau.

Indsættes z = - 1 fås $\mu_{(1)} = A_0 - B_1$,

og z = 1 $\mu_{(a)} = A_0 + B_1$ (2)

Af ligningssystemerne (1) og (2) ses, at $\mathbf{A_0} = \bar{\mu}$ **og** $\mathbf{B_1} = \tilde{A}$ (hældningskoefficienten).

k = 2

<u>A og B er kvalitativ (lavt, højt niveau).</u>

Der gælder

$$\begin{aligned} \mu_{(1)} &= \bar{\mu} - \tilde{A} - \tilde{B} + \widetilde{AB}, \\ \mu_{(a)} &= \bar{\mu} + \tilde{A} - \tilde{B} - \widetilde{AB}, \\ \mu_{(b)} &= \bar{\mu} - \tilde{A} + \tilde{B} - \widetilde{AB}, \qquad (3) \\ \mu_{(ab)} &= \bar{\mu} + \tilde{A} + \tilde{B} + \widetilde{AB}, \end{aligned}$$

<u>A og B er kvantitative med x_1 og x_2 hver på 2 niveauer.</u>
x_1 og x_2 normeres som tidligere nævnt
Regressionsligningen bliver $Y = A_{00} + B_{10} z_1 + B_{01} z_2 + B_{11} z_1 z_2$.
hvor $z_1 = -1$ for A på lavt niveau, og 1 for A på højt niveau.
og $z_2 = -1$ for B på lavt niveau, og 1 for B på højt niveau.
Ved at indsætte -1 og 1 fås

$$\begin{aligned} \mu_{(1)} &= A_{00} - B_{10} - B_{01} + B_{11} \\ \mu_{(a)} &= A_{00} + B_{10} - B_{01} - B_{11} \\ \mu_{(b)} &= A_{00} - B_{10} + B_{01} - B_{11} \qquad (4) \\ \mu_{(ab)} &= A_{00} + B_{10} + B_{01} + B_{11} \end{aligned}$$

Af ligningssystemerne (3) og (4) ses, at $A_{00} = \bar{\mu}$, $B_{10} = \tilde{A}$, $B_{01} = \tilde{B}$ og $B_{11} = \widetilde{AB}$.

k= 3: Med 3 faktorer A, B og C på 2 niveauer, bliver regressionsligningen

$$Y = A_{000} + B_{100} z_1 + B_{010} z_2 + B_{001} z_3 + B_{110} z_1 z_2 + B_{101} z_1 z_3 + B_{011} z_2 z_3 + B_{111} z_1 z_2 z_3. \qquad (1)$$

hvor $z_1 = -1$ for A på lavt niveau, og 1 for A på højt niveau,
$z_2 = -1$ for B på lavt niveau, og 1 for B på højt niveau og
$z_3 = -1$ for C på lavt niveau, og 1 for C på højt niveau .
På samme måde som for k = 1 og k = 2 kan man vise, at
$A_{000} = \bar{\mu}$, $B_{100} = \tilde{A}$, $B_{010} = \tilde{B}$, $B_{001} = \tilde{C}$, $B_{110} = \widetilde{AB}$,
$B_{101} = \widetilde{AC}$, $B_{011} = \widetilde{BC}$ $B_{111} = \widetilde{ABC}$.

Tilsvarende relationer findes naturligvis for k = 4 osv.

2^k faktorforsøg hvor der er en blanding af kvantitative og kvalitative faktorer

a) A og B kvantitative, C kvalitativ

For C (z_3) kvalitativ omskrives regressionsligningen (1) til

$$Y = (A_{000} + B_{001} z_3) + (B_{100} + B_{101} z_3) z_1 + (B_{010} + B_{011} z_3) z_2 + (B_{110} + B_{111} z_3) z_1 z_2$$

Vi har følgelig for C på højt niveau ($z_3 = 1$) et polynomium,
og for C på lavt niveau ($z_3 = -1$) et andet polynomium.
Ved at sammenholde disse ligninger, ses umiddelbart, hvorledes der skal testes:
Eksempel: Ønskes testet om koefficienterne til produktleddet er ens, ses, at det må betyde at vi skal teste om ABC =0 (da $B_{111} = \widetilde{ABC}$)

Se eventuelt videre SF III side 24.

b) A kvantitativ, B og C kvalitativ
På ganske analog måde, fås de resultater, som kan ses i SF III side 28 nederst og side 29 øverst.

Appendix A. Typiske ordrer i Statgraphics

a1. Indledning

I kurset i forsøgsplanlægning anvendes fortrinsvis følgende 4 Statgraphics programmer.

Compare: Sidedelte variansanalyse med lutter fixed kvalitative (catagorical) faktorer.
Relate: Regressionsanalyse med lutter kvantitative faktorer
Experimental design: 2^k faktorforsøg (screening design)
GLM: Mixed analyser, trinvise, sidedelte, blandende kvalitative og kvantitative analyser.

GLM (General linear models) kan i princippet benyttes til alle former for analyser, men de 3 øvrige programmer er indenfor deres anvendelsesområde lettere at anvende, idet de har indbygget nogle arbejdsbesparende hjælpefunktioner.

Endvidere anvendes **Describe** i visse situationer, bl.a. ved tegning af normalfordelingsplot og beregning af sandsynligheder (P - værdier).

a2 Generelle forhold ved opstart.

Efter at have startet Statgraphics, står man med en typisk Windows skærm med nogle menubjælker og ikoner.
Øverst er en menubjælke med navnene File, Edit, Plot, Describe, Compare osv. Trykkes på en af disse fremkommer en rullemenu, som man skal vælge fra.
Nedenunder denne er der en menubjælke med en række ikoner, hvoraf nogle er kendte fra Windows (eksempelvis Print), andre er genveje til almindeligt anvendte programmer, så man ikke behøver at gå over den øverste menubjælke.
På skærmen er der et regneark (hvis ikke så vælg ikonen "Untitled" forneden på skærmen og vælg "Restore")
Nederst på skærmen er igen en menubjælke, som starter men "Untitled Co...", StatAdvisor osv.
Når man har indtastet data, valgt eksempelvis Compare og derefter har specificeret hvilken analyse der skal foretages, vil der yderligere dukke en menubjælke op foroven som henfører til de valgmuligheder man har indenfor dette specielle program.
De fire første ikoner i denne menubjælke går igen i de fleste menuer.
"Input Dialog" : Er rød og giver mulighed for at ændre den statiske model, hvad der er "Dependent variable" , "Factors" osv.
"Tabular Options": Er gul og giver en række valgmuligheder for analyser.
"Graphical Options": Er blå giver forskellige grafiske muligheder for vurdering af analysen.
"Save Results": Er sort og giver muligheder for at gemme resultater i dataarket til brug ved senere analyser.

a2.1 Indtastning af data

Placer cursor på kolonne 1 (Col 1)og tryk først på venstre musetast for at vælge kolonnen, og derefter på højre musetast for at få en lille menu frem.
Vælg "Modify column", og giv søjlen et relevant navn
Hvis faktoren er et stof og er en "karaktervariabel" (ikke talvariabel), så giv søjlen navnet "stof" og vælg "Type" som "Character".
Placer cursor på cellen i første række og første søjle, tryk på venstre musetast, og skriv et stofnavn eksempelvis "T1". Brug piletast til at gå ned til næste række og skriv T2 osv.
Gentag proceduren med søjle 2. Hvis denne er en variabel med lutter tal, så skal man ikke ændre Type, da den automatisk er sat til "numeric". Lad søjlen blive kaldt "Urenhed"
Resultatet ser således ud:

```
Stof     Urenhed
T1        108
T1        110
T2        105
T2        110
T3        108
T3        111
T4        117
T4        119
osv.
```

Når regnearket er udfyldt med alle data, trykkes på funktionstasten F12, og man kan enten gemme data på harddisken i et passende katalog, eller hvis det er i databaren på en medbragt diskette med a:\sf2eks1. Bemærk, at filnavnet bliver sf2eks1.sf3, dvs. automatisk får endelsen .sf3.
Man kan eventuelt "minimize" datafilen.

a3. Compare

<u>Anvendelse</u>: Sidedelte variansanalyser med lutter fixed kvalitative (categorical) faktorer.

a3.1 Ensidet variansanlyse.

<u>Opstilling af variansanalysetabel.</u>
Lad starten af indtastningen i regnearke være

```
Stof     Urenhed
T1        108
T1        110
T2        105
```

Vælg (Compare | Analysis of Variance | One-Way Anova)
Der fremkommer nu en tabel, der skal udfyldes:
Vælg (Klik på "Urenhed"|klik dernæst på pilen ved "Dependent variable"|Klik på "stof"|klik på pilen "Factor"| OK)
Der fremkommer en "Analysis Summary" med nogle statistiske betragtninger.
Endvidere fremkommer et "scatterplot". Scatterplottet kan man om ønsket fjerne ved med cursoren på tabellen hurtigt 2 gange at trykke på venstre musetast.
Vælg(Tabular Options|Anova Tables| OK) (fjern evt. krydset ved "Analysis summary")

<u>Opstilling af konfidensintervaller.</u>
Vælg(Tabular options |Tables of Means | OK). Herved fås LSD-intervaller.

Ønskes de sædvanlige konfidensintervaller, så:
Vælg (Med cursor på udskrift, tryk på højre musetast|Pane options|Confidence intervals (pooled s)|OK).
Plot af konfidensintervaller: LSD-intervaller: Vælg (Blå ikon =Graphics options | Means plot| OK). Ønskes de almindelige konfidensintervaller, så:
Vælg (Med cursor på tegning, tryk på højre musetast|Pane options|Confidence intervals (pooled s)|OK).

a3.2 Flersidet variansanlyse.

Opstilling af variansanalysetabel

Lad starten af indtastningen i regnearket være

```
A  B  C D haerdetid
1  1  1 1   23.8
2  1  1 1   45.8
```

hvor A, B C og D er de 4 faktorer, og haerdetid er forsøgsvariablen.
Vælg (Compare | Analysis of Variance | Multifactor ANOVA | OK).
Der fremkommer nu en tabel der skal udfyldes:
Vælg (“haerdetid”|pilen ved “Dependent variable”|A | pilen “Factors” |B | pilen “Factors | C | pilen “Factors |D | pilen “Factors | OK)
Der fremkommer en “Analysis Summary ” og et “Scatterplot”.
Vælg (gul ikon = Tabular options| Anova Tables |OK).
Der fremkommer en variansanalysetabel uden vekselvirkning.
Vælg (Cursor i tabellen | højre musetast |Analysis options | Maximum Order Interaction til 4 | OK)
Der fremkommer en tabel med vekselvirkninger.
Pooling af virkninger: Det antages, at 3 og 4 faktorvekselvirkninger er nul.

Vælg (Cursor i tabellen | højre musetast |Analysis options | Maximum Order Interaction til 2 | OK) Derved fremkommer en ny tabel.
Lad os antage, at kun vekselvirkningerne AD og BD har betydning.
Vælg (Cursor i tabellen | højre musetast |Analysis options | Exclude). Herved fås et lille vindue med 2 kolonner (“Include” og “Exclude”), hvor man ved dobbeltklik på en vekselvirkning flytter den til den modsatte kolonne. Vi “excluderer” nu alle tofaktorvekselvirkningerne, der ikke har betydning. Derved fremkommer en ny tabel.
Viser det sig, at en hovedvirkning ikke har betydning (eksempelvis C) slettes denne af modellen: Vælg (rød ikon = Input dialog | Slet C | OK).
Dette har den lidt kedelige virkning, at der i visse udskrifter sker en omdøbning, så faktor D bliver kaldt C.

<u>Optimal værdi af forsøgsvariabel og de dertil svarende værdier af faktorerne:</u>
a) Antagelse: Der er vekselvirkning.
Vi laver derfor 95% konfidensintervaller for hver af behandlingerne.
Vælg (Compare | Analysis of Variance | Multifactor ANOVA | OK)., og indtast modellen.
Vælg (gul ikon = Tabular options | Tables of Means | OK).
<u>Grafisk fremstilling:</u>
Vælg (Compare | Analysis of Variance | Multifactor ANOVA | OK)., og indtast modellen.
Vælg (blå ikon = Graphics Options | Interaction Plot |Cursor på figur | højre musetast | Pane options | Confidence intervals |OK)
Eventuelt skifte til " second Factor" , hvis det giver en mere overskuelig tegning.

b) Antagelse: En eller flere hovedvirkninger ønskes undersøgt.
Vælg(Tabular options |Tables of Means | OK),
Den fremkomne tabel er over 95% "LSD" konfidensintervaller.
Ønskes de sædvanlige 95% konfidensintervaller så
Vælg (Med cursor på udskrift, tryk på højre musetast|Pane options|Confidence intervals (pooled s)|OK).
<u>Grafisk fremstilling:</u>
Vælg (blå ikon = Graphics Options | Means Plot |Cursor på figur | højre musetast | Pane options | Confidence intervals og vælg faktor, der ønskes undersøgt|OK)

<u>Kontrol af model</u>
<u>Residualplot af model:</u>
Vælg (blå ikon = Graphics options |Residuals versus Predicted | OK).

<u>Grafisk test for normalitet</u> (normalfordelingsplot):
Vælg (sort ikon = Save Results |Save Residuals | OK).
Residualerne bliver nu gemt som en søjle i data under navnet "RESIDUALS", da vi ikke har ændret navnet under "Target Variables"
Vælg (Describe |Numerical Data |One Variable Analysis |RESIDUALS |Pilen Data| OK).
Vælg (blå ikon = Graphics options |Normal Probability Plot| OK).
Ønskes linien placeret mere"ligeligt" kan man eventuelt med cursoren på figuren trykke på højre musetast og på den fremkomne menu vælge "Least Squares".

Undersøgelse af varianshomogenitet.

Lad os antage, at vi har 6 behandlinger, som hver gentages 2 gange, eksempelvis således:

		Karburator	
		K_1	K_2
Olieblanding	O_1	× ×	× ×
	O_2	× ×	× ×
	O_3	× ×	× ×

Man går ind i regnearket og danne en ekstra søjle "behandlinger" . Da der er 6 behandlinger (celler) med 2 tal i hver bliver den:

karburator	oliebland	benzinforb	behandlinger
k1	o1	830	1
k1	o1	860	1
k1	o2	940	2
k1	o2	990	2
k1	o3	855	3
k1	o3	815	3
k2	o1	810	4
k2	o1	840	4
k2	o2	1050	5
k2	o2	1020	5
k2	o3	930	6
k2	o3	910	6

Vælg (Compare | Analysis of Variance | One Way ANOVA |OK).

Vælg (Klik på "benzinforb" og pil ved "Dependent variable"|Klik på "behandlinger" og på pil ved "Factor"| OK)

Vælg (gul ikon = Tabular options | Variance check | OK).

fjern evt. krydset ved "Analysis summary

a4. Experimental design: Anvendelse: 2^k faktorforsøg (screening design).

a4.1 Indtastning af data

Lad os antage vi har et forsøg med 4 faktorer, stabilisator, konserveringsmiddel, syrekoncentration og temperatur, og forsøgsvariablen er "udbytte".
Vælg (Speciel | Experimental Design | Create Design | Screening | sæt "No of response" = 1 | "No of experimental Factor"=4 | OK).

1) Hvis man vil beholde Statgraphics standardnavne trykkes på OK og ved næste menu også på OK.
2) Vil man skifte til andre "egne" navne skal man i næste menu"Factor Definition Options" under "Name" skrives "Stabilisator". Under "Factor" markeres B, og under "Name" skrives "konserveringsmiddel", osv.
   Efter OK får man menuen "Response Definition Options" hvor vi kan skrive "udbytte".

I den næste menu fastholdes krydset ved "Display Blocked Designs", hvis der indgår blokke.
Ved at trykke på pilen i højre side af den fremhævede linie ses nu alle de indbyggede muligheder.

1) Den ønskede forsøgsplan findes blandt mulighederne.
   Vi vælger denne, og vil på den næste menu "Screening Design Options" se antal runs. Hvis der eksempelvis er 3 fuldstændige blokke, vil der være 3 gentagelser. Man sætter så "Replicate Design" til 2 (3 blokke giver 2 ekstra udførelser), hvorved vi ser antal runs ændre sig til det forventede antal.
   Endvidere sletter vi krydset ved Randomize, hvis vi ønsker forsøgene skrevet i standardrækkefølge. Vælges Randomize, bliver rækkefølgen af forsøgene tilfældig, sådan som det bør være, hvis man selv skulle lave forsøgene (randomiseret). Tryk på OK.
   Der ses nu en plan med Statgraphics standardnavne såsom -1 for lav og 1 for højt niveau.
   Vælg (gul ikon = Tabular Options | Alias Structure | OK).
   Derved fremkommer en udskrift af aliasstrukturerne, og vi kan her se, om forsøgsplanen opfylder de stillede krav.
   Valgte vi "Randomize" kan vi få skrevet en randomiseret arbejdsplan ud (behandlingerne er skrevet i tilfældig (randomiseret) rækkefølge) ved ordren:
   Vælg (gul ikon = Tabular Options | Worksheet | OK).
   Vi går nu ned i bunden af skærmen og vælger et regneark "Untitled". Vi ser at regnearkets faktorer allerede er udfyldt, så vi indtaster blot "udbytterne" .

2) Den ønskede forsøgsplan findes **ikke** blandt mulighederne.
   Man vælger "User specified design" (findes nederst)
   I den næste menu "Screening Design Options" sletter vi krydset ved Randomize, hvis vi ønsker forsøgene skrevet i standardrækkefølge.
   Der ses nu en udskrift "Design Summary" med en oversigt over navne, niveauer m.m.
   Vi går nu ned i bunden af skærmen og vælger et regneark "Untitled". Vi bliver her nødt til at udfylde alle faktorers niveauer i overensstemmelse med den forsøgsplan man på forhånd har konstrueret.

Den indtastede datafil bør gemmes (saves) (den får automatisk "endelsen" .sfx), og kan så senere hentes frem ved ordren: Vælg (Speciel | Experimental Design | Open Design | OK).

a4.2 Bestemmelse af model.

Vælg (Speciel | Experimental Design | Analyze Design).
Tryk på "udbytte" og derefter på $\Rightarrow$ under "Data" for at fortælle, at det er denne variabel der skal analyseres.
Dette resulterer i en tabel over virkningerne, samt et histogram (Pareto charts" over dem.
Hvis de er sammenblandede kan man her også se aliasrelationer.
Variansanlysetabel: Vælg (gul ikon = Tabular Options | ANOVA | OK).
Pooling af virkninger: Skal nogle vekselvirkninger eksempelvis AB og AC udelukkes, så
Vælg (Cursor i udskrift | højre musetast | Analysis Options | Exclude | dobbeltklik på AB, AC (hvorved de rykker over på højre side) | OK | OK).

a4.3. "Optimal" indstilling af faktorerne:

Efter at man har fundet den endelige model bestemmes faktorernes indstillinger:
1) Vælg (gul ikon = Tabular Options | Optimization | OK).
2) For at få de relevante virkninger
   Vælg (gul ikon = Tabular Options | Analysis Summary | OK).
3) Til tolkning af vekselvirkninger tegnes interaction plot:
   Vælg (blå ikon = Graphics plot |Interaction plot | OK).

a4.4 Bestemmelse af optimal værdi og dertil hørende konfidensinterval:

Indgår den optimale indstilling af faktorerne ikke blandt de oprindelige data tilføjes den i regnearket som den sidste observation.
Derefter kan optimal værdi og konfidensinterval findes ved ordren
Vælg (gul ikon = Tabular Options | Predictions | OK).

a4.5 Kontrol af model

Residualplot af model:
Vælg (blå ikon = Graphics options |Diagnostic plot | OK).
Grafisk test for normalitet:
Vælg (Højreklik på graf|Pane Options|Normal Probability Plot of Residuals| OK).
Undersøgelse af varianshomogenitet.
Kan ske som under Compare.

a5. GLM (General Linear Models):

Anvendelse: Mixed analyser indeholdende trinvise faktorer og analyser med blandende kvalitative og kvantitative faktorer.

a5.1 Indtastning af data

Foregår som ved Compare.

a5.2 Bestemmelse af model.

a5.2.1 Split-Plot forsøg

Som eksempel tages landbrugsforsøget med pløjemetoder i eksempel 3.7.
Modellen er $Y = \mu + P + M(P) + G + P*G + M(P)*G + \varepsilon$
Vælg (lilla GLM ikon) eller (Special | Advanced Regression | General Linear Methods).
Den fremkomne tabel udfyldes med
Dependent Variable : Udbytte
Catagorical Factors : ploejemetode, markstrimmel , goedning,
I det følgende vil de variable noget uhensigtsmæssigt blive omdøbt til A, B, C osv.
På næste tavle der hedder "GLM Model Specifikation" skrives
A (svarer til ploejemetode)
B(A) (svarer til at markstrimmel er undertrin af ploejemetode)
C (svarer til goedning)
A*C
Pooling
Lad os antage, at vekselvirkningen A*C = ploejemetode $\cdot$ goedning er 0.
Vælg (gul ikon = input dialog | OK | fjern A*C | OK)
Valg af teststørrelse
Fremgår det af varianskomponenttabellen, at B(A)=markstrimmel(ploejemetode) kan benyttes til testning af A= ploejemetode sker det på følgende måde:.
Vælg (Cursor i udskrift| højre musetast | Analysis options).
I det fremkomne felt under "Factor" sættes blå streg på A. I feltet med Error Term vælges B(A). Derved ses i nederste felt "Selections" , at der står A - B(A) som tegn på at A testes mod B(A).

For "fixed factors" finde den værdi der giver den optimale værdi af forsøgsvariablen:
Vælg (gul ikon = Tabular options | Tables of Means | OK).

a5.2.2 Forsøg med både kvantitative og kvalitative faktorer.

Som eksempel vælges et forsøg (eksempel 5.4), hvor man undersøger sammenhængen mellem mængden Y og to kvalitative faktorer A og B og en kvantitativ faktor T, idet man som udgangspunkt antager, at Y for hver af de 6 kombinationer af de 2 kvalitative faktorer A og B kan beskrives ved et andengradspolynomium i T, dvs. af formen

$Y = \alpha_{ij} + \beta_{1ij}T + \beta_{2ij}T^2$, hvor i = 1,2 (*additiv*) , j = 1, 2, 3 (*katalysator*)

<u>Variansanlysetabel</u>
Vælg ikonen "GLM" eller vælg (Special | Advanced Regression | General Linear Models")
Den fremkomne tabel udfyldes med
"Dependent variables" : Y
"Categorical Factors" : A og B
og "Quantitative Factors" : T
I "GLM specifications" hvor A= A, B = B og C = T udfyldes med
A, B, C, A*B, A*C, B*C, A*B*C, C*C, A*C*C, B*C*C, A*B*C*C.

<u>Pooling</u>
Dette sker som ovenfor, eksempelvis
Vælg (gul ikon = input dialog | OK | fjern A*C*C | OK)

<u>Opskriv regressionsligninger.</u>
Vælg (gul ikon = Tabular Options | Model Coefficients | OK).

<u>Konfidensintervaller for en regressionskoefficient</u> β : Ses af tabel for "Model Coefficients".

<u>Estimat af middelværdi og 95% konfidensinterval for givne værdier af faktorerne:</u>
Indsæt de givne værdier i sidste række af inddata, og
Vælg (gul ikon = Tabular Options | Report | OK).

a5.3 Kontrol af model.

<u>Residualplot</u>: (Blå ikon =Graphical options |Residual plots |OK).
Vi får hermed "Studentized Residuals"
Ønskes de sædvanlige residualer:
(Cursor på figur | højre musetast |Pane Options |Vælg Residuals, Scatter plot, Predicted value).
<u>Normalfordelingsplot</u>: (Cursor på figur | højre musetast |Normal Probability Plot |OK)

a6. Relate :

Anvendelse: Regressionsanalyse med lutter kvantitative faktorer

a6.1. Indtastning af data

Foregår som ved Compare.

a6.2 Regressionsanalyse med 1 faktor

a6.2.1. Enkelt Regressionsanalyse.

Regressionsanalysetabel opstilles: Vælg (Relate | Simple Regression |indsæt t og y |OK). Har vi gentagelser ignoreres den fremkomne udskrift i første omgang, og vi udfører en "Lack of fit test": Vælg (gul ikon = Tabular Options | Lack of Fit test | OK).

Residualplot: Vælg (blå ikon= Graphical options | Residual versus predicted | OK)
Tegningen viser de såkaldte "studentized residuals" .
Sædvanlige residualer fås af: (Cursor på tegning | højre musetast |Residuals).

Undersøgelse af Outliers: Vælg(Tabular Options | Unusual Residuals|OK).

Normalfordelingsplot: Som under Compare

Varianshomogenitet: Som under Compare

Konfidensinterval for hældningskoefficienten $\tilde{\beta}_1$:

Vælg (Relate | Polynomial Regression | indsæt t og y | cursor i udskrift | højre musetast | Analysis Options |sæt order til 1 | gul ikon = Tabular options | Confidence intervals |OK).

95% konfidensinterval for det til $t = 100$ svarende værdi af y:
Vælg (Relate | Simple Regression |indsæt t og y |OK).
Vælg (Gul ikon = Tabular Options | Forecast | OK | Cursor på udskrift | højre musetast | Pane Options) . Sæt i skemaet "Forecast at x" til 100 og resten stryges.

Transformation af model. Vælge eksponentiel model
Vælg (med cursor på udskrift, højre musetast|Analysis options|Exponential|OK)

Finde "bedste" model:
Vælg(gul ikon = Tabular options|Comparison of Alternative Models|OK)

a6.2.2 Polynomial Regressionsanalyse.

Vælg (Relate| Polynomial Regression | indsæt x og T i den fremkomne tavle | OK)
For at få en trediegradsmodel:
Vælg (Cursor på udskrift | højre musetast | Analysis Options |order til 3 | OK).
De øvrige ordrer er de samme som ved "Enkelt regressionsanalyse."

a6.3 Multipel Regressionsanalyse.

a6.3.1 Lineær model

Regressionsanalysetabel opstilles: Vælg (Relate | Multiple Regression |indsæt de variable |OK).
De øvrige ordrer er som ovenfor beskrevet, idet man dog ikke kan lave en “lack of fit test” direkte.

a6.3.2 Polynomial model

Lad os antage modellen er $Y = \alpha_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2$.
Vælg ikonen "Multiple Regression" eller Vælg (Relate | Multiple Regression).
I den fremkomne tavle under “Indenpendent Variable” indsættes x1, x2, x1^2, x2^2, x1*x2

a7. Describe

a7.1. Beregning af *P* - værdier.

Vi ønsker at finde P - værdien $= P(X > 3.9814)$, hvor X er F - fordelt med et frihedsgradstal for tæller på 3 og nævner på 6.
Vælg(Describe|Distributions|Probability Distributions|F(Variance Ratio)|OK)
Vælg (Tabular Options|Cumulative Distributions|med cursor på udskrift, højre musetast|Pane Options|Random Variable = 3.9814|OK|med cursor på udskrift, højre musetast|Analysis Options|Numerator = 3, Denominator =6|OK)

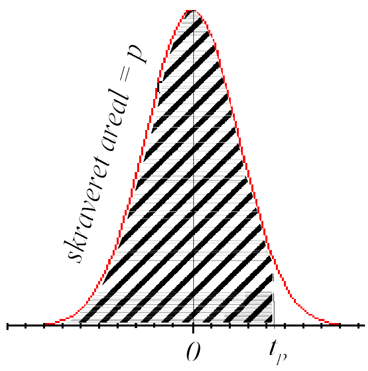
a7.2. Beregning af en t-fraktil

Vi ønsker at finde fraktilen $t_{0.975}(8)$.
Vælg(Describe|Distributions|Probability Distributions|Student’s t|OK)
Vælg (Tabular Options|Inverse CDF|med cursor på udskrift, højre musetast|Pane Options|CDF = 0.975|OK|med cursor på udskrift, højre musetast|Analysis Options|Degree og Fredom = 8|OK).
Resultat = 2.228

STATISTISKE TABELLER

Tabel 1 Fraktiler t_p i t-fordelingen t(f).

$$P(T \# t_p) = p \text{ , hvor } T = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$



f \ p	0.60	0.75	0.90	0.95	0.975	0.99	0.995	0.999	0.9995
1	0.33	1.00	3.08	6.31	12.7	31.8	63.7	318	637
2	0.29	0.82	1.89	2.92	4.30	6.97	9.93	22.3	31.6
3	0.28	0.74	1.64	2.35	3.18	4.54	5.84	10.2	12.9
4	0.27	0.74	1.53	2.13	2.78	3.75	4.60	7.17	8.61
5	0.27	0.72	1.48	2.02	2.57	3.37	4.03	5.89	6.86
6	0.27	0.72	1.44	1.94	2.45	3.14	3.71	5.21	5.96
7	0.26	0.71	1.42	1.90	2.37	3.00	3.50	4.79	5.41
8	0.26	0.71	1.40	1.86	2.31	2.90	3.36	4.50	5.04
9	0.26	0.70	1.38	1.83	2.26	2.82	3.25	4.30	4.78
10	0.26	0.70	1.37	1.81	2.23	2.76	3.17	4.14	4.59
11	0.26	0.70	1.36	1.80	2.20	2.72	3.11	4.03	4.44
12	0.26	0.70	1.36	1.78	2.18	2.68	3.06	3.93	4.32
13	0.26	0.69	1.35	1.77	2.16	2.65	3.01	3.85	4.22
14	0.26	0.69	1.35	1.76	2.15	2.62	2.98	3.79	4.14
15	0.26	0.69	1.34	1.75	2.13	2.60	2.95	3.73	4.07
16	0.26	0.69	1.34	1.75	2.12	2.58	2.92	3.69	4.02
17	0.26	0.69	1.33	1.74	2.11	2.57	2.90	3.65	3.97
18	0.26	0.69	1.33	1.73	2.10	2.55	2.88	3.61	3.92
19	0.26	0.69	1.33	1.73	2.09	2.54	2.86	3.58	3.88
20	0.26	0.69	1.33	1.73	2.09	2.53	2.85	3.55	3.85
21	0.26	0.69	1.32	1.72	2.08	2.52	2.83	3.53	3.82
22	0.26	0.69	1.32	1.72	2.07	2.51	2.82	3.51	3.79
23	0.26	0.69	1.32	1.71	2.07	2.50	2.81	3.49	3.77
24	0.26	0.69	1.32	1.71	2.06	2.49	2.80	3.47	3.75
25	0.26	0.68	1.32	1.71	2.06	2.49	2.79	3.45	3.73
26	0.26	0.68	1.32	1.71	2.06	2.48	2.78	3.44	3.71
27	0.26	0.68	1.31	1.70	2.05	2.47	2.77	3.42	3.69
28	0.26	0.68	1.31	1.70	2.05	2.47	2.76	3.41	3.67
29	0.26	0.68	1.31	1.70	2.05	2.46	2.76	3.40	3.66
30	0.26	0.68	1.31	1.70	2.04	2.46	2.75	3.39	3.65
40	0.26	0.68	1.30	1.68	2.02	2.42	2.70	3.31	3.55
50	0.26	0.68	1.30	1.68	2.01	2.40	2.68	3.26	3.50
60	0.25	0.68	1.30	1.67	2.00	2.39	2.66	3.23	3.46
80	0.25	0.68	1.29	1.66	1.99	2.37	2.64	3.20	3.42
100	0.25	0.68	1.29	1.66	1.98	2.37	2.63	3.17	3.39
120	0.25	0.68	1.29	1.66	1.98	2.36	2.62	3.16	3.37
200	0.25	0.68	1.29	1.65	1.97	2.35	2.60	3.13	3.34
500	0.25	0.68	1.28	1.65	1.97	2.33	2.59	3.11	3.31
∞	0.25	0.67	1.28	1.65	1.96	2.33	2.58	3.09	3.29

Eksempler: For t(27) er $P(X \leq 3.42)=0.999$. $t_{0.95}(1)=6.31$. $t_{0.05}(10) = -t_{0.95}(10) = -1.81$.

Tabel 2 95%-fraktil $F_{0.95}$ i F-fordelingen $F(f_T.f_N)$. $P(F \le F_{0.95}) = 0.95$

f_N \ f_T	1	2	3	4	5	6	7	8	9	10	12	14	16	18	20	25	30	40	60	120	4
1	161.5	199.5	215.7	224.6	230.5	234.0	236.8	238.9	240.5	241.9	243.9	245.4	246.5	247.3	248.0	249.3	250.1	251.1	252.2	253.3	253.7
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.42	19.43	19.44	19.45	19.46	19.46	19.47	19.48	19.49	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.71	8.69	8.67	8.66	8.63	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.87	5.84	5.82	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.64	4.60	4.58	4.56	4.52	4.50	4.46	4.43	4.40	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.96	3.92	3.90	3.87	3.83	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.53	3.49	3.47	3.44	3.40	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.24	3.20	3.17	3.15	3.11	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.03	2.99	2.96	2.94	2.89	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.86	2.83	2.80	2.77	2.73	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.74	2.70	2.67	2.65	2.60	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.64	2.60	2.57	2.54	2.50	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.55	2.52	2.48	2.46	2.41	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.48	2.44	2.41	2.39	2.34	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.42	2.38	2.35	2.33	2.28	2.25	2.20	2.16	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.37	2.33	2.30	2.28	2.23	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.33	2.29	2.26	2.23	2.18	2.15	2.10	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.29	2.25	2.22	2.19	2.14	2.11	2.06	2.02	1.97	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.26	2.21	2.18	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.23	2.18	2.15	2.12	2.07	2.04	1.99	1.95	1.90	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.20	2.16	2.12	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.17	2.13	2.10	2.07	2.02	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.15	2.11	2.08	2.05	2.00	1.96	1.91	1.86	1.81	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.13	2.09	2.05	2.03	1.97	1.94	1.89	1.84	1.79	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.11	2.07	2.04	2.01	1.96	1.92	1.87	1.82	1.77	1.71
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.04	1.99	1.96	1.93	1.88	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.95	1.90	1.87	1.84	1.78	1.74	1.69	1.64	1.58	1.51
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03	1.95	1.89	1.85	1.81	1.78	1.73	1.69	1.63	1.58	1.51	1.44
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.86	1.82	1.78	1.75	1.69	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83	1.78	1.73	1.69	1.66	1.60	1.55	1.50	1.43	1.35	1.25
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.69	1.64	1.60	1.57	1.51	1.46	1.39	1.32	1.22	1.00

OPGAVER

Opgave 1 (ensidet variansanalyse)
Fire forskellige typer teknik til blanding af cement ønskes undersøgt med hensyn til resultatets trykstyrke. Følgende data blev opnået:

Blandingstek-nik	Trykstyrke (psi)			
B1	3129	3000	2865	2890
B2	3200	3300	2975	3250
B3	2800	2900	2985	3050
B4	2600	2700	2600	2765

Undersøg om forskellen i blandingsteknik har betydning for trykstyrken , og angiv i bekræftende fald den (de) blandingsteknik(er) der har størst trykstyrke.

Opgave 2 (dimensionering)(eksamen 1973).
Ved et fuldstændigt randomiseret forsøg ønskes virkningerne af 4 forskellige behandlinger B_1, B_2, B_3 og B_4 sammenlignet. Forsøget skal dimensioneres således, at en eventuel forskel på 30 enheder mellem to middelværdier konstateres ved analysen af forsøgsresultaterne med en sandsynlighed på mindst 95%, når der udføres en sædvanlig 5%- signifikanstest.

1) Bestem antallet af gentagelser af hver behandling, når forsøgsfejlens spredning skønnes at være højst ca 9 enheder.
2) Hvilke konklusioner vedrørende behandlingernes virkninger vil du drage, hvis du finder
   a) signifikans?
   b) ikke finder signifikans?
3) Hvilke yderligere analyser af forsøgsresultaterne vil du udføre i de ovennævnte tilfælde a) og b)?

Opgave 3 (tosidet variansanalyse)
Man ønsker at undersøge den virkning som 2 faktorer (typen af glas og fosfor) har på skarpheden af billedet på en TV-skærm. Responsvariablen er den strøm (i microampere) som er nødvendig for at opnå et specifik skarpheds niveau.
Data er vist i nedenstående tabel:

		Fosfortype		
		1	2	3
Glastype	1	280 290 285	300 310 295	290 285 290
	2	230 235 240	260 240 235	220 225 230

Spørgsmål 1: Undersøg om forudsætningen om varianshomogenitet er opfyldt
Idet de sædvanlige regressionsanalyseforudsætninger antages opfyldt, ønskes følgende spørgsmål belyst:
Spørgsmål 2: Har glastype og fosfortype indflydelse på skarpheden?
Spørgsmål 3: Ud fra svaret i spørgsmål 1 skal angives, hvilken glastype og fosfortype der giver den største skarphed (giver den mindste respons)

Opgave 4 (dimensionering)
Et forsøg planlægges som et randomiseret forsøg med en fuldstændig faktorstruktur. Faktorerne A og B på henholdsvis 3 og 5 niveauer vides ikke at vekselvirke. Indledningsvis ønskes foretaget en dimensionering af forsøget.
Bagatelgrænsen i forsøget sættes til ca $\Delta = 1.0$, $\alpha = 5\%$ og $\beta = 10\%$. Fra tidligere forsøg vides, at forsøgsfejlens spredning er $\sigma \approx 0.3$.
Der ønskes bestemt antallet af gentagelser af hver behandling i forsøget.

Opgave 5 (flersidet variansanalyse) (øvelse 11 i SF1 omskrevet).
Ved en produktion ønskedes virkningen af 3 faktorer undersøgt:
A: Maskiner (3 forskellige anvendes)
B: Benyttet halvfabrikata (2 muligheder)
C: Produktionstidspunkt (2 muligheder)
Der var på forhånd en vis formodning om, at tidsfaktoren ingen virkning har.
Følgende observationer af produktionsudbyttet foretoges (kodede tal).

		Faktor C			
		C_1		C_2	
		B_1	B_2	B_1	B_2
Faktor A	A_1	35 37 38 37	24 28 26 24	32 38 36 35	27 30 24 26
	A_2	42 40 39 44	36 35 36 37	38 43 42 43	33 35 39 34
	A_3	42 47 48 49	41 47 40 48	50 43 48 47	40 44 39 50

Tallene er indtastet i Statgraphics og følgende tresidede variansanalysetabel er derved fremkommet.

```
Analysis of Variance for udbytte - Type III Sums of Squares
--------------------------------------------------------------------------------
Source                  Sum of Squares     Df    Mean Square    F-Ratio    P-Value
--------------------------------------------------------------------------------
MAIN EFFECTS
 A:A                           1597,63      2        798,813     100,03     0,0000
 B:B                            468,75      1         468,75      58,70     0,0000
 C:C                          0,333333      1       0,333333       0,04     0,8393

INTERACTIONS
 AB                             92,625      2        46,3125       5,80     0,0066
 AC                          0,0416667      2      0,0208333       0,00     0,9974
 BC                          0,0833333      1      0,0833333       0,01     0,9192
 ABC                           10,0417      2        5,02083       0,63     0,5390

RESIDUAL                         287,5     36        7,98611
--------------------------------------------------------------------------------
TOTAL (CORRECTED)               2457,0     47
```

Spørgsmål 1: Forklar hvordan du ud fra variansanalysetabellen kan se, at nulhypotesen H_0: ABC = 0 må accepteres.

Spørgsmål 2: Foretag en pooling, og test derefter nulhypotesen H_0: AB = 0 ved almindelig regning med lommeregner og brug af F tabel.

I spørgsmål 3 og 4 må du (men skal ikke) bruge Statgraphics:

Spørgsmål 3: Test om de øvrige 2 faktorvekselvirkninger kan antages at være nul.

Spørgsmål 4: Ud fra konklusionen i spørgsmål 1,2 og 3 skal man teste, om der er nogen af de 3 faktorer A, B og C der ikke har nogen virkning.

Spørgsmål 5: Hvis man af økonomiske grunde ønsker at benytte halvfabrikata B_1 hvilke maskiner og hvilket produktionstidspunkt skal da foretrækkes.
Du skal indsætte de korrekte tal i formlen for konfidensintervallet (t værdi, s osv), men kan eventuelt lade Statgraphics udregne intervallerne.

Opgave 6 (planlægning af randomiseret blokforsøg).
På en ingeniørskole ønsker man at sammenligne effektiviteten af undervisningen, når man underviser efter tre forskellige undervisningsmaterialer. En række studerende meldte sig frivilligt til forsøget. I det følgende er angivet 12 studerende ordnet efter studentereksamensgennemsnit.

Navn	JK	AL	TS	BS	DT	HN	MO	FD	PJ	KM	SR	RA
Snit	6.3	6.8	7.3	7.3	7.9	8.2	8.4	8.5	9.0	10.2	11.1	11.2

1) Hvordan ville du opdele disse studenter på tre hold med 4 på hver hold?
2) Hvordan ville du gøre det, hvis karaktererne gik fra 7.8 til 8.2 ?

Opgave 7. (randomiseret blokforsøg).
Følgende resultater blev opnået fra et eksperiment, hvor man ville undersøge om der var forskel på de resultater, som 5 analyseapparater gav, når man analyserede kvælstofindholdet i jordprøver. På hver af 3 dage blev en portion jord udvalgt og delt i 5 dele, som ved lodtrækning blev givet til analyse i hver sin maskine. Resultaterne var:

		Maskiner				
		P	Q	R	T	U
	Tirsdag	376	379	399	373	376
	Onsdag	372	374	409	387	386
	Torsdag	332	339	365	350	342

Undersøg på dette grundlag om der er forskel mellem analyseapparaterne, og angiv i bekræftende fald hvilke der er forskellige.
Mener du, at det i denne situation var en god ide at foretage forsøget som et blokforsøg?

Opgave 8. (randomiseret blokforsøg).
Fire forskellige produktionsmetoder P, Q, R, og T ønskes sammenlignet med hensyn til det procentiske udbytte ved udvinding af et metal fra et bestemt mineral. Da man ved forsøget er nødt til at benytte forskellige råvarepartierer, og er bange for, at det vil give stor spredning, vælger man at lave et fuldstændigt randomiseret blokforsøg med råvarepartier som blokke. Nedenstående skema angiver resultatet af dette forsøg.

	Metode P	Metode Q	Metode R	Metode T
Råvareparti 1	2.5 2.7	4.7 3.3	2.8 3.3	5.5 5.0
Råvareparti 2	4.6 4.3	7.9 5.9	5.1 6.9	7.2 6.8
Råvareparti 3	4.7 3.9	4.7 4.4	4.4 3.7	6.4 5.7

Undersøg på grundlag af disse oplysninger, om der er forskel på metoderne.

Opgave 9 (planlægning).
Man ønsker at undersøge øl's kemiske stabilitet, når det tilsættes et additiv. Det skal undersøges hvilke af 4 additiver T_1, T_2, T_3 og T_4 der giver størst stabilitet. Der er følgende forhold at tage i betragtning.

a) Hver behandling T_1, T_2, T_3 og T_4 skal udføres mindst 2 gange.
b) Øl er i tanke af forskellig konstruktion og det kan muligvis betyde noget for stabiliteten.
c) Hver ølcharge rummer kun øl til 4 tanke. Ølcharge kan muligvis betyde noget for stabiliteten.
d) Hvert forsøg tager 1 uge. Opbevares øl i mere end 2 til 3 uger forandres det så meget, at det er uegnet til forsøget.

Udarbejd om muligt en forsøgsplan i hvert af følgende 3 tilfælde.
Forsøgsplanen skal indbefatte en beskrivelse af sammenhængen mellem behandlinger og forsøgsenheder (hvorledes vil du tilrettelægge forsøget?) samt en skitse af en variansanalysetabel med angivelse af frihedsgrader.

1. Man tror ikke at tilfældene b og c spiller en væsentlig rolle for heterogeniteten i forsøget.
2. Man tror at b), men ikke c spiller en væsentlig rolle for heterogeniteten i forsøget.
3. Man tror at både b og c spiller en væsentlig rolle for heterogeniteten i forsøget.

Opgave 10. **(planlægning).**
En dækfabrikant ønsker at måle slidstyrken af 4 forskellige dæktyper P, Q, R og T. Dette skal ske ved, at de forskellige typer dæk monteres på hjulene af vogne, der så et stort antal gange gennemkører en udvalgt strækning på en bestemt tid og efter et bestemt hastighedsmønster. Derefter måles sliddet ved at måle den formindskede dæktrådstykkelse.
Dækfabrikanten kan stille 4 biler til rådighed til forsøget, men da der i alt skal køres ca 15000 km tager det så lang tid, at de 4 biler skal køre samtidig, og hver bil kan kun nå at køre 1 tur.
Dækfabrikanten henvender sig til din gruppe og beder jer om at planlægge forsøget.
Angiv på samme måde som i afsnit 1.5 ,

a) navnet på forsøgsplanen,
b) skitser hvorledes forsøget i praksis skal udføres (randomiseringen) og
c) angiv hvorledes forsøgsresultaterne skal analyseres (skitsere variansanalysetabel med angivelse af frihedsgrader) i hvert af følgende tilfælde:

1) det antages, at hverken vogne eller hjulplaceringer spiller en rolle.
2) det antages, at vogne men ikke hjulplaceringer spiller en rolle.
3) det antages, at både vogne og hjulplaceringer spiller en rolle.

Opgave 11. (romersk kvadratforsøg).(eksamensopgave juni 86)
Ved en enzymproduktion havde man valget mellem anvendelse af 4 forskellige råmaterialer R_1 R_2, R_3 og R_4. Man ønskede ved et forsøg at foretage en sammenligning mellem disse med hensyn til det opnåede procesudbytte.
Hvert delforsøg krævede en laborants fuldtidsmedvirken og varede 5 dage. Af ressourcernæssige grunde måtte forsøgsarbejdet fordeles på flere uger, idet man kunne få 4 laboranter stillet til disposition. Man kunne på forhånd ikke udelukke, at forsøgsresultaterne kunne afhænge af forsøgstidspunktet og af, hvilken laborant der medvirkede.
Forsøgsplan og forsøgsresultater var (kodede tal):

		Laboranter							
		1		2		3		4	
Uger	1	R_1	58	R_4	108	R_2	75	R_3	76
	2	R_4	60	R_3	26	R_1	-15	R_2	20
	3	R_3	-16	R_2	23	R_4	24	R_1	-24
	4	R_2	17	R_1	31	R_3	17	R_4	41

Foretag en statistisk analyse til undersøgelse af, om de 4 råmaterialer i middel giver samme udbytte, og foretag i benægtende fald en sammenligning af de 4 råmaterialers virkning. Undersøg tillige, om procesudbyttet i middel må antages at være påvirket af forsøgstidspunkt og/eller laborant.

Opgave 12. **(fuldstændigt 2^k faktorforsøg)** (eksamensopgave 1, juni 1991)
Som led i en serie forsøg med henblik på optimering af styrken af et plastprodukt ønskedes ved et fuldstændigt 2^4- faktorforsøg med n = 1 virkningen af følgende faktorer undersøgt:

		Lavt niveau	Højt niveau
A:	Materialetype	Type 1	Type 2
B:	Fabrikationsmetode	Metode 1	Metode 2
C:	Kemisk farvestof	Ikke tilsat	Tilsat
D:	Opbevaring	Kort tid	Lang tid

Ved forsøget måltes værdien af en parameter, som var mål for produktets styrke.
Forsøgsplan og forsøgsresultater (kodede tal) var følgende:

a 30.9	b 39.8	abd 50.2	bcd 26.2	ac 26.0	bc 30.4	(1) 29.2	ad 23.5
abc 28.4	acd 21.0	bd 40.0	c 23.9	cd 20.4	ab 50.5	abcd 24.7	d 24.3

Det forudsættes at alle faktorer er gruppe 2 faktorer (kun hovedvirkninger og 2-faktorvekselvirkninger kan være forskellige fra nul).

1) Foretag en en statistisk analyse med estimation og testning af de mulige faktorvirkninger og drag konklusioner om disse sidste.
2) Bestem, idet man ønsker størst mulig middelværdi, de(n) optimale kombination(er) af de 4 faktorers niveauer.
3) Opstil et 95%-konfidensinterval for de(n) middelværdi(er), som formodes at være den største middelværdi.

Opgave 13. **(fuldstændigt 2 ^{k} faktorforsøg med blokke)** (eksamensopgave 1, juni 1973)
Et forsøg med faktorerne A og B har en fuldstændig 2^2- faktorstruktur og udføres som et randomiseret blokforsøg med 2 blokke. Forsøgsplan og forsøgsresultatar er:

Blok 1	(1)	2	b	-1	ab	1	a	2
Blok 2	ab	-1	(1)	0	a	2	b	-2

Foretag en analyse af forsøgsresultaterne og angiv, hvilken kombination af faktorniveauerne der ifølge analysen må antages at give
a) den største middelværdi
b) den mindste middelværdi

Opgave 14 (konstruktion af forsøgsplan)
Virkningerne af 6 faktorer T_1, T_2, T_3, T_4, T_5 og T_6 ønskes undersøgt ved et screeningsforsøg. Man kan på forhånd udelukkes, at faktorerne T_4 og T_5 vekselvirker indbyrdes eller med de andre faktorer. Derimod skal det være muligt at teste alle øvrige 2-faktorvekselvirkninger, ligesom alle 6 hovedvirkninger skal kunne testes.
Foretag en hensigtsmæssig omdøbning til A, B osv. og opstil derefter en forsøgsplan med færrest mulige delforsøg. Angiv herunder
a) Den underliggende fuldstændige faktorstruktur
b) Relationerne hvormed nye faktorer er indført i den underliggende struktur
c) Forsøgets behandlinger.
d) Skitser en variansanlysetabel med angivelse af frihedsgrader

Opgave 15. (konstruktion af partiel forsøgsplan ved Statgraphics)
Virkningen af 6 faktorer A, B, C, D, E og F påtænkes undersøgt ved et fuldstændigt randomiseret forsøg med en partiel 2^k faktorstruktur. Om de 6 faktorer vides på forhånd, at kun hovedvirkninger og 2-faktorvirkninger hvori faktor A indgår kan være forskellig fra nul.
Konstruer ved benyttelse af Statgraphics en forsøgsplan med det færrest mulige antal forsøg.

Opgave 16. (analyse af et partielt forsøg ved brug af Statgraphics).
Virkningen af 5 faktorer A, B, C, D og E påtænkes undersøgt ved et fuldstændigt randomiseret forsøg med en partikulær 2^k faktorstruktur. Det antages, at bortset fra hovedvirkninger, kan der kun tænkes at forekomme 2-faktor vekselvirkningerne AC og BC. Da der skal estimeres 7 virkninger, ville man normalt være nødt til at udføre et 2^{5-1} faktorforsøg , men da vi fra tidligere forsøg har kendskab til forsøgsfejlens spredning: $\sigma \approx 2{,}2$, nøjes man med at udføre et 2^{5-2} faktorforsøg med aliasrelationerne D = AB og E = ABC.
Opstillet i standardrækkefølge (efter den underliggende fuldstændige faktorstruktur i faktorerne A, B og C) er forsøgsresultaterne på responsvariablen Y:

d	77.2	ae	84.4	be	80.9	abd	70.3	cde	75.2	ac	75.0	bc	62.1	abcde	86.7

1) Find hvilke faktorer der har virkning.
2) Man ønsker den størst mulige middelværdi af responsvariablen. Angiv de niveauer de i spørgsmål 1 fundne faktorer skal indstilles på, for at give den største middelværdi. Hvis der er flere kombinationer af niveauer der alle (indenfor usikkerheden) må formodes at være lige gode, skal de alle angives sammen med relevante 95% konfidensintervaller.

Opgave 17 (analyse af et partielt forsøg)
Virkningerne af 7 faktorer ønskedes undersøgt ved et partielt faktorforsøg. Om 3 af faktorerne kunne forudsættes, at kun hovedvirkninger kunne være forskellige fra nul, medens for de 4 øvrige også 2-faktorvekselvirkninger eventuelt kunne være forskellige fra nul. De 4 sidste faktorer identificeredes derfor med bogstaverne A,B,C og D og de 3 første med bogstaverne E,F og G.
Der udførtes et fuldstændigt randomiseret forsøg med en $\frac{1}{8} \cdot 2^7$ - faktorstruktur, hvor denne sidste er fremkommet ved, at faktorerne E,F og G indførtes i en fuldstændig 2^4 - faktorstruktur med faktorerne A,B,C og D ved relationerne:
E = ABC F = BCD G = ABCD
1) Bestem forsøgets behandlinger.
2) Idet behandlingerne anføres i standardrækkefølge efter A,B, C og D, og uden gentagelser, var forsøgsresultaterne:

15.3	18.4	26.1	26.3	13.5	15.7	18.8	17.3	21.0	22.3	18.9	15.5	9.6	10.5	23.1	25.0

2a) Find hvilke faktorer der har virkning
2b) Angiv de niveauer de pågældende faktorer skal indstilles på, for at give det største resultat, og angiv et 95% konfidensinterval for dette middeludbytte.

Opgave 18. (simpleste models princip).
Virkningen af 5 faktorer A, B, C, D og E ønskes undersøgt. Det vides om faktor E, at den ikke vekselvirker med nogle af de øvrige faktorer, mens man intet ved om de øvrige.
Da der er 11 mulige faktorvirkninger der skal estimeres, ville man normalt udføre et 2^{5-1} - faktorforsøg, men af ressourcemæssige grunde valgte man indledningsvis at udføre et 2^{5-2} - faktorforsøg, i håb om at det ville være unødvendigt at udføre et supplerende forsøg.
Ved forsøget indlagdes faktorerne D og E ved aliasrelationerne D = -BC og E = ABC.
Opstillet i standardrækkefølge (efter den underliggende fuldstændige faktorstruktur i faktorerne A, B og C) er forsøgsresultaterne på responsvariablen Y:

43,8	31,1	44,5	33,4	45,2	48,0	45,7	46,0

Ud fra tidligere forsøg vides at forsøgsfejlens spredning er ca 1.2.
1) Udfør (ved at benytte simpleste models princip) testninger til vurdering af, hvilke faktorer der må formodes at have indflydelse på responsvariablen Y.
2) Man ønsker den størst mulige middelværdi af responsvariablen. Angiv med baggrund i den i spørgsmål 1 foretagne analyse de niveauer faktorerne skal indstilles på, for at give det største middelværdi. Hvis der er flere kombinationer af niveauer der alle (indenfor usikkerheden) må formodes at være lige gode, skal de alle angives sammen med relevante 95% konfidensintervaller.

Opgave 19 (planlægning af et konfunderet partielt forsøg) (E87)
Virkningerne af 6 faktorer T_1, T_2, T_3, T_4, T_5 og T_6 ønskes undersøgt ved et screeningsforsøg. Man har kun råd til at lave 16 delforsøg, og det er af tidsmæssige grunde nødvendigt at arbejde med 4 blokke på 4 forsøgsenheder. Det vides at kun hovedvirkninger og 2 - faktorvekselvirkningerne T_4T_1, T_4T_2, T_4T_3 og T_4T_5 kan være forskellig fra nul.
Foretag en hensigtsmæssig omdøbning til A, B osv. og opstil derefter en forsøgsplan med færrest mulige delforsøg. Angiv herunder
a) Den underliggende fuldstændige faktorstruktur
b) Relationerne hvormed nye faktorer er indført i den underliggende struktur
c) Forsøgets behandlinger.
d) Skitser en variansanlysetabel med angivelse af frihedsgrader

Opgave 20. (analyse af et konfunderet partielt forsøg).
Virkningen af 6 faktorer A, B, C, D, E og F ønskes undersøgt. Man ved, at ingen af faktorerne vekselvirker med andre faktorer. Da man kun kan udføre 4 forsøg pr. apparat, indføres blokke på 4 forsøgsenheder.
I en fuldstændig 2^4 struktur med faktorerne A, B, C og D indføres E = - ABCD og F = ACD.
Endvidere indføres blokkene ved at konfundere ABC og CD.
Opstillet i standardrækkefølge (efter den underliggende fuldstændige faktorstruktur i faktorerne A, B, C og D) er forsøgsresultaterne på responsvariablen Y:

27.9	43.7	24.3	46.1	37.2	13.3	20.7	29.6	27.5	25.0	30.3	20.7	17.4	52.8	35.2	36.6

1) Udfør testninger til vurdering af, hvilke faktorer der må formodes at have indflydelse på responsvariablen Y
2) Angiv de niveauer de pågældende faktorer skal indstilles på, for at give den største værdi af responsvariablen Y.
3) Gennemsnittet af de 4 forsøgsresultater, som indgår i den blok, som har den højeste værdi af responsvariablen er 40,5(ses umiddelbart ved udregning).
   Beregn et estimat og et 95%- konfidensinterval for middelværdien af Y, når behandlingen fundet i spørgsmål 2 anvendes i denne blok.
   Beregningerne skal både foretages ved hjælp af de relevante formler (med lommeregner) og ved hjælp af Statgraphics. Vink: Det kan oplyses, at blokgennemsnittet har en varians på $\frac{\sigma^2}{4}$.

Opgave 21 (trinvis variansanalyse) (hentet fra SF I øvelse 19).
I hvert af 5 tilfældigt valgte autoriserede reparationsværksteder udvalgtes tilfældigt 2 mekanikere, og for hver af disse måltes 3 gange den tid, som en bestemt reparation tog. Følgende resultater fandtes (kodede tal):

Værksteder							
1		2		3		4	
Mekaniker		Mekaniker		Mekaniker		Mekaniker	
1	2	3	4	5	6	7	8
6 2 0	1 10 0	10 7 12	4 1 9	10 11 6	7 12 15	5 10 8	0 8 6

Udfør en trinvis "random factor" variansanalyse og drag konklusioner.
Estimér alle varianskomponenter.

Opgave 22 (fixed og random, sidedelt[crossed] og trindelt[nested]). (udarbejdet af B. Hellesen) En stor virksomhed har 3 **afdeling**er A1, A2, A3. Fra hver afdeling vælges tilfældigt 7 medarbejdere, dvs. i alt 21 **person**er B1, B2, ... , B21 (21 "hovedplots"). Man ønsker at undersøge, om et **kursus** har en virkning på medarbejdernes ydelse Y. Man vil også gerne undersøge, om afdelingerne har en virkning på ydelsen . Hver persons ydelse måles før og efter, at vedkommende har været på kurset (2 "split-plots" i hvert hovedplot"). De målte ydelser var:

Dag	Person	Før kursus C1	Efter kursus C2
Afdeling A1	B1	26.25	29.50
	B2	24.33	27.62
	B3	22.52	25.71
	B4	29.33	31.55
	B5	28.90	31.25
	B6	25.13	29.07
	B7	29.33	31.15
Afdeling A2	B8	27.47	28.74
	B9	25.19	26.11
	B10	23.53	25.45
	B11	24.57	25.58
	B12	26.88	27.70
	B13	27.86	28.82
	B14	28.09	28.99
Afdeling A3	B15	22.27	22.52
	B16	21.55	21.79
	B17	23.31	23.53
	B18	30.03	30.21
	B19	27.17	28.65
	B20	28.09	28.33
	B21	27.55	27.86

Den fuldstændige model for forsøget er Y = μ + A + B(A) + C + A*C + B(A)*C + ε

dvs. $$y_{ijkm} = \mu + A_i + B_{j(i)} + C_k + AC_{ik} + BC_{jk(i)} + e_{m(ijk)} \ .$$

a) Udarbejd en tabel for varianskomponentsammensætningen.

b) Undersøg, om der er vekselvirkning A*C og hovedvirkninger C og A.
(Vink: Da Statgraphics her ikke accepterer en "Residual" med 0 frihedsgrader, kan du udelade modelkomponenten B(A)*C, så den kommer til at optræde som "Residual" -- og herved fås nyttige F-ratios, jævnfør resultatet i spørgsmål a).

c) Har det en positiv effekt at sende medarbejderne på kurset?

d) På hvilken afdeling får man mest ud af kurset?

Facitliste:a) (14,2,0,0,0,1), (0,2,0,0,0,1), (0,0,21,0,1,1), (0,0,0,7,1,1), (0,0,0,0,1,1), (0,0,0,0,0,1).
b) SAK_total = 306.787

Opgave 23(fixed og random, sidedelt[crossed] og trindelt[nested]) (udarbejdet af B. Hellesen)
For at opnå den højst mulige ydelse Y af styrkesportsdamer ønsker man at finde den bedste slags **trøje** (C1 eller C2 ?).
Der udvælges tilfældigt 2 styrke**sport**sgrene (A1 og A2), og fra hver sportsgren vælges tilfældigt 2 damer, dvs. i alt 4 **damer** (D1, D2, D3 og D4). Der udvælges tilfældigt 2 **uger** (B1 og B2) i sportssæsonen, og fra hver uge vælges tilfældigt 2 dage, dvs. i alt 4 **dage** (E1, E2, E3 og E4). Man laver derpå et forsøg med sportsgrene (A) og uger (B) som blokke. De målte data for ydelsen var:

Uge			B1		B2	
Dag			E1	E2	E3	E4
Sport A1	Dame D1	C1	30	21	19	20
		C2	29	31	17	19
	Dame D2	C1	25	23	17	17
		C2	18	20	7	11
Sport A2	Dame D3	C1	27	27	18	16
		C2	19	23	12	16
	Dame D4	C1	27	19	16	18
		C2	16	19	11	9

Modellen for forsøget er Y = μ + A + B + C + D(A) + E(B) + C*D(A) + C*E(B) ,

dvs. $y_{ijkmnp} = \mu + A_i + B_j + C_k + D_{m(i)} + E_{n(j)} + CD_{km(i)} + CE_{kn(j)} + e_{p(ijkmn)}$.

a) Udarbejd en tabel for varianskomponentsammensætningen.
b) Synes modellen god? Undersøg, om der er vekselvirkninger CD og CE.
c) Undersøg, om der er hovedvirkninger A, B, C, D og E.
d) Estimér varianskomponenterne σ_{CE}^2 og σ_A^2 .

(Hovedplots: Kombinationerne af damer og dage kan opfattes som 16 hovedplots, men virkningen heraf kan i denne opgave ikke skelnes fra vekselvirkningen D(A)*E(B), og er i øvrigt negligeret i ovennævnte model.)

Opgave 24. (Split-plot) (SFII eksempel 71)
Indflydelsen af 4 forskellige lakeringsmetoder (L) og 3 forskellige ovntemperaturer (*T*) på overfladekvaliteten af lakerede metalemner skal undersøges. Hver behandling gentages 2 gange, så man i alt kommer op på 24 delforsøg.
Forsøget kunne udføres som et fuldstændigt randomiseret forsøg, hvor 6 tilfældigt udtagne metalemner lakeres med lak L_1, 6 andre med lak L_2 osv. Vi vælger så tilfældig af de tre temperaturer T_1, T_2 eller T_3, opvarmer ovnen hertil, placerer et tilfældigt af de 24 lakerede metalemner i ovnen og venter den nødvendige forsøgstid. Derefter gentages forsøget med en anden ovntemperatur og et andet metalemne.
Dette betyder imidlertid, at man mange gange skal opvarme og afkøle ovnen, hvilket tager tid og er dyrt. Da ovnen kan rumme mindst 4 metalemner pr. forsøg ønskes anvendt et split-plot forsøg hvor det er nok med 6 temperaturskift (antal gange ovnen skal opvarmes/afkøles).
1) Beskriv et split-plot forsøg, hvor der også på en skitse vises den resulterende plan.
2) Angiv forsøgsenheder og faktorer.
3) Angiv modellen for observationerne.

Opgave 25 (Split-plot)
Forsøgsresultaterne af det i opgave 23 fremgår af nedenstående skema:

	Temperatur	T1			T2			T3	
	Ovnkørsler	O1	O2		O3	O4		O5	O6
Lakerings-metoder	L1	14	21		10	12		21	17
	L2	18	27		16	17		26	21
	L3	12	17		8	7		15	16
	L4	17	22		11	13		23	20

Foretag en statistisk analyse af forsøget.

Opgave 26. (Split-plot) (SFII eksempel 75)
Forsøgsresultaterne af det i eksempel 4.4 side 101 angivne forsøg fremgår af nedenstående skema:

Hovedplotfaktor: Fodertyper		F1		F2		F3		F4	
Underplotfaktor: Dosis		D1	D2	D1	D2	D1	D2	D1	D2
Blokfaktor:Kuld	K 1	23	17	21	19	25	17	26	17
	K 2	34	19	33	22	33	23	35	14
	K 3	29	21	30	16	29	18	28	22
	K 4	40	23	36	23	38	19	42	25
	K 5	25	15	20	14	19	12	28	16

Foretag en statistisk analyse af forsøget.

Opgave 27 (Split-plot). (omskrivning af SFII øvelse 90)
Korrosionsbestandigheden X af to forskellige trælakker ønskes sammenlignet ved benyttelse af to forskellige træsorter og under påvirkning af to forskellige klimaer.
Fire tilfældigt udvalgte brædder af hver træsort males på en tilfældig valgt side med den ene slags lak og på den anden side med den anden slags lak. Derpå ophænges to tilfældigt valgte brædder af hver træsort i det ene klima og de resterende brædder i det andet klima. Efter ét år nedtages alle brædderne, og der foretages en måling af korrosionsbestandigheden under fuldstændigt randomiserede omstændigheder.
Forsøgsresultaterne blev følgende (mindst tal giver størst korrosionsbestandighed):

		Klima 1		Klima 2	
		Brædt 1	Brædt 3	Brædt 5	Brædt 7
Træsort 1	bræddeside 1	lak 1: 8,7	lak 1: 8,4	lak 1: 19,4	lak 1: 15,8
	bræddeside 2	lak 2: 12.4	lak 2: 12,6	lak 2: 21,6	lak 2: 18,5
		Brædt 2	Brædt 4	Brædt 6	Brædt 8
Træsort 2	bræddeside 1	lak 1: 6,1	lak 1: 9,8	lak 1: 22,8	lak 1: 17,3
	bræddeside 2	lak 2: 8,6	lak 2: 12,1	lak 2: 25,8	lak 2: 19,0

Der benyttes forkortelserne K for klima, T for træsort, B for brædt og L for lak.
Modellen for forsøget er Y = μ + K + T + KT + B(KT) + L + KL +TL + KTL + B(KT)L + ε
eller $x_{ijkm} = \mu + K_i + T_j + KT_{ij} + B_{(ij)k} + L_m + KL_{im} + TL_{jm} + KTL_{ijm} + BL_{(ij)km} + \varepsilon_{(ijkm)n}$

Spørgsmål 1: Opstil en varianskomponenttabel for forsøget.
Vink: Der er to hovedfaktorer i dette forsøg og brædt er undertrin til både klima og træsort

Spørgsmål 2. Foretag på basis heraf testningerne til vurdering af, om/hvorledes korrosions-bestandigheden afhænger af træsorter, klimaer og trælakker.

Opgave 28 (enkelt regressionsanalyse, transformation).
Udbyttet af en bestemt kemisk proces er en statistisk variabel, hvis middelværdi afhænger af temperaturen *T*. Følgende observationer foretoges:

Temperaturer (^{0} C)	50^0	60^0	70^0	80^0	100^0	125^0	150^0	200^0
Udbytter (gram)	0 2 0	3 1 2	2 3 5	12 10 8	19 22 19	44 43 41	76 75 80	160 156 157

1) Det ønskes ved en regressionsanalysetest vist, at en model, hvorefter *Y*'s middelværdi er en lineær funktion af *T*, ikke indenfor temperaturområdet [50 ; 200] giver en tilstrækkelig god beskrivelse af *Y*'s variation.
2) Benyt menupunktet"Comparison of Alternative Models" til at finde en model der bedre beskriver talmaterialet.
   Angiv i denne model sammenhængen mellem udbyttet og temperaturen, og angiv et estimat og et 95% konfidensinterval for middeludbyttet svarende til $T = 110^0$.

Opgave 29 (polynomial regressionsanalyse) (examen januar 1988)
Ved et fuldstændigt randomiseret forsøg foretoges følgende observationer mellem den ikke-statistiske variabel x og den statistiske variabel Y:

x							
10	20	30	40	50	60	70	80
72,5 70,4	78,3 74,5	79,6 78,6	78,9 82,4	76,9 77,3	76,6 78,2	68,9 70,0	66,4 63,5

1. Bestem ved en polynomial regressionsanalyse det polynomium i x af lavest mulig grad, der giver en tilfredsstillende beskrivelse af Y's variation.
2. Opstil et 95% konfidensinterval for middelværdien af Y, når $x = 45$.
3. Find den værdi x_{m} som giver den største y - værdi . Angiv endvidere den til x_{m} svarende estimerede middelværdi $\widetilde{Y}_m$ og et 95% konfidensinterval for $\widetilde{Y}_m$.

Opgave 30 (regressionsanalyse uden gentagelser)
Følgende sammenhørende data er 25 målinger mellem den jævnstrøm (y) en vindmølle udvikler og vindhastigheden (x).

x	5.00	6.00	3.40	2.70	10.0	9.70	9.55	3.05	8.15	6.20	2.90	6.35	4.60
y	1.582	1.822	1.057	0.500	2.236	2.386	2.294	0.588	2.166	1.866	0.653	1.930	1.562
x	5.80	7.40	3.60	7.85	8.80	7.00	5.45	9.10	10.20	4.10	3.95	2.45	
y	1.737	2.088	1.137	2.179	2.112	1.800	1.501	2.303	2.310	1.194	1.144	0.123	

1) Vurder grafisk om en ret linie kan siges at være en god model for forsøget (benyt både en figur med indtegnet regressionslinie og en figur med residualerne indtegnet til vurderingen).
2) Benyt menupunktet"Comparison of Alternative Models" til at få en ide om hvilken model, der bedre kan beskrive data, og foretag på samme måde som i spørgsmål 1) en vurdering af om modellen indenfor forsøgsområdet.
3) Angiv en ligning for den fundne kurve, og angiv et estimat og et 95% konfidensinterval for y svarende til en vindhastighed på $x = 7$.
4) Foretag på samme måde som i spørgsmål 1) en vurdering af om et polynomium af anden grad er en god model indenfor forsøgsområdet.
5) Idet det antages, at andengradsmodellen fra spørgsmål 4 er en rimelig god model, skal en ligning for andengradsmodellen angives, og på det grundlag et estimat og et 95% konfidensinterval for y svarende til en vindhastighed på $x = 7$.

Opgave 31. (multipel regressionsanalyse med gentagelser).

En fabrik fremstiller salpetersyre ved oxidering af ammoniak med luft. I løbet af processen ledes kvælstofoxider under afkøling ind i en absorptionskolonne, idet absorptionen i gennemstrømmende salpetersyre afhænger af kølevandstemperaturen x_1 (°C) , lufttemperaturen x_2 (°C) og salpetersyrekoncentrationen x_3

Man ønsker at teste, om sammenhængen mellem mængden Y af ikke-absorberede kvælstofoxider i et givet tidsrum og x_1, x_2 og x_3 (approksimativt) var lineær, og ønskede i bekræftende fald at estimere denne sammenhæng.

Følgende observationer af Y (kodede tal) fandtes:

x_3		x_1		x_2: 10	20	30	40
x_3	-5	x_1	5	32 30	50 47	62 66	88 82
			10	48 51	60 63	73 77	96 99
			15	54 56	70 75	90 83	102 110
	5	x_1	5	32 35	49 52	64 60	86 84
			10	50 55	64 60	82 77	92 94
			15	57 53	74 77	88 86	108 105

1) Vis ved en testning at sammenhængen mellem y og x_1, x_2 og x_3 kan være lineær dvs. af formen $Y = \alpha_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$.
2) Undersøg, om modellen kan reduceres, dvs om nogle af regressionskoefficienterne kunne være nul.
3) Giv et estimat for regressionskoefficienterne i den endelige model, og opskriv ligningen.
4) Angiv et 95% konfidensinterval for β_1 .
5) Angiv et estimat for Y i tilfældet $x_1 = 8$, $x_2 = 20$ og $x_3 = 4$, og angiv et 95% konfidensinterval for denne værdi.

Opgave 32. (partiel regressionsanalyse med gentagelser).
Ved en given produktion ønskes undersøgt, hvorledes mængden Y af et uønsket biprodukt afhang af mængderne x_1, x_2 og x_3 af tre tilsætningsstoffer. Følgende forsøg blev foretaget (kodede tal):

(x_1, x_2, x_3)	(1,1,1)	(2,9,4)	(3,3,9)	(4,7,5)	(5,5 7)	(6,3,3)	(7,6,2)	(8,9,6)
y	30 34	85 87	55 57	75 80	76 68	56 52	85 80	106 109

Det forudsættes, at regressionsforudsætningerne er opfyldt.
1) Vis ved en sædvanlig 5% test, at en lineær model i de tre variable kan beskrive Y's variation.
2) Reducer om muligt modellen, og bestem regressionsligningen.
3) Bestem et estimat for Y i tilfældet $x_1 = 4$, $x_2 = 5$ og $x_3 = 6$, og angiv et 95% konfidensinterval for denne værdi.

Opgave 33 (partiel regressionsanalyse uden gentagelser).
Den tid (y) det tager inden en bestemt maskinkomponent svigter kan tænkes at afhænge af den spænding (x_1) , den temperatur (x_2) som komponenten udsættes for under kørslen, samt motorens omdrejningshastighed pr. minut (x_3)
Det forløbne år har givet de data, som er vist i følgende tabel:

(x_1, x_2, x_3)	(110,60,750)	(110,82,850)	(110,60,1000)	(110,82,1100)	(120,60,750)
y	2145	2155	2220	2225	2360

(x_1, x_2, x_3)	(120,82,850)	(120,60,1000)	(130,82,1100)	(115,66,840)	(115,66,880)
y	2266	2334	2340	2212	2180

Det forudsættes, at regressionsforudsætningerne er opfyldt.
1) Vurder ud fra forklaringsgraden og grafisk, om en lineær model i de tre variable ,dvs. af formen
   $Y = \alpha_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ er rimelig.

Det antages i det følgende, at ovenstående model gælder.
2) Undersøg om modellen kan reduceres, dvs. om nogle af koefficienterne kan antages at være 0.
3) Angiv regressionsligningen i den endelige model.
4) Bestem et estimat for Y i tilfældet $x_1 = 125$, $x_2 = 70$ og $x_3 = 900$, og angiv et 95% konfidensinterval for denne værdi.

Opgave 34 (partiel regressionsanalyse uden gentagelser).
Det formodes, at den producerede mængde Y af et stof ved en given produktion er en lineær funktion af de anvendte mængder x_1, x_2, og x_3 af tre råvarer.
Følgende ikke særligt systematiske observationer foreligger:

x_2		0.1			0.2			0.4			0.5		
x_3		0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.5
x_1	0.1	6.37				6.70							
	0.2			8.02							7.54		
	0.3		8.70				9.08						
	0.5				10.30				10.40				
	1.0							15.34					16.12

1) Vurder på basis af disse observationer, om en lineær model i x_1 , x_2 og x_3 er rimelig.
2) Foretag så vidt mulig en reduktion af modellen, og angiv til sidst regressionsligningen for den endelige model.
3) Beregn et 95% konfidensinterval for regressionskoefficienterne i den endelige model.
4) Beregn et 95% konfidensinterval for middelværdien af Y hvis $x_1 = 0.3$, $x_2 = 0.4$ og $x_3 = 0.1$.

Opgave 35 (polynomial regressionsanalyse) (eksamen juni 87 omskrevet).
Det ønskes undersøgt, hvorledes sammenhængen mellem middeludbyttet af en kemisk produkt afhænger af reaktionstemperaturen x_1 og syrekoncentrationen x_2 .
Forsøgsresultaterne (udbyttet i %) fremgår af følgende skema:

		x_2		
		0.2 n	0.6 n	1.0 n
x_1	20	240,6 210,4	103,4 85,2	44,0 64,8
	60	275,2 321,4	195,5 176,3	132,0 140,2
	100	385,7 377,1	244,2 251,2	204,7 231,3

Det formodes på forhånd, at middelværdien af Y kan beskrives som et polynomium af højst 2 grad i x_1 og x_2: $Y = \alpha_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2$.
1) Vurdér om ovennævnte 2.gradsmodel kan antages at kunne beskrive sammenhængen mellem Y og de 2 uafhængige variable.

I det følgende antages ovennævnte model at gælde.
2) Test om en lineær model $Y = \alpha_0 + \beta_1 x_1 + \beta_2 x_2$ i de variable kan antages at kunne beskrive sammenhængen mellem Y og de 2uafhængige variable.
3) Reducer om muligt modellen, og angiv regressionsligningen i den endelige model.
4) Bestem et estimat og et 95% konfidensinterval for middeludbyttet ved 80^0 og en syrekoncentration på 0.3 n.

Opgave 36 (2^k faktorforsøg med 3 kvantitative faktorer og en blokfaktor) (hentet fra SFIII eksempel 2 side 19).
Ved et forsøg ønskedes undersøgt, om og hvorledes brudstyrken af metalstænger af en bestemt legering ændredes ved tilsætning af små mængder af metallerne mangan, chrom og nikkel til legeringen. Forsøget udførtes som et fuldstændigt 2^3 faktorforsøg med laboratorier som blokke, idet 2 laboratorier medvirkede ved forsøget. Resultaterne var:

	Mangan x_1 i %	Chrom x_2 i %	Nikkel x_3 i %	Brudstyrke Lab. 1	Brudstyrke Lab. 2
(1)	0.2	0.1	0	7.8	8.2
a	0.4	0.1	0	11.9	12.3
b	0.2	0.2	0	9.0	10.6
ab	0.4	0.2	0	12.2	9.0
c	0.2	0.1	1.5	-2.4	-3.7
ac	0.4	0.1	1.5	1.6	0.2
bc	0.2	0.2	1.5	-1.8	-5.0
abc	0.4	0.2	1.5	-3.0	-7.6

Det formodes på forhånd, at middeludbyttet for hvert laboratorium kan beskrives ved en 1. gradsmodel: $Y = r_i + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$, hvor $i = 1$ for lab 1 og $i = 2$ for lab 2 .
Det forudsættes, at blokkriterium og behandlinger ikke vekselvirker.

1) Test, om ovennævnte 1. gradsmodel kan antages at kunne beskrive sammenhængen mellem Y og de 3 uafhængige variable.
2) Reducer om muligt modellen, og angiv regressionsligningerne i den endelige model.
3) Angiv et estimat for middelbrudstyrken Y ved en tilsætning af 0.3% mangan, 0.1% chrom og 1% nikkel. Angiv endvidere et 95% konfidensinterval for dette estimat. Hvis der er forskel på laboratorierne, skal man vælge laboratoriet der giver den største brudstyrke.

Opgave 37 (2^k faktorforsøg med én kvantitativ og to kvalitative faktorer) (eksamenspgave 1974).
På en farve- og lakfabriks laboratorium ønskedes et farvestofs renhed undersøgt i afhængighed af tre filtreringsbetingelser, som hver varieredes på 2 niveauer. De 3 filtreringsbetingelser var:
1) væskekoncentration x (faktor A) 20% eller 40%,
2) opbevaring i varmt lokale før filtrering eller ingen opbevaring før filtrering(faktor B),
3) tilstedeværelse eller ikke tilstedeværelse af butanol (faktor C) .

Det vides, at for ethvert af de 4 niveaukombinationer af B og C kan renheden beskrives ved et 1. grads polynomium $Y = \alpha_{ij} + \beta_{ij} x$, hvor i = 0,1 (opbevaring ja/nej) og j = 0,1 (butanol nej/ja).
Der udførtes følgende fuldstændigt 2^3 faktorforsøg:

	Faktorer	Renhed

	Koncentration x i %	Opbevaring B	Tilstedeværelse af butanol C	
(1)	20	ja	nej	24 22
a	40	ja	nej	46 41
b	20	nej	nej	3 4
ab	40	nej	nej	26 24
c	20	ja	ja	21 21
ac	40	ja	ja	47 50
bc	20	nej	ja	-2 3
abc	40	nej	ja	29 26

1) Der er en formodning om, at de fire regressionslinier er parallelle. Dette ønskes testet og hvis det er tilfældet ønskes undersøgt om regressionslinierne eventuelt er vandret.
2) Test dernæst, om konstantleddet afhænger af opbevaringsforhold og tilsætning af butanol
3) Kontroller grafisk den fundne model og analysens forudsætninger.
4) Opskriv på basis af ovennævnte resultater de estimerede regressionsligninger.
5) Angiv et estimat for middelværdien af y ved filtrering under tilstedeværelsen af butanol efter opbevaring i et varmt lokale ved en væskekoncentration på 35%. Angiv endvidere et 95% konfidensinterval for den fundne middelværdi.

Opgave 38 (polynomial regression med blokfaktor) (SFIII eksempel 1 side 12)
Udbyttet ved en kemisk proces ønskedes undersøgt ved temperaturerne 100° C, 150°C, 200°C og 250°C. Det besluttedes at udføre et forsøg med 3 gentagelser af hver behandling. Idet forsøget baseredes på 3 forskellige råvarecharger, udførtes det som et randomiseret blokforsøg med råvarecharger som blokke.
Det formodedes på forhånd, at middeludbyttet for hver råvarecharge kan beskrives ved et polynomium af 2. grad.
Den model, der foreslås for den i'te råvarecharge er følgelig $Y = r_i + \beta_1 T + \beta_2 T^2$, $i = 1, 2, 3.$
Det forudsættes endvidere, at blokkriterium og temperatur ikke vekselvirker.
Forsøgsplan og forsøgsresultater var:

Råvarecharge 1	250° 41.0	100° 39.4	200° 41.5	150° 40.7
Råvarecharge 2	150° 47.0	200° 47.0	100° 43.4	250° 46.6
Råvarecharge 3	200° 44.7	250° 42.9	100° 41.0	150° 44.5

1) Test om ovennævnte 2. gradsmodel kan antages at kunne beskrive sammenhængen mellem Y og T.
2) Reducer om muligt modellen, og angiv regressionsligningerne i den endelige model.
3) Bestem den temperatur i intervallet fra 100° til 250°C, hvor udbyttet er størst.

Opgave 39 (2×4 faktorforsøg med 1 kvantitativ og 1 kvalitativ faktor) (eksamen juni 1975). En kemisk proces til fremstilling af et stof kan udføres på 2 forskellige måder; i begge tilfælde afhænger procesudbyttet af processens temperatur *T*. I et fuldstændigt randomiseret forsøg til undersøgelse af hvilket temperaturniveau der er optimalt for hvert produktionsområde, fandtes følgende resultater (udbytte i %):

		Temperatur T (^{0} C)			
		80	100	120	140
Produktions-måde	1	72.3 71.9	74.0 74.8	71.4 72.1	65.3 66.2
	2	69.6 70.4	72.8 73.4	74.2 73.7	72.9 72.6

1) Test, om det for hver produktionsmåde kan antages, at sammenhængen mellem udbyttet Y og temperaturen T kan beskrives ved et polynomium af højst 2. grad:
   $Y = \alpha_i + \beta_{1i}T + \beta_{2i}T^2$, hvor $i = 1$ for produktionsmåde 1 og $i = 2$ for produktionsmåde 2 .
2) Benyt endvidere de bedst mulige estimater for disse sammenhænge til bestemmelse af den for hver produktionsmetode optimale temperatur. Angiv endvidere for hver produktionsmetode et estimat for det maksimale procesudbytte med angivelse af et 95% konfidensinterval.

Opgave 40 (3×3×2 faktorforsøg med 2 kvantitative og 1 kvalitativ faktor) (eksamen juni 1977 omskrevet).
I et fuldstændigt randomiseret forsøg undersøgtes virkningen af 3 faktorer på udbyttet af en kemisk proces. Faktorerne er A: væskekoncentration x , B: procestemperatur *T*, mens C er produktionsmetoden (2 metoder). Det vides, at faktor C (metode) ikke vekselvirker med faktorerne A og B. Følgende resultater fandtes (10 · (udbytte% - 80)):

		C					
		Metode 1			Metode 2		
		T			T		
		50	70	90	50	70	90
x	20	1	88	61	28	131	107
	30	69	183	188	110	230	236
	40	42	171	183	67	203	220

1) Test om et andengradspolynomium i to variable for hvert niveau af C kan beskrive faktorerne A og B's indvirkning på udbyttet: $Y = \alpha_i + \beta_1 x + \beta_2 T + \beta_3 x \cdot T + \beta_4 x^2 + \beta_5 T^2$.
   hvor $i = 1$ for metode 1 og $i = 2$ for metode 2 .
2. Undersøg, om modellen kan reduceres.
3. Opskriv ligningerne for den endelige model.
4. Angiv et estimat for Y i tilfældet $x = 25$, $T = 75$ og C er på niveau 2.

Opgave 41 (kovariansanalyse) (SF I øvelse 68)
Ved et foderforsøg med svin af en bestemt race undersøgtes virkningen på vægtforøgelsen i løbet af en 2 ugers-periode af to faktorer
(faktor A: Uden eller med fodertilskud A; faktor B:Uden eller med fodertilskud B).
Ved forsøget benyttedes i alt 40 svin (opdelt i 4 lige store grupper), idet man for hvert svin målte dels vægtforøgelsen *y*, dels svinets samlede foderforbrug *x*. Forsøgsresultaterne var (kodede tal):

		B			
		uden tilskud B		med tilskud B	
		y	*x*	*y*	*x*
A	Uden tilskud A	291	536	494	650
		381	631	450	597
		407	661	417	561
		253	508	389	548
		238	508	244	404
		444	704	356	496
		356	605	464	613
		341	600	355	505
		414	664	454	598
		192	456	295	459
	Med tilskud A	171	395	243	363
		342	550	411	516
		373	575	475	577
		451	641	608	704
		442	636	562	658
		441	644	473	569
		325	533	409	516
		315	523	411	519
		319	523	448	556
		420	631	390	487

Foretag en tosidet kovariansanalyse og undersøg i forbindelse hermed forudsætningen om parallelle regressionslinier. Foretag endvidere til sammenligning en analyse af forsøgsresultaterne uden inddragelse af den kovariable *x*. Hvilken konklusion kan drages af sammenligningen?

STIKORDSREGISTER