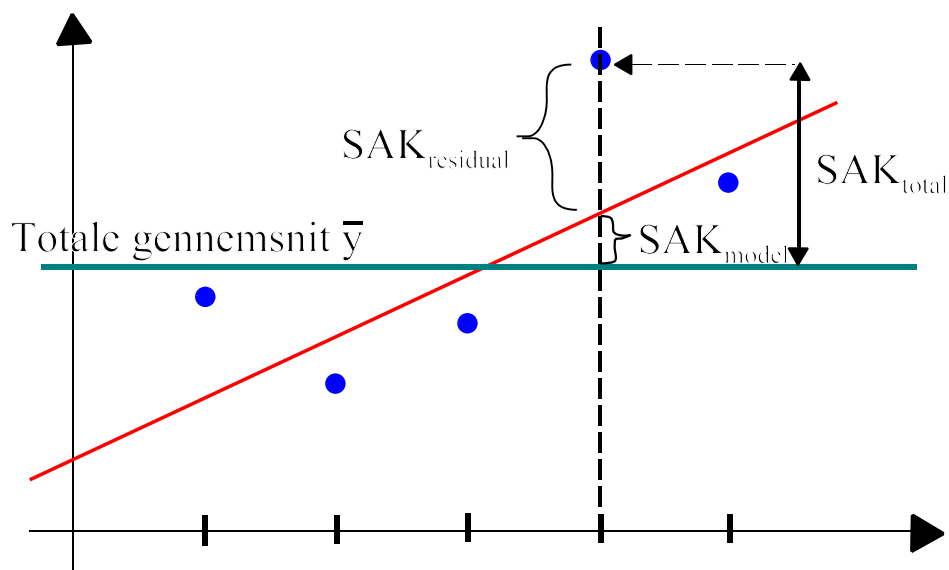


MOGENS ODDERSHEDE LARSEN

VIDEREGÅENDE STATISTIK II

Regressionsanalyse
(TI-89 og Statgraphics)



DANMARKS TEKNISKE UNIVERSITET
6. udgave 2005

FORORD

Dette notat kan læses på baggrund af en statistisk viden svarende til lærebogen “M. Oddershede Larsen : Statistiske grundbegreber”.

Notatet er bygget op således, at de væsentligste begreber søges forklaret anskueligt og ved hjælp af et stort antal eksempler.

Det forudsættes, at man har en lommeregner med de statistiske fordelinger indlagt. Der vil derfor i ringe omfang blive benyttet statistiske tabeller.

I “Statistiske grundbegreber” er der i appendix A en “brugsanvisning” på hvorledes dette kan gøres med bl.a lommeregnerne TI-83 og HP48G.

Sidst i dette notat (side 85) findes en tilsvarende “brugsanvisning” for TI-89.

Sidst i hvert kapitel findes en oversigt over de vigtigste formler samt nogle opgaver. En facitliste til opgaverne findes bagerst i notatet.

Fordelen ved direkte at bruge formlerne til løsning af eksemplerne er, at man derved opnår en større forståelse. De mere avancerede programmer bliver let en “sort” kasse, hvor der på mystisk vis dukker et facit op, som man ikke rigtig har noget forhold til.

Imidlertid er det naturligvis også vigtigt, at man kender de muligheder mere avancerede regnemidler tilbyder.

Lommeregneren TI-89 er en god “statistiklommeregner, så i slutningen af hvert kapitel bliver de samme eksempler som i hovedteksten regnet ved benyttelse af dens indbyggede programmer.

Ved behandling af store datamængder og ved mere regneteknisk komplicerede analyser er en lommeregner som TI - 89 ikke nok. Her er det nødvendigt at benytte en PC med et passende statistisk software. I dette notat er anvendt statistikprogrammet “Statgraphics”, således, at alle eksemplerne efter hvert kapitel også er regnet med dette program.

Der findes mange andre udmærkede statistikprogrammer. I et ganske tilsvarende notat (som kan findes på nedenstående adresse) er Statgraphics eksemplerne således udskiftet med SAS-JUMP. Udskrifterne fra sådanne statistikprogrammer afviger ikke væsentligt fra hinanden, så skulle man i undervisningen benytte et tredje statistikprogram, kan de studerende uden vanskelighed på basis af disse udskrifter tolke egne udskrifter.

Data foreligger ofte som en fil i et regneark som eksempelvis Excel. Disse regneark har indbygget en del statistik bl.a. de almindeligste testfunktioner.

I notatet “Videregående statistik regnet med Excel” er en række af disse statistiske muligheder gennemgået.

Andre notater i samme “serie” er noterne

Videregående Statistik I: Sammenligning af to eller flere kvalitative variable

Videregående Statistik III: Ikke parametriske metoder. Statistisk kvalitetskontrol

Noterne (som både findes i en Statgraphics og en SAS-JUMP version) et søgt udarbejdet, så de kan læses uafhængigt af hinanden.

Alle de nævnte noter kan i pdf-format findes på adressen www.larsen-net.dk

August 2006 Mogens Oddershede Larsen.

INDHOLD

13 REGRESSIONSANALYSE

13.1	Indledning	1
13.2	Enkelt Regressionsanalyse	2
13.2.1	Forklaring af metode og formler	2
13.2.2	Beregning af enkelt regressionsanalyse med én y - observation for hver x - værdi	11
13.2.3	Enkelt regressionsanalyse med flere y - observationer for hver x - værdi	14
13.2.4	Transformation af data inden regressionsanalyse foretages	17
13.3	Polynomial regressionsanalyse	19
13.3.1	Indledning	19
13.3.2	Forklaring af metoder ved beregning	19
13.4	Multipel regressionsanalyse	20
13.4.1	Indledning	20
13.4.2	Analyse med én y - observation for hver x - værdi	20
13.4.3	Analyse med flere y -observationer for hver x - værdi	26

Oversigter

13.1.	Formler til beregning af enkelt regressionsanalyse uden gentagelser	26
13.2	Formler til beregning af enkelt regressionsanalyse med lige mange gentagelser	28
13.3	Transformation til lineær model i Statgraphics	30
13.4	Formler til beregning af multipel regressionsanalyse	31

Statistikprogrammer

13A	Eksempler regnet med TI-89	35
13B	Eksempler regnet med Statgraphics	43

Opgaver	63
---------	----

GRUNDLÆGGENDE OPERATIONER PÅ TI - 89	71
--------------------------------------	----

STATISTISKE TABELLER

TABEL 1.	Fraktiler i U - fordelingen	73
TABEL 2.	Fraktiler i t - fordelingen	73

FACITLISTE	74
------------	----

STIKORD	75
---------	----

13 REGRESSIONSANALYSE

13.1 Indledning

I dette kapitel betragtes forsøg, hvor man har målt sammenhørende værdier af to eller flere variable. Det følgende eksempel demonstrerer et sådant tilfælde.

Eksempel 13.1.

I et spinderi udtrykkes garnets kvalitet bl.a. ved en norm for den forventede trækstyrke. Kvaliteten anses således for at være i orden, hvis middeltrækstyrken mindst er lig med 10 måleenheder (me). Ved uldgarn opfylder garnets naturlige trækstyrke ikke det nævnte kvalitetskrav, hvorfor der tilsættes en vis mængde kunstfibre, hvilket forøger trækstyrken. Herved sker der dog det, at andre kvalitetsegenskaber, såsom elasticitet og isoleringsevne, forringes. Man har eksperimenteret med forskellige tilsatte mængder kunstfibre x og registreret garnets trækstyrke y ved disse forskellige mængder. Herved fremkom følgende observationsmateriale:

Mængde x (i gram) af kunstfibre pr. kg uld	40	50	55	60	70	75	80	85	90	95	100	105	110	120	130
Trækstyrke (me): Y	4.5	6.5	5.4	7.0	8.2	8.0	7.1	8.9	8.2	10.3	9.6	10.8	10.5	11.2	12.0



Mængden af kunstfibre x er blevet bestemt på forhånd (har fået ganske bestemte værdier), så den er ikke en statistisk variabel. Trækstyrken Y synes derimod udover mængden af kunstfibre også at være påvirket af andre ukendte og ukontrollable “støjfaktorer”. Y må derfor opfattes som en statistisk variabel.

I andre situationer er både X og Y statistiske variable. Dette gælder eksempelvis, hvis man ønsker at undersøge om der er en sammenhæng mellem personers højde Y og vægt X , og derfor for en række personer måler sammenhørende værdier af højde og vægt.

Målet med en regressionsanalyse er at finde en funktionssammenhæng mellem den uafhængige variabel y og de afhængige variable.

I eksempel 13.1 ville man umiddelbart sige, at da man har 15 punktpar, så vil et polynomium af fjortende grad $y = a_{14}x^{14} + a_{13}x^{13} + \dots + a_1x + a_0$ gå igennem alle punkter, og det må derfor være en god model. Dette er imidlertid ikke tilfældet, da y -værdierne jo er resultater af forsøg der er påvirket af ukontrollable støjfaktorer. Polynomiets koefficienter vil derfor afspejle disse tilfældige udsving, og det giver derfor en ganske meningsløs model. Endvidere er modellen alt for matematisk kompliceret til at kunne bruges i praksis. Vi søger derfor i regressionsanalysen en enklere model, som tager rimeligt hensyn til støjen ved målingerne.

Er den ene variabel som i eksempel 13.1 en (kontrolleret) ikke statistisk variabel, så har man mulighed for hver x -værdi, at foretage gentagne målinger af den statistiske variabel Y (randomiseret). Dette giver mulighed for at beregne et estimat for den spredning der skyldes støjen, hvilket (som det vises i afsnit 13.2.3) kan udnyttes ved testning af den foreslåede model.

Lineær model.

Ved en lineær model forstås her en model, der er lineær med hensyn til parametrene.

Eksempelvis er såvel funktionen $y = f(x) = a + bx + cx^2$ som $y = g(P, T) = a + b \cdot P + c \cdot T$ lineære i de 3 parametre a , b og c . Som et eksempel på en model der ikke er lineær i parametrene kan nævnes $y = a + bx^c$.

Ved en regressionsanalyse søger man at finde den model, som indeholder det færreste antal koefficienter (parametre), og som samtidig beskriver dataene tilstrækkelig godt.

13.2. Enkelt regressionsanalyse.

Vi vil i dette kapitel betragte det ved anvendelserne meget ofte forekomne tilfælde, som kaldes enkelt regressionsanalyse, og hvor modellen er lineær i 2 parametre.

Som eksempler herpå kan nævnes $y = a + bx$ og $\ln y = a + b \cdot \ln x$.

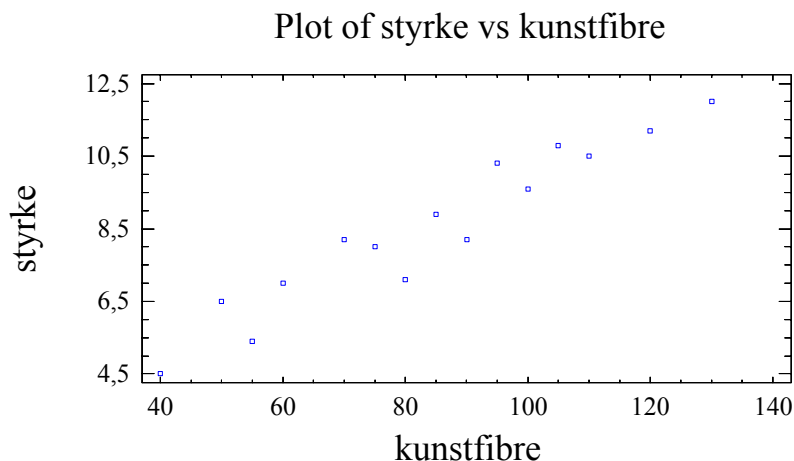
Benyttes et statistikprogram som Statgraphics til beregningsarbejdet, kan metoderne herfra nemt overføres til mere komplicerede regressionsmodeller (som det vises i afsnit 13.3 og 13.4 hvor modellerne $Y = \alpha_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \dots + \beta_p x^p$ og $Y = \alpha_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$ betragtes).

13.2.1. Forklaring af metode og formler.

Vi betragter igen eksempel 13.1

Regressionslinie og regressionskoefficienter.

Afsættes de målte punktpar (x_i, y_i) i et koordinatsystem for at få et overblik over forløbet, fås følgende tegning:



Punkterne ligger ikke eksakt på en ret linie, men det synes rimeligt at antage, at afvigelserne fra en ret linie kan forklares ved den tilfældige variation (støjen).

Derfor er det nærliggende at antage, at **middelværdien** af den statistiske variable Y er en lineær funktion af x af formen $E(Y|x) = \beta_0 + \beta_1 x$. (1)

$E(Y|x)$ skal læses middelværdien af Y for fastholdt x .

Vi vil i det følgende ofte i ligningen (1) kort skrive Y eller μ fremfor $E(Y|x)$.

Ligning (1) kaldes **regressionsligningen** (eller den teoretiske regressionsligning), grafen kaldes for **regressionslinien** (eller den teoretiske regressionslinie), og konstantledet β_0 og hældningskoefficienten β_1 kaldes **regressionskoefficienterne**.

Mens middelværdien af Y ligger på regressionslinien, kan den aktuelle observerede værdi af Y ikke forventes at ligge på den.

For et punktpar (x_i, y_i) gælder derfor, at $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, hvor ε_i kaldes den i 'te residual.

Bestemmelse af regressionslinien ved mindste kvadraters metode.

På basis af en række sammenhørende værdier af x og y bestemmes estimer $\hat{\beta}_0$ og $\hat{\beta}_1$ for regressionskoefficienterne β_0 og β_1 ved "mindste kvadraters metode". Værdierne $\hat{\beta}_0$ og $\hat{\beta}_1$ kaldes de empiriske regressionskoefficienter. Kan det ikke misforstås, så kort blot regressionskoefficienterne.

Det følgende eksempel viser metoden anvendt på et så (urealistisk) lille talekseksempel, at regningerne kan gennemføres uden anvendelse af et egentligt regressionsprogram. De angivne metoder kan imidlertid umiddelbart generaliseres til mere komplicerede eksempler, men vil så på grund af de omfattende beregninger af tidsmæssige grunde kræve anvendelse af et egentligt regressionsprogram.

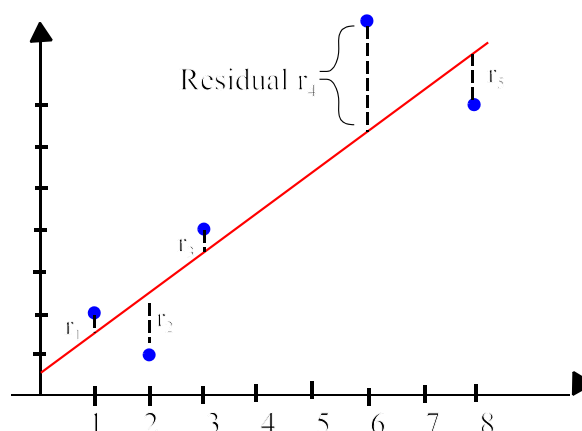
Eksempel 13.2. Bestemmelse af regressionskoefficienter ved mindste kvadraters metode.

I et medicinsk forsøg måles på en forsøgsperson sammenhørende værdier af en bestemt medicin i blodet (i %) og reaktionstiden. Resultaterne var:

x	1	2	3	6	8
y	2	1	4	9	7

Bestem ved mindste kvadraters metode et estimat for regressionslinien. ◆

Residual. Ved et punkts residual til en linie forstås den "lodrette" afstand fra punktet til linien (se tegningen).



På figur 13.1 er afsat de 5 punkter, og indtegnet en ret linie.

Figur 13.1 Residualer

Mindste Kvadraters metode. Regressionslinien $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ bestemmes som den af alle mulige rette linier, for hvilket summen af kvadratet af residualerne til linien er mindst.

I eksempel 13.2 er kvadratsummen $r_1^2 + r_2^2 + r_3^2 + r_4^2 + r_5^2$.

Løsningen af dette optimeringsproblem er angivet nedenfor (med petit).

Bestemmelse af regressionsligningen ved mindste kvadraters metode

LØSNING:

I vort tilfælde hvor vi har 5 punkter, indsættes vi disse i ligningen $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$. Dette giver:

$$2 = \hat{\beta}_0 + \hat{\beta}_1 \cdot 1, \quad 1 = \hat{\beta}_0 + \hat{\beta}_1 \cdot 2, \quad 4 = \hat{\beta}_0 + \hat{\beta}_1 \cdot 3, \quad 9 = \hat{\beta}_0 + \hat{\beta}_1 \cdot 6, \quad 7 = \hat{\beta}_0 + \hat{\beta}_1 \cdot 8.$$

De 5 ligninger med 2 ubekendte $\hat{\beta}_0$ og $\hat{\beta}_1$ kan i matrixnotation skrives:

$$Y = X \cdot B \text{ hvor } Y = \begin{pmatrix} 2 \\ 1 \\ 4 \\ 9 \\ 7 \end{pmatrix}, \quad X = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 6 \\ 1 & 8 \end{pmatrix} \text{ og } B = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix}$$

De søgte værdier af $\hat{\beta}_0$ og $\hat{\beta}_1$ findes som den løsning til dette overbestemte ligningssystem som giver den mindste RMS - fejl. Løsningen er (se evt. "Matematik for ingeniører" bind 3) bestemt ved

normalligningssystemet $X^T \cdot X \cdot B = X^T \cdot Y$.

Da matricen $X^T \cdot X$ er en kvadratisk symmetrisk matrix, som sædvanligvis ved regressionsanalyse har en invers matrix $(X^T \cdot X)^{-1}$, så er løsningen $B = (X^T \cdot X)^{-1} \cdot X^T \cdot Y$.

I vort taleksempel er

$$B = \left(\begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 6 & 8 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 6 \\ 1 & 8 \end{pmatrix} \right)^{-1} \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 6 & 8 \end{pmatrix} \begin{pmatrix} 2 \\ 1 \\ 4 \\ 9 \\ 7 \end{pmatrix} = \begin{pmatrix} 5 & 20 \\ 20 & 114 \end{pmatrix}^{-1} \begin{pmatrix} 23 \\ 126 \end{pmatrix} = \frac{1}{170} \begin{pmatrix} 114 & -20 \\ -20 & 5 \end{pmatrix} \begin{pmatrix} 23 \\ 126 \end{pmatrix} = \begin{pmatrix} \frac{102}{170} \\ \frac{170}{170} \end{pmatrix} = \begin{pmatrix} 0.6 \\ 1.0 \end{pmatrix}$$

Regressionsligningen bliver følgelig $\hat{y} = 0.6 + 1.0x$



I praksis vil man benytte et færdigt program til bestemmelse af regressionskoefficienterne. Dataene indtastes enten i et statistikprogram som Statgraphics, eller i en lommeregner med regressionsprogram som TI83 eller TI-89. I "Statistiske Grundbegreber" er der i appendix A for en række lommeregnere angivet, hvorledes man kan finde et estimat for korrelationskoefficienten r . Samtidig hermed beregner programmerne regressionskoefficienterne.

I ovenstående tilfælde vil man således af disse programmer finde, at $\hat{\beta}_0 = 0.6$ og $\hat{\beta}_1 = 1.0$

Liniens ligning bliver altså $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = 0.6 + 1.0x$

Vurdering af om model beskriver data godt.

Det er altid muligt ved mindste kvadraters metode at finde en sådan "mindste kvadraters linie". Det er den af alle rette linier, der har den mindste kvadratsum af residualerne, men det betyder ikke nødvendigvis, at linien så også er en rimelig model, som kan anvendes til at beskrive sammenhængen.

Til vurdering heraf vil man

1) se på en **tegning**.

“Mindste kvadraters linie” tegnes i et koordinatsystem sammen med punkterne. Hvis den lineære model beskriver dataene godt, skal punkterne fordeles sig tilfældigt omkring linien. I mere komplicerede tilfælde, er det nødvendigt i stedet at afsætte residualerne (i et såkaldt residualplot). Residualerne bør så fordeles sig tilfældigt omkring den vandrette 0 - linie

2) foretage nogle **beregninger** til støtte for denne vurdering.

a) Er modellen med gentagelser dvs. til hver x - værdi er målt mere end én y -værdi, har man mulighed for at foretage en egentlig statistisk test (en såkaldt “lack of fit test”). Denne test vil blive beskrevet i afsnit 13.2.3.

b) Er modellen uden gentagelser, som i eksempel 13.1, hvor der til hver x - værdi kun er målt én y -værdi har man ingen mulighed for at lave en statistisk test, men må nøjes med at foretage en beregning og vurdering af den såkaldte forklaringsgrad r^2 (også kaldet determinationskoefficient). Den angiver et talmæssigt mål for hvor tæt punkterne ligger på linien. Sædvanligvis finder man, at den fundne model på tilfredsstillende måde beskriver data, hvis forklaringsgraden er på over 70% samtidig med, at tegningen viser, at punkterne fordeles sig tilfældigt omkring den fundne regressionskurve.

3) undersøge om der er **outliers**, dvs. om enkelte målinger afviger kraftigt fra den almindelige tendens. Dette kan skyldes fejlmålinger, og sådanne punkter kan i uheldige tilfælde på grund af et stort bidrag til residualsommen få regressionslinien til at dreje. Det er dog klart, at man ikke blot kan stryge sådanne “ubehagelige” punkter. Det må kun ske, hvis man er sikker på, at punktet skyldes en fejl af en eller anden art ved målingen.

Transformation. Hvis man ikke finder, at en ret linie beskriver data godt nok, så er det jo muligt, at en anden kurve bedre beskriver sammenhængen. Eksempelvis er det jo velkendt fra matematikken, at graferne for eksponentialfunktioner og potensfunktioner ved en passende logaritmisk transformation kan blive til rette linier. Det giver naturligvis lidt mere komplicerede regninger, men statistikprogrammer og også en del lommeregnerne kan dog let foretage en regressionsanalyse også i sådanne tilfælde. I eksempel 13.5 er et sådant eksempel gennemgået.

Sammenhæng mellem korrelationskoefficient og forklaringsgrad.

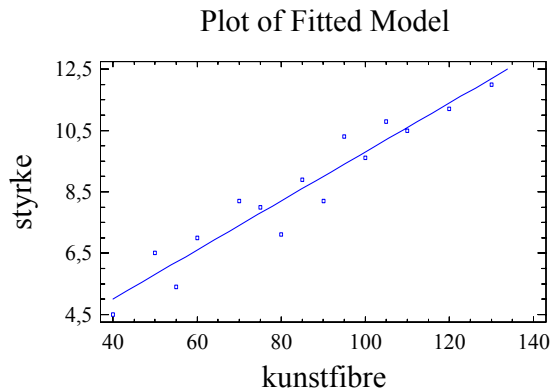
Hvis både X og Y er normalfordelte statistiske variable (som eksempelvis når man aflæser sammenhørende værdier af højde og vægt for en række personer) angiver korrelationskoefficienten ρ (jævnfør eventuelt “Statistiske Grundbegreber kapitel 10) en størrelse mellem -1 og 1 som kan anvendes til at angive, om der er en sammenhæng (korrelation) mellem X og Y . Er korrelationskoefficienten positiv har punkterne en voksende tendens, hvis den er negativ har de en aftagende tendens. Et estimat for ρ er størrelsen r . Kvadreres den er r^2 den samme som forklaringsgraden. Hvis kun Y er en statistisk variabel (som det er tilfældet i eksempel 13.1) eksisterer ρ ikke, men r kan naturligvis stadig udregnes, men er så ikke estimat for ρ .

Ekstrapolation. Selv om modellen synes på tilfredsstillende måde at beskrive data, så er det jo faktisk kun sikkert indenfor måleområdet. Man skal være yderst forsigtig med at ekstrapolere, dvs. på basis af modellen for x - værdier udenfor måleområdet beregne hvad y er.

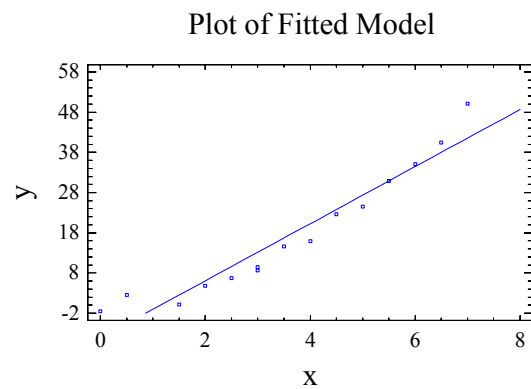
At man ikke alene kan stole på forklaringsgraden illustreres ved følgende eksempel.

Eksempel 13.3 .Grafisk vurdering af model.

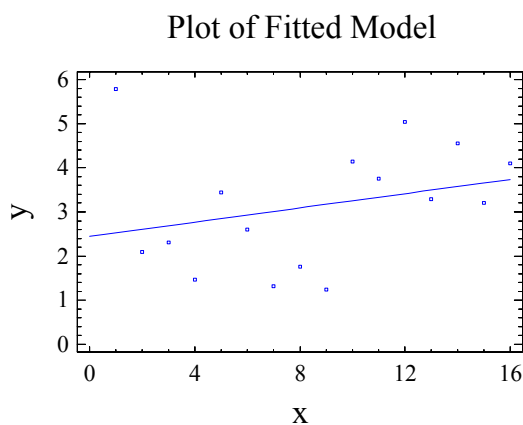
De følgende 4 figurer afspejler forskellige muligheder.



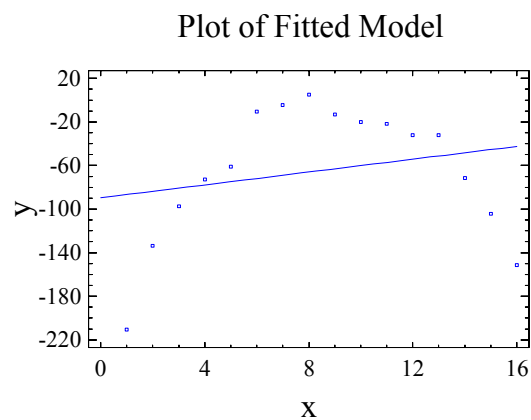
Figur 13.2a: $r = 0.959$ $r^2 = 91.9\%$



Figur 13.2b: $r = 0.962$ $r^2 = 92.6\%$



Figur 13.2c: $r = 0.278$ $r^2 = 7.73\%$



Figur 13.2d: $r = 0.229$ $r^2 = 5.24\%$

I figur 13.2a synes den lineære model at kunne beskrive dataene godt, idet punkterne fordeles sig tilfældigt omkring linien, og forklaringsgraden $r^2 = 91.9\%$ er høj.

I figur 13.2b er forklaringsgraden også høj, og punkterne ligger da også tæt ved linien. Imidlertid ligger punkterne ikke tilfældigt omkring linien. Yderpunkterne ligger over og de midterste punkter under linien, så det er næppe rimeligt at anvende en ret linie som model. I stedet kunne man overveje en eksponentialfunktion eller et andengradspolynomium.

I figur 13.2c er der næppe nogen relation mellem x og y . Er x og y uafhængige (ingen relation mellem x og y) vil punkterne fordele sig tilfældigt omkring gennemsnitslinien $y = \bar{y}$, og forklaringsgraden være 0. Vi ser, at regressionslinien er næsten vandret, og forklaringsgraden ringe.

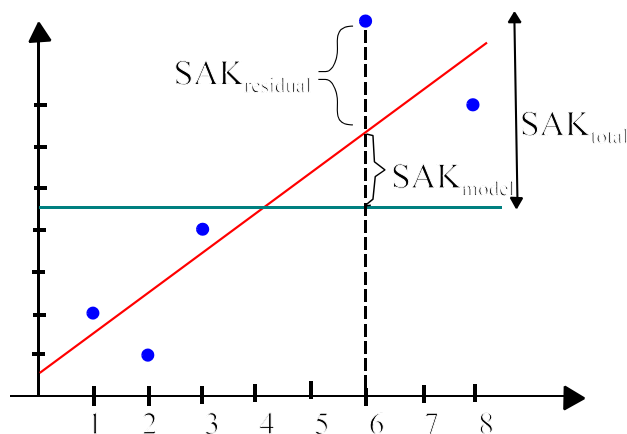
I figur 13.2d er forklaringsgraden også lille, men alligevel må vi antage at der er en sammenhæng mellem x og y . Den er blot ikke lineær, men muligvis en parabel. ♦

Definition og beregning af forklaringsgrad.

I praksis vil man lade en lommeregner eller en PC med et statistikprogram beregne de enkelte statistiske størrelser.

Ved tolkningen af de fremkomne størrelser vil en anskuelig forståelse af størrelserne dog være nyttig.

I det følgende vil vi derfor definere nogle fundamentale definitioner, og søge at anskueliggøre dem dels på figur 13.3 dels ved at foretage beregningerne på tallene i eksempel 13.2.



Figur 13.3. SAK - størrelser

Definitioner:

SAK_{total} = sum af kvadrater af residualerne til den vandrette linie $y = \bar{y}$

Data i eksempel 13.2 giver: $\bar{y} = \frac{2 + 1 + 4 + 9 + 7}{5} = 4.6$

De 5 punkters "residualer" til den vandrette linie $y = \bar{y}$.

$$r_1 = 2 - 4.6 = -2.6, r_2 = 1 - 4.6 = -3.6, r_3 = 4 - 4.6 = -0.6, r_4 = 9 - 4.6 = 4.4, r_5 = 7 - 4.6 = 2.4.$$

Vi får $SAK_{total} = r_1^2 + r_2^2 + \dots + r_5^2 = 2.6^2 + 3.6^2 + 0.6^2 + 4.4^2 + 2.4^2 = 45.2$.

$SAK_{residual}$ = sum af kvadrater på de enkelte punkters afstand fra den fundne regressionslinje.

Af eksempel 13.2 fås følgende residualer til den fundne regressionslinje $y = 0.6 + 1.0x$:

$$r_1 = 2 - (0.6 + 1 \cdot 1) = 0.4, r_2 = 1 - (0.6 + 1 \cdot 2) = -1.6, r_3 = 4 - (0.6 + 1 \cdot 3) = 0.4,$$

$$r_4 = 9 - (0.6 + 1 \cdot 6) = 2.4, r_5 = 7 - (0.6 + 1 \cdot 8) = 0.4.$$

$$SAK_{residual} = r_1^2 + r_2^2 + \dots + r_5^2 = 0.4^2 + (-1.6)^2 + 0.4^2 + 2.4^2 + 0.4^2 = 11.2$$

SAK_{model} = sum af kvadrater af "regressionsliniens afstand" fra det totale gennemsnit $\bar{y}$.

Af eksempel 13.2 fås residualerne for regressionslinien $y = 0.6 + 1.0x$'s "afstand" fra det totale gennemsnit $\bar{y} = 4.6$.

$$r_1 = 0.6 + 1 \cdot 1 - 4.6 = -3.0, r_2 = 0.6 + 1 \cdot 2 - 4.6 = -2.0, r_3 = 0.6 + 1 \cdot 3 - 4.6 = -1.0,$$

$$r_4 = 0.6 + 1 \cdot 6 - 4.6 = 2.0, r_5 = 0.6 + 1 \cdot 8 - 4.6 = 4.0.$$

$$SAK_{model} = r_1^2 + r_2^2 + \dots + r_5^2 = (-3.0)^2 + (-2.0)^2 + (-1.0)^2 + 2.0^2 + 4.0^2 = 34$$

Der gælder generelt, at $SAK_{total} = SAK_{residual} + SAK_{model}$ (jævnfør, at $45.2 = 11.2 + 34$)

Forklaringsgraden r^2 er bestemt ved

$$r^2 = \frac{SAK_{model}}{SAK_{total}} = \frac{SAK_{total} - SAK_{residual}}{SAK_{total}} = 1 - \frac{SAK_{residual}}{SAK_{total}} \quad (r^2 = \frac{34}{45.2} = 0.752)$$

Anskuelig forklaring:

Hvis Y er uafhængig af x vil regressionslinien stille sig næsten vandret, dvs. $y \approx \bar{y}$.

Det betyder igen at $SAK_{\text{residual}} \approx SAK_{\text{total}}$ og dermed at $r^2 \approx 0$.

Hvis derimod Y er lineært afhængig af x vil regressionslinien have en hældning forskellig fra nul.

Det betyder igen at $SAK_{\text{residual}} \ll SAK_{\text{total}}$, og dermed at $r^2 \approx 1$.

Man siger også, at den fundne model "forklarer" $r^2 \cdot 100\%$ af den "totale variation"

I eksempel 13.2 forklarer den fundne model således "75.2% af den totale variation."

Forudsætninger for regressionsanalyse.

De foregående betragtninger kræver ingen statistiske forudsætninger, idet man jo altid ved mindste kvadraters metode kan beregne regressionskoefficienterne, beregne forklaringsgrad, tegne kurver og punkter ind i et koordinatsystem og så herudfra vurdere om modellen er acceptabel.

Ønsker vi at foretage en nøjere statistisk analyse som eksempelvis at teste "om Y er uafhængig af x , dvs. af om $\beta_1 = 0$ ", eller opstille konfidensintervaller for β_1 må observationerne opfylde visse krav.

- 1) De enkelte observationer y_i er indbyrdes uafhængige (eksempelvis hvis der udføres flere målinger for samme mængde medicin skal de være indbyrdes uafhængige, ligesom det også skal gælde målinger baseret på forskellige mængder medicin.
- 2) 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 konstant varians σ^2 .

Da $V(Y) = V(\beta_0 + \beta_1 x + \varepsilon) = V(\beta_0 + \beta_1 x) + V(\varepsilon) = 0 + \sigma^2$ betyder kravet om konstant varians, at variansen af Y skal være den samme uafhængig af x 's værdi. Man siger kort, at der forudsættes varianshomogenitet.

Kravene i punkt 1 kan opfyldes ved en hensigtsmæssig forsøgsplan. I eksempel 13.2 skal man således være sikker på at den foregående dosis medicin er ude af blodet inden man foretager en ny indsprøjtning, ligesom forsøgene skal være randomiseret. Man kan nok i dette tilfælde betvivle uafhængigheden, hvis man udfører forsøgene på samme person.

Kravene i punkt 2 er sædvanligvis opfyldt i praksis, da det heldigvis er sådan, at analysen stadig er gyldig, selv om der forekommer mindre afvigelser. Hvis der er det samme antal gentagelser for hver x -værdi er analysen så robust overfor afvigelser, at man kun behøver at kontrollere det, hvis man har en stærk mistanke til store forskelle.

Er der derimod et forskelligt antal gentagelser, bliver analysen følsom overfor afvigelser.

Er man i tvivl om kravet om normalitet er rimeligt opfyldt, kan man få et indtryk af, om der er alvorlige afvigelser, ved at tegne et normalfordelingsplot (vises i afsnit 13B eksempel 13.4)

Har vi for hver x -værdi målt flere y -værdier, har vi mulighed for at teste kravet om varianshomogenitet (vises også i afsnit 13B eksempel 13.4).

Variansanalysetabel.

Til hjælp ved analysen udarbejdes en såkaldt “variansanalysetabel”, som på overskuelig form samler de væsentligste teststørrelser.

En sådan er angivet nedenfor. Her er N antallet af punkter, og i parentes er angivet de tilsvarende engelske betegnelser : SS = Sum of Squares, df = degree of freedom.

Variation (Source)	SAK (SS)	f (df)	$s^2 = \frac{\text{SAK}}{f}$	$F = \frac{s_{\text{model}}^2}{s_{\text{residual}}^2}$	P - værdi
Model	$\text{SAK}_{\text{model}}$	1	$s_{\text{model}}^2 = \frac{\text{SAK}_{\text{model}}}{1}$	$F_{\text{model}} = \frac{s_{\text{model}}^2}{s_{\text{residual}}^2}$	$P(Z > F_{\text{model}})$
Residual	$\text{SAK}_{\text{residual}}$	$N - 2$	$s_{\text{residual}}^2 = \frac{\text{SAK}_{\text{residual}}}{N - 2}$		
Total	$\text{SAK}_{\text{total}}$	$N - 1$			

Den statistiske variabel Z er F - fordelt med frihedsgraderne $(f_T, f_N) = (1, N - 2)$.

Vi har i forbindelse med beregningen af forklaringsgrad forklaret betydningen af SAK’størrelserne .

Frihedsgradstallet for $\text{SAK}_{\text{model}}$ er $f_{\text{model}} = (\text{antal parametre i ligningen}) - 1 = 2 - 1 = 1$

Frihedsgradstallet for $\text{SAK}_{\text{residual}}$ er $f_{\text{residual}} = (\text{totale antal punkter}) - 2$

(jævnfør, at det overbestemte ligningssystem har 5 ligninger og 2 ubekendte).

Frihedsgradstallet for $\text{SAK}_{\text{total}}$ er $f_{\text{total}} = (\text{totale antal punkter}) - 1$

Der gælder generelt, at $\text{SAK}_{\text{total}} = \text{SAK}_{\text{residual}} + \text{SAK}_{\text{model}}$, og $f_{\text{total}} = f_{\text{residual}} + f_{\text{model}}$

Test af om Y er uafhængig af x

Lad os antage, at de i afsnit 13.2.4 angivne forudsætninger er opfyldt, og at vi har fundet (ved at betragte tegning + forklaringsgrad eller tegning + “lack of fit test”), at modellen $E(Y|x) = \beta_0 + \beta_1 x$ gælder.

Hvis Y er uafhængig af x betyder det, at regressionslinien er vandret, eller at hældningskoefficienten β_1 er 0.

Vi får altså: $H_0: Y \text{ er uafhængig af } x \Leftrightarrow H_0: \text{Regressionslinien er vandret} \Leftrightarrow H_0: \beta_1 = 0$.

Metode 1: F - test.

Hvis modellen gælder så burde punkterne (uanset om H_0 er sand eller ej) ligge eksakt på en ret linie (og dermed $s_{\text{residual}}^2 = 0$), hvis ikke forsøgsresultaterne havde været påvirket af ”støjen” .

Et estimat for forsøgsfejlens (støjens) varians σ^2 er derfor s_{residual}^2 .

Er H_0 sand, så burde (jævnfør definitionen af $\text{SAK}_{\text{model}}$) s_{model}^2 være nul. Når det ikke er tilfældet skyldes det, at forsøgsresultaterne har været påvirket af “støjen”. Af samme grund som før må derfor også s_{model}^2 være et estimat for σ^2 .

Vi har følgelig, at hvis H_0 er sand, så er $F_{\text{model}} = \frac{s_{\text{model}}^2}{s_{\text{residual}}^2} \approx 1$.

Det kan vises, at hvis nulhypotesen ikke er sand, så vil $F_{\text{model}} > 1$, og at F_{model} er F-fordelt med en tællerfrihedsgrad på 1 og en nævnerfrihedsgrad på $N - 2$.

Testen bliver følgelig en ensidet F -test, dvs.

H_0 forkastes, hvis P -værdi $= P(F > F_{\text{model}}) < \alpha$

Metode 2. t -test.

$H_0: Y$ er uafhængig af $x \Leftrightarrow H_0: \beta_1 = 0$

Lad $t = \frac{\hat{\beta}_1}{s_{\beta_1}}$, hvor $s_{\beta_1} = \frac{|\hat{\beta}_1| \cdot s_{\text{residual}}}{s_{\text{model}}}$ er et estimat for spredningen på β_1 .

Det kan vises, at t er t -fordelt med $N - 2$ frihedsgrader.

Idet T er t -fordelt med $N - 2$ frihedsgrader, gælder det derfor, at

H_0 forkastes, hvis P -værdi $= P(T > |t|) < \alpha$

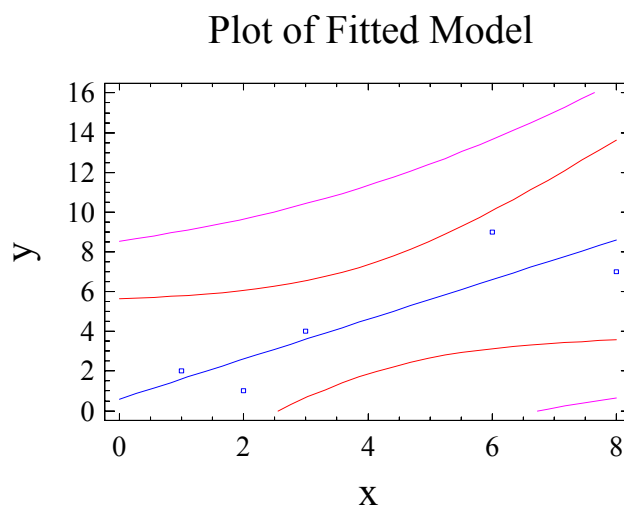
På tilsvarende måde kan man teste $H_0: \beta_1 \geq 0$ og $H_0: \beta_1 \leq 0$ ved ensidede test.

Hvis begge variable X og Y er statistiske variable kan man tilsvarende teste korrelationen ρ ved ovennævnte t -test.

Konfidensintervaller og prædiktionsintervaller.

Et led i analysen kan være, at udregne et 95% konfidensinterval for β_1 . Endvidere vil man ofte være interesseret i en speciel værdi for x , for hvilken man ønsker beregnet såvel den tilsvarende "forventede" y -værdi (predicted value) som et 95%-konfidensinterval for middelværdien $\tilde{\mu}$ og et 95% prædiktionsinterval for én ny observation.

På figur 13.4 er her tegnet 95% konfidensintervaller for middelværdierne (de inderste buede kurver), og 95% prædiktionsintervaller (de yderste to kurver). Man ser tydeligt, at konfidensintervallerne er smallest omkring "centrum" $(\bar{x}, \bar{y})$.



Figur 13.4. Konfidensintervaller og prædiktionsintervaller

I oversigt 13.1 er angivet formler for bl.a. disse konfidensintervaller.

Konfidensinterval for β_1 :
$$\left[\hat{\beta}_1 - t_{1-\frac{\alpha}{2}}(N-2) \cdot s_{\beta_1} ; \hat{\beta}_1 + t_{1-\frac{\alpha}{2}}(N-2) \cdot s_{\beta_1} \right]$$

Til en given x - værdi er $\hat{\mu}$ et estimat for middelværdien μ for Y .

$$\hat{\mu} = \hat{\beta}_0 + \hat{\beta}_1 x \quad \text{og} \quad \hat{V}(\mu) = s_{\text{residual}}^2 \cdot \left(\frac{1}{N} + \frac{(x - \bar{x})^2 (\hat{\beta}_1)^2}{SAK_{\text{model}}} \right).$$

Konfidensinterval for μ :
$$\left[\hat{\mu} - t_{1-\frac{\alpha}{2}}(N-2) \cdot \sqrt{\hat{V}(\mu)} ; \hat{\mu} + t_{1-\frac{\alpha}{2}}(N-2) \cdot \sqrt{\hat{V}(\mu)} \right]$$

13.2.2 Beregning af enkelt regressionsanalyse med én y-observation for hver x - værdi.

Vi vil i dette afsnit se på det tilfælde, hvor der for hver x - værdi kun er målt én y - værdi.

I eksempel 13.4 vises hvorledes man kan foretage beregningerne ved anvendelse af formlerne i oversigt 13.1.

I afsnit 13A og 13B vises, hvorledes beregningerne foretages med anvendelse af henholdsvis TI-89 og Statgraphics.

Eksempel 13.4. Enkelt regressionsanalyse regnet ved anvendelse af formler.

I et spinderi udtrykkes garnets kvalitet bl.a. ved en norm for den forventede trækstyrke. Kvaliteten anses således for at være i orden, hvis middeltrækstyrken mindst er lig med 10 måleenheder (me). Ved uldgarn opfylder garnets naturlige trækstyrke ikke det nævnte kvalitetskrav, hvorfor der tilsættes en vis mængde kunstfibre, hvilket forøger trækstyrken. Herved sker der dog det, at andre kvalitetsegenskaber, såsom elasticitet og isoleringsevne, forringes. Man har eksperimenteret med forskellige tilsatte mængder kunstfibre x og registreret garnets trækstyrke y ved disse forskellige mængder. Herved fremkom følgende observationsmateriale:

Mængde x (i gram) af kunstfibre pr kg uld	40	50	55	60	70	75	80	85	90	95	100	105	110	120	130
Trækstyrke (me): Y	4.5	6.5	5.4	7.0	8.2	8.0	7.1	8.9	8.2	10.3	9.6	10.8	10.5	11.2	12.0

- 1) Beregn forklaringsgraden r^2 og anvend denne samt en figur på lommeregnerens grafiske display (eller figur 11.2a) til vurdering af modellen.
 - 2) Opskriv regressionsligningen.
- Der antages i det følgende at regressionsforudsætningerne er opfyldt.
- 3) Opskriv en variansanalysetabel, og test om Y er uafhængig af x .
 - 4) Angiv et konfidensinterval for hældningskoefficienten β_1 .
 - 5) Opstil et 95% konfidensinterval for middeltrækstyrken svarende til x - værdien 100.
 - 6) Opstil et 95% prædeterminationsinterval for middeltrækstyrken svarende til x - værdien 100.

Løsning:

De 15 punktpar indtastes i en lommeregner, og et regressionsprogram aktiveres.

Blandt de beregnede størrelser findes: $\bar{x} = 84.333$ og $s_x = 26.38$, $\bar{y} = 8.5467$ og $s_y = 2.198$.

Endvidere findes estimater for regressionskoefficienter: $\hat{\beta}_0 = 1.8087$ og $\hat{\beta}_1 = 0.0799$ og korrelationskoefficienten $r = 0.958802$.

1) Forklaringsgraden er $r^2 = 0.958802^2 = \underline{0.919301}$.

Da punkterne på figuren (se figur 11.2a) fordeler sig tilfældigt omkring linien, der ikke er enkelte punkter (outliers) der afviger voldsomt fra linien, og forklaringsgraden samtidig er tæt på 1, så er den lineære model acceptabel.

1) Den estimerede regressionsligningen bliver $y = 0.0799x + 1.8087$

3) Man oversigt 13.1 fås

$$SAK_{\text{total}} = (N - 1) \cdot s_y^2 = (15 - 1) \cdot 2.198^2 = 67.657,$$

$$SAK_{\text{model}} = r^2 \cdot SAK_{\text{total}} = 0.9588^2 \cdot 67.657 = 62.197 \quad \text{og}$$

$$SAK_{\text{residual}} = SAK_{\text{total}} - SAK_{\text{model}} = 67.657 - 62.197 = 5.4599$$

Variansanalysetabel udfyldes:

Variation (Source)	SAK (SS)	f (df)	$s^2 = \frac{SAK}{f}$	F
Model	62.1974	1	62.1974	148.09
Residual	5.4599	13	0.4200	
Total	67.657	14		

$$H_0: y \text{ er uafhængig af } x \Leftrightarrow H_0: \text{Model} = 0 \Leftrightarrow H_0: \beta_1 = 0.$$

Metode 1: P - værdi = $P(F < 148.09) = FCdf(148.09, \infty, 1, 13) = 1.77 \cdot 10^{-8}$

Da P - værdi = $1.77 \cdot 10^{-8} < 0.001$ forkastes H_0 (stærkt) dvs.

Konklusion: Y er ikke uafhængig af x .

Metode 2:

$$s_{\beta_1} = \frac{|\hat{\beta}_1| \cdot s_{\text{residual}}}{s_{\text{model}}} = \frac{0.0799 \cdot \sqrt{0.4200}}{\sqrt{62.1974}} = 0.00656, \quad t = \frac{\hat{\beta}_1}{s_{\beta_1}} = \frac{0.0799}{0.00656} = 12.169$$

$$P\text{-værdi} = P(T > 12.169) = tCdf(12.169, \infty, 13) = 8.83 \cdot 10^{-9}$$

Da P - værdi = $8.83 \cdot 10^{-9} < 0.0005$ forkastes H_0 (stærkt) dvs.

Konklusion: Y er ikke uafhængig af x .

Mange statistikprogrammer (og det gælder TI-89 og Statgraphics) beregner ved to-sidede test begge "haler", dvs. P - værdien bliver dobbelt så stor, og man skal så altid sammenligne med signifikansniveauet α

3) Konfidensinterval for β_1 :

$$\hat{\beta}_1 \pm t_{0.975}(15-2) \cdot s_{\beta_1} = 0.0799 \pm 2.16 \cdot 0.00656 = 0.0799 \pm 0.01418$$

$$\underline{\underline{[0.0657; 0.0941]}}$$

5) Til x - værdien 100 er et estimat for middelværdien $\hat{\mu} = 0.0799 \cdot 100 + 1.8087 = 9.80$

$$V(\mu) = s_{residual}^2 \cdot \left(\frac{1}{N} + \frac{(x - \bar{x})^2 \cdot (\hat{\beta}_1)^2}{SAK_{model}} \right) = 0.4200 \left(\frac{1}{15} + \frac{(100 - 84.33)^2 \cdot 0.0799^2}{62.1974} \right) = 0.0386.$$

Konfidensinterval for μ :

$$\hat{\mu} \pm t_{0.975}(15-2) \cdot \sqrt{V(\mu)} = 9.80 \pm 2.16 \cdot \sqrt{0.0386} = 9.80 \pm 0.4243.$$

$$\underline{\underline{[9.38; 10.22]}}$$

6) Prædiktionsinterval for 1 ny observation svarende til x - værdien 100.

$$Q = s_{residual}^2 \cdot \left(1 + \frac{1}{N} + \frac{(x - \bar{x})^2 \cdot (\hat{\beta}_1)^2}{SAK_{model}} \right) = 0.4200 \left(1 + \frac{1}{15} + \frac{(100 - 84.33)^2 \cdot 0.0799^2}{62.1974} \right) = 0.4586$$

$$\hat{\mu} \pm t_{0.975}(15-2) \cdot \sqrt{Q} = 9.80 \pm 2.16 \cdot \sqrt{0.4586} = 9.80 \pm 1.463. \quad \underline{\underline{[8.34; 11.26]}} \quad \blacklozenge$$

Eksempel 13.3 er i afsnit 13A og 13B regnet med anvendelse af TI-89 og Statgraphics.

13.2.3. Enkelt Regressionsanalyse med flere y - værdier for hver x - værdi.

I forsøgsplanlægning hvor man selv fastlægger sine x - niveauer, er det ofte muligt for hver x - værdi, at foretage flere målinger af y - værdien. Vi siger kort at analysen er “med gentagelser”. Dette er således tilfældet i følgende eksempel:

Eksempel 13.5 (regressionsanalyse med gentagelser)

Metalpladers overflader oxideres i en ovn ved 200° C. Med henblik på en undersøgelse af sammenhængen mellem det oxiderede lags tykkelse y (i ångstrøm) og tiden t (i minutter) foretog man følgende målinger:

Tiden t	20	30	40	60	70	90	100	120	150	180
Tykkelse y	4.2	7.4	8.8	13.6	13.1	14.9	20.0	23.1	27.5	32.9
	4.9	6.9	8.2	12.0	12.4	16.8	21.2	25.2	25.1	32.4

Fordelen herved er, at man nu kan få et estimat for forsøgsfejls spredning (“støjen”), som kan anvendes til at teste, om den lineære model kan accepteres, når man tager støjen i betragtning. Endvidere kan man, hvis man finder det nødvendigt, teste om der er varianshomogenitet. Alle andre test udføres på samme måde som beskrevet i forrige afsnit.

Forklaring af metode og formler**Test af model.**

For hver x - værdi beregnes gennemsnittet af de dertil hørende y - værdier. Disse “gennemsnitspunkter” bør ligge tæt på linien hvis modellen er god. Hvis modellen er den rigtige, så er den eneste grund til at “gennemsnitspunkterne ikke ligger eksakt på linien, at der er støj.

Vi kan derfor beregne et estimat (kaldet $s_{\text{lack of fit}}^2$) for variansen af denne støj ud fra de afvigelser som gennemsnitspunkterne har. Hvorledes denne beregnes ses i oversigt 13.2.

Da vi samtidig ud fra gentagelserne kan beregne et andet estimat for støjen (kaldet s_0^2), har vi mulighed for at teste de to varianser mod hinanden, ved en sædvanlig F - test

$$F_{\text{lack of fit}} = \frac{s_{\text{lack of fit}}^2}{s_0^2}$$

Får vi her en forkastelse, kan “gennemsnitspunkternes” afvigelser fra linien ikke forklares alene ved støjen, og vi må derfor forkaste modellen.

Test af varianshomogenitet.

Som tidligere nævnt, er analysen robust overfor afvigelser fra kravet om varianshomogenitet (konstant varians σ^2), hvis der er lige mange gentagelser (som i dette forsøg). Man vil derfor kun foretage en vurdering af dette krav, hvis man ud fra forsøgets natur mener, at varianserne kan tænkes at være voldsomt forskellige. Har man ingen gentagelser, må man i alle tilfælde basere sig på sin viden om, at “det er overvejende sandsynligt at der er nogenlunde samme varians”.

Vi vil derfor sædvanligvis i eksempler og øvelser undlade at foretage en testning.

Er der ikke lige mange gentagelser, så bør man dog foretage en nøjere undersøgelse af kravet, da analysen så er følsom overfor afvigelser.

Simplificeret F - test. En enkel måde er at foretage “en simpel F - test” mellem den største og mindste varians (se eventuelt oversigt 13.2). Får vi en accept herved går vi ud fra der er varianshomogenitet.

Får vi en forkastelse, kan dette skyldes, at vi jo har set på det værst tænkelige tilfælde, og ikke har taget hensyn til de mellemliggende målinger. Vi må derfor så gå over til at benytte metoder, der tager dette hensyn.

“Bartlett's test”. Denne test er beskrevet i oversigt 13.2. Den er meget benyttet, men har den svaghed, at den i højere grad end selve regressionsanalysen er følsom overfor afvigelser fra normalitet.

Statgraphics kan foretage Bartlett's test, som beskrevet i afsnit 13B.

Eksempel 13.6. Regressionsanalyse (med gentagelser)

Metalpladers overflader oxideres i en ovn ved 200°C . Med henblik på en undersøgelse af sammenhængen mellem det oxiderede lags tykkelse y (i ångstrøm) og tiden t (i minutter) foretog man følgende målinger:

Tiden t	20	30	40	60	70	90	100	120	150	180
Tykkelse y	4.2	7.4	8.8	13.6	13.1	14.9	20.0	23.1	27.5	32.9
	4.9	6.9	8.2	12.0	12.4	16.8	21.2	25.2	25.1	32.4

Det antages, at målingerne y er værdier af uafhængige normalfordelte variable med samme varians.

1) Foretag en testning af forudsætningen om varianshomogenitet.

2) Det formodes på forhånd, at der er en lineær sammenhæng mellem x og y .

Test denne formodning ved en "lack of fit" test, og bestem i bekræftende fald ligningen for den empiriske regressionslinje.

Det antages i det følgende, at der er en lineær sammenhæng mellem x og y .

3) I litteraturen vedrørende dette problem påstås, at hældningskoefficienten β_1 er 0.15

Test om dette på et signifikansniveau på 5% kan være sandt.

4) Angiv et 95% konfidensinterval for middelværdien af tykkelsen y , når $t = 100$ minutter.

Løsning:

1) Lad de 10 varianser (svarende til de 10 t -værdier) benævnes $\sigma_1^2, \sigma_2^2, \dots, \sigma_{10}^2$.

Test af nulhypotesen $H_0: \sigma_1^2 = \sigma_2^2 = \dots = \sigma_{10}^2$

Der foretages en "simplificeret F - test".

Estimer for varianserne beregnes:

Tiden t	20	30	40	60	70	90	100	120	150	180
Tykkelse y	4.2	7.4	8.8	13.6	13.1	14.9	20.0	23.1	27.5	32.9
	4.9	6.9	8.2	12.0	12.4	16.8	21.2	25.2	25.1	32.4
s^2	0.245	0.125	0.180	1.280	0.245	1.805	0.720	2.205	2.880	0.125

$$F = \frac{s_{\max}^2}{s_{\min}^2} = \frac{2.880}{0.125} = 23.04 \quad . \quad P\text{-værdi} = P(F > 23.04) = FCdf(23.04, \infty, 1, 1) = 0.1308$$

Da P -værdi = 0.1308 > 0.025 accepteres nulhypotesen, dvs. vi vil i det følgende antage, at forskellene i varians ikke er så store, at det ødelægger analysen.

$$\text{Vi får } s_e^2 = \frac{s_1^2 + s_2^2 + \dots + s_k^2}{k} = \frac{9.81}{10} = 0.981.$$

2) De $N = n \cdot k = 2 \cdot 10 = 20$ punktpar (x_i, y_{ij}) indtastes i lommeregner.

Regressionsprogram aktiveres, og blandt beregnede størrelser findes:

Estimer for: Regressionskoefficienter: $\hat{\beta}_0 = 1.654$ og $\hat{\beta}_1 = 0.1730$,

korrelationskoefficient $r = 0.9892$, gennemsnit $\bar{x} = 86.0$ og spredning $s_y = 8.9412$.

13. Regressionsanalyse

Man udregner $SAK_{\text{total}} = (N - 1) \cdot s_y^2 = 19 \cdot 8.9412^2 = 1518.982$,

$$SAK_{\text{model}} = r^2 \cdot SAK_{\text{total}} = 0.9892^2 \cdot 1518.982 = 1486.44$$

$$SAK_e = (N - k) \cdot s_0^2 = (20 - 10) \cdot 0.981 = 9.81$$

$$SAK_{\text{lack of fit}} = SAK_{\text{total}} - SAK_{\text{model}} - SAK_0 = 22.732$$

Udfylder variansanalysetabel:

Variation (Source)	SAK (SS)	f (df)	$s^2 = \frac{SAK}{f}$	F
Model	1486.44	1	1486.44	
Lack of fit	22.732	8	2.8415	$F_{\text{lack of fit}} = \frac{2.8410}{0.981} = 2.90$
Gentagelser (error)	9.81	10	0.981	
Total	1518.98	19		

H_0 : Lineær model gælder $\Leftrightarrow H_0: (x_i, \mu_i)$ ligger på en ret linie

Da P -værdi $= P(F > 2.90) = FCdf(2.90, \infty, 8, 10) = 0.05891 > 0.05$ accepteres H_0 ,

dvs. vi vil i det følgende antage, at den lineære model gælder.

Vi er dog så tæt på forkastelse, at en nærmere undersøgelse kan være rimelig.

Af hensyn til de følgende regninger "pooles" s_e^2 og s_{residual}^2

$$s_{\text{residual}}^2 = \frac{SAK_e + SAK_{\text{lack of fit}}}{f_0 + f_{\text{lack of fit}}} = \frac{9.81 + 22.7328}{8 + 10} = 1.8079 \quad f_{\text{residual}} = 18$$

Den empiriske regressionslinie bliver: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot t = 1.654 + 0.1730 \cdot t$.

3) $H_0: \beta_1 = 0.15$

$$\text{Lad } s_{\beta_1} = \frac{|\hat{\beta}_1| \cdot s_{\text{residual}}}{s_{\text{model}}} = \frac{0.1730 \cdot \sqrt{1.8079}}{\sqrt{1486.40}} = 0.0060$$

$$t = \frac{\hat{\beta}_1 - a}{s_{\beta_1}} = \frac{0.1730 - 0.15}{0.0060} = 3.812 \quad \text{er } t \text{ - fordelt med } f = f_{\text{residual}} = 18$$

Da P -værdi $= P(T > 3.81) = tCdf(3.81, \infty, 18) = 0.00064 < 0.025$ forkastes H_0 , dvs. data giver ikke den i litteraturen angivne hældningskoefficient.

4) 95% konfidensinterval for middelværdien af tykkelsen y , når $t = 100$ minutter.

$$\text{Beregner hjælpe størrelserne } \hat{\mu} = \hat{\beta}_0 + \hat{\beta}_1 t = 1.645 + 0.1730 \cdot 100 = 18.95$$

$$\hat{V}(\hat{\mu}) = s_{\text{residual}}^2 \cdot \left(\frac{1}{N} + \frac{(x_0 - \bar{x})^2 (\hat{\beta}_1)^2}{SAK_{\text{model}}} \right) = 1.8079 \left(\frac{1}{20} + \frac{(100 - 86)^2 \cdot 0.1730^2}{1486.44} \right) = 0.09753$$

Konfidensinterval for μ

$$\left[\hat{\mu} - t_{1-\frac{\alpha}{2}}(N-2) \cdot \sqrt{\hat{V}(\hat{\mu})}; \hat{\mu} + t_{1-\frac{\alpha}{2}}(N-2) \cdot \sqrt{\hat{V}(\hat{\mu})} \right]$$

$$= [18.95 - 2.10 \cdot 0.3123; 18.95 + 2.10 \cdot 0.3123] = [18.29; 19.61]$$



Det samme eksempel er i afsnit 13A og 13B beregnet med anvendelse af henholdsvis TI-89 og Statgraphics.

13.2.4. Transformation af data inden lineær regressionsanalyse kan foretages.

Ligger punkterne ikke tilnærmelsesvis på en ret linie, er det muligt, at man ved at vælge en passende transformation kan føre problemet over i en lineær model i de transformerede data.

På visse lommeregnere såsom TI83, TI-89 og HP48G kan man vælge mellem forskellige ofte anvendbare modeller. Man kan så hurtigt grafisk og ud fra forklaringsgraden se hvilken transformation, der bedst svarer til de opgivne data.

I oversigt 13.3 er angivet en liste med kommentarer over de mest almindelige transformationer. Har man på forhånd en viden om, at en bestemt transformation skal anvendes, kan man uden større besvær foretage den pågældende transformation på dataene og så udføre regressionsanalysen på de transformerede data.

Dette illustreres ved følgende eksempel.

Eksempel 13.7. Transformation af udtryk.

Ved et forsøg ønskedes en bestemt luftarts adiabatexponent γ bestemt ved, at luftarten adiabatisk komprimeres til forskellige forudvalgte rumfang v , idet de tilsvarende værdier af trykket P målt. Man formodede på forhånd, at der gælder regressionsmodellen $P = c \cdot v^{-\gamma}$.

Ved forsøget fandtes følgende resultater:

$v \text{ cm}^3$	100	150	200	250	300	350	400	450	500	550	600
$p \text{ kp/cm}^2$	29.58	15.42	11.67	7.48	7.29	3.90	3.63	1.69	2.95	2.16	2.11

- 1) Begrund, at formodningen er rimelig
- 2) Angiv ligningen for den fundne model

Løsning:

- 1) Idet funktionen er en potensfunktion, tages logaritmen på begge sider (se evt. oversigt 13.3 punkt 7).

$$\text{Vi får } P = c \cdot v^{-\gamma} \Leftrightarrow \ln P = \ln c - \gamma \ln v \Leftrightarrow y = \ln c - \gamma \cdot x$$

$$\text{hvor } y = \ln P \text{ og } x = \ln v$$

Ved at tage logaritmen fås nu tabellen:

$x = \ln v$	4.605	5.010	5.298	5.522	5.703	5.859	5.992	6.109	6.215	6.310	6.397
$y = \ln P$	3.387	2.736	2.457	2.012	1.987	1.361	1.289	0.525	1.082	0.770	0.747

De 11 punktpar indtastes i lommeregner.

13. Regressionsanalyse

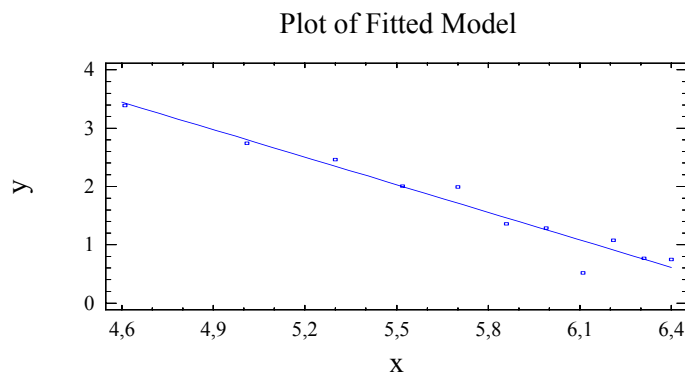
Regressionsprogram aktiveres, og blandt beregnede størrelser findes:

Estimer for regressionskoefficienter: $\hat{\beta}_0 = \ln(c) = 10.6555$ og $\hat{\beta}_1 = -v = -1.5689$

og korrelationskoefficient $r = -0.9734$

Forklaringsgraden er $r^2 = 0.97219^2 = 0.9474$.

Indtegnes punkterne og regressionslinien i lommeregnerens grafiske display fås følgende figur, som viser, at punkterne på figuren fordeler sig tilfældigt omkring linien og der ikke er enkelte punkter (outliers) der afviger voldsomt fra linien. Da forklaringsgraden samtidig er tæt på 1, så er den lineære model acceptabel.



Figur 13.6 $r = -0.973$ $r^2 = 94.63$

2) Den estimerede regressionsligning bliver

$$\ln(\hat{P}) = 10.6555 - 1.5689 \cdot \ln(v) \Leftrightarrow \hat{P} = e^{10.6555} \cdot v^{-1.5689} \Leftrightarrow \underline{\underline{\hat{P} = 42425.38 \cdot v^{-1.5689}}}$$

13.3 Polynomial regressionsanalyse

13.3.1 Indledning

Ved en polynomial regressionsanalyse er den statistiske model

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \dots + \beta_p x^p.$$

hvor den variable Y skal opfylde de sædvanlige regressionsforudsætninger.

Som det ses, er den i afsnit 13.2 betragtede enkelte regression et specialtilfælde. Den statistiske analyse da også meget beslægtet hermed.

Det man søger er altid den "enkleste" model der giver en tilstrækkelig god beskrivelse af Y indenfor det foreliggende variationsområde for x . 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. At andet lige vil man derfor foretrække de i forrige afsnit nævnte "transformerede" modeller som alle kun har 2 parametre fremfor eksempelvis et andengradspolynomium $Y = \beta_0 + \beta_1 x + \beta_2 x^2$ hvor man skal bestemme tre parametre β_0 , β_1 og β_2 . Blandt polynomierne vil man naturligvis foretrække et af lavest grad.

13.3.2. Forklaring af metoder ved beregninger

Fremgangsmåden er meget analog med den for enkelt regressionsanalyse. Et specielt problem er imidlertid at finde frem til det polynomium af lavest mulig grad, der giver en tilstrækkelig god beskrivelse af Y indenfor det foreliggende variationsområde for x .

Fremgangsmåden afhænger af, om der til hver x - værdi svarer én eller flere y -værdier.

Lad os derfor betragte to eksempler herpå

Eksempel 13.8 (polynomial regression 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

Eksempel 13.9 (polynomial regression med gentagelser)

Samme problem som i eksempel 13.6, men man fik nu følgende forsøgsresultater:

x g/l	0	1	2	3	4	5	6	7	8
T minutter	740	610	470	440	400	440	530	420	450
	710	650	540	420	450	480	470	480	490

I eksempel 13.8 vil et 17-gradspolynomium gå eksakt gennem de 18 punkter, og $r^2 = 100\%$. Tilsvarende vil i eksempel 13.9 et 8-gradspolynomium gå eksakt gennem de 9 "gennemsnitspunkter". Det er imidlertid klart, at sådanne modeller dels er alt for komplicerede til de fleste praktiske formål, dels følger kurven alle de tilfældige variationer, som vi netop ikke bør tage hensyn til. I stedet vil man i begge eksempler starte med det simpleste polynomium $Y = \beta_0 + \beta_1 x + \beta_2 x^2$ (idet vi antager vi allerede har forkastet en førstegradsmodel).

I eksempel 13.8 vil man beregne forklaringsgraden r^2 , mens man i eksempel 13.9 vil lave en "lack of fit" test.

Lad P - værdien ved "lack of fit" testen være $P_{\text{lack of fit}}$.

Hvis $P_{\text{lack of fit}} > \alpha$ og et residualplot viser, at residualerne ligger tilfældigt omkring 0-linien, og der ikke synes at være outliers, så accepteres modellen. I modsat fald vil vi gå en grad op til en trediegradsmodel, og så igen lave en "lack of fit" test, osv.

I eksempel 13.8 har man ikke tilsvarende et mål for, hvornår r^2 er stor nok.

Det er altid muligt at øge r^2 ved at addere flere led til modellen. 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))

Når r^2 ikke stiger væsentligt, og R-squared (adjusted for d.f.) begynder at falde, er man tæt ved den "bedste" model. Man ser så yderligere på et residualplot. Muligvis kan man uden skade reducere graden lidt.

¹ r^2 (adjusted) = $\frac{(n-1) \cdot r^2 + 1 - k}{n - k} = \frac{s_{\text{total}}^2 - s_0^2}{s_0^2}$, hvor k er antal parametre i modellen (incl konstantled).

Beregningerne er så komplicerede, at man er nødt til at foretage dem ved hjælp af matricer.

Lad os eksempelvis antage at vi har fundet en trediegradsmodel $Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$

Denne omskrives til $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ ved at sætte $x_1 = x$, $x_2 = x^2$, $x_3 = x^3$. Derefter foretage beregningerne som en multipel regression i 3 variable (kan ses i næste afsnit).

Det er naturligvis betydeligt nemmere at benytte et færdigt statistikprogram. I afsnit 13.A og 13B er eksemplerne 13.8 og 13.9 derfor løst ved anvendelse af henholdsvis TI - 89 og Statgraphics.

13.4 Multipel regressionsanalyse.

13.4.1 Indledning

Vi vil i dette afsnit behandle det tilfælde, hvor der indgår mere end 1 kvantitativ variabel. Vi vil begrænse os til at se på modeller, hvor de variable indgår lineært. Et eksempel herpå er modellen $Y = \alpha_0 + \beta_1 x_1 + \beta_2 x_2$, hvor parametrene er α_0, β_1 og β_2 .

Da beregningerne er meget omfattende er man nødt til at foretage dem ved hjælp af matricer. I oversigt 13.4 findes de nødvendige matrixformler. Det er dog langt lettere at have et statistikprogram til rådighed, så i afsnit 13A og 13B er eksemplerne regnet med henholdsvis TI - 89 og Statgraphics.

Den statistiske analyse af om modellen er acceptabel afhænger som før af, om der er gentagelser eller ej.

13.4.2 Multipel regressionsanalyse med én y - værdi for hver x - værdi.

Som beskrevet i de tidligere afsnit vurderer man om modellen er acceptabel ved

- 1) at se på et residualplot. Residualerne bør ligge "tilfældigt" omkring 0. Hvis dette ikke er tilfældet, bør man være på vagt, og eventuelt gå op til en model hvor de variable indgår af anden eller højere grad.
- 2) at se på "forklaringsgraden" r^2

Vi vil illustrere metoden ved følgende eksempel.

Eksempel 13.10 (multipel regressionsanalyse 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	22	91	100	836
-1	20	90	95	789
7	21	88	110	883
16	19	87	88	790
18	20	91	94	816
23	19	94	99	859
27	23	87	97	831
29	21	86	96	832
24	22	88	110	897
16	23	91	105	872
10	20	90	100	842
3	20	89	98	821

- 1) Vurder ud fra forklaringsgraden 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.
- 4) Angiv 95% konfidensintervaller for de regressionskoefficienter der indgår i ovenstående model.
- 5) Angiv et 95% konfidensinterval for Y i punktet $(x_1, x_2, x_3, x_4) = (0, 20, 90, 100)$

Løsning:

Vi vil her anvende de i oversigt 13.4 angivne matrixformler til løsningen og foretage matrixberegningerne ved hjælp af TI - 89.

Som det fremgår af løsningen er regningerne selv med matricer temmelig arbejdskrævende, så det må anbefales anvende et færdigt statistikprogram. I afsnit 13A og afsnit 13B er samme eksempel således regnet med henholdsvis TI-89 og Statgraphics.

- 1) Indsættes de 12 "punkter" i ligning $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$ fås 12 ligninger

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4}, \quad i = 1, 2, \dots, 12.$$

til bestemmelse af de 5 ubekendte $\beta_0, \beta_1, \beta_2, \beta_3$ og β_4 .

Disse 12 ligninger kan i matrixnotation skrives $\bar{y} = X \cdot \hat{\beta}$

13. Regressionsanalyse

$$\text{hvor } \bar{y} = \begin{bmatrix} 836 \\ 789 \\ 883 \\ 790 \\ 816 \\ 859 \\ 831 \\ 832 \\ 897 \\ 972 \\ 842 \\ 821 \end{bmatrix}, X = \begin{bmatrix} 1 & -4 & 22 & 91 & 100 \\ 1 & -1 & 20 & 90 & 95 \\ 1 & 7 & 21 & 88 & 110 \\ 1 & 16 & 19 & 87 & 88 \\ 1 & 18 & 20 & 91 & 94 \\ 1 & 23 & 19 & 94 & 99 \\ 1 & 27 & 23 & 87 & 97 \\ 1 & 29 & 21 & 86 & 96 \\ 1 & 24 & 22 & 88 & 110 \\ 1 & 16 & 23 & 91 & 105 \\ 1 & 10 & 20 & 90 & 100 \\ 1 & 3 & 20 & 89 & 98 \end{bmatrix}, \hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \\ \hat{\beta}_4 \end{bmatrix}$$

Løsningen til det overbestemte normalligningssystemet er $\hat{\beta} = (X^T \cdot X)^{-1} \cdot X^T \cdot \bar{y}$

Ti-89 - ordrer.

Matricerne X og $\bar{y}$ indtastes:

APPS, Data/Matrix editor, New, Udfyld Type = Matrix, Variable = x, antal rækker=12 og søjler = 5, ENTER, ENTER. Udfyld skemaet med matricen x, Home

Nu er matricen X indtastet.

APPS, Data/Matrix editor, New, Udfyld Type = Matrix, Variable = y, antal rækker=12 og søjler = 1, ENTER. Udfyld skemaet med matricen y, Home

Nu er også matricen y indtastet

Det antages, at matricerne er gemt i "current Folder" så der skrives kun x og ikke VAR-link, x (x, MATH, 4: MATRIX, ENTER, 1:T, * x) ^ -1 * (x, MATH, 4: MATRIX, ENTER, 1:T, * y) ENTER, STO → a

$$\text{Resultat: } \begin{bmatrix} 175.499 \\ 1.02664 \\ -0.793015 \\ 1.91131 \\ 14.98226 \end{bmatrix} \text{ som er blevet gemt i matricen a.}$$

Heraf fås $\hat{\beta}_0 = 175.499$, $\hat{\beta}_1 = 1.02664$, $\hat{\beta}_2 = -0.793015$, $\hat{\beta}_3 = 1.91131$, $\hat{\beta}_4 = 14.98226$

Da vi ikke har gentagelser, ser man på forklaringsgraden $r^2 = \frac{SAK_{\text{model}}}{SAK_{\text{total}}}$.

Da $SAK_{\text{total}} = SAK_{\text{model}} + SAK_{\text{residual}}$ fås $r^2 = 1 - \frac{SAK_{\text{residual}}}{SAK_{\text{total}}}$

$SAK_{\text{total}} = (12 - 1) \cdot s_{\text{total}}^2 = 11 \cdot \text{Variance}(y) = 12674$ (vælg Catalog, Variance)

Ifølge oversigt 13.4 er $SAK_{\text{residual}} = (\bar{y} - X\hat{\beta})^T \cdot (\bar{y} - X\hat{\beta}) = (y - x \cdot a)^T \cdot (y - x \cdot a)$

(y,-, x,*, a) MATH, 4: MATRIX, 1:T, ENTER,*, (y,-, x,*, a)

resultat 437.906

$$r^2 = 1 - \frac{SAK_{\text{residual}}}{SAK_{\text{total}}} = 1 - \frac{437.906}{12674} = \underline{\underline{0.9654}}$$

Da forklaringsgraden er tæt på 1 antages modellen at være tilfredsstillende.

- 2) Da vi nu mener, at modellen er rimelig, udfyldes som sædvanlig en variansanalysetabel:

Variation (Source)	SAK (SS)	f (df)	$s^2 = \frac{\text{SAK}}{f}$	$F = \frac{s_{\text{model}}^2}{s_{\text{residual}}^2}$
Model	12236.1	4	3059.02	48.899
Residual	437.906	7	62.558	
Total	1152,182	11		

Vi ønsker at teste $H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$

P -værdi = $P(F > 48,899) = \text{Fcdf}(48,899, \infty, 4, 7) = 0.000034$

Da P -værdi = $0.000034 < 0.001$ forkastes H_0 (stærkt), dvs. vi ved, at mindst én af regressionskoefficienterne er forskellig fra nul.

For at kunne undersøge om nogle af koefficienterne kan være 0, beregnes først kovariansmatricen

$\sigma^2 \cdot (X^T \cdot X)^{-1}$, idet vi erstatter σ^2 med sit estimat $s_{\text{residual}}^2 = 62.558$

(x, MATH, 4: MATRIX, 1: T, ENTER, *, x) ^ (-1) * 62.558

Af diagonalelementerne fås

$V(\hat{\beta}_0) = 12964.7$, $V(\hat{\beta}_1) = 0.051447$, $V(\hat{\beta}_2) = 4.49561$, $V(\hat{\beta}_3) = 1.31311$, $V(\hat{\beta}_4) = 0.20161$

Lad $s(\hat{\beta}_i) = \sqrt{V(\hat{\beta}_i)}$ (kaldet "standard error" for $\hat{\beta}_i$)

Nulhypotese $H_0: \beta_i = 0$ mod den alternative hypotese $H: \beta_i \neq 0$

Teststørrelsen $t_i = \frac{\hat{\beta}_i}{s(\hat{\beta}_i)}$, er t -fordelt med $12 - 5 = 7$ frihedsgrader

Af værdierne for $\hat{\beta}_i$ og $s(\hat{\beta}_i) = \sqrt{V(\hat{\beta}_i)}$ ses umiddelbart, at den numerisk mindste t -værdi må være t_2 eller t_3 .

Vi får: $t_2 = \frac{-0.793015}{\sqrt{4.49561}} = -0.37401$, $t_3 = \frac{1.91131}{\sqrt{1.31311}} = 1.6679$

Nulhypotese $H_0: \beta_2 = 0$ mod den alternative hypotese $H: \beta_2 \neq 0$

Da P -værdi = $P(T > 0.37401) = \text{tCdf}(0.37401, \infty, 7) = 0.3597 > 0.025$ accepteres H_0 ,

dvs. vi kan bortkaste x_2

Da regressionskoefficienterne afhænger af hinanden, skal man bortkaste én koefficient ad gangen (dvs. ikke på én gang bortkaste alle regressionskoefficienter, der har P -værdier større end 0).

Vi betragter nu $Y = \beta_0 + \beta_1 x_1 + \beta_3 x_3 + \beta_4 x_4$

$$\bar{y} = X \cdot \hat{\beta} \quad \text{hvor} \quad \bar{y} = \begin{bmatrix} 836 \\ 789 \\ 883 \\ 790 \\ 816 \\ 859 \\ 831 \\ 832 \\ 897 \\ 972 \\ 842 \\ 821 \end{bmatrix}, \quad X = \begin{bmatrix} 1 & -4 & 91 & 100 \\ 1 & -1 & 90 & 95 \\ 1 & 7 & 88 & 110 \\ 1 & 16 & 87 & 88 \\ 1 & 18 & 91 & 94 \\ 1 & 23 & 94 & 99 \\ 1 & 27 & 87 & 97 \\ 1 & 29 & 86 & 96 \\ 1 & 24 & 88 & 110 \\ 1 & 16 & 91 & 105 \\ 1 & 10 & 90 & 100 \\ 1 & 3 & 89 & 98 \end{bmatrix}, \quad \hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_3 \\ \hat{\beta}_4 \end{bmatrix}$$

13. Regressionsanalyse

APPS, Data/Matrix editor, Current, sæt cursor på 3 søjle, F6, 2: Delete, 3: column, ENTER

Vi beregner igen $\hat{\beta} = (X^T \cdot X)^{-1} \cdot X^T \cdot \bar{y}$

(x, MATH, 4: MATRIX, ENTER, 1:T, * x) ^ -1 * (x, MATH, 4: MATRIX, ENTER, 1:T, * y) ENTER, STO
→ a

Resultat: $\begin{bmatrix} 157.317 \\ 1.01695 \\ 2.03594 \\ 4.88827 \end{bmatrix}$ som er blevet gemt i matricen a.

Heraf fås $\hat{\beta}_0 = 157.317$, $\hat{\beta}_1 = 1.01695$, $\hat{\beta}_3 = 2.03594$, $\hat{\beta}_4 = 4.88827$

$SAK_{\text{residual}} = (y - x \cdot a)^T \cdot (y - x \cdot a)$

(y, -, x, *, a) MATH, 4: MATRIX, 1:T, ENTER, *, (y, -, x, *, a)

resultat 446.657

$s_{\text{residual}}^2 = \frac{446.657}{12 - 4} = 55.8321$

Kovariansmatricen $\sigma^2 \cdot (X^T \cdot X)^{-1}$

(x, MATH, 4: MATRIX, 1:T, ENTER, *, x) ^ (-1) * 55.8321

Af diagonalelementerne fås

$V(\hat{\beta}_0) = 9461.59$, $V(\hat{\beta}_1) = 0.045316$, $V(\hat{\beta}_3) = 1.07284$, $V(\hat{\beta}_4) = 0.123568$

Det ses umiddelbart, at t - værdien for β_3 er den numerisk største. Vi får: $t_3 = \frac{2.03594}{\sqrt{1.07284}} = 1.96561$

Nulhypotese $H_0: \beta_3 = 0$ mod den alternative hypotese $H: \beta_3 \neq 0$

Da P - værdi = $P(T > 1.96561) = t\text{Cdf}(1.96561, \infty, 8) = 0.042459 > 0.025$ accepteres H_0 , dvs vi kan bortkaste x_3

Vi betragter nu $Y = \beta_0 + \beta_1 x_1 + \beta_4 x_4$

$\bar{y} = X \cdot \hat{\beta}$

hvor $\bar{y} = \begin{bmatrix} 836 \\ 789 \\ 883 \\ 790 \\ 816 \\ 859 \\ 831 \\ 832 \\ 897 \\ 972 \\ 842 \\ 821 \end{bmatrix}$, $X = \begin{bmatrix} 1 & -4 & 100 \\ 1 & -1 & 95 \\ 1 & 7 & 110 \\ 1 & 16 & 88 \\ 1 & 18 & 94 \\ 1 & 23 & 99 \\ 1 & 27 & 97 \\ 1 & 29 & 96 \\ 1 & 24 & 110 \\ 1 & 16 & 105 \\ 1 & 10 & 100 \\ 1 & 3 & 98 \end{bmatrix}$, $\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_4 \end{bmatrix}$

APPS, Data/Matrix editor, Current, sæt cursor på 3 søjle, F6, 2: Delete, 3: column, ENTER

Vi beregner igen $\hat{\beta} = (X^T \cdot X)^{-1} \cdot X^T \cdot \bar{y}$

(x, MATH, 4: MATRIX, ENTER, 1:T, * x) ^ -1 * (x, MATH, 4: MATRIX, ENTER, 1:T, * y) ENTER, STO
→ a

Resultat: $\begin{bmatrix} 335.652 \\ 0.901697 \\ 4.94017 \end{bmatrix}$ som er blevet gemt i matricen a.

Heraf fås $\hat{\beta}_0 = 335.652$, $\hat{\beta}_1 = 0.901697$, $\hat{\beta}_4 = 4.94017$

$$SAK_{\text{residual}} = (y - x \cdot a)^T \cdot (y - x \cdot a)$$

(y,-, x,*, a) MATH, 4:MATRIX, 1:T, ENTER,*,(y,-, x,*, a)

resultat 662.372

$$s_{\text{residual}}^2 = \frac{662.372}{12-3} = 73.5968$$

Kovariansmatricen $\sigma^2 \cdot (X^T \cdot X)^{-1}$ (x, MATH, 4:MATRIX, 1:T, ENTER,*,x)^(-1)*73.5968

Af diagonalelementerne fås $V(\hat{\beta}_0) = 1621.51$, $V(\hat{\beta}_1) = 0.055203$, $V(\hat{\beta}_4) = 0.161966$

$$\text{Vi får: } t_1 = \frac{0.901697}{\sqrt{0.055203}} = 3.83777, \quad t_4 = \frac{4.94017}{\sqrt{0.161966}} = 12.2752$$

Nulhypotese $H_0: \beta_1 = 0$ mod den alternative hypotese $H: \beta_1 \neq 0$

Da P -værdi $= P(T > 3.83777) = \text{tCdf}(3.83777, \infty, 9) = 0.00199 < 0.025$ forkastes H_0 .

Modellen kan ikke reduceres mere.

3) Regressionsligningen bliver $\underline{Y = 335.652 + 0.901697x_1 + 4.94017x_4}$

4) $100(1 - \alpha)\%$ konfidensinterval for β_i

$$\hat{\beta}_1 \pm t_{1-\frac{\alpha}{2}}(n-p) \cdot s(\hat{\beta}_1) = 0.901697 \pm t_{0.975}(12-3) \cdot \sqrt{0.055203} = 0.901698 \pm 0.5315 \quad \underline{\underline{[0.3702; 1.4332]}}$$

5) 95% konfidensinterval for Y i punktet $(x_1, x_2, x_3, x_4) = (0, 20, 90, 100)$

$$\text{Lad } \bar{x}_0 = \begin{bmatrix} 1 \\ 0 \\ 100 \end{bmatrix} \quad \hat{y}_0 = 335.652 + 0.901697 \cdot 0 + 4.94017 \cdot 100 = 829.67$$

Idet $\bar{x}_0^T \cdot (X^T \cdot X)^{-1} \cdot \bar{x}_0 = 0.23074$ fås

$$r_{\text{kon}} = t_{1-\frac{\alpha}{2}}(n-p) \cdot \sqrt{s_{\text{residual}}^2 \cdot \bar{x}_0^T \cdot (X^T \cdot X)^{-1} \cdot \bar{x}_0} = t_{0.975}(9) \cdot \sqrt{73.5968 \cdot 0.23074} = 9.3221$$

$$[829.67 - 9.3221; 829.67 + 9.3221] = \underline{\underline{[820.34; 829.99]}}$$



13.4.3 Multipel regressionsanalyse med flere y -værdier for hver x -værdi.

Det vil sædvanligvis være klogere at måle ‘en gang i hvert målepunkt end eksempelvis at måle 2 gange i det halve antal målepunkter. Man vil derfor sjældent være ude for at skulle foretage en “lack of fit” test når man har mange variable, og eksempelvis Statgraphics har da heller ikke dette indbygget ved “multiple Regression”

Beregningerne er ganske de samme som ved enkelt regression. Man beregner et estimat s_e^2 for støjens varians, beregner SAK_{residual} og har så $SAK_{\text{lack of fit}} = SAK_{\text{residual}} - SAK_e$. Metoden vises i afsnit 13A (TI-89) og 13B (Statgraphics)

OVERSIGT 13.1. Formler til beregning af enkelt regressionsanalyse uden gentagelser.

I denne oversigt vises hvorledes man kan beregne en enkelt regressionsanalyse uden gentagelser, blot man har en lommeregner med regressionsprogram. I eksempel 13.5 er formlerne anvendt på et konkret eksempel.

Forudsætning:

Data :

x	x_1	x_2	x_3			x_N
y	y_1	y_2	y_3	...		y_N

De N - værdier er uafhængige observationer af statistisk uafhængig normalfordelte variable Y_i med samme varians σ^2 .

Det antages endvidere at man har fundet, at data kan beskrives ved en lineær model.

Vi har derfor at **middelværdien** af den statistiske variable Y er en lineær funktion af x af formen $E(Y|x) = \beta_0 + \beta_1 x$

Beregninger:

1) De N punktpar indtastes i lommeregner.

Regressionsprogram aktiveres, og blandt beregnede størrelser findes estimater for

regressionskoefficienter: $\hat{\beta}_0$ og $\hat{\beta}_1$, korrelationskoefficient r , gennemsnit $\bar{x}$, spredning s_y .

2) Udfylder variansanalysetabel:

Udregner $SAK_{\text{total}} = (N - 1) \cdot s_y^2$, $SAK_{\text{model}} = r^2 \cdot SAK_{\text{total}}$ og

$$SAK_{\text{residual}} = SAK_{\text{total}} - SAK_{\text{model}}.$$

Variation (Source)	SAK (SS)	f (df)	$s^2 = \frac{SAK}{f}$	F
Model	SAK_{model}	1	$s_{\text{model}}^2 = \frac{SAK_{\text{model}}}{1}$	$F_{\text{model}} = \frac{s_{\text{model}}^2}{s_{\text{residual}}^2}$
Residual	SAK_{residual}	$N - 2$	$s_{\text{residual}}^2 = \frac{SAK_{\text{residual}}}{N - 2}$	
Total	SAK_{total}	$N - 1$		

Test:

Lad α være signifikansniveau.

1) H_0 : Regressionslinien er vandret $\Leftrightarrow H_0$: y er uafhængig af $x \Leftrightarrow H_0$: Model = 0 $\Leftrightarrow H_0$: $\beta_1 = 0$

Metode 1. H_0 forkastes, hvis P -værdi = $P(Z < F_{\text{model}}) < \alpha$, hvor Z er F -fordelt

$$(f_T, f_N) = (1, N - 2).$$

Metode 2. Lad $t = \frac{\hat{\beta}_1}{s_{\beta_1}}$, hvor $s_{\beta_1} = \frac{|\hat{\beta}_1| \cdot s_{\text{residual}}}{s_{\text{model}}}$ er et estimat for spredningen på β_1 .

Det kan vises, at t er t -fordelt med $N - 2$ frihedsgrader.

Lad T være t -fordelt med $N - 2$ frihedsgrader

H_0 forkastes, hvis P -værdi = $P(T > |t|) < \frac{\alpha}{2}$.

En fordel ved denne metode er, at man også kan teste H_0 : $\beta_1 > 0$ og H_0 : $\beta_1 < 0$ ved ensidede test.

Hvis begge variable X og Y er statistiske variable kan man tilsvarende teste korrelationen ρ ved ovennævnte t -test.

2) H_0 : $\beta_1 = a$, hvor a er en given konstant.

Lad $t = \frac{\hat{\beta}_1 - a}{s_{\beta_1}}$, hvor $s_{\beta_1} = \frac{|\hat{\beta}_1| \cdot s_{\text{residual}}}{s_{\text{model}}}$

H_0 forkastes, hvis $|t| > t_{1-\frac{\alpha}{2}}(N-2)$ (for $a = 0$ svarer det til ovennævnte metode 2).

Konfidensinterval for β_1 : $\left[\hat{\beta}_1 - t_{1-\frac{\alpha}{2}}(N-2) \cdot s_{\hat{\beta}_1}, \hat{\beta}_1 + t_{1-\frac{\alpha}{2}}(N-2) \cdot s_{\hat{\beta}_1} \right]$ hvor

$$s_{\beta_1} = \frac{|\hat{\beta}_1| \cdot s_{\text{residual}}}{s_{\text{model}}}.$$

Lad $\hat{\mu} = E(Y|x = x_0)$ være et estimat for middelværdien μ for Y for en given værdi $x = x_0$.

Konfidensinterval for μ $\left[\hat{\mu} - t_{1-\frac{\alpha}{2}}(N-2) \cdot \sqrt{\hat{V}(\mu)}; \hat{\mu} + t_{1-\frac{\alpha}{2}}(N-2) \cdot \sqrt{\hat{V}(\mu)} \right]$

hvor $\hat{\mu} = \hat{\beta}_0 + \hat{\beta}_1 x_0$, $\hat{V}(\mu) = s_{\text{residual}}^2 \cdot \left(\frac{1}{N} + \frac{(x_0 - \bar{x})^2 (\hat{\beta}_1)^2}{SAK_{\text{model}}} \right)$

Prædistributionsinterval: (Konfidensinterval) for 1 ny observation for en given x -værdi:

$\left[\hat{\mu} - t_{1-\frac{\alpha}{2}}(N-2) \cdot \sqrt{\hat{Q}}; \hat{\mu} + t_{1-\frac{\alpha}{2}}(N-2) \cdot \sqrt{\hat{Q}} \right]$, hvor $\hat{Q} = s_{\text{residual}}^2 \cdot \left(1 + \frac{1}{N} + \frac{(x - \bar{x})^2 (\hat{\beta}_1)^2}{SAK_{\text{model}}} \right)$

OVERSIGT 13.2. Formler til beregning af enkelt regressionsanalyse med lige mange gentagelser

Forudsætning: Data :

x	x_1	x_2	x_3			x_k
y	y_{11}	y_{21}	y_{31}	...		y_{k1}
	y_{12}	y_{22}	y_{32}			y_{k2}
	...	...	...			...
	y_{1n}	y_{2n}	y_{3n}			y_{kn}

y_{ij} - værdierne er uafhængige observationer af statistisk uafhængig normalfordelte variable Y_i . For hver af de k x - værdier er der lige mange gentagelser n af y - værdier, dvs. i alt $N=n \cdot k$ observationer.

Der antages, at der er varianshomogenitet (ønskes dette testet se under punkt b)

Lad α være signifikansniveau.

Beregninger:

a) "Lack of fit test:

H_0 : Lineær model gælder $\Leftrightarrow H_0: (x_i, \mu_i)$ ligger på en ret linie

$\Leftrightarrow H_0$: Residual for gennemsnitspunkter = 0.

- 1) For hver x - værdi x_i indtastes de n y -værdier, og man beregner spredningen s_i .

Der beregnes et estimat for den fælles varians $s_e^2 = \frac{s_1^2 + s_2^2 + \dots + s_k^2}{k}$.

s_e^2 har $k(n-1) = N - k$ frihedsgrader.

- 2) De $N=n \cdot k$ punktpar (x_i, y_{ij}) indtastes i lommeregner.

Regressionsprogram aktiveres, og blandt beregnede størrelser findes estimater for:

regressionskoefficienter: $\hat{\beta}_0$ og $\hat{\beta}_1$, korrelationskoefficient r , gennemsnit $\bar{x}$, spredning s_y .

- 3) Man udregner $SAK_{\text{total}} = (N-1) \cdot s_y^2$, $SAK_{\text{model}} = r^2 \cdot SAK_{\text{total}}$ og $SAK_e = (N-k) \cdot s_e^2$

$$SAK_{\text{lack of fit}} = SAK_{\text{total}} - SAK_{\text{model}} - SAK_e$$

- 4) Udfylder variansanalysetabel:

Variation (Source)	SAK (SS)	f (df)	$s^2 = \frac{SAK}{f}$	F
Model	SAK_{model}	1	$s_{\text{model}}^2 = \frac{SAK_{\text{model}}}{1}$	$\left(F_{\text{model}} = \frac{s_{\text{model}}^2}{s_{\text{residual}}^2} \right)$
Lack of fit	$SAK_{\text{lack of fit}}$	$k - 2$	$s_{\text{lack of fit}}^2 = \frac{SAK_{\text{lack of fit}}}{k - 2}$	$F_{\text{lack of fit}} = \frac{s_{\text{lack of fit}}^2}{s_0^2}$
Gentagelser (error)	SAK_e	$N - k$	$s_e^2 = \frac{SAK_e}{N - k}$	
Total	SAK_{total}	$N - 1$		

- 5) H_0 forkastes, hvis P -værdi $= P(Z > F_{\text{lack of fit}}) < \alpha$, hvor Z er F -fordelt
 $(f_T, f_N) = (k - 2, N - k)$.

Såfremt H_0 accepteres (og et residualplot også virker rimelig) fortsætter testningen:

Da såvel s_e^2 som $s_{\text{lack of fit}}^2$ nu er et udtryk for forsøgsfejls varians σ^2 , foretages en pooling:

$$s_{\text{residual}}^2 = \frac{(N - k) \cdot s_e^2 + (k - 2) \cdot s_{\text{lack of fit}}^2}{N - 2} = \frac{SAK_e + SAK_{\text{lack of fit}}}{N - 2},$$

og F_{model} beregnes (se variansanalysetabel).

- 6) Formlerne for de forskellige test svarer nu fuldstændig til formlerne i afsnit 11.1.

b) Varianshomogenitet

$H_0: \sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2$. Test for at de variable Y_i har samme varians σ^2

- 1) **Simplificeret F-test.** Lad den største værdi af de estimerede k varianser være s_{max}^2 og den mindste være s_{min}^2 .

Beregn teststørrelsen $F = \frac{s_{\text{max}}^2}{s_{\text{min}}^2}$.

Lad Y være F -fordelt med frihedsgraderne $f_{\text{tæller}} = f_{\text{nævner}} = n - 1$

H_0 forkastes, hvis P -værdi $= P(Y > F) < \frac{\alpha}{2}$.

Hvis nulhypotesen accepteres, så antages kravet om varianshomogenitet at være opfyldt.

Hvis nulhypotesen forkastes, må anvendes en test med større styrke såsom Bartlett's test

- 2) **Bartlett's test.** Denne test er beregningsmæssigt vanskelig, og har den svaghed, at den er særdeles følsom overfor afvigelser fra normalitet.

$$\text{Beregn teststørrelsen } \chi^2 = \frac{(N - k) \cdot \ln \left(\frac{\sum_{i=1}^k (n_i - 1) \cdot s_i^2}{N - k} \right) - \sum (n_i - 1) \cdot \ln(s_i^2)}{1 - \frac{1}{N - k} - \sum_{i=1}^k \frac{1}{n_i - 1}}$$

Lad Y være χ^2 -fordelt med frihedsgrad $k - 1$. H_0 forkastes, hvis P -værdi $= P(Y > \chi^2) < \alpha$.

Oversigt 13.3. Transformation til lineær model i Statgraphics.

Nr	Model	Kommentar
1	Linear model: $Y = a + b \cdot X$	
2	Exponential model: $Y = \exp(a + b \cdot X)$	$Y = e^{a+bX} = e^a \cdot e^{bX} \Leftrightarrow \ln(Y) = a + bX$ Sættes $Z = \ln(Y)$ fås $Z = a + b \cdot X$
3	Reciprocal-Y model: $Y = 1/(a + b \cdot X)$	$Y = \frac{1}{a + b \cdot X} \Leftrightarrow \frac{1}{Y} = a + b \cdot X$ Sættes $Z = \frac{1}{Y}$ fås $Z = a + b \cdot X$
4	Reciprocal-X model: $Y = a + b/X$	$Y = a + \frac{b}{X}$. Sættes $W = \frac{1}{X}$ fås $Y = a + b \cdot W$
5	Double reciprocal model: $Y = 1/(a + b/X)$	$Y = \frac{1}{a + \frac{b}{X}} \Leftrightarrow \frac{1}{Y} = a + b \frac{1}{X}$ Sættes $Z = \frac{1}{Y}$ og $W = \frac{1}{X}$ fås $Z = a + b \cdot W$
6	Logarithmic-X model: $Y = a + b \cdot \ln(X)$	Sættes $W = \ln(X)$ fås $Y = a + b \cdot W$
7	Multiplicative model: $Y = a \cdot X^b$	$Y = a \cdot X^b \Leftrightarrow \ln(Y) = \ln(a) + b \cdot \ln(X)$ Sættes $Z = \ln(Y)$ og $W = \ln(X)$ fås $Z = a + b \cdot W$
8	Square root-X model: $Y = a + b \cdot \sqrt{X}$	$Y = a + b \cdot \sqrt{X}$. Sættes $W = \sqrt{X}$ fås $Y = a + b \cdot W$
9	Square root-Y model: $Y = (a + b \cdot X)^2$	$Y = (a + b \cdot X)^2 \Leftrightarrow \sqrt{Y} = a + b \cdot X$ Sættes $Z = \sqrt{Y}$ fås $Z = a + b \cdot X$
10	S-curve model: $Y = \exp(a + b/X)$	$Y = e^{a+\frac{b}{X}} \Leftrightarrow \ln(Y) = a + \frac{b}{X}$ Sættes $Z = \ln(Y)$ og $W = \frac{1}{X}$ fås $Z = a + b \cdot W$
11	Logistic model: $Y = \exp(a + b \cdot X)/(1 + \exp(a + b \cdot X))$	$Y = \frac{e^{a+bX}}{1 + e^{a+bX}} \Leftrightarrow \ln\left(\frac{1}{Y} + 1\right) = a + bX$ Sættes $Z = \ln\left(\frac{1}{Y} + 1\right)$ fås $Z = a + b \cdot X$
12	Log probit model: $Y = \text{normal}(a + b \cdot \ln(X))$	$Y = \Phi(a + b \cdot \ln(X)) \Leftrightarrow \Phi^{-1}(Y) = a + b \cdot \ln(X)$ Sættes $Z = \Phi^{-1}(Y)$ og $W = \ln(X)$ fås $Z = a + b \cdot W$

OVERSIGT 13.4. Formler til beregning af multipel regressionsanalyse.

I denne oversigt vises hvorledes man kan beregne en multipel regressionsanalyse, blot man har en "matematiklommeregner" med et matrixprogram.

I eksempel 13.8 er formlerne anvendt på et konkret eksempel.

Lad der være givet k uafhængige variable og N observationer

$(x_{i1}, x_{i2}, \dots, x_{ik}, y_i)$, $i = 1, 2, \dots, N$ og $N > k$.

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 $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k$, (1)

hvor $\beta_0, \beta_1, \beta_2, \dots, \beta_k$ er regressionskoefficienterne

Bestemmelse af estimater for regressionskoefficienterne

Modellen kan i matrixnotation skrives $\bar{y} = X \cdot \bar{\beta}$

$$\text{hvor } \bar{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdot & \cdot & \cdot & x_{1k} \\ 1 & x_{21} & x_{22} & \cdot & \cdot & \cdot & x_{2k} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & x_{N1} & x_{N2} & \cdot & \cdot & \cdot & x_{Nk} \end{bmatrix}, \quad \bar{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}.$$

Vi ønsker ved mindste kvadraters metode, at finde en vektor $\hat{\beta}$, der er et estimat for vektoren $\bar{\beta}$.

Løsningen til et sådant overbestemt ligningssystem $X \cdot \bar{\beta} = \bar{y}$ er (se eventuelt B. Hellesten, M. Oddershede Larsen: Bind III, kapitel 3) bestemt ved

$$X^T \cdot X \cdot \tilde{\beta} = X^T \cdot \bar{y} \quad (\text{kaldet } \mathbf{normalligningssystemet}). \quad (3)$$

Matricen $X^T \cdot X$ er en kvadratisk symmetrisk matrix, som sædvanligvis ved regressionsproblemer ikke er singulær. Der eksisterer derfor en invers matrix $(X^T \cdot X)^{-1}$, hvorved løsningen til normalligningssystemet (3) bliver

$$\underline{\underline{\hat{\beta} = (X^T \cdot X)^{-1} \cdot X^T \cdot \bar{y}}} \quad (4)$$

Herved er regressionskoefficienterne bestemt..

Beregning af variansanalysetabel.

Variation (Source)	SAK (SS)	f (df)	$s^2 = \frac{SAK}{f}$	F
Model	SAK_{model}	k	$s^2_{\text{model}} = \frac{SAK_{\text{model}}}{k}$	$F_{\text{model}} = \frac{s^2_{\text{model}}}{s^2_{\text{residual}}}$
Residual	SAK_{residual}	$N - k - 1$	$s^2_{\text{residual}} = \frac{SAK_{\text{residual}}}{n - k - 1}$	
Total	SAK_{total}	$N - 1$		

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. $r_i = y_i - \hat{y}_i$.

$$SAK_{\text{residual}} = \sum_{i=1}^n r_i^2.$$

Sættes $\bar{r} = \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_N \end{bmatrix}$ kan vi foretage følgende omskrivning

$$SAK_{\text{residual}} = \sum_{i=1}^n r_i^2 = \bar{r}^T \cdot \bar{r} = (\bar{y} - X\hat{\beta})^T \cdot (\bar{y} - X\hat{\beta}).$$

$$SAK_{\text{total}} = (n-1) \cdot s_y^2 \quad \text{hvor } s_y^2 \text{ er spredningen på } y \text{ - værdierne}$$

$$SAK_{\text{model}} = SAK_{\text{total}} - SAK_{\text{residual}}.$$

Vurdering af model

Har man ikke gentagelser kan man beregne forklaringsgraden $r^2 = \frac{SAK_{\text{model}}}{SAK_{\text{total}}}$

(se vurdering i eksempel 13.3)

Har man gentagelser kan man foretage en “lack of fit test” (se hvordan i eksempel 13.4)

Undersøgelse af om modellen kan reduceres.

1) $H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$ mod H : Mindst en af koefficienterne er forskellig fra 0.

Teststørrelse F_{model} er F - fordelt med tællerfrihedsgrad k og nævnerfrihedsgrad $N - k - 1$.

Hvis P - værdi $= P(F > F_{\text{model}}) < \alpha$ forkastes H_0 , dvs. y er ikke uafhængig af x - værdierne

- 2) Forkastes H_0 vil man dernæst undersøge om nogle af koefficienterne kunne være 0, dvs. teste nulhypoteserne $H_0: \beta_i = 0$ mod den alternative hypotese $H: \beta_i \neq 0$

Teststørrelsen er $t_i = \frac{\hat{\beta}_i}{s(\hat{\beta}_i)}$, som kan vises at være t -fordelt med $n - p$ frihedsgrader,

hvor p er antal regressionskoefficienter

H_0 forkastes, hvis P -værdi $= P(T > |t_i|) < \frac{\alpha}{2}$.

Beregningen af $s(\hat{\beta}_i)$ (kaldet "standard error" for $\hat{\beta}_i$) beregnes på følgende måde.

For den såkaldte kvadratiske symmetriske $p \times p$ kovariansmatrix $\sigma^2 \cdot (X^T \cdot X)^{-1}$ (p er antal regressionskoefficienter) gælder, at

a) diagonalelementerne er varianterne for regressionskoefficienterne og

b) elementerne udenfor diagonalen C_{ij} angiver kovariansen mellem β_i og β_j

Vi beregner derfor $\sigma_{\text{residual}}^2 \cdot (X^T \cdot X)^{-1}$, idet vi erstatter σ^2 med sit estimat s_{residual}^2

Af diagonalelementerne fås $V(\hat{\beta}_0)$, $V(\hat{\beta}_1)$, $V(\hat{\beta}_2)$, $V(\hat{\beta}_3)$ osv.

Konfidensintervaller.

$100(1 - \alpha)\%$ konfidensinterval for β_i .

$$\hat{\beta}_i - t_{1-\frac{\alpha}{2}}(N-p) \cdot s(\hat{\beta}_i) \leq \beta_i \leq \hat{\beta}_i + t_{1-\frac{\alpha}{2}}(N-p) \cdot s(\hat{\beta}_i)$$

Konfidensinterval for et til punkt $\bar{x}_0$ svarende værdi $\hat{y}_0$.

Lad $\bar{x}_0 = \begin{bmatrix} 1 \\ x_{01} \\ x_{02} \\ \vdots \\ x_{0k} \end{bmatrix}$

$$\left[\hat{y}_0 - t_{1-\frac{\alpha}{2}}(N-p) \cdot \sqrt{s_{\text{residual}}^2 \cdot \bar{x}_0^T \cdot (X^T \cdot X)^{-1} \cdot \bar{x}_0}; \hat{y}_0 + t_{1-\frac{\alpha}{2}}(N-p) \cdot \sqrt{s_{\text{residual}}^2 \cdot \bar{x}_0^T \cdot (X^T \cdot X)^{-1} \cdot \bar{x}_0} \right]$$

Forklaring på formlen

Forklaringen bygger for simpelhedens skyld på det enkle regressionspolynomium $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$.

Er $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ de estimerede værdier, og indsættes punktet (x_{01}, x_{02}) i ligningen, fås den dertil svarende estimerede y -værdi $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_{01} + \hat{\beta}_2 x_{02}$.

13. Regressionsanalyse

Ifølge reglerne for varians af en linearkombination fås

$$V(\hat{\beta}_0 + \hat{\beta}_1 x_{01} + \hat{\beta}_2 x_{02}) = V(\hat{\beta}_0) + x_{01}^2 \cdot V(\hat{\beta}_1) + x_{02}^2 \cdot V(\hat{\beta}_2) + 2x_{01} \cdot V(\hat{\beta}_0, \hat{\beta}_1) + 2x_{02} \cdot V(\hat{\beta}_0, \hat{\beta}_2) + 2x_{01} \cdot x_{02} \cdot V(\hat{\beta}_1, \hat{\beta}_2) .$$

Sættes $\bar{x}_0 = \begin{bmatrix} 1 \\ x_{01} \\ x_{02} \end{bmatrix}$ og idet kovariansmatricen

$$\text{er } \sigma^2 \cdot (X^T \cdot X)^{-1} = \begin{bmatrix} V(\hat{\beta}_0) & V(\hat{\beta}_0, \hat{\beta}_1) & V(\hat{\beta}_0, \hat{\beta}_2) \\ V(\hat{\beta}_1, \hat{\beta}_0) & V(\hat{\beta}_1) & V(\hat{\beta}_1, \hat{\beta}_2) \\ V(\hat{\beta}_2, \hat{\beta}_0) & V(\hat{\beta}_2, \hat{\beta}_1) & V(\hat{\beta}_2) \end{bmatrix} \text{ ses, at}$$

$$V(\hat{\beta}_0 + \hat{\beta}_1 x_{01} + \hat{\beta}_2 x_{02}) = \sigma^2 \cdot \begin{bmatrix} 1 & x_{01} & x_{02} \end{bmatrix} \cdot \begin{bmatrix} V(\hat{\beta}_0) & V(\hat{\beta}_0, \hat{\beta}_1) & V(\hat{\beta}_0, \hat{\beta}_2) \\ V(\hat{\beta}_1, \hat{\beta}_0) & V(\hat{\beta}_1) & V(\hat{\beta}_1, \hat{\beta}_2) \\ V(\hat{\beta}_2, \hat{\beta}_0) & V(\hat{\beta}_2, \hat{\beta}_1) & V(\hat{\beta}_2) \end{bmatrix} \cdot \begin{bmatrix} 1 \\ x_{01} \\ x_{02} \end{bmatrix}$$

$$\Leftrightarrow V(\hat{\beta}_0 + \hat{\beta}_1 x_{01} + \hat{\beta}_2 x_{02}) = \sigma^2 \cdot \bar{x}_0^T \cdot (X^T \cdot X)^{-1} \cdot \bar{x}_0 .$$

Konfidensintervallet bliver følgelig

$$\left[\hat{y}_0 - t_{1-\frac{\alpha}{2}}(n-p) \cdot \sqrt{s_{\text{residual}}^2 \cdot \bar{x}_0^T \cdot (X^T \cdot X)^{-1} \cdot \bar{x}_0} ; \hat{y}_0 + t_{1-\frac{\alpha}{2}}(n-p) \cdot \sqrt{s_{\text{residual}}^2 \cdot \bar{x}_0^T \cdot (X^T \cdot X)^{-1} \cdot \bar{x}_0} \right]$$

13A Eksempler regnet på TI - 89.

1. Indledning.

Det forudsættes, at man kender de grundlæggende operationer på lommeregneren.

I Appendix "Grundlæggende operationer på TI - 89" er beskrevet, hvorledes man beregner sandsynligheden for forskellige fordelinger, beregner gennemsnit og spredning, samt hvorledes man tester og beregner konfidensintervaller for funktion af 1 variabel. Dette forudsættes ligeledes bekendt.

2. Enkelt regressionsanalyse

2.1. Enkelt Regressionsanalyse uden gentagelser

Eksempel 13.4 (uden gentagelser)

Tilsætning af en vis mængde kunstfibre forøger et garns trækstyrke. Man har eksperimenteret med forskellige tilsatte mængder kunstfibre x og registreret garnets trækstyrke y ved disse forskellige mængder. Herved fremkom følgende observationsmateriale:

Mængde x (i gram) af kunstfibre pr kg uld	40	50	55	60	70	75	80	85	90	95	100	105	110	120	130
Trækstyrke : Y	4.5	6.5	5.4	7.0	8.2	8.0	7.1	8.9	8.2	10.3	9.6	10.8	10.5	11.2	12.0

- 1) Find r^2 og anvend denne samt en figur på lommeregnerens grafiske display til vurdering af modellen.
- 2) Opskriv regressionsligningen.
- 3) Test om y er uafhængig af x
- 4) Find 95% konfidensinterval for hældningen β
- 5) Find 95% konfidensinterval for y svarende til $x = 100$
- 6) Find 95% Prædiktionsinterval for 1 ny observation svarende til x - værdien 100.

Løsning:

- 1) APPS, STAT/LIST hvorefter data indtastes i list1(x - værdier) og list 2(y -værdier)

F4: Calc, 3. Regressions, 1:linReg(a+bx), Udfylder lister,

Da vi ønsker at tegne regressionslinien så StoreReqn to: $y1(x)$, ENTER,

Af udskriften fås umiddelbart $r^2 = 0.9193$

Man kan nu tegne linien ved at vælge "GRAPH".

Vi ønsker imidlertid punkterne tegnet med, så vi vælger

F2: Plots,, 1: Plot Setup, F1: Define, Behold Scatter og Box, indsæt listerne , ENTER, ENTER, F5

Linien vises sammen med punkterne.

Tegningen på lommeregnerens display viser, at punkterne fordeler sig tilfældigt omkring linien.

(Ønskes i stedet tegnet et residualplot, så findes residualerne i sidste kolonne.)

Outliers: Ingen punkter synes at afvige voldsomt fra linien

I sidste kolonne er residualerne beregnet, og den største residual er 1.1.

Da $s = s_{\text{residual}} = 0.648068$ er afvigelsen betydelig mindre end $2 \cdot s = 2 \cdot s_{\text{residual}}$ hvilket bekræfter, at der ikke er outliers.

Da forklaringsgraden samtidig er tæt på 1, er den lineære model acceptabel.

- 2) Regressionskoefficienterne ses i den ovennævnte udskrift eller ved at vælge "Y=" hvorefter man finder $y = 1.8087 + 0.0799x$
- 3) $H_0: Y$ er uafhængig af $x \Leftrightarrow H_0: \text{Regressionslinien er vandret} \Leftrightarrow H_0: \beta_1 = 0$.

APPS, STAT/LIST, F6, A:LinRegTest, Udfyld lister, Alternate Hyp= $\beta \& \rho \neq 0$, ENTER

Af udskriften ses, at P -værdi $= 1.7 \cdot 10^{-8}$ svarende til $t = 12.1693$.

Da P -værdi $= 1.76 \cdot 10^{-8} < 0.001$ forkastes $H_0: \beta = 0$ (stærkt)

Konklusion: Y er ikke uafhængig af x .

Andre udskrifter der kan være af interesse i andre sammenhænge:

$s = s_{\text{residual}} = 0.648068$, $df = f_{\text{residual}} = 13$, $\text{SE Slope} = s_{\tilde{\beta}_1} = 0.006565$

- 4) 95% konfidensinterval for β_1 :

F7: LinRegTInt: Udfyld menu:

Resultat: $[0.0657; 0.09413]$

Andre udskrifter der kan være af interesse i andre sammenhænge:

ME = radius i konfidensinterval for $\tilde{\beta}_1 = 0.014184$

- 5) Konfidensinterval for y svarende til $x = 100$:

Som under punkt 4) men Interval=Response, x Value = 100

Resultat: $[9.37; 10.22]$

Andre udskrifter der kan være af interesse i andre sammenhænge:

$y_{\text{hat}} = y$'s værdi for $x = 100 = 9.79839$,

ME = radius i konfidensintervallet $= 0.424332$, $\text{SE} = s_{\tilde{\mu}} = 0.677179$

- 6) 95% Prædiktionsinterval for 1 ny observation svarende til x -værdien 100.

Som under punkt 5. Se nederst i udskrift

Resultat: $[8.335; 11.263]$.

2.2. Enkelt Regressionsanalyse med gentagelser

Eksempel 13.6 Regressionsanalyse med gentagelser

Givet følgende målinger

Tiden t	20	30	40	60	70	90	100	120	150	180
Tykkelse	4.2	7.4	8.8	13.6	13.1	14.9	20.0	23.1	27.5	32.9
y	4.9	6.9	8.2	12.0	12.4	16.8	21.2	25.2	25.1	32.4

- 1) Foretag en testning af forudsætningen om varianshomogenitet.
- 2) Det formodes på forhånd, at der er en lineær sammenhæng mellem x og y .
Test denne formodning ved en "lack of fit" test, og bestem i bekræftende fald ligningen for den empiriske regressionslinie.
- 3) Det påstås at hældningskoefficienten β_1 er 0.15
Test om dette på et signifikansniveau på 5% kan være sandt.
- 4) Angiv et 95% konfidensinterval for middelværdien af tykkelsen y , når $t = 100$ minutter.

Løsning:

- 1) Test af nulhypotesen
- $H_0: \sigma_1^2 = \sigma_2^2 = \dots = \sigma_{10}^2$

TI-89 har intet "program" til testning af varianshomogenitet, så man må foretage en "simplificeret F-test ved at anvende formler.

APPS, STAT/LIST hvorefter data indtastes i list1(x- værdier) og list 2(y-værdier)

list1	list2
20	4.2
20	4.9
30	7.4
30	6.9
osv	osv.

Spredningerne beregnes i list3 ved HOME, CATALOG, Variance({4.2,4.9}) osv.

list1	list2	list3	Kommentar
20	4.2	0.245	Variance({4.2,4.9})
20	4.9	0.125	Variance({7.4,6.9})
30	7.4		osv.
30	6.9		
osv	osv.		

Forholdet $F = \frac{s_{\max}^2}{s_{\min}^2}$ beregnes:

HOME,MATH, list,max(list3)/MATH, list,min(list3), ENTER

Resultat: 23.04

P-værdi = FCdf(24.04,1,1)=0.1308

Da P - værdi = 0.1308 > 0.05, accepteres nulhypotesen

dvs. vi vil i det følgende antage, at der er varianshomogenitet.

- 2)
- H_0
- : Lineær model gælder
- $\Leftrightarrow H_0: (x_i, \mu_i)$
- ligger på en ret linie

Der udføres en "Lack of Fit" test.

- 1) Pooler de 10 spredninger sammen til et fælles
- s_e^2
- . Lad
- $A = \sum_{i=1}^{10} s_i^2$
- :

HOME,MATH, list, SUM, list3), ENTER, STO A

Resultat : 9.81

$$s_e^2 = \frac{9.81}{10} = 0.981, f_e = N - k = 20 - 10 = 10,$$

- 2) APPS, STAT/LIST, F6, A:LinRegTest,Udfyld lister, Alternate Hyp=
- $\beta \& \rho \neq 0$
- ,ENTER

Heraf: $s = s_{\text{residual}} = 1.3445$, $df = f_{\text{residual}} = 18$ og $\text{SAK}_{\text{residual}} = (1.3445)^2 \cdot 18 = 32.5382$

$$\text{SAK}_{\text{lack of fit}} = \text{SAK}_{\text{residual}} - \text{SAK}_e = 32.538 - 9.81 = 22.728$$

Vi kan nu udfylde skemaet

Variation	SAK	f	s^2	F
Lack of fit	22.728	8	2.841	2.896
Gentagelser (Pure Error)	9.81	10	0.981	
Residual	32.538	18	1.3445	

P - værdi = FCdf(2.896, ∞ , 8, 10) = 0.0591. Da P - værdi = 0.0591 > 0.05 accepteres H_0 ,

dvs. vi vil i det følgende antage, at den lineære model gælder.

Vi er dog så tæt på forkastelse, at en nærmere undersøgelse kan være rimelig.

3) $H_0: \beta_1 = 0.15$

Af udskriften fås $s_{\beta_1} = \text{SE SLOPE} = 0.0060$

$$t = \frac{\hat{\beta}_1 - a}{s_{\beta_1}} = \frac{0.1730 - 0.15}{0.0060} = 3.812 \text{ er } t - \text{fordelt med } f = f_{\text{residual}} = 18$$

Da P -værdi $= P(T > 3.81) = \text{tCdf}(3.81, \infty, 18) = 0.00064 < 0.025$ forkastes H_0 , dvs. data giver ikke den i litteraturen angivne hælningskoefficient.

4) 95% konfidensinterval for middelværdien af tykkelsen y , når $t = 100$ minutter.

F7: LinRegTInt: Udfyld menu: Interval=Response, x Value = 100

Resultat [18.3 ; 19.61]

$y_{\text{hat}} = y$'s værdi for $x = 100 = 18.95$.

2.3 Transformation af data

Eksempel 13.7. Transformation af udtryk.

Ved et forsøg komprimeres en luftart til forskellige forudvalgte rumfang v , idet de tilsvarende værdier af trykket P målt. Man formoder, at der gælder regressionsmodellen $P = c \cdot v^{-\gamma}$. Ved forsøget fandtes følgende resultater:

$v \text{ cm}^3$	100	150	200	250	300	350	400	450	500	550	600
$P \text{ kp/cm}^2$	29.58	15.42	11.67	7.48	7.29	3.90	3.63	1.69	2.95	2.16	2.11

1) Begrund, at formodningen er rimelig

2) Angiv ligningen for den fundne model

3) Find middelværdien af P og et 95% konfidensinterval svarende til $v = 375$.

Løsning:

1) APPS, STAT/LIST, Data indtastes i list1(v -værdier) og list 2(P -værdier)

F4: Calc, 3Regressions, 9:PowerReg, Udfylder lister,

Da vi ønsker at tegne regressionslinien så StoreReqn to: $y1(x)$, ENTER,

Af udskriften fås $r^2 = 0.9464$.

Grafen kan tegnes som under punkt 2.1.

På lommeregnerens display ses, at punkterne fordeler sig tilfældigt omkring grafen. Da forklaringsgraden samtidig er tæt på 1, så er den lineære model acceptabel.

2) Af udskriften fås $P = 4.266 \cdot 10^6 \cdot V^{-1.5696}$

3) Vi er nu nødt til at lave to lister svarende til $\ln(v)$ og $\ln(P)$

HOME, $\ln(\text{list1})$, STO, list3, ENTER

$\ln(\text{list2})$, STO, list4, ENTER

APPS, STAT/LIST, F7, 7: LinRegTint, Udfylder lister og sæt x Value = $\ln(375)$, ENTER

Finder $y_{\text{hat}} = 5.96293$ og CInt = [5.8 ; 6.126]

Heraf fås $P_{375} = e^{5.96292} = 389.75$ og 95% konfidensinterval: $[e^{5.8}; e^{6.126}] = [330.3; 457.6]$ [

3. Polynomial Regressionsanalyse

3.1. Polynomial Regressionsanalyse uden gentagelser

Eksempel 13.8. Polynomial regressionsanalyse uden gentagelser.

Et forsøg udføres, for at finde hvordan størkningstiden T (i minutter) afhænger af antal gram x af et additiv. 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 $T = \alpha_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \dots + \beta_p x^p$ 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.
- 3) angivet regressionsligningen $T = a_0 + b_1 x + b_2 x^2 + b_3 x^3 + \dots + b_p x^p$ hvor $a_0, b_1, \dots, b_p$ er estimater for $\alpha_0, \beta_1, \dots, \beta_p$, for den model, man i spørgsmål 3 har fundet frem til.

Løsning:

- 1) APPS, STAT/LIST, Data indtastes i list1(x- værdier) og list 2(T-værdier)
F4: Calc, 3Regressions, 9:QUADREG, Udfylder lister,
Da vi ønsker at tegne regressionslinien så StoreReqn to: y1(x), ENTER,
Af udskriften fås $r^2 = 0.7876$.
Tilsvarende kan man vælge CubicReg og QuartReg
CubicReg fås $r^2 = 0.8412$ altså en ret stor forbedring.
QuartReg fås $r^2 = 0.8463$, dvs. ingen væsentlig forbedring
Heraf slutes, at en trediegradsmodel må være en acceptabel model
- 2) Man kan nu tegne kurven ved at vælge "GRAPH".
Vi ønsker imidlertid punkterne tegnet med, så vi vælger
F2: Plots,,1: Plot Setup,F1: Define, Behold Scatter og Box, indsæt listerne , ENTER, ENTER,F5
Kurven vises sammen med punkterne.
Tegningen på lommeregnerens display viser, at punkterne fordeler sig tilfældigt omkring kurven.
- 3) Dette kan man kun gøre ved at omskrive trediegradsmodellen
 $T = a_0 + b_1 x + b_2 x^2 + b_3 x^3$ til $T = a_0 + b_1 x_1 + b_2 x_2 + b_3 x_3$ ved at sætte $x_1 = x, x_2 = x^2, x_3 = x^3$
Derefter foretage beregningerne som beskrevet under multipel regression i 3 variable (se eventuelt afsnit 4)
- 4) Man kan derefter skrive ligningen op.

3.2. Polynomial Regressionsanalyse med gentagelser

En "lack of fit" test kræver, at man som nævnt ovenfor omskriver til multipel regression.

4. Multipel Regressionsanalyse

Eksempel 13.10 (uden gentagelser)

Lad der foreligge følgende observationer.

x_1	x_2	x_3	x_4	Y
- 4	22	91	100	836
-1	20	90	95	789
7	21	88	110	883
16	19	87	88	790
18	20	91	94	816
23	19	94	99	859
27	23	87	97	831
29	21	86	96	832
24	22	88	110	897
16	23	91	105	872
10	20	90	100	842
3	20	89	98	821

- 1) Vurder ud fra forklaringsgraden og grafisk om en lineær model er rimelig.
- 2) Undersøg om modellen kan reduceres.
- 3) Angiv regressionsligningen i den endelige model
- 4) Angiv 95 % konfidensinterval for regressionskoefficienterne i den endelige model.
- 4) Find et 95% konfidensinterval for Y i punktet $(x_1, x_2, x_3, x_4) = (0, 20, 90, 100)$

Løsning:

- 1) APPS, STAT/LIST , navngiv lister x_1, x_2, x_3, x_4, y , og indtast data, F4: Calc, 3. Regressions, D:MultReg, Num of ind Vars = 4, Udfylder lister, ENTER, ENTER

Vi får regressionskoefficienterne og $r^2 = 0.9654$

Da vi ønsker at tegne et residualplot, så vælges

F2,1:Plot Setup, F1:Define, Vælg i Statvar yhatlist og resid , F5:ZoomStat

Residualplottet på displayet viser, at punkterne fordeler sig tilfældigt omkring linien.

Outliers: APPS, STAT/LIST, F6:Test, B:MultRegTests, Udfyld menuen (er nok allerede sket), ENTER, I listen (efter inddata) findes listen sresid, som indeholder "Studentized residuals"..

Da kun en enkelt værdi numerisk er større end 2 og ingen er over 3, antages, at der ikke er outliers

Da yderligere forklaringsgraden er tæt ved 1 vurderes modellen at være rimelig god.

- 2) $H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$, H : Mindst en af regressionskoefficienterne er forskellig fra 0.

I Udskriften fra MultRegTests findes en P -værdi på 0.000034

Da P -værdi = 0.000034 < 0.001 forkastes H_0 (stærkt), dvs. mindst en af regressionskoefficienterne er forskellig fra 0.

I P-list findes P -værdierne for de enkelte regressionskoefficienter.

Denne findes også i Stat/list, hvor man ser, at nr. 3 giver den største P -værdi på 0.7195. Da den første P -værdi svarer til konstantleddet svarer nr. 3 til x_2 .

$H_0: \beta_2 = 0$ accepteres, da P -værdien = 0.7195 > 0.05.

x_2 -leddet bortkastes. **Bemærk, at man kun eliminerer én variabel ad gangen**

Vælg F6:MultRegTests, og udfyld menuen med kun 3 variable.

Man ser, at nu er den største P - værdi ud for x_3 og P -værdien = 0.0849.

$H_0: \beta_3 = 0$ accepteres da, da P -værdien = 0.0849 > 0.05

x_3 slettes nu af modellen.

Vælg F6:MultRegTests, og udfyld menuen med kun 2 variable.

Nu ses, at alle P - værdier er mindre end 0.05, dvs. modellen kan ikke reduceres mere.

3) I blist findes koefficienterne, dvs. $\hat{y} = 335.65 + 0.9017 x_1 + 4.94 x_4$.

4) 95% konfidensinterval for β_1 :
$$\left[\hat{\beta}_1 - t_{1-\frac{\alpha}{2}}(f_{\text{residual}}) \cdot s_{\hat{\beta}_1} ; \hat{\beta}_1 + t_{1-\frac{\alpha}{2}}(f_{\text{residual}}) \cdot s_{\hat{\beta}_1} \right]$$

I Statvar findes $s_{\hat{\beta}_4}$ i "selist" = 0.23495. $f_{\text{residual}} = df_{\text{err}} = 9$

$[0.9017 - t_{0.975}(9) \cdot 0.23495; 0.9017 + t_{0.975}(9) \cdot 0.23495] = [0.371; 1.432]$

Tilsvarende findes $s_{\hat{\beta}_4}$ i "selist" = 0.40245 og dermed 95% konfidensinterval for β_4

$[4.94 - t_{0.975}(9) \cdot 0.40245; 4.94 + t_{0.975}(9) \cdot 0.40245] = [4.03; 5.85]$

5) Vælg F7:Ints, 8: MultRegInt, Udfyld menuer heraf x Value List={0,100}, ENTER

Man finder 95% konfidensinterval: [820.34 ; 838.99]

Eksempel 13.11. Multipel regression med gentagelser

Det formodes, at $Y = \alpha_0 + \beta_1 x_1 + \beta_2 x_2$. Følgende observationer foreligger:

(x_1, x_2)	(4,3)	(5,4)	(5,6)	(6,6)	(7,8)	(9,1)
y	33.0	44.4	50.0	57.4	73.0	87.1
	35.7	46.9	52.1	59.1	75.0	89.3

1) Test om man på basis af ovennævnte observationer kan få en accept heraf

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 for den fundne model.

Løsning:

1) Lack of fit test:

APPS, STAT/LIST , navngiv lister x_1 , x_2 , y , og indtast data

F6:Test, B:MultRegTests, Udfyld menuen, ENTER,

Blandt udskrifterne findes for ERROR: $SS = SAK_{\text{residual}} = 44.3823$, $df = f_{\text{residual}} = 9$

Et estimat for den poolede SAK:

CATALOG, VARIANCE, Variance({ $x_1[1]$, $x_2[2]$ })+Variance({ $x_1[3]$, $x_2[4]$ })+...+Variance({ $x_1[11]$, $x_2[12]$ })

Resultat: $SAK_e = 14.84$ $f_e = k(n-1) = 6(2-1) = 6$

Da $SAK_{\text{residual}} = SAK_e + SAK_{\text{lack}}$, fås $SAK_{\text{lack}} = 44.3823 - 14.84 = 29.5423$ med $f_{\text{lack}} = 9 - 6 = 3$

Man kan nu udarbejde følgende tabel:

Variation	SAK	f	s^2	F
Lack of fit	29.5423	3	9.8474	3.9814
Gentagelser (error)	14.84	6	2.47333	
Residual	4438.23	9		

P -værdi = $P(F > 3.9814) = \text{FCdf}(3.9814, \infty, 3, 6) = 0.071$

Da P -værdi = $0.071 > 0.05$, accepteres den lineære model ” $Y = \alpha_0 + \beta_1 x_1 + \beta_2 x_2$ ”.

2) Residualplottet på displayet viser, at punkterne fordeler sig tilfældigt omkring linien.

3) Af resultaterne i “Plist” ses, at

1) $H_0: \beta_1 = 0$ forkastes, da P -værdi = $0.0023 < 0.05$

2) $H_0: \beta_2 = 0$ forkastes, da P -værdi = $0.00391 < 0.05$

Modellen kan ikke reduceres.

4) Af resultaterne i “Blist fås at Ligningen bliver $\hat{y} = \underline{\underline{1.469 + 5.653 x_1 + 3.761 x_2}}$.

13B Eksempler regnet på Statgraphics.

1. Indledning

I "Grundlæggende statistik: appendix A" er beskrevet nogle grundlæggende operationer, hvorledes man beregner sandsynligheden for forskellige fordelinger og beregner gennemsnit og spredning. Dette forudsættes bekendt

2. Enkelt Regressionsanalyse

2.1. Regressionsanalyse uden gentagelser (1 faktor)

Eksempel 13.4 (uden gentagelser)

Tilsætning af en vis mængde kunstfibre forøger et garns trækstyrke. Man har eksperimenteret med forskellige tilsatte mængder kunstfibre x og registreret garnets trækstyrke y ved disse forskellige mængder. Herved fremkom følgende observationsmateriale:

Mængde x (i gram) af kunstfibre pr kg uld	40	50	55	60	70	75	80	85	90	95	100	105	110	120	130
Trækstyrke : Y	4.5	6.5	5.4	7.0	8.2	8.0	7.1	8.9	8.2	10.3	9.6	10.8	10.5	11.2	12.0

- 1) Find r^2 og anvend denne samt en figur til vurdering af modellen.
- 2) Opskriv regressionsligningen.
- 3) Test om y er uafhængig af x
- 4) Find 95% konfidensinterval for hældningen β
- 5) Find 95% konfidensinterval for y svarende til $x = 100$
- 6) Find 95% Prædistributionsinterval for 1 ny observation svarende til x - værdien 100.

Løsning:

Data indtastes

kunstfibre styrke

40 4,5

50 6,5

osv.

- 1) Variansanalysetabel opstilles:

Vælg (Relate \ Simple Regression \ indsæt "styrke" i y og "kunstfibre" i x \ OK)

Der fremkommer følgende tabel og figur:

Regression Analysis - Linear model: $Y = a + b \cdot X$

Dependent variable: styrke Independent variable: kunstfibre

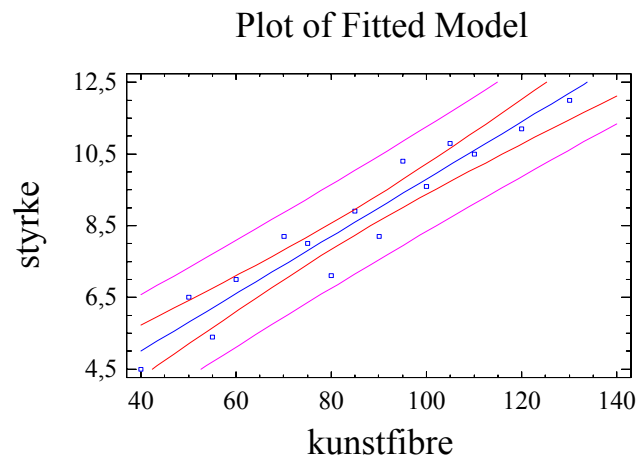
Parameter	Estimate	Standard Error	T Statistic	P-Value
Intercept	1,80866	0,578421	3,12688	0,0080
Slope	0,0798974	0,00656548	12,1693	0,0000

Analysis of Variance

Særk	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	62,1974	1	62,1974	148,09	0,0000
Residual	5,4599	13	0,419992		
Total (Corr.)	67,6573	14			
Correlation Coefficient = 0,958802				R-squared = 91,9301 percent	

Af udskriften ses, at forklaringsgraden “R-squared” er 91,93 %, hvilket er tilfredsstillende, da modellen altså “forklarer” 91,93% af variationen.

Samtidig med udskriften fremkommer følgende figur:



Af figuren ses, at punkterne fordeler sig tilfældigt omkring linien.

Ønsker man at gøre tegningen mere overskuelig kan man fjerne prediction og konfidens kurver ved følgende ordrer

Vælg(cursor på udskrift, højre musetast\Pane Options\fjern markering ved “Prediction limits” og “Confidence limits” \OK)

Outliers. Af ovenstående figur ses, at der næppe er nogen “outliers” (punkter der afviger så kraftigt fra det generelle billede, at man kunne frygte de var fejlmålinger), da ingen punkter falder udenfor 95% prædistributionslinierne (de yderste linier).

Da undersøgelse af outliers er vigtig, kan Statgraphics beregne såkaldte “Studentized Residuals”, som ligesom prædistributionsintervallerne tager i betragtning, at spredningen er mindre tæt ved “midtpunktet” end langt fra det.

Vælg(Tabul ar Options \ Unusual Residuals\OK).

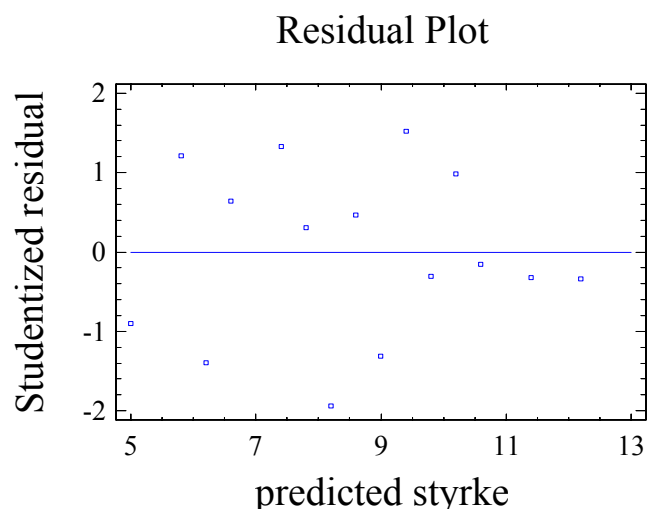
Heraf fremgår, at da ingen “Studentized Residuals”, numerisk er større end 2 er der ingen outliers. Det er også muligt at få dem tegnet.

Vælg (blå ikon= Graphical options\ Residual versus predicted\ OK)

Den fremkomne tegning viser, at disse “studentized residuals” ligger indenfor 3 enheder ja endda indenfor 2 enheder på hver side, hvilket igen viser, at der ikke er nogen outliers.

Sædvanlige residualer fås af:
(Cursor på tegning\ højre muse-
tast\ Residuals).

Konklusion: Modellen synes tilstrækkeligt godt at beskrive data indenfor måleområdet.



- 2) Af udskriften ved "Estimate" og "Slope" aflæses $\hat{\beta}_0 = 1.80866$ og $\hat{\beta}_1 = 0.0798$.

Regressionsligningen bliver derfor $\hat{y} = 1.80866 + 0.07989 \cdot x$

- 3) $H_0: Y$ er uafhængig af $x \Leftrightarrow H_0$: Regressionslinien er vandret $\Leftrightarrow H_0: \beta_1 = 0$.

Det ses, ud for "Model", at F -Ratio = 148.09 og at P -value = 0.0000, dvs.

H_0 forkastes Y er ikke uafhængig af x .

- 4) Konfidensinterval for β_1 :

Vælg (Relate \ Polynomial Regression \ indsæt "styrke" i y og "kunstfibre" i x \ OK \ Cursor i udskrift \ Analysis Options \ sæt order til 1 \ Gul ikon = Tabular options \ Confidence intervals \ OK).

Der fremkommer følgende udskrift:

95,0% confidence intervals for coefficient estimates

Parameter	Estimate	Standard Error	Lower Limit	Upper Limit
CONSTANT	1,80866	0,578421	0,559049	3,05826
kunstfibre	0,0798974	0,00656548	0,0657135	0,0940813

Heraf aflæses [0.0657 ; 0.0941]

- 5) 95% konfidensinterval for middeltrækstyrken svarende til x - værdien 100.

Vælg (Relate \ Simple Regression \ indsæt "styrke" i y og "kunstfibre" i x \ OK)

Vælg (Gul ikon = Tabular Options \ Forecasts \ OK \ Cursor på udskrift \ højre musetast \ Pane Options). Sæt i det fremkomne skema "Forecast at x " til 100 og stryg resten.

Der fremkommer følgende udskrift:

Predicted Values

X	Predicted Y	95,00% Prediction Limits		95,00% Confidence Limits	
		Lower	Upper	Lower	Upper
100,0	9,79839	8,33543	11,2614	9,37406	10,2227

Heraf aflæses, at den forventede middeltrækstyrke er 9.793 og 95% konfidensintervallet er [9.37;10.22]

- 6) 95% prædistributionsinterval for middeltrækstyrken svarende til x - værdien 100.

Af udskriften i punkt 5 fås : [8.34;11.26]



- 7) *Supplerende spørgsmål:*

Giv en grafisk vurdering af om kravet normalitet er opfyldt ved hjælp af et normalfordelingsplot.

Løsning:

Der skal vurderes om residualerne er tilnærmelsesvis normalfordelte. Der tegnes et normalfordelingsplot:

Vælg (sort ikon = Save Results \ Save Residuals \ OK).

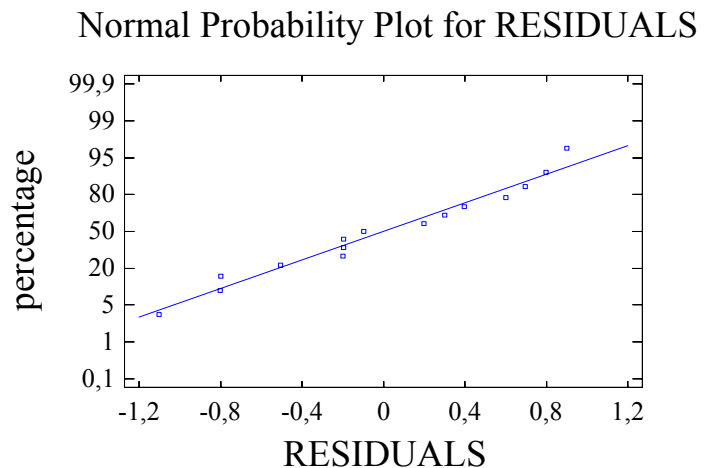
Residualerne bliver nu gemt som en søjle i data under navnet "RESIDUALS", da vi ikke har ændret navnet under "Target Variables"

Vælg (Describe \ Numerical Data \ One Variable Analysis \ RESIDUALS \ Pilen Data \ OK).

Vælg (blå ikon = Graphics options \ Normal Probability Plot \ OK).

Vælg (Med cursor på tegning, højre musetast \ Pane Options \ Using Least Squares \ OK).

Det ses, at residualerne ligger tilnærmelsesvis på en ret linie, så betingelsen om normalitet synes opfyldt.



2.2. Regressionsanalyse med gentagelser (1 faktor)

Eksempel 13.6 Regressionsanalyse med gentagelser

Givet følgende målinger

Tiden t	20	30	40	60	70	90	100	120	150	180
Tykkelse	4.2	7.4	8.8	13.6	13.1	14.9	20.0	23.1	27.5	32.9
y	4.9	6.9	8.2	12.0	12.4	16.8	21.2	25.2	25.1	32.4

- 1) Foretag en testning af forudsætningen om varianshomogenitet.
- 2) Det formodes på forhånd, at der er en lineær sammenhæng mellem x og y.
Test denne formodning ved en “lack of fit” test, og bestem i bekræftende fald ligningen for den empiriske regressionslinie.
- 3) Det påstås at hældningskoefficienten β_1 er 0.15
Test om dette på et signifikansniveau på 5% kan være sandt.
- 4) Angiv et 95% konfidensinterval for middelværdien af tykkelsen y, når $t = 100$ minutter.

Løsning:

Data indtastes i Statgraphics på sædvanlig måde:

```

x      y
20    4,2
20    4,9
30    7,4
30    6,9
40    8,8
40    8,2
...    ...
...    ...
180   32,9
180   32,4

```

- 1) Undersøgelse af varianshomogenitet.

Test af nulhypotesen $H_0: \sigma_1^2 = \sigma_2^2 = \dots = \sigma_{10}^2$

Vælg (Compare \ Analysis of Variance \ One Way ANOVA \ OK \ tabel udfyldes med y (“Dependent variable” og t (“Factors”) \ OK)

Vælg (gul ikon=Tabular options \ Variance check (fjern eventuelt krydset ved "Analysis summary") \ OK).

Der fremkommer følgende udskrift:

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

Da begge P - værdier er større end 0.05 accepteres H_0 , dvs. vi vil i det følgende antage, at kravet om varianshomogenitet er opfyldt.

- 2) H_0 : Lineær model gælder $\Leftrightarrow H_0: (x_i, \mu_i)$ ligger på en ret linie

Vælg (Relate \ Simple Regression \ indsæt x og y \ OK)

Regression Analysis - Linear model: $Y = a + b \cdot X$

Dependent variable: y

Independent variable: t

Parameter	Estimate	Standard Error	T 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			

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			

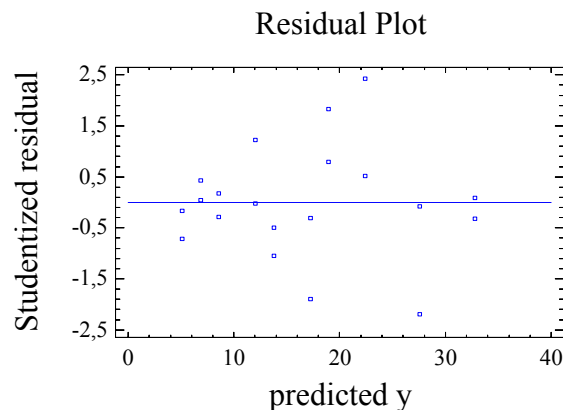
Af udskriften ses, at P - value ud for "lack of fit" er 0.0591. På et signifikansniveau på 5%, ses, at H_0 må accepteres, dvs. vi kan antage, at indenfor måleområdet giver førstegradsmodellen en rimelig god beskrivelse af resultaterne.

Af udskriften ved "Estimate" og "Slope" aflæses $\hat{\beta}_0 = 1.6542$ og $\hat{\beta}_1 = 0.1730$

Regressionsligningen bliver derfor $\hat{y} = 1.6542 + 0.1730 \cdot x$

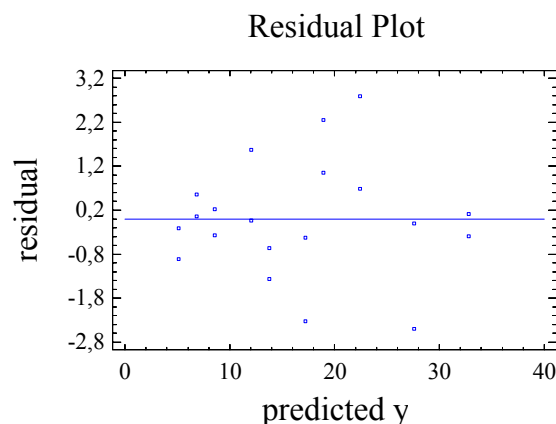
Da vi er tæt på forkastelse er det ekstra vigtigt yderligere at vurdere om den lineære model er rimelig ved at betragte et residualplot og i den forbindelse undersøge om der er outliers.

Vælg (blå ikon= Graphical options \ Residual versus predicted \ OK)



Da ingen af punkterne afvige mere end 3 fra midterlinien, så synes der ikke at være nogen outliers. Det skal bemærkes, at “Statgraphics” sætter grænsen ved 2, og derfor mener at der er 2 punkter man bør se nærmere på.

De sædvanlige residualer giver følgende tegning:



Idet det er middelværdierne der skal ligge på den rette linie, ses at netop gennemsnitspunkterne fordeler sig tilfældigt omkring linien.

Betragtes par af punkter, der svarer til samme x-værdi, ses, at afstandene er noget forskellige, dvs. der er nogen forskel på varianserne, men altså ikke mere end, at der ifølge testene er en rimelig varianshomogenitet.

3) $H_0: \beta_1 = 0.15$

Metode 1: Af udskriften ud for Slope fås, at “Standard Error “ = $s_{\hat{\beta}_1} = 0.0060$

$$t = \frac{\hat{\beta}_1 - a}{s_{\hat{\beta}_1}} = \frac{0.1730 - 0.15}{0.0060} = 3.812 \text{ er } t\text{-fordelt med } f = f_{\text{residual}} = 18$$

Da P -værdi = $P(T > 3.81) = \text{tCdf}(3.81, \infty, 18) = 0.00064 < 0.025$ forkastes H_0 , dvs.
data giver ikke den i litteraturen angivne hældningskoefficient..

Metode 2: Man finder et 95% konfidensinterval for β_1

Vælg (Relate \ Polynomial Regression \ indsæt y og x \ OK \ Cursor i udskrift \ Analysis Options \ sæt order til 1 \ Gul ikon = Tabular options \ Confidence intervals \ OK).

Man får følgende udskrift.

95,0% confidence intervals for coefficient estimates

Parameter	Estimate	Standard Error	Lower Limit	Upper Limit
CONSTANT	1,65415	0,599582	0,394468	2,91382
t	0,172975	0,00603212	0,160302	0,185648

Da konfidensintervallet ikke indeholder 0.15, må der konkluderes, at hælningskoefficienten ikke kan være 0.15.

- 4) Find det til $t = 100$ svarende 95% konfidensinterval for tykkelsen y .

Vælg (Relate \ Simple Regression \ indsæt y og x \ OK)

Vælg (Gul ikon = Tabular Options \ Forecasts \ OK \ Cursor på udskrift \ højre musetast \ Pane Options).

Sæt i det fremkomne skema "Forecast at x" til 100 og stryg resten.

Predicted Values

X	Predicted Y	95,00% Prediction Limits		95,00% Confidence Limits	
		Lower	Upper	Lower	Upper
100,0	18,9517	16,0518	21,8515	18,2956	19,6077

Vi får følgelig $\hat{y}_{100} = 18.95$ og 95% konfidensinterval [18,296 ; 19,608]



2.3. Transformation af model.

Eksempel 13.7. Transformation af udtryk.

Ved et forsøg komprimeres en luftart til forskellige forudvalgte rumfang v , idet de tilsvarende værdier af trykket P målt. Man formoder, at der gælder regressionsmodellen $P = c \cdot v^{-\gamma}$.

Ved forsøget fandtes følgende resultater:

$v \text{ cm}^3$	100	150	200	250	300	350	400	450	500	550	600
$P \text{ kp/cm}^2$	29.58	15.42	11.67	7.48	7.29	3.90	3.63	1.69	2.95	2.16	2.11

- 1) Begrund, at formodningen er rimelig
- 2) Angiv ligningen for den fundne model
- 3) Find middelværdien af P og et 95% konfidensinterval svarende til $v = 375$.

Løsning:

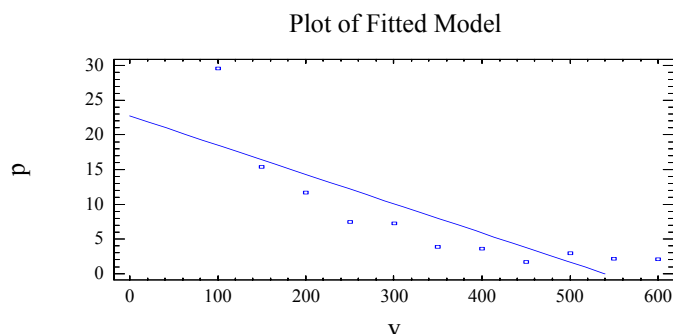
Data indtastes

p	v
100	29,58
150	15,42
200	11,67
osv.	

- 1) Finde "bedste" model:

Vælg (Relate \ Simple Regression \ indsæt p og v \ OK)

På figuren fjernes for at gøre figuren tydeligere "Prediction limits" og "Confidence limits"



Det ses tydeligt, at en model af typen $y = a + b x$ ikke er god.

For at få en vurdering af hvilken model der passer bedst

Vælg(gul ikon = Tabular options\Comparison of Alternative Models\OK)

Comparison of Alternative Models

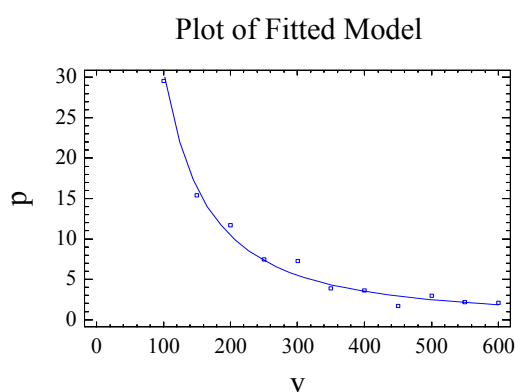
Model	Correlation	R-Squared
Reciprocal-X	0,9913	98,27%
Multiplicative	-0,9728	94,64%
Exponential	-0,9449	89,29%
Logarithmic-X	-0,9367	87,74%
S-curve	0,9348	87,39%
Square root-Y	-0,9063	82,15%
Reciprocal-Y	0,8984	80,71%
Square root-X	-0,8880	78,86%
Linear	-0,8323	69,28%
Double reciprocal	-0,7661	58,70%
Logistic	<no fit>	
Log probit	<no fit>	

“Reciprocal - X” og “Multiplicative” har de højeste forklaringsgrader.

Da vi af fysiske grunde mente, at en potensfunktion var den bedste, vælges den multiplicative.

Vælg (med cursor på udskrift, højre musetast\Analysis options\Multiplicative\OK)

Vi får følgende tegning:



Vi ser, at punkterne fordeles sig tilfældigt omkring kurven, så forhåndsformodningen om en potensfunktion synes korrekt.

Den tilsvarende variansanalysetabel er:

Regression Analysis - Multiplicative model: $Y = a \cdot X^b$

Dependent variable: p

Independent variable: v

Parameter	Estimate	Standard Error	T Statistic	P-Value
Intercept	10,6611	0,716487	14,8797	0,0000
Slope	-1,56967	0,124502	-12,6076	0,0000

NOTE: intercept = $\ln(a)$

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	8,02667	1	8,02667	158,95	0,0000
Residual	0,454476	9	0,0504973		
Total (Corr.)	8,48115	10			

Heraf aflæses ligningen : $\hat{P} = a \cdot v^b = e^{10,6611} \cdot v^{-1,56967} \Leftrightarrow \hat{P} = 42663,54 \cdot v^{-1,56967}$



3. Polynomial Regressionsanalyse

3.1. Polynomial Regressionsanalyse uden gentagelser

Eksempel 13.8 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 $T = \alpha_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \dots + \beta_p x^p$ 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 $T = a_0 + b_1 x + b_2 x^2 + b_3 x^3 + \dots + b_p x^p$ hvor $a_0, b_1, \dots, b_p$ er estimer for $\alpha_0, \beta_1, \dots, \beta_p$, for den model, man i spørgsmål 3 har fundet frem til.

Løsning:

- 1) Vurdering ud fra forklaringsgrad om polynomiums grad.

Data indtastes,

Vælg (Relate | Polynomial Regression \ indsæt x og T i den fremkomne tavle \ OK)

Først fås en variansanalysetabel svarende til en andengradsmodel (det er altid startværdien for Statgraphics).

13. Regressionsanalyse

Polynomial Regression Analysis

Dependent variable: T

Parameter	Estimate	Standard Error	T 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

Vælg (Cursor på udskrift\ højre musetast\ Analysis Options\order til 3\ OK)

Dette giver følgende variansanalysetabel.

Polynomial Regression Analysis

Dependent variable: T

Parameter	Estimate	Standard Error	T 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.

Vælg (Cursor på udskrift\ højre musetast\ Analysis Options\order til 4\ OK)

Dette giver følgende variansanalysetabel.

Polynomial Regression Analysis

Dependent variable: T

Parameter	Estimate	Standard Error	T 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 $0.5232 > 0.05$.

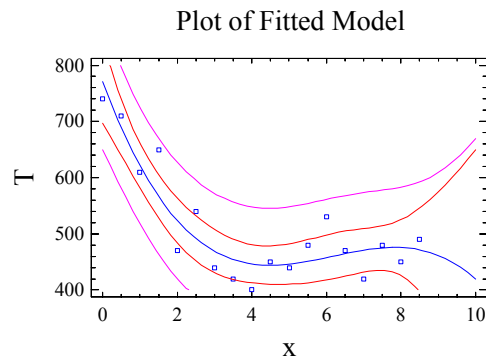
Bemærk iøvrigt, at selv om alle P -værdier for koefficienterne er større end 0,05, kan vi ikke deraf slutte, at vi kan reducere modellen til en konstant model. Man må kun bortkaste et led af gangen.

Det anførte tyder på, at en trediegradsmodel er en acceptabel model.

2): Grafisk kontrol af model.

Man bør altid som en ekstra kontrol indtegne kurven og punkterne.

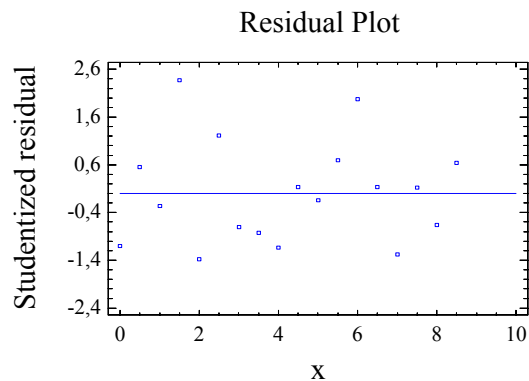
I Statgraphics sker det automatisk på højre side af udskriften. Den fremkomne tegning har indtegnede linier for konfidensgrænser og prædiktationsgrænser (predictionlimits). De 16 punkter synes at fordele sig rimeligt omkring den fundne kurve og ingen af punkterne falder udenfor prædiktationsgrænserne.



Vi danner et studentized residualplot.

Vælg (blå ikon= Graphical options\ Residual versus predicted\ OK)

Heraf ses, at der ikke er nogle “outliers”, da ingen værdier er over 3 (og kun et enkelt over 2)



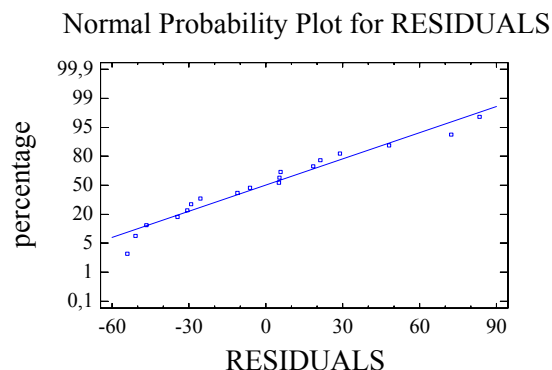
For at undersøge om kravet til normalfordeling er rimeligt opfyldt tegnes et normalfordelingsplot

Vælg (Describe\Numerical Data \One Variable Analysis\RESIDUALS\Pilen Data| OK).

Vælg (blå ikon = Graphics options\Normal Probability Plot| OK).

Vælg (Med cursor på tegning, højre musetast\Pane Options \Using Least Squares\OK)

Normalfordelingsplotet viser, at residualerne ligger rimelig tilfældigt omkring en ret linie, så forudsætningen om normalitet er opfyldt.



Samlet konklusion : Grafisk synes trediegradsmodellen at være rimelig.

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

4) Angiv regressionsligningen.

Ligningen ses af udskriften for trediegradsmodellen at være:

$$\hat{T} = 770.70 - 179.27x + 31.336x^2 - 1.629x^3$$



3.2. Polynomial Regressionsanalyse med gentagelser

Eksempel 13.9

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 minutter	740	610	470	440	400	440	530	420	450
	710	650	540	420	450	480	470	480	490

- 1) Vurder på basis af ovennævnte observationer ud fra forklaringsgraden, hvilket polynomium $T = \alpha_0 + \beta_1x + \beta_2x^2 + \beta_3x^3 + \dots + \beta_px^p$ 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.
- 3) angivet regressionsligningen for den model, man i de forrige spørgsmål har fundet frem til.
- 4) angivet et 95% konfidensinterval for regressionskoefficienten β_p til leddet af højest grad.
- 5) 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 $\hat{T}_m$, og et 95% konfidensinterval for $\hat{T}_m$.

Løsning:

1).Data indtastes,

Vælg (Relate| Polynomial Regression\ indsæt x og T i den fremkomne tavle\ OK)

Man får en variansanalysetabel svarende til en andengradsmodel (det er altid startværdien for Statgraphics).

Vi laver en “lack of fit test”

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	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 $0,0579 > 0,05$ accepteres andengradsmodellen.

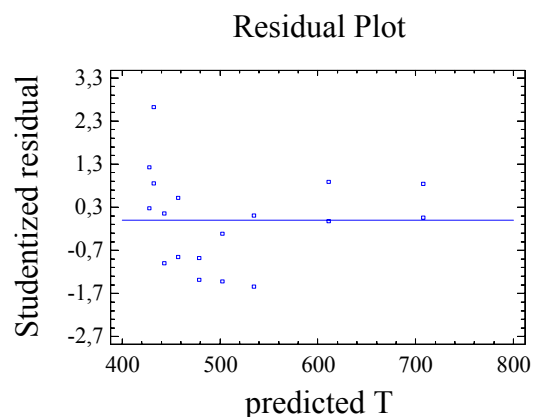
Da vi er meget tæt på forkastelse, kunne en trediegradsmodel dog overvejes.

2) Vi betragter nu et “studentized residualplot”:

Vælg (blå ikon= Graphical options\ Residual versus predicted\ OK)

Punkternes “gennemsnit” synes ikke at ligge helt tilfældigt omkring linien, så selv om punkterne ligger så tæt på linien at andengradsmodellen med nød og næppe kan accepteres, så er det næppe tilrådeligt at anvende modellen i hvert fald ikke ud over det angivne interval (at ekstrapolere)

Selv om en enkelt værdi har en “studentized residual” på mere end 2, så vil vi ikke betragte det som en outliers (skal være over 3)



Følgelig dannes en trediegradsmodel.

Vælg (Cursor på udskrift \ højre musetast \ Analysis Options \ order til 3 \ OK).

Polynomial Regression Analysis

Dependent variable: T

Parameter	Estimate	Standard Error	T 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			

13. Regressionsanalyse

Da vi har gentagelser, ignoreres den fremkomne udskrift og vi foretager en “lack of fit” test. 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	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.

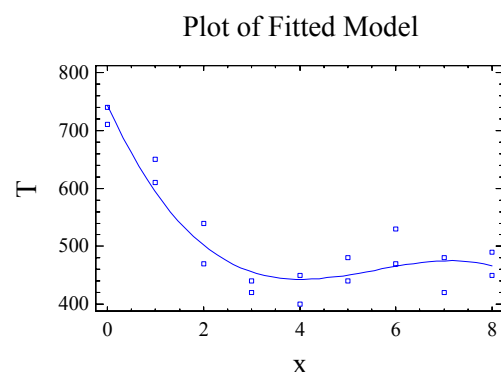
Som beskrevet under enkel regressionsanalyse, viser udskriften ved “Lack of Fit Test”, at

$$F = \frac{s_{lack}^2}{s_0^2} = 1.61 \text{ og denne værdi er ikke stor nok til at forkaste trediegradsmodellen.}$$

- 2) Som grafisk kontrol tegnes 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

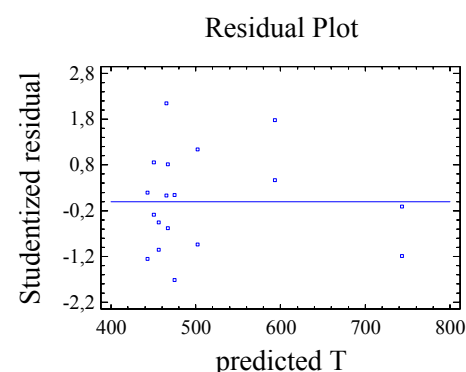


Vi betragter nu et “studentized residualplot”:

Vælg (blå ikon= Graphical options\ Residual versus predicted\ OK)

Vi ser, igen, at punkternes “gennemsnit” ligger tilfældigt omkring linien.

Vi ser, at der ikke er outliers, da alle værdier holder sig under 2



Samlet konklusion : Trediegradsmodellen synes at være rimelig.

- 3).Ligningen ses af udskriften at være: $\hat{T} = 743.23 - 183.26x + 35.52x^2 - 2.117x^3$

4) Opstilling af 95% konfidensinterval for β_3 .

Vælg (tabular options\ Confidence Intervals\ OK)

95,0% confidence intervals for coefficient estimates

Parameter	Estimate	Standard Error	Lower Limit	Upper Limit
CONSTANT	743,232	24,5077	690,668	795,796
x	-183,255	28,2854	-243,921	-122,588
x^2	35,5231	8,54009	17,2064	53,8398
x^3	-2,117	0,700505	-3,61944	-0,614566

Et 95% konfidensinterval for β_3 er følgelig [-3.62 ; -0.61].

Som forventet indeholder konfidensintervallet ikke 0.

5) Af figuren "Plot of fitted Model" ses, at den værdi x_m som giver den mindste størkningstid må være ca 4 og $\hat{T}_m \approx 440$ minutter.En mere præcis værdi fås ved at differentiere udtrykket for T .

$$T' = -3 \cdot 2.117x^2 + 2 \cdot 35.52x - 183.26 \quad T' = 0 \Leftrightarrow x = 4.206 \vee x = 6.671$$

Dette giver $x_m = \underline{4.21}$

Vælg (tabular options\ Forecasts\ højre musetast\ Pane options\ x til 4.21\ OK)

Vi får da følgende tabel:

Predicted Values

X	Predicted Y	95,00% Prediction Limits		95,00% Confidence Limits	
		Lower	Upper	Lower	Upper
4,21	443,378	358,125	528,63	414,535	472,221

Vi får følgelig $\hat{T}_m = 443.38$ minutter og 95% konfidensinterval [414.5 ; 472.2] ◆**4. Multipel Regressionsanalyse.****Eksempel 13.10 (multipel regressionsanalyse 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

$$Y = \alpha_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4.$$

13. Regressionsanalyse

Følgende observationer fra det forløbne år foreligger

x_1	x_2	x_3	x_4	Y
-4	22	91	100	836
-1	20	90	95	789
7	21	88	110	883
16	19	87	88	790
18	20	91	94	816
23	19	94	99	859
27	23	87	97	831
29	21	86	96	832
24	22	88	110	897
16	23	91	105	872
10	20	90	100	842
3	20	89	98	821

- 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.
- 4) Angiv 95% konfidensintervaller for de regressionskoefficienter der indgår i ovenstående model.
- 5) Angiv et 95% konfidensinterval for Y i punktet $(x_1, x_2, x_3, x_4) = (0, 20, 90, 100)$

Løsning:

- 1). Vurdering af om lineær model gælder.

Indtast data i 4 x - søjler og en Y- søjle.

Vælg ikonen "Multiple Regression" eller Vælg (Relate\ Multiple Regression).

Indsæt de variable i den fremkomne tavle.

Dette medfører følgende udskrift.

Multiple Regression Analysis

Dependent variable: Y

Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	175,499	113,863	1,54133	0,1671
x1	1,02664	0,226819	4,52626	0,0027
x2	-0,793015	2,12028	-0,374014	0,7195
x3	1,91131	1,14591	1,66794	0,1393
x4	4,98226	0,44901	11,0961	0,0000

Analysis of Variance

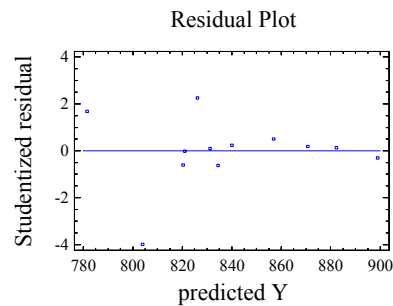
Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	12236,1	4	3059,02	48,90	0,0000
Residual	437,906	7	62,558		

Total (Corr.) 12674,0 11

R-squared = 96,5448 percent R-squared (adjusted for d.f.) = 94,5705 percent

Vælg (blå ikon= Graphical options\ Residual versus predicted\ OK)

Tegning af residualerne viser, at punkterne synes at placere sig rimeligt tilfældigt omkring 0-linien, Man kan også se, at der er ingen outliers.



Såvel ud fra R-squared på 96.55% som af residualplottet må konkluderes, at den lineære model er rimelig.

2) Mulig reduktion af modellen

$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$, H: Mindst en af regressionskoefficienterne er forskellig fra 0.

Af ovenstående udskrift ses ud for model, at P - Value = 0.0000

Da P -værdi = 0.000034 < 0.001 forkastes H_0 (stærkt), dvs mindst en af regressionskoefficienterne er forskellig fra 0.

Vi ser nu regressionskoefficienterne

Den β størrelse, der har størst P -værdi er β_2 .

$H_0: \beta_2 = 0$ accepteres, da P -værdien = 0.7195 > 0.05.

x_2 -leddet bortkastes.

Bemærk, at man kun eliminerer én variabel ad gangen.

Vi eliminerer nu x_2 :

Vælg (rød ikon = Input dialog \ slet x_2 i tavle \ OK).

Vi får :

Multiple Regression Analysis

Dependent variable: Y

Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	157,317	97,2707	1,61731	0,1445
x_1	1,01695	0,212876	4,7772	0,0014
x_3	2,03594	1,03578	1,96561	0,0849
x_4	4,88827	0,351523	13,906	0,0000

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	12227,3	3	4075,78	73,00	0,0000
Residual	446,657	8	55,8321		

Total (Corr.) 12674,0 11

R-squared = 96,4758 percent R-squared (adjusted for d.f.) = 95,1542 percent

13. Regressionsanalyse

Da P -værdien for x_3 er $0.0849 > 0.05$ eliminerer vi nu x_3 .

Multiple Regression Analysis

Dependent variable: Y

Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	335,652	40,268	8,33547	0,0000
x_1	0,901697	0,234953	3,83777	0,0040
x_4	4,94017	0,40245	12,2752	0,0000

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	12011,6	2	6005,81	81,60	0,0000
Residual	662,372	9	73,5968		

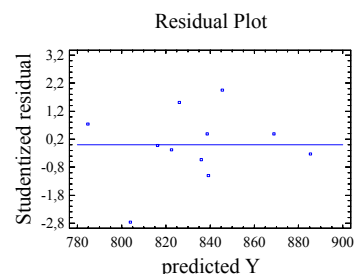
Total (Corr.) 12674,0 11

R-squared = 94,7738 percent R-squared (adjusted for d.f.) = 93,6124 percent

Det er nu ikke muligt at reducere modellen mere.

Vælg (blå ikon= Graphical options\ Residual versus predicted\ OK)

Residualplottet viser, at punkterne fordeler sig tilfældigt omkring linien, så vi antager at modellen er god nok.



3) Ligningen bliver $\hat{y} = 335.65 + 0.9017 \cdot x_1 + 4.9402 \cdot x_4$

4) Vælg (tabular options\ Confidence Intervals\ OK)

95,0% confidence intervals for coefficient estimates

Parameter	Estimate	Standard Error	Lower Limit	Upper Limit
CONSTANT	335,652	40,268	244,56	426,745
x_1	0,901697	0,234953	0,370195	1,4332
x_4	4,94017	0,40245	4,02977	5,85058

Konfidensintervallerne bliver $\beta_1: [0.3702; 1.4332]$, $\beta_4: [4.030; 5.851]$

5) Nederst i datafilen indsættes de ønskede x - værdier

Vælg (tabular options\ Report\ OK)

Regression Results for Y

Row	Fitted Value	Std. 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
13	829,67	9,51728	808,14	851,199	820,348	838,992

Vi har derfor, at $\hat{y} = 829.67$ og et 95% konfidensinterval er $[820.35; 838.99]$



Eksempel 13.11 (multipel regression med gentagelser)

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)	(5,6)	(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 for den fundne model.

Løsning:

- 1) Indtast data i 2 x - søjler og en Y- søjle.

Vælg ikonen "Multiple Regression" eller Vælg (Relate\ Multiple Regression).

Indsæt de variable i den fremkomne tavle.

Dette medfører følgende udskrift.

Multiple Regression Analysis

Dependent variable: y

Parameter	Estimate	Standard Error	T 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 at spalte SAK_{residual} fra ovenstående tabel op i $SAK_{\text{gentagelser}} = SAK_0$ og en $SAK_{\text{lack of fit}}$.

For at finde $SAK_{\text{gentagelser}}$ dannes en ekstra søjle "behandlinger", og med denne og y foretager en "ensidet variansanalyse"

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

osv.

Vælg (Compare\ Analysis of Variance \ One Way ANOVA \ Udfyld tavle med y og behandlinger \ OK).

Resultatet bliver

13. Regressionsanalyse

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 $SAK_{\text{residual}} = SAK_0 + SAK_{\text{lack}}$, fås $SAK_{\text{lack}} = 44.3823 - 14.84 = 29.5423$ med $f_{\text{lack}} = 9 - 6 = 3$.

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		

P -værdi = $P(F > 3.9814) = FCdf(3.9814, \infty, 3, 6) = 0.071$.

Da P -værdi = $0.071 > 0.05$, accepteres den lineære model ” $Y = \alpha_0 + \beta_1 x_1 + \beta_2 x_2$ ”.

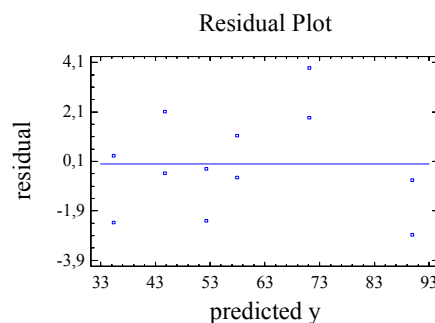
2. Grafisk kontrol af model.

For at få et overblik over punkternes placering tegnes et residualplot.

Vælg (blå ikon= Graphical options\ Residual versus predicted\ OK)\cursor på figur, højre musetast, vælg residuals<ok)

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.



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.

Heraf ses:

$H_0: \beta_1 = 0$ forkastes, da $P\text{-value} = 0.0029 < 0.05$

$H_0: \beta_2 = 0$ forkastes, da $P\text{-value} = 0.039 < 0.05$

Det ses, at vi ikke kan reducere modellen yderligere.

4. Opstilling af regressionsligning

Ligningen bliver $\hat{y} = 1.469 + 5.653x_1 + 3.761x_2$.

5. 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 "Independent Variable" indsættes x_1 , x_2 , x_1^2 , x_2^2 , $x_1 * x_2$

Der fremkommer en udskrift analogt med den der fremkommer hvis model var lineær i de variable som under afsnit 7.5, og en reduktion af model vil ske efter de samme retningslinier.

OPGAVER

Opgave 13.1

Nedenstående tabel angiver sammenhørende værdier af den "radiale" afbøjning X (i milliradianer) og den totale energiflux Y (i kilowatt) på et solvarmeanlæg.

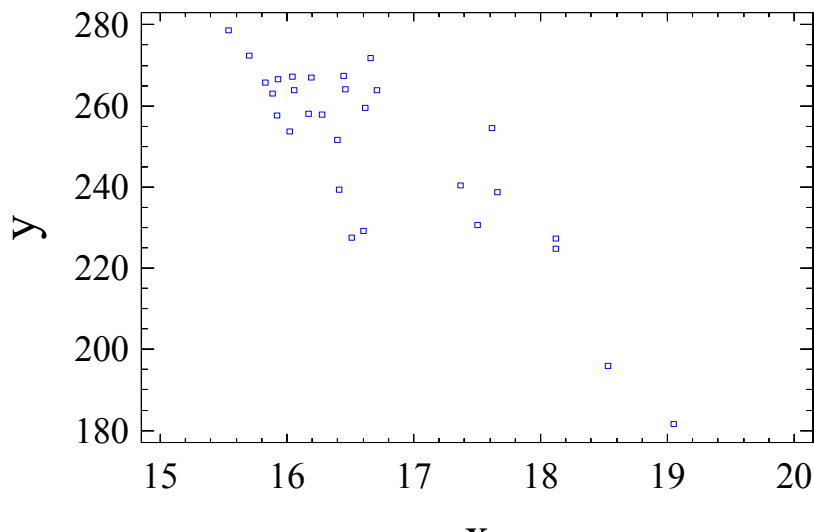
x	16.66	16.46	17.66	17.50	16.40	16.28	16.06	15.93	16.60	16.41	16.17	15.92	16.04	16.19	16.71
y	271.8	264.0	238.8	230.7	251.6	257.9	263.9	266.5	229.1	239.3	258.0	257.6	267.3	267.0	263.8
x	16.62	17.37	18.12	18.53	15.54	15.70	16.45	17.62	18.12	19.05	16.51	16.02	15.89	15.83	
y	259.6	240.4	227.2	196.0	278.7	272.3	267.4	254.5	224.7	181.5	227.5	253.6	263.0	265.8	

Der er på en lommeregner med regressionsprogram fundet følgende hjælpe størrelser:

	x	y
Antal	29	29
Gennemsnit	16,7021	249,638
Varians	0,824796	524,355
Spredning	0,908183	22,8988
Korrelationskoefficient	-0,848837	
Regressionskoefficienter	$\hat{\beta}_0 = 607.103,$	$\hat{\beta}_1 = -21.4025$

Punkterne er afsat på nedenstående figur.

Plot of y vs x



- 1) Begrund i ord på baggrund af figuren og forannævnte oplysninger, om du finder det sandsynligt, at der er uafhængighed mellem X og Y .

I det følgende antages, at en lineær model $y = \beta_0 + \beta_1 x$ gælder.

- 2) Angivet 99% konfidensinterval for hældningen β_1 . (Bemærk, der ønskes et 99% interval).
- 3) Beregn et 95% konfidensinterval for middelfluxen y i det tilfælde, hvor den radiale afbøjning x er 16.5 milliradianer.

Hvis et statistikprogram er til rådighed, er det tilladeligt at anvende dette.

Opgave 13.2

Man ønskede på et universitet at undersøge om der var en sammenhæng mellem de point de studerende fik ved en indledende prøve i matematik, og de point de fik ved den afsluttende prøve i matematik.

Resultaterne var

Student	1	2	3	4	5	6	7	8	9	10
Indledende prøve x	39	43	21	64	57	47	28	75	34	52
Afsluttende prøve y	65	78	52	82	92	89	73	98	56	75

- 1) Find en ligning for regressionslinien m , og tegn i et koordinatsystem såvel punkterne som linien m .
- 2) Man forventer en positiv korrelation mellem x og y . Finder du at dette er tilfældet? Det antages i det følgende at forudsætningerne for en regressionsanalyse er opfyldt.
- 3) Test om y er uafhængig af x (signifikansniveau $\alpha = 0.01$)
- 4) Find et 95% konfidensinterval for hældningskoefficienten β_1 .
- 5) En student har opnået 50 point ved den indledende prøve. Forudsig indenfor hvilket interval denne student pointtal vil ligge ved den afsluttende prøve. (signifikansniveau $\alpha = 0.05$)
- 6) Angiv et 95% konfidensinterval for middelværdien af det pointtal som studenter opnår ved den afsluttende prøve, når de alle ved den indledende prøve har opnået 50 point.

Opgave 13.3

Ved en kemisk proces vides reaktionshastigheden v at afhænge af mængden x af et bestemt additiv, som virker som katalysator.

Man formoder, at der (approksimativt) gælder sammenhængen

$$v = \alpha_0 + \beta_1 \cdot \sqrt{x}. \quad (1)$$

Ud fra teoretiske overvejelser forventes det yderligere, at $\beta_1 = \text{ca } 45$. Ved et fuldstændigt randomiseret forsøg fandtes følgende observationer:

Tilsat mængde additiv x	0	0.2	0.5	1.0	1.5	2.0
Reaktionshastighed v	1	17	29	41	50	58

- 1) Foretag en vurdering af, om model (1) kan antages at gælde.
- 2) Test, idet det forudsættes, at model (1) gælder, nulhypotesen $H_0: \beta_1 = 45$.
- 3) Opstil et 95% - konfidensinterval for β_1 .
- 4) Opstil et 95%~konfidensinterval for middelværdien af reaktionshastigheden ved en additivtilsætning på 1.75.

Opgave 13.4

I et forsøg undersøgte et ventilationsanlægs effektivitet. Målingerne foretoges ved at fylde et lokale med gas og vente til koncentrationen var stabil. Herefter startedes ventilationsanlægget og gaskoncentrationen C_t målte til forskellige tidspunkter t .

Følgende resultater fandtes:

t (min. efter anlæggets start)	2.67	4.59	6.75	7.67	11.34	14.34	16.25	18.25	23.09
C_t [ppm]	34	28	26	22	16	14	12	10	8

Følgende 2 modeller for funktionssammenhængen overvejes:

Model 1 (lineært henfald): $E(C_t|t) = \alpha_1 + \beta_1 \cdot t$

Model 2 (eksponentielt henfald): $E(C_t|t) = \alpha_2 \cdot e^{\beta_2 \cdot t}$

- 1) Indtegn punkterne i et koordinatsystem og vælg den af de to modeller du vurderer giver den bedste beskrivelse.
- 2) Beregn determinationskoefficienterne for hver af modellerne og vælg den model, for hvilken determinationskoefficienten er størst. Er der overensstemmelse med valget i spørgsmål 1)?
- 3) Antag, at model 2 gælder.
Bestem et 95%-konfidensinterval for halveringstiden $t_{0.5}$ bestemt ved $t_{0.5} = -\frac{\ln(2)}{\beta_2}$.

Opgave 13.5

I et organisk-kemisk laboratorium undersøgte man forskellige reaktionskinetiske processer. Ud fra teoretiske overvejelser har man fundet frem til, at "middeludbyttet" (angivet i %-enheder) af en bestemt kemisk forbindelse for $t > 5$ er approksimativt bestemt ved et udtryk af formen

$$(1) \quad y = 100 - \alpha_0 \cdot e^{\beta_1 \cdot t}, \text{ hvor } t \text{ angiver reaktionstiden og } y \text{ procesudbyttet.}$$

For at efterprøve rigtigheden af de teoretiske overvejelser udførte man et forsøg med følgende resultater:

t	6.5	8.2	11.1	13.6	16.4	18.5	20.7	23.0	25.8	28.5	33.3
y	39.5	64.7	65.6	72.9	88.0	92.7	92.5	95.9	96.3	98.3	99.2

- 1) Omskriv ovennævnte udtryk for modellen således, at regressionsmodellen kan gøres lineær i parametrene ved en logaritmisk transformation.
- 2) Foretag den logaritmiske transformation og vurder såvel grafisk som ud fra forklaringsgraden om den formodede model (1) kan accepteres.
- 3) Foretag, idet det forudsættes, at modellen (1) gælder, en estimation af parametrene α_0 og β_1 .
- 4) Opstil et 95% - konfidensinterval for middelværdien af udbyttet y svarende til $t = 20$.

Opgave 13.6

Ved en standardisering af et bestemt hormonpræparat behandler man et mindre antal mus med doser af forskellig størrelse og registrerer i hvert tilfælde tiden t , indtil musen dør. Fra tidligere undersøgelser ved man, at t er normalfordelt med konstant varians og med en middelværdi, som er en lineær funktion af logaritmen til dosis.

Til brug for standardiseringen af et produktionsparti af præparatet blev foretaget 5 delforsøg, som gav følgende resultater:

dosis (antal enheder)	1585	2239	2884	5248	6918
t (timer)	8.70	6.20	8.22	2.94	3.88

- 1) Angiv et estimat for regressionslinien, hvor t er en funktion af (titals)-logaritmen til dosis.
- 2) Opstil et 95% - konfidensinterval for koefficienten til logaritmen til dosis.
- 3) Opstil et 95% - konfidensinterval for middelværdien af t for en dosis på 6300 enheder.

Opgave 13.7

Man har erfaring for, at jerns viskositet Y under smeltning afhænger af jernets siliciumindhold x . Man besluttede sig til at foretage et forsøg med henblik på at undersøge denne sammenhæng nærmere. Ved forsøget foretoges 3 viskositetsmålinger for hver af 5 forskellige værdier af siliciumindholdet. Forsøgsresultaterne var:

	x				
	1.25	1.50	1.75	2.00	2.25
Y	47.5	60.0	65.0	72.5	77.5
	55.0	55.0	67.5	75.0	85.0
	37.5	50.0	70.0	75.0	75.0

- 1) Angiv forudsætningerne for at kunne udføre en variansanalyse.
- 2) Foretag en testning af forudsætningen om varianshomogenitet.
- 3) Test om der er en lineær sammenhæng mellem jerns viskositet og siliciumindholdet, og angiv i bekræftende fald ligningen for den empiriske regressionslinie.

Det antages i det følgende, at der er en lineær sammenhæng mellem x og y .

- 4) Foretag en testning af om regressionslinien er vandret.
- 5) Angiv et 95% konfidensinterval for hældningskoefficienten
- 6) Angiv et 95% konfidensinterval for middelværdien af middelviskositeten y , når $x = 1.60$.

Opgave 13.8

Koncentrationsbestemmelse af stoffet aprindin kan foretages ved hjælp af en gaskromatograf. Ved denne metode indsprøjtes en del af prøven indeholdende aprindin i gaskromatografen, og den såkaldte tophøjde bestemmes. Såfremt de laboratorietekniske procedurer er korrekt udført, skal tophøjden, bortset fra tilfældige udsving, være proportional med koncentrationen i prøven. I et eksperiment fremstillede man 12 prøver med kendte koncentrationer af aprindin og målte tophøjderne. Resultaterne fremgår af nedenstående tabel.

	Koncentration x ($\mu\text{g} / \text{ml}$)					
	0.5	1	2	3	4	5
Tophøjde	46	120	232	310	440	550
Y	55	90	241	318	435	539

- 1) Bestem den lineære regressionslinie for Y på x
- 2) Test, om en sådan lineær regression kan beskrive data.
- 3) Test, om tophøjden y kan antages at være proportional med koncentrationen x , dvs. $y = a \cdot x$.

Opgave 13.9

I et fuldstændigt randomiseret forsøg undersøgte rotters vægttilvækst som funktion af mængde riboflavin tilsat foderet. Forsøgsresultaterne var:

	Tilsætning x af riboflavin i $\mu\text{g} / \text{dag}$				
	2.5	5	10	20	40
Tilvækst Y i g/uge for 20 rotter (4 for hver af de 5 tilsætninger)	11.1 3.4 7.9 6.1	9.3 14.5 9.1 14.2	13.8 18.5 20.3 13.2	22.8 25.8 23.0 21.5	22.5 26.6 19.3 26.5

Følgende 2 modeller for funktionssammenhængen overvejes:

Model 1: $y = \alpha_0 + \beta_1 x$

Model 2: $y = \beta_0 + \beta_1 \ln(x)$

- 1) Foretag en grafisk testning af de to modeller og vælg den, der vurderes at give den bedste beskrivelse.
- 2) Udfør en testning af den i spørgsmål 1) valgte model.
- 3) Opstil, idet det forudsættes, at den i spørgsmål 2) udførte testning giver accept af den valgte model, et estimat og et 95% - konfidensinterval for middelvægttilvæksten ved tilsætning af $30 \mu\text{g} / \text{dag}$.

Opgave 13.10

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 regressionslinje og en figur med residualerne indtegnet til vurderingen).
- 2) Benyt menupunktet "Comparison of Alternative Models" til at få en idé 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 13.11

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	78.3	79.6	78.9	76.9	76.6	68.9	66.4
70.4	74.5	78.6	82.4	77.3	78.2	70.0	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 $\hat{Y}_m$ og et 95% konfidensinterval for $\hat{Y}_m$.

Opgave 13.12

Ved nogle forsøg med målinger af det tryk, som udgår fra jetmotorer, målt for udvalgte værdier af ændringen i udstødningsdysens vinkel x værdier af ændringen i trykket Y . Resultaterne var:

x	4	5	6,5	7	7,3	7,5
Y	24.6	27.2	67.3	77.8	80.6	83.4
(i%)	24.7	29.8	64.1	78.1	82.1	82.2

- Bestem ved en polynomial regressionsanalyse det polynomium af lavest grad, der giver en tilfredsstillende beskrivelse af Y 's variation.
- Opstil et 95% konfidensinterval for middelværdien af Y , når $x = 6$.

Opgave 13.13

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.

- Vurder ud fra forklaringsgraden og grafisk, om en lineær model i de tre variable, dvs af formen $Y = \alpha_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_3$ er rimelig.

Det antages i det følgende, at ovenstående model gælder.

- Undersøg om modellen kan reduceres.
- Angiv regressionsligningen i den endelige model.
- 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 13.14

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	85	55	75	76	56	85	106
	34	87	57	80	68	52	80	109

Det forudsættes, at regressionsforudsætningerne er opfyldt.

- Vis ved en sædvanlig 5% test, at en lineær model i de tre variable kan beskrive Y 's variation.
- Reducer om muligt modellen, og bestem regressionsligningen.
- 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 13.15

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_1	0.1	0.3	0.2	0.5	0.1	0.3	1.0	0.5	0.2	1.0
x_2	0.1	0.1	0.1	0.2	0.2	0.2	0.4	0.4	0.5	0.5
x_3	0.1	0.3	0.5	0.1	0.3	0.5	0.1	0.3	0.1	0.5
y	6.37	8.70	8.02	10.30	6.70	9.08	15.34	10.40	7.54	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 tilsidst 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 13.16

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 (aproksimativt) var lineær, og ønskede i bekræftende fald at estimere denne sammenhæng. Følgende observationer af Y (kodede tal) fandtes:

				x_2			
				10	20	30	40
x_3	-5	x_1	5	32	50	62	88
				30	47	66	82
			10	48	60	73	96
				51	63	77	99
		15		54	70	90	102
				56	75	83	110
	5	x_1	5	32	49	64	86
				35	52	60	84
			10	50	64	82	92
				55	60	77	94
		15		57	74	88	108
				53	77	86	105

- 1) Vis ved en testning at sammenhængen mellem Y og x_1 , x_2 og x_3 kan være lineær dvs. af formen $Y = \alpha_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_3$.
- 2) Undersøg, om modellen kan reduceres, dvs. om nogle af regressionskoefficienterne kunne være 0.
- 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.

Grundlæggende operationer på TI - 89.

1. Indledning.

Sædvanligvis vil man i statistik skulle analysere en eller flere kolonner af tal. I sådanne tilfælde benyttes

“Statistikmetoden”: Vælg APPS, Stats/List, indtast data i eksempelvis “list1”, og vælg en relevant “F- knap”.

Der fremkommer nu en menu, som er næsten selvforklarende.

Ønskes et resultat indsat på indtastningslinien: HOME, Var-Link.

I StatsVar mappen markeres den ønskede størrelse, ENTER .

Som eksempel se under afsnit 3: Beregning af gennemsnit m.m.

Skal man beregne sandsynligheder f.eks $P(X < 0.87)$, hvor X har en kendt sandsynlighedsfordeling, så er der 2 fremgangsmåder:

Statistikmetoden: Vælg APPS, Stats/list, F% og vælg den ønskede fordeling

“Sandsynlighedsmetoden”: Vælg HOME, CATALOG,, F3, vælg den ønskede fordeling (benyt evt. ALPHA,+ forbogstav for hurtigt at komme til det ønskede navn).

ENTER bevirker at funktionen indsættes på indsætningslinien.

Man indsætter nu parametrene (nederst på skærmen kan man se rækkefølgen af parametrene), og ved et tryk på ENTER fås resultatet.

Oprette og vælge en mappe (“Folder”):

Det kan være praktisk at oprette én eller flere mapper til at have sine data gemt i.

VAR-Link, F1, 5: Create Folder, Skriv navn på folder.

Vælg F2: View, og under Folder vælg den mappe som du vil arbejde i

Vælg en mappe som den aktuelle mappe: MODE, Current Folder,

2. Sandsynlighedsfordelinger.

Normalfordeling $n(\mu, \sigma)$

a) Find $p = P(a \leq X \leq b)$, hvor a, b, μ, σ er givne konstanter.

HOME, Catalog, F3 ,vælg normcdf, ENTER

normcdf(a, b, μ, σ)

b) Find fraktilen x_p : $P(X \leq x_p) = p$, hvor p, μ, σ er givne konstanter.

HOME, Catalog, F3

$x_p = \text{invNorm}(p, \mu, \sigma)$

t - fordeling.

Lad T være t - fordelt med frihedsgradstallet f .

a) Find $p = P(a \leq T \leq b)$, hvor a og b er givne konstanter.

tCdf(a, b, f)

b) Find fraktilen $t_\alpha(f)$: $P(T \leq t_\alpha(f)) = \alpha$ (α given konstant).

invT(α, f)

F-fordeling.

Lad F være F -fordelt med tællerfrihedsgradstallet f_T og nævnerfrihedsgradstallet f_N .

a) Find $p = P(a \leq F \leq b)$, hvor a og b er givne konstanter. $FCdf(a, b, f_T, f_N)$

b) Find fraktilen $F_\alpha(f_T, f_N)$: $P(F \leq F_\alpha(f_T, f_N)) = \alpha$ $invF(\alpha, f_T, f_N)$

3 Beregning af gennemsnit, varians og spredning

APPS, Stats/List, Data indtastes i "list1", F4, 1: 1-Var Stats,

I menu sættes "List" til "List1" (Benyt evt. Var-Link til at finde List1)

Eksempel: (hentet fra eksempel 5.1 side 53)

Forsøg nr	1	2	3	4	5	6	7	8	9	10	11	12
x	68.8	70.7	70.3	70.1	70.7	68.7	69.2	68.9	70.0	69.6	71	69.1

Udskriften består af en række statistiske størrelser, blandt hvilke man må finde det ønskede resultat. Man finder $\bar{x}=69.7583$, $s_x=0.8163$

Skal man regne videre med et resultat, f.eks finde variansen, skal resultatet indsættes på indtastningslinien:

Vælg Home og Var-Link. I StatsVar-mappen markerer man den ønskede størrelse. Tryk på ENTER indsætter så størrelsen på indsætningslinien.

Ønskes beregning af en enkelt størrelse f.eks gennemsnit = mean, spredning= stdDev eller varians = Variance

HOME, CATALOG,, F3,

Eksempel : Variance(list1)

Resultat: 0.664242

4. Hypotesetest og konfidensintervaller for 1 variabel.

APPS, STAT/LIST hvorefter eventuelle data indtastes i list1, list 2 osv.

4.1. Normalfordeling.

a1) Hypotesetest; σ kendt: F6, 1: Z-Test

I menu: Er data givet i en liste vælges Data.

Er kun gennemsnit (og σ) kendt vælges Stats.

Eksempel: Vælges ovennævnte data i list1, $\sigma=1$ og $H: \mu > 69.2$, fås P-værdi=2.48%.

a2) Konfidensinterval σ kendt: F7, 1: Z-Interval

Eksempel: Vælges ovennævnte data i list1 og er $\sigma=1$, fås $C\ Int=[69.2; 70.3]$

b1) σ ukendt: F6, 2: T-Test . Derefter som under punkt a)

b2) σ ukendt: F7, 2: T-Interval . Derefter som under punkt a)

4.2. Binomialfordeling.

a1) Hypotesetest: F6, 5: 1-Prop-ZTest

Menu selvforklarende. Bemærk: Kræver der kan approksimeres til normalfordeling

a2) Konfidensinterval: F7, 5: 1-Prop-ZInt

Bemærk: Kan kun benyttes, hvis kan approksimere til normalfordeling

Eksempel: Er af 100 forsøg de 85 en succes, så fås $C\ Int=[0.78; 0.93]$

Poissonfordeling: findes ikke, så her må formel for konfidensinterval benyttes

STATISTISKE TABELLER

Tabel 1 Fraktiler u_p i U-fordelingen $n(0,1)$. $P(U \leq u_p) = p$.

Bemærk: $u_p = -u_{1-p}$ Eksempler: $u_{0.975} = 1.960$

p	0.90	0.95	0.975	0.99	0.995	0.999	0.9995
u_p	1.282	1.645	1.960	2.326	2.576	3.090	3.291

Tabel 2 Fraktiler t_p i t -fordelingen $t(f)$. $P(T \leq t_p) = p$, hvor $T = \frac{(\bar{X} - \mu)\sqrt{n}}{s}$

Eksempler: For $t(27)$ er $P(X \leq 3.42) = 0.999$. $t_{0.95}(1)=6.31$. $t_{0.05}(10) = -t_{0.95}(10) = -1.81$.

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

FACITLISTE

KAPITEL 13

- 13.1** (1) nej (2) $[-28.51 ; -14.30]$ (3) $[249.2 ; 258.8]$
- 13.2** (1) $y = 40.78 + 0.766x$ (2) ja (3) $F = 19.14$, P -værdi = 0.024
(4) $[0.362 ; 1.170]$ (5) $[57.91 ; 100.21]$ (6) $[72.50 ; 85.62]$
- 13.3** (1) $r^2 = 0.9984$ (2) $t = -6.73$, P -værdi = 0.00123 (3) $[38.92 ; 42.48]$
(4) $[52.73 ; 55.25]$
- 13.4** (1) Model 2, (2) 0.9293, 0.9883 (3) $[8.703 ; 10.584]$
- 13.5** (1) - (2) $r^2 = 0.980$ (3) $\tilde{\alpha}_0 = 171.8$ $\tilde{\beta}_1 = -0.1585$ (4) $[91.7 ; 93.7]$
- 13.6** (1) $t = 35.435 - 8.380 \cdot \log(\text{dosis})$ (2) $[-17.39 ; 0.6277]$ (3) $[0.257 ; 6.929]$
- 13.7** (1) - (2) P -værdi = 0.2639 (3) $P(\text{lack of fit}) = 0.6099$, $y = 5.583 + 33.667x$
(4) $F = 84.41$, P -værdi = 0.000 (5) $[25.75 ; 41.58]$ (6) $[56.41 ; 62.49]$
- 13.8** (1) $Y = -0.1438 + 108.959x$ (2) $F = 2.57$, P -værdi = 0.144
(3) $t = -0.019$, P -værdi = 0.9845
- 13.9** (1) $P(\text{lack of fit}) = 0.0026$, model 2. (2) $F = 1.20$, $P(\text{lack of fit}) = 0.3535$
(3) 23.55, $[21.35 ; 25.76]$
- 13.10** (1) - (2) - (3) $Y = 2.9764 - \frac{6.9158}{x}$, $Y = 1.988$, $[1.944 ; 2.033]$
(3) 23.55, $[21.35 ; 25.76]$
(4) - (5) $Y = -1.144 + 0.7197x - 0.0379x^2$, $Y = 2.036$, $[1.959 ; 2.114]$.
- 13.11** (1) $Y = 65.2536 + 0.7385x - 0.00938x^2$, (2) $Y = 79.45$, $[78.05 ; 80.93]$
(3) $Y_m = 79.79$, $[78.38 ; 81.20]$
- 13.12** (1) $Y = 630.26 - 344.31x + 62.00x^2 - 3.443x^3$,
(2) $Y = 52,7531$, $[51.09; 54.41]$
- 13.13** (1) $r^2 = 0.7231$, OK, (2) ja, (3) $Y = 1131.88 + 9.5846x_1$,
(4) 2330, $[2267 ; 2393]$.
- 13.14** (1) $F_{lack} = 2.26$, $P(\text{lack of fit}) = 0.1517$ (2) $Y = 22.537 + 2.888x_1 + 6.586x_2$
(3) 67.02, $[64.66 ; 69.39]$
- 13.15.** (1) - (2) $y = 5.199 + 9.964x_1 + 1.689x_3$ (3) $[9.64 ; 10.29]$, $[1.07 ; 2.31]$ (4) 8.357, $[8.202 ; 8.513]$
- 13.16.** (1) - (2) - (3) $y = 6.146 + 2.301x_1 + 1.622x_2$ (4) $[2.06 ; 2.55]$ (5) 57.03, $[55.83 ; 58.23]$

STIKORDSREGISTER

A

B

Bartlett's test 15, 29

C

Comparison of Alternative Methods 50

D

determinationskoefficient 5

E

ekstrapolation 5

enkelt regressionsanalyse 2

formler 2

med gentagelser 14

formler 16, 28

med Statgraphics 46

på TI -89 36

uden gentagelser 11

formler 12, 26

med Statgraphics 43

på TI - 89 35

F

facitliste 74

F - test 9

simplificeret 15, 29

fordeling, tabel over

fraktiler for normalfordeling 73

fraktiler for t - fordeling 73

forklaringsgrad 5, 7

forudsætninger for regressionsanalyse 8

fraktiler beregnet ved

fraktiltabel for

normalfordeling 73

t - fordeling 73

G

H

I,J

K

konfidensinterval

regressionskoefficient 10, 11

formler 11, 16, 27, 34

med Statgraphics 45, 49

på TI - 89 36, 38

for den til x svarende værdi af Y

formler 11, 16, 27, 33

med Statgraphics 45, 49

på TI - 89 36, 38

korrelationskoefficient 5

L

lack of fit test 5, 28

lineær model 2

M

mindste kvadraters metode 3, 4

multipl regression 20, 31, 40, 41

N

normalfordeling

plot 8, 45

tabel 73

normalligningssystem 4, 22, 31

O

opgaver 63

oversigter

enkelt regression uden gentagelser 26

enkelt regression med gentagelser 28

transformation til lineær model 30

multipl regression 31

outliers 5, 35, 44

P

polynomial regressionsanalyse

med gentagelser 19, 39, 54

uden gentagelser 19, 39, 61

poolet estimat for varians 16

prædestinationsinterval 10, 27

Stikord

Q

R

residual 3
 plot 44, 48
 studentized 15
regressionsanalyse
 enkelt 2
 med gentagelser
 formler 16, 28
 med Statgraphics 46
 på TI - 89 36
 uden gentagelser
 formler 11, 12, 26
 med Statgraphics 43
 på TI - 89 35
 forudsætninger 8
 multipel
 med gentagelser 25, 32, 60
 uden gentagelser 20, 32, 57
 polynomial
 med gentagelser 19, 39, 54
 uden gentagelser 19, 39, 51
regressionskoefficienter 3, 23
regressionsligning 3
regressionslinie 3

S

SAK 7, 9
Statgraphics
 løsning
 lineær regression 43, 46
 multipel regression 57, 60
 polynomial regression 51, 54
statistisk uafhængige 8
studentized residualer 44
støj 9

T

tabel for fraktil
 normalfordeling 73
 t - fordeling 73
TI - 89
 Grundlæggende operationer 71
 løsning
 enkelt regression 35
 multipel regression 40
 polynomial regression 39

transformation 17, 30, 39, 49

U

afhængige observationer 8

V

variabeltransformation 17, 30, 39, 49
variensanalysetabel 9, 26, 32
varienshomogenitet 8, 29, 47

W

X