Afdeling for Teoretisk Statistik
Institut for Matematiske Fag
Aarhus Universitet

Statistik 1 og Statistik α
Jørgen Granfeldt
1. november 2002

Meddelelse 9

Forelæsningerne i sidste uge

(Onsdag den 23. oktober) Afsluttede lineær regression og begyndte på Kapitel 4 *Lineære normale modeller*, hvor jeg nåede til afsnit 4.1.2 *Test*.

(Fredag den 25. oktober) Kapitel 4 *Lineære normale modeller*: Afsluttede Afsnit 4.1 og begyndte på Afsnit 4.2 *Sammenligning af flere regressionslinjer*.

Forelæsningerne i denne uge

(Onsdag den 30. oktober) Afsluttede Afsnit 4.2 *Sammenligning af flere regressionslinjer*.

(Fredag den 1. november) Afsnit 4.3 *Tosidet variansanalyse*.

Forelæsningerne i næste uge

(Onsdag den 6. november og fredag den 8. november) Afslutning af *Tosidet variansanalyse* og derefter Kapitel 6 *Multinomialfordelte data*.

Øvelserne i denne uge

(Uge 44: 28. oktober-3. november) 3.14, 3.15, 3.16.

Øvelserne i de tre næste uger

(Uge 45: 4. november-10. november) 4.3, 4.4, 4.5 (regnes færdig) samt opgaven sidst på denne meddelelse.

(Uge 46: 11. november-17. november) 4.6, 4.8, 4.9, 4.10.

(Uge 47: 18. november- 24. november) 6.1, 6.4, 6.9, 6.10

Afleveringsopgaver

(Uge 44) Ingen afleveringsopgave.

(Uge 45) Ingen afleveringsopgave.

(Uge 46) Ingen afleveringsopgave.

(Uge 47) Ingen afleveringsopgave.

Udleveret materiale

2. obligatoriske opgave til aflevering fredag den 22. november.

Nogle SAS vink

Testsandsynligheder for de test vi bruger kan beregnes nøjagtigt i SAS ved at benytte SASs probability functions:

`PROBNORM(x)`	returns the probability from the standard normal distribution
`PROBT(x,df<,nc>)`	returns the probability from a Student's t distribution
`PROBCHI(x,df<,nc>)`	returns the probability from a chi-squared distribution
`PROBF(x,ndf,ddf<,nc>)`	returns the probability from an F distribution

Funktionerne bruges i et `DATA` trin, og man må bruge `PROC PRINT` for at se resultatet, som i nedenstående eksempel.

```
DATA p;
pu  =2*(1-probnorm(1.044));
pt  =2*(1-probt(1.429,39));
pchi=2*(probchi(39*0.5331,39));
* de tre ovenstaaende er fra Eksempel 3.1;
* nedenstående er fra Eksempel 3.3 side 3.58;
pf=1-probf(21.46,3,48);
PROC PRINT DATA=p;RUN;
```

giver følgende udskrift:

```
                          The SAS System                          1

        Obs        pu         pt        pchi          pf

         1     0.29649    0.16097    0.014865    5.8597E-9
```

Opgave til øvelserne i næste uge

Følgende opgave uddyber sammenhængen mellem lineær regression som behandlet i Afsnit 3.3 og behandlet som lineær normal model i Eksempel 4.4

Lad altså $x_1,\ldots,x_n$ være realisationer af uafhængige normalfordelte stokastiske variable $X_1,\ldots,X_n$, hvor

$$X_i \sim N(\mu_i,\sigma^2),\ i=1,\ldots,n.$$

I lineær regression er $\mu_i = \alpha+\beta t_i$, så

$$\begin{aligned}\mu &= (\alpha+\beta t_1,\ldots,\alpha+\beta t_n)^* = \alpha(1,\ldots,1)^* + \beta(t_1,\ldots,t_n)^* \\ &= \alpha\mathbf{e}+\beta\mathbf{t},\end{aligned}$$

hvor

$$\mathbf{e} = (1,\ldots,,1)^* \qquad \text{og} \qquad \mathbf{t} = (t_1,\ldots,t_n)^*.$$

Underrummet svarende til lineær regression er altså $L = \text{span}(\mathbf{e}, \mathbf{t})$.

(1) Vis, at dimensionen af L er 2, hvis der er mindst to forskellige værdier af den uafhængige variabel t.

(2) Hvornår er $\mathbf{e}$ og $\mathbf{t}$ ortogonale?

(3) Vis, at vektorerne $\mathbf{e}$ og $\mathbf{u}$, hvor

$$\mathbf{u} = \mathbf{t} - \bar{t}.\mathbf{e},$$

er ortogonale. Vis at $||\mathbf{e}||^2 = n$ og $||\mathbf{u}||^2 = SSD_t$.

(4) Find den ortogonale projektion på L og find på den måde maksimum likelihood estimaterne for α og β.

Betragt nu to regressionslinjer, det vil sige modellen

$$X_{hi} \sim N(\alpha_h + \beta_h t_{hi}, \sigma^2),\ i = 1, \ldots, n_h,\ h = 1, 2,$$

og definer vektorerne $\mathbf{e}_1$ og $\mathbf{e}_2$ ved henholdsvis

$$(\mathbf{e}_1)_j = \begin{cases} 1 & \text{for } j = 1, \ldots, n_1, \\ 0 & \text{for } j = n_1 + 1, \ldots, n_1 + n_2, \end{cases}$$

og

$$(\mathbf{e}_2)_j = \begin{cases} 0 & \text{for } j = 1, \ldots, n_1, \\ 1 & \text{for } j = n_1 + 1, \ldots, n_1 + n_2, \end{cases}$$

samt vektorerne $\mathbf{t}_1$ og $\mathbf{t}_2$ ved henholdsvis

$$(\mathbf{t}_1)_j = \begin{cases} t_{1j} & \text{for } j = 1, \ldots, n_1, \\ 0 & \text{for } j = n_1 + 1, \ldots, n_1 + n_2, \end{cases}$$

og

$$(\mathbf{t}_2)_j = \begin{cases} 0 & \text{for } j = 1, \ldots, n_1, \\ t_{2(j-n_1)} & \text{for } j = n_1 + 1, \ldots, n_1 + n_2. \end{cases}$$

Underrummet svarende til to regressionslinjer uden bånd på parametrene er $L_1 = \text{span}(\mathbf{e}_1, \mathbf{e}_2, \mathbf{t}_1, \mathbf{t}_2)$.

(5) Vis, at underrummet svarende til parallelle regressionslinjer er $L_2 = \text{span}(\mathbf{e}_1, \mathbf{e}_2, \mathbf{t}_1 + \mathbf{t}_2)$. Find den ortogonale projektion på L_2 og dermed maksimum likelihood estimaterne for afskæringer og for hældningen i den model.

(Vink: Lad jer inspirere af (3) find ortogonale vektorer, som udspænder L_2)

STAT 1, TØ nr. 8

opg. 3.14

Da X_i erne er uafh. gælder, at $f_{\underline{X}}(\underline{x}) = \prod_{i=1}^{n} f_{X_i}(x_i)$

Da $X_i \sim N(\underline{t}_i^* \underline{\beta}, \sigma^2)$ følger, at like. fkt. er

$$L(\underline{\beta}, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_i - \underline{t}_i^* \underline{\beta})^2} = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \underline{t}_i^* \underline{\beta})^2}$$

Dvs. $\ell(\underline{\beta}, \sigma^2) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \underline{t}_i^* \underline{\beta})^2$

$$\frac{\partial \ell}{\partial \beta_j} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \underline{t}_i^* \underline{\beta})\underline{t}_{ij} \quad , \; j = 1, \dots, k$$

$$\frac{\partial \ell}{\partial \beta_j} = 0 \quad , \; j = 1, \dots, k \iff \sum_{i=1}^{n} \underline{t}_i^* \underline{\beta}\, \underline{t}_{ij} = \sum_{i=1}^{n} x_i \underline{t}_{ij} \quad j = 1, \dots, k$$

$$\iff \sum_{i=1}^{n}\sum_{r=1}^{k} \underline{t}_{ir} \underline{\beta}_r \underline{t}_{ij} = \sum_{i=1}^{n} \underline{t}_{ij} x_i \quad , \; j = 1, \dots, k$$

$$\iff \sum_{r=1}^{k}\left(\sum_{i=1}^{n} \underline{t}_{ir} \underline{t}_{ij}\right)\underline{\beta}_r = \sum_{i=1}^{n} \underline{t}_{ij} x_i \quad , \; j = 1, \dots, k$$

$$\underset{(*)}{\iff} T^* T \underline{\beta} = T^* \underline{x}$$

$$T = \begin{pmatrix} - \underline{t}_1 - \\ \vdots \\ - \underline{t}_n - \end{pmatrix}$$

$$T^* T \underline{\beta} = \underbrace{\begin{pmatrix} | & & | \\ \underline{t}_1 & \cdots & \underline{t}_n \\ | & & | \end{pmatrix}}_{k \times n} \underbrace{\begin{pmatrix} - \underline{t}_1 - \\ \vdots \\ - \underline{t}_n - \end{pmatrix}}_{n \times k} \underbrace{\begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}}_{k \times 1}$$

$$(T^* T \underline{\beta})_j = \sum_{r=1}^{k}\left(\sum_{i=1}^{n} \underline{t}_{ij} \underline{t}_{ir}\right)\underline{\beta}_r$$

Idet T har fuld rang gælder, at $(T^*T) \in Gl_{k\times k}(\mathbb{R})$
dvs. $\hat{\underline{\beta}} = (T^*T)^{-1}(T^*T)\hat{\underline{\beta}} = (T^*T)^{-1}T^*\underline{x}$

2)

Lad $i,j \in \{1, \ldots, k\}$

$$\frac{\partial^2 \ell}{\partial \beta_i \partial \beta_j}(\underline{\beta}, \sigma^2) = \frac{-1}{\sigma^2}\sum_{r=1}^{n} t_{ri}\, t_{rj}$$

Dvs. jacobimatricen er $-\frac{1}{\sigma^2}T^*T$
som jo er neg. definit. (dvs. løsn. til lihe. lign. er et maks. pkt.)

3)

Idet $\underline{X}_i$'s middelværdi er $\underline{t}_i^*\underline{\beta}$ bliver den estimerede værdi (udfra maks. lihe. est.) $\underline{t}_i^*\hat{\underline{\beta}} = \underline{t}_i^*(T^*T)^{-1}T^*\underline{x}$
Altså er den estimerede middelværdivektor for $\underline{X}$
$\hat{\underline{x}} = T(T^*T)^{-1}T^*\underline{x}$

Idet $\underline{x} \sim\sim N(T\underline{\beta}, \sigma^2 I_{n\times n})$

og $T(T^*T)^{-1}T^*T\underline{\beta} = T\underline{\beta}$
$T(T^*T)^{-1}T^*\sigma^2 I_{n\times n}\, T(T^*T)^{-1}T^* = \sigma^2 T(T^*T)^{-1}T^*$

følger af proposition 9.3, at $\hat{\underline{x}} \sim\sim N(T\underline{\beta}, \sigma^2 T(T^*T)^{-1}T^*)$

$H := T(T^*T)^{-1}T^*$, er jvf. sætn. 6.11 i F&B projektionsmatricen for proj. ned på underrummet udspændt af søjlerne i T

4)

$$\ell(\underline{\beta}, \sigma^2) \leq \ell(\hat{\underline{\beta}}, \sigma^2) \quad , \forall \sigma^2$$

$$\frac{\partial \ell}{\partial \sigma^2}(\hat{\underline{\beta}}, \sigma^2) = -\frac{n}{2}\frac{1}{\sigma^2} + \frac{2}{(2\sigma^2)^2}\sum_{i=1}^{n}(x_i - \underline{t}_i^* \hat{\underline{\beta}})^2$$

$$\begin{cases} = 0 \quad , \iff \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \underline{t}_i^* \hat{\underline{\beta}})^2 \\ > 0 \quad , \quad \sigma^2 < \text{ — " — } \\ < 0 \quad , \quad \sigma^2 > \text{ — " — } \end{cases}$$

Altså er maks. likel. est. for σ^2

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \underline{t}_i^* \hat{\underline{\beta}})^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \underline{t}_i^*(T^*T)^{-1}T^*\underline{x})^2$$

$$= \frac{1}{n}\Big(\underbrace{\sum_{i=1}^{n} x_i^2}_{a)} + \underbrace{\sum_{i=1}^{n}(\underline{t}_i^*(T^*T)^{-1}T^*\underline{x})^2}_{b)} - 2\underbrace{\sum_{i=1}^{n} x_i \underline{t}_i^*(T^*T)^{-1}T^*\underline{x}}_{c)}\Big)$$

$$a) = \underline{x}^*\underline{x}$$

$$b) = \sum_{i=1}^{n} \underline{x}^* T (T^*T)^{-1} \underline{t}_i \underline{t}_i^* (T^*T)^{-1} T^* \underline{x}$$

$$= \underline{x}^* T (T^*T)^{-1} \Big(\sum_{i=1}^{n} \underline{t}_i \underline{t}_i^*\Big)(T^*T)^{-1} T^* \underline{x}$$

$$\underset{(*)}{=} \underline{x}^* T (T^*T)^{-1}(T^*T)(T^*T)^{-1} T^* \underline{x}$$

$$= \underline{x}^* T (T^*T)^{-1} T^* \underline{x}$$

(*):

$$\underbrace{\Big(\sum_{i=1}^{n} \underline{t}_i \underline{t}_i^*\Big)}_{k \times k}{}_{rs} = \sum_{i=1}^{n} \underline{t}_{i_r} \underline{t}_{i_s} = (T^*T)_{rs}$$

c) $= \underline{x}^* T (T^* T)^{-1} T^* \underline{x}$

Dvs. $\hat{\sigma}^2 = \frac{1}{n} \left(\underline{x}^* \left(I_n - T(T^*T)^{-1} T^* \right) \underline{x} \right)$

5)

$r := \underline{x} - T\hat{\underline{\beta}} = \underline{x} - T(T^*T)^{-1} T^* \underline{x} = (I_n - T(T^*T)^{-1}T^*)\underline{x}$

Idet $\underline{x} \sim\sim N(T\beta, \sigma^2 I_n)$

og $(I_n - T(T^*T)^{-1}T^*)T\underline{\beta} = (T - T(T^*T)^{-1}(T^*T))\underline{\beta} = \underline{0}$

samt $(I_n - T(T^*T)^{-1}T^*)\sigma^2 I_n (I_n - T(T^*T)^{-1}T^*)^*$

$= \sigma^2 (I_n - T(T^*T)^{-1}T^*)$ følger af prop. 9.3, at

$r \sim\sim N(0, \sigma^2 (I_n - T(T^*T)^{-1}T^*))$

6)

Idet $\hat{\underline{\beta}} = (T^*T)^{-1}T^* \underline{x}$ og $\underline{x} \sim\sim N(T\underline{\beta}, \sigma^2 I_n)$ og

$(T^*T)^{-1}T^*T\underline{\beta} = \underline{\beta}$ samt at

$(T^*T)^{-1}T^* \sigma^2 I_n ((T^*T)^{-1}T^*)^* = \sigma^2 (T^*T)^{-1}T^*T(T^*T)^{-1}$

$= \sigma^2 (T^*T)^{-1}$

følger af prop. 9.3, at

$\hat{\underline{\beta}} \sim\sim N(\underline{\beta}, \sigma^2 (T^*T)^{-1})$

opgave 3.15

1)

$$T = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \qquad r \sim\sim N(0, \sigma^2 (I_n - \tfrac{1}{n}\begin{pmatrix} 1 & \cdots & 1 \\ \vdots & & \vdots \\ 1 & \cdots & 1 \end{pmatrix}))$$

$$\hat{\beta} \sim\sim N(\beta, \frac{\sigma^2}{n})$$

(Vi ser, at covariansmatricen for residualerne ikke er en diagonalmatrice, dvs. residualerne er korrelerede og dermed ikke uafh.)

2)

Antag at der er n_i obs. i den i'te obs. række, da har T k søjler og i den i'te søjle står der 1-taller i indgangene fra (ikke med) $\sum_{j<i} n_j$ til og med $\sum_{j \leq i} n_j$, 0 ellers.

$$\underset{k \times n}{T^*} \underset{n \times k}{T} = \begin{pmatrix} n_1 & & 0 \\ & n_i & \\ 0 & & n_k \end{pmatrix} \qquad (T^*T)^{-1} = \begin{pmatrix} \frac{1}{n_1} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{n_k} \end{pmatrix}$$

$T(T^*T)^{-1} =$ i'te søjle som i T multipliceret m. $\frac{1}{n_i}$

$$\underset{n \times k}{T(T^*T)^{-1}} \underset{k \times n}{T^*} = \begin{pmatrix} \frac{1}{n_1^2} & \cdots & \frac{1}{n_1^2} & & 0 \\ \vdots & & & & \\ \frac{1}{n_k^2} & \cdots & \frac{1}{n_k^2} & & \\ 0 & & & \ddots & \\ & & & & \frac{1}{n_k^2} \end{pmatrix}$$

Dvs. $r \sim\sim N(0, \sigma^2 \begin{pmatrix} 1-\frac{1}{n_1^2} & & & & \\ & 1-1/n_1^2 & & & \\ & & \ddots & & \\ & & & 1-1/n_k^2 & \\ & & & & -\frac{1}{n_k^2} \end{pmatrix})$

Altså er residualerne kun identisk fordelt, hvis $n_i = n_j \quad \forall i,j$

$$\hat{\beta} \sim\sim N(\beta, \begin{pmatrix} \sigma^2/n_1 & & 0 \\ & \ddots & \\ 0 & & \sigma^2/n_k \end{pmatrix})$$

3)

Idet vi går udfra at $\underline{\beta} = (\alpha, \beta)$ ser vi, at

T har 2 søjler, 1. søjle har kun 1-tallene i indgangene mens den i'te indgang i 2. søjle er t_i, hvor t_i er den forklarende variable til den i'te obs.

$$\underset{2\times n}{T^*}\ \underset{n\times 2}{T} = \begin{pmatrix} n & \Sigma t_i \\ \Sigma t_i & \Sigma t_i^2 \end{pmatrix} \qquad |T^* T| = n\Sigma t_i^2 - (\Sigma t_i)^2$$

$$(T^* T)^{-1} = \frac{1}{n\Sigma t_i^2 - (\Sigma t_i)^2}\begin{pmatrix} \Sigma t_i^2 & -\Sigma t_i \\ -\Sigma t_i & n \end{pmatrix}$$

F&B s.269
setn. 4.6

$$\underset{n\times 2}{T(T^* T)^{-1}} = \frac{1}{n\Sigma t_i^2 - (\Sigma t_i)^2}\begin{pmatrix} \Sigma t_j^2 - t_1 \Sigma t_j & n t_1 - \Sigma t_j \\ & \\ \Sigma t_j^2 - t_n \Sigma t_j & n t_n - \Sigma t_j \end{pmatrix}$$

$$\left(\underset{n\times n}{T(T^* T)^{-1} T^*}\right)_{rs} = \frac{1}{n\Sigma t_i^2 - (\Sigma t_i)^2}\left((\Sigma t_j^2 - t_r \Sigma t_j) + t_s (n t_r - \Sigma t_j)\right)$$

Der er ikke nogen simpel måde at skrive covariansmatricerne fra fordelingerne af hh. $\underline{\varepsilon}$ og $\hat{\underline{\beta}}$ – det ovenstående indsættes.

opg. 3.16

1)

Se udskriften

2) residual mod pred. tyder på en kvadratisk sammenhæng.

3) Se udskriften → 2)

4, 5) → se udskriften

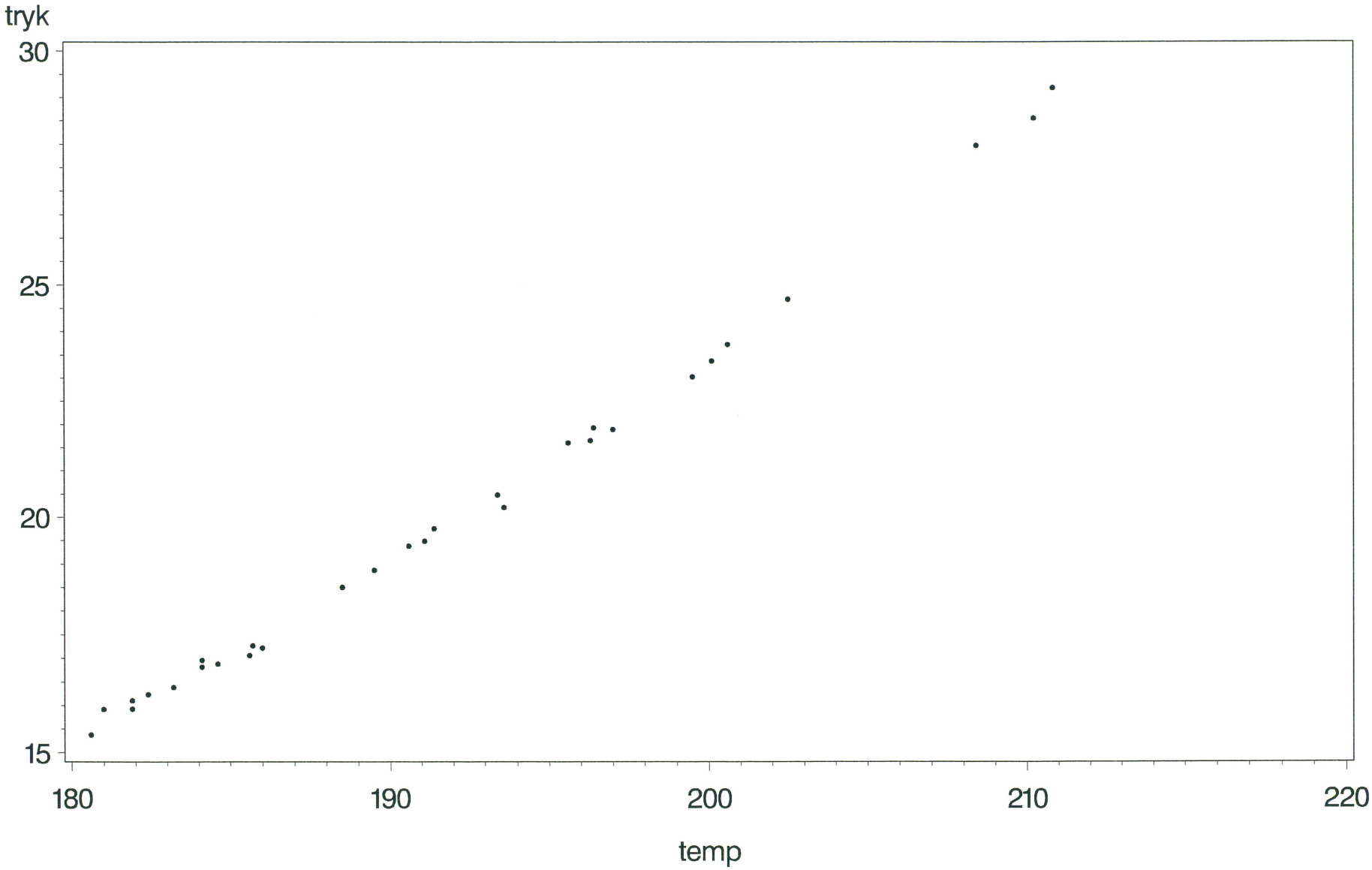
```
OPTIONS PS=45 LS=74 NODATE PAGENO=1;

DATA hooker;
INPUT temp tryk @@;
DATALINES;
18 6 15.376 181.0 15.919 181.9 16.106 181.9 15.928 182.4 16.235
183.2 16.385 184.1 16.959 184.1 16.817 184.6 16.881 185.6 17.062
185.7 17.267 186.0 17.221 188.5 18.507 189.5 18.869 189.5 18.869
190.6 19.386 191.1 19.490 191.4 19.758 193.4 20.480 193.6 20.212
195.6 21.605 196.3 21.654 196.4 21.928 197.0 21.892 199.5 23.030
200.1 23.369 200.6 23.726 202.5 24.697 208.4 27.972 210.2 28.559
210.8 29.211
;

/* nogle GOPTIONS */
GOPTIONS DEV=win CTEXT=black GACCESS=gsasfile GSFMODE=replace;
GOPTIONS TARGETDEVICE=PS300A4 ROTATE=landscape
         HORIGIN=2 cm VORIGIN=2 cm
         CBACK=white GUNIT=pct HTITLE=4 HTEXT=2.7
         FTEXT=swiss CTEXT=black CSYMBOL=black;
OPTIONS NODATE PAGESIZE=45;
PA  off;

OPTIONS PAGENO=1 LS=80;


AXIS1  ORDER=(180 to 220 by 10)  WIDTH=2 LENGTH=85 pct;
AXIS2  ORDER=(15 to 30 by 5)     WIDTH=2 LENGTH=70 pct;


/*-------------------------------------------*/

TITLE1 'tryk mod temperatur';
PROC GPLOT DATA=hooker;
PLOT tryk*temp/
HAXIS=axis1 VAXIS=axis2;
SYMBOL1 i=none  v=dot  c=black;
RU  QUIT;

TITLE1 'tryk mod temperatur - lineaert fit';
PROC GPLOT DATA=hooker;
PLOT tryk*temp/
HAXIS=axis1 VAXIS=axis2;
SYMBOL1 I=rl  V=dot  C=black;
RUN; QUIT;
/* I=rl (interpolate=regressionsline) tegner estimeret regressionslinje.
   Den estimerede regressionslinje angives i log-vinduet.
*/


PROC GLM DATA=hooker;
MODEL tryk=temp/SS1;
OUTPUT OUT=udlin P=pred R=resid STUDENT=stud_res;
RU
/* variablene resid og stud_res i datasaettet udlin indeholder
   henholdsvis residualer og normerede (studentized) residualer
   for en specificerede model
*/
```

```
TITLE1 'residual mod predikteret - lineaert fit';
PROC GPLOT DATA=udlin;
PLOT resid*pred;
SYMBOL1 I=none  V=dot  C=black;
RU

TITLE1 'tryk mod temperatur - linaert fit';
TITLE2 '95% konfidensinterval for middelvaerdien';
PROC GPLOT DATA=hooker;
PLOT tryk*temp/
HAXIS=axis1 VAXIS=axis2;
SYMBOL1 I=rlclm95  V=dot  C=black;
RUN; QUIT;

TITLE1 'tryk mod temperatur - lineaert fit';
TITLE2 '95% prediktionsinterval for ny observation';
PROC GPLOT DATA=hooker;
PLOT tryk*temp/
HAXIS=axis1 VAXIS=axis2;
SYMBOL1 I=rlcli95  V=dot  C=black;
RU   QUIT;

TITLE1 'Fraktildiagram af residualer fra lineaert fit';
PROC CAPABILITY DATA=udlin GRAPHICS;
VAR resid;
   QQPLOT resid/ NORMAL(MU=EST SIGMA=EST);
RUN; QUIT;
/* Ude i verden klarer man sig uden kenskab til normgraf-macroen
   for eksempel ved hjaelp af PROC CAPABILITY.
   Vores fraktildiagram hedder et QUANTILE-QUANTILE plot (QQ).
   Hvis man vil se efter flere muligheder i PROC CAPABILITY i HELP MENUEN,
   skal man se under QUALITY IMPROVEMENT, mens de fleste andre procedurer
   vi bruger findes under DATA ANALYSIS.
*/

/*-----------------------------------------------------------*/
/* PROC CAPABILITY kan ikke lave fraktildiagrammer for t-fordelte data,
   men det kraever kun nogle faa ekstra kommandoer at lave et saadant selv
*/

DATA stud_res;
SET udlin;
n_tot=31;
KEEP n_tot stud_res;
RUN;


PROC SORT DATA=stud_res;
BY  stud_res;
RUN;

DATA stud_res;
SET stud_res;
t_akt=tinv(((_n_ -.5)/n_tot),29);

TITLE1 'Fraktildiagram af normerede residualer fra lineaert fit';
PROC GPLOT DATA=stud_res;
```

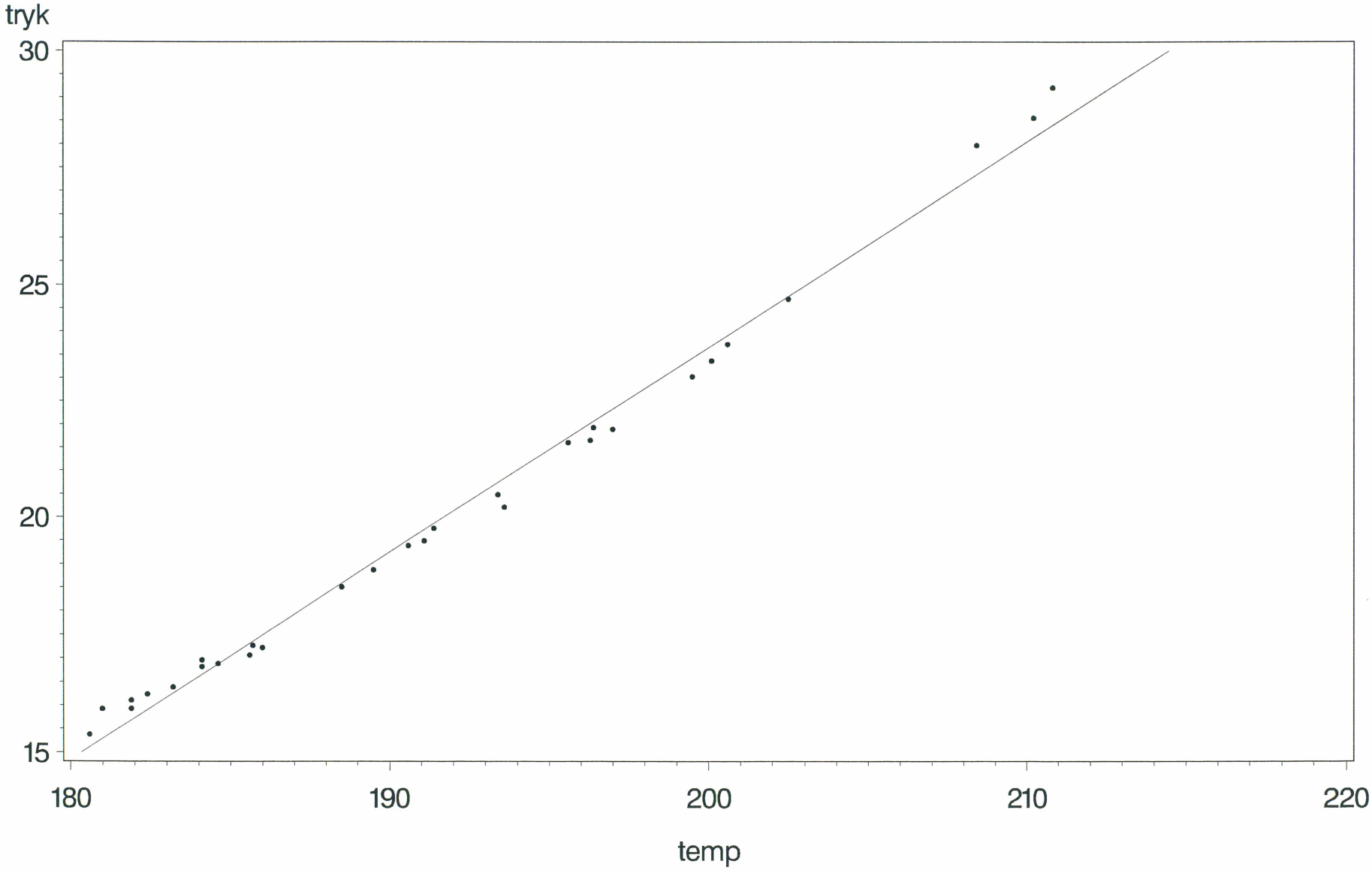
```
   PLOT t_frakt*stud_res;
SYMBOL1 I=rl  V=dot  C=black;
LABEL stud_res='normerede residualer'
      t_frakt='t-fraktil';
RUN; QUIT;

/*-------------------------------------------------------*/


PROC GLM DATA=hooker;
MODEL tryk=temp temp*temp/SS1;
OUTPUT OUT=udkvad p=pred r=resid student=stud_res;
RUN;

TITLE1 'residual mod predikteret - kvadratisk fit';
PROC GPLOT DATA=udkvad;
PLOT resid*pred;
SYMBOL1 I=none  V=dot  C=black;
RUN;

TITLE1 'Fraktildiagram af residualer fra kvadratisk fit';
PROC CAPABILITY DATA=udkvad GRAPHICS;
VAR resid;
   QQPLOT resid/ NORMAL(MU=EST SIGMA=EST);
RUN; QUIT;

/*-------------------------------------------------------*/

DATA stud_res;
SET udkvad;
n_tot=31;
KEEP n_tot stud_res;
RUN;


PROC SORT DATA=stud_res;
BY  stud_res;
RUN;

DATA stud_res;
SET stud_res;
t_frakt=tinv(((_n_ -.5)/n_tot),28);

TITLE1 'Fraktildiagram af normerede residualer fra kvadratisk fit';
PROC GPLOT DATA=stud_res;
   PLOT t_frakt*stud_res;
SYMBOL1 i=rl  v=dot  c=black;
LABEL stud_res='normerede residualer'
      t_frakt='t-fraktil';
RUN; QUIT;

/*-------------------------------------------------------*/

TITLE1 'tryk mod temperatur - kvadratisk fit';
PROC GPLOT DATA=hooker;
PLOT tryk*temp/
HAXIS=axis1 VAXIS=axis2;
SYMBOL1 I=rq  V=dot  C=black;
```

```
RUN; QUIT;

TITLE1 'tryk mod temperatur - kvadratisk fit';
TITLE2 '95% konfidensinterval for middelvaerdien';
PROC GPLOT DATA=hooker;
PL   tryk*temp/
HAXIS=axis1 VAXIS=axis2;
SYMBOL1 I=rqclm95  V=dot  C=black;
RUN; QUIT;

TITLE1 'tryk mod temperatur - kvadratisk fit';
TITLE2 '95% prediktionsinterval for ny observation';
PROC GPLOT DATA=hooker;
PLOT tryk*temp/
HAXIS=axis1 VAXIS=axis2;
SYMBOL1 I=rqcli95  V=dot  C=black;
RUN; QUIT;


TITLE1;
TITLE2;
TI  E3;
TITLE4;
TITLE5;
footnote1;
footnote2;
```

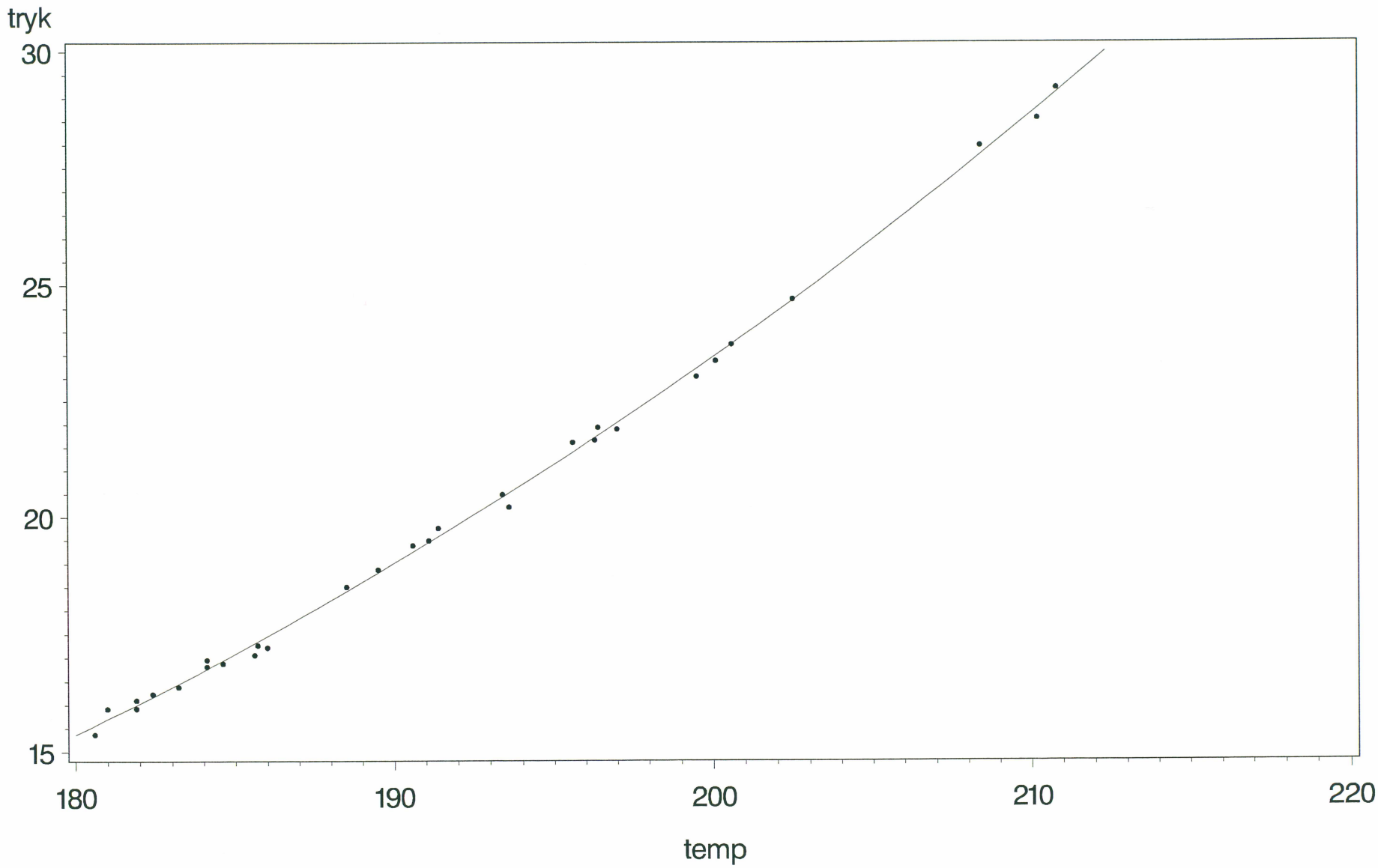
tryk mod temperatur



tryk mod temperatur – lineaert fit

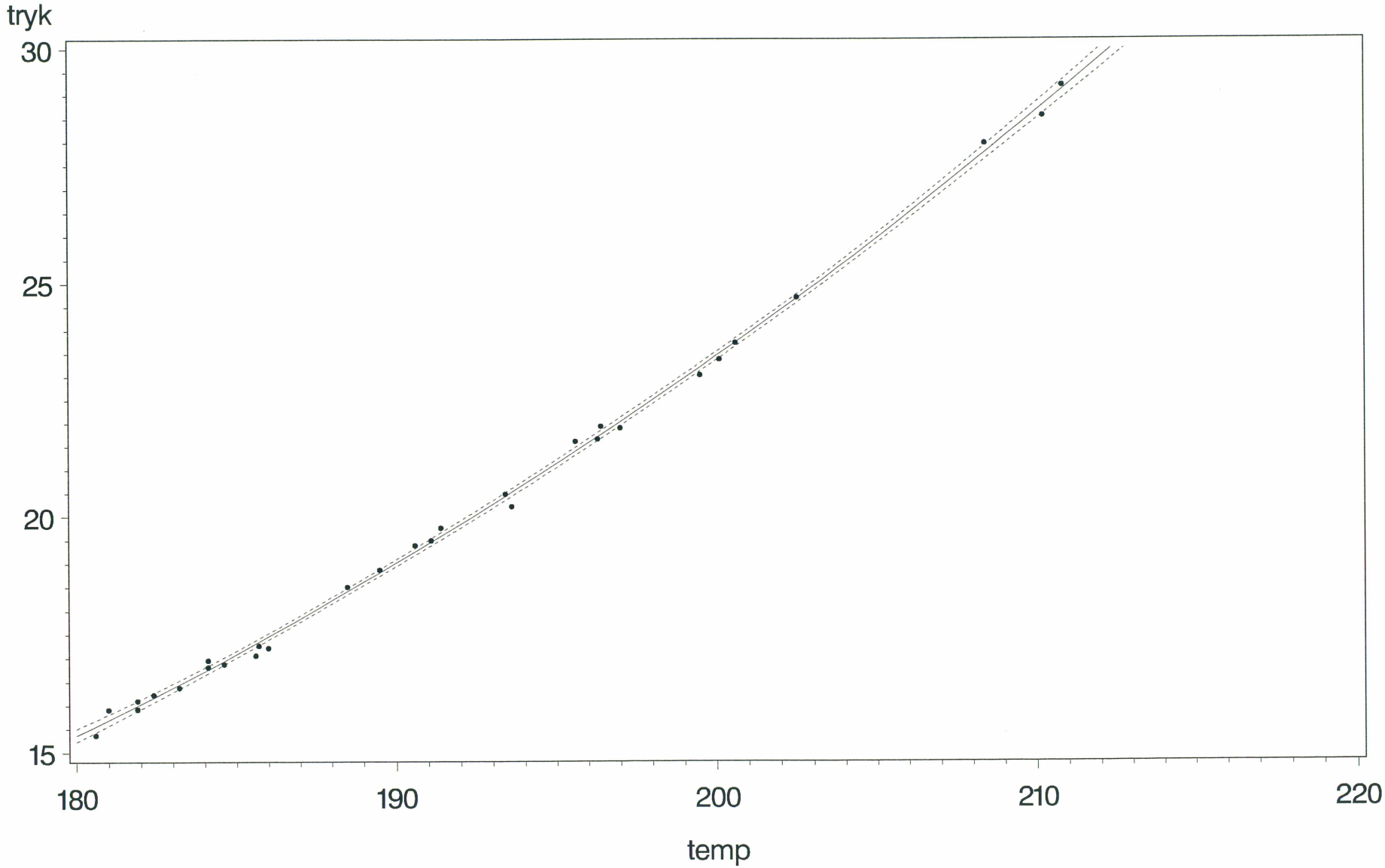


tryk mod temperatur – kvadratisk fit



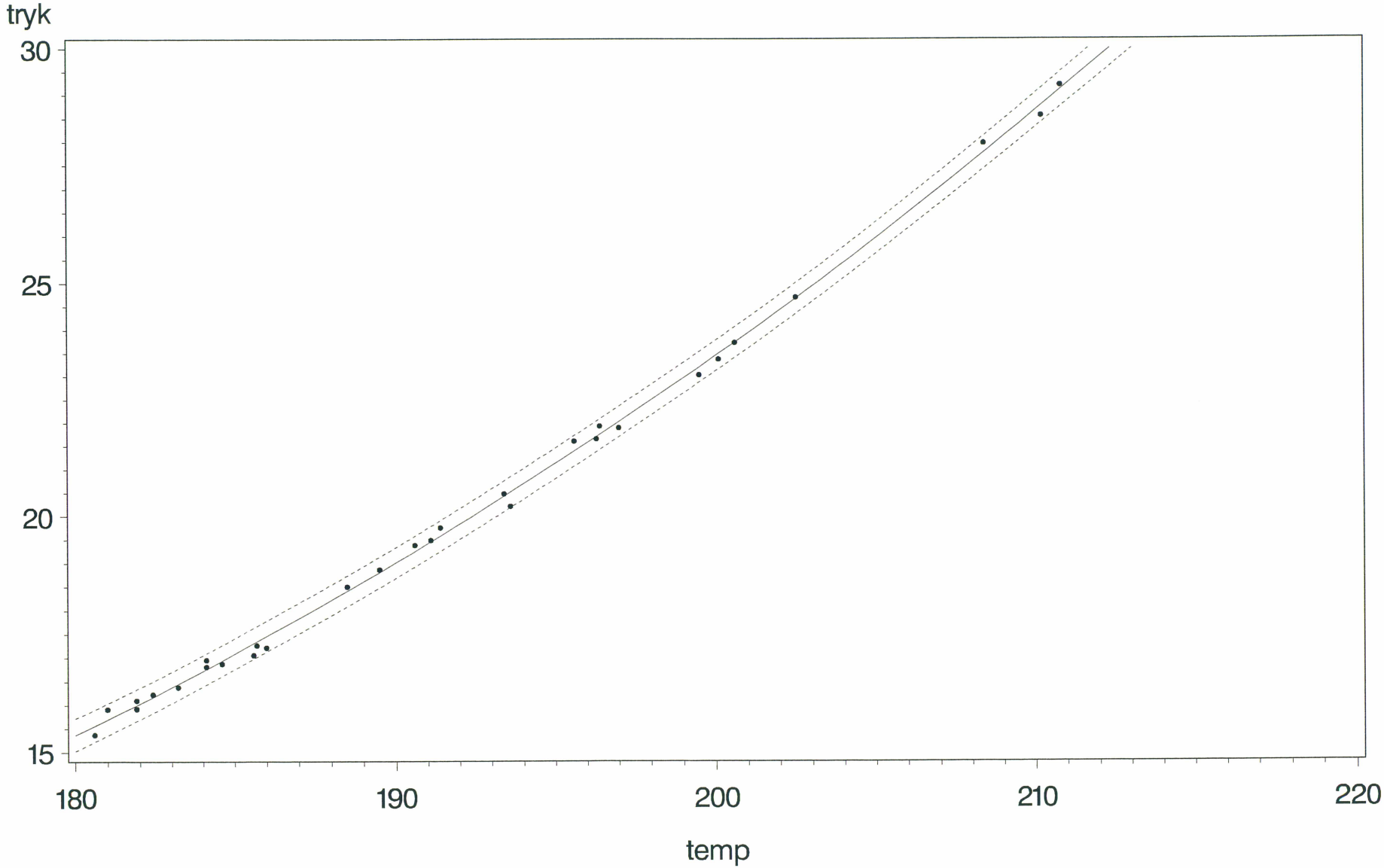
tryk mod temperatur – kvadratisk fit

95% konfidensinterval for middelvaerdien



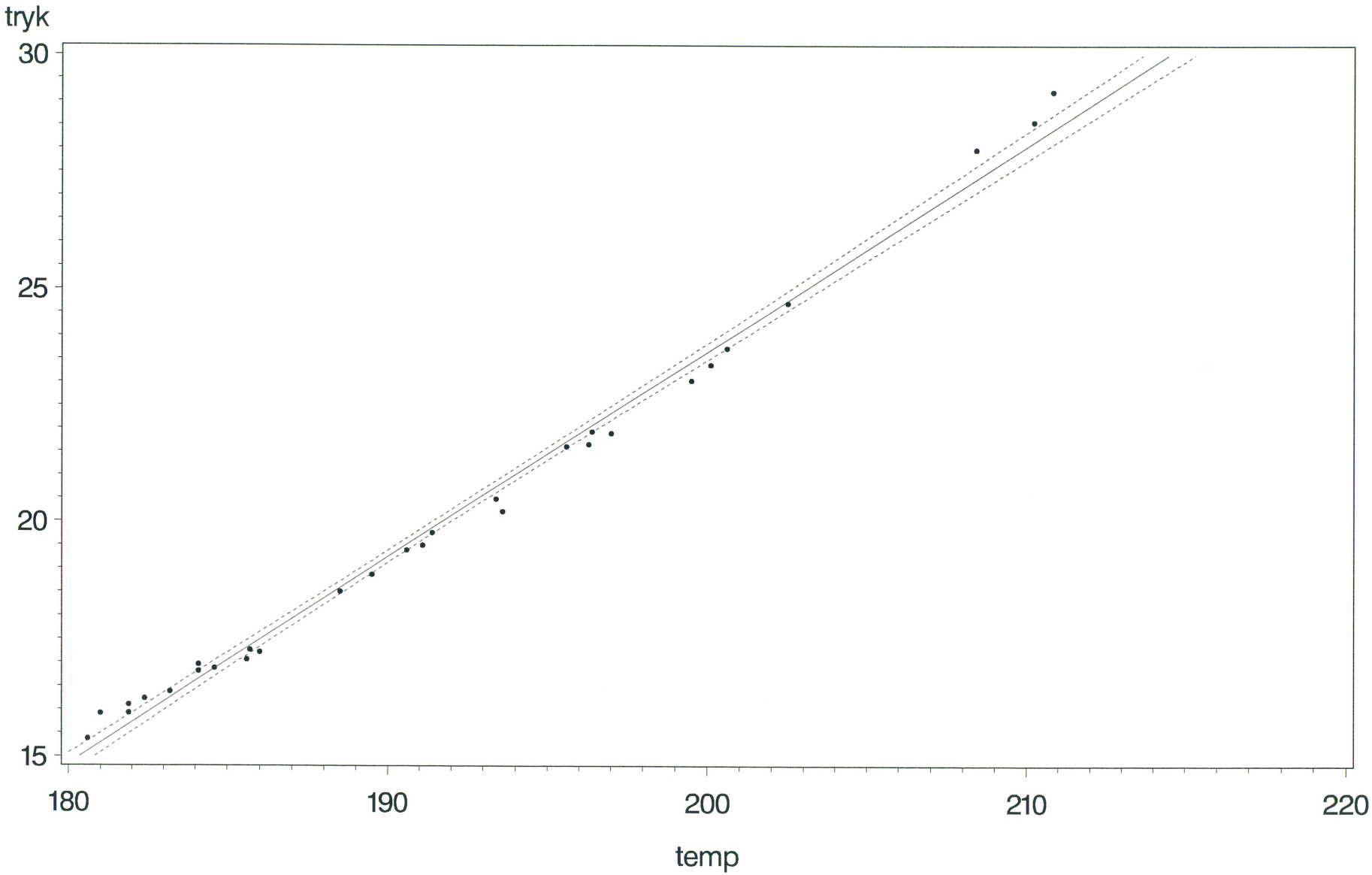
tryk mod temperatur – kvadratisk fit

95% prediktionsinterval for ny observation



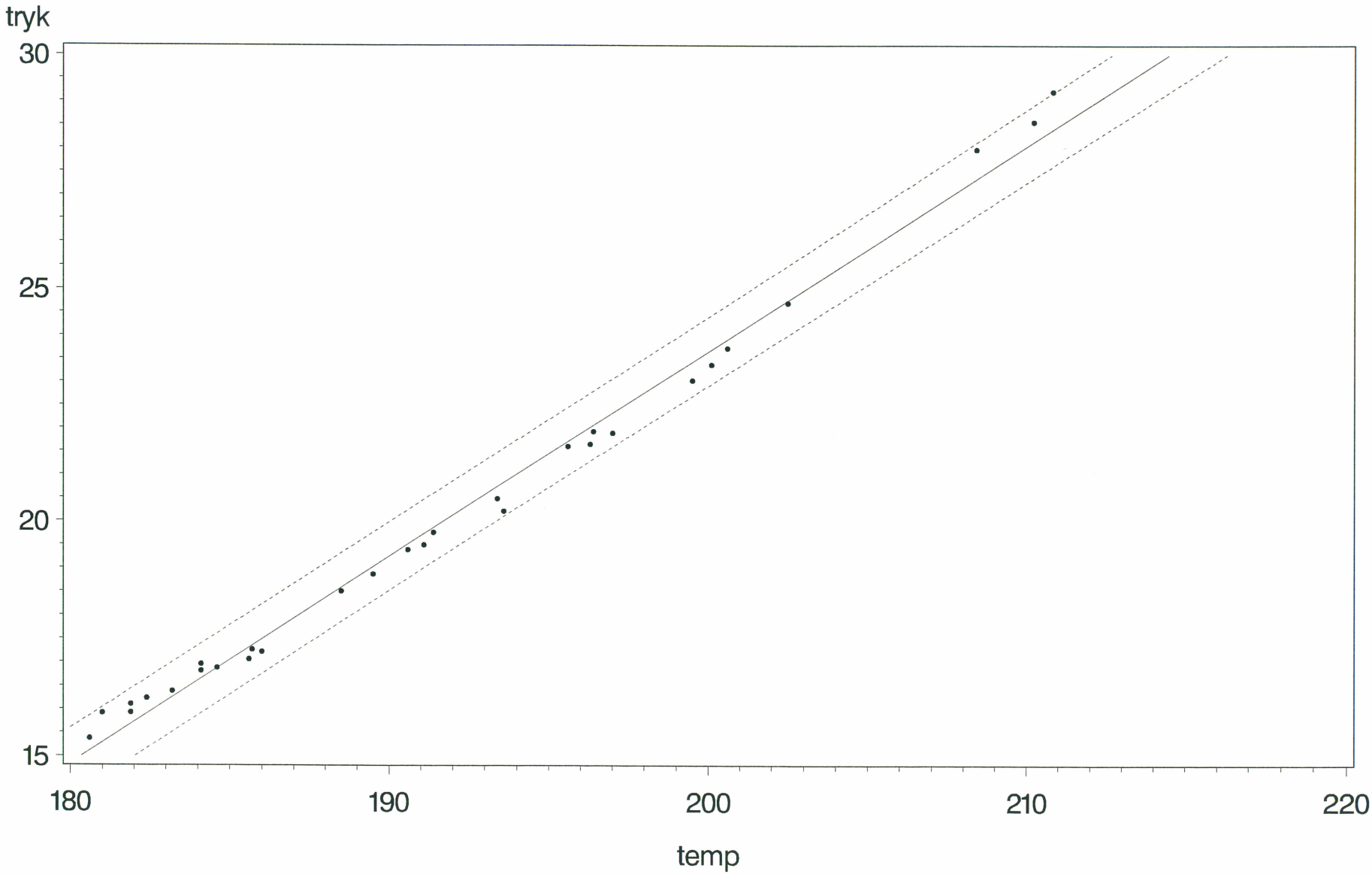
tryk mod temperatur – linaert fit

95% konfidensinterval for middelvaerdien

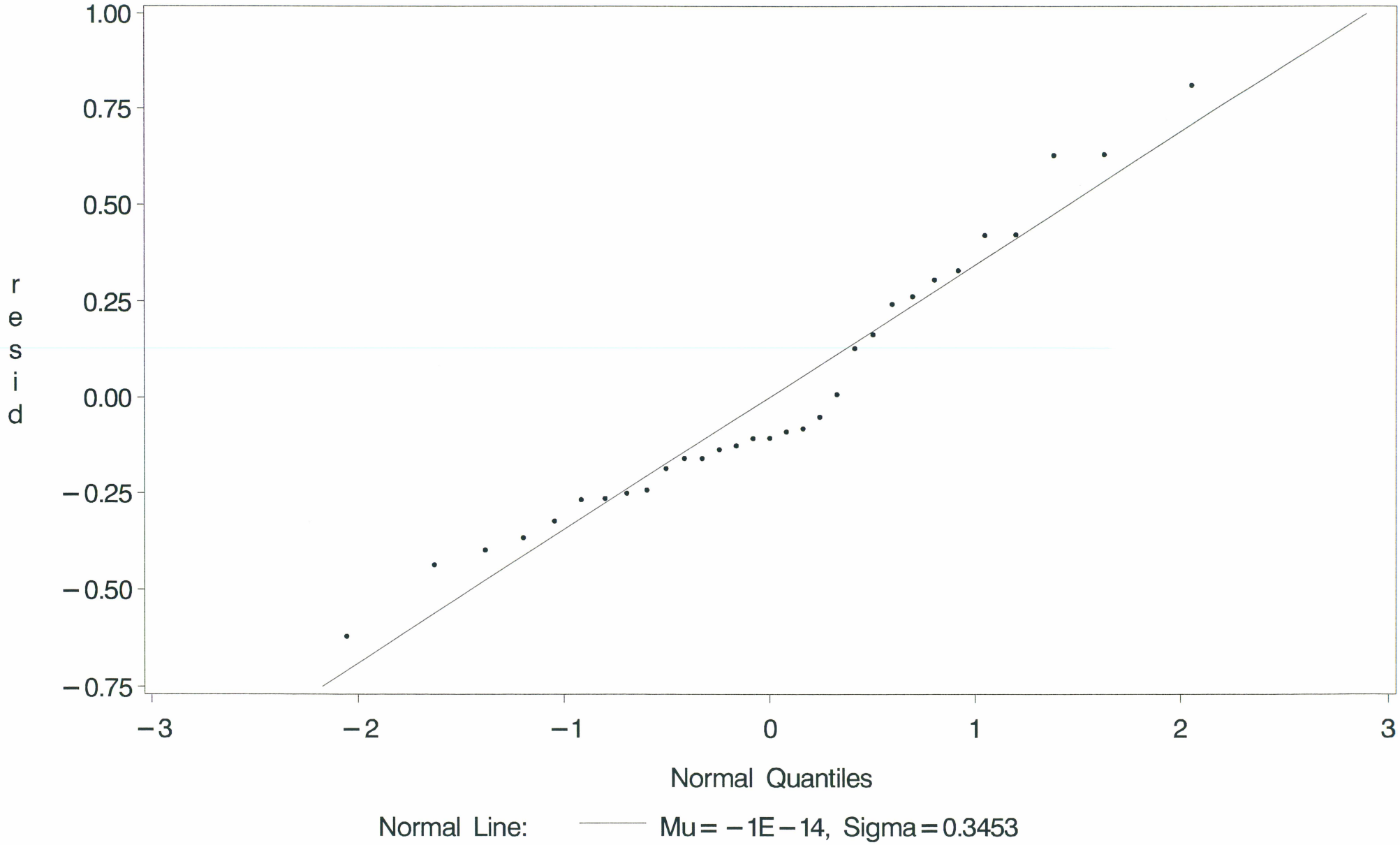


tryk mod temperatur – lineaert fit

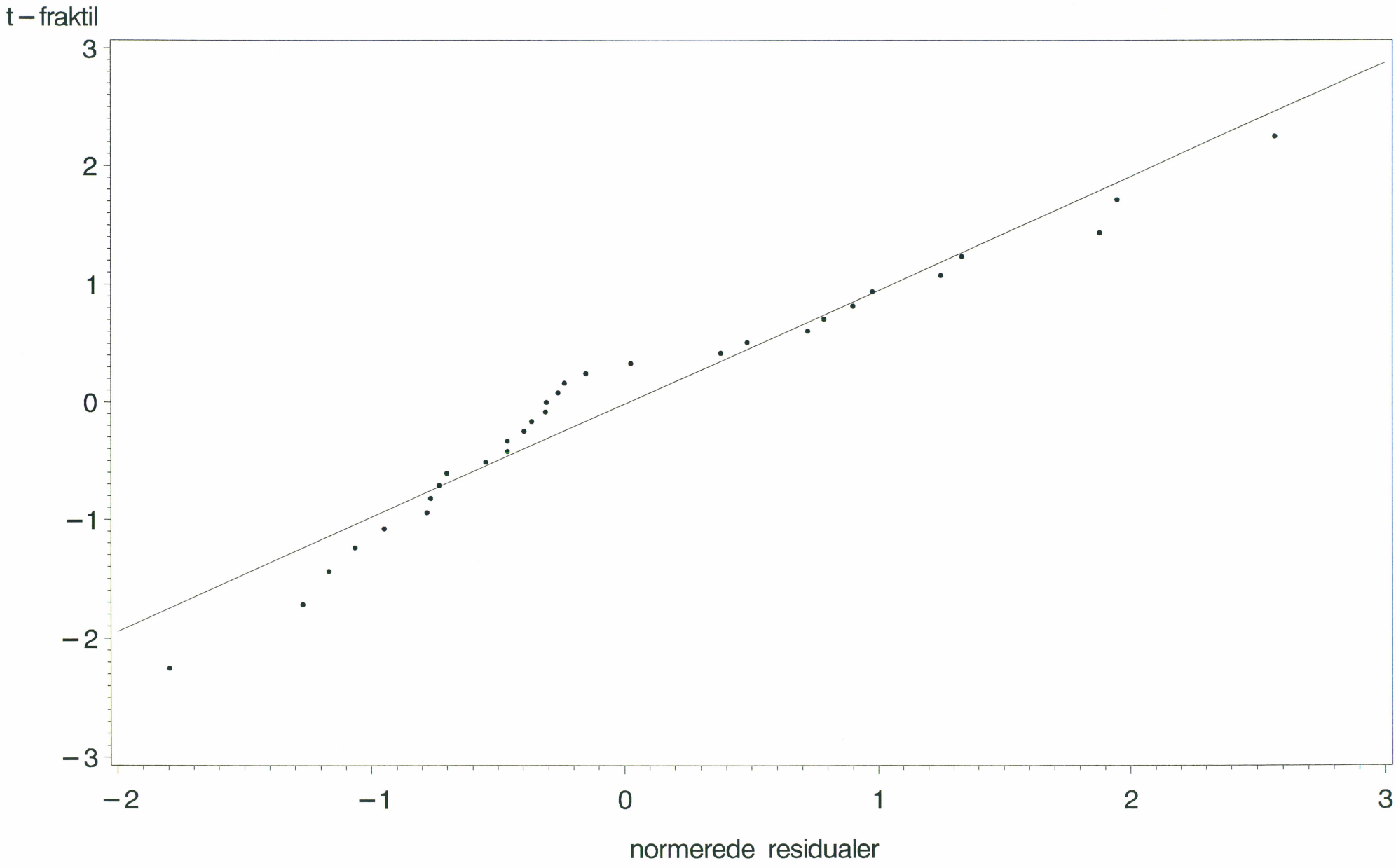
95% prediktionsinterval for ny observation



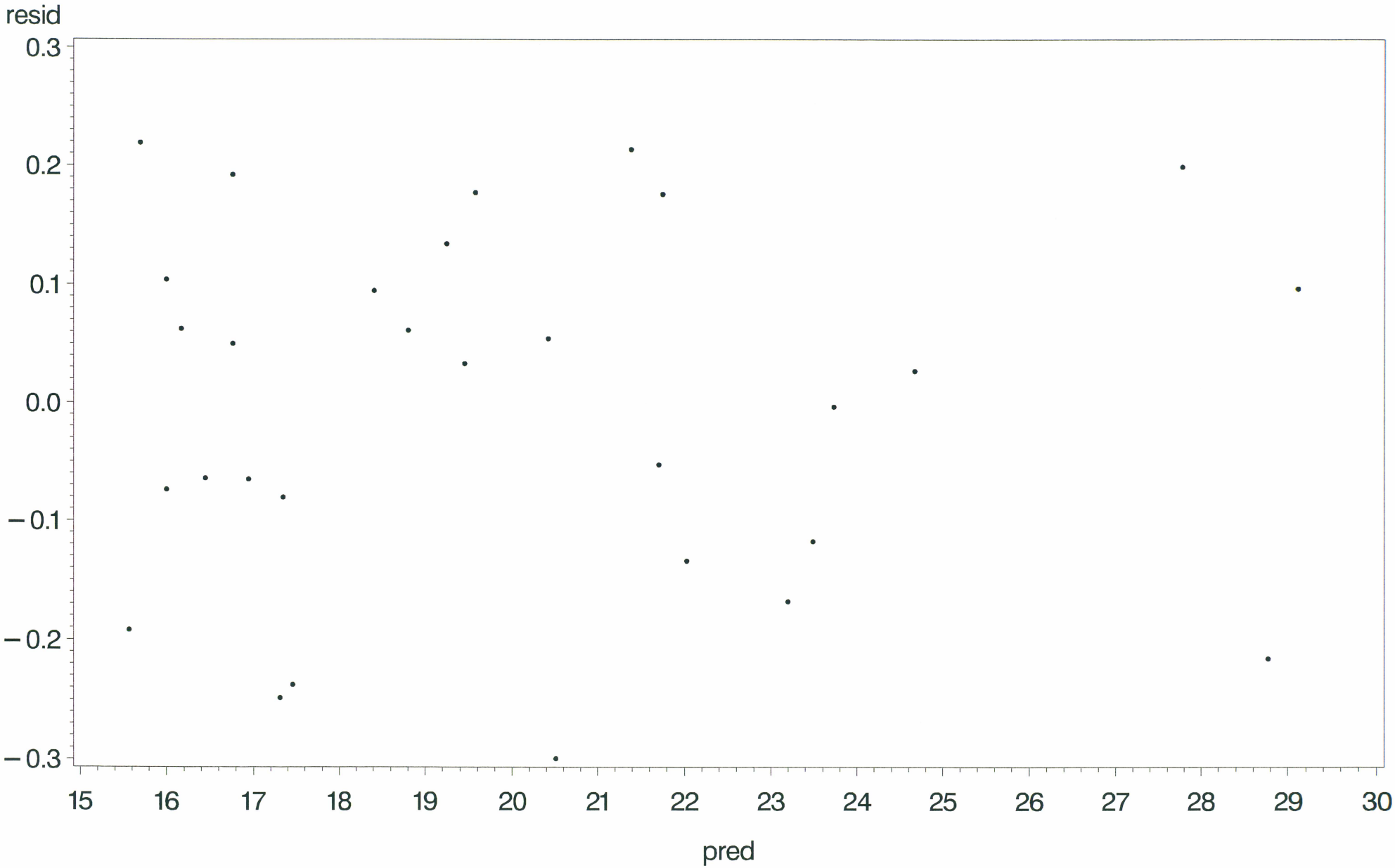
Fraktildiagram af residualer fra lineaert fit



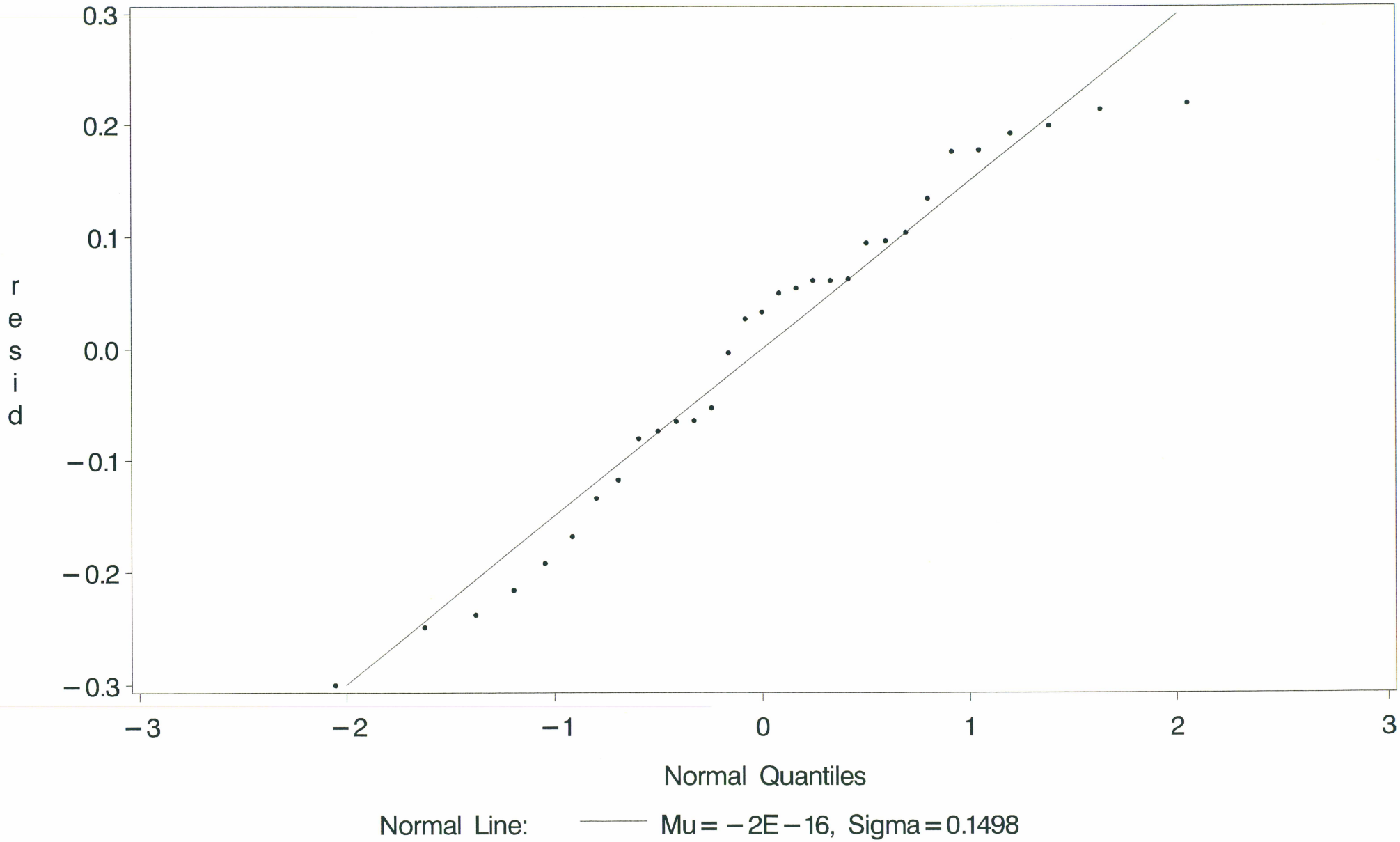
Fraktildiagram af normerede residualer fra lineaert fit



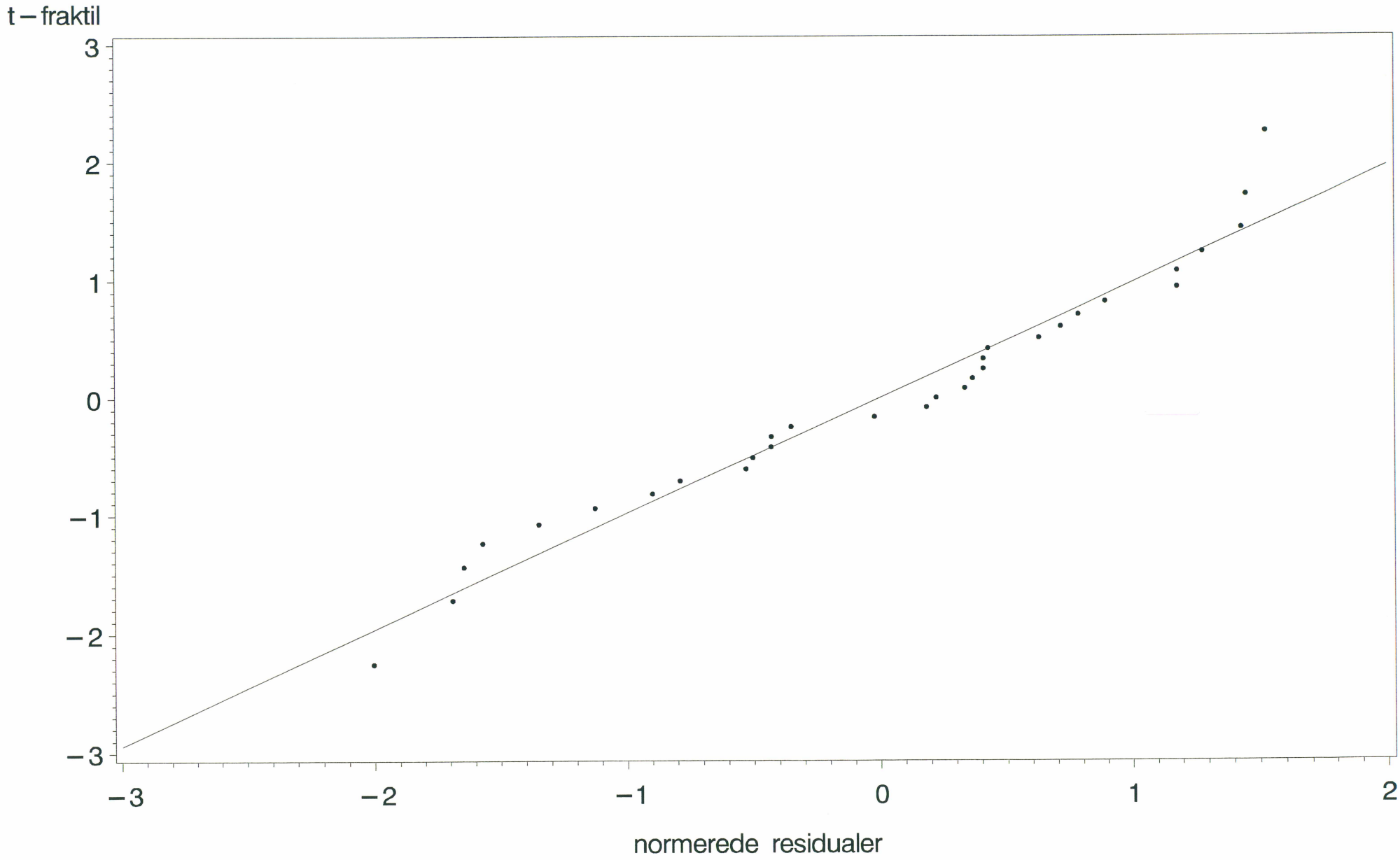
residual mod predikteret – kvadratisk fit



Fraktildiagram af residualer fra kvadratisk fit



Fraktildiagram af normerede residualer fra kvadratisk fit



The GLM Procedure

Dependent Variable: tryk

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	442.8124656	442.8124656	3590.24	<.0001
Error	29	3.5767972	0.1233378		
Corrected Total	30	446.3892628			

R-Square	Coeff Var	Root MSE	tryk Mean
0.991987	1.752103	0.351195	20.04419

Source	DF	Type I SS	Mean Square	F Value	Pr > F
temp	1	442.8124656	442.8124656	3590.24	<.0001

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	-64.34555959	1.40981805	-45.64	<.0001
temp	0.43996609	0.00734273	59.92	<.0001

The CAPABILITY Procedure
Variable: resid

Moments

N	31	Sum Weights	31
Mean	-1.094E-14	Sum Observations	-3.393E-13
Std Deviation	0.34529201	Variance	0.11922657
Skewness	0.6347196	Kurtosis	-0.1718209
Uncorrected SS	3.57679724	Corrected SS	3.57679724
Coeff Variation	-3.1549E15	Std Error Mean	0.06201628

Basic Statistical Measures

Location		Variability	
Mean	-0.00000	Std Deviation	0.34529
Median	-0.10595	Variance	0.11923
Mode	-0.15901	Range	1.43158
		Interquartile Range	0.51383

Tests for Location: Mu0=0

Test	-Statistic-		-----p Value------	
Student's t	t	-176E-15	Pr > |t|	1.0000
Sign	M	-3.5	Pr >= |M|	0.2810
Signed Rank	S	-12	Pr >= |S|	0.8185

Quantiles (Definition 5)

Quantile	Estimate
100% Max	0.811708543
99%	0.811708543
95%	0.630697923
90%	0.423688195
75% Q3	0.263684357
50% Median	-0.105949377
25% Q1	-0.250146075

The CAPABILITY Procedure
Variable: resid

Quantiles (Definition 5)

Quantile	Estimate
10%	-0.365783202
5%	-0.435759462
1%	-0.619874768
0% Min	-0.619874768

Extreme Observations

--------Lowest-------		-------Highest-------	
Value	Obs	Value	Obs
-0.619874768	20	0.421728445	3
-0.435759462	24	0.423688195	30
-0.397674679	25	0.628627151	29
-0.365783202	22	0.630697923	2
-0.322654330	26	0.811708543	31

The GLM Procedure

De ndent Variable: tryk

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	445.7163234	222.8581617	9272.79	<.0001
Error	28	0.6729395	0.0240336		
Corrected Total	30	446.3892628			

R-Square	Coeff Var	Root MSE	tryk Mean
0.998492	0.773429	0.155028	20.04419

Source	DF	Type I SS	Mean Square	F Value	Pr > F
temp	1	442.8124656	442.8124656	18424.8	<.0001
temp*temp	1	2.9038578	2.9038578	120.83	<.0001

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	85.65333116	13.66030896	6.27	<.0001
temp	-1.10484296	0.14057613	-7.86	<.0001
temp*temp	0.00396884	0.00036106	10.99	<.0001

The CAPABILITY Procedure
Variable: resid

Moments

N	31	Sum Weights	31
Mean	-1.719E-16	Sum Observations	-5.329E-15
Std Deviation	0.14977088	Variance	0.02243132
Skewness	-0.3271156	Kurtosis	-0.9002559
Uncorrected SS	0.67293947	Corrected SS	0.67293947
Coeff Variation	-8.7124E16	Std Error Mean	0.02689964

Basic Statistical Measures

Location		Variability	
Mean	-0.00000	Std Deviation	0.14977
Median	0.03310	Variance	0.02243
Mode	0.06122	Range	0.51876
		Interquartile Range	0.22185

Tests for Location: Mu0=0

Test	-Statistic-		-----p Value------	
Student's t	t	-639E-17	Pr > |t|	1.0000
Sign	M	1.5	Pr >= |M|	0.7201
Signed Rank	S	0	Pr >= |S|	1.0000

Quantiles (Definition 5)

Quantile	Estimate
100% Max	0.2189394570
99%	0.2189394570
95%	0.2135125043
90%	0.1919750841
75% Q3	0.1040339183
50% Median	0.0331034099
25% Q1	-0.1178165906

The CAPABILITY Procedure
Variable: resid

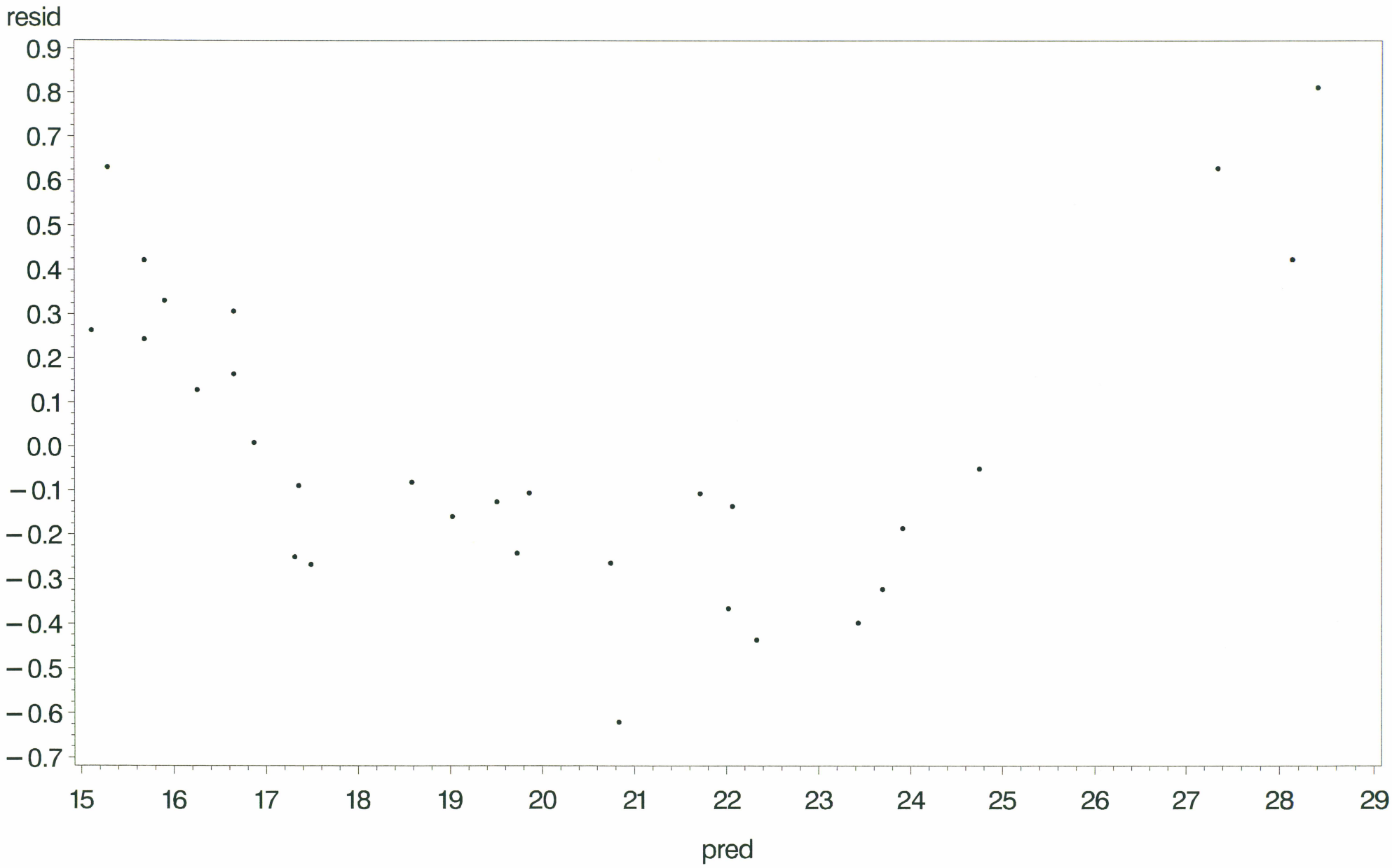
Quantiles (Definition 5)

Quantile	Estimate
10%	-0.2159121001
5%	-0.2486830245
1%	-0.2998250945
0% Min	-0.2998250945

Extreme Observations

--------Lowest-------		-------Highest-------	
Value	Obs	Value	Obs
-0.299825095	20	0.177131426	18
-0.248683024	10	0.191975084	7
-0.237674842	12	0.198815295	29
-0.215912100	30	0.213512504	21
-0.191944103	1	0.218939457	2

residual mod predikteret – lineaert fit



Afdeling for Teoretisk Statistik
Institut for Matematiske Fag
Aarhus Universitet

Statistik 1 og Statistik α
Jørgen Granfeldt
25. oktober 2002

Meddelelse 8

Forelæsningerne i sidste uge

(Onsdag den 9. oktober) Begyndte på Afsnit 3.3 *Lineær regression.* Blev næsten færdig med afsnittet om beregninger, som begynder side 3.74.

(Fredag den 11. oktober) Fortsatte med lineær regression og blev næsten helt færdig med afsnittet. Næste emne er Kapitel 4 *Lineære normale modeller.*

Forelæsningerne i denne uge

(Onsdag den 23. oktober) Afslutter lineær regression og begyndte på Kapitel 4 *Lineære normale modeller*, hvor jeg nåede til afsnit 4.1.2 *Test.*

(Fredag den 25. oktober) Kapitel 4 *Lineære normale modeller.*

Forelæsningerne i næste uge

(Onsdag den 30. oktober og fredag den 1. november) Efter Kapitel 4 begynder jeg på Kapitel 6.

Øvelserne i denne uge

(Uge 43: 21. oktober-27. oktober) 3.12, 3.17 (regnes færdig med brug af lineær regression), 3.20, 4.5 (Kun spørgsmålene 1), 2) og 3)).

Øvelserne i de to næste uger

(Uge 44: 28. oktober-3. november) 3.14, 3.15, 3.16.

(Uge 45: 4. november-10. november) 4.3, 4.4, 4.5 (regnes færdig).

(Uge 46: 11. november-17. november) 4.6, 4.8, 4.9, 4.10.

Afleveringsopgaver

(Uge 44) Ingen afleveringsopgave.

(Uge 45) Ingen afleveringsopgave.

Udleveret materiale

Intet

Nogle SAS vink fra Introduktion til SAS

Danske bogstaver i grafik kan man bruge hvis man fortæller hvilket

tegnsæt man bruger. På Unix skriver man:

```
GOPTIONS KEYMAP=ISO8859 DEVMAP=ISO8859;
```

og i Windows versionen skriver man

```
GOPTIONS KEYMAP=winansi DEVMAP=winansi;
```

På Unix skriver man:

```
GOPTIONS DEV=xcolor;
```

for at få vist grafik i et vindue. Den tilsvarende kommando i Windows er

```
GOPTIONS DEV=win;
```

Hjælp i SAS:

Ofte har man bare brug for at blive mindet om syntaksen for en procedure, for eksempel PROC GPLOT. I saa fald skriver man `help gplot` i kommandovinduet i TOOLBOX vinduet, og hjælp åbner det relevante sted.

Udskrivning af grafik fra SAS

er fuldstændig problemfrit! Læs eksemplerne 32, 33 og 34 i den blå introduktion til SAS.

Udskrivning af udskrifter fra SAS på postscript printer

Man kan bruge `a2ps`.

```
a2ps -PH2 fil
```

Som default klemmes to sider sammen på et ark, men man kan selv være herre over det, og også kontrollere fontstørrelsen.

```
a2ps -F15 -nP -1 -p output.sas > outputa.ps
```

Denne kommando laver ascii filen `output.lst` til postscript filen `outputa.lst` og trykker den ikke (`-nP`).

Kun 1 side per ark (`-1`), og i portrait (`-p`).

`-F15` vælger fontstørrelsen.

`-F15` er fint til transparanter, men er lidt stort til almindelig udskrivning.

Man kan desværre ikke stole på at `a2ps` virker på Linux.

STAT1, TØ nr. 7

opgave 3.12

$X_i \sim N(\mu, \sigma^2) \quad i = 1, \ldots, n$

$Y \sim N(\mu, \sigma^2) \quad$ uafh.

1)

Påstand: $\dfrac{Y - \bar{X}_\cdot}{s(X)\sqrt{1+1/n}} \sim t(n-1)$

Bevis:

Jvf. opg. 4.10: $\bar{X}_\cdot$, $s(X)$, Y uafh. (*)

$Y - \bar{X}_\cdot \sim N(0, \sigma^2 + \frac{1}{n}\sigma^2)$ jvf. 3.66-67

$\Downarrow$ 3.66

$\dfrac{Y - \bar{X}_\cdot}{\sqrt{1+1/n}} \sim N(0,1)$

Jvf. SS2 er $(Y - \bar{X}_\cdot)/\sqrt{1+1/n} \perp\!\!\!\perp s(X)$ pga. (*)
da $(n-1)s^2(X) \sim \sigma^2 \chi^2(n-1)$ jvf. 3.78 følger det
ønskede af def. s. 3.116.

$\square$

Dvs. med ss. $1-\alpha$ holder flg. (t-ford. symmetrisk)

$$\bar{X}_\cdot - t_{1-\alpha/2}(n-1)\, s(X)\sqrt{1+1/n} \le Y \le \bar{X}_\cdot + t_{1-\alpha/2}(n-1)\, s(X)\sqrt{1+1/n}$$

(↑ et prediktionsinterval, uden vedier af X indsættes)

2)

Jf. s.5 i Stat. Tables : $t_{1-\alpha/2}(n-1) \xrightarrow[n\to\infty]{} 1{,}96$ $(\alpha = 0{,}05)$

Jf. Afl. 19 i SS2 : $s(\underline{X}) \xrightarrow[n\to\infty]{} \sigma$, $\bar{X}. \xrightarrow[n\to\infty]{} \mu$

Og $\sqrt{1+1/n} \xrightarrow[n\to\infty]{} 1$ Alt i alt kan vi konkludere, at

i grænsen gælder, at $Y \in [\mu - 2\sigma, \mu + 2\sigma]$ m. ss. 95%.

konfidensintervallet for μ (3.19) udtrykker at i grænsen bestemmer $\bar{x}.$ middelværdien μ med stor precision.

3)

Jf. 3.3.3 : $\hat{\alpha} + \hat{\beta} t \sim\sim N(\alpha + \beta t, \sigma^2 (\frac{1}{n} + \frac{(t+\bar{t}.)^2}{SSD_t}))$

og $S_{02}^2 \sim\sim \sigma^2 \chi^2(f_{02}) / f_{02}$, $f_{02} = n-2$

Dvs. hvis $Y \sim N(\alpha + \beta t, \sigma^2)$, gælder $(\hat{\alpha}(\underline{X}), \hat{\beta}(\underline{X}))$

betegner estimatorerne hvori vi har indsat s.v.) :

$Y - (\hat{\alpha}(\underline{X}) + \hat{\beta}(\underline{X}) t) \sim N(0, \sigma^2 (1 + \frac{1}{n} + \frac{(t+\bar{t}.)^2}{SSD_t}))$

Er $Y \perp\!\!\!\perp X_1, \ldots, X_n$ gælder jf. opg. 4.11 , at

$Y - (\hat{\alpha}(\underline{X}) + \hat{\beta}(\underline{X}) t) \perp\!\!\!\perp S_{02}^2(\underline{X})$, af def. s. 3.116 følger, at

$$\frac{Y - (\hat{\alpha}(\underline{X}) + \hat{\beta}(\underline{X}) t)}{\sqrt{1 + \frac{1}{n} + \frac{(t-\bar{t}.)^2}{SSD_t}}\; S_{02}(\underline{X})} \sim t(n-2)$$

Hvilket giver flg. $1-\alpha$ prediktionsinterval for Y:

$$\left[\hat{\alpha}+\hat{\beta}t-\sqrt{1+\frac{1}{n}+\frac{(t-\bar{t})^2}{SSD_t}}\,s_{0_2}\,t_{1-\alpha/2}(n-2);\ \hat{\alpha}+\hat{\beta}t+\sqrt{1+\frac{1}{n}+\frac{(t-\bar{t})^2}{SSD_t}}\,s_{0_2}\,t_{1-\alpha/2}(n-2)\right]$$

Fortolkningen er, at måler vi mange værdier med t som forklarende var. (eller måler flere på t) og har vi en stor stikprøve, da vil obs. af Y ligge i int $[\alpha+\beta t-2\sigma;\ \alpha+\beta t+2\sigma]$ med stor ($\sim 95\%$) ss.

opgave 3.20

1)

Jvf. s. 3.22 : $\ell(\mu) = -\frac{n}{2}\ln(2\pi\sigma_0^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$

$$\frac{\partial \ell}{\partial \mu}(\mu) = \frac{1}{\sigma_0^2}\sum_{i=1}^{n}(x_i-\mu) \quad \begin{cases} =0 \iff \mu = \frac{1}{n}\sum_{i=1}^{n}x_i & (1) \\ >0 \ , \ \mu < \frac{1}{n}\sum_{i=1}^{n}x_i & (2) \\ <0 \ , \ \mu > \frac{1}{n}\sum_{i=1}^{n}x_i & (3) \end{cases}$$

Bevis:

(1) klar

(2):

$$\mu < \frac{1}{n}\sum_{i=1}^{n}x_i \implies \frac{\partial \ell}{\partial \mu}(\mu) > \frac{1}{\sigma_0^2}\sum_{i=1}^{n}\left(x_i - \frac{1}{n}\sum_{i=1}^{n}x_i\right) = 0$$

(3) tilsvarende

$\square$

2)

Jvf. s. 3.23 : $\ell(\mu,\sigma^2) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$

Jvf. 1) gælder $\ell(\mu,\sigma^2) \le \ell(\bar{x}_\cdot,\sigma^2)$ $\forall \mu,\sigma^2$, dvs.

vi skal blot maks. $\ell(\bar{x}_\cdot,\sigma^2)$ som fkt. af σ^2.

$$\frac{\partial \ell}{\partial \sigma^2}(\sigma^2) = -\frac{n}{2}\frac{2\pi}{2\pi\sigma^2} + \frac{2}{(2\sigma^2)^2}\sum_{i=1}^{n}(x_i-\bar{x}_\cdot)^2$$

$$\begin{cases} =0 \iff \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x}_\cdot)^2 & (1) \\ >0 \ , \ \sigma^2 < \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x}_\cdot)^2 & (2) \\ <0 \ , \ \sigma^2 > \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x}_\cdot)^2 & (3) \end{cases}$$

3)

Jf. s. 3.73 (for fastholdt σ^2):

$$\ell(\alpha,\beta) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\{x_i-(\alpha+\beta t_i)\}^2$$

$$\frac{\partial \ell}{\partial \alpha}(\alpha,\beta) = \frac{1}{\sigma^2}\sum_{i=1}^{n}\{x_i-(\alpha+\beta t_i)\}$$

$$\frac{\partial \ell}{\partial \beta}(\alpha,\beta) = \frac{1}{\sigma^2}\sum_{i=1}^{n}\{x_i-(\alpha+\beta t_i)\}t_i$$

$$\frac{\partial^2 \ell}{\partial \alpha^2}(\alpha,\beta) = -\frac{n}{\sigma^2} \qquad \frac{\partial^2 \ell}{\partial \beta^2}(\alpha,\beta) = \frac{-1}{\sigma^2}\sum_{i=1}^{n}t_i^2$$

$$\frac{\partial^2 \ell}{\partial \alpha\partial \beta}(\alpha,\beta) = \frac{-1}{\sigma^2}\sum_{i=1}^{n}t_i$$

Idet $\frac{\partial^2 \ell}{\partial \alpha^2}(\alpha,\beta) < 0$ og

$$\left(\frac{\partial^2 \ell}{\partial \alpha\partial \beta}(\alpha,\beta)\right)^2 = \frac{1}{\sigma^4}\left(\sum_{i=1}^{n}t_i\right)^2 \underset{*}{<} \left(-\frac{n}{\sigma^2}\right)\left(\frac{-1}{\sigma^2}\sum_{i=1}^{n}t_i^2\right)$$

$$= \frac{\partial^2 \ell}{\partial \alpha^2}(\alpha,\beta)\cdot\frac{\partial^2 \ell}{\partial \beta^2}(\alpha,\beta)$$

* Jf. uligh. 3) på US 16 i SS2 idet vi antager, at $USS_t - \frac{S_t^2}{n} \neq 0$ (hvilket er rimeligt, idet vi må have mere end En t-værdi)

Af sætn. 1+3 s. 707+709 i 'Calculus' af Robert Adams 3. ed. følger at ℓ og dermed L har et globalt maksimum i $(\hat{\alpha},\hat{\beta})$

opgave 4.5

1)

Tegningerne viser, at en god model er

$M_0: x_{ij} \sim\sim N(\alpha_i + \beta_i t_{ij}, \sigma_i^2) \quad i=1,2 \quad , j=1,\dots,7$

($i=1$ bymidte $i=2$ villahave.)

2)

$$\hat{\beta}_1 \underset{3.49}{=} \frac{SP_{xt} - \frac{S_x S_t}{n}}{USS_t - \frac{S_t^2}{n}} = -0{,}166964$$

$$\hat{\alpha}_1 = \frac{S_x}{n} - \hat{\beta}\,\frac{S_t}{n} = 20{,}61464$$

$$s_1^2 \underset{.37175}{=} \frac{1}{n-2}\left(SSD_x - \frac{SPD_{xt}^2}{SSD_t}\right)$$

$$= \frac{1}{n-2}\left(USS_x - \frac{S_x^2}{n} - \frac{\left(SP_{xt} - \frac{S_x S_t}{n}\right)^2}{USS_t - \frac{S_t^2}{n}}\right)$$

$$= 0{,}00719548$$

3)

$$s_2^2 = 0{,}00510936$$

$H_{0\sigma}: \sigma_1^2 = \sigma_2^2$

$$F = \frac{s_1^2}{s_2^2} = 1{,}40829$$

$$\varepsilon(\underline{x}) = 1 - F_{F(7-2,7-2)}(F) \geq 1 - 0{,}7 = 0{,}3 > 0{,}05$$

Dvs. vi reducere til $M_1: x_{ij} \sim\sim N(\alpha_i + \beta_i t_{ij}, \sigma^2)$

4)

$H_{0\beta}: \beta_1 = \beta_2$

$$F = \frac{\frac{0{,}07811879 - 0{,}06152421}{11 - 10}}{0{,}00615242} = 2{,}697244$$

$\varepsilon(\underline{x}) = 1 - F_{F(1,10)}(F) > 1 - 0{,}9 = 0{,}1 > 0{,}05$

Dvs. hypotesen forkastes ikke og vi reducere til

$M_2: x_{ij} \sim\sim N(\alpha_i + \beta t_{ij}, \sigma^2)$

$H_{0\alpha}: \alpha_1 = \alpha_2$

$$F = \frac{\frac{4{,}42829136 - 0{,}07811879}{12 - 11}}{0{,}00710171} = 612{,}5528$$

$\varepsilon(\underline{x}) = 1 - F_{F(1,11)}(F) < 1 - 0{,}9995 < 0{,}05$

Hypotesen forkastes, altså er slutmodellen M_2

Estimaterne er

$\hat{\alpha}_2 = 18{,}01935714 \qquad \hat{\alpha}_1 = 19{,}13421428$

$\hat{\beta} = -0{,}1497500$

$\hat{\sigma}^2 = 0{,}00710171$

```
TITLE1 'Opgave 4_5';

GOPTIONS DEV=win CTEXT=black GACCESS=gsasfile GSFMODE=replace;
GOPTIONS TARGETDEVICE=PS300A4;

GOPTIONS TARGETDEVICE=PS300A4 ROTATE=landscape
         HORIGIN=2 cm VORIGIN=2 cm
         CBACK=white GUNIT=pct HTITLE=4 HTEXT=2.7
         FTEXT=swiss CTEXT=black CSYMBOL=black;
OPTIONS NODATE PAGESIZE=45 PAGENO=1;
PAGE off;

DATA bly;
INPUT gruppe aar bly log_bly;
CARDS;
1 83 865   6.763
1 84 721   6.581
1 85 671   6.509
1 86 514   6.242
1 87 381   5.943
1 88 370   5.914
1 89 343   5.838
2 83 257   5.549
2 84 201   5.303
2 85 206   5.328
2 86 189   5.242
2 87 145   4.977
2 88 130   4.868
2 89 112   4.719
;

/* Standardberegninger */
DATA ber;
SET bly;
xt=log_bly*aar;
FORMAT xt 10.3;
RUN;

PROC MEANS DATA=ber FW=13 VARDEF=DF N USS CSS SUM ;
VAR aar log_bly xt;
BY gruppe;
RUN;


DATA logbly_1;
SET bly;
IF gruppe=1;
RUN;

DATA logbly_2;
SET bly;
IF gruppe=2;
RUN;

TITLE ' ';
TITLE2 'aar mod log_bly BYMIDTE';
PROC PLOT DATA=logbly_1;
   PLOT log_bly*aar;
```

```
RUN;

TITLE2 'aar mod log_bly VILLAKVARTER';
PROC PLOT DATA=logbly_2;
   PLOT log_bly*aar;
RUN;  QUIT;


PROC GPLOT DATA=bly;
PLOT log_bly*aar=gruppe;
SYMBOL1 I=rl V=dot L=2;
SYMBOL2 I=rl V=circle L=2;
RUN;


TITLE2 'to regressionslinjer';
TITLE3 'separate beregninger';
PROC GLM DATA=bly;
   MODEL log_bly = aar /SOLUTION  SS1;
   BY gruppe;
RUN;


TITLE2 'to regressionslinjer';
PROC GLM DATA=bly;
   CLASS gruppe;
   MODEL log_bly = gruppe aar aar*gruppe /SOLUTION  SS1;
RUN;


TITLE2 'to regressionslinjer';
TITLE3 'samme haeldning';
PROC GLM DATA=bly;
   CLASS gruppe;
   MODEL log_bly = aar gruppe /SOLUTION  SS1;
RUN;


TITLE2 'to regressionslinjer';
TITLE3 'sammenfaldende';
PROC GLM DATA=bly;
   MODEL log_bly = aar  / SOLUTION SS1;
RUN;

QUIT;
```

---------- gruppe=1 ----------

The MEANS Procedure

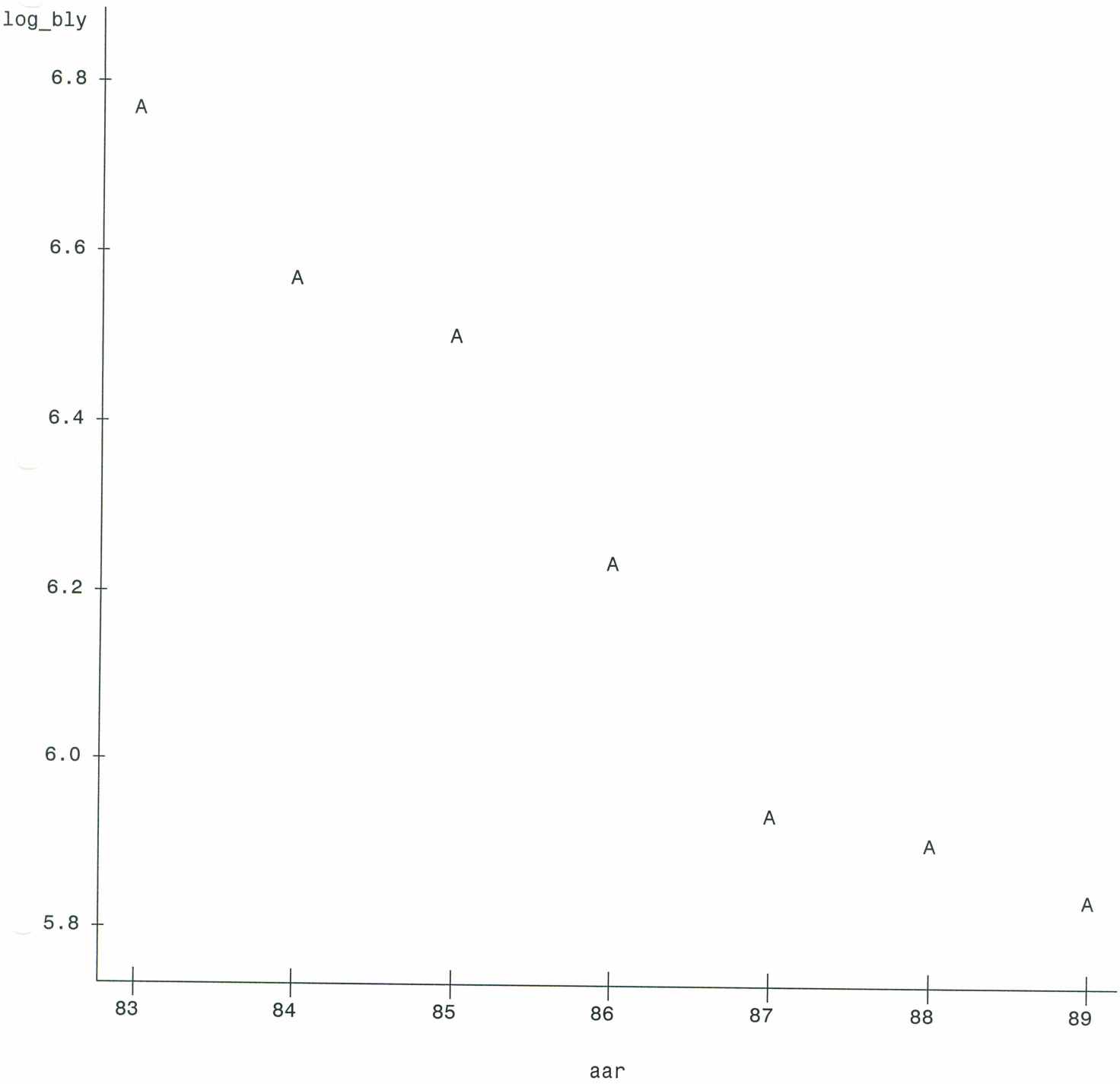
Variable	N	USS	Corrected SS	Sum
aar	7	51800.0	28.0000000	602.0000000
log_bly	7	274.7542640	0.8165354	43.7900000
xt	7	2023098.1	2081.7663657	3761.2650000

---------- gruppe=2 ----------

Variable	N	USS	Corrected SS	Sum
aar	7	51800.0	28.0000000	602.0000000
log_bly	7	185.5162720	0.5173869	35.9860000
xt	7	1366274.6	1302.2711857	3091.0850000

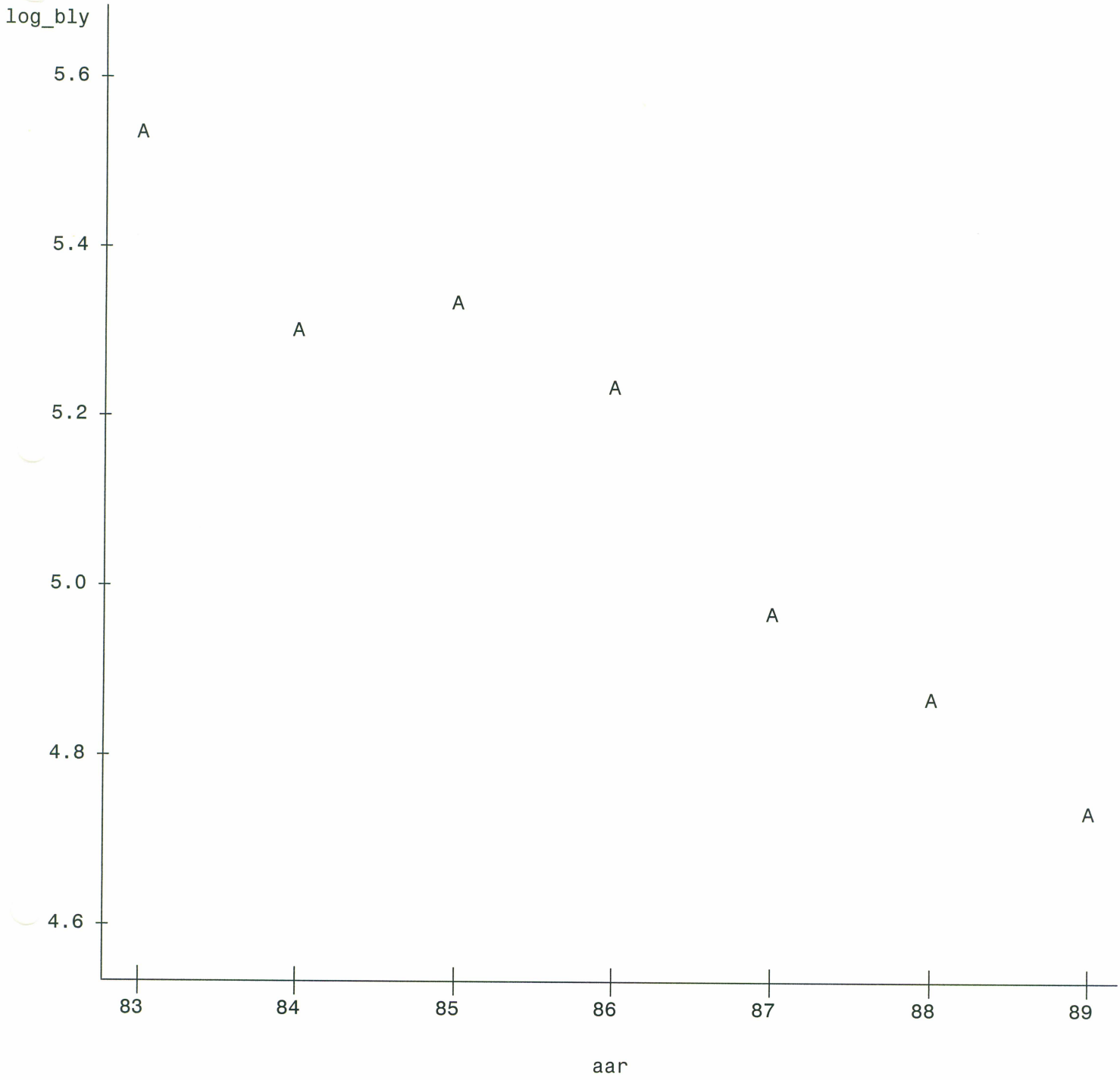
aar mod log_bly BYMIDTE

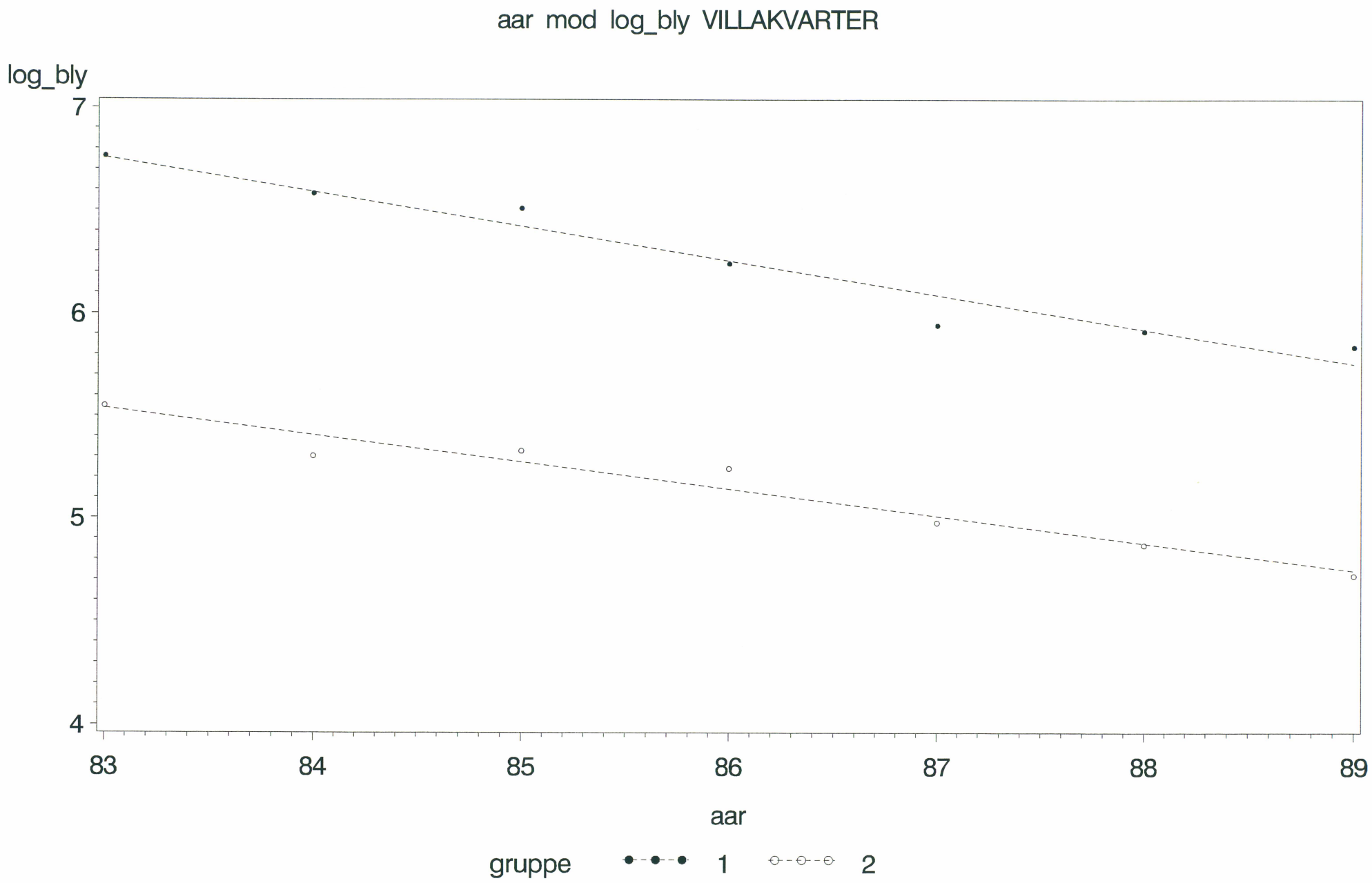
Plot of log_bly*aar. Legend: A = 1 obs, B = 2 obs, etc.



aar mod log_bly VILLAKVARTER

Plot of log_bly*aar. Legend: A = 1 obs, B = 2 obs, etc.




aar mod log_bly VILLAKVARTER
log_bly
7
6
5
4
83
84
85
86
87
88
89
aar
gruppe
1
2

to regressionslinjer
separate beregninger

---------- gruppe=1 ----------

The GLM Procedure

Dependent Variable: log_bly

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	0.78055804	0.78055804	108.48	0.0001
Error	5	0.03597739	0.00719548		
Corrected Total	6	0.81653543			

R-Square	Coeff Var	Root MSE	log_bly Mean
0.955939	1.355979	0.084826	6.255714

Source	DF	Type I SS	Mean Square	F Value	Pr > F
aar	1	0.78055804	0.78055804	108.48	0.0001

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	20.61464286	1.37900768	14.95	<.0001
aar	-0.16696429	0.01603064	-10.42	0.0001

to regressionslinjer
separate beregninger

---------- gruppe=2 ----------

The GLM Procedure

Dependent Variable: log_bly

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	0.49184004	0.49184004	96.26	0.0002
Error	5	0.02554682	0.00510936		
Corrected Total	6	0.51738686			

R-Square	Coeff Var	Root MSE	log_bly Mean
0.950623	1.390426	0.071480	5.140857

Source	DF	Type I SS	Mean Square	F Value	Pr > F
aar	1	0.49184004	0.49184004	96.26	0.0002

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	16.53892857	1.16203786	14.23	<.0001
aar	-0.13253571	0.01350842	-9.81	0.0002

to regressionslinjer

The GLM Procedure

Dependent Variable: log_bly

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	5.62257064	1.87419021	304.63	<.0001
Error	10	0.06152421	0.00615242		
Corrected Total	13	5.68409486			

R-Square	Coeff Var	Root MSE	log_bly Mean
0.989176	1.376508	0.078437	5.698286

Source	DF	Type I SS	Mean Square	F Value	Pr > F
gruppe	1	4.35017257	4.35017257	707.07	<.0001
aar	1	1.25580350	1.25580350	204.12	<.0001
aar*gruppe	1	0.01659457	0.01659457	2.70	0.1316

Parameter		Estimate		Standard Error	t Value	Pr > |t|
Intercept		16.53892857	B	1.27514591	12.97	<.0001
gruppe	1	4.07571429	B	1.80332864	2.26	0.0474
gruppe	2	0.00000000	B	.	.	.
aar		-0.13253571	B	0.01482327	-8.94	<.0001
aar*gruppe	1	-0.03442857	B	0.02096327	-1.64	0.1316
aar*gruppe	2	0.00000000	B	.	.	.

NOTE: The X'X matrix has been found to be singular, and a generalized inverse was used to solve the normal equations. Terms whose estimates are followed by the letter 'B' are not uniquely estimable.

to regressionslinjer
samme haeldning

The GLM Procedure

Dependent Variable: log_bly

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	5.60597607	2.80298804	394.69	<.0001
Error	11	0.07811879	0.00710171		
Corrected Total	13	5.68409486			

R-Square	Coeff Var	Root MSE	log_bly Mean
0.986257	1.478894	0.084272	5.698286

Source	DF	Type I SS	Mean Square	F Value	Pr > F
aar	1	1.25580350	1.25580350	176.83	<.0001
gruppe	1	4.35017257	4.35017257	612.55	<.0001

Parameter		Estimate		Standard Error	t Value	Pr > |t|
Intercept		18.01935714	B	0.96899289	18.60	<.0001
aar		-0.14975000		0.01126127	-13.30	<.0001
gruppe	1	1.11485714	B	0.04504508	24.75	<.0001
gruppe	2	0.00000000	B	.	.	.

NOTE: The X'X matrix has been found to be singular, and a generalized inverse was used to solve the normal equations. Terms whose estimates are followed by the letter 'B' are not uniquely estimable.

to regressionslinjer
sammenfaldende

The GLM Procedure

Dependent Variable: log_bly

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1.25580350	1.25580350	3.40	0.0899
Error	12	4.42829136	0.36902428		
Corrected Total	13	5.68409486			

R-Square	Coeff Var	Root MSE	log_bly Mean
0.220933	10.66064	0.607474	5.698286

Source	DF	Type I SS	Mean Square	F Value	Pr > F
aar	1	1.25580350	1.25580350	3.40	0.0899

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	18.57678571	6.98311697	2.66	0.0208
aar	-0.14975000	0.08117709	-1.84	0.0899

Afdeling for Teoretisk Statistik
Institut for Matematiske Fag
Aarhus Universitet

Statistik 1 og Statistik α
Jørgen Granfeldt
11. oktober 2002

Meddelelse 7

Forelæsningerne i sidste uge

(Onsdag den 2. oktober) Begyndte på Afsnit 3.2.2 *k observationsrækker* og nåede til nederst side 3.55.

(Fredag den 4. oktober) Afsluttede Afsnit 3.2.2 *k observationsrækker.*

Forelæsningerne i denne uge

(Onsdag den 9. oktober) Begyndte på Afsnit 3.3 *Lineær regression.* Blev næsten færdig med afsnittet om beregninger, som begynder side 3.74.

(Fredag den 11. oktober) Fortsætter med lineær regression og bliver måske helt færdig med afsnittet. Næste emne er Kapitel 4 *Lineære normale modeller.*

Forelæsningerne i ugen efter efterårsferien

(Onsdag den 23. oktober og fredag den 25. oktober) Kapitel 4 *Lineære normale modeller.*

Øvelserne i denne uge

(Uge 41: 7. oktober-13. oktober) Opgave 3.9, 3.17 (men spring over spørgsmål 3) og besvar spørgsmål 4) alene med metoden fra Afsnit 3.2.2), 3.19, 3.13.

Øvelserne i de to næste uger

(Uge 42: 14. oktober-20. oktober) Efterårsferie.

(Uge 43: 21. oktober-27. oktober) 3.12, 3.17 (regnes færdig med brug af lineær regression), 3.20, 4.5 (Kun spørgsmålene 1), 2) og 3)).

Afleveringsopgaver

(Uge 42) Ingen afleveringsopgave. Efterårsferie.

(Uge 43) Ingen afleveringsopgave.

Udleveret materiale

I uge 41 blev planen for de obligatoriske opgaver udleveret. Der var en enkelt trykfejl. Den rettede plan er her:

	Stilles	Afleveres
Obligatorisk opgave 1	Fredag den 4. oktober	Fredag den 1. november
Obligatorisk opgave 2	Fredag den 1. november	Fredag den 22. november
Obligatorisk opgave 3	Fredag den 22. november	Fredag den 13. december

Udlån af SAS cd Michael Sørensen gør stadig opmærksom på at der sædvanligvis opstår stort pres på udlån af SAS når afleveringsfristen for den første obligatoriske opgave nærmer sig. Man kan låne en SAS cd og installationsvejledning hos Michael Sørensen, som har kontor H2.11.

STAT 1, TØ nr. 6

opgave 3.9

Like. fkt. er i den oprindelige model

$$L(\underline{\mu}, \underline{\sigma}^2, \underline{x}) = \prod_{i=1}^{k} \prod_{j=1}^{n_i} \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-(x_{ij}-\mu_i)^2/2\sigma_i^2}$$

$$= \prod_{i=1}^{k} \left(\frac{1}{2\pi\sigma_i^2}\right)^{n_i/2} e^{-\frac{1}{2\sigma_i^2}\sum_{j=1}^{n_i}(x_{ij}-\mu_i)^2} = \prod_{i=1}^{k} L_i(\mu_i, \sigma_i^2)$$

hvor L_i er like. fkt. for den i'te obstrekke set som en obs. rekke, det er klart, at $L(\underline{\mu}, \underline{\sigma}^2) \leq \prod_{i=1}^{k} L_i(\hat{\mu}_i, \hat{\sigma}_i^2)$

Dvs. $\underline{\mu} = (\hat{\mu}_1, \ldots, \hat{\mu}_k)$, $\underline{\sigma}^2 = (\hat{\sigma}_1^2, \ldots, \hat{\sigma}_k^2)$ maksimerer like. fkt.

Under nulhypotesen ser vi

$$L(\underline{\mu}, \sigma^2, \underline{x}) \leq L(\hat{\underline{\mu}}, \sigma^2, \underline{x}) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n_\cdot}{2}} e^{-\frac{1}{2\sigma^2}\sum_{ij}(x_{ij}-\hat{\mu}_i)^2}$$

$$\ell(\hat{\underline{\mu}}, \sigma^2) = -\frac{n_\cdot}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\hat{\mu}_i)^2$$

$$\frac{\partial \ell}{\partial \sigma^2}(\hat{\underline{\mu}}, \sigma^2)\Big|_{\hat{\mu}_i = \bar{x}_i} = -\frac{n_\cdot}{2}\frac{2\pi}{2\pi\sigma^2} + \frac{2}{(2\sigma^2)^2}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_{i\cdot})^2$$

$$\begin{cases} = 0 & \Longleftrightarrow \sigma^2 = \frac{1}{n_\cdot}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_{i\cdot})^2 = \frac{1}{n_\cdot}\sum_{i=1}^{k} SSD_{(i)} \\ > 0, & \sigma^2 < \text{———} + \text{———} \\ < 0, & \sigma^2 > \text{———} - \text{———} \end{cases}$$

Altså er maks. like. est. for σ^2 $\hat{\sigma}^2 = \frac{1}{n_\cdot}\sum_{i=1}^{k} SSD_{(i)}$

Jf. 3.78: $SSD_{(i)} \sim\sim \sigma^2\chi^2(n_i - 1)$, pga. uafh. har vi (3.75) $\sum SSD_{(i)} \sim\sim \sigma^2\chi^2(\sum n_i - 1)$, dermed ser vi, at et middelværdirt est. for σ^2 er

$$s_1^2 = \frac{n_\cdot}{n_\cdot - k}\hat{\sigma}^2 = \frac{\sum SSD_{(i)}}{\sum f_{(i)}} = \frac{\sum f_{(i)} s_{(i)}^2}{\sum f_{(i)}}$$

opg. 3.17

1)

Se vedlagte fraktildiagram

2)

$M_0: \quad X_{ij} \sim\sim N(\mu_i, \sigma_i^2) \qquad i=1,\dots,4 \qquad j=1,\dots,n_i$

$H_{02}: \quad \sigma_1^2 = \dots = \sigma_4^2$

Jf. vedlagte udskrift: $\quad \varepsilon(\underline{x}) = 0{,}37179 > 0{,}05$

Altså forkaste vi ikke hypotesen, og reducere dermed til

$M_1: \quad X_{ij} \sim\sim N(\mu_i, \sigma^2) \qquad i=1,\dots,4, \; j=1,\dots,n_i$

3)

$H_{01}: \quad \mu_i = \alpha t_i + \beta \quad , i=1,\dots,4$

$$F = \frac{\frac{SSD_{02} - SSD_1}{f_{02} - f_1}}{s_1} = 2{,}00809$$

se udskriften

$$\varepsilon(x) = 1 - F_{F(2,242)}(F) > 1 - 0{,}9 = 0{,}1 > 0{,}05$$

Altså forkastes hypotesen ikke, og vi reducere til

$M_2: \quad X_{ij} \sim\sim N(\alpha t_i + \beta, \sigma^2)$

4)

Vi skal teste

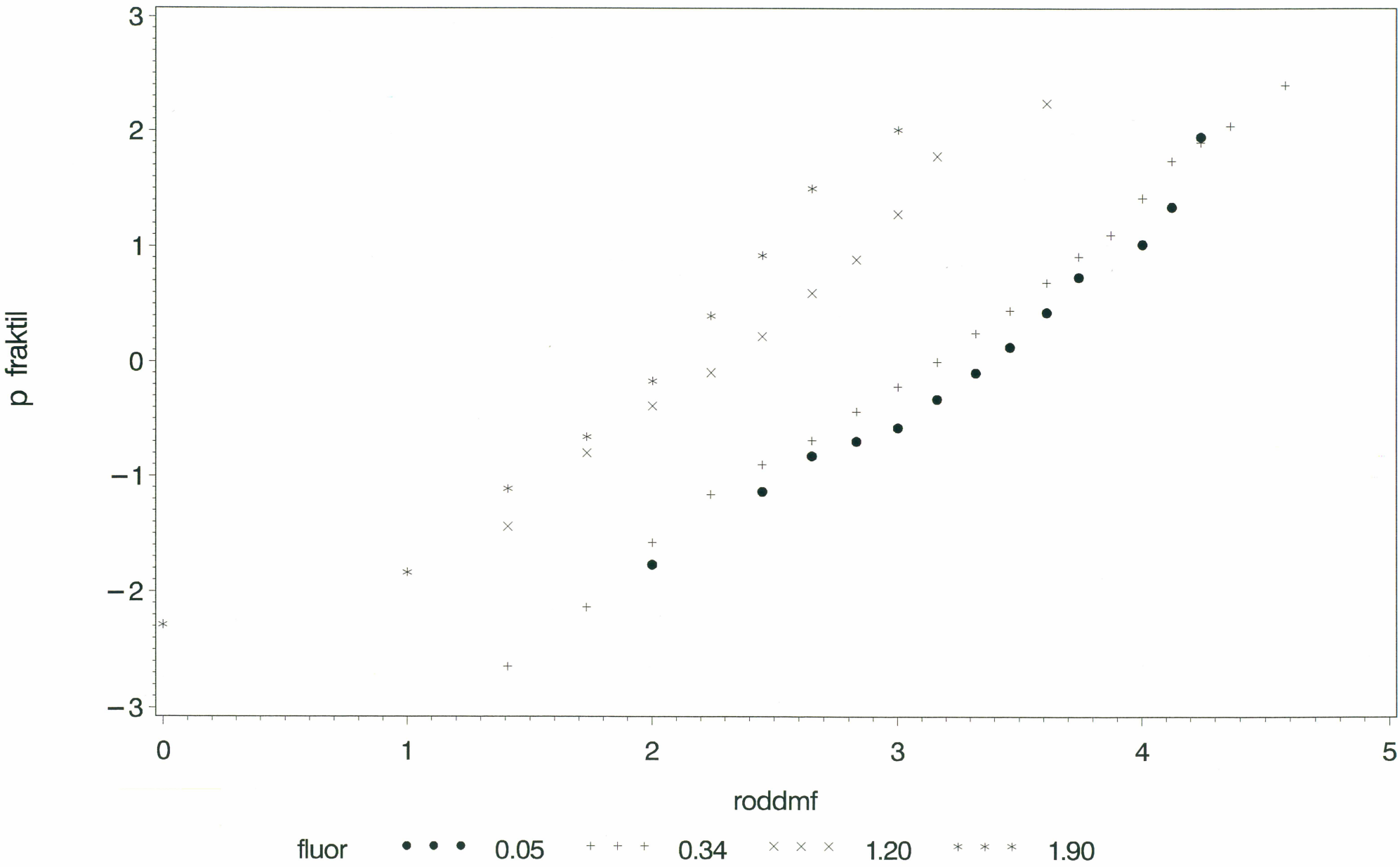
$H_{0\beta}: \beta = 0$

$t(x) = -12{,}26$, $\varepsilon(x) \leq 0{,}001 < 0{,}01$ (se udskrift)

Hypotesen forkastes

```
TITLE1 'Opgave 3.17';

OPTIONS PAGENO=1;

/* Nogle GOPTIONS */
GOPTIONS DEV=win CTEXT=black GACCESS=gsasfile GSFMODE=replace;
GOPTIONS TARGETDEVICE=PS300A4 ROTATE=landscape
         HORIGIN=2 cm VORIGIN=2 cm
         CBACK=white GUNIT=pct HTITLE=4 HTEXT=2.7
         FTEXT=swiss CTEXT=black CSYMBOL=black;
OPTIONS NODATE PAGESIZE=45;
PAGE off;

DATA fluor;
INFILE 'H:\Documents and Settings\Ejer\Skrivebord\EKSAMENSOPGAVER\opgaver\kapitel_03\opgave_1
INPUT dmf fluor;
RUN;

DATA fluor;
SET fluor;
roddmf=sqrt(dmf);
roddmf=round(roddmf,0.01);
/*  dette sikrer, at SAS regner med samme tal, som i opgaven
    nemlig kvadratrod med to decimaler                        */
RUN;


%INCLUDE 'H:\Documents and Settings\Ejer\Skrivebord\EKSAMENSOPGAVER\sasmacros\normgraf.mac';
%normgraf(data=fluor,var=roddmf,bygrp=fluor,tabel=ja,grafik=ja);

%INCLUDE 'H:\Documents and Settings\Ejer\Skrivebord\EKSAMENSOPGAVER\sasmacros\bartlett.mac';
%bartlett(data=fluor,var=roddmf,gruppe=fluor);

DATA fluor;
SET fluor;
xt = roddmf*fluor;
RUN;

PROC MEANS DATA=fluor VARDEF=DF
                SUM N MEAN CSS USS;
VAR fluor roddmf xt;
RUN;

PROC GLM DATA=fluor;
CLASS fluor;
MODEL roddmf=fluor / SS1 SOLUTION;
RUN;

PROC GLM DATA=fluor;
MODEL roddmf=fluor / SS1;
RUN;
QUIT;
```

Opgave 3.17



Macro "BARTLETT"

Dataset: FLUOR
Responsvariabel: RODDMF
Gruppevariabel: FLUOR

Beregningsskema:

i	ni	Si	USSi	Si2/ni	SSDi	fi	Estimeret varians	Gennem-snit
0.05	39	127.910	434.8519	419.5120	15.3399	38	0.40368	3.2797
0.34	122	379.420	1233.2858	1179.9962	53.2896	121	0.44041	3.1100
1.2	40	90.840	219.0978	206.2976	12.8002	39	0.32821	2.2710
1.9	45	90.050	193.0717	180.2001	12.8716	44	0.29254	2.0011
	246	688.220	2080.3072	1986.0059	94.3013	242	0.38967	

Bartletts test:

C	-2lnQ	Ba	Prob > chi
1.00876	3.15896	3.13153136	0.37179

The GLM Procedure

Dependent Variable: roddmf

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	60.6125306	20.2041769	51.85	<.0001
Error	242	94.3013019	0.3896748		
Corrected Total	245	154.9138325			

R-Square	Coeff Var	Root MSE	roddmf Mean
0.391266	22.31305	0.624239	2.797642

Source	DF	Type I SS	Mean Square	F Value	Pr > F
fluor	3	60.61253064	20.20417688	51.85	<.0001

Parameter		Estimate		Standard Error	t Value	Pr > |t|
Intercept		2.001111111	B	0.09305611	21.50	<.0001
fluor	0.05	1.278632479	B	0.13656904	9.36	<.0001
fluor	0.34	1.108888889	B	0.10887376	10.19	<.0001
fluor	1.2	0.269888889	B	0.13565143	1.99	0.0478
fluor	1.9	0.000000000	B	.	.	.

NOTE: The X'X matrix has been found to be singular, and a generalized inverse was used to solve the normal equations. Terms whose estimates are followed by the letter 'B' are not uniquely estimable.

The GLM Procedure

Dependent Variable: roddmf

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	59.0475234	59.0475234	150.29	<.0001
Error	244	95.8663092	0.3928947		
Corrected Total	245	154.9138325			

R-Square	Coeff Var	Root MSE	roddmf Mean
0.381164	22.40505	0.626813	2.797642

Source	DF	Type I SS	Mean Square	F Value	Pr > F
fluor	1	59.04752337	59.04752337	150.29	<.0001

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	3.331935769	0.05913215	56.35	<.0001
fluor	-0.742871187	0.06059695	-12.26	<.0001

opg. 3.19

1)

$M_0 : x_{ij} \sim N(\mu_i, \sigma_i^2) \quad i=1,\dots,4 \quad j=1,\dots,5$

$H_{0\sigma^2} : \sigma_1^2 = \dots = \sigma_4^2$

Jf. vedlagte udskrift: $\varepsilon(\underline{x}) = 0{,}80682 > 0{,}05$

dvs. hypotesen forkastes ikke

2)

$M_1 : x_{ij} \sim N(\mu_i, \sigma^2) \quad i=1,\dots,4 \quad j=1,\dots,5$

$H_{0\mu} : \mu_1 = \dots = \mu_4$

Jf. udskriftene:

$$F = \frac{\frac{700{,}55 - 515{,}375}{19-16}}{32{,}2109375} = 1{,}91627$$

$$\varepsilon(\underline{x}) = 1 - F_{F(19-16,16)}(F) \geq 1 - 0{,}9 = 0{,}1 > 0{,}05$$

Hypotesen forkastes ikke

Modellen bliver $M_2 : x_{ij} \sim N(\mu, \sigma^2)$

3)

M_0 : $X_{ijk} \sim\sim N(\mu_i, \sigma_i^2)$ $i=1,2$ $j=1,\dots,4$ $k=1,\dots,5$

$H_{0\sigma^2}$: $\sigma_1^2 = \sigma_2^2$

$\varepsilon(\underline{x}) = 0,8176 > 0,05$ jf. udskriften, dvs. vi kan reducere til

M_1 : $X_{ijk} \sim\sim N(\mu_i, \sigma^2)$

4)

$H_{0\mu}$: $\mu_1 = \mu_2$

$\varepsilon(\underline{x}) = 0,3194 > 0,05$

Vi kan reducere til slutmodellen

M : $X_{ijk} \sim\sim N(\mu, \sigma^2)$ $i=1,2$, $j=1,\dots,4$, $k=1,\dots,5$

```
TITLE1 'Opgave 3_19';

OPTIONS NODATE PAGENO=1 LS=76 PS=33;

DATA opg_z;
INFILE 'H:\Documents and Settings\Ejer\Skrivebord\EKSAMENSOPGAVER\opgaver\kapitel_03\opgave_1
INPUT tid lag $ myk_pct;
RUN;

%INCLUDE 'H:\Documents and Settings\Ejer\Skrivebord\EKSAMENSOPGAVER\sasmacros\bartlett.mac';
%bartlett(data=opg_z,var=myk_pct,gruppe=tid,bygruppe=lag);

DATA mor;
SET opg_z;
IF lag='m';
RUN;

DATA blegsand;
SET opg_z;
IF lag='b';
RUN;

TITLE1 'mor';
PROC SORT DATA=mor;
     BY tid;
RUN;

%INCLUDE 'H:\Documents and Settings\Ejer\Skrivebord\EKSAMENSOPGAVER\sasmacros\normgraf.mac';
AXIS2 ORDER=(30 TO 60 BY 10);
%normgraf(data=mor,var=myk_pct,bygrp=tid);

PROC PLOT DATA=mor;
     PLOT myk_pct*tid;
RUN;

%bartlett(data=mor,var=myk_pct,gruppe=tid);

TITLE2 '4 observationsraekker';
PROC GLM DATA=mor;
     CLASS tid;
MODEL myk_pct=tid /SS1;
RUN;

TITLE2 '1 observationsraekke';
PROC GLM DATA=mor;
     MODEL myk_pct= /SS1;
RUN;

TITLE1 'blegsand';
PROC SORT DATA=blegsand;
     BY tid;
RUN;

%normgraf(data=blegsand,var=myk_pct,bygrp=tid);

PROC PLOT DATA=blegsand;
     PLOT myk_pct*tid;
RUN;
```

```
%bartlett(data=blegsand,var=myk_pct,gruppe=tid);

TITLE2 '4 observationsraekker';
PROC GLM DATA=blegsand;
    CLASS tid;
MODEL myk_pct=tid /SS1;
RUN;

TITLE2 '1 observationsraekke';
PROC GLM DATA=blegsand;
    MODEL myk_pct= /SS1;
RUN;

TITLE1 'mor og blegsand';
TITLE2 '2 observationsraekker';
PROC GLM DATA=opg_z;
     CLASS lag;
MODEL myk_pct=lag /SS1;
RUN;

PROC TTEST DATA=opg_z;
CLASS lag;
VAR myk_pct;
RUN;

%normgraf(data=opg_z,var=myk_pct);
RUN;
```

Macro "BARTLETT"

Dataset: MOR
Responsvariabel: MYK_PCT
Gruppevariabel: TID

Beregningsskema:

i	ni	Si	USSi	Si2/ni	SSDi	fi	Estimeret varians	Gennem-snit
1	5	222.000	9931.1250	9856.8000	74.3250	4	18.58125	44.4000
2	5	200.250	8168.6875	8020.0125	148.6750	4	37.16875	40.0500
3	5	225.750	10290.5625	10192.6125	97.9500	4	24.48750	45.1500
4	5	189.000	7338.6250	7144.2000	194.4250	4	48.60625	37.8000
	20	837.000	35729.0000	35213.6250	515.3750	16	32.21094	

Bartletts test:

mor 5

C	-2lnQ	Ba	Prob > chi
1.10417	1.07875	0.97698297	0.80682

4 observationsraekker

The GLM Procedure

Dependent Variable: myk_pct

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	185.1750000	61.7250000	1.92	0.1677
Error	16	515.3750000	32.2109375		
Corrected Total	19	700.5500000			

R-Square	Coeff Var	Root MSE	myk_pct Mean
0.264328	13.56145	5.675468	41.85000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
tid	3	185.1750000	61.7250000	1.92	0.1677

1 observationsraekke

The GLM Procedure

Dependent Variable: myk_pct

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	35028.45000	35028.45000	950.03	<.0001
Error	19	700.55000	36.87105		
Uncorrected Total	20	35729.00000			

R-Square	Coeff Var	Root MSE	myk_pct Mean
0.000000	14.50933	6.072154	41.85000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Intercept	1	35028.45000	35028.45000	950.03	<.0001

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	41.85000000	1.35777488	30.82	<.0001

Macro "BARTLETT"

Dataset: BLEGSAND
Responsvariabel: MYK_PCT
Gruppevariabel: TID

Beregningsskema:

i	ni	Si	USSi	Si2/ni	SSDi	fi	Estimeret varians	Gennem-snit
1	5	212.500	9110.6250	9031.2500	79.3750	4	19.84375	42.5000
2	5	220.250	9817.8125	9702.0125	115.8000	4	28.95000	44.0500
3	5	237.250	11415.0625	11257.5125	157.5500	4	39.38750	47.4500
4	5	204.750	8545.1875	8384.5125	160.6750	4	40.16875	).9500
	20	874.750	38888.6875	38375.2875	513.4000	16	32.08750	

Bartletts test:

blegsand 14

C	-2lnQ	Ba	Prob > chi
1.10417	0.61548	0.5574124	0.90611

The GLM Procedure

Dependent Variable: myk_pct

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	115.9093750	38.6364583	1.20	0.3401
Error	16	513.4000000	32.0875000		
Corrected Total	19	629.3093750			

R-Square	Coeff Var	Root MSE	myk_pct Mean
0.184185	12.95132	5.664583	43.73750

Source	DF	Type I SS	Mean Square	F Value	Pr > F
tid	3	115.9093750	38.6364583	1.20	0.3401

1 observationsraekke

The GLM Procedure

Dependent Variable: myk_pct

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	38259.37813	38259.37813	1155.12	<.0001
Error	19	629.30937	33.12155		
Uncorrected Total	20	38888.68750			

R-Square	Coeff Var	Root MSE	myk_pct Mean
0.000000	13.15835	5.755132	43.73750

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Intercept	1	38259.37813	38259.37813	1155.12	<.0001

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	43.73750000	1.28688667	33.99	<.0001

2 observationsraekker

The TTEST Procedure

Statistics

Variable	lag	N	Lower CL Mean	Mean	Upper CL Mean	Lower CL Std Dev	Std Dev
myk_pct	b	20	41.044	43.738	46.431	4.3767	5.7551
myk_pct	m	20	39.008	41.85	44.692	4.6178	6.0722
myk_pct	Diff (1-2)		-1.9	1.8875	5.6746	4.8346	5.9158

Statistics

Variable	lag	Upper CL Std Dev	Std Err	Minimum	Maximum
myk_pct	b	8.4058	1.2869	34.25	55
myk_pct	m	8.8688	1.3578	28.25	50.5
myk_pct	Diff (1-2)	7.6241	1.8707		

T-Tests

Variable	Method	Variances	DF	t Value	Pr > |t|
myk_pct	Pooled	Equal	38	1.01	0.3194
myk_pct	Satterthwaite	Unequal	37.9	1.01	0.3194

2 observationsraekker

The TTEST Procedure

Equality of Variances

Variable	Method	Num DF	Den DF	F Value	Pr > F
myk_pct	Folded F	19	19	1.11	0.8176

2 observationsraekker

The GLM Procedure

Dependent Variable: myk_pct

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	35.626563	35.626563	1.02	0.3194
Error	38	1329.859375	34.996299		
Corrected Total	39	1365.485938			

R-Square	Coeff Var	Root MSE	myk_pct Mean
0.026091	13.82390	5.915767	42.79375

Source	DF	Type I SS	Mean Square	F Value	Pr > F
lag	1	35.62656250	35.62656250	1.02	0.3194

3.13

$\underline{X}$ s.v. (n-dim.)

$E[\underline{X}] = \mu$, $\operatorname{Var} \underline{X} = \Sigma$

$\xi \in \mathbb{R}^k$, $B \in \operatorname{Mat}_{k \times n}(\mathbb{R})$

Påstand: $E[\xi + B\underline{X}] = \xi + B\mu$, $\operatorname{Var}(\xi + B\underline{X}) = B \Sigma B^*$

Bevis:

$$B = \begin{pmatrix} — b_1 — \\ — b_i — \\ — b_k — \end{pmatrix} , \quad \xi = \begin{pmatrix} \xi_1 \\ \vdots \\ \xi_k \end{pmatrix} \quad \text{osv.}$$

$E[(\xi + B\underline{X})_i] = E[\xi_i + b_i \underline{X}]$

$= \xi_i + b_i \mu$, jvf. lin. af $E[\cdot]$

$\operatorname{Var}(\xi + B\underline{X})_{ij}$

$= E[((\xi + B\underline{X})_i - \xi_i - b_i\mu)((\xi + B\underline{X})_j - \xi_j - b_j\mu)]$

$= E[(b_i\underline{X} - b_i\mu)(b_j\underline{X} - b_j\mu)]$

$= E[b_i(\underline{X} - \mu) b_j (\underline{X} - \mu)]$

$= E[b_i(\underline{X} - \mu)(\underline{X} - \mu)^* b_j^*]$

$= b_i E[(\underline{X} - \mu)(\underline{X} - \mu)^*] b_j^* = (B \Sigma B^*)_{ij}$

jvf. lin. af $E[\cdot]$, idet $E[(\underline{X} - \mu)(\underline{X} - \mu)^*] = \operatorname{Var}(\underline{X})$

$\square$

Afdeling for Teoretisk Statistik
Institut for Matematiske Fag
Aarhus Universitet

Statistik 1 og Statistik α
Jørgen Granfeldt
4. oktober 2002

Meddelelse 6

Forelæsningerne i sidste uge

(Onsdag den 25. september) Forsatte gennemgangen af Afsnit 3.2.1 *To observationsrækker* og nåede til nederst side 3.36 i gennemgangen af *to observationsrækker med varianshomogenitet.*

(Fredag den 27. september) Afsluttede gennemgangen af Afsnit 3.2.1 *To observationsrækker* bortset fra F-testet for $H_{0\sigma^2}$: $\sigma^2 = \sigma_0^2$ og Statistical Tables side 3.43, som blev gennemgået tirsdag den 2. oktober. Gennemgangen var lidt enklere end i bogen, se sidst på denne meddelelse.

Forelæsningerne i denne uge

(Onsdag den 2. oktober) Begyndte på Afsnit 3.2.2 *k observationsrækker* og nåede til nederst side 3.55.

(Fredag den 4. oktober) Afslutter Afsnit 3.2.2 *k observationsrækker*. Næste punkt er Afsnit 3.3 *Lineær regression.*

Forelæsningerne i næste uge

(Onsdag den 9. oktober og fredag den 11. oktober) Afsnit 3.3 *Lineær regression.*

Øvelserne i denne uge

(Uge 40: 30. september-6. oktober) Opgave 3.5, 3.6 (afleveringsopgaven) regnes færdig, 3.7, 3.8 samt opgaven sidst på denne meddelelse. (Bemærk at programmet er ændret i forhold til programmet for uge 40 på meddelelse 4.)

Øvelserne i de to næste uger

(Uge 41: 7. oktober-13. oktober) Opgave 3.9, 3.17 (men spring over spørgsmål 3) og besvar spørgsmål 4) alene med metoden fra Afsnit 3.2.2) 3.19, 3.13.

(Uge 42: 14. oktober-20. oktober) Efterårsferie

(Uge 43: 21. oktober-27. oktober) 3.12, 3.17 (regnes færdig med brug af lineær regression), 3.20, 4.5 (Kun spørgsmålene 1), 2) og 3)).

Afleveringsopgaver

Instuktorerne og jeg synes at for få benytter sig af tilbuddet om at få rettet afleveringsopgaver og vi opfordrer til at I benytter lejligheden i næste uge som opvarmning til den obligatoriske opgave.

(Uge 41) Følgende opgave:

I tabellen nedenfor findes konditallene for to grupper fodboldspillere i

samme klub. Den ene gruppe spiller i serie 6 og den anden spiller i serie 1.

	kondital									
serie 6	56.0	61.8	64.1	64.9	65.2	66.3	66.9	68.9	70.6	70.8
	72.4	73.1	73.9	74.1	74.1	75.3	78.7			
serie 1	68.9	75.2	74.3	72.9	72.0	63.9	76.3	76.3	75.4	66.0
	68.4	64.1	73.1	64.9	73.4	76.2	79.4	69.4	79.8	72.1

Analyser data med henblik på om der er forskel på konditallene for de to grupper.

Et SAS program, der indholder data, kan ses i en browser på adressen

`http://home.imf.au.dk/statbib/atskurser/stat1/meddelelser/aktuelle/afl41.sas`

(Uge 42) Ingen afleveringsopgave. Efterårsferie.

Udleveret materiale Plan over de obligatoriske opgaver, samt første obligatoriske opgave blev udleveret ved forelæsningen fredag den 4. oktober, men er også på nettet som

`http://www.imf.au.dk/statbib/atskurser/stat1/meddelelser/aktuelle/oblg_opg1.ps`

Udlån af SAS cd Michael Sørensen gør opmærksom på at der sædvanligvis opstår stort pres på udlån af SAS når afleveringsfristen for den første obligatoriske opgave nærmer sig. Man kan låne en SAS cd og installationsvejledning hos Michael Sørensen, som har kontor H2.11.

The F-test of $H_{0\sigma^2}$: $\sigma_1^2 = \sigma_2^2$ and Statistical Tables

Tables of the F-distribution are often limited to the upper tail of the distribution, i.e. to fractiles greater than 0.5. This is sufficient because the F-distributions has the property

$$F_{F(f_1,f_2)}(x) = 1 - F_{F(f_2,f_1)}(\frac{1}{x}), \quad x > 0. \tag{1}$$

The identity follows because if $X \sim F(f_1, f_2)$, then $1/X \sim F(f_2, f_1)$, so for any $x > 0$

$$F_{F(f_1,f_2)}(x) = P[X \leq x] = P\left[\frac{1}{X} \geq \frac{1}{x}\right] = 1 - P\left[\frac{1}{X} \leq \frac{1}{x}\right] = 1 - F_{F(f_2,f_1)}(\frac{1}{x}).$$

The fact that the calculation of p-values previously relied on tables has influenced the way the F-test for the identity of two variances is presented even in statistical packages of today.

In Example 3.2 the F-test for $H_{0\sigma^2}$: $\sigma_1^2 = \sigma_2^2$ is given as

$$F = \frac{s_{(1)}^2}{s_{(2)}^2} = \frac{763.1397}{4034.873} = 0.1891$$

and the significance probability is

$$p_{obs}(\mathbf{x}) = 2F_{F(16,10)}(0.1891) = 0.0033.$$

Few, if any, tables af the F-distribution can be used to look up the value of 0.1891 directly, but using (1) the significance probability can be calculated as

$$p_{obs}(\mathbf{x}) = 2F_{F(16,10)}(0.1891) = 2\left[1 - F_{F(10,16)}(\frac{1}{0.1891})\right] = 2\left[1 - F_{F(10,16)}(5.29)\right] = 0.0033$$

where 5.29 belong to the upper tail of the $F(10, 16)$ distribution. If instead of F its reciprocal had been calculated

$$\frac{1}{F} = \frac{s_{(2)}^2}{s_{(1)}^2} = \frac{4034.873}{763.1397} = 5.29,$$

it should have been evaluated in an $F(10, 16)$ distribution and the significance probability would be calculated as

$$p_{obs}(\mathbf{x}) = 2\left[1 - F_{F(10,16)}(5.29)\right] = 0.0033.$$

It follows that regardless of whether the test statistic F or its reciprocal F^{-1} is calculated, the same significance probability is obtained, and its calculation is always based on the larger of F or F^{-1}. This has led to the practise always to have the larger of $s_{(1)}^2$ and $s_{(2)}^2$ in the numerator of the F-test statistic so a value greater than 1 is obtained.

STAT, TØ nr. 5

opg. 3.5

1)

Likelihood fkt. for 2 uafh. nom. fordelt obs. med middelværdi μ_1, μ_2 og fælles varians σ^2 er

$$L(\mu_1, \mu_2, \sigma^2, \underline{x}) = \prod_{i=1}^{2} \prod_{j=1}^{n_i} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_{ij}-\mu_i)^2/2\sigma^2}$$

$\underset{(*)}{=} L_1(\mu_1, \sigma^2) \cdot L_2(\mu_2, \sigma^2)$, hvor L_1, L_2 er likelihoodfkt. for hh. 1. og 2. obs. række.

Da $\overline{x}_{1.}$, $\overline{x}_{2.}$ er maks. likelihood est. for hh. μ_1, μ_2 når vi ser på obs. række hver for sig følger dermed af (*), at de også maksimere L, dvs. $\hat{\mu}_1 = \overline{x}_{1.}$, $\hat{\mu}_2 = \overline{x}_{2.}$

$$l(\mu_1, \mu_2, \sigma^2) = \frac{-(n_1+n_2)}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{ij} (x_{ij}-\mu_i)^2$$

$$\leq \frac{-(n_1+n_2)}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{ij} (x_{ij}-\hat{\mu}_i)^2$$

$$= -\frac{n_1+n_2}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}(SSD_{(1)} + SSD_{(2)})$$

$$\frac{\partial l}{\partial \sigma^2}(\hat{\mu}_1, \hat{\mu}_2, \sigma^2) = -\frac{n_1+n_2}{2} \cdot 2\pi \frac{1}{2\pi\sigma^2} + \frac{2}{4(\sigma^2)^2}(SSD_{(1)} + SSD_{(2)})$$

$$\begin{cases} = 0 & \Longleftrightarrow \sigma^2 = (SSD_{(1)} + SSD_{(2)})/(n_1+n_2) \\ > 0, & \sigma^2 < (SSD_{(1)} + SSD_{(2)})/(n_1+n_2) \\ < 0, & \sigma^2 > \text{——} \end{cases}$$

Altså maks. like. est. for σ^2 er $\hat{\sigma}^2 = \dfrac{SSD_{(1)} + SSD_{(2)}}{n_1+n_2}$

2)

$H_0: \mu_1 = \mu_2$

Under hypotesen kan likelihood fkt. betragtes som likelihood fkt. for en normalfordelt obs. række dvs. maks. like. est for μ $(=\mu_1=\mu_2)$ er $\hat{\mu} = \overline{x}_{..}$

Det giver flg. like. ratio test

$$Q(\underline{x}) = \frac{L(\hat{\mu}, \hat{\sigma}^2)}{L(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}^2)} = \frac{\prod_{i=1}^{2} \prod_{j=1}^{n_i} \frac{1}{\sqrt{2\pi\hat{\sigma}^2}} e^{-(x_{ij}-\hat{\mu})^2/2\hat{\sigma}^2}}{\prod_{i=1}^{2} \prod_{j=1}^{n_i} \frac{1}{\sqrt{2\pi\hat{\sigma}^2}} e^{-(x_{ij}-\hat{\mu}_i)^2/2\hat{\sigma}^2}}$$

$$= e^{-\sum_{ij}(x_{ij}-\hat{\mu})^2/2\hat{\sigma}^2 + \sum_{ij}(x_{ij}-\hat{\mu}_i)^2/2\hat{\sigma}^2}$$

Vi ser at

$$-2\ln Q = \frac{1}{\hat{\sigma}^2}\left(\sum_{ij}\left((x_{ij}-\hat{\mu}_i)^2 - (x_{ij}-\hat{\mu})^2\right)\right)$$

$$= \frac{1}{\hat{\sigma}^2}\left(\sum_{ij}\left((x_{ij}-\hat{\mu}_i)^2 - (x_{ij}-\hat{\mu}_i+\hat{\mu}_i-\hat{\mu})^2\right)\right)$$

$$= \frac{1}{\hat{\sigma}^2}\left(\sum_{ij}\left(-(\hat{\mu}_i-\hat{\mu})^2 - 2(x_{ij}-\hat{\mu}_i)(\hat{\mu}_i-\hat{\mu})\right)\right)$$

$$= -\frac{1}{\hat{\sigma}^2}\sum_{ij}(\hat{\mu}_i-\hat{\mu})^2 = -\frac{1}{\hat{\sigma}^2}\sum_{ij}\left(\hat{\mu}_i - \frac{n_1}{n_1+n_2}\hat{\mu}_1 - \frac{n_2}{n_1+n_2}\hat{\mu}_2\right)^2$$

$$= -\frac{1}{\hat{\sigma}^2}\left(n_1\left(\left(1-\frac{n_1}{n_1+n_2}\right)\hat{\mu}_1 - \frac{n_2}{n_1+n_2}\hat{\mu}_2\right)^2 + n_2\left(\left(1-\frac{n_2}{n_1+n_2}\right)\hat{\mu}_2 - \frac{n_1}{n_1+n_2}\hat{\mu}_1\right)^2\right)$$

$$= -\frac{1}{\hat{\sigma}^2}(\hat{\mu}_1-\hat{\mu}_2)^2 \cdot \frac{n_1 n_2^2 + n_2 n_1^2}{(n_1+n_2)^2} = -\frac{(\hat{\mu}_1-\hat{\mu}_2)^2}{\hat{\sigma}^2}\frac{n_1 n_2}{n_1+n_2}$$

(*)

$$= - \left(\frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\hat{\sigma}^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \right)^2 \qquad \frac{n_1 n_2}{n_1 + n_2} = \left(\frac{1}{n_1} + \frac{1}{n_2} \right)^{-1}$$

$$\{ \underline{y} \mid Q(\underline{y}) \leq Q(\underline{x}) \} = \{ \underline{y} \mid -2 \ln Q(\underline{x}) \leq -2 \ln Q(\underline{y}) \}$$

$$= \{ \underline{y} \mid -t(\underline{x})^2 \leq -t(\underline{y})^2 \} = \{ \underline{y} \mid |t(\underline{y})| \leq |t(\underline{x})| \}$$

(*)

Vi har hermed jf. s. 3.35 og s. 1.10 vist at t-testet s. 3.35 er ækvivalent med like. ratio. testet

5)

$H_0: \pi = \pi_0 = 1/4$, $(n=10,\ x=1)$

Helt anlog med 3) fås

$$Q(x) = \frac{(1/4)^x (3/4)^{n-x}}{(x/n)^x (1-x/n)^{n-x}} = \left(\frac{n}{4x}\right)^x \left(\frac{3n}{4(n-x)}\right)^{n-x}$$

Ved inspektion fås

$$\{y \mid Q(y) \leq Q(x)\} = \{1,5,6,7,8,9\}$$

Dvs. $\varepsilon(x) = 1 - P(X=0) - P(X=2) - P(X=3) - P(X=4)$

$$= 1 - (1-1/4)^{10} - \frac{10\cdot 9}{2}(1/4)^2(1-1/4)^8 - 10\cdot 3\cdot 4(1/4)^3(1-1/4)^7$$

$$- 10\cdot 3\cdot 7(1/4)^4(1-1/4)^6 \simeq 0{,}2658 > 0{,}05$$

Dvs. vi forkaster <u>ikke</u> hypotesen $(\pi = 1/4)$

opg. 3.7

$x_i \sim\sim N(\mu, \sigma^2) \qquad i = 1, \dots, n$

1)

Påstand:

Under $H_0: \sigma^2 = \sigma_0^2$ er $Q(\underline{x}) = r^{n/2} e^{-n/2(r-1)}$

hvor $r = \hat{\sigma}^2/\sigma_0^2 = 1/n \sum_{i=1}^{n} (x_i - \bar{x}_.)^2/\sigma_0^2$

Bevis:

$$Q(\underline{x}) = \frac{L(\hat{\mu}, \sigma_0^2; \underline{x})}{L(\hat{\mu}, \hat{\sigma}^2; \underline{x})}$$

$$= \frac{\left(\frac{1}{2\pi\sigma_0^2}\right)^{n/2} e^{-\frac{1}{2\sigma_0^2}\sum_{i=1}^{n}(x_i - \hat{\mu})^2}}{\left(\frac{1}{2\pi\hat{\sigma}^2}\right)^{n/2} e^{-\frac{1}{2\hat{\sigma}^2}\sum_{i=1}^{n}(x_i - \hat{\mu})^2}}$$

$$= \left(\frac{\hat{\sigma}^2}{\sigma_0^2}\right)^{n/2} e^{-\frac{1}{2}\left(\frac{1}{\sigma_0^2} - \frac{1}{\hat{\sigma}^2}\right)\sum_{i=1}^{n}(x_i - \hat{\mu})^2}$$

$$\underset{s.324}{=} \left(\frac{\hat{\sigma}^2}{\sigma_0^2}\right)^{n/2} e^{-\frac{1}{2}\left(\frac{1}{\sigma_0^2} - \frac{1}{\hat{\sigma}^2}\right) n\hat{\sigma}^2} = \left(\frac{\hat{\sigma}^2}{\sigma_0^2}\right)^{n/2} e^{-\frac{n}{2}\left(\frac{\hat{\sigma}^2}{\sigma_0^2} - 1\right)}$$

$$= r^{n/2} e^{-n/2(r-1)}$$

$\square$

2)

Påstand: like. ratio. test. forkaste for store og små værdier af r.

Bevis:

$f: \mathbb{R}_+ \to \mathbb{R}_+ : \quad r \in \mathbb{R}_+ : \; f(r) = r^{n/2} e^{-n/2(r-1)}$

$$f'(r) = \frac{n}{2} r^{n/2-1} e^{-n/2(r-1)} + r^{n/2} (-n/2) e^{-n/2(r-1)}$$

$$= \frac{n}{2} r^{n/2-1} (1-r) e^{-n/2(r-1)}$$

$f'(r) < 0 \quad , \; r > 1$

$f'(r) = 0 \quad , \; r = 1$

$f'(r) > 0 \quad , \; r < 1$

$$\varepsilon(\underline{x}) = P(\underline{y} \mid Q(\underline{y}) \leq Q(\underline{x}))$$

$$= P(\underline{y} \mid f(s) \leq f(\bar{x}) \;, \; s = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y}_.)^2 / \sigma_0^2)$$

$\square$

3)

$n = 10$, $r = 2{,}11364$

3.96 :

$$\frac{n}{n-1} r = \frac{10}{10-1} 2{,}11364 = 2{,}3485 > 1$$

$$\varepsilon(r) = 2\left(1 - F_{\chi^2(f)/f}\left(\frac{n}{n-1} r\right)\right)$$

$$= 2\left(1 - F_{\chi^2(9)/9}\left(\frac{n}{n-1} r\right)\right) < 0{,}05$$

Dvs. hypotesen forkastes.

V.hj.a like. ratio. testet :

$$\varepsilon(r) \underset{\substack{\text{Jf. 2)} \\ \text{og vinket}}}{=} P\left(r \in (0; 0{,}36973) \cup (2{,}11364; \infty)\right)$$

$$= P\left(\frac{10}{9} r \in (0; 0{,}4105) \cup (2{,}3485; \infty)\right)$$

$$\geq 0{,}05 + (1 - 0{,}99) = 0{,}06 > 0{,}05$$

idet $\frac{10}{9} r \sim \chi^2(9)/9$

Dvs. hypotesen forkastes <u>ikke</u>

4)

$n = 13$, $r = 2$

3.96 :

$$\frac{n}{n-1} r = \frac{13}{12} 2 > 1$$

$$\varepsilon(r) = 2\left(1 - F_{\chi^2(12)/12}\left(\tfrac{13}{12} \cdot 2\right)\right) < 2(1 - 0{,}975) = 0{,}05$$

Dvs. hypotesen forkastes

V.hj.a like.ratio. testet:

$$\varepsilon(r) \underset{\substack{\text{Jf. 2)} \\ \text{og vinkel}}}{=} P\left(r \in (0; 0{,}40638) \cup (2; \infty)\right)$$

$$= P\left(\tfrac{13}{12} r \in (0; 0{,}440245) \cup (2{,}16667; \infty)\right)$$

$$\geq 0{,}05 + (1 - 0{,}99) = 0{,}06$$

idet $\frac{13}{12} r \sim \chi^2(12)/12$

Dvs hypotesen forkastes <u>ikke</u>

opg. 3.8

Kyst:

obs	kom.antal	ss
0,1	1	10%
0,5	2	30%
0,8	3	50%
1,2	4	70%
2,1	5	90%

$S = 4,7 \quad USS = 6,75$

By:

obs	kom.antal	ss
2,0	1	6,25%
5,0	3	25%
14,0	4	43,75%
15,0	5	56,25%
17,0	6	68,75%
22,0	7	81,25%
25,0	8	93,75%

$S = 105 \quad USS = 1873$

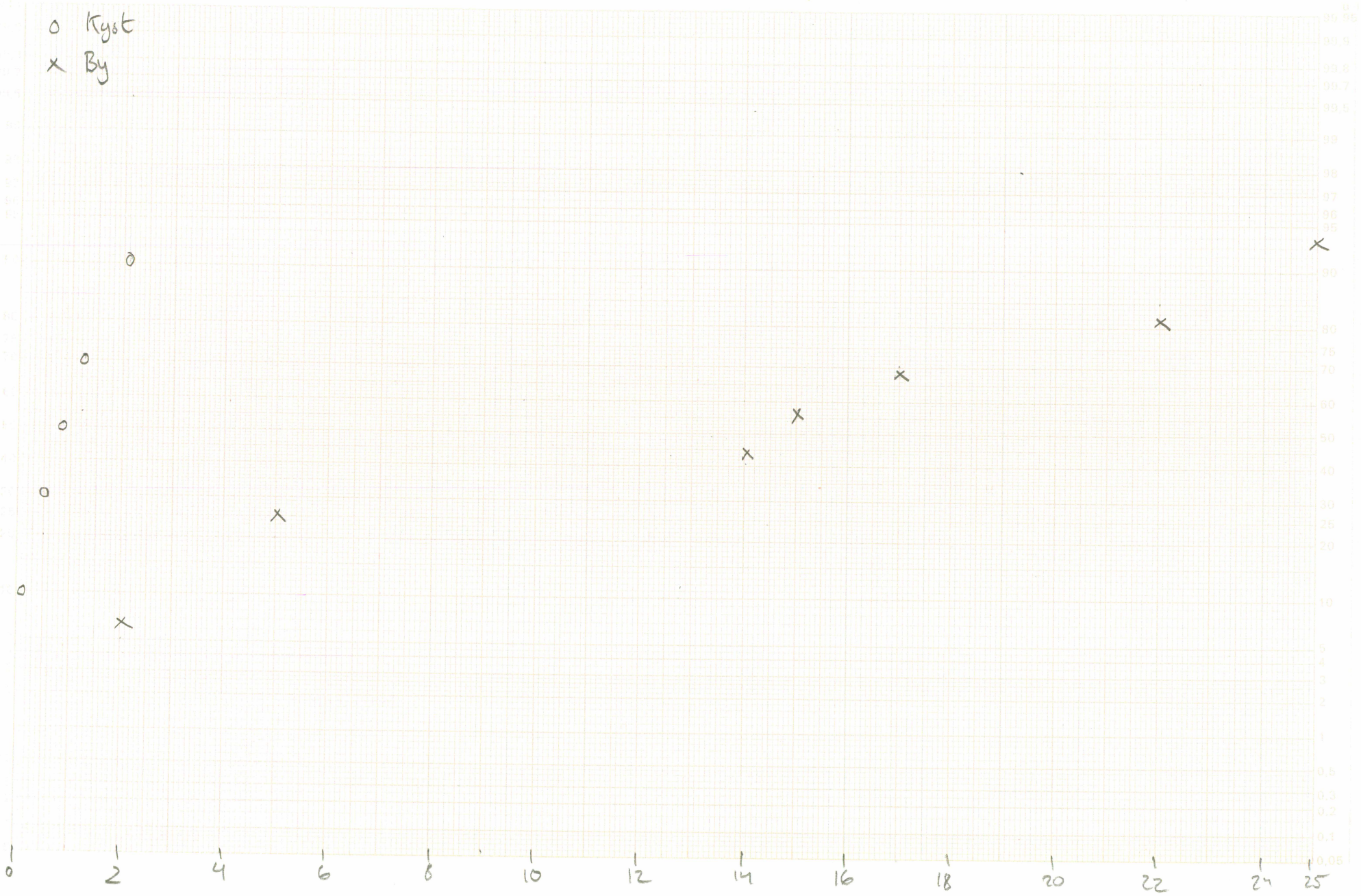
Som vedlagte fraktildiagram peger på, kan vi antage, at vi har to normalfordelte obs.vækker – intet tyder dog på ens middelværdi eller varians. (hæld. + afskæring)

H_0: $\sigma^2_{kyst} = \sigma^2_{by}$

$$F = \frac{s^2_{by}}{s^2_{kyst}} = \frac{70,696}{0,583} \simeq 121,2632$$

$$\varepsilon(\underline{x}) = 2(1 - F_{F(7,4)}(F)) < 2(1 - 0,9995) = 0,001$$

Hypotesen forkastes


o Kyst
x By
0
2
4
6
8
10
12
14
16
18
20
22
24
25

$H_{0\mu}: \mu_{kyst} = \mu_{by}$

$$t(\underline{x}) = \frac{\hat{\mu}_{by} - \hat{\mu}_{kyst}}{\sqrt{s_{by}^2/8 + s_{kyst}^2/5}} = 4{,}0722$$

$$\varepsilon(x) = 2(1 - F_{t(\bar{f})}|t(\underline{x})|) \leq 2(1 - 0{,}995) = 0{,}01 < 0{,}0$$

$$\bar{f} \underset{s.3.51}{=} 7{,}18$$

Dvs. hypotesen forkastes og slutmodellen er derfor to normalfordelte obs. rækker med forskellig middelværdi og varians

95% konfidensint. for $\mu_{by} - \mu_{kyst}$ (jvf. s.3.50)

$[5{,}108 \; ; \; 19{,}262]$

Dvs. der er en <u>betydelig</u> forskel på luftforureningen v. kyst og i by

OPG. på US

Vi vil antage, at vi har to normalfordelte obs. rækker

Dvs. grundmodellen ($i=1$ 1998, $i=2$ 1999) er

M_0: $x_{ij} \sim\sim N(\mu_i, \sigma_i^2)$ $j=1,\dots,n_i$ $(n_1=16, n_2=14)$

$H_{0\sigma}$: $\sigma_1^2 = \sigma_2^2$

(Notation se s. 3.67)

$f_1 = 16-1+14-1 = 28$, $f_{(1)} = 15$, $f_{(2)} = 13$

$$s_{(1)}^2 = \frac{1}{16-1}\left(USS_{(1)} - \frac{S_{(1)}^2}{16}\right) = 0{,}13215625$$

$$s_{(2)}^2 = \frac{1}{14-1}\left(USS_{(2)} - \frac{S_{(2)}^2}{14}\right) = 0{,}29549505$$

$s_1^2 = 0{,}20799212$

$-2\ln Q(\underline{x}) = 2{,}237754$

$$C = 1 + \frac{1}{3}\left(\frac{1}{15} + \frac{1}{13} - \frac{1}{28}\right) = 1{,}03595849$$

$$Ba = \frac{-2\ln Q(\underline{x})}{C} = 2{,}1601$$

$\varepsilon(\underline{x}) = 1 - F_{\chi^2(1)}(Ba) \geq 1 - 0{,}9 = 0{,}1 > 0{,}05$

Altså forkaster vi ikke hypotesen og dermed reducerer vi til flg. model:

M_1: $x_{ij} \sim N(\mu_i, \sigma^2)$

$H_{0\mu}: \mu_1 = \mu_2$

$s_2^2 = 0{,}1935005$

$F = \frac{s_2^2}{s_1^2} = 0{,}930326$

$\varepsilon(x) = 1 - F_{F(1,28)}(F) \geq 1 - 0{,}7 = 0{,}3 > 0{,}05$

Altså forkaster vi <u>ikke</u> hypotesen og reducerer dermed til modellen. –

$M_2: x_{ij} \sim N(\mu, \sigma^2)$

$\mu \leftarrow \bar{x}_{..} = 3{,}892$, $\sigma^2 \leftarrow s^2 = 0{,}20749$

95% konfidensint. for μ, σ^2 (s. 3.30):

μ: $[3{,}6719 ; 4{,}0121]$, σ^2: $[0{,}1317 ; 0{,}3761]$

STAT 1, aflæveringsopgave nr. 3, Gunna Hellmund.

Udfra vedlagte fraktildiag. har vi som udgangspkt. modellen

$M_0: X_{ij} \sim\sim N(\mu_i, \sigma_i^2)$ (i=1 serie 6, i=2 serie 1)

$H_{0\sigma^2}: \sigma_1^2 = \sigma_2^2$

Vi benytte vedlagt standardberegninger

$f_1 = 20 + 17 - 2 = 35$

$$s_1^2 = \frac{USS_{(1)} - s_{(1)}^2/n_1 + USS_{(2)} - s_{(2)}^2/n_2}{f_1} = 27{,}8760$$

$$C = 1 + \frac{1}{3}\left(\frac{1}{16} + \frac{1}{19} - \frac{1}{35}\right) = 1{,}0289$$

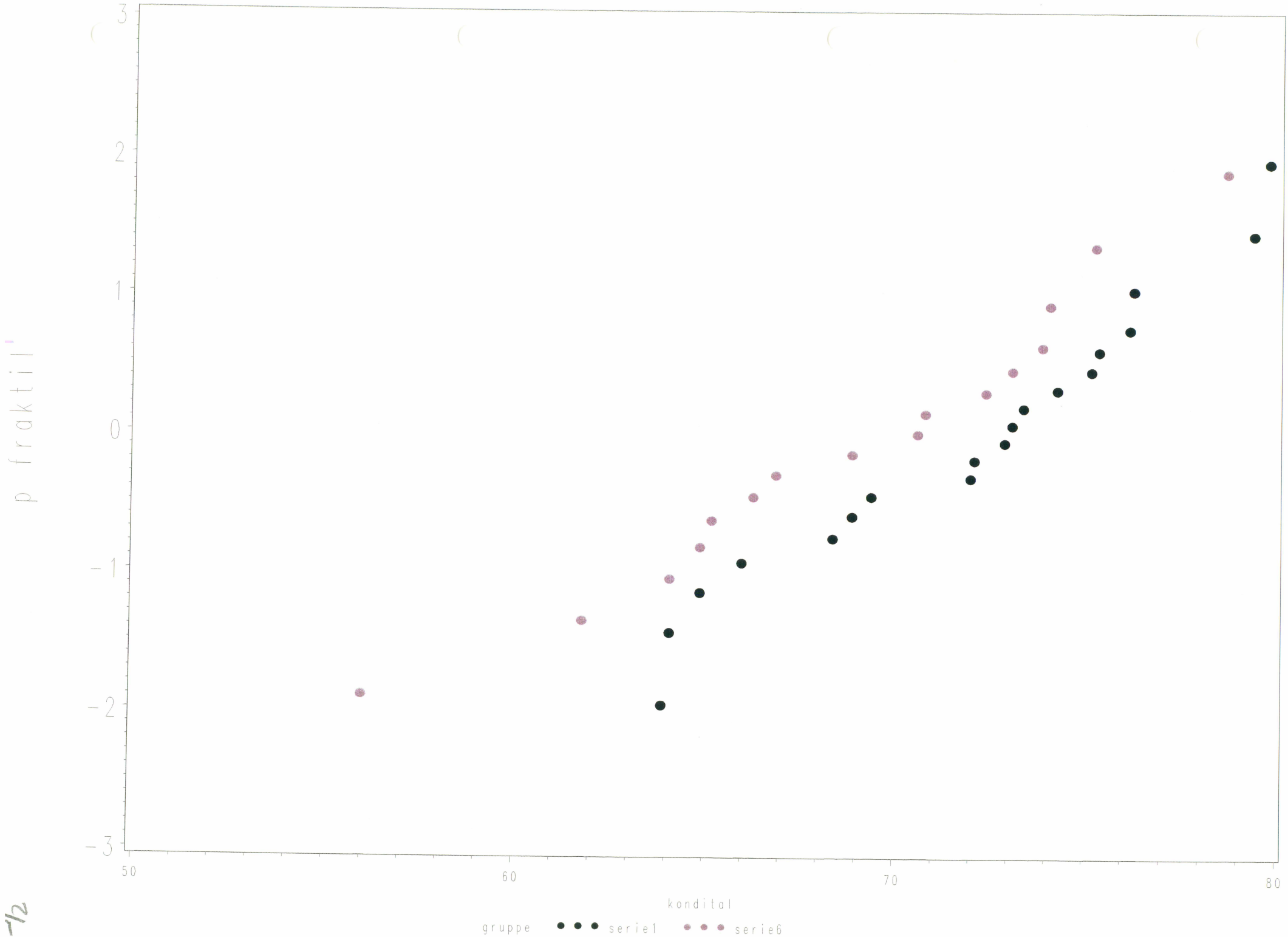
$-2 \cdot \ln Q(x) = 0{,}52998$

$Ba = 0{,}5151$

$\varepsilon(x) = 1 - F_{\chi^2(1)}(Ba) \geq 1 - 0{,}6 = 0{,}4 > 0{,}05$

Dvs. vi forkaste ikke hypotesen og vor model blir derfor

$M_1: X_{ij} \sim\sim N(\mu_i, \sigma^2)$



-12

---------- gruppe=serie1 ----------

The MEANS Procedure

Analysis Variable : kondital

Sum	Mean	USS	Corrected SS	Variance
1442.00	72.1000000	104413.06	444.8600000	23.4136842

---------- gruppe=serie6 ----------

Analysis Variable : kondital

Sum	Mean	USS	Corrected SS	Variance
1177.10	69.2411765	82034.59	530.8011765	33.1750735

The TTEST Procedure

Statistics

Variable	gruppe	N	Lower CL Mean	Mean	Upper CL Mean	Lower CL Std Dev	Std Dev	Upper CL Std Dev	Std Err
kondital	serie1	20	69.835	72.1	74.365	3.6798	4.8388	7.0674	1.082
kondital	serie6	17	66.28	69.241	72.203	4.2897	5.7598	8.766	1.397
kondital	Diff (1-2)		-0.677	2.8588	6.3947	4.2823	5.2798	6.8871	1.7417

T-Tests

Variable	Method	Variances	DF	t Value	Pr > |t|
kondital	Pooled	Equal	35	1.64	0.1097
kondital	Satterthwaite	Unequal	31.4	1.62	0.1157

Equality of Variances

Variable	Method	Num DF	Den DF	F Value	Pr > F
kondital	Folded F	16	19	1.42	0.4645

$H_{0\mu}: \mu_1 = \mu_2$

$s_2^2 = 75,10207$

$F = \frac{s_2^2}{s_1^2} = 2,6941$

$\varepsilon(x) = 1 - F_{F(1,35)}(F) > 1 - 0,9 = 0,1 > 0,05$

Dvs. hypotesen forkastes ikke og modellen er derfor

$M_2: x_{ij} \sim\sim N(\mu, \sigma^2)$

Estimat og konfidens int. i slutmodellen:

$\mu \leftarrow \bar{x}_{..} = 70,79$

$[68,983 ; 72,509]$

$\sigma^2 \leftarrow s^2 = \frac{USS_{(1)} + USS_{(2)} - (S_{(1)} + S_{(2)})^2/37}{37-1} = 29,188$

$[19,3155 ; 49,3316]$

Vi kan altså konkludere at der ikke er signifikant forskel på konditallene i serie 6 og serie 1.

Dette ses desuden også når man betragter konfidensintervallet for $\mu_2 - \mu_1$: (jvf. s. 3.50): $[-0,75909 ; 6,47674]$