Afdeling for Teoretisk Statistik
Institut for Matematiske Fag
Aarhus Universitet

Statistik 1 og Statistik α
Jørgen Granfeldt
22. november 2002

Meddelelse 12

Forelæsningerne i sidste uge

(Onsdag den 13. november) Kapitel 6 *Multinomialfordelte data*. Hypoteser, estimation under hypotese og test af hypotese som på side 6.7 til 6.10. Desuden Eksempel 6.1 side 6.18 og 6.19.

(Fredag den 15. november) Kapitel 6 *Multinomialfordelte data*. Konfidensinterval for sandsynlighed og uafhængighed af inddelingskriterier til nederst side 6.24.

Forelæsningerne i denne uge

(Onsdag den 20. november) Kapitel 6 *Multinomialfordelte data*. Afsluttede uafhængighed af inddelingskriterier nåede frem til midt på side 6.29 i Afsnit 6.3 *Inferens i flere multinomialfordelinger*.

(Fredag den 22. november) Afslutter Kapitel 6 *Multinomialfordelte data*.

Forelæsningerne i næste uge

(Onsdag den 27. og fredag den 29. november) Kapitel 7 *Poissonfordelte data*.

Øvelserne i denne uge

(Uge 47: 18. november - 24. november) 4.11, 6.1, 6.4, 6.9, 6.10.

Øvelserne i de tre næste uger

(Uge 48: 25. november - 1. december) 6.11, 7.6, 7.8, opgave 1 og 2 på Meddelelse 11, samt eventuelle manglende opgaver fra tidligere øvelsesgange.

(Uge 49: 2. december - 8. december) 7.14, 7.16, 7.17 samt opgaverne 1 og 2, som stilles på Meddelelse 12.

Afleveringsopgaver

(Uge 47) Ingen afleveringsopgave.

(Uge 48) Ingen afleveringsopgave.

(Uge 49) Eksamensopgave 2 i Statistik α Sommeren 2002. Opgavesættet for sommeren 2002 findes på adressen

`http://home.imf.au.dk/statbib/atskurser/statalfa/eksamen/2002/`

(Uge 50) Ingen afleveringsopgave.

Udleveret materiale

3. obligatoriske opgave

Gamle eksamensopgaver

Tre tidligere eksamenssæt i Statistik α kan findes på adressen

`http://home.imf.au.dk/statbib/atskurser/statalfa/eksamen`

Opgave 1

Ved 10 prøjsiske kavalleriregimenter optalte man over 20 på hinanden følgende år (1875-94) antallet af soldater, der omkom som følge af hestespark. Resultatet er gengivet i Tabel 1.

Antal dødsfald	*Antal regimentår*
0	109
1	65
2	22
3	3
4	1

Tabel 1: Antal døde ved hestespark i 10 prøjsiske regimenter 1875-94.

1) Vis, at det kan antages at antallet af dødsfald per år er Poissonfordelt.

2) Estimer middelværdien og angiv et 95% konfidensinterval for middelværdien.

Opgave 2

Det foreliggende materiale stammer fra en amerikansk undersøgelse af en art plettyfus *Rocky Mountain Spotted Fever*, der spredes med blodmider som smittebærere.

Blandt 747 tilfælde af sygdommen i den vestlige del af USA var der 210 dødsfald, mens der blandt 661 tilfælde i den østlige del af USA var 122 dødsfald. Umiddelbart kunne det derfor se ud som om der var en "vestlig" og en "østlig" type af sygdommer, hvor den "vestlige" var den alvorligste. Imidlertid viser mikrobiologiske undersøgelser ingen forskel mellem de "vestlige" og de "østlige" bakteriestammer.

I Tabel 2 er angivet det totale antal tilfælde og antallet af tilfælde med dødelig udgang i henholdsvis den vestlige og den østlige del af USA delt op i aldersgrupper.

1) Undersøg om der på grundlag af dette materiale kan antages at være forskel på den "vestlige" og den "østlige" type plettyfus.

	"vestlig" type		*"østlig" type*	
alder i år	*totalt antal tilfælde*	*tilfælde med dødelig udgang*	*totalt antal tilfælde*	*tilfælde med dødelig udgang*
under 15	108 95	13	310 270	40
15-39	264 224	40	189 168	21
40 og derover	375 218	157	162 101	61
	747	210	661	122

Tabel 2: Totalt antal tilfælde og tilfælde med dødelig udgang af *Rocky Mountain Spotted Fever* af henholdsvis "vestlig" og "østlig" type .

Opgave 2

Data stammer fra en undersøgelse af sammenhængen mellem anvendelse af fysisk afstraffelse, politisk tilhørsforhold og oplevelse af fysisk afstraffelse i barndommen. Undersøgelsen er foretaget af Gallup Instituttet.

Første del af undersøgelsen omfatter 162 kvinder, der alle har oplevet fysisk afstraffelse i barndommen. I tabellen nedenfor er de opdelt efter om de selv anvender fysisk afstraffelse (ja, nej) og deres politiske tilhørsforhold, der er opgjort som enten socialdemokratiet (s), partier, der anses for at være til venstre (v) eller til højre (h) for dette parti.

	v	s	h	Ialt
ja	12	27	58	97
nej	7	28	30	65
Ialt	19	55	88	162

1) Vis, at det kan antages, at der er uafhængighed mellem anvendelse af fysisk afstraffelse og politisk tilhørsforhold.

2) Angiv et 95% konfidensinterval for sandsynligheden for at anvende fysisk afstraffelse, når man selv har oplevet fysisk afstraffelse i barndommen.

I anden del af undersøgelsen undersøgte man de samme spørgsmål, men nu for ialt 75 kvinder, der *ikke* havde oplevet fysisk afstraffelse i barndommen. Blandt disse kvinder var der også uafhængighed mellem anvendelse af fysisk afstraffelse og politisk tilhørsforhold.

3) Blandt de 75 kvinder, der ikke havde oplevet fysisk afstraffelse i barndommen, anvendte 23 selv fysisk afstraffelse. Undersøg om sandsynligheden for at anvende fysisk afstraffelse afhænger af, om man selv har oplevet fysisk afstraffelse i barndommen.

Opgave 3

Diffusions MRI (Magnetic Resonance Imaging) er en moderne scanningsmetode, der blandt andet kan anvendes til at bestemme axonernes forløb i den menneskelige hjerne inden kirurgiske indgreb. Metoden måler diffusion af vandmolekyler i en bestemt retning ved hjælp af magnetfelter.

Gentagne bestemmelser af diffusionen i 3 forskellige retninger er vist i nedenstående tabel for et bestemt område af hjernen. Disse data udgør en del af en større undersøgelse af en forsøgsperson.

Opgaven fortsættes

3 obligatoriske opgave, Stat 1, Gunnar Hellmund, 19960312

Opgave 3a

1)

Likelihoodfunktionen under H_1 er:

$$L(p) = \frac{n!}{x_1!\, x_2!\, x_3!}\, p^{x_1} (1-2p)^{x_2} p^{x_3} = \frac{n!}{x_1!\, x_2!\, x_3!}\, p^{x_1+x_3} (1-2p)^{x_2}$$

$$= \frac{n!}{x_1!\, x_2!\, x_3!\, 2^{x_1+x_3}} (2p)^{x_1+x_3} (1-2p)^{x_2}$$

Jf. sætn. 6.1 har $2p \longmapsto (2p)^{x_1+x_3}(1-2p)^{x_2}$ maksimum i $\hat{2p} = \frac{x_1+x_3}{n}$, dvs. maksimum likelihood estimatet $\hat{p} = \frac{x_1+x_3}{2n}$

Ja, for $2p \in \Pi$ def. p.6.5. Det bør overvejes, at dette svarer til den afstuckte def. mgd.

2)

Jf. side 6.1, 6.7, 6.9 fås, at det app. likelihood ratio test er

$$-2 \ln Q(\underline{x}) = 2 \sum x_i \cdot \ln\left(\frac{x_i}{e_i}\right)$$

$$= 2\left(x_1 \cdot \ln\left(\frac{x_1}{n\frac{x_1+x_3}{2n}}\right) + x_2 \cdot \ln\left(\frac{x_2}{n(1-2\frac{x_1+x_3}{2n})}\right) + x_3 \cdot \ln\left(\frac{x_3}{n\frac{x_1+x_3}{2n}}\right)\right)$$

$$= 2\left(x_1 \cdot \ln\left(\frac{2x_1}{x_1+x_3}\right) + x_3 \cdot \ln\left(\frac{2x_3}{x_1+x_3}\right)\right) \simeq 3{,}2672$$

Testss. for H_1 under H_0 bliver $\varepsilon(x) = 1 - F_{\chi^2(3-1-1)}(3{,}2672) > 0{,}05$

Dvs. H_1 ~~må~~ kan! antages at være sand under H_0

Testet brugbart hvis $e_i > 5$ og H_1 er en glat hypotese

3)

Jf. side 1.17 fås $-2 \ln Q(\underline{x}) = -2\left(\ell\left(\frac{1}{3}\right) - \ell(\hat{p})\right)$

$$= 2\left(x_1 \ln(\hat{p}) + x_2 \ln(1-2\hat{p}) + x_3 \ln(\hat{p}) - (x_1+x_2+x_3)\ln\left(\frac{1}{3}\right)\right)$$

$$= 2\left((x_1+x_3) \cdot \ln\left(\frac{x_1+x_3}{2n}\right) + x_2 \ln\left(\frac{x_2}{n}\right) + n \cdot \ln(3)\right) \simeq 3{,}1732$$

Testss. for H_2 under H_1 bliver $\varepsilon(x) = 1 - F_{\chi^2(1-0)}(3{,}1732) > 0{,}05$ (forv. > 5)

Dvs. H_2 må antages at være sand under H_1

4)

Jf. kap. 6 $-2\ln Q(x) = 2\sum x_i \ln\left(\frac{x_i}{e_i}\right) = 2\sum x_i \ln\left(\frac{x_i}{\frac{1}{3}n}\right)$

$= 2\sum x_i \cdot \ln\left(\frac{3x_i}{n}\right) \simeq 6{,}4404$ V

Testss. for H_2 under H_0 bliver $\varepsilon(x) = 1 - F_{\chi^2(3-1-0)}(6{,}4404) < 0{,}05$ V

Dvs. H_2 under H_0 forkastes V

Resultaterne i 2)-4) kommer ikke bag på os, da følgen af modeller i 2)-3) er en anden end i 4). Generelt kan man sige at udregningerne i 2)-3) er at foretrække frem for 4), idet vi i 2)-3) smider færest informationer væk pr. test (V)

(Hellere en række af hvidlegne, fremfor en forbundet legn.) Hm

Trist at du har fjernet mørsomhederne. Men jeg synes man kunne sige noget mere til dette også af serios art. Ex: $-2\ln Q_{02} = (-2\ln Q_{01}) + (-2\ln Q_{12})$, og hvis du havde angivet de kontrolte test-ssh'er istedet for blot $\geq 0{,}05$ ville der også være bedre basis for at overvåge resultatet af de to test-procedurer.

Opgave 3b

1)

$$L_2(\lambda_{..}, \underline{\tau}) = \prod_{ij} e^{-\lambda_{..}\tau_i\tau_j} \frac{(\lambda_{..}\tau_i\tau_j)^{x_{ij}}}{x_{ij}!} = e^{-\lambda_{..}\sum_{ij}\tau_i\tau_j} \lambda_{..}^{\sum_{ij} x_{ij}} \prod_i \tau_i^{\sum x_{ij}} \prod_j \tau_j^{\sum_i x_{ij}} \prod_{ij} \frac{1}{x_{ij}!}$$

$$= e^{-\lambda_{..}} \lambda_{..}^{x_{..}} \prod_i \tau_i^{x_{i.}+x_{.i}} \prod_{ij} \frac{1}{x_{ij}!}$$, sætn. 7.1 & 6.1 giver: $\hat{\lambda} = x_{..}$, $\hat{\tau}_i = \frac{x_{i.}+x_{.i}}{2x_{..}}$

Noget spinkel redegørelse.

2)

$$\ell_2(\hat{\lambda}_{..}, \hat{\underline{\tau}}) = -\hat{\lambda}_{..} + x_{..}\ln(\hat{\lambda}_{..}) + \sum_i (x_{i.}+x_{.i})\cdot\ln(\hat{\tau}_i) - \sum_{ij}\ln(x_{ij}!)$$

$$= -x_{..} + x_{..}\ln(x_{..}) + \sum_i (x_{i.}+x_{.i})\cdot\ln(x_{i.}+x_{.i}) - \sum_i (x_{i.}+x_{.i})\cdot\ln(2x_{..}) - \sum_{ij}\ln(x_{ij}!)$$

$$= -x_{..} + x_{..}\ln(x_{..}) + \sum_i (x_{i.}+x_{.i})\cdot\ln(x_{i.}+x_{.i}) - 2x_{..}\ln(2x_{..}) - \sum_{ij}\ln(x_{ij}!)$$ V (*)

$$-2\ln Q = 2(\ell_1(\hat{\lambda}_{..}, \hat{\rho}_i, \hat{\sigma}_j) - \ell_2(\hat{\lambda}_{..}, \hat{\tau}_i))$$

$$\underset{(7.44)+*}{=} 2\Big(-x_{..} + \sum_i x_{i.}\ln(x_{i.}) + \sum_j x_{.j}\ln(x_{.j}) - x_{..}\ln(x_{..}) - \sum_{ij}\ln(x_{ij}!)$$ V

$$- \Big(-x_{..} + x_{..}\ln(x_{..}) + \sum_i (x_{i.}+x_{.i})\ln(x_{i.}+x_{.i}) - 2x_{..}\ln(2x_{..}) - \sum_{ij}\ln(x_{ij}!)\Big)$$

$$= 2\Big(\sum_i x_{i.}\ln(x_{i.}) + \sum_j x_{.j}\ln(x_{.j}) - \sum_i (x_{i.}+x_{.i})\cdot\ln(x_{i.}+x_{.i}) + 2x_{..}\ln(2)\Big)$$ V

3)

Da vi går udfra, at det totale antal opererede er tilfældigt, er det udfra kap. 7.2 klart (snarere: nærmeligt!) at grundmodellen er $M_0: x_{ij} \sim\sim po(\lambda_{ij})$ $i,j=1$

$H_0: M_0 \to M_1$ $\quad (M_1: x_{ij} \sim\sim po(\lambda_{..}\rho_i\sigma_j)$ $i,j=1,..,4)$

Testss for H_0 er jf. udskriften $\varepsilon(x) = 1 - F_{\chi^2(16-7)}(-2\ln Q(x)) = 0{,}888 > 0{,}05$

Dvs. data kan beskrives v.hj.a. den multiplikative Poissonmodel.

4)

Vi skal undersøge $H_0: M_1 \to M_2$, $M_2: x_{ij} \sim\sim po(\lambda_{..}\tau_j)$, $i=1,..,4$

Jf. 2) har vi, at den app. likelihood ratio teststørrelse er (Q er ikke appr.!)

$$-2\ln Q(x) = 2\left(\sum_i x_{i.}\ln(x_{i.}) + \sum_j x_{.j}\ln(x_{.j}) - \sum_i (x_{i.}+x_{.i})\ln(x_{i.}+x_{.i})\right) + 2x_{..}\ln(2)$$

$$= 2(889{,}87 + 885{,}87 - 2077{,}99 + 303{,}60) \simeq 2{,}6810$$

Dvs. testss. for H_0 er $\varepsilon(x) = 1 - F_{\chi^2(7-4)}(2{,}6810) > 0{,}4 > 0{,}05$

Altså har behandlingen samme virkning for højre og venstre øje.

Extra:

Konfidensintervaller for hh. $\lambda_{..}$ og $\lambda_{..}\tau_4^2$ $(\alpha = 0{,}05)$:

$u_{0{,}975} = 1{,}960$, jf. (7.19-20) har vi:

$C_{0{,}95}(\lambda_{..}) = [191{,}85; 249{,}99]$

$C_{0{,}95}(\lambda_{..}\tau_4^2) = [21{,}01; 42{,}83]$

er dette meningsfuldt? (estimation af τ_4 (og ikke $\tau_4\tau_4$!) efter 6.27-6.28 forekommer mig mere informativt.)

Vurderer vi $C_{0{,}95}(\lambda_{..}\tau_4^2)$ i forhold til $C_{0{,}95}(\lambda_{..})$, ser vi, at behandlingen ikke ligefrem er effektiv.

(Når et mål for 'effektivitet' er, at operationen er vellykket (efter en observationsperiode på 1 md.) på begge øjne: Hmm.

En besvarelse, hvor du har været lidt for 'økonomisk' til at opnå en rimelig udnyttelse af dit potentiale. (Dog, lige netop:)

Godkendt 11/12-02 Morten Sørensen

STAT1, TØ nr. 11

opgave 4.11

$L = \text{span}(\underline{e}, \underline{t})$, $\underline{e} = (1, \ldots, 1)^*$, $\underline{t} = (t_1, \ldots, t_n)^*$

Definer $\underline{u} = \underline{t} - \bar{t} \cdot \underline{e}$

(Jf. opg. på US 9 er $\underline{u} \perp \underline{e}$)

Det er klart, at $sp(\underline{e}, \underline{u}) = sp(\underline{e}, \underline{t}) = L$

Idet $\|\underline{e}\|^2 = n$, $\|\underline{u}\|^2 = USS_t - \frac{S_t^2}{n}$ er en ortonormal basis for L: $(\frac{1}{\sqrt{n}}\underline{e}, (USS_t - S_t^2/n)^{-1/2}\,\underline{u}) = (\underline{b}_1, \underline{b}_2)$

Påstand: $\bar{\underline{X}}.$, $\hat{\beta}(\underline{X})$, $S_{02}^2(\underline{X})$ er uafh.

Bevis:

Som i beviset på s. 4.52-53 ser vi, at vi kan vælge $\underline{b}_3, \ldots, \underline{b}_n$ så $(\underline{b}_1, \ldots, \underline{b}_n)$ er en ortonormalbasis for $\mathbb{R}^n$ og $\underline{b}_1^* \underline{X}, \ldots, \underline{b}_n^* \underline{X}$ er uafh.

Projektionen ned på L er bestemt ved matricen $\begin{pmatrix} \underline{b}_1^* \\ \underline{b}_2^* \end{pmatrix}$ og idet $\underline{b}_1^* \underline{X}$, $\underline{b}_2^* \underline{X}$ og $(\underline{b}_3^* \underline{X}, \ldots, \underline{b}_n^* \underline{X})$ er uafh.

$\bar{\underline{X}}. = \frac{1}{\sqrt{n}} \underline{b}_1^* \underline{X}$, $\hat{\beta}(\underline{X}) = (USS_t - S_t^2/n)^{-1/2}\, \underline{b}_2^* \underline{X}$

$$S_{02}^2(\underline{X}) = \frac{1}{n-2} \| \underline{X} - (\underline{b}_1^* \underline{X})\underline{b}_1 - (\underline{b}_2^* \underline{X})\underline{b}_2 \|^2$$

$$= \frac{1}{n-2} \sum_{i=3}^{n} (\underline{b}_i^* \underline{X})^2 \quad \text{uafh.}$$

$\square$

Jf. prop. 9.3 følger (under hypotesen om regg.)

$$\underline{b}_1^* \underline{X} \sim N\left(\sum_{i=1}^{n} (\alpha + \beta t_i)/\sqrt{n}\ ,\ \sigma^2\right)$$

, dvs. $\overline{X}_\cdot \sim N\left(\sum_{i=1}^{n} (\alpha + \beta t_i)/n\ ,\ \sigma^2/n\right)$

$$\sim N(\alpha + \beta \overline{t}_\cdot\ ,\ \sigma^2/n)$$

$$\underline{b}_2^* \underline{X} \sim N\left(\beta / (USS_t - S_t^2/n)^{-1/2}\ ,\ \sigma^2\right)$$

idet $\left\langle \underline{t} - \overline{t}_\cdot \underline{e}\ ,\ \begin{pmatrix} \alpha + \beta t_1 \\ \vdots \\ \alpha + \beta t_n \end{pmatrix} \right\rangle$

$$= \alpha S_t + \beta USS_t - \alpha \overline{t}_\cdot n - \beta \overline{t}_\cdot S_t$$

$$= \beta (USS_t - S_t^2/n) = \beta SSD_t$$

, dvs. $\hat{\beta}(\underline{X}) \sim N(\beta\ ,\ \sigma^2/SSD_t)$

Ifølge modellen er $E[\underline{b}_i^* \underline{X}] = 0$, $i = 3, \ldots, n$

Dvs. $\underline{b}_i^* \underline{X} \sim N(0, \underline{b}_i^* \underline{b}_i \sigma^2) \sim N(0, \sigma^2)$, $i = 3, \ldots, n$

Jf. 3.75 gælder $(\underline{b}_i^* \underline{X}/\sigma)^2 \sim \chi^2(1)$, $i = 3, \ldots, n$

og dermed jf. 3.75 $\frac{1}{\sigma^2} \sum_{i=3}^{n} (\underline{b}_i^* \underline{X})^2 \sim \chi^2(n-2)$

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

opgave 6.1

Påstand:

Antag $x_1 + \dots + x_k = n$ $\quad \forall i = 1, \dots, k: \; x_i \in \mathbb{N}$

$\pi_1 + \dots + \pi_k = 1$ $\quad \forall i = 1, \dots, k: \; \pi_i \in (0,1)$

Funktionen $(\pi_1, \dots, \pi_k) \longmapsto \pi_1^{x_1} \cdot \dots \cdot \pi_k^{x_k}$

antager sit maksimum i $(\pi_1, \dots, \pi_k) = (x_1/n, \dots, x_k/n)$

Bevis:

Induktion:

$k=2$:

$(\pi, \pi-1) \longmapsto \pi^x (1-\pi)^{n-x}$ (kald fkt. g)

$$D[g] = x \pi^{x-1} (1-\pi)^{n-x} - \pi^x (n-x)(1-\pi)^{n-x-1}$$

$$= \pi^{x-1} (1-\pi)^{n-x-1} (x - \pi x - \pi n + \pi x)$$

$$= \pi^{x-1} (1-\pi)^{n-x-1} (x - \pi n)$$

Vi ser at g har maksimum i $\hat{\pi} = x/n$

$(\hat{\pi}, 1-\hat{\pi}) = (x/n, \frac{n-x}{n})$

Antag at påstanden er sand for $k \geq 2$

$$g : (\pi_1, \dots, \pi_{k+1}) \longmapsto \pi_1^{x_1} \cdot \dots \cdot \pi_{k+1}^{x_{k+1}}$$

$$\pi_1 + \dots + \pi_{k+1} = 1 \quad , \quad x_1 + \dots + x_{k+1} = n$$

$$\pi_1^{x_1} \cdot \dots \cdot \pi_{k+1}^{x_{k+1}} = \left(\frac{\pi_1}{\pi}\right)^{x_1} \cdot \dots \cdot \left(\frac{\pi_k}{\pi}\right)^{x_k} \pi^{n - x_{k+1}} (1-\pi)^{x_{k+1}}$$

$$\pi := \pi_1 + \dots + \pi_k$$

Af induktionsantagelsen og tilfældet $k=2$ følger

$$\pi_1^{x_1} \cdot \dots \cdot \pi_{k+1}^{x_{k+1}} \leq \left(\frac{x_1}{n - x_{k+1}}\right)^{x_1} \cdot \dots \cdot \left(\frac{x_k}{n - x_{k+1}}\right)^{x_k} \left(\frac{n - x_{k+1}}{n}\right)^{n - x_{k+1}} \left(\frac{x_{k+1}}{n}\right)^{x_{k+1}}$$

$$= \left(\frac{x_1}{n}\right)^{x_1} \cdot \dots \cdot \left(\frac{x_{k+1}}{n}\right)^{x_{k+1}}$$

$\square$

opgave 6.4

$\pi :]0,1[\longrightarrow \Pi$

$\pi(p) : (p^2, 2p(1-p), (1-p)^2)^*$

$]0,1[\subseteq \mathbb{R}$, $\dim \mathbb{R} = 1 \leq 3 = \dim \mathbb{R}^3$

og $]0,1[$ er et område (åbn kurvesammenhængende) af $\mathbb{R}$

Det er klart, at $\pi \in C^\infty(]0,1[)$, specielt C^2

og at π er en-entydig.

$$i(p) \underset{5.6.52}{=} \left\{ \frac{1}{p^2}(2p)^2 + \frac{1}{2p(1-p)}(2(1-p)-2p)^2 + \frac{1}{(1-p)^2}(-2(1-p))^2 \right.$$

$$= 4 + \frac{2(1-2p)^2}{p(1-p)} + 4 = 8 + \frac{2(1-2p)^2}{p(1-p)}$$

Da $i(p)$ er 'invertibel' for alle værdier af p ($\in]0,1[$

er hypotesen om Hardy-Weinberg ligevægt

en <u>glat</u> hypotese (af orden 1).

Antag modellen för Hardy-Weinberg ligvikt:

M_1: $(X_1, X_2, X_3) \sim m(n, (p^2, 2p(1-p), (1-p)^2))$

Påstand: $2X_1 + X_2 \sim b(2n, p)$

Bevis:

$$P(2X_1 + X_2 = z) = \sum_{\{(x_1,x_2,x_3):\ 2x_1+x_2=z\}} \binom{n}{x_1\ x_2\ x_3} p^{2x_1} 2^{x_2} p^{x_2} (1-p)^{x_2} (1-p)^{2x_3}$$

$$= \sum_{2x_1+x_2=z} \binom{n}{x_1\ x_2\ x_3} 2^{x_2} \quad p^{2x_1+x_2} (1-p)^{x_2+2x_3}$$

$$= \sum_{2x_1+x_2=z} \binom{n}{x_1\ x_2\ x_3} 2^{x_2} \quad p^{z} (1-p)^{2n-z}$$

$$= \binom{2n}{z} p^{z} (1-p)^{2n-z}$$, dvs. $2X_1 + X_2 \sim b(2n, p)$ idet

$$\sum_{2x_1+x_2=z} \binom{n}{x_1\ x_2\ x_3} = \binom{2n}{z} \qquad \left((p+1)^{2n} = (p^2+2p+1)^n \right)$$

$\square$

opgave 6.9

Grundmodellen er:

$M_0 : (X_{i1}, X_{i2}, X_{i3})^* \sim m(n_i, (\pi_{i1}, \pi_{i2}, \pi_{i3})^*)$

(i=1 mavesår , i=2 maverhvæft , i=3 raske , i1 = type O i2 = type A i3 = type B)

1)

$H_0 : \underline{\pi}_2 = \underline{\pi}_3$

$\underline{x}$	O	A	B	Σ
i=2	383	416	84	883
i=3	2892	2625	570	6087
Σ	3275	3041	654	6970

Jf. (6.46) fås:

$\underline{e}$	O	A	B	Σ
i=2	414,90	385,25	82,85	883
i=3	2860,10	2655,75	571,15	6087
Σ	3275	3041	654	6970

Jf. (6.47) fås $-2\ln Q(\underline{x}) = 5,64$

$\underset{(6.48)}{\varepsilon(\underline{x})} = 1 - F_{\chi^2((3-1)(3-1))}(-2\ln Q(\underline{x})) \geq 1-0,7=0,3 > 0,05$

Hypotesen forkastes ikke og vi reducerer til

$$M_1 : (Y_{i1}, Y_{i2}, Y_{i3})^* \sim m(n_{i1}(\tilde{\pi}_{i1}, \tilde{\pi}_{i2}, \tilde{\pi}_{i3})^*)$$

($i=1$ mave sår, $i=2$ mavekræft, raske)

2)

$$H_0 : \underline{\tilde{\pi}}_1 = \underline{\tilde{\pi}}_2$$

En test helt efter samme metode som i 1) udføres, jvf. udskriften: $-2\ln Q(\underline{x}) = 34{,}9192$

$\varepsilon(x) < 0{,}0001 < 0{,}05$

Dvs. hypotesen forkastes.

```
OPTIONS NODATE PAGENO=1 PS=65;

TITLE1 'Opgave 6.9';

DATA ab0;
INPUT sygdom$ blodtype$ antal;
gruppe=2;
IF sygdom='mave_s' THEN gruppe=1;
CARDS;
mave_s 0  983
mave_s A  679
mave_s B  134
mave_k 0  383
mave_k A  416
mave_k B   84
rask   0 2892
rask   A 2625
rask   B  570
;
RUN;

DATA ab0_23;
SET ab0;
IF gruppe=2;
RUN;

TITLE2 'spoergsmaal 1';
PROC FREQ DATA=ab0_23 ORDER=DATA;
WEIGHT antal;
TABLES sygdom*blodtype/EXPECTED CHISQ NOROW NOCOL NOPERCENT;
RUN;

TITLE2 'spoergsmaal 2';
PROC FREQ DATA=ab0 ORDER=DATA;
WEIGHT antal;
TABLES gruppe*blodtype/EXPECTED CHISQ NOROW NOCOL NOPERCENT;
RUN;
```

Opgave 6.9
spoergsmaal 1

The FREQ Procedure

Table of sygdom by blodtype

sygdom	blodtype			
Frequency Expected	0	A	B	Total
mave_k	383 414.9	416 385.25	84 82.853	883
rask	2892 2860.1	2625 2655.7	570 571.15	6087
Total	3275	3041	654	6970

Statistics for Table of sygdom by blodtype

Statistic	DF	Value	Prob
Chi-Square	2	5.6361	0.0597
Likelihood Ratio Chi-Square	2	5.6387	0.0596
Mantel-Haenszel Chi-Square	1	3.3524	0.0671
Phi Coefficient		0.0284	
Contingency Coefficient		0.0284	
Cramer's V		0.0284	

Sample Size = 6970

Opgave 6.9
spoergsmaal 2

The FREQ Procedure

Table of gruppe by blodtype

gruppe blodtype

Frequency Expected	O	A	B	Total
1	983 872.39	679 762.16	134 161.45	1796
2	3275 3385.6	3041 2957.8	654 626.55	6970
Total	4258	3720	788	8766

Statistics for Table of gruppe by blodtype

Statistic	DF	Value	Prob
Chi-Square	2	34.9192	<.0001
Likelihood Ratio Chi-Square	2	35.0014	<.0001
Mantel-Haenszel Chi-Square	1	31.8555	<.0001
Phi Coefficient		0.0631	
Contingency Coefficient		0.0630	
Cramer's V		0.0631	

Sample Size = 8766

```
OPTIONS NODATE PAGENO=1 PS=65;

%INCLUDE 'H:\Documents and Settings\Ejer\Skrivebord\EKSAMENSOPGAVER\sasmacros\hw.mac';

TITLE1 'Opgave 6.10';
TITLE2 'Lolland';
%hw(27,30,12);
RUN;

TITLE1 'Opgave 6.10';
TITLE2 'Bornholm';
%hw(14,20,52);
RUN;

TITLE1 'Opgave 6.10';
TITLE2 'Aaland';
%hw(0,5,75);
RUN;


DATA kendt;
p   =0.609**2;
p2l=2*0.609*(1-0.609);
p3l=(1-0.609)**2;
p1a=0.0313**2;
p2a=2*0.0313*(1-0.0313);
p3a=(1-0.0313)**2;
RUN;

TITLE2 'kendte frekvenser';
PROC PRINT DATA=kendt NOOBS;
RUN;

/* Da de samme beregninger skal laves 3 gange laves en macro*/
/* som benytter maksimaliserings proceduren i NLPNRR i      */
/* PROC IML                                                  */
%macro regn;

DATA regn1;
MERGE kendt obs;
RUN;

PROC IML;
/*indlaesning af data i PROC IML*/
use regn1 VAR _ALL_;
read VAR _ALL_ into aaa;
pl=j(3,1,0); pl[1]=aaa[1]; pl[2]=aaa[2]; pl[3]=aaa[3];
pa=j(3,1,0); pa[1]=aaa[4]; pa[2]=aaa[5]; pa[3]=aaa[6];
x=j(3,1,0); x[1]=aaa[7]; x[2]=aaa[8]; x[3]=aaa[9];
/* options for maksimalisering */
opt=j(1,5,0); opt[1]=1;
bc=j(2,1,0); bc[1]=0; bc[2]=1;
alfa0=0.5;


START l(alfa) global(x,pl,pa);
l=x[1]*log((1-alfa)*pl[1]+alfa*pa[1])
 +x[2]*log((1-alfa)*pl[2]+alfa*pa[2])
```

```
 +x[3]*log((1-alfa)*pl[3]+alfa*pa[3]);
return(l);
FINISH l;

CALL NLPNRR(rc,alfar,"l",alfa0,opt,bc);
alfa=alfar;
print alfa;


n=sum(x);
e=j(3,1,0);
e[1]=n*((1-alfa)*pl[1]+alfa*pa[1]);
e[2]=n*((1-alfa)*pl[2]+alfa*pa[2]);
e[3]=n*((1-alfa)*pl[3]+alfa*pa[3]);
x_e=x-e;
PRINT x e [format=8.4] x_e [format=8.4];

_2lnQ=0;
DO i=1 TO 3;
   _2lnQ=_2lnQ+2*x[i]*log(x[i]/e[i]);
END;
testss=1-PROBCHI(_2lnQ,1);
print _2lnQ [format=8.4] testss [format=8.4];
QUIT;

%mend regn;

DATA obs;
x1=14;
x2=20;
x3=52;
RUN;

TITLE2 'Bornholm lokalitet 1';
%regn;

DATA obs;
x1=12;
x2=12;
x3=54;
RUN;

TITLE2 'Bornholm lokalitet 2';
%regn;

DATA obs;
x1=2;
x2=4;
x3=66;
RUN;

TITLE2 'Bornholm lokalitet 3';
%regn;
```

Opgave 6.10
Lolland

Hardy-Weinberg 1

genotype	observeret	forventet	observeret-forventet
AA	27	25.5652	1.43478
Aa	30	32.8696	-2.86957
aa	12	10.5652	1.43478

Opgave 6.10
Lolland

Hardy-Weinberg 2

n	p	antal A gener	-2lnQ	testss. -2lnQ	chi-i-anden	testss. chi-i-anden
69	0.6087	84	0.5238	0.4692	0.5259	0.4683

Opgave 6.10
Bornholm

Hardy-Weinberg 1

genotype	observeret	forventet	observeret-forventet
AA	14	6.6977	7.3023
Aa	20	34.6047	-14.6047
aa	52	44.6977	7.3023

Hardy-Weinberg 2

n	p	antal A gener	-2lnQ	testss. -2lnQ	chi-i-anden	testss. chi-i-anden
86	0.2791	48	14.452	0.0001	15.318	0.0001

Opgave 6.10
Aaland

Hardy-Weinberg 1

genotype	observeret	forventet	observeret-forventet
AA	0	0.0781	-0.07813
Aa	5	4.8438	0.15625
aa	75	75.0781	-0.07813

Opgave 6.10
Aaland

Hardy-Weinberg 2

n	p	antal A gener	-2lnQ	testss. -2lnQ	chi-i-anden	testss. chi-i-anden
80	0.0313	5	.	.	0.0832	0.7729

Opgave 6.10
kendte frekvenser

p1l	p2l	p3l	p1a	p2a	p3a
0.37088	0.47624	0.15288	.00097969	0.060641	0.93838

Opgave 6.10
Bornholm lokalitet 1

ALFA

0.5740962

X	E	X_E
14	13.6329	0.3671
20	20.4375	-0.4375
52	51.9296	0.0704

_2LNQ	TESTSS
0.0193	0.8894

Opgave 6.10
Bornholm lokalitet 2

ALFA

0.6749509

X	E	X_E
12	9.4548	2.5452
12	15.2670	-3.2670
54	53.2782	0.7218

_2LNQ	TESTSS
1.3956	0.2375

Opgave 6.10
Bornholm lokalitet 3

ALFA

0.9491384

X	E	X_E
2	1.4251	0.5749
4	5.8881	-1.8881
66	64.6868	1.3132

_2LNQ	TESTSS
0.9153	0.3387

STAT1, aflevringsopgave uge 49, Gunna Hellmund

1)

$M_0: \; x_{ij} \sim\sim m(n_i, (\pi_{i1}, \pi_{i2}, \pi_{i3})^*) \quad i=1,2, \; j=1,2,3$

Vi har vist det ønskede hvis flg. hypotese kan accepteres:

$H_0: \; \pi_{ij} = \rho_i \sigma_j$

Under hypotesen gælder:

$\underline{e}$	v	s	h
ja	11,38	32,93	52,69
nej	7,62	22,07	35,31

(jf. (6.46))

$-2\ln Q(\underline{x}) = 4{,}0506$ (jf. (6.47))

$\varepsilon(\underline{x}) = 1 - F_{\chi^2(4)}(-2\ln Q(\underline{x})) \geq 1 - 0{,}7 = 0{,}3 > 0{,}05$ (jf. (6.48))

Vi forkaster altså ikke H_0.

2)

Jf. (6.40) er $x_{i\cdot} \sim b(162, p)$

Det ønskede konf.int. er derfor givet udfra formlerne i (6.27) – (6.28) og er:

$[0{,}5702; 0{,}6711]$

3)

Lad $y_{1\cdot}$ betegne det antal kvinder, der ikke selv har oplevet afstraffelse, men som selv anvender den. Da gælder (jf. (6.40)) at $y_{1\cdot} \sim b(75, q)$

Hvis der <u>ikke</u> er en sammenhæng mellem hvorvidt man afstraffer og hvorvidt man selv er blevet afstraffet accepterer vi

$H_0: p = q$

	ja	nej	Σ
1)	97	65	162
2)	23	52	75
Σ	120	117	237

1) afstraffede 2) ikke-afstraffede

e	ja	nej
1)	82,03	79,97
2)	37,97	37,03

(jf. (6.46))

$$-2 \ln Q(\underline{x}) = 17{,}82 \qquad \text{(jf. (6.47))}$$

$$\varepsilon(\underline{x}) = 1 - F_{\chi^2(1)}(-2 \ln Q(\underline{x})) \leq 1 - 0{,}995 < 0{,}05$$

(jf. (6.48))

Hypotesen forkastes, altså peger materialet på en sammenhæng mellem hvordan man afstraffer og selv er blevet afstraffet.

Afdeling for Teoretisk Statistik
Institut for Matematiske Fag
Aarhus Universitet

Statistik 1 og Statistik α
Preben Blæsild
Jørgen Granfeldt
15. november 2002

Meddelelse 11

Forelæsningerne i sidste uge

(Onsdag den 6. november) Afsnit 4.4.1 *Fordelingsresultater i den lineære normale model.*

(Fredag den 8. november) Afslutning af *Tosidet variansanalyse* og derefter Kapitel 6 *Multinomialfordelte data.* Jeg nåede til øverst side 6.7.

Forelæsningerne i denne uge

(Onsdag den 13. november) Kapitel 6 *Multinomialfordelte data.* Hypoteser, estimation under hypotese og test af hypotese som på side 6.7 til 6.10. Desuden Eksempel 6.1 side 6.18 og 6.19.

(Fredag den 15. november) Kapitel 6 *Multinomialfordelte data.* Konfidensintervaller, test for uafhængighed.

Forelæsningerne i næste uge

(Onsdag den 20. og fredag den 22. november) Kapitel 6 og derefter kapitel 7 *Poissonfordelte data*

Øvelserne i denne uge

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

Øvelserne i de tre næste uger

(Uge 47: 18. november - 24. november) 4.11, 6.1, 6.4, 6.9, 6.10.

(Uge 48: 25. november - 1. december) 6.11, 7.6, 7.8, opgave 1 og 2 denne meddelelse, samt eventuelle manglende opgaver fra tidligere øvelsesgange.

(Uge 49: 2. december - 8. december) (Foreløbigt program) 7.14, 7.16, 7.17 samt opgaver som stilles på Meddelelse 12.

Afleveringsopgaver

(Uge 47) Ingen afleveringsopgave.

(Uge 48) Ingen afleveringsopgave.

(Uge 49) Eksamensopgave 2 i Statistik α Sommeren 2002. Opgavesættet for sommeren 20002 findes på adressen

`http://home.imf.au.dk/statbib/atskurser/statalfa/eksamen/2002/`

Webcasting af forelæsningerne i Statistik 1/α

Meddelelse fra Gunnar Hellmund:

Ifølge aftale med dekanen er det meningen, at forelæsningerne i Statistik 1/α skal webcastes. Det varer et stykke tid før forelæsningerne bliver tilgængelige via internettet, derfor bliver optagelserne af forelæsningerne midlertidig tilgængelige påCDere. (Et tilsvarende tilbud gives i Sandsynlighedsteori 2.) Hver CD koster 20 kroner og indeholder 2 ugers forelæsninger i RealMedia8 formatet, som kan læses på (nyere) Windows, Linux og Apple computere ved hjælp af gratis software. Inddelingen af efterårssemestret er:

CD1 uge 36-37

CD2 uge 38-39

CD3 uge 40-41

CD4 uge 43-44

CD5 uge 45-46

osv.

Optagelserne er komprimeret væsentligt (fra 30 til 10 frames/sec.) og lydoptagelsen er svag, så man skal skrue helt op for volumen i softwaren og på hardwaren for tydeligt at høre, hvad der bliver sagt. Med andre ord kan optagelserne ikke erstatte den oplevelse, det er selv at møde op til forelæsningerne. Har man været syg, eller ønsker man at repetere det gennemgåede stof, er optagelserne gode. Bestil CDere ved at e-maile til `hellmund@imf.au.dk` – angiv navn, fag (Stat1 eller SS2) og numrene på de ønskede CDere. Bestilte CDere kan afhentes til næstkommende onsdagsforelæsning i Stat1 (dog tidligst 2 dage efter bestilling). Evt. spørgsmål stilles via email til `hellmund@imf.au.dk`.

Første obligatoriske opgave

Instruktorerne fortæller at første obligatoriske opgave forlængst er rettet, men at en del opgaver ikke er afhentet. Det drejer sig i mange tilfælde om opgaver, som ikke kan godkendes. Kontakt instruktorerne for at aftale en afleveringsforretning, hvor I aftaler, hvad der skal laves ekstra for at opgaven kan godkendes.

Instruktorernes e-mail adresser er

Morten: `fenger@imf.au.dk`

Jacob: `jj@imf.au.dk`

Berit: `lindum@imf.au.dk`

Jørgen: `jvhansen@imf.au.dk`

Som bekendt er afleveringsfristen for 2. opgave på fredag. Hvis første opgave ikke er godkendt, og man ikke har ulejliget sig med at indhente respons på den første opgave, og ikke har gået til konsultation, så er er anden obligatoriske opgave nok heller ikke til umiddelbar godkendelse.

Det ser ud som om at nogen har overset at det obligatoriske ved opgaverne både er at aflevere dem og at få dem godkendt. Hvis ikke opgaverne er godkendt samlet, kan man ikke gå til eksamen.

Man kan ikke tillade sig at disponere efter at instruktorerne har ubegræset tid til at hjæpe med gamle obligatoriske opgaver i december.

Opgaver til Øvelserne i uge 48

Opgave 1

I tabellen nedenfor er de 953 og 740 tvillingefødsler i Danmark i årene 1956 og 1969 fordelt efter børnenes køn (*DD* 2 drenge, *DP* 1 dreng og 1 pige, *PP* 2 piger).

år/køn	*DD*	*DP*	*PP*	*i alt*
1956	322	350	281	953
1969	272	246	222	740
i alt	594	596	503	1693

1) Vis, at det kan antages, at fordelingen på køn er identisk for de to år.

Hvis $r \in]0,1[$ betegner sandsynligheden for, at et tvillingepar er enægget og $p \in]0,1[$ sandsynligheden for, at et nyfødt barn er en dreng, er sandsynlighedsvektoren $\pi = (\pi_1, \pi_2, \pi_3)$ for (*DD, DP, PP*) givet ved afbildningen

$$\begin{array}{rcl} \pi: \ ]0,1[^2 & \to & \Pi = \{(\pi_1,\pi_2,\pi_3) \in R^3_+ : \pi_1+\pi_2+\pi_3 = 1\} \\ (p,r) & \to & \pi(p,r), \end{array} \tag{1}$$

hvor

$$\begin{aligned} \pi_1(p,r) &= rp + (1-r)p^2, \\ \pi_2(p,r) &= 2(1-r)p(1-p), \\ \pi_3(p,r) &= r(1-p) + (1-r)(1-p)^2. \end{aligned}$$

Dette skal **ikke** vises.

2) Vis, at den inverse afbildning til (1) er givet ved

$$\begin{aligned} p(\pi) &= \frac{2\pi_1+\pi_2}{2}, \\ r(\pi) &= 1 - \frac{2\pi_2}{(2\pi_1+\pi_2)(2\pi_3+\pi_2)}. \end{aligned}$$

3) Estimer p og r på grundlag af (x_1, x_2, x_3), den observerede fælles fordeling af kønnene i de to betragtede år.

4) Vis, at under hypotesen $H_0 : p = 1/2$ er

$$\hat{r} = \frac{x_1 + x_3 - x_2}{n}$$

maksimum likelihood estimatet for r, og test derefter hypotesen.

Opgave 2

En yndet opgave i sandsynlighedsregning er at beregne sandsynligheden for at der i en forsamling på for eksempel 30 personer er mindst to, der har fødselsdag på samme dag. Den forudsætning, der mere eller mindre stiltiende gøres er, at der bliver født lige mange børn hver dag.

1) I tabellen nedenfor er fødslerne i 1988 fordelt efter fødselsmåned. Undersøg, om disse tal stemmer med en hypotese om, at der kan forventes lige mange fødsler hver dag året igennem.

Måned	*Levendefødte*	*Dødfødte*
Januar	4500	31
Februar	4471	27
Marts	5201	19
April	5061	18
Maj	5169	24
Juni	4989	23
Juli	5160	27
August	5288	24
September	5158	31
Oktober	4716	22
November	4474	23
December	4657	23
Ialt	58844	292

2) Overvej, om sandsynligheden for sammenfald af fødselsdage bliver større eller mindre i et inhomogent materiale.

STAT 1, ~~TO~~ nr. 10

opgave 4.6

1)

$M_0: \; X_{ijk} \sim\sim N(\mu_{ij}, \sigma_{ij}^2)$ (i = fosfor niv., j = nitrat niv.)

$H_{0\sigma^2}: \; \sigma_{ij}^2 = \sigma^2 \; \forall i,j.$

Se s. 3.67-68 for beregningsformler:

$C = 1,0\overline{77}$

$-2 \ln Q(\underline{x}) = 9,532908$

$Ba = 8,844967$

$\varepsilon(x) = 1 - F_{\chi^2(6-1)}(Ba) \geq 1 - 0,9 = 0,1 > 0,05$

Hypotesen forkastes ikke og vi reducerer til

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

2)

$H_{0_1}: \; \mu_{ij} = \alpha_i + \beta_j$

$$F = \frac{\frac{112693,5556 - 111816,8333}{32-30}}{3727,2278} = 0,117611$$

$\varepsilon(x) = 1 - F_{F(32-30,30)}(F) \geq 1 - 0,5 = 0,5 > 0,05$

Vi reducerer dermed til

$M_2: \; x_{ijk} \sim\sim N(\alpha_i + \beta_j, \sigma^2)$

3)

$H_\beta: \; x_{ijk} \sim\sim N(\alpha_i, \sigma^2)$

$$F = \frac{\frac{180727,5833 - 112693,5556}{33 - 32}}{3521,6736} = 19,31866$$

$\varepsilon(x) = 1 - F_{F(33-32,32)}(F) \leq 1 - 0,9995 < 0,05$

Hypotesen forkastes og modellen er stadig M_2

4)

$H_{0\alpha}: \; x_{ijk} \sim\sim N(\beta_j, \sigma^2)$

$$F = \frac{\frac{270334,7222 - 112693,5556}{34 - 32}}{3521,6736} = 22,381570$$

$\varepsilon(x) = 1 - F_{F(34-32,32)}(F) \leq 1 - 0,9995 < 0,05$

$H_{0\alpha}$ forkastes, modellen er M_2

5)

95% konfidensint. for $\beta_1 - \beta_2$:

$t_{0,975}(32) \simeq 2,04$

Endepkt. er $\hat{\beta}_1 - \hat{\beta}_2 \pm$ Std. av. $\cdot t_{0,975}(32)$

$= -86,9444444 \pm 19,78123019 \cdot 2,04$

Konf. int.: $[-127,2982 ; -46,5907]$

opgave 4.8

2)

Jf udskriften ser vi at en rimelig statmodel er

M_0: $x_{ijk} \sim\sim N(\mu_{ij}, \sigma_{ij}^2)$ $i=1,2$ $j=1,2$ $k=1,\dots,10$

(i=niveau af A, j=niveau af B)

$H_{0\sigma^2}$: $\sigma_{ij}^2 = \sigma^2$; $i,j=1,2$

$\varepsilon(x) = 0{,}74534 > 0{,}05$ (se udskriften fra Bartlett max.)

Dvs. vi forkaster ikke hypotesen og reducerer til

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

3)

H_{02}: $x_{ijk} \sim\sim N(\alpha_i + \beta_j, \sigma^2)$

$$F = \frac{\frac{2{,}75168656 - 2{,}74254471}{37-36}}{0{,}07618180} = 0{,}1200$$

$$\varepsilon(\underline{x}) = 1 - F_{F(37-36,36)}(F) \geq 1 - 0{,}5 = 0{,}5 > 0{,}05$$

Dvs. vi kan reducere til

M_2: $x_{ijk} \sim\sim N(\alpha_i + \beta_j, \sigma^2)$

$H_{0\beta}$: $x_{ij} \sim\sim N(\alpha_i, \sigma^2)$

$$F = \frac{\frac{5,86707810 - 2,75168656}{38-37}}{0,07436991} = 41,8905$$

$$\varepsilon(x) = 1 - F_{F(38-37,37)}(F) \leq 1 - 0,9995 < 0,05$$

Dus. $H_{0\beta}$ fouhastes

$H_{0\alpha}$: $x_{ij} \sim\sim N(\beta_j, \sigma^2)$

$$F = \frac{\frac{4,73538884 - 2,75168656}{38-37}}{0,07436991} = 26,6734$$

$$\varepsilon(\underline{x}) = 1 - F_{F(38-37,37)}(F) \leq 1 - 0,995 < 0,05$$

$H_{0\alpha}$ fouhastes.

4)

H_{0z}: $\beta_2 - \beta_1 = \alpha_2 - \alpha_1$

$$F = \frac{\frac{2,81527108 - 2,75168656}{38-37}}{0,07436991} = 0,854976$$

$$\varepsilon(x) = 1 - F_{F(38-37,37)}(F) \geq 1 - 0,7 = 0,3 > 0,05$$

Altså reducerer vi til modellen

$M_3: \quad x_{ij} \sim\sim N(\alpha + \beta \cdot t_{ij}, \sigma^2)$

5)

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

$H_0: \quad \mu_{12} - \mu_{11} = \mu_{21} - \mu_{11} \iff \mu_{12} = \mu_{21}$

$$F = \frac{\frac{2{,}80612922 - 2{,}74254471}{37-36}}{0{,}07618180} = 0{,}834642$$

$\varepsilon(x) = 1 - F_{F(37-36,36)}(F) \geq 1 - 0{,}7 = 0{,}3 > 0{,}05$

Dvs. vi kan ikke forkaste hypotesen og reducerer til

$M_2^*: \quad x_{ij} \sim\sim N(\mu_{ij}, \sigma^2), \ \mu_{12} = \mu_{21}$

6)

$H_0: \quad \mu_{22} = \mu_{21} + \mu_{12}$

$$F = \frac{\frac{2{,}81527108 - 2{,}80612922}{38-37}}{0{,}07584133} = 0{,}120539$$

$\varepsilon(x) = 1 - F_{F(38-37,37)}(F) \geq 1 - 0{,}5 = 0{,}5 > 0{,}05$

Vi reducerer til $M_3: \ x_{ij} \sim\sim N(\alpha + \beta t_{ij}, \sigma^2)$

7)

$M_2, M_2^*, M_3 \subseteq M_1$

$M_3 \subseteq M_2$

$M_3 \subseteq M_2^*$

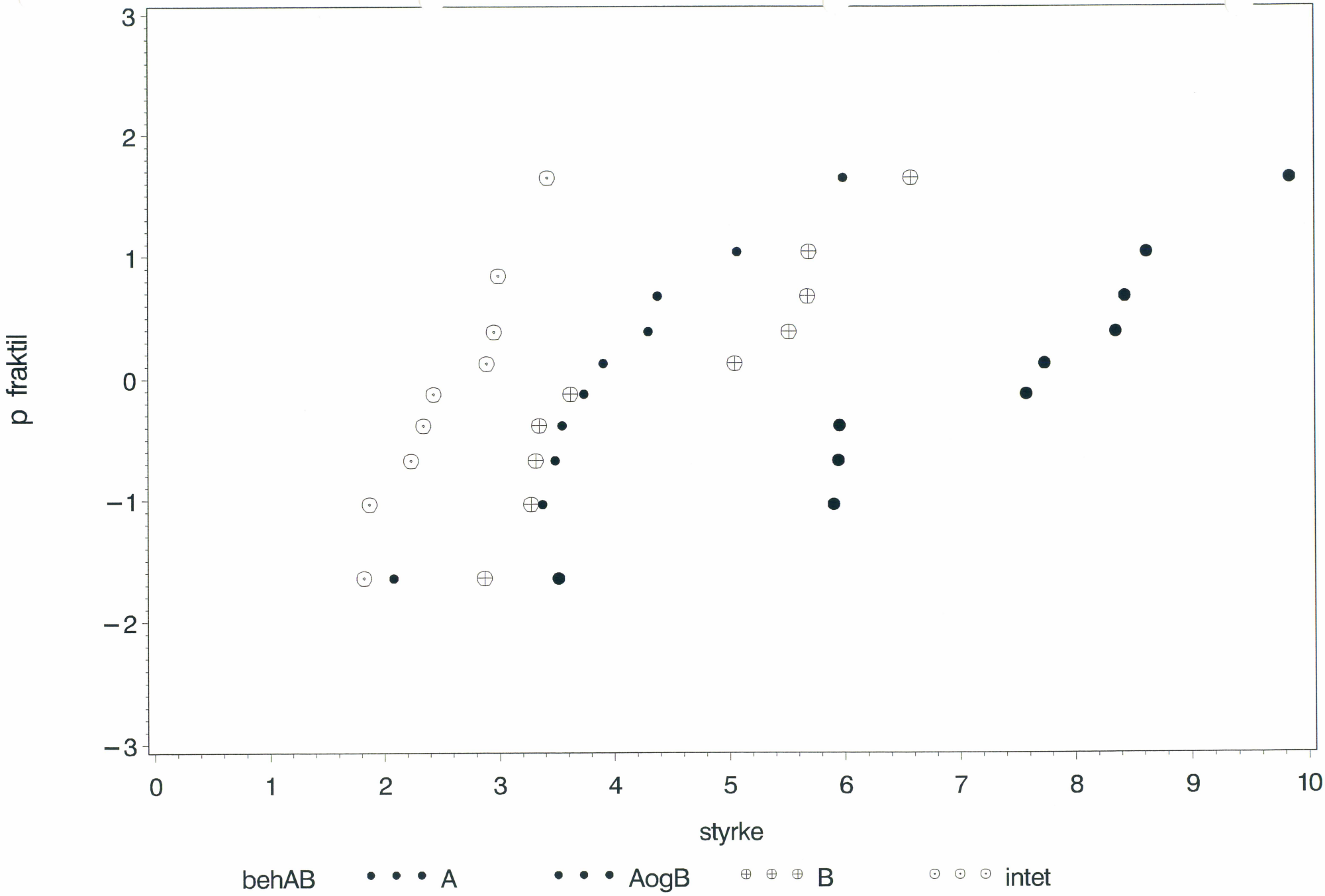
```
OPTIONS NODATE PAGENO=1 LS=78;

/* Nogle GOPTIONS */
/* FILENAME GSASFILE '~/atskurser/stat1/opgaver/kapitel_04/opgave_08.ps'; */
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 fliser;
INFILE 'H:\Documents and Settings\Ejer\Skrivebord\EKSAMENSOPGAVER\opgaver\kapitel_04\opgave_0
INPUT nummer A B styrke;
lnstyrke=log(styrke);
RUN;


DATA fliser;
SET fliser;
LENGTH behA $ 13;
LENGTH behB $ 13;
LENGTH behAB $ 8;
lnstyrke=log(styrke);
 gruppe = (A+1)/2 + B+1;
IF A=-1 THEN behA='A ikke tilsat';
ELSE IF A=1 THEN behA='A tilsat';
IF B=-1 THEN behB='B ikke tilsat';
ELSE IF B=1 THEN behB='B tilsat';
IF gruppe=0 THEN AB=0;
ELSE IF gruppe=1 OR gruppe=2 THEN AB=10;
ELSE IF gruppe=3 THEN AB=20;
IF gruppe=0 THEN behAB='intet';
ELSE IF gruppe=1 THEN behAB='A';
ELSE IF gruppe=2 THEN behAB='B';
ELSE IF gruppe=3 THEN behAB='AogB';
TILSAT=AB;
LENGTH bh $ 13;
/*--Dette har jeg tilføjet, Gunnar--*/
IF gruppe=0 THEN bh='Intet tilsat';
ELSE IF gruppe=1 or gruppe=2 THEN bh='10 kg tilsat';
ELSE IF gruppe=3 THEN bh='20 kg tilsat';
/*----*/
RUN;


PROC PRINT DATA=fliser;
VAR a b styrke;
RUN;


SYMBOL1  i=NONE v=CIRCLE H=2.7;

SYMBOL2  i=NONE v=dot    H=2.7;

SYMBOL3  i=NONE v=+      H=2.7;
```

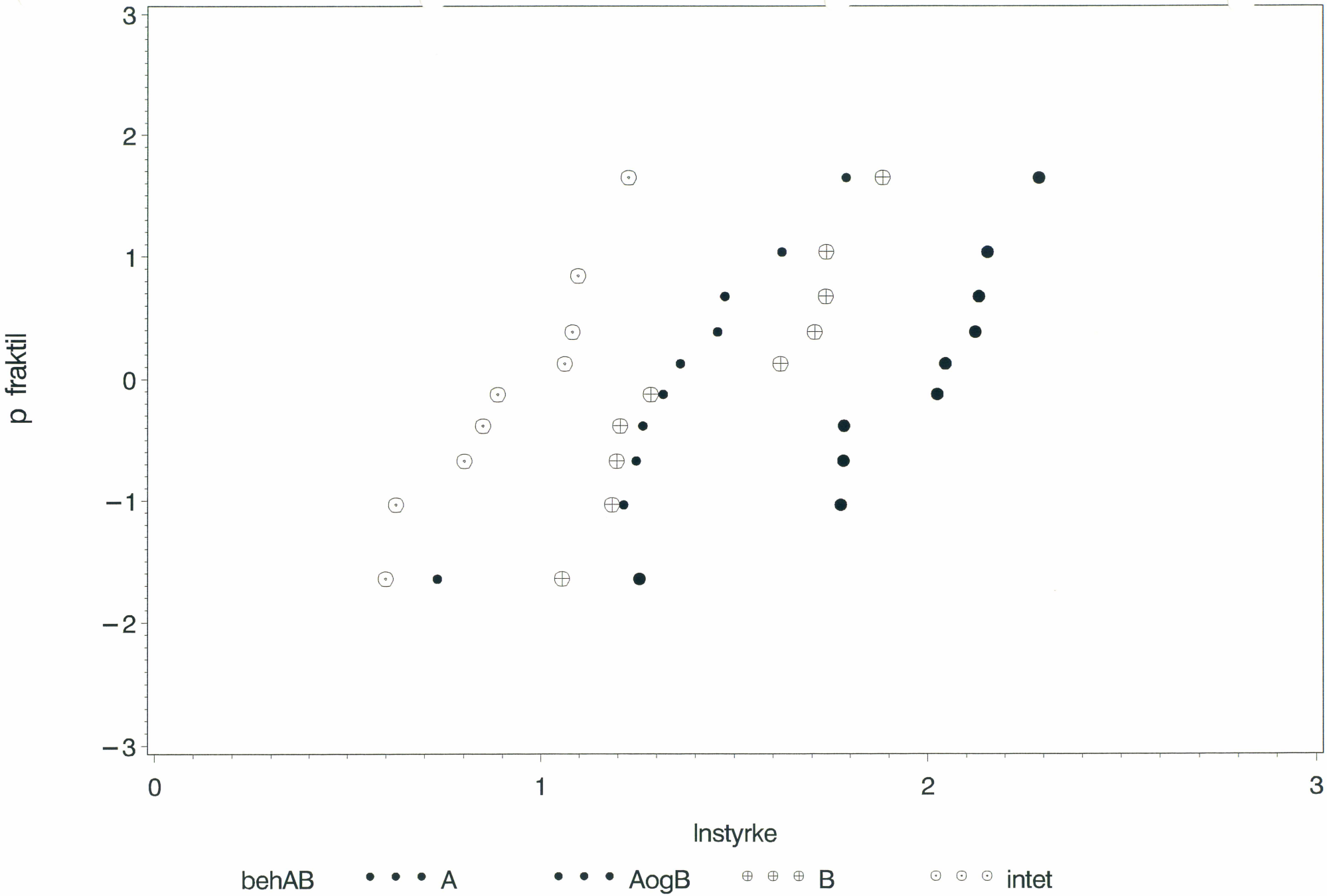
```
SYMBOL4  i=NONE v=-      H=2.7;

/*
AXIS1
   ORDER=0 to 3 by 1
   COLOR=black
   WIDTH=2.0
   STYLE=1
   ;

AXIS2
   COLOR=black
   ORDER=0 to 6 by 1
   WIDTH=2.0
   STYLE=1
   ;

AXIS2
   COLOR=black
   ORDER=0 to 6 by 1
   WIDTH=2.0
   STYLE=1
   ;
*/


TITLE1;
%INCLUDE 'H:\Documents and Settings\Ejer\Skrivebord\EKSAMENSOPGAVER\sasmacros\normgraf.mac';
%INCLUDE 'H:\Documents and Settings\Ejer\Skrivebord\EKSAMENSOPGAVER\sasmacros\bartlett.mac';
AXIS2
   COLOR=black
   ORDER=0 TO 10 BY 1
   WIDTH=2.0
   STYLE=1
   ;
%normgraf(data=fliser,var=styrke,bygrp=behAB,grafik=ja);
AXIS2
   COLOR=black
   ORDER=0 to 3 by 1
   WIDTH=2.0
   STYLE=1
   ;
%normgraf(data=fliser,var=lnstyrke,bygrp=behAB,grafik=ja);
RUN;
TITLE1 ' ';


AXIS1
   ORDER=0 to 3 by 1
   COLOR=black
   WIDTH=2.0
   STYLE=1
   ;


AXIS2
   COLOR=black
```

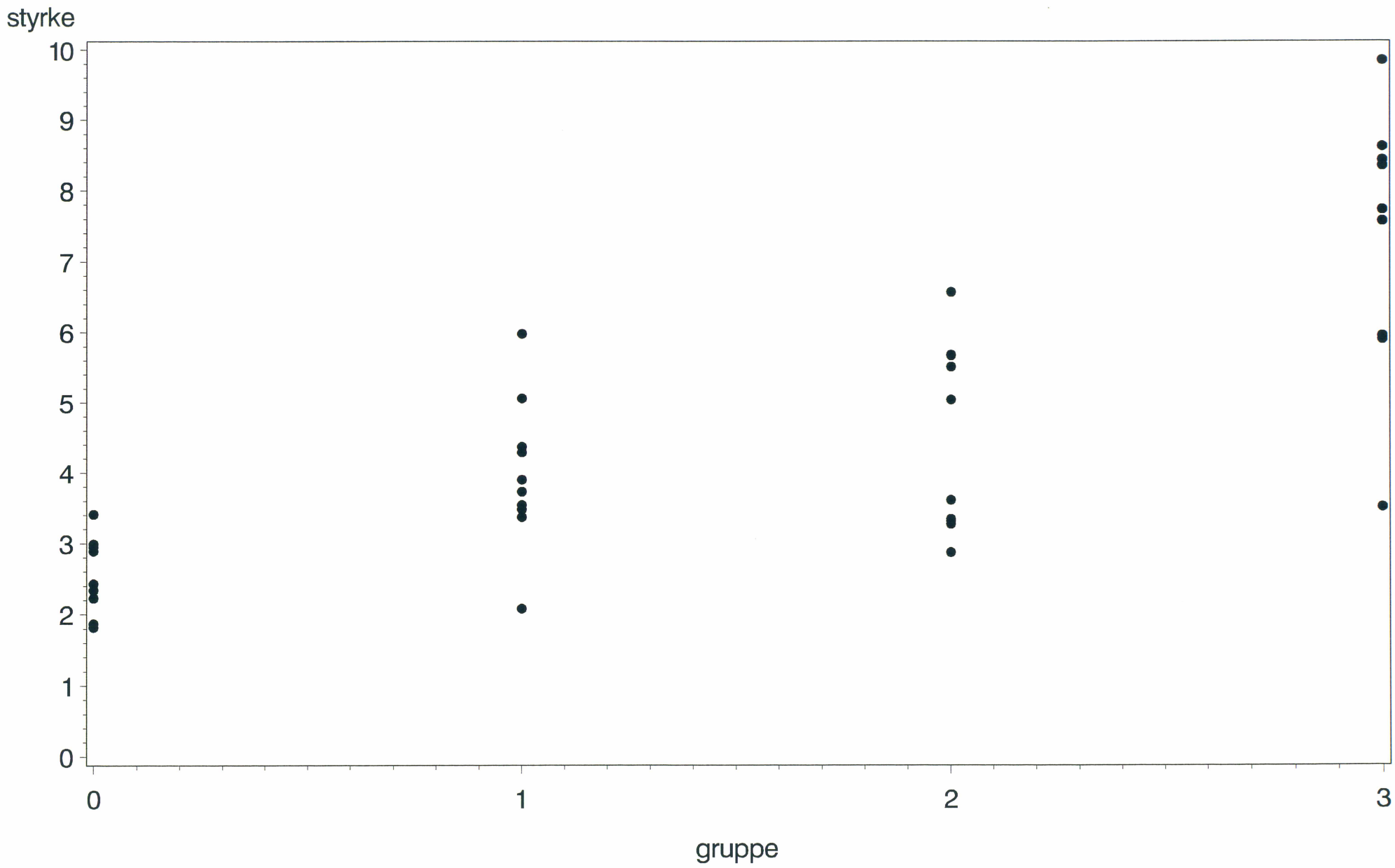
```
   ORDER=0 to 10 by 1
   WIDTH=2.0
   STYLE=1
   ;
PROC GPLOT DATA=fliser ;
   PLOT styrke * GRUPPE /
      HAXIS=axis1
      VAXIS=axis2
      ;
RUN; QUIT;

AXIS2
   COLOR=black
   ORDER=0 to 3 by 1
   WIDTH=2.0
   STYLE=1
   ;
PROC GPLOT DATA=fliser ;
   PLOT lnstyrke * GRUPPE  /
      HAXIS=axis1
      VAXIS=axis2
      ;
RUN; QUIT;



OPTIONS LS=82;
%bartlett(data=fliser,gruppe=gruppe,var=styrke);
%bartlett(data=fliser,gruppe=gruppe,var=lnstyrke);

/*-----------------------------------------------*/
/*  kontroltegninger for additivitet            */

SYMBOL1 C=black H=2 I=rl
   V=circle L=1
   ;

SYMBOL2 C=black  H=2 I=rl
   V=triangle L=2
   ;

PROC GPLOT DATA=fliser;
PLOT styrke*A=B;
RUN;

PROC GPLOT DATA=fliser;
PLOT styrke*B=A;
RUN;

PROC GPLOT DATA=fliser;
PLOT lnstyrke*A=B;
RUN;

PROC GPLOT DATA=fliser;
PLOT lnstyrke*B=A;
RUN;
```

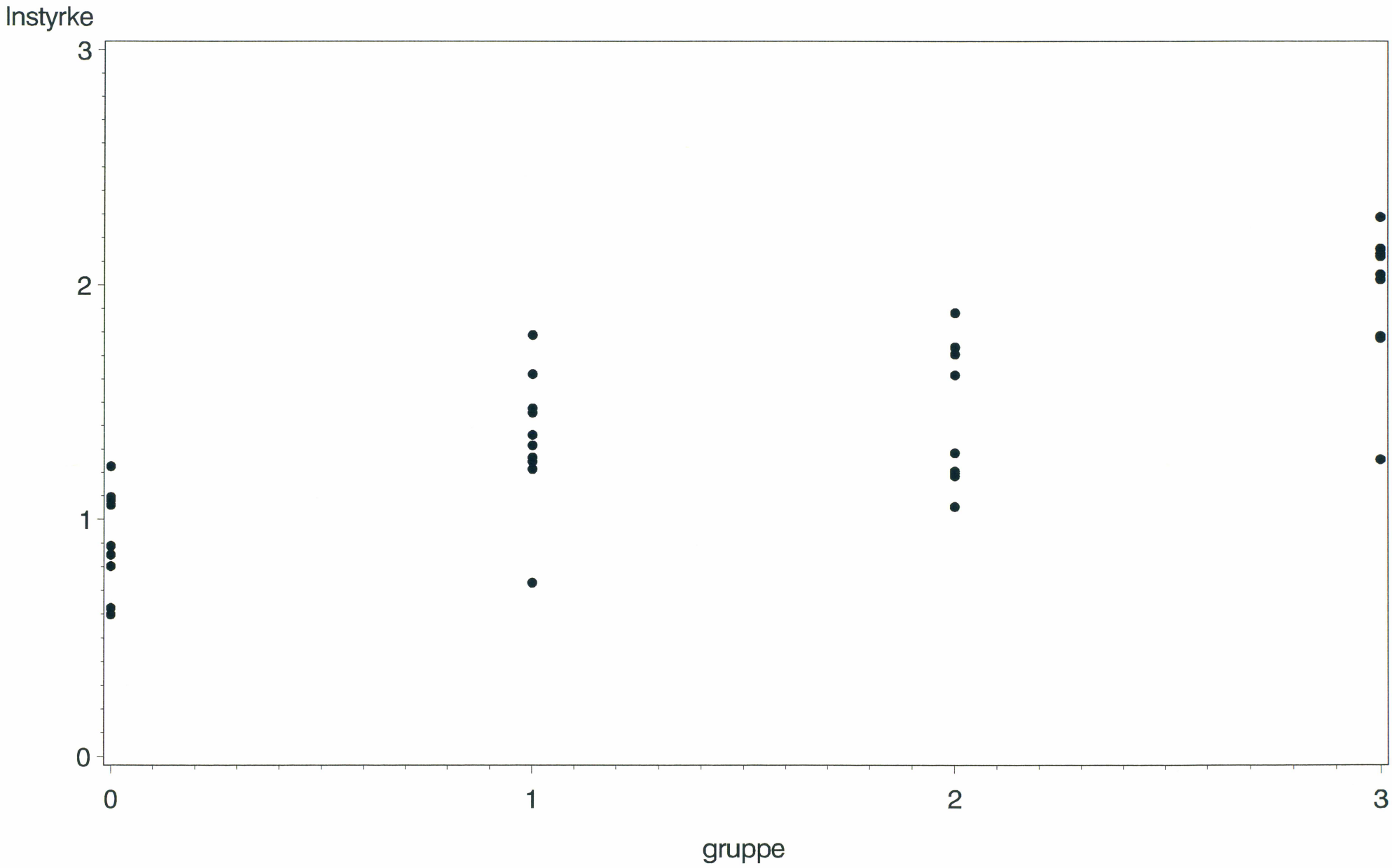
```
/*-----------------------------------------------*/
/*          Tosidet variansanalyse               */

TITLE 'M1';
PROC GLM DATA=fliser;
CLASS behA behB;
MODEL lnstyrke = behA behB behA*behB/ SS1 SOLUTION;
RUN;

TITLE 'M2';
PROC GLM DATA=fliser;
CLASS behA behB;
MODEL lnstyrke = behA behB/ SS1 SOLUTION;
RUN;

TITLE 'M3r';
PROC GLM DATA=fliser;
CLASS behA;
MODEL lnstyrke = behA/ SS1 SOLUTION;
RUN;

TITLE 'M3s';
PROC GLM DATA=fliser;
CLASS behB;
MODEL lnstyrke = behB/ SS1 SOLUTION;
RUN;
QUIT;

/*-----------------------------------------------*/
/*              Ensidet variansanalyse           */
/*          Ser paa ens virkning af A og B       */

TITLE 'Ens virkning af A og B';
PROC GLM DATA=fliser;
CLASS AB gruppe;
MODEL lnstyrke = tilsat AB gruppe/ SS1 SOLUTION;
RUN;

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


%bartlett(data=fliser,gruppe=tilsat,var=lnstyrke);

AXIS2
   COLOR=black
   ORDER=0 to 3 by 1
   WIDTH=3.0
   STYLE=1
   ;

AXIS4
   ORDER=0 to 20 by 5
   COLOR=black
   WIDTH=3
   label=(h=2 'kg tilsat')
   ;

SYMBOL1 c=black
```

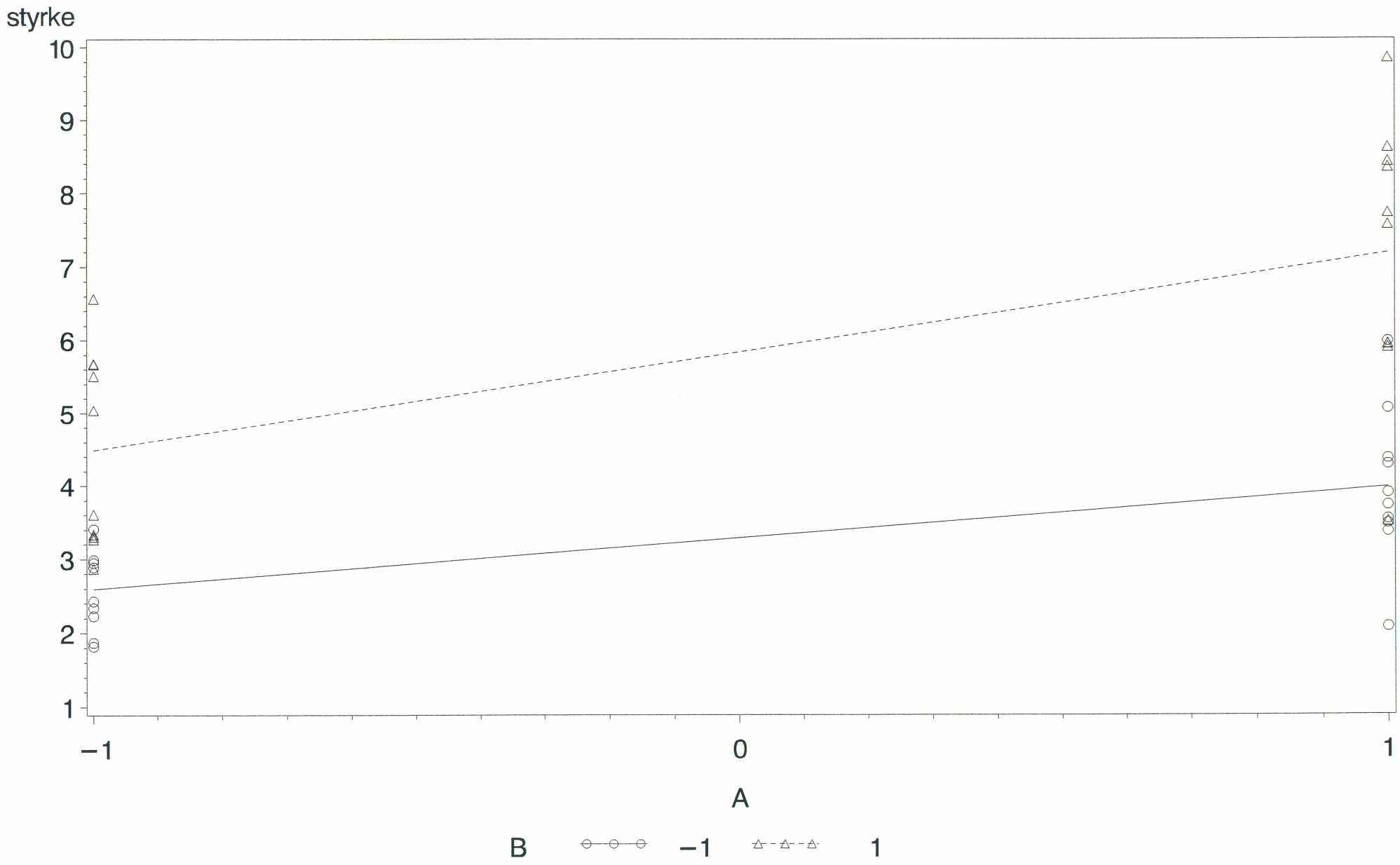
```
   i=rl
   v=CIRCLE
   ;

TITLE1 'Logtraekstyrke mod maengde tilsat stof';
PROC GPLOT DATA=fliser ;
   PLOT lnstyrke * tilsat /
      HAXIS=axis4
      VAXIS=axis2
      ;
RUN; QUIT;



TITLE1 'Regression';
PROC GLM DATA=fliser;
MODEL lnstyrke = tilsat;
RUN;  QUIT;


/*--standardberegninger---*/

DATA beregn;
SET fliser;
t=tilsat;
x=lnstyrke;
xt=t*x;
KEEP t x xt;
RUN;

PROC MEANS DATA=beregn SUM USS CSS VAR;
VAR x t xt;
RUN;


/*--Dette har jeg tilføjet, Gunnar--*/
TITLE '3 obs. rækker';
PROC GLM DATA=fliser;
CLASS bh;
MODEL lnstyrke = bh/ SS1 SOLUTION;
RUN;
QUIT;
/*----*/
```

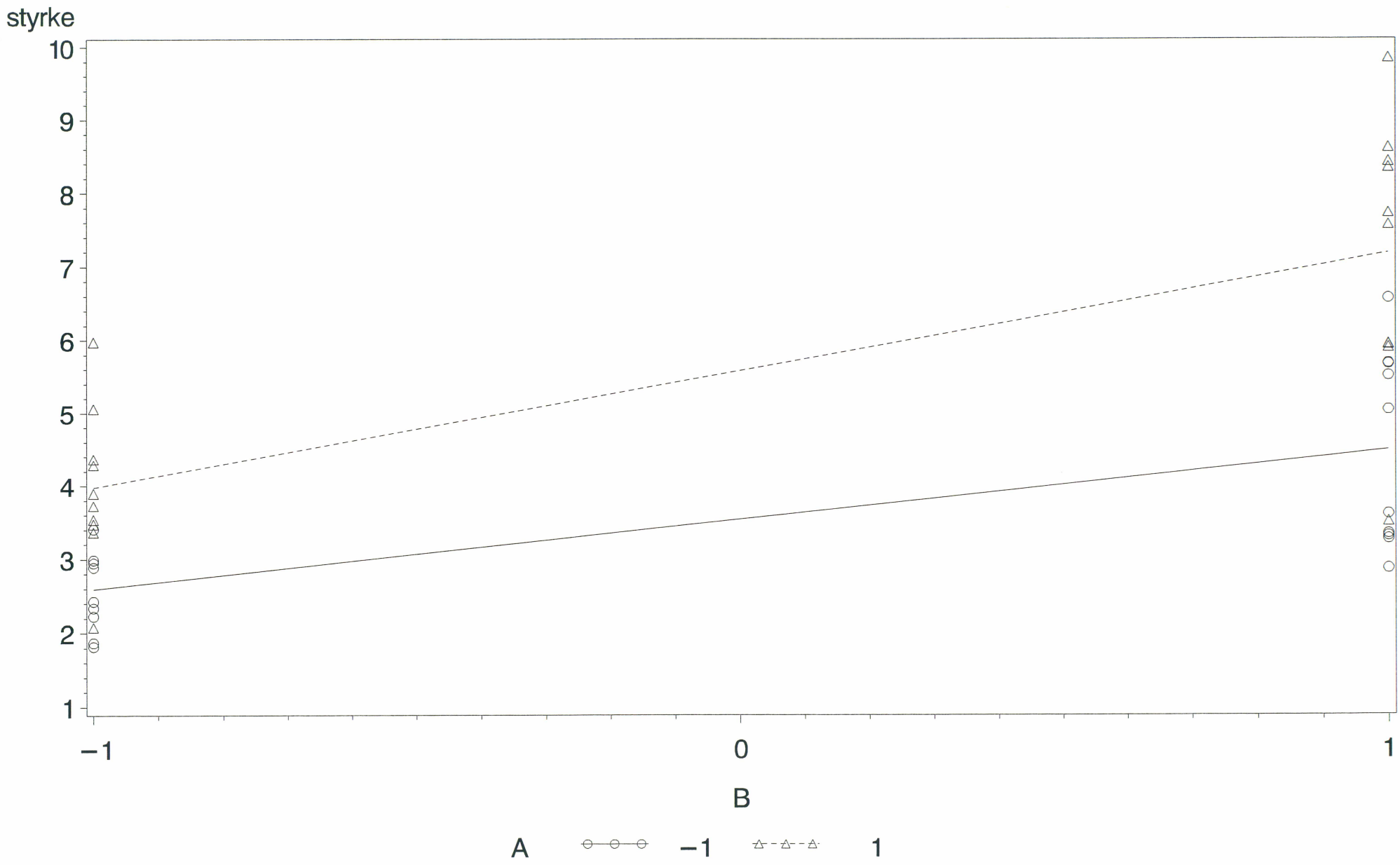

p fraktil
3
2
1
0
−1
−2
−3
0
1
2
3
4
5
6
7
8
9
10
styrke
behAB
A
AogB
B
intet

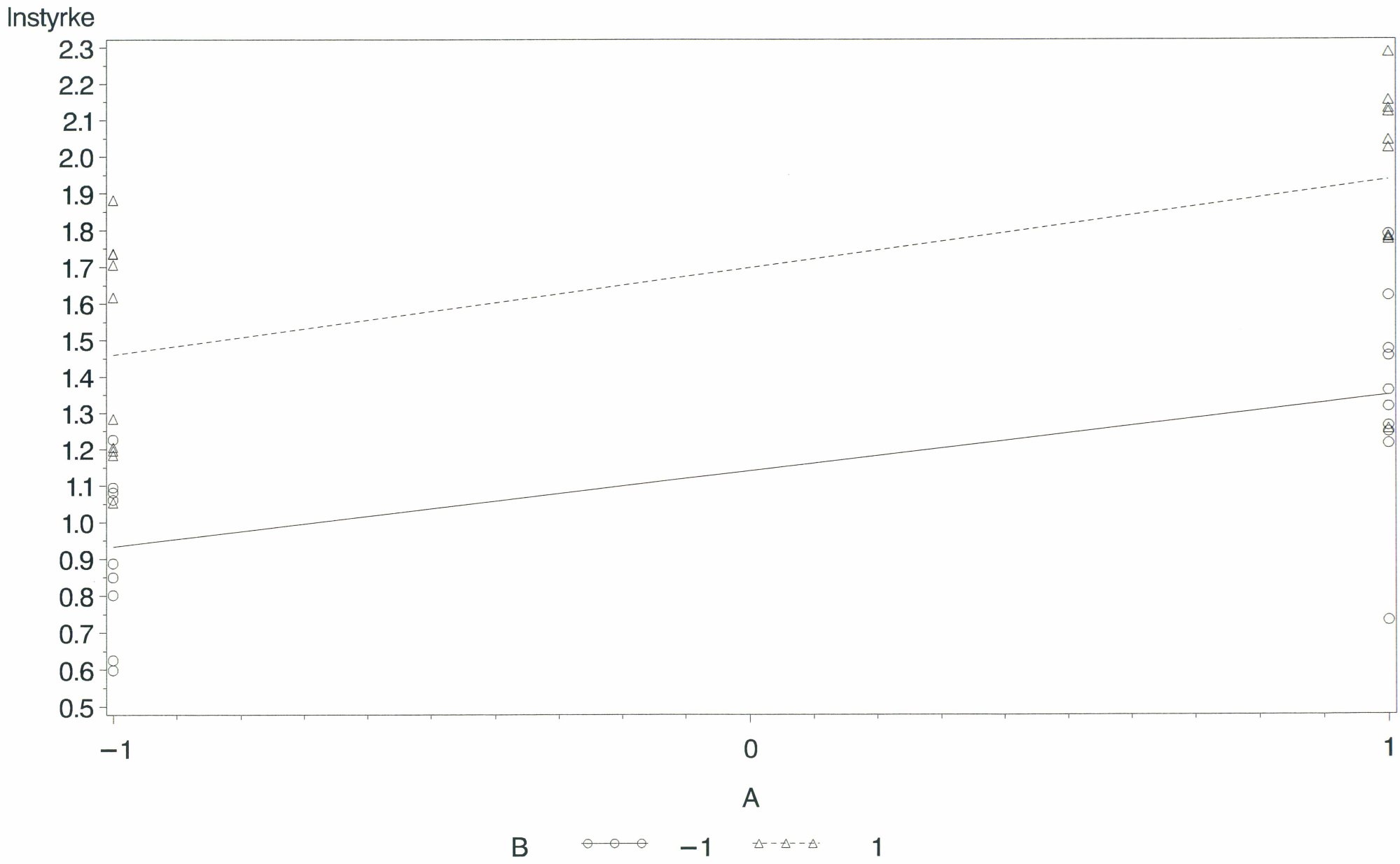

3
2
1
0
−1
−2
−3
p fraktil
0
1
2
3
lnstyrke
behAB
A
AogB
B
intet

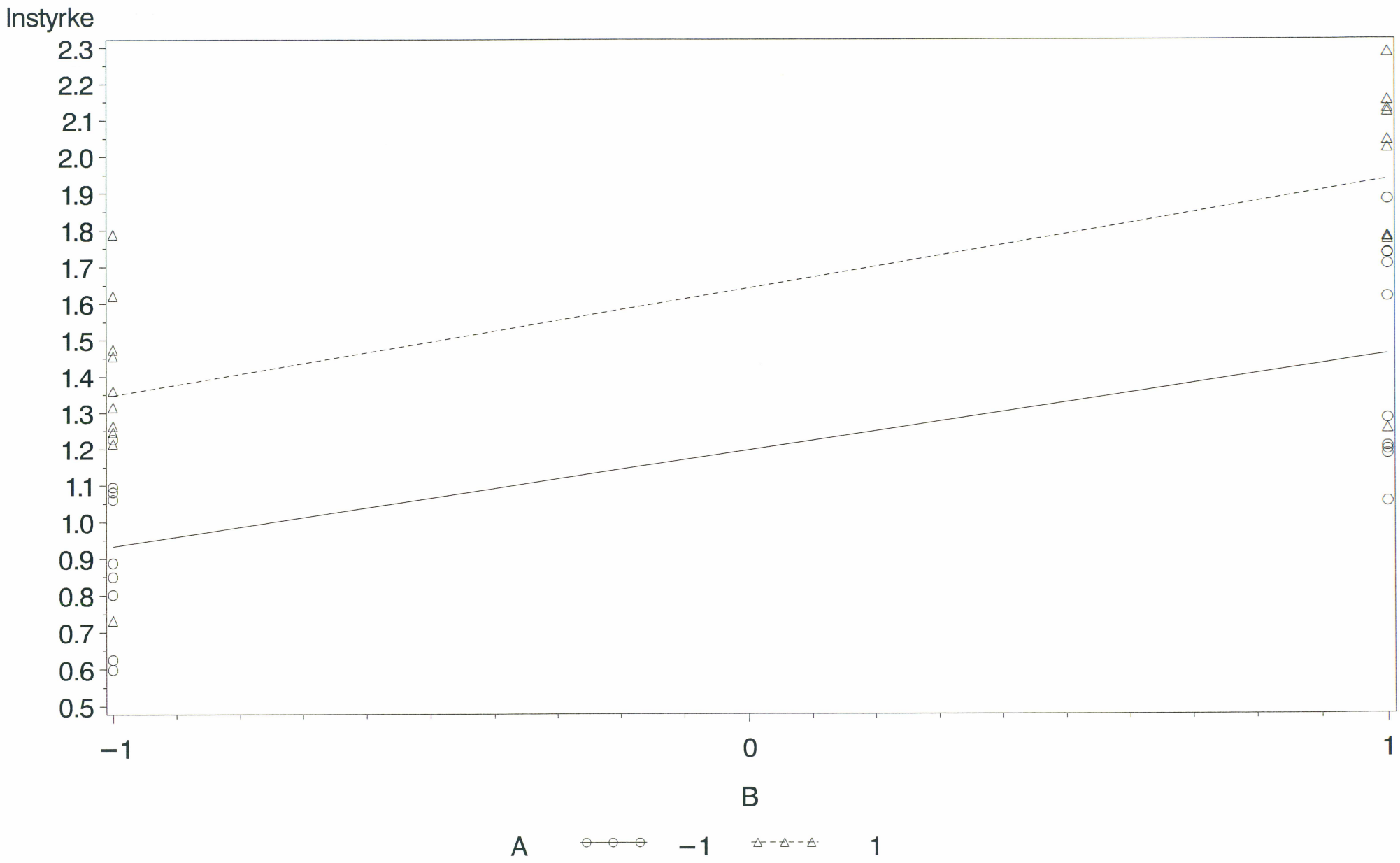

styrke
10
9
8
7
6
5
4
3
2
1
0
0
1
2
3
gruppe


Instyrke
3
2
1
0
0
1
2
3
gruppe

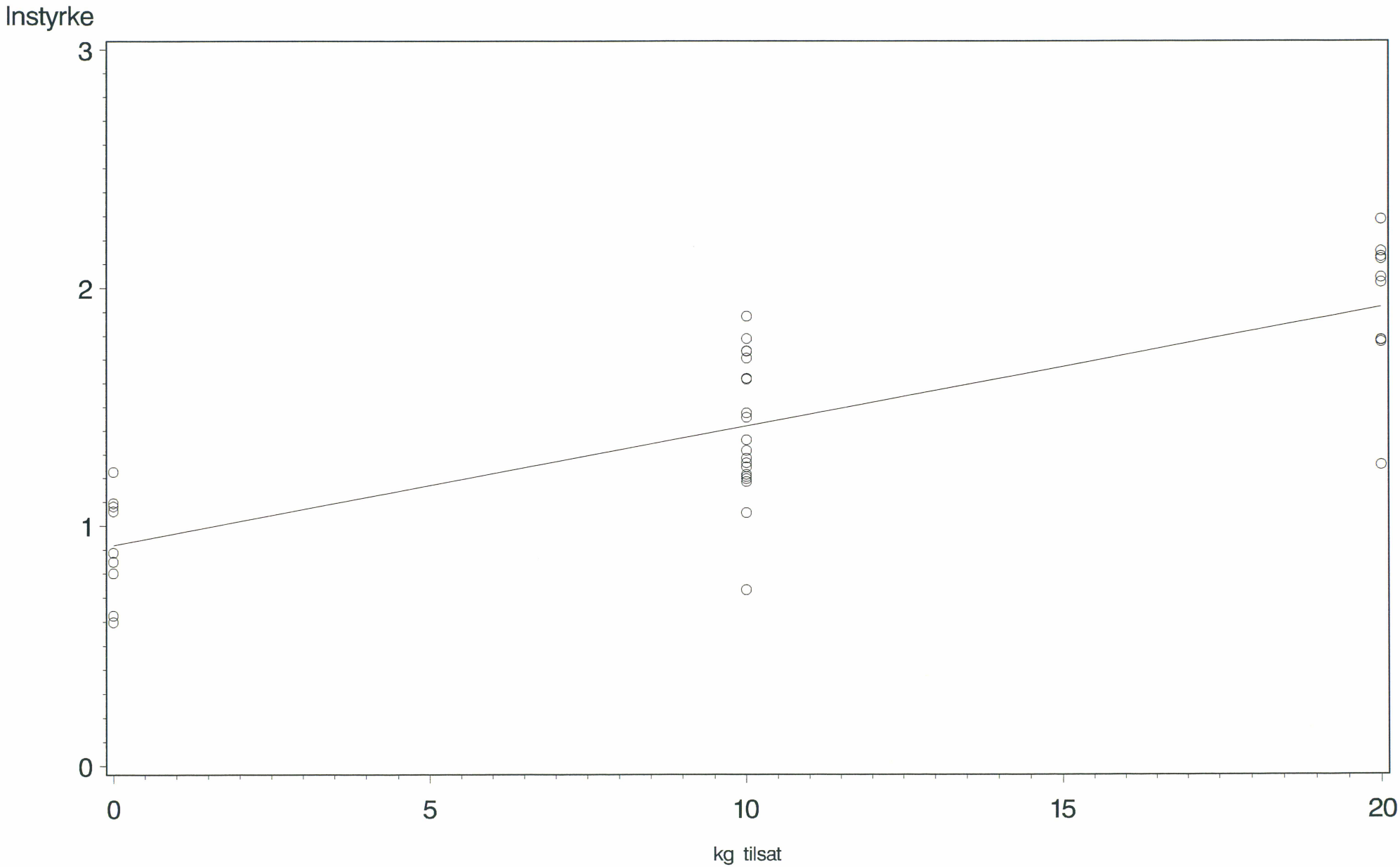

styrke
10
9
8
7
6
5
4
3
2
1
−1
0
1
A
B
−1
1


styrke
10
9
8
7
6
5
4
3
2
1
−1
0
1
B
A
−1
1


Instyrke
2.3
2.2
2.1
2.0
1.9
1.8
1.7
1.6
1.5
1.4
1.3
1.2
1.1
1.0
0.9
0.8
0.7
0.6
0.5
−1
0
1
A
B
−1
1


Instyrke
2.3
2.2
2.1
2.0
1.9
1.8
1.7
1.6
1.5
1.4
1.3
1.2
1.1
1.0
0.9
0.8
0.7
0.6
0.5
−1
0
1
B
A
−1
1

Logtraekstyrke mod maengde tilsat stof



Macro "BARTLETT"

Dataset: FLISER
Responsvariabel: LNSTYRKE
Gruppevariabel: GRUPPE

Beregningsskema:

i	ni	Si	USSi	Si2/ni	SSDi	fi	Estimeret varians	Gennem-snit
0	10	9.325	9.1054	8.6958	0.4095	9	0.04550	0.9325
1	10	13.477	18.8736	18.1620	0.7116	9	0.07907	1.3477
2	10	14.604	22.1504	21.3287	0.8216	9	0.09129	1.4604
3	10	19.361	38.2830	37.4832	0.7998	9	0.08886	1.9361
	40	56.767	88.4123	85.6698	2.7425	36	0.07618	

Bartletts test:

C	-2lnQ	Ba	Prob > chi
1.04630	1.28902	1.23198757	0.74534

The GLM Procedure

Dependent Variable: lnstyrke

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	5.10823527	1.70274509	22.35	<.0001
Error	36	2.74254471	0.07618180		
Corrected Total	39	7.85077998			

R-Square	Coeff Var	Root MSE	lnstyrke Mean
0.650666	19.44875	0.276011	1.419168

Source	DF	Type I SS	Mean Square	F Value	Pr > F
behA	1	1.98370188	1.98370188	26.04	<.0001
behB	1	3.11539154	3.11539154	40.89	<.0001
behA*behB	1	0.00914186	0.00914186	0.12	0.7311

Parameter		Estimate		Standard Error	t Value
Intercept		1.936058552	B	0.08728218	22.18
behA	A ikke tilsat	-0.475623185	B	0.12343565	-3.85
behA	A tilsat	0.000000000	B	.	.
behB	B ikke tilsat	-0.588392428	B	0.12343565	-4.77

Parameter		Pr > |t|
Intercept		<.0001
behA	A ikke tilsat	0.0005
behA	A tilsat	.
behB	B ikke tilsat	<.0001

The GLM Procedure

Dependent Variable: lnstyrke

Parameter	Estimate	Standard Error	t Value
behB B tilsat	0.000000000 B	.	.
behA*behB A ikke tilsat B ikke tilsat	0.060471005 B	0.17456437	0.35
behA*behB A ikke tilsat B tilsat	0.000000000 B	.	.
behA*behB A tilsat B ikke tilsat	0.000000000 B	.	.
behA*behB A tilsat B tilsat	0.000000000 B	.	.

Parameter	Pr > |t|
behB B tilsat	.
behA*behB A ikke tilsat B ikke tilsat	0.7311
behA*behB A ikke tilsat B tilsat	.
behA*behB A tilsat B ikke tilsat	.
behA*behB A tilsat B tilsat	.

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: lnstyrke

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	5.09909341	2.54954671	34.28	<.0001
Error	37	2.75168656	0.07436991		
Corrected Total	39	7.85077998			

R-Square	Coeff Var	Root MSE	lnstyrke Mean
0.649502	19.21607	0.272708	1.419168

Source	DF	Type I SS	Mean Square	F Value	Pr > F
behA	1	1.98370188	1.98370188	26.67	<.0001
behB	1	3.11539154	3.11539154	41.89	<.0001

Parameter		Estimate		Standard Error	t Value	Pr > |t|
Intercept		1.920940801	B	0.07468429	25.72	<.0001
behA	A ikke tilsat	-0.445387682	B	0.08623799	-5.16	<.0001
behA	A tilsat	0.000000000	B	.	.	.
behB	B ikke tilsat	-0.558156926	B	0.08623799	-6.47	<.0001
behB	B tilsat	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: lnstyrke

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1.98370188	1.98370188	12.85	0.0009
Error	38	5.86707810	0.15439679		
Corrected Total	39	7.85077998			

R-Square	Coeff Var	Root MSE	lnstyrke Mean
0.252676	27.68759	0.392934	1.419168

Source	DF	Type I SS	Mean Square	F Value	Pr > F
behA	1	1.98370188	1.98370188	12.85	0.0009

Parameter		Estimate	Standard Error	t Value	Pr > |t|
Intercept		1.641862338 B	0.08786262	18.69	<.0001
behA	A ikke tilsat	-0.445387682 B	0.12425651	-3.58	0.0009
behA	A tilsat	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: lnstyrke

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	3.11539154	3.11539154	25.00	<.0001
Error	38	4.73538844	0.12461549		
Corrected Total	39	7.85077998			

R-Square	Coeff Var	Root MSE	lnstyrke Mean
0.396826	24.87437	0.353009	1.419168

Source	DF	Type I SS	Mean Square	F Value	Pr > F
behB	1	3.11539154	3.11539154	25.00	<.0001

Parameter		Estimate		Standard Error	t Value	Pr > |t|
Intercept		1.698246960	B	0.07893525	21.51	<.0001
behB	B ikke tilsat	-0.558156926	B	0.11163131	-5.00	<.0001
behB	B tilsat	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: lnstyrke

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	5.10823527	1.70274509	22.35	<.0001
Error	36	2.74254471	0.07618180		
Corrected Total	39	7.85077998			

R-Square	Coeff Var	Root MSE	lnstyrke Mean
0.650666	19.44875	0.276011	1.419168

Source	DF	Type I SS	Mean Square	F Value	Pr > F
TILSAT	1	5.03550890	5.03550890	66.10	<.0001
AB	1	0.00914186	0.00914186	0.12	0.7311
gruppe	1	0.06358451	0.06358451	0.83	0.3670

Parameter		Estimate		Standard Error	t Value	Pr > |t|
Intercept		0.9848121821	B	0.19516890	5.05	<.0001
TILSAT		0.0475623185	B	0.01234356	3.85	0.0005
AB	0	-.0522982378	B	0.21379682	-0.24	0.8081
AB	10	0.0000000000	B	.	.	.
AB	20	0.0000000000	B	.	.	.
gruppe	0	0.0000000000	B	.	.	.
gruppe	1	-.1127692433	B	0.12343565	-0.91	0.3670
gruppe	2	0.0000000000	B	.	.	.
gruppe	3	0.0000000000	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: lnstyrke

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	5.03550890	5.03550890	67.97	<.0001
Error	38	2.81527108	0.07408608		
Corrected Total	39	7.85077998			

R-Square	Coeff Var	Root MSE	lnstyrke Mean
0.641402	19.17937	0.272188	1.419168

Source	DF	Type I SS	Mean Square	F Value	Pr > F
TILSAT	1	5.03550890	5.03550890	67.97	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
TILSAT	1	5.03550890	5.03550890	67.97	<.0001

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	0.9173961929	0.07454164	12.31	<.0001
TILSAT	0.0501772304	0.00608630	8.24	<.0001

The GLM Procedure

Dependent Variable: lnstyrke

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	5.04465076	2.52232538	33.26	<.0001
Error	37	2.80612922	0.07584133		
Corrected Total	39	7.85077998			

R-Square	Coeff Var	Root MSE	lnstyrke Mean
0.642567	19.40524	0.275393	1.419168

Source	DF	Type I SS	Mean Square	F Value	Pr > F
bh	2	5.04465076	2.52232538	33.26	<.0001

Parameter		Estimate	Standard Error	t Value	Pr > |t|
Intercept		0.932513944 B	0.08708693	10.71	<.0001
bh	10 kg tilsat	0.471536801 B	0.10665927	4.42	<.0001
bh	20 kg tilsat	1.003544608 B	0.12315951	8.15	<.0001
bh	Intet tilsat	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.

opgave 4.9

$X_1, \ldots, X_n$ uafh. $\quad i = 1, \ldots, n : X_i \sim \Gamma(\alpha_i, \beta)$

$Y_i := \sum_{j=1}^{i} X_j \Big/ \sum_{j=1}^{i+1} X_j \quad , i = 1, \ldots, n-1$

$Y_n := \sum_{j=1}^{n} X_i$

Påstand: $\quad Y_i \sim B(\alpha_1 + \ldots + \alpha_i, \alpha_{i+1}) \quad i = 1, \ldots, n-1$

$Y_n \sim \Gamma(\alpha_1 + \ldots + \alpha_n, \beta)$

$Y_1, \ldots, Y_n$ uafh.

Bevis:

I afl. nr: 15 i SSZ har vi vist, at

$Z_1 \perp\!\!\!\perp Z_2 \quad , Z_i \sim \Gamma(\gamma_i, \beta) : \frac{Z_1}{Z_1 + Z_2} \sim B(\gamma_1, \gamma_2)$

Af de udregninger (hvor vi bl.a. brugte trans. sætn.)

følger også at tætheden for $(Z_1 + Z_2, Z_1/Z_1+Z_2)$ splitter

op således, at $Z_1 + Z_2 \perp\!\!\!\perp Z_1/Z_1+Z_2$

Jf. 4.23 i Hoffmann er $\sum_{j=1}^{i} X_j \sim \Gamma(\alpha_1 + \ldots + \alpha_i, \beta)$

(specielt er $Y_n \sim \Gamma(\alpha_1 + \ldots + \alpha_n, \beta)$) ved for givet

$i = 1, \ldots, n-1$ at sætte $Z_1 = \sum_{j=1}^{i-1} X_j$ og $Z_2 = X_i$

følger det ønskede, idet $X_1 + \ldots + X_{i-1} \perp\!\!\!\perp X_i$

$\square$

opgave 4.10

$X_1, \ldots, X_n$ uafh. $\quad X_i \sim N(\mu, \sigma^2) \quad i = 1, \ldots, n$

Påstand: $\sum_{i=1}^{n} (X_i - \overline{X}_\cdot)^2 \sim \sigma^2 \chi^2(n-1)$

og $\quad \overline{X}_\cdot \perp\!\!\!\perp \sum_{i=1}^{n} (X_i - \overline{X}_\cdot)^2$

Bevis:

Idet projektionen af $(X_1, \ldots, X_n)$ ned på $L_1 = sp\left(\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}\right)$ er $\overline{X}_\cdot$ er påstanden en direkte konsekvens af sætn. 4.1 (1). $\square$

Afdeling for Teoretisk Statistik
Institut for Matematiske Fag
Aarhus Universitet

Statistik 1 og Statistik α
Jørgen Granfeldt
8. november 2002

Meddelelse 10

Forelæsningerne i sidste uge (Onsdag den 30. oktober) Afsluttede Afsnit 4.2 *Sammenligning af flere regressionslinjer.*

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

Forelæsningerne i denne uge (Onsdag den 6. november) Afsnit 4.4.1 *Fordelingsresultater i den lineære normale model.*

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

Forelæsningerne i næste uge (Onsdag den 13. november of fredag den 15. november) Kapitel 6 *Multinomialfordelte data.*

Øvelserne i denne uge

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

Øvelserne i de tre næste uger

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

(Uge 47: 18. november - 24. november) 4.11, 6.1, 6.4, 6.9, 6.10.

(Uge 48: 25. november - 1. december) 6.11, 7.6, 7.8 samt supplerende opgaver som stilles på Meddelelse 11.

Afleveringsopgaver

(Uge 46) Ingen afleveringsopgave.

(Uge 47) Ingen afleveringsopgave.

(Uge 48) Måske en afleveringsopgave.

Udleveret materiale

Intet

Nogle SAS vink

Følgende uddrag er fra den den engelske version af bio- og geostatistik bøgerne. Det forklarer Type I kvadratsummerne, som gives i udskriften fra *PROC GLM*. Type I kvadratsummerne kan være en hjæp til at kontrollere egne beregninger af F-tests for reduktioner mellem lineære normale modeller. Sidehenvisningerne er rettet, så de passer med den danske udgave.

Type I Sums of Squares in *PROC GLM*

The model statement used to specify model M_1 in the program on page 4.20 was

```
MODEL blodtryk = alder gruppe alder*gruppe/
                        SS1 SOLUTION;
```

and the option `SS1` gave the following table in the output listing on page 4.24

Source	DF	Type I SS	Mean Square	F Value	Pr > F
alder	1	3912.654942	3912.654942	207.53	<.0001
gruppe	1	393.079752	393.079752	20.85	0.0001
alder*gruppe	1	0.403936	0.403936	0.02	0.8849

This is the table we have in mind when we speak of the type I sums of squares, although strictly speaking, the type I sums of squares are those in the column labeled `Type I SS`. We will refer to it as the `SS1` table. An entry in the `Mean Square` column is the entry in the same row of the `Type I SS` column divided by the entry in the same row in the `DF` column.

The entries of the `SS1` table are related to some of the F-test for model reduction that were used in the analysis of Example 3.4 in Section 4.2. For easy reference we copy the F-test for the reduction from M_1 to M_2^* that was given on page 4.17:,

$$F = \frac{\dfrac{452.892448 - 452.488513}{25-24}}{18.853688} = \frac{0.403935}{18.853688} = 0.02142, \tag{1}$$

and the F-test for the reduction M_2^* from to M_3:

$$F = \frac{\dfrac{845.972201 - 452.892448}{26-25}}{18.115698} = \frac{393.079753}{18.115698} = 21.70. \tag{2}$$

The bottom row of the `SS1` table gives the F-test and related quantities for the reduction from M_1 to M_2^* and in (1) we recognize the degrees of freedom for the numerator of the F-test as the entry in the `DF` column, the numerator of the F-test as the entry in the `Mean Square` column, apart from the effect of rounding that shows in the last digit, the F-test statistic in the `F Value` column, and the significance probability based on the $F(1,24)$-distribution in the `Pr > F` column.

The row above which is labeled `group` **does not** give the F-test for the reduction from M_2^* to M_3, but it still gives related quantities that may be useful; the degrees of freedom for the numerator of the F-test is the entry in the `DF` column and the numerator of the F-test is the entry in the `Mean Square` column. The `F value` is not the F-test in (2) but the only difference is that the estimated variance in M_1, $s_1^2 = 18.853688$, is used in the denominator instead of the estimated variance in M_2^*, $s_{02}^{2*} = 18.115698$. The `Pr > F` is computed using the appropriate $F(1,24)$-distribution.

The entries in the top row of the `SS1` table are similar to the `group` row. The table gives relevant quantities for the numerator of the F-test for the reduction from M_3 to M_∞, the model for a single sample, but s_1^2 is used in the denominator to compute the F-value.

Type I Sums of Squares in *PROC GLM*

What has already been described for model M_1 in Example **??** will now be described in general terms. We will consider the table of type I sums of squares which is derived from the model formula,

$$\texttt{MODEL x = A}_k\ \texttt{A}_{k-1}\ \ldots\ \texttt{A}_2\ \texttt{A}_1\texttt{;}$$

where $\texttt{A}_i$ $i = 1,\ldots,k$ may be a factor, i.e., it appears in the `CLASS` stamement of *PROC GLM*, or a variable or a product of variables and factors. A model formula with k terms defines a sequence of $k+1$ simpler models, $M_1 \to M_2 \to \cdots \to M_k \to M_{k+1}$, where M_1 is defined by the full model formula and the following models are derived from M_1 by successively removing terms from the model formula from the right, first A_1, then A_2, and so on.

M_1: `MODEL` $\texttt{x = A}_k\ \texttt{A}_{k-1}\ \ldots\ \texttt{A}_2\ \texttt{A}_1\texttt{;}$

M_2: `MODEL` $\texttt{x = A}_k\ \texttt{A}_{k-1}\ \ldots\ \texttt{A}_2\texttt{;}$

M_i: `MODEL` $\texttt{x = A}_k\ \texttt{A}_{k-1}\ \ldots\ \texttt{A}_i\texttt{;}$

M_k: `MODEL` $\texttt{x = A}_k\texttt{;}$

M_{k+1}: `MODEL` `x = ;`

The empty model M_{k+1} is the model for one sample where all observations have the same mean. This is the model we have denoted M_∞, because it is the same model regardless of the number of terms in the original model formula.

Using the notation from linear normal models, we can explain the contents of the table of sums of squares of Type I and this has been done in in Table 1. Recall that n denotes the number of

Tabel 1: Explanation of Type I sums of squares. SS, DF, and MS refer to sums of squares, degrees of freedom, and mean square in the same row. In the last column the function F denotes the distribution function of the test statistic, $F = MS/s_1^2$, in a given row. In the row A_i it is an F-distribution with $d_{i+1} - d_i$ degrees of freedom in the numerator and $n - d_1$ degrees of freedom in the denominator.

Source	DF	Type I SS	Mean Square	F Value	Pr > F
A_k	$d_{k+1} - d_k$	$\|\|P_k(\mathbf{x}) - P_{k+1}(\mathbf{x})\|\|^2$	SS/DF	MS/s_1^2	$1 - F(F)$
A_{k-1}	$d_k - d_{k-1}$	$\|\|P_{k-1}(\mathbf{x}) - P_k(\mathbf{x})\|\|^2$	SS/DF	MS/s_1^2	$1 - F(F)$
...	...	...	...	...	...
A_i	$d_{i+1} - d_i$	$\|\|P_i(\mathbf{x}) - P_{i+1}(\mathbf{x})\|\|^2$	SS/DF	MS/s_1^2	$1 - F(F)$
...	...	...	...	...	...
A_1	$d_2 - d_1$	$\|\|P_1(\mathbf{x}) - P_2(\mathbf{x})\|\|^2$	SS/DF	MS/s_1^2	$1 - F(F)$

observations, L_i refers to the linear subspace of the model M_i and d_i is the dimension of L_i. The orthogonal projection on L_i is denoted by P_i and, finally, s_{0i}^2 denotes the estimate of the variance in M_i, except in M_1, where it has been denoted by s_1^2. Note that the sum of the type I sums of squares gives $\|P_1(\mathbf{x}) - P_\infty(\mathbf{x})\|^2$, which is the `MODEL Sum of Squares` in the analysis of variance table from M_1.

If a particular decreasing sequence of models is of interest, it may be possible to formulate the model in such a way in the `MODEL` statement that the interesting sequence of models corresponds to the sequence in the `SS1` table. The `Type I SS` column will usually be studied first, and it will be noticed which terms in the model give the largest contributions to the `MODEL Sum of Squares`, and then the `Pr > F` column will be consulted to see which contributions are significant.

STAT TØ uv. 9

opgave 4.3

1)

Se udskriften for plottet af x mod t

M: $x_i \sim\sim N(\alpha + \beta t_i, \sigma^2)$

2)

$$\hat{\beta} \underset{s.3.110}{=} \frac{SP_{xt} - \frac{S_x S_t}{n}}{USS_t - \frac{S_t^2}{n}} = 4{,}651685$$

$$\hat{\alpha} \underset{s.3.110}{=} \frac{1}{n}\left(S_x - \hat{\beta} S_t\right) = -1{,}407445$$

$$s^2 \underset{s.3.110}{=} \frac{1}{n-2}\left(SSD_x - \frac{SPD_{xt}^2}{SSD_t}\right) = 0{,}0237285$$

3)

Se udskriften

Modellen for 2. reg. linie:

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

($i=1$ model A $\quad i=2$ model B)

4)

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

$$F = \frac{s_2^2}{s_1^2} = 1{,}235167$$

$$\varepsilon(\underline{x}) = 2\left(1 - F_{F(n_2-2,\, n_1-2)}(F)\right)$$

$$\geq 2(1-0{,}7) = 0{,}6 > 0{,}05$$

Altså forkastes hypotesen <u>ikke</u> og vi reducere til:

$M_1: \quad x_{ij} \sim\sim N(\alpha_i + \beta_i t_{ij}, \sigma^2)$

5)

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

$$\underset{(4.12)}{F} = \frac{\dfrac{0{,}39637311 - 0{,}39507512}{16-15}}{0{,}02633834} = 0{,}0492814$$

$$\varepsilon(\underline{x}) = 1 - F_{F(16-15,\, 15)}(F) \geq 1 - 0{,}5 = 0{,}5 > 0{,}05$$

Dvs. hypotesen forkastes ikke.

Modellen reduceres til

$M_2: \; x_{ij} \sim\sim N(\alpha + \beta_i t_{ij}, \sigma^2) \quad i=1,2, \; j=1,\dots,n_i$

6)

95% konfidensinterval for α er:

$[-1{,}518161644 ; -1{,}268800684]$

$t_{0{,}975}(16) = 2{,}120$,

endepkt.: $-1{,}393481264 \pm 0{,}05881150 \cdot 2{,}120$

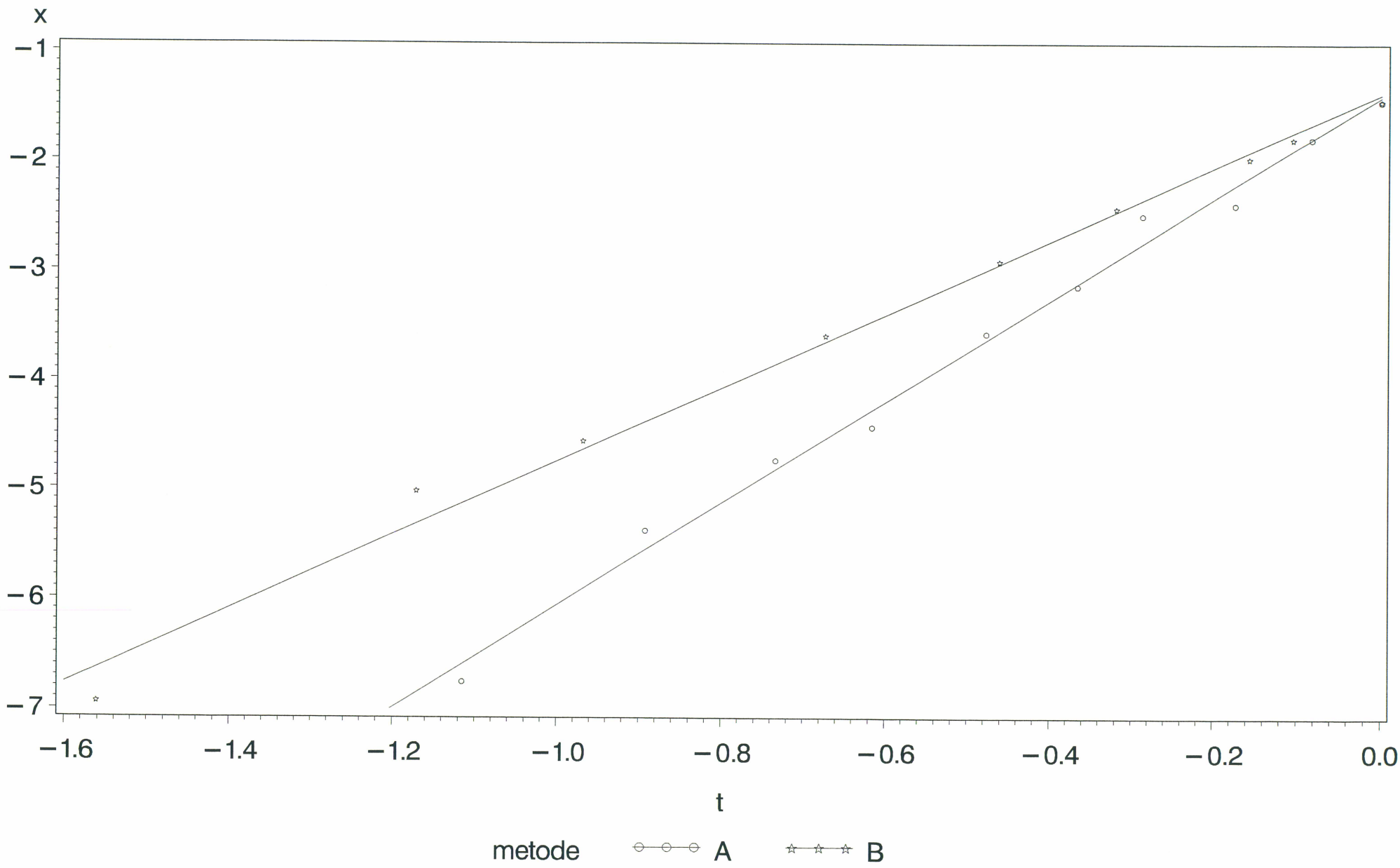
```
TITLE1 'Opgave 4_3';

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;


/* Bo Elberlings forsoeg havde tre cementeringsmetoder,
   A, B og C, men i opgaven benyttes kun A og B */

DATA data;
INPUT metode$ c bofit d;
CARDS;
   A    0 4.77E-06 4.68E-06
   B    0 4.77E-06 4.71E-06
   A    0.08 3.28E-06 3.32E-06
   B    0.1 3.42E-06 3.32E-06
   B    0.145 2.91E-06 2.79E-06
   A    0.16 2.17E-06 1.82E-06
   A    0.25  1.3E-06 1.65E-06
   B    0.274 1.74E-06 1.76E-06
   A    0.307 9.13E-07 8.68E-07
   B    0.37 1.11E-06 1.09E-06
   A    0.38 5.53E-07 5.63E-07
   A    0.46 2.97E-07 2.38E-07
   B    0.49 5.72E-07 5.53E-07
   A    0.52 1.74E-07 1.75E-07
   A    0.59 8.57E-08 9.18E-08
   B    0.62 2.26E-07 2.09E-07
   A    0.672 3.13E-08 2.33E-08
   B    0.69 1.19E-07 1.32E-07
   B    0.79  3.5E-08 1.94E-08
   C    0    4.77E-06 4.72E-06
   C    0.099 3.37E-06 3.31E-06
   C    0.21 2.17E-06 2.36E-06
   C    0.307 1.4E-06 1.55E-06
   C    0.39  9.11E-07 1.06E-06
   C    0.55 3.29E-07 3.86E-07
   C    0.67 1.16E-07 1.42E-07
   C    0.76 4E-08 5.33E-08
;

DATA data;
SET data;
k=0.00002;
x=log(d)-log(k);
t=log(1-c);
x=round(x,0.001);
t=round(t,0.001);
xt=x*t;
RUN;
```

```
PROC SORT DATA=data;
BY metode;
RUN;

L  A abdata;
SET data;
IF metode='A' OR metode='B';
RUN;

PROC PRINT DATA=abdata;
RUN;

PROC MEANS DATA=abdata VARDEF=DF N SUM USS VAR CSS;
VAR x t xt;
BY METODE;
RUN;



symbol1 c=DEFAULT
   i=rl
   v=circle
   cv=BLACK
   ;
symbol2 c=DEFAULT
   i=rl
   v==
   cv=BLACK
   ;
symbol3 c=DEFAULT
   i=rl
   v=circle
   cv=BLACK
   ;

axis1
   color=BLACK
   order=-1.6 to 0.0 by 0.2
   width=2.0
   style=1
   label=(color=BLACK  )
   ;
axis2
   color=BLACK
   order=-7 to -1 by 1
   width=2.0
   style=1
   label=(color=BLACK  )
   ;
axis3
   color=BLACK
   order=-7 to -1 by 1
   width=2.0
   style=1
   label=(color=BLACK  )
   ;
TITLE1 'x mod t _ Metode A og B';
```

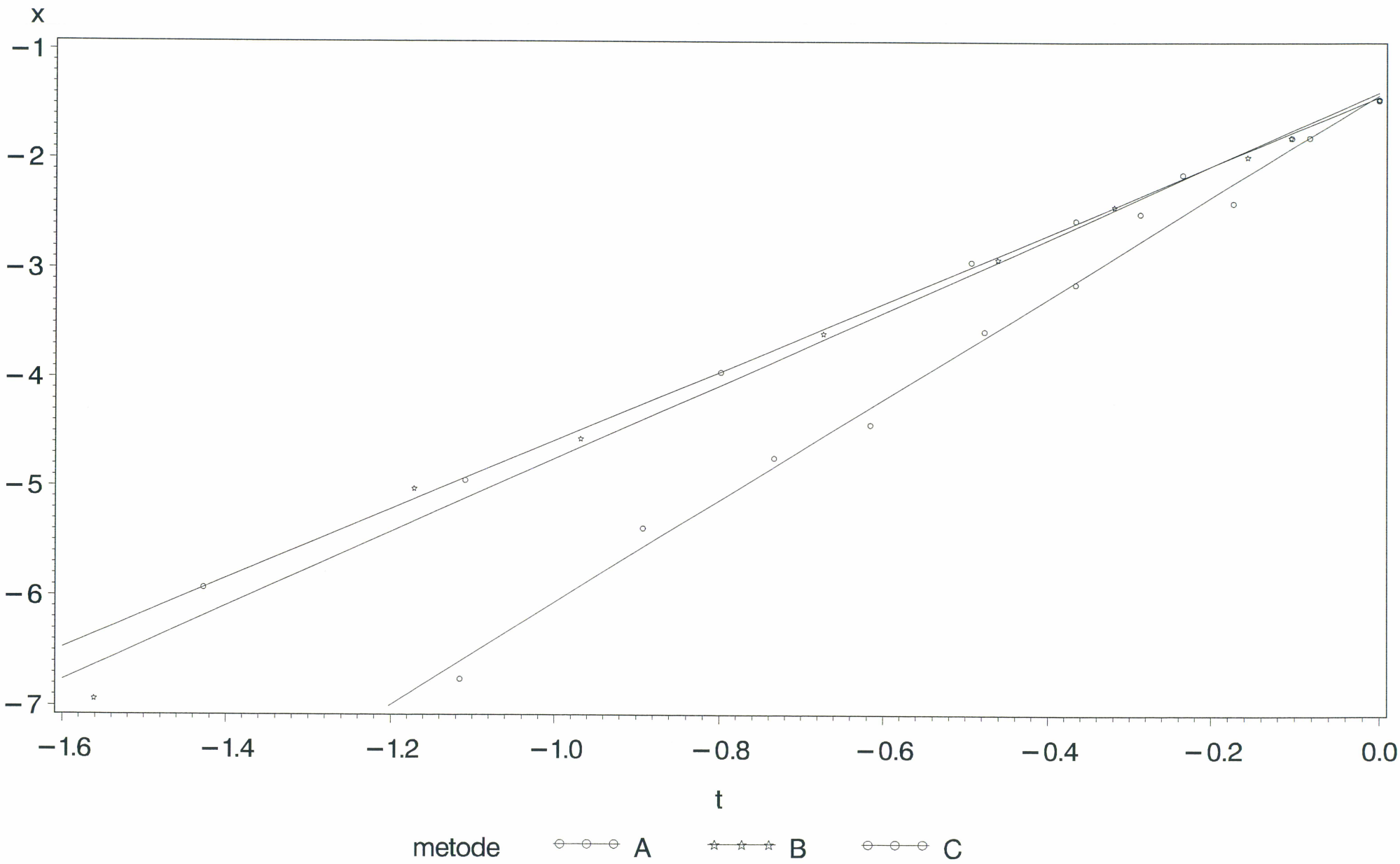
```
PROC GPLOT DATA=abdata;
     PLOT x*t=metode/
          haxis=axis1
          vaxis=axis2;
RUN;

TITLE1 'x mod t - Metode A, B og C';
PROC GPLOT DATA=data;
     PLOT x*t=metode/
          haxis=axis1
          vaxis=axis2;
RUN;

TITLE1 '2 separate regressionslinier';
PROC GLM DATA= abdata;
     MODEL x=t/SS1 ;
     BY metode;
RUN;

TITLE1 '2 regressionslinier';
PROC GLM DATA=abdata;
        CLASS metode;
        MODEL x= metode t metode*t/SS3 SOLUTION;
        OUTPUT OUT=ud1 p=predx;
RUN;

goptions reset=(axis, legend, pattern, symbol, title, footnote);

symbol1 c=DEFAULT
   i=NONE
   v=DOT
   cv=BLACK
   ;
symbol2 c=DEFAULT
   i=join
   l=1
   v=NONE
   cv=BLACK
   ;
axis1
   color=BLACK
   order=-1.6 to 0 by 0.2
   width=2.0
   style=1
   label=(color=BLACK  )
   ;
axis2
   color=BLACK
   order=-7 to -1 by 1
   width=2.0
   style=1
   label=(color=BLACK  )
   ;


TITLE1 '2 regressionslinjer';
proc gplot data=ud1 ;
   plot x*t predx*t /
```

```
      overlay
      haxis=axis1
      vaxis=axis2
      ;
run;

TITLE1 'samme intercept';
PROC GLM DATA=abdata;
CLASS metode;
MODEL x= t metode*t/SS1 SOLUTION;
        OUTPUT OUT=ud2 p=pred2x;
RUN;

TITLE1 '2 regressionslinjer samme afskaering';
proc gplot data=ud2 ;
   plot x*t pred2x*t /
      overlay
      haxis=axis1
      vaxis=axis2
      ;
run;


QUIT;
```

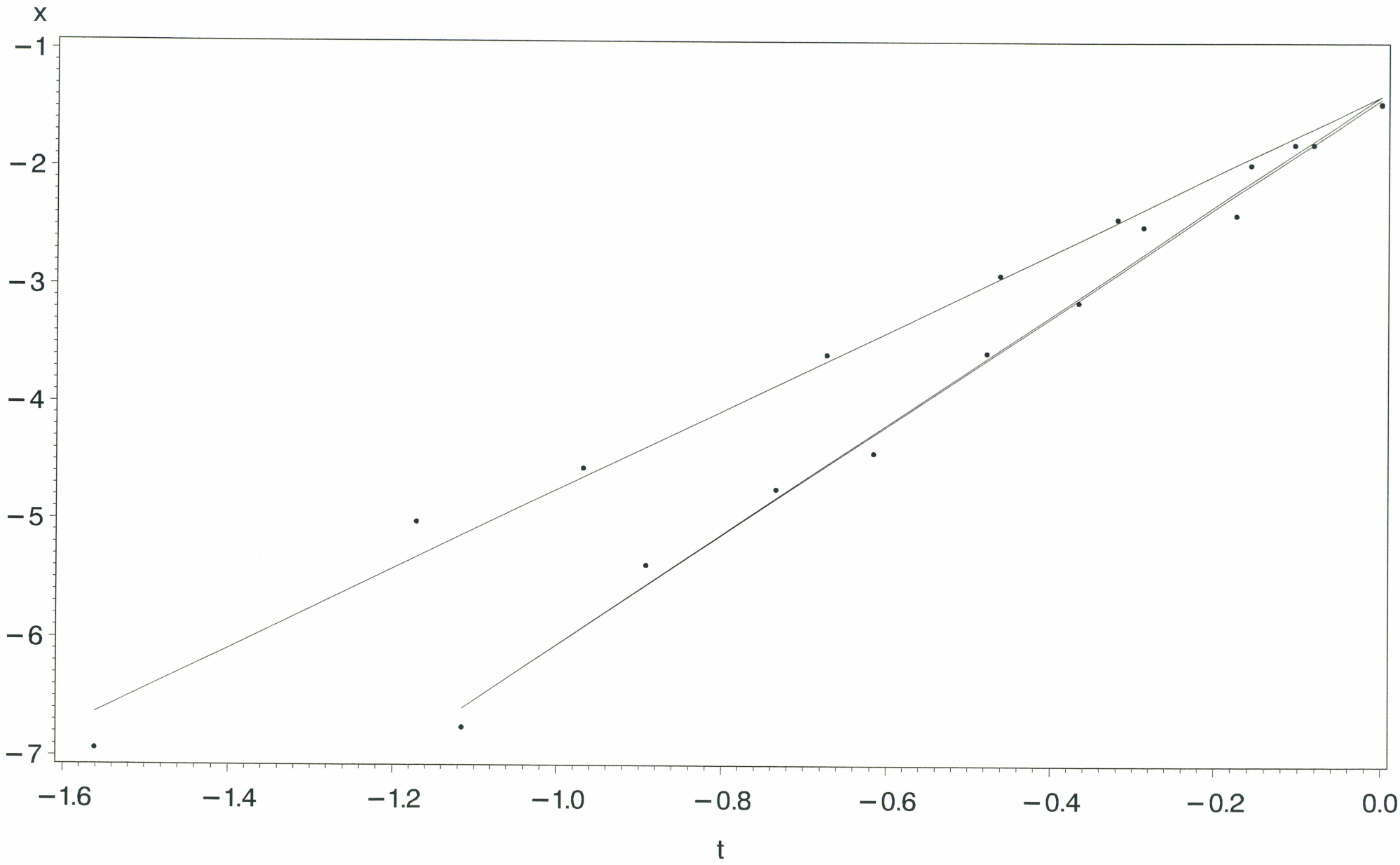
x mod t _ Metode A og B



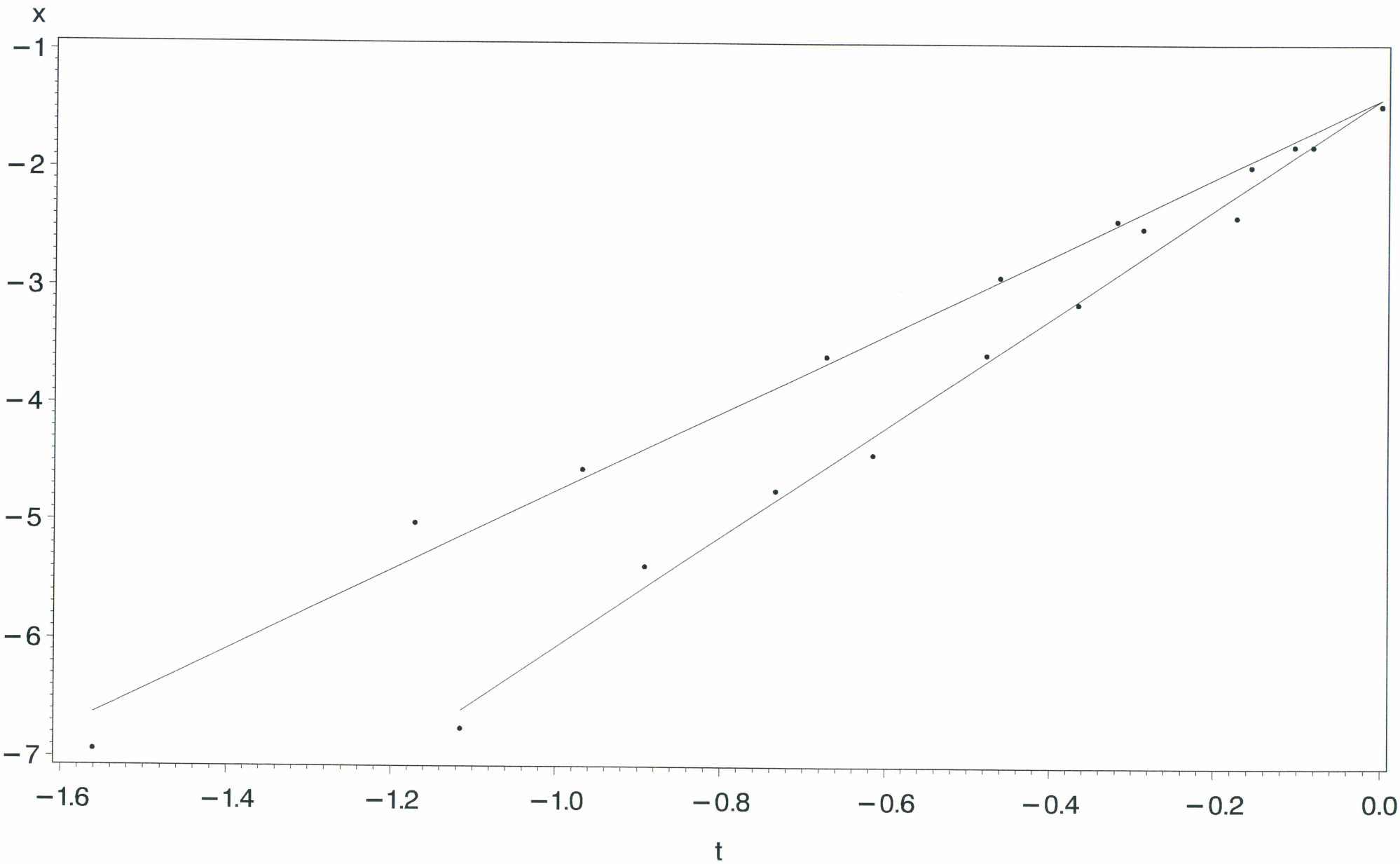
x mod t – Metode A, B og C



2 regressionslinjer



2 regressionslinjer samme afskaering



Obs	metode	c	bofit	d	k	x	t	xt
1	A	0.000	.000004770	.000004680	.00002	-1.452	0.000	0.0000
2	A	0.080	.000003280	.000003320	.00002	-1.796	-0.083	0.1491
3	A	0.160	.000002170	.000001820	.00002	-2.397	-0.174	0.4171
4	A	0.250	.000001300	.000001650	.00002	-2.495	-0.288	0.7186
5	A	0.307	.000000913	.000000868	.00002	-3.137	-0.367	1.1513
6	A	0.380	.000000553	.000000563	.00002	-3.570	-0.478	1.7065
7	A	0.460	.000000297	.000000238	.00002	-4.431	-0.616	2.7295
8	A	0.520	.000000174	.000000175	.00002	-4.739	-0.734	3.4784
9	A	0.590	.000000086	.000000092	.00002	-5.384	-0.892	4.8025
10	A	0.672	.000000031	.000000023	.00002	-6.755	-1.115	7.5318
11	B	0.000	.000004770	.000004710	.00002	-1.446	0.000	0.0000
12	B	0.100	.000003420	.000003320	.00002	-1.796	-0.105	0.1886
13	B	0.145	.000002910	.000002790	.00002	-1.970	-0.157	0.3093
14	B	0.274	.000001740	.000001760	.00002	-2.430	-0.320	0.7776
15	B	0.370	.000001110	.000001090	.00002	-2.910	-0.462	1.3444
16	B	0.490	.000000572	.000000553	.00002	-3.588	-0.673	2.4147
17	B	0.620	.000000226	.000000209	.00002	-4.561	-0.968	4.4150
18	B	0.690	.000000119	.000000132	.00002	-5.021	-1.171	5.8796
19	B	0.790	.000000035	.000000019	.00002	-6.938	-1.561	10.8302

---------- metode=A ----------

The MEANS Procedure

Variable	N	Sum	USS	Variance	Corrected SS
x	10	-36.1560000	156.5995860	2.8748836	25.8739524
t	10	-4.7470000	3.4403830	0.1318869	1.1869821
xt	10	22.6847200	104.2922117	5.8702844	52.8325596

---------- metode=B ----------

Variable	N	Sum	USS	Variance	Corrected SS
x	9	-30.6600000	130.5931820	3.2680978	26.1447820
t	9	-5.4170000	5.5494330	0.2861251	2.2890009
xt	9	26.1594710	179.7301027	12.9618472	103.6947779

---------- metode=A ----------

The GLM Procedure

Dependent Variable: x

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	25.68412415	25.68412415	1082.42	<.0001
Error	8	0.18982825	0.02372853		
Corrected Total	9	25.87395240			

R-Square	Coeff Var	Root MSE	x Mean
0.992663	-4.260446	0.154041	-3.615600

Source	DF	Type I SS	Mean Square	F Value	Pr > F
t	1	25.68412415	25.68412415	1082.42	<.0001

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	-1.407445142	0.08293097	-16.97	<.0001
t	4.651684975	0.14138826	32.90	<.0001

---------- metode=B ----------

The GLM Procedure

Dependent Variable: x

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	25.93953512	25.93953512	884.67	<.0001
Error	7	0.20524688	0.02932098		
Corrected Total	8	26.14478200			

R-Square	Coeff Var	Root MSE	x Mean
0.992150	-5.026430	0.171234	-3.406667

Source	DF	Type I SS	Mean Square	F Value	Pr > F
t	1	25.93953512	25.93953512	884.67	<.0001

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	-1.380503390	0.08887290	-15.53	<.0001
t	3.366341055	0.11317911	29.74	<.0001

The GLM Procedure

Dependent Variable: x

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	51.83043730	17.27681243	655.96	<.0001
Error	15	0.39507512	0.02633834		
Corrected Total	18	52.22551242			

R-Square	Coeff Var	Root MSE	x Mean
0.992435	-4.614954	0.162291	-3.516632

Source	DF	Type III SS	Mean Square	F Value	Pr > F
metode	1	0.00129798	0.00129798	0.05	0.8273
t	1	50.25131283	50.25131283	1907.91	<.0001
t*metode	1	1.29137146	1.29137146	49.03	<.0001

Parameter		Estimate		Standard Error	t Value	Pr > |t|
Intercept		-1.380503390	B	0.08423145	-16.39	<.0001
metode	A	-0.026941753	B	0.12136275	-0.22	0.8273
metode	B	0.000000000	B	.	.	.
t		3.366341055	B	0.10726826	31.38	<.0001
t*metode	A	1.285343920	B	0.18356419	7.00	<.0001
t*metode	B	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: x

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	51.82913931	25.91456966	1046.07	<.0001
Error	16	0.39637311	0.02477332		
Corrected Total	18	52.22551242			

R-Square	Coeff Var	Root MSE	x Mean
0.992410	-4.475744	0.157395	-3.516632

Source	DF	Type I SS	Mean Square	F Value	Pr > F
t	1	48.31378156	48.31378156	1950.23	<.0001
t*metode	1	3.51535776	3.51535776	141.90	<.0001

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	-1.393481264		0.05881150	-23.69	<.0001
t	3.353672887	B	0.08808971	38.07	<.0001
t*metode A	1.317279283	B	0.11058219	11.91	<.0001
t*metode B	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.

opgave 4.4

1)

Se udskrift

2)

Modellen er:

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

Udregninger (s. 3.110)

$SSD_x = 0{,}448567863636$

$SSD_t = 223{,}909545455$

$SPD_{xt} = 6{,}86482272727$ $SSD_{02} = 0{,}238099877642$

$\hat{\beta}_1 = 0{,}0306589$

$\hat{\alpha}_1 = 0{,}876544$

$s_1^2 = 0{,}0119050$

3)

Udskriftene giver:

$\hat{\alpha}_2 = 0{,}8644323996$, $\hat{\beta}_2 = 0{,}0316959523$

$s_2^2 = 0{,}00849728$

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

$$F = \frac{s_1^2}{s_2^2} = 1{,}401036$$

$$\varepsilon(\underline{x}) = 2(1 - F_{(22-2,\,17-2)}(F)) \geq 2(1-0{,}9) = 0{,}2 > 0{,}05$$

Hypotesen forkastes ikke og vi reducerer til

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

4)

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

$$F = \frac{\frac{0{,}36564741 - 0{,}36555909}{36-35}}{0{,}01044455} = 0{,}00845608$$

$$\varepsilon(\underline{x}) = 1 - F_{(36-35,\,35)}(F) \geq 1 - 0{,}5 = 0{,}5 > 0{,}05$$

Dvs. hypotesen forkastes ikke og vi reducerer til

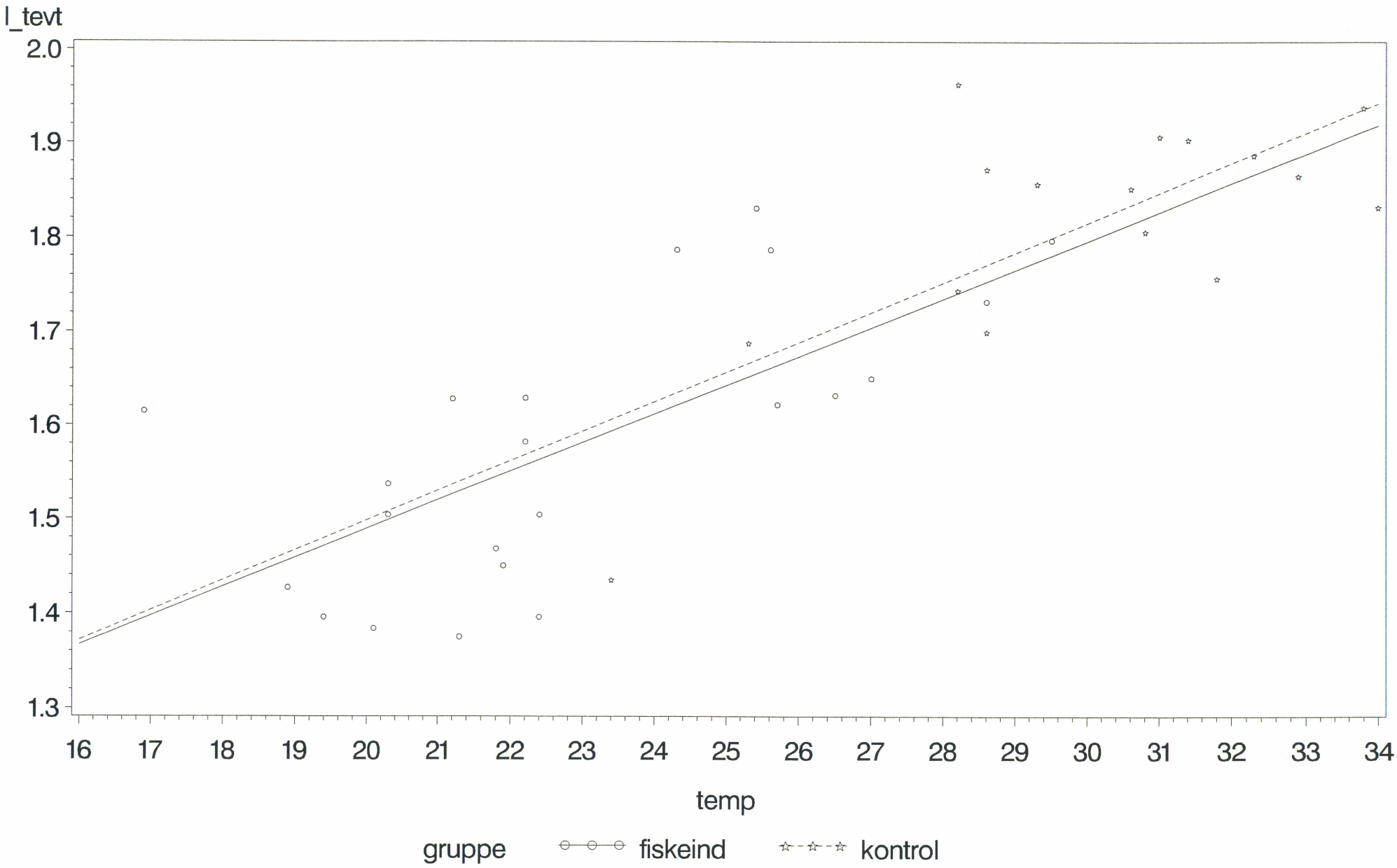
$M_2 : x_{ij} \sim\sim N(\alpha_i + \beta t_{ij}, \sigma^2)$

```
TITLE1 'Opgave 4_4';

/* 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 tevtdat;
INPUT gruppe $ temp l_tevt;
xt=temp*l_tevt;
CARDS;
fiskeind 21.8 1.468
fiskeind 21.3 1.375
fiskeind 25.4 1.831
fiskeind 24.3 1.787
fiskeind 18.9 1.427
fiskeind 22.2 1.629
fiskeind 21.9 1.450
fiskeind 21.2 1.628
fiskeind 28.6 1.732
fiskeind 20.3 1.537
fiskeind 25.7 1.622
fiskeind 25.6 1.787
fiskeind 22.4 1.396
fiskeind 19.4 1.396
fiskeind 26.5 1.632
fiskeind 22.2 1.582
fiskeind 22.4 1.504
fiskeind 27.0 1.650
fiskeind 29.5 1.797
fiskeind 20.1 1.384
fiskeind 20.3 1.504
fiskeind 16.9 1.615
kontrol 31.0 1.907
kontrol 31.4 1.904
kontrol 28.2 1.744
kontrol 31.4 1.904
kontrol 25.3 1.688
kontrol 23.4 1.435
kontrol 32.9 1.865
kontrol 28.2 1.963
kontrol 28.6 1.872
kontrol 30.8 1.806
kontrol 28.6 1.700
kontrol 31.8 1.757
kontrol 33.8 1.938
kontrol 29.3 1.857
kontrol 30.6 1.852
kontrol 34.0 1.832
kontrol 32.3 1.887
;
```

```
PROC MEANS DATA=tevtdat N SUM USS ;
VAR xt temp l_tevt;
BY gruppe;
RUN;

TITLE2 'to regressionslinjer';
PROC GLM DATA=tevtdat;
CLASS gruppe;
MODEL l_tevt = temp/SS1 SOLUTION;
BY gruppe;
RUN; QUIT;

OPTIONS ps=80;
TITLE2 'to regressionslinjer';
PROC GLM DATA=tevtdat;
CLASS gruppe;
MODEL l_tevt = gruppe temp gruppe*temp/SS1 SOLUTION;
RUN; QUIT;

TITLE2 'Samme haeldning';
TITLE3 'M2*';
PROC GLM DATA=tevtdat;
CLASS gruppe;
MODEL l_tevt = gruppe temp/SS1 SOLUTION;
RUN; QUIT;

TITLE2 'Sammenfaldende regressionslinjer';
TITLE3 'M3';
PROC GLM DATA=tevtdat;
CLASS gruppe;
MODEL l_tevt = temp/SS1 SOLUTION;
RUN; QUIT;
TITLE2 ' ';
TITLE3 ' ';



symbol1 c=DEFAULT
   i=rl
   v=circle
   cv=BLACK
   L=1
   ;
symbol2 c=DEFAULT
   i=rl
   v==
   cv=BLACK
   L=2
   ;

axis1
   color=BLACK
   order=1.3 to 2.0 by 0.1
   width=2.0
   style=1
   label=(color=BLACK  )
   ;
axis2
```

```
  color=BLACK
  order=16 to 35 by 1
  width=2.0
  style=1
  label=(color=BLACK  )
  ;


TITLE1 'logaritmen til TEVT mod hudtemperatur';
PROC GPLOT;
PLOT l_tevt*temp=gruppe;
RUN;
```

logaritmen til TEVT mod hudtemperatur



---------- gruppe=fiskeind ----------

The MEANS Procedure

Variable	N	Sum	USS
xt	22	802.4084000	30619.00
temp	22	503.9000000	11765.51
l_tevt	22	34.7330000	55.2840810

---------- gruppe=kontrol ----------

Variable	N	Sum	USS
xt	17	934.3499000	52346.65
temp	17	511.6000000	15525.84
l_tevt	17	30.9110000	56.4630390

Opgave 4_4
to regressionslinjer

---------- gruppe=fiskeind ----------

The GLM Procedure

Dependent Variable: l_tevt

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	0.21046799	0.21046799	17.68	0.0004
Error	20	0.23809988	0.01190499		
Corrected Total	21	0.44856786			

R-Square	Coeff Var	Root MSE	l_tevt Mean
0.469200	6.911065	0.109110	1.578773

Source	DF	Type I SS	Mean Square	F Value	Pr > F
temp	1	0.21046799	0.21046799	17.68	0.0004

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	0.8765443168	0.16862514	5.20	<.0001
temp	0.0306589106	0.00729169	4.20	0.0004

---------- gruppe=kontrol ----------

The GLM Procedure

Dependent Variable: l_tevt

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	0.13029031	0.13029031	15.33	0.0014
Error	15	0.12745922	0.00849728		
Corrected Total	16	0.25774953			

R-Square	Coeff Var	Root MSE	l_tevt Mean
0.505492	5.069625	0.092181	1.818294

Source	DF	Type I SS	Mean Square	F Value	Pr > F
temp	1	0.13029031	0.13029031	15.33	0.0014

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	0.8644323996	0.24461936	3.53	0.0030
temp	0.0316959523	0.00809446	3.92	0.0014

Opgave 4_4
to regressionslinjer

The GLM Procedure

Dependent Variable: l_tevt

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	0.89092665	0.29697555	28.43	<.0001
Error	35	0.36555909	0.01044455		
Corrected Total	38	1.25648574			

R-Square	Coeff Var	Root MSE	l_tevt Mean
0.709062	6.071756	0.102199	1.683179

Source	DF	Type I SS	Mean Square	F Value	Pr > F
gruppe	1	0.55016835	0.55016835	52.68	<.0001
temp	1	0.34066998	0.34066998	32.62	<.0001
temp*gruppe	1	0.00008832	0.00008832	0.01	0.9273

Parameter		Estimate	Standard Error	t Value	Pr > |t|
Intercept		0.8644323996 B	0.27120370	3.19	0.0030
gruppe	fiskeind	0.0121119172 B	0.31384340	0.04	0.9694
gruppe	kontrol	0.0000000000 B	.	.	.
temp		0.0316959523 B	0.00897413	3.53	0.0012
temp*gruppe	fiskeind	-.0010370418 B	0.01127747	-0.09	0.9273
temp*gruppe	kontrol	0.0000000000 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.

Opgave 4_4
Samme haeldning
M2*

The GLM Procedure

Dependent Variable: l_tevt

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	0.89083833	0.44541917	43.85	<.0001
Error	36	0.36564741	0.01015687		
Corrected Total	38	1.25648574			

R-Square	Coeff Var	Root MSE	l_tevt Mean
0.708992	5.987556	0.100781	1.683179

Source	DF	Type I SS	Mean Square	F Value	Pr > F
gruppe	1	0.55016835	0.55016835	54.17	<.0001
temp	1	0.34066998	0.34066998	33.54	<.0001

Parameter		Estimate	Standard Error	t Value	Pr > |t|
Intercept		0.8841947908 B	0.16313114	5.42	<.0001
gruppe	fiskeind	-.0163623457 B	0.05043704	-0.32	0.7475
gruppe	kontrol	0.0000000000 B	.	.	.
temp		0.0310392661	0.00535950	5.79	<.0001

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.

Opgave 4_4
Sammenfaldende regressionslinjer
M3

The GLM Procedure

Dependent Variable: l_tevt

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	0.88976939	0.88976939	89.77	<.0001
Error	37	0.36671635	0.00991125		
Corrected Total	38	1.25648574			

R-Square	Coeff Var	Root MSE	l_tevt Mean
0.708141	5.914715	0.099555	1.683179

Source	DF	Type I SS	Mean Square	F Value	Pr > F
temp	1	0.88976939	0.88976939	89.77	<.0001

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	0.8403776825	0.09036825	9.30	<.0001
temp	0.0323675730	0.00341614	9.47	<.0001

Opgave på medd. 9

1)

En trivialitet

2)

$\underline{e} \perp \underline{t} \iff \langle \underline{e}, \underline{t} \rangle = 0 \iff \sum t_i = 0$

3)

$\underline{u} := \underline{t} - \bar{t} \cdot \underline{e}$

Påstand: $\underline{e} \perp \underline{u}$

Bevis:

$\langle \underline{e}, \underline{u} \rangle = \langle \underline{e}, \underline{t} \rangle - \bar{t} \cdot \langle \underline{e}, \underline{e} \rangle = \sum t_i - n\bar{t} = 0$ $\square$

Påstand: $\|\underline{e}\|^2 = n$, $\|\underline{u}\|^2 = SSD_t$

Bevis:

$\|\underline{e}\|^2 = \langle \underline{e}, \underline{e} \rangle = n$

$\|\underline{u}\|^2 = \langle \underline{u}, \underline{u} \rangle = \langle \underline{u}, \underline{t} - \bar{t} \cdot \underline{e} \rangle \underset{\underline{u} \perp \underline{e}}{=} \langle \underline{u}, \underline{t} \rangle$

$= \langle \underline{t} - \bar{t} \cdot \underline{e}, \underline{t} \rangle = \sum t_i^2 - \bar{t} \cdot \sum t_i = \sum t_i^2 - \frac{1}{n} (\sum t_i)^2$

$= SSD_t$

$\square$

4)

Jf. sætn. 6.11 i F&B svarer projektion på L til den lin. trans. med matrix

$$T(T^*T)^{-1}T^* \quad , \quad T = (\underline{e}\ \underline{t})$$

$$\underset{2\times n}{T^*}\ \underset{n\times 2}{T} = \begin{pmatrix} n & \Sigma t_i \\ \Sigma t_i & \Sigma t_i^2 \end{pmatrix}$$

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

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

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

Jf. opg. 3.14 ser vi, at

$$\hat{\alpha} = \left[(T^*T)^{-1}T^*\underline{x}\right]_1 = \frac{1}{n\,SSD_t}\sum_{j=1}^{n}\left(\Sigma t_i^2 - t_j\Sigma t_i\right)x_j$$

$$= \frac{1}{n\,SSD_t}\left(S_x \cdot USS_t - S_t \cdot SP_{xt}\right)$$

$$= -\frac{S_t}{n}\frac{SPD_{xt}}{SSD_t} - \frac{S_t}{n}\frac{S_xS_t/n}{SSD_t} + \frac{S_x USS_t}{n\,SSD_t} = \frac{1}{n}\left(S_x - \frac{SPD_{xt}}{SSD_t}S_t\right)$$

$$\hat{\beta} = \left[(T^*T)^{-1}T^*\underline{x}\right]_2 = \frac{1}{n\,SSD_t}\sum_{j=1}^{n}\left(t_jx_jn - x_j\Sigma t_i\right)$$

$$= \frac{1}{n\,SSD_t}\left(n\,SP_{xt} - S_xS_t\right) = \frac{SPD_{xt}}{SSD_t}$$

Dette stemmer overens m. resultater på s. 3.73.

5)

At underrummet svarende til parallelle regressionslinier (modellen $x_{ij} \sim\sim N(\alpha_i + \beta t_{ij}, \sigma^2)$) er

$L_2 = sp(\underline{e}_1, \underline{e}_2, \underline{t}_1 + \underline{t}_2)$ er tivielt

Definer $\underline{u}_1 = \underline{t}_1 - \bar{t}_1 \cdot \underline{e}_1$, $\underline{u}_2 = t_2 - \bar{t}_2 \cdot \underline{e}_2$

Vi ser, at $e_1, e_2, \underline{u}_1 + \underline{u}_2$ er ort. og udspænder L_2

Jvf. sætn. 6.11. i F&B er den til L_2 hørende projektionsmatrix

$$T(T^* T)^{-1} T^* \quad , \quad T = (\underline{e}_1 \ \underline{e}_2 \ \underline{u}_1 + \underline{u}_2)$$

$$T^* T = \begin{pmatrix} n_1 & & 0 \\ & n_2 & \\ 0 & & \|\underline{u}_1 + \underline{u}_2\|^2 \end{pmatrix}$$

$$(T^* T)^{-1} = \begin{pmatrix} 1/n_1 & & 0 \\ & 1/n_2 & \\ 0 & & 1/\|\underline{u}_1 + \underline{u}_2\|^2 \end{pmatrix}$$

$$(T^* T)^{-1} T^* = \begin{pmatrix} 1/n_1 & \cdots & 1/n_1 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1/n_2 & \cdots & 1/n_2 \\ \frac{(\underline{u}_1 + \underline{u}_2)_1}{\|\underline{u}_1 + \underline{u}_2\|^2} & \cdots & & & & \frac{(\underline{u}_1 + \underline{u}_2)_n}{\|\underline{u}_1 + \underline{u}_2\|^2} \end{pmatrix}$$

$$\left(T(T^* T)^{-1} T^*\right)_{rs} = \frac{1}{n_1} \cdot \underline{e}_{1r} + \frac{1}{n_2} \cdot \underline{e}_{2r} + \frac{(\underline{u}_1 + \underline{u}_2)_r (\underline{u}_1 + \underline{u}_2)_s}{\|\underline{u}_1 + \underline{u}_2\|^2}$$

$$((T^* T)^{-1} T^* \underline{x})_1 = \bar{x}_{1\cdot} \quad , ((T^* T)^{-1} T^* \underline{x})_2 = \bar{x}_{2\cdot}$$

$$((T^* T)^{-1} T^* \underline{x})_3 = \sum_{i=1}^{2} \sum_{j=1}^{n_i} (t_{ij} - \bar{t}_{i\cdot}) x_{ij} \Big/ \sum_{i=1}^{2} \sum_{j=1}^{n_i} (t_{ij} - \bar{t}_{i\cdot})^2$$

Dvs. jvf. opgave 3.14

$$\hat{\alpha}_1 = \bar{x}_{1\cdot} - \bar{t}_{1\cdot} \left(\frac{SPD_{xt_1} + SPD_{xt_2} - S_{x_1} S_{t_1}/n_1 - S_{x_2} S_{t_2}/n_2}{SSD_{t_1} + SSD_{t_2}} \right)$$

$$\hat{\alpha}_2 = \bar{x}_{2\cdot} - \bar{t}_{2\cdot} \left(\text{———} \quad , \quad \text{———} \right)$$

$$\hat{\beta} = \frac{SPD_{xt_1} + SPD_{xt_2} - S_{x_1} S_{t_1}/n_1 - S_{x_2} S_{t_2}/n_2}{SSD_{t_1} + SSD_{t_2}}$$