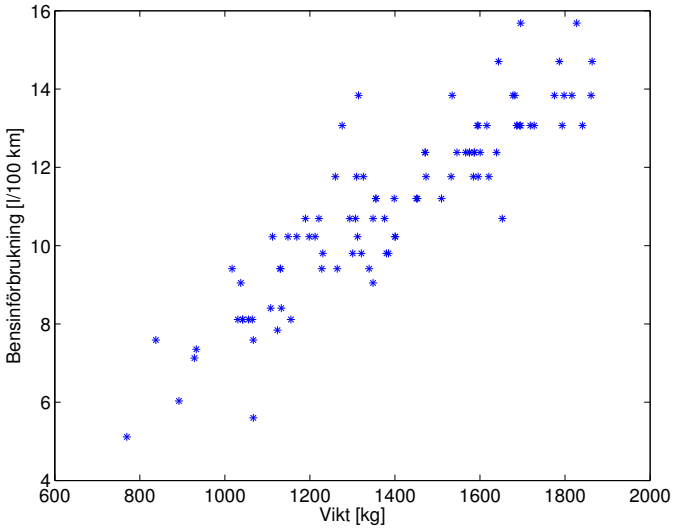
UTDRAG UR FÖRELÄSNINGSANTECKNINGAR I STATISTIKTEORI
LINJÄR REGRESSION OCH STOKASTISKA VEKTORER
MATEMATISK STATISTIK AK FÖR F, E, D, I, C, II; FMS 012
JOAKIM LÜBECK, MARS 2014

Innehåll

4 Enkel linjär regression

Med regressionsanalys kan man studera samband mellan olika variabler. I den enklaste formen, enkel linjär regression, studerar vi en variabel y som beror linjärt av en variabel x och där vi som vanligt har en slumpmässig störning eller avvikelse. Det kan till exempel vara en situation som beskrivs i figur 4.1.



Figur 4.1: Ett slumpmässigt urval av bilar där $y =$ "bensinförbrukning i stadskörning" är plottad mot $x =$ "vikt". Man kan kanske tänka sig ett linjärt samband mellan x och y som beskriver hur stor bensinförbrukning en "medelbil" av en viss vikt har.

Vi använder följande modell där y_i är n st oberoende observationer av

$$Y_i = \alpha + \beta x_i + \varepsilon_i, \quad \text{där } \varepsilon_i \in N(0, \sigma), \text{ oberoende av varandra}$$

så observationerna är $Y_i \in N(\alpha + \beta x_i, \sigma) = N(\mu_i, \sigma)$, dvs de är normalfördelade med väntevärde på den okända regressionslinjen $\mu(x) = \alpha + \beta x$ och med samma standardavvikelse σ som avvikelserna ε_i kring linjen har, se figur 4.2. Tidigare har vi haft modeller där observationerna (vi kallade dem oftast X_i men här är Y_i naturligare) $Y_i = \mu + \varepsilon_i$ hade samma väntevärde men nu är observationernas väntevärde en linjär funktion av x.

4.1 Punktskattningar och deras fördelning

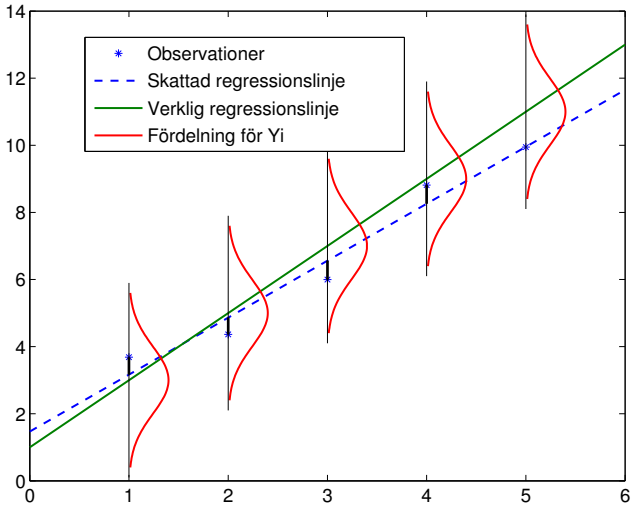
Skattningarna och deras fördelning härleds i appendix A, så här ges en kortfattad och lite mer lättläst sammanfattning.

ML- och MK-skattningarna av regressionslinjens lutning, β, och intercept, α, ges av

$$\beta^* = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}, \qquad \alpha^* = \bar{y} - \beta^* \bar{x}.$$

Eftersom β^* är en linjär funktion av observationerna Y_i ($\beta^* = \sum c_i Y_i$ där $c_i = (x_i - \bar{x})/S_{xx}$) och även α^* en linjär funktion av β^* och observationerna så är dessa skattningar normalfördelade med väntevärde och standardavvikelse enligt

$$\beta^* \in N(\beta, \frac{\sigma}{\sqrt{S_{xx}}}), \qquad \alpha^* \in N(\alpha, \sigma\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}).$$



Figur 4.2: Sann regressionslinje, observationer och skattad regressionslinje. Residualerna är markerade som de lodräta avstånden mellan observationerna och den skattade regressionslinjen.

De är dock *inte* oberoende av varandra. Man kan däremot visa att β^* och $\bar{Y}$ är oberoende[1] av varandra så kovariansen mellan α^* och β^* blir

$$C(\alpha^*, \beta^*) = C(\bar{Y} - \beta^*\bar{x}, \beta^*) = C(\bar{Y}, \beta^*) - \bar{x}C(\beta^*, \beta^*) = 0 - \bar{x}V(\beta^*) = -\bar{x}\frac{\sigma^2}{S_{xx}}.$$

Vi ser att kovariansen är negativ då $\bar{x} > 0$ och positiv då $\bar{x} < 0$, vilket verkar rimligt(?!)

Den, för väntevärdesriktighet, korrigerade ML-skattningen av variansen ges av

$$(\sigma^2)^* = s^2 = \frac{Q_0}{n-2}$$

där Q_0 är residualkvadratsumman

$$Q_0 = \sum_{i=1}^{n}(y_i - \alpha^* - \beta^* x_i)^2 = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$$

dvs summan av lodräta kvadratiska avvikelser från observationerna till den skattade regressionslinjen. Dessutom är

$$\frac{Q_0}{\sigma^2} \in \chi^2(n-2), \quad \frac{\beta^* - \beta}{d(\beta^*)} = \frac{\beta^* - \beta}{s/\sqrt{S_{xx}}} \in t(n-2) \quad \text{och} \quad \frac{\alpha^* - \alpha}{d(\alpha^*)} = \frac{\alpha^* - \alpha}{s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}} \in t(n-2).$$

Vi ser att $n-2$ som delas med i variansskattningen återkommer som parameter i χ^2- och t-fördelningen.

[1] Vi visar inte här att β^* och $\bar{Y}$ är oberoende av varandra, men det faktum att regressionslinjen alltid går genom punkten $(\bar{x}, \bar{y})$ gör det kanske troligt; om β över- eller underskattas påverkas inte $\bar{Y}$ av detta.

För att räkna ut kvadratsummorna S_{xx}, S_{yy} och S_{xy} "för hand" kan man ha användning av sambanden

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\Big(\sum_{i=1}^{n} x_i\Big)^2$$
$$S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{1}{n}\Big(\sum_{i=1}^{n} y_i\Big)^2$$
$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\Big(\sum_{i=1}^{n} x_i\Big)\Big(\sum_{i=1}^{n} y_i\Big).$$

Naturligtvis har vi även t.ex om s_x^2 är stickprovsvariansen för x-dataserien $S_{xx} = (n-1)s_x^2$.

4.2 Intervallskattningar

Eftersom skattningarna av α och β är normalfördelade får vi direkt konfidensintervall med konfidensgraden $1 - a$ (α är upptagen) precis som tidigare enligt

$$I_\beta = \beta^* \pm t_{a/2}(f)d(\beta^*) = \beta^* \pm t_{a/2}(n-2) \cdot \frac{s}{\sqrt{S_{xx}}}$$
$$I_\alpha = \alpha^* \pm t_{a/2}(f)d(\alpha^*) = \alpha^* \pm t_{a/2}(n-2) \cdot s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}.$$

Om σ skulle råka vara känd används naturligtvis den i stället för s och då även λ- i stället för t-kvantiler. För variansen blir intervallet

$$I_{\sigma^2} = \left(\frac{Q_0}{\chi^2_{a/2}(n-2)}, \frac{Q_0}{\chi^2_{1-a/2}(n-2)}\right) = \left(\frac{(n-2)s^2}{\chi^2_{a/2}(n-2)}, \frac{(n-2)s^2}{\chi^2_{1-a/2}(n-2)}\right).$$

4.3 Skattning av punkt på linjen

För ett givet värde x_0 är Y's väntevärde $E(Y(x_0)) = \alpha + \beta x_0 = \mu_0$, dvs en punkt på den teoretiska regressionslinjen. μ_0 skattas med motsvarande punkt på den skattade regressionslinjen som $\mu_0^* = \alpha^* + \beta^* x_0$. Vi ser direkt att skattningen är väntevärdesriktig samt att den måste vara normalfördelad (linjär funktion av två normalfördelade skattningar). Ett enkelt sätt att bestämma skattningens varians får vi om vi återigen utnyttjar att β^* och $\bar{Y}$ är oberoende av varandra (men inte av α^*)

$$V(\mu_0^*) = V(\alpha^* + \beta^* x_0) = [\alpha^* = \bar{Y} - \beta^*\bar{x}] = V(\bar{Y} + \beta^*(x_0 - \bar{x})) = [\text{ober}] =$$
$$= V(\bar{Y}) + (x_0 - \bar{x})^2 V(\beta^*) = \frac{\sigma^2}{n} + (x_0 - \bar{x})^2 \frac{\sigma^2}{S_{xx}} = \sigma^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right) \qquad \Longrightarrow$$
$$\mu_0^* \in N\left(\mu_0, \sigma\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}\right).$$

Vi får således ett konfidensintervall för μ_0 med konfidensgraden $1 - a$ som

$$I_{\mu_0} = \mu_0^* \pm t_{a/2}(f)d(\mu_0^*) = \alpha^* + \beta^* x_0 \pm t_{a/2}(n-2)s\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}.$$

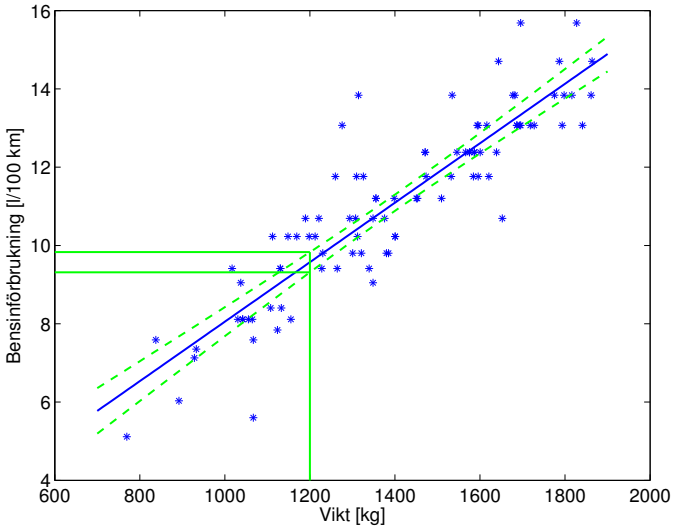
Exempel 4.1. I ett slumpmässigt urval av bilar avsattes $y =$ ”bensinförbrukning i stadskörning” som funktion av $x =$ ”vikt” i en linjär regressionsmodell $Y_i = \alpha + \beta x_i + \varepsilon_i$, $\varepsilon_i \in N(0, \sigma)$. Parametrarna skattas enligt resultaten i avsnitt 4.1 till $\alpha^* = 0.46$, $\beta^* = 0.0076$ samt $\sigma^* = 1.009$.

β är ett mått på hur mycket y beror av x, om vikten ökas med ett kg skattas ökningen av bensinförbrukningen med $\beta^* = 0.0076$ liter per 100 kilometer. Ett 95% konfidensintervall för β blir $I_\beta = [0.0068, 0.0084]$.

Antag att vi är speciellt intresserade av bilar som väger $x_0 = 1200$ kg. En skattning av medelförbrukingen μ_0 för denna typ av bilar blir då $\mu_0^* = \alpha^* + \beta^* x_0 = 9.57$ $\ell/100$ km. Ett 95% konfidensintervall för μ_0 blir med ovanstående uttryck $I_{\mu_0} = [9.32, 9.83]$. Detta intervall täcker alltså med sannolikhet 95% den sanna medelförbrukningen för 1200 kg’s bilar.

Observera att intervallet inte ger någon information om individuella 1200 kg bilars variation, så det är inte till så mycket hjälp till att ge någon uppfattning om en framtida observation (den 1200 kg bil du tänkte köpa?). Till detta behövs ett *prediktionsintervall*, se nästa avsnitt.

I figur 4.3 är konfidensintervallen förutom för 1200 kg bilar även plottat som funktion av vikten. I formeln för konfidensintervallet ser man att det är som smalast då $x_0 = \bar{x}$ vilket även kan antydas i figuren. Man ser även att observationerna i regel inte täcks av konfidensintervallen för linjen.



Figur 4.3: Bensinförbrukning enligt exempel 1.1. Skattad regressionslinje (–), konfidensintervall för linjen som funktion av vikt (- -). Konfidensintervall för linjen då vikten är x_0 =1200 kg är markerat (–).

□

4.4 Prediktionsintervall för observationer

Intervallet ovan gäller väntevärdet för Y då $x = x_0$. Om man vill uttala sig om *en* framtida observation av Y för $x = x_0$ blir ovanstående intervall i regel för smalt. Om α, β och σ vore kända så skulle intervallet $\alpha + \beta x_0 \pm \lambda_{a/2}\sigma$ täcka en framtida observation Y med sannolikhet $1 - a$.

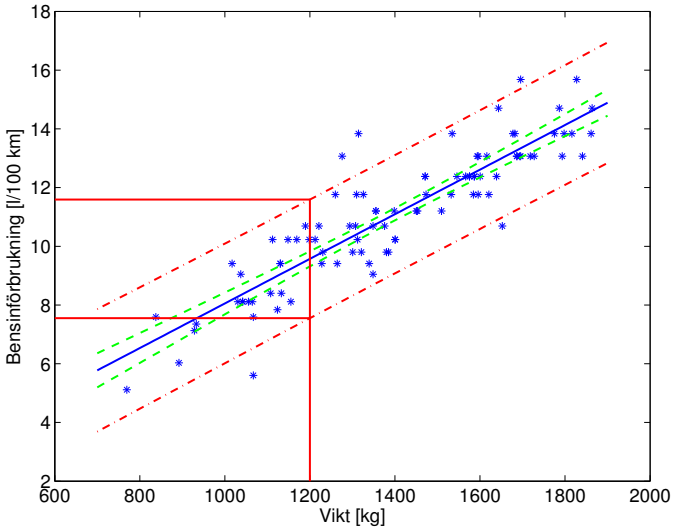
Eftersom regressionslinjen skattas med $\mu_0^* = \alpha^* + \beta^* x_0$ kan vi få hur mycket en framtida observation Y_0 varierar kring den skattade linjen som

$$V(Y_0 - \alpha^* - \beta^* x_0) = V(Y_0) + V(\alpha^* + \beta^* x_0) = \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right).$$

Vi kan alltså få ett *prediktionsintervall* med prediktionsgraden $1-p$ för en framtida observation som

$$I_{Y(x_0)} = \alpha^* + \beta^* x_0 \pm t_{p/2}(n-2)s\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}}.$$

Observera att det bara är första ettan under kvadratroten som skiljer mellan prediktionsintervallet och I_{μ_0}.



Figur 4.4: Bensinförbrukning enligt exempel 1.1. Skattad regressionslinje (–), konfidensintervall för linjen som funktion av vikt (- -), prediktionsintervall för framtida observationer som funktion av vikt (-.). Prediktionsintervall för en framtida observation då vikten är x_0 =1 200 kg är markerat (–).

Ett prediktionsintervall för bensinförbrukningen hos en 1200 kg bil enligt exempel 1.1 blir $[7.6, 11.6]$ vilket är betydligt bredare än intervallet för väntevärdet. I figur 4.4 ses detta intervall och prediktionsintervallen som funktion av x_0.

4.5 Kalibreringsintervall

Om man observerat ett värde y_0 på y, vad blir då x_0? Man kan lösa ut x_0 ur $y_0 = \alpha^* + \beta^* x_0$ och får

$$x_0^* = \frac{y_0 - \alpha^*}{\beta^*}$$

Denna skattning är inte normalfördelad, men vi kan t.ex använda Gauss approximationsformler för att få en skattninga av $d(x_0^*)$ och konstruera ett approximativt intervall

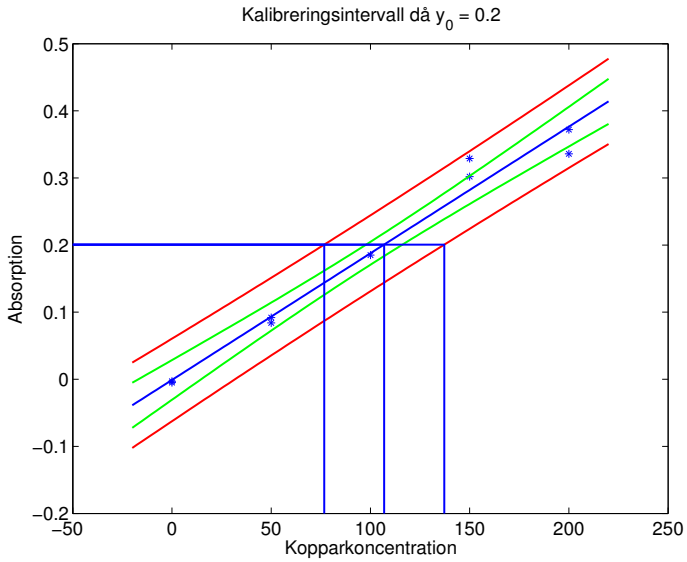
$$I_{x_0} = x_0^* \pm t_{a/2}(n-2)d(x_0^*) = \frac{y_0-\alpha^*}{\beta^*} \pm t_{a/2}(n-2)\cdot\frac{s}{|\beta^*|}\sqrt{1+\frac{1}{n}+\frac{(y_0-\bar{y})^2}{(\beta^*)^2 S_{xx}}}.$$

Ett annat sätt att konstruera kalibreringsintervallet är att dra en linje $y = y_0$ och ta skärningspunkterna med prediktionsintervallet som gränser i kalibreringsintervallet. Ett analytiskt uttryck för detta blir efter lite arbete

$$I_{x_0} = \bar{x} + \frac{\beta^*(y_0-\bar{y})}{c} \pm \frac{t_{p/2}(n-2)\cdot s}{c}\sqrt{c(1+\frac{1}{n})+\frac{(y_0-\bar{y})^2}{S_{xx}}}$$

$$c = (\beta^*)^2 - \frac{(t_{p/2}(n-2)\cdot s)^2}{S_{xx}}.$$

Uttrycket gäller då β är signifikant skild från noll (se kapitel 4.6) annars är det inte säkert att linjen skär prediktionsintervallen. Grafiskt konstrueras detta intervall enligt figur 4.5.



Figur 4.5: Kalibreringsintervall konstruerat som skärning med prediktionsintervall. I försöket har man för ett par prover med kända kopparkoncentrationer mätt absorption med atomabsorptionsspektrofotometri. Kalibreringsintervallet täcker med ungefär 95% sannolikhet den rätta kopparkoncentrationen för ett prov med okänd kopparhalt där absorptionen uppmätts till 0.2.

4.6 Modellvalidering

4.6.1 Residualanalys

Modellen vi använder baseras på att avvikelserna från regressionslinjen är likafördelade ($\varepsilon_i \in \mathrm{N}(0, \sigma)$) och oberoende av varandra vilket medför att även observationerna Y_i är normalfördelade och oberoende. Dessa antaganden används då vi tar fram fördelningen för skattningarna. För att övertyga sig om att antagandena är rimliga kan det vara bra att studera avvikelserna mellan observerade y-värden och motsvarande punkt på den skattade linjen, de s.k. *residualerna*

$$e_i = y_i - (\alpha^* + \beta^* x_i), \quad i = 1, \ldots, n,$$

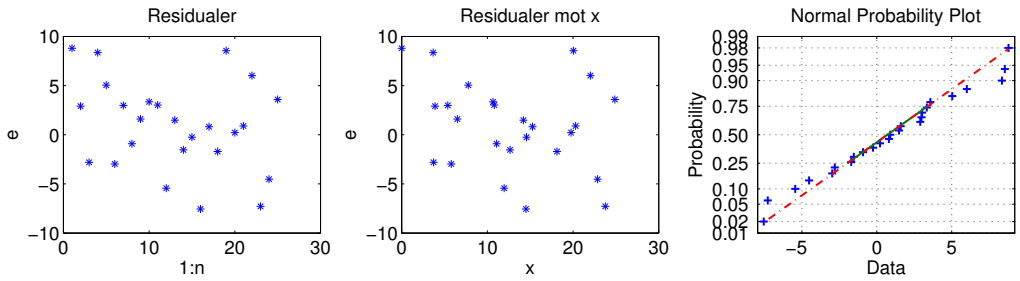
eftersom dessa är observationer av ε_i. Residualerna bör alltså se ut att komma från en och samma normalfördelning samt vara oberoende av dels varandra, samt även av alla x_i. I figur 4.6 visas några exempel på residualplottar som ser bra ut medan de i figur 4.7 ser mindre bra ut.

4.6.2 Är β signifikant?

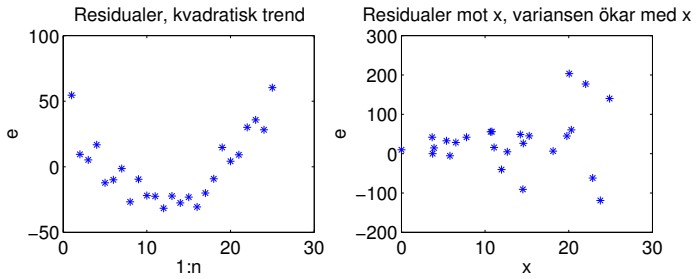
Eftersom β anger hur mycket y beror av x är det även lämpligt att ha med följande hypotestest i en modellvalidering

$$H_0: \ \beta = 0$$
$$H_1: \ \beta \neq 0$$

t.ex. genom att förkasta H_0 om punkten 0 ej täcks av I_β, eller om $|T| > t_{p/2}(n-2)$ där $T = (\beta^* - 0)/d(\beta^*)$. Om H_0 inte kan förkastas har y inget signifikant beroende av x och man kan kanske använda modellen $Y_i = \mu + \varepsilon_i$ i stället.



Figur 4.6: Bra residualplottar. Residualerna plottade i den ordning de kommer, mot x samt i en normalfördelningsplott. De verkar kunna vara oberoende normalfördelade observationer.



Figur 4.7: Residualplottar där man ser en tydlig kvadratisk trend i den vänstra figuren och i den högra ser man att variansen ökar med ökat x

4.7 Linjärisering av några icke linjära samband

Vissa typer av exponential- och potenssamband med multiplikativa fel kan logaritmeras för att få en linjär relation. T.ex. fås när man logaritmerar

$$z_i = a \cdot e^{\beta x_i} \cdot \varepsilon_i' \qquad \xrightarrow{\ln} \qquad \underbrace{\ln z_i}_{y_i} = \underbrace{\ln a}_{\alpha} + \beta \cdot x_i + \underbrace{\ln \varepsilon_i'}_{\varepsilon_i}$$

ett samband på formen $y_i = \alpha + \beta x_i + \varepsilon_i$. Man logaritmerar således z_i-värdena och skattar α och β som vanligt och transformerar till den ursprungliga modellen med $a^* = e^{\alpha^*}$. Observera att de multiplikativa felen ε_i' bör vara lognormalfördelade (dvs $\ln \varepsilon_i' \in N(0, \sigma)$). En annan typ av samband är

$$z_i = a \cdot t_i^{\beta} \cdot \varepsilon_i' \qquad \xrightarrow{\ln} \qquad \underbrace{\ln z_i}_{y_i} = \underbrace{\ln a}_{\alpha} + \beta \underbrace{\ln t_i}_{x_i} + \underbrace{\ln \varepsilon_i'}_{\varepsilon_i}$$

där man får logaritmera både z_i och t_i för att få ett linjärt samband.

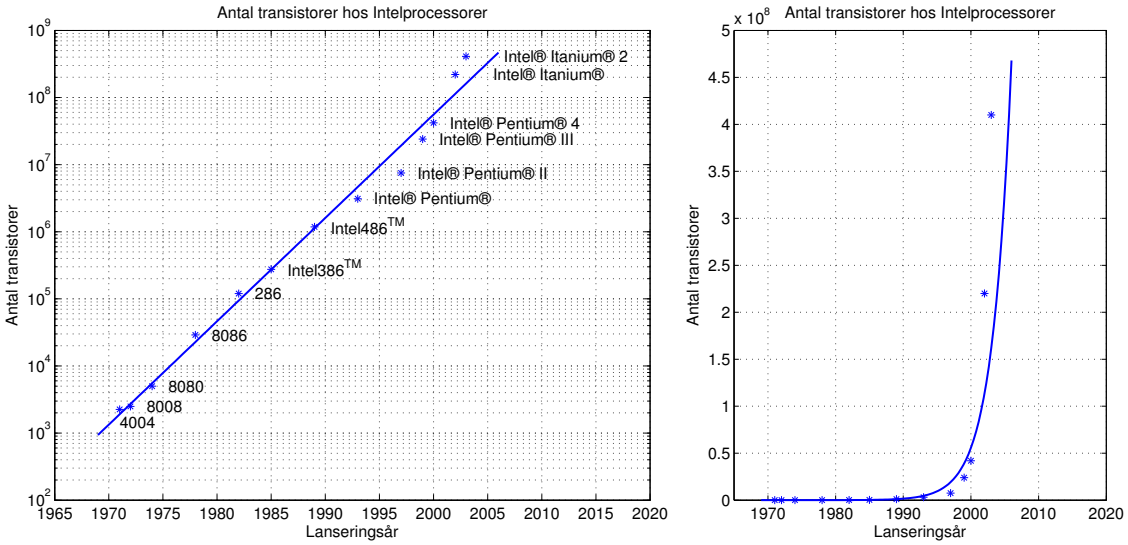
I figur 4.8 ses ett exempel där logaritmering av y-värdena ger ett linjärt samband.

4.8 Centrerad modell

I vissa sammanhang har man tagit fasta på att β^* och $\bar{Y}$ är oberoende av varandra, t.ex I Blom bok B [3] som tidigare användes i kursen, och använt följande parametrisering av regressionslinjen

$$Y_i = \alpha_c + \beta(x_i - \bar{x}) + \varepsilon_i$$

dvs man centrerar x-värdena genom att subtrahera deras medelvärde. I denna framställning blir $\alpha_c^* = \bar{y}$ och β^* skattas på samma sätt som med vår framställning. En fördel med detta är, som vi sett tidigare, att α_c^*



Figur 4.8: Antal transistorer på en cpu mot lanseringsår med logaritmisk y-axel i vänstra figuren. Till höger visas samma sak i linjär skala. Det skattade sambandet är $y = 5.13 \cdot 10^{-301} \cdot e^{0.35x}$.

och β^* blir oberoende av varandra och därmed blir variansberäkningar där de båda är inblandade enklare eftersom man inte behöver ta med någon kovariansterm.

En annan fördel med denna framställning är att man ofta får beräkningar som är mindre känsliga för numeriska problem, men detta är i regel ändå inget problem om man använder programpaket som är designade med numeriska aspekter i åtanke. Vi ska senare se (i kapitel 6) att man t.ex. kan använda operatorn \ i Matlab för att skatta regressionsparametrarna och då har man ingen nytta av att centrera utan Matlab gör själv ett bättre jobb för att få en numeriskt stabil lösning , se t.ex [2].

Anm. I figur 4.8 kan man inte räkna ut regressionslinjen för årtal efter 2002 i Matlab med den formel som står i figurtexten ($e^{0.35 \cdot 2002}$ är för stort för att kunna representeras med flyttalsaritmetik med dubbel precision). Man kan då t.ex. använda en centrerad modell, eller låta x vara t.ex antal år efter år 1900.

5 Stokastiska vektorer

I nästa avsnitt skall vi bygga ut vår linjära regressionsmodell genom att låta y vara en funktion av flera variabler. Man får då en mycket rationell framställning genom att använda en modell skriven på matrisform och vi behöver därför först en kort introduktion till begreppen stokastisk vektor, väntevärdesvektor och kovariansmatris tillsammans med lite räkneregler.

En *stokastisk vektor* är en kolonnvektor där elementen är stokastiska variabler (dvs en n-dimensionell stokastisk variabel).

$$\boldsymbol{X} = [X_1, \ldots, X_n]^T.$$

Med väntevärdet av en stokastisk vektor (eller matris) menas att väntevärdena bildas elementvis. *Väntevärdesvektorn* $\boldsymbol{\mu}$ till en stokastisk vektor $\boldsymbol{X}$ blir då

$$\boldsymbol{\mu} = E(\boldsymbol{X}) = [E(X_1), \ldots, E(X_n)]^T = [\mu_1, \ldots, \mu_n]^T.$$

För att hålla reda på varianser för och kovarianser mellan alla element i en stokastisk vektor kan vi definiera

kovariansmatrisen $\boldsymbol{\Sigma}$ för en stokastisk vektor som (jämför med definitionerna av varians och kovarians)

$$\boldsymbol{\Sigma} = V(\boldsymbol{X}) = E[(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})^T] =$$

$$= \begin{bmatrix} E[(X_1-\mu_1)(X_1-\mu_1)] & E[(X_1-\mu_1)(X_2-\mu_2)] & \cdots & E[(X_1-\mu_1)(X_n-\mu_n)] \\ E[(X_2-\mu_2)(X_1-\mu_1)] & E[(X_2-\mu_2)(X_2-\mu_2)] & \cdots & E[(X_2-\mu_2)(X_n-\mu_n)] \\ \vdots & \vdots & \ddots & \vdots \\ E[(X_n-\mu_n)(X_1-\mu_1)] & E[(X_n-\mu_n)(X_2-\mu_2)] & \cdots & E[(X_n-\mu_n)(X_n-\mu_n)] \end{bmatrix} =$$

$$= \begin{bmatrix} V(X_1) & C(X_1,X_2) & \cdots & C(X_1,X_n) \\ C(X_2,X_1) & V(X_2) & \cdots & C(X_2,X_n) \\ \vdots & \vdots & \ddots & \vdots \\ C(X_n,X_1) & C(X_n,X_2) & \cdots & V(X_n) \end{bmatrix}.$$

Vi ser att kovariansmatrisen är symmetrisk, har varianser som diagonalelement, $\boldsymbol{\Sigma}_{ii} = V(X_i)$, samt att den har kovarianser mellan alla olika element i $\boldsymbol{X}$ för övrigt, $\boldsymbol{\Sigma}_{ij} = \boldsymbol{\Sigma}_{ji} = C(X_i, X_j)$.

Om vi bildar en linjär funktion av $\boldsymbol{X}$ som

$$\boldsymbol{Y} = A\boldsymbol{X} + \boldsymbol{b}$$

där A är en deterministisk (dvs icke stokastisk) $r \times n$-matris och $\boldsymbol{b}$ en deterministisk kolonnvektor med r element fås väntevärdesvektorn till $\boldsymbol{Y}$

$$\boldsymbol{\mu}_Y = E(\boldsymbol{Y}) = E(A\boldsymbol{X} + \boldsymbol{b}) = AE(\boldsymbol{X}) + \boldsymbol{b} = A\boldsymbol{\mu}_X + \boldsymbol{b}$$

dvs motsvarande räkneregel som i det endimensionella fallet. Kovariansmatrisen för $\boldsymbol{Y}$ blir

$$\begin{aligned}\boldsymbol{\Sigma}_Y = V(\boldsymbol{Y}) &= E[(\boldsymbol{Y}-\boldsymbol{m}_Y)(\boldsymbol{Y}-\boldsymbol{m}_Y)^T] = E[(A\boldsymbol{X}+\boldsymbol{b}-A\boldsymbol{m}_X-\boldsymbol{b})(A\boldsymbol{X}+\boldsymbol{b}-A\boldsymbol{m}_X-\boldsymbol{b})^T] = \\ &= [(AB)^T = B^TA^T] = E[A(\boldsymbol{X}-m_X)(\boldsymbol{X}-m_X)^TA^T] = AE[(\boldsymbol{X}-m_X)(\boldsymbol{X}-m_X)^T]A^T = \\ &= AV(\boldsymbol{X})A^T = A\boldsymbol{\Sigma}_X A^T.\end{aligned}$$

Jämför med räkneregeln $V(aX+b) = a^2V(X)$ i det endimensionella fallet.

Sammanfattningsvis har vi följande räkneregler för en linjär funktion av $\boldsymbol{X}$

$$\begin{aligned}E(A\boldsymbol{X}+\boldsymbol{b}) &= AE(\boldsymbol{X}) + \boldsymbol{b} \\ V(A\boldsymbol{X}+\boldsymbol{b}) &= AV(\boldsymbol{X})A^T.\end{aligned}$$

Speciellt har vi om alla element i $\boldsymbol{X}$ är normalfördelade (inte nödvändigtvis oberoende av varandra), en s.k. *multivariat normalfördelning*[2], så blir resultatet av en linjär funktion av $\boldsymbol{X}$ också normalfördelat (eftersom en linjär funktion av $\boldsymbol{X}$ består av linjärkombinationer av elementen i $\boldsymbol{X}$). Vi har alltså fått kraftfulla verktyg för att hantera beroende stokastiska variabler.

6 Multipel regression

Antag att y beror linjärt av k st. förklarande variabler $x_1, \ldots, x_k$.

$$y = \alpha + \beta_1 x_1 + \ldots + \beta_k x_k$$

[2] Definitionen av multivariat normalfördelning är egentligen lite striktare än så, se t.ex avsnitt 6.6 i kursboken Blom et al. [1].

Vi gör n observationer av y och har då en y-dataserie och k st. x-dataserier med n värden i varje serie. Framställningen blir lite renare om vi här kallar α för β_0.

Vi har alltså följande linjära regressionsmodell

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \varepsilon_i, \quad i = 1, \ldots, n$$

där $\varepsilon_i \in N(0, \sigma)$ samt oberoende av varandra.

6.1 Matrisformulering

För att få en enhetlig framställning som ser likadan ut oberoende av hur många variabler y beror av visar det sig fördelaktigt att skriva modellen på matrisform. Ovanstående uttryck kan då skrivas som

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

där $\boldsymbol{y}$ och $\boldsymbol{\varepsilon}$ är $n \times 1$-vektorer, $\boldsymbol{\beta}$ en $1 \times (k+1)$-vektor och X en $n \times (k+1)$-matris

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \qquad X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}, \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}, \qquad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Matrisen X har alltså en kolonn med ettor och sedan en kolonn för vardera x-dataserie. $\boldsymbol{\varepsilon}$, och därmed $\boldsymbol{y}$, har en multivariat normalfördelning med väntevärdesvektor $\mathbf{0}$ respektive $X\boldsymbol{\beta}$ samt samma kovariansmatris $\sigma^2 I$, där I är en enhetsmatris.

6.2 MK-skattning av $\boldsymbol{\beta}$

Parametrarna $\beta_0, \ldots, \beta_k$, dvs elementen i $\boldsymbol{\beta}$ kan skattas med minsta–kvadrat–metoden genom att minimera $Q(\boldsymbol{\beta})$ med avseende på elementen i $\boldsymbol{\beta}$.

$$Q(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - E(Y_i))^2 = \sum_{i=1}^{n} (y_i - \beta_0 x_{i0} - \beta_1 x_{i1} - \ldots - \beta_k x_{ik})^2 = (\boldsymbol{y} - X\boldsymbol{\beta})^T (\boldsymbol{y} - X\boldsymbol{\beta}).$$

där vi låter x_{i0} vara ettorna i matrisen X. Derivatan av Q med avseende på ett element β_ℓ i $\boldsymbol{\beta}$ blir då

$$\frac{\partial Q}{\partial \beta_\ell} = -2 \sum_{i=1}^{n} (y_i - \beta_0 x_{i0} - \beta_1 x_{i1} - \ldots - \beta_k x_{ik}) x_{i\ell}.$$

Denna derivata satt till noll för varje element i $\boldsymbol{\beta}$ kan skrivas på matrisform

$$X^T(\boldsymbol{y} - X\beta) = \mathbf{0} \qquad \Longleftrightarrow \qquad X^T\boldsymbol{y} = X^T X\boldsymbol{\beta}.$$

MK-skattningen av elementen i $\boldsymbol{\beta}$ fås som lösningen till detta ekvationssystem, som kallas normalekvationerna, och ges av

$$\boldsymbol{\beta}^* = (X^T X)^{-1} X^T \boldsymbol{y}$$

om matrisen $X^T X$ är inverterbar, dvs $\det(X^T X) \neq 0$.

En väntevärdesriktig skattning av observationernas varians σ^2 ges av

$$s^2 = \frac{Q_0}{n-(k+1)} \quad \text{där } Q_0 = (\boldsymbol{y} - X\boldsymbol{\beta}^*)^T(\boldsymbol{y} - X\boldsymbol{\beta}^*).$$

Q_0 är alltså residualkvadratsumman (minimivärdet på $Q(\boldsymbol{\beta})$) och $k+1$ är antalet skattade parametrar i densamma.

Exempel 6.1. I ett experiment har man ansatt en linjär regressionsmodell där y beror av två variabler x_1 och x_2. Bestäm MK-skattningarna av β_1 och β_2 om

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ 1 & x_{31} & x_{32} \\ 1 & x_{41} & x_{42} \end{bmatrix} = \begin{bmatrix} 1 & 4 & 1 \\ 1 & 4 & 2 \\ 1 & 4 & 3 \\ 1 & 4 & 4 \end{bmatrix} \text{ och } \boldsymbol{y} = \begin{bmatrix} 3.1 \\ 5.0 \\ 4.5 \\ 6.6 \end{bmatrix}.$$

Lsg. MK-skattningarna fås som $\boldsymbol{\beta}^* = (X^TX)^{-1}X^T\boldsymbol{y}$, men i

$$X^TX = \begin{bmatrix} 4 & 16 & 10 \\ 16 & 64 & 40 \\ 10 & 40 & 30 \end{bmatrix}$$

är de två första kolonnerna parallella, dvs $\det(X^TX) = 0$, och normalekvationerna saknar en entydig lösning. Man bör alltså inte mäta y för ett enda värde på x_1 eller x_2 (och inte bara för samma värden på x_1 och x_2) eftersom det resulterar i parallella kolonner i X och därmed i X^TX. □

Exempel 6.2. Regression genom origo. Bestäm MK-skattningen av β i modellen $y_i = \beta x_i + \varepsilon_i$.

Lsg I de regressionsmodeller vi hittills sett har vi haft ett intercept med (α eller β_0) för att inte tvinga regressionslinjen (eller planet) genom origo. I den här modellen skall linjen gå genom origo så vi kan använda matrisformuleringen men utan att ta med någon kolonn med ettor i X-matrisen (som här blir en vektor). Vi har alltså

$$X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad \boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad X^TX = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \sum_{i=1}^{n} x_i^2$$

Skattningen av elementen i $\boldsymbol{\beta}$ (dvs det enda elementet β) blir

$$\beta^* = (X^TX)^{-1}X^T\boldsymbol{y} = (X^TX)^{-1} \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}.$$

□

6.3 Skattningarnas fördelning

Skattningarna av elementen i $\boldsymbol{\beta}$ är linjära funktioner av $\boldsymbol{y}$ och är därmed normalfördelade. För $\boldsymbol{\beta}^*$ fås väntevärdesvektorn enligt räknereglerna i avsnitt 5 till

$$E(\boldsymbol{\beta}^*) = E[(X^TX)^{-1}X^T\boldsymbol{Y}] = (X^TX)^{-1}X^TE(\boldsymbol{Y}) = (X^TX)^{-1}X^TX\boldsymbol{\beta} = \boldsymbol{\beta}$$

och kovariansmatrisen blir

$$\begin{aligned} V(\boldsymbol{\beta}^*) &= V[(X^TX)^{-1}X^T\boldsymbol{y}] = (X^TX)^{-1}X^T \;\; \sigma^2 I \;\; [(X^TX)^{-1}X^T]^T = \\ &= \sigma^2 \underbrace{(X^TX)^{-1}\,X^TX}_{=I}[(X^TX)^{-1}]^T = \sigma^2[(X^TX)^{-1}]^T = [\text{en kovariansmatris är symmetrisk}] = \\ &= \sigma^2(X^TX)^{-1}. \end{aligned}$$

Varianserna för elementen i $\boldsymbol{\beta}^*$ återfinns alltså som diagonalelementen i kovariansmatrisen $\sigma^2(X^TX)^{-1}$ och de övriga elementen är kovarianser

$$\sigma^2(X^TX)^{-1} = \begin{bmatrix} V(\beta_0^*) & C(\beta_0^*,\beta_1^*) & \cdots & C(\beta_0^*,\beta_k^*) \\ C(\beta_1^*,\beta_0^*) & V(\beta_1^*) & \cdots & C(\beta_1^*,\beta_k^*) \\ \vdots & \vdots & \ddots & \vdots \\ C(\beta_k^*,\beta_0^*) & C(\beta_k^*,\beta_1^*) & \cdots & V(\beta_k^*) \end{bmatrix}.$$

Ett konfidensintervall för en parameter β_ℓ blir således

$$I_{\beta_\ell} = \beta_\ell^* \pm t_{\alpha/2}(n-(k+1))d(\beta_\ell^*)$$

där $d(\beta_\ell^*)$ är roten ur motsvarande diagonalelement i den skattade kovariansmatrisen $s^2(X^TX)^{-1}$.

För residualkvadratsumman gäller dessutom

$$\frac{Q_0}{\sigma^2} \in \chi^2(n-(k+1)).$$

Exempel 6.3. I West Virginia har man under ett antal år räknat antalet frostdagar på olika orter. I vektorn `y` finns medelantalet frostdagar per år, i `x1` ortens höjd över havet (ft) och `x2` nordlig breddgrad (°).

x_1	2375	1586	1459	680	604	1298	3242	1426	550	2250	675	2135	635	1649	2727	1053	2424	789	659	673
x_2	39.27	38.63	39	39.17	38.35	39.47	37.58	37.37	39.38	37.8	38.05	38.23	39.65	39.1	38.66	39.48	37.97	38.8	40.1	37.67
y	73	29	28	25	11.5	32.5	64	13	23	37	26	73	24.7	41	56	34	37	16	41	12

Skatta parametrarna i modellen

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$$

samt gör 95% konfidensintervall för var och en av parametrarna.

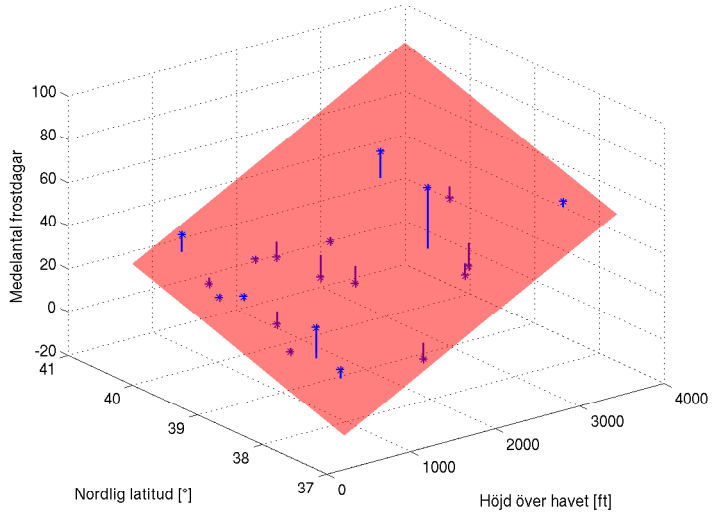
I Matlab kan beräkningarna göras enligt:

```
X = [ones(size(y)) x1 x2];
```

$\boldsymbol{\beta}$ skattas med `inv(X'*X)*X'*y` men det är i regel dumt att räkna ut en invers för att använda till att lösa ett ekvationssystem. I Matlab kan man i stället lösa (det överbestämda) ekvationssystemet i minsta-kvadratmening med operatorn \

```
beta = X\y
beta =
 -399.6582
    0.0212
   10.4411
```

Vi ser att antalet frostdagar ökar i genomsnitt med $\beta_1^* = 0.02$ dagar då höjden över havet ökas en fot och med $\beta_2^* = 10.4$ dagar då breddgraden ökas en enhet. En plot över det skattade regressionsplanet kan ses i figur 6.1.



Figur 6.1: En plot över skattat regressionsplan i exempel 6.3. Från observationerna har dragits en lodrät linje till det skattade regressionsplanet (residualerna).

Residualkvadratsumman Q_0 fås ur

```
Q0 = (y-X*beta)'*(y-X*beta)
Q0 =
   1.7798e+03
```

och med hjälp av en skattning av kovariansmatrisen, `V`, kan man göra konfidensintervall för parametrarna β_i^*

```
n = length(y);
s2 = Q0/(n-3);
V = s2*inv(X'*X)
V =

    1.66e4   -0.1722 -424.9661
   -0.1722    9.5e-6    0.0041
 -424.9661    0.0041   10.8320
```

För t.ex β_1 blir konfidensintervallet

$$I_{\beta_1} = \beta_1^* \pm t_{p/2}(n-3)d(\beta_1^*)$$

som i Matlab kan räknas ut som (β_1^* är element 2 i vektorn `beta`)

```
kvantil = tinv(1-0.05/2, n-3);
d = sqrt(V(2,2));
Ib1 = beta(2) + [-1 1] * kvantil * d
Ib1 =
    0.0146    0.0277
```

och övriga intervall:

```
Ib2 = beta(3) + [-1 1] * kvantil * sqrt(C(3,3))
Ib2 =
    3.4972   17.3849
Ib0 = beta(1) + [-1 1] * kvantil * sqrt(C(1,1))
Ib0 =
    -672.2605 -127.0559
```

□

6.4 Skattning av punkt på ”planet”

För att skatta Y-s väntevärde i en punkt $x_0 = (x_{01}, x_{02}, \ldots, x_{0k})$ kan vi bilda radvektorn $\boldsymbol{x}_0 = [1\ x_{01}\ x_{02} \ldots x_{0k}]$. Punkskattningen blir

$$\mu^*(\boldsymbol{x}_0) = \boldsymbol{x}_0\boldsymbol{\beta}^*$$

som är normalfördelad och dess varians, enligt räknereglerna för kovariansmatris, blir

$$V(\mu^*(\boldsymbol{x}_0)) = V(\boldsymbol{x}_0\boldsymbol{\beta}^*) = \boldsymbol{x}_0 V(\boldsymbol{\beta}^*)\boldsymbol{x}_0^T = \sigma^2\boldsymbol{x}_0(X^TX)^{-1}\boldsymbol{x}_0^T.$$

Observera att vi här har tagit hänsyn till att elementen i $\boldsymbol{\beta}$ *inte* är oberoende av varandra, kovarianserna mellan dem ingår ju i kovariansmatrisen vi räknar med.

Ett konfidensintervall för $\mu(\boldsymbol{x}_0)$ blir således

$$I_{\mu(\boldsymbol{x}_0)} = \mu^*(\boldsymbol{x}_0) \pm t_{\alpha/2}(n-(k+1))s\sqrt{\boldsymbol{x}_0(X^TX)^{-1}\boldsymbol{x}_0^T}.$$

Vill man i stället göra ett prediktionsintervall får man som tidigare lägga till en etta under kvadratroten.

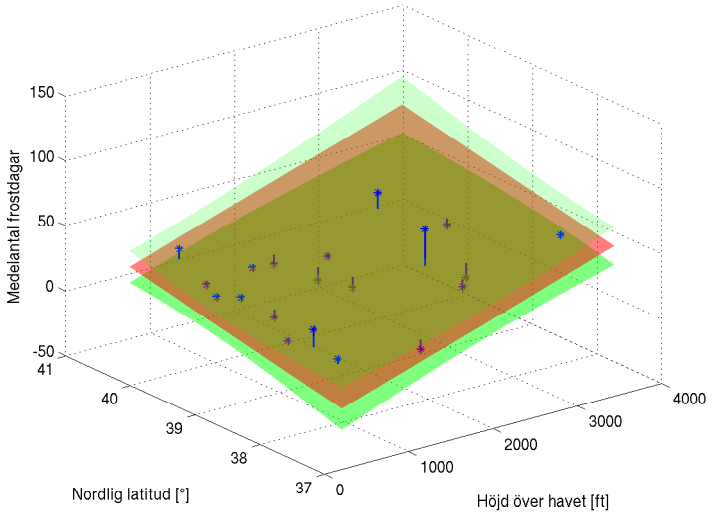
Exempel 6.4. (forts ex. 6.3) Gör ett konfidensintervall för medelantalet frostdagar på en höjd av 3000 ft och 39° nordlig breddgrad.

Lsg. I Matlab blir beräkningarna

```
x0 = [1 3000 39];
mu0 = x0*beta
mu0 =
   71.0234
Vmu0 = x0 * V * x0'
Vmu0 =
   33.4553
dmu0 = sqrt(Vmu0)
dmu0 =
```

```
    5.7841
Imu0 = mu0 + [-1 1] * kvantil * dmu0
Imu0 =
   58.8201   83.2266
```

En plot över konfidensintervallen som funktion av x_1 och x_2 kan ses i figur 6.2.



Figur 6.2: Konfidensintervall plottade som funktion av x_1 och x_2 i exempel 6.4.

□

6.5 Modellvalidering

För att övertyga sig om att modellen är rimlig bör man liksom tidigare förvissa sig om att residualerna verkar vara oberoende observationer av $N(0, \sigma)$. Plotta residualerna

- ”Som de kommer”, dvs mot $1, 2, \ldots, n$. Ev. ett histogram
- Mot var och en av x_i-dataserierna
- I en normalfördelningsplot

För var och en av $\beta_1, \ldots, \beta_k$ (obs i regel ej β_0) bör man kunna förkasta H_0 i testet

$$H_0 : \ \beta_i = 0$$
$$H_1 : \ \beta_i \neq 0$$

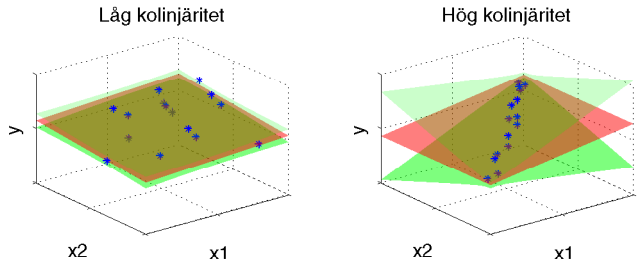
eftersom β_i anger ”hur mycket y beror av variabeln x_i”.

I exempel 6.3 kan vi se att β_1 och β_2 båda är signifikant skilda från noll, varken I_{β_1} eller I_{β_2} täckte punkten noll.

Anm. För att testa om alla parametrar i modellen är signifikanta bör man göra ett simultant test H_0 : alla $\beta_i = 0$ mot H_1 : något $\beta_i \neq 0$. Detta kan utföras med ett F-test men det ligger utanför ramen för denna kurs.

6.6 Kolinjäritet mellan förklarande variabler

I exempel 6.2 såg vi att man inte kan välja värdena på de förklarande variablerna hur som helst. T.ex. om man väljer samma värden på alla x-variabler så blir inte X^TX inverterbar. För att kunna få en skattning av t.ex. ett regressionsplan ”stabil” bör man om möjligt välja sina (x_{1i}, x_{2i})-värden så att de blir utspridda i (x_1, x_2)-planet och inte klumpar ihop sig längs en linje. Detta ger ”en mer stabil grund” åt regressionsplanet. Se figur 6.3.



Figur 6.3: I vänstra figuren är värdena på x_1 och x_2 valda så att de har låg korrelation mellan varandra och ger en stabil grund för regressionsplanet. I högra figuren är korrelationen hög och regressionsplanet ”får en sämre grund att stå på”, dvs osäkerheten blir stor i vissa riktningar. Konfidensplanen är inritade i figuren.

6.7 Stegvis regression

Om inte alla β_i är signifikant skilda från noll bör man reducera sin modell, dvs ta bort en eller flera x-variabler, skatta parametrarna i den reducerade modellen och eventuellt upprepa förfarandet. Vilka variabler skall man då ta bort?

- x-variabler med hög kolinjäritet (korrelation) bör inte båda vara med i modellen.
- x-variabler med hög korrelation med Y är bra att ha med.

Har man sedan flera signifikanta modeller att välja mellan kan man beakta saker som

- Litet s, dvs residualerna avviker lite från skattat ”plan”.
- Med få variabler blir modellen enklare att hantera, men man bör ha tillräckligt många för att beskriva y väl.

6.8 Polynomregression

Med matrisframställningen kan man även enkelt hantera vissa situationer där y inte beror linjärt av en variabel x utan beskrivs av t.ex ett polynom

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \ldots + \beta_k x_i^k + \varepsilon_i$$

Matrisen X blir

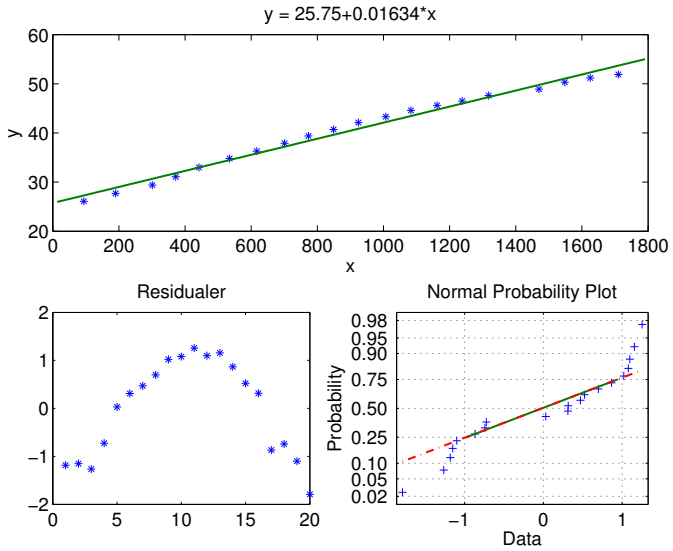
$$X = \begin{pmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^k \\ 1 & x_2 & x_2^2 & \cdots & x_2^k \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^k \end{pmatrix}$$

Skattningar av parametrar blir på samma sätt som tidigare.

Exempel 6.5. I en fysiklaboration i kretsprocesser uppmättes följande där $x =$ ”tid i sekunder” och $y =$ ”temperatur i $^\circ C$” i en värmepump.

x_i: 94 , 190, 301, 372, 442, 535, 617, 701, 773, 849, 924, 1007, 1083, 1162, 1238, 1318, 1470, 1548, 1625, 1710
y_i: 26.1, 27.7, 29.4, 31.1, 33 , 34.8, 36.3, 37.9, 39.4, 40.7, 42.1, 43.3, 44.6, 45.6, 46.5, 47.6, 48.9, 50.3, 51.2, 51.9

I figur 6.4 ser man att det inte passar så bra med en enkel linjär regressionsmodell $Y_i = \alpha + \beta x_i + \varepsilon_i$.

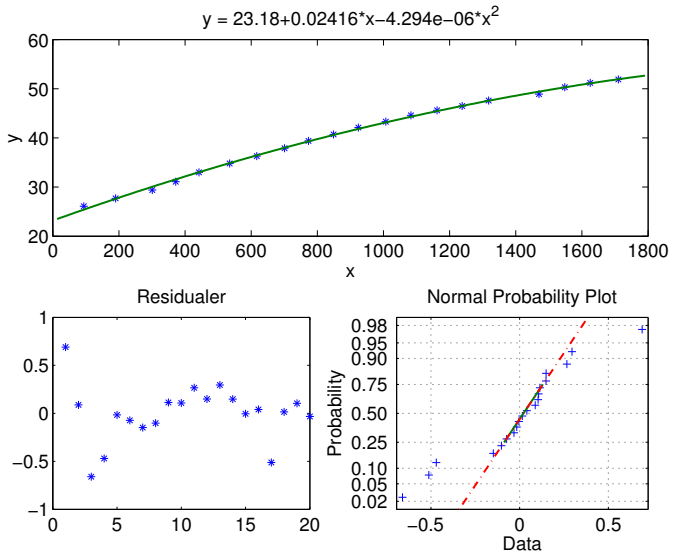


Figur 6.4: Data från kretsprocesslaborationen anpassat till en förstagradsmodell. Residualplotten visar tydligt att modellen inte är lämplig.

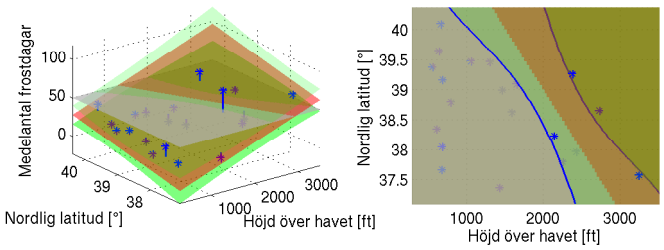
Om man däremot ansätter en andragradsmodell, $Y_i = \alpha + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$, passar data bättre till modellen. Se figur 6.5. □

6.9 Kalibreringsområde

Motsvarigheten till kalibreringsintervall blir i regel ganska besvärligt att hantera analytiskt då man har en funktion av flera variabler. Men med inspiration av metoden med skärningen av prediktionsintervallen i avsnitt 4.5 kan man ganska enkelt göra kalibreringområden då y är en linjär funktion av två variabler. Man plottar in planet $y = y_0$ och tar skärningarna med prediktionsplanen som kalibreringsområde. I figur 6.6 visas det område där man i genomsnitt har 50 frostdagar.



Figur 6.5: Data från kretsprocesslaborationen anpassat till en andragradsmodell. Residualplotten ser betydligt bättre ut även om de kanske inte riktigt är normalfördelade; de tre minsta residualerna är lite för små och den största lite för stor. Parametrarna β_1 och β_2 är signifikanta.



Figur 6.6: I vänstra figuren är regressionsplanet från exempel 6.3 plottat tillsammans med prediktionsplanen och planet $y = 50$. I högra figuren är samma plott sedd ovanifrån och kalibreringsområdet syns som skärningen mellan planet $y = 50$ och prediktionplanen.

A ML- och MK skattningar av parametrarna i enkel linjär regression

A.1 Några hjälpresultat

Vi börjar med ett par användbara beteckningar och räkneregler för de summor och kvadratsummor som kommer att ingå i skattningarna. Då alla summor nedan löper från 1 till n avstår jag från att skriva ut summationsindexen.

Först har vi att en ren summa av avvikelser av ett antal observationer kring sitt medelvärde är noll

$$\sum(x_i - \bar{x}) = \sum x_i - n\bar{x} = [\bar{x} = \frac{1}{n}\sum x_i] = \sum x_i - \sum x_i = 0 \tag{A.1}$$

Några beteckningar för kvadratiska- och korsavvikelser kring medelvärde

$$S_{xx} = \sum(x_i - \bar{x})^2, \quad S_{xy} = \sum(x_i - \bar{x})(y_i - \bar{y}), \quad S_{yy} = \sum(y_i - \bar{y})^2$$

där vi känner igen den första och sista från stickprovsvarianserna för x resp. y, $s_x^2 = S_{xx}/(n-1)$ och motsvarande för y. Dessa summor kan skrivas på ett antal former, t.ex kan S_{xy} utvecklas till

$$S_{xy} = \sum(x_i - \bar{x})(y_i - \bar{y}) = \sum x_i(y_i - \bar{y}) - \bar{x}\sum(y_i - \bar{y}) = \sum x_i(y_i - \bar{y}) \quad \text{eller}$$
$$S_{xy} = \sum(x_i - \bar{x})(y_i - \bar{y}) = \sum(x_i - \bar{x})y_i - \bar{y}\sum(x_i - \bar{x}) = \sum(x_i - \bar{x})y_i$$

där sista summan i andra leden blir noll enligt (A.1). Motsvarande räkneregler gäller för S_{xx} och S_{yy} och vi har sammanfattningsvis

$$S_{xy} = \sum(x_i - \bar{x})(y_i - \bar{y}) = \sum x_i(y_i - \bar{y}) = \sum(x_i - \bar{x})y_i \tag{A.2}$$
$$S_{xx} = \sum(x_i - \bar{x})^2 = \sum x_i(x_i - \bar{x}) \quad \text{och motsvarande för } S_{yy} \tag{A.3}$$

A.2 Punktskattningar

ML-skattning av α, β och σ^2 då y_i är oberoende observationer av $Y_i \in N(\alpha + \beta x_i, \sigma)$ fås genom att maximera likelihood-funktionen

$$L(\alpha, \beta, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_1 - \alpha - \beta x_1)^2}{2\sigma^2}} \cdot \ldots \cdot \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_n - \alpha - \beta x_n)^2}{2\sigma^2}} = (2\pi)^{n/2} \cdot (\sigma^2)^{n/2} \cdot e^{-\frac{1}{2\sigma^2}\sum(y_i - \alpha - \beta x_i)^2}$$

Hur än σ väljs så kommer L att maximeras med avseende på α och β då $\sum(y_i - \alpha - \beta x_i)^2$ är minimal, och eftersom det är just denna kvadratsumma som minimeras med MK-metoden så blir skattningarna av α och β de samma vid de två metoderna. Med ML-metoden kan vi dessutom skatta σ^2 varför vi väljer den. Logaritmeras likelihoodfunktionen fås

$$\ln L(\alpha, \beta, \sigma^2) = \frac{n}{2}\ln(2\pi) + \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum(y_i - \alpha - \beta x_i)^2$$

Deriveras denna med avseende på var och en av parametrarna och sedan sättes till noll fås ekvationssystemet

$$\frac{\partial \ln L}{\partial \alpha} = \frac{1}{\sigma^2}\sum(y_i - \alpha - \beta x_i) = 0 \tag{A.4}$$
$$\frac{\partial \ln L}{\partial \beta} = \frac{1}{\sigma^2}\sum(y_i - \alpha - \beta x_i)x_i = 0 \tag{A.5}$$
$$\frac{\partial \ln L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum(y_i - \alpha - \beta x_i)^2 = 0 \tag{A.6}$$

att lösa med avseende på α, β och σ^2. Eftersom vi kan förlänga de två första ekvationerna med σ^2 och därmed bli av med den kan vi använda dessa till att skatta α och β. (A.4) och (A.5) kan formas om till

$$\begin{aligned}\sum y_i &= n\alpha + \beta \sum x_i \\ \sum x_i y_i &= \alpha \sum x_i + \beta \sum x_i^2\end{aligned} \tag{A.7}$$

Delas första ekvationen med n fås

$$\bar{y} = \alpha + \beta\bar{x} \quad \Longleftrightarrow \quad \alpha = \bar{y} - \beta\bar{x} \tag{A.8}$$

som vi kan stoppa in i (A.7) som då blir

$$\begin{aligned}&\sum x_i y_i = \bar{y}\sum x_i - \beta\bar{x}\sum x_i + \beta\sum x_i^2 \quad \Longleftrightarrow \\ &\sum x_i y_i = \beta(\sum x_i^2 - \bar{x}\sum x_i) + \bar{y}\sum x_i \quad \Longleftrightarrow \\ &\beta = \frac{\sum x_i y_i - \bar{y}\sum x_i}{\sum x_j^2 - \bar{x}\sum x_j} = \frac{\sum x_i(y_i - \bar{y})}{\sum x_j(x_j - \bar{x})} = [\text{(A.2)}] = \frac{\sum (x_i - \bar{x})y_i}{\sum x_j(x_j - \bar{x})} = [\text{(A.2) och (A.3)}] = \frac{S_{xy}}{S_{xx}}\end{aligned} \tag{A.9}$$

Detta resultat tillsammans med (A.8) ger ML-skattningarna av α och β

$$\beta^* = \frac{S_{xy}}{S_{xx}}, \quad \alpha^* = \bar{y} - \beta^*\bar{x}$$

Dessa värden insatta i (A.6) förlängd med σ^4 ger

$$(\sigma^2)^* = \frac{1}{n}\sum (y_i - \alpha^* - \beta^* x_i)^2$$

som dock inte är väntevärdesriktig utan korrigeras till

$$(\sigma^2)^* = s^2 = \frac{1}{n-2}\sum (y_i - \alpha^* - \beta^* x_i)^2 = \frac{Q_0}{n-2}$$

som är det. Q_0 som är summan av kvadratiska avvikelser från observationerna y_i till motsvarande punkt på den skattade linjen kallas *residualkvadratsumma* och den kan skrivas på formen

$$Q_0 = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$$

A.3 Skattningarnas fördelning

Om vi börjar med β^* och utgår från (A.9)

$$\beta^* = \frac{S_{xy}}{S_{xx}} = \frac{\sum (x_i - \bar{x})y_i}{\sum x_j(x_j - \bar{x})} = \sum c_i y_i \quad \text{där} \quad c_i = \frac{x_i - \bar{x}}{S_{xx}} \tag{A.10}$$

den är alltså en linjär funktion av de normalfördelade observationerna och därmed är skattningen normalfördelad. Väntevärdet blir

$$\begin{aligned}E(\beta^*) = E(\sum c_i Y_i) &= \sum c_i E(Y_i) = \sum c_i(\alpha + \beta x_i) = \frac{1}{S_{xx}}\sum (x_i - \bar{x})(\alpha + \beta x_i) \\ &= \frac{\alpha}{S_{xx}}\sum (x_i - \bar{x}) + \frac{\beta}{S_{xx}}\sum (x_i - \bar{x})x_i = 0 + \beta\frac{S_{xx}}{S_{xx}} = \beta\end{aligned}$$

där vi i näst sista ledet åter använde hjälpresultaten (A.2) och (A.3). Skattningen är alltså väntevärdesriktig och dess varians blir

$$V(\beta^*) = V(\sum c_i Y_i) = \sum c_i^2 V(Y_i) = \sum c_i^2 \sigma^2 = \frac{\sigma^2}{S_{xx}^2} \sum (x_i - \bar{x})^2 = \frac{\sigma^2}{S_{xx}}$$

dvs

$$\beta^* = \frac{S_{xy}}{S_{xx}} \quad \text{är en observation av} \quad \beta^* \in N(\beta, \frac{\sigma}{\sqrt{S_{xx}}})$$

$\alpha^* = \bar{y} - \beta^* \bar{x}$ är även den normalfördelad eftersom den är en linjär funkton av normalfördelningar. Väntevärdet blir

$$E(\alpha^*) = E(\bar{Y}) - \bar{x}E(\beta^*) = E(\frac{1}{n}\sum Y_i) - \bar{x}\beta = \frac{1}{n}\sum(\alpha + \beta x_i) - \bar{x}\beta =$$
$$= \frac{1}{n}\sum \alpha + \frac{\beta}{n}\sum x_i - \bar{x}\beta = \alpha + \beta\bar{x} - \bar{x}\beta = \alpha$$

så även α^* är väntevärdesriktig. Innan vi beräknar dess varians har vi nytta av att $\bar{Y}$ och β^* är oberoende av varandra. Vi visar här att de är okorrelerade, vilket räcker för variansberäkningen. Återigen visar det sig fördelaktigt att uttrycka β^* enligt (A.10)

$$C(\bar{Y}, \beta^*) = C(\frac{1}{n}\sum Y_i, \sum c_j Y_j) = \frac{1}{n}\sum_i \sum_j c_j C(Y_i, Y_j) = [Y_i \text{ är ober. av } Y_j \text{ då } i \neq j] =$$
$$= \frac{1}{n}\sum c_i C(Y_i, Y_i) = \frac{1}{n}\sum c_i V(Y_i) = \frac{\sigma^2}{n}\sum c_i = \frac{\sigma^2}{nS_{xx}}\sum(x_i - \bar{x}) = 0$$

där vi återigen känner igen (A.1) i sista steget. Variansen för α^* blir

$$V(\alpha^*) = V(\bar{Y} - \beta^*\bar{x}) = V(\bar{Y}) + \bar{x}^2 V(\beta^*) - 2\bar{x}C(\bar{Y}, \beta^*) = \frac{\sigma^2}{n} + \bar{x}^2\frac{\sigma^2}{S_{xx}} + 0$$

dvs

$$\alpha^* = \bar{y} - \beta^*\bar{x} \quad \text{är en observation av} \quad \alpha^* \in N(\alpha, \sigma\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}})$$

α^* och β^* är dock inte oberoende av varandra. Kovariansen mellan dem är

$$C(\alpha^*, \beta^*) = C(\bar{Y} - \beta^*\bar{x}, \beta^*) = C(\bar{Y}, \beta^*) - \bar{x}C(\beta^*, \beta^*) = 0 - \bar{x}V(\beta^*) = -\bar{x}\frac{\sigma^2}{S_{xx}}.$$

För variansskattningen och residualkvadratsumman gäller

$$(\sigma^2)^* = s = \frac{1}{n-2}\sum(y_i - \alpha^* - \beta^* x_i)^2 = \frac{Q_0}{f}, \qquad \frac{Q_0}{\sigma^2} \in \chi^2(f)$$

Referenser

[1] Gunnar Blom, Jan Enger, Gunnar Englund, Jan Grandell och Lars Holst. *Sannolikhetsteori och statistikteori med tillämpningar*. Studentlitteratur, Lund, 2005.

[2] The Math Works, Inc., Natick, Mass. *MATLAB. Reference Guide*, 1993.

[3] Gunnar Blom och Björn Holmquist. *Statistikteori med tillämpningar, bok B*. Studentlitteratur, Lund, 1998.