Föreläsning 13, Matematisk statistik 7.5 hp för E
Multipel linjär regression

Stas Volkov

Linjär regression

Vi har n st par av mätvärden (x_ℓ, y_ℓ), $\ell = 1, \ldots, n$ där y_ℓ är observationer av

$$Y_\ell = \alpha + \beta x_\ell + \varepsilon_\ell$$

där ε_ℓ är oberoende av varandra, och $\varepsilon_\ell \in \mathsf{N}(0, \sigma)$.

Parameterskattningarna

Skattningarna av α^*, β^* och $(\sigma^2)^*$ är

$$\alpha^* = \bar{y} - \beta^* \cdot \bar{x}, \qquad \beta^* = \frac{\sum_{\ell=1}^{n}(x_\ell - \bar{x})(y_\ell - \bar{y})}{\sum_{\ell=1}^{n}(x_\ell - \bar{x})^2} = \frac{S_{xy}}{S_{xx}},$$

$$(\sigma^2)^* = s^2 = \frac{Q_0}{n-2}, \quad Q_0 = \sum_{\ell=1}^{n}(y_\ell - \alpha^* - \beta^* x_\ell)^2 = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$$

Skattningarnas fördelning:

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

Men de är inte oberoende av varandra.
Konfidens-, prediktions- och kalibrerings-intervall ($f = n - 2$):

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

$$I_\beta = \beta^* \pm t_{a/2}(f) \cdot \frac{s}{\sqrt{S_{xx}}},$$

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

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

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

Linjärisering av exponentiella samband (vid behov)

För att få ett linjärt samband

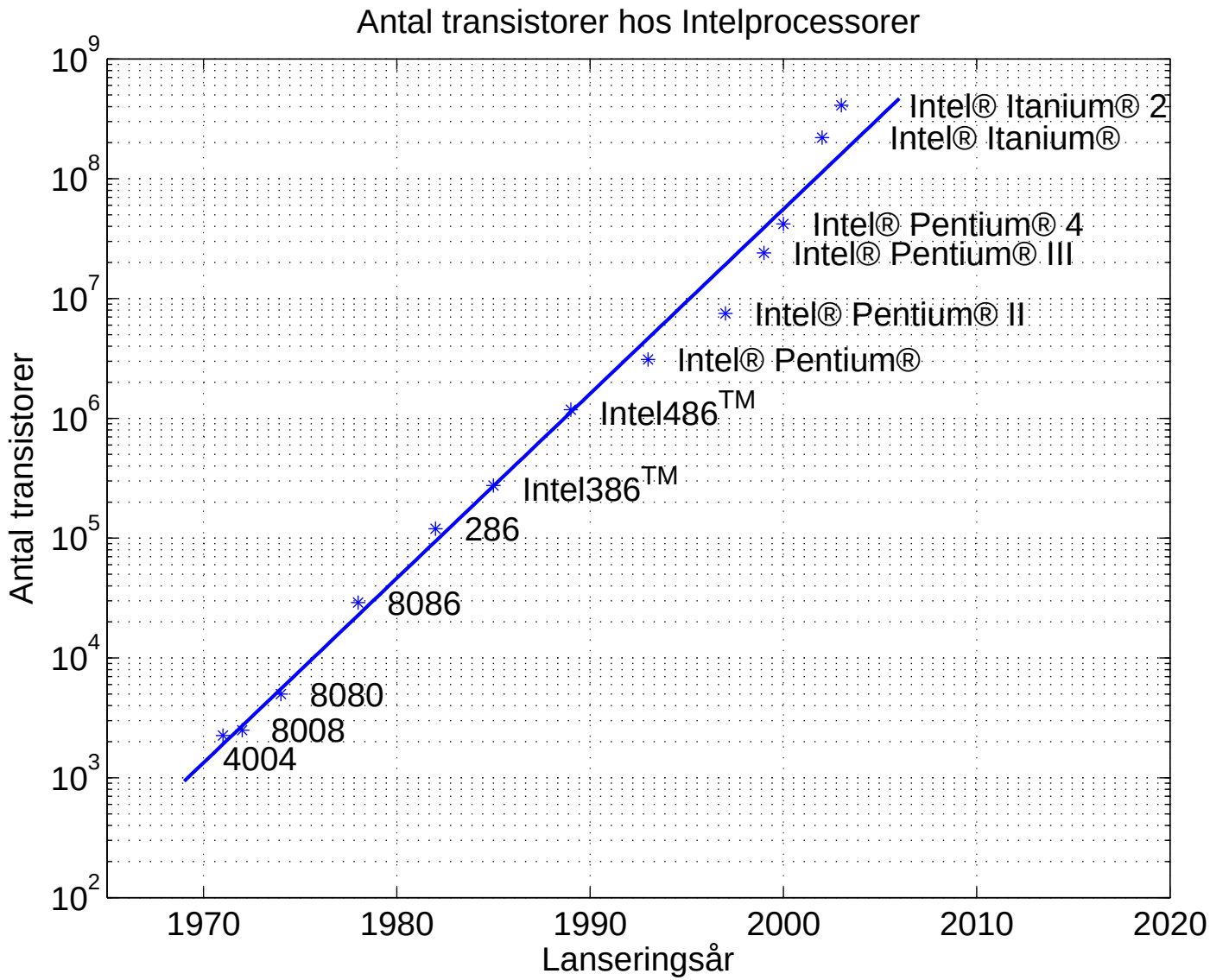
$$y_\ell = \alpha + \beta x_\ell + \varepsilon_\ell$$

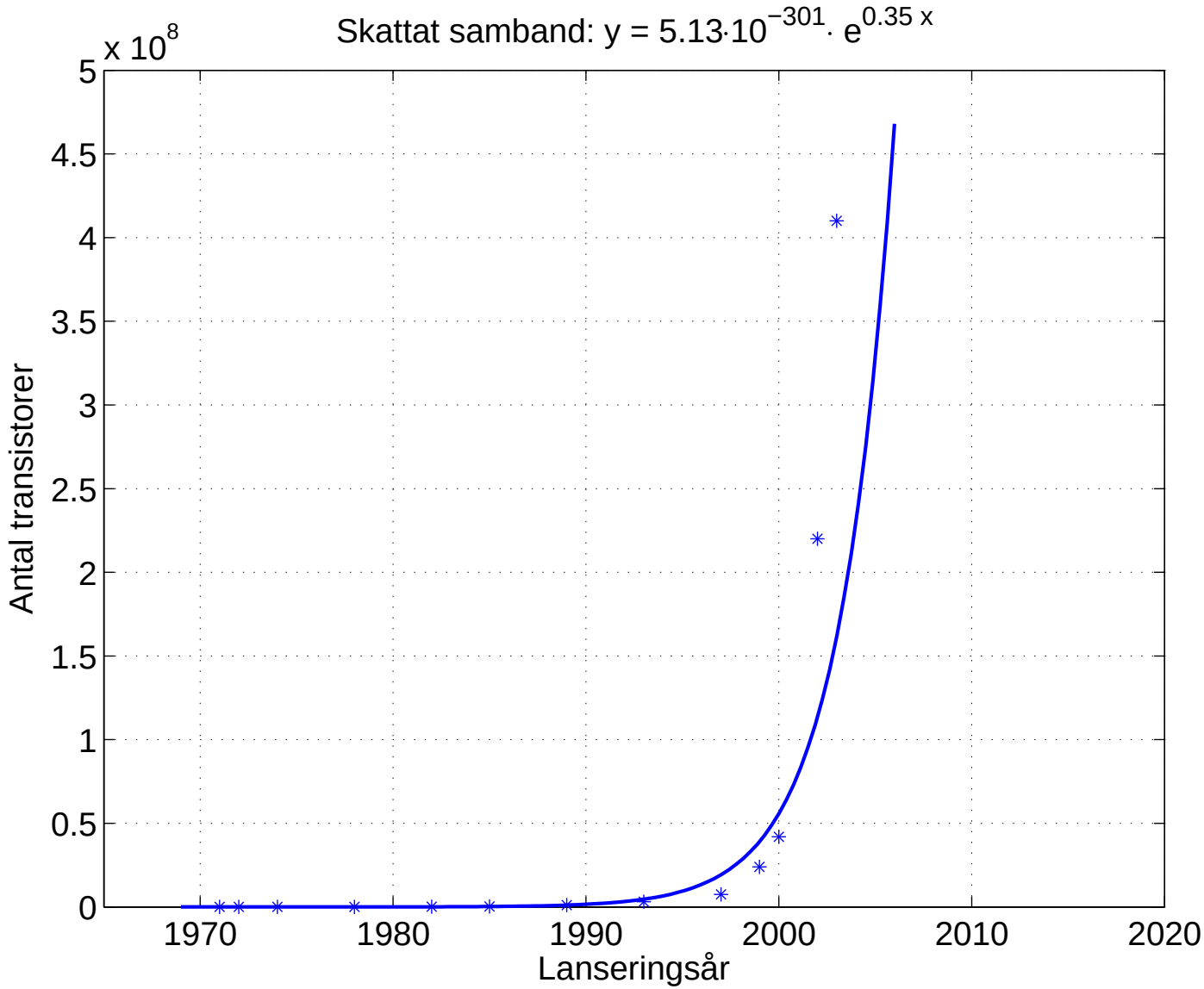
kan vissa exponent- och potenssamband logaritmeras.

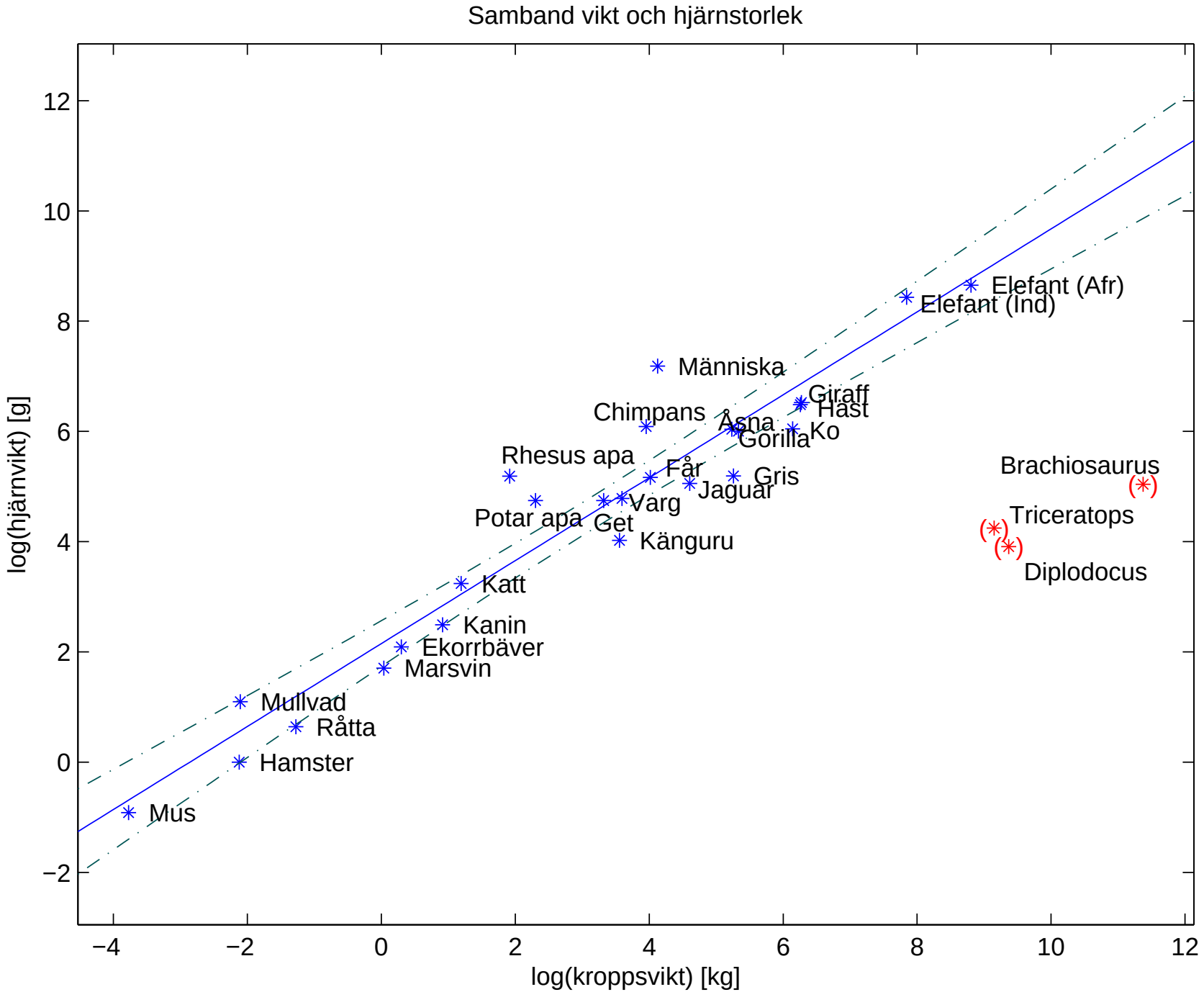
$$z_\ell = a \cdot e^{\beta x_\ell} \cdot \varepsilon'_\ell \quad \xrightarrow{\ln} \quad \underbrace{\ln z_\ell}_{y_\ell} = \underbrace{\ln a}_{\alpha} + \beta \cdot x_\ell + \underbrace{\ln \varepsilon'_\ell}_{\varepsilon_\ell}$$

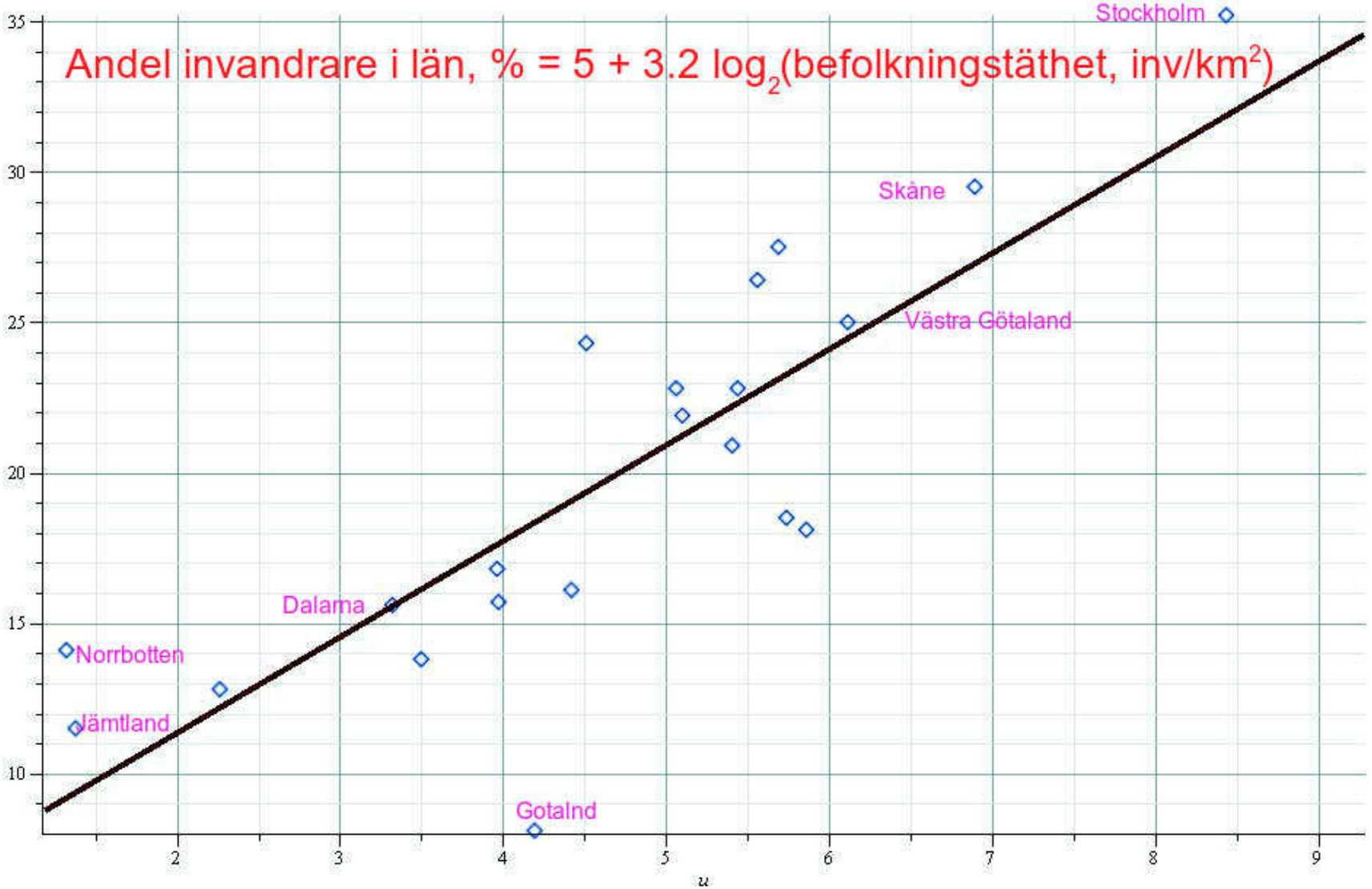
$$z_\ell = a \cdot t_\ell^{\beta} \cdot \varepsilon'_\ell \quad \xrightarrow{\ln} \quad \underbrace{\ln z_\ell}_{y_\ell} = \underbrace{\ln a}_{\alpha} + \beta \underbrace{\ln t_\ell}_{x_\ell} + \underbrace{\ln \varepsilon'_\ell}_{\varepsilon_\ell}$$

Om de multiplikativa felen, ε'_ℓ, är lognormalfördelade blir då $\ln \varepsilon'_\ell \in \mathsf{N}$ och vi kan använda linjär regression för att skatta $\ln \alpha$ och β.


Antal transistorer hos Intelprocessorer
Antal transistorer
Lanseringsår
10^9
10^8
10^7
10^6
10^5
10^4
10^3
10^2
1970
1980
1990
2000
2010
2020
Intel® Itanium® 2
Intel® Itanium®
Intel® Pentium® 4
Intel® Pentium® III
Intel® Pentium® II
Intel® Pentium®
Intel486™
Intel386™
286
8086
8080
8008
4004


Skattat samband: y = 5.13·10^−301· e^0.35 x
x 10^8
5
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
Antal transistorer
1970
1980
1990
2000
2010
2020
Lanseringsår


Samband vikt och hjärnstorlek
log(hjärnvikt) [g]
log(kroppsvikt) [kg]
12
10
8
6
4
2
0
−2
−4
−2
0
2
4
6
8
10
12
Människa
Elefant (Afr)
Elefant (Ind)
Giraff
Häst
Chimpans
Åsna
Ko
Gorilla
Rhesus apa
Får
Gris
Jaguar
Brachiosaurus
Varg
Potar apa
Get
Triceratops
Känguru
Katt
Diplodocus
Kanin
Ekorrbäver
Marsvin
Mullvad
Råtta
Hamster
Mus


Andel invandrare i län, % = 5 + 3.2 log2(befolkningstäthet, inv/km2)
Stockholm
Skåne
Västra Götaland
Dalarna
Norrbotten
Jämtland
Gotalnd
35
30
25
20
15
10
2
3
4
5
6
7
8
9
u

Multipel linjär regression

Modellen kan utökas med flera x-variabler:

$$y_\ell = \beta_0 + \beta_1 x_{\ell 1} + \ldots + \beta_k x_{\ell k} + \varepsilon_\ell, \quad \ell = 1, \ldots, n, \quad \varepsilon_\ell \in \mathrm{N}(0, \sigma)$$

som kan skrivas på matrisform som

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

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

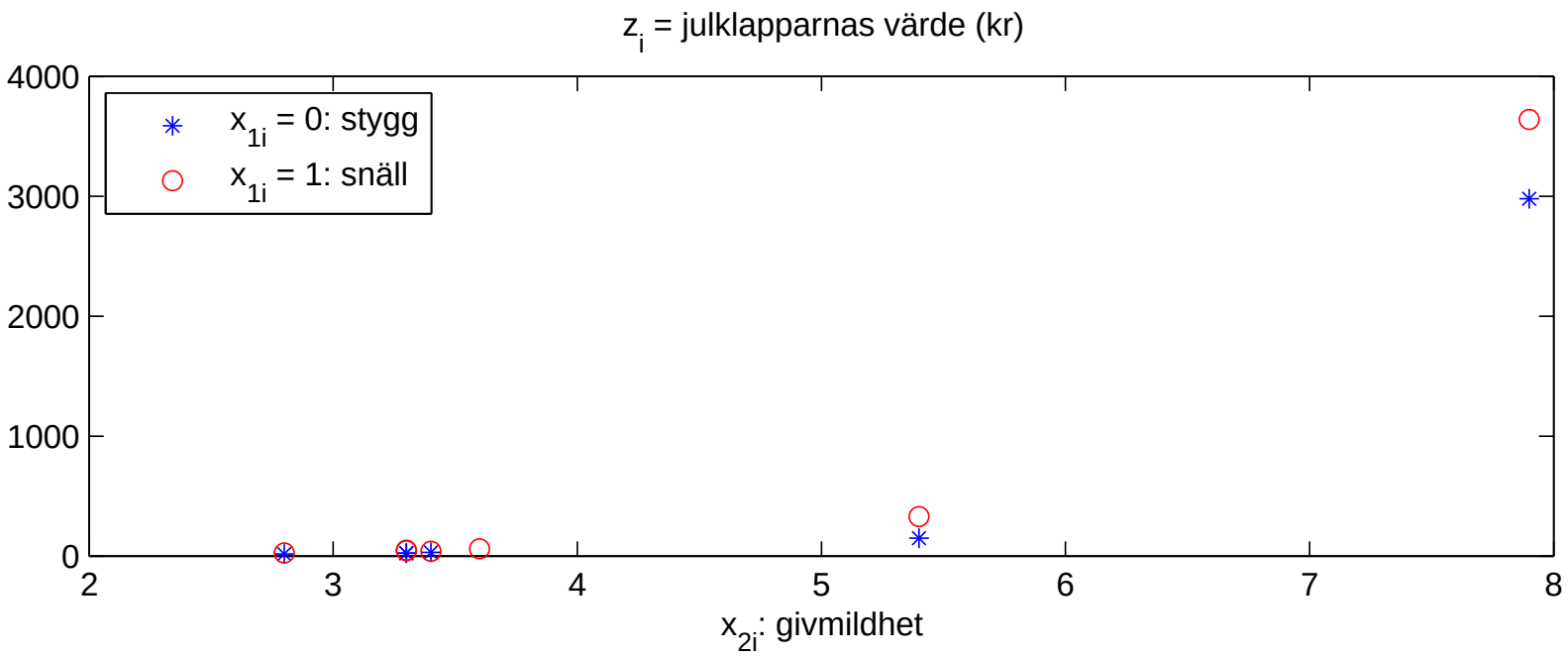
dvs.

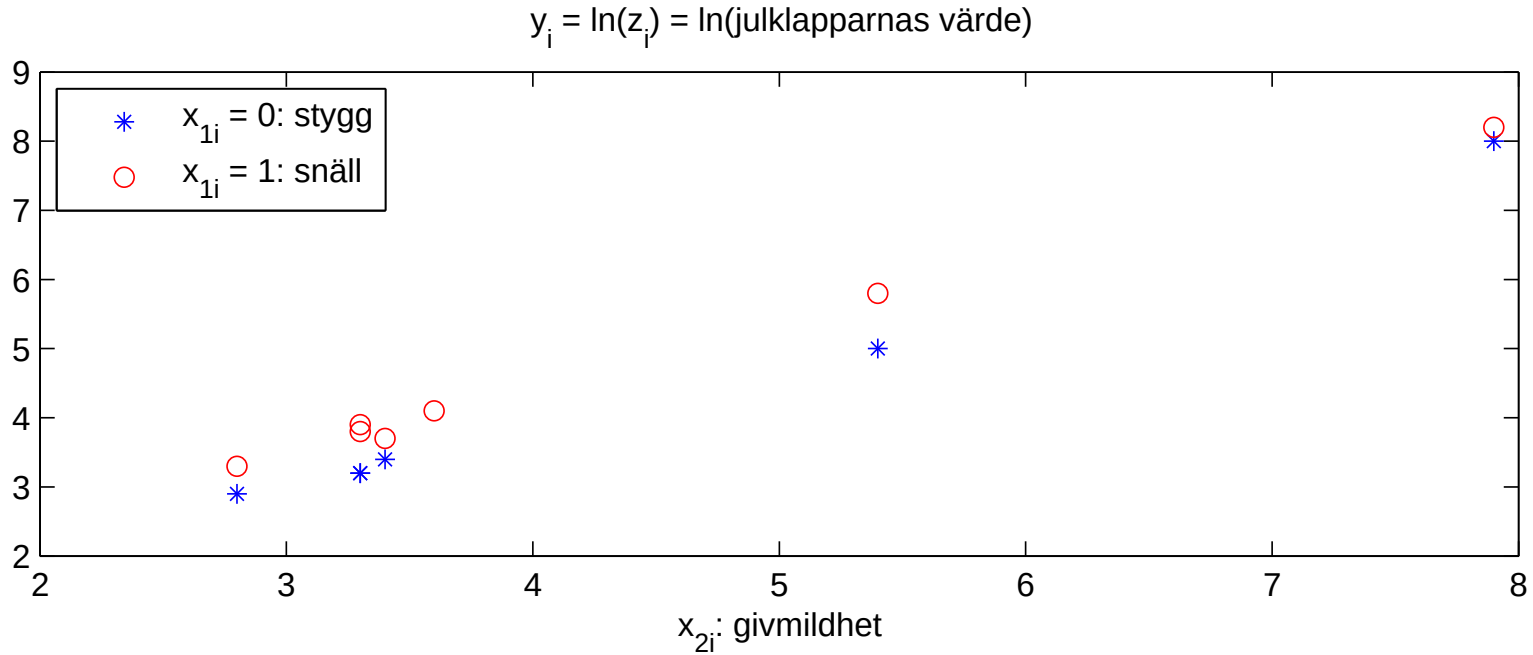
$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\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

Exempel – Julklappar:

En liten flicka vill undersöka om det lönar sig att vara snäll. Hon har därför noterat värdet på de julklappar hon fick från olika släktningar i år, när hon varit snäll, och i fjol då hon var stygg. Hon har insett att värdet på julklapparna också till stor del beror på givarens ekonomi och allmänna generositet. Hon räknar därför också ut ett lämpligt mått på givmildhet.

	värde (kr.)		ln(värde)		
Släkting	i fjol	i år	i fjol	i år	givmildhet
Storebror	24.5	49.5	3.2	3.9	3.3
Lillebror	18.	27.	2.9	3.3	2.8
Mormor och morfar	2981.	3641.	8.0	8.2	7.9
Farmor och farfar	30.	40.	3.4	3.7	3.4
Mamma och pappa	148.	329.50	5.0	5.8	5.4
Moster	24.5	44.5	3.2	3.8	3.3
Kusin	?	62.	?	4.1	3.6


z_i = julklapparnas värde (kr)
x_{1i} = 0: stygg
x_{1i} = 1: snäll
4000
3000
2000
1000
0
2
3
4
5
6
7
8
x_{2i}: givmildhet


$y_i = \ln(z_i)$ = ln(julklapparnas värde)
x_{1i} = 0: stygg
x_{1i} = 1: snäll
9
8
7
6
5
4
3
2
2
3
4
5
6
7
8
x_{2i}: givmildhet

Lämplig regressionsmodell:

$$\ln z_\ell = y_\ell = \alpha + \beta_1 \cdot x_{1\ell} + \beta_2 \cdot x_{2\ell} + \varepsilon_\ell, \qquad \ell = 1, \ldots, 13$$

där e^{β_1} = relativa ökningen i värde när flickan är snäll.

Responsvariabeln:

$$z_\ell = \text{värdet (i kronor) av julklapp } \ell,$$
$$\Longrightarrow y_\ell = \ln z_\ell = \text{logaritmerat värde på julklapp } \ell$$

Förklarande variablerna:

$$x_{1\ell} = \begin{cases} 0 & \text{för alla fjolårets julklappar (då hon varit stygg)} \\ 1 & \text{för alla årets julklappar (då hon varit snäll)} \end{cases},$$

$x_{2\ell}$ = givmildheten hos givaren av julklapp ℓ

$\varepsilon_\ell \in \mathsf{N}(0, \sigma)$ oberoende

- Testa på nivån 5%, om det lönar sig att vara snäll, dvs om β_1 är signifikant större än 0.
- Gör ett tvåsidigt 95% prediktionsintervall för värdet på Kusinens julklapp i fjol, dvs. då den lilla flickan varit stygg.

Modell med matriser: $\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{\varepsilon}$ där $n = 13$ samt $k = 2$ och

$$\boldsymbol{Y} = \begin{bmatrix} 3.2 \\ 3.9 \\ 2.9 \\ 3.3 \\ 8.0 \\ 8.2 \\ 3.4 \\ 3.7 \\ 5.0 \\ 5.8 \\ 3.2 \\ 3.8 \\ 4.1 \end{bmatrix} \quad \boldsymbol{X} = \begin{bmatrix} 1 & 0 & 3.3 \\ 1 & 1 & 3.3 \\ 1 & 0 & 2.8 \\ 1 & 1 & 2.8 \\ 1 & 0 & 7.9 \\ 1 & 1 & 7.9 \\ 1 & 0 & 3.4 \\ 1 & 1 & 3.4 \\ 1 & 0 & 5.4 \\ 1 & 1 & 5.4 \\ 1 & 0 & 3.3 \\ 1 & 1 & 3.3 \\ 1 & 1 & 3.6 \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_{12} \\ \varepsilon_{13} \end{bmatrix}$$

Skattning av parametrarna

Skattning av β

ML- och MK-skattningar av $\beta_0, \ldots, \beta_k$ (elementen i $\boldsymbol{\beta}$) blir

$$\boldsymbol{\beta}^* = \underbrace{(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}}_{\text{(k+1) x (k+1)}} \underbrace{\boldsymbol{X}^\mathsf{T}}_{\text{k x n}} \underbrace{\boldsymbol{y}}_{\text{n x 1}}$$

En väntevärdesriktig skattning av σ^2 ges av (korrigerad ML)

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

Q_0 är alltså residualkvadratsumman och $k+1$ är antalet skattade parametrar i Q_0 (obs. tumregeln!).

Skattningar:

$$\boldsymbol{X}^T\boldsymbol{X} = \begin{bmatrix} 13 & 7 & 55.8 \\ 7 & 7 & 29.7 \\ 55.8 & 29.7 & 278.5 \end{bmatrix} \qquad \boldsymbol{X}^T\boldsymbol{y} = \begin{bmatrix} 58.5 \\ 32.8 \\ 289.09 \end{bmatrix}$$

$$(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \begin{bmatrix} 0.6530 & -0.1786 & -0.1118 \\ -0.1786 & 0.3098 & 0.0028 \\ -0.1118 & 0.0028 & 0.0257 \end{bmatrix}$$

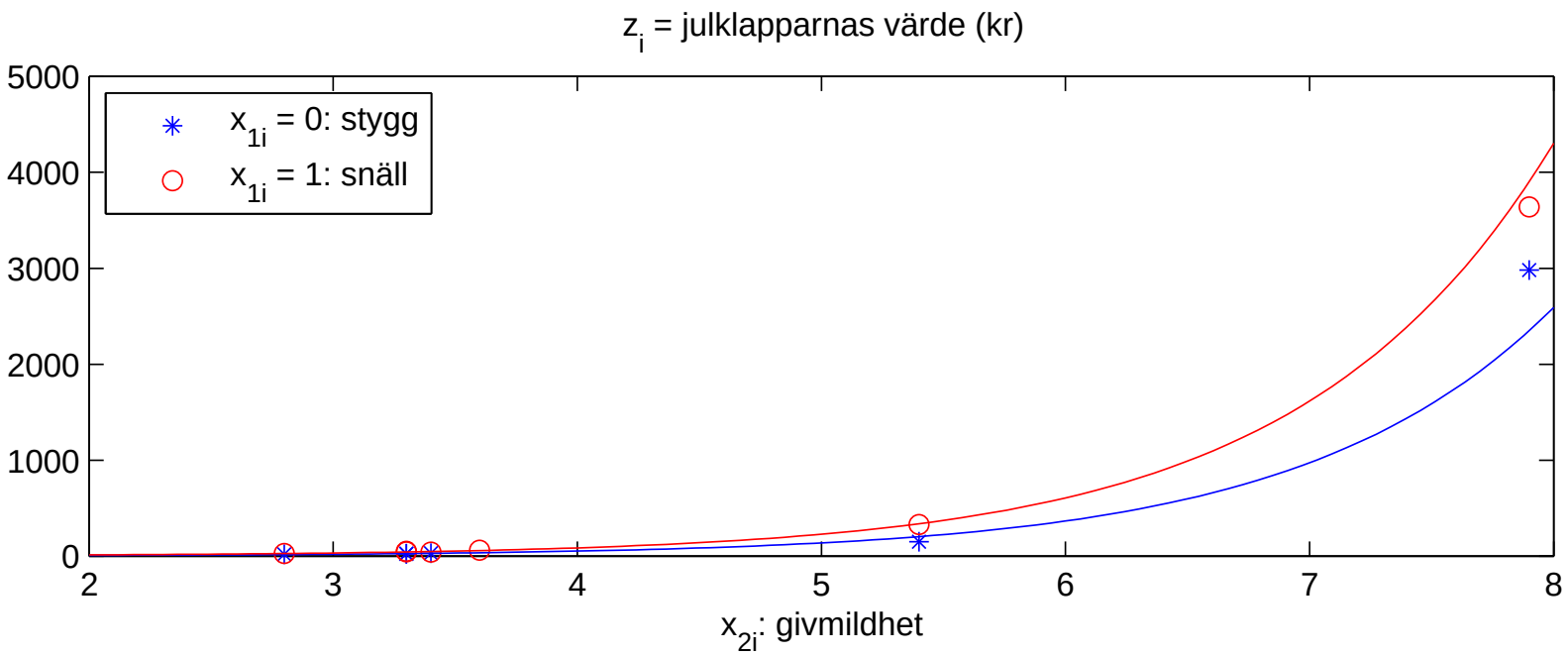
så

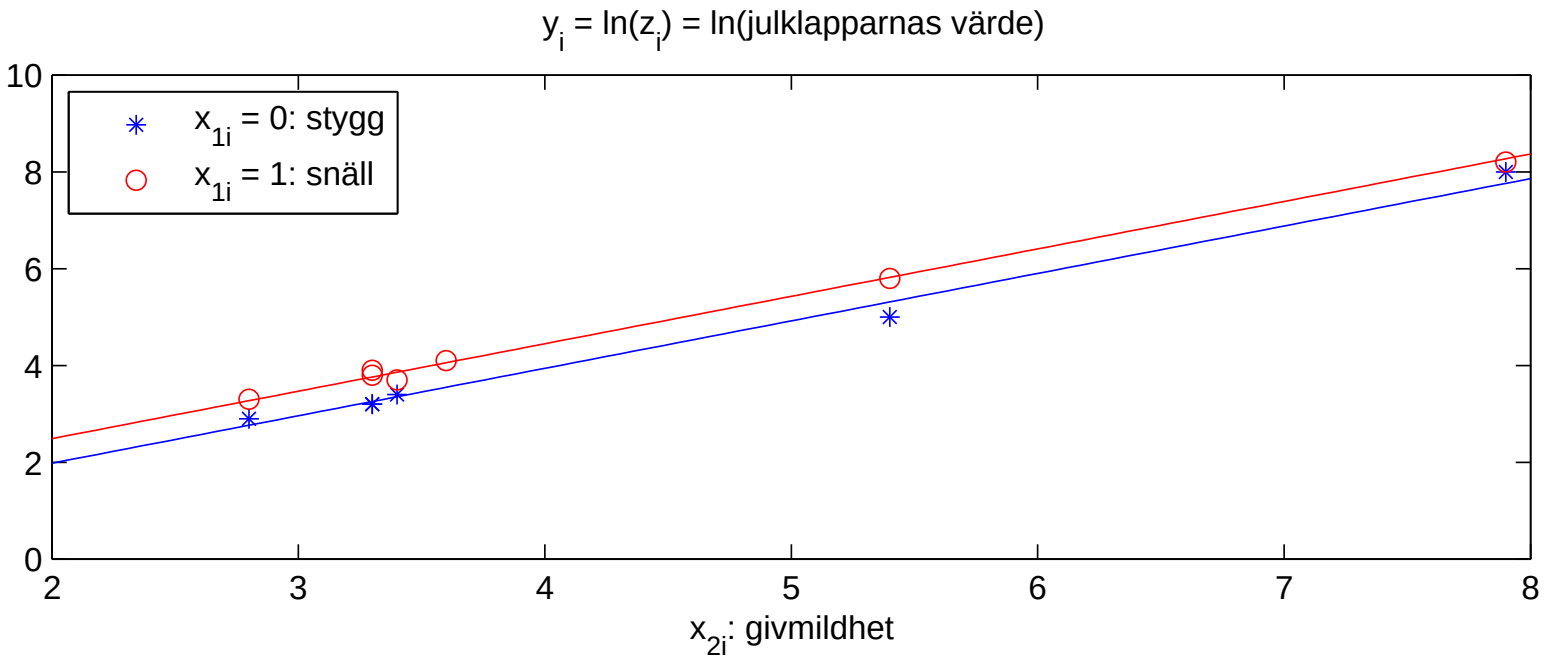
$$\beta^* = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} = \begin{bmatrix} \beta_0^* \\ \beta_1^* \\ \beta_2^* \end{bmatrix} = \begin{bmatrix} 0.0208 \\ 0.5074 \\ 0.9799 \end{bmatrix}$$

$$Q_0 = (\boldsymbol{y} - \boldsymbol{X}\beta^*)^T(\boldsymbol{y} - \boldsymbol{X}\beta^*) = 0.2347$$

$$\sigma^* = s = \sqrt{\frac{Q_0}{f}} = 0.1532$$

samt $f = n - (k+1) = 13 - 3 = 10$


z_i = julklapparnas värde (kr)
x_{1i} = 0: stygg
x_{1i} = 1: snäll
5000
4000
3000
2000
1000
0
2
3
4
5
6
7
8
x_{2i}: givmildhet


$y_i = \ln(z_i)$ = ln(julklapparnas värde)
x_{1i} = 0: stygg
x_{1i} = 1: snäll
10
8
6
4
2
0
2
3
4
5
6
7
8
x_{2i}: givmildhet

Skattningarnas fördelning

Skattningarna av β är linjära funktioner av $\boldsymbol{Y}$ och är därmed normalfördelade

$$\beta_\ell^* \in \mathsf{N}\left(\beta_\ell, \mathsf{D}(\beta_\ell^*)\right),$$

där $\mathsf{D}(\beta_\ell^*)$ ges av roten ur diagonalelementen i **kovariansmatrisen**

$$\mathsf{V}(\beta^*) = \underbrace{\sigma^2(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}}_{\substack{(k+1)\times(k+1)\\ \text{matris}}} = \begin{bmatrix} \mathsf{V}(\beta_0^*) & \mathsf{C}(\beta_0^*, \beta_1^*) & \cdots & \mathsf{C}(\beta_0^*, \beta_k^*) \\ \mathsf{C}(\beta_1^*, \beta_0^*) & \mathsf{V}(\beta_1^*) & \cdots & \mathsf{C}(\beta_1^*, \beta_k^*) \\ \vdots & \vdots & \ddots & \vdots \\ \mathsf{C}(\beta_k^*, \beta_0^*) & \mathsf{C}(\beta_k^*, \beta_1^*) & \cdots & \mathsf{V}(\beta_k^*) \end{bmatrix}.$$

För residualkvadratsumman gäller

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

Konfidensintervall och hypotestest för β_ℓ

Konfidensintervall för β_ℓ blir alltså

$$I_{\beta_\ell} = \beta_\ell^* \pm t_{a/2}(f) \cdot d(\beta_\ell^*) = \beta_\ell^* \pm t_{a/2}(f) \cdot s \cdot \sqrt{\left[(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\right]_{\ell,\ell}}$$

där $[(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}]_{\ell,\ell}$ är diagonalelement nr ℓ, $f = n - k - 1$.
Obs! det första elementet har nummer $\ell = 0$.
Intervallet kan användas för att testa hypotesen

$$H_0 : \beta_\ell = 0 \quad \text{kontra} \quad H_1 : \beta_\ell \neq 0$$

Alternativt kan man naturligtvis använda

$$T = \frac{\beta_\ell^* - 0}{d(\beta_\ell^*)} \qquad \text{och förkasta } H_0 \text{ om } |T| > t_{a/2}(n - (k+1)).$$

(1) Vi vill testa $H_0 : \beta_1 = 0$ mot $H_1 : \beta_1 > 0$ på signifikansnivån $a = 0.05$. Medelfelet blir

$$d(\beta_1^*) = s\sqrt{[(\boldsymbol{X}^T\boldsymbol{X})^{-1}]_{1,1}} = 0.1532 \cdot \sqrt{0.3098} = 0.0853$$

således (obs! ensidig...)

$$T = \frac{\beta_1^* - 0}{d(\beta^*)} = \frac{0.5074}{0.0853} = 5.9496 > t_a(f) = t_{0.05}(10) = 1.81$$

(dvs. ”för mycket”) kan H_0 förkastas. Ja, det lönar sig att vara snäll. Hur mycket lönar det sig? Ett tvåsidigt konfidensintervall för β_1 blir

$$I_{\beta_1} = \beta_1^* \pm t_{a/2}(f) \cdot d(\beta_1^*) = 0.5074 \pm \underbrace{t_{0.025}(10)}_{2.23} \cdot 0.0853$$

$$= (0.3174,\ 0.6974) \Rightarrow I_{e^{\beta_1}} = (e^{0.3174},\ e^{0.6974}) = (1.37,\ 2.01)$$

Att vara snäll ökar värdet på julklapparna med i genomsnitt 37-101%!

Skattning av punkt på ”planet”

$\boldsymbol{Y}$-s väntevärde i en punkt $\boldsymbol{x}_0 = \begin{bmatrix} 1 & x_{01} & x_{02} & \cdots & x_{0k} \end{bmatrix}$ ges nu av

$$\mu^*(\boldsymbol{x}_0) = \beta_0^* + \sum_{\ell=1}^{k} \beta_\ell^* x_{0i} = \boldsymbol{x}_0 \beta^*$$

med $$V(\mu^*(\boldsymbol{x}_0)) = \boldsymbol{x}_0 V(\beta^*) \boldsymbol{x}_0^T = \sigma^2 \underbrace{\boldsymbol{x}_0}_{1\,x\,k} \underbrace{\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}}_{k\,x\,k} \underbrace{\boldsymbol{x}_0^T}_{k\,x\,1}$$

Ett **konfidensintervall** för $\mu^*(\boldsymbol{x}_0)$ blir således (med $f = n - (k+1)$)

$$I_{\mu^*(\boldsymbol{x}_0)} = \underbrace{\boldsymbol{x}_0}_{1\,x\,k} \underbrace{\beta^*}_{k\,x\,1} \pm t_{a/2}(f) \cdot s \cdot \sqrt{\boldsymbol{x}_0 \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1} \boldsymbol{x}_0^T}$$

För **prediktionsintervallet** får man lägga till en etta under kvadratroten, som tidigare

$$I_{Y(\boldsymbol{x}_0)} = \boldsymbol{x}_0 \beta^* \pm t_{a/2}(f) \cdot s \cdot \sqrt{1 + \boldsymbol{x}_0 \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1} \boldsymbol{x}_0^T}$$

(2) Prediktionsintervall för Kusinens julklapp:
Vi har $\boldsymbol{x_0} = \begin{bmatrix} 1 & 0 & 3.6 \end{bmatrix}$ och skattningen

$$\mu^*(\boldsymbol{x_0}) = \boldsymbol{x_0}\beta^* = 1 \cdot \beta_0^* + 0 \cdot \beta_1^* + 3.6 \cdot \beta_2^* = 3.5484,$$
$$e^{\mu^*(\boldsymbol{x_0})} = e^{3.5484} = 35.76\,\text{kr},$$

$$\boldsymbol{x_0}(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{x_0}^\mathsf{T} = 0.18,$$

$$\text{därför } I_{\boldsymbol{Y}(\boldsymbol{x_0})} = 3.5484 \pm 2.23 \cdot 0.1532\sqrt{1 + 0.18} = (3.21,\ 3.89)$$

Omräknat till kronor blir det

$$I_{e^{Y(\boldsymbol{x_0})}} = (e^{3.21},\ e^{3.89}) = (25.69,\ 48.94)\,\text{kr}$$

Modellvalidering

Precis som för enkel regression bör man undersöka **residualerna**

$$\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{X}\beta^*,$$

och förvisssa sig om att de verkar vara oberoende och $\mathsf{N}(0, \sigma)$-fördelade. Plotta residualerna

- ”Som de kommer”, dvs mot $1, 2, \ldots, n$. Ev. ett histogram
- Mot var och en av $\boldsymbol{x}_\ell$-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

$$H_0 : \beta_\ell = 0 \qquad\qquad H_1 : \beta_\ell \neq 0$$

eftersom β_ℓ anger ”*hur mycket* $\boldsymbol{y}$ *ändrar sig när vi ändrar* $\boldsymbol{x}_\ell$”.

Kolinjäritet (ex. två variabler, motsv. för fler)

Man bör om möjligt välja sina $(x_{1\ell}, x_{2\ell})$-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.