Sannolikhet och statistik

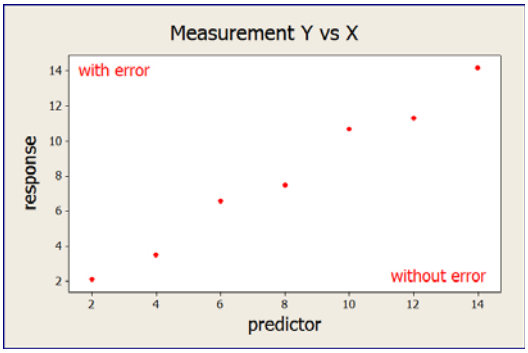
Regressionsanalys

VT 2009

Uwe.Menzel@math.uu.se

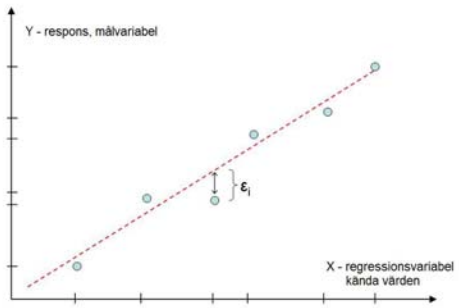
http://www.math.uu.se/∼uwe/

Linjär regression



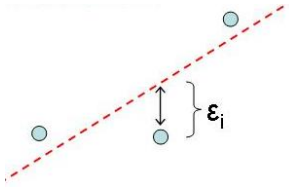
Figur: Mätpunkter: (x_i, y_i)

Linjär regression - kalibrering av en våg



Figur: Modell för linjär regression

Modell för linjär regression



Figur: Modell för linjär regression

Modell:

$y_i = \alpha + \beta x_i + \epsilon_i$

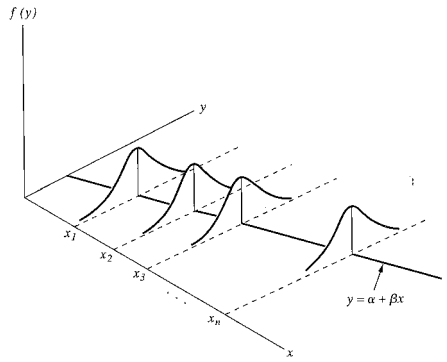
där $\epsilon_i \in N(0, \sigma)$ ("*brus*")

$E(y_i) = \alpha + \beta x_i = \mu_i$

x_i = regressionsvariabel

y_i = målvariabel (s.v.)

Punktskattningar för intercept α och lutning β



Figur: Modell för linjär regression

Punktskattningar för intercept α och lutning β

Modell: $y_i = \alpha + \beta x_i + \epsilon_i$

Minsta-kvadrat-metoden:

$$Q(\alpha, \beta) = \sum_{i=1}^{n} \epsilon_i^2 \implies \text{Minimum}$$

$$Q(\alpha, \beta) = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 \implies \text{Minimum}$$

$$\frac{\partial Q}{\partial \alpha} = -2 \sum_{i=1}^{n} (y_i - \alpha - \beta x_i) = 0$$

$$\frac{\partial Q}{\partial \beta} = -2 \sum_{i=1}^{n} x_i (y_i - \alpha - \beta x_i) = 0$$

Punktskattningar för intercept α och lutning β

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

$$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}; \quad S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

För varje x_0 får man nu en skattning för målvariablen:

$\mu_0^* = \alpha^* + \beta^* \cdot x_0$

Skattning(!): α^*, β^* och μ_0^* är slumpvariabler.

Nytt försök: förändrade $\epsilon_i \Rightarrow y_i \Rightarrow S_{xy} \Rightarrow \beta^*$, α^* och μ_0^*

Exempel 14.2: Bilirubinhalt (x) och proteinkoncentration (y) i ryggmärgsvätskan hos nyfödda.

x	0.14	0.08	0.07	0.26	0.08	0.02	0.03	0.22	0.06	0.23
y	83	65	71	140	135	30	30	128	80	168
x	0.29	0.04	0.13	0.14	0.07	0.05	0.13	0.06	0.05	0.08
y	139	88	121	125	56	98	101	96	73	116

$\bar{x} = 0.1115 \qquad \sum x_i^2 = 0.3701$

$S_{xx} = \sum x_i^2 - n\bar{x}^2 = 0.3701 - 20 \cdot 0.1115^2 = 0.1215$

$\bar{y} = 97.15 \qquad \sum y_i^2 = 215061$

$S_{yy} = \sum y_i^2 - n\bar{y}^2 = 215061 - 20 \cdot 97.15^2 = 26299$

$\sum x_i y_i = 259.79$

$S_{xy} = \sum x_i y_i - n\bar{x}\bar{y} = 259.79 - 20 \cdot 0.1115 \cdot 97.15 = 43.1455$

Exempel 14.2: Bilirubinhalt (x) och proteinkoncentration (y) i ryggmärgsvätskan hos nyfödda.

$$\beta^* = \frac{S_{xy}}{S_{xx}} = \frac{43.1455}{0.1215} = 355.2 \quad \text{lutning}$$

$$\alpha^* = \bar{y} - \beta^* \cdot \bar{x} = 97.15 - 355.2 \cdot 0.1115 = 57.5$$

Skattning för målvariablen:

$$\mu^* = \alpha^* + \beta^* \cdot x = 57.5 + 355.2 \cdot x$$

För varje givet x kan nu den förväntade koncentrationen beräknas.

Hur stor är det minimala Q:et?

Vi sökte *de* värdena för α och β som minimerar uttrycket Q:

$$Q(\alpha, \beta) = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 \implies \text{Minimum}$$

... och fick lösningarna α^* och β^*. Hur stor är Q_0 ?

$$\begin{aligned} Q_0 = Q(\alpha^*, \beta^*) &= \sum_{i=1}^{n} (y_i - \alpha^* - \beta^* x_i)^2 \\ &= \sum_{i=1}^{n} [y_i - \bar{y} + \beta^* \cdot \bar{x} - \beta^* x_i]^2 \\ &= \sum_{i=1}^{n} [(y_i - \bar{y}) - \beta^* (x_i - \bar{x})]^2 \\ &= \ldots \\ &= S_{yy} - S_{xy}^2 / S_{xx} \end{aligned}$$

Q_0 används för att skatta σ

Vi använde modellen:

$y_i = \alpha + \beta x_i + \epsilon_i \quad \text{med} \quad \epsilon_i \in N(0, \sigma)$

$$Q_0 = S_{yy} - S_{xy}^2 / S_{xx} \quad \textit{minsta } Q_0$$

$$E[Q_0] = (n-2) \cdot \sigma^2 \quad \textit{väntevärde}$$

$$s^2 = Q_0/(n-2) \quad \textit{väntevärdesriktig skattning för } \sigma^2$$

$$s = \sqrt{Q_0/(n-2)} \quad \textit{skattning för } \sigma$$

⋆ Skattning β^* är en linjärkombination av y_i:na

$$\begin{aligned} \beta^* &= \frac{1}{S_{xx}} S_{xy} = \frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \\ &= \frac{1}{S_{xx}} S_{xy} = \frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \bar{x}) \cdot y_i \\ &= \sum_{i=1}^{n} c_i \cdot y_i \quad \text{med} \quad c_i = \frac{x_i - \bar{x}}{S_{xx}} \quad (c_i \textit{ ingen s.v.!}) \end{aligned}$$

⋆ Skattning α^* är en linjärkombination av y_i:na

$$\begin{aligned}
\alpha^* &= \bar{y} - \beta^*\bar{x} = \frac{1}{n}\sum_{i=1}^{n} y_i - \bar{x}\sum_{i=1}^{n} c_i \cdot y_i \\
&= \sum_{i=1}^{n}\left(\frac{1}{n} - c_i\bar{x}\right) y_i \\
&= \sum_{i=1}^{n} d_i \cdot y_i \quad \text{med} \quad d_i = \frac{1}{n} - c_i\bar{x} \quad (d_i \textit{ inte någon} \text{ s.v.!})
\end{aligned}$$

Fördelningar för skattningarna β^* och μ_0^*

$$\beta^* = \sum_{i=1}^{n} c_i \cdot y_i \quad \text{med} \quad c_i = \frac{x_i - \bar{x}}{S_{xx}}$$
$$\alpha^* = \sum_{i=1}^{n} d_i \cdot y_i \quad \text{med} \quad d_i = \frac{1}{n} - c_i\bar{x}$$

$\epsilon_i \in N(0,\sigma)$ ... som vi hade antagit
$y_i = \alpha + \beta x_i + \epsilon_i$ ... var modell
$y_i \in N(\alpha + \beta x_i, \sigma)$ ... y_i linjärkombination av ϵ_i

Slumpvariablerna y_i är normalfördelade
$\Longrightarrow$ α^*, β^* och μ_0^* är också normalfördelade!

Väntevärde för skattningen β^*

$$\begin{aligned}
E(\beta^*) &= E\left(\sum_{i=1}^{n} c_i y_i\right) = \sum_{i=1}^{n} c_i E(y_i) = \sum_{i=1}^{n} c_i(\alpha + \beta x_i) \\
&= \alpha\sum_{i=1}^{n}\frac{x_i - \bar{x}}{S_{xx}} + \beta\sum_{i=1}^{n}\frac{x_i - \bar{x}}{S_{xx}}x_i \quad \text{ty} \quad c_i = \frac{x_i - \bar{x}}{S_{xx}} \\
&= \frac{\alpha}{S_{xx}} \cdot \sum_{i=1}^{n}(x_i - \bar{x}) + \frac{\beta}{S_{xx}}\sum_{i=1}^{n}(x_i - \bar{x})\,x_i \\
&= \frac{\alpha}{S_{xx}} \cdot 0 + \frac{\beta}{S_{xx}}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x}) \quad \text{ty} \quad \sum_{i=1}^{n}(x_i - \bar{x})\,\bar{x} = 0 \\
&= 0 + \frac{\beta}{S_{xx}}S_{xx} \\
&= \beta \quad \text{väntevärdesriktigt}
\end{aligned}$$

Varians för skattningen β^*

$$\begin{aligned}
V(\beta^*) &= V\left(\sum_{i=1}^{n} c_i y_i\right) = \sum_{i=1}^{n} V(c_i y_i) \quad \text{bruset oberoende!} \\
&= \sum_{i=1}^{n} c_i^2 V(y_i) = \sum_{i=1}^{n} c_i^2\sigma^2 \\
&= \sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{S_{xx}}\right)^2\sigma^2 \\
&= \frac{\sigma^2}{S_{xx}^2}\sum_{i=1}^{n}(x_i - \bar{x})^2 \\
&= \frac{\sigma^2}{S_{xx}}
\end{aligned}$$

Variansen för β^* är liten om x-värdena är "utspredda".

Fördelning för β^*

$$E(\beta^*) = \beta$$

$$V(\beta^*) = \frac{\sigma^2}{S_{xx}}$$

$$D(\beta^*) = \frac{\sigma}{\sqrt{S_{xx}}}$$

$$\beta^* \in N\left(\beta, \frac{\sigma}{\sqrt{S_{xx}}}\right)$$

Fördelningar för α^* och μ_0^*

$$\mu_0^* \in N\left(\alpha + \beta x_0,\ \sigma \cdot \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}\right)$$

för $x_0 = 0$ erhåller man fördelningen för α:

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

Fördelning för skattningarna β^* och μ_0^*

Skattningarna β^* och μ_0^* är normalfördelade med:

$$E(\beta^*) = \beta \quad V(\beta^*) = \frac{\sigma^2}{S_{xx}} \quad \text{med} \quad S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$E(\mu_0^*) = \mu_0 \quad V(\mu_0^*) = \sigma^2\left(1/n + (x_0 - \bar{x})^2/S_{xx}\right)$$

Det betyder:

$$\beta^* \in N\left(\beta,\ \frac{\sigma}{\sqrt{S_{xx}}}\right)$$

$$\mu_0^* \in N\left(\mu_0,\ \sigma\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}\right)$$

Intervallskattning för lutningen β när σ är känd

$$\beta^* \in N\left(\beta,\ \frac{\sigma}{\sqrt{S_{xx}}}\right)$$

$$\frac{\beta^* - \beta}{\sigma/\sqrt{S_{xx}}} \in N(0,1) \qquad \textit{referensvariabel}$$

$$P\left(-\lambda_{\alpha/2} \le \frac{\beta^* - \beta}{\sigma/\sqrt{S_{xx}}} \le \lambda_{\alpha/2}\right) = 1 - \alpha$$

$$I_\beta = \beta^*_{obs} \pm \lambda_{\alpha/2} \cdot \frac{\sigma}{\sqrt{S_{xx}}}$$

Intervallskattning för målvariabeln μ_0 när σ är känd

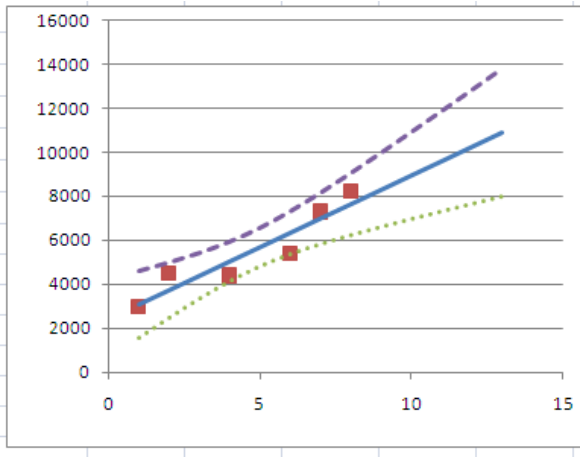
$$\mu_0^* \in N\left(\alpha + \beta x_0,\ \sigma \cdot \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}\right)$$

$$\frac{\mu_0^* - \mu_0}{\sigma \cdot \sqrt{\cdots}} \in N(0,1) \qquad \textit{referensvariabel}$$

$$P\left(-\lambda_{\alpha/2} \le \frac{\mu_{obs} - \mu_0}{\sigma \cdot \sqrt{\cdots}} \le \lambda_{\alpha/2}\right) = 1 - \alpha$$

$$I_{\mu_0} = \alpha + \beta x_0 \pm \lambda_{\alpha/2} \cdot \sigma \cdot \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

Konfidensintervall för målvariabeln μ_0



Intervallskattning för lutningen β när σ är okänd

$$\beta^* \in N(\beta, \frac{\sigma}{\sqrt{S_{xx}}})$$

$$Z = \frac{\beta^* - \beta}{\sigma/\sqrt{S_{xx}}} \in N(0,1) \qquad \textit{inte någon referensvariabel}$$

$$W = \frac{Q_0}{\sigma^2} = (n-2) \cdot \frac{S^2}{\sigma^2} \in \chi^2(n-2)$$

$$\frac{Z}{\sqrt{\frac{W}{n-2}}} \in t(n-2)$$

$$\frac{\frac{(\beta^* - \beta)\sqrt{S_{xx}}}{\sigma}}{\frac{S}{\sigma}} = \frac{\beta^* - \beta}{S/\sqrt{S_{xx}}} \in t(n-2) \qquad \textit{referensvariabel}$$

Allmän metod

$$\frac{\beta^* - \beta}{S/\sqrt{S_{xx}}} \in t(n-2) \qquad \textit{referensvariabel}$$

$$P\left(-t_{\alpha/2}(n-2) \le \frac{\beta_{obs} - \beta}{s/\sqrt{S_{xx}}} \le t_{\alpha/2}(n-2)\right) = 1 - \alpha$$

$$I_\beta = \beta_{obs} \pm t_{\alpha/2}(n-2) \cdot \frac{s}{\sqrt{S_{xx}}} \quad \text{med} \quad s = \sqrt{Q_0/(n-2)}$$

Exempel 14.2: Bilirubinhalt (x) och proteinkoncentration (y).

$S_{xx} = 0.1215 \quad S_{yy} = 26299 \quad S_{xy} = 43.1455 \quad \beta^* = 355.2$

Sökes: 95% konfidensintervall för lutningen β:

$Q_0 = S_{yy} - \frac{S_{xy}^2}{S_{xx}} = 26299 - \frac{43.1455^2}{0.1215} = 10973$

$s = \sqrt{Q_0/(n-2)} = \sqrt{10973/(20-2)} = 24.7$

$d = \frac{s}{\sqrt{S_{xx}}} = \frac{24.7}{\sqrt{0.1215}} = 70.9$

$t_{\alpha/2}(n-2) = t_{0.025}(18) = 2.1$

$I_\beta = \beta^*_{obs} \pm t_{\alpha/2}(n-2) \cdot d = 355.2 \pm 2.1 \cdot 70.9 = (210, 500)$

Intervallskattning för målvariabeln μ_0 när σ är okänd

$$\mu_0^* \in N\left(\mu_0,\ \sigma \cdot \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}\right)$$

$$Z = \frac{\mu_0^* - \mu_0}{\sigma \cdot \sqrt{\cdots}} \in N(0,1) \quad \textit{inte någon referensvariabel}$$

$$W = \frac{Q_0}{\sigma^2} = (n-2) \cdot \frac{S^2}{\sigma^2} \in \chi^2(n-2)$$

$$\frac{Z}{\sqrt{\frac{W}{n-2}}} \in t(n-2)$$

$$\frac{\frac{\mu_0^* - \mu_0}{\sigma \cdot \sqrt{\cdots}}}{\frac{S}{\sigma}} = \frac{\mu_0^* - \mu_0}{S \cdot \sqrt{\cdots}} \in t(n-2) \quad \textit{referensvariabel}$$

Allmän metod

$$\frac{\mu_0^* - \mu_0}{S \cdot \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}} \in t(n-2) \quad \textit{referensvariabel}$$

$$P\left(-t_{\alpha/2}(n-2) \le \frac{\mu_{obs} - \mu_0}{s \cdot \sqrt{\cdots}} \le t_{\alpha/2}(n-2)\right) = 1 - \alpha$$

$$I_{\mu_0} = \mu_{obs} \pm t_{\alpha/2}(n-2) \cdot s \cdot \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

$\mu_{obs} = \alpha + \beta x_0 \qquad s = \sqrt{\frac{Q_0}{n-2}} \qquad Q_0 = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$

Exempel 14.2: Bilirubinhalt (x) och proteinkoncentration (y).

$\bar{x} = 0.1115 \quad S_{xx} = 0.1215 \quad \alpha^* = 57.5 \quad \beta^* = 355.2$

$s = 24.7$ (samma $s = \sqrt{Q_0/(n-2)}$ men annat d)

Sökes: 95% konfidensintervall för μ_0 när $x_0 = 0.2$:

$\mu_0^* = \alpha^* + \beta^* \cdot x_0 = 57.5 + 355.2 \cdot 0.2 \approx 129$

$$d = s \cdot \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} = 24.7 \sqrt{\frac{1}{20} + \frac{(0.2 - 0.1115)^2}{0.1215}} = 8.38$$

$I_{\mu_0} = \mu^*_{obs} \pm t_{\alpha/2}(n-2) \cdot d = 129 \pm t_{0.025}(18) \cdot 8.38 = (110, 150)$

Intervallskattning för β , μ_0 och α

σ känd	σ okänd
$\frac{\beta^* - \beta}{D} \in N(0,1)$ $D = \frac{\sigma}{\sqrt{S_{xx}}}$ $I_\beta = \beta^* \pm \lambda_{\alpha/2} \cdot D$	$\frac{\beta^* - \beta}{d} \in t(n-2)$ $d = \frac{s}{\sqrt{S_{xx}}}$ $I_\beta = \beta^* \pm t_{\alpha/2} \cdot d$
$\frac{\mu_0^* - \mu_0}{D} \in N(0,1)$ $D = \sigma\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$ $I_{\mu_0} = \mu^* \pm \lambda_{\alpha/2} \cdot D$	$\frac{\mu_0^* - \mu_0}{d} \in t(n-2)$ $d = s\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$ $I_{\mu_0} = \mu^* \pm t_{\alpha/2} \cdot d$

$$s = \sqrt{Q_0/(n-2)}$$

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

$$S_{xy} = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$$

Sammanfattning

- Punktskattning för lutning, intercept och målvariabel
- Punktskattning för σ ("brus")
- Väntevärde och varians för skattningarna β^* och μ_0^*
- Intervallskattning för lutning och målvariabel med känt σ
- Intervallskattning för lutning och målvariabel med skattad σ