

# Sannolikhet och statistik

## Regressionsanalys

HT 2008

Uwe.Menzel@math.uu.se

<http://www.math.uu.se/~uwe/>

## Linjär regression

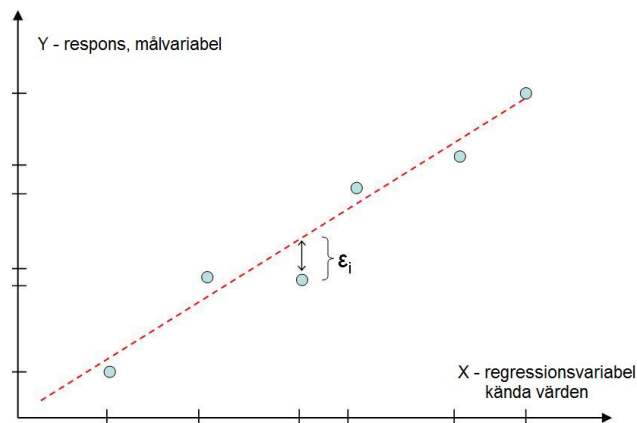
**Samband** mellan två storheter ofta av intresse:

- ▶ ... solarium  $\Leftrightarrow$  hudcancer
- ▶ ... BNP  $\Leftrightarrow$  växelkurs
- ▶ ... rökandet  $\Leftrightarrow$  livslängd
- ▶ ... vikt av en pendel  $\Leftrightarrow$  period av en svängning

**Mätpunkter:**  $(x_i, y_i)$

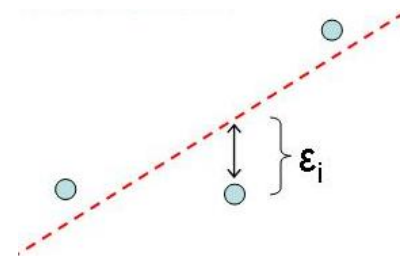
**Linjär regression:**  $y = \alpha + \beta x$

## Linjär regression - kalibrering av en våg



Figur: Modell för linjär regression

## Modell för linjär regression



Figur: Modell för linjär regression

**Modell:**

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

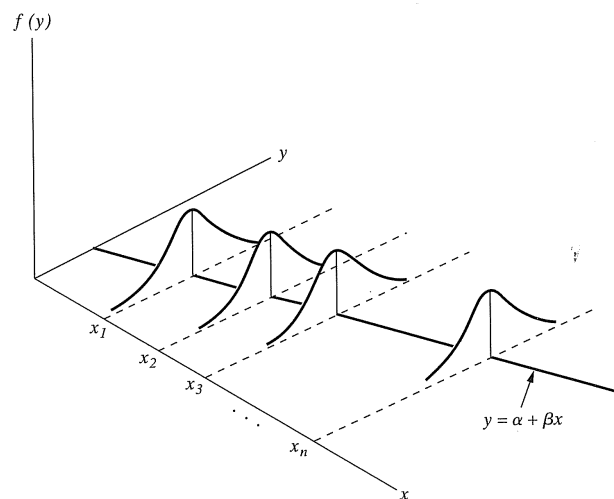
där  $\epsilon_i \in N(0, \sigma)$  ("brus")

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

$x_i$  = regressionsvariabel

$y_i$  = målvariabel (s.v.)

## Punktskattningar för intercept $\alpha$ och lutning $\beta$



Figur: Modell för linjär regression

## Punktskattningar för intercept $\alpha$ och lutning $\beta$

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

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

För varje  $x_0$  får man nu en skattning för målvariabeln:

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

**Skattning(!):**  $\alpha^*$ ,  $\beta^*$  och  $\mu_0^*$  är slumpvariabler.

Nytt försök: förändrade  $\epsilon_i \Rightarrow y_i \Rightarrow S_{xy} \Rightarrow \beta^*$ ,  $\alpha^*$  och  $\mu_0^*$

## Punktskattningar för intercept $\alpha$ och lutning $\beta$

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

**Minsta-kvadrat-metoden:**

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

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

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

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

**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  |

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

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

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

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

$$\sum x_i y_i = 259.79$$

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

**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 målvariabeln:**

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

**För varje** givet  $x$  kan nu den förväntade koncentrationen beräknas.

$Q_0$  används för att skatta  $\sigma$

**Vi använde** modellen:

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

**ML-skattning** för  $\sigma^2$  är  $Q_0/n$  (jämför  $Q = \sum_{i=1}^n \epsilon_i^2$ )

**Korrigerade ML-skattning:**

$$s^2 = Q_0/(n-2) \quad \text{resp.} \quad s = \sqrt{Q_0/(n-2)} \quad \text{väntevärdesriktigt}$$

Hur stor är det minimala  $Q$ ?

**Vi sökte** de värdena för  $\alpha$  och  $\beta$  som minimerar uttrycket  $Q$ :

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

... och fick lösningarna  $\alpha^*$  och  $\beta^*$ . Hur stor är  $Q_0$ ?

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

★ Skattning  $\beta^*$  är en linjärkombination av  $y_i$ :na

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

★ Skattning  $\alpha^*$  är en linjärkombination av  $y_i$ :na

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

Skattningar  $\alpha^*$  och  $\beta^*$  är linjärkombinationer av  $y_i$ :na

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

**Möjlighet** att dra slutsatser (se nere)

Väntevärde för skattningen  $\beta^*$

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

Varians för skattningen  $\beta^*$

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

**Variansen** för  $\beta^*$  är liten om x-värdena är "utspredda".

## Väntevärde och varians för skattningen $\mu_0^*$

**Skattningen** för funktionsvärdet för  $x_0$  var:  $\mu_0^* = \alpha^* + \beta^* \cdot x_0$

$$\begin{aligned} E(\mu_0^*) &= E(\alpha^* + \beta^* \cdot x_0) = E\left(\sum_{i=1}^n (d_i + c_i x_0) Y_i\right) \\ &= \dots \\ &= \alpha + \beta x_0 = \mu_0 \quad \text{väntevärdesriktigt} \\ V(\mu_0^*) &= V(\alpha^* + \beta^* \cdot x_0) = V\left(\sum_{i=1}^n (1/n + c_i(x_0 - \bar{x})) Y_i\right) \\ &= \dots \\ &= \sigma^2 \left(1/n + (x_0 - \bar{x})^2 / S_{xx}\right) \end{aligned}$$

**Variansen** är stor när  $x_0$  ligger långt ifrån  $\bar{x}$ .

För  $x_0 = 0$  får man skattningen för interceptet  $\alpha^*$ .

## Fördelning för skattningarna $\beta^*$ och $\mu_0^*$

**Skattningarna**  $\beta^*$  och  $\mu_0^*$  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:

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

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

## Fördelningar för skattningarna $\beta^*$ och $\mu_0^*$

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

$\epsilon_i \in N(0, \sigma)$  ... 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  $\epsilon_i$

**Slumpvariablerna**  $y_i$  är normalfördelade

$\Rightarrow \alpha^*, \beta^*$  och  $\mu_0^*$  är också normalfördelade!

## Intervallskattning för lutningen $\beta$ när $\sigma$ är känd

**Fördelning** för  $\beta^*$  med känd  $\sigma_0$ :

$$\beta^* \in N(\beta, D) \quad \text{med} \quad D = \frac{\sigma_0}{\sqrt{S_{xx}}} \quad \Rightarrow \quad \frac{\beta^* - \beta}{D} \in N(0, 1)$$

**Referensvariabeln** omfattas med kvantilerna:

$$P\left(-\lambda_{\alpha/2} < \frac{\beta^* - \beta}{D} \leq \lambda_{\alpha/2}\right) = 1 - \alpha$$

$$P(\beta_{obs}^* - \lambda_{\alpha/2} \cdot D < \beta \leq \beta_{obs}^* + \lambda_{\alpha/2} \cdot D) = 1 - \alpha$$

**Konfidensintervall** med konfidensgrad  $1 - \alpha$ :

$$I_\beta = \beta_{obs}^* \pm \lambda_{\alpha/2} \cdot D$$

## Intervallskattning för målvariabeln $\mu_0$ när $\sigma$ är känd

Fördelning för  $\mu_0^*$  med känd  $\sigma_0$ :

$$\mu_0^* \in N(\mu_0, D) \text{ med } D = \sigma_0 \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \Rightarrow \frac{\mu_0^* - \mu_0}{D} \in N(0, 1)$$

Referensvariabeln omfattas med kvantilerna:

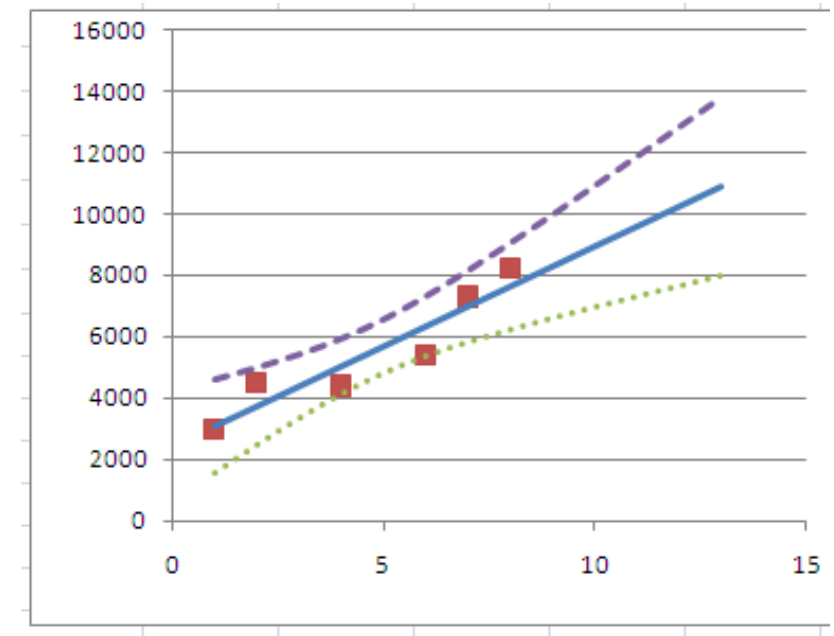
$$P\left(-\lambda_{\alpha/2} < \frac{\mu_0^* - \mu_0}{D} \leq \lambda_{\alpha/2}\right) = 1 - \alpha$$

$$P(\mu_{obs}^* - \lambda_{\alpha/2} \cdot D < \mu_0 \leq \mu_{obs}^* + \lambda_{\alpha/2} \cdot D) = 1 - \alpha$$

Konfidensintervall med konfidsgrad  $1 - \alpha$ :

$$I_{\mu_0} = \mu_{obs}^* \pm \lambda_{\alpha/2} \cdot D$$

## Konfidensintervall för målvariabeln $\mu_0$



Figur: Skattning för målvariabeln med konfidensintervall

## Intervallskattning för lutningen $\beta$ när $\sigma$ är okänd

Är  $\sigma$  okänd skattas det med  $s = \sqrt{Q_0/(n-2)}$

Referensvariabel:

$$\frac{\beta^* - \beta}{d} \in t(f) \text{ med } d = \frac{s}{\sqrt{S_{xx}}} \text{ och } f = n - 2$$

Referensvariabeln omfattas med kvantilerna:

$$P\left(-t_{\alpha/2} < \frac{\beta^* - \beta}{d} \leq t_{\alpha/2}\right) = 1 - \alpha$$

$$P(\beta_{obs}^* - t_{\alpha/2} \cdot d < \beta \leq \beta_{obs}^* + t_{\alpha/2} \cdot d) = 1 - \alpha$$

Konfidensintervall med konfidsgrad  $1 - \alpha$ :

$$I_{\beta} = \beta_{obs}^* \pm t_{\alpha/2}(n-2) \cdot d$$

**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  $\beta$ :

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

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

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

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

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

## Intervallskattning för målvariabeln $\mu_0$ när $\sigma$ är okänd

Är  $\sigma$  okänd skattas det med  $s = \sqrt{Q_0/(n-2)}$

En referensvariabel är:

$$\frac{\mu_0^* - \mu_0}{d} \in t(n-2) \quad \text{med} \quad d = s \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

Referensvariabeln omfattas med kvantilerna:

$$P\left(-t_{\alpha/2} < \frac{\mu_0^* - \mu_0}{d} \leq t_{\alpha/2}\right) = 1 - \alpha$$

$$P(\mu_{obs}^* - t_{\alpha/2} \cdot d < \mu \leq \mu_{obs}^* + t_{\alpha/2} \cdot d) = 1 - \alpha$$

Konfidensintervall med konfidensgrad  $1 - \alpha$ :

$$I_{\mu_0} = \mu_{obs}^* \pm t_{\alpha/2}(n-2) \cdot d$$

## Intervallskattning för $\beta$ och $\mu_0$

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

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

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

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

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

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

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

**Sökes:** 95% konfidensintervall för  $\mu_0$  när  $x_0 = 0.2$ :

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

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

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

## Sammanfattning

- Punktskattning för lutning, intercept och målvariabel
- Punktskattning för  $\sigma$  ("brus")
- Väntevärde och varians för skattningarna  $\beta^*$  och  $\mu_0^*$
- Intervallskattning för lutning och målvariabel med känt  $\sigma$
- Intervallskattning för lutning och målvariabel med skattad  $\sigma$