

Grundläggande matematisk statistik

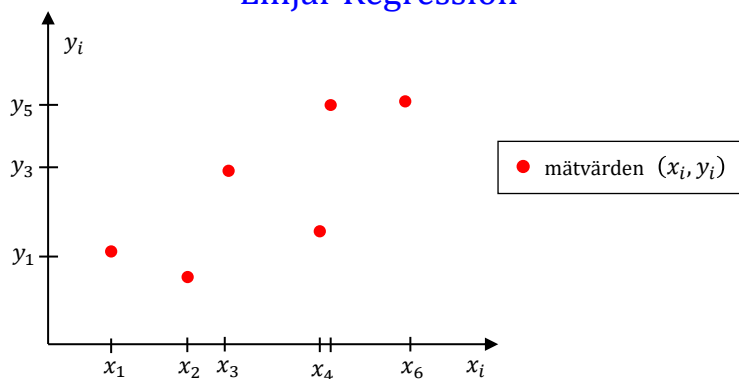
Linjär Regression

Uwe Menzel, 2018

uwe.menzel@matstat.org

www.matstat.org

Linjär Regression



x_i : oberoende variabler (prediktorer, kovariater)

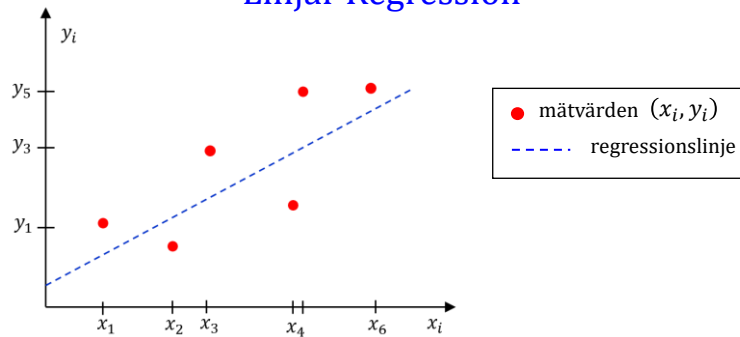
- **kända**, inga slumpvariabler!

y_i : beroende variabler (målvariabler, responsvariabler)

- behäftade med **brus** (mätningar), **slumpvariabler**

www.matstat.org

Linjär Regression



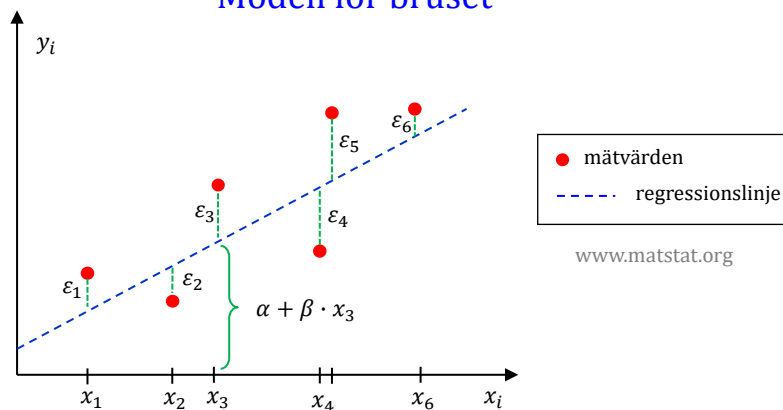
Vi antar att det finns ett **linjärt samband mellan x och y** :

$$y = \alpha + \beta \cdot x$$

- α och β är okända, vi söker skattningar för dem
- skattningarna beräknas baserat på ett stickprov
- stickprov = mätningar av talpar (x_i, y_i)

www.matstat.org

Modell för bruset



www.matstat.org

Modell för att hantera bruset:

$$y_i = \alpha + \beta \cdot x_i + \varepsilon_i \quad i = 1, 2, \dots, n \quad n = \text{antal mätningar (stickprovsstorlek)}$$

ε_i = "brus"

$$\varepsilon_i \sim N(0, \sigma)$$

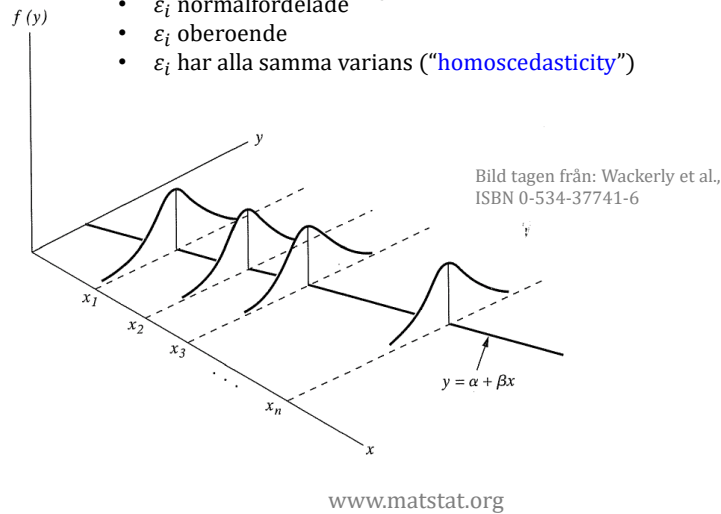
$$E(\varepsilon_i) = 0$$

$$V(\varepsilon_i) = \sigma^2$$

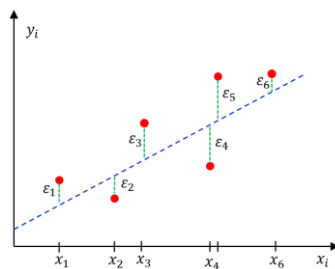
Modell för bruset

Antaganden:

- ε_i normalfördelade
- ε_i oberoende
- ε_i har alla samma varians ("homoscedasticity")



Att hitta den bästa regressionslinjen



Minsta-kvadrat metoden ("least squares method"):

$$Q = \sum_{i=1}^n \varepsilon_i^2 \Rightarrow \text{Min}$$

Idè

Q : kvadratsumma
("sum of squares")

- regressionslinjen (- - - -) ska läggas så att Q minimeras
- använder kvadraten av ε_i så att positiva och negativa värden inte jämnar ut varandra

Att bestämma α och β så att Q minimeras:

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

www.matstat.org

Att hitta den bästa regressionslinjen

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

$$\Rightarrow \left. \begin{aligned} \frac{\partial Q}{\partial \alpha} &= -2 \sum_{i=1}^n (y_i - \alpha - \beta x_i) = 0 \\ \frac{\partial Q}{\partial \beta} &= -2 \sum_{i=1}^n x_i (y_i - \alpha - \beta x_i) = 0 \end{aligned} \right\} \begin{array}{l} \text{partiella deriveringar} \\ \text{måsta vara noll} \end{array}$$

$$\Rightarrow \left. \begin{aligned} \sum_{i=1}^n y_i - n \cdot \alpha - \beta \sum_{i=1}^n x_i &= 0 \\ \sum_{i=1}^n y_i \cdot x_i - \alpha \cdot \sum_{i=1}^n x_i - \beta \sum_{i=1}^n x_i^2 &= 0 \end{aligned} \right\} \begin{array}{l} \text{linjära ekvationer för } \alpha \text{ och } \beta \\ \text{(allt annat känt från stickprovet)} \\ \text{två ekvationer, två variabler} \quad \checkmark \end{array}$$

Uwe Menzel, 2017

Att hitta den bästa regressionslinjen

Lösningar:

$$\beta^* = \frac{S_{xy}}{S_{xx}}$$

$$\alpha^* = \bar{y} - \beta^* \bar{x}$$

(med ekvation 1)

skattningar för α och β

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

$$S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - n\bar{x}^2$$

OBS!: α^* och β^* är slumpvariabler (beror på stickprovet) :

nytt stickprov ändrar $\varepsilon_i \rightarrow y_i \rightarrow S_{xy} \rightarrow \alpha^*$

För varje x_0 får man nu en **skattning av målvariabeln** (x_0 måste inte stämma överens med någon mätt x_i):

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

www.matstat.org

Skattning av brusets varians

Vi ville minimera Q . Hur stort blev det?: $Q(\alpha, \beta) = \sum_i (y_i - \alpha - \beta \cdot x_i)^2$

Låt oss mata in de värdena som minimerar Q , alltså α^* och β^* :

$$\begin{aligned} Q_0 &= \sum_{i=1}^n (y_i - \alpha^* - \beta^* x_i)^2 = \sum_{i=1}^n [y_i - \bar{y} + \underbrace{\beta^* \cdot \bar{x} - \beta^* x_i}_{-\alpha^*}]^2 \\ &= \sum_{i=1}^n [(y_i - \bar{y}) - \beta^* (x_i - \bar{x})]^2 = \dots = S_{yy} - S_{xy}^2 / S_{xx} \quad (\text{helt känd från stickprovet}) \end{aligned}$$

Man kan använda **Maximum-Likelihood-metoden** för att hitta en skattning för σ^2 . Detta ger Q_0/n . Skattningen måste dock korrigeras för att bli väntevärdesriktig:

$$\boxed{s_r^2 = \frac{Q_0}{n-2}} \quad E(s_r^2) = \sigma^2 \quad \begin{array}{l} s_r^2 \text{ är en väntevärdesriktig skattning för } \sigma^2 \\ (\text{och } s_r \text{ en väntevärdesriktig skattning för } \sigma) \end{array}$$

$$(\sigma^2)^* = s_r^2 \quad \sigma^* = s_r \quad \text{skattning för varians/std.avvikelse av bruset } \varepsilon$$

Denna formel är åskadlig: ju större $Q = \sum_i \varepsilon_i^2$, desto större σ (se bild ovan).

www.matstat.org

Att hitta den bästa regressionslinjen

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

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

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

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

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

$$\sum x_i y_i = 259.79$$

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

www.matstat.org

Exempel, fortsättning

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

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

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

Skattning för målvariabeln:

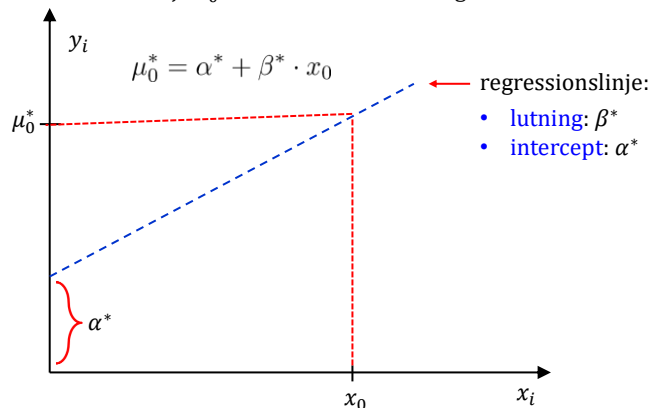
även för x som aldrig mätts

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

För varje givet x kan nu den förväntade koncentrationen beräknas.

Beräkning av regressionslinjen

För varje x_0 får man nu en skattning av målvariabeln:



- x_0 kan även ligga mellan de mätta x_i -värdena
- α^* och β^* är slumpvariabler $\rightarrow \mu_0$ är en slumpvariabel (förändras när nytt stickprov tas).

Fördelningarna för de skattade variablerna

Vi har sett att $\alpha^*, \beta^*, \mu_0^*, \dots$ är slumpvariabler. Vilken fördelning har de?

$$y_i = \alpha + \beta \cdot x_i + \varepsilon_i \quad i = 1, 2, \dots, n$$

$y_i \sim N$ $\longleftrightarrow$ $\varepsilon_i \sim N$

målvariablerna y_i är normalfördelade, eftersom bruset ε_i är normalfördelat

Man kan räkna ut att (lite krångligt, se appendixet):

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

α^* och β^* är linjärkombinationer av normalfördelade slumpvariabler (y_i)

$\Rightarrow \boxed{\alpha^*, \beta^*, \mu_0^* \sim N}$

$\Rightarrow$ Regressionsvariablerna $\alpha^*, \beta^*, \mu_0^*$ är normalfördelade.

www.matstat.org

Fördelningarna för de skattade variablerna

Regressionsvariablerna $\alpha^*, \beta^*, \mu_0^*$ är normalfördelade. Vi kan räkna ut deras väntevärde, varians och standardavvikelse:

$$E(\beta^*) = \dots = \underline{\underline{\beta}} \quad \text{väntevärdesriktig skattning (se appendix för härledning)}$$

$$V(\beta^*) = \dots = \underline{\underline{\frac{\sigma^2}{S_{xx}}}} \quad \text{konsistent, eftersom } S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 \rightarrow \infty \text{ när } n \rightarrow \infty$$

$$E(\mu_0^*) = E(\alpha^* + \beta^* \cdot x_0) = E\left(\sum_{i=1}^n (d_i + c_i \cdot x_0) \cdot Y_i\right) = \dots = \alpha + \beta x_0 = \underline{\underline{\mu_0}}$$

$$\begin{aligned}
 V(\mu_0^*) &= V(\alpha^* + \beta^* \cdot x_0) = V\left(\sum_{i=1}^n \left(\frac{1}{n} + c_i \cdot (x_0 - \bar{x})\right) \cdot Y_i\right) \quad \text{mit } d_i = \frac{1}{n} - c_i \cdot \bar{x} \\
 &= \dots = \underline{\underline{\sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}} \quad (\text{variansen är stor när } x_0 \text{ ligger långt ifrån } \bar{x})
 \end{aligned}$$

Väntevärde och varians för α^* följer från uttrycket för μ_0^* genom att sätta $x_0 = 0$.

Uwe Menzel, 2017

Fördelningarna för de skattade variablerna

Vi vet nu att $\alpha^*, \beta^*, \mu_0^*$ är normalfördelade, och vi vet vilka väntevärden och vilka standardavvikelser de har. Det betyder att vi känner till deras fördelning:

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

fördelning för skattningen av
lutningen (om $\varepsilon_i \sim N$)

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

fördelning för skattningen av
interceptet (om $\varepsilon_i \sim N$)

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

fördelning för skattningen av
målvariabeln (om $\varepsilon_i \sim N$)
för $x_0 = 0$ blir detta fördelningen för α

Alla skattningar är väntevärdesriktiga och konsistenta.

www.matstat.org

Konfidensintervall för β

Vi vill räkna ut ett intervall som omfattar det sanna värdet β med en viss sannolikhet (oftast 95%) = **konfidensintervall**.

Vi utnyttjar fördelningen för β^* , som vi har fått ovan:

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

fördelning för skattningen
av lutningen

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

Om σ var känt hade detta varit en
referensvariabel.

Om σ inte är känt **ersätts σ med respektive skattning s_r** . Som vi har sett förut (t. ex. föreläsning intervallskattning, **F9**) måste därför fördelningen ändras:

$$\sigma \rightarrow s_r$$

$$\frac{\beta^* - \beta}{s_r/\sqrt{S_{xx}}} \sim t(n - 2)$$

Om σ är okänt är detta en referensvariabel (som följer
t-fördelningen med $n - 2$ frihetsgrader).

www.matstat.org

Konfidensintervall för β

$$\frac{\beta^* - \beta}{s_r / \sqrt{S_{xx}}} \sim t(n-2) \quad \text{referensvariabel} \quad \sigma \rightarrow s_r$$

1. Hitta en referensvariabel
2. Stäng in referensvariabeln mellan kvantilgränserna
3. Stäng in parametern som ska skattas mellan kända storheter genom att omforma i parentesen $P(\dots)$

$$I_\beta = \beta^* \pm t_{\alpha/2}(n-2) \cdot \frac{s_r}{\sqrt{S_{xx}}}$$

$(1 - \alpha) \cdot 100\%$ **konfidensintervall** för lutningen (när σ är okänd)

punktskattning kvantil medelfel

Analogt kan vi härleda ett konfidensintervall för interceptet α :

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

$(1 - \alpha) \cdot 100\%$ KI för interceptet (σ okänd)

Lutningen är viktigast: om lutningen inte är noll (dvs. om KI:et inte går över noll) betyder det att det finns ett samband mellan x och y .

www.matstat.org

Exempel, KI för β

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 :

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

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

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

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

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

www.matstat.org

Konfidensintervall för μ_0

$$\frac{\mu_0^* - \mu_0}{d} \sim t(n-2) \quad \text{referensvariabel wenn } \sigma \rightarrow s_r$$

$$\text{där } d = s_r \cdot \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \quad \text{medelfel}$$

1. Hitta en **referensvariabel**
2. Stäng in referensvariabeln mellan kvantilgränserna
3. Stäng in parametern som ska skattas mellan kända storheter genom att omforma i parentesen $P(\dots)$

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

$(1 - \alpha) \cdot 100\%$ konfidensintervall för
målvariabeln (σ okänd)

punktskattning

kvantil

medelfel

www.matstat.org

Sammanfattning: KI för regression

a) σ känd

www.matstat.org

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

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

$$I_{\alpha} = \alpha^* \pm \lambda_{\alpha/2} \cdot \sigma \cdot \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}$$

$$S_{xx} = \sum_{i=1}^n x_i^2 - n\bar{x}^2$$

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

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

$$s_r = \sqrt{\frac{Q_0}{n-2}}$$

b) σ känd

$$I_{\beta} = \beta^* \pm t_{\alpha/2}(n-2) \cdot \frac{s_r}{\sqrt{S_{xx}}}$$

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

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

fas från I_{μ_0} för $x_0 = 0$

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

Uwe Menzel, 2017

Exempel, KI för μ_0

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 μ_0 när $x_0 = 0.2$:

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

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

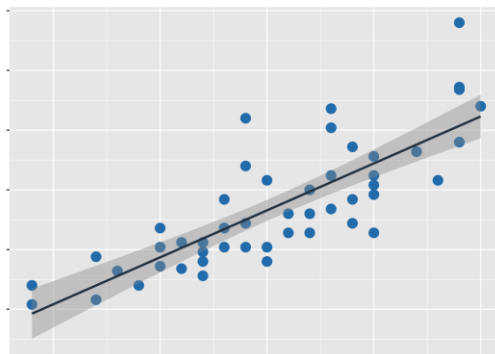
$$I_{\mu_0} = \mu_0^* \pm t_{\alpha/2}(n-2) \cdot d = 129 \pm t_{0.025}(18) \cdot 8.38 = \underline{\underline{(110, 150)}}$$

www.matstat.org

Konfidensintervall för μ_0

$$I_{\mu_0} = \mu_0^* \pm t_{\alpha/2}(n-2) \cdot s_r \cdot \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \quad (1 - \alpha) \cdot 100\% \text{ KI för målvariabeln}$$

x_0 är den punkt på x -axeln för den vi vill ha ett KI för målvariabeln. Ovanstående formel visar att konfidensintervallet blir bredare om x_0 är längre bort från $\bar{x}$ (från medelvärde av alla x -värden), eftersom medelfelet blir större. Konfidensintervallet har därför följande (typisk) form:



- längre bort från $\bar{x}$ blir skattningarna mera osäkra (KI:et blir bredare)
- en extrapolation långt bort från alla x_i är därför mycket osäker

Bild från
<https://rpubs.com/aaronsc32/regression-confidence-prediction-intervals>

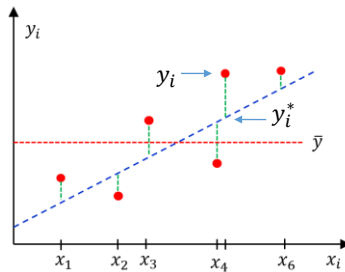
www.matstat.org

Förklaringsgrad (R^2)

Hur bra "fungerade" regressionslinjen ?

$$R^2 = \frac{SS_{reg}}{SS_{tot}} = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum_i (y_i - y_i^*)^2}{\sum_i (y_i - \bar{y})^2} = \frac{S_{xy}^2}{S_{xx} \cdot S_{yy}}$$

Förklaringsgrad,
definition



$$SS_{tot} = \sum_i (y_i - \bar{y})^2 \quad \text{total "Sum of Squares"}$$

$$SS_{reg} = \sum_i (y_i^* - \bar{y})^2 \quad \text{"Sum of Squares" för regression}$$

$$SS_{res} = \sum_i (y_i - y_i^*)^2 \quad \text{"Sum of Squares" för residuer}$$

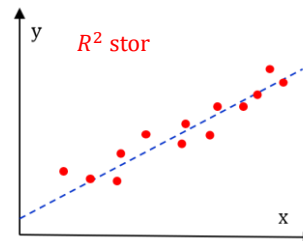
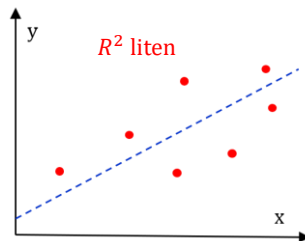
$$SS_{tot} = SS_{reg} + SS_{res}$$

www.matstat.org

Förklaringsgrad (R^2)

Hur bra "fungerade" regressionslinjen ?

$$R^2 = \frac{SS_{reg}}{SS_{tot}} = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum_i (y_i - y_i^*)^2}{\sum_i (y_i - \bar{y})^2} = \frac{S_{xy}^2}{S_{xx} \cdot S_{yy}}$$



dålig anpassning $\longrightarrow 0 \leq R^2 \leq 1 \longleftarrow$ bra anpassning

$$R^2 = r_{xy}^2 \quad (\text{för enkel linjär regression})$$

$$r_{xy} = \frac{c_{xy}}{S_x \cdot S_y} \quad \begin{array}{l} \text{korrelationskoefficient} \\ c_{xy} = \text{kovarians} \end{array}$$

www.matstat.org

Icke-linjärt sammanhang mellan x och y

- vi kan ändå använda linjär regression
- med "linjär" menas att koefficienterna α och β uppträder linjärt i modellformeln

Exempel 1: kvadratisk samband

$$y_i = \alpha + \beta \cdot \underbrace{x_i^2}_{z_i = x_i^2} + \varepsilon_i \quad \begin{array}{l} \text{kvadratisk term för } x \\ x_i \text{ är ju kända, } z_i = \text{"nya" mätvärden} \end{array}$$

$$y_i = \alpha + \beta \cdot z_i + \varepsilon_i \quad \text{vanlig linjär regression} \rightarrow \alpha^*, \beta^* \text{ beräknas}$$

$$\Rightarrow \mu_0^* = \alpha^* + \beta^* \cdot z_0 = \alpha^* + \beta^* \cdot x_0^2 \quad \text{parabel } \mu_0(x_0)$$

Exempel 2: exponentiellt samband

$$y_i = \alpha + \beta \cdot \underbrace{e^{-x_i^2}}_{z_i = e^{-x_i^2}} + \varepsilon_i$$



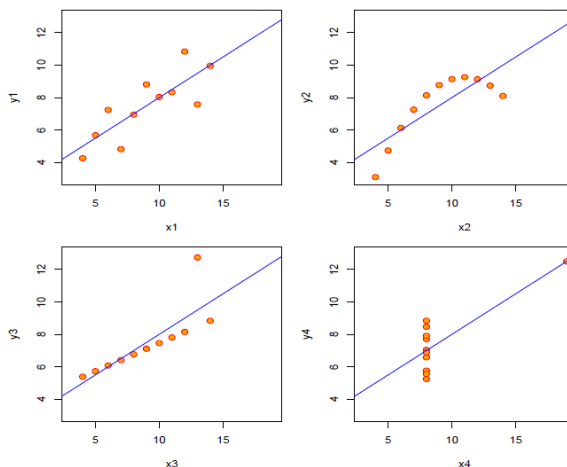
OBS!: I R kan man beräkna icke-linjära samband direkt: funktion `lm()`

www.matstat.org

Motivering för linjär regression

- det måste kollas om linjär regression kan användas
- "rent tekniskt" fungerar det (nästan) alltid att räkna ut en regressionslinje (dvs. att räkna ut α^* och β^*)

Anscombe's 4 Regression data sets

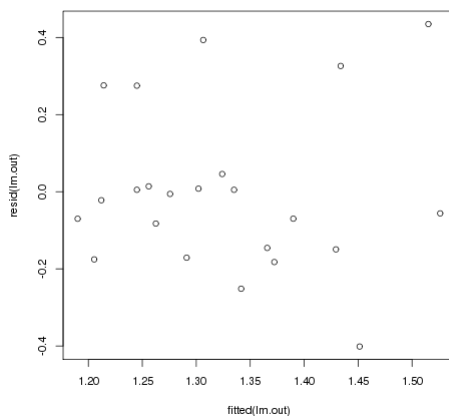


- Alla fyra stickprov ger samma regressionslinje!
- Det är därför viktigt att rita ett diagram!
- Dessutom borde värdet på R^2 inte vara för litet.

www.matstat.org

Motivering för linjär Regression

Vi antog att $\varepsilon_i \sim N(0, \sigma)$, dvs. att alla ε_i har samma varians $V(\varepsilon_i) = \sigma^2$. Dessutom borde väntevärdet $E(\varepsilon_i)$ vara noll enligt modellen. För att (uppskattningsvis) kolla om detta stämmer kan man rita ε_i mot x -värden:



Bruset ε_i ("residuer") borde skingras runt noll, och inte bli systematiskt större eller mindre med x (borde t. ex inte se ut som en strut)

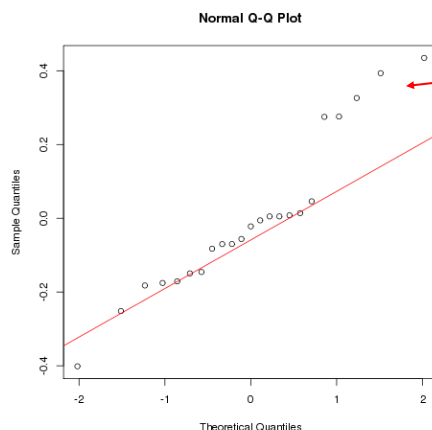


```
residuals = resid(lm.out)
plot(residuals, x)
```

www.matstat.org

Motivering för linjär regression

Vi antog att $\varepsilon_i \sim N(0, \sigma)$, dvs. att bruset (ε_i) är normalfördelat. För att kolla om detta stämmer kan man rita en så kallad **QQ-plot** med residuerna:



inte bra!

flertalet av punkterna borde ligga på den röda linjen

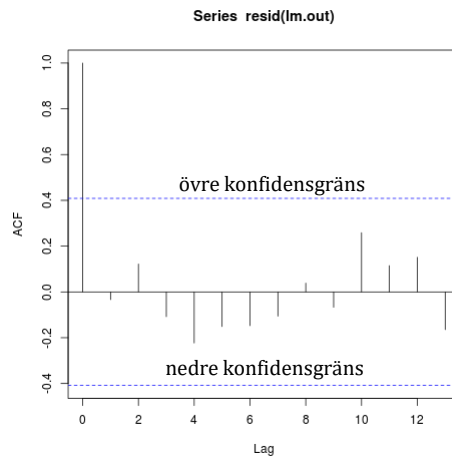


```
residuals = resid(lm.out)
qqnorm(residuals)
qqline(residuals)
```

www.matstat.org

Motivering för linjär regression

Vi antog att brusets värden ε_i är oberoende. För att kolla om detta stämmer kan man rita ett autokorrelations-diagram:



Flertalet av värden på autokorrelation (helst alla utom värdet för "Lag = 0") borde vara små, dvs. ligga innanför konfidensgränserna (blå linje -----)



```
auto.cor = acf(resid(lm.out))
plot(auto.cor)
```

www.matstat.org

Appendix

Linjär Regression

Uwe Menzel, 2018

uwe.menzel@matstat.org

www.matstat.org

Skattningen β^* är en linjärkombination av y_i

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

Uwe Menzel, 2017

Skattningen α^* är en linjärkombination av y_i

$$\begin{aligned}
 \alpha^* &= \bar{y} - \beta^* \bar{x} = \frac{1}{n} \sum_{i=1}^n y_i - \bar{x} \sum_{i=1}^n c_i \cdot y_i \\
 &= \sum_{i=1}^n \left(\frac{1}{n} - c_i \bar{x} \right) y_i \\
 &= \sum_{i=1}^n d_i \cdot y_i \quad \text{med} \quad d_i = \frac{1}{n} - c_i \bar{x} \quad (d_i \text{ är inte några slumpvariabler})
 \end{aligned}$$

www.matstat.org

Skattningen β^* är väntevärdesriktig

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

Uwe Menzel, 2017

Skattningen β^* är konsistent

$$\begin{aligned}
 V(\beta^*) &= V\left(\sum_{i=1}^n c_i y_i\right) = \sum_{i=1}^n V(c_i y_i) \quad \text{pga. oberoendet !} \\
 &= \sum_{i=1}^n c_i^2 V(y_i) = \sum_{i=1}^n c_i^2 \sigma^2 \\
 &= \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{S_{xx}}\right)^2 \sigma^2 \\
 &= \frac{\sigma^2}{S_{xx}^2} \sum_{i=1}^n (x_i - \bar{x})^2 \\
 &= \frac{\sigma^2}{S_{xx}} \quad \text{för att } S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 \\
 S_{xx} &= \sum_{i=1}^n (x_i - \bar{x})^2 \rightarrow \infty \text{ när } n \rightarrow \infty
 \end{aligned}$$

Variansen för β^* är liten om x - värdena är "utspridda" ($\rightarrow$ stor S_{xx}) !

www.matstat.org

Väntevärde och varians för skattningen för μ_0

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

$$\begin{aligned} E(\mu_0^*) &= E(\alpha^* + \beta^* \cdot x_0) = E\left(\sum_{i=1}^n (d_i + c_i x_0) Y_i\right) \\ &= \dots \\ &= \alpha + \beta x_0 = \mu_0 \quad \text{väntevärdesriktigt!} \end{aligned}$$

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

www.matstat.org

Referensvariabel för beräkningen av KI:et för β

$$\beta^* \sim N\left(\beta, \frac{\sigma}{\sqrt{S_{xx}}}\right) \Rightarrow Z = \frac{\beta^* - \beta}{\sigma/\sqrt{S_{xx}}} \sim N(0, 1)$$

$$\varepsilon_i \sim N(0, \sigma) \quad V(\varepsilon_i) = \sigma^2 \quad V\left(\frac{\varepsilon_i}{\sigma}\right) = \frac{1}{\sigma^2} \cdot V(\varepsilon_i) = 1 \Rightarrow \frac{\varepsilon_i}{\sigma} \sim N(0, 1)$$

$$Q_0 = \sum_{i=1}^n \varepsilon_i^2 \Rightarrow \frac{Q_0}{\sigma^2} = \sum_{i=1}^n \left(\frac{\varepsilon_i}{\sigma}\right)^2 \sim \chi^2(n-2)$$

$$\frac{Z}{\sqrt{\frac{\chi^2(n)}{n}}} \sim t(n) \quad \text{gäller allmänt för alla } n \Rightarrow \frac{Z}{\sqrt{\frac{\chi^2(n-2)}{n-2}}} \sim t(n-2)$$

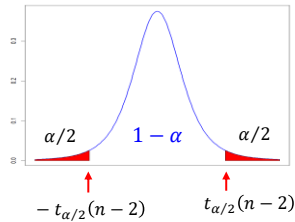
$$\frac{Q_0}{\sigma^2} \sim \chi^2(n-2) \Rightarrow \frac{Z}{\sqrt{\frac{Q_0}{\sigma^2(n-2)}}} \sim t(n-2) \quad \text{mata in ovanstående uttrycket för } Z \rightarrow$$

$$\frac{\frac{\beta^* - \beta}{\sigma/\sqrt{S_{xx}}}}{\sqrt{\frac{Q_0}{\sigma^2(n-2)}}} \sim t(n-2) \Rightarrow \boxed{\frac{\beta^* - \beta}{s_r/\sqrt{S_{xx}}} \sim t(n-2)} \quad \text{med } s_r = \sqrt{Q_0/(n-2)}$$

Uwe Menzel, 2017

Konfidensintervall för β

$$\frac{\beta^* - \beta}{s_r / \sqrt{S_{xx}}} \sim t(n-2) \quad \text{referensvariabel}$$



$$P\left(-t_{\alpha/2}(n-2) < \frac{\beta^* - \beta}{\frac{s_r}{\sqrt{S_{xx}}}} \leq +t_{\alpha/2}(n-2)\right) = 1 - \alpha$$

stäng in referensvariabeln mellan
kvantilgränserna (se bild)
→ omforma i parantesen $P(\dots)$

$$P\left(\beta^* - t_{\alpha/2}(n-2) \cdot \frac{s_r}{\sqrt{S_{xx}}} < \beta \leq \beta^* + t_{\alpha/2}(n-2) \cdot \frac{s_r}{\sqrt{S_{xx}}}\right) = 1 - \alpha$$

$$I_\beta = \beta^* \pm t_{\alpha/2}(n-2) \cdot \frac{s_r}{\sqrt{S_{xx}}} \quad \text{konfidensintervall för } \beta$$

Uwe Menzel, 2017

Sammanhang mellan R^2 och Pearsons korrelationskoefficient

$$R^2 = \frac{S_{xy}^2}{S_{xx} \cdot S_{yy}} \quad \text{Förklaringsgrad}$$

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} \quad \text{Pearson-korrelation}$$

Förklaringsgrad R^2 och Pearsons korrelationskoefficient är detsamma för enkel linjär regression (när vi bara har en oberoende variabel).

www.matstat.org