Grundlagen der Mathematischen Statistik

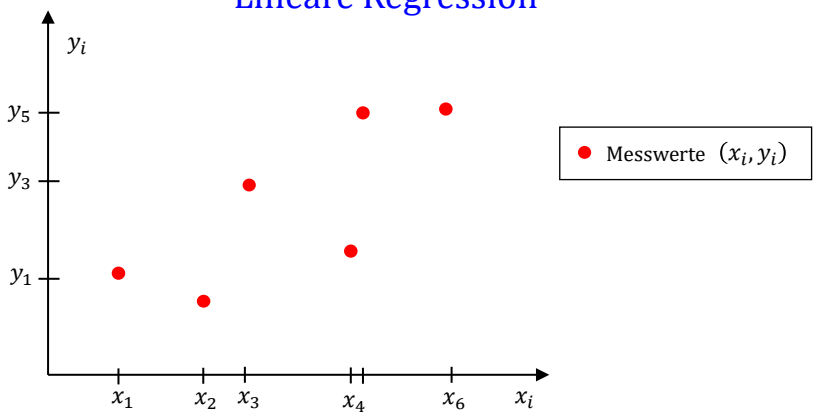
Lineare Regression

Uwe Menzel, 2018

uwe.menzel@matstat.org

www.matstat.org

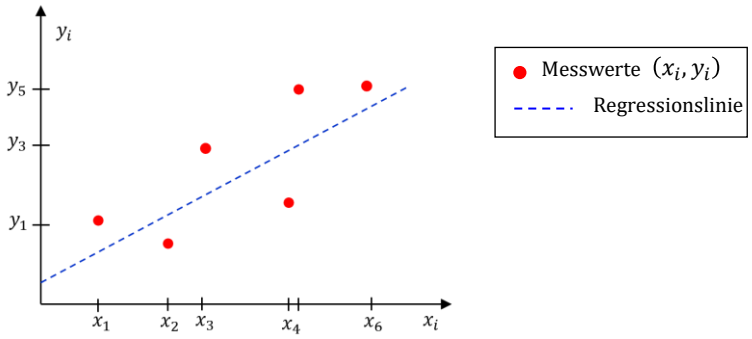
Lineare Regression



$\boldsymbol{x_i}$: unabhängige Variablen (Prediktoren, Kovariate)
- bekannt, keine Zufallsvariablen!

$\boldsymbol{y_i}$: abhängige Variablen (Zielvariablen, Responsvariablen)
- behaftet mit Rauschen (Messung), Zufallsvariablen

Lineare Regression

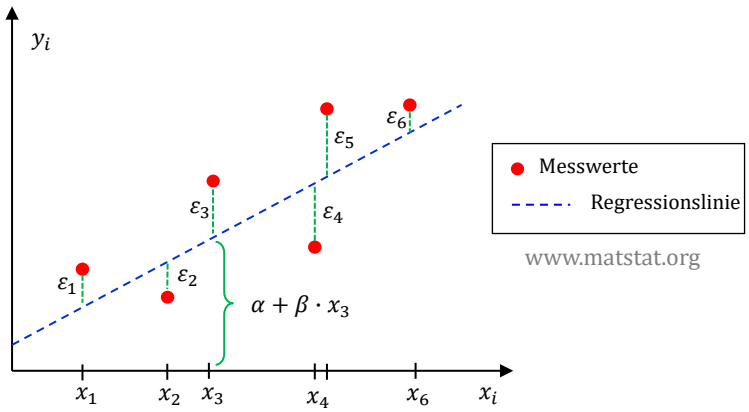


Wir nehmen an, dass es einen linearen Zusammenhang zwischen x und y gibt:

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

- α und β sind unbekannt, wir suchen Schätzungen für beide Parameter
- die Schätzungen werden auf der Basis einer Stichprobe gemacht
- Stichprobe = Messung von Zahlenpaaren (x_i, y_i)

Modell für den Rauschterm ε



Modell für den Rauschterm ε:

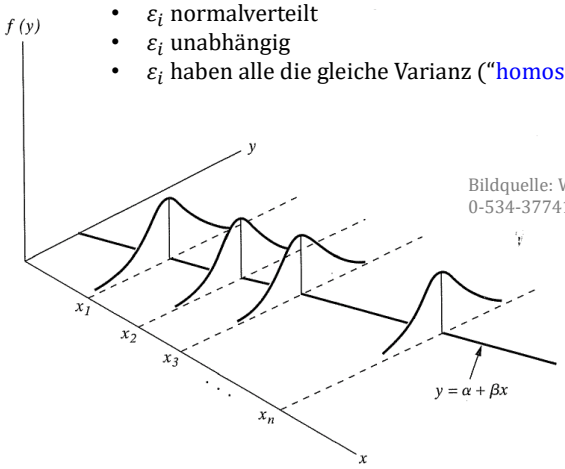
$y_i = \alpha + \beta \cdot x_i + \varepsilon_i \quad i = 1, 2, \cdots, n \quad n$ = Anzahl Messungen (Stichprobengröße)

ε_i = "Rauschen" $\quad \boxed{\varepsilon_i \sim N(0, \sigma)} \quad E(\varepsilon_i) = 0 \quad V(\varepsilon_i) = \sigma^2$

Modell den Rauschterm

Annahmen:

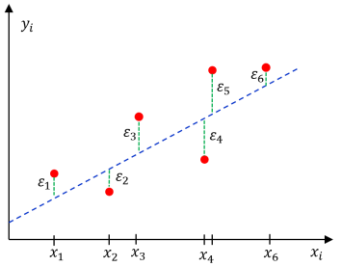
- ε_i normalverteilt
- ε_i unabhängig
- ε_i haben alle die gleiche Varianz ("homoscedasticity")



Bildquelle: Wackerly et al., ISBN 0-534-37741-6

www.matstat.org

Berechnung der Regressionslinie



Methode der kleinsten Quadrate:
(least squares method):

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

Idee

Q : Quadratsumme
("sum of squares")

- Regressionslinie (------) soll so gelegt werden, dass Q minimiert wird
- es wird das Quadrat der ε_i benutzt um auszuschließen, dass sich positive und negative Terme aufheben

Berechnung von α und β zur Minimierung von Q :

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

Uwe Menzel, 2017

Berechnung der Regressionslinie

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

$$\Rightarrow\quad \begin{aligned}\frac{\partial Q}{\partial \alpha} &= -2\sum_{i=1}^{n}(y_i-\alpha-\beta x_i)=0\\ \frac{\partial Q}{\partial \beta} &= -2\sum_{i=1}^{n}x_i(y_i-\alpha-\beta x_i)=0\end{aligned}$$

setze partielle Ableitungen Null

$$\Rightarrow\quad \begin{aligned}\sum_{i=1}^{n}y_i-n\cdot\alpha-\beta\sum_{i=1}^{n}x_i&=0\\ \sum_{i=1}^{n}y_i\cdot x_i-\alpha\cdot\sum_{i=1}^{n}x_i-\beta\sum_{i=1}^{n}x_i^2&=0\end{aligned}$$

lineare Gleichungen für α und β (alle anderen Größen sind aus der Stichprobe bekannt)
2 Gleichungen, 2 Variablen ✓

Uwe Menzel, 2017

Berechnung der Regressionslinie

Lösungen: $\boxed{\beta^*=\frac{S_{xy}}{S_{xx}}}$ $\boxed{\alpha^*=\bar{y}-\beta^*\bar{x}}$ Schätzungen für α und β

(mit Gleichung 1)

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

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

Achtung!: α^*und β^* sind Zufallsvariablen (beruhen auf einer Stichprobe) : neue Stichprobe ändert $\varepsilon_i \rightarrow y_i \rightarrow S_{xy} \rightarrow \alpha^*$

Für jedes x_0 erhält man jetzt eine **Schätzung der Responsvariablen** (x_0 muss nicht mit irgend einem der gemessenen x_i übereinstimmen):

$$\boxed{\mu_0^*=\alpha^*+\beta^*\cdot x_0}$$

www.matstat.org

Schätzung der Varianz des Rauschterms

Wir wollten Q minimieren. Wie groß wurde es?: $Q(\alpha, \beta) = \sum_i (y_i - \alpha - \beta \cdot x_i)^2$

Einsetzen der Werte welche Q minimieren, also α^* und β^*, ergibt:

$$Q_0 = \sum_{i=1}^{n} (y_i - \alpha^* - \beta^* x_i)^2 = \sum_{i=1}^{n} [y_i - \underbrace{\bar{y} + \beta^* \cdot \bar{x}}_{-\alpha^*} - \beta^* x_i]^2$$

$$= \sum_{i=1}^{n} [(y_i - \bar{y}) - \beta^* (x_i - \bar{x})]^2 = \ldots = S_{yy} - S_{xy}^2 / S_{xx}$$

(vollständig bekannt aus der Stichprobe)

Man kann die Maximum-Likelihood-Methode anwenden, um eine Schätzung für σ^2 zu finden. Diese liefert Q_0/n. Diese Schätzung muss jedoch korrigiert werden, um erwartungstreu zu werden:

$$s_r^2 = \frac{Q_0}{n-2} \qquad E\left(s_r^2\right) = \sigma^2$$

s_r^2 ist eine erwartungstreue Schätzung für σ^2 (und s_r eine erwartungstreue Schätzung für σ)

$(\sigma^2)^* = s_r^2 \qquad \sigma^* = s_r$ Schätzung für Varianz/Standardabw. des Rauschterms ε

Diese Formel ist anschaulich: je größer $Q = \sum_i \varepsilon_i^2$, desto größer σ (siehe Bild oben).

Berechnung der Regressionslinie

Beispiel: Bilirubingehalt (x) und Proteinkonzentration (y) in der Rückenmarksflüssigkeit von Neugeborenen

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

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

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

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

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

$\sum x_i y_i = 259.79$

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

Beispiel, Fortsetzung

Beispiel: Bilirubingehalt (x) und Proteinkonzentration (y) in der Rückenmarksflüssigkeit von Neugeborenen

$$\beta^* = \frac{S_{xy}}{S_{xx}} = \frac{43.1455}{0.1215} = 355.2$$ Anstieg

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

Schätzung für die Responsvariable μ:

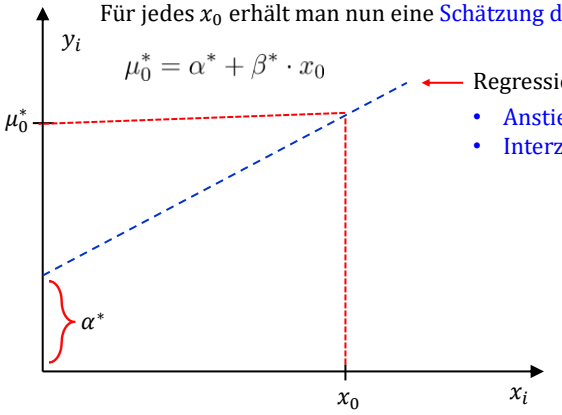
auch für x, die nicht gemessen wurden

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

Für jedes x kann nun die erwartete Konzentration berechnet werden.

Berechnung der Regressionslinie

Für jedes x_0 erhält man nun eine Schätzung der Responsvariablen:



Regressionslinie:

- Anstieg: β^*
- Interzept: α^* (y-Achsenabschnitt)

- x_0 kann auch zwischen den ursprünglich gemessenen x_i-Werten liegen
- α^* und β^* sind Zufallsvariablen → μ_0 ist eine Zufallsvariable (verändert sich wenn eine neue Stichprobe genommen wird).

Verteilung der Schätzer

Wir haben gesehen dass $\alpha^*, \beta^*, \mu_0^*, \ldots$ Zufallsvariablen sind.
Welche Verteilung haben sie ?

$$y_i = \alpha + \beta \cdot x_i + \varepsilon_i \qquad i = 1, 2, \cdots, n$$

$y_i \sim N$ $\Leftarrow$ $\varepsilon_i \sim N$

Die Responsvariable y_i ist normalverteilt, weil der Rauschterm ε_i normalverteilt ist.

Man kann ausrechnen (etwas umständlich, siehe Anhang):

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

$$\alpha^* = \sum_{i=1}^{n} d_i \cdot y_i \quad \text{mit} \quad d_i = \frac{1}{n} - c_i \bar{x}$$

α^* und β^* sind Linearekombinationen von normalverteilten Zufallsvariablen (y_i)

$\Rightarrow$ $\alpha^*, \beta^*, \mu_0^* \sim N$

$\Rightarrow$ die Regressionsvariablen $\alpha^*, \beta^*, \mu_0^*$ sind normalverteilt (wenn der Rauschterm normalverteilt ist).

Verteilung der Schätzer

Die Regressionsvariablen $\alpha^*, \beta^*, \mu_0^*$ sind normalverteilt. Wir können deren Erwartungswert, Varianz und Standardabweichung berechnen:

$E(\beta^*) = \ldots = \beta$ erwartungstreue Schätzung (siehe Anhang für Herleitung)

$V(\beta^*) = \ldots = \frac{\sigma^2}{S_{xx}}$ konsistent, denn $S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 \to \infty$ *wenn* $n \to \infty$

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

$$V(\mu_0^*) = V(\alpha^* + \beta^* \cdot x_0) = V\left(\sum_{i=1}^{n} \left(\frac{1}{n} + c_i \cdot (x_0 - \bar{x})\right) \cdot Y_i\right) \qquad \text{mit } d_i = \tfrac{1}{n} - c_i \cdot \bar{x}$$

$$= \ldots = \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)$$

(die Varianz ist groß wenn x_0 weit weg von $\bar{x}$ liegt)

Erwartungswert und Varianz für α^* folgen aus der Formel für μ_0^* mit $x_0 = 0$.

Verteilung der Schätzer

Wir wissen nun, dass $\alpha^*, \beta^*, \mu_0^*$ normalverteilt sind, und wir kennen deren Erwartungswerte und Standardabweichungen. Das bedeutet, dass deren Verteilung vollständig bekannt ist:

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

Verteilung für Schätzung des Anstiegs (wenn $\varepsilon_i \sim N$)

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

Verteilung für Schätzung des Interzeptes (wenn $\varepsilon_i \sim N$)

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

Verteilung für Schätzung der Responsvariablen (wenn $\varepsilon_i \sim N$)
für $x_0 = 0$ wird dies die Verteilung für α

Alle Schätzer sind erwartungstreu und konsistent.

Konfidenzintervall für β

Wir wollen ein Intervall ausrechnen, welches den wahren Wert von β mit einer gewissen Wahrscheinlichkeit (oft 95%) einschließt = Konfidenzintervall.

Wir nutzen die Verteilung für β^* aus, die wir oben berechnet haben:

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

Verteilung für die Schätzung des Anstiegs der Regressionsgeraden

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

Wäre σ bekannt, so wäre dies eine Referenzvariable.

Wenn σ nicht bekannt ist, wird σ mit seiner Schätzung s_r ersetzt. Wie wir schon zuvor gesehen haben (z. B. Vorlesung Intervallschätzung, **F9**) verändert sich damit die Verteilung des Ausdrucks :

$\sigma \rightarrow s_r$

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

Wenn σ unbekannt ist, ist dies eine Referenzvariable (die t-verteilt mit f = $n - 2$ Freiheitsgraden ist).

Konfidenzintervall für β

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

1. Finde eine Referenzvariable
2. Schließe die Referenzvariable zwischen deren Quantilgrenzen ein
3. Schließe den zu schätzenden Parameter durch Umformen in der Klammer $P(\dots)$ zwischen bekannten Größen ein

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

$(1-\alpha)\cdot 100\%$ **Konfidenzintervall** für den Anstieg der Regressionlinie (wenn σ unbekannt ist)

Punktschätzer — Quantil — Standardfehler

Analog kann ein Konfidenzintervall für das Interzept α hergeleitet werden:

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

$(1-\alpha)\cdot 100\%$ KI für das Interzept (σ unbekannt)

Meist ist der Anstieg wichtiger: wenn der Anstieg verschieden von Null ist (d.h. wenn das KI nicht die Null enthält), dann existiert ein Zusammenhang zwischen x und y.

Beispiel, KI für β

Beispiel: Bilirubingehalt (x) und Proteinkonzentration (y)

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

Gesucht: 95% Konfidenzintervall für den Anstieg der Regressionslinie

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

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

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

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

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

Konfidenzintervall für μ_0

$$\frac{\mu_0^* - \mu_0}{d} \sim t\,(n-2) \qquad \text{Referensvariable} \qquad \sigma \to s_r$$

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

1. Finde eine Referenzvariable
2. Schließe die Referenzvariable zwischen deren Quantilgrenzen ein
3. Schließe den zu schätzenden Parameter durch Umformen in der Klammer $P(\dots)$ zwischen bekannten Größen ein

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

$(1-\alpha) \cdot 100\%$ Konfidenzintervall für die Responsvariable (wenn σ unbekannt)

Punktschätzer — Quantil — Standardfehler

www.matstat.org

Zusammenfassung: KI für Regression

a) σ bekannt

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

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

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

b) σ unbekannt

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

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

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

www.matstat.org

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

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

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

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

ergibt sich aus I_{μ_0} für $x_0 = 0$

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

Uwe Menzel, 2017

Beispiel, KI für μ_0

Beispiel: Bilirubingehalt (x) und Proteinkonzentration (y)

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

$s = 24.7$

Gesucht: 95% Konfidenzintervall für μ_0 für $x_0 = 0.2$

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

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

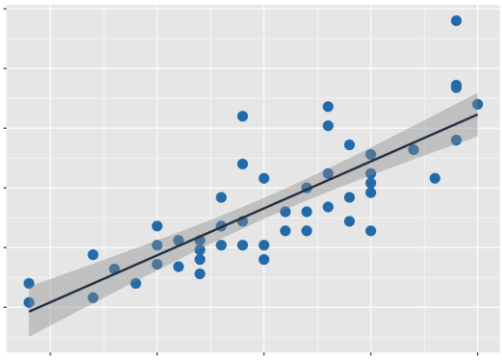
$$I_{\mu_0} = \mu^*_{obs} \pm t_{\alpha/2}(n-2) \cdot d = 129 \pm t_{0.025}(18) \cdot 8.38 = (110, 150)$$

Konfidenzintervall für μ_0

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

$(1 - \alpha) \cdot 100\%$ KI für die Responsvariable

x_0 ist der Punkt auf der x-Achse für den wir ein KI für die Responsvariable haben wollen. Die oben stehende Formel zeigt, dass das Konfidenzintervall desto breiter wird, je weiter der Punkt x_0 vom Mittelpunkt $\bar{x}$ aller x-Werte entfernt ist. Das Konfidenzintervall hat daher die folgende typische Form:



- weiter weg von $\bar{x}$ wird die Schätzung unsicherer (das KI wird breiter)
- eine Extrapolation weit weg von allen x_i ist daher sehr unsicher

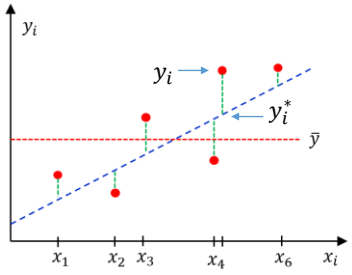
Bildquelle:
https://rpubs.com/aaronsc32/regression-confidence-prediction-intervals

Bestimmtheitsmaß (R^2)

Wie gut "funktioniert" die Regressionslinie ?

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

Bestimmtheits-maß, Definition



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

$$SS_{reg} = \sum_i (y_i^* - \bar{y})^2$$ Quadratsumme für Regression

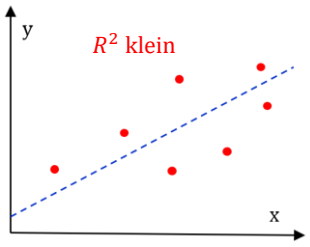
$$SS_{res} = \sum_i (y_i - y_i^*)^2$$ Quadratsumme für Residuen

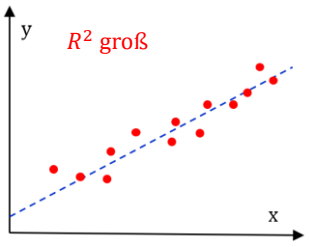
$$SS_{tot} = SS_{reg} + SS_{res}$$

Bestimmtheitsmaß (R^2)

Wie gut "funktioniert" die Regressionslinie ?

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





schlechte Anpassung → $0 \leq R^2 \leq 1$ ← gute Anpassung

$R^2 = r_{xy}^2$ (für einfache lineare Regression) $\quad r_{xy} = \frac{c_{xy}}{s_x \cdot s_y}$ Korrelationskoeffizient

c_{xy} = Kovarianz

Nichtlinearer Zusammenhang zwischen x und y

- es kann trotzdem lineare Regression angewendet werden
- mit "linear" ist gemeint, dass die Koeffizienten α und β linear im Modell auftreten !

Beispiel 1: Quadratischer Zusammenhang

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

quadratischer Term für x

x_i sind bekannt, z_i = "neue" Messwerte

$y_i = \alpha + \beta \cdot z_i + \varepsilon_i$ gewöhnliche lineare Regression → α^*, β^* berechnen

$\Rightarrow \mu_0^* = \alpha^* + \beta^* \cdot z_0 = \alpha^* + \beta^* \cdot x_0^2$ Parabel $\mu_0(x_0)$

Beispiel 2: Exponentieller Zusammenhang

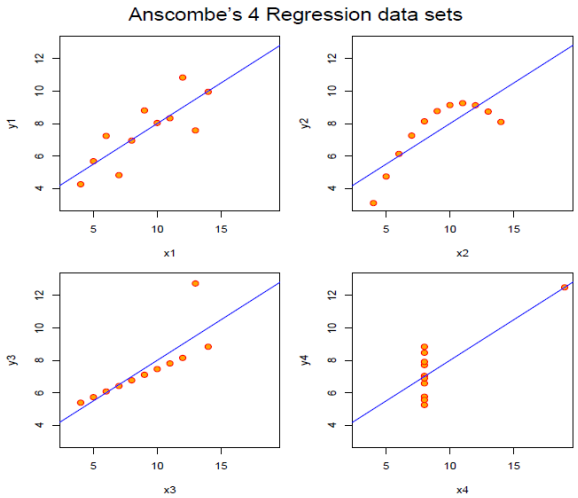
$y_i = \alpha + \beta \cdot \underbrace{e^{-x_i^2}}_{z_i = e^{-x_i^2}} + \varepsilon_i$

!!: In R können nicht-lineare Ausdrücke direkt formuliert werden: Funktion lm()

www.matstat.org

Validierung der durchgeführten Regression

- es muss kontrolliert werden, ob lineare Regression angewendet werden kann
- "rein technisch" funktioniert es (fast) immer, eine Regressionslinie auszurechnen (d. h. α^*und β^* auszurechnen)

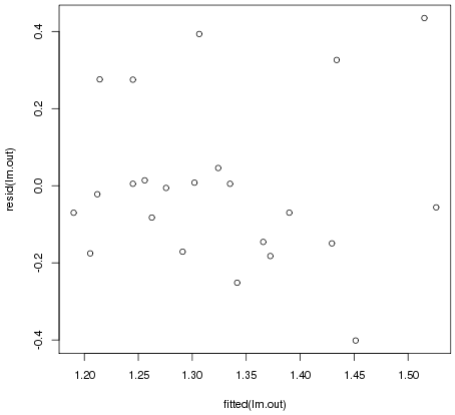


- Alle vier Stichproben ergeben die gleiche Regressionslinie!
- Es ist daher wichtig, immer eine Grafik anzufertigen!
- Außerdem sollte der Wert von R^2 nicht zu klein sein.

www.matstat.org

Validierung der durchgeführten Regression

Unser Modell setzt voraus, dass $\varepsilon_i \sim N(0, \sigma)$, d. h. dass alle ε_i die gleiche Varianz haben: $V(\varepsilon_i) = \sigma^2$. Außerdem soll der Erwartungswert der ε_i Null sein. Um dies (grob) zu überprüfen, kann man die ε_i über x plotten:

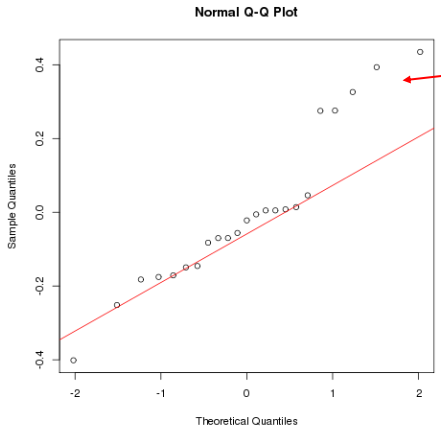


Die Werte ε_i des Rauschterms ("Residuen") sollten um Null verteilt liegen und sollten sich nicht systematisch mit x vergrößern oder verkleinern (es sollte sich z. B. kein trichterförmiges Bild ergeben).

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


Validierung der durchgeführten Regression

Wir haben angenommen, dass $\varepsilon_i \sim N(0, \sigma)$, d. h. dass der Rauschterm (ε_i) normalverteilt ist. Um zu überprüfen, ob dies stimmt, kann man einen sog. QQ-plot mit den Residuen zeichnen:



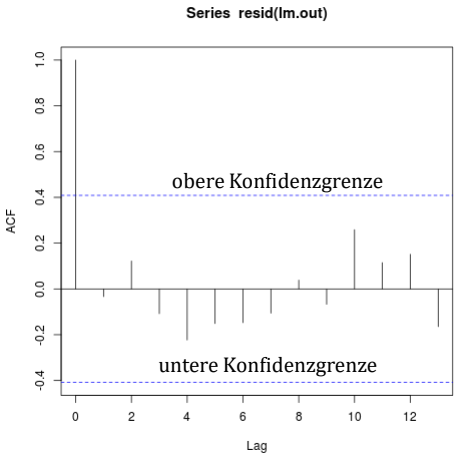
nicht gut!

Die Mehrzahl der Punkte sollte auf der roten Geraden liegen.

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

Validierung der durchgeführten Regression

Wir haben angenommen, dass die Werte des Rauschterms ε_i unabhängig sind. Um zu kontrollieren, ob dies stimmt, kann man die Autokorrelation der ε_i zeichnen:



Die Mehrzahl der Werte der Autokorrelation (möglichst alle bis auf den Wert für "$Lag = 0$") sollte klein sein, d. h. innerhalb der Konfidenzgrenzen liegen (blaue Linie -------)

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


Anhang

Lineare Regression

Uwe Menzel, 2018

uwe.menzel@matstat.org

www.matstat.org

Der Schätzer β^* ist eine Linearkombination der y_i

$$
\begin{aligned}
\beta^* &= \frac{1}{S_{xx}} S_{xy} = \frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \bar{x})\,(y_i - \bar{y}) \\
&= \frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i y_i - \bar{x} y_i - x_i \bar{y} + \bar{x}\bar{y}) \\
&= \frac{1}{S_{xx}} \left[\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} \bar{x} y_i - n\bar{x}\bar{y} + n\bar{x}\bar{y} \right] \\
&= \frac{1}{S_{xx}} \left[\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} \bar{x} y_i \right] \\
&= \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{S_{xx}} \cdot y_i \right) \\
&= \sum_{i=1}^{n} c_i \cdot y_i \quad \text{mit} \quad c_i = \frac{x_i - \bar{x}}{S_{xx}} \quad \text{(die } c_i \text{ sind keine Zufallsvariablen)}
\end{aligned}
$$

Der Schätzer α^* ist eine Linearekombination der y_i

$$
\begin{aligned}
\alpha^* &= \bar{y} - \beta^* \bar{x} = \frac{1}{n} \sum_{i=1}^{n} y_i - \bar{x} \sum_{i=1}^{n} c_i \cdot y_i \\
&= \sum_{i=1}^{n} \left(\frac{1}{n} - c_i \bar{x} \right) y_i \\
&= \sum_{i=1}^{n} d_i \cdot y_i \quad \text{mit} \quad d_i = \frac{1}{n} - c_i \bar{x} \quad \text{(die } d_i \text{ sind keine Zufallsvariablen)}
\end{aligned}
$$

Der Schätzer β^* ist erwartungstreu

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

Uwe Menzel, 2017

Der Schätzer β^* ist konsistent

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

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

Die Varianz für β^* ist klein wenn die x - Werte "ausgebreitet" sind (→ großes S_{xx}) !

www.matstat.org

Erwartungswert und Varianz für den Schätzer μ_0

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

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

erwartungstreu !

$$V(\mu_0^*) = V(\alpha^* + \beta^* \cdot x_0) = V\left[\sum_{i=1}^{n} \left(\frac{1}{n} + c_i \cdot (x_0 - \bar{x})\right) \cdot Y_i\right] = \dots$$

$$\dots = \sigma^2 \cdot \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)$$

Referensvariabel für die Berechnung eines KI für β

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

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

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

$$\frac{Z}{\sqrt{\frac{\chi^2(n)}{n}}} \sim t(n) \quad \text{gilt allgemein für alle } n \quad \Rightarrow \quad \frac{Z}{\sqrt{\frac{\chi^2(n-2)}{n-2}}} \sim t(n-2)$$

$$\frac{Q_0}{\sigma^2} \sim \chi^2(n-2) \quad \Rightarrow \quad \frac{Z}{\sqrt{\frac{Q_0}{\sigma^2(n-2)}}} \sim t(n-2)$$

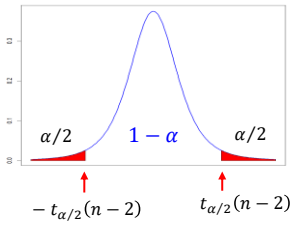
setze nun obigen Ausdruck für Z ein

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

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

Konfidenzintervall für β

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



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

Einschließen zwischen Quantilen (siehe Abb.)
→ Umformen in der Klammer $P(\ldots)$

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

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

Uwe Menzel, 2017

Zusammenhang zwischen R^2 und Pearsons Korrelations-Koeffizient

$$R^2 = \frac{S_{xy}^2}{S_{xx}\cdot S_{yy}}$$ Bestimmtheitsmaß

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

Bestimmtheitsmaß R^2 und Pearsons Korrelations-Koeffizient sind für einfache lineare Regression (wenn wir nur eine unabhängige Variable haben) gleich.

www.matstat.org