Anmerkungen zur Vorlesung Statistik vom 08.11.2016

Inhalt:

- (einfache lineare) Regressionsanalyse
  - Regressionsmodell
  - Regressionskoeffizienten, Regressionsgerade
  - Gütemaß: Determinationskoeffizient/Bestimmtheitsmaß R^2
  - R: Umsetzung, Darstellung (Plot) und Outputinterpretation
  - Residualanalyse
- Klausur SS2016: Aufgabe 2 (WDH Korrelation / Regression)

Aufgabensammlung Statistik:

25.10.2016 Hausaufgabe A23-A31 ohne Regressions- aspekte
08.11.2016 Hausaufgabe A27-A31 Regressions- aspekte A32-A33

Ausgangsdaten

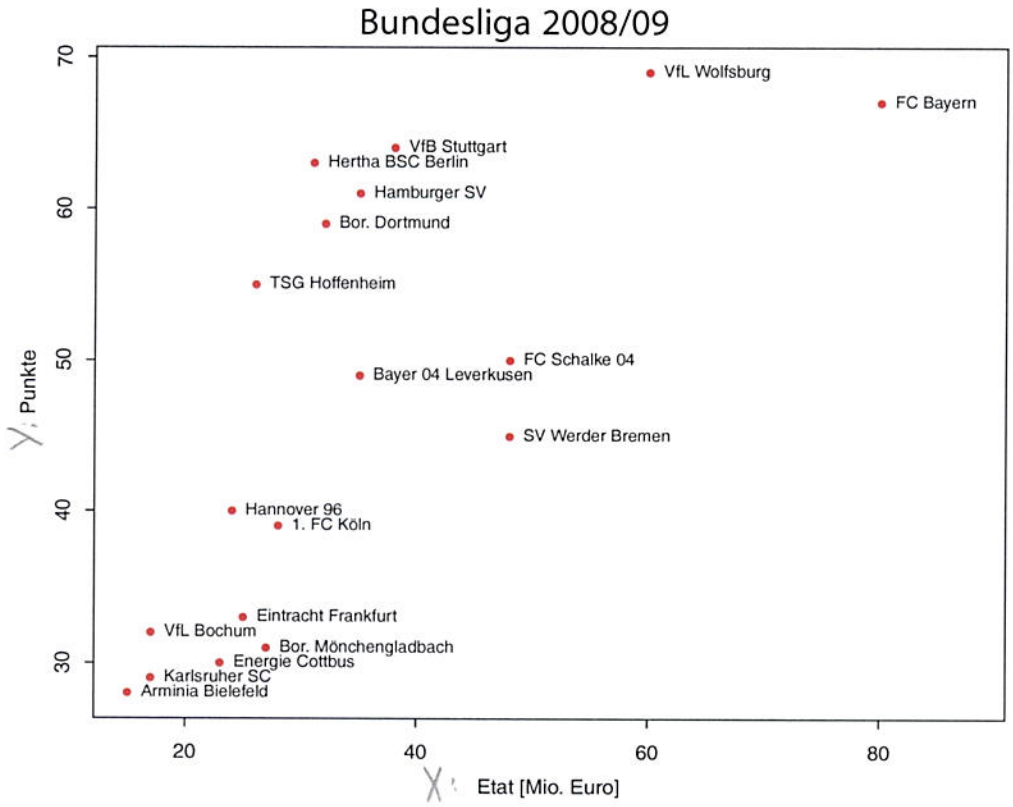
Bundesliga 2008/2009

- Gegeben: Daten zu den 18 Vereinen der ersten Bundesliga in der Saison 2008/09
- Merkmale: **Vereinssetat** für Saison (nur direkte Gehälter und Spielergehälter)
- und **Ergebnispunkte** in Tabelle am Ende der Saison

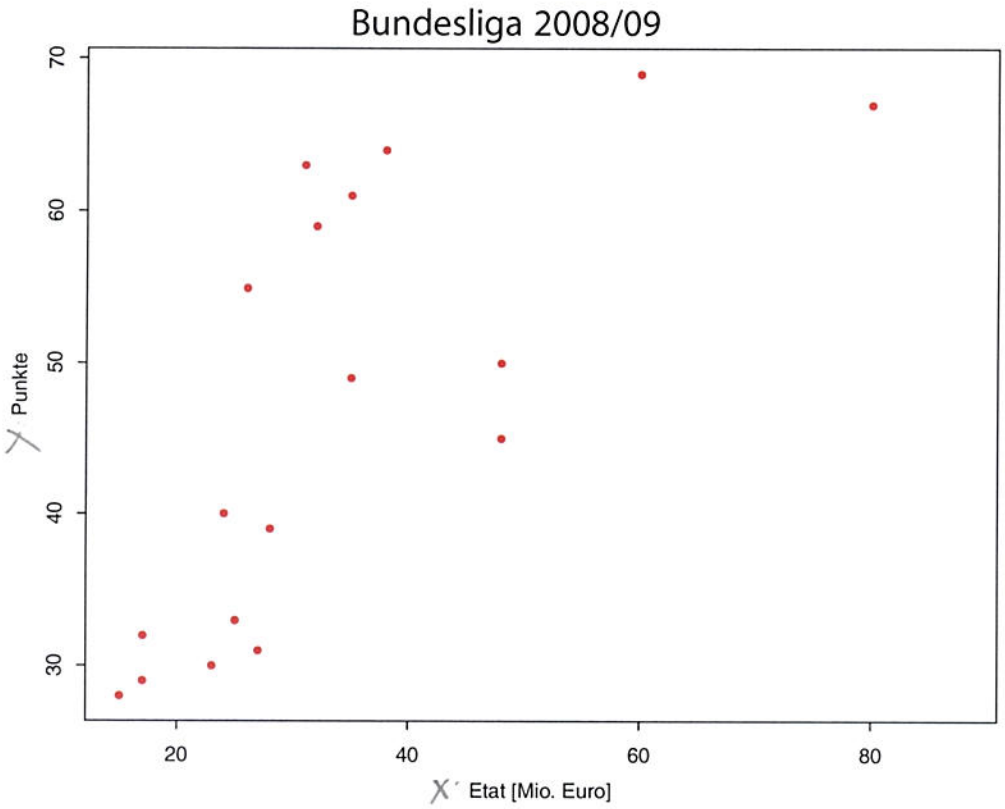
	Etat	Punkte
FC Bayern	80	67
VfL Wolfsburg	60	69
SV Werder Bremen	48	45
FC Schalke 04	48	50
VfB Stuttgart	38	64
Hamburger SV	35	61
Bayer 04 Leverkusen	35	49
Bor. Dortmund	32	59
Hertha BSC Berlin	31	63
1. FC Köln	28	39
Bor. Mönchengladbach	27	31
TSG Hoffenheim	26	55
Eintracht Frankfurt	25	33
Hannover 96	24	40
Energie Cottbus	23	30
VfL Bochum	17	32
Karlsruher SC	17	29
Arminia Bielefeld	15	28

(Quelle: Welt)

Darstellung der Daten in Streuplot



Darstellung der Daten in Streuplot



Trend als lineares Modell

- Kann man die **Tabellenpunkte** näherungsweise über einfache Funktion **in Abhängigkeit des Vereinsetats** darstellen?
- Allgemein: Darstellung einer Variablen Y als Funktion von X:

$$y = f(x)$$

- Dabei:
  - X heißt Regressor bzw. unabhängige Variable
  - Y heißt Regressand bzw. abhängige Variable
- Wichtiger (und einfachster) Spezialfall: f beschreibt einen linearen Trend:

$$y = a + b\,x$$

- Dabei anhand der Daten zu schätzen: a (Achsenabschnitt) und b (Steigung)
- Schätzung von a und b: Lineare Regression

Fehlerquadratsumme

- Pro Datenpunkt gilt mit Regressionsmodell:

$$y_i = a + bx_i + \epsilon_i$$

- Dabei: ϵ_i ist jeweils Fehler (der Grundgesamtheit),
- mit $e_i = y_i - (\hat{a} + \hat{b}x_i)$: Abweichung (Residuen) zwischen gegebenen Daten der Stichprobe und durch Modell geschätzten Werten
- Modell gut wenn alle Residuen e_i zusammen möglichst klein
- Einfache Summe aber nicht möglich, denn e_i positiv oder negativ
- Deswegen: Summe der Quadrate von e_i
- Prinzip der kleinsten Quadrate: Wähle a und b so, dass

$$Q(a,b) = \sum_{i=1}^{n} [y_i - (a + bx_i)]^2 \to \min$$

(y_i: tats. Beobachtung, $\hat{y}_i$: „Prognosewert")

Notwendige Bedingung für Minimum:

$\frac{\partial Q}{\partial a} \stackrel{!}{=} 0$ und $\frac{\partial Q}{\partial b} \stackrel{!}{=} 0$

Beste Lösung

- Beste und eindeutige Lösung:

$$\hat{b} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

$$\hat{a} = \bar{y} - \hat{b}\,\bar{x}$$

$(\bar{x}, \bar{y})$ ist Datenschwerpunkt und liegt immer auf der Regressionsgerade!

- Regressionsgerade:

$$\hat{y} = \hat{a} + \hat{b}\,x$$

Bundesligabeispiel

- Berechnung eines linearen Modells der Bundesligadaten
- dabei: Punkte $\hat{=} y$ und Etat $\hat{=} x$:

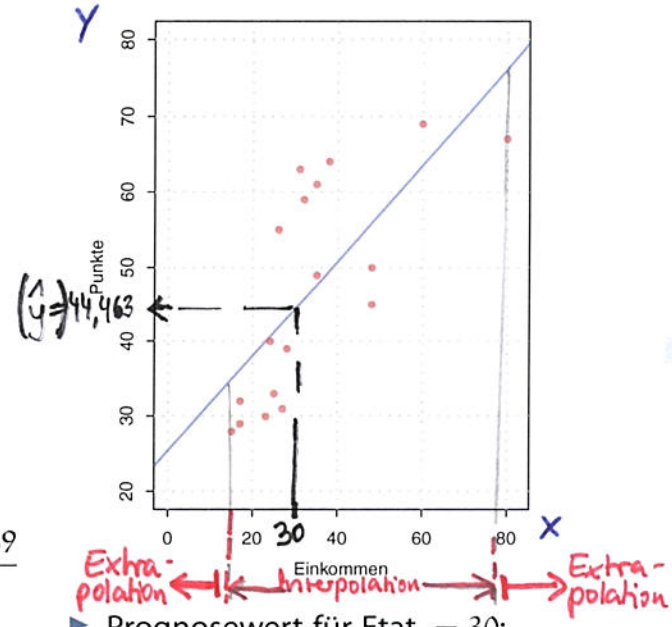
$\bar{x}$	33,83
$\bar{y}$	46,89
$\sum x_i^2$	25209
$\sum x_i y_i$	31474
n	18

$$\Rightarrow \hat{b} = \frac{31474 - 18 \cdot 33{,}83 \cdot 46{,}89}{25209 - 18 \cdot 33{,}83^2} \approx 0{,}634$$

$$\Rightarrow \hat{a} = 46{,}89 - \hat{b} \cdot 33{,}83 \approx 25{,}443$$

Regressionsgerade!

- Modell: $\hat{y} = 25{,}443 + 0{,}634 \cdot x$

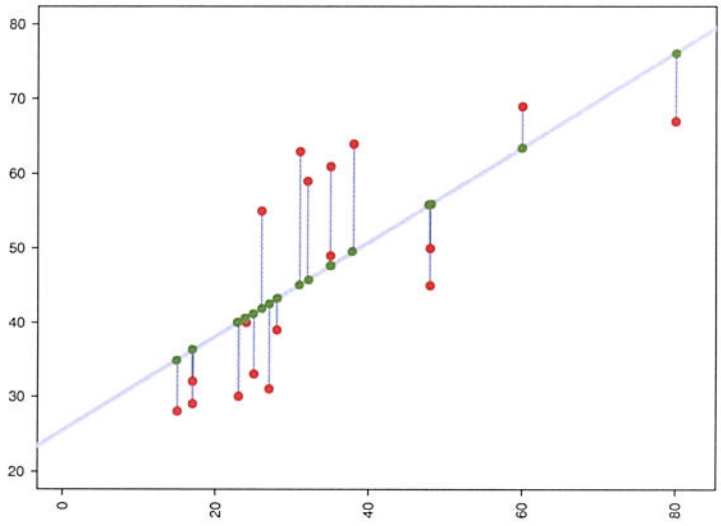


- Prognosewert für Etat $= 30$:

$$\hat{y}(30) = 25{,}443 + 0{,}634 \cdot 30 \approx 44{,}463$$

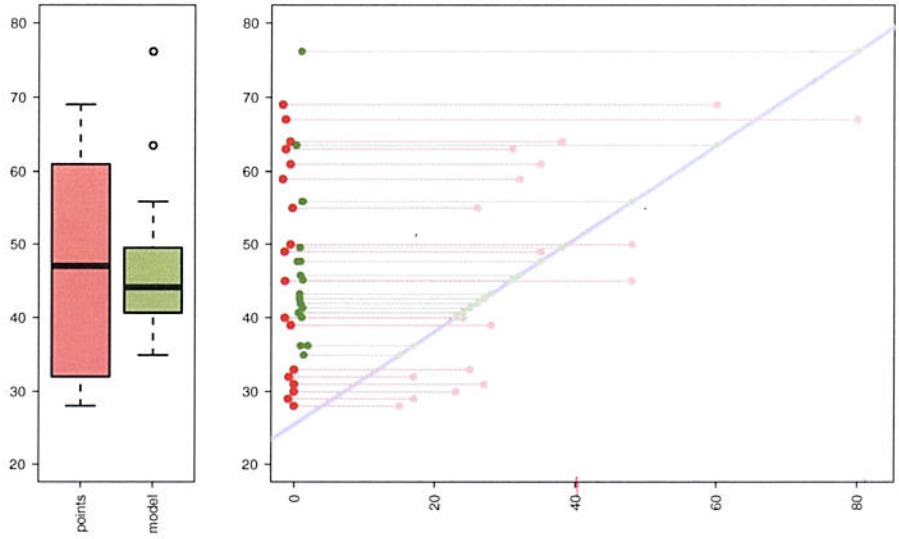
Varianz und Information

- **Varianz** der Daten in abhängiger Variablen y_i als Repräsentant des **Information**sgehalts
- Ein Bruchteil davon kann in Modellwerten $\hat{y}_i$ abgebildet werden



Varianz und Information

- **Varianz** der Daten in abhängiger Variablen y_i als Repräsentant des **Information**sgehalts
- Ein Bruchteil davon kann in Modellwerten $\hat{y}_i$ abgebildet werden



- Empirische Varianz (mittlere quadratische Abweichung) für „rot" bzw. „grün" ergibt jeweils

$$\frac{1}{18}\sum_{i=1}^{18}(y_i-\bar{y})^2 \approx 200{,}77 \qquad \text{bzw.} \qquad \frac{1}{18}\sum_{i=1}^{18}(\hat{y}_i-\bar{y})^2 \approx 102{,}78$$

Determinationskoeffizient

- Gütemaß für die Regression: Determinationskoeffizient (Bestimmtheitskoeffizient):

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} = \frac{\sum_{i=1}^{n}\hat{y}_i^2 - n\bar{y}^2}{\sum_{i=1}^{n}y_i^2 - n\bar{y}^2} = r^2 \in [0;1]$$

Prognosewerte; tatsächliche Werte; „erklärte Streuung / Gesamtstreuung"; Bravais-Pearson Korrelationskoeff. zum Quadrat!

- Mögliche Interpretation von R^2:
  Durch die Regression erklärter Anteil der Varianz
- $R^2 = 0$ wird erreicht wenn X, Y unkorreliert
  $R^2 = 1$ wird erreicht wenn $\hat{y}_i = y_i \; \forall i$ (alle Punkte auf Regressionsgerade)
- Im (Bundesliga-)Beispiel:

$$R^2 = \frac{\sum_{i=1}^{18}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{18}(y_i-\bar{y})^2} \approx \frac{102{,}78}{200{,}77} \approx 51{,}19\,\%$$

d.h. ca. 51% der in der abh. Variable Y vorhandenen Information kann über die Regression erklärt werden (Rest: nicht berücksichtigte oder zufällige Effekte)

TR:

Verwenden von SOLVE (nur fx-991DE X)

SOLVE verwendet das Newton'sche Näherungsverfahren zum Lösen von Gleichungen. Beachten Sie, dass SOLVE nur im Berechnungen-Modus verwendet werden kann. SOLVE unterstützt die Eingabe von Gleichungen in den folgenden Formaten.

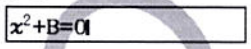
Beispiele: $y = x + 5$; $x = \sin(M)$; $xy + C$ (Behandelt wie $xy + C = 0$)

Hinweis

- Wenn eine Gleichung Eingabefunktionen mit einer öffnenden Klammer (wie z. B. bei sin und log) enthält, dürfen Sie die schließende Klammer nicht weglassen.
- Nach dem Drücken von SHIFT CALC (SOLVE) bis zum Beenden der SOLVE-Funktion durch Drücken von AC sollten Sie die linearen Eingabeverfahren zur Eingabe verwenden.

Zum Lösen von $x^2 + b = 0$ in Bezug auf x, wenn $b = -2$

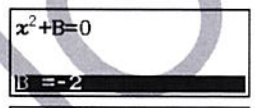
ALPHA) (x) x² + ALPHA ▪▪▪ (B) ALPHA CALC (=) 0



SHIFT CALC (SOLVE)

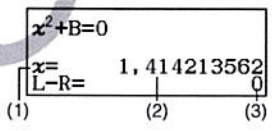
Geben Sie einen Anfangswert für x ein (in diesem Fall 1): 1 =

Weisen Sie B den Wert -2 zu: ◀▶ 2 =



Legen Sie die zu lösende Variable fest (Hier möchten wir x lösen, also verschieben Sie die Hervorhebung auf x.): ▲

Lösen Sie die Gleichung: =



(1) Lösungsvariable
(2) Lösung
(3) Ergebnis für (Linke Seite) – (Rechte Seite)

- Lösungen werden immer in der Dezimalform angezeigt.
- Je näher das Ergebnis für (Linke Seite) – (Rechte Seite) bei 0 liegt, desto genauer ist die Lösung.

Wichtig

- Bei Verwendung der Funktion SOLVE wird versucht, mit einer bestimmten Anzahl von Schritten eine Lösung zu finden. Wenn keine Lösung bestimmt werden kann, wird der Bestätigungsbildschirm „Fortfahren:[=] " angezeigt, der nachfragt, ob Sie fortfahren möchten. Drücken Sie =, um den Vorgang fortzusetzen, oder AC, um die SOLVE-Operation abzubrechen.
- Je nach der Eingabe für den Anfangswert für x (Lösungsvariable) können mit SOLVE möglicherweise keine Lösungen bestimmt werden. Wählen Sie in diesem Fall einen Anfangswert, der näher bei der Lösung liegt.
- Mit SOLVE kann nicht in jedem Fall die korrekte Lösung ermittelt werden, auch wenn eine existiert.
- Da bei der Funktion SOLVE das Newton'sche Verfahren verwendet wird, wird selbst dann, wenn mehrere Lösungen vorhanden sind, nur eine Lösung ermittelt.

- Aufgrund von Beschränkungen des Newton'schen Verfahrens sind Lösungen für Gleichungen wie die folgenden in der Regel schwer zu bestimmen: $y = \sin x$; $y = e^x$; $y = \sqrt{x}$.

Statistische Berechnungen

Führen Sie die nachstehenden Schritte aus, um eine statistische Berechnung zu starten.

1. Drücken Sie MENU, wählen Sie das Symbol für den Statistik-Modus und drücken Sie dann =.
2. Wählen Sie im eingeblendeten „Typ auswählen"-Bildschirm einen statistischen Berechnungstyp aus.

Zur Wahl dieses statistischen Berechnungstyps:	Drücken Sie diese Taste:
Einzelvariable (x)	1 (1 Variable)
Variablenpaar (x; y), lineare Regression	2 (y=a+bx)
Variablenpaar (x; y), quadratische Regression	3 (y=a+bx+cx²)
Variablenpaar (x; y), logarithmische Regression	4 (y=a+b·ln(x))
Variablenpaar (x; y), exponentielle Regression mit e	▼ 1 (y=a·e^(bx))
Variablenpaar (x; y), exponentielle Regression mit ab	▼ 2 (y=a·b^x)
Variablenpaar (x; y), Potenzregression	▼ 3 (y=a·x^b)
Variablenpaar (x; y), inverse Regression	▼ 4 (y=a+b/x)

- Verwenden Sie eine der oben genannten Tastenoperationen, um den Statistik-Editor anzuzeigen.

Hinweis: Wenn Sie den Berechnungstyp ändern möchten, nachdem Sie den Statistik-Modus aufgerufen haben, drücken Sie die Tasten OPTN 1 (Typ auswählen), um den Auswahlbildschirm für den Berechnungstyp aufzurufen.

Eingeben von Daten mit dem Statistik-Editor

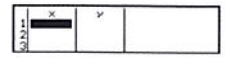
Der Statistik-Editor zeigt eine, zwei oder drei Spalten an: Einzelvariable (x), Einzelvariable und Häufigkeit (x; Freq), Variablenpaar (x; y), Variablenpaar und Häufigkeit (x; y; Freq). Die Anzahl der Datenzeilen, die eingegeben werden kann, hängt von der Anzahl der Spalten ab: 160 Zeilen für eine Spalte, 80 Zeilen für zwei Spalten, 53 Zeilen für drei Spalten.

Hinweis

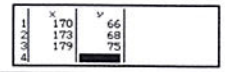
- Sie können die Freq-Spalte (Häufigkeit) verwenden, um die Anzahl (Häufigkeit) von identischen Datenelementen einzugeben. Die Anzeige der Freq-Spalte kann in der Statistik-Einstellung im Setup-Menü eingeschaltet (angezeigt) oder ausgeschaltet (nicht angezeigt) werden.
- Wenn Sie während der Anzeige des Statistik-Editors die Taste AC drücken, wird ein Bildschirm für statistische Berechnungen zur Durchführung von Berechnungen auf Grundlage der eingegebenen Daten angezeigt. Was Sie zum Zurückkehren vom Bildschirm für statistische Berechnungen zum Statistik-Editor tun müssen, hängt vom gewählten Berechnungstyp ab. Drücken Sie OPTN 3 (Daten), falls Sie Einzelvariable gewählt haben, oder OPTN 4 (Daten), falls Sie Variablenpaar gewählt haben.

Bsp. 1: Auswahl von logarithmischer Regression und Eingabe der folgenden Daten: (170; 66), (173; 68), (179; 75)

OPTN 1 (Typ auswählen) 4 (y=a+b·ln(x))



170 = 173 = 179 = ▼ ▶
66 = 68 = 75 =



Wichtig: Alle aktuell in den Statistik-Editor eingegebenen Daten werden gelöscht, wenn Sie den Statistik-Modus verlassen, zwischen der statistischen Rechnungsart mit einer Variablen oder einem Variablenpaar umschalten oder die Statistik-Einstellung im Setup-Menü ändern.

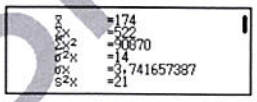
Eine Zeile löschen: Stellen Sie den Cursor im Statistik-Editor auf die zu löschende Zeile und drücken Sie dann DEL.

Eine Zeile einfügen: Stellen Sie den Cursor im Statistik-Editor auf die Position, an der Sie eine Zeile einfügen möchten, und drücken Sie die folgenden Tasten: OPTN 2 (Editor) 1 (Zeile einfügen).

Den kompletten Inhalt im Statistik-Editor löschen: Verwenden Sie hierzu im Statistik-Editor die folgenden Tasten: OPTN 2 (Editor) 2 (Alles löschen).

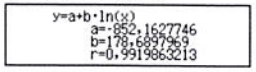
Anzeigen von statistischen Werten auf Grundlage eingegebener Daten

Vom Statistik-Editor aus:
OPTN 3 (1-Variab-Berech oder 2-Variab-Berech)
Vom Bildschirm für statistische Berechnungen aus:
OPTN 2 (1-Variab-Berech oder 2-Variab-Berech)



Anzeigen der Ergebnisse von Regressionsberechnungen auf Grundlage eingegebener Daten (nur Variablenpaar-Daten)

Vom Statistik-Editor aus:
OPTN 4 (Regression)
Vom Bildschirm für statistische Berechnungen aus: OPTN 3 (Regression)



Berechnen statistischer Werte aus den eingegebenen Daten

Sie können die Operationen in diesem Abschnitt verwenden, um statistische Werte aufzurufen, die den Variablen (σ_x, Σx^2 usw.) auf Grundlage der mit dem Statistik-Editor eingegebenen Daten zugewiesen wurden. Sie können die Variablen auch in Berechnungen verwenden. Die Operationen in diesem Abschnitt werden auf dem Bildschirm für statistische Berechnungen ausgeführt, der erscheint, wenn Sie während der Anzeige des Statistik-Editors AC drücken.

Nachfolgend werden die unterstützten statistischen Variablen und die für ihren Aufruf zu drückenden Tasten dargestellt. Für statistische Rechnungen mit einer Variablen sind die mit einem Sternchen (*) gekennzeichneten Variablen verfügbar.

Summierung: Σx^*, Σx^{2*}, Σy, Σy^2, Σxy, Σx^3, $\Sigma x^2 y$, Σx^4

OPTN ▼ 1 (Addition) 1 bis 8

Anzahl der Stichproben: n^* / **Mittelwert:** $\bar{x}^*$, $\bar{y}$ / **Varianz der Grundgesamtheit:** ${\sigma^2}_x^*$, ${\sigma^2}_y$ / **Grundgesamtheits-Standardabweichung:** ${\sigma}_x^*$, ${\sigma}_y$ / **Stichprobenvarianz:** ${s^2}_x^*$, ${s^2}_y$ / **Stichproben-Standardabweichung:** ${s}_x^*$, ${s}_y$

OPTN ▼ 2 (Variable) 1 bis 8, ▼ 1 bis ▼ 3

Minimalwert: $\min(x)^*$, $\min(y)$ / **Maximalwert:** $\max(x)^*$, $\max(y)$

Wenn die statistische Berechnung mit Einzelvariable ausgewählt ist:

OPTN ▼ 3 (Minimum/Maximum) 1, 5

Wenn die statistische Berechnung mit Variablenpaar ausgewählt ist:

OPTN ▼ 3 (Minimum/Maximum) 1 bis 4

Erstes Quartil: Q_1^* / **Median:** Med^* / **Drittes Quartil:** Q_3^* (Nur bei statistischen Berechnungen mit Einzelvariable)

OPTN ▼ 3 (Minimum/Maximum) 2 bis 4

Regressionskoeffizienten: a, b / **Korrelationskoeffizient:** r / **Schätzwerte:** $\hat{x}$, $\hat{y}$

OPTN ▼ 4 (Regressionen) 1 bis 5

Regressionskoeffizienten für quadratische Regression: a, b, c / **Schätzwerte:** $\hat{x}_1$, $\hat{x}_2$, $\hat{y}$

OPTN ▼ 4 (Regressionen) 1 bis 6

• $\hat{x}$, $\hat{x}_1$, $\hat{x}_2$ und $\hat{y}$ sind Befehle mit einem Argument unmittelbar davor.

Bsp. 2: Geben Sie die Daten x = {1; 2; 2; 3; 3; 3; 4; 4; 5} für eine einzelne Variable ein, verwenden Sie dabei die Freq-Spalte, um die Anzahl der Wiederholungen für jedes Element anzugeben {x_n; freq_n} = {1;1, 2;2, 3;3, 4;2, 5;1} und berechnen Sie den Mittelwert.

SHIFT MENU (SETUP) ▼ 2 *1 oder ▼ 3 *2 (Statistik) 1 (Ein)

*1: fx-87DE X *2: fx-991DE X

OPTN 1 (Typ auswählen) 1 (1 Variable)

1 = 2 = 3 = 4 = 5 = ▼ ▶
1 = 2 = 3 = 2 =

AC OPTN ▼ 2 (Variable) 1 ($\bar{x}$) = 3

Bsp. 3: Berechnen Sie die Korrelationskoeffizienten für die logarithmische Regression für folgende Variablenpaar-Daten und bestimmen Sie die Regressionsformel: (x; y) = (20; 3150), (110; 7310), (200; 8800), (290; 9310). Legen Sie Fix 3 (drei Dezimalstellen) für die Ergebnisse fest.

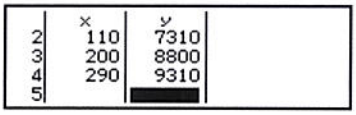
SHIFT MENU (SETUP) ▼ 2 *1 oder ▼ 3 *2 (Statistik) 2 (Aus)

*1: fx-87DE X *2: fx-991DE X

SHIFT MENU (SETUP) 3 (Zahlenformat) 1 (Fix) 3

OPTN 1 (Typ auswählen) 4 (y=a+b·ln(x))

20 = 110 = 200 = 290 = ▼ ▶
3150 = 7310 = 8800 = 9310 =



AC OPTN ▼ 4 (Regressionen) 3 (r) = 0,998

AC OPTN ▼ 4 (Regressionen) 1 (a) = -3857,984

AC OPTN ▼ 4 (Regressionen) 2 (b) = 2357,532

Schätzwerte berechnen

Anhand der mit einer statistischen Berechnung mit Variablenpaar erhaltenen Regressionsformel kann der Schätzwert von y für einen gegebenen x-Wert berechnet werden. Der entsprechende x-Wert (zwei

Beispiel: **Taschenrechnerfunktion**

vgl. F 71

i	x_i	y_i
1	2	4
2	4	3
3	3	6
4	9	7
5	7	8

Gesucht: i) $\hat{a}$, $\hat{b}$, R^2 $(=r^2)$

ii) Regressionsgerade

iii) Prognosewert für $x = 6$

i) $\hat{a} = 3{,}1$ $\hat{b} = 0{,}5$ $r \approx 0{,}7030$

$\rightarrow R^2 \approx 0{,}4942$

ii) $\hat{y} = 3{,}1 + 0{,}5 \cdot x$

iii) $\hat{y}(6) = 3{,}1 + 0{,}5 \cdot 6 = 6{,}1$

fx-991 DE X:

MENU → Statistik → [=] → [2]

[6]

<Dateneingabe>

[mittels [=] abschließen

[^] [v] [>] [<] Pfeile nutzen]

[OPTN] → [4] ⇝ $a = 3{,}1$

$b = 0{,}5$

$r = 0{,}7030$

(1) Regressionsanalyse mit R am Beispiel

Umsetzung mit R

R-Code

x_i	y_i
2	4
4	3
3	6
9	7
7	8

1) Regression
2) Plot mit Regressionsgerade

(2) Anscombe - Daten (F.106 - 108)

(3) Regressionsanalyse mit Falldaten (F.109)

Output - interpretation

1) summary (meine Regression)
   - ↳ Outputinterpretation
     - i) $\hat{a}$, $\hat{b}$, R^2 ?
     - ii) Regressionsgerade ?
     - iii) Prognosewert für AlterV = 50
     - iv) r ?
2) plot (meine Regression)
   - ↳ Residualanalyse zur Modellprüfung

Regression: 4 eindimensionale Beispiele

- Berühmte Daten aus den 1970er Jahren:

	①	②	③	④	①	②	③	④
i	x_{1i}	x_{2i}	x_{3i}	x_{4i}	y_{1i}	y_{2i}	y_{3i}	y_{4i}
1	10	10	10	8	8,04	9,14	7,46	6,58
2	8	8	8	8	6,95	8,14	6,77	5,76
3	13	13	13	8	7,58	8,74	12,74	7,71
4	9	9	9	8	8,81	8,77	7,11	8,84
5	11	11	11	8	8,33	9,26	7,81	8,47
6	14	14	14	8	9,96	8,10	8,84	7,04
7	6	6	6	8	7,24	6,13	6,08	5,25
8	4	4	4	19	4,26	3,10	5,39	12,50
9	12	12	12	8	10,84	9,13	8,15	5,56
10	7	7	7	8	4,82	7,26	6,42	7,91
11	5	5	5	8	5,68	4,74	5,73	6,89

(Quelle: Anscombe (1973))

in R: data(anscombe)
anscombe[order(anscombe$x1),] # Daten sortieren, um Ersten Eindruck der Datenlage zu haben und der Datentabelle

Regression: 4 eindimensionale Beispiele

Statistik

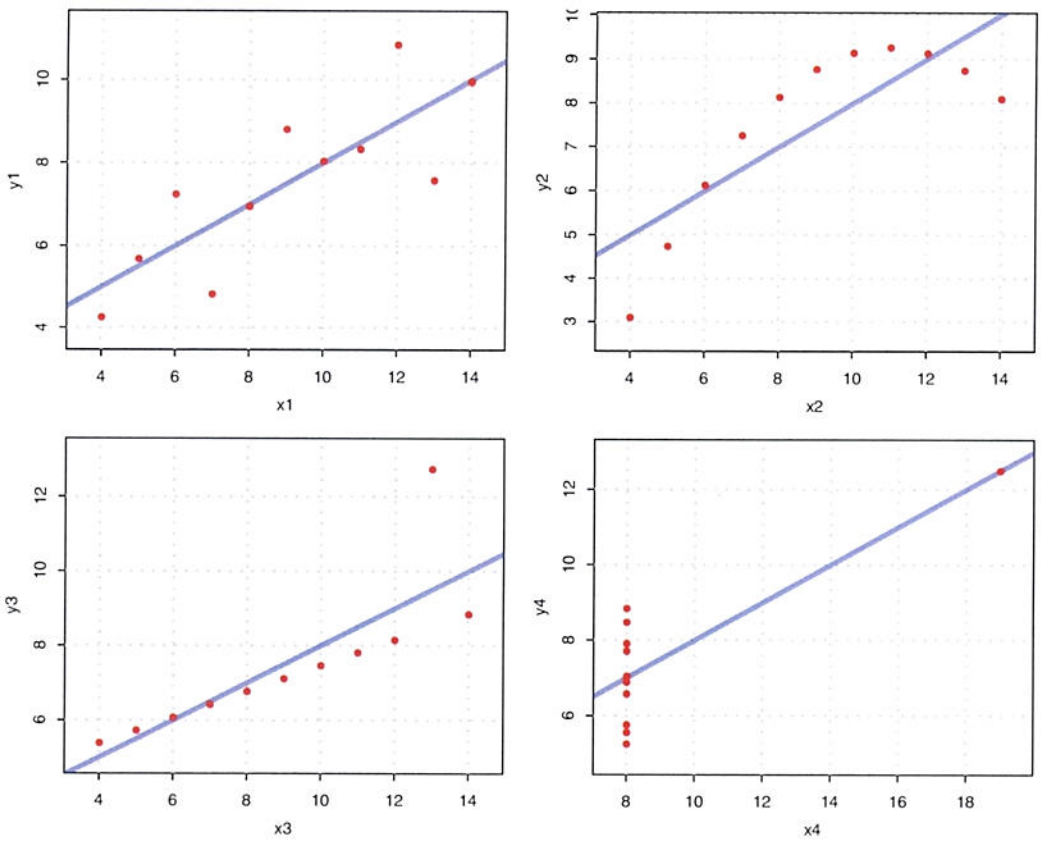
1. Einführung
2. Deskriptive Statistik
Häufigkeiten
Lage und Streuung
Konzentration
Zwei Merkmale
Korrelation
Preisindizes
Lineare Regression
3. W-Theorie
4. Induktive Statistik
Quellen
Tabellen

- In folgender Tabelle: Jeweils Ergebnisse der linearen Regressionsanalyse
- dabei: x_k unabhängige Variable und y_k abhängige Variable
- Modell jeweils: $y_k = a_k + b_k x_k$

k	$\hat{a}_k$	$\hat{b}_k$	R_k^2
1	3,0001	0,5001	0,6665
2	3,0010	0,5000	0,6662
3	3,0025	0,4997	0,6663
4	3,0017	0,4999	0,6667

[?] Wie lautet r?
1.) selbes Vorzeichen wie $\hat{b}$ (Steigungsparameter)
2.) $r = \sqrt{R^2}$ (hier alle positiv!)

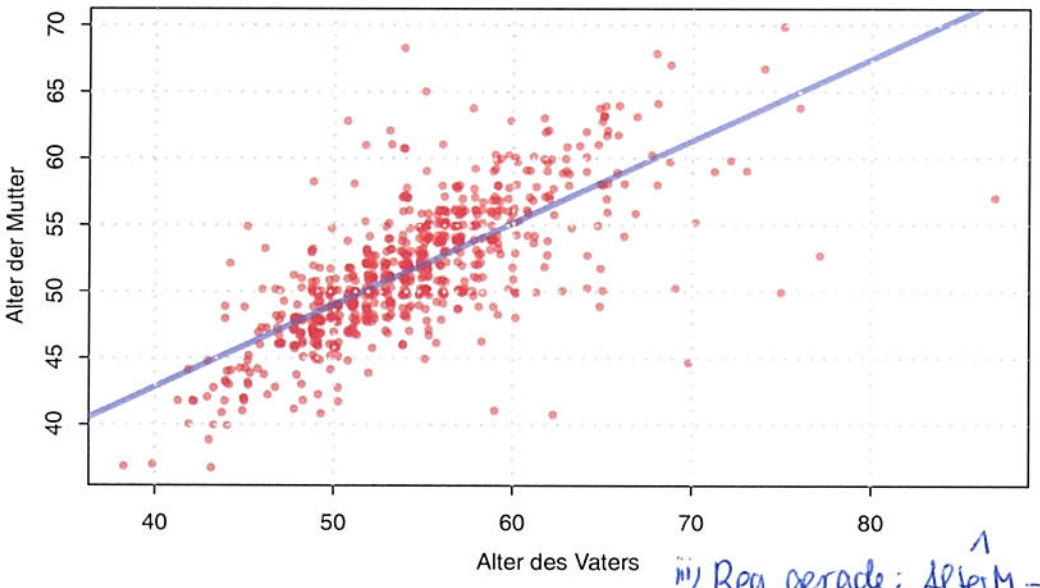
Plot der Anscombe-Daten



Beispieldaten

```
meineRegression = lm(AlterM ~ AlterV)
meineRegression

plot(AlterV, AlterM,
     xlab="Alter des Vaters",
     ylab="Alter der Mutter")
abline(meineRegression)
```

```
##
## Call:
## lm(formula = AlterM ~ AlterV)
##
## Coefficients:
## (Intercept)       AlterV
##     18.2234       0.6159
```



aktuelle Daten:

i) $\hat{\alpha} = 18,2309$

$\hat{b} = 0,6153$

$R^2 = 0,5211$

ii) $\rightarrow r = +\sqrt{0,5211} = \ldots$

iii) Reg. gerade: $\widehat{AlterM} = 18,2309 + 0,6153 \cdot AlterV$

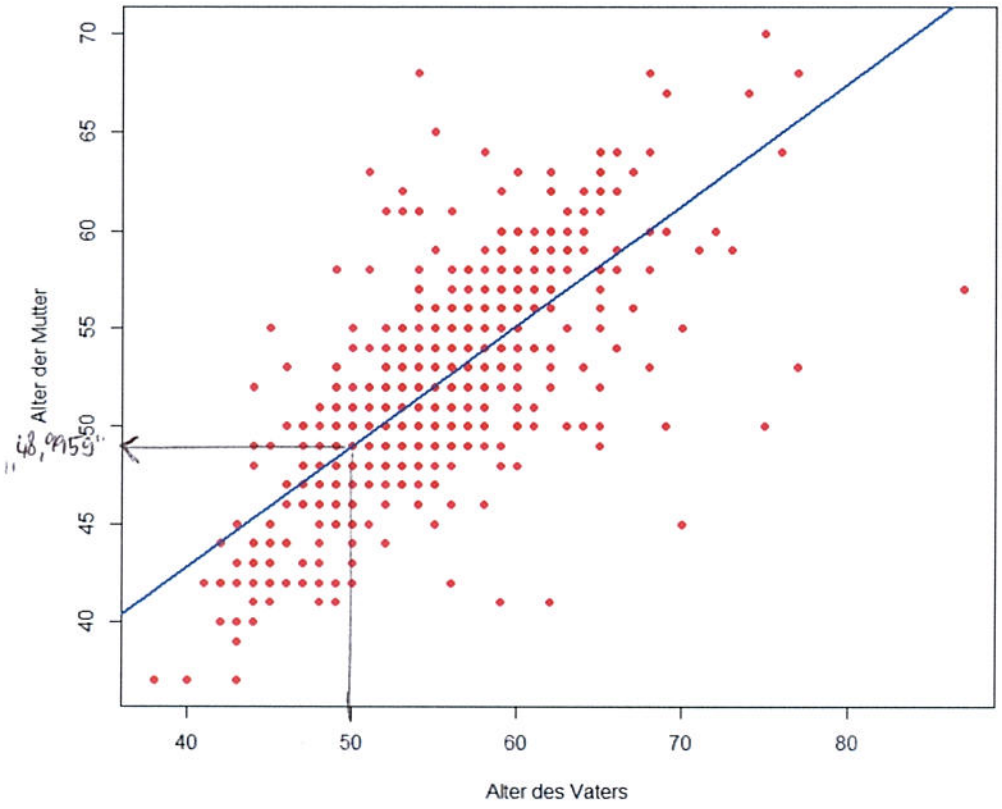
iv) Prognose: $\widehat{AlterM}(50) = 18,2309 + 0,6153 \cdot 50$

$\approx 48,9959$

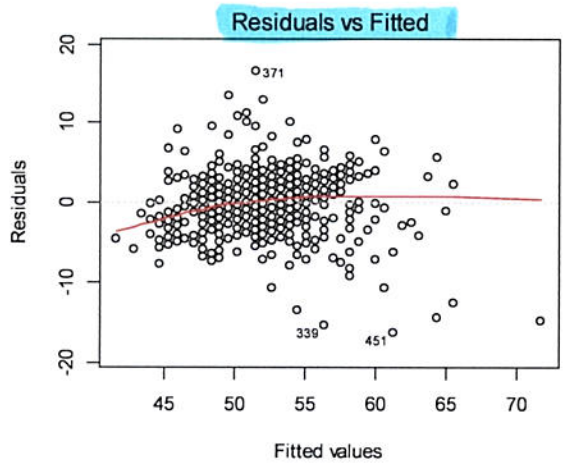
③ Regressionsanalyse mit aktuellen Fallzahlen (R-Outputinterpretation)

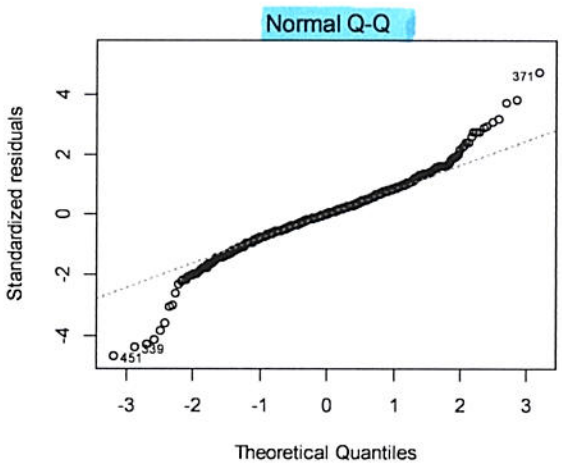
```
> meineRegression=lm(AlterM~AlterV)
>
> plot(AlterV, AlterM, xlab="Alter des Vaters", ylab="Alter der Mutter", col=rgb(1,0,0,0.7), pch=16)
> abline(meineRegression, col="blue", lwd=2)
>
> meineRegression

Call:
lm(formula = AlterM ~ AlterV)

Coefficients:
(Intercept)       AlterV
    18.2309       0.6154
```

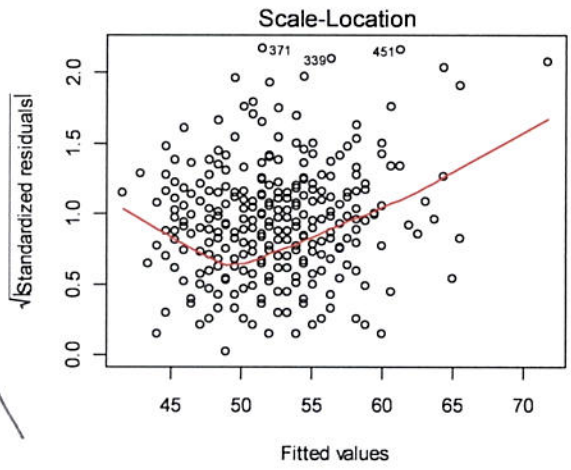
```
> summary(meineRegression)

Call:
lm(formula = AlterM ~ AlterV)

Residuals:
     Min       1Q   Median       3Q      Max
-16.3055  -1.9213   0.0015   1.9247  16.5401

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 18.2309     1.2190   14.96   <2e-16 ***
AlterV       0.6153     0.0223   27.60   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.51 on 700 degrees of freedom
  (4 observations deleted due to missingness)
Multiple R-squared:  0.5211,    Adjusted R-squared:  0.5204
F-statistic: 761.7 on 1 and 700 DF,  p-value: < 2.2e-16
```

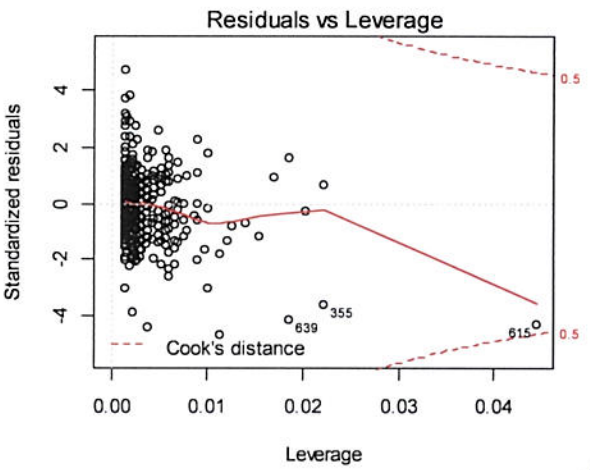


```
> par(mfrow=c(2,2))
> plot(meineRegression)
```









Cook's Distanz

- Oft Kritisch: Einzelne Punkte, die Modell stark beeinflussen
- Idee: Was würde sich ändern, wenn solche Punkte weggelassen würden?
- Cook-Distanz: Misst den Effekt eines gelöschten Objekts
- Formel für ein lineares Modell mit einem unabh. Merkmal:

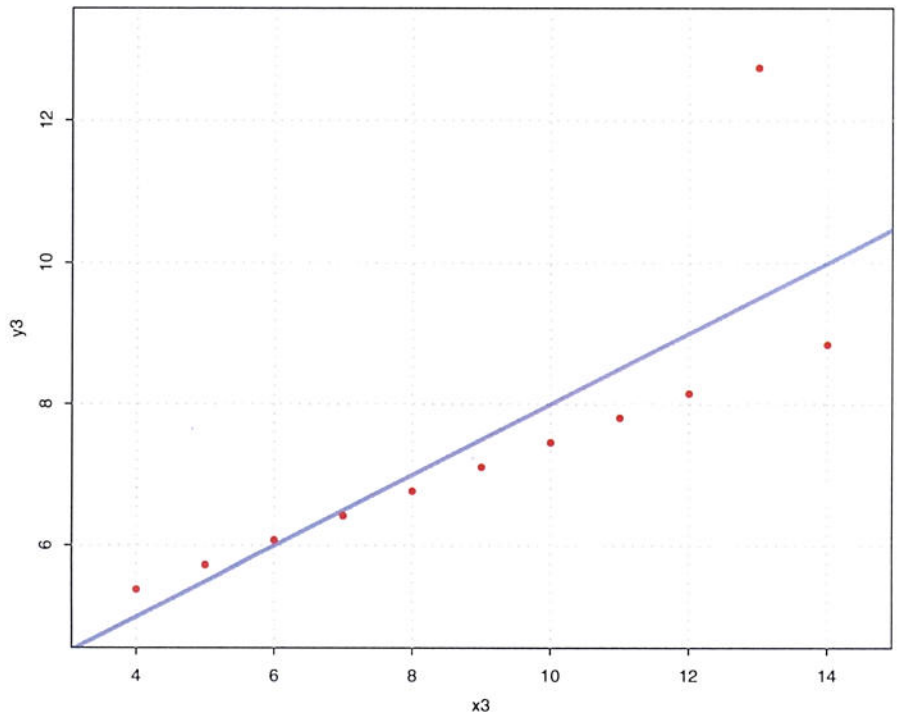
$$D_i = \frac{\sum_{j=1}^{n} (\hat{y}_j - \hat{y}_{j(\text{ohne } i)})^2}{\text{MSE}}$$

- Dabei bedeutet:
  - $\hat{y}_j$: Prognosewert des kompletten Modells für das j-te Objekt
  - $\hat{y}_{j(\text{ohne } i)}$: Prognosewert des Modells ohne Objekt i für das j-te Objekt
  - $\text{MSE} = \frac{1}{n} \cdot \sum(\hat{y}_i - y_i)^2$: Normierender Term (Schätzwert für Fehlerstreuung)

ohne F. 110-111

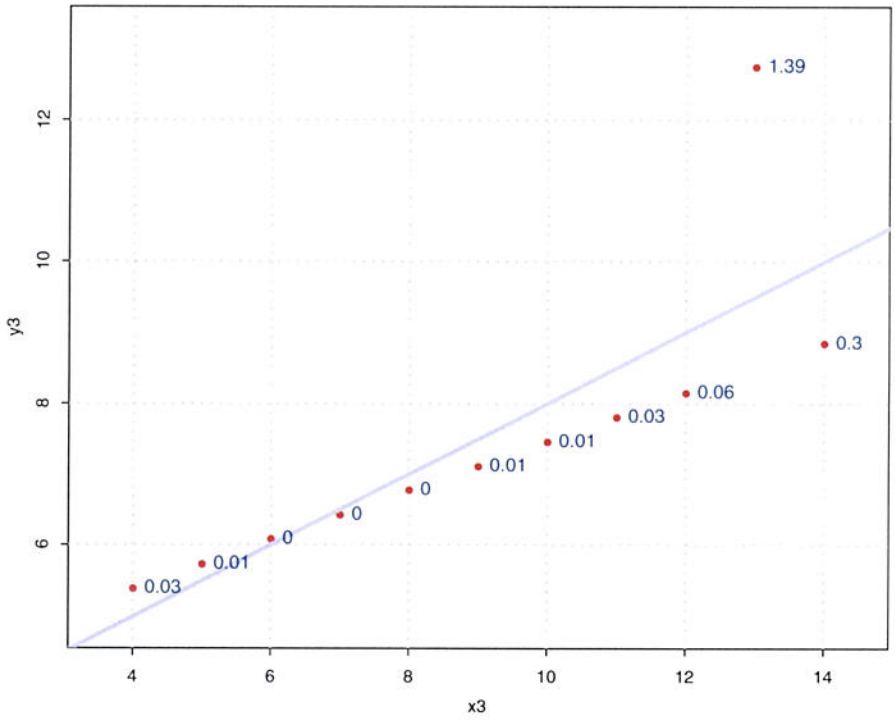
Ausreißer?

- Anscombe-Daten: Regressionsmodell Nr. 3



Ausreißer?

- Anscombe-Daten: Regressionsmodell Nr. 3
- Darstellung der Cook-Distanz neben Punkten
- Faustformel: Werte über 1 sollten genau untersucht werden

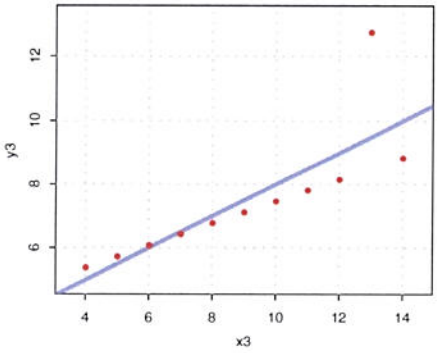


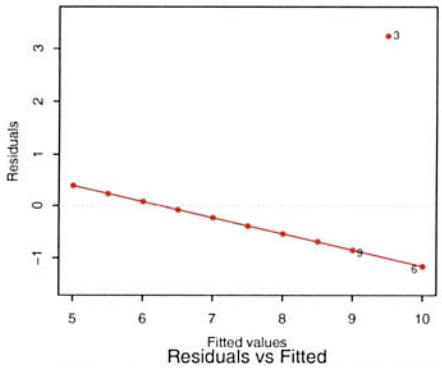
Residualanalyse

- Oft aufschlussreich: Verteilung der Residuen e_i
- Verbreitet: Graphische Darstellungen der Residuen
- Z.B.: e_i über $\hat{y}_i$

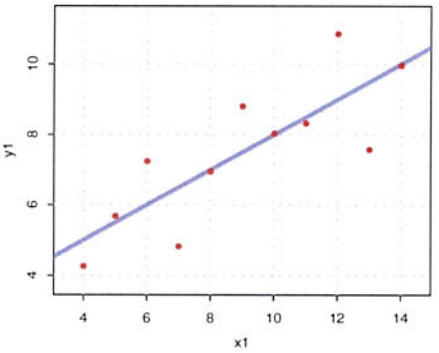
Residualanalyse

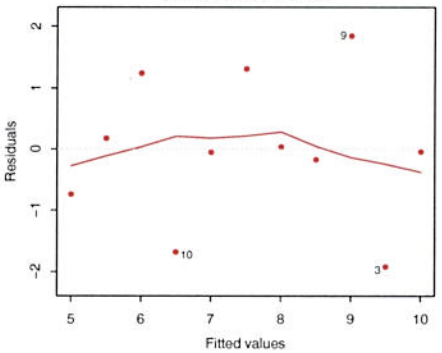
- Oft aufschlussreich: Verteilung der Residuen e_i
- Verbreitet: Graphische Darstellungen der Residuen
- Z.B.: e_i über $\hat{y}_i$

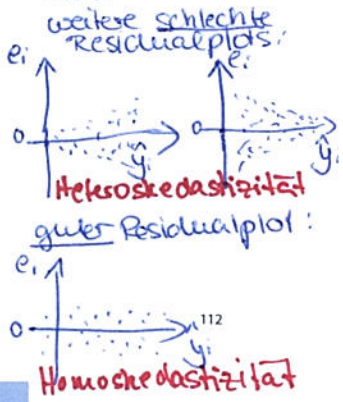




schlecht! Trend / Tendenz in Residuen!







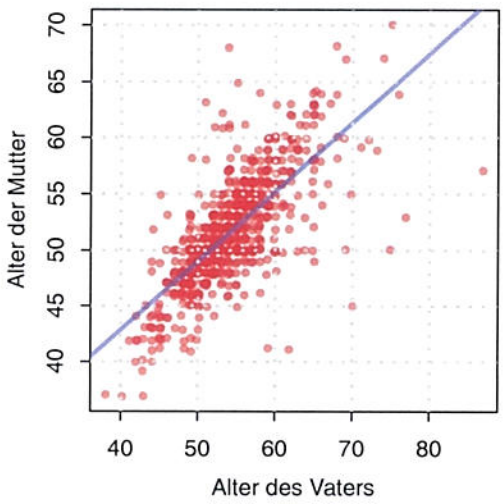
Statistik

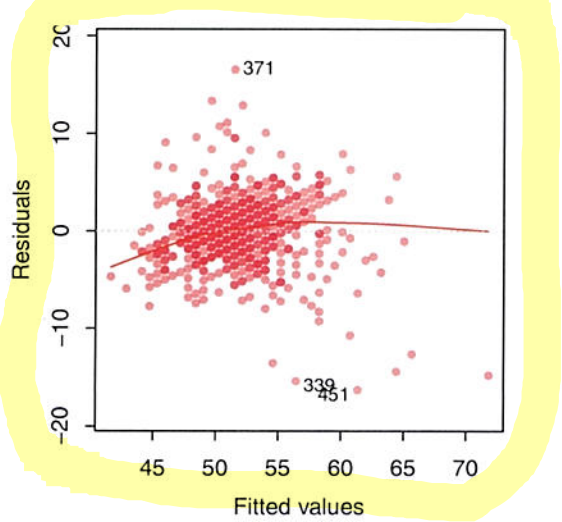
Residualanalyse

R:
plot (meine Regression)

Wichtige Eigenschaften der Residuenverteilung

- Möglichst keine systematischen Muster
- Keine Änderung der Varianz in Abhängigkeit von $\hat{y}_i$ (Homoskedastizität) } 1. Plot
- Nötig für inferentielle Analysen: Näherungsweise Normalverteilung der Residuen (q-q-plots) } 2. Plot nicht abgebildet





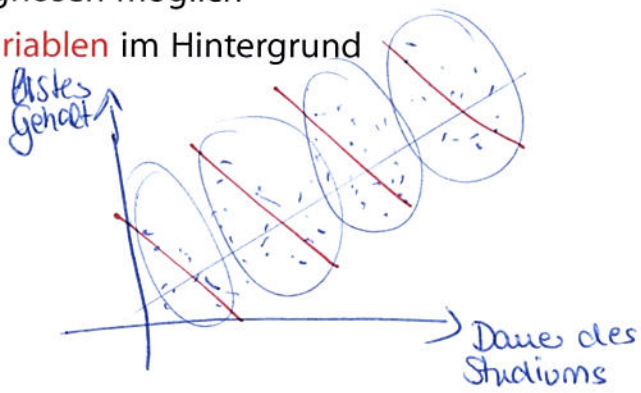
1. Einführung
2. Deskriptive Statistik
Häufigkeiten
Lage und Streuung
Konzentration
Zwei Merkmale
Korrelation
Preisindizes
Lineare Regression
3. W-Theorie
4. Induktive Statistik
Quellen
Tabellen

Kausalität versus Korrelation

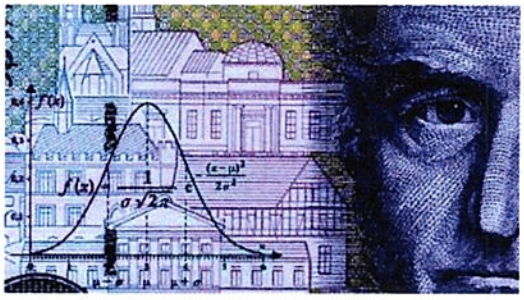
Exkurs: Kausalität vs. Korrelation

- Meist wichtig für sinnvolle Regressionsanalysen:
- Kausale Verbindung zwischen unabhängigem und abhängigem Merkmal (Wirkungsbeziehung)
- Sonst bei Änderung der unabhängigen Variablen keine sinnvollen Prognosen möglich
- Oft: Latente Variablen im Hintergrund

Beispiel:



Statistik: Table of Contents



Aufgabe 2

18 Punkte

Stefan Jumper nimmt seit 8 Jahren am gleichen Marathon teil und hat jedes Jahr seine Trainings- und Ergebnisdaten dokumentiert. Er hat dazu pro Jahr jeweils die Ergebniszeit im Marathon (Merkmal *Ergebnis Marathon*, in Minuten) sowie die durchschnittliche Anzahl gelaufener Trainingskilometer in den 8 bzw. 16 Wochen vor dem Marathon (Merkmal *8-Wochen-Trainings-Durchschnitt* bzw. *16-Wochen-Trainings-Durchschnitt*, jeweils in km pro Woche) in der folgenden Tabelle festgehalten:

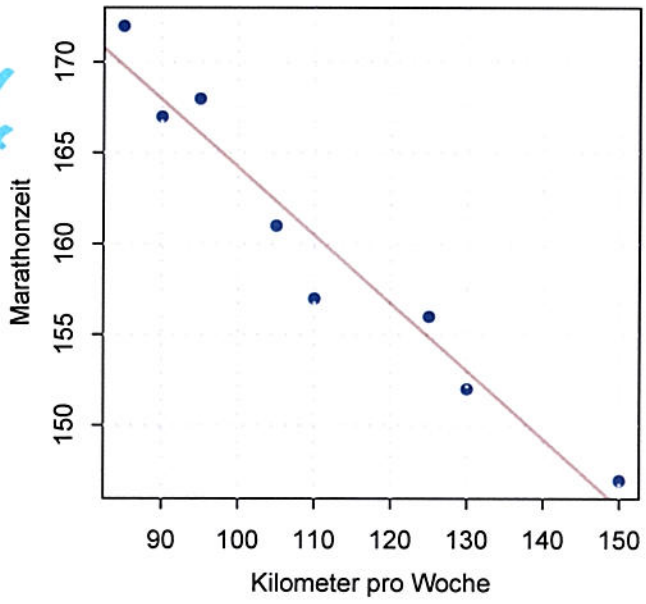
Jahr	1	2	3	4	5	6	7	8
Ergebnis Marathon	172	161	156	147	152	167	157	168
8-Wochen-Trainings-Durchschnitt	85	105	125	150	130	90	110	95
16-Wochen-Trainings-Durchschnitt	70	80	120	125	125	100	105	75

Sie haben Stefan laufend von den interessanten Themen der Statistik-Vorlesung berichtet, die Sie in diesem Semester besucht haben. Stefan wünscht sich daraufhin von Ihnen eine statistische Auswertung seiner Ergebnisse.

a) Stefan vermutet, dass der 8-Wochen-Trainings-Durchschnitt eine bessere Prognose für seine gelaufene Marathonzeit darstellt als der 16-Wochen-Trainings-Durchschnitt.
Berechnen Sie jeweils einen geeigneten Korrelationskoeffizienten, vergleichen Sie die beiden Ergebnisse und geben Sie an, ob Stefans Vermutung durch das Ergebnis gestützt wird.

b) Bestimmen Sie den Funktionsterm eines linearen Regressionsmodells für das Ergebnis im Marathon in Abhängigkeit vom 8-Wochen-Trainings-Durchschnitt.

c) Zeichnen Sie die Datenpunkte sowie die Regressionsgerade in das nebenstehende Koordinatensystem ein. Beschriften Sie dazu auch die Achsen geeignet. (Hinweis: Der Schnittpunkt der Koordinatenachsen soll nicht bei (0, 0) liegen)



d) Welche Marathonzeit erwartet Stefan auf Grundlage des linearen Modells für einen 8-Wochen-Trainings-Durchschnitt von 175 km pro Woche?

e) Korrigieren Sie den folgenden R-Code mit dem ein wenig begabter Student einen Teil der obige Aufgabe lösen wollte:

$$r = \frac{\sum x_i y_i - n \cdot \bar{x}\bar{y}}{\sqrt{\sum x_i^2 - n\bar{x}^2} \cdot \sqrt{\sum y_i^2 - n\bar{y}^2}}$$

a) Bravais-Pearson-Korrelationskoeffizient
8-Wo./Mar.: $r \approx -0{,}9728$ → Betragsmäßig größer → stärkerer (negativer) Zusammenhang zw. 8-Wo. und Marathonzeit
16-Wo./Mar.: $r \simeq -0{,}8805$

b) $\hat{Marathon} = 201{,}6290 - 0{,}3742 \cdot 8\text{-Wo.}$
$\hat{a}$ $\hat{b}$

c) y (Marathonzeit)
170
150
145 80 90 ... 150 x (8-Wo.Ø)

d) $\hat{Marathon}(175) = 201{,}6290 - 0{,}3742 \cdot 175 \simeq 136.144$

```
Zeit = c[172:161:156:147:152:167:157:168]
8Wochendurchschnitt == seq(85,105,125,150,130,90,110,95)
cor(Zeit, 8Wochendurchschnitt, method = "spearman")
z <- LM(8Wochendurchschnitt~Zeit)
plotter(Zeit,8Wochendurchschnitt, color = "blau", blue
        ylab = Kilometer pro Woche", ylab == "Zeit")
abline(z)
```

Lösungshinweis:

```
Marathon    = c(172, 161, 156, 147, 152, 167, 157, 168)
WochenKM.8  = c(85, 105, 125, 150, 130, 90, 110, 95)
WochenKM.16 = c(70, 80, 120, 125, 125, 100, 105, 75)

cor(Marathon, WochenKM.8,  method = "pearson")
cor(Marathon, WochenKM.16, method = "pearson")

Marathon.Regression = lm(Marathon ~ WochenKM.8)
Marathon.Regression

plot(WochenKM.8, Marathon, col = "blue", pch=16, cex=1.5,
     xlab = "Kilometer pro Woche", ylab = "Marathonzeit")
grid()
abline(Marathon.Regression, lwd=2, col="#90000050")
```

zu c) 2 geeignete Punkte der Regressionsgerade:

$(\bar{x} \mid \bar{y}) = (111{,}25 \mid 160)$

$(90 \mid 161{,}951)$

$(140 \mid 149{,}241)$