Anmerkung zu den letzten HA: die Aufgaben 27-32 sind auch mit R zu rechnen

Statistik

HA 20.4.2016: Aufgaben 34-40

für Betriebswirtschaft, Internationales Management, Wirtschaftsinformatik und Informatik

Sommersemester 2016

Veranstaltungen zur Statistik für BW/IM Sommersemester 2016				
Was?	Wer?	Wann?	Wo?	ab wann?
Vorlesung Statistik	Etschberger	Mi, 14.00-17.00	B2.14	16.03.2016
Vorlesung Statistik PLUS	Etschberger/Jansen	- Blocktermin -	?	?
Übung Statistik	Etschberger	Mi, 17.00-18.30	A1.10	30.03.2016
Übung Statistik	Jansen	Di, 11.30-13.00	W1.06	22.03.2016
Übung Statistik	Jansen	Di, 14.00-15.30	W2.14	22.03.2016
Übung Statistik	Jansen	Mi, 11.30-13.00	W2.11	30.03.2016
Übung Statistik	Jansen	Do, 14.00-15.30	W2.14	31.03.2016
Übung Statistik	Schneller	Do, 14.00-15.30	W3.03	31.03.2016
Übung Statistik	Schneller	Do, 15.30-17.00	W3.03	31.03.2016
Übung Statistik	Wins	Di, 14.00-15.30	J3.19	22.03.2016
Übung Statistik	Wins	Di, 15.30-17.00	J3.19	22.03.2016
Offener Statistikraum	Etschberger/Tutoren	Mo, 14.00-17.45	B3.05	04.04.
Veranstaltungen für Teilnehmer der WiMa-Klausur im Juli 2016				
Was?	Wer?	Wann?	Wo?	ab wann?
Tutorium Mathematik	Burkart	Do 13.30-15.00	W1.06	07.04.2016
Tutorium Mathematik	Burkart	Do 15.00-16.15	W1.06	07.04.2016
Offener Matheraum	Jansen/Tutoren	?	?	?

	HSA Statistik SS 2016 Sessionlist		
1			
2			
3	Datum	Statistik für IM/BW	Nr.
4	Mittwoch, 16. März 2016	Einführung, R Installation, Rstudio Einführung, Skalen	1
5	Mittwoch, 23. März 2016	univ. deskr. Stat., Quantile, Plots	2
6	Mittwoch, 30. März 2016	Streuung, Konzentrationsmaße	3
7	Mittwoch, 6. April 2016	Kontingenztabellen, Mosaikplots, Korrelation	4
8	Mittwoch, 13. April 2016	Preisindizes, lineare Regression	5
9	Mittwoch, 20. April 2016	Kombinatorik, Wahrscheinlichkeit	6
10	Mittwoch, 27. April 2016	Wahrscheinlichkeit, diskrete Zufallsvariablen	7
11	Mittwoch, 4. Mai 2016	Pyramid	
12	Mittwoch, 11. Mai 2016	Binomial-, Hypergeom.-, Poisson-Verteilung	8
13	Mittwoch, 18. Mai 2016	Stetige ZV, Gleichverteilung	9
14	Mittwoch, 25. Mai 2016	Normalverteilung, Verteilungsparameter	10
15	Mittwoch, 1. Juni 2016	Schätzfunktionen und Punktschätzer	11
16	Mittwoch, 8. Juni 2016	Konfidenzintervalle	12
17	Mittwoch, 15. Juni 2016	Tests	13
18	Mittwoch, 22. Juni 2016	Puffer, WH, Fragen zur Probekl.	14
19	Mittwoch, 29. Juni 2016	AW Prüfungswoche	

Prof. Dr. Stefan Etschberger
Hochschule Augsburg

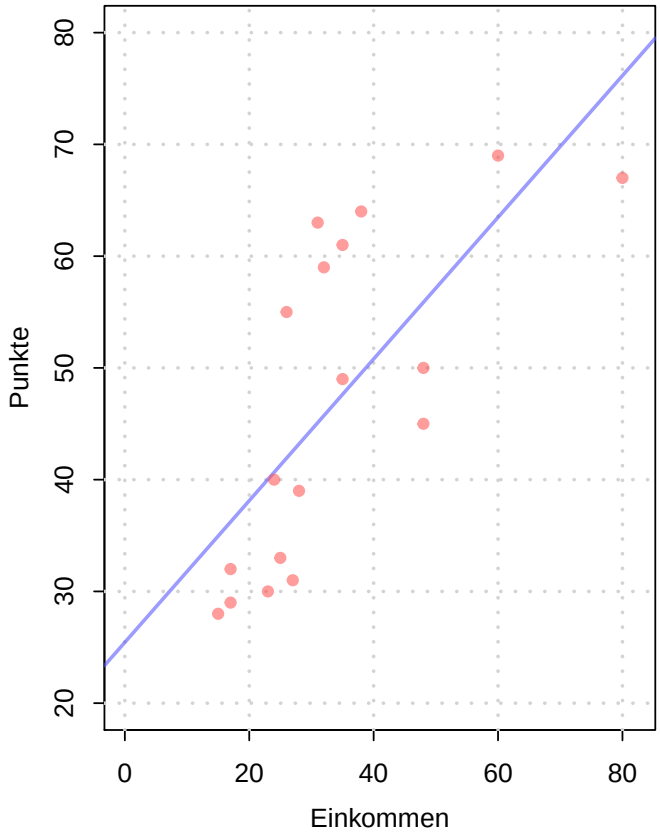
Bundesligabeispiel

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

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

$$\Rightarrow \hat{b} = \frac{31474 - 18 \cdot 33{,}83 \cdot 46{,}89}{25209 - 18 \cdot 33{,}83^2}$$
$$\approx 0{,}634$$
$$\Rightarrow \hat{a} = 46{,}89 - \hat{b} \cdot 33{,}83$$
$$\approx 25{,}443$$

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

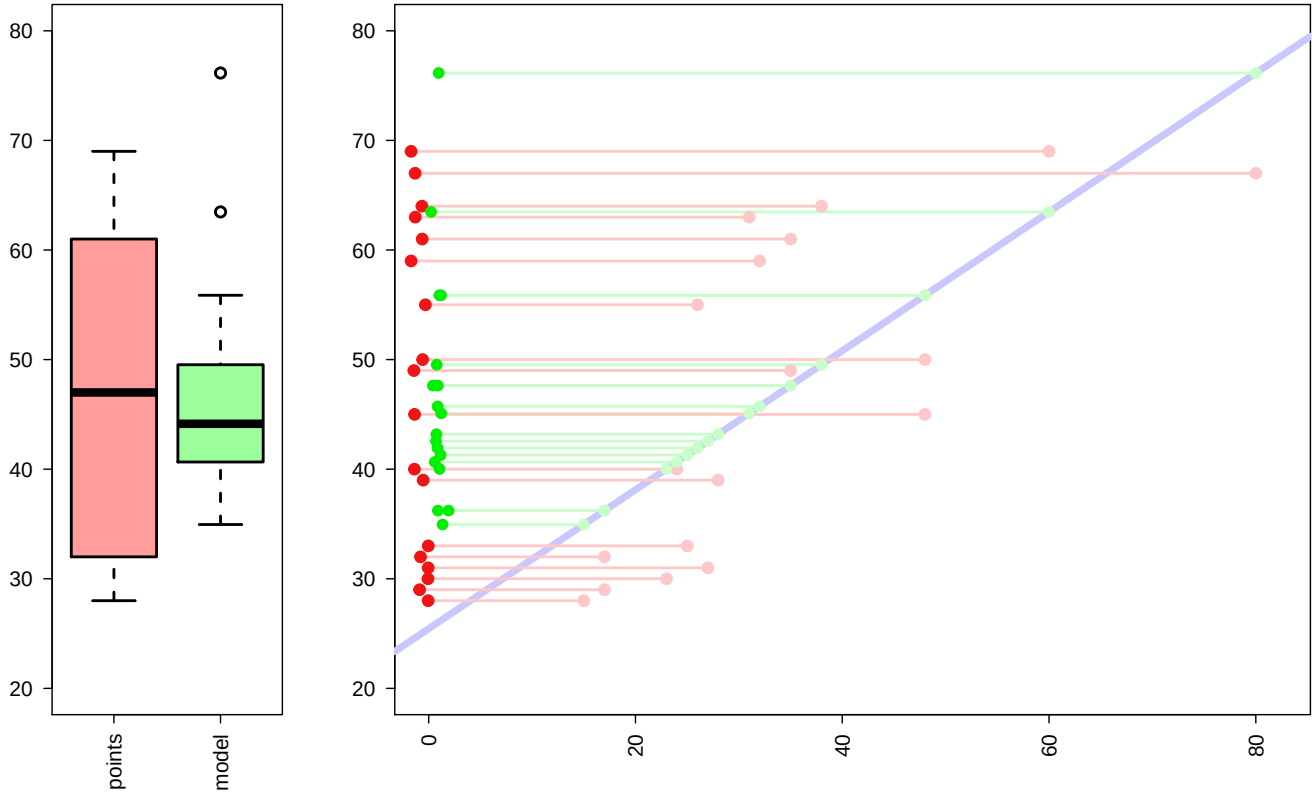


- Prognosewert für Etat $= 30$:

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

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



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

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

Determinationskoeffizient

- Gütemaß für die Regression: Determinationskoeffizient (Bestimmtheitskoeffizient):

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

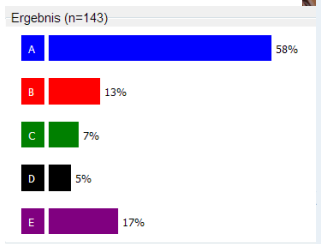
- Mögliche Interpretation von R^2:
  Durch die Regression erklärter Anteil der Varianz
- $R^2 = 0$ wird erreicht wenn X, Y unkorreliert
  $R^2 = 1$ wird erreicht wenn $\hat{y}_i = y_i \; \forall i$ (alle Punkte auf Regressionsgerade)
- Im (Bundesliga-)Beispiel:

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

Regression: 4 eindimensionale Beispiele

Umfrage: Ich habe

A) Zahlen und Grafik richtig
B) Gerade falsch eingezeichnet, sonst alles richtig
C) Fehler in den Zahlen (Grafik konsequent)
D) Fehler in Zahlen und Grafik
E) alles falsch
F) nichts gemacht



- Berühmte Daten aus den 1970er Jahren:

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

Aufgaben:

- Modellparameter (a, b)
- R^2
- Streuplot mit eingezeichneter Gerade

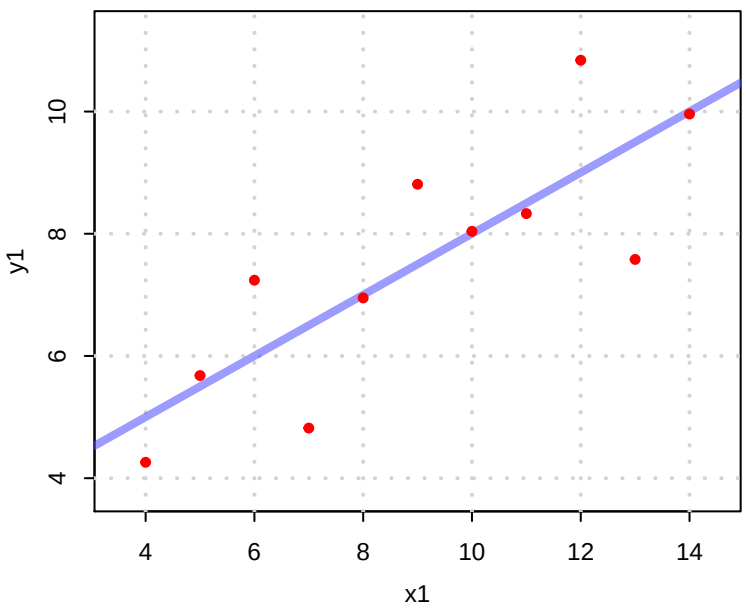
(Quelle: Anscombe (1973))

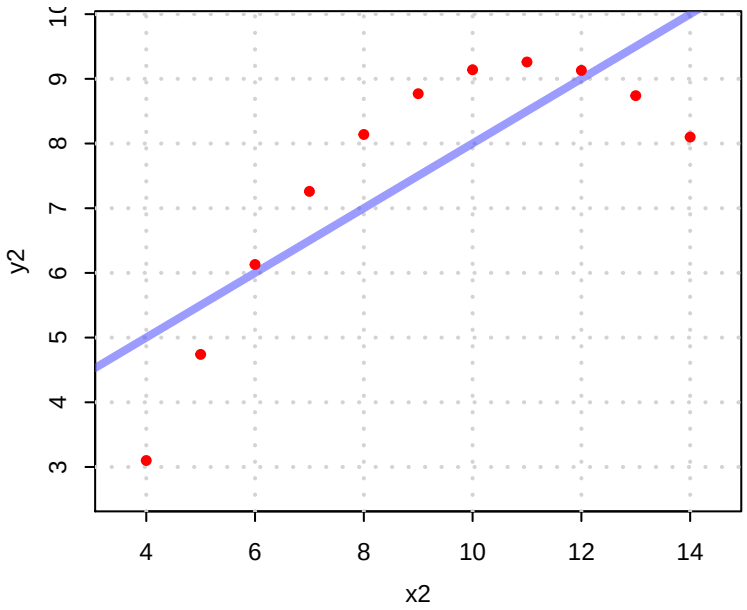
Regression: 4 eindimensionale Beispiele

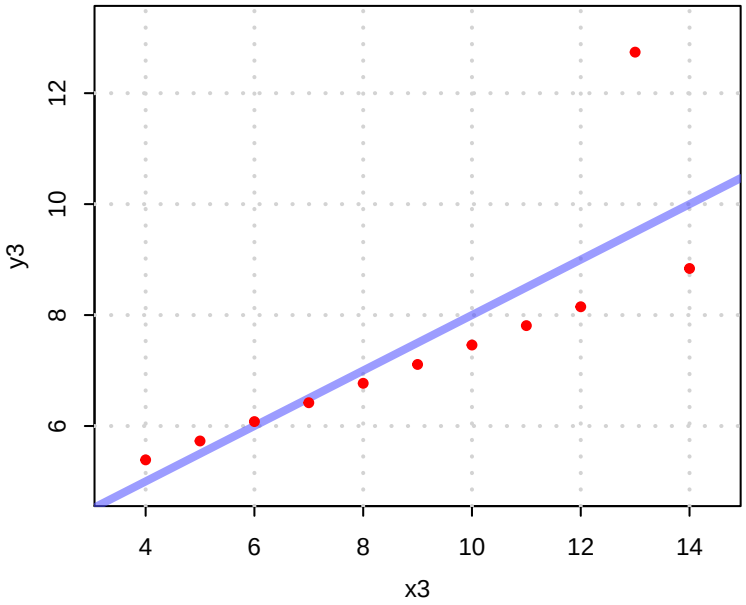
- In folgender Tabelle: Jeweils Ergebnisse der linearen Regressionsanalyse
- dabei: x_k unabhängige Variable und y_k abhängige Variable
- Modell jeweils: $y_k = a_k + b_k x_k$

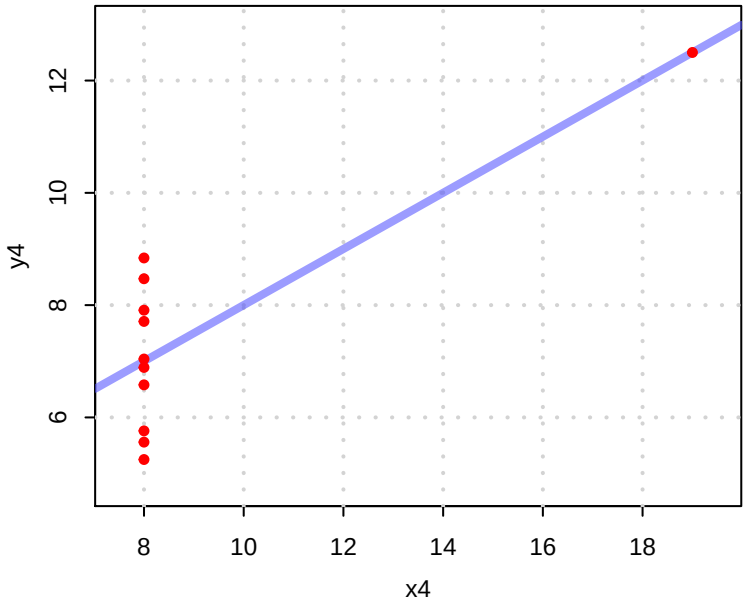
k	$\hat{a}_k$	$\hat{b}_k$	R_k^2
1	3,0001	0,5001	0,6665
2	3,0010	0,5000	0,6662
3	3,0025	0,4997	0,6663
4	3,0017	0,4999	0,6667

Plot der Anscombe-Daten







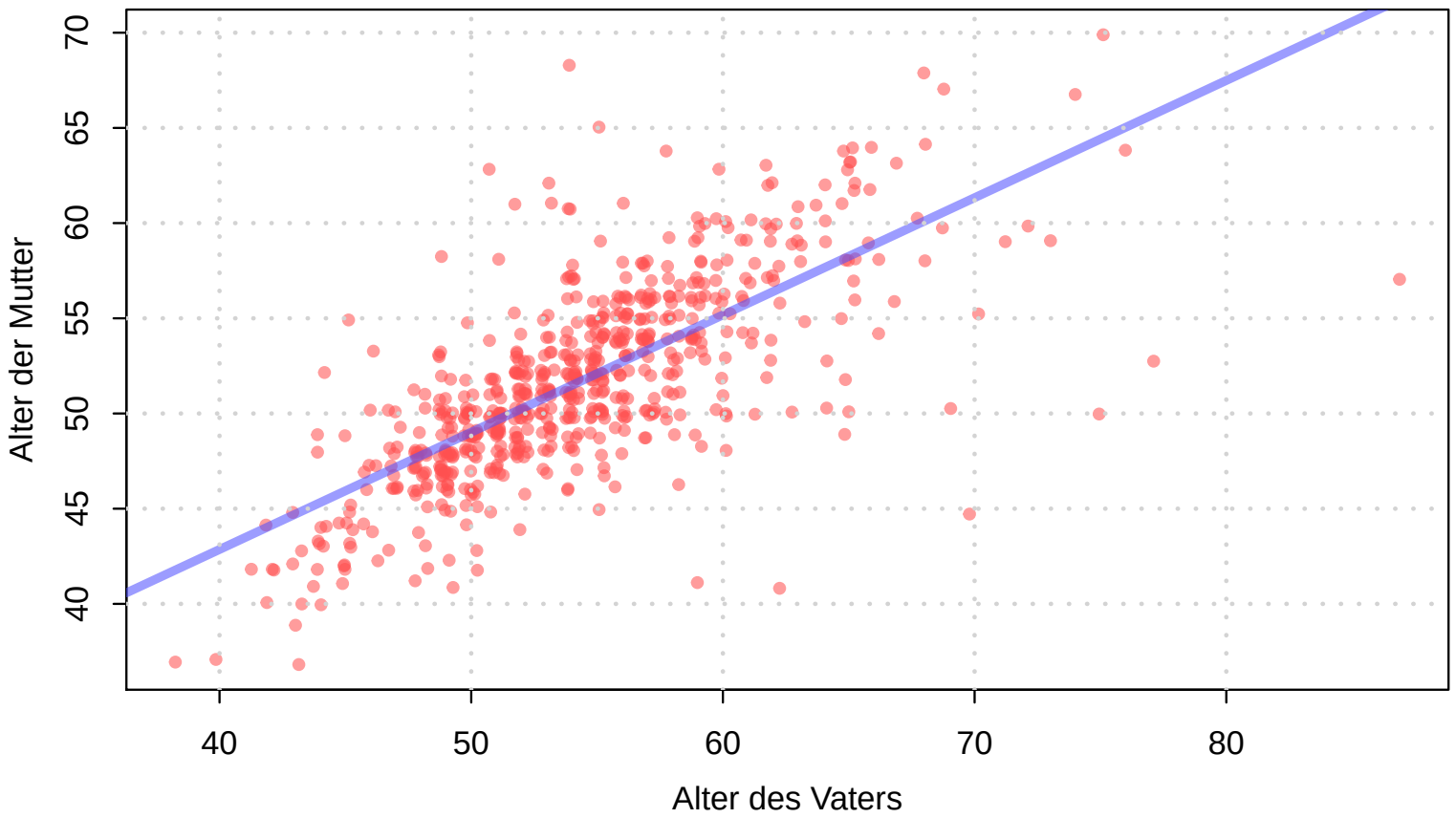


Beispieldaten

```
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
```



Cook's Distanz

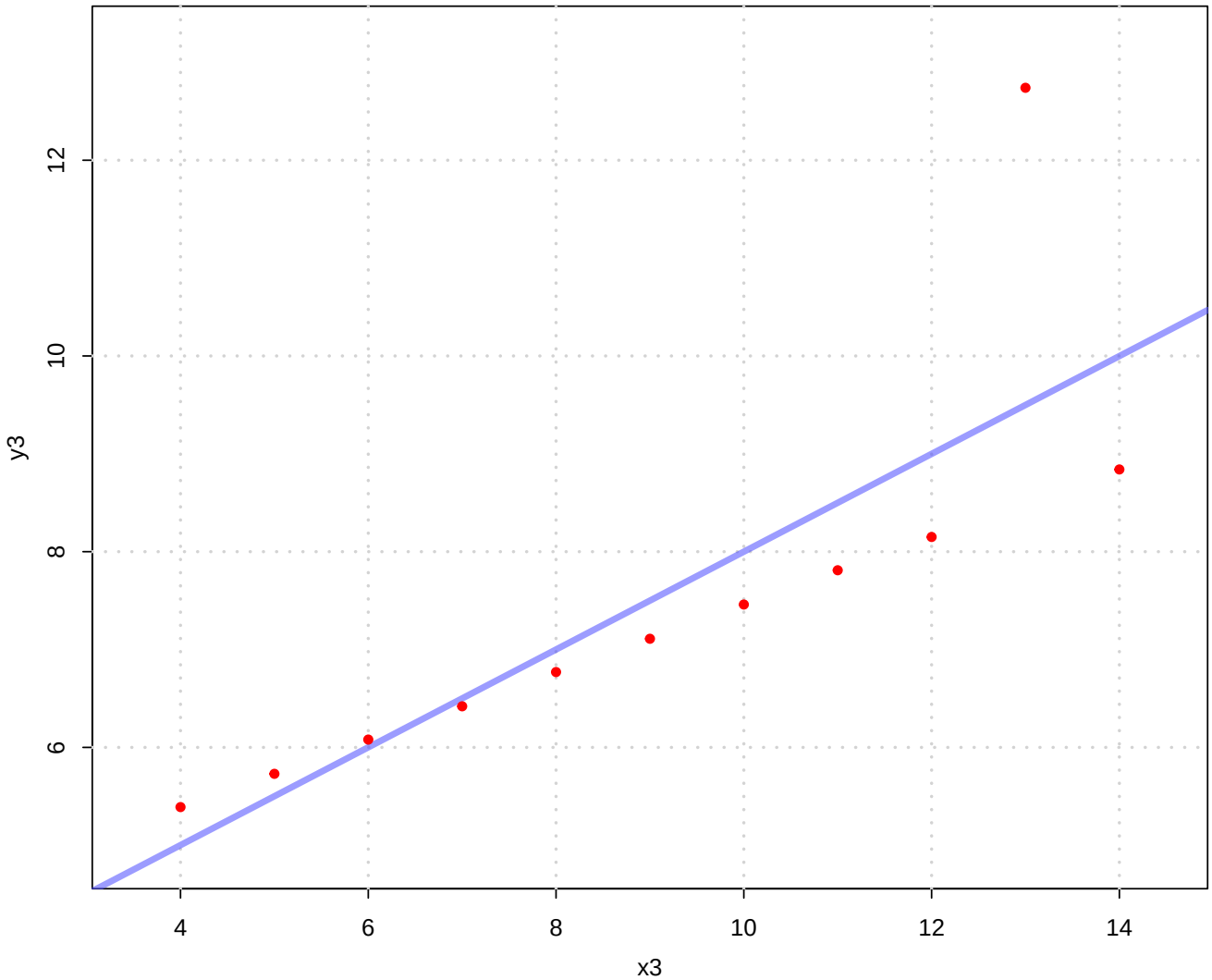
- Oft Kritisch: Einzelne Punkte, die Modell stark beeinflussen
- Idee: Was würde sich ändern, wenn solche Punkte weggelassen würden?
- Cook-Distanz: Misst den Effekt eines gelöschten Objekts
- Formel für ein lineares Modell mit einem unabh. Merkmal:

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

- 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)

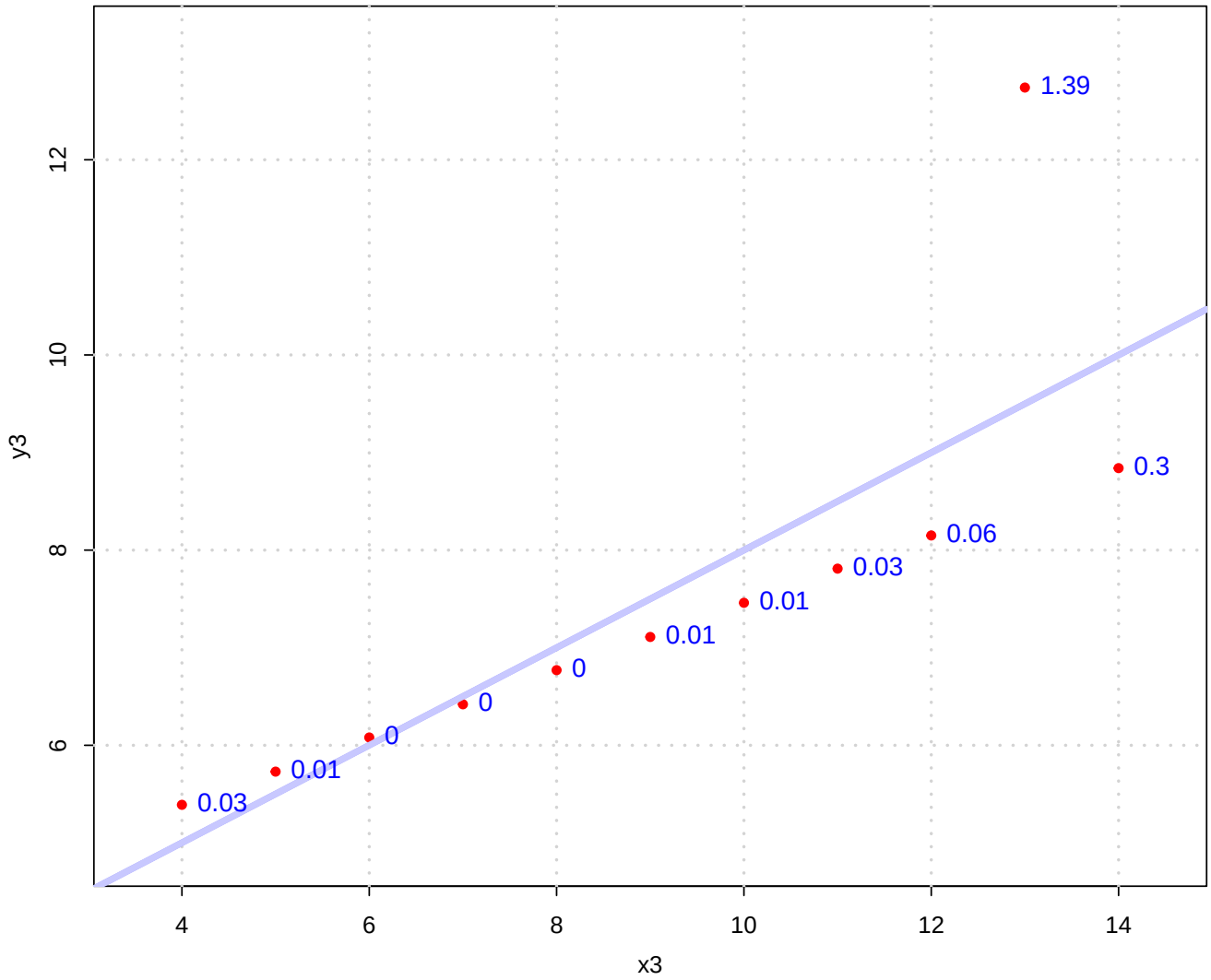
Ausreißer?

- 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

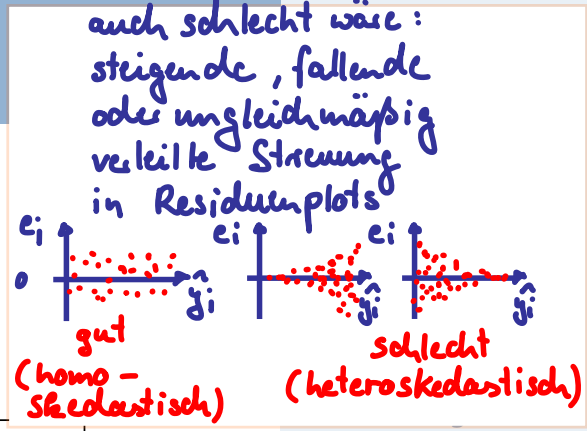


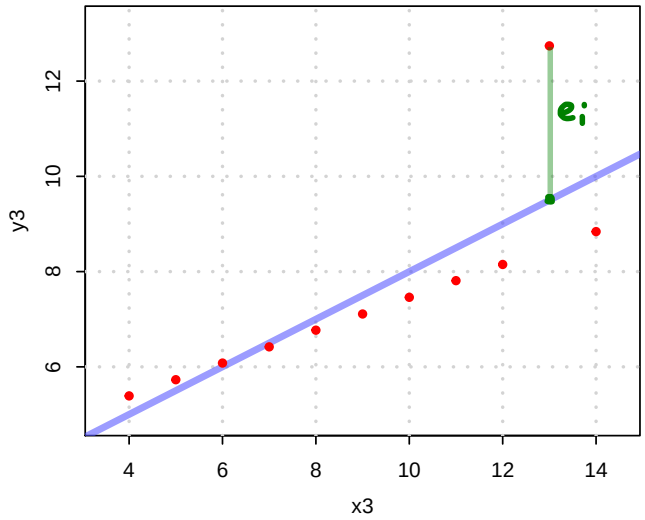
Residualanalyse

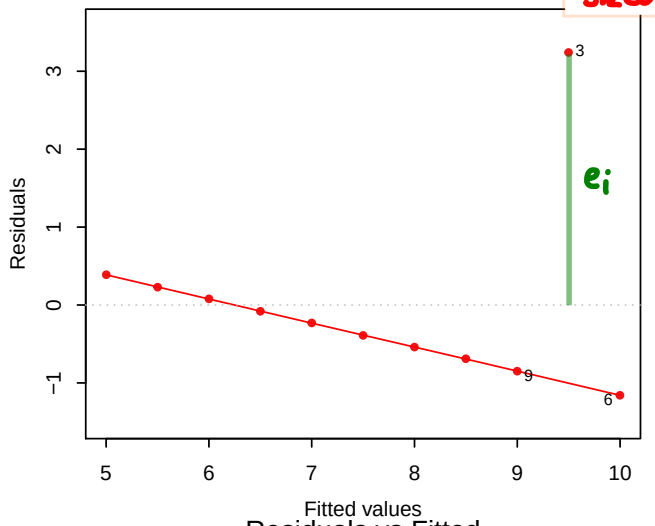
- Oft aufschlussreich: Verteilung der Residuen e_i
- Verbreitet: Graphische Darstellungen der Residuen
- Z.B.: e_i über $\hat{y}_i$

Residualanalyse

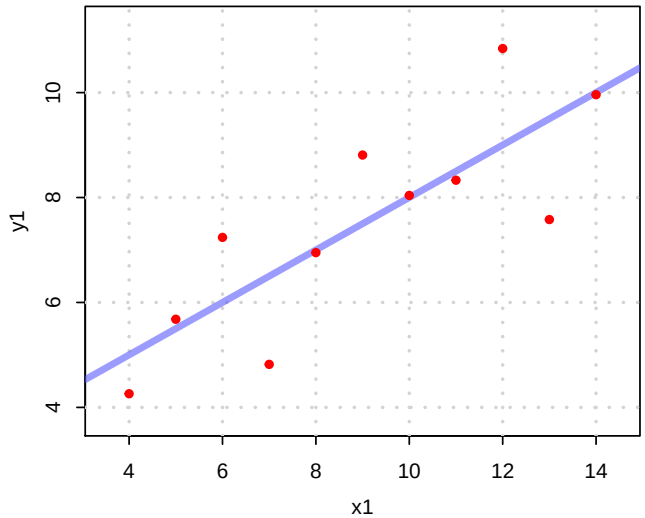
- Oft aufschlussreich: Verteilung der Residuen e_i
- Verbreitet: Graphische Darstellungen der Residuen
- Z.B.: e_i über $\hat{y}_i$

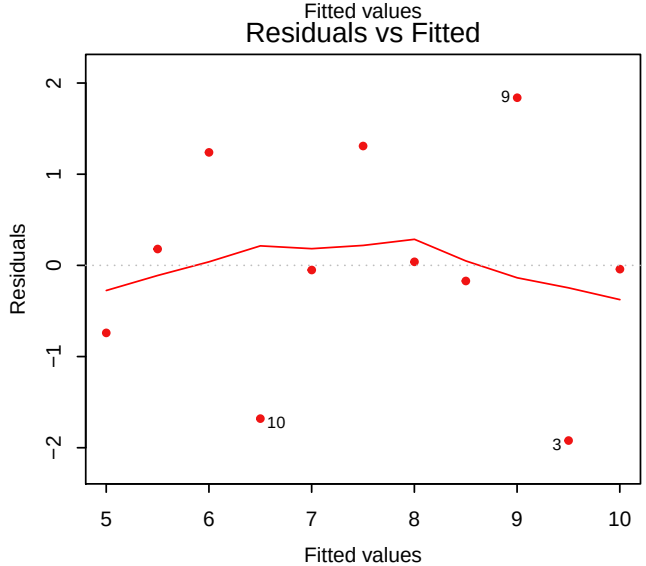






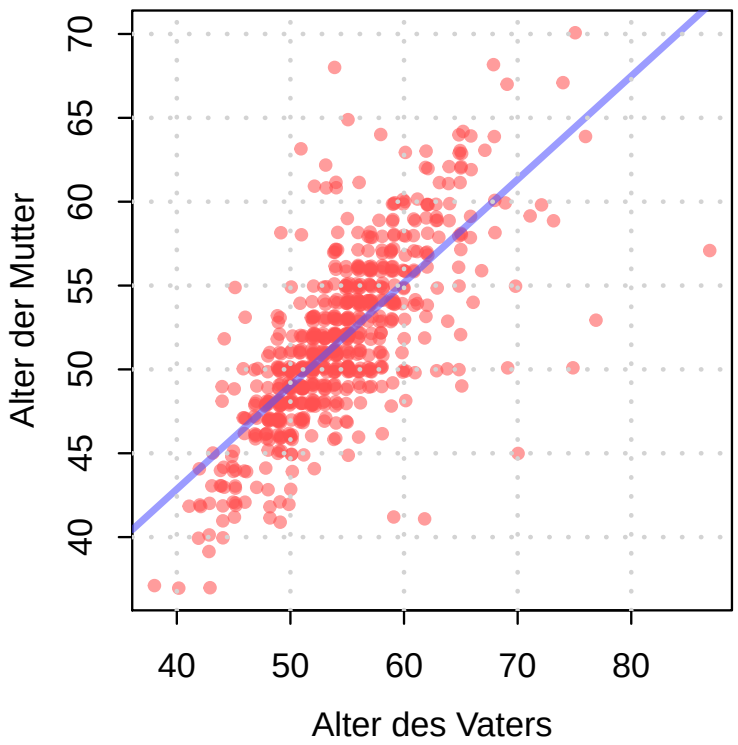
schlecht! Tendenz oder nicht gleichmäßig

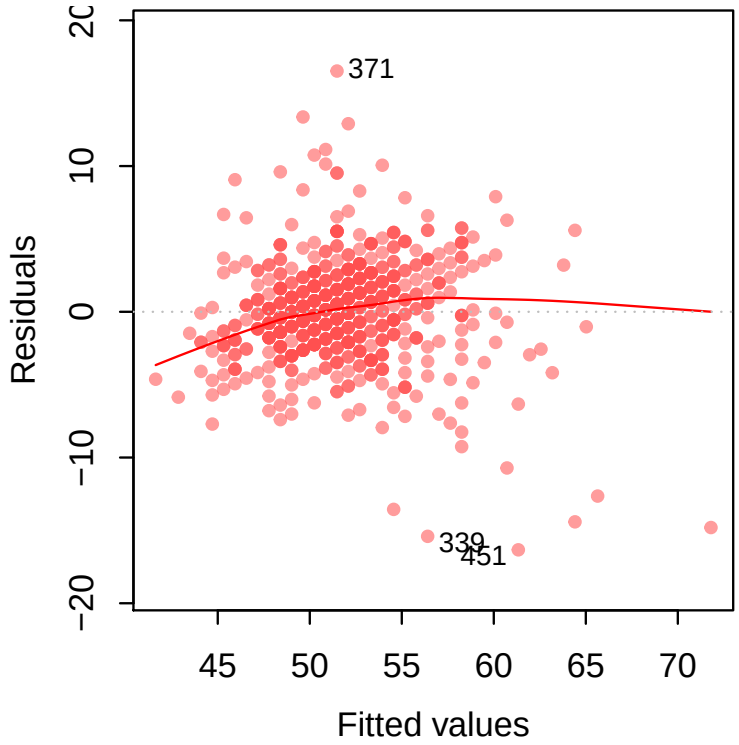


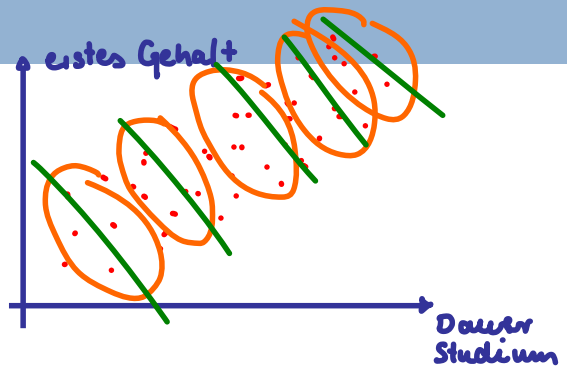


Wichtige Eigenschaften der Residuenverteilung

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







Exkurs: Kausalität vs. Korrelation

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

Statistik: Table of Contents

Kombinatorik

Permutationen ≙ Anordnungen von Objekten

Beispiel: Außendienstmitarbeiter, 50 Stationen bei Kunden; gesucht: Anzahl möglicher Routen

$$50 \cdot 49 \cdot 48 \cdot \ldots \cdot 3 \cdot 2 \cdot 1 = 50! \approx 3 \cdot 10^{64}$$

allgemein: n (unterschiedliche) Objekte kann man auf

$$n!$$

verschiedene Arten anordnen

(R: `factorial(n)`)

Beispiel: M IIII SSSS PP

ges: Anzahl der (unterscheidbaren) Anordnungen

M I SS I SS I PP I
M I SS I SS I PP I
⋮

$4! \cdot 4! \cdot 2!$ ≙ Anzahl der Permutationen pro „Wort", also die nicht unterscheidbaren Variante

- $4!$: Vert. der „I"
- $4!$: Vert. der „S"
- $2!$: Vert. der „P"

⇒ unterscheidbare Buchstabenanordnungen

$$\frac{11!}{4! \cdot 4! \cdot 2!} = 34650$$

allgemein: n Objekte, in k Gruppen zu jeweils $n_1, \ldots, n_k$ nicht unterscheidbaren Elementen

⇒ Anzahl der (unterscheidbaren) Permutationen

$$\frac{n!}{\prod_{i=1}^{k} (n_i!)}$$

Kombinationen ≙ Auswahl von k aus n Elementen

Beispiel: Anzahl der Möglichkeiten für EC-Karten-Pin

$$10 \cdot 10 \cdot 10 \cdot 10 = 10000 = 10^4$$

allgemein: n^k mit Wiederholung, mit Reihenfolge

Beispiel: Top-3-Liste von 100 Mitarbeitern (Bester(r), Zweitbester(r), 3.-Beste(r))

$$100 \cdot 99 \cdot 98 = \frac{100!}{97!} = \frac{100!}{(100-3)!} = 970\,200$$

allgemein: $\frac{n!}{(n-k)!}$ ohne WH, mit Reihenfolge

[TR: $\frac{100!}{(100-3)!} = 100$ nPr $3 = 970\,200$]

Beispiel: Wahl eines 3-köpfigen, gleichberechtigten Teams aus 250 Leuten

$$\frac{250 \cdot 249 \cdot 248}{3!} = \frac{250!}{(250-3)! \cdot 3!} = \binom{250}{3} = 2.573.000$$

allgemein: $\binom{n}{k}$ ohne WH ohne Reihenfolge

[im TR: $\binom{250}{3} = 250$ nCr 3 = 2.573.000]

Beispiel: B, S, W (3 Sorten Getränke)
10 Fläschchen sollen gekauft werden

B B B | S S S S | W W W
B B B B B B B B B B | |
| S S S S S S S S S | W
B | S S | W W W W W W W
F F F | F F F F | F F F

hier: n = 3
k = 10

insgesamt: $\frac{12 \cdot 11}{2 \cdot 1} = 66 = \frac{(10+3-1)!}{2! \cdot 10!}$

allgemein: $\frac{(k+n-1)!}{(n-1)!\,k!} = \binom{n+k-1}{k}$ mit WH ohne R.F.

Kombinatorik: Anzahl von Kombinationen bei Auswahl

2-mal Würfeln, das heißt Auswahl von $k = 2$ aus $n = 6$ Zahlen.

→

(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

- mit WH, mit RF: alle Möglichkeiten, $6^2 = 36$
- ohne WH, mit RF: Diagonale entfällt, $36 - 6 = 30 = 6 \cdot 5 = \frac{6!}{(6-2)!}$
- ohne WH, ohne RF: Hälfte des letzten Ergebnisses: $\frac{30}{2} = 15 = \frac{6!}{4!2!} = \binom{6}{2}$
- mit WH, ohne RF: Letztes Ergebnis plus Diagonale, $15 + 6 = 21 = \binom{7}{2}$

Auswahl von k aus n Dingen		
	mit Wiederholung	ohne Wiederholung
mit Reihenfolge	n^k	$\frac{n!}{(n-k)!}$
ohne Reihenfolge	$\binom{n+k-1}{k}$	$\binom{n}{k}$

- Zufallsvorgang: Geschehen mit ungewissem Ausgang, z.B. Münzwurf
- Elementarereignis ω (kleines omega): Ein möglicher Ausgang, z.B. „ Kopf “
  Elementarereignisse schließen sich gegenseitig aus („ Kopf “ oder „ Zahl “)!
- Ergebnismenge Ω: Menge aller ω
- **Beispiel:** Werfen zweier Würfel:

$$\Omega: \quad \left\{ \begin{array}{cccc} (1,1) & (1,2) & \cdots & (1,6) \\ (2,1) & (2,2) & \cdots & (2,6) \\ \vdots & \vdots & \ddots & \vdots \\ (6,1) & (6,2) & \cdots & (6,6) \end{array} \right\}$$

$$\Rightarrow \Omega = \{(x_1, x_2): \; x_1, x_2 \in \{1, \ldots, 6\}\}$$

- Ereignis A: Folgeerscheinung eines Elementarereignisses
- Formal:

$$A \subset \Omega$$

$\Omega = \{1,2,3\}$
$\omega = 1 \Rightarrow \omega \in \Omega$
$A = \{x \in \Omega : x < 2\} = \{1\}$
$A \subset \Omega$
$A = \{\omega\} \subset \Omega$

- Ereignisse schließen sich nicht gegenseitig aus!
- **Beispiel:** Werfen zweier Würfel:

Ereignis	verbal	formal
A	Augensumme $= 4$	$\{(1,3),\ (2,2),\ (3,1)\}$
B	Erste Zahl $= 2$	$\{(2,1),\ (2,2),\ \ldots,\ (2,6)\}$

- Wahrscheinlichkeit $P(A)$: Chance für das Eintreten von A
- Laplace-Wahrscheinlichkeit:

$$P(A) = \frac{|A|}{|\Omega|} = \frac{\text{Anzahl der für } A \text{ günstigen Fälle}}{\text{Anzahl aller möglichen Fälle}}$$