Statistik

für Betriebswirtschaft, Internationales Management, Wirtschaftsinformatik und Informatik

Sommersemester 2017

HSA Statistik SS 2017 Sessionlist		
Datum	Statistik für BW/IM/I/Winf	Nr.
15.03.2017	Einführung Statistik	1
22.03.2017	Differentialrechnung, 2-dim Diff.Rechnung	2
29.03.2017	univ. deskr. Stat., Quantile, Plots	3
05.04.2017	Streuung, Konzentrationsmaße	4
12.04.2017	Kontingenztabellen, Mosaikplots, Korrelation	5
19.04.2017	Preisindizes, lineare Regression	6
26.04.2017	Wahrscheinlichkeitsbegriff	7
03.05.2017	Bedingte Wahrscheinlichkeit, Bayes	8
10.05.2017	diskrete Zufallsvariablen	9
17.05.2017	Stetige ZV, Gleichverteilung	10
24.05.2017	Pyramid	
31.05.2017	Normalverteilung, Verteilungsparameter	11
07.06.2017	Schätzfunktionen und Punktschätzer	12
14.06.2017	Konfidenzintervalle	13
21.06.2017	Wiederholung, Besprechung Probeklausur	14
28.06.2017	Prüfungswoche	15

Stefan Etschberger

Kontingenzkoeffizient

- Gegeben: Kontingenztabelle mit k Zeilen und l Spalten (vgl. hier)
- Vorgehensweise:
  1. Ergänze Randhäufigkeiten

$$h_{i\cdot} = \sum_{j=1}^{l} h_{ij} \qquad \text{und} \qquad h_{\cdot j} = \sum_{i=1}^{k} h_{ij}$$

  2. Berechne theoretische Häufigkeiten

$$\tilde{h}_{ij} = \frac{h_{i\cdot} \cdot h_{\cdot j}}{n}$$

  3. Berechne

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{(h_{ij} - \tilde{h}_{ij})^2}{\tilde{h}_{ij}}$$

χ^2 hängt von n ab! $(h_{ij} \mapsto 2 \cdot h_{ij} \Rightarrow \chi^2 \mapsto 2 \cdot \chi^2)$

	Verletzung			
	leicht	schwer	tödlich	
angegurtet	264 (239.4)	90 (111.6)	6 (9)	360
nicht angegurtet	2 (26.6)	34 (12.4)	4 (1)	40
	266	124	10	400

$266 \cdot \frac{360}{400}$

$266 \cdot \frac{40}{400}$

„chi-Quadrat"

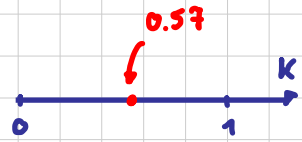
$$\chi^2 = \frac{(264-239.4)^2}{239.4} + \frac{(90-111.6)^2}{111.6} + \frac{(6-9)^2}{9}$$

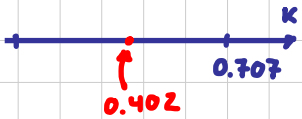
$$+ \frac{(2-26.6)^2}{26.6} + \frac{(34-12.4)^2}{12.4} + \frac{(4-1)^2}{1}$$

$$\approx 77.08$$

	leicht	schwer	tot	
Gurt	264	90	6	360
kein Gurt	2	34	4	40
	266	124	10	400

	leicht	schwer	tot	
Gurt	239,4	111,6	9	360
kein Gurt	26,6	12,4	1	40
	266	124	10	400





	leicht	schwer	tot	
Gurt	2,52781955	4,18064516	1	
kein Gurt	22,7503759	37,6258065	9	
chi^2	77,0846471			
K	0,40196313			
Kmax	0,70710678			
K*	0,57			

④ Kontingenzkoeffizient:

$$K = \sqrt{\frac{\chi^2}{n+\chi^2}} \qquad \in [0; K_{max}]$$

wobei

$$K_{max} = \sqrt{\frac{M-1}{M}} \qquad \text{mit} \qquad M = \min\{k, l\}$$

⑤ Normierter Kontingenzkoeffizient:

$$K_* = \frac{K}{K_{max}} \qquad \in [0; 1]$$

$K_* = +1 \iff$
bei Kenntnis von x_i kann y_i erschlossen werden u.u.

Beispiel

X : Staatsangehörigkeit (d,a)
Y : Geschlecht (m,w)

h_{ij}	m	w	$h_{i\cdot}$
d	30	30	60
a	10	30	40
$h_{\cdot j}$	40	60	100

$\Rightarrow$

$\tilde{h}_{ij}$	m	w
d	24	36
a	16	24

wobei $\tilde{h}_{11} = \frac{60 \cdot 40}{100} = 24$ usw.

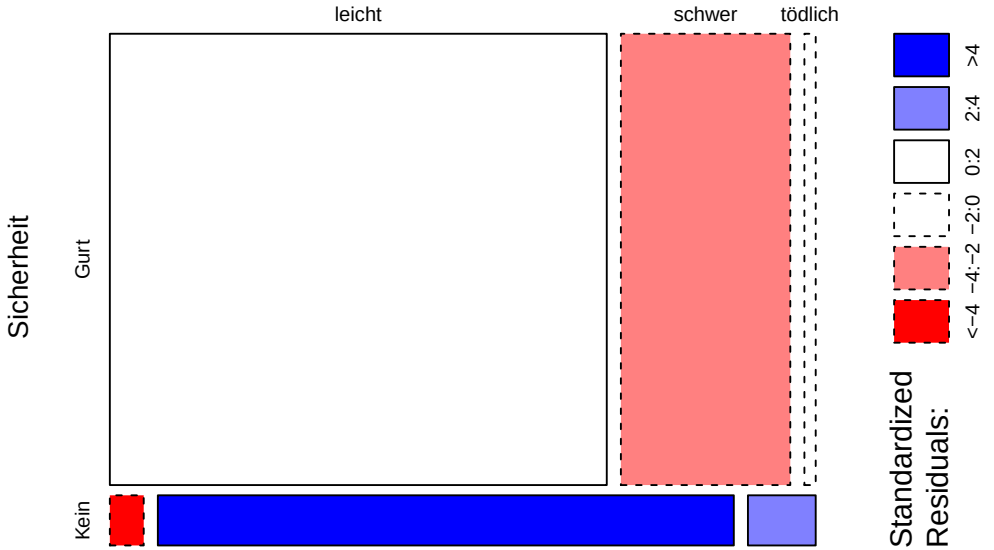
$$\chi^2 = \frac{(30-24)^2}{24} + \frac{(30-36)^2}{36} + \frac{(10-16)^2}{16} + \frac{(30-24)^2}{24} = 6{,}25$$

$$K = \sqrt{\frac{6{,}25}{100+6{,}25}} = 0{,}2425; \quad M = \min\{2,2\} = 2; \quad K_{max} = \sqrt{\frac{2-1}{2}} = 0{,}7071$$

$$K_* = \frac{0{,}2425}{0{,}7071} = 0{,}3430$$

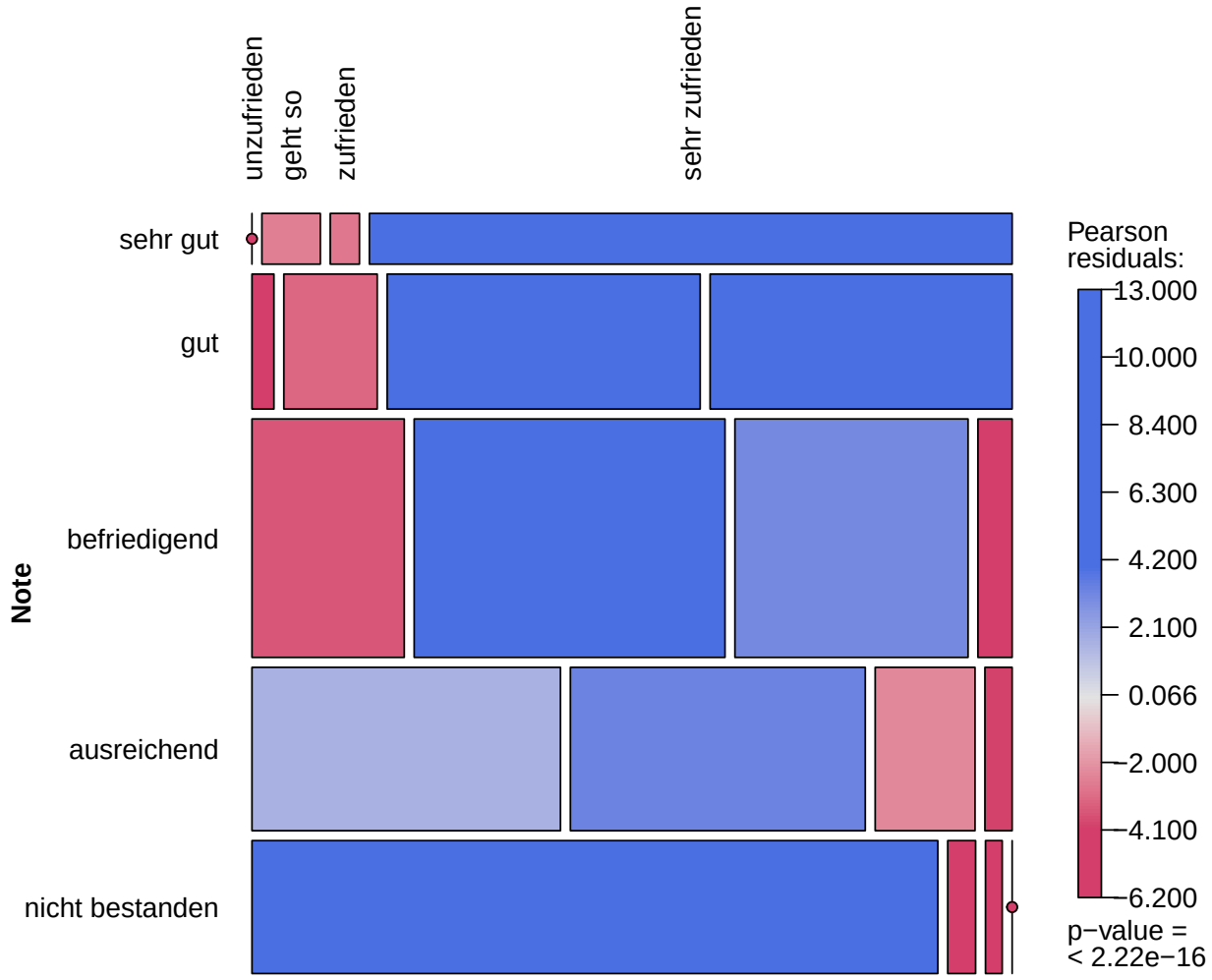
Beispiel Autounfälle

	Verletzung			
	leicht	schwer	tödlich	
angegurtet	264	90	6	360
nicht angegurtet	2	34	4	40
	266	124	10	400



Mosaikplot Autounfälle

```r
Data.complete = na.omit(MyData[,c("MatheZufr", "NoteMathe")])
Noten.complete =
  ordered(cut(Data.complete$NoteMathe, breaks=c(0,1.5,2.5,3.5,4.1,5.0)),
          labels=c("sehr gut", "gut", "befriedigend", "ausreichend", "nicht bestanden"))
tab = table("Note"=Noten.complete, "Zufrieden mit Leistung"=Data.complete$MatheZufr)
require(vcd)
mosaic(tab, shade = TRUE, gp_args = list(interpolate = function(x) pmin(x/4, 1)), labeling_args =
   list(rot_labels = c(90,0,0,0), just_labels = c("left", "left", "right", "right"),
        offset_varnames = c(left = 5, top=5.5), offset_labels = c(right = 3)),
 margins = c(right = 1, bottom = 3, left=6, top=5))
```



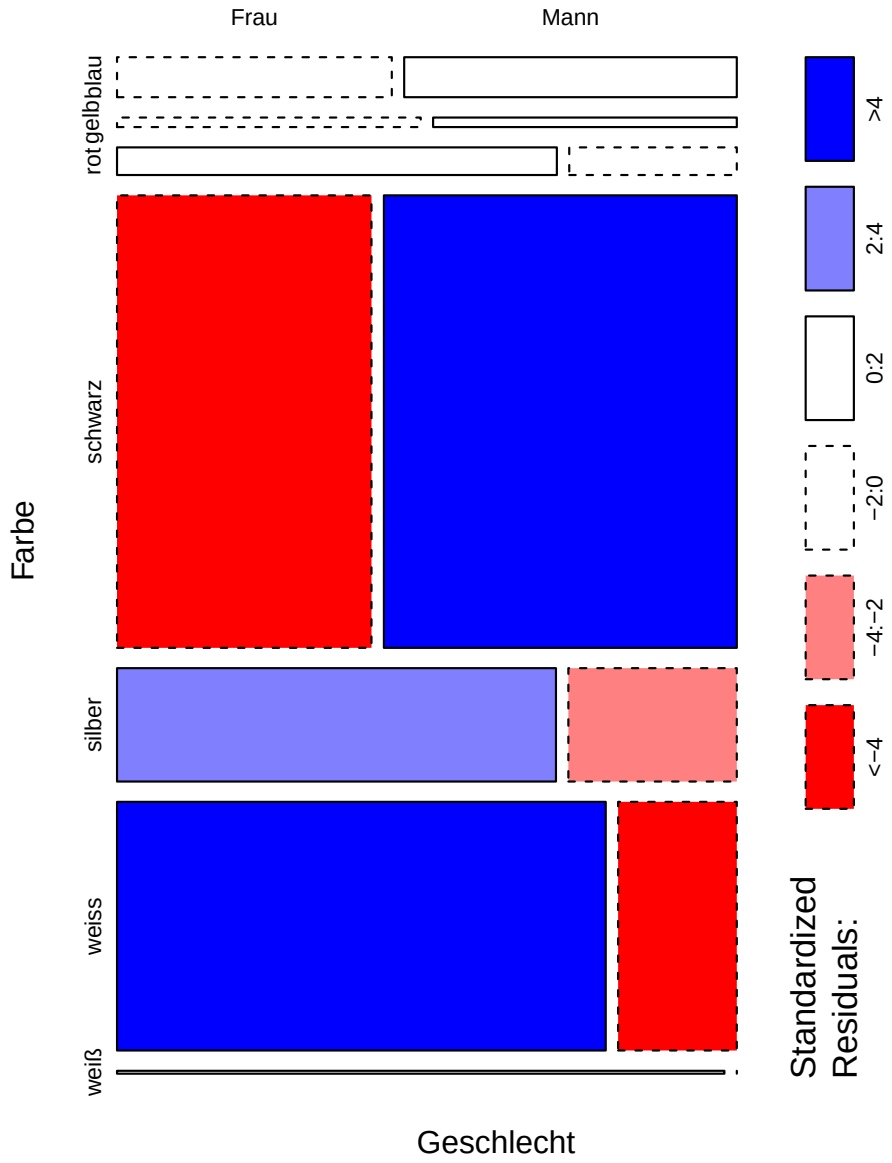
„Note in Matheklausur“ gegen „Zufrieden mit Leistung“

Mosaicplot Geschlecht, Wunschfarbe für Smartphone

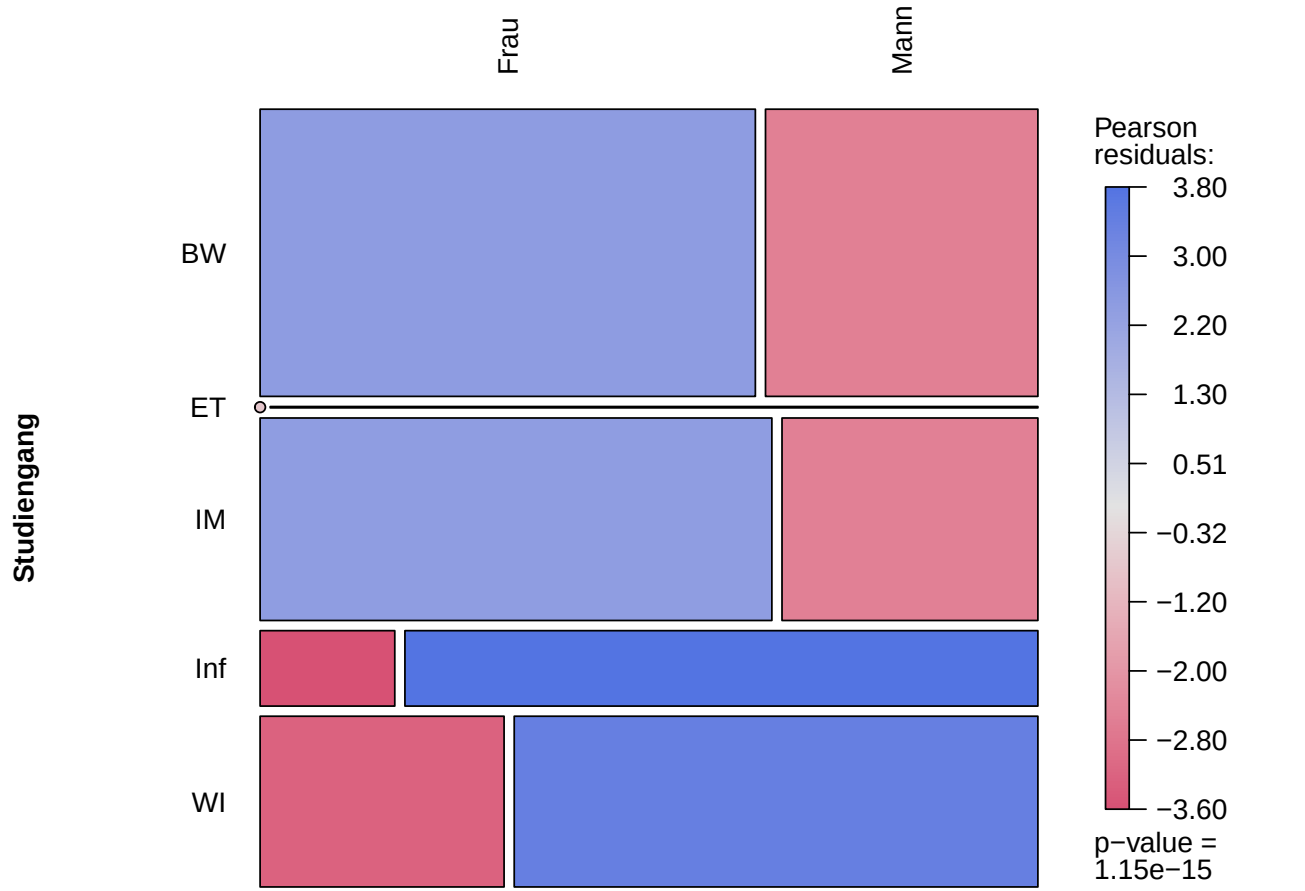
```
tab = table(Farbe, Geschlecht)
tab
```

```
##          Geschlecht
## Farbe     Frau Mann
##   blau      19   23
##   gelb       5    5
##   rot       21    8
##   schwarz  199  276
##   silber    86   33
##   weiss    210   51
##   weiß       3    0
```

```
mosaicplot(t(tab), shade = TRUE,
           sort=2:1, main="")
```




```
require(vcd)
Data.complete = na.omit(MyData[,c("Geschlecht", "Studiengang")])
with(Data.complete, {
  tab = table("Studiengang"=Studiengang, "Geschlecht"=Geschlecht)
mosaic(tab, shade = TRUE, gp_args = list(interpolate = function(x) pmin(x/4, 1)), labeling_args =
   list(rot_labels = c(90,0,0,0), just_labels = c("left", "left", "right", "right"),
        offset_varnames = c(left = 5, top=5.5), offset_labels = c(right = 3)),
 margins = c(right = 1, bottom = 3, left=6, top=5))
})
```



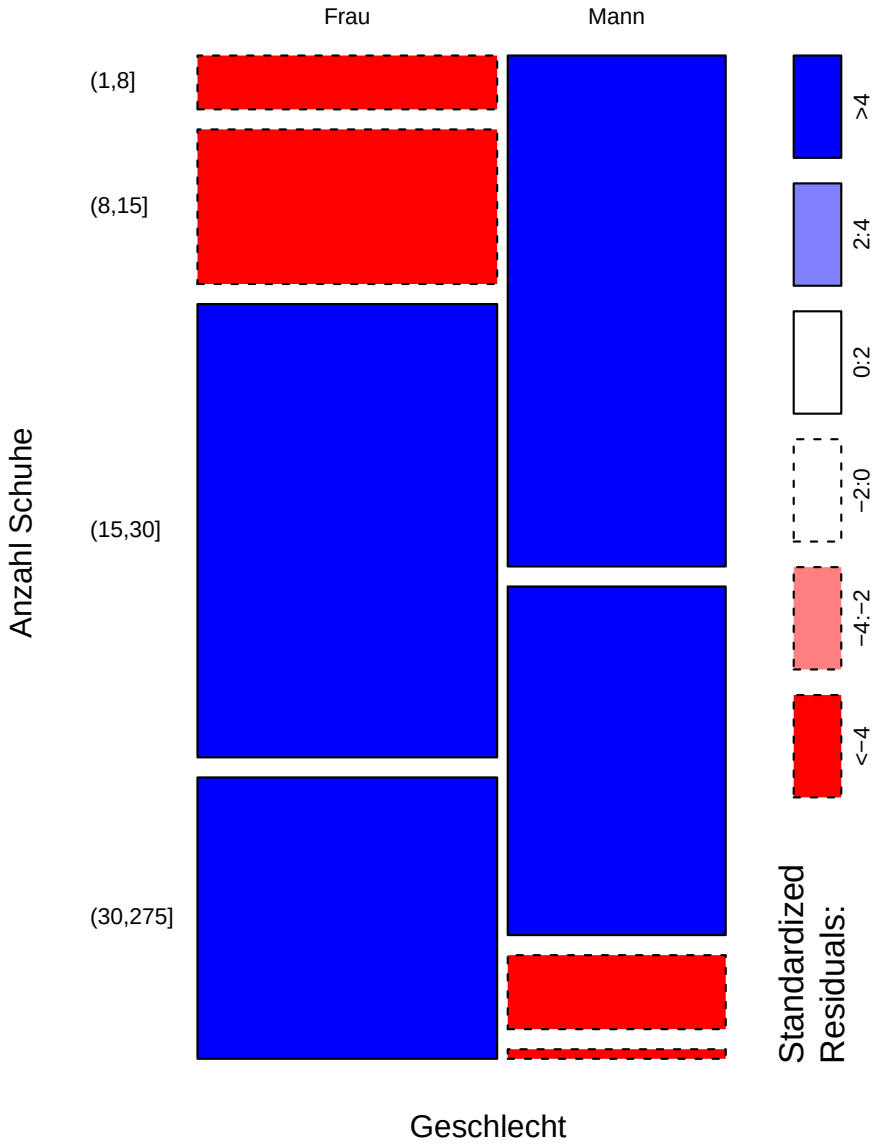
~~„Note in Matheklausur“ gegen „Zufrieden mit Leistung“~~

Mosaicplot Geschlecht, Anzahl Schuhe

```r
tab = table(
  "Anzahl Schuhe" =
  cut(AnzSchuhe,
      breaks =
        quantile(
          AnzSchuhe,
          probs = (0:4)/4
        )
  ),
  Geschlecht)
tab
```

```
##                Geschlecht
## Anzahl Schuhe Frau Mann
##       (1,8]      31   214
##       (8,15]     89   146
##       (15,30]   261    31
##       (30,275]  162     4
```

```r
mosaicplot(t(tab), shade = TRUE,
           main="", las=1)
```



- Preismesszahl: Misst Preisveränderung eines einzelnen Gutes:

$$\frac{\text{Preis zum Zeitpunkt } j}{\text{Preis zum Zeitpunkt } i}$$

dabei: j: Berichtsperiode, i: Basisperiode

- Preisindex: Misst Preisveränderung mehrerer Güter (Aggregation von Preismesszahlen durch Gewichtung)
- Notation:

„damals"

$p_0(i)$: Preis des i-ten Gutes in Basisperiode 0
$p_t(i)$: Preis des i-ten Gutes in Berichtsperiode t „heute"
$q_0(i)$: Menge des i-ten Gutes in Basisperiode 0
$q_t(i)$: Menge des i-ten Gutes in Berichtsperiode t

Preisindizes

- Gleichgewichteter Preisindex:

$$P_{0t}^{G} = \frac{1}{n}\sum_{i=1}^{n}\frac{p_t(i)}{p_0(i)} = \sum_{i=1}^{n}\frac{p_t(i)}{p_0(i)}\cdot g(i) \qquad \text{mit} \qquad g(i) = \frac{1}{n}$$

Nachteil: Auto und Streichhölzer haben gleiches Gewicht
Lösung: Preise mit Mengen gewichten!

- Preisindex von Laspeyres:

$$P_{0t}^{L} = \frac{\sum_{i=1}^{n} p_t(i)q_0(i)}{\sum_{i=1}^{n} p_0(i)q_0(i)} = \sum_{i=1}^{n}\frac{p_t(i)}{p_0(i)}\cdot g_0(i) \qquad \text{mit} \qquad g_0(i) = \frac{p_0(i)\,q_0(i)}{\sum_{j=1}^{n} p_0(j)\,q_0(j)}$$

- Preisindex von Paasche:

$$P_{0t}^{P} = \frac{\sum_{i=1}^{n} p_t(i)q_t(i)}{\sum_{i=1}^{n} p_0(i)q_t(i)} = \sum_{i=1}^{n}\frac{p_t(i)}{p_0(i)}\cdot g_t(i) \qquad \text{mit} \qquad g_t(i) = \frac{p_0(i)\,q_t(i)}{\sum_{j=1}^{n} p_0(j)\,q_t(j)}$$

Warenkorb: Kartoffeln und Kaffee

	1950		2013	
	Preis (€)	Menge pro Woche	Preis (€)	Menge pro Woche
1 kg Kartoffeln	0,04	3,58	1,10	1,25
100 g Kaffeebohnen	3,00	0,25	0,70	1,31

$$P^{L}_{1950,2013} = \frac{1{,}10 \cdot 3{,}58 + 0{,}70 \cdot 0{,}25}{0{,}04 \cdot 3{,}58 + 3{,}00 \cdot 0{,}25} \approx 4{,}6048$$

d.h. 360% Preissteigerung in 63 Jahren

durchschnittlich pro Jahr

$\sqrt[63]{4.6048} \approx 1.0245$

$\hat{=}$ 2,45 % „Inflation" pro Jahr

$$P^{P}_{1950,2013} = \frac{1{,}10 \cdot 1{,}25 + 0{,}70 \cdot 1{,}31}{0{,}04 \cdot 1{,}25 + 3{,}00 \cdot 1{,}31} \approx 0{,}5759$$

d.h. -42.4 % in 63 Jahren

pro Jahr im Durchschnitt

$\sqrt[63]{0.5759} \approx 0.9913$

$\hat{=}$ -0.0087

$\hat{=}$ -0.87 %

Idealindex von Fisher:

$$P_{0t}^{F} = \sqrt{P_{0t}^{L} P_{0t}^{P}}$$

Marshall-Edgeworth-Index:

$$P_{0t}^{ME} = \frac{\sum_{i=1}^{n} p_t(i)[q_0(i) + q_t(i)]}{\sum_{i=1}^{n} p_0(i)[q_0(i) + q_t(i)]}$$

Preisindex von Lowe:

$$P_{0t}^{LO} = \frac{\sum_{i=1}^{n} p_t(i) q(i)}{\sum_{i=1}^{n} p_0(i) q(i)}$$

wobei $q(i) \triangleq$ { Durchschn. Menge von Gut i über alle (bekannten) Perioden

Warenkorb: Kartoffeln und Kaffee

	1950		2013	
	Preis (€)	Menge pro Woche	Preis (€)	Menge pro Woche
1 kg Kartoffeln	0,04	3,58	1,10	1,25
100 g Kaffeebohnen	3,00	0,25	0,70	1,31

$$P^{F}_{1950,2013} \approx \sqrt{4{,}6048 \cdot 0{,}5759} = 1{,}6284$$

$$P^{ME}_{1950,2013} = \frac{1{,}10 \cdot (3{,}58 + 1{,}25) + 0{,}70 \cdot (0{,}25 + 1{,}31)}{0{,}04 \cdot (3{,}58 + 1{,}25) + 3{,}00 \cdot (0{,}25 + 1{,}31)} = 1{,}3143$$

$$P^{Lo}_{1950,2013} = \frac{1{,}10 \cdot 2{,}5 + 0{,}70 \cdot 0{,}75}{0{,}04 \cdot 2{,}5 + 3{,}00 \cdot 0{,}75} = 1{,}3936$$

Annahme bei P^{LO}: Durchschn. Mengen bei Kartoffeln bzw. Kaffebohnen von 1950 bis 2013 sind 2,5 bzw. 0,75.

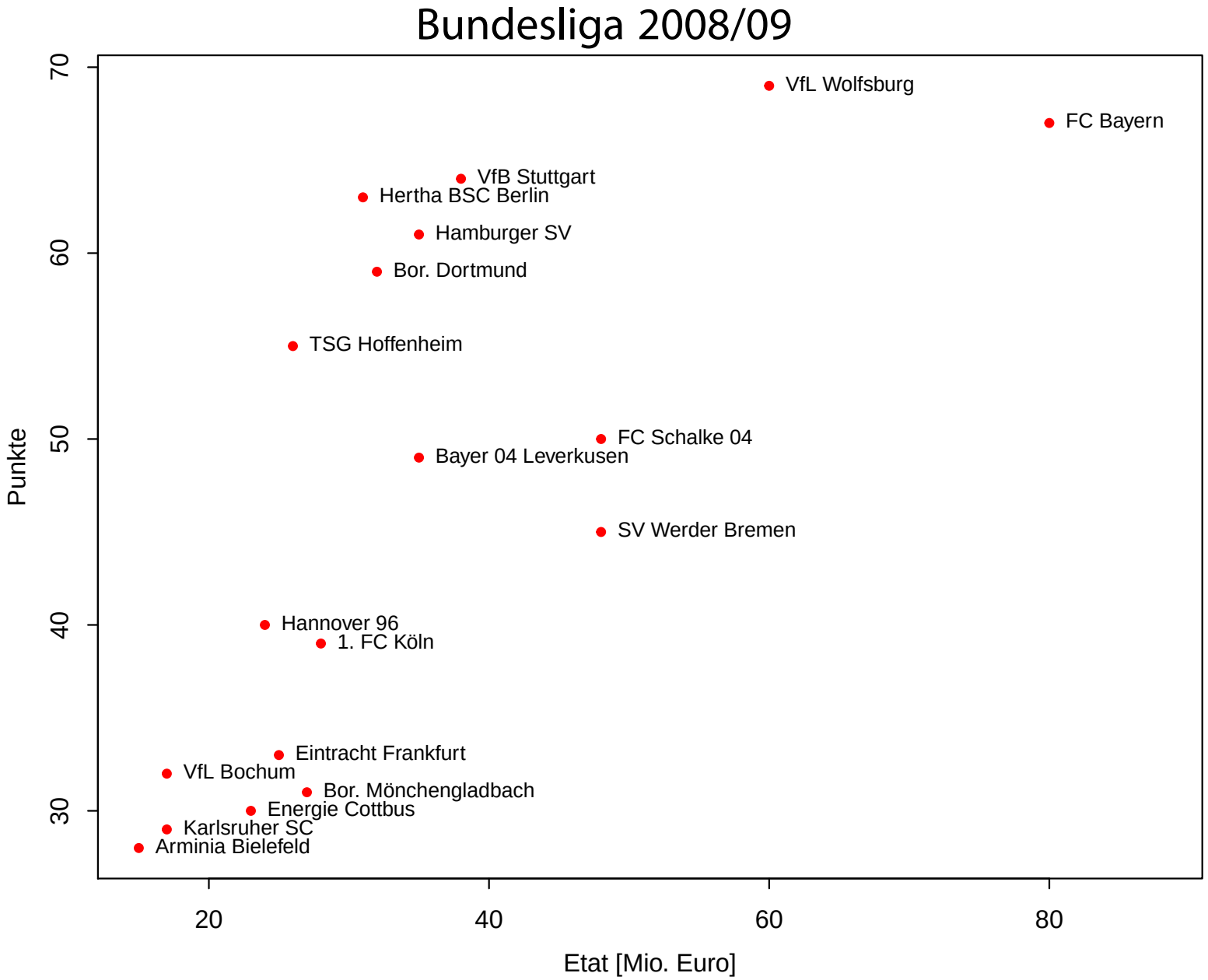
Bundesliga 2008/2009

- 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

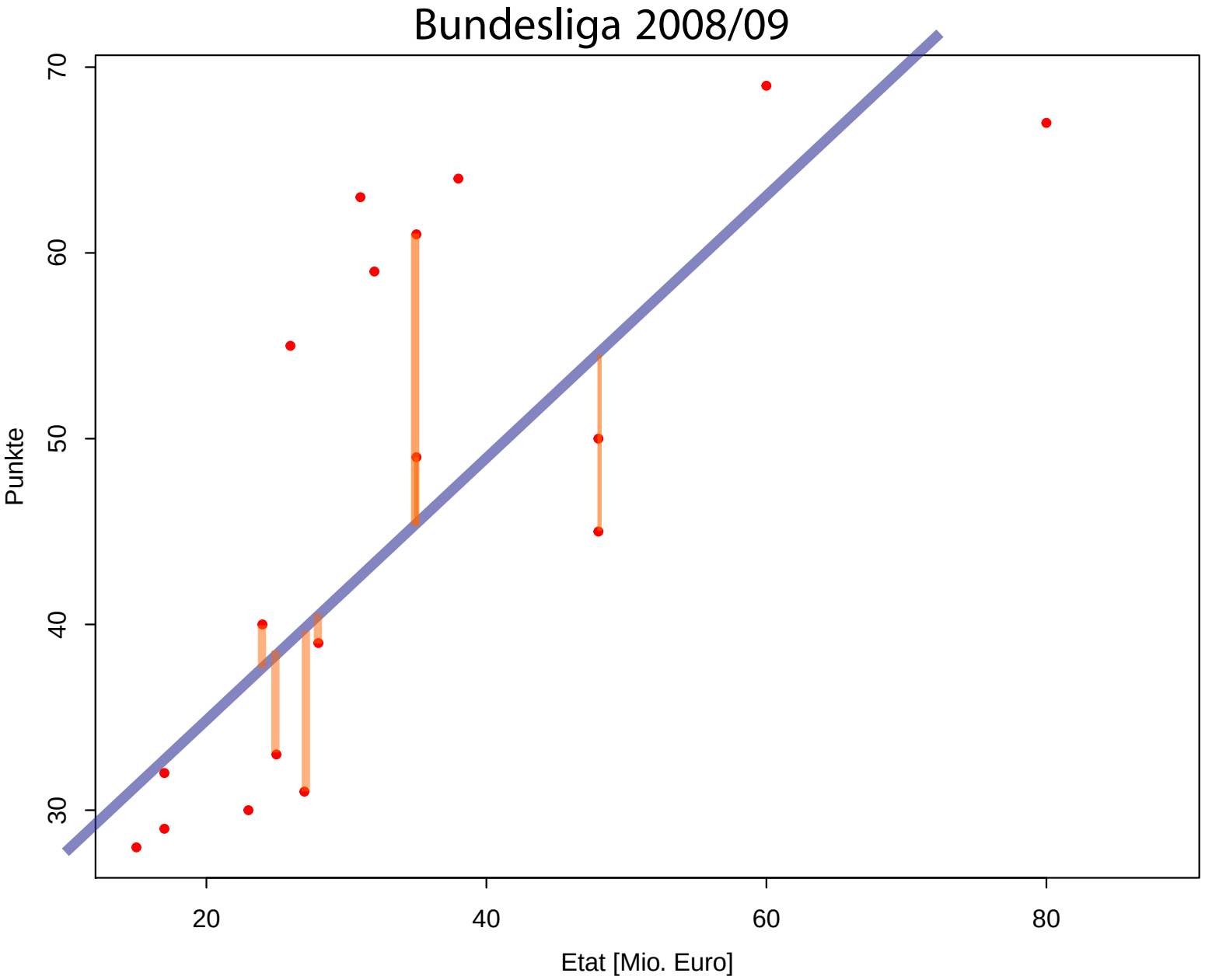
	Etat	Punkte
FC Bayern	80	67
VfL Wolfsburg	60	69
SV Werder Bremen	48	45
FC Schalke 04	48	50
VfB Stuttgart	38	64
Hamburger SV	35	61
Bayer 04 Leverkusen	35	49
Bor. Dortmund	32	59
Hertha BSC Berlin	31	63
1. FC Köln	28	39
Bor. Mönchengladbach	27	31
TSG Hoffenheim	26	55
Eintracht Frankfurt	25	33
Hannover 96	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

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

Beste Lösung

- Beste und eindeutige Lösung:

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

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

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

- Regressionsgerade:

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

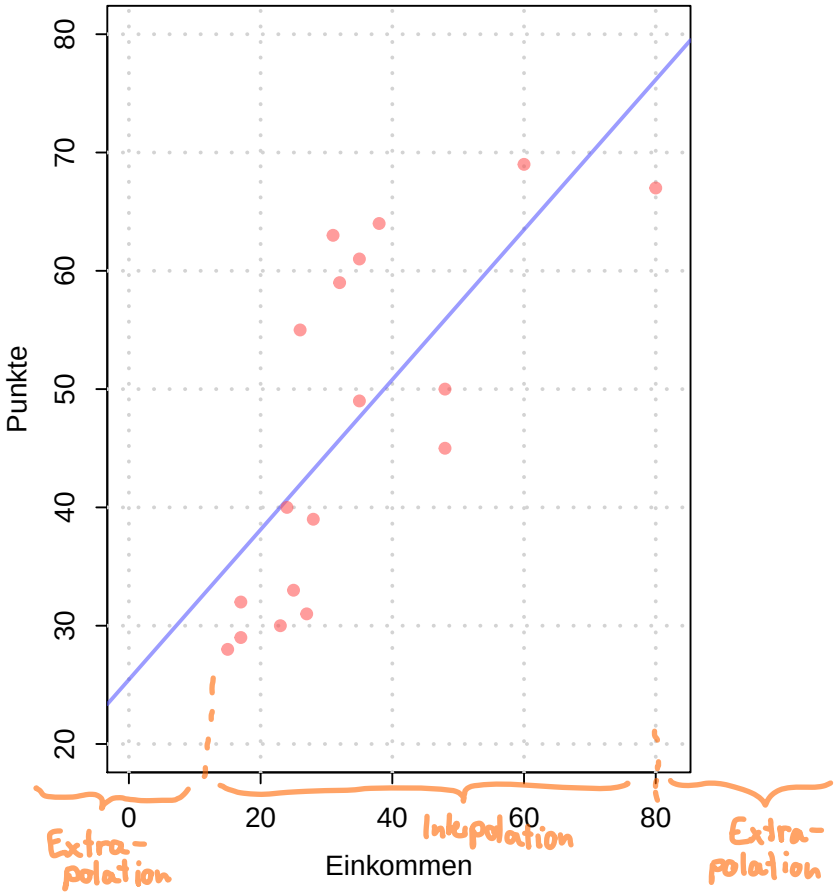
Bundesligabeispiel

- 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

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

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

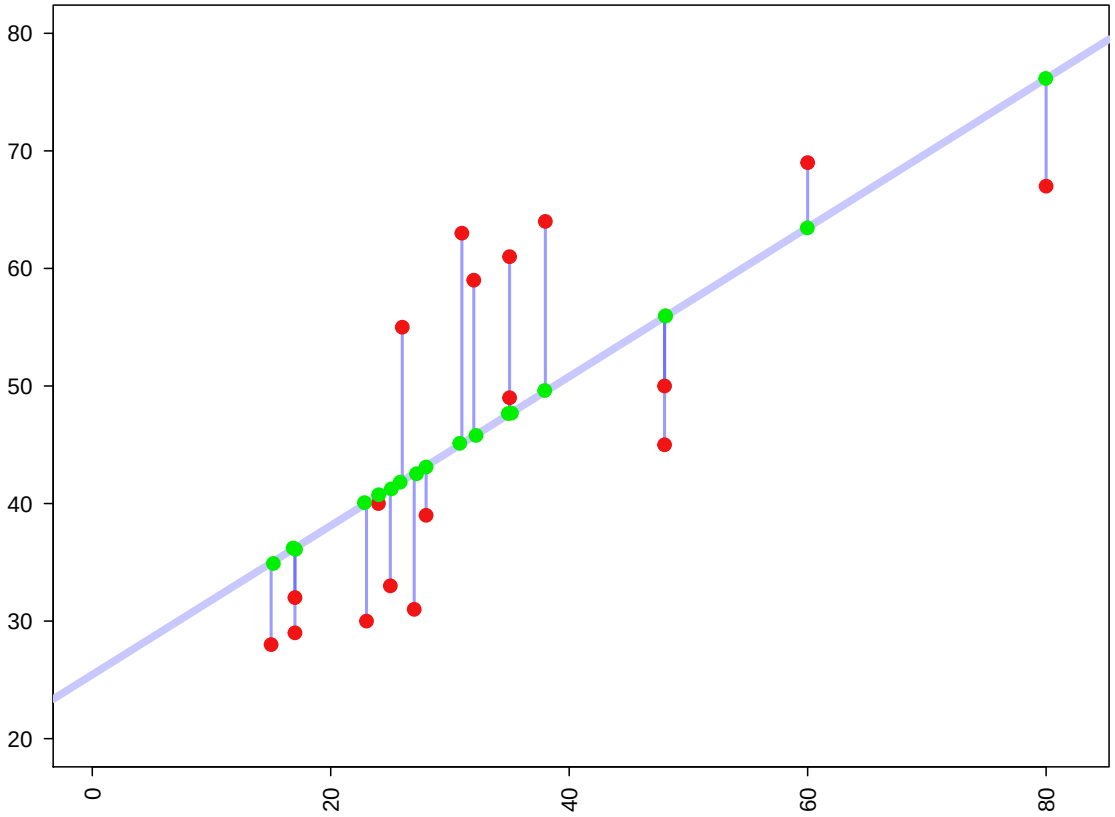


- Prognosewert für Etat $= 30$:

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

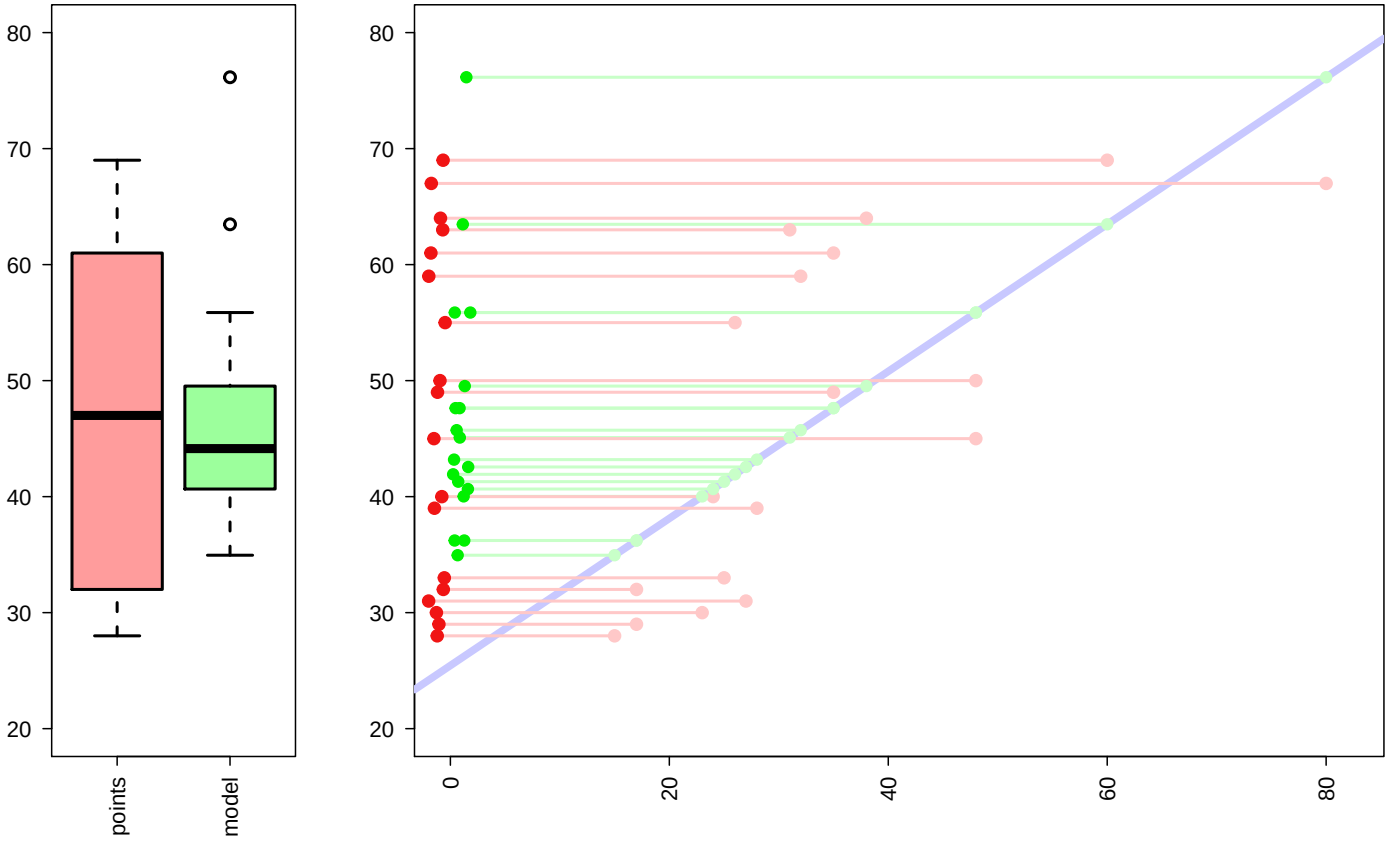
Varianz und Information

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



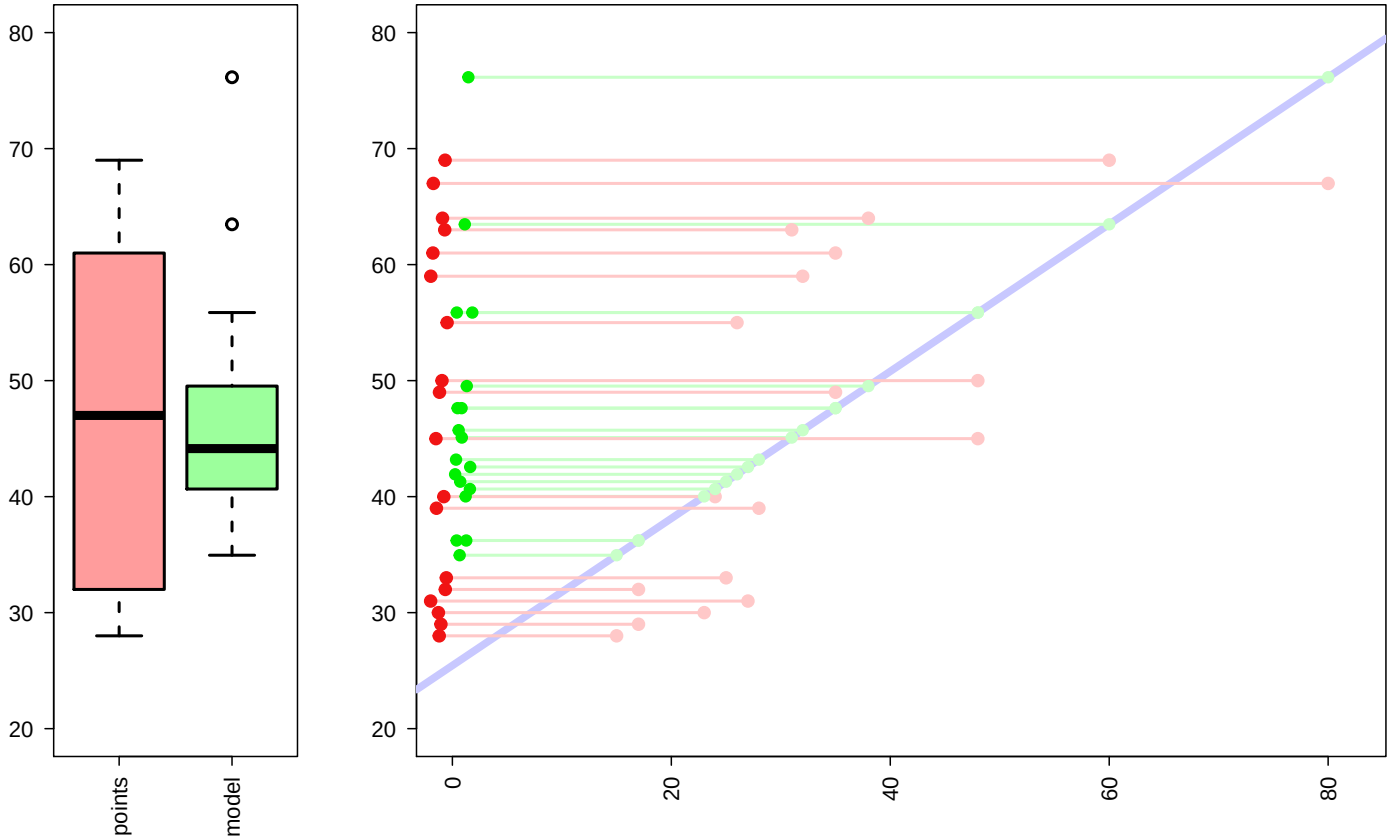
Varianz und Information

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



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

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

Determinationskoeffizient

- Gütemaß für die Regression: Determinationskoeffizient (Bestimmtheitskoeffizient):

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

- 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:

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

Regression: 4 eindimensionale Beispiele

- 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

(Quelle: Anscombe, (1973))

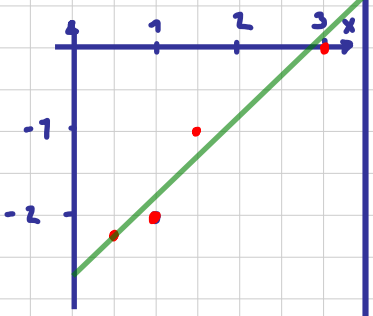
Lineare Regression:

① mit TR
② mit R
③ mit Excel

① ▸ Mode → STAT → A+Bx

▸ Dateneingabe:

x	y
1	-2
1.5	-1
0.5	-2.2
3	0



▸ Modellparameter: AC → Shift → STAT → REG

a, b, R^2 $\quad \hat{y} = -2.7 + 0.914x$

-2.7 $\quad$ 0.914 $\quad$ 0.95

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