Statistik

für Betriebswirtschaft und internationales Management

Sommersemester 2015

Prof. Dr. Stefan Etschberger
Hochschule Augsburg

Korrelationsrechnung

- Frage: Wie stark ist der Zusammenhang zwischen X und Y?
- Dazu: Korrelationskoeffizienten
- Verschiedene Varianten: Wahl abhängig vom Skalenniveau von X und Y:

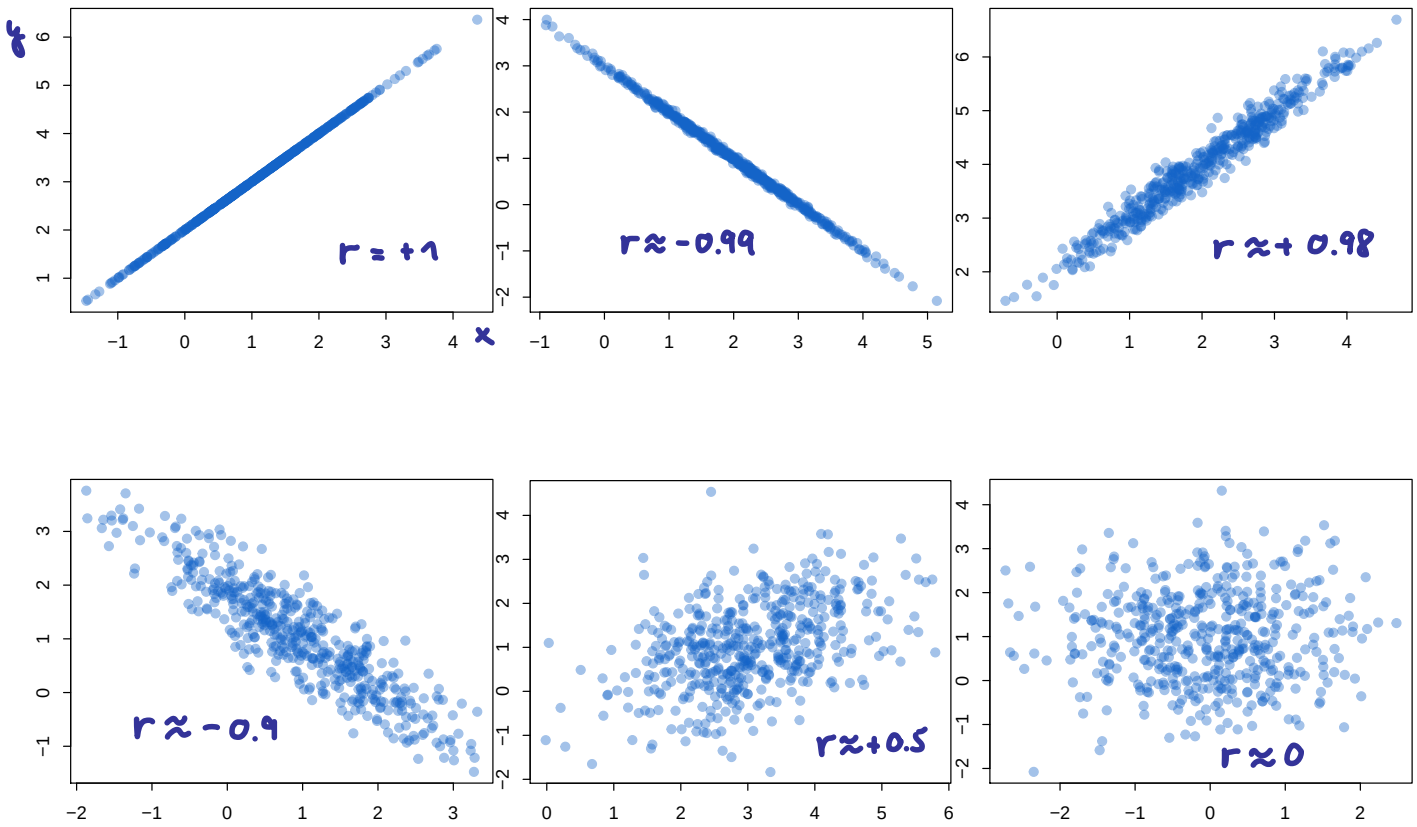
Skalierung von X	Skalierung von Y: kardinal (Zahlen, metrisch)	ordinal	nominal
kardinal	Bravais-Pearson-Korrelationskoeffizient	Rangkorrelationskoeffizient von Spearman	Kontingenzkoeffizient
ordinal			
nominal			

Korrelationskoeffizient von Bravais und Pearson

Bravais-Pearson-Korrelationskoeffizient

Voraussetzung: X, Y kardinalskaliert

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} = \frac{\sum_{i=1}^{n}x_iy_i - n\bar{x}\bar{y}}{\sqrt{\sum_{i=1}^{n}x_i^2 - n\bar{x}^2}\sqrt{\sum_{i=1}^{n}y_i^2 - n\bar{y}^2}} \in [-1;+1]$$



Korrelationskoeffizient von Bravais und Pearson

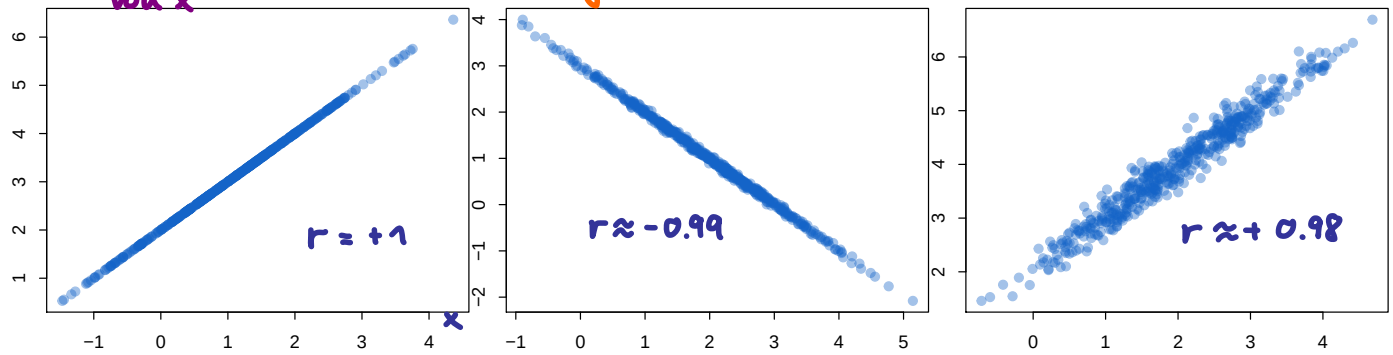
Bravais-Pearson-Korrelationskoeffizient

Voraussetzung: X, Y kardinalskaliert

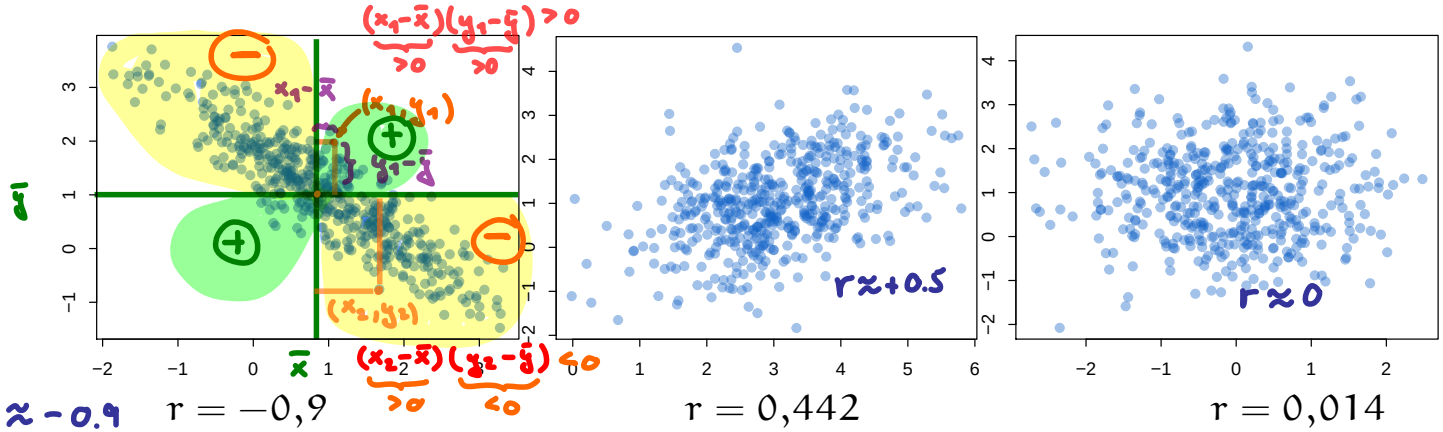
Kovarianz von x und y

$$r = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2}} = \frac{\sum_{i=1}^{n}x_iy_i - n\bar{x}\bar{y}}{\sqrt{\sum_{i=1}^{n}x_i^2 - n\bar{x}^2}\sqrt{\sum_{i=1}^{n}y_i^2 - n\bar{y}^2}} \in [-1;+1]$$

Standardabw. von x — Std. abw. von y



$r = 1$ $\quad r = -0{,}999$ $\quad r = 0{,}979$



$r \approx -0.9$ $\quad r = -0{,}9$ $\quad r = 0{,}442$ $\quad r = 0{,}014$

Korrelation mit TR

Mode → STAT → A+Bx

x	y	FREQ
2	4	1
4	3	1
3	6	1
9	7	1
7	8	1

AC Shift → STAT → REG → r

0,7029

Rangkorrelation: nötig, sobald 2 ordinale Merkmale oder 1 ord. und 1 metr. Merkmal

Beispiel:

Rang Preis	Preis (kardinal)	Qualität (ordinal)		Rang Qual.
2	4,99	spitze	1	1
6	17,99	geht so	5 →	5,5
4	9,99	geht so	6 →	5,5
5	10,20	gut	2 →	3
3	5,79	gut	3 →	3
1	4,80	gut	4 →	3

Rangkorrelationskoeffizient von Spearman

≙ Bravais-Pearson-Korr. koeff. der Rangnummern

r_{sp} =

Bravais-Pearson-Korrelationskoeffizient

Im Beispiel:

i	x_i	y_i	x_i^2	y_i^2	$x_i y_i$
1	2	4	4	16	8
2	4	3	16	9	12
3	3	6	9	36	18
4	9	7	81	49	63
5	7	8	49	64	56
$\sum$	25	28	159	174	157

$$\Rightarrow \quad \bar{x} = 25/5 = 5 \qquad \bar{y} = 28/5 = 5{,}6$$

$$r = \frac{157 - 5 \cdot 5 \cdot 5{,}6}{\sqrt{159 - 5 \cdot 5^2}\sqrt{174 - 5 \cdot 5{,}6^2}} = 0{,}703$$

(deutliche positive Korrelation)

Rangkorrelationskoeffizient von Spearman

- Voraussetzungen: X, Y (mindestens) ordinalskaliert, Ränge eindeutig (keine Doppelbelegung von Rängen)
- Vorgehensweise:
  1. Rangnummern R_i (X) bzw. R'_i (Y) mit $R_i^{(\prime)} = 1$ bei größtem Wert usw.
  2. Berechne

$$r_{SP} = 1 - \frac{6 \sum_{i=1}^{n} (R_i - R'_i)^2}{(n-1)\, n\, (n+1)} \in [-1; +1]$$

- Hinweise:
  - $r_{SP} = +1$ wird erreicht bei $R_i = R'_i \qquad \forall\, i = 1, \ldots, n$
  - $r_{SP} = -1$ wird erreicht bei $R_i = n + 1 - R'_i \quad \forall\, i = 1, \ldots, n$

Im Beispiel:

x_i	R_i	y_i	R'_i
2	5	4	4
4	3	3	5
3	4	6	3
9	1	7	2
7	2	8	1

$$r_{SP} = 1 - \frac{6 \cdot [(5-4)^2 + (3-5)^2 + (4-3)^2 + (1-2)^2 + (2-1)^2]}{(5-1) \cdot 5 \cdot (5+1)} = 0{,}6$$

- 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! $(h_{ij} \mapsto 2 \cdot h_{ij} \Rightarrow \chi^2 \mapsto 2 \cdot \chi^2)$

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.

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

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

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

Graphische Repräsentation von Kontingenztabellen

(chi-Quadrat)

$$\chi^2 = \frac{(264-239{,}4)^2}{239{,}4} + \frac{(90-111{,}6)^2}{111{,}6} + \frac{(6-9)^2}{9} + \frac{(2-26{,}6)^2}{26{,}6} + \frac{(34-12{,}4)^2}{12{,}4} + \frac{(4-1)^2}{1}$$

χ^2 ist nicht nach oben begrenzt

$\Rightarrow$ Kontingenzkoeffizient

$$k = \sqrt{\frac{\chi^2}{n+\chi^2}}$$

$$0 \leq k \leq k_{max}$$

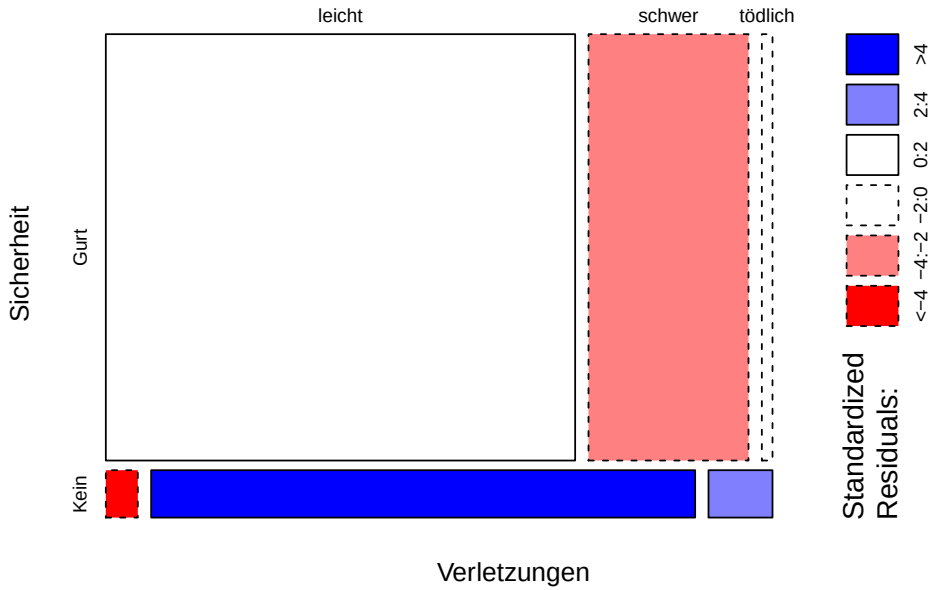
$$k_{max} = \sqrt{\frac{m-1}{m}}$$

$m = \min\{\text{Spaltenzahl}^3; \text{Zeilenzahl}^2\}$

(hier: $m=2 \Rightarrow k_{max} = \sqrt{\frac{2-1}{2}} = \sqrt{\frac{1}{2}} \approx 0{,}70$

Beispiel Autounfälle

	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



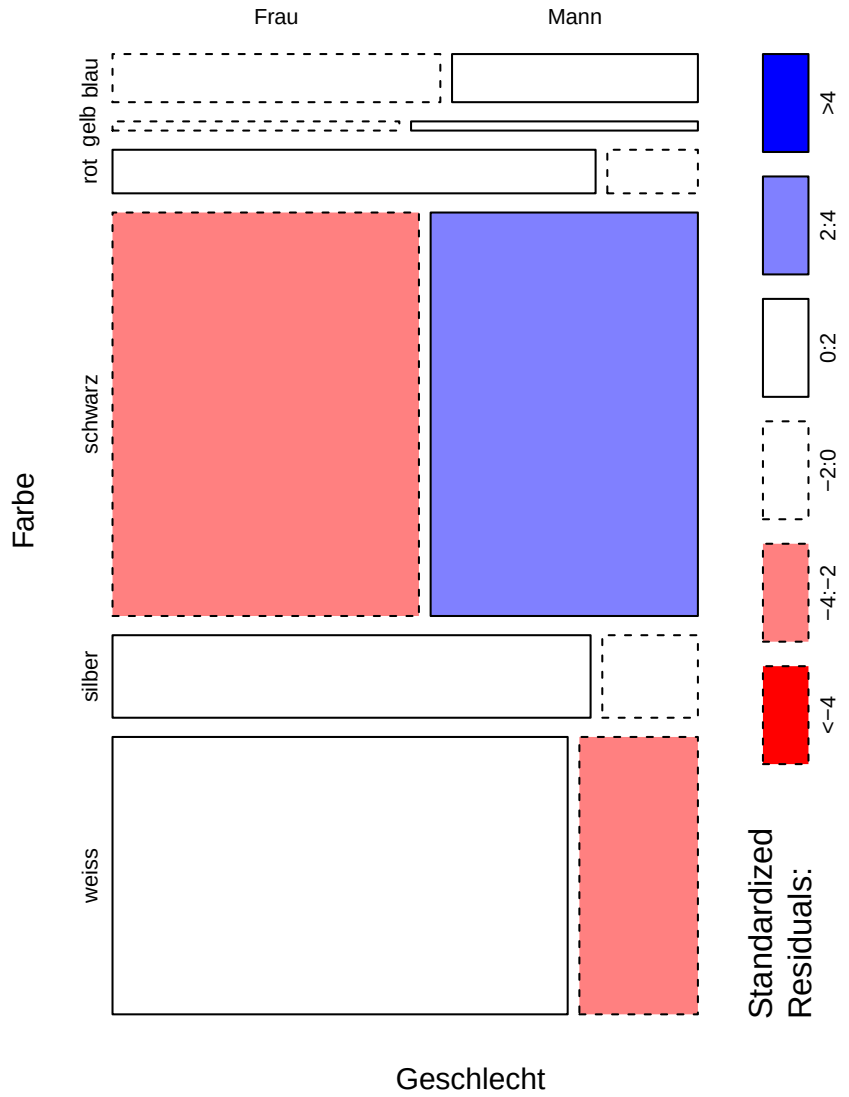
Mosaikplot Autounfälle

Mosaicplot Geschlecht, Wunschfarbe für Smartphone

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

```
##          Geschlecht
## Farbe     Frau Mann
##   blau      12    9
##   gelb       2    2
##   rot       16    3
##   schwarz   94   82
##   silber    30    6
##   weiss     96   25
```

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

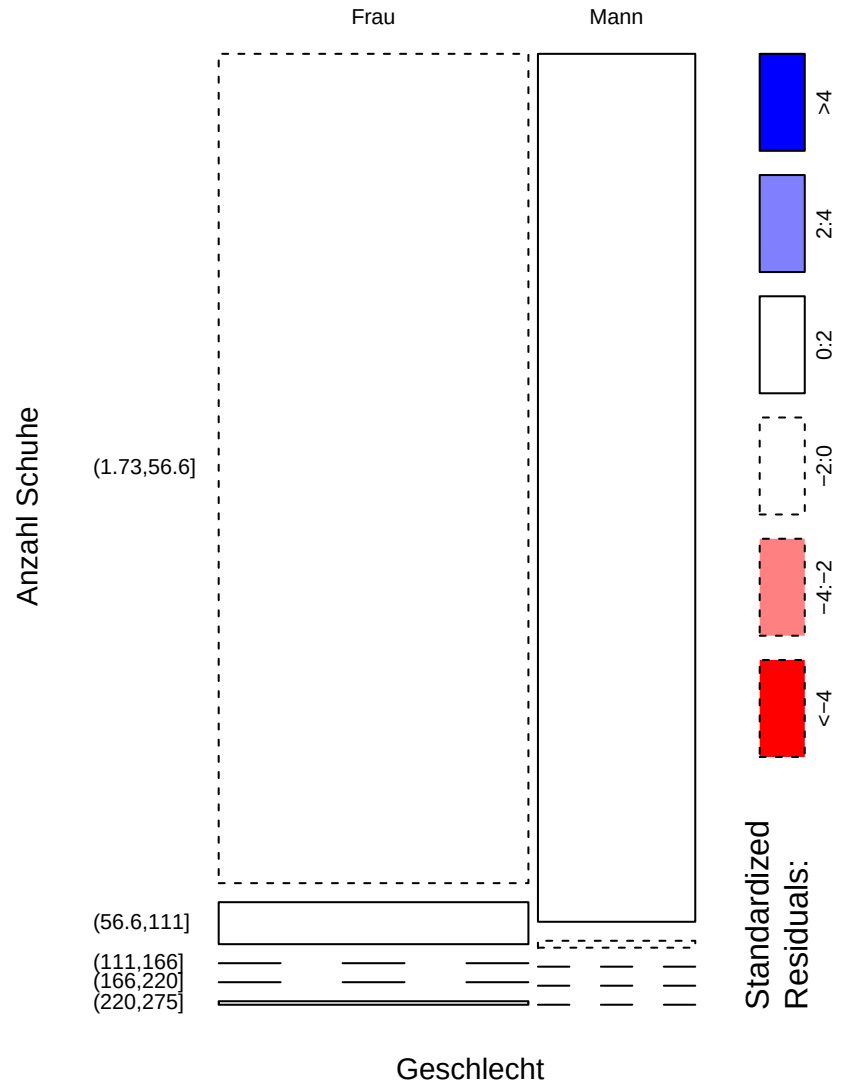


Mosaicplot Geschlecht, Anzahl Schuhe


```
tab = table(
  "Anzahl Schuhe"=cut(AnzSchuhe, 5),
  Geschlecht)
tab
```

```
##                Geschlecht
## Anzahl Schuhe Frau Mann
##   (1.73,56.6]  237  126
##   (56.6,111]    12    1
##   (111,166]      0    0
##   (166,220]      0    0
##   (220,275]      1    0
```

```
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:

$p_0(i)$: Preis des i-ten Gutes in Basisperiode 0
$p_t(i)$: Preis des i-ten Gutes in Berichtsperiode t
$q_0(i)$: 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:

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

↳ Mengen „damals"

- Preisindex von Paasche:

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

↳ Mengen „heute"

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

$$P^{L}_{1950,2013} = \frac{1{,}10 \cdot 3{,}58 + 0{,}70 \cdot 0{,}25}{0{,}04 \cdot 3{,}58 + 3{,}00 \cdot 0{,}25} \approx 4{,}6048 \;\hat{=}\; 360\,\%$$

pro Jahr ∅ : $\sqrt[63]{4{,}6048} \approx 1{,}0245 \;\hat{=}\; 2{,}45\,\%$

$$P^{P}_{1950,2013} = \frac{1{,}10 \cdot 1{,}25 + 0{,}70 \cdot 1{,}31}{0{,}04 \cdot 1{,}25 + 3{,}00 \cdot 1{,}31} \approx 0{,}5759 \;\hat{=}\; -42{,}4\,\%$$

pro Jahr ∅ : $\sqrt[63]{0{,}5759} \approx 0{,}9912 \;\hat{=}\; -0{,}88\,\%$

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 $q(i) \hat{=}$ 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

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)

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

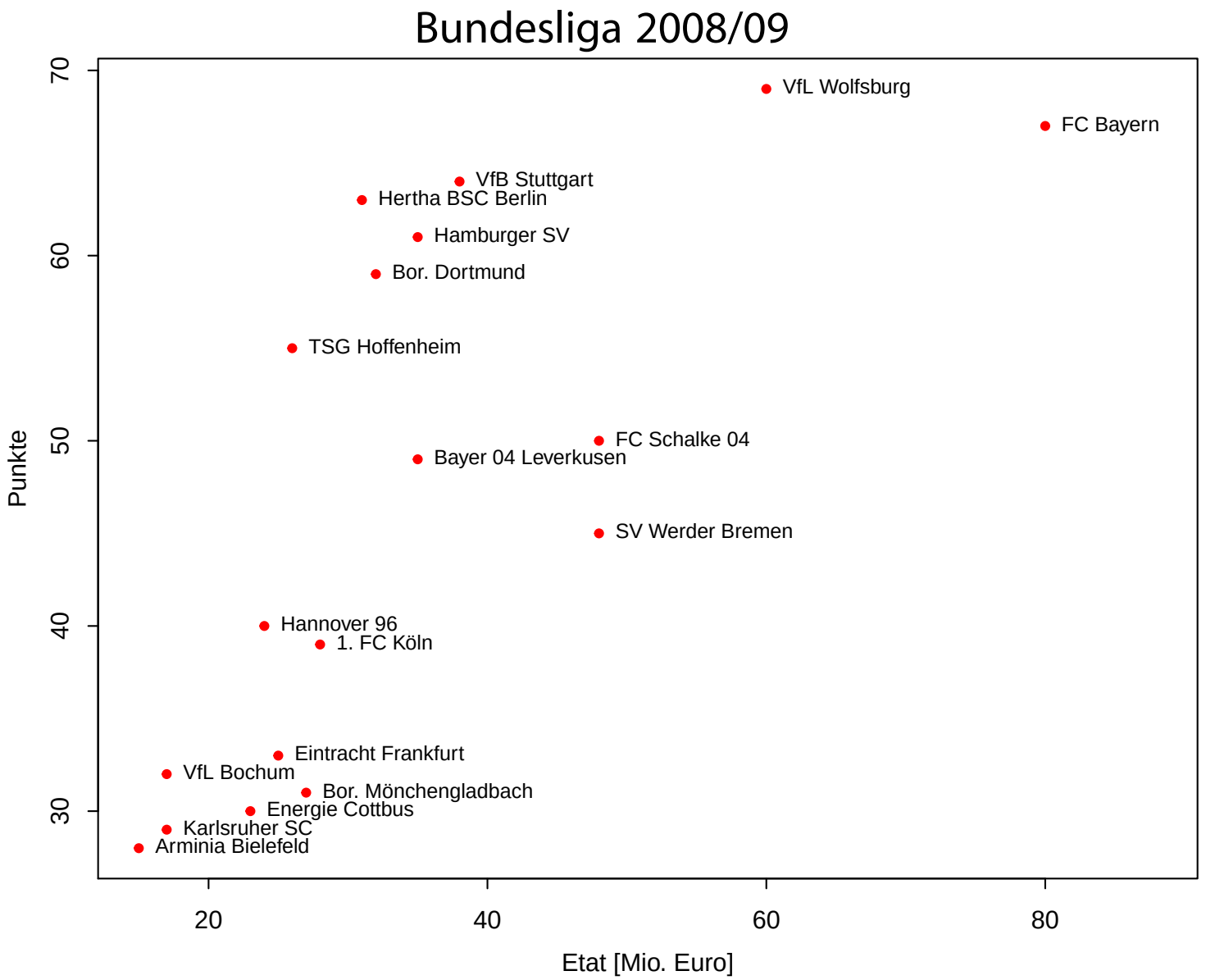
Bundesliga 2008/2009

- 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

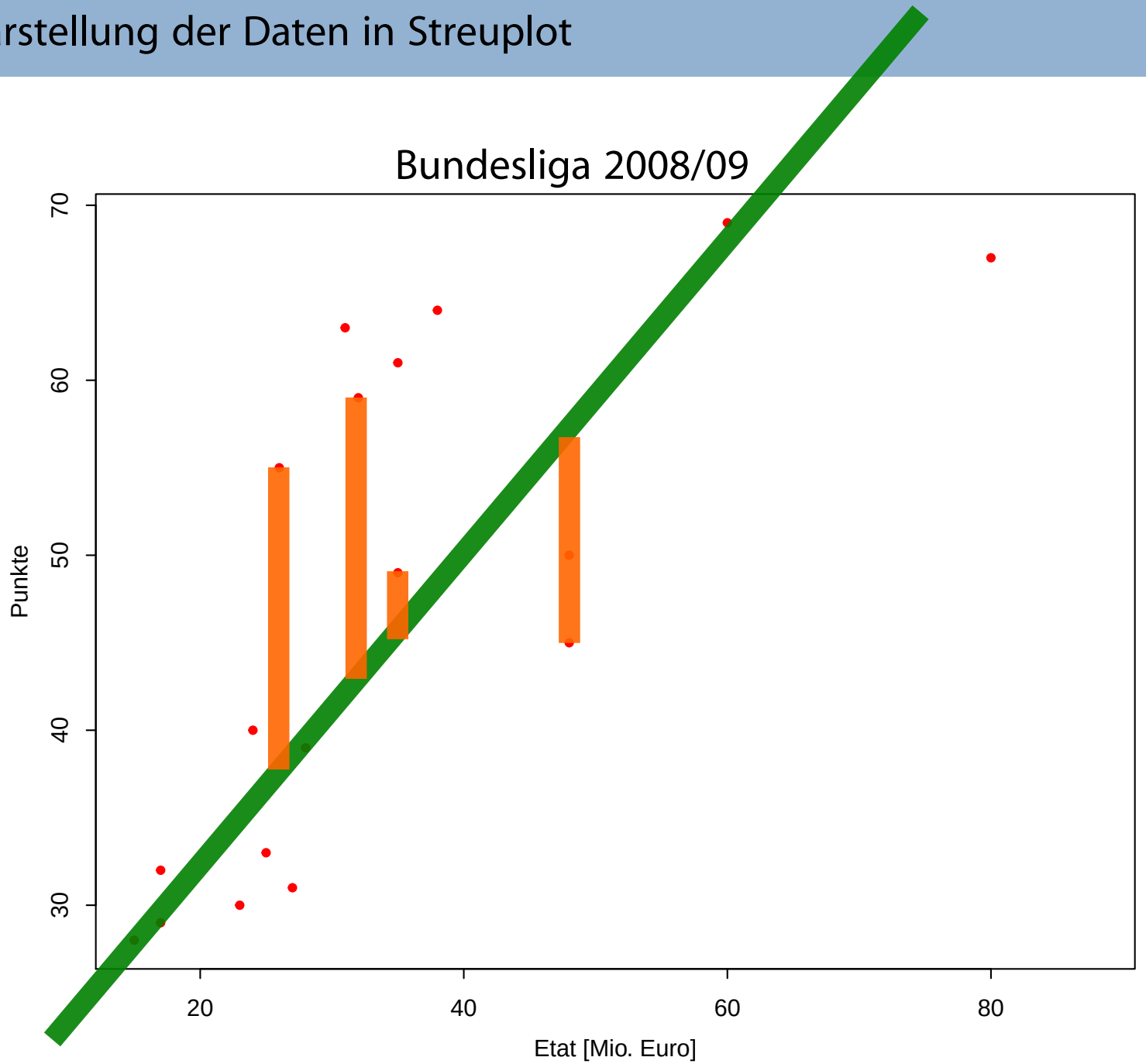
	Etat	Punkte
FC Bayern	80	67
VfL Wolfsburg	60	69
SV Werder Bremen	48	45
FC Schalke 04	48	50
VfB Stuttgart	38	64
Hamburger SV	35	61
Bayer 04 Leverkusen	35	49
Bor. Dortmund	32	59
Hertha BSC Berlin	31	63
1. FC Köln	28	39
Bor. Mönchengladbach	27	31
TSG Hoffenheim	26	55
Eintracht Frankfurt	25	33
Hannover 96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

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

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

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

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

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

$f(x,y)$
∇f , H_f

$$\begin{pmatrix} \frac{\partial Q}{\partial a} \\ \frac{\partial Q}{\partial b} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

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 $\widehat{=} y$ und Etat $\widehat{=} x$:

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

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

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

$$\approx 0{,}634$$

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

$$\approx 25{,}443$$

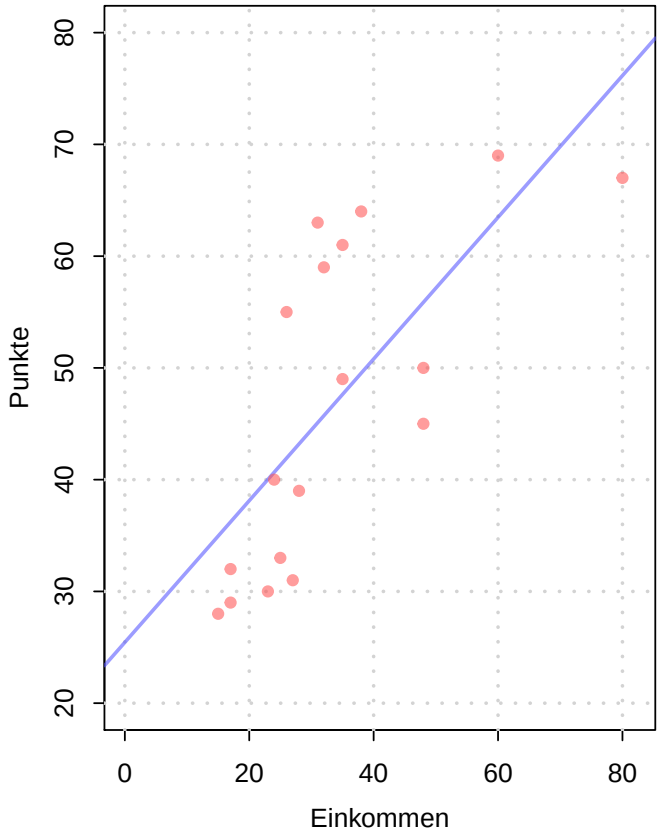
Bundesligabeispiel

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



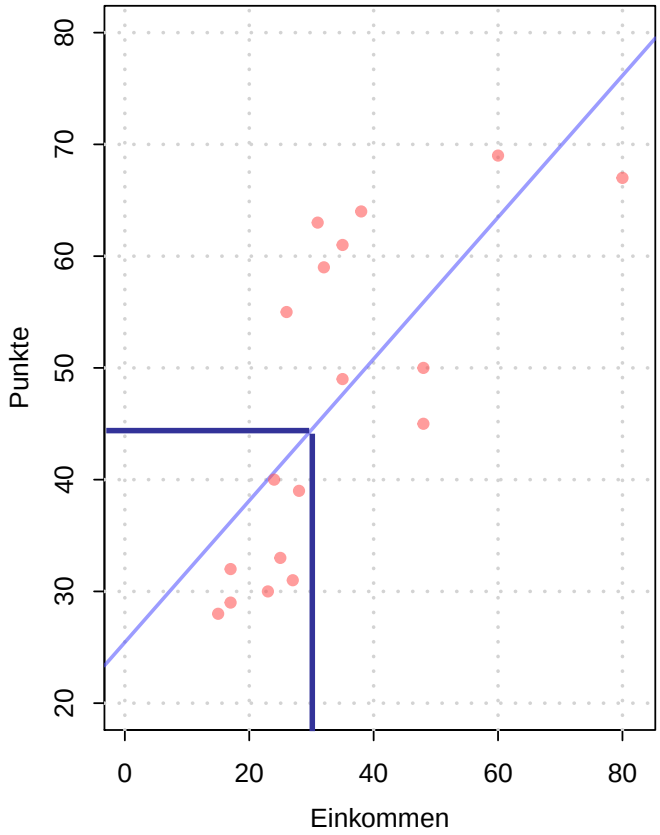
Bundesligabeispiel

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

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

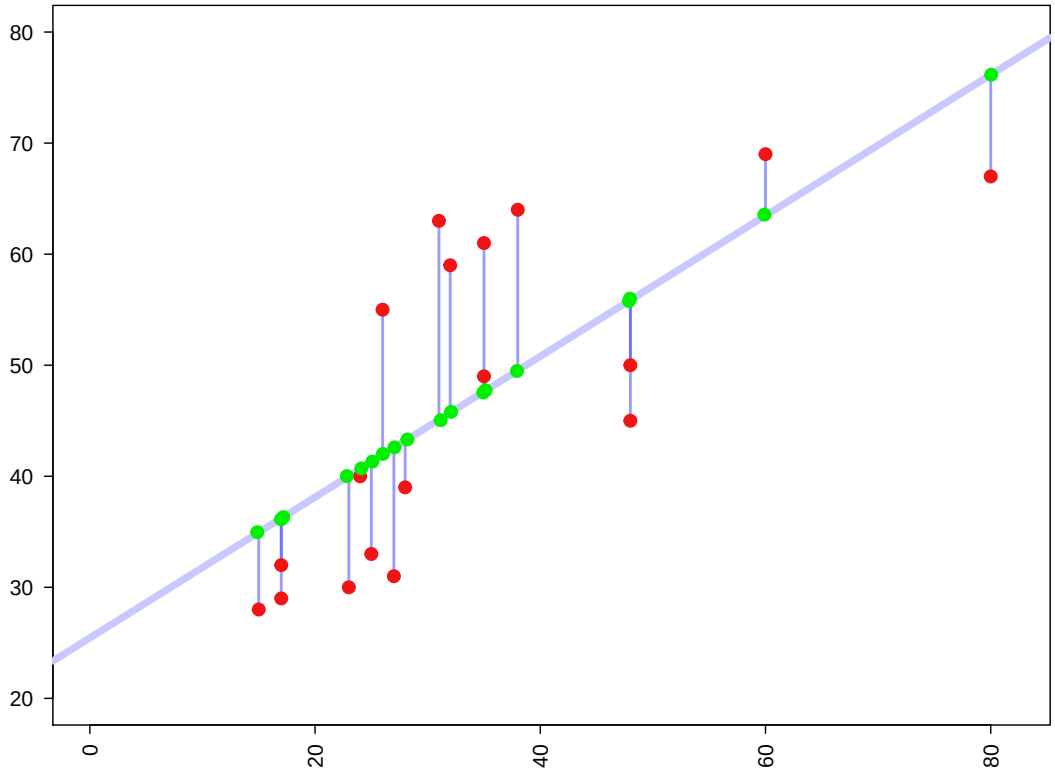


- Prognosewert für Etat $= 30$:

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

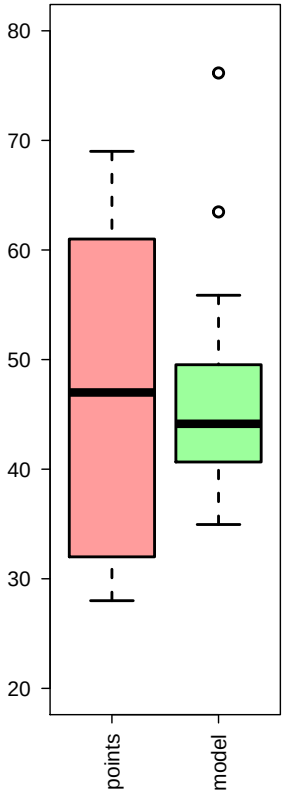
Varianz und Information

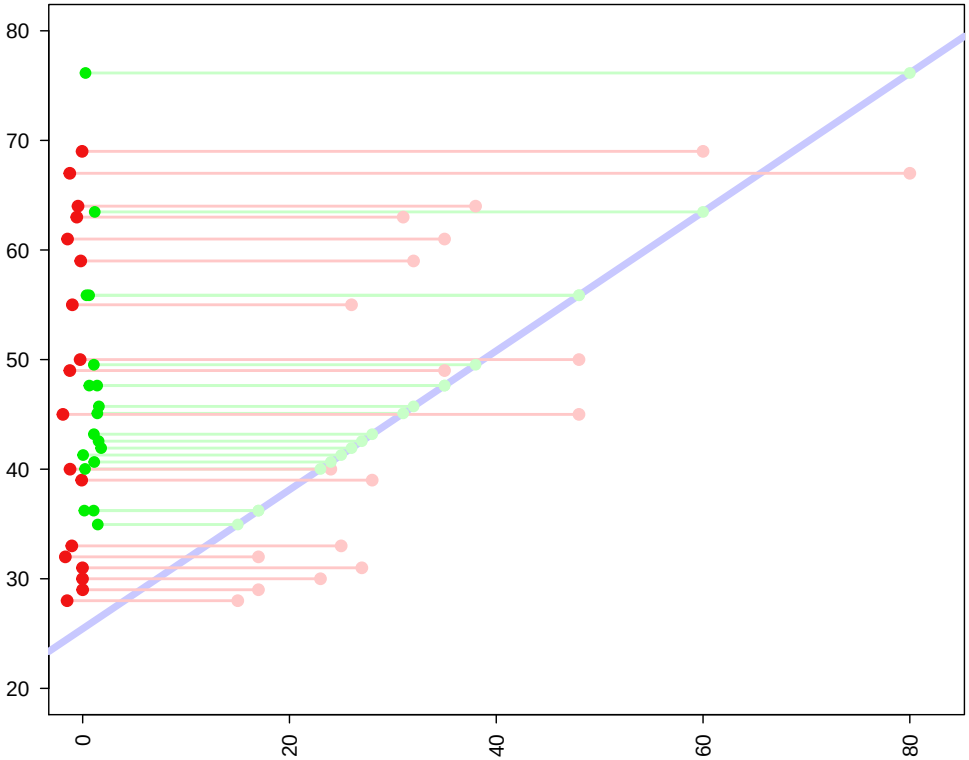
- **Varianz** der Daten in abhängiger Variablen y_i als Repräsentant des **Informations**gehalts
- Ein Bruchteil davon kann in Modellwerten $\hat{y}_i$ abgebildet werden



Varianz und Information

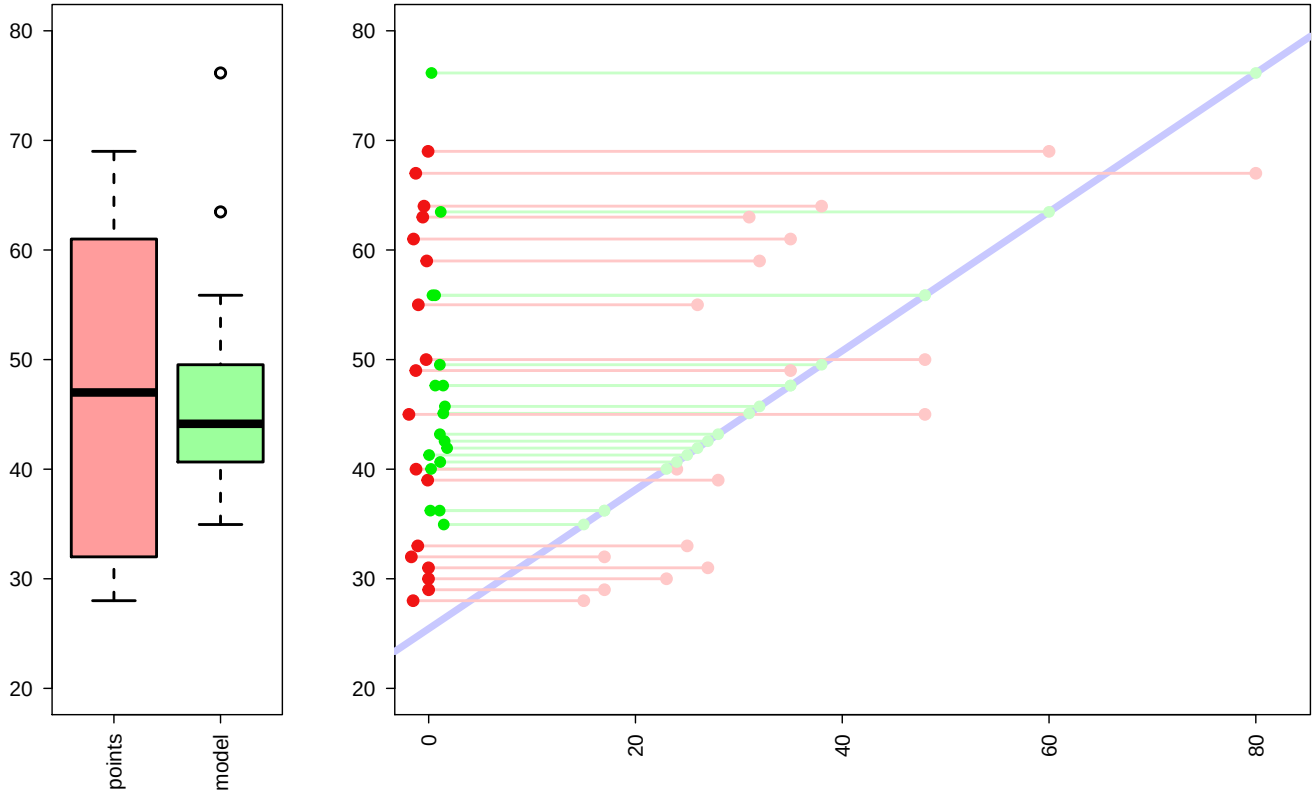
- **Varianz** der Daten in abhängiger Variablen y_i als Repräsentant des **Information**sgehalts
- Ein Bruchteil davon kann in Modellwerten $\hat{y}_i$ abgebildet werden





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

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

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:

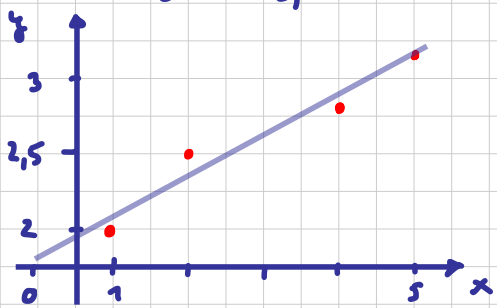
$$R^2 = \frac{\sum_{i=1}^{18} (\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{18} (y_i - \overline{y})^2} \approx \frac{102{,}78}{200{,}77} \approx 51{,}19\,\%$$

Beispiel (Regression)

① TR: Mode → STAT → A+Bx

②

x	y
1	2
2	2,5
4	2,8
5	3,2



③ AC ; shift → STAT → REG

↳ a 1,82

↳ b 0,27

$\hat{y} = 1{,}82 + 0{,}27x$

Determinationskoeffizient

$R^2 = (r^2) \approx 0{,}950$