Hausaufgabe: 21, 22, 24, 27-33

Statistik

für Betriebswirtschaft, Internationales Management, Wirtschaftsinformatik und Informatik

Sommersemester 2016

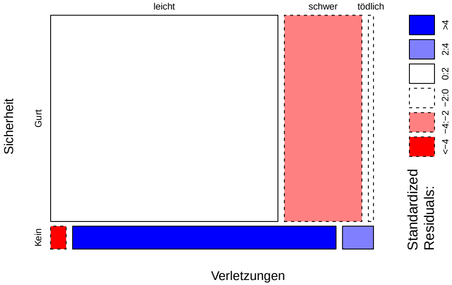
Veranstaltungen zur Statistik für BW/IM Sommersemester 2016				
Was?	Wer?	Wann?	Wo?	ab wann?
Vorlesung Statistik	Etschberger	Mi, 14.00-17.00	B2.14	16.03.2016
Vorlesung Statistik PLUS	Etschberger/Jansen	- Blocktermin -	?	?
Übung Statistik	Etschberger	Mi, 17.00-18.30	A1.10	30.03.2016
Übung Statistik	Jansen	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	?	?	?

1	HSA Statistik SS 2016 Sessionlist		
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

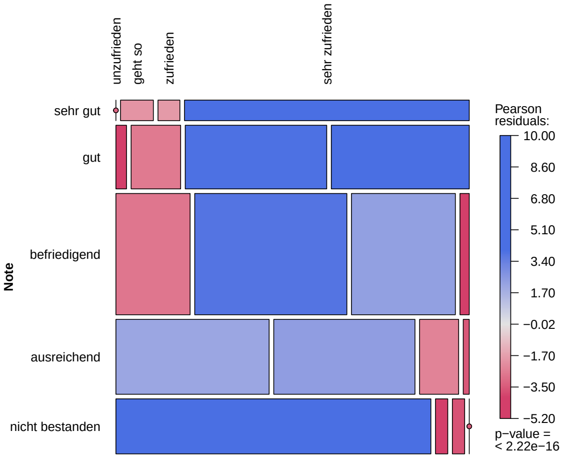
Beispiel Autounfälle

	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))
```



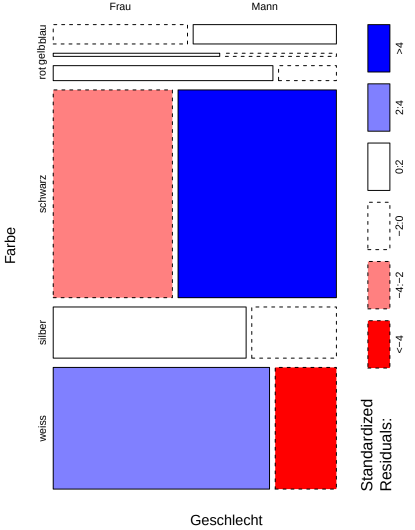
„Note in Matheklausur“ gegen „Zufrieden mit Leistung“

Mosaicplot Geschlecht, Wunschfarbe für Smartphone

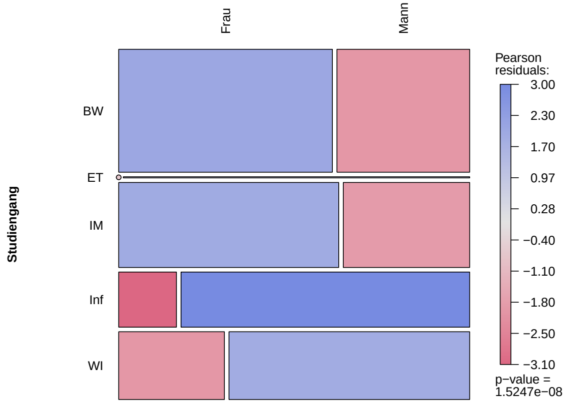
```r
tab = table(Farbe, Geschlecht)
tab

##          Geschlecht
## Farbe     Frau Mann
##   blau      15   16
##   gelb       3    2
##   rot       19    5
##   schwarz  143  190
##   silber    57   25
##   weiss    152   43
```

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



```r
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“~~

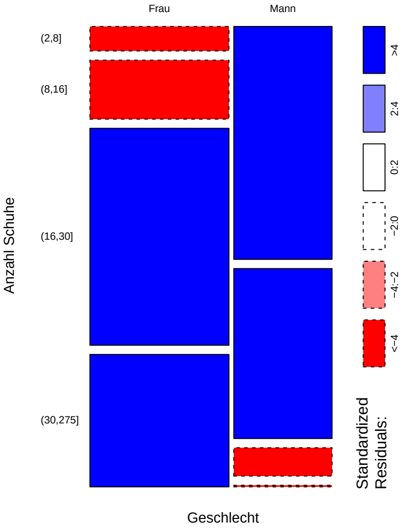
Studiengang vs. Geschlecht

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
##      (2,8]      22  148
##      (8,16]     53  108
##      (16,30]   195   18
##      (30,275]  119    1
```

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



Preisindizes

- 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:

$$\begin{aligned}
p_0(i):&\quad \text{Preis des } i\text{-ten Gutes in Basisperiode } 0\\
p_t(i):&\quad \text{Preis des } i\text{-ten Gutes in Berichtsperiode } t\\
q_0(i):&\quad \text{Menge des } i\text{-ten Gutes in Basisperiode } 0\\
q_t(i):&\quad \text{Menge des } i\text{-ten Gutes in Berichtsperiode } t
\end{aligned}$$

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:

Warenkorb von „damals"

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

- Preisindex von Paasche:

Warenkorb von „heute"

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

$100\,€ \cdot 1{,}03 \cdot 1{,}05 \cdot 1{,}10 = 118{,}97$

$100\,€ \cdot q \cdot q \cdot q = 118.97$

$q^3 = 1{,}1897$

$q = \sqrt[3]{1{,}1897} \approx 1{,}059$

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. 360,48 % Preissteigerung von 1950 bis 2013

durchschnittlich:

$\sqrt[63]{4{,}6048} \approx 1{,}025 \mathrel{\hat{=}} 2{,}5\,\%$ Infl. 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. $-42{,}41\,\%$ Preisentw. von 1950-2013

durchschnittlich:

$\sqrt[63]{0.5759} \approx 0.9913 \mathrel{\hat{=}} -0.87\,\%$ Infl. pro Jahr

Weitere Preisindizes

Idealindex von Fisher:

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

Marshall-Edgeworth-Index:

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

Preisindex von Lowe:

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

wobei $q(i) \triangleq \begin{cases} \text{Durchschn. Menge von} \\ \text{Gut } i \text{ über alle (bekannten)} \\ \text{Perioden} \end{cases}$

Weitere Preisindizes: Beispiel

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.

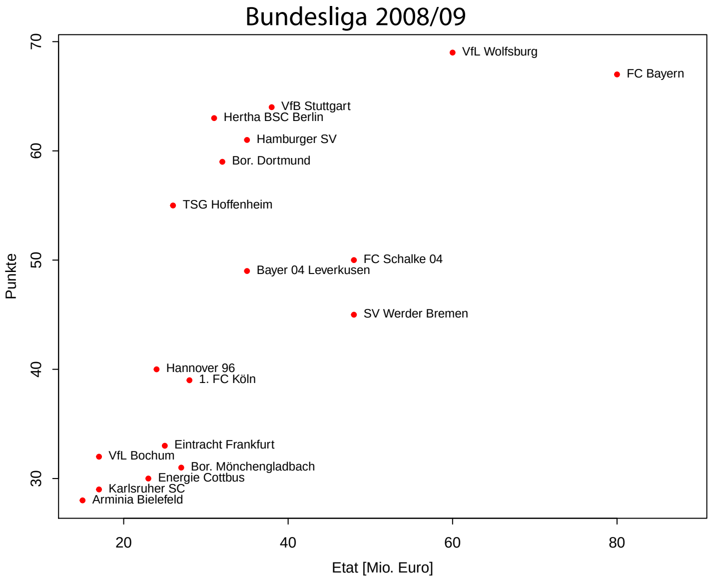
Bundesliga 2008/2009

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

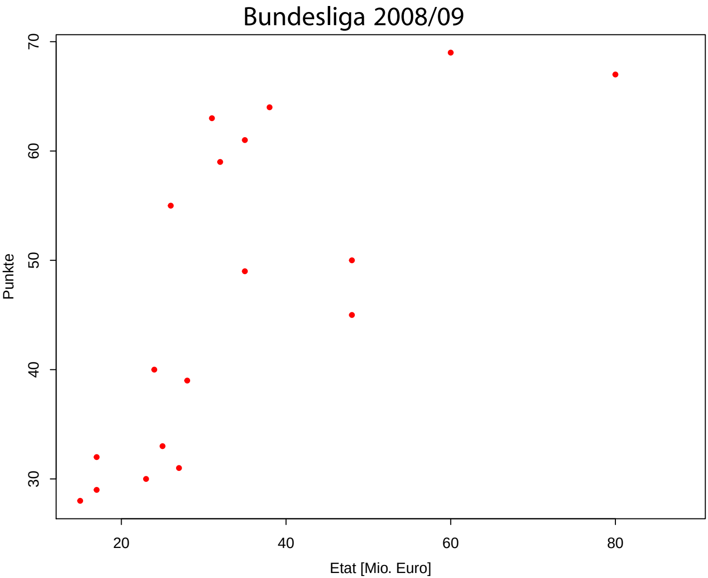
	Etat	Punkte
FC Bayern	80	67
VfL Wolfsburg	60	69
SV Werder Bremen	48	45
FC Schalke 04	48	50
VfB Stuttgart	38	64
Hamburger SV	35	61
Bayer 04 Leverkusen	35	49
Bor. Dortmund	32	59
Hertha BSC Berlin	31	63
1. FC Köln	28	39
Bor. Mönchengladbach	27	31
TSG Hoffenheim	26	55
Eintracht Frankfurt	25	33
Hannover 96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

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

notwendig für Minimum:

$$\nabla Q = 0$$

$$\frac{\partial Q}{\partial a} = \sum 2[y_i - (a + bx_i)] \cdot (-1) = 0$$

$$\frac{\partial Q}{\partial b} = \sum 2[y_i - (a + bx_i)] \cdot (-x_i) = 0$$

Beste Lösung

- Beste und eindeutige Lösung:

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

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

- Regressionsgerade:

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

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

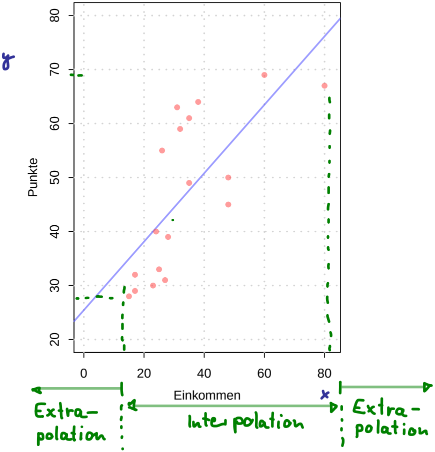
$$\Rightarrow \hat{b} = \frac{31474 - 18 \cdot 33{,}83 \cdot 46{,}89}{25209 - 18 \cdot 33{,}83^2}$$

$$\approx 0{,}634$$

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

$$\approx 25{,}443$$

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



Bundesligabeispiel

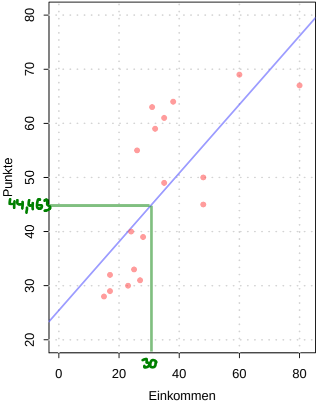
- Berechnung eines linearen Modells der Bundesligadaten
- dabei: Punkte $\hat{=}\, y$ und Etat $\hat{=}\, x$:

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

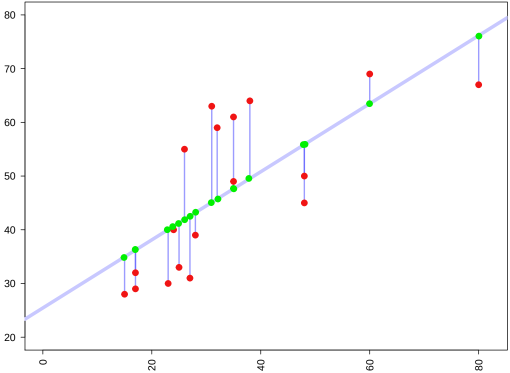


- Prognosewert für Etat $= 30$:

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

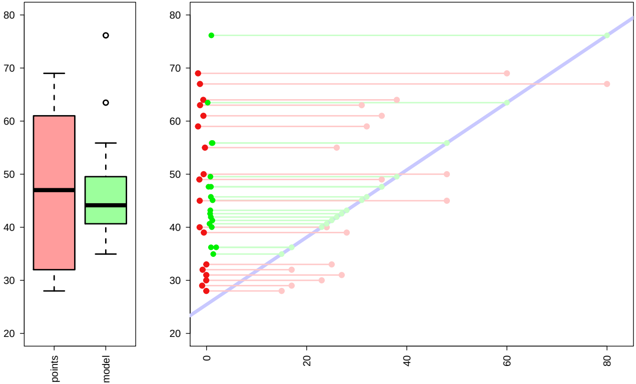
Varianz und Information

- **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



- 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):

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

→ Bravais-Pearson 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:

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

→ d.h. ca. 51% der in der abhängigen Variable vorhandenen Information kann über die Regression erklärt werden (die restlichen 49% sind andere Effekte)

Beispiel: Körpergröße in Abh. der Größe des Vaters

gesucht: Regressionsmodell: $G = a + b \cdot V$

unabh. Variable: x

abh. Variable: y

TR: ▶ Mode → STAT → A+BX

▶ Daten eingeben

▶ AC

▶ Shift → STAT → REG

a: 53,968 b: 0.709

	G y	V x	
	Groesse	GroesseV	Geschlecht
451	173	172	Mann
620	189	193	Mann
233	190	180	Mann
315	178	180	Mann
548	184	183	Mann

$G = 53{,}968 + 0.709 \cdot V$

$R^2 = r^2 \approx 0.547$