

# Statistik

## 09. Lineare Modelle

Roland Schäfer

Institut für Germanistische Sprachwissenschaft  
Friedrich-Schiller-Universität Jena

stets aktuelle Fassungen: <https://github.com/rsling/VL-Statistik>

- 1 Lineare Modelle
  - Korrelation und Signifikanz
  - Lineare Regression
  - Multiple Regression
  - ANOVA und LMs

■ In R

- 2 Lineare Modelle und ANOVA
- 3 Nächste Woche | Überblick

LMs

- Gravetter & Wallnau 2007
- Zuur u. a. 2009
- Maxwell & Delaney 2004

- Pearson-Korrleation ( $r, r^2$ )

- Pearson-Korrelation ( $r, r^2$ )
- Signifikanztests mit Korrelationen

- Pearson-Korrelation ( $r$ ,  $r^2$ )
- Signifikanztests mit Korrelationen
- Unterschied von Pearsons  $r$  zu Spearman's Rang-Korrelation

- Pearson-Korrelation ( $r$ ,  $r^2$ )
- Signifikanztests mit Korrelationen
- Unterschied von Pearsons  $r$  zu Spearman's Rang-Korrelation
- Unterschiede zwischen Korrelation und Regression

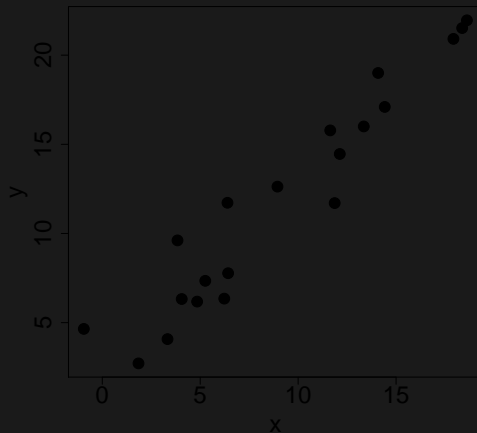


- Pearson-Korrelation ( $r$ ,  $r^2$ )
- Signifikanztests mit Korrelationen
- Unterschied von Pearsons  $r$  zu Spearmans Rang-Korrelation
- Unterschiede zwischen Korrelation und Regression
- Berechnung linearer Regressionsmodelle

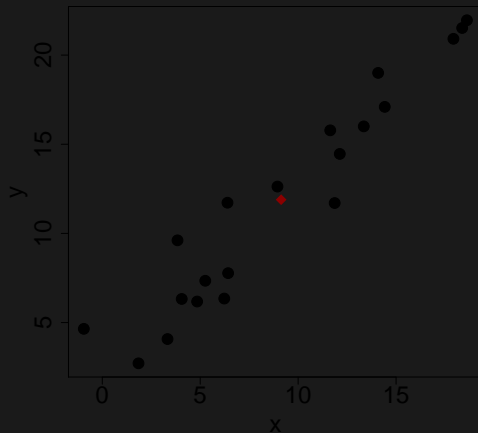
- Pearson-Korrelation ( $r$ ,  $r^2$ )
- Signifikanztests mit Korrelationen
- Unterschied von Pearsons  $r$  zu Spearmans Rang-Korrelation
- Unterschiede zwischen Korrelation und Regression
- Berechnung linearer Regressionsmodelle
- Signifikanztests für Modell und Koeffizienten

# Korrelationen | Zusammenhänge zwischen numerischen Variablen

## Bivariate Korrelationskoeffizienten | ab Ordinalskala

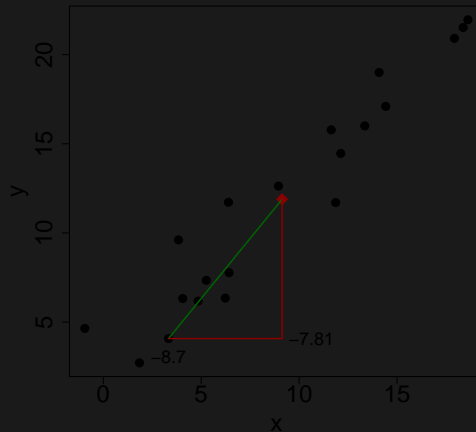


Koordinate von  $\langle \bar{x}, \bar{y} \rangle$  | Mittel der beiden gemessenen Variablen



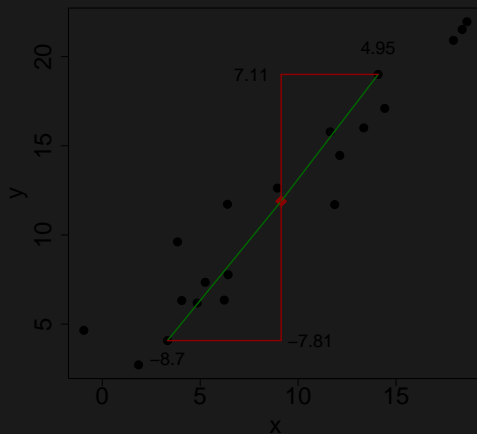
# Kovarianz | Illustration 2

Punktvarianzen |  $x_3 - \bar{x} = -7.81$  und  $y_3 - \bar{y} = -5.80$  |  $-7.81 \cdot -5.80 = 45.30$



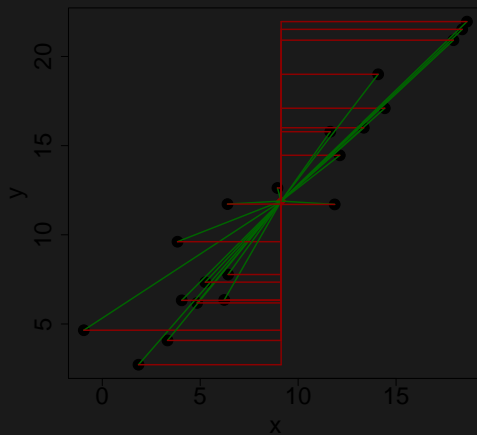
# Kovarianz | Illustration 3

Punktvarianzen |  $x_{17} - \bar{x} = 4.95$  und  $y_{17} - \bar{y} = 7.11$  |  $4.95 \cdot 7.11 = 35.19$

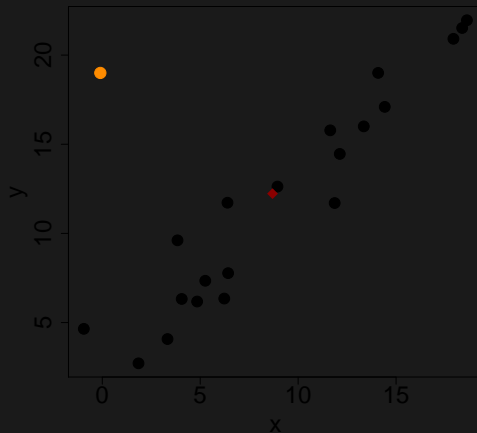


# Kovarianz | Illustration 4

Puntvarianzen für alle  $\langle x_i, y_i \rangle$   $\text{cov}(x, y) = 34.52$



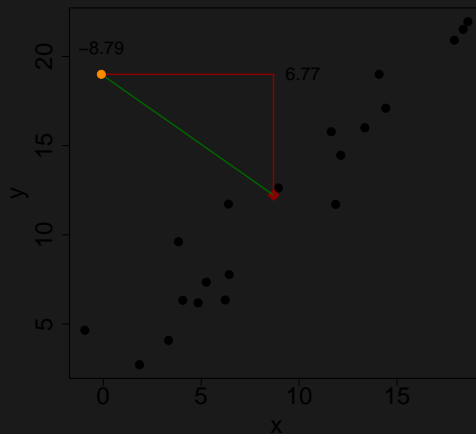
Ausreißer bei ansonsten positiver Kovarianz | Negatives Produkt der Punktvarianzen





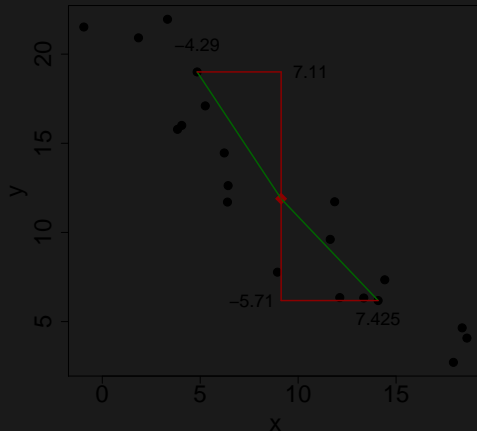
# Kovarianz | Illustration 6

Punktvarianzen |  $x_{21} - \bar{x} = 6.77$  und  $y_{21} - \bar{y} = -8.79$  |  $6.77 \cdot -8.79 = -59.51$



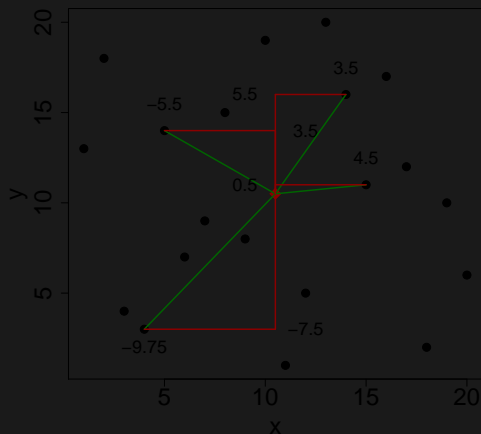
# Negative Kovarianz

Tendenziell negative Abhängigkeit | Punktvarianzen überwiegend |  $\text{cov}(x, y) = -33.77$



# Kovarianz nahe Null

Ohne Abhängigkeit | Kovarianz nahe 0 |  $\text{cov}(x, y) = -1.74$



Kovarianz | Kombination der Abweichung der Messpunkte vom jeweiligen Mittel

$$\text{cov}(x, y) = \frac{\sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})}{n-1}$$

Summe der Produkte | Der Zählerterm |  $SP(x, y) = \sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})$

- $x_i - \bar{x} > 0$  und  $y_i - \bar{y} > 0$  | Beitrag zur Kovarianz **positiv**

Kovarianz | Kombination der Abweichung der Messpunkte vom jeweiligen Mittel

$$\text{cov}(x, y) = \frac{\sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})}{n-1}$$

Summe der Produkte | Der Zählerterm |  $SP(x, y) = \sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})$

- $x_i - \bar{x} > 0$  und  $y_i - \bar{y} > 0$  | Beitrag zur Kovarianz **positiv**
- $x_i - \bar{x} < 0$  und  $y_i - \bar{y} < 0$  | Beitrag zur Kovarianz **positiv**

Kovarianz | Kombination der Abweichung der Messpunkte vom jeweiligen Mittel

$$\text{cov}(x, y) = \frac{\sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})}{n-1}$$

Summe der Produkte | Der Zählerterm |  $SP(x, y) = \sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})$

- $x_i - \bar{x} > 0$  und  $y_i - \bar{y} > 0$  | Beitrag zur Kovarianz **positiv**
- $x_i - \bar{x} < 0$  und  $y_i - \bar{y} < 0$  | Beitrag zur Kovarianz **positiv**
- $x_i - \bar{x} > 0$  und  $y_i - \bar{y} < 0$  | Beitrag zur Kovarianz **negativ**

Kovarianz | Kombination der Abweichung der Messpunkte vom jeweiligen Mittel

$$\text{cov}(x, y) = \frac{\sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})}{n-1}$$

Summe der Produkte | Der Zählerterm |  $SP(x, y) = \sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})$

- $x_i - \bar{x} > 0$  und  $y_i - \bar{y} > 0$  | Beitrag zur Kovarianz **positiv**
- $x_i - \bar{x} < 0$  und  $y_i - \bar{y} < 0$  | Beitrag zur Kovarianz **positiv**
- $x_i - \bar{x} > 0$  und  $y_i - \bar{y} < 0$  | Beitrag zur Kovarianz **negativ**
- $x_i - \bar{x} < 0$  und  $y_i - \bar{y} > 0$  | Beitrag zur Kovarianz **negativ**

Korrelationskoeffizient | Im Gegensatz zur Kovarianz skalenunabhängig

$$r(x, y) = \frac{\text{cov}(x, y)}{s(x) \cdot s(y)}$$

Pearson-Korrelation



- Maß der Varianzerklärung durch  $r$ :  $r^2$  (vgl. t-Test)

- Maß der Varianzerklärung durch  $r$ :  $r^2$  (vgl. t-Test)
- Signifikanztest möglich: Schluss auf Korrelation in der Grundgesamtheit

- Maß der Varianzerklärung durch  $r$ :  $r^2$  (vgl. t-Test)
- Signifikanztest möglich: Schluss auf Korrelation in der Grundgesamtheit
- $df_r = n - 2$

- Maß der Varianzerklärung durch  $r$ :  $r^2$  (vgl. t-Test)
- Signifikanztest möglich: Schluss auf Korrelation in der Grundgesamtheit
- $df_r = n - 2$
- Unter der NULL (keine Korrelation) t-verteilt:

$$t = r \sqrt{\frac{n-2}{1-r^2}}$$

- Maß der Varianzerklärung durch  $r$ :  $r^2$  (vgl. t-Test)
- Signifikanztest möglich: Schluss auf Korrelation in der Grundgesamtheit
- $df_r = n - 2$
- Unter der NULL (keine Korrelation) t-verteilt:  
$$t = r \sqrt{\frac{n-2}{1-r^2}}$$
- ...oder Tabellen (z. B. G&W, B.6)

- Intervallskalierung

- Intervallskalierung
- lineare Abhängigkeit

- Intervallskalierung
- lineare Abhängigkeit
- bei kleinen  $n$ : Normalverteilung für  $x$  und  $y$



- Intervallskalierung
  - lineare Abhängigkeit
  - bei kleinen  $n$ : Normalverteilung für  $x$  und  $y$
- 
- wenn nicht: Spearmans Rang-Korrelation

- mathematisch nicht andere als eine Pearson-Korrleation

- mathematisch nicht andere als eine Pearson-Korrleation
- vorher: Umrechnung der rohen x,y-Werte in Ränge

- mathematisch nicht andere als eine Pearson-Korrleation
- vorher: Umrechnung der rohen  $x, y$ -Werte in Ränge
- bei gleichen Werten: alle gleichen Werte bekommen Rang-Mittel

# Werte in Ränge umrechnen

Ein Beispiel zur Umwandlung in Ränge:

Index:	1	2	3	4	5
Messwerte x:	4	7	3	1	3
Messwerte y:	9	12	11	2	8

Statt der Messwerte arbeitet man mit den Rängen der Messwerte an den jeweiligen Indexen.

Index:	1	2	3	4	5
Ränge der Messwerte x:	4	5	2.5	1	2.5
Ränge der Messwerte y:	3	5	4	1	2

Wenn  $\text{Rang}(x_i)$  der Rang für  $x_i$  in  $x$  ist:

Spearman's Rang-Korrelation:

$$r_S = 1 - \frac{6 \sum_{i=1}^n (\text{Rang}(x_i) - \text{Rang}(y_i))^2}{n(n^2 - 1)}$$

- Korrelation: Stärke des Zusammenhangs

# Unterschiede zwischen Korrelation und Regression

- Korrelation: Stärke des Zusammenhangs
- Regression: genaue Funktion zur Modellierung des Zusammenhangs



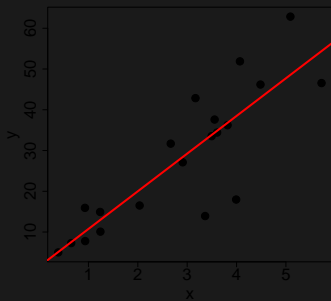
# Unterschiede zwischen Korrelation und Regression

- Korrelation: Stärke des Zusammenhangs
- Regression: genaue Funktion zur Modellierung des Zusammenhangs
- Korrelation: Diagnostik/Test

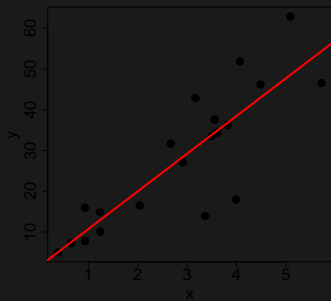
# Unterschiede zwischen Korrelation und Regression

- Korrelation: Stärke des Zusammenhangs
- Regression: genaue Funktion zur Modellierung des Zusammenhangs
- Korrelation: Diagnostik/Test
- Regression: Vorhersage (und Test)

# Spezifikation der Funktion für die Regressionsgerade

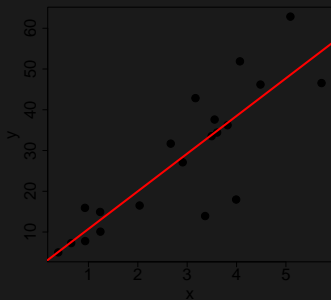


# Spezifikation der Funktion für die Regressionsgerade



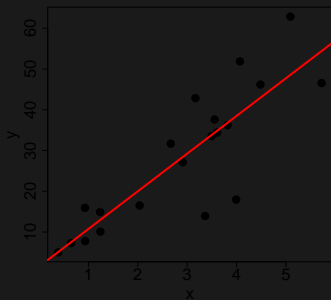
- Schnittpunkt mit der y-Achse (Intercept):  $a$

# Spezifikation der Funktion für die Regressionsgerade



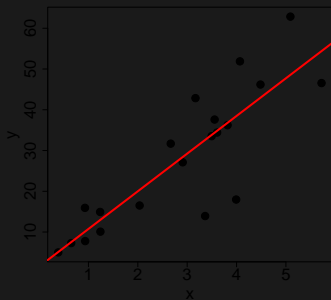
- Schnittpunkt mit der  $y$ -Achse (Intercept):  $a$
- Steigung (Slope):  $b$  ( $b$  heißt auch Koeffizient)

# Spezifikation der Funktion für die Regressionsgerade



- Schnittpunkt mit der y-Achse (Intercept):  $a$
- Steigung (Slope):  $b$  ( $b$  heißt auch Koeffizient)
- Regressionsgleichung (=Modell):  $\hat{y} = b \cdot x + a$

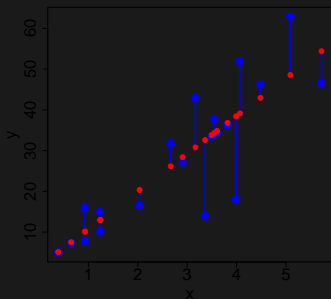
# Spezifikation der Funktion für die Regressionsgerade



- Schnittpunkt mit der y-Achse (Intercept):  $a$
- Steigung (Slope):  $b$  ( $b$  heißt auch Koeffizient)
- Regressionsgleichung (=Modell):  $\hat{y} = b \cdot x + a$
- Für jeden beobachteten Wert:  $y_i = b \cdot x_i + a + e_i$  ( $e_i$  als Fehlerterm)

# Idee der kleinsten Quadrate

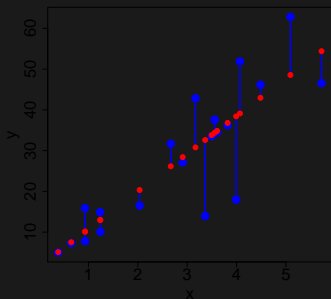
Die vom Modell vorhergesagten Werte (rot, auf der Regressionsgerade) sollen insgesamt einen so geringen Abstand wie möglich zu den Beobachtungen (blau) haben.





# Idee der kleinsten Quadrate

Die vom Modell vorhergesagten Werte (rot, auf der Regressionsgerade) sollen insgesamt einen so geringen Abstand wie möglich zu den Beobachtungen (blau) haben.



Die Summe der quadrierten negativen und positiven Differenzen (blau) soll minimiert werden (=kleinste Quadrate): Minimierung von  $E = \sum e^2$

- Slope/Steigung:  $b = \frac{\sum(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{SP(x,y)}{SQ(x)}$

# Berechnung der Regressionsgleichung

- Slope/Steigung:  $b = \frac{\sum(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{SP(x,y)}{SQ(x)}$
- Intercept:  $a = \bar{y} - b \cdot \bar{x}$

# Berechnung der Regressionsgleichung

- Slope/Steigung:  $b = \frac{\sum(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{SP(x,y)}{SQ(x)}$
- Intercept:  $a = \bar{y} - b \cdot \bar{x}$
- Der Beweis, dass dies die Gerade mit den kleinsten Quadraten schätzt, erfordert bereits erheblichen mathematischen Aufwand, den wir uns sparen.

# Berechnung der Regressionsgleichung

- Slope/Steigung:  $b = \frac{\sum(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{SP(x,y)}{SQ(x)}$
- Intercept:  $a = \bar{y} - b \cdot \bar{x}$
- Der Beweis, dass dies die Gerade mit den kleinsten Quadraten schätzt, erfordert bereits erheblichen mathematischen Aufwand, den wir uns sparen.
- Determinationskoeffizient:  $r^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$

- Wie stark variiert der Fehler für Stichproben einer Größe?

- Wie stark variiert der Fehler für Stichproben einer Größe?

- $SF_{residual} = \sqrt{\frac{\sum e^2}{n-2}}$

- Wie stark variiert der Fehler für Stichproben einer Größe?
- $SF_{residual} = \sqrt{\frac{\sum e^2}{n-2}}$
- Je kleiner  $SF_{residual}$ , desto besser das Modell.



- Wie stark variiert der Fehler für Stichproben einer Größe?
- $SF_{residual} = \sqrt{\frac{\sum e^2}{n-2}}$
- Je kleiner  $SF_{residual}$ , desto besser das Modell.
- Beachte:  $n$  wird größer (größere Stichprobe):  $SF_{residual}$  wird kleiner.

# Standardfehler für die Gleichung

- Wie stark variiert der Fehler für Stichproben einer Größe?
- $SF_{residual} = \sqrt{\frac{\sum e^2}{n-2}}$
- Je kleiner  $SF_{residual}$ , desto besser das Modell.
- Beachte:  $n$  wird größer (größere Stichprobe):  $SF_{residual}$  wird kleiner.
- Und: Fehler  $e$  werden kleiner:  $SF_{residual}$  wird kleiner.

- Wie bei ANOVA:  $F = \frac{\text{erklärte Varianz}}{\text{zufällige Varianz}} = \frac{s_{\text{regression}}^2}{s_{\text{residual}}^2}$

- Wie bei ANOVA:  $F = \frac{\text{erklärte Varianz}}{\text{zufällige Varianz}} = \frac{s_{\text{regression}}^2}{s_{\text{residual}}^2}$
- zufällige Varianz:  $s_{\text{residual}}^2 = \frac{(1-r^2) \cdot \text{SQ}(y)}{1}$

- Wie bei ANOVA:  $F = \frac{\text{erklärte Varianz}}{\text{zufällige Varianz}} = \frac{S_{\text{regression}}^2}{S_{\text{residual}}^2}$
- zufällige Varianz:  $S_{\text{residual}}^2 = \frac{(1-r^2) \cdot \text{SQ}(y)}{1}$
- erklärte Varianz:  $S_{\text{regression}}^2 = \frac{r^2 \cdot \text{SQ}(y)}{n-2}$

- Wie bei ANOVA:  $F = \frac{\text{erklärte Varianz}}{\text{zufällige Varianz}} = \frac{S_{\text{regression}}^2}{S_{\text{residual}}^2}$
- zufällige Varianz:  $S_{\text{residual}}^2 = \frac{(1-r^2) \cdot \text{SQ}(y)}{1}$
- erklärte Varianz:  $S_{\text{regression}}^2 = \frac{r^2 \cdot \text{SQ}(y)}{n-2}$
- Freiheitsgrade sind immer  $df_1 = 1$  und  $df_2 = n - 1$ .

- Wie bei ANOVA:  $F = \frac{\text{erklärte Varianz}}{\text{zufällige Varianz}} = \frac{s_{\text{regression}}^2}{s_{\text{residual}}^2}$
- zufällige Varianz:  $s_{\text{residual}}^2 = \frac{(1-r^2) \cdot \text{SQ}(y)}{1}$
- erklärte Varianz:  $s_{\text{regression}}^2 = \frac{r^2 \cdot \text{SQ}(y)}{n-2}$
- Freiheitsgrade sind immer  $df_1 = 1$  und  $df_2 = n - 1$ .
- Beachte:  $r^2$  ist in  $[0..1]$  und teilt die Varianz von  $y$  auf.

- Für  $b$  und  $a$  kann je ein Standardfehler angegeben werden.



- Für  $b$  und  $a$  kann je ein Standardfehler angegeben werden.

- $SF(b) = \frac{\sqrt{\frac{\sum e^2}{n-1}}}{\sqrt{SQ(x)}}$

- Für  $b$  und  $a$  kann je ein Standardfehler angegeben werden.

- $SF(b) = \frac{\sqrt{\frac{\sum e^2}{n-1}}}{\sqrt{SQ(x)}}$

- Unter der NULL:  $b = 0$  ist dann t-verteilt:

$$t = \frac{b}{SF(b)}$$

- Design bei einfachem LM:

- Design bei einfachem LM:
  - eine intervallskalierte Abhängige

- Design bei einfachem LM:
  - eine intervallskalierte Abhängige
  - eine Unabhängige

- Design bei einfachem LM:
  - eine intervallskalierte Abhängige
  - eine Unabhängige
- wie bei mehrfaktorieller ANOVA:

- Design bei einfachem LM:
  - eine intervallskalierte Abhängige
  - eine Unabhängige
- wie bei mehrfaktorieller ANOVA:
  - oft interessiert mehrfaktorielle Abhängigkeit

Mehrere Koeffizienten im allgemeinen linearen Modell:



Mehrere Koeffizienten im allgemeinen linearen Modell:

$$\hat{y} = b_1 \cdot x_1 + b_2 \cdot x_2 \dots b_n \cdot x_n + a$$

Mehrere Koeffizienten im allgemeinen linearen Modell:

$$\hat{y} = b_1 \cdot x_1 + b_2 \cdot x_2 \dots b_n \cdot x_n + a$$

Konzeptuell bleibt die Berechnung aller Werte und Tests gleich, die Mathematik wird ungleich komplizierter.

Mehrere Koeffizienten im allgemeinen linearen Modell:

$$\hat{y} = b_1 \cdot x_1 + b_2 \cdot x_2 \dots b_n \cdot x_n + a$$

Konzeptuell bleibt die Berechnung aller Werte und Tests gleich, die Mathematik wird ungleich komplizierter.

Man schreibt  $R^2$  statt  $r^2$ .

Die Residuen  $E$  müssen normalverteilt sein.  
(als Diagnostik für: Die Messwerte müssen normalverteilt sein.)

- Missverständnis: Test aller Residuen auf Normalität

Die Residuen  $E$  müssen normalverteilt sein.  
(als Diagnostik für: Die Messwerte müssen normalverteilt sein.)

- Missverständnis: Test aller Residuen auf Normalität
- denn: Für jedes  $x_i$  müssen die  $e$  normalverteilt sein.

Die Residuen  $E$  müssen normalverteilt sein.  
(als Diagnostik für: Die Messwerte müssen normalverteilt sein.)

- Missverständnis: Test aller Residuen auf Normalität
- denn: Für jedes  $x_i$  müssen die  $e$  normalverteilt sein.
- erfordert mehrere Messungen pro  $x_i$  oder Intervallbildung

Die Residuen  $E$  müssen normalverteilt sein.  
(als Diagnostik für: Die Messwerte müssen normalverteilt sein.)

- Missverständnis: Test aller Residuen auf Normalität
- denn: Für jedes  $x_i$  müssen die  $e$  normalverteilt sein.
- erfordert mehrere Messungen pro  $x_i$  oder Intervallbildung
- größere Stichproben, kleinere Probleme

Die Residuen  $E$  müssen normalverteilt sein.  
(als Diagnostik für: Die Messwerte müssen normalverteilt sein.)

- Missverständnis: Test aller Residuen auf Normalität
- denn: Für jedes  $x_i$  müssen die  $e$  normalverteilt sein.
- erfordert mehrere Messungen pro  $x_i$  oder Intervallbildung
- größere Stichproben, kleinere Probleme
- visuelle Diagnose: Q-Q-Plots (hier nicht behandelt)



Jedes  $y_i$  darf nur von  $x_i$  abhängen,  
niemals zusätzlich von  $x_j$  mit  $i \neq j$ .

- mathematisch: nicht-lineare Abhängigkeit

Jedes  $y_i$  darf nur von  $x_i$  abhängen,  
niemals zusätzlich von  $x_j$  mit  $i \neq j$ .

- mathematisch: nicht-lineare Abhängigkeit
- konzeptuell: Zeitserien

Jedes  $y_i$  darf nur von  $x_i$  abhängen,  
niemals zusätzlich von  $x_j$  mit  $i \neq j$ .

- mathematisch: nicht-lineare Abhängigkeit
- konzeptuell: Zeitserien
- konzeptuell: Sequenzen in Texten

Jedes  $y_i$  darf nur von  $x_i$  abhängen,  
niemals zusätzlich von  $x_j$  mit  $i \neq j$ .

- mathematisch: nicht-lineare Abhängigkeit
- konzeptuell: Zeitserien
- konzeptuell: Sequenzen in Texten
- Lösung: andere Modellspezifikation

Die Residuen müssen homoskedastisch verteilt sein.

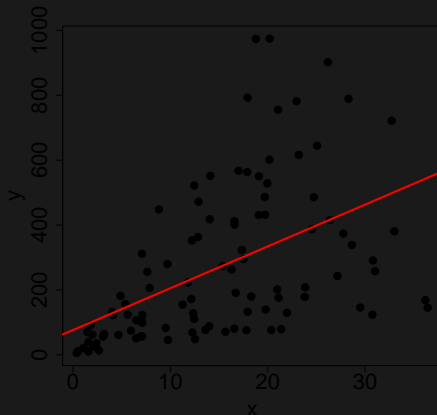
- Bedeutung: Die Varianz der  $e$  muss über alle  $x$  homogen sein.

Die Residuen müssen homoskedastisch verteilt sein.

- Bedeutung: Die Varianz der  $e$  muss über alle  $x$  homogen sein.
- vgl. die Forderung der „Varianzhomogenität“ bei t-Test und ANOVA

# Darstellung heteroskedastischer Residuen

Hier wird die Varianz der Residuen mit steigendem  $x$  immer größer.  
Ein lineares Modell versagt hier wegen verletzter Verteilungsannahmen.



- mehr Daten ziehen, Daten transformieren



- mehr Daten ziehen, Daten transformieren
- generalisierte lineare Modelle (GLM)  
legen andere Verteilungsannahmen zugrunde

- mehr Daten ziehen, Daten transformieren
- generalisierte lineare Modelle (GLM)  
legen andere Verteilungsannahmen zugrunde
- (generalisiert) additive Modelle (GAM)  
schätzen Smoothingfunktionen für Koeffizienten

# ANOVA als Modell mit kategorialen Regressoren

$n$  Gruppen der ANOVA können als  $n$  dichotome Variablen dargestellt werden:

		ANOVA-Gruppen		
		$A_1$	$A_2$	$A_3$
Regressor	$x_1 =$	1	0	0
	$x_2 =$	0	1	0
	$x_3 =$	0	0	1

Normale Modellspezifikation:

$$\hat{y} = b_1x_1 + b_2x_2 + \dots + b_nx_n + a$$

Normale Modellspezifikation:

$$\hat{y} = b_1x_1 + b_2x_2 + \dots + b_nx_n + a$$

Da jeweils nur eins der  $x_i = 1$  und alle anderen immer 0 werden, wird einfach der Wert des entsprechenden  $\beta_i$  (plus  $a$ ) vorhergesagt.

# Spearman's Rang-Korrelation in R

Die Funktion `cor()` hat ein Argument `method`, das als "spearman" angegeben werden kann.

```
> cor(x, y, method = "spearman")
```

- Modellformeln:  $y \sim x$   
„y abhängig von x“

- Modellformeln:  $y \sim x$   
„y abhängig von x“
- Mehrere Unabhängige:  $y \sim x_1 + x_2$



- Modellformeln:  $y \sim x$   
„y abhängig von x“
- Mehrere Unabhängige:  $y \sim x_1 + x_2$
- Mehrere Unabhängige mit Interaktion:  $y \sim x_1 * x_2$

- Modellformeln:  $y \sim x$   
„y abhängig von x“
- Mehrere Unabhängige:  $y \sim x_1 + x_2$
- Mehrere Unabhängige mit Interaktion:  $y \sim x_1 * x_2$
- Mehrere Unabhängige nur Interaktion:  $y \sim x_1 : x_2$

- Modellformeln:  $y \sim x$   
„y abhängig von x“
- Mehrere Unabhängige:  $y \sim x_1 + x_2$
- Mehrere Unabhängige mit Interaktion:  $y \sim x_1 * x_2$
- Mehrere Unabhängige nur Interaktion:  $y \sim x_1 : x_2$
- Lineares Modell schätzen und speichern:  

```
> m <- lm(y~x)
```

- Modellformeln:  $y \sim x$   
„y abhängig von x“
- Mehrere Unabhängige:  $y \sim x_1 + x_2$
- Mehrere Unabhängige mit Interaktion:  $y \sim x_1 * x_2$
- Mehrere Unabhängige nur Interaktion:  $y \sim x_1 : x_2$
  
- Lineares Modell schätzen und speichern:  

```
> m <- lm(y~x)
```
- Ausgabe Evaluation:  

```
> summary(m)
```

Interpretieren Sie diese Ausgabe anhand der Folien:

```
Call:
lm(formula = y ~ x)

Residuals:
Min       1Q   Median       3Q      Max
-20.4298  -2.4920  -0.2625   3.8038  14.2922

Coefficients:
(Intercept)  1.513      4.321  0.350      0.73
x            9.242      1.333  6.933 1.77e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.008 on 18 degrees of freedom
Multiple R-squared:  0.7275, Adjusted R-squared:  0.7124
F-statistic: 48.06 on 1 and 18 DF,  p-value: 1.768e-06
```

## Lineare Modelle und ANOVA

# Äquivalenz von ANOVA und LM

Binäre Kodierung der Gruppenzugehörigkeit

Hier: drei Gruppen von einem Faktor (einfaktorielle ANOVA mit drei Gruppen)

		ANOVA-Gruppen		
		$A_1$	$A_2$	$A_3$
Regressor	$x_1 =$	1	0	0
	$x_2 =$	0	1	0
	$x_3 =$	0	0	1

$$\hat{y} = b_1 x_1 + b_2 x_2 + \dots + b_n x_n + a \quad (1)$$

Kleinste Quadrate | für jeden Koeffizienten  $b_i$  jeweils Mittelwert der Gruppe  $i$  ( $\bar{y}_i$ )  
Außerdem erstmal  $a = 0$  | dann:

$$\hat{y} = \bar{y}_1 x_1 + \bar{y}_2 x_2 + \dots + \bar{y}_n x_n \quad (2)$$

Allgemeines Mittel als Intercept |  $a = \bar{Y}$

Koeffizienten = Abweichung Gruppenmittel vom Gesamtmittel

$$\hat{y} = (\bar{y}_1 - \bar{Y})x_1 + (\bar{y}_2 - \bar{Y})x_2 + \dots + (\bar{y}_n - \bar{Y})x_n + \bar{Y} \quad (3)$$



## Beispiel für einen Datenpunkt aus Gruppe 2

Entsprechend in ANOVA  $A_2$  | Unabhängige im LM:  $x_1 = 0$ ,  $x_2 = 1$  und  $x_3 = 0$   
Schätzung für den  $y$ -Wert:

$$\begin{aligned}\hat{y} &= (\bar{y}_1 - \bar{Y})0 + (\bar{y}_2 - \bar{Y})1 + \dots + (\bar{y}_n - \bar{Y})0 + \bar{Y} \\ &= 0 + (\bar{y}_2 - \bar{Y}) + \dots + 0 + \bar{Y} \\ &= \bar{y}_2 - \bar{Y} + \bar{Y} \\ &= \bar{y}_2\end{aligned}\tag{4}$$

Jeder  $\hat{y}$ -Wert | Mittel der beobachteten  $y$ -Werte der Gruppe, zu der er gehört

Das ergibt für ausschließlich nominale Unabhängige in der Tat den Schätzer mit den kleinsten Quadraten (s. Maxwell & Delaney, Kap. 3).

Kernfrag | Bringen Abweichungen der Gruppenmittel eine Verbesserung der Vorhersage gegenüber dem Gesamt-Mittel?

Methode | Vergleich der Residuen  $E_f$  für volles Modell (mit Gruppenmitteln)  
vs. Residuen  $E_r$  für reduziertes Modell (ohne Gruppenmittel)

Gleichung 5 für volles und Gleichung 6 für reduziertes Modell

$$\hat{y}_f = (\bar{y}_1 - \bar{Y})x_1 + (\bar{y}_2 - \bar{Y})x_2 + \dots + (\bar{y}_n - \bar{Y})x_n + \bar{Y} \quad (5)$$

$$\hat{y}_r = \bar{Y} \quad (6)$$

$$F = \frac{\frac{E_r - E_f}{df_r - df_f}}{\frac{E_f}{df_f}} = \frac{\text{erklärte Varianz}}{\text{zufällige Varianz}} \quad (7)$$

- 1 Residuen = Maß für die Varianz
- 2 Residuen des vollen Modells = Maß für die Varianz, die trotz der Erklärungskraft des vollen Modells bleibt (= unerklärte Varianz)
- 3 Residuen des reduzierten Modells = Maß für die Gesamtvarianz (Abweichungen vom Gesamt-Mittel)
- 4 Zähler | trotz der Erklärung verbleibende Varianz – vollen Varianz abgezogen (= erklärte Varianz)
- 5 Quotient insgesamt | Zählervarianz in Bezug zur Gesamtvarianz (klassischer F-Quotient, s. ANOVA)

Nächste Woche | Überblick

- 1 Inferenz
- 2 Deskriptive Statistik
- 3 Nichtparametrische Verfahren
- 4 z-Test und t-Test
- 5 ANOVA
- 6 Freiheitsgrade und Effektstärken
- 7 Power und Severity
- 8 Lineare Modelle
- 9 Generalisierte Lineare Modelle
- 10 Gemischte Modelle

- Gravetter, Frederick J. & Larry B. Wallnau. 2007. *Statistics for the Behavioral Sciences*. 7. Aufl. Belmont: Thomson.
- Maxwell, Scott E. & Harold D. Delaney. 2004. *Designing experiments and analyzing data: a model comparison perspective*. Mahwa, New Jersey, London: Taylor & Francis.
- Zuur, Alain F., Elena N. Ieno, Neil Walker, Anatoly A. Saveliev & Graham M. Smith. 2009. *Mixed effects models and extensions in ecology with R*. Berlin etc.: Springer.

## Kontakt

Prof. Dr. Roland Schäfer  
Institut für Germanistische Sprachwissenschaft  
Friedrich-Schiller-Universität Jena  
Fürstengraben 30  
07743 Jena

<https://rolandschaefer.net>  
[roland.schaefer@uni-jena.de](mailto:roland.schaefer@uni-jena.de)

## Creative Commons BY-SA-3.0-DE

Dieses Werk ist unter einer Creative Commons Lizenz vom Typ *Namensnennung - Weitergabe unter gleichen Bedingungen 3.0 Deutschland* zugänglich. Um eine Kopie dieser Lizenz einzusehen, konsultieren Sie

<http://creativecommons.org/licenses/by-sa/3.0/de/> oder wenden Sie sich brieflich an Creative Commons, Postfach 1866, Mountain View, California, 94042, USA.