TUTORIUM Datenauswertung

REGRESSIONSANALYSE

Agenda

- Regressionsanalyse – Basics
- Regressionsanalyse in R
- Interpretation von Regressionsanalysen
- Prognosen mit Regressionsanalysen
- Drittvariablenkontrolle mit Regressionsanalysen

Regressionsanalyse – Basics

Die Regressionsanalyse

– Ist ein Verfahren, das benutzt wird, um
  - Den Einfluss mindestens einer unabhängigen Variablen auf eine abhängige Variable zu untersuchen → Kausalannäherung
  - Scheinkorrelationen zu vermeiden → Drittvariablenkontrolle
  - Hypothesen über die Zukunft aufzustellen → Prognosen

– Gibt es in verschiedenen Arten:
  - Linear vs. Nicht-linear
  - Einfach (eine UV) vs. multipel (mehr als eine UV)

Lineare Regressionsanalyse – Basics

Die lineare Regressionsanalyse

- Stellt ein lineares Gleichungsmodell auf, sodass
  - Eine abhängige Variable durch eine Menge unabhängiger Variablen geschätzt wird
  - die Summe der quadrierten Abstände der abhängigen Variable von der geschätzten abhängigen Variable minimal ist
- Erfordert metrisch skalierte abhängige und unabhängige Variablen

Lineare Regressionsanalyse – Gleichungsmodell

$\hat{y}_i = a + b_1 * x_1 + \ldots + b_n * x_n$

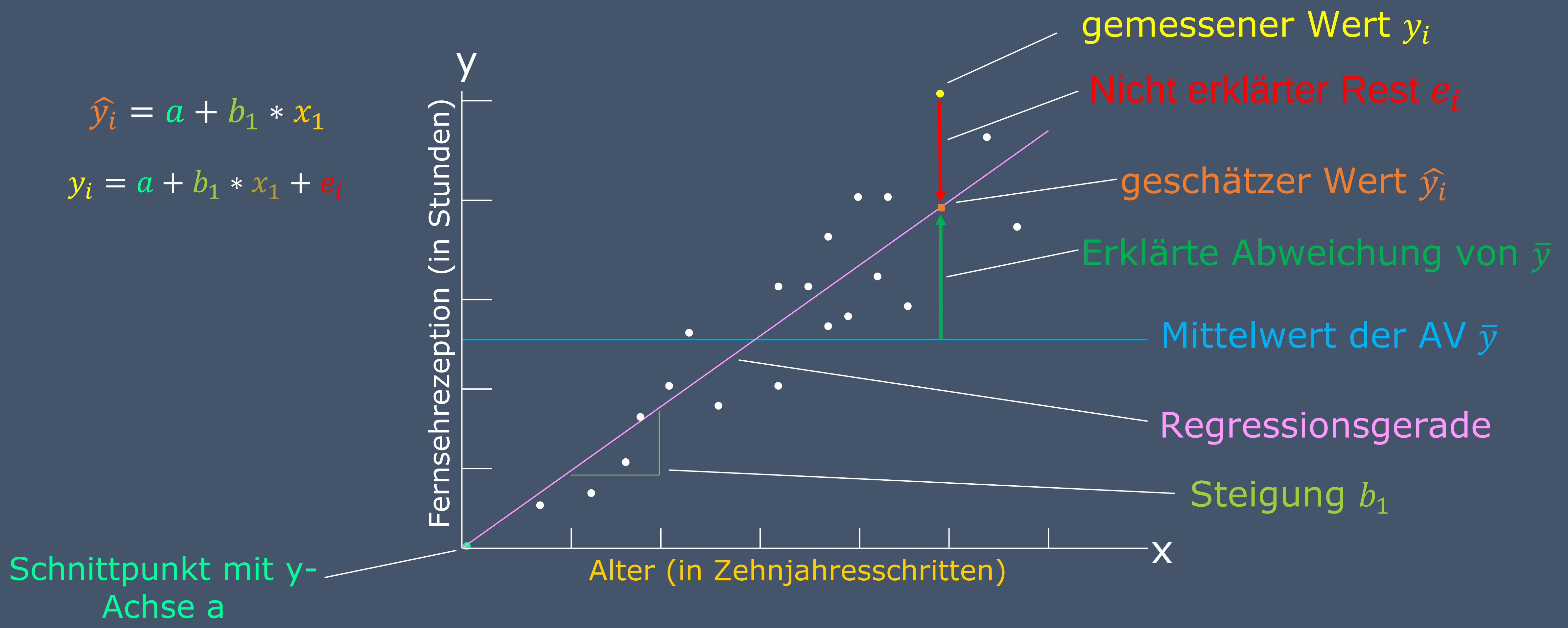
$y_i = a + b_1 * x_1 + \ldots + b_n * x_n + e_i$

Variable	Bedeutung	Name in R
y_i	Tatsächlich gemessener Wert der AV für Fall i	[Name der AV in R]
$\hat{y}_i$	Geschätzter Wert der AV für Fall i	-
a	Schnittpunkt der Regressionsgeraden mit der y-Achse	Intercept
b_1	Steigung der AV, wenn x_1 um 1 steigt (Regressionskoeffizient)	Estimate
x_1	Erste unabhängige Variable der Gleichung	[Name der ersten UV in R]
e_i	Abweichung des geschätzten Wertes $\hat{y}_i$ vom tatsächlich gemessenen Wert y_i	Residuals

LINEARE REGRESSIONSANALYSE – BEISPIEL

$\widehat{y_i} = a + b_1 * x_1$

$y_i = a + b_1 * x_1 + e_i$



Der Determinationskoeffizient R^2

Der Determinationskoeffizient R^2

- ist ein Maß dafür, welchen Anteil der Varianz der AV das Modell erklärt:

Interpretation	Mathematisch	graphisch
Das Modell erklärt so viel Varianz wie der Mittelwert	$R^2 = 0$	die Regressionsgerade ist identisch mit der Mittelwertgeraden
Das Modell erklärt die gesamte Varianz der AV	$R^2 = 1$	Alle Punkte liegen auf der Regressionsgeraden

- Sollte größer als 0,1 sein
- Ist für einfache Regression identisch mit r^2
- Wird bei multipler Regression an die Anzahl der UVs angepasst → adjusted R^2

Der Betakoeffizient β

Der Betakoeffizient β

- Ist der standardisierte Regressionskoeffizient b
- Wird berechnet, weil b von der Maßeinheit und der Streuungsbreite abhängt
- Macht die Effekte der unabhängigen Variablen auf die abhängige Variable vergleichbar
- lässt sich für die nte UV so berechnen: $\beta_n = \frac{b_n}{s_{x_n} * s_y}$

Regressionsanalyse in R: Beschreibung der Daten

```
#Laden der Pakete
library(knitr)
library(mosaic)
# Laden und Auswählen der Daten
load("daten_x.RData")
datensatz <- subset(daten_x, Bedingung)
#Berechnung der Lage- und Streuungsmaße
kable(round(rbind("AV-Label" = favstats(datensatz$AV),
                  "UV1-Label" = favstats(datensatz$UV1),
                  "UV2-Label" = favstats(datensatz$UV2),
                  "UV3-Label" = favstats(datensatz$UV3)), 2))
```

Regressionsanalyse in R: Beschreibung der Daten – Beispiel

```
#Laden der Pakete
library(knitr)
library(mosaic)
# Laden und Auswählen der Daten
load("daten_2019.RData")
datensatz <- subset(daten_2019, jahr > 2000)
#Berechnung der Lage- und Streuungsmaße
kable(round(rbind("Radiokonsum" = favstats(datensatz$radio_minuten),
                  "Zeitungskonsum" = favstats(datensatz$tz_minuten),
                  "Jahrgang" = favstats(datensatz$soz_jahrgang),
                  "Haushaltsgröße" = favstats(datensatz$soz_haushalt)), 2))
```

	min	Q1	median	Q3	max	mean	sd	n	missing
Radiokonsum	0	10	45	120	720	93.09	127.05	482	6
Zeitungskonsum	0	0	15	30	180	22.10	24.80	476	12
Jahrgang	1943	1963	1971	1991	2004	1974.09	16.00	488	0
Haushaltsgröße	0	1	2	3	19	1.83	1.51	486	2

Regressionsanalyse in R: Regressionsmodell

```
#Laden der Pakete
library(knitr)
library(QuantPsyc)
# Laden und Auswählen der Daten
load("daten_x.RData")
datensatz <- subset(daten_x, Bedingung)
#Regressionsmodell
regression = lm(AV ~ UV1 + UV2 + UV3, data= datensatz)
summary(regression)
lm.ergebnis <- summary(regression)
coeff <- round(lm.ergbnis$coefficients[-1,], 2)
Std.coeff <- round(lm.beta(regression), 2)
coeff.gesamt <- data.frame(coeff, Std.coeff)
kable(coeff.gesamt)
```

Regressionsanalyse in R: Regressionsmodell – Beispiel

```
#Laden der Pakete
library(knitr)
library(QuantPsyc)
# Laden und Auswählen der Daten
load("daten_2019.RData")
datensatz <- subset(daten_2019, jahr > 2000)
#Regressionsmodell
regression = lm(radio_minuten ~ tz_minuten + soz_jahrgang + soz_haushalt, data= datensatz)
summary(regression)
lm.ergebnis = summary(regression)
coeff = round(lm.ergebnis$coefficients[-1,], 2)
std.coeff = round(lm.beta(regression), 2)
coeff.gesamt = data.frame(coeff, std.coeff)
kable(coeff.gesamt)
```

AV UV1 UV2 UV3

```
Call:
lm(formula = radio_minuten ~ tz_minuten + soz_jahrgang + soz_haushalt,
    data = datensatz)
```

Verteilung der nicht erklärten Abweichungen vom gemessenen Wert

```
Residuals:
    Min      1Q  Median      3Q     Max
-160.22  -68.69  -39.30   13.47  620.25
```

Signifikanzwerte der Regressionskoeffizienten

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4177.6372   792.7439   5.270 2.09e-07 ***
tz_minuten      0.1519     0.2573   0.590    0.555
soz_jahrgang   -2.0702     0.4007  -5.166 3.55e-07 ***
soz_haushalt   -0.6954     3.7695  -0.184    0.854
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

a; x_1, b_1; x_2, b_2; x_3, b_3

signifikant

R² < 0,1 → schlechtes Modell!

```
Residual standard error: 123.3 on 467 degrees of freedom
  (17 observations deleted due to missingness)
Multiple R-squared:  0.07579,   Adjusted R-squared:  0.06985
F-statistic: 12.77 on 3 and 467 DF,  p-value: 4.959e-08
```

p < 0.05: signifikant!

Unabhängige Variablen Regressionkoeffizienten b

Signifikanzwerte der Regressionskoeffizienten

Standardisierte Regressionskoeffizienten β

	Estimate	Std..Error	t.value	Pr...t..	std.coeff
tz_minuten	0.15	0.26	0.59	0.56	0.03
soz_jahrgang	-2.07	0.40	-5.17	0.00	-0.26
soz_haushalt	-0.70	3.77	-0.18	0.85	-0.01

Regression – Kausalannäherung

Der kausale Einfluss einer UV auf eine AV ist zu vermuten dann und nur dann, wenn…

- Die Vermutung über den Effekt auf einer Theorie basiert
- Alle möglichen Einflussfaktoren miterhoben und ins Regressionsmodell miteinbezogen wurden
- Der Regressionskoeffizient der UV dennoch signifikant ist

Regression – Prognosen

Prognosen

- Sind Schätzungen über die (möglicherweise zukünftige) Ausprägung einer AV aufgrund mindestens einer UV
- Können anhand von Regressionsgleichungen aufgestellt werden, indem man die Werte der UVs für den betrachteten Fall in die Regressionsgleichung einsetzt
- Beispiel: Wetterprognosen

Prognosen – Beispiel I

```
Call:
lm(formula = daten_2019$tv_minuten ~ daten_2019$soz_jahrgang +
    daten_2019$tz_minuten)

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)             3869.8680   503.2073   7.690 8.64e-14 ***
daten_2019$soz_jahrgang   -1.9134     0.2541  -7.531 2.60e-13 ***
daten_2019$tz_minuten      0.5965     0.1651   3.614 0.000334 ***
```

Form: $\hat{y}_i = a + b_1 * x_1 + \ldots + b_n * x_n$

Eingesetzt: $\widehat{tv\_minuten}_i = 3870 - 1{,}9 * soz\_jahrgang + 0{,}6 * tz\_minuten$

Prognosen – Beispiel II

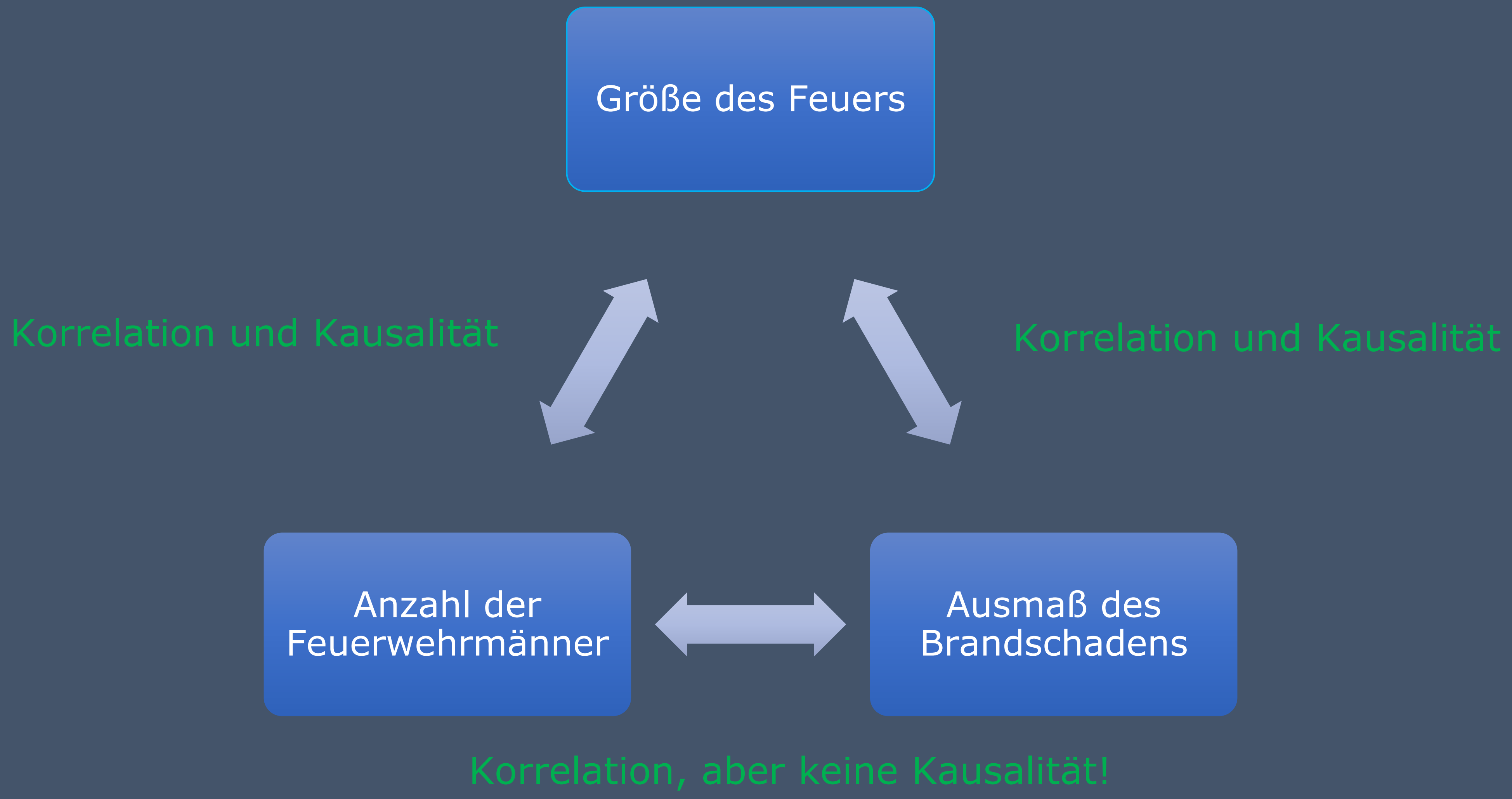
Formel: $\widehat{tv\_minuten}_i = 3870 - 1{,}9 * soz\_jahrgang + 0{,}6 * tz\_minuten$

Aufgabe	Peter ist 1990 geboren und liest 45 Minuten Zeitung pro Tag. Schätzt, wie lange er pro Tag fern schaut.
Rechnung	$\widehat{tv\_minuten}_{Peter} = 3870 - 1{,}9 * 1990 + 0{,}6 * 45 = 116$
Lösung	Peter schaut voraussichtlich ca. 2 Stunden fern pro Tag.

Regression – Drittvariablenkontrolle

Die Drittvariablenkontrolle

- Prüft den Einfluss einer möglichen Störvariable auf eine unabhängige Variable
- Deckt damit Scheinkorrelationen auf
- Kann durchgeführt werden, indem die Störvariable als weitere UV in die Regressionsgleichung miteinbezogen wird:
  - Der Regressionskoeffizient der UV ist signifikant → keine Scheinkorrelation
  - Der Regressionskoeffizient der UV ist nicht signifikant → Scheinkorrelation


Größe des Feuers
Korrelation und Kausalität
Korrelation und Kausalität
Anzahl der Feuerwehrmänner
Ausmaß des Brandschadens
Korrelation, aber keine Kausalität!

	Anzahl der Feuerwehrmänner	Höhe des Brandschadens	Größe des Feuers
Anzahl der Feuerwehrmänner	1.00	0.79	0.92
Höhe des Brandschadens	0.79	1.00	0.84
Größe des Feuers	0.92	0.84	1.00

→ Alle Variablen korrelieren sehr stark miteinander, auch die Höhe des Brandschadens mit der Anzahl der Feuerwehrmänner!

AV UV Störvariable

```
Call:
lm(formula = brandschaden ~ anzahl_feuerwehr + grösse_feuer)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.0408 -0.2424  0.4968  0.5087  1.3089 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)  
(Intercept)        0.4677     0.8497   0.550   0.5892  
anzahl_feuerwehr   0.1245     0.3276   0.380   0.7087  
grösse_feuer       0.7629     0.3593   2.124   0.0487 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.484 on 17 degrees of freedom
Multiple R-squared:  0.7092,	Adjusted R-squared:  0.6749 
F-statistic: 20.73 on 2 and 17 DF,  p-value: 2.761e-05
```

Nicht signifikant → Scheinkorrelation!

Signifikant → keine Scheinkorrelation!

Extrem gutes Modell!

Signifikantes Modell!

FRAGEN?

Selbststudium

- Forschungsbericht (nur EFB):
  - Regressionsmodell: www_minuten ~ tz_minuten + soz_haushalt + soz_jahrgang
  - Inhaltliche Sätze zu den Ergebnissen
  - Vergleich tz_minuten im Regressionsmodell und Korrelation tz_minuten, www_minuten → Scheinkorrelation?
- Fragen zur Wiederholungssitzung sammeln und bis Freitag, <u>29.06.2019, 23:59 Uhr</u> per Mail zuschicken
- Forschungsbericht bis Sonntag, <u>30.06.2019, 23:59 Uhr</u> ins Learnweb hochladen

Bis nächste Woche!