Tutorium Datenauswertung

Assoziationsmaße

Agenda

- Assoziationsmaße
- Pearson‘s r per Hand
- Probleme mit Pearson‘s r
- Pearson‘s r mit R
- Scatterplots in R

Assoziationsmaße

Assoziationsmaße

- Messen die Stärke des Zusammenhangs einer Zusammenhangshypothese
- Erstrecken sich (bis auf V) von -1 bis 1
- Gibt es für verschiedene Skalenniveaus:
  - Nominal: Cramer‘s V
  - Ordinal: Spearman‘s ρ
  - Intervall: Kendall‘s τ
  - Rational: Pearson‘s r

Kovarianz und Varianz

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$ | Formel Varianz

$$\leftrightarrow s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) * (x_i - \bar{x})}{n-1}$$ | Formel Varianz (umgeformt)

$$cov\,(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) * (y_i - \bar{y})}{n-1}$$ | Formel Kovarianz

Kovarianz – Beispiel

Fall i	1	2	3	4	5
Radiokonsum x_i	60	120	30	120	180
Zeitungskonsum y_i	15	0	10	20	15

$$\bar{x} = \frac{60 + 120 + 30 + 120 + 180}{5} = 102 \qquad \bar{y} = \frac{15 + 0 + 10 + 20 + 15}{5} = 12$$

Fall i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x}) * (y_i - \bar{y})$
1	60 – 102 = -42	15 – 12 = 3	-42 * 3 = -126
2	120 – 102 = 18	0 – 12 = -12	18 * (-12) = -216
3	30 – 102 = -72	10 – 12 = -2	-72 * (-2) = 144
4	120 – 102 = 18	20 – 12 = 8	18 * 8 = 144
5	180 – 102 = 78	15 – 12 = 3	78 * 3 = 234
			$-126 + (-216) + 144 + 144 + 234 = 180$

$$cov\,(x, y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) * (y_i - \bar{y})}{n - 1} = \frac{180}{5 - 1} = 45$$

$$\sum_{i=1}^{n} (x_i - \bar{x}) * (y_i - \bar{y})$$

Pearson's r

Pearson's r

- Ist die durch das Produkt der Standardabweichungen standardisierte Kovarianz
- Ist ein sehr gängiges Zusammenhangsmaß in den Kommunikationswissenschaften
- Gibt an, wie stark zwei metrische Variablen zusammenhängen:
  - -1 → annähernd perfekter negativer linearer Zusammenhang
  - 0 → kein Zusammenhang
  - 1 → annähernd perfekter positiver linearer Zusammenhang

Pearson's r – Formeln und Interpretation

$$r = \frac{cov\,(x,y)}{s_x * s_y} = \frac{\frac{\sum_{i=1}^{n}(x_i - \bar{x}) * (y_i - \bar{y})}{n-1}}{\sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} * \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}}$$

r-Wert (negativ)	R-Wert (positiv)	Stärke des Einflusses
-0,1 ≤ r ≤ 0	0 ≤ r ≤ 0,1	nicht vorhanden
-0,2 ≤ r ≤ -0,1	0,2 ≤ r ≤ 0,1	schwach
-0,4 ≤ r ≤ -0,2	0,2 ≤ r ≤ 0,4	mittel
-0,8 ≤ r ≤ -0,4	0,4 ≤ r ≤ 0,8	stark
-1,0 ≤ r ≤ -0,8	0,8 ≤ r ≤ 1,0	(annähernd) linear

Pearson's r – Beispiel

$cov\,(x,y) = 45, \quad \bar{x} = 102, \quad \bar{y} = 12$

Fall i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1	60 – 102 = -42	1764
2	120 – 102 = 18	324
3	30 – 102 = -72	5184
4	120 – 102 = 18	324
5	180 – 102 = 78	6084
		$\sum$ = 13680

Fall i	$y_i - \bar{y}$	$(y_i - \bar{y})^2$
1	15 – 12 = 3	9
2	0 – 12 = -12	144
3	10 – 12 = -2	4
4	20 – 12 = 8	64
5	15 – 12 = 3	9
		$\sum$ =230

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

$$s_x = \sqrt{\frac{13680}{5-1}} \approx 58{,}48$$

$$s_y = \sqrt{\frac{230}{5-1}} \approx 7{,}58$$

$$r = \frac{cov\,(x,y)}{s_x * s_y} = \frac{45}{58{,}48 * 7{,}58} \approx 0{,}10$$

$r \leq 0{,}1 \rightarrow$ Kein Zusammenhang!

Pearson's r – Aufgabe

Aufgabe: Prüft, ob die Variablen Radiokonsum und Internetkonsum zusammenhängen, und gebt, falls das so ist, an, wie stark sie zusammenhängen.

$$cov\,(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) * (y_i - \bar{y})}{n-1}, \qquad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i\,, \qquad \mathrm{r} = \frac{cov\,(x,y)}{s_x * s_y}, \qquad s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

$\bar{y} = 192$ $\qquad$ $s_y \approx 115{,}41$

Fall i		1	2	3	4	5
Radiokonsum	x_i	120	180	60	45	60
Internetkonsum	y_i	180	240	120	60	360

Pearson's R – Lösung I: Mittelwert und Standardabweichung (x)

Fall i		1	2	3	4	5
Radiokonsum	x_i	120	180	60	45	60

$$\bar{x} = \frac{120 + 180 + 60 + 45 + 60}{5} = 93$$

$$s_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{12780}{5-1}} \approx 56{,}52$$

Fall i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1	120 – 93 = 27	27² = 729
2	180 – 93 = 87	87² = 7569
3	60 – 93 = -33	-33² = 1089
4	45 – 93 = -48	-48² = 2304
5	60 – 93 = -33	-33² = 1089
		$\sum = 12780$

Pearson's R – Lösung II: Mittelwert und Standardabweichung (Y)

Fall i	1	2	3	4	5
Internetkonsum y_i	180	240	120	60	360

$$\bar{y} = \frac{180 + 240 + 120 + 60 + 360}{5} = 192$$

$$s_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}} = \sqrt{\frac{53280}{5-1}} \approx 115{,}41$$

Fall i	$y_i - \bar{y}$	$(y_i - \bar{y})^2$
1	180 – 192 = -12	-12² = 144
2	240 – 192 = 48	48² = 2304
3	120 – 192 = -72	-72² = 5184
4	60 – 192 = -132	-132² = 17424
5	360 – 192 = 168	168² = 28224
		$\sum = 53280$

Pearson's r – Lösung III: Kovarianz (x,y)

Fall i		1	2	3	4	5
Radiokonsum	x_i	120	180	60	45	60
Internetkonsum	y_i	180	240	120	60	360

Fall i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x}) * (y_i - \bar{y})$
1	120 – 93 = 27	180 – 192 = -12	27 * (-12) = -324
2	180 – 93 = 87	240 – 192 = 48	87 * 48 = 4176
3	60 – 93 = -33	120 – 192 = -72	-33 *(-72) = 2376
4	45 – 93 = -48	60 – 192 = -132	-48 * (-132) = 6336
5	60 – 93 = -33	360 – 192 = 168	-33 * 168 = -5544
			$\sum = 7020$

$$cov\,(x,y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) * (y_i - \bar{y})}{n-1} = \frac{7020}{5-1} = 1755$$

Pearson's r – Lösung IV: Standardisierung, Interpretation

$$cov\,(x,y) = 1755, \qquad s_x \approx 56{,}52, \qquad s_y \approx 115{,}41$$

$$\mathrm{r} = \frac{cov\,(x,y)}{s_x * s_y} = \frac{1755}{56{,}52 * 115{,}41} \approx 0{,}27$$

→ Es gibt einen mittelstarken positiven Zusammenhang zwischen den Variablen Radiokonsum und Internetkonsum.

Pearson's r – Probleme

Pearson's r

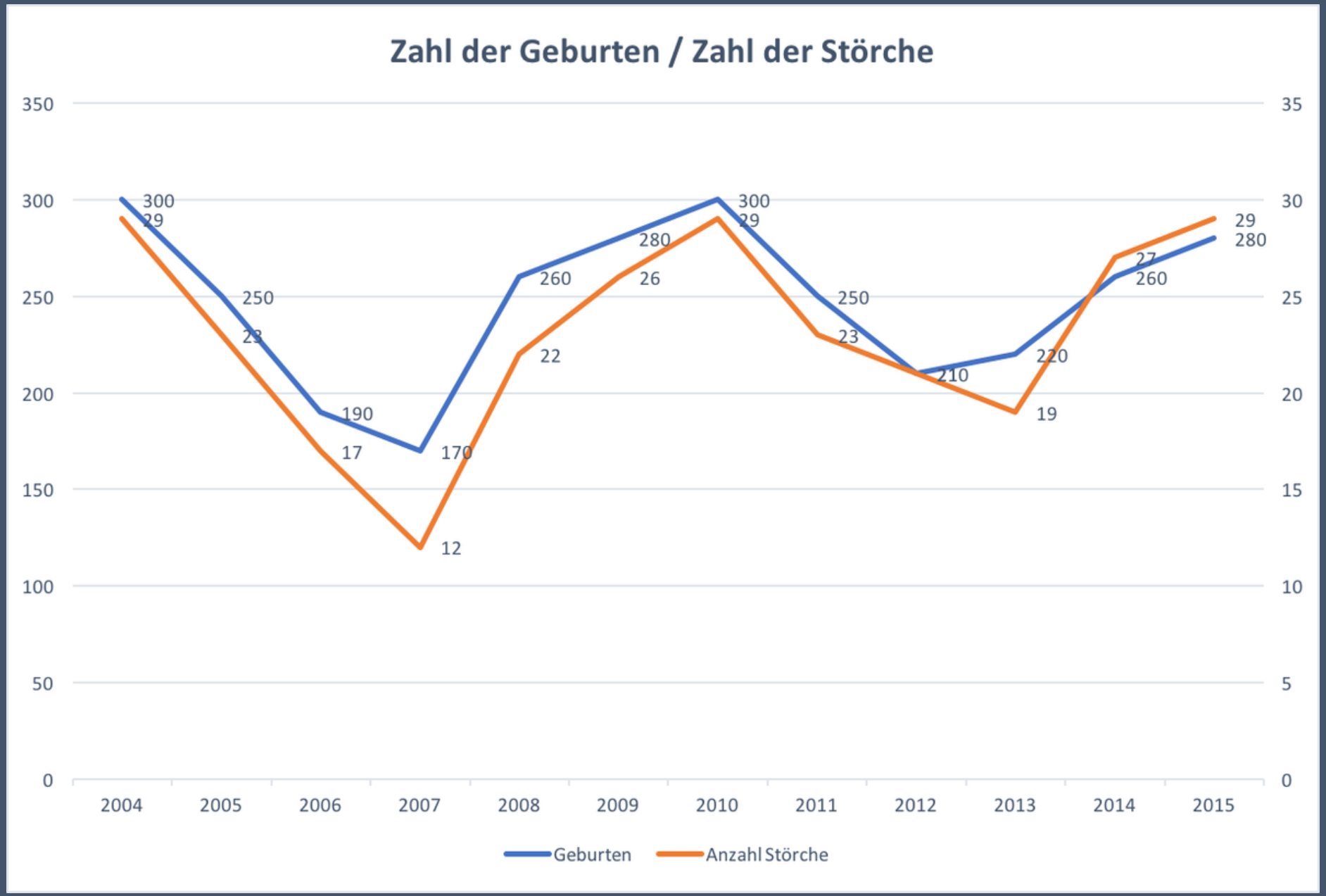
- Funktioniert nur für metrische und stetige Variablen
- Benötigt große Stichprobenzahlen (n>100) oder kleinere Stichprobenzahlen, wobei beide Variablen normalverteilt sind
- Ist extrem ausreißerempfindlich
- Kann deshalb ersetzt werden durch
  - Kendall's τ
  - Spearman's ρ
- Sagt nicht, ob die Variablen *kausal* zusammenhängen

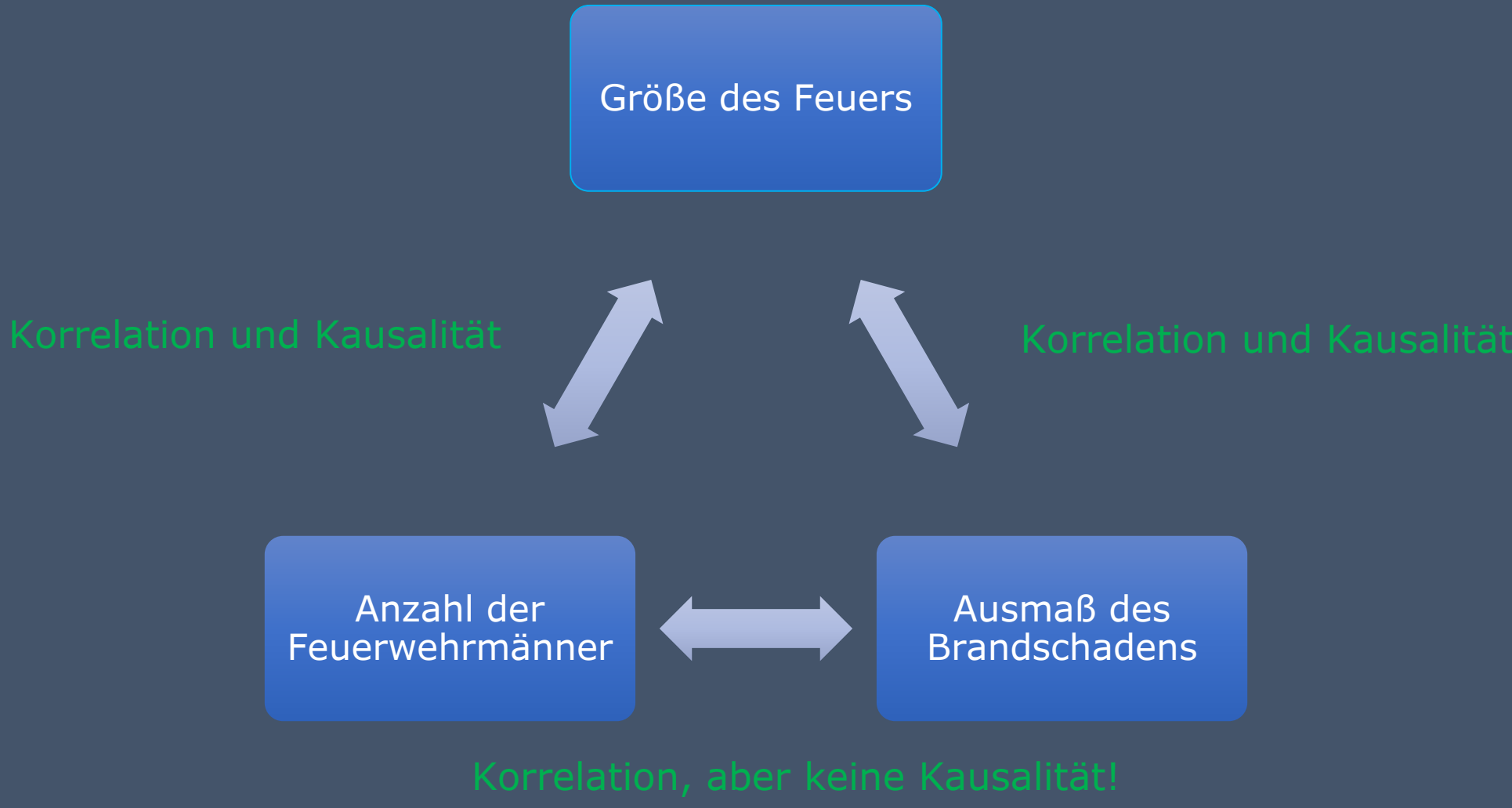
Korrelationen und Kausalität

<u>Die Korrelation zwischen zwei Variablen</u>

- ist notwendige Bedingung für einen Kausalzusammenhang dieser Variablen
- Ist dafür aber nicht hinreichend, denn
  - Die Korrelation könnte zufällig zustande gekommen sein
  - Die Variablen könnten aufgrund einer dritten Variable zusammenhängen
  - Die Korrelation sagt nichts über die Richtung aus (beeinflusst x y oder andersherum?)

→ Korrelationen dürfen nie als Kausalzusammenhänge interpretiert werden!


Zahl der Geburten / Zahl der Störche
350
300
250
200
150
100
50
0
35
30
25
20
15
10
5
0
300
29
250
23
190
17
170
12
260
22
280
26
300
29
250
23
210
19
220
27
260
29
280
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
Geburten
Anzahl Störche


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

Korrelationen in R: r-Matrix

```
#Laden und Auswahlen der Daten
library(knitr)
library(Hmisc)
load("daten_x.RData")
datensatz <- subset(daten_x, Bedingung)
#Berechnung
matrix= rcorr(cbind(datensatz$var1, datensatz$var2, datensatz$var3))
#Korrelationsmatrix zusammenstellen
rmatrix = data.frame(rmatrix$r, row.names = c("var1label", "var2label", "var3label"))
colnames(rmatrix) = c("var1label", "var2label", "var3label")
kable(round(rmatrix, 2), caption = "Name der Graphik")
```

Korrelationen in R: r Matrix - Beispiel

```
#Laden und Auswählen der Daten
library(knitr)
library(Hmisc)
load("daten_2019.RData")
datensatz <- subset(daten_2019, jahr>2000)
#Berechnung
matrix = rcorr(cbind(datensatz$tv_minuten, datensatz$tz_minuten, datensatz$www_minuten))
#Korrelationsmatrix zusammenstellen
rmatrix = data.frame(matrix$r, row.names = c("TV-Konsum", "Zeitungskonsum", "Inernetkonsum"))
colnames(rmatrix) = c("TV-Konsum", "Zeitungskonsum", "Inernetkonsum")
kable(round(rmatrix, 2), caption = "Korrelationsmatrix")
```

Korrelationsmatrix

	TV-Konsum	Zeitungskonsum	Inernetkonsum
TV-Konsum	1.00	0.33	-0.20
Zeitungskonsum	0.33	1.00	-0.15
Inernetkonsum	-0.20	-0.15	1.00

Korrelationen in R: p-Matrix

```
#Laden von Daten und Paketen
library(knitr)
library(Hmisc)
load("daten_x.RData")
datensatz <- subset(daten_x, Bedingung)
#p-Matrix erstellen
pmatrix = data.frame(matrix$P, row.names = c("var1label", "var2label", " var3label "))
colnames(pmatrix) = c("var1label", "var2label", "var1label")
kable(round(pmatrix, 2), caption = "Überschrift")
```

Korrelationen in R: p-Matrix

```
#p-Matrix erstellen
matrix = rcorr(cbind(datensatz$tv_minuten, datensatz$tz_minuten, datensatz$www_minuten))
pmatrix = data.frame(matrix$P, row.names = c("TV-Konsum", "Zeitungskonsum", "Internetkonsum"))
colnames(pmatrix) = c("TV-Konsum", "Zeitungskonsum", "Internetkonsum")
kable(round(pmatrix, 2), caption = "p-Matrix")
```

p-Matrix

	TV-Konsum	Zeitungskonsum	Haushaltsgröße
TV-Konsum	NA	0	0
Zeitungskonsum	0	NA	0
Haushaltsgröße	0	0	NA

Korrelationen in r: n-Matrix

```
#n-Matrix erstellen
nmatrix = data.frame(matrix$n, row.names = c(" var1label ", "var2label", "var3label"))
colnames(nmatrix) = c("var1label", "var2label", "var3label")

#n-Matrix ausgeben
kable(nmatrix, caption = "Überschrift")
```

Korrelationen in R: n-Matrix

```
#n-Matrix erstellen
nmatrix = data.frame(matrix$n, row.names = c("TV-Konsum", "Zeitungskonsum", "Internetkonsum"))
colnames(nmatrix) = c("TV-Konsum", "Zeitungskonsum", "Internetkonsum")

#n-Matrix ausgeben
kable(nmatrix, caption = "n-Matrix")
```

n-Matrix

	TV-Konsum	Zeitungskonsum	Haushaltsgröße
TV-Konsum	483	473	481
Zeitungskonsum	473	476	474
Haushaltsgröße	481	474	486

Der Determinationskoeffizient R^2

<u>Der Determinationskoeffizient</u>

- Gibt an, wie groß der Anteil der Varianz ist, den die beiden Variablen teilen → wie gut passt das Modell?
- Sagt dennoch nichts über Kausalität aus
- Kann näherungsweise auch für Spearman‘s ρ und Kendall‘s τ angegeben werden
- Formel: $R^2 = r^2 = r * r$

Scatterplots

Scatterplots

- Sind Diagramme, die den Zusammenhang zwischen zwei numerischen Variablen darstellen
- Stellen jeden Fall als einen Punkt dar, der aus den Ausprägungen der beiden Variablen besteht: $(x_i \mid y_i)$
- Können durch eine Regressionslinie ergänzt werden → Veranschaulichung des Determinationskoeffizienten

Scatterplots in R

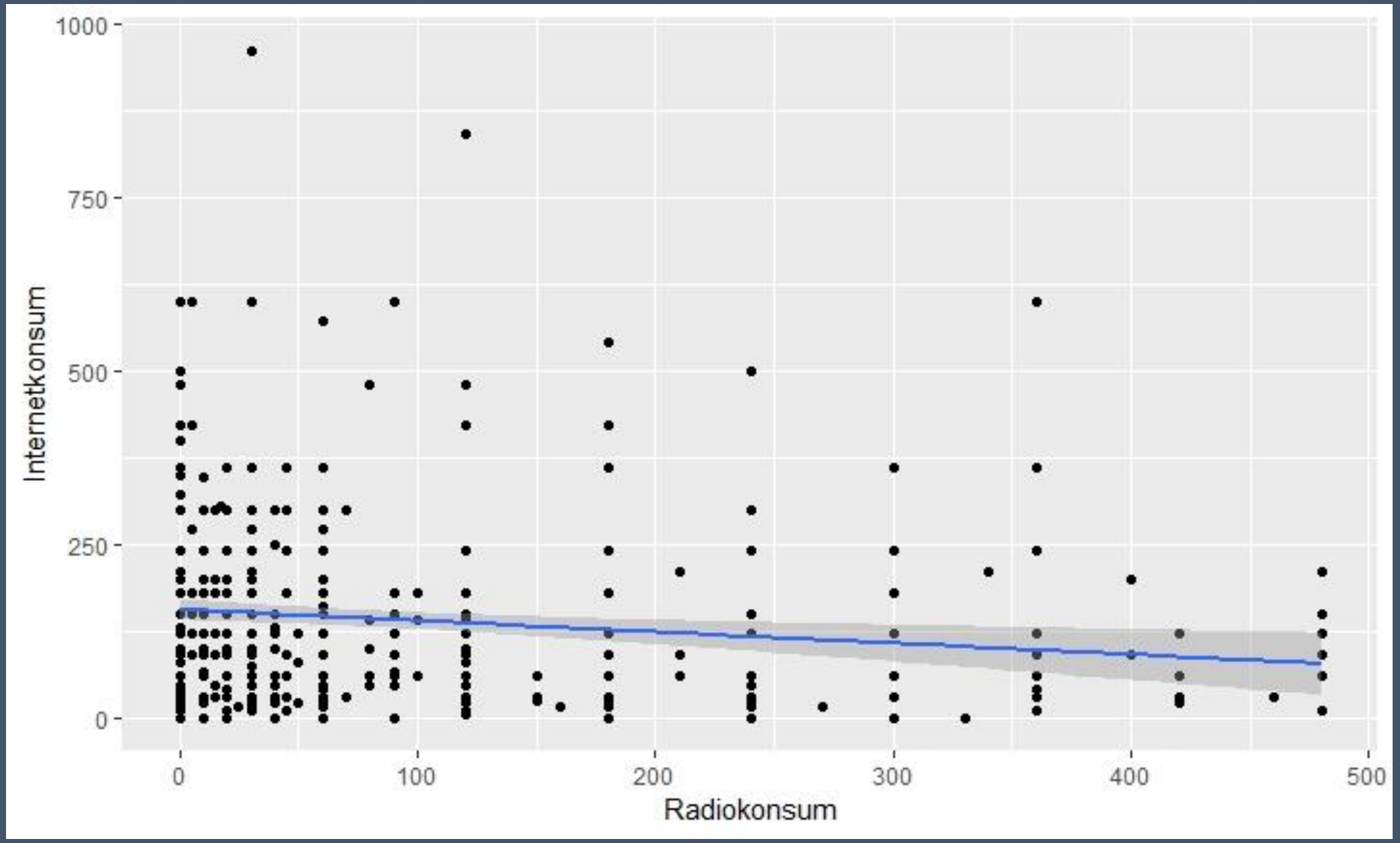
```
#Laden von Daten und Paketen
load("daten_x.RData")
library(ggplot2)
datensatz <- subset(daten_x, Bedingung)

#Scatterplot erstellen
ggplot(daten_x) + geom_point() +
geom_smooth(method = "lm") + aes(var1, var2) + labs
(x="var1label", y="var2label")
```

Scatterplots in R – Beispiel

```
#Laden von Daten und Paketen
load("daten_2019.RData")
library(ggplot2)
datensatz <- subset(daten_2019, radio_minuten<500 & www_minuten <500)

#Scatterplot erstellen
ggplot(datensatz) + geom_point() + geom_smooth(method = "lm") + aes(radio_minuten, www_minuten) + labs (x="Radiokonsum", y="Internetkonsum")
```


1000
750
500
250
0
Internetkonsum
0
100
200
300
400
500
Radiokonsum

Spearman's ρ

Spearman's ρ

- Ist ein Assoziationsmaß für ordinalskalierte Variablen
- Funktioniert genauso wie Pearson's r, nur arbeitet mit den Rangplätzen der Variablen
- Liefert kleinere Werte als Pearson's r
- benutzt man
  - Bei metrischen Variablen mit vielen Ausreißern
  - Bei natürlicherweise ordinalskalierten Variablen (bspw. Zustimmung: überhaupt nicht – voll und ganz)
- Formel: $\rho = \frac{cov(rp(x),(rp(y))}{s(rp(x))*s(rp(y))}$

Spearman's ρ – Rangplätze bestimmen

Fall i	1	2	3	4	5
Radiokonsum x_i	60	120	30	120	180
Zeitungskonsum y_i	15	0	10	20	15

Fall i	1	2	3	4	5
Radiokonsum x_i	4	3	5	2	1
Zeitungskonsum y_i	2	5	4	1	3

FRAGEN?

Selbststudium

- Forschungsbericht (EFB und ZFB)
  - r-, n- und p-Matrix für www_minuten, tz_minuten und soz_haushalt
  - Inhaltliche Sätze zu den Ergebnissen
- Anfangen: Wiederholung
- Fragen zur Wiederholungssitzung sammeln und bis Freitag, 29.06.2019, 23:59 Uhr per Mail zuschicken

Bis nächste Woche!