LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

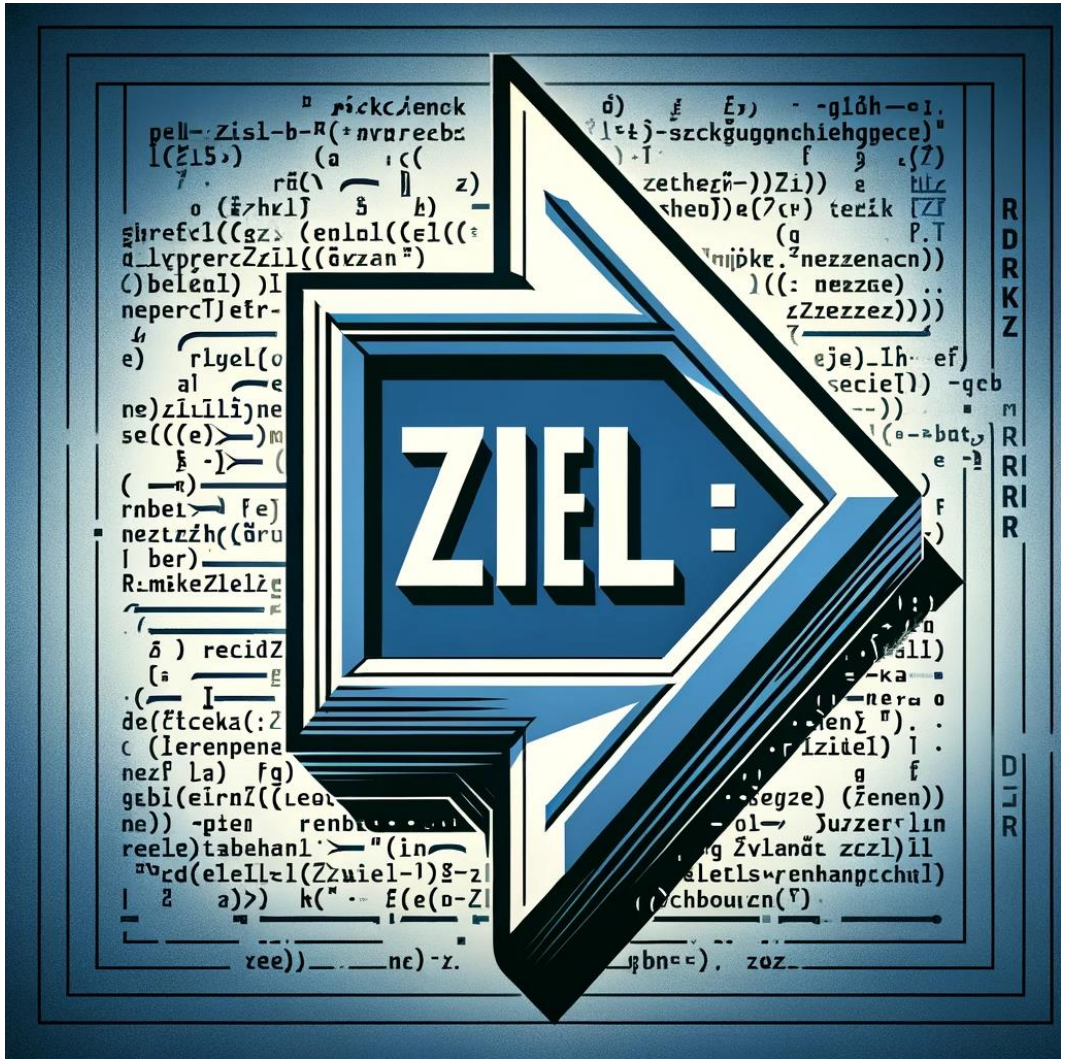
Datenanalyse

Sitzung 9: Regression

Institut für Kommunikationswissenschaft und Medienforschung
Ludwig-Maximilians-Universität München

- Korrelation ≠ Kausalität
- Verschiedene Korrelationen möglich
- Korrelation häufig verstanden als Pearson-Korrelation (r)
- Zusammenhangsmaß zwischen zwei metrischen Variablen
- Wertebereich zwischen -1 und +1
- Auch partieller Korrelationskoeffizient berechenbar

Ablauf der Sitzung

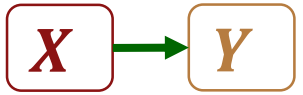


1. Kurze Wiederholung: Regression in der Theorie
2. Übungsblatt Aufgabe 1: Regression berechnen (Hausaufgabe)
3. Regression in R
4. Übungblatt Aufgabe 2: Einfache Regression in R
5. Aufgabe 3: Multiple Regression in R
6. Aufgabe 4: Standardisierte und unstandardisierte Koeffizienten in R
7. Aufgabe 5: Signifikanztest der Regressionskoeffizienten in R

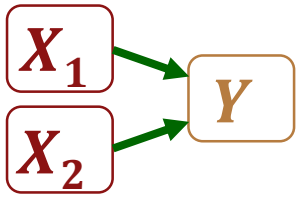
1. Kurze Wiederholung:
Regression

Grundlagen der Regression

- „Zurückführen“ einer abhängigen Variable Y („Kriterium“) auf eine unabhängige Variable X oder auf mehrere X_1, X_2 etc. („Prädiktoren“)
- Die Regression setzt (anders als die Korrelation) einen Zusammenhang voraus, der in eine bestimmte Richtung verläuft.
  - Einfache bzw. bivariate Regression: ein Prädiktor



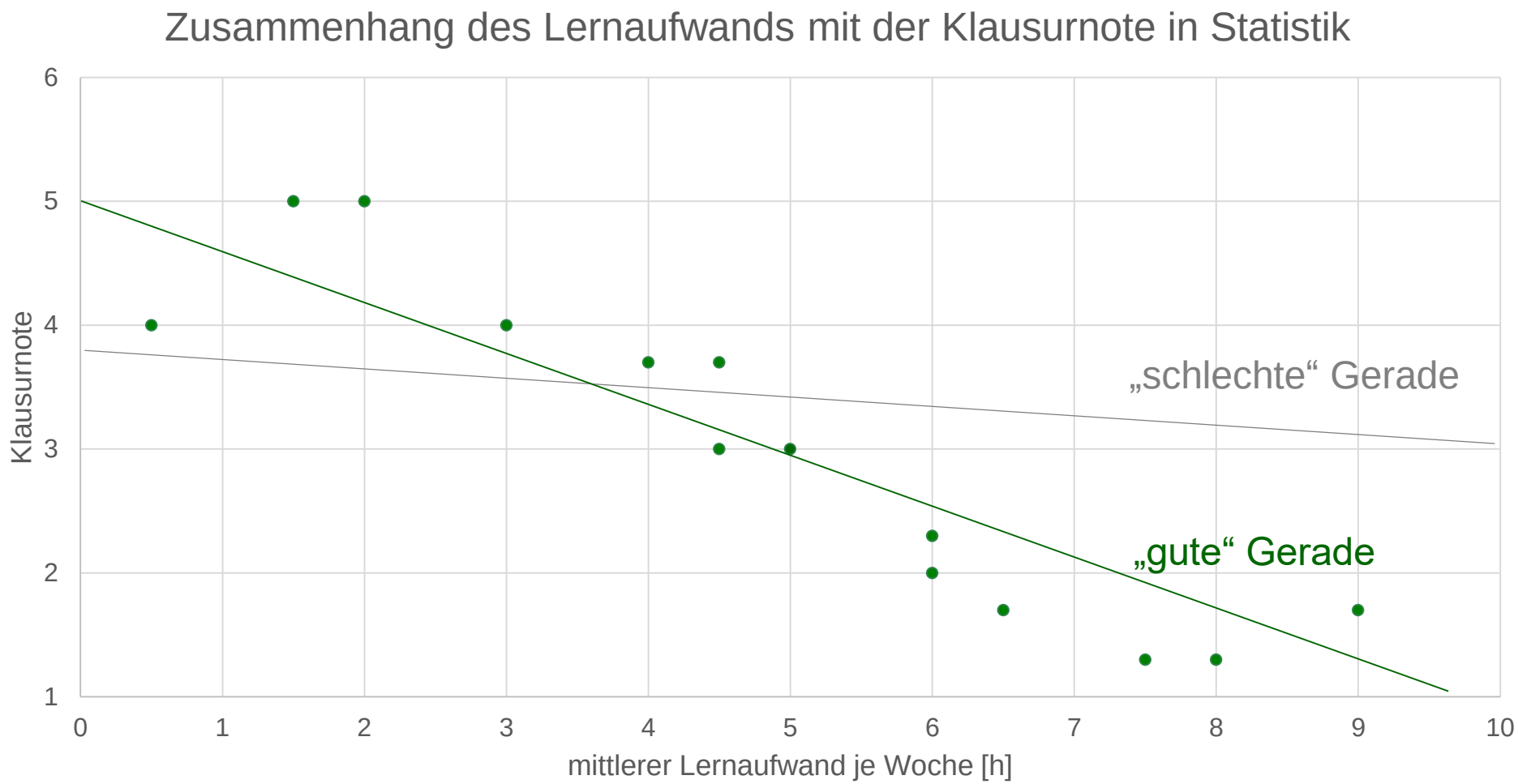
  - Multiple Regression: zwei und mehr Prädiktoren



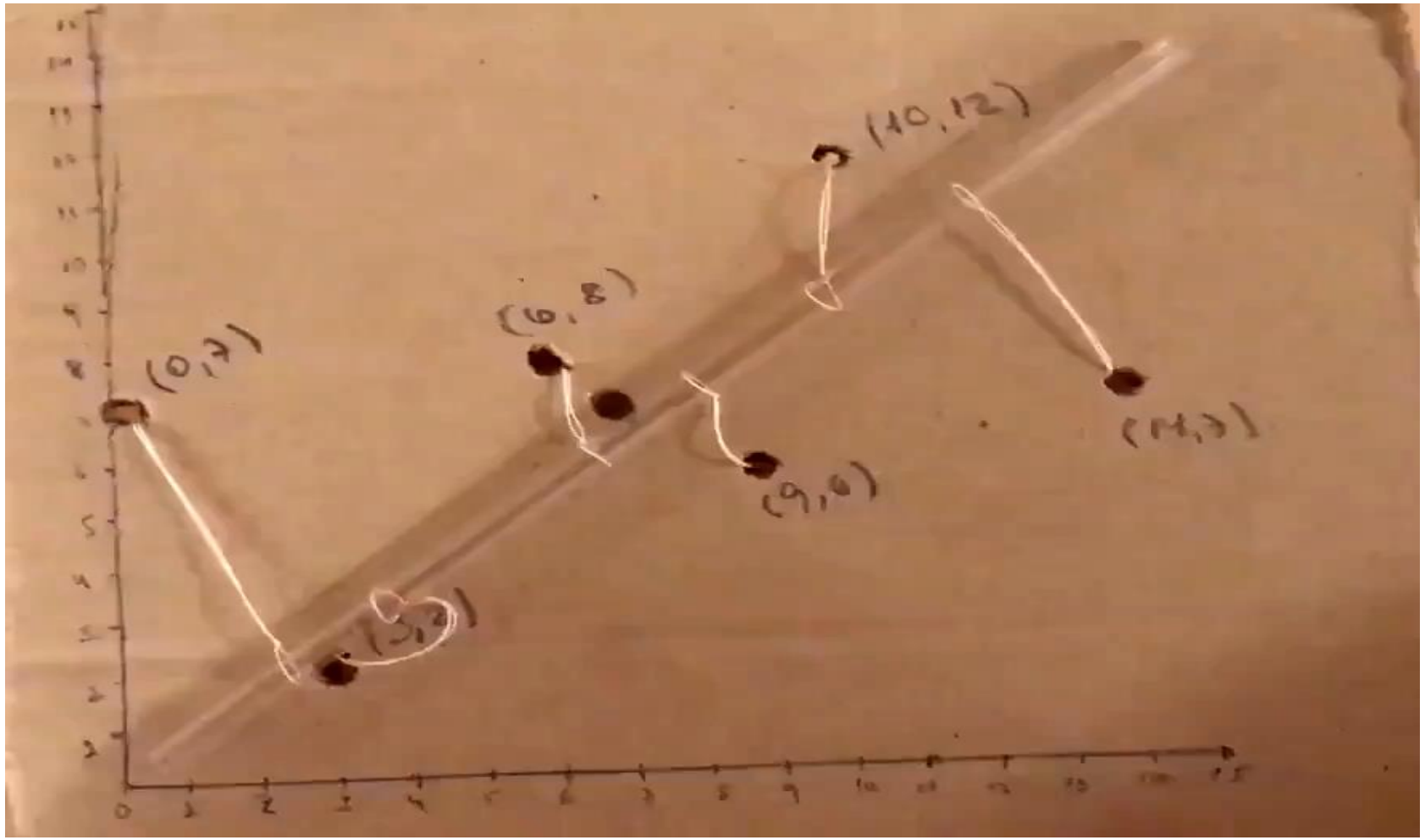
- Kausaler Zusammenhang kann auch durch Regression nicht belastbar getestet werden
- Ziel einfacher & multipler lineare Regression

→ Schätzung einer linearen Funktion, die die Punktewolke möglichst gut abbildet (Regressionsgerade)

Grundlagen der Regression – Regressionsgerade



Grundlagen der Regression – Methode der kleinsten Quadrate

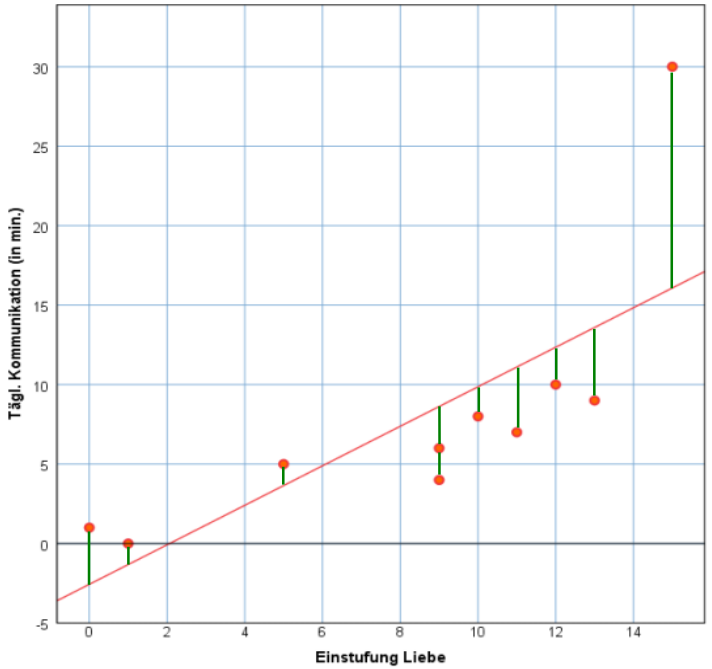


Grundlagen der Regression – Methode der kleinsten Quadrate

- Regressionsgerade wird so geschätzt, dass die Summe der quadrierten Residuen (= Vorhersagefehler) minimiert wird:

$$\sum_{i=1}^{N} e_i^2 \qquad \text{mit: } e_i = y_i - \hat{y}_i$$

→ bestmögliche Anpassung der Geraden an die beobachten Werte



Grundlagen der Regression – Regressionsfunktion

- Geradengleichung:

$$y = a + b \cdot x$$

- Einfache lineare Regressionsfunktion:

$$Y = a + b \cdot X + e$$

- Multiple lineare Regressionsfunktion für k Prädiktoren:

$$Y = a + b_1 \cdot X_1 + b_2 \cdot X_2 + \ \ldots \ + b_k \cdot X_k + e$$

X = beobachteter *X*-Wert
a = Schnittpunkt mit der *Y*-Achse (Achsenabschnitt bzw. „**Konstante**“)
b = **Steigung** der Regressionsgeraden
e = Residuum bzw. Abweichung des beobachteten Werts y_i vom vorhergesagten Wert $\hat{y}_i$ (e=„error“; Fehler)

Grundlagen der Regression – Regressionsfunktion

- Berechnung der Parameter

$$b = \frac{s_{XY}}{s_X^2} = \frac{Kovarianz}{Varianz\ X}$$

gibt an, in welche **Richtung** und um wie viele Einheiten (**Stärke**) sich Y verändert, wenn sich X um eine Einheit verändert

$$a = \overline{y} - b \cdot \overline{x}$$

gibt Y-Wert für X = 0 an (= Konstante)

Grundlagen der Regression – Regressionsfunktion

- Standardisierung des Regressionskoeffizienten *b*
  - Unstandardisierte Regressionskoeffizienten sind abhängig vom Wertebereich von X und Y
  - Standardisierung sinnvoll, wenn Regressionskoeffizienten hinsichtlich ihrer Einflussstärke miteinander verglichen werden sollen (z.B. bei der multiplen Regression)
  - Berechnung:

$$\beta = b \cdot \frac{s_X}{s_Y}$$

  - gibt Veränderung in Standardabweichung als Maßeinheit an

Anpassungsgüte & Varianzzerlegung (R^2)

- Die Anpassungsgüte:
  - Wird oft auch als „Bestimmtheitsmaß“ bezeichnet
  - Wertebereich: zwischen 0 und 1 (= zwischen 0 und 100%)
  - Bemisst den Anteil der Varianz von Y, der durch einen linearen Zusammenhang zwischen X und Y „erklärt“ werden kann

$$R^2 = \frac{s^2_{\hat{Y}}}{s^2_Y}$$

Anpassungsgüte & Varianzzerlegung (R^2)

- Vereinfachte Berechnung
  - Bei einer einfachen (!) linearen Regression lässt sich R^2 einfacher berechnen
  - über die Formel der **Korrelation**:

$$R^2 = r_{XY}^2 = \left(\frac{s_{XY}}{s_X \cdot s_Y}\right)^2 = \left(\frac{31{,}0}{4{,}99 \cdot 8{,}38}\right)^2 = 0{,}74^2 = 0{,}549$$

  - über den **standardisierten Regressionskoeffizienten**:

$$R^2 = \beta^2 = \left(b \cdot \frac{s_X}{s_Y}\right)^2 = \left(1{,}24 \cdot \frac{4{,}99}{8{,}38}\right)^2 = 0{,}74^2 = 0{,}549$$

Regressionskoeffizient b: Signifikanztest

- Formulierung der Hypothesen: zweiseitiges Problem

H_0: $b = 0$ $(bzw.\ t = 0)$

H_1: $b \neq 0$ $(bzw.\ t \neq 0)$

Regressionskoeffizient b: Signifikanztest

- Berechnung der Prüfgröße

$$t = \frac{b}{SE_b} \quad \text{mit} \quad df = N - 2$$

$$SE_b = \frac{s_e}{\sqrt{QS_X}}$$

$$s_e = \sqrt{\frac{\sum_{i=1}^{N} e_i^2}{N-2}}$$

$$QS_X = \sum_{i=1}^{N} (x_i - \overline{x})^2$$

- **Testentscheidung nach festgelegtem Signifikanzniveau:**
  - Kritischer Wert kann wie gewohnt unter Berücksichtigung des Signifikanzniveaus und der Freiheitsgrade aus der t-Tabelle gelesen werden
  - H_0 wird abgelehnt, wenn $|t| > t_{krit}$

Wichtige Take-Aways

- **Regression**: Berechnung eines Einflusses **einer oder mehrerer** metrischer Variablen (**Prädiktoren/unabhänhige Variablen**) auf **eine** einzige weitere Variable (**abhängige Variable**)
- **b:** **Un**standardisierter Regressionkoeffizient: Um wie viel verändert sich das Kriterium, **wenn der Prädiktor um eine Zähleinheit erhöht wird**.
- **beta:** Standardisierter Regressionkoeffizient: gibt Veränderung in **Standardabweichung als Maßeinheit** an
- **R^2:** Güte/Erklärungskraft/**Anpassungsgüte**/Varianzauflärung des Modells

2. Übungsblatt: Aufgabe 1

Übungsblatt: Aufgabe 1

a) Pierre (siehe Übungsblatt 8, Aufgabe 1) nimmt nun an, dass die TV-Nutzungsdauer einen linearen Einfluss auf das Vertrauen in andere Menschen hat. Welchen Wert für das Vertrauen würden wir für Personen prognostizieren, die täglich 6 Stunden fernsehen?

b) Wie gut kann das Regressionsmodell das interpersonale Vertrauen "erklären"?

ID	TV-Nutzung	Vertrauen
1	1	5
2	2	6
3	3	4

Übungsblatt: Aufgabe 1

c) Können wir annehmen, dass ein Einfluss der TV-Nutzungsdauer auch in der Grundgesamtheit vorhanden ist? (Signifikanzniveau α=0,05. Die Residuen sind normalverteilt und unkorreliert, ihre Varianz ist unabhängig von X. Ausreißer liegen nicht vor.)

d) Pierres Cousine Pamela führt dieselbe Studie mit einer anderen Stichprobe durch, erfasst das Vertrauen aber anders als Pierre auf einer 7-stufigen Ratingskala. Sie ermittelt eine Kovarianz von s_{XY}= – 0,3, sowie Standardabweichungen für das Vertrauen von s_Y = 0,75 und für die TV-Nutzungsdauer erneut von s_X = 1,0. Hat Pamela oder hat Pierre einen stärkeren Einfluss der TV-Nutzungsdauer (für ihre Stichprobe!) ermittelt?

Lösung für Aufgabe 1a

- Pierre (siehe Übung 8) nimmt nun an, dass die TV-Nutzungsdauer einen linearen Einfluss auf das Vertrauen in andere Menschen hat. Welchen Wert für das Vertrauen würden wir für Personen prognostizieren, die täglich 6 Stunden fernsehen?

→ Für Prognose **Schätzung Regressionsgleichung** notwendig

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

ID	TV-Nutzung	Vertrauen
1	1	5
2	2	6
3	3	4

Lösung für Aufgabe 1a

- Berechnung b

$$b = \frac{s_{XY}}{s_X^2} = \frac{Kovarianz}{Varianz\ X}$$

→ gibt an, in welche **Richtung** und um wie viele Einheiten (**Stärke**) sich Y (**Vertrauen**) verändert, wenn sich X (**TV-Nutzung**) um eine Einheit verändert (= unstandardisierter Regressionskoeffizient)

- Berechnung a

$$a = \overline{y} - b \cdot \overline{x}$$

→ gibt Y-Wert für X = 0 an (= Konstante)

Lösung für Aufgabe 1a

- **Berechnung *b***

Kovarianz s_{XY} (siehe Übung 8)

ID	TV-Nutzung (x)	Vertrauen (y)	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$
1	1	5	– 1,0	0,0	0,0
2	2	6	0,0	1,0	0,0
3	3	4	1,0	– 1,0	– 1,0
	$\bar{x} = 2$	$\bar{y} = 5$			
	$s_X = 1$	$s_Y = 1$			$\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}) = -1$

$$\boldsymbol{s_{XY}} = \frac{1}{3-1} \cdot -1 = \mathbf{-0,50}$$

Lösung für Aufgabe 1a

- **Berechnung *b***

Varianz Prädiktor s_X^2

ID	TV-Nutzung (x)	Vertrauen (y)	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$
1	1	5	– 1,0	0,0	0,0
2	2	6	0,0	1,0	0,0
3	3	4	1,0	– 1,0	– 1,0
	$\bar{x} = 2$	$\bar{y} = 5$			
	$s_X = 1$	$s_Y = 1$			$\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}) = -1$

$$s_X^2 = 1^2 = 1$$

Lösung für Aufgabe 1a

- **Berechnung *b***

$$\boldsymbol{b} = \frac{s_{XY}}{s_X^2} = \frac{-0{,}50}{1} = \mathbf{-0{,}50}$$

Lösung für Aufgabe 1a

- **Berechnung *a***
- Mittelwerte $\overline{y}$ & $\overline{x}$

ID	TV-Nutzung (x)	Vertrauen (y)	$(x_i - \overline{x})$	$(y_i - \overline{y})$	$(x_i - \overline{x})(y_i - \overline{y})$
1	1	5	– 1,0	0,0	0,0
2	2	6	0,0	1,0	0,0
3	3	4	1,0	– 1,0	– 1,0
	$\bar{x} = 2$	$\bar{y} = 5$			
	$s_X = 1$	$s_Y = 1$			$\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}) = -1$

Lösung für Aufgabe 1a

- **Berechnung *a***

$$\boldsymbol{a} = \overline{y} - b \cdot \overline{x} = 5 - (-0.50) \cdot 2 = 5 + 1 = \boldsymbol{6}$$

Lösung für Aufgabe 1a

- **Geschätzte Regressionsgleichung**

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

$$\hat{\boldsymbol{y}} = \mathbf{6} + (-\mathbf{0.50} \cdot \boldsymbol{x})$$

$$\hat{\boldsymbol{y}} = \mathbf{6} - \mathbf{0.50} \cdot \boldsymbol{x}$$

- Gemäß *b*: Für jede zusätzliche Stunde TV-Nutzung (= 1 Einheit X) reduziert sich das interpersonale Vertrauen geschätzt um 0.5 Skalenpunkte.
- Gemäß *a*: Für Personen, die nicht Fernsehen (x = 0), schätzen wir das interpersonale Vertrauen auf 6 (von 10) Skalenpunkten.

Lösung für Aufgabe 1a

- …Welchen Wert für das Vertrauen würden wir für Personen prognostizieren, die täglich 6 Stunden fernsehen?

$$\hat{y} = 6 - 0,50 \cdot x = 6 - 0,50 \cdot 6 = 3$$

Wir würden einen Wert von **3** Skalenpunkten für interpersonales Vertrauen prognostizieren

Lösung für Aufgabe 1b

- Wie gut kann das Regressionsmodell Vertrauen „erklären“?

Bestimmtheitsmaß R^2

Lösung für Aufgabe 1b

- Bestimmtheitsmaß ***R²*** berechnen

Variante 1 → über Varianzen

ID	TV-Nutzung (x)	Vertrauen (y)	vorhergesagte Werte Vertrauen ($\hat{y}$)
1	1	5	5.5
2	2	6	5
3	3	4	4.5
	$s_X^2 = 1$	$s_Y^2 = 1$	$s_{\hat{y}}^2 = 0.25$

mit $\hat{y} = 6 - 0.50 \cdot x$

$$s^2 = \frac{1}{N-1} \sum_{i=1}^{N} (\hat{y}_i - \bar{y})^2$$

$$R^2 = \frac{s_{\hat{Y}}^2}{s_Y^2} = \frac{0{,}25}{1} = 0{,}25$$

Regressionsmodell (d.h. hier: TV-Nutzung) kann 25% der Varianz in interpersonalem Vertrauen „erklären“

Lösung für Aufgabe 1c

- Können wir annehmen, dass ein Einfluss der TV-Nutzungsdauer auch in der Grundgesamtheit vorhanden ist?
  - Signifikanzniveau α=0,05
  - Die Residuen sind normalverteilt und unkorreliert, ihre Varianz ist unabhängig von X
  - Ausreißer liegen nicht vor
    (→ bedeutet: Voraussetzungen für den Signifikanztest für *b* sind erfüllt)

Signifikanztest für *b*

Lösung für Aufgabe 1c

- **Prüfgröße t berechnen**

$$t = \frac{b}{SE_b}$$

ID	TV-Nutzung (x)	Vertrauen (y)	$\hat{y}$	$e = y - \hat{y}$	$(x_i - \overline{x})^2$
1	1	5	5.5	-0,50	
2	2	6	5	1	
3	3	4	4.5	-0,50	
	$s_X^2 = 1$	$s_Y^2 = 1$	$s_{\hat{y}}^2 = 0.25$		

$$SE_b = \frac{s_e}{\sqrt{QS_X}}$$

Berechnung Standardschätzfehler *(s_e):*

$$s_e = \sqrt{\frac{\sum_{i=1}^{N} e_i^2}{N-2}} = \sqrt{\frac{-0{,}50^2 + 1^2 + (-0{,}50^2)}{3-2}} = \sqrt{\frac{0{,}25 + 1 + 0{,}25}{1}} = 1{,}22$$

Lösung für Aufgabe 1c

- **Prüfgröße t berechnen**

$$t = \frac{b}{SE_b}$$

ID	TV-Nutzung (x)	Vertrauen (y)	$\hat{y}$	$e = y - \hat{y}$	$(x_i - \overline{x})^2$
1	1	5	5.5	-0,50	1
2	2	6	5	1	0
3	3	4	4.5	-0,50	1
	$s_X^2 = 1$	$s_Y^2 = 1$	$s_{\hat{y}}^2 = 0.25$		$QS_X = 2$

mit $\overline{x} = 2$

$$SE_b = \frac{s_e}{\sqrt{QS_X}}$$

Berechnung Standardfehler Regressionskoeffizient *b* (SE_b):

$$SE_b = \frac{s_e}{\sqrt{QS_X}} = \frac{1{,}22}{\sqrt{2}} = \mathbf{0,87}$$

Lösung für Aufgabe 1c

- **Prüfgröße t berechnen (Fortsetzung)**

$$t = \frac{b}{SE_b} = \frac{-0{,}50}{0{,}87} = -0{,}57$$

- **Kritischer Wert für Prüfgröße *t***

$df = N - 2 = 1$, Signifikanzniveau 5% (α=0,05), zweiseitig

$t_{\text{krit}} = 12{,}71$

	einseitig		zweiseitig	
df	α = 5%	α = 1%	α = 5%	α = 1%
1	6,31	31,82	12,71	63,66

- **Testentscheidung**

$|t| = 0{,}57 < t_{\text{krit}}$ → H_0 wird beibehalten!

Lösung für Aufgabe 1c

- Können wir annehmen, dass ein Einfluss der TV-Nutzungsdauer auch in der Grundgesamtheit vorhanden ist?
  (Signifikanzniveau α=0,05. Die Residuen sind normalverteilt und unkorreliert, ihre Varianz ist unabhängig von X. Ausreißer liegen nicht vor.)

Nein, wir können nicht annehmen, dass die TV-Nutzungsdauer das interpersonale Vertrauen in der Grundgesamtheit beeinflusst, weil sich der geschätzte Regressionskoeffizient nicht statistisch signifikant von 0 unterscheidet ($b = – 0{,}50$; $p > 0{,}05$).

Lösung für Aufgabe 1d

- Pierres Cousine Pamela führt dieselbe Studie mit einer anderen Stichprobe (*N*=10) durch, aber erfasst das Vertrauen auf einer 7-stufigen Ratingskala. Sie ermittelt eine Kovarianz von s_{XY}= – 0,30, sowie Standardabweichungen für das Vertrauen von s_y= 0,75 und für die TV-Nutzungsdauer von s_x = 1,0. Hat damit Pamela oder Pierre einen stärkeren Einfluss der TV-Nutzungsdauer (für ihre Stichprobe!) ermittelt?

→ **Berechnung des standardisierten Regressionskoeffizienten β**

Lösung für Aufgabe 1d

- **Berechnung β für Pierres Studie**

$$\beta_{Pi} = b \cdot \frac{s_X}{s_Y} = -0{,}50 \cdot \frac{1}{1} = -0{,}50$$

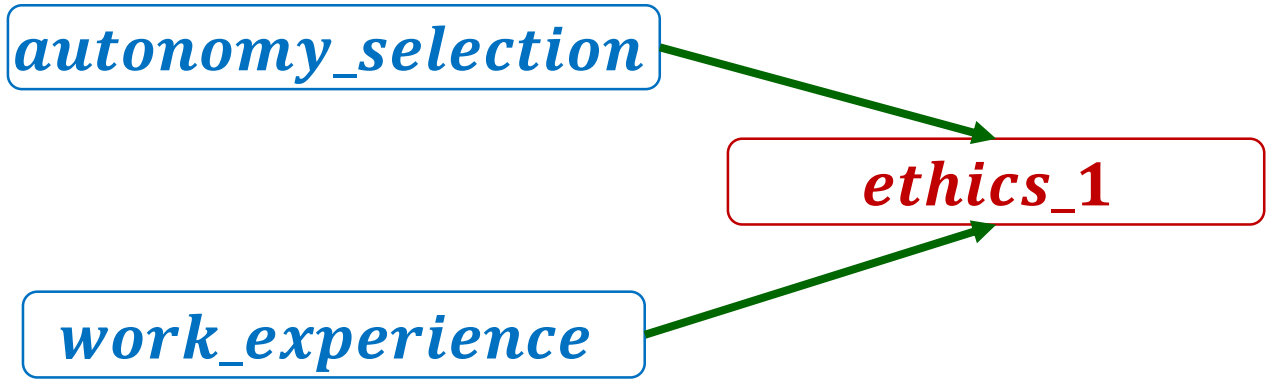
- **Berechnung β für Pamelas Studie**

$$\beta_{Pa} = b \cdot \frac{s_X}{s_Y} = \frac{s_{XY}}{s_X^2} \cdot \frac{s_X}{s_Y} = \frac{s_{XY}}{s_X \cdot s_Y} = \frac{-0{,}30}{1 \cdot 0{,}75} = -0{,}40$$

→ Pierre hat (deskriptiv!) einen stärkeren Einfluss der TV-Nutzungsdauer auf das interpersonale Vertrauen beobachtet

3. Regression in R

Regression in R mit…



ethics_1

Agreement with statement "Journalists should always adhere to codes of professional ethics, regardless of situation and context", scale from 1 (*strongly disagree*) to 5 (*strongly agree*) (*reverse-coded!*)

autonomy_selection

Autonomy in news story selection, scale from 1 (*no freedom at all*) to 5 (*complete freedom*)

work_experience

Work experience as a journalist in years

Hypothese: Je höher die wahrgenommene Autonomie von Journalist:innen bei der Auswahl von Nachrichten, desto geringer ist auch ihre Zustimmung zur Aussage "Journalisten sollten sich unabhängig von der Situation und dem Kontext immer an berufsethische Regeln halten".

Kontrollvariable: Arbeitserfahrung in Jahren

Regression in R

Code - Einfache lineare Regression

```
 9  WoJ %>%
10    regress(
11      ethics_1,
12      autonomy_selection)
```

Abhängige Variable

Unbhängige Variable

Output – Einfache lineare Regression

Variable	B	StdErr	beta	t	p
* <chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1 (Intercept)	1.96	0.127	NA	15.4	4.57e-49
2 autonomy_selection	-0.0852	0.0321	-0.0766	-2.66	7.98e- 3
# F(1, 1195) = 7.061023, p = 0.007983, R-square = 0.005874					

F-Test & R-Quadrat der gesamten Regression

Signifikanztests der einzelnen Regressions-koeffizienten

Unstandardisierter (b) & standardisierter (beta) Regressionskoeffizient

---

Code – Multiple lineare Regression

```
15  WoJ %>% regress(
16    ethics_1,
17    autonomy_selection,
18    work_experience
19  )
```

Output – Multiple lineare Regression

Variable	B	StdErr	beta	t	p
* <chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1 (Intercept)	2.03	0.129	NA	15.8	5.32e-51
2 autonomy_selection	-0.0692	0.0325	-0.0624	-2.13	3.32e- 2
3 work_experience	-0.00762	0.00239	-0.0934	-3.19	1.46e- 3
# F(2, 1181) = 8.677001, p = 0.000182, R-square = 0.014482					

Regression in empirischer Forschung berichten

Interpretation einfach lineare Regression:

„Der **Prädiktor,** nämlich die **wahrgenommene Autonomie** bei der Auswahl von Nachrichten beeinflusst die Zustimmung zur Aussage „Journalisten sollten sich unabhängig von der Situation und dem Kontext immer an berufsethische Regeln halten“ kaum. Der Einfluss ist zwar **statistisch signifikant**, jedoch so schwach, dass er in der Praxis keine Relevanz hat ($R^2 < 0{,}01$). Nur etwa **0,6% der Varianz** der abhängigen Variable lassen sich auf den Prädiktor zurückführen (df = 1195; b = -,069; β = -,08; t = -2.66; p < ,01; R^2 = ,006).“

4. Übungsblatt: Aufgabe 2

Übungsblatt: Aufgabe 2

Geben Sie die Daten aus Pierres Studie in R ein und führen Sie die in Regressionsanalyse in R durch.

- Welche Regressionsgleichung wird geschätzt?
- Wie können Sie aus dem R-Output die Regressionsgleichung ablesen?
- Beschreiben Sie einen möglichen Einfluss der TV-Nutzungsdauer auf das interpersonale Vertrauen ausführlich (Signifikanzniveau α=0,05)!

ID	TV-Nutzung	Vertrauen
1	1	5
2	2	6
3	3	4

Lösung: Aufgabe 2

Regressionsgleichung als R Output

Code

```
32  ergebnis_regression <- data %>%
33    regress(
34      vertrauen,
35      tv_nutzung
36    )
37
38  ergebnis_regression
```

Output

```
# A tibble: 2 × 6
  Variable          B StdErr   beta      t     p
* <chr>         <dbl>  <dbl>  <dbl>  <dbl> <dbl>
1 (Intercept)   6      1.87   NA      3.21  0.192
2 tv_nutzung   -0.500  0.866 -0.500 -0.577 0.667
# F(1, 1) = 0.333333, p = 0.666667, R-square = 0.250000
```

→ Geschätzte Regressionsgleichung: $\hat{y} = 6{,}0 - 0{,}50 \cdot x$

→ Gleich der händisch berechneten Regressionsgleichung

Lösung: Aufgabe 2

Beschreibung des Einflusses des Prädiktors

Output

```
# A tibble: 2 × 6
  Variable          B StdErr   beta      t     p
* <chr>         <dbl>  <dbl>  <dbl>  <dbl> <dbl>
1 (Intercept)  6      1.87   NA      3.21  0.192
2 tv_nutzung  -0.500  0.866 -0.500 -0.577 0.667
# F(1, 1) = 0.333333, p = 0.666667, R-square = 0.250000
```

- *b* = – 0,50, p = ,667
- Für jede zusätzliche Stunde TV-Nutzung ***sinkt*** das interpersonale Vertrauen geschätzt um **0.5 Skalenpunkte** (negativer Einfluss).
- Dieser Einfluss ist aber **nicht** signifikant, es kann also für die Grundgesamtheit nicht angenommen werden, dass der Prädiktor einen Erklärungsbeitrag leistet (b ≠ 0).

5. Aufgabe 3

Aufgabe 3

Korrelation & Regression

Laden Sie aus dem **tidycomm-package** den Datensatz **incvlcomments.**
Nutzen Sie die Hilfe-Funktion um einen Eindruck der Daten zu bekommen. (`?incvlcomments`)

Incivil Comments Data

Description

A dataset of a preregistered factorial survey experiment with a nationally representative sample of 964 German online users. Participants were presented with manipulated user comments that included statements associated with incivil discourse (such as profanity and attacks on arguments) and intolerant discourse (such as offensive stereotyping and violent threats). Participants rated the comments, e.g. offensiveness, harm to society, and their intention to delete the comment containing the statement.

offensiveness
: Rate statement whether the comment is being perceived as offensive & hostile (Scale from 1 to 7)

adequacy
: Rate statement whether the comment is being perceived as necessary & accurate (Scale from 1 to 7)

harm_to_society
: Rate statement whether the comment is being perceived as harmful to society (Scale from 1 to 7)

deletion_intention
: Whether the participant wants to delete the comment

similarity_poster
: How similar the participant feels to the person who created the post (Scale from 1 to 7)

similarity_group
: How similar the participant feels to the group of people criticized in the post (Scale from 1 to 7)

attitude_gender
: Rate agreement with statements on gender policies (Scale from 1 to 7)

attitude_abortion
: Rate agreement with statements on abortion (Scale from 1 to 7)

attitude_migration
: Rate agreement with statements on migration (Scale from 1 to 7)

attitude_climate
: Rate agreement with statements on climate change (Scale from 1 to 7)

left_right_placement
: Placement on a political spectrum from left to right (Scale from 1 to 9)

freedom_of_speech
: Rate agreement with statements about the freedom of speech and expression (Scale from 1 to 7)

Übungsblatt: Aufgabe 3

Korrelation & Regression

a) Wie stark hängen die Einstellung zu Abtreibung und die Einstellung zum Klimawandel zusammen?
b) Wie stark hängen die Einstellung zu Migrationsfragen und die Einstellung die Links-rechts Selbsteinschätzung zusammen?

c) Wie verändert sich die wahrgenommene Feindseligkeit eines Kommentars in Abhängigkeit des wahrgenommenen gesellschaftlichen Schadens, den dieser Kommentar anrichten kann?
d) Wie verändert sich die die Links-rechts Selbsteinschätzung in Abhängigkeit der Einstellung zu Abtreibung, des Alters und der Einstellung zu Gender-Fragen?

e) Visualisieren Sie die Ergebnisse der Aufgaben c) und d) in Streudiagrammen.

offensiveness
: Rate statement whether the comment is being perceived as offensive & hostile (Scale from 1 to 7)

adequacy
: Rate statement whether the comment is being perceived as necessary & accurate (Scale from 1 to 7)

harm_to_society
: Rate statement whether the comment is being perceived as harmful to society (Scale from 1 to 7)

deletion_intention
: Whether the participant wants to delete the comment

similarity_poster
: How similar the participant feels to the person who created the post (Scale from 1 to 7)

similarity_group
: How similar the participant feels to the group of people criticized in the post (Scale from 1 to 7)

attitude_gender
: Rate agreement with statements on gender policies (Scale from 1 to 7)

attitude_abortion
: Rate agreement with statements on abortion (Scale from 1 to 7)

attitude_migration
: Rate agreement with statements on migration (Scale from 1 to 7)

attitude_climate
: Rate agreement with statements on climate change (Scale from 1 to 7)

left_right_placement
: Placement on a political spectrum from left to right (Scale from 1 to 9)

freedom_of_speech
: Rate agreement with statements about the freedom of speech and expression (Scale from 1 to 7)

Lösung: Aufgabe 3 a) + b)

Datensatz: Inzivile Kommentare

Signifikanzniveau α=0,05

```
40  # Uebung 3)
41  incvlcomments <- incvlcomments
42  ?incvlcomments
43
44  # Uebung 3a + 3b)
45  ergebnis_3a <- incvlcomments %>% correlate(
46    attitude_abortion,
47    attitude_climate,
48    attitude_migration,
49    left_right_placement,
50    method = "pearson",
51  ) %>% to_correlation_matrix()
```

Beispiel-Interpretation Korrelation:
„Zwischen der Einstellung zu Schwangerschaftsabbrüchen und dem Klimawandel besteht ein geringer positiver Zusammenhang.“

"Zwischen der Einstellung zu Migration und der Einstellung zum Klimawandel besteht ein mittlerer positiver Zusammenhang."

"Zwischen der der Einstellung zu Migration und der link-rechts Selbsteinschätzung besteht ein mittlerer negativer Zusammenhang."

```
# A tibble: 4 × 5
  r                   attitude_abortion attitude_climate attitude_migration left_right_placement
* <chr>                           <dbl>            <dbl>              <dbl>                <dbl>
1 attitude_abortion               1                0.105              0.0929              -0.159
2 attitude_climate                0.105            1                  0.405               -0.314
3 attitude_migration              0.0929           0.405              1                   -0.274
4 left_right_placement           -0.159           -0.314             -0.274                1
```

Lösung: Aufgabe 3 c)

Datensatz: Inzivile Kommentare

Signifikanzniveau α=0,05

```
36  # Uebung 3c)
37  ergebnis_3c <- incvlcomments %>% regress(
38    offensiveness,
39    harm_to_society,
40  )
41
42  ergebnis_3c
```

```
# A tibble: 2 × 6
  Variable            B StdErr   beta     t         p
* <chr>           <dbl>  <dbl>  <dbl> <dbl>     <dbl>
1 (Intercept)     1.19  0.0526 NA      22.5 1.30e-105
2 harm_to_society 0.771 0.0112  0.742  68.7 0
# F(1, 3854) = 4713.567306, p = 0.000000, R-square = 0.550164
```

Beispiel-Interpretation einfach Regression:
"Für jede Einheit auf der Skala der Variable "harm to society" steigt die Variable "offensiveness" um 0,771 Skalenpunkte."

"Die unabhängige Variable erklärt 55 Prozent der Varianz der abhängigen Variable. Die Schätzung ist statistisch höchst signifikant."

"Für jede Standardabweichung der Variable "harm to society" steigt die Variable "offensiveness" um 0,742 Standardabweichungen."

Lösung: Aufgabe 3 d)

Datensatz: Inzivile Kommentare

Signifikanzniveau α=0,05

```
44  # Uebung 3d)
45  ergebnis_3d <- incvlcomments %>% regress(
46    left_right_placement,
47    attitude_abortion,
48    age,
49    attitude_gender
50  )
51
52  ergebnis_3d
```

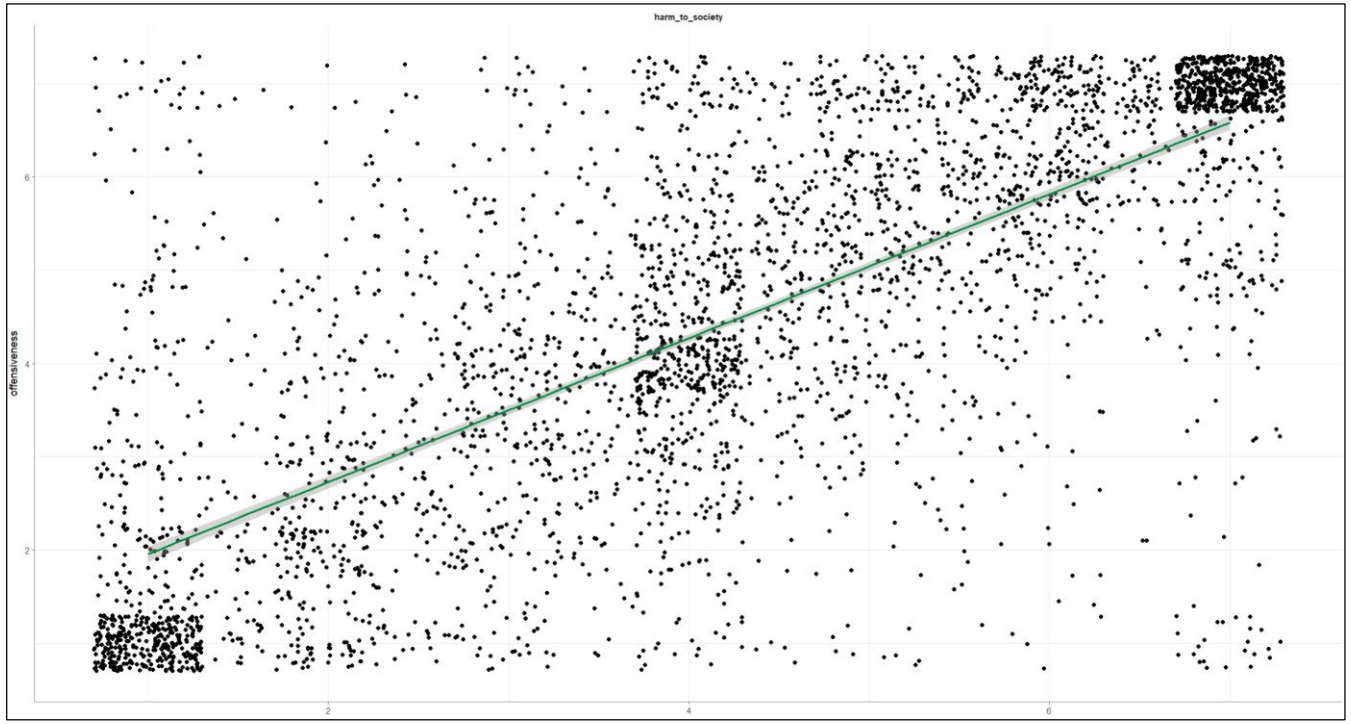
```
# A tibble: 4 × 6
  Variable                B  StdErr    beta      t        p
* <chr>                <dbl>   <dbl>   <dbl>  <dbl>    <dbl>
1 (Intercept)         6.56    0.133   NA       49.1  0
2 attitude_abortion  -0.151   0.0166  -0.143   -9.11 1.32e-19
3 age                -0.00697 0.00176 -0.0623  -3.96 7.68e- 5
4 attitude_gender    -0.151   0.0126  -0.189  -12.0  1.62e-32
# F(3, 3852) = 84.560481, p = 0.000000, R-square = 0.061788
```

Beispiel-Interpretation multiple Regression:
"Unter Berücksichtigung der Variablen Alter sowie der Einstellung zu gendergerechter Sprache gilt: Für jede Einheit auf der Skala der Variable "Einstellung zu Abtreibung" sinkt die abhängige Variable der links-rechts Selbsteinschätzung um 0,151 Skalenpunkte."

Lösung: Aufgabe 3 e)

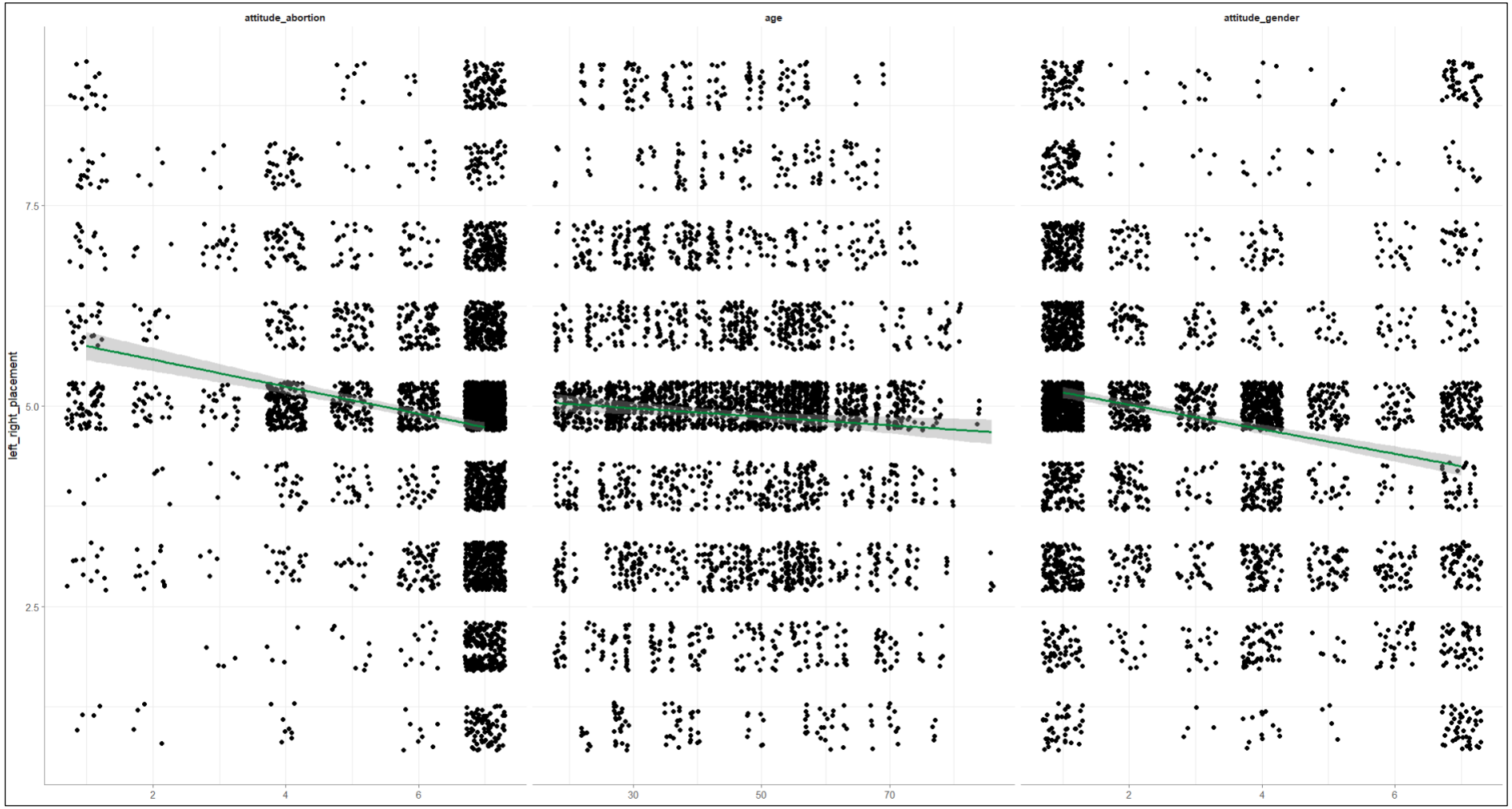
Datensatz: Inzivile Kommentare

```
55  # Uebung 3e)
56
57  ergebnis_3e1 <- ergebnis_3c %>% visualize()
58  ergebnis_3e1
59
60  ergebnis_3e2 <- ergebnis_3d %>% visualize()
61  ergebnis_3e2
```



Lösung: Aufgabe 3 e)

Datensatz: Inzivile Kommentare



Ergänzung: Voraussetzungsprüfung - Multikollinearität

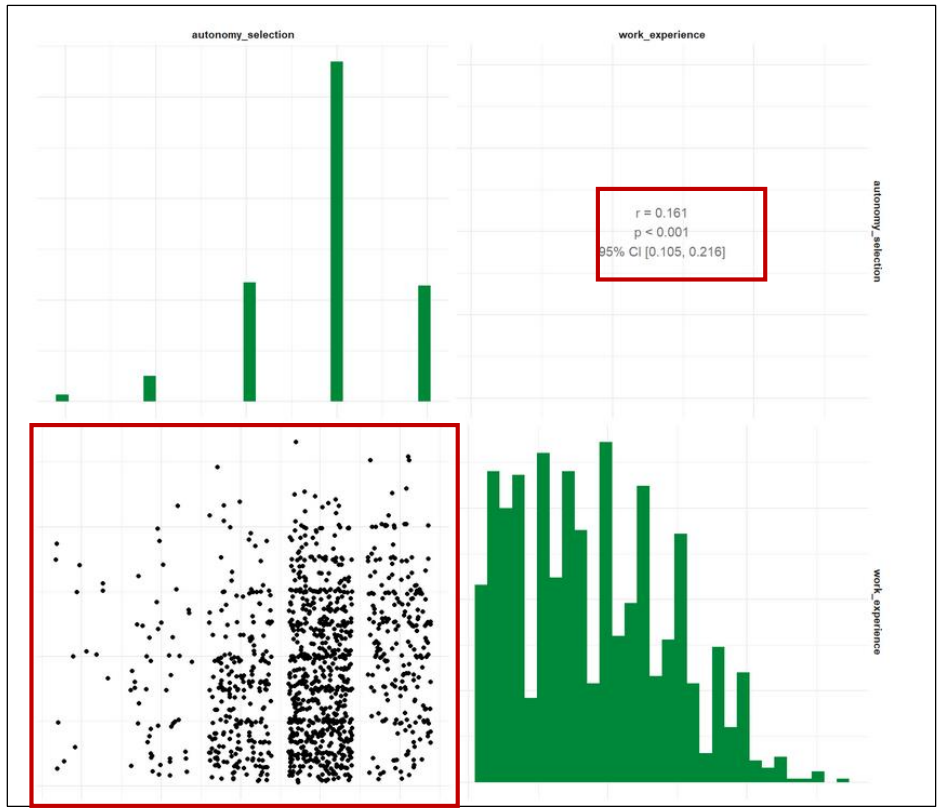
```
WoJ %>% regress(ethics_1, autonomy_selection, work_experience,
                check_multicollinearity = TRUE, #Variance Inflation Factor (VIF)
                )
```

```
# A tibble: 3 × 8
  Variable                B  StdErr    beta     t        p   VIF tolerance
* <chr>                <dbl>   <dbl>   <dbl> <dbl>    <dbl> <dbl>     <dbl>
1 (Intercept)       2.03    0.129   NA      15.8  5.32e-51 NA       NA
2 autonomy_select… -0.0692  0.0325  -0.0624 -2.13 3.32e- 2  1.03     0.974
3 work_experience  -0.00762 0.00239 -0.0934 -3.19 1.46e- 3  1.03     0.974
# F(2, 1181) = 8.677001, p = 0.000182, R-square = 0.014482
```

VIF und **tolerance** sollten zwischen 0,25 und 4 liegen

Ergänzung: Voraussetzungsprüfung - Multikollinearität

```
WoJ %>%
  regress(ethics_1, autonomy_selection, work_experience) %>%
  visualize(which = "correlogram")
```



Streudiagramm und **Korrelation** der beiden Prädiktoren sollten **keinen zu starken Zusammenhang** andeuten.

Vielen Dank für Ihre Aufmerksamkeit!