Multiple lineare Regression

Methoden der empirischen Kommunikations- und Medienforschung

Marko Bachl
Freie Universität Berlin

Fragen zur Übung?

AGENDA

1. Grundlagen der multiplen Regression
2. Kategorielle Prädiktoren
3. “Große” Regressionsmodelle
4. Annahmen und ihre Überprüfung
5. Übungen

DATEN DER HEUTIGEN SITZUNG

POLITICAL COMMUNICATION
2021, VOL. 38, NO. 4, 407–425
https://doi.org/10.1080/10584609.2020.1784328

Routledge
Taylor & Francis Group

Why Don't We Learn from Social Media? Studying Effects of and Mechanisms behind Social Media News Use on General Surveillance Political Knowledge

Patrick F. A. van Erkel and Peter Van Aelst

Department of Political Science, University of Antwerp, Antwerp, Belgium

ABSTRACT

Does exposure to news affect what people know about politics? This old question attracted new scholarly interest as the political information environment is changing rapidly. In particular, since citizens have new channels at their disposal, such as Twitter and Facebook, which increasingly complement or even replace traditional channels of information. This study investigates to what extent citizens have knowledge about daily politics and to what extent news on social media can provide this knowledge. It does so by means of a large online survey in Belgium (Flanders), in which we measured what people know about current political events, their so-called general surveillance knowledge. Our findings demonstrate that unlike following news via traditional media channels, citizens do not gain more political knowledge from following news on social media. We even find a negative association between following the news on Facebook and political knowledge. We further investigate why this is the case. Our data demonstrate that this lack of learning on social media is not due to a narrow, personalized news diet, as is often suggested. Rather, we find evidence that following news via social media increases a feeling of information overload, which decreases what people actually learn, especially for citizens who combine news via social media with other news sources.

KEYWORDS

Political knowledge; social media; filter bubbles; information overload

(Van Erkel, 2020; Van Erkel & Van Aelst, 2021)

MODELLE DER HEUTIGEN SITZUNG (HIER: M1 & M4)

Table 4. OLS regression.

Political knowledge	Model 1 b(SE)	Model 2 b(SE)	Model 3 b(SE)	Model 4 b(SE)
Radio	−.009(.02)	−.019(.03)	−.018(.03)	−.007(.02)
TV	.082(.03)**	.105(.04)**	.105(.04)**	.086(.02)**
Newspapers	.082(.02)**	.083(.03)**	.083(.03)**	.082(.02)**
News websites	.062(.02)**	.069(.03)*	.070(.03)*	.063(.02)**
Twitter	−.059(.04)	−.064(.04)	−.063(.04)	−.052(.04)
Facebook	−.075(.02)**	−.075(.02)**	−.037(.06)	−.069(.02)**
Female	−.479(.07)**	−.525(.08)**	−.525(.08)**	−.461(.07)**
Age	.016(.00)**	.016(.00)**	.016(.00)**	.017(.00)**
Level of education (ref = Low)				
• Middle	.276(.11)*	.249(.13)*	.245(.13)*	.270(.11)*
• High	.482(.11)**	.504(.13)**	.496(.13)**	.476(.11)**
Political interest	.175(.01)**	.159(.02)**	.159(.02)**	.173(.01)**
Personalized news environment scale		−.032(.02)	−.010(.04)	
Facebook*Personalized news environment scale			−.008(.01)	
Information overload				−.033(.01)**
Constant	.463(.22)**	.684(.28)**	.577(.32)**	.665(.23)**
N	**993**	**779**	**779**	**993**
Adjusted R^2	**.37**	**.35**	**.35**	**.38**

*p < .05, **p < .01

(Van Erkel & Van Aelst, 2021)

Grundlagen der multiplen Regression

GRUNDGEDANKE DER MULTIPLEN REGRESSION

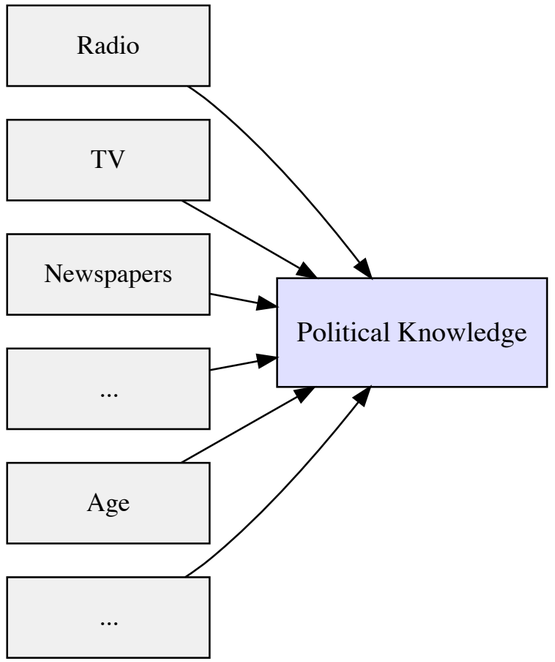
Die Berücksichtigung mehrerer Prädiktoren in einer Analyse ermöglicht

- Eine Verbesserung der Erklärungsleistung des gesamten Modells
- Eine Verbesserung der Vorhersagequalität des gesamten Modells
- Den Vergleich der Bedeutung von Zusammenhängen
- Die Berücksichtigung von Drittvariablen bei Vergleichen

- Mit zusätzlichen Annahmen: Schätzen von kausalen Effekten mit nicht-experimentellen Daten
- In komplexeren Modellen: Modellierung von Wechselwirkungen mehrerer Einflüsse (Moderation)
- In komplexeren Modellvergleichen: Modellierung von mehrstufigen Zusammenhängen (X → M → Y, Mediation)

GRUNDGEDANKE DER MULTIPLEN REGRESSION

- Schätzung der aV durch die Linearkombination mehrerer uV
- Die Regressionskoeffizienten werden nach der Methode der kleinsten Quadrate (OLS) so geschätzt, dass die Quadratsumme der Residuen minimiert wird:
  $\beta = (X^T X)^{-1} X^T Y$
- Im einfachen, linearen Fall wird ein additives Modell angenommen, d.h., wir gehen davon aus, dass sich die Einflüsse der Prädiktoren auf die aV aufsummieren
- Die Koeffizienten b_i werden als partielle Regressionskoeffizienten bezeichnet.



- Regressionsgerade:
  $\text{Political Knowledge} = b_0 + b_1 \times \text{Radio} + b_2 \times \text{TV} + b_3 \times \text{Newspapers} + b_4 \times \text{Age} + \ldots + \varepsilon$

REGRESSIONSTABELLE

Parameter	Coefficient	95% CI	t(990)	p	Std. Coef.	Std. Coef. 95% CI	Fit
(Intercept)	0.99	(0.67, 1.31)	6.07	< .001	0.00	(-0.06, 0.06)	
Age	0.02	(0.02, 0.03)	8.24	< .001	0.24	(0.19, 0.30)	
Newspapers	0.22	(0.18, 0.27)	9.31	< .001	0.28	(0.22, 0.33)	
R2 (adj.)							0.16

Text auf der nächsten Folie

REGRESSIONSTABELLE - ORIENTIERUNGSHILFE

- **Parameter**: Name des Prädiktors; *(Intercept)* = Konstante = Schnittpunkt mit y-Achse wenn alle Prädiktoren = 0
- **Coefficient**: Partielle Regressionskoeffizienten in Einheiten der Variablen (hier: Jahre, Skalenpunkte, Zahl richtige Antworten)
- **95% CI**: 95% Konfidenzintervalle um *Coefficient*; Werte in diesem Intervall können wir plausibel mit den Daten vereinbaren
- **t(990)**: t-Werte = $\text{Coefficient}/\text{SE}_{\text{Coefficient}}$; Freiheitsgrade in Klammern; Daraus ergibt sich der p-Wert.
- **p**: p-Wert; Wie wahrscheinlich wäre es, diesen oder einen noch größeren Regressionskoeffizienten zu ermitteln, wenn der Regressionskoeffizient in der Grundgesamtheit = 0 wäre? Wenn $p < \alpha$, dann nennen wir den Koeffizienten “statistisch signifikant”.
- **Std. Coef.**, **Std. Coef. 95% CI**:
  - Standardisierter partieller Regressionskoeffizienten mit ihren 95% Konfidenzintervallen.
  - Alle *quasi-metrischen* Prädiktoren und die abhängige Variable werden vor der Schätzung Z-standardisiert (Mittelwert = 0, SD = 1).
  - *Vorsicht*: Manchmal werden auch standardisierte Koeffizienten für binäre Prädiktoren berichtet; schwer zu interpretieren, da binäre Prädiktoren keine SD haben.
  - Liegt (fast) immer im Bereich $[-1; +1]$; *ähnlich* (nicht identisch!) Korrelationskoeffizient *r*
  - Zusammenhänge in Einheiten von Standardabweichungen sind zwischen Prädiktoren, die mit unterschiedlichen Einheiten gemessen wurden, vergleichbar.
- **Fit**, **R2 (adj.)**: Anteil erklärter Varianz durch das gesamte Modell (alle Prädiktoren gemeinsam), angepasst um die zufällige Erklärung durch jeden weiteren Prädiktor. Details in 4 Folien.

INTERPRETATION: VERGLEICH

Parameter	Coefficient	95% CI	t(990)	p	Std. Coef.	Std. Coef. 95% CI	Fit
(Intercept)	0.99	(0.67, 1.31)	6.07	< .001	0.00	(-0.06, 0.06)	
Age	0.02	(0.02, 0.03)	8.24	< .001	0.24	(0.19, 0.30)	
Newspapers	0.22	(0.18, 0.27)	9.31	< .001	0.28	(0.22, 0.33)	
R2 (adj.)							0.16

- Zwei Personen, die in allen x_k dieselbe Ausprägung haben und sich in x_1 um einen Skalenpunkt unterscheiden, unterscheiden sich in y um b_1 Punkte.
- Wir vergleichen zwei Personen, die sich im Alter um ein Jahr unterscheiden und die gleich häufig Zeitungen nutzen. Die ältere Person beantwortet 0.02 Fragen mehr korrekt als die jüngere Person.
- Wir vergleichen zwei gleich alte Personen, deren Zeitungsnutzung sich um einen Skalenpunkt unterscheidet. Die Person, die sich häufiger über Zeitungen informiert, beantwortet 0.22 Fragen mehr korrekt.
- Entsprechende Interpretation mit standardisierten partiellen Regressionskoeffizienten: Standardabweichungen statt Punkte / Jahre / Fragen

INTERPRETATION: VERÄNDERUNG, INTERVENTION

Parameter	Coefficient	95% CI	t(990)	p	Std. Coef.	Std. Coef. 95% CI	Fit
(Intercept)	0.99	(0.67, 1.31)	6.07	< .001	0.00	(-0.06, 0.06)	
Age	0.02	(0.02, 0.03)	8.24	< .001	0.24	(0.19, 0.30)	
Newspapers	0.22	(0.18, 0.27)	9.31	< .001	0.28	(0.22, 0.33)	
R2 (adj.)							0.16

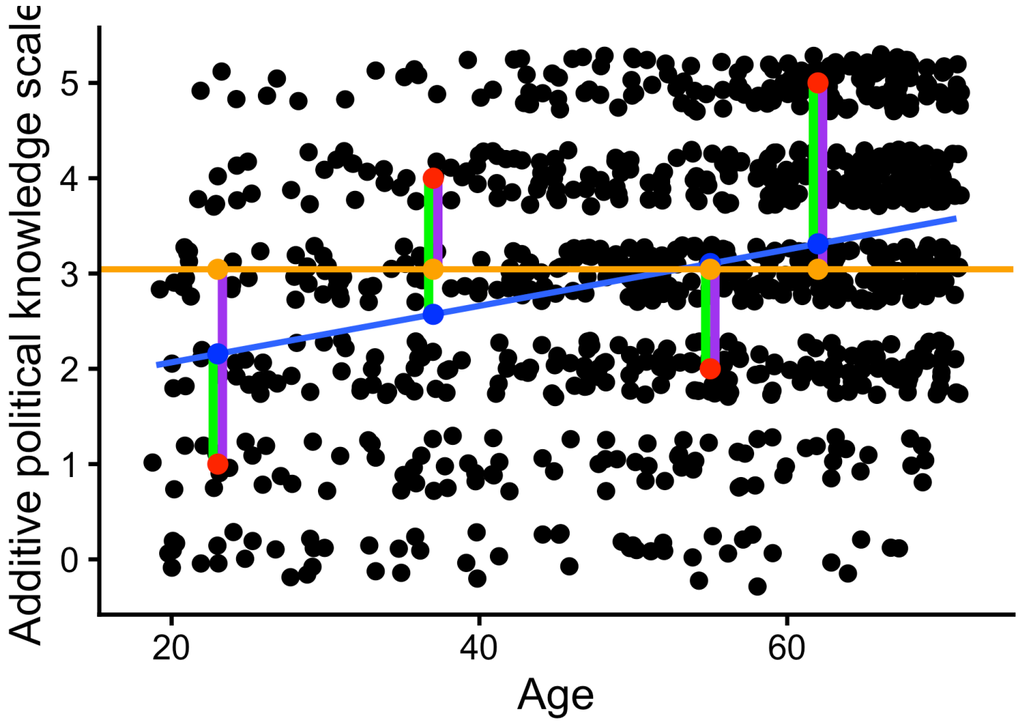
- Wenn x_1 um einen Punkt steigt und alle anderen x_k konstant gehalten werden, steigt y um b_1 Punkte.
- Wenn eine Person um ein Jahr älter wird und ihr Zeitungsnutzungsverhalten nicht verändert, beantwortet sie 0.02 Fragen mehr korrekt (Annahme: Kontrolle von Zeitungsnutzung deckt alle alternativen Ursachen von politischem Wissen ab).
- Wenn eine Person ihre Zeitungsnutzung um einen Skalenpunkt steigert, dann beantwortet sie unmittelbar (im Sinne von: nichts durch weiteres Lebensalter gelernt) 0.22 Fragen mehr korrekt (Annahme: Kontrolle von Alter deckt alle alternativen Ursachen von politischem Wissen ab).
- Entsprechende Interpretation mit standardisierten partiellen Regressionskoeffizienten: Standardabweichungen statt Punkte / Jahre / Fragen

LINEARE REGRESSION: R^2

Beispiel für vier Befragte:

Age	y	yhat	e	e2	e_M	e_M2
23	1	2.2	-1.2	1.3	-2	4.2
37	4	2.6	1.4	2.0	1	0.9
55	2	3.1	-1.1	1.2	-1	1.1
62	5	3.3	1.7	2.9	2	3.8

- y: Beobachteter Wert
- $\hat{y}$: Vorhergesagter Wert
- e: Residuum, Vorhersagefehler
- $\bar{y}$: Mittelwert
- e_m: Abweichung vom Mittelwert



$$R^2 = \frac{\Sigma(y_i-\bar{y})^2-\Sigma(y_i-\hat{y})^2}{\Sigma(y_i-\bar{y})^2}$$

- $R^2 = .09$: Anteil der Varianz, die das Regressionsmodell erklärt; 0 (Modell erklärt keine Varianz) bis 1 (perfekter linearer Zusammenhang). Vergleich mit Mittelwert als einfachstem Modell von y.

INTERPRETATION: KORRIGIERTES (ADJUSTED) R^2

Parameter	Coefficient	95% CI	t(990)	p	Std. Coef.	Std. Coef. 95% CI	Fit
(Intercept)	0.99	(0.67, 1.31)	6.07	< .001	0.00	(-0.06, 0.06)	
Age	0.02	(0.02, 0.03)	8.24	< .001	0.24	(0.19, 0.30)	
Newspapers	0.22	(0.18, 0.27)	9.31	< .001	0.28	(0.22, 0.33)	
R2 (adj.)							0.16

- $\text{korr. | adj. } R^2 = R^2 - \frac{n \times (1-R^2)}{n-k-1}$, mit Fallzahl n und Zahl der Prädiktoren k
- Maß für die Varianzerklärung des gesamten Regressionsmodells (aller Prädiktoren gemeinsam)
- 0 (Modell erklärt keine Varianz) bis 1 (aV ist perfekte Linearkombination der Prädikoren)
- Korrektur für den Umstand, dass die Linearkombination von vielen Prädiktoren Varianz in der unabhängigen Variable alleine dadurch erklärt, dass sie typisch für die Fälle in der Stichprobe ist.

Fragen?

Kategorielle Prädiktoren

Vergleiche Abschnitt Regression und Mittelwertvergleich in Einheit zur bivariaten linearen Regresssion

MODELLE DER HEUTIGEN SITZUNG (HIER: M1 & M4)

Table 4. OLS regression.

Political knowledge	Model 1 b(SE)	Model 2 b(SE)	Model 3 b(SE)	Model 4 b(SE)
Radio	−.009(.02)	−.019(.03)	−.018(.03)	−.007(.02)
TV	.082(.03)**	.105(.04)**	.105(.04)**	.086(.02)**
Newspapers	.082(.02)**	.083(.03)**	.083(.03)**	.082(.02)**
News websites	.062(.02)**	.069(.03)*	.070(.03)*	.063(.02)**
Twitter	−.059(.04)	−.064(.04)	−.063(.04)	−.052(.04)
Facebook	−.075(.02)**	−.075(.02)**	−.037(.06)	−.069(.02)**
Female	−.479(.07)**	−.525(.08)**	−.525(.08)**	−.461(.07)**
Age	.016(.00)**	.016(.00)**	.016(.00)**	.017(.00)**
Level of education (ref = Low)				
• Middle	.276(.11)*	.249(.13)*	.245(.13)*	.270(.11)*
• High	.482(.11)**	.504(.13)**	.496(.13)**	.476(.11)**
Political interest	.175(.01)**	.159(.02)**	.159(.02)**	.173(.01)**
Personalized news environment scale		−.032(.02)	−.010(.04)	
Facebook*Personalized news environment scale			−.008(.01)	
Information overload				−.033(.01)**
Constant	.463(.22)**	.684(.28)**	.577(.32)**	.665(.23)**
N	**993**	**779**	**779**	**993**
Adjusted R^2	**.37**	**.35**	**.35**	**.38**

*p < .05, **p < .01

(Van Erkel & Van Aelst, 2021)

KATEGORIELLE PRÄDIKTOREN

- Die lineare Regressionsanalyse (bzw. das allgemeine lineare Modell) ist ein sehr flexibles Werkzeug.
- Bekannte Verfahren zum Vergleich von Gruppenmittelwerten können als Spezialfälle der linearen Regression betrachtet werden. Sie sind statistisch äquivalent, für spezifische Anwendungen teils einfacher zu verwenden.
  - T-Test: Vergleich von 2 Gruppenmittelwerten
  - Varianzanalyse: Traditionelle mehrfaktorielle Experimentaldesigns
- Für alle weitergehenden Analysen sind die lineare Regression und ihre Erweiterungen zu empfehlen, da sie flexibel angepasst werden können.

- Erweiterung des bisherigen Modells zur Erklärung von politischem Wissen um zwei kategorielle Prädiktoren:
  - Wiederholung: Gender (zweistufig, binär)
  - Education (dreistufig, ordinal)

WIEDERHOLUNG: BINÄRE PRÄDIKTOREN (BEISPIEL)

- $\mathbf{Gender}$ wird in eine *Dummy*-Variable $\mathbf{female}$ recodiert (0 = not female [hier: male], 1 = female)
- Regressionsgerade: $Y = b_0 + b_1 \times \mathrm{female} + \varepsilon$
- Wenn $\mathbf{female} = 0$: $Y = b_0 + \varepsilon$
- b_0 ist Mittelwert der Männer
- b_1 ist Differenz zwischen Männern und Frauen

Regression

Parameter	Coefficient	95% CI	t(991)	p	Fit
(Intercept)	3.44	(3.33, 3.55)	60.48	< .001	
Gender (female)	-0.84	(-1.00, -0.67)	-10.14	< .001	
R2					0.09

PRÄDIKTOREN MIT k AUSPRÄGUNGEN

- um k Gruppen zu vergleichen, werden $k-1$ Prädiktor-Variablen erstellt
- die $k-1$ Variablen werden in das Modell aufgenommen:
  $Y = b_0 + b_1 \times X_1 + b_2 \times X_2 + \ldots + b_{k-1} \times X_{k-1} + \varepsilon.$
- Bei Dummy-Codierung:
  - in der Referenzgruppe (alle $X_1 = X_2 = \ldots = 0$) ergibt sich $Y = b_0$
  - b_1 ist die Differenz zwischen der Gruppe 1 und der Referenzgruppe, b_2 die Differenz zwischen der Gruppe 2 und der Referenzgruppe, …
- $k-1$ paarweise Vergleiche in einem Modell gleichzeitig
- Mehr Vergleiche: Modell mehrmals schätzen oder Post-Hoc-Tests

Alternativen zur Dummy-Codierung: UCLA Statistical Methods and Data Analysis (nicht Teil der Vorlesung)

DUMMY-CODIERUNG MIT k AUSPRÄGUNGEN (BEISPIEL)

Niedrige Bildung als Referenz

Dummy-Variablen

Zugehörigkeit	Middle	High
Lower	0	0
Middle	1	0
High	0	1

Mittelwerte

Lower	Middle	High
2.58	2.97	3.25

Regression

Parameter	Coefficient	95% CI	t(990)	p	Fit
(Intercept)	2.58	(2.35, 2.81)	22.39	< .001	
Education (Middle)	0.39	(0.13, 0.65)	2.92	0.004	
Education (High)	0.67	(0.41, 0.93)	5.09	< .001	

Mittlere Bildung als Referenz

Dummy-Variablen

Zugehörigkeit	Lower	High
Middle	0	0
Lower	1	0
High	0	1

Mittelwerte

Lower	Middle	High
2.58	2.97	3.25

Regression

Parameter	Coefficient	95% CI	t(990)	p	Fit
(Intercept)	2.97	(2.84, 3.10)	44.41	< .001	
Education (Lower)	-0.39	(-0.65, -0.13)	-2.92	0.004	
Education (High)	0.28	(0.10, 0.46)	3.03	0.002	

ALTERNATIVE: POST-HOC-VERGLEICHE

- Erst Modell schätzen, dann relevante (oder alle) Vergleiche betrachten

Contrast	Estimate	Std. Error	z	Pr(>|z|)	S	2.5 %	97.5 %
High - Lower	0.669	0.131	5.09	< 0.001	21.4	0.4109	0.926
High - Middle	0.279	0.092	3.03	0.00241	8.7	0.0988	0.459
Middle - Lower	0.389	0.133	2.92	0.00348	8.2	0.1282	0.651

Type: response

- Korrektur der *p*-Werte für Mehrfach-Vergleiche (hier: nach Bonferroni) wird empfohlen, um α-Fehler durch viele Vergleiche unwahrscheinlicher zu machen.

Estimate	Std. Error	z	Pr(>|z|)	S	2.5 %	97.5 %
0.669	0.131	5.09	< 0.001	19.8	0.3619	0.975
0.279	0.092	3.03	0.00724	7.1	0.0645	0.494
0.389	0.133	2.92	0.01045	6.6	0.0785	0.700

Details zu Korrekturverfahren z.B. Bortz & Schuster (2010), S. 232; nicht klausurrelevant)

Fragen?

“Große” Regressionsmodelle

ZU REPRODUZIERENDE MODELLE (HIER: M1 & M4)

Table 4. OLS regression.

Political knowledge	Model 1 b(SE)	Model 2 b(SE)	Model 3 b(SE)	Model 4 b(SE)
Radio	−.009(.02)	−.019(.03)	−.018(.03)	−.007(.02)
TV	.082(.03)**	.105(.04)**	.105(.04)**	.086(.02)**
Newspapers	.082(.02)**	.083(.03)**	.083(.03)**	.082(.02)**
News websites	.062(.02)**	.069(.03)*	.070(.03)*	.063(.02)**
Twitter	−.059(.04)	−.064(.04)	−.063(.04)	−.052(.04)
Facebook	−.075(.02)**	−.075(.02)**	−.037(.06)	−.069(.02)**
Female	−.479(.07)**	−.525(.08)**	−.525(.08)**	−.461(.07)**
Age	.016(.00)**	.016(.00)**	.016(.00)**	.017(.00)**
Level of education (ref = Low)				
• Middle	.276(.11)*	.249(.13)*	.245(.13)*	.270(.11)*
• High	.482(.11)**	.504(.13)**	.496(.13)**	.476(.11)**
Political interest	.175(.01)**	.159(.02)**	.159(.02)**	.173(.01)**
Personalized news environment scale		−.032(.02)	−.010(.04)	
Facebook*Personalized news environment scale			−.008(.01)	
Information overload				−.033(.01)**
Constant	.463(.22)**	.684(.28)**	.577(.32)**	.665(.23)**
N	**993**	**779**	**779**	**993**
Adjusted R^2	**.37**	**.35**	**.35**	**.38**

*p < .05, **p < .01

(Van Erkel & Van Aelst, 2021)

MODELLE SCHÄTZEN

```r
m1 <- lm(Political_knowledge ~ Radio + Television + Newspapers + Online_news_sites + Twitter +
  Facebook + Gender + Age + Education + Political_interest, data = d)

m4 <- lm(Political_knowledge ~ Radio + Television + Newspapers + Online_news_sites + Twitter +
  Facebook + Gender + Age + Education + Political_interest + Information_overload, data = d)
```

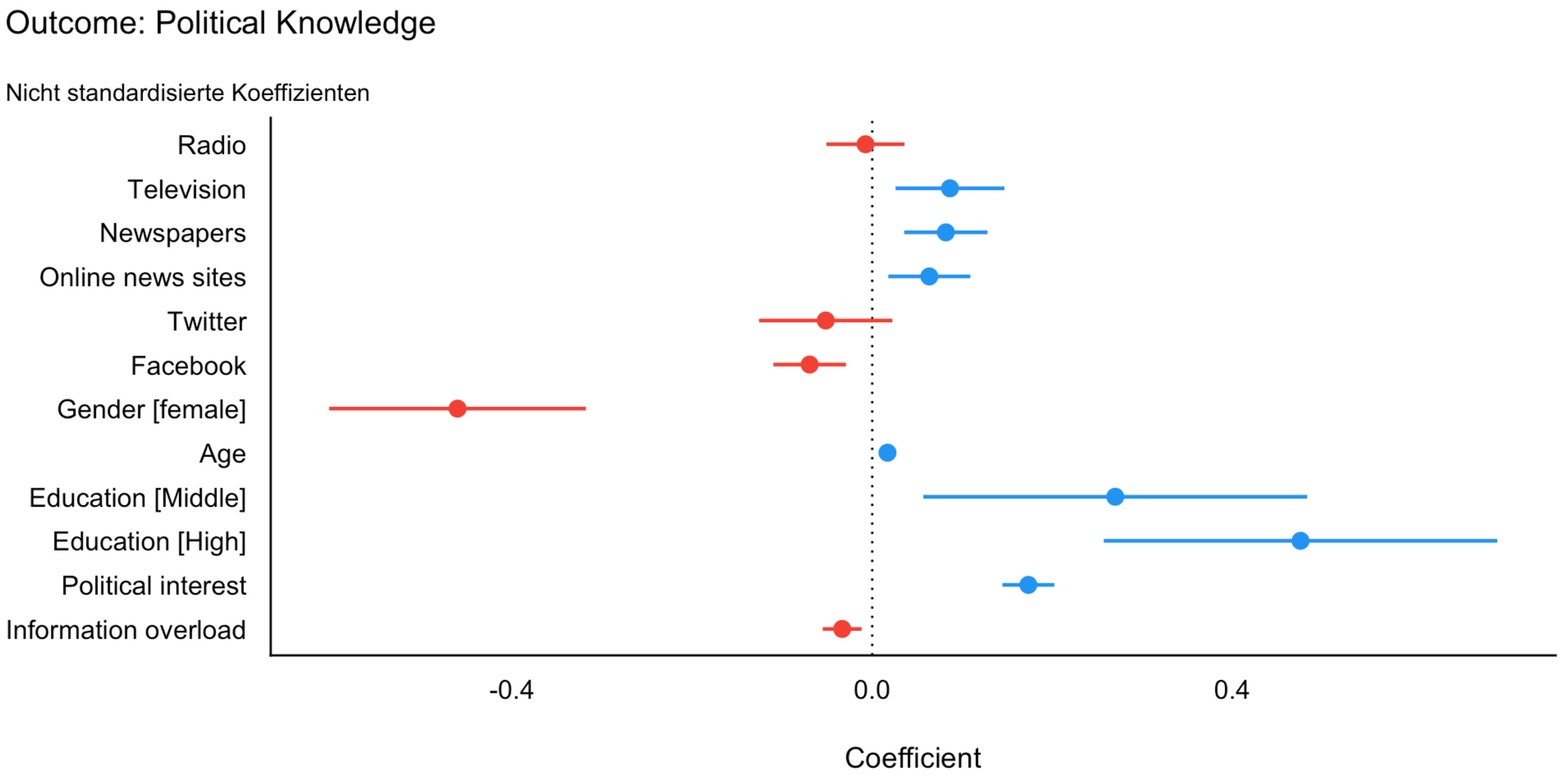
MODELS 1 & 4

	Model 1	Model 4
(Intercept)	0.46 (0.22)*	0.67 (0.23)**
Radio	-0.01 (0.02)	-0.01 (0.02)
Television	0.08 (0.03)**	0.09 (0.03)**
Newspapers	0.08 (0.02)***	0.08 (0.02)***
Online_news_sites	0.06 (0.02)**	0.06 (0.02)**
Twitter	-0.06 (0.04)	-0.05 (0.04)
Facebook	-0.07 (0.02)***	-0.07 (0.02)***
Genderfemale	-0.48 (0.07)***	-0.46 (0.07)***
Age	0.02 (0.00)***	0.02 (0.00)***
EducationMiddle	0.28 (0.11)*	0.27 (0.11)*
EducationHigh	0.48 (0.11)***	0.48 (0.11)***
Political_interest	0.18 (0.01)***	0.17 (0.01)***
Information_overload		-0.03 (0.01)**
Num.Obs.	993	993
R2 Adj.	0.371	0.376

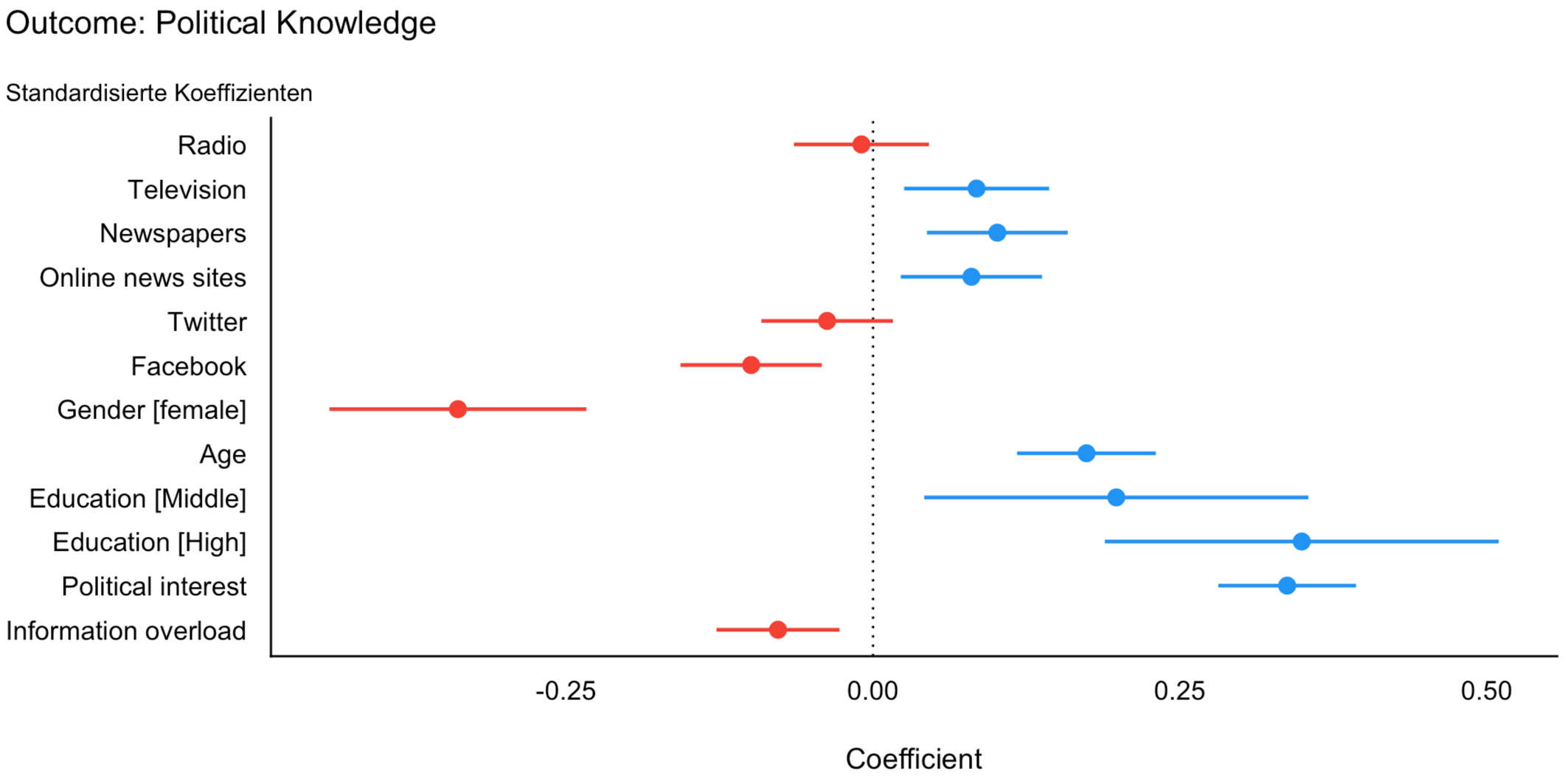
MODEL 4: AUSFÜHRLICHE REGRESSIONSTABELLE

Parameter	Coefficient	95% CI	t(980)	p	Std. Coef.	Std. Coef. 95% CI	Fit
(Intercept)	0.67	(0.21, 1.12)	2.90	0.004	-0.08	(-0.22, 0.06)	
Radio	-0.01	(-0.05, 0.04)	-0.34	0.736	-0.01	(-0.06, 0.05)	
Television	0.09	(0.03, 0.15)	2.80	0.005	0.08	(0.03, 0.14)	
Newspapers	0.08	(0.04, 0.13)	3.46	< .001	0.10	(0.04, 0.16)	
Online news sites	0.06	(0.02, 0.11)	2.73	0.006	0.08	(0.02, 0.14)	
Twitter	-0.05	(-0.13, 0.02)	-1.37	0.171	-0.04	(-0.09, 0.02)	
Facebook	-0.07	(-0.11, -0.03)	-3.38	< .001	-0.10	(-0.16, -0.04)	
Gender (female)	-0.46	(-0.60, -0.32)	-6.34	< .001	-0.34	(-0.44, -0.23)	
Age	0.02	(0.01, 0.02)	6.03	< .001	0.17	(0.12, 0.23)	
Education (Middle)	0.27	(0.06, 0.48)	2.48	0.013	0.20	(0.04, 0.35)	
Education (High)	0.48	(0.26, 0.69)	4.27	< .001	0.35	(0.19, 0.51)	
Political interest	0.17	(0.14, 0.20)	11.79	< .001	0.34	(0.28, 0.39)	
Information overload	-0.03	(-0.05, -0.01)	-3.03	0.003	-0.08	(-0.13, -0.03)	
R2 (adj.)							0.38

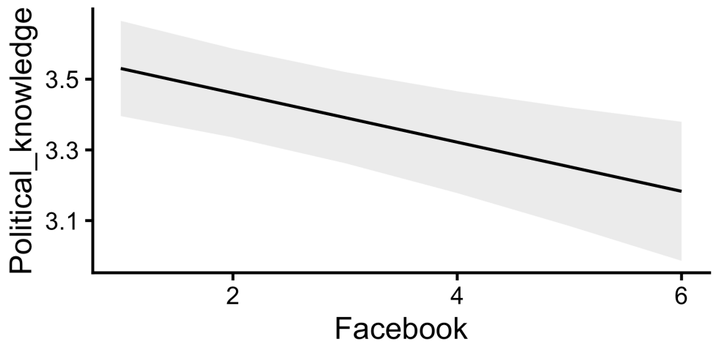
MODEL 4: KOEFFIZIENTENPLOT NICHT STANDARDISIERT



MODEL 4: KOEFFIZIENTENPLOT STANDARDISIERT



MODEL 4: VORHERSAGE FÜR EINZELNE PRÄDIKTOREN



Facebook	Estimate	Std. Error	z	Pr(>|z|)	S	2.5 %	97.5 %
1	3.53	0.0687	51.4	<0.001	Inf	3.40	3.66
2	3.46	0.0640	54.1	<0.001	Inf	3.34	3.59
3	3.39	0.0657	51.6	<0.001	Inf	3.26	3.52
4	3.32	0.0734	45.2	<0.001	Inf	3.18	3.47
5	3.25	0.0855	38.1	<0.001	1050.3	3.08	3.42
6	3.18	0.1003	31.7	<0.001	731.9	2.99	3.38

Type: response

- Vorhergesagte Werte, wenn andere Prädiktoren den Mittelwert bzw. Modus haben.

MODEL 4: VERGLEICHE FÜR EINZELNE PRÄDIKTOREN

Term	Contrast	Estimate	Std. Error	z	Pr(>|z|)	S	2.5 %	97.5 %
Education	High - Lower	0.476	0.1114	4.27	< 0.001	15.7	0.2575	0.694
Education	High - Middle	0.206	0.0757	2.72	0.00654	7.3	0.0575	0.354
Education	Middle - Lower	0.270	0.1086	2.48	0.01297	6.3	0.0570	0.483
Facebook	(x + sd) - (x - sd)	-0.270	0.0799	-3.38	< 0.001	10.5	-0.4271	-0.114
Gender	female - male	-0.461	0.0727	-6.34	< 0.001	32.0	-0.6030	-0.318
Newspapers	(x + sd) - (x - sd)	0.276	0.0796	3.46	< 0.001	10.9	0.1198	0.432

Type: response

Average counterfactual comparison

1. Compute predictions for every row of the dataset in the counterfactual world where all observations belong to the treatment condition.
2. Compute predictions for every row of the dataset in the counterfactual world where all observations belong to the control condition.
3. Take the differences between the two vectors of predictions.
4. Average the unit-level estimates across the whole dataset, or within subgroups.

Fragen?

Annahmen und ihre Überprüfung

Siehe Einheit zur bivariaten linearen Regression für Details

ANNAHMEN UND IHRE ÜBERPRÜFUNG

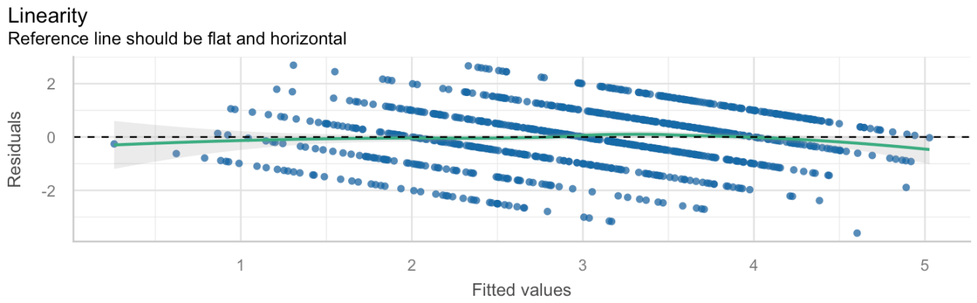
Statistische Annahmen

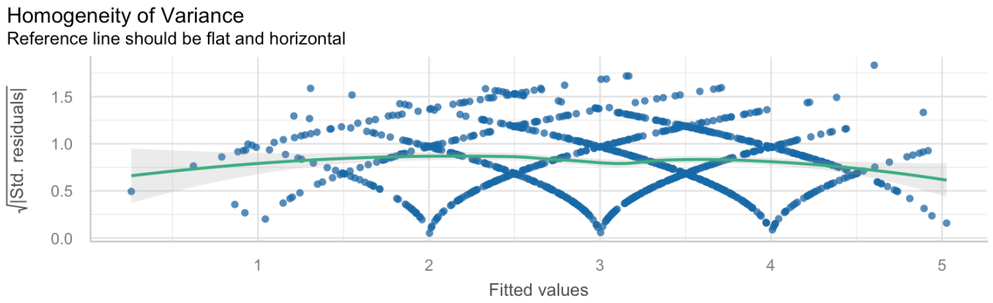
- Linearität und Additivität der Zusammenhänge
- Normalverteilung und Homoskedastizität der Residuen
- Unabhängigkeit der Residuen
- keine einflussreichen Ausreißer
- **keine Multikollinearität**

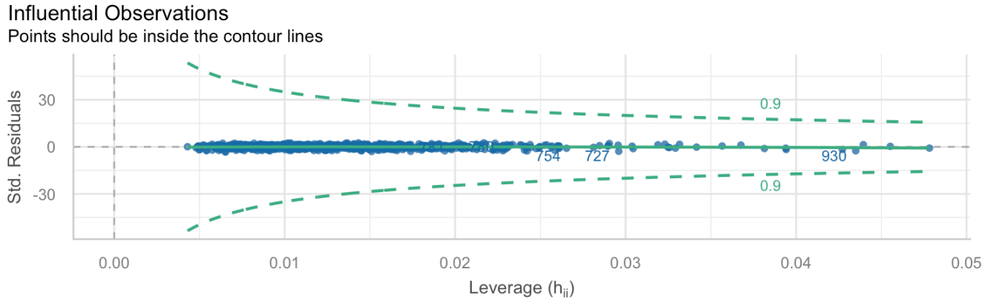
Kausalannahmen

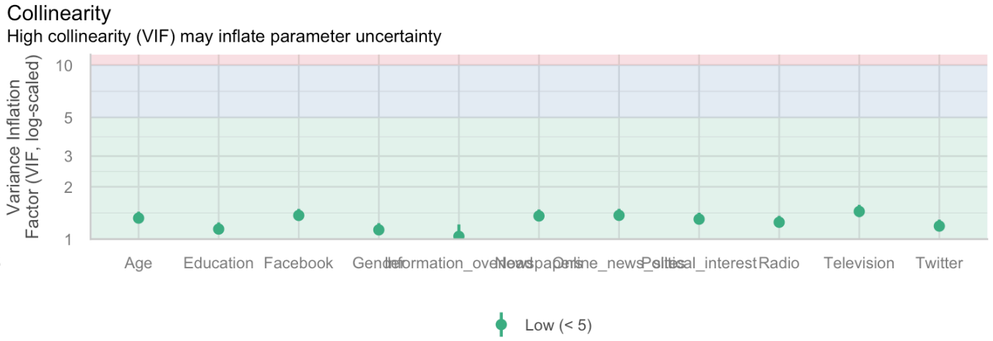
- korrekt spezifiziertes Modell; keine fehlenden oder überflüssigen Variablen
- Besprechen wir in eigener Einheit

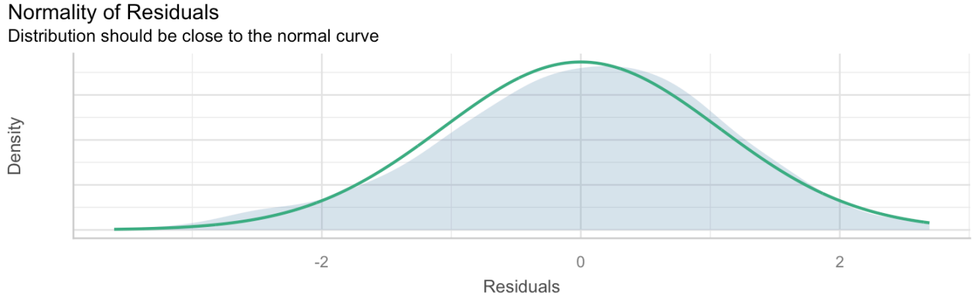
MODEL 4 AUS VAN ERKEL & VAN AELST (2021)





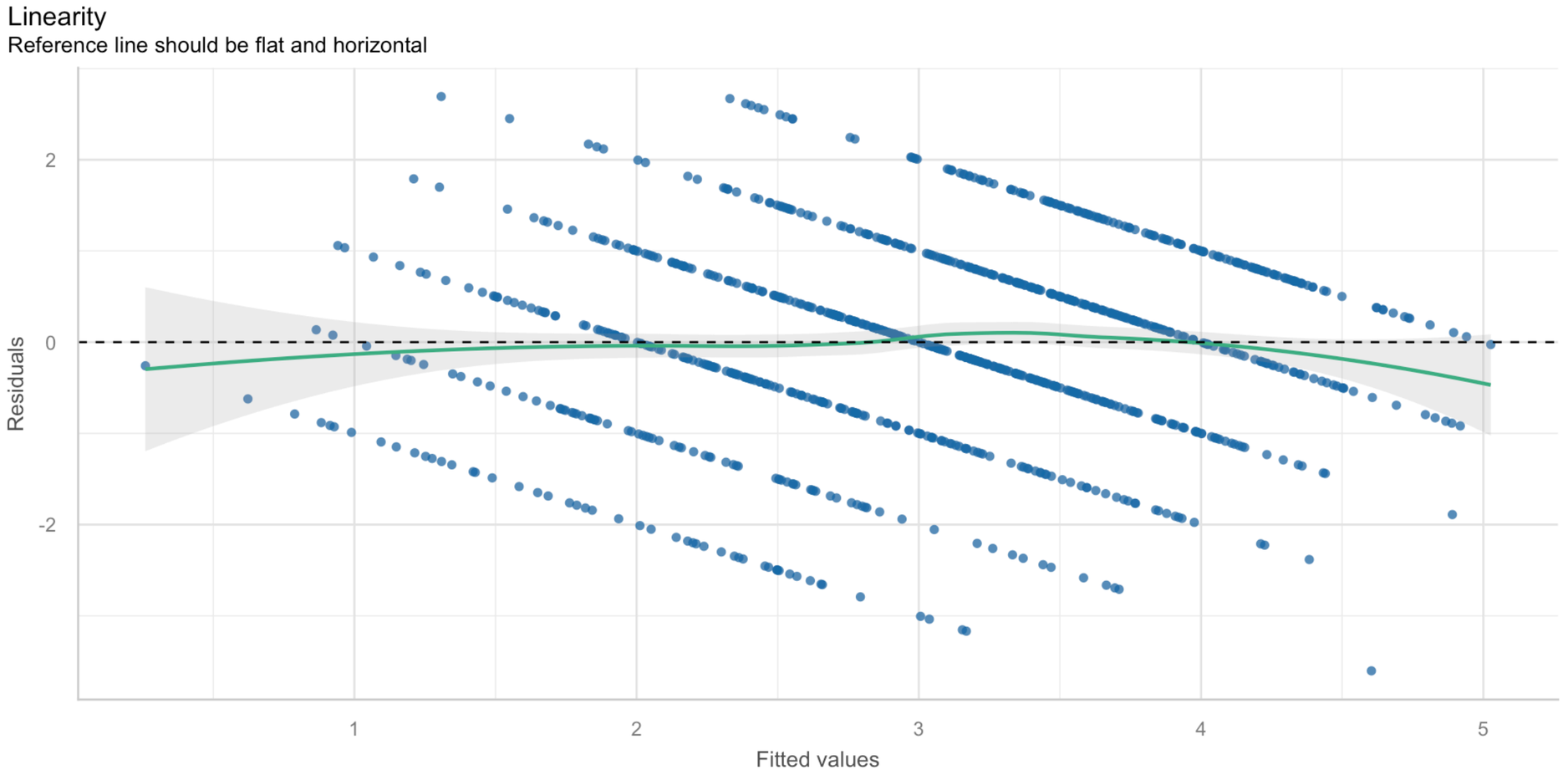




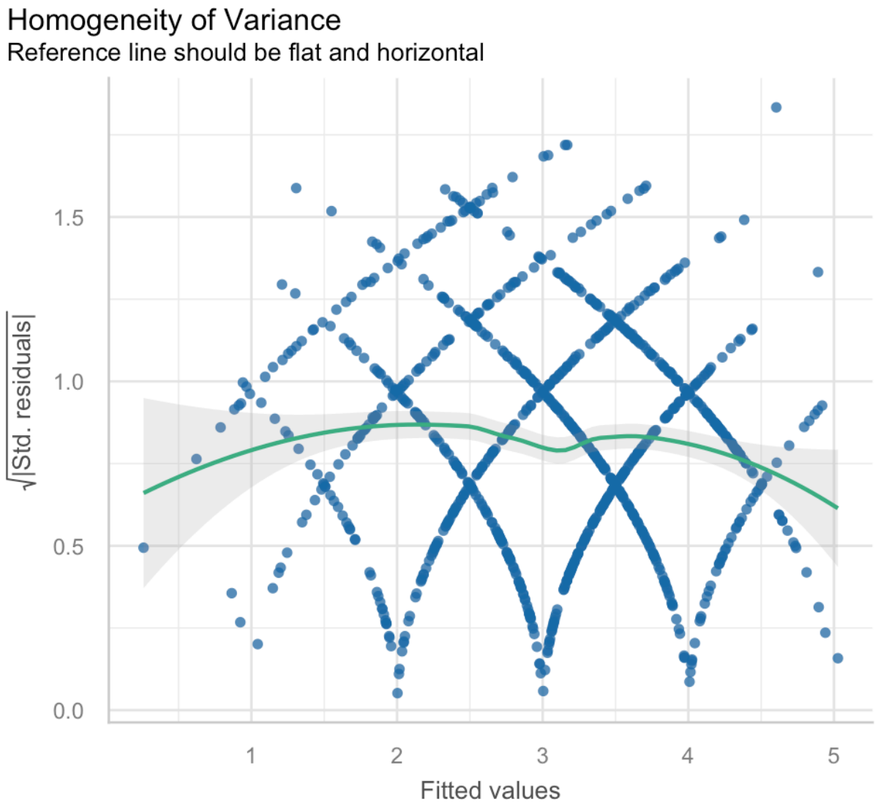


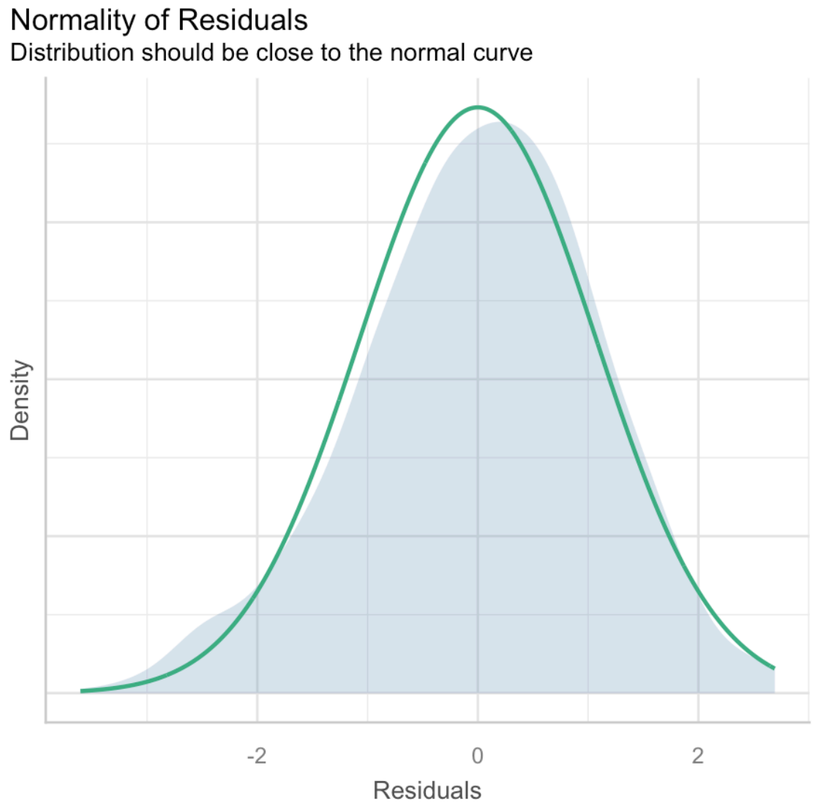
LINEARITÄT

```
$NCV
```

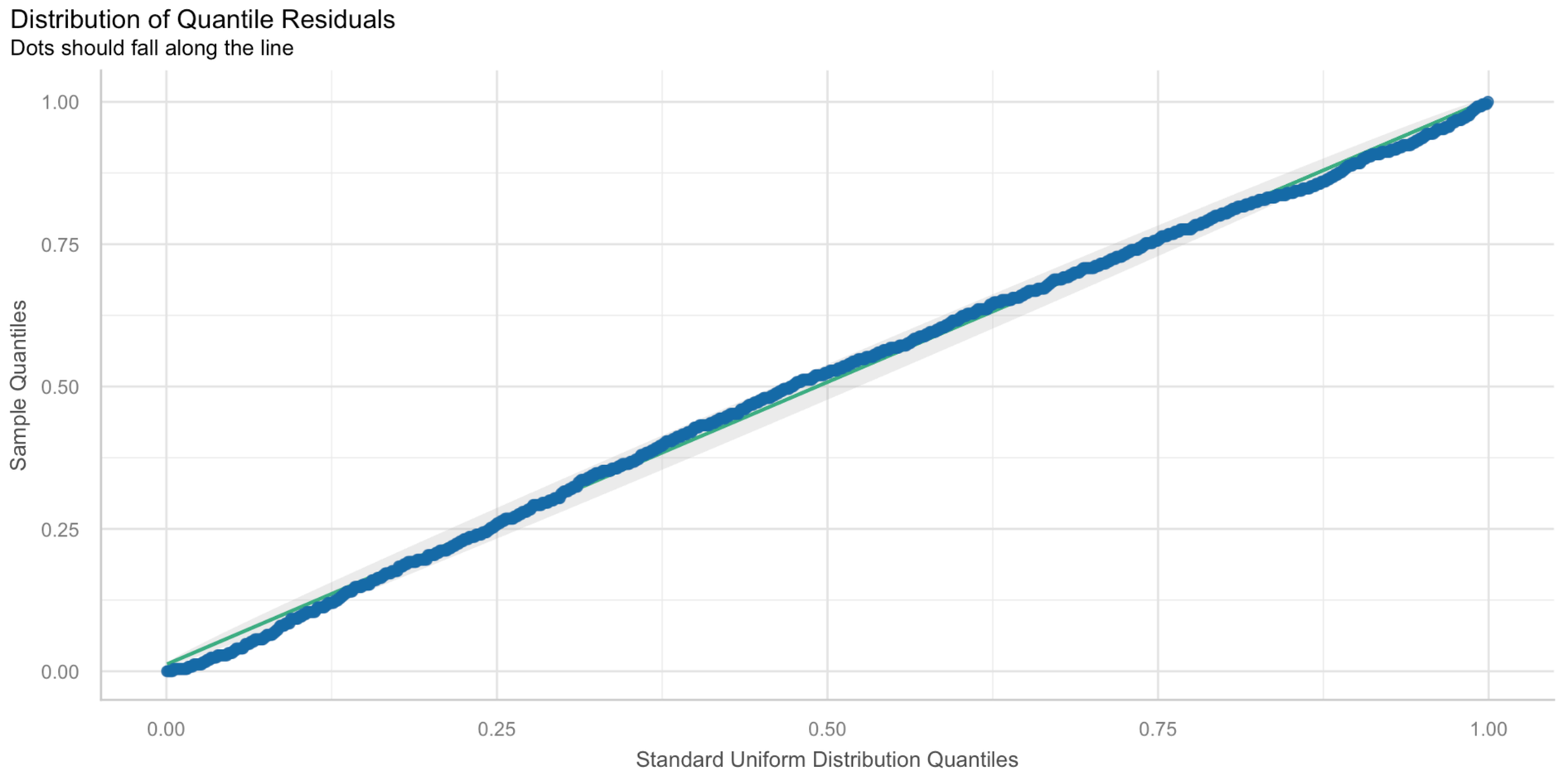


NORMALVERTEILUNG UND HOMOSKEDASTIZITÄT DER RESIDUEN

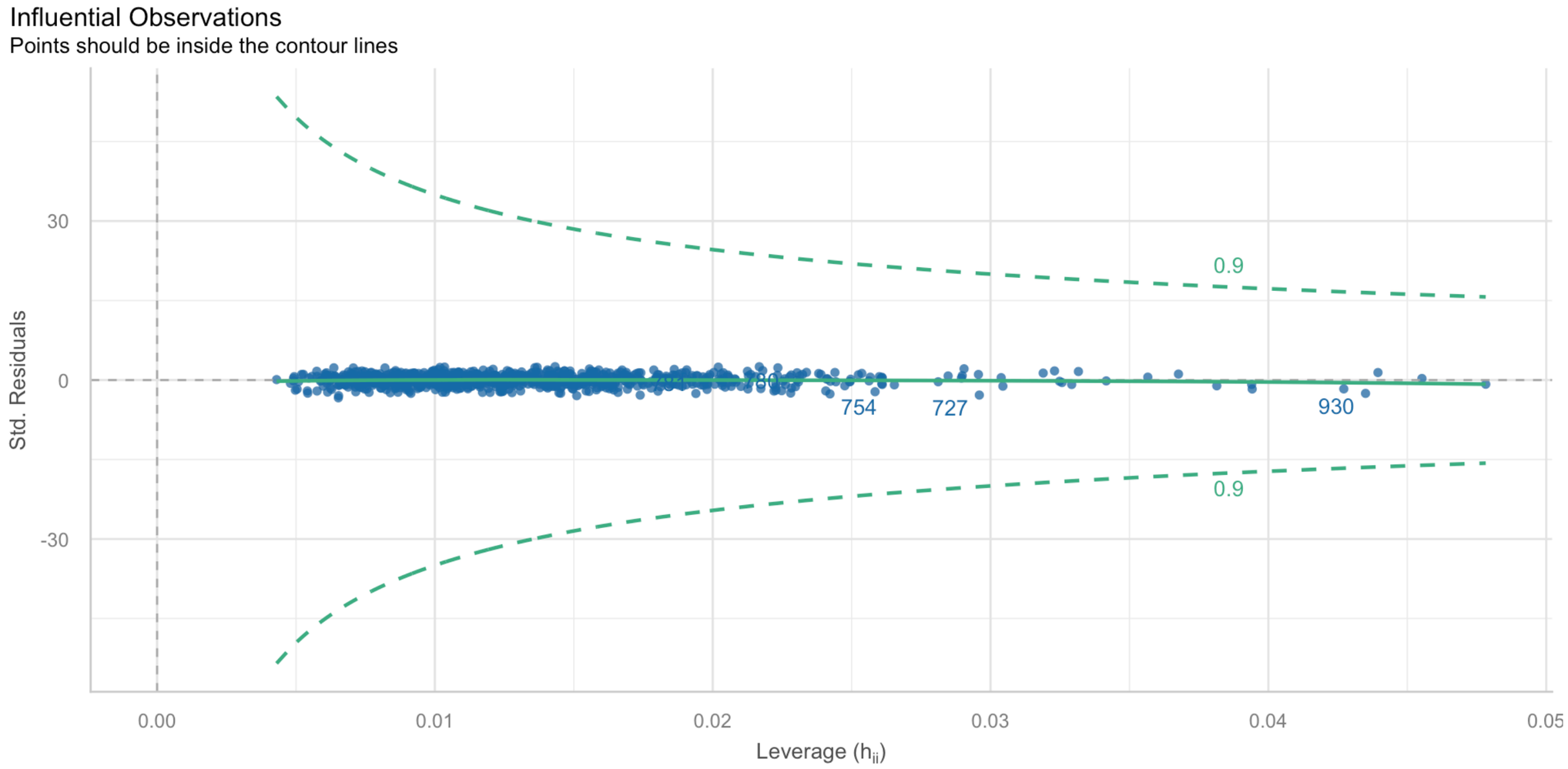




UNABHÄNGIGKEIT DER RESIDUEN



KEINE EINFLUSSREICHEN AUSREISSER



KEINE MULTIKOLLINEARITÄT

- **Annahme**: Prädiktorvariablen X korrelieren nicht zu stark miteinander
- **Diagnose**: Korrelationsmatrix der Prädiktoren, Variance Inflation Factor (VIF)
- **Verletzung**: Prädiktorvariablen korrelieren stark miteinander
- **Konsequenz der Verletzung**: falsche Standardfehler, ineffiziente Schätzung
- **Lösung**: Ggf. Ausschluss von Prädiktorvariablen, falls uns Koeffizient der betroffenen Prädiktoren überhaupt interessiert.

MULTIKOLLINEARITÄT I

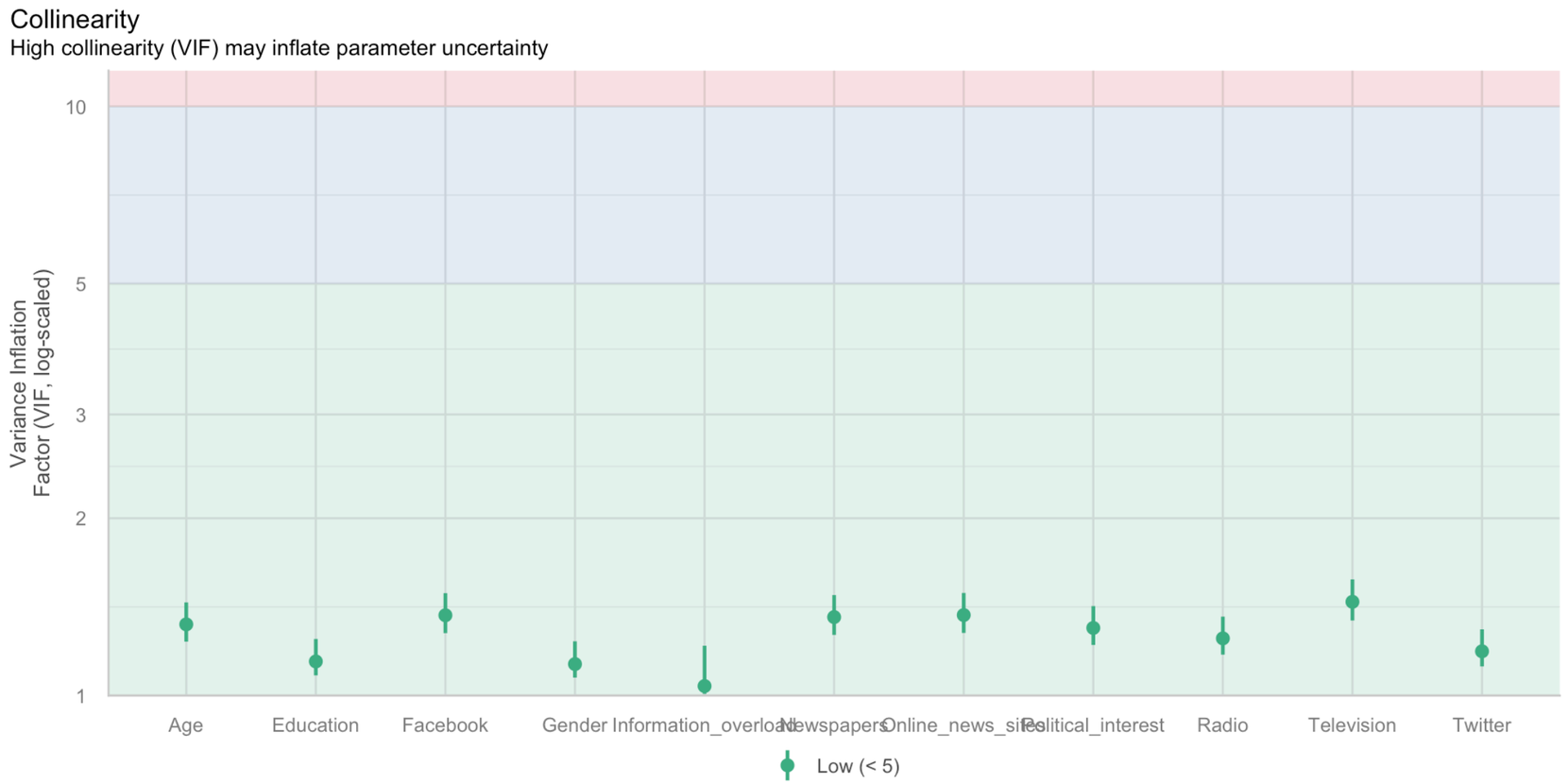
Correlation Matrix (pearson-method)

Parameter	Political_knowledge	Radio	Television	Newspapers	Online_news_sites	Twitter	Facebook	A
Political_knowledge		0.14***	0.26***	0.33***	0.22***	-0.04	-0.13***	0.
Radio	0.14***		0.39***	0.28***	0.22***	0.09	0.18***	(
Television	0.26***	0.39***		0.31***	0.28***	0.04	0.19***	0.
Newspapers	0.33***	0.28***	0.31***		0.33***	0.09	0.06	0.
Online_news_sites	0.22***	0.22***	0.28***	0.33***		0.23***	0.27***	-
Twitter	-0.04	0.09	0.04	0.09	0.23***		0.33***	-0.
Facebook	-0.13***	0.18***	0.19***	0.06	0.27***	0.33***		-0.
Age	0.30***	0.05	0.24***	0.21***	-0.08	-0.15***	-0.25***	
Political_interest	0.49***	0.20***	0.30***	0.33***	0.32***	0.05	0.06	0.
Information_overload	-0.09*	0.05	0.08	0.02	0.04	0.09	0.12**	(

p-value adjustment method: Holm (1979)

MULTIKOLLINEARITÄT II

```
$VIF
```



- Aber: Kalnins & Praitis Hill (2025): The VIF score. What is it good for? Absolutely nothing
- Korrelationen zwischen Prädiktoren sind zunächst ein empirischer Befund, mit dem wir umgehen müssen.

Fragen?

Übungsaufgaben

Fragen?

Nächste Einheit

~~Digitale Verhaltensdaten und Webtracking~~

Multiple lineare Regression III: Moderation

Danke — und erholsame Ferien 

Marko Bachl

marko.bachl@fu-berlin.de

LITERATUR

Arel-Bundock, V. (2025). *Model to meaning: How to interpret statistical models with R and Python* (1. Aufl.). Chapman; Hall/CRC. https://doi.org/10.1201/9781003560333

Bortz, J., & Schuster, C. (2010). *Statistik für Human- und Sozialwissenschaftler* (7.). Springer. https://doi.org/10.1007/978-3-642-12770-0

Kalnins, A., & Praitis Hill, K. (2025). The VIF score. What is it good for? Absolutely nothing. *Organizational Research Methods*, *28*(1), 58–75. https://doi.org/10/pffj

Van Erkel, P. F. A. (2020). *„Replication data for "Why don't we learn from social media?"* (Version V2) [Dataset]. Harvard Dataverse. https://doi.org/10.7910/DVN/D0COF1

Van Erkel, P. F. A., & Van Aelst, P. (2021). Why don't we learn from social media? Studying effects of and mechanisms behind social media news use on general surveillance political knowledge. *Political Communication*, *38*(4), 407–425. https://doi.org/10/ghk94s