Anwendungsorientierte Analyseverfahren

Multiple Regression

Prof. Dr. Michael Scharkow

Sommersemester 2024

Fragen zur praktischen Übung?

WIEDERHOLUNG: FACEBOOK-NEWS

Parameter	Coefficient	95% CI	t(520)	p	Std. Coef.	Fit
(Intercept)	3.51	(3.31, 3.72)	34.14	< .001	0.37	
modus (Post)	-0.72	(-1.04, -0.40)	-4.36	< .001	-0.55	
modus (Chronik)	-0.63	(-0.93, -0.34)	-4.27	< .001	-0.49	
modus (DM)	-0.68	(-0.97, -0.38)	-4.50	< .001	-0.52	
AICc						1742.69
R2						0.06
R2 (adj.)						0.05
Sigma						1.27

Outcome: Rezeptionswahrscheinlichkeit, Referenzkategorie: Tagging, Daten: Kümpel, 2019

MULTIPLE REGRESSION UND DAS GLM

- im GLM können mehrere Prädiktorvariablen in einem Modell kombiniert werden
- die Regressionskoeffizienten B sind wie sonst auch zu interpretieren, mit der Annahme, dass die anderen Prädiktoren sich nicht ändern (“ceteris paribus”)
- der Intercept B_0 ist der erwartete Wert von Y, wenn **alle** Prädiktoren $X = 0$ sind
- das R^2 ist die durch *alle* Prädiktoren erklärte Varianz in Y, d.h. nicht mehr nur die quadrierte Korrelation von X und Y
- der F-Test ist nun ein Omnibustest, d.h. er prüft, ob *irgendeine* Variable X einen signifikanten Zusammenhang mit Y hat

MULTIPLE VS. VIELE BIVARIATE REGRESSIONEN

- in klassischen Befragungsstudien haben wir oft mehrere plausible Prädiktoren
- eine multiple Regression vermeidete unnötig viele Einzeltests
- technisch ist es trivial, mehrere Prädiktoren ins Modell zu nehmen
- die Zusammenhänge zwischen Prädiktoren und Outcome werden unter Berücksichtigung der anderen Variablen im Modell geschätzt (Drittvariablenkontrolle)
- ein Modell mit mehreren Prädiktoren kann besser Y voraussagen als ein Modell mit weniger Prädiktoren

REGRESSIONSKOEFFIZIENTEN

- bei multiplen Regressionen sollten standardisierte und unstandardisierte Regressionskoeffizienten berichtet werden
- **unstandardisierte** Koeffizienten lassen sich
  1. in der Originalmetrik von X und Y interpretieren *und*
  2. sind beim Vergleich von Modellen mit verschiedenen Y, aber denselben X sinnvoll
- **standardisierte Koeffizienten** sind sinnvoll, um
  1. generell die Größe der Effekte abschätzen und
  2. *innerhalb* desselben Modells den relativen Einfluss verschiedener Variablen vergleichen zu können

DRITTVARIABLEN & MULTIKOLLINEARITÄT

- durch Hinzunahme einer weiteren X_2 Prädiktorvariable wird deren gemeinsamer Einfluss auf X_1 und Y in der Schätzung berücksichtigt
- eine statistische Berücksichtigung ist jedoch keinesfalls mit einer kausalen Berücksichtigung oder Konstanthaltung zu verwechseln
- wenn X_1 und X_2 untereinander korrelieren, sprechen wir von Multikollinearität
- obwohl die Schätzer B_1 und B_2 unverzerrt sind, wird die Präzision bei Multikollinearität geringer, d.h. die Standardfehler größer
- Multikollinearität ist *kein* statistisches Problem und kann daher auch nicht statistisch (z.B. durch Zentrierung) gelöst werden, sondern nur durch Änderungen in der Messung oder Variablenauswahl

MODELLGÜTE: R^2 UND F-TEST

- im bivariaten Fall entspricht das R^2 dem Determinationskoeffizienten r^2, also der quadrierten Korrelation
- alternative Herleitung als Verhältnis von erklärter (modellierter) und nicht erklärter (Residual-) Varianz
- Beispiel Nullmodell: keine Erklärungskraft, d.h. Residualvarianz $\text{var}(\epsilon) = \text{var}(Y)$, also Gesamtvarianz von Y
- $R2 = 1 - \text{var}(\epsilon)/\text{var}(Y)$, d.h. Anteil erklärter Varianz, den *alle* Prädiktoren zusammen ermöglichen
- F-Test: ist der Anteil erklärter Varianz signifikant von 0 verschieden

KORRIGIERTES R^2

- ein lineares Regressionsmodell wird durch Hinzunahme einer zusätzlichen Prädiktorvariable *nie* schlechter
- d.h. das naive R^2 kann durch zusätzliche Prädiktoren nur ansteigen oder sich schlimmstenfalls nicht (sichtbar) ändern, aber nie sinken
- würde man nun verschiedene Modelle miteinander vergleichen, würde das komplexere (= mehr Prädiktoren) Modell besser abschneiden, obwohl wir erkenntnistheoretisch eher an Modellsparsamkeit interessiert sind
- daher betrachten wir bei multiplen Regressionen immer ein (durch die Anzahl Prädiktoren k) korrigiertes $R^2 = 1 - (1 - R^2)\frac{n-1}{n-k-1}$

SCHRITTWEISE ODER HIERARCHISCHE REGRESSION

- manchmal werden Regressionsmodelle schrittweise geschätzt, d.h. einzelne Prädiktoren (oder Prädiktorenblöcke) nacheinander in das Modell eingeführt
- Differenz Δ im R^2 und/oder ein sog. partieller F-Test durchgeführt, der prüft, ob die Hinzunahme der Prädiktoren das Modell signifikant verbessert hat
- in der Schätzung der Regression gibt es **keine Reihenfolge-Effekte**, d.h. das finale Modell ist immer dasselbe, egal, ob man mit X_1 oder X_2 startet
- grundsätzlich nur die Regressionskoeffizienten aus dem finalen Modell interpretieren, das alle theoretisch postulierten Variablen enthält
- Vorteil schrittweise Testung: Übersichtlichkeit der Darstellung, Nachteil: unnötig viele Zwischenschritte, voreilige Interpretationen

BEISPIEL KÜMPEL, 2019

Table 2 Hierarchical OLS Regression Analysis Predicting Intention to Read the Article

	Intention to Read the Article		
Predictors	$r_{zero\text{-}order}$	$\beta_{upon\text{-}entry}$	β_{final}
Block 1 ($\Delta R^2_{adj.} = .17^{***}$)			
Gender[a]	.09	.06	.08
Age	-.10*	-.06	-.07
Education[b]	.05	.03	.04
Topical interest	.41***	.40***	.36***
Evaluation *Tagesschau*	.05	-.04	-.04
Duty to keep informed	.12*	.09	.08
Block 2 ($\Delta R^2_{adj.} = .08^{***}$)			
Tag[c]	.24***	.22***	.22***
Tie strength	.22***	.20***	.23***

PARTIELLER F-TEST

- globaler F-Test als Modellvergleich: Residualvarianz mein Modell vs. Nullmodell
- partieller F-Test: Modell 1 vs. Modell 2 (wobei Modell 2 auch alle Prädiktoren von 1 enthalten muss)
- Interpretation, wenn der F-Test signifikant ist: Modell 2 kann sig. mehr Varianz in Y erklären als Modell 1
- Modell 2 ist damit auch signifikant besser darin, Y vorauszusagen

KITCHEN-SINK REGRESSION

- oft ist es verführerisch, einfach möglichst viele (plausible) Prädiktoren ins Modell aufzunehmen
- Problem 1: Gefahr von Multikollinearität steigt, weil viele Variablen untereinander korrelieren
- Problem 2: durch falsche Einbeziehung von sog. Collider-Variablen, die von X und Y beeinflusst werden, werden die Schätzungen verzerrt (vgl. Sitzung zu Annahmen)
- Problem 3: sehr umfangreiche Regressionstabellen, die gelesen werden müssen

BEISPIEL: VAN ERKEL & VAN AELST, 2021

> Does exposure to news affect what people know about politics? This old question attracted new scholarly interest as the political informa- tion 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.

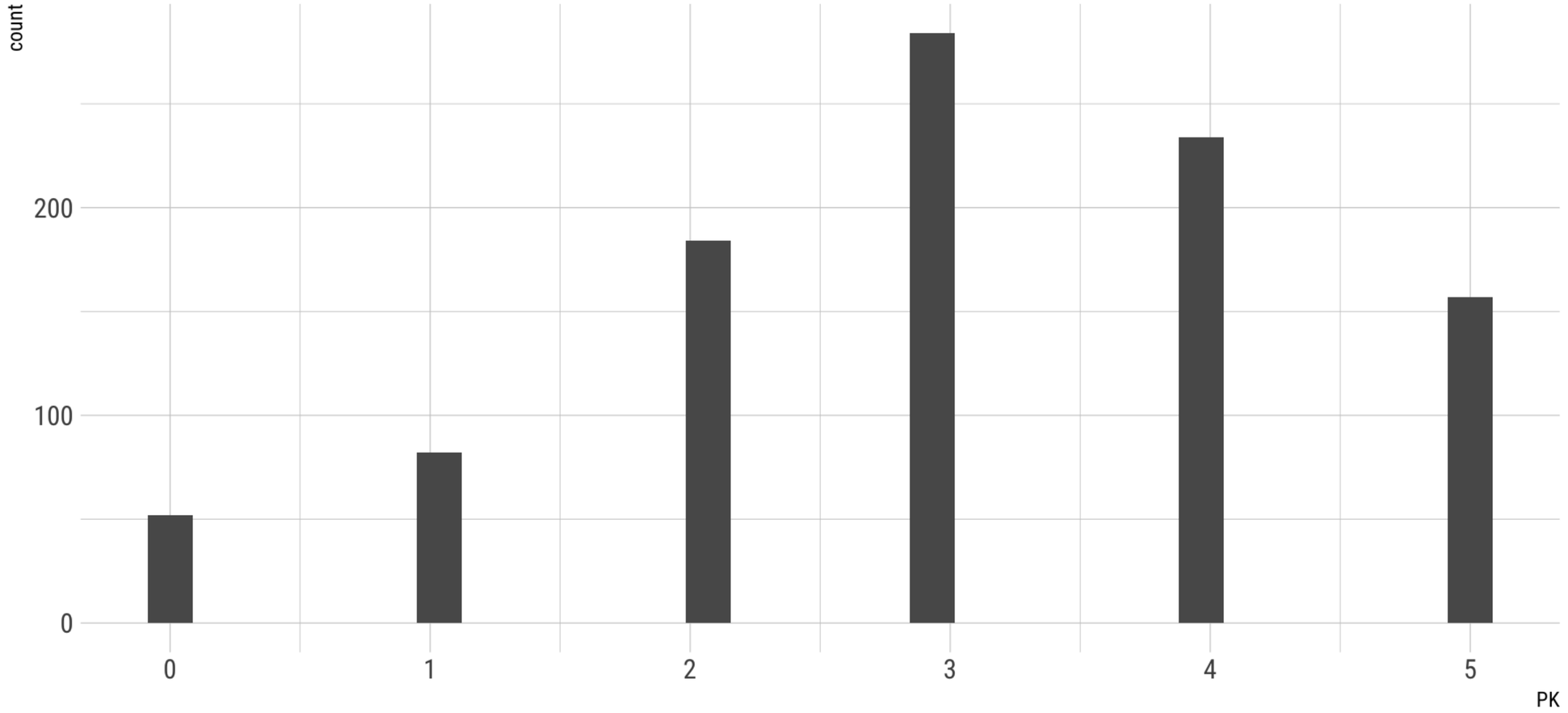
DATEN

Age	Gender	Education	TV	Newspaper	Websites	Facebook	PK
65	male	Lower	5	3	5	5	3
50	female	High	5	2	2	5	5
23	female	Middle	5	1	5	4	4
71	male	High	5	3	4	1	5
45	female	High	5	5	5	5	2

DESKRIPTIVSTATISTIK

Variable	Summary
Mean Age (SD)	52.98 (13.96)
Gender [female], %	47.7
Education [Lower], %	13.7
Education [Middle], %	40.7
Education [High], %	45.6
Mean TV (SD)	4.43 (1.33)
Mean Newspaper (SD)	3.52 (1.69)
Mean Websites (SD)	3.44 (1.72)
Mean Facebook (SD)	2.69 (1.95)
Mean PK (SD)	3.04 (1.36)

OUTCOME-VARIABLE



REGRESSIONSMODELL I (NUR SOZIODEMOGRAPHIE)

Parameter	Coefficient	95% CI	t(988)	p	Std. Coef.	Fit
(Intercept)	1.35	(0.96, 1.74)	6.81	< .001	-0.20	
Gender (female)	-0.73	(-0.89, -0.58)	-9.23	< .001	-0.54	
Age	0.03	(0.02, 0.03)	9.46	< .001	0.28	
Education (Middle)	0.51	(0.27, 0.75)	4.23	< .001	0.38	
Education (High)	0.89	(0.66, 1.13)	7.43	< .001	0.66	
AICc						3217.50
R2						0.20
R2 (adj.)						0.20
Sigma						1.22

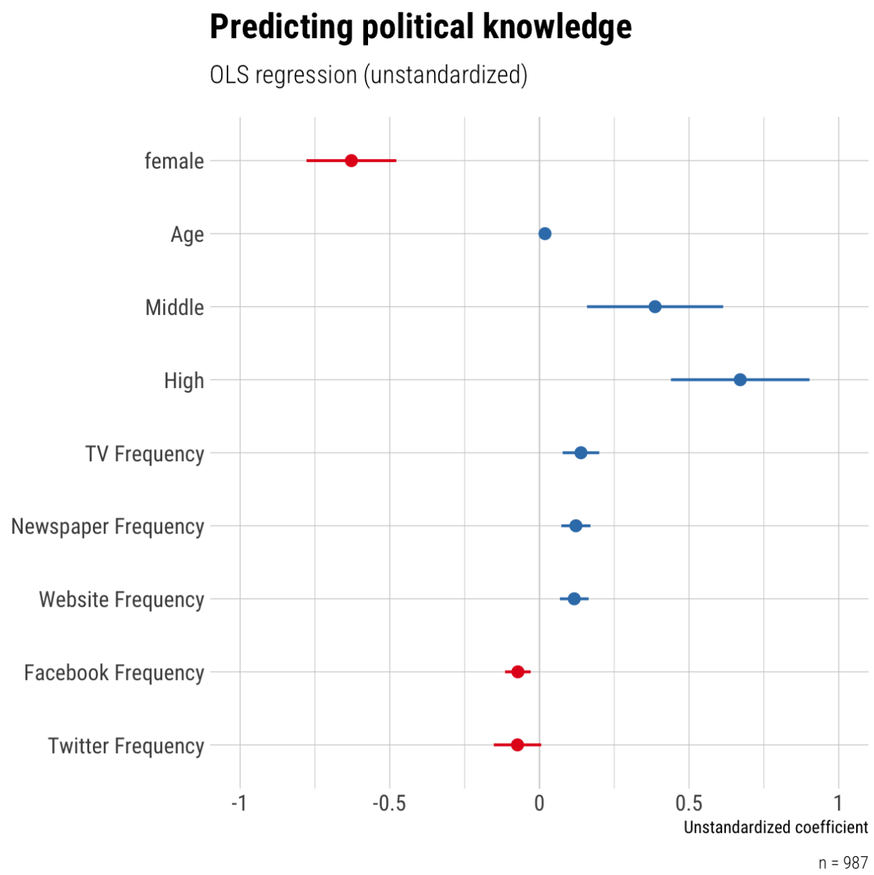
REGRESSIONSMODELL II (MEDIENNUTZUNG)

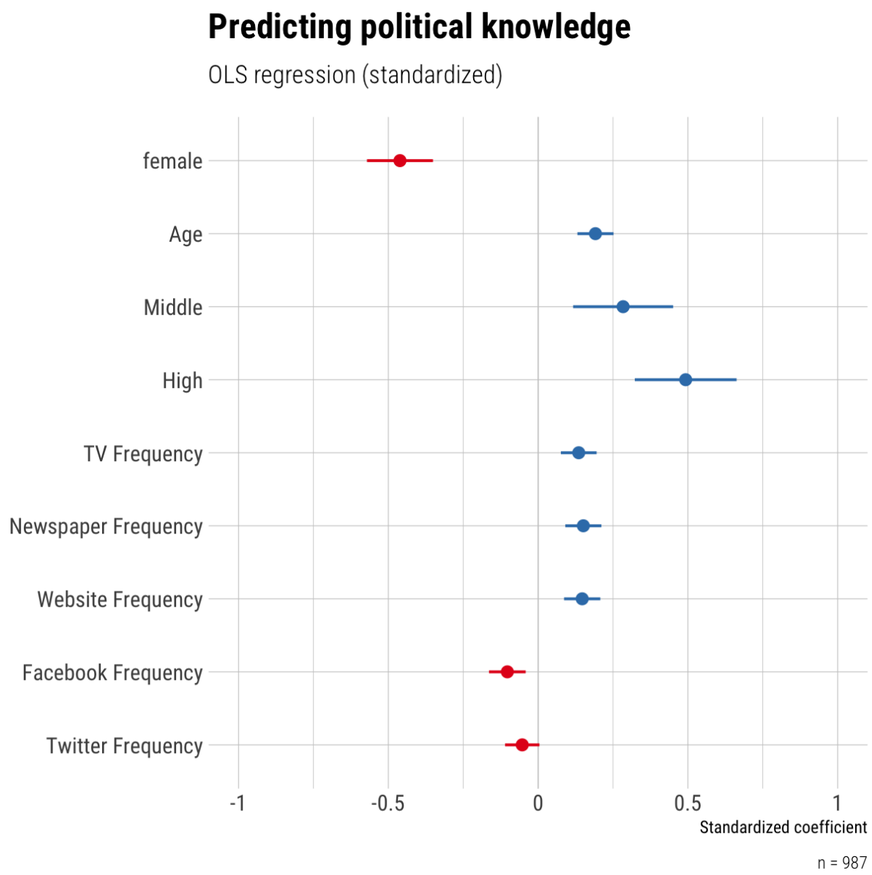
Parameter	Coefficient	95% CI	t(983)	p	Std. Coef.	Fit
(Intercept)	0.74	(0.28, 1.20)	3.18	0.002	-0.12	
Gender (female)	-0.63	(-0.78, -0.48)	-8.21	< .001	-0.46	
Age	0.02	(0.01, 0.02)	6.23	< .001	0.19	
Education (Middle)	0.39	(0.16, 0.61)	3.33	< .001	0.28	
Education (High)	0.67	(0.44, 0.90)	5.69	< .001	0.49	
TV	0.14	(0.08, 0.20)	4.46	< .001	0.14	
Newspaper	0.12	(0.07, 0.17)	4.91	< .001	0.15	
Websites	0.12	(0.07, 0.16)	4.77	< .001	0.15	
Facebook	-0.07	(-0.11, -0.03)	-3.29	0.001	-0.10	
Twitter	-0.07	(-0.15, 0.01)	-1.81	0.070	-0.05	
AICc						3114.32
R2						0.29
R2 (adj.)						0.28
Sigma						1.15

PARTIELLER F-TEST (MODELLVERBESSERUNG)

Modell	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	988	1466.83	NA	NA	NA	NA
2	983	1308.59	5	158.24	23.77	0

BONUS: VISUALISIERUNG DER ERGEBNISSE





Fragen?

LITERATUR

van Erkel, P. F., & 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.