Verallgemeinerte lineare Modelle

Methoden der empirischen Kommunikations- und Medienforschung

Marko Bachl
Freie Universität Berlin

ORGA

- Freitag, 3.7., 10 UHR, Online Q&A zur Klausur (Ersatz für 8. 6.)
- Bitte an Selbstreport der Hausaufgaben denken — Voraussetzung für Bestätigung der aktiven Teilnahme

Fragen zur letzten Sitzung?

Verallgemeinerte lineare Modelle

AGENDA (2 SITZUNGEN!)

1. Warum *verallgemeinerte* lineare Modelle?
2. Binäre Variablen als aV: Logistische Modelle
3. Zählvariablen als aV: Poisson- und negativ-binomiale Modelle
4. Fazit

WARUM *VERALLGEMEINERTE* LINEARE MODELLE?

Unsere **abhängige** Variable ist nicht metrisch und kontinuierlich, z.B.

- Binäre Variable (0; 1): falsch, richtig: kommt (nicht) vor
- Zählvariable (count): Anzahl (nicht-negative ganze Zahlen)

→ Funktion zwischen Prädiktoren und aV ist nicht unbedingt linear

→ Lineares Modell sagt Werte voraus, die es nicht geben kann

VAN ERKEL & VAN AELST (2021): WISSENSFRAGEN

Table 1. Overview of the political knowledge items, their coverage and the Mokken scale analysis.

	% correct answers	Item H	Newspaper articles	News articles on Facebook (number of shares)
Which two politicians were recently appointed by the king as informateurs to form a federal government? [National politics]	83.1%	0.53	162	28 (518)
Goedele Liekens [note: A Flemish celebrity) was elected in the Federal parliament? For which party?[National politics]	68.7%	0.44	60	13 (177)
Who was leading the investigation on the Russian interference in the past US elections?[International politics]	40.7%	0.54	71	15 (18)
During the campaign politicians spoke a lot about a minimum amount for pensions. What was this amount? [National politics]	87.9%	0.50	74	13 (146)
How did Groen [Note: the green party] score at the elections of May 26 2019 in comparison to the 2014 elections? [National politics]	62.8%	0.33	84	50 (1055)
In the Netherlands the PvdA (Labor party) won the European elections. Who was the list puller for this party in the European elections? [European elections].	24.1%	0.58	54	13 (74)

Newspaper articles: articles that appeared in 10 Flemish newspapers (incl. online versions) obtained via Gopress. News articles on Facebook: articles posted on Facebook by 18 different media outlets including digital only platforms, online version of newspapers and the news websites of the two main TV broadcasters. Data obtained via Crowdtangle. Between brackets: the total number of times these articles were shared by users.

(Van Erkel & Van Aelst, 2021)

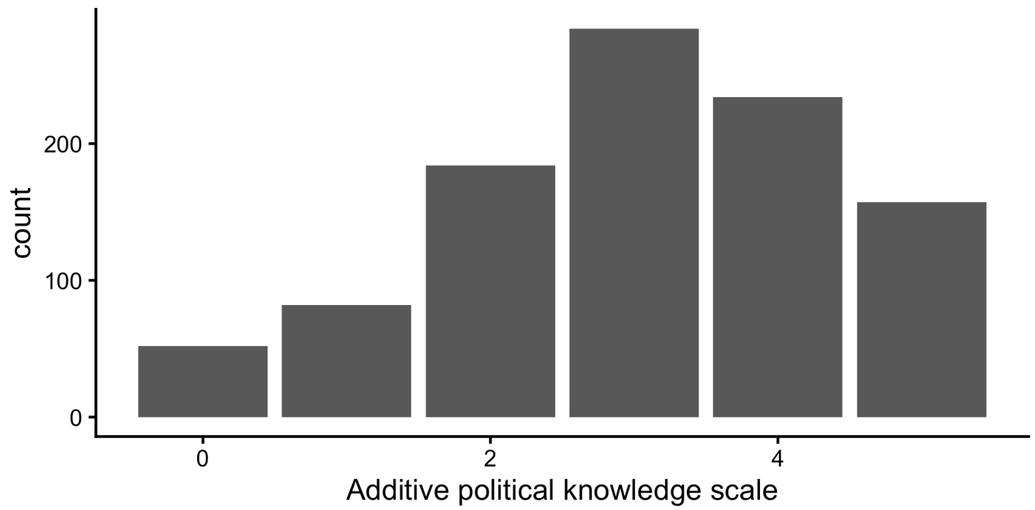
VAN ERKEL & VAN AELST (2021): WISSENSFRAGEN

Einzelne Fragen: binär

Descriptive Statistics

Variable	Summary
PK1 [correct], %	83.1
PK2 [correct], %	68.7
PK3 [correct], %	40.7
PK4 [correct], %	87.9
PK5 [correct], %	62.8
PK6 [correct], %	24.1

Wissensindex: Zählvariable



PK5 ist nicht Teil des Index (Van Erkel & Van Aelst, 2021, S. 414)

FÄHNRICH ET AL. (2020): SOCIAL MEDIA METRIKEN

id	likes_count	comments_count	shares_count
18058830773_315819235204024	105	9	10
140105122708_10152295975152709	16	3	2
103256838688_10153391508503689	330	10	30
49702985881_10151656671260882	10	0	3
16761458703_10153232709798704	161	1	6
374809010319_10151487539830320	1331	35	204
18058830773_10151884419680774	62	2	7
374809010319_10152497007995320	49	1	4
55528788113_10151241262778114	390	17	21
58323112191_10151982898231845	7	0	0

Fragen?

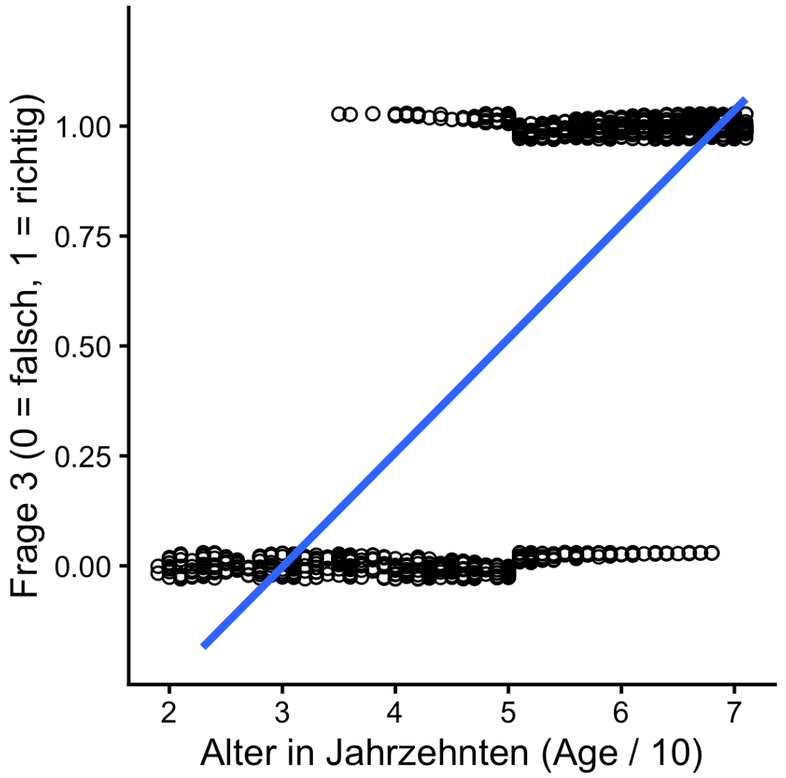
Binäre Variablen als aV: Logistische Modelle

Simuliertes Beispiel mit sehr starkem Zusammenhang!

LINEARE REGRESSION MIT BINÄRER AV

Simuliertes Beispiel!

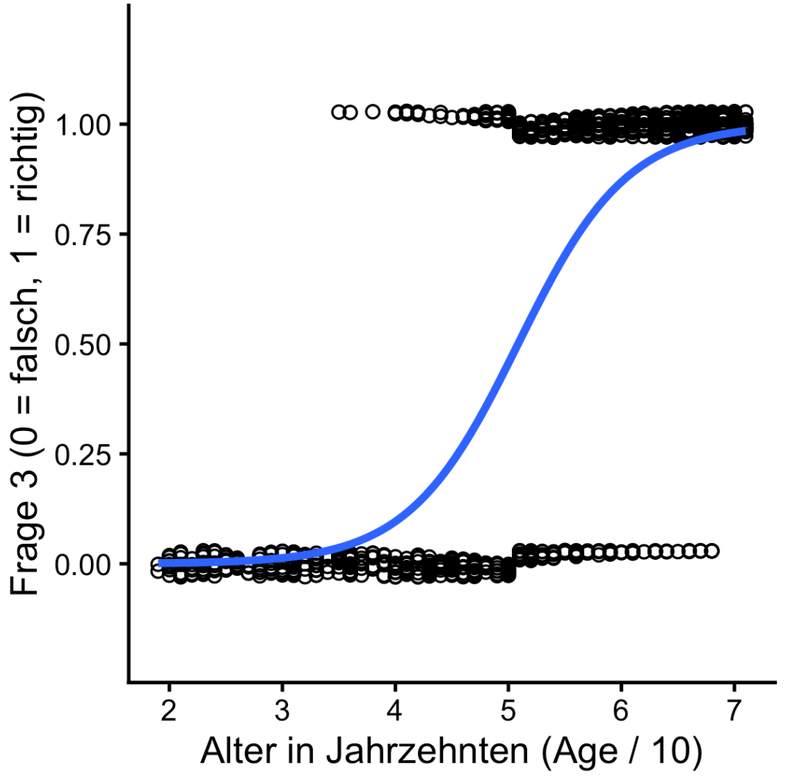
- Beispiel: Richtige Antwort auf *eine* Wissensfrage nach Alter (in Jahrzehnten)
- Vorhergesagte Wahrscheinlichkeiten außerhalb von [0; 1] möglich.
- Linearer Zusammenhang ist nicht gegeben, d.h. wir verstoßen gegen eine Annahme der linearen Regression
- Manchmal Einsatz als Linear Probability Model (LPM) mit einfacher Interpretation



LOGISTISCHE REGRESSION

Simuliertes Beispiel!

- aV: binär
- Logit-Funktion: $P(Y_i) = \frac{1}{1+e^{-(b_0+b_1X_i)}}$
- Die Logit-Funktion transformiert alle Werte in einen Bereich von 0-1.
- Die Werte von 0-1 können als Wahrscheinlichkeiten interpretiert werden.



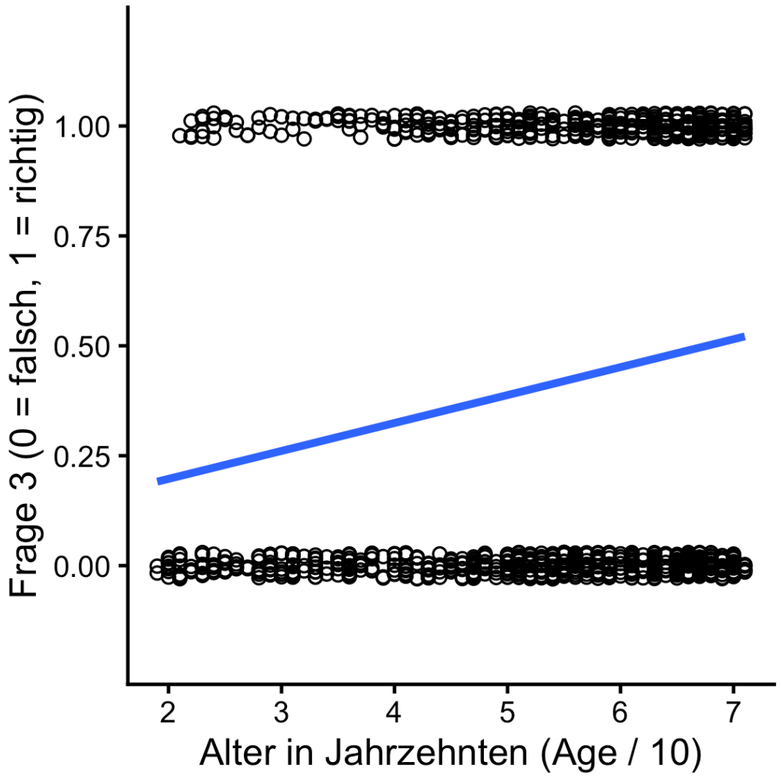
LINEARE VS. LOGISTISCHE REGRESSION

$$Logit(Y_i) = b_0 + b_1 X_i$$

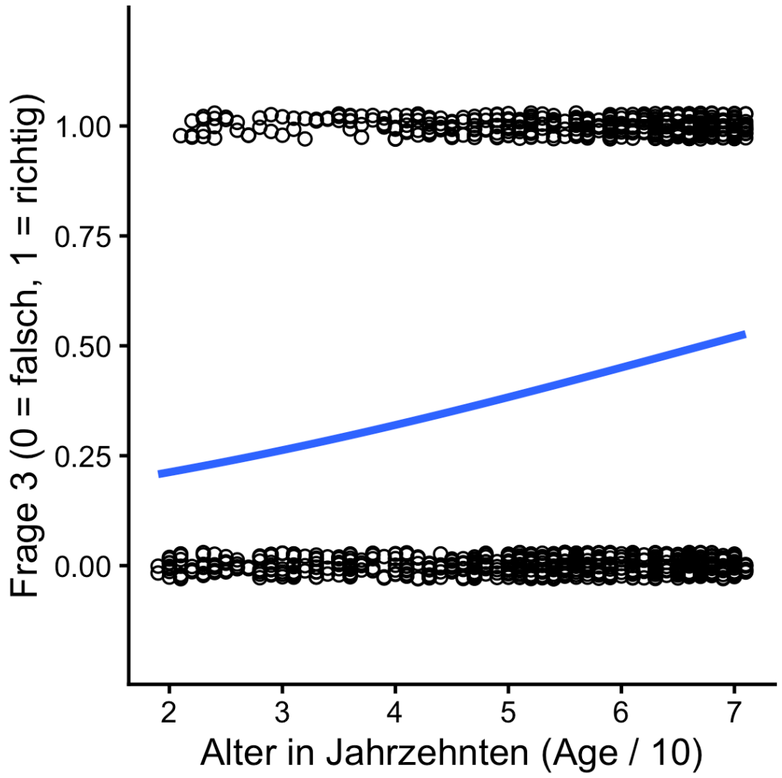
- Gleiche Annahmen zu Unabhängigkeit der Fälle und Multikollinearität
- Gleiche Logik der Modellspezifikation mit Nullmodell und Prädiktorvariablen
- Gleiche Logik der statistischen Inferenz (Standardfehler, Test-Statistik, Konfidenzintervalle)
- Statistisch unterschiedliche, konzeptionell ähnliche R^2-Maße
- Unterschiedliche Interpretation der Koeffizienten

DATEN AUS VAN ERKEL & VAN AELST (2021)

Lineare Regression



Logistische Regression



- Bei realistischen Zusammenhängen sind die Unterschiede weniger deutlich.

NULL-MODELL UND INTERCEPT

```r
m0 <- glm(
  PK3 ~ 1, # Null-Modell nur mit Intercept (Konstante)
  family = binomial(link = "logit"), # aV ist binär, logit-Link = Logistische Regression
  data = d
)
```

Parameter	Coefficient	95% CI	z	p	Fit
(Intercept)	-0.38	(-0.50, -0.25)	-5.84	< .001	
Tjur's R2					0

$$Logit(Y) = b_0$$

$$P(Y) = \frac{1}{1+e^{-(\beta_0)}}$$

Invers logit Formel

```r
1 / (1 + exp(--0.38))
```

```
[1] 0.4061269
```

Invers logit Funktion

```r
plogis(-0.38)
```

```
[1] 0.4061269
```

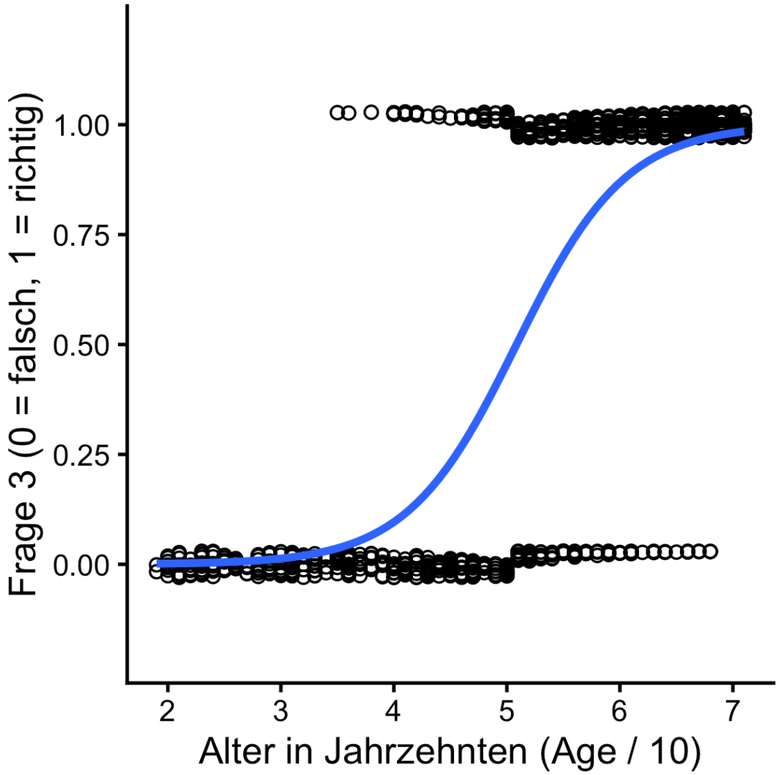
Descriptive Statistics

Variable	Summary
PK3 [correct], %	40.7

Fragen?

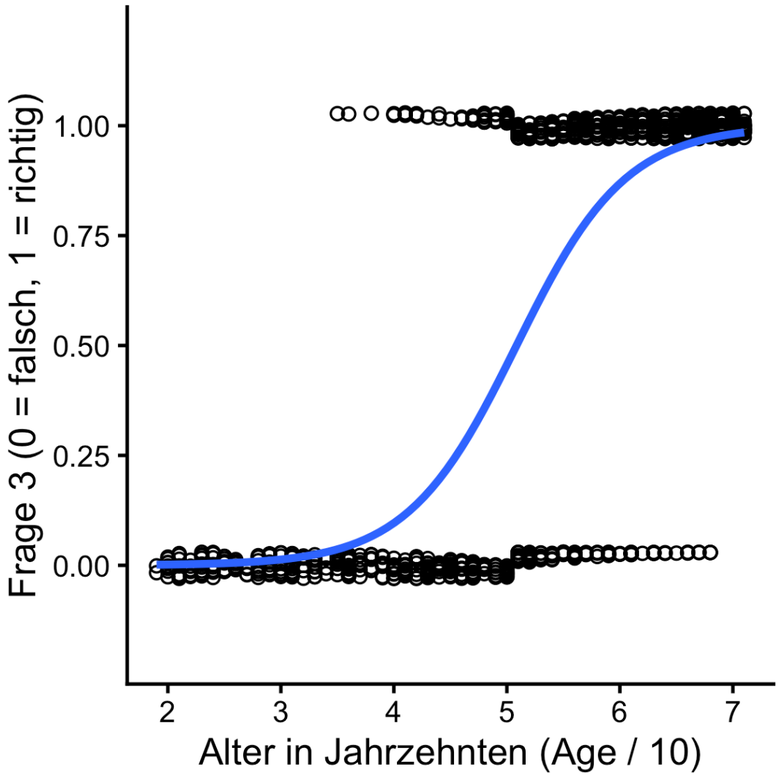
KOEFFIZIENTEN IN LOGISTISCHER REGRESSION

- $Logit(Y_i) = b_0 + b_1 X_i$
- Logit-Funktion: $P(Y_i) = \frac{1}{1+e^{-(b_0+b_1 X_i)}}$
- Koeffizienten sind bei logistischer Regression schwer direkt interpretierbar
- Interpretation als (Änderung von) Wahrscheinlichkeiten ist **falsch**, u.a. weil diese nicht konstant sind
- 3 Möglichkeiten, logistische Modelle zu interpretieren:
  - Geteilt-durch-4-Regel (Gelman & Hill, 2006, S. 82)
  - Average counterfactual comparison (Arel-Bundock, 2025) (siehe Einheit Multiple lineare Regression)
  - Odds Ratios (Gelman & Hill, 2006, S. 82–83)



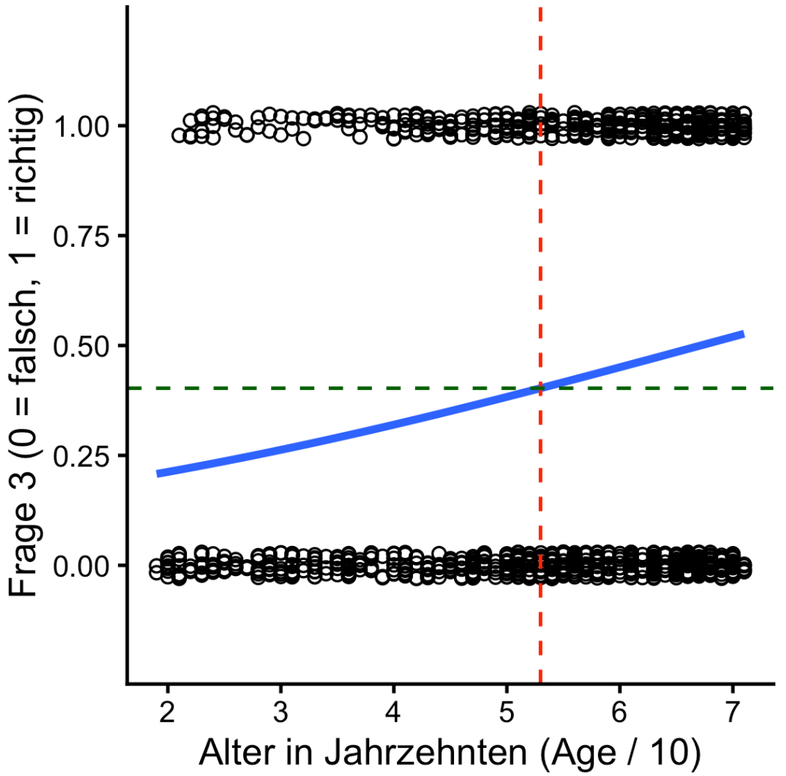
KOEFFIZIENTEN IN LOGISTISCHER REGRESSION

- Die Veränderung der Wahrscheinlichkeiten bei einer Änderung von X ist nicht in allen Fällen gleich groß.
- Interpretation im logistischen Modell hängt (auch) am Intercept und an den anderen Prädiktoren.
- Intercept durch Inverse Logit Funktion zu Baseline-Wahrscheinlichkeit umrechnen
- Geteilt-durch-4-Regel für schnelle Interpretation:
- $B/4$ ist eine *Obergrenze* für die Änderung in der Wahrscheinlichkeit $P(Y = 1)$, wenn X sich um eine Einheit ändert, in Prozentpunkten.
- Abweichung von dieser Regel am Rande der Verteilung von X größer als in der Mitte



MODELL MIT EINEM PRÄDIKTOR

```r
d <- d |>
  mutate(
    Age10_c = Age10 - mean(Age10)
  )
m1 <- glm(
  PK3 ~ Age10_c, # Regressionsgleichung
  family = binomial(link = "logit"), # aV ist binär, logit-Lir
  data = d
)
```



- Alle Prädiktoren, für die 0 kein sinnvoller Wert ist, werden um ihren Mittelwert zentriert, da wir den Intercept interpretieren wollen.

GETEILT-DURCH-4 REGEL

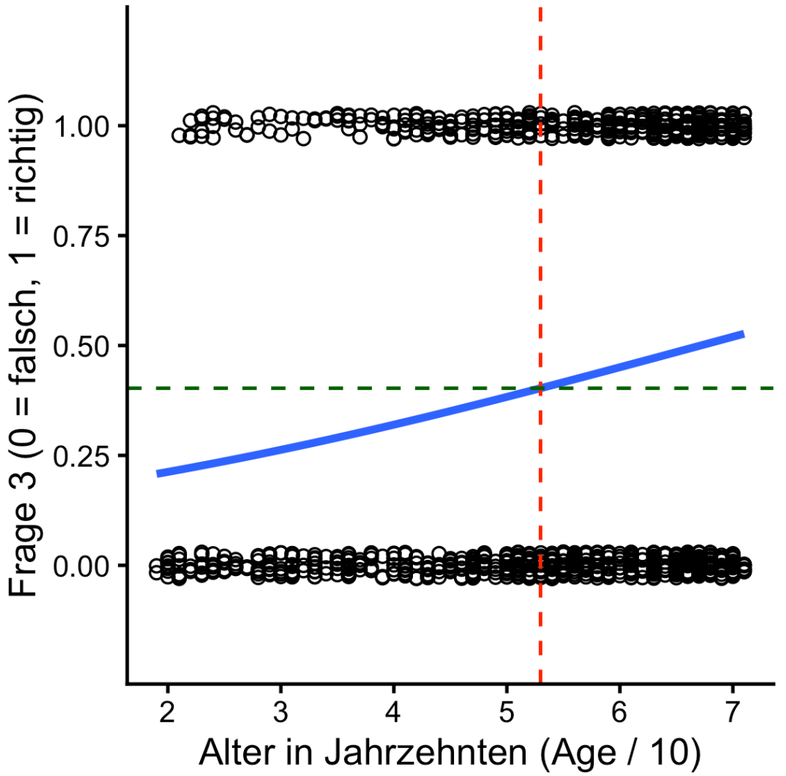
Parameter	Coefficient	95% CI	z	p	Fit
(Intercept)	-0.39	(-0.52, -0.26)	-5.97	< .001	
Age10 c	0.28	(0.18, 0.38)	5.60	< .001	
Tjur's R2					0.03

```
1 plogis(-0.39)
```

```
[1] 0.4037173
```

- Personen, die das durchschnittliche Alter 53 haben, beantworten die Frage mit einer Wahrscheinlichkeit von 40% richtig.

- **Geteilt-durch-4-Regel**: Wir vergleichen 2 Personen, deren Alter sich um 10 Jahre unterscheidet. Die ältere Person beantwortet die Frage mit einer ca. 7 Prozentpunkte (0.28/4) größeren Wahrscheinlichkeit richtig.



AVERAGE COUNTERFACTUAL COMPARISON

Um den Mittelwert

```
 Estimate Std. Error   z Pr(>|z|)    S 2.5 % 97.5 %
    0.066      0.011 5.9   <0.001 28.2 0.044  0.088

Term: Age10_c
Type: response
Comparison: +1
```

```
 Estimate Std. Error    z Pr(>|z|)    S  2.5 % 97.5 %
   −0.063       0.01 −6.1   <0.001 30.2 −0.083 −0.043

Term: Age10_c
Type: response
Comparison: +−1
```

An den Rändern der Verteilung

```
 Estimate Std. Error   z Pr(>|z|)    S 2.5 % 97.5 %
    0.049     0.0052 9.6   <0.001 69.9 0.039   0.06

Term: Age10_c
Type: response
Comparison: −2.397583081571 − −3.397583081571
```

```
 Estimate Std. Error   z Pr(>|z|)    S 2.5 % 97.5 %
    0.069      0.012 5.6   <0.001 25.4 0.045  0.094

Term: Age10_c
Type: response
Comparison: 1.802416918429 − 0.802416918429003
```

- Im Durchschnitt beantwortet eine um 10 Jahre ältere Person die Frage mit einer um 6-7 Prozentpunkten (um das mittlere Alter herum sowie im Vergleich der jüngeren Personen) bzw. 4 Prozentpunkten (Vergleich der älteren Personen) höheren Wahrscheinlichkeit richtig.

ODDS RATIOS — EXP(B)

- Häufig wird der Wert Exp(B) berichtet, auch als Odds Ratio (OR) bekannt
- Odds = $\frac{P(Y=1)}{1-P(Y=1)}$ = “Chance” oder “Risiko”, nicht “Wahrscheinlichkeit”!
- Exp(B) = 1 bedeutet kein Unterschied, Exp(B) > 1 bedeutet eine *erhöhte(s)* Chance oder Risiko, Exp(B) < 1 eine *niedrigere(s)* Chance oder Risiko
- Exp(B) = .5 heißt: Mit jeder Einheit mehr X besteht ein halb so großes Risiko
- Exp(B) = 2 heißt: Mit jeder Einheit mehr X verdoppelt sich das Risiko

ODDS RATIOS — EXP(B)

```
1 m1 |>
2   report_table(metrics = "R2", include_effectsize = FALSE, exponentiate = TRUE)
```

Parameter	Coefficient	95% CI	z	p	Fit
(Intercept)	0.67	(0.59, 0.77)	-5.97	< .001	
Age10 c	1.32	(1.20, 1.46)	5.60	< .001	
Tjur’s R2					0.03

- Wir vergleichen 2 Personen, deren Alter sich um 10 Jahre unterscheidet. Die ältere Person hat eine etwa 1.3-mal so große Chance, die Frage richtig zu beantworten.
- Vorsicht: Der exponenzierte Intercept hat hier keine interpretierbare Bedeutung.

Fragen?

MODELL MIT MEHREREN PRÄDIKTOREN

```r
d <- d |>
  mutate(
    Online_news_sites_c = Online_news_sites - mean(Online_news_sites),
    Twitter_c = Twitter - mean(Twitter),
    Facebook_c = Facebook - mean(Facebook),
    Age10_c = Age10 - mean(Age10)
  )
m2 <- glm(
  PK3 ~ Online_news_sites_c + Twitter_c + Facebook_c + Gender + Age10_c, # Regressionsgleichung
  family = binomial(link = "logit"), # aV ist binär, logit-Link = Logistische Regression
  data = d
)
```

- Alle Prädiktoren, für die 0 kein sinnvoller Wert ist, werden um ihren Mittelwert zentriert, da wir den Intercept interpretieren wollen.

GETEILT-DURCH-4 REGEL

Parameter	Coefficient	95% CI	z	p	Fit
(Intercept)	0.13	(-0.05, 0.31)	1.42	0.157	
Online news sites c	0.19	(0.11, 0.28)	4.50	< .001	
Twitter c	-0.03	(-0.18, 0.12)	-0.42	0.675	
Facebook c	-0.11	(-0.19, -0.04)	-2.82	0.005	
Gender (female)	-1.20	(-1.48, -0.92)	-8.35	< .001	
Age10 c	0.19	(0.09, 0.30)	3.55	< .001	
Tjur's R2					0.14

```
1 plogis(0.13)
```

```
[1] 0.5324543
```

- Männer mit durchschnittlichem Alter und durchschnittlicher Online-Medien-Nutzung beantworten die Frage mit einer Wahrscheinlichkeit von 53% richtig.

- Wir vergleichen 2 Personen, deren Nutzung von Online-Nachrichten-Seiten sich um einen Skalenpunkt unterscheidet und die auf den übrigen Prädiktoren identische Werte aufweisen. Die Person, die häufiger Online-Nachrichten nutzt, beantwortet die Frage mit einer ca. 5 Prozentpunkte (0.19/4) größeren Wahrscheinlichkeit richtig.
- Frauen beantworten die Frage mit einer um ca. 30 Prozentpunkte (1.20/4) geringeren Wahrscheinlichkeit richtig als gleich alte Männer mit identischer Online-Mediennutzung.

AVERAGE COUNTERFACTUAL COMPARISON

```
                Term      Contrast Estimate Std. Error     z Pr(>|z|)    S  2.5 %  97.5 %
 Age10_c             +1              0.0406     0.0113  3.60   <0.001 11.6  0.019  0.0627
 Facebook_c          +1             −0.0235     0.0081 −2.89   0.0039  8.0 −0.039 −0.0076
 Gender              female − male  −0.2664     0.0306 −8.69   <0.001 58.0 −0.326 −0.2063
 Online_news_sites_c +1              0.0406     0.0088  4.63   <0.001 18.0  0.023  0.0578
 Twitter_c           +1             −0.0067     0.0159 −0.42   0.6745  0.6 −0.038  0.0245

Type: response
```

- Personen mit einer Online-Nachrichten-Seitennutzung von einem Skalenpunkt über dem Stichprobenmittelwert beantworten die Frage mit einer um 4 Prozentpunkten größeren Wahrscheinlichkeit richtig als Personen mit durchschnittlicher Online-Nachrichten-Seitennutzung.
- Im Vergleich zu Männern beantworten Frauen die Frage mit einer 27 Prozentpunkte geringeren Wahrscheinlichkeit richtig.

ODDS RATIOS — EXP(B)

```
m2 |>
  report_table(metrics = "R2", include_effectsize = FALSE, exponentiate = TRUE)
```

Parameter	Coefficient	95% CI	z	p	Fit
(Intercept)	1.14	(0.95, 1.36)	1.42	0.157	
Online news sites c	1.21	(1.12, 1.32)	4.50	< .001	
Twitter c	0.97	(0.83, 1.12)	-0.42	0.675	
Facebook c	0.89	(0.82, 0.97)	-2.82	0.005	
Gender (female)	0.30	(0.23, 0.40)	-8.35	< .001	
Age10 c	1.21	(1.09, 1.35)	3.55	< .001	
Tjur’s R2					0.14

- Wir vergleichen 2 Personen, deren Nutzung von Online-Nachrichten-Seiten sich um einen Skalenpunkt unterscheidet und die auf den übrigen Prädiktoren identische Werte aufweisen. Die Person, die häufiger Online-Nachrichten nutzt, hat eine 1.2-mal so große Chance, die Frage richtig zu beantworten.
- Frauen haben eine 0.3-mal so große Chance, die Frage richtig zu beantworten, wie gleich alte Männer mit identischer Online-Mediennutzung.

Fragen?

MODELLGÜTE: PSEUDO R^2

- Pseudo R^2 vergleichen die Passung des Null-Modells mit zu beurteilendem Modell.
- Verbreitet: McFadden, Nagelkerke, Cox & Snell, Tjur
- Empfehlung für logistische Regression: Tjur's R^2
- Interpretation wie R^2 in linearer Regression, wobei es je nach Maß und Modell sein kann, dass 0 und/oder 1 nicht erreicht werden können.

MODELLGÜTE: PSEUDO R^2

Parameter	Coefficient	95% CI	z	p	Fit
(Intercept)	0.13	(-0.05, 0.31)	1.42	0.157	
Online news sites c	0.19	(0.11, 0.28)	4.50	< .001	
Twitter c	-0.03	(-0.18, 0.12)	-0.42	0.675	
Facebook c	-0.11	(-0.19, -0.04)	-2.82	0.005	
Gender (female)	-1.20	(-1.48, -0.92)	-8.35	< .001	
Age10 c	0.19	(0.09, 0.30)	3.55	< .001	
Tjur’s R2					0.14

The model’s explanatory power is moderate (Tjur’s R2 = 0.14)

Fragen?

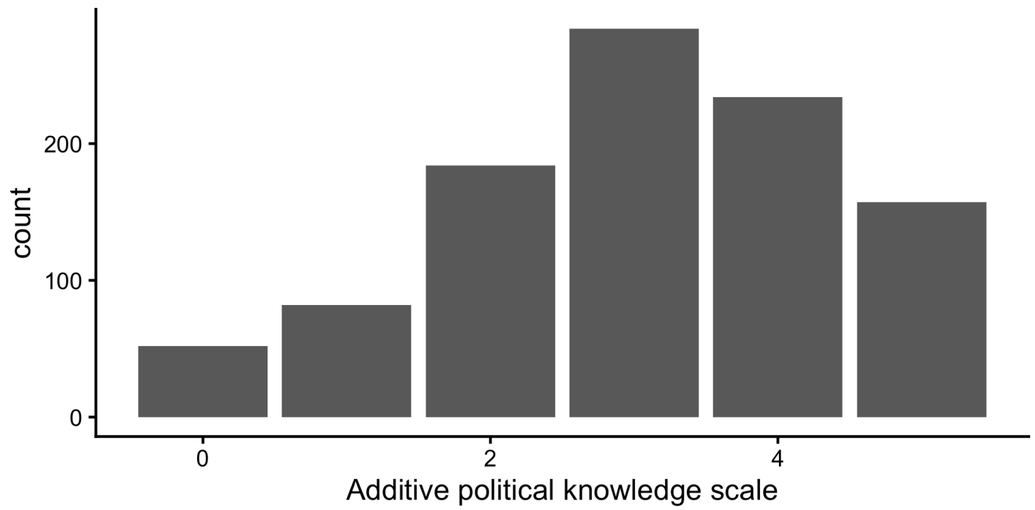
VAN ERKEL & VAN AELST (2021): WISSENSFRAGEN

Einzelne Fragen: binär

Descriptive Statistics

Variable	Summary
PK1 [correct], %	83.1
PK2 [correct], %	68.7
PK3 [correct], %	40.7
PK4 [correct], %	87.9
PK5 [correct], %	62.8
PK6 [correct], %	24.1

Wissensindex: Zählvariable



PK5 ist nicht Teil des Index (Van Erkel & Van Aelst, 2021, S. 414)

AUSBLICK: BINOMIALE ZÄHLVARIABLE

```r
m5q_v1 <- glm(
  cbind(Political_knowledge, 5 - Political_knowledge) ~ Age10_c, # Zahl der richtig und falsch beantwortete
  family = binomial(link = "logit"), # eine Frage ist binär, logit-Link = Logistische Regression
  data = d
)

m5q_v2 <- glm(
  Political_knowledge / 5 ~ Age10_c, # Anteil der richtig beantworteten Fragen als aV
  family = binomial(link = "logit"), # eine Frage ist binär, logit-Link = Logistische Regression
  weights = rep(5, nrow(d)), # Zahl der Fragen als Gewicht
  data = d
)
```

- Besondere Art der Zählvariable: ganze, nicht-negative Zahl *mit bekanntem Maximum* = Erfolge in einer bekannten Zahl von binären Tests
- Modellierung des Index Politisches Wissen (Zahl der richtigen Fragen, [0; 5]) als Funktion der Prädiktoren.
- Dazu entweder Eingabe der Zahl der richtig und falsch beantworteten Fragen als aV oder Anteil der richtig beantworteten Fragen als aV und Zahl der Fragen als Gewichte
- Es wird die Wahrscheinlichkeit geschätzt, **eine** weitere “durchschnittliche” Frage richtig zu beantworten.

AUSBLICK: BINOMIALE ZÄHLVARIABLE

Parameter	Coefficient	95% CI	z	p	Fit
(Intercept)	0.45	(0.40, 0.51)	15.34	< .001	
Age10 c	0.25	(0.21, 0.29)	11.78	< .001	

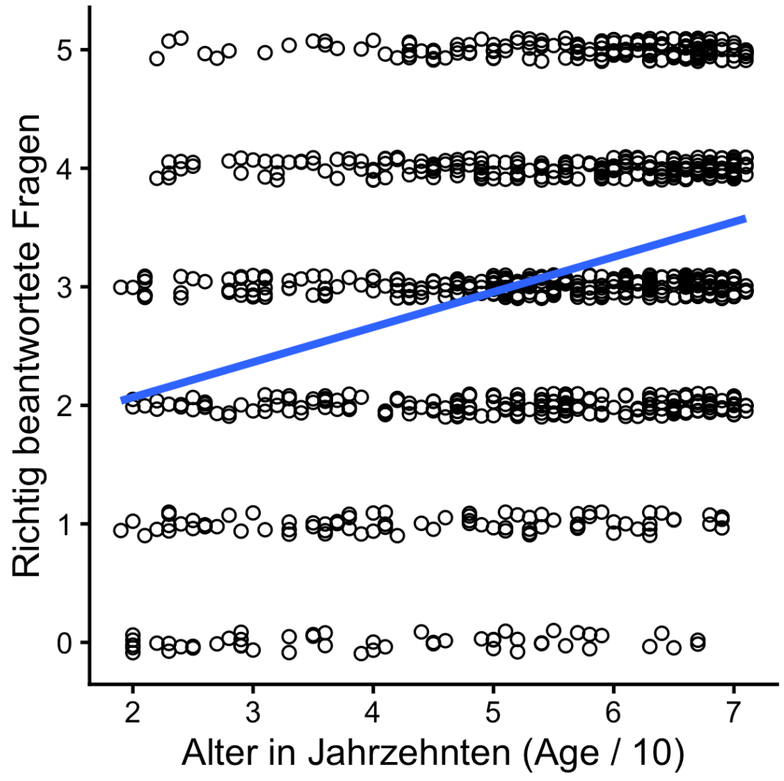
```
1  round(plogis(coef(m5q_v1))[1], 2)
```

```
(Intercept)
       0.61
```

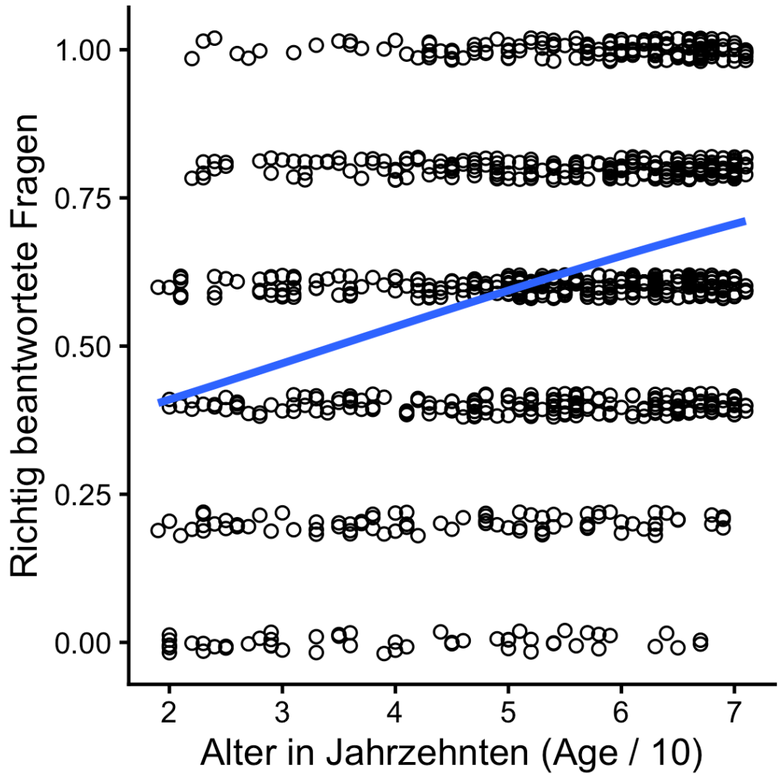
- **Invers logit des Intercept**: Eine Person mit durchschnittlichem Alter beantwortet eine “durchschnittliche” Frage mit einer Wahrscheinlichkeit von 61% richtig.
- **Geteilt-durch-4-Regel**: Wir vergleichen 2 Personen, deren Alter sich um 10 Jahre unterscheidet. Die ältere Person beantwortet eine “durchschnittliche” Frage aus dem Test mit einer ca. 6 Prozentpunkte (0.25/4) größeren Wahrscheinlichkeit richtig.

AUSBLICK: BINOMIALE ZÄHLVARIABLE

Lineare Regression



Logistische Regression



- Hier macht es praktisch kein Unterschied.

Fragen?

Zählvariablen als aV: Poisson- und negativ-binomiale Modelle

POISSON- UND NEGATIV-BINOMIALE MODELLE

- Zählvariablen: nicht-negative ganze Zahlen; Typisch für Variablen, bei denen etwas gezählt wird und keine natürliche Grenze nach oben besteht; z.B. Social-Media-Metriken, Mediennutzungsdauer und -häufigkeit.

- Zwei verbreitete Modelle:
  - Poisson-Verteilung: Annahme Erwarteter Mittelwert = Erwartete Varianz
  - Negative Binomial-Verteilung: Zusätzlicher Dispersionsparameter, damit erwarteter Mittelwert von erwarteter Varianz abweichen darf.

- Wichtige Begriffe: Über- bzw. Unter**dispersion** = Fehlende Passung der theoretisch erwarteten Varianz zur tatsächlichen Streuung der Datenpunkte um ihre Vorhersagewerte.

POISSON- UND NEGATIV-BINOMIALE MODELLE

- Vereinfachte Entscheidungsregeln:
  - Entscheidung lässt sich erst anhand der geschätzten Modelle prüfen. Deskriptives Verhältnis von Mittelwert zu Varianz kann aber häufig schon auf angemessenes Modell hinweisen.
  - Vor allem im Bereich von Social-Media-Daten ist die Annahme der Poisson-Regression sehr selten erfüllt. Daher müssen wir die negativ-binomiale Regression zumindest im Vergleich testen; Die Interpretation unterscheidet sich am Ende aber kaum.

Beispiel: Zahl der Kommentare zu Posts der UC San Francisco

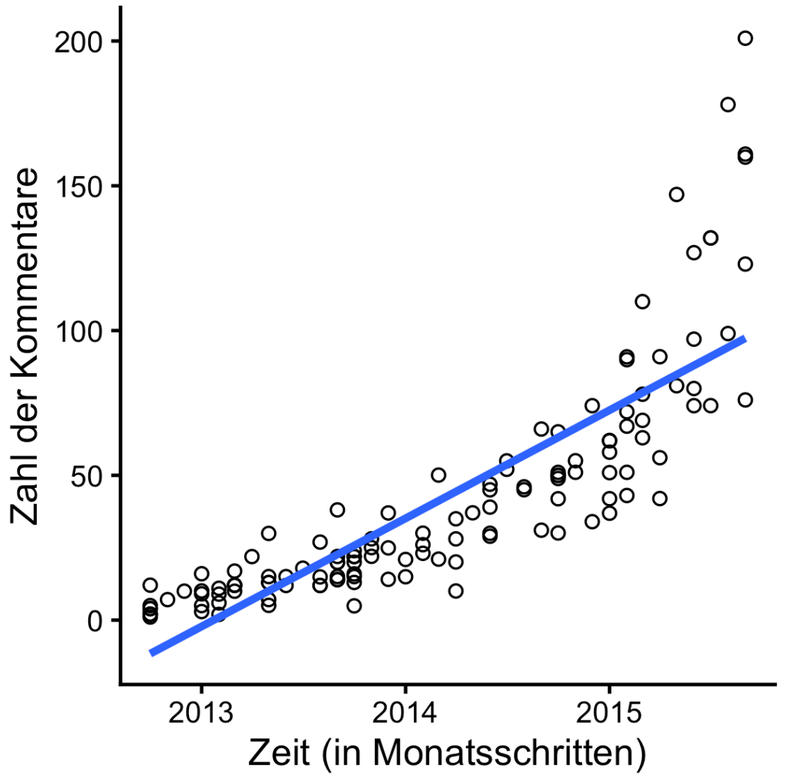
M	Var	n
3.55	28.85	128

Simuliertes Beispiel mit bekanntem datengenerierenden Poisson-Prozess!

LINEARE REGRESSION MIT ZÄHLVARIABLE ALS AV

Simuliertes Beispiel!

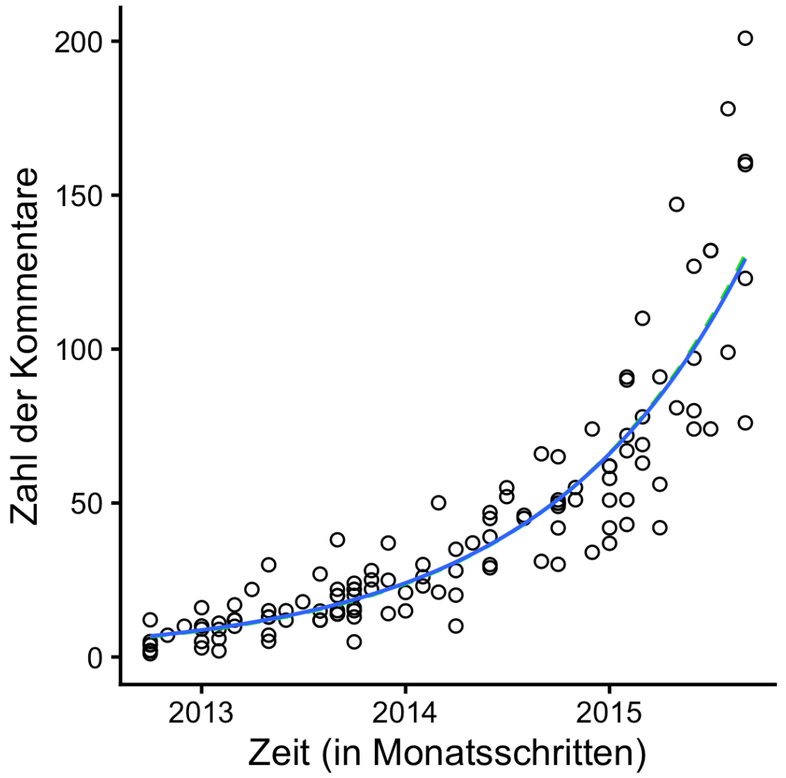
- Beispiel: Zahl der Kommentare zu Posts der UC San Francisco im Zeitverlauf (Monate)
- Vorhergesagte Anzahl kleiner als 0.
- Linearer Zusammenhang und Homoskedastizität sind nicht gegeben, d.h. wir verstoßen gegen Annahmen der linearen Regression.



POISSON- UND NEGATIV-BINOMIALE REGRESSION

Simuliertes Beispiel!

- aV: Zähldaten (nicht-negative ganze Zahlen)
- Log-Link-Funktion:
  - $\log(Y_i) = b_0 + b_1 X_i$
  - $Y_i = e^{b_0 + b_1 X_i}$
- Die Exponentialfunktion transformiert die Vorhersagewerte in einen Bereich von 0 bis ∞.



LINEARE VS. POISSON- UND NEGATIV-BINOMIALE REGRESSION

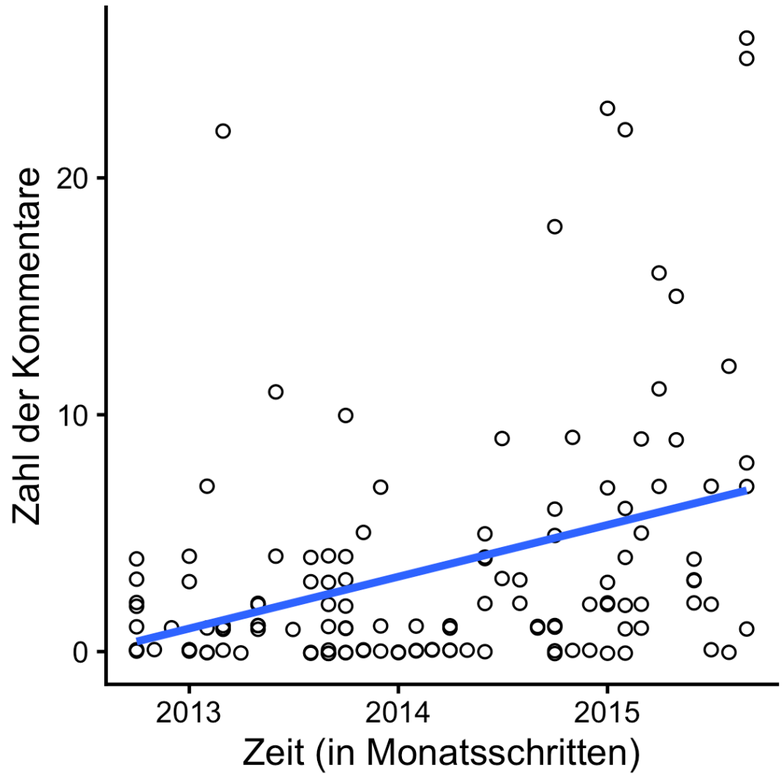
$$\log(Y_i) = b_0 + b_1 X_i$$

$$Y_i = e^{b_0 + b_1 X_i}$$

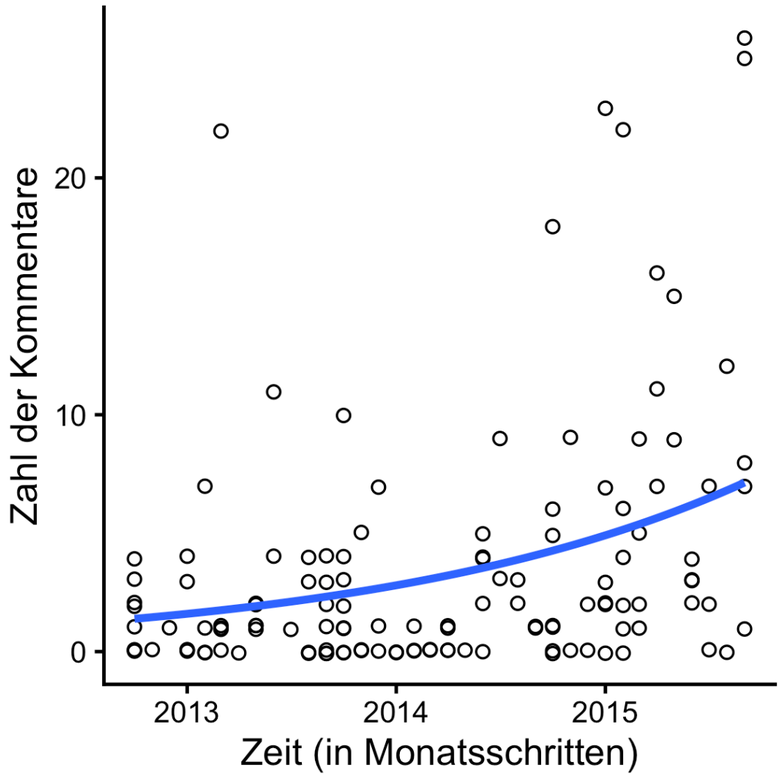
- Gleiche Annahmen zu Unabhängigkeit der Fälle und Multikollinearität
- Spezifische Verteilungsannahme: Varianz entspricht dem erwarteten Mittelwert (Poisson) bzw. dem Mittelwert mal einem Dispersionsparameter (Negativ-binomial))
- Gleiche Logik der Modellspezifikation mit Nullmodell und Prädiktorvariablen
- Gleiche Logik der statistischen Inferenz (Standardfehler, Test-Statistik, Konfidenzintervalle)
- Statistisch unterschiedliche, konzeptionell ähnliche Pseudo-R^2-Maße
- Unterschiedliche Interpretation der Koeffizienten (multiplikative Effekte nach Exponenzieren: e^{b_1})

DATEN AUS FÄHNRICH ET AL. (2020)

Lineare Regression



Negativ-binomiale Regression



- Bei realistischen Zusammenhängen sind die Unterschiede weniger deutlich.

NULL-MODELL UND INTERCEPT

```r
nbm0 <- glm.nb( # Negativ-binomiale Regression
  comments_count ~ 1, # Null-Modell nur mit Intercept (Konstante)
  data = d2_RU
)
```

Parameter	Coefficient	95% CI	z	df	p	Fit
(Intercept)	1.27	(1.02, 1.52)	10.06	Inf	< .001	
R2_Nagelkerke						0.00

$$\log(Y_i) = b_0$$

$$Y_i = e^{b_0}$$

Intercept exponenzieren

```r
exp(coef(nbm0))
```

```
(Intercept)
   3.554687
```

Mittelwert der Variable

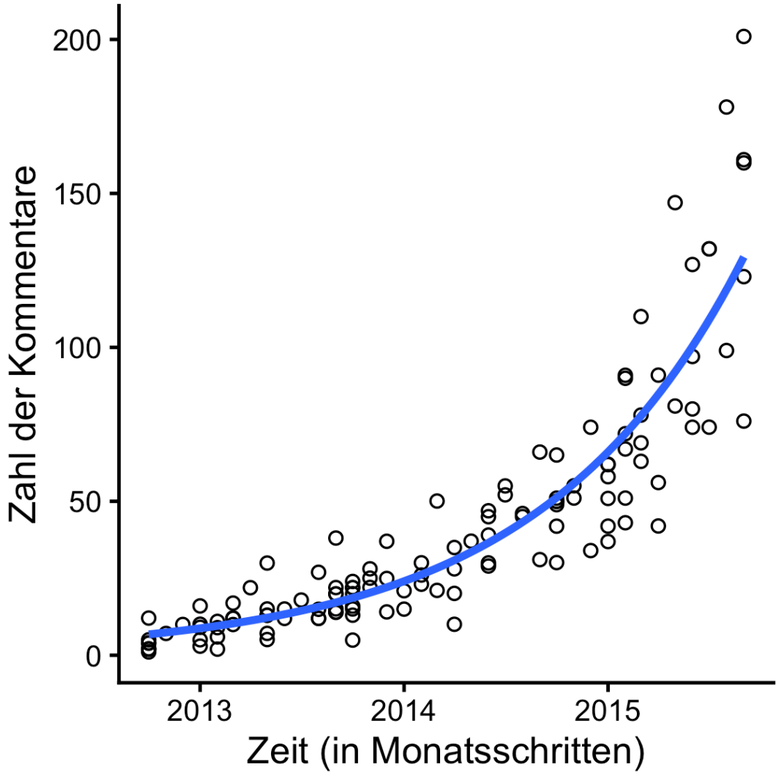
```r
mean(d2_RU$comments_count)
```

```
[1] 3.554688
```

Fragen?

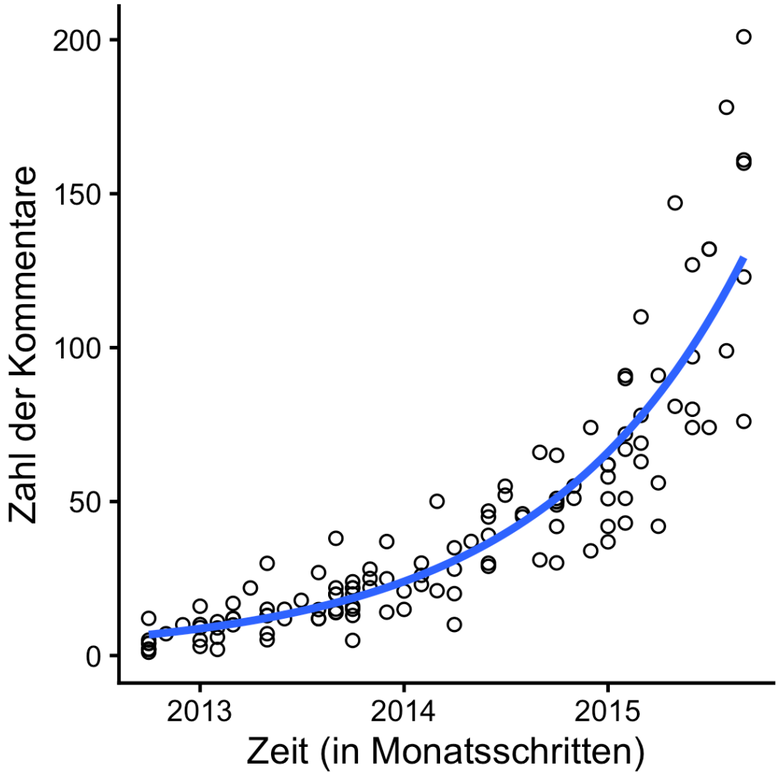
KOEFFIZIENTEN IN DER POISSON- UND NEGATIV-BINOMIALEN REGRESSION

- Koeffizienten sind bei Modellen mit Log-Link schwer direkt interpretierbar, da multiplikativer Effekt vom Ausgangsniveau abhängt.
- 2 Möglichkeiten, Modelle mit Log-Link zu interpretieren:
  - Exponenzierter Koeffizient (e^{b_1}) als Wachstumsfaktor, $e^{b_1} - 1$ als Wachstumsrate
  - Average counterfactual comparison (Arel-Bundock, 2025) (siehe Einheit Multiple lineare Regression)



NICHT STANDARDISIERTE KOEFFIZIENTEN

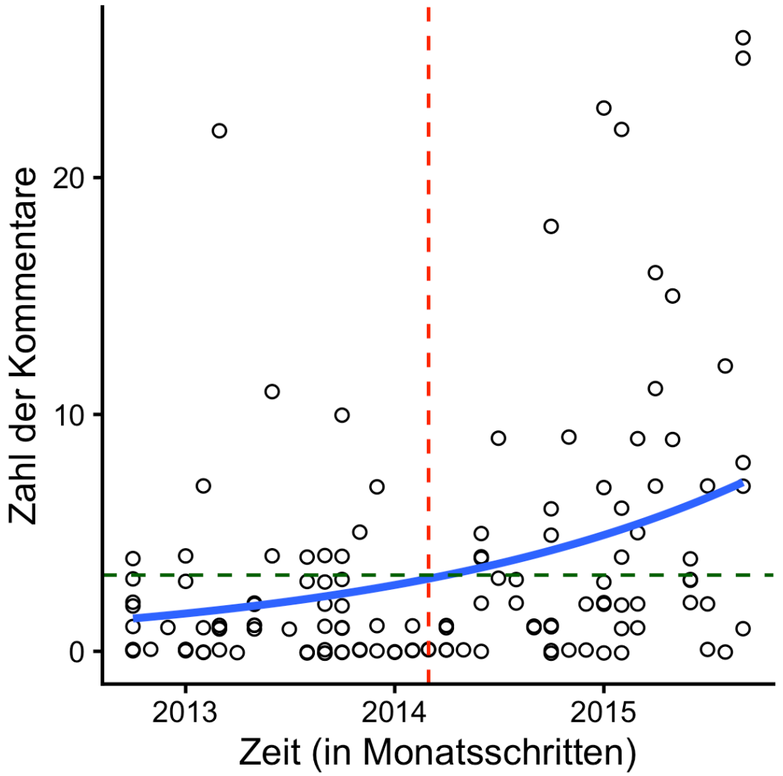
- Die Veränderung der aV bei einer Änderung von X ist nicht in allen Fällen gleich groß.
- Interpretation hängt (auch) am Intercept und an den anderen Prädiktoren.
- Intercept durch Exponential-Funktion zu Baseline umrechnen
- Koeffizienten zu Wachstumsfaktoren (e^{b_1}) oder Wachstumsraten ($e^{b_1} - 1$) umrechnen



MODELL MIT EINEM PRÄDIKTOR

```r
d2_RU <- d2_RU |>
  mutate(ym_int_half = ym_int - (max(ym_int) + 1) / 2)

nbm1 <- glm.nb( # Negativ-binomiale Regression
  comments_count ~ ym_int_half, # Regressionsgleichung
  data = d2_RU
)
```

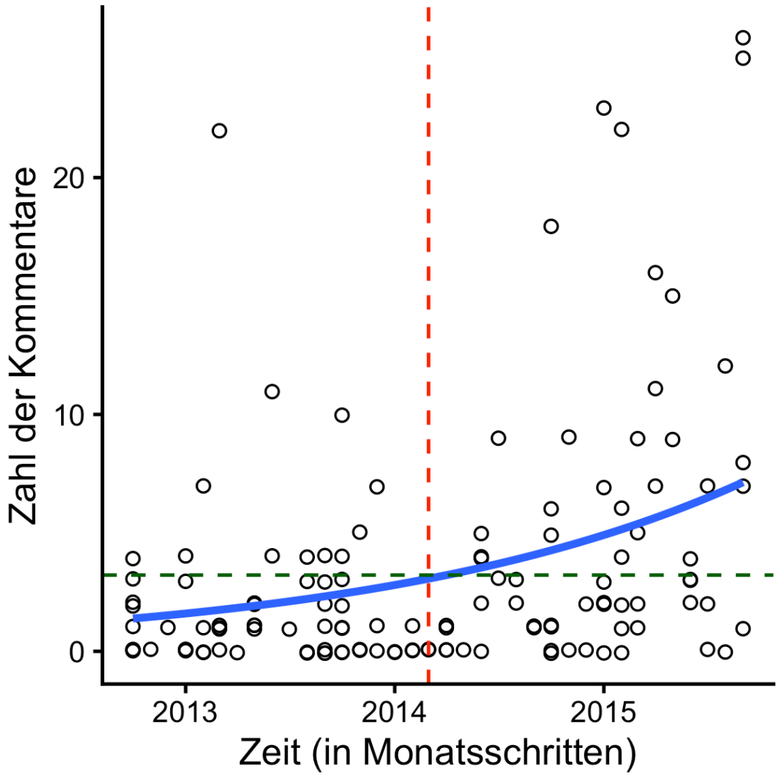


- Alle Prädiktoren, für die 0 kein sinnvoller Wert ist, werden um ihren Mittelwert zentriert, da wir den Intercept interpretieren wollen. In diesem Fall nehmen wir die Hälfte des Beobachtungszeitraums als Referenz.

EXPONENZIERTE KOEFFIZIENTEN

```
1 nbm1 |>
2   report_table(exponentiate = TRUE, metrics = "R2", include_ef
```

Parameter	Coefficient	95% CI	z	df	p	Fit
(Intercept)	3.22	(2.57, 4.07)	9.97	Inf	< .001	
ym int half	1.05	(1.03, 1.07)	4.20	Inf	< .001	
R2_Nagelkerke						0.21



- In der Mitte des Beobachtungszeitraums liegt die erwartete Zahl der Kommentare bei 3.2.
- Wachstumsfaktor: Die erwartete Zahl der Kommentare wächst pro Monat im das 1.05-fache.
- Wachstumsrate: Die erwartete Zahl der Kommentare wächst pro Monat um 5%.

Fragen?

AVERAGE COUNTERFACTUAL COMPARISON

Um die Mitte des Untersuchungszeitraums

```
 Estimate Std. Error   z Pr(>|z|)    S 2.5 % 97.5 %
     0.15      0.041 3.7   <0.001 12.3 0.073   0.24

Term: ym_int_half
Type: response
Comparison: 1 - 0
```

```
 Estimate Std. Error   z Pr(>|z|)    S 2.5 % 97.5 %
     0.15      0.038 3.9   <0.001 13.1 0.073   0.22

Term: ym_int_half
Type: response
Comparison: 0 - -1
```

Am Anfang und am Ende des Untersuchungszeitraums

```
 Estimate Std. Error   z Pr(>|z|)    S 2.5 % 97.5 %
    0.066     0.0082 8.1   <0.001 50.1  0.05  0.083

Term: ym_int_half
Type: response
Comparison: -17 - -18
```

```
 Estimate Std. Error   z Pr(>|z|)   S 2.5 % 97.5 %
     0.33       0.14 2.3    0.022 5.5 0.047   0.61

Term: ym_int_half
Type: response
Comparison: 17 - 16
```

- Zu Beginn des Untersuchungszeitraums steigt die erwartete Zahl der Kommentare um 0.07 pro Monat. In der Mitte des Untersuchungszeitraums sind es 0.15 weitere Kommentare pro Monat. Vom vorletzten auf den letzten Monat steigt die erwartete Zahl um knapp 0.33 Kommentare.

Fragen?

MODELL MIT MEHREREN PRÄDIKTOREN

```r
d2_RU <- d2_RU |>
  mutate(
    ym_int_half = ym_int - (max(ym_int) + 1) / 2,
    word_count_100_c = (word_count - mean(word_count, rm.na = TRUE)) / 100
  )
nbm2 <- glm.nb(
  comments_count ~ ym_int_half + topic_research + topic_teaching + word_count_100_c, # Regressionsgleichung
  data = d2_RU
)
```

- Alle Prädiktoren, für die 0 kein sinnvoller Wert ist, müssen transformiert werden, da wir den Intercept interpretieren wollen.

EXPONENZIERTE KOEFFIZIENTEN

```
1 nbm2 |>
2   report_table(exponentiate = TRUE, metrics = "R2", include_effectsize = FALSE)
```

Parameter	Coefficient	95% CI	z	df	p	Fit
(Intercept)	2.54	(1.82, 3.61)	5.31	Inf	< .001	
ym int half	1.05	(1.03, 1.08)	4.40	Inf	< .001	
topic research (yes)	1.23	(0.77, 1.96)	0.89	Inf	0.376	
topic teaching (yes)	1.81	(1.00, 3.45)	1.92	Inf	0.055	
word count 100 c	1.12	(0.31, 4.29)	0.19	Inf	0.848	
R2_Nagelkerke						0.26

- In der Mitte des Beobachtungszeitraums liegt die erwartete Zahl der Kommentare für einen durchschnittlich langen Post, der keines der beiden Themen enthält, bei 2.5.
- Wenn die übrigen Eigenschaften gleich bleiben, wächst die erwartete Zahl der Kommentare pro Monat um 5%.
- Die Koeffizienten der drei weiteren Prädiktoren sind nicht statistisch signifikant.

AVERAGE COUNTERFACTUAL COMPARISON

```
            Term Contrast Estimate Std. Error    z Pr(>|z|)    S  2.5 % 97.5 %
 topic_research   yes - no     0.68      0.776 0.88     0.38  1.4 -0.838   2.20
 topic_teaching   yes - no     2.52      1.644 1.53     0.13  3.0 -0.706   5.74
 word_count_100_c +1           0.39      2.141 0.18     0.86  0.2 -3.809   4.58
 ym_int_half      1 - 0        0.16      0.041 3.82   <0.001 12.9  0.076   0.24

Type: response
```

- Die Zahl der Kommentare steigt über den Untersuchungszeitraum hinweg an. Im hypothetischen Vergleich, in dem alle Posts in der Mitte des Untersuchungszeitraums oder einen Monat danach erschienen wären, erhielten sie im späteren Monat 0.16 zusätzliche Kommentare.
- Die übrigen Vergleiche zeigen keine statistischen signifikanten Unterschiede.

Fragen?

MODELLGÜTE: PSEUDO R^2

- Pseudo R^2 vergleichen die Passung des Null-Modells mit zu beurteilendem Modell.
- Verbreitet: McFadden, Nagelkerke, Cox & Snell, Tjur
- Empfehlung für Poisson- und negativ-binomiale Regression: Nagelkerke's R^2
- Interpretation wie R^2 in linearer Regression, wobei es je nach Maß und Modell sein kann, dass 0 und/oder 1 nicht erreicht werden können.

MODELLGÜTE: PSEUDO R^2

Parameter	Coefficient	95% CI	z	df	p	Fit
(Intercept)	2.54	(1.82, 3.61)	5.31	Inf	< .001	
ym int half	1.05	(1.03, 1.08)	4.40	Inf	< .001	
topic research (yes)	1.23	(0.77, 1.96)	0.89	Inf	0.376	
topic teaching (yes)	1.81	(1.00, 3.45)	1.92	Inf	0.055	
word count 100 c	1.12	(0.31, 4.29)	0.19	Inf	0.848	
R2_Nagelkerke						0.26

The model's explanatory power is substantial (Nagelkerke's R2 = 0.26)

Fragen?

MODELLVERGLEICH: POISSON- ODER NEGATIV-BINOMIALE REGRESSION?

```r
pm2 <- glm(
  comments_count ~ ym_int_half + topic_research + topic_teaching + word_count_100_c, # Regressionsgleichung
  family = poisson(link = "log"), # aV Zählvariable, log-Link = Poisson-Regression
  data = d2_RU
)
```

Parameter	Poisson	Negativ-binomial
(Intercept)	2.46 (2.10, 2.88)	2.54 (1.80, 3.59)
ym int half	1.06 (1.05, 1.07)	1.05 (1.03, 1.08)
topic research (yes)	1.31 (1.08, 1.58)	1.23 (0.78, 1.96)
topic teaching (yes)	1.72 (1.37, 2.17)	1.81 (0.99, 3.33)
word count 100 c	1.10 (0.66, 1.83)	1.12 (0.36, 3.48)
Observations	128	128

- Die Koeffizienten sind ähnlich.
- Die Inferenzstatistik unterscheidet aber sich deutlich. Die Unterschiede nach den Themen sind in der Poisson-Regression statistisch signifikant.

MODELLVERGLEICH: POISSON- ODER NEGATIV-BINOMIALE REGRESSION?

Poisson-Regression

```
# Overdispersion test

       dispersion ratio =   5.822
  Pearson's Chi-Squared = 716.094
                  p-value = < 0.001
```

Negativ-binomiale Regression

```
# Overdispersion test

 dispersion ratio = 0.933
          p-value = 0.976
```

Modellvergleich

Name	Model	AIC (weights)	BIC (weights)
pm2	glm	876.9 (<.001)	891.1 (<.001)
nbm2	negbin	589.2 (>.999)	606.3 (>.999)

#Df	LogLik	Df	Chisq	Pr(>Chisq)
5	-433.44			
6	-288.59	1	289.70	0.00

- Annahme der Poisson-Regression zur Verteilung der Varianz entsprechend des Erwartungswerts klar nicht erfüllt. Standardfehler des Modells zu klein, Inferenzstatistik falsch. Statistisch signifikante Unterschiede nach Themen wären sehr wahrscheinlich fälschliche Ablehnung der H_0.
- Negativ-binomiale Regression passt wesentlich besser zu den Daten. Nicht signifikante Ergebnisse für Themen sind verlässlicher.

Fragen?

FAZIT

- Die lineare Regression kann verallgemeinert werden, um Verteilung der aV und andere Link-Funktionen zwischen Prädiktoren und aV zu ermöglichen.
- Abwägung: Komplexere Modelle mit weniger intuitiver Interpretation der Koeffizienten als Preis für bessere Annäherung an datengenerierenden Prozess. Oder: Wann ist die lineare Regression gut genug, wann nicht?
- Vieles, was wir für die lineare Regression gelernt haben, lässt sich auf verallgemeinerte lineare Modelle übertragen: Kausale Annahmen, Moderation & Interaktion, Mediation, Pfadmodelle, Strukturgleichungsmodelle, Mehrebenenmodelle.

HAUSAUFGABE (FREIWILLIG)

1. Reproduzieren Sie die verallgemeinerten linearen Modelle aus der Vorlesung. Interpretieren Sie die Ergebnisse.

Nächste Sitzungen

Wiederholung und Fragen

Danke

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

Fähnrich, B., Vogelgesang, J., & Scharkow, M. (2020). Evaluating universities' strategic online communication: how do Shanghai Ranking's top 50 universities grow stakeholder engagement with Facebook posts? *Journal of Communication Management*, *24*(3), 265–283. https://doi.org/10/gmjs23

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

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