Statistik II - Sitzung 11

Lena Masch

Institut für Politikwissenschaft

16. Dezember 2024

Statistik II - Sitzung 11

Die Schätzung der logistischen Regression

- Der Unterschied zwischen OLS und logistischer Regression betrifft auch das Schätzverfahren
- Bei der OLS-Regression ging es um die Schätzung einer linearen Gerade, welche die verbleibenden Fehler(quadrate) minimiert
- Die logistische Regression hingegen basiert auf dem Maximum-Likelihood (ML)-Schätzverfahren

Die Schätzung der logistischen Regression

- Was bedeutet Maximum-Likelihood-Schätzung? (nach Wenzelburger et al. 2014)
- Vereinfacht: Im Gegensatz zur OLS-Regression wird im ML-Verfahren eine iterative (d.h., schrittweise annähernde) Schätzung vorgenommen
- Schritt I - Grundsätzliche Sch¨atzung des Modells aufgrund von Startwerten
- Schritt II - Verbesserung der grundlegenden Sch¨atzung durch Herantasten an die korrekte Vorhersage der Wahrscheinlichkeiten der tatsächlich beobachteten Stichprobenwerte

Die Schätzung der logistischen Regression

- Was bedeutet Maximum-Likelihood-Schätzung? (nach Wenzelburger et al. 2014)
- Vereinfacht: Im Gegensatz zur OLS-Regression wird im ML-Verfahren eine iterative (d.h., schrittweise annähernde) Schätzung vorgenommen
- "Mithilfe des Verfahrens werden die geschätzten Parameter (die Koeffizienten) so gewählt, dass die Wahrscheinlichkeit maximiert wird, die tatsächlich empirisch beobachteten Werte zu erhalten."(Wenzelburger et al. 2014: 64))

Die Schätzung der logistischen Regression

- Was bedeutet Maximum-Likelihood-Schätzung? (nach Wenzelburger et al. 2014)
- Vereinfacht: Im Gegensatz zur OLS-Regression wird im ML-Verfahren eine iterative (d.h., schrittweise annähernde) Schätzung vorgenommen
- Ende des Schätzungsprozesses, wenn keine Verbesserung der vorhergesagten Wahrscheinlichkeiten mehr erreicht werden kann
- Meistens - Interpretation des Grads an Verbesserung über LogLikelihood-Wert

Die Schätzung der logistischen Regression

- Was bedeutet der LogLikelihood-Wert (LL)?
- Der LogLikelihood-Wert bezeichnet den Wert, an dem das Maximum der Schätzung erreicht wird (daher: Maximum-Likelihood)
- Meist wird der negative LogLikelihood-Wert verwendet -LL (bei SPSS: -2LL)
- Das bedeutet: Je geringer der absolute LL-Wert, desto besser die Schätzung
  - Je geringer der absolute LL-Wert, desto besser die Schätzung ODER
  - Je kleiner der LL-Betrag, desto besser ist das Erklärungsmodell

Die Schätzung der logistischen Regression

- Die Schätzung des ML-Modells wird so lange durch das Rechenprogramm vorangetrieben, bis sich der -LL-Wert nicht mehr verändert
- D.h. sehr vereinfacht: Das Programm berechnet nacheinander ahnliche Modelle und vergleicht anhand des -LL-Werts, welches ¨ Modell die beste Annäherung an die beobachten Stichproben-Werte (Wahrscheinlichkeiten) ergibt
- Je komplexer das theoretische Modell (je mehr Variablen, je mehr Fall-Ebenen etc.), desto mehr Schätzungen mussen berechnet werden

Die Schätzung der logistischen Regression

- Grundlegend lassen sich dann über die -LL-Werte auch unterschiedlich theoretisch spezifizierte Modelle vergleichen
- Berechne ich ein Modell A mit wenigen unabhängigen Variablen und ein Modell B mit vielen unabhängigen Variablen, so kann ich anhand der Reduzierung des -LL-Werts zwischen A und B erkennen, ob sich durch die Hereinnahme mehrerer UVs in Modell B die Modellgüte des Erklärungsmodells verbessert hat
- nur möglich für Berechnungen mit dem selben Datensatz

Die Schätzung der logistischen Regression

- Der Pseudo-R2-Wert (oder McFadden's R2) ist ein weiteres Maß zur Beurteilung der Modellgüte
- Aber - er ist nicht wie der R2-Wert in der OLS-Regression interpretierbar, sondern lässt nur die Interpretation eines Anstiegs der Erklärungskraft / der Modellgute relativ zu einem anderen Modell zu!

Die Schätzung der logistischen Regression

- Weitere Maße zur Beurteilung der (relativen) Modellgüte sind die AIC- und BIC-Werte. Je niedriger die AIC-/BIC-Werte, desto besser das Modell
  - AIC = Akaike Information Criterion
  - BIC = Bayes Information Criterion
- Grundlegend basieren aber alle diese Maße auf den (-)LL-Werten!

Zentrale Probleme der logistischen Regression

- Grundlegendes Problem, das auftauchen kann: Die ML-Sch¨atzung konvergiert nicht.
- Vereinfacht gesagt bedeutet das, dass das Rechenprogramm keine eindeutig beste Schätzung ermitteln kann und damit kein Modell mit dem geringsten -LL-Wert ausgeben kann

Zentrale Probleme der logistischen Regression

- Mögliche Ursachen des Konvergenz-Problems (nach Wenzelburger et al. 2014)
  - Nicht korrekt spezifizierte Variablen
  - Zu geringe Fallzahlen
  - Zu ungleiche Skalierung der unabhängigen Variablen
  - AV-Ausprägungen zu ungleich verteilt
- Konvergiert das Schätzmodell nicht, ist daher immer zunächst Hinterfragung des theoretischen Erklärungsmodells bzw. der verwendeten Variablen und Fall-Verteilungen notwendig!

Beispiel: Parteineigung zur AfD

- Fragestellung(en)
  - Wie lässt sich eine Parteineigung zur AfD erklären?
  - Welche Faktoren beeinflussen das berichtete Neigung ?
  - Inwiefern beeinflussen populistische Einstellungen die Neigung?
- Analyse anhand von Sekundärdaten (ALLBUS 2018)

Operationalisierung der Variablen

- Die abhängige Variable (DV): **Parteineigung AfD (0/1)**
  - Populistische Einstellungen werden als Mittelwert aus mehreren Items gemessen (Index).
  - Der Index fasst Aussagen zur Unterstützung populistischer Ideologien zusammen (z. B. Anti-Establishment, einfacher Bürger*innen besser geeeignet).
- Die unabhängigen Variablen (UV):
  - **Geschlecht**: Binäre Variable (männlich = 0, weiblich = 1).
  - **Alter**: Metrische Variable, gemessen in Jahren.
  - **Bildungsniveau**: Kategorische Variable, die den höchsten Abschluss angibt in drei Kategorien (niedrig, mittel, hoch).
  - **Einkommen**: Einkommen als metrische Variable (monatliches Nettoeinkommen in €).

Beispiel: Parteineigung zur AfD

- Modell 1 (ohne populistische Einstellungen)

```
Call:
glm(formula = afd_vote ~ gender + education + income + age, family = binomial(
= "logit"),
    data = za)

Coefficients:
                  Estimate Std. Error z value   Pr(>|z|)
(Intercept)     -1.2697946  0.4307744  -2.948     0.0032 **
genderweiblich  -0.5634167  0.2465879  -2.285     0.0223 *
educationmittel  1.3629068  0.2976573   4.579 0.00000468 ***
educationniedrig 1.4662111  0.3555805   4.123 0.00003733 ***
income          -0.0003115  0.0001340  -2.325     0.0201 *
age             -0.0325824  0.0073566  -4.429 0.00000947 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 664.29  on 1499  degrees of freedom
Residual deviance: 610.19  on 1494  degrees of freedom
  (1977 observations deleted due to missingness)
AIC: 622.19

Number of Fisher Scoring iterations: 6
```

Beispiel: Parteineigung zur AfD

- Modell 2 (mit populistischen Einstellungen)

```
Call:
glm(formula = afd_vote ~ gender + education + income + age +
    pop_index, family = binomial(link = "logit"), data = za)

Coefficients:
                    Estimate  Std. Error z value             Pr(>|z|)
(Intercept)      -7.07413013  0.82423290  -8.583 < 0.0000000000000002 ***
genderweiblich   -0.52816926  0.26098482  -2.024               0.0430 *
educationmittel   0.57561762  0.32246534   1.785               0.0743 .
educationniedrig  0.41402869  0.38082897   1.087               0.2770
income           -0.00008247  0.00014030  -0.588               0.5567
age              -0.04182117  0.00811138  -5.156          0.000000252 ***
pop_index         1.80364641  0.19662355   9.173 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 664.29  on 1499  degrees of freedom
Residual deviance: 507.77  on 1493  degrees of freedom
  (1977 observations deleted due to missingness)
AIC: 521.77

Number of Fisher Scoring iterations: 7
```

Beispiel: Parteineigung zur AfD

- Modell 2 (mit populistischen Einstellungen)

Logistic Regression Results

Coefficients, Confidence Intervals, Z-Statistics, and P-Values

Term	Coefficient (Logit)	Conf. Interval (Low)	Conf. Interval (High)	Z-Value	P-Value
(Intercept)	−7.074	−8.748	−5.512	−8.583	0.000
genderweiblich	−0.528	−1.046	−0.021	−2.024	0.043
educationmittel	0.576	−0.047	1.223	1.785	0.074
educationniedrig	0.414	−0.329	1.169	1.087	0.277
income	0.000	0.000	0.000	−0.588	0.557
age	−0.042	−0.058	−0.026	−5.156	0.000

Beispiel: Parteineigung zur AfD

- Modell 2 (mit populistischen Einstellungen)

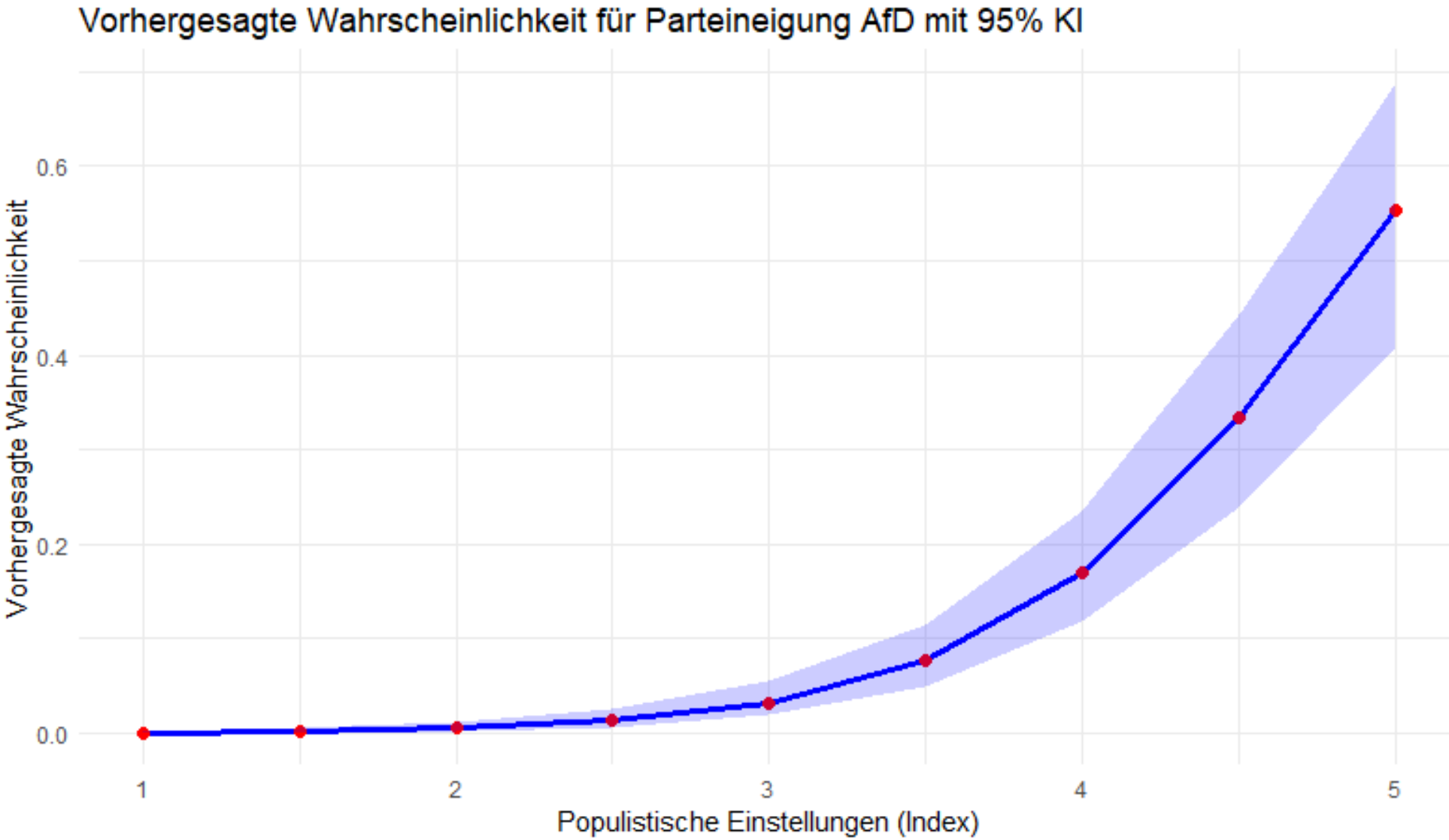
Logistic Regression Results

Exponentiated Coefficients (Odds Ratios) and Confidence Intervals

Term	Odds Ratio	Conf. Interval (Low)	Conf. Interval (High)
(Intercept)	0.001	0.000	0.004
genderweiblich	0.590	0.351	0.979
educationmittel	1.779	0.954	3.397
educationniedrig	1.513	0.720	3.219
income	1.000	1.000	1.000
age	0.959	0.944	0.974
pop_index	6.074	4.170	9.034

Beispiel: Parteineigung zur AfD

- Modell 2 (mit populistischen Einstellungen)



Die Schätzung der logistischen Regression

- Im Beispiel verbessert sich die Modellgüte durch die Hinzunahme einer weiteren Variable (Index populistischer Einstellungen)
- Sichtbar ist dies
  - an der Reduzierung des LogLikelihood-Wertes zwischen den Modellen: Modell 1 (ohne populistische Einstellungen) und Modell 2 (mit populistischen Einstellungen)
  - an der Reduzierung der Deviance
  - an der Reduzierung des AICs
  - an der Erhöhung des Pseudo-R^2-Wertes

Model	**Loglikelihood**	**Deviance**	**AIC**	**McFadden Pseudo R^2**
Modell 1	-305.10	610.19	622.19	0.081
Modell 2	-253.88	507.77	521.77	0.236

Tabelle: Modellvergleich

Verwendete Literatur

- Wenzelburger G, Jäckle S, König P (2014). Weiterführende statistische Methoden fur Politikwissenschaftler. München: Oldenbourg.

Ausblick

- Tutorien vor Weihnachten: logistische Regression
- Klausurübungsaufgabe auf Learnweb (ab 20.12.2024 einsehbar)
- Klausurübungsaufgabe wird in den Tutorien in der ersten Januarwoche gemeinsam diskutiert und bearbeitet
- Blick in Literatur zur Nachbereitung (s. Learnweb und Bibliothek)