

Kategoriale abhängige Variablen: „Logit-“ und „Probit“-Modelle

Statistik II

Wiederholung
Exkurs
Binäre abhängige Variablen
Interpretation
Zusammenfassung

Wiederholung

Literatur

Annahmen und

Annahmeverletzungen

Exkurs

Funktionen

Exponenten, Wurzeln usw.

Binäre abhängige Variablen

Das Problem

Das binäre Logit-Modell

Interpretation

Zusammenfassung



Zum Nachlesen/Vorbereiten

- ▶ Agresti ch. 15:

Was passiert, wenn Annahme 1 nicht erfüllt ist?

„Die abhängige Variable ist intervallskaliert und unbeschränkt.
Variablen werden ohne Fehler gemessen“

- ▶ Abhängige Variable hat häufig wenig diskrete Ausprägungen (Ratingskalen)
 - ▶ Erwartete Werte außerhalb des gültigen Wertebereichs
 - ▶ Modelle für ordinale Daten
 - ▶ In der Literatur wenig diskutiert, häufig wird angenommen, daß Modell relativ robust ist
- ▶ Alle sozialwissenschaftlichen Variablen fehlerbehaftet
 - ▶ Relativ unproblematisch, wenn Fehler voneinander unabhängig und Stichprobe groß
 - ▶ Fehler bei y wird von ϵ absorbiert, OLS weniger effizient
 - ▶ Fehler bei x schwächt im bivariaten Fall Zusammenhang ab, multivariat auf jeden Fall bias

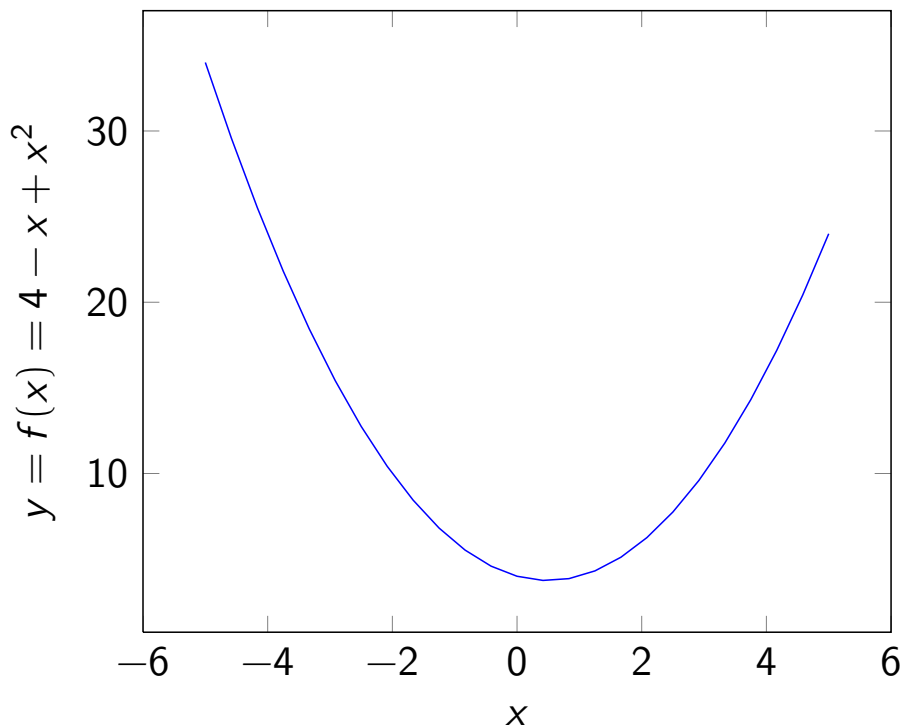
Annahmen und Annahmeverletzungen

- ▶ Ergebnisse auf Grundlage einer Stichprobe nur Schätzungen
- ▶ Schätzverfahren setzen Annahmen voraus
- ▶ Wenn Annahmen nicht zutreffen
 - ▶ Verzerrte Parameterschätzungen
 - ▶ Ineffiziente (und/oder inkonsistente) Parameterschätzungen
 - ▶ **Zu optimistische Standardfehler**
- ▶ Annahmen in Politikwissenschaft häufig verletzt
 - ▶ Z. B. Abhängigkeiten zwischen Beobachtungen (Zeitreihen, Panel ...)
 - ▶ Kategoriale abhängige Variablen
- ▶ Erweiterungen/Ergänzungen des linearen Modells

Was ist eine Funktion?

- ▶ „Abbildungsvorschrift“
- ▶ \approx Berechnungsvorschrift
- ▶ Ordnet jedem Wert der x -Variable(n) genau einen Wert zu
 - ▶ Einstellige Funktionen
 - ▶ Mehrstellige Funktionen
- ▶ Allgemeine Formulierung: $f(x_1, x_2, \dots)$
 - ▶ Lineare Funktion besteht nur aus Konstanten und Produkten von x_1, x_2, \dots
 - ▶ Nicht-lineare Funktion: andere Elemente
- ▶ Bisher y als lineare Funktion von x_1, \dots
- ▶ Alle Funktionen graphisch darstellbar (ggf. mehrdimensional)
- ▶ Steigung der Funktion in einem Punkt: (1.) Ableitung

Nicht-lineare Funktionen: z. B. Polynome



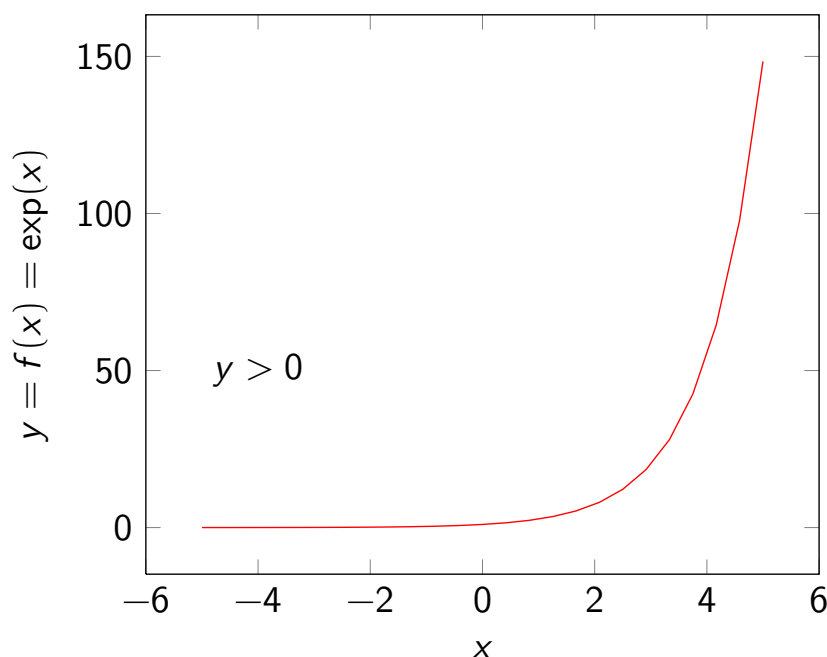
Exponenten

- ▶ Basis und Exponent
- ▶ Ganzzahlige positive Exponenten: $x^3 = x \cdot x \cdot x$
- ▶ Exponent 1 oder 0: $x^1 = x$; $x^0 = 1$
- ▶ Negative Exponenten: $x^{-1} = \frac{1}{x}$; $x^{-3} = \frac{1}{x^3}$
- ▶ Rationale Exponenten
 - ▶ Quadratwurzel aus x : Mit sich selbst multiplizieren, um x zu erhalten
 - ▶ n -te Wurzel aus x : n -mal mit sich selbst multiplizieren, um x zu erhalten
 - ▶ Nenner = n -te Wurzel: $\sqrt[n]{x} = x^{\frac{1}{n}}$
 - ▶ Kompletter Bruch: $\sqrt[5]{x^4} = x^{\frac{4}{5}}$
- ▶ $x^{-\frac{4}{5}} = \frac{1}{x^{\frac{4}{5}}} = \frac{1}{\sqrt[5]{x^4}}$

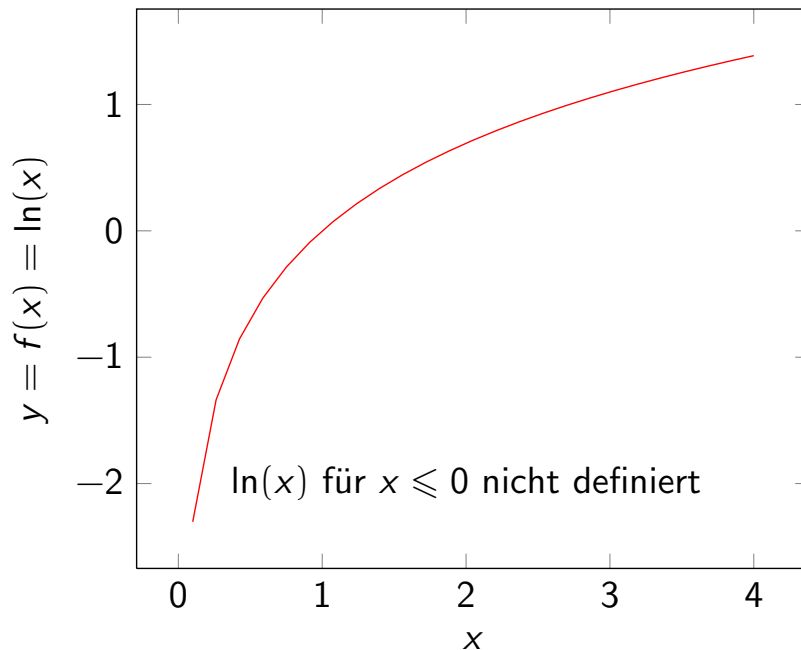
Was ist der natürliche Logarithmus?

- ▶ Logarithmus Umkehrfunktion zur Exponentialfunktion
- ▶ „Natürlicher“ Logarithmus (Funktionsname $\ln()$ oder \log_e) basiert auf Eulerscher Zahl $e = 2.71828182\dots$
- ▶ e wichtige Konstante in vielen statistischen Verteilungen und Herleitungen (z. B. Normalverteilung)
- ▶ $e^3 = \exp(3) \approx 20.0855\dots$
- ▶ $\ln(20.0855) \approx 3$
- ▶ Natürlicher Logarithmus von x gibt Antwort auf die Frage: Wie oft muß ich e mit sich selbst multiplizieren, um x zu erhalten?

Potenzen zur Basis e : $y = \exp(x)$



Natürlicher Logarithmus = Umkehrfunktion: $y = \ln(x)$



Binäre Variablen in der Politikwissenschaft

- ▶ Wahlabsicht in den USA: Republikanisch (0) vs. Demokratisch (1)
- ▶ Land in bestimmtem Jahr in Bürgerkrieg verwickelt: ja (1) vs. nein (0)
- ▶ Parteibindung vorhanden: ja (1) vs. nein (0)
- ▶ Politisches System eine Demokratie: ja (1) vs. nein (0)
- ▶ Wertorientierungen: postmaterialistisch (1) vs. nicht-postmaterialistisch (0)
- ▶ Wahlabsicht zugunsten der CDU: ja (1) vs. nein (0)
- ▶ Viele relevante Variablen binär (oder dichotom)
- ▶ **Wie modellieren?**

Strategie I: „Lineares Wahrscheinlichkeitsmodell“

- ▶ Beispiel: Wahlverhalten für CDU durch Sympathie für Merkel zu erklären?
- ▶ Zweitstimme in Umfrage → binäre Variable CDU-Wahl
- ▶ Für jeden Befragten 0 (nein) oder 1 (ja)
- ▶ Mittelwert der Dummy-Variablen entspricht relativer Häufigkeit bzw. Wahrscheinlichkeit der CDU-Wahl
- ▶ Warum?

$$\text{Mittelwert cduwahl} = \frac{10 \times 1 + 77 \times 0}{87} \approx 0.115 = \frac{10}{87}$$

- ▶ Gesamtwahrscheinlichkeit der CDU-Wahl ca. 11.5 Prozent
- ▶ Mittelwert der Dummy-Variablen in Sympathie-Gruppen = Anteil der CDU-Wähler in Sympathie-Gruppen =
- ▶ Konditionaler Mittelwert = Konditionale Wahrscheinlichkeit der CDU-Wahl in den Gruppen ($n = 80$)

```
. tabstat cduwahl, by (polsympangela Merkel)
```

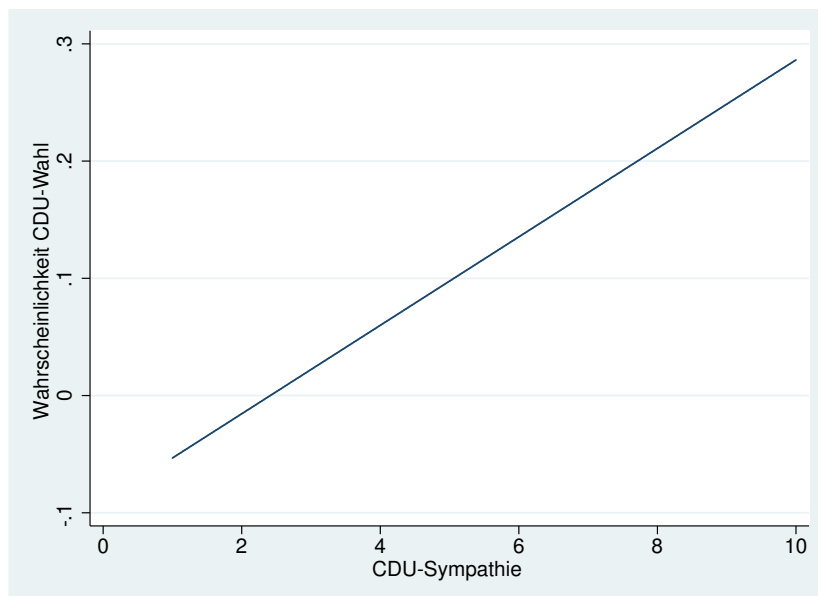
Statistik II

Logistische Regression (11/27)

```
by categories of: polsympangela Merkel (polsymp [Angela Merkel] )
```

polsympangela Merkel	mean
0	0
1	0

Probleme?



Wie kommt man zum Modell?

- ▶ Problem: CDU-Wahl bzw. deren Wahrscheinlichkeit auf Wertebereich $[0;1]$ beschränkt
- ▶ *Transformation* der Variablen
- ▶ 1. Schritt: Statt Wahrscheinlichkeiten „odds“ betrachten
 - ▶ (Entsprechen in etwa Wettquoten beim Sport)
 - ▶ $\text{odds}(p) = \frac{p}{1-p}$; im Beispiel $\frac{0.125}{0.875} \approx 0.143$
 - ▶ Wertebereich von 0 bis (fast) ∞
 - ▶ Variieren über Ausprägungen der unabhängigen
 - ▶ z. B. 0.07 (6.6%) und 1.99 (66.6%)
- ▶ 2. Schritt: Von diesen odds wird der natürliche Logarithmus bestimmt (Logarithmierung)

Wie kommt man zum Modell? II

- ▶ Die logarithmierten Odds werden als Logits bezeichnet
- ▶ Wertebereich von (fast) $-\infty$ bis (fast) $+\infty$
- ▶ Im Beispiel Logits zwischen -2.66 (6.6%) und 0.688 (66.6%)
- ▶ **Nicht-lineares Verhältnis zur Wahrscheinlichkeit**
 - ▶ Wahrscheinlichkeit von 50% entspricht Logit von 0
 - ▶ Positiver Logit – größere Wahrscheinlichkeit
 - ▶ Negativer Logit – kleinere Wahrscheinlichkeit
- ▶ **Logit-Modell: linearer Zusammenhang zwischen x und Logit**
 - ▶ $\text{logit}(\text{cduwahl}) = \beta_0 + \beta_1 \times \text{merkelsympathie}$
 - ▶ Schätzung der Koeffizienten/Standardfehler mit speziellem iterativen Verfahren (Maximum Likelihood)

In Stata

```
. logit cduwahl polsympangelamerkel
Iteration 0:  log likelihood = -30.141613
Iteration 1:  log likelihood = -27.289467
Iteration 2:  log likelihood = -26.993016
Iteration 3:  log likelihood = -26.991918
Iteration 4:  log likelihood = -26.991918

Logistic regression              Number of obs   =          80
                                LR chi2(1)      =           6.30
                                Prob > chi2      =          0.0121
                                Pseudo R2       =          0.1045

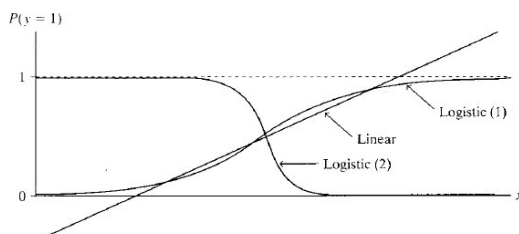
Log likelihood = -26.991918
```

	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
polysympang~1	.408557	.1822747	2.24	0.025	.0513051	.7658089
_cons	-4.604574	1.355321	-3.40	0.001	-7.260955	-1.948194

- ▶ Interpretation?
- ▶ Richtung und Signifikanz

Nicht-Linearität

- ▶ Zusammenhang zwischen x und y nicht-linear, aber monoton
 - ▶ Mehr x , mehr y (positiver Zusammenhang) bzw.
 - ▶ Mehr x , weniger y (negativer) Zusammenhang
 - ▶ **Aber nicht mit konstanter Rate**
- ▶ S-förmiger Zusammenhang
- ▶ Veränderung in Wahrscheinlichkeit *nicht* proportional zu Veränderung in x
 - ▶ Großer Effekt, wenn Wahrscheinlichkeit im mittleren Bereich
 - ▶ Kleiner Effekt, wenn Wahrscheinlichkeit sehr groß oder sehr gering



Was ist mit den zufälligen Fehlern?

- ▶ Im linearen Regressionsmodell zufällige Normalverteilung um konditionalen Mittelwert
- ▶ Separater Parameter (σ_ϵ^2)
- ▶ Für Logit-Modell Binomialverteilung um konditionale Wahrscheinlichkeit
- ▶ Varianz hängt ab von erwarteter Wahrscheinlichkeit (Heteroskedastizität)
- ▶ Ist durch Modell fixiert und wird nicht separat geschätzt
- ▶ Probit-Modelle sind sehr ähnlich, haben lediglich eine andere Link- bzw. Varianzfunktion

Interpretation Logit-Koeffizienten

- ▶ Modell nur in den Logits linear
- ▶ Interpretation von Richtung (Vorzeichen)
- ▶ Interpretation von Signifikanz
- ▶ Logits sind *sehr* unanschaulich

Interpretation Odds/Odd-Ratios

$$\begin{aligned}\text{logit}(\text{cduwahl}) &= \beta_0 + \beta_1 \times \text{merkelsympathie} \\ e^{\text{logit}(\text{cduwahl})} &= \text{odds}(\text{cduwahl}) = e^{(\beta_0 + \beta_1 \times \text{merkelsympathie})} \\ &= e^{\beta_0} \times e^{\beta_1 \text{merkelsympathie}}\end{aligned}$$

- ▶ Multiplikative Darstellung des Modells
- ▶ Für $x = 0$: odds = anti-logarithmierte Konstante
- ▶ $e^{\beta_1} = \exp(\beta_1) =$ „Effektkoeffizient“
- ▶ Veränderung von x um eine Einheit multipliziert die odds mit dem Effektkoeffizienten
- ▶ Findet sich manchmal in (älterer) Literatur, nicht sehr anschaulich

Wie kommt man von Logits zu Wahrscheinlichkeiten?

Logit

$$\text{Logit} = \beta_0 + \beta_1 x_1 = \ln \left(\frac{p}{1-p} \right)$$

Wie nach p auflösen?

Logarithmus loswerden

$$\exp(\text{Logit}) = \frac{p}{1-p}$$

p auf eine Seite bringen, ausmultiplizieren

Interpretation Wahrscheinlichkeiten

$$p = \frac{e^{\beta_0 + \beta_1 x_1}}{1 + e^{\beta_0 + \beta_1 x_1}}$$

- ▶ Odds auch nicht wirklich anschaulich
- ▶ Klarste Interpretation: erwartete Wahrscheinlichkeiten
- ▶ 1. Teil der Transformation auch umkehren
- ▶ Veränderung der Wahrscheinlichkeit nicht proportional zur Veränderung von x bzw. abhängig vom *Niveau* von x (und ggf. anderer x_2, \dots) → S-förmiger Zusammenhang

Erweiterung des Modells

$$\text{logit}(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots$$

- ▶ Logistische Regression ebenfalls multivariat möglich
- ▶ Mehrere unabhängige Variablen wirken linear-additiv auf den Logit
- ▶ Wirkung einer Veränderung von x_1 um eine Einheit auf die Wahrscheinlichkeit von $y = 1$ hängt ab vom
 - ▶ Niveau von x_1 **und**
 - ▶ vom Niveau von x_2, \dots
- ▶ Am besten graphisch darstellbar

CDU-Wahl II

- ▶ CDU-Wahl als Funktion von
 - ▶ Merksympathie
 - ▶ Links-Rechts-Selbsteinstufung
- ▶ $\text{logit}(\text{cduwahl}) = \beta_0 + \beta_1 \text{merkelsympathie} + \beta_2 \text{LRS}$

In Stata

```
. logit cduwahl polsympangelamerkel lrsselbstselbst
Iteration 0:  log likelihood = -29.870914
Iteration 1:  log likelihood = -22.149578
Iteration 2:  log likelihood = -19.757701
Iteration 3:  log likelihood = -19.605749
Iteration 4:  log likelihood = -19.604952
Iteration 5:  log likelihood = -19.604952

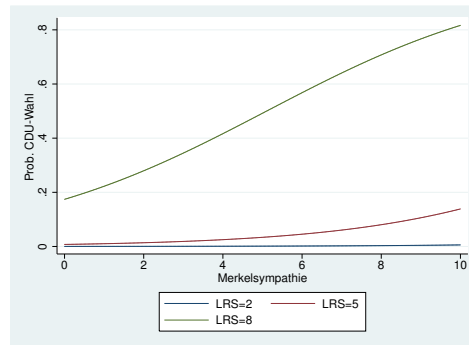
Logistic regression               Number of obs   =          78
                                LR chi2(2)       =         20.53
                                Prob > chi2      =          0.0000
                                Pseudo R2       =          0.3437

Log likelihood = -19.604952
```

cduwahl	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
polisympang~1	.3049494	.2132139	1.43	0.153	-.1129421 .7228409
lrsselbsts~t	1.106581	.3888137	2.85	0.004	.3445196 1.868641
_cons	-10.40907	3.036782	-3.43	0.001	-16.36105 -4.457082

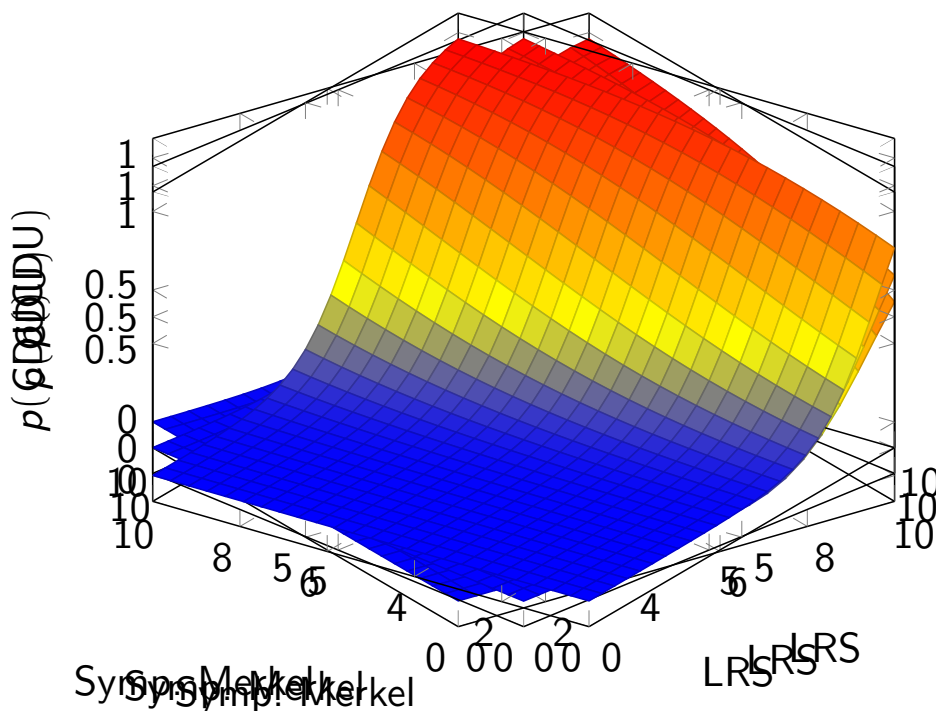
Graphische Darstellung

- ▶ Wie wirkt Merksympathie für
 - ▶ Linke (LRS=2)
 - ▶ Zentristen (LRS=5)
 - ▶ Rechte (LRS=8)?
- ▶ Wirkung von Sympathie ...
 - ▶ Fast linear für Rechte
 - ▶ Schwach bei Zentristen
 - ▶ Praktisch nicht vorhanden bei Linken
- ▶ Implizite Interaktion auf der Ebene der Wahrscheinlichkeiten



Statistik II Logistische Regression (25/27)

Multivariate Nicht-Linearität



Zusammenfassung

- ▶ Viele politikwissenschaftlich interessante Variablen dichotom
- ▶ Lineares Modell problematisch
- ▶ Logit-Modell als gute Alternative
- ▶ Interpretation erfordert Sorgfalt