

Nicht-kontinuierliche abhängige Variablen: Das generalisierte lineare Modell und die Parameterschätzung via Maximum Likelihood

Regressionsmodelle für Politikwissenschaftler

Wiederholung

Interaktionseffekte

Varianz-Kovarianz-Matrix

Das generalisierte lineare Modell

Probleme mit dem Standardmodell

GLM und Exponentialfamilie

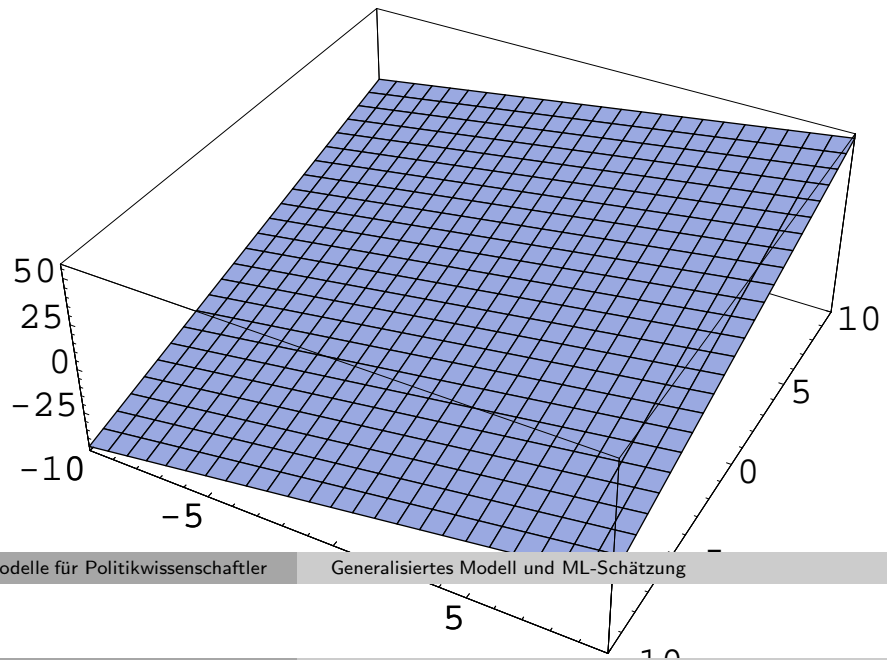
Lineares und Logit-Modell

ML-Schätzung

Fazit

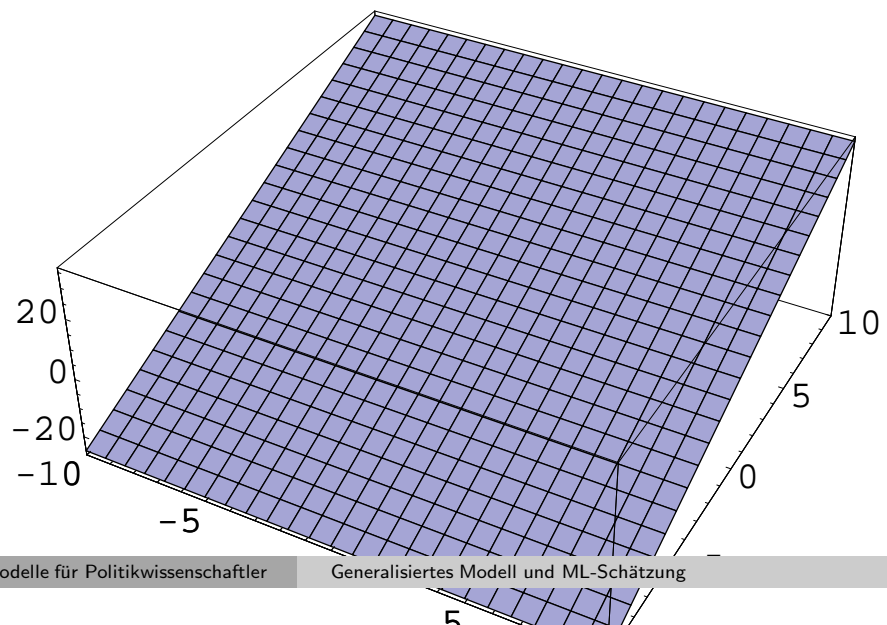
Inhaltliche Bedeutung von Interaktionseffekt und Komponenten?

- ▶ In einem Regressionsmodell mit zwei unabhängigen Variablen x_1, x_2 , die linear additiv zusammenwirken, liegen die erwarteten Werte in einer nicht-gewölbten Fläche



Inhaltliche Bedeutung von Interaktionseffekt und Komponenten?

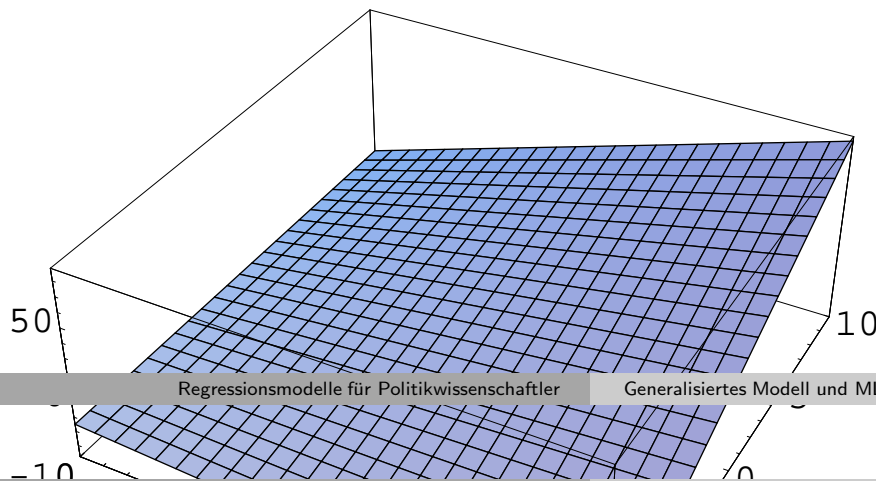
- ▶ Der Anstieg in Richtung einer Achse entspricht der Stärke des Effekts. Wenn x_1 keinen Effekt hat, steigt die Fläche in diese Richtung nicht an



Inhaltliche Bedeutung von Interaktionseffekt und Komponenten?

- ▶ Durch die Interaktion kommt es zu einer „Krümmung“ der Fläche, d. h. die Steilheit des Anstiegs (Effektstärke) hängt vom Niveau der jeweils anderen Variablen ab
- ▶ Je nach Wertebereich Umkehrung der Wirkungsrichtung möglich
- ▶ Auch Signifikanz der Wirkung von x_1 hängt vom Niveau von x_2 ab und umgekehrt

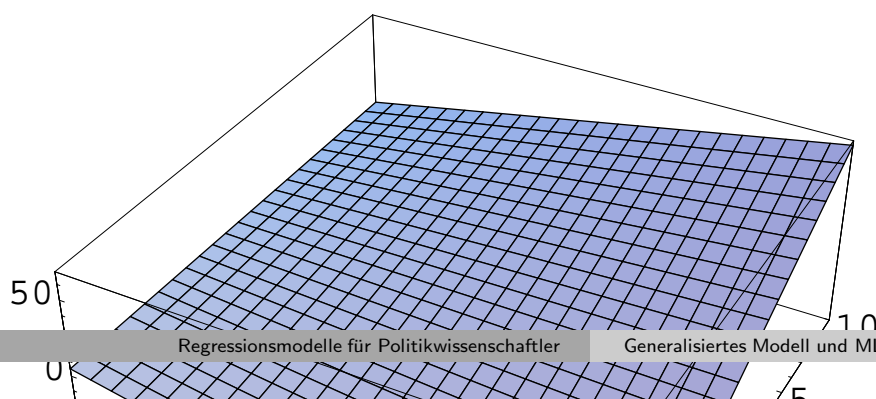
$$y = 5 + 2x_1 + 3x_2 + 0,3x_1x_2$$



Inhaltliche Bedeutung von Interaktionseffekt und Komponenten?

- ▶ Läßt man einen der Haupteffekte weg, erzwingt man, daß für diesen ein Wert von null geschätzt wird
- ▶ Das bedeutet, daß z. B. x_1 für den Fall, daß $x_2 = 0$ keine Wirkung hat (*nur* für diesen Fall, wegen Produktterm)
- ▶ Fläche muß bei $x_2 = 0$ in Richtung von x_1 gerade sein – vermutlich nicht der gewünschte Effekt
- ▶ (Realistischerweise verzerrte Schätzungen für andere Effekte)

$$y = 5 + 0x_1 + 3x_2 + 0,3x_1x_2$$



Was ist die Bedeutung der Varianz-Kovarianz-Matrix?

- ▶ Diese Matrix enthält *Schätzungen*
- ▶ Für die Varianz der Parameterschätzungen um den wahren Wert in der Grundgesamtheit auf der Hauptdiagonalen
- ▶ Und für die Kovarianzen zwischen diesen Schätzungen in den übrigen Zellen (symmetrisch)
- ▶ Positive Kovarianzen zwischen Schätzungen (z. B. $\text{cov}(\hat{\beta}_1, \hat{\beta}_2)$): In einer Stichprobe mit einer überdurchschnittlich hohen Schätzung für β_1 wird häufig auch die Schätzung für β_2 überdurchschnittlich hoch sein
- ▶ Unterdurchschnittliche Schätzung für β_2 ?
- ▶ Negative Kovarianz?
- ▶ Kommt zustande durch Zusammenhänge zwischen unabhängigen Variablen

Probleme im linearen Modell?

- ▶ Viele interessante abhängige politikwissenschaftliche Variablen dichotom (z. B. Wahl der extremen Rechten)
- ▶ Klassische Lösung: 0/1 kodieren, Regression rechnen
- ▶ Interpretation der vorhergesagten Werte als Wahrscheinlichkeiten / erwartete relative Häufigkeiten

Warum macht das Probleme?

- ▶ Wahrscheinlichkeiten auf Intervall $[0;1]$ beschränkt. Kleinere / größere Werte?
- ▶ Selbst wenn y über die beobachteten Werte von x innerhalb des Intervalls und linear mit x verbunden ...
- ▶ Varianz von $\epsilon = p(p - 1)$ hängt von erwartetem Wert ab – Konsequenz?
- ▶ Symmetrische Verteilung von ϵ um y unmöglich
- ▶ Wenn y nahe null fast nur positive Werte von ca. eins möglich, wenn y nahe 1, fast nur negative Werte von nahe -1 möglich Korrelation zwischen x und ϵ – Konsequenz?

Wie läßt sich das lineare Modell erweitern?

- ▶ Das generalisierte Modell hat drei Komponenten
 1. Eine systematische Komponente $\mathbf{X}\beta$ beziehungsweise $\beta_0 x_0 + \beta_1 x_1 \dots$
 2. Den konditionalen Mittelwert von y , der entweder mit $\mathbf{X}\beta$ identisch *oder durch eine non-lineare Funktion* (link, $\theta(\mu)$) *mit diesem Ausdruck verbunden ist*
 3. Eine Verteilung, die die Streuung von y um den konditionalen Mittelwert beschreibt und deren Varianz *normalerweise eine Funktion des konditionalen Mittelwertes ist*
- ▶ Für diese Verteilung wird ein Mitglied der Exponentialfamilie gewählt (u. a. Normalverteilung, Bernoulli-Verteilung, Poissonverteilung)
- ▶ „Kanonischer“ Link
- ▶ Große Flexibilität + konveniente mathematische Eigenschaften, u. a. existiert für Likelihood-Funktion ein globales Maximum

Was ist die Exponentialfamilie?

- ▶ Zur Erinnerung: e = Eulersche Zahl: 2,71...
- ▶ $e^x = x$ -mal $e \times e \dots = \exp(x)$
- ▶ Die Umkehrfunktion ist der natürliche Logarithmus $\ln(\exp(x)) = x$
- ▶ Ein negativer Exponent steht für den Kehrwert, d. h. $a^{-m} = \frac{1}{a^m}$
- ▶ Mit Hilfe rationaler Exponenten lassen sich Wurzeln ausdrücken, d. h. z. B. $a^{1/3} = \sqrt[3]{a}$
- ▶ Zwei Potenzen mit gleicher Basis kann man multiplizieren, indem man die Exponenten addiert

Was ist die Exponentialfamilie?

- ▶ Grundsätzlich: Modellierung der konditionalen Verteilung von y (diskret oder stetig)
- ▶ Normalverteilung normalerweise so definiert:

$$(f(y|\mu, \sigma^2)) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2}$$

- ▶ π und e sind Konstanten
- ▶ Achtung: hier y statt x

Was ist die Exponentialfamilie?

- ▶ Ergibt durch Umformung:

$$f(y|\mu, \sigma^2) = \exp\left(\frac{y\mu - \frac{\mu^2}{2}}{\sigma^2} - \frac{1}{2}\left(\frac{y^2}{\sigma^2} + \ln(2\pi\sigma^2)\right)\right)$$

- ▶ Generelle Form:

$$f(y|\theta, \phi) = \exp\left(\frac{y\theta - b(\theta)}{\phi} + c(y, \phi)\right)$$

- ▶ θ ist eine Funktion des Mittelwertes μ
- ▶ b ist eine Funktion von θ (und damit von μ)
- ▶ ϕ ist ein Parameter, der die Varianz definiert
- ▶ c ist eine Funktion des betreffenden y -Wertes und des Varianz-Parameters

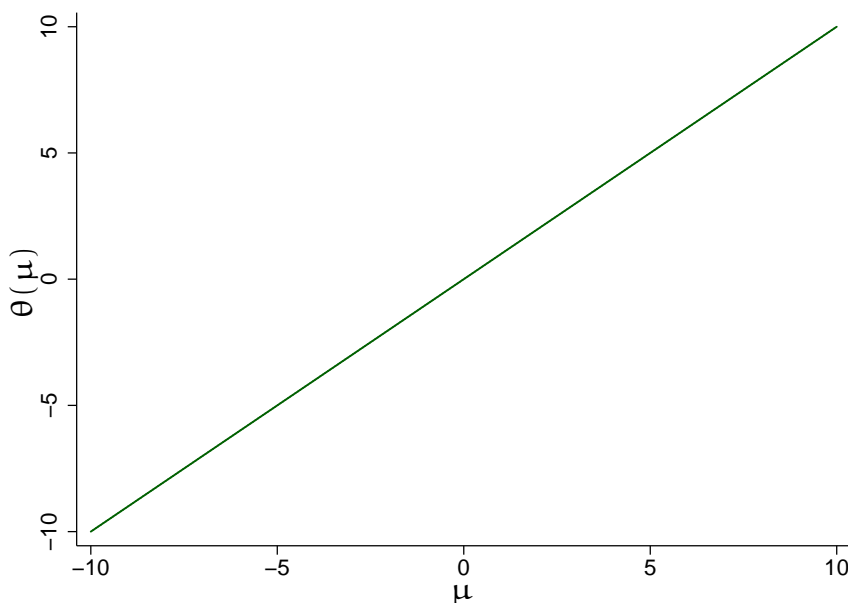
Was ist die Exponentialfamilie?

- ▶ Die Funktion $\theta(\mu)$ ist die kanonische Link-Funktion
- ▶ Die zweite Ableitung der Funktion $b(\theta)$ ist die Varianzfunktion, d. h. Varianz hängt vom Mittelwert ab
- ▶ Besonderheiten der Normalverteilung
 - ▶ Kanonische Link-Funktion = Identität: $\theta(\mu) = \mu$
 - ▶ Zweite Ableitung von $b(\theta) = b'' = 1$, d. h. Varianz ist konstant σ^2 und nicht vom Mittelwert abhängig
- ▶ Normalverteilung besonders einfacher Spezialfall der Exponentialfamilie
- ▶ Lineare Regression besonders einfacher Spezialfall des generalisierten Modells

Wie läßt sich das lineare Modell als generalisiertes Modell rekonstruieren?

- ▶ Identitäts-Link
- ▶ y ist für ein gegebenes μ normalverteilt mit einem separaten Varianzparameter σ^2

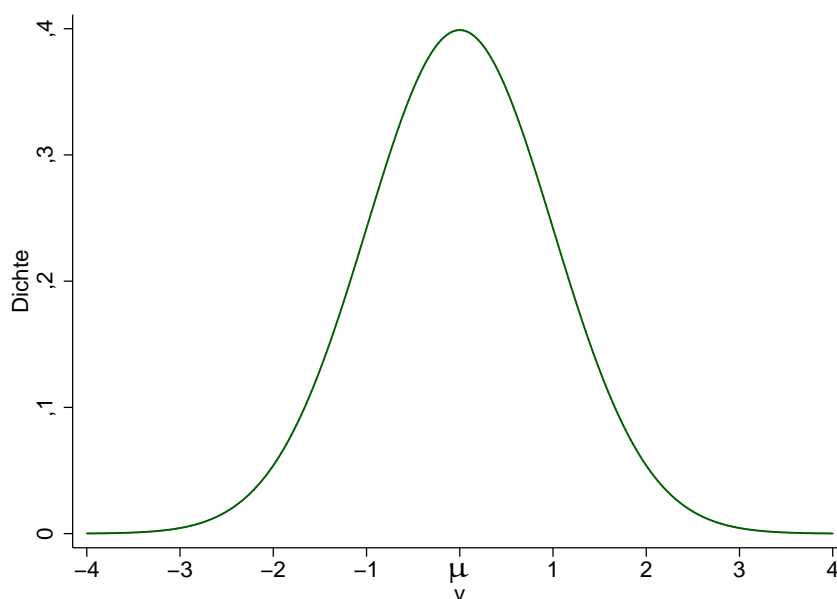
Was heißt Identitäts-Link?



Wie läßt sich das lineare Modell als generalisiertes Modell rekonstruieren?

- ▶ Identitäts-Link
- ▶ y ist für ein gegebenes μ normalverteilt mit einem separaten Varianzparameter σ^2

Was heißt Normalverteilung von y ?



Was gibt es sonst noch?

- ▶ Sehr viele interessante Variablen dichotom (unser Ausgangspunkt), d. h. $y = 0$ oder $y = 1$
- ▶ Solche Variablen werden allgemein durch Binomialverteilung beschrieben. . .

$$f(y|n, p) = \binom{n}{y} p^y (1 - p)^{n-y}$$

- ▶ . . . die sich für unsere Zwecke meistens auf die Bernoulli-Verteilung ($n = 1$) reduzieren läßt . . .
- ▶ . . . die eine sehr einfache exponentielle Form hat:

$$\begin{aligned} f(y|\pi) &= \exp \left(y \ln \left(\frac{\pi}{1 - \pi} \right) - \ln \left(1 + \frac{\pi}{1 - \pi} \right) \right) \\ &= \exp \left(y \ln \left(\frac{\pi}{1 - \pi} \right) + \ln (1 - \pi) \right) \\ &= \pi \left(\frac{\pi}{1 - \pi} \right)^{y-1} \end{aligned} \quad \text{Achtung: statt } \mu \text{ meistens } \pi$$

Wieso ist das besonders einfach?

- ▶ Die kanonische Funktion $\theta = \ln \left(\frac{\pi}{1 - \pi} \right)$, d. h. die beliebte Logit-Transformation
- ▶ $b(\theta) = \ln (1 + \exp (\theta))$, die zweite Ableitung davon (Varianzfunktion) ist $\pi(1 - \pi)$
- ▶ $\phi = 1$, $c(y, \phi) = 0$

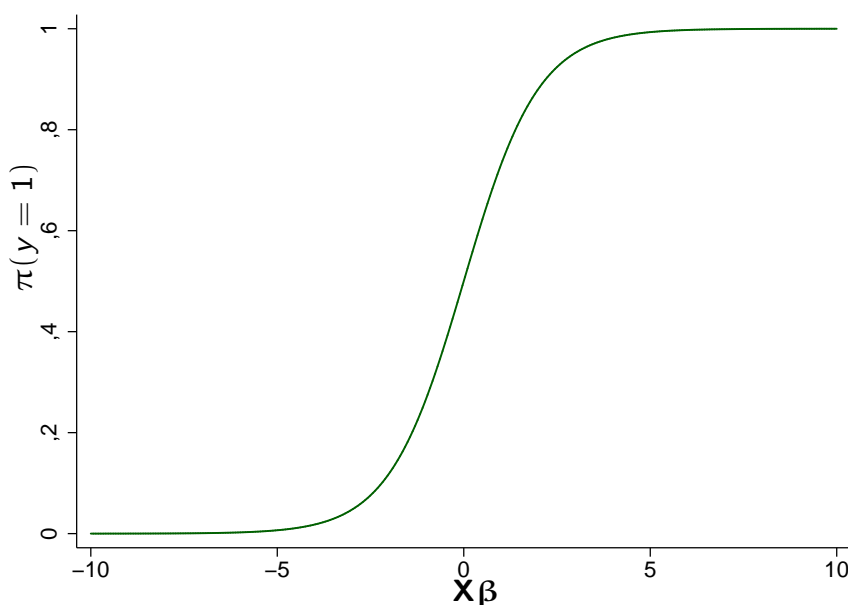
Was war noch mal die Logit-Transformation?

- ▶ Wahrscheinlichkeiten auf das Intervall $[0;1]$ beschränkt
- ▶ $\mathbf{X}\beta$ kann zwischen $+\infty$ und $-\infty$ liegen

$$\text{logit } p = \ln \left(\frac{p}{1-p} \right)$$
$$\text{invlogit}(\text{logit}(p)) = \frac{\exp(\text{logit}(p))}{1 + \exp(\text{logit}(p))}$$

- ▶ Logits können (fast) zwischen $+\infty$ und $-\infty$ liegen →
Wahrscheinlichkeiten zwischen (fast) null und (fast) eins darstellbar

Wie sieht der logistische Link aus?



Was heißt Bernoulliverteilung von y ?

- ▶ Wenn $\pi = 0,3$ ist die Wahrscheinlichkeit für $y = 1$ 0,3
- ▶ Und die Wahrscheinlichkeit für $y = 0$ gleich 0,7

Was gibt es sonst noch?

- ▶ Für die Zahl der Ereignisse pro Zeitraum die Poisson-Verteilung mit dem Parameter λ , der Varianz und Mittelwert definiert
- ▶ Survival Analysis, z. B. mit der Exponentialfunktion
- ▶ Vieles andere mehr
- ▶ Alle diese Modelle haben
 - ▶ Dieselbe Struktur (systematischer Teil, konditionale Verteilung von y , deren Varianz von μ abhängt, nicht-linearer Link zwischen $\mathbf{X}\beta$ und μ)
 - ▶ Erfreuliche Eigenschaften des zugehörigen Schätzverfahrens (ML)

Was ist die Grundidee der ML-Schätzung?

- ▶ Die Daten werden als gegeben angesehen
- ▶ Nun variiert man die Parameterschätzungen ...
- ▶ ... bis man solche Werte gefunden hat, die *am wahrscheinlichsten* die beobachteten Daten hervorgebracht haben können
- ▶ ML-Schätzungen sind
 1. Asymptotisch unverzerrt und effizient
 2. Konsistent
 3. Bei großen Stichproben approximativ normalverteilt

Wie funktioniert das beim normalen linearen Modell?

- ▶ Die konditionale Normal-Verteilung von y sagt uns, wie wahrscheinlich bestimmte Werte bei gegebenem μ sind
- ▶ Wahrscheinlichkeit der Daten bei gegebenen Parametern
- ▶ Erwarteter Wert: Wert, der bei gegebenen Parametern am wahrscheinlichsten ist
- ▶ Das läßt sich gedanklich umdrehen: Wir fragen nach Werten, die bei gegebenen Daten die Wahrscheinlichkeit der Parameter maximieren

Wie funktioniert das beim normalen linearen Modell?

- ▶ Allerdings reden wir hier nicht mehr über Wahrscheinlichkeiten
 - ▶ Auf Wertebereich 0/1 normiert
 - ▶ y ist eine Zufallsvariable
- ▶ Weil wir innerhalb des frequentistischen Ansatzes davon ausgehen, daß Parameter in der Grundgesamtheit festliegen
- ▶ Sondern über die Plausibilität von Parameterschätzungen – Likelihood
- ▶ Likelihood nicht auf Bereich von 0/1 beschränkt
- ▶ Likelihood-Funktion: Funktion der Parameterschätzungen und der Daten
- ▶ Maximum für diese Funktion suchen

Welches ist die Likelihood-Funktion im linearen Modell?

- ▶ Für jeden individuellen Fall i ist die Likelihood-Funktion die Dichtefunktion der Normalverteilung

$$\begin{aligned} (f(y_i | \mu_i, \sigma^2)) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2} \left(\frac{y_i - \mu_i}{\sigma^2}\right)^2\right) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2} \left(\frac{y_i - \mathbf{X}_i\boldsymbol{\beta}}{\sigma^2}\right)^2\right) \end{aligned}$$

Welches ist die Likelihood-Funktion im linearen Modell?

- ▶ Da die Fälle voneinander unabhängig sind, ergibt sich die *gemeinsame* Likelihood-Funktion für alle Fälle in der Stichprobe durch Multiplikation der individuellen Funktionen

$$\begin{aligned} f(y_1, y_2 \dots y_n | \boldsymbol{\beta}, \sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2} \left(\frac{y_1 - \mathbf{X}_1\boldsymbol{\beta}}{\sigma^2}\right)^2\right) \\ &\times \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2} \left(\frac{y_2 - \mathbf{X}_2\boldsymbol{\beta}}{\sigma^2}\right)^2\right) \dots \\ &\times \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2} \left(\frac{y_n - \mathbf{X}_n\boldsymbol{\beta}}{\sigma^2}\right)^2\right) \\ &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2} \left(\frac{y_i - \mathbf{X}_i\boldsymbol{\beta}}{\sigma^2}\right)^2\right) \end{aligned}$$

Welches ist die Likelihood-Funktion im linearen Modell?

- ▶ Der erste Faktor in diesem Produkt ist eine Konstante:

$$\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n = \frac{1}{(2\pi\sigma^2)^{n/2}}$$

- ▶ Der zweite Faktor ist ein Produkt von Potenzen mit gleicher Basis (e), deshalb kann man die Exponenten addieren
- ▶ So erhält man die altbekannten SAQ
- ▶ In Matrix-Form:

$$\begin{aligned} f(\mathbf{y} | \boldsymbol{\beta}, \sigma^2) &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{2\sigma^2}\right) \\ &= L(\mathbf{y} | \boldsymbol{\beta}, \sigma^2) \end{aligned}$$

Was passiert dann?

- ▶ Meistens ist es einfacher, nicht mit der Likelihood-Funktion selbst, sondern mit deren Logarithmus zu rechnen (Log-Likelihood)
- ▶ Logarithmus ist monotone Transformation, deshalb führt Maximierung zum selben Ergebnis
- ▶ Nimmt man auf beiden Seiten den Logarithmus, erhält man

$$\begin{aligned} & \ln (L(\mathbf{y}|\boldsymbol{\beta}, \sigma^2)) \\ &= -\frac{n}{2} \ln (2\pi\sigma^2) - \frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{2\sigma^2} \end{aligned}$$

- ▶ Log-Likelihood wird maximal, wenn $(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$ minimal wird
- ▶ Analytische Lösung bereits bekannt, OLS ist für lineares Modell der ML-Schätzer
- ▶ Wenn Fläche der Likelihood-Funktion in der Nähe des Maximums stark gewölbt, präzise Schätzungen möglich (kleine Standardfehler)

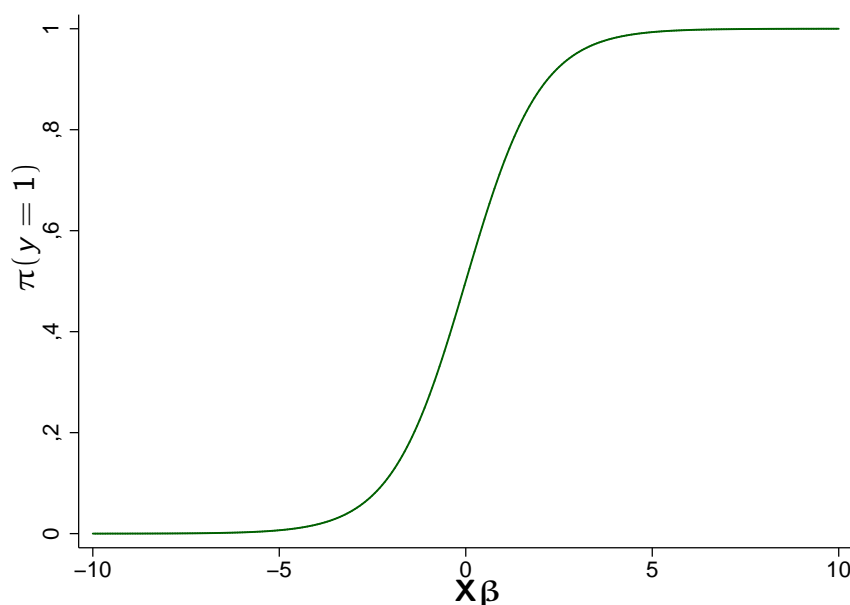
Was ist die Devianz?

- ▶ Beschreibt, wie gut Modell und Daten zueinander passen
- ▶ Absoluter Wert uninteressant, wichtig für Vergleich (Differenz) zwischen konkurrierenden Modellvarianten (vor allem Nullmodell (keine Parameter) und saturiertes Modell (pro Fall ein Parameter))
- ▶ Differenz zwischen zwei Devianzen ist χ^2 -verteilt – Test möglich (sind alle Parameter in der Grundgesamtheit gleich null), aber Problem beim linearen Modell
- ▶ Formal entspricht die Devianz eines Modells der doppelten Differenz zwischen der Log-Likelihood des saturierten Modells und des Modells, das analysiert wird
- ▶ Wegen unterschiedlicher Log-Likelihood-Funktionen unterschiedliche Berechnung der Devianz bei verschiedenen Modelltypen
- ▶ Im linearen Fall: SAQ/σ^2
- ▶ Für alle Modelle können Pearson- und Devianz-Residuen berechnet werden, um Modellfit zu prüfen

Wie schätzt man die Parameter des logistischen Modells?

- ▶ Nochmal: Varianz hängt jetzt vom konditionalen Mittelwert, d. h. von der Wahrscheinlichkeit, daß $y = 1$ ab
- ▶ Kanonischer Link: Logit-Funktion. Andere Links möglich (Probit, Clog-Log)
- ▶ D. h. Zusammenhang zwischen $\mathbf{X}\beta$ und μ nicht-linear, sondern über Funktion θ (Logit-Transformation) vermittelt

Was impliziert der Logit-Link?



Wie schätzt man die Parameter des logistischen Modells?

- ▶ Die Bernoulli-Verteilung ist gegeben durch

$$f(y|\pi) = \exp \left(y \ln \left(\frac{\pi}{1-\pi} \right) - \ln \left(1 + \exp \left(\frac{\pi}{1-\pi} \right) \right) \right)$$

- ▶ π ist eine nichtlineare Funktion von Daten und Parametern:
 $\frac{\exp(\mathbf{X}\boldsymbol{\beta})}{1+\exp(\mathbf{X}\boldsymbol{\beta})}$ (Umkehr der Logit-Funktion)
- ▶ Wir drehen die Überlegung wieder um und suchen die Likelihood von π für gegebene Daten. . .
- ▶ Müssen aber so umformen, daß das ganze eine Funktion des Vektors $\boldsymbol{\beta}$ ist, für die wir uns interessieren
- ▶ Wenn man von der resultierenden Likelihood-Funktion wiederum den Logarithmus sucht, erhält man

$$\sum_{i=1}^n (-y_i \ln(1 + \exp(-\mathbf{X}_i\boldsymbol{\beta})) - (1 - y_i) \ln(1 + \exp(\mathbf{X}_i\boldsymbol{\beta})))$$

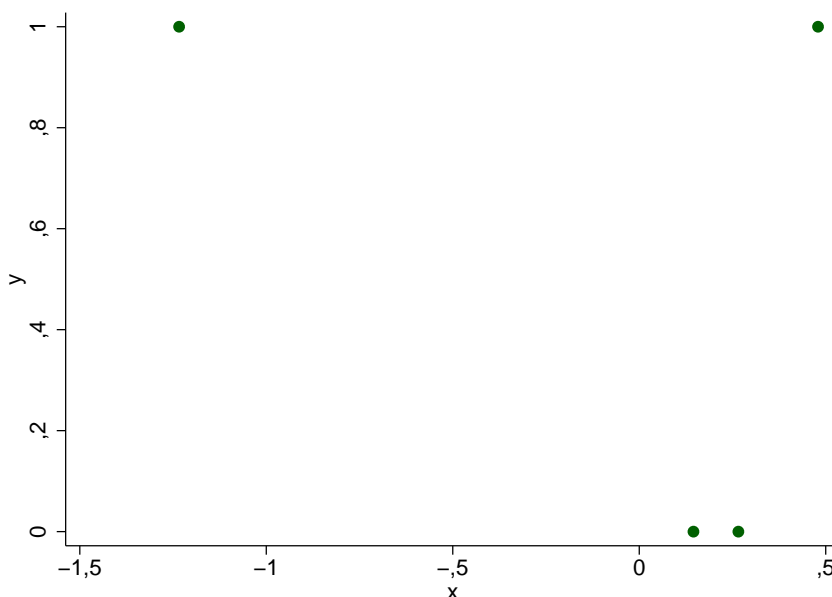
Wie schätzt man die Parameter des logistischen Modells?

- ▶ Die Log-Likelihood-Funktion ist differenzierbar
- ▶ Die erste Ableitung heißt Score-Funktion
- ▶ Diese kann auf null gesetzt werden
- ▶ Aber die Lösung ist nicht mit analytischen Methoden zu finden → numerische Methoden

Wie funktioniert diese numerische Schätzung?

- ▶ Mit dem Newton-Raphson-Algorithmus kann man iterativ die Nullstelle(n) einer Funktion finden: $x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$ (basiert auf Taylor-Serie, die den Zusammenhang zwischen zwei Funktionswerten approximiert)
- ▶ D. h. man teilt den Wert der Funktion durch den Wert der ersten Ableitung an dieser Stelle,
- ▶ zieht das Ergebnis vom Ausgangswert ab und
- ▶ bewegt sich so auf die Nullstelle zu
- ▶ Dabei werden die Schritte immer kleiner
- ▶ Wenn Differenz zwischen x_n und x_{n+1} Grenzwert unterschreitet, wird der Algorithmus abgebrochen

Ein Beispiel?



Wie geht der Computer vor?

y	x	$\mathbf{X}\tilde{\boldsymbol{\beta}}$	LL
0	0.15	-1.07	-0.30
0	0.27	-0.87	-0.35
1	0.48	-0.53	-0.99
1	-1.23	-3.27	-3.31

- ▶ Startwerte für β_0, β_1 : $-0.13; -1.6$
- ▶ Initiale LL -4.95
- ▶ LL nach vier Iterationen -2.36
- ▶ Parameterschätzungen nach vier Iterationen $-.04; -1.56$

Mögliche Probleme?

- ▶ Startwerte und lokale Extremwerte
- ▶ Achtung: Da f bereits die erste Ableitung der Log-Likelihood-Funktion ist, ist f' die (partielle) zweite Ableitung der Log-Likelihood-Funktion → Hesse-Matrix
- ▶ Strukturelle Ähnlichkeit zwischen iterativer Likelihood-Schätzung und Iterativer Gewichteter Kleinster Quadrate-Schätzung für heteroskedastische Daten (IWLS)
- ▶ ML-Schätzung für die hier vorgestellten Modelle normalerweise unproblematisch, sehr schnelle und sichere Konvergenz

Was ist das Fazit für heute?

- ▶ Viele politikwissenschaftliche Probleme erfüllen nicht die Voraussetzungen des linearen Modells
- ▶ GLM erweitert das lineare Modell um eine ganze Klasse von alternativen Modellen
- ▶ Für Stichproben von normalem Umfang ist ML ein zumindest intuitiv nachvollziehbares Verfahren, um zu exzellenten Parameterschätzungen zu gelangen

Was ist die Übung für heute?

- ▶ Die Wahlbeteiligung (w , gemessen 0/1) hängt (u. a.) vom politischen Interesse (i , gemessen von 0-4) und von der formalen Bildung (b , gemessen von 0-2) ab und soll durch das einfach Logit Modell $\text{logit}(w) = .08 + i \times 0.4 + b \times 0.5$ beschrieben werden
- ▶ Errechnen Sie die Wahrscheinlichkeit der Wahlbeteiligung für einen Befragten mit mittlerer Bildung (1) und mittlerem Interesse (2)
- ▶ Um wieviele Prozentpunkte steigt diese Wahrscheinlichkeit an, wenn das politische Interesse um zwei Skalenpunkte steigt?
- ▶ Wie groß ist die Differenz für einen Befragten mit hoher (2) formaler Bildung?
- ▶ Was fällt Ihnen dabei auf?
- ▶ Hinweis zur Berechnung: Die Exponentialfunktion e^x erreichen Sie in STATA mit `exp(x)`. Sie können sich die Wahrscheinlichkeit, die einem Logit von 0 entspricht, z. B. mit `display exp(.0)/(1+exp(.0))` anzeigen lassen