Maximum-Likelihood Schätzung

VL Forschungsmethoden

Lernziele

1. Grundzüge der Likelihood-Schätzung
2. Struktur des Generalisierten Linearen Modells

Einführung/Wiederholung

Wiederholung

- Kausalität ist komplex
- Kontrafaktische Definition
- Prüfung vor allem eine Frage des Designs

Was ist ein Modell?

- Allgemein: (stark) vereinfachte Annahmen über “datengenerierenden Prozeß”, System von Beziehungen zwischen Variablen
- Kann oft als statistisches (Regressions)-Modell formuliert werden
- Sozialer Prozeß $\leftrightarrow$ Parameter und zufällige Einflüsse

Was ist ein Schätzverfahren?

Inferenzproblem

- Mathematisches Modell ($y = \beta_0 + \beta_1 x_1 \ldots$)
  - Unsere Vorstellung über Zustand einer Grundgesamtheit oder
  - den DGP selbst
- Wie können wir mit Daten auf Parameter des Modells schließen?

Was ist ein Schätzverfahren?

Inferenzproblem

- Mathematisches Modell ($y = \beta_0 + \beta_1 x_1 \ldots$)
  - Unsere Vorstellung über Zustand einer Grundgesamtheit oder
  - den DGP selbst
- Wie können wir mit Daten auf Parameter des Modells schließen?

Schätzverfahren

- Schätzt Parameter
- Auf Grundlage der Daten
- *Berücksichtigt Unsicherheit* (durch beschränkte Information)

Welche Eigenschaften hat ein gutes Verfahren?

Maß für Probleme: MSE

Mean Squared Error (MSE) $=(\text{bias}(\hat{\beta}))^2 + \text{Varianz}(\hat{\beta})$

Welche Eigenschaften hat ein gutes Verfahren?

Maß für Probleme: MSE

Mean Squared Error (MSE) $=(\text{bias}(\hat{\beta}))^2 + \text{Varianz}(\hat{\beta})$

Gute Eigenschaften

1. Asymptotisch unverzerrt: kein systematischer Fehler (Mittelwert über Parameterschätzungen)
2. Effizient: Schätzungen haben relative geringe Varianz um wahren Wert (Kehrwert MSE)
3. Konsistenz: Wahrscheinlichkeit einer relevanten Differenz zwischen Schätzung und Parameter kann durch zusätzliche Beobachtungen beliebig verringert werden
    - Wenn bias und Varianz bei steigendem *n* gegen null streben …
    - Hinreichende Bedingung für Konsistenz

Zwei Ansätze

„Frequentistisch“

- Annahme, daß sich essentiell identische Stichprobenziehungen beliebig oft wiederholen lassen
- Wahrscheinlichkeit als langfristige relative Häufigkeit
- *Verteilung* von *Stichprobenkennwerten* bzw. *Parameterschätzungen*
- Konkreter Wert/Schätzung als zufälliger Wert aus bekannter Zufallsverteilung
- Klassische Konfidenzintervalle und Hypothesentests

Zwei Ansätze

„Frequentistisch“

- Annahme, daß sich essentiell identische Stichprobenziehungen beliebig oft wiederholen lassen
- Wahrscheinlichkeit als langfristige relative Häufigkeit
- *Verteilung* von *Stichprobenkennwerten* bzw. *Parameterschätzungen*
- Konkreter Wert/Schätzung als zufälliger Wert aus bekannter Zufallsverteilung
- Klassische Konfidenzintervalle und Hypothesentests

„Bayesianisch“

- Wahrscheinlichkeit als subjektive Überzeugung einer Forscherin (“beliefs”)
- Systematische Inkorporation bisheriges Wissens (“prior beliefs”), beschrieben durch *Wahrscheinlichkeitsdichte* (Verteilung)
- Modellschätzungen integrieren neue Daten und bisheriges Wissen (“posterior beliefs”)
- “Credible Intervals”

Unterschiede/Gemeinsamkeiten

- Große philosophische Unterschiede
- Bayesianische Methoden flexibler
- Sehr ähnliche Ergebnisse in großen Stichproben (kein prior knowledge)
- Beide Verfahren basieren auf dem Likelihood-Prinzip

Likelihood-Schätzung und Generalisiertes Lineares Modell

Drei Arten von Wahrscheinlichkeiten

1. Unbedinge Wahrscheinlichkeit (marginal probability)
    - Wie wahrscheinlich ist Ausprägung/Intervall von X insgesamt
    - über alle Ausprägungen/gesamtes Intervall von Y hinweg?

Drei Arten von Wahrscheinlichkeiten

1. Unbedinge Wahrscheinlichkeit (marginal probability)
   - Wie wahrscheinlich ist Ausprägung/Intervall von X insgesamt
   - über alle Ausprägungen/gesamtes Intervall von Y hinweg?
2. Gemeinsame Wahrscheinlichkeit (joint probability)
   - Wie wahrscheinlich ist gemeinsames Auftreten einer Ausprägung/eines Intervalls von X
   - mit Ausprägung/Intervall von Y?
   - (Produkt der marginalen Wahrscheinlichkeiten wenn X und Y unabhängig)

Drei Arten von Wahrscheinlichkeiten

1. Unbedinge Wahrscheinlichkeit (marginal probability)
   - Wie wahrscheinlich ist Ausprägung/Intervall von X insgesamt
   - über alle Ausprägungen/gesamtes Intervall von Y hinweg?
2. Gemeinsame Wahrscheinlichkeit (joint probability)
   - Wie wahrscheinlich ist gemeinsames Auftreten einer Ausprägung/eines Intervalls von X
   - mit Ausprägung/Intervall von Y?
   - (Produkt der marginalen Wahrscheinlichkeiten wenn X und Y unabhängig)
3. Bedingte Wahrscheinlichkeit (conditional probability)
   - Wie wahrscheinlich ist Ausprägung/Intervall von X
   - wenn Ausprägung/Intervall von Y bekannt?
     - = marginale Wahrscheinlichkeit, wenn unabhängig
     - Sonst gemeinsame Wahrscheinlichkeit durch Wahrscheinlichkeit von Y

Unbedingte vs bedingte Wahrscheinlichkeit

Ethnizität und PI

	+D	-D		↓
+B	60	40	100	0.1
-B	360	540	900	0.9
	420	580		
→	.42	.58		

Wahrscheinlichkeiten

- Gemeinsame Wahrscheinlichkeit $Pr(-B \cap -D)$: 0.54
- Konditionale Wahrscheinlichkeit

$$Pr(-D| + B) = \frac{Pr(-D \cap +B)}{Pr(+B)} = \frac{.04}{.10} = 0.4$$

- Konditionale Wahrscheinlichkeiten über Bayes' Rule verbunden

Satz von Bayes und Likelihood-Schätzung

$$Pr(A|B) = \frac{Pr(A \cap B)}{Pr(B)} = \frac{Pr(B|A) \times Pr(A)}{Pr(B)}$$

Übertragung:

$$Pr(Parameter|Daten) = \frac{Pr(Daten|Parameter) \times Pr(Parameter)}{Pr(Daten)}$$

Proportionalität:

$$Pr(Parameter|Daten) \propto Pr(Daten|Parameter) \times Pr(Parameter)$$

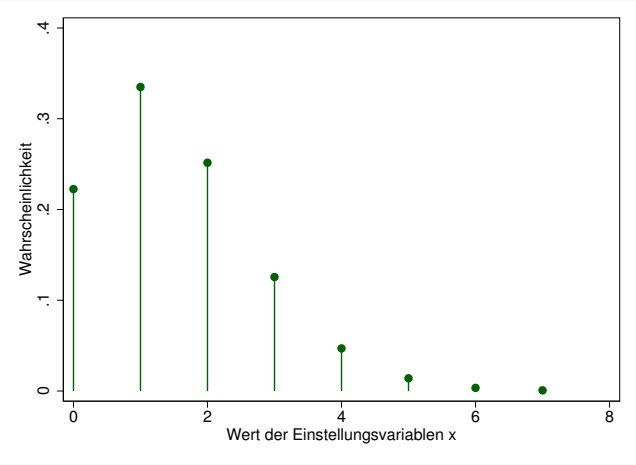
Grundgedanke Likelihood-Schätzung

- Wenn Daten als gegeben betrachtet werden (konditional) …
- Wahrscheinlichkeitsverteilung der Daten für gegebene Parameter → Proportionalität → Quasi-Wahrscheinlichkeit von Parameterwerten für gegebene Daten = Likelihood
- Likelihood
  - Funktion von Daten, Modell und Parameterschätzungen
  - Keine echte Wahrscheinlichkeit (fehlende Informationen)
  - Aber *proportional* zu Wahrscheinlichkeit → Vergleich von Schätzwerten möglich → welche Parameterschätzungen sind mehr oder weniger plausibel?
- Regel: Maximiere Likelihood, um gute Schätzungen zu erhalten → Wahrscheinlichkeitsverteilungen

Was ist eine Zufallsvariable?

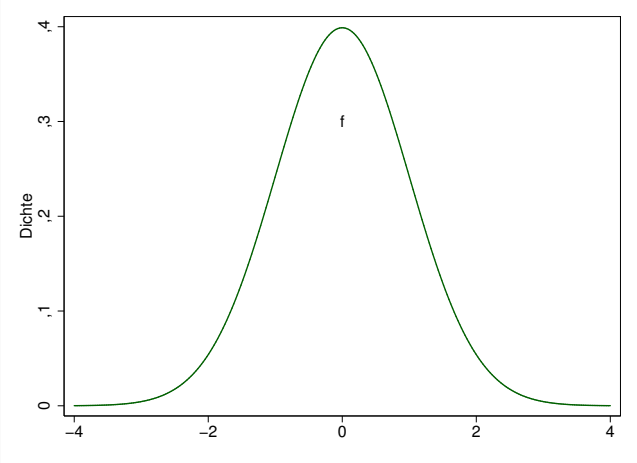
- Zufallsvariablen werden durch Verteilungsfunktionen beschrieben
- Diskrete Variablen:
  - Säulen, die die Wahrscheinlichkeit eines Ereignisses repräsentieren
  - Die **Summe der Säulen** ergibt 1
- Stetige Zufallsvariablen:
  - Stetige Verteilungsfunktion („Dichte“)
  - Wahrscheinlichkeit eines *exakten* Wertes gleich null
  - Statt dessen Wahrscheinlichkeiten für *Intervalle*, indem das Integral (Fläche unter der Dichtekurve) bestimmt wird
  - Die **Gesamtfläche ist 1**

Verteilung einer diskreten Zufallsvariablen

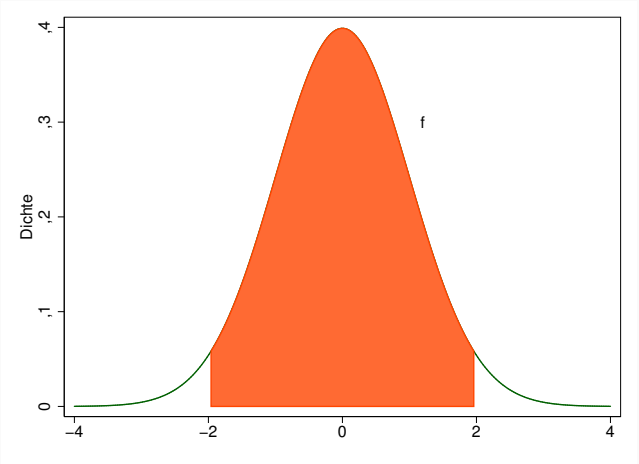


Verteilung einer stetigen Zufallsvariablen

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



Verteilung einer stetigen Zufallsvariablen



$\int_{-1,96}^{1,96} f(x|\mu, \sigma^2) = 0,95$

Wie läßt sich das lineare Modell erweitern?

- Das generalisierte Modell hat drei Komponenten

vorwärts

Wie läßt sich das lineare Modell erweitern?

- Das generalisierte Modell hat drei Komponenten
  1. Eine **Zufallsverteilung**, die die Streuung von Y um den konditionalen Mittelwert μ beschreibt und deren Varianz *normalerweise eine Funktion des konditionalen Mittelwertes* μ *ist*

vorwärts

Wie läßt sich das lineare Modell erweitern?

- Das generalisierte Modell hat drei Komponenten
  1. Eine **Zufallsverteilung**, die die Streuung von *Y* um den konditionalen Mittelwert μ beschreibt und deren Varianz *normalerweise eine Funktion des konditionalen Mittelwertes* μ *ist*
  2. Eine **systematische Komponente (Prädiktor) X**β beziehungsweise $\beta_0 x_0 + \beta_1 x_1 \ldots$

vorwärts

Wie läßt sich das lineare Modell erweitern?

- Das generalisierte Modell hat drei Komponenten
  1. Eine **Zufallsverteilung**, die die Streuung von *Y* um den konditionalen Mittelwert μ beschreibt und deren Varianz *normalerweise eine Funktion des konditionalen Mittelwertes* μ *ist*
  2. Eine **systematische Komponente (Prädiktor) X**β beziehungsweise $\beta_0 x_0 + \beta_1 x_1 \ldots$
  3. Eine (normalerweise non-lineare) **Funktion** θ (link), die Prädiktor und konditionalen Mittelwert verbindet $\mathbf{X}\beta = \theta(\mu)$

vorwärts

Wie läßt sich das lineare Modell erweitern?

- Das generalisierte Modell hat drei Komponenten
  1. Eine **Zufallsverteilung**, die die Streuung von *Y* um den konditionalen Mittelwert μ beschreibt und deren Varianz *normalerweise eine Funktion des konditionalen Mittelwertes* μ *ist*
  2. Eine **systematische Komponente (Prädiktor) X**β beziehungsweise $\beta_0 x_0 + \beta_1 x_1 \ldots$
  3. Eine (normalerweise non-lineare) **Funktion** θ (link), die Prädiktor und konditionalen Mittelwert verbindet $\mathbf{X}\beta = \theta(\mu)$
- Für die Verteilung wird ein Mitglied der Exponentialfamilie gewählt, das zu den Daten paßt

vorwärts

Wie läßt sich das lineare Modell erweitern?

- Das generalisierte Modell hat drei Komponenten
  1. Eine **Zufallsverteilung**, die die Streuung von *Y* um den konditionalen Mittelwert μ beschreibt und deren Varianz *normalerweise eine Funktion des konditionalen Mittelwertes* μ *ist*
  2. Eine **systematische Komponente (Prädiktor) X**β beziehungsweise $\beta_0 x_0 + \beta_1 x_1 \ldots$
  3. Eine (normalerweise non-lineare) **Funktion** θ (link), die Prädiktor und konditionalen Mittelwert verbindet $\mathbf{X}\beta = \theta(\mu)$
- Für die Verteilung wird ein Mitglied der Exponentialfamilie gewählt, das zu den Daten paßt
- Zu jeder Verteilung gehört ein „kanonischer“ Link

vorwärts

Was ist die Exponentialfamilie?

- Grundsätzlich: Modellierung der konditionalen Verteilung von y (diskret oder stetig)

Was ist die Exponentialfamilie?

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

Was ist die Exponentialfamilie?

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

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

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

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:

Was ist die Exponentialfamilie?

- Ergibt durch Umformung:

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

Was ist die Exponentialfamilie?

- Ergibt durch Umformung:

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

- Generelle Form:

Was ist die Exponentialfamilie?

- Ergibt durch Umformung:

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

- Generelle Form:

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

Was ist die Exponentialfamilie?

- Ergibt durch Umformung:

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

- 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

Was ist die Exponentialfamilie?

- Ergibt durch Umformung:

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

- 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)

Was ist die Exponentialfamilie?

- Ergibt durch Umformung:

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

- 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

Was ist die Exponentialfamilie?

- Ergibt durch Umformung:

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

- 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

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

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

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: $\theta(\mu) = \mu$

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

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

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 gebenes μ normalverteilt mit einem separaten Varianzparameter σ^2

Wie läßt sich das lineare Modell als generalisiertes Modell rekonstruieren?

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

Bisher ...

$$y = \mathbf{X}\beta + \epsilon \tag{1}$$

$$\epsilon \sim N(0, \sigma^2) \tag{2}$$

Wie läßt sich das lineare Modell als generalisiertes Modell rekonstruieren?

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

Bisher ...

$$y = \mathbf{X}\beta + \epsilon \tag{1}$$
$$\epsilon \sim N(0, \sigma^2) \tag{2}$$

GLM

$$y \sim N(\mu, \sigma^2) \tag{3}$$
$$\theta(\mu) = \mu = \mathbf{X}\beta \tag{4}$$

@StuartJRitchie

- Logistische Regression
- Poisson-Regression
- …
- gehören alle zum GLM

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: $f(y|n, p) = \binom{n}{y} p^y (1-p)^{n-y}$
- … reduziert sich für uns meist zu Bernoulli-Verteilung ($n = 1$) $\rightarrow$ $f(y|p) = p^y (1-p)^{1-y}$; kanonischer Link: Logit-Funktion vorwärts

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: $f(y|n, p) = \binom{n}{y} p^y (1-p)^{n-y}$
- … reduziert sich für uns meist zu Bernoulli-Verteilung ($n = 1$) $\rightarrow$ $f(y|p) = p^y (1-p)^{1-y}$; kanonischer Link: Logit-Funktion vorwärts
- … die eine sehr einfache exponentielle Form hat:

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: $f(y|n, p) = \binom{n}{y} p^y (1-p)^{n-y}$
- … reduziert sich für uns meist zu Bernoulli-Verteilung ($n = 1$) $\rightarrow$ $f(y|p) = p^y (1-p)^{1-y}$; kanonischer Link: Logit-Funktion vorwärts
- … die eine sehr einfache exponentielle Form hat:

$$
\begin{aligned}
f(y|\pi) &= \exp\left(y \ln\left(\frac{\pi}{1-\pi}\right) - \ln\left(1 + \frac{\pi}{1-\pi}\right)\right) \\
&= \pi \left(\frac{\pi}{1-\pi}\right)^{y-1}
\end{aligned}
$$

Achtung: statt μ meistens π

Wieso ist das besonders einfach?

- Die kanonische Funktion $\theta = \ln\left(\frac{\pi}{1-\pi}\right)$, d. h. die beliebte Logit-Transformation

Wieso ist das besonders einfach?

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

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 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
  - Diesselbe 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 (Maximierung der Likelihood)
- Oft: einfacher den Logarithmus der Likelihood zu maximieren (Log-Likelihood)
- ML-Schätzungen sind
  1. Asymptotisch unverzerrt und effizient
  2. Konsistent
  3. Bei großen Stichproben approximativ normalverteilt

Welches ist die Likelihood-Funktion im linearen Modell?

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

Welches ist die Likelihood-Funktion im linearen Modell?

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

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

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

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

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

Welches ist die Likelihood-Funktion im linearen Modell?

- Der erste Faktor in dieser Produktkette ist eine Konstante:

Welches ist die Likelihood-Funktion im linearen Modell?

- Der erste Faktor in dieser Produktkette ist eine Konstante:

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

Welches ist die Likelihood-Funktion im linearen Modell?

- Der erste Faktor in dieser Produktkette 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

Welches ist die Likelihood-Funktion im linearen Modell?

- Der erste Faktor in dieser Produktkette 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

Welches ist die Likelihood-Funktion im linearen Modell?

- Der erste Faktor in dieser Produktkette 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:

Welches ist die Likelihood-Funktion im linearen Modell?

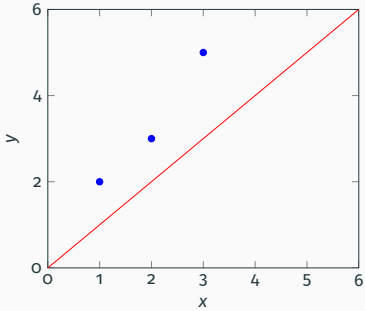
- Der erste Faktor in dieser Produktkette ist eine Konstante:

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

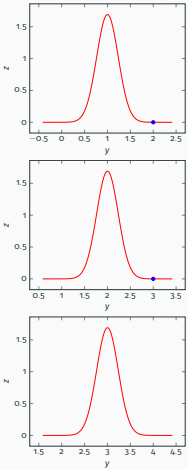
- Der zweite Faktor ist ein Produkt von Potenzen mit gleicher Basis (e), deshalb kann man die Exponenten addieren
- So erhält man die altbekannten SAQ
- In Matrix-Form:

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

Beispiel lineare Regression

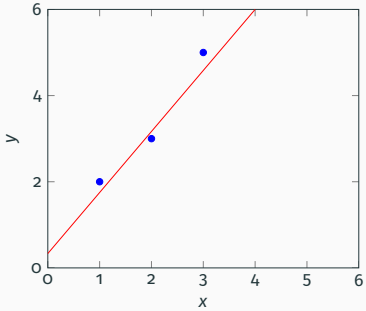


$\hat{\beta}_0 = 0; \hat{\beta}_1 = 1$

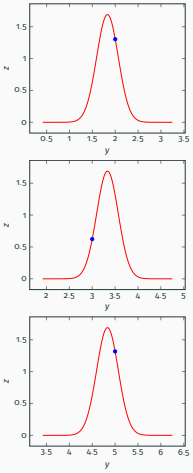


$L \approx 0$

Beispiel lineare Regression



$\hat{\beta}_0 = \frac{1}{3}; \hat{\beta}_1 = 1,5$



$L \approx 1.0710952$

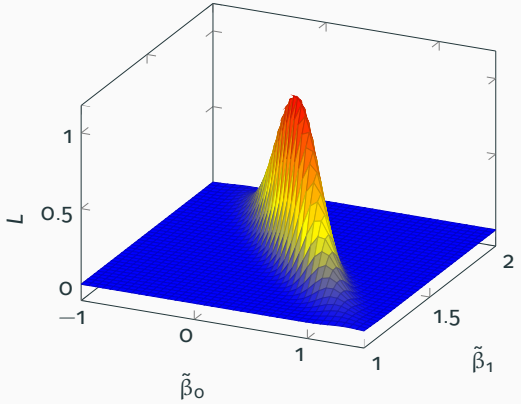
Was passiert dann?

- 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

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

- Log-Likelihood wird maximal, wenn $(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})$ minimal wird
- Wenn Fläche der Likelihood-Funktion in der Nähe des Maximums stark gewölbt, präzise Schätzungen möglich

Wieso Fläche?



$\hat{\beta}_0$ hat größeren Standardfehler als $\hat{\beta}_1$

Was ist die Devianz?

- Beschreibt, wie gut Modell und Daten zueinander passen
- 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)
- Formal entspricht die Devianz eines Modells der doppelten Differenz zwischen der Log-Likelihood des saturierten Modells und des Modells, das analysiert wird
- Unterschiedlicher Log-Likelihood-Funktionen $\rightarrow$ unterschiedliche Berechnung der Devianz
- Im linearen Fall: SAQ/σ^2

Wie schätzt man die Parameter des logistischen Modells?

- Nochmal: Varianz hängt jetzt vom konditionalen Mittelwert, d. h. von der Wahrscheinlichkeit, daß $y = 1$ ab
- Diese Wahrscheinlichkeit wird meistens mit π bezeichnet (steht hier **nicht** für die Kreiszahl)
- Kanonischer Link: Logit-Funktion. Andere Links möglich (z.B. Probit)
- D. h. Zusammenhang zwischen $\mathbf{X}\beta$ und μ nicht-linear, sondern über Funktion θ (Logit-Transformation) vermittelt

Wie schätzt man die Parameter des logistischen Modells?

- Die Bernoulli-Verteilung ist gegeben durch $f(y|\pi) = \pi^y(1-\pi)^{1-y}$
- π ist eine nichtlineare Funktion von Daten und Parametern: $\frac{\exp(\mathbf{X}\beta)}{1+\exp(\mathbf{X}\beta)}$ (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 β ist, für die wir uns interessieren
- Wenn man von der resultierenden Likelihood-Funktion wiederum den Logarithmus sucht, erhält man

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

Wie schätzt man die Parameter des logistischen Modells?

- Diese Log-Likelihood-Funktion ist differenzierbar
- Die erste Ableitung heißt Score-Funktion
- Diese kann auf null gesetzt werden
- (Am Maximum einer Funktion ist die erste Ableitung =0)
- Aber die Lösung ist nicht mit analytischen Methoden zu finden $\rightarrow$ 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)

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)
- (Achtung: $f(x_n)$ ist hier die Score-Funktion, d.h. die erste Ableitung der Log-Likelihood-Funktion)

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)
- (Achtung: $f(x_n)$ ist hier die Score-Funktion, d.h. die erste Ableitung der Log-Likelihood-Funktion)
- D. h. man teilt den Wert der Funktion durch den Wert der ersten Ableitung an dieser Stelle,

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)
- (Achtung: $f(x_n)$ ist hier die Score-Funktion, d.h. die erste Ableitung der Log-Likelihood-Funktion)
- D. h. man teilt den Wert der Funktion durch den Wert der ersten Ableitung an dieser Stelle,
- zieht das Ergebnis vom Ausgangswert ab und

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)
- (Achtung: $f(x_n)$ ist hier die Score-Funktion, d.h. die erste Ableitung der Log-Likelihood-Funktion)
- 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

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)
- (Achtung: $f(x_n)$ ist hier die Score-Funktion, d.h. die erste Ableitung der Log-Likelihood-Funktion)
- 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

Wie funktioniert diese numerische Schätzung?

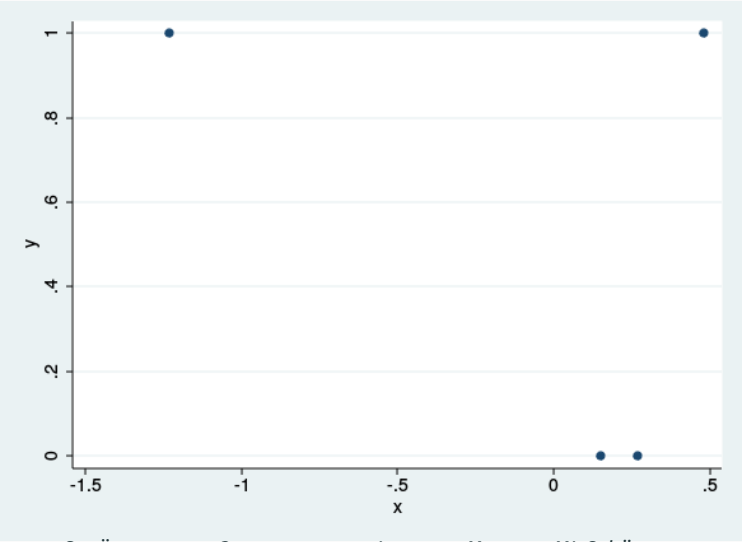
- Mit dem Newton-Raphson-Algorithmus kann man iterativ die Nullstelle(n) einer Funktion finden: $x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$ (basiert auf Taylor-Serie, die den Zusammenhang zwischen zwei Funktionswerten approximiert)
- (Achtung: $f(x_n)$ ist hier die Score-Funktion, d.h. die erste Ableitung der Log-Likelihood-Funktion)
- 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

Wie geht der Computer vor?

y	x	$\boldsymbol{X}\tilde{\beta}$	LL
0	0.15	0	-0.693
0	0.27	0	-0.693
1	0.48	0	-0.693
1	-1.23	0	-0.693

- Startwerte für $\beta_0, \beta_1 : 0; 0$
- Initiale LL -2.77

Wie geht der Computer vor?



Wie geht der Computer vor?

y	x	$\boldsymbol{X}\tilde{\beta}$	LL
0	0.15	-0.274	-0.566
0	0.27	-0.461	-0.489
1	0.48	-0.789	-1.163
1	-1.23	1.879	-0.142

- LL nach vier Iterationen -2.36
- Parameterschätzungen nach vier Iterationen -.04; -1.56

Wie geht der Computer vor?

```
. logit y x,trace
------------------------------------------------------------------------------
Iteration 0:
Parameter vector:
        y:      y:
        x   _cons
r1      0       0
                                              log likelihood = -2.7725887
------------------------------------------------------------------------------
Iteration 1:
Parameter vector:
             y:          y:
             x       _cons
r1  -1.453236   -.1198919
                                              log likelihood = -2.3639066
------------------------------------------------------------------------------
Iteration 2:
Parameter vector:
             y:          y:
             x       _cons
r1  -1.558763   -.0404538
                                              log likelihood = -2.3599661
------------------------------------------------------------------------------
Iteration 3:
Parameter vector:
             y:          y:
             x       _cons
r1  -1.569448   -.0374612
                                              log likelihood = -2.3599503
------------------------------------------------------------------------------
Iteration 4:
Parameter vector:
             y:          y:
             x       _cons
r1  -1.569508   -.0374432
                                              log likelihood = -2.3599503
------------------------------------------------------------------------------

Logistic regression                             Number of obs     =          4
                                                LR chi2(1)        =       0.83
                                                Prob > chi2       =     0.3636
Log likelihood = -2.3599503                     Pseudo R2         =     0.1488

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |  -1.569508    2.04407    -0.77   0.443    -5.575813    2.436796
       _cons |  -.0374432   1.122075    -0.03   0.973    -2.236669    2.161783
------------------------------------------------------------------------------
```

Mögliche Probleme?

- Achtung: Da f bereits die erste Ableitung der Log-Likelihood-Funktion ist, ist f' die (partielle) zweite Ableitung der Log-Likelihood-Funktion $\rightarrow$ Hesse-Matrix
- Inverse der Hesse-Matrix $\rightarrow$ Varianz-Kovarianz-Matrix der Koeffizienten (Standardfehler)
- Startwerte und lokale Extremwerte
- ML-Schätzung für die hier vorgestellten Modelle normalerweise unproblematisch, sehr schnelle und sichere Konvergenz

Zusammenfassung/Ausblick

Zusammenfassung

- Im GLM ist y ein Zufallswert aus einer statistischen Verteilung
- Der wesentliche Parameter dieser Verteilung ist μ
- μ ist über eine (nicht-lineare) Linkfunktion θ mit dem linearen Prediktor $\mathbf{X}\beta$ verbunden
- Mit ML können β und evtl. weitere Parameter iterativ geschätzt werden (+Standardfehler)
- Grundlage für sehr viele moderne Verfahren

Ausblick

- Nächste Woche: Analyse sozialer Netzwerke
- Literatur zur Einführung:
  - Ebook von Hanneman/Riddle:
    `http://www.faculty.ucr.edu/~hanneman/nettext/`
  - Scott (2000): Social Network Analysis