Angewandte Statistik

Julian Hinz — Universität Bielefeld

Session 5

Parametrische Regression

Logistische Regression

Lernziel

(Vor-)letzte Woche

- *Parametrische* Regression: Poisson Regression
- Nichtparametrische Regression: Splines

Heute

- Logistische Regression

Motivierendes Beispiel: Donner Party

April 1846: eine Gruppe von insg. 87 Siedlern — die “Donner Party” — macht sich in Illinois auf den Weg nach Kalifornien, in der Hoffnung auf ein besseres Leben.

Donner Party

- Für diese Trecks gab es nur begrenztes Zeitfenster: April–September
- Donner Party kam aber erst im Mai richtig in Gang (also sehr spät)
- Neue Route gen Westen wurde ausprobiert: Desaster, in den Rockies eingeschneit

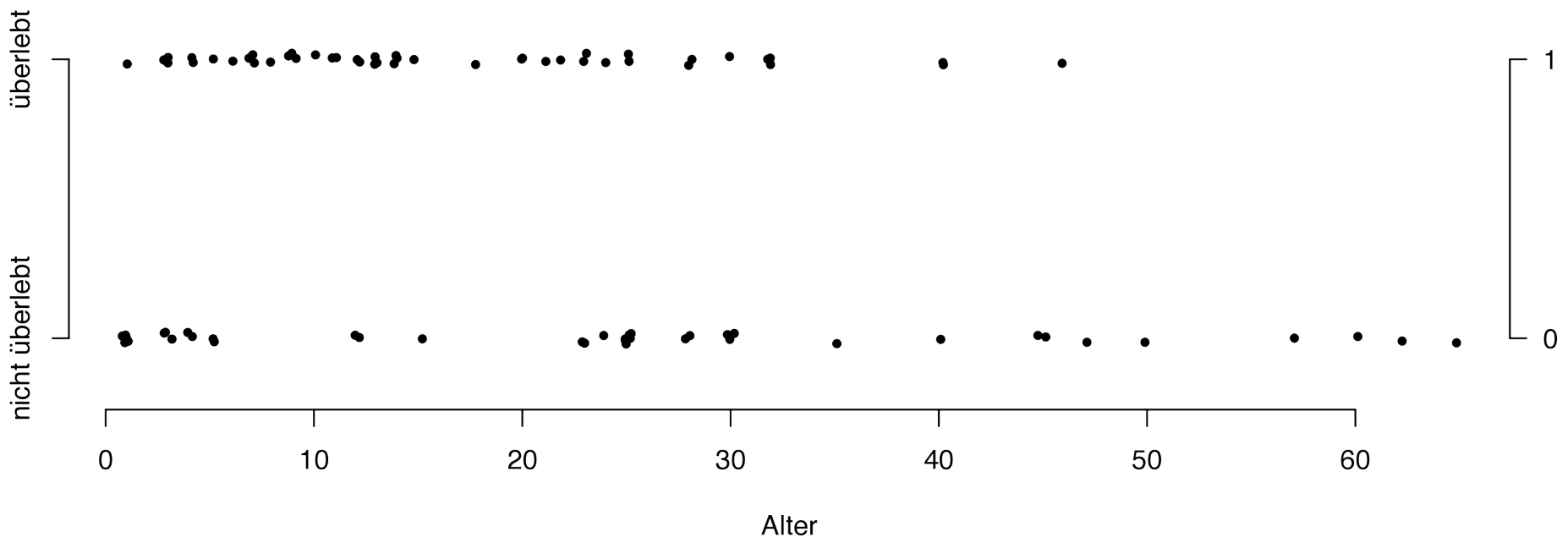
Donner Party

insgesamt 40 der 87 Siedler kamen ums Leben

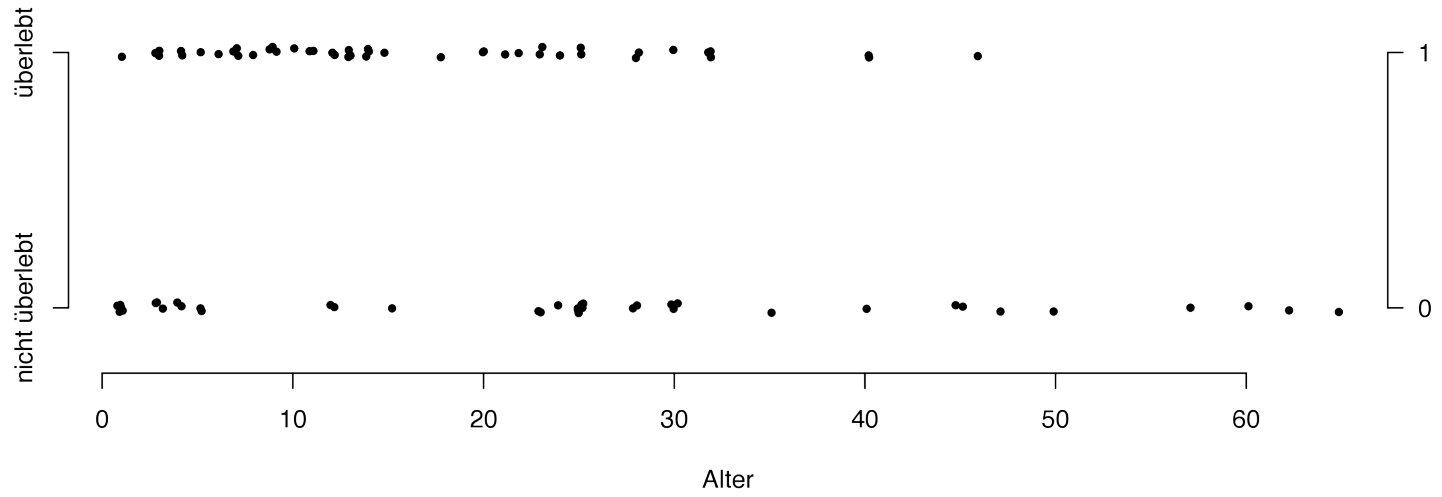
	Überlebt	Geschlecht	Alter	Familiengröße
Antoine	Nein	M	23	1
Edward	Ja	M	13	9
Isabella	Ja	W	1	9
James	Ja	M	4	9
Elisabeth	Nein	W	45	16
Margaret	Nein	W	1	4
⋮	⋮	⋮	⋮	⋮
Sarah	Ja	W	22	12

Quelle: Grayson (1990, Journal of Anthrop. Research)

Donner Party



Donner Party



- jüngeren Reisenden scheinen eher überlebt zu haben
- Auch Familiengröße und Geschlecht Einfluss auf Überlebenswahrscheinlichkeit
  - 56% der Männer, aber nur 30% der Frauen starben

Regressionsmodell

- Zusammenhang zwischen Überlebenswahrscheinlichkeit und Alter, Geschlecht & Familiengröße quantifizieren
- Definiere

$$Y_i = \begin{cases} 1 & \text{falls das } i\text{-te Mitglied der Reisegruppe überlebt hat;} \\ 0 & \text{sonst, d.h. falls das } i\text{-te Mitglied der Reisegruppe gestorben ist,} \end{cases}$$

- x_i das Alter dieser i-ten Person, …

Regressionsmodell

- Zielvariable Y_i also binär
- Bernoulli-Variablen mit Erfolgswahrscheinlichkeit $\pi_i = P(Y_i = 1)$
- Insbesondere ist

$$E(Y_i) = \pi_i \cdot 1 + (1 - \pi_i) \cdot 0 = \pi_i$$

Modellierung der Überlebenswahrscheinlichkeit

- Wir suchen also nach geeigneten Regressionsmodell der Form

$$\pi_i = P(Y_i = 1) = E(Y_i) = f(x_i)$$

Modellierung der Überlebenswahrscheinlichkeit

- Warum nicht einfach das herkömmliche lineare Regressionsmodell verwenden?

$$E(Y_i) = \beta_0 + \beta_1 x_i$$

- Linearer Prädiktor, $\eta_i = \beta_0 + \beta_1 x_i$, würde für bestimmte x_i-Werte Überlebenswahrscheinlichkeiten $\pi_i < 0$ und $\pi_i > 1$ vorhersagen!

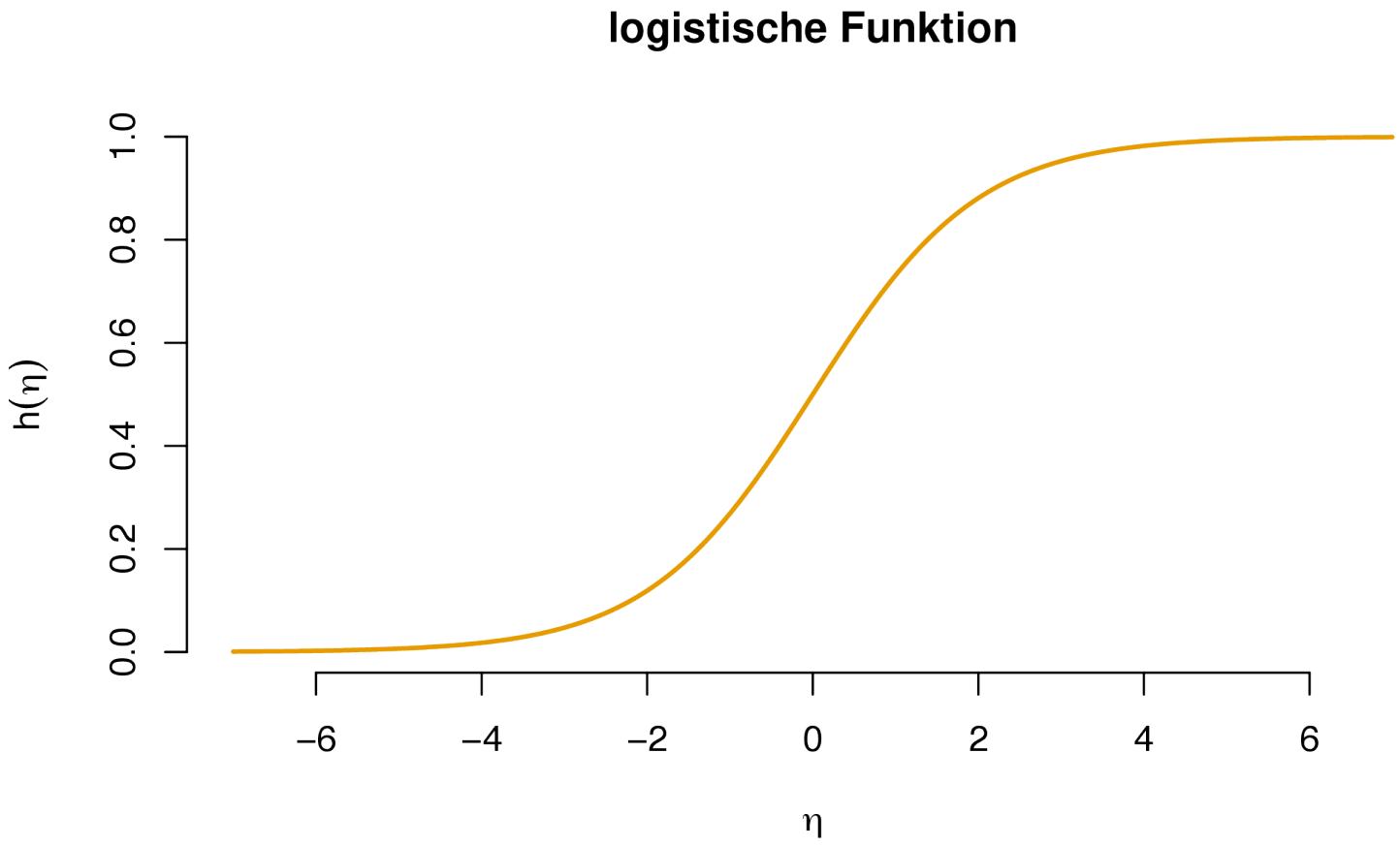
Bernoulli-verteilte Zielgrößen

- Wie bei der Poissonregression transformieren wir den linearen Prädiktor:

$$\pi_i = P(Y_i = 1) = E(Y_i) = h(\beta_0 + \beta_1 x_i)$$

- Idee: Hohe Werte des linearen Prädiktors η_i sollen hohen Überlebenswahrscheinlichkeiten π_i entsprechen
- Wir suchen also Funktion h von $\mathbb{R}$ nach $[0, 1]$, so dass:
  - Alle Werte in $[0, 1]$ angenommen werden
  - Die Funktion streng monoton wachsend ist

Logistische Funktion



Logistische Funktion

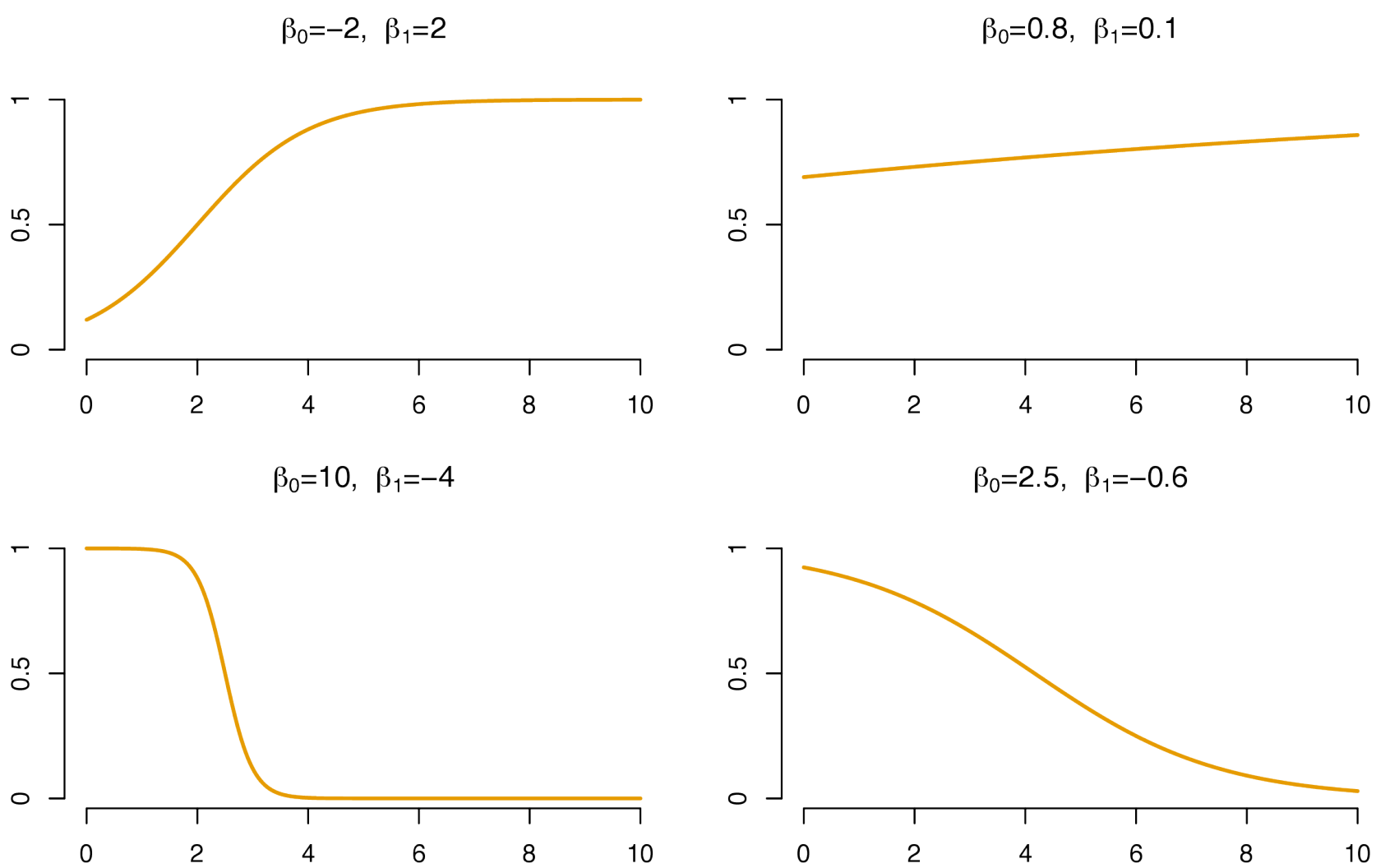
$$h : \mathbb{R} \longrightarrow [0, 1], \qquad \eta \mapsto \frac{e^{\eta}}{1 + e^{\eta}}$$

- Streng monoton wachsend
- Mit $h(0) = 0.5$,
- $\lim_{\eta \to -\infty} h(\eta) = 0$, und
- $\lim_{\eta \to \infty} h(\eta) = 1$.
- h ist die **logistische Funktion** (`plogis()` in R)

Einfache logistische Regression

$$\pi_i = P(Y_i = 1) = E(Y_i) = \frac{e^{\beta_0+\beta_1 x_i}}{1 + e^{\beta_0+\beta_1 x_i}}$$

- π_i also Anwendung der logistischen Funktion auf üblichen linearen Prädiktor
- Vorsicht bei Interpretation:
  - Vorzeichen von β_1 gibt an, ob Effekt von x_i positiv oder negativ
  - (absolute) Größe von β_1 zeigt wie steil die Regressionsfunktion ist
  - kein **marginaler Effekt**


β_0=−2, β_1=2
β_0=0.8, β_1=0.1
β_0=10, β_1=−4
β_0=2.5, β_1=−0.6
0
0.5
1
0
2
4
6
8
10

Logistische Regression

- logistisches Regressionsmodell für unabhängige Zufallsvariablen Y_i mit Wertebereich $\{0, 1\}$ hat die Form

$$
\begin{aligned}
Y_i &\sim Bern(\pi_i); \\
\pi_i &= P(Y_i = 1) = E(Y_i) = \frac{e^{\eta_i}}{1 + e^{\eta_i}}; \\
\eta_i &= \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}
\end{aligned}
$$

- lineare Prädiktor $\eta = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$ ebenso flexibel wie im linearen Modell

Parameterschätzung

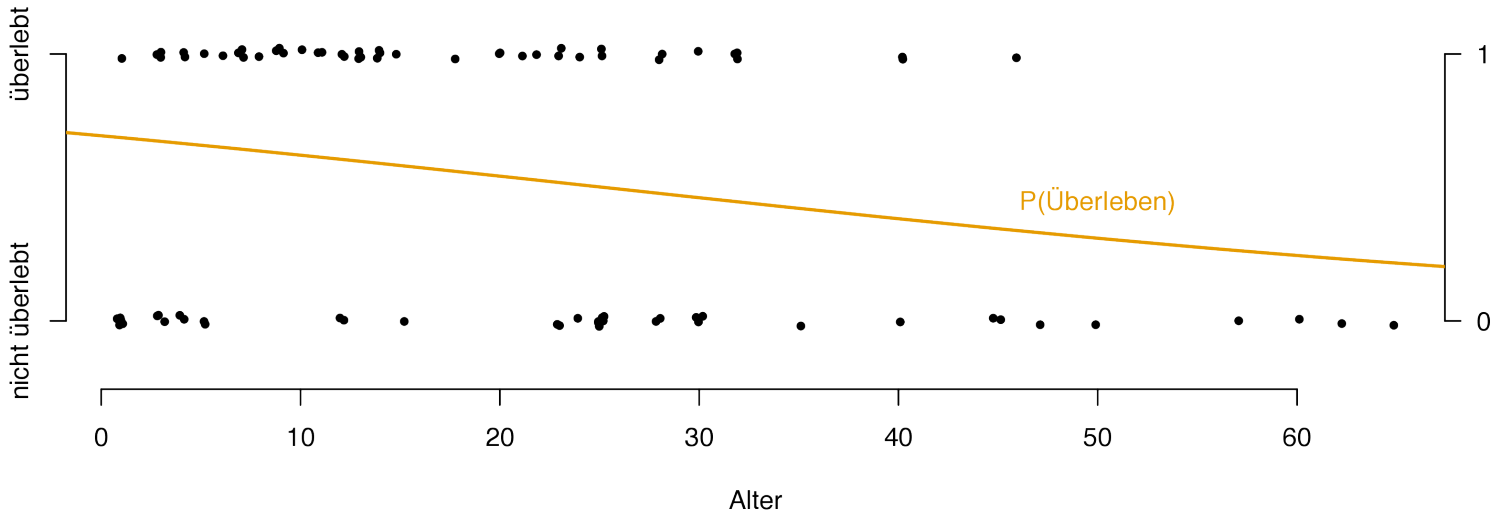
- Wie im Poissonregressionsmodell besteht Problem, dass die Varianz der Y_i nicht konstant ist

$$\text{Var}(Y_i) = \pi_i(1 - \pi_i)$$

mit $\pi_i = \frac{e^{\eta_i}}{1+e^{\eta_i}}$ und $\eta_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$

- Varianz maximal für $\pi_i = 0.5$ (d.h. für $\eta_i = 0$) und klein für $\pi_i \approx 0$ ($\eta_i \to -\infty$) oder $\pi_i \approx 1$ ($\eta_i \to \infty$)
- Schätzung der Parameter: Methode der iterativ gewichteten kleinsten Quadrate

Logistische Regression im Beispiel




```
mod <- glm(ueberleben ~ alter, family=binomial)
summary(mod)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.81733    0.37549   2.177   0.0295 *
alter       -0.03237    0.01509  -2.145   0.0320 *
```

Hypothesentest im logistischen Regressionsmodell

- Exakt wie bei Poissonregression beziehen sich die im R-Output gegebenen p-Werte auf den Test von

$$H_0 : \beta_j = 0 \quad \text{vs.} \quad H_1 : \beta_j \neq 0$$

Unter H_0 (d.h. für $\beta_j = 0$) gilt

$$Z = \frac{\widehat{\beta}_j}{\widehat{\sigma}_{\widehat{\beta}_j}} \sim N(0, 1) \text{ approx.}$$

- Es handelt sich also auch hier um einen approximativen Gaußtest

Interpretation der Modellparameter

- Wie können wir hier $\hat{\beta}_1 = -0.03237$ interpretieren?
- “Wenn sich der Wert der erklärenden Variable um eins erhöht, dann … (?)”
- **Odds**:

$$\text{Odds}(\text{Erfolg}) = \frac{P(\text{Erfolg})}{P(\text{kein Erfolg})}$$

- Beispiel: wenn 90%-Chance die Klausur zu bestehen, dann Odds 9/1

Interpretation der Modellparameter

- **Odds Ratio**: Maß für die Änderung der Odds bei Änderung der erklärenden Variablen x um 1:

$$\text{Odds Ratio} = \frac{\text{Odds(Erfolg für } x+1)}{\text{Odds(Erfolg für } x)}$$

- Für einfaches logistisches Modell, $E(Y) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$, erhalten wir

$$\begin{aligned}\text{Odds Ratio} &= \frac{\text{Odds}^{x+1}}{\text{Odds}^{x}} \\ &= \frac{e^{\beta_0+\beta_1(x+1)}}{e^{\beta_0+\beta_1(x)}} \\ &= e^{\beta_1}\end{aligned}$$

Interpretation der Modellparameter

$$\begin{aligned}\text{Odds Ratio} &= \frac{\text{Odds}^{x+1}}{\text{Odds}^{x}} \\ &= \frac{e^{\beta_0+\beta_1(x+1)}}{e^{\beta_0+\beta_1(x)}} \\ &= e^{\beta_1}\end{aligned}$$

- Interpretation: Odds zu überleben verringern sich um den Faktor e^{β_1} für jedes zusätzliche Lebensjahr
- Interpretation im Donner Party Beispiel: $e^{\widehat{\beta}_1} = e^{-0.03237} = 0.968$, d.h. Odds zu überleben verringern sich um Faktor 0.968 für jedes zusätzliche Lebensjahr

Mehrere erklärende Variablen

```
mod <- glm(ueberleben ~ alter + geschlecht + fam.groesse, family=binomial)
summary(mod)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.14052    0.57884   0.243   0.8082
alter       -0.02829    0.01558  -1.816   0.0694 .
geschlecht   0.91151    0.49551   1.840   0.0658 .
fam.groesse  0.02942    0.04395   0.669   0.5033
```

(hierbei ist Geschlecht $= 1$ für die weiblichen Reisenden; sonst $= 0$)

Polynome im Prädiktor

Genauere Betrachtung des Streudiagramms zeigt:

- Babys & kleine Kinder sowie ältere Mitglieder hatten besonders niedrige Überlebenswahrscheinlichkeit
- Teenager/junge Erwachsene hatten höchste Überlebenswahrscheinlichkeit

Genau wie im linearen Regressionsmodell kann man quadratischen Term im Prädiktor abbilden

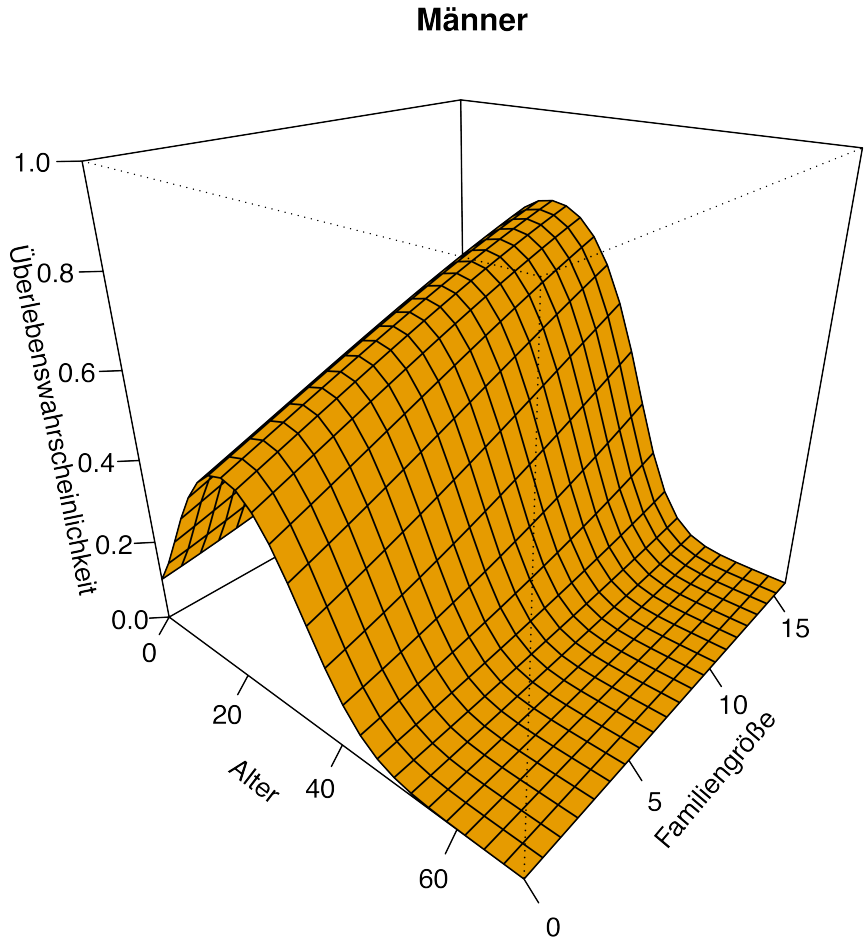
$$\eta_i = \beta_0 + \beta_1 \cdot \text{alter}_i + \beta_2 \cdot \text{alter}_i^2 + \beta_3 \cdot \text{geschlecht}_i + \beta_4 \cdot \text{fam.groesse}_i$$

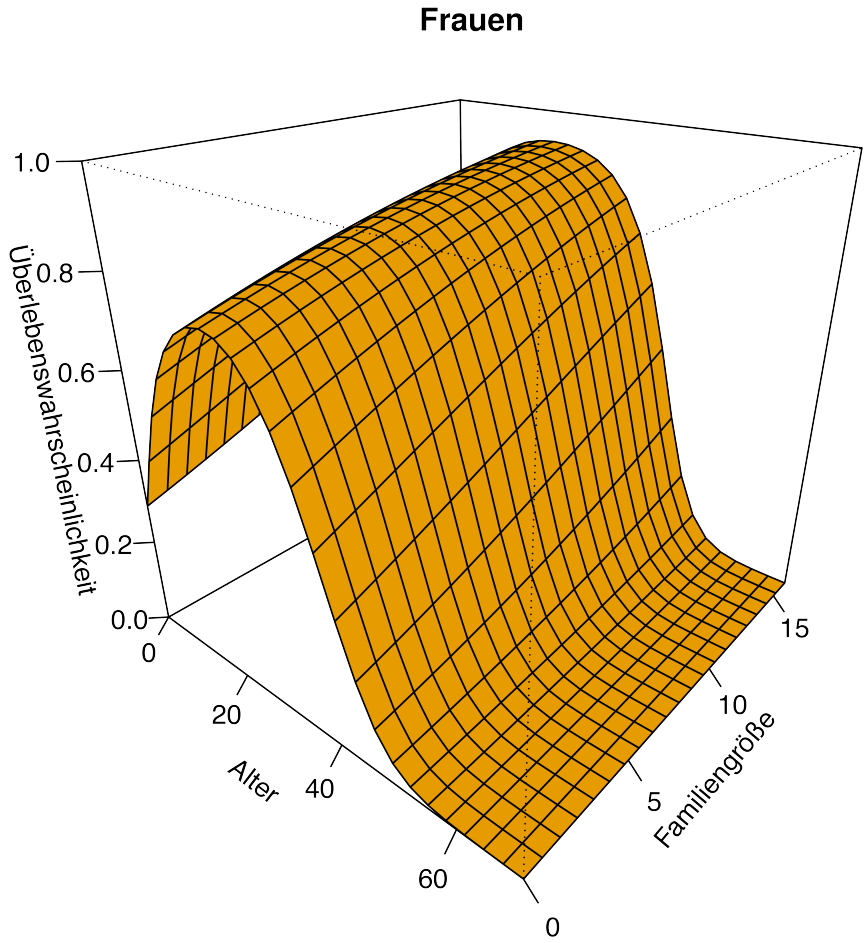
Polynome im Prädiktor

```
mod <- glm(ueberleben ~ alter + I(alter^2) + geschlecht + fam.groesse, family=binomial)
summary(mod)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.164255   0.941828  -2.298  0.02157 *
alter        0.199589   0.074621   2.675  0.00748 **
I(alter^2)  -0.004871   0.001672  -2.913  0.00358 **
geschlecht   1.271679   0.580658   2.190  0.02852 *
fam.groesse  0.091228   0.051350   1.777  0.07564 .
```

Polynome im Prädiktor





Weitere Anwendungen von logistischer Regression

- Logistische Regression immer dann sinnvoll, wenn wir **binäre Zielvariablen** modellieren
- Beispiele gibt es hier sehr viele:
  - Überleben eines Krebspatienten (ja/nein) ~ Blutwerte, Stadium, Alter, …
  - Produktkauf (ja/nein) ~ vergangene Käufe, Gesamtumsatz, …
  - E-Mail Spam (ja/nein) ~ Nutzung farbiger Schrift, Wörter wie “Gewinne”, …
  - Teilnahme an Wahl (ja/nein) ~ sozioökonomische Faktoren
  - Arbeitslosigkeit (ja/nein) ~ Bildung, Herkunft, Alter, …
  - Bestehen der Klausur (ja/nein) ~ Fachsemester, Abinote, Lernaufwand, …
  - Kreditwürdigkeit (ja/nein) ~ Einkommen, Alter, Lebensumstände, …

Beispiel Kreditscoring

SCHUFA nutzt logistische Regression, um Verbraucherinnen und Verbrauchern ihre individuellen **Kreditausfallswahrscheinlichkeiten** zuzuordnen: Meine SCHUFA

In Kurzform:

- Die SCHUFA hat Daten zu mehreren Millionen Kreditverträgen.
- Insbesondere enthalten:
  - Information, ob Kredit letztendlich zurückgezahlt wurde.
  - Sehr viele weitere personen- und kreditbezogene Variablen.
- Hieraus wird ein **logistisches Regressionsmodell mit Zielvariable Kreditrückzahlung ja/nein** gebildet.
- Dieses wird für neue Kundinnen und Kunden eingesetzt, um die Kreditausfallswahrscheinlichkeit zu ermitteln.

Beispielsdatensatz zu Kreditscoring

Relevante Variablen:

- Kunde, kreditwürdig, laufendes Konto bei Bank, Laufzeit des Kredits, Höhe des Kredits
- $>$ 3 Jahre in Job, Rate $>$ 25% des Einkommens, $>$ 3 Jahre in derz. Wohnung
- Alter, problem. Zahlungen in Vergangenheit

Hier:

- Stichprobe von 1000 Kunden einer süddeutschen Bank
- davon 30% kreditunwürdig und 70% kreditwürdig
- `kredit = 0`: nicht vereinbarungsgemäß zurückgezahlt

1. Beispielsrechnung

Nun kommt ein 24-jähriger in die Bank und möchte einen Kredit über 10 Tsd. DM aufnehmen. Er gibt an:

- kein laufendes Konto
- Laufzeit des Kredits soll 72 Monate sein
- er hat gerade erst einen neuen Job aufgenommen
- er ist auch gerade umgezogen
- die zu zahlende Rate entspricht 35% seines Einkommens
- er hat noch nie zuvor einen Kredit aufgenommen

Kreditausfallwahrscheinlichkeit gemäß des Modells: `1-plogis(eta)`

2. Beispielsrechnung

Als nächstes kommt ein 55-jähriger in die Bank und möchte einen Kredit über 4 Tsd. DM aufnehmen. Er gibt an:

- laufendes Konto
- Laufzeit des Kredits soll 12 Monate sein
- er arbeitet seit mehr als 3 Jahren im selben Job
- er lebt seit mehr als 3 Jahren in der derzeitigen Wohnung
- die zu zahlende Rate entspricht <25% seines Einkommens
- er hat alle bisher aufgenommenen Kredite wie vereinbart zurückgezahlt

Übersicht behandelte parametrischen Regressionsmodelle

- Lineare Regression (mit Annahme der Normalität):

$$Y_i \sim N(\mu_i, \sigma^2), \quad \mu_i = E(Y_i) = \eta_i$$

- Poissonregression:

$$Y_i \sim Po(\lambda_i), \quad \lambda_i = E(Y_i) = e^{\eta_i}$$

- Logistische Regression:

$$Y_i \sim Bern(\pi_i), \quad \pi_i = E(Y_i) = \frac{e^{\eta_i}}{1 + e^{\eta_i}}$$

- hier immer: η_i linearer Prädiktor für Beobachtung $i, i = 1 \ldots, n$:

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \quad + \beta_p x_{ip}$$