Kapitel 2

Parametrische Regression

2.3 Logistische Regression

Fragen? https://www.menti.com/alneo5om9sq5.

Motivierendes Beispiel: Donner Party

April 1846: eine Gruppe von insgesamt 87 Siedler*innen — die Donner Party — macht sich in Illinois auf den Weg nach Kalifornien, in der Hoffnung auf ein besseres Leben.

Fragen? https://www.menti.com/alneo5om9sq5.

- für diese Trecks gab es nur ein begrenztes Zeitfenster: April-September
- die Donner Party kam aber erst im Mai richtig in Gang (also sehr spät)
- eine neue Route Richtung Westen wurde ausprobiert, stellte sich aber schnell als Desaster heraus — letztendlich wurden sie in den Rockies eingeschneit

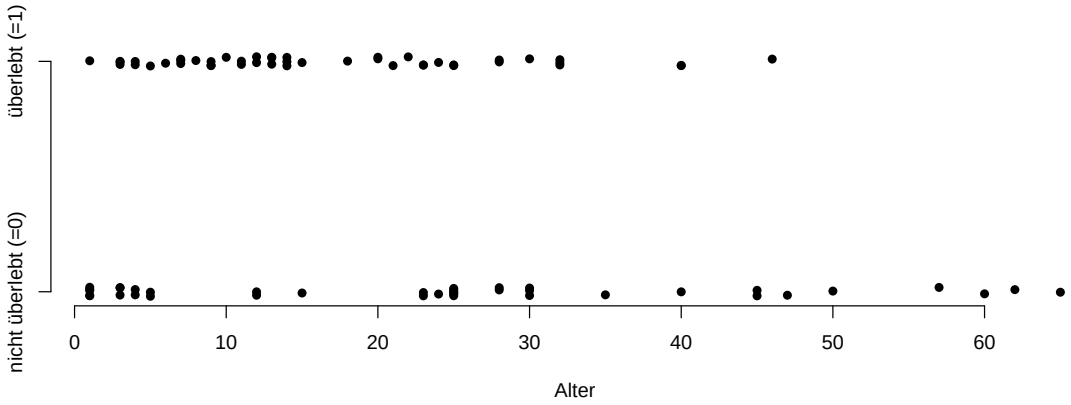
Fragen? https://www.menti.com/alneo5om9sq5.

⇝ **insgesamt 40 der 87 Siedler*innen kamen ums Leben**

Ueberleben	Geschlecht	Alter	Familiengroesse
Nein	M	23	1
Ja	M	13	9
Ja	W	1	9
Ja	M	4	9
...	...	...	...
Ja	M	23	1
Nein	M	24	2
Ja	W	25	2

Quelle: Grayson (1990, Journal of Anthropological Research)

Fragen? https://www.menti.com/alneo5om9sq5.



- die jüngeren Reisenden scheinen eher überlebt zu haben
- auch Familiengröße[1] und Geschlecht[2] scheinen Einfluss auf die Überlebensw'keit gehabt zu haben

[1] empirische Korrelation: 0.17

[2] 56% der Männer, aber nur 30% der Frauen starben

Fragen? https://www.menti.com/alneo5om9sq5.

Modellierung der Überlebenswahrscheinlichkeit

Ziel: per Regressionsmodell den Zusammenhang zwischen Überlebenswahrscheinlichkeit und Alter, Geschlecht und Familiengröße quantifizieren.

Sei dazu zunächst

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

und sei x_{i1} das Alter der i-ten person.

Die Zielvariable Y_i ist also binär, sodass wir die als **Ausgang einer Bernoulli-Variablen** mit Erfolgswahrscheinlichkeit $\pi_i = P(Y_i = 1)$ betrachten. Insbesondere ist

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

Wir wollen jetzt $\pi_i = E(Y_i)$ als Funktion von x_{i1} modellieren!

Fragen? https://www.menti.com/alneo5om9sq5.

Wir suchen also nach einem geeigneten Regressionsmodell der Form

$$\pi_i = P(Y_i = 1) = E(Y_i) = f(x_{i1})$$

Warum nicht einfach das herkömmliche lineare Regressionsmodell,

$$E(Y_i) = \beta_0 + \beta_1 x_{i1}?$$

Der lineare Prädiktor, $\eta_i = \beta_0 + \beta_1 x_{i1}$, würde für bestimmte x_{i1}-Werte Überlebenswahrscheinlichkeiten $\pi_i < 0$ und $\pi_i > 1$ vorhersagen.

Regressionsmodell für Bernoulli-verteilte Zielgrößen

Wie bei der Poissonregression[3] transformieren wir den linearen Prädiktor:

$$\pi_i = P(Y_i = 1) = E(Y_i) = h(\underbrace{\beta_0 + \beta_1 x_{i1}}_{\eta_i})$$

Idee: hohe (niedrige) Werte des linearen Prädiktors η_i sollen hohen (niedrigen) Überlebenswahrscheinlichkeiten π_i entsprechen.

Wir suchen demnach eine Funktion h von $\mathbb{R}$ nach $[0, 1]$, so dass

- alle Werte in $[0, 1]$ angenommen werden
- die Funktion streng monoton wachsend ist

[3] bei der Poissonregression haben wir den linearen Prädiktor η_i einfach exponenziert

Fragen? https://www.menti.com/alneo5om9sq5.

Logistische Funktion

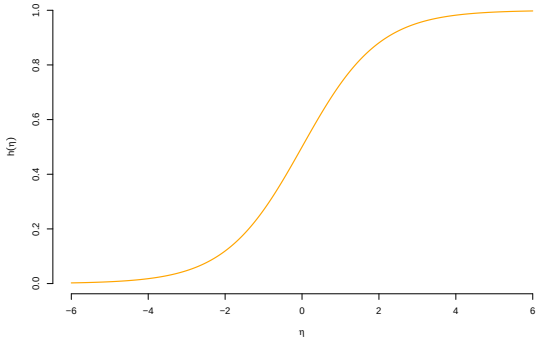
Die Funktion

$$\mathbb{R} \to [0, 1]$$
$$\eta \to \frac{e^\eta}{1 + e^\eta}$$

ist

- streng monoton wachsend
- mit $h(0) = 0.5$,
- $\lim_{\eta \to -\infty} h(\eta) = 0$ und
- $\lim_{\eta \to \infty} h(\eta) = 1$

Man nennt h die **logistische Funktion** (in R: `plogis()`).



Fragen? https://www.menti.com/alneo5om9sq5.

Einfache logistische Regression

Entsprechend ergibt sich das Modell

$$\pi_i = P(Y_i = 1) = E(Y_i) = \frac{e^{\beta_0 + \beta_1 x_{i1}}}{1 + e^{\beta_0 + \beta_1 x_{i1}}}$$

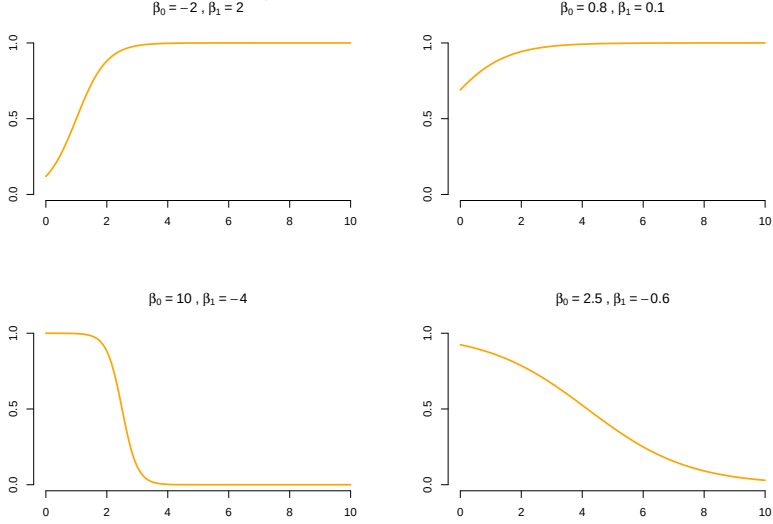
Letztendlich ergibt sich π_i also einfach durch Anwendung der logistischen Funktion auf den üblichen linearen Prädiktor.

Erste Feststellungen hinsichtlich der Interpretation:

- das Vorzeichen von β_1 gibt an, ob der Effekt von x_{i1} positiv oder negativ ist
- die (absolute) Höhe von β_1 kontrolliert, wie steil die Regressionsfunktion ist

Fragen? https://www.menti.com/alneo5om9sq5.

Wie die Regressionsfunktion $e^{\beta_0+\beta_1 x_{i1}}/(1 + e^{\beta_0+\beta_1 x_{i1}})$ aussehen könnte



Fragen? https://www.menti.com/alneo5om9sq5.

Logistische Regression (allgemeiner Fall mit p erklärenden Variablen)

Das **logistische Regressionsmodell**, für unabhängige Zufallsvariablen Y_i, jeweils mit Wertebereich $\{0, 1\}$, hat die Form

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

Der lineare Prädiktor $\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$ ist wieder ebenso flexibel wie im linearen Modell was quadratische Terme, Interaktionen etc. angeht.

Fragen? https://www.menti.com/alneo5om9sq5.

Parameterschätzung

Wie schon im Poissonregressionsmodell besteht das Problem, dass die Varianz der Y_i nicht konstant ist:

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

wobei $\pi_i = e^{\eta_i}/(1 + e^{\eta_i})$, mit dem Prädiktor $\eta_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$.

Die Varianz ist maximal für $\pi_i = 0.5$ (d.h. für $\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} = 0$) und klein für $\pi_i \approx 0$ ($\eta_i \to -\infty$) oder $\pi_i \approx 1$ ($\eta_i \to \infty$).

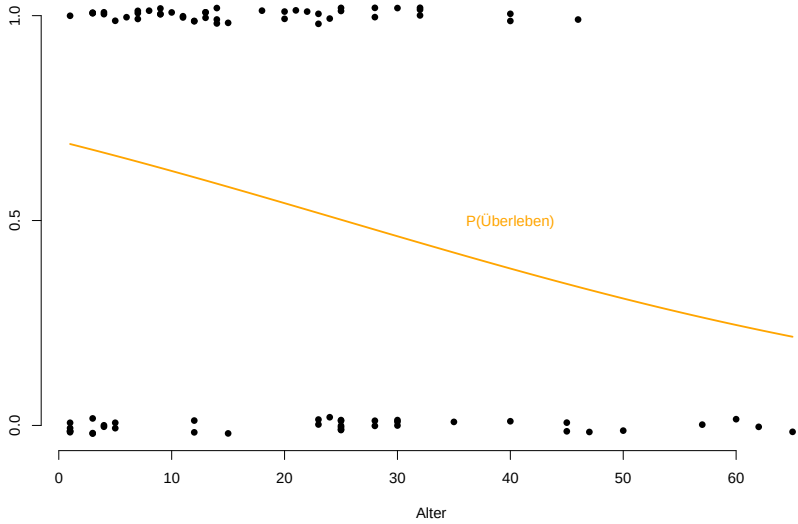
$\rightsquigarrow$ zur Schätzung der Parameter verwenden wir wieder die Methode der iterativ gewichteten kleinsten Quadrate.

Fragen? https://www.menti.com/alneo5om9sq5.

```r
donner_party = read.csv("https://tinyurl.com/mrkjynja")
mod = glm(Ueberleben ~ Alter, family = "binomial", data = donner_party)
summary(mod)
```

```
Call:
glm(formula = Ueberleben ~ Alter, family = "binomial", data = donner_party)

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 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 117.26  on 84  degrees of freedom
Residual deviance: 112.25  on 83  degrees of freedom
AIC: 116.25

Number of Fisher Scoring iterations: 4
```

Fragen? https://www.menti.com/alneo5om9sq5.



Fragen? https://www.menti.com/alneo5om9sq5.

Interpretation der Modellparameter

Wie können wir hier $\hat{\beta}_1 = -0.03237$ interpretieren? Wir wollen wieder eine Aussage treffen wie

"Wenn sich der Wert der erklärenden Variable um eins erhöht, dann...

Hierzu definieren wir zunächst die **Odds** eines Erfolges[4],

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

Falls Sie z.B. eine 90%-Wahrscheinlichkeit haben, die Klausur zu bestehen, dann sind Ihre Odds 9/1 (“9 zu 1” — Sie bestehen in 9 von 10 Fällen).

[4] in unserem Beispiel: zu überleben

Fragen? https://www.menti.com/alneo5om9sq5.

Interpretation der Modellparameter (in der Vorlesung)

Wir betrachten jetzt die **Odds Ratio** als ein Maß für die Änderung der Odds bei einer Änderung der erklärenden Variablen x um eine Einheit:

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

Für das einfache logistische Modell, $E(Y) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$ erhalten wir:

Odds(Erfolg) =

Odds Ratio =

Interpretation im Donner Party Beispiel:

- $e^{\hat{\beta}_1} = e^{-0.03237} = 0.968$, d.h. die Odds zu überleben verringern sich um den Faktor 0.968 für jedes zusätzliche Lebensjahr (bzw. verringern sich um 3.2%)

Fragen? https://www.menti.com/alneo5om9sq5.

Logistische Regression mit mehreren erklärenden Variablen

```r
mod = glm(Ueberleben ~ Alter + Geschlecht + Familiengroesse, family = "binomial", data = donner_party)
summary(mod)
```

```
Call:
glm(formula = Ueberleben ~ Alter + Geschlecht + Familiengroesse,
    family = "binomial", data = donner_party)

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 .
Familiengroesse 0.02942    0.04395   0.669   0.5033
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 117.26  on 84  degrees of freedom
Residual deviance: 107.39  on 81  degrees of freedom
AIC: 115.39

Number of Fisher Scoring iterations: 4
```

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

Fragen? https://www.menti.com/alneo5om9sq5.

Polynomiale Terme im Prädiktor

Genauere Betrachtung des Streudiagramms (Slide 6):

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

Genau wie im linearen Regressionsmodell kann man einen solchen Effekt durch einen quadratischen Term im Prädiktor abbilden:

$$\eta_i = \beta_0 + \beta_1 \cdot \text{Alter}_i + \beta_2 \cdot \text{Alter}_i^2 + \beta_3 \cdot \text{Geschlecht}_i + \beta_4 \cdot \text{Familiengroesse}_i$$

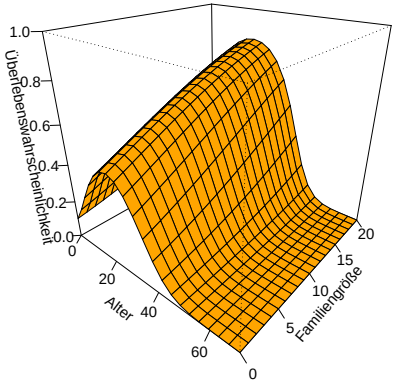
Fragen? https://www.menti.com/alneo5om9sq5.

```r
mod = glm(Ueberleben ~ Alter + I(Alter^2) + Geschlecht + Familiengroesse, family = "binomial",
          data = donner_party)
summary(mod)
```

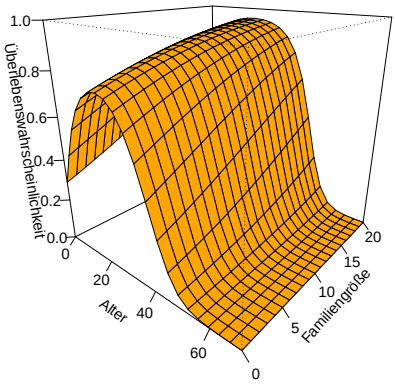
```
Call:
glm(formula = Ueberleben ~ Alter + I(Alter^2) + Geschlecht +
    Familiengroesse, family = "binomial", data = donner_party)

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 *
Familiengroesse  0.091228   0.051350   1.777  0.07564 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 117.258  on 84  degrees of freedom
Residual deviance:  94.571  on 80  degrees of freedom
AIC: 104.57

Number of Fisher Scoring iterations: 5
```

Fragen? https://www.menti.com/alneo5om9sq5.

Männer



Frauen



Fragen? https://www.menti.com/alneo5om9sq5.

Weitere Anwendungen von logistischer Regression

Logistische Regression ist immer dann sinnvoll, wenn wir **binäre Zielvariablen** modellieren wollen.

Beispiele gibt es hier sehr viele:

- Produktkauf (ja/nein) $\sim$ vergangene Käufe, personalisierte Werbung, ...
- Teilnahme an Wahl (ja/nein) $\sim$ sozioökonomische Faktoren, ...
- Arbeitslosigkeit (ja/nein) $\sim$ Bildung, Herkunft, Alter, ...
- Kreditwürdigkeit (ja/nein) $\sim$ Einkommen, Alter, Lebensumstände, ...

Fragen? https://www.menti.com/alneo5om9sq5.

Beispiel Kreditscoring

Die SCHUFA beispielsweise nutzt logistische Regression, um Verbraucher*innen ihre individuellen **Kreditausfallswahrscheinlichkeiten** zuzuordnen:

https://www.meineschufa.de/de/scoring

- 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 Kund*innen eingesetzt, um die Kreditausfallswahrscheinlichkeit zu ermitteln

Fragen? https://www.menti.com/alneo5om9sq5.

Beispielsdatensatz zum Kreditscoring

ID	Hoehe	Geschlecht	Alter	Zahlungsausfall
1	20000	0	24	1
2	120000	0	26	1
3	90000	0	34	0
4	50000	0	37	0
5	50000	1	57	0
6	50000	1	37	0
...	...	...	...	...

Fragen? https://www.menti.com/alneo5om9sq5.

```
Call:
glm(formula = Zahlungsausfall ~ Hoehe + Geschlecht + Alter, family = "binomial",
    data = kredit_daten)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.117e+00  5.626e-02 -19.855  < 2e-16 ***
Hoehe       -3.363e-06  1.272e-07 -26.445  < 2e-16 ***
Geschlecht   1.552e-01  2.870e-02   5.409 6.33e-08 ***
Alter        8.657e-03  1.499e-03   5.773 7.77e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 31705  on 29999  degrees of freedom
Residual deviance: 30863  on 29996  degrees of freedom
AIC: 30871

Number of Fisher Scoring iterations: 4
```

Beispielrechnung:

Ein 31-jähriger Mann kommt in die Bank und möchte einen Kredit über 100,000 Dollar aufnehmen.

Kreditausfallwahrscheinlichkeit nach unserem Modell:

```
plogis(mod$coef[1] + mod$coef[2] * 100000 + mod$coef[3] * 1 + mod$coef[4] * 31)
```

```
(Intercept)
  0.2631175
```

$\rightsquigarrow$ Das wäre ein sehr risikoreicher Kredit...

Fragen? https://www.menti.com/alneo5om9sq5.

Kurzübersicht der behandelten Regressionsmodelle

Lineare Regression (mit Annahme der Normalität des Fehlerterms):

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

Poissonregression:

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

Logistische Regression:

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

In allen drei Fällen ist η_i der lineare Prädiktor für Beobachtung i, $i = 1, \ldots, n$:

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

Fragen? https://www.menti.com/alneo5om9sq5.