Regressionsmodelle und Parameterschätzverfahren

Jonas Nick

No Institute Given

Zusammenfassung. Die folgende Arbeit führt in Regressionsmodelle, wie lineare, nichtlineare und logistische Regression ein. Anschließend werden die zugehörigen Kostenfunktionen mit der Maximum Likelihood Methode und Bayes Inferenz bestimmt und Verfahren zu ihrer Optimierung vorgestellt. Zum Schluss wird die praktische Anwendung auf einen wirklichen Datensatz gezeigt.

1 Problembeschreibung

Charakteristisch für überwachtes maschinelles Lernen, zu der auch die Regression gehört, ist das Beschreiben der Beziehung von Zielvariable und erklärender Variable[1] aus vorliegenden Daten, also Realisierungen von Zufallsvariablen. Folgende Notation wird für die Daten genutzt:

$$x = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^n, \ \ y \in \mathbb{R} \wedge y \in G(Gruppen)$$

Hierbei bezeichnet x einen Vektor von n erklärenden Merkmalen und y ein Zielmerkmal. Bei der Regression sind also die Daten aus einem metrischen Raum, wobei das Zielmerkmal auch auf der Nominalskala sein kann, wie später bei der logistischen Regression gezeigt wird. [2] Wie sich gleich zeigen wird, ist es außerdem günstig x_0 als 1 zu definieren.

Das Regressionsmodell stellt y durch die Summe einer Hypothese von x und einem Fehlerterm ϵ dar.

$$y = h(x) + \epsilon$$

Das Ziel einer Regressionsanalyse besteht grundsätzlich darin, den Fehler ϵ (auch Residuum genannt) möglichst klein zu halten. Denn meist ist man daran

[1] werden auch abhängige und unabhängige Variable oder Prädiktor genannt

[2] Es kann sich auch um einen Vektor von Zielvariablen handeln. Das wird hier jedoch nicht weiter behandelt.

interessiert, aus gänzlich neuen erklärenden Variablen die Zielvariable vorrauszusagen. Es soll also das y vorrausgesagt werden, dass die höchste Wahrscheinlichkeit besitzt, wenn der neue Datenpunkt x und die vorherigen Erfahrungen X und Y vorliegen.

$$\operatorname*{argmax}_y p(y|x, X, Y)$$

2 Regressionsverfahren

2.1 Lineare Regression

Im linearen Regressionsmodell gibt es einen Parametervektor θ, dessen Skalarprodukt mit Merkmalsvektor x die Hypothese darstellt.

$$\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \vdots \\ \theta_n \end{bmatrix} \in \mathbb{R}^{n+1}$$

$$h_\theta(x) = \theta^T x$$

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$$

Hat ein Immobilienmakler beispielsweise eine Erhebung von Häusern gemacht, bei denen er jeweils Preis und Größe des Hauses protokollierte, so lassen sich die Datenpunkte wie in Abbildung 1 darstellen. Die Hausgröße ist hier beispielsweise erklärendes Merkmal der Preis ist das Zielmerkmal. Da es nur ein erklärendes Merkmal x_1 gibt, kann man die Hypothese als Geradengleichung auffassen, bei der Achsenabschnitt θ_0 und Steigung θ_1 gefunden worden sind.

2.2 Nichtlineare Regression

Die nichtlineare Regression ist die Linearkombination von beliebigen Funktionen ϕ von x, die sogenannten Basisfunktionen. Nichtlineare Modelle sind für Daten wie in 2, bei denen das Anlegen einer Gerade nur geringe Erfolge erzielen kann, wesentlich bessere Modelle.

$$h_\theta(x) = \sum_{i=0}^{n} \theta_i \phi_i(x)$$

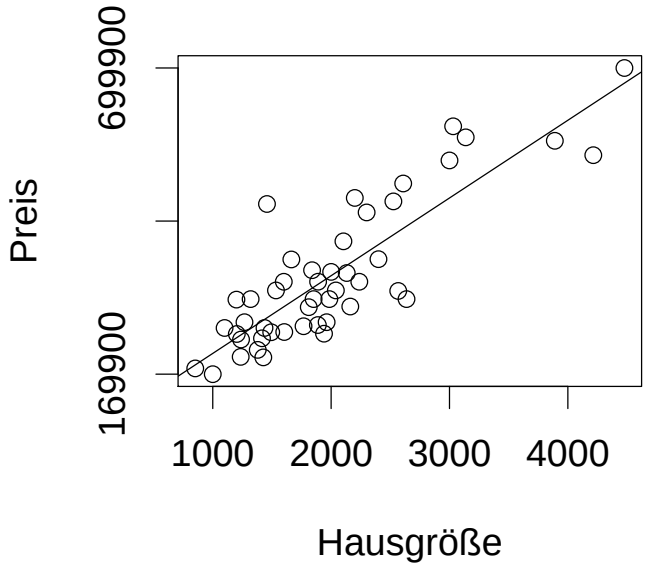

Abb. 1. Lineare Regression, Preis abhängig von Hausgröße

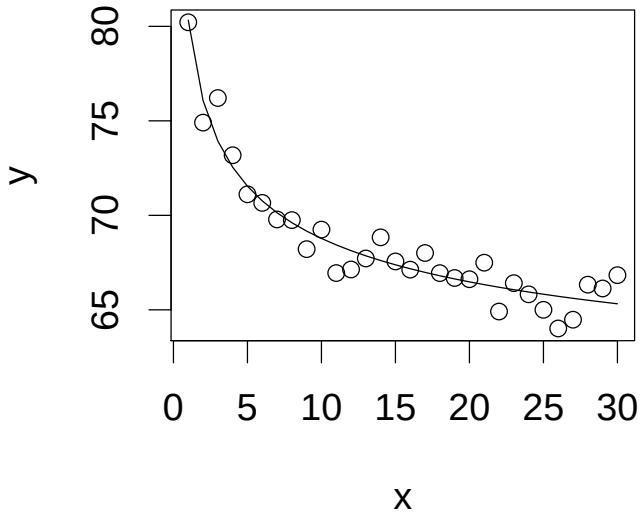

Abb. 2. Nichtlineare Regression mit Basisfunktionen: $\phi_1(x) = x_1$, $\phi_2(x) = x_1^- 0.2$

2.3 Logistische Regression

Falls es sich um ein Klassifikationsproblem handelt, das Zielmerkmal also nicht metrisch ist sondern nur abgrenzbare Kategorien annehmen kann, wendet man logistische Regression an. Im folgenden werden wir uns auf zwei Kategorien beschränken. Man nehme an, dass $y \in \{0, 1\}$ und die Kategorien jeweils 0 und 1 entsprechen. Wenn die erklärenden Variablen x vorliegen folgt die abhangige Variable y dann einer Bernoulli Verteilung, die mit der Wahrscheinlichkeit für Kategorie 1 ('Erfolg') p parameterisiert ist.

$$y \in \{0, 1\}$$
$$y|x \sim Bernoulli(p)$$

Die Hypothese ähnelt den vorangegangen, nur wird die bisherige lineare Hypothese[3] mithilfe einer Verbindungsfunktion g in die kontinuierliche Menge $[0, 1]$ abgebildet, um diese als Wahrscheinlichkeit für Erfolg zu interpretieren. Da y Bernoulli verteilt ist, ist diese Wahrscheinlichkeit auch gleich dem Erwartungswert für y.

$$h_\theta(x) = g(\theta^T x) = p = \mathbb{E}(y|x)$$

Die angesprochene Verbindungsfunktion, ist hier die namensgebende logistische Funktion (Abbildung 3).

$$g(z) = \frac{1}{1 + e^{-z}}$$

Die Hypothese hat nun genau die Eigenschaft dass ihre logarithmierten Odds[4] (logit) $\pi(x)$ äquivalent zur linearen Regressionshypothese ist.

$$\pi(x) = \ln \frac{h_\theta(x)}{1 - h_\theta(x)} = \theta^T x$$

Das heißt die Interpretation von θ_i im logistischen Model ist der geschätzte additive Effekt auf den logit für eine Veränderung des i-ten erklärenden Mermals. Schlussendlich bildet eine Schwellenfunktion $t(x)$ den kontinuierlichen Wert von h_θ auf Nominalniveau ab.

$$t(x) = \begin{cases} 1 & \text{falls } h_\theta(x) \geq 0,5 \\ 0 & \text{sonst} \end{cases}$$

Zur Klassifikation mehrer Variablen wird die sogenannte einer-gegen-alle Klassifikation angewandt, indem für jede Klasse i für die Wahrscheinlichkeit, dass $y = i$ eine Hypothese $h_\theta^{(i)}(x)$ aufgestellt wird [Bishop(2006), Ch. 2.2]

[3] Nichtlineare Hypothesen sind auch möglich, werden hier jedoch nicht im Speziellen behandelt

[4] Odds ist eine Möglichkeit Wahrscheinlichkeiten p anzugeben und folgendermaßen definiert: $\frac{p}{1-p}$

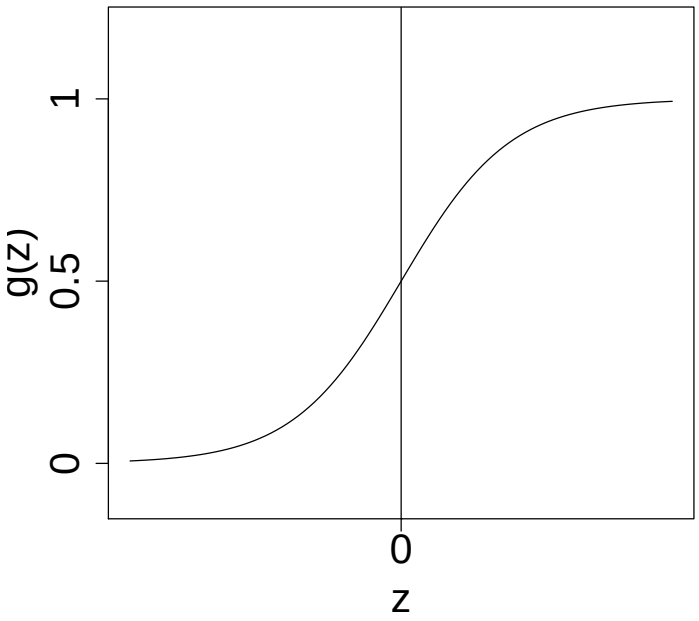

Abb. 3. Logistische Funktion

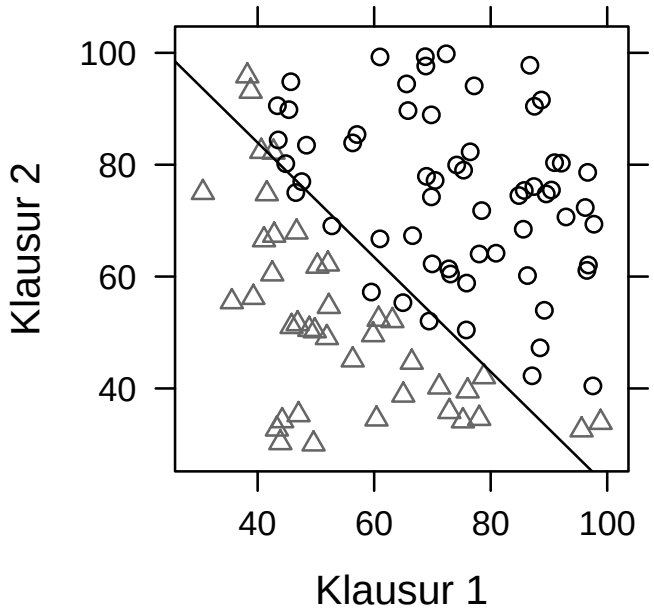

Abb. 4. Entscheidungsgrenze für Zulassung von Studenten, basierend auf Klausurnoten

Aus den Koeffizienzen lässt sich dann eine Entscheidungsgrenze berechnen. Da $\theta^T x > 0 \Leftrightarrow y = 1$ lässt sich als Entscheidungsgrenze eine n-1 dimensionale Hyperebene berechnen. Sie teilt den Raum der erklärenden Variablen in zwei Halbräume, die jeweils eine Kategorie von y repräsentieren. [5] Die Entscheidungsgrenze für nichtlineare Regression ist analog zu berechnen.

3 Parameterschätzung

Als nächstes wird gezeigt, wie man die Parameter θ findet, die die Daten am besten beschreiben, deren Fehler ϵ also möglichst gering ist. Anstelle der Beziehung eines Vektors x von erklärenden Merkmalen zu einer Zielvariablen y, liegen bei praktischen Problemen m viele Zielvariablen Y und deren zugehörige erklärende Merkmale X vor. Dabei bezeichnet X die sogenannte Designmatrix, in der die erklärenden Merkmale für ein y zeilenweise angeordnet sind.

$$X = \begin{bmatrix} -{x^{(1)}}^T- \\ -{x^{(2)}}^T- \\ \vdots \\ -{x^{(m)}}^T- \end{bmatrix} \in \mathbb{R}^{mxn},\ Y = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(m)} \end{bmatrix} \in \mathbb{R}^m$$

Nun werden die Parameter gesucht, die die beste Hypothese für alle Daten X, Y bildet. Dazu bedient man sich des Begriffs der Kostenfunktion, dessen Minimum die besten Parameter beschreibt.

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}}\, J(\theta)$$

Im Folgenden widmen wir uns der Herleitung einer Kostenfunktion.

3.1 Maximum Likelihood Methode

Zunächst allgemein zur Methode: bei der Maximum Likelihood Schätzung wird derjenige Parameter $\hat{\delta}$ ausgewählt, gemäß dessen die Realisierung der beobachteten Daten am wahrscheinlichsten ist. Es bezeichne $\mathcal{D} = (d^{(1)}, d^{(2)} \dots d^{(m)})$ Realisierungen von Zufallsvariablen mit zugehöriger Wahrscheinlichkeitsdichte $p(\mathcal{D}|\delta)$. Dann definiert man die Likelihood von δ folgendermaßen:

$$L(\delta) := p(\mathcal{D}|\delta)$$

[5] Beispiel mit zwei erklärenden Variablen x_1 und x_2 (Abbildung 4):

$$\begin{aligned} \theta^T x &= 0 \\ \theta_0 + \theta_1 x_1 + \theta_2 x_2 &= 0 \\ f(x_1) = x_2 &= -\frac{\theta_0}{\theta_2} - \frac{\theta_1}{\theta_2} x_1 \end{aligned}$$

Wähle zu den Beobachtungen $\mathcal{D}$ als Parameterschätzung denjenigen Parameter $\hat{\delta}$, für den die Likelihood maximal ist, d.h.

$$\hat{\delta} = \underset{\delta}{\operatorname{argmax}}\, L(\delta)$$

Die Wahrscheinlichkeitsdichte der Daten muss also bekannt sein, bzw. vorausgesetzt werden um den Maximum Likelihood Parameter $\hat{\delta}$ zu bestimmen. Will man beispielsweise Erwartungswert μ sowie Standardabweichung σ aus normalverteilten Daten schätzen, so lässt sich die Wahrscheinlichkeitsdichte folgendermaßen darstellen:

$$L(\mu, \sigma) = p(d|\mu, \sigma) = \mathcal{N}(\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(d-\mu)^2}{2\sigma^2})$$

Sucht man nun das Maximum dieser Funktion [Fahrmeir et al.(2011)Fahrmeir, Künstler, Pigeot, and Tutz, Kapitel 9.3.1] jeweils für μ und σ, so landet man bei den bekannten Formeln für Mittelwert und Standardabweichung. [6]

Die Anwendung der Maximum Likelihood auf das Regressionsproblem erfordert folgende Annahme über die Verteilung der Zielvariable y gegeben der erklärenden Merkmale und dem Parametervektor θ. Zielvariable y ist normalverteilt, wobei der Erwartungswert der Hypothese entspricht. Man findet nun die Hypothese, gemäß deren die Realisierung des Zielmerkmals möglichst wahrscheinlich ist. (Abbildung 5).

$$L(\theta) = p(y|x, \theta) = \mathcal{N}(h_\theta(x), \sigma_1)$$

Aus der Annahme folgt außerdem, dass die Residuen normalverteilt sind. Die Standardabweichung σ_1 lässt sich berechnen, ist jedoch hier nicht weiter von Interesse.
Da die einzelnen Datenpunkte als unabhängig in den Daten betrachtet werden, gilt für die gesamten Daten:

$$p(Y|X, \theta) = \prod_{i=1}^{m} \mathcal{N}(h_\theta(x^{(i)}), \sigma_1)$$

Es ist üblich anstatt der Likelihood die log-Likelihood zu berechnen, da diese leichter abzuleiten ist, die Extremwerte aber nicht verschoben sind. Logarithmieren und Einsetzen der Gleichung für die Normalverteilung ergibt:

$$\ln L(\theta) = p(Y|X, \theta) = -\frac{\sigma_1}{2} \sum_{n=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2 + \frac{m}{2} \ln \sigma_1 - \frac{m}{2} \ln(2\pi)$$

[6]

$$\mu = \bar{x} = \frac{1}{n} \sum_{i=1}^{n}$$

$$\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})}$$

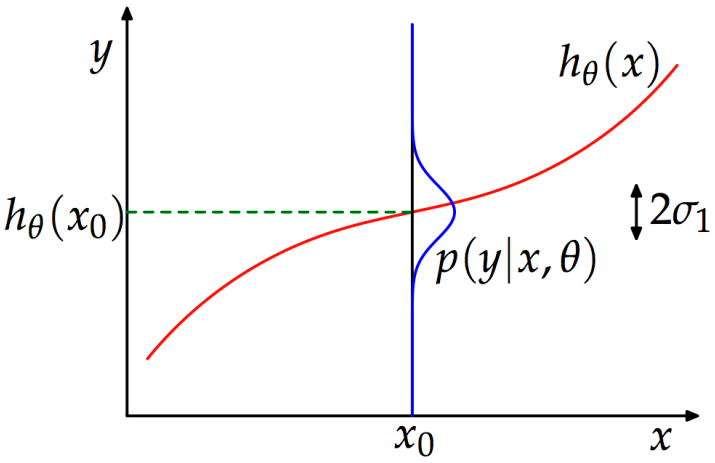

Abb. 5. Annahme für die Maximum Likelihood Schätzung bei der Regression. Grafik verändert aus [Bishop(2006), Figure 1.16]

Die letzen beiden Summanden der Gleichung hängen nicht von θ ab, daher fallen sie weg wenn man das Maximum bezüglich θ sucht. Multiplikation mit einem positiven Faktor verschiebt die Position des Maximums nicht, daher kann man die Multiplikation mit σ_1 vernachlässigen. Desweiteren kann man anstatt log-Likelihood zu maximieren auch die negative log-Likelihood minimieren. Dann erhält man folgende Kostenfunktion:

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2$$

Im Grunde beschreibt diese Kostenfunktion Abweichung von realem Wert und vorhergesagtem Wert.
Gesucht ist nun das Minimum der Kostenfunktion, also der minimale Mittelwert der Fehlerquadrate. Daher bildet man die partielle Ableitung von $J(\theta)$ für jeden Parameter θ_i und sucht dessen Nullpunkt.

$$\frac{\partial}{\partial \theta_j} J(\theta) = \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})\, x_j^{(i)} \stackrel{!}{=} 0$$

Durch Einsetzen der linearen Hypothese und Umformen in Matrixschreibweise erhält man ein lineares System von Normalgleichungen, mit dem sich der Parametervektor direkt berechnen lässt [Bishop(2006), Ch. 3.1.1].

$$\begin{aligned} \Rightarrow \theta &= X^+ y \\ &= (X^T X)^{-1} X^T y \end{aligned}$$

X^+ bezeichnet hier die *Moore-Penrose Pseudoinverse* von X, die das Konzept von Invertierbarkeit auf nichtquadratische Matrizen erweitert und allgemein zum Berechnen von optimalen Lösungen mit kleinster euklidischer Norm bei linearen Ausgleichsproblemen verwendet wird. Es kann allerdings vorkommen, dass $X^T X$ nicht invertierbar ist [7]- häufig aufgrund redundanter erklärende Merkmale, also solche die nicht linear unabhängig von anderen sind, oder wenn es mehr Merkmale als Datenpunkte gibt ($m < n$).

3.2 Gradientenverfahren

Da die Matrix $X^T X$ eine nxn Matrix ist und die Invertierung einer Matrix eine asymptotisch untere Schranke von $\Omega(n^2)$ (Strassen's Algorithmus $O(n^{2.81})$, siehe [Cormen et al.(2009)Cormen, Leiserson, Rivest, and Stein, S. 827]) hat, wird die analytischen Lösung auf Probleme mit vielen erklärenden Merkmalen nicht angewandt, sondern das meist wesentlich schnellere Verfahren des Gradientenabstiegs. Es findet das Minimum der Kostenfunktion dadurch, dass iterativ so lange in Richtung des Gefälles der Funktion abgestiegen wird, bis der Werteunterschied so gering ist, dass man sich sicher in der Nähe des Minimums befindet:

repeat until convergence {
$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta)$
}

Die Parameter θ_j sind anfangs zufällig initialisiert. Wichtig ist die Wahl der Lernrate α, die bestimmt wie weit in der Funktion gesprungen wird. Falls nämlich α zu groß gewählt wurde, so kann es sein, dass der Algorithmus systematisch das Minimum überspringt und der Algorithmus daher nicht konvergiert. Falls das Gradientenverfahren ein Minimum findet, so ist es global, da die maximum likelihood Kostenfunktion für lineare Regression konvex ist. [Bishop(2006), Ch. 1.6.1]

Es gibt elaboriertere Optimierungsverfahren, wie konjugierte Gradienten oder das BFGS Verfahren, die keine explizite Lernrate benötigen und unter Umständen schneller konvergieren.

3.3 Logistische Regression

Bei der logistischen Regression wendet man nicht die kleinste Quadrate Kostenfunktion an, da diese multiple lokale Minima haben könnte. Stattdessen ist folgende Kostenfunktion üblich:

[7] Beweis:
Spaltenvektoren von X abhängig $\Rightarrow \det(X) = 0 \Rightarrow \det(X^T)\det(X) = 0 \Rightarrow \det(X^T X) = 0 \Rightarrow X^T X$ nicht invertierbar

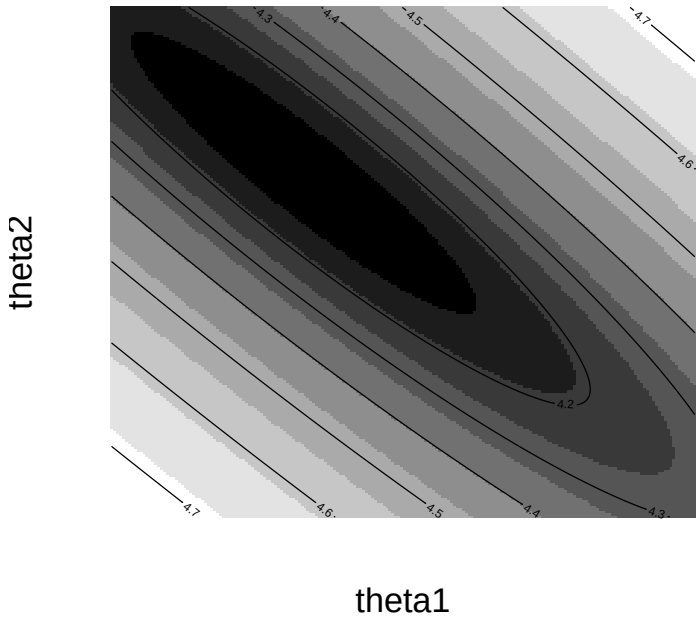

Abb. 6. Kostenfunktion der Hauspreis Regression (Abbildung 1)

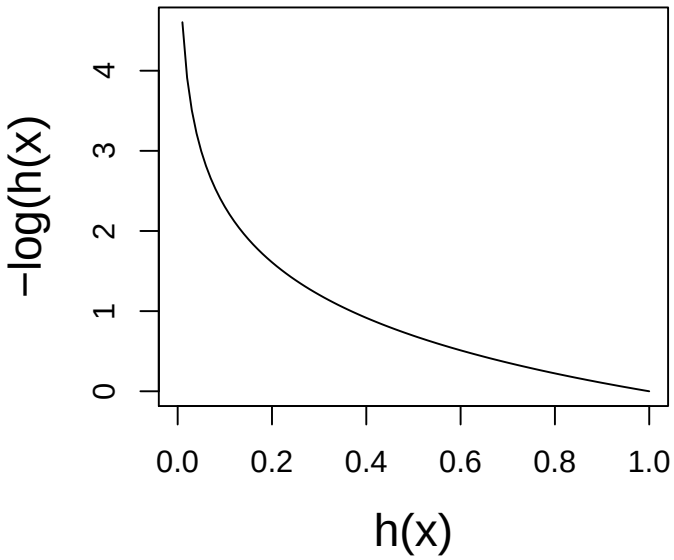

Abb. 7. Logistische Kostenfunktion für $y = 1$

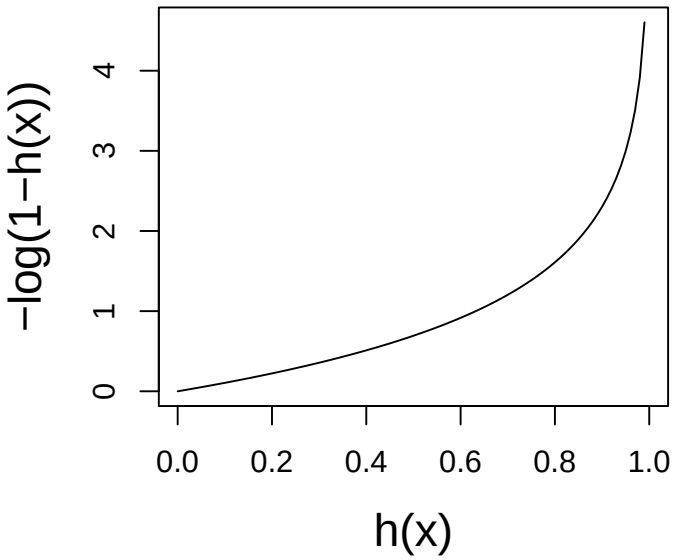

Abb. 8. Logistische Kostenfunktion für $y = 0$

$$J(\theta) = \frac{1}{m} \sum_{i=1}^{m} Cost(h_\theta(x^{(i)}), y^{(i)})$$

$$Cost(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & if\ y = 1 \\ -\log(1 - h_\theta(x)) & if\ y = 0 \end{cases}$$

Diese Kostenfunktion entspricht der intuitiven Vorstellung, dass Hypothesen, die in der Nähe der wahren Werte $y = 0$ oder $y = 1$ liegen geringe Kosten haben, im Gegensatz zu Hypothesen, die einen größeren Abstand zum wahren Wert haben (siehe Abbildung 7 und 8). Man kann keine geschlossene Form zur Minimumsuche angeben. Da diese Kostenfunktion aber konvex ist, lässt sich das Gradientenverfahren komfortabel anwenden. Folgende Formel ist äquivalent zur vorangegangen, lässt sich allerdings leichter ableiten.

$$J(\theta) = -\frac{1}{m} \sum_{i=1}^{m} y^{(i)} \log h_\theta(x^{(i)}) + (1 - y^{(i)}) \log(1 - h_\theta(x^{(i)}))$$

4 Bayes'sche Lineare Regression

Die Bayes'sche Lineare Regression ist eine Erweiterung der Linearen Regression, bei der anstelle der Maximum Likelihood die Bayes'sche Parameterschätzung eingesetzt wird. Dies führt dazu, dass die Regression weniger anfällig für Überanpassung wird.
Es wird nun zunächst allgemein das Konzept der Bayes'schen Schätzung eingeführt, um diese dann auf die lineare Regression anzuwenden.

4.1 Bayes Schätzer

Bei der Maximum Likelihood Methode wird der Parameter so gewählt, dass die vorliegende Beobachtung der Daten am wahrscheinlichsten ist. Tatsächlich wäre es aber umgekehrt wesentlich intuitiver, den Parameter zu wählen, der unter den vorliegenden Daten am wahrscheinlichsten ist (a posteriori Wahrscheinlichkeit). Dieses Vorgehen ist mit Bayes Inferenz möglich. Wie gehabt sei $p(D|\delta)$ die Wahrscheinlichkeitsdichte der Daten $\mathcal{D}$, gegeben Parameter δ, und $L(\delta) = p(\mathcal{D}|\delta)$ die Likelihoodfunktion.

Die Erweiterung ist, dass für das unbekannte δ eine a priori Dichte

$$p(\delta)$$

spezifiziert wird. Nun kann die a posteriori Dichte des Parameters über den Satz von Bayes bestimmt werden durch

$$p(\delta|\mathcal{D}) = \frac{p(d|\delta)p(\delta)}{p(\mathcal{D})} = \frac{L(\delta)p(\delta)}{\int L(\delta)p(\delta)d\delta}$$

Hierbei ist der Nenner nicht von besonderem Interesse, da er nicht von δ abhängt. Der Maximum a posteriori (MAP) Schätzer ist derjenigen Parameterwert $\hat{\delta}_{MAP}$, für den die a posteriori Dichte maximal wird, d.h.:

$$\hat{\delta}_{MAP} = \underset{\delta}{\operatorname{argmax}}\, L(\delta)p(\delta)$$

4.2 Bayes'sche Lineare Regression

Der für die Regression benötigte Parametervektor θ wird nun mithilfe der Bayes Inferenz geschätzt, wozu eine a priori Verteilung angenommen wird:

$$p(\theta) = \mathcal{N}(0, \sigma_0)$$

Diese Annahme drückt unter anderem aus, Parameter nahe 0 wahrscheinlicher sind und daher näher beieinander liegen. Bei der Maximum Likelihood Methode wird dagegen impliziert, dass der Parameter einer Gleichverteilung folgt. Für die Likelihood von θ wird die selbe Verteilungsannahme getroffen wie bei der Maximum Likelihood Schätzung für Lineare Regression.

$$L(\theta) = p(y|x, \theta) = \mathcal{N}(h_\theta(x), \sigma_1)$$

Gemäß dem Bayes Theorem ist die a posteriori Wahrscheinlichkeitsdichte proportional zum Produkt aus Likelihood und a priori Dichte.

$$p(\theta|x, y) \propto p(y|x, \theta)p(\theta)$$

Anwendung des negativen ln und Einsetzen der Normalverteilungsfunktion ergibt folgende Kostenfunktion, deren Minimum beim MAP $\hat{\theta}$ zu finden ist.

$$J(\theta) = \frac{\sigma_1}{2} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2 + \frac{\sigma_0}{2} \sum_{j=1}^{n} \theta_j^2$$

Diese Kostenfunktion ähnelt sehr der Kleinsten Quadrate Kostenfunktion, außer dass durch den zweiten Summanden ganz allgemein größere Parameter θ höhere Kosten verursachen. Dieser wird Regularisierungsterm genannt und er hilft zu vermeiden, dass das Modell überangepasst ist (Overfitting, Abbildung 9), also die Trainingsdaten zwar sehr gut approximiert, aber für neue Daten schlechte Voraussagen gemacht werden, da der Einfluss einzelner erklärender Merkmale zu groß ist. Es gibt eine handvoll von Regularisierungsverfahren, aber eine gängige Erweiterung wird *Tikhonov Regularisierung* genannt. Angewandt auf die kleinste Quadrate Kostenfunktion bedeutet folgendes:

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2 + \lambda \sum_{j=1}^{n} \theta_j^2$$

Der Koeffizient λ beschreibt die relative Wichtigkeit des Regularisierungsterms im Vergleich zu den Fehlerquadraten und wird beispielsweise mit Kreuzvalidierung gefunden. Für diese Kostenfunktionen lässt sich das Minimum analog mit Normalgleichung oder Gradientenabstieg berechnen [Bishop(2006), Ch. 1.1].

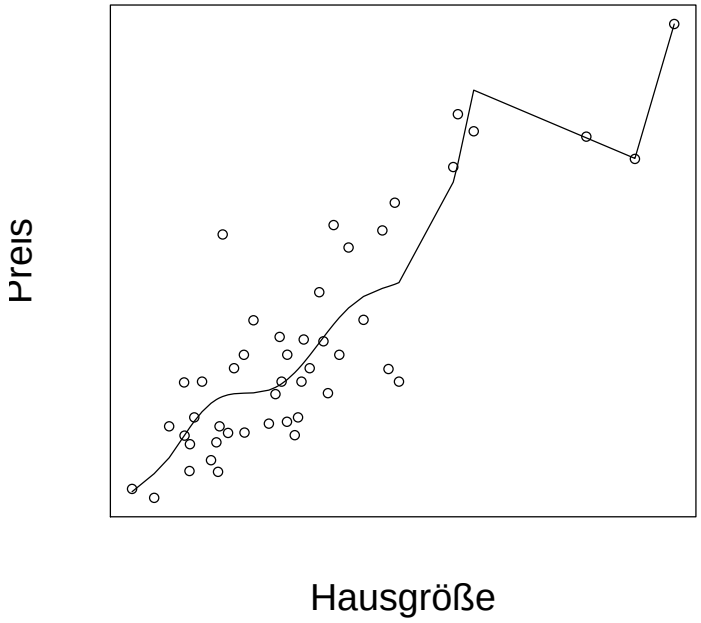

Abb. 9. Polynomielle Regression 10. Grades des Hauspreises ohne Regularisierung

Vorraussage eines neuen y, basierend auf neuen erklärenden Merkmalen x und vorherigen Daten X und Y war ein anfänglich gestelltes Problem. Mit Bayes'schen Mitteln lässt sich dazu die a posteriori Dichte des Zielmerkmals ermitteln. Diese wird durch Herausmarginalisieren von θ aus den bekannten Wahrscheinlichkeitsdichten berechnet:

$$p(y|x, X, Y) = \int p(y|x, \theta)p(\theta|X, Y)d\theta$$

Aus dieser Dichte lässt sich nun dasjenige y mit der höchsten Wahrscheinlichkeit vorraussagen.

5 Zusammenfassung

- Ein Regressionsmodell verknüpft Zielvariable y mit einer Funktion von erklärendem Merkmal x.
- Lineare Regression ist die Linearkombination von nach θ gewichteten erklärenden Merkmalen.
- Kleinste Quadrate Kostenfunktion ergibt sich aus der Maximum Likelihood Parameterschätzung.
- Logistische Regression klassifiziert mittels Verbindungsfunktion.
- Bayes Regression nimmt eine a priori Verteilung der Koeffizienten θ an, wodurch sich Überanpassung teilweise vermeiden lässt.

6 Anwendung: Titanic-Datensatz

Ein häufig angewandter Demonstrationsdatensatz für Klassifikationen ist eine Studie über das Sinken der Titanic [British Board of Trade(1990)] , in dem für 1309 Passagiere jeweils angegeben ist, ob er überlebt hat, sowie zusätzliche Informationen über den Passagier, wie ökonomischer Status (Beförderungsklasse, `pclass`), Geschlecht (`sex`), Alter (`age`), Anzahl der Geschwister an Bord (`sibsp`) und Anzahl der Eltern an Bord (`parch`).

	survived	pclass	sex	age	sibsp	parch
1	0	3	male	22	1	0
2	1	1	female	38	1	0
3	1	3	female	26	0	0
4	1	1	female	35	1	0
5	0	3	male	35	0	0
6	0	3	male	NA	0	0

Die Untersuchung des Datensatzes ist historisch motiviert und kann Aufschluss über soziale Struktur der damaligen Gesellschaft und den ungefähren Ablauf der Katastrophe geben. Anhand der Abbildungen 10, 11 und 12 kann man erkennen, dass Passagiere mit hohem ökonomischer Status , Frauen und Kinder eher überlebt haben. Mit diesen drei erklärenden Variablen werden nun im Folgenden mithilfe der Open Source Statistik Software R [R Development Core Team(2012)] eine logistische Regression durchgeführt.

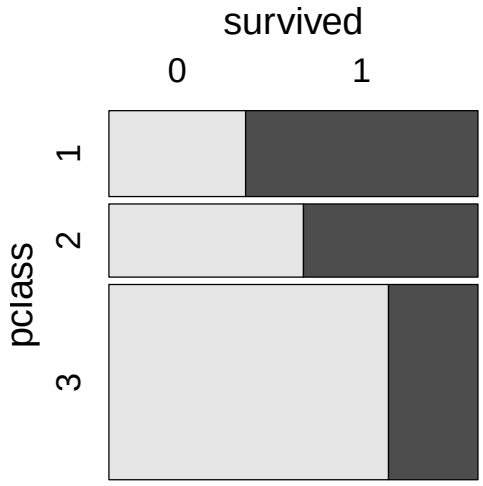

Abb. 10. Mosaikplot Überleben gegeben Beförderungsklasse

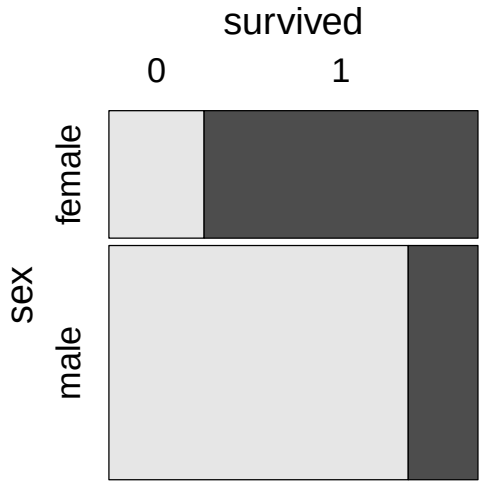

Abb. 11. Mosaikplot Überleben gegeben Geschlecht

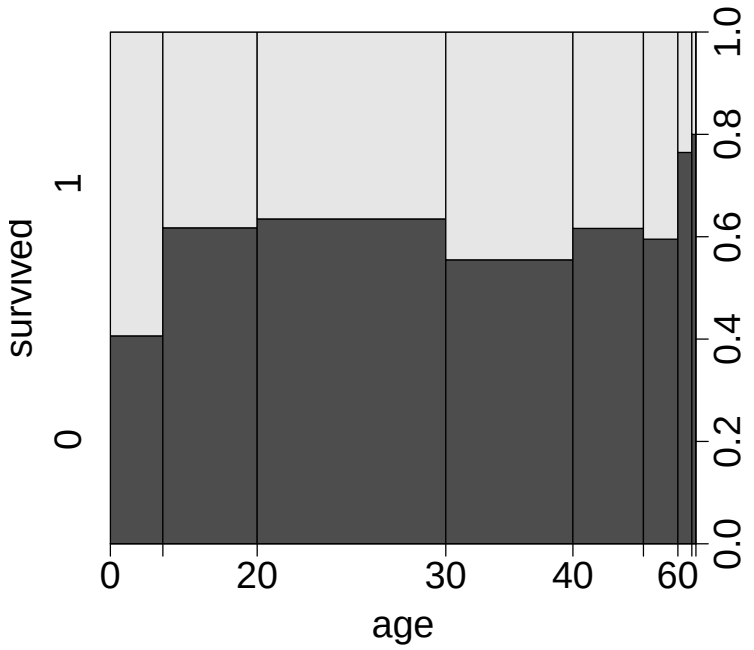

Abb. 12. Mosaikplot Überleben gegeben Alter

```
>model <- glm(survived ~ age + sex + I(pclass==1)
                + I(pclass==2),data=titanicTrainData,
                family=binomial("logit"))
```

Der Ausdruck liest sich folgendermaßen: In der Variable model wird das logistische Regressionsmodell gespeichert, dass mit der Funktion `glm` erzeugt wird. Der erste Parameter spezifiziert folgende `formula`:
Die Zielvariable survived soll anhand einer linearen Kombination der erklärenden Merkmale `age`, `sex` und `pclass` dargestellt werden. Da `sex` und `pclass` Faktorvariablen sind, also nicht kontinuierlich, wird im Modell für jede Faktorstufe eine zusätzliche erklärende Variable eingeführt die 0 oder 1 sein kann, wobei 1 bedeutet, dass die Variable die Faktorstufe annimmt. Dann können Koeffizienten für jede Faktorstufe ermittelt werden, wobei eine Faktorstufe als Standard angenommen wird. Zum Beispiel geht das Merkmal Geschlecht so in das Modell ein, dass nur im Fall 'männlich' der zugehörige θ-Wert addiert wird und der Fall 'weiblich' als Standardfall angenommen wird.
Der zweite Parameter gibt den Datensatz an und der dritte bestimmt, dass das glm (generalized linear model) die logit-Funktion als Verbindungsfunktion nutzen soll, damit die logistische Regression durchgeführt wird.
Informationen über geschätzte Koeffizienten und Gütekriterien lassen sich mit dem Befehl summary() aufrufen.

```
>summary(model)
Call:
glm(formula = survived ~ age + sex + I(pclass == 1) + I(pclass ==
    2), family = binomial("logit"), data = titanicTrainData)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7303  -0.6780  -0.3953   0.6485   2.4657

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)         1.196387   0.252649   4.735 2.19e-06 ***
age                -0.036985   0.007656  -4.831 1.36e-06 ***
sexmale            -2.522781   0.207391 -12.164  < 2e-16 ***
I(pclass == 1)TRUE  2.580625   0.281442   9.169  < 2e-16 ***
I(pclass == 2)TRUE  1.270826   0.244048   5.207 1.92e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 964.52  on 713  degrees of freedom
Residual deviance: 647.28  on 709  degrees of freedom
  (177 observations deleted due to missingness)
AIC: 657.28

Number of Fisher Scoring iterations: 5
```

Hier sind zunächst die Quantile der Residuen oder Fehler ϵ aufgetragen. Diese sollten einen Median Nahe 0 haben und bei linearer Regression annähernd normalverteilt sein. Dann sieht man zeilenweise Achsenabschnitt und die erklärenden Merkmale, wobei die erste Spalte (unter `Estimate`) genau der ermittelte θ Vektor ist. Diese sind allerdings im Gegensatz bei der logistischen Regression nicht sehr einfach zu interpretieren. Betrachtet man zum Beispiel Geschlecht, so wird im Falle des Mannes etwa 2.5 vom linearen Modell subtrahiert, hat daher nach der logistischen Transformation eine geringere Wahrscheinlichkeit für 'Erfolg' und spiegelt wider, dass Männer eher seltener überlebt haben. Genauer: die Koeffizienten sind logarithmische Quotenverhältnisse (Odds ratio) und geben an, wie sich die log-Odds für 'Erfolg' mit jeder Veränderung der erklärenden Variable ändert. [Hosmer and Lemeshow(2000), S.47] Das heisst im Falle männlich ändert sich die Odds der Wahrscheinlichkeit zu überleben um den Faktor $e^{-2.522781} \approx 0.08$
Die folgenden Spalten geben Standardfehler, z-Werte und die Wahrscheinlichkeit mit der sich das geschätzte θ für die Variable von 0 unterscheidet. `Null deviance` repräsentiert die Abweichung des Modells von den realen Daten ohne unabhängige Variablen (Null Modell) und `Residual deviance` hängt von der Summe der Residuale des gesamten Modells ab [Baayen(2008), S. 217]. `AIC` ist das Akaike Information Criterion und stellt ein wichtiges Kriterium zur Auswahl des Modells dar, da sowohl Anpassungsgüte, als auch Komplexität des Modells (siehe Overfitting) in die Beurteilung mit einfließt. Es sollte möglichst klein sein, da es den Verlust von Information darstellt.

```
>cat(testModel())
Trainingsdaten: 78.8515406162465% korrekte Vorhersagen
Testdaten: 76.07% korrekte Vorhersagen
```

Diese Präzision auf dem Testdatensatz lässt darauf schließen, dass die drei bisherigen Variablen einen Großteil der Daten erklären.
Zur Verbesserung des Modells werden wir den Datensatz noch etwas genauer anschauen. In Abbildung 12 erkennt man, dass die Beziehung zwischen Alter und Überleben nicht linear ist, sondern eher logarithmisch oder polynomiell. Abbildung 13 zeigt, dass Personen mit Geschwistern an Bord eher überlebt haben. Desweiteren sieht man in Abbildung 14, dass eine Interaktion zwischen Geschlecht und Beförderungsklasse vorliegt, beispielsweise haben Frauen in der zweiten Klasse im Gegensatz zu Männern ähnlich häufig überlebt wie in der ersten Klasse. Diese Erkenntnisse sollen nun in das Modell einfließen.

```
Call:
glm(formula = survived ~ I(log(age)) + pclass:sex + sibsp, family = binomial("logit"),
    data = titanicTrainData)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.5090  -0.6339  -0.4179   0.3846   2.3867

Coefficients: (1 not defined because of singularities)
```

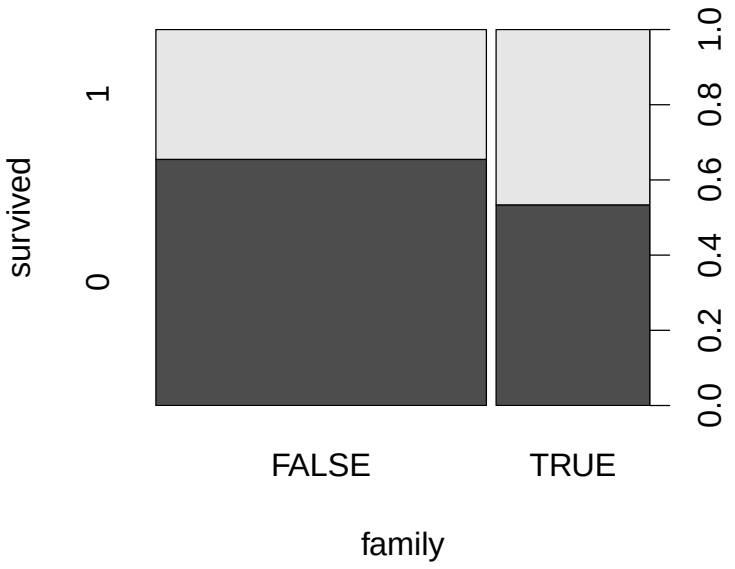

Abb. 13. Mosaikplot: Überleben gegeben Geschwister an Bord

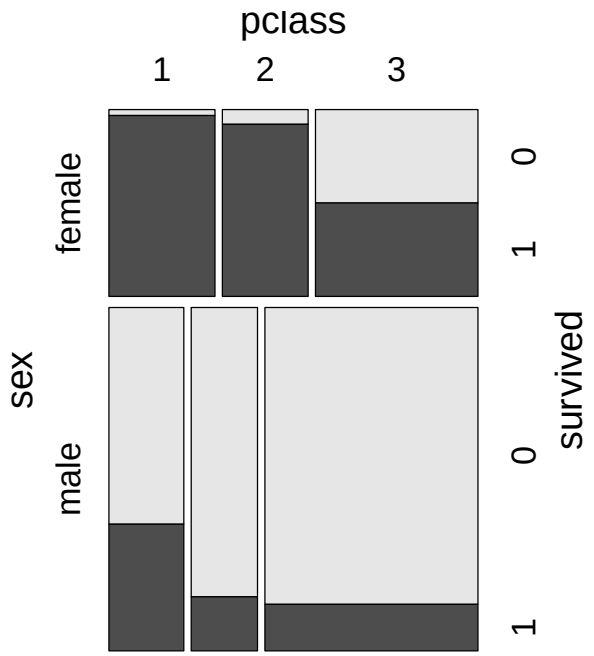

Abb. 14. Mosaikplot: Überleben gegeben Geschlecht und Beförderungsklasse

```
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)         1.7400     0.5584   3.116  0.00183 **
I(log(age))        -1.0651     0.1663  -6.406 1.50e-10 ***
sibsp              -0.5358     0.1344  -3.987 6.68e-05 ***
pclass1:sexfemale   5.6885     0.6328   8.989  < 2e-16 ***
pclass2:sexfemale   4.5471     0.4776   9.520  < 2e-16 ***
pclass3:sexfemale   1.5393     0.2845   5.412 6.25e-08 ***
pclass1:sexmale     1.8860     0.2967   6.356 2.07e-10 ***
pclass2:sexmale    -0.1326     0.3669  -0.362  0.71768
pclass3:sexmale         NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 964.52  on 713  degrees of freedom
Residual deviance: 586.38  on 706  degrees of freedom
  (177 observations deleted due to missingness)
AIC: 602.38

Number of Fisher Scoring iterations: 6
```

Der Interaktionsterm wird durch einen Doppelpunkt spezifiziert und bewirkt, dass für jede Kombination von Geschlecht und Beförderungsklasse ein Koeffizient geschätzt wird der auf die log-Odds addiert wird, wenn die Kombination zutrifft. Fast alle unabhängigen Variablen sind signifikant[8], die Residual deviance hat sich stark verkleinert, ebenso wie der AIC. Die Funktion drop1 berechnet Residual deviance und AIC wenn jeweils eine Variable fehlt.

```
>drop1(model)

Single term deletions

Model:
survived ~ I(log(age)) + pclass:sex + sibsp
           Df Deviance    AIC
<none>          586.38 602.38
I(log(age))  1   640.26 654.26
sibsp        1   605.29 619.29
pclass:sex   5   949.05 955.05
```

Man erkennt, dass das Modell ohne die Interaktion von Geschlecht und Beförderungsklasse am schlechtesten abschneidet, aber auch die beiden anderen erklärenden Variablen senken die Kriterien, wenn sie nicht fehlen.

```
>cat(testModel())

Trainingsdaten: 81.5126050420168% korrekte Vorhersagen

Testdaten: 74.6% korrekte Vorhersagen
```

[8] `NA` bedeutet not available, warum die Koeffizienten nicht ermittelt werden können ist mir noch nicht bekannt.

Zusammenfassend lässt sich sagen, dass das zweite Modell überangepasst ist, da es zwar auf dem Trainingsdatensatz bessere Ergebnisse erzielt, aber auf dem Testdatensatz schlechter als das einfachere Modell abschneidet. Allerdings ist auch das Modell nicht zufriedenstellend. Denn eine sehr naive Klassifikation, bei der Frauen immer überleben und Männer nie überleben hat eine höhere Präzision, nämlich $76,5\%$. [9] Random forests erzielen eine Genauigkeit von 77.5% Eine Erklärung für das schlechte Abschneiden der logistischen Regression in diesem Vergleich könnte darin liegen, dass auch schon das erste, einfachere Modell überangepasst und deshalb das Naive Modell besser war. Auch scheint Regression generell nicht sehr gut mit faktoriellen erklärenden Variablen zurecht zu kommen, da lediglich der Achsenabschnitt verändert wird, komplexere Interaktionen aber nicht berücksichtigt werden, die beispielsweise Entscheidungsbäume im Allgemeinen besser modellieren.

Literatur

[Baayen(2008)] H. Baayen. *Analyzing linguistic data: A practical introduction to statistics using R.* Cambridge University Press, 2008.

[Bishop(2006)] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics).* Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006. ISBN 0387310738.

[British Board of Trade(1990)] British Board of Trade. Report on the loss of the titanic (reprint), 1990.

[Cormen et al.(2009)Cormen, Leiserson, Rivest, and Stein] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to Algorithms (3. ed.).* MIT Press, 2009. ISBN 978-0-262-03384-8.

[Fahrmeir et al.(2011)Fahrmeir, Künstler, Pigeot, and Tutz] L. Fahrmeir, R. Künstler, I. Pigeot, and G. Tutz. *Statistik.* Springer Verlag, siebte edition, 2011.

[Hosmer and Lemeshow(2000)] D. W. Hosmer and S. Lemeshow. *Applied logistic regression (Wiley Series in probability and statistics).* Wiley-Interscience Publication, 2 edition, 2000. ISBN 0471356328.

[Ng(2012)] A. Ng. Machine learning. Coursera, 2012. URL https://class.coursera.org/ml-2012-002/wiki/view?page=CourseInformation.

[R Development Core Team(2012)] R Development Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2012. URL http://www.R-project.org/. ISBN 3-900051-07-0.

[9] https://www.kaggle.com/c/titanic-gettingStarted/leaderboard