Katrin Mauthner, BSc

Nichtparametrische Regressionsmodelle für ordinale Zielgrößen

MASTERARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieurin

Masterstudium Operations Research und Statistik

eingereicht an der

Technischen Universität Graz

Betreuer:
Ao.Univ.-Prof. Dipl.-Ing. Dr.techn. Herwig Friedl

Institut für Statistik

Graz, März 2015

EIDESSTATTLICHE ERKLÄRUNG

AFFIDAVIT

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt, und die den benutzten Quellen wörtlich und inhaltlich entnommenen Stellen als solche kenntlich gemacht habe. Das in TUGRAZonline hochgeladene Textdokument ist mit der vorliegenden Masterarbeit identisch.

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly indicated all material which has been quoted either literally or by content from the sources used. The text document uploaded to TUGRAZonline is identical to the present master's thesis.

Datum/Date	Unterschrift/Signature

ZUSAMMENFASSUNG

In dieser Arbeit werden zwei wichtige Konzepte der statistischen Modellierung mit dem Ziel vorgestellt, die beiden Ideen zu kombinieren und auf einen Datensatz basierend auf dem österreichischen Silikonregister anzuwenden. Die *kategorielle Datenanalyse* befasst sich mit Regressionsmodellen für Response-Variablen mit mehreren Stufen. Mit *nichtparametrischen Regressionsmodellen* können stetige erklärende Variablen flexibel in ein Modell eingebunden werden, indem die Variable als glatte Funktion in den Prädiktor aufgenommen wird. In Bezug auf das Silikonregister sind wir am Einfluss der Zeit auf das Auftreten einer sogenannten Kapselkontraktur interessiert. Gesucht sind Regressionsmodelle für ordinale Responses, wobei die Zeit nichtparametrisch im Prädiktor inkludiert ist.

ABSTRACT

This thesis introduces two important concepts of statistical modelling with the aim of combining both ideas and applying them to data based on the Austrian Breast Implant Register. Categorical Data Analysis *deals with regression models for response variables with several categories.* Nonparametric Regression Models *allow for continuous covariates to be included in a flexible way by incorporating a smooth function of the variable into the predictor. Concerning the Breast Implant Register we are interested in the influence of time on the occurrence of a so-called capsular fibrosis. We search for regression models for ordinal responses where the time is included nonparametrically in the predictor.*

Inhaltsverzeichnis

1. Einleitung

1.1. Das Silikonregister

Schon im Jahre 1996 hat die Österreichische Gesellschaft für plastische, ästhetische und rekonstruktive Chirurgie entschieden ein Silikonregister für Brustimplantate umzusetzen. Das Ziel war die Qualitätssicherung- und verbesserung für Patientinnen und Ärzte. Mit diesem Gedanken wird das Register bis heute weitergeführt. Alle Fachärzte, die Implantate einsetzen, sind aufgefordert sich am Register zu beteiligen. Zwischen 2004 und 2012 wurden 13116 Implantate registriert, vgl. Wurzer et al. (2014).

Eine Operation ist immer mit gewissen Risiken verbunden. Der Körper bildet z. B. als Immunreaktion eine Bindegewebshülle (Kapsel) um das Implantat. In manchen Fällen kann sich diese Kapsel zusammen ziehen und verhärten, wobei es oft auch zu Verschiebungen des Implantats kommt. Man spricht von einer Kapselkontraktur (Kapselfibrose). Eine Einteilung nach dem Schweregrad der Kontraktur erfolgt nach Baker in vier Stufen, den sogenannten Baker-Stufen.

Wir sind in dieser Arbeit daran interessiert, wie sich gewisse Faktoren auf die Entwicklung einer Kapselkontraktur auswirken. Der dazu herangezogene Datensatz stammt aus dem Silikonregister und enthält aktuell 17327 Einträge. In dieser Arbeit sind wir aber nur an Patientinnen mit der Operationsart `Impaltatwechsel` und der Implantatart `Mammaimplantat` interessiert, wodurch sich diese Anzahl verringert. Die Daten liegen im Daten-Frame `mydata` vor, welcher 3534 Zeilen hat. Für genaue Informationen wie dieser generiert wurde siehe Anhang A.1. Im Datensatz gibt es die Variable `kontr` mit den Kategorien {`BakerI`, `BakerII`, `BakerIII`, `BakerIV`}, für deren Auftrittswahrscheinlichkeiten wir ein möglichst gutes Regressionsmodell finden wollen. Folgende Größen aus dem Datensatz können als mögliche erklärende Variablen betrachtet werden:

`dau`: Dauer zwischen der Erst- und Revisionsoperation in Jahren bzw. Zeit zwischen der Revisionsoperation und der letzten Operation vor dieser Revisionsoperation. Es gibt Werte zwischen 0 und ungefähr 40 Jahren.

`vol`: Füllvolumen des Implantats in cm^3, Werte zwischen 80 und 1200 cm^3.

`oberfl`: Oberflächenbeschaffenheit des Implantats, entweder `glatt`, `texturiert` oder `Polyurethan`-beschichtet.

`lumen`: Anzahl der verschiedenen Typen von Silikongelen im Implantat. Entweder `single`, `double` oder `triple`. Falls nicht `single`, sind die Silikongele räumlich getrennt.

`fuel`: Füllmaterial des Implantats, z. B. `Silikongel` oder `Kochsalzlösung`.

`lage`: Lage des Implantats, z. B. `submuskulär` oder `subglandulär`.

`opzug`: Einschnittstelle bei der Operation, z. B. `inframammär` oder durch eine bestehende Narbe.

`antib`: Verabreichung von Antibiotika während der Operation, z. B. `systemisch` oder in der `Implantathöhle`.

`ster`: Verabreichung von Entzündungshemmern vor der Operation, entweder `systemisch`, in der `Implantathöhle` oder `keine`.

`drain`: Während der Operation wurde eine Drainage gesetzt (`j`) oder nicht (`n`).

`prim`: Motivation für das Implantat, entweder kosmetisch oder rekonstruktiv.

Das Hauptinteresse bei der folgenden statistischen Analyse liegt am Einfluss der Variable `dau`, welche nichtparametrisch ins Modell eingehen soll.

1.2. Ausblick

Das folgende Kapitel 2 startet mit einer Einleitung und Wiederholung wichtiger Konzepte für generalisierte lineare Modelle, welche die Grundlage für die in dieser Arbeit betrachteten Regressionsmodelle bilden. Kapitel 3 beschäftigt sich mit kategorieller Datenanalyse und zeigt wie generalisierte lineare Modelle für binomiale, multinomiale und ordinale Responses angewendet werden können. In Kapitel 4 findet man eine Einführung in nichtparametrische Regression zunächst für lineare und dann auch für generalisierte lineare Modelle. Es werden verschiedene Ansätze zur Basiskonstruktion sowie die Idee der pönalisierten Least-Squares-Schätzung und automatische Glättungsparameterbestimmung diskutiert. In Kapitel 5 werden diese beiden Konzepte dann kombiniert und auf verschiedene Aspekte des Silikonregisters in Zusammenhang mit dem Auftreten einer Kapselkontraktur angewendet. Die Modellierung erfolgt mit der Software `R` (R Core Team, 2014).

2. Generalisierte Lineare Modelle

Lineare Regressionsmodelle sind ein wichtiges Tool in der statistischen Modellierung. Um ein lineares Regressionsmodell verwenden zu können, müssen wir allerdings stets annehmen, dass die Response-Variable normalverteilt ist. In vielen Anwendungen ist dies aber nicht der Fall. Man denke z. B. an unser Silikonregister, wofür wir Risikowahrscheinlichkeiten für verschiedene Gruppen schätzen wollen. Die Response-Variable stellt in diesem Fall also Häufigkeiten für Kategorien dar, was uns zu einer Multinomialverteilung führen wird. Generalisierte Lineare Modelle stellen eine Verallgemeinerung der Theorie der linearen Regressionsmodelle dar. Response-Variablen in generalisierten linearen Modellen können aus einer bestimmten Verteilungsfamilie, die die Normalverteilung umfasst, stammen und es wird nicht mehr der Erwartungswert selbst, sondern eine Funktion des Erwartungswertes modelliert. Falls die Quelle nicht explizit erwähnt wird, vergleiche für diesen Abschnitt Hinkley, Reid und Snell (1991) und McCullagh und Nelder (1983).

2.1. Die lineare Exponentialfamilie

Definition 1 (Exponentialfamilie)**.** *Eine Zufallsvariable Y stammt aus der* ***linearen Exponentialfamilie****, falls man die Dichte- oder Wahrscheinlichkeitsfunktion folgendermaßen schreiben kann*

$$f(y,\theta) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}. \tag{2.1.1}$$

Dabei sind $b(\cdot)$, $c(\cdot)$ und $a(\cdot)$ bekannte Funktionen und θ und ϕ Parameter. Meistens hat $a(\phi)$ die Form $a(\phi) = w\phi$, wobei w ein bekanntes Gewicht ist. Der Parameter ϕ wird ***Dispersionsparameter*** *genannt und ist in manchen Fällen bekannt. Man spricht dann von einer einparametrigen linearen Exponentialfamilie mit* ***kanonischem Parameter*** *θ.*

Bemerkung. Wir werden den Dispersionsparameter ϕ stets als bekannt annehmen.

Die zur linearen Exponentialfamilie gehörende Log-Likelihood-Funktion ist

$$l(\theta, y) = \log f(y,\theta) = \frac{y\theta - b(\theta)}{w\phi} + c(y,\phi). \tag{2.1.2}$$

Die folgenden zwei Identitäten gelten unter bestimmten Regularitätsbedingungen und insbesondere für Verteilungen aus der Exponentialfamilie, siehe Cox und Hinkley (1974, S.107 ff.):

$$\mathbb{E}\left[\frac{\partial l(\theta, Y)}{\partial \theta}\right] = 0 \tag{2.1.3}$$

$$\mathbb{E}\left[\frac{\partial^2 l(\theta, Y)}{\partial \theta^2}\right] + \mathbb{E}\left[\left(\frac{\partial l(\theta, Y)}{\partial \theta}\right)^2\right] = 0 \tag{2.1.4}$$

Damit können wir zeigen, dass für die Exponentialfamilie $\mathbb{E}[Y] = b'(\theta)$ und $\mathbb{V}\mathrm{ar}[Y] = w\phi b''(\theta)$ gilt:

Aus den Eigenschaften (2.1.2) und (2.1.3) folgt

$$\begin{aligned}\mathbb{E}\left[\frac{\partial l(\theta, Y)}{\partial \theta}\right] &= \mathbb{E}\left[\frac{Y - b'(\theta)}{w\phi}\right] \\ &= \frac{\mathbb{E}[Y] - b'(\theta)}{w\phi} \\ &= 0.\end{aligned}$$

Damit folgt

$$\mathbb{E}[Y] = b'(\theta).$$

Aus den Eigenschaften (2.1.2) und (2.1.4) folgt

$$\begin{aligned}\mathbb{E}\left[\frac{\partial^2 l(\theta, Y)}{\partial \theta^2}\right] + \mathbb{E}\left[\left(\frac{\partial l(\theta, Y)}{\partial \theta}\right)^2\right] &= \mathbb{E}\left[-\frac{b''(\theta)}{w\phi}\right] + \mathbb{E}\left[\frac{(Y - b'(\theta))^2}{(w\phi)^2}\right] \\ &= -\frac{b''(\theta)}{w\phi} + \frac{\mathbb{E}[(Y - \mathbb{E}[Y])^2]}{(w\phi)^2} \\ &= -\frac{b''(\theta)}{w\phi} + \frac{\mathbb{V}\mathrm{ar}[Y]}{(w\phi)^2} \\ &= 0.\end{aligned}$$

Damit folgt

$$\mathbb{V}\mathrm{ar}[Y] = w\phi b''(\theta),$$

wobei wir $V(\mu) = b''(\theta)$ setzen und **Varianzfunktion** nennen.

Beispiele für Verteilungen aus der einparametrigen linearen Exponentialfamilie.

1. Normalverteilung

   Wir verifizieren nun, dass die Dichte einer Normalverteilung zur Exponentialfamilie gehört. Es sei $Y \sim N(\mu, \sigma^2)$. Für eine Realisierung $y \in \mathbb{R}$ und ein festes σ^2 ist die Dichte gegeben durch

$$
\begin{aligned}
f(y, \mu, \sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\} \\
&= \exp\left\{-\frac{1}{2}\log(2\pi\sigma^2) - \frac{(y-\mu)^2}{2\sigma^2}\right\} \\
&= \exp\left\{-\frac{1}{2}\log(2\pi\sigma^2) - \frac{y^2}{2\sigma^2} + \frac{y\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2}\right\}.
\end{aligned}
$$

Mit $\theta = \mu$, $\phi = \sigma^2$, $w = 1$, $b(\theta) = \frac{1}{2}\theta^2$ und $c(y, \phi) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{y^2}{2\sigma^2}$ ist die Dichte einer Normalverteilung also von der Form (2.1.1).

Wir rechnen noch nach, dass $b'(\theta) = \theta = \mu = \mathbb{E}[Y]$ und $w\phi b''(\theta) = w\phi = \sigma^2 = \mathbb{V}\mathrm{ar}[Y]$ ist.

2. Binomialverteilung, siehe Agresti (2013, S.132).

   Es sei $nY \sim Bin(n, \pi)$. Die Zufallsvariable Y ist also der *relative Anteil* an Erfolgen bei einer Reihe von n unabhängigen Versuchen eines Bernoulliexperiments mit Erfolgswahrscheinlichkeit π. Demnach ist nY die *absolute Häufigkeit* der Erfolge bei n Versuchen. Für Realisierungen $y = 0, \frac{1}{n}, \frac{2}{n}, \ldots, 1$ ist die Wahrscheinlichkeitsfunktion gegeben als

$$
\begin{aligned}
f(y, \pi) &= \binom{n}{ny} \pi^{ny}(1-\pi)^{n-ny} \\
&= \exp\left\{\log\binom{n}{ny} + ny\log(\pi) + (n-ny)\log(1-\pi)\right\} \\
&= \exp\left\{ny\log\left(\frac{\pi}{1-\pi}\right) + n\log(1-\pi) + \log\binom{n}{ny}\right\} \\
&= \exp\left\{\frac{y\log\left(\frac{\pi}{1-\pi}\right) + \log(1-\pi)}{\frac{1}{n}} + \log\binom{n}{ny}\right\}.
\end{aligned}
$$

   Mit $\theta = \log\left(\frac{\pi}{1-\pi}\right)$, $\phi = 1$, $w = \frac{1}{n}$, $b(\theta) = -\log(1-\pi) = \log\left(\frac{1}{1-\pi}\right) = \log(1 + \exp\{\theta\})$ und $c(y, \phi) = \log\binom{n}{ny} = \log\binom{1/w}{y/w}$ ist dies von der gewünschten Form (2.1.1).

   Wir verifizieren, dass

$$
b'(\theta) = \frac{\exp\{\theta\}}{1 + \exp\{\theta\}} = \pi = \mathbb{E}[Y]
$$

   und

$$
\begin{aligned}
w\phi b''(\theta) &= \frac{1}{n}\frac{\exp\{\theta\}(1+\exp\{\theta\}) - \exp\{\theta\}^2}{(1+\exp\{\theta\})^2} \\
&= \frac{1}{n}\frac{\exp\{\theta\}}{(1+\exp\{\theta\})^2} \\
&= \frac{1}{n}\pi(1-\pi) = \mathbb{V}\mathrm{ar}[Y].
\end{aligned}
$$

Bemerkung. Weitere Beispiele für Verteilungen aus der linearen Exponentialfamilie sind u. a. die Poisson-, Gamma- oder die Inverse Normalverteilung.

2.2. Modellspezifikation

Die Beobachtung $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^\top$ sei Realisierung des Response-Vektors $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)^\top$ mit Erwartungswert $\mathbb{E}[\boldsymbol{Y}] = \boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_n)^\top$. Die erklärenden Variablen seien in einer $(n \times p)$-Design-Matrix gegeben, wobei eine Spalte genau einer erklärenden Variable entspricht. Es sei $\boldsymbol{x}_i$ ein Spaltenvektor, dessen Einträge genau die i-te Zeile der Design-Matrix sind. Der Vektor enthält also die Werte aller erklärenden Variablen für die i-te Beobachtung.

Wir wollen ein parametrisches Modell für den Erwartungswert $\boldsymbol{\mu}$ der Response $\boldsymbol{Y}$ finden. Dies sei von der Form

$$\mathbb{E}[\boldsymbol{Y}] = \boldsymbol{\mu}(\beta_0, \beta_1, \ldots, \beta_{p-1}),$$

wobei $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_{p-1})^\top$ ein unbekannter Parametervektor der Dimension p ist. Das Ziel ist es, den Parametervektor $\boldsymbol{\beta}$ und damit auch den Erwartungswert $\boldsymbol{\mu}$ zu schätzen.

Definition 2 (Generalisiertes Lineares Modell)**.** *Ein* ***Generalisiertes Lineares Modell****, kurz GLM, besteht typischerweise aus drei Komponenten:*

1. ***Verteilungsannahme****: Wir nehmen an, dass die Y_i unabhängig verteilt sind mit $\mathbb{E}[Y_i] = \mu_i$ und aus der Exponentialfamilie (2.1.1) stammen, d. h. $Y_i \overset{ind}{\sim}$ Exponentialfamilie(θ_i) für $i = 1, \ldots, n$.*
2. ***Linearer Prädiktor****: Der Vektor $\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_n)^\top$ wird linearer Prädiktor genannt, wobei $\eta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}$ für $i = 1, \ldots, n$.*
3. ***Link-Funktion****: Die Link-Funktion g ist eine monotone und differenzierbare Funktion, die den Erwartungswert und den linearen Prädiktor durch $g(\boldsymbol{\mu}) = \boldsymbol{\eta}$ in Beziehung setzt.*

Im Unterschied zum klassischen linearen Regressionsmodell wird bei GLMs eine Funktion des Erwartungswertes der Response linear modelliert. Des Weiteren kann die Varianzstruktur durch die Varianzfunktion $V(\boldsymbol{\mu}) = b''(\boldsymbol{\theta})$ vom Erwartungswert abhängen.

Es sei Y_i aus der Exponentialfamilie mit Dichte wie in (2.1.1). Dann gilt $\mathbb{E}[Y_i] = \mu_i = b'(\theta_i)$, wie wir schon gezeigt haben. Die spezielle Link-Funktion $g(\cdot) = (b')^{-1}(\cdot)$ heißt **kanonische Link-Funktion**. Bei Verwendung der kanonischen Link-Funktion wird der kanonische Parameter θ_i der Exponentialfamilie durch den linearen Prädiktor η_i modelliert, denn aus

$$\mu_i = b'(\theta_i) \text{ und } \eta_i = g(\mu_i) \quad \text{folgt dann} \quad \eta_i = \theta_i.$$

Beispiele für kanonische Link-Funktionen.

1. Normalverteilung

   Es gilt $b'(\theta_i) = \theta_i$, also $b'(\cdot) = id(\cdot)$ und $(b')^{-1}(\cdot) = id(\cdot)$. Die identische Abbildung ist die kanonische Link-Funktion der Normalverteilung.

2. Binomialverteilung

   Es gilt $\pi_i = b'(\theta_i) = \frac{\exp\{\theta_i\}}{1+\exp\{\theta_i\}}$ und somit $\theta_i = \log(\frac{\pi_i}{1-\pi_i}) = logit(\pi_i)$. Die Funktion $logit(\cdot)$ ist also kanonische Link-Funktion der Binomialverteilung und wird **Logit-Link** genannt.

Link-Funktionen dienen u. a. dazu, gewisse Voraussetzungen an die Erwartungswerte sicherzustellen. Bei der Modellierung von Wahrscheinlichkeiten sollen diese z. B. in $(0, 1)$ liegen. Um dies zu erreichen, verwendet man als Link-Funktion die inverse Abbildung einer beliebigen Verteilungsfunktion. Will man nur positive gefittete Werte zulassen, wie beispielsweise bei der Modellierung von Zählvariablen, wird der **Log-Link** verwendet.

Beispiele.

1. Bei der Normalverteilung gibt es durch die identische Abbildung als Link-Funktion keine Einschränkung. Die Erwartungswerte können aus ganz $\mathbb{R}$ sein.

2. Die kanonische Link-Funktion der Binomialverteilung ist die inverse Abbildung der Verteilungsfunktion der logistischen Verteilung $L(0, 1)$. Für Details siehe Anhang A.4.

2.3. Maximum-Likelihood-Schätzung

Die Schätzung des Parametervektors $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_{p-1})^\top$ erfolgt durch Maximum-Likelihood-Schätzung (ML-Schätzung), d. h. es wird die Likelihood-Funktion der Stichprobe maximiert. In der Praxis wird einfachheitshalber meistens die Log-Likelihood-Funktion maximiert. Wegen der Monotonie des Logarithmus liefert dies dieselbe Maximalstelle.

Die Responses seien $Y_i \overset{ind}{\sim} Exponentialfamilie(\theta_i)$ mit $\mathbb{E}[Y_i] = \mu_i$ für $i = 1, \ldots, n$ und wir betrachten das Modell $g(\mu_i) = \eta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}$. Die gemeinsame Dichte ist wegen der Unabhängigkeit der Y_i gegeben durch

$$
\begin{aligned}
f(\boldsymbol{y}, \boldsymbol{\theta}) &= \prod_{i=1}^{n} f(y_i, \theta_i) \\
&= \prod_{i=1}^{n} \exp\left\{ \frac{y_i\theta_i - b'(\theta_i)}{w_i\phi} + c(y_i, \phi) \right\} \\
&= \exp\left\{ \sum_{i=1}^{n} \left(\frac{y_i\theta_i - b'(\theta_i)}{w_i\phi} + c(y_i, \phi) \right) \right\}
\end{aligned}
$$

und die Log-Likelihood-Funktion ist

$$
l(\boldsymbol{\theta}, \boldsymbol{y}) = \sum_{i=1}^{n} l_i(\theta_i, y_i) = \sum_{i=1}^{n} \left(\frac{y_i\theta_i - b'(\theta_i)}{w_i\phi} + c(y_i, \phi) \right).
$$

Den ML-Schätzer für $\boldsymbol{\beta}$ erhalten wir durch Nullsetzen der Score-Funktionen, d. h.

$$
\frac{\partial l(\boldsymbol{\theta}, \boldsymbol{y})}{\partial \boldsymbol{\beta}} \stackrel{!}{=} \mathbf{0},
$$

wobei $\mathbf{0}$ der p-dimensionale Nullvektor ist. Mittels Kettenregel erhalten wir

$$
\frac{\partial l(\boldsymbol{\theta}, \boldsymbol{y})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \left(\frac{\partial l_i(\theta_i, y_i)}{\partial \theta_i} \, \frac{\partial \theta_i}{\partial \mu_i} \, \frac{\partial \mu_i}{\partial \eta_i} \, \frac{\partial \eta_i}{\partial \boldsymbol{\beta}} \right).
$$

Wir berechnen die einzelnen Faktoren in der Summe separat:

$$
\begin{aligned}
&\frac{\partial l_i(\theta_i, y_i)}{\partial \theta_i} = \frac{y_i - b'(\theta_i)}{w_i\phi} = \frac{y_i - \mu_i}{w_i\phi}, \\
&\frac{\partial \theta_i}{\partial \mu_i} = \frac{1}{b''(\theta_i)} = \frac{1}{V(\mu_i)}, \text{ da } \quad \frac{\partial \mu_i}{\partial \theta_i} = \frac{\partial b'(\theta_i)}{\partial \theta_i} = b''(\theta_i), \\
&\frac{\partial \mu_i}{\partial \eta_i} = \frac{1}{g'(\mu_i)}, \text{ da } \quad \frac{\partial \eta_i}{\partial \mu_i} = \frac{\partial g(\mu_i)}{\partial \mu_i} = g'(\mu_i), \\
&\frac{\partial \eta_i}{\partial \boldsymbol{\beta}} = \boldsymbol{x}_i.
\end{aligned}
$$

Insgesamt erhalten wir die Score-Funktionen

$$
\frac{\partial l(\boldsymbol{\theta}, \boldsymbol{y})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \frac{y_i - \mu_i}{w_i\phi V(\mu_i)} \frac{\boldsymbol{x}_i}{g'(\mu_i)} \tag{2.3.1}
$$

bzw. das Score-Gleichungssystem

$$
\sum_{i=1}^{n} \frac{y_i - \hat{\mu}_i}{w_i\phi V(\hat{\mu}_i)} \frac{\boldsymbol{x}_i}{g'(\hat{\mu}_i)} \stackrel{!}{=} \mathbf{0}. \tag{2.3.2}
$$

Dies ist im Allgemeinen ein nichtlineares System und kann nur iterativ gelöst werden. Verwendet wird dazu ein Newton-Raphson-Verfahren und iterative gewichtete Least-Squares-Schätzung, für Details siehe Anhang A.2. Als Lösung erhalten wir den Parameterschätzer von $\boldsymbol{\beta}$, den wir mit $\hat{\boldsymbol{\beta}}$ bezeichnen. Der geschätzte Erwartungswert wird mit $\hat{\boldsymbol{\mu}} = g^{-1}(\boldsymbol{X}\hat{\boldsymbol{\beta}})$ bezeichnet.

Bemerkung 1. Wenn wir konstante Varianz, also $V(\mu_i) = 1$ und $w_i = 1$ voraussetzen, was bei der Normalverteilung der Fall ist, vereinfachen sich die Gleichungen (2.3.2) bei Verwendung der kanonischen Link-Funktion zu

$$\sum_{i=1}^{n} \frac{y_i - \hat{\mu}_i}{\sigma^2} \boldsymbol{x}_i \stackrel{!}{=} \mathbf{0}, \quad \Longleftrightarrow \quad \sum_{i=1}^{n} (y_i - \hat{\mu}_i)\boldsymbol{x}_i \stackrel{!}{=} \mathbf{0}.$$

Dies ist äquivalent zu den Normalengleichungen, die wir bei der Minimierung der Fehlerquadratsumme bei einem klassischen linearen Regressionsmodell erhalten, denn in diesem Fall ist

$$SSE(\boldsymbol{\beta}) = \quad \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2$$

und das Score-Gleichungssystem ist dann

$$\sum_{i=1}^{n} 2(y_i - \hat{\mu}_i)(-\boldsymbol{x}_i) \stackrel{!}{=} \mathbf{0} \quad \Longleftrightarrow \quad \sum_{i=1}^{n} (y_i - \hat{\mu}_i)\boldsymbol{x}_i \stackrel{!}{=} \mathbf{0}.$$

Bemerkung 2. Für kanonische Link-Funktionen gilt $\eta_i = \theta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}$ für $i = 1, \ldots, n$ und die Score-Funktion (2.3.1) lässt sich vereinfachen zu

$$\frac{\partial l(\boldsymbol{\theta}, \boldsymbol{y})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \left(\frac{\partial l_i(\theta_i, y_i)}{\partial \theta_i} \frac{\partial \theta_i}{\partial \boldsymbol{\beta}} \right) = \sum_{i=1}^{n} \frac{y_i - \mu_i}{w_i \phi} \boldsymbol{x}_i. \tag{2.3.3}$$

2.4. Modellselektion und Modellvalidierung

Vergleiche für diesen Abschnitt v. a. Hinkley et al. (1991, S.66 ff.).

Zur Beurteilung der Anpassungsgüte eines Modells und in weiterer Folge zum Vergleich zweier GLMs brauchen wir ein Gütekriterium („Goodness-Of-Fit-Kriterium“). Als Maß dafür definieren wir die Deviance eines GLMs, die die Abweichung der Anpassung unseres aktuellen Modells mit ML-Schätzer $\hat{\boldsymbol{\mu}}$ vom saturierten Modell beschreibt. Das saturierte Modell enthält für jede Beobachtung y_i einen Parameter μ_i und liefert daher perfekte Anpassung. Der ML-Schätzer in diesem Fall ist $\hat{\mu}_i = y_i$ für $i = 1, \ldots, n$ (Beweis siehe Anhang A.3). Die Deviance ist die doppelte logarithmierte Likelihood-Quotienten-Teststatistik basierend auf diesen zwei Modellen.

Definition 3 (Deviance)**.** *Die* ***skalierte Deviance*** *ist definiert als*

$$\frac{1}{\phi}D(\hat{\boldsymbol{\mu}}, \boldsymbol{y}) = -2\{l(\hat{\boldsymbol{\mu}}, \boldsymbol{y}) - l(\boldsymbol{y}, \boldsymbol{y})\}, \tag{2.4.1}$$

wobei $l(\hat{\boldsymbol{\mu}}, \boldsymbol{y})$ *die maximierte Log-Likelihood des aktuellen Modells und* $l(\boldsymbol{y}, \boldsymbol{y})$ *die Log-Likelihood des saturierten Modells ist.*

Aus der Definition folgt, dass die Deviance des saturierten Modells Null ist. Aus $l(\boldsymbol{y}, \boldsymbol{y}) \geq l(\hat{\boldsymbol{\mu}}, \boldsymbol{y})$ folgt, dass die Deviance stets positiv ist, denn

$$\frac{1}{\phi}D(\hat{\boldsymbol{\mu}}, \boldsymbol{y}) = -2\{l(\hat{\boldsymbol{\mu}}, \boldsymbol{y}) - l(\boldsymbol{y}, \boldsymbol{y})\} \geq 0.$$

Prinzipiell gilt, je größer die Deviance, desto größer die Abweichung des aktuellen Modells vom saturierten Modell, welches eine perfekte Anpassung liefert. Deshalb versuchen wir, die Deviance zu minimieren. Die Minimierung der Deviance ist dabei äquivalent zu Maximierung der Log-Likelihood des aktuellen Modells.

Vergleich zweier verschachtelter Modelle („nested models“)

Es seien zwei ineinander verschachtelte GLMs, Modell A und Modell B, gegeben, wobei Modell A ein Untermodell von Modell B ist. Modell A habe q Parameter, Modell B habe p Parameter und es gelte $q < p$. Die Spalten der $(n \times q)$-Design-Matrix zu Modell A sind also im Spaltenraum der $(n \times p)$-Design-Matrix zu Modell B enthalten. Es seien $\hat{\boldsymbol{\mu}}_A$ und $\hat{\boldsymbol{\mu}}_B$ die ML-Schätzer unter dem jeweiligen Modell. Modell A ist das einfachere Modell und hat eine kleinere Menge von Parametern. Einfachere Modelle haben eine größere Deviance, denn es gilt

$$l(\hat{\boldsymbol{\mu}}_A, \boldsymbol{y}) \leq l(\hat{\boldsymbol{\mu}}_B, \boldsymbol{y}) \quad \text{bzw.} \quad D(\hat{\boldsymbol{\mu}}_B, \boldsymbol{y}) \leq D(\hat{\boldsymbol{\mu}}_A, \boldsymbol{y}).$$

Wir wollen nun testen, ob Modell A gegenüber Modell B zu bevorzugen ist. Dies entspricht dem Test einer Hypothese der Form

$$H_0 : \beta_q = \cdots = \beta_{p-1} = 0 \quad \text{vs.} \quad H_1 : \beta_q, \ldots, \beta_{p-1} \text{ beliebig.}$$

Die Likelihood-Quotienten-Teststatistik ist dann die Deviance-Differenz der beiden Modelle

$$-2\{l(\hat{\boldsymbol{\mu}}_A, \boldsymbol{y}) - l(\hat{\boldsymbol{\mu}}_B, \boldsymbol{y})\} = \frac{1}{\phi}\{D(\hat{\boldsymbol{\mu}}_A, \boldsymbol{y}) - D(\hat{\boldsymbol{\mu}}_B, \boldsymbol{y})\}.$$

Unter gewissen Regularitätsbedingungen gilt, dass diese Teststatistik asymptotisch χ^2_{p-q}-verteilt ist. Der Vergleich mit dem entsprechenden Quantil dieser Verteilung liefert Aufschluss darüber, ob die entsprechenden Terme im Modell relevant sind.

Bemerkung.

- Obige Betrachtungen gehen davon aus, dass ϕ bekannt ist. Ist dies nicht der Fall, muss auch dieser Parameter geschätzt werden (für Details dazu siehe Anhang A.2). Parametertests erfolgen dann ähnlich wie bei klassischen linearen Regressionsmodellen mittels eines F-Tests.
- Die χ^2-Approximation hält im Allgemeinen nur für die Deviance-Differenz zweier verschachtelter Modelle. Die Verteilung der Deviance eines einzelnen Modells kann unter Umständen weit von einer χ^2-Verteilung entfernt sein.

3. Kategorielle Datenanalyse

In unserem Datensatz aus dem Silikonregister von Österreich wurden für Patientinnen verschiedene Faktoren erhoben. Es wurde z. B. festgestellt ob eine Kapselkontraktur aufgetreten ist oder nicht. Dies ist in der Variable `kontr_flag` mit den zwei Stufen {`yes`, `no`} gespeichert. Im Falle einer Kapselkontraktur wurde der Schweregrad der Kontraktur in der Variable `kontr` mit den vier Stufen {`BakerI`, `BakerII`, `BakerIII`, `BakerIV`} festgehalten. Wir werden diese Faktoren als Response-Variable verwenden und die Auftrittswahrscheinlichkeiten für die verschiedenen Stufen des jeweiligen Faktors schätzen. Man spricht von kategorieller Datenanalyse.

In diesem Abschnitt soll die Theorie der kategoriellen Datenanalyse schrittweise aufgebaut werden. Wir beginnen dabei mit Modellen für zweistufige Faktoren (z. B. `kontr_flag`). Die Response-Variable ist dann also eine binäre Variable und als Verteilung resultiert die Binomialverteilung. Man spricht von binomialen Regressionsmodellen. Dieses werden wir erweitern zu multinomialen Response-Modellen, welche Faktoren mit mehr als zwei Stufen als Response-Variable haben können. Im letzten Schritt sollen dann auch Response-Variablen berücksichtigt werden, deren Stufen eine Ordnung haben, was uns zu ordinalen Regressionsmodellen führen wird. Ein Beispiel dafür ist die Variable `kontr` aus dem Silikonregister, deren Stufen eine stärker werdende Kapselkontraktur bezeichnen. `BakerI` steht beispielsweise für eine leichte Kapselkontraktur und `BakerIV` für eine sehr starke Kapselkontraktur. Zusätzlich soll zur Theorie anhand von einfachen Beispielen auch die Umsetzung in `R` erklärt werden. Vergleiche für diesen Abschnitt hauptsächlich Agresti (2013), Fahrmeir, Kneib, Lang und Marx (2013) und eventuell explizit erwähnte Quellen für einzelne Unterabschnitte.

3.1. Binomiale Regression

Wir wollen uns zuerst mit binomialer Regression beschäftigen und auch die binäre Regression als Spezialfall betrachten. Man spricht von binomialer Regression, falls die Response-Variablen aus einer Binomialverteilung stammen. Diese gehört zur Exponentialfamilie und wir können das bekannte Framework aus Abschnitt 2 zur Parameterschätzung und Modellvalidierung verwenden.

3.1.1. Modelle für binomiale Responses

Wir sprechen von binärer Regression, wenn $Y_i \in \{0, 1\}$ für $i = 1, \ldots, n$. Die Response-Variablen sind hier Bernoulli-Variablen, d. h.

$$Y_i \overset{ind}{\sim} Bin(1, \pi_i).$$

Wir wollen den Erwartungswert $\mathbb{E}[Y_i]$ modellieren, der in diesem Fall die (unbekannte) Erfolgswahrscheinlichkeit ist, denn

$$\mathbb{E}[Y_i] = \mathbb{P}[Y_i = 1] = \pi_i.$$

Falls nur Faktoren als erklärende Variablen vorliegen, können verschiedene Beobachtungen die gleiche Design-Zeile haben. Wir können die Daten gruppieren, indem wir diese Beobachtungen zusammenfassen. Die Gruppen entsprechen den verschiedenen Spezifikationen der Design-Zeilen.

Es seien zunächst binäre Response-Variablen $Y_{ij} \overset{ind}{\sim} Bin(1, \pi_i)$ für $i = 1, \ldots, n$ und $j = 1, \ldots, n_i$ gegeben. Wir haben also n Gruppen mit je n_i Beobachtungen. Es seien Y_i für $i = 1, \ldots, n$ neue Response-Variablen, die den relativen Anteil an Beobachtungen mit $Y_{ij} = 1$ für eine bestimmte Gruppe angeben, d. h. $Y_i = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}$. Es gilt also $Y_i \in [0, 1]$ und $n_i Y_i \overset{ind}{\sim} Bin(n_i, \pi_i)$, wobei wieder $\mathbb{E}[Y_i] = \pi_i$.

Zusammenfassend haben wir bei der binomialen Regression also Response-Variablen mit $n_i Y_i \overset{ind}{\sim} Bin(n_i, \pi_i)$ für $i = 1, \ldots, n$ gegeben. Ist $n_i = 1$ für alle $i = 1, \ldots, n$ sprechen wir von binärer Regression.

Bei der Modellierung des Erwartungswertes $\mathbb{E}[y_i] = \pi_i$ muss sichergestellt werden, dass die gefitteten Werte $\hat{\pi}_i$ in $(0, 1)$ sind. Dies wird durch das allgemeine **binomiale Modell**

$$\pi_i = F(\boldsymbol{x}_i^\top \boldsymbol{\beta})$$

erreicht, wobei F eine Verteilungsfunktion ist. Die Link-Funktion g ist dann die Umkehrfunktion von F. Wie immer ist $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{i,p-1})^\top$ die i-te Zeile der Design-Matrix und $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_{p-1})^\top$ der Parametervektor. Der lineare Prädiktor sei $\eta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}$.

Die Verwendung der Verteilungsfunktion der logistischen Verteilung $L(0, 1)$ (siehe Anhang A.4) führt uns zum sogenannten **Logit-Modell**,

$$\eta_i = \text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1 - \pi_i}\right) \quad \text{bzw.} \quad \pi_i = \frac{\exp\{\eta_i\}}{1 + \exp\{\eta_i\}}. \tag{3.1.1}$$

Das Logit-Modell wird in der Praxis häufig verwendet, da der Logit-Link die kanonische Link-Funktion der Binomialverteilung ist. Man spricht dann von **logistischer Regression**. Alternativ könnte man auch die Umkehrfunktion anderer Verteilungsfunktionen als Link-Funktion verwenden. Bei der Verwendung der Umkehrfunktion der Standardnormalverteilung resultiert beispielsweise das **Probit-Modell**

$$\eta_i = \Phi^{-1}(\pi_i) \quad \text{bzw.} \quad \pi_i = \Phi(\eta_i). \tag{3.1.2}$$

Ein weiteres Beispiel ist der Complementary-Log-Log-Link. Er resultiert als Umkehrfunktion der Verteilungsfunktion einer Zufallsvariable $-Z$, wobei Z aus einer Gumbel-Verteilung (siehe Anhang A.5) stammt. Das **Complementary-Log-Log-Modell** ist

$$\eta_i = \log(-\log(1-\pi_i)) \quad \text{bzw.} \quad \pi_i = 1 - \exp\{-\exp\{\eta_i\}\}. \tag{3.1.3}$$

Abbildung 3.1 zeigt einen Vergleich der drei verschiedenen Link-Funktionen. Während Logit-Link und Probit-Link symmetrisch um 0.5 sind, ist der Complementary-Log-Log-Link eine asymmetrische Link-Funktion, deren Annäherung an die 1 schneller als die Annäherung an 0 ist.

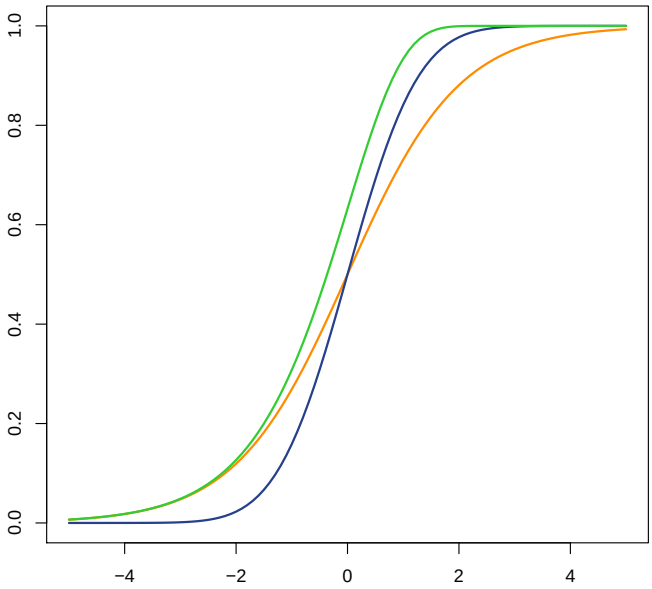


Abbildung 3.1.: Vergleich der verschiedenen Link-Funktionen. Logit-Link (orange), Probit-Link (blau) und Complementary-Log-Log-Link (grün).

3.1.2. Interpretation der Parameter beim Logit-Modell

Das Logit-Modell hat den weiteren Vorteil, dass sich die Parameter als spezielle Odds und Odds-Ratios interpretieren lassen.

Definition 4 (Odds, Log-Odds)**.** *Es sei Y eine binäre Response-Variable mit $\mathbb{P}[Y=1] = \pi$. Die **Odds** für das Eintreten von $Y=1$ gegenüber $Y=0$ bei gegebenen Werten $\boldsymbol{x}$ für die erklärenden Variablen sind definiert als*

$$\text{odds}(\boldsymbol{x}) = \frac{\mathbb{P}[Y=1|\boldsymbol{x}]}{\mathbb{P}[Y=0|\boldsymbol{x}]} = \frac{\pi}{1-\pi}.$$

*Die **logarithmierten Odds** (Log-Odds) sind*

$$\log\left(\frac{\pi}{1-\pi}\right) = \text{logit}(\pi).$$

Definition 5 (Odds-Ratio, Log-Odds-Ratio)**.** *Das **Odds-Ratio** dient dem Vergleich der Odds für zwei Vektoren $\boldsymbol{x}$ und $\tilde{\boldsymbol{x}}$ mit verschiedenen Werten für die erklärenden Variablen und ist definiert als Quotient der jeweiligen Odds,*

$$\frac{\text{odds}(\boldsymbol{x})}{\text{odds}(\tilde{\boldsymbol{x}})} = \frac{\mathbb{P}[Y=1|\boldsymbol{x}]/\mathbb{P}[Y=0|\boldsymbol{x}]}{\mathbb{P}[Y=1|\tilde{\boldsymbol{x}}]/\mathbb{P}[Y=0|\tilde{\boldsymbol{x}}]}.$$

*Das **logarithmierte Odds-Ratio** (Log-Odds-Ratio) ist die Logit-Differenz der beiden Erfolgswahrscheinlichkeiten*

$$\log\left(\frac{\text{odds}(\boldsymbol{x})}{\text{odds}(\tilde{\boldsymbol{x}})}\right) = \text{logit}(\mathbb{P}[Y=1|\boldsymbol{x}]) - \text{logit}(\mathbb{P}[Y=1|\tilde{\boldsymbol{x}}]).$$

Das Logit-Modell liefert ein *additives* Modell für die Log-Odds und ein *multiplikatives* Modell für die Odds, da für $\boldsymbol{x} = (1, x_1, \ldots, x_{p-1})^\top$

$$\text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \boldsymbol{x}^\top\boldsymbol{\beta} = \beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1}x_{p-1}$$

bzw.

$$\frac{\pi}{1-\pi} = \exp\{\boldsymbol{x}^\top\boldsymbol{\beta}\} = \exp\{\beta_0\}\exp\{\beta_1 x_1\}\cdots\exp\{\beta_{p-1}x_{p-1}\}.$$

Angenommen x_j ist stetig für ein $j \in \{1, \ldots, p-1\}$ und wird um eine Einheit erhöht, während die restlichen Werte im Vektor $\boldsymbol{x}$ gleich bleiben. Dann multiplizieren sich die Odds für das Eintreten von $Y=1$ gegenüber $Y=0$ mit $\exp\{\beta_j\}$. In Abhängigkeit vom Vorzeichen des Parameters β_j erhält man unterschiedliche Fälle. Ist $\beta_j < 0$, folgt $\exp\{\beta_j\} < 1$ und die Odds verringern sich. Umgekehrt erhöhen sich die Odds für $\beta_j > 0$, denn dann ist $\exp\{\beta_j\} > 1$. Falls $\beta_j = 0$, folgt $\exp\{\beta_j\} = 1$ und die Odds bleiben gleich.

Für $\boldsymbol{x} = (1, x_1, \ldots, x_j+1, \ldots, x_{p-1})^\top$ und $\tilde{\boldsymbol{x}} = (1, x_1, \ldots, x_j, \ldots, x_{p-1})^\top$ ist das Odds-Ratio gegeben als

$$\frac{\text{odds}(\boldsymbol{x})}{\text{odds}(\tilde{\boldsymbol{x}})} = \frac{\exp\{\beta_0\}\cdots\exp\{\beta_j\}^{x_j+1}\cdots\exp\{\beta_{p-1}\}^{x_{p-1}}}{\exp\{\beta_0\}\cdots\exp\{\beta_j\}^{x_j}\cdots\exp\{\beta_{p-1}\}^{x_{p-1}}} = \exp\{\beta_j\}.$$

Der Parameter β_j kann also auch als spezielles logarithmiertes Odds-Ratio interpretiert werden.

Bemerkung. Für 2×2-Kontingenztafeln kann das Odds-Ratio verwendet werden um stochastische Unabhängigkeit der zwei Faktoren zu charakterisieren: Das Odds-Ratio ist genau dann 1 (bzw. das Log-Odds-Ratio 0), wenn die zwei Faktoren stochastisch unabhängig sind. Beweis siehe Anhang A.10.

Beispiel. Die Interpretation der Parameter bzw. der Odds soll anhand eines Beispiels erklärt werden. Die binäre Response-Variable gebe an, ob bei einer Patientin eine Kapselkontraktur auftritt ($Y = 1$) oder nicht ($Y = 0$). Dazu verwenden wir die Variable `kontr_flag` aus dem Silikonregister. Wir modellieren die Wahrscheinlichkeit π für eine Kapselkontraktur in Abhängigkeit *einer stetigen* erklärenden Variable x, die der Variable `dau` aus dem Silikonregister entspricht. Dann ist $1 - \pi$ die Wahrscheinlichkeit, dass keine Kapselkontraktur auftritt. Wir schätzen ein Logit-Modell der Form

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x.$$

Die Parameterschätzer, die wir z. B. durch ML-Schätzung in `R` erhalten, sind $\hat{\beta}_0 \approx -0.1$ und $\hat{\beta}_1 \approx 0.04$. Dann sind

$$\hat{\pi} = \hat{\pi}(x) = \frac{\exp\{\hat{\beta}_0 + \hat{\beta}_1 x\}}{1 + \exp\{\hat{\beta}_0 + \hat{\beta}_1 x\}}$$

die geschätzten Wahrscheinlichkeiten, dass eine Kapselkontraktur auftritt in Abhängigkeit der Dauer. Das Vorzeichen von $\hat{\beta}_1$ bestimmt, ob $\hat{\pi}(x)$ monoton steigend oder fallend in x ist, denn

$$\frac{\partial \hat{\pi}(x)}{\partial x} = \frac{\exp\{\hat{\beta}_0 + \hat{\beta}_1 x\}\hat{\beta}_1}{(1 + \exp\{\hat{\beta}_0 + \hat{\beta}_1 x\})^2} = \begin{cases} > 0 & \text{falls } \hat{\beta}_1 > 0, \\ < 0 & \text{falls } \hat{\beta}_1 < 0. \end{cases}$$

Abbildung 3.2 zeigt die geschätzten Wahrscheinlichkeiten $\hat{\pi}(x)$ in Abhängigkeit von der Dauer für die geschätzten Parameter $\hat{\beta}_0 \approx -0.1$ und $\hat{\beta}_1 \approx 0.04$. Die zweite Kurve stellt die hypothetischen Wahrscheinlichkeiten dar, wenn wir das Vorzeichen des Slope-Parameters ändern. Die Werte der Variable x sind stets größer als Null. Zur Veranschaulichung wurde in der Abbildung 3.2 aber auch der theoretische Verlauf der Kurven im negativen Bereich als punktierte Linie eingezeichnet.

Die geschätzten Odds unter dem Logit-Modell sind

$$\widehat{\text{odds}}(x) = \frac{\hat{\pi}}{1 - \hat{\pi}} = \exp\{\hat{\beta}_0\} \exp\{\hat{\beta}_1\}^x.$$

Für $\hat{\beta}_0 \approx -0.1$ und $\hat{\beta}_1 \approx 0.04$ sind die Odds

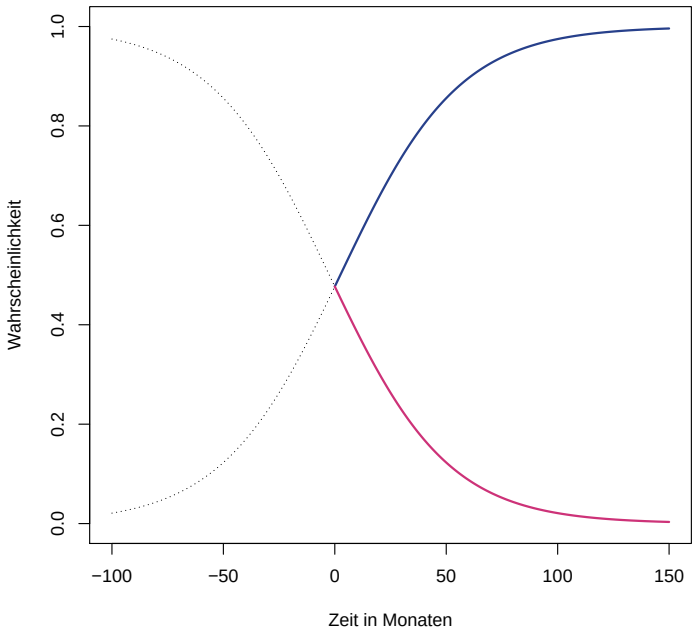


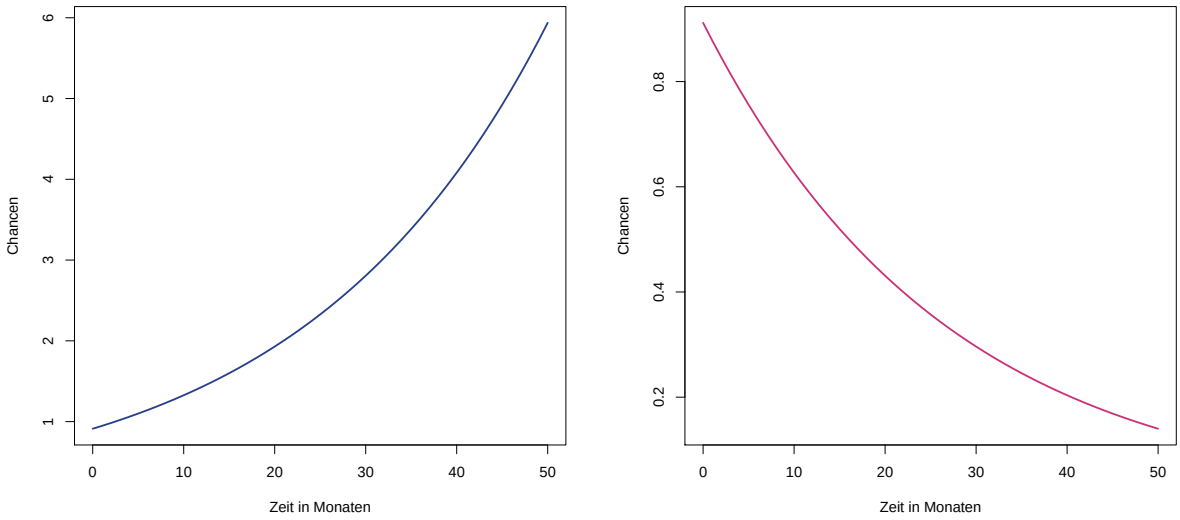
Abbildung 3.2.: Wahrscheinlichkeiten einer Kapselkontraktur für das Logit-Modell mit den geschätzten Parametern $\hat{\beta}_0 \approx -0.1$ und $\hat{\beta}_1 \approx 0.04$ (blau) bzw. mit den hypothetischen Parametern $\bar{\beta}_0 = -0.1$ und $\bar{\beta}_1 \approx -0.04$ (violett).

$$\widehat{\text{odds}}(x) = \exp\{-0.1\} \cdot \exp\{0.04\}^x = 0.91 \cdot 1.04^x.$$

Abbildung 3.3 (a) zeigt die geschätzten Odds in Abhängigkeit der Dauer x, welche sich folgendermaßen interpretieren lassen:

- Bei einer Patientin mit einer Dauer von einem Monat gilt $\widehat{\text{odds}}(\frac{1}{12}) \approx 0.9$. Die Odds eine Kapselkontraktur zu erleiden sind 0.9-mal so groß wie (10 % kleiner als) die Odds keine Kapselkontraktur zu erleiden.
- Bei einer Patientin mit einer Dauer von zweieinhalb Jahren gilt $\widehat{\text{odds}}(2.5) \approx 1$. Die Odds eine Kapselkontraktur zu erleiden bzw. keine zu erleiden sind gleich groß, die zugehörigen Wahrscheinlichkeiten sind jeweils $\frac{1}{2}$.
- Bei einer Patientin mit einer Dauer von 32 Jahren gilt $\widehat{\text{odds}}(32) \approx 3$. Die Odds eine Kapselkontraktur zu erleiden sind etwa 3-mal so groß wie die Odds keine Kapselkontraktur zu erleiden.

Die Interpretation von $\hat{\beta}_1$ als Odds-Ratio liefert folgende Erkenntnisse. Das Odds-Ratio für eine Patientin mit einer Dauer von $(x+1)$ Jahren gegenüber einer Patientin mit x Jahren



(a) $\hat{\beta}_0 \approx -0.1$ und $\hat{\beta}_1 \approx 0.04$

(b) $\bar{\beta}_0 \approx -0.1$ und $\bar{\beta}_1 \approx -0.04$

Abbildung 3.3.: Geschätzte Chancen für eine Kapselkontraktur unter dem Logit-Modell (links) und hypothetisches Verhalten bei negativem Steigungsparameter (rechts).

zwischen den zwei Operationen ist $\exp(\hat{\beta}_1) \approx 1.04$. Die Odds eine Kapselkontraktur zu erleiden nehmen also pro Jahr multiplikativ um den Faktor 1.04 zu (erhöhen sich jährlich um 4 %). Vergleicht man die Odds von Patientinnen mit zweieinhalb bzw. 32 Jahren zwischen den Operationen erhält man

$$\frac{\widehat{\text{odds}}(2.5)}{\widehat{\text{odds}}(32)} \approx \frac{1}{3}.$$

Die Odds einer Patientin nach 2.5 Jahren eine Kapselkontraktur zu erleiden sind ein Drittel der Odds einer Patientin nach 32 Jahren.

Angenommen die Parameter seien gegeben als $\bar{\beta}_0 = -0.1$ und $\bar{\beta}_1 = -0.04$, dann sind die Odds

$$\widehat{\text{odds}} = \exp\{-0.1\} \cdot \exp\{-0.04\}^x = 0.91 \cdot 0.96^x.$$

Abbildung 3.3 (b) zeigt die geschätzten Odds in Abhängigkeit der Dauer für diesen Fall. Im Vergleich zum Beispiel mit positivem Parameterschätzer $\hat{\beta}_1$ sehen wir hier, dass sich die Odds verringern. Sie sind alle kleiner als Eins, d. h. die Odds eine Kapselkontraktur zu erleiden sind für alle Patientinnen kleiner als die Odds keine Kapselkontraktur zu erleiden.

Das Odds-Ratio für $(x+1)$ gegenüber x ist $\exp(\bar{\beta}_1) \approx 0.96$. Die Odds eine Kapselkontraktur zu erleiden nehmen in diesem Fall pro Jahr um den Faktor 0.96 ab.

Dieses Beispiel hat die Interpretation der Parameter und diverser Odds bei einer stetigen erklärenden Variable gezeigt. Wir wollen uns nun überlegen, was sich an der Interpretation ändert falls wir anstatt einer stetigen erklärenden Variable einen erklärenden Faktor haben. In obigem Beispiel könnte das beispielsweise die Variable `oberfl` mit den Kategorien {`glatt`, `Polyurethan`, `texturiert`} sein.

Es sei dazu x ein erklärender Faktor mit L Stufen, wobei mit $j = 1, \ldots, L$ die Faktorstufen bezeichnet werden. Die Stufe 1 wird als Referenzstufe verwendet und es werden $L - 1$ Indikatorvariablen für $j = 2, \ldots, L$ definiert:

$$x^{(j)} = \begin{cases} 1 & \text{falls } x = j, \\ 0 & \text{sonst.} \end{cases}$$

Eine Design-Zeile ist dann gegeben als $\boldsymbol{x} = (1, x^{(2)}, \ldots, x^{(L)})^\top$, wobei maximal eine der Variablen $x^{(j)}$ Eins ist. Der Parametervektor sei $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_L)^\top$ und der lineare Prädiktor ist dann

$$\log\left(\frac{\pi}{1-\pi}\right) = \boldsymbol{x}^\top \boldsymbol{\beta} = \begin{cases} \beta_1 + \beta_j & \text{falls } x^{(j)} = 1 \text{ für ein } j > 1, \\ \beta_1 & \text{falls } x^{(j)} = 0 \text{ für alle } j > 1. \end{cases}$$

Die Odds für $Y = 1$ gegenüber $Y = 0$ sind dann gegeben als

$$\text{odds}(x) = \begin{cases} \exp\{\beta_0\}\exp\{\beta_j\} & \text{für ein } x \text{ aus Kategorie } j = 2, \ldots, L, \\ \exp\{\beta_0\} & \text{für } x \text{ aus der Referenzkategorie 1.} \end{cases}$$

Die Parameter können als Odds-Ratio interpretiert werden. Es sei $\boldsymbol{x}$ eine Design-Zeile mit $x^{(j)} = 1$ für ein $j > 1$ und $\tilde{\boldsymbol{x}}$ eine Design-Zeile mit $\tilde{x}^{(j)} = 0$ für alle $j > 1$. Dann ist

$$\frac{\text{odds}(\boldsymbol{x})}{\text{odds}(\tilde{\boldsymbol{x}})} = \exp\{\beta_j\}.$$

Das Odds-Ratio $\exp\{\beta_j\}$ beschreibt die Odds in der Kategorie j gegenüber den Odds in der Referenzkategorie L.

Falls $\boldsymbol{x}$ eine Design-Zeile mit $x^{(j)} = 1$ für ein $j > 1$ und $\tilde{\boldsymbol{x}}$ eine Design-Zeile mit $\tilde{x}^{(k)} = 1$ für ein $k > 1,\ k \neq j$ ist, dann ist das Odds-Ratio

$$\frac{\text{odds}(\boldsymbol{x})}{\text{odds}(\tilde{\boldsymbol{x}})} = \frac{\exp\{\beta_j\}}{\exp\{\beta_k\}}$$

und es beschreibt die Odds in Kategorie j gegenüber den Odds in Kategorie k. Das logarithmierte Odds-Ratio wird beschrieben durch die Differenz $\beta_j - \beta_k$ der beiden Parameter.

Bemerkung. Natürlich könnte auch jede beliebige andere Stufe des Faktors als Referenzkategorie verwendet werden. In `R` wird aber standardmäßig die erste (alphabetische Reihenfolge) Stufe als Referenzklasse verwendet.

ML-Schätzung für logistische Regressionsmodelle

Es seien $n_i Y_i \overset{ind}{\sim} Bin(n_i, \pi_i)$ mit $\mathbb{E}[Y_i] = \pi_i$ für $i = 1, \ldots, n$ gegeben. Für Beobachtungen y_i ist die gemeinsame Wahrscheinlichkeitsfunktion

$$f(\boldsymbol{y}, \boldsymbol{\pi}) = \prod_{i=1}^{n} \binom{n_i}{n_i y_i} \pi_i^{n_i y_i} (1 - \pi_i)^{n_i - n_i y_i}$$

und als Log-Likelihood-Funktion resultiert

$$l(\boldsymbol{\pi}, \boldsymbol{y}) = \sum_{i=1}^{n} \left\{ n_i y_i \log\left(\frac{\pi_i}{1 - \pi_i}\right) + n_i \log(1 - \pi_i) + \log\binom{n_i}{n_i y_i} \right\}.$$

Es ist $\log\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{x}_i^\top \boldsymbol{\beta}$ und $\log\left(\frac{1}{1-\pi_i}\right) = \log(1 + \exp\{\boldsymbol{x}_i^\top \boldsymbol{\beta}\})$. Damit folgt

$$l(\boldsymbol{\beta}, \boldsymbol{y}) = \sum_{i=1}^{n} \left\{ n_i y_i (\boldsymbol{x}_i^\top \boldsymbol{\beta}) - n_i \log(1 + \exp\{\boldsymbol{x}_i^\top \boldsymbol{\beta}\}) + c(y_i) \right\}.$$

Als Score-Funktionen erhalten wir damit

$$\begin{aligned} \frac{\partial l(\boldsymbol{\beta}, \boldsymbol{y})}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^{n} \left\{ n_i y_i \boldsymbol{x}_i - n_i \frac{\exp\{\boldsymbol{x}_i^\top \boldsymbol{\beta}\} \boldsymbol{x}_i}{1 + \exp\{\boldsymbol{x}_i^\top \boldsymbol{\beta}\}} \right\} \\ &= \sum_{i=1}^{n} (n_i y_i - n_i \pi_i) \boldsymbol{x}_i \end{aligned}$$

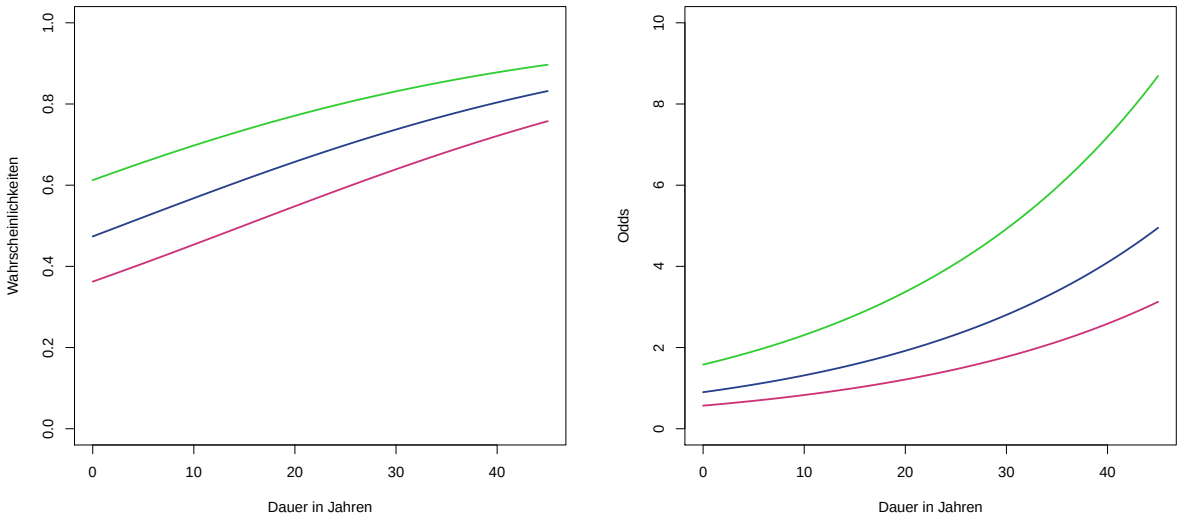
und das Score-Gleichungssystem ist dann

$$\sum_{i=1}^{n} (n_i y_i - n_i \hat{\pi}_i) \boldsymbol{x}_i \overset{!}{=} \mathbf{0}. \tag{3.1.4}$$

Dies hängt durch $\hat{\pi}_i$ von $\hat{\boldsymbol{\beta}}$ ab und ist ein nichtlineares System in $\hat{\boldsymbol{\beta}}$, welches wir iterativ mittels Newton-Raphson lösen.

Binomiale Modelle in `R`

Binomiale Modelle können in `R` mittels der Funktion `glm` geschätzt werden, indem das `family`-Argument auf `binomial` gesetzt wird. Als Prädiktoren verwenden wir der Einfachheit halber zunächst nur die stetige Variable `dau` und den Faktor `oberfl` und nehmen an, dass es keine Interaktion der beiden gibt (Einfluss von `dau` ist auf allen Stufen des Faktors



(a) Wahrscheinlichkeiten (b) Odds

Abbildung 3.4.: Geschätzte Wahrscheinlichkeiten und Odds für das Logit-Modell. Oberfläche des Implantats `glatt` (violett), `Polyurethan` (grün), `texturiert` (blau).

`oberfl` gleich). Die Stufen der Response-Variable `kontr_flag` sind mit {`yes`, `no`} codiert, sodass die Wahrscheinlichkeiten für eine Kapselkontraktur modelliert werden.

```
mod.logit <- glm(kontr_flag ~ dau + oberfl, family=binomial)
```

Das Modell wurde mit `mod.logit` benannt, denn defaultmäßig wird in `R` für die Binomialverteilung der Logit-Link verwendet. Die geschätzten Wahrscheinlichkeiten und Odds für dieses Modell können folgendermaßen berechnet werden:

```
g <- expand.grid(oberfl=c(levels(oberfl)), dau=c(1,5,10,30))
wsk.logit <- predict(mod.logit, g, type='response')
odds <- exp(predict(mod.logit, g, type='link'))
```

Abbildung 3.4 zeigt die vorhergesagten Wahrscheinlichkeiten und Odds für das Logit-Modell für die verschiedenen Stufen des Faktors `oberfl` in Abhängigkeit der Dauer. Die Wahrscheinlichkeiten und Odds steigen prinzipiell mit zunehmender Dauer an und sind bei einer glatten Oberfläche des Implantats am niedrigsten und bei einer Polyurethan-Oberfläche am höchsten.

Ein Probit-Modell kann geschätzt werden, indem das `link`-Argument für die Binomialverteilung auf `probit` gesetzt wird.

```
mod.probit <- glm(kontr_flag ~ dau + oberfl, family=binomial(link=probit))
```

Analog kann man durch setzen des `link`-Arguments auf `cloglog` ein Complementary-Log-Log-Modell schätzen.

```
mod.log <- glm(kontr_flag ~ dau + oberfl, family=binomial(link=cloglog))
```

Tabelle 3.1 zeigt einen Vergleich der geschätzten Wahrscheinlichkeiten für einige Kombinationen der erklärenden Variablen.

oberfl	dau	Logit-Modell	Probit-Modell	Cloglog-Modell
glatt	1	0.371	0.372	0.378
Polyurethan	1	0.621	0.621	0.620
texturiert	1	0.483	0.484	0.488
glatt	5	0.407	0.408	0.408
Polyurethan	5	0.656	0.656	0.656
texturiert	5	0.521	0.521	0.521
glatt	10	0.454	0.454	0.446
Polyurethan	10	0.698	0.698	0.700
texturiert	10	0.568	0.567	0.564
glatt	30	0.639	0.637	0.616
Polyurethan	30	0.831	0.838	0.858
texturiert	30	0.737	0.738	0.740

Tabelle 3.1.: Geschätze Wahrscheinlichkeiten für eine Kapselkontraktur.

3.2. Multinomiale Response-Modelle

In diesem Abschnitt wollen wir das binomiale Regressionsmodell erweitern, sodass auch multinomiale Responses betrachtet werden können. Diese dürfen im Gegensatz zu binomialen Responses, die stets aus einer der zwei generischen Kategorien {`Erfolg`, `Misserfolg`} stammen, aus einer von mehreren Kategorien sein. Anstatt der Binomialverteilung kommen die Responses aus der Multinomialverteilung. Beispiele aus dem Silikonregister sind die Variable `kontr` mit den Kategorien {`BakerI`, `BakerII`, `BakerIII`, `BakerIV`} oder die Variable `oberfl` mit Kategorien {`glatt`, `Polyurethan`, `texturiert`}. Zwischen diesen zwei Variablen gibt es einen wesentlichen Unterschied. Die Kategorien für die Variable `oberfl` sind nicht geordnet (wir können nicht von vornherein sagen, dass eine Oberflächenbeschaffenheit besser ist als die anderen), während die Kategorien für die Variable `kontr` geordnet sind (Probleme werden mit zunehmender Baker-Stufe immer schwerwiegender). Wir unterscheiden daher Modelle für *nominale* Responses und Modelle für *ordinale* Responses. Hauptquelle für diesen Abschnitt ist Agresti (2013) und zusätzliche Quellen werden explizit erwähnt.

3.2.1. Die Multinomialverteilung

Es werden n unabhängige Versuche eines Zufallsexperiments durchgeführt, wobei $\{1, \dots, c\}$ die möglichen Ausgänge sind. Die Wahrscheinlichkeiten für die möglichen Ausgänge seien gleich für alle n Versuche und gegeben als $\boldsymbol{\pi} = (\pi_1, \dots, \pi_c)^\top$, wobei $\sum_{j=1}^{c} \pi_j = 1$ gelte. Es sei Y_j die Anzahl der Versuche mit Ausgang in Kategorie j für $j = 1, \dots, c$. Dann ist der Zufallsvektor $\boldsymbol{Y} = (Y_1, \dots, Y_c)^\top$ multinomialverteilt mit Wahrscheinlichkeitsfunktion

$$f(\boldsymbol{y}, \boldsymbol{\pi}) = \binom{n}{y_1, \dots, y_c} \prod_{j=1}^{c} \pi_j^{y_j},$$

wobei $\sum_{j=1}^{c} y_j = n$. Wir schreiben $\boldsymbol{Y} = (Y_1, \dots, Y_c)^\top \sim M(n, \boldsymbol{\pi})$.

Der Multinomialkoeffizient ist hierbei definiert als

$$\binom{n}{y_1, \dots, y_c} = \frac{n!}{y_1! \cdots y_c!}.$$

Er gibt die Anzahl der Möglichkeiten an, die n Beobachtungen in die c Kategorien einzuteilen, sodass jeweils genau y_j Beobachtungen in Kategorie j sind.

Für $j = 1, \dots, c$ ist $Y_j \sim Bin(n, \pi_j)$ und es gilt daher

$$\mathbb{E}[Y_j] = n\pi_j \quad \text{und} \quad \mathbb{V}\text{ar}[Y_j] = n\pi_j(1 - \pi_j).$$

Des Weiteren gilt für die Kovarianz zwischen Y_j und Y_k für $j \neq k \in \{1, \dots, c\}$

$$\mathbb{C}\text{ov}[Y_j, Y_k] = -n\pi_j\pi_k.$$

Für einen Beweis siehe Anhang A.7.

Bemerkung. Die Binomialverteilung ist ein Spezialfall der Multinomialverteilung mit $c = 2$, denn in diesem Fall ist $y_2 = n - y_1$ und es gilt

$$\binom{n}{y_1, y_2} \pi_1^{y_1} \pi_2^{y_2} = \binom{n}{y_1} \pi_1^{y_1} (1 - \pi_1)^{n - y_1}.$$

3.2.2. Modell für nominale Responses

Die Beobachtungen seien aus einer von c Kategorien, d. h. $Y_i \in \{1, \dots, c\}$ für $i = 1, \dots, n$. Haben die Kategorien dabei keinerlei Ordnung, dann liegen sie auf der sogenannten Nominalskala. Ein Beispiel ist die Variable `fuel` aus dem Silikonregister mit den Kategorien {`Hydrogel`, `Kochsalzlösung`, `Silikongel`, `gemischt`, `andere`}.

Wir assoziieren den Vektor $\boldsymbol{Y}_i = (Y_{i1}, \dots, Y_{ic})^\top$ mit Y_i, wobei dessen Einträge gegeben sind als

$$Y_{ij} = \begin{cases} 1 & \text{falls } Y_i = j, \\ 0 & \text{sonst.} \end{cases}$$

Die dazugehörigen Wahrscheinlichkeiten für die Kategorien seien $\boldsymbol{\pi}_i = (\pi_{i1}, \ldots, \pi_{ic})^\top$. Es gilt also $\sum_{j=1}^{c} Y_{ij} = 1$ und $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{ic})^\top \sim M(1, \boldsymbol{\pi}_i)$.

Falls es Beobachtungen gibt, die die gleiche Design-Zeile haben, können wir die Daten gruppieren. Die Zufallsvariable Y_{ij} ist dann der relative Anteil an Beobachtungen in Kategorie j für die Gruppe i. In der Gruppe i sei dann $n_i \boldsymbol{Y}_i = (n_i Y_{i1} \ldots, n_i Y_{ic})^\top \sim M(n_i, \boldsymbol{\pi}_i)$, wobei $n_i Y_{ij}$ die absolute Häufigkeit der Beobachtungen aus Kategorie j und n_i die Gesamtanzahl der Beobachtungen in dieser Gruppe ist. Es gilt $\mathbb{E}[Y_{ij}] = \pi_{ij}$. Haben alle Beobachtungen verschiedene Design-Zeilen resultiert der Spezialfall mit $n_i = 1$ für alle i.

Geschätzt werden sollen die Wahrscheinlichkeiten für die Kategorien mit fixen Werten $\boldsymbol{x}_i$ für die erklärenden Variablen, also

$$\pi_{ij} = \pi_j(\boldsymbol{x}_i), \quad \text{für } j = 1, \ldots, c-1.$$

Daraus ergibt sich

$$\pi_c(\boldsymbol{x}_i) = \pi_{ic} = 1 - \pi_{i1} - \cdots - \pi_{ic-1}.$$

Ein **multikategorielles Logit-Modell** ist gegeben als

$$\log\left(\frac{\pi_{ij}}{\pi_{ic}}\right) = \boldsymbol{x}_i^\top \boldsymbol{\beta}_j \quad \text{für } j = 1, \ldots, c-1. \tag{3.2.1}$$

Die Kategorie c ist die Referenzkategorie. Alle anderen Kategorien erhalten ihren eigenen Parametervektor $\boldsymbol{\beta}_j = (\beta_{j0}, \ldots, \beta_{j,p-1})^\top$. Es werden die $(c-1)$ sogenannten **Baseline-Category-Logits** modelliert. Diese geben die Log-Odds für $Y_i = j$ gegenüber $Y_i = c$ an. Ein solches Modell nennen wir auch **Baseline-Category-Modell**.

Das Modell (3.2.1) ist äquivalent zu

$$\frac{\pi_{ij}}{\pi_{ic}} = \exp\{\boldsymbol{x}_i^\top \boldsymbol{\beta}_j\}. \tag{3.2.2}$$

Durch Summieren über $l = 1, \ldots, c-1$ erhalten wir

$$\frac{1}{\pi_{ic}} \sum_{l=1}^{c-1} \pi_{il} = \sum_{l=1}^{c-1} \exp\{\boldsymbol{x}_i^\top \boldsymbol{\beta}_l\}.$$

Mit $\sum_{l=1}^{c-1} \pi_{il} = 1 - \pi_{ic}$ folgt

$$\frac{1}{\pi_{ic}}(1 - \pi_{ic}) = \sum_{l=1}^{c-1} \exp\{\boldsymbol{x}_i^\top \boldsymbol{\beta}_l\}.$$

Daraus erhalten wir durch einfache Umformungen

$$\pi_{ic} = \frac{1}{1 + \sum_{l=1}^{c-1} \exp\{\boldsymbol{x}_i^\top \boldsymbol{\beta}_l\}}. \tag{3.2.3}$$

Für $j = 1, \ldots, c-1$ folgt aus (3.2.2)

$$\pi_{ij} = \frac{\exp\{\boldsymbol{x}_i^\top \boldsymbol{\beta}_j\}}{1 + \sum_{l=1}^{c-1} \exp\{\boldsymbol{x}_i^\top \boldsymbol{\beta}_l\}}. \tag{3.2.4}$$

Durch die Baseline-Category-Logits sind die Log-Odds für alle $\binom{c}{2}$ Paare von Kategorien beschrieben, denn für $j, k \in \{1, \ldots, c-1\}$ mit $j \neq k$ ist

$$\log\left(\frac{\pi_{ij}}{\pi_{ik}}\right) = \log\left(\frac{\pi_{ij}/\pi_{ic}}{\pi_{ik}/\pi_{ic}}\right) = \log\left(\frac{\pi_{ij}}{\pi_{ic}}\right) - \log\left(\frac{\pi_{ik}}{\pi_{ic}}\right).$$

Bemerkung. Für den binomialen Fall mit zwei Kategorien für die Response-Variable und einem linearen Prädiktor der Form

$$\boldsymbol{x}^\top \boldsymbol{\beta}_j = \beta_{j0} + \beta_{j1} x$$

mit einer stetigen Prädiktorvariable x haben wir gesehen, dass die Wahrscheinlichkeiten $\pi_j(\boldsymbol{x})$ in Abhängigkeit des Vorzeichens von β_{j1} monoton fallend oder steigend in x_i sind (siehe Abschnitt 3.1). Dies muss für multikategorielle Responses nicht der Fall sein, wie nachfolgende Überlegung zeigt.

Es sei $c = 3$. Dann ist laut (3.2.4)

$$\pi_j(\boldsymbol{x}) = \frac{\exp\{\boldsymbol{x}^\top \boldsymbol{\beta}_j\}}{1 + \exp\{\boldsymbol{x}^\top \boldsymbol{\beta}_1\} + \exp\{\boldsymbol{x}^\top \boldsymbol{\beta}_2\}} \quad \text{für } j = 1, 2.$$

und mit (3.2.3)

$$\pi_3(\boldsymbol{x}) = \frac{1}{1 + \exp\{\boldsymbol{x}^\top \boldsymbol{\beta}_1\} + \exp\{\boldsymbol{x}^\top \boldsymbol{\beta}_2\}}.$$

Damit folgt

$$\frac{\partial \pi_3(\boldsymbol{x})}{\partial x} = -\frac{\exp\{\boldsymbol{x}^\top \boldsymbol{\beta}_1\}\beta_{11} + \exp\{\boldsymbol{x}^\top \boldsymbol{\beta}_2\}\beta_{21}}{(1 + \exp\{\boldsymbol{x}^\top \boldsymbol{\beta}_1\} + \exp\{\boldsymbol{x}^\top \boldsymbol{\beta}_2\})^2}.$$

Falls β_{11} und β_{21} beide positiv sind, ist dieser Ausdruck negativ und $\pi_3(\boldsymbol{x})$ ist monoton fallend in x. Sind beide negativ, ist $\pi_3(\boldsymbol{x})$ monoton steigend. Haben β_{11} und β_{21} jedoch unterschiedliches Vorzeichen, dann ist $\pi_3(\boldsymbol{x})$ nicht monoton. Abbildung 3.5 veranschaulicht dies mithilfe von gewählten Werten für die Parameter für den Fall von unterschiedlichen Vorzeichen für β_{11} und β_{21}.

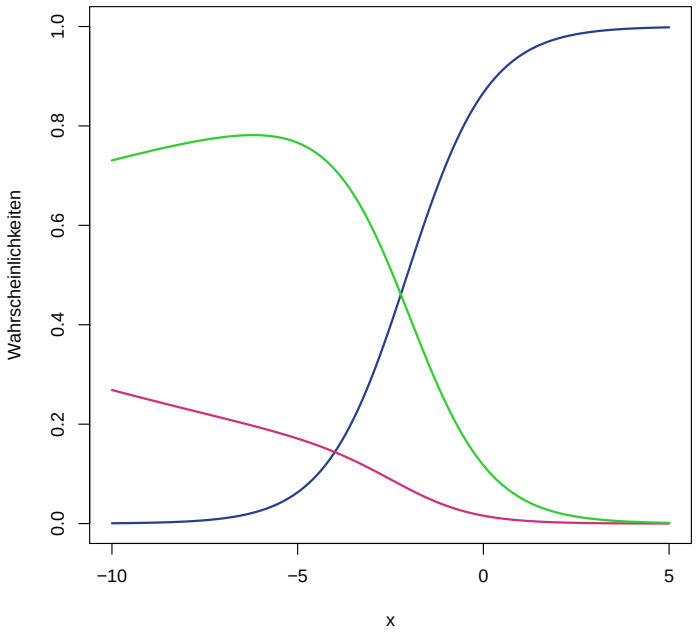


Abbildung 3.5.: Wahrscheinlichkeiten für ein multikategorielles Logit-Modell mit $\beta_{10} = 2$, $\beta_{11} = 0.9$ und $\beta_{20} = -2$, $\beta_{21} = -0.1$, $\pi_1(\boldsymbol{x})$ (blau), $\pi_2(\boldsymbol{x})$ (violett) und $\pi_3(\boldsymbol{x})$ (grün).

Mulitkategorielles Logit-Modell in R

Wir schätzen ein multikategorielles Logit-Modell für die Variable `kontr` aus dem Silikonregister. Dabei ignorieren wir zunächst die Ordnung der Kategorien, sodass wir eine nominale Response-Variable haben. Man kann für diesen Zweck das R-Package `VGAM` zum Schätzen von Vektor-generalisierten linearen und additiven Modellen verwenden. Als Prädiktoren werden wieder nur die stetige Variable `dau` und der Faktor `oberfl` mit den Stufen {`glatt`, `Polyurethan`, `texturiert` } (Abk. `Pol`, `text`) verwendet. Die Schätzung erfolgt durch den Aufruf der Funktion `vglm`.

```
library(VGAM)
mod.logit <- vglm(kontr ~ dau + oberfl, family=multinomial)
```

Die Baseline-Category ist `BakerIV`. Der Befehl `coef(mod.logit)` liefert uns die Schätzungen für die Parametervektoren für die Baker-Stufen 1-3, welche für unser Modell gegeben sind als

$$\hat{\boldsymbol{\beta}}_1 = (1.57, -0.11, -1.20, -1.14)^\top,$$
$$\hat{\boldsymbol{\beta}}_2 = (0.78, -0.12, 0.17, 0.51)^\top,$$
$$\hat{\boldsymbol{\beta}}_3 = (0.04, -0.08, 0.38, 1.45)^\top.$$

Es sei $\hat{\pi}_j$ die geschätzte Wahrscheinlichkeit, dass eine Patientin mit Werten $\boldsymbol{x}$ für die erklärenden Variablen eine Kapselkontraktur der Baker-Stufe j hat. Dann sind die Baseline-Category-Logits

$$\log\left(\frac{\hat{\pi}_1}{\hat{\pi}_4}\right) = \boldsymbol{x}^\top\boldsymbol{\beta}_1 = 1.57 - 0.11\ \texttt{dau} - 1.20\ (\texttt{oberfl} = \texttt{Pol}) - 1.14\ (\texttt{oberfl} = \texttt{text})$$

$$\log\left(\frac{\hat{\pi}_2}{\hat{\pi}_4}\right) = \boldsymbol{x}^\top\boldsymbol{\beta}_2 = 0.78 - 0.12\ \texttt{dau} + 0.17\ (\texttt{oberfl} = \texttt{Pol}) + 0.51\ (\texttt{oberfl} = \texttt{text})$$

$$\log\left(\frac{\hat{\pi}_3}{\hat{\pi}_4}\right) = \boldsymbol{x}^\top\boldsymbol{\beta}_3 = 0.04 - 0.08\ \texttt{dau} + 0.38\ (\texttt{oberfl} = \texttt{Poly}) + 1.45\ (\texttt{oberfl} = \texttt{text})$$

Die Baseline-Category-Logits und Odds für eine Kapselkontraktur der Baker-Stufe j gegenüber Baker-Stufe 4 für einige Kombinationen der erklärenden Variablen können wir in `R` folgendermaßen berechnen:

```
g <- expand.grid(oberfl=c(levels(oberfl)), dau=c(1,5,10,30))
baseline.logits <- predict(mod.logit, newdata=g, type='link')
odds <- exp(baseline.logits)
```

Dabei gibt `odds` die Chancen an, dass eine Patientin eine Kapselkontraktur der Baker-Stufe j anstatt der Baker-Stufe 4 hat. Die Wahrscheinlichkeiten für die verschiedenen Baker-Stufen sind

```
wsk <- predict(mod.logit, newdata=g, type='response')
```

Das Odds-Ratio für die Chance einer Kapselkontraktur in Baker-Stufe 1 gegenüber Baker-Stufe 4 wenn die Dauer um ein Jahr erhöht wird ist für eine glatte Implantatoberfläche gegeben als

$$\frac{\exp\{1.57\}\exp\{-0.11\}^{\texttt{dau}+1}}{\exp\{1.57\}\exp\{-0.11\}^{\texttt{dau}}} = \exp\{-0.11\} \approx 0.90.$$

Die Odds für Baker-Stufe 1 gegenüber Baker-Stufe 4 verringern sich pro Jahr um 10 %. Bei beliebig gegebener Dauer ist das Odds-Ratio für eine glatte Oberfläche gegenüber einer texturierten Oberfläche

$$\frac{\exp\{1.57\}\exp\{-0.11\}^{\texttt{dau}}}{\exp\{1.57\}\exp\{-0.11\}^{\texttt{dau}}\exp\{-1.14\}} = \frac{1}{\exp\{-1.14\}} \approx 3.$$

Die Odds für eine Kapselkontraktur der Baker-Stufe 1 gegenüber Baker-Stufe 4 sind bei glatter Oberfläche etwa dreimal so groß wie bei einer texturierten Oberfläche. Wir können die Parameter (bzw. die Exponenten der Parameter) also wie bei binomialen Modellen als spezielle Odds-Ratios interpretieren.

Alternativ zur `vglm`-Funktion kann man auch die Funktion `multinom` aus dem `nnet`-Package zum Schätzen von multikategoriellen Logit-Modellen verwenden.

```
library(nnet)
mod.nnet <- multinom(kontr ~ dau + oberfl)
```

Achtung, hier ist die Baseline-Category im Unterschied zu vorher `BakerI` und wir erhalten Schätzungen für die Parametervektoren der Baker-Stufen 2-4. Die geschätzten Wahrscheinlichkeiten stimmen aber natürlich mit den von der Funktion `vglm` geschätzten Wahrscheinlichkeiten überein.

3.2.3. ML-Schätzung für multikategorielle logistische Regressionsmodelle

Die i-te Beobachtung $n_i\boldsymbol{Y}_i$ sei multinomialverteilt, d. h. $n_i\boldsymbol{Y}_i = (n_iY_{i1},\ldots,n_iY_{ic})^\top \overset{ind}{\sim} M(n_i,\boldsymbol{\pi}_i)$ für $i = 1,\ldots,n$. Dann ist die Wahrscheinlichkeitsfunktion gegeben als

$$f(\boldsymbol{y}_i,\boldsymbol{\pi}_i) = \binom{n_i}{n_iy_{i1},\ldots,n_iy_{ic}} \prod_{j=1}^{c} \pi_{ij}^{n_iy_{ij}}.$$

Die Log-Likelihood-Funktion ist dann

$$\begin{aligned} l(\boldsymbol{\pi}_i,\boldsymbol{y}_i) &= \log\binom{n_i}{n_iy_{i1},\ldots,n_iy_{ic}} + \sum_{j=1}^{c} n_iy_{ij}\log(\pi_{ij}) \\ &= c(\boldsymbol{y}_i) + \sum_{j=1}^{c-1} n_iy_{ij}\log(\pi_{ij}) + \left(n_i - \sum_{j=1}^{c-1} n_iy_{ij}\right)\log(\pi_{ic}) \\ &= c(\boldsymbol{y}_i) + \sum_{j=1}^{c-1} n_iy_{ij}\log\left(\frac{\pi_{ij}}{\pi_{ic}}\right) + n_i\log(\pi_{ic}), \end{aligned} \tag{3.2.5}$$

wobei der Term $c(\boldsymbol{y}_i)$ unabhängig von $\boldsymbol{\pi}_i$ ist und deshalb vernachlässigt werden kann. Die Log-Likelihood der gesamten Stichprobe ist dann

$$l(\boldsymbol{\pi}, \boldsymbol{y}) = \sum_{i=1}^{n} \left\{ \sum_{j=1}^{c-1} n_i y_{ij} \log\left(\frac{\pi_{ij}}{\pi_{ic}}\right) + n_i \log(\pi_{ic}) \right\}.$$

Mit (3.2.1) und (3.2.3) folgt als Log-Likelihood

$$l(\boldsymbol{\pi}, \boldsymbol{y}) = \sum_{i=1}^{n} \left\{ \sum_{j=1}^{c-1} n_i y_{ij} (\boldsymbol{x}_i^\top \boldsymbol{\beta}_j) - n_i \log\left(1 + \sum_{j=1}^{c-1} \exp\{\boldsymbol{x}_i^\top \boldsymbol{\beta}_j\}\right) \right\}.$$

Die ML-Schätzer $\hat{\boldsymbol{\beta}}_j$ für die Parametervektoren $\boldsymbol{\beta}_j$ werden wieder mittels Newton-Raphson berechnet.

Bemerkung. Multikategorielle logistische Modelle können auch als multivariate GLMs für Verteilungen aus der multivariaten linearen Exponentialfamilie gesehen werden, vgl. Tutz (2012). Diese haben eine Dichte der Form

$$f(\boldsymbol{y}_i, \boldsymbol{\theta}_i) = \exp\left\{ \frac{\boldsymbol{y}_i^\top \boldsymbol{\theta}_i - b(\boldsymbol{\theta}_i)}{w_i \phi} + c(\boldsymbol{y}_i, \phi) \right\}.$$

Dabei sind nun $\boldsymbol{y}_i$ und $\boldsymbol{\theta}_i$ Vektoren, w_i ein bekanntes Gewicht und ϕ ein Dispersionsparameter, den wir als bekannt annehmen.

Wir zeigen zuerst, dass die Multinomialverteilung zu dieser Exponentialfamilie gehört. Es sei $n_i \boldsymbol{Y}_i = (n_i Y_{i1}, \ldots, n_i Y_{ic})^\top \sim M(n_i, \boldsymbol{\pi}_i)$ und $\boldsymbol{\theta}_i = (\theta_{i1}, \ldots, \theta_{ic})^\top$, dann ist die Log-Likelihood einer Beobachtung wie in (3.2.5)

$$l(\boldsymbol{\pi}_i, \boldsymbol{y}_i) = \sum_{j=1}^{c-1} n_i y_{ij} \log\left(\frac{\pi_{ij}}{\pi_{ic}}\right) + n_i \log(\pi_{ic}) + \log\binom{n_i}{n_i y_{i1}, \ldots, n_i y_{ic}}.$$

Mit $\boldsymbol{\theta}_i = (\log(\frac{\pi_{i1}}{\pi_{ic}}), \ldots, \log(\frac{\pi_{i,c-1}}{\pi_{ic}}), 0)^\top$, $b(\boldsymbol{\theta}_i) = -\log(\pi_{ic})$, $\phi = 1$, $w_i = \frac{1}{n_i}$ und $c(\boldsymbol{y}_i, \phi) = \log\binom{n_i}{n_i y_{i1}, \ldots, n_i y_{ic}}$ ist dies eine multivariate lineare Exponentialfamilie. Weil $\pi_{ic} = 1 - \sum_{j=1}^{c-1} \pi_{ij}$ gilt, hat diese Exponentialfamilie $(c-1)$ frei schätzbare Parameter.

Für einen Response-Vektor $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{ic})^\top$ wofür wir $\boldsymbol{\mu}_i = \mathbb{E}[\boldsymbol{Y}_i]$ schätzen wollen, ist ein multivariates Modell gegeben als

$$g(\boldsymbol{\mu}_i) = \boldsymbol{X}_i \boldsymbol{\beta}, \tag{3.2.6}$$

wobei hier $g(\cdot)$ eine vektorwertige Funktion und $\boldsymbol{X}_i$ die Design-Matrix zur Beobachtung i ist, deren j-te Zeile die Prädiktorvariablen für Y_{ij} enthält.

Wir zeigen nun noch, dass das multikategorielle logistische Modell (3.2.1) in dieses Schema passt. Es ist $n_i \boldsymbol{Y}_i = (n_i Y_{i1}, \ldots, n_i Y_{ic-1})^\top \sim M(n_i, \boldsymbol{\pi}_i)$ für $i = 1, \ldots, n$. Wir setzen $\boldsymbol{\mu}_i$, $g(\cdot)$, $\boldsymbol{X}_i$ und $\boldsymbol{\beta}$ beim multivariaten Modell folgendermaßen:

$$\boldsymbol{\mu}_i = \boldsymbol{\pi}_i = (\pi_{i1}, \ldots, \pi_{ic})^\top,$$

$$g(\boldsymbol{\mu}_i) = \left(\log\left(\frac{\pi_{i1}}{\pi_{ic}}\right), \ldots, \log\left(\frac{\pi_{i,c-1}}{\pi_{ic}}\right), 0 \right)^\top,$$

$$\boldsymbol{X}_i = \begin{pmatrix} \boldsymbol{x}_i^\top & 0 & \ldots & 0 \\ 0 & \boldsymbol{x}_i^\top & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & \boldsymbol{x}_i^\top \\ 0 & \ldots & \ldots & 0 \end{pmatrix}$$

und

$$\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \ldots, \boldsymbol{\beta}_{c-1}^\top)^\top,$$

wobei $\boldsymbol{\beta}_j$ der Parametervektor für die j-te Kategorie ist. Damit ist

$$\boldsymbol{X}_i\boldsymbol{\beta} = (\boldsymbol{x}_i^\top\boldsymbol{\beta}_1, \ldots, \boldsymbol{x}_i^\top\boldsymbol{\beta}_{c-1}, 0)^\top$$

und laut dem multivariaten Modell (3.2.6) gilt dann

$$\log\left(\frac{\pi_{ij}}{\pi_{ic}}\right) = \boldsymbol{x}_i^\top\boldsymbol{\beta}_j \quad \text{für } j = 1, \ldots, c-1,$$

was dem Modell (3.2.1) entspricht.

3.2.4. Modelle für ordinale Responses

Sind die c Kategorien für eine Beobachtung Y mit einer Ordnung versehen, liegt diese auf der Ordinalskala. Die Abstände zwischen den Kategorien sind dabei nicht genau quantifiziert. Es gibt verschiedene Ansätze, die Ordnung der Kategorien im Modell zu berücksichtigen. Vergleiche für diesen Abschnitt auch Tutz (2012, Kapitel 9).

Kumulative Modelle

Es sei $Y \in \{1, \ldots, c\}$ mit zugehöriger Design-Zeile $\boldsymbol{x}$. Ein Ansatz, die Ordnung der Kategorien zu berücksichtigen, ist die Betrachtung von **kumulativen Wahrscheinlichkeiten**

$$\mathbb{P}[Y \leq j|\boldsymbol{x}] = \pi_1(\boldsymbol{x}) + \cdots + \pi_j(\boldsymbol{x}).$$

Es folgt sofort

$$\mathbb{P}[Y \leq j|\boldsymbol{x}] \leq \mathbb{P}[Y_i \leq k|\boldsymbol{x}], \quad \text{für } j < k.$$

Die **kumulativen Logits** sind definiert als

$$\begin{aligned}\text{logit}(\mathbb{P}[Y \leq j|\boldsymbol{x}]) &= \log\left(\frac{\mathbb{P}[Y \leq j|\boldsymbol{x}]}{1-\mathbb{P}[Y \leq j|\boldsymbol{x}]}\right) \\ &= \log\left(\frac{\pi_1(\boldsymbol{x})+\cdots+\pi_j(\boldsymbol{x})}{\pi_{j+1}(\boldsymbol{x})+\cdots+\pi_c(\boldsymbol{x})}\right), \quad \text{für } j = 1,\ldots,c-1.\end{aligned}$$

Das **kumulative Logit-Modell** hat die Form

$$\text{logit}(\mathbb{P}[Y \leq j|\boldsymbol{x}]) = \alpha_j + \boldsymbol{x}^\top\boldsymbol{\beta} \quad \text{für } j = 1,\ldots,c-1. \tag{3.2.7}$$

Jeder der $c-1$ Logits hat in diesem Modell einen eigenen Intercept-Parameter, aber der Parametervektor $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_{p-1})^\top$ bleibt gleich für alle Logits. Modell (3.2.7) kann auch geschrieben werden als

$$\mathbb{P}[Y \leq j|\boldsymbol{x}] = \frac{\exp\{\alpha_j + \boldsymbol{x}^\top\boldsymbol{\beta}\}}{1+\exp\{\alpha_j + \boldsymbol{x}^\top\boldsymbol{\beta}\}}.$$

Abbildung 3.6 zeigt die kumulativen Wahrscheinlichkeiten eines Modells für eine Response-Variable mit $c = 4$ Kategorien und nur einer stetigen erklärenden Variable x. Die Parameter sind dabei $\alpha_1 = -1$, $\alpha_2 = 1$, $\alpha_3 = 4$ und $\beta = -2$. Alle drei Kurven haben dieselbe Form, da β gleich ist für alle $c-1$ Logits. Sie sind aufgrund der verschiedenen Intercepts nur horizontal versetzt.

In Abbildung 3.7 sind die dazugehörigen Wahrscheinlichkeiten für die einzelnen Kategorien dargestellt. Dabei gilt

$$\begin{aligned}&\pi_1(\boldsymbol{x}) = \mathbb{P}[Y \leq 1|\boldsymbol{x}], \\ &\pi_j(\boldsymbol{x}) = \mathbb{P}[Y \leq j|\boldsymbol{x}] - \mathbb{P}[Y_i \leq j-1|\boldsymbol{x}] \quad \text{für } j = 2,\ldots,c-1 \quad \text{und} \\ &\pi_c(\boldsymbol{x}) = 1 - \mathbb{P}[Y \leq c-1|\boldsymbol{x}].\end{aligned}$$

Eigenschaften des kumulativen Logit-Modells

Für die Intercept-Parameter muss

$$\alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_{c-1}$$

gelten. Dies folgt sofort aus der Monotonie der kumulativen Wahrscheinlichkeiten und aus der Monotonie der Logit-Funktion.

Es seien $\hat{\boldsymbol{x}}$ und $\tilde{\boldsymbol{x}}$ zwei Vektoren mit verschiedenen Werten für die Prädiktoren. Das Odds-Ratio von kumulativen Wahrscheinlichkeiten

$$\frac{\mathbb{P}[Y \leq j|\hat{\boldsymbol{x}}]/(1-\mathbb{P}[Y \leq j|\hat{\boldsymbol{x}}])}{\mathbb{P}[Y \leq j|\tilde{\boldsymbol{x}}]/(1-\mathbb{P}[Y \leq j|\tilde{\boldsymbol{x}}])}$$

heißt **kumulatives Odds-Ratio**. Das logarithmierte kumulative Odds-Ratio ist proportional zur Differenz $(\hat{\boldsymbol{x}} - \tilde{\boldsymbol{x}})$, denn

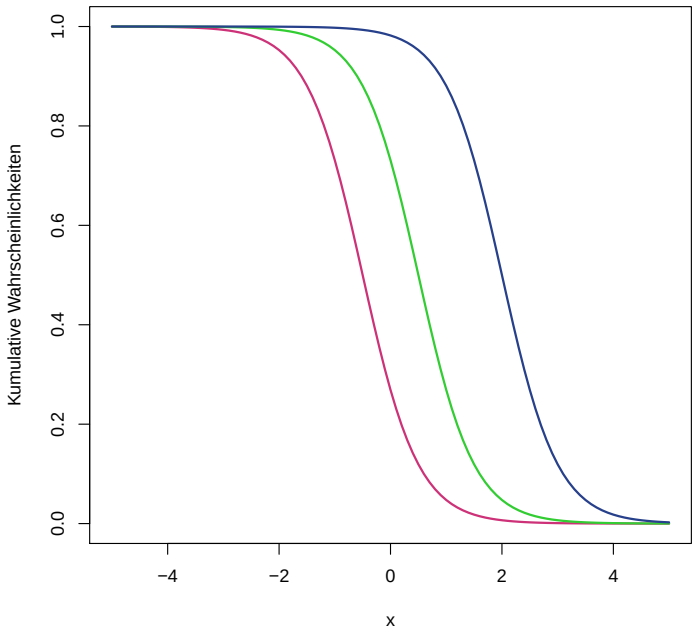


Abbildung 3.6.: Kumulative Wahrscheinlichkeiten $\mathbb{P}[Y_i \leq 1|\boldsymbol{x}]$ (violett), $\mathbb{P}[Y_i \leq 2|\boldsymbol{x}]$ (grün), $\mathbb{P}[Y_i \leq 3|\boldsymbol{x}]$ (blau).

$$\begin{aligned}\log\left(\frac{\mathbb{P}[Y \leq j|\hat{\boldsymbol{x}}]/(1-\mathbb{P}[Y \leq j|\hat{\boldsymbol{x}}])}{\mathbb{P}[Y \leq j|\tilde{\boldsymbol{x}}]/(1-\mathbb{P}[Y \leq j|\tilde{\boldsymbol{x}}])}\right) &= \text{logit}(\mathbb{P}[Y \leq j|\hat{\boldsymbol{x}}]) - \text{logit}(\mathbb{P}[Y \leq j|\tilde{\boldsymbol{x}}]) \\ &= \alpha_j + \hat{\boldsymbol{x}}^\top\boldsymbol{\beta} - (\alpha_j + \tilde{\boldsymbol{x}}^\top\boldsymbol{\beta}) \\ &= (\hat{\boldsymbol{x}} - \tilde{\boldsymbol{x}})^\top\boldsymbol{\beta}.\end{aligned}$$

Die Odds für $Y \leq j$ bei $\hat{\boldsymbol{x}}$ sind also $\exp\{(\hat{\boldsymbol{x}} - \tilde{\boldsymbol{x}})^\top\boldsymbol{\beta})\}$-mal den Odds für $Y \leq j$ bei $\tilde{\boldsymbol{x}}$. Der Faktor $\exp\{(\hat{\boldsymbol{x}} - \tilde{\boldsymbol{x}})^\top\boldsymbol{\beta})\}$ ist unabhängig von j und somit gleich für alle Kategorien. Deshalb nennt man ein solches Modell auch **Proportional-Odds-Modell**. Bei einem Modell mit einer stetigen erklärenden Variable x kann $\exp\{\beta\}$ als spezielles kumulatives Odds-Ratio interpretiert werden. Es werden die Odds für $Y \leq j$ für $x = \hat{x}$ mit den Odds für $x = \hat{x} + 1$ verglichen.

Des Weiteren kann ein Modell für $\text{logit}(\mathbb{P}[Y \leq j|\boldsymbol{x}])$ für ein fixes j als gewöhnliches logistisches Modell für binäre Responses gesehen werden, indem man Beobachtungen in den Kategorien $\{1, \ldots, j\}$ als Erfolg und Beobachtungen in den Kategorien $\{j+1, \ldots, c\}$ als Misserfolg sieht. Das Schätzen der unbekannten Erfolgswahrscheinlichkeit entspricht dann dem Schätzen der kumulativen Wahrscheinlichkeit $\mathbb{P}[Y \leq j|\boldsymbol{x}]$.

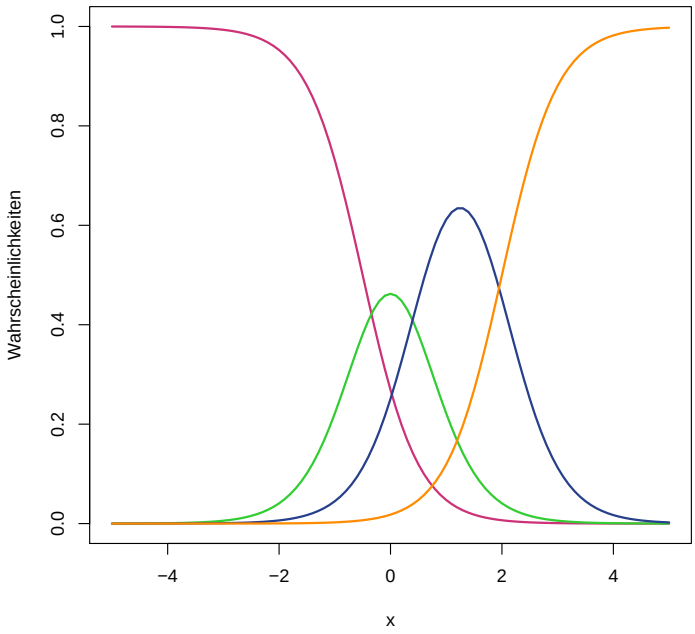


Abbildung 3.7.: Wahrscheinlichkeiten für die Kategorien $\pi_1(\boldsymbol{x})$ (violett), $\pi_2(\boldsymbol{x})$ (grün), $\pi_3(\boldsymbol{x})$ (blau) und $\pi_4(\boldsymbol{x})$ (orange).

Alternative Motivation des kumulativen Logit-Modells

Eine alternative Herleitung des kumulativen Logit-Modells erfolgt über Schwellenwerte, vgl. Fahrmeir et al. (2013, S. 334). Wir nehmen dabei an, dass der Beobachtung Y eine nicht beobachtbare (latente) Zufallsvariable U zugrunde liegt. Diese sei gegeben als

$$U = -\boldsymbol{x}^\top \boldsymbol{\beta} + \varepsilon.$$

Dabei enthält $\boldsymbol{x}$ gegebene Werte für die erklärenden Variablen, $\boldsymbol{\beta}$ ist der Parametervektor und ε ein Zufallsfehler mit Verteilungsfunktion F. Es besteht folgender Zusammenhang zwischen Y und U:

$$Y = j \quad \Longleftrightarrow \quad \alpha_{j-1} < U \leq \alpha_j \quad \text{für } j = 1, \ldots, c,$$

wobei $-\infty = \alpha_0 < \alpha_1 < \cdots < \alpha_c = \infty$ latente, geordnete Schwellenwerte („**thresholds**“) sind. Damit folgt

$$\begin{aligned}\mathbb{P}[Y \leq j|\boldsymbol{x}] &= \mathbb{P}[U \leq \alpha_j] \\ &= \mathbb{P}[-\boldsymbol{x}^\top\boldsymbol{\beta} + \varepsilon \leq \alpha_j] \\ &= \mathbb{P}[\varepsilon \leq \alpha_j + \boldsymbol{x}^\top\boldsymbol{\beta}] \\ &= F(\alpha_j + \boldsymbol{x}^\top\boldsymbol{\beta}) \quad \text{für } j = 1, \ldots, c-1.\end{aligned} \tag{3.2.8}$$

Dies ist ein kumulatives Modell mit Werten $\boldsymbol{x}$ für die erklärenden Variablen und Parametern $\alpha_1, \ldots, \alpha_{c-1}$ und $\boldsymbol{\beta}$. Nehmen wir für die ε die logistische Verteilung $L(0,1)$ an, so resultiert das kumulative Logit-Modell, denn dann folgt

$$\mathbb{P}[Y \leq j|\boldsymbol{x}] = \frac{\exp\{\alpha_j + \boldsymbol{x}^\top\boldsymbol{\beta}\}}{1 + \exp\{\alpha_j + \boldsymbol{x}^\top\boldsymbol{\beta}\}}.$$

Dieser Sachverhalt ist in Abbildung 3.8 für ein Modell mit einer stetigen erklärenden Variable x veranschaulicht. Die Schwellen wurden mit $\alpha_1 = -1$, $\alpha_2 = 1$, $\alpha_3 = 4$ gewählt und sind als horizontale punktierte Linien eingezeichnet. Es sei $\beta = -2$. Für die Werte $x = 0, 1, 2$ ist die Dichte der logistischen Verteilung der zugrunde liegenden latenten Zufallsvariable U gezeichnet. Die schraffierten Bereiche unter den Kurven entsprechen den Wahrscheinlichkeiten $\pi_j(x)$. Man vergleiche dies auch mit den Wahrscheinlichkeiten in der Abbildung 3.7.

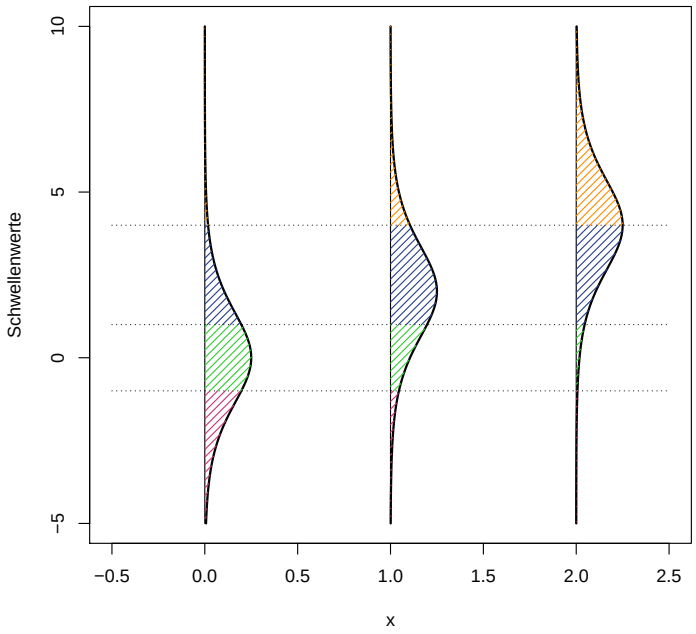


Abbildung 3.8.: Logistische Dichtefunktion der latenten Variable für $x = 0$, $x = 1$ und $x = 2$. Schraffierte Bereiche: $\pi_1(x)$ (violett), $\pi_2(x)$ (grün), $\pi_3(x)$ (blau), $\pi_4(x)$ (orange).

Bemerkung. Natürlich ist es auch möglich andere Verteilungen für den Zufallsfehler ε zu wählen. Verwendet man die Standardnormalverteilung, d. h. $\varepsilon \sim N(0,1)$, resultiert das **kumulative Probit-Modell**,

$$\mathbb{P}[Y \leq j|\boldsymbol{x}] = \Phi(\alpha_j + \boldsymbol{x}^\top \boldsymbol{\beta}).$$

Das kumulative Probit-Modell liefert ähnliche geschätzte Wahrscheinlichkeiten wie das kumulative Logit-Modell. Die Standardfehler der Parameter beim Probit-Modell sind kleiner als beim Logit-Modell. Dies liegt an der geringeren Standardabweichung der Normalverteilung $N(0,1)$ verglichen mit der logistischen Verteilung $L(0,1)$, zu sehen in Abbildung 3.9.

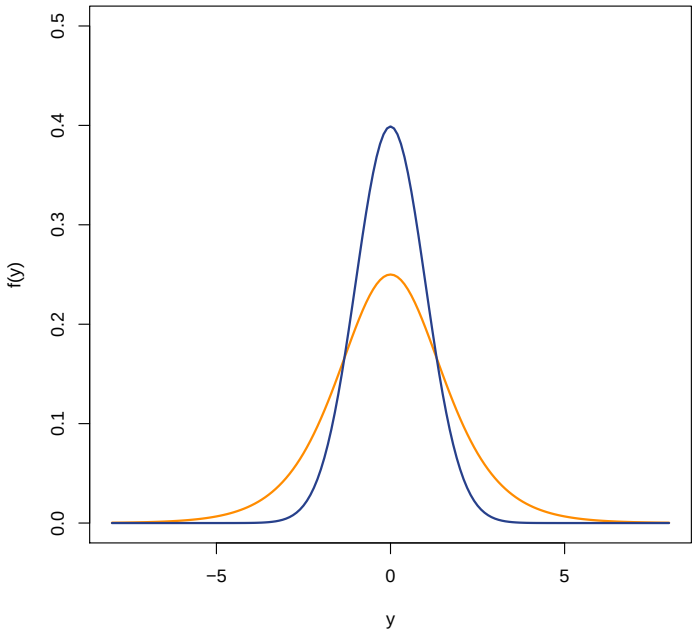


Abbildung 3.9.: Dichtefunktion der Standardnormalverteilung (blau) und der standard-logistischen Verteilung (orange).

In der Praxis findet auch die Gumbel-Verteilung Verwendung. Diese führt uns zum **kumulativen Complementary-Log-Log-Modell**

$$\mathbb{P}[Y \leq j|\boldsymbol{x}] = 1 - \exp\{-\exp\{\alpha_j + \boldsymbol{x}^\top \boldsymbol{\beta}\}\},$$

oder äquivalent

$$\log(-\log(\mathbb{P}[Y > j|\boldsymbol{x}])) = \alpha_j + \boldsymbol{x}^\top \boldsymbol{\beta}.$$

Es gilt die Eigenschaft

$$1 - \mathbb{P}[Y \leq j|\hat{\boldsymbol{x}}] = (1 - \mathbb{P}[Y \leq j|\tilde{\boldsymbol{x}})^{\exp\{(\hat{\boldsymbol{x}}-\tilde{\boldsymbol{x}})^\top \boldsymbol{\beta}\}},$$

denn

$$\begin{aligned}
(1 - \mathbb{P}[Y \leq j|\tilde{\boldsymbol{x}}])^{\exp\{(\hat{\boldsymbol{x}}-\tilde{\boldsymbol{x}})^\top \boldsymbol{\beta}\}} &= \exp\{-\exp\{\alpha_j + \tilde{\boldsymbol{x}}^\top \boldsymbol{\beta}\}\}^{\exp\{(\hat{\boldsymbol{x}}-\tilde{\boldsymbol{x}})^\top \boldsymbol{\beta}\}} \\
&= \exp\{-\exp\{\alpha_j + \tilde{\boldsymbol{x}}^\top \boldsymbol{\beta}\} \exp\{(\hat{\boldsymbol{x}} - \tilde{\boldsymbol{x}})^\top \boldsymbol{\beta}\}\} \\
&= \exp\{-\exp\{\alpha_j + \tilde{\boldsymbol{x}}^\top \boldsymbol{\beta} + \hat{\boldsymbol{x}}^\top \boldsymbol{\beta} - \tilde{\boldsymbol{x}}^\top \boldsymbol{\beta}\}\} \\
&= \exp\{-\exp\{\alpha_j + \hat{\boldsymbol{x}}^\top \boldsymbol{\beta}\}\} \\
&= (1 - \mathbb{P}[Y \leq j|\hat{\boldsymbol{x}}]).
\end{aligned}$$

Kumulative Modelle in Proportional-Odds-Form sind gegeben durch (3.2.8). Sie können verallgemeinert werden zu Modellen mit einem eigenen Effekt β_j für jede Kategorie, d. h.

$$\mathbb{P}[Y \leq j|\boldsymbol{x}] = F(\alpha_j + \boldsymbol{x}^\top \boldsymbol{\beta}_j), \text{ für } j = 1, \ldots, c-1.$$

Die Kurven $F(\alpha_j + \boldsymbol{x}^\top \boldsymbol{\beta}_j)$ sind dann nicht mehr parallel und können sich auch schneiden. Dann ist aber die Ordnung der kumulativen Wahrscheinlichkeiten verletzt. Mittels eines Hypothesentests können wir prüfen, ob die Proportional-Odds-Annahme gerechtfertigt ist. Die dazugehörigen Hypothesen haben folgende Form

$$H_0 : \boldsymbol{\beta}_1 = \cdots = \boldsymbol{\beta}_{c-1} \quad \text{vs.} \quad H_1 : \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{c-1} \text{ beliebig.}$$

Beispiel. Zum Schätzen eines kumulativen Logit-Modells in `R` kann wieder die Funktion `vglm` aus dem `VGAM`-Package verwendet werden. Als Prädiktoren definieren wir zunächst wieder nur `dau` und `oberfl`. Durch Setzen des `family`-Arguments auf `cumulative` erhalten wir ein kumulatives Modell. Zusätzlich setzen wir noch `parallel=T`, um ein Modell in Proportional-Odds-Form zu erhalten. Die Response-Variable wird als geordneter Faktor `ord.kontr` übergeben, welcher eine Ordnung für die Levels des Faktors `kontr` aus dem Silikonregister enthält.

```
ord.kontr <- ordered(kontr)
mod.kum <- vglm(ord.kontr ~ dau + oberfl, family=cumulative(parallel=T))
coef(mod.kum)
```

Als Schätzungen für die Schwellenwerte resultieren

$$\hat{\alpha}_1 = 0.00, \quad \hat{\alpha}_2 = 1.51, \quad \hat{\alpha}_3 = 3.40$$

und als Schätzung für den Parametervektor ergibt sich

$$\hat{\boldsymbol{\beta}} = (-0.08, -1.28, -1.38)^\top.$$

Die Intercept-Parameter sind monoton steigend und es gibt einen globalen Parameterschätzer $\hat{\boldsymbol{\beta}}$ für alle Kategorien. Standardmäßig wird der Logit-Link verwenden. Durch setzen des `link`-Arguments können auch Probit-Link oder Complementary-Log-Log-Link ausgewählt werden.

```
vglm(ord.kontr ~ dau + oberfl, family=cumulative(link=probit, parallel=T))
```

```
vglm(ord.kontr ~ dau + oberfl, family=cumulative(link=cloglog, parallel=T))
```

Alternativ kann auch die Funktion `polr` aus der `MASS`-Library zum Schätzen von Modellen in Proportional-Odds-Form verwendet werden. Auch hier übergeben wir wieder einen geordneten Faktor als Response und schätzen standardmäßig ein Logit-Modell.

```
mod.polr <- polr(ord.kontr ~ dau + oberfl)
```

Das Objekt `mod.polr$zeta` beinhaltet die Schätzungen für die Schwellenwerte, welche hier mit ζ anstatt α bezeichnet werden und mit den Schätzungen der `vglm`-Funktion übereinstimmen:

$$\hat{\zeta}_1 = 0.00, \quad \hat{\zeta}_2 = 1.51, \quad \hat{\zeta}_3 = 3.40.$$

Die Schätzung für den Parametervektor ist in `mod.polr$coef` enthalten:

$$\hat{\boldsymbol{\beta}} = (0.08, 1.28, 1.38)^\top.$$

Die Werte des Vektors stimmen bis auf das Vorzeichen mit den Schätzungen der `vglm`-Funktion überein. Dies liegt an der internen Definition des Modells. Für die Funktion `vglm` ist dieses wie in (3.2.7) definiert:

$$\text{logit}(\mathbb{P}[Y \leq j|\boldsymbol{x}]) = \alpha_j + \boldsymbol{x}^\top\boldsymbol{\beta}.$$

Für die Funktion `polr` ist die interne Definition

$$\text{logit}(\mathbb{P}[Y \leq j|\boldsymbol{x}]) = \zeta_j - \boldsymbol{x}^\top\boldsymbol{\beta}.$$

Die beiden Definitionen stimmen also bis auf das Minus, welches der Grund für das unterschiedliche Vorzeichen der Schätzungen des Parametervektors $\hat{\boldsymbol{\beta}}$ ist, überein.

Durch setzen des `method`-Arguments können auch mit der Funktion `polr` der Probit- oder Complementary-Log-Log-Link verwendet werden.

```
mod.polr <- polr(ord.kontr ~ dau + oberfl, method='probit')
```

```
mod.polr <- polr(ord.kontr ~ dau + oberfl, method='cloglog')
```

Sequenzielle Modelle

Oft müssen die Kategorien einer ordinalen Response-Variable nacheinander durchlaufen werden, wie beispielsweise bei der Variable `kontr` aus dem Silikonregister mit den Kategorien `BakerI, BakerII, BakerIII, BakerIV`. In so einem Fall bietet es sich an, ein sequenzielles Modell zu verwenden, vgl. auch Tutz (2012). Für eine Response $Y \in \{1, \dots, c\}$ werden die Wahrscheinlichkeiten für die Kategorien schrittweise geschätzt. Der Prozess startet dabei mit einer Entscheidung zwischen der Kategorie $\{1\}$ und den Kategorien $\{2, \dots, c\}$, modelliert durch

$$\pi_1(\boldsymbol{x}) = F(\alpha_1 + \boldsymbol{x}^\top \boldsymbol{\beta}).$$

Dabei ist $\boldsymbol{x}$ wieder ein Vektor mit Werten für die erklärenden Variablen und F eine Verteilungsfunktion. Der Prozess stoppt, falls $Y = 1$. Andernfalls setzen wir mit einer Entscheidung zwischen der Kategorie $\{2\}$ und den Kategorien $\{3, \dots, c\}$ fort,

$$\mathbb{P}[Y = 2 | Y \geq 2, \boldsymbol{x}] = F(\alpha_2 + \boldsymbol{x}^\top \boldsymbol{\beta}).$$

Allgemein werden beim **sequenziellen Modell** die bedingten Wahrscheinlichkeiten

$$\mathbb{P}[Y = j | Y \geq j, \boldsymbol{x}] = F(\alpha_j + \boldsymbol{x}^\top \boldsymbol{\beta}), \quad \text{für } j = 1, \dots, c-1$$

modelliert. Dies entspricht im sequenziellen Prozess einer Entscheidung zwischen der Kategorie $\{j\}$ und den Kategorien $\{j+1, \dots, c\}$. Es ist $\mathbb{P}[Y = c | Y \geq c, \boldsymbol{x}] = 1$. Eine Beobachtung Y ist in Kategorie j, wenn im sequenziellen Prozess auf allen Stufen l mit $l < j$ eine Entscheidung für die Kategorien $\{l+1, \dots, c\}$ getroffen wird, was mit Wahrscheinlichkeit

$$\prod_{l=1}^{j-1} \mathbb{P}[Y > l | Y \geq l, \boldsymbol{x}]$$

geschieht, und auf Stufe j eine Entscheidung für Kategorie $\{j\}$ fällt, was mit Wahrscheinlichkeit

$$\mathbb{P}[Y = j | Y \geq j, \boldsymbol{x}]$$

vorkommt. Für die Wahrscheinlichkeiten der Kategorien gilt deshalb

$$\mathbb{P}[Y = j | \boldsymbol{x}] = \pi_j(\boldsymbol{x}) = \mathbb{P}[Y = j | Y \geq j, \boldsymbol{x}] \prod_{l=1}^{j-1} \mathbb{P}[Y > l | Y \geq l, \boldsymbol{x}].$$

Des Weiteren folgt aus der Definition der bedingten Wahrscheinlichkeit, dass

$$\mathbb{P}[Y = j | Y \geq j, \boldsymbol{x}] = \frac{\pi_j(\boldsymbol{x})}{\pi_j(\boldsymbol{x}) + \dots + \pi_c(\boldsymbol{x})}.$$

Verwenden wir als Verteilungsfunktion die logistische Verteilung $L(0,1)$, resultiert das **sequenzielle Logit-Modell**

$$\mathbb{P}[Y = j|Y \geq j, \boldsymbol{x}] = \frac{\exp\{\alpha_j + \boldsymbol{x}^\top \boldsymbol{\beta}\}}{1 + \exp\{\alpha_j + \boldsymbol{x}^\top \boldsymbol{\beta}\}}, \quad \text{für } j = 1, \ldots, c-1$$

oder äquivalent

$$\text{logit}(\mathbb{P}[Y = j|Y \geq j, \boldsymbol{x}]) = \log\left(\frac{\mathbb{P}[Y = j|Y \geq j, \boldsymbol{x}]}{\mathbb{P}[Y > j|Y \geq j, \boldsymbol{x}]}\right) = \alpha_j + \boldsymbol{x}^\top \boldsymbol{\beta}.$$

Natürlich können auch wieder andere Verteilungsfunktionen, wie z. B. die Standardnormalverteilung oder die Gumbel-Verteilung verwendet werden.

Beispiel. Man kann ein sequenzielles Modell in Proportional-Odds-Form mit der `VGAM`-Funktion `vglm` durch setzen des `family`-Arguments auf `sratio` schätzen. Dabei steht `sratio` für Stopping-Ratio (wir schätzen die Wahrscheinlichkeit, dass der Prozess auf Stufe j stoppt).

```
mod.seq <- vglm(ord.kontr ~ dau + oberfl, family=sratio(parallel=T))
```

Adjacent-Category-Modelle

Eine weitere Möglichkeit ist es, benachbarte Kategorien miteinander in Beziehung zu setzen. Dies liefert ein allgemeines **Adjacent-Category-Modell**

$$\mathbb{P}[Y = j|Y \in \{j, j+1\}, \boldsymbol{x}] = F(\eta_j(\boldsymbol{x})),$$

wobei F wieder eine beliebige Verteilungsfunktion ist und $\eta_j(\boldsymbol{x})$ der lineare Prädiktor für die Werte $\boldsymbol{x}$ der erklärenden Variable. Der lineare Prädiktor kann dabei je nach Anwendung in Proportional-Odds-Form sein oder auch nicht.

Bei Verwendung der logistischen Verteilung $L(0,1)$ folgt

$$\mathbb{P}[Y = j|Y \in \{j, j+1\}, \boldsymbol{x}] = \frac{\exp\{\eta_j(\boldsymbol{x})\}}{1 + \exp\{\eta_j(\boldsymbol{x})\}}$$

und dies liefert die sogenannten **Adjacent-Category-Logits**

$$\begin{aligned} \text{logit}(\mathbb{P}[Y = j|Y \in \{j, j+1\}], \boldsymbol{x}) &= \log\left(\frac{\mathbb{P}[Y = j|Y \in \{j, j+1\}, \boldsymbol{x}]}{\mathbb{P}[Y = j+1|Y \in \{j, j+1\}, \boldsymbol{x}]}\right) \\ &= \log\left(\frac{\pi_j(\boldsymbol{x})/(\pi_j(\boldsymbol{x}) + \pi_{j+1}(\boldsymbol{x}))}{\pi_{j+1}(\boldsymbol{x})/(\pi_j(\boldsymbol{x}) + \pi_{j+1}(\boldsymbol{x}))}\right) \\ &= \log\left(\frac{\pi_j(\boldsymbol{x})}{\pi_{j+1}(\boldsymbol{x})}\right), \quad \text{für } j = 1, \ldots, c-1. \end{aligned} \tag{3.2.9}$$

Die Adjacent-Category-Logits sind äquivalent zu den Baseline-Category-Logits, denn

$$\log\left(\frac{\pi_j}{\pi_c}\right) = \log\left(\frac{\pi_j}{\pi_{j+1}}\right) + \log\left(\frac{\pi_{j+1}}{\pi_{j+2}}\right) + \cdots + \log\left(\frac{\pi_{c-1}}{\pi_c}\right)$$

und

$$\log\left(\frac{\pi_j}{\pi_{j+1}}\right) = \log\left(\frac{\pi_j}{\pi_c}\right) - \log\left(\frac{\pi_{j+1}}{\pi_c}\right).$$

Man beachte aber, dass Adjacent-Category-Logits nur für ordinale Response-Variablen definiert sind, während Baseline-Category-Logits auch für nominale Responses existieren. Adjacent-Category-Modelle können auch als Baseline-Category-Modelle geschrieben werden. Um von der Ordnung der Kategorien zu profitieren und ein einfacheres Modell zu erhalten, werden oft Modelle mit einem gemeinsamen Parameter $\boldsymbol{\beta}$ für alle Kategorien („Proportional-Odds-Form“) verwendet. Betrachte das Adjacent-Category-Modell

$$\log\left(\frac{\pi_j(\boldsymbol{x})}{\pi_{j+1}(\boldsymbol{x})}\right) = \alpha_j + \boldsymbol{x}^\top\boldsymbol{\beta}, \quad \text{für } j = 1, \ldots, c-1,$$

dann erhalten wir daraus ein Baseline-Category-Modell durch

$$\begin{aligned}
\log\left(\frac{\pi_j(\boldsymbol{x})}{\pi_c(\boldsymbol{x})}\right) &= \log\left(\frac{\pi_j(\boldsymbol{x})}{\pi_{j+1}(\boldsymbol{x})}\right) + \log\left(\frac{\pi_{j+1}(\boldsymbol{x})}{\pi_{j+2}(\boldsymbol{x})}\right) + \cdots + \log\left(\frac{\pi_{c-1}(\boldsymbol{x})}{\pi_c(\boldsymbol{x})}\right) \\
&= \alpha_j + \boldsymbol{x}^\top\boldsymbol{\beta} + \alpha_{j+1} + \boldsymbol{x}^\top\boldsymbol{\beta} + \cdots + \alpha_{c-1} + \boldsymbol{x}^\top\boldsymbol{\beta} \\
&= \left(\sum_{l=j}^{c-1} \alpha_l\right) + (c-j)\boldsymbol{x}^\top\boldsymbol{\beta} \\
&= \delta_j + \boldsymbol{u}_j^\top\boldsymbol{\beta},
\end{aligned}$$

mit

$$\delta_j = \sum_{l=j}^{c-1} \alpha_l \quad \text{und} \quad \boldsymbol{u}_j = (c-j)\boldsymbol{x}.$$

Dies ist also ein Baseline-Category-Modell mit verschiedenen Intercept-Parametern δ_j für alle Baseline-Category-Logits, aber wieder mit gleichem Effektparameter $\boldsymbol{\beta}$. Man beachte, dass hier allerdings mit einer veränderten Design-Matrix gearbeitet wird. Anstatt der Design-Zeile $\boldsymbol{x}$ wird beim j-ten Baseline-Category-Logit mit $\boldsymbol{u}_j$ gerechnet.

Beispiel. Für ein Adjacent-Category Modell definiert man das `family`-Argument der Funktion `vglm` als `acat`. Standardmäßig wird dann

$$\frac{\mathbb{P}(Y = j+1)}{\mathbb{P}(Y = j)}$$

geschätzt. Wir setzen deshalb zusätzlich `reverse=T`, sodass wir

$$\frac{\mathbb{P}(Y = j)}{\mathbb{P}(Y = j + 1)}$$

gemäß unserer Definition (3.2.9) schätzen.

```
mod.acat <- vglm(ord.kontr ~ dau + oberfl, family=acat(reverse=T, parallel=T))
```

Die Tabellen 3.2, 3.3 und 3.4 dienen einem Vergleich der geschätzten Wahrscheinlichkeiten der Baker-Stufen für die verschiedenen Modelle, welche sehr ähnliche Schätzungen liefern.

oberfl	dau	BakerI	BakerII	BakerIII	BakerIV
glatt	1	0.479	0.327	0.159	0.035
Polyurethan	1	0.204	0.332	0.348	0.115
texturiert	1	0.188	0.323	0.363	0.126
glatt	5	0.399	0.351	0.202	0.048
Polyurethan	5	0.156	0.299	0.392	0.153
texturiert	5	0.143	0.286	0.404	0.167

Tabelle 3.2.: Geschätzte Wahrscheinlichkeiten beim kumulativen Modell.

oberfl	dau	BakerI	BakerII	BakerIII	BakerIV
glatt	1	0.314	0.396	0.254	0.036
Polyurethan	1	0.185	0.329	0.378	0.109
texturiert	1	0.183	0.328	0.379	0.110
glatt	5	0.256	0.377	0.308	0.058
Polyurethan	5	0.146	0.288	0.409	0.156
texturiert	5	0.145	0.287	0.410	0.158

Tabelle 3.3.: Geschätzte Wahrscheinlichkeiten beim sequenziellen Modell.

oberfl	dau	BakerI	BakerII	BakerIII	BakerIV
glatt	1	0.410	0.345	0.205	0.039
Polyurethan	1	0.214	0.322	0.345	0.119
texturiert	1	0.195	0.314	0.359	0.132
glatt	5	0.344	0.348	0.250	0.058
Polyurethan	5	0.163	0.296	0.382	0.159
texturiert	5	0.147	0.285	0.393	0.174

Tabelle 3.4.: Geschätzte Wahrscheinlichkeiten beim Adjacent-Category-Modell.

4. Nichtparametrische Regression

Nichtparametrische Regression erlaubt eine gewisse Flexibilität bei der Modellierung mit stetigen erklärenden Variablen. Diese fließen nicht linear in den Prädiktor ein, sondern als glatte Funktion, welche mit den Daten geschätzt wird. Dieser Abschnitt beginnt mit univariater Glättung für Modelle mit normalverteilten Responses ohne und mit Strafterm und behandelt anschließend lineare additive Modelle. Zum Schluss werden die Ideen für generalisierte additive Modelle erweitert. Der Abschnitt folgt Fahrmeir et al. (2013) und Wood (2006).

Um die Glättungsterme vom parametrischen Teil des Modells zu unterscheiden, bezeichnen wir stetige erklärende Variablen mit Glättung als z und die dazugehörigen Parameter mit γ. Des Weiteren unterscheiden wir ab jetzt in der Notation nicht mehr explizit zwischen Zufallsvariablen und deren Realisierungen. Bisher wurden Großbuchstaben für Zufallsvariablen und Kleinbuchstaben für die Realisierungen verwendet. In diesem Abschnitt werden beide mit Kleinbuchstaben bezeichnet.

4.1. Univariate Glättung

Im einfachsten Fall der nichtparametrischen Regression haben wir Daten der Form (y_i, z_i) für $i = 1, \dots, n$ gegeben. Dabei ist y_i eine Beobachtung und z_i eine stetige erklärende Variable, die auf dem Bereich $[a, b]$ definiert ist. Wir nehmen an, dass y_i durch eine Funktion der Variable z_i beschrieben werden kann, d. h.

$$y_i = f(z_i) + \varepsilon_i, \quad \text{für } i = 1, \dots, n.$$

Dabei ist ε_i ein Zufallsfehler für den $\mathbb{E}[\varepsilon_i] = 0$ und $\mathbb{V}\text{ar}[\varepsilon_i] = \sigma^2$ für $i = 1, \dots, n$ gilt. Für y_i gilt dann $\mathbb{E}[y_i] = f(z_i)$ und $\mathbb{V}\text{ar}[y_i] = \sigma^2$. Für manche Anwendungen, z. B. bei der Konstruktion von Konfidenzintervallen, wird zusätzlich noch $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$ angenommen, woraus $y_i \sim N(f(z_i), \sigma^2)$ folgt.

4.1.1. Polynomielle Splines

Durch verschiedene Annahmen bzw. Voraussetzungen für die Funktion f resultieren verschiedene Modelle und Ansätze. Es wird jeweils eine Basis $B_1(z), \dots, B_d(z)$ eines Raums von Funktionen, von dem f ein Element sein soll, gewählt. Wir nehmen also folgende Darstellung für f an

$$f(z_i) = \sum_{j=1}^{d} \gamma_j B_j(z_i).$$

Das nichtparametrische Modell

$$y_i = \sum_{j=1}^{d} \gamma_j B_j(z_i) + \varepsilon_i \tag{4.1.1}$$

ist damit wieder ein lineares Modell in den Parametern $\gamma_1, \ldots, \gamma_d$. Definieren wir die $(n \times d)$-Design-Matrix unseres Modells als

$$\boldsymbol{Z} = \begin{pmatrix} B_1(z_1) & \ldots & B_d(z_1) \\ \vdots & & \vdots \\ B_1(z_n) & \ldots & B_d(z_n) \end{pmatrix},$$

dann folgt das Modell in Matrixschreibweise

$$\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}. \tag{4.1.2}$$

Der Parametervektor $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_d)^\top$ kann mittels der Least-Squares (LS)-Methode geschätzt werden und ist dann

$$\hat{\boldsymbol{\gamma}} = (\boldsymbol{Z}^\top \boldsymbol{Z})^{-1} \boldsymbol{Z}^\top \boldsymbol{y}. \tag{4.1.3}$$

Die geschätzten Erwartungswerte sind

$$\hat{\mathbb{E}}[\boldsymbol{y}] = \hat{f}(\boldsymbol{z}) = (\hat{f}(z_1), \ldots, \hat{f}(z_n))^\top = \boldsymbol{Z}\hat{\boldsymbol{\gamma}}.$$

Im Folgenden werden verschiedene Basen für die Konstruktion der Design-Matrix eingesetzt. Für einen Vergleich verwenden wir ein Beispiel mit simulierten Daten aus Fahrmeir et al. (2013, S. 414), welche in Abbildung 4.1 zu sehen sind.

Ein erster Ansatz ist die Verwendung einer **polynomiellen Basis**

$$B_1(z) = 1, \ldots, B_d(z) = z^d.$$

Dabei wird f als Polynom vom Grad d angenommen

$$f(z_i) = \gamma_1 + \gamma_2 z_i + \cdots + \gamma_{d+1} z_i^d$$

und das nichtparametrische Modell

$$y_i = \gamma_1 + \gamma_2 z_i + \cdots + \gamma_{d+1} z_i^d + \varepsilon_i$$

ist ein lineares Regressionsmodell der Form (4.1.1). Abbildung 4.2 zeigt das Ergebnis einer LS-Schätzung anhand des Beispiels mit den simulierten Daten.

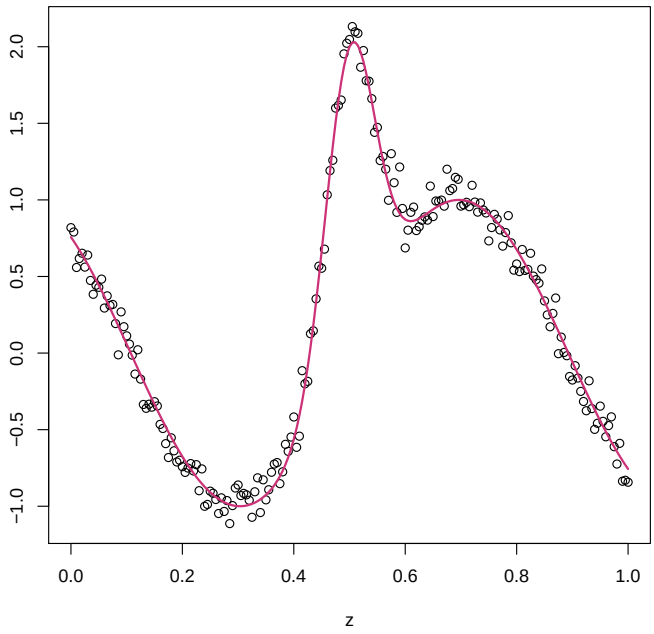


Abbildung 4.1.: Daten $y = f(z) + \varepsilon$ als Kreise, wobei $f(z) = \sin(2(4z-2)) + 2\exp\{-(16)^2(z-0.5)^2\}$ und $\varepsilon \sim N(0, 0.3^2)$ und wahre Funktion $f(z)$ in violett.

Wir sehen, dass polynomielle Modelle in der Praxis oft nicht ausreichend sind. Das lokale Maximum wird nur schlecht gefittet, aber eine Erhöhung des Grades führt zu unerwünschten Verwackelungen. Deshalb ist es besser statt einem global definierten Polynom stückweise Polynome zu verwenden.

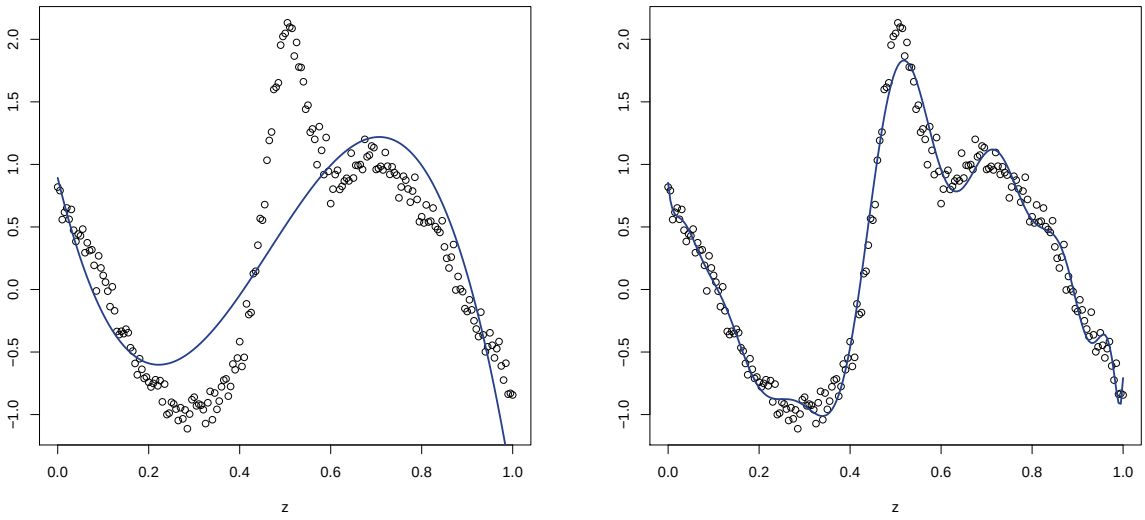
Diese Idee wird durch die **Truncated-Power (TP)-Basis** umgesetzt, welche aus den folgenden $d = l + m - 1$ Basisfunktionen besteht

$$B_1(z) = 1, \quad B_2(z) = z, \quad \ldots, \quad B_{l+1}(z) = z^l,$$
$$B_{l+2}(z) = (z - \kappa_2)_+^l, \quad \ldots, \quad B_d(z) = (z - \kappa_{m-1})_+^l.$$

Dabei sind $a = \kappa_1 < \cdots < \kappa_m = b$ fixe Knoten, die wir zunächst als bekannt annehmen, und die sogenannten **Truncated-Power-Funktionen** sind gegeben als

$$(z - \kappa_j)_+^l = \begin{cases} (z - \kappa_j)^l & \text{falls } z \geq \kappa_j, \\ 0 & \text{sonst.} \end{cases}$$

Damit kann das nichtparametrische Regressionsmodell, in dem f ein polynomieller Spline ist, folgendermaßen dargestellt werden



(a) Grad $d = 3$ (b) Grad $d = 20$

Abbildung 4.2.: Least-Squares-Schätzung eines polynomiellen Regressionsmodells.

$$\begin{aligned} y_i &= \sum_{j=1}^{d} \gamma_j B_j(z_i) + \varepsilon_i \\ &= \gamma_1 + \gamma_2 z_i + \cdots + \gamma_{l+1} z_i^l + \gamma_{l+2}(z_i - \kappa_2)_+^l + \cdots + \gamma_d (z_i - \kappa_{m-1})_+^l. \end{aligned}$$

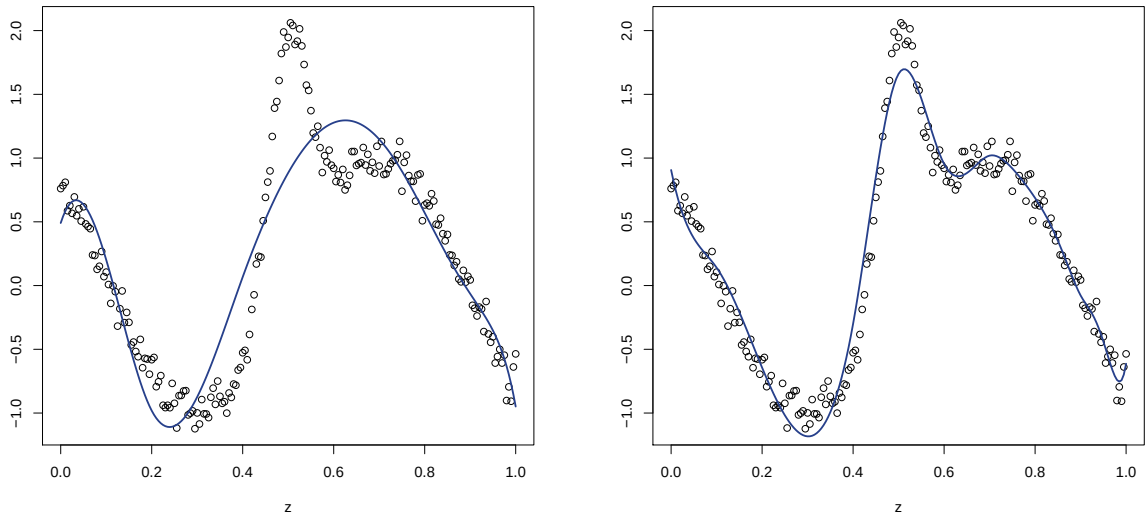
Der erste Teil des Modells ist also ein globales Polynom vom Grad l und die Truncated-Power-Funktionen im zweiten Teil des Modells beschreiben die lokale Abweichung zu diesem Polynom auf den, durch die Knoten definierten, $m-1$ Intervallen. Abbildung 4.3 zeigt das Beispiel mit den simulierten Daten, wobei zur Schätzung eine TP-Basis mit $l = 3$ und $m = 6$ bzw. $m = 11$ Knoten verwendet wurde.

Die TP-Basis besteht immer noch aus global definierten Basisfunktionen und die Basisfunktionen sind nicht nach oben beschränkt. Außerdem kommt es schnell zu Kollineritäten der Basisfunktionen, insbesondere wenn zwei Knoten sehr nahe beieinander liegen. Dies führt dazu, dass die TP-Basis numerisch instabil ist. Zu bevorzugen ist eine Basis, welche ausschließlich aus lokal definierten Basisfunktionen besteht.

Dies führt uns zur **Basic-Spline (B-Spline)-Basis** mit den rekursiv definierten Basisfunktionen

$$B_j^0(z) = \begin{cases} 1 & \text{falls } \kappa_j \leq z < \kappa_{j+1} \\ 0 & \text{sonst} \end{cases} \qquad j = 1, \ldots, d-1$$

und



(a) Grad $l = 3$, $m = 6$ Knoten

(b) Grad $l = 3$, $m = 11$ Knoten

Abbildung 4.3.: Least-Squares-Schätzung eines Modells mit TP-Basis.

$$B_j^l(z) = \frac{z - \kappa_{j-l}}{\kappa_j - \kappa_{j-l}} B_{j-1}^{l-1}(z) + \frac{\kappa_{j+1} - z}{\kappa_{j+1} - \kappa_{j+1-l}} B_j^{l-1}(z), \quad j = 1, \ldots, d-1.$$

Diese Rekursion ist bekannt als Cox-de Boor-Rekursion, vgl. auch De Boor (2001, S. 90). Um diese Rekursion anwenden zu können, brauchen wir zusätzlich zu den m Knoten $a = \kappa_1 < \cdots < \kappa_m = b$ noch $2l$ äußere Knoten, welche außerhalb des Intervalls $[a, b]$ liegen. Diese Knoten seien

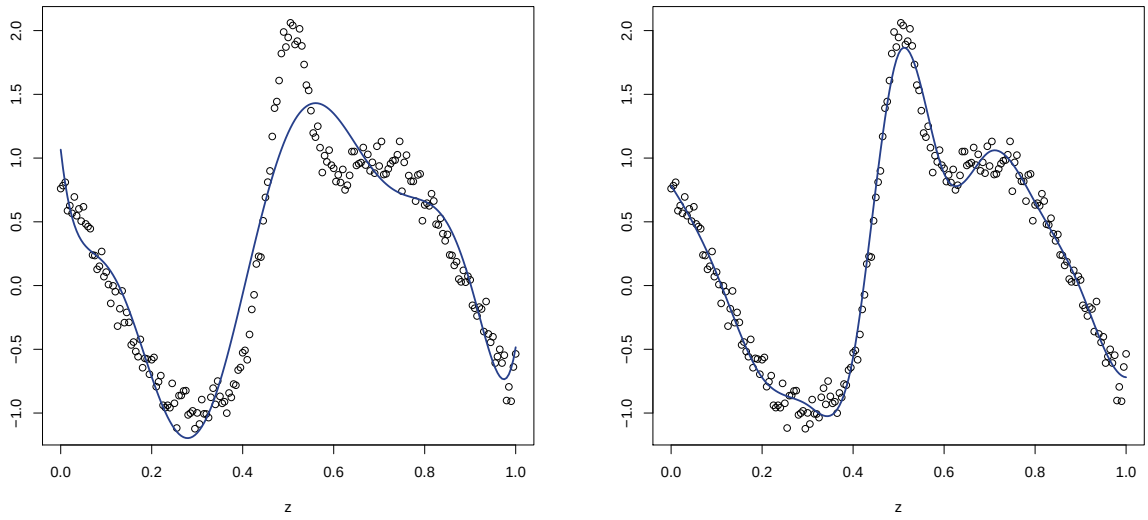
$$\kappa_{1-l} < \cdots < \kappa_0 < a \quad \text{und} \quad b < \kappa_{m+1} < \cdots < \kappa_{m+l}.$$

Die Basisfunktion $B_j^l(z)$ ist dabei nur auf genau $(l + 1)$ benachbarten Intervallen positiv und die Basisfunktionen sind nach oben beschränkt und deshalb numerisch stabiler. Abbildung 4.4 zeigt das Beispiel mit den simulierten Daten, wobei zur Schätzung eine B-Spline-Basis mit $l = 3$ und $m = 6$ bzw. $m = 11$ Knoten verwendet wurde.

Bemerkung. Sowohl die TP-Basis als auch die B-Spline-Basis spannen den Raum der polynomiellen Splines auf, siehe De Boor (2001).

Definition 6 (Polynomielle Splines)**.** *Eine Funktion $f : [a, b] \to \mathcal{R}$ heißt **polynomieller Spline** vom Grad $l \geq 0$ mit Knoten $a = \kappa_1 < \cdots < \kappa_m = b$, falls sie zwei Bedingungen erfüllt:*

1. *Die Funktion $f(z)$ ist $(l - 1)$-mal stetig differenzierbar. Im Spezialfall $l = 1$ muss $f(z)$ also stetig sein (aber nicht differenzierbar) und im Fall $l = 0$ gibt es keine Glattheitsbedingungen.*



(a) Grad $l = 3$, $m = 6$ Knoten

(b) Grad $l = 3$, $m = 11$ Knoten

Abbildung 4.4.: Least-Squares-Schätzung eines Modells mit B-Spline-Basis.

2. *Die Funktion $f(z)$ ist ein Polynom von Grad l auf jedem durch die Knoten definierten Intervall $[\kappa_j, \kappa_{j+1})$.*

Der Grad des Polynoms kontrolliert die globale Glattheit der Funktion, während die Anzahl der Knoten die Vielfältigkeit der schätzbaren Funktionen bestimmt. Wir müssen also noch folgende Fragen beantworten:

1. Wie soll der Grad des Splines gewählt werden?
2. Wie viele Knoten sollen verwendet werden und wie sollen sie verteilt werden?

Im obigen Beispielen mit den simulierten Daten haben wir sowohl bei der TP-Basis als auch bei der B-Spline-Basis $l = 3$ gewählt. Es stellt sich heraus, das dies eine recht gute Wahl ist, denn die sogenannten **kubischen Splines** besitzen eine Optimalitätseigenschaft, wie wir später noch sehen werden. Die Lage der Knoten kann z. B. äquidistant oder durch Verwendung der $\frac{j-1}{m-1}$-Quantile ($j = 1, \ldots, m$) der erklärenden Variable z erfolgen. Um die Anzahl der Knoten zu steuern wird ein Strafterm eingeführt.

4.1.2. Penalized Splines

Wie wir anhand des Beispiels mit den simulierten Daten sehen, hängt die Qualität der geschätzten Funktion sehr stark von der Anzahl der Knoten ab. Die Wahl der Anzahl der Knoten wollen wir in gewisser Weise automatisieren. Eine allgemeine Vorgehensweise ist die folgende:

1. Wähle f als polynomiellen Spline mit einer großzügigen Anzahl an Knoten (ca. 20-40). Die Lage der Knoten ist dann nicht mehr so wesentlich und oft werden aufgrund der Einfachheit äquidistante Knoten verwendet.

2. Führe einen Strafterm ein, sodass unnötige Verwackelungen der geschätzten Funktion („overfitting“) vermieden werden.

Wir können die Parameter durch Minimierung des Penalized-Least-Squares (PLS)-Kriteriums

$$\|\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\gamma}\|^2 + \lambda \boldsymbol{\gamma}^\top \boldsymbol{P} \boldsymbol{\gamma} \tag{4.1.4}$$

schätzen, wobei (4.1.4) auch pönalisierte Fehlerquadratsumme genannt wird. Dabei ist der Parameter $\lambda \geq 0$ und heißt **Smoothing-Parameter**. Er steuert den Trade-Off zwischen Glattheit der geschätzten Funktion und Anpassung an die Daten. Die Glattheit der geschätzten Funktion wird nun nicht mehr durch die Anzahl der Knoten, sondern durch die Wahl des Parameters λ kontrolliert. Die Matrix $\boldsymbol{P}$ nennen wir **Strafmatrix** („penalty matrix“). Den PLS-Schätzer erhalten wir durch Nullsetzen der zu (4.1.4) gehörenden γ-Score-Funktion

$$-2\boldsymbol{Z}^\top \boldsymbol{y} + 2\boldsymbol{Z}^\top \boldsymbol{Z}\boldsymbol{\gamma} + 2\lambda \boldsymbol{P}\boldsymbol{\gamma},$$

welcher demnach gegeben ist durch

$$\hat{\boldsymbol{\gamma}} = (\boldsymbol{Z}^\top \boldsymbol{Z} + \lambda \boldsymbol{P})^{-1} \boldsymbol{Z}^\top \boldsymbol{y}. \tag{4.1.5}$$

Die geschätzten Erwartungswerte sind wieder

$$\hat{\mathbb{E}}[\boldsymbol{y}] = \hat{f}(\boldsymbol{z}) = (\hat{f}(z_1), \ldots, \hat{f}(z_n))^\top = \boldsymbol{Z}\hat{\boldsymbol{\gamma}}.$$

Für $\lambda = 0$ erhalten wir den unbestraften LS-Schätzer (4.1.3).

Strafmatrix für TP-Basis

Eine TP-Basis besteht aus $l+1$ Basisfunktionen, die ein globales Polynom darstellen, und aus weiteren $m-2$ Basisfunktionen, welche die lokale Abweichung zu diesem Polynom auf den Intervallen mithilfe von Truncated-Power-Funktionen beschreiben. Ein Spline ist aber genau dann glatt, wenn alle Koeffizienten ähnliche Werte haben. Eine zu große Abweichung vom globalen Polynom auf den Intervallen ist nicht wünschenswert. Um die Abweichung zu kontrollieren führen wir deshalb folgende Strafe für die zu den TP-Funktionen gehörenden Parameter ein:

$$\sum_{j=l+2}^{d} \gamma_j^2.$$

Dadurch wird der Effekt der individuellen stückweisen Funktionen reduziert und somit Overfitting vermieden. Einen Strafterm dieser Form nennt man **Ridge-Penalty**. Wir minimieren also die pönalisierte Fehlerquadratsumme

$$\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{d}\gamma_j B_j(z_i)\right)^2 + \lambda \sum_{j=l+2}^{d}\gamma_j^2,$$

wobei $\{B_1(z), \ldots, B_d(z)\}$ die TP-Basis ist. Mit der Strafmatrix

$$\boldsymbol{P} = \text{diag}(\underbrace{0, \ldots, 0}_{l+1}, \underbrace{1, \ldots, 1}_{m-2})$$

kann man dies in Matrixschreibweise wie in (4.1.4) schreiben. Für $\lambda = 0$ verschwindet der Effekt der Strafe und wir erhalten dieselben Ergebnisse wie bei der gewöhnlichen LS-Schätzung ohne Strafterm. Für $\lambda \to \infty$ wird das Resultat vom Strafterm dominiert und es folgt $\gamma_j = 0$ für $j = l+2, \ldots, d$. Die geschätzte Funktion $\hat{f}(z_i)$ ist in diesem Fall ein Polynom von Grad l.

Strafmatrix für B-Spline-Basis

Bei Verwendung der B-Spline-Basis gibt es keine Aufteilung des Modells in ein globales Polynom und die Abweichung davon. Deshalb wählen wir hier einen anderen Zugang als bei Verwendung der TP-Basis. Der Strafterm basiert hier auf den (quadrierten) zweiten Ableitungen der Funktion f. Diese ist als Änderung der Steigung f' ein Maß für die Krümmung einer Funktion. Eine zu starke Krümmung soll bestraft werden und der Strafterm ist dann

$$\int (f''(z))^2 dz. \tag{4.1.6}$$

Wir minimieren die pönalisierte Fehlerquadratsumme

$$\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{d}\gamma_j B_j(z_i)\right)^2 + \lambda \int (f''(z))^2 dz,$$

wobei $\{B_1(z), \ldots, B_d(z)\}$ die B-Spline-Basis ist und

$$f''(z) = \sum_{j=1}^{d}\gamma_j B_j''(z) = \boldsymbol{\gamma}^\top \boldsymbol{b}(z)$$

mit $\boldsymbol{b}(z) = (B_1''(z), \ldots, B_d''(z))^\top$ gilt. Damit folgt

$$\begin{aligned}\int (f''(z))^2 dz &= \int \boldsymbol{\gamma}^\top \boldsymbol{b}(z)\boldsymbol{b}(z)^\top \boldsymbol{\gamma}\, dz \\ &= \boldsymbol{\gamma}^\top \left(\int \boldsymbol{b}(z)\boldsymbol{b}(z)^\top\, dz\right)\boldsymbol{\gamma} \\ &= \boldsymbol{\gamma}^\top \boldsymbol{P}\boldsymbol{\gamma}.\end{aligned}$$

Mit der Strafmatrix $\boldsymbol{P} = \int \boldsymbol{b}(z)\boldsymbol{b}(z)^\top dz$ folgt also wieder die Matrixschreibweise (4.1.4) für das PLS-Kriterium.

In De Boor (2001, S. 115) wurde gezeigt, dass

$$\frac{\partial}{\partial z} B_j^l(z) = l \left(\frac{1}{\kappa_j - \kappa_{j-l}} B_{j-1}^{l-1}(z) - \frac{1}{\kappa_{j+1} - \kappa_{j+1-l}} B_j^{l-1}(z) \right)$$

gilt. Daraus folgt als Ableitung für den polynomiellen Spline

$$\frac{\partial}{\partial z} \sum_{j=1}^{d} \gamma_j B_j^l(z) = l \cdot \sum_{j=1}^{d} \frac{\gamma_j - \gamma_{j-1}}{\kappa_j - \kappa_{j-l}} B_{j-1}^{l-1}(z).$$

Man kann also die Ableitung als Funktion in den Differenzen 1. Ordnung der Parameter darstellen. Analog kann die zweite Ableitung als Funktion in den Differenzen 2. Ordnung dargestellt werden. Durch Schätzen der Parameter γ_j erhält man also nicht nur eine Schätzung für die Funktion $f(z)$, sondern auch eine Schätzung für $f''(z)$. Deshalb wird ein Strafterm der Form (4.1.6) auch als **Difference-Penalty** bezeichnet. Die Strafmatrix kann dann mithilfe der sogenannten Difference-Matrizen geschrieben werden. In unserem Fall brauchen wir $\boldsymbol{D}_2$, die Difference-Matrix der Ordnung 2, welche gegeben ist als folgende Bandmatrix:

$$\boldsymbol{D}_2 = \begin{pmatrix} 1 & -2 & 1 & & & \\ & 1 & -2 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & -2 & 1 \end{pmatrix}.$$

Als Strafmatrix folgt dann

$$\boldsymbol{P} = \boldsymbol{D}_2^\top \boldsymbol{D}_2 = \begin{pmatrix} 1 & -2 & 1 & & & & \\ -2 & 5 & -4 & 1 & & & \\ 1 & -4 & 6 & -4 & 1 & & \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & 1 & -4 & 6 & -4 & 1 \\ & & & 1 & -4 & 5 & -2 \\ & & & & 1 & -2 & 1 \end{pmatrix}.$$

Vergleich der beiden Ansätze

Wir haben nun zwei Ansätze für P-Splines kennengelernt. Zum einen können wir eine TP-Basis mit einer Ridge-Penalty verwenden zum anderen eine B-Spline-Basis mit Difference-Penalty. Eilers und Marx (1996) sind Verfechter des ersten Ansatzes in Verbindung mit äquidistant verteilten Knoten. Ruppert, Wand und Carroll (2003) haben sich für den zweiten Ansatz eingesetzt, wobei sie die Knoten basierend auf Quantilen verteilen. In Eilers und Marx (2010) werden beide Methoden verglichen und Vor- und Nachteile aufgezeigt.

Eilers und Marx haben festgestellt, dass bei Anwendung von P-Splines eine äquidistante Verteilung der Knoten zu bevorzugen ist. Der Strafterm stellt sicher, dass eine hinreichend glatte Funktion resultiert. TP-Basisfunktionen können numerisch instabil sein, wie wir schon wissen. Auch Eilers und Marx sehen dies als Nachteil der TP-Basis an. Außerdem sind die Strafmatrizen bei einer Difference-Penalty dünn besetzt, was Vorteile bei sehr großen Problemen bringt. Des Weiteren weisen B-Spline Basisfunktionen bei der multivariaten Glättung mittels Tensorprodukten bessere Eigenschaften auf.

4.1.3. Optimalitätseigenschaft von natürlichen kubischen Splines

In den vorherigen Abschnitten haben wir angenommen, dass f ein polynomieller Spline ist. Schon im Beispiel mit den simulierten Daten haben wir gesehen, dass kubische Splines sehr gut zu passen scheinen. Dies wollen wir nun rechtfertigen.

Wir nehmen hier zunächst nur an, dass f eine zweimal stetig differenzierbare Funktion ist, sodass die pönalisierte Fehlerquadratsumme definiert ist. Wir suchen dann jene zweimal stetig differenzierbare Funktion f, die

$$\sum_{i=1}^{n}(y_i - f(z_i))^2 + \lambda \int (f''(z))^2 dz$$

minimiert. Die Lösung dieses Minimierungsproblems ist ein natürlicher kubischer Spline (Beweis siehe Anhang A.8).

Definition 7 (Natürlicher kubischer Spline)**.** *Die Funktion $f(z)$ ist ein **natürlicher kubischer Spline** basierend auf den Knoten $a \leq \kappa_1 < \cdots < \kappa_m \leq b$, falls*

1. *$f(z)$ ein kubischer polynomieller Spline für die Knoten ist und*
2. *die Randbedingung $f''(a) = f''(b) = 0$ erfüllt ist.*

Abbildung (4.5) zeigt das Ergebnis der PLS-Schätzung für unser Beispiel mit den simulierten Daten. Dabei wurden kubische Splines als Basis verwendet und die maximale Basisdimension mit 20 festgelegt. Die Schätzung des Glättungsparameters erfolgte mittels generalisierter Cross-Validation (siehe Abschnitt 4.1.5).

4.1.4. Lineare Glätter

Glätter, für die man die geschätzte Funktion $\hat{f(\boldsymbol{z})} = (\hat{f}(z_1), \ldots, \hat{f}(z_n))^\top$ als Linearkombination der Beobachtungen schreiben kann, nennt man lineare Glätter. Dann gilt

$$\hat{f}(\boldsymbol{z}) = \boldsymbol{S}\boldsymbol{y}.$$

Dabei ist $\boldsymbol{S}$ eine $(n \times n)$-Matrix und wird **Glättungsmatrix** genannt. Bei der Verwendung von polynomiellen Splines ist $\hat{\boldsymbol{\gamma}}$ gegeben wie in (4.1.3) und es folgt

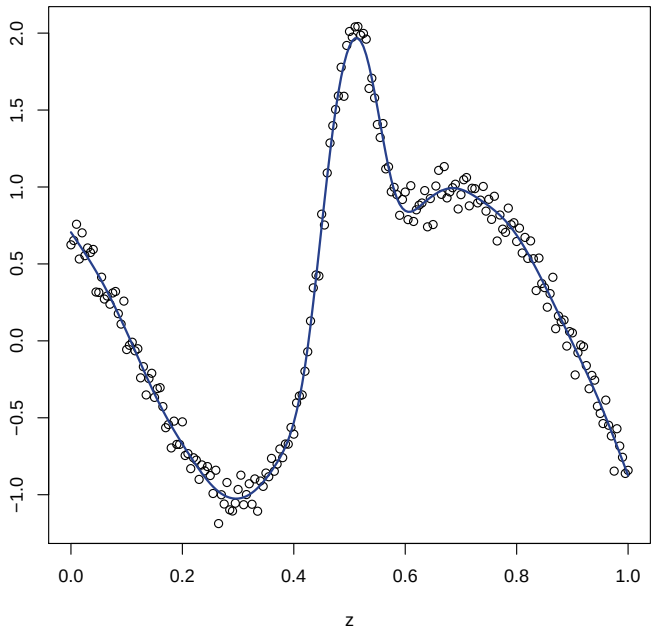


Abbildung 4.5.: PLS-Schätzung mit kubischen Splines.

$$\hat{f}(\boldsymbol{z}) = \boldsymbol{Z}\hat{\boldsymbol{\gamma}} = \boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}\boldsymbol{Z}^\top \boldsymbol{y} = \boldsymbol{S}\boldsymbol{y}.$$

Falls P-Splines verwendet werden ist $\hat{\boldsymbol{\gamma}}$ gegeben wie in (4.1.5) und es folgt

$$\hat{f}(\boldsymbol{z}) = \boldsymbol{Z}\hat{\boldsymbol{\gamma}} = \boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{Z} + \lambda \boldsymbol{P})^{-1}\boldsymbol{Z}^\top \boldsymbol{y} = \boldsymbol{S}\boldsymbol{y}.$$

Die Glättungsmatrix ist also gegeben durch

$$\boldsymbol{S} = \boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}\boldsymbol{Z}^\top \quad \text{bzw.} \quad \boldsymbol{S} = \boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{Z} + \lambda \boldsymbol{P})^{-1}\boldsymbol{Z}^\top.$$

Im Falle von polynomiellen Splines entspricht die Glättungsmatrix also der Hat-Matrix, im Falle von P-Splines kommt noch ein Strafterm dazu.

Es sei $\tilde{z}$ ein neuer Wert der erklärenden Variable, für den wir den Funktionswert voraussagen wollen. Dieser ist dann

$$\hat{f}(\tilde{z}) = \boldsymbol{s}(\tilde{z})^\top \boldsymbol{y},$$

wobei

$$\boldsymbol{s}(\tilde{z})^\top = (B_1(\tilde{z}), \ldots, B_d(\tilde{z}))(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}\boldsymbol{Z}^\top$$

bzw.

$$\boldsymbol{s}(\tilde{z})^\top = (B_1(\tilde{z}), \ldots, B_d(\tilde{z}))(\boldsymbol{Z}^\top \boldsymbol{Z} + \lambda \boldsymbol{P})^{-1} \boldsymbol{Z}^\top.$$

4.1.5. Schätzung des Glättungsparameters

Zur optimalen Schätzung des Glättungsparameters λ gibt es mehrere Ansätze. Einerseits kann man den Glättungsparameter basierend auf Optimalitätskriterien, wie z.B. die Minimierung des MSE bei der Vorhersage für neue Beobachtungen, geschätzt werden, andererseits kann man ein nichtparametrisches Regressionsmodell als Mixed-Model sehen. Dann erhält man den Glättungsparameter mittels Maximum-Likelihood (ML)- oder Restricted-Maximum-Likelihood (REML)-Schätzung.

Schätzung mittels Cross-Validation

Für die Schätzung des Glättungsparameters λ brauchen wir ein Maß für die Abweichung der geschätzten Funktion $\hat{f}(z)$ von der wahren Funktion $f(z)$. Ein Maß für die punktweise Abweichung ist beispielsweise der mittlere quadratische Fehler (MSE)

$$\mathrm{MSE}(\hat{f}(z)) = \mathbb{E}[(\hat{f}(z) - f(z))^2].$$

Durch Mitteln der MSE-Terme für alle n Punkte z_i erhalten wir

$$\mathrm{M} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[(\hat{f}(z_i) - f(z_i))^2].$$

Da die wahre Funktion $f(z)$ aber unbekannt ist setzt man y_i statt $f(z_i)$ als naive Approximation ein. Dann würde M aber durch $\hat{f}(z_i) = y_i$ minimiert werden. Deshalb verwendet man in der Regel den mittleren quadratischen Fehler der bei der Vorhersage („prediction“) für eine neue Beobachtung y^* resultiert. In der Praxis wird dazu der Datensatz in k Teilmengen $\{T_1, \ldots, T_k\}$ aufgeteilt. In jedem von k Schritten wird nun eine der Teilmengen als Validierungsdatensatz (=„neue Beobachtungen“) verwendet. Die restlichen $k-1$ Teilmengen dienen wie gewohnt zum schätzen des Erwartungswertes. Oft wird ein Spezialfall, die sogenannte Leaving-One-Out-Cross-Validation, angewandt. Dabei wird in jedem Schritt nur eine Beobachtung ausgelassen. Die Prozedur funktioniert dann wie folgt:

1. Iteriere für $i = 1, \ldots, n$.
2. Entferne die Beobachtung (y_i, z_i) aus dem Datensatz.
3. Schätze die glatte Funktion $\hat{f}^{(-i)}$ mit den restlichen $n-1$ Beobachtungen.
4. Es sei $\hat{f}^{(-i)}(z_i)$ die Vorhersage für y_i.

Bilden des arithmetischen Mittels liefert das gewöhnliche **Cross-Validation-Kriterium** (OCV-Kriterium)

$$\nu_o = \frac{1}{n}\sum_{i=1}^{n}(\hat{f}^{(-i)}(z_i) - y_i)^2. \tag{4.1.7}$$

Wir bestimmen λ durch Minimierung von (4.1.7). Dabei ist λ hier in $\hat{f}^{(-i)}$ enthalten. Es gilt

$$\begin{aligned}
\mathbb{E}[\nu_o] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(\hat{f}^{(-i)}(z_i) - y_i)^2\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[(\hat{f}^{(-i)}(z_i) - f(z_i) - \varepsilon_i)^2] \\
&= \frac{1}{n}\sum_{i=1}^{n}\left\{\mathbb{E}[(\hat{f}^{(-i)}(z_i) - f(z_i))^2] - 2\mathbb{E}[(\hat{f}^{(-i)}(z_i) - f(z_i))\varepsilon_i] + \mathbb{E}[\varepsilon_i^2]\right\}.
\end{aligned}$$

Dabei sind $\hat{f}^{(-i)}(z_i)$ und ε_i unabhängig, weil die Beobachtung y_i bei der Schätzung von $\hat{f}^{(-i)}(z_i)$ nicht verwendet wurde. Der Erwartungswert des Mischterms ist also wegen $\mathbb{E}[\varepsilon_i] = 0$ gleich Null. Des Weiteren ist $\mathbb{E}[\varepsilon_i^2] = \sigma^2$. Mit $\hat{f}^{(-i)}(z) \approx \hat{f}(z)$ für große Stichproben folgt

$$\mathbb{E}[\nu_o] \approx \mathrm{M} + \sigma^2.$$

Dadurch können wir also rechtfertigen, dass die Minimierung von (4.1.7) zur Bestimmung des Glättungsparameters λ angebracht ist.

Auf den ersten Blick muss man beim OCV-Kriterium n nichtparametrische Modelle anpassen, was viel Rechenaufwand erfordern kann. Wir können aber glücklicherweise eine effizientere Methode herleiten. Betrachten wir dazu das Penalized-Least-Squares-Kriterium (4.1.4) für die Probleme die während der Cross-Validierung auftreten

$$\sum_{\substack{l=1\\l\neq i}}^{n}(y_l - \hat{f}^{(-i)}(z_l))^2 + \lambda\boldsymbol{\gamma}^\top\boldsymbol{P}\boldsymbol{\gamma}. \tag{4.1.8}$$

Wir minimieren (4.1.8), um den i-ten Term des OCV-Kriteriums zu finden. Die Schätzer ändern sich nicht, wenn wir Null zu (4.1.8) addieren und mit

$$y_l^* = y_l \text{ für } l \neq i \quad \text{und} \quad y_i^* = \hat{f}^{(-i)}(z_i)$$

erhalten wir dann

$$\sum_{l=1}^{n}(y_l^* - \hat{f}^{(-i)}(z_l))^2 + \lambda\boldsymbol{\gamma}^\top\boldsymbol{P}\boldsymbol{\gamma}.$$

Anpassen des Modells liefert $\hat{f}^{(-i)}(z_i)$ als i-te Vorhersage und die Glättungsmatrix $\boldsymbol{S}$, welche mit der Glättungsmatrix eines Modells mit allen Daten entspricht. Einen Beweis dafür liefert das sogenannte Leaving-One-Out-Lemma, siehe Wahba (1990, Abschnitt 4.2). Wenn $\boldsymbol{s}_i$ die i-te Zeile und s_{ii} das i-te Diagonalelement der Glättungsmatrix ist, folgt

$$\begin{aligned}\hat{f}^{(-i)}(z_i) &= \boldsymbol{s}_i \boldsymbol{y}^* \\ &= \boldsymbol{s}_i \boldsymbol{y} - s_{ii} y_i + s_{ii} \hat{f}^{(-i)}(z_i) \\ &= \hat{f}(z_i) - s_{ii} y_i + s_{ii} \hat{f}^{(-i)}(z_i).\end{aligned}$$

Durch Multiplikation mit (-1) und Addition von y_i erhält man

$$y_i - \hat{f}^{(-i)}(z_i) = y_i - \hat{f}(z_i) + s_{ii}(y_i - \hat{f}^{(-i)}(z_i)),$$

woraus durch einfache Umformungen

$$y_i - \hat{f}^{(-i)}(z_i) = \frac{y_i - \hat{f}(z_i)}{1 - s_{ii}}$$

folgt. Dies liefert als OCV-Kriterium

$$\nu_o = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{y_i - \hat{f}(z_i)}{1 - s_{ii}} \right)^2.$$

Für die Berechnung von ν_o muss also *nur* ein nichtparametrisches Modell geschätzt und die Glättungsmatrix $\boldsymbol{S}$ berechnet werden. Die explizite Berechnung von $\boldsymbol{S}$ kann vor allem für große Datensätze aufwändig werden. Ersetzt man die Diagonalelemente s_{ii} durch ihr Mittel erhält man das **generalisierte Cross-Validation-Kriterium** (GCV-Kriterium)

$$\nu_g = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{y_i - \hat{f}(z_i)}{1 - \operatorname{tr}(\boldsymbol{S})/n} \right)^2.$$

Die Berechnung von ν_g ist effizienter als die Berechnung von ν_o, denn zur Bestimmung der Spur von $\boldsymbol{S}$ müssen wir die Glättungsmatrix nicht explizit ausrechnen. Des Weiteren ist das GCV-Kriterium invariant bzgl. orthogonaler Transformation, was für das OCV-Kriterium nicht der Fall ist, siehe Wood (2006, Abschnitt 4.5.2).

Schätzung mittels ML und REML

Ausgangspunkt ist ein Modell wie in (4.1.2)

$$\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon} \quad \text{mit } \boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}),$$

wofür die pönalisierte Fehlerquadratsumme (4.1.4) minimiert wird. Der Strafterm hat die Form

$$\lambda \boldsymbol{\gamma}^\top \boldsymbol{P} \boldsymbol{\gamma},$$

wobei $\boldsymbol{P}$ eine symmetrische, positiv semidefinite Matrix ist (gilt für beide Strafmatrizen in Abschnitt 4.1.2). In diesem Abschnitt folgt zunächst eine Einführung in Mixed-Models und Methoden zur Varianzschätzung für diese (ML, REML). Danach wird gezeigt wie ein nichtparametrisches Modell als Mixed-Model geschrieben werden kann, um dann die ML- oder REML-Methode für die Schätzung des Glättungsparameters λ zu verwenden.

In Mixed-Models wird der lineare Prädiktor $\boldsymbol{\eta}$ um einen zufälligen Effekt $\boldsymbol{\gamma}$ erweitert. Deshalb werden solche Modelle oft auch Random-Effect-Models genannt, vgl. Fahrmeir et al. (2013, Kapitel 7) für eine Einführung in Mixed-Models. Lineare Mixed-Models haben die Form

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{U}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}, \quad \text{mit } \boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}) \text{ und } \boldsymbol{\gamma} \sim N(\boldsymbol{0}, \boldsymbol{G}), \tag{4.1.9}$$

wobei $\boldsymbol{\gamma}$ und $\boldsymbol{\varepsilon}$ als unabhängig angenommen werden. Dabei ist $\boldsymbol{\beta}$ ein fixer Effekt mit Design-Matrix $\boldsymbol{X}$ und $\boldsymbol{\gamma}$ ein zufälliger Effekt mit Design-Matrix $\boldsymbol{U}$. Im *konditionalen* Modell für $\boldsymbol{y}$ gegeben $\boldsymbol{\gamma}$ wird $\boldsymbol{\gamma}$ als fix angesehen und es ist

$$\boldsymbol{y}|\boldsymbol{\gamma} \sim N(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{U}\boldsymbol{\gamma}, \boldsymbol{\Sigma}). \tag{4.1.10}$$

Um die *marginale* Verteilung von $\boldsymbol{y}$ zu erhalten, schreiben wir (4.1.9) um, sodass

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}^*, \quad \text{mit } \boldsymbol{\varepsilon}^* = \boldsymbol{U}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}.$$

Dann ist der Erwartungswert von $\boldsymbol{\varepsilon}^*$ gegeben als

$$\mathbb{E}[\boldsymbol{\varepsilon}^*] = \boldsymbol{0}$$

und die Kovarianzmatrix ist wegen der Unabhängigkeit von $\boldsymbol{\gamma}$ und $\boldsymbol{\varepsilon}$

$$\boldsymbol{V} \coloneqq \mathbb{C}\text{ov}[\boldsymbol{\varepsilon}^*] = \mathbb{C}\text{ov}[\boldsymbol{U}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}] = \mathbb{C}\text{ov}[\boldsymbol{U}\boldsymbol{\gamma}] + \mathbb{C}\text{ov}[\boldsymbol{\varepsilon}] = \boldsymbol{U}\boldsymbol{G}\boldsymbol{U}^\top + \boldsymbol{\Sigma}.$$

Daraus folgt

$$\boldsymbol{y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{V}) \tag{4.1.11}$$

als marginale Verteilung für $\boldsymbol{y}$.

Angenommen $\boldsymbol{\Sigma}$ und $\boldsymbol{G}$ (und somit $\boldsymbol{V}$) sind bekannt, dann können Schätzer für die Parameter $\boldsymbol{\beta}$ und $\boldsymbol{\gamma}$ durch Maximierung der gemeinsamen Log-Likelihood-Funktion von $\boldsymbol{y}$ und $\boldsymbol{\gamma}$ gefunden werden. Es gilt für die gemeinsame Dichte

$$\begin{aligned} f(\boldsymbol{y}, \boldsymbol{\gamma}) &= f(\boldsymbol{y}|\boldsymbol{\gamma}) f(\boldsymbol{\gamma}) \\ &\propto \exp\left\{-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{U}\boldsymbol{\gamma})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{U}\boldsymbol{\gamma})\right\} \cdot \exp\left\{-\frac{1}{2}\boldsymbol{\gamma}^\top \boldsymbol{G}^{-1}\boldsymbol{\gamma}\right\} \end{aligned}$$

und die gemeinsame Log-Likelihood ist dann

$$-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{U}\boldsymbol{\gamma})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{U}\boldsymbol{\gamma})-\frac{1}{2}\boldsymbol{\gamma}^\top\boldsymbol{G}^{-1}\boldsymbol{\gamma}.$$

Die Maximierung dieser Log-Likelihood ist äquivalent zur Minimierung des folgenden PLS-Kriteriums

$$(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{U}\boldsymbol{\gamma})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{U}\boldsymbol{\gamma})+\boldsymbol{\gamma}^\top\boldsymbol{G}^{-1}\boldsymbol{\gamma}.$$

Der erste Term entspricht dabei einem gewichteten LS-Kriterium. Der zweite Term bezieht die Zufälligkeit von $\boldsymbol{\gamma}$ mit ein, wobei Abweichungen des Vektors $\boldsymbol{\gamma}$ vom Nullvektor bestraft werden.

In der Praxis sind $\boldsymbol{\Sigma}$ und $\boldsymbol{G}$ meistens unbekannt. Es sei $\boldsymbol{\vartheta}$ ein Vektor, der alle Parameter enthält, die in $\boldsymbol{\Sigma}$ und $\boldsymbol{G}$ vorkommen. Wir schreiben dann $\boldsymbol{\Sigma}(\boldsymbol{\vartheta})$ und $\boldsymbol{G}(\boldsymbol{\vartheta})$, um die Abhängigkeit deutlich zu machen. Es gibt nun zwei Ansätze zur Schätzung von $\boldsymbol{\vartheta}$.

1) Der Parameter $\boldsymbol{\vartheta}$ kann mittels ML-Schätzung basierend auf dem marginalen Modell $\boldsymbol{y}\sim N(\boldsymbol{X}\boldsymbol{\beta},\boldsymbol{V}(\boldsymbol{\vartheta}))$ geschätzt werden. Die dazugehörige Log-Likelihood ist

$$\log L(\boldsymbol{\beta},\boldsymbol{\vartheta})=-\frac{1}{2}\log(|\boldsymbol{V}(\boldsymbol{\vartheta})|)-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^\top\boldsymbol{V}(\boldsymbol{\vartheta})^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}). \tag{4.1.12}$$

Maximierung von (4.1.12) bezüglich $\boldsymbol{\beta}$ ergibt

$$\hat{\boldsymbol{\beta}}(\boldsymbol{\vartheta})=(\boldsymbol{X}^\top\boldsymbol{V}(\boldsymbol{\vartheta})^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{V}(\boldsymbol{\vartheta})^{-1}\boldsymbol{y}$$

und durch Einsetzen von $\hat{\boldsymbol{\beta}}(\boldsymbol{\vartheta})$ in (4.1.12) erhalten wir die Profile-Log-Likelihood

$$l_P(\boldsymbol{\vartheta})=-\frac{1}{2}\log(|\boldsymbol{V}(\boldsymbol{\vartheta})|)-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}(\boldsymbol{\vartheta}))^\top\boldsymbol{V}(\boldsymbol{\vartheta})^{-1}(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}(\boldsymbol{\vartheta})).$$

Die Maximierung von $l_P(\boldsymbol{\vartheta})$ bezüglich $\boldsymbol{\vartheta}$ liefert den ML-Schätzer $\hat{\boldsymbol{\vartheta}}_{\mathrm{ML}}$.

2) Anstatt der Log-Likelihood kann auch die sogenannte restringierte Log-Likelihood

$$l_R(\boldsymbol{\vartheta})=\log\left(\int L(\boldsymbol{\beta},\boldsymbol{\vartheta})d\boldsymbol{\beta}\right)$$

verwendet werden. Es kann gezeigt werden, dass

$$l_R(\boldsymbol{\vartheta})=l_P(\boldsymbol{\vartheta})-\frac{1}{2}\log|\boldsymbol{X}^\top\boldsymbol{V}(\boldsymbol{\vartheta})^{-1}\boldsymbol{X}|.$$

Maximierung von $l_R(\boldsymbol{\vartheta})$ bezüglich $\boldsymbol{\vartheta}$ liefert dann den restringierten ML-Schätzer $\hat{\boldsymbol{\vartheta}}_{\mathrm{REML}}$.

Falls $\boldsymbol{\varepsilon}\sim N(\boldsymbol{0},\sigma^2\boldsymbol{I})$ und $\boldsymbol{\gamma}\sim N(\boldsymbol{0},\tau^2\boldsymbol{I})$ dann folgt als PLS-Kriterium für das Mixed-Model

$$(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{U}\boldsymbol{\gamma})^\top(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{U}\boldsymbol{\gamma})+\frac{\sigma^2}{\tau^2}\boldsymbol{\gamma}^\top\boldsymbol{\gamma}. \tag{4.1.13}$$

Der Vektor $\boldsymbol{\vartheta}$ ist dann $\boldsymbol{\vartheta} = (\sigma^2, \tau^2)^\top$ und kann mittels ML oder REML geschätzt werden. Der Quotient der Varianzen σ^2/τ^2 wird später die Rolle des Glättungsparameters λ spielen.

Nun werden wir zeigen, wie wir ein nichtparametrisches Modell (4.1.2) als Mixed-Model schreiben können. Für eine TP-Basis mit Ridge-Penalty ist die Darstellung als Mixed-Model recht einfach herzuleiten. Das PLS-Kriterium für das nichtparametrische Modell ist in diesem Fall

$$\|\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\gamma}\|^2 + \lambda \sum_{j=l+2}^{d} \gamma_j^2.$$

Wir teilen den Parametervektor $\boldsymbol{\gamma}$ auf in

$$\boldsymbol{\beta} = (\gamma_1, \ldots, \gamma_{l+1})^\top \quad \text{und} \quad \tilde{\boldsymbol{\gamma}} = (\gamma_{l+2}, \ldots, \gamma_d)^\top.$$

Dabei ist $\boldsymbol{\beta}$ der Parametervektor zum globalen Polynom und $\tilde{\boldsymbol{\gamma}}$ der Parametervektor zu den TP-Funktionen, die bestraft werden. Wir teilen auch die Design-Matrix entsprechend auf in

$$\boldsymbol{Z} = (\boldsymbol{X}|\boldsymbol{U}).$$

Dann kann das Modell geschrieben werden als

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{U}\tilde{\boldsymbol{\gamma}} + \boldsymbol{\varepsilon} \quad \text{mit } \boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I}) \tag{4.1.14}$$

und wir haben eine Mixed-Model Formulierung, wenn wir die Parameter $\boldsymbol{\beta}$ als fix und die Parameter $\tilde{\boldsymbol{\gamma}}$ als zufällig mit $\tilde{\boldsymbol{\gamma}} \sim N(\boldsymbol{0}, \tau^2\boldsymbol{I})$ ansehen.

Gehen wir von einer B-Spline-Basis mit Difference-Penalty aus ist die Darstellung als Mixed-Model etwas komplizierter. Hier tritt das Problem auf, dass die Strafmatrix $\boldsymbol{P}$ meist ein Rangdefizit hat und nicht invertierbar ist. Der direkte Ansatz, bei dem wir

$$\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I}), \quad \boldsymbol{\gamma} \sim N(\boldsymbol{0}, \tau^2\boldsymbol{P})$$

annehmen würden, funktioniert also nicht. Die Verteilung von $\boldsymbol{\gamma}$ ist degeneriert und es sei $\operatorname{rk}(\boldsymbol{P}) = r < d$. Wir können aber eine Aufteilung von $\boldsymbol{\gamma}$ gemäß dem Satz in Anhang A.6 vornehmen und führen dazu folgende Transformation durch:

$$\boldsymbol{\gamma} = (\tilde{\boldsymbol{X}}|\tilde{\boldsymbol{U}}) \begin{pmatrix} \boldsymbol{\beta} \\ \tilde{\boldsymbol{\gamma}} \end{pmatrix} = \tilde{\boldsymbol{X}}\boldsymbol{\beta} + \tilde{\boldsymbol{U}}\tilde{\boldsymbol{\gamma}}.$$

Dabei ist $\tilde{\boldsymbol{\gamma}}$ ein multivariat normalverteilter Zufallsvektor mit $\tilde{\boldsymbol{\gamma}} \sim N_r(\boldsymbol{0}, \tau^2\tilde{\boldsymbol{U}}^\top\boldsymbol{P}\tilde{\boldsymbol{U}})$. Als Spalten von $\tilde{\boldsymbol{X}}$ wählen wir eine Basis des Kerns von $\boldsymbol{P}$, woraus

$$\tilde{\boldsymbol{X}}^\top\boldsymbol{P} = \boldsymbol{0}$$

folgt. Die Matrix $\tilde{\boldsymbol{U}}$ bestimmen wir mithilfe der Spektralzerlegung von $\boldsymbol{P}$, welche existiert, da wir $\boldsymbol{P}$ als symmetrisch vorausgesetzt haben. Es sei

$$\boldsymbol{P} = \boldsymbol{\Gamma}\boldsymbol{\Omega}_+\boldsymbol{\Gamma}^\top,$$

wobei $\boldsymbol{\Omega}_+$ eine Diagonalmatrix mit den positiven Eigenwerten von $\boldsymbol{P}$ und $\boldsymbol{\Gamma}$ die orthonormale Matrix bestehend aus den korrespondierenden Eigenvektoren ist. Definieren wir nun

$$\tilde{\boldsymbol{U}} = \boldsymbol{L}(\boldsymbol{L}^\top\boldsymbol{L})^{-1}, \quad \text{mit } \boldsymbol{L} = \boldsymbol{\Gamma}\boldsymbol{\Omega}_+^{1/2},$$

dann folgt daraus

$$\begin{aligned}\tilde{\boldsymbol{U}}^\top\boldsymbol{P}\tilde{\boldsymbol{U}} &= (\boldsymbol{L}^\top\boldsymbol{L})^{-1}\boldsymbol{L}^\top\boldsymbol{P}\boldsymbol{L}(\boldsymbol{L}^\top\boldsymbol{L})^{-1} \\ &= (\boldsymbol{L}^\top\boldsymbol{L})^{-1}\boldsymbol{L}^\top\boldsymbol{L}\boldsymbol{L}^\top\boldsymbol{L}(\boldsymbol{L}^\top\boldsymbol{L})^{-1} = \boldsymbol{I}_r.\end{aligned}$$

Wenn wir $\boldsymbol{P} = \boldsymbol{D}_2^\top\boldsymbol{D}_2$ als Strafmatrix verwenden, können wir auch $\boldsymbol{L} = \boldsymbol{D}_2^\top$ definieren und wir erhalten auch $\tilde{\boldsymbol{U}}^\top\boldsymbol{P}\tilde{\boldsymbol{U}} = \boldsymbol{I}_r$. Der Vektor $\tilde{\boldsymbol{\gamma}}$ ist also ein iid-Zufallsvektor und es gilt

$$\tilde{\boldsymbol{\gamma}} \sim N_r(\boldsymbol{0}, \tau^2\boldsymbol{I}_r).$$

Das Modell kann dann als Mixed-Model geschrieben werden, denn

$$\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon} = \boldsymbol{Z}(\tilde{\boldsymbol{X}}\boldsymbol{\beta} + \tilde{\boldsymbol{U}}\tilde{\boldsymbol{\gamma}}) + \boldsymbol{\varepsilon} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{U}\tilde{\boldsymbol{\gamma}} + \boldsymbol{\varepsilon}.$$

Für den Strafterm folgt

$$\begin{aligned}\boldsymbol{\gamma}^\top\boldsymbol{P}\boldsymbol{\gamma} &= (\tilde{\boldsymbol{X}}\boldsymbol{\beta} + \tilde{\boldsymbol{U}}\tilde{\boldsymbol{\gamma}})^\top\boldsymbol{P}(\tilde{\boldsymbol{X}}\boldsymbol{\beta} + \tilde{\boldsymbol{U}}\tilde{\boldsymbol{\gamma}}) \\ &= \boldsymbol{\beta}^\top\underbrace{\tilde{\boldsymbol{X}}^\top\boldsymbol{P}}_{=\boldsymbol{0}}\tilde{\boldsymbol{X}}\boldsymbol{\beta} + 2\boldsymbol{\beta}^\top\underbrace{\tilde{\boldsymbol{X}}^\top\boldsymbol{P}}_{=\boldsymbol{0}}\tilde{\boldsymbol{U}}\tilde{\boldsymbol{\gamma}} + \tilde{\boldsymbol{\gamma}}^\top\underbrace{\tilde{\boldsymbol{U}}^\top\boldsymbol{P}\tilde{\boldsymbol{U}}}_{=\boldsymbol{I}_r}\tilde{\boldsymbol{\gamma}} \\ &= \tilde{\boldsymbol{\gamma}}^\top\tilde{\boldsymbol{\gamma}}.\end{aligned}$$

Sowohl für eine TP-Basis als auch für eine B-Spline-Basis ist also eine Darstellung des nichtparametrischen Modells (4.1.2) als Mixed-Model möglich und das PLS-Kriterium des Mixed-Models hat in beiden Fällen die Form

$$\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{U}\tilde{\boldsymbol{\gamma}}\|^2 + \lambda\tilde{\boldsymbol{\gamma}}^\top\tilde{\boldsymbol{\gamma}}. \tag{4.1.15}$$

Für ein Mixed-Model mit Formulierung wie in (4.1.14) können die Koeffizienten durch Minimierung des PLS-Kriteriums

$$\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{U}\tilde{\boldsymbol{\gamma}}\|^2 + \frac{\sigma^2}{\tau^2}\tilde{\boldsymbol{\gamma}}^\top\tilde{\boldsymbol{\gamma}} \tag{4.1.16}$$

geschätzt werden, siehe (4.1.13). Vergleichen wir nun die beiden Kriterien (4.1.15) und (4.1.16) sehen wir, dass wir den Glättungsparameter als Verhältnis der Fehlervarianz σ^2 und der Varianz des zufälligen Effekts τ^2 sehen können. Nun können wir entweder ML- oder REML-Schätzung verwenden um $\hat{\sigma}^2$ und $\hat{\tau}^2$ zu erhalten. Dann ist:

$$\hat{\lambda} = \frac{\hat{\sigma}^2}{\hat{\tau}^2}.$$

Für Details zur Schätzung des Glättungsparameters mittels ML- oder REML-Methode siehe Wood (2011).

4.2. Additive Modelle

In Abschnitt 4.1 haben wir nichtparametrische Modelle für eine stetige Response-Variable mit einer stetigen erklärenden Variablen betrachtet. Die Ideen aus diesem Abschnitt sollen hier für mehrere erklärende Variable erweitert werden. Es seien $(y_i, z_{i1}, \ldots, z_{iq}, x_{i1}, \ldots, x_{ip})$ für $i = 1, \ldots, n$ gegeben. Dabei bezeichnet y_i die Beobachtungen, $z_{i1}, \ldots, z_{iq}$ sind stetige erklärende Variablen, die nichtparametrisch ins Modell eingehen sollen, und $x_{i1}, \ldots, x_{ip}$ sind erklärende Variablen, die über einen linearen Prädiktor ins Modell einfließen. Ein Modell der Form

$$y_i = f_1(z_{i1}) + \cdots + f_q(z_{iq}) + \boldsymbol{x}_i^\top \boldsymbol{\beta} + \varepsilon_i \tag{4.2.1}$$

nennen wir **semiparametrisches Modell**. Es gelte wieder $\mathbb{E}[\varepsilon_i] = 0$ und $\mathbb{V}\mathrm{ar}[\varepsilon_i] = \sigma^2$, woraus

$$\mathbb{E}[y_i] = \mu_i = f_1(z_{i1}) + \cdots + f_q(z_{iq}) + \boldsymbol{x}_i^\top \boldsymbol{\beta} \quad \text{und} \quad \mathbb{V}\mathrm{ar}[y_i] = \sigma^2$$

folgt. Modelle ohne zusätzlichen parametrischen Teil heißen in der Literatur typischerweise **additive Modelle**,

$$y_i = f_1(z_{i1}) + \cdots + f_q(z_{iq}) + \varepsilon_i. \tag{4.2.2}$$

Wir werden sowohl Modelle der Form (4.2.1) als auch der Form (4.2.2) als additive Modelle bezeichnen.

Bei Modellen mit mehreren additiven Glättungstermen kommt es zu einem Identifizierbarkeitsproblem. Addieren wir beispielsweise eine Konstante $C \neq 0$ zur Funktion $f_1(z_1)$ und subtrahieren C von der Funktion $f_2(z_2)$ resultiert derselbe Prädiktor, denn

$$f_1(z_{i1}) + f_2(z_{i2}) = f_1(z_{i1}) + C + f_2(z_{i2}) - C.$$

Die Funktionen werden deshalb typischerweise um Null zentriert, d. h.

$$\sum_{i=1}^{n} f_1(z_{i1}) = \cdots = \sum_{i=1}^{n} f_q(z_{iq}) = 0.$$

Für die Funktionen f_k, $k = 1, \ldots, q$, werden wieder passende Basisfunktionen gewählt, sodass

$$f_k(z_{ik}) = \sum_{j=1}^{d_k} \gamma_{kj} B_j(z_{ik}).$$

Als Basisfunktionen können z. B. die TP-Basis oder die B-Spline-Basis gewählt werden und es ist sogar möglich verschiedene Basisfunktionen für die verschiedenen Glättungsterme zu wählen. Es sei

$$\boldsymbol{Z}_k = \begin{pmatrix} B_1(z_{1k}) & \dots & B_{d_k}(z_{1k}) \\ \vdots & & \vdots \\ B_1(z_{nk}) & \dots & B_{d_k}(z_{nk}) \end{pmatrix},$$

die k-te Design-Matrix und $\boldsymbol{\gamma}_k = (\gamma_{k1}, \dots, \gamma_{kd_k})^\top$ der k-te Parametervektor. Dann kann die Funktion $\boldsymbol{f}_k = (f_k(z_{1k}), \dots, f_k(z_{nk}))^\top$ dargestellt werden als

$$\boldsymbol{f}_k = \boldsymbol{Z}_k \boldsymbol{\gamma}_k$$

und es folgt das additive Modell in Matrixschreibweise

$$\boldsymbol{y} = \boldsymbol{Z}_1 \boldsymbol{\gamma}_1 + \dots + \boldsymbol{Z}_q \boldsymbol{\gamma}_q + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

Jeder Glättungsterm erhält einen eigenen Strafterm

$$\lambda_k \boldsymbol{\gamma}_k^\top \tilde{\boldsymbol{P}}_k \boldsymbol{\gamma}_k \quad \text{für } k = 1, \dots, q$$

wobei die Strafmatrizen $\tilde{\boldsymbol{P}}_k$ wie in Abschnitt 4.1.2 abhängig von der gewählten Basis sind. Wir minimieren die pönalisierte Fehlerquadratsumme

$$\|\boldsymbol{y} - \boldsymbol{Z}_1 \boldsymbol{\gamma}_1 + \dots + \boldsymbol{Z}_q \boldsymbol{\gamma}_q + \boldsymbol{X}\boldsymbol{\beta}\|^2 + \sum_{k=1}^{q} \lambda_k \boldsymbol{\gamma}_k^\top \tilde{\boldsymbol{P}}_k \boldsymbol{\gamma}_k. \tag{4.2.3}$$

Wir setzen $\boldsymbol{P}_k = \lambda_k \tilde{\boldsymbol{P}}_k$ und definieren die Strafmatrix des gesamten Modells als Blockmatrix

$$\boldsymbol{P} = \begin{pmatrix} \boldsymbol{P}_1 & \boldsymbol{0} & \dots & \dots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{P}_2 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \boldsymbol{P}_q & \boldsymbol{0} \\ \boldsymbol{0} & \dots & \dots & \boldsymbol{0} & \boldsymbol{0} \end{pmatrix}.$$

Es sei $\boldsymbol{Z} = (\boldsymbol{Z}_1 | \dots | \boldsymbol{Z}_q | \boldsymbol{X})$ die Design-Matrix und $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^\top, \dots, \boldsymbol{\gamma}_q^\top, \boldsymbol{\beta}^\top)^\top$ der Parametervektor des gesamten Modells. Es ist d_k die Dimension der Basis des k-ten Glättungsterms und p die Anzahl der Parameter des parametrischen Teils des Modells. Dann hat das gesamte additive Modell

$$r = d_1 + \dots + d_q + p$$

Parameter, die Design-Matrix $\boldsymbol{Z}$ ist eine $(n \times r)$-Matrix und die Strafmatrix $\boldsymbol{P}$ ist eine $(r \times r)$-Matrix. Man kann (4.2.3) äquivalent schreiben als

$$\|\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\gamma}\|^2 + \boldsymbol{\gamma}^\top \boldsymbol{P}\boldsymbol{\gamma}.$$

Dies hat nun ähnliche Form wie (4.1.4) und kann mittels Penalized-Least-Squares-Schätzung gelöst werden. Hier sind die λ-Parameter schon in der Matrix $\boldsymbol{P}$ enthalten.

4.3. Freiheitsgrade des Modells

Wir wollen uns nun überlegen, wie viele Freiheitsgrade unser Modell hat. Vergleiche dazu Wood (2006). Bei linearen Regressionsmodellen $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$ mit einer $(n\times p)$-Design-Matrix $\boldsymbol{X}$ war die Anzahl der Parameter gleich der Spur der Hat-Matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$, nämlich

$$\operatorname{tr}(\boldsymbol{H}) = p.$$

Das Schätzen eines nichtparametrischen Modells mit q Glättungstermen entspricht dem Schätzen eines linearen Modells mit $d_1 + \cdots + d_k + p$ (vielen) Parametern, wie wir schon gesehen haben. Durch die Bestrafung von zu verwackelten Funktionen verringert sich die Anzahl der Parameter aber. Als Analogon zur Hat-Matrix verwenden wir die Glättungsmatrix $\boldsymbol{S} = \boldsymbol{Z}(\boldsymbol{Z}^\top\boldsymbol{Z} + \boldsymbol{P})^{-1}\boldsymbol{Z}^\top$, um die sogenannten **effektiven Freiheitsgrade** eines additiven Modells zu definieren,

$$\mathrm{edf} = \operatorname{tr}(\boldsymbol{S}).$$

Bemerkung. Die $(n \times n)$-Matrix $\boldsymbol{S}$ muss nicht explizit ausgerechnet werden, um die Spur zu bestimmen, denn wegen

$$\begin{aligned}\operatorname{tr}(\boldsymbol{S}) &= \operatorname{tr}(\boldsymbol{Z}(\boldsymbol{Z}^\top\boldsymbol{Z} + \boldsymbol{P})^{-1}\boldsymbol{Z}^\top) \\ &= \operatorname{tr}(\boldsymbol{Z}^\top\boldsymbol{Z}(\boldsymbol{Z}^\top\boldsymbol{Z} + \boldsymbol{P})^{-1})\end{aligned}$$

kann die Spur auch mithilfe des Produkts zweier kleinerer $(r \times r)$-Matrizen bestimmt werden. Für Rechenregeln bzgl. der Spur siehe Anhang A.9.

Wir können die effektiven Freiheitsgrade des additiven Modells aufteilen in effektive Freiheitsgrade zu den einzelnen Parametern. Wir definieren dazu die $(r \times n)$-Matrix

$$\boldsymbol{Q} = (\boldsymbol{Z}^\top\boldsymbol{Z} + \boldsymbol{P})^{-1}\boldsymbol{Z}^\top.$$

Dann ist

$$\operatorname{tr}(\boldsymbol{S}) = \operatorname{tr}(\boldsymbol{Z}\boldsymbol{Q}).$$

Es sei $\boldsymbol{Q}_l^0$ eine $(r\times n)$-Matrix, in der alle bis auf die l-te Zeile der Matrix $\boldsymbol{Q}$ auf Null gesetzt wurden. Damit folgt

$$\begin{aligned}\operatorname{tr}(\boldsymbol{ZQ}) &= \operatorname{tr}(\boldsymbol{Z}(\boldsymbol{Q}_1^0 + \cdots + \boldsymbol{Q}_r^0)) \\ &= \operatorname{tr}\left(\sum_{l=1}^{r} \boldsymbol{ZQ}_l^0\right) \\ &= \sum_{l=1}^{r} \operatorname{tr}(\boldsymbol{ZQ}_l^0).\end{aligned}$$

Des Weiteren ist

$$\operatorname{tr}(\boldsymbol{ZQ}_l^0) = (\boldsymbol{QZ})_{ll}.$$

Wir sehen also $\operatorname{tr}(\boldsymbol{ZQ}_l^0)$ als effektiven Freiheitsgrad zum l-ten Parameter für alle $l = 1, \ldots, r$ und berechnen diesen als l-tes Diagonalelemente der Matrix $\boldsymbol{QZ}$.

4.4. Generalisierte additive Modelle

Additive Modelle können erweitert werden, sodass die Responses aus der Exponentialfamilie stammen können. Es seien die Responses

$$y_i \overset{ind}{\sim} \text{ Exponantialfamile}(\theta_i), \quad i = 1, \ldots, n$$

und der additive Prädiktor sei

$$\eta_i = f_1(z_{i1}) + \cdots + f_q(z_{iq}) + \boldsymbol{x}_i^\top \boldsymbol{\beta}.$$

Die monotone, differenzierbare Link-Funktion g verbindet dies mittels

$$\eta_i = g(\mu_i),$$

wobei $\mu_i = \mathbb{E}[y_i]$ ist. Man nennt solche Modelle **generalisierte additive Modelle** (GAM) und diese haben die Form

$$g(\mu_i) = f_1(z_{i1}) + \cdots + f_q(z_{iq}) + \boldsymbol{x}_i^\top \boldsymbol{\beta}, \tag{4.4.1}$$

wobei die einzelnen Glättungsfunktionen wieder die Form

$$f_k(z_{ik}) = \sum_{j=1}^{d_k} \gamma_{kj} B_j(z_{ik})$$

haben. Das Modell in Matrixschreibweise ist dann

$$g(\boldsymbol{\mu}) = \boldsymbol{Z}_1\boldsymbol{\gamma}_1 + \cdots + \boldsymbol{Z}_q\boldsymbol{\gamma}_q + \boldsymbol{X}\boldsymbol{\beta}.$$

Wenn wir $\boldsymbol{Z} = (\boldsymbol{Z}_1 | \ldots | \boldsymbol{Z}_q | \boldsymbol{X})$ und $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^\top, \ldots, \boldsymbol{\gamma}_d^\top, \boldsymbol{\beta}^\top)^\top$ setzen erhalten wir ein GLM der gewohnten Form

$$g(\boldsymbol{\mu}) = \boldsymbol{Z}\boldsymbol{\gamma}.$$

Wir maximieren hier die pönalisierte Log-Likelihood-Funktion

$$l(\boldsymbol{\gamma}, \boldsymbol{y}) - \frac{1}{2}\sum_{k=1}^{q} \lambda_k \boldsymbol{\gamma}_k^\top \boldsymbol{P}_k \boldsymbol{\gamma}_k,$$

wobei $l(\boldsymbol{\gamma}, \boldsymbol{y})$ die Log-Likelihood-Funktion des GLMs ist und wir noch den Strafterm abziehen. Gelöst wird dies mit einem pönalisierten IRLS-Verfahren, welches wie in Anhang A.2 funktioniert. Wir müssen nur den Strafterm in der Iterationsvorschrift beachten.

5. Modellierung in R

In Zusammenhang mit dem Silikonregister sind wir an Modellen für das Auftreten einer Kapselkontraktur interessiert. Dabei können verschiedene Fragestellungen betrachtet werden. Je nach Fragestellung verwenden wir dann Response-Variablen mit einer bestimmten Anzahl an Stufen und versuchen herauszufinden welche erklärenden Variablen Einfluss haben.

Fragestellung 1: Wie groß sind die Wahrscheinlichkeiten für die einzelnen Baker-Stufen im Falle einer Kapselkontraktur?

Zur Beantwortung dieser Frage suchen wir ein Modell für die Variable `kontr` mit den Stufen {`BakerI`, `BakerII`, `BakerIII`, `BakerIV`}. Es werden nur Patientinnen bei denen eine Kapselkontraktur aufgetreten ist betrachtet. Ist bei einer Patientin keine Kapselkontraktur aufgetreten, dann steht `NA` an der entsprechenden Stelle in `kontr` und diese Zeile wird nicht zur Modellierung herangezogen. Der Stichprobenumfang ist in diesem Fall `n=1949`.

Fragestellung 2: Ist eine Kapselkontraktur eingetreten oder nicht?

Hier suchen wir ein Modell für die binäre Variable `kontr_flag`. Diese hat die zwei Stufen {`yes`, `no`}, wobei in der Stufe `yes` alle Patientinnen mit einer Kapselkontraktur auf einer der vier Baker-Stufen zusammengefasst sind. Hier werden alle Zeilen des Datensatzes zur Modellierung verwendet und der Stichprobenumfang ist `n=3534`.

Man kann auch versuchen beide Fragestellungen gleichzeitig zu beantworten und kommt dann zu folgender Fragestellung.

Fragestellung 3: Ist eine Kapselkontraktur eingetreten und wenn ja wie stark ist diese?

Dazu verwandeln wir die `NA`-Einträge in der Variable `kontr` zu einer neuen Stufe `no` (keine Kapselkontraktur ist aufgetreten) und erhalten eine neue Response-Variable mit 5 Stufen, die wir `kontr5` nennen. Der Stichprobenumfang ist auch hier `n=3534`.

Zur Modellfindung wird jeweils das Paket `mgcv` verwendet. Im Anschluss daran werden die Ergebnisse für die gefundenen Modelle mit den Schätzungen eines zweiten Pakets, `VGAM`, verglichen. Es wurde `R Version 3.0.3`, sowie `mgcv 1.8-4.` und `VGAM 0.9-5` verwendet.

5.1. Softwarepaket mgcv

Das `mgcv`-Paket (**M**ixed **G**AM **C**omputation **V**ehicle) ist ein Softwarepaket von Simon N. Wood zur Schätzung von generalisierten additiven Modellen und generalisierten additiven Mischmodellen (Mixed-Models). Besonders ist, dass eine automatische Schätzung

des Glättungsparameters möglich ist. Einen schnellen Überblick kann man sich in Wood (2001) verschaffen. Für theoretische Details sowie eine praktische Einführung mit vielen Beispielen siehe Wood (2006).

Zum Schätzen der Parameter in den Modellen verwenden wir die Funktion `gam` aus dem `mgcv`-Paket. Diese funktioniert im Wesentlichen gleich wie die Funktion `glm`. Es wird eine Formel übergeben, die das Modell spezifiziert, welche aber im Unterschied zur `glm`-Funktion nun auch Glättungsterme für stetige Variablen enthalten darf. Für eine stetige erklärende Variable `z` kann ein Glättungsterm mittels `s(z)` in den Prädiktor inkludiert werden. Dabei werden standardmäßig sogenannte Thin-Plate-Regression-Splines zur Basiskonstruktion verwendet. Diese sind allerdings rechnerisch aufwändig. Wir benützen deshalb kubische Splines und setzen dazu `bs='cr'` in der Funktion `s()`. Kubische Splines können nur bei der univariaten Glättung angewendet werden. Zur multivariaten Glättung mittels Tensor-Produkten könnte der Glättungsterm `te()` verwendet werden. Es ist möglich verschiedene Glättungsterme, die sich in der Konstruktion der Basis unterscheiden, für verschiedene erklärende Variablen einzusetzen.

Im `family`-Argument wird die Verteilung unserer Response angegeben. Wir werden `binomial` für das Modell mit der binären Response `kontr_flag` und `ocat` (ordered categories) für die Modelle mit den ordinalen Responses `kontr` und `kontr5` verwenden.

Mit dem `method`-Argument der Funktion `gam` können wir angeben welche Methode zur Schätzung der Glättungsparameter verwendet werden soll. Zur Auswahl stehen u. a. die Default-Methode `'REML'` für Restricted-Maximum-Likelihood-Schätzung, `'ML'` für Maximum-Likelihood-Schätzung sowie `'GCV.cp'` für generalisierte Cross-Validation. Wir werden stets die ML-Methode auswählen, da diese laut Wood (2015) (vgl. Erklärung zur Funktion `summary.gam`) in Simulationen die besten Ergebnisse und die zuverlässigsten p-Werte liefert. Wir werden p-Werte aus dem `summary` und `anova` verwenden. Diese sind approximativ und eher zu klein. Für Glättungsterme liefern die beiden dasselbe Ergebnis. Es wird ein Test auf Gleichheit zu Null für die einzelnen Glättungsterme durchgeführt, welcher auf dem approximativen Verteilungsresultat

$$\hat{f}(\boldsymbol{z}) \sim N(f(\boldsymbol{z}), \boldsymbol{V}_f)$$

basiert, wobei $\boldsymbol{V}_f$ die Bayes'sche Kovarianzmatrix ist. Für Details siehe Wood (2013). Die p-Werte für die parametrischen Terme des Modells basieren auf dem Verteilungsresultat

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \mathbb{V}\mathrm{ar}(\hat{\boldsymbol{\beta}})),$$

womit Wald-Statistiken berechnet werden. Dabei wird wieder die Bayes'sche Kovarianzmatrix verwendet und die p-Werte sind wieder approximativ. Für Faktoren sollte unbedingt das Ergebnis in `anova` (anstatt in `summary`) verwendet werden, da im `summary` nur p-Werte für die Abweichung zur Referenzstufe dargestellt sind. Man beachte auch, dass die p-Werte im `anova`-Output für die Funktion `gam` anders zu interpretieren sind als wir es für die Funktion `glm` gewohnt sind. Dort wurden p-Werte zum Vergleich von verschachtelten Modellen

berechnet, die sequenziell zu interpretieren waren. Für die Funktion gam werden p-Werte für den Test auf Relevanz der einzelnen Terme zusätzlich zu allen anderen enthaltenen Termen berechnet. Des Weiteren sollte der anova-Befehl für gam nur auf ein einzelnes Modell angewandt werden, da beim Vergleich zweier Modelle mittels anova die p-Werte nicht sehr zuverlässig sind.

Modellfindung allgemein

Die allgemeine Vorgehensweise bei der Modellfindung ist dieselbe für alle drei Fragestellungen. Wir beginnen mit einem Modell mit Glättungsterm für die Dauer dau, wobei kubische Splines als Basis verwendet werden und die Basisdimension passend gewählt wird. Zur Glättungsparameterschätzung verwenden wir die ML-Methode. Dann testen wir, ob die zweite stetige Variable, das Volumen vol, einen signifikanten Einfluss hat und wenn ja, wie dieses am besten ins Modell eingehen soll. Danach fügen wir zunächst einzeln die restlichen Prädiktoren, die nun alle Faktoren sind, hinzu. Wir testen welche zusätzlich zur Dauer und ggf. zum Volumen relevant sind und notieren die dazugehörigen p-Werte. Jene mit signifikantem p-Wert sind unsere Kandidaten für das Modell. Nun inkludieren wir die Kandidaten aufsteigend nach ihrem p-Wert und prüfen, ob sie zusätzlich zu diversen zuvor hinzugefügten Prädiktoren immer noch relevant sind.

5.1.1. Modell für die ordinalen Baker-Stufen

Wir starten bei der Modellfindung für die Variable kontr mit einem Modell, welches nur einen Glättungsterm für die Variable dau enthält. Die Response-Variable ist ein geordneter Faktor mit den vier Stufen {BakerI, BakerII, BakerIII, BakerIV}, wofür im mgcv-Paket die Funktion ocat verfügbar ist. Im Argument R der Funktion ocat muss man die Anzahl der Kategorien angeben. Die Response wird als numerischer Vektor übergeben, wobei die einzelnen Kategorien mit {1,2,3,4} codiert sind. Damit die originalen Daten erhalten bleiben und es zu keinen Verwechslungen kommt, erzeugen wir den Vektor num.kontr, den wir dann der Funktion gam als Response übergeben können.

```
library(plyr)
num.kontr <- as.numeric(revalue(kontr, c("BakerI"="1", "BakerII"="2", "BakerIII"=
    "3", "BakerIV"="4")))
```

Wir passen ein Modell mit Glättungsterm für die Dauer an, wobei wir kubische Splines als Basis verwenden. Standardmäßig ist die maximale Dimension der Basis 10 und wir bleiben anfangs dabei. Durch Setzen des Arguments k in der Funktion s() könnte man dies ändern.

```
mod.dau <- gam(num.kontr ~ s(dau, bs='cr'), family=ocat(R=4), method='ML')
summary(mod.dau)

Family: Ordered Categorical(-1,0.5,2.39)
Link function: identity
```

```
Formula:
num.kontr ~ s(dau, bs = "cr")

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.10982    0.04203   26.41   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
         edf Ref.df Chi.sq p-value
s(dau) 3.675  4.382    229  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Deviance explained =   40%
-ML = 2440.2  Scale est. = 1         n = 1949
```

Das Ergebnis des `summary`-Befehls liefert einige Informationen. Es werden die Schwellenwerte (Intercepts) des kumulativen Modells angezeigt, wobei der erste Wert immer auf -1 festgesetzt und die restlichen Werte geschätzt werden. Wir erhalten Parameterschätzer, Standardfehler und approximative p-Werte für die erklärenden Variablen im parametrischen Teil des Modells (in diesem Fall nur der Intercept-Parameter). Für den Glättungsterm werden die effektiven Freiheitsgrade und der approximative p-Wert beim Test auf Relevanz des Terms angezeigt. Des Weiteren erhalten wir noch den Prozentsatz der vom Modell erklärten Deviance, den (negativen) Wert des ML-Kriteriums an der Optimalstelle, den Scale-Parameter (der in unserem Fall immer 1 ist, prinzipiell aber auch geschätzt werden könnte) sowie den Stichprobenumfang.

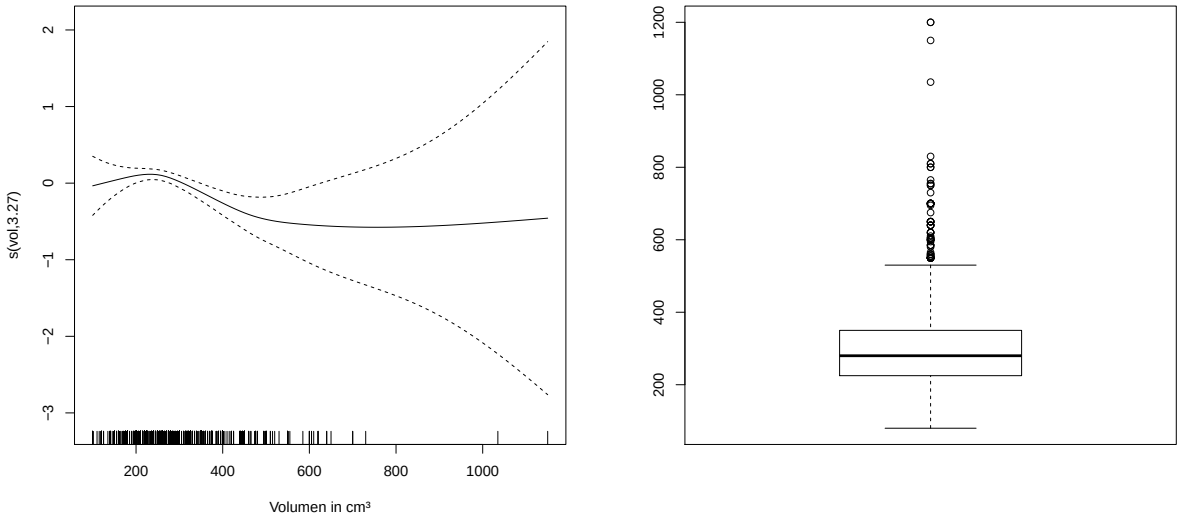
```
anova(mod.dau)

Family: Ordered Categorical(-1,0.5,2.39)
Link function: identity

Formula:
num.kontr ~ s(dau, bs = "cr")

Approximate significance of smooth terms:
         edf Ref.df Chi.sq p-value
s(dau) 3.675  4.382    229  <2e-16
```

Auch der `anova`-Befehl liefert uns die Schwellenwerte des Modells, sowie effektive Freiheitsgrade und den p-Wert für den Glättungsterm der Variable Dauer. Dieser ist höchst signifikant und ist auf alle Fälle in unserem Modell enthalten. Die effektiven Freiheitsgrade sind ungefähr 3.68, weshalb wir bei der maximalen Basisdimension von `k=10` bleiben.

Im nächsten Schritt versuchen wir den Einfluss des Füllvolumens `vol` herauszufinden. Dazu fügen wir zunächst einen Glättungsterm für das Volumen zu unserem Modell hinzu.

```
mod.vol <- gam(num.kontr ~ s(dau, bs='cr') + s(vol, bs='cr'), family=ocat(R=4),
    method='ML')
```

Die effektiven Freiheitsgrade sind ungefähr 3.26, d. h. eine maximale Basisdimension von `k=10` passt auch hier. Der `anova`-Befehl liefert einen signifikanten p-Wert von 0.00408 und wir wissen nun, dass das Volumen zusätzlich zur Dauer Einfluss auf die Baker-Stufe hat. Abbildung 5.1(a) zeigt die geschätzte Funktion für das Volumen. Es fällt sofort die Trompetenform des Konfidenzbandes auf. Dieses ist relativ eng für Füllvolumen bis etwa 400 cm^3 und wird dann immer breiter. Das liegt daran, dass es für größer werdendes Volumen immer weniger Beobachtungen gibt, wie man im Boxplot (Abbildung 5.1(b)) erkennen kann. Beispielsweise haben nur vier Patientinnen ein Volumen von über 1000 cm^3 (zwei Patientinnen mit 1200 cm^3). Dies hat die extreme Trompetenform zur Folge.



(a) Geschätzte Funktion für das Volumen

(b) Boxplot für das Volumen

Abbildung 5.1.: Schätzung für das Modell für die ordinalen Baker-Stufen mit Glättungstermen für Dauer und Volumen.

Wir testen nun den Einfluss des Volumens in einem Modell in dem `vol` parametrisch eingeht. Die Form der geschätzten nichtparametrischen Funktion legt einen kubischen Einfluss nahe. Allerdings ist auch eine Gerade im Konfidenzband enthalten. Wir testen auf einen linearen, quadratischen und kubischen Einfluss. In Tabelle 5.1 sind die p-Werte für die entsprechenden Tests dargestellt. Sowohl ein linearer als auch ein quadratischer oder kubischer Einfluss des Volumens sind gerechtfertigt. Der p-Wert für den linearen Einfluss ist mit etwa 0.2% sogar kleiner als jener beim Test auf einen nichtparametrischen Einfluss (etwa 0.4%). Zusätzlich zum linearen Einfluss ist weder der quadratische noch der kubi-

sche Term signifikant. Wir entscheiden uns, nicht zuletzt auch der Einfachheit halber, das Füllvolumen linear in den Prädiktor einfließen zu lassen.

Prädiktor	p-Wert für den letzten Term
`s(dau) + vol`	0.002
`s(dau) + vol`^{2}	0.004
`s(dau) + vol`^{3}	0.045
`s(dau) + vol + vol`^{2}	0.932
`s(dau) + vol + vol`^{3}	0.725

Tabelle 5.1.: p-Werte für Volumsterme in einem Modell für die ordinalen Baker-Stufen.

Die restlichen erklärenden Variablen sind alle Faktoren. Wir testen mittels `anova`, welche davon zusätzlich zur Dauer und zum Volumen signifikant sind. Die dazugehörigen p-Werte sind in Tabelle 5.2 dargestellt. Jene mit signifikantem p-Wert sind Kandidaten in das Modell aufgenommen zu werden, die anderen werden verworfen.

Prädiktor	p-Wert für den letzten Term
`s(dau) + vol + oberfl`	0.006
`s(dau) + vol + lumen`	0.490
`s(dau) + vol + fuel`	0.240
`s(dau) + vol + lage`	0.921
`s(dau) + vol + opzug`	0.169
`s(dau) + vol + antib`	0.188
`s(dau) + vol + ster`	0.425
`s(dau) + vol + drain`	$1.89 \cdot 10^{-6}$
`s(dau) + vol + prim`	0.271

Tabelle 5.2.: p-Werte für erklärende Faktoren im Modell für die ordinalen Baker-Stufen.

Demnach bleiben `drain` und `oberfl` als Kandidaten für unser Modell übrig. Wir fügen zuerst `drain` hinzu, da der p-Wert viel kleiner ist und testen ob die Oberflächenbeschaffenheit `oberfl` zusätzlich noch immer signifikant ist. Dies ist der Fall und das finale Modell enthält neben dem Glättungsterm für die Dauer auch die stetige erklärende Variable `vol`, sowie die Faktoren `drain` und `oberfl`.

```
mod.kontr <- gam(num.kontr ~ s(dau, bs='cr') + vol + drain + oberfl, family=ocat(
    R=4), method='ML')
```

Die geschätzte Funktion für die Dauer ist in Abbildung 5.2 abgebildet und der lineare Prädiktor hat folgende Form:

$$\begin{aligned}\hat{\eta} = \hat{f}(\texttt{dau}) + 0.31 - 0.0013\ \texttt{vol} - 0.64\ (\texttt{drain} = \texttt{n}) \\ + 1.14\ (\texttt{oberfl} = \texttt{Polyurethan}) + 1.29\ (\texttt{oberfl} = \texttt{texturiert}).\end{aligned}$$

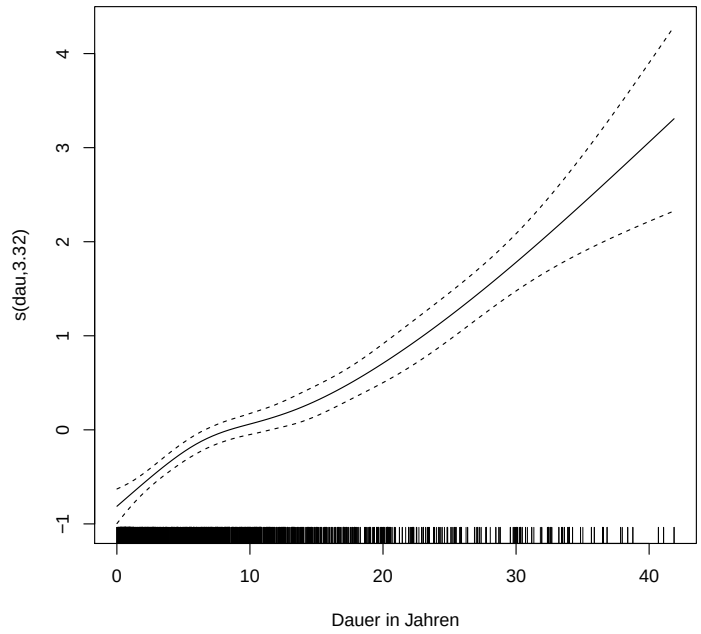


Abbildung 5.2.: Geschätzte Funktion für die Dauer unter dem finalen Modell für die ordinalen Baker-Stufen.

Für die Interpretation der Parameter ist die interne Definition des linearen Prädiktors wichtig. Wir haben diesen für ein kumulatives Logit-Modell wie in (3.2.7) definiert:

$$\text{logit}(\mathbb{P}[Y \leq j|\boldsymbol{x}]) = \alpha_j + \boldsymbol{x}^\top \boldsymbol{\beta}.$$

Im mgcv-Paket wird intern die Definition

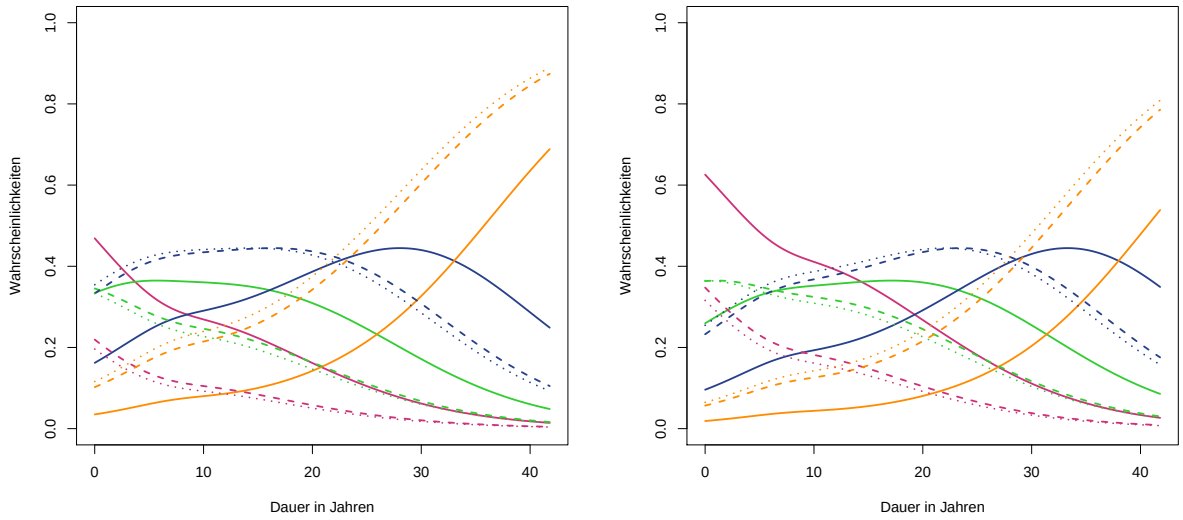
$$\text{logit}(\mathbb{P}[Y \leq j|\boldsymbol{x}]) = \alpha_j - \boldsymbol{x}^\top \boldsymbol{\beta}$$

verwendet. Demnach hat ein *negatives* Vorzeichen eines mittels mgcv geschätzten Parameters eine *Vergrößerung* des linearen Prädiktors und somit auch eine Vergrößerung der kumulativen Wahrscheinlichkeiten zur Folge. Da

$$\pi_1(\boldsymbol{x}) = \mathbb{P}[Y \leq 1|\boldsymbol{x}] \quad \text{und} \quad \pi_4(\boldsymbol{x}) = 1 - \mathbb{P}[Y \leq 3|\boldsymbol{x}]$$

gilt, können wir damit die Wahrscheinlichkeiten für Baker-Stufe 1 und Baker-Stufe 4 interpretieren. Eine Vergrößerung des linearen Prädiktors bewirkt eine Vergrößerung der Wahrscheinlichkeit für Baker-Stufe 1. Umgekehrt muss die Wahrscheinlichkeit für die Baker-Stufe 4 sinken. Über die Wahrscheinlichkeiten für die Baker-Stufen 2 und 3 kann mithilfe der Parameter nicht direkt etwas ausgesagt werden.

Die geschätzten Wahrscheinlichkeiten für die einzelnen Baker-Stufen in Abhängigkeit der Dauer sind für die verschiedenen Faktorkombinationen in Abbildung 5.3 dargestellt, wobei das Volumen auf den Median von 280 cm^3 fixiert wurde.



(a) Patientinnen mit Drainage

(b) Patientinnen ohne Drainage

Abbildung 5.3.: Geschätzte Wahrscheinlichkeiten für die Baker-Stufen bei medianem Volumen von 280 cm^3: `BakerI` (violett), `BakerII` (grün), `BakerIII` (blau), `BakerIV` (orange), Oberfläche `glatt` (durchgehende Linien), Oberfläche `Polyurethan` (strichlierte Linien), Oberfläche `texturiert` (punktierte Linien).

5.1.2. Binomiales Modell für `kontr_flag`

Bei der Modellfindung für `kontr_flag` gehen wir analog wie bei der Modellfindung für `kontr` vor. Die Response-Variable ist hier eine binäre Variable und wir geben im `family`-Argument der Funktion `gam` deshalb `binomial` an. Als Methode für die Glättungsparameterschätzung wählen wir wieder `ML` aus. Wir schätzen ein Modell mit einem Glättungsterm für die Variable Dauer:

```
mod.dau <- gam(kontr_flag ~ s(dau, bs='cr'), family=binomial, method='ML')
anova(mod.dau)

Family: binomial
Link function: logit

Formula:
kontr_flag ~ s(dau, bs = "cr")
```

```
Approximate significance of smooth terms:
         edf Ref.df Chi.sq p-value
s(dau) 5.530  6.343  117.3  <2e-16
```

Der anova-Befehl liefert einen höchst signifikanten p-Wert für den Glättungsterm der Dauer und die effektiven Freiheitsgrade sind etwa 5.53, weshalb wir auch hier bei der maximalen Basisdimension von k=10 bleiben.

Nun untersuchen wir den Einfluss des Volumens und betrachten zuerst ein nichtparametrisches Modell mit einem Glättungsterm für das Volumen zusätzlich zum Glättungsterm für die Dauer. Abbildung 5.4 zeigt die geschätzte Funktion für das Volumen. Deutlich ist wieder die Trompetenform des Konfidenzbandes zu erkennen. Im Unterschied zur geschätzten Funktion für das Volumen im Modell für kontr (Abbildung 5.1(a)) ist hier aber eine Gerade nicht im Konfidenzband enthalten. Trotzdem werden auch parametrische Modelle mit dem Volumen überprüft. Tabelle 5.3 zeigt die p-Werte für diese Tests.

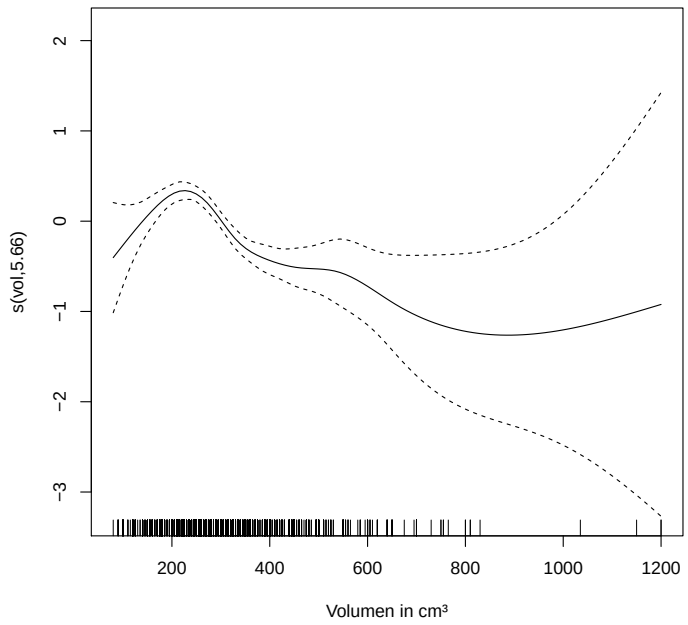


Abbildung 5.4.: Geschätzte Funktion für vol in einem binomialen Modell für die Auftrittswahrscheinlichkeiten einer Kapselkontraktur mit Glättungstermen für die Dauer und das Volumen.

Sowohl ein nichtparametrisches Modell als auch die parametrischen Modelle mit vol, $\texttt{vol}^2$ und $\texttt{vol}^3$ wären den p-Werten zufolge passend. Zusätzlich zum linearen Term sind $\texttt{vol}^2$ und $\texttt{vol}^3$ jeweils nicht signifikant. Wir entscheiden uns hier aufgrund des sehr kleinen p-Wertes (kleiner als alle anderen) und der Form des Konfidenzbandes in Abbildung 5.4 dafür, einen Glättungsterm für das Volumen in das Modell aufzunehmen.

Prädiktor	p-Wert für den letzten Term
`s(dau) + s(vol)`	$5.53 \cdot 10^{-15}$
`s(dau) + vol`	$2.97 \cdot 10^{-13}$
`s(dau) + vol`^{2}	$3.61 \cdot 10^{-11}$
`s(dau) + vol`^{3}	$4.6 \cdot 10^{-7}$
`s(dau) + vol + vol`^{2}	0.964
`s(dau) + vol + vol`^{3}	0.456

Tabelle 5.3.: p-Werte für Volumsterme in binomialen Modellen für die Auftrittswahrscheinlichkeiten einer Kapselkontraktur.

Nun überprüfen wir, welche der erklärenden Faktoren zusätzlich zu den beiden Glättungstermen für die Dauer und das Volumen noch signifikant sind und notieren die p-Werte in Tabelle 5.4. Demnach sind die Variablen `oberfl`, `fuel`, `lage`, `opzug` und `drain` die Kandidaten für unser Modell. Wir fügen diese nacheinander (nach aufsteigendem p-Wert) ins Modell hinzu und testen, ob sie zusätzlich zu schon enthaltenen Termen signifikant sind. Dies ist der Fall für alle Kandidaten. Wir schätzen das finale Modell mit:

```
mod.kontr_flag <- gam(kontr_flag ~ s(dau, bs='cr') + s(vol, bs='cr') +  drain +
    lage + opzug + oberfl + fuel, family=binomial, method='ML')
```

Prädiktor	p-Wert für den letzten Term
`s(dau) + s(vol) + oberfl`	0.006
`s(dau) + s(vol) + lumen`	1
`s(dau) + s(vol) + fuel`	0.019
`s(dau) + s(vol) + lage`	$2.55 \cdot 10^{-9}$
`s(dau) + s(vol) + opzug`	$2.84 \cdot 10^{-7}$
`s(dau) + s(vol) + antib`	1
`s(dau) + s(vol) + ster`	0.209
`s(dau) + s(vol) + drain`	$5.84 \cdot 10^{-15}$
`s(dau) + s(vol) + prim`	0.089

Tabelle 5.4.: p-Werte für erklärende Faktoren in Modellen für die Auftrittswahrscheinlichkeit einer Kapselkontraktur.

Abbildung 5.5 zeigt die geschätzten Funktionen für die Dauer und das Volumen. Der Prädiktor ist:

$$\begin{aligned}\hat{\eta} =& \hat{f}_1(\texttt{dau}) + \hat{f}_2(\texttt{vol}) + 0.76 - 0.73\ (\texttt{drain} = \texttt{n}) - 0.042\ (\texttt{lage} = \texttt{intermuskulär}) \\ &+ 0.32\ (\texttt{lage} = \texttt{subcutan}) + 0.58\ (\texttt{lage} = \texttt{subglandulär}) \\ &- 0.037\ (\texttt{lage} = \texttt{submuskulär}) + 1.09\ (\texttt{opzug} = \texttt{bestehendeNarbe}) \\ &+ 0.89\ (\texttt{opzug} = \texttt{inframammär}) + 0.81\ (\texttt{opzug} = \texttt{periareolär}) \\ &+ 1.45\ (\texttt{opzug} = \texttt{T-Schnitt}) - 0.30\ (\texttt{opzug} = \texttt{transareolär}) \\ &+ 0.69\ (\texttt{opzug} = \texttt{vertikal}) + 0.29\ (\texttt{oberfl} = \texttt{Polyurethan}) \\ &- 0.31\ (\texttt{oberfl} = \texttt{texturiert}) - 1.72\ (\texttt{fuel} = \texttt{gemischt}) \\ &- 1.31\ (\texttt{fuel} = \texttt{Hydrogel}) - 2.16\ (\texttt{fuel} = \texttt{Kochsalzlösung}) \\ &- 1.26\ (\texttt{fuel} = \texttt{Silikongel}).\end{aligned}$$

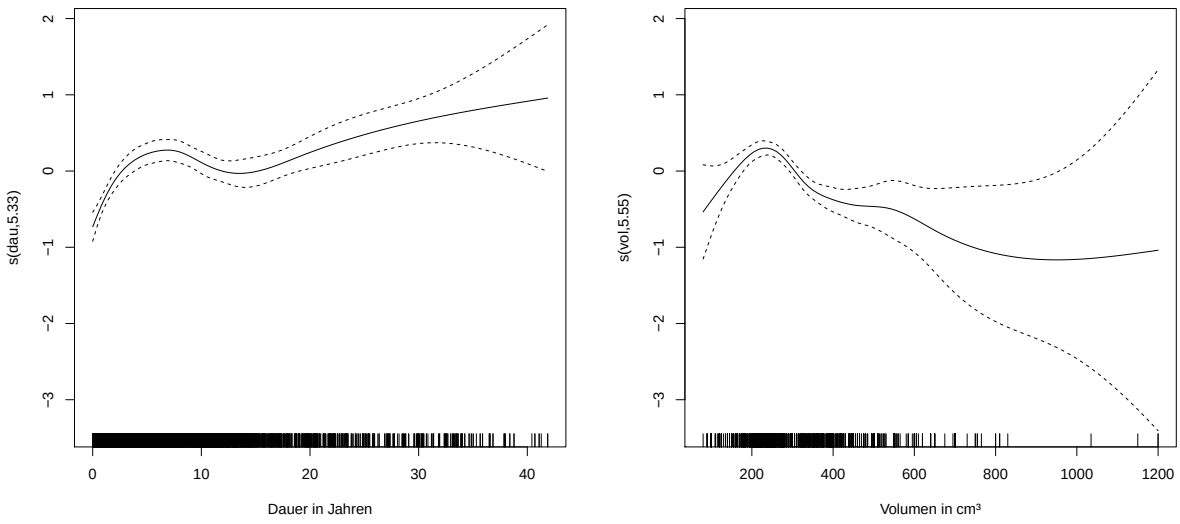


Abbildung 5.5.: Geschätzte Funktionen für `dau` und `vol` unter dem finalen Modell für die Auftrittswahrscheinlichkeit einer Kapselkontraktur.

Abbildung 5.6 zeigt die geschätzten Wahrscheinlichkeiten in Abhängigkeit der Dauer, wobei das Volumen auf den Median von 280 und die restlichen Faktoren jeweils auf die häufigste Stufe fixiert wurden. Die Wahrscheinlichkeit für eine Kapselkontraktur erhöht sich zunächst während der ersten 10 Jahre, um danach wieder etwas zu sinken. Nach ca. 15 Jahren steigt sie stetig an. Dieser Trend ist auch für andere Faktorkombinationen erkennbar. Die Linien für die geschätzten Wahrscheinlichkeiten sind annähernd parallel zu den Linien in Abbildung 5.6. Die Intercepts unterscheiden sich jedoch deutlich, d. h. je nach Faktorkombination ist ein gewisses Grundrisiko zu Beginn gegeben, welches sich entsprechend des oben beschriebenen Trends im Laufe der Zeit entwickelt.

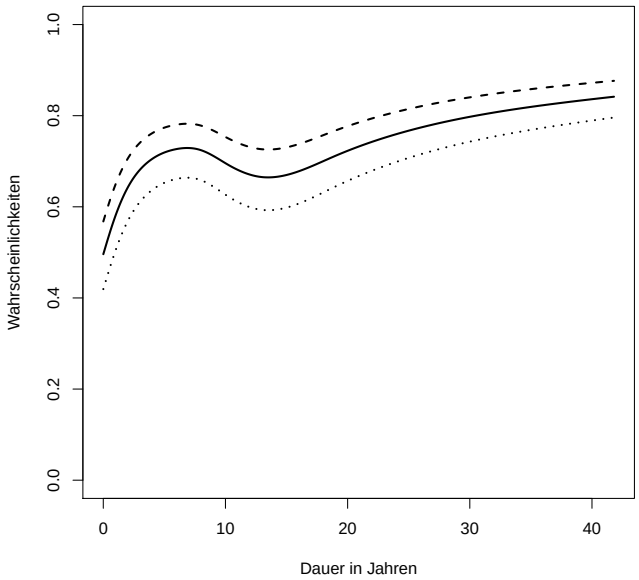


Abbildung 5.6.: Geschätzte Wahrscheinlichkeiten für eine Kapselkontraktur bei medianem Volumen von 280 cm^3: Oberfläche `glatt` (durchgehende Linie), Oberfläche `Polyurethan` (strichlierte Linie), Oberfläche `texturiert` (punktierte Linie), Drainage `j`, Lage `submuskulär`, OP-Zugang `bestehendeNarbe`, Füllung `Silikongel`.

5.1.3. Modell für die 5-stufige Response `kontr5`

Es soll auch ein Modell gefunden werden, in dem der Einfluss der erklärenden Variablen auf das Auftreten einer Kapselkontraktur und gleichzeitig das Auftreten der einzelnen Baker-Stufen untersucht wird. Dazu erstellen wir zunächst den 5-stufigen Faktor `kontr5`, indem die `NA`-Einträge der Variable `kontr` zu `no` (für keine Kapselkontraktur) umcodiert werden. Da wir wieder die `family`-Funktion `ocat` verwenden wollen, wandeln wir dies anschließend noch in einen numerischen Vektor mit Einträgen {1,2,3,4,5} um und nennen diesen `num.kontr5`. Dabei steht 1 für keine Kapselkontraktur, 2 für Baker-Stufe 1 usw.

```
tmp <- kontr
tmp <- as.character(tmp)
tmp[is.na(tmp)] <- 'no'
kontr5 <- factor(tmp, levels=c('no', 'BakerI', 'BakerII', 'BakerIII', 'BakerIV'),
     ordered=is.ordered(tmp))

library(plyr)
num.kontr5 <- as.numeric(revalue(kontr5, c('no'='1', 'BakerI'='2', 'BakerII'='3',
     'BakerIII'='4', 'BakerIV'='5')))
```

Zuerst suchen wir ein nichtparametrisches Modell mit der Dauer als erklärende Variable. Eine maximale Basisdimension von `k=10` liefert 5.49 effektive Freiheitsgrade, weshalb wir beim Default-Wert von `k=10` bleiben. Wir erhalten einen höchst signifikanten p-Wert beim Test auf Signifikanz des Glättungsterms.

```
mod.dau <- gam(num.kontr5 ~ s(dau, bs='cr'), family=ocat(R=5), method='ML')
anova(mod.dau)

Family: Ordered Categorical(-1,-0.71,-0.11,1.26)
Link function: identity

Formula:
num.kontr5 ~ s(dau, bs = "cr")

Approximate significance of smooth terms:
         edf Ref.df Chi.sq p-value
s(dau) 5.468  6.281    274  <2e-16
```

Nun wird der Einfluss des Volumens zusätzlich zur Dauer untersucht. Tabelle 5.5 zeigt die p-Werte des letzten Modellterms für verschiedene Modelle mit `vol`. Aufgrund dieser p-Werte erkennen wir, dass das Volumen einen signifikanten Einfluss hat. Ein nichtparametrisches Modell ist gerechtfertigt und Abbildung 5.7 zeigt die geschätzte Funktion unter diesem Modell. Auch ein lineares, quadratisches oder kubisches Modell wäre adäquat aber ein quadratischer und kubischer Term zusätzlich zu einem linearen Term ist jeweils nicht signifikant. Wir wählen ein Modell mit nichtparametrischem Einfluss für `vol`, da das Konfidenzband in Abbildung 5.7 v. a. für kleinere Volumina sehr eng ist und keine Gerade enthält.

Prädiktor	p-Wert für den letzten Term
`s(dau) + s(vol)`	≈ 0
`s(dau) + vol`	$5.65 \cdot 10^{-15}$
`s(dau) + vol`^{2}	$5.09 \cdot 10^{-13}$
`s(dau) + vol`^{3}	$1.77 \cdot 10^{-8}$
`s(dau) + vol + vol`^{2}	0.649
`s(dau) + vol + vol`^{3}	0.657

Tabelle 5.5.: p-Werte für Volumsterme in Modellen für die 5-stufige Response.

Nun untersuchen wir, welche erklärenden Faktoren zusätzlich zu den Glättungstermen für die Dauer und das Volumen noch relevant sind. Tabelle 5.6 listet die p-Werte für diese Tests für die einzelnen Faktoren auf.

Demnach gibt es viele Kandidaten für das Modell, darunter auch welche, die in den vorigen Modellen keine Rolle gespielt haben. Die Variable `antib` hat hier z. B. einen höchst signifikanten p-Wert war aber weder im Modell für `kontr` noch im Modell für `kontr_flag` enthalten. Wir fügen die Kandidaten aufsteigend nach dem p-Wert ins Modell ein und

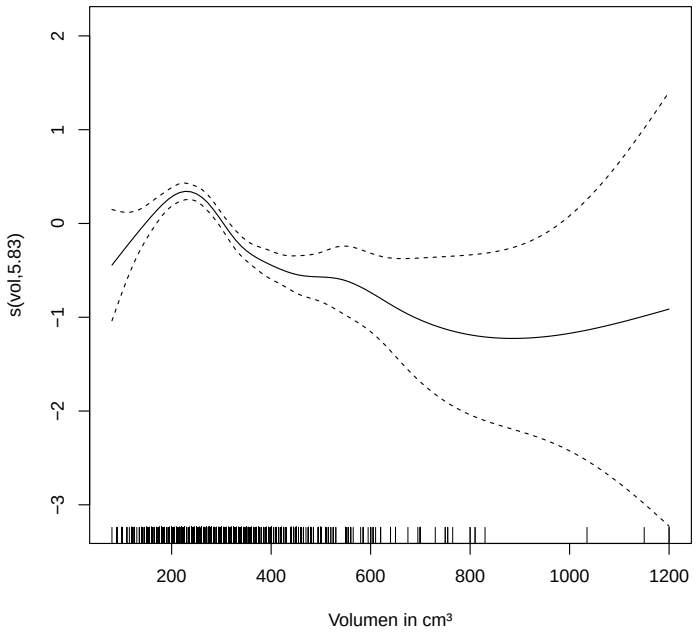


Abbildung 5.7.: Geschätzte Funktion für `vol` unter einem Modell für die 5-stufige Response mit Glättungstermen für die Dauer und das Volumen.

überprüfen welche zusätzlich zu zuvor hinzugefügten erklärenden Variablen signifikant sind. Dabei fallen `fuel` und `lumen` als Kandidaten weg, alle anderen Kandidaten werden ins Modell aufgenommen. Wir schätzen das finale Modell mit:

```
mod.kontr5 <- gam(num.kontr5 ~ s(dau, bs='cr') + s(vol, bs='cr') + antib + drain
    + lage + opzug + oberfl, family=ocat(R=5), method='ML')
```

Abbildung 5.8 zeigt die geschätzte Funktion für die Dauer unter dem finalen Modell. Der Prädiktor ist dann:

$$\begin{aligned}\hat{\eta} =& \hat{f}_1(\texttt{dau}) + \hat{f}_2(\texttt{vol}) - 1.01 - 0.16\ (\texttt{antib} = \texttt{Implantatinhalt}) \\ & - 0.97\ (\texttt{antib} = \texttt{keine}) - 0.82\ (\texttt{antib} = \texttt{systemisch}) \\ & - 0.21\ (\texttt{antib} = \texttt{systemisch+Implantathöhle}) - 0.70\ (\texttt{drain} = \texttt{n}) \\ & - 0.07\ (\texttt{lage} = \texttt{intermuskulär}) + 0.33\ (\texttt{lage} = \texttt{subcutan}) \\ & + 0.25\ (\texttt{lage} = \texttt{subglandulär}) - 0.08\ (\texttt{lage} = \texttt{submuskulär}) \\ & + 0.91\ (\texttt{opzug} = \texttt{bestehendeNarbe}) + 0.82\ (\texttt{opzug} = \texttt{inframammär}) \\ & + 0.60\ (\texttt{opzug} = \texttt{periareolär}) + 1.10\ (\texttt{opzug} = \texttt{T-Schnitt}) \\ & - 0.32\ (\texttt{opzug} = \texttt{transareolär}) + 0.79\ (\texttt{opzug} = \texttt{vertikal}) \\ & + 0.73\ (\texttt{oberfl} = \texttt{Polyurethan}) + 0.15\ (\texttt{oberfl} = \texttt{texturiert}).\end{aligned}$$

Prädiktor	p-Wert für den letzten Term
`s(dau) + s(vol) + oberfl`	0.002
`s(dau) + s(vol) + lumen`	0.011
`s(dau) + s(vol) + fuel`	0.003
`s(dau) + s(vol) + lage`	$2.50 \cdot 10^{-8}$
`s(dau) + s(vol) + opzug`	$1.33 \cdot 10^{-7}$
`s(dau) + s(vol) + antib`	≈ 0
`s(dau) + s(vol) + ster`	0.105
`s(dau) + s(vol) + drain`	≈ 0
`s(dau) + s(vol) + prim`	0.218

Tabelle 5.6.: p-Werte für erklärende Faktoren.

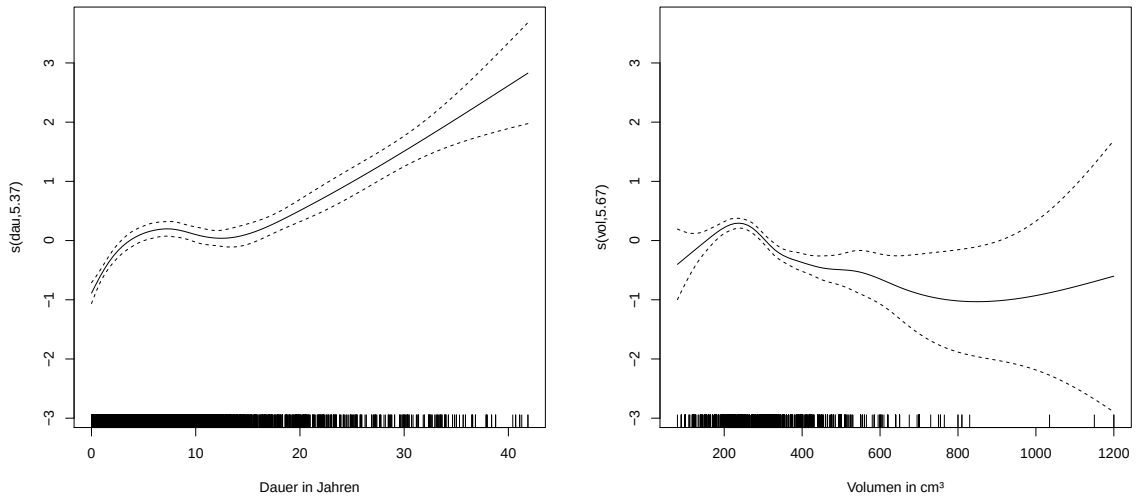


Abbildung 5.8.: Geschätzte Funktionen für `dau` und `vol` unter dem finalen Modell für die 5-stufige Response.

In Abbildung 5.9 sind die geschätzten Wahrscheinlichkeiten für die verschiedenen Kategorien in Abhängigkeit von der Dauer zu sehen, wobei das Volumen wieder auf den Median von 280 cm^3 und die erklärenden Faktoren jeweils auf die Stufe mit den meisten Beobachtungen fixiert wurden. Zu Beginn ist die Wahrscheinlichkeit für keine Kapselkontraktur am höchsten (ca. 60%) und sinkt dann im Laufe der Zeit (mit einer kurzen Ausnahme). Die Wahrscheinlichkeit für eine Kapselkontraktur der Baker-Stufe 4 verhält sich genau gegensätzlich. Sie ist zu Beginn sehr niedrig (< 10 %) und steigt nach 15 Jahren stetig an. Die Wahrscheinlichkeiten für die Baker-Stufen 1 und 2 sind sehr gering (beide ca. 10%, wobei die Wahrscheinlichkeit für Stufe 2 etwas höher ist) und beide sinken mit der Zeit ab. Die Wahrscheinlichkeit für Baker-Stufe 3 erhöht sich bis ca. 30 Jahre und sinkt dann wieder ab. Werden die erklärenden Faktoren auf andere Stufen fixiert resultieren ähnliche Grafiken.

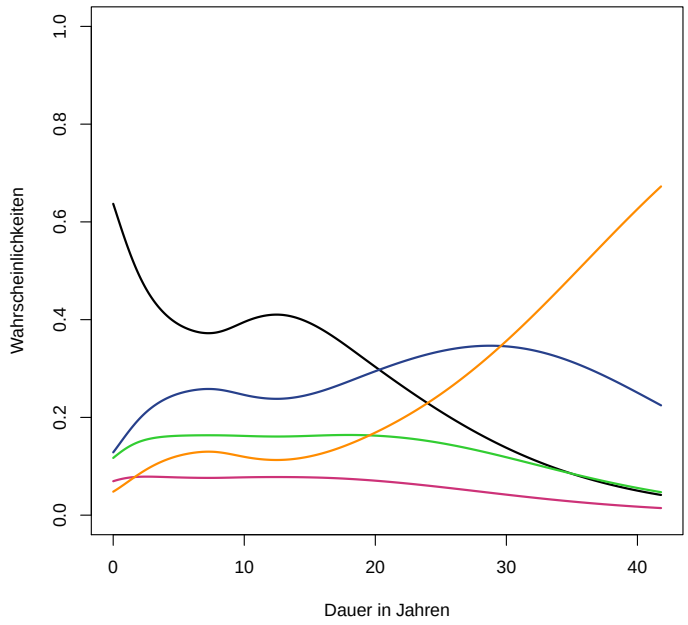


Abbildung 5.9.: Geschätzte Wahrscheinlichkeiten für die 5-stufige Response bei medianem Volumen von 280 cm^3: `BakerI` (violett), `BakerII` (grün), `BakerIII` (blau), `BakerIV` (orange), keine Kapselkontraktur (schwarz), Antibiotika `systemisch`, Drainage `j`, Lage `submuskulär`, OP-Zugang `bestehendeNarbe`, Oberfläche `texturiert`.

5.2. Softwarepaket VGAM

Eine Alternative zum mgcv-Paket stellt das VGAM-Package dar. Dabei handelt es sich um ein Softwarepaket zum Schätzen von *vektorgeneralisierten linearen und additiven Modellen*. Dies ist eine relativ große Klasse von Modellen, welche auch unsere additiven Modelle umfasst. Einen recht guten Überblick über die Möglichkeiten der Modellierung mit VGAM kann man sich in Yee (2010) verschaffen. Die Funktion vglm des VGAM-Pakets haben wir schon in Abschnitt 3.2 zum Schätzen von parametrischen Modellen mit kategoriellen Responses verwendet. Für Modelle mit Glättungstermen werden wir nun die Funktion vgam verwenden.

Modellierung mit VGAM

Der Aufruf der Funktion vgam funktioniert wie gewohnt durch Übergabe einer Modellformel und Angabe einer Verteilung im family-Argument. Da wir die Modelle mit den mittels mgcv gefundenen Modellen vergleichen wollen, schätzen wir kumulative Modelle in Proportional-Odds-Form und setzen dazu `family = cumulative(parallel=T)`. Dies liefert ein kumulatives Modell, welches wegen dem parallel-Argument in Proportional-Odds-Form ist. Die Modellformel darf Glättungsterme enthalten, welche wie beim mgcv-Paket mit s() eingegeben werden. Die Funktion s() des VGAM-Packages unterscheidet sich aber von der entsprechenden Funktion des mgcv-Pakets. Die beiden Pakete sollten deshalb nicht gleichzeitig ausgeführt werden.

Die Funktion s() des VGAM-Pakets repräsentiert einen kubischen Spline mit Strafterm. Allerdings ist im Gegensatz zum mgcv-Paket keine automatische Glättungsparameterschätzung implementiert. Dieser ist hier fix und muss vom Benutzer eingegeben werden. Dazu kann man entweder das Argument df oder das Argument spar angeben (beides gleichzeitig liefert eine Fehlermeldung). Mittels df werden die effektiven Freiheitsgrade festgelegt. Setzen von df=1 liefert eine Gerade. Der Default-Wert ist k=4. Alternativ kann im spar-Argument der Glättungsparameter angegeben werden. Ob nun df oder spar spezifiziert wird, bleibt dem Benutzer überlassen. Beides hat natürlich den gleichen Effekt. Ein großer Glättungsparameter reduziert die effektiven Freiheitsgrade, während ein kleiner Glättungsparameter mehr effektive Freiheitsgrade zulässt. Abbildung 5.10 zeigt diesen Zusammenhang für ein Modell für die ordinalen Baker-Stufen mit einem Glättungsterm für die Dauer.

Für die Funktion vgam ist kein anova-Befehl implementiert. Der summary-Befehl existiert zwar, liefert aber weniger Informationen als der summary-Befehl für die Funktion gam des mgcv-Pakets. Es gibt beispielsweise keine p-Werte für den parametrischen Teil des Modells. Es ist nicht ganz klar wie die Teststatistiken und dazugehörigen p-Werte für die Glättungsterme berechnet werden. Die Entscheidung über die Relevanz der Terme ist aber in beiden Paketen dieselbe.

Das VGAM-Paket hat aber auch einen großen Vorteil. Neben kumulativen Modellen kann eine Vielzahl von weiteren Modellformen geschätzt werden. Es ist beispielsweise möglich

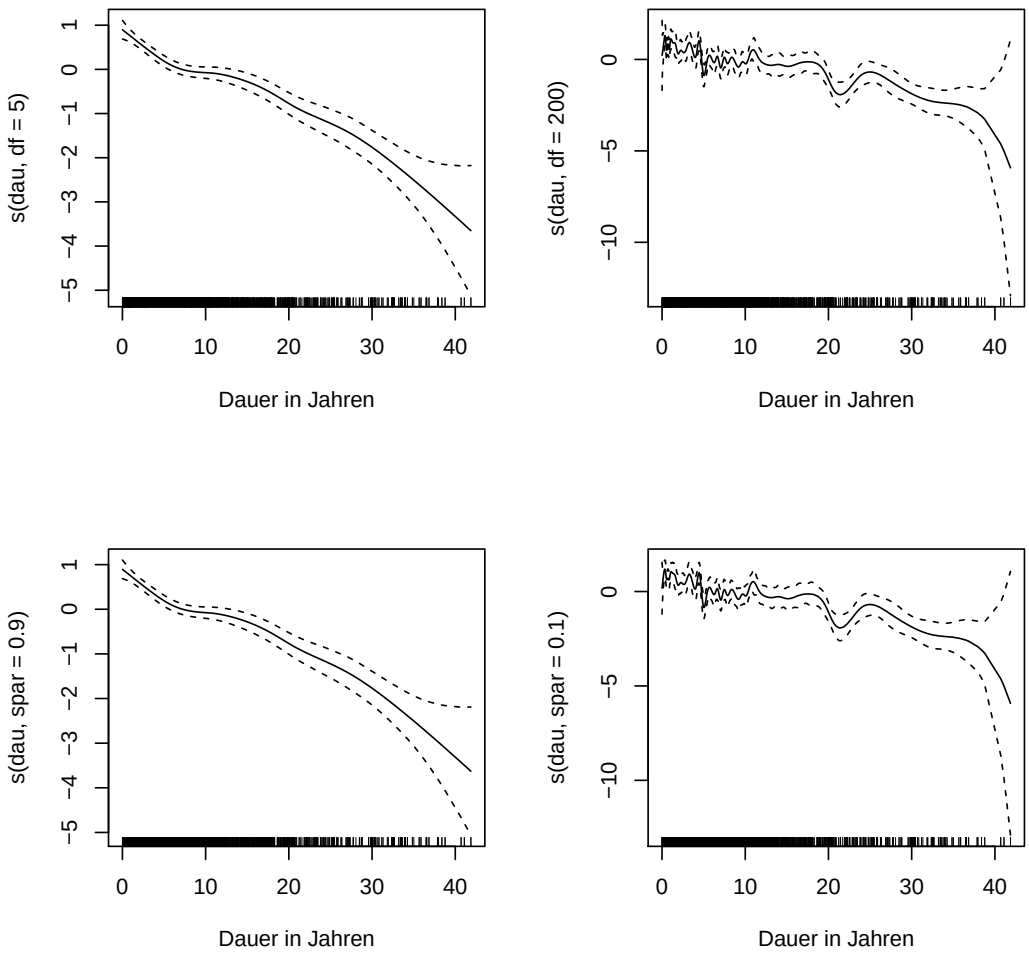


Abbildung 5.10.: Geschätzte Funktion für ein Modell für die ordinalen Baker-Stufen mit Glättungsterm für die Dauer und unterschiedlichen Werten für die effektiven Freiheitsgrade bzw. den Glättungsparameter: `df=5` (oben links) entspricht etwa `spar=0.9` (unten links), `df=200` (oben rechts) entspricht ungefähr `spar=0.1` (unten rechts).

sequenzielle Modelle oder Adjacent-Category-Modelle zu betrachten. Dazu setzen wir das `family`-Argument für ein sequenzielles Modell auf `sratio` bzw. für ein Adjacent-Category-Modell auf `acat`, wie wir schon in Abschnitt 3.2 gesehen haben. Ein Vergleich der unter dem finalen Modell für `kontr` geschätzten Wahrscheinlichkeiten bei Verwendung der verschiedenen Modellformen ist in Abbildung 5.11 zu sehen. Tabelle 5.7 fasst die wesentlichen Vor- und Nachteile der zwei vorgestellten Softwarepakete noch einmal zusammen.

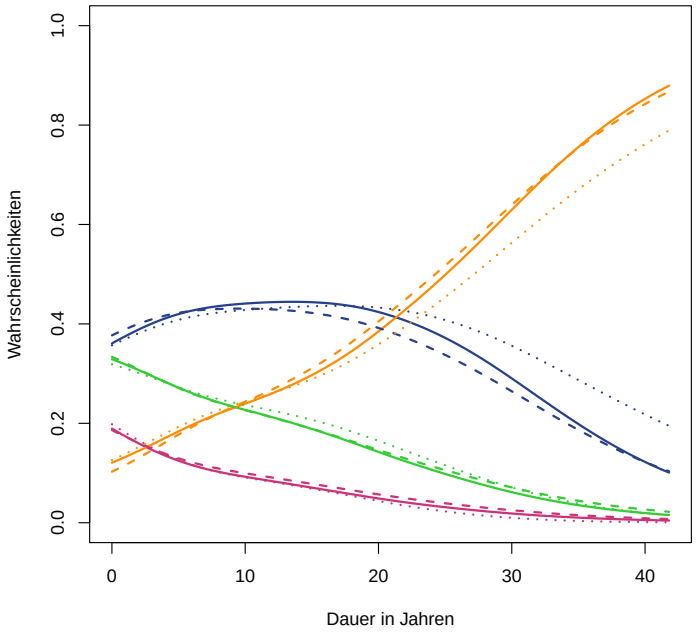


Abbildung 5.11.: Geschätzte Wahrscheinlichkeiten unter dem finalen Modell für die ordinalen Baker-Stufen bei medianem Volumen von 280 cm^3: kumulatives Modell (durchgehende Linien), sequenzielles Modell (strichlierte Linien) und Adjacent-Category-Modell (punktierte Linien), `BakerI` (violett), `BakerII` (grün), `BakerIII` (blau), `BakerIV` (orange), Oberfläche `texturiert`, Drainage `j`.

`mgcv`		`VGAM`	
+	Automatische Schätzung der Glättungsparameter möglich	–	Glättungsparameter bzw. effektive Freiheitsgrade sind fix und müssen gesetzt werden
+	`anova`-Befehl verfügbar	–	kein `anova` vorhanden
–	nur kumulative Modelle in Proportional-Odds-Form	+	auch viele andere Modellformen möglich

Tabelle 5.7.: Vergleich des `mgcv`- und `VGAM`-Pakets.

Vergleich der Modelle für das Silikonregister

Es sollen nun die in den Abschnitten 5.1.1 - 5.1.3 gefundenen Modelle auch mittels der Funktion `vgam` geschätzt und die Ergebnisse verglichen werden. Da es für die Funktion `vgam` keine automatische Glättungsparameterschätzung gibt, setzen wir das `df`-Argument der Funktion `s()` für `vgam` auf den Wert des für die Funktion `s()` unter `mgcv` geschätzten effektiven Freiheitsgrades.

Bei Modellen mit geordneten Responses muss diese bei Verwendung der Funktion `vgam` auch als geordneter Faktor übergeben werden. Deshalb erstellen wir zuerst die Variable `ord.kontr`, welche die Stufen der Faktors `kontr` mit einer Ordnung enthält, und schätzen dann das Modell mit `ord.kontr` als Response.

```
ord.kontr <- as.ordered(kontr)
mod.kontr.vgam <- vgam(ord.kontr ~ s(dau, df=3.3) + vol + drain + oberfl, family=
    cumulative(parallel=T))
summary(mod.kontr.vgam)

Call:
vgam(formula = ord.kontr ~ s(dau, df = 3.3) + vol + drain + oberfl,
    family = cumulative(parallel = T))

Number of linear predictors:    3

Names of linear predictors: logit(P[Y<=1]), logit(P[Y<=2]), logit(P[Y<=3])

Dispersion Parameter for cumulative family:   1

Residual deviance:  4830.954 on 5836.833 degrees of freedom

Log-likelihood: -2415.477 on 5836.833 degrees of freedom

Number of iterations:  7

DF for Terms and Approximate Chi-squares for Nonparametric Effects

                 Df Npar Df Npar Chisq    P(Chi)
(Intercept):1     1
(Intercept):2     1
(Intercept):3     1
s(dau, df = 3.3)  1      2.2     9.2758 0.0116553
vol               1
drain             1
oberfl            2
```

Abbildung 5.12 zeigt die von `vgam` geschätzten Wahrscheinlichkeiten für die vier Baker-Stufen. Zum Vergleich wurden auch die mit der `mgcv`-Funktion `gam` geschätzten Wahrscheinlichkeiten als dünne punktierte Linien eingezeichnet. Die Ergebnisse sind nahezu identisch.

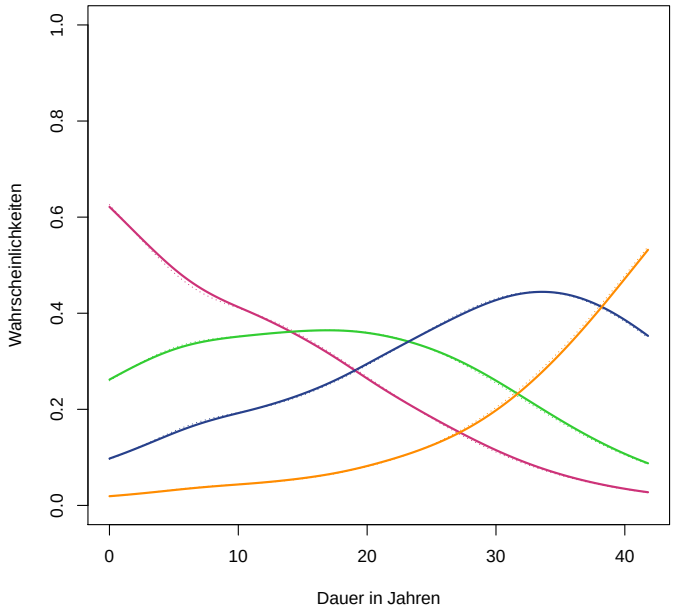


Abbildung 5.12.: Geschätzte Wahrscheinlichkeiten beim Modell für die ordinalen Baker-Stufen: mittels `vgam` aus dem `VGAM`-Package (dicke durchgehende Linien), mittels `gam` aus dem `mgcv`-Package (dünne punktierte Linien), `BakerI` (violett), `BakerII` (grün), `BakerIII` (blau), `BakerIV` (orange).

Wir können auch die Schätzer für die Parameter vergleichen:

```
coef(mod.kontr.vgam)
    (Intercept):1     (Intercept):2     (Intercept):3  s(dau, df = 3.3)
     -0.579295013       0.948859040       2.860093458      -0.079153784
              vol            drainn oberflPolyurethan  oberfltexturiert
      0.001350855       0.640069462      -1.153407839      -1.293493995
```

Die ersten drei Intercept-Parameter sind die geschätzten Schwellenwerte. Hier wird der erste Schwellenwert nicht wie beim `mgcv`-Paket auf -1 festgesetzt und wir erhalten deshalb andere Schätzer. Die Abstände zwischen den Schwellen sind aber gleich wie bei den Schätzungen unter `mgcv`. Der Parameter zu `s(dau, df = 3.3)` ist der globale Intercept-Parameter. Auch dieser unterscheidet sich von der Schätzung unter dem `mgcv`-Paket, aber addiert mit den Schwellenwerten der jeweiligen Kategorie ergeben sich in bei beiden Paketen wieder die gleichen Intercepts. Die restlichen Parameter entsprechen bis aufs Vorzeichen genau den unter `mgcv` geschätzten Parametern. Dies liegt wieder an der internen Definition des linearen Prädiktors (hier $\boldsymbol{\eta}_j(\boldsymbol{x}) = \alpha_j + \boldsymbol{x}^\top \boldsymbol{\beta}$), die mit unserer Definition wie in (3.2.7) übereinstimmt. Im `mgcv`-Paket steht statt dem Plus ein Minus, was zu den unterschiedlichen Vorzeichen der Parameterschätzer führt.

Beim Modell für `kontr_flag` setzen wir das `family`-Argument auf `binomialff`. Diese Funktion ist laut Yee (2015) weitgehend gleich mit der uns bekannten Funktion `binomial`, erlaubt aber zusätzlich einen frei wählbaren Dispersionsparameter. Wir schätzen das Modell durch:

```
mod.kontr_flag.vgam <- vgam(kontr_flag ~ s(dau, df=5.3) + s(vol, df=5.6) + drain
    + lage + opzug + oberfl + fuel, family=binomialff)
```

Das Ergebnis ist in Abbildung 5.13 dargestellt. Dabei sind sowohl die mit `vgam` als auch die mit `gam` geschätzten Wahrscheinlichkeiten eingezeichnet. Auch hier stimmen die geschätzten Funktionen nahezu überein. Dasselbe gilt bis auf die Intercept-Parameter auch für die Parameterschätzer.

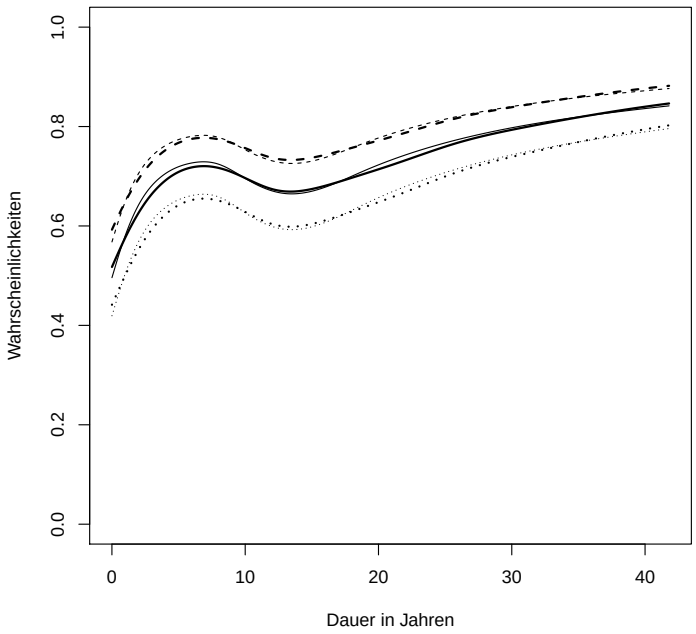


Abbildung 5.13.: Geschätzte Wahrscheinlichkeiten einer Kapselkontraktur bei medianem Volumen von 280 cm^3 beim binomialen Modell: mittels `vgam` aus dem `VGAM`-Package (dicke Linien), mittels `gam` aus dem `mgcv`-Package (dünne Linien), Oberfläche `glatt` (durchgehende Linien), Oberfläche `Polyurethan` (strichlierte Linien), Oberfläche `texturiert` (punktierte Linien), Drainage `j`, Lage `submuskulär`, OP-Zugang `bestehendeNarbe`, Füllung `Silikongel`.

Für ein Modell mit der Variable `kontr5` erstellen wir zuerst wieder die Variable mit geordneten Stufen und nennen diese `ord.kontr5`. Wir schätzen das Modell mit:

```
ord.kontr5 <- as.ordered(kontr5)
mod.kontr5.vgam <- vgam(ord.kontr5 ~ s(dau, df=5.4) + s(vol, df=5.7) + antib +
    drain + lage + opzug + oberfl, family=cumulative(parallel=T))
```

Die geschätzten Wahrscheinlichkeiten sind in Abbildung 5.14 dargestellt. Zum Vergleich sind auch wieder die mit dem `mgcv`-Paket geschätzten Wahrscheinlichkeiten eingezeichnet. Auch hier sind die Ergebnisse fast identisch. Die Parameterschätzer (ausgenommen Intercepts) stimmen wieder bis auf das Vorzeichen überein.

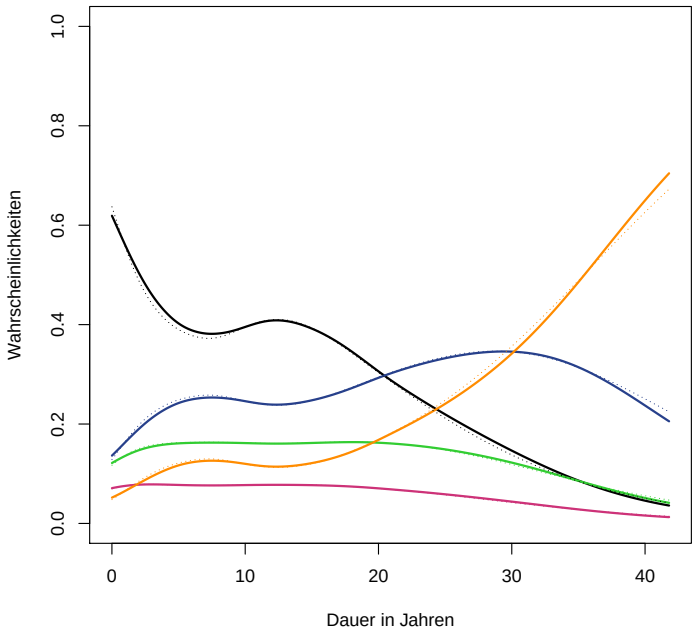


Abbildung 5.14.: Geschätzte Wahrscheinlichkeiten beim Modell für die 5-stufige Response: mittels `vgam` aus dem `VGAM`-Package (dicke durchgehende Linien), mittels `gam` aus dem `mgcv`-Package (dünne punktierte Linien), `BakerI` (violett), `BakerII` (grün), `BakerIII` (blau), `BakerIV` (orange).

6. Resümee

In diesem Abschnitt werden die Ergebnisse der statistischen Analysen noch einmal kurz zusammengefasst. Wir haben Regressionsmodelle für drei verschiedene Responses betrachtet, um verschiedene Fragestellungen in Bezug auf eine Kapselkontraktur zu beantworten. Es wurde jeweils der Einfluss von diversen erklärenden Variablen untersucht. Wir interpretieren jeweils die unter dem `mgcv`-Paket geschätzten Parameter.

Ergebnis für `kontr`

Im ordinalen Modell für die vier Baker-Stufen haben die Dauer zwischen den Operationen, das Volumen sowie die Oberflächenbeschaffenheit des Implantats und das Setzen einer Drainage während der Operation Einfluss.
Die Dauer wirkt dabei nichtparametrisch ins Modell ein. Prinzipiell wird die Wahrscheinlichkeit für eine Kapselkontraktur in Baker-Stufe 1 mit zunehmender Dauer kleiner. Umgekehrt steigt die Wahrscheinlichkeit für Baker-Stufe 4 mit zunehmender Dauer an.
Das Volumen hat linearen Einfluss, wobei das Vorzeichen des Parameterschätzers negativ ist. Mit zunehmendem Volumen vergrößert sich die Wahrscheinlichkeit für Baker-Stufe 1, während sie sich für Baker-Stufe 4 verringert.
Für das Setzen einer Drainage ist die Referenz die Stufe `j`. Wurde bei einer Patientin keine Drainage gesetzt vergrößert sich die Wahrscheinlichkeit für Baker-Stufe 1. Die Wahrscheinlichkeit für Baker-Stufe 4 ist dann für Patientinnen ohne Drainage kleiner als für Patientinnen mit Drainage. Für Patientinnen ist es demnach besser, wenn keine Drainage gesetzt werden muss.
Für die Oberfläche ist die Referenz `glatt`. `Polyurethan`-beschichtete und `texturierte` Oberflächen verringern die Wahrscheinlichkeit für Baker-Stufe 1, wobei die Verringerung für eine `texturierte` Oberfläche etwas größer ist. Die Wahrscheinlichkeiten für Baker-Stufe 4 verhalten sich umgekehrt und eine Ordnung der Levels nach aufsteigenden Wahrscheinlichkeiten für Baker-Stufe 4 ist die folgende:

$$\texttt{glatt} < \texttt{Polyurethan} < \texttt{texturiert}.$$

Ergebnis für `kontr_flag`

Beim binomialen Modell für die Variable `kontr_flag` wurde die Auftrittswahrscheinlichkeit für eine Kapselkontraktur modelliert. Ein größerer linearer Prädiktor impliziert hier eine größere Auftrittswahrscheinlichkeit. In unserem Modell haben die stetigen erklärenden Variablen Dauer und Volumen einen nichtparametrischen Einfluss. Zusätzlich sind noch die

Oberfläche, Füllung und Lage des Implantats, der OP-Zugang und das Setzten einer Drainage relevant.
Die Auftrittswahrscheinlichkeit für eine Kapselkontraktur ist zu Beginn klein und wird über die Dauer gesehen größer. Umgekehrt ist die Wahrscheinlichkeit, dass keine Kapselkontraktur auftritt anfangs größer und wird mit der Zeit kleiner.
Das Setzen einer Drainage verringert den linearen Prädiktor. Damit ist die Auftrittswahrscheinlichkeit für eine Kapselkontraktur bei Patientinnen ohne Drainage niedriger als bei Patientinnen mit Drainage.
Bezüglich der Lage des Implantats ist `biplane` die Referenz. Eine `subcutane` und `subglanduläre` Lage vergrößern die Auftrittswahrscheinlichkeit, `submuskuläre` und `intermuskuläre` Lage verringern diese. Die folgende Ordnung der Levels des Faktors `lage` nach aufsteigenden Auftrittswahrscheinlichkeiten für eine Kapselkontraktur gilt:

$$\texttt{intermuskulär} < \texttt{submuskulär} < \texttt{biplane} < \texttt{subcutan} < \texttt{subglandulär}.$$

Für den OP-Zugang ist ein `axillärer` Zugang die Referenz. Ein `transareolärer` Zugang verringert die Auftrittswahrscheinlichkeit für eine Kapselkontraktur, die restlichen Faktorstufen vergrößern selbige. Die Levels des Faktors `opzug` sind nach aufsteigenden Auftrittswahrscheinlichkeiten folgendermaßen geordnet:

$$\begin{aligned}\texttt{transareolär} < \texttt{axillär} < \texttt{vertikal} < \texttt{periareolär} \\ < \texttt{inframammär} < \texttt{bestehendeNarbe} < \texttt{T-Schnitt}.\end{aligned}$$

Eine `texturierte` Oberfläche verringert die Auftrittswahrscheinlichkeit einer Kapselkontraktur bzgl. der Referenz `glatt`, während eine `Polyurethan`-beschichtete Oberfläche diese vergrößert. Die Levels des Faktors `oberfl` nach aufsteigenden Auftrittswahrscheinlichkeiten sind also:

$$\texttt{texturiert} < \texttt{glatt} < \texttt{Polyurethan}.$$

Für die Füllung des Implantats ist `andere` die Referenz. Die restlichen Füllungen verringern die Auftrittswahrscheinlichkeit. Es gilt folgende Ordnung der Levels des Faktors `fuel` nach aufsteigenden Auftrittswahrscheinlichkeiten:

$$\texttt{Kochsalzlösung} < \texttt{gemischt} < \texttt{Hydrogel} < \texttt{Silikongel} < \texttt{andere}.$$

Ergebnis für `kontr5`

Im Modell mit der 5-stufigen Response sind die Dauer, das Volumen, die Verabreichung von Antibiotika, die Lage des Implantats, der OP-Zugang und die Oberfläche relevant. Der Einfluss der Dauer und des Volumens ist nichtparametrisch.
Mithilfe der Parameter können wir hier die Wahrscheinlichkeit für keine Kapselkontrakur (1. Kategorie der Response) und die Wahrscheinlichkeit für eine Kontraktur der Baker-Stufe 4 (5. Kategorie der Response) interpretieren. Zu Beginn sind die Wahrscheinlichkeiten

für keine Kapselkontraktur am höchsten und für Baker-Stufe 4 niedrig. Mit zunehmender Dauer kehrt sich dies genau um.
Für die Verabreichung von Antibiotika ist die Verabreichung in der `Implantathöhle` die Referenz. Für alle anderen Levels des Faktors vergrößert sich die Wahrscheinlichkeit für keine Kapselkontraktur bzgl. der Referenz und es verringert sich die Wahrscheinlichkeit für Baker-Stufe 4. Die Werte der Parameterschätzung implizieren folgende Ordnung der Levels nach aufsteigender Wahrscheinlichkeit für Baker-Stufe 4:

$$\texttt{keine} < \texttt{systemisch} < \texttt{systemisch+Implantathöhle} < \texttt{Implantatinhalt} < \texttt{Implantathöhle}.$$

Für das Setzen einer Drainage ist wieder `j` die Referenz und für Patientinnen mit keiner Drainage verringert sich die Wahrscheinlichkeit für eine Kapselkontraktur auf Baker-Stufe 4.
Die Referenz für die Lage des Implantats ist `biplane`. Eine `intermuskuläre` und `submuskuläre` Lage verringern die Wahrscheinlichkeit für Baker-Stufe 4, während sie sich für `subcutane` und `subglanduläre` Lage vergrößert. Es gilt folgende Ordnung der Levels des Faktors `lage` nach aufsteigender Wahrscheinlichkeit für eine Kapselkontraktur auf Baker-Stufe 4:

$$\texttt{submuskulär} < \texttt{intermuskulär} < \texttt{biplane} < \texttt{subglandulär} < \texttt{subcutan}.$$

Bei Patientinnen mit `subcutaner` Lage tritt am ehesten eine Kontraktur der Baker-Stufe 4 ein.
Für den OP-Zugang ist die Referenz `axillär`. Ein `transareolärer` OP-Zugang verringert die Wahrscheinlichkeit für Baker-Stufe 4, alle anderen Levels vergrößern diese Wahrscheinlichkeit gegenüber der Referenz. Wir haben folgende Ordnung der Levels des Faktors `opzug` nach aufsteigender Wahrscheinlichkeit für Baker-Stufe 4:

$$\texttt{transareolär} < \texttt{axillär} < \texttt{periareolär} < \texttt{vertikal} < \texttt{inframammär} < \texttt{bestehendeNarbe} < \texttt{T-Schnitt}.$$

Die Wahrscheinlichkeiten für keine Kapselkontraktur verhalten sich gegensätzlich.
Für die Oberflächenbeschaffenheit haben eine `Polyurethan`-beschichtete und `texturierte` Oberfläche wieder größere Wahrscheinlichkeiten für Baker-Stufe 4 als die Referenz `glatt`. Hier gilt die folgende Ordnung der Levels nach aufsteigenden Wahrscheinlichkeiten für Baker-Stufe 4:

$$\texttt{glatt} < \texttt{texturiert} < \texttt{Polyurethan}.$$

Der Faktor Oberfläche ist in allen drei Modellen relevant. Es fällt aber auf, dass die Wirkung der Oberflächenbeschaffenheit für die verschiedenen Responses unterschiedlich ist (verschiedene Ordnung der Levels).

A. Anhang

A.1. Datenaufbereitung

Die Daten lagen ursprünglich als `Excel`-Datei vor, welche als `csv`-Datei, in der die einzelnen Felder durch ein Komma getrennt sind, exportiert wurde. Einige Felder haben am Ende von Bezeichnungen viele Leerzeichen enthalten. Deshalb wurde die `csv`-Datei in einem ersten Schritt in einem Texteditor geöffnet und die überflüssigen Leerzeichen mittels „Suchen & Ersetzen“ gelöscht. Die bearbeitete Datei wurde unter dem Namen `IMPLANTATE_201402.txt` gespeichert. Die Textdatei wurde ins `Microsoft Excel` importiert und es wurden Spaltenüberschriften eingefügt. Das funktioniert auch mit der freien Version `OpenOffice Calc`. Diese Datei wurde unter dem Namen `IMPLANTATE_201402.xlsx` gespeichert und auch als `csv`-Datei `IMPLANTATE_201402.csv` exportiert.

Mithilfe der Datei `ReadIn.R` wurde die `csv`-Datei in `R` zunächst als Daten-Frame `mydata` eingelesen. Bevor damit aber gearbeitet werden kann, mussten noch ein paar Manipulationen des Datensatzes durchgeführt werden. Der ursprüngliche Datensatz enthält 17327 Zeilen und 70 Spalten. Da für uns nur Patienten mit Operationsart `Implantatwechsel` und Implantatart `Mammaimplantat` von Interesse sind, wurden alle Zeilen aus dem Daten-Frame gelöscht, für die das nicht zutrifft. Im Datensatz gibt es die zwei Spalten `REV_OPDATUM` und `REV_LETZTESOPDATUM`. Erstere gibt dabei das Datum einer Revisionsoperation an und letztere gibt das Datum der letzten Operation vor dieser Operation an. Wir betrachten nur Patientinnen bei denen die erste Operation nach inkl. 1970 stattgefunden hat. Frühere Operationen betrachten wir als Tippfehler und löschen diese. Des Weiteren gibt es noch in der Spalte `IMPLANTAT_FUELLVOLUMEN` einen extremen Ausreißer mit über 3000 cm^3 Füllvolumen. Auch diesen betrachten wir als Tippfehler und entfernen die entsprechende Zeile aus dem Datensatz. Dabei werden beim Entfernen jeweils auch alle Levels von Faktoren gelöscht, für die es keine Beobachtungen mehr gibt.

Der Datensatz soll eine neue Spalte `DAUER` enthalten, die die Zeit zwischen der Revisionsoperation und der Operation davor in Jahren angibt. Die Werte für diese Spalte sollen aus den beiden vorhandenen Spalten `REV_OPDATUM` und `REV_LETZTESOPDATUM` ermittelt werden. Aus diesen zwei Spalten konstruieren wir die neue Spalte `DAUER` mithilfe der `ISOdate`-Funktion in `R`. Damit kann für zwei Daten die Anzahl der Tage, die dazwischen liegen, berechnet werden. Die Anzahl der Tage wurden dann noch durch 365.25 dividiert, um auch Schaltjahre zu berücksichtigen. Dabei tritt das Problem auf, dass manche Einträge der Spalte `REV_LETZTESOPDATUM` nur das Jahr oder nur Jahr und Monat beinhalten. Für alle diese Fälle wurden der 01.07. des Jahres bzw. der 15. des Monats als Stichtag willkürlich

festgesetzt. Nun erhalten wir die Spalte `DAUER`, die allerdings auch nichtpositive Einträge enthält. Wir betrachten auch diese als Tippfehler und löschen die entsprechenden Zeilen. Der Datensatz hat nun noch 3534 Zeilen.

Im Datensatz gibt es auch eine Spalte `PRIMAERINDIKATION`, die den Grund angibt, weshalb es zu einem Implantat gekommen ist. Dieser Faktor hat sehr viele verschiedene Levels, wobei die Einträge oft ziemlich ähnlich sind. Es gibt beispielsweise die Levels `tubuläreDeformität` und `TubuläreDeformität`. Zur Vereinfachung, werden wir beim Grund deshalb nur zwischen kosmetisch oder rekonstruktiv entscheiden und erstellen dazu eine neue Spalte `PRIMAERIND`. Dazu gibt es die Datei `Diagnosen_Zuordnung.csv`, die von Paul Wurzer erstellt wurde. Diese hat zwei Spalten, wobei die erste Spalte `Diagnose` alle Levels des Faktors `PRIMAERINDIKATION` aus dem Datensatz enthält. In der zweiten Spalte `Gruppe` wurden diesen Levels die neuen Levels `Cosmetic` bzw. `Reconstructive` zugeordnet. Falls die Zuordnung nicht bekannt war, wurde `NA` eingefügt.

Die originale Codierung der binomialen Variablen ist `j`, wenn bei einer Patientin das entsprechende Problem (z. B. eine Kapselkontraktur) aufgetreten ist, und `n`, wenn das Problem nicht aufgetreten ist. In `R` werden bei binomialen Responses standardmäßig die Wahrscheinlichkeiten für die lexikographisch letzte aller Stufen modelliert. D. h. es würden bei der Codierung mit `j` und `n` die Wahrscheinlichkeiten für das Nicht-Eintreten des Problems geschätzt werden. Für die Interpretation ist es aber einfacher die Wahrscheinlichkeiten für das Eintreten eines Problems zu schätzen. Deshalb codieren wir die binomialen Response-Variable `kontr_flag` um, sodass nun `yes` für das Eintreten eines Problems und `no` für das Nicht-Eintreten steht. Nun werden genau die gewünschten Wahrscheinlichkeiten modelliert, da `yes` alphabetisch nach `no` kommt. Aus Gründen der besseren Lesbarkeit wurden außerdem die Variablen umbenannt, sodass die Namen nun aus Kleinbuchstaben bestehen und teilweise verkürzt sind. In Tabelle A.1 sind die originalen und die neuen Variablennamen gegenübergestellt.

Originaler Name	Neuer Name	Originaler Name	Neuer Name
KAPSELKONTRAKTUR	kontr	IMPLANTAT_LAGE	lage
KAPSELKONTRAKTUR_FLAG	kontr_flag	OPZUGANG	opzug
DAUER	dau	ANTIBIOTIKA	antib
IMPLANTAT_FUELLVOLUMEN	vol	STEROIDE	ster
IMPLANTAT_OBERFLAECHE	oberfl	DRAINAGE	drain
IMPLANTAT_LUMEN	lumen	PRIMAERIND	prim
IMPLANTAT_FUELLUNG	fuel		

Tabelle A.1.: Originale und neue Variablennamen.

Nun haben wir alle Spalten, die wir brauchen und löschen aus dem Datensatz alle nicht relevanten Spalten. Der fertige Daten-Frame hat nun 3534 Zeilen und 13 Spalten. Nach einem `attach`-Befehl kann man sich noch mittels geeigneter Plots und mittels des `summary`-Befehls einen Überblick über die im Datensatz vorhandenen Variablen verschaffen.

```
summary(mydata)
```

```
 kontr_flag      kontr           vol                oberfl        lumen
 no :1585   BakerI  : 243   Min.   :  80.0   glatt      :  61   double:  84
 yes:1949   BakerII : 488   1st Qu.: 225.0   Polyurethan: 114   single:3446
            BakerIII: 767   Median : 280.0   texturiert :3359   triple:   4
            BakerIV : 451   Mean   : 292.7
            NA's    :1585   3rd Qu.: 350.0
                            Max.   :1200.0

             fuel                  lage                  opzug
 andere        :   5   biplane      : 135   axillär        :  61
 gemischt      :  30   intermuskulär:  67   bestehendeNarbe:1582
 Hydrogel      :  19   subcutan     : 114   inframammär    :1539
 Kochsalzlösung:  64   subglandulär :1186   periareolär    : 215
 Silikongel    :3416   submuskulär  :2032   T-Schnitt      :  57
                                            transareolär   :  34
                                            vertikal       :  46
                       antib                   ster      drain
 Implantathöhle           : 345   Implantathöhle:  11   j:3001
 Implantatinhalt          :   3   keine         :3497   n: 533
 keine                    : 229   systemisch    :  26
 systemisch               :2478
 systemisch+Implantathöhle: 479


             prim           dau
 Reconstructive:1655   Min.   : 0.00274
 Cosmetic      : 523   1st Qu.: 1.80287
 NA's          :1356   Median : 6.43121
                       Mean   : 8.17054
                       3rd Qu.:11.52704
                       Max.   :41.87543
```

A.2. Parameterschätzung bei GLMs

Vergleiche zu diesem Abschnitt Agresti (2013, S. 143 ff.).

Wir wollen die Maximalstelle $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \ldots, \hat{\beta}_{p-1})^\top$ der Log-Likelihood-Funktion zu einem GLM bestimmen. Dies geschieht iterativ mittels des **Newton-Raphson-Verfahrens**. Die Log-Likelihood-Funktion ist gegeben als

$$l(\boldsymbol{\theta}, \boldsymbol{y}) = \sum_{i=1}^{n} \left(\frac{y_i \theta_i - b(\theta_i)}{w_i \phi} + c(y_i, \phi) \right),$$

wobei $\boldsymbol{\theta}$ durch die Beziehungen $b'(\boldsymbol{\theta}) = \boldsymbol{\mu}$ und $g(\boldsymbol{\mu}) = \boldsymbol{X}\boldsymbol{\beta}$ von $\boldsymbol{\beta}$ abhängt. Deshalb schreiben wir von nun an $l(\boldsymbol{\beta}, \boldsymbol{y})$ statt $l(\boldsymbol{\theta}, \boldsymbol{y})$.

Es seien

$$\boldsymbol{u} = \frac{\partial l(\boldsymbol{\beta}, \boldsymbol{y})}{\partial \boldsymbol{\beta}} \quad \text{und} \quad \boldsymbol{H} = \frac{\partial^2 l(\boldsymbol{\beta}, \boldsymbol{y})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top}.$$

Die Iterationsvorschrift für die Berechnung von $\hat{\boldsymbol{\beta}}$ lautet dann

$$\boldsymbol{u}^{(t)} + \boldsymbol{H}^{(t)}(\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}) = \mathbf{0}. \tag{A.2.1}$$

Dabei ist $\boldsymbol{\beta}^{(t)}$ der Wert für $\hat{\boldsymbol{\beta}}$ in Schritt t der Iteration, wobei $t = 0, 1, \ldots$. Es sei $\boldsymbol{u}^{(t)}$ der Vektor $\boldsymbol{u}$ ausgewertet an der Stelle $\boldsymbol{\beta}^{(t)}$ und $\boldsymbol{H}^{(t)}$ die Hessematrix $\boldsymbol{H}$ ausgewertet an $\boldsymbol{\beta}^{(t)}$. Im Iterationsschritt $t+1$ wird daraus mit Iterationsvorschrift (A.2.1) der neue Wert $\boldsymbol{\beta}^{(t+1)}$ berechnet, vorausgesetzt $\boldsymbol{H}^{(t)}$ ist regulär.

Wenn wir nun genauer betrachten wie der Vektor $\boldsymbol{u}$ bzw. die Matrix $\boldsymbol{H}$ aussehen, können wir uns eine explizite Iterationsvorschrift herleiten. Für den Vektor $\boldsymbol{u}$ erhalten wir zunächst (vgl. Gleichung (2.3.2))

$$\frac{\partial l(\boldsymbol{\beta}, \boldsymbol{y})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \frac{y_i - \mu_i}{w_i \phi V(\mu_i)} \frac{\boldsymbol{x}_i}{g'(\mu_i)}.$$

Für die Matrix $\boldsymbol{H}$ lässt sich nicht so einfach eine geschlossene Form finden. Deshalb wird in der Praxis oft mit dem Erwartungswert $\mathbb{E}[\boldsymbol{H}]$ gerechnet. Dies nennt man **Fisher-Scoring-Methode**. Den negativen Erwartungswert $-\mathbb{E}[\boldsymbol{H}]$ nennt man **Fisher-Information**.

Es ist

$$\mathbb{E}\left[\frac{\partial^2 l(\boldsymbol{\beta}, \boldsymbol{Y})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top}\right] = \sum_{i=1}^{n} \mathbb{E}\left[\frac{\partial^2 l_i(\boldsymbol{\beta}, \boldsymbol{Y})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top}\right]$$

und wir verwenden zunächst wieder die Identität (2.1.4)

$$\mathbb{E}\left[\frac{\partial^2 l_i(\boldsymbol{\beta}, \boldsymbol{Y})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top}\right] = -\mathbb{E}\left[\left(\frac{\partial l_i(\boldsymbol{\beta}, \boldsymbol{Y})}{\partial \boldsymbol{\beta}}\right)\left(\frac{\partial l_i(\boldsymbol{\beta}, \boldsymbol{Y})}{\partial \boldsymbol{\beta}^\top}\right)\right].$$

Damit sind die einzelnen Summanden

$$\begin{aligned}
\mathbb{E}\left[\frac{\partial^2 l_i(\boldsymbol{\beta}, \boldsymbol{Y})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top}\right] &= -\mathbb{E}\left[\left(\frac{Y_i - \mu_i}{w_i \phi V(\mu_i)} \frac{\boldsymbol{x}_i}{g'(\mu_i)}\right)\left(\frac{Y_i - \mu_i}{w_i \phi V(\mu_i)} \frac{\boldsymbol{x}_i^\top}{g'(\mu_i)}\right)\right] \\
&= -\frac{\boldsymbol{x}_i \boldsymbol{x}_i^\top \mathbb{E}[(Y_i - \mu_i)^2]}{(w_i \phi V(\mu_i))^2 (g'(\mu_i))^2} \\
&= -\frac{\boldsymbol{x}_i \boldsymbol{x}_i^\top}{w_i \phi V(\mu_i)(g'(\mu_i))^2}
\end{aligned}$$

und es folgt

$$\mathbb{E}\left[\frac{\partial^2 l(\boldsymbol{\beta}, \boldsymbol{Y})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top}\right] = -\sum_{i=1}^{n} \frac{\boldsymbol{x}_i \boldsymbol{x}_i^\top}{w_i \phi V(\mu_i)(g'(\mu_i))^2}.$$

Des Weiteren setzen wir

$$\tilde{w_i} = \frac{1}{w_i V(\mu_i)(g'(\mu_i))^2},$$

womit wir

$$\frac{\partial l(\boldsymbol{\beta}, \boldsymbol{y})}{\partial \boldsymbol{\beta}} = \frac{1}{\phi} \sum_{i=1}^{n} \tilde{w_i} g'(\mu_i)(y_i - \mu_i)\boldsymbol{x}_i \quad \text{und} \quad \mathbb{E}\left[\frac{\partial^2 l(\boldsymbol{\beta}, \boldsymbol{Y})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top}\right] = -\frac{1}{\phi} \sum_{i=1}^{n} \tilde{w_i} \boldsymbol{x}_i \boldsymbol{x}_i^\top$$

erhalten.

Wir definieren die beiden Diagonalmatrizen

$$\boldsymbol{W} = \text{diag}(\tilde{w_1}, \ldots, \tilde{w_n}) \quad \text{und} \quad \boldsymbol{G} = \text{diag}(g'(\mu_1), \ldots, g'(\mu_n)).$$

Man beachte, dass beide Matrizen durch $\boldsymbol{\mu}$ von $\boldsymbol{\beta}$ abhängig sind. Wir bezeichnen die Matrizen in Iteration t, welche von $\boldsymbol{\beta}^{(t)}$ abhängen, mit $\boldsymbol{W}^{(t)}$ bzw. $\boldsymbol{G}^{(t)}$.

Damit erhalten wir für den Vektor $\boldsymbol{u}$ bzw. für die Matrix $\mathbb{E}(\boldsymbol{H})$ an der Stelle $\boldsymbol{\beta}^{(t)}$

$$\boldsymbol{u}^{(t)} = \frac{1}{\phi} \boldsymbol{X}^\top \boldsymbol{W}^{(t)} \boldsymbol{G}^{(t)} (\boldsymbol{y} - \boldsymbol{\mu}^{(t)}) \quad \text{und} \quad \mathbb{E}[\boldsymbol{H}^{(t)}] = -\frac{1}{\phi} \boldsymbol{X}^\top \boldsymbol{W}^{(t)} \boldsymbol{X}.$$

Wenn wir in Iterationsvorschrift (A.2.1) anstelle der Hessematrix $\boldsymbol{H}^{(t)}$ deren Erwartungswert $\mathbb{E}[\boldsymbol{H}^{(t)}]$ einsetzen erhalten wir

$$\boldsymbol{u}^{(t)} + \mathbb{E}[\boldsymbol{H}^{(t)}](\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}) = \boldsymbol{0}$$

und durch Umformung ergibt sich

$$\mathbb{E}[\boldsymbol{H}^{(t)}]\boldsymbol{\beta}^{(t+1)} = \mathbb{E}[\boldsymbol{H}^{(t)}]\boldsymbol{\beta}^{(t)} - \boldsymbol{u}^{(t)}.$$

Durch Einsetzen der Matrixdarstellungen von $\boldsymbol{u}^{(t)}$ und $\mathbb{E}[\boldsymbol{H}^{(t)}]$ erhalten wir

$$-\frac{1}{\phi} \boldsymbol{X}^\top \boldsymbol{W}^{(t)} \boldsymbol{X} \boldsymbol{\beta}^{(t+1)} = -\frac{1}{\phi} \boldsymbol{X}^\top \boldsymbol{W}^{(t)} \boldsymbol{X} \boldsymbol{\beta}^{(t)} - \frac{1}{\phi} \boldsymbol{X}^\top \boldsymbol{W}^{(t)} \boldsymbol{G}^{(t)} (\boldsymbol{y} - \boldsymbol{\mu}^{(t)}).$$

Dies vereinfacht sich zu folgender Vorschrift, welche nun unabhängig von ϕ ist,

$$\boldsymbol{X}^\top \boldsymbol{W}^{(t)} \boldsymbol{X} \boldsymbol{\beta}^{(t+1)} = \boldsymbol{X}^\top \boldsymbol{W}^{(t)} \left(\boldsymbol{X} \boldsymbol{\beta}^{(t)} + \boldsymbol{G}^{(t)} (\boldsymbol{y} - \boldsymbol{\mu}^{(t)}) \right). \tag{A.2.2}$$

Definieren wir die sogenannten Pseudobeobachtungen $\boldsymbol{z}^{(t)} = \boldsymbol{X} \boldsymbol{\beta}^{(t)} + \boldsymbol{G}^{(t)} (\boldsymbol{y} - \boldsymbol{\mu}^{(t)})$ und multiplizieren Gleichung (A.2.2) von Links mit $(\boldsymbol{X}^\top \boldsymbol{W}^{(t)} \boldsymbol{X})^{-1}$ erhalten wir

$$\boldsymbol{\beta}^{(t+1)} = (\boldsymbol{X}^\top \boldsymbol{W}^{(t)} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{W}^{(t)} \boldsymbol{z}^{(t)}. \tag{A.2.3}$$

In Iterationsschritt $(t+1)$ sind $\boldsymbol{z}^{(t)}$, $\boldsymbol{W}^{(t)}$ und $\boldsymbol{G}^{(t)}$ aus dem vorherigen Iterationsschritt bekannt. Der Parameterschätzer $\hat{\boldsymbol{\beta}}$ ist der Wert von $\boldsymbol{\beta}^{(t)}$, den wir bei Konvergenz im letzten Schritt der Iteration erhalten.

Bemerkung 1. Die Iterationsvorschrift (A.2.3) ist vergleichbar mit dem Least-Squares-Schätzer eines linearen Modells $\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}$, wofür $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y}$ die explizite Form des Parameterschätzers ist. Hier rechnen wir in jedem Schritt der Iteration mit den Pseudobeobachtungen $\boldsymbol{z}^{(t)}$ und haben noch zusätzlich eine Gewichtsmatrix $\boldsymbol{W}^{(t)}$. Deshalb nennt man dieses Verfahren **iteratives gewichtetes Least-Squares-Verfahren**.

Bemerkung 2. Für kanonische Link-Funktionen ist das Score-System gegeben durch (2.3.3). Die Elemente der Fisher-Information sind dann

$$\begin{aligned} -\mathbb{E}\left[\frac{\partial^2 l(\boldsymbol{\beta}, \boldsymbol{Y})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top}\right] &= \mathbb{E}\left[\sum_{i=1}^{n} \frac{b''(\theta_i)}{w_i \phi} \boldsymbol{x}_i \frac{\partial \theta_i}{\partial \boldsymbol{\beta}^\top}\right] \\ &= \mathbb{E}\left[\sum_{i=1}^{n} \frac{V(\mu_i)}{w_i \phi} \boldsymbol{x}_i \boldsymbol{x}_i^\top\right] \\ &= \sum_{i=1}^{n} \frac{V(\mu_i)}{w_i \phi} \boldsymbol{x}_i \boldsymbol{x}_i^\top, \end{aligned}$$

d. h. im Falle einer kanonischen Link-Funktion gilt $\mathbb{E}[\boldsymbol{H}] = \boldsymbol{H}$ und die Fisher-Scoring-Methode stimmt mit dem Newton-Raphson-Verfahren überein.

Eigenschaften des Schätzers

Entwickeln wir $l(\hat{\boldsymbol{\beta}}, \boldsymbol{y})$ um den wahren Parameter $\boldsymbol{\beta}$ erhalten wir

$$l(\hat{\boldsymbol{\beta}}, \boldsymbol{y}) \approx l(\boldsymbol{\beta}, \boldsymbol{y}) + \boldsymbol{u}^\top (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + \frac{1}{2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \boldsymbol{H} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}).$$

Durch Lösen von

$$\frac{\partial l(\hat{\boldsymbol{\beta}}, \boldsymbol{y})}{\partial \hat{\boldsymbol{\beta}}} \approx \boldsymbol{u} + \boldsymbol{H}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \stackrel{!}{=} \boldsymbol{0}$$

nach $\hat{\boldsymbol{\beta}}$ ergibt sich

$$\hat{\boldsymbol{\beta}} \approx \boldsymbol{\beta} - \boldsymbol{H}^{-1}\boldsymbol{u}.$$

Dabei werden $\boldsymbol{u}$ und $\boldsymbol{H}$ an der Stelle $\boldsymbol{\beta}$ betrachtet, wobei $\boldsymbol{\beta}$ aber der unbekannte wahre Parameter ist. Durch Einsetzen für $\boldsymbol{u}$ und $\boldsymbol{H}$ folgt

$$\hat{\boldsymbol{\beta}} \approx \boldsymbol{\beta} + (\boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{G} (\boldsymbol{y} - \boldsymbol{\mu}).$$

Damit gilt

$$\mathbb{E}[\hat{\boldsymbol{\beta}}] \approx \boldsymbol{\beta} + (\boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{G} \underbrace{(\mathbb{E}[\boldsymbol{Y}] - \boldsymbol{\mu})}_{=\mathbf{0}}, \text{ also } \mathbb{E}[\hat{\boldsymbol{\beta}}] \approx \boldsymbol{\beta}.$$

Für die Varianz von $\boldsymbol{Y}$ gilt aufgrund der Definition von $\tilde{w}_i$, dass

$$\mathbb{V}\mathrm{ar}[Y_i] = \frac{\phi}{\tilde{w}_i (g'(\mu_i))^2}, \quad i = 1, \ldots, n \quad \text{und} \quad \mathbb{V}\mathrm{ar}[\boldsymbol{Y}] = \phi(\boldsymbol{G}\boldsymbol{W}\boldsymbol{G})^{-1}.$$

Daraus folgt, dass die Varianz von $\hat{\boldsymbol{\beta}}$ die Inverse der Informationsmatrix ist, denn

$$\mathbb{V}\mathrm{ar}[\hat{\boldsymbol{\beta}}] \approx (\boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{G}\ \mathbb{V}\mathrm{ar}[\boldsymbol{Y}]\ ((\boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{G})^\top,$$

also

$$\mathbb{V}\mathrm{ar}[\hat{\boldsymbol{\beta}}] \approx \phi(\boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X})^{-1}.$$

Fahrmeir und Kaufmann (1985) haben gezeigt, dass unter gewissen Regularitätsbedingungen

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} N_p(\mathbf{0}, n\mathbb{V}\mathrm{ar}(\hat{\boldsymbol{\beta}}))$$

gilt. Als *Schätzer* für die Varianz von $\hat{\boldsymbol{\beta}}$ wird ein Plug-in-Schätzer verwendet, wobei in der Matrix $\boldsymbol{W}$ der Schätzer $\hat{\boldsymbol{\beta}}$ für $\boldsymbol{\beta}$ eingesetzt wird, d.h.

$$\widehat{\mathbb{V}\mathrm{ar}}[\hat{\boldsymbol{\beta}}] = \phi(\boldsymbol{X}^\top \hat{\boldsymbol{W}} \boldsymbol{X})^{-1}.$$

Zu beachten ist, dass $\widehat{\mathbb{V}\mathrm{ar}}[\hat{\boldsymbol{\beta}}]$ aber auch vom Dispersionsparameter ϕ abhängen kann. Meistens ist dieser bekannt und wir können bei der Schätzung der Varianz von $\hat{\boldsymbol{\beta}}$ wie oben vorgehen. Ist ϕ allerdings unbekannt, so wird für gewöhnlich auch dafür ein Schätzer eingesetzt. Für ein $(1-\alpha)$-Konfidenzintervall folgt

$$\mathbb{P}\left[-z_{1-\frac{\alpha}{2}} \leq \frac{\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}}{\sqrt{\widehat{\mathbb{V}\mathrm{ar}}[\hat{\boldsymbol{\beta}}]}} \leq z_{1-\frac{\alpha}{2}} \right] = 1 - \alpha$$

und wir erhalten als approximatives Intervall

$$\hat{\boldsymbol{\beta}} \pm z_{1-\frac{\alpha}{2}} \sqrt{\widehat{\mathbb{V}\mathrm{ar}}[\hat{\boldsymbol{\beta}}]}.$$

Schätzung des Dispersionsparameters

Definition 8 (Generalisierte Pearson Statistik)**.** *Die* ***Pearson Statistik*** *ist definiert als*

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{w_i V(\hat{\mu}_i)}.$$

Unter bestimmten Regularitätsbedingungen gilt $X^2 \overset{a}{\sim} \phi\chi^2_{n-p}$.

Für die Exponentialfamilie gilt $\mathbb{V}\text{ar}[Y_i] = w_i \phi V(\mu_i)$ für alle $i = 1, \ldots, n$. Demnach ist

$$\phi = \frac{\mathbb{V}\text{ar}[Y_i]}{w_i V(\mu_i)} \quad \text{für alle } i = 1, \ldots, n.$$

Man kann nun $(y_i - \hat{\mu}_i)^2$ als Schätzer für $\mathbb{V}\text{ar}[Y_i]$ sehen und erhält als Schätzer für ϕ die mittlere biaskorrigierte Größe

$$\tilde{\phi} = \frac{1}{n-p} \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{w_i V(\hat{\mu}_i)} = \frac{1}{n-p} X^2.$$

Wegen $X^2 \overset{a}{\sim} \phi\chi^2_{n-p}$ folgt

$$\mathbb{E}[\tilde{\phi}] = \frac{1}{n-p}\mathbb{E}[X^2] \approx \frac{\phi(n-p)}{n-p} = \phi.$$

A.3. Das saturierte Modell

Das saturierte (volle) Modell dient als Vergleichsmodell bei der Bestimmung der Anpassungsgüte eines GLMs. Das saturierte Modell an sich hat wenig Aussagekraft, denn es liefert zwar perfekte Anpassung, allerdings auch keine Datenreduktion.

Es seien $Y_i \overset{ind}{\sim} Exponentialfamilie(\theta_i)$ für $i = 1, \ldots, n$ gegeben. Die Log-Likelihood ist gegeben als

$$l(\boldsymbol{\theta}, \boldsymbol{y}) = \sum_{i=1}^{n} \left(\frac{y_i\theta_i - b(\theta_i)}{w_i\phi} + c(y_i, \phi) \right).$$

Die Score-Funktionen nach μ_i für $i = 1, \ldots, n$ sind

$$\begin{aligned} \frac{\partial l(\boldsymbol{\theta}, \boldsymbol{y})}{\partial \mu_i} &= \sum_{i=1}^{n} \left(\frac{\partial l_i(\theta_i, y_i)}{\partial \theta_i} \, \frac{\partial \theta_i}{\partial \mu_i} \right) \\ &= \sum_{i=1}^{n} \frac{y_i - b'(\theta_i)}{w_i \phi V(\mu_i)} \\ &= \sum_{i=1}^{n} \frac{y_i - \mu_i}{\mathbb{V}\text{ar}[Y_i]}. \end{aligned}$$

Im saturierten Modell gibt es keinerlei Einschränkungen für μ_i und die Score-Funktionen sind Null für $\hat{\mu}_i = y_i$ für $i = 1, \ldots, n$.

A.4. Die Logistische Verteilung

Siehe Balakrishnan (1991). Die Dichte einer logistischen Zufallsvariable $X \sim L(\mu, \tau)$ mit Parametern $-\infty < \mu < \infty$ und $\tau > 0$ ist gegeben durch

$$f(x, \mu, \tau) = \frac{\exp\{-\frac{x-\mu}{\tau}\}}{\tau(1 + \exp\{-\frac{x-\mu}{\tau}\})^2}, \quad -\infty < x < \infty.$$

Die Verteilungsfunktion ist

$$F(x, \mu, \tau) = \frac{\exp\{\frac{x-\mu}{\tau}\}}{1 + \exp\{\frac{x-\mu}{\tau}\}}.$$

Es gilt $\mathbb{E}[X] = \mu$ und $\mathbb{V}\text{ar}[X] = \tau^2\pi^2/3$.

Abbildung A.1 zeigt Dichte- und Verteilungsfunktion der logistischen Verteilung für verschiedene Werte von μ und τ. Ist $X \sim L(0, 1)$, dann hat X die Verteilungsfunktion

$$F(x, 0, 1) = \frac{\exp\{x\}}{1 + \exp\{x\}},$$

was der Umkehrfunktion der Logit-Link-Funktion entspricht.

A.5. Die Gumbel-Verteilung

Siehe Kotz und Nadarajah (2000). Es sei X eine Zufallsvariable aus der Gumbel-Verteilung, auch Extremwertverteilung Typ 1 genannt. Dann hat X die Dichtefunktion

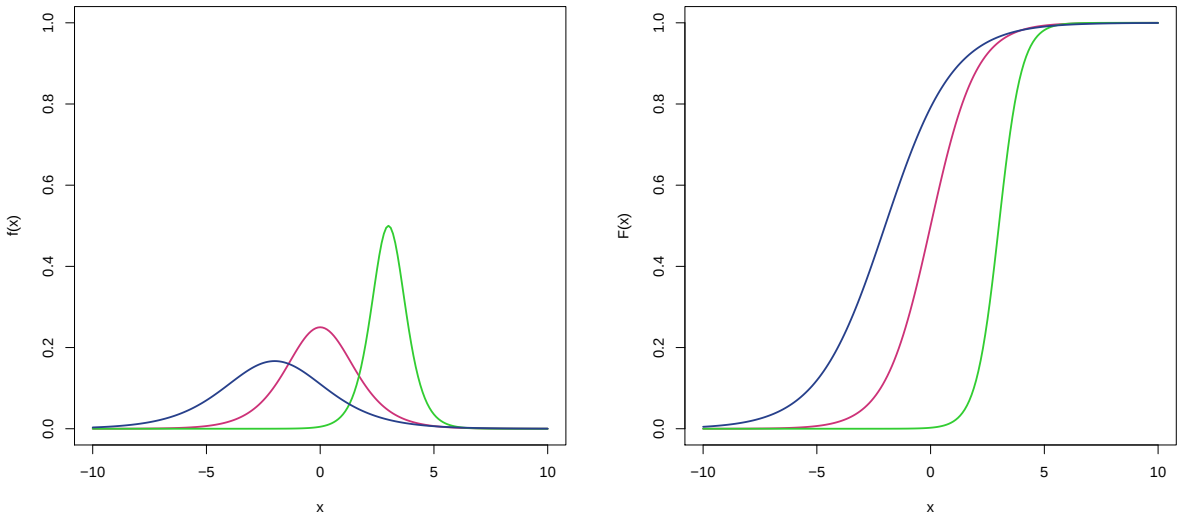
$$f(x, \mu, \tau) = \frac{1}{\tau} \exp\left\{-\frac{x-\mu}{\tau}\right\} \exp\left\{-\exp\left\{-\frac{x-\mu}{\tau}\right\}\right\}, \quad x \in \mathbb{R},\ \mu \in \mathbb{R} \text{ und } \tau > 0.$$

Die Verteilungsfunktion ist

$$F(x, \mu, \tau) = \exp\left\{-\exp\left\{-\frac{x-\mu}{\tau}\right\}\right\}.$$

Es gilt $\mathbb{E}[X] = \mu + \gamma\tau$ und $\mathbb{V}\text{ar}[X] = \tau^2\pi^2/6$, wobei $\gamma \approx 0.57722$ die Euler-Mascheroni-Konstante ist. Abbildung A.2 zeigt Dichte- und Verteilungsfunktion der Gumbel-Verteilung für verschiedene Werte von μ und τ.

Ist die Zufallsvariable X aus der Extremwertverteilung, dann gilt dies auch für die Zufallsvariable $-X$, welche dann die Verteilungsfunktion



(a) Dichtefunktion

(b) Verteilungsfunktion

Abbildung A.1.: Logistische Verteilung für $\mu = 0$ und $\tau = 1$ (violett), $\mu = 3$ und $\tau = 0.5$ (grün) und $\mu = -2$ und $\tau = 1.5$ (blau).

$$1 - F(-x, \mu, \tau) = 1 - \exp\left\{-\exp\left\{-\frac{-x-\mu}{\tau}\right\}\right\}$$

hat. Falls $\mu = 0$ und $\tau = 1$ ist die Verteilungsfunktion von $-X$ gegeben als

$$1 - F(-x, 0, 1) = 1 - \exp\{-\exp\{x\}\},$$

was der Umkehrfunktion des Complementary-Log-Log-Links entspricht.

A.6. Die Multivariate Normalverteilung

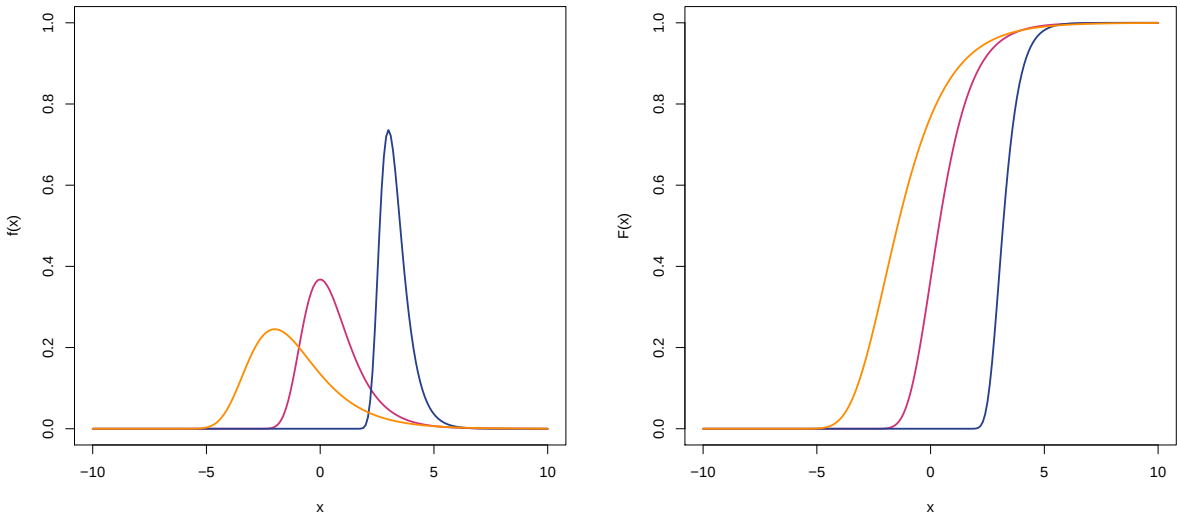
Siehe Fahrmeir et al. (2013, S. 648).

Definition 9 (Multivariate Normalverteilung)**.** *Es sei* $\boldsymbol{X} = (X_1, \ldots, X_d)^\top$ *ein d-dimensionaler Zufallsvektor. Dann ist* $\boldsymbol{X}$ *multivariat normalverteilt, wenn* $\boldsymbol{X}$ *die folgende Dichte hat:*

$$f(\boldsymbol{x}) = (2\pi)^{-\frac{d}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\},$$

wobei $\boldsymbol{\mu} \in \mathbb{R}^d$ *und* $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$ *positiv definit.*

Dann folgt $\mathbb{E}[\boldsymbol{X}] = \boldsymbol{\mu}$ *und* $\mathbb{C}ov[\boldsymbol{X}] = \boldsymbol{\Sigma}$ *und wir schreiben*



(a) Dichtefunktion

(b) Verteilungsfunktion

Abbildung A.2.: Gumbel-Verteilung für $\mu = 0$ und $\tau = 1$ (violett), $\mu = 3$ und $\tau = 0.5$ (blau) und $\mu = -2$ und $\tau = 1.5$ (orange).

$$\boldsymbol{X} \sim N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

Wenn der Index d aus dem Zusammenhang klar ist, wird er auch oft weggelassen.

Definition 10 (Singuläre multivariate Normalverteilung)**.** *Es sei $\boldsymbol{X} \sim N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Dann folgt $\boldsymbol{X}$ einer singulären (degenerierten) multivariaten Normalverteilung, wenn $\mathrm{rk}(\boldsymbol{\Sigma}) = r < d$ ist. Die Dichte wird dann oft mithilfe der Precision-Matrix $\boldsymbol{R}$ mit $\mathrm{rk}(\boldsymbol{R}) = r < p$ ausgedrückt:*

$$f(\boldsymbol{x}) \propto \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{R}(\boldsymbol{x}-\boldsymbol{\mu})\right\}.$$

Die Matrix $\boldsymbol{R}$ kann als generalisierte Inverse von $\boldsymbol{\Sigma}$ gewählt werden.

Satz 1. *Es sei $\boldsymbol{X}$ singulär multivariat normalverteilt, d. h. $\boldsymbol{X} \sim N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ mit $\mathrm{rk}(\boldsymbol{\Sigma}) = \mathrm{rk}(\boldsymbol{R}) = r < d$. Ist $(\boldsymbol{G}|\boldsymbol{H})$ eine orthogonale Matrix, wobei die Spalten der $(d \times r)$-Matrix $\boldsymbol{G}$ eine Basis des Spaltenraums von $\boldsymbol{\Sigma}$ und die Spalten von $\boldsymbol{H}$ eine Basis des Kerns von $\boldsymbol{\Sigma}$ sind. Wir betrachten folgende Transformation:*

$$\begin{pmatrix}\boldsymbol{Y}_1\\ \boldsymbol{Y}_2\end{pmatrix} = (\boldsymbol{G}|\boldsymbol{H})^\top \boldsymbol{X} = \begin{pmatrix}\boldsymbol{G}^\top \boldsymbol{X}\\ \boldsymbol{H}^\top \boldsymbol{X}\end{pmatrix}.$$

Dann ist $\boldsymbol{Y}_1$ der stochastische Teil von $\boldsymbol{X}$ und ist nicht-singulär normalverteilt mit

$$\boldsymbol{Y}_1 \sim N_r(\boldsymbol{G}^\top \boldsymbol{\mu}, \boldsymbol{G}^\top \boldsymbol{\Sigma} \boldsymbol{G})$$

und hat die Dichte

$$f(\boldsymbol{y}_1) = \frac{1}{(2\pi)^{\frac{r}{2}} (\prod_{i=1}^{r} \lambda_i)^{\frac{1}{2}}} \exp\left\{ -\frac{1}{2} (\boldsymbol{y}_1 - \boldsymbol{G}^\top \boldsymbol{\mu})^\top (\boldsymbol{G}^\top \boldsymbol{\Sigma} \boldsymbol{G})^{-1} (\boldsymbol{y}_1 - \boldsymbol{G}^\top \boldsymbol{\mu}) \right\},$$

wobei $\lambda_1, \ldots, \lambda_r$ die r positiven Eigenwerte von $\boldsymbol{\Sigma}$ sind. Der Vektor $\boldsymbol{Y}_2$ ist der deterministische Teil von $\boldsymbol{X}$ und es gilt

$$\mathbb{E}[\boldsymbol{Y}_2] = \boldsymbol{H}^\top \boldsymbol{\mu} \quad \textit{und} \quad \mathbb{V}ar[\boldsymbol{Y}_2] = \boldsymbol{0}.$$

Hier ist $\mathbb{V}ar[\boldsymbol{Y}_2]$ der Vektor der Varianzen von $\boldsymbol{Y}_2$.

A.7. Kovarianz der Multinomialverteilung

Es sei $(Y_1, \ldots, Y_c)^\top \sim M(n, \pi)$. Zu zeigen ist, dass $\mathbb{C}\mathrm{ov}[Y_j, Y_k] = -n\pi_j\pi_k$ für $j \neq k$ ist. Per definitionem ist

$$\mathbb{C}\mathrm{ov}[Y_j, Y_k] = \mathbb{E}[Y_j Y_k] - \mathbb{E}[Y_j]\mathbb{E}[Y_k].$$

Die beiden Erwartungswerte $\mathbb{E}[Y_j] = n\pi_j$ und $\mathbb{E}[Y_k] = n\pi_k$ sind dabei bekannt. Zu berechnen ist der Erwartungswert $\mathbb{E}[Y_j Y_k]$. Dazu definieren wir die Dummy-Variablen

$$\tilde{Y}_{ij} = \begin{cases} 1 & \text{falls Versuch } i \text{ Augsgang } j \text{ hat,} \\ 0 & \text{sonst.} \end{cases}$$

Wegen $\sum_{i=1}^{n} \tilde{Y}_{ij} = Y_j$ folgt damit

$$\begin{aligned} \mathbb{E}[Y_j Y_k] &= \mathbb{E}\left[\sum_{i=1}^{n} \tilde{Y}_{ij} \sum_{i'=1}^{n} \tilde{Y}_{i'k} \right] \\ &= \sum_{i=1}^{n} \sum_{i'=1}^{n} \mathbb{E}[\tilde{Y}_{ij} \tilde{Y}_{i'k}] \\ &= \sum_{i} \sum_{i' \neq i} \mathbb{E}[\tilde{Y}_{ij}] \mathbb{E}[\tilde{Y}_{i'k}] \\ &= n(n-1)\pi_j\pi_k. \end{aligned}$$

Die vorletzte Gleichung folgt dabei aus

$$\mathbb{E}[\tilde{Y}_{ij} \tilde{Y}_{i'k}] = 0 \quad \text{für } i' = i, \text{ da} \sum_{j=1}^{c} \tilde{Y}_{ij} = 1$$

und

$$\mathbb{E}[\tilde{Y}_{ij}\tilde{Y}_{i'k}] = \mathbb{E}[\tilde{Y}_{ij}]\mathbb{E}[\tilde{Y}_{i'k}] \quad \text{für } i' \neq i, \text{ da } \tilde{Y}_{ij} \text{ und } \tilde{Y}_{i'k} \text{ unabhängig sind.}$$

Damit folgt insgesamt

$$\begin{aligned}\mathbb{C}\text{ov}[Y_j, Y_k] &= n(n-1)\pi_j\pi_k - n^2\pi_j\pi_k \\ &= n\pi_j\pi_k[(n-1)-n] \\ &= -n\pi_j\pi_k.\end{aligned}$$

A.8. Natürliche kubische Splines

In diesem Abschnitt folgt ein Beweis der Optimalitätseigenschaft von natürlichen kubischen Splines aus Abschnitt 4.1.3. Vergleiche dazu Green und Silverman (1994).

Wir starten damit zu zeigen, dass natürliche kubische Splines die glattesten Interpolierenden der Daten (z_i, y_i) für $i = 1, \ldots, n$ sind, d. h. sie lösen

$$\min_f \quad \int (f''(z))^2 dz. \tag{A.8.1}$$

Es sei $f(z)$ ein natürlicher kubische Spline für die Punkte (z_i, y_i), $i = 1, \ldots, n$, mit $z_i < z_{i+1}$. Es gilt $f(z_i) = y_i$ und $f''(z_1) = f''(z_n) = 0$. Es sei $g(z)$ ein beliebige zweimal stetig differenzierbare Funktion mit $g(z_i) = y_i$. Setze $h(z) = g(z) - f(z)$.

Als Hilfsresultat berechnen wir zunächst das Integral mit den gemischten Termen. Durch partielle Integration und die Randbedingungen erhalten wir

$$\begin{aligned}\int_{z_1}^{z_n} f''(t)h''(t)dt &= [f''(t)h'(t)]_{z_1}^{z_n} - \int_{z_1}^{z_n} f'''(t)h'(t)dt \\ &= [\underbrace{f''(z_n)}_{=0} h'(z_n) - \underbrace{f''(z_1)}_{=0} h'(z_1)] - \sum_{i=1}^{n-1} \int_{z_i}^{z_{i+1}} f'''(t)h'(t)dt \\ &= -\sum_{i=1}^{n-1} f'''(z_i^+) \int_{z_i}^{z_{i+1}} h'(t)dt \\ &= -\sum_{i=1}^{n-1} f'''(z_i^+)[\underbrace{h(z_{i+1})}_{=0} - \underbrace{h(z_i)}_{=0}] = 0.\end{aligned}$$

Dabei wurde verwendet, dass f auf jedem Intervall $[z_i, z_{i+1})$ eine Polynom dritten Grades und demnach f''' konstant auf den Intervallen ist. Der Wert sei $f'''(z_i^+)$. Des Weiteren ist $h(z_i) = 0$ für alle $i = 1, \ldots, n$, weil f und g in z_i den gleichen Funktionswert haben.

Mit dem Hilfsresultat zeigen wir nun, dass f das Optimierungsproblem (A.8.1) löst.

$$\begin{aligned}\int_{z_1}^{z_n}(g''(t))^2dt &= \int_{z_1}^{z_n}(h''(t)+f''(t))^2dt \\ &= \int_{z_1}^{z_n}(h''(t))^2dt + 2\int_{z_1}^{z_n}h''(t)f''(t)dt + \int_{z_1}^{z_n}(f''(t))^2dt \\ &= \int_{z_1}^{z_n}(h''(t))^2dt + \int_{z_1}^{z_n}(f''(t))^2dt \\ &\geq \int_{z_1}^{z_n}(f''(t))^2dt.\end{aligned}$$

Wir betrachten nun das Optimierungsproblem

$$\min_f \quad \sum_{i=1}^{n}(y_i - f(z_i))^2 + \lambda\int(f''(z))^2dz. \tag{A.8.2}$$

In Abschnitt 4.1.3 wurde behauptet, dass die Lösung dieses Problems ein natürlicher kubischer Spline ist. Nehmen wir an, dass g eine Lösung des Optimierungsproblems (A.8.2), aber kein natürlicher kubischer Spline ist. Es sei $\tilde{g}$ die natürliche kubische Spline Interpolierende von $(z_i, g(z_i))$, dann gilt $\tilde{g}(z_i) = g(z_i)$. Daraus folgt, dass

$$\sum_{i=1}^{n}(y_i - \tilde{g}(z_i))^2 = \sum_{i=1}^{n}(y_i - g(z_i))^2$$

und wegen obigem Resultat folgt mit $\lambda > 0$

$$\sum_{i=1}^{n}(y_i - \tilde{g}(z_i))^2 + \lambda\int(\tilde{g}''(z))^2dz < \sum_{i=1}^{n}(y_i - g(z_i))^2 + \lambda\int(g''(z))^2dz,$$

was ein Widerspruch ist.

A.9. Rechenregeln für die Spur

Definition 11 (Spur)**.** *Es sei $A = (a_{ij})$ eine $(n \times n)$-Matrix. Dann ist die **Spur** von A die Summe der Diagonalelemente, also*

$$\mathrm{tr}(A) = \sum_{i=1}^{n} a_{ii}.$$

Die folgenden Eigenschaften der Spur wurden in Abschnitt 4.3 verwendet:

1. Es seien A, B zwei $(n \times n)$-Matrizen, dann gilt
$$\mathrm{tr}(A + B) = \mathrm{tr}(A) + \mathrm{tr}(B).$$
2. Es seien $A \in (n \times m)$ und $B \in (m \times n)$, dann gilt
$$\mathrm{tr}(AB) = \mathrm{tr}(BA).$$

A.10. Odds-Ratio und stochastische Unabhängigkeit

Es seien X und Y zwei zweistufige Faktoren, deren Stufen jeweils mit $\{0, 1\}$ codiert sind. Die Randwahrscheinlichkeiten seien

$$\mathbb{P}[X = 1] = p_x \quad \text{und} \quad \mathbb{P}[X = 0] = 1 - p_x,$$

sowie

$$\mathbb{P}[Y = 1] = p_y \quad \text{und} \quad \mathbb{P}[Y = 0] = 1 - p_y.$$

Die gemeinsame Verteilung von X und Y sei gegeben durch

$$\mathbb{P}[X = i, Y = j] = p_{ij}, \quad \text{für } i = 1, 2 \text{ und } j = 1, 2.$$

Dies impliziert folgende Kontingenztafel:

	$Y = 1$	$Y = 0$	
$X = 1$	p_{11}	p_{10}	p_x
$X = 0$	p_{01}	p_{00}	$1 - p_x$
	p_y	$1 - p_y$	1

Aus dem Satz der totalen Wahrscheinlichkeit und der Definition der bedingten Wahrscheinlichkeit folgt:

$$p_x = p_{11} + p_{10}, \quad 1 - p_x = p_{01} + p_{00}, \quad p_y = p_{11} + p_{01}, \quad \text{und} \quad 1 - p_y = p_{10} + p_{00}.$$

Wir wollen zeigen, dass das Odds-Ratio genau dann 1 ist, wenn X und Y stochastisch unabhängig sind.

Beweis.
1) Annahme: Das Odds-Ratio ist 1. Dann gilt

$$1 = \frac{p_{11} \cdot p_{00}}{p_{01} \cdot p_{10}} \quad \Longleftrightarrow \quad p_{11} \cdot p_{00} = p_{01} \cdot p_{10}. \tag{A.10.1}$$

Wir müssen nun zeigen, dass das Produkt der Randwahrscheinlichkeiten die Zellwahrscheinlichkeiten ergibt. Dies können wir z. B. für p_{11} machen:

$$\begin{aligned} p_x \cdot p_y &= (p_{11} + p_{10}) \cdot (p_{11} + p_{01}) \\ &= p_{11}^2 + p_{11}p_{01} + p_{10}p_{11} + p_{10}p_{01} \\ &\overset{(A.10.1)}{=} p_{11} \underbrace{(p_{11} + p_{01} + p_{10} + p_{00})}_{=1} = p_{11}. \end{aligned}$$

Analog kann man für die anderen Zellen zeigen, dass

$$p_x \cdot (1 - p_y) = p_{10}, \quad (1 - p_x) \cdot p_y = p_{01} \quad \text{und} \quad (1 - p_x) \cdot (1 - p_y) = p_{00},$$

woraus dann die stochastische Unabhängigkeit von X und Y folgt.

2) Annahme: X und Y sind stochastisch unabhängig. Dann ist die gemeinsame Wahrscheinlichkeit das Produkt der Randwahrscheinlichkeiten und als Kontingenztabelle erhalten wir dann:

	$Y = 1$	$Y = 0$	
$X = 1$	$p_x \cdot p_y$	$p_x \cdot (1 - p_y)$	p_x
$X = 0$	$(1 - p_x) \cdot p_y$	$(1 - p_x) \cdot (1 - p_y)$	$1 - p_x$
	p_y	$1 - p_y$	1

Dann ist das Odds-Ratio gleich 1, denn

$$\frac{p_x \cdot p_y \cdot (1 - p_x) \cdot (1 - p_y)}{p_x \cdot (1 - p_y) \cdot (1 - p_x) \cdot p_y} = 1.$$

D. h. falls X und Y stochastisch unabhängig sind ist das Odds-Ratio gleich 1. Die Chancen für $X = 1$ sind gegeben $Y = 1$ oder $Y = 0$ gleich.

Literatur

Agresti, A. (2013). *Categorical Data Analysis* (Dritte Aufl.). Hoboken, NJ: John Wiley & Sons.

Balakrishnan, N. (1991). Statistics: A Series of Textbooks and Monographs. In *Handbook of the Logistic Distribution.* CRC Press.

Cox, D. R. & Hinkley, D. V. (1974). *Theoretical Statistics.* London: Chapman and Hall.

De Boor, C. (2001). *A Practical Guide to Splines.* New York, NY: Springer.

Eilers, P. & Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, *11* (2), 89–121.

Eilers, P. & Marx, B. D. (2010). Splines, knots, and penalties. *Wiley Interdisciplinary Reviews. Computational Statistics*, *2* (6), 637–653.

Fahrmeir, L. & Kaufmann, H. (1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *Annals of Statistics*, *13* (1), 342–368.

Fahrmeir, L., Kneib, T., Lang, S. & Marx, B. (2013). *Regression. Models, Methods and Applications.* Heidelberg: Springer.

Green, P. & Silverman, B. (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach.* CRC Press.

Hinkley, D. V., Reid, N. & Snell, E. J. (Hrsg.). (1991). *Statistical Theory and Modelling. In Honour of Sir David Cox, FRS.* London: Chapman and Hall, Ltd.

Kotz, S. & Nadarajah, S. (2000). *Extreme Value Distributions. Theory and Applications.* London: Imperial College Press.

McCullagh, P. & Nelder, J. A. (1983). *Generalized Linear Models* (Zweite Aufl.). London: Chapman & Hall.

R Core Team. (2014). R: A language and environment for statistical computing [Software-Handbuch]. Vienna, Austria. Zugriff auf `http://www.R-project.org/`

Ruppert, D., Wand, P. & Carroll, R. J. (2003). *Semiparametric Regression.* Cambridge University Press.

Tutz, G. (2012). *Regression for Categorical Data.* Cambridge University Press.

Wahba, G. (1990). *Spline Models for Observational Data* (Bd. 59). Philadelphia, PA: Society for Industrial and Applied Mathematics (SIAM).

Wood, S. N. (2001). mgcv: GAMs and generalized ridge regression for R. *R News*, *1* (2), 20–25.

Wood, S. N. (2006). *Generalized Additive Models: An Introduction with R.* Chapman and Hall/CRC.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)*, *73* (1), 3–36.

Wood, S. N. (2013). On p-values for smooth components of an extended generalized additive model. *Biometrika*, *100* (1), 221–228.

Wood, S. N. (2015). Package 'mgcv' [Software-Handbuch]. Zugriff auf `http://cran.r-project.org/web/packages/mgcv/mgcv.pdf`

Wurzer, P., Rappl, T., Friedl, H., Kamolz, L.-P., Spendel, S., Hoflehner, H. & Parvizi, D. (2014). The Austrian breast implant register: Recent trends in implant-based breast surgery. *Aesthetic Plastic Surgery*, *38*, 1109–1115.

Yee, T. W. (2010). The VGAM package for categorical data analysis. *Journal of Statistical Software*, *32* (10), 1–34.

Yee, T. W. (2015). Package 'VGAM' [Software-Handbuch]. Zugriff auf `http://cran.r-project.org/web/packages/VGAM/VGAM.pdf`