Christof PETERNELL, BSc

# Statistische Modellierung von Leistungen einer Krankenversicherung

MASTERARBEIT
zur Erlangung des akademischen Grades eines Diplom-Ingenieur

Masterstudium Finanz- und Versicherungsmathematik


Technische Universität Graz

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

Graz, März 2014

## Statutory Declaration

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 marked all material which has been quoted either literally or by content from the used sources.

Graz, _____ _____

Date Signature

## Eidesstattliche Erklärung[1]

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.

Graz, am _____ _____

Datum Unterschrift

[1] Beschluss der Curricula-Kommission für Bachelor-, Master- und Diplomstudien vom 10.11.2008; Genehmigung des Senates am 1.12.2008

## Zusammenfassung

In dieser Masterarbeit werden die Leistungen einer privaten Krankenversicherung, die innerhalb zweier Tarifklassen (ambulante Kosten und Spitalskosten) an die Versicherungsnehmer ausbezahlt werden, analysiert und passende Modelle gesucht, um deren Verhalten zu beschreiben. Dazu werden sowohl einige Erwartungswertmodelle als auch Methoden der Quantilen Regression vorgestellt und angewendet. Der behandelte Datensatz beinhaltet Leistungen dreier aufeinanderfolgender Jahre (2010, 2011, 2012) sowohl weiblicher als auch männlicher Versicherter. Es soll dabei auf den Verlauf der Leistungsansprüche über die betrachteten Kalenderjahre, als auch auf die geschlechterspezifischen Unterschiede eingegangen werden. Eine gemeinsame Modellierung beider Geschlechter wird durch die Einführung der Unisex-Richtlinie durch Solvency II angeregt. Zusätzlich werden einige Leistungsarten hervorgehoben, die den Großteil der angefallenen Leistungen verursachen.

Für die Schätzung der Modelle und die Generierung der Graphiken wurde das Statistikprogramm „R 3.0.2“ verwendet. Die Theorie zu den Methoden der Quantilen Regression beruht auf dem gleichnamigen Buch von Roger Koenker (Koenker, 2005).



## Abstract

In this master thesis the payments of a private health insurance company within two tariff classes (ambulant treatments and treatments in hospital) are estimated by the use of quantile regression models and mean models. The data set contains observations of three following years (2010, 2011 and 2012) of male and female insurants, for which a trend of payments over the years is analysed. Also the gap between male and female insurants is considered by using the different model classes. Since Solvency II there have to be unisex insurance rates, so the male and female insurants are also considered together to fit the estimated models. In addition the insurance benefit types that cause the biggest payments in the insurance portfolio are analysed.

For estimating the models and generating the plots the statistical program „R 3.0.2“ is used. The theory of quantile regression is based on the book „Quantile Regression“ by Roger Koenker (Koenker, 2005).

# Inhaltsverzeichnis

# 1 Lineare Regression

Der Begriff *Regressionsanalyse* beinhaltet statistische Analyseverfahren, deren Ziel es ist, eine Beziehung zwischen der abhängigen Variable $Y$ (auch *Responsevariable* genannt) und der unabhängigen Variable $x$ (beziehungsweise *Prädiktor*, *erklärende Variable* oder auch *Kovariable* genannt) festzustellen (Fahrmeir et al., 2013).

In der linearen Regression werden einige fundamentale Annahmen getroffen:

1. Für jeden einzelnen Wert $x$ ist die Responsevariable $Y$ eine Zufallsvariable, deren Erwartungswert von $x$ abhängt. Die $x$-Werte werden als feste und bekannte Konstanten angenommen.
2. Der Erwartungswert von $Y$ lässt sich als deterministische Funktion in $x$ schreiben.

## 1.1 Multiple Lineare Regression

Es wird ein lineares Model mit $(p-1)$ erklärenden Variablen betrachtet:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

mit Responsevektor $\mathbf{y} = (y_1, \ldots, y_n)^\top$, Parametervektor $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{p-1})^\top$ und Fehlervektor $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\top$. Die $(n \times p)$-dimensionale Designmatrix des Modells besitzt die Form

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1,p-1} \\ 1 & x_{21} & x_{22} & \cdots & x_{2,p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{n,p-1} \end{bmatrix}.$$

Für den Fehlerterm $\boldsymbol{\varepsilon}$ wird folgende $n$-dimensionale Normalverteilung angenommen:

$$\boldsymbol{\varepsilon} \sim \mathcal{N}_n\left(\mathbf{0}, \sigma^2 \mathbb{I}_n\right).$$

Dadurch ergibt sich für den Responsevektor

$$\mathbf{y} \sim \mathcal{N}_n\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbb{I}_n\right).$$

Man ist nun an einem Schätzer des Parametervektors $\boldsymbol{\beta}$ interessiert. Dafür wird der *Maximum-Likelihood Schätzer* betrachtet. Dazu muss die Likelihood-Funktion

maximiert werden.

*Definition* 1.1. Sei Y eine Zufallsvariable mit zugehöriger Dichte- oder Wahrscheinlichkeitsfunktion $f(y|\,\boldsymbol{\theta})$, wobei $\boldsymbol{\theta}$ ein ein- oder mehrdimensionaler unbekannter Parameter ist. Seien $y_1, \ldots, y_n$ Realisierungen dieser Zufallsvariablen. Dann ist die *Likelihood-Funktion* definiert als

$$\mathrm{L}\left(\boldsymbol{\theta} \mid y_1, \ldots, y_n\right) = f\left(y_1, \ldots, y_n|\,\boldsymbol{\theta}\right).$$

Zur Berechnung des Maximum-Likelihood (ML) Schätzers muss die Likelihood-Funktion zweimal bezüglich $\boldsymbol{\theta}$ abgeleitet werden. Daher ist es oft einfacher, statt der Likelihood-Funktion die Log-Likelihood-Funktion zu betrachten, welche dieselben optimalen Parameterwerte aufweist.

Unter Annahme der Normalverteilung entspricht der ML Schätzer dem *Least-Squares-Schätzer*, der durch Minimierung der *Sum of Squared Errors* (SSE) bestimmt wird:

$$\begin{aligned}\mathrm{SSE}(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) &= \mathbf{y}^\top\mathbf{y} - \mathbf{y}^\top\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}^\top\boldsymbol{X}^\top\mathbf{y} + \boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} \\ &= \mathbf{y}^\top\mathbf{y} - 2\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{y} + \boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}.\end{aligned}$$

Hier wurde die Tatsache verwendet, dass alle Terme skalarwertig sind. Ableiten und Nullsetzen ergibt den Maximum-Likelihood Schätzer

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{y}.$$

Da der Schätzer $\hat{\boldsymbol{\beta}}$ eine Linearkombination der normalverteilten Response $\mathbf{y}$ ist, ist dieser selbst wieder normalverteilt, mit

$$\begin{aligned}\mathbb{E}\left(\hat{\boldsymbol{\beta}}\right) &= \boldsymbol{\beta} \\ \mathrm{Var}\left(\hat{\boldsymbol{\beta}}\right) &= \sigma^2\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}.\end{aligned}$$

Für den Vektor der *Fitted Values* gilt

$$\hat{\boldsymbol{\mu}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{H}\mathbf{y},$$

wobei $\mathbf{H}$ die symmetrische $(n \times n)$ Hat-Matrix ist:

$$\mathbf{H} = \mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top.$$

Für die Responsevarianz $\sigma^2$ erhält man als ML Schätzer

$$\hat{\sigma}^2 = \frac{1}{n}\mathrm{SSE}\left(\hat{\boldsymbol{\beta}}\right).$$

Da dieser Schätzer jedoch verzerrt ist, wird stattdessen folgender erwartungstreuer Schätzer verwendet:

$$\hat{\sigma}^2 = \frac{1}{n-p}\mathrm{SSE}\left(\hat{\boldsymbol{\beta}}\right).$$

## Beispiel

Als Beispieldatensatz wird nun der in `R` inkludierte Datensatz `trees` betrachtet. Darin sind Daten über 31 Kirschbäume enthalten. Die Prädiktoren sind das verwendbare *Holzvolumen* `Volume` in feet$^3$, die *Baumhöhe* `Height` in feet und der *Umfang* `Girth` in inches auf einer Höhe von 4.5 feet über dem Boden (1 inch = 2.54cm, 1 foot = 30.48cm). Es soll das mittlere verwendbare Holzvolumen aus den beobachteten Größen `Height` und `Girth` vorhergesagt werden.

Einlesen, Fitten und Plotten des Fits

```
1 data(trees)
2 attach(trees)
3 mod <- lm(Volume~Girth + Height)
4 summary(mod)
5 plot(fitted(mod), Volume)
```

Der Befehl `summary()` ergibt folgendes Listing:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -57.9877     8.6382  -6.713 2.75e-07 ***
Girth         4.7082     0.2643  17.816  < 2e-16 ***
Height        0.3393     0.1302   2.607   0.0145 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.882 on 28 degrees of freedom
Multiple R-squared: 0.948,      Adjusted R-squared: 0.9442
F-statistic:   255 on 2 and 28 DF,  p-value: < 2.2e-16
```

Die durch das Modell gefittete Regressionsgerade hat also die Form

$$\hat{\mathbb{E}}\left(\mathrm{Volume}\right) = -57.99 + 4.71 \cdot \mathrm{Girth} + 0.34 \cdot \mathrm{Height}.$$

Der p-Wert des Prädiktors `Height` von 0.01 deutet darauf hin, dass `Height` nicht gerade höchstsignifikant für das Modell ist. Für ein Signifikanzniveau von $\alpha = 0.01$ würde der

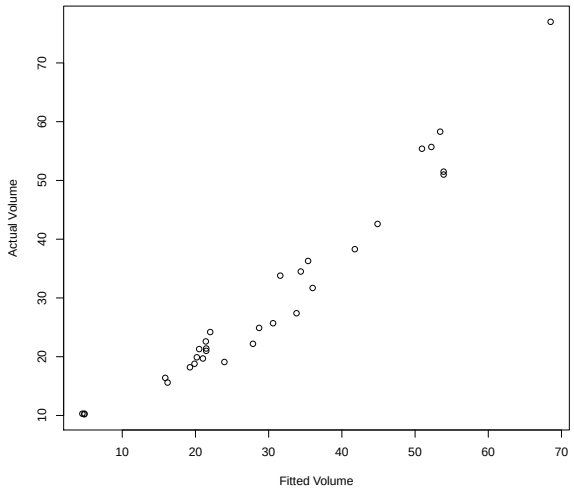


Abbildung 1.1: Scatterplot des beobachteten Volumens gegenüber dem gefitteten Volumen durch die Prädiktoren `Girth` und `Height`.

Prädiktor verworfen, daher wäre er nicht signifikant von null verschieden.

Abbildung 1.1 zeigt den Fit des betrachteten Modells mittels Scatterplot des beobachteten Volumen gegenüber dem durch das Modell gefitteten mittleren Volumen.

Nun wird der Prädiktor `Height` entfernt und das Modell neuerdings gefittet.

Einlesen, Fitten und Plotten des Fits

```
1 mod.2 <-lm(Volume~Girth)
2 summary(mod.2)
3 plot(Girth, Volume)
4 lines(Girth, fitted(mod.2), col="red")
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -36.9435     3.3651  -10.98 7.62e-12 ***
Girth         5.0659     0.2474   20.48  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.252 on 29 degrees of freedom
Multiple R-squared: 0.9353,     Adjusted R-squared: 0.9331
F-statistic: 419.4 on 1 and 29 DF,  p-value: < 2.2e-16
```

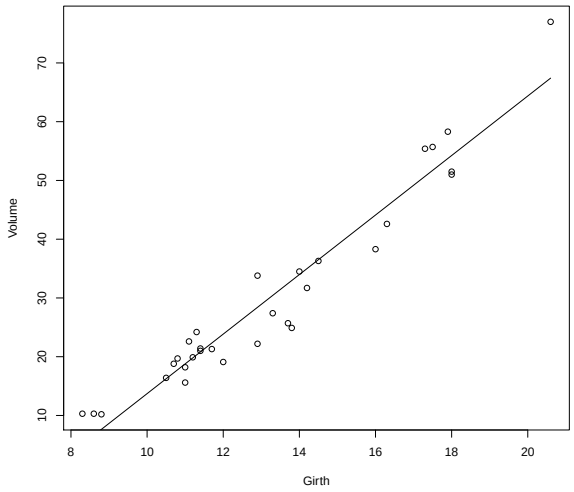


Abbildung 1.2: Scatterplot des beobachteten Volumens gegenüber `Girth` und Plot der gefitteten Regressionsgerade.

Die durch das neue Modell berechnete Regressionsgerade hat nun die Form

$$\hat{\mathbb{E}}\,(\texttt{Volume}) = -36.96 + 5.07 \cdot \texttt{Alter}.$$

Abbildung 1.2 zeigt den Plot der Daten und die durch das Modell gefittete Regressionsgerade.

Klassische lineare Regression beruht auf der Annahme, dass die Zielgrößen normalverteilt sind und eine konstante Varianz aufweisen. Falls die Daten diese Annahme jedoch nicht unterstützen, besteht die Möglichkeit einer Transformation der Daten, beziehungsweise allgemein der Übergang zu *Generalisierten Linearen Modellen*.

## 1.2 Generalisierte Lineare Modelle

Generalisierte Lineare Modelle (GLM) entsprechen einer Verallgemeinerung des klassischen, linearen Regressionsmodells. Im Gegensatz zu linearen Modellen, bei denen angenommen wird, dass die Response normalverteilt ist und eine konstante Varianz $\sigma^2$ besitzt, kann für die Zielvariable im Generalisierten Linearen Modell jede Verteilung aus der Klasse der einparametrigen linearen Exponentialfamilie angenommen werden. Mitglieder dieser Exponentialfamilie sind die Normal-, Poisson-, Gamma- und Binomialverteilung, um nur einige Beispiele zu nennen (Aitkin et al., 2009).

Falls die zugrunde liegenden Daten keine Normalverteilung beziehungsweise keine konstante Varianz vorweisen, können die Methoden der klassischen Regressionsanalyse nicht angewendet werden. Eine Möglichkeit, dieses Problem zu lösen, wäre die Verwendung einer Transformation, um die geforderten Bedingungen approximativ zu erfüllen. Die *Box-Cox Transformation* (Box & Cox, 1964) ist ein Beispiel, um eine Stabilisierung der Varianz zu erzielen.

Für eine positive Response $Y > 0$ sei

$$Y(\lambda) = \begin{cases} \frac{Y^\lambda - 1}{\lambda}, & \text{falls } \lambda \neq 0, \\ \log(Y), & \text{falls } \lambda = 0. \end{cases}$$

Strebt der *Transformationsparameter* $\lambda$ gegen null, lässt sich zeigen, dass $Y(\lambda)$ eine stetige Funktion in $\lambda$ ist:

$$\begin{aligned} \lim_{\lambda \to 0} \frac{Y^\lambda - 1}{\lambda} &= \lim_{\lambda \to 0} \frac{\exp(\lambda \log(Y)) - 1}{\lambda} \overset{L'Hospital}{=} \lim_{\lambda \to 0} \frac{\exp(\lambda \log(Y)) \log(Y)}{1} \\ &= \log(Y). \end{aligned}$$

Für die Daten $(Y_i, x_i)$ wird nun angenommen, dass ein $\lambda$ existiert, sodass $Y(\lambda)$ einer Normalverteilung mit Erwartungswert $\mathbf{x}^\top \boldsymbol{\beta}$ und Varianz $\sigma^2$ folgt. Um diesen Parameter $\lambda$ zu schätzen, wird die Maximum-Likelihood Methode angewendet. Dabei werden für jedes $\lambda$ die ML Schätzer für $\boldsymbol{\beta}$ und $\sigma^2$ bestimmt und in der Likelihood-Funktion von $\lambda$ angewendet, wodurch man es eigentlich mit einer *profile Likelihood-Funktion* zu tun hat. Es werden likelihoodbasierte Konfidenzintervalle für $\lambda$ bestimmt.

Im GLM werden, anstatt der Anwendung einer Transformation, die Annahmen des Modells verallgemeinert. Die Responsevariable $Y$ wird nun nicht auf die Normalverteilung eingeschränkt, sondern auf die Verteilungen der *einparametrigen, linearen Exponentialfamilie* erweitert.

*Definition* 1.2. Eine Zufallsvariable $Y$ sei aus einer Verteilung mit Dichte- oder Wahrscheinlichkeitsfunktion

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

für spezielle, bekannte Funktionen $b(\cdot)$ und $c(\cdot)$, mit $\phi > 0$ (*Dispersion*). Kann $\phi$ als feste Größe betrachtet werden, so bezeichnet man $f(y, \theta)$ als *einparametrige, lineare Exponentialfamilie* mit kanonischem Parameter $\theta$.

Durch die Erweiterung auf die einparametrige Exponentialfamilie wird die Varianz als Funktion des Erwartungswertes modelliert. Um dies zu sehen, müssen zuerst einige Eigenschaften der Score-Funktion gezeigt werden, die auch innerhalb der Exponential-

familie anwendbar sind.

*Definition* 1.3 (Score-Funktion). Die *Score-Funktion* ist definiert als Ableitung der Log-Likelihood Funktion:

$$\frac{\partial \log f\left(y,\theta\right)}{\partial \theta}.$$

*Satz* 1.1. Für die Score-Funktion gilt:

$$\begin{aligned}\mathbb{E}\left[\frac{\partial \log f(Y,\theta)}{\partial \theta}\right] &= 0,\\ \mathbb{E}\left[\left(\frac{\partial \log f(Y,\theta)}{\partial \theta}\right)^2\right] &= \mathbb{E}\left[-\frac{\partial^2 \log f(Y,\theta)}{\partial \theta^2}\right].\end{aligned}$$

*Beweis.* Für den Beweis werden folgende Eigenschaften verwendet:

$$\begin{aligned}\frac{\partial \log f(y,\theta)}{\partial \theta} &= \frac{1}{f(y,\theta)}\frac{\partial f(y,\theta)}{\partial \theta},\\ \int_{\mathbb{R}} f(y,\theta)\mathrm{d}y &= 1.\end{aligned}$$

Da sich innerhalb der Exponentialfamilie die Reihenfolge von Integration und Differentiation vertauschen lässt, folgt

$$\begin{aligned}\mathbb{E}\left[\frac{\partial \log f(Y,\theta)}{\partial \theta}\right] = \mathbb{E}\left[\frac{\partial f(Y,\theta)}{\partial \theta}\frac{1}{f(Y,\theta)}\right] &= \int_{\mathbb{R}} \frac{\partial f(y,\theta)}{\partial \theta}\frac{1}{f(y,\theta)} f(y,\theta)\mathrm{d}y\\ &= \int_{\mathbb{R}} \frac{\partial f(y,\theta)}{\partial \theta}\mathrm{d}y\\ &= \frac{\partial}{\partial \theta}\underbrace{\int_{\mathbb{R}} f(y,\theta)\mathrm{d}y}_{=1} = 0.\end{aligned}$$

Für die zweite Eigenschaft wird zunächst folgende Nebenrechnung durchgeführt:

$$\begin{aligned}\frac{\partial^2 \log f(y,\theta)}{\partial \theta^2} &= \frac{\partial}{\partial \theta}\left(\frac{\partial}{\partial \theta} f(y,\theta)\frac{1}{f(y,\theta)}\right)\\ &= \frac{\partial^2}{\partial \theta^2} f(y,\theta)\frac{1}{f(y,\theta)} + \frac{\partial}{\partial \theta} f(y,\theta)\frac{-1}{f^2(y,\theta)}\frac{\partial}{\partial \theta} f(y,\theta).\end{aligned}$$

Damit folgt

$$\begin{aligned}\mathbb{E}\left[-\frac{\partial^2 \log f(Y,\theta)}{\partial\theta^2}\right] &= \mathbb{E}\left[-\frac{\partial^2}{\partial\theta^2} f(Y,\theta)\frac{1}{f(Y,\theta)} + \frac{\partial}{\partial\theta} f(Y,\theta)\frac{\partial}{\partial\theta} f(Y,\theta)\frac{1}{f^2(Y,\theta)}\right] \\ &= -\int_{\mathbb{R}} \frac{\partial^2}{\partial\theta^2} f(y,\theta)\frac{1}{f(y,\theta)} f(y,\theta)\mathrm{d}y \\ &+ \int_{\mathbb{R}} \frac{\partial}{\partial\theta}\log f(y,\theta)\frac{\partial}{\partial\theta}\log f(y,\theta) f(y,\theta)\mathrm{d}y \\ &= \underbrace{-\frac{\partial^2}{\partial\theta^2}\underbrace{\int_{\mathbb{R}} f(y,\theta)\mathrm{d}y}_{=1}}_{=0} + \mathbb{E}\left[\left(\frac{\partial \log f(Y,\theta)}{\partial\theta}\right)^2\right].\end{aligned}$$

□

Mit Hilfe dieses Satzes lassen sich für die Exponentialfamilie folgende Eigenschaften zeigen:

$$\begin{aligned}\mathbb{E}\left[\frac{\partial \log(f(Y,\theta))}{\partial\theta}\right] &= \mathbb{E}\left[\frac{\partial}{\partial\theta}\left(\frac{Y\cdot\theta - b(\theta)}{\phi} + c(Y,\phi)\right)\right] \\ &= \frac{1}{\phi}\mathbb{E}\left[Y - b'(\theta)\right] = 0,\end{aligned}$$

womit folgt, dass

$$\mathbb{E}\left[Y\right] = \mu := b'(\theta). \tag{1.1}$$

Nun wird die Varianz der Response betrachtet. Mit Satz 1.1 folgt

$$\begin{aligned}&\mathbb{E}\left[\left(\frac{\partial \log f(Y,\theta)}{\partial\theta}\right)^2\right] + \mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\log f(Y,\theta)\right] \\ =&\frac{1}{\phi^2}\mathrm{Var}(Y) - \frac{b''(\theta)}{\phi} = 0.\end{aligned}$$

Damit folgt

$$\mathrm{Var}(Y) = \phi\cdot b''(\theta) =: \phi\cdot \mathrm{V}(\mu). \tag{1.2}$$

Die Funktion $\mathrm{V}(\mu)$ wird als *Varianzfunktion*, $\phi$ als *Dispersionsparameter* bezeichnet. Die Funktion $b(\theta)$ wird *Kumulantenfunktion* genannt.

Der Zufallsvektor $Y = (Y_1,\ldots,Y_n)^\top$ besitzt eine Verteilung aus der Exponentialfamilie, mit $\mathbb{E}(Y_i) = \mu_i$ und $\mathrm{Var}\left(Y_i\right) = \phi\cdot V(\mu_i)$. Der Vektor $\mathbf{x}_i = (x_{i0},\ldots,x_{ip-1})^\top$ bezeichnet

den Vektor der erklärenden Variablen, die zu einer Designmatrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\top$ zusammengefasst werden, $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{p-1})^\top$ ist der Vektor der unbekannten Parameter, $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)^\top$ beschreibt den Vektor der linearen Prädiktoren.

Unter Annahme der Existenz von $\mathbb{E}(Y)$ und $\mathrm{Var}(Y)$ wird bei den Generalisierten Linearen Modellen eine Parametrisierung der Form

$$\begin{aligned} Y_i &\overset{ind}{\sim} \text{Exponentialfamilie}(\theta_i), \quad \mathbb{E}(Y_i) = \mu_i = \mu(\theta_i) \\ \eta_i &= \mathbf{x}_i^\top \boldsymbol{\beta} \\ g(\mu_i) &= \eta_i \end{aligned}$$

betrachtet, wobei $g(\cdot)$ die *Linkfunktion* beschreibt.

Die Linkfunktion zeigt eine wesentliche Änderung im GLM gegenüber dem klassischen linearen Modell. Während im linearen Modell der Erwartungswert direkt durch eine Linearkombination der erklärenden Variablen modelliert wird, modelliert das GLM eine Funktion des Erwartungswerts der Response. Für den Parametervektor $\boldsymbol{\beta}$ erhält man als Maximum-Likelihood Schätzung durch Anwendung der Newton-Raphson Methode im $(t+1)$-ten Iterationsschritt

$$\boldsymbol{\beta}^{(t+1)} = \left(\mathbf{X}^\top \mathbf{W}^{(t)} \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{W}^{(t)} \mathbf{z}^{(t)},$$

mit Diagonalmatrix $\mathbf{W} = \mathrm{diag}(w_i)$, $1/w_i = \phi V(\mu_i) \left(g'(\mu_i)\right)^2$ und *Pseudobeobachtungen* $\mathbf{z} = \mathbf{X}\boldsymbol{\beta} + \mathbf{D}(\mathbf{y} - \boldsymbol{\mu})$, wobei $\mathbf{D} = \mathrm{diag}(d_i)$, $d_i = g'(\mu_i)$. Weiters lässt sich zeigen, dass bei Konvergenz $(t \to \infty)$ für den resultierenden ML Schätzer $\hat{\boldsymbol{\beta}}$ gilt:

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \overset{n\to\infty}{\sim} \mathcal{N}_p\left(\mathbf{0}, n\left(\mathbf{X}^\top \mathbf{W} \mathbf{X}\right)^{-1}\right).$$

Um die Güte der Modellanpassung zu beschreiben, werden in den Generalisierten Linearen Modellen die Log-Likelihood Funktionen betrachtet. Es wird dabei das volle, *saturierte* Modell mit $n$ Parametern mit dem aktuellen Modell verglichen.

Beim saturierten Modell entspricht die Anzahl der Parameter der Anzahl der Beobachtungen. Dadurch wird ein ML Schätzer ohne Einschränkungen bestimmt. Es gilt $\hat{\mu}_i = y_i$, $i = 1, \ldots, n$. Das saturierte Modell liefert also eine bestmögliche Anpassung an die Daten. Es ist natürlich nicht sinnvoll, da es lediglich die Datenpunkte wiederholt, jedoch hilfreich, um die Güte von kleineren Modellen zu beurteilen. Dazu wird der Begriff der *Deviance* eingeführt.

*Definition* 1.4 (Skalierte Deviance). Die *skalierte Deviance* ist definiert als

$$\frac{1}{\phi} D(\mathbf{y}, \hat{\boldsymbol{\mu}}) = -2\left(\log f(\mathbf{y}, \hat{\boldsymbol{\mu}}) - \log f(\mathbf{y}, \mathbf{y})\right).$$

Dies entspricht dem $(-2)$-fachen der Differenz der Log-Likelihood-Funktion des betrachteten Modells und jener des saturierten Modells.

Unter Annahme der Normalverteilung gilt, dass das Maß der Deviance der Fehlerquadratsumme im linearen Modell entspricht:

*Beispiel* 1.1. Sei $Y_i \overset{ind}{\sim} \mathcal{N}(\mu_i, \sigma^2)$, mit $\sigma^2$ konstant. Der ML Schätzer $\hat{\boldsymbol{\mu}}$, der $\log f(\mathbf{y}, \boldsymbol{\mu})$ maximiert, minimiert auch die Deviance, da $\log f(\mathbf{y}, \mathbf{y})$ unabhängig von $\boldsymbol{\mu}$ ist. Es gilt:

$$\begin{aligned}
\log f(\mathbf{y}, \boldsymbol{\mu}) &\overset{ind}{=} \log \prod_{i=1}^{n} f(y_i, \mu_i) = \log \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{1}{2}\left(\frac{y_i-\mu_i}{\sigma}\right)^2} \\
&= \sum_{i=1}^{n} \log \left( \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{y_i-\mu_i}{\sigma}\right)^2} \right) \\
&= \sum_{i=1}^{n} \left\{ \log \left( \frac{1}{\sqrt{2\pi\sigma^2}} \right) + \log \left( e^{-\frac{1}{2}\left(\frac{y_i-\mu_i}{\sigma}\right)^2} \right) \right\} \\
&= \sum_{i=1}^{n} \left\{ -\frac{1}{2} \log \left(2\pi\sigma^2\right) - \frac{1}{2} \frac{(y_i-\mu_i)^2}{\sigma^2} \right\} \\
&= -\frac{n}{2} \log \left(2\pi\sigma^2\right) - \frac{1}{2} \sum_{i=1}^{n} \frac{(y_i-\mu_i)^2}{\sigma^2}.
\end{aligned}$$

Für das saturierte Modell ergibt sich folgende Log-Likelihood Funktion:

$$\log f(\mathbf{y}, \mathbf{y}) = \log \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \underbrace{e^{-\frac{1}{2}\left(\frac{y_i-y_i}{\sigma^2}\right)^2}}_{=1} = -\frac{n}{2} \log \left(2\pi\sigma^2\right).$$

Dadurch folgt die skalierte Deviance ($\phi = \sigma^2$)

$$\begin{aligned}
\frac{1}{\sigma^2} D(\mathbf{y}, \hat{\boldsymbol{\mu}}) &= -2 \left( \log f(\mathbf{y}, \hat{\boldsymbol{\mu}}) - \log f(\mathbf{y}, \mathbf{y}) \right) \\
&= \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\sigma^2} = \frac{1}{\sigma^2} \mathrm{SSE}\left(\hat{\boldsymbol{\beta}}\right) \sim \chi^2_{n-p}.
\end{aligned}$$

## Beispiel

Als Beispieldatensatz wird nun wieder der Datensatz `trees` der Kirschbäume verwendet.

Unter der Annahme der Normalverteilung (konstante Varianz) für `Volume` wird das Modell mit der Linkfunktion $g(\mu) = \mu^{1/3} = \eta$ angenommen, wobei der lineare Prädiktor durch $\eta = \beta_0 + \beta_1 \cdot \text{Girth} + \beta_2 \cdot \text{Height}$ spezifiziert wird.

Fitten des Modells mittels GLM

```
glm(Volume ~ Girth + Height, family = quasi(link=power(1/3),
             variance="constant"), trace=T, epsilon=1e-8, maxit=15)
```

```
Deviance = 185.3623 Iterations - 1
Deviance = 184.1578 Iterations - 2
Deviance = 184.1577 Iterations - 3
Deviance = 184.1577 Iterations - 4

Coefficients:
(Intercept)        Girth       Height
     2.0965       1.8491       0.3429

Degrees of Freedom: 30 Total (i.e. Null);  28 Residual
Null Deviance:      8106
Residual Deviance: 184.2        AIC: NA
```

Mittels der Funktion `glm()` werden Generalisierte Lineare Modelle in `R` geschätzt. Die `R`-Ausgabe `Coefficients` gibt die durch die Iteration gefundenen ML Schätzer wieder. Die `Null Deviance` gibt den Wert der Deviance an, wenn das saturierte Modell mit einem Intercept-only-Modell (das iid-Modell) verglichen wird. Der hohe Wert der Deviance deutet auf eine schlechte Modellanpassung hin. Die `Residual Deviance` gibt an, um wieviel besser der Deviance-Wert wird, wenn man die Terme des aktuellen Modells zum iid-Modell hinzufügt.

Das resultierende Modell hat die Form

$$\hat{\mathbb{E}}\,(\text{Volume})^{1/3} = 2.10 + 1.85 \cdot \text{Girth} + 0.34 \cdot \text{Height}.$$

Bei den bisher betrachteten Modellen handelt es sich durchwegs um (verallgemeinert) lineare Modelle, d.h. die Form der Erwartungswertfunktion ist a priori linear festgelegt. Im Folgenden werden Modelle betrachtet, die die Klasse der GLM verallgemeinern, indem ein additiver Term bestehend aus *glatten* Funktionen in den Prädiktorvariablen hinzugefügt wird.

# 2 Generalisierte Additive Modelle

*Definition* 2.1. Ein *Generalisiertes Additives Modell* (GAM) ist ein GLM mit linearem Prädiktor, addiert mit einer Summe von Glättungsfunktionen in den Prädiktorvariablen. Allgemein besitzt ein GAM folgende Struktur (Wood, 2006):

$$g(\mu_i) = \mathbf{x}_i^{*\top}\boldsymbol{\theta} + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}, x_{4i}) + \ldots,$$

wobei

$$Y_i \sim \text{Exponentialfamilie}(\theta_i), \quad \mu_i := \mathbb{E}(Y_i).$$

$Y_i$ bezeichnet die Responsevariable, $\mathbf{x}_i^{*\top}$ ist die $i$-te Zeile der Modellmatrix der strikt-parametrischen Modellkomponenten $\boldsymbol{\theta}$, $f_j$ seien die Glättungsfunktionen in den Prädiktorvariablen $x_k$.

Die Anwendung von GAMs ermöglicht eine sehr flexible Modellierung der Abhängigkeit der Responsevariablen von den erklärenden Variablen, es treten jedoch zwei neue Fragen auf: Die Glättungsfunktionen müssen bestimmt werden und es muss spezifiziert werden, wie glatt sie sein sollen.

Im Folgenden wird aus Gründen der Übersicht der strikt-parametrische Teil vernachlässigt, da Schätzungen für diesen Teil bereits ausführlich zuvor behandelt wurden. Weiters beschränkt man sich zu Beginn auf den univariaten Fall, das heißt, man betrachtet eine Glättungsfunktion $f$ in einem Prädiktor $x$. Von Interesse ist es nun, eine Schätzung des *nichtparametrischen* Teil des GAMs zu finden. Das Modell reduziert sich also im ersten Schritt auf ein univariates, nichtparametrisches Regressionsmodell der Form:

$$Y_i = f(x_i) + \varepsilon_i, \quad i = 1, \ldots, n.$$

Als Link-Funktion wurde hier die identische Abbildung verwendet. Später in Kapitel 2.5 wird dann ein erstes Anwendungsbeispiel eines vollständigen GAMs vorgestellt.

Es stellt sich nun die Frage nach einer vernünftigen Darstellung der Glättungsfunktion $f(x)$. Sei $b_j(x)$ die $j$-te *Basisfunktion*, dann gilt

$$f(x) = \sum_{j=1}^{q+1} b_j(x)\beta_j, \tag{2.1}$$

wobei $\beta_j$ die zu schätzenden Parameter sind. Durch diese Darstellung entsteht eine

Glättungsfunktion vom Grad $q$.

Für das Beispiel einer polynomialen Basis dritter Ordnung gilt $q + 1 = 4$ und $b_1(x) = 1$, $b_2(x) = x$, $b_3(x) = x^2$ und $b_4(x) = x^3$.

Häufig reicht polynomiale Regression nicht aus, um die Daten zu beschreiben. Ein Polynom niedrigeren Grades lässt oft nicht genügend Flexibilität zu, während Polynome höheren Grades zwar ausreichend Flexibilität bieten, doch wird die Schätzung oft sehr „rau“ und Overfitting ist ein häufiges Problem. Vor allem an den Enden der gefitteten Funktion tritt das Problem des Oszillierens auf, falls ein hoher Polynomgrad verwendet wird.

Folgendes Beispiel (Wood, 2006) zeigt die Probleme einer polynomialen Basis, wenn eine Funktion $f$ über einen gesamten Bereich geschätzt werden soll, hier das Intervall $[0, 1]$. Zu diesem Zweck werden transformierte, gleichverteilte Zufallsvariablen modelliert.

Simulieren, Fitten und Plotten der Daten

```
1 set.seed(1)
2 x <- sort(runif(40)*10)^.5
3 y <- sort(runif(40))^0.1
4 xx <- seq(min(x), max(x), length=200)
5 plot(x, y, cex.lab=1.5)
6 b <- lm(y~poly(x, 5))
7 lines(xx,predict(b, data.frame(x=xx)))
8 b <- lm(y~poly(x, 10))
9 lines(xx, predict(b, data.frame(x=xx)), col=2)
```

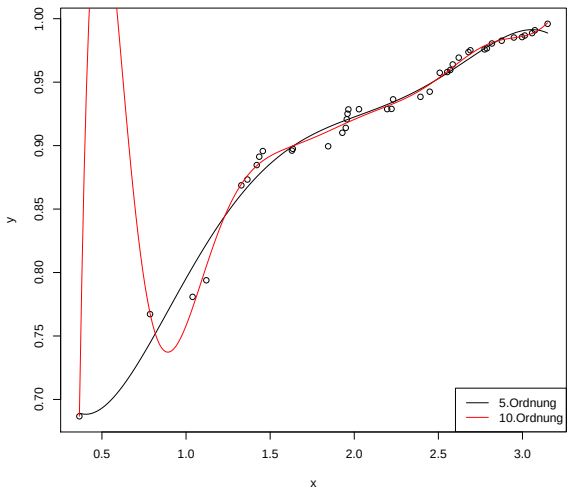


Abbildung 2.1: Probleme bei Verwendung polynomieller Basen.

In Abbildung 2.1 erkennt man die beschriebenen Probleme. Der Grund dafür liegt bei der Taylorapproximation, die nur im Konvergenzbereich, d.h. hinreichend nahe am Entwicklungspunkt, konvergiert.

Eine bessere Alternative, um die Funktion $f$ zu modellieren, sind die sogenannten *Polynom-Splines* (Fahrmeir et al., 2013).

## 2.1 Polynom-Splines

*Definition* 2.2 (Polynom-Splines). Eine Funktion $f : [a, b] \to \mathbb{R}$ heißt Polynom-Spline vom Grad $q \geq 0$ zu den Knoten $a = x_1^* < \cdots < x_k^* = b$, falls sie die folgenden Bedingungen erfüllt:

1. $f(z)$ ist $(q-1)$-mal stetig differenzierbar.
2. $f(z)$ ist auf den durch die Knoten gebildeten Intervallen $[x_j^*, x_{j+1}^*)$ ein Polynom vom Grad $q$.

Es werden also stückweise Polynome an den Knotenpunkten zusammengesetzt. An diesen Knoten müssen die angeführten *Glattheitsbedingungen* erfüllt werden.

Durch eine Erhöhung des Splinegrads $q$ erhält man eine glattere Darstellung des Polynom-Splines. Entgegengesetzt dazu führt eine Erhöhung der Knotenanzahl zu einer raueren Darstellung, da eine größere Anzahl an stückweisen Polynomen verwendet wird, jedoch auch zu einer höheren Flexibilität der Schätzung. Üblicherweise wird kein höherer Splinegrad als kubisch verwendet.

Im Folgenden soll erläutert werden, warum die Verwendung von Splines eine sinnvolle Methode ist, um Funktionen zu approximieren. Im Speziellen werden für den kubischen Spline einige angeneheme Eigenschaften gezeigt.

Betrachte eine Menge von Punkten $\{(x_i, y_i) : i = 1, \ldots, n\}$ mit $x_i < x_{i+1}$. Der *natürliche kubische Spline* $g(x)$ ist eine Funktion, die auf den Intervallen $[x_i, x_{i+1}]$ aus Teilen kubischer Polynome besteht, die an den Enden der Intervalle so verbunden werden, dass der Spline stetig bis zur zweiten Ableitung ist. Für solch natürlichen kubischen Spline gilt $g(x_i) = y_i$ und $g''(x_1) = g''(x_n) = 0$.

Der folgende Satz zeigt, dass natürliche kubische Splines die glattesten Interpolierenden sind:

*Satz* 2.1. Unter allen Funktionen, die stetig auf $[x_1, x_n]$ sind, absolut stetige erste Ableitung haben und $\{x_i, y_i\}$ interpolieren, ist $g(x)$ die glatteste Funktion im Sinne von:

$$\min \quad \mathcal{J}(f) = \int_{x_1}^{x_n} f''(x)^2 \mathrm{d}x.$$

*Beweis.* Der Beweis beruht auf Schoenberg (1964).

Sei $f(x)$ eine Interpolierende von $\{x_i, y_i\}$ (ungleich $g(x)$). Sei $h(x) = f(x) - g(x)$. Umformen ergibt

$$f(x) = h(x) + g(x) \Rightarrow f''(x) = h''(x) + g''(x).$$

Damit folgt

$$\begin{aligned}\int_{x_1}^{x_n} f''(x)^2 \mathrm{d}x &= \int_{x_1}^{x_n} \left(g''(x) + h''(x)\right)^2 \mathrm{d}x \\ &= \int_{x_1}^{x_n} g''(x)^2 \mathrm{d}x + 2\int_{x_1}^{x_n} g''(x)h''(x)\mathrm{d}x + \int_{x_1}^{x_n} h''(x)^2 \mathrm{d}x.\end{aligned}$$

Für das gemischte Integral gilt

$$\begin{aligned}\int_{x_1}^{x_n} g''(x)h''(x)\mathrm{d}x &= h'(x)g''(x)\Big|_{x_1}^{x_n} - \int_{x_1}^{x_n} g'''(x)h'(x)\mathrm{d}x \\ &= -h'(x_n)\underbrace{g''(x_n)}_{=0} - h'(x_1)\underbrace{g''(x_1)}_{=0} - \int_{x_1}^{x_n} g'''(x)h'(x)\mathrm{d}x \\ &= -\int_{x_1}^{x_n} g'''h'(x)\mathrm{d}x.\end{aligned}$$

Die Funktion $g(x)$ besteht stückweise aus kubischen Polynomen, deshalb ist $g'''(x)$ konstant auf $(x_i, x_{i+1})$ $\forall i$:

$$\begin{aligned}-\int_{x_1}^{x_n} g'''(x)h'(x)\mathrm{d}x &= -\sum_{i=1}^{n-1}\int_{x_i}^{x_{i+1}} g'''(x)h'(x)\mathrm{d}x \\ &\overset{konst.}{=} -\sum_{i=1}^{n-1} g'''\left(x_i^+\right)\int_{x_i}^{x_{i+1}} h'(x)\mathrm{d}x \\ &= -\sum_{i=1}^{n-1} g\left(x_i^+\right)\left(\underbrace{h(x_{i+1})}_{=0} - \underbrace{h(x_i)}_{=0}\right) = 0.\end{aligned}$$

Der Punkt $x_i^+$ repräsentiert das Argument der Funktion $g'''$ auf dem Intervall $(x_i, x_{i+1})$. Die letzte Gleichung ergibt sich aus der Tatsache, dass $h(x_i) = f(x_i) - g(x_i) = 0$ $\forall i$, da

$f(x)$ und $g(x)$ beide Interpolierende und somit äquivalent in $x_i$ sind.

Somit folgt

$$\int_{x_1}^{x_n} f''(x)^2 \mathrm{d}x = \int_{x_1}^{x_n} g''(x)^2 \mathrm{d}x + \underbrace{\int_{x_1}^{x_n} h''(x)^2 \mathrm{d}x}_{\geq 0} \geq \int_{x_1}^{x_n} g''(x)^2 \mathrm{d}x.$$

Die Gleichheit gilt genau dann, wenn $h''(x) = 0$ für $x_1 < x < x_n$. Wegen $h(x_1) = h(x_n) = 0$ ist dies genau dann der Fall, wenn $h(x) = 0$ auf $[x_1, x_n]$, denn:

$$\begin{aligned} h''(x) &= 0, \quad x_1 < x < x_n \\ \Rightarrow h'(x) &= c, \quad x_1 < x < x_n \\ \Rightarrow h(x) &= c \cdot x + c_1, \quad x_1 < x < x_n \end{aligned}$$

Es folgt für die Funktion $h$:

$$\begin{aligned} h(x_1) &= c \cdot x_1 + c_1 \stackrel{!}{=} 0, \\ h(x_n) &= c \cdot x_n + c_1 \stackrel{!}{=} 0. \end{aligned}$$

Die Differenz der beiden Gleichungen führt zu

$$c \underbrace{(x_1 - x_n)}_{\neq 0} = 0 \Rightarrow c = 0 \Rightarrow c_1 = 0 \Rightarrow h(x) = 0.$$

Das heißt, jede Interpolierende, die nicht identisch mit $g(x)$ ist, hat ein größeres Integral seiner zweiten Ableitung. Somit ist der natürliche kubische Spline die glatteste Interpolierende. □

Die kubischen Splines werden berechnet durch Minimierung bezüglich $g$ von

$$\sum_{i=1}^{n} \{y_i - g(x_i)\}^2 + \lambda \int g''(x)^2 \mathrm{d}x. \tag{2.2}$$

Der Parameter $\lambda \geq 0$ heißt *Glättungsparameter* und kontrolliert den Trade-off zwischen Modell-Fit und Glattheit des Modells.

*Satz* 2.2. Die durch Minimierung von (2.2) erhaltene Funktion ist der kubische Glättungsspline $g(x)$.

*Beweis.* Sei $f^*(x)$ eine weitere Funktion, die (2.2) minimiert. Dann kann $\{x_i, f^*(x_i)\}$ durch den kubischen Glättungsspline $g(x)$ interpoliert werden. Dadurch haben $f^*$ und $g$ dieselbe Quadratsumme in (2.2). Wegen der zuvor gezeigten Eigenschaften von $g$ muss deren Integral der zweiten Ableitung kleiner sein. Also entspricht $g$ einer kleineren Lösung von (2.2) als $f^*$, was zu einem Widerspruch führt. Also ist der kubischer Glättungsspline die einzige Funktion, die (2.2) minimiert. □

Es wurden nun einige Eigenschaften von Splines gezeigt, die deren Anwendung rechtfertigen. Ein Problem ist, dass so viele freie Parameter auftreten, wie es zu glättende Datenpunkte gibt, was zu einem enormen Rechenaufwand führen kann. Daher finden in der Praxis *penalisierte Regressions-Splines* Anwendung, auf die später noch näher eingegangen wird.

Neben der Anzahl muss auch die *Lokalisierung* der Knoten a priori bestimmt werden. Die drei gängisten Varianten für die Wahl der Knoten sind:

- *Äquidistante Knoten*: Die Knoten werden gleichmäßig auf den gesamten Wertebereich verteilt. Als Beispiel wird das Intervall $[a, b]$ in $(k-1)$ Intervalle der Breite $h$ zerlegt, mit
$$h = \frac{b-a}{k-1}.$$
Durch diese Wahl von $h$ ergeben sich die äquidistant verteilten Knotenpunkte
$$x_j^* = a + (j-1) \cdot h, \quad j = 1, \ldots, k.$$
- *Quantilsbasierte Knoten*: Als Knoten werden die $\left(\frac{j-1}{k-1}\right)$-Quantile der beobachteten Werte $x_i$ verwendet. Dadurch werden viele Knoten gerade dort positioniert, wo die Dichte der $x_i$ hoch ist.
- Weiters ist eine Wahl der Knotenpunkte durch *Beobachtung des Streudiagramms* der Daten möglich.

Eine erste Möglichkeit der Anwendung von Polynom-Splines ist die in Wood (2006) vorgestellte Anwendung einer kubischen Spline-Basis. Seien die Knotenpunkte des Splines, $\{x_i^*\colon i = 1, \ldots, q-2\}$, gegeben, wobei $q$ die Dimension der Basis bezeichnet. Eine Möglichkeit, die kubischen Splines zu definieren, wurde in Gu (2002) vorgestellt.

Die Basisfunktionen werden definiert als $b_1(x) = 1$, $b_2(x) = x$, $b_{i+2} = R(x, x_i^*)$ für $i = 1, \ldots, q-2$, wobei

$$R(x,z) = \left[(z-1/2)^2 - 1/12\right]\left[(x-1/2)^2 - 1/12\right]/4 - \left[(|x-z|-1/2)^4 - 1/2(|x-z|-1/2)^2 + 7/240\right]/24. \tag{2.3}$$

Durch Anwendung dieser kubischen Spline-Basis erhält man ein lineares Modell $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, wobei die $i$-te Zeile der Modellmatrix $\mathbf{X}$ folgende Gestalt besitzt:

$$\mathbf{x}_i^\top = \left(1, x_i, R(x_i, x_1^*), R(x_i, x_2^*), \ldots, R(x_i, x_{q-2}^*)\right).$$

Die korrespondierenden Parameter können mittels der Least-Squares-Methode geschätzt werden.

Es soll nun eine Anwendung dieser Basis illustriert werden. Das folgende Beispiel stammt aus Wood (2006). Darin wird der Zusammenhang von Hubraumgröße und Verschleiß des Automotors betrachtet. Dazu liegen Messwerte von 19 Volvo-Motoren vor.

*Beispiel* 2.1 (Anwendungsbeispiel der kubischen Spline-Basis). Zunächst werden die Daten in `R` eingelesen und die Hubraumgrößen auf das Einheitsintervall $[0, 1]$ skaliert, um eine bessere numerische Stabilität zu erreichen.

Plot der Volvo-Daten

```
1 library(gamair)
2 data(engine)
3 attach(engine)
4 x <- size - min(size); x <- x/max(x)
5 x <- sort(x, decreasing = FALSE)
6 plot(x, wear, xlab="Scaled engine size", ylab="Wear index")
```

Es handelt sich hierbei um einen in `R` standardmäßig implementierten Datensatz `engine`, der innerhalb des Packetes `gamair` aufgerufen werden kann.

Nun wird die Funktion `rk()` definiert, um die Basisfunktion der kubischen Splines zu bestimmen.

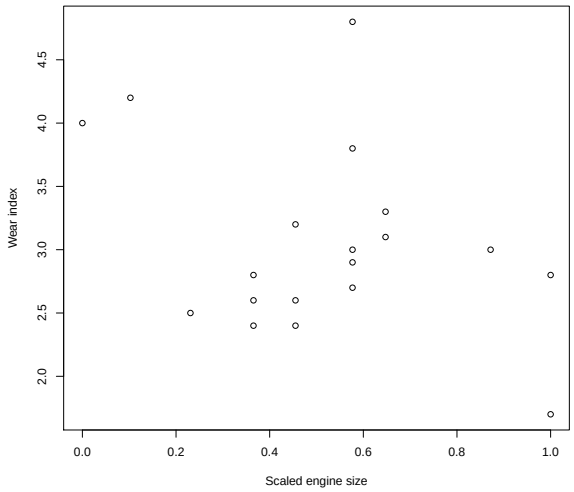


Abbildung 2.2: Scatterplot der Volvo-Datensatzes: Verschleiß des Motors in Abhängigkeit vom Hubraum des Motors.

Implementation der Funktion rk

```
1 rk <- function(x,z)
2 {
3   ((z-0.5)^2-1/12)*((x-0.5)^2-1/12)/4-
4   ((abs(x-z)-0.5)^4-(abs(x-z)-0.5)^2/2+7/240)/24
5 }
```

Um die Modellmatrix $\mathbf{X}$ zu berechnen, wird eine Funktion `spl.X()` geschrieben, die als Input die Vorhersagewerte `x` und die Knotenpunkte `xk` benötigt.

Implementation der Funktion spl.X

```
1 spl.X <- function(x,xk)
2 {
3   q <- length(xk)+2
4   n <- length(x)
5   X <- matrix(1, n, q)
6   X[,2] <- x
7   X[,3:q] <- outer(x, xk, FUN=rk)
8   X
9 }
```

Die Matrix `X` wird zunächst als Einsmatrix initialisiert (Zeile 5). Dann werden die $x$-Werte auf die zweite Spalte der Matrix geschrieben (Zeile 6). Zum Schluß wird die Funktion `rk()` auf den Rest der Matrix angewendet (Zeile 7).

Nun fehlt nur noch eine Knotenmenge $x_i^*$, um das Modell aufzustellen. Im folgenden wird eine Rang 6 Basis verwendet, d.h. $q = 6$, weshalb vier Knotenpunkte benötigt werden, die gleichmäßig auf dem Einheitsintervall $[0, 1]$ verteilt werden. Damit lässt sich ein erstes Modell fitten:

Fit mittels globalen kubischen Spline-Basen

```
1 xk <- 1:4/5
2 X <- spl.X(x,xk)
3 mod.1 <- lm(wear~X-1)
4 xp <- 0:100/100
5 Xp <- spl.X(xp,xk)
6 plot(x, wear, xlab="Scaled engine size", ylab="Wear index")
7 lines(xp, Xp %*% coef(mod.1))
```

In Zeile 2 wird die Modellmatrix mit den neuen Knotenpunkten `xk` berechnet. Dann wird das Modell bestimmt (Zeile 3). Der Vektor `xp` entspricht den Vorhersagewerten, an denen das Modell berechnet werden soll. In Zeile 8 wird das entstandene Modell geplottet.

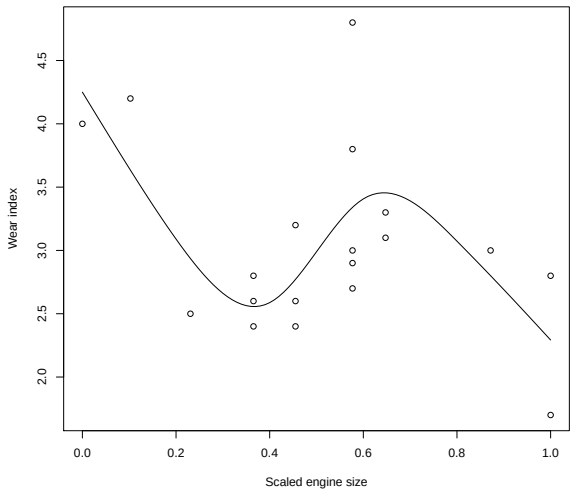


Abbildung 2.3: Gefittetes Modell für den Datensatz der Volvo-Automotoren.

Abbildung 2.3 zeigt den Scatterplot des Datensatzes und den Plot des gefitteten Modells. Der Fit schaut ziemlich gut aus, jedoch war die Wahl des Grades der Glätte, der durch die Basisdimension $q$ bestimmt wird, beliebig. Dies kann für verschiedene Wahlen von $q$ erhebliche Unterschiede bedeuten, wie Abbildung 2.4 zeigt. Das Problem der passenden Wahl von $q$ wird im Folgenden betrachtet.

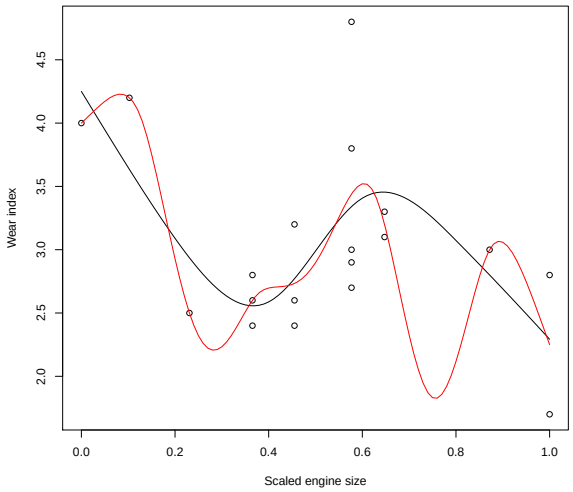


Abbildung 2.4: Grad der Glättung wird durch den Parameter $q$ kontrolliert; Schwarz: $q$=6, Rot: $q$=9

## 2.2 Penalisierte Regression-Splines (PRS)

Eine Möglichkeit der Wahl von $q$ wäre die *Backward-Selection*, d.h. mit einem „maximalen“ $q$-Wert zu starten und schrittweise den Parameter $q$ zu verringern, solange sich das Modell dadurch verbessert. Dieses Vorgehen ist jedoch im aktuellen Fall problematisch, da ein Modell mit $(k-1)$ gleichmäßig verteilten Knoten nicht unbedingt „nested“ in einem Modell mit $k$ gleichmäßig verteilten Knoten sein muss, d.h. dass eine Modell muss nicht Teilmenge des anderen Modells sein, was zu Widersprüchen führen würde.

Eine alternative Methode wäre es, die Basisdimension zu fixieren, groß genug, um ausreichend Flexibilität zuzulassen (Faustregel: meistens im Bereich 20 bis 40) und die Glätte des Modells durch eine hinzugefügte „*Schwankungs-Bestrafung*“ zu kontrollieren.

Statt das Modell durch Minimierung von

$$\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2$$

zu fitten, minimiert man folgenden Term:

$$\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \int_0^1 \left[f''(x)\right]^2 \mathrm{d}x.$$

Die zweite Ableitung gilt als Maß für die Krümmung einer Funktion, wodurch das Integral, die sogenannte *Spline-Strafe*, Modelle bestraft, die zu stark schwanken. Für $\lambda \to \infty$ folgt eine Geradenschätzung der Funktion $f$, während $\lambda = 0$ eine unbestrafte Schätzung des Regressions-Splines zur Folge hat.

Für den Strafterm lässt sich folgende Eigenschaft zeigen (Wood, 2006):

*Satz* 2.3. Da die Funktion $f$ linear im Parameter $\boldsymbol{\beta}$ ist, kann der Strafterm als quadratische Form in $\boldsymbol{\beta}$ dargestellt werden :

$$\int_0^1 \left[f''(x)\right]^2 \mathrm{d}x = \boldsymbol{\beta}^\top \mathbf{S} \boldsymbol{\beta},$$

wobei $\mathbf{S}$ eine Matrix der bekannten Koeffizienten ist, die sogenannte *Strafmatrix*.

*Beweis.* Es gilt

$$f(x) = \sum_{j=1}^{q} \beta_j b_j(x).$$

Nun lässt sich die zweite Ableitung der Funktion $f$ darstellen als

$$f''(x) = \boldsymbol{\beta}^\top \mathbf{d}(x),$$

wobei $\mathrm{d}_j(x) = b_j''(x)$. Daher folgt

$$\int_0^1 \left[f''(x)\right]^2 \mathrm{d}x = \int \boldsymbol{\beta}^\top \mathbf{d}(x)\mathbf{d}(x)^\top \boldsymbol{\beta} \mathrm{d}x = \boldsymbol{\beta}^\top \mathbf{S} \boldsymbol{\beta},$$

mit $\mathbf{S} = \int \mathbf{d}(x)\mathbf{d}(x)^\top \mathrm{d}x$. □

Es lässt sich für die Strafmatrix $\mathbf{S}$ zeigen, dass die ersten beiden Zeilen und Spalten nur aus Nullen bestehen, während für die übrigen Einträge gilt (Gu, 2002):

$$\mathbf{S}_{i+2,j+2} = R(x_i^*, x_j^*), \quad i, j = 1, \ldots, q-2.$$

Aufgrund der Form der Funktion $R(.\ ,\ .)$ ist sofort ersichtlich, dass die Matrix $\mathbf{S}$ eine quadratische $(q \times q)$-Matrix ist, denn

$$R(x_i, x_j) = R(x_j, x_i) \quad \forall i, j.$$

Das Problem der Parameterschätzung der Penalisierten Regressions-Splines lässt sich also darstellen durch

$$\min_{\boldsymbol{\beta}} \left\{ \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \boldsymbol{\beta}^\top \mathbf{S} \boldsymbol{\beta} \right\}.$$

Damit wurde das Schätzen des optimalen Grades an Glättung des Modells als Problem des Schätzens des optimalen Glättungsparameters $\lambda$ dargestellt.

Nun interessiert man sich für die Form des Least-Squares Schätzer $\hat{\beta}$. Sei $\lambda$ gegeben, dann gilt:

$$\begin{aligned} S_p &= \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\boldsymbol{\beta}^\top\mathbf{S}\boldsymbol{\beta} \\ &= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^\top\mathbf{S}\boldsymbol{\beta} \\ &= \mathbf{y}^\top\mathbf{y} - \mathbf{y}^\top\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{y} + \boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} + \lambda\boldsymbol{\beta}^\top\mathbf{S}\boldsymbol{\beta} \\ &= \mathbf{y}^\top\mathbf{y} - 2\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{y} + \boldsymbol{\beta}^\top\left(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{S}\right)\boldsymbol{\beta}. \end{aligned}$$

Hier wurde die Tatsache ausgenützt, dass alle Ausdrücke Skalare sind. Ableiten von $S_p$ nach $\boldsymbol{\beta}$ und Nullsetzen ergibt den penalisierten Least-Squares Schätzer $\hat{\boldsymbol{\beta}}$:

$$\begin{aligned} \mathbf{0} &\stackrel{!}{=} -2\mathbf{X}^\top\mathbf{y} + 2\left(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{S}\right)\boldsymbol{\beta} \\ \Rightarrow \hat{\boldsymbol{\beta}} &= \left(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{S}\right)^{-1}\left(\mathbf{X}^\top\mathbf{y}\right). \end{aligned}$$

Damit ergibt sich für die Hat-Matrix $\mathbf{H}$:

$$\hat{\mathbb{E}}(\mathbf{y}) = \hat{\boldsymbol{\mu}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \underbrace{\mathbf{X}\left(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{S}\right)^{-1}\mathbf{X}^\top}_{\mathbf{H}}\mathbf{y}.$$

Um eine bessere numerische Stabilität zu erreichen, lässt sich die Zielfunktion alternativ darstellen.

*Satz* 2.4. Für die Zielfunktion der penalisierten Regressions-Splines gilt:

$$\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\boldsymbol{\beta}^\top\mathbf{S}\boldsymbol{\beta} = \left\|\begin{pmatrix}\mathbf{y}\\ \mathbf{0}\end{pmatrix} - \begin{pmatrix}\mathbf{X}\\ \sqrt{\lambda}\mathbf{B}\end{pmatrix}\boldsymbol{\beta}\right\|^2,$$

mit $\mathbf{B}^\top\mathbf{B} = \mathbf{S}$.

*Beweis.*

$$\begin{aligned} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\boldsymbol{\beta}^\top\mathbf{S}\boldsymbol{\beta} &= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^\top\mathbf{B}^\top\mathbf{B}\boldsymbol{\beta} \\ &= \begin{pmatrix}\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\\ \mathbf{0} - \sqrt{\lambda}\mathbf{B}\boldsymbol{\beta}\end{pmatrix}^\top\begin{pmatrix}\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\\ \mathbf{0} - \sqrt{\lambda}\mathbf{B}\boldsymbol{\beta}\end{pmatrix} \\ &= \left\|\begin{pmatrix}\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\\ \mathbf{0} - \sqrt{\lambda}\mathbf{B}\boldsymbol{\beta}\end{pmatrix}\right\|^2 = \left\|\begin{pmatrix}\mathbf{y}\\ \mathbf{0}\end{pmatrix} - \begin{pmatrix}\mathbf{X}\\ \sqrt{\lambda}\mathbf{B}\end{pmatrix}\boldsymbol{\beta}\right\|^2. \end{aligned}$$

□

Die Matrix $\mathbf{B}$ entspricht der Wurzel der Strafmatrix $\mathbf{S}$, d.h. es gilt $\mathbf{B}^\top\mathbf{B} = \mathbf{S}$. Sie lässt

sich unter anderem durch *Eigen-Zerlegung* bestimmen.

Die Minimierung dieser neuen Fehlerquadratsummen entspricht der Least-Squares-Bedingung für ein erweitertes, nicht-penalisiertes Modell. Der Vektor $\mathbf{y}$ wurde mit dem Nullvektor augmentiert, die Modellmatrix $\mathbf{X}$ mit der Matrix $\sqrt{\lambda}\mathbf{B}$.

Um die Penalisierten Regressions-Splines anwenden zu können, muss zunächst eine Funktion geschrieben werden, um die Strafmatrix $\mathbf{S}$ zu berechnen. Als Input benötigt die Funktion `spl.S()` die Knotenpunkte `xk`:

Funktion zur Bestimmung der Strafmatrix $\mathbf{S}$

```
1 spl.S <- function(xk)
2 {
3   q <- length(xk)+2
4   S <- matrix(0,q,q)
5   S[3:q,3:q] <- outer(xk,xk,FUN=rk)
6   S
7 }
```

Die Matrix `S` wird mit Nullen initialisiert (Zeile 4), um dann mittels der `R`-Funktion `outer()` die Nicht-Nulleinträge einzutragen.

Zuletzt muss die Matrix $\mathbf{B} = \sqrt{\mathbf{S}}$ berechnet werden. Dies geschieht mittels der Funktion `mat.sqrt()`:

Funktion zur Bestimmung der Wurzel Strafmatrix $\mathbf{B}$

```
1 mat.sqrt <- function(S)
2 {
3   d <- eigen(S, symmetric=TRUE)
4   rS <- d$vectors %*% diag(d$values^0.5) %*% t(d$vectors)
5 }
```

Nun lässt sich eine erste Funktion schreiben, um penalisierte Regressions-Splines zu fitten:

Fitten der penalisierten Regressions-Splines

```
1 prs.fit <- function(y,x,xk,lambda)
2 {
3   q <- length(xk)+2
4   n <- length(x)
5   Xa <- rbind(spl.X(x,xk), mat.sqrt(spl.S(xk))*sqrt(lambda))
6   y[(n+1):(n+q)] <- 0
7   lm(y~Xa-1)
8 }
```

Der Parameter $q$ entspricht der Dimension der Basis, $n$ gibt die Anzahl der vorhandenen Daten an. In Zeile 5 wird die augmentierte Matrix $\begin{pmatrix}\mathbf{X}\\ \sqrt{\lambda}\mathbf{B}\end{pmatrix}$ berechnet. Zeile 6 erweitert den Datenvektor $\mathbf{y}$ mit dem Vektor $\mathbf{0}$. Zum Schluß wird das Modell gefittet und die

penalisierten Regressions-Splines ausgegeben.

Um die Funktion `prs.fit` anwenden zu können, müssen zuvor die Basisdimension $q$, die Knotenpunkte $x_j^*$ und der Glättungsparameter $\lambda$ bestimmt werden.

Unter der Voraussetzung, dass $q$ groß genug gewählt wird, damit die Basis ausreichend flexibel ist, ist weder die exakte Wahl von $q$, noch die Wahl der Knotenpunkte von großer Bedeutung für den Modellfit. Es ist dann die Wahl des Glättungsparameters $\lambda$, der den größten Einfluß auf die Modellflexibilität und die Form von $\hat{f}(x)$ hat. Diese Eigenschaft soll im Folgenden illustriert werden (Wood, 2006).

*Beispiel* 2.2. Sei $q = 9$ und die Knotenpunkte gleichmäßig auf $[0, 1]$ verteilt. Durch Fitten des Modells mit unterschiedlichen Glättungsparametern erkennt man, dass der Grad der Glättung nun in erster Linie von $\lambda$ abhängt:

Fitten der penalisierten Regressions-Splines

```
xk <- 1:7/8
mod.2 <- prs.fit(wear, x, xk, 0.000001)
Xp <- spl.X(xp,xk)
plot(x,wear)
lines(xp,Xp %*% coef(mod.2))
```

Durch unterschiedliche Wahlen von $\lambda$ erhält man unterschiedliche Grade an Glättung, wie in Abbildung (2.5) ersichtlich ist. Die Wahl des Glättungsparameters $\lambda = 0.01$ ist für den vorhandenen Datensatz zu groß, da der Fit zu wenig Flexibilität aufweist. Im Gegensatz dazu würde eine Wahl von $\lambda = 1 \cdot 10^{-6}$ zuviel Schwankung zulassen, wodurch der erhaltene Fit schwer zu interpretieren ist, da zuviel Schwankung aufgezeigt wird. In beiden Fällen wird die Schätzung $\hat{f}$ recht weit von der wahren Funktion $f$ entfernt sein. Der für das aktuelle Beispiel vernünftige Wert des Glättungsparameters scheint $\lambda = 1 \cdot 10^{-4}$ zu sein.

Es stellt sich nun die Frage nach der passenden Wahl von $\lambda$. Diese Frage wird im folgenden Unterkapitel behandelt. Dabei wird die Notation $f_i \equiv f(x_i)$ verwendet.

## 2.3 Wahl des Glättungsparameters $\lambda$: Kreuzvalidierung

Ein passendes Kriterium für die Wahl von $\lambda$ wäre die Minimierung von

$$M = \frac{1}{n} \sum_{i=1}^{n} \left( \hat{f}_i - f_i \right)^2 .$$

Aber die wahre Funktion $f$ ist unbekannt! Eine Lösung dieses Problems ist die sogenannte *Kreuzvalidierung* (Wood, 2006).

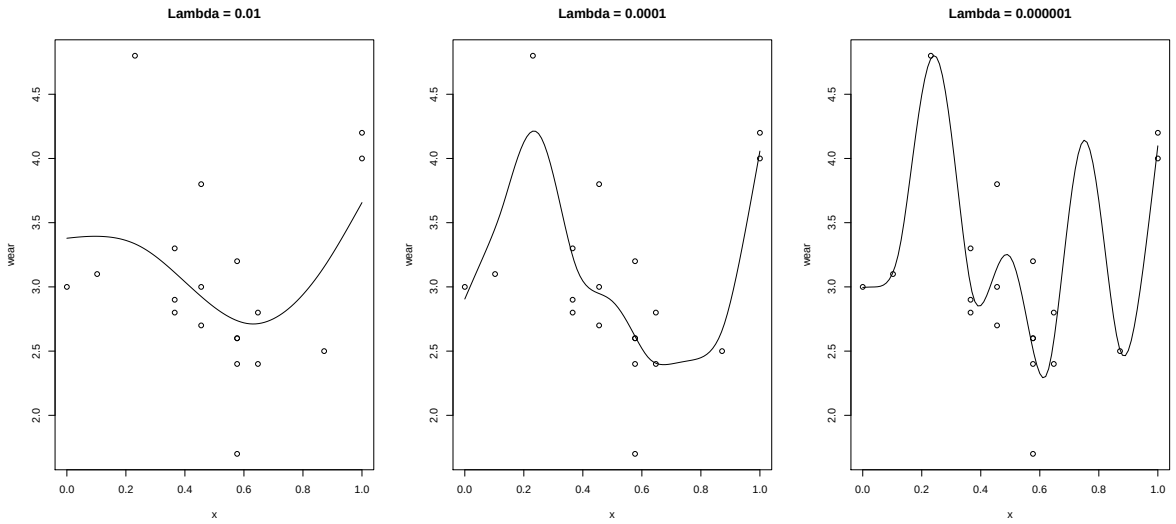


Abbildung 2.5: Unterschiedliche $\lambda$ ergeben unterschiedliche Grade der Glättung.

Das Ziel von Kreuzvalidierungsverfahren ist es, die Güte eines Modells zu bestimmen. In der Regel gibt es dazu Kennzahlen, wie zum Beispiel das *Akaike Informationskriterium* (AIC). Diese Kennzahlen basieren jedoch zum Teil auf asymptotischer Theorie, das heißt, es wird eine sehr große Stichprobenmenge vorausgesetzt. Ein weiteres Problem entsteht, wenn mit denselben Daten das Modell gefittet wird, mit denen auch der Fehler geschätzt wird. Dadurch wird nämlich die Modellqualität überschätzt (*In-Sample-Error*). Man benötigt also Testdaten, um diesem Problem vorzubeugen.

Bei Kreuzvalidierungsverfahren wird der Datensatz in zwei Teile geteilt. Der erste Teil wird dazu verwendet, um die Modellparameter zu schätzen, mit dem zweiten Teil wird der Modellfehler geschätzt (*Out-of-Sample-Error*). Im Folgenden wird der Spezialfall betrachtet, dass genau ein Stichprobenpunkt zur Bestimmung des Modells ausgelassen wird (*Leave-one-Out*).

Sei $\hat{f}^{[-i]}$ das ohne den Datenpunkt $y_i$ gefittete Modell. Dann betrachte man den *Ordinary Cross Validation* (OCV)-Score

$$\nu_o = \frac{1}{n} \sum_{i=1}^{n} \left( \hat{f}_i^{[-i]} - y_i \right)^2 .$$

Einsetzen des Datenpunktes $y_i = f_i + \varepsilon_i$ und umformen liefert

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

Laut Annahme gilt $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. Da $\varepsilon_i$ und $\hat{f}_i^{[-i]}$ unabhängig sind, ist der Erwartungswert des mittleren Terms Null und es folgt

$$\mathbb{E}(\nu_o) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}\left(\hat{f}_i^{[-i]} - f_i\right)^2\right) + \sigma^2.$$

Wegen $\hat{f}_i^{[-i]} \approx \hat{f}_i$ folgt, dass $\mathbb{E}(\nu_o) \approx \mathbb{E}(M) + \sigma^2$, mit Gleichheit für große Stichproben. Wenn also zur Bestimmung von $\hat{f}$ die Minimierung von $M$ optimal wäre, stellt die Minimierung von $\nu_o$ eine vernünftige Methode dar.

Die Berechnung des OCV-Scores ist jedoch sehr aufwendig (für jedes $i$ muss $\hat{f}_i^{[-i]}$ berechnt werden). Man kann aber zeigen, dass gilt:

$$\nu_o = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{f}_i - y_i\right)^2 / (1 - H_{ii})^2,$$

wobei $H_{ii}$ der $i$-te Eintrag der Hauptdiagonale der Hat-Matrix ist.

Aufgrund von numerischen Vorteilen wird in der Praxis statt dem OCV häufig die *generalisierte Kreuzvalidierung* verwendet. Hier werden die Gewichte $1 - H_{ii}$ durch ihr Mittel $\mathrm{tr}(\mathbb{I} - \mathbf{H})/n$ ersetzt und man erhält den *Generalized Cross Validation* (GCV)-Score

$$\nu_g = \frac{n\sum_{i=1}^{n}\left(y_i - \hat{f}_i\right)^2}{\left[\mathrm{tr}(\mathbb{I} - \mathbf{H})\right]^2}.$$

Es folgt eine direkte Anwendung der Suche nach dem optimalen Glättungsparameter mittels Minimierung des GCV-Scores. Dazu wird abermals der Datensatz der Volvo-Automotoren verwendet.

Berechnung des Glättungsparameters durch Minimierung des GCV-Scores

```
lambda <- 1e-8; n <- length(wear); V <- 0
for (i in 1:60)
{
  mod <- prs.fit(wear, x, xk, lambda)
  trA <- sum(influence(mod)$hat[1:n])
  rss <- sum((wear-fitted(mod)[1:n])^2)
  V[i] <- n*rss/(n-trA)^2
```

```
8   lambda <- lambda*1.5
9 }
10 which.min(V)
```

```
[1] 29
```

Als Startwert wird für den Glättungsparameter $\lambda = 1 \cdot 10^{-8}$ gewählt. In jedem Schleifendurchlauf wird $\lambda$ um den Faktor 1.5 erhöht. Der Vektor `V` entspricht dem GCV-Score für den jeweiligen Glättungsparameter. In Zeile 4 werden die penalisierten Regressions-Splines gefittet. Der `R`-Befehl `which.min()` gibt den Index an, bei dem der Vektor `V` sein Minimum annimmt. Im aktuellen Beispiel ist dies bei $i = 29$ der Fall. D.h. der als optimal berechnete Glättungsparameter entspricht dem Wert $\hat{\lambda} = 1.5^{28} \times 10^{-8} \approx 9 \cdot 10^{-4}$.

Nun soll das Modell mit dem gewählten Glättungsparameter geschätzt werden.

Schätzen des Modells

```
1 i <- which.min(V)
2 mod.3 <- prs.fit(wear, x, xk, 1.5^(i-1)*1e-8)
3 Xp <- spl.X(xp, xk)
4 plot(x,wear, cex.axis=1.5, cex.lab=2)
5 lines(xp, Xp %*% coef(mod.3))
```

Abbildung (2.6) zeigt den mittels Minimierung des GCV-Scores erhaltenen, optimalen Fit. Dieser scheint die zugrunde liegenden Daten ziemlich gut darzustellen.

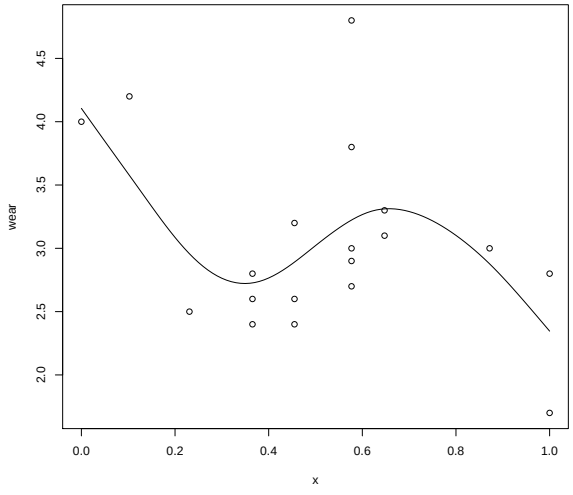


Abbildung 2.6: Fit beruhend auf den durch Minimierung des GCV-Scores erhaltenen optimalen Glättungsparameter $\hat{\lambda} \approx 9 \cdot 10^{-4}$.

Die vorgestellte Methode der Anwendung kubischer Splines zeigt bereits einen ziemlich guten Fit, sie hat jedoch einen erheblichen Nachteil: Es handelt sich bei den verwendeten Basen um *globale* Basen, wie Abbildung 2.7 zeigt (Peternell, 2013):

Darstellung der globalen Basen

```
beta <- coef(mod.1)
plot(x, wear)
lines(xp,Xp%*%coef(mod.1), lwd=2)
for(i in 1:6)
{
lines(xp, beta[i]*Xp[,i], lty=2)
}
```

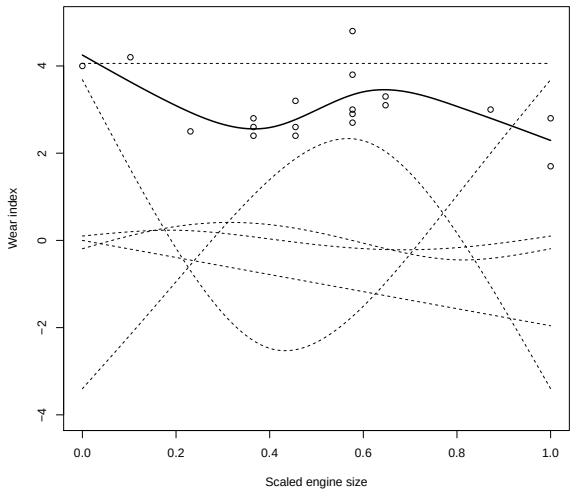


Abbildung 2.7: Kubische Spline-Basen und deren Summe, die geschätzte Funktion $\hat{f}(x)$.

Die gestrichelten Linien stellen die einzelnen Basisfunktionen dar, die durch Multiplikation mit dem gefitteten Regressionsparameter und anschließender Summe den Fit ergeben, dargestellt in Abbildung 2.7 als durchgehende Linie.

Durch die Globalität dieser Basen tritt ein nicht unwesentlicher Nachteil auf: Änderungen der einzelnen Koeffizienten wirken sich global auf die gesamte Funktion aus. Um dies zu umgehen, ist eine Verwendung von lokalen Basen vorteilhaft.

Es existieren zwei äquivalente Methoden zur Darstellung von lokalen Polynom-Splines: *Trunkierte Potenzen* und *B-Splines*. Da jedoch bei durch trunkierte Potenzen erzeugte Basen numerische Probleme entstehen können und diese Basisfunktionen annäherend lineare Abhängigkeit aufweisen, wird im Folgenden nur auf B-Splines eingegangen.

## 2.4 Polynom Splines mittels B-Splines

Für die Basisfunktion der *B-Splines* muss die Notation etwas abgeändert werden. Da diese Funktionen im Gegensatz zu den Basisfunktionen der Glättungsfunktion in Gleichung 2.1 *nicht* global, sondern lokal sind, kann hier nicht einfach über die Basisdimension summiert werden. Stattdessen wird über die $k$ Knotenpunkte summiert, auf denen der B-Spline definiert ist.

Ein Spline der Ordnung $q = (m + 1)$ kann dargestellt werden als

$$f(x) = \sum_{j=1}^{k} B_j^m(x)\beta_j,$$

wobei die $k$-parametrischen Basisfunktionen $B_j^m$ rekursiv definiert werden (Wood, 2006):

$$B_j^m(x) = \frac{x - x_j}{x_{j+m+1} - x_j} B_j^{m-1}(x) + \frac{x_{j+m+2} - x}{x_{j+m+2} - x_{j+1}} B_{j+1}^{m-1}(x), \quad j = 1, \ldots, k,$$

mit

$$B_j^{-1}(x) = \begin{cases} 1, & x_j \leq x < x_{j+1}, \\ 0, & \text{sonst.} \end{cases}$$

Ein wesentlicher Vorteil der B-Splines ist ihre *strikte Lokalität*. Jede Basisfunktion ist nur auf Intervallen zwischen $(m + 3)$ benachbarten Knoten ungleich null.

Um eine $k$-parametrische B-Spline Basis zu konstruieren, müssen zunächst $(k + m + 2)$ Knoten $x_1 < x_2 < \ldots < x_{k+m+2}$ bestimmt werden. Das Intervall, worüber der Spline berechnet wird, liegt dann in $[x_{m+2}, x_k]$, d.h. die ersten und letzten $m+1$ Knotenpunkte sind beliebig.

Mittels folgendem `R`-Code lassen sich B-Spline-Basisfunktion der Ordnung 3 an den Punkten $x$ auswerten:

Auswertung der B-Splines

```
1  bspline <- function(x,xk,i,m=2)
2  {
3    if (m==-1)
4    {
5      res <- as.numeric(x < xk[i+1] & x >= xk[i])
6    }
7    else
8    {
9      z0 <- (x - xk[i])/(xk[i+m+1] - xk[i])
10     z1 <- (xk[i+m+2] - x)/(xk[i+m+2] - xk[i+1])
11     res <- z0*bspline(x,xk,i,m-1) + z1*bspline(x,xk,i+1,m-1)
```

```
12   }
13   res
14 }
```

Als Input erhält die Funktion `bspline` die Auswertungspunkte `x`, sowie die Knotenpunkte `xk`. Dann werden die Basisfunktionen rekursiv definiert, wobei der Befehl `as.numeric` in Zeile 5 die Eins widergibt, sofern der logische Ausdruck `x < xk[i+1] & x >= k[i]` wahr ist.

B-Splines werden in erster Linie dazu verwendet, um sogenannte *P-Splines* (Penalisierte Splines) zu konstruieren (Wood, 2006). Dies sind Glättungen, die für gewöhnlich auf gleichmäßig verteilten Knoten definiert werden. Zusätzlich wird ein Strafterm eingefügt, der die Schwankung der Funktion kontrollieren soll.

Der Begriff *P-Spline* wurde in Verbindung mit dem folgenden Strafterm $\mathcal{P}$ eingeführt:

$$\mathcal{P} = \sum_{j=1}^{k-1} (\beta_{j+1} - \beta_j)^2 = \beta_1^2 - 2\beta_1\beta_2 + 2\beta_2^2 - 2\beta_2\beta_3 + \cdots + \beta_k^2,$$

wobei natürlich auch eine andere Strafform angewendet werden kann.

Hier werden also die quadrierten Differenzen benachbarter Parameterwerte $\beta_j$ bestraft. Verallgemeinert lässt sich dies darstellen als

$$\mathcal{P} = \boldsymbol{\beta}^\top \underbrace{\begin{bmatrix} 1 & -1 & 0 & \cdot & \cdot \\ -1 & 2 & -1 & \cdot & \cdot \\ 0 & -1 & 2 & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix}}_{=\mathbf{S}} \boldsymbol{\beta} \tag{2.4}$$

(2.5)

Mittels `R` lässt sich die Strafmatrix $\mathbf{S}$ in (2.4) wie folgt bestimmen:

Berechnung der Strafmatrix

```
1 q <- 6
2 B <- diff(diag(q),differences=1)
3 S <- t(B) %*% B
```

Als Beispiel wurde hier eine Basis mit Dimension $q = 6$ betrachtet. Es wurde sowohl die Wurzel $\mathbf{B}$, als auch die Strafmatrix $\mathbf{S}$ selbst berechnet. Strafen höherer Ordnung können verwendet werden, indem man den Parameter `differences` erhöht.

Die Funktion `diff()` hat als Defaultwert einen lag von 1, d.h. es werden zwei aufeinanderfolgende Werte betrachtet. Falls der Eingabewert `x` eine Matrix ist, wird der Differenzenoperator auf jede Spalte angewendet:

$$\mathbf{X}[(1 + \text{lag}) : n] - \mathbf{X}[1 : (n - \text{lag})].$$

Um die R-Funktion `diff()` etwas näher zubringen, sei der lag = 1 und die betrachtete Matrix die $(4 \times 4)$ Einheitsmatrix:

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Nun wird darauf die Funktion `diff` angewendet, mit Parameter `differences` = 2.

Zunächst wird die erste Spalte berechnet:

$$\begin{bmatrix} x_2 - x_1 \\ x_3 - x_2 \\ x_4 - x_3 \end{bmatrix} = \begin{bmatrix} -1 \\ 0 \\ 0 \end{bmatrix}.$$

Analog ergeben die Spalten zwei bis vier

$$\begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}.$$

Wegen `differences` = 2 wird das ganze Prozedere nochmal durchgeführt und man erhält

$$\tilde{\mathbf{X}} = \begin{bmatrix} 1 & -2 & 1 & 0 \\ 0 & 1 & -2 & 1 \end{bmatrix}.$$

Ein Nachteil bei der betrachteten Bestrafungsmethode ist, dass der Strafterm nicht so einfach interpretiert werden kann wie bei der zuvor betrachteten Spline-Strafe.

Es soll nun anhand eines kleinen Beispiels demonstriert werden, wie man relativ einfach B-Splines inklusive Strafterm berechnen kann (Wood, 2006).

*Beispiel* 2.3 (Anwendungsbeispiel der P-Splines). Zunächst soll die Modellmatrix (X) und die Wurzel der Strafmatrix (B) berechnet werden. Dazu wird als Input der Prädiktorvektor `x`, die Basisdimension `q` und die Ordnung der B-Splines `m+1`, sowie der Straf-Differenz `p.m` benötigt:

Berechnung der Modellmatrix und der Wurzel der Strafmatrix

```
1 library(splines)
2 pspline.XB <- function(x, q = 10, m = 2, p.m = 2)
```

```
{
  k <- seq(min(x), max(x), length = q-m)
  dk <- k[2] - k[1]
  k <- c(k[1] - dk*((m+1):1), k, k[q-m] + dk*(1:(m+1)))
  X <- splineDesign(k, x, ord = m+2)
  B <- diff(diag(q),differences = p.m)
  list(X=X, B=B)
}
```

Nun werden 100 gleichmäßig auf dem Intervall $(0, 1)$ verteilte $x$-Werte simuliert. Darauf soll die Funktion `pspline.XB` angewendet werden, um die Basisfunktion einer Rang 9 B-Spline Basis in den $x$-Punkten auszuwerten. Es handelt sich hierbei um einen kubischen B-Spline mit einem Strafterm zweiter Ordnung:

Berechnung der Modellmatrix und der Wurzel der Strafmatrix

```
n <- 100
x <- sort(runif(n))
ps <- pspline.XB(x, q=9, m=2, p.m=2)
par(mfrow=c(3,3))
for (i in 1:9) plot(x, ps$X[,i], type="l")
```

Abbildung (2.8) zeigt die dadurch entstehenden Rang 9 B-Spline Basisfunktionen.

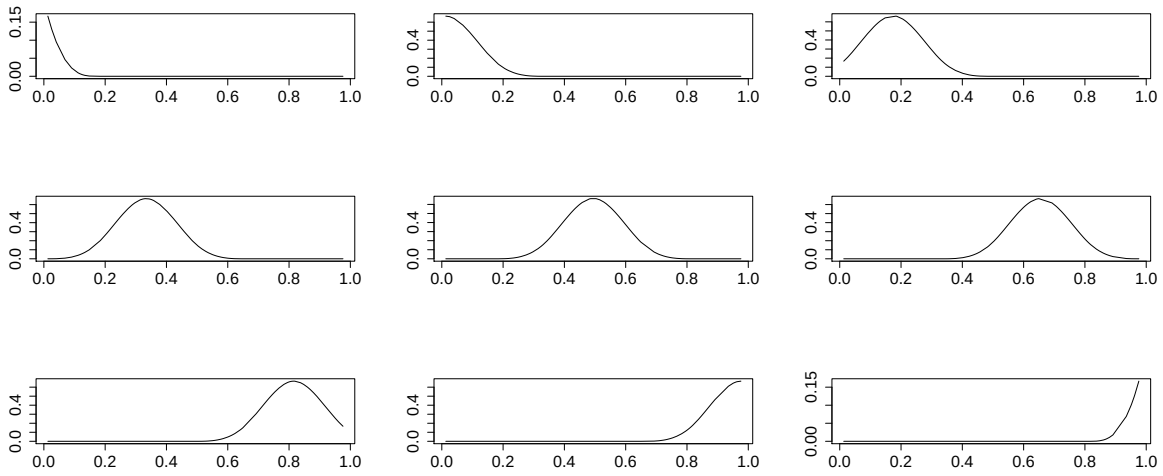

Abbildung 2.8: Rang 9 B-Spline Basisfunktionen ausgewertet über 100 gleichmäßig verteilten $x$-Werten in $(0, 1)$.

P-Splines sind sehr einfach anzuwenden und erlauben einiges an Flexibilität, da jede beliebige Strafordnung mit jeder beliebigen Ordnung der B-Spline Basis kombiniert werden

kann. Falls keine gleichmäßig verteilten Knoten vorliegen, sind sie etwas komplizierter anzuwenden.

## 2.5 Anwendungsbeispiel der Funktion `gam()`

Als Beispiel eines GAMs betrachte man folgendes Modell für den Datensatz `trees`:

$$\log\left\{\mathbb{E}\left[\text{Volume}_i\right]\right\} = f_1(\text{Girth}_i) + f_2(\text{Height}_i), \quad \text{Volume}_i \sim \text{Gamma}. \tag{2.6}$$

Als Link-Funktion wird hier der Logarithmus angewendet. Dies erscheint äußerst natürlich, da man erwartet, dass das Volumen ein Produkt einer Funktion des Umfangs mit einer Funktion der Höhe ist.

Statt mittels penalisierten Least-Squares werden GAMs mittels penalisierter Likelihood-Maximierung gefittet. Dazu wird das Konvergenzschema der *Penalized Iteratively re-weighted least squares* (P-IRLS) verwendet.

Sei die im Iterlationsschritt $(t)$ aktuelle Parameterschätzung $\boldsymbol{\beta}^{(t)}$ und die dazugehörige Schätzung des Erwartungswert des Responsevektors $\boldsymbol{\mu}^{(t)}$ gegeben.

1. Berechne

$$\omega_i^{(t)} \propto \frac{1}{V\left(\mu_i^{(t)}\right) g'\left(\mu_i^{(t)}\right)^2}, \quad z_i^{(t)} = g'\left(\mu_i^{(t)}\right)\left(y_i - \mu_i^{(t)}\right) + \mathbf{x}_i^\top \boldsymbol{\beta}^{(t)},$$

   wobei $\text{Var}(Y_i) = \phi V\left(\mu^{(t)}\right)$, $\mathbf{x}_i^\top$ entspricht der $i$-ten Zeile der Modellmatrix $\mathbf{X}$.

2. Minimiere folgendes Kriterium bezüglich $\boldsymbol{\beta}$:

$$\left\| \begin{bmatrix} \sqrt{\mathbf{W}} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \left( \begin{bmatrix} \mathbf{z} \\ \mathbf{0} \end{bmatrix} - \begin{bmatrix} \mathbf{X} \\ \mathbf{B} \end{bmatrix} \boldsymbol{\beta} \right) \right\|^2,$$

   wobei $\mathbf{W}$ eine Diagonalmatrix ist, mit $\mathbf{W} = \text{diag}\left(w_i^{(t)}\right)$ und $\mathbf{z} = \left(z_1^{(t)}, \ldots, z_n^{(t)}\right)^\top$, $\mathbf{B}$ ist die Quadratwurzel von $\mathbf{B}^\top\mathbf{B} = \lambda_1\mathbf{S}_1 + \lambda_2\mathbf{S}_2$. Die Matrizen $\mathbf{S}_1$ und $\mathbf{S}_2$ entsprechen den Strafmatrizen für die jeweiligen Glättungsfunktionen der Prädiktorvariablen `Height` und `Girth`. Minimierung der Bedingung führt zur Schätzung $\boldsymbol{\beta}^{(t+1)}$.

Im aktuellen Beispiel ist die Linkfunktion $g = \log$, wodurch $g'(\mu_i) = \mu_i^{-1}$ folgt. Für die Gamma-Verteilung gilt $\text{V}(\mu_i) = \mu_i^2$. Daher ergibt sich für das Modell (2.6):

$$\omega_i = 1 \quad \text{und} \quad z_i^{(t)} = \frac{\left(y_i - \mu_i^{(t)}\right)}{\mu_i^{(t)}} + \mathbf{x}_i^\top \boldsymbol{\beta}^{(t)}.$$

Zur Anwendung von GAMs wurde in R das Paket mgcv (*Mixed GAM Computation Vehicle*) (Wood, 2006) implementiert. Darin befindet sich die Funktion gam(), mit der sich Generalisierte Additive Modelle fitten lassen.

Fitten eines GAMs

```
1 library(mgcv)
2 data(trees)
3 attach(trees)
4 gam.mod <- gam(Volume ~ s(Girth, bs="ps") + s(Height, bs="ps"),
5                family=Gamma(link = "log"), data=trees)
```

Zunächst muss in der Funktion gam() der funktionelle Zusammenhang zwischen Volume und den Prädiktorvariablen Girth und Height angegeben werden. Durch die Funktion s() wird die Glättung innerhalb des GAMs definiert, wobei mittels bs="ps" P-Splines verwendet werden. Mittels dem Befehl family werden die Verteilung und Link-Funktion des Modells spezifiziert.

Anhand des summary-Outputs erkennt man, dass die Funktion gam() standardmäßig das Modell durch Minimierung des GCV-Score bestimmt wird:

Interpretation des GAMs

```
1 summary(gam.mod)
```

```
Family: Gamma
Link function: log

Formula:
Volume ~ s(Girth, bs = "ps") + s(Height, bs = "ps")

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.27572    0.01496     219   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
            edf Ref.df      F  p-value
s(Girth)  2.318  2.865 229.84  < 2e-16 ***
s(Height) 1.000  1.000  31.41 6.29e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.973   Deviance explained = 97.8%
GCV score = 0.0080567  Scale est. = 0.0069344  n = 31
```

Um den `summary()`-Output interpretieren zu können, muss zunächst der Begriff des *Estimated Degree of Freedom* geklärt werden. Dazu wird der Begriff des *Effective Degree of Freedom* benötigt (Wood, 2006).

Für ein Modell ohne Strafterm hat der Parameterschätzer $\tilde{\boldsymbol{\beta}}$ die Form

$$\tilde{\boldsymbol{\beta}} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{y}.$$

Für ein Modell mit Strafterm ändert sich der Parameterschätzer zu

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= \left(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{S}\right)^{-1} \mathbf{X}^\top \mathbf{y} \\
&= \left(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{S}\right)^{-1} \mathbf{X}^\top \mathbf{X} \underbrace{\left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{y}}_{=\tilde{\boldsymbol{\beta}}} \\
&= \mathbf{F} \tilde{\boldsymbol{\beta}}.
\end{aligned}$$

Die Matrix $\mathbf{F}$ bildet also den unbestraften Parameterschätzer auf $\hat{\boldsymbol{\beta}}$ ab. Es gilt

$$\frac{\partial \hat{\boldsymbol{\beta}}}{\partial \tilde{\boldsymbol{\beta}}} = \mathbf{F}.$$

Das $j$-te Hauptdiagonalelement $F_{jj}$ gibt also an, wieviel sich der bestrafte Schätzer $\hat{\boldsymbol{\beta}}$ ändert, wenn sich der unbestrafte Schätzer $\tilde{\boldsymbol{\beta}}$ um eins ändert. Man sagt, $F_{jj}$ misst den *Effective Degree of Freedom* des $j$-ten penalisierten Parameters.

Wird ein einzelner Parameter geschätzt, verliert man einen Freiheitsgrad. Schätzt man zum Beispiel einen kubischen Glättungsspline, so ist im Vornhinein nicht klar, wie viele Freiheitsgrade verloren gehen. Eine Geradenschätzung ist beispielsweise sehr glatt, die Glättungsfunktion folgt kaum dem Muster des Scatterplots der Daten. Hier muß nur ein Slope-Parameter (der Parameter, der die Steigung der Gerade festlegt) geschätzt werden. Eine sehr raue Schätzung könnte jeden Datenpunkt des Scatterplots beinhalten und dadurch approximativ soviele Freiheitsgrade verlieren, wie Datenpunkte vorhanden sind. Der *Estimated Degree of Freedom* gibt somit den Effective Degree of Freedom für die jeweils geschätzte Glättungsfunktion an. Je glatter die Glättungsfunktion geschätzt wird, desto größer ist der Estimated Degree of Freedom.

Um nun eine Aussage über die Signifikanz der Glättungsfunktion $f(x)$ treffen zu können, wird sie in einen Intercept-Term und in einen Glättungsterm $\tilde{f}(x)$ aufgeteilt:

$$f(x) = \text{Intercept} + \tilde{f}(x).$$

Der Glättungsterm $\tilde{f}(x)$ unterscheideit sich von der Glättungsfunktion $f(x)$ also genau um den Intercept. Er gibt an, wie stark der Erwartungswert (d.h. die Glättungsfunktion

$f(x)$) von einer horizontalen Gerade abweicht, unter der Bedingung, dass er glatt in $x$ ist.

Prinzipiell spricht ein sehr kleiner *Estimated Degree of Freedom* gegen einen Glättungsterm. Eine Anleitung dafür, wann ein Glättungsterm $\tilde{f}(x)$ aus dem Modell entfernt werden soll, findet man in Wood (2001). Dafür stellt er drei Fragen zur Verfügung:

1. Ist der *Estimated Degree of Freedom* (`edf`) des Terms nahe eins?
2. Überdeckt das Band der Konfidenzintervalle zu $\tilde{f}(x)$ die null über den Gesamtbereich der Werte von x?
3. Verringert sich der GCV-Score durch Entfernen des Terms?

Können alle drei Fragen mit ja beantwortet werden, soll der Term entfernt werden. Wird nur die erste Frage bejaht, soll der Glättungsterm durch einen parametrischen, linearen Term ersetzt werden. Sollten Korrelationen zwischen einigen erklärenden Variablen bestehen, sollten Terme nur einzeln entfernt werden, beginnend bei jenem Term, bei dem die Null am ehesten von dessen Konfidenzband überdeckt wird.

Man kann mittels folgendem Codeschnipsel die geschätzten Glättungsparameter der beiden Glättungsterme ausgeben.

Ausgabe der geschätzten Glättungsparameter

```
1 gam.mod$sp
```

```
    s(Girth)    s(Height)
1.727378e+01 6.348960e+06
```

Mittels `plot()` lassen sich die durch das Modell erhaltenen Glättungsterme in Abhängigkeit von den linearen Prädiktoren darstellen, wie Abbildung 2.9 zeigt.

Plotten der geschätzten Glättungsfunktionen samt Konfidenzband

```
1 plot(gam.mod)
```

Schließlich lässt sich die Güte des Fits noch mit einem Plot des beobachteten `Volume` gegen das durch das Modell gefitte Volumen darstellen.

Vergleich des beobachteten Volumen und dem geschätzten Volumen

```
1 plot(gam.mod$fitted, Volume)
```

Abbildung 2.10 zeigt deutlich eine 45°-Gerade, was für den Fit des Modells spricht.

Nun wird noch eine Möglichkeit dargestellt, mittels Approximation durch die Normalverteilung 95% Konfidenzintervalle für log-lineare GAMs zu berechnen. Dafür wird das `trees`-Beispiel auf den Prädiktor `Girth` reduziert.

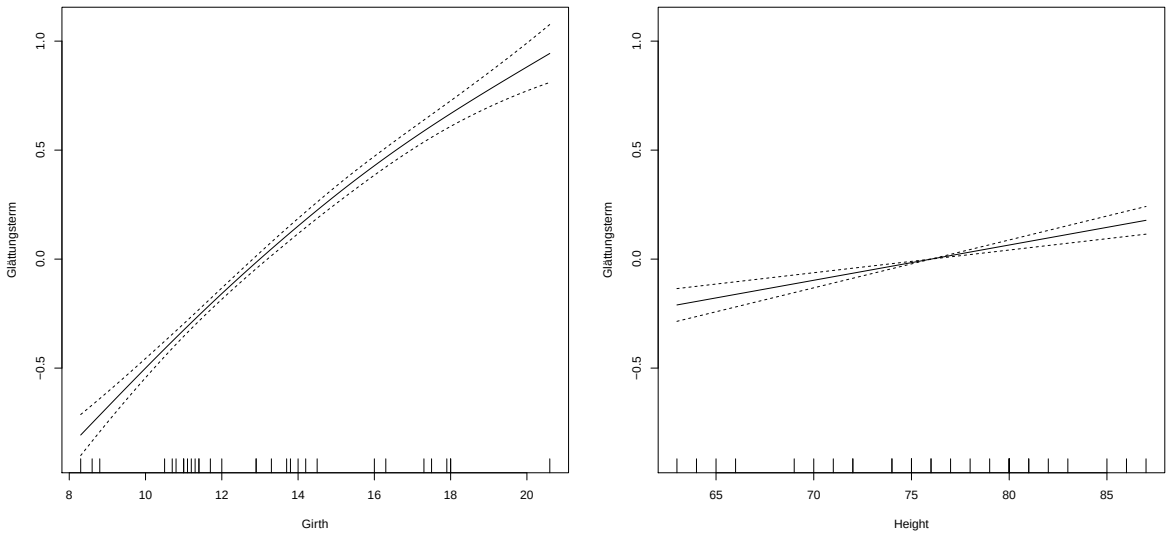


Abbildung 2.9: Glättungsterme bezüglich den linearen Prädiktoren des GAMs zum Datensatz `trees`. Der Glättungsterm des Prädiktors Girth (links) besitzt einen EDF von 2.32, für den Prädiktor Height (rechts) gilt EDF = 1.

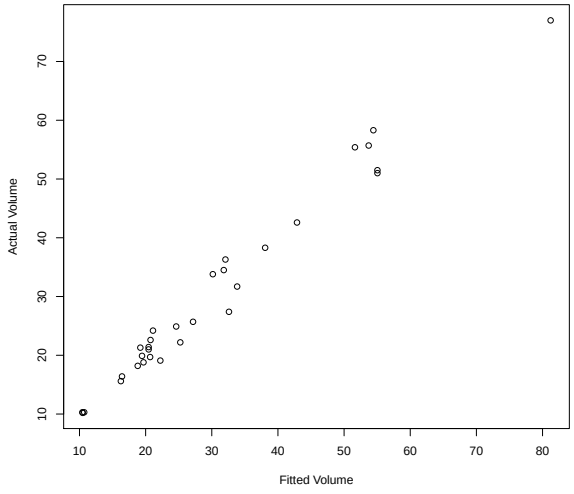


Abbildung 2.10: Scatterplots des beobachteten Volumens gegenüber dem mit dem GAM gefitteten Volumen.

Berechnung von Konfidenzintervallen für ein log-lineares GAM

```
gam.mod2 <- gam(Volume~s(Girth, bs="ps"), family=Gamma(link = "log"),
                data=trees)
newd <- data.frame(Girth)

mu <- predict.gam(gam.mod2, newd, type="lpmatrix")%*%coef(gam.mod2)
se <- as.numeric(predict.gam(gam.mod2, newd, se="TRUE")$se.fit)

plot(Girth, Volume)
lines(Girth, fitted(gam.mod2))
lines(Girth, exp(mu+1.96*se), lty=2)
lines(Girth, exp(mu-1.96*se), lty=2)
```

Da die Erwartungswerte des Modells logarithmiert werden, müssen die berechneten Konfidenzintervalle exponiert werden. Abbildung 2.11 zeigt die Schätzung des Erwartungswerts von `Volumen` samt der dazugehörigen 95% Konfidenzintervalle.

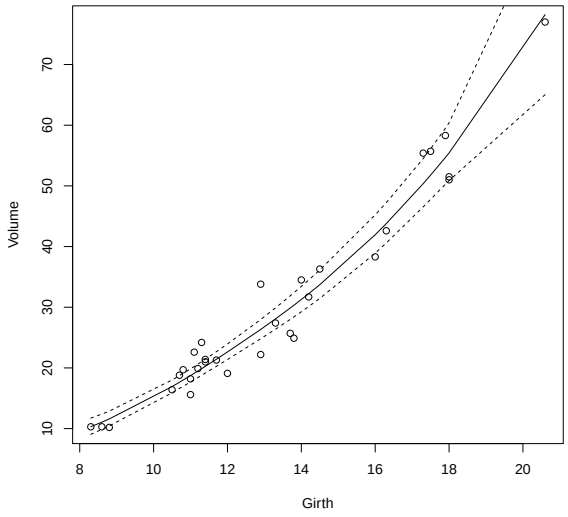


Abbildung 2.11: Log-lineares GAM von `Volumen` mit dem Prädiktor `Girth` inklusive der durch Approximation mittels Normalverteilung erzeugten 95% Konfidenzintervalle.

# 3 Quantile Regression

Mit Methoden der klassischen Regression wird der Erwartungswert $\mathbb{E}(Y \mid x)$ der Response $Y$ durch die erklärende Variable $x$ modelliert, beziehungsweise für Generalisierte Regressionsmodelle eine Funktion des Erwartungswertes. Dies bietet jedoch nur eine recht eingeschränkte Sichtweise des Verhältnisses zwischen $Y$ und $x$, da man zum Beispiel an deren Beziehung zueinander bei unterschiedlichen Punkten der Verteilung von $Y$ interessiert sein kann. Mittels geschätzten Quantilsfunktionen lässt sich ein viel ausführlicheres Bild der Regression darstellen, wobei die Schätzer zusätzlich noch wesentlich unempfindlicher gegenüber Ausreißern sind als herkömmliche Erwartungswert-Schätzer und keine Verteilungsannahmen getroffen werden müssen.

Zunächst wird der Begriff des *Quantils* definiert (Genschel & Becker, 2005).

*Definition* 3.1 (Quantil). Sei $Y$ eine Zufallsvariable mit Verteilungsfunktion $F$ und $\tau \in (0,1)$. Dann ist das $\tau$-*Quantil* $q_\tau$ für $0 < \tau < 1$ definiert durch

$$\mathbb{P}(Y \leq q_\tau) \geq \tau \quad \text{und} \quad \mathbb{P}(Y \geq q_\tau) \geq 1 - \tau. \tag{3.1}$$

Ist $q_\tau$ nicht eindeutig bestimmbar, wählt man den kleinsten Wert, der die Bedingung (3.1) erfüllt:

$$q_\tau = F^{-1}(\tau) = \inf\{y : F(y) \geq \tau\} = \inf\{y : \mathbb{P}(Y \leq y) \geq \tau\}.$$

Die Funktion $F^{-1}(y)$ wird als *Quantilsfunktion* bezeichnet.

Die Quantile einer Verteilungsfunktion $F$ lassen sich durch ein einfaches Optimierungsproblem bestimmen, auf das die gesamte Theorie der Quantilen Regression (QR) aufgebaut ist. Dazu betrachtet man die sogenannte *Verlustfunktion* (Englisch: Loss Function)

$$\rho_\tau(y) = y\left(\tau - \mathbb{1}_{(y<0)}\right),$$

wobei $\tau \in (0,1)$ (Koenker, 2005). Für negative Argumente $y < 0$ ergibt sich folgende Gestalt:

$$\rho_\tau(y) = y\left(\tau - \mathbb{1}_{(y<0)}\right) = y(\tau - 1) = (1-\tau)|y|.$$

Für positive Argumente $y > 0$ folgt

$$\rho_\tau(y) = y\left(\tau - \mathbb{1}_{(y<0)}\right) = \tau y.$$

Abbildung 3.1 zeigt eine Darstellung der Verlustfunktion für $\tau = 0.25$.

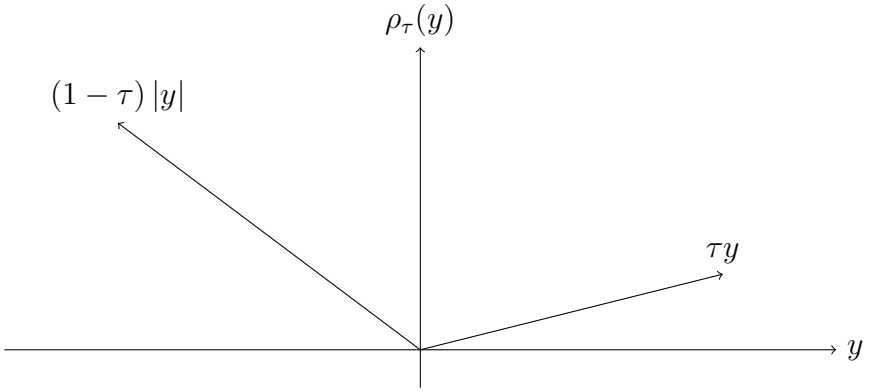


Abbildung 3.1: Verlustfunktion $\rho_\tau(y)$ für $\tau = 0.25$.

Gesucht sei nun jenes $\hat{y}$, dass den erwarteten Verlust minimiert:

$$\min_{\hat{y}} \mathbb{E}_{\rho_\tau} (Y - \hat{y}) . \tag{3.2}$$

Es gilt

$$\mathbb{E}_{\rho_\tau} (Y - \hat{y}) = (\tau - 1) \int_{-\infty}^{\hat{y}} (y - \hat{y}) \, \mathrm{d}F(y) + \tau \int_{\hat{y}}^{\infty} (y - \hat{y}) \, \mathrm{d}F(y). \tag{3.3}$$

Ableiten der Gleichung (3.3) bezüglich $\hat{y}$ unter der Annahme, dass Integration und Differentiation vertauschbar sind, und Nullsetzen ergibt

$$0 = (1 - \tau) \int_{-\infty}^{\hat{y}} \mathrm{d}F(y) - \tau \int_{\hat{y}}^{\infty} \mathrm{d}F(y) = F(\hat{y}) - \tau. \tag{3.4}$$

Da $F$ eine monotone Funktion ist, minimiert jedes Element $\{y : F(y) = \tau\}$ den erwarteten Verlust. Umformen von Gleichung (3.4) ergibt $\hat{y} = F^{-1}(\tau)$, sofern es eine eindeutige Lösung gibt. Andernfalls erhält man ein Intervall an $\tau$-Quantilen, von denen das kleinste Element gewählt wird. Dies nutzt die Tatsache aus, dass die Quantilsfunktion linksseitig stetig ist. Somit wurde gezeigt, dass durch das Optimierungsproblem (3.2) die Quantile einer Verteilung bestimmt werden können.

Oft ist die tatsächliche Verteilungsfunktion $F(y)$ einer Zufallsvariable $Y$ unbekannt. Es ist dann sinnvoll, zu der sogenannten *Empirischen Verteilungsfunktion* überzugehen.

*Definition* 3.2 (Empirische Verteilungsfunktion). Die *Empirische Verteilungsfunktion* ist definiert durch

$$F_n(y) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{(y_i \leq y)},$$

wobei $y_1, \ldots, y_n$ Beobachtungswerten der Zufallsvariable $Y$ entsprechen.

Durch Ersetzen der Verteilungsfunktion $F$ mit seinem empirischen Gegenpart folgt für den erwarteten Verlust:

$$\int_{-\infty}^{\infty} \rho_\tau \left(y - \hat{y}\right) \mathrm{d}F_n(y) = \frac{1}{n} \sum_{i=1}^{n} \rho_\tau(y_i - \hat{y}).$$

Durch Minimierung erhält man das $\tau$-Stichprobenquantil. Falls $\tau n$ ein ganzzahliger Wert ist, erhält man ein Intervall von Lösungen $\{y : F_n(y) = \tau\}$.

Das Problem, das $\tau$-Stichprobenquantil zu finden, kann dargestellt werden durch

$$\min_{q_\tau \in \mathbb{R}} \sum_{i=1}^{n} \rho_\tau \left(y_i - q_\tau\right). \tag{3.5}$$

Anders als bei der Kleinsten-Quadrate Methode der klassischen Regression ist es hier nicht möglich, den optimalen Schätzer durch Ableiten und Nullsetzen zu bestimmen. Dies kommt daher, dass die Zielfunktion in (3.5) an sehr vielen Stellen nicht differenzierbar ist. Das Problem kann jedoch als lineares Optimierungsproblem dargestellt und gelöst werden.

Sei $\mathbf{y}$ der Vektor der Responses. Zunächst werden die beiden Schlupfvariablen $\{u_i, v_i : i = 1, \ldots, n\}$ definiert. Diese stellen den positiven und negativen Teil des Residuenvektors $\mathbf{r} = \mathbf{y} - \mathbf{1}_n q_\tau$ dar, wobei $\mathbf{1}_n$ den $n$-dimensionalen Einsvektor $(1, \ldots, 1)^\top \in \mathbb{R}^n$ bezeichnet. Es gilt also $\mathbf{r} = \mathbf{u} - \mathbf{v}$, mit $\mathbf{u} = (u_1, \ldots, u_n)^\top$, $\mathbf{v} = (v_1, \ldots, v_n)^\top \in \mathbb{R}^n_+$. Mittels des Residuenvektors $\mathbf{r}$ lässt sich eine alternative Darstellung für den Responsevektor $\mathbf{y}$ bestimmen:

$$\mathbf{y} = \mathbf{r} + \mathbf{1}_n q_\tau = \mathbf{u} - \mathbf{v} + \mathbf{1}_n q_\tau.$$

Die Verlustfunktion $\rho_\tau$ an der Stelle $r_i$ mit $r_i = y_i - q_\tau$ ergibt

$$\begin{aligned}
\rho_\tau \left(r_i\right) &= r_i \left(\tau - \mathbb{1}_{(r_i < 0)}\right) \\
&= \left(u_i - v_i\right) \tau - \left(u_i - v_i\right) \mathbb{1}_{(u_i = 0)} \\
&= u_i \tau - v_i \tau + v_i \\
&= \tau u_i + (1 - \tau) v_i.
\end{aligned}$$

In der zweiten Zeile wurde die Tatsache ausgenutzt, dass das Residuum $r_i$ negativ und somit der positive Anteil $u_i$ null ist.

Bis jetzt war Y eine Zufallsvariable und damit galt für die Quantilsfunktion $Q_Y(\tau) = q_\tau$. Ab jetzt sollen die Quantile einem linearen (multiplen) Regressionsmodell folgen, das heißt für die Quantilsfunktion des Responsvektors $\mathbf{y}$ gilt $Q_{\mathbf{y}}(\tau \mid \mathbf{X}) = \mathbf{X}\boldsymbol{\beta}$. Die Matrix $\mathbf{X}$ entspricht der Designmatrix des Regressionsmodells; der Vektor $\boldsymbol{\beta}$ bezeichnet den Vektor der unbekannten Parameter.

Nun lässt sich das zu (3.5) äquivalente Lineare Programm formulieren (Koenker, 2005):

$$\min_{(q_\tau, \mathbf{u}, \mathbf{v}) \in \mathbb{R} \times \mathbb{R}_+^{2n}} \left\{ \tau \mathbf{1}_n^\top \mathbf{u} + (1-\tau)\mathbf{1}_n^\top \mathbf{v} \mid \mathbf{1}_n q_\tau + \mathbf{u} - \mathbf{v} = \mathbf{y} \right\}. \tag{3.6}$$

Sei die Quantilsfunktion der Stichprobe $\mathbf{y}$ gegeben durch $Q_{\mathbf{y}}(\tau \mid \mathbf{x}) = \mathbf{x}^\top \boldsymbol{\beta}(\tau)$. Um die Regressionsquantile berechnen zu können, wird das Lineare Programm dann folgendermaßen angepasst:

$$\min_{(\boldsymbol{\beta}, \mathbf{u}, \mathbf{v}) \in \mathbb{R}^p \times \mathbb{R}_+^{2n}} \left\{ \tau \mathbf{1}_n^\top \mathbf{u} + (1-\tau)\mathbf{1}_n^\top \mathbf{v} \mid \mathbf{X}\boldsymbol{\beta} + \mathbf{u} - \mathbf{v} = \mathbf{y} \right\}.$$

Die Matrix $\mathbf{X}$ bezeichnet die $(n \times p)$ Designmatrix der Regression. Der Residuenvektor besitzt nun die Form $\mathbf{r} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$. Die geschätzte Quantilsfunktion entspricht $Q_{\mathbf{y}}(\tau \mid \mathbf{X}) = \mathbf{X}\hat{\boldsymbol{\beta}}(\tau)$; der Vektor $\hat{\boldsymbol{\beta}}(\tau)$ ist der *quantilsspezifische Regressionskoeffizient* zum Niveau $\tau$.

Das betrachtete Lineare Programm besitzt eine polyedrische Restriktionsmenge, weshalb es sich mittels Simplex-Verfahren lösen lässt (Dantzig, 1966). Geometrisch betrachtet startet man bei einer beliebigen Ecke des Polyeders, um schlußendlich die optimale Ecke zu finden. Jede Ecke des Polyeders entspricht dabei einer *Basislösung*.

Sei $h \in \mathcal{H}$ eine $p$-elementige Teilmenge der $n$ Beobachtungen, $\bar{h}$ das Komplement von $h$ bezüglich $\mathcal{N}$, $\mathcal{N} = \{1, \ldots, n\}$ und $\mathbf{X}(h)$ die Untermatrix der Designmatrix mit den Zeilen $\{\mathbf{x}_i : i \in h\}$. Dann lässt sich der Begriff der Basislösung definieren.

*Definition* 3.3 (Basislösung). Jede *Basislösung*, die durch die Punkte $\{(x_i, y_i) : i \in h\}$ geht, hat die Form

$$\mathbf{b}(h) = \mathbf{X}(h)^{-1}\mathbf{y}(h).$$

Dabei wird voraussgesetzt, dass die Matrix $\mathbf{X}(h)$ nicht singulär ist. Es gibt $\binom{n}{p}$ solche Basislösungen.

Im Folgenden werden Optimalitätsbedingungen präsentiert, um die geschätzte

Quantilsfunktion für das Problem

$$\min_{\mathbf{b}\in\mathbb{R}^p} \sum_{i=1}^{n} \rho_\tau \left( y_i - \mathbf{x}_i^\top \mathbf{b} \right) \tag{3.7}$$

zu bestimmen. Dazu wird zunächst der Begriff der *allgemeinen Lage* definiert.

*Definition* 3.4. Die Beobachtungen $(\mathbf{y}, \mathbf{X})$ liegen in *allgemeiner Lage*, wenn für ein beliebiges $h \in \mathcal{H}$ gilt:

$$y_i - \mathbf{x}_i^\top \mathbf{b}(h) \neq 0 \quad \text{für ein beliebiges } i \notin h.$$

Besitzt die Zufallsvariable $\mathbf{y}$ eine Dichte bezüglich dem Lebesgue-Maß, so sind die Beobachtungen $(\mathbf{y}, \mathbf{X})$ mit Wahrscheinlichkeit 1 in allgemeiner Lage.

Mit dem Begriff der allgemeinen Lage lassen sich nun Voraussetzungen für die Existenz von (eindeutigen) Lösungen festsetzen.

*Satz* 3.1. Ist $(\mathbf{y}, \mathbf{X})$ in allgemeiner Lage, dann existiert eine Lösung zu dem QR-Problem (3.7) der Form $\mathbf{b}(h) = \mathbf{X}(h)^{-1}\mathbf{y}(h)$ genau dann, wenn für ein $h \in \mathcal{H}$ gilt:

$$-\tau \mathbf{1}_p \leq \xi(h) \leq (1-\tau)\, \mathbf{1}_p, \tag{3.8}$$

wobei $\boldsymbol{\xi}^\top(h) = \sum_{i\in\bar{h}} \psi_\tau \left( y_i - \mathbf{x}_i^\top \mathbf{b}(h) \right) \mathbf{x}_i^\top \mathbf{X}(h)^{-1}$ und $\psi_\tau(u) = \tau - \mathbb{1}_{(u<0)}$. Die Lösung $\mathbf{b}(h)$ ist *eindeutig* genau dann, wenn die Ungleichungen (3.8) strikt erfüllt werden. Andernfalls ist die Lösungsmenge eine konvexe Hülle, die mehrere Lösungen der Form $\mathbf{b}(h)$ beinhaltet.

Die Regressionsquantile besitzen sogenannte *Equivarianz-Eigenschaften*. Sei $\hat{\boldsymbol{\beta}}(\tau; \mathbf{y}, \mathbf{X})$ der durch die Beobachtungen $(\mathbf{y}, \mathbf{X})$ geschätzte quantilsspezifische Regressionskoeffizient zum Niveau $\tau$. Dann gilt folgender Satz (Koenker & Bassett, 1978):

*Satz* 3.2. Sei $\mathbf{A}$ eine nichtsinguläre $(p \times p)$ Matrix, $\boldsymbol{\gamma} \in \mathbb{R}^p$ und $a > 0$. Dann gilt für ein beliebiges $\tau \in (0, 1)$:

1. $\hat{\boldsymbol{\beta}}(\tau; a\mathbf{y}, \mathbf{X}) = a\hat{\boldsymbol{\beta}}(\tau; \mathbf{y}, \mathbf{X})$
2. $\hat{\boldsymbol{\beta}}(\tau; -a\mathbf{y}, \mathbf{X}) = -a\hat{\boldsymbol{\beta}}(1-\tau; \mathbf{y}, \mathbf{X})$
3. $\hat{\boldsymbol{\beta}}(\tau; \mathbf{y} + \mathbf{X}\boldsymbol{\gamma}, \mathbf{X}) = \hat{\boldsymbol{\beta}}(\tau; \mathbf{y}, \mathbf{X}) + \boldsymbol{\gamma}$
4. $\hat{\boldsymbol{\beta}}(\tau; \mathbf{y}, \mathbf{X}\mathbf{A}) = \mathbf{A}^{-1}\hat{\boldsymbol{\beta}}(\tau; \mathbf{y}, \mathbf{X})$.

Dies sind wertvolle Eigenschaften des Schätzers. Wenn man zum Beispiel die Temperatur einer Flüssigkeit modelliert, ermöglicht es unter der Annahme, dass die Modellmatrix $\mathbf{X}$ einen Intercept enthält, die Skala der Messungen von °Fahrenheit auf °Celsius zu

tansformieren. Dies würde die Schätzung $\hat{\boldsymbol{\beta}}(\tau; \mathbf{y}, \mathbf{X})$ zu $\frac{5}{9}\left(\hat{\boldsymbol{\beta}}(\tau; \mathbf{y}, \mathbf{X})\right) - 32\boldsymbol{\gamma})$ verändern, wobei $\boldsymbol{\gamma}$ hier dem Einheitsvektor $\mathbf{e}_1$ entspricht.

Quantilsfunktionen besitzen eine weitere Eigenschaft: Die *Equivarianz bezüglich monotoner Transformation*. Sei $h(\cdot)$ eine steigende (nichtfallende) Funktion auf $\mathbb{R}$. Dann gilt für eine beliebige Zufallsvariable $Y$:

$$Q_{h(Y)}(\tau) = h\left(Q_Y(\tau)\right).$$

Das bedeutet, dass die Quantile der transformierten Zufallsvariable $h(Y)$ den transformierten Quantilen der Zufallsvariable $Y$ entsprechen. Dies ist recht einfach zu zeigen: Sei $\min_y \mathbb{P}(Y \leq y) \geq \tau$, das heißt, $y$ ist das $\tau$-Quantil der Verteilung $F_Y$: $Q_Y(\tau) = y$. Für monotone Funktionen $h$ gilt

$$\min_y \mathbb{P}(Y \leq y) \geq \tau \Leftrightarrow \min_y \mathbb{P}(h(Y) \leq h(y)) \geq \tau.$$

Damit folgt, dass $h(y)$ das $\tau$-Quantil der Verteilung $F_{h(Y)}$ ist: $Q_{h(Y)}(\tau) = h(y)$. Es folgt somit $Q_{h(y)}(\tau) = h(y) = h\left(Q_Y(\tau)\right)$.

Diese angenehme Eigenschaft erscheint auf den ersten Blick selbstverständlich, für die Erwartungswertschätzer der Kleinsten-Quadrate Methoden gilt sie für beliebige monotone Funktionen jedoch nicht:

$$\mathbb{E}\left(h(Y)\right) \neq h\left(\mathbb{E}\left(Y\right)\right).$$

Der Grund dafür liegt in der Jensen'schen Ungleichung (Klenke, 2006):

*Satz* 3.3 (Jensen'sche Ungleichung). Sei $Y$ eine integrierbare Zufallsvariable und $h$ eine konvexe Funktion, dann gilt

$$h\left(\mathbb{E}\left(Y\right)\right) \leq \mathbb{E}\left(h(Y)\right).$$

Für eine konkave Funktion $h$ gilt

$$h\left(\mathbb{E}\left(Y\right)\right) \geq \mathbb{E}\left(h(Y)\right).$$

Ein weiterer angenehmer Aspekt der geschätzten Quantilsfunktion ist es, dass sie es ermöglicht, beliebige Teile der Verteilung von Y zu betrachten, ohne globale Verteilungsannahmen zu treffen. Es wird nur lokale Information nahe dem zu betrachtenden Quantil verwendet. Diese unabhängige Schätzung der unabhängigen Quantile kann jedoch manchmal auch zu schwerwiegenden Problemen führen, nämlich wenn sich die geschätzten Quantile kreuzen. Dieses sogenannte *Quantile Crossing* würde einer der grundlegendsten Annahmen einer Verteilungsfunktion widersprechen, dass ihre Inverse monoton steigend ist.

Erfreulicherweise lässt es sich zeigen, dass dieses Kreuzen von Quantilsschätzungen nur für sehr extreme Werte des Designraumes auftreten kann. Im Zentrum $\bar{\mathbf{x}}$ ist die Funktion $\hat{Q}_{\mathbf{y}}(\tau \mid \bar{\mathbf{x}}) = \bar{\mathbf{x}}^\top \hat{\boldsymbol{\beta}}(\tau)$ monoton in $\tau$ (Koenker, 2005).

*Satz* 3.4. Die Pfade von $\hat{Q}_{\mathbf{y}}(\tau \mid \bar{\mathbf{x}})$ sind mononton steigend für $\tau \in (0,1)$.

*Beweis.* Man zeigt

$$\tau_1 < \tau_2 \Rightarrow \bar{\mathbf{x}}^\top \hat{\boldsymbol{\beta}}(\tau_1) \leq \bar{\mathbf{x}}^\top \hat{\boldsymbol{\beta}}(\tau_2).$$

Es gilt für ein $\mathbf{b} \in \mathbb{R}^p$:

$$\begin{aligned}
&\rho_{\tau_2}\left(y_i - \mathbf{x}_i^\top \mathbf{b}\right) - \rho_{\tau_1}\left(y_i - \mathbf{x}_i^\top \mathbf{b}\right) \\
=&\left(y_i - \mathbf{x}_i^\top \mathbf{b}\right)\left(\tau_2 - \mathbb{1}_{\left(y_i - \mathbf{x}_i^\top \mathbf{b} < 0\right)}\right) - \left(y_i - \mathbf{x}_i^\top \mathbf{b}\right)\left(\tau_1 - \mathbb{1}_{\left(y_i - \mathbf{x}_i^\top \mathbf{b} < 0\right)}\right) \\
=&\left(y_i - \mathbf{x}_i^\top \mathbf{b}\right)\left(\tau_2 - \tau_1 - \mathbb{1}_{\left(y_i - \mathbf{x}_i^\top \mathbf{b} < 0\right)} + \mathbb{1}_{\left(y_i - \mathbf{x}_i^\top \mathbf{b} < 0\right)}\right) \\
=&\left(y_i - \mathbf{x}_i^\top \mathbf{b}\right)(\tau_2 - \tau_1).
\end{aligned}$$

Damit folgt

$$\begin{aligned}
&\sum_{i=1}^{n}\left[\rho_{\tau_2}\left(y_i - \mathbf{x}_i^\top \mathbf{b}\right) - \rho_{\tau_1}\left(y_i - \mathbf{x}_i^\top \mathbf{b}\right)\right] \\
=&\sum_{i=1}^{n}\left(y_i - \mathbf{x}_i^\top \mathbf{b}\right)(\tau_2 - \tau_1) = (\tau_2 - \tau_1)\sum_{i=1}^{n}\left(y_i - \mathbf{x}_i^\top \mathbf{b}\right) \\
=&n(\tau_2 - \tau_1)\left(\bar{\mathbf{y}} - \bar{\mathbf{x}}^\top \mathbf{b}\right).
\end{aligned} \tag{3.9}$$

Nun gilt für $\mathbf{b} = \hat{\boldsymbol{\beta}}(\tau_k)$, $k = 1, 2$:

$$\begin{aligned}
&\sum_{i=1}^{n}\rho_{\tau_1}\left(y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}(\tau_1)\right) + n(\tau_2 - \tau_1)\left(\bar{\mathbf{y}} - \bar{\mathbf{x}}^\top \hat{\boldsymbol{\beta}}(\tau_2)\right) \\
\overset{(*)}{\leq}&\sum_{i=1}^{n}\rho_{\tau_1}\left(y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}(\tau_2)\right) + n(\tau_2 - \tau_1)\left(\bar{\mathbf{y}} - \bar{\mathbf{x}}^\top \hat{\boldsymbol{\beta}}(\tau_2)\right) \\
=&\sum_{i=1}^{n}\rho_{\tau_2}\left(y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}(\tau_2)\right).
\end{aligned}$$

Die letzte Gleichung folgt durch Umformung von (3.9). Die Ungleichung $(*)$ folgt aus der Tatsache, dass $\hat{\boldsymbol{\beta}}(\tau_1)$ optimal für $\rho_{\tau_1}$ ist. Wird $\hat{\boldsymbol{\beta}}(\tau_1)$ also durch $\hat{\boldsymbol{\beta}}(\tau_2)$ ersetzt, vergrößert sich die Summe.

Es folgt weiters

$$\begin{aligned}\sum_{i=1}^{n} \rho_{\tau_2}\left(y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}\left(\tau_2\right)\right) &\leq \sum_{i=1}^{n} \rho_{\tau_2}\left(y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}\left(\tau_1\right)\right) \\ &= \sum_{i=1}^{n} \rho_{\tau_1}\left(y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}\left(\tau_1\right)\right) + n\left(\tau_2 - \tau_1\right)\left(\bar{\mathbf{y}} - \bar{\mathbf{x}}^\top \hat{\boldsymbol{\beta}}\left(\tau_1\right)\right).\end{aligned}$$

Zusammenfassend folgt

$$\begin{aligned}&\sum_{i=1}^{n} \rho_{\tau_1}\left(y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}\left(\tau_1\right)\right) + n\left(\tau_2 - \tau_1\right)\left(\bar{\mathbf{y}} - \bar{\mathbf{x}}^\top \hat{\boldsymbol{\beta}}\left(\tau_2\right)\right) \\ \leq &\sum_{i=1}^{n} \rho_{\tau_1}\left(y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}\left(\tau_1\right)\right) + n\left(\tau_2 - \tau_1\right)\left(\bar{\mathbf{y}} - \bar{\mathbf{x}}^\top \hat{\boldsymbol{\beta}}\left(\tau_1\right)\right) \\ \Leftrightarrow &n\left(\tau_2 - \tau_1\right)\left(\bar{\mathbf{x}}^\top \hat{\boldsymbol{\beta}}\left(\tau_2\right) - \bar{\mathbf{x}}^\top \hat{\boldsymbol{\beta}}\left(\tau_1\right)\right) \geq 0.\end{aligned}$$

Wegen $\tau_2 - \tau_1 > 0$ folgt $\left(\bar{\mathbf{x}}^\top \hat{\boldsymbol{\beta}}\left(\tau_2\right) - \bar{\mathbf{x}}^\top \hat{\boldsymbol{\beta}}\left(\tau_1\right)\right) \geq 0$ und damit $\bar{\mathbf{x}}^\top \hat{\boldsymbol{\beta}}\left(\tau_2\right) \geq \bar{\mathbf{x}}^\top \hat{\boldsymbol{\beta}}\left(\tau_1\right)$. □

Natürlich folgt aus der Monotonie von $\hat{Q}_{\mathbf{y}}\left(\tau \mid \mathbf{x}\right)$ an der Stelle $\mathbf{x} = \bar{\mathbf{x}}$ nicht die Monotonie der Funktion bei anderen $x$-Werten, jedoch ist es offensichtlich, dass bei Linearität von $\hat{Q}_{\mathbf{y}}$ in den Variablen $\mathbf{x}$ ein mögliches Kreuzen der Quantile hinreichend weit entfernt von $\bar{\mathbf{x}}$ sein muss.

## 3.1 Nichtparametrische Quantile Regression

Eine Möglichkeit, die Flexibilität der Schätzung zu verbessern, ist die Anwendung von nichtparametrischen Modellen. Es werden zwei Möglichkeiten vorgestellt, nichtparametrische Modelle mittels Quantile Regression zu schätzen (Koenker, 2005):

- Anwendung von B-Splines für QR-Modelle
- Anwendung von Straftermen zur Bildung von QR-Modellen.

Sei die Stichprobe $\{(x_i, y_i), i = 1, \ldots, n\}$ gegeben. An der Stelle $x$ sei eine Schätzung für die $\tau$-Quantilsfunktion der Response $y$ gesucht:

$$g(x) = Q_y\left(\tau \mid x\right).$$

Die Anwendung der B-Splines ist eine sehr einfache Methode, um nichtparametrische Regressionsquantile zu berechnen. Durch Darstellung der gesuchten Funktion $g(x)$ mittels B-Splines der Ordnung $(m+1)$ folgt

$$g(x) = \sum_{j=1}^{k} B_j^m(x)\beta_j.$$

Somit hat die zu minimierende Zielfunktion die Form

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho_\tau \left( y_i - \sum_{j=1}^{k} B_j^m(x)\beta_j \right).$$

Durch Minimierung bezüglich $\boldsymbol{\beta}$ erhält man einen Schätzer der Funktion $g(x)$ zum Niveau $\tau$.

Die Implementation dieser Modellklasse in `R` lässt sich mittels der Funktion `rq` des Pakets `quantreg` durchführen (Koenker, 2005).

Berechnung nichtparametrischer QR-Modelle

```
1 library(quantreg)
2 library(splines)
3 rq(y~bs(x,knots = k))
```

Um mittels der Funktion `bs()` B-Splines zu berechnen, muss das Paket `splines` (Bates & Venables, 2013) geladen werden. Durch den Befehl `degree` wird der Grad der stückweisen Polynome bestimmt. Standardmäßig ist dieser drei, wodurch kubische Splines erzeugt werden. Der Parameter `knots` bestimmt die Anzahl der Knotenpunkte, die die B-Spline Basis definieren. Alternativ lassen sich die Knotenpunkte auch durch den Freiheitsgrad `df` festlegen. Dann ergeben sich die Knotenpunkte automatisch durch `df` − `degree`.

Für die Funktion `rq()` wird zunächst die zu schätzende Regressionsformel definiert. Mittels dem Parameter `tau` wird das gesuchte Quantilsniveau festgesetzt. Der Default-Wert ist hier `tau` = 0.5, wodurch der Median der Verteilungsfunktion von $y$ geschätzt wird.

Ein wesentlicher Vorteil der Anwendung der B-Splines ist die Tatsache, dass das Modell immer noch linear in den Parametern ist, und somit relativ einfach berechnet werden kann. Durch den Befehl `method` wird der Algorithmus festgelegt, mit dem die Quantile berechnet werden. Der Default-Wert ist `method="br"`, wodurch eine modifizierte Version des Barrodale-Roberts Algorithmus verwendet wird (Koenker & D'Orey, 1987).

Der BR-Algorithmus (Barrodale & Roberts, 1973) ist eine Spezialform des Simplex-Algorithmus zur Lösung des *Least Absolute Deviation* (LAD) Problems

$$\min_{f} \sum_{i=1}^{n} |y_i - f(x_i)|.$$

Eine Modifizierung des Algorithmus' ist sehr effizient zur Berechnung der Regressionsquantile für Probleme von einer Größe bis zu einigen tausend Beobachtungen. Durch Anwendung dieser Methode werden innerhalb der Funktion `rq()` Konfidenz-

intervalle der geschätzten Parameter mittels der von Koenker (2005) beschriebenen Rang-Inversions-Methode berechnet.

Für sehr große Probleme ist das Innere-Punkte-Verfahren von Frisch-Newton geeigneter (Fiacco & McCormick, 1968). Es lässt sich durch den Befehl `method="fn"` anwenden. Im Gegensatz zum Simplex-Verfahren, das die optimale Lösung eines linearen Optimierungsproblems in den Ecken sucht, nähert sich das Innere-Punkte-Verfahren vereinfacht ausgedrückt der Optimallösung per Newton-Richtung durch das Innere des Polyeders. Da im Laufe dieser Arbeit lediglich der BR-Algorithmus verwendet wird, wird auf eine nähere Betrachtung des FN-Algorithmus verzichtet.

Für extrem große Probleme lässt sich eine erweiterte Version des FN-Algorithmus `method="pfn"` anwenden.

Es wird nun ein Beispiel der Anwendung nichtparametrischer Regressionsquantile mittels B-Splines dargestellt. Koenker (2005) demonstriert in seinem Buch ein Beispiel anhand des Datensatzes `mcycle`. Dabei handelt es sich um Messungen eines simulierten Motorradunfalls, bei dem die Beschleunigung der Helme der verwendeten Crashtest-Dummies betrachtet wurde. Der Datensatz beinhaltet die Variablen `times` (gemessen in Millisekunden nach dem Aufprall) und `accel` (Beschleunigung, gemessen in $g \approx 9.81 m/s^2$).

Berechnung nichtparametrischer QR-Modelle für den Datensatz mcycle

```
1  library(MASS)
2  data(mcycle)
3  attach(mcycle)
4  plot(times, accel, type="n")
5  points(times, accel)
6  X <- model.matrix(accel~bs(times, df=15))
7  for(tau in 1:3/4)
8  {
9    fit <- rq(accel~bs(times, df=15), tau=tau, data=mcycle)
10   accel.fit <- X %*% fit$coef
11   lines(times, accel.fit)
12 }
```

Abbildung 3.2 zeigt die geschätzten Quantilsfunktionen und den Scatterplot des Datensatzes. Es wurden 12 Knotenpunkte gewählt, die auf Quantile der $x$-Werte verteilt werden.

Für die ersten Millisekunden nach dem Aufprall erkennt man kaum Variabilität, die jedoch mit zunehmender Dauer ansteigt, erkennbar an den größer werdenden Niveauunterschieden der jeweiligen Quantilsschätzungen. Bis zu einem Zeitpunkt von 50 Millisekunden nach dem Aufprall zeigt das Modell für alle betrachteten Quantile gute Schätzungen. Danach werden die Daten sehr dünn und die geschätzten Quantile zeigen zum Schluß gar keinen Unterschied mehr, da nur mehr eine Beobachtung vorhanden ist.

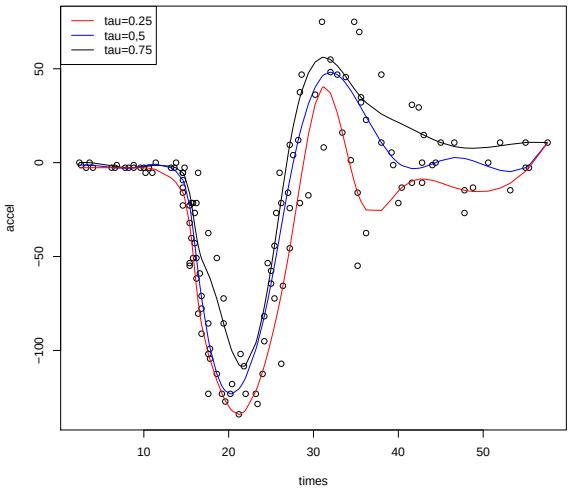


Abbildung 3.2: Nichtparametrisches QR-Modell mittels B-Splines für den Datensatz `mcycle`.

Eine weitere Möglichkeit der Anwendung von B-Splines zum Erstellen von nichtparametrischen QR-Modellen ist das Nutzen der Funktion `cobs()` (*CO*nstrained *B*-*S*plines Nonparametric Regression Quantiles) aus dem gleichnamigen `R`-Paket `cobs` (He & Ng, 2011). Mit dieser Funktion lassen sich „bedingte“ Quantilskurven durch die Anwendung von linearen oder quadratischen Splines erzeugen. Durch den Befehl `constraint` werden zusätzliche Bedingungen an die zu schätzende Kurve gestellt:

- wachsend (”`increase`”)
- fallend (”`decrease`”)
- konvex (”`convex`”)
- konkav (”`concave`”)
- periodisch (”`periodic`”) oder
- keine zusätzliche Bedingung (”`none`”).

Durch den Befehl `degree` wird der Grad der Splines bestimmt: 1 erzeugt lineare Splines, 2 quadratische Splines. Der Parameter `lambda` entspricht dem Strafterm des Modells. Für $\lambda = 0$ wird kein Strafterm betrachtet, wodurch unbestrafte Regressions-B-Splines erzeugt werden. Für $\lambda > 0$ werden Glättungs-B-Splines mit dem gegebenen Strafterm $\lambda$ berechnet. Wird jedoch ein Wert $\lambda < 0$ eingegeben, wird der optimale Strafterm mittels einem modifizierten AIC bestimmt. Alternativ lässt sich dies auch durch ein modifiziertes Schwarzsches Informationskriterium (SIC) oder einem Bayesschen Informationskriterium bestimmen. Bis zur Version 1.1-6 des Pakets `cobs` wurde

standardmäßig das SIC verwendet, für alle nachfolgenden Versionen ist AIC das Default-Kriterium. Die aktuellste Version ist 1.2-2 aus dem Jahr 2011.

Nun wird auch die Funktion `cobs` für den Datensatz `mcycle` angewendet. Dabei soll der optimale Strafterm $\lambda$ automatisch bestimmt werden. Um eine akzeptable Schätzung der Daten zu erreichen, muss der Bereich der betrachteten $\lambda$-Werte adaptiert werden. Mittels `lambda.lo` und `lambda.hi` werden die untere beziehungsweise obere Schranke für die Gittersuche nach dem optimalen $\lambda$ festgelegt. Mittels `lambda.length` wird die Anzahl der betrachteten Punkte in der Gittersuche bestimmt. Diese Parameter können bei Erhalten eines nicht-zufriedenstellenden Ergebnisses angepasst werden.

Berechnung nichtparametrischer QR-Modelle für den Datensatz mcycle

```
1 library(MASS)
2 data(mcycle)
3 attach(mcycle)
4 library(cobs)
5 fit0.75 <- cobs(times, accel, nknots=100, lambda=-1,lambda.lo=0.01,
6                 lambda.hi=10,lambda.length=200, tau=0.75,
7                 constraint="none")
```

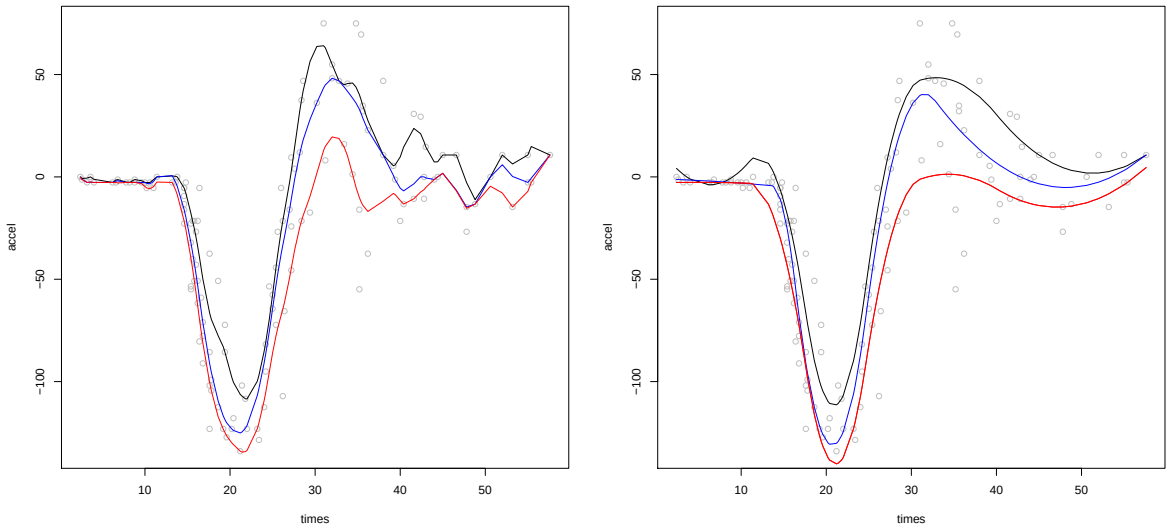


Abbildung 3.3: Nichtparametrisches QR-Modell durch die Funktion `cobs()` für den Datensatz `mcycle` zu den Niveaus $\tau \in \left\{\frac{1}{4}, \frac{1}{2}, \frac{3}{4}\right\}$. Links: Bestimmung des optimalen Strafterms $\lambda$ bezüglich AIC-Wert . Rechts: Wahl des Parameters $\lambda = 0$, d.h. Modellierung ohne Strafterm.

Abbildung 3.3 zeigt das geschätzte Modell durch Ermittlung des optimalen Strafterms $\lambda$ (links), beziehungsweise ohne Strafterm (rechts). Durch Ignorieren eines Strafterms

erhält man eine deutlich glattere Schätzung, jedoch ist Quantile Crossing bei `times` $\approx 10$ erkennbar. Durch Bestimmung des optimalen Strafterms ist zwar kein Kreuzen der Quantilsschätzungen sichtbar, jedoch vereinen sich die unterschiedlichen Schätzungen in einigen Punkten und es wird allgemein eine rauere Schätzung abgebildet. Man erkennt in Abbildung 3.3 in der linken Graphik einen deutlichen Unterschied der unbestraften Schätzung des nichtparametrischen QR-Modells im Vergleich zu dem nichtparametrischen QR-Modell durch Anwendung der B-Splines in Abbildung 3.2. Der Grund dafür ist, dass die Funktion `cobs()` einen maximalen Grad von 2 zulässt, wobei zuvor kubische B-Splines angewendet wurden.

Nun werden nichtparametrische QR-Modelle durch *Anwendung von Straftermen* betrachtet. Dabei wird der Fokus auf den univariaten Fall gelegt.

Für eine alternative Betrachtung nichtparametrischer Quantilsfunktionen minimiert man folgenden Term:

$$\min_{g \in \mathcal{G}} \sum_{i=1}^{n} \rho_\tau \left(y_i - g(x_i)\right) - \lambda \int \left(g''(x)\right)^2 \mathrm{d}x.$$

Der Raum $\mathcal{G}$ bezeichnet den Sobolev Raum der Funktionen $g(\cdot)$ mit quadratisch integrierbarer zweiter Ableitung. Mittels dem Parameter $\lambda$ wird eine zu starke Schwankung der Funktion $g(\cdot)$ bestraft. Je größer $\lambda$ ist, desto mehr nähert sich die geschätzte Funktion $g(\cdot)$ einer Gerade an. Durch die zusätzliche Bedingung, dass die vierte Ableitung von $g(\cdot)$ null ist fast überall, erhält man stückweise kubische Polynome. Bosch et al. (1995) beschreiben ein Inneres-Punkte-Verfahren zur Lösung dieses Problems.

Eine weitere wichtige Gruppe von Straftermen ist jene, die man durch Anwendung der *Totalen Variation* erhält.

*Definition* 3.5. Die *Totale Variation* einer Funktion $f : [a, b] \to \mathbb{R}$ ist definiert als

$$V(f) = \sup_{a \leq x_1 \leq \ldots \leq x_n \leq b} \sum_{i=1}^{n} \left| f(x_{i+1}) - f(x_i) \right|.$$

Wird kein solches reelle Supremum gefunden, wird die Totale Variation unendlich gesetzt.

Man kann zeigen, dass für eine Funktion $g(\cdot)$, deren erste Ableitung absolut stetig ist, gilt (Natanson, 1974):

$$V\left(g'\right) = \int_a^b \left|g''\right| \mathrm{d}x.$$

Gesucht sei nun eine stetige Funktion $g(\cdot)$, die die Funktion $V\left(g'\right)$ minimiert und in den Punkten $\{(x_i, y_i) : i = 1, \ldots, n\}$ interpoliert.

Man kann zeigen, dass so eine Lösung durch Minimierung des Problems

$$\min_g \sum_{i=1}^{n} \rho_\tau \left( y_i - g(x_i) \right) - \lambda V(g')$$

für Funktionen $g : [a, b] \to \mathbb{R}$ mit absolut stetiger erster Ableitung gefunden wird. Die Lösungen sind stückweise lineare Funktionen mit Knoten in den Punkten $x_i$.

Zur Bestimmung des optimalen Parameters $\lambda$ schlagen Koenker et al. (1994) eine abgewandelte Version des Schwarz-Kriteriums vor. Der Parameter $\lambda$ wird durch Minimierung von

$$\mathrm{SIC}(\lambda) = \log \left( \frac{1}{n} \sum \rho_\tau (y_i - \hat{g}_\lambda(x_i)) \right) + \frac{1}{2n} p_\lambda \log n$$

bestimmt. Der Wert $p_\lambda$ bezeichnet die Anzahl der Punkte, die durch die geschätzte Funktion interpoliert werden.

Mittels der Funktion `rqss()` aus dem Paket `quantreg` lassen sich solche QR-Modelle schätzen. Hier wird standardmäßig der Strafterm der Totalen Variation angewendet.

Nun wird diese Modellklasse anhand eines kleinen Beispiels demonstriert (Koenker, 2005). Dazu wird der Datensatz `Mammals` aus dem Paket `quantreg` betrachtet. Er beinhaltet 107 Beobachtungen über die maximale Geschwindigkeit von Säugetieren. Für die unterschiedlichen Beobachtungen existieren die Variablen `weight` (durchschnittliches Gewicht in kg der Spezies), `speed` (maximal gemessene Geschwindigkeit), `hoppers` (eine logische Variable, die „hüpfende“ Spezien betrachtet, z.B. Kängurus) und `specials` (eine logische Variable für Spezien, deren Geschwindigkeit nicht für ihr Überleben wichtig ist, z.B. Nilpferd, Mensch).

Berechnung nichtparametrischer QR-Modelle für den Datensatz Mammals

```
1  library(quantreg)
2  library(MatrixModels)
3
4  data(Mammals)
5  attach(Mammals)
6  x <- log(weight)
7  y <- log(speed)
8  plot(x, y, type = "n")
9  points(x[hoppers], y[hoppers], pch = "h", col = "red")
10 points(x[specials], y[specials], pch = "s", col = "blue")
11 others <- (!hoppers & !specials)
12 points(x[others], y[others], col = "black", cex = 0.75)
13 fit <- rqss(y ~ qss(x, lambda = 1), tau = 0.9)
14
15 daten= data.frame(log.speed= fitted(fit), log.weight= x)
```

```
daten= daten[order(daten$log.weight), ]

lines(unique(daten$log.weight), unique(daten$log.speed))
```

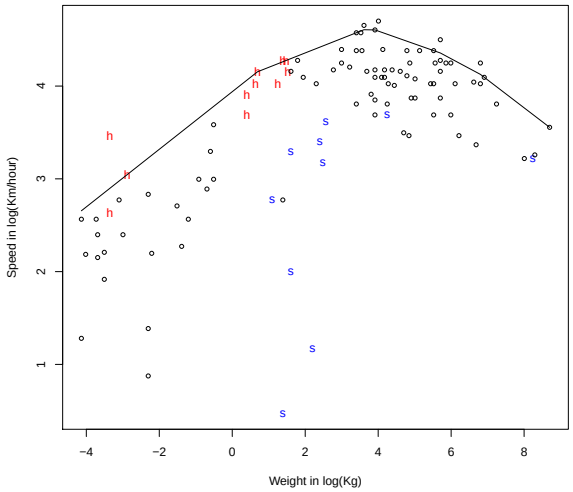


Abbildung 3.4: Nichtparametrisches QR-Modell mittels Straftermen Totaler Variation für den Datensatz `Mammals`.

Zunächst wird Gewicht und Geschwindigkeit der Beobachtungen logarithmiert. Zur Illustration werden die Datenpunkte der `hoppers` und der `specials` extra hervorgehoben. Es ist wenig überraschend, dass „Hüpfer“ tendenziell leichter und schneller sind als jene Säugetiere, die nicht von ihrer Geschwindigkeit profitieren. Mittels der Funktion `qss()` wird der nichtparametrische Teil und der Strafterm des Modells definiert. Damit dem geschätzten Quantil eine gewisse Glattheit erhalten bleibt, wird der Glättungsparameter $\lambda = 1$ gesetzt. Abbildung 3.4 zeigt das geschätzte 90%-Quantil unter Anwendung nichtparametrischer QR-Methoden für einen Strafterm der Totalen Variation. Es lässt sich ein guter Fit der Daten erkennen, jedoch ist der stückweise lineare Verlauf der geschätzten Kurve sichtbar, der nicht für alle Datenstrukturen vorteilhaft ist.

# 4 Analyse von Leistungen einer Krankenversicherung

Die zugrunde liegenden Daten beschreiben die Klientel einer Versicherung. Es befinden sich 184930 Beobachtungen im Datensatz, die durch die folgenden Variablen beschrieben werden:

- *Leistung*: Zu erbringende Versicherungsleistung an die versicherte Person
- *LNR*: Leistungsnummer zur Identifizierung der Personen, die die unterschiedlichen Leistungen konsumieren
- *Jahr*: Betrachtetes Kalenderjahr (2010, 2011, 2012), in dem der Versicherungsfall auftritt
- *TKZ*: Tarifklassen innerhalb des Portfolios:
  - 1: Ambulante Kosten
  - 2: Spitalskosten
- *Alter*: Alter der versicherten Person in Jahren
- *Geschlecht*: Geschlecht der versicherten Person (M,W)
- *LKZ*: Zu erbringende Leistungsarten durch den Versicherer. Es wird zwischen 32 Leistungsarten unterschieden.

Nachdem die Daten in `R` eingelesen werden, lassen sie sich mittels dem Befehl `summary()` analysieren. Der Datensatz beinhaltet Versicherungsdaten 81395 männlicher und 103535 weiblicher Personen im Alter zwischen null und 106 Jahren für die aufeinanderfolgenden Kalenderjahre 2010, 2011 und 2012.

Wird die Variable `Leistung` sämtlicher Versicherungsnehmer näher betrachtet, erkennt man, dass einige wenige Versicherte mit sehr hohen Leistungsanforderungen den Großteil des finanziellen Schadens ausmachen, während die Mehrheit der betrachteten Leistungen null ist, erkennbar durch den Median der Daten.

```
Min.  1st Qu. Median  Mean   3rd Qu.  Max.
0.0     0.0     0.0   589.2  216.2   47540.0
```

Ziel ist es, die Leistungen der Versicherten passend zu modellieren. Dabei sollen die Daten bezüglich der Tarifklassen, Geschlechter und Kalenderjahre kategorisiert werden.

| | Männer | | | |
|---|---|---|---|---|
| TKZ | 2010 | 2011 | 2012 | Max. Alter |
| 1 | 5522 | 8374 | 11531 | 70 |
| 2 | 18691 | 18668 | 18609 | 104 |

| | Frauen | | | |
|---|---|---|---|---|
| TKZ | 2010 | 2011 | 2012 | Max. Alter |
| 1 | 6552 | 9869 | 12987 | 70 |
| 2 | 24933 | 24756 | 24438 | 106 |

Tabelle 4.1: Anzahl versicherter Personen, aufgeteilt bezüglich Geschlecht, Tarifklasse und Jahr.

Zur Modellierung der Daten werden unterschiedliche Methoden angewendet.

Die Anzahl der versicherten Personen für die jeweilige Unterteilung ist in Tabelle 4.1 ersichtlich. Es treten deutlich häufiger Aufwendungen in Tarifklasse 2 auf. Geschlechterspezifisch gesehen treten über alle Kategorien häufiger Kosten für weibliche Versicherte auf.

Die Anzahl der Datenpunkte in den unterschiedlichen Kategorien ist jedoch immer noch sehr hoch. Dies wird anhand des Scatterplots der 5522 männlichen Leistungen der Tarifklasse 1 für das Kalenderjahr 2010 in Abbildung 4.1 illustriert.

Aus diesem Grund wird mittels der Funktion `tapply()` eine Mittelung der Daten durchgeführt.

Mittelung der Daten
```
1 leistung.m <- tapply(Leistung, Alter, mean)
```

Die Funktion `tapply()` wendet hier den Befehl `mean()` auf die Leistungen bezüglich dem Prädiktor `Alter` an, das heißt, für jede Altersstufe wird der jeweilige Mittelwert generiert.

Nun soll ein Scatterplot der gemittelten Leistungen pro Altersklasse erzeugt werden. Dazu müssen zunächst die Altersstufen berechnet werden.

Ausgabe der Altersstufen
```
1 alter.m <- unique(sort(Alter))
```

Durch die Funktion `sort()` werden die Altersstufen der vorhandenen Klienten aufsteigend sortiert. Die Funktion `unique()` eliminiert mehrmals vorkommende Werte.

Abbildung 4.1 zeigt die Scatterplots der Portfolios der männlichen Leistungen beider Tarifklassen für das Kalenderjahr 2010. Es befinden sich Beobachtungen versicherter

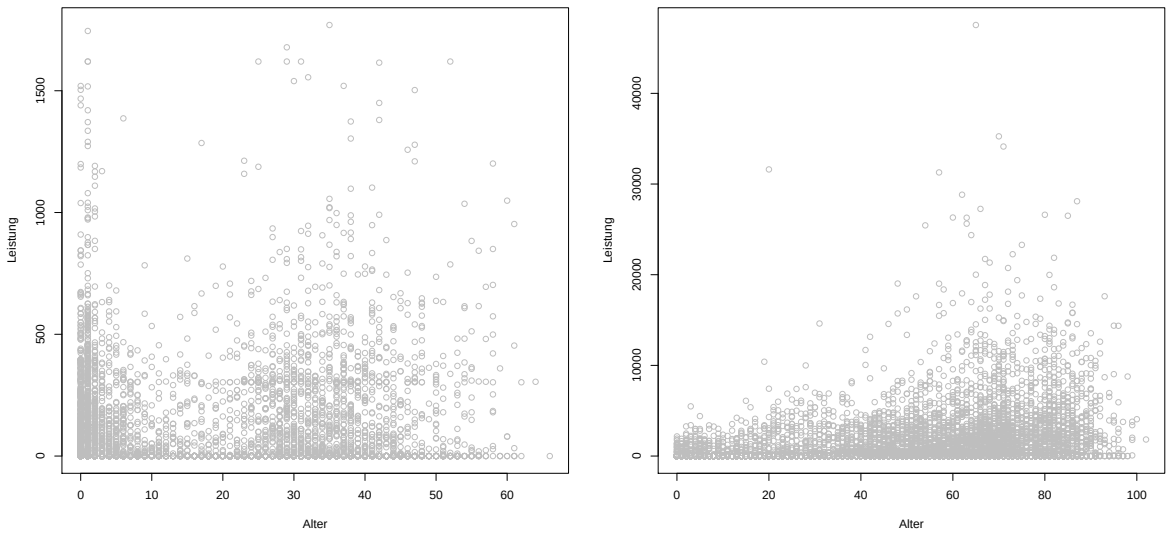


Abbildung 4.1: Scatterplot der männlichen Leistungen der Tarifklassen 1 (links) und 2 (rechts) für das Kalenderjahr 2010.

Personen im Alter von null bis 66 Jahren in der Tarifklasse 1, wobei keine 63- und 65-Jährigen vorhanden sind. In der Tarifklasse 2 sind Versicherungsnehmer im Alter von null bis 102 Jahren vorhanden.

Zunächst werden die gemittelten männlichen Leistungen beider Kategorien betrachtet. Abbildung 4.2 zeigt die Leistungen der männlichen Versicherten für das Kalenderjahr 2010. Es lässt sich allgemein ein höheres Niveau der Zahlungen in der Tarifklasse 2 feststellen, erkennbar an der Skala der y-Achse. Für TKZ=2 sind viel höhere Altersstufen vertreten, als dies in Tarifklasse 1 der Fall ist.

Für die gemittelten Leistungen der weiblichen Versicherten zeigt Abbildung 4.3 ein ähnliches Bild. Die Tarifklasse 2 weist ein viel höheres Niveau auf. Es ist für TKZ=2 deutlich das Phänomen des „Geburten-Hügels“ feststellbar: Da im Altersbereich von 20-40 Jahren vermehrt Schwangerschaften auftreten, steigt auch der Leistungsanspruch der Frauen in diesem Bereich infolge der anfallenden Geburtskosten deutlich an. Sowohl für die Mittelwerte der weiblichen als auch der männlichen Daten ist eine „Trichterform“ erkennbar, das heißt, die Schwankungsbreite steigt mit dem Alter der Versicherten. Das Verhalten der Leistungen in der Tarifklasse 1 ist für männliche und weibliche Versicherte sehr ähnlich, auch bezüglich der Höhe der anfallenden Kosten.

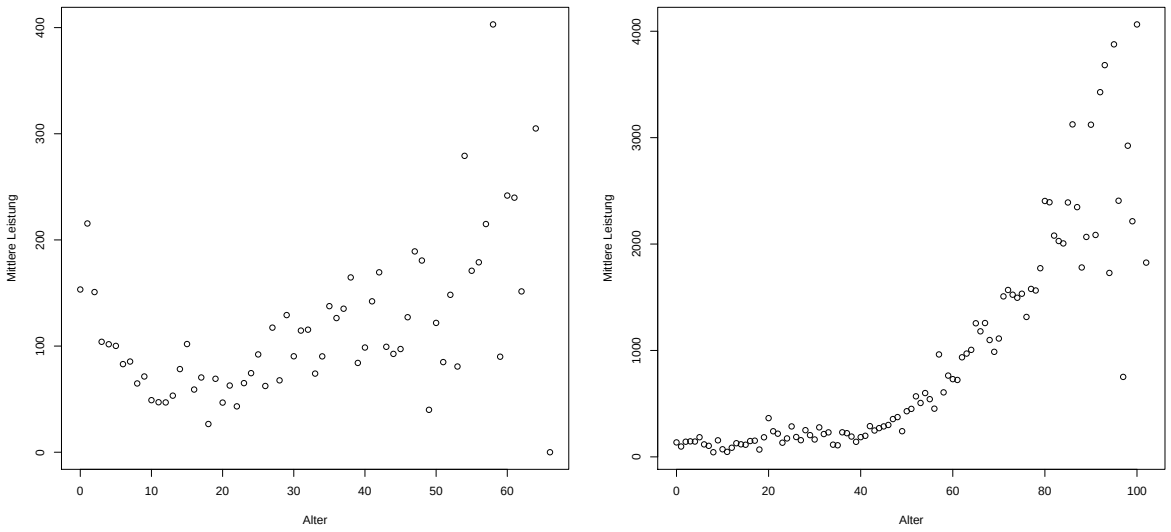


Abbildung 4.2: Gemittelte Leistungen der männlichen Versicherten des Kalenderjahres 2010 für die Tarifklassen 1 (links) und 2 (rechts).

Diese Eigenschaften lassen sich auch für die anderen Kalenderjahre feststellen.

Da eine Behandlung jeder einzelnen Kategorie auf Kosten der Übersicht gehen würde, wird der Fokus auf zwei Kategorien gesetzt: Die männlichen Leistungen des Jahres 2010 für die Tarifklasse 1 und die weiblichen Daten desselben Jahres für die Tarifklasse 2.

Zunächst wird die Dichte der Daten betrachtet: Es wird analysiert, wieviele Personen in der jeweiligen Altersstufe vorhanden sind. Werden die Daten zu dünn, werden sie ab einem gewissen Alter abgeschnitten, um eine sinnvolle Modellierung zu erhalten.

Ermittlung der Dichte der Daten

```
1 tapply(Leistung, Alter, length)
```

Der männliche Datensatz der Tarifklasse 1 für das Jahr 2010 wird ab dem Alter von 60 Jahren abgeschnitten, da hier nur mehr wenige Datenpunkte existieren, was auch in Abbildung 4.1 ersichtlich ist. Für die weiblichen Leistungen der Tarifklasse 2 desselben Jahres werden Leistungen bis zu einem Alter von 85 Jahren betrachtet.

Als nächstes werden die aufgetreten geschlechterspezifischen Leistungsarten in der jeweiligen Tarifklasse analysiert.

Die drei am häufigsten vorkommenden Leistungsarten in der Tarifklasse 1 für das Kalenderjahr 2010 in Tabelle 4.2 sind 39 (Arzt- und Facharztkosten), 17 (Kosten für

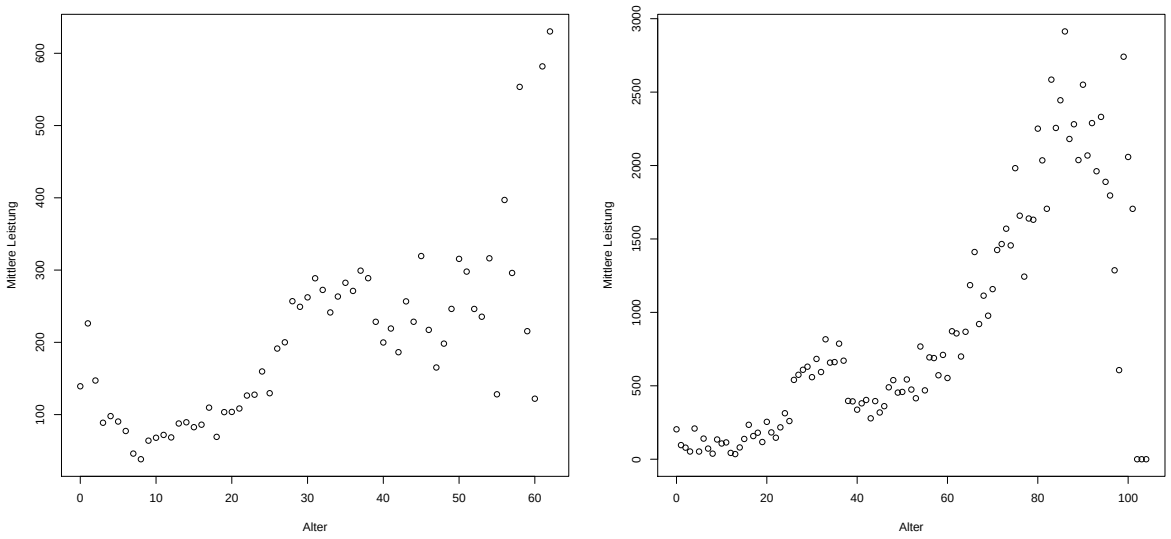


Abbildung 4.3: Gemittelte Leistungen der weiblichen Versicherten des Kalenderjahres 2010 für die Tarifklassen 1 (links) und 2 (rechts).

Medikamente) und 117 (Rezeptgebühren), sowohl für männliche, als auch weibliche Versicherte. Diese verursachen auch den Großteil der anfallenden Kosten. Es treten für Frauen mehr Versicherungsfälle auf, wodurch auch die anfallenden Leistungen höher sind. Frauen und Männer konsumieren in der Tarifklasse 1 dieselben Leistungsarten, der einzige geschlechterspezifische Unterschied ist hier die Leistungsart 16 (Ärztliche Sonderleistungskosten/ambulant) bei den Männern, was jedoch wegen dem lediglich einmaligen Auftreten vernachlässigbar ist.

Tabelle 4.3 zeigt die verursachten Kosten innerhalb der Tarifklasse 2 für das Kalenderjahr 2010. Für beide Geschlechter werden am meisten Leistungen der Arten 23 (Kurbehandlung), 2 (Krankenhaus-Operationsgebühren) und 3 (Interne Krankenhausbehandlungskosten) angefordert. Für die weiblichen Versicherten befindet sich keine Anforderung der Leistungsart 39 (Arzt- und Facharztkosten) im Datensatz, während eine männliche Person eine Leistung dieser Art angefordert hat. Im Gegensatz dazu ist keine Beobachtung der Leistungsarten 42 (Beratungsgespräch), 14 (Ambulante Chemotherapie) und 17 (Medikamente) für Männer vorhanden; für weibliche Versicherte gibt es jeweils eine Beobachtung. Für die Mehrheit der Beobachtungen in Tarifklasse 2 lässt sich wieder folgern, dass Frauen mehr Leistungen beanspruchen und damit höhere Kosten verursachen.

Nun werden einige bereits eingeführten Modellklassen auf die Datensätze angewendet. Es werden dabei sowohl die gemittelten Daten, als auch die Originaldaten

modelliert und Unterschiede analysiert.

| TKZ = 1 | Männer | | Frauen | |
|---|---|---|---|---|
| Leistungsart | Häufigkeit | Summe | Häufigkeit | Summe |
| 16 | 1 | 7129 | — | — |
| 20 | 1 | 9676 | 2 | 3211 |
| 30 | 3 | 15848 | 10 | 94909 |
| 57 | 22 | 174118 | 29 | 198969 |
| 42 | 29 | 233051 | 121 | 991536 |
| 15 | 157 | 1314258 | 320 | 2698066 |
| 23 | 455 | 3678788 | 768 | 5940993 |
| 18 | 554 | 4452739 | 958 | 7376912 |
| 117 | 1281 | 10644110 | 1919 | 15787858 |
| 17 | 1393 | 11921956 | 2028 | 17029145 |
| 39 | 1667 | 14148588 | 3021 | 26324207 |

Tabelle 4.2: Anzahl und Summe der erbrachten Leistungen, aufgeteilt nach Leistungsarten der Tarifklasse 1 für das Kalenderjahr 2010.

## 4.1 Anwendung der globalen kubischen Splines (GKS)

Zunächst werden *globale kubische Splines* betrachtet. Es werden als Basisdimension $q = 6$ und äquidistant verteilte Knotenpunkte $x_1^* = 1/5, \ldots, x_4^* = 4/5$ gewählt. Als Erstes werden die gemittelten Leistungen für die Leistungen der Männer des Kalenderjahres 2010 modelliert.

Fit und Plot der globalen kubischen Splines

```
1 x <- alter.m - min(alter.m)
2 x <- x/max(x)
3 xk <- 1:4/5
4 X <- spl.X(x,xk)
5 mod.1 <- lm(leistung.m~X-1)
6 plot(alter.m, leistung.m)
7 lines(alter.m, fitted(mod.1))
```

In Zeile 1 und 2 wird das Alter auf das Einheitsintervall $[0, 1]$ skaliert, damit eine bessere numerische Stabilität erzielt werden kann. Dann werden die Knotenpunkte und die Modellmatrix bestimmt, um anschließend das Modell unter Normalverteilungsannahme

| TKZ = 2 | Männer | | Frauen | |
|---|---|---|---|---|
| Leistungsart | Häufigkeit | Summe | Häufigkeit | Summe |
| 42 | — | — | 1 | 4298 |
| 14 | — | — | 1 | 5518 |
| 17 | — | — | 1 | 7115 |
| 13 | 2 | 410 | 3 | 36281 |
| 5 | 1 | 1167 | 42 | 58022 |
| 59 | 1 | 4072 | 1 | 3441 |
| 37 | 1 | 9343 | 3 | 16185 |
| 7 | 2 | 12877 | 9 | 71712 |
| 39 | 1 | 14891 | — | — |
| 1 | 4 | 25403 | 9 | 71248 |
| 27 | 16 | 149580 | 19 | 153142 |
| 6 | 27 | 187936 | 16 | 131546 |
| 45 | 23 | 195109 | 29 | 243936 |
| 4 | 51 | 357619 | 107 | 734846 |
| 57 | 79 | 614656 | 114 | 904967 |
| 8 | 114 | 880451 | 115 | 931566 |
| 48 | 110 | 1014335 | 258 | 2387289 |
| 10 | 102 | 1409742 | 267 | 3680000 |
| 12 | 321 | 2472957 | 462 | 3621710 |
| 24 | 369 | 3146518 | 490 | 4296848 |
| 23 | 616 | 6378467 | 1003 | 10217754 |
| 2 | 1575 | 10944487 | 2947 | 20759940 |
| 3 | 1596 | 11340212 | 3444 | 24827309 |

Tabelle 4.3: Anzahl und Summe der erbrachten Leistungen, aufgeteilt nach Leistungsarten der Tarifklasse 2 für das Kalenderjahr 2010.

für die Mittelwerte zu schätzen. Die geschätzte Kurve wird mit den gemittelten Leistungen verglichen.

Abbildung 4.4 zeigt eine gute Schätzung der Erwartungswerte der Daten feststellen. In der rechten Abbildung sind zusätzlich die 95% Konfidenzintervalle der Schätzungen für die jeweiligen Altersstufen abgebildet. Das Modell erzeugt ein *Akaike Informationskriterium* (AIC-Wert) von 648.97.

Nun wird dasselbe Modell auf die ungemittelten Daten angewendet. Die Basisdimension und die Wahl der Knotenpunkte bleiben dabei unverändert.

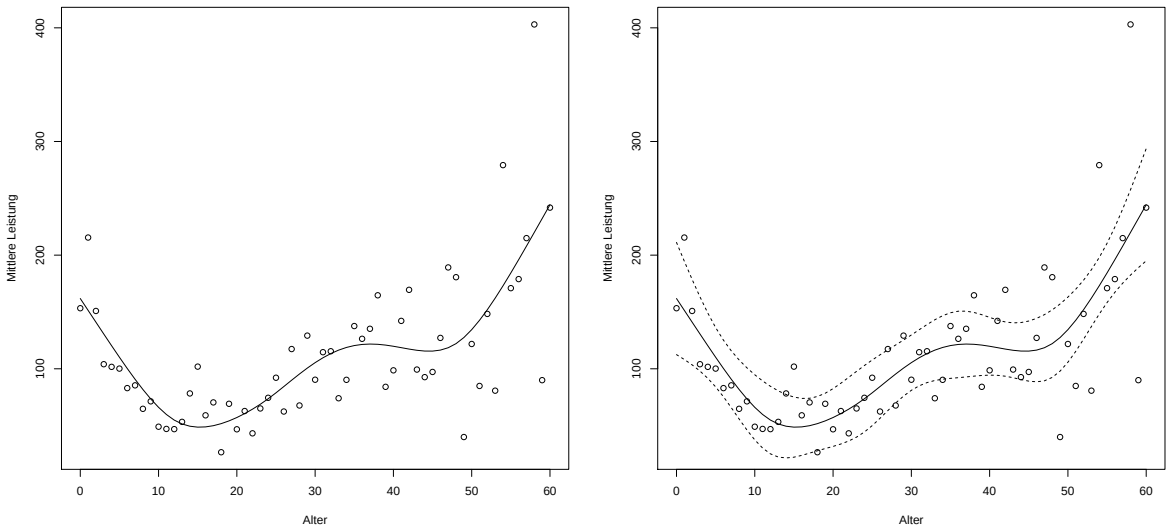


Abbildung 4.4: Schätzung der gemittelten männlichen Daten des Jahres 2010 durch globale kubische Splines mit Basis der Ordnung $q = 6$. Zusätzlich wird in der rechten Abbildung die 95% Konfidenzintervalle der Schätzung abgebildet.

Globale kubische Splines für ungemittelte Daten

```
1  x <- Alter - min(Alter); x <- x/max(x)
2  xk <- 1:4/5
3  X <- spl.X(x,xk)
4  mod.2 <- lm(Leistung~X-1)
5
6  daten.gks <- data.frame(leistung.gks = fitted(mod.2), alter.gks = Alter)
7  attach(daten.gks)
8  daten.gks <- daten.gks[order(alter.gks), ]
9
10 plot(Alter, Leistung)
11 lines(unique(alter.gks), unique(leistung.gks))
12 points(alter.m, leistung.m, pch=2, col="red")
```

Das Schätzen des Modells funktioniert analog zum gemittelten Modell. Für die ungemittelten Daten kommen jedoch mehrere Datenpunkte pro Altersstufe vor, sodass die geschätzten Werte nach dem Alter sortiert werden müssen. Dies geschieht in Zeile 8. Mittels dem Befehl `unique()` werden dann mehrfach vorkommende Werte entfernt. Abbildung 4.5 zeigt den Fit dieses Modells. Als Vergleich werden die Mittelwerte der jeweiligen Altersklasse rot gekennzeichnet.

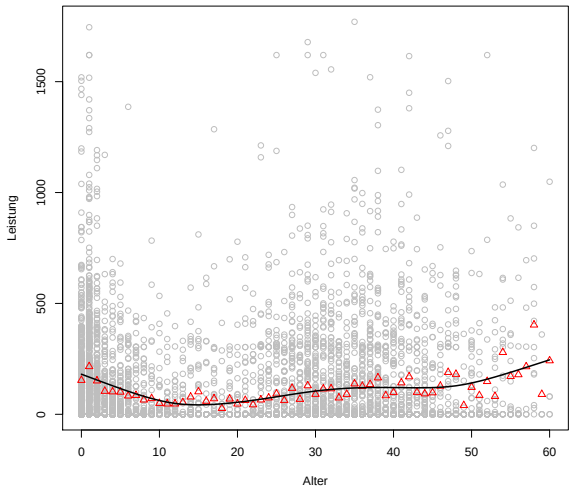


Abbildung 4.5: Schätzung der ungemittelten männlichen Daten durch globale kubische Splines mit Basis der Ordnung $q = 6$.

Obwohl die Originaldaten zur Modellierung verwendet wurden, passt sich die resultierende Schätzung gut an die Mittelwerte der Leistungen an. Das Modell der Originaldaten erzeugt einen AIC-Wert von 74604.88.

Jetzt werden die beiden Modelle, die durch Anwendung der gemittelten beziehungsweise ungemittelten Daten entstanden sind, miteinander verglichen. Im Hintergrund werden die arithmetischen Mittel der Leistungen zum Vergleich dargestellt. Abbildung 4.6 zeigt das sehr analoge Verhalten der beiden Modelle durch Anwendung der globalen kubischen Splines. Betrachtet man jedoch zusätzlich die jeweiligen Konfidenzintervalle (im aktuellen Fall das 95% Konfidenzintervall), erkennt man, dass die Schätzung durch Anwendung auf die Originaldaten aufgrund der engeren Konfidenzintervalle über den gesamten Bereich der $x$-Werte wesentlich präziser ist als die Schätzwerte bei Anwendung der Modellklasse auf die gemittelten Leistungen. Dies ist kein überraschendes Ergebnis, da bei den Modellen der Originaldaten eine viel höhere Anzahl an Beobachtungen vorhanden sind, um die Schätzungen zu berechnen.

Die Berechnung des 95% Konfidenzintervalls wird beispielhaft an dem Modell der gemittelten Leistungen `mod.1` illustriert.

Berechnung der 95% Konfidenzintervalle

```
1 new <- c(0:60)
2 pred <- predict(mod.1, newdata = data.frame(x=new),
3                 interval = c("confidence"), level = 0.95, type="response")
```

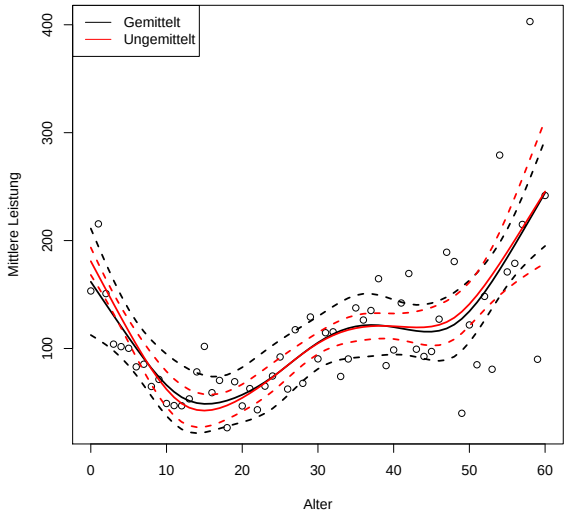


Abbildung 4.6: Vergleich der globalen kubischen Splines für TKZ = 1 der männlichen Versicherten des Jahres 2010, angewendet auf die ungemittelten (rot) und gemittelten Daten (schwarz). Zusätzlich werden die jeweiligen 95% Konfidenzintervalle aufgetragen.

Im Folgenden wird die Modellklasse der globalen kubischen Splines auch auf den weiblichen Datensatz des Kalenderjahres 2010 für TKZ = 2 angewendet. Abbildung 4.7 zeigt einen äußerst akzeptablen Fit der Daten; auch der „Geburten-Hügel“ ist bei der Schätzung gut zu erkennen, obwohl die Mittelwerte in diesem Bereich jedoch ein wenig unterschätzt werden. Offensichtlich ist dieses Modell nicht ausreichend flexibel, um diesen plötzlichen Anstieg der Leistungsansprüche zufriedenstellend zu modellieren, was als Indiz dafür angesehen werden kann, eine komplexere Modellklasse als die globalen kubischen Splines zu verwenden. Das Modell erzeugt einen AIC-Wert von 1123.26.

Als nächstes wird das Modell auf die ungemittelten Daten angewendet.

Abbildung 4.8 (links) zeigt das Modell der globalen kubischen Splines bei Anwendung auf die ungemittelten weiblichen Leistungen der Tarifklasse 2. Da das geschätzte Modell im Vergleich zu den Originaldaten ein sehr niedriges Niveau hat, wird zusätzlich eine genauere Betrachtung des Modells erzeugt, indem es im rechten Teil der Abbildung „hervorgehoben“ wird, das heißt, der Bereich der $y$-Achse wird eingeschränkt. Es lassen sich kaum Unterschiede zu den Mittelwerten der Altersklassen erkennen. Lediglich der „Geburten-Hügel“ wird wie beim Modell der gemittelten Leistungen etwas unterschätzt. Der AIC-Wert des Modells entspricht 422774.90.

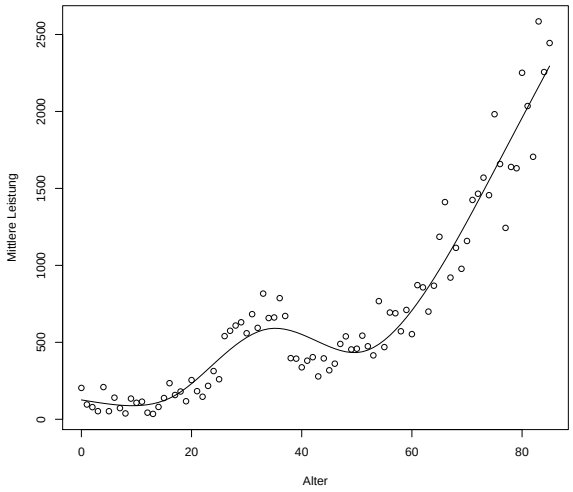


Abbildung 4.7: Schätzung der gemittelten Daten durch globale kubische Splines für TKZ = 2 der weibliche Versicherten des Kalenderjahres 2010.

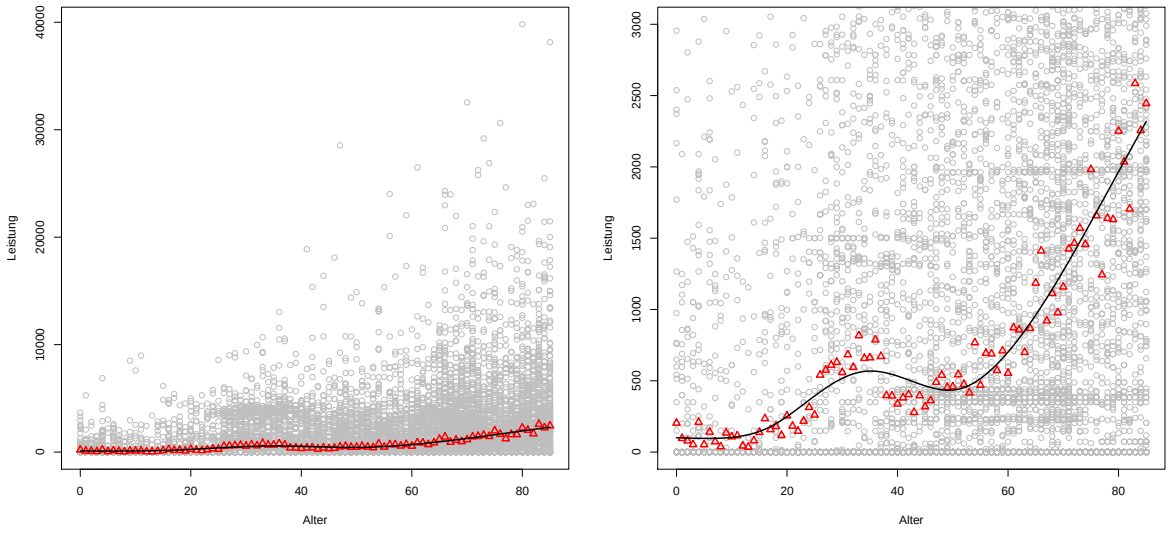


Abbildung 4.8: Schätzung der ungemittelten weiblichen Daten durch globale kubische Splines mit Basis der Ordnung $q = 6$ und eine vergrößerte Ansicht des geschätzten Modells.

Abbildung 4.9 zeigt einen Vergleich der beiden erzeugten Modelle. Im Hintergrund werden die gemittelten Leistungen abgebildet. Auch hier werden auf den ersten Blick zwei praktisch identische Schätzungen erzeugt, mit minimalen Unterschieden für junge Versicherte und im Bereich des „Geburten-Hügels“. Doch abermals besitzt die Schätzung bei Anwendung der Modellklasse auf die Originaldaten engere Konfidenzintervalle (auch wenn es auf Grund des höheren Niveaus der $y$-Skala nicht so deutlich erkennbar ist), was für eine präzisere Schätzung im Vergleich zur Schätzung durch die gemittelten Leistungen schließen lässt.

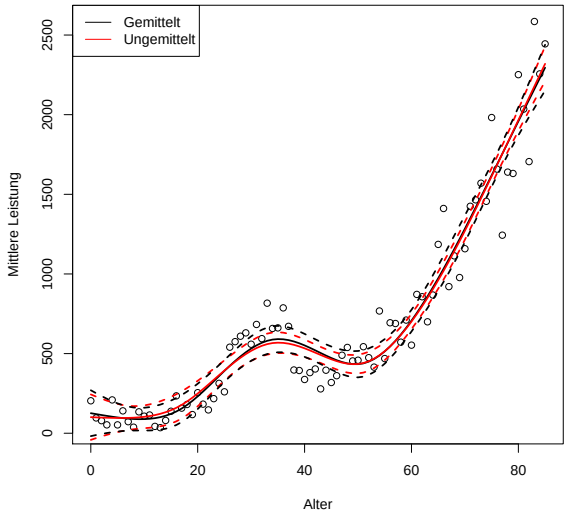


Abbildung 4.9: Vergleich der globalen kubischen Splines für TKZ = 2 der weiblichen Versicherten des Jahres 2010, angewendet auf die ungemittelten (rot) und gemittelten Daten (schwarz). Zusätzlich werden die jeweiligen 95% Konfidenzintervalle aufgetragen.

## 4.2 Penalisierte Regressions-Splines

Nun kommen die *Penalisierten Regressions-Splines* zur Anwendung. Dabei werden die globalen kubischen Splines mit einem Strafterm versehen. Die optimalen Glättungsparameter bestimmt man durch Minimierung des GCV-Scores.

Zunächst wird TKZ = 1 der gemittelten Daten für männliche Versicherte des

Kalenderjahres 2010 betrachtet.

Berechnung des optimalen Glättungsparameter

```
x <- alter.m - min(alter.m)
x <- x/max(x)
lambda <- 1e-8; n <- length(leistung.m); V <- 0
for (i in 1:60)
{
  mod <- prs.fit(leistung.m,x,xk,lambda)
  trA <- sum(influence(mod)$hat[1:n])
  rss <- sum((leistung.m-fitted(mod)[1:n])^2)
  V[i] <- n*rss/(n-trA)^2
  lambda <- lambda*1.5
}
which.min(V)
```

```
[1] 31
```

Der als optimal berechnete Glättungsparameter ist $\hat{\lambda} = 1 \cdot 10^{-8} \cdot 1.5^{30} \approx 1.92 \cdot 10^{-3}$. Zur Anwendung der PRS werden ein paar gleichmäßig verteilte Knotenpunkte gewählt, um das Modell zu schätzen: $x_1^* = 1/8, \ldots, x_7^* = 7/8$.

Schätzen und Plotten eines PRS-Modells

```
xk <- 1:7/8
xp <- 0:max(alter.m)/max(alter.m)
mod.13 <- prs.fit(leistung.m,x,xk, 10^(-8)*1.5^(30))
Xp <- spl.X(xp,xk)
plot(alter.m, leistung.m)
lines(alter.m, Xp%*%coef(mod.13))
```

Der AIC-Wert des Modells entspricht 744.14.

Abbildung 4.10 zeigt den erhaltenen Fit durch Anwendung der Modellklasse auf die gemittelten männlichen Leistungen. Zum Vergleich werden die globalen kubischen Splines der gemittelten Leistungen aufgetragen. Man erkennt einen guten Fit des Modells über den gesamten Bereich der Daten bei Anwendung der PRS. Durch PRS und mittels GCV-Score minimierten Glättungsparameter erhält man eine glattere Schätzung als durch Anwendung der globalen kubischen Splines.

Die PRS-Modelle lassen sich nicht effizient auf die ungemittelten Daten anwenden, da der Rechenaufwand des GCV-Algorithmus zu hoch ist, um vernünftige Laufzeiten zu garantieren.

Nun wird die Modellklasse auf die gemittelten Leistungen der weiblichen Versicherten in der Tarifklasse 2 angewendet. Es ergibt sich ein optimaler Glättungsparameter von

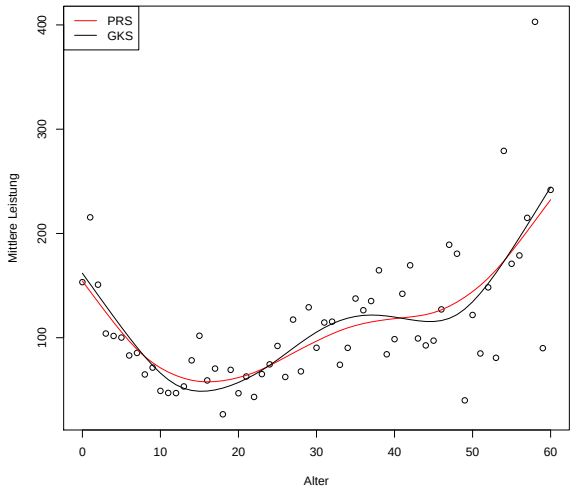


Abbildung 4.10: Penalisierte Regression-Splines (rot), im Vergleich zu den globalen kubischen Splines (schwarz), angewendet auf die gemittelten Leistungen der Tarifklasse 1 für Männer im Jahr 2010.

$\hat{\lambda} = 1 \cdot 10^{-8} \cdot 1.5^{22} \approx 7.50 \cdot 10^{-5}$, es wird also ein sehr flexibles Modell zugelassen, da der niedrige Wert kaum Bestrafung bedeutet.

Abbildung 4.11 zeigt den Vergleich der PRS mit den globalen kubischen Splines für die gemittelten Leistungen von TKZ = 2 der weiblichen Leistungen des Jahres 2010. Es lässt sich ein besserer Fit bei den PRS feststellen; vor allem der „Geburten-Hügel“ lässt sich mit den PRS nahezu perfekt abbilden. Der AIC-Wert des Modells entspricht 1223.21.

## 4.3 Anwendung der P-Splines

Im Folgenden werden *P-Splines* auf die Daten angewendet, also B-Splines, die mit einem Strafterm versehen werden, der aus quadrierten Differenzen benachbarter Parameterwerte besteht.

Die einfachste Methode, P-Splines mittels `R` zu modellieren, ist die Anwendung der Funktion `gam()`. Dazu muss das Paket `mgcv` geladen werden.

Schätzen und Plotten von P-Splines

```
1 library(mgcv)
2 gam(leistung.m ~ s(alter.m, bs="ps"))
```

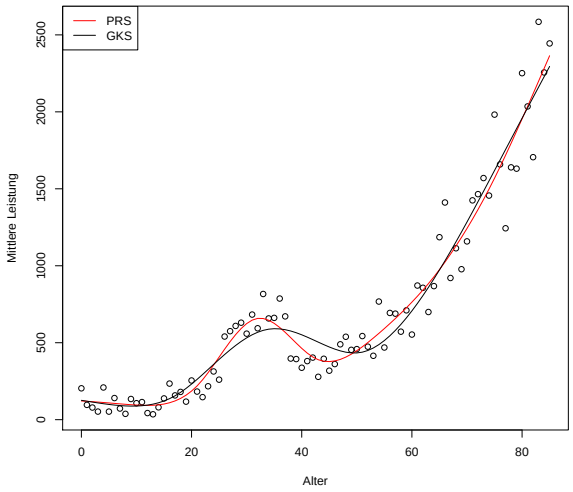


Abbildung 4.11: Penalisierte Regression-Splines (rot), im Vergleich zu den globalen kubischen Splines (schwarz), angewendet auf die gemittelten Leistungen der Tarifklasse 2 für Frauen im Jahr 2010.

Durch die Funktion `s()` lässt sich der Glättungsterm des Modells definieren. Mit der Angabe von `bs="ps"` werden P-Splines festgelegt. Standardmäßig sind dies kubische B-Splines mit einem Strafterm zweiter Ordnung. Durch die Funktion `fitted()` können die geschätzten Leistungen pro Altersklasse dargestellt werden.

Abbildung 4.12 zeigt das geschätzte P-Splines Modell. Zum Vergleich wurde das PRS-Modell der gemittelten Leistungen dargestellt. Die beiden Modelle sind praktisch identisch. Es lassen sich kaum Unterschiede der gefitteten Modelle feststellen. Der AIC-Wert des geschätzten P-Spline Modells entspricht 648.51.

P-Splines lassen sich auch bezüglich der ungemittelten Daten darstellen. Abbildung 4.13 zeigt das durch die Originaldaten geschätzte P-Spline Modell, sowie den Vergleich zu dem mit den gemittelten Werten erzeugten Modell. Es sind einige kleine Unterschiede bemerkbar, die darauf zurückzuführen sind, dass für das gemittelte Modell jeder Leistungspunkt gleich gewichtet wird, egal wieviel Versicherte sich in der jeweiligen Altersstufe befinden, während beim ungemittelten Modell jeder einzelne Datenpunkt berücksichtigt wird. Das ungemittelte Modell besitzt einen AIC-Werte von 74601.39. Aufgrund der geringen Unterschiede der beiden Modelle ist hier wohl wieder jenes der gemittelten Leistungen zu bevorzugen.

Im Folgenden werden die P-Splines für Tarifklasse 2 der weiblichen Leistungen des Jahres 2010 angewendet. Abbildung 4.14 zeigt den Vergleich des PRS-Modells

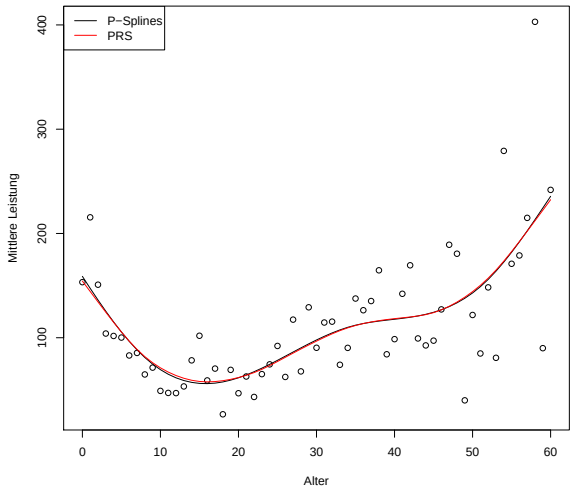


Abbildung 4.12: Vergleich der Penalisierten Regression-Splines (rot) und der P-Splines (schwarz) bezüglich der gemittelten Daten für die Tarifklasse 1 der männlichen Versicherten des Jahres 2010.

mit dem Modell der P-Splines für die gemittelten Leistungen. Die Schätzung des PRS-Modells ist vor allem im Bereich des „Geburten-Hügels“ deutlich besser als jene der P-Splines. Das P-Spline Modell der gemittelten weiblichen Leistungen der Tarifklasse 2 hat einen AIC-Wert von 1122.21.

In Abbildung 4.15 werden die P-Splines auf die Originaldaten der weiblichen Versicherten des Jahres 2010 für die Tarifklasse 2 angewendet. Im Plot des Modells gegenüber den Originaldaten lässt sich das Modell kaum erkennen. Wenn man jedoch wieder die beiden Modelle der gemittelten und ungemittelten Daten vergleicht, erkennt man die Analogie der beiden Modelle. Aus diesem Grund ist das Modell der gemittelten Daten wohl dem Modell der Ausgangsdaten aufgrund des geringeren Rechenaufwands wieder vorzuziehen. Der AIC-Wert des Modells der Originaldaten entspricht 422770.10.

## 4.4 Anwendung von GAMs

Nun sollen *Generalisierte Additive Modelle* auf die Daten angewendet werden. Da die Gammaverteilung eine klassische Verteilung in der Versicherungsmathematik darstellt, wird sie im Folgenden zur Anwendung kommen. Sie ist jedoch nur für positive Werte geeignet, deshalb müssten die Nullwerte aus den Leistungen entfernt werden. Dies trifft jedoch für einen Großteil der Daten zu, weshalb dies das Ergebnis zu stark verfälscht.

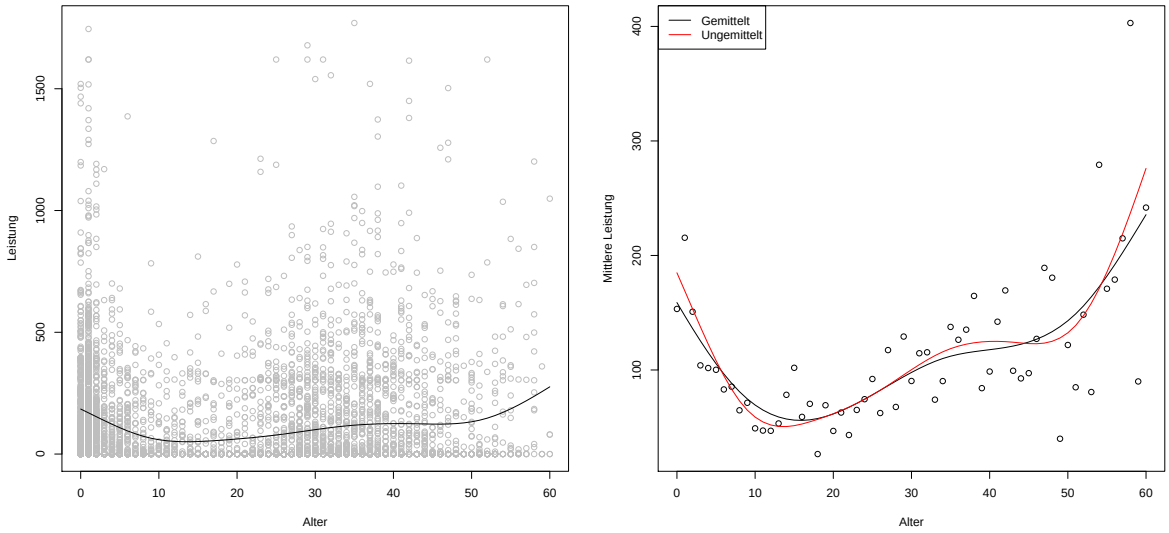


Abbildung 4.13: Schätzung der Originaldaten mittels P-Splines (links) und Vergleich mit dem P-Spline Modell der gemittelten Leistungen (rechts) für TKZ = 1 der männlichen Leistungen des Jahres 2010. Schwarz: Modell der gemittelten Leistungen. Rot: Modell der Originaldaten.

Das Problem lässt sich aber mit der R-Funktion `quasi()` lösen. Für die Gammaverteilung kann man zeigen, dass sie eine quadratische Varianzfunktion und eine kanonische inverse Linkfunktion besitzt:

$$V(\mu) = \mu^2, \quad g(\mu) = \frac{1}{\mu}.$$

Für positive Responses, wie sie in dieser Arbeit betrachtet werden, ist jedoch eine Log-Linkfunktion passender, da die inverse Linkfunktion auch negative Erwartungswerte generieren kann, was bei der Gammaverteilung nicht möglich sein sollte. Diese Eigenschaften werden dem Modell mittels `quasi()` übergeben.

Zunächst wird das log-lineare GAM unter Annahme der Gammaverteilung auf die gemittelten Leistungen für Männner in der Tarifklasse 2 des Jahres 2010 angewendet.

GAM für gemittelte Daten

```
1 rownames(leistung.m) <- NULL
2 gam(leistung.m~s(alter.m), family = quasi(link = log, variance="mu^2"))
```

Im Vergleich zum jeweiligen PRS-Modell des Datensatzes erkennt man in Abbildung 4.16 kaum gravierende Unterschiede. Der Plot des Glättungsterms $f(x)$ zeigt, dass die null nicht über den gesamten Prädiktorbereich von den 95% Konfidenzintervallen überdeckt

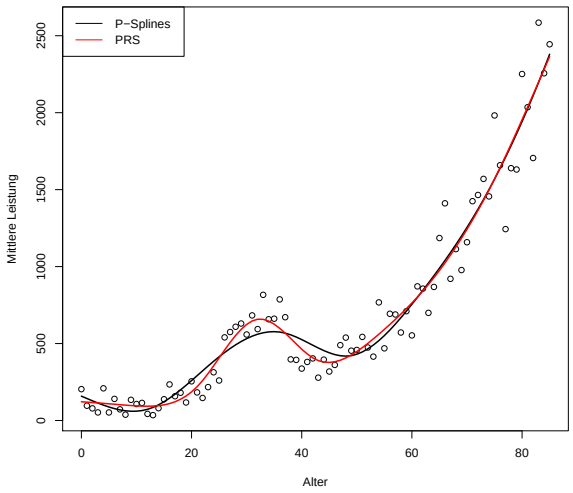


Abbildung 4.14: Vergleich der Penalisierten Regression-Splines (rot) und der P-Splines (schwarz) bezüglich der gemittelten weiblichen Daten des Jahres 2010 für TKZ = 2.

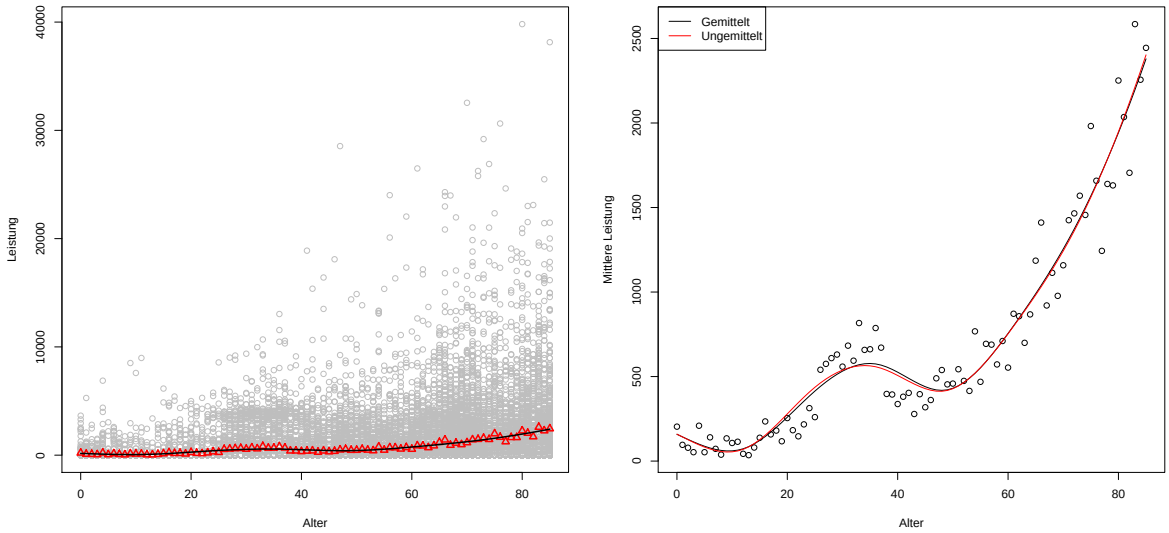


Abbildung 4.15: Anwendung der P-Splines auf die ungemittelten weiblichen Leistungen (rot) der Tarifklasse 2 vom Jahr 2010 und Vergleich mit dem jeweiligen Modell der Mittelwerte (schwarz).

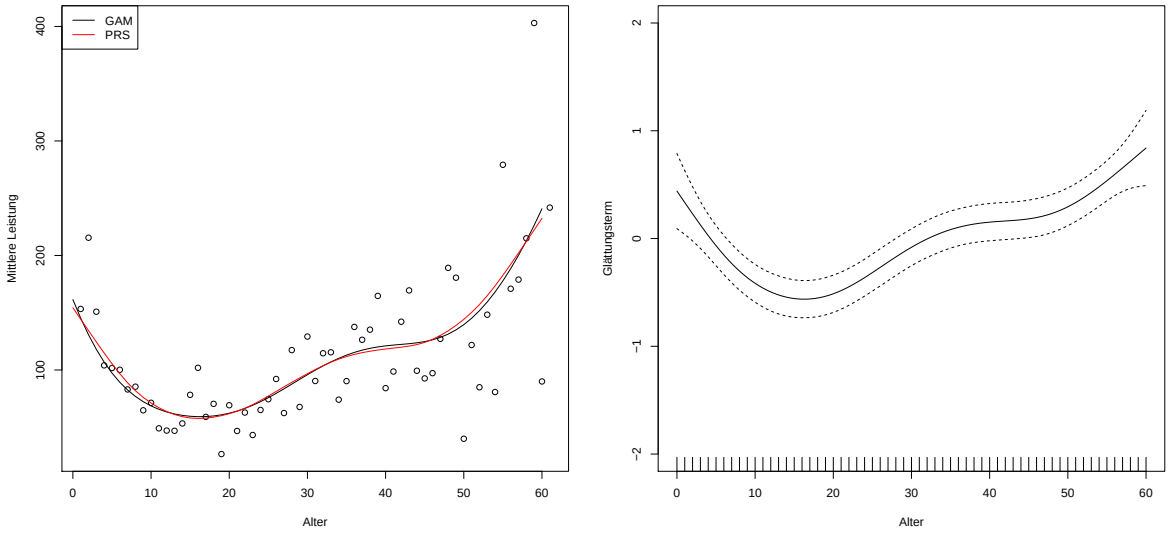


Abbildung 4.16: Anwendung eines GAMs unter Annahme der Gammaverteilung auf die gemittelten Leistungen von TKZ = 1 für männliche Versicherte des Jahres 2010. Links: Vergleich des Modells mit dem jeweiligen PRS-Modell. Rechts: Plot des geschätzten Glättungsterms und der jeweiligen 95% Konfidenzintervalle. Der EDF wird mit 4.84 geschätzt.

wird. Dies und der geschätzte *Estimated Degree of Freedom* von 4.84 sprechen für die Signifikanz der Funktion $f(x)$.

```
Family: quasi
Link function: log

Formula:
leistung.m ~ s(alter.m)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.64418    0.04383     106   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
                             edf Ref.df     F  p-value
s(alter.m) 4.84  5.915 13.36 2.52e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

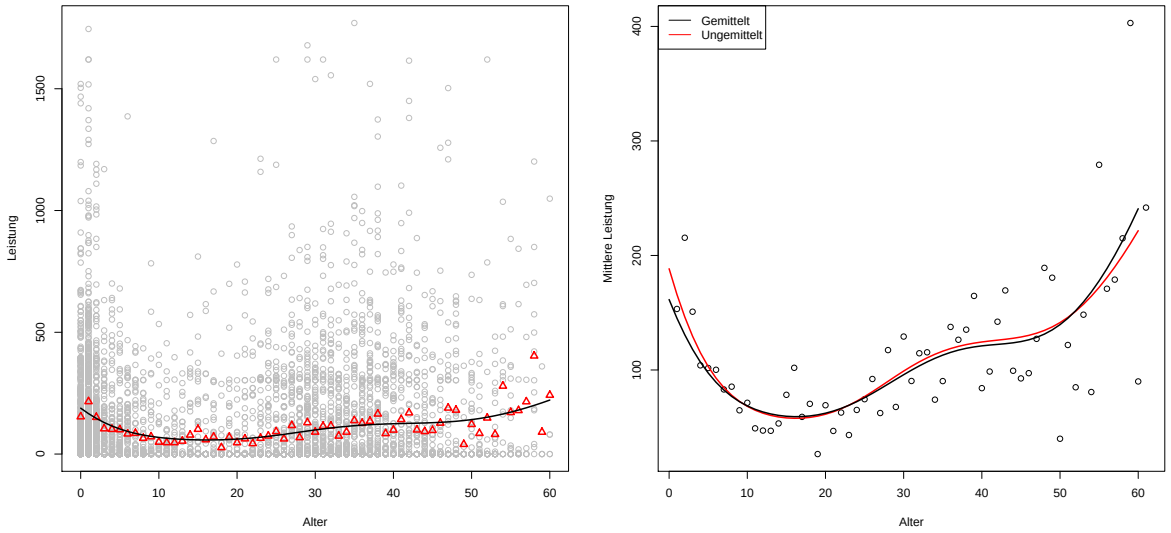


Abbildung 4.17: GAM der ungemittelten Leistungen für männliche Versicherte in der Tarifklasse 2 des Kalenderjahres 2010 (links) und Vergleich mit dem jeweiligen Modell der Mittelwerte (rechts).

```
R-sq.(adj) =  0.484   Deviance explained = 60.4%
GCV score = 0.12961  Scale est. = 0.1172    n = 61
```

Zum Vergleich wird nun das Modell bezüglich der ungemittelten Daten modelliert.

Schätzung eines GAMs

```
1 gam(Leistung~s(Alter), family = quasi(link = log, variance="mu^2"),
2     data=men2010.1)
```

Der linke Teil der Abbildung 4.17 zeigt das GAM der ungemittelten Leistungen. Zusätzlich werden die Mittelwerte rot gekennzeichnet. Der rechte Teil der Abbildung vergleicht die Modelle der gemittelten und ungemittelten Leistungen miteinander. Es handelt sich abermals um praktisch identische Modelle.

Abbildung 4.18 zeigt den Glättungsterm des Modells. Die 95% Konfidenzintervalle des Glättungsterms $f(x)$ sind für die Originaldaten enger als jene für das Modell der gemittelten Leistungen. Der Plot des Glättungsterms und der `edf`, den man anhand der `summary()`-Ausgabe des Modells erhält, sprechen abermals für die Signifikanz des Glättungsterms.

Nun wird das log-lineare GAM auf weiblichen Originaldaten der Tarifklasse 2 des

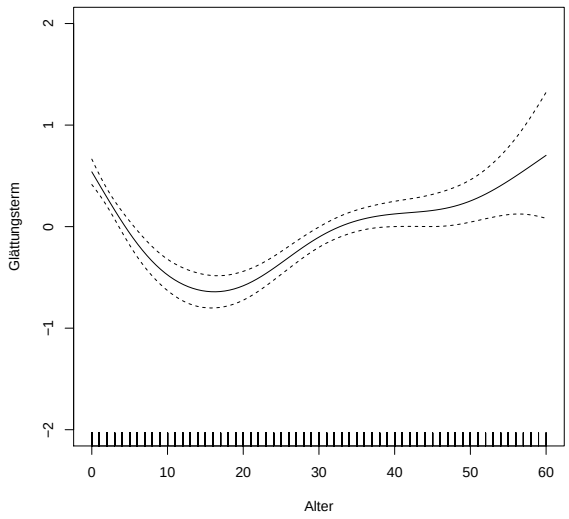


Abbildung 4.18: Glättungsterm des GAMs für die ungemittelten männlichen Leistungen der Tarifklasse 1. Die Schätzung des Freiheitsgrades entspricht 5.09.

Jahres 2010 angewendet. Abbildung 4.19 stellt den Vergleich des durch Anwendung auf die gemittelten Leistungen erzeugten GAMs mit dem jeweiligen PRS-Modell dar. Beide Modelle zeigen einen ziemlich guten Fit der Daten, jedoch scheint das PRS-Modell den „Geburten-Hügel“ noch etwas besser darzustellen als das GAM. Zusätzlich wird der Glättungsterm des Modells in der rechten Graphik abgebildet. Sowohl der Plot von $f(x)$ als auch der geschätzte Freiheitsgrad von 7.46 sprechen für die Signifikanz des Glättungsterms.

Im Folgenden werden als Vergleich die ungemittelten Daten mittels GAM modelliert. Abbildung 4.20 zeigt das analoge Verhalten des GAM bei Anwendung gemittelter beziehungsweise ungemittelter Leistungen.

Bei Betrachtung des Glättungsterms in Abbildung 4.21 bekommt man ein bekanntes Bild zu sehen: Weder die Abbildung der Konfidenzintervalle des Glättungsterms noch der geschätzte Freiheitsgrad sprechen gegen die Signifikanz des Glättungsterms. Aufgrund der optisch äquivalenten Schätzung der Daten durch die Modelle der gemittelten beziehungsweise ungemittelten Daten ist für GAMs das Modell der Mittelwerte zu bevorzugen, da der Rechenaufwand deutlich geringer ist.

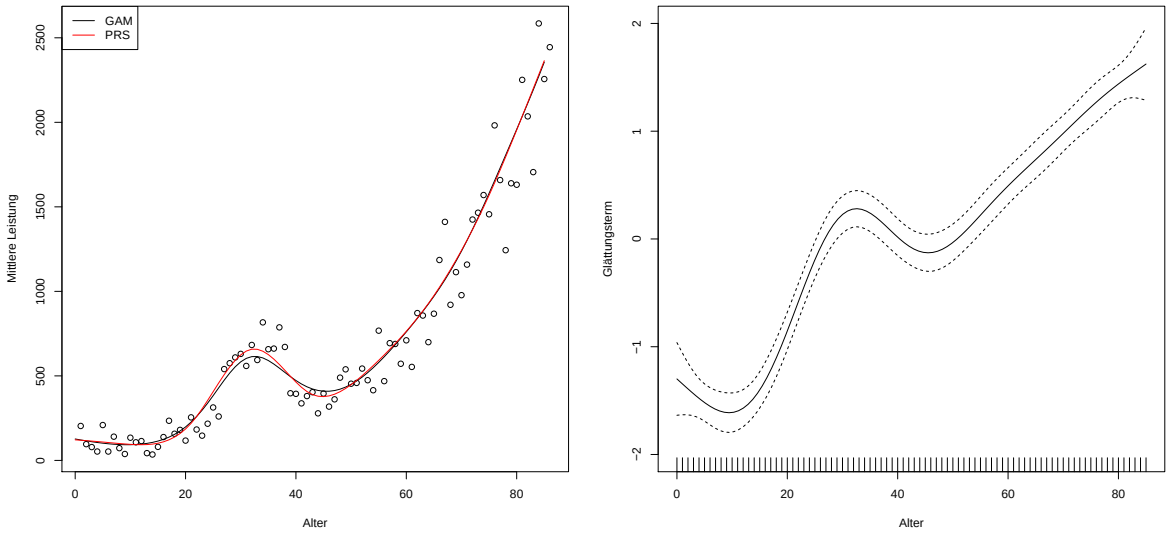


Abbildung 4.19: Links: Anwendung eines GAMs unter Annahme der Gammaverteilung auf die gemittelten weiblichen Leistungen für TKZ = 2 des Jahres 2010 (schwarz) und Vergleich mit dem jeweiligen PRS Modell (rot). Rechts: Darstellung des geschätzten Glättungsterms und der jeweiligen 95% Konfidenzintervalle.

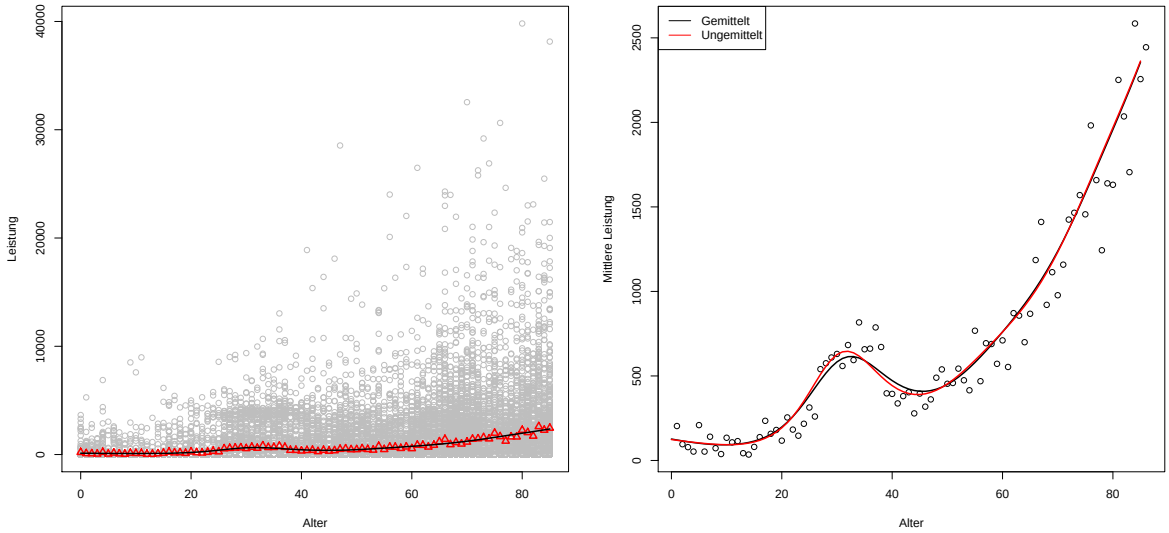


Abbildung 4.20: GAM der ungemittelten Leistungen für TKZ = 2 der weiblichen Versicherten des Jahres 2010 (links) und Vergleich mit dem jeweiligen Modell der Mittelwerte (rechts).

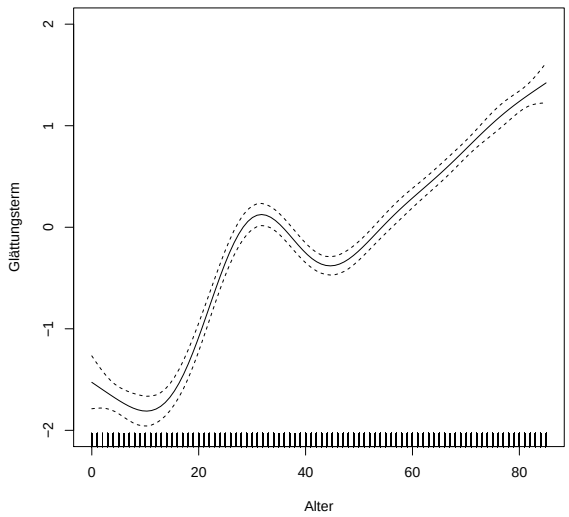


Abbildung 4.21: Schätzung des Glättungsterms für das GAM bei Anwendung auf die ungemittelten weiblichen Daten für TKZ = 2 des Kalenderjahres 2010. Der *Estimated Degree of Freedom* entspricht 7.46.

## 4.5 Fazit der Erwartungswertmodelle

Zusammenfassend werden nun die Ergebnisse der betrachteten Erwartungswertmodelle dokumentiert. Für die männlichen Leistungen lassen sich durchwegs vernünftige Schätzungen mit allen Modellklassen feststellen. Aufgrund des Scatterplots der arithmetischen Mittel der Leistungen für Frauen erkennt man jedoch einen komplexeren Zusammenhang der Leistung mit dem jeweiligen Alter der versicherten Person, wodurch höhere Anforderungen an das Modell gestellt werden müssen.

Für die Schätzung der Leistungen durch *globale kubische Splines* erkennt man sowohl für das durch die gemittelten Leistungen geschätzte Modell in Abbildung 4.7 als auch für das Modell der Originaldaten in Abbildung 4.8 keine zufriedenstellende Schätzung. Im kritischen Bereich des „Geburten-Hügels“ bei Frauen im Alter zwischen 20 und 40 ist das Modell offensichtlich nicht flexibel genug, um den komplexen Verlauf der Daten akzeptabel darzustellen.

Werden die globalen kubischen Splines mit einem Strafterm versehen, erhält man das *Penalisierte Regressions-Splines Modell*. Durch Anwendung dieser Modellklasse lässt sich eine äußerst gute Schätzung erkennen. Im Vergleich zu dem GKS-Modell in Abbildung 4.11 ergibt sich eine deutlich bessere Schätzung der Beobachtungen im Altersbereich 20-40 Jahre. Eine Einführung eines Strafterms inklusive Berechnung des optimalen Glättungsparameter macht das Modell offensichtlich bereits flexibel genug für die vorhandene Datenlage. Ein Nachteil dieser Modellklasse ist der Rechenaufwand zur Ermittlung dieses Glättungsparameters, was eine Betrachtung dieser Methoden für die ungemittelten Leistungsdaten unmöglich macht.

Die Anwendung der *P-Splines*, das heißt von B-Splines in Verbindung mit einem Strafterm, der die quadrierten Differenzen benachbarter Parameterwerte bestraft, führt zu keiner verbesserten Schätzung. Obwohl die Lokalität der B-Splines eine relativ flexible Schätzung ermöglichen sollte, wird auch hier der „Geburten-Hügel“ etwas zu niedrig geschätzt. Dies erkennt man in Abbildung 4.14, wo die Schätzung der PRS zum Vergleich besser dargestellt wird als das P-Splines Modell für die gemittelten Daten. Auch bei Anwendung der Modellkasse auf die Ausgangsdaten erkennt man in Abbildung 4.15 kaum Unterschiede.

In Abbildung 4.19 erhält man durch Anwendung eines log-lineare GAMs erstmals eine annähernd äquivalente Schätzung zu dem bisher optimalen PRS-Modell. Man erkennt zwar immer noch eine höhere Schätzung des „Geburten-Hügels“ für das PRS-Modell, jedoch lässt sich die Modellklasse der GAM auch auf die ungemittelten Leistungen anwenden, wodurch in Abbildung 4.20 auch hier eine höhere Schätzung des kritischen Bereichs erkennbar ist. Aufgrund der ähnlichen Schätzungen und dem geringeren Rechenaufwand der GAMs, was eine einfachere Anwendung ermöglicht, wird hier

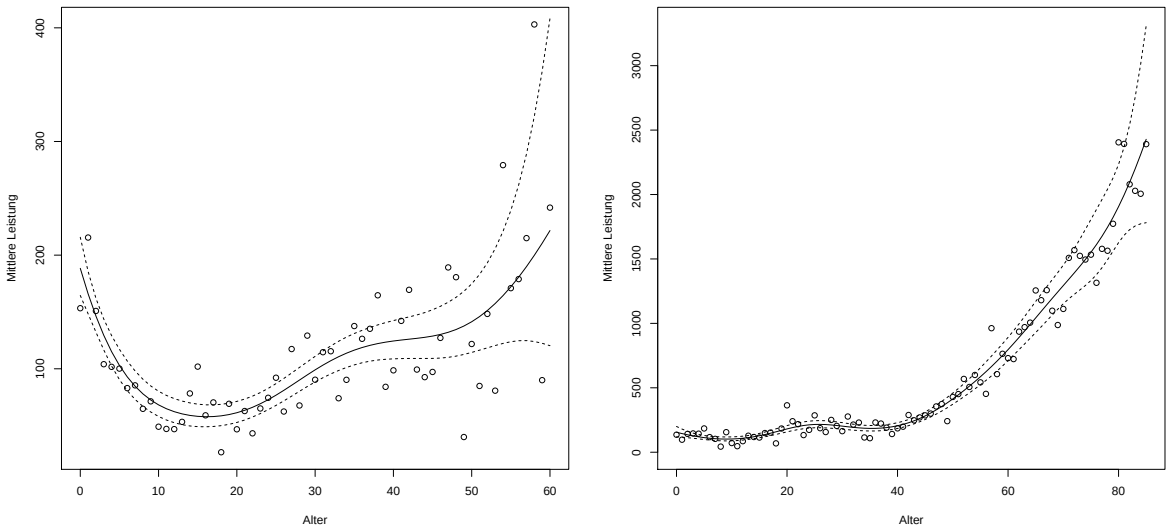


Abbildung 4.22: Vergleich der männlichen Leistungen für Tarifklasse 1 (links) und 2 (rechts) durch Anwendung eines GAMs auf die ungemittelten Leistungen. Zusätzlich werden die 95% Konfidenzintervalle für die jeweiligen Altersstufen abgebildet.

die GAM-Schätzung als optimale Modellklasse innerhalb der Erwartungswertmodelle angesehen.

Anhand der Modellklasse der GAM soll nun zum Abschluß der Erwartungswertmodelle eine Gegenüberstellung der Tarifklassen 1 und 2 für beide Geschlechter durchgeführt werden. Aufgrund der starken Ähnlichkeit der geschätzten Kurven bei Anwendung der Modellklasse auf die gemittelten bzw. ungemittelten Daten werden die Modelle lediglich auf die Originaldaten angewendet, da hier die jeweiligen Konfidenzintervalle etwas enger sind. Die 95% Konfidenzintervalle werden mittels Approximation durch die Normalverteilung erzeugt.

Abbildung 4.22 zeigt ein deutlich höheres Niveau der Erwartungswertschätzungen innerhalb der Tarifklasse 2, erkennbar an dem Maximalwert der $y$-Skala der rechten Graphik. In Tarifklasse 1 ist eine vergleichsweise hohe Schätzung für Kleinkinder erkennbar, mit einem am höchsten geschätzten Erwartungswert von 188.39 für 0-Jährige. Ab der Altersstufe 20 zeigt das geschätzte Modell einen steigenden Trend, mit einem deutlichen Anstieg im Altersbereich von ca. 30-40 Jahren. Die höchste Schätzung ist für 60-jährige männliche Versicherte mit 221.63 erkennbar.

Tarifklasse 2 zeigt generell ein wachsendes Verhalten der geschätzten Erwartungs-

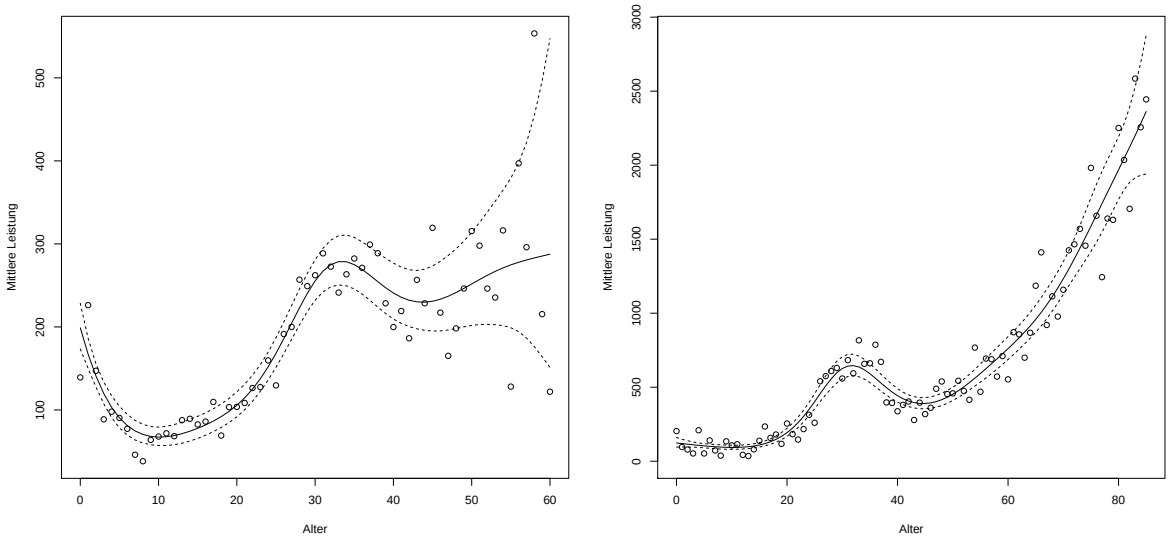


Abbildung 4.23: Vergleich der weiblichen Leistungen für Tarifklasse 1 (links) und 2 (rechts) durch Anwendung eines GAMs auf die ungemittelten Leistungen.

wertfunktion mit zunehmendem Alter der versicherten Personen. Im Bereich von circa 20-40 Jahren ist ein leichter Anstieg des geschätzten Leistungsanspruchs feststellbar. Das Maximum der Schätzung wird im Alter von 85 Jahren mit 2385.78 erreicht.

Für den oberen Altersbereich der Daten verbreitern sich vor allem die Konfidenzintervalle für die männlichen Leistungsdaten der Tarifklasse 1 des Jahres 2010 deutlich, da hier sehr wenige Datenpunkte vorhanden sind. Es befinden sich nur mehr fünf Beobachtungen für 60-Jährige Männer im Datensatz; zum Vergleich existieren noch 306 Beobachtungen 85-jähriger Männer für Tarifklasse 2.

Auch für die weiblichen Leistungen des Kalenderjahres 2010 sind die beanspruchten Leistungen in der Tarifklasse 2 deutlich höher als jene in der Tarifklasse 1. Für TKZ=1 in Abbildung 4.23 erkennt man eine vergleichsweise hohe Erwartungswertschätzung für Kleinkinder mit einem höchsten Wert von 199.05 für 0-Jährige. Für die Altersstufen von 30-40 Jahren steigt die Schätzung deutlich an, mit einem Maximum von 278.67 für 33-Jährige. Das globale Maximum wird für die 60-jährigen weiblichen Versicherten mit 287.62 geschätzt.

Die Erwartungswertschätzungen der Tarifklasse 2 zeigen den bereits erwähnten „Geburten-Hügel“ im Bereich der Altersklassen von circa 20-40 Jahren mit einem Maximum von 645.87 für 32-jährige weibliche Versicherte. Das globale Maximum wird

für den Erwartungswert der 85-Jährigen mit 2362.63 geschätzt.

Für die weiblichen Daten lasst sich für den oberen Altersbereich analog zu den männlichen Daten feststellen, dass die Konfidenzintervalle in der Tarifklasse 1 deutlich breiter werden. Hier existieren lediglich noch 13 Beobachtungen für 60-jährige Frauen in TKZ = 1; für Tarifklasse 2 befinden sich noch 306 85-Jährige im Datensatz, wodurch kein so extremer Anstieg der Breite der Konfidenzintervalle feststellbar ist.

Im Allgemeinen stellen die Erwartungswertmodelle gute Möglichkeiten dar, die Leistungen der Versicherung zu beschreiben. Auch für die Kalenderjahre 2011 und 2012 kommt man zu ähnlichen Ergebnissen und damit denselben Schlüssen. Sowohl die Betrachtung der Modelle für die Originaldaten als auch die gemittelten Leistungen liefern gute und sehr ähnliche Ergebnisse.

## 4.6 Anwendung von QR-Modellen

Im Folgenden werden Modelle der *Quantilen Regression* auf die Daten angewendet. Aufgrund des Scatterplots der Daten werden nichtparametrische Modelle genutzt. Da die Struktur der Daten nicht für einen stückweisen linearen Verlauf eines geschätzten Modells spricht, wird dafür die Methode der B-Splines verwendet. Die Anwendung der QR-Modelle erfolgt dabei lediglich auf die Originaldaten.

Die Anwendung von Modellen der Quantilen Regression bei gemittelten Leistungen führt zu keiner sinnvollen Schätzung. Abbildung 4.24 zeigt die Darstellung von QR-Modellen für die gemittelten männlichen Leistungen der Tarifklasse 1 des Jahres 2010 für Quantilniveaus $\tau \in \{0.50, 0.70, 0.90\}$. Aufgrund der geringen Anzahl an Datenpunkten ist „Quantile-Crossing“ in den Schätzungen ersichtlich. Zusätzlich liefern die gefitteten Quantilsfunktionen *keine* Information über die Ausgangsdaten. Somit kann keinerlei Aussage über die Verteilung der Response gemacht werden, was jedoch eine wesentliche Rechtfertigung der Anwendung Quantiler Regression ist.

Um nun ein QR-Modell zu schätzen, muss zunächst das Paket `quantreg` geladen werden (Koenker, 2005).

Laden des Pakets quantreg

```
1 library(quantreg)
```

Mittels der Funktion `rq()` lassen sich Quantile Regressionsmodelle schätzen. Durch das Argument `tau` kann das Niveau des zu berechnenden Quantils bestimmt werden. Um nichtparametrische Modelle zu erzeugen, wird die Funktion `bs()` aus dem Paket `splines` verwendet. Dies führt dazu, dass B-Splines zur Modellierung benutzt werden. Standardmäßig sind dies kubische Splines. Das Argument `df` spezifiziert die Anzahl der zu betrachtenden Knoten.

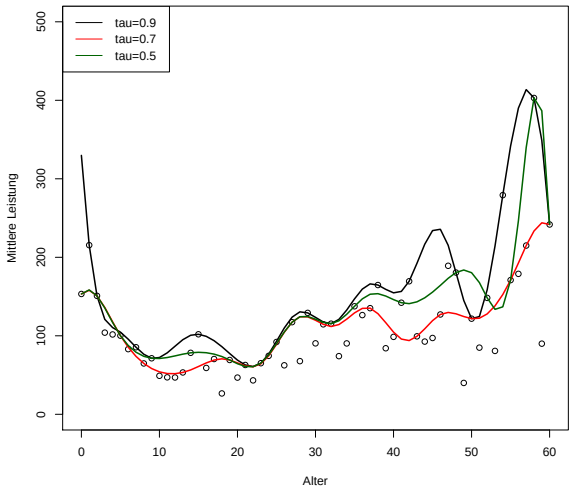


Abbildung 4.24: QR-Modelle für die männlichen gemittelten Leistungen von TKZ = 1 des Jahres 2010 zu den Quantilniveaus $\tau \in \{0.50, 0.70, 0.90\}$.

Fitten und Plotten des QR-Modells

```
plot(Alter, Leistung)
points(alter.m, leistung.m)
X <- model.matrix(Leistung~bs(Alter, df=11))
fit <- rq(Leistung~bs(Alter, df=11), tau=0.9)
qr.fit <- X %*% fit$coef
lines(Alter, qr.fit)
```

Abbildung 4.25 zeigt die Schätzung eines nichtparametrischen QR-Modells zum Niveau $\tau = 0.90$ für die männlichen Leistungen der Tarifklasse 1 aus dem Kalenderjahr 2010. Zum Vergleich werden die gemittelten Leistungen aufgetragen.

Da die Schätzung für das hohe Quantilniveau von $\tau = 0.9$ deutlich über den jeweiligen Mittelwerten liegt, schaut die Schätzung durchaus vernünftig aus. Prinzipiell ist ein steigender Leistungsbedarf mit zunehmendem Alter erkennbar, ausgenommen einer relativ hohen Schätzung für Kleinkinder (mit einem Maximum von 532.44 für Einjährige). Für das maximale Alter von 60 Jahren erhält man eine sehr große Schätzung, wobei jedoch hinzugefügt werden muss, dass hier die Daten bereits relativ dünn sind und sich nur noch fünf Beobachtungen im Datensatz befinden. Die Qualität der Schätzungen sowohl durch Erwartungswertmodelle als auch durch Regressionsquantilen leiden also für den oberen Altersbereich aufgrund einer geringen Dichte der Daten. Der AIC-Wert des QR-Modells entspricht 80419.64.

Abbildung 4.26 zeigt die Schätzung von QR-Modellen zu unterschiedlichen Quan-

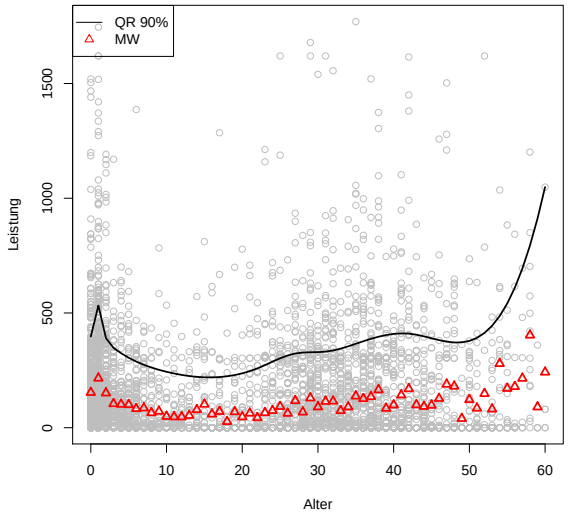


Abbildung 4.25: Nichtparametrisches QR-Modell zum Niveau $\tau = 0.90$ für die männlichen Leistungen von TKZ = 1 des Jahres 2010.

tilniveaus. Zum Vergleich dazu wird ein loglineares Gamma-GAM der ungemittelten Daten aufgetragen. Erfreulicherweise tritt das Problem des *Quantile Crossings* nicht auf. Man erkennt eine gewisse Parallelität der unterschiedlichen QR-Modelle. Dies bedeutet, dass der Einfluß vom Alter in jedem Leistungssegment annähernd gleichbleibend eingeschätzt wird. Das Modell für den Median (also das 50%-Quantil) liegt deutlich unter den Mittelwerten der Leistungen. Dies hängt damit zusammen, dass viele Leistungen den Wert null besitzen, das heißt, der betroffene Versicherte hat keinerlei Leistungsansprüche wahrgenommen.

Nun wird ein QR-Modell zum Niveau $\tau = 0.90$ für die weiblichen Leistungen der Tarifklasse 2 im Kalenderjahr 2010 geschätzt. Abbilung 4.27 zeigt die erhaltene Quantilsschätzung. Sie liegt erst ab der Altersstufe 20 deutlich über den gemittelten Leistungen. Für Kleinkinder ist auch bei den Frauen ein leichter Anstieg mit einem Maximum von 252.80 für 4-Jährige zu erkennen. Der größte Wert des „Geburten-Hügels“ wird mit 2844.21 für 32-jährige Frauen geschätzt. Generell ist wieder ein steigender Trend der geschätzten 90% Quantile der Leistungen mit einem Maximum von 7851.93 für 85-Jährige feststellbar. Da sich immer noch 306 Beobachtungen für weibliche Versicherte im Alter von 85 Jahren im Datensatz befinden, kann man hier von einer vernünftigen Schätzung ausgehen. Der AIC-Wert des Modells entspricht 443487.80.

Nun werden auch für den weiblichen Datensatz mehrere geschätzte Quantile miteinander verglichen. Da jedoch so viele nullen in den Leistungen auftreten, lässt sich aus numerischen Gründen das Modell für die 50%-Quantile hier nicht berechnen.

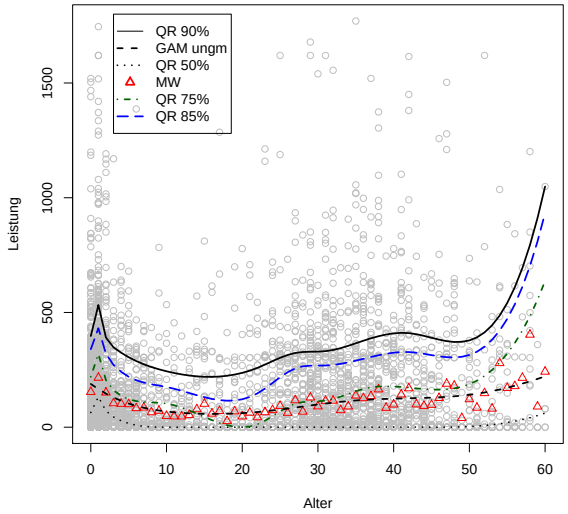


Abbildung 4.26: Vergleich von nichtparametrischen QR-Modellen für die männlichen Originaldaten von TKZ = 1 des Jahres 2010. Zusätzlich ist ein log-lineares GAM der selben Daten aufgetragen.

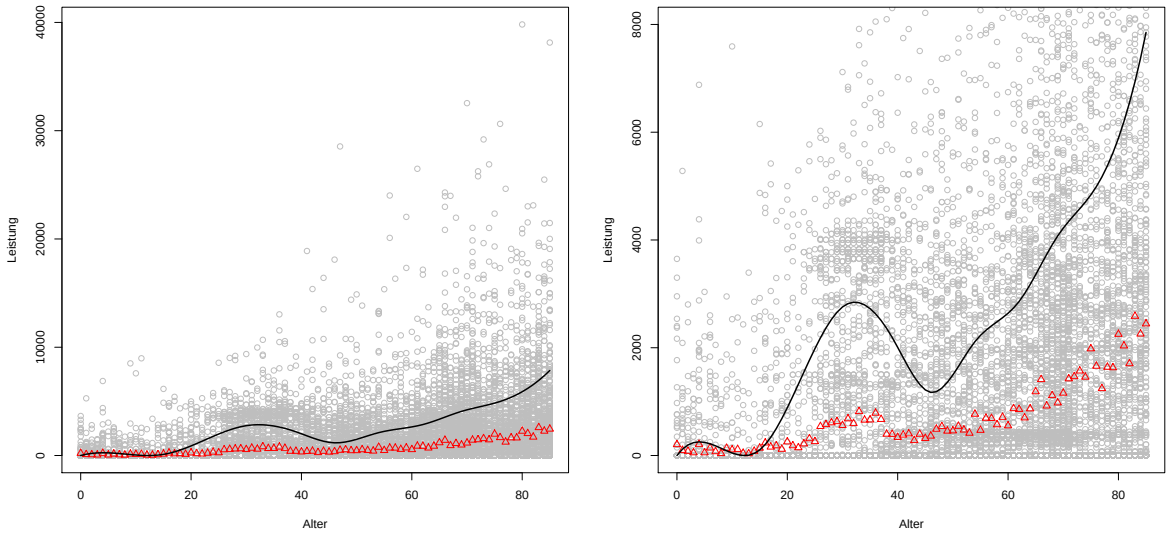


Abbildung 4.27: Nichtparametrisches QR-Modell zum Niveau $\tau = 0.90$ für die weiblichen Leistungen der Tarifklasse 2 des Kalenderjahres 2010 (links) und der gezoomte Bereich (rechts).

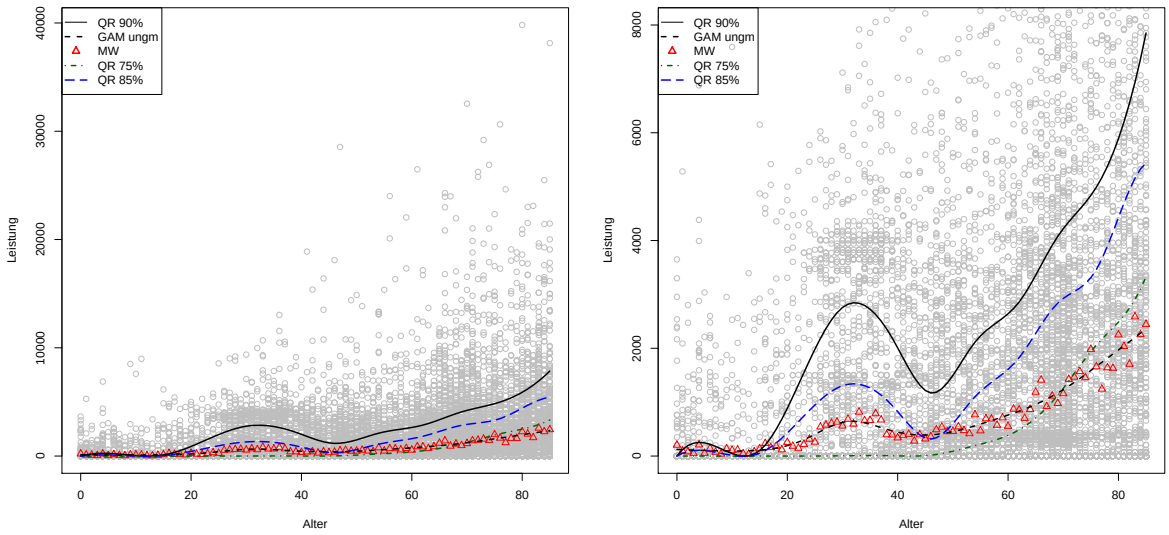


Abbildung 4.28: Vergleich unterschiedlicher QR-Modelle und einem log-linearen Gamma-GAM der ungemittelten weiblichen Leistungen für TKZ = 2 des Jahres 2010 (links) und gezoomte Darstellung (rechts).

Erfreulicherweise ist auch für die weiblichen Originaldaten der Tarifklasse 2 des Jahres 2010 kein *Quantile-Crossing* feststellbar (die sichtbaren Kreuzungen sind mit dem GAM, welche jedoch kein Problem darstellen). Der „Geburten-Hügel“ ist in Abbildung 4.28 erst für größere Quantilniveaus deutlich erkennbar (ab $\tau = 0.85$). Daraus lässt sich schließen, dass Geburten auf niedrigere Quantile offensichtlich kaum Einfluß haben. Durch Geburten verursachte Kosten sind also generell sehr hoch (im Vergleich zu den anderen Kosten). Das 75%-Quantil ist erst ab einem Alter von 50 deutlich von null verschieden. Dies bestätigt die Vermutung, dass sehr viele versicherte Personen in dieser Tarifklasse keine Leistungen angefordert haben. Die unterschiedlichen Quantilsmodelle zeigen ab einem Alter von 50 Jahren ein annäherend paralleles Verhalten, wodurch der Prädiktor Alter auf jedes Leistungssegment einen gleichmäßigen Einluß hat.

Zum Abschluß werden die geschätzten Quantilsfunktion noch mit den jeweiligen Stichprobenquantilen verglichen. Als Niveau wird dafür $\tau = 0.90$ betrachtet.

Mittels der R-Funktion `tapply()` lassen sich auch die Stichprobenquantile für das jeweilige Alter berechnen.

Fitten und Plotten des QR-Modells

```
1 tapply(Leistung, Alter, quantile, probs=0.9)
```

Durch den Parameter `probs` wird das Quantilniveau festgelegt, das für die Leistungen der jeweiligen Altersstufen berechnet werden soll.

Abbildung 4.29 zeigt den Vergleich der geschätzten Quantilsfunktionen zum Niveau $\tau = 0.90$ für die männlichen Leistungen der Tarifklasse 1 und die weiblichen Leistungen der Tarifklasse 2 des Jahres 2010 mit den 90%-Stichprobenquantilen der jeweiligen Altersstufen. Die geschätzte Quantilsfunktion passt sich den jeweiligen

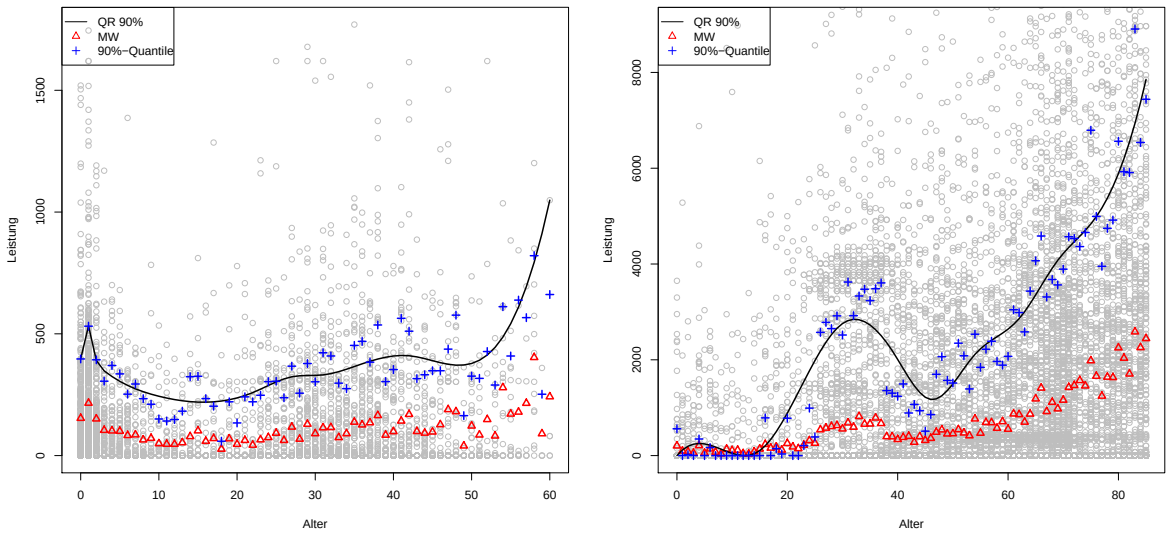


Abbildung 4.29: Vergleich der nichtparametrischen QR-Modelle der männlichen Leistungen der Tarifklasse 1 (links) und der weiblichen Leistung aus Tarifklasse 2 (rechts), jeweils aus dem Jahr 2010. Schwarz: Geschätzte 90%-Quantilsfunktion. Rot: Mittelwerte der jeweiligen Altersstufe. Blau: 90% Stichprobenquantile für die jeweilige Altersstufe.

Stichprobenquantilen ziemlich gut an, wobei die mittels QR geschätzten Quantile ein glatteres Verhalten als die „rohen“ Quantile der Daten aufweisen.

## 4.7 Betrachtung der Leistungen über die Kalenderjahre

Es soll nun eine Veränderung der Leistungen über die Kalenderjahre betrachtet werden, um feststellen zu können, ob ein Trend über die Jahre erkennbar ist. Zu diesem Zweck werden die männlichen Leistungen für die Tarifklasse 1 und die weiblichen Aufwendungen der Tarifklasse 2 verwendet.

Stellvertretend werden GAMs betrachtet. Dafür wird der Prädiktor `Jahr` ins Modell aufgenommen.

Betrachten der Leistungen über die Jahre

```
library(mgcv)
fit <- gam(Leistung.1~s(Alter.1) + as.factor(Jahr.1),
           family = quasi(link = log, variance="mu^2"))
summary(fit)
```

```
Family: quasi
Link function: log

Formula:
Leistung.1 ~ s(Alter.1) + as.factor(Jahr.1)

Parametric coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)            4.70877    0.02970 158.533   <2e-16 ***
as.factor(Jahr.1)2011  0.09582    0.03829   2.503   0.0123 *
as.factor(Jahr.1)2012  0.08272    0.03619   2.285   0.0223 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
             edf Ref.df     F p-value
s(Alter.1) 7.528  8.424 66.87  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.0276   Deviance explained = 2.15%
GCV score = 4.6977  Scale est. = 4.6957    n = 24413
```

Hierbei wird angenommen, dass es für die drei Kalenderjahre eine Struktur gibt, die sich nur multiplikativ unterscheidet.

Aufgrund der Log-Linkfunktion gilt für den Erwartungswert $\mu$ und den linearen Prädiktor $\eta$ folgender Zusammenhang:

$$\log \mu = \eta \Leftrightarrow \mu = \exp(\eta).$$

Im Output des Befehls `summary()` beschreibt der Intercept den geschätzten Referenzwert des Jahres 2010, $\alpha_{2011} :=$ `as.factor(Jahr.1)2011` $\approx 0.10$ bezeichnet die logarithmierte Änderung des Jahr 2011 und $\alpha_{2012} :=$ `as.factor(Jahr.1)2012` $\approx 0.08$ für das Jahr 2012, jeweils verglichen mit dem Jahr 2010.

Für die Erwartungswertschätzung des Jahres 2011 gilt:

$$\hat{\mu}_{2011} = \exp(\alpha_{2011}) \cdot \widehat{\tilde{f}(Alter)} \cdot \exp(4.71) = \exp(\alpha_{2011}) \cdot \mu_{2010} \approx 1.10 \cdot \hat{\mu}_{2010}.$$

Die Ausgaben der Versicherung im Jahre 2011 liegen um etwa 10% höher als im Jahre 2010 bei gleichem Altersverlauf. Betrachtet man den p-Wert, erkennt man die Signifikanz des Unterschieds zwischen den Erwartungswertverläufen der Jahre 2010 und 2011: Der angegebene Wert von circa 0.01 entspricht dem p-Wert des zweiseitigen Tests

$$\mathrm{H}_0 : \exp(\alpha_{2011}) = 1 \text{ gegen } \mathrm{H}_1 : \exp(\alpha_{2011}) \neq 1,$$

oder äquivalent

$$\mathrm{H}_0 : \alpha_{2011} = 0 \text{ gegen } \mathrm{H}_1 : \alpha_{2011} \neq 0.$$

Gleichbedeutend könnte man diesen Test aufgrund der Modellannahmen auch folgendermaßen aufstellen:

$$\mathrm{H}_0 : \log(\mu_{2010}) - \log(\mu_{2011}) = 0 \text{ gegen } \mathrm{H}_1 : \log(\mu_{2010}) - \log(\mu_{2011}) \neq 0.$$

Für den viel interessanteren einseitigen Test $\mathrm{H}_0$: $\exp(\alpha_{2011}) \leq 1$ gegen $\mathrm{H}_1$: $\exp(\alpha_{2011}) > 1$ ergibt sich ein p-Wert von circa $5 \cdot 10^{-3}$. Dieser ist höchstsignifikant und spricht daher dafür, dass die Erwartungswertschätzung des Jahres 2011 höher ist als jene des Jahres 2010.

Für das Jahr 2012 ergibt sich

$$\hat{\mu}_{2012} = \exp(\alpha_{2012}) \cdot \hat{\mu}_{2010} \approx 1.09 \cdot \hat{\mu}_{2010}.$$

Dies entspricht einer circa 9% höheren Kurve für das Jahr 2012 im Vergleich zum Jahr 2010. Der einseitige Hypothesentest ergibt einen p-Wert von circa 0.01, was wieder für einen signifikanten Anstieg der Erwartungswertschätzung des Jahres 2012 im Vergleich zum Jahr 2010 spricht.

Diese Ergebnisse sind auch in Abbildung 4.30 ersichtlich. Zum Vergleich werden die gemittelten Leistungen der jeweiligen Jahre geplottet.

Damit kann nachgewiesen werden, dass die geschätzten Leistungen in den beiden Jahren 2011 und 2012 sich signifikant bezüglich jenen im Jahre 2010 erhöht haben.

Nun wird das Modell auf die weiblichen Leistungen der Tarifklasse 2 angewendet.

Betrachten der Leistungen über die Jahre

```
1 library(mgcv)
2 fit <- gam(Leistung.2~s(Alter.2) + as.factor(Jahr.2),
```

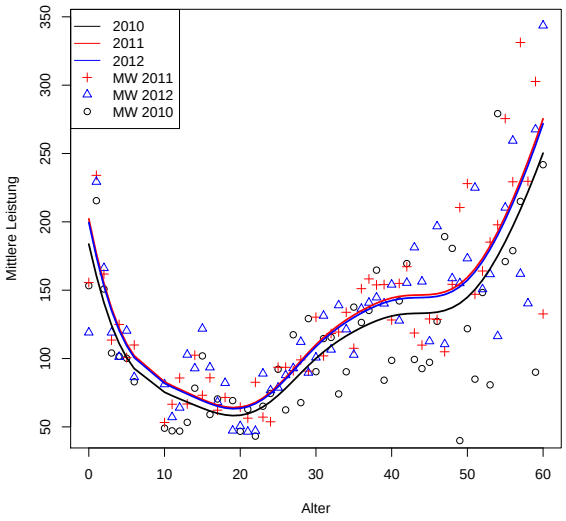


Abbildung 4.30: Betrachtung der männlichen Leistungen der Tarifklasse 1 über die Kalenderjahre 2010, 2011 und 2012 mittels GAM.

```
            family = quasi(link = log, variance="mu^2"))
summary(fit)
```

```
Family: quasi
Link function: log

Formula:
Leistung.2 ~ s(Alter.2) + as.factor(Jahr.2)

Parametric coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)            6.3425592  0.0194399 326.265   <2e-16 ***
as.factor(Jahr.2)2011  0.0342483  0.0275498   1.243    0.214
as.factor(Jahr.2)2012 -0.0004969  0.0276542  -0.018    0.986
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
             edf Ref.df     F p-value
s(Alter.2) 8.777  8.985 682.8  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

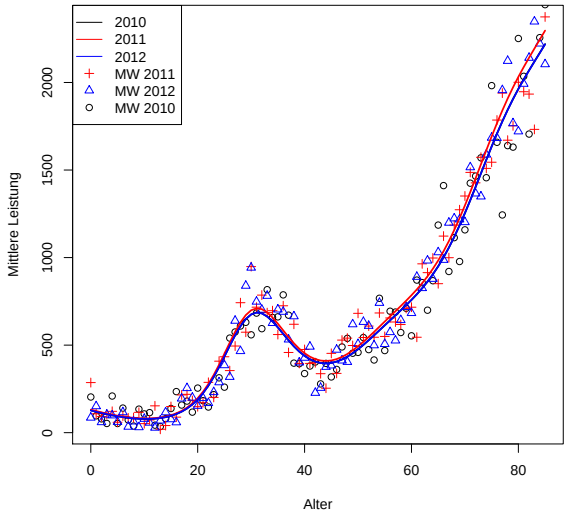


Abbildung 4.31: Betrachtung der weiblichen Leistungen der Tarifklasse 2 über die Kalenderjahre 2010, 2011 und 2012 mittels individueller loglinearer GAMs.

```
R-sq.(adj) =  0.0684   Deviance explained = 6.74%
GCV score = 8.8107  Scale est. = 8.8092    n = 69241
```

Für die weiblichen Leistungen ergibt sich die höchste Schätzung für das Kalenderjahr 2011, gefolgt von den Jahren 2010 und 2012. Abbildung 4.31 zeigt die Unterschiede der jeweiligen Kurven: Die geschätzte Kurve für das Jahr 2011 ist am höchsten, zwischen den anderen beiden Modellen ist kein Unterschied erkennbar. Im Gegensatz zum Ergebnis bei den Männern zuvor, scheint es hier keine relevanten Unterschiede zwischen den zu erwartenden Leistungen in den drei Kalenderjahren zu geben. Dies erkennt man an den jeweiligen p-Werten.

In den bisherigen Vergleichen der Leistungen für die jeweiligen Kalenderjahre wurde die Form der Modellkurve festgehalten. Dies wird sich im Folgenden ändern.

Es wird für jedes Jahr ein eigenes Modell erzeugt. Dafür werden wieder GAMs verwendet. Stellvertretend wird der Code für das Jahr 2010 präsentiert.

Darstellung der Leistung über die Kalenderjahre

```
1 gam(Leistung2010~s(Alter2010), family=quasi(link=log, variance="mu^2"))
```

Durch die getrennte Modellierung der Leistungen für die jeweiligen Kalenderjahre wird eine eigene Form der geschätzten Kurve für jedes Jahr ermöglicht.

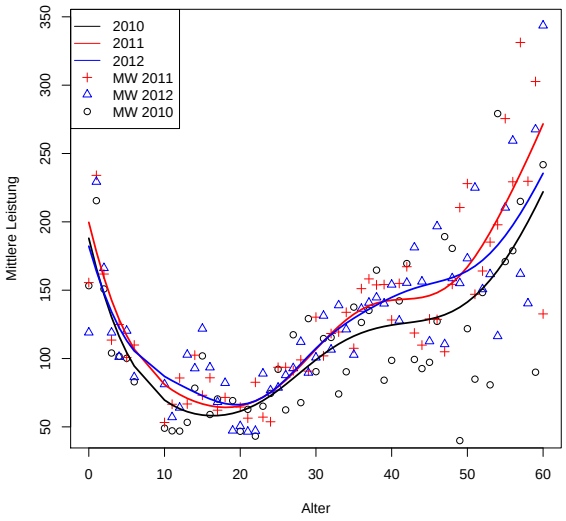


Abbildung 4.32: Betrachtung der männlichen Leistungen der Tarifklasse 1 über die Kalenderjahre 2010, 2011 und 2012 mittels individueller loglinearer GAMs.

Abbildung 4.32 zeigt die erhaltenen Modelle inklusive der jeweiligen gemittelten Leistungen. Das Modell der Leistungen des Kalenderjahres 2010 wird am niedrigsten geschätzt. Der Unterschied für die Jahre 2011 und 2012 ist nicht so offensichtlich. Bei jungen Versicherten erkennt man für die Daten von 2012 eine höhere Schätzung; für Versicherte höheren Alters ist die Schätzung des Jahres 2011 ab einem Alter von circa 50 Jahren am höchsten.

Selbige Modelle werden als nächstes für die weiblichen Leistungen der Tarifklasse 2 über die Kalenderjahre 2010, 2011 und 2012 betrachtet.

Abbildung 4.33 zeigt die Anwendung von GAMs auf die weiblichen Leistungen der Tarifklasse 2. Es sind kaum Unterschiede feststellbar, lediglich für den „Geburten-Hügel“ lässt sich ein niedrigeres Niveau für die Daten des Jahres 2010 erkennen; die Schätzungen für die Jahre 2011 und 2012 sind nahezu identisch.

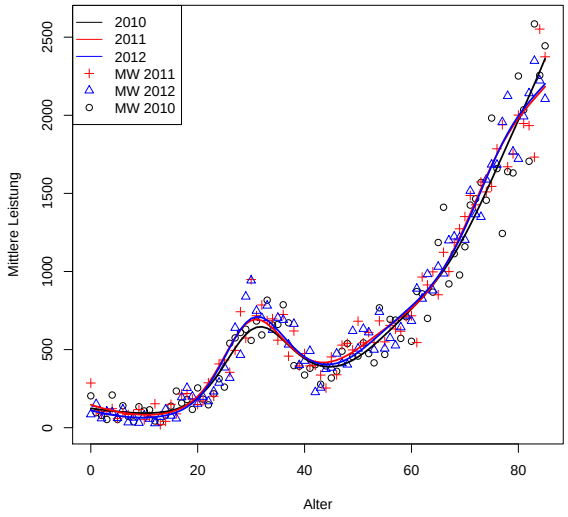


Abbildung 4.33: Betrachtung der weiblichen Leistungen der Tarifklasse 2 über die Kalenderjahre 2010, 2011 und 2012 mittels individueller loglinearer GAMs.

## 4.8 Gemeinsame Betrachtung beider Geschlechter

Solvency II ist eine neue Reform der EU-Komission, die vor allem Solvabilitätsvorschriften für die Eigenmittelausstattung von Versicherungsunternehmen regelt (EU, 2013). Eine darin enthaltene Neuerung ist die *„Unisex-Regelung“*, die es verbietet, geschlechterdifferenzierte Tarife in der Versicherungsbranche anzubieten. Dies begründet ein gemeinsames Betrachten der männlichen und weiblichen Leistungen der jeweiligen Tarifklasse.

Die Kombination der männlichen und weiblichen Leistungen wird im Folgenden als „*globale Daten*“ bezeichnet.

Berechnung und Darstellung der globalen Daten

```
1 leistung.m.global <- tapply(daten$leistung, daten$alter, mean)
2 alter.m.global <- unique(daten$alter)
```

Wenn man in Abbildung 4.34 die gemittelten Leistungen der Tarifklasse 1 betrachtet, erscheinen die weiblichen Daten (+) ab der Altersstufe 20 deutlich höher als die männlichen Werte (△). Für die zweite Tarifklasse (rechts) erkennt man einen sichtbar höheren Trend der weiblichen Leistungen zwischen den Altersstufen 20 und 40 Jahren, den „Geburten-Hügel“.

Zunächst werden *globale kubische Splines* auf die gemeinsamen Daten angewen-

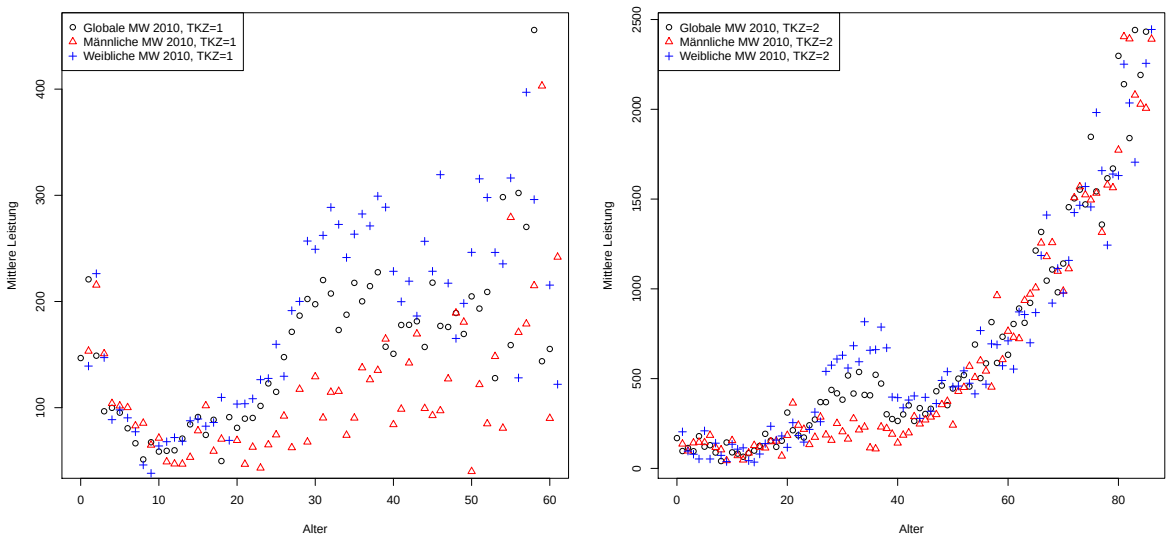


Abbildung 4.34: Vergleich der globale Mittelwerte (○) mit den männlichen (△) und weiblichen Mittelwerten (+) des Kalenderjahres 2010 für die Tarifklassen 1 (links) und 2 (rechts).

det. Abbildung 4.35 zeigt die Anwendung der globalen kubischen Splines auf die globalen Daten der beiden Tarifklassen sowohl für die gemittelten Daten, als auch auf die Originaldaten, inklusive der gemeinsamen Mittelwerte. Für die Tarifklasse 1 sind leichte Unterschiede feststellbar, vor allem im oberen Bereich des Prädiktors `Alter`. Bei Tarifklasse 2 lässt sich praktisch kein Unterschied zwischen den beiden Modellen feststellen. Generell ist ein guter Fit für Tarifklasse 1 feststellbar, während die betrachtete Modellklasse offensichtlich nicht flexibel genug ist, um die Daten der Tarifklasse 2 zufriedenstellend zu schätzen. Betrachtet man sich die jeweiligen 95% Konfidenzintervalle über den gesamten Bereich des Prädiktors `Alter`, so sind diejenigen für das Modell der Originaldaten enger als jene bei Anwendung der gemittelten Leistungen. Tabelle 4.4 zeigt die AIC-Werte der jeweiligen Modelle.

| TKZ | Global, Gemittelt | Global, Ungemittelt |
|---|---|---|
| 1 | 647.57 | 167919.90 |
| 2 | 1078.18 | 749909.50 |

Tabelle 4.4: AIC-Werte der globalen kubischen Splines der globalen Modelle des Jahres 2010.

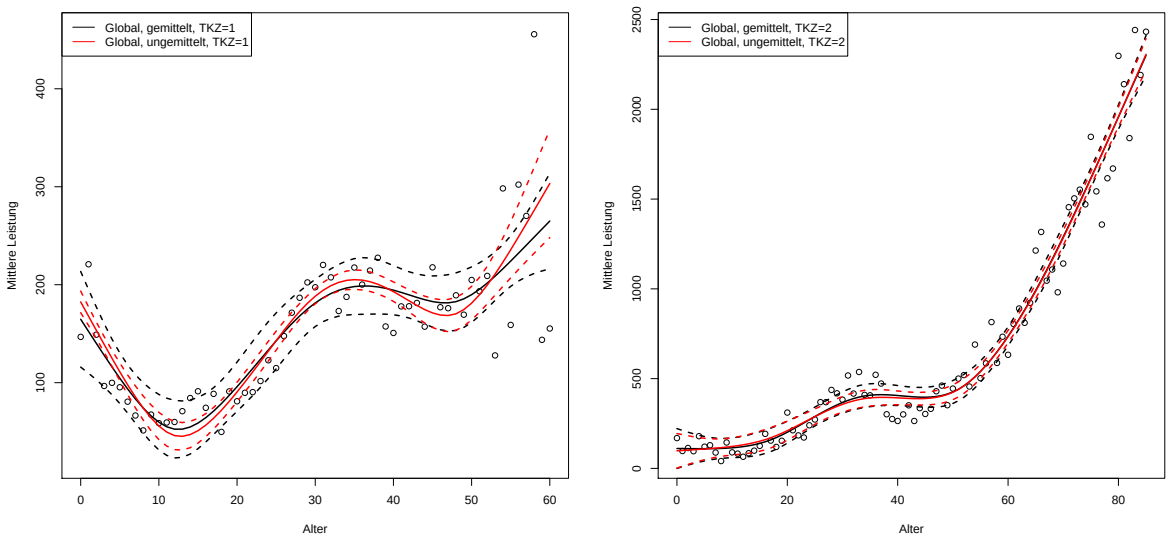


Abbildung 4.35: Globale kubische Splines für die beiden Tarifklassen (TKZ = 1 links, TKZ = 2 rechts) der globalen Daten des Kalenderjahres 2010 samt der jeweiligen 95% Konfidenzintervalle bei Anwendung auf die gemittelten (schwarz) und ungemittelten (rot) Leistungen.

Zur Anwendung der *Penalisierten Regression-Splines* muss der optimale Glättungsparameter $\hat{\lambda}$ durch Minimierung des GCV-Scores berechnet werden.

Zunächst werden die globalen Daten der Tarifklasse 1 für das Jahr 2010 betrachtet.

Der als optimal berechnete Wert ist $\hat{\lambda} = 10^{-8} \cdot 1.5^{27} \approx 5.68 \cdot 10^{-4}$, was also kaum Bestrafung zufolge hat. Man erkennt in Abbildung 4.36 (links) eine gute Schätzung der Daten. Zum Vergleich werden die Mittelwerte der Leistungen aufgetragen. Der AIC-Wert des Modells beträgt 738.19. Da der Rechenaufwand der Minimierung des GCV-Scores ziemlich hoch ist, lässt sich diese Modellklasse jedoch nicht auf die Originaldaten anwenden.

Für die globalen Daten der Tarifklasse 2 ergibt sich durch Anwendung der Penalisierten Regression-Splines ein optimaler Glättungsparameter von $\hat{\lambda} \approx 2.50 \cdot 10^{-4}$. Abbildung 4.36 (rechts) zeigt eine ziemlich gute Schätzung des Modells bei Anwendung der gemittelten, globalen Daten. Auch der „Geburten-Hügel“ lässt sich mit Hilfe dieser Modellklasse vernünftig schätzen. Der AIC-Wert des Modells ergibt 1179.83.

Nun werden auch die *P-Splines* auf die globalen Daten angewendet. Dazu wird abermals die Funktion `gam()` verwendet.

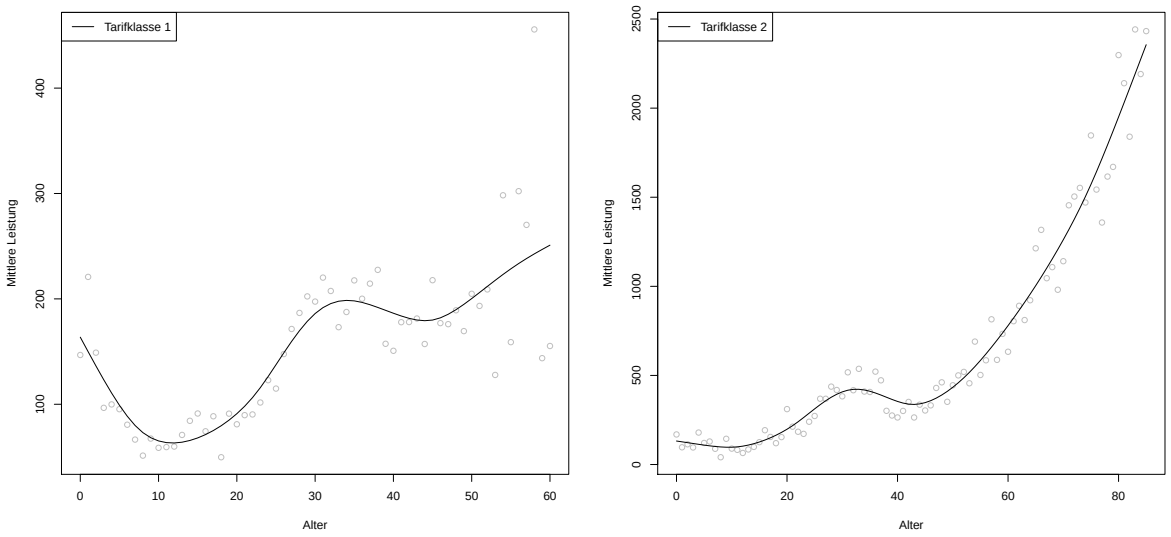


Abbildung 4.36: Penalisierte Regression-Splines der globalen gemittelten Daten für beide Tarifklassen des Jahres 2010 bei Anwendung auf die gemeinsamen Leistungen.

P-Splines bezüglich der globalen Daten

```
1 gam(leistung.m.global~s(alter.m.global, bs="ps"))
2 gam(leistung.global~s(alter.global, bs="ps"))
```

Abbildung 4.37 zeigt die Anwendung der P-Splines auf die globalen Daten beider Tarifklassen sowohl auf die Originaldaten, als auch die gemittelten Leistungen. Für Tarifklasse 1 ist ein deutlicher Unterschied der beiden Kurven im oberen Bereich zu erkennen. Dies ist damit zu begründen, dass die beiden niedrigen Mittelwerte das gemittelte Modell nach unten ziehen, was für die ungemittelten Leistungen nicht auftritt. Hier ist also das Modell der Originaldaten zu bevorzugen, da es praktisch keinen Sinn macht, dass die Leistungen für ein Alter von 60 Jahren derart nach unten gezogen werden.

Werden die P-Splines auf die Leistungen der Tarifklasse 2 angewendet, erkennt man kaum Unterschiede bei Anwendung auf die Originaldaten oder auf die gemittelten Leistungen. Es lässt sich für beide Modelle eine gute Schätzung feststellen. Tabelle 4.5 zeigt die AIC-Werte der globalen P-Spline Modelle.

Als nächstes werden *GAM*s auf die globalen Daten angewendet. Dazu wird wieder mit der Funktion `quasi()` die Gammaverteilung als Modell verwendet.

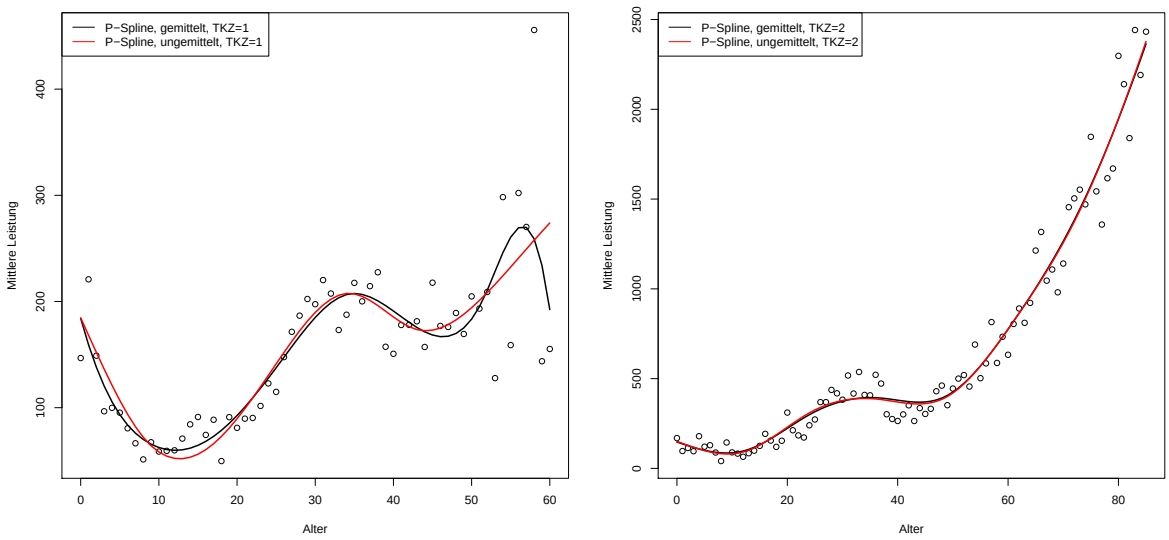


Abbildung 4.37: P-Splines der globalen Leistungen beider Tarifklassen (TKZ = 1 links, TKZ = 2 rechts) des Jahres 2010.

| TKZ | Global, Gemittelt | Global, Ungemittelt |
|---|---|---|
| 1 | 641.83 | 167914.80 |
| 2 | 1074.21 | 749900.60 |

Tabelle 4.5: AIC-Werte der P-Splines der globalen Modelle des Jahres 2010.

GAM bezüglich der globalen Daten

```
gam(leistung.global~s(alter.global, bs="cr"),
    family = quasi(link = log, variance="mu^2"))
```

Abbildung 4.38 zeigt die Anwendung von GAMs auf die globalen Daten der Tarifklasse 1. Es wird sowohl das Modell der gemittelten Leistungen, als auch jenes der Originaldaten abgebildet. Es gibt kaum Unterschiede zwischen den beiden Modellen. Lediglich am oberen Ende der Altersstufen ist eine minimale Abweichung der beiden Modelle ersichtlich.

Zum Vergleich der beiden Modelle wird der geschätzte Freiheitsgrad des Glättungsterms betrachtet. Diesen erhält man durch den `summary()`-Output der Modelle.

```
Family: quasi
Link function: log

Formula:
```

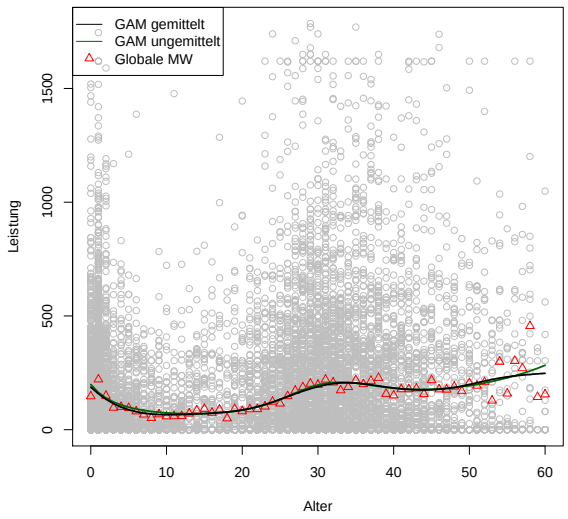


Abbildung 4.38: GAM für globale Leistungen des Jahres 2010 der Tarifklasse 1

```
leistung.global ~ s(alter.global, bs = "cr")

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.00304    0.01921   260.4   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
                         edf Ref.df     F p-value
s(alter.global) 6.856  7.891 44.92  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.0311   Deviance explained = 2.53%
GCV score =  4.454  Scale est. = 4.4511    n = 12058
```

Für das Modell der Originaldaten wird ein EDF von 6.86 grschätzt; er scheint also äußerst signifikant zu sein. Das durch die gemittelten Leistungen erzeugte Modell liefert eine Schätzung des EDFs von 7.18 und somit dasselbe Ergebnis.

Im Folgenden werden loglineare GAMs auf die globalen Daten des Kalenderjahres 2010 für die Tarifklasse 2 angewendet.

Abbildung 4.39 zeigt die Schätzungen des GAMs für die Originaldaten, als auch

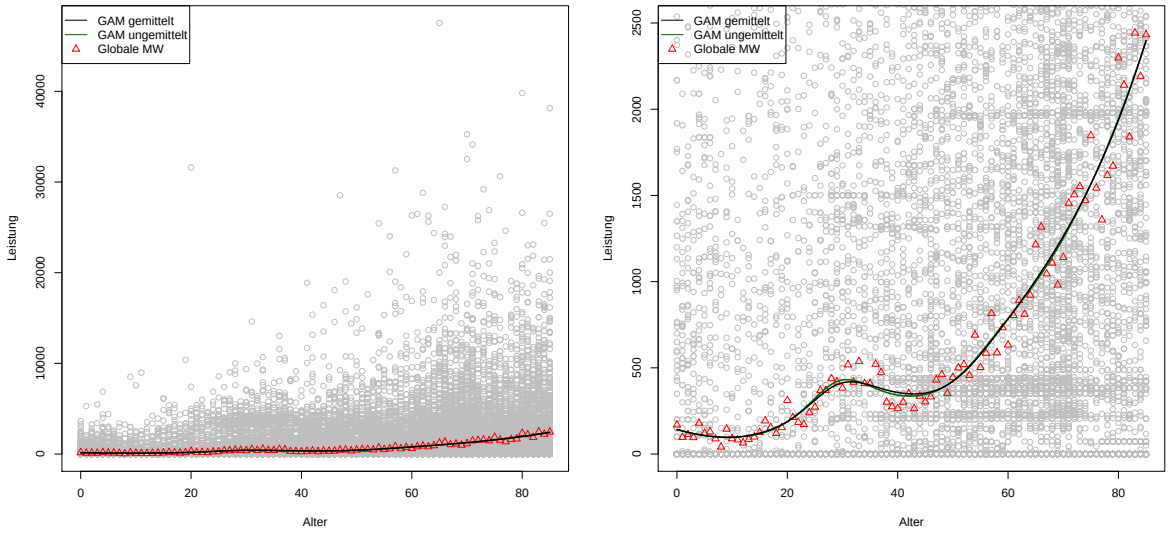


Abbildung 4.39: GAM für globale Leistungen der Tarifklasse 2 für das Kalenderjahr 2010 (links) und gezoomter Ausschnitt (rechts).

die gemittelten Daten. Um einen besseren Blick auf die geschätzten Modelle werfen zu können, wird der betrachtete Bereich in der rechten Graphik eingeschränkt. Es lassen sich kaum Unterschiede der beiden Modelle feststellen. Die Schätzung der ungemittelten Daten fällt im Bereich des „Geburten-Hügels“ etwas höher aus. Beide Modelle passen sich deutlich an die gemittelten Leistungen an. Die `summary()`-Ausgabe des Modells der Mittelwerte zeigt, das der Glättungsterm des Modells signifikant ist, da der *Expected Degree of Freedom* einen Wert von 6.07 besitzt. Der Anteil der durch das Modell erklärten Deviance ist mit 95.2% sehr hoch, was ein weiterer Indikator für die Qualität des Modells ist. Auch für das Modell der Originaldaten erkennt man, dass die Parameter höchstsignifikant sind (EDF = 6.18). Der Anteil der erklärten Deviance des Modells ist 7.21%.

Nun werden Modelle der *Quantilen Regression* auf die globalen Daten angewendet. Aus den bereits diskutierten Gründen werden hier nur die Originaldaten betrachtet.

Schätzen und Plotten von QR-Modellen

```
1 X <- model.matrix(leistung.global~bs(alter.global, df=11))
2 fit <- rq(leistung.global~bs(alter.global, df=11), tau=0.9)
```

Da die Mehrheit der Leistungen null sind, ist der geschätzte Median in Abbildung 4.40 durchgehend sehr nahe dem Nullniveau. Grundsätzlich ist ein steigender Leistungsanspruch mit zunehmendem Alter zu erkennen. Für einjährige Versicherte und Personen im Alter von 25-35 Jahre erkennt man einen deutlichen Anstieg der Schätzung. Diese Effekte sind für höhere Quantilniveaus besser sichtbar, wodurch diese Wirkung

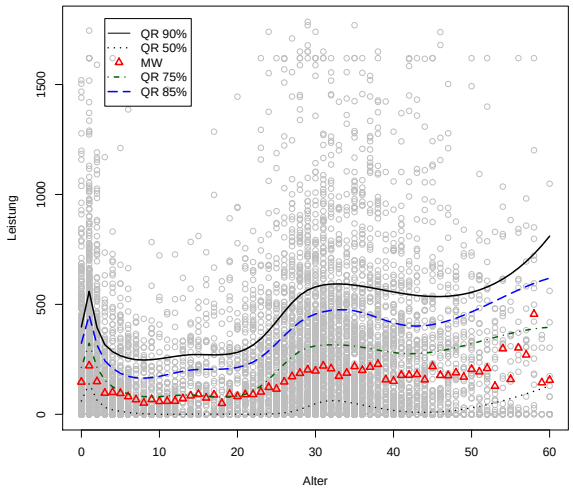


Abbildung 4.40: QR-Modelle zu den Niveaus $\tau \in \{0.50, 0.75, 0.85, 0.90\}$ der globalen Daten des Kalenderjahres 2010 für die Tarifklasse 1.

offensichtlich das obere Leistungssegment stärker beeinflusst. Es ist erfreulicherweise kein Quantile-Crossing feststellbar. Ab einem Niveau von $\tau = 0.75$ sind die geschätzten QR-Modelle deutlich über den gemittelten Leistungswerten der Daten. Die QR-Modelle zeigen ein paralleles Verhalten für Niveaus ab 75%, weshalb das Alter wieder auf diese Leistungssegmente einen gleichmäßigen Einfluß besitzt. Der AIC-Wert für das 90%-Quantil entspricht 180647.20.

Nun werden selbige Modelle für die Tarifklasse 2 gebildet.

Da das Niveau sämtlicher Modelle sehr niedrig ist, lässt sich in Abbildung 4.41 kaum etwas erkennen. Darum werden die QR-Modelle in der rechten Graphik nochmals vergrößert dargestellt.

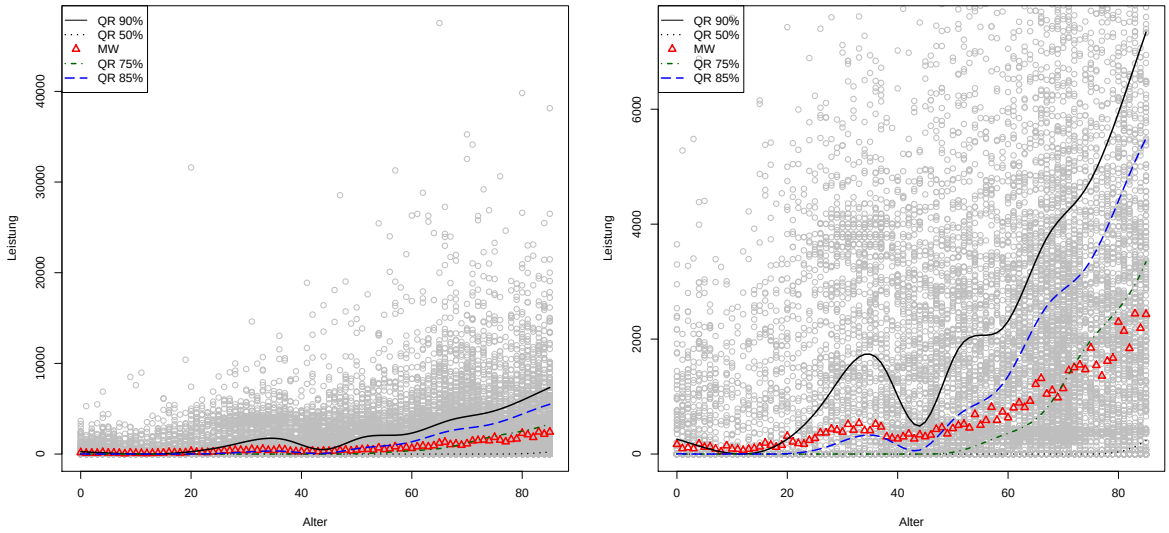


Abbildung 4.41: Diverse QR-Modelle der globalen Daten des Kalenderjahres 2010 für die Tarifklasse 2 (links) und Ausschnitt (rechts).

Erfreulicherweise ist wieder kein „Quantile Crossing“ feststellbar. Das 50%-Quantil ist praktisch null über den gesamten Bereich der Daten. Das 75%-Quantil ist bis zur Altersklasse 50 null, besitzt dann aber einen steigenden Trend, annäherend parallel den Schätzungen für größere $\tau$-Werte. Für das 85%-Quantil ist zum ersten mal ein Anstieg durch den „Geburten-Hügel“ erkennbar, dessen Einfluss jedoch für das 90%-Quantil noch stärker sichtbar wird. Die Modelle mit einem höheren Quantil-Niveau zeigen eine parallele Tendenz, wobei dieses Verhalten nicht mehr so stark erkennbar ist wie für Tarifklasse 1. Der AIC-Wert des 90% QR-Modells entspricht 784288.00.

## 4.9 Betrachtung unterschiedlicher Leistungsarten der Tarifklasse 2

Es wird nun die Tarifklasse 2 der Spitalskosten bezüglich unterschiedlicher Leistungsarten (`LKZ`) unterschieden. Dabei wird keine Unterteilung bezüglich dem Geschlecht vorgenommen. Man kann zeigen, dass der Großteil der Kosten dieser Tarifklasse durch die Leistungsarten 1, 2 und 3 entstehen. Diese entsprechen den *Krankenhaus-Pflegegebühren*, den *Krankenhaus-Operationsgebühren* und den *Internen Krankenhausbehandlungskosten*.

Durch die Betrachtung der jeweiligen Leistungsarten müssen Leistungen der Höhe null aus dem Datensatz entfernt werden, da diese nichtauftretenden Leistungsansprüche keiner Leistungsart zugeteilt werden können.

Die Leistungsposten 1 und 3 werden gemeinsam modelliert, da sie vorwiegend zusammen auftreten, während Leistungsart 2 gesondert modelliert wird. Da für höhere Alter nur mehr wenige Beobachtungen vorhanden sind, betrachtet man die Daten abermals lediglich bis zu einem Alter von 85 Jahren. Zunächst werden die jeweiligen Daten des Kalenderjahres 2010 analysiert. Es wird dabei nicht mehr geschlechterspezifisch unterschieden, sondern man betrachtet die globalen Daten.

Für die Leistungen der Leistungsarten 1/3 befinden sich 4383 Beobachtungen im Datensatz. Man erhält anhand des `summary()`-Befehls folgenden Ausdruck:

```
Min. 1st Qu.  Median   Mean   3rd Qu.  Max.
 23    2097    7768    7165   10360   15640
```

Bereits das erste Quartil der Leistungen zeigt ein sehr hohes Niveau an. Es sind also größtenteils sehr hohe Leistungsansprüche in diesem Datensatz vorhanden. Für Leistungsart 2 befinden sich 4263 Beobachtungen im Datensatz. Man erhält folgende statistische Kenngrößen:

```
Min.  1st Qu. Median   Mean   3rd Qu.  Max.
 26    2136    6319    6986   10350   15640
```

Auch hier werden also sehr hohe Leistungsansprüche beobachtet. Auffallend ist weiters, dass für die Leistungsarten 1/3 beziehungsweise 2 dasselbe Maximum von 15640 auftritt.

Abbildung 4.42 stellt die gemittelten Leistungen der gesamten Tarifklasse 2, sowie gesondert die Mittelwerte für die Leistungsarten 1/3 und der Leistungsart 2 dar. Man erkennt einen relativ konstanten Verlauf der gemittelten Leistungen über den gesamten Prädiktor `Alter`. Für Altersstufen kleiner 20 Jahre reißen die Mittelwerte der Leistungsklasse 2 etwas nach unten und die Mittelwerte der Leistungsklassen 1/3 etwas nach oben aus.

Im Folgenden werden die Leistungsposten bezüglich der betrachteten Modellklassen analysiert. Zunächst werden dafür die *globalen kubischen Splines* auf die Daten angewendet.

Abbildung 4.43 zeigt die Schätzung durch Anwendung der globalen kubischen Splines auf die gemittelten und ungemittelten Daten für die Leistungsarten 1/3 (links) und 2 (rechts).

Für die Modelle der Leistungsarten 1/3 links sind kaum Unterschiede erkennbar. Lediglich im unteren Bereich des Prädiktors `Alter` ist die Schätzung der gemittelten Daten etwas höher. Die Modelle haben einen leichten Abwärtstrend. Sie sind für niedrige Altersstufen etwas höher als die allgemeinen Mittelwerte der Tarifklasse.

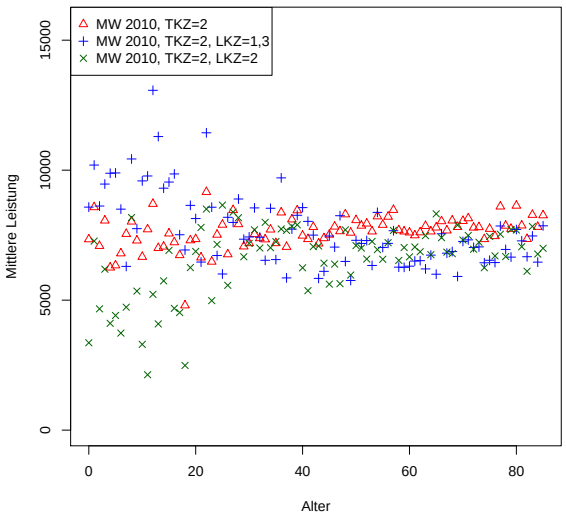


Abbildung 4.42: Mittelwerte der globalen Daten der gesamten Tarifklasse 2 (△), der Leistungsarten 1/3 (+) sowie der Leistungsart 2 (×) des Jahres 2010.

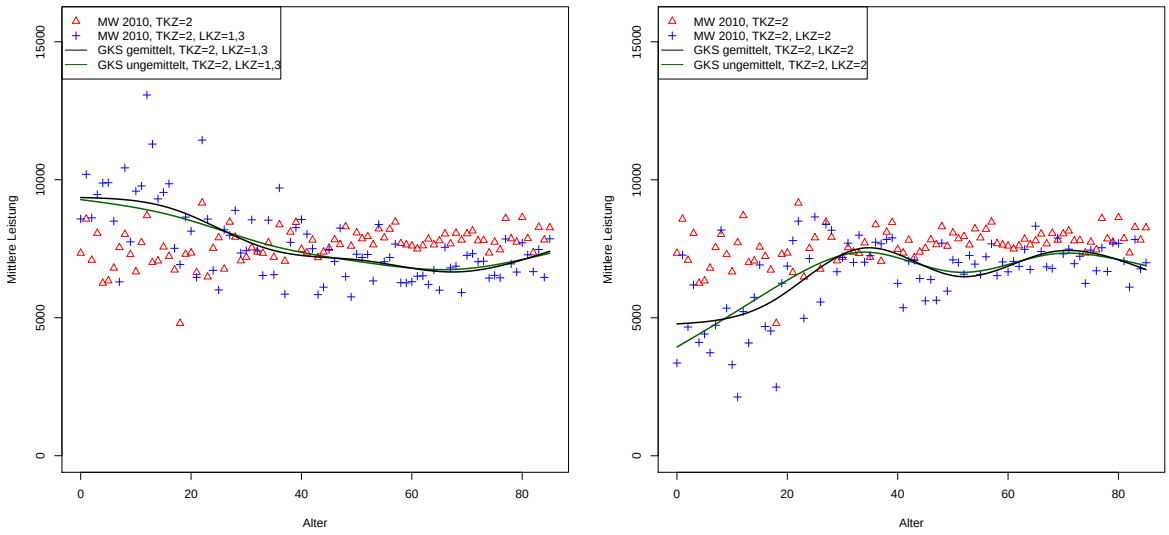


Abbildung 4.43: Globale kubische Splines der globalen Daten der Tarifklasse 2: Links die Modelle der Leistungsarten 1/3, rechts für Leistungsart 2 bei Anwendung auf die gemittelten (grün) und ungemittelten (schwarz) Leistungen. Zusätzlich werden die globalen Mittelwerte der gesamten Tarifklasse 2 (△) und jene der Leistungsarten 1/3 (+) abgebildet.

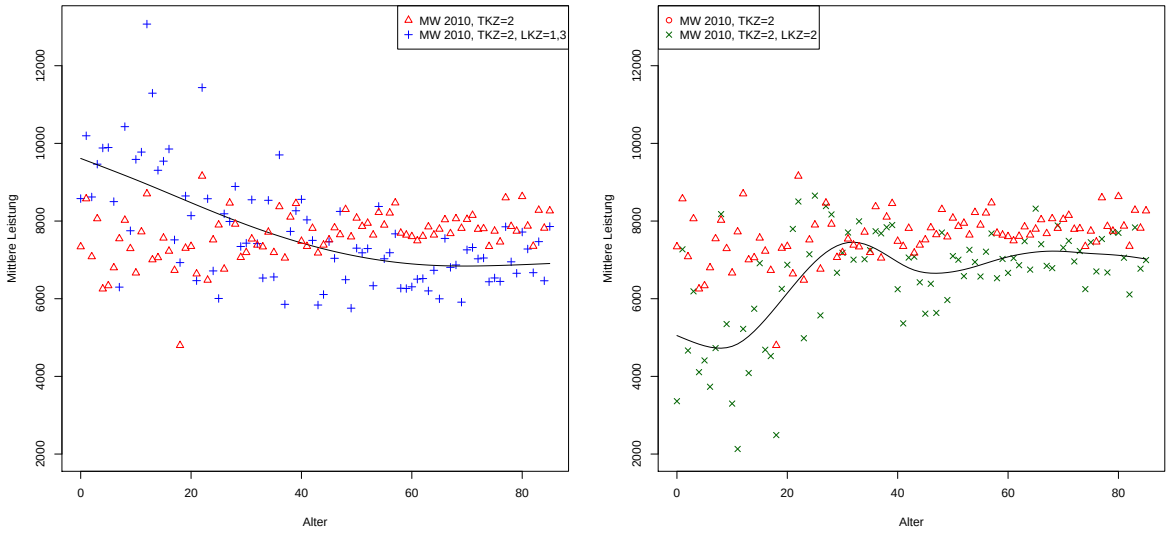


Abbildung 4.44: Penalisierte Regression Splines der globalen Daten der Tarifklasse 2: Links das Modell der Leistungsarten 1/3, rechts für Leistungsart 2 bei Anwendung auf die gemittelten Leistungen. Zusätzlich werden die globalen Mittelwerte der gesamten Tarifklasse 2 (△) sowie jene der Leistungsarten 1/3 (+) links und der Leistungsart 2 (×) rechts abgebildet.

Für versicherte Personen höheren Alters gleichen sich die Schätzungen wieder an die globalen Mittelwerten der Tarifklasse 2 an.

Für die Leistungen der Leistungsart 2 erkennt man steigende Modelle für zunehmende Altersstufen. Im unteren Bereich wird das Modell für die Mittelwerte etwas höher geschätzt. Allgemein lässt sich kaum ein Unterschied zwischen den beiden Modellen feststellen. Die Schätzungen bezüglich der Leistungsart 2 haben generell ein niedrigeres Niveau als die globalen Mittelwerte der Tarifklasse 2.

Nun werden die *Penalisierten Regression Splines* auf die Daten angewendet; zunächst auf die gemittelten Leistungen der Leistungsarten 1/3.

Durch Minimierung des GCV-Scores ergibt sich für die Leistungsklassen 1/3 ein Glättungsparameter von $\hat{\lambda} = 10^{-8} \cdot 1.5^{38} \approx 0.05$.

Abbildung 4.44 (links) zeigt das geschätzte PRS Modell für die arithmetischen Mittel der Leistungen der Leistungsarten 1/3. Es lässt sich für die Leistungsklassen 1/3 ein fallender Trend des Modells mit steigenden Altersklassen feststellen. Zum Vergleich werden die Mittelwerte der Tarifklasse 2 abgebildet.

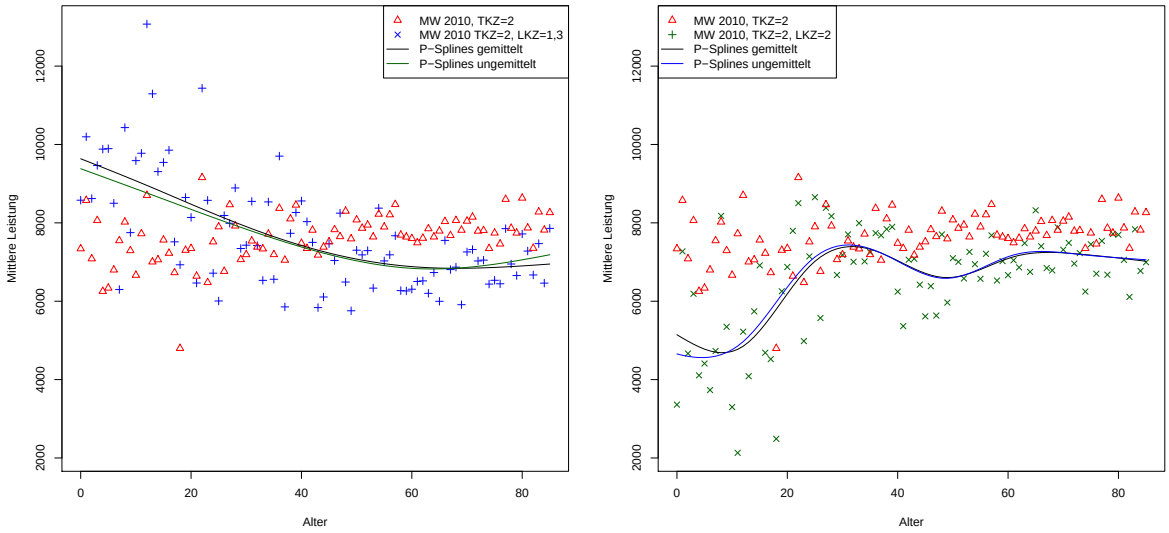


Abbildung 4.45: P-Splines der globalen Daten der Tarifklasse 2: Links das Modell der Leistungsarten 1/3, rechts für Leistungsart 2 bei Anwendung auf die gemittelten und ungemittelten Leistungen.

Für die Leistungsklasse 2 ergibt sich ein Glättungsparameter von $\hat{\lambda} = 10^{-8} \cdot 1.5^{38} \approx 5.7 \cdot 10^{-4}$. Abbildung 4.44 (rechts) zeigt einen steigenden Trend des Modells mit zunehmendem Alter.

Als nächstes werden die *P-Splines* verwendet. Es lassen sich in Abbildung 4.45 leichte Unterschiede bei Nutzung der gemittelten beziehungsweise ungemittelten Daten feststellen für die Leistungsklassen 1/3 feststellen.

Auch für die Leistungsart 2 sind in Abbildung 4.45 nur marginale Unterschiede im unteren Bereich der Altersstufen für die Modelle der gemittelten beziehungsweise ungemittelte Leistungen feststellbar.

Nun werden loglineare *GAMs* auf die Daten angewendet. Als Verteilung wird die Gamma-Verteilung angenommen.

Abbildung 4.46 (rechts) gibt den Glättungsterm des Modells wieder. Sowohl durch Betrachtung der Konfidenzintervalle des Terms, als auch durch die Schätzung des EDF-Wertes lässt sich die Signifikanz des Glättungsterms $\tilde{f}(x)$ folgern.

Abbildung 4.46 (links) zeigt die Schätzung des Modells. Man erkennt einen Abwärtstrend

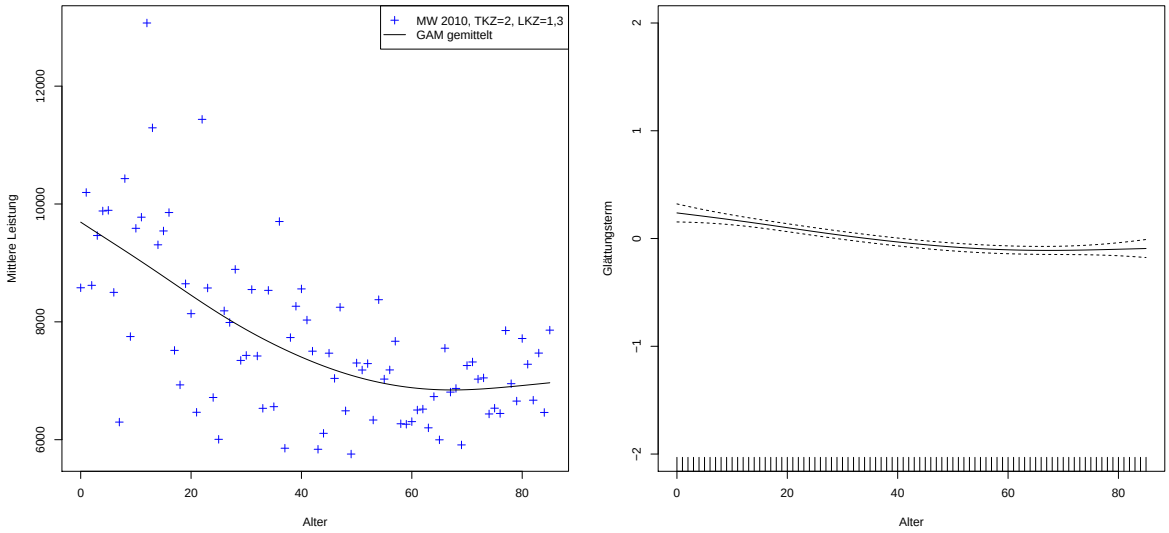


Abbildung 4.46: Links: Schätzung eines loglinearen GAMs der globalen arithmetischen Mittel der Leistungen der Tarifklasse 2 für die Leistungsarten 1/3. Rechts: Jeweiliger Glättungsterm des Modells mit einem EDF von 2.50.

des Modells für steigende Altersstufen und einen guten Fit an die gemittelten Leistungen.

Für das GAM der ungemittelten Daten wird der Glättungsterm in Abbildung 4.47 (links) dargestellt. Im Zusammenhang mit der `summary()`-Ausgabe folgt die Signifikanz des Glättungsterms des Modells.

Abbildung 4.47 (rechts) zeigt den Fit des Modells durch Anwendung eines loglinearen GAMs auf die ungemittelten Daten. Zum Vergleich wird wieder das Modell der ungemittelten Daten geplottet. Es lässt sich kaum ein Unterschied der beiden Modelle feststellen.

Nun wird auch für Leistungsart 2 ein GAM der gemittelten Leistungen betrachtet. Abbildung 4.48 (rechts) zeigt die Darstellung des Glättungsterms. Zusätzlich ist der geschätzte EDF an der Ordinatenachse ablesbar. Die Tatsache, dass die null nicht über den gesamten Bereich des Prädiktors `Alter` von den Konfidenzinervallen überdeckt wird und der EDF-Wert sprechen für die Signifikanz des Glättungsterms.

Abbildung 4.48 (links) zeigt die Schätzung des GAMs für die Leistungart 2 der Tarifklasse 2 sowohl unter Anwendung der Orginaldaten, als auch der arithmetischen Mittel der Leistungen. Es sind leichte Unterschiede der beiden Modelle im Altersbereich von circa 0-25 Jahren erkennbar. Für versicherte Personen höheren Alters werden die

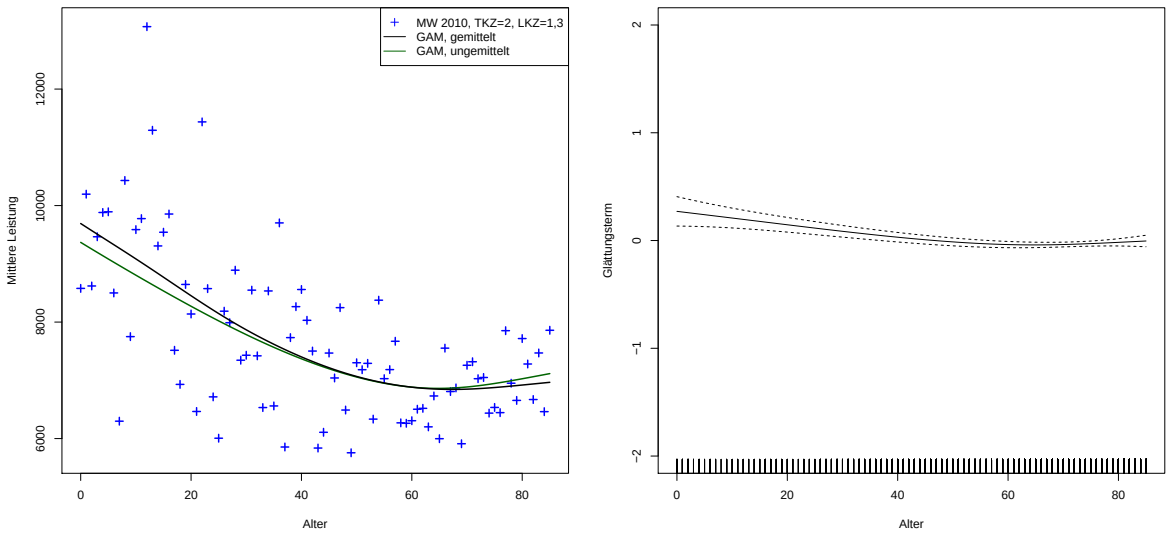


Abbildung 4.47: Links: Schätzung eines GAMs der globalen Originaldaten (schwarz) im Vergleich mit dem Modell der gemittelten Leistungen (grün) der Tarifklasse 2 für die Leistungsarten 1/3 unter Annahme der Quasigammaverteilung. Rechts: Glättungsterm eines GAMs der globalen Originaldaten unter Annahme der Gammaverteilung der Tarifklasse 2 und Leistungsart 1/3. Schätzung des Freiheitsgrades: EDF = 2.40.

beiden Modelle praktisch identisch geschätzt.

Zum Abschluß werden Modelle der *Quantile Regression* auf die betrachteten Leistungsarten der Tarifklasse 2 angewendet. Abbildung 4.49 links illustriert die Anwendung einiger QR-Modelle zu den Niveaus $\tau = \{0.50, 0.75, 0.85, 0.90\}$ für die Leistungsarten 1 und 3.

Erstmals ist bereits das geschätzte 50%-Quantil (gepunktete Linie) über den gesamten Bereich des Prädiktors `Alter` von null verschieden. Man erkennt einen höheren Einfluß von `Alter` für die niedrigeren Altersstufen. Der Verlauf der geschätzten Medianfunktion ist ähnlich zu jenem der jeweiligen Mittelwerte. Allgemein zeigt das QR-Modell für $\tau = 0.50$ einen linearen, fallenden Trend.

Das 75%-Quantil spiegelt ein ähnliches Ergebnis wieder. Für junge Versicherte zeigt der Plot eine sehr hohe Schätzung, die bis zu einem Alter von ungefähr 40 Jahren stark absinkt und dann stagniert, um dann für Versicherte ab einem Alter von ungefähr 70 Jahren wieder leicht anzusteigen.

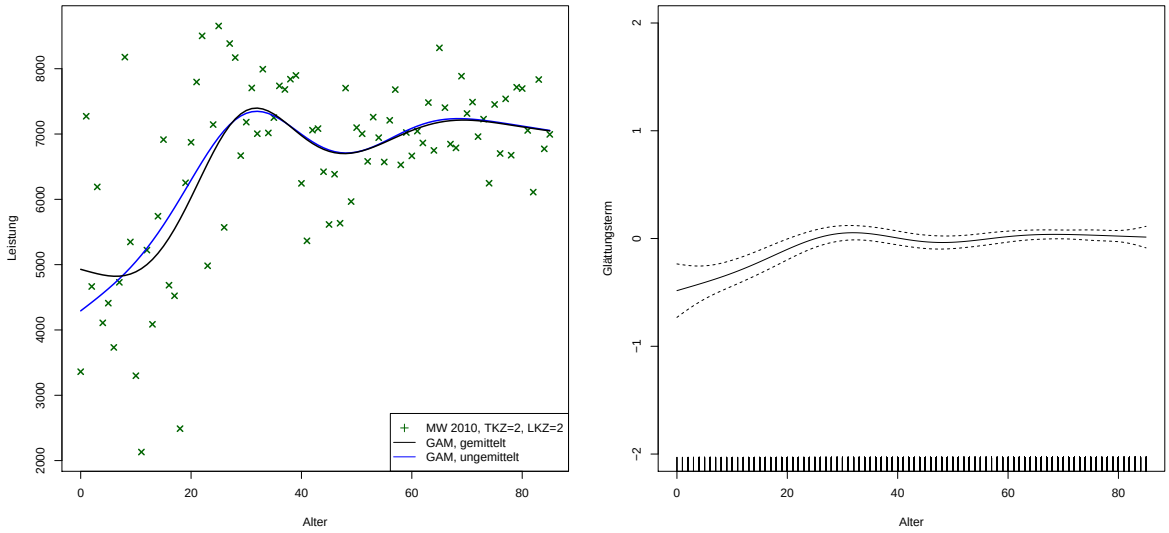


Abbildung 4.48: Links: Schätzung eines GAMs der globalen Originaldaten (schwarz) im Vergleich zum jeweiligen Modell der gemittelten Leistungen (grün). Zusätzlich werden die Mittelwerte der Leistungsart 2 der Tarifklasse 2 des Jahres 2010 dargestellt ($\times$). Rechts: Glättungsterm des jeweiligen loglinearen GAMs der globalen Originaldaten der Tarifklasse 2 und Leistungsart 2 unter Annahme der Quasigammaverteilung mit EDF = 5.77.

Für die 85%- und 95%-Quantile zeigt Abbildung 4.49 (links) ein paralleles Verhalten, wodurch ein gleichbleibender Einfluß des Prädiktors `Alter` für die oberen Leistungssegmente festzustellen ist. Allgemein erkennt man einen leicht fallender Trend der Schätzungen. Erfreulicherweise ist über alle Quantilniveaus $\tau$ kein Kreuzen der Quantilsschätzungen festzustellen.

Für die Leistungsart 2 werden die Schätzungen in Abbildung 4.49 rechts geplottet.

Das geschätzte 50%-Quantil besitzt ein sehr niedriges Niveau für den unteren Bereich des Prädiktors `Alter`. Bis zum Alter von circa 30 Jahren ist ein deutlicher Anstieg der Schätzung erkennbar; für ältere Versicherte beginnt die Quantilsfunktion zu oszillieren. Der Verlauf des Medians ist ähnlich zu dem der jeweiligen Mittelwerte.

Das gefittete 75%-Quantil zeigt einen viel flachere Kurve als der geschätzte Median. Es ist allgemein ein steigender Trend mit einem Maximum ungefähr beim Alter von 70 Jahren.

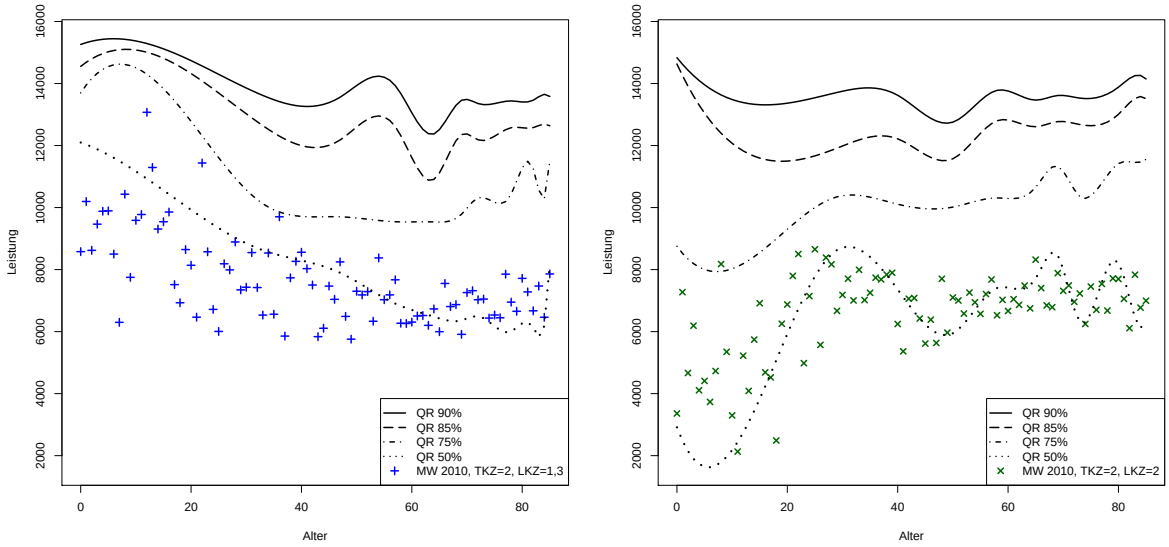


Abbildung 4.49: QR-Modelle der Leistungsarten 1/3 (links) und der Leistungsart 2 (rechts) zu den Niveaus $\tau \in \{0.50, 0.75, 0.85, 0.90\}$ der Tarifklasse 2 für die globalen Daten des Jahres 2010. Zusätzlich werden die globalen Mittelwerte der Leistungsarten 1/3 (+) beziehungsweise der Leistungsart 2 ($\times$) aufgetragen.

Die Quantile der Niveaus 85% und 95% zeigen annähernd eine parallele Schätzung. Es ist also ein gleichbleibender Trend des Einflußes des Prädiktors `Alter` für das obere Leistungssegment fesstellbar. Es werden hier die Leistungsansprüche für sehr junge und sehr alte Versicherte recht hoch geschätzt, dazwischen ist ein leichter Abstieg der Kurve erkennbar.

# 5 Schlussfolgerung

Im Folgenden sollen ein paar positive und negative Aspekte der betrachteten Modellklassen hervorgehoben werden.

Prinzipiell treten vor allem bei der Verwendung Quantiler Regressionsmodelle für die vorhandene Datensituation einige Probleme auf, die berücksichtigt werden müssen.

Werden nichtparametrische QR-Modelle mit Hilfe von B-Splines geschätzt, darf der Freiheitsgrad dieser Splines nicht zu hoch gewählt werden, da die Designmatrix sonst singulär werden kann. Der Grund dafür ist, dass im Datensatz sehr wenige „eindeutige“ Punkte sind, d.h. es gibt sehr viele Beobachtungen $x$-Jähriger, wodurch mehrere übereinstimmende Knotenpunkte auftreten, da die `R`-Funktion `bs()` standardmäßig als Knotenpunkte Quantile des Prädiktors `Alter` wählt.

Weiters wurde auf den Vergleich der Konfidenzintervalle der geschätzten Quantilsfunktionen verzichtet:

Bestimmen von KIV für QR-Modelle

```
1 predict(qr.mod, newdata, interval="confidence", level=0.95)
```

Die beschriebene Methode mittels der Funktion `predict()` funktioniert zwar für die Leistungen der Tarifklasse 1, jedoch lässt sich selbiges für die Tarifklasse 2 aufgrund von Speicherproblemen mit dem Programm `R` nicht berechnen.

Wegen der vielen nullen in den Daten ist aufgrund von numerischen Problemen oft keine Quantilsfunktion für Niveaus $\tau \leq 0.50$ berechenbar, wie dies zum Beispiel für die weiblichen Leistungen der Tarifklasse 2 der Fall ist.

Ansonsten ist die Anwendung von Quantiler Regression sehr intuitiv, wenn auch die Theorie dahinter zunächst gewöhnungsbedürftig erscheint. Die Interpretation der Schätzungen ist im Vergleich zu herkömmlichen Erwartungswertmodellen etwas komplizierter, jedoch erhält man dafür einen ausführlicheren Blick auf die geschätzte Verteilung der Response. Während Erwartungswertmodelle das Zentrum der Verteilung modelliert, ist es bei QR-Modellen möglich, extreme Ereignisse zu schätzen. Zusätzlich muss für die Schätzung von Quantilsfunktionen keinerlei Annahme über die Verteilung der Responsevariablen getroffen werden, während dies bei den Erwartungswertmodellen schon der Fall sein muss. Aber auch Erwartungswertmodelle liefern gute Schätzungen für die vorhandene Datenlage.

Tabelle 5.1 fasst die AIC-Werte der betrachteten Modelle nochmals zusammen.

Für die loglinearen GAMs musste die Quasi-Gammaverteilung angewendet werden, da die Annahme einer Gamma-Verteilung nullen im Datensatz verbietet. Dadurch kann für diese Modelle kein AIC-Wert berechnet werden, da für dessen Berechnung eine Log-Likelihood-Funktion benötigt wird, jedoch nur eine *Quasi*-Log-Likelihood Funktion zur Verfügung steht.

Aus zuvor beschriebenen Gründen macht eine Betrachtung von QR-Modellen nur für die ungemittelten Leistungen Sinn, weshalb auch kein AIC-Wert für die gemittelten Leistungen vorhanden ist. Der Algorithmus zur Berechnung des optimalen Glättungsparameters für PRS-Modelle ist aus numerischen Gründen nicht auf die Originaldaten anwendbar, weshalb auch hier kein AIC-Wert vorliegt.

Sowohl für die geschlechterspezifischen Modelle als auch für die globalen Daten werden die niedrigsten AIC-Werte bei den P-Spline Modellen beobachtet, wobei die Unterschiede vernachlässigbar sind.

| | Männer | | Frauen | |
|---|---|---|---|---|
| AIC-Wert | Gemittelt | Ungemittelt | Gemittelt | Ungemittelt |
| GKS | 648.97 | 74604.88 | 1123.26 | 422774.90 |
| PRS | 744.14 | — | 1223.21 | — |
| P-Splines | 648.51 | 74601.39 | 1122.21 | 422770.10 |
| loglineare GAMs | — | — | — | — |
| QR 90% | — | 80419.64 | — | 443487.80 |

| | Globale Daten, 2010 | | | |
|---|---|---|---|---|
| | TKZ = 1 | | TKZ = 2 | |
| AIC-Wert | Gemittelt | Ungemittelt | Gemittelt | Ungemittelt |
| GKS | 647.57 | 167919.90 | 1078.18 | 749909.50 |
| PRS | 738.19 | — | 1179.83 | — |
| P-Splines | 641.83 | 167914.80 | 1074.21 | 749900.60 |
| loglineare GAMs | — | — | — | — |
| QR 90% | — | 180647.20 | — | 784288.00 |

Tabelle 5.1: Vergleich der AIC-Werte für die unterschiedlichen Modelle.

# Literaturverzeichnis


Aitkin, M., Francis, B., Hinde, J. & Darnell, R. (2009). *Statistical Modelling in R.* Oxford University Press.

Barrodale, I. & Roberts, F. (1973). An improved algorithm for discrete $L_1$ linear approximation. *SIAM Journal of Numerical Analysis*, *10*, 839-848.

Bates, D. M. & Venables, W. N. (2013). *Regression spline functions and classes.* R Documentation. Zugriff auf `http://127.0.0.1:20742/library/splines/html/splines-package.html`

Bosch, R., Ye, Y. & Woodworth, G. (1995). A convergent algorithm for quantile regression with smoothing splines. *Computational Statistics and Data Analysis*, *19*, 613-630.

Box, G. & Cox, D. (1964). An analysis of transformations. *Journal of the Royal Statistical Society Series B*, *26*, 211-252.

Dantzig, G. (1966). *Lineare Programmierung und Erweiterungen.* Springer Verlag.

EU. (2013, September). *Künftig geltende Vorschriften (Solvency II, Omnibus II).* Zugriff auf `http://ec.europa.eu/internal_market/insurance/solvency/future/index_de.htm`

Fahrmeir, L., Kneib, T., Lang, S. & Marx, B. (2013). *Regression.* Springer Verlag.

Fiacco, A. & McCormick, G. (1968). *Nonlinear Programming: Sequential Unconstrained Minimization Techniques.* New York: Wiley.

Genschel, U. & Becker, C. (2005). *Schließende Statistik, Grundlegende Methoden.* Springer.

Gu, C. (2002). *Smoothing Spline ANOVA Models.* Springer.

He, X. & Ng, P. (2011). *COBS: Qualitatively Constrained Smoothing via Linear Programming.* Zugriff auf `http://franke.nau.edu/pin-ng/working/cobs_cs.pdf`

Klenke, A. (2006). *Wahrscheinlichkeitstheorie.* Springer.

Koenker, R. (2005). *Quantile Regression.* Cambridge University Press.

Koenker, R. & Bassett, G. (1978). Regression quantiles. *Econometrica*, *46*, 33-50.

Koenker, R. & D'Orey, V. (1987). Computing regression quantiles. *Journal of the Royal Statistical Society. Series C*, *36*, 383-393.

Koenker, R., Ng, P. & Portnoy, S. (1994). Quantile smoothing splines. *Biometrika*, *81*, 673-680.

Natanson, I. (1974). *Theory of Functions of a Real Variable*. Frederick Ungar Publishing Co.

Peternell, C. (2013). *Einführung in Generalisierte Additive Modelle* (Bericht). Institut für Statistik. Technische Universität Graz.

Schoenberg, I. (1964). Spline functions and the problem of graduation. *Proceedings of the National Academy of Sciences*, *52*, 947-950.

Wood, S. (2001). mgcv: GAMs and Generalized Ridge Regression for R. *R News*, *1*, 20-25.

Wood, S. (2006). *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC.