Michele Soranno, BSc

Gewichtete Regressionsmodelle für heteroskedastische Verschleißraten

MASTERARBEIT

zur Erlangung des akademischen Grades eines Diplomingenieurs

Masterstudium Technische Mathematik: Operations Research und Statistik

eingereicht an der

Technischen Universität Graz

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

Institut für Statistik

Graz, Juli 2018

Eidesstattliche Erklärung

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

Affidavit

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

Datum	Unterschrift

Zusammenfassung

Heteroskedastische Responsevariablen in einem gewöhnlichen linearen Modell wirken sich negativ auf die Eigenschaften des kleinsten Quadrate Schätzers aus. Als größter Problempunkt stellt sich die Verzerrung der Varianz-Kovarianz-Matrix heraus. Dieser Komplikation kann jedoch mit Hilfe robuster Standardfehler gegengesteuert werden. Als Erweiterung zum gewöhnlichen linearen Modell beschäftigt sich diese Masterarbeit mit dem gewichteten linearen Regressionsmodell und seinen Komponenten. Diese Modelle haben den Vorteil, dass im Falle passend gewählter Gewichte, der Heteroskedastizität entgegen gewirkt werden kann. In der Praxis sind Gewichte meistens nicht vorab bekannt. Daher werden zwei Techniken präsentiert, wie diese aus den vorhandenen Daten geschätzt werden können. Um das Problem der Heteroskedastizität zusätzlich aus einem alternativen Blickwinkel zu betrachten, werden generalisierte lineare Modelle (GLMs) vorgestellt. Der Effekt heteroskedastischer Responsevariablen kann hierfür durch so genannte prior-weights ausgeglichen werden. Es wird außerdem auf einen Spezialfall der GLMs, nämlich die Gammaregression, genauer eingegangen. Ausgangspunkt für den praktischen Abschnitt der Masterarbeit stellt eine Kooperation der Technischen Universität Graz mit dem Unternehmen AC^2T *research GmbH* dar. Im Zuge dieser Zusammenarbeit wurde ein Einzylinder-Forschungsmotor untersucht, wobei das Hauptaugenmerk in der Analyse der Reibung und des Verschleißes der Kolbengruppe lag. Dabei wurde die Verschleißhöhe des Kolbenringes für unterschiedliche Motoreinstellungen gemessen. Ausgehend von diesen Messdaten versucht man mittels gewichteter Regressionsmodelle die Verschleißrate durch die Motoreinstellungen zu erklären. Unter all den betrachteten Modellen überzeugte ein gewichtetes Gammamodel mit log-Link.

Abstract

Heteroscedastic response variables in a linear regression model have a negative impact on the properties of the least squares estimator. The biggest problem is that the traditional formula of the variance-covariance matrix does not hold, which can lead to incorrect inference. However, this complication can be circumvented by using robust standard errors. As an extension to the linear regression model, this thesis presents the weighted linear regression model and its components. If the weights of this model are chosen in a suitable way, one can counteract the negative effects of heteroscedastic responses. In practice, weights are usually not known in advance. Therefore two techniques are introduced, on how one can estimate weights from the existing data. Moreover, the problem of heteroscedasticity is handled from a different point of view, namely by generalized linear models (GLMs). In this case the effect of heteroscedastic response variables can be countered by prior-weights. More than that, a special case of GLMs, the Gammaregression, is discussed in detail. The starting point for the practical part of this thesis is a cooperation between the Graz University of Technology and the company *AC^{2}T research GmbH*. As part of this collaboration a single cylinder research engine was examined. The primary focus was the analysis of wear and friction of the piston group. The wear of the piston ring was measured for different engine settings. Finally, the previously elaborated theory is applied to the measured data. Amongst all the considered models a weighted gamma model with log-link is the most convincing one.

Danksagung

Ich möchte mich bei meinem Betreuer Herrn Prof. Friedl für das interessante Masterarbeitsthema und seine fachliche Unterstützung recht herzlich bedanken. Durch seine Hilfsbereitschaft konnten alle Fragen und Probleme gelöst werden.

Des Weiteren bedanke ich mich bei Herrn Dipl. Ing. Walch für die zur Verfügung gestellten Messdaten, sowie für die Besichtigung des Motorenprüfstandes der Technischen Universität Graz.

Ganz besonders möchte mich bei meiner Familie, insbesondere meinen Eltern bedanken, die mir mein Studium ermöglicht und mich stets unterstützt haben. Weiters danke ich meiner Freundin für ihre motivierenden Worte und das Korrekturlesen meiner Arbeit und auch meinen Freunden und Studienkollegen, die mir eine unvergessliche Studienzeit ermöglicht haben.

INHALTSVERZEICHNIS

KAPITEL 1

EINLEITUNG

Eine zentrale Komponente von vielen Kraftfahrzeugen ist der Verbrennungsmotor. Dieses Antriebsystem wird seit über 100 Jahren kontinuierlich verbessert und weiterentwickelt. Auch die Technische Universität Graz forscht an Verbrennungsmotoren und besitzt am Institut für Maschinenelemente und Entwicklungsmethodik einen Motorenprüfstand, an dem Motoren untersucht werden können. Im Zuge einer Kooperation der Technischen Universität Graz mit dem Unternehmen AC^2T *research GmbH* wird ein Verbrennungsmotor untersucht, bei dem besonders die Analyse der Reibung und des Verschleißes von Interesse ist.

Diese Arbeit gliedert sich in einen theoretischen und einen praktischen Teil, wobei der theoretische Abschnitt in zwei Kapitel aufgeteilt ist. Zu Beginn des Kapitels 2 werden das einfachste Regressionsmodell, das so genannte Gewöhnliche Lineare Modell, und seine Komponenten vorgestellt. Dieses Modell kann durch seine Einfachheit punkten, jedoch bergen seine restriktiven Annahmen einige Nachteile. In der Praxis verletzt der nicht beobachtbare statistische Fehler manchmal die Annahme der konstanten Varianz, auch als Homoskedastiziät bekannt. Folglich beschäftigt sich Abschnitt 2.1.2 mit den Auswirkungen, die durch die Missachtung der Modellannahme entstehen. Wird die Modellannahme der Homoskedastizität relaxiert, entsteht aus dem gewöhnlichen linearen Modell das gewichtete lineare Modell. Der Abschluss dieses Kapitels behandelt das neue Regressionsmodell sowie seine Komponenten.

Kapitel 3 beschäftigt sich mit den Generalisierten Linearen Modellen. Diese stellen eine Verallgemeinerung der gewöhnlichen linearen Regressionsmodelle aus Kapitel 2 dar. Am Beginn werden Parameterschätzer hergeleitet und in den nachfolgen-

den Abschnitten Maße für die Güte der Modellanpassung ermittelt. Der Abschluss des theoretischen Abschnitt behandelt einen Spezialfall der Generalisierten Linearen Modelle, nämlich die Gammaregression. Die Gammaverteilungsannahme wird speziell für den folgenden Praxisteil benötigt.

Der praktische Abschnitt befasst sich mit den Verschleißdaten eines Kolbenringes, die am Motorenprüfstand für diverse Motoreinstellungen ermittelt wurden. Da der Verschleiß einen Lebensdauer mindernden Effekt auf diverse Bauteile des Motors hat, ist es wünschenswert, diesen besonders klein zu halten. Zu Beginn werden so genannte Verschleißraten ermittelt und im Anschluss wird die Theorie der Kapitel 2 und 3 auf die neu bestimmten Koeffizienten angewandt. Mit Hilfe der zuvor diskutierten Modelle kann der Zusammenhang von Verschleißrate und Motoreinstellung geschätzt werden. Somit kann aufgezeigt werden, inwiefern sich die Motoreinstellung auf die Verschleißrate auswirkt.

KAPITEL 2

LINEARE REGRESSION

2.1 Gewöhnliche Lineare Regression

Die Gewöhnliche Lineare Regression ist ein statistisches Verfahren, das den Zusammenhang zwischen einer abhängigen Variable y und einer oder mehreren erklärenden Variablen $x_1, \ldots, x_{p-1}$ modelliert. Gibt es nur eine erklärende Variable, so spricht man von einfacher linearer Regression, andernfalls von multipler linearer Regression. Die abhängige Variable y wird auch als *Response* und die erklärenden Variablen x_j werden als *Prädiktoren* bezeichnet.

Das Ziel der linearen Regression ist es, eine lineare Funktion $f(x_1, \ldots, x_{p-1})$ der Prädiktoren $x_1, \ldots, x_{p-1}$ zu finden, die den Erwartungswert der Response y am besten beschreibt. Liegen die Daten als p-dimensionales Tupel $(y_i, x_{i1}, \ldots, x_{ip-1})$ für $i = 1, \ldots, n$ vor, so lässt sich das Modell als

$$y_i = f(x_1, \ldots, x_{p-1}) + \epsilon_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{ip-1} + \epsilon_i = \boldsymbol{x}_i' \boldsymbol{\beta} + \epsilon_i$$

darstellen, wobei y_i die i-te Responsevariable, $\boldsymbol{x}_i' = (1, x_{i1}, \ldots, x_{ip-1})'$ die bekannten Prädiktoren, $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{p-1})'$ die unbekannten Parameter und ϵ_i der nicht beobachtbare statistische Fehler mit den Eigenschaften $\mathbb{E}(\epsilon_i) = 0$, $\mathrm{Var}(\epsilon_i) = \sigma^2$ und $\mathrm{Cov}(\epsilon_i, \epsilon_j) = 0$ für $i \neq j$ sind. Das multiple lineare Regressionsmodell (MLR) kann auch kompakter in Matrixschreibweise dargestellt werden. Es lautet

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{2.1}$$

mit dem Responsevektor $\boldsymbol{y} = (y_1, \ldots, y_n)'$, dem unbekannten Prädiktorvektor $\boldsymbol{\beta}$, dem nicht beobachtbaren statistischen Fehlervektor $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)'$ und der $n \times p$

Designmatrix

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p-1} \\ 1 & x_{21} & \cdots & x_{2p-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np-1} \end{pmatrix}.$$

Das MLR (2.1) besitzt folgende Modellannahmen:

1. Die Designmatrix $\boldsymbol{X}$ hat vollen Spaltenrang, d.h. $\text{rank}(\boldsymbol{X}) = p$.
2. $\mathbb{E}(\boldsymbol{\epsilon}) = \boldsymbol{0}$.
3. $\text{Var}(\boldsymbol{\epsilon}) = \mathbb{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}') = \sigma^2 \boldsymbol{I}_n$.

Man spricht von gewöhnlicher Normalregression, falls

4. $\boldsymbol{\epsilon} \sim Normal(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$ erfüllt ist.

Durch die Annahme der Normalverteilung für den nicht beobachtbaren statistischen Fehler $\boldsymbol{\epsilon}$ folgt $\boldsymbol{y} \sim Normal(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n)$.

Nach dem Aufstellen des Modells (2.1) ist man an der Schätzung der unbekannten Parameter $\boldsymbol{\beta}$ und σ^2 interessiert. Mit Hilfe des geschätzten Parametervektors $\hat{\boldsymbol{\beta}}$ lassen sich die gefitteten Werte $\widehat{\mathbb{E}(y_i)} = \hat{\mu}(x_i) = \hat{\mu}_i = \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}$ ermitteln. Die Abweichung des gefitteten Wert $\hat{\mu}_i$ von der Beobachtung y_i wird als rohes Residuum $r_i = y_i - \hat{\mu}_i$ bezeichnet. Der Parametervektor $\boldsymbol{\beta}$ kann mit Hilfe der Methode der Kleinsten Quadrate geschätzt werden. Das Kleinste Quadrate Kriterium fordert jene Kurve $\hat{\boldsymbol{\mu}}$ zu wählen, die die Summe der quadrierten Abweichungen zwischen den beobachteten Werten und den gefitteten Werten möglichst klein hält. Das Ziel lautet somit, die Summe der quadrierten rohen Residuen

$$\text{SSE}(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2 = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$$

zu minimieren. Der Kleinste Quadrate Schätzer $\hat{\boldsymbol{\beta}}$, kurz KQ-Schätzer, minimiert $\text{SSE}(\boldsymbol{\beta})$ und lautet

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}. \tag{2.2}$$

Der zweite unbekannte Parameter σ^2 kann mittels Maxiumum-Likelihood-Methode geschätzt werden. Da die Response $\boldsymbol{y}$ normalverteilt ist, folgt für die Log-Likelihood-Funktion

$$\log f(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\text{SSE}(\boldsymbol{\beta}). \tag{2.3}$$

Anschließend werden die partiellen Ableitungen der Log-Likelihood-Funktion nach $\boldsymbol{\beta}$ und σ^2 gebildet und gleichzeitig Null gesetzt, um die bekannten Normalengleichungen zu erhalten (Fahrmeir, Kneib und Lang, 2009, S. 92). Durch Umformen dieser Gleichungen erhält man die Maximum-Likelihood-Schätzer

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \qquad \text{und} \qquad \hat{\sigma}^2 = \frac{1}{n}\text{SSE}(\hat{\boldsymbol{\beta}}).$$

Der MLE $\hat{\boldsymbol{\beta}}$ ist ein erwartungstreuer Schätzer für $\boldsymbol{\beta}$ und stimmt mit dem KQ-Schätzer (2.2) überein. Der MLE $\hat{\sigma}^2$ hingegen ist ein verzerrter Schätzer für die Varianz. Ein unverzerrter Varianzschätzer ist durch

$$S^2 = \frac{1}{n-p}\text{SSE}(\hat{\boldsymbol{\beta}})$$

gegeben (Fahrmeir et al., 2009, S. 94), wobei p dem Rang der Designmatrix $\text{rank}(\boldsymbol{X}) = p$ entspricht.

Des Weiteren ist der KQ-Schätzer $\hat{\boldsymbol{\beta}}$, laut dem Satz von Gauß-Markow (siehe Anhang Satz A.10), der beste lineare unverzerrte Schätzer (engl. BLUE für Best Linear Unbiased Estimator). Bester deshalb, weil er in der Klasse der linearen erwartungstreuen Schätzer die kleinste Varianz-Kovarianz-Matrix besitzt. Die Varianz-Kovarianz-Matrix des KQ-Schätzers lautet

$$\text{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}.$$

2.1.1 Feststellen von Heteroskedastizität

Die Annahme konstanter Varianz (*Homoskedastizität*) des statistischen Fehlers $\text{Var}(\boldsymbol{\epsilon}) = \sigma^2\boldsymbol{I}_n$ ist in der Praxis sehr oft nicht gegeben. Häufig ist das Gegenteil der Fall. Die Varianz des statistischen Fehlers lautet nun

$$\text{Var}(\boldsymbol{\epsilon}) = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & 0 & 0 \\ \vdots & 0 & \ddots & 0 \\ 0 & \cdots & 0 & \sigma_n^2 \end{pmatrix} \tag{2.4}$$

und man spricht von sogenannter *Heteroskedastizität*.

Es ist hilfreich, sich über eventuelle Heteroskedastizität des Modells im Klaren zu sein, um diesem Problem mit geeigneten Maßnahmen entgegenzusteuern.

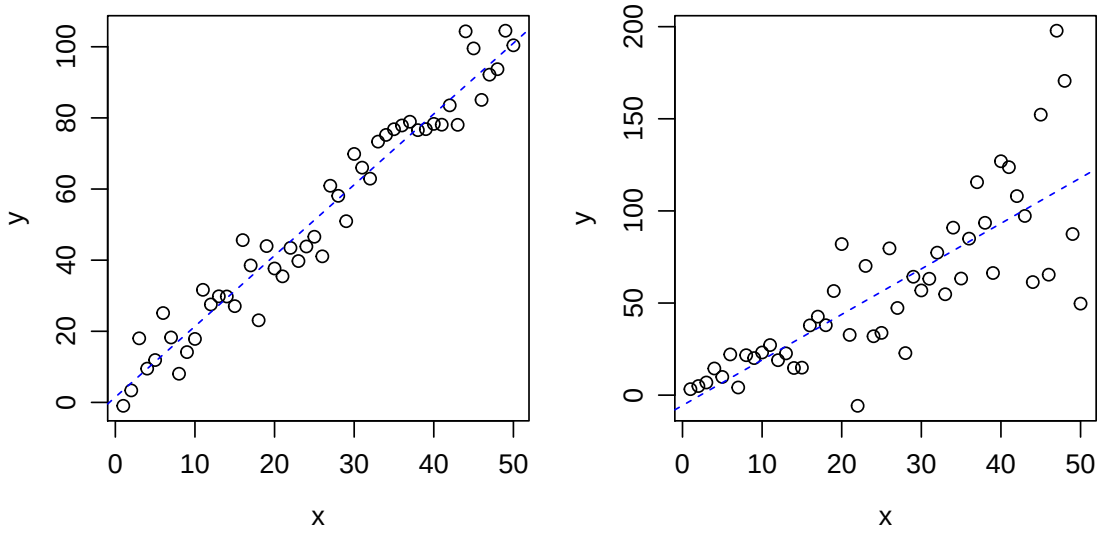

Abbildung 2.1: Regressionen mit unterschiedlicher Varianz von ϵ. Homoskedatizität (links) und Heteroskedastizität (rechts).

Betrachte zunächst die Regressionsgerade $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ für $i = 1, \ldots, n$. Links in der Abbildung 2.1 ist die Streuung der Punkte um die Regressionsgerade in vertikaler Richtung konstant und es ist kein Muster erkennbar. Daher kann man Homoskedastizität annehmen. Im rechten Scatterplot hingegen, steigt mit zunehmenden x die Streuung von y. Somit liegt eindeutig Heteroskedastizität vor. Oftmals ist das Feststellen von Heteroskedastizität nicht so eindeutig wie im rechten Plot der Abbildung 2.1. Als nützliches Hilfsmittel erweisen sich hierbei eine Analyse der Residuen sowie einige Hypothesentests.

Residuen

Die rohen Residuen sind ein hilfreiches Werkzeug, um festzustellen, ob die Modellannahmen in einem MLR (2.1) korrekt erfüllt sind. Üblicherweise werden die rohen Residuen in diversen Diagnoseplots untersucht. Diese sind definiert als Differenz zwischen beobachteten und gefitteten Werten und sind durch

$$\begin{aligned} \boldsymbol{r} &= \boldsymbol{y} - \hat{\boldsymbol{\mu}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} \\ &\overset{(2.2)}{=} \boldsymbol{y} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}, \end{aligned} \tag{2.5}$$

gegeben, wobei $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ die so genannte *Hat-Matrix* ist. Die rohen Residuen haben jedoch den Nachteil, dass ihre Varianzen $\mathrm{Var}(r_i) = \sigma^2(1 - h_{ii})$ nicht konstant und somit heteroskedastisch sind.

Alternativ werden in den Diagnoseplots die studentisierten Residuen betrachtet. Diese Residuen genügen einer Student-t-Verteilung mit $n - p - 1$ Freiheitsgraden, haben Erwartung Null und konstante Varianz. Im homoskedastischen Fall sollten diese Residuen also konstant um die Null streuen. Die studentisierten Residuen sind aufgrund ihrer homoskedastischen Eigenschaft besser geeignet, etwaige Heteroskedastizität der Response aufzuzeigen.

Zur Aufdeckung heteroskedastischer Fehler können daher die rohen oder studentisierten Residuen in einem Scatterplot gegen die gefitteten Werte $\hat{\mu}_i$ bzw. die Prädiktorvariablen x_{ij} geplottet werden.

Hypothesentests

Im Laufe der Jahre wurde eine Vielzahl von Hypothesentests entwickelt, die ein gegebenes Regressionsmodell auf Heteroskedastizität überprüfen. Viele dieser Tests waren zwar in der Lage Heteroskedastizität aufzuspüren, testeten jedoch in keinster Weise die Annahme, dass die Varianz des statistischen Fehlers ϵ_i nicht von den Prädiktoren x_i abhing. Infolgedessen empfiehlt Wooldridge (2005, Kap. 8) den Breusch-Pagan und den White-Test auf Heteroskedastizität.

Breusch-Pagan-Test auf Heteroskedastizität

Ist man sich im Klaren, von welchen Prädiktoren die Varianz des nicht beobachtbaren statistischen Fehlers ϵ_i abhängt, sollte die Heteroskedastizität mittels Breusch-Pagan-Test überprüft werden. Ausgehend von einem gewöhnlichen MLR

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{ip-1} + \epsilon_i, \tag{2.6}$$

für $i = 1, \ldots, n$, wird die Varianz

$$\mathrm{Var}(\epsilon_i) = \sigma_i^2 = h\left(\alpha_0 + \sum_{j=1}^{q} \alpha_j z_{ij}\right)$$

überprüft, wobei z_{ij} die Variablen sind, die die Varianz beeinflussen. Diese können natürlich auch mit den Prädiktoren des Modells (2.6) übereinstimmen. Die Hypothesen lauten

$$H_0: \quad \alpha_1 = \alpha_2 = \cdots = \alpha_q = 0 \qquad \text{vs.} \qquad H_1: \quad \alpha_t \neq 0 \text{ mit } t \in \{1, \ldots, q\}.$$

Unter der Nullhypothese geht man davon aus, dass $\sigma_i^2 = h(\alpha_0)$ für alle i übereinstimmt und somit die Annahme der Homoskedastizität erfüllt ist. Falls H_0 verworfen wird, so entspricht die Varianz einer linearen Funktion der $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{iq})'$. Die Vorgehensweise dieses Tests ist wie folgt:

1. Im ersten Schritt wird eine gewöhnliche lineare Regression für das Modell (2.6) gerechnet. Dadurch erhält man die rohen Residuen r_i des Modells.

2. Mit den zuvor ermittelten Residuen wird eine Hilfsregression zwischen der abhängigen Variable r_i^2 und den erklärenden Variablen $\boldsymbol{z}_i$ kalkuliert, d.h.

$$r_i^2 = \alpha_0 + \alpha_1 z_{i1} + \cdots + \alpha_q z_{iq} + \tilde{\epsilon}_i.$$

   Anschließend wird das Bestimmtheitsmaß $R^2_{r^2}$ der Hilfsregression ermittelt.

3. Die Teststatistik des Breusch-Pagan-Hypothesentests lautet

$$\text{LM} = nR^2_{r^2}$$

   und ist unter H_0 asymptotisch χ^2_q verteilt. Die Nullhypothese der Homoskedastizität wird also verworfen, falls die Teststatistik LM im kritischen Bereich der χ^2-Verteilung mit q Freiheitsgraden realisiert.

Die hier vorgestellte Version des Breusch-Pagan-Tests stammt von Koenker (1981) und ist eine robuste Modifikation des ursprünglichen Tests von Breusch und Pagan (1979). Dieser Test lässt sich in `R` mit Hilfe des Kommandos `bptest` aus dem Packet `lmtest` ausführen (Zeileis und Hothorn, 2002).

White-Test auf Heteroskedastizität

Dieser Hypothesentest basiert auf White (1980) und sollte verwendet werden, falls kaum Wissen über die Struktur der Heteroskedastizität vorhanden ist. Die Hypothesen dieses Tests lauten

$$H_0: \quad \sigma_i^2 = \sigma^2 \ \forall i \qquad \text{vs.} \qquad H_1: \quad \exists i \in \{1, \ldots, n\} : \sigma_i^2 \neq \sigma^2.$$

Die Vorgehensweise des White-Tests auf Heteroskedatizität ist wie folgt:

1. Im ersten Schritt wird wie im Breusch-Pagan-Hypothesentest eine gewöhnliche lineare Regression des Modells

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{ip-1} + \epsilon_i \tag{2.7}$$

   gerechnet, um deren rohe Residuen r_i zu erhalten.

2. Anschließend wird eine Hilfsregression kalkuliert, wobei zusätzlich die quadratischen Terme und Interaktionsterme aller erklärenden Variablen berücksichtigt werden. Wenn z.B. im Schritt Eins eine Gleichung mit zwei Parametern geschätzt wird

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i,$$

so ist die Hilfsregression durch

$$r_i^2 = \delta_0 + \delta_1 x_{i1} + \delta_2 x_{i2} + \delta_3 x_{i1}^2 + \delta_4 x_{i2}^2 + \delta_5 x_{i1} x_{i2} + \tilde{\epsilon}_i \tag{2.8}$$

gegeben. Anschließend wird das Bestimmtheitsmaß $R_{r^2}^2$ der Hilfsregression ermittelt.

3. Die Teststatistik des White-Hypothesentes lautet

$$\mathrm{LM} = nR_{r^2}^2$$

und es lässt sich zeigen, dass LM unter H_0 asymptotisch χ_q^2 verteilt ist, wobei q die Anzahl der Parameter der Hilfsregression (2.8) ist.

Auch dieser Test lässt sich mit dem `R`-Befehl `bptest` ausführen, wobei zusätzlich dem Funktionsparameter `varformula=` das Regressionsmodell aus Schritt 2 übergeben werden muss.

2.1.2 Auswirkungen der Heteroskedastizität

Die Verletzung der Varianzeigenschaft des MLRs (2.1) hat spürbare Folgen für den KQ-Schätzer. Bei Verletzung der Homoskedastizitätseigenschaft sind nicht mehr automatisch alle Voraussetzungen des Gauß-Markow-Theorems erfüllt. Der KQ-Schätzer ist daraufhin nicht mehr der beste lineare Schätzer. Laut den Sätzen A.8 und A.9 verliert er die Eigenschaften der Erwartungstreue und Konsistenz jedoch nicht.

Die Folgen für die Varianz-Kovarianz-Matrix des KQ-Schätzers sind weitaus schwerwiegender. Betrachte zuvor die Differenz

$$\begin{aligned}
\hat{\boldsymbol{\beta}} - \mathbb{E}(\hat{\boldsymbol{\beta}}) &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} - \boldsymbol{\beta} \\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) - \boldsymbol{\beta} \\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\epsilon}.
\end{aligned} \tag{2.9}$$

Mittels (2.9) folgt für die Varianz-Kovarianz-Matrix

$$\begin{aligned}
\mathrm{Var}(\hat{\boldsymbol{\beta}}) &= \mathbb{E}((\hat{\boldsymbol{\beta}} - \mathbb{E}(\hat{\boldsymbol{\beta}}))(\hat{\boldsymbol{\beta}} - \mathbb{E}(\hat{\boldsymbol{\beta}}))') \\
&\overset{(2.9)}{=} \mathbb{E}((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\epsilon}\boldsymbol{\epsilon}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}) \\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathbb{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}')\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\,\mathrm{Var}(\boldsymbol{\epsilon})\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}.
\end{aligned} \tag{2.10}$$

Nachdem die Varianz nicht konstant ist, gilt $\mathrm{Var}(\boldsymbol{\epsilon}) \neq \sigma^2 \boldsymbol{I}_n$ und somit kann (2.10) nicht weiter vereinfacht werden. Daher ist der KQ-Schätzer nicht mehr asymptotisch

effizient und die Varianz-Kovarianz-Matrix, $\text{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$, liefert verzerrte Schätzungen für die Standardfehler der Koeffizienten. Somit sind alle darauf beruhenden Hypothesentests ungültig. Jedoch bewies White (1980), dass die Matrix

$$n^{-1}\sum_{i=1}^{n}\mathbb{E}(\epsilon_i^2)\boldsymbol{x}_i\boldsymbol{x}_i'$$

konsistent durch

$$n^{-1}\sum_{i=1}^{n}r_i^2\boldsymbol{x}_i\boldsymbol{x}_i'$$

geschätzt werden kann, wobei r_i dem i-ten Residuum entspricht. Dadurch ergibt sich die heteroskedastisch konsistente Varianz-Kovarianz-Matrix

$$\widehat{\text{Var}}_{HC}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\text{diag}[r_i^2]\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}. \tag{2.11}$$

Die Standardfehler, welche auf (2.11) beruhen, werden in der Fachliteratur auch als *robuste Standardfehler* oder *Eicker–Huber–White Standardfehler* bezeichnet. Diese können in die Teststatistik des F-Tests zum Test allgemeiner linearer Hypothesen eingesetzt werden. Dadurch erhält man asymptotisch korrekte Tests sowie asymptotisch korrekte Konfidenzintervalle. Robuste Standardfehler haben jedoch den Nachteil, dass sie in der Regel größer sind als Standardfehler der gewöhnlichen linearen Regression. Es wird also asymptotische Gültigkeit durch erhöhte Varianzen der Schätzer erkauft.

Weitere robuste Standardfehler

Der zuvor ermittelte robuste Standardfehler (2.11) besitzt nur asymptotische Eigenschaften und ist somit für kleine Stichprobenumfänge nicht geeignet. Aus diesem Grund ermittelten MacKinnon und White (1985) alternative Schätzer, die asymptotisch äquivalent zu (2.11) sind und dieser Matrix in kleinen Stichprobenumfängen überlegen sind. Die heteroskedastisch konsistente Varianz-Kovarianz-Matrix (2.11) wird fortan mit

$$\text{HC0} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\text{diag}[r_i^2]\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \tag{2.12}$$

bezeichnet. Eine bias korrigierte Modifikation von (2.12) ist durch

$$\text{HC1} = \frac{n}{n-p}\text{HC0} \tag{2.13}$$

gegeben, wobei p die Anzahl der Parameter des linearen Modells ist. Weitere robuste Varianz-Kovarianz-Matrizen lauten

$$\text{HC2} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\text{diag}\left[\frac{r_i^2}{1-h_{ii}}\right]\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \tag{2.14}$$

oder

$$\text{HC3} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\text{diag}\left[\frac{r_i^2}{(1-h_{ii})^2}\right]\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}. \tag{2.15}$$

Die Einträge der Diagnoalmatrix sind Brüche, wobei im Zähler r_i das i-te Residuum und im Nenner h_{ii} der i-te Eintrag der Hat-Matrix ist.

Im Falle von Heteroskedastizität sind Hypothesentests und Konfidenzintervalle, die auf den Standardfehlern des gewöhnlichen KQ-Schätzer basieren, ungültig und können zu falschen Schlussfolgerungen führen. Die Varianz-Kovarianz-Matrix (2.10) kann nun durch einen der Schätzer HC0, HC1, HC2 oder HC3 konsistent geschätzt werden.

Long und Ervin (2000) haben eine Monte Carlo Simulationsstudie durchgeführt, um das Verhalten der zuvor genannten Matrizen für kleine Stichprobenumfänge zu überprüfen. Im Zuge dieser Studie wurden Daten simuliert, Regressionsgleichungen geschätzt und Hypothesentests berechnet. Des Weiteren wurde gezählt, wie oft die korrekte Nullhypothese der Hypothesentests verworfen wurde (Type 1 Error). Das beste Ergebnis lieferte die Matrix HC3, da sie sogar in sehr kleinen Stichprobenumfängen ($n = 25$) überzeugen konnte. Die robusten Matrizen HC0 bis HC3 können in `R` aus dem Package `sandwich` mit dem Kommando `vcovHC` ermittelt werden (Zeileis, 2006). Die Default-Einstellung ist HC3. Möchte man eine andere robuste Matrix berechnen, so kann man dies mit Hilfe des Parameters `type=` machen.

2.2 Gewichtete Lineare Regression

Die gewichtete lineare Regression ist eine weitere Möglichkeit, wie man dem Problem der Heteroskedastizität entgegen wirken kann. Im multiplen linearen Regressionsmodell (2.1) geht man davon aus, dass der statistische Fehler unkorreliert und homoskedastisch ist, d.h.

$$\text{Var}(\epsilon_i) = \sigma^2.$$

Im Falle von Heteroskedastizität benötigt man jedoch eine allgemeinere Annahme. Die Varianz sei nun durch

$$\text{Var}(\epsilon_i) \overset{(2.4)}{=} \sigma_i^2 = \frac{\sigma^2}{w_i} \tag{2.16}$$

gegeben, wobei w_i bekannte positive Konstanten sind. Die Varianz ist dennoch durch eine unbekannte Zahl σ^2 charakterisiert. In der Matrixschreibweise ist dafür die

Varianz-Kovarianz-Matrix der Response und des statistischen Fehlers durch

$$\mathrm{Var}(\boldsymbol{y}) = \mathrm{Var}(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{W}^{-1}$$

gegeben, wobei $\boldsymbol{W}$ eine $n \times n$ positiv definite Diagonalmatrix und $\boldsymbol{W}^{-1}$ deren Inverse ist, mit

$$\boldsymbol{W} = \begin{pmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & 0 & 0 \\ \vdots & 0 & \ddots & 0 \\ 0 & \cdots & 0 & w_n \end{pmatrix}, \quad \boldsymbol{W}^{-1} = \begin{pmatrix} \frac{1}{w_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{w_2} & 0 & 0 \\ \vdots & 0 & \ddots & 0 \\ 0 & \cdots & 0 & \frac{1}{w_n} \end{pmatrix}.$$

Das gewichtete multiple lineare Regressionsmodell lautet

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{2.17}$$

mit den Modellannahmen:

1. Die Designmatrix $\boldsymbol{X}$ hat vollen Spaltenrang, d.h. $\mathrm{rank}(\boldsymbol{X}) = p$.
2. $\mathbb{E}(\boldsymbol{\epsilon}) = \boldsymbol{0}$.
3. $\mathrm{Var}(\boldsymbol{\epsilon}) = \mathbb{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}') = \sigma^2 \boldsymbol{W}^{-1}$.

Man spricht von allgemeiner Normalregression, falls

4. $\boldsymbol{\epsilon} \sim Normal(\boldsymbol{0}, \sigma^2 \boldsymbol{W}^{-1})$ erfüllt ist.

Durch die Annahme der Normalverteilung für den nicht beobachtbaren statistischen Fehler $\boldsymbol{\epsilon}$ folgt laut Fahrmeir et al. (2009, S. 125) das $\boldsymbol{y} \sim Normal(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{W}^{-1})$.

2.2.1 Transformiertes Modell

Ausgehend vom gewichteten linearen Regressionsmodell (2.17) versucht man mittels einer Datentransformation die Eigenschaften des nicht beobachtbaren statistischen Fehlers $\boldsymbol{\epsilon}$ so zu ändern, dass eine konsistente Schätzung der Varianz-Kovarianz-Matrix möglich ist (siehe Folgerung aus (2.10)). Dabei orientieren wir uns an Fahrmeir et al. (2009, S. 125 ff.).

Nachdem $\boldsymbol{W}$ eine positiv definite Matrix ist, existiert eine nicht singuläre Matrix $\boldsymbol{W}^{1/2}$, so dass $\boldsymbol{W} = \boldsymbol{W}^{1/2}(\boldsymbol{W}^{1/2})'$ ist. Da $\boldsymbol{W}$ eine Diagonalmatrix ist, ist natürlich $\boldsymbol{W}^{1/2}$ auch eine $n \times n$ Diagonalmatrix, wobei $\sqrt{w_i}$ dem i-ten Diagonalelement entspricht. Es wird die Varianz-Kovarianz-Matrix des Zufallsvektors $\boldsymbol{W}^{1/2}\boldsymbol{\epsilon}$ untersucht,

wofür gilt

$$\begin{aligned}\mathrm{Var}(\boldsymbol{W}^{1/2}\boldsymbol{\epsilon}) &= \boldsymbol{W}^{1/2}\,\mathrm{Var}(\boldsymbol{\epsilon})\boldsymbol{W}^{1/2} \\ &= \boldsymbol{W}^{1/2}\sigma^2\boldsymbol{W}^{-1}\boldsymbol{W}^{1/2} \\ &= \sigma^2\boldsymbol{W}^{1/2}\boldsymbol{W}^{-1/2}\boldsymbol{W}^{-1/2}\boldsymbol{W}^{1/2} \\ &= \sigma^2\boldsymbol{I}. \end{aligned} \tag{2.18}$$

Die Varianz des Zufallsvektors $\boldsymbol{W}^{1/2}\boldsymbol{\epsilon}$ ist für jeden Eintrag identisch. Somit konnte durch die Multiplikation der Matrix $\boldsymbol{W}^{1/2}$ die Heteroskedastizität von $\boldsymbol{\epsilon}$ behoben werden. Aufgrund dessen wird die Regressionsgleichung (2.17) mit der Gewichtsmatrix $\boldsymbol{W}^{1/2}$ von links multipliziert. Dadurch resultiert folgende Gleichung

$$\underbrace{\boldsymbol{W}^{1/2}\boldsymbol{y}}_{\boldsymbol{y}^*} = \underbrace{\boldsymbol{W}^{1/2}\boldsymbol{X}}_{\boldsymbol{X}^*}\boldsymbol{\beta} + \underbrace{\boldsymbol{W}^{1/2}\boldsymbol{\epsilon}}_{\boldsymbol{\epsilon}^*}.$$

Die Transformation hat die Heteroskedastizität von $\boldsymbol{\epsilon}$ behoben und das Modell

$$\boldsymbol{y}^* = \boldsymbol{X}^*\boldsymbol{\beta} + \boldsymbol{\epsilon}^* \tag{2.19}$$

kann mit den bekannten Werkzeugen der gewöhnlichen linearen Regression gehandhabt werden. Betrachtet man den KQ-Schätzer des Modells (2.19), so ergibt sich:

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= ((\boldsymbol{X}^*)'\boldsymbol{X}^*)^{-1}(\boldsymbol{X}^*)'\boldsymbol{y}^* \\ &= ((\boldsymbol{W}^{1/2}\boldsymbol{X})'(\boldsymbol{W}^{1/2}\boldsymbol{X}))^{-1}(\boldsymbol{W}^{1/2}\boldsymbol{X})'(\boldsymbol{W}^{1/2}\boldsymbol{y}) \\ &= (\boldsymbol{X}'\boldsymbol{W}^{1/2}\boldsymbol{W}^{1/2}\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{W}^{1/2}\boldsymbol{W}^{1/2}\boldsymbol{y}) \\ &= (\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}\boldsymbol{y}.\end{aligned}$$

Der KQ-Schätzer mit dieser Gestalt ist in der Fachliteratur als *Aitken-* oder *gewichteter KQ-Schätzer* bekannt. Aus Gründen der Verständlichkeit wird der gewichtete KQ-Schätzer mit

$$\hat{\boldsymbol{\beta}}_w = (\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}\boldsymbol{y} \tag{2.20}$$

bezeichnet.

2.2.2 Gewichteter Kleinster Quadrate Schätzer

Um die Regressionskoeffizienten $\boldsymbol{\beta}$ des Modells (2.17) zu schätzen, verwendet man die gewichtete Methode der kleinsten Quadrate. Mit diesem Hilfsmittel versucht man, die Fehlerquadratsumme $\mathrm{SSE_w}(\boldsymbol{\beta})$ zu minimieren. Im gewichteten Fall sieht diese wie folgt aus

$$\mathrm{SSE_w}(\boldsymbol{\beta}) = \sum_{i=1}^{n} w_i(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2 = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{W}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}).$$

Das Ziel besteht also darin, einen Vektor $\boldsymbol{\beta}$ zu finden, der $\mathrm{SSE_w}(\boldsymbol{\beta})$ minimiert. Wir werden nun zeigen, dass der zuvor ermittelte gewichtete KQ-Schätzer $\hat{\boldsymbol{\beta}}_w$ die Fehlerquadratsumme $\mathrm{SSE_w}(\boldsymbol{\beta})$ minimiert. Die Vorgehensweise lautet wie folgt

$$\begin{aligned}\mathrm{SSE_w}(\boldsymbol{\beta}) &= (\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{W}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) \\ &= \boldsymbol{y}'\boldsymbol{W}\boldsymbol{y} - \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{W}\boldsymbol{y} - \boldsymbol{y}'\boldsymbol{W}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}\boldsymbol{\beta} \\ &= \boldsymbol{y}'\boldsymbol{W}\boldsymbol{y} - 2\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{W}\boldsymbol{y} + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}\boldsymbol{\beta}.\end{aligned}$$

Bei den Termen $\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{W}\boldsymbol{y}$ und $\boldsymbol{y}'\boldsymbol{W}\boldsymbol{X}\boldsymbol{\beta}$ handelt es sich um Skalare. Damit ist der eine Term das Transponierte vom anderen und man kann diese zusammenfassen.

Durch Ableiten des $\mathrm{SSE_w}(\boldsymbol{\beta})$ nach dem Parametervektor $\boldsymbol{\beta}$ erhält man

$$\frac{\partial}{\partial \boldsymbol{\beta}} \mathrm{SSE_w}(\boldsymbol{\beta}) = -2\boldsymbol{X}'\boldsymbol{W}\boldsymbol{y} + 2\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}\boldsymbol{\beta}.$$

Nochmaliges Differenzieren liefert

$$\frac{\partial^2}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'} \mathrm{SSE_w}(\boldsymbol{\beta}) = 2\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}.$$

Das Nullsetzen der ersten Ableitung liefert die so genannten Normalengleichungen

$$(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})\hat{\boldsymbol{\beta}}_w = \boldsymbol{X}'\boldsymbol{W}\boldsymbol{y}. \tag{2.21}$$

Da $\boldsymbol{X}$ vollen Spaltenrang besitzt und $\boldsymbol{W}$ positiv definit ist, folgt, dass die Matrix $\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}$ ebenfalls positiv definit ist (siehe Satz A.2). Somit ist $\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}$ regulär und daher invertierbar (siehe Satz A.1) und wir schreiben

$$\hat{\boldsymbol{\beta}}_w = (\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}\boldsymbol{y}.$$

Die Matrix $\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}$ der zweiten Ableitung ist positiv definit, somit handelt es sich bei dem gewichteten KQ-Schätzer um die Minimalstelle von $\mathrm{SSE_w}$.

Eigenschaften des gewichteten KQ-Schätzers

Da alle Annahmen (GM.1 - GM.4) des Gauß-Markow-Theorems für das Modell (2.19) erfüllt sind, folgt, dass $\hat{\boldsymbol{\beta}}_w$ der beste lineare erwartungstreue (engl. BLUE) Schätzer ist.

Die Varianz-Kovarianz-Matrix des gewichteten KQ-Schätzers lautet im heteroskedastischen Fall

$$\begin{aligned}\mathrm{Var}(\hat{\boldsymbol{\beta}}_w) &= \mathrm{Var}((\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}\boldsymbol{y}) \\ &= (\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}\,\mathrm{Var}(\boldsymbol{y})((\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W})' \\ &= (\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}\sigma^2\boldsymbol{W}^{-1}(\boldsymbol{W}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}) \\ &= \sigma^2(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1} \\ &= \sigma^2(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}. \end{aligned} \tag{2.22}$$

Somit fließen die bekannten Gewichte w_i in die Varianz-Kovarianz-Matrix ein.

2.2.3 Maximum-Likelihood-Schätzer

Die Maximum-Likelihood-Methode dient zur Konstruktion von Parameterschätzern bei gegebener Verteilung. Die Maximum-Likelihood-Schätzer besitzen eine Vielzahl von wünschenswerten Eigenschaften. Sie sind z.B. unter allgemeinen Bedingungen konsistent, asymptotisch normalverteilt und asymptotisch effizient. Im Anschluss werden die Maximum-Likelihood-Schätzer für die Parameter $\boldsymbol{\beta}$ und σ^2 hergeleitet.

Seien $y_1, \ldots, y_n \overset{ind}{\sim} Normal(\boldsymbol{x}_i'\boldsymbol{\beta}, \sigma_i^2)$, wobei $\sigma_i^2 = \frac{\sigma^2}{w_i}$ sind, mit $\boldsymbol{\beta}$ und σ^2 unbekannt. Die Gewichte w_i sind bekannte positive Konstanten. Die Likelihood-Funktion ist gegeben durch

$$\begin{aligned} L(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}, \boldsymbol{W}) &= \prod_{i=1}^n f(y_i|\boldsymbol{\beta}, \sigma_i^2) = \prod_{i=1}^n \left(\frac{1}{2\pi\sigma_i^2}\right)^{1/2} \exp\left\{-\frac{(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2}{2\sigma_i^2}\right\} \\ &= \prod_{i=1}^n \left(\frac{1}{2\pi\sigma^2}\right)^{1/2} \prod_{i=1}^n \sqrt{w_i} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n w_i(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2\right\} \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left\{-\frac{1}{2\sigma^2}\,\mathrm{SSE_w}(\boldsymbol{\beta})\right\} \prod_{i=1}^n w_i^{1/2}. \end{aligned}$$

Der Maximum-Likelihood-Schätzer (MLE) maximiert die (Log-) Likelihood-Funktion

$$\log f(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2, \boldsymbol{W}) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\,\mathrm{SSE_w}(\boldsymbol{\beta}) + \frac{1}{2}\sum_{i=1}^n \log w_i.$$

Die Maximierung von $\log f(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2, \boldsymbol{W})$ in $\boldsymbol{\beta}$ ist somit unabhängig von den Werten σ^2 und $\boldsymbol{W}$ und äquivalent mit der Minimierung von $\mathrm{SSE_w}(\boldsymbol{\beta})$. Laut (2.20) minimiert der gewichtete KQ-Schätzer $\hat{\boldsymbol{\beta}}_w$ die gewichtete Fehlerquadratsumme $\mathrm{SSE_w}(\boldsymbol{\beta})$ und daher maximiert $\hat{\boldsymbol{\beta}}_w$ die Log-Likelihood-Funktion. Daher entspricht der MLE $\hat{\boldsymbol{\beta}}_w$

dem KQ-Schätzer $\hat{\boldsymbol{\beta}}_w$ und besitzt auch dessen Eigenschaften.

Die Bestimmung des Maximum-Likelihood-Schätzers der Varianz erfolgt durch die partielle Ableitung der Log-Likelihood-Funktion nach σ^2 für gegebenes $\hat{\boldsymbol{\beta}}_w$.

$$\frac{\partial \log f(\boldsymbol{y}|\hat{\boldsymbol{\beta}}_w, \sigma^2, \boldsymbol{W})}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \operatorname{SSE_w}(\hat{\boldsymbol{\beta}}_w).$$

Das Nullsetzen der Ableitung liefert die Normalengleichung

$$\frac{n}{2\hat{\sigma}_w^2} = \frac{1}{2\hat{\sigma}_w^4} \operatorname{SSE_w}(\hat{\boldsymbol{\beta}}_w).$$

Durch elementare Umformungen ergibt sich der Maximum-Likelihood-Schätzer der Varianz

$$\hat{\sigma}_w^2 = \frac{1}{n} \operatorname{SSE_w}(\hat{\boldsymbol{\beta}}_w). \tag{2.23}$$

Wie im gewöhnlichen linearen Modell ist dieser Schätzer verzerrt. Ein unverzerrter Schätzer ist laut Fahrmeir et al. (2009, S. 126) durch

$$\hat{\sigma}^2 = \frac{1}{n-p} \operatorname{SSE_w}(\hat{\boldsymbol{\beta}}_w) \tag{2.24}$$

gegeben.

2.2.4 Residuen im gewichteten Regressionsmodell

Die Analyse der Residuen ist für jedes Regressionsmodell wichtig. Mit diversen Hypothesentests bzw. Diagnoseplots kann überprüft werden, ob die Modellannahmen eines MLRs korrekt erfüllt worden sind. Ausgehend vom gewichteten Regressionsmodell (2.17) befasst man sich mit den rohen Residuen. Diese sind definiert als

$$\begin{aligned} \boldsymbol{r} &= \boldsymbol{y} - \hat{\boldsymbol{\mu}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_w \\ &\overset{(2.20)}{=} \boldsymbol{y} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}\boldsymbol{y} \\ &= (\boldsymbol{I} - \boldsymbol{H}_w)\boldsymbol{y}, \end{aligned} \tag{2.25}$$

wobei $\boldsymbol{H}_w = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}$ die gewichtete Hat-Matrix ist. Die Residuen (2.25) bringen, laut Weisberg (2005, S. 171), jedoch schwerwiegende Defizite mit sich. Erstens entspricht die Summe der quadrierten Residuen nicht der Residuenquadratsumme, da die Gewichte ignoriert werden. Zweitens hängt die Varianz des i-ten Residuums vom i-ten Gewicht w_i ab. Beide Probleme können durch die nachfolgende Definition, dem gewichteten Residuum, gelöst werden

$$r_i^* := \sqrt{w_i}(y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_w) = \sqrt{w_i} \cdot r_i. \tag{2.26}$$

Bei der Residuenanalyse des gewichteten linearen Regressionsmodells sind die gewichteten Residuen (2.26) den rohen Residuen (2.25) vorzuziehen.

2.2.5 Schätzung der Gewichte

Die Theorie der gewichteten linearen Regression basiert darauf, dass die Struktur der Heteroskedastizität und somit die Gewichtsmatrix $\boldsymbol{W}$ bekannt sind. Daraus resultiert, dass der gewichtete KQ-Schätzer der beste lineare erwartungstreue Schätzer (BLUE) ist. In der Praxis ist jedoch die Gewichtsmatrix $\boldsymbol{W}$ oftmals unbekannt und muss zusätzlich zum Koeffizientenvektor $\boldsymbol{\beta}$ und der Varianz σ^2 aus den Daten geschätzt werden. In der englischsprachigen Literatur werden gewichtete KQ-Schätzer, welche auf geschätzte Gewichte $\widehat{\boldsymbol{W}}$ basieren, auch als **FGLS-Schätzer** (Feasible Generalized Least Squares Estimator) bezeichnet. Dieser lautet

$$\hat{\boldsymbol{\beta}}_{FGLS} = (\boldsymbol{X}'\widehat{\boldsymbol{W}}\boldsymbol{X})^{-1}\boldsymbol{X}'\widehat{\boldsymbol{W}}\boldsymbol{y}. \tag{2.27}$$

Ein problematischer Aspekt von geschätzten Gewichten ist, dass die Voraussetzungen des Gauß-Markow-Theorems nicht mehr erfüllt sind. Damit ist (2.27) nicht mehr BLUE. Des Weiteren ist der FGLS-Schätzer im Allgemeinen verzerrt, bleibt jedoch konsistent (Wooldridge, 2005, S. 295).

Unter der Voraussetzung, dass die Gewichtsmatrix konsistent geschätzt wurde, d.h. $\widehat{\boldsymbol{W}} \xrightarrow{P} \boldsymbol{W}$, lässt sich zeigen, dass $\hat{\boldsymbol{\beta}}_{FGLS}$ asymptotisch äquivalent zu $\hat{\boldsymbol{\beta}}_w$ ist. Somit ist der FGLS-Schätzer asymptotisch BLUE, d.h. asymptotisch erwartungstreu, asymptotisch effizient und konsistent. Dies ist jedoch nur gegeben, wenn die korrekte Struktur der Heteroskedastizität gewählt wurde. Im Allgemeinen ist es jedoch kaum möglich, Aussagen über endliche Stichprobeneigenschaften des $\hat{\boldsymbol{\beta}}_{FGLS}$ zu treffen (Baltagi, 2008, S. 226).

Wird die Gewichtsmatrix $\boldsymbol{W}$ jedoch nicht konsistent geschätzt, sind die Standardfehler des FGLS-Schätzers verzerrt. Die asymptotische Varianz-Kovarianz-Matrix des FGLS-Schätzers lautet

$$\mathrm{AVar}(\hat{\boldsymbol{\beta}}_{FGLS}) = (\boldsymbol{X}'\widehat{\boldsymbol{W}}\boldsymbol{X})^{-1}\boldsymbol{X}'\widehat{\boldsymbol{W}}\,\mathrm{Var}(\boldsymbol{\epsilon})\widehat{\boldsymbol{W}}\boldsymbol{X}(\boldsymbol{X}'\widehat{\boldsymbol{W}}\boldsymbol{X})^{-1}. \tag{2.28}$$

Diese kann nun auch im gewichteten Fall durch einen der heteroskedastisch konsistenten Schätzer HC0 bis HC3 geschätzt werden (Greene, 2002, S. 319). Die robuste Matrix HC0 ist im gewichteten Fall durch

$$\mathrm{HC0}_w = (\boldsymbol{X}'\widehat{\boldsymbol{W}}\boldsymbol{X})^{-1}\left[\sum_{i=1}^{n}(\hat{w}_{ii}r_i)^2\boldsymbol{x}_i'\boldsymbol{x}_i\right](\boldsymbol{X}'\widehat{\boldsymbol{W}}\boldsymbol{X})^{-1}$$

gegeben, wobei r_i das i-te rohe Residuum ist. Kann die Gewichtsmatrix $\boldsymbol{W}$ aufgrund von fehlendem Wissen über die Struktur der Heteroskedastizität nicht konsistent geschätzt werden, wird standardmäßig gewöhnliche lineare Regression mit robusten

Standardfehlern (2.11) – (2.15) verwendet (Greene, 2002, S. 319).

Angesichts der Existenz von heteroskedastisch robusten Standardfehler stellt sich jedoch die Frage, inwiefern die gewichtete lineare Regression mit geschätzten Gewichten einen Vorteil gegenüber der gewöhnlichen linearen Regression hat. Laut Greene (2002, S. 319) ist in eindeutigen Fällen von Heteroskedastizität die gewöhnliche lineare Regression extrem ineffizient. Des Weiteren ist es laut Wooldridge (2005, S. 291) in Fällen von starker Heteroskedastizität oftmals besser, gewichtete lineare Regression mit falsch gewählten Gewichten zu verwenden, als das Problem der Heteroskedastizität zu ignorieren, wie es bei der gewöhnlichen linearen Regression der Fall wäre.

Im Anschluss werden zwei Ansätze betrachtet, um die Gewichtsmatrix $\boldsymbol{W}$ zu schätzen.

Direkte Schätzung im Falle wiederholter Messungen

Ein Ziel der gewichteten linearen Regression ist es, jedem Datenpunkt einen bestimmten Einfluss zuzuweisen, so dass bei der Bestimmung der Parameterschätzer die Datenpunkte in angemessener Art und Weise in die Berechnung einfließen. Beobachtungen mit großen Varianzen tragen weniger Information zur Schätzung bei und sollten daher weniger stark gewichtet werden als jene mit kleinerer Varianz.

Optimale Resultate erhält man, wenn die Gewichte w_i die Kehrwerte der Varianzen σ_i^2 sind, d.h.

$$w_i = \frac{1}{\sigma_i^2},$$

mit $\mathrm{Var}(y_i) = \sigma_i^2$. Unglücklicherweise sind diese optimalen Gewichte, die von den echten Varianzen jedes Datenpunktes abhängig sind, unbekannt.

Im Falle mehrerer Beobachtungen n_i der Response für eine bestimmte Prädiktoreinstellung i, kann die Stichprobenvarianz s^2 dieser Response für die Schätzung der Gewichte herangezogen werden.

Das geschätzte Gewicht der j-ten Beobachtung in der i-ten Einstellung ist dann

$$\hat{w}_{ij} = \frac{1}{s_i^2} \qquad \text{mit} \qquad s_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2,$$

für $i = 1, \ldots, k$. Die Gewichtsmatrix $\widehat{\boldsymbol{W}}$ entspricht in diesem Fall der Diagonalmatrix $\mathrm{diag}(\hat{w}_{11}, \ldots, \hat{w}_{1n_1}, \ldots, \hat{w}_{k1}, \ldots, \hat{w}_{kn_k})$.

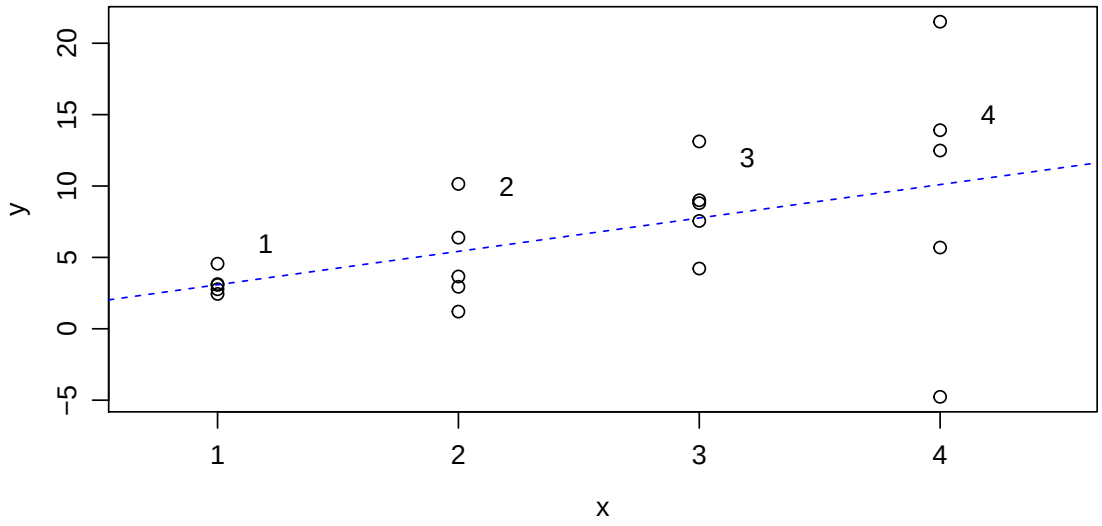

Abbildung 2.2: Für jede der $k = 4$ Einstellungen existieren $n_j = 5$ Beobachtungen. Die daraus resultierenden Punktwolken weisen nicht idente Streuungen auf.

Diese Methode der Gewichtsschätzung hat jedoch einen Nachteil. Gewichte, die durch die direkte Methode geschätzt werden, sind sehr variabel. Der daraus resultierende Einfluss der einzelenen Datenpunkte auf die Parameterschätzer ist nicht mehr korrekt. Um passende Gewichte zu erhalten, benötigt die direkte Methode jedoch eine sehr große Anzahl von Replikationen der Response für jede Prädiktorkombination (NIST/SEMATECH, 2012, Kap. 4).

Zweistufiges Schätzverfahren

Eine weitere Möglichkeit, die Gewichte zu schätzen, liefert das zweistufige Schätzverfahren (Fahrmeir et al., 2009, S. 133 f.). Zweistufiges Schätzverfahren deshalb, da zwei multiple lineare Regressionsschätzungen in drei Schritten durchgeführt werden.

Allgemein gilt für den nicht beobachtbaren statistischen Fehler $\mathbb{E}(\epsilon_i) = 0$. Der Verschiebungssatz liefert $\mathrm{Var}(\epsilon_i) = \mathbb{E}(\epsilon_i^2) - \mathbb{E}(\epsilon_i)^2 = \sigma_i^2$. Daher kann ϵ_i^2 als

$$\epsilon_i^2 = \sigma_i^2 + v_i$$

dargestellt werden, wobei v_i die Differenz zwischen dem quadrierten statistischen Fehler ϵ_i^2 und dessen Erwartungswert σ_i^2 ist. Da die Varianz σ_i^2 heteroskedastisch ist, hängt diese für gewöhnlich von einer oder mehreren Prädiktorvariablen ab. Daher resultiert folgende Annahme für die Varianz:

$$\sigma_i^2 = \alpha_0 + \alpha_1 z_{i1} + \alpha_2 z_{i2} + \cdots + \alpha_q z_{iq} = \boldsymbol{z}_i' \boldsymbol{\alpha}.$$

Der Vektor $\boldsymbol{z}_i$ besteht somit aus jenen Variablen, die einen Einfluss auf die Varianz haben. Der Schätzer $\hat{\boldsymbol{\alpha}}$ und die geschätzten Gewichte $\hat{w}_i$ für das lineare Modell zwischen $\boldsymbol{y}$ und $\boldsymbol{x}$ werden in drei Schritten ermittelt.

1. Schritt
Im ersten Schritt wird eine gewöhnliche ungewichtete lineare Regression zwischen $\boldsymbol{y}$ und $\boldsymbol{x}$ durchgeführt. Dadurch erhält man den ungewichteten KQ-Schätzer $\hat{\boldsymbol{\beta}}$ und die Residuen $r_i = y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}$ des ungewichteten Modells.

2. Schritt
Um den Koeffizientenvektor $\boldsymbol{\alpha}$ zu schätzen, ist eine lineare Regression zwischen dem quadrierten statistischen Fehler ϵ_i^2 als Zielgröße und dem Vektor $\boldsymbol{z}_i'$ als abhängige Variable nötig. Da ϵ_i^2 nicht beobachtbar ist, müssen die ϵ_i durch die zuvor ermittelten Residuen r_i geschätzt werden.

Der Schätzer $\hat{\boldsymbol{\alpha}}$ ergibt sich nun aus einer gewöhnlichen Regression zwischen den quadrierten Residuen r_i^2 und dem Vektor $\boldsymbol{z}_i$. Da $\boldsymbol{z}_i$ jener Vektor ist, der die Varianzen beeinflusst, wird dieser in den meisten Fällen identisch mit dem Prädiktorvektor $\boldsymbol{x}$ sein. Dieses Verfahren ist jedoch in der Praxis nicht immer anwendbar, da $\boldsymbol{z}_i'\hat{\boldsymbol{\alpha}}$ auch negative Werte annehmen kann, was in negativen Schätzern für die Varianzen resultiert. Um positive Varianzen zu garantieren, betrachtet man alternativ ein loglineares Modell. Die Varianzen σ_i^2 sind nun wie folgt gegeben:

$$\log(\sigma_i^2) = \boldsymbol{z}_i'\boldsymbol{\alpha}.$$

Aufgrund dessen ergibt sich ein neues Regressionsmodell, aus dem der Vektor $\boldsymbol{\alpha}$ geschätzt werden kann

$$\log(r_i^2) = \boldsymbol{z}_i'\boldsymbol{\alpha} + \tilde{v}_i.$$

3. Schritt
Im abschließenden Schritt werden mit Hilfe des zuvor ermittelten Vektors $\hat{\boldsymbol{\alpha}}$ die Gewichte für die Regression zwischen $\boldsymbol{y}$ und $\boldsymbol{x}$ durch

$$\hat{w}_i = \frac{1}{\exp(\boldsymbol{z}_i'\hat{\boldsymbol{\alpha}})}$$

geschätzt. Somit kann eine gewichtete Regressionsanalyse durchgeführt werden.

2.2.6 Akaike's Informationskriterium

Möchte man zwischen unterschiedlichen statistischen Modellen mit diversen Prädiktoren und Parametern auswählen, so muss ein Mittelweg zwischen zu großer

Modellkomplexität von Parametern und guter Datenanpassung gefunden werden. Ein beliebtes Werkzeug für die Modellauswahl ist das Akaike Informationskriterium (AIC). Hierbei ist die negative Likelihood zum Modell ein Maß für die Anpassungsgüte, während p ein Maß für die Modellkomplexität ist. Das AIC lautet im gewichteten Fall

$$\text{AIC} = 2[-\log L(\hat{\boldsymbol{\beta}}_w, \hat{\sigma}_w^2, \boldsymbol{y}, \boldsymbol{W}) + (p+1)]. \tag{2.29}$$

Das Einsetzen des gewichteten KQ-Schätzer (2.20) und des gewichteten Varianzschätzers (2.23) liefert

$$\begin{aligned}
\text{AIC} &= n\log(2\pi) + n\log(\hat{\sigma}_w^2) + \frac{1}{\hat{\sigma}_w^2}\,\text{SSE}_\text{w}(\hat{\boldsymbol{\beta}}_w) - \sum_{i=1}^{n}\log w_i + 2(p+1) \\
&= n\log(2\pi) + n\log\left(\frac{\text{SSE}_\text{w}(\hat{\boldsymbol{\beta}}_w)}{n}\right) + \frac{n}{\text{SSE}_\text{w}(\hat{\boldsymbol{\beta}}_w)}\,\text{SSE}_\text{w}(\hat{\boldsymbol{\beta}}_w) \\
&\quad - \sum_{i=1}^{n}\log w_i + 2(p+1) \\
&= n(\log(2\pi)+1) + n\log\left(\frac{\text{SSE}_\text{w}(\hat{\boldsymbol{\beta}}_w)}{n}\right) - \sum_{i=1}^{n}\log w_i + 2(p+1).
\end{aligned} \tag{2.30}$$

In `R` kann das Informationskriterium mit dem Kommando `AIC(mod)` für ein gewichtetes lineares Modell `mod` bestimmt werden.

2.3 Monte Carlo Simulation

Im Folgenden wird eine Monte Carlo (MC) Simulation mit $N = 1000$ Replikationen durchgeführt. Das Ziel ist es, die Auswirkungen von Heteroskedastizität auf die diversen KQ-Schätzer sowie auf deren Standardfehler aufzuzeigen. Die Daten werden aus dem Modell

$$y_i = 10 + 3 \cdot x_i + \epsilon_i$$

mit $\epsilon_i \overset{ind}{\sim} Normal(0, 4x_i^2)$ für $i = 1, \ldots, n = 100$ generiert. Die Werte des Prädiktors x_i liegen auf dem Einheitsintervall, beginnend bei 0.01 und fortführend bis 1 mit äquidistanten Abständen.

In jeder Replikation der MC-Simulation werden eine gewöhnliche sowie zwei gewichtete lineare Regressionen gerechnet. Die Gewichtung erfolgt einmal mit geschätzten Gewichten, die aus dem zweistufigen Verfahren ermittelt werden, sowie mit exakten Gewichten $w_i = \frac{1}{x_i^2}$. Wie im nachfolgenden `R`-Code ersichtlich, werden in jeder

Replikation diverse Steigungsschätzer und deren Standardfehler berechnet. Zusätzlich werden die heteroskedastisch konsistenten Standardfehler HC0 bis HC3 aus dem Package `sandwich` bestimmt (Zeileis, 2006).

```
> library(sandwich); set.seed(501);
# Initialisierung der Dataframes
> ols.beta <- data.frame(Steigung=numeric())
> wls.beta <- ols.beta; fgls.beta <- ols.beta
> ols.se <- data.frame("SE Steigung"=numeric())
> hc0.se <- ols.se; hc1.se <- ols.se; hc2.se <- ols.se;
> hc3.se <- ols.se; wls.se <- ols.se; fgls.se <- ols.se;
> fgls.hc0.se <- ols.se; fgls.hc3.se <- ols.se
# Initialisierung Simulation
> N.rep <- 1000; n <- 100; x <- (1:100)/n
> a <- 10; b <-3; sigma2 <- 4; sigma2.eps <- sigma2*x^2;
> wts.exakt <- 1/x^2
> for(i in 1:N.rep){
   eps <- rnorm(n,mean=rep(0,n),sd=sqrt(sigma2.eps))
   y <- a+b*x + eps
   # OLS
   mod <- lm(y ~ x)
   ols.beta[i,] <- mod$coefficients[2]
   ols.se[i,] <- sqrt(diag(vcov(mod)))[2]
   hc0.se[i,] <- sqrt(diag(vcovHC(mod,type="HC0")))[2]
   hc1.se[i,] <- sqrt(diag(vcovHC(mod,type="HC1")))[2]
   hc2.se[i,] <- sqrt(diag(vcovHC(mod,type="HC2")))[2]
   hc3.se[i,] <- sqrt(diag(vcovHC(mod,type="HC3")))[2]
   # WLS exakt
   mod.w <- lm(y ~ x, weights = wts.exakt)
   wls.beta[i,] <- mod.w$coefficients[2]
   wls.se[i,] <- sqrt(diag(vcov(mod.w)))[2]
   # FGLS zweistufige Methode
   res <- residuals(mod)
   mod.res.lm <- lm(log(res^2) ~ x)
   wts <- 1/exp(fitted(mod.res.lm))
   mod.fgls <- lm(y ~ x, weights = wts)
   fgls.beta[i,] <- mod.fgls$coefficients[2]
   fgls.se[i,] <- sqrt(diag(vcov(mod.fgls)))[2]
   fgls.hc0.se[i,] <- sqrt(diag(vcovHC(mod.fgls,type="HC0")))[2]
   fgls.hc3.se[i,] <- sqrt(diag(vcovHC(mod.fgls,type="HC3")))[2]
}
```

Anschließend werden die entsprechenden MC-Standardabweichungen der diversen Steigungsschätzer ermittelt.

```
> beta.sd <- apply(ols.beta,2,sd); beta.sd
0.4411332
> beta.wls.sd <- apply(wls.beta,2,sd);  beta.wls.sd
0.21528594
> beta.fgls.sd <- apply(fgls.beta,2,sd); beta.fgls.sd
0.27402899
```

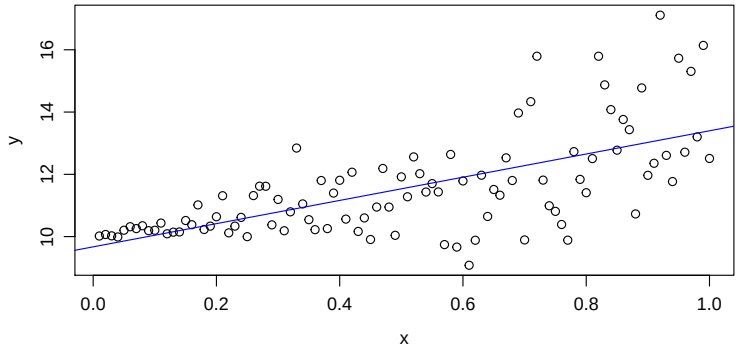

Abbildung 2.3: Eine exemplarische Datensituation mit Regressionsgerade einer Replikation der MC-Simulation.

In Abbildung 2.3 sind eine generierte Datensituation und die dazugehörige Regressionsgerade ersichtlich. Mit steigendem x nimmt auch die Varianz der Response y zu.

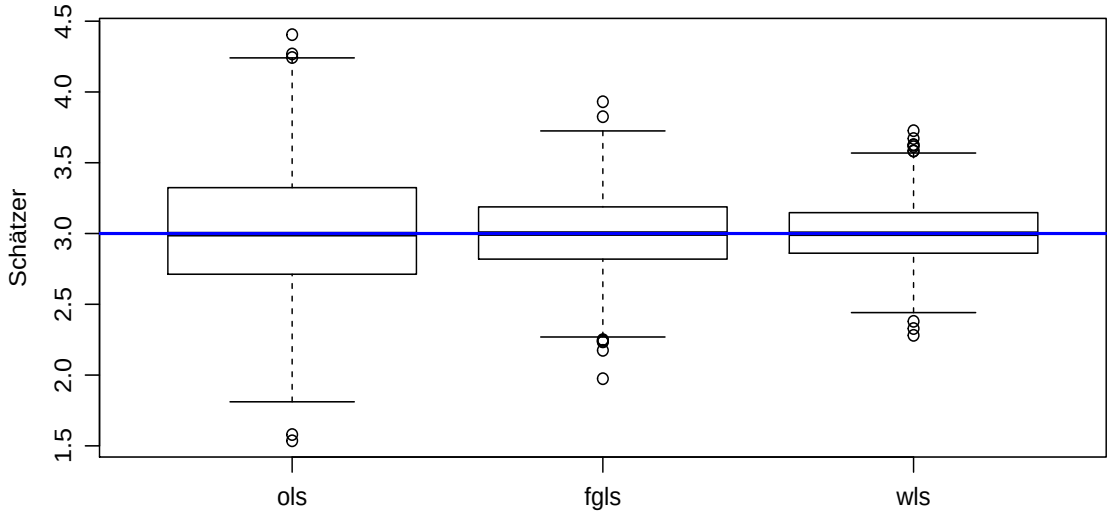

Abbildung 2.4: Boxplots der 1000 Steigungsschätzer $\hat{\boldsymbol{\beta}}_1$ bei heteroskedastischen Responses y_i. Die blaue Linie entspricht dem wahren Wert $\beta_1 = 3$.

Links in Abbildung 2.4 befindet sich der Boxplot der gewöhnlichen Schätzer (OLS), mittig der Boxplot der FGLS-Schätzer und rechts der Boxplot der gewichteten Schätzer (WLS). In allen drei Fällen gibt es ein paar Ausreißer an den Enden, jedoch wird für alle drei Verfahren $\beta_1 = 3$ im Median unverzerrt geschätzt. Die Heteroskedastizität wirkt sich auf die interquartilen Bereiche (engl. IQR) der einzelnen Boxplots aus. Der Boxplot der WLS-Schätzer weist aufgrund der korrekt gewählten Gewichte den kleinsten IQR auf, wogegen der Boxplot der OLS-Schätzer mit Abstand den größten IQR besitzt, da keine Gewichte gewählt wurden, um der nicht konstanten Varianz von ϵ_i entgegenzuwirken.

Wie bereits aus Abschnitt 2.1.2 bekannt, lautet die Varianz-Kovarianz-Matrix des gewöhnlichen KQ-Schätzers im Falle von Heteroskedastizität

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\,\mathrm{Var}(\boldsymbol{\epsilon})\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}.$$

Die wahre Standardabweichung kann mit dem nachfolgenden R-Code errechnet werden.

```
> X <- matrix(c(rep(1,n),x),nrow=n, ncol=2)
> XtX_inv <- solve(t(X) %*% X)
> var_eps <- diag(x^2,n,n)
> sqrt(diag(XtX_inv %*% t(X) %*% var_eps %*% X %*% XtX_inv) * sigma2)[2]
0.4409248
```

Es ist sofort ersichtlich, dass die Standardabweichung des Steigungsschätzers $sd(\hat{\beta}_1) = 0.4409$ mit dem zuvor ermittelten MC-Schätzer für den gewöhnlichen Fall $\widehat{sd}_{MC}(\hat{\beta}_1) = 0.4411$ nahezu identisch ist.

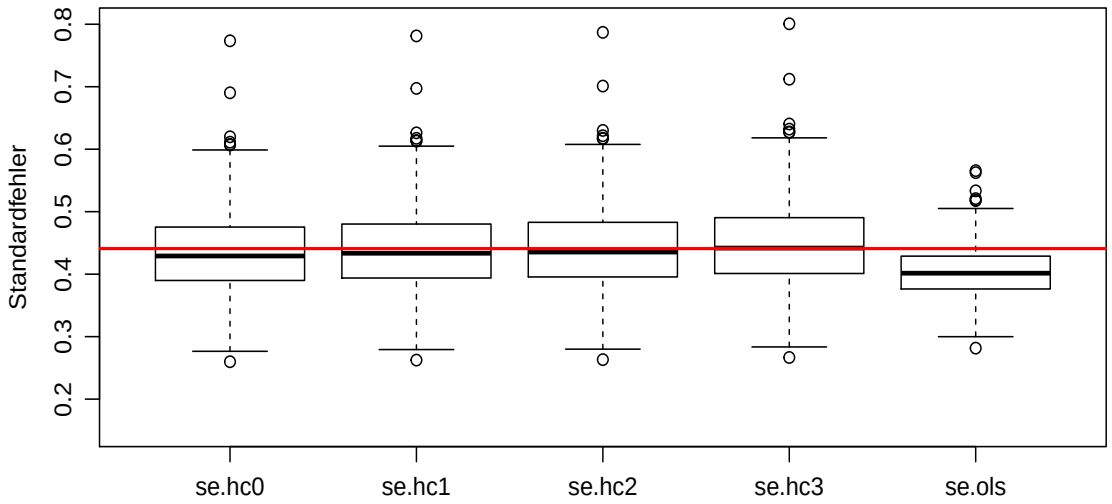

Abbildung 2.5: Boxplots der 1000 Standardfehler des ungewichteten Steigungsschätzer $\widehat{sd}(\hat{\beta}_1)$. Die rote Linie entspricht der wahren Standardabweichung von 0.4409.

In Abbildung 2.5 fällt sofort auf, dass der Standardfehler des gewöhnlichen Schätzers (se.ols) im Median verzerrt ist und den wahren Wert (rote Linie) unterschätzt. Die heteroskedastisch robusten Standardfehler se.hc0 bis se.hc3 weichen im Median nur minimal vom wahren Standardfehler (rote Linie) ab, wobei das beste Ergebnis im Median vom robusten Standardfehler se.hc3 erzielt wird. Dieser stimmt mit dem wahren Wert im Median fast perfekt überein.

Wie bereits aus Abschnitt 2.2.2 bekannt, ist der gewichtete KQ-Schätzer der beste lineare erwartungstreue Schätzer und besitzt somit die kleinste Varianz-Kovarianz-Matrix. Diese ist durch

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}_w) = \sigma^2(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}$$

gegeben und kann mit dem nachfolgenden R-Code berechnet werden.

```
> W <- diag(wts.exakt, n,n)
> XtWX_inv <- solve(t(X) %*% W %*% X)
> sqrt(diag(XtWX_inv) * sigma2)[2]
0.21881549
```

Auch hier sind der zu Beginn errechnete MC-Schätzer für den gewichteten Fall $\widehat{sd}_{MC}(\hat{\beta}_{w1}) = 0.2153$ und die Standardabweichung des gewichteten Steigungsschätzers $sd(\hat{\beta}_{w1}) = 0.2188$ nahezu ident.

Die Qualität des FGLS-Schätzers hängt davon ab, ob die Gewichtsmatrix $\boldsymbol{W}$ konsistent geschätzt wird. Ist dies nicht der Fall, wird der Standardfehler des FGLS-Schätzers verzerrt sein. Diesem Problem kann auch im gewichteten Fall mit robusten Matrizen $\mathrm{HC0}_w$ bis $\mathrm{HC3}_w$ entgegen gewirkt werden. Die asymptotische Varianz-Kovarianz-Matrix ist laut (2.28) durch

$$\mathrm{AVar}(\hat{\boldsymbol{\beta}}_{FGLS}) = (\boldsymbol{X}'\widehat{\boldsymbol{W}}\boldsymbol{X})^{-1}\boldsymbol{X}'\widehat{\boldsymbol{W}}\,\mathrm{Var}(\boldsymbol{\epsilon})\widehat{\boldsymbol{W}}\boldsymbol{X}(\boldsymbol{X}'\widehat{\boldsymbol{W}}\boldsymbol{X})^{-1}$$

gegeben und von den geschätzten Gewichten $\widehat{\boldsymbol{W}}$ abhängig.

In der nachfolgenden Abbildung 2.6 sind die Boxplots der Standardfehler des FGLS-Schätzers (se.fgls) sowie die zwei robusten Schätzer der Standardfehler se.fgls.hc0 und se.fgls.hc3 abgebildet. Die rote Linie entspricht der MC-Standardabweichung aller 1000 FGLS-Schätzer (0.274) und somit einer Approximation des entsprechenden „wahren Standardfehlers“ für den FGLS-Fall.

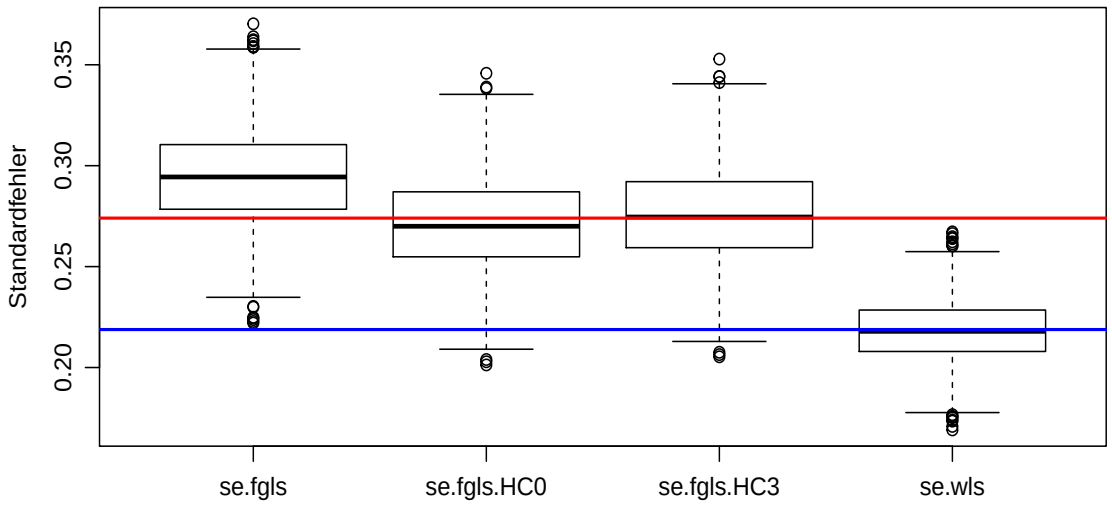

Abbildung 2.6: Boxplots der 1000 Standardfehler des FGLS-Steigungsschätzers $\widehat{sd}(\hat{\beta}_{FGLS1})$ sowie der 1000 Standardfehler des gewichteten KQ-Schätzers $\widehat{sd}(\hat{\beta}_{w1})$. Die rote Linie entspricht der MC-Standardabweichung der 1000 FGLS-Schätzer (0.274) und die blaue Linie entspricht der wahren Standardabweichung im gewichteten Fall (0.2188).

Da die Gewichtsmatrix nicht konsistent geschätzt wurde, wird der Standardfehler des FGLS-Schätzers verzerrt sein. Dies spiegelt sich auch in Abbildung 2.6 wider, denn der Standardfehler se.fgls überschätzt die MC-Standardabweichung im FGLS-Fall (rote Linie). Die heteroskedastisch robusten Standardfehler se.fgls.hc0 und se.fgls.hc3 weichen im Median nur minimal von der MC-Standardabweichung (rote Linie) ab, wobei das beste Ergebnis vom robusten Standardfehler HC3 (se.fgls.hc3) erzielt wird. Die robuste Matrix HC3 konnte somit im gewichteten sowie im ungewichteten Fall überzeugen.

Weiters sind in Abbildung 2.6 der Boxplot der Standardfehler des gewichteten Schätzers (se.wls) und eine horizontale blaue Linie abgebildet. Diese blaue Linie entspricht der Standardabweichung des gewichteten Schätzers (0.2188). Die Standardfehler des gewichteten Schätzers schätzen die Standardabweichung im Median nahezu exakt. Außerdem ist in den Abbildungen 2.5 und 2.6 aufgrund der korrekt gewählten Gewichte der IQR des Boxplots se.wls der kleinste.

KAPITEL 3

GENERALISIERTE LINEARE MODELLE

Bei Generalisierten Linearen Modellen (GLMs) handelt es sich um eine Verallgemeinerung der linearen Modelle aus Kapitel 2. Die gewöhnliche lineare Regression basiert auf den Annahmen von stochastisch unabhängigen normalverteilten Responses mit Erwartung $\mu_i = \boldsymbol{x}_i'\boldsymbol{\beta}$ und konstanter Varianz σ^2. Werden Anzahlen oder binäre Daten (z.B. krank/gesund) als Responses betrachtet, so ist die Normalverteilungsannahme fehl am Platz. Auch Abhängigkeiten der Varianzstruktur vom Erwartungswert können nur sehr schlecht mittels linearen Modellen dargestellt werden. Nelder und Wedderburn (1972) führten die Klasse der GLMs ein, um eine Modellierung der oben genannten Situation zu gewährleisten. Ein GLM besteht laut McCullagh und Nelder (1989) aus drei Modellkomponenten.

1. Stochastische Komponente: $y_i \overset{ind}{\sim}$ Exponentialfamilie(θ_i) für $i = 1, \ldots, n$.
   Die Responsevariable stammt aus der einparametrigen linearen Exponentialfamilie mit $\mathbb{E}(y_i) = \mu_i$.
2. Systematische Komponente: $\eta_i = \boldsymbol{x}_i'\boldsymbol{\beta}$.
   Der Vektor der linearen Prädiktoren wird mit $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)$ bezeichnet.
3. Linkfunktion: $g(\mu_i) = \eta_i$.
   Die Verknüpfung zwischen stochastischer und systematischer Komponente erfolgt mittels einer monotonen zweimal stetig differenzierbaren Funktion $g(\cdot)$.

3.1 Lineare Exponentialfamilie

Die stochastische Komponente des GLM stammt aus der linearen Exponentialfamilie. In diesem Abschnitt werden einige Kenngrößen dieser Familie von Verteilungen diskutiert.

Definition 3.1. Die Verteilung einer Variable gehört zur linearen Exponentialfamilie (kurz LEF), wenn sich die Dichtefunktion in folgender Form schreiben lässt:

$$f(y|\theta) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right), \tag{3.1}$$

wobei $a(.) > 0$, $b(.)$ und $c(.)$ spezielle bekannte Funktionen sind. Ist der Parameter ϕ bekannt, so wird $f(y|\theta)$ als einparametrige lineare Exponentialfamilie in kanonischer Form und ϕ als kanonischer Parameter bezeichnet.

Sei die Verteilung der Zufallsvariable y ein Mitglied der linearen Exponentialfamilie. Der Erwartungswert und die Varianz von y können laut McCullagh und Nelder (1989, S. 28 ff.) mittels der bekannten Eigenschaften

$$\mathbb{E}\left(\frac{\partial \log f(y|\theta)}{\partial\theta}\right) = 0 \text{ und} \tag{3.2}$$

$$\mathrm{Var}\left(\frac{\partial \log f(y|\theta)}{\partial\theta}\right) = \mathbb{E}\left(\left(\frac{\partial \log f(y|\theta)}{\partial\theta}\right)^2\right) = \mathbb{E}\left(-\frac{\partial^2 \log f(y|\theta)}{\partial\theta^2}\right) \tag{3.3}$$

ermittelt werden. Wendet man (3.2) und (3.3) auf die Dichtefunktion (3.1) an, so ergeben sich

$$\mathbb{E}\left(\frac{\partial \log f(y|\theta)}{\partial\theta}\right) = \frac{1}{a(\phi)}\mathbb{E}(y - b'(\theta)) = \frac{1}{a(\phi)}\left(\mathbb{E}(y) - b'(\theta)\right) = 0$$

und

$$\mathbb{E}\left(\left(\frac{\partial \log f(y|\theta)}{\partial\theta}\right)^2\right) + \mathbb{E}\left(\frac{\partial^2 \log f(y|\theta)}{\partial\theta^2}\right) = \frac{1}{a^2(\phi)}\mathrm{Var}(y) - \frac{1}{a(\phi)}b''(\theta) = 0.$$

Somit folgt, dass die ersten beiden Momente der Zufallsvariable y

- $\mathbb{E}(y) = b'(\theta) = \mu$ und
- $\mathrm{Var}(y) = a(\phi) \cdot b''(\theta)$

lauten. Da die Varianz von y vom Erwartungswert $\mu = b'(\theta)$ abhängig ist, kann diese alternativ als $\mathrm{Var}(y) = a(\phi)V(\mu)$ dargestellt werden, wobei $V(\mu)$ die so genannte *Varianzfunktion* und ϕ der *Dispersionsparameter* sind.

Betrachte eine n-elementige Stichprobe $(y_1, \ldots, y_n)$ aus der einparametrigen LEF. Um den Dispersionsparameter schätzen zu können, muss die Annahme $a_i(\phi) = a_i \cdot \phi$

getroffen werden, wobei a_i bekannte Gewichte sind. Dies hat zur Folge, dass die Varianzen der y_i als

$$\mathrm{Var}(y_i) = a_i \cdot \phi \cdot b''(\theta) \tag{3.4}$$

geschrieben werden können und somit nur von einem einzigen Dispersionsparameter abhängig sind.

In der nachfolgenden Tabelle 3.1 sind die wichtigsten Mitglieder der LEF aufgelistet (Wood, 2006, S. 61).

Verteilung	θ	ϕ	$c(y,\phi)$	$b(\theta)$	$V(\mu)$
$y \sim \text{Normal}(\mu,\sigma^2)$	μ	σ^2	$-\frac{1}{2}\left[\frac{y^2}{\phi} + \log(2\pi\phi)\right]$	$\frac{\theta^2}{2}$	1
$y \sim \text{Poisson}(\mu)$	$\log(\mu)$	1	$-\log y!$	$\exp(\theta)$	μ
$y \sim \text{Gamma}(\mu,\nu)$	$-\frac{1}{\mu}$	$\frac{1}{\nu}$	$\nu\log(\nu y) - \log(y\Gamma(\nu))$	$-\log(-\theta)$	μ^2
$my \sim \text{Binomial}(m,\pi)$	$\log\left(\frac{\pi}{1-\pi}\right)$	1	$\log\binom{m}{my}$	$\log(1+\exp(\theta))$	$\mu(1-\mu)$
$y \sim \text{InvGauss}(\mu,\sigma^2)$	$-\frac{1}{2\mu^2}$	σ^2	$-\frac{1}{2}\left[\frac{1}{\phi y} + \log(2\pi\phi y^3)\right]$	$-(-2\theta)^{1/2}$	μ^3

Tabelle 3.1: Mitglieder der LEF. Die Parametrisierung der Gammaverteilung erfolgt so, dass die Varianz als μ^2/ν gegeben ist.

3.1.1 Unterschiedliche Dispersionsparameter

Für den Fall, dass n unabhängige Responses $y_1, \ldots, y_n$ aus derselben Exponentialfamilie unterschiedliche Dispersionsparameter ϕ_i aufweisen, ist es möglich, diesem Problem mit Gewichten w_i (englisch prior weights) entgegenzuwirken. Die Dichtefunktion ist laut (3.1) gegeben als

$$f(y_i|\theta_i,\phi_i) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{a(\phi_i)} + c(y_i,\phi_i)\right).$$

Durch die Gewichte w_i ist es möglich, die unterschiedlichen Dispersionsparameter ϕ_i vergleichbar zu machen, d.h. wir nehmen an, dass

$$\phi_i = \frac{\phi}{w_i},$$

wobei ϕ der eigentliche Dispersionsparameter ist. Laut Venables und Ripley (2012, S. 183) entsprechen die Koeffizienten a_i den reziproken Gewichten w_i, d.h.

$$a_i = \frac{1}{w_i}.$$

Somit folgt für die Dichtefunktion (3.1)

$$f(y_i|\theta_i, \phi, w_i) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{a_i\phi} + c(y_i, a_i\phi)\right).$$

Die Gewichte können mit bekannten Schätzmethoden aus dem Abschnitt 2.2.5 ermittelt werden. In `R` werden die Koeffizienten a_i, laut Venables und Ripley (2012, S. 188), mittels des Kommandos `weights=` zugewiesen.

3.2 Linkfunktion

Die Linkfunktion ist eine monotone, zweimal stetig differenzierbare Funktion, die den linearen Prädiktor η_i mit dem Erwartungswert μ_i einer Response y_i verknüpft. Im klassischen linearen Modell aus Kapitel 2 entspricht $\mu_i = \eta_i = \boldsymbol{x}_i'\boldsymbol{\beta}$. Dies kann natürlich auch für das GLM der Fall sein. Durch die Wahl der so genannten *kanonischen Linkfunktion*

$$g(\cdot) = (b')^{-1}(\cdot) \tag{3.5}$$

wird

$$\boldsymbol{x}_i'\boldsymbol{\beta} = \eta_i = g(\mu_i) = g(b'(\theta_i)) \overset{(3.5)}{=} (b')^{-1}(b'(\theta_i)) = \theta_i$$

gewährleistet. Die kanonische Linkfunktion ist vom Mitglied der LEF abhängig. Für normalverteilte Responses entspricht die Linkfunktion gleich der Identitätsfunktion, d.h. $g(\cdot) = id(\cdot)$, wogegen $g(\cdot) = \log(\cdot)$ für poissonverteilte Zufallsvariablen gilt. Ein Vorteil der kanonischen Linkfunktion ist, dass sich die Scorefunktion in der Maximum-Likelihood-Schätzung vereinfacht. Es können jedoch auch andere Funktionen als Linkfunktion gewählt werden. Für die Binomialverteilung gilt $0 < \mu < 1$ und die Linkfunktion sollte daher das Intervall $(0, 1)$ auf die reelle Zahlengerade abbilden. Als Linkfunktionen wird für gewöhnlich eine der nachfolgenden Funktionen verwendet (McCullagh und Nelder, 1989, S. 31).

1. Die kanonische Linkfunktion ist der so genannte logit-link:

$$\eta = \log\left(\frac{\mu}{1-\mu}\right).$$

Weitere Linkfunktionen im Falle der Binomalverteilung lauten:

2. probit-link:

$$\eta = \Phi^{-1}(\mu),$$

   wobei $\Phi(\cdot)$ die Standardnormalverteilung ist.

3. complementary log-log-link:

$$\eta = \log(-\log(1-\mu)).$$

3.3 Maximum-Likelihood-Schätzung

Gegeben seien n unabhängige Responses $y_1, \ldots, y_n$ aus der selben Exponentialfamilie mit Parametern (θ_i, ϕ). Die Log-Likelihood-Funktion der Stichprobe ist durch

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

gegeben. Die Herleitung des Maximum-Likelihood-Schätzers $\hat{\boldsymbol{\beta}}$ orientiert sich an Dobson (2002, Kap. 4.3). Um diesen zu ermitteln, wird die Log-Likelihood-Funktion nach β_j abgeleitet. Es folgt

$$\frac{\partial l(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \beta_j} = U_j = \sum_{i=1}^{n} \underbrace{\frac{\partial l(\theta_i|y_i)}{\partial \theta_i}}_{1} \cdot \underbrace{\frac{\partial \theta_i}{\partial \mu_i}}_{2} \cdot \underbrace{\frac{\partial \mu_i}{\partial \beta_j}}_{3}. \tag{3.6}$$

Wir betrachten die drei Faktoren separat.

1. Das Bilden der partiellen Ableitung nach θ_i liefert

$$\frac{\partial l(\theta_i|y_i)}{\partial \theta_i} = \frac{y_i - b'(\theta_i)}{a_i\phi} = \frac{y_i - \mu_i}{a_i\phi}.$$

2. Mit Hilfe der Inversenregel der Differentiation folgt

$$\frac{\partial \mu_i}{\partial \theta_i} = \frac{\partial b'(\theta_i)}{\partial \theta_i} = b''(\theta_i) \Longrightarrow \frac{\partial \theta_i}{\partial \mu_i} = \frac{1}{b''(\theta_i)} = \frac{1}{V(\mu_i)}.$$

3. Mit der systematischen Komponente $\eta_i = \boldsymbol{x}_i'\boldsymbol{\beta}$ folgt

$$\frac{\partial \mu_i}{\partial \beta_j} = \frac{\partial \mu_i}{\partial \eta_i} \cdot \frac{\partial \eta_i}{\partial \beta_j} = \frac{\partial \mu_i}{\partial \eta_i} \cdot x_{ij}.$$

Setzt man die drei Terme in (3.6) ein, so folgt

$$U_j = \sum_{i=1}^{n} \left[\frac{y_i - \mu_i}{a_i\phi V(\mu_i)} x_{ij} \left(\frac{\partial \mu_i}{\partial \eta_i} \right) \right] = \sum_{i=1}^{n} \left[\frac{y_i - \mu_i}{\mathrm{Var}(y_i)} x_{ij} \left(\frac{\partial \mu_i}{\partial \eta_i} \right) \right]. \tag{3.7}$$

Die **Informationsmatrix** $\boldsymbol{\mathcal{J}}$ wird durch die Elemente der Varianz-Kovarianz-Matrix der U_j gebildet, das heißt

$$\mathcal{J}_{jk} = \mathbb{E}(U_j U_k) = \mathbb{E}\left\{ \sum_{i=1}^{n} \left[\frac{y_i - \mu_i}{\mathrm{Var}(y_i)} x_{ij} \left(\frac{\partial \mu_i}{\partial \eta_i} \right) \right] \sum_{l=1}^{n} \left[\frac{y_l - \mu_l}{\mathrm{Var}(y_l)} x_{lk} \left(\frac{\partial \mu_l}{\partial \eta_l} \right) \right] \right\}.$$

Da die y_i paarweise unabhängig sind, gilt $\mathbb{E}\left((y_i-\mu_i)(y_l-\mu_l)\right)=0$ für $i\neq l$. Im Falle $i=l$ gilt $\mathbb{E}\left((y_i-\mu_i)^2\right)=\mathrm{Var}(y_i)$ und $\mathcal{J}_{jk}$ lässt sich wie folgt vereinfachen:

$$\mathcal{J}_{jk}=\sum_{i=1}^{n}\left[\frac{\mathbb{E}\left((y_i-\mu_i)^2\right)}{\mathrm{Var}(y_i)^2}x_{ij}x_{ik}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2\right]=\sum_{i=1}^{n}\left[\frac{x_{ij}x_{ik}}{\mathrm{Var}(y_i)}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2\right]. \tag{3.8}$$

Da das Gleichungssystem (3.7) nicht linear in $\boldsymbol{\beta}$ ist, kann dieses nur mittels numerischer Verfahren gelöst werden. Hierfür eignet sich die Newton-Raphson-Methode mit der Iterationsvorschrift

$$\boldsymbol{b}^{(m)}=\boldsymbol{b}^{(m-1)}+(\boldsymbol{\mathcal{J}}^{(m-1)})^{-1}\cdot\boldsymbol{U}^{(m-1)}.$$

Der Vektor $\boldsymbol{b}^{(m)}$ der obigen Gleichung beinhaltet die Schätzer der Parameter $\boldsymbol{\beta}=(\beta_0,\ldots,\beta_{p-1})$ in der m-ten Iteration. Die Elemente U_j (3.6) werden im Vektor $\boldsymbol{U}^{(m-1)}$ zusammengefasst. Durch Multiplikation der Matrix $\boldsymbol{\mathcal{J}}^{(m-1)}$ von links erhält man

$$\boldsymbol{\mathcal{J}}^{(m-1)}\boldsymbol{b}^{(m)}=\boldsymbol{\mathcal{J}}^{(m-1)}\boldsymbol{b}^{(m-1)}+\boldsymbol{U}^{(m-1)}. \tag{3.9}$$

Die Einträge der Matrix $\boldsymbol{\mathcal{J}}^{(m-1)}$ sind durch (3.8) gegeben. Somit lässt sich $\boldsymbol{\mathcal{J}}^{(m-1)}$ als Matrixprodukt

$$\boldsymbol{\mathcal{J}}^{(m-1)}=\boldsymbol{X'}\boldsymbol{W}^{(m-1)}\boldsymbol{X}, \tag{3.10}$$

schreiben, wobei $\boldsymbol{W}^{(m-1)}$ eine $n\times n$ Diagonalmatrix mit Einträgen

$$w_{ii}^{(m-1)}=\frac{1}{\mathrm{Var}^{(m-1)}(y_i)}\left(\frac{\partial\mu_i^{(m-1)}}{\partial\eta_i^{(m-1)}}\right)^2=\frac{1}{a_i\phi V\left(\mu_i^{(m-1)}\right)}\left(\frac{\partial\mu_i^{(m-1)}}{\partial\eta_i^{(m-1)}}\right)^2$$

ist. Die rechte Seite der Newton-Raphson-Iterationsvorschrift (3.9) ist ein Vektor bestehend aus den Einträgen

$$\sum_{k=0}^{p-1}\sum_{i=1}^{n}\frac{x_{ij}x_{ik}}{\mathrm{Var}^{(m-1)}(y_i)}\left(\frac{\partial\mu_i^{(m-1)}}{\partial\eta_i^{(m-1)}}\right)^2 b_k^{(m-1)}+\sum_{i=1}^{n}\frac{\left(y_i-\mu_i^{(m-1)}\right)x_{ij}}{\mathrm{Var}^{(m-1)}(y_i)}\left(\frac{\partial\mu_i^{(m-1)}}{\partial\eta_i^{(m-1)}}\right).$$

In Matrixschreibweise kann die rechte Seite von (3.9) als

$$\boldsymbol{X'}\boldsymbol{W}^{(m-1)}\boldsymbol{z}^{(m-1)} \tag{3.11}$$

geschrieben werden, wobei $\boldsymbol{z}^{(m-1)}$ ein n-dimensionaler Vektor bestehend aus den Elementen

$$z_i=\sum_{k=0}^{p-1}x_{ik}b_k^{(m-1)}+\left(y_i-\mu_i^{(m-1)}\right)\left(\frac{\partial\eta_i^{(m-1)}}{\partial\mu_i^{(m-1)}}\right)$$

ist. Aus (3.10) und (3.11) folgt, dass die iterative Gleichung (3.9) als

$$\boldsymbol{X}\boldsymbol{W}^{(m-1)}\boldsymbol{X}\boldsymbol{b}^{(m)} = \boldsymbol{X}'\boldsymbol{W}^{(m-1)}\boldsymbol{z}^{(m-1)} \tag{3.12}$$

geschrieben werden kann. Die Gleichungen (3.12) der Newton-Raphson-Vorschrift haben dieselbe Gestalt wie die Normalengleichungen des gewichteten linearen Modells (2.21). In diesem Fall müssen diese Normalengleichungen jedoch durch ein iteratives Verfahren gelöst werden, da $\boldsymbol{z}$ und $\boldsymbol{W}$ von $\boldsymbol{b}$ abhängig sind. Der MLE $\hat{\boldsymbol{\beta}}$ wird somit durch eine iterative gewichtete kleinste Quadrate Prozedur (engl. **Iterative Weighted Least Squares** procedure) ermittelt. Die Iterationsvorschrift wird so lange durchgeführt, bis der Unterschied der Vektoren $\boldsymbol{b}^{(k-1)}$ und $\boldsymbol{b}^{(k)}$ für zweie aufeinanderfolgenden Iterationen $k-1$ und k vernachlässigbar klein ist. Als MLE $\hat{\boldsymbol{\beta}}$ wird dann der Vektor $\boldsymbol{b}^{(k)}$ der finalen Iteration k verwendet.

3.3.1 Asymptotische Eigenschaften des MLE

Um die asymptotischen Momente des Maximum-Likelihood-Schätzers $\hat{\boldsymbol{\beta}}$ zu ermitteln, wird eine Taylor Entwicklung der Scorefunktion $\frac{\partial l(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \boldsymbol{\beta}}|_{\boldsymbol{b}}$ des Parametervektors $\boldsymbol{b}$ an der Entwicklungsstelle $\boldsymbol{\beta}$ durchgeführt. Hierbei orientieren wir uns an Dobson (2002, Kap 5.4). Die ersten beiden Glieder der Taylorreihe sind durch

$$\left.\frac{\partial l(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{b}} = \left.\frac{\partial l(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\beta}} + (\boldsymbol{b}-\boldsymbol{\beta})\left.\frac{\partial^2 l(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'}\right|_{\boldsymbol{\beta}}$$

gegeben. Um die Schreibweise zu vereinfachen, bezeichne mit $\boldsymbol{U}(\boldsymbol{\beta}) = \frac{\partial l(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \boldsymbol{\beta}}|_{\boldsymbol{\beta}}$ die Scorefunktion und $\boldsymbol{H}(\boldsymbol{\beta}) = \frac{\partial^2 l(\boldsymbol{\theta}|\boldsymbol{y})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'}|_{\boldsymbol{\beta}}$ die Hessematrix in $\boldsymbol{\beta}$. Wird die Hessematrix $\boldsymbol{H}(\boldsymbol{\beta})$ durch deren Erwartungswert $\mathbb{E}(\boldsymbol{H}(\boldsymbol{\beta})) = -\boldsymbol{\mathcal{J}}(\boldsymbol{\beta})$ geschätzt, erhaltet man das Gleichungssystem

$$\boldsymbol{U}(\boldsymbol{b}) = \boldsymbol{U}(\boldsymbol{\beta}) - \boldsymbol{\mathcal{J}}(\boldsymbol{\beta})(\boldsymbol{b}-\boldsymbol{\beta}).$$

Hierbei entspricht $\boldsymbol{\mathcal{J}}(\boldsymbol{\beta})$ der Informationsmatrix, die aus den Einträgen (3.8) besteht. Da der MLE $\hat{\boldsymbol{\beta}}$ die Log-Likelihood-Funktion $l(\boldsymbol{\beta}|\boldsymbol{y})$ maximiert, folgt an der Stelle $\boldsymbol{b} = \hat{\boldsymbol{\beta}}$ für die Scorefunktion $\boldsymbol{U}(\hat{\boldsymbol{\beta}}) = \boldsymbol{0}$. Daraus resultiert die Gleichung

$$\boldsymbol{U}(\boldsymbol{\beta}) = \boldsymbol{\mathcal{J}}(\boldsymbol{\beta})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})$$

oder äquivalent

$$(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}) = \boldsymbol{\mathcal{J}}(\boldsymbol{\beta})^{-1}\boldsymbol{U}(\boldsymbol{\beta}),$$

falls $\boldsymbol{\mathcal{J}}(\boldsymbol{\beta})$ eine nicht singuläre Matrix ist. Wird die Matrix $\boldsymbol{\mathcal{J}}(\boldsymbol{\beta})$ als Konstante betrachtet, folgt für den Erwartungswert von $(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})$:

$$\mathbb{E}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}) = \mathbb{E}(\boldsymbol{\mathcal{J}}(\boldsymbol{\beta})^{-1}\boldsymbol{U}(\boldsymbol{\beta})) = \boldsymbol{\mathcal{J}}(\boldsymbol{\beta})^{-1}\mathbb{E}(\boldsymbol{U}(\boldsymbol{\beta})) = \boldsymbol{0}.$$

Weil der Erwartungswert der Scorefunktion gleich Null ist, folgt dass der MLE $\hat{\boldsymbol{\beta}}$ zumindest asymptotisch ein erwartungstreuer Schätzer für den Parameter $\boldsymbol{\beta}$ ist. Die Varianz-Kovarianz-Matrix des MLE $\hat{\boldsymbol{\beta}}$ ist durch

$$\text{Var}(\hat{\boldsymbol{\beta}}) = \mathbb{E}\left((\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\right) = \boldsymbol{\mathcal{J}}(\boldsymbol{\beta})^{-1}\mathbb{E}\left(\boldsymbol{U}(\boldsymbol{\beta})\boldsymbol{U}(\boldsymbol{\beta})'\right)\left(\boldsymbol{\mathcal{J}}(\boldsymbol{\beta})^{-1}\right)'$$

gegeben. Die Varianz-Kovarianz-Matrix der Scorefunktion entspricht der Informationsmatrix $\boldsymbol{\mathcal{J}}(\boldsymbol{\beta}) = \mathbb{E}\left(\boldsymbol{U}(\boldsymbol{\beta})\boldsymbol{U}(\boldsymbol{\beta})'\right)$. Da die Informationsmatrix $\boldsymbol{\mathcal{J}}(\boldsymbol{\beta})$ symmetrisch ist, gilt $\boldsymbol{\mathcal{J}}(\boldsymbol{\beta})^{-1} = \left(\boldsymbol{\mathcal{J}}(\boldsymbol{\beta})^{-1}\right)'$. Somit lautet die Varianz-Kovarianz-Matrix des MLE $\hat{\boldsymbol{\beta}}$ schlussendlich

$$\text{Var}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\mathcal{J}}(\boldsymbol{\beta})^{-1}.$$

Fahrmeir und Kaufmann (1985) zeigten sogar

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{D} Normal(\mathbf{0}, n\boldsymbol{\mathcal{J}}(\boldsymbol{\beta})^{-1})). \tag{3.13}$$

Für Mitglieder der LEF mit bekanntem Dispersionsparameter ϕ kann das Resultat (3.13) verwendet werden, um Konfidenzintervalle für die Parameter β_j zu konstruieren. Ist ϕ jedoch unbekannt (z.B. Normalverteilung), muss der Dispersionsparameter geschätzt werden. Aufgrund dessen basieren die Konfidenzintervalle der β_j auf den Quantilen der Student-t-Verteilung

3.4 Güte der Modellanpassung

Nachdem die Daten gefitted worden sind, sollte überprüft werden, wie gut das GLM die Datensituation beschreibt. Die Deviance bietet eine Möglichkeit, die Güte der Modellanpassung zu bewerten (McCullagh und Nelder, 1989, S. 33 ff.). Hier betrachtet man die Likelihood-Quotienten-Teststatistik zu den Hypothesen

$$H_0: \ \boldsymbol{\mu} = g^{-1}(\boldsymbol{x}'\boldsymbol{\beta}) \text{ gegen } H_1: \ \boldsymbol{\mu} \neq g^{-1}(\boldsymbol{x}'\boldsymbol{\beta}).$$

Unter der Nullhypothese wird behauptet, dass das betrachtete Modell das korrekte ist. Die Likelihood-Quotienten-Teststatistik wird gebildet aus dem Quotienten

$$\Lambda(\boldsymbol{y}) = \frac{\sup\limits_{\boldsymbol{\mu}=g^{-1}(\boldsymbol{\eta})} L(\boldsymbol{\mu}|\boldsymbol{y})}{\sup\limits_{\boldsymbol{\mu}} L(\boldsymbol{\mu}|\boldsymbol{y})},$$

wobei im Zähler die eingeschränkte Likelihood-Funktion unter H_0 und im Nenner die uneingeschränkte Likelihood-Funktion maximiert werden. Wie wir bereits wissen, maximiert der MLE $\hat{\boldsymbol{\beta}}$ die Likelihood-Funktion unter H_0 (siehe Abschnitt

3.3). Die uneingeschränkte Likelihood-Funktion nimmt ihr Maximum in $\hat{\mu}_i = y_i$ für $i = 1, \ldots, n$ an. Es folgt somit der Quotient

$$\Lambda(\boldsymbol{y}) = \frac{L(\hat{\boldsymbol{\mu}}|\boldsymbol{y})}{L(\boldsymbol{y}|\boldsymbol{y})}.$$

Unter bestimmten Voraussetzungen lässt sich $-2\log(\Lambda(\boldsymbol{y}))$ durch eine χ^2-verteilte Zufallsvariable annähern. Durch Bilden dieses Termes erhält man die Statistik

$$D^*(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = \frac{1}{\phi} D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = -2(l(\hat{\boldsymbol{\mu}}|\boldsymbol{y}) - l(\boldsymbol{y}|\boldsymbol{y})). \tag{3.14}$$

Diese wird *skalierte Deviance* genannt und große Werte von (3.14) sprechen gegen H_0.

Wie in der nachfolgenden Herleitung ersichtlich, entspricht die skalierte Deviance für normalverteilte Responses einer Verallgemeinerung der Fehlerquadratsumme des linearen Modells. Betrachte hierfür zuerst die Log-Likelihood-Funktion der Stichprobe $y_i \overset{ind}{\sim} Normal(\mu_i, \sigma^2)$ für $i = 1, \ldots, n$. Laut (2.3) ist diese für bekanntes σ^2 gegeben als

$$l(\boldsymbol{\mu}|\boldsymbol{y}) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu_i)^2.$$

Daher lautet die skalierte Deviance (3.14) für normalverteilte Responses

$$\begin{aligned}
\frac{1}{\phi} D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) &= -2(l(\hat{\boldsymbol{\mu}}, \boldsymbol{y}) - l(\boldsymbol{y}, \boldsymbol{y})) \\
&= -2\left[-\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2 + \frac{n}{2}\log(2\pi\sigma^2)\right] \\
&= \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2 = \frac{1}{\sigma^2}\mathrm{SSE}(\hat{\boldsymbol{\beta}}).
\end{aligned}$$

Es ist bekannt, dass $\frac{1}{\sigma^2}\mathrm{SSE}(\hat{\boldsymbol{\beta}})$ einer χ^2_{n-p}-Verteilung folgt (Fahrmeir et al., 2009, S. 107) und somit ist die skalierte Deviance, im Falle normalverteilter Responses, eine exakte χ^2_{n-p} verteilte Zufallsvariable. Für weitere Mitglieder der Exponentialfamilie mit bekanntem Dispersionsparameter ist $D^*(\boldsymbol{y}, \hat{\boldsymbol{\mu}})$ approximativ χ^2_{n-p} verteilt mit Erwartungswert $n - p$ (Wood, 2006, S. 69).

Ist der Dispersionsparameter ϕ unbekannt, besteht die Möglichkeit, diesen mittels der mittleren Deviance zu schätzen. Da die skalierte Deviance durch eine χ^2_{n-p} Verteilung approximiert werden kann, folgt

$$\frac{1}{\phi}\mathbb{E}(D(\boldsymbol{y}, \hat{\boldsymbol{\mu}})) \approx n - p \quad \Longleftrightarrow \quad \phi \approx \frac{\mathbb{E}(D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}))}{n - p}.$$

Somit lässt sich der unbekannte Dispersionsparameter durch die mittlere Deviance

$$\hat{\phi} = \frac{D(\boldsymbol{y}, \hat{\boldsymbol{\mu}})}{n-p} \tag{3.15}$$

erwartungstreu schätzen.

Ist ϕ nicht bekannt, kann die Güte der Modellanpassung mittels der *generalisierten Pearson Statistik* bewertet werden.

Definition 3.2. Für n unabhängige Beobachtungen $y_1, \ldots, y_n$ mit Erwartungswerten μ_i und Varianzen $\text{Var}(y_i) = a_i \phi V(\mu_i)$ ist die *generalisierte Pearson Statistik* definiert als

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{a_i V(\hat{\mu}_i)}.$$

Die biaskorrigierte Version von X^2, die so genannte *mittlere Pearson Statistik*

$$\hat{\phi} = \frac{1}{n-p} X^2, \tag{3.16}$$

wird zur Schätzung des Dispersionsparameters ϕ verwendet. Des Weiteren ist diese Statistik approximativ χ^2_{n-p} verteilt und sollte unter dem korrekten Modell im Mittel etwa der Anzahl der Freiheitsgrade $n - p$ entsprechen.

3.5 Residuen

Mittels einer Residuenanalyse kann überprüft werden, ob die Modellannahmen korrekt erfüllt worden sind. Diesbezüglich sollten laut Wood (2006, S. 72 ff.) die Pearson bzw. die Deviance Residuen untersucht werden.

Die Pearson Residuen sind definiert als

$$r_i^P = \frac{y_i - \hat{\mu}_i}{\sqrt{a_i V(\hat{\mu}_i)}} \quad \text{für } i = 1, \ldots, n$$

und entsprechen den rohen Residuen skaliert um einen Faktor, der proportional zur geschätzten Standardabweichung der y_i ist. Die Pearson Residuen erhalten ihren Namen aufgrund ihrer Verbindung zur generalisierten Pearson Statistik X^2 (3.16). Sie entsprechen den nicht quadrierten Summanden von X^2, d.h.

$$X^2 = \sum_{i=1}^{n} \left(r_i^P\right)^2.$$

Wurde das Modell korrekt gewählt, dann sollten die Pearson Residuen circa Erwartung Null und Varianz ϕ besitzen (McCullagh und Nelder, 1989, S. 37). Falls die Pearson Residuen gegen die Fitted-Values oder einen Prädiktor geplottet werden, so sollte kein Trend im Erwartungswert und in der Varianz ersichtlich sein. In der Praxis jedoch kann die Verteilung der Pearson Residuen sehr asymmetrisch um die Null sein. Der Verhaltensunterschied zwischen Pearson Residuen und den rohen Residuen (2.25) des gewöhnlichen linearen Modells ist deswegen größer als erhofft. Aufgrund dessen werden die so genannten Deviance Residuen den Pearson Residuen in diesem Aspekt vorgezogen. Die Deviance Residuen sind definiert als

$$r_i^D = \text{sign}(y_i - \hat{\mu}_i)\sqrt{d_i},$$

wobei d_i die i-te Komponente der Deviance $D(\boldsymbol{y}, \hat{\boldsymbol{\mu}})$ ist, das heißt

$$D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = \sum_{i=1}^{n} d_i.$$

Wird die Deviance für ein Modell berechnet, in dem alle Parameter bekannt sind, dann ist die skalierte Deviance (3.14) χ^2_n verteilt. Dies würde bedeuten, dass ein einzelner Summand $d_i \sim \chi^2_1$ verteilt ist und somit $r_i^D \sim N(0,1)$. Somit kann man von einem Modell mit guter Anpassungsgüte davon ausgehen, dass sich die Deviance Residuen wie standardnormalverteilte Zufallsvariablen verhalten (Wood, 2006, S. 73).

3.6 Parametertests

Wie bereits bei linearen Modellen, kann auch hier das Konzept der verschachtelten Modelle (engl. nested models) verwendet werden, um die Signifikanz einzelner Parameter zu überprüfen. Gegeben seien zwei Modelle M_0 und M_1 wobei M_0 ein Untermodell von M_1 ist. Ob die Parameter des Obermodells für die Regression benötigt werden, kann mit den nachfolgenden Hypothesen überprüft werden. Diese lauten wie folgt

$$\begin{aligned} H_0 &: \ \eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{q-1} x_{iq-1} \text{ gegen} \\ H_1 &: \ \eta_i = \beta_0 + \cdots + \beta_{q-1} x_{iq-1} + \beta_q x_{iq} + \cdots + \beta_{p-1} x_{ip-1}, \end{aligned}$$

was äquivalent ist zu

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

Die Deviance der beiden Modelle ist durch

$$\begin{aligned} D(M_0) = D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}_0) &= -2\phi(l(\hat{\boldsymbol{\mu}}_0, \boldsymbol{y}) - l(\boldsymbol{y}, \boldsymbol{y})) \\ D(M_1) = D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}_1) &= -2\phi(l(\hat{\boldsymbol{\mu}}_1, \boldsymbol{y}) - l(\boldsymbol{y}, \boldsymbol{y})) \end{aligned}$$

gegeben, wobei die Deviancedifferenz

$$D(M_0) - D(M_1) = D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}_0) - D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}_1) = -2\phi \left(l(\hat{\boldsymbol{\mu}}_0, \boldsymbol{y}) - l(\hat{\boldsymbol{\mu}}_1, \boldsymbol{y}) \right)$$

den Unterschied der Anpassungsgüte der beiden Modelle beschreibt. Falls der Dispersionsparameter ϕ bekannt ist, dann entspricht die skalierte Deviancedifferenz $(D(M_0) - D(M_1))/\phi$ unter H_0 der Likelihood-Quotienten-Teststatistik. Da die skalierte Deviance asymptotisch χ^2-verteilt ist, folgt aus $D(M_0)/\phi \sim \chi^2_{n-q}$ und $D(M_1)/\phi \sim \chi^2_{n-p}$ dass $(D(M_0) - D(M_1))/\phi \sim \chi^2_{p-q}$ ist. Daher sollte bei bekanntem Dispersionsparameter laut Wood (2006, S. 70), die Teststatistik

$$\frac{D(M_0) - D(M_1)}{\phi} = \frac{D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}_0) - D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}_1)}{\phi} \overset{H_0}{\sim} \chi^2_{p-q}$$

verwendet werden. Im Falle eines unbekannten Dispersionparameters ϕ, muss ϕ durch die mittlere Deviance (3.15) geschätzt werden und es wird folgende Teststatistik herangezogen

$$\frac{\left(D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}_0) - D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}_1)\right)/(p-q)}{D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}_1)/(n-p)} \overset{H_0}{\sim} F_{p-q,n-p}.$$

Ein weiterer Parametertest für statistische Modelle ist durch den so genannten **Wald-Test** gegeben. Dessen Teststatistik basiert auf den asymptotischen Verteilungseigenschaften des MLE $\hat{\boldsymbol{\beta}}$ (siehe Abschnitt 3.3.1) und entspricht laut Dobson (2002)

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})' \boldsymbol{\mathcal{J}}(\hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim \chi^2_p.$$

Um die individuellen Koeffizienten des MLE $\hat{\boldsymbol{\beta}}$ zu testen, kann die Nullhypothese $H_0 : \beta_j = 0$ für $j = 0, \ldots, p-1$ mit der Teststatistik

$$\left(\frac{\hat{\beta}_j}{\hat{se}(\hat{\beta}_j)} \right)^2 \overset{H_0}{\sim} \chi^2_1$$

überprüft werden (Myers et al., 2010, S. 217 ff.).

3.6.1 AIC für GLMs

Wie bereits im Abschnitt 2.2.6 erwähnt, ist das Akaike Informationskriterium (AIC) ein beliebtes Werkzeug für die Modellauswahl. Möchte man sich zwischen mehreren konkurrierenden Modellen entscheiden, wählt man jenes mit dem kleinsten AIC-Wert. Das Kriterium ist für GLMs mit bekanntem Dispersionsparameter ϕ als

$$\text{AIC} = 2[-l(\hat{\boldsymbol{\mu}}|\boldsymbol{y}) + p]$$

gegeben, wobei p die Anzahl der Parameter des GLMs ist. Muss ϕ geschätzt werden, so erhöht sich der Strafterm von p auf $p+1$ (Wood, 2006, S. 68). In `R` kann das Informationskriterium mit dem Kommando `AIC(mod)` für ein GLM `mod` bestimmt werden.

3.7 Gammaregression

Die Gammaverteilung findet Anwendung in der Finanz und Versicherungsmathematik bei der Modellierung kleiner und mittlerer Schäden, sowie in der Warteschlangentheorie um Reparaturzeiten oder Bedienzeiten zu charakterisieren. Weitere Anwendungsbereiche sind Regressionsprobleme mit positiven Responses in der die Varianz quadratisch vom Erwartungswert abhängig ist (Myers et al., 2010, S. 213 ff.).

3.7.1 Kennzahlen der Gammaverteilung

Die Gammaverteilung ist ein bekanntes Mitglied der LEF. Sei $y \sim Gamma(\alpha, \beta)$ mit Gestaltsparameter α und Rate-Parameter β, dann ist ihre Dichte als

$$f(y|\alpha,\beta) = \exp(-\beta y)\beta^{\alpha} y^{\alpha-1}\frac{1}{\Gamma(\alpha)}, \quad \alpha,\beta > 0, y \geq 0$$

definiert. Alternativ existiert jedoch eine reparametrisierte Darstellung der obigen Funktion. Durch die Wahl $\mu = \nu/\beta$ mit $\nu = \alpha$ gilt nun $y \sim Gamma(\nu, \mu)$. Die dazugehörige Dichtefunktion lautet

$$\begin{aligned} f(y|\mu,\nu) &= \exp\left(-\frac{\nu}{\mu}y\right)\left(\frac{\nu}{\mu}\right)^{\nu} y^{\nu-1}\frac{1}{\Gamma(\nu)} \\ &= \exp\left(-\frac{\nu}{\mu}y + \nu\log\nu - \nu\log\mu + (\nu-1)\log y - \log\Gamma(\nu)\right) \\ &= \exp\left(\frac{y\left(-\frac{1}{\mu}\right) + \log\frac{1}{\mu}}{\frac{1}{\nu}} + \nu\log\nu + (\nu-1)\log y - \log\Gamma(\nu)\right), \quad \mu,\nu,y > 0. \end{aligned} \tag{3.17}$$

Mit $\theta = -\frac{1}{\mu}$ und $\phi = \frac{1}{\nu}$ führt dies zur Exponentialfamilie mit

$$a = 1, \quad b(\theta) = -\log(-\theta), \quad c(y,\phi) = \frac{1}{\phi}\log\left(\frac{1}{\phi}\right) + \left(\frac{1}{\phi} - 1\right)\log y - \log\Gamma\left(\frac{1}{\phi}\right).$$

Dies entspricht genau den Termen die in der Tabelle 3.1 ersichtlich sind. Als Erwartungswert und Varianz erhält man, aufgrund der Mitgliedschaft der LEF (siehe

Abschnitt 3.1)

$$\mathbb{E}(y) = b'(\theta) = -\frac{1}{\theta} = \mu \text{ und}$$
$$\text{Var}(y) = a\phi b''(\theta) = \phi\frac{1}{\theta^2} = \frac{1}{\nu}\mu^2.$$

Unter der Annahme, dass $y_i \overset{ind}{\sim} Gamma(\nu, \mu_i)$ für $i = 1, \ldots, n$ sind, folgt dass y_i konstante Variationskoeffizienten C_i besitzen, d.h.

$$C_i = \frac{\sqrt{\text{Var}(y_i)}}{\mathbb{E}(y_i)} = \frac{\sqrt{\frac{\mu_i^2}{\nu}}}{\mu_i} = \frac{1}{\sqrt{\nu}}.$$

In der nachfolgenden Abbildung 3.1 ist die Dichtefunktion für unterschiedliche Parameterwerte von ν abgebildet.

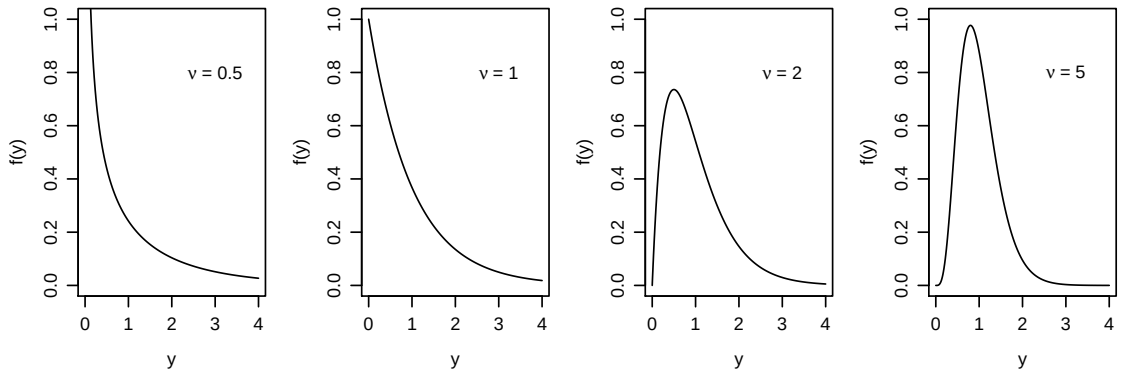

Abbildung 3.1: Die Dichte der Gammaverteilung für $\nu = 0.5$, 1, 2 und 5, $\mu = 1$.

Für $0 < \nu < 1$ besitzt die Dichtefunktion eine Polstelle im Ursprung und fällt monoton für $y \to \infty$. Der Spezialfall $\nu = 1$ entspricht der Exponentialverteilung und der Grenzfall $\nu \to \infty$ entspricht der Normalverteilung. Für Werte $\nu \geq 1$ ist die Dichte rechtsschief mit einem Schiefekoeffizienten von $2\nu^{-1/2}$.

3.7.2 Unterschiedliche Dispersionsparameter

Sind die Dispersionsparameter ϕ_i für n unabhängige gammaverteilte Responses y_i unterschiedlich, so kann man die unterschiedlichen Dispersionsparameter mit Gewichten w_i vergleichbar machen, d.h.

$$\phi_i = \frac{\phi}{w_i} = \frac{\frac{1}{\nu}}{w_i} = \frac{1}{w_i} \cdot \frac{1}{\nu} = \frac{1}{\nu_i}.$$

Sei y_i eine der zuvor genannten unabhängigen gammaverteilten Zufallsvariablen mit individuellem Gestaltsparameter $\nu_i = w_i\nu$, d.h. $y_i \overset{ind}{\sim} Gamma(\nu_i, \mu_i)$, dann ist ihre Dichtefunktion, laut (3.17), gegeben als

$$f(y_i|\nu_i, \mu_i) = \exp\left(\frac{y_i\left(-\frac{1}{\mu_i}\right) + \log\frac{1}{\mu_i}}{\frac{1}{\nu_i}} + \nu_i \log\nu_i + (\nu_i - 1)\log y_i - \log\Gamma(\nu_i)\right).$$

Mit $\phi = \frac{1}{\nu}$ und $\theta_i = \mu_i$ führt das zur Exponentialfamilie mit

$$a_i = \frac{1}{w_i}, \quad b(\theta_i) = -\log(-\theta_i),$$
$$c(y_i, \phi) = \frac{1}{a_i\phi}\log\left(\frac{1}{a_i\phi}\right) + \left(\frac{1}{a_i\phi} - 1\right)\log y_i - \log\Gamma\left(\frac{1}{a_i\phi}\right).$$

Der Erwartungswert und die Varianz von y_i lauten

$$\mathbb{E}(y_i) = b'(\theta_i) = -\frac{1}{\theta_i} = \mu_i \text{ und}$$
$$\mathrm{Var}(y_i) = a_i\phi b''(\theta_i) = \frac{\phi}{w_i}\frac{1}{\theta_i^2} = \frac{1}{w_i}\frac{1}{\nu}\mu_i^2.$$

3.7.3 Linkfunktion der Gammaregression

Wie bereits im Abschnitt 3.2 erwähnt, verknüpft die Linkfunktion $g(.)$ den linearen Prädiktor η mit dem Erwartungswert μ. Der kanonische Link ist $g(\cdot) = (b')^{-1}(\cdot)$ und entspricht im Falle der Gammaregression der negativen reziproken Funktion, d.h.

$$g(\mu) = -\frac{1}{\mu}.$$

Für viele Problemstellungen ist dieser Link jedoch unpassend, da $\mu > 0$ gelten muss. Aus diesem Grund wird häufig der Logarithmus der reziproken Funktion vorgezogen. Die alternative Linkfunktion lautet somit

$$g(\mu) = \log(\mu).$$

3.7.4 Schätzung des Dispersionsparameters

Der Dispersionsparameter lässt sich mit Hilfe der mittleren Pearson Statistik (3.16)

$$\hat{\phi} = \frac{1}{n-p}\sum_{i=1}^{n}\frac{(y_i - \hat{\mu}_i)^2}{a_i V(\hat{\mu}_i)}$$

schätzen. Für gammaverteilte Responses y_i ist die Varianzfunktion durch $V(\hat{\mu}_i) = \hat{\mu}_i^2$ gegeben und das liefert für den geschätzten Dispersionsparameter

$$\hat{\phi} = \frac{1}{n-p} \sum_{i=1}^{n} \frac{1}{a_i} \left(\frac{y_i - \hat{\mu}_i}{\hat{\mu}_i} \right)^2 .$$

3.7.5 Deviance der Gammaregression

Die Log-Likelihood-Funktion der Stichprobe $y_i \overset{ind}{\sim} Gamma(\nu_i, \mu_i)$ mit $i = 1, \ldots, n$ ist für bekanntes ν_i gegeben als

$$\begin{aligned} l(\boldsymbol{\theta}|\boldsymbol{y}) &= \sum_{i=1}^{n} \left(\frac{y_i \theta_i - b(\theta_i)}{a_i \phi} + c(y_i, \phi) \right) \\ &\overset{(3.17)}{=} \sum_{i=1}^{n} \left\{ \nu_i \left[-\frac{y_i}{\mu_i} - \log(\mu_i) \right] + \nu_i \log \nu_i + (\nu_i - 1) \log y_i - \log \Gamma(\nu_i) \right\} . \end{aligned}$$

Daher folgt für die skalierte Deviance (3.14)

$$\begin{aligned} \frac{1}{\phi} D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) &= -2(l(\hat{\boldsymbol{\mu}}, \boldsymbol{y}) - l(\boldsymbol{y}, \boldsymbol{y})) \\ &= -2 \sum_{i=1}^{n} \left[\nu_i \left(-\frac{y_i}{\hat{\mu}_i} - \log(\hat{\mu}_i) \right) + \nu_i \log \nu_i + (\nu_i - 1) \log y_i - \log \Gamma(\nu_i) \right. \\ &\quad \left. - \nu_i \left(-1 - \log(y_i) \right) - \nu_i \log \nu_i - (\nu_i - 1) \log y_i + \log \Gamma(\nu_i) \right] \\ &= -2 \sum_{i=1}^{n} \left[\nu_i \left(-\frac{y_i}{\hat{\mu}_i} - \log \hat{\mu}_i + 1 + \log y_i \right) \right] \\ &= -2 \sum_{i=1}^{n} \left[\nu_i \left(\log \left(\frac{y_i}{\hat{\mu}_i} \right) - \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i} \right) \right] \\ &= -2\nu \sum_{i=1}^{n} w_i \left(\log \left(\frac{y_i}{\hat{\mu}_i} \right) - \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i} \right) . \end{aligned}$$

Für diese gilt asymptotisch $\frac{1}{\phi} D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) \sim \chi^2_{n-p}$ mit $\phi = \frac{1}{\nu}$.

3.8 Monte Carlo Simulation

Im Folgenden wird eine MC-Simulation mit $N = 1000$ Replikationen durchgeführt. Das Ziel ist es, die Auswirkungen unterschiedlicher Gestaltsparameter innerhalb einer Stichprobe auf die Maximum Likelihood Schätzer und deren Standardfehler aufzuzeigen. Das generalisierte lineare Modell besteht aus folgenden Komponenten:

- Link Funktion $g(\cdot) = \log(\cdot)$.
- Systematischen Komponente $\eta_i = 1 + 2x_i$.
- Unterschiedlichen Dispersionsparameter

$$\phi_i = \begin{cases} 1, & \text{falls } i = 1, \ldots, 25 \\ \frac{1}{2}, & \text{falls } i = 26, \ldots, 50 \\ \frac{1}{4}, & \text{falls } i = 51, \ldots, 75 \\ \frac{1}{8}, & \text{falls } i = 76, \ldots, 100. \end{cases}$$

Dies hat unterschiedliche Gestaltsparameter $v_i \in \{1, 2, 4, 8\}$ zur Folge.

Die Werte des Prädiktors x_i liegen auf dem Einheitsintervall, beginnend bei 0.01 und fortführend bis 1 mit äquidistanten Abständen. Daraus folgend werden die Responses y_i aus dem GLM

$$y_i \overset{ind}{\sim} Gamma(\nu_i, \exp(1 + 2x_i))$$

für $i = 1, \ldots, n = 100$ erzeugt.

In jeder Replikation der MC-Simulation werden ein ungewichtetes sowie ein gewichtetes GLM gerechnet. Die Gewichtung erfolgt mit passend gewählten „prior weights“, so dass der Gestaltsparameter für die gesamte Stichprobe einheitlich ist ($\nu = 2$). Wie im nachfolgenden R-Code ersichtlich, werden in jeder Replikation die Steigungsschätzer und deren Standardfehler sowie der geschätzte Dispersionsparameter des Modells bestimmt.

```
> set.seed(501)
# Initialisierung der Dataframes und Vektoren
> g.beta <- data.frame(Steigung=numeric())
> g.w.beta <- g.beta
> g.se <- data.frame("SE Steigung"=numeric())
> g.w.se <- g.se
> phi.hat <- c(); phi.hat.w <- c()

# Initialisierung Simulation
> N.rep <- 1000; n <- 100;
> x <- (1:100)/n; eta <- 1 + 2 * x; mu <- exp(eta);
> nu.global <- rep(2,times=n)  # Globaler Gestaltsparameter
> wts <- 1/rep(c(2,1,0.5,0.25), each=25)
> nu <- nu.global*wts # Individueller Gestaltsparameter
> theta <- mu / nu;
> phi <- 1/nu
```

```
> for(i in 1:N.rep){
  y <- rgamma(n, shape = nu, scale = theta)

  # Ungewichtet
  g <- glm(y ~ x, family = Gamma(link=log))
  g.beta[i,] <- g$coefficients[2]
  g.se[i,] <- sqrt(diag(vcov(g)))[2] # SE der MLE
  phi.hat[i] <- summary(g)$dispersion

  # Gewichtet
  g.w <- glm(y ~ x, family = Gamma(link=log), weights = wts)
  g.w.beta[i,] <- g.w$coefficients[2]
  g.w.se[i,] <- sqrt(diag(vcov(g.w)))[2] # SE der MLE
  phi.hat.w[i] <- summary(g.w)$dispersion
}
```

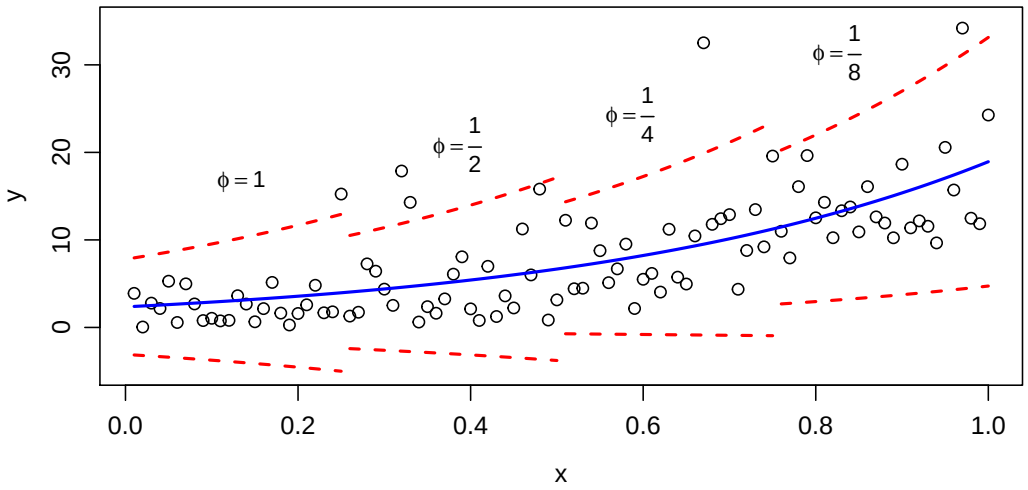

Abbildung 3.2: Eine exemplarische Datensituation mit Regressionskurve einer Replikation der MC-Simulation.

In Abbildung 3.2 sind eine generierte Datensituation und die dazugehörige Regressionskurve ersichtlich. Mit steigendem x nimmt auch die Varianz der Response y quadratisch zu. Die roten Linien entsprechen zwei Mal der Standardabweichung der Response y, die von den unterschiedlichen Dispersionsparametern ϕ_i abhängt. Mit steigendem Index i nimmt auch der Dispersionsparameter ϕ_i ab. Dadurch verkleinert sich die Standardabweichung der Response y. Dieser Effekt ist in den vier Sprungstellen der eingezeichneten Standardabweichung ersichtlich.

Mit Hilfe des anschließenden R-Codes wird die MC-Standardabweichung bestimmt. Diese entspricht der Standardabweichung der 1000 Schätzer und ist somit eine MC-Approximation der wahren Standardabweichung des Schätzers $\hat{\beta}_1$. Diese Berechnung wird für die Schätzer der ungewichteten und der gewichteten GLMs durchgeführt.

```
> beta.sd <- apply(g.beta,2,sd); beta.sd
0.257767
> beta.w.sd <- apply(g.w.beta,2,sd); beta.w.sd
0.2129822
```

Nachfolgende Abbildung 3.3 beinhaltet links einen Boxplot der 1000 Schätzer aus den gewichteten GLMs und rechts einen Boxplot, der die Schätzer der ungewichteten GLMs veranschaulicht. Man erkennt sofort, dass der wahre Wert $\beta_1 = 2$ (blaue Linie) im Median beides mal unverzerrt geschätzt wird. Die korrekt gewählten prior weights wirken sich auf den IQR des Boxplots aus. Dieser ist im gewichteten Fall kleiner als jener des ungewichteten Falls.

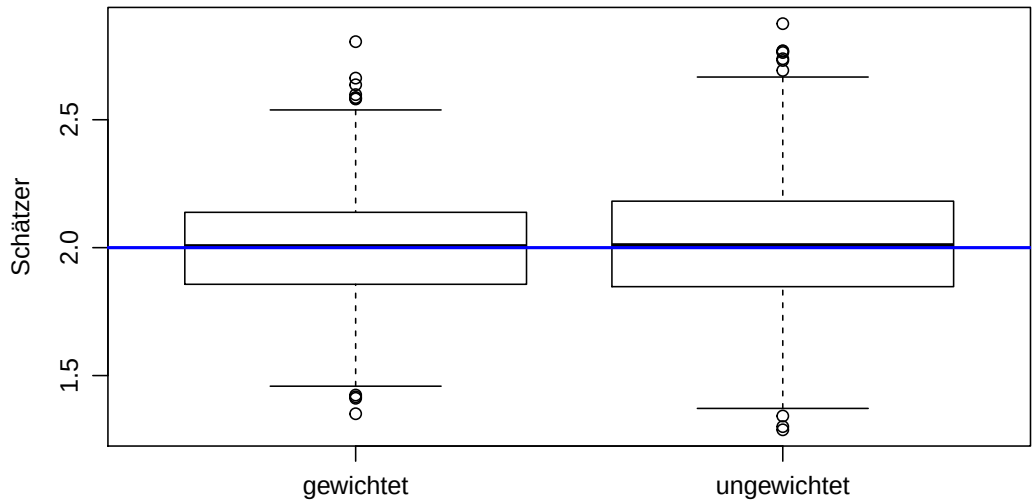

Abbildung 3.3: Boxplots der 1000 Schätzer für ungewichtetes und gewichtetes GLM. Die blaue Linie entspricht dem wahren Wert $\beta_1 = 2$.

Abbildung 3.4 beinhaltet Boxplots der Standardfehler der Schätzer $\widehat{sd}(\hat{\beta}_1)$. Auch hier gibt es eine Unterscheidung zwischen Schätzer, die aus einem ungewichteten und einem gewichteten GLM stammen.

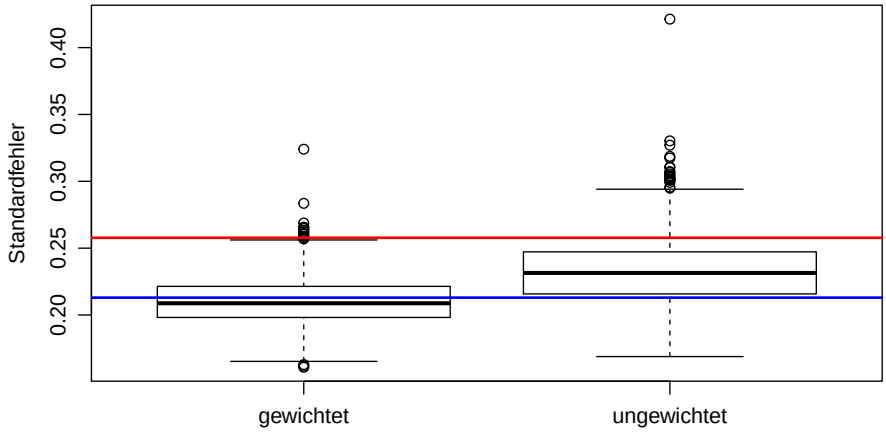

Abbildung 3.4: Boxplots der 1000 Standardfehler der Schätzer für ungewichtetes und gewichtetes GLM. Die blaue Linie entspricht der MC-Standardabweichung im gewichteten Fall (0.213). Die rote Linie entspricht der MC-Standardabweichung im ungewichteten Fall (0.2578).

In beiden Boxplots gibt es einige Ausreißer an den oberen Enden. Die Standardfehler der Schätzer der gewichteten GLMs schätzen die MC-Standardabweichung (blaue Linie) im Median fast exakt. Ein nicht so gutes Ergebnis liefert der rechte Boxplot. Hier unterschätzen die Standardfehler der Schätzer der ungewichteten GLMs die MC-Standardabweichung (rote Linie) im Median ein wenig. Die prior weights wirken sich somit auf die Verzerrtheit der Schätzung sowie auf die Länge des IQR der Boxplots aus.

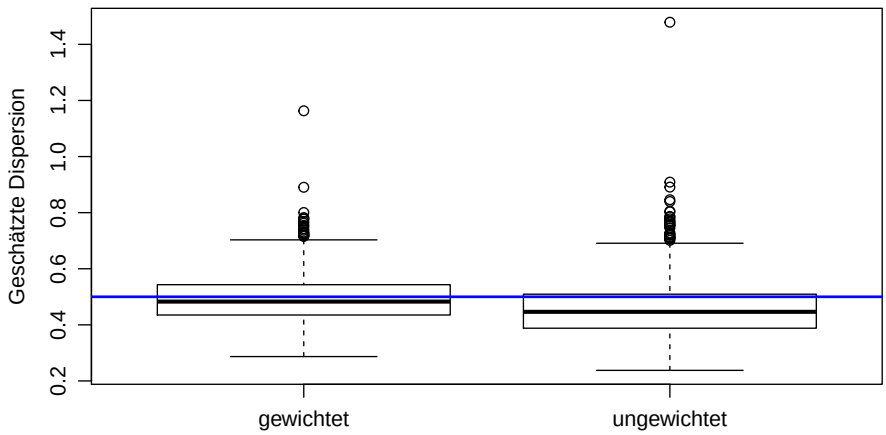

Abbildung 3.5: Boxplots der 1000 geschätzten Dispersionen für ungewichtetes und gewichtetes GLM. Die blaue Linie entspricht dem Wert des Dispersionsparameters $\phi = 0.5$.

Abschließend werden die Boxplots der geschätzten Dispersionsparameter gegenüber-

gestellt. Wie in Abbildung 3.5 ersichtlich, gibt es in beiden Fällen zahlreiche Ausreißer an den oberen Enden. Auffallend ist, dass die Dispersionsparameter, die aus den gewichteten GLMs kommen, die blaue Linie ($\phi = 0.5$) im Median exakt schätzen, wogegen die Dispersionsparameter der ungewichteten GLMs im Median eindeutig verzerrt sind.

KAPITEL 4

ANALYSE DES VERSCHLEISSBETRAGES

Der Verbrennungsmotor ist eine zentrale Komponente vieler Kraftfahrzeuge. Das Antriebssystem erfährt seit seiner Erfindung vor über 100 Jahren kontinuierliche Verbesserungen. Bei der Entwicklung neuer Motoren wird laut Hick et al. (2016) die Betrachtung von Reibung, Verschleiß sowie Blow-By an Bedeutung gewinnen. Im Zuge einer Kooperation der Technischen Universität Graz mit dem Unternehmen *AC²T research GmbH* werden speziell diese drei Aspekte anhand eines Einzylinder-Forschungsmotors untersucht.

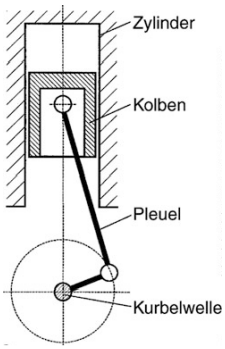

Abbildung 4.1: Vereinfachte Darstellung eines Motors (van Basshuysen und Schäfer, 2017, S. 11, Abb. 2.2).

Bei laufendem Motor bewegt sich die Kolbengruppe, bestehend aus Kolben, Kolbenring (Kompressionsring) und Kolbenbolzen im Zylinder ständig auf und ab. Während dieser oszillierenden Bewegung ist die Kolbengruppe besonders starken thermischen, chemischen und tribologischen Belastungen ausgesetzt. Aus diesem Grund liegt der

Fokus der Untersuchung in der Analyse der Reibung und des Verschleißes der Kolbengruppe (Hick et al., 2016).

Abbildung 4.2: Exemplarischer Kolben für einen Ottomotor mit Direkteinspritzung (van Basshuysen und Schäfer, 2017, S. 110, Abb. 7.6).

4.1 Motorkenngrößen

Motorkenngrößen sind ein wichtiges Hilfsmittel der Leistungs- und Verbrauchsbeurteilung und werden beim Vergleich verschiedener Aggregate herangezogen. Die Definitionen der nachfolgenden Motorkenngrößen stammen aus van Basshuysen und Schäfer (2017, Kap. 3).

Drehzahl

Die Drehzahl ist eine Größe, die die Häufigkeit einer Drehbewegung angibt und lautet

$$n = \frac{\text{Anzahl Umdrehungen der Kurbelwelle}}{\text{Zeit}}. \tag{4.1}$$

Häufig wird die Drehzahl in Umdrehungen pro Minuten (min^{-1}) angegeben.

Drehmoment

Das Drehmoment drückt die Drehwirkung einer Kraft auf einen drehbar gelagerten Körper aus. Das Pleuel setzt die Auf-/Abbewegung des Kolbens in eine Rotationsbewegung der Kurbelwelle um. Dabei lässt sich die Kraft F_{ST}, die auf die Pleuelstange wirkt, in die Tangentialkraft F_T und die Radialkraft F_R zerlegen.

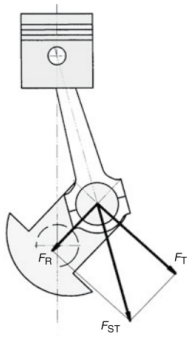

Abbildung 4.3: Aufteilung der Stangenkraft F_{ST} (van Basshuysen und Schäfer, 2017, S. 68, Abb. 6.15).

Das Motordrehmoment M_d ist von der Tangentialkraft F_T und dem Radius der Kurbelwelle r abhängig. Somit ergibt sich

$$M_d = F_T \cdot r. \tag{4.2}$$

Die international verwendete Einheit ist Newtonmeter (Nm).

4.2 Verschleiß

Verschleiß ist nach van Basshuysen und Schäfer (2017, S. 520) gegeben als „der fortschreitende Materialverlust aus der Oberfläche eines festen Körpers, hervorgerufen durch mechanische Ursachen, das heißt Kontakt und Relativbewegung eines festen, flüssigen oder gasförmigen Gegenkörpers“.

Der Verschleiß kann zahlenmäßig als Verschleißbetrag in unterschiedlichen Dimensionen angegeben werden. Hierbei unterscheidet man zwischen Verschleißhöhe W_h, Verschleißfläche W_q und Verschleißvolumen W_V.

Häufiger erfolgt die Angabe des Verschleißbetrages in Relation zu einer Bezugsgröße (Zeit oder Weg). Daraus resultieren sogenannte Verschleißraten wie die Verschleißgeschwindigkeit (zeitbezogene Verschleißrate) oder die Verschleißintensität (wegbezogene Verschleißrate).

Die zeitliche Entwicklung des Verschleißbetrages lässt sich in einem Verschleiß-Zeit Diagramm veranschaulichen. Im nachfolgenden Bild 4.4 ist der Ablauf des Verschleißbetrages, der unter typischen Versuchsbedingungen erzeugt wird, vereinfacht dargestellt (Czichos und Habig, 2015, S. 276).

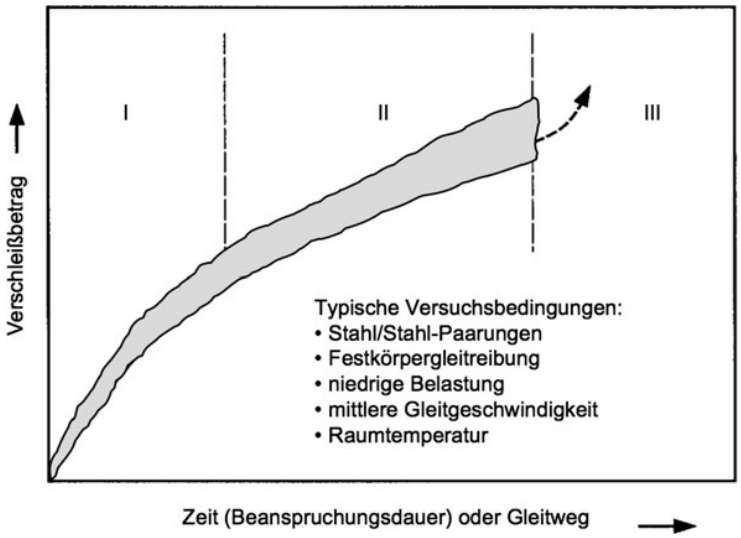

Abbildung 4.4: Typisches Verschleiß-Zeit Diagramm (Czichos und Habig, 2015, S. 276, Abb. 8.38).

Der Verlauf des Verschleißbetrages kann zeitlich gesehen in drei Bereiche gegliedert werden. Der erste Zeitabschnitt, auch *degressiver Bereich* genannt, ist durch einen quadratwurzelförmigen Anstieg gekennzeichnet. Im zweiten Bereich nimmt der Verschleißbetrag linear über die Zeit zu. Der abschließende *progressive Bereich* weist einen exponentiellen Anstieg des Verschleißbetrages auf.

Im Maschinenbetrieb ist der Verschleiß unvermeidbar und wirkt sich negativ auf die Lebensdauer der Komponenten aus. Um Kosten zu sparen und die Lebensdauer der Maschinen zu erhöhen, ist eine Reduktion des Verschleißes wünschenswert.

4.3 Problemstellung

Unter der Leitung von Herrn Dipl. Ing. Walch wurde mit Hilfe eines Versuchsplanes der lineare Verschleißbetrag W_h des Kolbenrings für unterschiedliche Drehzahl-Drehmoment Kombinationen ermittelt. Die Verschleißmessung erfolgte mit der Radio-Isotope Concentration Methode (RIC) (siehe Hick et al., 2016, S. 11). In der nachfolgenden Tabelle 4.1 sind 21 Motoreinstellungen und die Anzahl an erhobenen Verschleißbeträgen ersichtlich.

Im Versuchsplan wurden auch Messvorgänge für manche Motoreinstellungen wiederholt. Beispielsweise besitzen die Messvorgänge M1 bis M4 nahezu die exakten Drehzahl-Drehmoment Einstellungen. Analoges gilt für die Messvorgänge M6 bis

	Motoreinstellung				Motoreinstellung		
Nummer	Drehzahl	Drehmoment	Anzahl	Nummer	Drehzahl	Drehmoment	Anzahl
M1	2500	34.70	723	M12	1350	33.20	364
M2	2500	34.50	546	M13	1350	33.20	1805
M3	2500	34.80	703	M14	2000	10.70	659
M4	2500	35.20	786	M15	2200	8.40	1661
M5	2499	21.70	803	M16	2000	22.50	791
M6	1500	23.00	694	M17	1500	−7.40	1379
M7	1502	23.00	442	M18	2999	21.10	729
M8	1500	23.00	3467	M19	1000	23.50	1680
M9	1500	34.90	790	M20	2700	−10.10	1095
M10	1500	11.30	1622	M21	1000	−6.40	1277
M11	3000	9.60	820				

Tabelle 4.1: Übersicht der Messvorgänge M1 bis M21 mit ihrer Motoreinstellung.

M8 und die beiden Messvorgänge M12 und M13.

Während eines Messvorganges für eine bestimmte Drehzahl-Drehmoment Kombination wird alle 30 Sekunden der Verschleißbetrag ermittelt. Daraus resultieren 120 Messungen pro Stunde. Wird daraufhin der Verschleißbetrag differenziert, ergeben sich sogenannte Verschleißraten (siehe Abschnitt 4.2).

Die erhobenen Verschleißbeträge fallen zeitlich gesehen in den zweiten Bereich und sollten daher linear über die Zeit zunehmen. Das erste Ziel des praktischen Abschnitts ist somit die Schätzung des Steigungskoeffizienten und dessen dazugehöriger Standardfehler. Hierfür eignet sich das einfache lineare Regressionsmodell.

4.4 Ermittlung der Verschleißraten

Bevor die Verschleißrate mittels einer einfachen linearen Regression ermittelt werden kann, müssen die Motoreinstellungen auf etwaige Ausreißer untersucht werden. In der nachfolgenden Abbildung 4.5 sind links das Drehmoment (Nm) und rechts die Drehzahl (n) des fünften Messvorganges abgebildet.

Wie in beiden Grafiken ersichtlich, gibt es vier Zeitbereiche, in der die Drehzahl und das Drehmoment von den eingestellten 2500 Umdrehungen und 21.70 Nm abweichen. Vergleicht man die Indizes auf der x-Achse der beiden Scatterplots, so erkennt man, dass die fehlerhaften Einstellungen dieselben Indizes aufweisen. Da die fehlerhaften Einstellungen das Endresultat verfälschen könnten, werden diese Datenzeilen aus der zukünftigen Analyse ausgeschlossen.

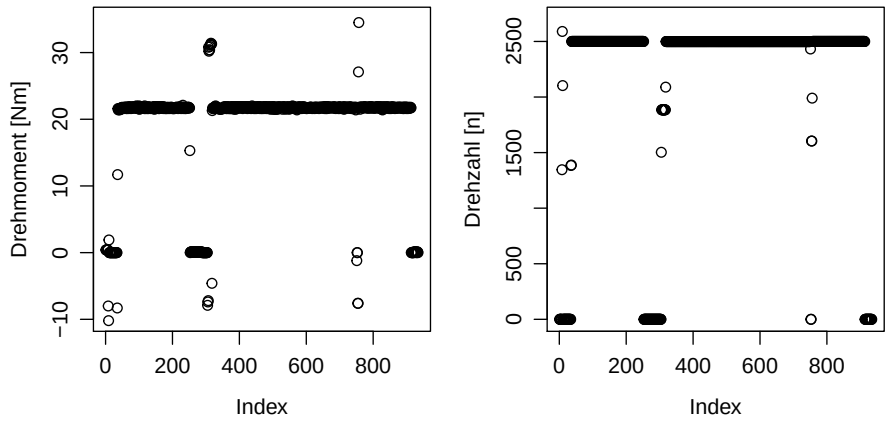

Abbildung 4.5: Drehmoment (links) und Drehzahl (rechts) des fünften Messvorganges.

Die Motoreinstellungen des achten Messvorganges sind ebenfalls nicht gänzlich korrekt. In diesem Fall gibt es einen sehr kurzen Zeitbereich, in der die Drehzahl von 1500 Umdrehungen auf Null Umdrehungen springt. Gleiches gilt für das Drehmoment. Auch hier überschneiden sich die Indizes der fehlerhaften Motoreinstellungen und werden aus der anstehenden Analyse ausgeschlossen.

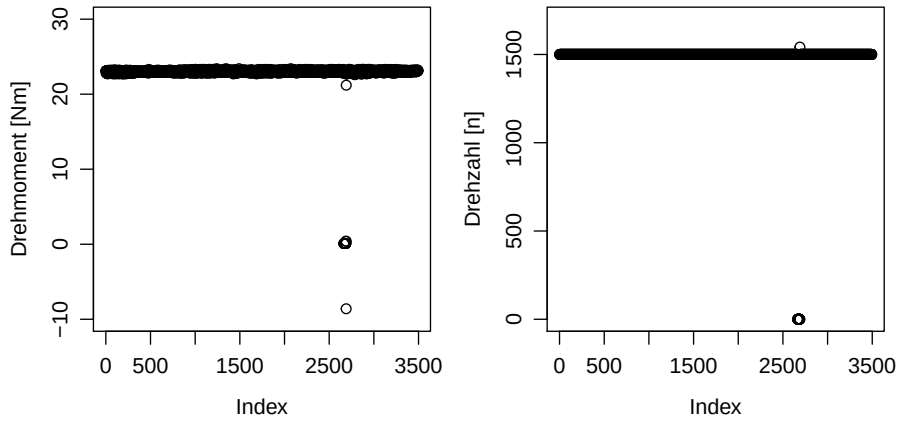

Abbildung 4.6: Drehmoment (links) und Drehzahl (rechts) des achten Messvorganges.

Die Motoreinstellungen der anderen 19 Messvorgänge weisen keine besonderen Abweichungen auf.

Mit der R-Funktion `readFile` (siehe Anhang - Abschnitt A.2) wird ein bereinigter

Datensatz eingelesen, wobei der Funktionsparameter `i` für den Index des Dateinamens des i-ten Messverfahrens steht. Wie bereits am Beginn des Abschnitts 4.3 erwähnt, erfolgt die Messung des Verschleißbetrages im 30 Sekunden-Takt. Dadurch resultieren 120 Messungen des Verschleißbetrages pro Stunde. Folglich benötigt man für die Messung eines Verschleißbetrages 1/120 Stunden.

Im Anschluss wird die Verschleißrate für den ersten Messvorgang M1 ermittelt. Die Bestimmung der Steigungsparameter der restlichen 20 Messvorgänge erfolgt analog, wobei die Ergebnisse im Abschnitt A.2 des Anhangs aufgelistet sind.

```
> df <- readFile(1)
> mod <- lm(df$verschleiss ~ df$zeit)
> summary(mod)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.22888    0.07949   78.36   <2e-16 ***
df$zeit      0.28260    0.02283   12.38   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Man sieht in der `summary`-Ausgabe, dass alle Parameter sehr signifikant ist. Vergeht eine Stunde, so erhöht sich der mittlere Verschleißbetrag um 0.2826 nm.

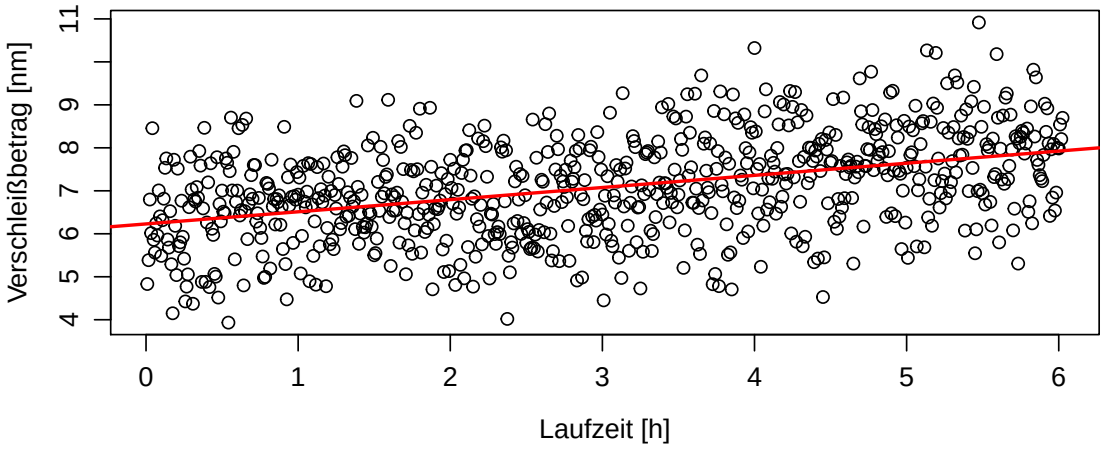

Abbildung 4.7: Gemessene Verschleißbeträge des ersten Messvorganges werden über die Zeit aufgetragen. Die rote Linie entspricht der Regressionsgerade.

Zusätzlich wird das Modell `mod` nun mit Hilfe von graphischen Diagnoseplots weiter untersucht.

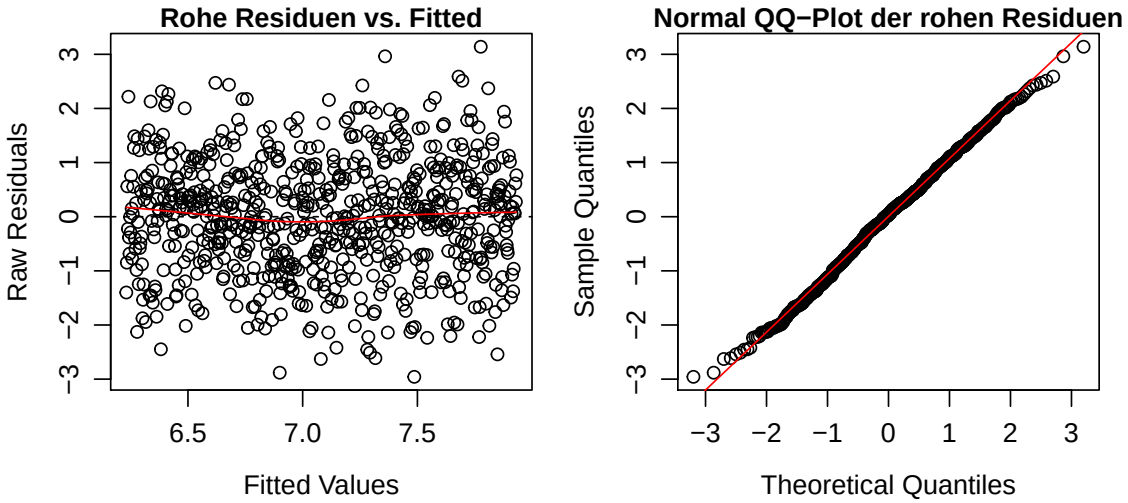

Abbildung 4.8: Diagnoseplots für das Modell `mod`.

Die Residuen bilden im Plot der *Rohen Residuen vs. Fitted Values* ein „horizontales Band“ um die Null-Linie. Daraus kann man auf Homoskedastizität des nicht beobachtbaren statistischen Fehlers schließen. Des Weiteren wird die Annahme der Homoskedastizität durch den Breusch-Pagan-Hypothesentest bestätigt. Dieser verwirft H_0 (Homoskedastizität) mit einem p-Wert von 0.2648 nicht. Die rohen Residuen sind, wie im *Normal QQ-Plot* ersichtlich, eindeutig normalverteilt.

Die Verschleißdaten, die mittels der Messmethode im Kolbenring erhoben wurden, sind nicht wahrheitsgetreu. Aus diesem Grund müssen die berechneten Verschleißraten und die dazugehörigen Standardfehler (siehe Tabelle A.1 im Anhang) mit einem Skalierungsfaktor von

$$c = \frac{1200}{330} = 3.636364$$

multipliziert werden (Hick et al., 2016, S. 17).

Mit dem nachfolgenden `R`-Code werden die korrigierten Parameter ermittelt. Diese stellen im anschließenden Kapitel 5 die Grundlage für weitere Berechnungen dar.

```
> skal.faktor <- 1200/330
> df <- data.frame(wear.rate=numeric(),"se"=numeric(),
                   "dz"=numeric(), "dm"=numeric())

> for(i in 1:21){
   aktueller.df <- readFile(i)
```

```
   mod <- lm(aktueller.df$verschleiss ~ aktueller.df$zeit);
   df[i,1] <- mod$coefficients[2] * faktor
   df[i,2] <- sqrt(diag(vcov(mod)))[2] * faktor
   p.val <- bptest(mod)$p.value
   df[i,3] <- median(aktueller.df$drehzahl)
   df[i,4] <- median(aktueller.df$drehmoment)
   if(p.val < 0.05){
     df[i,2] <- sqrt(diag(vcovHC(mod,type = "HC3")))[2] * skal.faktor
   }
 }
> attach(df)
```

Im Anschluss werden die korrigierten Verschleißraten $\hat{\beta}_1$ der Tabelle 4.2 etwas näher betrachtet. Die Messvorgänge M1 bis M4 besitzen in etwa die selbe Motoreinstellung und die Verschleißbeträge wurden über einen ähnlichen Zeitraum gemessen. Somit befinden sich die dazugehörigen Steigungskoeffizienten auch alle in derselben Größenordnung von 1 nm/h.

	Dz. [n]	Dm. [Nm]	Rate	SE		Dz. [n]	Dm. [Nm]	Rate	SE
1	2500	34.7	1.0276	0.0830	12	1350	33.2	3.0335	0.3051
2	2500	34.5	1.0307	0.1276	13	1350	33.2	1.6153	0.0327
3	2500	34.8	0.9520	0.0915	14	2000	10.7	0.4777	0.1388
4	2500	35.2	0.9427	0.0784	15	2200	8.4	0.4935	0.0401
5	2499	21.7	1.0563	0.0772	16	2000	22.5	1.1346	0.1116
6	1500	23.0	0.1114	0.0996	17	1500	-7.4	0.4570	0.0503
7	1502	23.0	0.3874	0.1754	18	2999	21.1	1.1259	0.1301
8	1500	23.0	0.5033	0.0101	19	1000	23.5	0.2826	0.0354
9	1500	34.9	0.8623	0.0878	20	2700	-10.1	0.5927	0.0780
10	1500	11.3	0.4131	0.0302	21	1000	-6.4	0.1875	0.0612
11	3000	9.6	1.1934	0.0917					

Tabelle 4.2: Übersicht der korrigierten Verschleißrate inklusive Standardfehler für eine bestimmte Motoreinstellung.

Die Messvorgänge M6 bis M8 weisen ebenfalls die gleichen Motoreinstellungen auf, die Dauer der erhobenen Verschleißbeträge ist jedoch stark unterschiedlich. Auffallend ist, dass der Steigungskoeffizient des sechsten Messvorganges viel kleiner ist, als jener des siebenten oder achten Messvorganges.

Hingegen unterscheiden sich die Verschleißraten $\hat{\beta}_1$ der Messvorgänge M12 und M13 stark. Obwohl gleiche Motoreinstellungen gegeben sind, ist die Verschleißgeschwin-

digkeit im zwölften Messvorgang doppelt so groß wie jene des dreizehnten Messvorganges.

In der Grafik 4.9 sind die Steigungskoeffizienten und die dazugehörigen Standardfehler aus der Tabelle 4.2 abgebildet. Auffallend ist, dass die Verschleißrate des Messvorganges 12 mit 3.0335 nm/h und dessen Standardfehler mit 0.3051 mit Abstand am größten ist. Die restlichen Verschleißraten befinden sich zwischen 0.1 und 1.5 nm/h unabhängig von der jeweiligen Drehzahl-Drehmoment Einstellung.

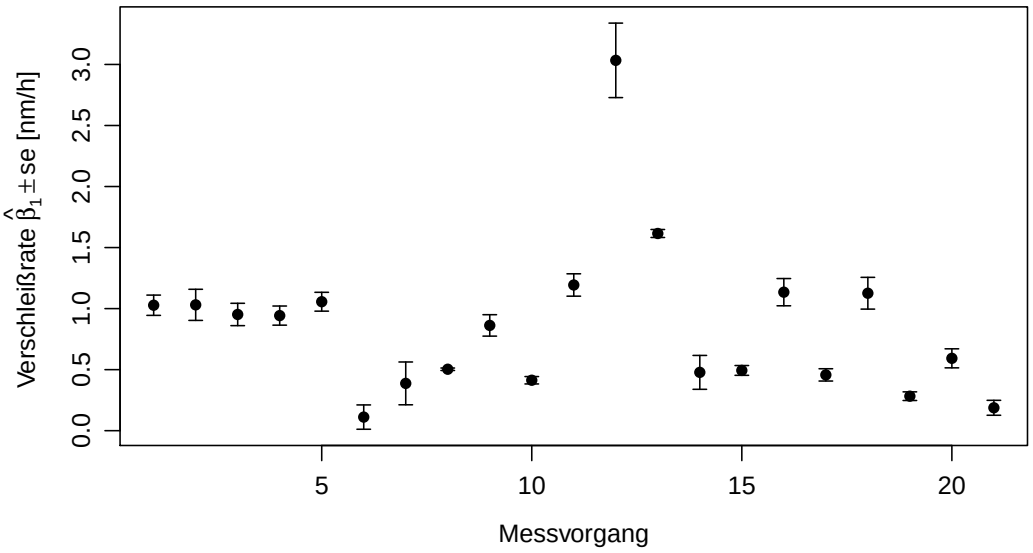

Abbildung 4.9: Graphische Veranschaulichung der unterschiedlichen Steigungskoeffizienten inklusive Standardfehler.

KAPITEL 5

GEWICHTETE REGRESSIONSMODELLE

Im Kapitel 4 wurden die unterschiedlichen Verschleißgeschwindigkeiten (nm/h) und deren Standardfehler ermittelt. Von Interesse ist nun der Zusammenhang von zu erwartender Verschleißgeschwindigkeit und der Drehzahl dz sowie dem Drehmoment dm. Das Ziel ist also die Erstellung eines Regressionsmodells, das die Verschleißgeschwindikeit aus den Größen dz und dm vorhersagt.

Aufgrund der nicht konstanten Standardfehler der Verschleißgeschwindigkeiten folgt, dass die Responsevariablen des angehenden Regressionsmodells heteroskedastisch sind. Wie bereits aus Abschnitt 2.2 bekannt, kann mit passend gewählten Gewichten der Heteroskedastizität entgegen gewirkt werden. Für gewöhnlich müssen die Gewichte aus den Daten geschätzt werden, im Optimalfall sind diese bereits vorab bekannt. Im vorherigen Kapitel wurde geschätzt, wie unterschiedlich stark die Responsevariablen streuen. Somit kann diese Information für die Gewichtung verwendet werden.

5.1 Gewichtete lineare Regression

Unter der Annahme einer Normalverteilung mit nicht konstanter Varianz für die Responsevariablen y_i betrachtet man zuerst ein gewichtetes lineares Regressionsmodell. Den heteroskedastischen Responsevariablen wird mit Gewichten w_i entgegengewirkt. Laut Abschnitt 2.2.5 erhält man optimale Resultate, wenn die Gewichte w_i die Kehrwerte der Varianzen σ_i^2 sind, d.h.

$$w_i = \frac{1}{\sigma_i^2},$$

mit $\text{Var}(y_i) = \sigma_i^2$. Als nächstes soll der Zusammenhang zwischen Verschleißgeschwindigkeit und Drehzahl sowie Drehmoment mit Hilfe von Scatterplots veranschaulicht werden.

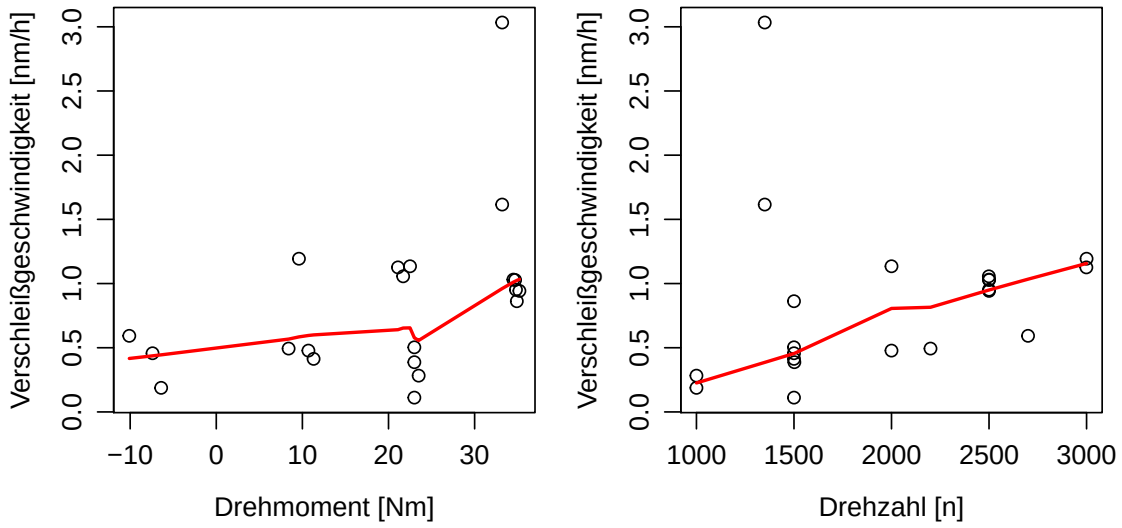

Abbildung 5.1: Scatterplots der Verschleißgeschwindigkeit gegen Drehmoment/Drehzahl inklusive lowess-Kurve.

Wie im linken Scatterplot ersichtlich, scheint es einen quadratischen Zusammenhang zwischen Drehmoment und Verschleißgeschwindigkeit zu geben. In der rechten Abbildung vermuten wir einen linearen Zusammenhang zwischen Drehzahl und Verschleißgeschwindigkeit. Im Anschluss werden die Abhängigkeiten mittels gewichteter linearer Regressionsmodelle untersucht.

```
> wts <- 1/(se^2)
> summary(lm(wear.rate ~ dm + I(dm^2), weights = wts))

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.3089290  0.1346861   2.294   0.0341 *
dm          -0.0198362  0.0122581  -1.618   0.1230
I(dm^2)      0.0013340  0.0004017   3.321   0.0038 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Falls die Verschleißgeschwindigkeit `wear.rate` durch das Drehmoment erklärt wird, ist der quadratische Drehmomentterm `I(dm^2)` signifikant. Erhöht man das Drehmoment von 0 auf 10 Nm, so reduziert sich die Verschleißgeschwindigkeit im Mittel

um 0.065 nm/h. In der linken Abbildung 5.2 erkennt man, dass es ab 10 Nm wieder zu einer Geschwindigkeitszunahme kommt. Bei einem Anstieg des Drehmoments von 10 auf 20 Nm nimmt die mittlere Verschleißgeschwindigkeit um 0.0719 nm/h zu.

```
> summary(lm(wear.rate ~ dz, weights = wts))

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.2128337  0.3198968   0.665    0.514
dz          0.0002360  0.0001999   1.181    0.252
```

Hingegen scheint es keinen direkten Zusammenhang zwischen `wear.rate` und `dz` zu geben. Versucht man die Verschleißrate `wear.rate` durch die reziproke Drehzahl `I(1/dz)` oder die wurzeltransformierte Form der Drehzahl `sqrt(dz)` zu erklären, so sind diese Terme ebenfalls nicht signifikant. Im rechten Scatterplot 5.2 ist die zuvor errechnete Regressionsgerade eingezeichnet. Nimmt die Drehzahl um 1000 Umdrehungen/Minute zu, dann steigt die mittlere Verschleißgeschwindigkeit um 0.236 nm/h.

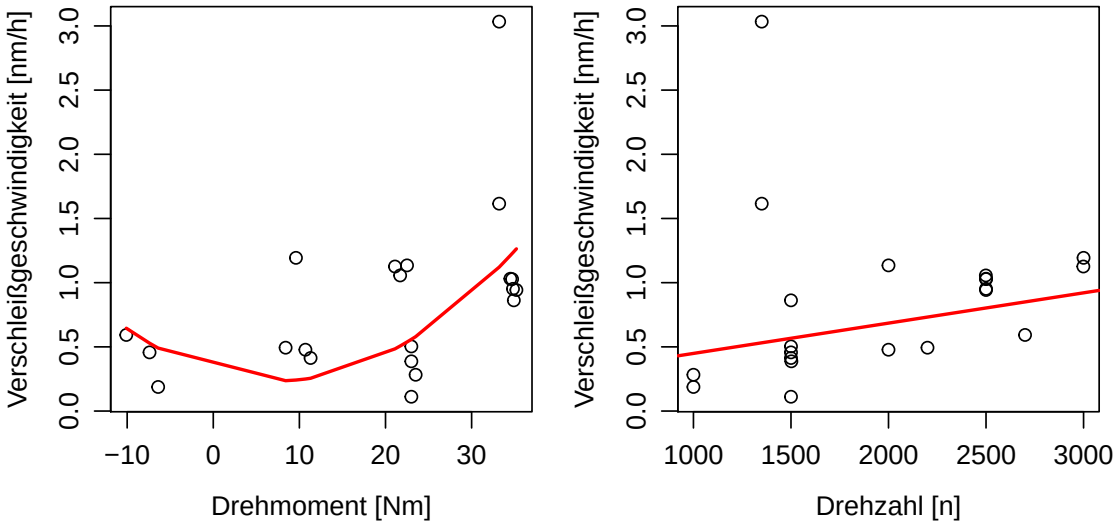

Abbildung 5.2: Scatterplots der Verschleißgeschwindigkeit gegen Drehmoment/Drehzahl inklusive den zuvor ermittelten Regressionskurven.

Im Anschluss wird ein gewichtetes Regressionsmodell betrachtet, welches eine Linearkombination der beiden Prädiktoren `dm` und `dz` annimmt. Die Vermutung ist, dass die Drehzahl in Interaktion mit dem linearen sowie quadratischen Drehmomentterm einen Einfluss auf die Verschleißrate hat.

```
> m1 <- lm(wear.rate ~ dm * dz + I(dm^2) + I(dm^2):dz, weights = wts)
> summary(m1)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.697e-01  3.008e-01  -2.226 0.041737 *
dm          -8.814e-02  2.790e-02  -3.159 0.006482 **
dz           5.662e-04  1.668e-04   3.394 0.004003 **
I(dm^2)      4.958e-03  9.944e-04   4.986 0.000163 ***
dm:dz        4.014e-05  1.571e-05   2.554 0.022001 *
dz:I(dm^2)  -2.049e-06  5.372e-07  -3.815 0.001691 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.492 on 15 degrees of freedom
Multiple R-squared:  0.8002, Adjusted R-squared:  0.7336
F-statistic: 12.02 on 5 and 15 DF,  p-value: 8.255e-05
```

Man erkennt in der `summary`-Ausgabe, dass alle Parameter signifikant sind. Die zuvor aufgestellte Vermutung hat sich somit bestätigt. Der Drehzahlterm konnte in Kombination mit dem Drehmoment an Signifikanz zunehmen. Bei gleichbleibender Drehzahl (1000 n) und Erhöhung des Drehmoments von 10 auf 20 Nm, nimmt die mittlere Verschleißgeschwindigkeit um 0.393 nm/h zu.

Die Differenz zwischen dem R^2 (0.8002) und dem R^2_{adj} (0.7336) weist auf ein leichtes Overfitten der Daten hin. Die Nullhypothesen der t-Tests (H_0: $\beta_j = 0$ für $j = 1, \ldots, 5$) werden aufgrund der kleinen p-Werte alle verworfen. Daher kann kein Parameter aus dem Modell entfernt werden. Der R^2-Wert weist außerdem darauf hin, dass 80.02% der Variabilität der Daten durch das Modell erklärt werden. Des Weiteren wird die Nullhypothese des Overall-F-Tests ($H_0 : \beta_1 = \cdots = \beta_5 = 0$) mit einem p-Wert von nahezu Null eindeutig verworfen.

Nachdem kein Prädiktor aus dem Modell `m1` entfernt werden kann, wird eine Residuenanalyse anhand der gewichteten Residuen r_i^* durchgeführt. Diese sind im gewichteten Regressionsmodell laut (2.26) durch

$$r_i^* = \sqrt{w_i} \cdot r_i$$

gegeben. In `R` erhält man r_i^* für ein gegebenes Modellobjekt durch folgendes Kommando:

```
> res <- residuals(m1) * sqrt(wts)
```

Die graphischen Diagnoseplots der Residuenanalyse werden mit dem nachfolgenden R-Code erzeugt.

```
> fit <- fitted(m1);
> layout(matrix(c(1,2,3), 1, 3, byrow = TRUE))

> plot(fit,res, xlab="Fitted Values", ylab="Wtd. Residuals",
  main="Wtd. Residuals vs. Fitted")
> abline(0, 0,lty = 2)
> lines(lowess(x=fit,y=res), col="red", lwd="1")

> plot(fit,wear.rate, xlab="Fitted Values", ylab="Observed",
  main="Observed vs. Fitted")
> abline(0,1, col="red")

> qqnorm(res,main="Normal QQ-Plot Wtd. Res")
> qqline(res,col = 2)
```

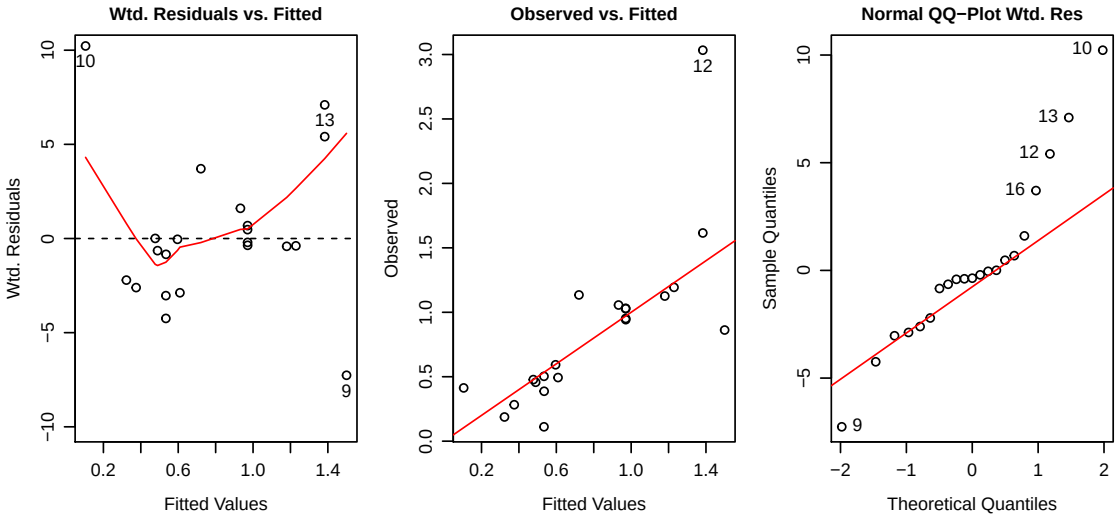

Abbildung 5.3: Diagnoseplots des Modells `m1`.

Im Scatterplot *Weighted Residuals vs. Fitted Values* gibt es jeweils einen Ausreißer am oberen bzw. am unteren Ende. Die gewichteten Residuen streuen um die Null-Linie. Die Gewichte konnten somit der Heteroskedastizität scheinbar entgegenwirken. Wie im *Observed vs. Fitted Plot* ersichtlich, befinden sich alle Werte in der Nähe der ersten Mediane. Dies zeugt von einer guten Modellanpassung. Einzig der zwölfte Wert wird stark unterschätzt. Da es im *Normal QQ-Plot der gewichteten*

Residuen am oberen Ende zahlreiche Ausreißer gibt, scheinen die gewichteten Residuen nicht normalverteilt zu sein.

Die vorherige Analyse des Modells `m1` konnte bestätigen, dass die Drehzahl `dz` nur in Kombination mit `dm` und `I(dm^2)` einen Einfluss auf `wear.rate` hat. Da das Modell `m1` jedoch nur den linearen Drehzahlterm `dz` beinhaltet, wird das anschließende Modell `m2` die wurzeltransformierte Drehzahl `sqrt(dz)` zusätzlich inkludieren. Wir vermuten, dass die wurzeltransformierte Drehzahl `sqrt(dz)` und der quadrierte Term `dz`, wiederum in Kombination mit `dm` und `I(dm^2)` einen Einfluss auf die Verschleißrate `wear.rate` haben werden.

```
> m2 <- lm(wear.rate ~ dm * sqrt(dz) + I(dm^2) * dz, weights = wts)
> summary(m2)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.703e-01  2.378e+00  -0.240  0.81390
dm          -1.503e-01  5.618e-02  -2.675  0.01812 *
sqrt(dz)    -5.542e-03  1.124e-01  -0.049  0.96138
I(dm^2)      4.825e-03  1.019e-03   4.735  0.00032 ***
dz           6.354e-04  1.321e-03   0.481  0.63793
dm:sqrt(dz)  3.183e-03  1.372e-03   2.321  0.03590 *
I(dm^2):dz  -1.961e-06  5.497e-07  -3.568  0.00309 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.709 on 14 degrees of freedom
Multiple R-squared:  0.7951, Adjusted R-squared:  0.7073
F-statistic: 9.056 on 6 and 14 DF,  p-value: 0.0003676
```

Im Vergleich zum Modell `m1` gibt es hier ein paar nicht signifikante Parameter. Dies spiegelt sich auch in der Differenz zwischen R^2 (0.7951) und R^2_{adj} (0.7073) wider. Der Abstand der beiden Werte weist auf ein leichtes Overfitten der Daten hin. Auffallend ist, dass die Drehzahlterme `sqrt(dz)` und `dz` scheinbar keinen direkten Einfluss auf die mittlere Verschleißrate `wear.rate` haben. Da jedoch beide Interaktionsterme signifikant sind, verbleiben `sqrt(dz)` und `dz` im Modell.

Bei gleichbleibender Drehzahl (1000 n) und Erhöhung des Drehmoments von 10 auf 20 Nm, kommt es auch hier zu einer Geschwindigkeitszunahme. Im Mittel erhöht sich nämlich die Verschleißgeschwindigkeit um 0.363 nm/h. Der hohe R^2-Wert weist darauf hin, dass 79.51% der Variabilität der Daten durch das Modell erklärt sind. Im Anschluss erfolgt die Analyse mit den graphischen Diagnoseplots.

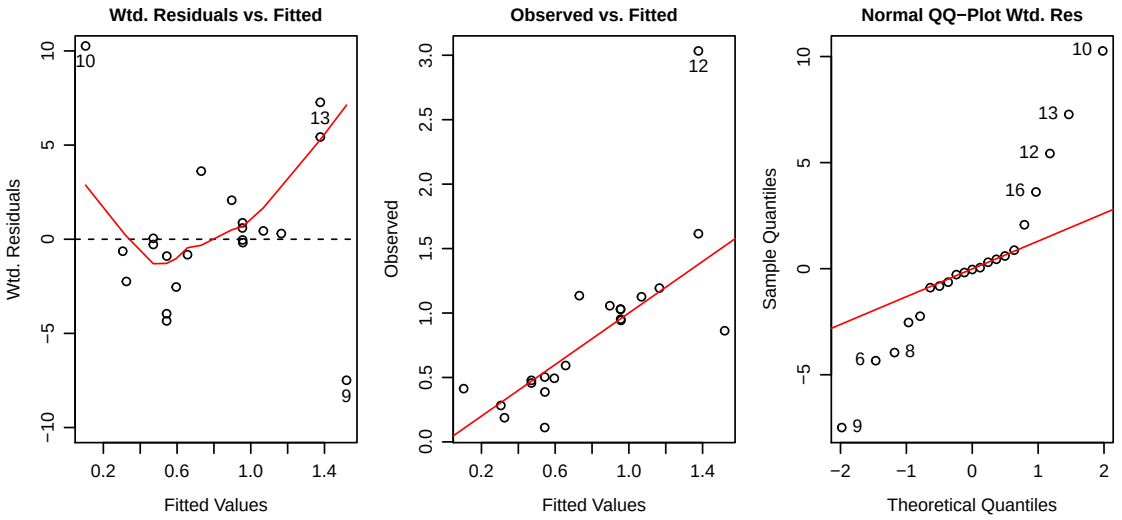

Abbildung 5.4: Diagnoseplots des Modells `m2`.

Im Scatterplot *Weighted Residuals vs. Fitted Values* gibt es wie zuvor jeweils einen Ausreißer am oberen bzw. am unteren Ende. Zusätzlich streuen die Residuen um die Nulllinie und es ist kein Muster erkennbar. Die Gewichtung konnte somit die störende Varianzinhomogenität scheinbar beheben. Wie im *Observed vs. Fitted Plot* erkennbar, liegen die Punkte in der Nähe der ersten Mediane. Auch hier wird der Punkt Nummer 12 stark unterschätzt. Im *Normal QQ-Plot der gewichteten Residuen* gibt es an beiden Enden zahlreiche Ausreißer. Das vorliegende Muster deutet darauf hin, dass die gewichteten Residuen zu sehr konzentriert sind, um normalverteilt zu sein.

Im nächsten Schritt wird das bessere der zwei erhobenen Modelle ermittelt. Ein beliebtes Werkzeug, um sich zwischen mehreren konkurrierenden Modellen zu entscheiden, ist das Akaike Informationskriterium (siehe Abschnitt 2.2.6). Dieses findet einen Mittelweg zwischen zu großer Modellkomplexität und guter Datenanpassung. Der AIC-Wert kann in `R` für ein gegebenes Modell mit Hilfe der Funktion `AIC` bestimmt werden.

```
> AIC(m1,m2)

    df       AIC
m1   7 20.19724
m2   8 22.72618
```

Es besteht natürlich auch die Möglichkeit, diesen Wert händisch mit den Formeln (2.29) oder (2.30) zu berechnen.

```
> p <- length(m1$coefficients)
> -2 * logLik(m1) + 2 * (p+1) # AIC m1
20.19724

> n <- length(wear.rate)
> p <- length(m2$coefficients)
> sse.w <- sum((residuals(m2)*sqrt(wts))^2)
> n * (log(2*pi) + 1 + log(sse.w/n)) + 2*(p + 1) - sum(log(wts)) # AIC m2
22.72618
```

Weil das Modell `m1` den kleineren AIC-Wert aufweist, entscheiden wir uns für dieses gewichtete lineare Regressionsmodell.

Im Anschluss erfolgt die graphische Veranschaulichung der Regressionskurve des Modells `m1`. Diese Grafik wurde mit Hilfe des `R`-Pakets `scatterplot3d` gezeichnet (Ligges und Mächler, 2003).

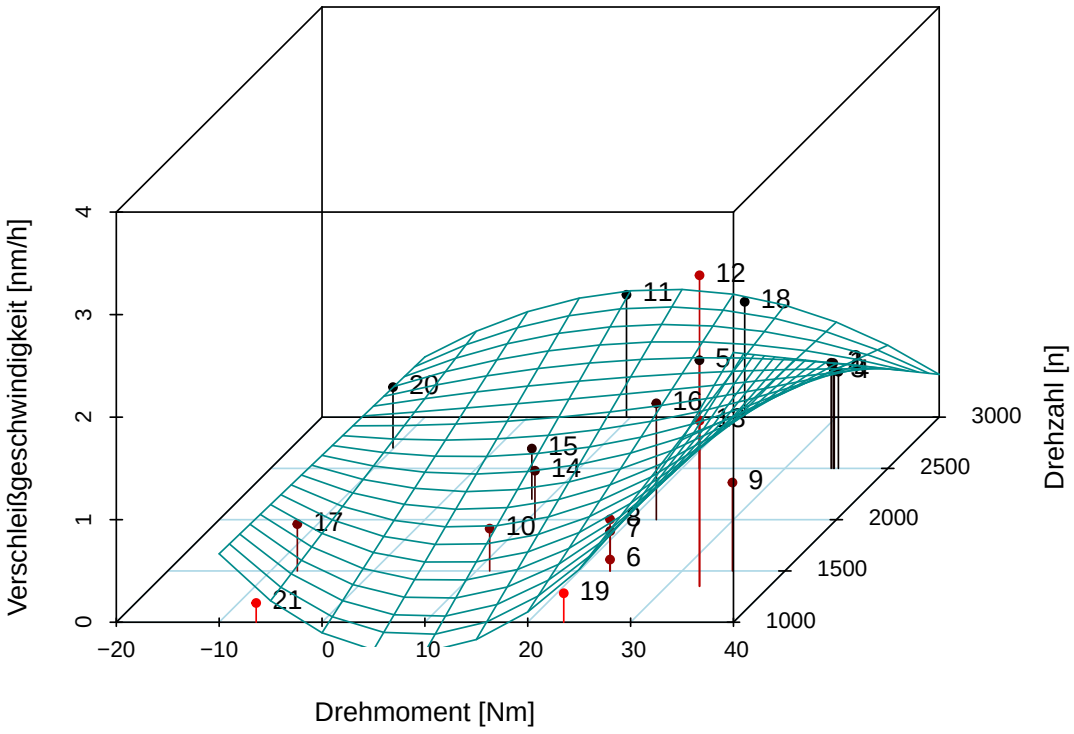

Abbildung 5.5: Graphische Darstellung der Regressionskurve des Modells `m1`.

Die gefitteten Werte in Abbildung 5.5 sind durch

$$\hat{\mu} = -0.6697 - 0.08814 \cdot dm + 0.00057 \cdot dz + 0.004958 \cdot (dm)^2 + 4.014 \cdot 10^{-5} \cdot dm \cdot dz - 2.049 \cdot 10^{-6} \cdot (dm)^2 \cdot dz$$

gegeben.

In Abbildung 5.5 erkennt man auf den ersten Blick, dass die höchste geschätzte Verschleißgeschwindigkeit mit einem hohen Drehmoment (40 Nm) und einer niedrigen Drehzahl (1000 n) erzielt wird. Weiters nimmt die geschätzte Verschleißgeschwindigkeit für eine hohe Drehzahl (3000 n) mit wachsendem Drehmoment zu. Nachdem der Hochpunkt bei ungefähr 20 Nm erreicht wird, kommt es anschließend zu einer Geschwindigkeitsreduktion. Sehr auffällig sind die negativen gefitteten Werte im Bereich 0 bis 15 Nm und 1000 n. Die geschätzte Verschleißgeschwindigkeit sollte für jede Motoreinstellung einen positiven Wert annehmen. Daher konnte das Modell `m1` in dieser Hinsicht nicht überzeugen.

In der Tabelle 5.1 sind einige gefittete Verschleißgeschwindigkeiten und 95%-Konfidenzintervalle für die Erwartungswerte aufgelistet. Diese Werte werden für weiterführende Berechnungen benötigt, wobei diese Problematik nicht im Rahmen der Masterarbeit abgehandelt wird.

```
> dm.grid <- rep(c(-10,25,40),times=3) # 3 x 3 Grid
> dz.grid <- rep(c(1000,2000,3000),each=3)
> m.conf <- predict(m1, new=data.frame(dm.grid, dz.grid),
                    interval="confidence")
```

			Grenze	
Drehmoment	Drehzahl	Fitted Value $\hat{\mu}$	Untere	Obere
−10 Nm	1000 n	0.6674	0.1090	1.2257
25 Nm	1000 n	0.5145	0.3440	0.6850
40 Nm	1000 n	2.6307	1.9177	3.3436
−10 Nm	2000 n	0.6273	0.2130	1.0416
25 Nm	2000 n	0.8035	0.6497	0.9572
40 Nm	2000 n	1.5237	1.1142	1.9331
−10 Nm	3000 n	0.5872	−0.2655	1.4398
25 Nm	3000 n	1.0924	0.6816	1.5031
40 Nm	3000 n	0.4167	−0.5140	1.3474

Tabelle 5.1: Exemplarische gefittete Werte $\hat{\mu}$ aus dem Modell `m1` inklusive 95%-Konfidenzintervalle für die Erwartungswerte μ.

5.2 Gewichtete Gammaregression

Im Abschnitt 5.1 wurde angenommen, dass die Response y normalverteilt ist. Alternativ könnte man die Annahme treffen, dass die abhängige Variable y aus einer rechtsschiefen Verteilung stammt, bei der man zusätzlich die Varianz in Abhängigkeit vom Erwartungswert modellieren kann. Für die gegebene Problemstellung würde sich die Gammaverteilung bestens eignen. Da zusätzlich alle Responsewerte positiv sind und einen positiven Erwartungswert aufweisen müssen, eignet sich der Logarithmus als Linkfunktion.

Im Modell `m1` wurde untersucht, ob die Drehzahl in Interaktion mit dem linearen und quadratischen Drehmomentterm einen Einfluss auf die Verschleißrate hat. Das nachfolgende Modell `g1` überprüft, unter der Annahme der Gammaverteilung, ob sich die Verschleißrate durch die gleiche Linearkombination erklären lässt. Das betrachtete Modell ist somit durch die Linkfunktion $g(\mu) = \log(\mu)$ und der systematischen Komponente

$$\eta = \beta_0 + \beta_1 \cdot dm + \beta_2 \cdot dz + \beta_3 \cdot (dm)^2 + \beta_4 \cdot dm \cdot dz + \beta_5 \cdot (dm)^2 \cdot dz$$

gegeben. Da die Responsevariablen heteroskedastisch sind, impliziert das unterschiedliche Dispersionsparameter ϕ_i. Diesem Problem versucht man mittels prior weights entgegenzuwirken. Weil die Standardabweichungen der Responses vorab geschätzt wurden, ist bekannt, wie stark die einzelnen Respones streuen. Somit können diese als Gewichte herangezogen werden. Die prior weights sind somit durch

$$w_i = \frac{1}{sd(y_i)}$$

gegeben und werden in `R` für ein `glm`-Objekt mit dem Parameter `weights=` zugewiesen.

```
> prior.wts <- 1/se
> g1 <- glm(wear.rate ~ dm * dz + I(dm^2) + I(dm^2):dz,
               family = Gamma(link=log) , weights = prior.wts)
```

Nachdem die Daten gefittet worden sind, sollte überprüft werden, wie gut das GLM die Datensituation beschreibt.

```
> deviance(g1)/summary(g1)$dispersion
17.65765
```

Die skalierte Deviance $D^*(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = 17.66$ ist approximativ χ^2_{n-p} verteilt und sollte unter dem korrekten Modell in etwa ihrem Freiheitsgrad von $n-p = 15$ entsprechen.

Da beide Werte ungefähr gleich groß sind, zeugt dies von einer guten Modellanpassung und es spricht nichts gegen die Annahme der Gammaverteilung.

Als erstes wird ein Blick auf das Modell `g1` geworfen, um darin nicht signifikante Parameter mittels Wald-Test zu entfernen.

```
> summary(g1)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.598e+00  4.045e-01  -6.423 1.15e-05 ***
dm          -7.872e-02  3.853e-02  -2.043 0.059029 .
dz           9.049e-04  2.107e-04   4.295 0.000639 ***
I(dm^2)      5.263e-03  1.434e-03   3.670 0.002274 **
dm:dz        3.855e-05  1.970e-05   1.957 0.069255 .
dz:I(dm^2)  -2.195e-06  7.009e-07  -3.132 0.006859 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Gamma family taken to be 2.091069)

    Null deviance: 141.051  on 20  degrees of freedom
Residual deviance:  36.923  on 15  degrees of freedom
AIC: -215.23
```

Man erkennt in der `summary`-Ausgabe, dass alle Parameter von Bedeutung sind und es bietet sich ein ähnliches Bild wie im Modell `m1`. Auch hier konnte der Drehzahlterm in Kombination mit dem Drehmoment an Signifikanz erlangen. Des Weiteren nimmt bei gleichbleibender Drehzahl (1000 n) und Zunahme des Drehmoments von 10 auf 20 Nm die mittlere Verschleißgeschwindigkeit um 0.056 nm/h zu.

Im Anschluss wird das Modell `g1` nun mittels graphischer Diagnoseplots weiter untersucht.

```
> res.dev <- residuals(g1,type="deviance"); fit <- fitted(g1);
> par(mfrow=c(1,3))

> plot(fit,res.dev, xlab="Fitted Values", ylab="Deviance Residuals",
       main = "Deviance Residuals vs. Fitted")
> abline(0, 0,lty = 2)
> lines(lowess(x=fit,y=res.dev), col="red", lwd="1")

> plot(fit, wear.rate, xlab="Fitted Values", ylab="Observed",
       main="Observed vs. Fitted")
```

```
> abline(0,1, col="red")

> qqnorm(res.dev,main="Normal QQ-Plot of Deviance Residuals")
> qqline(res.dev,col = 2)
```

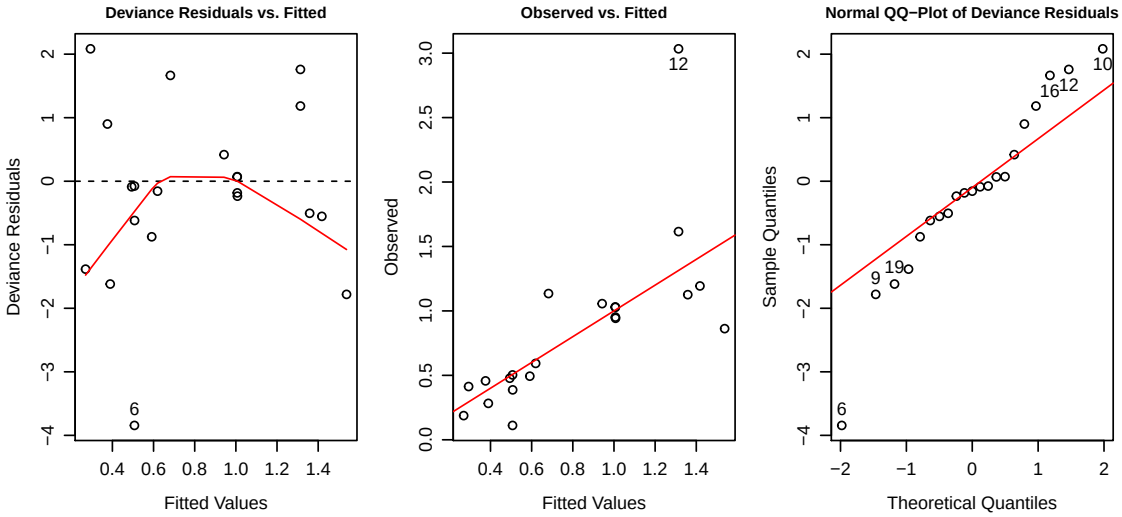

Abbildung 5.6: Diagnoseplots des Modells `g1`.

Laut Abschnitt 3.5 sollten sich die Deviance Residuen unter einem Modell mit guter Anpassungsgüte wie standardnormalverteilte Zufallsvariablen verhalten. Man erkennt, dass im *Deviance Residuals vs. Fitted Values Plot* die Punkte kein Muster bilden und um die Nulllinie streuen. Mit der Nummer 6 gibt es einen sehr starken Ausreißer am unteren Ende. Dieser Ausreißer ist ebenfalls im *Normal QQ-Plot der Deviance Residuen* ersichtlich. Die Deviance Residuen scheinen normalverteilt zu sein, obwohl es leichte Abweichungen am oberen Ende gibt. Außerdem stellt dieser Plot eine starke Verbesserung zu den zuvor betrachteten Normal QQ-Plots dar. Wie im *Observed vs. Fitted Plot* erkennbar, befinden sich alle Punkte in der Nähe der ersten Mediane. Dies zeugt von einer guten Modellanpassung. Einzig die Nummer 12 wird stark unterschätzt.

Da der quadratische Drehzahlterm `I(dz^2)` bis jetzt außer Acht gelassen wurde, wird dieser Term in das zu untersuchende Modell `g2` aufgenommen. Von Interesse ist hierbei, ob ein rein linearer und ein rein quadratischer Interaktionsterm zwischen Drehzahl und Drehmoment einen signifikanten Einfluss auf die Verschleißrate `wear.rate` hat. Das Modell `g2` besteht somit aus der Linkfunktion $g(\mu) = \log(\mu)$ und der systematischen Komponente

$$\eta = \beta_0 + \beta_1 \cdot dm + \beta_2 \cdot dz + \beta_3 \cdot (dm)^2 + \beta_4 \cdot (dz)^2 + \beta_5 \cdot dm \cdot dz + \beta_6 \cdot (dm)^2 \cdot (dz)^2.$$

Die Gammaregression wird mit dem nachfolgenden R-Code gerechnet.

```
> prior.wts <- 1/se
> g2 <- glm(wear.rate ~ dm * dz + I(dm^2)*I(dz^2),
                family = Gamma(link=log) , weights = prior.wts)

> deviance(g2)/summary(g2)$dispersion
17.84054
```

Nachdem die Daten gefittet worden sind, wird die Güte der Modellanpassung überprüft. Die skalierte Deviance $D^*(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = 17.84$ ist unter dem korrekten Modell approximativ χ^2_{n-p} verteilt und sollte ungefähr der Anzahl der Freiheitsgraden $n - p = 14$ entsprechen. Somit spricht nichts gegen die Gammaverteilungsannahme. Im Anschluss wird mittels Parametertests (siehe Abschnitt 3.6) überprüft, ob ein Untermodell von g2 ausreicht.

```
> g2.red <- update(g2, . ~ . - I(dm^2):I(dz^2))
> anova(g2,g2.red, test = "F")

Model 1: wear.rate ~ dm * dz + I(dm^2) * I(dz^2)
Model 2: wear.rate ~ dm + dz + I(dm^2) + I(dz^2) + dm:dz
  Resid. Df Resid. Dev Df Deviance      F  Pr(>F)
1        14     34.317
2        15     56.588 -1  -22.271 11.578 0.00429 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Aus der Analysis of Deviance folgt, dass die quadratische Interaktion ein relevanter Modellparameter ist. Durch die Entfernung des besagten Terms verschlechtert sich die skalierte Deviance. Das Untermodell g2.red ist somit für die vorhandene Datensituation nicht ausreichend.

```
> summary(g2)
```

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -3.307e+00  9.320e-01  -3.549 0.003210 **
dm              -7.981e-02  3.580e-02  -2.229 0.042694 *
dz               1.681e-03  9.906e-04   1.697 0.111816
I(dm^2)          3.405e-03  8.008e-04   4.252 0.000806 ***
I(dz^2)         -1.911e-07  2.472e-07  -0.773 0.452333
dm:dz            3.816e-05  1.828e-05   2.088 0.055582 .
I(dm^2):I(dz^2) -5.764e-10  1.678e-10  -3.434 0.004030 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Gamma family taken to be 1.923533)

    Null deviance: 141.051  on 20  degrees of freedom
Residual deviance:  34.317  on 14  degrees of freedom
AIC: -241.96
```

Nachdem es sich beim betrachteten GLM um ein hierarchisches Modell handelt und beide Interaktionsterme `dm:dz` und `I(dm^2):I(dz^2)` signifikant sind, verbleiben die nicht einflussreichen Drehzahlparameter `dz` und `I(dz^2)` im Modell.

Im Anschluss erfolgt wieder die Untersuchung mittels graphischer Diagnoseplots.

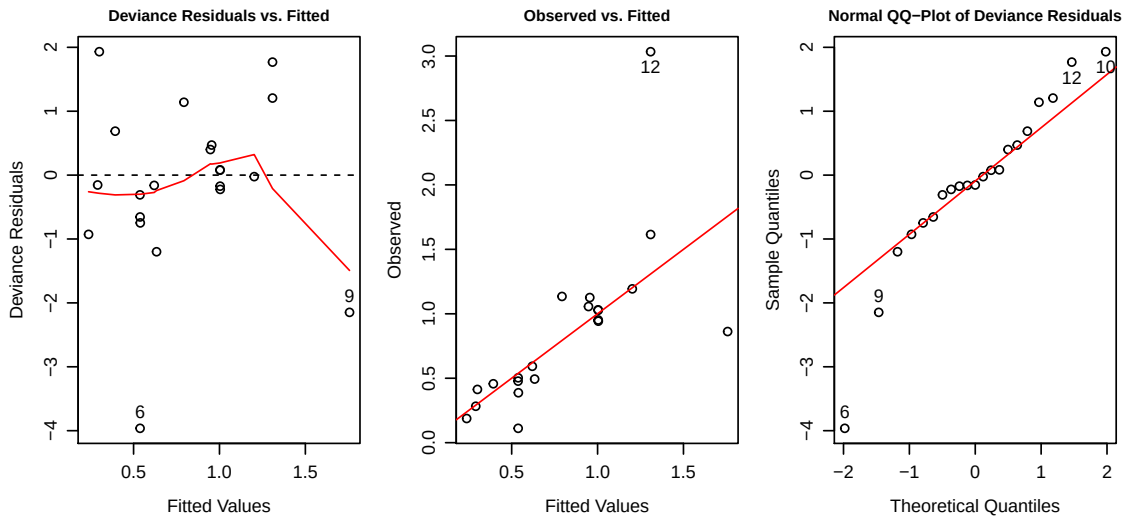

Abbildung 5.7: Diagnoseplots des Modells `g2`.

Im *Deviance Residuals vs. Fitted Values Plot* erkennt man, dass die einzelnen Punkte kein Muster bilden und um die Null streuen. Es gibt bloß einen starken Ausreißer am unteren Ende. Auch hier scheinen die Deviance Residuen normalverteilt zu sein. Wie im *Normal QQ-Plot der Deviance Residuen* ersichtlich, bilden nahezu alle Punkte eine Gerade. Lediglich am unteren Ende scheint mit der Nummer 6 der bekannte Ausreißer auf. Fast alle Punkte befinden sich im *Observed vs. Fitted Plot* in der Nähe der ersten Mediane. Einzig die Nummer 12 wird ebenfalls, wie in all den anderen Modellen, stark unterschätzt.

Im nächsten Schritt wird das bessere der zwei erhobenen GLMs ermittelt. Hierfür verwendet man wieder das Akaike Informationskriterium (siehe Abschnitt 3.6.1).

```
> AIC(g1,g2)
   df       AIC
g1  7 -215.2335
g2  8 -241.9614
```

Da das Modell `g2` den kleineren AIC-Wert aufweist, entscheidet man sich für dieses Modell.

Anschließend wird in Abbildung 5.8 die Regressionskurve des GLMs `g2` geplottet. Die gefitteten Werte sind durch

$$\begin{aligned}\hat{\mu} = \exp\big(&-3.307 - 0.0798 \cdot dm + 0.00168 \cdot dz + 0.003405 \cdot (dm)^2 \\ &- 1.911 \cdot 10^{-7} \cdot (dz)^2 + 3.816 \cdot 10^{-5} \cdot dm \cdot dz \\ &- 5.764 \cdot 10^{-10} \cdot (dm)^2 \cdot (dz)^2\big)\end{aligned}$$

gegeben.

Man erkennt in Abbildung 5.8, dass dank des log-Links alle gefitteten Werte des Modells `g2` positiv sind. Dies ist sicherlich ein Vorteil gegenüber dem Modell `m1`. Die größte geschätzte Verschleißgeschwindigkeit wird auch hier für ein hohes Drehmoment und eine niedrige Drehzahl erzielt. Im Gegensatz zur Abbildung 5.5 steigt hier die Regressionskurve für großes Drehmoment und niedrige Drehzahl viel stärker an. Außerdem erscheint dieser Kurventeil etwas „künstlich“ und ist wohl durch die zwölfte Messung motiviert. Der Kurvenverlauf für hohe Drehzahlen (3000 n) ähnelt jenem aus Abbildung 5.5.

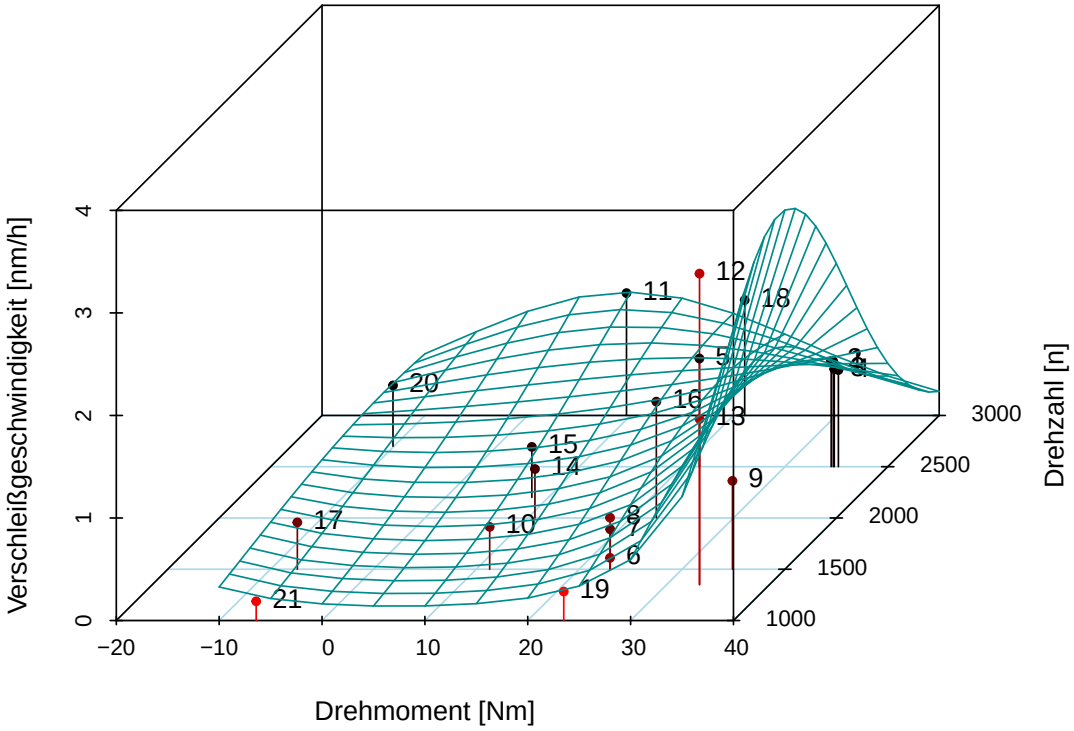

Abbildung 5.8: Graphische Darstellung der Regressionskurve des Modells `g2`.

Im Anschluss werden für einige exemplarische Motoreinstellungen die gefitteten Werte und 95%-Konfidenzintervalle für die Erwartunswerte bestimmt. Die ermittelten Werte sind in der Tabelle 5.2 aufgelistet.

```
> ki.pred <- predict(g2, newdata= data.frame(dm=dm.grid,
  dz=dz.grid), type = "response", se.fit = T)
> g.fit <- ki.pred$fit
> g.oben <- g.fit + qnorm(0.975)*ki.pred$se.fit
> g.unten <- g.fit - qnorm(0.975)*ki.pred$se.fit
```

Vergleicht man die gefitteten Werte der Tabellen 5.1 (Modell `m1`) und 5.2 (Modell `g2`), so erkennt man, dass sich diese für zahlreiche Motoreinstellungen sehr unterscheiden. Die ersten beiden gefitteten Werte des Modells `g2` aus der Tabelle 5.2 (0.3269 und 0.3358) sind nur halb so groß wie die geschätzten Werte (0.6674 und 0.5145) für vergleichbare Motoreinstellungen des Modells `m1`.

			Grenze	
Drehmoment	Drehzahl	Fitted Value $\hat{\mu}$	Untere	Obere
−10 Nm	1000 n	0.3269	0.1012	0.5526
25 Nm	1000 n	0.3358	0.1944	0.4773
40 Nm	1000 n	2.8336	0.5465	5.1208
−10 Nm	2000 n	0.5684	0.2707	0.8660
25 Nm	2000 n	0.8957	0.6689	1.1225
40 Nm	2000 n	2.4820	1.1834	3.7805
−10 Nm	3000 n	0.6008	0.0512	1.1505
25 Nm	3000 n	0.7930	0.3540	1.2320
40 Nm	3000 n	0.2345	−0.0738	0.5429

Tabelle 5.2: Exemplarische gefittete Werte $\hat{\mu}$ aus dem Modell `g2` inklusive 95%-Konfidenzintervalle für die Erwartungswerte μ.

Bei Betrachtung der beiden Abbildungen 5.5 und 5.8 sieht man, dass die größte Verschleißgeschwindigkeit bei einem hohen Drehmoment und niedriger Drehzahl erzielt wird. Die Regressionskurve des Modells `g2` nimmt jedoch für ein hohes Drehmoment mit steigender Drehzahl viel langsamer ab als die Kurve des Modells `m1`. Daher unterscheiden sich die gefitteten Werte 2.4820 (Modell `g2`) und 1.5237 (Modell `m1`) für die Motoreinstellung 40 Nm und 2500 n so stark.

Ein weiter Nachteil des Modells `m1` ist, dass die gefitteten Werte im Bereich 0 bis 15 Nm für eine niedrige Drehzahl (1000 n) kurz negativ sind. Durch die Wahl des Logarithmus als Linkfunktion, werden positive Erwartungswerte der Responsevariablen im Modell `g2` garantiert. Somit entsprechen die geschätzten Werte des Modells `g2` eher der Realität.

5.3 Reduzierter Datenumfang

Weil das GLM `g2` den kleinsten AIC-Wert unter den zuvor untersuchten Modellen aufweist, wird eine Gammaregression mit den gleichen Prädiktoren und Log-Link für eine reduzierte Datensituation gerechnet. In Abbildung 5.7 weist die sechste Beobachtung ein großes Deviance-Residuum auf, welches womöglich das Endresultat verfälschen könnte. Außerdem erkennt man in Abbildung 4.9, dass die Verschleißrate und der dazugehörige Standardfehler des zwölften Messvorgangs viel größer sind, als bei den restlichen Messvorgängen. Dies spiegelt sich auch in den *Observed vs. Fitted Plots* aller untersuchten Regressionsmodelle wider. Man erkennt, dass die zwölfte Verschleißrate immer sehr stark unterschätzt wird. Daher werden diese

beiden Datenzeilen nun aus der Analyse ausgeschlossen.

```
# Entferne die Datenzeilen
> df <- df[c(-6,-12),]
> attach(df)
> prior.weights <- 1/se


> g2.small <- glm(wear.rate ~ dm * dz + I(dm^2)*I(dz^2),
          family = Gamma(link=log) , weights = prior.weights)

> deviance(g2.small)/summary(g2.small)$dispersion
12.46875
```

Die skalierte Deviance $D^*(\boldsymbol{y}, \hat{\boldsymbol{\mu}})$ beträgt 12.47 und sollte unter dem korrekten Modell in etwa dem Freiheitsgrad von $n - p = 12$ entsprechen. Nach der Entfernung der störenden Beobachtungen stimmen die beiden Werte nahezu überein. Dies zeugt von einer sehr guten Modellanpassung.

```
> summary(g2.small)

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -3.344e+00  7.386e-01  -4.527 0.000693 ***
dm              -7.084e-02  2.849e-02  -2.487 0.028606 *
dz               1.777e-03  7.855e-04   2.262 0.043073 *
I(dm^2)          3.094e-03  6.443e-04   4.802 0.000432 ***
I(dz^2)         -2.227e-07  1.961e-07  -1.135 0.278464
dm:dz            3.518e-05  1.450e-05   2.425 0.032001 *
I(dm^2):I(dz^2) -5.339e-10  1.340e-10  -3.984 0.001813 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Gamma family taken to be 1.205226)

    Null deviance: 108.705  on 18  degrees of freedom
Residual deviance:  15.028  on 12  degrees of freedom
AIC: -508.96
```

Ausgehend vom Modell `g2.small` versucht man wieder mittels Wald-Test einzelne Prädiktionsparameter zu eliminieren. Man erkennt im `summary`-output, dass mit

den gewählten prior weights nur der quadratische Drehzahlterm `I(dz^2)` nicht signifikant ist. Da jedoch hierarchische Modelle betrachtet werden und der quadratische Interaktionsterm `I(dm^2):I(dz^2)` einflussreich ist, verbleibt der Parameter `I(dz^2)` im Modell.

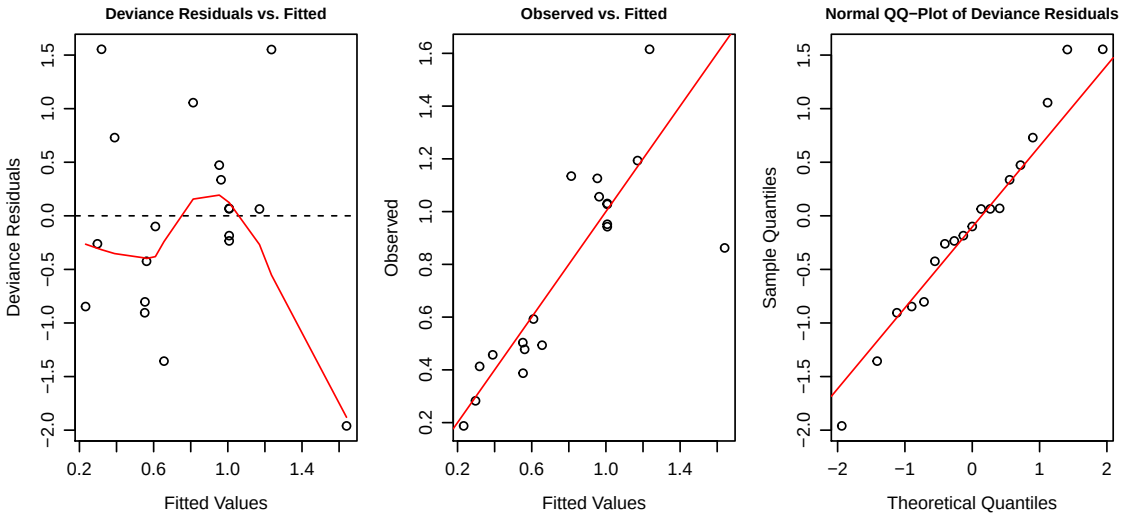

Abbildung 5.9: Diagnoseplots des Modells `g2.small`.

Die Entfernung der beiden Datenzeilen liefert ein stark verändertes Bild. Wie im *Deviance Residuals vs. Fitted Values Plot* erkennbar, streuen die Deviance Residuen um die Null und es gibt keine groben Ausreißer. Im *Observed vs. Fitted Values Plot* sieht man, dass sich fast alle Punkte in der Nähe der ersten Mediane befinden. Die Deviance-Residuen im *Normal QQ-Plot der Deviance Residuen* scheinen nun auch eindeutiger normalverteilt zu sein. Abschließend wird noch die Regressionskurve geplottet.

Auf den ersten Blick sind sich die beiden Regressionskurven der Modelle `g2.small` und `g2` sehr ähnlich. Der parabelförmige „künstliche“ Kurvenabschnitt ist anscheinend nicht nur durch die zwölfte Messung motiviert. Die Entfernung der beiden störenden Datenzeilen hat sich somit nicht sehr stark auf die Gestalt der Regressionskurve ausgewirkt.

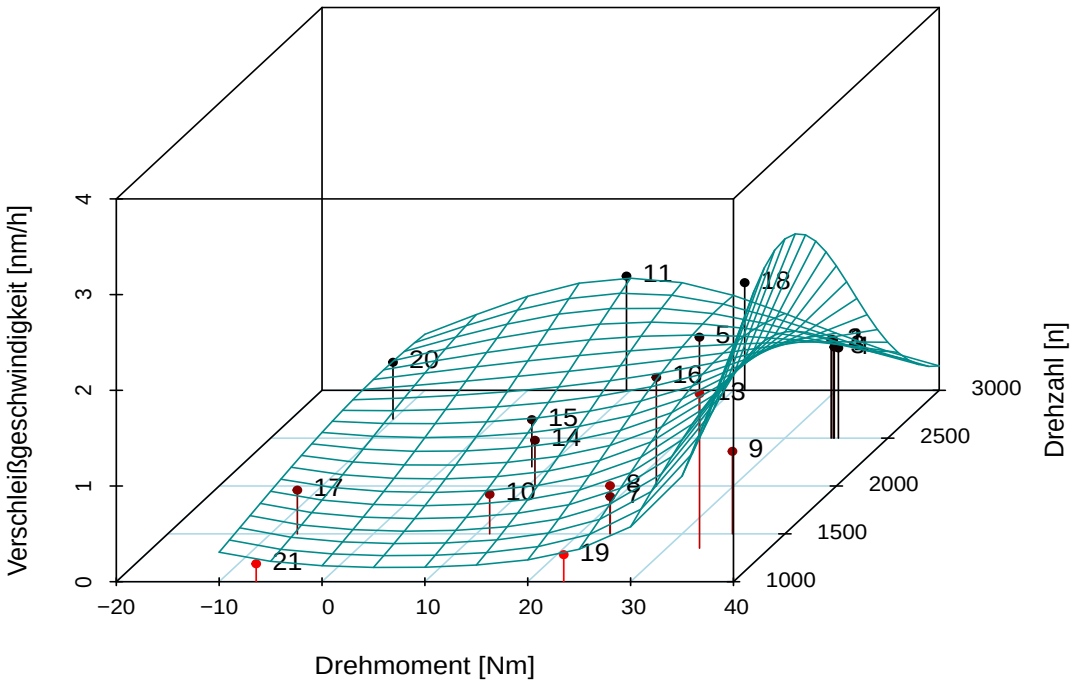

Abbildung 5.10: Graphische Darstellung der Regressionskurve des Modells `g2.small`.

Im Anschluss werden noch gefittete Werte und 95%-Konfidenzintervalle der Erwartungswerte μ für bestimmte Motoreinstellungen ermittelt.

Wie in Tabelle 5.3 ersichtlich, erhält man nach wie vor die größte geschätzte Verschleißgeschwindigkeit für ein großes Drehmoment und eine niedrige Drehzahl. Vergleicht man die Werte der Tabellen 5.2 (Modell `g2`) und 5.3 (Modell `g2.small`), so erkennt man, dass alle gefitteten Werte nahezu übereinstimmen. Einen Unterschied liefert nur die Motoreinstellung: Drehmoment 40 Nm und Drehzahl 1000 n. Hierbei ist die geschätzte Verschleißgeschwindigkeit des Modells `g2` um ungefähr 0.40 nm/h größer als die Modells `g2.small`. Die Entfernung der zwölften Beobachtung hat sich zumindest auf die Höhe des „künstlichen“ Kurvenabschnittes ausgewirkt.

Ein Vergleich der Breiten der Konfidenzintervalle in den Tabellen 5.2 und 5.3 verdeutlicht, dass für jede Motoreinstellung die Konfidenzintervalle der gefitteten Werte des zuletzt betrachteten Modells kleiner sind. Aufgrund der kleineren Breite ist die Schätzung $\hat{\mu}$ präziser. Dies lässt sich durch die sehr gute Modellanpassung von `g2.small` erklären.

			Grenze	
Drehmoment	Drehzahl	Fitted Value $\hat{\mu}$	Untere	Obere
−10 Nm	1000 n	0.3082	0.1394	0.4769
25 Nm	1000 n	0.3392	0.2257	0.4527
40 Nm	1000 n	2.4104	0.8287	3.9921
−10 Nm	2000 n	0.5597	0.3275	0.7918
25 Nm	2000 n	0.9101	0.7256	1.0945
40 Nm	2000 n	2.3000	1.3323	3.2677
−10 Nm	3000 n	0.5852	0.1612	1.0093
25 Nm	3000 n	0.8026	0.4502	1.1551
40 Nm	3000 n	0.2547	−0.0117	0.5211

Tabelle 5.3: Exemplarische gefittete Werte $\hat{\mu}$ aus dem Modell `g2.small` und 95%-Konfidenzintervalle für die Erwartungswerte μ.

KAPITEL 6

RÜCKBLICK

In dieser Masterarbeit wird das Problem der Heteroskedastizität vorgestellt. Ausgangspunkt dafür ist das gewöhnliche lineare Modell. Dabei werden zuerst die Auswirkungen untersucht, die durch die Verletzung der Modellannahme der konstanten Varianz entstehen. Als größter Problempunkt erweist sich die Verzerrung der Varianz-Kovarianz-Matrix des kleinsten Quadrate Schätzers (KQ-Schätzer). Infolgedessen werden die robusten Standardfehler HC0 bis HC3 vorgestellt, die im Falle vorherrschender Heteroskedastizität dem gewöhnlichen Standardfehler vorgezogen werden sollten. Als Erweiterung des gewöhnlichen linearen Modells werden das gewichtete lineare Modell und seine Komponenten vorgestellt. Es wird festgehalten, dass im Falle bekannter Gewichte bei vorherrschender Heteroskedastizität der gewichtete KQ-Schätzer der beste lineare unverzerrte erwartungstreue Schätzer ist (engl. BLUE). Da jedoch Gewichte nicht für jede Problemstellung vorab gegeben sind, werden zwei Techniken präsentiert, um diese aus den vorhandenen Daten zu schätzen. Abschließend wird eine Monte Carlo Simulationsstudie durchgeführt, in der die Varianz-Kovarianz-Matrizen der jeweiligen KQ-Schätzer bei Heteroskedastizität ermittelt werden, um die zuvor bestimmten theoretischen Ergebnisse zu veranschaulichen.

Weil die Normalverteilungsannahme der Responsevariablen nicht für jede Datensituation passend ist, werden zusätzlich generalisierte lineare Modelle (GLM) behandelt. Hierbei ist es ausreichend, dass die Verteilung der Response aus der linearen Exponentialfamilie stammt. Die Problematik der Heteroskedastizität wird auch für GLMs betrachtet. Mit Hilfe passender prior-weights kann den Auswirkungen heteroskedastischer Responses entgegen gewirkt werden. Problemstellungen, die positive Responsevariablen verlangen und in der zusätzlich die Varianzstruktur Abhängig-

keiten vom Erwartungswert aufweisen, können mittels Gammaregression analysiert werden. Da die Gammaverteilung ein Mitglied der linearen Exponentialfamilie ist, wird dieser Spezialfall der GLMs in dieser Masterarbeit diskutiert. Abschließend wird in einer Monte Carlo Simulationsstudie aufgezeigt, wie sich im Falle der Gammaregression die Wahl der prior-weights auf die Varianz-Kovarianz-Matrix des Maximum-Likelihood-Schätzers auswirkt.

Der praktische Abschnitt der Masterarbeit beschäftigt sich mit der Analyse von Verschleißdaten, die im Kolbenring eines Einzylinder-Forschungsmotors gemessen wurden. Den Ausgangspunkt stellen hierbei Verschleißbeträge dar, die für bestimmte Drehmoment-Drehzahl Einstellungen des Motors über die Zeit erhoben wurden. Davon ausgehend werden zuerst einfache lineare Regressionsmodelle gerechnet, um den linearen Anstieg des Verschleißbetrages über die Zeit zu schätzen. Dadurch erhält man 21 geschätzte Verschleißgeschwindigkeiten und deren Standardfehler. Mit den gewonnenen Ergebnissen soll versucht werden die Verschleißgeschwindigkeit durch die Drehzahl und das Drehmoment zu erklären. Da die Verschleißgeschwindigkeiten unterschiedlich starke Streuung aufweisen, ist die Responsevariable der behandelten Regressionsmodelle heteroskedastisch. Um der Problematik der Heteroskedastizität entgegenzuwirken, werden zwei gewichtete lineare Regressionsmodelle und zwei gewichtete Gammaregressionen gerechnet. Unter den vier betrachteten gewichteten Regressionsmodellen kann das Modell `g2` überzeugen. Die systematische Komponente dieses Modells besteht aus einem linearen und quadratischen Drehzahl-/Drehmomentterm sowie einer rein linearen und rein quadratischen Interaktion. Es wird zusätzlich der Logarithmus als Linkfunktion gewählt, um positive Erwartungswerte zu garantieren. Abschließend werden noch 95%-Konfidenzintervalle für die Erwartungswerte bestimmt.

ANHANG A

ALLGEMEINES

A.1 Matrixalgebra und Mathematische Statistik

Die Sätze A.1 und A.2 des Anhangs stammen aus Fahrmeir et al. (2009, S. 455 ff.).

Satz A.1. Sei $\boldsymbol{A}$ positiv definit. Dann gilt

1. $\boldsymbol{A}$ ist regulär (und damit invertierbar).
2. $\boldsymbol{A}^{-1}$ ist positiv definit.

Satz A.2. Sei $\boldsymbol{A}$ eine $n \times n$ und $\boldsymbol{B}$ eine $n \times p$ Matrix dann gilt: Ist $\boldsymbol{A}$ positiv definit und hat $\boldsymbol{B}$ vollen Spaltenrang, so ist auch $\boldsymbol{B}'\boldsymbol{A}\boldsymbol{B}$ positiv definit.

Die nachfolgenden Definitionen und Sätze stammen aus Casella und Berger (2001).

Definition A.3. Sei T ein Punktschätzer für einen Parameter θ. Der Bias (Grad der Verzerrung) des Schätzers ist die erwartete Differenz zwischen dem Punktschätzer und seinem zu schätzendem Parameter θ, d.h.

$$\text{bias}(T, \theta) = \mathbb{E}(T - \theta).$$

Ein Schätzer, dessen Bias gleich Null ist, nennt man **erwartungstreu** oder **unverzerrt**.

Definition A.4. Eine Folge von Schätzer $(T_n)_{n\in\mathbb{N}}$ heißt konsistente Schätzfolge für den Parameter θ, wenn für alle $\epsilon > 0$ und jedes $\theta \in \Theta$ gilt

$$\lim_{n\to\infty} P(|T_n - \theta| < \epsilon) = 1.$$

Somit konvergiert die konsistente Folge von Schätzer T_n in Wahrscheinlichkeit gegen den Parameter θ.

Satz A.5. (**Cramér-Rao Ungleichung**)
Gegeben sei eine Stichprobe $\boldsymbol{X}$ mit gemeinsamer Dichte $f(\boldsymbol{x}|\theta)$ und sei T ein Schätzer der Stichprobe mit endlicher Varianz, der die Vertauschungsrelation

$$\frac{\partial}{\partial\theta}\int T(\boldsymbol{x})\cdot f(\boldsymbol{x}|\theta)\,\mathrm{d}\boldsymbol{x} = \int T(\boldsymbol{x})\cdot\frac{\partial}{\partial\theta}f(\boldsymbol{x}|\theta)\,\mathrm{d}\boldsymbol{x}$$

erfüllt. Dann gilt

$$\mathrm{Var}(T(\mathbf{X})) \geq \frac{\left(\frac{\partial}{\partial\theta}\mathbb{E}[T(\mathbf{X})]\right)^2}{\mathcal{J}(\theta)},$$

wobei $\mathcal{J}(\theta) := \mathbb{E}\left[(\frac{\partial}{\partial\theta}\log f(\boldsymbol{x}|\theta))^2\right]$ die Fisher-Information ist.

Definition A.6. Eine Folge von Schätzer $(T_n)_{n\in\mathbb{N}}$ heißt asymptotisch erwartungstreu für den Parameter θ, wenn für alle Werte von θ gilt:

$$\lim_{n\to\infty}\mathbb{E}(T_n) = \theta.$$

Definition A.7. Eine Folge von Schätzer $(T_n)_{n\in\mathbb{N}}$ ist asymptotisch effizient für einen Parameter $\tau(\theta)$, falls

$$\sqrt{n}\,(T_n - \tau(\theta)) \xrightarrow{D} N(0, v_1(\theta))$$

mit

$$v_1(\theta) = \frac{(\tau'(\theta))^2}{\mathcal{J}(\theta)},$$

wobei $\mathcal{J}(\theta)$ die Fisher-Information einer Beobachtung ist. Somit erreicht die asymptotische Varianz der Folge von Schätzer T_n die Cramér-Rao Schranke.

Wooldridge (2005, Kap. 3) formulierte den Satz von Gauß-Markow und dessen Voraussetzungen folgendermaßen:

GM.1: Linear in den Parametern
Gegeben sei eine Stichprobe vom Umfang n, dessen Daten als p-dimensionales Tupel der Form $(y_i, x_{i1}, \ldots, x_{ip-1})$ für $i = 1, \ldots, n$ vorliegen. Das dazugehörige lineare Modell lautet

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1}x_{ip-1} + \epsilon_i,$$

wobei y_i die i-te Responsevariable, $\boldsymbol{x}_i' = (1, x_{i1}, \ldots, x_{ip-1})'$ die bekannten Prädiktoren, $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{p-1})'$ die unbekannten Parameter und $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)'$ der nicht beobachtbare statistische Fehler sind.

GM.2: Keine Abhängigkeit
In der Stichprobe gibt es keine exakte lineare Abhängigkeit zwischen den Prädiktoren.

GM.3: Erwartung Null
Der nicht beobachtbare statistische Fehler $\boldsymbol{\epsilon}$ besitzt Erwartung Null, d.h.

$$\mathbb{E}(\boldsymbol{\epsilon}) = \mathbf{0}.$$

GM.4: Homoskedastizität
Der nicht beobachtbare statistische Fehler $\boldsymbol{\epsilon}$ besitzt konstante Varianz, d.h.

$$\mathrm{Var}(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{I}_n.$$

Satz A.8. (Erwartungstreue)
Seien die Gauß-Markow-Annahmen GM.1 – GM.3 erfüllt, dann ist der Kleinste Quadrate Schätzer $\hat{\boldsymbol{\beta}}$ ein erwartungstreuer Schätzer.

Satz A.9. (Konsistenz)
Seien die Gauß-Markow-Annahmen GM.1 – GM.3 erfüllt, dann ist der Kleinste Quadrate Schätzer $\hat{\boldsymbol{\beta}}$ ein konsistenter Schätzer, d.h. es gilt

$$\hat{\boldsymbol{\beta}} \xrightarrow{P} \boldsymbol{\beta}.$$

Satz A.10. (**Satz von Gauß-Markow**)
Seien die Gauß-Markow-Annahmen GM.1 – GM.4 erfüllt, dann ist der beste lineare erwartungstreue Schätzer (engl. BLUE für Best Linear Unbiased Estimator) durch den Kleinsten-Quadrate-Schätzer gegeben.

Die Beweise der Sätze A.8 – A.10 sind in Wooldridge (2005) auffindbar.

A.2 Auswertung der Messergebnisse

Dieser Abschnitt beinhaltet einige grundlegende `R`-Funktionen und Resultate, die in Kapitel 4 benötigt werden. Mit der nachfolgenden `R`-Funktion `readFile` kann ein bereinigter Datensatz eingelesen werden, wobei der Funktionsparameter `i` für den Index des Dateinamens des i-ten Messverfahrens steht.

```
> dateinamen <- c("File_NR1.csv","File_NR2.csv","File_NR3.csv",
 "File_NR4.csv","File_NR5.csv","File_NR6.csv","File_NR7.csv",
 "File_NR8.csv","File_NR9.csv","File_NR10.csv","File_NR11.csv",
 "File_NR12.csv","File_NR13.csv","File_NR14.csv","File_NR15.csv",
 "File_NR16.csv","File_NR17.csv","File_NR18.csv","File_NR19.csv",
 "File_NR20.csv","File_NR21.csv");
```

```
> readFile <- function(i){
   zeit_einheit <- 1/120
   akt_datei <- read.table(paste("Messwerte/",dateinamen[i],sep=""),
      sep = ";", header = FALSE, dec=".",
      colClasses = c("NULL",rep("numeric", 3)))

  akt_datei <- akt_datei[complete.cases(akt_datei), ]
  n <- length(akt_datei$V2)
  akt_datei[,4] <- seq(from=zeit_einheit,to=n*zeit_einheit,by=zeit_einheit)
  colnames(akt_datei) <- c("verschleiss","drehzahl","drehmoment","zeit")
  return (akt_datei)
}
```

Der anschließende R-Code berechnet ein einfaches lineares Regressionsmodell anhand der Verschleißbeträge des jeweiligen Messvorganges. Dadurch können der Steigungskoeffizient $\hat{\beta}_1$ und dessen Standardfehler ermittelt werden. Falls Heteroskedastizität vorherrscht, wird der robuste Standardfehler „HC3“ bestimmt.

```
> df <- data.frame(Steigung=numeric(),SE=numeric(),p_Wert=numeric())
> for(i in 1:21){
   akt_df <- readFile(i)
   mod <- lm(akt_df$verschleiss ~ akt_df$zeit);
   df[i,1] <- mod$coefficients[2]
   df[i,2] <- sqrt(diag(vcov(mod)))[2]
   p.val <- bptest(mod)$p.value
   df[i,3] <- p.val
   if(p.val < 0.05){
      df[i,2] <- sqrt(diag(vcovHC(mod,type = "HC3")))[2]
   }
 }
```

In den Abbildungen A.1 bis A.3 sind Scatterplots ersichtlich, in denen der gemessene Verschleißbetrag gegen die Zeit aufgetragen wird. Zusätzlich ist für jede Motoreinstellung die Regressionsgerade eingezeichnet.

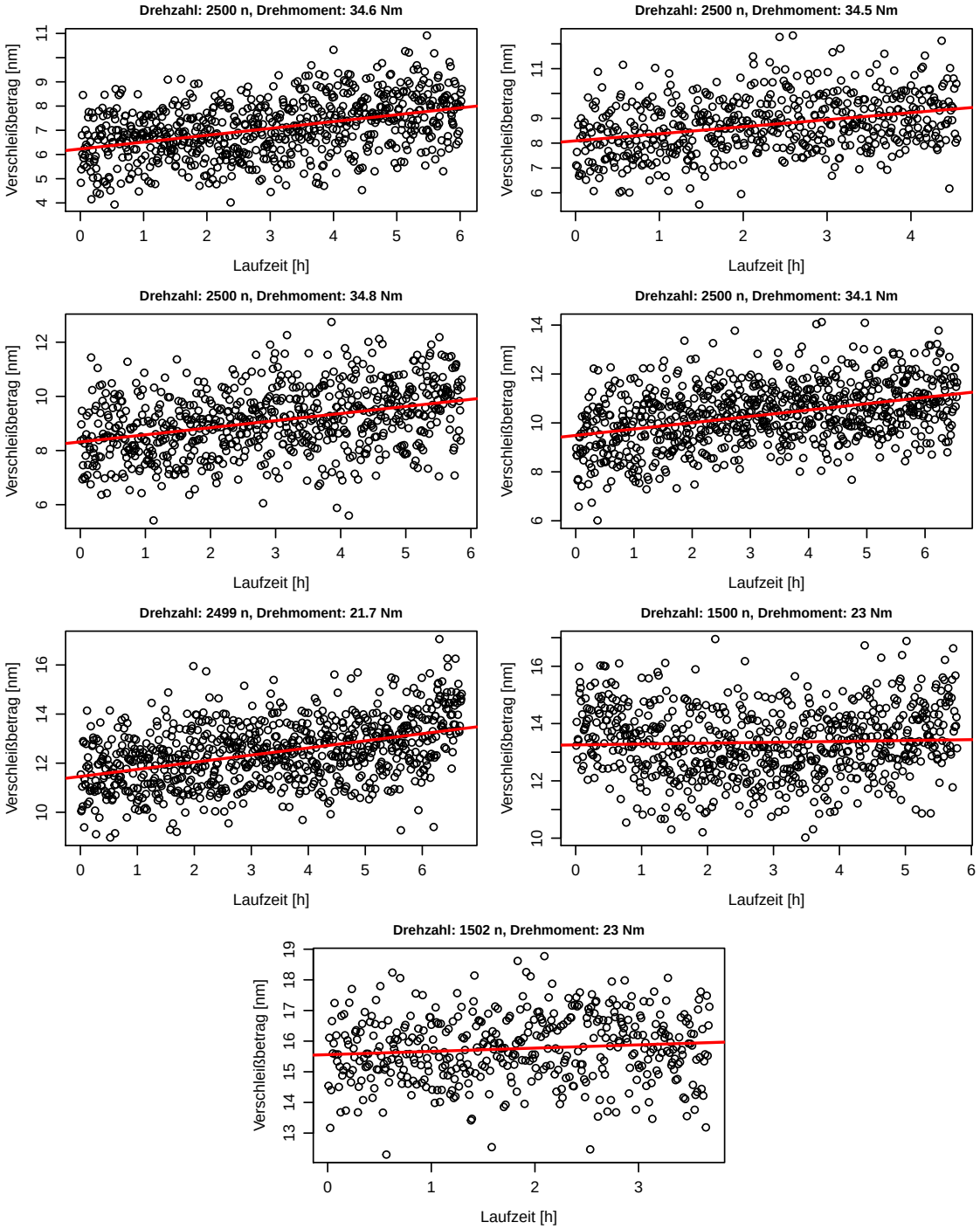

Abbildung A.1: Gemessene Verschleißbeträge der Messvorgänge $1-7$ werden über die Zeit aufgetragen. Die rote Linie entspricht der Regressionsgeraden.

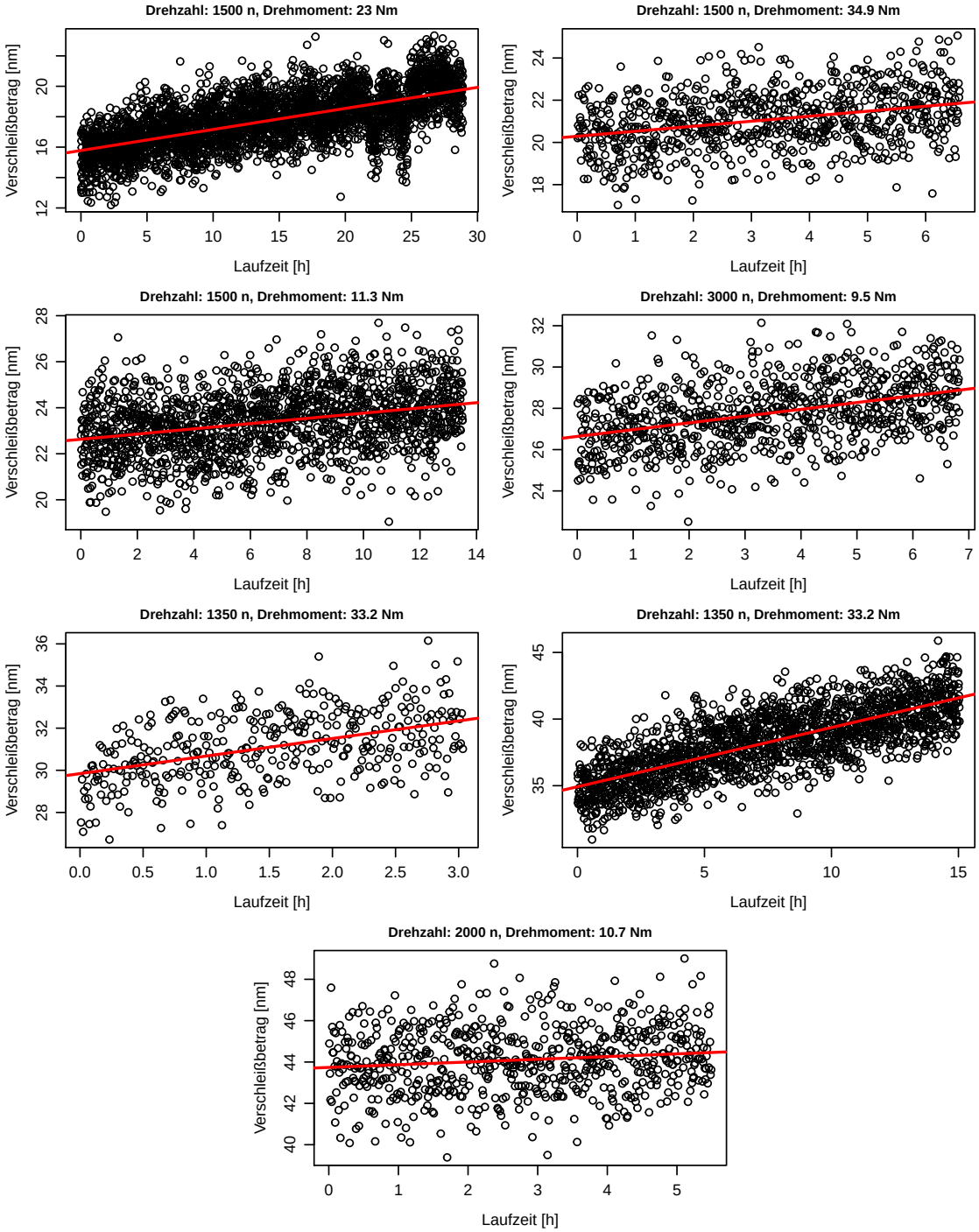

Abbildung A.2: Gemessene Verschleißbeträge der Messvorgänge 8 – 14 werden über die Zeit aufgetragen. Die rote Linie entspricht der Regressionsgeraden.

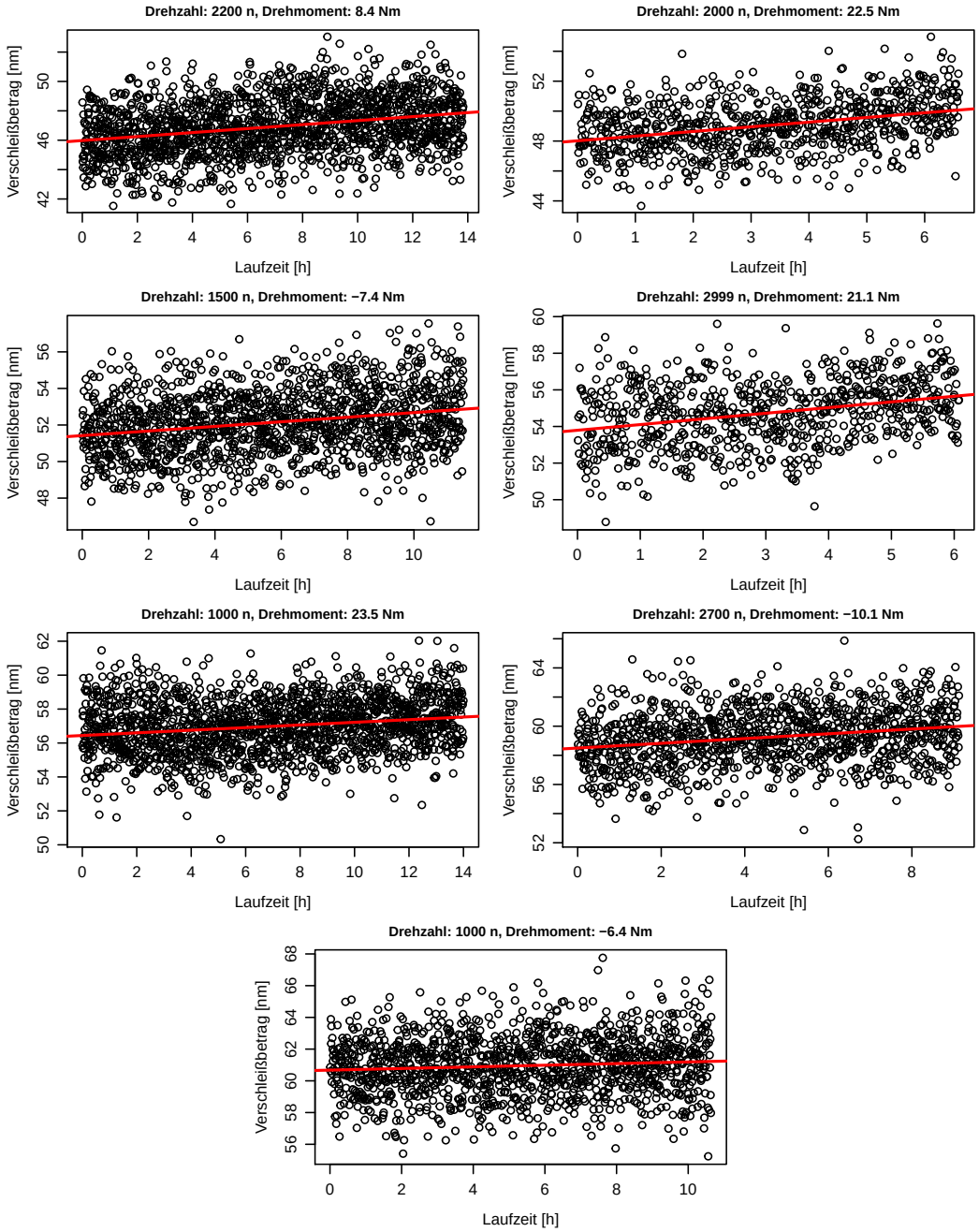

Abbildung A.3: Gemessene Verschleißbeträge der Messvorgänge $15 - 21$ werden über die Zeit aufgetragen. Die rote Linie entspricht der Regressionsgeraden.

In Tabelle A.1 sind alle Steigungskoeffizienten und deren Standardfehler aufgelistet. Wie im zuvor gezeigten `R`-Code ersichtlich, werden die rohen Residuen jedes Modells mittels des Breusch-Pagan-Tests auf Heteroskedastizität überprüft. Der Hypothesentest verwirft H_0 (Homoskedastizität) für die Modelle M8, M18 und M19 mit p-Werten von 0.002781, 0.035786 und 0.009733. In jenen Fällen wurde der robuste Standardfehler HC3 bestimmt.

Nummer	Steigung	Standardfehler	Nummer	Steigung	Standardfehler
M1	0.2826	0.0228	M12	0.8342	0.0839
M2	0.2835	0.0351	M13	0.4442	0.0070
M3	0.2618	0.0252	M14	0.1314	0.0382
M4	0.2593	0.0216	M15	0.1357	0.0110
M5	0.2905	0.0212	M16	0.3120	0.0307
M6	0.0306	0.0274	M17	0.1257	0.0138
M7	0.1065	0.0482	M18	0.3096	0.0358
M8	0.1384	0.0028	M19	0.0777	0.0097
M9	0.2371	0.0241	M20	0.1630	0.0214
M10	0.1136	0.0083	M21	0.0516	0.0168
M11	0.3282	0.0252			

Tabelle A.1: Übersicht der Steigungskoeffizienten $\hat{\beta}_1$ inklusive Standardfehler $se(\hat{\beta}_1)$.

LITERATURVERZEICHNIS

Baltagi, B. H. (2008). *Econometrics* (Fourth Edition). Springer-Verlag, Berlin, Heidelberg.

Breusch, T. S. und Pagan, A. R. (1979). A simple test for heteroscedasticity and random coefficient variation. *Econometrica: Journal of the Econometric Society*, *47*, 1287–1294.

Casella, G. und Berger, R. L. (2001). *Statistical Inference* (Second Edition). Duxbury Press, Pacific Grove.

Czichos, H. und Habig, K.-H. (2015). *Tribologie-Handbuch* (Vierte Auflage). Springer, Wiesbaden.

Dobson, A. (2002). *An Introduction to Generalized Linear Models* (Second Edition). Chapman & Hall/CRC, Boca Raton, Florida.

Fahrmeir, L. und Kaufmann, H. (1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *The Annals of Statistics*, *13*, 342–368.

Fahrmeir, L., Kneib, T. und Lang, S. (2009). *Regression: Modelle, Methoden und Anwendungen* (Zweite Auflage). Springer–Verlag, Berlin, Heidelberg.

Greene, W. (2002). *Econometric Analysis* (Fifth Edition). Prentice Hall, Upper Saddle River, New Jersey.

Hick, H., Walch, S., Edtmayer, J., Lösch, S., Jech, M. und Wopelka, T. (2016). *Implementierung einer kombinierten Verschleiß und Reibanalyse an einem Einzylinder Forschungsmotor.* Paper vorgestellt bei Reibungsminimierung im Antriebsstrang 2016. Esslingen, Germany.

Koenker, R. (1981). A note on studentizing a test for heteroscedasticity. *Journal of Econometrics*, *17*, 107–112.

Ligges, U. und Mächler, M. (2003). Scatterplot3d - an R package for visualizing multivariate data. *Journal of Statistical Software*, *8*, 1–20.

Long, J. S. und Ervin, L. H. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician*, *54*, 217–224.

MacKinnon, J. G. und White, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*, *29*, 305–325.

McCullagh, P. und Nelder, J. A. (1989). *Generalized Linear Models* (Second Edition). Chapman & Hall, London.

Myers, R. H., Montgomery, D. C., Vining, G. G. und Robinson, T. J. (2010). *Generalized Linear Models: with Applications in Engineering and the Sciences* (Second Edition). John Wiley & Sons, Inc., Hoboken, New Jersey.

Nelder, J. A. und Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, *135*, 370–384.

NIST/SEMATECH. (2012). *e-handbook of statistical methods.* Zugriff am 2018.01.18 auf `http://www.itl.nist.gov/div898/handbook/`

van Basshuysen, R. und Schäfer, F. (2017). *Handbuch Verbrennungsmotor Grundlagen, Komponenten, Systeme, Perspektiven* (Achte Auflage). Springer Fachmedien Wiesbaden GmbH, Wiesbaden.

Venables, W. N. und Ripley, B. D. (2012). *Modern Applied Statistics with S* (Fourth Edition). Springer-Verlag New York.

Weisberg, S. (2005). *Applied Linear Regression* (Third Edition). John Wiley & Sons, Inc., Hoboken, New Jersey.

White, H. (1980). A heteroskedasticity–consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, *48*, 817–838.

Wood, S. N. (2006). *Generalized Additive Models: An Introduction with R.* Chapman & Hall/CRC, Boca Raton, Florida.

Wooldridge, J. (2005). *Introductory Econometrics: A Modern Approach* (Third Edition). South-Western College Pub.

Zeileis, A. (2006). Object-oriented computation of sandwich estimators. *Journal of Statistical Software*, *16*, 1–16.

Zeileis, A. und Hothorn, T. (2002). Diagnostic checking in regression relationships. *R News*, *2*, 7–10.