

2. Intervalldaten und generalisierte lineare Modelle

In diesem Kapitel wird die Regression mit Intervalldaten behandelt. Im ersten Abschnitt wird dabei vorerst auf generalisierte lineare Modelle eingegangen. In den Abschnitten 2.4 und 2.5 werden neue Ansätze [Augustin, 2012] für die Bestimmung des Identifizierungsbereichs von Modellparametern in generalisierten linearen Modellen vorgestellt. Zuletzt wird für den Spezialfall der linearen Regression mit skalarer unabhängiger Variable die analytische Lösung besprochen.

2.1. Generalisierte lineare Regression

In diesem Abschnitt werden vorerst einige Aspekte generalisierter linearer Modelle [Nelder und Wedderburn, 1972] behandelt (siehe hierzu auch [McCullagh und Nelder, 1986; Dobson und Barnett, 2008]). Im klassischen linearen Modell wird der Mittelwert μ direkt durch den linearen Prädiktor

$$\eta = \mathbf{x}^T \boldsymbol{\beta}$$

bestimmt. Es gilt also $\mu = \eta$. Außerdem ist der Störterm ϵ um diesen Mittelwert mit fester Varianz normalverteilt (siehe beispielsweise [Fahrmeir et al., 2009]). Das bedeutet, dass die Streuung um die Regressionsgerade unabhängig von x ist. Durch quadratische oder anders transformierte Terme kann man zwar auch im linearen Modell eine nicht-lineare Regressionskurve erzeugen, im Allgemeinen könnte es aber von Interesse sein, nicht nur auf die Normalverteilung beschränkt zu sein. Insbesondere gibt es viele Situationen, in denen beispielsweise negative Werte inhaltlich nicht interpretiert werden können. Ein anderes Beispiel wären diskrete Verteilungen, bei denen $y \in \mathbb{Y} \subset \mathbb{Z}$ gilt, also nur ganze Zahlen annehmen kann, während $x \in \mathbb{R}$ ist. Die Normalverteilung kann in beiden Beispielen die Eigenschaften der Variablen nicht abbilden.

In generalisierten linearen Modellen ist die Einschränkung der Normalverteilung nicht mehr gegeben. Die Voraussetzung an die Verteilung ist lediglich, dass diese zur Exponentialfamilie gehört (siehe beispielsweise [Young und Smith, 2005], Kapitel 5 und [Dobson und Barnett, 2008], Kapitel 3). Diese Voraussetzung wird durch die meisten wichtigen Verteilungen erfüllt.

Eine weitere wichtige Verallgemeinerung in generalisierten linearen Modellen ist, dass der Erwartungswert nicht mehr direkt durch den linearen Prädiktor festgelegt werden muss, sondern durch eine Link-Funktion $g(\mu)$, für die

$$g(\mu) = \eta = \mathbf{x}^T \beta$$

gilt. Löst man nach μ auf, erhält man die Response-Funktion $h(\eta)$ mit

$$\mu = g^{-1}(\eta) = h(\eta).$$

Dadurch lassen sich sehr flexible Zusammenhänge modellieren. Für generalisierte lineare Modelle besteht außerdem eine umfassende Theorie zum Schätzen der Parameter und Testen von Hypothesen. Im Unterschied zum linearen Modell können die Parameter aber im Allgemeinen nicht analytisch bestimmt, sondern müssen mit numerischen Verfahren gesucht werden. Dies liegt vor allem daran, dass die Maximum-Likelihood-Methode eingesetzt wird. Die Berechnung der Parameter ist also aufwendiger und weniger zuverlässig. Es bestehen aber höchst effiziente numerische Verfahren zur Bestimmung der Parameter.

Bei der praktischen Umsetzung in den Kapiteln 3 und 4 werden sich die Erläuterungen der Methoden und die Simulationsbeispiele auf die lineare Regression und die generalisierte lineare Regression mit Exponentialverteilung und log-Link beschränken. Andere generalisierte Modelle können aber analog behandelt werden. Es wäre jedoch zu untersuchen, ob die angewandten Optimierungsverfahren auch dort effizient sind.

Im Folgenden wird das generalisierte lineare Modell mit Exponentialverteilung beschrieben. Dieses stellt einen Spezialfall des Modells mit Gamma-Verteilung dar, wenn der zweite Parameter der Gamma-Verteilung auf 1 gesetzt wird (siehe beispielsweise [Dobson und Barnett, 2008]). Die Dichtefunktion der Exponentialverteilung ist durch

$$f(y; \lambda) = \lambda \exp(-\lambda y) \tag{2.1}$$

gegeben, wobei für den Erwartungswert und die Varianz

$$\mu = \frac{1}{\lambda} \quad \text{und} \quad \sigma^2 = \frac{1}{\lambda^2}$$

gilt. Die Varianz ist also abhängig von dem Erwartungswert. Der natürliche Link (siehe beispielsweise [Dobson und Barnett, 2008], Kapitel 3) ist für die Exponentialverteilung $1/\lambda$. Problematisch ist dabei aber, dass für die Exponentialverteilung nur positive Werte von λ erlaubt sind, also $\lambda > 0$ gelten muss. Mit

$$\mu = \frac{1}{\lambda} = \eta$$

ist dies im Allgemeinen nicht gegeben, da der lineare Prädiktor negativ werden kann. Häufig wird daher eine logarithmische Link-Funktion eingesetzt, für die

$$\log(\mu) = \eta$$

gilt. Diese wird auch im Weiteren beschrieben und für die Beispiele verwendet. Aus dieser Wahl der Link-Funktion folgt

$$\mu = \exp(\eta) \Rightarrow \lambda = \frac{1}{\exp(\eta)}.$$

Für die Likelihood-Funktion gilt mit der Response-Funktion $h(\eta) = \exp(\eta)$

$$\begin{aligned} f(\mathbf{y}; \boldsymbol{\eta}) &= \prod_{i=1}^n \left(\frac{1}{h(\eta_i)} \exp \left(-\frac{y_i}{h(\eta_i)} \right) \right) \\ &= \prod_{i=1}^n \left(\frac{1}{\exp(\eta_i)} \exp \left(-\frac{y_i}{\exp(\eta_i)} \right) \right). \end{aligned} \quad (2.2)$$

Durch Logarithmieren erhält man die log-Likelihood-Funktion

$$\begin{aligned} l(\mathbf{y}; \boldsymbol{\eta}) &= \log(f(\mathbf{y}; \boldsymbol{\eta})) \\ &= \sum_{i=1}^n \left(-\log(\exp(\eta_i)) - \frac{y_i}{\exp(\eta_i)} \right) \\ &= \sum_{i=1}^n \left(-\eta_i - \frac{y_i}{\exp(\eta_i)} \right). \end{aligned} \quad (2.3)$$

Setzt man den linearen Prädiktor ein, ergibt sich

$$l(\mathbf{x}, \mathbf{y}; \boldsymbol{\beta}) = \sum_{i=1}^n \left(-\mathbf{x}_i^T \boldsymbol{\beta} - \frac{y_i}{\exp(\mathbf{x}_i^T \boldsymbol{\beta})} \right). \quad (2.4)$$

Schließlich erhält man durch Ableiten nach den einzelnen Parametern die Score-Funktion

$$\frac{\partial l(\mathbf{x}, \mathbf{y}; \boldsymbol{\beta})}{\partial \beta_k} = \sum_{i=1}^n \left(\frac{x_{ik} y_i}{\exp(\mathbf{x}_i^T \boldsymbol{\beta})} - x_{ik} \right). \quad (2.5)$$

2.2. Regression mit Intervalldaten

Bei Ungenauigkeit in den Daten geht man implizit davon aus, dass es ein theoretisches, exaktes Modell gibt. Unter Umständen kann dieses aber mit den gegebenen

Beobachtungen (auch mit unendlich vielen Daten, also $n \rightarrow \infty$) nicht exakt bestimmt werden. In diesem Fall sind die Parameter des Modells nicht oder nur partiell identifiziert. Mit einem Modell, das tatsächlich Intervalldaten erzeugt, also auch theoretisch keine skalaren Werte liefert, unterscheidet sich die Inferenz deutlich von der im ersten Fall: Die Intervalle sollten im zweiten Fall als zusätzliche Information (bei den unabhängigen Variablen) und als zu schätzende Größe (bei den abhängigen Variablen) betrachtet werden [Diamond, 1990; Ferraro et al., 2010; de A. Lima Neto und de A. T. de Carvalho, 2010; D’Urso und Gastaldi, 2010; Blanco-Fernández et al., 2011]. Ein Identifizierungsbereich ist für letzteren Fall nicht sinnvoll interpretierbar.

Ein Beispiel für den zweiten Fall wäre das natürliche Auftreten einer physikalischen Größe (die unabhängige Variable) an einer bestimmten Position mit einer bestimmten Ausdehnung, die zur Erklärung einer anderen Größe (die abhängige Variable) mit einer bestimmten Position und einer bestimmten Ausdehnung herangezogen wird. Beide Größen bestünden aus Intervallen. Was geschätzt werden müsste, wäre also die Position und Ausdehnung der abhängigen Variablen mit der verfügbaren Information. Die verfügbare Information besteht wiederum aus der Position und Ausdehnung der unabhängigen Variablen. Die Schätzung wäre ein klassisches multivariates Regressionsmodell mit zwei unabhängigen Variablen. Ein lineares Modell könnte mit

$$\begin{aligned} y_{pos} &= \beta_0 + \beta_1 x_{pos} + \beta_2 x_{ext} + \epsilon_{pos}, \\ y_{ext} &= \beta_0 + \beta_1 x_{pos} + \beta_2 x_{ext} + \epsilon_{ext} \end{aligned} \quad (2.6)$$

aufgestellt werden, wobei ϵ_{pos} und ϵ_{ext} die jeweiligen Störterme mit einer bestimmten Verteilung für die Position und die Ausdehnung der abhängigen Variablen sind. Beispielsweise könnte eine Normalverteilung angenommen werden.

Sind die Größen im Modell jedoch theoretisch exakt, können aber nicht beobachtet werden, so ist das Intervall der abhängigen Variablen keine zu schätzende Größe. Dieses ist als Ungenauigkeit der Beobachtung zu interpretieren und damit eine Störung, die eine exakte Schätzung der eigentlich nicht intervallwertigen Größen nicht ermöglicht. Daher muss diese Ungenauigkeit bei der Schätzung beachtet werden [Rohwer und Pötter, 2001; Marino und Palumbo, 2002; Gioia und Lauro, 2005]. Die Ungenauigkeit soll aber nicht wie im ersten Beispiel geschätzt werden, da die theoretische Größe exakt ist. Man will vielmehr einen Bereich bestimmen, in dem die exakte Größe liegen kann: der Identifizierungsbereich. Ein derartiges lineares Modell könnte beispielsweise durch zusätzliche unbeobachtete Variablen aufgestellt werden. Die theoretische abhängige Variable könnte durch das theoretische Modell

$$y = \beta_0 + \beta_1 x + \epsilon \quad (2.7)$$

mit $\epsilon \sim N(0, \sigma^2)$ beschrieben werden. Man beobachtet aber nur zwei Grenzen $\underline{y}$ und $\overline{y}$, deren Verteilungen in Abhängigkeit von y unbekannt sind. Außerdem könnten

auch die unabhängigen Variablen nur als Intervalle $\underline{x}$ und $\bar{x}$ vorliegen. Von Interesse ist in beiden Fällen aber nicht die Verteilung der Grenzen, sondern die Schätzung der Parameter β_0 und β_1 .

Die Beiden Modelle (2.6) und (2.7) unterscheiden sich also deutlich in ihrer Struktur (siehe beispielsweise [Cattaneo und Wiencierz, 2012]). Außerdem ist die Interpretation der Intervalle sehr unterschiedlich und damit auch der Ansatz zur Schätzung.

In den nächsten Abschnitten werden verschiedene Ansätze besprochen um Identifizierungsbereiche für die Parameter bei gegebenen Intervalldaten zu bestimmen¹.

2.3. Notation

Die Daten des Regressionsproblems werden zur Designmatrix der unabhängigen Variablen $\mathbf{x}$ und dem Vektor der abhängigen Variablen $\mathbf{y}$ zusammengefasst. Dabei hat $\mathbf{x}$ die Dimension $n \times (p + 1)$ mit n Beobachtungen der p unabhängigen Variablen, und $\mathbf{y}$ die Dimension $n \times 1$.

Im Weiteren sind die abhängigen und unabhängigen Variablen keine Skalare, sondern als Intervalle $\mathfrak{X}_i = [\underline{x}_i, \bar{x}_i]$ und $\mathfrak{Y}_i = [\underline{y}_i, \bar{y}_i]$ gegeben. $\mathfrak{X}$ und $\mathfrak{Y}$ fassen wie auch im skalaren Fall die einzelnen Beobachtungen zusammen. Es gilt

$$\begin{aligned}\mathfrak{X} &= (\mathfrak{X}_1, \dots, \mathfrak{X}_n), \\ \mathfrak{Y} &= (\mathfrak{Y}_1, \dots, \mathfrak{Y}_n)\end{aligned}$$

wobei

$$\begin{aligned}\mathfrak{X}_i &= [\underline{x}_i, \bar{x}_i], \\ \mathfrak{Y}_i &= [\underline{y}_i, \bar{y}_i]\end{aligned}$$

für $i = 1, \dots, n$ ist.

Wenn die Regressoren $\mathfrak{X}_i$ aus mehreren Beobachtungen bestehen, ergibt sich jeweils ein Vektor aus 2-Tupeln:

$$\mathfrak{X}_i = ([\underline{x}_{i1}, \bar{x}_{i1}], \dots, [\underline{x}_{ip}, \bar{x}_{ip}]) \text{ für } i = 1, \dots, n.$$

Offensichtlich ist mit den Intervalldaten eine klassische Schätzung nicht möglich, da diese mit skalaren Werten durchgeführt wird. Aus den jeweiligen Intervallen könnten bestimmte Werte ausgewählt werden. Es ist aber gerade nicht bekannt, welche die wahren Werte in diesen Intervallen sind, die für die Inferenz verwendet werden

¹Die Ansätze in den Abschnitten 2.3, 2.4 und 2.5 basieren auf dem unveröffentlichten Manuskript [Augustin, 2012]. Auch die Notation und Darstellung orientiert sich an dieser Quelle.

sollen. Daher ist das Ziel, alle zulässigen Schätzwerte für einen Modellparameter ϑ zu finden. Ist diese Menge konvex, so erhält man ein Intervall

$$I(\vartheta) := [\underline{\vartheta}, \bar{\vartheta}],$$

als Identifizierungsbereich, in dem alle Schätzwerte liegen, die aus einer beliebigen Kombination von skalaren Werten aus den jeweiligen Intervallen der Daten hervorgehen würden. Ein Schätzwert $\hat{\vartheta}$ ist also zulässig, wenn sich Daten $x_i \in \mathfrak{X}_i$ und $y_i \in \mathfrak{Y}_i$ für $i = 1, \dots, n$ finden lassen, mit denen das Schätzverfahren zum Schätzwert $\hat{\vartheta}$ führen.

2.4. Formulierung als Optimierungsproblem

Für die lineare Regression kann der Schätzer für die Modellparameter durch das KQ-Kriterium bestimmt werden. Dabei werden die quadrierten Abstände der Beobachtungen zum Erwartungswert des Modells minimiert. Bei der generalisierten linearen Regression kann dieses Kriterium hingegen nicht angewandt werden. Es ist jedoch bekannt, dass die Maximum-Likelihood-Methode für die lineare Regression den gleichen Schätzer liefert wie das KQ-Kriterium und diese zudem auch für die generalisierte lineare Regression geeignet ist.

Bei der Maximum-Likelihood-Schätzung wird $\hat{\vartheta}$ so gewählt, dass die beobachteten Daten mit diesem Parameter die größte Wahrscheinlichkeit hätten. Dazu wird die Likelihood-Funktion $L(\vartheta; \mathbf{x}, \mathbf{y})$ logarithmiert und anschließend nach ϑ abgeleitet. Dies ergibt die Score-Funktion

$$s(\vartheta; \mathbf{x}, \mathbf{y}) := \frac{\partial l(\vartheta; \mathbf{x}, \mathbf{y})}{\partial \vartheta}, \quad (2.8)$$

wobei $l(\vartheta; \mathbf{x}, \mathbf{y})$ die logarithmierte Likelihood-Funktion der Daten ist. Für das Maximum von $f(\vartheta; \mathbf{x}, \mathbf{y})$ gilt $s(\vartheta; \mathbf{x}, \mathbf{y}) = 0$. Durch Auflösen nach ϑ erhält man $\hat{\vartheta}$ als Schätzwert (siehe beispielsweise [Young und Smith, 2005], Kapitel 8). Diese Score-Funktion gibt damit ein Kriterium für die Gültigkeit eines Schätzwertes vor [Augustin, 2012]. Lassen sich also Daten $x_i \in \mathfrak{X}_i$ und $y_i \in \mathfrak{Y}_i$ für $i = 1, \dots, n$ und ein $\hat{\vartheta}$ finden, für das $s(\hat{\vartheta}; \mathbf{x}, \mathbf{y}) = 0$ gilt, so ist $\hat{\vartheta} \in I(\vartheta)$.

Allgemeiner kann auch nur angenommen werden, dass es eine Schätzfunktion $\Psi(\vartheta; \mathbf{x}, \mathbf{y})$ gibt, die eine eindeutige Nullstelle hat, welche den Schätzer liefert. Für das Intervall der zulässigen Schätzwerte erhält man so

$$\underline{\vartheta} = \min_{\vartheta, \mathbf{x}, \mathbf{y}} \vartheta, \quad \bar{\vartheta} = \max_{\vartheta, \mathbf{x}, \mathbf{y}} \vartheta \quad (2.9)$$

unter den Nebenbedingungen

$$\begin{aligned}\Psi(\vartheta; \mathbf{x}, \mathbf{y}) &= 0 \\ x_i &\in \mathfrak{X}_i, \quad i = 1, \dots, n \\ y_i &\in \mathfrak{Y}_i, \quad i = 1, \dots, n\end{aligned}$$

als zu lösendes Optimierungsproblem [Augustin, 2012]. Außer der Score-Funktion auf Basis der logarithmierten Likelihood-Funktion wären auch andere Ansätze möglich, wie beispielsweise der M-Schätzer [Godambe, 1991; Shapiro, 2000], der ein Verfahren liefert, um entsprechende Schätzfunktionen zu konstruieren.

2.5. Optimierung mit Strafterm

Da es sich bei der Schätzfunktion um eine komplexe, und insbesondere nicht-lineare Nebenbedingung handelt, ist ein zweiter Ansatz, die Gleichheitsnebenbedingung mit in die Zielfunktion als Strafterm zu integrieren [Augustin, 2012]. So kann man

$$\underline{\vartheta} = \underset{\vartheta; \mathbf{x}, \mathbf{y}}{\operatorname{argmin}} \left(\vartheta + \rho \cdot (\Psi(\vartheta; \mathbf{x}, \mathbf{y}))^2 \right), \quad (2.10)$$

$$\overline{\vartheta} = \underset{\vartheta; \mathbf{x}, \mathbf{y}}{\operatorname{argmax}} \left(\vartheta - \rho \cdot (\Psi(\vartheta; \mathbf{x}, \mathbf{y}))^2 \right) \quad (2.11)$$

berechnen, wobei ρ so groß gewählt werden muss, dass die Nebenbedingung

$$\Psi(\vartheta; \mathbf{x}, \mathbf{y}) = 0$$

implizit erfüllt wird. Weiterhin müssen bei der Optimierung die Nebenbedingungen

$$\begin{aligned}x_i &\in \mathfrak{X}_i, \quad i = 1, \dots, n \\ y_i &\in \mathfrak{Y}_i, \quad i = 1, \dots, n\end{aligned}$$

explizit als Restriktionen eingehalten werden. Diese sind aber ausschließlich linear und lassen sich als Box-Constraints beschreiben. Auf der anderen Seite wird die zu optimierende Funktion deutlich komplexer und ist im Allgemeinen nicht mehr als Polynom darstellbar.

Die erste Formulierung stellt die allgemeine Lösung für das Problem dar. Als Alternative zu diesem Problem mit komplexen Nebenbedingungen kann der zweite Ansatz mit komplexer Zielfunktion verwendet werden. Doch auch hier handelt es sich nicht unbedingt um ein einfacheres Optimierungsproblem, da beispielsweise nicht klar ist, ob die Zielfunktion konvex ist.

In beiden Fällen ist das Optimierungsproblem $n(p+1) + q$ dimensional, wobei n die Anzahl der Beobachtungen, p die Anzahl der abhängigen Variablen x_i , und q die

Dimension des Vektorraums von ϑ ist. Dadurch wird das Problem bei vielen Beobachtungen äußerst rechenintensiv, weshalb effiziente Optimierungsverfahren benötigt werden.

2.6. Analytische Lösung für die lineare Regression

Sind nur die y -Werte als Intervalle gegeben und die x -Werte als Skalare, so können die Schranken für die Parameterschätzer im linearen Modell analytisch bestimmt werden. Für β_1 wird das Verfahren zur Bestimmung der Grenzen von Rohwer und Pötter beschrieben [Rohwer und Pötter, 2001]. Hierbei ist β_0 zwar im Modell enthalten, es wird aber kein Identifizierungsbereich für diesen Parameter bestimmt. Der Ansatz lässt sich aber leicht für β_0 erweitern. Ebenfalls lässt sich das eindimensionale Modell ohne β_0 aus der Grundidee herleiten.

Vorerst wird hier der Ansatz zur Bestimmung der Grenzen des Identifizierungsbereichs von β_1 vorgestellt. Dabei wird das Intervall der abhängigen Variablen y durch

$$y = \underline{y} + \delta$$

beschrieben, wobei $0 \leq \delta \leq \bar{y} - \underline{y}$ gilt. Für die unabhängige Variable gilt hingegen

$$x = \underline{x} = \bar{x}.$$

Diese ist also ein Skalar oder ein triviales Intervall, das nur aus einem Wert besteht. Wird nun die geschlossene Form des Parameterschätzers β_1 herangezogen, sind die zulässigen Schätzer durch

$$\begin{aligned} \hat{\beta}_1(\delta_1, \dots, \delta_n) &= \frac{n \sum_{i=1}^n x_i (\underline{y}_i + \delta_i) - \sum_{i=1}^n x_i \sum_{i=1}^n (\underline{y}_i + \delta_i)}{n \sum_{i=1}^n x_i^2 - \sum_{i=1}^n x_i \sum_{i=1}^n x_i} \\ &= \frac{v}{w} + \frac{n}{w} \sum_{i=1}^n \left(x_i - \frac{1}{n} \sum_{j=1}^n x_j \right) \delta_i \end{aligned} \quad (2.12)$$

gegeben, wobei

$$v := n \sum_{i=1}^n x_i \underline{y}_i - \sum_{i=1}^n x_i \sum_{i=1}^n \underline{y}_i$$

und

$$w := n \sum_{i=1}^n x_i^2 - \sum_{i=1}^n x_i \sum_{i=1}^n x_i$$

ist [Rohwer und Pötter, 2001]. Dabei kann man sehen, dass

$$\hat{\beta}_1(\delta_1, \dots, \delta_n) \propto \sum_{i=1}^n \left(x_i - \frac{1}{n} \sum_{j=1}^n x_j \right) \delta_i \quad (2.13)$$

gilt. Wählt man δ_i jeweils so, dass die einzelnen Summanden in (2.13) minimal oder maximal sind, so erhält man auch in der Summe das Minimum und Maximum. Da (2.13) proportional zum Parameterschätzer ist, erhält man dadurch zugleich den minimalen und maximalen zulässigen Parameterschätzer. Es lässt sich leicht erkennen, dass ein Summand minimal ist, wenn

$$\delta_i = \begin{cases} \bar{y}_i - \underline{y}_i & \text{wenn } x_i < \tilde{x} \\ 0 & \text{sonst} \end{cases}$$

gewählt wird und maximal ist, wenn

$$\delta_i = \begin{cases} \bar{y}_i - \underline{y}_i & \text{wenn } x_i > \tilde{x} \\ 0 & \text{sonst} \end{cases}$$

gewählt wird [Rohwer und Pötter, 2001], wobei

$$\tilde{x} := \frac{1}{n} \sum_{j=1}^n x_j$$

das arithmetische Mittel von x ist. Das heißt, bei der Maximierung wird für x -Werte, die größer als der Mittelwert sind, die Obergrenze $\bar{y}$ und für x -Werte, die kleiner als der Mittelwert sind, die Untergrenze $\underline{y}$ selektiert. Für die Minimierung wird genau die umgekehrte Auswahl getroffen. Anschließend müssen mit den nun gegebenen Beobachtungen nur noch die Parameterschätzer β_0 und β_1 berechnet werden.

Ist β_0 nicht im Modell enthalten, lässt sich die Idee von Rohwer und Pötter leicht übertragen. Der Parameterschätzer für β ist hier durch

$$\begin{aligned} \hat{\beta}(\delta_1, \dots, \delta_n) &= \frac{\sum_{i=1}^n x_i (\underline{y}_i + \delta_i)}{\sum_{i=1}^n x_i^2} \\ &\propto \sum_{i=1}^n x_i (\underline{y}_i + \delta_i) = \sum_{i=1}^n x_i \underline{y}_i + \sum_{i=1}^n x_i \delta_i \\ &\propto \sum_{i=1}^n x_i \delta_i \end{aligned} \tag{2.14}$$

gegeben (siehe auch Kapitel 3, Abschnitt 3.4). Hier muss also nur noch untersucht werden, ob x_i positiv oder negativ ist. Es wird dementsprechend für das Minimum

$$\delta_i = \begin{cases} \bar{y}_i - \underline{y}_i & \text{wenn } x_i < 0 \\ 0 & \text{sonst} \end{cases}$$

und für das Maximum

$$\delta_i = \begin{cases} \bar{y}_i - \underline{y}_i & \text{wenn } x_i > 0 \\ 0 & \text{sonst} \end{cases}$$

gewählt.

Zuletzt wird nun noch eine Vorschrift für die Bestimmung der Grenzen des Identifizierungsintervalls von β_0 hergeleitet. Für den Parameterschätzer $\hat{\beta}_0$ gilt mit skalaren Werten für x und y

$$\hat{\beta}_0 = \frac{1}{n} \sum_{i=1}^n y_i - \frac{1}{n} \sum_{i=1}^n x_i \hat{\beta}_1,$$

wobei der Parameterschätzer für $\hat{\beta}_1$ eingesetzt werden kann (siehe beispielsweise [Rohwer und Pötter, 2001], Abschnitt 9.2.3). Mit obigen Definitionen können die zulässigen Parameterschätzer für β_0 durch

$$\begin{aligned} \hat{\beta}_0(\delta_1, \dots, \delta_n) &= \frac{1}{n} \sum_{i=1}^n (\underline{y}_i + \delta_i) - \frac{v\tilde{x}}{w} - \frac{n\tilde{x}}{w} \sum_{i=1}^n (x_i - \tilde{x}) \delta_i \\ &\propto \frac{1}{n} \sum_{i=1}^n \delta_i - \frac{n\tilde{x}}{w} \sum_{i=1}^n (x_i - \tilde{x}) \delta_i \\ &= \sum_{i=1}^n \left(\frac{1}{n} - \frac{n\tilde{x}}{w} (x_i - \tilde{x}) \right) \delta_i \end{aligned} \quad (2.15)$$

bestimmt werden. Aus (2.15) kann nun die Vorschrift hergeleitet werden. Für das Minimum setzt man

$$\delta_i = \begin{cases} \bar{y}_i - \underline{y}_i & \text{wenn } \frac{1}{n} - \frac{n\tilde{x}}{w} (x_i - \tilde{x}) < 0 \\ 0 & \text{sonst} \end{cases}$$

und für das Maximum

$$\delta_i = \begin{cases} \bar{y}_i - \underline{y}_i & \text{wenn } \frac{1}{n} - \frac{n\tilde{x}}{w} (x_i - \tilde{x}) > 0 \\ 0 & \text{sonst} \end{cases},$$

um die Grenzen des Intervalls für $\hat{\beta}_0$ zu erhalten.

Intervalldaten und generalisierte lineare Modelle

Seitz, M.

2015, XVII, 110 S. 31 Abb., Softcover

ISBN: 978-3-658-08745-6