Statistische Modelle mit nicht-ignorierbar fehlender Zielgröße und Anwendung in der Reject Inference

---

Dissertation

zur Erlangung des akademischen Grades
Doktor der Naturwissenschaften
der Technischen Universität Dortmund

Der Fakultät Statistik
der Technischen Universität Dortmund
vorgelegt von

Michael Bücker

Dortmund, 2011

technische universität
dortmund

Gutachter	Prof. Dr. Walter Krämer
	Prof. Dr. Claus Weihs
Tag der mündlichen Prüfung	21. April 2011

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungs- und Symbolverzeichnis

Abkürzungen

AEL	Adjusted Empirical Likelihood
AUROC	Area under ROC
BDSG	Bundesdatenschutzgesetz
EL	Empirische Likelihood
GLM	Generalisierte Lineare Modelle
KQ	Kleinste Quadrate
MAR	Missing at random
MCAR	Missing completely at random
ML	Maximum Likelihood
MEL	Maximum Empirical Likelihood
MNAR	Missing not at random
MSE	Mean Square Error (Mittlerer Quadratischer Fehler)
ROC	Receiver Operating Characteristic
SEL	Semi-Empirische Likelihood

Symbole

$\mathbb{1}_A(x)$	Indikatorfunktion der Menge A an der Stelle x

$\|\cdot\|$	Euklidische Norm
$X \perp\!\!\!\perp Y$	X und Y sind stochastisch unabhängig
$\xrightarrow{d}$	Konvergenz in Verteilung bzw. schwache Konvergenz
$\xrightarrow{p}$	Konvergenz in Wahrscheinlichkeit
$\mathcal{B}(n, \pi)$	Binomialverteilung mit Parametern n und π
$\mathcal{E}(\theta, \psi)$	Exponentialfamilie mit kanonischem Parameter θ und Dispersionsparameter ψ
$F_n(\boldsymbol{x})$	Empirische Verteilungsfunktion an der Stelle $\boldsymbol{x}$
$\mathbb{I}$	Einheitsmatrix
$\lambda_{\max}(\boldsymbol{A})$	größter Eigenwert von $\boldsymbol{A}^\top \boldsymbol{A}$
$L(F)$	Empirische Likelihood-Funktion
$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	Normalverteilung mit Erwartungswertvektor $\boldsymbol{\mu}$ und Kovarianzmatrix $\boldsymbol{\Sigma}$
$o_p(\cdot)$, $O_p(\cdot)$	Stochastische Landau Notation (vgl. Owen 2001, Anhang A.1)
$R(\boldsymbol{\theta})$	Empirischer Profil-Likelihood-Quotient
$R(F)$	Empirischer Likelihood-Quotient
R^2_{McF}	Pseudo-R^2 von McFadden
$R^\star(\boldsymbol{\theta})$	Adjusted Empirical Profil-Likelihood-Quotient
$\boldsymbol{T}(F)$	Parametrisierung von F
$W(\boldsymbol{\theta})$	Empirische Likelihood-Quotienten-Teststatistik für den Parameter $\boldsymbol{\theta}$
χ^2_q	χ^2-Verteilung mit q Freiheitsgraden
$\chi^2_{q,\alpha}$	α-Quantil der χ^2-Verteilung mit q Freiheitsgraden

Kapitel 1

Einleitung

Fehlende Daten stellen in der angewandten Statistik ein allgegenwärtiges Problem dar. Ein Beispiel dafür ist die so genannte „Reject Inference", die sich mit fehlenden Informationen bei der Einschätzung von Kreditrisiken beschäftigt. Im Kreditscoring werden Bankkunden durch den Kreditgeber hinsichtlich ihrer Bonität bewertet, um das Risiko eines Ausfalls einzuschätzen und Kreditanträge mit hohem Ausfallrisiko ablehnen zu können. Diese Bewertung – der so genannte Score – trägt zur Entscheidungsfindung der Bank bei. Um eine Regel zu finden, mit deren Hilfe sich Kunden möglichst gut nach ihrer Rückzahlungsfähigkeit einstufen lassen, werden in der Praxis unterschiedlichste Modellierungsansätze verwendet. Üblicherweise werden Scoring-Modelle anhand von Daten der Bestandskunden geschätzt. Vollständige Daten liegen allerdings nur für diejenigen Kunden vor, denen bereits ein Kredit gewährt wurde. Diese Bestandskunden wurden als kreditwürdig eingeschätzt. Die Daten der abgewiesenen Kreditnehmer können nicht berücksichtigt werden, da für sie unbekannt ist, ob sie einen Kredit zurückzahlen könnten. Dies kann unter Umständen zu einer Verzerrung der Parameterschätzungen des Scoring-Modells führen, so dass es möglicherweise unzutreffende Prognosen liefert und damit nicht zur Anwendung auf Neukunden geeignet ist.

Das Kreditscoring ist nur ein Beispiel für Modelle mit teilweise nicht beobachtbarer Zielgröße. Daneben gibt es zahlreiche andere Umstände, welche fehlende Daten verursachen. Grundsätzlich sollte stets die Ursache für das Fehlen berücksichtigt werden. Je nachdem, ob das Fehlen der Daten von Kovariablen oder der Zielgröße selbst abhängt, lassen sich verschiedene Aussagen über die Ignorierbarkeit der feh-

lenden Beobachtungen treffen. Solche Annahmen über den Prozess, der die fehlenden Daten verursacht, sind nicht immer überprüfbar. Da sie aber die anschließende Modellbildung entscheidend mitbestimmen, spielen sie eine wichtige Rolle.

Zum Umgang mit fehlenden Daten wurden bereits zahlreiche Ansätze vorgeschlagen. Dabei gehen die meisten Autoren von der Annahme aus, dass das Fehlen der Ausprägungen der Zielgröße ignorierbar ist, d. h. unabhängig von der Zielgröße bei gegebenen Realisationen der Kovariablen. Die Rechtfertigung dieser Annahme ist dabei im Zusammenhang zu untersuchen. Im Kreditscoring trifft sie zu, falls die Gewährung des Kredits unabhängig von der tatsächlichen Rückzahlfähigkeit ist – gegeben die übrigen Merkmale der Kreditnehmer sind bekannt. Das ist aber nur der Fall, falls die Entscheidung zur Kreditvergabe ausschließlich aufgrund der Scoremerkmale getroffen wird. Dies bedeutet, dass das Scoring-Modell exakt die gleichen Variablen verwendet und keine zusätzlichen Einflussgrößen bei der Kreditvergabe eine Rolle spielen. Im Allgemeinen ist diese Annahme jedoch verletzt, da die Entscheidung für oder gegen eine Kreditvergabe nicht allein anhand der Scores, sondern auch mittels des persönlichen Eindrucks des zuständigen Bankangestellten getroffen wird. Eine nur auf den Score basierende Entscheidung ist durch das Bundesdatenschutzgesetz untersagt. Daher kann die Vernachlässigung der Daten abgelehnter Kunden zu Verzerrungen bei der Modellanpassung führen.

Die vorliegende Arbeit entwickelt einen allgemeinen Modellierungsansatz, welcher erlaubt, dass teilweises Fehlen der Beobachtungen der Zielvariablen auch von der Zielgröße selbst abhängt. Dieser umfasst Lineare und Generalisierte Lineare Modelle und verallgemeinert einen Ansatz von Qin, Leung und Shao (2002). Deren Methode ermöglicht es, den Erwartungswert einer nicht-ignorierbar fehlenden Zufallsvariablen mit Hilfe von Kovariablen konsistent zu schätzen. Hierzu verwenden die Autoren ein Modell, dessen hybride Likelihood aus einer parametrischen und einer Empirischen Likelihood zusammengesetzt ist. Dabei steuert der parametrische Teil den Prozess des Fehlens der Daten. Qin, Leung und Shao (2002) schätzen die empirische Verteilung aller, auch der unbeobachteten Daten. Als Schätzer für den Erwartungswert verwenden sie ein gewichtetes arithmetisches Mittel, gewissermaßen den Erwartungswert dieser empirischen Verteilung. Sie zeigen, dass der resultierende Schätzer konsistent und asymptotisch normalverteilt ist.

Die Methode von Qin, Leung und Shao (2002) lässt sich auf allgemeinere Schätzprobleme übertragen. Generalisierte Lineare Modelle verknüpfen den Erwartungswert

der Zielgröße mit einer Linearkombination der Kovariablen über eine Link-Funktion. Die Schätzung der Modellparameter erfolgt überlicherweise durch die Maximum-Likelihood-Methode. Die Idee, welche dieser Arbeit zu Grunde liegt, ist die Schätzung der Parameter durch Scoregleichungen, die wie bei Qin, Leung und Shao (2002) gewichtet werden. Damit lassen sich die Parameter Generalisierter Linearer Modelle mit teilweise fehlender Zielgröße konsistent schätzen. Außerdem sind die resultierenden Parameterschätzer asymptotisch normalverteilt.

Bei Vorliegen von ignorierbar fehlenden Daten sind sowohl der gewöhnliche ML-Schätzer als auch der neue Schätzer konsistent. Wenn die fehlenden Daten jedoch nicht vernachlässigt werden können, ist das herkömmliche Modell falsch spezifiziert. Dann ist der ML-Schätzer verzerrt und beide Schätzungen unterscheiden sich voneinander. Dies lässt sich mit Hilfe eines Hausman-Tests überprüfen. Damit kann die Hypothese der Ignorierbarkeit fehlender Daten beziehungsweise der korrekten Modellspezifikation überprüft werden.

Die Arbeit ist wie folgt strukturiert: Zunächst wird in Kapitel 2 das Problem fehlender Daten analysiert und in unterschiedliche Situationen klassifiziert. Außerdem werden verschiedene existierende Methoden zur Behandlung fehlender Daten diskutiert. Kapitel 3 thematisiert das Kreditscoring – im Besonderen die Reject Inference. Empirische Likelihood-Verfahren sind das Thema von Kapitel 4. Da der Ansatz von Qin, Leung und Shao (2002) auf der Theorie der Empirischen Likelihood beruht, werden die Idee und die wichtigsten Ergebnisse der Forschung auf diesem Gebiet dargestellt und ein Fokus auf ihre Anwendung bei fehlenden Daten gelegt. Kapitel 5 umfasst schließlich eine Beschreibung des Modells von Qin, Leung und Shao (2002) sowie dessen Erweiterung auf die Modellschätzung. Außerdem wird der Hausman-Test erläutert, mit dem sich die Ignorierbarkeit testen lässt. In einer Simulationsstudie in Kapitel 6 werden die Eigenschaften der hergeleiteten Schätzer und des Hausman-Tests für Lineare und Logistische Regressionsmodelle untersucht. Des Weiteren folgt in Kapitel 7 eine Anwendung auf einen realen Datensatz eines Finanzinstituts. Kapitel 8 fasst schließlich die Erkenntnisse der Arbeit zusammen und gibt einen kurzen Ausblick auf weitere Forschungsmöglichkeiten.

Kapitel 2

Fehlende Daten

Nahezu jeder Anwender statistischer Verfahren wird früher oder später mit fehlenden Merkmalsausprägungen in einem Datensatz konfrontiert. Fehlende Daten treten in den unterschiedlichsten statistischen Fragestellungen auf. Aufgrund der Vielzahl verschiedener Situationen gibt es keine angemessene universelle Vorgehensweise – kein Allgmeinrezept – zur Behandlung dieser Problematik. In der Literatur wurden bereits zahlreiche Methoden und Ideen vorgeschlagen, wie mit fehlenden Daten umgegangen werden kann. Einen Überblick über diese Arbeiten liefern Rubin (1976), Little und Rubin (1987), Schafer (2000), Tsiatis (2006) sowie Copas und Li (2002).

Neben dem in der Einleitung beschriebenen Kreditscoring gibt es in der Praxis unzählige Situationen fehlender Daten und genauso vielseitig sind die Gründe für das Fehlen. Dabei müssen diese Gegebenheiten differenziert betrachtet werden. So lassen sich grob Situationen unterscheiden, in denen entweder Daten rein zufällig fehlen – zum Beispiel durch Transkriptionsfehler oder Ausfall von Messgeräten – oder in denen die Daten systematisch fehlen. Die unterschiedlichen Umstände lassen sich im Wesentlichen hinsichtlich der Abhängigkeit des Fehlens von den beobachteten und unbeobachteten Größen gegeneinander abgrenzen. Im nächsten Abschnitt wird diese Differenzierung näher beleuchtet.

Wir betrachten im Folgenden das Fehlen von Daten in einem Regressionskontext. Uns interessiert die Verteilung von $Y|R, \boldsymbol{X}$ mit Zielgröße Y und Kovariablen $\boldsymbol{X}$ beziehungsweise Parameter dieser Verteilung. Fehlende Daten tauchen lediglich für die Zielvariable Y auf. Eine Übersicht zur Behandlung fehlender Kovariablen bietet Ibrahim et al. (2005).

2.1 Klassifizierung von Situationen fehlender Daten

Die Klassifizierung der Mechanismen, durch die fehlende Daten entstehen, erfolgte erstmals durch Rubin (1976) bzw. Little und Rubin (1987). Sie soll hier in einem Regressionskontext erläutert werden. Sei

$$P(Y|R, \boldsymbol{X})$$

die Verteilung der Zielgröße Y gegeben die Kovariablen $\boldsymbol{X}$ bzw.

$$P(Y|R, \boldsymbol{X}, \boldsymbol{\beta})$$

die parametrische Verteilung mit Parametervektor $\boldsymbol{\beta}$. Fehlende Daten tauchen lediglich für die abhängige Variable Y auf. Eine binäre Zufallsvariable R gebe an, ob Y beobachtet werden kann oder nicht, das heißt

$$R = \begin{cases} 1 & Y \text{ ist beobachtbar} \\ 0 & \text{die Beobachtung von } Y \text{ fehlt.} \end{cases}$$

Es besteht natürlich in der Praxis die Möglichkeit, dass auch Kovariablen teilweise nicht beobachtet werden können. Dieser Fall soll hier jedoch nicht weiter diskutiert werden. Zusätzlich sei

$$P(R|Y, \boldsymbol{X})$$

die Verteilung von $R|Y, \boldsymbol{X}$, also die Verteilung des Fehlens bei gegebenen abhängigen und unabhägigen Variablen. Auch hier kann eine parametrische Variante

$$P(R|Y, \boldsymbol{X}, \boldsymbol{\theta})$$

mit Parametervektor $\boldsymbol{\theta}$ unterstellt werden. Im Folgenden werden nun drei Situationen mit unterschiedlichen Annahmen betrachtet, wobei dazu im Weiteren die parametrisierten Verteilungen von $Y|R, \boldsymbol{X}$ und $R|Y, \boldsymbol{X}$ verwendet werden. Die nichtparametrischen Definitionen ergeben sich vollkommen analog.

Definition 2.1 (missing completely at random (MCAR)):

Fehlende Beobachtungen von Y werden als *missing completely at random* (MCAR) bezeichnet, falls

$$P(R|Y, \boldsymbol{X}, \boldsymbol{\theta}) = P(R|\boldsymbol{\theta}), \tag{2.1}$$

d. h. falls R weder von Y noch von $\boldsymbol{X}$ abhängt.

In Situationen, in denen Daten MCAR sind, gilt also $\boldsymbol{X}, Y \perp\!\!\!\perp R$. Damit hängt die Verteilung der abhängigen Variable Y nicht von R, sondern lediglich von $\boldsymbol{X}$ ab und es gilt

$$P(Y|R, \boldsymbol{X}, \boldsymbol{\beta}) = P(Y|\boldsymbol{X}, \boldsymbol{\beta}). \tag{2.2}$$

Abbildung 2.1 stellt eine schematische Veranschaulichung der MCAR-Situation dar, wobei die Abhängigkeit der Variablen durch verbindende Linien ausgedrückt wird.

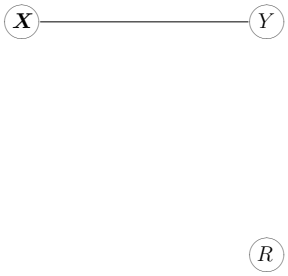

Abbildung 2.1 Schematische Darstellung MCAR (vgl. Smith und Elkan 2007).

Definition 2.1 stellt eine starke Anforderung an den Mechanismus des Fehlens dar. Wie im Anschluss diskutiert wird, ist eine solche Annahme in vielen Situationen nicht angemessen. Die nächste Definition lässt eine gewisse Abhängigkeit zwischen $\boldsymbol{X}$ und R zu.

Definition 2.2 (missing at random (MAR)):

Fehlende Ausprägungen von Y nennt man *missing at random* (MAR), wenn die Verteilung von R zwar von dem vollständig beobachtbaren Zufallsvektor $\boldsymbol{X}$ abhängt, aber bedingt auf $\boldsymbol{X}$ unabhängig von der Zielgröße Y ist, das heißt falls

$$P(R|Y, \boldsymbol{X}, \boldsymbol{\theta}) = P(R|\boldsymbol{X}, \boldsymbol{\theta}) \tag{2.3}$$

gilt.

Unter der MAR-Annahme gilt also $R|\boldsymbol{X} \perp\!\!\!\perp Y$. Dann folgt für die bedingte Verteilung von Y

$$P(Y|\boldsymbol{X}, R, \boldsymbol{\beta}) = P(Y|\boldsymbol{X}, \boldsymbol{\beta}). \tag{2.4}$$

Die abhängige Variable Y ist also bei gegebenen Kovariablen unabhängig von R: $Y|\boldsymbol{X} \perp\!\!\!\perp R$. In Abbildung 2.2 ist die MCAR-Situation grafisch dargestellt.

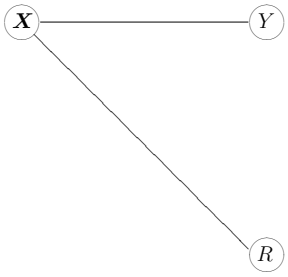

Abbildung 2.2 Schematische Darstellung MAR (vgl. Smith und Elkan 2007).

Die Bezeichnung *missing at random* kann irreführend sein, da damit keinesfalls ein rein zufälliges Fehlen von Beobachtungen gemeint ist (wie im MCAR-Fall). Wir können bisher feststellen, dass in MCAR- oder MAR-Situationen die Schätzung des Parametervektors $\boldsymbol{\beta}$ mit Hilfe der vollständigen Beobachtungen offenbar kein Problem darstellt. Die gesuchte Verteilung wird nicht durch das Fehlen beeinflusst. Allerdings liegen weniger Beobachtungen vor, so dass die Schätzung anhand der vollständigen Beobachtungspaare nicht so effizient ist wie in einer Situation ohne fehlende Daten.

Schließlich bleibt die Definition, in der keine Einschränkung der Abhängigkeitsstruktur zwischen den Zufallsvariablen vorausgesetzt wird.

Definition 2.3 (missing not at random (MNAR)):

Fehlende Ausprägungen von Y werden mit *missing not at random* (MNAR) bezeichnet, falls eine Abhängigkeit zwischen R, Y und $\boldsymbol{X}$ besteht. Es gilt dann

$$P(R|Y, \boldsymbol{X}, \boldsymbol{\theta}) \neq P(R|\boldsymbol{X}, \boldsymbol{\theta}). \tag{2.5}$$

Aus obiger Definition lässt sich (2.4) nicht folgern, so dass dann in der Regel

$$P(Y|R, \boldsymbol{X}, \boldsymbol{\beta}) \neq P(Y|\boldsymbol{X}, \boldsymbol{\beta})$$

gilt. Grafisch lässt sich der Zusammenhang wie in Abbildung 2.3 darstellen.

Für die Modellierung von $Y|\boldsymbol{X}$ ist vor allem von Interesse, ob man hierbei die fehlenden Daten ignorieren kann oder ob die Schätzer in einem parametrischen Modell sonst möglicherweise verzerrt sind. Die folgende Definition trennt diese beiden Fälle.

Definition 2.4 (Ignorierbarkeit):

Die Ausprägungen von Y heißen *nicht-ignorierbar fehlend*, falls

1. Y MNAR ist und
2. (im parametrischen Fall) die Parameter $\boldsymbol{\theta}$ und $\boldsymbol{\beta}$ in dem Sinne verschieden sind, dass der gemeinsame Parameterraum von $(\boldsymbol{\theta}, \boldsymbol{\beta})$ das kartesische Produkt der Parameterräume von $\boldsymbol{\theta}$ und $\boldsymbol{\beta}$ ist.

Ansonsten sind sie *ignorierbar fehlend* (Schafer 2000, S. 11). Die Ignorierbarkeit ist eine einfache Voraussetzung an die Verteilung von R dafür, dass eine Außerachtlassung der fehlenden Daten bei der Maximum-Likelihood-Schätzung von $P(Y|\boldsymbol{X}, \boldsymbol{\beta})$ keine Verzerrung verursacht (Rubin 1976).

Falls nicht-ignorierbar fehlende Daten vorliegen, muss der Mechanismus, der die fehlenden Daten erzeugt, explizit modelliert werden, um unverzerrte Schätzer für den interessierenden Parametervektor $\boldsymbol{\beta}$ zu erhalten (Schafer 2000, S. 28). Dies kann auf zwei unterschiedliche Arten geschehen, wie der nächste Abschnitt 2.2 zeigt.

Ein großes Problem bei der Definitionen von MAR und MNAR bzw. Ignorierbarkeit ist, dass diese beiden Situationen sich zwar theoretisch unterscheiden lassen, jedoch bisher keine allgemeine Methode zur Überprüfung der häufig unterstellten MAR-Annahme existiert. Da der Mechanismus der fehlenden Daten üblicherweise unbekannt ist, kann nicht ohne Weiteres festgestellt werden, ob das Fehlen der Zielgröße Y ignorierbar ist. Dies wäre möglich, wenn die fehlenden Werte vollständig oder teilweise beobachtet werden könnten. Bislang lässt sich die Entscheidung, ob MAR vorliegt, daher in Anwendungen nur aus Plausibilitätsüberlegungen heraus treffen.

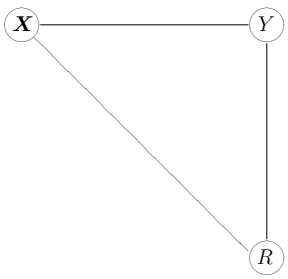

Abbildung 2.3 Schematische Darstellung MNAR (vgl. Smith und Elkan 2007).

Das nicht-ignorierbare Fehlen von Beobachtungen bringt große Schwierigkeiten mit sich und es gibt bisher nur wenige Verfahren, die diese Problematik berücksichtigen. Ein immer noch aktuelles Zitat aus Schafer (2000, S. 384) verdeutlicht diese Tatsache:

> „The assumption of ignorable nonresponse [...] is computationally very convenient because it allows us to conduct inferences without explicitly specifying a missingness mechanism. In many situations, however, ignorability is questionable or implausible, and it would be worthwhile to investigate nonignorable alternatives. [...] Broadly speaking, the literature on nonignorable methods has been rather sporadic, with extended discussion of special cases but relatively few overall themes. [...] The major barriers are no longer computational but conceptual; models for nonignorable response appropriate for wide classes of data (especially multivariate data) have not been proposed. Construction and evaluation of general models for nonignorable nonresponse are an important area for future study."

2.2 Selektionsmodelle und Pattern-Mixture Modelle

Es lassen sich zwei statistische Modellierungsansätze für nicht-ignorierbar fehlende Daten unterscheiden, die von verschiedenen Faktorisierungen der gemeinsamen Verteilung der Zufallsvariablen ausgehen. Dabei sind die beiden Modellansätze im Regressionskontext äquivalent, wenn die fehlenden Daten MCAR oder MAR sind. Im MNAR-Fall handelt es sich um verschiedene statistische Modelle.

Definition 2.5 (Selektionsmodell):

Mit den obigen Definitionen von Y, $\boldsymbol{X}$ und R nennt man die Faktorisierung

$$P(Y, R|\boldsymbol{X}) = P(Y|\boldsymbol{X}) \cdot P(R|Y, \boldsymbol{X})$$

ein *Selektionsmodell* (engl.: selection model).

Bei Selektionsmodellen betrachtet man die gesuchte Verteilung von $Y|\boldsymbol{X}$ getrennt von dem Prozess, welcher die fehlenden Daten produziert, also $R|Y, \boldsymbol{X}$. Bei der

Modellierung auf diese Weise ist allerdings problematisch, dass für beide Teile die Beobachtungen von Y nicht vollständig vorhanden sind.

Definition 2.6 (Pattern-Mixture Modell):

Mit den obigen Definitionen von Y, $\boldsymbol{X}$ und R nennt man die Faktorisierung

$$P(Y, R|\boldsymbol{X}) = P(Y|R, \boldsymbol{X}) \cdot P(R|\boldsymbol{X})$$

ein *Pattern-Mixture Modell*.

Im MCAR-Fall ist mit den Gleichungen (2.1) und (2.2) direkt klar, dass das Selektionsmodell und das Pattern-Mixture Modell äquivalent sind. Die Modelle lassen sich dann beide als

$$P(Y, R|\boldsymbol{X}) = P(Y|\boldsymbol{X}) \cdot P(R)$$

schreiben. Auch unter der MAR Annahme (2.3) gilt mit Gleichung (2.4) die Äquivalenz der beiden Modelle (vgl. Scheid 2005, S. 30 ff.), das heißt das Selektionsmodell vereinfacht sich zu

$$P(Y, R|\boldsymbol{X}) = P(Y|\boldsymbol{X}) \cdot P(R|Y, \boldsymbol{X}) = P(Y|\boldsymbol{X}) \cdot P(R|\boldsymbol{X})$$

und das Pattern-Mixture-Modell zu

$$P(Y, R|\boldsymbol{X}) = P(Y|R, \boldsymbol{X}) \cdot P(R|\boldsymbol{X}) = P(Y|\boldsymbol{X}) \cdot P(R|\boldsymbol{X}).$$

Falls die fehlenden Daten nicht-ignorierbar sind, handelt es sich bei den beiden Ansätzen um verschiedene statistische Modelle in dem Sinne, dass sie sich nicht wie oben vereinfachen lassen und die Faktorisierungen damit identisch sind.

2.3 Methoden zur Behandlung fehlender Daten

Es gibt unterschiedliche Wege, mit fehlenden Daten umzugehen. Dabei ist es wichtig, das Vorgehen nach der Art des Fehlens bzw. nach der Ignorierbarkeit auszurichten, da es ansonsten zu Verzerrungen kommen kann. In der Literatur wurden diesbezüglich bereits zahlreiche Methoden vorgeschlagen. Allgemein lässt sich festhalten, dass mit dem MCAR-Fall am leichtesten umzugehen ist, während der MNAR-Fall einige Schwierigkeiten mit sich bringt.

MCAR

Das Ziel der unter der MCAR-Annahme formulierten Verfahren beschränkt sich auf die Effizienzsteigerung der Schätzung. Falls nur die vollständigen Beobachtungen verwendet werden, führt dies häufig dazu, dass sich der Stichprobenumfang drastisch verringert. Ein gängiges Vorgehen ist dann die Ersetzung der fehlenden Werte durch möglichst gute Repräsentanten. In der Regel sind dies Mittelwerte oder lineare Prognosen aus anderen beobachteten Variablen. Solche Imputationsmethoden sind weit verbreitet. Die Multiple Imputation ist eine Erweiterung, welche von Rubin (1996) entwickelt wurde. Dabei wird ein fehlender Wert nicht nur einmal, sondern mehrfach durch verschiedene Werte ersetzt. Damit lässt sich aus den neuen Datensätzen auch die Unsicherheit, die durch die Ersetzung der fehlenden Werte entsteht, über die Varianz der Schätzung ermitteln.

Für den Klassifikations- bzw. Regressionskontext hat sich im Bereich des Maschinellen Lernens eine neue Disziplin entwickelt – das sogenannte Halbüberwachte Lernen. Im Gegensatz zum Überwachten Lernen sind nicht alle Ausprägungen der Zielgröße bekannt, aber sie sind eben auch nicht vollkommen unbekannt wie beim Unüberwachten Lernen. Ein Überblick zu Halbüberwachten Lernverfahren findet sich bei Chapelle, Schölkopf und Zien (2006).

MAR

Für die Bearbeitung von MAR Daten existieren zahlreiche statistische Verfahren. Anders als im Regressionskontext, in dem nur die Zielgröße teilweise fehlt, können MAR Daten in allgemeinen Schätzproblemen zu Verzerrungen führen. Dann ist es wichtig, das Fehlen der Daten zu berücksichtigen. Zu den Verfahren gehören unter anderem der EM-Algorithmus, Data Augmentation (vgl. Tanner und Wong 1987), Imputation und Multiple Imputation, Gewichtungsmethoden sowie Likelihood-basierte Verfahren. Einen Überblick geben Little und Rubin (1987) und Schafer (2000).

Da in der Regression unverzerrte Schätzungen mittels der vollständigen Beobachtungen möglich sind, bieten sich die gleichen Methoden zur Effizienzsteigerung wie im MCAR-Fall an.

MNAR

Veröffentlichungen über nicht-ignorierbar fehlende Daten sind immer noch selten. Für stetige Zielgrößen kann im MNAR-Fall die bekannte Arbeit von Heckman (1979), für kategorielle Zielgrößen das bivariate Probitmodell von Meng und Schmidt (1985) genannt werden. Kritik an diesen Modellen bezieht sich vor allem auf die starren Annahmen (vgl. Little 1985, Copas und Li 2002, einschl. Diskussion). Für kategorielle Zielgrößen und Kovariablen existieren Methoden von Fay (1986) und Baker und Laird (1988), sowie weitere darauf aufbauende Verfahren. Die Grenzen dieser Methoden zeigen Molenberghs et al. (1999) auf. Für Longitudinale Daten bieten Daniels und Hogan (2008, Kap. 8) einen Überblick über mögliche Verfahren, darunter viele Bayesianische Ansätze. Ibrahim und Lipsitz (1996) zeigen eine Möglichkeit auf, ein binäres Regressionsmodell zu schätzen, wenn die Zielgröße MNAR ist. Die Idee dabei ist, dass die fehlenden Ausprägungen der Zielgröße bei der Modellierung von R als fehlende Kovariablen behandelt werden können und hier der EM-Algorithmus angewendet wird. Ihr Ansatz lässt sich auf beliebige diskrete Zielgrößen verallgemeinern.

Kapitel 3

Reject Inference

Wenn Banken Kredite an Privat- oder Firmenkunden vergeben, ist dies stets mit dem Risiko verbunden, dass der Kreditnehmer bei Fälligkeit nicht in der Lage ist, das Darlehen zu tilgen. Im Interesse der Schuldner – vor allem aber auch der Bank selbst – ist diese Situation möglichst zu vermeiden. Der Geldgeber kann sein Risiko minimieren, indem er möglichst kein Geld an Kunden verleiht, die ein hohes Risiko für eine Zahlungsunfähigkeit aufweisen. Die Bank muss dazu eben dieses Risiko für jeden einzelnen Kunden einschätzen. Dies geschieht mit Hilfe eines Scoring-Modells. Das Risiko eines Kreditausfalls wird als die Wahrscheinlichkeit für die Zahlungsunfähigkeit eines Klienten modelliert. Dadurch kann die Bank Kreditanträge von potentiellen Debitoren mit hohem Ausfallrisiko erkennen und gegebenenfalls den Kredit ablehnen oder zu anderen Konditionen vergeben, um sich gegen das erhöhte Risiko abzusichern. Ihr eigenes Wagnis, dass Fehlbeträge durch „faule Kredite" entstehen, hält das Finanzinstitut dadurch gering und kann zudem entsprechend der eingegangenen Risiken Rücklagen bilden. Gleichzeitig werden die potentiellen Kunden vor einer Überschuldung geschützt.

3.1 Kreditscoring

Im Kreditscoring wird die Wahrscheinlichkeit eines Kreditausfalls mittels eines statistischen Modells geschätzt. Im Folgenden sei Y die binäre Zufallsvariable, die den Ausfall des Darlehens angibt. Dabei gilt $Y = 1$, falls der Kunde zahlungsfähig ist und $Y = 0$, falls er zahlungsunfähig ist. Aussagen über die Zielvariable Y sollen mit

Hilfe von Kovariablen getroffen werden, die im Kontext des Kreditscorings auch Scoremerkmale oder Scorevariablen genannt werden. Diese Variablen sind der Bank für sämtliche Kreditantragssteller bekannt. Sie sollen möglichst aussagekräftig bezüglich der Zahlungsfähigkeit sein, so dass die Entscheidung über die Kreditvergabe anhand dieser Merkmale getroffen werden kann. Typische Scorevariablen sind demographische Merkmale wie Wohnort, Beruf, Alter oder Familienstand, aber auch Daten der Bank wie die Dauer der bisherigen Kundenbeziehung und Anzahl der bereits aufgenommenen Kredite. Die verwendeten Einflussgrößen können sowohl Angaben aus den Kreditanträgen, Kundendaten der Bank als auch Informationen externer Auskunfteien wie der Schufa enthalten. Im Folgenden bezeichnet $\boldsymbol{X} = (X_1, \ldots, X_d)^\top$ den Zufallsvektor der d erhobenen Scoremerkmale. Das gesuchte Modell für die Bonität hat also die Form

$$P(Y = 1|X_1, \ldots, X_d) = f(X_1, \ldots, X_d), \tag{3.1}$$

wobei die Merkmale $X_1, \ldots, X_d$ zu einem Zeitpunkt vor Ausfall des Kredits beobachtet wurden (vgl. Henking, Bluhm und Fahrmeir 2006, Kap. 7). Die Ausfallwahrscheinlichkeit bei einem Kunden mit Scoremerkmalsausprägung $\boldsymbol{x}$ ist $P(Y = 0|\boldsymbol{X} = \boldsymbol{x}) = 1 - P(Y = 1|\boldsymbol{X} = \boldsymbol{x})$.

Das Modell (3.1) wird anhand von Bestandskunden geschätzt. Für diese sind die Ausprägungen des Merkmalsvektors $\boldsymbol{X}$ sowie Zahlungsfähigkeit Y bis zum aktuellen Zeitpunkt bekannt. Ist das Modell einmal geschätzt, können aktuelle Kreditanträge damit eingestuft werden.

Das Ziel des Kreditscorings ist die Entwicklung einer so genannten Scorekarte: einer Kategorisierung der Ausfallsrisiken auf einer vorgegebenen Skala. Entsprechend der prognostizierten Ausfallwahrscheinlichkeiten werden Punktwerte (Scores) vergeben. Anschließend kann für jeden Antragsteller anhand der Realisationen der Merkmale $\boldsymbol{X}$ sein Scorewert ermittelt werden und als Entscheidungshilfe bei der Zustimmung oder Ablehnung des Darlehens dienen. In dieser Arbeit soll es jedoch nicht um die konkrete Erstellung einer Scorekarte gehen – der Schwerpunkt liegt vielmehr auf der Bestimmung eines geeigneten Modells zur Schätzung der Ausfallwahrscheinlichkeiten.

Es gibt zahlreiche Methoden, die sich für die Vorhersage der Bonität nutzen lassen. Die Zielgröße Y ist beim Kreditscoring dichotom, während die Scoremerkmale unterschiedliche Skalenniveaus aufweisen können. Es lassen sich daher gängige Klassifikationsverfahren verwenden. Besonders häufig werden in der Praxis aufgrund der

guten Interpretierbarkeit und Verständlichkeit Logistische Regressionsmodelle und Lineare Diskriminanzanalyse eingesetzt. Aber auch Verfahren wie Neuronale Netze, Support Vector Machines, Entscheidungsbäume und Zufallswälder sind mögliche Ansätze.

Einen Überblick über statistische Methoden im Kreditscoring geben unter anderem Rosenberg und Gleit (1994). Sie gehen vor allem auf die Lineare Diskriminanzanalyse, Entscheidungsbäume sowie Expertensysteme und Neuronale Netze ein. Thomas (2000) gibt einen kurzen Rückblick auf die Geschichte des Kreditscorings und fasst Verfahren des Scorings wie die Fisher'sche Diskriminanzanalyse, Entscheidungsbäume, das Nächste-Nachbarn Verfahren, Logistische Regression und das Probitmodell zusammen.

Hand und Henley (1997) betonen, dass der Einsatz des Kreditscorings aufgrund des starken Zuwachses an aufgenommenen Krediten immer mehr an Bedeutung gewinnt. Denn Kreditscoring ist nicht nur im Bankwesen wichtig, sondern auch für Kreditkartenfirmen, im Versandhandel oder beim Finanzierungskauf. Dabei ist es unter anderem aus Kostengründen interessant, durch Kreditscoring eine vollständige oder teilweise Automatisierung der Kreditvergabeentscheidung zu erreichen. Die Autoren gehen auch auf das Problem der Reject Inference ein: nur für bestehende Kunden, denen ein Kredit gewährt wurde, ist bekannt, ob sie ihren Kredit zurückzahlen konnten oder nicht. Daher wird das Scoremodell üblicherweise nur anhand der Kunden geschätzt, die tatsächlich einen Kredit erhalten haben. Damit ist die Stichprobe, die der Schätzung des Modells zugrunde liegt, verzerrt bzw. selektiert. Es gibt Versuche, die Kenntnis über $\boldsymbol{X}$ für die abgelehnten Kunden (engl.: rejects) zu nutzen um diese Verzerrung zu vermeiden. Auf die Reject Inference wird im nächsten Abschnitt näher eingegangen.

Hand (2005) sowie Krämer und Bücker (2009) geben eine Übersicht über Verfahren, die zum Vergleich von Scorekarten geeignet sind. Da es eine Vielzahl an Methoden zur Konstruktion von Scorekarten gibt, sind verlässliche Vergleiche verschiedener Prognosemodelle von großer Bedeutung. Das beste Scoringmodell unter mehreren möglichen zu finden, ist im Allgemeinen nicht trivial. Auf diesen Aspekt des Kreditscorings wird in dieser Arbeit jedoch nicht näher eingegangen, vielmehr werden die Resultate der beiden genannten Arbeiten berücksichtigt.

3.2 Reject Inference

Die binäre Zufallsvariable R gebe an, ob ein Kunde einen Kredit erhalten hat ($R = 1$) oder ob ihm der Kredit verwehrt wurde ($R = 0$). Das bedeutet, dass die Zielvariable Y nur bei den Kunden beobachtet werden kann, für die $R = 1$ gilt. Nehmen wir an, die Stichprobe der Kunden, für die Y bekannt ist, enthalte n Beobachtungen Außerdem gebe es noch $N - n$ weitere Kunden, für die Y nicht beobachtbar ist. Insgesamt liegen also die Realisationen von $X_1, \ldots, X_N$ und $Y_1, \ldots, Y_n$ vor. Die Realisationen von $R_1, \ldots, R_n$ sind somit gleich 1, die von $R_{n+1}, \ldots, R_N$ gleich 0. Tabelle 3.1 verdeutlicht dieses Schema, wobei ein horizontaler Strich (–) für eine nicht beobachtete Realisation steht.

Tabelle 3.1 Schematische Darstellung einer Stichprobe für die Reject Inference.

Beobachtung	$\boldsymbol{X}$	Y	R
1	$\boldsymbol{x}_1$	y_1	1
2	$\boldsymbol{x}_2$	y_2	1
⋮	⋮	⋮	⋮
n	$\boldsymbol{x}_n$	y_n	1
$n+1$	$\boldsymbol{x}_{n+1}$	–	0
⋮	⋮	⋮	⋮
N	$\boldsymbol{x}_N$	–	0

Damit lässt sich das Kreditscoring als Problem fehlender Daten verstehen. Dabei können die Kovariablen für alle Kunden beobachtet werden, wohingegen die Zielgröße nur bei den akzeptierten Kunden vorliegt. Verfahren, die die verfügbaren Information über die abgelehnten Kunden (die rejects) ausnutzen, um Aussagen über deren unbekanntes Zahlungsverhalten zu treffen, werden unter dem Oberbegriff *Reject Inference* zusammengefasst. Wie der Name deutlich macht, soll mit den Daten der abgelehnten Kunden Inferenz betrieben werden. Dadurch sollen mögliche Verzerrungen, die in Abschnitt 3.1 und in Kapitel 2 schon besprochen wurden, verhindert werden.

Für die Reject Inference stellt sich nun zunächst die Frage, ob im Falle des Kreditscorings die fehlenden Beobachtungen der Zielgröße MCAR, MAR oder MNAR sind.

Außerdem ist nicht klar, ob die Reject Inference die Prognosegüte im Kreditscoring steigern kann. Zu diesen Fragestellungen gibt es in der Literatur bisher widersprüchliche Resultate. Beide Fragestellungen werden nun getrennt betrachtet.

Verbessert Reject Inference die Prognosegüte eines Scoring-Modells?

Crook und Banasik (2004) untersuchen, ob Reject Inference überhaupt in der Lage ist, die Vorhersagegüte der Scoring-Modelle zu verbessern. Für ihre Untersuchung verwenden sie einen Datensatz, der auch die vollständigen Informationen über die Kunden enthält, die eigentlich vom Kreditinstitut abgelehnt worden wären. Sie zeigen, dass mit Hilfe der Reject Inference Verbesserungen der Vorhersagegüte möglich sind. Verstraeten und Poel (2005) hingegen stellen fest, dass sich die mögliche Verzerrung der Modellschätzung nur moderat auf die Vorhersagegüte und die Profitabilität eines Scoring-Modells auswirkt.

Hand und Henley (1993) geben einen Überblick über Methoden, die in der Reject Inference Verwendung finden und hinterfragen, ob Reject Inference überhaupt nützlich ist. Sie kommen zu dem Schluss, dass dies nicht der Fall ist. Allerdings berücksichtigen sie bei ihren Überlegungen die Unterscheidung zwischen den möglichen Situationen fehlender Daten nicht und gehen implizit davon aus, dass diese MAR sind.

Sind die fehlenden Daten im Kreditscoring MCAR, MAR oder MNAR?

Da im Allgemeinen nicht anhand von Daten getestet werden kann, ob die fehlenden Daten MAR oder MNAR sind (vgl. Abschnitt 2.1, S. 8), kann die zutreffende Annahme zunächst nur über Plausibilitätsüberlegungen gefunden werden. Feelders (2000) untersucht dies und unterscheidet, unter welchen Annahmen im Kreditscoring die fehlenden Daten MCAR, MAR oder MNAR sind. Er stellt fest, dass der MCAR-Fall vorliegt, falls den Kunden rein zufällig ein Kredit bewilligt wird, unabhängig von ihren Scoremerkmalen oder der Zahlungsfähigkeit. Dies ist der Fall, wenn eine Bank gar kein Scoring betreibt und deshalb eine unrealistische Annahme.

Die MAR Annahme ist gerechtfertigt, falls die Bewilligung eines Kreditantrags *nur* anhand der Scoremerkmale getroffen wurde und es keine anderen Einflüsse auf die

Entscheidung gab. Die fehlenden Daten sind MNAR, wenn die Entscheidung teilweise anhand von Charakteristika getroffen wird, die nicht in den Scoremerkmalen enthalten sind. Dies ist zum Beispiel der Fall, wenn der Sachbearbeiter der Bank bei seiner Entscheidung nicht an den Scorewert gebunden ist, sondern auch den Gesamteindruck des potentiellen Kunden bei der Kreditvergabe einfließen lassen kann. So kann es zum Beispiel vorkommen, dass einem Kunden anhand des Scorewertes kein Geld geliehen würde, der Klient dem Kundenbetreuer jedoch sehr gut bekannt ist, so dass dieser keine Bedenken gegenüber einer Kreditvergabe hat. Dieses Vorgehen hat auch eine Grundlage im Datenschutzgesetz.

Rechtliche Bedenken gegenüber dem Kreditscoring treten immer wieder auf, da die Kreditvergabe mit Hilfe eines Scores für den Antragsteller häufig nicht transparent ist. Einen Überblick hierzu und zur rechtlichen Situation des Scoring in Deutschland bietet Weichert (2006). Er geht unter anderem auf das Verbot automatisierter Entscheidungen ein, das im § 6 des Bundesdatenschutzgesetzes (BDSG) festgehalten ist. Die genaue Formulierung des entscheidenden Paragraphen lautet:

Bundesdatenschutzgesetz (BDSG)

§ 6a Automatisierte Einzelentscheidung

(1) Entscheidungen, die für den Betroffenen eine rechtliche Folge nach sich ziehen oder ihn erheblich beeinträchtigen, dürfen nicht ausschließlich auf eine automatisierte Verarbeitung personenbezogener Daten gestützt werden, die der Bewertung einzelner Persönlichkeitsmerkmale dienen. Eine ausschließlich auf eine automatisierte Verarbeitung gestützte Entscheidung liegt insbesondere dann vor, wenn keine inhaltliche Bewertung und darauf gestützte Entscheidung durch eine natürliche Person stattgefunden hat.

[...]

Weichert (2006) stellt fest, dass dieser Paragraph auf das Kreditscoring anwendbar ist. Dabei wird das Scoring als solches keinesfalls allgemein durch den Gesetzgeber verboten, sondern lediglich einigen Regeln unterworfen. Es wird festgelegt, dass die Entscheidung der Kreditvergabe nicht allein automatisiert getroffen werden darf, sondern durch einen Sachbearbeiter überprüft und im Zweifelsfall auch redigiert werden muss. Folglich wird die Kreditvergabe als solche nicht nur vom Score und

damit von den Kovariablen abhängen, sondern es werden auch weitere Einflüsse eine Rolle spielen.

Außerdem wird bei der Modellierung von Kreditausfallrisiken in der Regel eine Modellwahl durchgeführt. Dabei werden diejenigen Merkmale selektiert, die einen tatsächlichen Einfluss auf die Zahlungsmoral haben. Allerdings sind diese ausgewählten unabhängigen Variablen möglicherweise nicht bei jeder neuen Schätzung der Ausfallrisiken dieselben, so dass auch in diesem Fall das Fehlen nicht nur von den verwendeten Scoremerkmalen abhängt.

Geht man schließlich davon aus, dass die zusätzlichen Einflussgrößen wie die Beurteilung durch den Sachbearbeiter oder zusätzliche Scorevariablen mit der Zielgröße korreliert sind, hängt das Fehlen der Daten von dem tatsächlichen Ausfallrisiko der Kunden ab. Das bedeutet letztendlich, dass die kompetente Einschätzung eines Sachbearbeiters und die Wahl sinnvoller Scoremerkmale gegen die Ignorierbarkeit der fehlenden Daten abgelehnter Kunden und für eine Abhängigkeit des Fehlens von der Zielgröße sprechen.

3.3 Bestehende Ansätze der Reject Inference

Die bestehenden Veröffentlichungen auf dem Gebiet der Reject Inference lassen sich grob in zwei Gruppen einteilen: Arbeiten, die Beobachtungen der abgelehnten Kreditnehmer als MAR auffassen und solche, die von einer Nicht-Ignorierbarkeit ausgehen.

Zur ersten Gruppe gehört zunächst die Arbeit von Feelders, Chang und McLachlan (1998). Mittels einer Diskriminanzanalyse mit Mischverteilungen mit zwei Komponenten pro Klasse erzielen die Autoren auf einem Datensatz aus der Kreditwirtschaft leicht bessere Ergebnisse als ohne die Daten der abgelehnten Kunden. Auch Feelders (1999, 2000) stellt eine neue Methode zur Reject Inference basierend auf der Schätzung von Mischverteilungen mit Hilfe des EM-Algorithmus' vor. In einer Simulationsstudie zeigt sich, dass bei geringer Stichprobengröße Verbesserungen bei der Vorhersage der Kreditausfälle möglich sind. Joanes (1993) verwendet zwei Ansätze zur Reject Inference (Augmentation und iterative Reklassifikation, vgl. McLachlan 1975) und geht damit implizit ebenso von der Ignorierbarkeit der Ausprägungen der abgelehnten Kreditnehmer aus.

Banasik und Crook (2007) untersuchen das Reject Inference Problem und diskutieren, in welchen Situationen die fehlenden Daten MCAR, MAR oder MNAR sind. Sie erläutern nicht-ignorierende Methoden wie das in Kapitel 2 erwähnte bivariate Probitmodell von Meng und Schmidt (1985) sowie Methoden der Gewichtung von Beobachtungen. Die Autoren verbinden beide Verfahren, können jedoch auf diese Weise die Vorhersagegüte nicht verbessern. Auch Boyes, Hoffman und Low (1989) verwenden das Modell von Meng und Schmidt (1985) und konzentrieren sich bei ihren Prognosen nicht nur auf das Ausfallrisiko der Kredite sondern auf erwartete Gewinne durch die Kreditvergabe und versuchen diese zu maximieren. Jacobson und Roszbach (2003) gehen ähnlich vor und untersuchen zudem mit einem Kreditdatensatz, inwiefern sich die Berücksichtigung der abgelehnten Kunden auf den Kreditbestand, das Risikopotential (Value at Risk) und Verluste durch Kreditausfälle der Bank auswirkt. Kim und Sohn (2007) stellen fest, dass das bivariate Probitmodell von Meng und Schmidt (1985) in der Reject Inference das Problem der Stichprobenverzerrung nicht vollständig lösen kann. Greene (1998) geht von einer verzerrten Modellschätzung im Kreditscoring durch Ignorieren der fehlenden Beobachtungen aus und schätzt ein Modell für das Kreditrisiko sowie die Höhe eines Kreditausfalls mit einem eigenen Modellansatz ähnlich dem von Heckman (1979). Eine Anwendung auf einen Datensatz zeigt, dass das neue Modell eine größere Trennschärfe besitzt. Smith und Elkan (2004) stellen die Reject Inference als Bayessches Netz mit verschiedenen Abhängigkeitsstrukturen dar (vgl. dazu auch die Abbildungen 2.1, 2.2 und 2.3). Für alle aufgezählten Fälle beschreiben sie Methoden zur Modellierung, für die nicht-ignorierbaren Fälle schlagen auch sie bivariate Probitmodelle (Meng und Schmidt 1985) vor. Smith und Elkan (2007) versuchen die Verzerrung nicht-ignorierbar fehlender Daten mit Hilfe einer Mischverteilung zu überwinden, die sie in die Richtung der fehlenden Beobachtungen „verschieben".

Insgesamt lässt sich feststellen, dass in der Literatur große Uneinigkeit über die Annahmen der fehlenden Beobachtungen herrscht. Fast alle Ansätze, die eine Ignorierbarkeit ablehnen, verwenden dabei Modelle, die auf der Arbeit von Heckman (1979) bzw. Meng und Schmidt (1985) basieren, welches wegen seiner starren Annahmen kritisiert wurde und sich in der Praxis häufig als unbrauchbar erweist. Aus den in Abschnitt 3.2 genannten Gründen gehen wir in dieser Arbeit von einer Nicht-Ignorierbarkeit aus und leiten dafür in Kapitel 5 ein neues Verfahren zur Reject Inference her.

Kapitel 4

Empirische Likelihood

Die Maximum-Likelihood-Methode ist wegen ihrer nützlichen Eigenschaften das wohl am häufigsten verwendete Verfahren zur Schätzung parametrischer Modelle. Auch zur Konstruktion von Tests können Likelihood-Methoden herangezogen werden. Der Nachteil parametrischer Likelihood-Inferenz ist allerdings die Voraussetzung einer bekannten Verteilungsfamilie. In der Regel ist die zugrundeliegende Verteilung unbekannt und es kann nicht immer davon ausgegangen werden, dass die Stichprobe aus einer der gängigen Verteilungsfamilien stammt. Bei einer fälschlicherweise angenommen Verteilungsklasse führen parametrische Likelihood-Methoden zu verzerrten Ergebnissen.

Dies begründet die Idee einer verteilungsfreien Likelihood-Theorie. Auch im nichtparametrischen Kontext erweist sich die Likelihood-Funktion als nützlich. Owen (1988, 1990, 2001) hat hierzu die Theorie der Empirischen Likelihood entwickelt. Diese wurde durch zahlreiche Arbeiten ausgebaut, so dass Empirische Likelihood-Verfahren heute in nahezu jedem Bereich der Statistik angewendet werden können. Es existiert eine Fülle von Einsatzmöglichkeiten in der Praxis. Owen (2001) gibt einen ausführlichen Überblick über Anwendungsgebiete, darunter (Generalisierte) Lineare Modelle, Glättungsverfahren sowie Modelle für abhängige Daten (z. B. Zeitreihen).

4.1 Empirische Likelihood

Im parametrischen Fall ist die Likelihood für eine Zufallsstichprobe das Produkt über die Dichte- bzw. Wahrscheinlichkeitsfunktion an der Stelle der Realisationen und wird als Funktion in dem interessierenden Parameter aufgefasst. Auf ähnliche Weise lässt sich auch das nichtparametrische Analogon definieren.

Definition 4.1 (Empirische Likelihood-Funktion):

Seien $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ Realisationen eines u.i.v. d-dimensionalen Zufallsvektors mit Verteilungsfunktion F. Dann nennt man

$$L(F) = \prod_{i=1}^{n} dF(\boldsymbol{x}_i) = \prod_{i=1}^{n} p_i$$

die *Empirische Likelihood-Funktion*, wobei $p_i = dF(\boldsymbol{x}_i) = P(\boldsymbol{X} \leq \boldsymbol{x}_i) - P(\boldsymbol{X} < \boldsymbol{x}_i)$ und die Verteilung F unbekannt sei. Im Falle von Bindungen in der Stichprobe (z. B. falls die zugrunde liegende Verteilung diskret ist), sind die p_i nicht als Sprünge der Verteilungsfunktion zu verstehen, sondern als Gewichte der einzelnen Realisationen (vgl. Owen 2001, S. 11f). In der Literatur wird jedoch im Allgemeinen einfach die obige Definition verwendet. Im Folgenden wollen wir daher zur Vereinfachung der Notation auf eine Unterscheidung zwischen Stichproben mit und ohne Bindungen verzichten.

Die Empirische Likelihood-Funktion wird häufig auch als *Nichtparametrische Likelihood-Funktion* bezeichnet. Aus der Definition ist direkt klar, dass $L(F) = 0$ ist für jede stetige Verteilung F. Um eine positive Empirische Likelihood (EL) zu erhalten muss jede Realisation der Stichprobenvariablen eine positive Wahrscheinlichkeit aufweisen. Eine plausible Eigenschaft der Empirischen Likelihood-Funktion ist, dass sie ihr Maximum in der empirischen Verteilung annimmt.

Satz 4.1 (Owen 2001, Theorem 2.1):

Die Empirische Likelihood-Funktion wird durch die Empirische Verteilungsfunktion

$$F_n(\boldsymbol{x}) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{(-\infty, x_1] \times \cdots \times (-\infty, x_d]}(\boldsymbol{x}_i)$$

maximiert.

Beweis: Die Stichprobe enthalte m verschiedene Werte $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m$, wobei $n_j \geq 1$ die Anzahl der Beobachtungen gleich $\boldsymbol{z}_j$ ist. Sei weiter $F \neq F_n$ eine ansonsten beliebige Verteilungsfunktion, $p_j = dF(\boldsymbol{z}_j) = P(\boldsymbol{X} \leq \boldsymbol{z}_j) - P(\boldsymbol{X} < \boldsymbol{z}_j)$ und $\widehat{p}_j = n_j/n$. Falls $p_j = 0$ für ein $j \in \{1, \ldots, m\}$, dann folgt $L(F) = 0 < L(F_n)$. Wir gehen also davon aus, dass alle $p_j > 0$ und für mindestens ein $j \in \{1, \ldots, m\}$ $p_j \neq \widehat{p}_j$. Weiter gilt $\log(x) \leq x - 1$ für alle $x > 0$, wobei die Gleichheit nur für $x = 1$ gilt. Daher folgt

$$\begin{aligned}
\log\left(\frac{L(F)}{L(F_n)}\right) &= \sum_{j=1}^{m} n_j \log\left(\frac{p_j}{\widehat{p}_j}\right) \\
&= n \sum_{j=1}^{m} \widehat{p}_j \log\left(\frac{p_j}{\widehat{p}_j}\right) \\
&< n \sum_{j=1}^{m} \widehat{p}_j \left(\frac{p_j}{\widehat{p}_j} - 1\right) \\
&\leq 0
\end{aligned}$$

und damit

$$L(F) < L(F_n).$$

■

Tests und Konfidenzintervalle können für parametrische Verteilungsfamilien anhand des Likelihood-Quotienten konstruiert werden. Analog lässt sich wieder ein Nichtparametrischer bzw. Empirischer Likelihood-Quotient definieren.

Definition 4.2 (Empirischer Likelihood-Quotient):

Der *Empirische Likelihood-Quotient* ist definiert als

$$R(F) = \frac{L(F)}{L(F_n)}.$$

Mit Satz 4.1 ist unmittelbar zu sehen, dass für den Empirischen Likelihood-Quotienten

$$R(F) = \prod_{i=1}^{n} np_i$$

folgt.

Nun interessieren wir uns für einen unbekannten Parameter(vektor) $\boldsymbol{\theta} = \boldsymbol{T}(F)$ der Verteilung F. Um Konfidenzintervalle oder Tests für den unbekannten Parameter $\boldsymbol{\theta}$ zu konstruieren, betrachtet man den Empirischen Profil-Likelihood-Quotienten.

Definition 4.3 (Empirischer Profil-Likelihood-Quotient):

Der *Empirische Profil-Likelihood-Quotient* für den unbekannten Parameter $\boldsymbol{\theta}$ ist gegeben durch

$$R(\boldsymbol{\theta}) = \sup_{p_1,\ldots,p_n} \{ R(F) | \, \boldsymbol{T}(F) = \boldsymbol{\theta}, F \in \mathcal{F} \}, \tag{4.1}$$

wobei $\mathcal{F}$ die Menge aller Verteilungsfunktionen ist.

Der Einfachheit halber betrachten wir nun als unbekannten Parameter zunächst den Erwartungswert $\boldsymbol{\mu}$ der Verteilung F. Weiter sei die gesamte Wahrscheinlichkeitsmasse von F auf die Ausprägungen der Stichprobe verteilt, es gelte also $\sum_{i=1}^n p_i = 1$ (vgl. dazu Owen 2001, S. 13f.). Wir können den Empirischen Profil-Likelihood-Quotienten für $\boldsymbol{\mu}$ dann als

$$R(\boldsymbol{\mu}) = \sup_{p_1,\ldots,p_n} \left\{ \prod_{i=1}^n np_i \middle| \, p_i \geq 0, \sum_{i=1}^n p_i = 1, \sum_{i=1}^n p_i \boldsymbol{x}_i = \boldsymbol{\mu} \right\} \tag{4.2}$$

ausdrücken.

Mit Hilfe von Lagrange Multiplikatoren lässt sich eine explizite Darstellung des Empirischen Profil-Likelihood-Quotienten finden. Es existiert ein eindeutiges Maximum für Gleichung (4.2), falls $\boldsymbol{\mu}$ innerhalb der konvexen Hülle der Beobachtungen $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ liegt. Dann kann das Maximum von $\prod_{i=1}^n np_i$ oder äquivalent von $\sum_{i=1}^n \ln(np_i)$ unter den Nebenbedingungen $p_i \geq 0$, $\sum_{i=1}^n p_i = 1$ und $\sum_{i=1}^n p_i \boldsymbol{x}_i = \boldsymbol{\mu}$ mit Hilfe der Lagrange-Methode gefunden werden.

Mittels Lagrange Multiplikatoren γ_1 und $\boldsymbol{\gamma}_2$ lässt sich $R(\boldsymbol{\mu})$ durch Maximierung von

$$G = \sum_{i=1}^n \ln(np_i) - \gamma_1 \left(\sum_{i=1}^n p_i - 1 \right) - \boldsymbol{\gamma}_2^\top \left(\sum_{i=1}^n p_i (\boldsymbol{x}_i - \boldsymbol{\mu}) \right)$$

bestimmen. Ableiten nach p_i, $i = 1, \ldots, n$ und Nullsetzen ergibt

$$\frac{\partial G}{\partial p_i} = \frac{1}{p_i} - \gamma_1 - \boldsymbol{\gamma}_2^\top (\boldsymbol{x}_i - \boldsymbol{\mu}) = 0.$$

Daraus folgt

$$\begin{aligned} 0 = \sum_{i=1}^n p_i \frac{\partial G}{\partial p_i} &= n - \gamma_1 \\ \Leftrightarrow \qquad \gamma_1 &= n. \end{aligned}$$

Sei $\boldsymbol{\gamma}_2 = n\boldsymbol{\lambda}$, so erhalten wir als Ausdruck

$$p_i = p_i(\boldsymbol{\mu}) = \frac{1}{n\,(1 + \boldsymbol{\lambda}^\top(\boldsymbol{x}_i - \boldsymbol{\mu}))}. \tag{4.3}$$

Der Parameter $\boldsymbol{\lambda}$ lässt sich in Abhängigkeit von $\boldsymbol{\mu}$ durch numerische Verfahren bestimmen, denn für $\boldsymbol{\lambda} = \boldsymbol{\lambda}(\boldsymbol{\mu})$ gilt

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\boldsymbol{x}_i - \boldsymbol{\mu}}{1 + \boldsymbol{\lambda}^\top(\boldsymbol{x}_i - \boldsymbol{\mu})} = \mathbf{0}. \tag{4.4}$$

Damit gilt für den Empirischen Profil-Likelihood-Quotienten aus Gleichung (4.1)

$$R(\boldsymbol{\mu}) = \prod_{i=1}^{n} \frac{1}{(1 + \boldsymbol{\lambda}^\top(\boldsymbol{x}_i - \boldsymbol{\mu}))}.$$

Das Maximum von $R(\boldsymbol{\mu})$ ist der Maximum-Empirical-Likelihood-Schätzer (MEL-Schätzer). Für den unbekannten Erwartungswert erhält man als MEL-Schätzer das arithmetische Mittel $\widehat{\boldsymbol{\mu}}_{\mathrm{MEL}} = \bar{\boldsymbol{x}}$. Dies lässt sich leicht verifizieren, denn $\prod_{i=1}^{n} p_i$ mit $\sum_{i=1}^{n} p_i = 1$ und $p_i \geq 0$ $(i = 1, \ldots, n)$ nimmt genau dann sein Maximum an, falls $p_i = 1/n$. Dann gilt auch $\boldsymbol{\mu} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i = \bar{\boldsymbol{x}}$ und Gleichung (4.4) ist erfüllt für $\boldsymbol{\lambda} = \boldsymbol{\lambda}(\boldsymbol{\mu}) = \mathbf{0}$ (vgl. Mittelhammer et al. 2000, Kap. 12).

Die Parameterschätzung ist jedoch nicht die eigentliche Intention der Empirischen Likelihood. Der große Nutzen des Empirischen Profil-Likelihood-Quotienten liegt in der Tatsache, dass sich mit ihm nichtparametrische Konfidenzintervalle und Signifikanztests für den Parameter $\boldsymbol{\mu}$ konstruieren lassen. Eine Aussage über die asymptotische Verteilung gibt der folgende Satz, der ein nichtparametrisches Pendant zu den Ergebnissen von Wilks (1938) bezüglich des Likelihood-Quotienten darstellt.

Satz 4.2 (Empirical Likelihood Theorem; Owen 1988, 1990):

Seien $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ u.i.v. Realisationen eines d-dimensionalen Zufallsvektors $\boldsymbol{X}$ mit Verteilungsfunktion F, Erwartungswert $\mathrm{E}(\boldsymbol{X}) = \boldsymbol{\mu}_0$ und finiter Kovarianzmatrix $\boldsymbol{V}_0$ mit Rang $q > 0$. Dann konvergiert die *Empirische Likelihood-Quotienten-Teststatistik*

$$W(\boldsymbol{\mu}_0) = -2\ln R(\boldsymbol{\mu}_0) = 2\sum_{i=1}^{n} \ln\left(1 + \boldsymbol{\lambda}^\top(\boldsymbol{x}_i - \boldsymbol{\mu}_0)\right)$$

für $n \to \infty$ in Verteilung gegen eine χ^2_q-verteilte Zufallsvariable.

Beweis: Vergleiche Owen (1988) für den univariaten Fall, Owen (1990) oder Owen (2001, Kap. 11) für vektorwertige Zufallsvariablen. ■

Mit Hilfe des Satzes 4.2 lassen sich nun asymptotische Konfidenzintervalle für den unbekannten Erwartungswert $\boldsymbol{\mu}$ finden. Diese enthalten alle Punkte $\boldsymbol{\mu}$ mit $W(\boldsymbol{\mu}) \leq c_\alpha$, wobei für c_α $P(\chi^2_q \leq c_\alpha) = \alpha$ gilt. Ein Empirischer Likelihood-Quotienten-Test zum Niveau $1-\alpha$ für die Hypothesen

$$\mathrm{H}_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0 \qquad \text{gegen} \qquad \mathrm{H}_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$$

hat entsprechend die Form

$$\varphi(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = \begin{cases} 0 & W(\boldsymbol{\mu}_0) \leq c_\alpha \\ 1 & W(\boldsymbol{\mu}_0) > c_\alpha. \end{cases}$$

Das beschriebene Vorgehen lässt sich, wie in Owen (1990) gezeigt, auf weitere Parameter $\boldsymbol{\theta} = \boldsymbol{T}(F)$ der Verteilung F verallgemeinern. Außerdem lässt sich die Empirische Likelihood auch auf Momenten- bzw. Schätzgleichungen übertragen, wie im im nächsten Kapitel deutlich wird.

4.2 Empirische Likelihood und allgemeine Schätzgleichungen

Qin und Lawless (1994) übertragen das Konzept der Empirischen Likelihood auf die in der Ökonometrie weit verbreiteten allgemeinen Schätzfunktionen bzw. Schätzgleichungen (vgl. Godambe 1991). Damit lassen sich empirische Likelihood-Quotienten für Parameter bestimmen, welche über Schätzgleichungen definiert werden. Eine Vielzahl von Schätz- und Testproblemen lässt sich so nichtparametrisch bearbeiten.

Seien wieder $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ u.i.v. Realisationen einer d-dimensionalen Zufallsvariablen mit unbekannter Verteilungsfunktion F und einem k-dimensionalen Parametervektor $\boldsymbol{\theta}$, der zu F gehöre. Außerdem liege die Information über $\boldsymbol{\theta}$ durch $l \geq k$ unabhängige unverzerrte Schätzfunktionen vor. Diese Funktionen $g_j(\boldsymbol{X}, \boldsymbol{\theta})$ $(j = 1, \ldots, l)$

werden so gewählt, dass der Erwartungswert der Schätzfunktionen verschwindet. Das bedeutet, es soll

$$\mathrm{E}\left(\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{\theta})\right) = \mathbf{0} \tag{4.5}$$

gelten, wobei

$$\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{\theta}) = \left(g_1(\boldsymbol{X}, \boldsymbol{\theta}), \ldots, g_l(\boldsymbol{X}, \boldsymbol{\theta})\right)^\top.$$

Analog zum letzten Abschnitt lassen sich die Nebenbedingungen

$$p_i \geq 0 \quad (i = 1, \ldots, n),$$

$$\sum_{i=1}^{n} p_i = 1$$

und

$$\sum_{i=1}^{n} p_i \boldsymbol{g}(\boldsymbol{x}_i, \boldsymbol{\theta}) = \mathbf{0}$$

definieren. Die Empirische Likelihood ist dann bezüglich dieser Nebenbedingungen zu maximieren. Dazu geht man wie im vorherigen Abschnitt vor. Es ergibt sich wie in Gleichung (4.3)

$$p_i = p_i(\boldsymbol{\theta}) = \frac{1}{n(1 + \boldsymbol{\lambda}^\top \boldsymbol{g}(\boldsymbol{x}_i, \boldsymbol{\theta}))},$$

wobei $\boldsymbol{\lambda} = \boldsymbol{\lambda}(\boldsymbol{\theta})$ wieder ein Lagrange Multiplikator ist, für den wie in Gleichung (4.4)

$$\frac{1}{n} \sum_{i=1}^{n} \frac{\boldsymbol{g}(\boldsymbol{x}_i, \boldsymbol{\theta})}{1 + \boldsymbol{\lambda}^\top \boldsymbol{g}(\boldsymbol{x}_i, \boldsymbol{\theta})} = \mathbf{0} \tag{4.7}$$

gilt. Das Empirische Likelihood Theorem (Satz 4.2) lässt sich entsprechend übertragen (vgl. Owen 2001, Seite 39 ff.).

Qin und Lawless (1994) stellen fest, dass für $l = k$ der Maximum Empirical Likelihood Schätzer $\widehat{\boldsymbol{\theta}}_{\mathrm{MEL}}$, welcher den Empirischen Profil-Likelihood-Quotienten

$$R(\boldsymbol{\theta}) = \prod_{i=1}^{n} \frac{1}{1 + \boldsymbol{\lambda}^\top \boldsymbol{g}(\boldsymbol{x}_i, \boldsymbol{\theta})}$$

maximiert, genau derjenige Wert ist, der die Schätzgleichungen

$$\sum_{i=1}^{n} \boldsymbol{g}(\boldsymbol{x}_i, \boldsymbol{\theta}) = \mathbf{0}$$

löst (vgl. MEL-Schätzer für den Erwartungswert, S. 25).

In der Ökonometrie liegen häufig überbestimmte Schätzgleichungen vor, also der Fall $l > k$. Qin und Lawless (1994) leiten die asymptotische Verteilung der Parameter, der empirischen Verteilungsfunktion und des Empirischen Likelihood-Quotienten her, die im folgenden Satz zusammengefasst sind.

Satz 4.3 (Qin und Lawless 1994):

Seien $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ Realisationen eines u.i.v. d-dimensionalen Zufallsvektors $\boldsymbol{X}$ mit unbekannter Verteilungsfunktion F. Zu F gehöre ein k-dimensionaler Parametervektor $\boldsymbol{\theta}$, wobei der wahre Parameter $\boldsymbol{\theta}_0$ durch $\mathrm{E}(\boldsymbol{g}(\boldsymbol{x}, \boldsymbol{\theta}_0)) = \mathbf{0}$ eindeutig bestimmt ist und $\boldsymbol{g}(\boldsymbol{x}, \boldsymbol{\theta}) \in \mathbb{R}^l$. Sei

$$\tilde{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} R(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \prod_{i=1}^{n} \left(1 + \boldsymbol{\lambda}^\top \boldsymbol{g}(\boldsymbol{x}, \boldsymbol{\theta})\right)^{-1}$$

und $\tilde{\boldsymbol{\lambda}} = \boldsymbol{\lambda}(\tilde{\boldsymbol{\theta}})$. Weiter sei Θ eine Umgebung von $\boldsymbol{\theta}_0$ und $G(\boldsymbol{x})$ eine integrierbare Funktion, also $\mathrm{E}(G(\boldsymbol{X})) < \infty$. Unter den Voraussetzungen

1. $\mathrm{E}\left(\partial \boldsymbol{g}(\boldsymbol{X}, \boldsymbol{\theta}_0)/\partial \boldsymbol{\theta}\right)$ hat Rang k,
2. $\mathrm{E}\left(\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{\theta}_0)\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{\theta}_0)^\top\right)$ ist positiv definit,
3. $\partial \boldsymbol{g}(\boldsymbol{x}, \boldsymbol{\theta})/\partial \boldsymbol{\theta}$ ist stetig $\forall \boldsymbol{\theta} \in \Theta$,
4. $\partial^2 \boldsymbol{g}(\boldsymbol{x}, \boldsymbol{\theta})/\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top$ ist stetig in $\boldsymbol{\theta}$ für alle $\boldsymbol{\theta} \in \Theta$,
5. $\|\boldsymbol{g}(\boldsymbol{x}, \boldsymbol{\theta})\|^3 \leq G(x)\ \forall \boldsymbol{\theta} \in \Theta$,
6. $\|\partial \boldsymbol{g}(\boldsymbol{x}, \boldsymbol{\theta})/\partial \boldsymbol{\theta}\| \leq G(x)\ \forall \boldsymbol{\theta} \in \Theta$,
7. $\left\|\partial^2 \boldsymbol{g}(\boldsymbol{x}, \boldsymbol{\theta})/\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top\right\| \leq G(x)\ \forall \boldsymbol{\theta} \in \Theta$

gilt

$$\sqrt{n}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{V}),$$

$$\sqrt{n}(\tilde{\boldsymbol{\lambda}} - \mathbf{0}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{U})$$

und

$$\sqrt{n}(\tilde{F}_n(\boldsymbol{x}) - F(\boldsymbol{x})) \xrightarrow{d} \mathcal{N}(0, W(\boldsymbol{x})),$$

wobei

$$\tilde{F}_n(\boldsymbol{x}) = \sum_{i=1}^{n} \tilde{p}_i \mathbb{1}_{(-\infty, x_1] \times \cdots \times (-\infty, x_d]}(\boldsymbol{x}_i),$$

$$\tilde{p}_i = \frac{1}{n\left(1 + \tilde{\boldsymbol{\lambda}}^\top \boldsymbol{g}(\boldsymbol{x}_i, \tilde{\boldsymbol{\theta}})\right)},$$

$$\boldsymbol{V} = \left[\mathrm{E}\left(\frac{\partial \boldsymbol{g}(\boldsymbol{X}, \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}\right)^\top \left(\mathrm{E}\left(\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{\theta}_0)\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{\theta}_0)^\top\right)\right)^{-1} \mathrm{E}\left(\frac{\partial \boldsymbol{g}(\boldsymbol{X}, \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}\right)\right]^{-1},$$

$$W(\boldsymbol{x}) = F(\boldsymbol{x})(1 - F(\boldsymbol{x})) - \boldsymbol{B}(\boldsymbol{x})\boldsymbol{U}\boldsymbol{B}(\boldsymbol{x})^\top,$$

$$\boldsymbol{B}(\boldsymbol{x}) = \mathrm{E}\left(\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{\theta}_0)\mathbb{1}_{(-\infty, x_1] \times \cdots \times (-\infty, x_d]}(\boldsymbol{X})\right),$$

und

$$\boldsymbol{U} = \left(\mathrm{E}\left(\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{\theta}_0)\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{\theta}_0)^\top\right)\right)^{-1} \cdot \left[\mathbb{I} - \mathrm{E}\left(\frac{\partial \boldsymbol{g}(\boldsymbol{X}, \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}\right)\boldsymbol{V}\mathrm{E}\left(\frac{\partial \boldsymbol{g}(\boldsymbol{X}, \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}\right)^\top \left(\mathrm{E}\left(\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{\theta}_0)\boldsymbol{g}(\boldsymbol{X}, \boldsymbol{\theta}_0)^\top\right)\right)^{-1}\right].$$

Dabei sind $\tilde{\boldsymbol{\theta}}$ und $\tilde{\boldsymbol{\lambda}}$ unkorreliert. Außerdem gilt

$$W(\boldsymbol{\theta}_0) = -2\ln\left(R(\boldsymbol{\theta}_0)/R(\tilde{\boldsymbol{\theta}})\right) \xrightarrow{d} \chi^2_k.$$

Beweis: Vgl. Qin und Lawless (1994). ■

Mit Hilfe von Satz 4.3 lassen sich asymptotische Konfidenzintervalle und Tests für die einzelnen Parameter sowie die empirische Verteilungsfunktion konstruieren.

4.3 Hybride Likelihoods

Als hybride Likelihoods bezeichnen wir im Folgenden solche, die aus parametrischen und empirischen Likelihoods zusammengesetzt sind (vgl. Owen 2001, Kap. 9). Seien zwei unabhängige u.i.v. Stichproben $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ und $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m$ gegeben. Dabei sei die Verteilung der Zufallsvariablen Y_j $(j = 1, \ldots, m)$ bis auf einen unbekannten Parameter(vektor) $\boldsymbol{\theta}$ bekannt mit zugehöriger Dichte $g(\boldsymbol{y}, \boldsymbol{\theta})$. Die Verteilung der Zufallsvariablen X_i $(i = 1, \ldots, n)$ sei unbekannt. Ein intuitiver Ansatz ist dann die Verwendung einer Likelihood mit nichtparametrischem Teil für die Verteilung F von $\boldsymbol{X}_i$ $(i = 1, \ldots, n)$ und parametrischem Teil für die Verteilung G von $\boldsymbol{Y}_j$ $(j = 1, \ldots, m)$, also

$$L(F, \boldsymbol{\theta}) = \prod_{i=1}^{n} dF(\boldsymbol{x}_i) \prod_{j=1}^{m} g(\boldsymbol{y}_j, \boldsymbol{\theta}).$$

Dann ist der Likelihood-Quotient

$$R(F, \boldsymbol{\theta}) = \prod_{i=1}^{n} np_i \prod_{j=1}^{m} \frac{g(\boldsymbol{y}_j, \boldsymbol{\theta})}{g(\boldsymbol{y}_j, \widehat{\boldsymbol{\theta}})},$$

wobei $\widehat{\boldsymbol{\theta}}$ der gewöhnliche parametrische Maximum Likelihood Schätzer ist. Qin (1994) leitet die asymptotische χ^2-Verteilung des Empirischen Likelihood-Quotienten für das Zwei-Stichproben-Testproblem $\mathrm{H}_0 : \boldsymbol{\mu_X} = \boldsymbol{\mu_Y}$ her, wobei $\boldsymbol{\mu_X}$ der Erwartungswert der $X_i \ (i = 1, \ldots, n)$ und $\boldsymbol{\mu_Y}$ der Erwartungswert der $Y_j \ (j = 1, \ldots, m)$ ist.

Bei obigem Vorgehen ließe sich ebenfalls die Information über einen Parameter(vektor) $\boldsymbol{\phi}$ berücksichtigen, die durch Schätzgleichungen

$$\mathrm{E}(\boldsymbol{h}(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{\phi})) = \iint \boldsymbol{h}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\phi}) dG(\boldsymbol{y}, \boldsymbol{\theta}) dF(\boldsymbol{x}) = \mathbf{0}$$

gegeben ist. Die asymptotische χ^2-Verteilung des Empirischen Likelihood-Quotienten $R(\boldsymbol{\phi}) = \max_{F,\boldsymbol{\theta}} R(F, \boldsymbol{\theta})$ mit Nebenbedingungen

$$\sum_{i=1}^{n} p_i \int \boldsymbol{h}(\boldsymbol{x}_i, \boldsymbol{y}, \boldsymbol{\phi}) dG(\boldsymbol{y}, \boldsymbol{\theta}) = \mathbf{0}$$

lässt sich ebenso herleiten (vgl. Owen 2001, Kap. 9).

Qin und Wong (1996) bestimmen den Empirische Likelihood-Quotienten und dessen Verteilung für Situationen, in denen die parametrische Verteilung einer Zufallsvariablen für bestimme Teilmengen des Trägers bekannt, aber in anderen Teilmengen unbekannt ist. So kann zum Beispiel ein Konfidenzintervall für den Erwartungswert einer Zufallsvariablen bestimmt werden, deren Verteilung in einer Umgebung des Erwartungswerts der Normalverteilung entspricht, an den Rändern jedoch unbekannt ist.

4.4 Empirische Likelihood und fehlende Daten

Für Situationen, in denen fehlende Daten vorliegen, existieren zahlreiche hybride Ansätze. So schlägt Qin (2000) eine ähnliche Vorgehensweise wie im vorigen Abschnitt vor. Angenommen es liegen u.i.v. Beobachtungspaare $(\boldsymbol{x}_i, y_i) \ (i = 1, \ldots, N)$ vor, wobei die y_i für $i = n+1, \ldots, N$ fehlen. Bei gegebenem $\boldsymbol{x}$ sei die parametrische

bedingte Verteilung von $Y|\boldsymbol{X} = \boldsymbol{x}$ bekannt und besitze die Dichte $f(y|\boldsymbol{x}, \boldsymbol{\theta})$. Die Randverteilung von $\boldsymbol{x}$ folge einer unbekannten Verteilung G. Um ein Konfidenzintervall für den Erwartungswert μ_Y von Y zu bestimmen, leitet man die Likelihood der beobachteten Daten

$$L(\boldsymbol{\theta}, G) = \prod_{i=1}^{n} f(y_i|\boldsymbol{x}_i, \boldsymbol{\theta}) dG(\boldsymbol{x}_i) \prod_{j=n+1}^{N} dG(\boldsymbol{x}_j) = \prod_{i=1}^{N} dG(\boldsymbol{x}_i) \prod_{j=1}^{n} f(y_j|\boldsymbol{x}_j, \boldsymbol{\theta}) \tag{4.8}$$

her. Es soll nun davon ausgegangen werden, dass zusätzliche Informationen über den unbekannten Parametervektor $\boldsymbol{\theta}$ über die Gleichungen

$$\mathrm{E}(\boldsymbol{h}(\boldsymbol{X}, \boldsymbol{\theta})) = \mathbf{0} \tag{4.9}$$

vorliegen. Falls zum Beispiel Y eine dichotome Variable ist, so könnte die Information aus Gleichung (4.9) dergestalt sein, dass die Randwahrscheinlichkeiten für $Y = 0$ und $Y = 1$

$$P(Y = 0) = \mathrm{E}(f(0|\boldsymbol{X}, \boldsymbol{\theta})) \quad \text{und} \quad P(Y = 1) = \mathrm{E}(f(1|\boldsymbol{X}, \boldsymbol{\theta})) \tag{4.10}$$

bekannt sind. Wenn die Verteilung $G(\boldsymbol{x})$ unbekannt ist, lässt sich wieder die Methode der Empirischen Likelihood anwenden. Die Likelihood aus Gleichung (4.8) ist dann

$$L(\boldsymbol{\theta}, G) = \prod_{i=1}^{N} p_i \prod_{j=1}^{n} f(y_j|\boldsymbol{x}_j, \boldsymbol{\theta})$$

und diese wird unter den Nebenbedingungen

$$p_i \geq 0, \quad \sum_{i=1}^{n} p_i = 1, \quad \sum_{i=1}^{n} p_i \boldsymbol{h}(\boldsymbol{x_i}, \boldsymbol{\theta}) = \mathbf{0}$$

maximiert. Als Logarithmus der Semi-Empirischen Likelihood ergibt sich durch die üblichen Umformungen

$$\ln L(\boldsymbol{\theta}, G) = -\sum_{i=1}^{N} \ln\left(1 + \boldsymbol{\lambda}^\top \boldsymbol{h}(\boldsymbol{x_i}, \boldsymbol{\theta})\right) + \sum_{j=1}^{n} \ln f(y_j|\boldsymbol{x}_j, \boldsymbol{\theta}),$$

wobei $\boldsymbol{\lambda}$ die Gleichung

$$\frac{1}{n} \sum_{i=1}^{n} \frac{\boldsymbol{h}(\boldsymbol{x_i}, \boldsymbol{\theta})}{1 + \boldsymbol{\lambda}^\top \boldsymbol{h}(\boldsymbol{x_i}, \boldsymbol{\theta})} = \mathbf{0}$$

erfüllt. Qin (1992, Kap. 5.3) zeigt, dass diese Log-Likelihood asymptotisch zumindest ein lokales Maximum in einer Umgebung des wahren Parameters $\boldsymbol{\theta}_0$ besitzt. Qin

(1992, 2000) leitet die asymptotische Normalverteilung der MEL-Schätzer für $\boldsymbol{\theta}$ und $\boldsymbol{\lambda}$ her. Außerdem zeigt er die asymptotische χ^2-Verteilung des Empirischen Profil-Likelihood-Quotienten für den Erwartungswert μ_Y, der durch

$$R(\mu_Y) = \sup_{G,\boldsymbol{\theta}} \left\{ R(G, \boldsymbol{\theta}) \,\middle|\, \int \boldsymbol{h}(\boldsymbol{x}, \boldsymbol{\theta}) dG(\boldsymbol{x}) = \mu_Y \right\}$$

gegeben ist, wobei

$$R(G, \boldsymbol{\theta}) = \prod_{i=1}^{N} N dG(\boldsymbol{x}_i) \prod_{i=1}^{n} \frac{f(y_i|\boldsymbol{x}_i, \boldsymbol{\theta})}{f(y_i|\boldsymbol{x}_i, \widehat{\boldsymbol{\theta}})}.$$

Es sei an dieser Stelle angemerkt, dass sich die obige Methode nur für MAR-Situationen eignen, da der fehlende Daten generierende Prozess nicht modelliert wird. Außerdem ist ungewiss, ob bei nicht beobachtbarer Zielgröße brauchbare Information in Form der Gleichung (4.9) oder (4.10) vorliegt.

Ähnlich dazu stellen Qin und Zhang (2007) einen semi-parametrischen Empirischen Likelihood Ansatz für fehlende Daten vor, die ebenfalls MAR sind. Dazu modellieren sie die Wahrscheinlichkeit der Beobachtbarkeit der Zielgröße Y durch ein parametrisches Modell

$$P(R = 1|\boldsymbol{X} = \boldsymbol{x}) = w(\boldsymbol{x}, \boldsymbol{\theta}).$$

Da alle Beobachtungen der Kovariablen vorliegen, kann $\boldsymbol{\theta}$ problemlos geschätzt werden. Mit zusätzlichen Nebenbedingungen maximieren die Autoren anschließend die Empirische Likelihood, um damit den Erwartungswert der Zielgröße zu schätzen und zeigen, dass ihr Schätzer doppelt robust ist (zur doppelten Robustheit vgl. z. B. Tsiatis 2006, Kap. 10). Ähnlich dazu entwickeln Qin, Leung und Shao (2002) ein Modell zur Schätzung des Erwartungswerts einer teilweise nicht beobachtbaren Zufallsvariable Y, wobei das Fehlen hier nicht-ignorierbar ist. Zur Schätzung kann der vollständig beobachtbare Zufallsvektor $\boldsymbol{X}$ hinzugezogen werden. Dieses Vorgehen wird in Kapitel 5 ausführlich beschrieben.

Chen und Qin (2006) stellen einen weiteren EL-Ansatz zur Schätzung des Zusammenhangs zwischen einer binären Zielgröße und einer Kovariable mit fehlenden Klassenzuordnungen vor. Dabei ist das Fehlen ignorierbar und die neu entwickelten EL-Schätzer steigern die Effizienz verglichen mit dem gewöhnlichen ML-Schätzer der vollständigen Stichprobe. Qin, Zhang und Leung (2009) erarbeiten einen vereinheitlichten EL-Ansatz für verschiedene Problemstellungen mit fehlenden Daten. Dazu gehören Situationen mit fehlenden Kovariablen, Surrogate Response oder Regression mit Double-Sampling Designs.

4.5 Algorithmen

Bei der Maximierung der Empirischen Likelihood können verschiedene Optimierungsalgorithmen wie der Newton-Raphson- oder der Fisher-Scoring-Algorithmus verwendet werden. Verschiedene Aspekte sollten dabei beachtet werden. So sollte zum Beispiel in jedem Schritt des Optimierungsalgorithmus' überprüft werden, ob $p_i > 0$ $(i = 1, \ldots, n)$, da eine Verletzung dieser wichtigen Nebenbedingung auftreten kann. In diesem Fall sollte die Schrittlänge des Iterationsschritts verkürzt werden. Auch ist es möglich, einen verschachtelten Algorithmus zu verwenden, in dem die Likelihood abwechselnd über die Lagrangeparameter und die interessierenden Parameter minimiert bzw. maximiert wird. Dies bietet sich vor allem an, wenn es sehr viele Nebenbedingungen gibt. Ausführliche Details zu Algorithmen für die Empirische Likelihood finden sich auch bei Owen (2001, Kap. 12).

Eine zuverlässige Methode zur Maximierung Empirischer Likelihoods ist die Adjusted Empirical Likelihood (AEL, vgl. Chen, Variyath und Abraham 2008; Liu und Chen 2010). Diese ermöglicht eine sinnvolle Schätzung auch dann, wenn das betrachtete numerische Problem keine Lösung besitzt. Dies kann zum Beispiel auftreten, falls keine vernünftigen Startwerte für den Optimierungsalgorithmus vorliegen. Die durch die Schätzgleichungen aufgespannte konvexe Hülle überdeckt dann nicht die Null so dass die Gleichungen keine Lösung besitzen. Zur Bestimmung von $R(\boldsymbol{\theta})$ sucht man ein $\boldsymbol{\lambda}$, so dass

$$\frac{1}{n} \sum_{i=1}^{n} \frac{\boldsymbol{g}(\boldsymbol{x}_i, \boldsymbol{\theta})}{1 + \boldsymbol{\lambda}^\top \boldsymbol{g}(\boldsymbol{x}_i, \boldsymbol{\theta})} = \mathbf{0},$$

wobei gefordert wird, dass $1 + \boldsymbol{\lambda}^\top \boldsymbol{g}(\boldsymbol{x}_i, \boldsymbol{\theta}) > 0$ für $i = 1, \ldots, n$ (vgl. Gleichung (4.7), S. 27). Eine notwendige und hinreichende Bedingung für die Existenz einer solchen Lösung ist, dass die Null ein innerer Punkt der konvexen Hülle der Werte $\{\boldsymbol{g}(\boldsymbol{x}_i, \boldsymbol{\theta}), i = 1, \ldots, n\}$ ist. Für $n \to \infty$ gilt für den wahren Parameter $\boldsymbol{\theta}_0$, dass die Null in der konvexen Hülle der Werte $\{\boldsymbol{g}(\boldsymbol{x}_i, \boldsymbol{\theta}_0), i = 1, \ldots, n\}$ liegt. Falls aber n klein ist oder $\boldsymbol{\theta}$ nicht in der Nähe der wahren $\boldsymbol{\theta}_0$ liegt, besteht die Möglichkeit, dass keine Lösung existiert. Bei der AEL wird nun eine künstliche Beobachtung hinzugefügt, so dass die konvexe Hülle die Null überdeckt.

Sei $\boldsymbol{g}_i = \boldsymbol{g}(\boldsymbol{x}_i, \boldsymbol{\theta})$ und $\bar{\boldsymbol{g}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{g}_i$. Weiter sei

$$\boldsymbol{g}_{n+1} = -a_n \bar{\boldsymbol{g}}, \tag{4.11}$$

wobei a_n eine positive Konstante ist.

Die konvexe Hülle der Werte $\{\boldsymbol{g}_i, i = 1, \ldots, n, n+1\}$ überdeckt nun bei geeigneter Wahl von a_n die Null für beliebige $\boldsymbol{\theta}$, so dass der Adjusted Empirical Likelihood-Quotient

$$R^\star(\boldsymbol{\theta}) = \sup_{p_1,\ldots,p_{n+1}} \left\{ \prod_{i=1}^{n+1} (n+1) p_i \,\middle|\, p_i \geq 0, \sum_{i=1}^{n+1} p_i = 1, \sum_{i=1}^{n+1} p_i \boldsymbol{g}_i = \mathbf{0} \right\}. \tag{4.12}$$

wohldefiniert ist.

Chen, Variyath und Abraham (2008) zeigen, dass der Adjusted Empirical Likelihood-Quotient $R^\star(\boldsymbol{\theta})$ die asymptotischen Eigenschaften von $R(\boldsymbol{\theta})$ beibehält, falls $a_n = o_p\left(n^{2/3}\right)$ und stellen einen Algorithmus vor, der zuverlässig konvergiert. Als Empfehlung für die Konstante a_n geben die Autoren $a_n = \max(1, \log(n)/2)$ an. Liu und Chen (2010) zeigen, wie die Konstante a_n so gewählt wird, dass (4.12) stets wohl definiert ist und sich die Verteilung besser an eine χ^2-Verteilung anpasst.

Eine Veranschaulichung der Vorgehensweise bietet die Abbildung 4.1. Dort sind die Realisationen zweier Schätzgleichungen $g_1(\boldsymbol{x}_i, \boldsymbol{\theta})$ und $g_2(\boldsymbol{x}_i, \boldsymbol{\theta})$ für $i = 1, \ldots, 50$ abgetragen. Der zusätzliche Punkt ergibt sich als arithmetisches Mittel dieser Werte multipliziert mit $-a_n = -\log(50)/2 = 1.956$, so dass die konvexe Hülle erweitert wird und schließlich den Ursprung überdeckt.

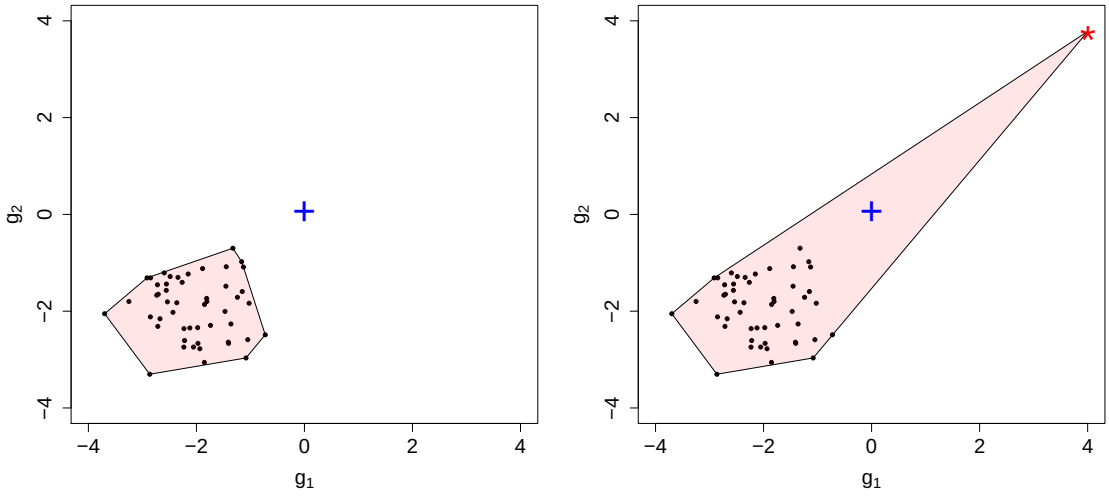

Abbildung 4.1 Konvexe Hülle (links) und angepasste konvexe Hülle (rechts) mit Nullpunkt (✚) und künstlichem Punkt (*) sowie $a_n = \log(n)/2$, in Anlehnung an Chen, Variyath und Abraham (2008).

Kapitel 5

Statistische Modelle mit nicht-ignorierbar fehlender Zielgröße

In diesem Kapitel wird ein Ansatz für die statistische Modellbildung in Situationen mit einer teilweise nicht beobachtbaren Zielgröße untersucht, wobei das Fehlen nicht-ignorierbar ist. In Kapitel 2 und 4 wurden bereits Ansätze zur Behandlung dieser Problematik in unterschiedlichen Spezialfällen vorgestellt. An dieser Stelle wird nun eine neue Methode entwickelt, welche es erlaubt, relativ allgemeine statistische Modelle mit fehlender Zielgröße zu schätzen. Diese beruht auf einem semi-empirischen Likelihood (SEL) Ansatz von Qin, Leung und Shao (2002) zur Schätzung des Erwartungswertes einer nicht-ignorierbar fehlenden Zufallsvariable, welcher hier zunächst erläutert wird. Es folgt eine kurze Beschreibung des Konzepts Generalisierter Linearer Modelle (GLM) mit einem Fokus auf der Maximum-Likelihood-Schätzung. Anschließend wird ein neuer Schätzer vorgeschlagen, der in Generalisierten Linearen Modellen mit fehlender Zielgröße angewendet werden kann. Dieser Schätzer erweist sich als konsistent und asymptotisch normalverteilt. Abschließend wird ein Hausman-Test vorgeschlagen, mit dem sich die Hypothese überprüfen lässt, ob eine abhängige Variable in einem Modell ignorierbar fehlt.

5.1 Schätzung des Erwartungswerts bei nicht-ignorierbar fehlenden Daten

Sei $Y \in \mathbb{R}$ eine Zufallsvariable und $\boldsymbol{X} \in \mathbb{R}^d$ ein Zufallsvektor von Kovariablen. Die Realisationen der Zielgröße Y können nicht vollständig beobachtet werden und die fehlenden Werte seien MNAR, d. h. die Wahrscheinlichkeit, dass diese Zielvariable fehlt, hängt von der Ausprägung der Zielvariablen selbst ab. Das Ziel ist, den Erwartungswert $\mathrm{E}(Y) =: \mu_Y$ konsistent zu schätzen.

Sei nun F die unbekannte Verteilungsfunktion von $(Y, \boldsymbol{X})$. Sei außerdem eine u.i.v. Stichprobe vom Umfang N gegeben, wobei die Realisation von Y bei $N - n$ Stichprobenvariablen fehlt. Ohne Beschränkung der Allgemeinheit nehmen wir dabei an, dass die ersten n Werte von Y beobachtbar sind, während die Werte $y_{n+1}, \ldots, y_N$ fehlen.

Das gewöhnliche arithmetische Mittel aus den vollständigen Beobachtungen ist nicht konsistent, da

$$\frac{1}{n} \sum_{i=1}^{n} y_i \xrightarrow{p} \mathrm{E}(Y|R = 1)$$

und dieser Grenzwert im Allgemeinen nicht identisch mit dem gesuchten Erwartungswert ist.

Die Zufallsvariable R gebe an, ob Y beobachtet wird oder nicht, das heißt

$$R = \begin{cases} 1, & \text{falls } Y \text{ beobachtbar} \\ 0, & \text{falls } Y \text{ nicht beobachtbar.} \end{cases}$$

Wie in Abschnitt 2.1 (S. 8) erwähnt, ist ein Wahrscheinlichkeitsmodell für den Prozess der fehlenden Daten $P(R|Y, \boldsymbol{X}, \boldsymbol{\theta})$ zu bestimmen, wobei die Abhängigkeit von Y im MNAR-Fall ausdrücklich erforderlich ist. Dies geschieht durch ein parametrisches Modell, z. B. durch ein Logistisches Regressionsmodell. Der Notation von Qin, Leung und Shao (2002) folgend bezeichnen wir dieses parametrische Modell allgemein als

$$w(Y, \boldsymbol{X}, \boldsymbol{\theta}) := P(R = 1|Y, \boldsymbol{X}, \boldsymbol{\theta}),$$

wobei w eine bekannte Funktion und $\boldsymbol{\theta}$ ein unbekannter Parametervektor sind. Dann ist die Likelihood von $(\boldsymbol{\theta}, F)$ bezüglich der Beobachtungen $\{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$

gegeben durch

$$L_n(\boldsymbol{\theta}, W, \boldsymbol{\lambda}_1, \lambda_2) = \prod_{i=1}^{n} w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) dF(y_i, \boldsymbol{x}_i) \cdot \prod_{i=n+1}^{N} \iint \left[1 - w(y, \boldsymbol{x}, \boldsymbol{\theta})\right] dF(y, \boldsymbol{x}). \tag{5.1}$$

Da die fehlenden Daten nicht in die Likelihood eingehen können, wird der nicht beobachtbare Teil der Realisationen von Y „herausintegriert". Das gleiche geschieht mit den zugehörigen Beobachtungen von $\boldsymbol{X}$. Die Likelihood ist die eines Selektionsmodells (vgl. Definition 2.5).

Es könnte auch die Likelihood $(\boldsymbol{\theta}, F)$ bezüglich aller beobachteten Werte der Stichprobe $\{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n), \boldsymbol{x}_{n+1}, \ldots, \boldsymbol{x}_N\}$

$$L_N(\boldsymbol{\theta}, W) = \prod_{i=1}^{n} w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) dF(y_i, \boldsymbol{x}_i) \cdot \prod_{i=n+1}^{N} \int \left[1 - w(y, \boldsymbol{x}_i, \boldsymbol{\theta})\right] dF(y, \boldsymbol{x}_i)$$

betrachtet werden, diese ist jedoch schwer zu behandeln.

Wir definieren nun die Randwahrscheinlichkeit für $\{R = 1\}$ als

$$W := P(R = 1) = \iint w(y, \boldsymbol{x}, \boldsymbol{\theta}) dF(y, \boldsymbol{x}).$$

Damit können wir nun die Likelihood aus Gleichung (5.1) umschreiben zu

$$\begin{aligned} L_n(\boldsymbol{\theta}, W, \boldsymbol{\lambda}_1, \lambda_2) &= \left[\prod_{i=1}^{n} w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) dF(y_i, \boldsymbol{x}_i)\right] \cdot (1 - W)^{N-n} \\ &= \left[\prod_{i=1}^{n} \frac{w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) dF(y_i, \boldsymbol{x}_i)}{W}\right] W^n (1 - W)^{N-n} \end{aligned} \tag{5.2}$$

Die zweite Darstellung der Likelihood in Gleichung (5.2) besteht aus zwei Termen. Der erste Teil stellt die auf $\{R = 1\}$ bedingte Likelihood dar und der zweite Teil entspricht der unbedingten Likelihood für die Anzahl fehlender Werte.

Ohne die Annahme einer expliziten parametrischen Verteilungsfamilie für die Verteilung F von $(Y, \boldsymbol{X})$ maximieren wir nun obige semiparametrische Likelihood (parametrischer Teil $w(Y, \boldsymbol{X}, \boldsymbol{\theta})$, nichtparametrischer Teil $F(y, \boldsymbol{x})$). Dazu bezeichnen wir mit $p_i := dF(y_i, \boldsymbol{x}_i)$ den Sprung der Verteilungsfunktion F an der Stelle $(y_i, \boldsymbol{x}_i)$ für $i = 1, \ldots, n$. Wir können die unbekannten Parameter $\boldsymbol{\theta}$ und W schätzen, indem wir die Likelihood aus Gleichung (5.2) maximieren unter den Nebenbedingungen

$$p_i \geq 0, \tag{5.3a}$$

$$\sum_{i=1}^{n} p_i = 1, \tag{5.3b}$$

$$\sum_{i=1}^{n} p_i(w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) - W) = 0, \tag{5.3c}$$

und

$$\sum_{i=1}^{n} p_i(\boldsymbol{x}_i - \boldsymbol{\mu}_{\boldsymbol{X}}) = \mathbf{0}, \tag{5.3d}$$

wobei $\mathrm{E}(\boldsymbol{X}) =: \boldsymbol{\mu}_{\boldsymbol{X}}$ als bekannt vorausgesetzt wird. Falls dieser Erwartungswert unbekannt ist, kann auch das arithmetische Mittel aller beobachteten $\boldsymbol{x}$-Werte $\bar{\boldsymbol{x}} = N^{-1} \sum_{i=1}^{N} \boldsymbol{x}_i$ anstelle von $\boldsymbol{\mu}_{\boldsymbol{X}}$ verwendet werden.

Die Bedeutung der Nebenbedingungen lässt sich folgendermaßen festhalten: Die ersten beiden Nebenbedingung (5.3a) und (5.3b) sind aus Kapitel 4 bekannt. Erstere stellt sicher, dass die Sprünge der Verteilungsfunktion F größer gleich Null sind und die Likelihood damit überhaupt positiv sein kann. Letztere garantiert, dass die Wahrscheinlichkeitsmasse vollständig auf der Stichprobe liegt. Nebenbedingung (5.3c) gewährleistet, dass

$$\sum_{i=1}^{n} w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) p_i = \sum_{i=1}^{n} P(R = 1 | y_i, \boldsymbol{x}_i) dF(y_i, \boldsymbol{x}_i) = P(R = 1) = W$$

gilt, dass also die Funktion w und die Wahrscheinlichkeit W die ihrer Definition entsprechende Eigenschaft erfüllen. Die letzte Nebenbedingung (5.3d) berücksichtigt, dass $\boldsymbol{\mu}_{\boldsymbol{X}}$ dem Erwartungswert $\mathrm{E}(\boldsymbol{X}) = \sum_{i=1}^{n} p_i \boldsymbol{x}_i$ der Verteilung F entspricht. Diese letzte Nebenbedingung ist die entscheidende zur Schätzung der unbekannten Parameter und gibt auch eine Idee, wie wir an gute Schätzungen gelangen. Die Abweichung der $\boldsymbol{x}$-Werte $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ der vollständigen Beobachtungspaare vom Erwartungswert $\boldsymbol{\mu}_{\boldsymbol{X}}$ bzw. dem arithmetischen Mittel aus allen $\boldsymbol{x}$-Werten $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ kann Aufschluss über die Lage der nicht beobachteten Realisationen $y_{n+1}, \ldots, y_N$ von Y geben, falls $\boldsymbol{X}$ und Y korreliert sind.

Wir lösen das Maximierungsproblem unter Nebenbedingungen nun mit Hilfe von Lagrange-Multiplikatoren, indem wir folgende Funktion bezüglich $\boldsymbol{\theta}$, W und p_i ($i =$

$1, \ldots, n)$ maximieren

$$\begin{aligned}
G_n =& \ln L_n(\boldsymbol{\theta}, W, \boldsymbol{\lambda}_1, \lambda_2) - \gamma_1 \left(\sum_{i=1}^{n} p_i - 1 \right) \\
& - \gamma_2 n \left(\sum_{i=1}^{n} p_i (w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) - W) \right) - \boldsymbol{\gamma}_3^\top n \left(\sum_{i=1}^{n} p_i (\boldsymbol{x}_i - \boldsymbol{\mu}_{\boldsymbol{X}}) \right) \\
=& \ln \left(\left[\prod_{i=1}^{n} \frac{w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) dF(y_i, \boldsymbol{x}_i)}{W} \right] W^n (1-W)^{N-n} \right) - \gamma_1 \left(\sum_{i=1}^{n} p_i - 1 \right) \\
& - \gamma_2 n \left(\sum_{i=1}^{n} p_i (w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) - W) \right) - \boldsymbol{\gamma}_3^\top n \left(\sum_{i=1}^{n} p_i (\boldsymbol{x}_i - \boldsymbol{\mu}_{\boldsymbol{X}}) \right) \\
=& \sum_{i=1}^{n} \ln \left(w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) \right) + \sum_{i=1}^{n} \ln p_i + (N-n) \ln (1-W) - \gamma_1 \left(\sum_{i=1}^{n} p_i - 1 \right) \\
& - \gamma_2 n \left(\sum_{i=1}^{n} p_i (w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) - W) \right) - \boldsymbol{\gamma}_3^\top n \left(\sum_{i=1}^{n} p_i (\boldsymbol{x}_i - \boldsymbol{\mu}_{\boldsymbol{X}}) \right).
\end{aligned}$$

Durch Ableiten nach p_i und anschließendes Nullsetzen erhält man

$$\begin{aligned}
& \frac{\partial G_n}{\partial p_i} = \frac{1}{p_i} - \gamma_1 - n\gamma_2 \left(w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) - W \right) - \boldsymbol{\gamma}_3^\top n \left(\boldsymbol{x}_i - \boldsymbol{\mu}_{\boldsymbol{X}} \right) = 0 \\
\Leftrightarrow \qquad & p_i = \frac{1}{\gamma_1 + \gamma_2 n \left(w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) - W \right) + \boldsymbol{\gamma}_3^\top n \left(\boldsymbol{x}_i - \boldsymbol{\mu}_{\boldsymbol{X}} \right)}.
\end{aligned}$$

Da außerdem

$$\begin{aligned}
n =& \sum_{i=1}^{n} p_i \frac{1}{p_i} = \sum_{i=1}^{n} p_i \left(\gamma_1 + \gamma_2 n \left(w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) - W \right) + \boldsymbol{\gamma}_3^\top n \left(\boldsymbol{x}_i - \boldsymbol{\mu}_{\boldsymbol{X}} \right) \right) \\
=& \gamma_1 \underbrace{\sum_{i=1}^{n} p_i}_{\overset{(5.3b)}{=} 1} + \gamma_2 n \underbrace{\sum_{i=1}^{n} p_i \left(w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) - W \right)}_{\overset{(5.3c)}{=} 0} + \boldsymbol{\gamma}_3^\top n \underbrace{\sum_{i=1}^{n} p_i \left(\boldsymbol{x}_i - \boldsymbol{\mu}_{\boldsymbol{X}} \right)}_{\overset{(5.3d)}{=} 0} = \gamma_1,
\end{aligned}$$

gilt

$$p_i = \frac{1}{n \left[1 + \boldsymbol{\lambda}_1^\top \left(\boldsymbol{x}_i - \boldsymbol{\mu}_{\boldsymbol{X}} \right) + \lambda_2 \left(w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) - W \right) \right]}, \tag{5.4}$$

wobei $\gamma_2 = \lambda_2$ und $\boldsymbol{\gamma}_3^\top = \boldsymbol{\lambda}_1$. Setzen wir nun (5.4) in die Likelihood (5.2) ein, so erhalten wir

$$L_n(\boldsymbol{\theta}, W, \boldsymbol{\lambda}_1, \lambda_2) = \left[\prod_{i=1}^{n} \frac{w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta})}{n \left[1 + \boldsymbol{\lambda}_1^\top \left(\boldsymbol{x}_i - \boldsymbol{\mu}_{\boldsymbol{X}} \right) + \lambda_2 \left(w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) - W \right) \right]} \right] \cdot (1-W)^{N-n}. \tag{5.5}$$

Die Log-Likelihood ist damit

$$\begin{aligned}\ln L_n(\boldsymbol{\theta}, W, \boldsymbol{\lambda}_1, \lambda_2) = &\sum_{i=1}^{n} \ln w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) + (N-n)\ln(1-W) - n\ln n \\ &- \sum_{i=1}^{n} \ln\left[1 + \boldsymbol{\lambda}_1^\top (\boldsymbol{x}_i - \boldsymbol{\mu}_{\boldsymbol{X}}) + \lambda_2 \left(w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) - W\right)\right].\end{aligned} \tag{5.6}$$

Leiten wir diese nach W, $\boldsymbol{\lambda}_1$, λ_2 und nach $\boldsymbol{\theta}$ ab und setzten die partiellen Ableitungen gleich Null, erhalten wir

$$\begin{aligned}\frac{\partial \ln L_n(\boldsymbol{\theta}, W, \boldsymbol{\lambda}_1, \lambda_2)}{\partial W} = &-\frac{N-n}{1-W} \\ &+ \sum_{i=1}^{n} \frac{\lambda_2}{1 + \boldsymbol{\lambda}_1^\top (\boldsymbol{x}_i - \boldsymbol{\mu}_{\boldsymbol{X}}) + \lambda_2 \left(w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) - W\right)} = 0,\end{aligned} \tag{5.7a}$$

$$\frac{\partial \ln L_n(\boldsymbol{\theta}, W, \boldsymbol{\lambda}_1, \lambda_2)}{\partial \boldsymbol{\lambda}_1} = -\sum_{i=1}^{n} \frac{\boldsymbol{x}_i - \boldsymbol{\mu}_{\boldsymbol{X}}}{1 + \boldsymbol{\lambda}_1^\top (\boldsymbol{x}_i - \boldsymbol{\mu}_{\boldsymbol{X}}) + \lambda_2 \left(w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) - W\right)} = \mathbf{0}, \tag{5.7b}$$

$$\frac{\partial \ln L_n(\boldsymbol{\theta}, W, \boldsymbol{\lambda}_1, \lambda_2)}{\partial \lambda_2} = -\sum_{i=1}^{n} \frac{w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) - W}{1 + \boldsymbol{\lambda}_1^\top (\boldsymbol{x}_i - \boldsymbol{\mu}_{\boldsymbol{X}}) + \lambda_2 \left(w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) - W\right)} = 0 \tag{5.7c}$$

und

$$\begin{aligned}\frac{\partial \ln L_n(\boldsymbol{\theta}, W, \boldsymbol{\lambda}_1, \lambda_2)}{\partial \boldsymbol{\theta}} = &\sum_{i=1}^{n} \frac{\partial \ln w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \\ &- \sum_{i=1}^{n} \frac{\lambda_2 \partial w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta})/\partial \boldsymbol{\theta}}{1 + \boldsymbol{\lambda}_1^\top (\boldsymbol{x}_i - \boldsymbol{\mu}_{\boldsymbol{X}}) + \lambda_2 \left(w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) - W\right)} = \mathbf{0}.\end{aligned} \tag{5.7d}$$

Aus (5.7a), (5.4) und der Nebenbedingung (5.3b) folgt

$$\begin{aligned}& \frac{N-n}{1-W} - \sum_{i=1}^{n} \frac{\lambda_2}{1 + \boldsymbol{\lambda}_1^\top (\boldsymbol{x}_i - \boldsymbol{\mu}_{\boldsymbol{X}}) + \lambda_2 \left(w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) - W\right)} = 0 \\ \Leftrightarrow \quad & \frac{N-n}{1-W} - \lambda_2 \sum_{i=1}^{n} n p_i = 0 \\ \Leftrightarrow \quad & \frac{N-n}{1-W} - \lambda_2 n = 0 \\ \Leftrightarrow \quad & \lambda_2 = \frac{N-n}{n(1-W)}.\end{aligned} \tag{5.8}$$

Zur Parameterschätzung müssen nun die Gleichungen (5.7b), (5.7c) und (5.7d) gelöst werden. Wir bezeichnen die resultierenden Schätzer mit $\widehat{\lambda}_1$, $\widehat{W}$ und $\widehat{\boldsymbol{\theta}}$ und $\widehat{\lambda}_2 = (N-n)/(n(1-\widehat{W}))$. Eine geschlossene Form kann nicht hergeleitet werden, weshalb die Bestimmung der Nullstellen numerisch erfolgt (zur Maximierung der Likelihood und Bestimmung der Schätzer vgl. Anhang A). Die wahren Werte der Parameter $\boldsymbol{\theta}$ und W bezeichnen wir im Folgenden mit $\boldsymbol{\theta}_0$ und W_0. Qin, Leung und Shao (2002) zeigen, dass die resultierenden Schätzer konsistent und asymptotisch normalverteilt sind. Damit lassen sich Tests für die Parameter konstruieren. Die Beweise von Qin, Leung und Shao (2002) werden an dieser Stelle kurz skizziert, da sie für die späteren Überlegungen nützlich sind.

Satz 5.1 (Qin, Leung und Shao 2002, Theorem 1):

Sei $\widehat{\boldsymbol{\gamma}} = \widehat{\boldsymbol{\lambda}}_1(1-\widehat{W})$. Sei weiter die Verteilung $F(y, \boldsymbol{x})$ nicht degeneriert und $w(y, \boldsymbol{x}, \boldsymbol{\theta}) > 0$ für alle $(y, \boldsymbol{x})$ aus dem Wertebereich von F. In einer Umgebung von $\boldsymbol{\theta}_0$ sei außerdem $\mathrm{E}\left(|X|^3 w(Y, \boldsymbol{X})\right) < \infty$, $\mathrm{E}\left(|w(Y, \boldsymbol{X})|^3\right) < \infty$ und $\partial^2 w(y, \boldsymbol{x}, \boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top$ existiere und sei beschränkt durch eine integrierbare Funktion. Dann existiert eine Folge $\{\widehat{\boldsymbol{\theta}}, \widehat{W}, \widehat{\boldsymbol{\gamma}}\}_N$, so dass für $N \to \infty$ gilt:

$$P\left(\widehat{\boldsymbol{\theta}}, \widehat{W} \text{ und } \widehat{\boldsymbol{\gamma}} \text{ sind Lösungen der Gleichungen (5.7b)-(5.7d) und (5.8)}\right) \to 1 \tag{5.9}$$

und

$$\sqrt{N}\begin{pmatrix} \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 \\ \widehat{W} - W_0 \\ \widehat{\boldsymbol{\gamma}} - \mathbf{0} \end{pmatrix} \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}), \tag{5.10}$$

wobei $\boldsymbol{\Sigma}$ im Folgenden näher spezifiziert wird.

Beweis: Zunächst werden einige Notationen benötigt, die auch in späteren Beweisen Verwendung finden. Sei

$$\begin{aligned} \boldsymbol{\gamma} &= \boldsymbol{\lambda}_1(1-W), & a_N &= \tfrac{N}{n} - \tfrac{1}{W_0}, \\ \boldsymbol{\eta} &= (\boldsymbol{\theta}^\top, W, \boldsymbol{\gamma}^\top)^\top, & \boldsymbol{\eta}_0 &= (\boldsymbol{\theta}_0^\top, W_0, 0)^\top, \end{aligned}.$$

Offenbar gilt

$$\begin{aligned} &1 + \boldsymbol{\lambda}_1^\top(\boldsymbol{x} - \boldsymbol{\mu}_{\boldsymbol{X}}) + \lambda_2\left(w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) - W\right) \\ =&\frac{1}{1-W}\Bigg[1 - \frac{W}{W_0} + \left(\frac{1}{W_0} - 1\right) w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) \\ &\qquad + \boldsymbol{\gamma}^\top(\boldsymbol{x} - \boldsymbol{\mu}_{\boldsymbol{X}}) + a_N\left(w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) - W\right)\Bigg]. \end{aligned}$$

Sei weiter

$$\begin{aligned}
&\boldsymbol{g}_i^1(\boldsymbol{\eta}, a_N) \\
=&\frac{\partial \ln w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \\
&- \frac{\left[a_N + \frac{1-W_0}{W_0}\right] \frac{\partial w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}}{1 - \frac{W}{W_0} + \left(\frac{1}{W_0} - 1\right) w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) + \boldsymbol{\gamma}^\top (\boldsymbol{x} - \boldsymbol{\mu}_{\boldsymbol{X}}) + a_N \left(w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) - W\right)},
\end{aligned}$$

$$\begin{aligned}
&g_i^2(\boldsymbol{\eta}, a_N) \\
=&\frac{w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) - W}{1 - \frac{W}{W_0} + \left(\frac{1}{W_0} - 1\right) w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) + \boldsymbol{\gamma}^\top (\boldsymbol{x} - \boldsymbol{\mu}_{\boldsymbol{X}}) + a_N \left(w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) - W\right)},
\end{aligned}$$

und

$$\begin{aligned}
&\boldsymbol{g}_i^3(\boldsymbol{\eta}, a_N) \\
=&\frac{\boldsymbol{x}_i - \boldsymbol{\mu}_{\boldsymbol{X}}}{1 - \frac{W}{W_0} + \left(\frac{1}{W_0} - 1\right) w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) + \boldsymbol{\gamma}^\top (\boldsymbol{x} - \boldsymbol{\mu}_{\boldsymbol{X}}) + a_N \left(w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) - W\right)}
\end{aligned}$$

sowie

$$\boldsymbol{g}(\boldsymbol{\eta}, a_N) = \sum_{i=1}^{n} \boldsymbol{g}_i(\boldsymbol{\eta}, a_N) = \sum_{i=1}^{n} \left(\boldsymbol{g}_i^1(\boldsymbol{\eta}, a_N)^\top, g_i^2(\boldsymbol{\eta}, a_N), \boldsymbol{g}_i^3(\boldsymbol{\eta}, a_N)^\top\right)^\top$$

Die Gleichungen (5.7b)-(5.7d) sind dann äquivalent zu

$$\boldsymbol{g}(\boldsymbol{\eta}, a_N) = \mathbf{0}.$$

Qin, Leung und Shao (2002) zeigen nun, dass für Werte $\left(\boldsymbol{\theta}^\top, W\right)^\top$ aus der Menge

$$B = \left\{ \left(\boldsymbol{\theta}^\top, W\right)^\top : \left\| \left(\boldsymbol{\theta}^\top, W\right)^\top - \left(\boldsymbol{\theta}_0^\top, W_0\right)^\top \right\| = N^{-\frac{1}{3}} \right\}$$

fast sicher

$$\ln L_n(\boldsymbol{\theta}, W, \boldsymbol{\lambda}_1, \lambda_2) < \ln L_n(\boldsymbol{\theta}_0, W_0, \boldsymbol{\lambda}_1, \lambda_2)$$

gilt.

Mit der Stetigkeit und Differenzierbarkeit von $\ln L_n(\boldsymbol{\theta}, W, \boldsymbol{\lambda}_1, \lambda_2)$ folgt, dass die Log-Likelihood ein lokales Maximum innerhalb der Kugel mit der Oberfläche B hat. Daraus folgt (5.9). Da B für $N \to \infty$ gegen $\left(\boldsymbol{\theta}_0^\top, W_0\right)^\top$ konvergiert, folgt die Konsistenz der Parameterschätzer.

Die Normalverteilung von $\boldsymbol{\eta}$ wird über eine Taylorentwicklung hergeleitet. Zunächst gilt

$$\boldsymbol{g}(\boldsymbol{\eta}_0, 0) + \frac{\partial \boldsymbol{g}(\boldsymbol{\eta}_0, 0)}{\partial \boldsymbol{\eta}} (\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0) + \frac{\partial \boldsymbol{g}(\boldsymbol{\eta}_0, 0)}{\partial a_N} (a_N - 0) + o_p\left(N^{\frac{1}{2}}\right) = \boldsymbol{g}(\widehat{\boldsymbol{\eta}}, a_N) = \mathbf{0}.$$

Damit folgt

$$\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0 = \boldsymbol{U}_N^{-1} \boldsymbol{\psi}_N + o_p\left(N^{-\frac{1}{2}}\right),$$

wobei

$$\boldsymbol{U}_N = -\frac{\partial \boldsymbol{g}(\boldsymbol{\eta}_0, 0)}{\partial \boldsymbol{\eta}} \qquad \text{und} \qquad \boldsymbol{\psi}_N = \boldsymbol{g}(\boldsymbol{\eta}_0, 0) + \frac{\partial \boldsymbol{g}(\boldsymbol{\eta}_0, 0)}{\partial a_N} (a_N - 0).$$

Es folgt mit dem Gesetz der Großen Zahlen

$$\frac{1}{N} \boldsymbol{U}_N \xrightarrow{p} \boldsymbol{U},$$

wobei

$$\boldsymbol{U} = W_0 \cdot \mathrm{E}_{Y, \boldsymbol{X} | R=1}(\boldsymbol{T})$$

mit

$$\boldsymbol{T} = \begin{pmatrix} \mathbf{0} & \frac{\partial w(Y, \boldsymbol{X}, \boldsymbol{\theta}_0)/\partial \boldsymbol{\theta}}{(1-W_0) w^2(Y, \boldsymbol{X}, \boldsymbol{\theta}_0)} & -\frac{W_0 (\boldsymbol{X} - \boldsymbol{\mu}_{\boldsymbol{X}}) \partial w(Y, \boldsymbol{X}, \boldsymbol{\theta}_0)/\partial \boldsymbol{\theta}}{(1-W_0) w^2(Y, \boldsymbol{X}, \boldsymbol{\theta}_0)} \\ -\frac{W_0^2 \partial w(Y, \boldsymbol{X}, \boldsymbol{\theta}_0)/\partial \boldsymbol{\theta}}{(1-W_0) w^2(Y, \boldsymbol{X}, \boldsymbol{\theta}_0)} & -\frac{W_0^2 (w(Y, \boldsymbol{X}, \boldsymbol{\theta}_0) - 1)}{(1-W_0)^2 w^2(Y, \boldsymbol{X}, \boldsymbol{\theta}_0)} & \frac{W_0^2 (\boldsymbol{X} - \boldsymbol{\mu}_{\boldsymbol{X}})(w(Y, \boldsymbol{X}, \boldsymbol{\theta}_0) - W_0)}{(1-W_0)^2 w^2(Y, \boldsymbol{X}, \boldsymbol{\theta}_0)} \\ \frac{W_0 (\boldsymbol{X} - \boldsymbol{\mu}_{\boldsymbol{X}}) \partial w(Y, \boldsymbol{X}, \boldsymbol{\theta}_0)/\partial \boldsymbol{\theta}}{(1-W_0) w^2(Y, \boldsymbol{X}, \boldsymbol{\theta}_0)} & -\frac{W_0 (\boldsymbol{X} - \boldsymbol{\mu}_{\boldsymbol{X}})}{(1-W_0)^2 w^2(Y, \boldsymbol{X}, \boldsymbol{\theta}_0)} & \frac{W_0^2 (\boldsymbol{X} - \boldsymbol{\mu}_{\boldsymbol{X}})^2}{(1-W_0)^2 w^2(Y, \boldsymbol{X}, \boldsymbol{\theta}_0)} \end{pmatrix}$$

Weiter gilt

$$\frac{1}{N} \frac{\partial \boldsymbol{g}(\boldsymbol{\eta}_0, 0)}{\partial a_N} \xrightarrow{p} \boldsymbol{h}$$

mit

$$\boldsymbol{h} = -\frac{W_0^3}{(1-W_0)^2} \mathrm{E}_{Y, \boldsymbol{X} | R=1} \begin{pmatrix} \frac{1 - W_0}{(w(Y, \boldsymbol{X}, \boldsymbol{\theta}_0))^2} \frac{\partial w(Y, \boldsymbol{X}, \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \\ \frac{(w(Y, \boldsymbol{X}, \boldsymbol{\theta}_0) - W_0)^2}{(w(Y, \boldsymbol{X}, \boldsymbol{\theta}_0))^2} \\ \frac{(\boldsymbol{X} - \boldsymbol{\mu}_{\boldsymbol{X}})(w(Y, \boldsymbol{X}, \boldsymbol{\theta}_0) - W_0)}{(w(Y, \boldsymbol{X}, \boldsymbol{\theta}_0))^2} \end{pmatrix}.$$

Dann folgt mit dem Zentralen Grenzwertsatz

$$\begin{aligned}
&\sqrt{N}\frac{1}{N}\boldsymbol{\psi}_N \\
=&\sqrt{N}\frac{1}{N}\left(\boldsymbol{g}(\boldsymbol{\eta}_0,0)+\frac{\partial \boldsymbol{g}(\boldsymbol{\eta}_0,0)}{\partial a_N}a_N\right) \\
=&\sqrt{N}\frac{1}{N}\sum_{i=1}^{N}\left[r_i\boldsymbol{g}_i(\boldsymbol{\eta}_0,0)+\boldsymbol{h}\left(\frac{1}{W_0}-\frac{r_i}{W_0^2}\right)\right]+o_p\left(1\right)\xrightarrow{p}\mathcal{N}(\mathbf{0},\boldsymbol{V}),
\end{aligned}$$

da

$$\begin{aligned}
a_N &= \frac{N}{n}-\frac{1}{W_0}=\frac{N}{n}\left(1-\frac{n}{W_0N}\right)=\left(\frac{N}{n}-\frac{1}{W_0}+\frac{1}{W_0}\right)\frac{1}{N}\sum_{i=1}^{N}\left(1-\frac{r_i}{W_0}\right) \\
&= \frac{1}{W_0}\frac{1}{N}\sum_{i=1}^{N}\left(1-\frac{r_i}{W_0}\right)+\underbrace{\left(\frac{N}{n}-\frac{1}{W_0}\right)}_{=o_p(1)}\underbrace{\frac{1}{N}\sum_{i=1}^{N}\left(1-\frac{r_i}{W_0}\right)}_{=O_p\left(N^{-1/2}\right)} \\
&= \frac{1}{W_0}\frac{1}{N}\sum_{i=1}^{N}\left(1-\frac{r_i}{W_0}\right)+o_p\left(N^{-\frac{1}{2}}\right).
\end{aligned}$$

Dabei ist

$$\boldsymbol{V}=\operatorname{Var}\left(R\boldsymbol{g}_i(\boldsymbol{\eta}_0,0)+\boldsymbol{h}\left(\frac{1}{W_0}-\frac{R}{W_0^2}\right)\right).$$

Schließlich folgt wiederum mit dem Zentralen Grenzwertsatz und dem Slutzky Theorem

$$\sqrt{N}\left(\widehat{\boldsymbol{\eta}}-\boldsymbol{\eta}_0\right)=\left(\frac{1}{N}\boldsymbol{U}_N\right)^{-1}\sqrt{N}\frac{1}{N}\boldsymbol{\psi}_N+o_p\left(1\right)\xrightarrow{p}\mathcal{N}(\mathbf{0},\boldsymbol{\Sigma}),$$

wobei

$$\boldsymbol{\Sigma}=\boldsymbol{U}^{-1}\boldsymbol{V}\left(\boldsymbol{U}^{-1}\right)^{\top}.$$

Falls $\boldsymbol{\mu_X}$ unbekannt ist und durch das arithmetische Mittel $\bar{\boldsymbol{X}}$ geschätzt wird, ergibt sich die asymptotische Normalverteilung analog, indem man in $\boldsymbol{g}_i$ und $\boldsymbol{h}$ jeweils $\boldsymbol{\mu_X}$ durch $\bar{\boldsymbol{X}}$ ersetzt. Dann hat $\boldsymbol{V}$ die Form

$$\boldsymbol{V}=\operatorname{Var}\left(R\boldsymbol{g}_i(\boldsymbol{\eta}_0,0)+\boldsymbol{h}\left(\frac{1}{W_0}-\frac{R}{W_0^2}\right)+\left(\mathbf{0}^{\top},0,\frac{W_0(\boldsymbol{\mu_X}-\boldsymbol{X})^{\top}}{1-W_0}\right)^{\top}\right).$$

Der zusätzliche Summand in $\boldsymbol{V}$ entsteht dadurch, dass

$$\boldsymbol{g}_i^3(\boldsymbol{\eta}_0,0)=\frac{W_0}{1-W_0}\frac{\boldsymbol{x}_i-\bar{\boldsymbol{x}}}{w(y_i,\boldsymbol{x}_i,\boldsymbol{\theta}_0)}=\frac{W_0}{1-W_0}\frac{\boldsymbol{x}_i-\boldsymbol{\mu_X}-(\bar{\boldsymbol{x}}-\boldsymbol{\mu_X})}{w(y_i,\boldsymbol{x}_i,\boldsymbol{\theta}_0)}$$

und damit

$$\begin{aligned}
&N^{-\frac{1}{2}} \sum_{i=1}^{N} r_i \boldsymbol{g}_i^3(\boldsymbol{\eta}_0, 0) \\
=&N^{-\frac{1}{2}} \frac{W_0}{1-W_0} \sum_{i=1}^{N} r_i \frac{\boldsymbol{x}_i - \bar{\boldsymbol{x}}}{w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}_0)} - N^{-\frac{1}{2}} \frac{W_0}{1-W_0} \sum_{i=1}^{N} r_i \frac{\bar{\boldsymbol{x}} - \boldsymbol{\mu}_{\boldsymbol{X}}}{w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}_0)}.
\end{aligned}$$

Mit den Schritten

$$\begin{aligned}
\frac{1}{N} \sum_{i=1}^{N} \frac{r_i}{w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}_0)} &= \frac{n}{N} \frac{1}{n} \sum_{i=1}^{n} \frac{1}{w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}_0)} \\
&\xrightarrow{p} W_0 \cdot \mathrm{E}_{Y, \boldsymbol{X}|R=1} \left(\frac{1}{w(Y, \boldsymbol{X}, \boldsymbol{\theta}_0)} \right) \\
&= W_0 \cdot \int \frac{1}{w(y, \boldsymbol{x}, \boldsymbol{\theta}_0)} \frac{w(y, \boldsymbol{x}, \boldsymbol{\theta}_0)}{W_0} dF(y, \boldsymbol{x}) \\
&= W_0 \cdot \frac{1}{W_0} \\
&= 1
\end{aligned}$$

und

$$N(\bar{\boldsymbol{x}} - \boldsymbol{\mu}_{\boldsymbol{X}}) = \sum_{i=1}^{N} (\boldsymbol{x}_i - \boldsymbol{\mu}_{\boldsymbol{X}})$$

folgt schließlich

$$\begin{aligned}
&N^{-\frac{1}{2}} \sum_{i=1}^{N} r_i \boldsymbol{g}_i^3(\boldsymbol{\eta}_0, 0) \\
=&N^{-\frac{1}{2}} \frac{W_0}{1-W_0} \sum_{i=1}^{N} r_i \frac{\boldsymbol{x}_i - \bar{\boldsymbol{x}}}{w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}_0)} - N^{-\frac{1}{2}} \frac{W_0}{1-W_0} \sum_{i=1}^{N} (\boldsymbol{x}_i - \boldsymbol{\mu}_{\boldsymbol{X}}) + o_p(1).
\end{aligned}$$

■

Satz 5.1 erlaubt es also unter anderem, den Parametervektor $\boldsymbol{\theta}$ des parametrischen Wahrscheinlichkeitsmodells für das Fehlen von Y konsistent zu schätzen. Allerdings muss dabei beachtet werden, dass das Modell nicht überparametrisiert wird, wie folgende Bemerkung deutlich macht.

Bemerkung 5.1:

Ein Defizit des hier angegeben Modells ist, dass in $w(y, \boldsymbol{x}, \boldsymbol{\theta})$ nur ein Parametervektor $\boldsymbol{\theta}$ der Länge $(d+1)$ geschätzt werden kann. Ansonsten ist die Anzahl der freien

Parameter größer als die Anzahl der Nebenbedingungen (vgl. Gleichung (5.3c) und (5.3d)). W ist kein freier Parameter sondern durch die Nebenbedingung (5.3c) und $\boldsymbol{\theta}$ festgelegt. Das bedeutet, dass zum Beispiel im Falle eines logistischen Regressionsmodells neben der Konstante und dem Parameter für Y nur $(d-1)$ weitere Parameter in dem Modell geschätzt werden können, also einer weniger als Komponenten in $\boldsymbol{X}$.

Das ursprüngliche Ziel von Qin, Leung und Shao (2002) ist die Schätzung von $\mathrm{E}(Y) = \mu_Y$. Als Schätzer schlagen sie

$$\widehat{\mu}_Y^{\mathrm{SEL}} := \sum_{i=1}^{n} \widehat{p}_i y_i$$

vor, wobei

$$\widehat{p}_i = \frac{1}{n\left[1 + \widehat{\boldsymbol{\lambda}}_1^{\top}\left(\boldsymbol{x}_i - \boldsymbol{\mu}_{\boldsymbol{X}}\right) + \widehat{\lambda}_2\left(w(y_i, \boldsymbol{x}_i, \widehat{\boldsymbol{\theta}}) - \widehat{W}\right)\right]}.$$

Der Schätzer entspricht damit dem gewogenen arithmetischen Mittel der Beobachtungen von Y, gewichtet mit den Werten der geschätzten empirischen Wahrscheinlichkeitsfunktion der Stichprobe. Zur Verteilung des Schätzers liefern die Autoren den folgenden Satz.

Satz 5.2 (Qin, Leung und Shao 2002, Theorem 2):

Unter den Voraussetzungen von Satz 5.1 und falls $\mathrm{E}(Y^2/w(Y, X, \boldsymbol{\theta}_0)) < \infty$ gilt

$$\sqrt{N}(\widehat{\mu}_Y^{\mathrm{SEL}} - \mu_Y) \xrightarrow{d} \mathcal{N}(0, \sigma^2).$$

Beweis: Für den Beweis und den Ausdruck für σ^2 vgl. Qin, Leung und Shao (2002).

■

Qin, Leung und Shao (2002) untersuchen in ihrer Arbeit mit Hilfe einer Simulationsstudie die Güte des Schätzers für $\boldsymbol{\theta}$ sowie die von $\widehat{\mu}_Y^{\mathrm{SEL}}$. Sie betrachten dafür unterschiedliche Situationen fehlender Daten sowie verschiedene Modelle für den Zusammenhang zwischen Y und $\boldsymbol{X}$.

Letztere unterscheiden sich maßgeblich in der Stärke der Korrelation von $\boldsymbol{X}$ und Y, wobei unter anderem ein Modell perfekter Korrelation zwischen Zielgröße und Kovariablen betrachtet wird, aber auch eines vollkommener Unabhängigkeit.

Die Autoren gehen in ihrer Simulation davon aus, dass das Fehlen der Beobachtungen nicht von $\boldsymbol{X}$ sondern lediglich von Y abhängt. Als Modell w für das Fehlen betrachten sie

$$w(y, \boldsymbol{x}, \boldsymbol{\theta}) = \frac{\exp(\theta_1 + \theta_2 y)}{1 + \exp(\theta_1 + \theta_2 y)}.$$

Sie wählen $\theta_2 = -0.2$ und $\theta_1 \in \{3, 2, 1, -1\}$, wobei θ_1 als bekannt vorausgesetzt wird und somit lediglich θ_2 zu schätzen ist. Wir werden im Folgenden sehen, dass diese Vorgehensweise problematisch ist.

Qin, Leung und Shao (2002) simulieren für die vorgeschlagenen Modelle Daten und schätzen θ_2 und μ_Y sowie die Verzerrung und die Varianz der Schätzer. Sie stellen fest, dass selbst bei vollkommen unkorrelierter Zielgröße und Kovariablen eine konsistente Schätzung möglich ist. Allerdings berücksichtigen sie dabei nicht, dass sie durch die Kenntnis von θ_1 unbeabsichtigt Information über die fehlenden Werte von Y einfließen lassen, die in der Regel nicht vorhanden sein dürfte.

Falls beide Parameter θ_1 und θ_2 unbekannt sind und geschätzt werden sollen, führt dies im unkorrelierten Fall ($\mathrm{Cor}(Y, \boldsymbol{X}) = \mathbf{0}$) bei endlichen Stichproben zu einer sehr schwachen Schätzung. Dies lässt sich leicht anhand einer eigenen kleinen Simulation verdeutlichen. Seien X und Y unabhängig standardnormalverteilt und

$$R|Y = y, X = x \sim \mathcal{B}(1, \pi) \quad \text{mit } \pi = \frac{\exp(\theta_1 + \theta_2 y)}{1 + \exp(\theta_1 + \theta_2 y)},$$

wobei $\theta_1 = 0$ und $\theta_2 = -3$. Nun werden 1 000 Beobachtungen dieser Zufallsvariablen generiert und anschließend die Parameter geschätzt. Dies erfolgt 1 000 mal. Im Mittel ergibt sich als geschätzte Verzerrung $\widehat{\mathrm{Bias}}(\widehat{\theta_1}) = 3.426$ ($\widehat{\mathrm{Var}}(\widehat{\theta_1}) = 996.520$) und $\widehat{\mathrm{Bias}}(\widehat{\theta_2}) = 3.059$ ($\widehat{\mathrm{Var}}(\widehat{\theta_2}) = 216.980$). Wir sehen also, dass die Schätzer in diesem Fall verzerrt sind und zudem eine große Streuung besitzen.

Auch aus Plausibilitätsgründen ist davon auszugehen, dass eine gute Schätzung der Parameter nur möglich ist, wenn Y und $\boldsymbol{X}$ korreliert sind. Da keine Information über die fehlenden Werte von Y vorliegt, kann diese nur über die Ausprägungen von $\boldsymbol{X}$ bezogen werden. Dies geschieht durch den Erwartungswert $\mathrm{E}(\boldsymbol{X})$ in Nebenbedingung (5.3d). Aus diesem Grund wird die Güte der Schätzung bei geringer Korrelation von Y und $\boldsymbol{X}$ unbefriedigend sein.

Qin, Leung und Shao (2002) haben letztendlich gezeigt, dass die geschätzten Werte der empirischen Wahrscheinlichkeitsfunktion der Stichprobe

$$\widehat{p_i} = \frac{1}{n\left[1 + \widehat{\boldsymbol{\lambda}}_1^\top (\boldsymbol{x}_i - \boldsymbol{\mu}_{\boldsymbol{X}}) + \widehat{\lambda}_2 \left(w(y_i, \boldsymbol{x}_i, \widehat{\boldsymbol{\theta}}) - \widehat{W}\right)\right]}$$

als Gewichte bei der konsistenten Schätzung des Erwartungswerts von Y unter gewissen Voraussetzungen nützlich sind. Die Wahrscheinlichkeitsmasse p_i der einzelnen Beobachtungen lässt sich jedoch auch in zahlreichen anderen statistischen Problemstellungen nutzen. Eine Schätzung der empirischen Verteilungsfunktion von Y beispielsweise (vgl. Satz 4.3) ist gegeben durch

$$\widehat{F}_n^{\mathrm{EL}}(y) = \sum_{i=1}^{n} \widehat{p}_i \mathbb{1}_{(-\infty,y]}(y_i).$$

Falls Y stetig ist, könnte auch eine Schätzung der zugehörigen Dichte von Interesse sein. Dies könnte durch einen gewichteten Kerndichteschätzer geschehen (vgl. Chen 1997). Ein solcher Kerndichteschätzer $\widehat{f}$ mit Kernfunktion K und Bandbreite h für eine u.i.v. Stichprobe stetiger Zufallsvariablen $y_1, \ldots, y_n$ ergibt sich dann, indem beim üblichen Schätzer

$$\widehat{f}(y) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{y - y_i}{h}\right)$$

die gleichmäßige Gewichtung für die einzelnen Beobachtungen durch die geschätzten Gewichte $\widehat{p}_i$ ersetzt werden. Dann erhält man

$$\widehat{f}^{\mathrm{EL}}(y) = \frac{1}{h} \sum_{i=1}^{n} \widehat{p}_i K\left(\frac{y - y_i}{h}\right).$$

Chen (1997) zeigt, dass die Varianz von $\widehat{f}^{\mathrm{EL}}$ unter gewissen Voraussetzungen geringer als die des gewöhnlichen Kerndichteschätzers ist.

Bei der Reject Inference ist die Wahrscheinlichkeit $P(Y = 0|\boldsymbol{X} = \boldsymbol{x})$ von Interesse. Mit vielen unterschiedlichen Verfahren lässt sich diese Wahrscheinlichkeit schätzen, wie zum Beispiel durch ein logistisches Regressionsmodell oder ein Probitmodell (vgl. Kapitel 3). Diese gehören zur Klasse der Generalisierten Linearen Modelle, auf deren ML-Schätzung im nächsten Abschnitt näher eingegangen wird. Mit der Idee, die Schätzung eines solchen parametrischen Modells durch die oben beschriebene Gewichtung zu „entzerren", befasst sich Abschnitt 5.3. Der Einsatz in nichtparametrischen Verfahren wie der schon erwähnten Kerndichteschätzung ist auch denkbar. Außerdem könnten die Gewichte zum Beispiel auch bei Diskriminanzanalysen, Klassifikations- oder Regressionsbäumen sowie Zufallswäldern nützlich sein.

5.2 Generalisierte Lineare Modelle

An dieser Stelle gibt ein kurzer Einschub einen Überblick über *Generalisierte Lineare Modelle* (GLM) sowie deren ML-Schätzung. Eine ausführliche Übersicht zu GLM bieten McCullagh und Nelder (1989) sowie Gill (2001).

Generalisierte Lineare Modelle stellen eine Verallgemeinerung klassischer Linearer Modelle dar. Dabei kann ein allgemeinerer Zusammenhang zwischen einer Zielvariablen Y und einem Vektor von Einflussgrößen $\boldsymbol{x} = (x_1, \ldots, x_d)^\top$ modelliert werden als in der klassischen Regression. Die Kovariablen $\boldsymbol{x}$ werden üblicherweise als deterministisch angenommen. GLM bestehen typischerweise aus drei verschiedenen Komponenten:

1. Die stochastische Komponente: Y ist eine Zufallsvariable mit $\mathrm{E}(Y) = \mu_Y$, wobei die Verteilung von Y einer Exponentialfamilie entstammt.

2. Die deterministische Komponente: Die Kovariablen $\boldsymbol{x} = (x_1, \ldots, x_d)^\top$ bilden einen linearen Prediktor
$$\eta = \boldsymbol{x}^\top \boldsymbol{\beta}.$$

3. Die Verknüpfung (der Link) zwischen der stochastischen und der deterministischen Komponente: die beiden ersten Komponenten sind durch eine *Linkfunktion* g bzw. eine *Responsefunktion* g^{-1} verknüpft
$$\eta = g(\mu_Y) \quad \text{bzw.} \quad \mu_Y = g^{-1}(\eta).$$

Die Zufallsvariable Y stammt aus einer *Exponentialfamilie* $\mathcal{E}(\theta, \psi)$ mit kanonischem Parameter θ und Dispersionsparameter ψ, falls sie eine Dichte bzw. Wahrscheinlichkeitsfunktion der Form
$$f(y|\theta, \psi) = \exp\left(\frac{y\theta - b(\theta)}{a(\psi)} - c(y, \psi)\right) \tag{5.12}$$
besitzt. Dabei sind $a(\cdot)$, $b(\cdot)$ und $c(\cdot)$ reelle Funktionen.

Die Log-Likelihood für (θ, ψ) bei gegebener Realisation y ist dann
$$\ln L(\theta, \psi|y) = \ln f(y|\theta, \psi) = \frac{y\theta - b(\theta)}{a(\psi)} - c(y, \psi).$$

Für den Erwartungswert von Y gilt
$$\mathrm{E}(Y) = \frac{\partial b(\theta)}{\partial \theta}$$

und für die Varianz

$$\mathrm{Var}(Y) = \frac{\partial^2 b(\theta)}{(\partial\theta)^2} a(\psi).$$

Die Linkfunktion verknüpft den Erwartungswert von Y mit dem linearen Prediktor η. Im klassischen Linearen Modell wird Y als normalverteilt angenommen, die Linkfunktion ist die Identität, so dass der Erwartungswert μ_Y mit dem linearen Prediktor gleichgesetzt wird. Da Y in diesem Fall Werte in $\mathbb{R}$ annehmen kann, ist dieser Zusammenhang plausibel. Im Falle einer binären Zufallsvariable oder Zählvariable als Zielgröße sollte der Erwartungswert allerdings strikt positiv sein und damit eine Linkfunktion gewählt werden, die dies garantiert. Für die verschiedenen Verteilungsklassen der Exponentialfamilie (wie etwa Normal-, Poisson-, Binomial- oder Gammaverteilung) gibt es sogenannte *kanonische Linkfunktionen*, für die $(\boldsymbol{X}_1^\top, \ldots, \boldsymbol{X}_n^\top)(Y_1, \ldots, Y_n)^\top$ eine suffiziente Statistik der Dimension von $\boldsymbol{\beta}$ ist (vgl. McCullagh und Nelder 1989, Kap. 2). Für diese kanonische Linkfunktion gilt dann $\theta = \eta$.

Wir betrachten im Folgenden Generalisierte Lineare Modelle für Exponentialfamilien mit kanonischer Linkfunktion. Diese haben die Log-Likelihoodfunktion

$$\ln L(\boldsymbol{\beta}|y, \boldsymbol{x}) = \frac{y\boldsymbol{x}^\top\boldsymbol{\beta} - b(\boldsymbol{x}^\top\boldsymbol{\beta})}{a(\psi)} - c(y, \psi).$$

Häufig interessiert man sich für den Parametervektor $\boldsymbol{\beta}$ und betrachtet ψ als Störparameter. Zur ML-Schätzung von $\boldsymbol{\beta}$ genügt es deshalb für eine Stichprobe von n Beobachtungen die Scoregleichungen

$$\boldsymbol{s}^n(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{x}_i \left[y_i - g^{-1}(\boldsymbol{x}_i^\top\boldsymbol{\beta})\right]\Big|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}^{\mathrm{ML}}} = \mathbf{0} \tag{5.13}$$

zu lösen.

5.3 Parameterschätzung in Modellen mit nicht-ignorierbar fehlender Zielgröße

Sei $f(y|\boldsymbol{x}, \boldsymbol{\beta})$ die parametrische bedingte Dichte von $Y|\boldsymbol{X}$. Diese kann zum Beispiel aus einer Exponentialfamilie stammen. Dann ergibt sich der Maximum-Likelihood-

Schätzer aus einer Stichprobe $\{(y_1, \boldsymbol{x}_1), \ldots, (y_N, \boldsymbol{x}_N)\}$ als Nullstelle der Scorefunktion

$$\boldsymbol{s}^N(\boldsymbol{\beta}) := \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{s}_i(\boldsymbol{\beta}) := \frac{1}{N}\sum_{i=1}^{N} \frac{\partial \ln f(y_i|\boldsymbol{x}_i, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}}.$$

Wenn die Beobachtungen $y_{n+1}, \ldots, y_N$ MNAR sind, betrachten wir die Likelihood, welche nur auf den vollständigen Beobachtungen $\{(y_1, \boldsymbol{x}_1), \ldots, (y_n, \boldsymbol{x}_n)\}$ basiert. Gewichten wir diese nun mit der geschätzten empirischen Wahrscheinlichkeitsfunktion $\widehat{p}_i$, erhalten wir die Scorefunktion

$$\boldsymbol{\Psi}^n(\boldsymbol{\beta}) := \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\psi}_i^n(\boldsymbol{\beta}) := \frac{1}{n}\sum_{i=1}^{n} n\widehat{p}_i \frac{\partial \ln f(y_i|\boldsymbol{x}_i, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}}. \tag{5.14}$$

Man kann zeigen, dass sich der Parameter $\boldsymbol{\beta}$ durch die Nullstelle dieser Scorefunktion konsistent schätzen lässt. Dazu betrachten wir zunächst folgendes Lemma.

Lemma 5.1:

Sei

$$\boldsymbol{s}_i(\boldsymbol{\beta}) := \frac{\partial \ln f(y_i|\boldsymbol{x}_i, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}}.$$

Dann gilt für alle $i \in \{1, \ldots, n\}$

$$n\widehat{p}_i \boldsymbol{s}_i(\boldsymbol{\beta}) \xrightarrow{p} \frac{W_0}{w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}_0)} \boldsymbol{s}_i(\boldsymbol{\beta}).$$

Beweis: Wegen

$$\begin{aligned}
\frac{n}{N} &\xrightarrow{p} W_0, \\
\widehat{W} &\xrightarrow{p} W_0, \\
\widehat{\boldsymbol{\theta}} &\xrightarrow{p} \boldsymbol{\theta}_0, \\
\widehat{\boldsymbol{\lambda}}_1 &\xrightarrow{p} \mathbf{0}, \\
\widehat{\lambda}_2 = \frac{N/n - 1}{1 - \widehat{W}} &\xrightarrow{p} \frac{1}{W_0}
\end{aligned}$$

gilt mit

$$\widehat{p}_i = \frac{1}{n\left[1 + \widehat{\boldsymbol{\lambda}}_1^\top (\boldsymbol{x}_i - \boldsymbol{\mu}_{\boldsymbol{X}}) + \widehat{\lambda}_2 \left(w(y_i, \boldsymbol{x}_i, \widehat{\boldsymbol{\theta}}) - \widehat{W}\right)\right]}$$

und dem Satz von der stetigen Abbildung, dass

$$n\widehat{p}_i \xrightarrow{p} \frac{W_0}{w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}_0)}$$

und damit

$$n\widehat{p}_i \boldsymbol{s}_i(\boldsymbol{\beta}) \xrightarrow{p} \frac{W_0}{w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}_0)} \boldsymbol{s}_i(\boldsymbol{\beta}).$$

■

Das folgende Lemma zeigt nun, dass wir von dem bedingten Erwartungswert bezüglich $Y, \boldsymbol{X}|R = 1$ mit Hilfe des Grenzwertes von $n\widehat{p}_i$ zu einem unbedingten Erwartungswert gelangen können.

Lemma 5.2:

Für eine integrierbare Funktion $g(y, \boldsymbol{x})$ gilt

$$\mathrm{E}_{Y,\boldsymbol{X}|R=1}\left(\frac{W_0}{w(Y, \boldsymbol{X}, \boldsymbol{\theta}_0)} g(Y, \boldsymbol{X})\right) = \mathrm{E}_{Y,\boldsymbol{X}}\left(g(Y, \boldsymbol{X})\right).$$

Beweis:

$$\begin{aligned}
\mathrm{E}_{Y,\boldsymbol{X}|R=1}\left(\frac{W_0}{w(Y, \boldsymbol{X}, \boldsymbol{\theta}_0)} g(Y, \boldsymbol{X})\right) &= \iint \frac{W_0}{w(y, \boldsymbol{x}, \boldsymbol{\theta}_0)} g(y, \boldsymbol{x}) \frac{w(y, \boldsymbol{x}, \boldsymbol{\theta}_0)}{W_0} dF(\boldsymbol{x}, y) \\
&= \iint g(y, \boldsymbol{x}) dF(\boldsymbol{x}, y) \\
&= \mathrm{E}_{Y,\boldsymbol{X}}\left(g(Y, \boldsymbol{X})\right).
\end{aligned}$$

■

Nun betrachten wir den Schätzer, den man durch Nullsetzen der Scorefunktion (5.14) erhält.

Satz 5.3:

Es gelten die folgenden Annahmen:

(A1) $\exists q > 1 : \limsup_{n\to\infty} \mathrm{E}_{Y,\boldsymbol{X}|R=1}\left(\|\boldsymbol{\psi}^n(\boldsymbol{\beta})\|^q\right) < \infty$.

(A2) $\mathrm{E}_{Y,\boldsymbol{X}|R=1}\left(\boldsymbol{\psi}^n(\boldsymbol{\beta})\right)$ ist gleichgradig stetig.

(A3) Der Wertebereich von $\boldsymbol{\beta}$ ist kompakt.

(A4) $\boldsymbol{\psi}^n(\boldsymbol{\beta})$ ist stetig in $\boldsymbol{\beta}$ für fast alle $\boldsymbol{x}, y$.

(A5) $\exists d(\boldsymbol{x}, y)$ mit $\mathrm{E}_{Y,\boldsymbol{X}|R=1}(d(\boldsymbol{X}, Y)) < \infty$ und $\|\boldsymbol{\psi}^n(\boldsymbol{\beta})\| \leq d(\boldsymbol{x}, y)$ für alle $\boldsymbol{\beta}$.

(A6) Bei der Ableitung nach $\boldsymbol{\beta}$ des Integrals von $f(y|\boldsymbol{x}, \boldsymbol{\beta}_0)$ über den Träger von $\boldsymbol{X}, Y$ kann die Reihenfolge von Integration und Differentiation vertauscht werden:

$$\iint \frac{\partial}{\partial \boldsymbol{\beta}} f(y|\boldsymbol{x}, \boldsymbol{\beta}_0) d\lambda_{\boldsymbol{X}} d\lambda_Y = \frac{\partial}{\partial \boldsymbol{\beta}} \iint f(y|\boldsymbol{x}, \boldsymbol{\beta}_0) d\lambda_{\boldsymbol{X}} d\lambda_Y,$$

wobei $\lambda_{\boldsymbol{X}}$ und λ_Y entsprechende Maße (Lebesgue- bzw. Zählmaß) sind.

(A7) Die Randverteilung von $\boldsymbol{X}$ mit Randdichte $f(\boldsymbol{x})$ ist unabhängig von $\boldsymbol{\beta}$.

(A8) $\mathrm{E}_{Y,\boldsymbol{X}}\left(\boldsymbol{s}(\boldsymbol{\beta})\right)$ hat eine eindeutige Nullstelle.

(A9) $\boldsymbol{\Psi}^n(\widehat{\boldsymbol{\beta}}^{\mathrm{SEL}}) = o_p(1)$.

Sei $\widehat{\boldsymbol{\beta}}^{\mathrm{SEL}}$ die Lösung des Gleichungssystems

$$\boldsymbol{\Psi}^n(\boldsymbol{\beta}) = \mathbf{0}.$$

Dann gilt $\widehat{\boldsymbol{\beta}}^{\mathrm{SEL}} \xrightarrow{p} \boldsymbol{\beta}_0$, das heißt $\widehat{\boldsymbol{\beta}}^{\mathrm{SEL}}$ ist schwach konsistent.

Beweis: Die Argumentation erfolgt über die Tatsache, dass es sich bei $\widehat{\boldsymbol{\beta}}^{\mathrm{SEL}}$ um einen M-Schätzer (bzw. Z-Schätzer) handelt. Die schwache Konsistenz gilt also mit van der Vaart (1998, Theorem 5.9), falls $\boldsymbol{\Psi}^n(\boldsymbol{\beta})$ gleichmäßig stochastisch gegen ein $\boldsymbol{\Psi}(\boldsymbol{\beta})$ mit $\boldsymbol{\beta}_0$ als einzige Nullstelle konvergiert.

Der Beweis erfolgt nun in vier Schritten:

1. Sei
$$\boldsymbol{\psi}_i(\boldsymbol{\beta}) := \frac{W_0}{w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}_0)} \frac{\partial \ln f(y_i|\boldsymbol{x}_i, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}}.$$
Dann gilt mit Lemma 5.1 für alle $i \in \{1, \ldots, n\}$ $\boldsymbol{\psi}_i^n(\boldsymbol{\beta}) \xrightarrow{p} \boldsymbol{\psi}_i(\boldsymbol{\beta})$.

2. Mit (A1) folgt die asymptotische gleichgradige Integrierbarkeit von $\boldsymbol{\psi}^n(\boldsymbol{\beta})$. Daher gilt mit van der Vaart (1998, Theorem 2.20) und Lemma 5.2 für alle $i \in \{1, \ldots, n\}$
$$\mathrm{E}_{Y,\boldsymbol{X}|R=1}\left(\boldsymbol{\psi}_i^n(\boldsymbol{\beta})\right) \xrightarrow{n\to\infty} \mathrm{E}_{Y,\boldsymbol{X}|R=1}\left(\boldsymbol{\psi}_i(\boldsymbol{\beta})\right) = \mathrm{E}_{Y,\boldsymbol{X}}\left(\boldsymbol{s}_i(\boldsymbol{\beta})\right).$$
Die gleichmäßige Konvergenz
$$\sup_{\boldsymbol{\beta}} \left\| \mathrm{E}_{Y,\boldsymbol{X}|R=1}\left(\boldsymbol{\psi}_i^n(\boldsymbol{\beta})\right) - \mathrm{E}_{Y,\boldsymbol{X}}\left(\boldsymbol{s}_i(\boldsymbol{\beta})\right) \right\| \xrightarrow{p} 0 \tag{5.15}$$
folgt mit der gleichgradigen Stetigkeit von $\mathrm{E}_{Y,\boldsymbol{X}|R=1}\left(\boldsymbol{\psi}_i^n(\boldsymbol{\beta})\right)$ (Voraussetzung (A2)).

3. Wegen (A3), (A4) und (A5) gilt mit dem Gleichmäßigen Gesetz der Großen Zahlen (Newey und McFadden 1994, Lemma 2.4), dass

$$\sup_{\boldsymbol{\beta}} \left\| \frac{1}{n} \sum_{i=1}^{n} \left[\boldsymbol{\psi}_{\boldsymbol{i}}^{n}(\boldsymbol{\beta}) - \mathrm{E}_{Y,\boldsymbol{X}|R=1} \left(\boldsymbol{\psi}_{\boldsymbol{i}}^{n}(\boldsymbol{\beta}) \right) \right] \right\| \xrightarrow{p} 0.$$

4. Schließlich folgt mit $\boldsymbol{\Psi}(\boldsymbol{\beta}) := \mathrm{E}_{Y,\boldsymbol{X}} \left(\boldsymbol{s}(\boldsymbol{\beta}) \right)$, dass

$$\begin{aligned}
&\sup_{\boldsymbol{\beta}} \left\| \boldsymbol{\Psi}^{n}(\boldsymbol{\beta}) - \boldsymbol{\Psi}(\boldsymbol{\beta}) \right\| \\
=& \sup_{\boldsymbol{\beta}} \left\| \frac{1}{n} \sum_{i=1}^{n} \left[\boldsymbol{\psi}_{\boldsymbol{i}}^{n}(\boldsymbol{\beta}) - \mathrm{E}_{Y,\boldsymbol{X}|R=1} \left(\boldsymbol{\psi}_{\boldsymbol{i}}^{n}(\boldsymbol{\beta}) \right) \right] \right. \\
&\qquad \left. + \frac{1}{n} \sum_{i=1}^{n} \mathrm{E}_{Y,\boldsymbol{X}|R=1} \left(\boldsymbol{\psi}_{\boldsymbol{i}}^{n}(\boldsymbol{\beta}) \right) - \mathrm{E}_{Y,\boldsymbol{X}} \left(\boldsymbol{s}_{i}(\boldsymbol{\beta}) \right) \right\| \\
\leq& \sup_{\boldsymbol{\beta}} \left\| \frac{1}{n} \sum_{i=1}^{n} \left[\boldsymbol{\psi}_{\boldsymbol{i}}^{n}(\boldsymbol{\beta}) - \mathrm{E}_{Y,\boldsymbol{X}|R=1} \left(\boldsymbol{\psi}_{\boldsymbol{i}}^{n}(\boldsymbol{\beta}) \right) \right] \right\| \\
&\qquad + \sup_{\boldsymbol{\beta}} \left\| \frac{1}{n} \sum_{i=1}^{n} \mathrm{E}_{Y,\boldsymbol{X}|R=1} \left(\boldsymbol{\psi}_{\boldsymbol{i}}^{n}(\boldsymbol{\beta}) \right) - \mathrm{E}_{Y,\boldsymbol{X}} \left(\boldsymbol{s}_{i}(\boldsymbol{\beta}) \right) \right\| \xrightarrow{p} 0.
\end{aligned}$$

Dabei gilt die Konvergenz des zweiten Summanden, da er ein Césaro-Mittel von (5.15) ist. Mit den Voraussetzungen (A6) und (A7) gilt schließlich

$$\begin{aligned}
\mathrm{E}_{Y,\boldsymbol{X}} \left(\boldsymbol{s}_{i}(\boldsymbol{\beta}_0) \right) &= \iint \frac{\partial f(y|\boldsymbol{x},\boldsymbol{\beta}_0)/\partial \boldsymbol{\beta}}{f(y|\boldsymbol{x},\boldsymbol{\beta}_0)} f(y,\boldsymbol{x}|\boldsymbol{\beta}_0) d\lambda_{\boldsymbol{X}} d\lambda_Y \\
&= \iint \frac{\partial f(y|\boldsymbol{x},\boldsymbol{\beta}_0)/\partial \boldsymbol{\beta} \cdot f(\boldsymbol{x})}{f(y,\boldsymbol{x}|\boldsymbol{\beta}_0)} f(y,\boldsymbol{x}|\boldsymbol{\beta}_0) d\lambda_{\boldsymbol{X}} d\lambda_Y \\
&= \frac{\partial}{\partial \boldsymbol{\beta}} \iint f(y|\boldsymbol{x},\boldsymbol{\beta}_0) f(x) d\lambda_{\boldsymbol{X}} d\lambda_Y \\
&= \frac{\partial}{\partial \boldsymbol{\beta}} \iint f(y,\boldsymbol{x}|\boldsymbol{\beta}_0) d\lambda_{\boldsymbol{X}} d\lambda_Y \\
&= \frac{\partial}{\partial \boldsymbol{\beta}} 1 \\
&= \mathbf{0},
\end{aligned}$$

das heißt $\mathrm{E}_{Y,\boldsymbol{X}} \left(\boldsymbol{s}_{i}(\boldsymbol{\beta}) \right)$ hat eine Nullstelle in $\boldsymbol{\beta}_0$. Dann folgt wegen (A8) und (A9) mit van der Vaart (1998, Theorem 5.9) die Behauptung.

■

In der Regel dürfte der Erwartungswert von $\boldsymbol{X}$ unbekannt sein. Allerdings ist auch dann eine konsistente Schätzung möglich, wie die folgende Bemerkung deutlich macht.

Bemerkung 5.2:

Wenn der Erwartungswert $\mathrm{E}(\boldsymbol{X}) = \boldsymbol{\mu}_{\boldsymbol{X}}$ unbekannt ist und durch das arithmetische Mittel aller realisierten $\boldsymbol{x}$-Werte $\bar{\boldsymbol{x}} = N^{-1} \sum_{i=1}^{N} \boldsymbol{x}_i$ ersetzt wird, bleibt Satz 5.3 gültig, da Lemma 5.1 auch in diesem Fall korrekt ist.

Die folgende Bemerkung untersucht die Voraussetzungen von Satz 5.3 genauer.

Bemerkung 5.3:

Die Annahmen (A1), (A2) und (A5) fordern die gleichgradige Integrierbarkeit, die gleichgradige Stetigkeit des Erwartungswerts bzw. die Beschränktheit durch eine integrierbare Funktion von $\boldsymbol{\psi}^n(\boldsymbol{\beta})$. Eine Vereinfachung dieser Annahmen für Generalisierte Lineare Modelle liefern die Sätze 5.4-5.9.

Die Voraussetzungen eines kompakten Wertebereichs für den unbekannten Parameter (A3), die Regularitätsbedingung (A6), die Existenz einer eindeutigen Nullstelle des Erwartungswerts in (A8) sowie die asymptotischen Existenz einer Nullstelle der Scoregleichungen (A9) sind typische Annahmen und in der Regel unproblematisch. Für Generalisierte Lineare Modelle ist $\boldsymbol{\psi}^n(\boldsymbol{\beta})$ stetig differenzierbar (vgl. (A4)). Auch die Annahme (A7), dass die Randverteilung von $\boldsymbol{X}$ unabhängig von $\boldsymbol{\beta}$ ist und damit selbst keine Informationen über den unbekannten Parameter enthält, ist plausibel.

Der folgende Satz untersucht nun, inwiefern die Annahme (A1) für GLM erfüllt ist. Sei dazu zunächst

$$\widehat{p} = \frac{1}{n\left[1 + \widehat{\boldsymbol{\lambda}}_1^{\top}\left(\boldsymbol{X} - \boldsymbol{\mu}_{\boldsymbol{X}}\right) + \widehat{\lambda}_2\left(w(Y, \boldsymbol{X}_i, \widehat{\boldsymbol{\theta}}) - \widehat{W}\right)\right]},$$

wobei die Parameter aus n Beobachtungen geschätzt wurden.

Satz 5.4:

Sei $Y \sim \mathcal{E}(\theta, \psi)$ (vgl. Gleichung (5.12)), also mit Dichte bzw. Wahrscheinlichkeitsfunktion

$$f(y|\theta, \psi) = \exp\left(\frac{y\theta - b(\theta)}{a(\psi)} - c(y, \psi)\right).$$

Von Interesse sei ein GLM mit kanonischer Linkfunktion g, das heißt es gilt $\theta = \eta = \boldsymbol{x}^{\top}\boldsymbol{\beta}$ und der Störparameter ψ wird vernachlässigt.

Dann ist Annahme (A1) für $q = 2$ erfüllt, falls

$$\begin{aligned} & \limsup_{n\to\infty} \mathrm{E}_{Y,\boldsymbol{X}|R=1}\left((n\widehat{p})^2 \left\|Y\boldsymbol{X}\right\|^2\right) < \infty \\ \text{und} \quad & \limsup_{n\to\infty} \mathrm{E}_{Y,\boldsymbol{X}|R=1}\left((n\widehat{p})^2 \left\|\boldsymbol{X}g^{-1}(\boldsymbol{X}^\top\boldsymbol{\beta})\right\|^2\right) < \infty \quad \forall\boldsymbol{\beta}. \end{aligned}$$

Beweis: Sei (vgl. (5.13))

$$\boldsymbol{\psi}^n(\boldsymbol{\beta}) = n\widehat{p}\boldsymbol{x}\left[y - g^{-1}(\boldsymbol{x}^\top\boldsymbol{\beta})\right],$$

wobei g die kanonische Linkfunktion der Exponentialfamilie ist.

Dann folgt

$$\begin{aligned} & \mathrm{E}_{Y,\boldsymbol{X}|R=1}\left(\left\|\boldsymbol{\psi}^n(\boldsymbol{\beta})\right\|^2\right) \\ = \; & \mathrm{E}_{Y,\boldsymbol{X}|R=1}\left(\left\|n\widehat{p}\boldsymbol{X}\left[Y - g^{-1}(\boldsymbol{X}^\top\boldsymbol{\beta})\right]\right\|^2\right) \\ \leq \; & \mathrm{E}_{Y,\boldsymbol{X}|R=1}\left((n\widehat{p})^2(\left\|Y\boldsymbol{X}\right\|^2 + \left\|\boldsymbol{X}g^{-1}(\boldsymbol{X}^\top\boldsymbol{\beta})\right\|^2)\right) \\ \leq \; & \mathrm{E}_{Y,\boldsymbol{X}|R=1}\left((n\widehat{p})^2(\left\|Y\boldsymbol{X}\right\|^2 + (n\widehat{p})^2\left\|\boldsymbol{X}g^{-1}(\boldsymbol{X}^\top\boldsymbol{\beta})\right\|^2)\right) \\ = \; & \mathrm{E}_{Y,\boldsymbol{X}|R=1}\left((n\widehat{p})^2\left\|Y\boldsymbol{X}\right\|^2\right) + \mathrm{E}_{Y,\boldsymbol{X}|R=1}\left((n\widehat{p})^2\left\|\boldsymbol{X}g^{-1}(\boldsymbol{X}^\top\boldsymbol{\beta})\right\|^2\right) \end{aligned}$$

und damit die Behauptung. ■

Für ein logistisches Regressionsmodell, welches in der Reject Inference von Interesse ist, kann die Annahme noch weiter vereinfacht werden.

Satz 5.5:

Sei Y die binomialverteilte Zielgröße eines GLM mit kanonischem logit-Link (Logistisches Regressionsmodell), also mit Wahrscheinlichkeitsfunktion

$$f(y|\boldsymbol{x},\boldsymbol{\beta}) = \left(\frac{\exp(\boldsymbol{x}^\top\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}^\top\boldsymbol{\beta})}\right)^y \left(\frac{1}{1+\exp(\boldsymbol{x}^\top\boldsymbol{\beta})}\right)^{1-y}.$$

Dann gilt

$$\limsup_{n\to\infty} \mathrm{E}_{Y,\boldsymbol{X}|R=1}\left(\left\|\boldsymbol{\psi}^n(\boldsymbol{\beta})\right\|^2\right) < \infty,$$

falls

$$\limsup_{n\to\infty} \mathrm{E}_{Y,\boldsymbol{X}|R=1}\left((n\widehat{p})^2\left\|\boldsymbol{X}\right\|^2\right) < \infty.$$

Beweis: Es gilt

$$\frac{\partial \ln f(y|\boldsymbol{x},\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \boldsymbol{x}\left(y - \frac{\exp(\boldsymbol{x}^\top\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}^\top\boldsymbol{\beta})}\right)$$

und damit

$$\boldsymbol{\psi}^n(\boldsymbol{\beta}) = n\widehat{p}\boldsymbol{x}\left(y - \frac{\exp(\boldsymbol{x}^\top\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}^\top\boldsymbol{\beta})}\right).$$

Es folgt

$$\begin{aligned}
& \mathrm{E}_{Y,\boldsymbol{X}|R=1}\left(\|\boldsymbol{\psi}^n(\boldsymbol{\beta})\|^2\right) \\
= \; & \mathrm{E}_{Y,\boldsymbol{X}|R=1}\left(\left\| n\widehat{p}\boldsymbol{X}\underbrace{\left(Y - \frac{\exp(\boldsymbol{X}^\top\boldsymbol{\beta})}{1+\exp(\boldsymbol{X}^\top\boldsymbol{\beta})}\right)}_{\in(-1,1)}\right\|^2\right) \\
\leq \; & \mathrm{E}_{Y,\boldsymbol{X}|R=1}\left(\|n\widehat{p}\boldsymbol{X}\|^2\right) \\
\leq \; & \mathrm{E}_{Y,\boldsymbol{X}|R=1}\left((n\widehat{p})^2\|\boldsymbol{X}\|^2\right).
\end{aligned}$$

und damit die Behauptung. ■

Nun untersuchen wir die Annahme (A2) der gleichgradigen Stetigkeit des Erwartungswerts $\mathrm{E}_{Y,\boldsymbol{X}|R=1}\left(\boldsymbol{\psi}^n(\boldsymbol{\beta})\right)$ für den Fall von Generalisierten Linearen Modellen.

Satz 5.6:

Unter den Voraussetzungen von Satz 5.4 ist $\mathrm{E}_{Y,\boldsymbol{X}|R=1}\left(\boldsymbol{\psi}^n(\boldsymbol{\beta})\right)$ gleichgradig stetig, falls ein $M < \infty$ existiert, so dass für alle $n \in \mathbb{N}$ und alle $\boldsymbol{\beta}, \boldsymbol{\beta}'$

$$\mathrm{E}_{Y,\boldsymbol{X}|R=1}\left(n\widehat{p}\lambda_{\max}\left(\boldsymbol{X}\frac{\partial g^{-1}(\boldsymbol{X}^\top\tilde{\boldsymbol{\beta}})}{\partial\boldsymbol{\beta}}\right)\right) \leq M$$

für eine Konvexkombination $\tilde{\boldsymbol{\beta}}$ von $\boldsymbol{\beta}$ und $\boldsymbol{\beta}'$, wobei $\lambda_{\max}(\boldsymbol{A})$ den größten Eigenwert von $\boldsymbol{A}^\top\boldsymbol{A}$ bezeichnet.

Beweis: Es gilt wieder

$$\boldsymbol{\psi}^n(\boldsymbol{\beta}) = n\widehat{p}\boldsymbol{x}\left[y - g^{-1}(\boldsymbol{x}^\top\boldsymbol{\beta})\right].$$

Die gleichgradige Stetigkeit von $\mathrm{E}_{Y,\boldsymbol{X}|R=1}\left(\boldsymbol{\psi}^n(\boldsymbol{\beta})\right)$ gilt, falls für alle $\varepsilon > 0$ ein $\delta > 0$ für alle $n \in \mathbb{N}$ und alle $\boldsymbol{\beta}, \boldsymbol{\beta}'$ existiert, so dass

$$\|\boldsymbol{\beta} - \boldsymbol{\beta}'\| \leq \delta \quad \Rightarrow \quad \left\|\mathrm{E}_{Y,\boldsymbol{X}|R=1}\left(\boldsymbol{\psi}^n(\boldsymbol{\beta})\right) - \mathrm{E}_{Y,\boldsymbol{X}|R=1}\left(\boldsymbol{\psi}^n(\boldsymbol{\beta}')\right)\right\| \leq \varepsilon.$$

Nun folgt

$$\begin{aligned}
& \left\| \mathrm{E}_{Y,\boldsymbol{X}|R=1}\left(\boldsymbol{\psi}^n(\boldsymbol{\beta})\right) - \mathrm{E}_{Y,\boldsymbol{X}|R=1}\left(\boldsymbol{\psi}^n(\boldsymbol{\beta}')\right) \right\| \\
= & \left\| \mathrm{E}_{Y,\boldsymbol{X}|R=1}\left(n\widehat{p}\boldsymbol{X}\left[g^{-1}(\boldsymbol{X}^\top\boldsymbol{\beta}') - g^{-1}(\boldsymbol{X}^\top\boldsymbol{\beta})\right]\right) \right\| \\
\leq & \mathrm{E}_{Y,\boldsymbol{X}|R=1}\left(n\widehat{p}\left\|\boldsymbol{X}\left[g^{-1}(\boldsymbol{X}^\top\boldsymbol{\beta}') - g^{-1}(\boldsymbol{X}^\top\boldsymbol{\beta})\right]\right\|\right) \\
\overset{(*)}{\leq} & \mathrm{E}_{Y,\boldsymbol{X}|R=1}\left(n\widehat{p}\left\|\boldsymbol{X}\frac{\partial g^{-1}(\boldsymbol{X}^\top\tilde{\boldsymbol{\beta}})}{\partial\boldsymbol{\beta}}(\boldsymbol{\beta}'-\boldsymbol{\beta})\right\|\right) \\
\leq & \|\boldsymbol{\beta}'-\boldsymbol{\beta}\|\,\mathrm{E}_{Y,\boldsymbol{X}|R=1}\left(n\widehat{p}\lambda_{\max}\left(\boldsymbol{X}\frac{\partial g^{-1}(\boldsymbol{X}^\top\tilde{\boldsymbol{\beta}})}{\partial\boldsymbol{\beta}}\right)\right) \\
\leq & \|\boldsymbol{\beta}'-\boldsymbol{\beta}\|\,M = \varepsilon.
\end{aligned}$$

Dabei gilt $(*)$ mit dem Mittelwertsatz, denn

$$g^{-1}(\boldsymbol{X}^\top\boldsymbol{\beta}') - g^{-1}(\boldsymbol{X}^\top\boldsymbol{\beta}) = \frac{\partial}{\partial\boldsymbol{\beta}}g^{-1}(\boldsymbol{X}^\top\tilde{\boldsymbol{\beta}})(\boldsymbol{\beta}'-\boldsymbol{\beta})$$

für eine Konvexkombination $\tilde{\boldsymbol{\beta}}$ von $\boldsymbol{\beta}$ und $\boldsymbol{\beta}'$. Mit $\delta = \varepsilon/M$ folgt die Behauptung.

■

Eine Vereinfachung für Logistische Regressionsmodelle lässt sich auch für die gleichgradige Stetigkeit angeben.

Satz 5.7:

Unter den Voraussetzungen von Satz 5.5 ist $\mathrm{E}_{Y,\boldsymbol{X}|R=1}\left(\boldsymbol{\psi}^n(\boldsymbol{\beta})\right)$ gleichgradig stetig, falls ein $M < \infty$ existiert, so dass für alle $n \in \mathbb{N}$

$$\mathrm{E}_{Y,\boldsymbol{X}|R=1}\left(n\widehat{p}\lambda_{\max}\left(\boldsymbol{X}\boldsymbol{X}^\top\right)\right) \leq M,$$

mit $\lambda_{\max}(\boldsymbol{A})$ der größte Eigenwert von $\boldsymbol{A}^\top\boldsymbol{A}$.

Beweis: Es gilt wieder

$$\frac{\partial \ln f(y|\boldsymbol{x},\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} = \boldsymbol{x}\left(y - \frac{\exp(\boldsymbol{x}^\top\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}^\top\boldsymbol{\beta})}\right)$$

und damit

$$\boldsymbol{\psi}^n(\boldsymbol{\beta}) = n\widehat{p}\boldsymbol{x}\left(y - \frac{\exp(\boldsymbol{x}^\top\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}^\top\boldsymbol{\beta})}\right).$$

Nun folgt wie im Beweis zu Satz 5.6

$$\begin{aligned}
& \left\| \mathrm{E}_{Y,\boldsymbol{X}|R=1}\left(\boldsymbol{\psi}^n(\boldsymbol{\beta})\right) - \mathrm{E}_{Y,\boldsymbol{X}|R=1}\left(\boldsymbol{\psi}^n(\boldsymbol{\beta}')\right) \right\| \\
= & \left\| \mathrm{E}_{Y,\boldsymbol{X}|R=1}\left(\boldsymbol{\psi}^n(\boldsymbol{\beta}) - \boldsymbol{\psi}^n(\boldsymbol{\beta}')\right) \right\| \\
= & \left\| \mathrm{E}_{Y,\boldsymbol{X}|R=1}\left(n\widehat{p}\boldsymbol{X}\left[\left(Y - \frac{\exp(\boldsymbol{X}^\top\boldsymbol{\beta})}{1+\exp(\boldsymbol{X}^\top\boldsymbol{\beta})} \right) - \left(Y - \frac{\exp(\boldsymbol{X}^\top\boldsymbol{\beta}')}{1+\exp(\boldsymbol{X}^\top\boldsymbol{\beta}')} \right) \right] \right) \right\| \\
= & \left\| \mathrm{E}_{Y,\boldsymbol{X}|R=1}\left(n\widehat{p}\boldsymbol{X}\left(\frac{\exp(\boldsymbol{X}^\top\boldsymbol{\beta}')}{1+\exp(\boldsymbol{X}^\top\boldsymbol{\beta}')} - \frac{\exp(\boldsymbol{X}^\top\boldsymbol{\beta})}{1+\exp(\boldsymbol{X}^\top\boldsymbol{\beta})} \right) \right) \right\| \\
\leq & \mathrm{E}_{Y,\boldsymbol{X}|R=1}\left(n\widehat{p} \left\| \boldsymbol{X}\left(\frac{\exp(\boldsymbol{X}^\top\boldsymbol{\beta}')}{1+\exp(\boldsymbol{X}^\top\boldsymbol{\beta}')} - \frac{\exp(\boldsymbol{X}^\top\boldsymbol{\beta})}{1+\exp(\boldsymbol{X}^\top\boldsymbol{\beta})} \right) \right\| \right) \\
\overset{(*)}{\leq} & \mathrm{E}_{Y,\boldsymbol{X}|R=1}\left(n\widehat{p} \left\| \boldsymbol{X}\boldsymbol{X}^\top(\boldsymbol{\beta}'-\boldsymbol{\beta}) \right\| \right) \\
= & \left\| \boldsymbol{\beta}'-\boldsymbol{\beta} \right\| \mathrm{E}_{Y,\boldsymbol{X}|R=1}\left(n\widehat{p}\lambda_{\max}\left(\boldsymbol{X}\boldsymbol{X}^\top \right) \right) \\
\leq & \left\| \boldsymbol{\beta}'-\boldsymbol{\beta} \right\| M = \varepsilon.
\end{aligned}$$

Dabei gilt $(*)$ mit dem Mittelwertsatz, denn

$$\begin{aligned}
\frac{\exp(\boldsymbol{X}^\top\boldsymbol{\beta}')}{1+\exp(\boldsymbol{X}^\top\boldsymbol{\beta}')} - \frac{\exp(\boldsymbol{X}^\top\boldsymbol{\beta})}{1+\exp(\boldsymbol{X}^\top\boldsymbol{\beta})} &= \frac{\partial}{\partial\boldsymbol{\beta}} \frac{\exp(\boldsymbol{X}^\top\tilde{\boldsymbol{\beta}})}{1+\exp(\boldsymbol{X}^\top\tilde{\boldsymbol{\beta}})} \cdot (\boldsymbol{\beta}'-\boldsymbol{\beta}) \\
&= \underbrace{\frac{\exp(\boldsymbol{X}^\top\tilde{\boldsymbol{\beta}})}{\left(1+\exp(\boldsymbol{X}^\top\tilde{\boldsymbol{\beta}})\right)^2}}_{\in(-1,1)} \boldsymbol{X}^\top(\boldsymbol{\beta}'-\boldsymbol{\beta})
\end{aligned}$$

für eine Konvexkombination $\tilde{\boldsymbol{\beta}}$ von $\boldsymbol{\beta}$ und $\boldsymbol{\beta}'$. Mit $\delta = \varepsilon/M$ folgt die Behauptung.

■

Nun folgt eine Untersuchung der Annahme (A5) für den Fall von Generalisierten Linearen Modellen.

Satz 5.8:

Unter den Voraussetzungen von Satz 5.4 ist Annahme (A5) erfüllt, falls

$$\mathrm{E}_{Y,\boldsymbol{X}|R=1}\left((n\widehat{p}) \left\| Y\boldsymbol{X} \right\| \right) < \infty \quad \text{und} \quad \mathrm{E}_{Y,\boldsymbol{X}|R=1}\left((n\widehat{p}) \left\| \boldsymbol{X} g^{-1}(\boldsymbol{X}^\top\boldsymbol{\beta}) \right\| \right) < \infty.$$

Beweis: Es gilt wieder

$$\boldsymbol{\psi}^n(\boldsymbol{\beta}) = n\widehat{p}\boldsymbol{x}\left[y - g^{-1}(\boldsymbol{x}^\top\boldsymbol{\beta}) \right].$$

Dann ist

$$\begin{aligned} & \mathrm{E}_{Y,\boldsymbol{X}|R=1}\left(\|\boldsymbol{\psi}^n(\boldsymbol{\beta})\|\right) \\ = \; & \mathrm{E}_{Y,\boldsymbol{X}|R=1}\left(\left\|n\widehat{p}\boldsymbol{X}\left[Y - g^{-1}(\boldsymbol{X}^\top\boldsymbol{\beta})\right]\right\|\right) \\ = \; & \mathrm{E}_{Y,\boldsymbol{X}|R=1}\left((n\widehat{p})\,\|Y\boldsymbol{X}\|\right) + \mathrm{E}_{Y,\boldsymbol{X}|R=1}\left((n\widehat{p})\left\|\boldsymbol{X}g^{-1}(\boldsymbol{X}^\top\boldsymbol{\beta})\right\|\right) \end{aligned}$$

und es folgt die Behauptung. ■

Satz 5.9:

Unter den Voraussetzungen von Satz 5.5 ist Annahme (A5) erfüllt, falls

$$\mathrm{E}_{Y,\boldsymbol{X}|R=1}\left((n\widehat{p})\,\|\boldsymbol{X}\|\right) < \infty.$$

Beweis: Es gilt

$$\frac{\partial \ln f(y|\boldsymbol{x},\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \boldsymbol{x}\left(y - \frac{\exp(\boldsymbol{x}^\top\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}^\top\boldsymbol{\beta})}\right)$$

und damit

$$\boldsymbol{\psi}^n(\boldsymbol{\beta}) = n\widehat{p}\boldsymbol{x}\left(y - \frac{\exp(\boldsymbol{x}^\top\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}^\top\boldsymbol{\beta})}\right).$$

Nun folgt

$$\begin{aligned} & \mathrm{E}_{Y,\boldsymbol{X}|R=1}\left(\|\boldsymbol{\psi}^n(\boldsymbol{\beta})\|\right) \\ = \; & \mathrm{E}_{Y,\boldsymbol{X}|R=1}\left(\left\|n\widehat{p}\boldsymbol{X}\underbrace{\left(Y - \frac{\exp(\boldsymbol{X}^\top\boldsymbol{\beta})}{1+\exp(\boldsymbol{X}^\top\boldsymbol{\beta})}\right)}_{\in(-1,1)}\right\|\right) \\ \leq \; & \mathrm{E}_{Y,\boldsymbol{X}|R=1}\left((n\widehat{p})\,\|\boldsymbol{X}\|\right) \end{aligned}$$

und damit die Behauptung. ■

Insgesamt zeigt sich, dass die Voraussetzungen (A1), (A2) und (A5) für GLM erfüllt sind, falls verschiedene erste und zweite Momente existieren. Es ist schwierig, allgemeine Aussagen über die Zufallsvariable $\widehat{p}$ und ihre Verteilung zu treffen, da diese auch entscheidend von den Zufallsvariablen $\boldsymbol{X}$ und Y abhängt. In Anwendungen zeigt sich jedoch, dass die $\widehat{p}_i$ um den Wert $1/n$ schwanken, so dass diese Momente in der Regel existieren dürften. Somit lassen sich zwar die Annahmen nicht vollkommen allgemein verifizieren, im Einzelfall können Sie jedoch durch Simulationen überprüft werden. Die Konsistenz wird für Lineare und Logistische Modelle in Kapitel 6 mit Hilfe einer Simulationsstudie näher untersucht.

5.4 Asymptotische Verteilung des Schätzers

Wir können nun zeigen, dass $\widehat{\boldsymbol{\beta}}^{\mathrm{SEL}}$ asymptotisch normalverteilt ist. Dabei gehen wir ähnlich vor wie Qin, Leung und Shao (2002) in ihrem Beweis zu Theorem 2. Der Notation der Autoren folgend sei wieder

$$\boldsymbol{\gamma} = \boldsymbol{\lambda}_1(1-W), \qquad \boldsymbol{\eta} = (\boldsymbol{\theta}^\top, W, \boldsymbol{\gamma}^\top)^\top, \qquad \boldsymbol{\eta}_0 = (\boldsymbol{\theta}_0^\top, W_0, \mathbf{0}^\top)^\top, \qquad a_N = \frac{N}{n} - \frac{1}{W_0}.$$

Dann ist

$$np_i = \frac{1-W}{1 - \frac{W}{W_0} + \frac{1-W_0}{W_0} w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) + \boldsymbol{\gamma}^\top(\boldsymbol{x}_i - \boldsymbol{\mu}_{\boldsymbol{X}}) + a_N(w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) - W)}.$$

Sei nun

$$\begin{aligned}\boldsymbol{\Xi}^n(\boldsymbol{\beta}, \boldsymbol{\eta}, a_N) &:= \frac{1}{n}\sum_{i=1}^n \boldsymbol{\xi}_i(\boldsymbol{\beta}, \boldsymbol{\eta}, a_N) \\ &:= \frac{1}{n}\sum_{i=1}^n \frac{1-W}{1 - \frac{W}{W_0} + \frac{1-W_0}{W_0} w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) + \boldsymbol{\gamma}^\top(\boldsymbol{x}_i - \boldsymbol{\mu}_{\boldsymbol{X}}) + a_N(w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) - W)} \boldsymbol{s}_i(\boldsymbol{\beta})\end{aligned}$$

und $\Xi_j^n(\boldsymbol{\beta}, \boldsymbol{\eta}, a_N)$ die j-te Komponente von $\boldsymbol{\Xi}^n(\boldsymbol{\beta}, \boldsymbol{\eta}, a_N)$. Dann ist

$$\boldsymbol{\Xi}^n(\boldsymbol{\beta}, \widehat{\boldsymbol{\eta}}, a_N) = \boldsymbol{\Psi}^n(\boldsymbol{\beta}).$$

Außerdem sei $\boldsymbol{\beta}_0$ der wahre Wert des unbekannten Parameters $\boldsymbol{\beta}$.

Satz 5.10:

Seien die Annahmen aus Satz 5.1 erfüllt und es gelte außerdem

(A1) $\frac{\partial^2 \Xi_j^n(\boldsymbol{\beta}^j, \boldsymbol{\eta}^j, a_n^j)}{\partial \boldsymbol{\beta}^\top \partial \boldsymbol{\beta}}$, $\frac{\partial^2 \Xi_j^n(\boldsymbol{\beta}^j, \boldsymbol{\eta}^j, a_n^j)}{\partial \boldsymbol{\eta}^\top \partial \boldsymbol{\eta}}$ und $\frac{\partial^2 \Xi_j^n(\boldsymbol{\beta}^j, \boldsymbol{\eta}^j, a_n^j)}{(\partial a_N)^2}$ existieren und sind beschränkt durch eine integrierbare Funktion $\forall\, j$.

(A2) $\mathrm{E}_{Y,\boldsymbol{X}}\left(\frac{\partial s(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}^\top}\right)$ existiert und ist invertierbar.

Dann folgt für den Schätzer $\widehat{\boldsymbol{\beta}}^{\mathrm{SEL}}$:

$$\sqrt{N}\left(\widehat{\boldsymbol{\beta}}^{\mathrm{SEL}} - \boldsymbol{\beta}_0\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \boldsymbol{F}^{-1}\boldsymbol{H}(\boldsymbol{F}^{-1})^\top\right),$$

wobei

$$\begin{aligned}\boldsymbol{F} &= \mathrm{E}_{Y,\boldsymbol{X}}\left(-\frac{\partial s(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}^\top}\right) \\ \text{und} \quad \boldsymbol{H} &= \mathrm{Var}\left(R\left(\frac{\boldsymbol{s}(\boldsymbol{\beta}_0)}{w(Y, \boldsymbol{X}, \boldsymbol{\theta}_0)}\right) + R\boldsymbol{b}_1\boldsymbol{U}^{-1}\boldsymbol{g}_i + \left(\boldsymbol{b}_1\boldsymbol{U}^{-1}\boldsymbol{h} + \boldsymbol{b}_2\right) a_N\right).\end{aligned}$$

Beweis: Eine komponentenweise Taylorentwicklung von $\boldsymbol{\Xi}^n(\widehat{\boldsymbol{\beta}}^{\mathrm{SEL}}, \widehat{\boldsymbol{\eta}}, a_N)$ um den Entwicklungspunkt $(\boldsymbol{\beta}_{\mathbf{0}}^\top, \boldsymbol{\eta}_0^\top, 0)$ ergibt

$$
\begin{aligned}
0 &= \Xi_j^n(\widehat{\boldsymbol{\beta}}^{\mathrm{SEL}}, \widehat{\boldsymbol{\eta}}, a_N) \\
&= \Xi_j^n(\boldsymbol{\beta}_0, \boldsymbol{\eta}_0, 0) \\
&\quad + \frac{\partial \Xi_j^n(\boldsymbol{\beta}_0, \boldsymbol{\eta}_0, 0)}{\partial \boldsymbol{\beta}^\top} \left(\widehat{\boldsymbol{\beta}}^{\mathrm{SEL}} - \boldsymbol{\beta}_0\right) + \frac{\partial \Xi_j^n(\boldsymbol{\beta}_0, \boldsymbol{\eta}_0, 0)}{\partial \boldsymbol{\eta}^\top} (\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0) + \frac{\partial \Xi_j^n(\boldsymbol{\beta}_0, \boldsymbol{\eta}_0, 0)}{\partial a_N} (a_N - 0) \\
&\quad + \frac{1}{2} \left(\widehat{\boldsymbol{\beta}}^{\mathrm{SEL}} - \boldsymbol{\beta}_0\right)^\top \frac{\partial^2 \Xi_j^n(\tilde{\boldsymbol{\beta}}^j, \tilde{\boldsymbol{\eta}}^j, \tilde{a}_n^j)}{\partial \boldsymbol{\beta}^\top \partial \boldsymbol{\beta}} \left(\widehat{\boldsymbol{\beta}}^{\mathrm{SEL}} - \boldsymbol{\beta}_0\right) \\
&\quad + \frac{1}{2} (\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0)^\top \frac{\partial^2 \Xi_j^n(\tilde{\boldsymbol{\beta}}^j, \tilde{\boldsymbol{\eta}}^j, \tilde{a}_n^j)}{\partial \boldsymbol{\eta}^\top \partial \boldsymbol{\eta}} (\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0) \\
&\quad + \frac{1}{2} (a_N - 0)^2 \frac{\partial^2 \Xi_j^n(\tilde{\boldsymbol{\beta}}^j, \tilde{\boldsymbol{\eta}}^j, \tilde{a}_n^j)}{(\partial a_N)^2} \\
&\overset{(*)}{=} \Xi_j^n(\boldsymbol{\beta}_0, \boldsymbol{\eta}_0, 0) \\
&\quad + \frac{\partial \Xi_j^n(\boldsymbol{\beta}_0, \boldsymbol{\eta}_0, 0)}{\partial \boldsymbol{\beta}^\top} \left(\widehat{\boldsymbol{\beta}}^{\mathrm{SEL}} - \boldsymbol{\beta}_0\right) + \frac{\partial \Xi_j^n(\boldsymbol{\beta}_0, \boldsymbol{\eta}_0, 0)}{\partial \boldsymbol{\eta}^\top} (\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0) + \frac{\partial \Xi_j^n(\boldsymbol{\beta}_0, \boldsymbol{\eta}_0, 0)}{\partial a_N} (a_N - 0) \\
&\quad + o_p(1) \left(\widehat{\boldsymbol{\beta}}^{\mathrm{SEL}} - \boldsymbol{\beta}_0\right) + o_p\left(N^{-\frac{1}{2}}\right) \\
&= \Xi_j^n(\boldsymbol{\beta}_0, \boldsymbol{\eta}_0, 0) + \left(\frac{\partial \Xi_j^n(\boldsymbol{\beta}_0, \boldsymbol{\eta}_0, 0)}{\partial \boldsymbol{\beta}^\top} + o_p(1)\right) \left(\widehat{\boldsymbol{\beta}}^{\mathrm{SEL}} - \boldsymbol{\beta}_0\right) \\
&\quad + \frac{\partial \Xi_j^n(\boldsymbol{\beta}_0, \boldsymbol{\eta}_0, 0)}{\partial \boldsymbol{\eta}^\top} (\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0) + \frac{\partial \Xi_j^n(\boldsymbol{\beta}_0, \boldsymbol{\eta}_0, 0)}{\partial a_N} (a_N - 0) + o_p\left(N^{-\frac{1}{2}}\right),
\end{aligned}
$$

wobei $\left(\tilde{\boldsymbol{\beta}}^{j\top}, \tilde{\boldsymbol{\eta}}^{j\top}, \tilde{a}_N^j\right)^\top$ eine Konvexkombination von $\left(\widehat{\boldsymbol{\beta}}^{\mathrm{SEL}\top}, \widehat{\boldsymbol{\eta}}^\top, a_N\right)^\top$ und $\left(\boldsymbol{\beta}_0^\top, \boldsymbol{\eta}_0^\top, 0\right)^\top$ ist. Dabei gilt $(*)$ wegen

$$
\begin{aligned}
\frac{\partial^2 \Xi_j^n(\tilde{\boldsymbol{\beta}}^j, \tilde{\boldsymbol{\eta}}^j, \tilde{a}_n^j)}{\partial \boldsymbol{\beta}^\top \partial \boldsymbol{\beta}} &= O_p(1), \\
\frac{\partial^2 \Xi_j^n(\tilde{\boldsymbol{\beta}}^j, \tilde{\boldsymbol{\eta}}^j, \tilde{a}_n^j)}{\partial \boldsymbol{\eta}^\top \partial \boldsymbol{\eta}} &= O_p(1) \\
\text{und} \quad \frac{\partial^2 \Xi_j^n(\tilde{\boldsymbol{\beta}}^j, \tilde{\boldsymbol{\eta}}^j, \tilde{a}_n^j)}{(\partial a_N)^2} &= O_p(1)
\end{aligned}
$$

sowie mit (vgl. Qin, Leung und Shao 2002)

$$
(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0) = O_p\left(N^{-\frac{1}{2}}\right) \quad \text{und} \quad (a_N - 0) = O_p\left(N^{-\frac{1}{2}}\right).
$$

Dann folgt

$$
\sqrt{N}\left(\widehat{\boldsymbol{\beta}}^{\mathrm{SEL}} - \boldsymbol{\beta}_0\right) = \sqrt{N} \boldsymbol{C}_n^{-1} \boldsymbol{\zeta}_n + o_p(1),
$$

wobei

$$\begin{aligned}\boldsymbol{\zeta}_n &= \boldsymbol{\Xi}^n(\boldsymbol{\beta}_0, \boldsymbol{\eta}_0, 0) + \frac{\partial \boldsymbol{\Xi}^n(\boldsymbol{\beta}_0, \boldsymbol{\eta}_0, 0)}{\partial \boldsymbol{\eta}^\top}(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0) + \frac{\partial \boldsymbol{\Xi}^n(\boldsymbol{\beta}_0, \boldsymbol{\eta}_0, 0)}{\partial a_N}(a_N - 0) \\ &= \frac{1}{n}\sum_{i=1}^{n}\left[\boldsymbol{\xi}_i(\boldsymbol{\beta}_0, \boldsymbol{\eta}_0, 0) + \frac{\partial \boldsymbol{\xi}_i(\boldsymbol{\beta}_0, \boldsymbol{\eta}_0, 0)}{\partial \boldsymbol{\eta}^\top}(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0) + \frac{\partial \boldsymbol{\xi}_i(\boldsymbol{\beta}_0, \boldsymbol{\eta}_0, 0)}{\partial a_N}(a_N - 0)\right].\end{aligned}$$

Die entsprechenden Ableitungen sind durch

$$\begin{aligned}\frac{\partial \boldsymbol{\xi}_i(\boldsymbol{\beta}_0, \boldsymbol{\eta}_0, 0)}{\partial \boldsymbol{\eta}^\top} &= \left(\frac{\partial \boldsymbol{\xi}_i(\boldsymbol{\beta}_0, \boldsymbol{\eta}_0, 0)}{\partial \boldsymbol{\theta}^\top}, \frac{\partial \boldsymbol{\xi}_i(\boldsymbol{\beta}_0, \boldsymbol{\eta}_0, 0)}{\partial W}, \frac{\partial \boldsymbol{\xi}_i(\boldsymbol{\beta}_0, \boldsymbol{\eta}_0, 0)}{\partial \gamma}\right)^\top, \\ \frac{\partial \boldsymbol{\xi}_i(\boldsymbol{\beta}_0, \boldsymbol{\eta}_0, 0)}{\partial \boldsymbol{\theta}^\top} &= -\frac{W_0 \cdot \partial w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}_0)/\partial \boldsymbol{\theta}^\top}{(w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}_0))^2}\boldsymbol{s}_i(\boldsymbol{\beta}_0), \\ \frac{\partial \boldsymbol{\xi}_i(\boldsymbol{\beta}_0, \boldsymbol{\eta}_0, 0)}{\partial W} &= \frac{W_0(1 - w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}_0))}{(1 - W_0)(w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}_0))^2}\boldsymbol{s}_i(\boldsymbol{\beta}_0), \\ \frac{\partial \boldsymbol{\xi}_i(\boldsymbol{\beta}_0, \boldsymbol{\eta}_0, 0)}{\partial \gamma} &= -\frac{W_0^2(\boldsymbol{x}_i - \boldsymbol{\mu}_{\boldsymbol{X}})}{(1 - W_0)(w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}_0))^2}\boldsymbol{s}_i(\boldsymbol{\beta}_0)\end{aligned}$$

und

$$\frac{\partial \boldsymbol{\xi}_i(\boldsymbol{\beta}_0, \boldsymbol{\eta}_0, 0)}{\partial a_N} = -\frac{W_0^2(w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}_0) - W_0)}{(1 - W_0)(w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}_0))^2}\boldsymbol{s}_i(\boldsymbol{\beta}_0)$$

gegeben. Außerdem ist

$$\boldsymbol{C}_n = -\left(\frac{\partial \boldsymbol{\Xi}^n(\boldsymbol{\beta}_0, \boldsymbol{\eta}_0, 0)}{\partial \boldsymbol{\beta}^\top} + o_p(1)\right).$$

Dabei konvergiert $\boldsymbol{C}_n$ gegen die Fisher-Informationsmatrix $\boldsymbol{F}$ von $f(y|\boldsymbol{x}, \boldsymbol{\beta})$, denn

$$\boldsymbol{\xi}_i(\boldsymbol{\beta}_0, \boldsymbol{\eta}_0, 0) = \frac{W_0}{w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}_0)}\boldsymbol{s}_i(\boldsymbol{\beta}_0)$$

und es folgt mit dem Gesetz der Großen Zahlen

$$\begin{aligned}\boldsymbol{C}_n &= -\left(\frac{\partial \boldsymbol{\Xi}^n(\boldsymbol{\beta}_0, \boldsymbol{\eta}_0, 0)}{\partial \boldsymbol{\beta}^\top} + o_p(1)\right) = -\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial \boldsymbol{\beta}^\top}\boldsymbol{\xi}_i(\boldsymbol{\beta}, \boldsymbol{\eta}, a_N) + o_p(1)\right) \\ &\xrightarrow{p} \mathrm{E}_{Y,\boldsymbol{X}|R=1}\left(-\frac{W_0}{w(Y, \boldsymbol{X}, \boldsymbol{\theta}_0)}\frac{\partial \boldsymbol{s}(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}^\top}\right) = \mathrm{E}_{Y,\boldsymbol{X}}\left(-\frac{\partial \boldsymbol{s}(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}^\top}\right) \\ &= \mathrm{E}_{Y,\boldsymbol{X}}\left(-\frac{\partial^2 \ln f(Y|\boldsymbol{X}, \boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^\top}\right) = \boldsymbol{F}.\end{aligned} \tag{5.16}$$

Seien

$$\boldsymbol{b}_1 = \mathrm{E}_{Y,\boldsymbol{X}|R=1}\left(\frac{\partial \boldsymbol{\xi}_i(\boldsymbol{\beta}_0, \boldsymbol{\eta}_0, 0)}{\partial \boldsymbol{\eta}^\top}\right) \quad \text{und} \quad \boldsymbol{b}_2 = \mathrm{E}_{Y,\boldsymbol{X}|R=1}\left(\frac{\partial \boldsymbol{\xi}_i(\boldsymbol{\beta}_0, \boldsymbol{\eta}_0, 0)}{\partial a_N}\right).$$

Dann gilt analog zum Beweis von Theorem 2 aus Qin, Leung und Shao (2002)

$$\begin{aligned}
\boldsymbol{\zeta}_n &= \frac{1}{n}\sum_{i=1}^{n}\left[\boldsymbol{\xi}_i(\boldsymbol{\beta}_0,\boldsymbol{\eta}_0,0) + \frac{\partial\boldsymbol{\xi}_i(\boldsymbol{\beta}_0,\boldsymbol{\eta}_0,0)}{\partial\boldsymbol{\eta}^\top}\left(\widehat{\boldsymbol{\eta}}-\boldsymbol{\eta}_0\right) + \frac{\partial\boldsymbol{\xi}_i(\boldsymbol{\beta}_0,\boldsymbol{\eta}_0,0)}{\partial a_N}\left(a_N-0\right)\right] \\
&= \frac{1}{W_0}\frac{1}{N}\sum_{i=1}^{N} r_i\left(\frac{W_0}{w(y_i,\boldsymbol{x}_i,\boldsymbol{\theta}_0)}\boldsymbol{s}_i(\boldsymbol{\beta}_0)\right) + \boldsymbol{b}_1\boldsymbol{U}_N^{-1}\boldsymbol{\phi}_N + \boldsymbol{b}_2 a_N + o_p\left(N^{-\frac{1}{2}}\right) \\
&= \frac{1}{N}\sum_{i=1}^{N} r_i\left(\frac{\boldsymbol{s}_i(\boldsymbol{\beta}_0)}{w(y_i,\boldsymbol{x}_i,\boldsymbol{\theta}_0)}\right) \\
&\quad + \frac{1}{N}\sum_{i=1}^{N}\left[r_i\boldsymbol{b}_1\boldsymbol{U}^{-1}\boldsymbol{g}_i + \left(\boldsymbol{b}_1\boldsymbol{U}^{-1}\boldsymbol{h}+\boldsymbol{b}_2\right)a_N\right] + o_p\left(N^{-\frac{1}{2}}\right),
\end{aligned}$$

und

$$\begin{aligned}
\mathrm{E}_{Y,\boldsymbol{X}|R=1}\left(\boldsymbol{\xi}_i(\boldsymbol{\beta}_0,\boldsymbol{\eta}_0,0)\right) &= \mathrm{E}_{Y,\boldsymbol{X}|R=1}\left(\frac{W_0}{w(Y,\boldsymbol{X},\boldsymbol{\theta}_0)}\boldsymbol{s}(\boldsymbol{\beta}_0)\right) \\
&= \mathrm{E}_{Y,\boldsymbol{X}}\left(\boldsymbol{s}(\boldsymbol{\beta}_0)\right) \\
&= \mathbf{0}
\end{aligned}$$

sowie außerdem

$$\begin{aligned}
\mathrm{E}_{Y,\boldsymbol{X}|R=1}\left(\boldsymbol{b}_1\boldsymbol{U}^{-1}\boldsymbol{g}_i\right) &= \boldsymbol{b}_1\boldsymbol{U}^{-1}\mathrm{E}_{Y,\boldsymbol{X}|R=1}\left(\boldsymbol{g}_i\right) \\
&= \mathbf{0}
\end{aligned}$$

und

$$\begin{aligned}
\mathrm{E}_{Y,\boldsymbol{X},R}\left(\left(\boldsymbol{b}_1\boldsymbol{U}^{-1}\boldsymbol{h}+\boldsymbol{b}_2\right)a_N\right) &= \left(\boldsymbol{b}_1\boldsymbol{U}^{-1}\boldsymbol{h}+\boldsymbol{b}_2\right)\mathrm{E}_{Y,\boldsymbol{X},R}\left(a_N\right) \\
&= \mathbf{0}.
\end{aligned}$$

Damit folgt

$$\sqrt{N}\boldsymbol{\zeta}_n \xrightarrow{p} \mathcal{N}\left(\mathbf{0},\boldsymbol{H}\right),$$

wobei

$$\boldsymbol{H} = \mathrm{Var}\left(R\left(\frac{\boldsymbol{s}(\boldsymbol{\beta}_0)}{w(Y,\boldsymbol{X},\boldsymbol{\theta}_0)}\right) + R\boldsymbol{d}_1\boldsymbol{g}_i + \boldsymbol{d}_2\frac{1}{W_0}\left(1-\frac{R}{W_0}\right)\right)$$

mit

$$\boldsymbol{d}_1 = \boldsymbol{b}_1\boldsymbol{U}^{-1} \quad \text{und} \quad \boldsymbol{d}_2 = \left(\boldsymbol{b}_1\boldsymbol{U}^{-1}\boldsymbol{h}+\boldsymbol{b}_2\right).$$

Schließlich folgt mit dem Slutzky-Theorem

$$\sqrt{N}\left(\widehat{\boldsymbol{\beta}}^{\mathrm{SEL}}-\boldsymbol{\beta}_0\right) \xrightarrow{p} \mathcal{N}\left(\mathbf{0},\boldsymbol{F}^{-1}\boldsymbol{H}\boldsymbol{F}^{-1}\right). \tag{5.17}$$

■

Bemerkung 5.4:

Falls $\boldsymbol{\mu_X}$ unbekannt ist und durch das arithmetische Mittel $\bar{\boldsymbol{X}}$ geschätzt wird, ergibt sich die asymptotische Normalverteilung analog, indem man in $\boldsymbol{g}_i$ und $\boldsymbol{h}$ jeweils $\boldsymbol{\mu_X}$ durch $\bar{\boldsymbol{X}}$ ersetzt. Dadurch ändert sich die Kovarianzmatrix, denn für die Komponente $\boldsymbol{H}$ gilt dann

$$\boldsymbol{H} = \mathrm{Var}\Bigg(R\left(\frac{\boldsymbol{s}(\boldsymbol{\beta}_0)}{w(Y,\boldsymbol{X},\boldsymbol{\theta}_0)}\right) + R\boldsymbol{d}_1\boldsymbol{g}_i + \boldsymbol{d}_1\left(\boldsymbol{0}^\top, 0, \frac{W_0(\boldsymbol{\mu_X}-\boldsymbol{X})^\top}{1-W_0}\right)^\top + \boldsymbol{d}_2\frac{1}{W_0}\left(1-\frac{R}{W_0}\right)\Bigg).$$

Dies lässt sich wie im Beweis zu Satz 5.1 zeigen.

Bemerkung 5.5:

Die Informationsmatrix $\boldsymbol{F}$ aus Gleichung (5.17) kann nicht direkt geschätzt werden, da hierfür die Schätzung des Erwartungswerts bezüglich der gemeinsamen Verteilung von $(Y, \boldsymbol{X})$ nötig ist. Da jedoch nur die Beobachtungen aus der Verteilung $Y, \boldsymbol{X}|R = 1$ vorliegen, lässt sich $\boldsymbol{F}$ über eine Schätzung des bedingten Erwartungswerts

$$\mathrm{E}_{Y,\boldsymbol{X}|R=1}\left(-\frac{W_0}{w(Y,\boldsymbol{X},\boldsymbol{\theta}_0)}\frac{\partial \boldsymbol{s}(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}^\top}\right)$$

approximieren (vgl. Gleichung 5.16).

5.5 Ein Hausman-Test

Zur Überprüfung, ob das in diesem Kapitel beschriebene Vorgehen in einer Anwendung zur Schätzung der Modellparameter gerechtfertigt ist, lässt sich der *Hausman-Test* heranziehen (Hausman 1978). Der Spezifikationstest überprüft die Nullhypothese, ob das vorgestellte Verfahren und die herkömmliche Schätzung des Modells ohne Berücksichtigung der MNAR-Situation identische Ergebnisse liefern. In diesem Fall liegt keine Missspezifikation bei der üblichen ML-Schätzung vor.

Sei nun $\widehat{\boldsymbol{\beta}}^{\mathrm{SEL}}$ der vorgeschlagene Schätzer für den unbekannten Parametervektor $\boldsymbol{\beta}$ und $\widehat{\boldsymbol{\beta}}^{\mathrm{ML}}$ der gewöhnliche ML-Schätzer, der anhand der vollständig vorhandenen Beobachtungspaare $(y_i, \boldsymbol{x}_i)$ $(i = 1, \ldots, n)$ bestimmt wurde. Unter der Nullhypothese, dass keine Missspezifikation vorliegt, ist $\widehat{\boldsymbol{\beta}}^{\mathrm{ML}}$ asymptotisch unverzerrt und effizient, das heißt die Kovarianzmatrix des Schätzers erreicht die Cramér-Rao-Schranke. Der

Schätzer $\widehat{\boldsymbol{\beta}}^{\mathrm{SEL}}$ hingegen ist unter H_0 ineffizient, dafür aber unter H_1 noch konsistent – im Gegensatz zu $\widehat{\boldsymbol{\beta}}^{\mathrm{ML}}$. Unter der Nullhypothese gilt damit

$$\widehat{\boldsymbol{q}} := \widehat{\boldsymbol{\beta}}^{\mathrm{SEL}} - \widehat{\boldsymbol{\beta}}^{\mathrm{ML}} \xrightarrow{p} \mathbf{0}, \tag{5.18}$$

während unter der Alternative der Grenzwert von $\widehat{\boldsymbol{q}}$ von Null verschieden ist.

Hausman (1978) zeigt, dass unter der Nullhypothese die Kovarianz der Grenzverteilungen von $\widehat{\boldsymbol{\beta}}^{\mathrm{ML}}$ und $\widehat{\boldsymbol{q}}$ Null ist. Dazu nutzt er die asymptotische Effizienz von $\widehat{\boldsymbol{\beta}}^{\mathrm{ML}}$ aus. Mit Hilfe dieser Feststellung lässt sich die Varianz von $\widehat{\boldsymbol{q}}$ bestimmen, denn es gilt

$$\widehat{\boldsymbol{q}} + \widehat{\boldsymbol{\beta}}^{\mathrm{ML}} = \widehat{\boldsymbol{\beta}}^{\mathrm{SEL}}$$

und damit

$$\mathrm{Var}(\widehat{\boldsymbol{q}}) = \mathrm{Var}(\widehat{\boldsymbol{\beta}}^{\mathrm{SEL}}) - \mathrm{Var}(\widehat{\boldsymbol{\beta}}^{\mathrm{ML}}) \geq \mathbf{0}, \tag{5.19}$$

wobei die Ungleichung im Sinne der Löwner-Ordnung zu verstehen ist, das heißt $\mathrm{Var}(\widehat{\boldsymbol{\beta}}^{\mathrm{SEL}}) - \mathrm{Var}(\widehat{\boldsymbol{\beta}}^{\mathrm{ML}})$ ist nichtnegativ definit.

Schließlich zeigt Hausman (1978), dass die Teststatistik

$$Q = \widehat{\boldsymbol{q}}^{\top} \widehat{\mathrm{Var}(\widehat{\boldsymbol{q}})}^{-1} \widehat{\boldsymbol{q}} \tag{5.20}$$

unter H_0 asymptotisch χ^2-verteilt ist mit ebensovielen Freiheitsgraden wie $\hat{\boldsymbol{q}}$ Komponenten besitzt. In endlichen Stichproben kann es vorkommen, dass $\widehat{\mathrm{Var}(\widehat{\boldsymbol{q}})}$ nicht invertierbar oder negativ definit ist. Für eine Diskussion zu Ersterem siehe Krämer und Sonnberger (1986, S.88-92). Das Problem der negativen Definitheit kann sogar asymptotisch auftreten. Schreiber (2008) beschreibt dieses Problem und zeigt Wege auf, wie man Abhilfe schaffen kann.

Im vorliegenden Fall ist der ML-Schätzer zwar asymptotisch effizient, in endlichen Stichproben ist die Varianz des SEL-Schätzers – auch unter H_0 – nicht zwangsläufig größer. Für den ML-Schätzer können nicht alle Beobachtungen zur Bestimmung genutzt werden, während der neue Schätzer auch die Informationen aus den Beobachtungen verwendet, für die nur die Ausprägungen der Kovariablen vorliegen. Da die Varianz von $\widehat{\boldsymbol{q}}$ dadurch im Allgemeinen nicht identisch ist mit der Differenz der Varianzen der beiden Schätzer, sollte $\mathrm{Var}(\widehat{\boldsymbol{q}})$ anderweitig geschätzt werden, zum Beispiel durch Resampling-Verfahren wie Bootstrap oder Jackknife. Mit Hilfe des Hausman-Tests lässt sich dann untersuchen, ob für die Zielgröße des Modells eine MNAR-Situation vorliegt oder nicht.

Kapitel 6

Simulationsstudie

Nachdem nun Schätzer von Parametern in statistischen Modellen mit nicht-ignorierbar fehlender Zielgröße hergeleitet wurden, untersucht dieses Kapitel die Anwendbarkeit des vorgeschlagenen Schätzverfahrens. Dabei wird zunächst in einer Simulationsstudie die Güte der Schätzer in Abhängigkeit des Prozesses der fehlenden Daten verglichen. Wir unterscheiden dazu Situationen mit großem und kleinem Anteil fehlender Beobachtungen. Da sich die Auswirkungen des Fehlens einiger Datenpunkte in einem Linearen Modell anschaulich darstellen lassen, wird zunächst ein klassisches Regressionsmodell betrachtet. Im Anschluss folgt eine Simulation im Logistischen Regressionsmodell. Abschließend geben weitere Simulationen Aufschluss über die Eigenschaften des Hausman-Tests in beiden Modellen. Dazu wird als Nullhypothese sowohl der MCAR- sowie der MAR-Fall berücksichtigt, da in beiden Fällen keine Missspezifikation vorliegt.

6.1 Lineares Regressionsmodell

An dieser Stelle verdeutlicht eine Simulation das vorgestellte Verfahren zur Schätzung eines einfachen Linearen Modells. Es gelten folgende Annahmen:

$$X_1, \ldots, X_N \sim \mathcal{N}(3, 4),$$
$$Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i \quad (i = 1, \ldots, N),$$

wobei $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbb{I})$. Die Verteilung von $R|Y, \boldsymbol{X}$ sei nur von Y abhängig und durch ein logistisches Modell

$$w(y, x, \boldsymbol{\theta}) = P(R = 1|y, x, \boldsymbol{\theta}) = \frac{\exp(\theta_1 + \theta_2 y)}{1 + \exp(\theta_1 + \theta_2 y)}$$

gegeben, d. h.

$$R_i|Y_i = y_i, X_i = x_i \sim \mathcal{B}(1, \pi_i) \quad \text{mit } \pi_i = \frac{\exp(\theta_1 + \theta_2 y_i)}{1 + \exp(\theta_1 + \theta_2 y_i)} \quad (i = 1, \ldots, N).$$

Abbildung 6.1 veranschaulicht einen nach obigem Schema konstruierten Datensatz für $N = 10\,000$, $\beta_1 = 2$, $\beta_2 = 1$, $\theta_1 = 3$, $\theta_2 = -1$. In diesem Beispiel ist $n = 2\,467$, es fehlen also 75,33% der Beobachtungen von Y. Dabei fehlen anteilig mehr große Realisationen von Y als kleine. Die beobachteten Werte sind rot hervorgehoben. Die eingezeichneten Geraden sind zum einen die geschätzte KQ-Gerade aus allen, also auch den nicht beobachtbaren Werten (durchgezogene dunkelblaue Linie) und die geschätzte KQ-Gerade aus den tatsächlich beobachteten Werten (durchgezogene grüne Linie). Es ist eine deutliche Abweichung zu erkennen. Schätzt man also ein gewöhnliches Lineares Modell aus den beobachteten Daten, ist die resultierende Schätzung in diesem Beispiel verzerrt. Die Schätzung mit Hilfe der Methoden aus Kapitel 5 führt zur dritten eingezeichneten Geraden (gestrichelte hellblaue Linie). Diese liegt nah der KQ-Geraden, die man durch die unverzerrte Schätzung aus allen Beobachtungen erhielte.

Eine systematische Untersuchung der Güte der Schätzer aus Kapitel 5 im Vergleich zum gewöhnlichen KQ-Schätzer bzw. ML-Schätzer (unter der Annahme normalverteilter Störgrößen sind beide identisch) soll nun Aufschluss über die Vorteile der neuen Methode geben. Wir wählen wie im oben erwähnten Beispiel $\beta_1 = 2$ und $\beta_2 = 1$. Wir setzen $\theta_2 = -1$ während der Parameter θ_1 variiert, um verschiedene Situationen fehlender Daten zu erfassen. Es werden 1 000 Datensätze jeweils vom Umfang $N = 10\,000$ erzeugt und die Parameterschätzer bestimmt. Deren Güte wird mittels der mittleren Verzerrung, Varianz sowie einer Schätzung des Mittleren Quadratischen Fehlers

$$\mathrm{MSE}(\boldsymbol{\beta}) = \mathrm{E}\left((\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})^\top (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})\right)$$

gemessen.

Tabelle 6.1 zeigt die Ergebnisse der Simulation. Bei den vier betrachteten Szenarien tritt jeweils ein unterschiedlicher Anteil fehlender Daten auf, wobei dieser zwischen

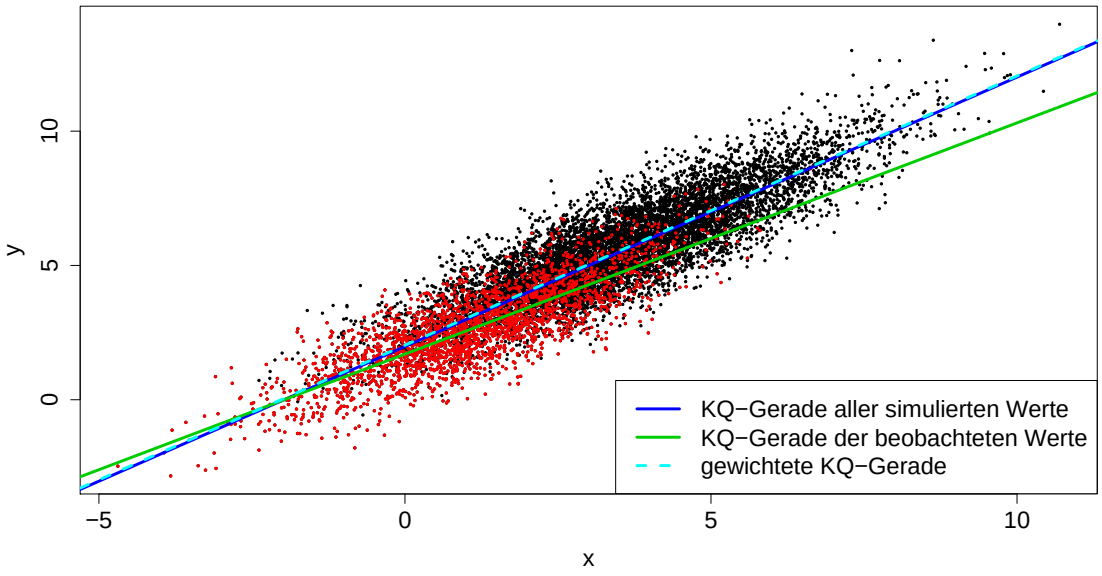

Abbildung 6.1 Beispieldatensatz für ein Lineares Modell mit nicht-ignorierbar fehlender Zielgröße. Dabei stellen die schwarzen Punkte die Werte dar, bei denen Y nicht beobachtet wird, die roten Punkte sind die vollständig beobachtbaren Werte.

75% und etwa 1% liegt. Falls ein großer Teil der Beobachtungen der Zielgröße nicht verfügbar ist, ist der gewöhnliche ML- bzw. KQ-Schätzer für β_1 und β_2 stark verzerrt, wohingegen die Verzerrung von $\widehat{\boldsymbol{\beta}}^{\text{SEL}}$ in allen Situationen deutlich geringer ist. Damit einher geht allerdings auch eine größere Varianz des SEL-Schätzers, wobei dieser Nachteil mit fallendem Anteil an unbeobachtbaren Werten an Bedeutung verliert. Der MSE verdeutlicht schließlich, dass der Schätzfehler durch den neuen Schätzer deutlich geringer ist. Dies ist auch in Abbildung 6.2 grafisch dargestellt. Für die vier betrachteten Situationen ist der MSE der Schätzer sowie der Anteil fehlender Daten abgetragen. Falls nur wenige Beobachtungspaare vollständig sind, ist der Vorteil des SEL-Schätzers besonders groß und nimmt immer mehr ab, je weniger Beobachtungen fehlen.

Tabelle 6.1 Ergebnisse der Simulation im Linearen Regressionsmodell für den ML-Schätzer aus den vollständig beobachteten Daten und den vorgestellten SEL-Schätzer.

Schätzung für	$\theta_1 = 3$	$\theta_1 = 5$	$\theta_1 = 7$	$\theta_1 = 9$
Anteil fehlender Daten	0.760	0.500	0.240	0.081
$\text{Bias}(\widehat{\beta}_1^{\text{ML}}) \cdot 10^3$	-287.707	-64.140	37.531	42.720
$\text{Bias}(\widehat{\beta}_1^{\text{SEL}}) \cdot 10^3$	34.362	9.763	3.409	1.005
$\text{Bias}(\widehat{\beta}_2^{\text{ML}}) \cdot 10^3$	-141.965	-116.642	-74.676	-35.333
$\text{Bias}(\widehat{\beta}_2^{\text{SEL}}) \cdot 10^3$	-11.136	-2.398	-1.137	-0.168
$\text{Var}(\widehat{\beta}_1^{\text{ML}}) \cdot 10^3$	0.549	0.414	0.381	0.317
$\text{Var}(\widehat{\beta}_1^{\text{SEL}}) \cdot 10^3$	12.685	2.700	0.954	0.390
$\text{Var}(\widehat{\beta}_2^{\text{ML}}) \cdot 10^3$	0.147	0.065	0.039	0.028
$\text{Var}(\widehat{\beta}_2^{\text{SEL}}) \cdot 10^3$	2.522	0.575	0.163	0.043
$\text{MSE}(\widehat{\boldsymbol{\beta}}^{\text{ML}}) \cdot 10^3$	103.624	18.198	7.405	3.417
$\text{MSE}(\widehat{\boldsymbol{\beta}}^{\text{SEL}}) \cdot 10^3$	16.497	3.372	1.128	0.434

6.2 Logistisches Regressionsmodell

Für die Simulationen des Logistischen Regressionsmodells gelten folgende Annahmen:

$$X_1, \ldots, X_N \sim N(0, 4),$$
$$Y_i \sim \mathcal{B}(1, \pi_i) \quad \text{mit } \pi_i = \frac{\exp(\beta_1 + \beta_2 x_i)}{1 + \exp(\beta_1 + \beta_2 x_i)} \quad (i = 1, \ldots, N).$$

Zunächst erzeugen wir Zufallszahlen $x_1, \ldots, x_N$ und bestimmen anschließend die Wahrscheinlichkeiten $\pi_i = P(Y_i = 1 | X_i = x_i)$ $(i = 1, \ldots, N)$. Mit Hilfe dieser Wahrscheinlichkeiten werden schließlich die Realisationen von Y generiert.

Die Verteilung von $R|Y, X$ sei wieder nur von Y abhängig und durch ein logistisches Modell

$$w(y, x, \boldsymbol{\theta}) = P(R = 1 | y, x, \boldsymbol{\theta}) = \frac{\exp(\theta_1 + \theta_2 y)}{1 + \exp(\theta_1 + \theta_2 y)}$$

gegeben, d. h. es gilt wieder

$$R_i | Y_i = y_i, X_i = x_i \sim \mathcal{B}(1, \pi_i) \quad \text{mit } \pi_i = \frac{\exp(\theta_1 + \theta_2 y_i)}{1 + \exp(\theta_1 + \theta_2 y_i)} \quad (i = 1, \ldots, N).$$

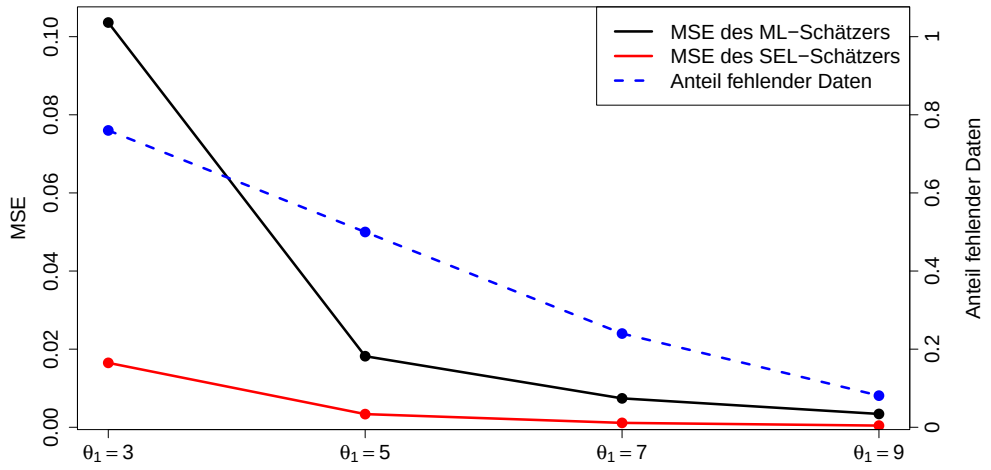

Abbildung 6.2 MSE des ML-Schätzers aus den vollständig beobachteten Daten und des vorgestellten SEL-Schätzers im Linearen Modell und der Zusammenhang zum Anteil fehlender Daten.

Wie im vorangegangenen Abschnitt wird nun die Güte der Schätzer aus Kapitel 5 im Vergleich zum gewöhnlichen ML-Schätzer untersucht. Sei dazu $\beta_1 = 2$ und $\beta_2 = -1$. Außerdem wählen wir $\theta_2 = 1$ und der Parameter θ_1 variiere wieder. Es werden wiederum 1 000 Datensätze jeweils vom Umfang $N = 10\,000$ erzeugt und die Parameterschätzer bestimmt.

Tabelle 6.2 zeigt die Ergebnisse der Simulation. Es werden fünf verschiedene Szenarien mit einem unterschiedlichen Anteil fehlender Daten betrachtet - dieser liegt hier zwischen 76.5% und etwa 1%. Bei großem Anteil fehlender Beobachtungen ist der gewöhnliche ML-Schätzer für β_1 auch hier stark verzerrt, wohingegen die Verzerrung des SEL-Schätzers in allen Situationen deutlich geringer ist. Der ML-Schätzer von β_2 scheint hingegen konsistent zu sein. Die Varianz von $\widehat{\boldsymbol{\beta}}^{\text{SEL}}$ für β_1 ist größer als die des ML-Schätzers, für den Parameter β_2 variiert die SEL-Schätzung nur leicht stärker. Wiederum wird die Differenz mit fallendem Anteil an unbeobachtbaren Werten geringer. Der MSE ist für den neuen Schätzer in allen Situationen geringer. Dies verdeutlicht auch Abbildung 6.3, welche für die verschiedenen Modelle den MSE sowie den Anteil an fehlenden Daten abträgt. Der Vorteil des SEL-Schätzers gegenüber dem ML-Schätzer ist bei vielen fehlenden Daten besonders groß, die Differenz des MSE nimmt mit zunehmendem Anteil vollständiger Beobachtungspaare ab.

Tabelle 6.2 Ergebnisse der Simulation im Logistischen Regressionsmodell für den ML-Schätzer aus den vollständig beobachteten Daten und den vorgestellten SEL-Schätzer.

Schätzung für	$\theta_1 = -2$	$\theta_1 = -1$	$\theta_1 = 0$	$\theta_1 = 1$	$\theta_1 = 2$
Anteil fehlender Daten	0.765	0.552	0.321	0.153	0.064
$\text{Bias}(\widehat{\beta}_1^{\text{ML}}) \cdot 10^3$	818.276	618.419	380.993	187.864	77.567
$\text{Bias}(\widehat{\beta}_1^{\text{SEL}}) \cdot 10^3$	3.200	-2.432	2.108	1.275	-0.202
$\text{Bias}(\widehat{\beta}_2^{\text{ML}}) \cdot 10^3$	-3.757	-0.568	-0.356	-0.127	0.543
$\text{Bias}(\widehat{\beta}_2^{\text{SEL}}) \cdot 10^3$	-4.291	-0.786	-0.322	-0.151	0.538
$\text{Var}(\widehat{\beta}_1^{\text{ML}}) \cdot 10^3$	11.883	5.406	2.778	2.149	1.834
$\text{Var}(\widehat{\beta}_1^{\text{SEL}}) \cdot 10^3$	21.521	8.537	4.089	2.475	1.988
$\text{Var}(\widehat{\beta}_2^{\text{ML}}) \cdot 10^3$	3.476	1.459	0.906	0.664	0.599
$\text{Var}(\widehat{\beta}_2^{\text{SEL}}) \cdot 10^3$	3.645	1.492	0.919	0.666	0.600
$\text{MSE}(\widehat{\boldsymbol{\beta}}^{\text{ML}}) \cdot 10^3$	684.933	389.301	148.836	38.102	8.448
$\text{MSE}(\widehat{\boldsymbol{\beta}}^{\text{SEL}}) \cdot 10^3$	25.169	10.026	5.007	3.140	2.586

6.3 Untersuchung des Hausman-Tests

Eine Anwendung des Hausman-Tests aus Abschnitt 5.5 soll nun zeigen, ob dieser Test die Missspezifikation des ML-Schätzers aufdecken kann. Unter der Nullhypothese liegt keine Missspezifikation des Modells vor. Das bedeutet, dass wir zwischen zwei verschiedenen Nullhypothesen unterscheiden können, nämlich MAR und MCAR. In beiden Fällen ist ein Regressionsmodell nicht falsch spezifiziert. Um also zu überprüfen, ob der Hausman-Test ein vorgegebenes Niveau einhält, können wir Daten aus einer MAR- oder MCAR-Situation generieren und überprüfen, wie häufig der Hausman-Test zu dem entsprechenden Signifikanzniveau tatsächlich verwirft. Um die Güte des Tests zu analysieren, bietet sich das gleiche Vorgehen bei Daten aus der Alternative (also MNAR) an. Dies geschieht nun getrennt für das Lineare Modell sowie für ein Logistisches Regressionsmodell. Dabei ist zu beachten, dass bei dem Hausman-Test sowohl die Nullhypothese als auch die Alternative sehr „breit" ist, das heißt, es gibt sehr viele unterschiedliche Spezifikationen aus den beiden Hypothesen. Daher erfolgt die Untersuchung des Tests hier nur exemplarisch.

Damit die Nullhypothesen MCAR und MAR überhaupt getrennt betrachtet werden

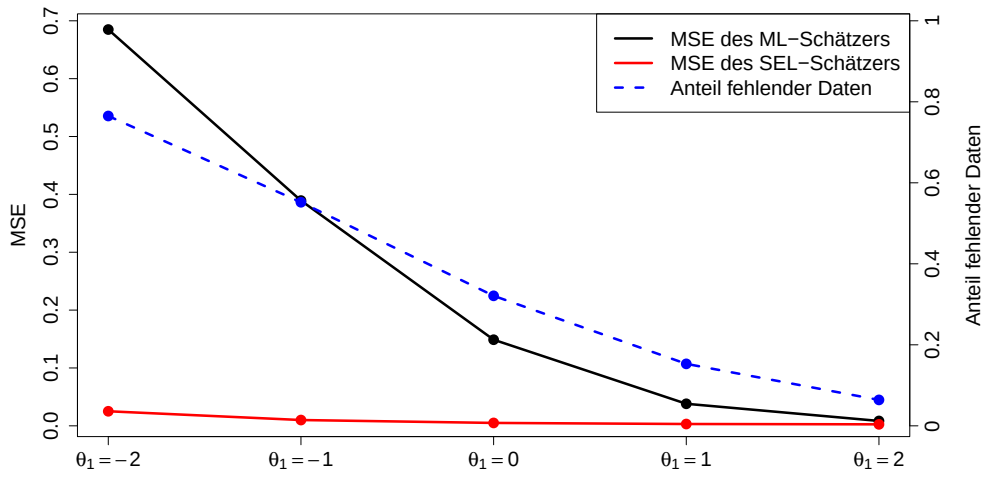

Abbildung 6.3 MSE des ML-Schätzers aus den vollständig beobachteten Daten und des vorgestellten SEL-Schätzers im Logistischen Regressionsmodell und der Zusammenhang zum Anteil fehlender Daten.

können, sollte im MAR-Fall das Fehlen auch von einer Kovariablen abhängen. Damit dies möglich ist, geben wir uns zwei Regressoren X_1 und X_2 vor, wobei im MAR-Fall R nur von X_1, die Zielgröße aber von beiden Einflussgrößen abhängt (vgl. dazu auch Bemerkung 5.1, S. 45).

In die Teststatistik des Hausman-Tests geht die Differenz der Varianzen der beiden Schätzer ein. In den Simulationen ist die resultierende Kovarianzmatrix aus Gleichung (5.19) aber in vielen Fällen nicht positiv definit. Dies führt unter Umständen zu einer negativen Teststatistik. Daher lassen sich Resampling-Verfahren wie Bootstrap oder Jackknife verwenden, um die Kovarianzmatrix der Differenz der Schätzer zu approximieren. In den folgenden Simulationen wird die Kovarianzmatrix per Bootstrap geschätzt. Dazu werden in jedem Simulationsdurchlauf $B = 200$ Bootstrap-Stichproben aller Beobachtungen generiert, für diese beide Schätzer bestimmt und die Varianz der Differenzen geschätzt. Alternativ lassen sich auch über die Residuen Bootstrap-Stichproben generieren und die Schätzer bestimmen.

Lineares Modell

Für die Simulationen wird dem Regressanden Y eine lineare Abhängigkeit von den beiden unabhängigen Variablen X_1 und X_2 mit einem additiven normalverteilten Störterm unterstellt. Es gilt für $i = 1, \ldots, N$ mit $N = 10\,000$

$$\begin{aligned} X_{1,i} &\sim \mathcal{N}(3, 4), \\ X_{2,i} &\sim \mathcal{N}(3, 4), \\ Y_i &= \beta_1 + \beta_2 X_{1,i} + \beta_2 X_{2,i} + \varepsilon_i, \\ (\varepsilon_1, \ldots, \varepsilon_N)^\top &= \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbb{I}) \end{aligned}$$

mit $\beta_1 = 2$ und $\beta_2 = \beta_3 = 1$. Die Verteilung von R ist durch ein Logistisches Regressionsmodell gegeben und hat die Form

$$R_i | Y_i = y_i, X_{1,i} = x_{1,i} \sim \mathcal{B}(1, \pi_i) \quad \text{mit } \pi_i = \frac{\exp(\theta_1 + \theta_2 y_i + \theta_3 x_{1,i})}{1 + \exp(\theta_1 + \theta_2 y_i + \theta_3 x_{1,i})}.$$

Der Parameter θ_1 variiere wieder, so dass Situationen mit unterschiedlich vielen fehlenden Daten vorliegen. Die Parameter θ_2 und θ_3 legen fest, ob die fehlenden Daten MCAR, MAR oder MNAR sind:

- MCAR: $\theta_2 = \theta_3 = 0$, $\theta_1 \in \{-1.3, -0.4, 0.4, 1.3\}$,
- MAR: $\theta_2 = 0$, $\theta_3 = -1$, $\theta_1 \in \{1, 2.5, 4, 5\}$,
- MNAR: $\theta_2 = \theta_3 = -1$, $\theta_1 \in \{8, 10, 12, 15\}$.

Für die unterschiedlichen Datensituationen wird nun jeweils die ML- und SEL-Schätzung bestimmt, die Varianz der Differenz per Bootstrap geschätzt und daraus die Teststatistik Q des Hausman-Tests aus Gleichung (5.20) berechnet. Dies wiederholen wir jeweils 1 000-fach. Zur Überprüfung, ob der Hausman-Test das Niveau einhält, beziehungsweise zur Untersuchung der Güte, vergleicht man die Ausprägungen der Teststatistik anschließend mit dem entsprechenden Quantil der χ^2-Verteilung, um die Verwerfungswahrscheinlichkeit

$$P\left(Q > \chi^2_{3;1-\alpha}\right)$$

Für die kritischen Werte der Grenzverteilung verwenden wir die üblichen Signifikanzniveaus $\alpha \in \{0.01, 0.05, 0.1\}$.

Tabelle 6.3 zeigt die Ablehnungshäufigkeiten unter der Nullhypothese (MCAR). Es zeigt sich, dass das Niveau für $\alpha = 0.01$ nicht ganz eingehalten wird, für größere

Tabelle 6.3 Relative Ablehnungshäufigkeiten des Hausman-Tests im Linearen Modell bei Gültigkeit der Nullhypothese (MCAR) bei 1 000 Simulationen.

	$\theta_1 = -1.3$	$\theta_1 = -0.4$	$\theta_1 = 0.4$	$\theta_1 = 1.3$
Anteil fehlender Daten	0.786	0.599	0.401	0.214
$\alpha = 0.01$	0.028	0.032	0.020	0.029
$\alpha = 0.05$	0.067	0.064	0.047	0.063
$\alpha = 0.10$	0.099	0.098	0.071	0.085

Niveaus liegen die Ablehnungshäufigkeiten nahe an den zu erwartenden Werten. Dies scheint unabhängig vom Anteil der fehlenden Daten zu sein.

Für den MAR-Fall ergeben sich die Ablehnungshäufigkeiten aus Tabelle 6.4. Das Niveau wird wieder weitestgehend eingehalten und ist unabhängig vom Anteil der fehlenden Daten. Für sehr kleines α lehnt der Test etwas zu häufig ab, das heißt, die Ränder der empirischen Verteilung der Teststatistik sind auch hier etwas dicker als die der χ^2_3-Verteilung.

Tabelle 6.5 verdeutlicht die Güte des Hausman-Tests, also die Wahrscheinlichkeit, H_0 bei Gültigkeit der Alternative abzulehnen. Die Daten wurden unter der MNAR Annahme generiert. Für nicht allzu großen Anteil fehlender Daten (weniger als 50%) wird die Nullhypothese in fast allen Fällen abgelehnt. Falls sehr viele Daten fehlen (73%), fällt die Ablehnungshäufigkeit drastisch ab. Dennoch liegt sie jeweils über dem geforderten Niveau.

Tabelle 6.4 Relative Ablehnungshäufigkeiten des Hausman-Tests im Linearen Modell bei Gültigkeit der Nullhypothese (MAR) bei 1 000 Simulationen.

	$\theta_1 = 1$	$\theta_1 = 2.5$	$\theta_1 = 4$	$\theta_1 = 5$
Anteil fehlender Daten	0.775	0.575	0.352	0.225
$\alpha = 0.01$	0.026	0.025	0.015	0.018
$\alpha = 0.05$	0.070	0.064	0.058	0.053
$\alpha = 0.10$	0.121	0.115	0.092	0.100

Tabelle 6.5 Relative Ablehnungshäufigkeiten des Hausman-Tests im Linearen Modell bei Gültigkeit der Alternative (MNAR) bei 1 000 Simulationen.

	$\theta_1 = 8$	$\theta_1 = 10$	$\theta_1 = 12$	$\theta_1 = 15$
Anteil fehlender Daten	0.730	0.581	0.419	0.208
$\alpha = 0.01$	0.158	0.645	0.951	0.992
$\alpha = 0.05$	0.319	0.820	0.981	0.997
$\alpha = 0.10$	0.422	0.882	0.985	0.998

Logistisches Regressionsmodell

Nun unterstellen wir ein Logistisches Regressionsmodell. Für $i = 1, \ldots, N$ mit $N = 10\,000$ gelte

$$\begin{aligned} X_{1,i} &\sim \mathcal{N}(0, 4), \\ X_{2,i} &\sim \mathcal{N}(0, 4), \\ Y_i | X_{1,i} = x_{1,i}, X_{2,i} = x_{2,i} &\sim \mathcal{B}(1, \pi_i) \\ \text{mit } \pi_i &= \frac{\exp(\beta_1 + \beta_2 x_{1,i} + \beta_3 x_{2,i})}{1 + \exp(\beta_1 + \beta_2 x_{1,i} + \beta_3 x_{2,i})}, \end{aligned}$$

wobei $\beta_1 = 2$, $\beta_2 = \beta_3 = -1$.

Die Verteilung von R sei wieder durch ein Logistisches Regressionsmodell gegeben und habe die Form

$$R_i | Y_i = y_i, X_{1,i} = x_{1,i} \sim \mathcal{B}(1, \pi_i) \quad \text{mit } \pi_i = \frac{\exp(\theta_1 + \theta_2 y_i + \theta_3 x_{1,i})}{1 + \exp(\theta_1 + \theta_2 y_i + \theta_3 x_{1,i})}.$$

Der Parameter θ_1 wird wieder jeweils so variiert, dass Situationen mit unterschiedlich vielen fehlenden Daten vorliegen. Die Parameter θ_2 und θ_3 legen fest, ob die fehlenden Daten MCAR, MAR oder MNAR sind:

- MCAR: $\theta_2 = \theta_3 = 0$, $\theta_1 \in \{-1.3, -0.4, 0.4, 1.3\}$,
- MAR: $\theta_2 = 0$, $\theta_3 = -1$, $\theta_1 \in \{-2, -0.5, 0.5, 2\}$,
- MNAR: $\theta_2 = \theta_3 = -1$, $\theta_1 \in \{-2, 0, 1.5, 3\}$.

Tabelle 6.6 Relative Ablehnungshäufigkeiten des Hausman-Tests im Logistischen Regressionsmodell bei Gültigkeit der Nullhypothese (MCAR) bei 1 000 Simulationen.

	$\theta_1 = -1.3$	$\theta_1 = -0.4$	$\theta_1 = 0.4$	$\theta_1 = 1.3$
Anteil fehlender Daten	0.786	0.599	0.401	0.214
$\alpha = 0.01$	0.047	0.046	0.031	0.035
$\alpha = 0.05$	0.082	0.084	0.059	0.064
$\alpha = 0.10$	0.114	0.116	0.090	0.086

Auch in diesem Fall simulieren wir wieder die unterschiedlichen Datensituationen 1 000-fach und berechnen jeweils die Teststatistik Q des Hausman-Tests aus Gleichung (5.20). Daraus bestimmen wir anschließend die relative Ablehnungshäufigkeit zu den üblichen Signifikanzniveaus $\alpha \in \{0.01, 0.05, 0.1\}$.

Die Tabellen 6.6, 6.7 und 6.8 spiegeln die Ergebnisse der Simulationen zum Logistischen Regressionsmodell wider. Es zeigt sich, dass im MCAR-Fall (Tabelle 6.6) das Niveau bei kleinem Fehler 2. Art überschritten und für $\alpha = 0.1$ besser eingehalten wird. Für MAR-Daten (Tabelle 6.7) ergibt sich ein ähnliches Bild, wobei die Ablehnungshäufigkeit näher am geforderten Niveau liegt. Dies gilt für die vier betrachteten Situationen gleichermaßen. Die Untersuchung der Güte des Hausman-Tests (Tabelle 6.8) deutet auf eine größere Macht als im Linearen Modell hin. Falls weniger als 40% der Daten fehlen, kann die Nullhypothese in (fast) allen Fällen abgelehnt werden. Selbst wenn sogar etwa 86% der Beobachtungen der Zielgröße fehlen, wird die Nullhypothese zum Niveau $\alpha = 0.01$ immer noch in über 25% der Fälle verworfen.

Tabelle 6.7 Relative Ablehnungshäufigkeiten des Hausman-Tests im Logistischen Regressionsmodell bei Gültigkeit der Nullhypothese (MAR) bei 1 000 Simulationen.

	$\theta_1 = -2$	$\theta_1 = -0.5$	$\theta_1 = 0.5$	$\theta_1 = 2$
Anteil fehlender Daten	0.775	0.575	0.425	0.225
$\alpha = 0.01$	0.010	0.017	0.018	0.028
$\alpha = 0.05$	0.067	0.059	0.063	0.076
$\alpha = 0.10$	0.116	0.100	0.110	0.125

Tabelle 6.8 Relative Ablehnungshäufigkeiten des Hausman-Tests im Logistischen Regressionsmodell bei Gültigkeit der Alternative (MNAR) bei 1 000 Simulationen.

	$\theta_1 = -2$	$\theta_1 = 0$	$\theta_1 = 1.5$	$\theta_1 = 3$
Anteil fehlender Daten	0.858	0.615	0.381	0.184
$\alpha = 0.01$	0.265	0.956	0.996	1.000
$\alpha = 0.05$	0.481	0.989	1.000	1.000
$\alpha = 0.10$	0.596	0.995	1.000	1.000

Kapitel 7

Anwendung

Ein Datensatz aus der Kreditvergabepraxis einer größeren deutschen Geschäftsbank, die aus betriebsinternen Gründen nicht näher genannt werden will, soll in diesem Kapitel Aufschluss darüber geben, inwiefern die vorgestellten Methoden in der Realität anwendbar sind. Der Datensatz enthält für alle Kunden die Ausprägungen der Scoremerkmale, wobei die Zielgröße, also ob ein Kunde den Kredit an die Bank zurückzahlen konnte oder nicht, für die abgelehnten Kunden nicht verfügbar ist. Die Logistische Regression ist eine sehr verbreitete Methode im Kreditscoring und soll daher hier zur Anwendung kommen.

In Kapitel 3 wurde argumentiert, warum die Bonität der abgelehnten Kunden als nicht-ignorierbar fehlend betrachtet werden kann. Eine Modellbildung ausschließlich auf Basis der Daten akzeptierter Kunden würde folglich eine verzerrte Schätzung liefern. Mit Hilfe des neuen Schätzverfahrens können nun die Parameter eines solchen Modells unverzerrt geschätzt werden. Anschließend lässt sich mit Hilfe des Hausman-Tests überprüfen, ob dieses Vorgehen gerechtfertigt ist oder ob nicht auch auf herkömmlichem Weg eine konsistente Schätzung möglich ist.

Ein Vergleich der Prognosegüte beider Methoden kann nur mit den vollständigen Daten erfolgen und ist daher für das herkömmliche Modell ohne Berücksichtigung des Fehlens zu optimistisch. Ein angemessener Vergleich ist nur möglich, falls auch Informationen über abgelehnte Kunden vorliegen. Da dies im vorliegenden Datensatz nicht der Fall ist, kann nur eine Gegenüberstellung der Prognosegüte anhand der vollständigen Daten durchgeführt werden.

7.1 Die Daten

Der verwendete Datensatz enthält neun kategorielle und vier stetige Scoremerkmale und als abhängige Variable die Bonität der Kunden. Insgesamt liegen $n = 4\,000$ der insgesamt $N = 9\,780$ Realisationen mit beobachtbarer Zielgröße vor, bei den übrigen 5 780 Beobachtungen fehlt die Ausprägung der Zielgröße. Die Einflussgrößen sind

- Alter in Jahren (metrisch)
- Familienstand (kategoriell)
  (Klassen: verheiratet, verwitwet, ledig, Lebensgem., geschieden, getrennt)
- Anzahl der Kinder (kategoriell)
  (Klassen: keine Kinder, 1 Kind, 2 Kinder, 3-6 Kinder)
- Beruf (kategoriell)
  (Klassen: Beamte, Angestellte, Facharbeiter, Selbständige, Rentner, Sonstiges)
- Arbeitsdauer in Monaten (metrisch)
- Haushaltseinkommen in Euro (metrisch)
- Mitantragsteller (kategoriell)
  (Klassen: kein Mitantragsteller, Mitantragsteller vorhanden)
- Kaufkraft (kategoriell)
  (Klassen: sehr hoch, hoch, mittel, niedrig, Sonstiges)
- Anzahl laufender Kredite (kategoriell)
  (Klassen: 1 Kredit, 2 Kredite, 3 Kredite, 4 oder mehr Kredite, kein Kredit, Sonstiges)
- Schufa Boni-Klasse (kategoriell)
  (Klassen: A-E, F-J, K-O, P, Sonstiges)
- Neukunde (kategoriell)
  (Klassen: Bestands- oder Neukunde)
- Haustyp (kategoriell)
  (Klassen: keine Familie, 1 Fam., 2 Fam., 3-5 Fam., 6-10 Fam., 11-14 Fam., 15-20 Fam., >20 Fam., Sonstiges)
- Wohndauer an der aktuellen Adresse in Monaten (metrisch)

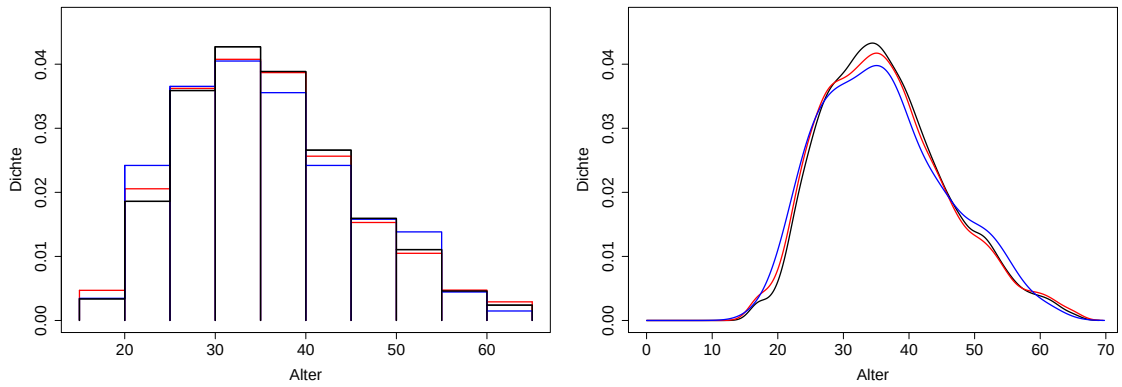

Abbildung 7.1 Einfluss der Variable „Alter" auf Kreditvergabe und Rückzahlung: Histogramm (links) und Kerndichteschätzung (rechts) für alle Kunden (schwarz), die abgelehnten Kunden (rot) und die schlechten Kunden (blau).

Die Abbildungen 7.1, 7.2 und B.1-B.11 (S. 105-110) veranschaulichen, inwiefern diese Variablen einen Einfluss auf die Akzeptanz der Bank und auf die Rückzahlfähigkeit der Kunden haben. Für die stetigen Merkmale sind jeweils Dichteschätzungen durch Histogramme und Kernschätzer abgebildet: farblich getrennt für alle Kunden, die abgelehnten Kunden und die Kunden mit Zahlungsschwierigkeiten. Dadurch lässt sich ermitteln, welche Kunden ein besonders hohes Ablehnungs- bzw. Ausfallrisiko hatten.

Für die kategoriellen Merkmale sind die Anteile der abgelehnten Kunden (unter allen Kunden) sowie die Anteile der ausgefallenen Kredite (unter allen vergebenen Krediten) in jeder Kategorie als Balken dargestellt. Die Skalierung ist so gewählt, dass der durchschnittliche Anteil abgelehnter Kunden bzw. ausgefallener Kredite durch eine gemeinsame Referenzlinie dargestellt werden kann. Dadurch lässt sich feststellen, in welcher Gruppe überdurchschnittlich viele Kunden abgelehnt bzw. Kredite nicht zurückgezahlt wurden.

Es zeigt sich, dass besonders Kunden mittleren Alters zwischen 30 und 40 seltener abgelehnt und zahlungsunfähig wurden (vgl. Abbildung 7.1). Abbildung 7.2 lässt sich entnehmen, dass der Familienstand offenbar kaum einen Einfluss auf die Vergabe eines Kredits durch die Bank hat, das Ausfallrisiko hingegen ist für verwitwete oder getrennt lebende Kunden deutlich erhöht, während Kunden in Lebensgemeinschaften sehr selten ausfallen. Mit wachsender Anzahl an Kindern scheint die Ablehnungswahrscheinlichkeit durch die Bank ebenfalls zu wachsen. Für die Fähigkeit zur

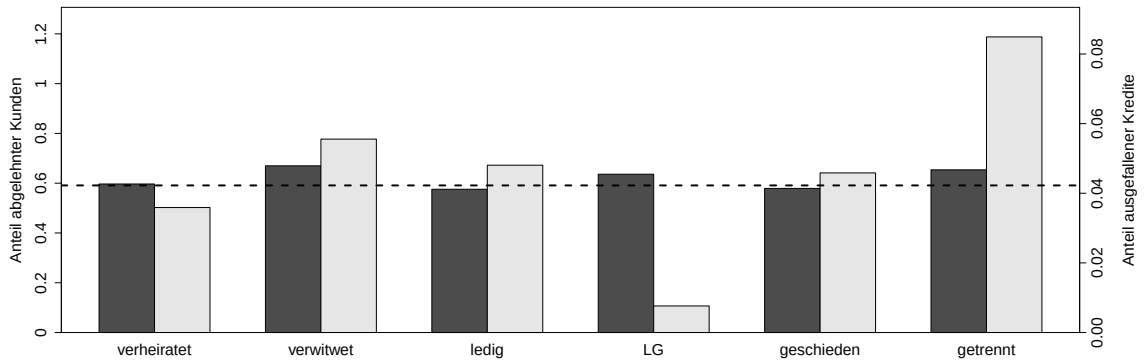

Abbildung 7.2 Einfluss der Variable „Familienstand" auf Kreditvergabe und Rückzahlung: die dunklen Balken geben den Anteil abgelehnter Kredite, die hellen Balken den Anteil ausgefallener Kredite in der jeweiligen Gruppe an, die Linie stellt den durchschnittlichen Anteil abgelehnter/ausgefallener Kredite dar.

Tilgung eines Kredits ergibt sich mit Abbildung B.1 ein anderes Bild. So fallen Kunden mit einem oder 3-6 Kindern seltener aus als solche mit keinen oder zwei Kindern. In der Gruppe der akzeptierten Kunden sind überdurchschnittlich viele Beamte und Angestellte, andere Berufsgruppen werden häufiger abgelehnt. Bei der Rückzahlung sind besonders Facharbeiter und Selbständige öfter säumig, während Beamte und Renter nur sehr selten in Zahlungsschwierigkeiten geraten (vgl. Abbildung B.2). Wie man in Abbildung B.3 erkennen kann, werden Kunden mit einer sehr geringen und sehr hohen Arbeitsdauer scheinbar häufig durch die Bank abgelehnt, auf das Ausfallrisiko scheint die Arbeitsdauer jedoch kaum einen Einfluss zu haben. Ein sehr hohes Einkommen führt mit einer geringen Wahrscheinlichkeit zu Ablehnung, obwohl nicht überdurchschnittlich viele Kunden dieser Gruppe den Kredit tilgen (Abbildung B.4). Ansonsten zeigt sich, dass unter den abgelehnten und zahlungsfähigen Kunden besonders viele mit niedrigem Einkommen sind. Das Vorhandensein eines Mitantragsstellers (Abbildung B.5) scheint für die Bank kein ausschlaggebendes Kriterium zur Vergabe eines Kredits zu sein, obwohl unter den Kunden mit einem Bürgen überdurchschnittlich viele kreditwürdig sind. Die Kaufkraft scheint bei Betrachtung von Abbildung B.6 ebenfalls für die Bank kein Entscheidungskriterium zu sein, obwohl die Ausfallrate für Kunden mit sehr hoher, hoher aber auch niedriger Kaufkraft geringer als der Durchschnitt ist. Die Anzahl der aufgenommen Kredite hat auf die Ablehnung nur einen moderaten Einfluss. Kunden ohne einen Kredit und mit vielen Krediten sowie mit der Angabe Sonstiges (keine Angabe) wer-

den häufiger abgelehnt als Kunden mit einem oder zwei Krediten (Abbildung B.7). Besonders häufig fallen Kunden mit sehr vielen Krediten oder der Angabe Sonstiges aus. Laut Abbildung B.8 scheinen die Schufa-Boni-Klassen ein guter Indikator für die Rückzahlfähigkeit der Kunden zu sein – so wächst das Risiko eines Ausfalls von der Kategorie A-E bis P schrittweise stark an. Die Entscheidung der Bank fiel dementsprechend aus, auch wenn die Ablehnungshäufigkeiten nicht ganz so stark ansteigen. Eine interessante Bedeutung hat die Variable Neukunde, welche in Abbildung B.9 veranschaulicht wird. So scheint zwischen Bestands- und Neukunden bei der Zahlungsfähigkeit kein großer Unterschied zu bestehen. Die Bank vergibt jedoch bevorzugt Kredite an Bestandskunden und lehnt neue Kunden häufiger ab. Die Variable Haustyp hat auf die Ablehnung durch die Bank nur einen geringen Einfluss. Gute Kunden mit überdurchschnittlicher Rückzahlwahrscheinlichkeit stammen aus den Gruppen mit Einfamilienhäusern, Häusern mit 3-5, 11-14 und mit über 20 Familien (Abbildung B.10). Die Wohndauer an der aktuellen Adresse scheint nicht auf die Entscheidung zur Kreditvergabe durch die Bank eingewirkt zu haben, wohl aber auf die Rückzahlfähigkeit, denn unter den Kunden mit kurzer Wohndauer befinden sich besonders viele Ausfälle (vgl. Abbildung B.11).

Nun soll ein Prognosemodell geschätzt werden. Für ein Logistisches Regressionsmodell werden – wie in Kapitel 5 beschrieben – Gewichte bestimmt, um unverzerrte Schätzer zu erhalten. Wie in Bemerkung 5.1 angesprochen, ist das Modell überparametrisiert, falls das Fehlen der Zielgröße von allen Kovariablen und der Zielgröße selbst abhängen darf. Daher gilt es herauszufinden, ob eine der Kovariablen möglicherweise keinen Einfluss auf das Fehlen hat. Dazu schätzen wir ein Logistisches Regressionsmodell, wobei die binäre Zielgröße hier R ist. Dieses Modell können wir anhand aller Beobachtungen schätzen, falls Y als Kovariable ausgeschlossen ist.

Die Ergebnisse der Schätzung sind in Tabelle B.1 angegeben. Sie zeigen, dass zum Beispiel die Einflussgröße Wohndauer keinen zum 5%-Niveau signifikant von Null verschiedenen Effekt auf die Akzeptanz hat und einen sehr großen p-Wert nahe eins aufweist. Dies spricht dafür, dass die Wohndauer keinen Einfluss auf die Entscheidung der Bank ausgeübt hat. Es ist möglich, dass die Wohndauer bei der zuvor erstellten Scorekarte keine Rolle gespielt und auch der Sachbearbeiter der Bank diese nicht in seine Entscheidung einbezogen hat. In der Praxis kann die Bank zur Klärung ihre internen Informationen einbeziehen. Da diese hier jedoch nicht vorliegen, gehen wir für das weitere Vorgehen davon aus, dass die Ablehnung eines Kunden unabhängig von seiner Wohndauer ist.

7.2 Ergebnisse

Mit Hilfe der neuen Methode lässt sich nun das Modell $w(y, \boldsymbol{x}, \boldsymbol{\theta})$ für das Fehlen der Zielgröße auch in Abhängigkeit der Zielgröße selbst schätzen. Tabelle B.2 präsentiert die Ergebnisse. Daneben enthält sie zum Vergleich auch die gewöhnliche Schätzung ohne Y als Kovariable. Es lässt sich feststellen, dass der zu Y gehörende Parameter positiv und signifikant von Null verschieden ist. Dies entspricht der Intuition, denn damit haben diejenigen Kunden eine größere Wahrscheinlichkeit Geld zu bekommen, die den Kredit tatsächlich zurückzahlen können. Einige der übrigen Parameterschätzungen unterscheiden sich deutlich von dem Modell ohne Bonität als unabhängige Variable – teilweise wechselt sogar das Vorzeichen. So zeigt sich, dass verwitwete Kunden eine niedrigere, ledige Kunden aber eine höhere Wahrscheinlichkeit haben, einen Kredit von der Bank zu bekommen. Auch haben Kunden mit vielen Kindern offenbar bessere Chancen auf einen Kredit. Ein Mitantragsteller beeinflusst die Entscheidung der Bank positiv. Teilweise ergeben sich auch verwunderliche Ergebnisse, zum Beispiel dass die Bank Kunden aus Wohngebieten mit hoher und mittlerer Kaufkraft mit größerer Wahrscheinlichkeit Geld leiht als solchen aus Gebieten mit sehr hoher Kaufkraft.

Von Interesse ist nun die Vorhersage der Bonität der Kunden. Die Schätzungen im Logistischen Regressionsmodell durch die ML-Methode und das neue SEL-Verfahren sind Tabelle B.3 zu entnehmen. Die Schätzungen unterscheiden sich durch die Berücksichtigung der MNAR-Situation teilweise deutlich. Für einige Parameter ändert sich auch hier das Vorzeichen der Schätzung, so hat beispielsweise nach dem neuen Modell die Arbeitsdauer einen positiven Effekt auf die Bonität, ein vorhandener Mitantragsteller einen negativen. Letzteres lässt sich so erklären, dass von Kunden mit hoher Ausfallwahrscheinlichkeit häufig eine Bürgschaft verlangt wird. Für die meisten Parameter ändert sich nur die Intensität des Effekts. So ist beispielsweise das Risiko für Selbständige nach dem neuen Modell deutlich erhöht im Vergleich zum gewöhnlichen Logit-Modell.

Als Maß zum Vergleich der Anpassungsgüte der beiden Modelle kann ein Pseudo-Bestimmtheitsmaß wie das häufig verwendete Pseudo-R^2 von McFadden (1973, S. 121) hinzugezogen werden. Dies ist definiert als

$$R^2_{\mathrm{McF}} = 1 - \frac{\log L(\widehat{\boldsymbol{\beta}})}{\log L(\widehat{\boldsymbol{\beta}}^{\star})},$$

wobei $L(\widehat{\boldsymbol{\beta}})$ die Likelihoodfunktion des vollen Modells, $L(\widehat{\boldsymbol{\beta}}^{\star})$ die Likelihoodfunktion des Nullmodells (ohne unabhängige Variablen außer der Konstante) ist. Für das gewöhnliche Logistische Regressionsmodell erhalten wir $R^2_{\text{McF}} = 0.103$, für das Modell aus Kapitel 5 ergibt sich $R^2_{\text{McF}} = 0.341$. Damit scheint die Anpassung des gewichteten Modells deutlich besser zu sein.

Nun soll mit Hilfe des Hausman-Tests aus Abschnitt 5.5 überprüft werden, ob das Vorgehen für die Bankdaten gerechtfertigt ist. Es ist also zu überprüfen, ob eine Missspezifikation in dem gewöhnlichen Logistischen Regressionsmodell vorliegt. Dabei tritt wieder das Problem auf, dass die Differenz der Kovarianzmatrizen von $\widehat{\boldsymbol{\beta}}^{\text{SEL}}$ und $\widehat{\boldsymbol{\beta}}^{\text{ML}}$ nicht positiv semi-definit ist. So ist die Realisation der Teststatistik $Q = -231.510$ dann negativ, $\widehat{\boldsymbol{q}}$ und $\widehat{\boldsymbol{\beta}}^{\text{ML}}$ sind offenbar nicht unabhängig (vgl. Gleichung (5.18)). Schreiber (2008) schlägt für so einen Fall unter anderem vor, als neue Teststatistik den Absolutwert $|Q| = 231.510$ zu wählen. Dies führt zur Ablehnung der Nullhypothese (kritischer Wert: $\chi^2_{41,0.95} = 56.942$, p-Wert: $p = 0.000$). Eine andere Möglichkeit ist nun, wie in den Simulation in Kapitel 6 die Varianz von $\widehat{\boldsymbol{q}}$ gesondert zu schätzen. Das in den Simulation verwendete Vorgehen per Bootstrap führt hier allerdings zu sehr starken Verzerrungen, was auf die Struktur der Daten zurückzuführen ist. Da es sich bei einigen Kovariablen um kategorielle Merkmale handelt, schwanken die Schätzungen der einzelnen Bootstrap-Stichproben sehr stark. Dieses Verhalten ist typisch für Modelle mit kategoriellen Kovariablen (vgl. auch Davison und Hinkley 1997, S. 333f). Außerdem kann das Problem auftauchen, dass Ausprägungen der Kovariablen in einigen Bootstrap-Stichproben nicht auftauchen und die zugehörigen Effekte dadurch nicht schätzbar sind. Aus diesen Gründen soll hier ein Jackknife-Schätzer verwendet werden, der nicht so anfällig gegenüber solchen Schwankungen ist, da für jede Schätzung jeweils nur eine einzelne Beobachtung weggelassen wird.

Einen Jackknife-Schätzer für die Varianz der Parameterschätzer in Generalisierten Linearen Modell leitet Shao (1992) oder auch Shao und Tu (1995, S. 3540f) als

$$\widehat{\text{Var}}_{\text{J}}(\widehat{\boldsymbol{\beta}}) = \frac{n-d}{n} \sum_{i=1}^{n} (\widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}})(\widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}})^{\top}$$

her, wobei $\widehat{\boldsymbol{\beta}}_{(i)}$ aus allen bis auf die i-te Beobachtung geschätzt wird und $\widehat{\boldsymbol{\beta}}$ aus der kompletten Stichprobe. Ein Schätzer für die Varianz von $\widehat{\boldsymbol{q}} := \widehat{\boldsymbol{\beta}}^{\text{SEL}} - \widehat{\boldsymbol{\beta}}^{\text{ML}}$ lässt sich

nun entsprechend durch

$$\widehat{\mathrm{Var}}_{\mathrm{J}}(\widehat{\boldsymbol{q}}) = \frac{n-d}{n} \sum_{i=1}^{n} (\widehat{\boldsymbol{q}}_{(i)} - \widehat{\boldsymbol{q}})(\widehat{\boldsymbol{q}}_{(i)} - \widehat{\boldsymbol{q}})^{\top}$$

bestimmen. Damit ergibt sich als Realisation der Teststatistik des Hausman-Tests $Q = 478.367$ (p-Wert: $p = 0.000$). Dieses Ergebnis bestätigt genau wie das obige Resultat die Vermutung, dass die Modellierung ohne Berücksichtigung der fehlenden Daten missspezifiziert ist.

7.3 Vergleich der Prognosegüte

Ein angemessener Vergleich der Prognosegüte des ML- und des SEL-Schätzers ist anhand des vorliegenden Datensatzes problematisch. Da die Realisationen, mit deren Hilfe die Güte der Prognose untersucht werden kann, aus der bedingten Verteilung von $Y, \boldsymbol{X}|R = 1$ stammen, liegt die Vermutung nahe, dass die Güte des neuen Modells anhand dieser Daten unterschätzt wird und für das herkömmliche Modell zu optimistisch ist. Ein adäquater Vergleich wäre dann möglich, wenn auch Beobachtungen aus der Verteilung $Y, \boldsymbol{X}|R = 0$ vorlägen. Da dies hier jedoch nicht der Fall ist, beschränken wir uns auf einen Vergleich auf Basis der vollständigen Beobachtungen.

Für den Vergleich werden zunächst Prognosen mittels Leave-one-out-Kreuzvalidierung bestimmt. Dazu wird eine Beobachtung bei der Parameterschätzung zurückgehalten und diese anschließend durch beide Modelle prognostiziert. Dies wiederholen wir für alle 4 000 Beobachtungen und erhalten Prognosen für die einzelnen Kunden. Die Güte lässt sich nun durch verschiedene Maße vergleichen. Für Kreditausfallprognosen werden dazu häufig ROC-Kurven hinzugezogen, welche die Falschpositiv-Rate gegen die Richtigpositiv-Rate für verschiedene Grenzwerte abtragen (vgl. Krämer und Bücker 2009). Die beiden resultierenden ROC-Kurven sind in Abbildung 7.3 dargestellt. Beide unterscheiden sich nicht maßgeblich, so dass schwer zu entscheiden ist, welches Modell die größere Vorhersagegüte besitzt.

Ein Vergleich der Flächen unter den ROC-Kurven (AUROC) als skalarwertiges Maß liefert eine fast identische Vorhersagegüte für beide Modelle (gewöhnliches Logistisches Regressionsmodell: AUROC = 0.661, neues Modell: AUROC = 0.662). Somit prognostizieren beide Modelle auf dem vorliegenden Datensatz mit gleicher Güte.

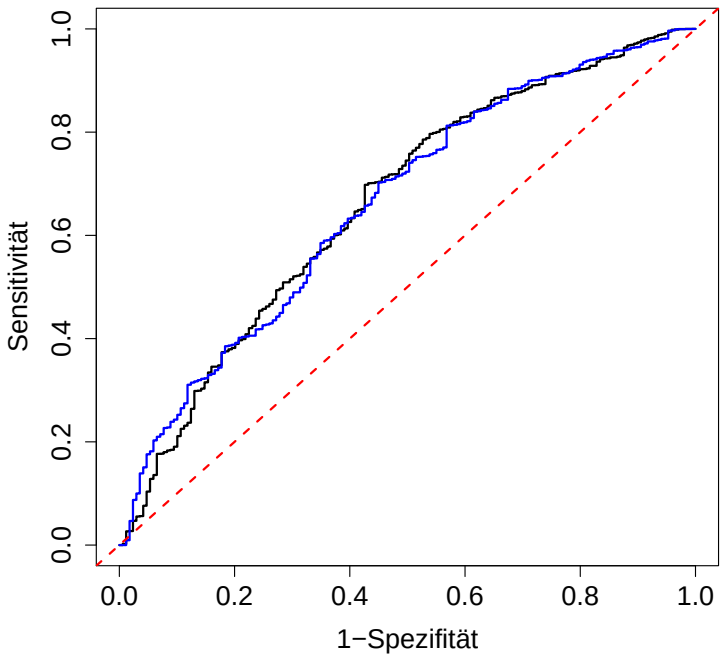

Abbildung 7.3 ROC-Kurven der Prognosen durch das herkömmliche Logistische Regressionsmodell (blau) und das neue Modell (schwarz).

Da die Parameterschätzungen des gewöhnlichen Logit-Modells jedoch offenbar verzerrt sind, liegt die Vermutung nahe, dass die Prognosen des neuen Modells für die unvollständigen Beobachtungen präziser sind.

Kapitel 8

Zusammenfassung

Fehlende Daten werden in empirischen Analysen häufig vernachlässigt. Bei der Modellierung von Kreditausfallwahrscheinlichkeiten beispielsweise werden die Daten der abgelehnten Kunden in der Regel ignoriert, da sich für sie die Zielgröße der Beobachtung entzieht. Ob dieses Vorgehen zu einer unverzerrten Modellanpassung führt, hängt von der Art des Fehlens ab. Die Unterscheidung zwischen nicht-ignorierbar und ignorierbar fehlenden Daten spielt hier die entscheidende Rolle. Falls das Fehlen der nur teilweise beobachteten Variable von den fehlenden Ausprägungen selbst abhängt, ist dies nicht-ignorierbar und eine Modellanpassung ohne Berücksichtigung dieser Tatsache führt zu Verzerrungen. Die vorliegende Arbeit beschäftigt sich mit der konsistenten Schätzung von statistischen Modellen mit nicht-ignorierbar fehlender Zielgröße.

Das neue Schätzverfahren verallgemeinert einen semi-empirischen Likelihood-Ansatz zur Schätzung des Erwartungswerts einer teilweise nicht-ignorierbar fehlenden Zufallsvariable mit Hilfe von vollständig beobachteten Kovariablen. Dazu schätzt man die gemeinsame empirische Verteilung der Zielgröße und der Kovariablen, die nicht auf die beobachteten Daten bedingt ist. Die resultierenden Werte der Wahrscheinlichkeitsfunktion liefern Gewichte zur Bestimmung eines gewichteten arithmetischen Mittels: dem unverzerrten Schätzer für den Erwartungswert. Eine naheliegende Vermutung ist, dass sich dieses Zwei-Schritt-Verfahren auch auf andere Schätzprobleme übertragen lässt.

Die Anpassung von statistischen Modellen (wie Generalisierten Linearen Modellen) geschieht in der Regel durch die Maximum-Likelihood-Methode. Die Parameter-

schätzer ergeben sich als Nullstellen der Scoregleichungen. Es zeigt sich, dass eine entsprechende Gewichtung der Scoregleichungen zu unverzerrten Parameterschätzern führt. Diese Schätzer sind außerdem asymptotisch normalverteilt. Eine Simulationsstudie verdeutlicht die Vorteile der gewichteten gegenüber der herkömmlichen Maximum-Likelihood-Schätzung: der mittlere quadratische Fehler ist vor allem in Situationen mit großem Anteil fehlender Daten wesentlich kleiner, sowohl in Linearen wie auch in Logistischen Regressionsmodellen.

Ein neuer Hausman-Test beantwortetet die Frage, ob das Fehlen der Zielgröße ignorierbar ist und damit ob eine unverzerrte Modellanpassung auch auf herkömmlichem Weg möglich ist. Der Spezifikationstest auf Ignorierbarkeit hält die vorgegebene Wahrscheinlichkeit für einen Fehler erster Art relativ gut ein und weist im Fall von nur wenigen fehlenden Daten eine hervorragende Güte auf. Selbst bei sehr hohem Anteil fehlender Daten ist die Güte noch annehmbar. Er eignet sich damit zur Aufdeckung nicht-ignorierbar fehlender Zielgrößen in statistischen Modellen. Die Schätzung der Varianz der Differenz der beiden Parameterschätzer stellt bei der Berechnung der Teststatistik eine Schwierigkeit dar. Man kann sich hier mir Resampling-Methoden behelfen, weitere Möglichkeiten könnten das Ziel zukünftiger Forschung sein.

Zu den möglichen Anwendungsgebieten der neuen Schätzmethode gehört die angesprochene Reject Inference. Viele andere Einsatzmöglichkeiten sind denkbar. Diese liegen in allen relevanten Bereichen der statistischen Modellbildung. So treten in medizinischen Untersuchungen oft fehlende Werte auf, falls ein Patient aus einer Studie ausscheidet. Auch in technischen Fragestellungen kann sich die Ausprägung der Zielgröße der Beobachtung entziehen, falls zum Beispiel ein Werkstück bei der Qualitätsprüfung zerstört wird. In vielen Situationen hängt das Fehlen dann von der unbeobachteten Ausprägung ab. Im Allgemeinen lässt sich ohne näheres Wissen über die Entstehung der fehlenden Beobachtungen nicht entscheiden, ob man diese ignorieren kann.

Zu den offenen Fragen für weitere Untersuchungen gehört die Vereinfachung der Methode von einer Zwei-Schritt-Schätzung zu einer direkten Schätzung der unbekannten Modellparameter. Dazu könnte die Semi-Empirische Likelihood weiter spezifiziert werden mit Hilfe der bedingten Verteilung der Zielgröße, gegeben die Kovariablen. Über entsprechende Nebenbedingungen ließen sich dann die Parameter – sofern identifizierbar – direkt schätzen. Eine weitere Frage ist, inwiefern sich die Be-

obachtungen der Kovariablen bei der Formulierung der Likelihood vollständig nutzen lassen. Die Schwierigkeit besteht dabei in der Handhabung der resultierenden Likelihood. Dennoch könnte dies zu wesentlichen Verbesserung der Schätzung führen, da die vollständige Information der Kovariablen und nicht nur deren Erwartungswert einfließen kann. Schließlich könnten weitere Untersuchungen die Übertragung der Adjusted Empirical Likelihood für die Maximierung der Likelihood klären. Falls die asymptotischen Eigenschaften der Schätzer erhalten bleiben, könnte somit die Konvergenz des Newton-Raphson-Algorithmus auch für kleine Stichproben und schlechte Startwerte gewährleistet werden.

Literaturverzeichnis

Baker, S. G. und Laird, N. M. (1988). Regression analysis for categorical variables with outcome subject to nonignorable nonresponse. *Journal of the American Statistical Association*, **83** (401), 62– 69.

Banasik, J. und Crook, J. (2007). Reject inference, augmentation, and sample selection. *European Journal of Operational Research*, **183**, 1582–1594.

Boyes, W. J., Hoffman, D. L. und Low, S. A. (1989). An econometric analysis of the bank credit scoring problem. *Journal of Econometrics*, **40** (1), 3–14.

Chapelle, O., Schölkopf, B. und Zien, A. (Hrsg.) (2006). *Semi-Supervised Learning*. MIT Press.

Chen, J., Variyath, M. A. und Abraham, B. (2008). Adjusted empirical likelihood and its properties. *Journal of Computational and Graphical Statistics*, **17** (2), 426–443.

Chen, S. X. (1997). Empirical likelihood-based kernel density estimation. *Australian & New Zealand Journal of Statistics*, **39** (1), 47–56.

Chen, S. X. und Qin, J. (2006). An empirical likelihood method in mixture models with incomplete classifications. *Statistica Sinica*, **16**, 1101–1115.

Copas, J. B. und Li, H. G. (2002). Inference for non-random samples (with discussion and rejoinder). *Journal of the Royal Statistical Society B*, **59** (1), 55–95.

Crook, J. und Banasik, J. (2004). Does reject inference really improve the performance of application scoring models? *Journal of Banking & Finance*, **28**, 857–874.

Daniels, M. J. und Hogan, J. W. (2008). *Missing Data In Longitudinal Studies*. Chapman & Hall/CRC.

Davison, A. C. und Hinkley, D. V. (1997). *Bootstrap Methods and their Application*. Cambridge University Press.

Fay, R. E. (1986). Causal models for patterns of nonresponse. *Journal of the American Statistical Association*, **81** (394), 354–365.

Feelders, A. J. (1999). Credit scoring and reject inference with mixture models. *International Journal of Intelligent Systems in Accounting, Finance & Management*, **8**, 271–279.

Feelders, A. J. (2000). Credit scoring and reject inference with mixture models. *International Journal of Intelligent Systems in Accounting, Finance & Management*, **9**, 1–8.

Feelders, A. J., Chang, S. und McLachlan, G. J. (1998). Mining in the presence of selectivity bias and its application to reject inference. In *KDD*, 199–203.

Gill, J. (2001). *Generalized linear models: a unified approach*. Sage Publications, Thousand Oaks.

Godambe, V. P. (1991). *Estimating Functions*. Oxford University Press.

Greene, W. H. (1998). Sample selection in credit-scoring models. *Japan and World Economy*, **10**, 299–316.

Hand, D. J. (2005). Good practice in retail credit scorecard assessment. *Journal of the Operational Research Society*, **56** (9), 1109–1117.

Hand, D. J. und Henley, W. E. (1993). Can reject inference ever work? *IMA Journal of Mathematics Applied in Business & Industry*, **5**, 45–55.

Hand, D. J. und Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: a review. *Journal of the Royal Statistical Society A*, **160** (3), 523–541.

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica*, **46** (6), 1251–1271.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, **47** (1), 153–161.

Henking, A., Bluhm, C. und Fahrmeir, L. (2006). *Kreditrisikomessung*. Springer, Berlin.

Ibrahim, J. G. und Lipsitz, S. R. (1996). Parameter estimation from incomplete data in binomial regression when the missing data mechanism is nonignorable. *Biometrics*, **52** (3), 1071–1078.

Ibrahim, J. G., Chen, M. H., Lipsitz, S. R. und Herring, A. H. (2005). Missing-data methods for generalized linear models: A comparative review. *Journal of the American Statistical Association*, **100**, 332–346.

Jacobson, T. und Roszbach, K. (2003). Bank lending policy, credit scoring and value-at-risk. *Journal of Banking & Finance*, **27** (4), 615–633.

Joanes, D. N. (1993). Reject inference applied to logistic regression for credit scoring. *IMA Journal of Mathematics Applied in Business and Industry*, **5** (1), 35–43.

Kim, Y. und Sohn, S.-Y. (2007). Technology scoring model considering rejected applicants and effect of reject inference. *Journal of the Operational Research Society*, **58**, 1341–1347.

Krämer, W. und Bücker, M. (2009). Statistischer Qualitätsvergleich von Kreditausfallprognosen. Technical Report 30/09, SFB 823, Technische Universität Dortmund.

Krämer, W. und Sonnberger, H. (1986). *The Linear Regression Model under Test*. Physica-Verlag, Heidelberg Wien.

Little, R. J. A. (1985). A note about models for selectivity bias. *Econometrica*, **53** (6), 1469–1474.

Little, R. J. A. und Rubin, D. B. (1987). *Statistical analysis with missing data*. Wiley series in probability and mathematical statistics : Applied probability and statistics. Wiley, New York.

Liu, Y. und Chen, J. (2010). Adjusted empirical likelihood with high-order precision. *The Annals of Statistics*, **38** (3), 1341–1362.

McCullagh, P. und Nelder, J. A. (1989). *Generalized linear models*. Chapman and Hall, London, 2. Auflage.

McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Hrsg.), *Frontiers in Econometrics*, 105–142. Academic Press, New York.

McLachlan, G. J. (1975). Iterative reclassification procedure for constructing an asymptotically optimal rule of allocation in discriminant analysis. *Journal of the American Statistical Association*, **70** (350), 365–369.

Meng, C.-L. und Schmidt, P. (1985). On the cost of partial observability in the bivariate probit model. *International Economic Review*, **26** (1), 71–85.

Mittelhammer, R. C., Judge, G. G. und Miller, D. J. (2000). *Econometric foundations*. Cambridge University Press.

Molenberghs, G., Goetghebeur, E. J. T., Lipsitz, S. R. und Kenward, M. G. (1999). Nonrandom missingness in categorical data: Strengths and limitations. *The American Statistician*, **53** (2), 110–118.

Newey, W. K. und McFadden, D. (1994). *Large sample estimation and hypothesis testing*, Band IV, Kap. 36, 2111–2245. Elsevier Science.

Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, **75**, 237–249.

Owen, A. B. (1990). Empirical likelihood ratio confidence regions. *The Annals of Statistics*, **18** (1), 90–120.

Owen, A. B. (2001). *Empirical Likelihood*. Chapman & Hall/CRC, Boca Raton.

Qin, J. (1992). *Empirical likelihood and semiparametric models*. Dissertation, University of Waterloo, Ontario, Canada.

Qin, J. (1994). Semi-empirical likelihood ratio confidence intervals for the difference of two sample means. *Annals of the Institute of Statistical Mathematics*, **46** (1), 117–126.

Qin, J. (2000). Combining parametric and empirical likelihoods. *Biometrika*, **87** (2), 484–490.

Qin, J. und Lawless, J. (1994). Empirical likelihood and general estimating equations. *The Annals of Statistics*, **22** (1), 300–325.

Qin, J. und Wong, A. (1996). Empirical likelihood in a semi-parametric model. *Scandinavian Journal of Statistics*, **23** (2), 209–219.

Qin, J. und Zhang, B. (2007). Empirical-likelihood-based inference in missing response problems and its application in observational studies. *Journal of the Royal Statistical Society B*, **69** (1), 101–122.

Qin, J., Leung, D. und Shao, J. (2002). Estimation with survey data under nonignorable nonresponse or informative sampling. *Journal of the American Statistical Association*, **97** (457), 193–200.

Qin, J., Zhang, B. und Leung, D. H. Y. (2009). Empirical likelihood in missing data problems. *Journal of the American Statistical Association*, **104** (488), 1492–1503.

Rosenberg, E. und Gleit, A. (1994). Quantitative Methods in Credit Management: A Survey. *Operations Research*, **42** (4), 589–613.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, **36** (3), 581–592.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 473–489.

Schafer, J. L. (2000). *Analysis of incomplete multivariate data*. Monographs on statistics and applied probability. Chapman & Hall/CRC, Boca Raton.

Scheid, S. (2005). *Selection Models for Nonignorable Missing Data*, Band 8 von *Anwendungsorientierte Statistik*. Peter Lang, Frankfurt a. M.

Schreiber, S. (2008). The Hausman test statistic can be negative even asymptotically. *Journal of Economics and Statistics (Jahrbücher für Nationalökonomie und Statistik)*, **228** (4), 394–405.

Shao, J. (1992). Jackknifing in generalized linear models. *Annals of the Institute of Statistical Mathematics*, **44** (4), 673–686.

Shao, J. und Tu, D. (1995). *The Jackknife and Bootstrap*. Springer, New-York.

Smith, A. T. und Elkan, C. (2004). A bayesian network framework for reject inference. In *KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 286–295, New York, NY, USA. ACM.

Smith, A. T. und Elkan, C. (2007). Making generative classifiers robust to selection bias. In *KDD '07: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, 657–666, New York, NY, USA. ACM.

Tanner, M. A. und Wong, W. H. (1987). The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association*, **82**, 528–550.

Thomas, L. C. (2000). A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. *International Journal of Forecasting*, **16** (2), 149–172.

Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data*. Springer, Berlin.

Vaart, van der, A. W. (1998). *Asymptotic statistics*. Cambridge University Press, Cambridge.

Verstraeten, G. und Poel, D. V. d. (2005). The impact of sample bias on consumer credit scoring performance and profitability. *The Journal of the Operational Research Society*, **56** (8), 981–992.

Weichert, T. (2006). Verbraucher-Scoring meets Datenschutz. *Datenschutz und Datensicherheit – DuD*, **30**, 399–404.

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, **9**, 60–62.

Anhang A

Der Newton-Raphson-Algorithmus

In Kapitel 5 ist zur Schätzung der Parameter die semi-empirische Likelihood zu maximieren. Dies soll durch das Newton-Verfahren realisiert werden. Die Log-Likelihood hat zunächst die Form

$$\begin{aligned}\ln L_n(\boldsymbol{\theta}, W, \boldsymbol{\lambda}_1, \lambda_2) = &\sum_{i=1}^{n} \ln w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) + (N-n)\ln(1-W) - n\ln n \\ &- \sum_{i=1}^{n} \ln z_i,\end{aligned}$$

wobei

$$z_i = \left[1 + \boldsymbol{\lambda}_1^\top (\boldsymbol{x}_i - \boldsymbol{\mu}) + \lambda_2 \left(w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) - W\right)\right] \qquad (i = 1, \ldots, n).$$

Wie in Kapitel 5 bereits gesehen, ist nach partiellem Ableiten nach dem unbekannten Parametervektor $\boldsymbol{\psi}^\star = (\boldsymbol{\lambda}_1^\top, \lambda_2, \boldsymbol{\theta}^\top, W)$ und Nullsetzen dieser Ableitungen das Gleichungssystem

$$\boldsymbol{f}^\star(\boldsymbol{\psi}^\star) = \begin{pmatrix} \mathbb{L}^{\boldsymbol{\lambda}_1} \\ \mathbb{L}^{\lambda_2} \\ \mathbb{L}^{\boldsymbol{\theta}} \\ \mathbb{L}^{W} \end{pmatrix} = \mathbf{0}$$

zu lösen, mit

$$\mathbb{L}^{\boldsymbol{\lambda}_1} := -\sum_{i=1}^{n} \frac{\boldsymbol{x}_i - \boldsymbol{\mu}}{z_i},$$

$$\mathbb{L}^{\lambda_2} := -\sum_{i=1}^{n} \frac{w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) - W}{z_i},$$
$$\mathbb{L}^{\boldsymbol{\theta}} := \sum_{i=1}^{n} \left[\frac{\partial \ln w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} - \frac{\lambda_2 \partial w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) / \partial \boldsymbol{\theta}}{z_i} \right]$$

und

$$\mathbb{L}^{W} := \frac{N/n - 1}{1 - W} - \lambda_2.$$

Dabei ist $\mathbb{L}^W = 0$ äquivalent zu $\lambda_2 = \frac{N/n-1}{1-W}$. Daher ist es nicht nötig, λ_2 explizit zu schätzen, da dieser Parameter durch W eindeutig festgelegt ist. Damit reduziert sich der Parametervektor zu $\boldsymbol{\psi} = (\boldsymbol{\lambda}_1^\top, \boldsymbol{\theta}^\top, W)$ und das zu lösende Gleichungssystem ist

$$\boldsymbol{f}(\boldsymbol{\psi}) = \begin{pmatrix} \mathbb{L}^{\boldsymbol{\lambda}_1} \\ \mathbb{L}^{\lambda_2} \\ \mathbb{L}^{\boldsymbol{\theta}} \end{pmatrix} = \mathbf{0}.$$

Die Jacobi-Matrix von $\boldsymbol{f}$ lautet

$$\boldsymbol{J}(\boldsymbol{\psi}) = \begin{pmatrix} \mathbb{L}^{\boldsymbol{\lambda}_1 \boldsymbol{\lambda}_1} & \mathbb{L}^{\boldsymbol{\lambda}_1 \boldsymbol{\theta}} & \mathbb{L}^{\boldsymbol{\lambda}_1 W} \\ \mathbb{L}^{\lambda_2 \boldsymbol{\lambda}_1} & \mathbb{L}^{\lambda_2 \boldsymbol{\theta}} & \mathbb{L}^{\lambda_2 W} \\ \mathbb{L}^{\boldsymbol{\theta} \boldsymbol{\lambda}_1} & \mathbb{L}^{\boldsymbol{\theta}\boldsymbol{\theta}} & \mathbb{L}^{\boldsymbol{\theta} W} \end{pmatrix}$$

mit

$$\mathbb{L}^{\boldsymbol{\lambda}_1 \boldsymbol{\lambda}_1} := \frac{\partial \mathbb{L}^{\boldsymbol{\lambda}_1}}{\partial \boldsymbol{\lambda}_1^\top} = \sum_{i=1}^{n} \frac{(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^\top}{z_i^2},$$
$$\mathbb{L}^{\lambda_2 \boldsymbol{\lambda}_1} := \frac{\partial \mathbb{L}^{\boldsymbol{\lambda}_2}}{\partial \lambda_1} = \sum_{i=1}^{n} \frac{(\boldsymbol{x}_i - \boldsymbol{\mu})^\top (w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) - W)}{z_i^2},$$
$$\mathbb{L}^{\boldsymbol{\lambda}_1 \boldsymbol{\theta}} := \frac{\partial \mathbb{L}^{\boldsymbol{\lambda}_1}}{\partial \boldsymbol{\theta}^\top} = \sum_{i=1}^{n} \frac{\lambda_2 (\boldsymbol{x}_i - \boldsymbol{\mu}) \cdot \partial w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) / \partial \boldsymbol{\theta}^\top}{z_i^2} = \left(\mathbb{L}^{\boldsymbol{\theta} \boldsymbol{\lambda}_1} \right)^\top,$$
$$\mathbb{L}^{\boldsymbol{\lambda}_1 W} := \frac{\partial \mathbb{L}^{\boldsymbol{\lambda}_1}}{\partial W} = -\sum_{i=1}^{n} \frac{\lambda_2 (\boldsymbol{x}_i - \boldsymbol{\mu})}{z_i^2},$$
$$\mathbb{L}^{\lambda_2 \boldsymbol{\theta}} := \frac{\partial \mathbb{L}^{\lambda_2}}{\partial \boldsymbol{\theta}^\top} = -\sum_{i=1}^{n} \frac{\frac{\partial w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}}{z_i} + \sum_{i=1}^{n} \frac{\lambda_2 \frac{\partial w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} (w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) - W)}{z_i^2},$$
$$\mathbb{L}^{\lambda_2 W} := \frac{\partial \mathbb{L}^{\lambda_2}}{\partial W} = \sum_{i=1}^{n} \frac{1}{z_i} + \sum_{i=1}^{n} \frac{\lambda_2 (w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) - W)}{z_i^2},$$

$$\begin{aligned}\mathbb{L}^{\boldsymbol{\theta}\boldsymbol{\theta}} := \frac{\partial \mathbb{L}^{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}^\top} = &\sum_{i=1}^{n} \frac{\partial^2 \ln w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \\ &- \sum_{i=1}^{n} \frac{\frac{\lambda_2 \partial^2 w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}}{z_i} \\ &+ \sum_{i=1}^{n} \frac{\left[\frac{\lambda_2 \partial w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right] \left[\frac{\lambda_2 \partial w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top}\right]}{z_i^2}\end{aligned}$$

und

$$\mathbb{L}^{\boldsymbol{\theta}W} := \frac{\partial \mathbb{L}^{\boldsymbol{\theta}}}{\partial W} = -\sum_{i=1}^{n} \frac{\lambda_2^2 \cdot \partial w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) / \partial \boldsymbol{\theta}}{z_i^2}.$$

Zur numerischen Bestimmung der Nullstelle von $\boldsymbol{f}$ beziehungsweise des Maximums der Log-Likelihood ist der k-te Schritt des Newton Verfahrens durch

$$\boldsymbol{\psi}^{(k+1)} = \boldsymbol{\psi}^{(k)} - \boldsymbol{J}(\boldsymbol{\psi}^{(k)})^{-1} \boldsymbol{f}(\boldsymbol{\psi}^{(k)})$$

gegeben.

Wählt man als parametrische Funktion w ein logistisches Regressionsmodell, so ergibt sich mit

$$\boldsymbol{d}_i := \begin{pmatrix} 1 \\ \boldsymbol{x}_i \\ y_i \end{pmatrix} \tag{A.1}$$

das Modell zu

$$w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) = \frac{\exp\left(\boldsymbol{d}_i^\top \boldsymbol{\theta}\right)}{1 + \exp\left(\boldsymbol{d}_i^\top \boldsymbol{\theta}\right)}. \tag{A.2}$$

Zur Bestimmung der Funktion $\boldsymbol{f}$ und der Jacobi-Matrix $\boldsymbol{J}$ benötigen wir die ersten und zweiten Ableitungen von w und $\ln w$ bezüglich $\boldsymbol{\theta}$. Mit (A.1) und (A.2) gilt

$$\begin{aligned}\frac{\partial w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} &= \boldsymbol{d}_i w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) \left[1 - w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta})\right], \\ \frac{\partial^2 w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} &= \boldsymbol{d}_i w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) \left[1 - w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta})\right]^2 \boldsymbol{d}_i^\top \\ &\quad - \boldsymbol{d}_i \left[w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta})\right]^2 \left[1 - w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta})\right] \boldsymbol{d}_i^\top, \\ \frac{\partial \ln w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} &= \boldsymbol{d}_i \left[1 - w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta})\right]\end{aligned}$$

und

$$\frac{\partial^2 \ln w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} = \boldsymbol{d}_i \left[-w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) \left[1 - w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) \right] \right] \boldsymbol{d}_i^\top.$$

Die Summen dieser Ausdrücke lassen sich vereinfachend in Matrixschreibweise formulieren. Seien $\boldsymbol{D}$ die $((p+1) \times n)$-Matrix

$$\boldsymbol{D} = \begin{pmatrix} 1 & 1 & \dots & 1 \\ \boldsymbol{x}_1 & \boldsymbol{x}_2 & \dots & \boldsymbol{x}_n \\ y_1 & y_2 & \dots & y_n \end{pmatrix},$$

$\boldsymbol{w}$ der Vektor

$$\boldsymbol{w} = \begin{pmatrix} w(y_1, \boldsymbol{x}_1, \boldsymbol{\theta}) \\ w(y_2, \boldsymbol{x}_2, \boldsymbol{\theta}) \\ \vdots \\ w(y_n, \boldsymbol{x}_n, \boldsymbol{\theta}) \end{pmatrix}$$

und $\boldsymbol{W}$ die $(n \times n)$-Diagonalmatrix

$$\boldsymbol{W} = \begin{pmatrix} w(y_1, \boldsymbol{x}_1, \boldsymbol{\theta}) & 0 & \dots & 0 \\ 0 & w(y_2, \boldsymbol{x}_2, \boldsymbol{\theta}) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & w(y_n, \boldsymbol{x}_n, \boldsymbol{\theta}) \end{pmatrix}.$$

Dann gilt

$$\begin{aligned} \sum_{i=1}^{n} \frac{\partial w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} &= \boldsymbol{D} \left[\boldsymbol{W} (\mathbb{I} - \boldsymbol{W}) \right], \\ \sum_{i=1}^{n} \frac{\partial^2 w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} &= \boldsymbol{D} \left[\boldsymbol{W} (\mathbb{I} - \boldsymbol{W})^2 - \boldsymbol{W}^2 (\mathbb{I} - \boldsymbol{W}) \right] \boldsymbol{D}^\top, \\ \sum_{i=1}^{n} \frac{\partial \ln w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} &= \boldsymbol{D} \left[\mathbb{I} - \boldsymbol{W} \right] \end{aligned}$$

und

$$\sum_{i=1}^{n} \frac{\partial^2 \ln w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} = \boldsymbol{D} \left[-\boldsymbol{W} (\mathbb{I} - \boldsymbol{W}) \right] \boldsymbol{D}^\top.$$

Damit lässt sich dann zum Beispiel $\mathbb{L}^{\boldsymbol{\theta}\boldsymbol{\theta}}$ leicht berechnen durch

$$\mathbb{L}^{\boldsymbol{\theta}\boldsymbol{\theta}} = \boldsymbol{D} \left[-\boldsymbol{W} (\mathbb{I} - \boldsymbol{W}) \right] \boldsymbol{D}^\top$$

$$
\begin{aligned}
&- \boldsymbol{D}\lambda_2^\top \boldsymbol{Z}^{-1}\left[\boldsymbol{W}(\mathbb{I}-\boldsymbol{W})^2 - \boldsymbol{W}^2(\mathbb{I}-\boldsymbol{W})\right]\boldsymbol{D}^\top \\
&+ \boldsymbol{D}\lambda_2^{2^\top} \boldsymbol{Z}^{-2}\left[\boldsymbol{W}(\mathbb{I}-\boldsymbol{W})\right],
\end{aligned}
$$

wobei

$$
\boldsymbol{z} = \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{pmatrix}
$$

und

$$
\boldsymbol{Z} = \begin{pmatrix} z_1 & 0 & \dots & 0 \\ 0 & z_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & z_n \end{pmatrix}.
$$

Für das Newton-Verfahren spielt die Wahl des Startparameters eine entscheidende Rolle. Dieser sollte möglichst in der Nähe des wahren Wertes liegen, um die Konvergenz des Algorithmus' zur Nullstelle zu garantieren. Für die Parameter W und $\boldsymbol{\lambda}_1$ lassen sich aufgrund ihrer Eigenschaften sinnvolle Startwerte finden. Da $W = P(R = 1)$ der Anteil fehlender Realisationen von Y ist, ist $W^{(0)} = n/N$ eine gute Schätzung und damit ein guter Startwert für W. Da außerdem $\boldsymbol{\lambda}_1 \xrightarrow{p} \mathbf{0}$, ist $\boldsymbol{\lambda}_1^{(0)} = \mathbf{0}$ ein sinnvoller Startwert für den Lagrangeparameter. Für $\boldsymbol{\theta}$ ist es schwieriger gute Schätzungen zu finden, die in der Nähe des wahren Parameters liegen. Falls ein Zusammenhang zwischen Y und $\boldsymbol{X}$ bekannt ist, kann dieser ausgenutzt werden. Falls $y_i \approx m(\boldsymbol{x}_i, \boldsymbol{\eta})$ $(i = 1, \dots, n)$, können mit Hilfe dieses Modells die fehlenden Beobachtungen der Zufallsvariable Y geschätzt und anschließend eine Schätzung des Parametervektors $\boldsymbol{\theta}$ mit Hilfe der neuen Beobachtungen durchgeführt werden. Diese liefern dann als Startwert $\boldsymbol{\theta}^{(0)}$. Dabei hängt in diesem Fall die Güte des Startwerts für $\boldsymbol{\theta}$ im Wesentlichen von der Güte des Modells m ab. Dieses kann zum Beispiel ein auf den beobachteten Daten geschätztes Regressionsmodell sein. Je stärker hier die Verzerrung durch die fehlenden Daten ist, desto geringer wird die Eignung der Startwerte sein.

Ein wichtiger Punkt bei der Schätzung durch Empirische Likelihoods ist, dass stets die Nebenbedingung $p_i > 0$ erfüllt sein muss. Dies sollte in jedem Schritt überprüft werden, andernfalls konvergiert der Algorithmus in der Regel nicht. Falls die Nebenbedingung verletzt wird, ist die Schrittlänge des Algorithmus' zu verringern. Abschließend gibt der folgende Pseudo-Code eine Möglichkeit zur Schätzung der

Parameter mit Hilfe des Newton-Raphson-Algorithmus an (vgl. Chen, Variyath und Abraham 2008):

1. Wähle als Startwerte $\boldsymbol{\lambda}_1^{(0)} = \mathbf{0}$, $W^{(0)} = n/N$ und $\boldsymbol{\theta}^{(0)}$ (wie oben beschrieben). Setze $\gamma = 1$, $\boldsymbol{\psi}^{(0)} = (\boldsymbol{\lambda}_1^{(0)\top}, \boldsymbol{\theta}^{(0)\top}, W^{(0)})$. Wähle außerdem als Iterationsschritt $k = 0$ und das Toleranzlevel $\varepsilon = 10^{-8}$.

2. Bestimme
$$\boldsymbol{\nabla} = -\boldsymbol{J}(\boldsymbol{\psi}^{(k)})^{-1}\boldsymbol{f}(\boldsymbol{\psi}^{(k)}).$$
Falls $\|\boldsymbol{\nabla}\| < \varepsilon$ beende die Iteration und gebe $\boldsymbol{\lambda}_1^{(k)}$, $W^{(k)}$, $\boldsymbol{\theta}^{(k)}$ und $\lambda_2^{(k)}$ als Schätzer aus. Sonst fahre mit Schritt 3 fort.

3. Berechne $\boldsymbol{\delta} = \gamma\boldsymbol{\nabla}$. Falls für
$$z_i(\boldsymbol{\psi}^{(k)}) = \left[1 + \boldsymbol{\lambda}_1^{(0)\top}(\boldsymbol{x}_i - \boldsymbol{\mu}_{\boldsymbol{X}}) + \frac{N/n - 1}{1 - W^{(k)}}\left(w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}^{(k)}) - W^{(k)}\right)\right]$$
die Ungleichung
$$z_i(\boldsymbol{\psi}^{(k)} + \boldsymbol{\delta}) \leq 0 \qquad (i = 1, \ldots, n)$$
gilt oder für
$$\ln L_n(\boldsymbol{\psi}^{(k)}) = \sum_{i=1}^{n} \ln w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}^{(k)}) + (N - n)\ln(1 - W^{(k)}) - \sum_{i=1}^{n} \ln z_i(\boldsymbol{\psi}^{(k)})$$
die Ungleichung
$$\ln L_n(\boldsymbol{\psi}^{(k)} + \boldsymbol{\delta}) < \ln L_n(\boldsymbol{\psi}^{(k)})$$
gilt, setze $\gamma = \gamma/2$ und wiederhole den Schritt. Sonst fahre mit dem nächsten Schritt fort.

4. Setze
$$\boldsymbol{\psi}^{(k+1)} = \boldsymbol{\psi}^{(k)} + \boldsymbol{\delta}$$
und
$$\gamma = (k + 1)^{-1/2}.$$
Erhöhe k um 1 und gehe zurück zu Schritt 2.

Ein noch zu untersuchender Punkt ist, inwiefern sich die Adjusted Empirical Likelihood (vgl. Abschnitt 4.5, S. 33) auf eine Semi-Empirische Likelihood übertragen lässt. So kann es beispielsweise dazu kommen, dass bei großem Anteil fehlender Daten der Nullvektor nicht in der konvexen Hülle der Punkte $\{\boldsymbol{x}_i - \boldsymbol{\mu}_{\boldsymbol{X}}, i = 1, \ldots, n\}$

liegt und damit keine Nullstelle der Gleichung (5.3d) existiert. Abbildung A.1 zeigt ein Beispiel für 1 000 Ausprägungen von zwei Kovariablen, wobei die konvexe Hülle der vollständigen Beobachtungen den unbedingten Erwartungswert nicht überdeckt. In diesem Fall fehlt ein sehr großer Anteil der Beobachtungen der Zielgröße, so dass selbst bei diesem relativ großen Stichprobenumfang keine Lösung existiert. Genauso kann es passieren, dass schlechte Startwerte der Parameter $\boldsymbol{\theta}$ und W dazu führen, dass die konvexe Hülle der Punkte $\{w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) - W, i = 1, \ldots, n\}$ nicht die Null enthält. Ebenso kann der Parametervektor $\boldsymbol{\theta}$ in einem zu großen Iterationsschritt so geschätzt werden, dass selbiges Problem auftritt. Dann wäre die Gleichung (5.3c) nicht lösbar.

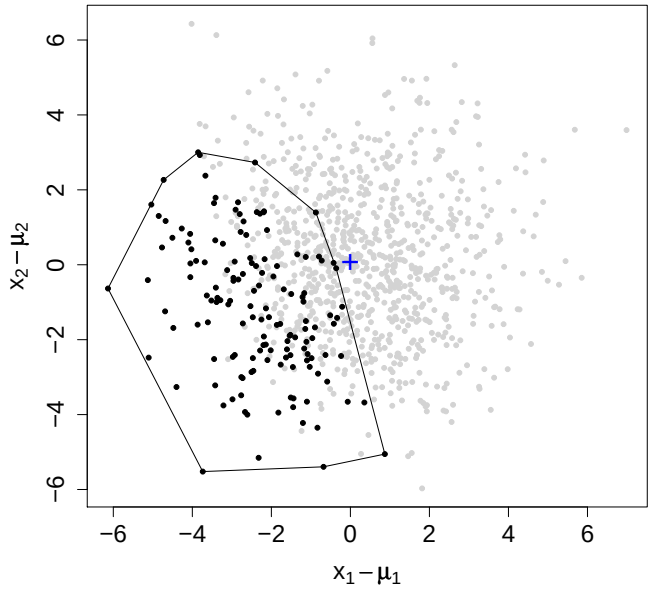

Abbildung A.1 Die Konvexe Hülle der vollständig beobachtbare Daten (schwarz) überdeckt den Nullpunkt (✚) nicht. Die Beobachtungen mit fehlender Zielgröße sind grau dargestellt.

Das Vorgehen der AEL lässt sich nicht ohne Weiteres übertragen, da die Likelihood (5.5) hier eine semi-empirische ist und die unbekannten Parameter sowie die Beobachtungen von $\boldsymbol{X}$ und Y enthält. Um diesem Problem zu begegnen könnte eine zusätzliche künstliche Beobachtung $(y_0, \boldsymbol{x}_0)$ herangezogen werden, so dass die Nebenbedingungen erfüllt sind. Wie eine solche Beobachtung gefunden werden kann, soll hier kurz erläutert werden.

Zunächst wähle wie in Gleichung (4.11) (S. 33)

$$\boldsymbol{x}_0 = -\frac{a_n}{n}\sum_{i=1}^{n}\left[\boldsymbol{x}_i - \boldsymbol{\mu}_{\boldsymbol{X}}\right] + \boldsymbol{\mu}_{\boldsymbol{X}} = (a_n + 1)\boldsymbol{\mu}_{\boldsymbol{X}} - \frac{a_n}{n}\sum_{i=1}^{n}\boldsymbol{x}_i.$$

Die zusätzliche Beobachtung von Y kann mittels der Umkehrfunktion von w bestimmt werden. Ist w zum Beispiel durch ein Logistisches Modell

$$w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) = \frac{\exp\left(\theta_1 + \theta_2 y_i + \boldsymbol{\theta}_3^\top \boldsymbol{x}_i\right)}{1 + \exp\left(\theta_1 + \theta_2 y_i + \boldsymbol{\theta}_3^\top \boldsymbol{x}_i\right)}$$

gegeben lässt sich die neue Beobachtung durch

$$\begin{aligned}
w(y_0, \boldsymbol{x}_0, \boldsymbol{\theta}) - W =& -\frac{a_n}{n}\sum_{i=1}^{n}\left[w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) - W\right] \\
\Leftrightarrow \theta_1 + \theta_2 y_0 + \boldsymbol{\theta}_3^\top \boldsymbol{x}_0 =& \log\left(\frac{-\frac{a_n}{n}\sum_{i=1}^{n}\left[w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) - W\right] + W}{1 + \frac{a_n}{n}\sum_{i=1}^{n}\left[w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) - W\right] - W}\right) \\
\Leftrightarrow y_0 =& \frac{1}{\theta_2}\left[\log\left(\frac{-\frac{a_n}{n}\sum_{i=1}^{n}\left[w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) - W\right] + W}{1 + \frac{a_n}{n}\sum_{i=1}^{n}\left[w(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}) - W\right] - W}\right) - \theta_1 - \boldsymbol{\theta}_3^\top \boldsymbol{x}_0\right]
\end{aligned}$$

herleiten.

Nun kann die Likelihood für die Beobachtungen $i = 0, 1, \ldots, N$ mit Hilfe des Newton-Raphson Algorithmus maximiert werden, wobei stets eine Lösung existiert. Es ist zu erwarten, dass sich die Asymptotik der Schätzer wie bei Chen, Variyath und Abraham (2008) nicht verändert, falls a_n geeignet gewählt wird, da nur eine einzige Beobachtungen hinzugefügt wird und diese für größer werdenden Stichprobenumfang an Bedeutung verliert. Dies ist jedoch noch zu untersuchen und könnte das Ziel zukünftiger Forschung sein.

Anhang B

Tabellen und Abbildungen

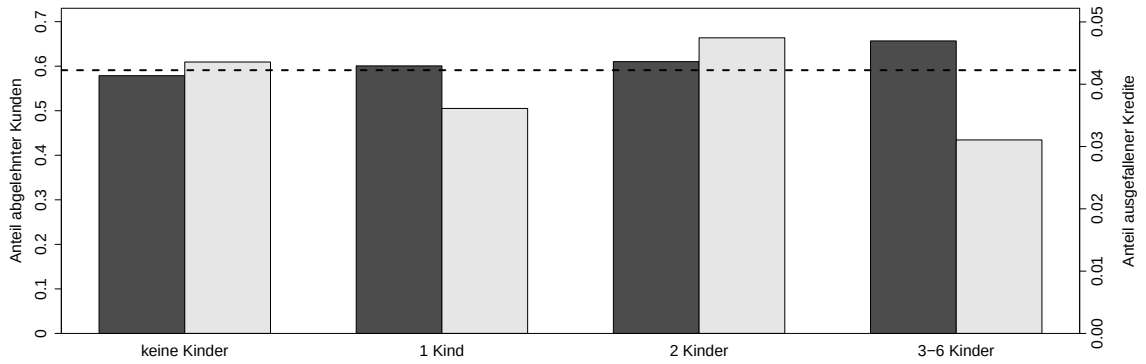

Abbildung B.1 Einfluss der Variable „Kinder" auf Kreditvergabe und Rückzahlung: die dunklen Balken geben den Anteil abgelehnter Kredite, die hellen Balken den Anteil ausgefallener Kredite in der jeweiligen Gruppe an, die Linie stellt den durchschnittlichen Anteil abgelehnter/ausgefallener Kredite dar.

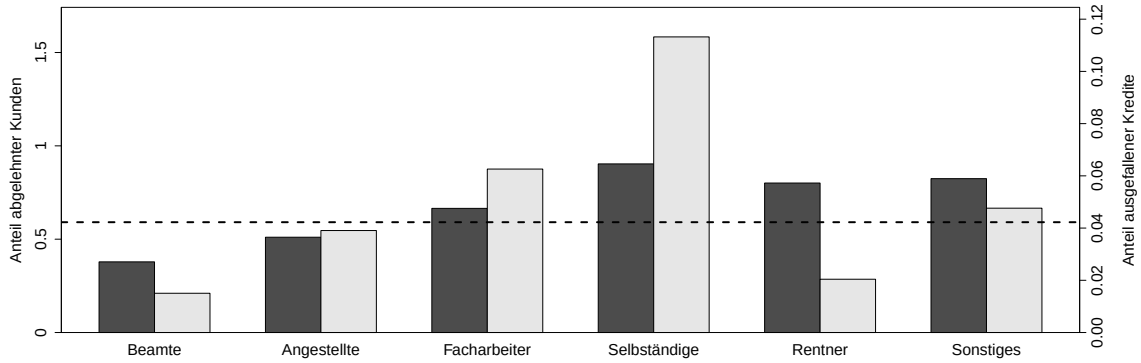

Abbildung B.2 Einfluss der Variable „Beruf" auf Kreditvergabe und Rückzahlung: die dunklen Balken geben den Anteil abgelehnter Kredite, die hellen Balken den Anteil ausgefallener Kredite in der jeweiligen Gruppe an, die Linie stellt den durchschnittlichen Anteil abgelehnter/ausgefallener Kredite dar.

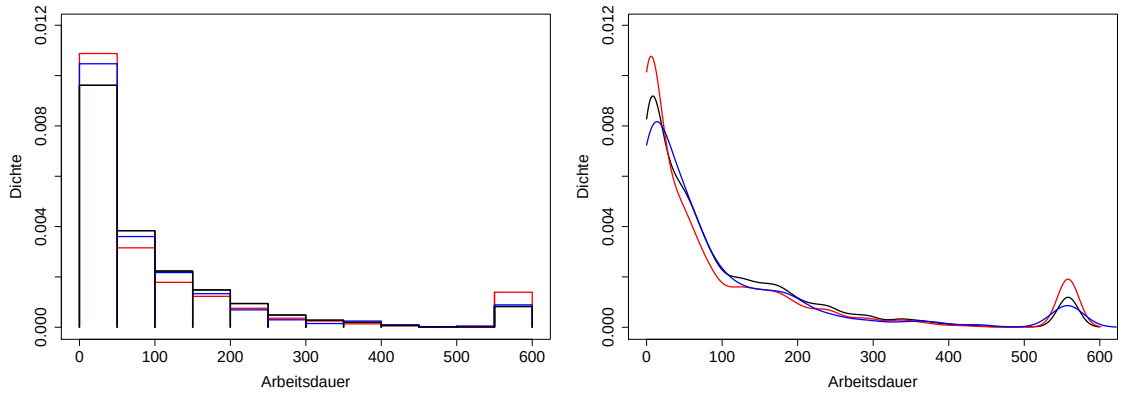

Abbildung B.3 Einfluss der Variable „Arbeitsdauer" auf Kreditvergabe und Rückzahlung: Histogramm (links) und Kerndichteschätzung (rechts) für alle Kunden (schwarz), die abgelehnten Kunden (rot) und die schlechten Kunden (blau).

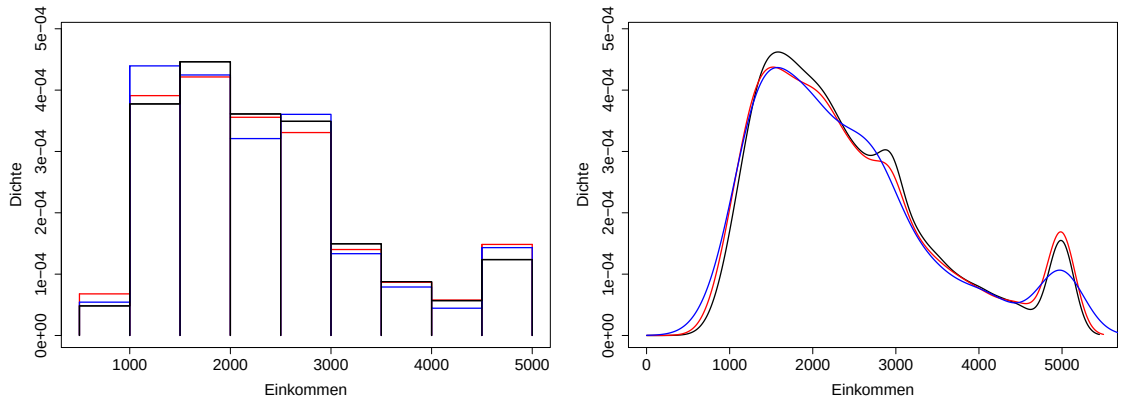

Abbildung B.4 Einfluss der Variable „Einkommen" auf Kreditvergabe und Rückzahlung: Histogramm (links) und Kerndichteschätzung (rechts) für alle Kunden (schwarz), die abgelehnten Kunden (rot) und die schlechten Kunden (blau).

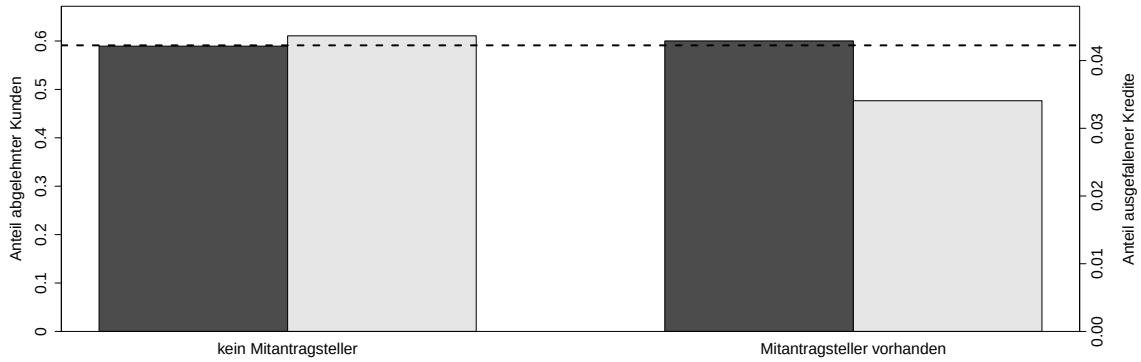

Abbildung B.5 Einfluss der Variable „Mitantragsteller" auf Kreditvergabe und Rückzahlung: die dunklen Balken geben den Anteil abgelehnter Kredite, die hellen Balken den Anteil ausgefallener Kredite in der jeweiligen Gruppe an, die Linie stellt den durchschnittlichen Anteil abgelehnter/ausgefallener Kredite dar.

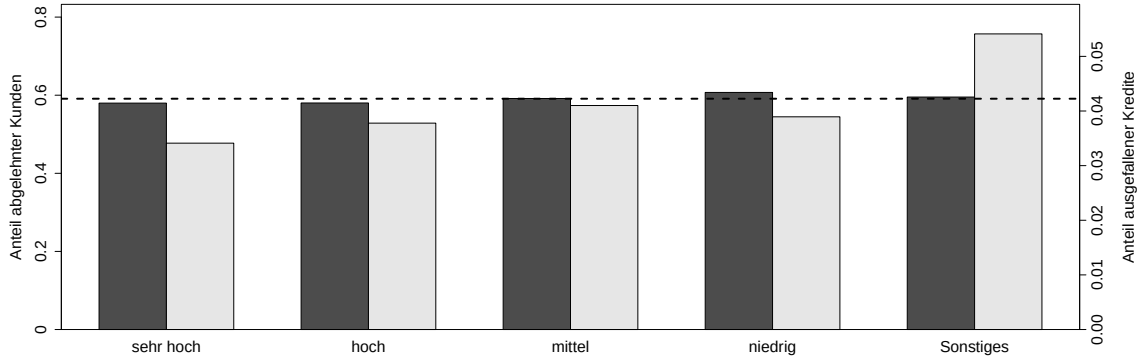

Abbildung B.6 Einfluss der Variable „Kaufkraft" auf Kreditvergabe und Rückzahlung: die dunklen Balken geben den Anteil abgelehnter Kredite, die hellen Balken den Anteil ausgefallener Kredite in der jeweiligen Gruppe an, die Linie stellt den durchschnittlichen Anteil abgelehnter/ausgefallener Kredite dar.

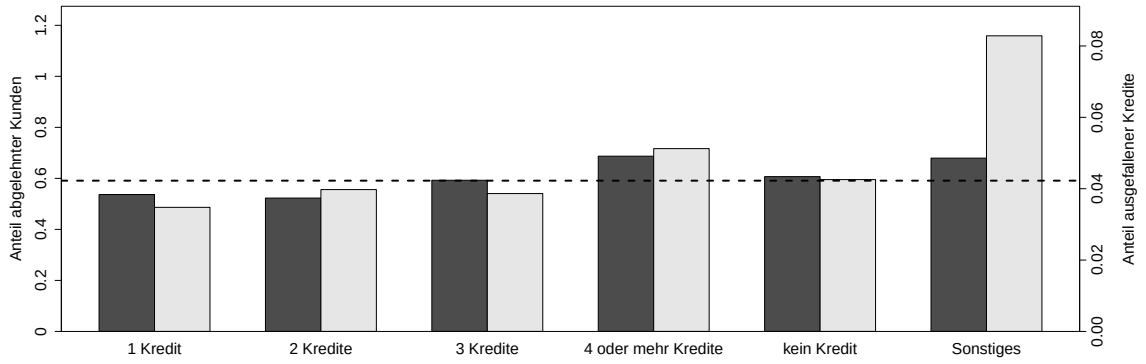

Abbildung B.7 Einfluss der Variable „Kredite" auf Kreditvergabe und Rückzahlung: die dunklen Balken geben den Anteil abgelehnter Kredite, die hellen Balken den Anteil ausgefallener Kredite in der jeweiligen Gruppe an, die Linie stellt den durchschnittlichen Anteil abgelehnter/ausgefallener Kredite dar.

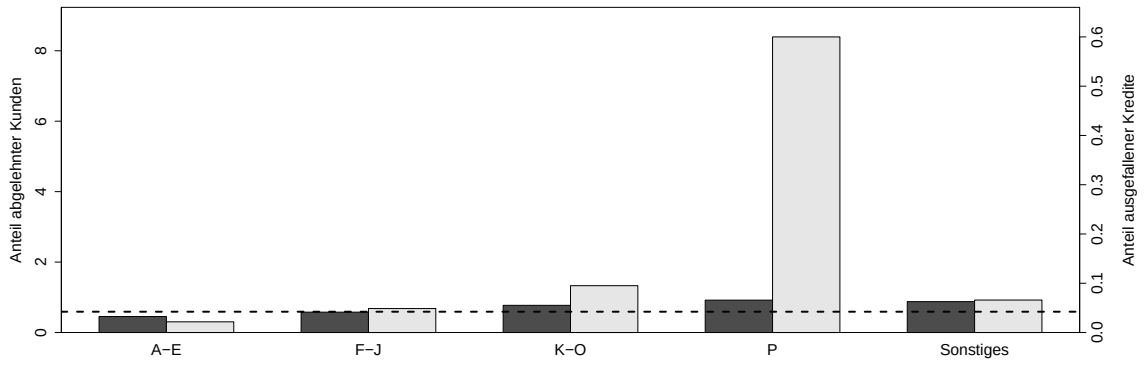

Abbildung B.8 Einfluss der Variable „Schufa" auf Kreditvergabe und Rückzahlung: die dunklen Balken geben den Anteil abgelehnter Kredite, die hellen Balken den Anteil ausgefallener Kredite in der jeweiligen Gruppe an, die Linie stellt den durchschnittlichen Anteil abgelehnter/ausgefallener Kredite dar.

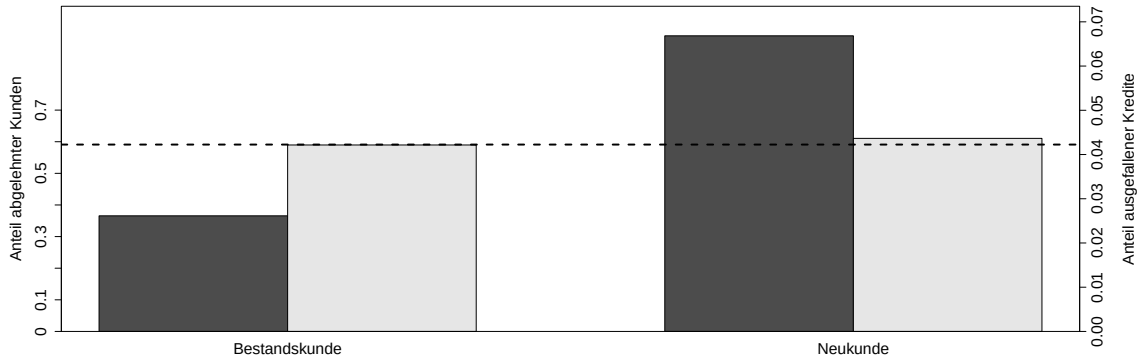

Abbildung B.9 Einfluss der Variable „Neukunde" auf Kreditvergabe und Rückzahlung: die dunklen Balken geben den Anteil abgelehnter Kredite, die hellen Balken den Anteil ausgefallener Kredite in der jeweiligen Gruppe an, die Linie stellt den durchschnittlichen Anteil abgelehnter/ausgefallener Kredite dar.

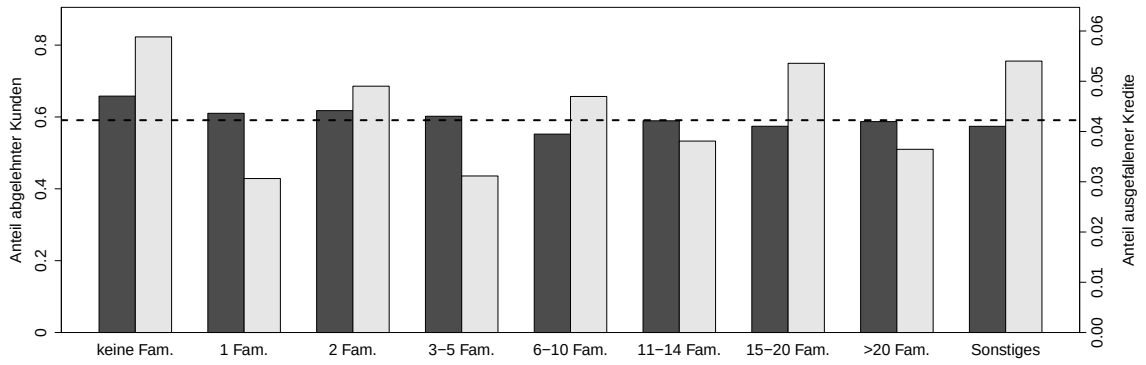

Abbildung B.10 Einfluss der Variable „Haustyp" auf Kreditvergabe und Rückzahlung: die dunklen Balken geben den Anteil abgelehnter Kredite, die hellen Balken den Anteil ausgefallener Kredite in der jeweiligen Gruppe an, die Linie stellt den durchschnittlichen Anteil abgelehnter/ausgefallener Kredite dar.

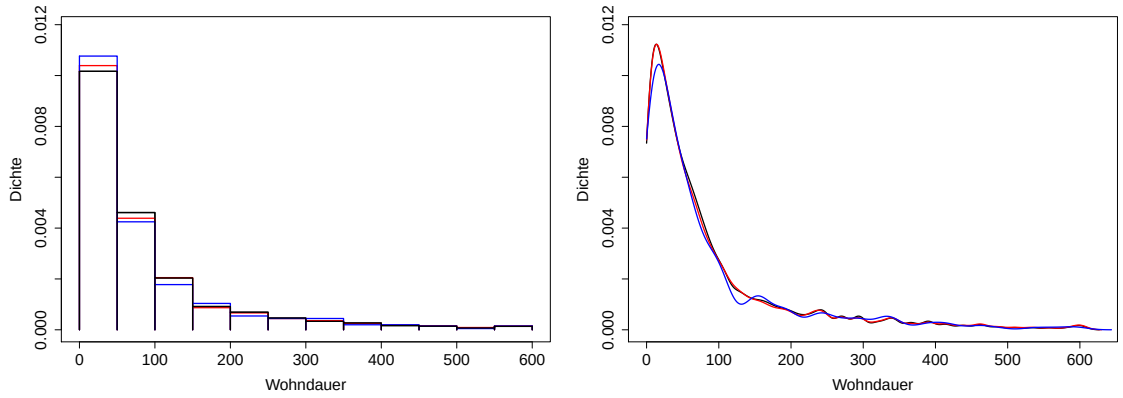

Abbildung B.11 Einfluss der Variable „Wohndauer" auf Kreditvergabe und Rückzahlung: Histogramm (links) und Kerndichteschätzung (rechts) für alle Kunden (schwarz), die abgelehnten Kunden (rot) und die schlechten Kunden (blau).

Tabelle B.1 Geschätztes Logistisches Regressionsmodell für die Akzeptanz eines Kunden durch die Bank in Abhängigkeit der Scorevariablen: angegeben ist jeweils der Schätzer mit Standardabweichung (sd) und p-Wert, wobei für kategorielle Variablen jeweils die Referenzklasse in Klammern angegeben ist.

		$\widehat{\boldsymbol{\theta}}^{\mathrm{ML}}$	sd	p
Konstante		1.676	0.286	0.000
Alter		0.017	0.004	0.000
Familie	verwitwet	0.435	0.302	0.150
(verheiratet)	ledig	-0.118	0.087	0.172
	Lebensgem.	-0.054	0.164	0.743
	geschieden	-0.081	0.114	0.478
	getrennt	-0.216	0.174	0.214
Kinder	1 Kind	-0.053	0.085	0.533
(keine Kinder)	2 Kinder	-0.241	0.098	0.014
	3-6 Kinder	0.495	0.147	0.001
Beruf	Angestellter	-0.561	0.114	0.000
(Beamter)	Facharbeiter	-0.675	0.128	0.000
	Selbständige	-3.147	0.162	0.000
	Rentner	-2.138	0.243	0.000
	Sonstiges	-2.440	0.218	0.000
Arbeitsdauer		-0.002	0.000	0.000
Einkommen		-0.000	0.000	0.564
Mitantragst.	vorhanden	-0.126	0.088	0.151
Kaufkraft	hoch	-0.107	0.104	0.308
(sehr hoch)	mittel	-0.207	0.099	0.037
	niedrig	-0.080	0.114	0.482
	Sonstiges	-0.386	0.138	0.005
Kredite	2 Kredite	0.319	0.090	0.000
(1 Kredit)	3 Kredite	0.223	0.106	0.035
	4 oder mehr	-0.077	0.099	0.435
	kein Kredit	-0.537	0.077	0.000

		$\widehat{\boldsymbol{\theta}}^{\mathrm{ML}}$	sd	p
	Sonstiges	0.454	0.150	0.002
Schufa	B-E	-0.286	0.065	0.000
(A-D)	F-J	-1.390	0.098	0.000
	K-M	-2.316	0.309	0.000
	P	-2.296	0.141	0.000
Neukunde	Neukunde	-3.401	0.077	0.000
Haustyp	1 Fam.	0.108	0.167	0.520
(keine Fam.)	2 Fam.	0.165	0.177	0.351
	3-5 Fam.	0.195	0.170	0.252
	6-10 Fam.	0.339	0.169	0.045
	11-14 Fam.	0.200	0.185	0.280
	15-20Fam.	0.167	0.229	0.466
	>20 Fam.	0.401	0.203	0.049
	Sonstiges	0.477	0.189	0.012
Wohndauer		0.000	0.000	0.939

Tabelle B.2 Geschätztes Logistisches Regressionsmodell für die Akzeptanz eines Kunden durch die Bank in Abhängigkeit der Scorevariablen ohne die Einflussgröße „Wohndauer", geschätzt mittels herkömmlicher ML-Methode und geschätztes Modell w: angegeben ist jeweils der Schätzer mit Standardabweichung (sd) und p-Wert, wobei für kategorielle Variablen jeweils die Referenzklasse in Klammern angegeben ist.

		$\widehat{\boldsymbol{\theta}}^{\mathrm{ML}}$	sd	p	$\widehat{\boldsymbol{\theta}}^{\mathrm{SEL}}$	sd	p
Konstante		1.676	0.286	0.000	-4.406	1.706	0.010
Bonität		–	–	–	6.845	2.987	0.022
Alter		0.017	0.004	0.000	0.144	0.062	0.020
Familie	verwitwet	0.435	0.302	0.150	-1.006	0.984	0.307
(verheiratet)	ledig	-0.118	0.087	0.172	0.314	0.371	0.398
	Lebensgem.	-0.054	0.163	0.741	-1.750	0.911	0.055
	geschieden	-0.081	0.113	0.473	-0.997	0.506	0.049
	getrennt	-0.217	0.174	0.211	2.495	1.370	0.068
Kinder	1 Kind	-0.053	0.085	0.533	-1.019	0.681	0.134
(keine Kinder)	2 Kinder	-0.241	0.097	0.013	-0.834	0.551	0.130
	3-6 Kinder	-0.496	0.147	0.001	0.665	1.015	0.512
Beruf	Angestellter	-0.561	0.114	0.000	-2.853	1.306	0.029
(Beamter)	Facharbeiter	-0.674	0.128	0.000	-0.953	0.932	0.307
	Selbständige	-3.146	0.162	0.000	-5.377	1.485	0.000
	Rentner	-2.137	0.242	0.000	-4.017	1.342	0.003
	Sonstiges	-2.439	0.218	0.000	-5.471	1.818	0.003
Arbeitsdauer		-0.002	0.000	0.000	-0.009	0.002	0.000
Einkommen		-0.000	0.000	0.562	-0.000	0.000	0.139
Mitantragst.	vorhanden	-0.126	0.088	0.150	0.647	0.417	0.121
Kaufkraft	hoch	-0.107	0.104	0.308	0.392	0.414	0.343
(sehr hoch)	mittel	-0.207	0.099	0.037	0.190	0.460	0.680
	niedrig	-0.080	0.114	0.483	-0.173	0.409	0.673
	Sonstiges	-0.387	0.137	0.005	2.236	1.871	0.232
Kredite	2 Kredite	0.319	0.090	0.000	0.895	0.500	0.074

		$\widehat{\boldsymbol{\theta}}^{\mathrm{ML}}$	sd	p	$\widehat{\boldsymbol{\theta}}^{\mathrm{SEL}}$	sd	p
(1 Kredit)	3 Kredite	0.223	0.106	0.035	-0.357	0.388	0.357
	4 oder mehr	-0.077	0.099	0.435	0.397	0.509	0.435
	kein Kredit	-0.537	0.077	0.000	0.512	0.500	0.306
	Sonstiges	0.453	0.150	0.003	1.915	1.255	0.127
Schufa	B-E	-0.287	0.065	0.000	0.262	0.279	0.348
(A-D)	F-J	-1.390	0.098	0.000	0.397	0.861	0.645
	K-M	-2.317	0.309	0.000	-0.122	0.732	0.868
	P	-2.296	0.141	0.000	-1.793	0.666	0.007
Neukunde	Neukunde	-3.401	0.077	0.000	-5.828	2.164	0.007
Haustyp	1 Fam.	0.109	0.167	0.515	0.864	0.528	0.102
(keine Fam.)	2 Fam.	0.165	0.177	0.351	0.808	0.595	0.175
	3-5 Fam.	0.195	0.170	0.253	1.163	0.632	0.066
	6-10 Fam.	0.338	0.169	0.045	1.586	0.722	0.028
	11-14 Fam.	0.199	0.185	0.281	0.611	0.554	0.270
	15-20Fam.	0.166	0.229	0.467	0.218	0.633	0.730
	>20 Fam.	0.400	0.203	0.049	1.118	0.576	0.052
	Sonstiges	0.477	0.189	0.012	-0.956	1.789	0.593

Tabelle B.3 Geschätztes Logistisches Regressionsmodell für die Bonität eines Kunden in Abhängigkeit der Scorevariablen: angegeben ist jeweils der Schätzer mit Standardabweichung (sd) und p-Wert, wobei für kategorielle Variablen jeweils die Referenzklasse in Klammern angegeben ist.

		$\widehat{\boldsymbol{\beta}}^{\mathrm{ML}}$	sd	p	$\widehat{\boldsymbol{\beta}}^{\mathrm{SEL}}$	sd	p
Konstante		5.726	0.930	0.000	4.638	0.983	0.000
Alter		-0.032	0.013	0.012	-0.053	0.012	0.000
Familie	verwitwet	0.264	0.873	0.762	0.997	0.806	0.216
(verheiratet)	ledig	-0.299	0.269	0.267	-0.585	0.287	0.042
	Lebensgem.	1.632	1.030	0.113	3.134	1.178	0.008
	geschieden	-0.121	0.336	0.718	-0.053	0.370	0.887
	getrennt	-0.663	0.424	0.118	-1.028	0.505	0.042
Kinder	1 Kind	0.007	0.271	0.980	0.240	0.333	0.471
(keine Kinder)	2 Kinder	-0.450	0.293	0.125	-1.033	0.450	0.022
	3-6 Kinder	-0.124	0.517	0.811	-0.754	0.325	0.020
Beruf	Angestellter	-0.826	0.447	0.065	-0.997	0.431	0.021
(Beamter)	Facharbeiter	-1.336	0.468	0.004	-1.788	0.301	0.000
	Selbständige	-1.958	0.578	0.001	-3.296	0.618	0.000
	Rentner	0.635	1.321	0.631	3.459	1.229	0.005
	Sonstiges	-0.621	0.876	0.478	-1.433	1.240	0.248
Arbeitsdauer		0.003	0.001	0.020	-0.002	0.001	0.156
Einkommen		0.000	0.000	0.090	0.001	0.000	0.000
Mitantragst.	vorhanden	0.102	0.284	0.719	-0.240	0.267	0.370
Kaufkraft	hoch	-0.188	0.327	0.564	-0.178	0.391	0.648
(sehr hoch)	mittel	-0.219	0.311	0.482	-0.941	0.348	0.007
	niedrig	-0.129	0.353	0.715	-0.067	0.414	0.871
	Sonstiges	-0.669	0.401	0.095	-2.248	0.385	0.000
Kredite	2 Kredite	-0.119	0.259	0.647	-0.058	0.340	0.865
(1 Kredit)	3 Kredite	-0.113	0.312	0.717	-0.222	0.450	0.623
	4 oder mehr	-0.365	0.290	0.208	-1.165	0.306	0.000
	kein Kredit	-0.294	0.239	0.218	-1.360	0.287	0.000

		$\widehat{\boldsymbol{\beta}}^{\mathrm{ML}}$	sd	p	$\widehat{\boldsymbol{\beta}}^{\mathrm{SEL}}$	sd	p
	Sonstiges	-0.800	0.353	0.024	-1.574	0.403	0.000
Schufa	B-E	-0.830	0.205	0.000	-0.922	0.229	0.000
(A-D)	F-J	-1.548	0.271	0.000	-2.741	0.295	0.000
	K-M	-4.202	0.615	0.000	-7.086	1.541	0.000
	P	-0.804	0.477	0.092	-2.143	0.393	0.000
Neukunde	Neukunde	0.068	0.329	0.836	-0.091	0.500	0.856
Haustyp	1 Fam.	0.340	0.496	0.494	0.959	0.741	0.196
(keine Fam.)	2 Fam.	-0.165	0.506	0.745	0.394	0.691	0.568
	3-5 Fam.	0.331	0.503	0.511	0.572	0.701	0.414
	6-10 Fam.	-0.069	0.481	0.886	0.682	0.772	0.377
	11-14 Fam.	0.124	0.543	0.820	0.875	0.770	0.256
	15-20Fam.	-0.165	0.633	0.794	-0.229	0.719	0.750
	>20 Fam.	0.334	0.604	0.580	1.551	0.953	0.104
	Sonstiges	0.256	0.525	0.626	2.686	0.612	0.000
Wohndauer		-0.001	0.001	0.562	0.001	0.001	0.326