Seminararbeit:

Fehler in der abhängigen Variable

Autor: Nina Markovic

Betreuer: Prof. Dr. Thomas Augustin

13. März 2015

Inhaltsverzeichnis

1 Einleitung

Hinsichtlich eines linearen Modells ist man oft der Meinung, dass ein Messfehler in den Einflussvariablen verheerendere Folgen als ein Messfehler in der Zielvariable hat. So sagten Abrevaya und Hausman (2004) aus, dass Messfehler dieser Art grundsätzlich ignoriert werden, da diese letztendlich über die Residuen absorbiert werden und somit vernachlässigbar sind. Inwieweit die Aussage berechtigt ist, wird in den nchsten Kapiteln untersucht. Die folgende Grafik verdeutlicht eine mögliche Auswirkung eines additiven Messfehlers der Response (weiteres in Kapitel 2.2.1):

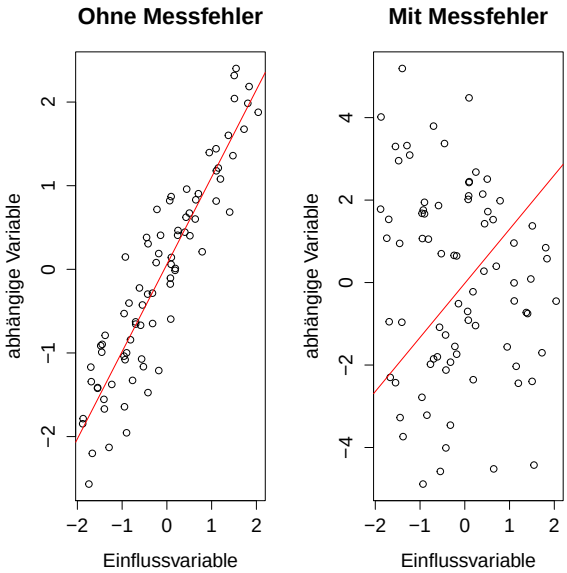

Abbildung 1.1: Simulation eines additiven Messfehlers um eine Gerade gestreuten Daten. Deutlich zu erkennen bei dem Modell mit Messfehler die größere Varianz.

Bei einem Modell mit einem additiven Messfehler streuen die Punkte deutlich mehr um die Regressionsgerade. Die zentrale Problemstellung, die sich hierbei herausstellt, ist, wie sich ein Messfehler in der Zielvariable letztendlich auf die Schätzung der Regressionskoefizienten auswirkt. Um auf die Aussage von Abrevaya und Hausman zurückzugreifen, lässt sich aussagen, dass das Nicht- Beachten eines Messfehlers in der Response eine oberflächliche Betrachtung ist. In der vorliegenden Seminararbeit wird der Umgang mit Messfehlern in einer Response- Variable thematisiert und auf Methoden eingegangen, die zu besseren Schätzungen der Regressionskoeffizienten führen.

2 Theorie

2.1 Auswirkungen eines Messfehlers in einer abhängigen Variable

Linearer Zusammenhang

Wie bereits in der Einleitung erwähnt, muss untersucht werden welche Auswirkungen ein Response- Messfehler auf ein Modell haben kann, wenn der Zusammenhang zwischen der Einflussvariablen und Zielvariablen ein linearer ist. Es ist offensichtlich, dass das Modell mit einem Response- Messfehler sich schlechter an die Daten anpasst, was sich schon an Abbildung 1.1 erkennbar macht: Die Punkte streuen wesentlich mehr um die Regressionsgerade und folglich ist die Anpassung der Regressionsgeraden an die Punkte schlechter. Hierbei wurde in der Simulation ein Messfehler V, mit $V \sim \mathbb{N}(0; 3.0)$ auf die Zielvariablen Y addiert.

Diese Auswirkung macht sich am stärksten am R^2 bemerkbar:

R^2 eines Modells ohne Messfehler	R^2 eines Modells mit Messfehler
0.8032	0.3102

Tabelle 2.1: Verglich der R^2

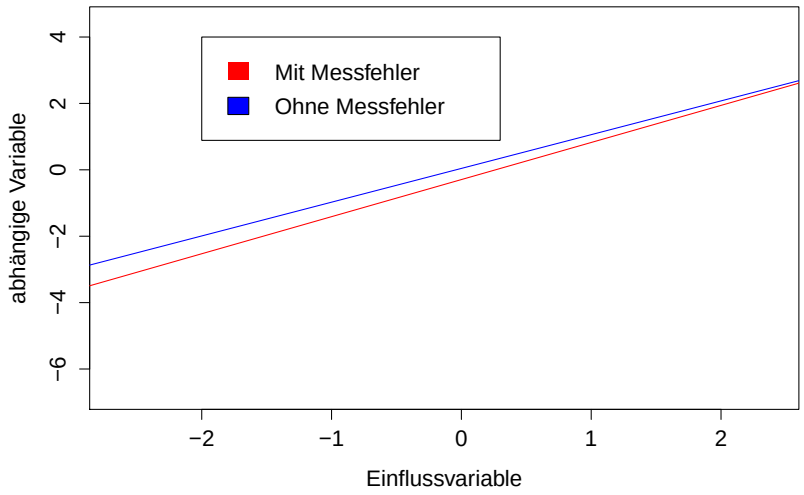

Abbildung 2.1: Vergleich von aus simulierten Daten mit und ohne Messfehler geschätzten Regressionsgeraden

Simuliert man die Daten mit einem Messfehler in der Response genügend oft und lässt darauf immer ein Modell schätzen, so streuen die Regressionsgeraden wesentlich mehr als bei einem Modell ohne Messfehler.

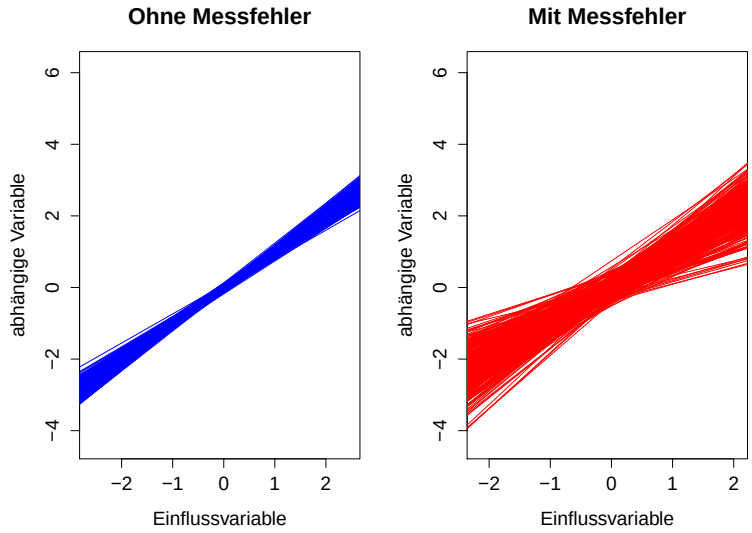

Abbildung 2.2: 350 simulierte Regressionsgeraden für jeweils ein Modell mit und ohne Messfehler.

Ein additiver Messfehler in einer Response- Variablen bewirkt, dass die Varianz der beobachteten Response Variablen deutlich steigt und in Folge dessen variieren die Re-

gressionsgeraden stärker.

Darüber hinaus muss noch untersucht werden, ob das aufgestellte z.B. lineare Modell mit einem Messfehler noch gültig ist. Denn ein Messfehler bewirkt eine höhere Streuung der Zielvariable und es wäre denkbar, dass in einem linearem Modell die meisten Daten nicht mehr durch das Modell erklärt werden könnten. Eine der gängigsten Methoden,um Modellverletzungen festzustellen, sind die Untersuchungen der Diagnostik-Plots. Rügamer (2014)

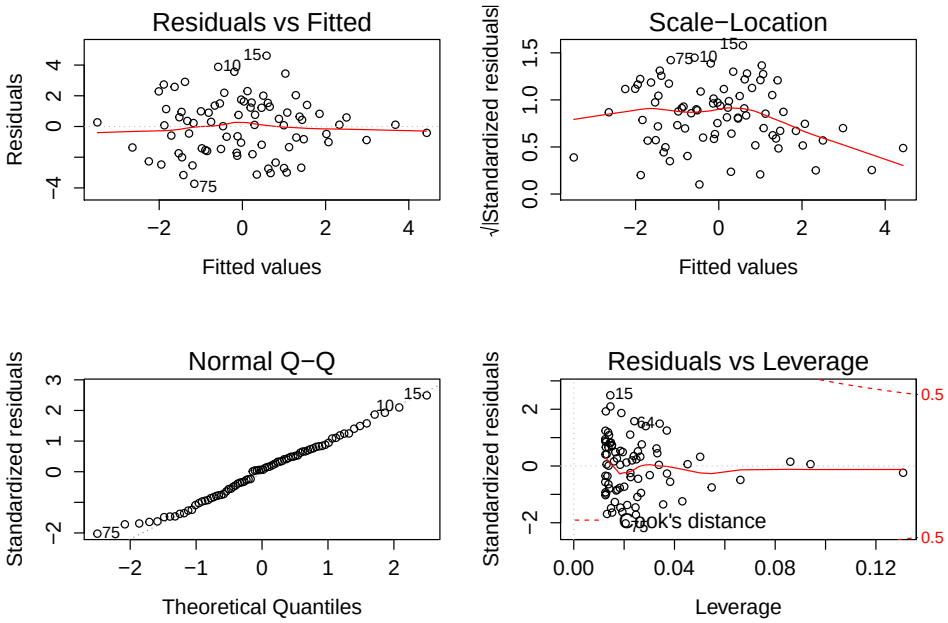

Abbildung 2.3: Diagnostik Plots eines Modells mit Response- Messfehler. Hier sind grobe Modellverletzungen Heteroskedastizität oder Nicht- Normalverteilung der Residuen nicht erkennbar.

Auf den ersten Blick sind keine groben Modellverletzungen, wie Nicht- Normalverteilung der Residuen oder Heteroskedastizität, zu erkennen. Auch in weiteren Untersuchungen auf Unkorreliertheit der Residuen und auf Linearität zwischen der Response und einer Einfluss-variable, sind keine Auffälligkeiten zu beobachten.

Besitzt ein Modell bereits heteroskedastische Eigenschaften, so verstärkt ein Messfehler laut Carroll et al. (2006, S.344) diese Eigenschaft,da ein unverzerrter Response Messfehler die Varianzfunktion verändert und somit sich auf die Residuen auswirkt. Das Ignorieren der heteroskedastische Eigenschaft führt, aufgrund hoher Varainzen, zu ineffizienten Schätzungen der Regressionsparameter.

Nicht linearer Zusammenhang

Ist der Zusammenhang einer Einflussgröße mit der Zielgröße nicht linear, sondern beispielsweise ein quadratischer Zusammenhang ($Y = \beta_0 + \beta_1 Z + \beta_2 Z^2$), so kann ein Messfehler bewirken, dass sich der Zusammenhang scheinbar als ein linearer herausstellt, obwohl dieser in Wahrheit nicht eintritt oder auch umgekehrt.

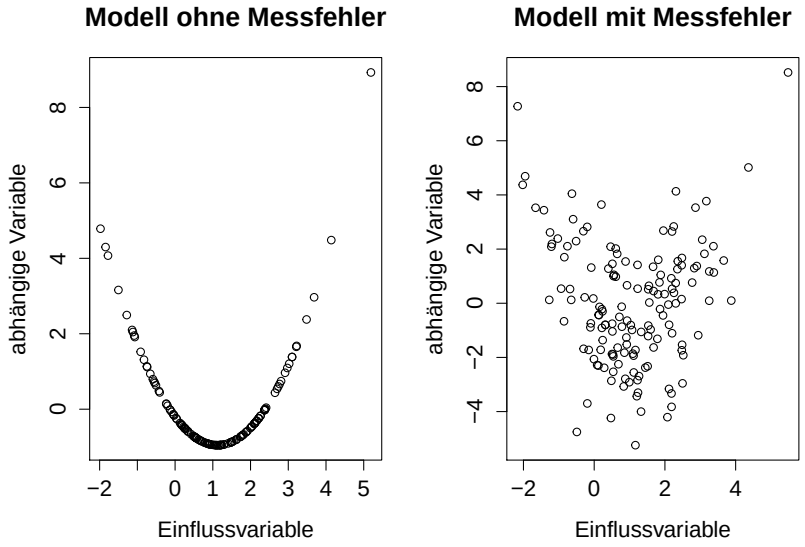

Abbildung 2.4: In dem Modell mit Messfehler ist der quadratischer Zusammenhang ($Y = \beta_0 + \beta_1 Z + \beta_2 Z^2$) nicht mehr erkennbar

Wird nun fälschlicherweise, aufgrund des Messfehlers, das Modell ($Y = \beta_0 + \beta_1 Z$) aufgestellt und die quadratische Einflussgröße Z^2 weggelassen, so werden Modellverletzungen, wie z.B. Nicht- Normalverteilung der Residuen, mit steigender Varianz des Messfehlers ausgeblendet und sind in einem Modell mit Response- Messfehler nur schwer erkennbar.

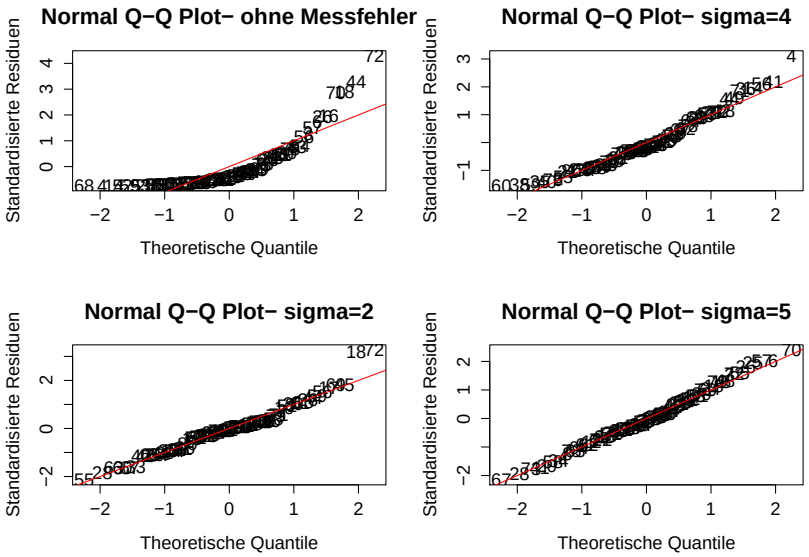

Abbildung 2.5: Modellverletzung in dem Modell ohne Messfehler sind deutlich zu erkennen. Im Modell mit Messfehler und einer hohen Varianz des Messfehlers sind die Modellverletzung kaum zu erkennen

Nach Carroll et al. (2006, S.341) kann bei strengen nicht- linearen Regressionsmodellen eine große Varianz des Messfehlers und somit auch der Zielvariable auch noch andere Folgen hervorrufen. Inferenz bei nicht linearen Modellen basiert oft auf einer Approximation durch ein lineares Regressionsmodell, beispielsweise durch Entwicklung einer Taylor Reihe für β um den den wahren Wert β_0:
$Y_i = m_Y(Z_i, \beta) + \epsilon_i \approx m_Y(Z_i, \beta_0) + f'(Z_i, \beta_0)(\beta - \beta_0) + \epsilon_i$
Der Fehler in der Taylor Approximation geht gegen Null, wenn sich β an β_0 nähert. Eine tendenzielle steigende Varianz des Messfehlers (σ_u)führt zu einer höheren Variation von $\hat{\beta}$ um β_0, welches zu Folge hat, dass die Approximation nicht mehr ganz so exakt ist. Erkennbar ist dieses an folgender Abbildung:

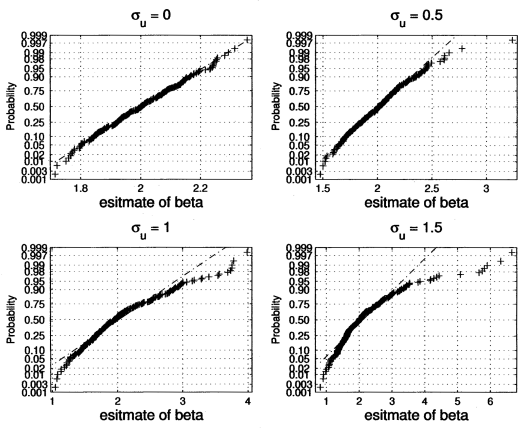

Abbildung 2.6: Abbild von $\hat{\beta}$ mit 250 Simulationen eines Exponentiellen Regressionsmodells (Carroll et al. (2006, S.341))

Eine steigende Varianz der Response führt nicht nur zu einer höheren Varianz von $\hat{\beta}$, sondern wirkt sich auch auf dessen Schiefe aus. (Vgl. Carroll et al. (2006, S.341))

2.2 Arten von Messfehlern in der Response Variablen

2.2.1 Additiver Messfehler

Angenommen Z bezeichnet die Einflussvariable und Y die wahre Zielvariable und folgen einem linearen Modell der Form:

$$Y_i = \beta_0 + \beta_1 Z_i + \varepsilon_i, \text{ wobei } \varepsilon_i \sim \mathbb{N}(0, \sigma^2)$$

S wird als die beobachtete fehlerbehaftete Zielvariable bezeichnet. Die Variable S setzt sich laut Carroll et al. (2006, S.340)zusammen aus der wahren Zielvariablen Y und einem zufälligen Fehler V:

$$S_i = Y_i + V_i \text{ wobei } V_i \sim \mathbb{N}(0; \tau^2) \text{ und } V_i \text{ iid}$$

Die Konklusion aus dem vorherigen Kapitel 2.1, über die höhere Varianz in der Zielvariable bedingt auf den Einflussvariablen verursacht durch einen Messfehler, lässt sich im Fall eines additiven Messfehlers leicht beweisen:

$$\mathbb{E}(S|Z) = \mathbb{E}(Y + V|Z) = \mathbb{E}(Y|Z) + \mathbb{E}(V) = \mathbb{E}(Y|Z)$$

$$\mathbb{V}(S|Z) = \mathbb{V}(Y + V|Z) = \mathbb{V}(Y|Z) + \mathbb{V}(V) + 2Cov(Y, V) = \mathbb{V}(Y|Z) + \mathbb{V}(V)$$

,da Y und V voneinander unabhängig sind.

Angenommen der Erwartungswert von Y ist gegeben durch m_Y(Z,B) und Varianz durch σ_Y^2. Der Erwartungswert der beobachteten Responsevariablen S ist identisch mit dem von Y, m_Y(Z,B),die neue Varianz jedoch addiert sich mit der Varianz des Messfehlers: $\sigma_{new,S}^2 = \sigma_Y^2 + \sigma_V^2$

Einige Konklusionen:

- In einem linearen Regressionsmodell mit einem homoskedästischen und unverzerrten Messfehler der Response, führt ein additiver Messfehler zu einer größeren Varianz der beobachteten Response Variablen,ohne Verzerrung in den Schätzungen zu erzeugen, da sich die Erwartungswerte $\mathbb{E}(\hat{\beta}_{messf}|Z) = \mathbb{E}(\hat{\beta}_{wahr}|Z)$ nicht unterscheiden.
- Die Steigerung der Varianz führt , nach Carroll et al. (2006)[S.341], zur Reduzierung der Güte einiger Tests (z.B. Konfidenzintervalle).

2.2.2 Linearer Messfehler

Der wahre lineare Zusammenhang zwischen einer Response Y und einer Einflussvariablen Z behält weiterhin die Form eines linearen Zusammenhangs:

$$Y_i = \beta_0 + \beta_1 Z_i + \varepsilon_i, \text{ wobei } \varepsilon_i \sim \mathbb{N}(0, \sigma^2)$$

Nun wird ein Modell spezifiziert, das den Zusammenhang zwischen der Zielvariablen mit Messfehler und der wahren Zielvariablen darstellt (Vgl. Carroll et al. (2006, S.343)). Angenommen die beobachtete fehlerhafte Variable S folgt einem linearen Modell abhängig von der wahren Zielvariablen Y. Dann lautet das lineare Modell von S bedingt auf (Y,Z) für eine Beobachtung i:

$$S_i = \gamma_0 + \gamma_1 Y_i + \varepsilon_i$$

Der neue lineare Zusammenhang der beobachteten Daten, d.h. der Zusammenhang zwischen der beobachteten Response- Variablen S und der Einflussvariablen Z, ergibt sich zu:

$$S_i = \gamma_0 + \beta_0\gamma_1 + \gamma_1\beta_1 Z_i + \varepsilon_i$$

Hinsichtlich der Parameterschätzung treten nun Verzerrungen auf, da ein linearer Messfehler bewirkt, dass sich der Erwartungswert von Y und S unterscheidet:

$$\mathbb{E}(S|Z) = \mathbb{E}(\gamma_0 + \gamma_1 Y|Z) \neq \mathbb{E}(Y|Z)$$

Folglich gilt:

$$\mathbb{E}(\hat{\beta}_{messf}|Z) \neq \mathbb{E}(\hat{\beta}_{wahr}|Z)$$

2.3 Allgemeine Likelihood Methoden

Als Dichte aller konkreten Beobachtungen, ist die Likelihood für $f_{S|Z}$ im Fall eines Messfehlers in der Response- Varaible deutlich komplizierter zu berechnen, da für die Variable Y ein Modell spezifiziert wurde, aber man nur S beobachtet hat. Es ist wichtig beide Punkte in der Likelihood zu berücksichtigen, um damit später die best mögliche Schätzung der Modellparameter zu gewährleisten. Angenommen in der Variablen S befindet sich ein linearer Messfehler, so beschreibt Pepe und Fleming (1991, S.109) den Erwartungswert von S gegeben Z für alle Beobachtung durch:

$$\mathbb{E}_{\beta,\gamma}(S|Z) = \int \mathbb{E}_{\beta}(Y|Z) dP_{\gamma}(S|Y,X)$$

2.3.1 Likelihood Methoden für Messfehler in einer diskrete Response- Variablen

Im Fall in dem die Varaiblen S,Y diskrete Variable sind und P ein Wahrscheinlichkeitsmaß, formuliert Carroll et al. (2006, S.353) die bedingte Dichte $f_{S|Z}(s|z,B,\gamma)$ als:

$$f_{S|Z}(s|z,B,\gamma) = \sum_{y} f_{Y|Z}(y|z,B) \cdot f_{S|Y,Z}(s|y,z,\gamma)$$

Am Einfachsten lässt sich dies an Wahrscheinlichkeiten verdeutlichen, denn eine bedingte Wahrscheinlichkeit ist auch eine Wahrscheinlichkeit.
Für die bedingte Wahrscheinlichkeit $P(S=s|Z=z)$ gilt:

$$P(S=s|Z=z) = \underbrace{\sum_{y} \underbrace{P(S=s|Y=y,Z=z)}_{:=(1)} \cdot \underbrace{P(Y=y|Z=z)}_{:=(2)}}_{:=(3)}$$

(1): die Wahrscheinlichkeit von S gegeben Y,Z
(2): die Wahrscheinlichkeit von Y gegeben Z
(3): summiert über alle möglichen y

Ist die Response Variable eine binäre Variable und der Zusammenhang zwischen Ziel- und Einflussvariable linear, so folgen wir einem Modell der Logistischen Regression. Der Messfehler in der Zielvariablen wird für den binären Fall Missklassifikation genannt. Im Gegensatz zum stetigen Fall muss noch folgendes unterschieden werden:

- Ein additiver Messfehler macht im binären Fall wenig Sinn. Denn im diskreten, binären Fall ergibt sich ein Messfehler durch z.B. Zuordnung einer Response- Variablen der Klasse 0, obwohl sie in Wahrheit der Klasse 1 entspricht und andersherum.
- Bei Missklassifikation entsteht Verzerrung, da sich der Erwartungswert der beobachteten Response und der wahren Response unterscheidet:

Der Erwartungswert einer wahren Response- Variablen in einem logistische Modell wird gebildet durch:

$$\mathbb{E}(Y|Z) = P(Y = 1|Z) = H(\beta_0 + \beta_1 Z),$$

wobei H(*) eine logistische Verteilungsfunktion darstellt. Nun werden Wahrscheinlichkeiten für die beiden Fälle der Missklassifikation definiert. Angenommen die Missklassifikation hängt nicht von der Einflussvariablen Z ab, so werden die Wahrscheinlichkeiten folgendermaßen definiert:

$\pi_0 = P(S = 1|Y = 0)$: Die Zuordnung einer Beobachtung zu Klasse 1, obwohl diese in Wahrheit der Klasse 0 entspricht z.B. der Patient wird anhand eines Testergebnisses als krank eingestuft, obwohl dieser in Wahrheit gesund ist (falsch positiv)

$\pi_1 = P(S = 0|Y = 1)$: Die Zuordnung einer Beobachtung zu Klasse 0, obwohl diese in Wahrheit der Klasse 1 entspricht z.B. der Patient wird als gesund eingestuft, obwohl der Patient tatsächlich krank ist (falsch negativ)

Der Erwartungswert $\mathbb{E}(S|Z)$ ergibt sich, nach J. Abrevaya und Scott-Morton (1998, S.241), zu:

$$P(S = 1|Z) = \underbrace{\pi_0}_{(1)} + \underbrace{(1 - \pi_0 - \pi_1) H(\beta_0 \dot{+} \beta_1 Z)}_{(2)} \neq P(Y = 1|Z)$$

(1): Dieser Term entspricht der Missklassifikations- Wahrscheinlichkeit, wie schon oben definiert, dass bei $S = 1$ falsch zugeordnet wird.

(2): Entspricht der Gegenwahrscheinlichkeit von (1), d.h. es wird richtig zugeordnet, weder π_0 noch π_1 treten ein und dieses wiederum multipliziert mit der Wahrscheinlichkeit für $S = 1$, wenn keine Missklassifikation vorliegt, also $H(X'\beta)$

Im Fall, dass alle Beobachtungen richtig zugeordnet werden, so betragen die Wahrscheinlichkeiten der Missklassifikation $\pi_0 = \pi_1 = 0$ und der Erwartungswert vereinfacht sich zu: $P(Y = 1|Z) = H(\beta_0 + \beta_1 Z)$

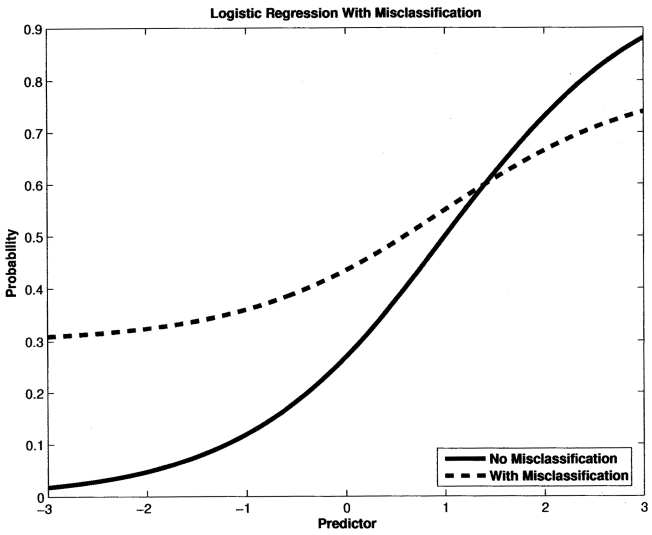

Abbildung 2.7: logistische Regression. Die Graphik zeigt die geschätzten Wahrscheinlichkeit für Y=1 in Abhängigkeit des Prädiktors.

Der nächste wichtige Schritt ist nun die bestmögliche Schätzung aller unbekannten Parameter π_0, π_1, β_0 und β_1. Als ein Lösungsansatz sieht J. Abrevaya und Scott-Morton (1998, S.242) das Aufstellen und maximieren der Log- Likelihoodfunktion.
Angenommen die logistische Verteilungsfunktion H(*) ist bekannt, so nimmt die Likelihood für die Beobachtungen $i = 1, ..., n$ folgende Form an:

$$L(\pi_0, \pi_1, \beta_0, \beta_1) = \prod_{i=1}^{n} \left[\pi_0 + (1 - \pi_0 - \pi_1) H(\beta_0 \dot{+} \beta_1 Z)\right]^{S_i}$$
$$+ \left[1 - \pi_0 - (1 - \pi_0 - \pi_1) H(\beta_0 \dot{+} \beta_1 Z)\right]^{(1-S_i)}$$

(Vgl. (J. Abrevaya und Scott-Morton; 1998, S.242))
Wird auf die gesamte Likelihoodfunktion der natürliche Logarithmus angewendet, so ergibt sich die Loglikelihood:

$$l(\pi_0, \pi_1, \beta_0, \beta_1) = \sum_{i=1}^{n} \left[S_i ln\left(\pi_0 + (1 - \pi_0 - \pi_1) H(\beta_0 \dot{+} \beta_1 Z)\right)\right]$$
$$[(1 - S_i) ln(1 - \pi_0 - (1 - \pi_0 - \pi_1) \cdot H(\beta_0 + \beta_1 Z))]$$

(Vgl. (J. Abrevaya und Scott-Morton; 1998, S.242))
Leitet man die Loglikelihood nach dem interessierenden Parameter ab, so ergibt sich die Scorefunktion. Der Schätzer für den unbekannten Parameter ergibt sich, indem die Scorefunktion gleich Null gesetzt wird und anschließend nach den Parametern aufgelöst wird. Der Vorteil hierbei ist, dass mit Hilfe der ML Schätzung zugleich die Modellparameter (β_0, β_1) geschätzt werden können.

In den seltensten Fällen, in denen die Missklassifikations- Wahrscheinlichkeiten bekannt sind, werden die Wahrscheinlichkeiten in die Likelihood eingesetzt und für β_0, β_1 ergeben sich nach Carroll et al. (2006, S.348), letztendlich konsistente und approximative Schätzer.

2.3.2 Likelihood Methoden für Messfehler in einer stetigen Response- Variablen

Man bezeichne $f_{S|Z}(s|z, B, \gamma)$ als die Dichte für alle Beobachtungen i=1,...,n und die Variable S sei gegeben durch eine stetige Response-Variable Y mit linearem Messfehler. Sei P weiterhin ein Wahrscheinlichkeitsmaß, so beschreibt sich die bedingte Dichten für den stetigen Fall durch:

$$f_{S|Z}(s|z, B, \gamma) = \int f_{Y|Z}(y|z, B) \cdot f_{S|Y,Z}(s|y, z, \gamma) d\lambda$$

wobei λ das Lebesgue- Maß ist.

Die Schätzer der ML- Methode aller Parameter berechnen sich nach Carroll et al. (2006, S.353) auch hier durch das Maximieren der Loglikelihoodfunktion. Oft erweist sich diese Methode als eine komplizierte Vorgehensweise, da die Berechnung der Likelihood- Funktion meistens sehr mühsam erfolgt. Ein weiterer Schwachpunkt der ML Schätzung ist die starke Sensibilität gegenüber der Annahme über die Verteilung der fehlerbehafteten Varaible S.
Die Variable S wird als eine 'surrogate' Response bezeichnet, wenn ihre Verteilung nur von der wahren Response Y abhängig ist und nicht von der Einflussgröße. In diesem Fall vereinfacht sich die bedingte Dichtefunktion zu $f_{S|Y,Z}(s|y, z, \gamma) = f_{S|Y}(s|y, \gamma)$.

2.4 Allgemeine Validierungsdaten

2.4.1 Validierungsdaten im Fall einer stetigen Response- Variablen

Validierungsdaten sind Daten, bei denen ein Teil wahre Beobachtungen enthält und der Rest sich aus fehlerbehafteten zusammensetzt. Oft werden die Beobachtungen per Zufall den Validierungsdaten zugeordnet.
Carroll et al. (2006, S.343) nennt eine Methode bezogen auf die Problematik aus 2.2, die Verzerrung in der Response Variablen S, verursacht durch einen linearen Messfehler, zu eliminieren. Man versucht hierbei mit Hilfe der Validierungsdaten Information über (γ_0, γ_1) zu gewinnen, um damit später die Variable S so anzupassen, dass keine Verzerrungen in den Parameterschätzungen mehr vorliegen.
Es wird, wie folgt vorgegangen:

1. Eine Modellgleichung wird auf Basis eines Anteils der Validierungsdaten zwischen wahrer Response- Variable Y und Einflussvariable Z aufgestellt, um dabei β_0, β_1 mittels KQ- Methode zu schätzen. Zugleich werden die Regressionskoeffizienten

(γ_0, γ_1) mit Bezug des Restdatensatzes geschätzt. Alle diese Schätzer bilden die Menge $\hat{B}_1$.

2. Nun werden die Schätzer γ_0, γ_1 verwendet, um einen neue unverzerrte Variable S' durch $(S - \gamma_0)/\gamma_1$ zu erzeugen. Der Zusammenhang zwischen S' und Y wird nochmals geschätzt und die Schätzer werden der Menge $\hat{B}_2$ zugeordnet.

3. Um die beste Kombination zwischen den beiden Schätzer $\hat{B}_1$ und $\hat{B}_2$ zu gewinnen, werden beide Schätzer miteinander multipliziert und anschließend mit deren Kovarianzmatrix gewichtet, die sich mit Hilfe von Bootstrap Methoden erzeugen lässt. Die Ergebnissmatrix liefert unverzerrte Schätzer.

(Carroll et al.; 2006, S.343)

2.4.2 Validierungsdaten im Fall einer diskreten Response-

Im diskreten Fall versucht man mit Hilfe der Validierungsdaten die Missklassifikations-Wahrscheinlichkeiten zu schätzen, bevor man die Likelihood Funktion für alle Beobachtungen aufstellt.
Ein möglicher Ansatz zur Schätzung der Missklassifikations Wahrscheinlichkeiten π_0, π_1 ist es, die Wahrscheinlichkeiten als ein Anteilsschätzer zwischen den Beobachtungen, die korrekt klassifiziert worden sind und den Beobachtungen dessen wahrer Wert (hier Y=1) beträgt, zu betrachten. Anschließend werden die geschätzten Missklassifikations Wahrscheinlichkeiten in die Likelihood -Funktion eingesetzt und durch Maximierung der Log- Likelihood Funktion die Regressionsparamter (β_0, β_1) geschätzt. Diese Art von Likelihood Funktionen werden Pseudo- Likelihood genannt. Ein Schwachpunkt dieser Methode ist, dass oft ein genaues Schätzen der Missklassifikations- Wahrscheinlichkeiten nur schwer mit den vorhandenen Daten durchführbar ist und somit sich für die Schätzungen der β_0, β_1 Ungenauigkeiten ergeben(Vgl. Carroll et al. (2006, S.347-349)). So zitierte auch Copas (1988), dass ein genaues und konsistentes Schätzen nur unter einem hohem Stichprobenumfang möglich ist.

2.4.3 Validierungsdaten im Bezug auf die allgemeinen Likelihood Methoden

Man hilft sich mit den Validierungsdaten, um zu einer vereinfachte Form der Likelihood-Funktion zu gelangen.
Dabei splittet (Carroll et al.; 2006, S.354) die Likelihood- Funktion für alle Beobachtungen in zwei Produkte:

$$\prod_{i=1}^{n} \left[f_{Y|Z}\left(y_i|z_i, B\right) f_{S|Y,Z}\left(s_i|y_i, z_i, \gamma\right) \right]^{1-\Delta_i} \left[f_{Y|Z}\left(y_i|z_i, B\right) f_{S|Y,Z}\left(s_i|y_i, z_i, \gamma\right) \right]^{\Delta_i}$$

mit

$$\Delta_i = \begin{cases} 1 & \text{wenn Beobachtung } i \in \text{Validierungsdaten} \\ 0 & sonst \end{cases}$$

In der obigen Funktion befindet sich eine kritische Komponente: die Dichtefunktion $f_{S|Y,Z}(s_i|y_i, z_i, \gamma)$, die oft von Hand aufwendig zu berechnen ist und somit nach einer Approximation dieser Dichte gesucht werden muss. Im Fall, dass die Response- Variable S eine ordinale Größe ist, sieht Carroll et al. (2006, S.354) als Lösungsansatz das Anwenden eines Multinomialen Regressionsmodels vor. Angenommen die Variable S besitzt die Ausprägungen $(1, ..., S)$,so lässt sich die Wahrscheinlichkeit für S bedingt auf Y,Z mit Hilfe eines multinomialen Regressionsmodels beschreiben durch:

$$P(S \geq s|Y, Z) = H(\gamma_{0s} + \gamma_1 Y + \gamma_2 Z), s = 1, ..., S$$

berträgt man die Wahrscheinlichkeit auf Dichten so erhält man eine Approximation für die bedingte Dichte $f_{S|Y,Z}(s_i|y_i, z_i, \gamma)$. Handelt es sich bei S um eine stetige Varaible, so hilft man sich, indem man die Variable S in Levels unterteilt, somit eine künstlich erzeugte ordinale Variable erhält, und anschließend das oben aufgeführte Vorgehen nochmals anwendet. Nach Carroll et al. (2006, S.354-S.355) bringt dieses Vorgehen noch Kritik mit sich:

- Annahme über die Verteilung des Messfehlers könnte verletzt sein.
- die Berechnung der Likelihood weiterhin sehr aufwendig

2.5 Complete Data Methode

Die Complete Data Methoden, auch **Complete- Cases** Schätzer genannt, knüpfen an das Verfahren der allgemeinen Validierunsdaten an. Man macht jedoch bei den Complete Data Methoden eine viel stärkere Anforderung an die Beobachtungen in den Validierungsdaten: In dem neuen Datensatz der Complete Data Methode befinden sich nur diejenige Beobachtungen, in denen Y auch beobachtbar ist, sprich man wählt sich nur die Beobachtungen aus, bei denen man mit Sicherheit sagen kann, dass sie ohne Fehler gemessen worden sind. Alle anderen Beobachtungen werden weggelassen.Da nun ein Teildatensatz mit der wahre Response Y und der Einflussvariable Z vorliegt, sieht die Methode der Complete Data vor ein Modell nur auf Basis der Validierungsdaten aufzustellen. (Carroll et al.; 2006, S.355) Bei Normalverteilung von $Y|Z$ und einem linearen Zusammenhang erfolgt die Schätzung mittels KQ- Methode, ansonsten behilft man sich mit der ML- Schätzung:
Sei $\pi(S, Z)$ die Wahrscheinlichkeit, dass eine Beobachtung den Validierungsdaten zugeordnet wird, so lässt sich die Likelihood Funktion für eine Beobachtung in den Validierungsdaten beschreiben durch:

$$f(Y, S|Z, \Delta = 1) = \frac{\pi(S, Z) f(S|Y, Z, \gamma) f(Y|Z, B)}{\sum_s \sum_y \pi(s, Z) f(s|y, Z, \gamma) f(y|Z, B)}$$

(Vgl. (Carroll et al.; 2006, S.355))

Inwieweit die Complete Data Methode hinsichtlich der Parameterschätzung keine Probleme hervorruft, wird im nächsten Kapitel beim Vergleich der Methoden besprochen. Fakt ist, dass diese Methode Informationsverlust verursacht und in Konkurrenz zu den Missklassifikations- Wahrscheinlichkeiten steht, da man diese Wahrscheinlichkeiten komplett ausblenden kann und diese somit nicht mehr berechnet werden müssen.

In manchen Fällen kann man sogar von zwei unabhängigen Datensätzen ausgehen: Im ersten Datensatz, in dem (S,Z) beobachtet wurden ($\Delta = 0$) und im zweiten, in dem nur die Variablen (Y,Z) auftauchen ($\Delta = 1$). Beispielweise kann man Y als das verifizierte Einkommen betrachten (z.B. mit Nachweis einer Lohnabrechnung) und S als das berichtete Einkommen, bzw. ohne jeglichen Nachweis über die tatsächliche Höhe des Einkommens. Aufgrund des Datenschutzes ist es unmöglich die Variablen Y und S gleichzeitig zu erheben. Speziell für solche Fälle nimmt die Likelihood- Funktion für alle Beobachtungen i=1,...,n folgenden Form an:

$$\prod_{i=n}^{n} \{f\left(Y_i|Z_i, B\right)\}^{\Delta_i} \{f\left(S_i|Z_i, B, \gamma\right)\}^{1-\Delta_i}$$

mit

$$\Delta_i = \begin{cases} 1 & \text{wenn Beobachtung } i \in \text{Validierungsdaten} \\ 0 & sonst \end{cases}$$

(Vgl.(Carroll et al.; 2006, S.356))

2.6 Vergleich der Methoden

Für den Vergleich der Methoden, wird der einfachheitshalber der Fall der binären Response betrachtet.
Folgende Abbildungen zeigen Kerndichten- Schätzer, die mit Hilfe der Methoden, Maximum Likelihood, Pseudolikelihood und Complete Cases, berechnet wurden. In der ersten Abbildung wurden die Validierungsdaten per Zufall gebildet und in der zweiten Abbildung wurden die Beobachtungen abhängig von S und Z den Validierungsdaten zugeordnet:

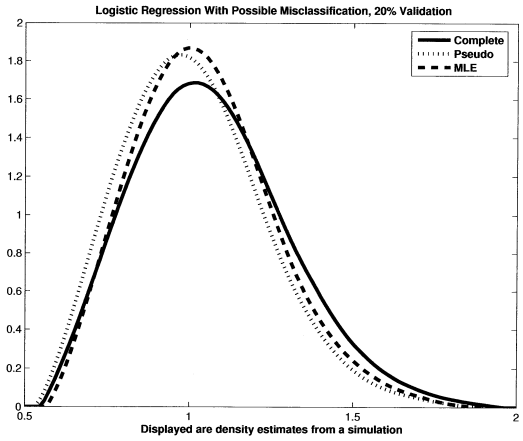

Abbildung 2.8: Vergleich der Methoden mit zufällig ausgewählten Validierungsdaten (Carroll et al. (2006, S.349))

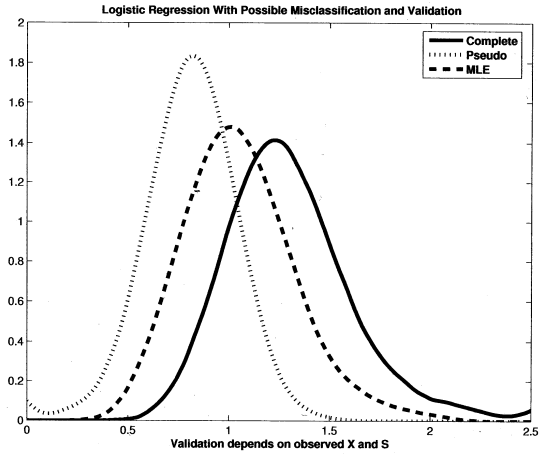

Abbildung 2.9: Vergleich der Methoden mit Auswahl der Validierungsdaten von S und Z abhängig (Carroll et al. (2006, S.350))

Es ist deutlich zu erkennen, dass die Schätzer der Methoden der Complete Data, Pseudolikelihood und Maximum- Likelihood nahezu identisch sind, jedoch Verzerrungen aufzeigen, wenn man die Auswahl der Validierungsdaten von S und Z abhängig macht. Die Schätzungen sind aufgrund der starken Verzerrungen ungültig und führen somit zu falscher Inferenz. Da die allgemeine Likelihood Methode sich nicht auf die Validierungsdaten bezieht sind keine Veränderungen zu erkennen. Am Beispiel des GVHR (Graft-versus-Host-Reaktion) Datensatzes ist deutlich zu erkennen, wie verheerend sich die Methode der Complete Date auf die Schätzung der β auswirkt, wenn die Auswahl in die Validierungsdaten mit einem festen $\pi(S, Z)$ festgelegt wird. Unter GVHR versteht man eine immunologische Reaktion, die in der Folge einer allogenen Knochenmark- oder

Stammzelltransplantation auftreten kann. Der Hauptuntersuchungspunkt dieser Studie war, wie sich das Alter eines Patienten auf das Auftreten einer GHVR auswirkt. Die Variable Y stellt dabei die Existenz einer chronischen GHVR und S die akute. Die Einflußgröße $Z = 0, 1$ hängt davon ab, ob der Patient jünger als 20 ist oder nicht. Die folgende Tabelle liefert die Validierungsdaten dieser Studie, die mit einem gesamten Anteil von $\pi(S, Z) = 1/3$ und abhängig von S und Z, erzeugt wurden:

Validation Data			
Z	**S**	**Y**	Count
0	0	0	19
0	0	1	5
0	1	0	7
0	1	1	14
1	0	0	28
1	0	1	27
1	1	0	8
1	1	1	24
Nonvalidation Data			
0	0	–	47

Abbildung 2.10: Validierungsdaten für das GHVR Beispiel (Carroll et al. (2006, S.351))

Zum Vergleich wird zu einem die Methode der Complete Data auf dem oben aufgeführten Teildatensatz angewendet und zum anderen die Maximum Likelihood Schätzung durchgeführt:

	Validation Data	MLE
$\widehat{\beta}_z$	0.66	1.13
Standard Error	0.37	0.38
p-value	0.078	0.004

Abbildung 2.11: Vergleich Complete Data Methode und ML Schätzung (Carroll et al. (2006, S.352))

Deutlich zu erkennen hierbei die starke Verzerrungen der Schätzungen von β. Bei einem Signifikanzniveau von $\alpha = 0.05$ ergeben sich sogar unterschiedliche Ergebnisse bezüglich des signifikanten Einflusses der Variable Z. (Carroll et al.; 2006, S. 351- 352)
Ist die Auswahl der Validierungsdaten nur von den Einflussvariablen abhängig, so ist es berechtigt bei der Methode der Complete Data das Modell nur auf Basis der (Y,Z)

-Daten aufzustellen und das Standard Verfahren anzuwenden. Die Schätzungen sind nun gültig, jedoch ist dies in der Praxis nur schwer durchführbar, da die Wahrscheinlichkeit in den Valiedierungsdaten nur gültige Werte zu erhalten sehr gering ist. Die fehlerbehaftet Variable S darf dann vollständig ignoriert werden. Dasselbe gilt auch, wenn man die Auswahl an Validierungsdaten komplett dem Zufall überlasst.

2.7 Semiparametrische Methoden

Semiparametrische Methoden knüpfen an die Problematik der Likelihood Methoden an. Diese Methode zielt darauf ab, die Empfindlichkeite der Verteilungsannahmen der beobachetetn Response S zu reduzieren, d.h. einen Schätzer der Dichte $f_{S|Y,Z}$ so zu modellieren, sodass dieser nicht- parametrisch ist. Im konkreten Fall in dem die Verteilung der Variable S linear von der Varaible Y abhänging ist, also $S = \gamma_0 + \gamma_1 Y$, soll die Dichtefunktion $f_{S|Y,Z}$ von γ_0, γ_1 unabhängig sein. Das Hauptproblem, welches sich oft bei einer Variable S mit Messfehler ergibt, ist, dass die Variable in den Validierungsdaten keine Information über die Verteilung von Y liefern kann. Die Idee beispielsweise von Pepe und Fleming (1991), ist es eine neue Variable K zu erheben, die die informative Komponente von S widerspiegelt. In dem Fall wird K auch Surrogat genannt, denn formal hängt ihre Verteilung nur noch von der wahren Response Y ab und nicht mehr von den Einflussgrößen.
Die Dichtefunktion ergibt sich zu $f_{S|Y,Z} = f_{K|Y,Z} = f_{K|Y}$ (siehe Kapitel 2.3). Genauer gesagt, sucht man sich hauptsächlich in den Validierungsdaten diejenige Beobachtungen aus, die für die Verteilungsannahme der wahren Variable Y nützlich sein könnte. Sei beispielsweise Y die Konzentration eines Medikamentes im Blut und die neue erhobene Variable K beschreibt die Dosierung eines Medikamentes. Da die Dosierung des Medikamentes billiger und einfacher zu erheben ist, ist sie vor allem auch genauer als die Konzentration des Medikamentes im Blut. Somit dient die Dichte von $f_{K|Y}$ als empirischer Schätzer für $\hat{f}_{S|Y,Z}$. Anschließend wird geschätzte Dichte in die allgemeine Likelihood - Funktion aus Kapitel 2.3.2 eingesetzt, um den Schätzer $\hat{f}_{S|Z}$ zu erhalten.

$$\hat{f}_{S|Z}(s|z,B) = \int f_{Y|Z}(y|z,B) \cdot f_{K|Y}(s|k) d\lambda$$

Der Schätzer setzt sich nun zusammen aus einem parametrischem Teil $f_{Y|Z}(y|z,B)$ und einem nicht- parametrischem $f_{K|Y}(s|k)$. Die Modellschätzer β_0, β_1 ergeben sich laut (Carroll et al.; 2006, S.356) durch Maximierung folgender Funktion:

$$\prod_{i=n}^{n} \{f(Y_i|Z_i,B)\}^{\Delta_i} \left\{\hat{f}(S_i|Z_i,B)\right\}^{1-\Delta_i}$$

(Vgl. (Carroll et al.; 2006, S.356))

2.8 Funktionelle Verfahren

Die bisher genannten Methoden (Likelihood Methode, Complete Data Methode) gehören zu den strukturellen Verfahren, da man bei diesen Verfahren Verteilungen ber die latente Variable voraussetzt und mit Hilfe dieser Verteilungsannahmen letztendlich zu den bestmöglichen Schätzer gelangt. Darüber hinaus existieren noch die funktionellen Verfahren, wie Simulation Extrapolation (SIMEX) und Regressionskalibrierung, bei denen die latenten Variablen als Unbekannte in die Berrechnungen mit einfließt.

Die zentrale Idee der Regressionskalibrierung ist hierbei eine naive Regression der fehlerbehaftete Zielvariablen auf die wahre Zielvariable zu schätzen und im Nachhinein diesen Zusammenhang zu verwenden, um die bestmögliche Modellparameter- Schätzung zu gewährleisten.(Vgl. Le (2014)) Hierzu eignet sich der Algorithmus nach Carroll et al. (2006), der jedoch für einen Messfehler in einer Einflussvariablen definiert ist. Die Form des Algorithmus der Regressionskalibrierung für einen Messfehler in einer abhängigen Variablen ist dem von Carroll et al. (2006) völlig analog:
Im ersten Schritt wird der Erwartungswert von $S|Y$, mit Hilfe einer Regression von S auf Y, geschätzt. Im zweiten Schritt wird die nicht beobachtbare Variable Y durch den Erwartungswert von $S|Y$ aus Schritt 1 ersetzt und anschließend die Standardanalyse durchgeführt, um die Parameterschätzer zu erhalten. Der dritte Schritt ist mit dem vom Carroll et al. (2006) identisch.

Auch das Verfahren der Simulation Extrapolation (SIMEX), wie es in der Seminararbeit von Hölzl (2015) beschrieben ist, lässt sich analog für einen Messfehler in der abhängigen Variablen anwenden. Hierbei unterscheidet sich nur der erste Schritt im Simulationsschritt des SIMEX Algorithmus. Der Explorationsschritt ist identisch. Im ersten Schritt des Simulationsschrittes werden Pseudo Daten für die Variable S simuliert, indem der wahren Variablen Y ein Messfehler $\sqrt{\lambda\sigma^2}U_{b,i}$ mit $U_{b,i} \sim N(0,1)$ addiert wird.

3 Fazit

Welche Schlussfolgerung aus den beschriebenen Methoden zur besseren Schätzung der Regressionsparameter in einem Modell mit Response Messfehler gezogen wird ist, dass alle Methoden sozusagen Verbesserungen anderer Methoden sind, da immer an bereits genannten Methoden angeknüpft wird. Es ist gut möglich, dass hinsichtlich dieser Methoden noch ein großes Potential zur Ausweitung und Verbesserung besteht. Beispielsweise wäre es denkbar Kombinationen verschiedener Methoden einzuführen. So könnte man in den semiparametrischen Methoden die Schätzung von $\hat{f}_{S|Y,Z}$, anstatt mit einer Surrogaten Variablen auch auf Basis der Validierungsdaten durchführen. Ob dieser Schätzer letztendlich der bessere Schätzer ist, lässt sich mit Hilfe des kleineren MSE (Mean Square Error) feststellen. Darüber hinaus lässt es sich im Allgemeinen aussagen, dass man zu den bestmöglichen Schätzern nur schwer gelangt, wenn ein Teildatensatz mit wahren Beobachtungen oder eine Surrogate nicht vorhanden ist.

Literaturverzeichnis

Abrevaya, J. und Hausman, J. A. (2004). Response error in a transformation model with an application to earnings-equation estimation, *The Econometrics Journal* **7**: 366–388.

Carroll, R. J., Ruppert, D., Stefanski, L. A. und Crainiceanu., C. M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*, 2nd edn, Chapman and Hall, Boca Raton, USA.

Copas, J. B. (1988). Binary regression models for contaminated data, *Journal of the Royal Statistical Society* **50**: 225–265.

Hölzl, A. (2015). Simulation extrapolation (simex). Seminararbeit, Institut für Statistik.

J. Abrevaya, J. A. H. und Scott-Morton, F. (1998). Misclassification of the dependent variable in a discrete-response setting, *Journal of Econometrics* **87**: 239–269.

Le, M. A. (2014). Regressionskalibrierung. Seminararbeit.

Pepe, M. S. und Fleming, T. R. (1991). A nonparametric method for dealing with mismeasured covariate data, *Journal of the American Statistical Associationt* **86**: 108–113.

Rügamer, D. (2014). Modelldiagnose. Folien zum Tutorium Lineare Modelle von Prof. Dr. Helmut Kchenhoff.

Erklärung zur Urheberschaft

Hiermit versichere ich, dass ich die vorliegende Bachelor/Master-Arbeit selbststa"ndig und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe.
München, den 13. März 2015

(Nina Markovic)