Ludwig-Maximilians University

Bachelorarbeit

Regressionskalibrierung versus korrigierte Scorefunktion

Vergleich zweier Verfahren zur Messfehlerkorrektur anhand einer Simulationsstudie

---

Author:
Minh-Anh Le

Supervisor:
Prof. Dr. Thomas Augustin

23. September 2015

Inhaltsverzeichnis

1 Einleitung

Ziel von Regressionsmodellen ist es, den Einfluss von Kovariablen X, Z auf eine Zielgröße Y zu untersuchen. In der Praxis ist es jedoch oft der Fall, dass die interessierenden Größen X nicht fehlerfrei erhoben werden können, da sie z.B. latente Variablen sind, d.h. nicht direkt gemessen werden können oder keine genauen Messinstrumente vorliegen. Somit werden statt X die fehlerbehafteten Variablen X^* erhoben, Z steht hierbei für fehlerfrei gemessenen Kovariablen. Wird die Tatsache, dass Messfehler vorliegen, ignoriert, d.h., wird eine naive Schätzung berechnet, so werden die Parametern des Regressionsmodels nicht konsistent geschätzt (Nakumara [1990, S.1]). Bei Messfehlern in den Kovariablen, werden durch die naive Regression die Kovariableneffekte tendenziell unterschätzt (für den Fall von einer Kovariable vgl. Kapitel 5). Daher ist der Umgang mit Messfehlern eine zentrale Problemstellung der Statistik. Einige Methoden wurden in der Vergangenheit entwickelt um dem Problem nicht konsistenter Schätzungen beim Vorliegen von Messfehlern entgegenzuwirken. Im Rahmen dieser Arbeit werden die Methoden der korrigierten Score Funktion und die Monte-Carlo korrigierte Score Funktion vorgestellt und deren Auswirkung auf die Parameterschätzung der linearen, logistischen und Poisson Regression untersucht. Diese Methoden sind sogenannte funktionelle Methoden, da sie keine oder höchstens minimale Annahmen über die Verteilung der unbeobachteten Variablen X treffen (Caroll, R. J. and Ruppert, D. and Stefanski, L. A. and Crainiceanu, C. M. [2006, S.25]). Im Gegensatz dazu werden bei strukturellen Methoden wie z.B. der Likelihood-, Quasilikelihood- und Bayes-Methode, Annahmen über die Verteilung der stochastischen, unbeobachteten Variablen X getroffen. Allerdings werden Annahmen über den Messfehler U getroffen, so ist es im Umgang mit Messfehlern notwendig zwischen differentiellen Fehler und nicht differentiellen Fehler zu unterscheiden. Letzteres ist eine Voraussetzung für die Anwendung der oben genannten Methoden (Caroll, R. J. and Ruppert, D. and Stefanski, L. A. and Crainiceanu, C. M. [2006, S.37,S.152]), da man bei differentiellen Fehler davon ausgeht, dass diese nicht mehr Informationen zu Y als X enthält, sodass gilt $E[Y|X, X^*, Z] = E[Y|X, Z]$ (Caroll, R. J. and Ruppert, D. and Stefanski, L. A. and Crainiceanu, C. M. [2006, S.36]). Eine weitere Annahme ist, dass ein klassisches additives Fehlermodell vorliegt, (Caroll, R. J. and Ruppert, D. and Stefanski, L. A. and Crainiceanu, C. M. [2006, S.37,S.152]). Explizit bedeutet das, dass $X^* = X + U$, U und X unabhängig sind und dass $U|X \sim N(0, \Sigma_{uu})$ gilt. Vorteile dieser Annahme können in Le [2015, S.19] nachgelesen werden. In dieser Arbeit werden Methoden zur Korrektur von additiven Messfehlern vorgestellt, jedoch können diese Methoden auch auf multiplikative Fehler angewendet werden (vgl. Caroll, R. J. and Ruppert, D. and Stefanski, L. A. and Crainiceanu, C. M. [2006]).

Ziel dieser Arbeit ist es einen Überblick über drei Verfahren zur Messfehlerkorrektur — der Regressionskalibrierung, der korrigierten Score Funktion und der Monte-Carlo korrigierten Score Funktion — zu geben und diese in einer Simulationsstudie, wozu Rstudio verwendet wurde (Rst), zu vergleichen. In Kapitel 2 und Appendix A werden zunächst Notationen eingeführt; in Kapitel 3 wird die Methode der Regressionskalibrie-

rung erläutert und in 3.4 für die lineare-,logistische- und Poisson Regressio untersucht. Die Methode der korrigierten Score Funktion wird in Kapitel 4 dargestellt und für die lineare und Poisson Regression in Abschnitt 4.3 hergeleitet. Die korrigierte Score Funktion für die logistische Regression ist ein Spezialfall und wird in 4.4 separat behandelt. Da das Aufstellen einer korrigierten Score Funktion nicht immer möglich ist, bietet die Monte-Carlo korrigierte Score in Kapitel 4.5 eine flexible Approximation. Anhand der Ausführung des Monte-Carlo-Averaging an der linearen (Kapitel 4.5.2) und Poisson Regression (Kapitel 4.5.3), kann die Idee des Algorithmus veranschaulicht werden. Die konkrete Anwendung der Korrekturverfahren hängt auch vom vorliegendem Datentyp ab, was in Kapitel 3.3 und 4.6 näher erläutert wird. Schließlich werden in einer Simulationsstudie in Kapitel 5 die unterschiedlichen Methoden verglichen und das Verhalten bei z.B. Annahmeverletzungen untersucht.

2 Notation

In dieser Arbeit wird die einfache Regression betrachtet, d.h. mit einer Zielgröße Y und einer Einflussgröße X, jedoch können die vorgestellten Methoden auch auf erweiterte Modelle angewendet werden, beispielsweise bei vorliegen von weiteren (fehlerfrei gemessenen) Einflussgrößen Z. Aus Gründen der notationellen Übersichtlichkeit und den Umfang dieser Arbeit nicht zu überschreiten, wird an einigen Stellen eine Einschränkung auf eine Einflussgröße X bzw. X^* vorgenommen. Neben der Zielgröße Y und der wahren Einflussgröße X, bezeichnet X^* die fehlerhafte Messung von X, und U den Messfehler mit der Messfehlervarianz σ_u^2 bzw. Messfehlerkovarianzmatrix $\Sigma_{uu} = \begin{pmatrix} 0 & 0 \\ 0 & \sigma_u^2 \end{pmatrix}$. Mit σ wird die Fehlervarianz des Fehlers ϵ im linearen Modell bezeichnet. Es wird auch zwischen Matrizen, Vektoren und einzelne Beobachtungen unterschieden, wobei Index i die i-te Beobachtungseinheit bezeichnet von insgesamt n Beobachtungen. Einzelne Beobachtungen werden durch Kleinbuchstaben und den Index i gekennzeichnet, etwa $y_i, x_i, z_i, \epsilon_i$. Vektoren werden mit einem Großbuchstaben bezeichnet, z.B. $X = (x_1, .., x_n)^t$, $Y = (y_1, ., y_n)^t$, $Z = (z_1, .., z_n)^t$ und $U = (u_1, .., u_n)^t$, mit Ausnahme von $\epsilon = (\epsilon_1, .., \epsilon_n)^t$ und dem Vektor $\beta = (\beta_0, \beta_1)^t$, wobei das Superscript t für die Transponierte steht. Matrizen sind mit Großbuchstaben und zwei Unterstrichen kenntlich gemacht, etwa die Designmatrix $\underline{\underline{X}} = \begin{pmatrix} 1 & x_1 \\ . & . \\ . & . \\ 1 & x_n \end{pmatrix}$, analog für $\underline{\underline{X}}^*$. Betrachtet man aus der Designmatrix nur einzelne Zeilen so werden Kleinbuchstaben mit einem Unterstrich und Index i benutzt, wie $\underline{x}_i^t = (1, x_i)$, analog $\underline{x}_i^{*t}$. Weitere Variablen und Notationen werden im Laufe der Arbeit eingeführt. Eine Übersicht aller verwendeten Variablen befindet sich im Appendix A.

3 Regressionskalibrierung

In diesem Kapitel wird die Methode der Regressionskalibrierung zusammengefasst erläutert. Detailliertere Informationen können in Caroll, R. J. and Ruppert, D. and Stefanski, L. A. and Crainiceanu, C. M. [2006, S.65 ff] und Le [2015] nachgelesen werden. Die Regressionskalibrierung ist eine einfache funktionelle Korrekturmethode für klassischem Messfehler bei einer stetigen Einflussgröße im Rahmen von generalisierten linearen Modellen (Caroll, R. J. and Ruppert, D. and Stefanski, L. A. and Crainiceanu, C. M. [2006, S.65 ff]). Sie erfordert keine Annahmen über die nicht beobachtbaren fehlerfreien Variablen X und ist auf alle generalisierten linearen Modelle anwendbar, sowohl beim Vorliegen von additiven als auch multiplikativen Messfehlern (siehe Caroll, R. J. and Ruppert, D. and Stefanski, L. A. and Crainiceanu, C. M. [2006, S.78ff.]).

3.1 Zentrale Idee der Regressionskalibrierung

Die Idee der Regressionskalibrierung ist einfach. Interessiert ist man am Hauptmodell, wobei der Erwartungswert der Zielgröße Y $E[Y|Z,X] = m_Y(Z,X,\beta)$ ist, d.h. die Zielgröße Y wird modelliert gegeben die Einflussgrößen (Z,X), wobei Z fehlerfrei gemessene Kovariablen sind. Statt der wahren Kovariable X wird jedoch die fehlerhafte Variable $X^* = X + U$ mit $U \sim N(0,\sigma_u^2)$ erhoben, sodass ein naives Modell $E[Y|Z,X^*] = m_Y(Z,X^*,\beta^*)$ berechnet werden kann. Wie Eingangs bereits beschrieben führt das Ignorieren von Messfehlern zu verzerrten Schätzungen. Daher wird X durch eine Regression von X auf (Z,X^*) geschätzt, d.h. durch $\hat{X} = \hat{m}_X(Z,X^*,\hat{\gamma})$. Anschließend kann $E[Y|Z,\hat{X}] = m_Y(Z,\hat{X},\beta_{Rk})$ berechnet und die gewöhnliche, jetzt näherungsweise unverzerrte Inferenz durchgeführt werden (Le [2015, S.2]).

3.2 Algorithmus der Regressionskalibrierung

Der Algorithmus der Regressionskalibrierung nach Caroll, R. J. and Ruppert, D. and Stefanski, L. A. and Crainiceanu, C. M. [2006, S.66] und Le [2015, S.3ff] ist wie folgt:

- Schritt 1: Schätze die unbeobachtete Variable X durch eine Regression von X auf (Z, X^*), also $E[X|Z,X^*] = m_X(Z,X^*,\gamma)$. Die Schätzung ist von γ, genauer von der Schätzung $\widehat{\gamma}$ abhängig.
- Schritt 2: Ersetze die nicht beobachtete Variable X durch die im vorherigen Schritt durchgeführte Schätzung, d.h. ersetze im Hauptmodell X durch $\hat{m}_X(Z,X^*,\widehat{\gamma})$. Schäze die Parameter wie üblich. Somit erhält man:

$$E[Y|Z,X^*] \approx m_Y(Z,\underbrace{\hat{m}_X(Z,X^*,\widehat{\gamma})}_{\widehat{X}},\beta_{RK})$$

  Beachte, dass nun keine Gleichheit mehr gilt, da das Regressionskalibrierungsmodell ein approximatives Arbeitsmodell für beobachtete Daten darstellt, es gilt somit

$\beta \approx \beta_{RK}$.

- Schritt 3: Korrigiere mit dem Bootstrap-Verfahren (Appendix B.3) oder der Sandwich-Methode die resultierenden Standardfehler Caroll, R. J. and Ruppert, D. and Stefanski, L. A. and Crainiceanu, C. M. [2006, Appendix A].

Im Gegensatz zur korrigierten Score Funktion in Kapitel 4 produziert die Regressionskalibrierung, abhängig vom Modell, nicht unbedingt konsistente Schätzer, reduziert jedoch den Bias zwischen dem resultierenden Mittelwert und dem wahren Mittelwert (Augustin et al. [2008, S.257] und Le [2015, S.5]). Die Schätzung im ersten Schritt ist abhängig von den vorliegenden Daten bzw. den zusätzlichen Informationen. Wie mit unterschiedlichen Informationen aus den Daten umgegangen werden kann, wird im folgenden Kapitel erläutert.

3.3 Schätzung nach verschiedenen Datentypen

Problematisch ist, dass eine Regression $m_X(Z, X^*, \gamma)$ ohne Beobachtungen von X nicht möglich ist. Daher muss man sich mit der vorliegenden Datenstruktur behelfen. Je nachdem welcher Datentyp also vorliegt ändert sich die Anwendung von Schritt 1 des Regressionskalibrierungsalgorithmus. Für diese Arbeit werden interne Validierungsdaten und Wiederholungsdaten näher betrachtet.

Interne Validierungsdaten Interne Validierungsdaten liegen vor, wenn für einen Teil der Beobachtungen zusätzlich zu X^* wahre X Werte vorliegen. Für diesen Fall schlägt Caroll, R. J. and Ruppert, D. and Stefanski, L. A. and Crainiceanu, C. M. [2006, S.70] vor, eine einfache Regression von X auf die Kovariablen (Z, X^*) aus den internen Validierungsdaten durchzuführen, anschließend sollen für Beobachtungen, die keine wahre X haben, $\hat{X}$ geschätzt werden. Das Einführen einer Dummyvariable, die angibt ob es sich um eine wahre oder geschätzte Beobachtung handelt, führt zu einer Verbesserung der Schätzung (Caroll, R. J. and Ruppert, D. and Stefanski, L. A. and Crainiceanu, C. M. [2006, S.70] und vgl. Ergebnis Le [2015, S.23]).

Wiederholungsdaten In den meisten Fällen jedoch, können keine wahren X-Werte erhoben werden. Es liegen oft nur eine oder mehrere fehlerhafte Messung, $X^*_{i1}, ..., X^*_{ik_i}$ von X_i vor, wobei $\bar{X}^*_i$ der zugehörige Mittelwert ist. Der Index k_i gibt an wie viele Messwiederholungen für Beobachtung i vorliegen. Caroll, R. J. and Ruppert, D. and Stefanski, L. A. and Crainiceanu, C. M. [2006, S.70 ff] und Le [2015, S.6 ff] beschreiben einen Algorithmus der eine lineare Approximation der Regressionskalibrierung liefert. Der Algorithmus ist selbst dann noch anwendbar, wenn die Fehlervarianz σ^2_u bzw. Messfehlerkovarianz Σ_{uu} aus externen Validierungsdaten nicht vorliegt. In diesem Fall werden diese Größen aus den vorliegenden Daten, die mindestens zwei Wiederholungsmessungen enthalten sollten, geschätzt. Die Formeln in diesem Abschnitt gelten allgemein für mehrere X-Variablen bzw. Z-Variablen. Hat man nur eine Variable X fehlerhaft erhoben, ein Spezialfall auf den sich diese Arbeit beschränkt, dann ist X ein Vektor und $\Sigma_{uu} = \sigma^2_u$,

analog für $\Sigma_{zz}, \Sigma_{xx}, \Sigma_{xz}$
Die beste lineare Approximation von X gegeben $(Z, \bar{X}^*)$ ist (Caroll, R. J. and Ruppert, D. and Stefanski, L. A. and Crainiceanu, C. M. [2006, S.70ff] und Le [2015, S.7]):

$$E[X|Z, \overline{X^*}] \approx \mu_x + (\Sigma_{xx}, \Sigma_{zx}) \begin{bmatrix} \Sigma_{xx} + \Sigma_{uu}/k & \Sigma_{xz} \\ \Sigma_{xz}{}^t & \Sigma_{zz} \end{bmatrix}^{-1} \begin{pmatrix} \overline{X^*} - \mu_{x^*} \\ Z - \mu_z \end{pmatrix}.$$

Wiederholungsmessungen ermöglichen die Schätzung der Kovarianz Matrix der Messfehler U_i

$$\widehat{\Sigma}_{uu} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{k_i} \left(X_{ij}^* - \overline{X_{i\cdot}^*}\right) \left(X_{ij}^* - \overline{X_{i\cdot}^*}\right)^t}{\sum_{i=1}^{n} (k_i - 1)}. \tag{1}$$

Wenn in der Regression nur eine Messung ($k_i = 1$) vorhanden ist, muss man auf die aus externen Daten geschätzte Kovarianz Matrix $\widehat{\Sigma}_{uu}$ zurückgreifen.
Weiterhin folgt auf Basis folgender Schätzungen

$$\begin{aligned}
\widehat{\mu}_z &= \overline{Z}, \\
\widehat{\mu}_x &= \widehat{\mu}_{x^*} = \sum_{i=1}^{n} k_i \overline{X_{i\cdot}^*} / \sum_{i=1}^{n} k_i, \\
\nu &= \sum_{i=1}^{n} k_i - \sum_{i=1}^{n} k_i^2 / \sum_{i=1}^{n} k_i, \\
\widehat{\Sigma}_{zz} &= (n-1)^{-1} \sum_{i=1}^{n} (Z_i - \overline{Z_\cdot})(Z_i - \overline{Z_\cdot})^t, \\
\widehat{\Sigma}_{xz} &= \sum_{i=1}^{n} k_i (\overline{X_{i\cdot}^*} - \widehat{\mu}_{x^*})(Z_i - \overline{Z_\cdot})^t / \nu, \\
\widehat{\Sigma}_{xx} &= \left[\left\{ \Sigma_{i=1}^{n} k_i (\overline{X_{i\cdot}^*} - \widehat{\mu}_{x^*})(\overline{X_{i\cdot}^*} - \widehat{\mu}_{x^*})^t \right\} - (n-1)\widehat{\Sigma}_{uu} \right] / \nu
\end{aligned}$$

die Gleichung der geschätzten Regressionskalibrierungsfunktion

$$\begin{aligned}
\widehat{E[X_i|Z_i, \overline{X_{i\cdot}^*}]} &\approx \widehat{\mu}_{x^*} + (\widehat{\Sigma}_{xx}, \widehat{\Sigma}_{zx}) \begin{bmatrix} \widehat{\Sigma}_{xx} + \widehat{\Sigma}_{uu}/k_i & \widehat{\Sigma}_{xz} \\ \widehat{\Sigma}_{xz}^t & \widehat{\Sigma}_{zz} \end{bmatrix}^{-1} \begin{pmatrix} \overline{X_{i\cdot}^*} - \widehat{\mu}_{x^*} \\ Z_i - \overline{Z_\cdot} \end{pmatrix} \\
&\approx m_{X_i}(Z_i, X_i, \widehat{\gamma}) = \widehat{X_i}.
\end{aligned} \tag{2}$$

Mit dem vorgestelltem Algorithmus können z.B. in der linearen Regression konsistente Schätzer und in der logistischen Regression approximativ konsistente Schätzer berechnet werden (Caroll, R. J. and Ruppert, D. and Stefanski, L. A. and Crainiceanu, C. M. [2006, S.72]).

3.4 Theoretische Anwendung der Regressionskalibrierung

Ohne die exakten korrigierten Formeln für jede Regression für die Regressionskalibrierung herzuleiten, kann trotzdem in statistischen Programmen diese Methode angewendet werden (Appendix D.1). Für ein besseres Verständnis der Korrekturmethode werden anhand des einfachen linearen, Poisson und logistischen Regressionsmodells die praktische Anwendung von Schritt 1 und 2 der Regressionskalibrierung im Folgenden gezeigt. Zur Übersichtlichkeit und in Anlehnung an den in der Praxis häufig auftretenden der Fall, sollen von Wiederholungsdaten mit einer Messung ($k_i = 1$) und einer extern geschätzten Messfehlervarianz ausgegangen werden. Die Schritte der Regressionskalibrierung, die für alle Regressionsmodelle gelten, seien wie folgt:

- Schritt 1: Aus den getroffenen Annahmen folgt $Z_i = \overline{Z_\cdot} = 0$, $\hat{\Sigma}_{zx} = 0$, $\hat{\Sigma}_{zz} = 0$, $\hat{\Sigma}_{xx} = \hat{\sigma}_x^2$ und $\hat{\Sigma}_{xx} + \hat{\Sigma}_{uu} = \hat{\Sigma}_{x^*x^*} = \hat{\sigma}_{x^*}^2$. Somit vereinfacht sich Formel (2) zur Schätzung der unbeobachteten Variable X zu

$$\widehat{E[X|\overline{X}^*]} \approx \underbrace{\hat{\mu}_{x^*}(1 - \frac{\widehat{\sigma}_x^2}{\widehat{\sigma}_{x^*}})}_{\widehat{\gamma_0}} + \underbrace{\frac{\widehat{\sigma}_x^2}{\widehat{\sigma}_{x^*}^2}}_{\widehat{\gamma_1}} X^* = \widehat{X} \tag{3}$$

  mit $\hat{\Sigma}_{uu} = \widehat{\sigma}_u^2$ z.B. aus externen Daten.

- Schritt 2: Ersetzte die unbeobachtete Variable X durch die Schätzungen im letzten Schritt, mit dem Wissen aus Caroll, R. J. and Ruppert, D. and Stefanski, L. A. and Crainiceanu, C. M. [2006, S.38] und Gustafson [2004, S.90]:

$$\begin{aligned} E(Y|X^*) &=^1 E(\{E(Y|X, X^*)\}|X^*) \\ &=^2 E(\{E(Y|X)\}|X^*) \end{aligned} \tag{4}$$

[1] iterierter Erwartungswert, Satz der totalen Wahrscheinlichkeit
[2] X^* differentieller Fehler (Gustafson [2004, S.90])
Nach Schritt 1 der Regressionskalibrierungsalgorithmus wird X durch eine lineare Regression durch X^* geschätzt. Es gelten somit folgende Annahmen aus der linearen Regression für
$X = \gamma_0 + \gamma_1 X^* + \tilde{\epsilon}_i$, mit $\tilde{\epsilon}_i \sim N(0, \tilde{\sigma}^2)$

$$X|X^* \sim N(\mu, \tilde{\sigma}^2)$$

Für die Poisson und logistische Regression werden zusätzlich folgende Transformatio-

nen benötigt

$$\pm(\beta_0 + \beta_1 X)|X^* \sim N(\pm(\beta_0 + \beta_1\mu), \beta_1^2\tilde{\sigma}^2) \tag{5}$$

$$\tilde{X} = exp(\pm(\beta_0 + \beta_1 X))|X^* \sim LN(\pm(\beta_0 + \beta_1\mu), \beta_1^2\tilde{\sigma}^2) \tag{6}$$

$$E[\tilde{X}|X^*] = exp(\pm(\beta_0 + \beta_1\mu) + \frac{1}{2}\beta_1^2\tilde{\sigma}^2) \tag{7}$$

An Formel (4) knüpfen die drei folgenden Unterkapiteln an.

3.4.1 Einfache lineare Regression

Für die einfache lineare Regression gilt $Y = \underline{\underline{X}}\beta + \epsilon$ mit $\epsilon \sim N(0, \sigma^2)$. Dann gilt vereinfacht für die interessierende Größe Y die Gleichung $E(Y|X) = \beta_0 + \beta_1 X$. Formel (4) lässt sich umformen zu:

$$\begin{aligned}
&=^3 E(\{\beta_0 + \beta_1 X\}|X^*) \\
&= \beta_0 + \beta_1 E(X|X^*) \\
&\approx \beta_{Rk_0} + \beta_{Rk_1}\widehat{E(X|X^*)} \\
&= \beta_{Rk_0} + \beta_{Rk_1}(\hat{\gamma}_0 + \hat{\gamma}_1 X^*) \\
&= \underbrace{\beta_{Rk_0} + \beta_{Rk_1}\hat{\gamma}_0}_{\beta_{naiv_0}} + \underbrace{\beta_{Rk_1}\hat{\gamma}_1}_{\beta_{naiv_1}} X^*.
\end{aligned} \tag{8}$$

[3] $E(Y|X) = \beta_0 + \beta_1 X$

Beachte, β_{naiv_0} und β_{naiv_1} sind die Parameter einer naiven Regression von Y auf X^*, d.h. wenn nicht berücksichtigt wird, dass Messfehler vorliegen. Unter der Bedingung, dass der Ausdruck 3 gilt, kann man die Schätzer der wahren Effekte β_0, β_1 wie folgt aus (8) extrahieren:

$$\widehat{\beta}_1 \approx \widehat{\beta}_{Rk_1} = \frac{\widehat{\beta}_{naiv_1}}{\widehat{\gamma}_1}, \qquad \widehat{\beta}_0 \approx \widehat{\beta}_{Rk_0} = \widehat{\beta}_{naiv_0} - \widehat{\gamma}_0\widehat{\beta}_{Rk1},$$

sodass im Grunde genommen die korrigierten Schätzer von β_0, β_1 durch eine Regression von Y auf X^* statt auf X geschätzt werden können.

3.4.2 Einfache Poisson Regression

Für die einfache Poisson Regression gilt $\lambda(X) = exp(\underline{\underline{X}}\beta)$. Und für die interessierende Größe Y die Gleichung $E[Y|X] = \lambda(X) = exp(\underline{\underline{X}}\beta)$. Formel (4) lässt sich umformen zu:

$$
\begin{aligned}
&= E[\lambda(X)|X^*] \\
&= E[exp(\underline{\underline{X}}\beta)|X^*] \\
&\approx \hat{E}[exp(\beta_{Rk0} + \beta_{Rk1}X)|X^*] \\
&=^3 exp((\beta_{Rk0} + \beta_{Rk1}\hat{\mu}_x + \frac{1}{2}\beta_{Rk1}^2\hat{\sigma}^2)) \\
&= exp((\beta_{Rk0} + \beta_{Rk1}(\hat{\gamma}_0 + \hat{\gamma}_1 X^*) + \frac{1}{2}\beta_{Rk1}^2\hat{\sigma}^2)) \\
&= exp((\underbrace{\beta_{Rk0} + \beta_{Rk1}\hat{\gamma}_0 + \frac{1}{2}\beta_{Rk1}^2\hat{\sigma}^2}_{\beta_{naiv0}} + \underbrace{\beta_{Rk1}\hat{\gamma}_1}_{\beta_{naiv1}} X^*)) \qquad (9)
\end{aligned}
$$

[3] nach (7)
Beachte, β_{naiv_0} und β_{naiv_1} sind die Parameter einer naiven Regression von Y auf X^*, d.h. wenn nicht berücksichtigt wird, dass Messfehler vorliegen. Unter der Bedingung, dass der Ausdruck (3) gilt, kann man die Schätzer der wahren Effekte β_0, β_1 wie folgt aus Formel (9) extrahieren:

$$
\widehat{\beta}_1 \approx \widehat{\beta}_{Rk_1} = \frac{\widehat{\beta}_{naiv_1}}{\widehat{\gamma}_1}, \qquad \widehat{\beta}_0 \approx \widehat{\beta}_{Rk_0} = \widehat{\beta}_{naiv_0} - \widehat{\gamma}_0\widehat{\beta}_{Rk1} - \frac{1}{2}\widehat{\beta}_{Rk1}^2\hat{\sigma}^2,
$$

sodass im Grunde genommen die korrigierten Schätzer von β_0, β_1 durch eine Regression von Y auf X^* statt auf X geschätzt werden können.

3.4.3 Einfache logistische Regression

Für die einfache logistische Regression gilt für die interessierende Größe Y die Gleichung $E[Y|X] = p(X) = P(Y = 1|X) = \frac{1}{1+exp(-\underline{\underline{X}}\beta)}$. Formel (4) lässt sich umformen zu:

$$
\begin{aligned}
&= E[p(X)|X^*] \\
&= E[\frac{1}{1 + exp(-\underline{\underline{X}}\beta)}|X^*] \\
&= \frac{1}{1 + E[exp(-\underline{\underline{X}}\beta)|X^*]} \\
&\approx \frac{1}{1 + \hat{E}[exp(-\beta_{Rk0} - \beta_{Rk1}X)|X^*]} \\
&=^3 \frac{1}{1 + exp(-(\beta_{Rk0} + \beta_{Rk1}\hat{\mu}_x) + \frac{1}{2}\beta^2_{Rk1}\hat{\sigma}^2)} \\
&= \frac{1}{1 + exp(-(\beta_{Rk0} + \beta_{Rk1}(\hat{\gamma}_0 + \hat{\gamma}_1 X^*)) + \frac{1}{2}\beta^2_{Rk1}\hat{\sigma}^2)} \\
&= \frac{1}{exp(-(\underbrace{\beta_{Rk0} + \beta_{Rk1}\hat{\gamma}_0 - \frac{1}{2}\beta^2_{Rk1}\hat{\sigma}^2}_{\beta_{naiv0}} + \underbrace{\beta_{Rk1}\hat{\gamma}_1}_{\beta_{naiv1}} X^*))}
\end{aligned}
\tag{10}
$$

[3] nach (7)
Beachte, β_{naiv_0} und β_{naiv_1} sind die Parameter einer naiven Regression von Y auf X^*, d.h. wenn nicht berücksichtigt wird, dass Messfehler vorliegen. Unter der Bedingung, dass der Ausdruck (3) gilt, kann man die Schätzer der wahren Effekte β_0, β_1 wie folgt aus Formel (10) extrahieren:

$$
\widehat{\beta}_1 \approx \widehat{\beta}_{Rk_1} = \frac{\widehat{\beta}_{naiv_1}}{\widehat{\gamma}_1}, \qquad \widehat{\beta}_0 \approx \widehat{\beta}_{Rk_0} = \widehat{\beta}_{naiv_0} - \widehat{\gamma}_0\widehat{\beta}_{Rk1} + \frac{1}{2}\hat{\beta}^2_{Rk1}\hat{\sigma}^2,
$$

sodass im Grunde genommen die korrigierten Schätzer von β_0, β_1 durch eine Regression von Y auf X^* statt auf X geschätzt werden können.

4 Korrigierte Score Funktion

Neben der Regressionskalibrierung ist auch die korrigierte Score Funktion eine funktionelle Korrekturmethode für klassische, additive oder multiplikative Fehler (Caroll, R. J. and Ruppert, D. and Stefanski, L. A. and Crainiceanu, C. M. [2006, S.151]), die keine Annahmen über die Verteilung der nicht beobachtbaren wahren Variablen benötigt (Buzas [2009, S.1] und Caroll, R. J. and Ruppert, D. and Stefanski, L. A. and Crainiceanu, C. M. [2006, S.151]). Abgesehen davon, dass diese Methode unabhängig von der Größe des Messfehlers ist (Buzas [2009, S.1] und Kapitel 5.3.1), führt diese bei der Kor-

rektur zu konsistenten Schätzern (Caroll, R. J. and Ruppert, D. and Stefanski, L. A. and Crainiceanu, C. M. [2006, S.151]) und ist eine unverzerrte Score Funktion (Nakumara [1990, S.128]).
In diesem Kapitel wird weiterhin vom Model $E[Y|Z,X] = m_Y(Z,X,\beta)$ bzw. $E[Y|Z,X^*] = m_Y(Z,X^*,\beta^*)$ mit additivem, normalverteilten Messfehler $U \sim N(0,\sigma_u^2)$ ausgegangen, sodass $X^* = X + U$ gilt.

4.1 Zentrale Idee der korrigierten Score Funktion

Das Prinzip und die Vorteile der nicht korrigierten Score Funktion $S()$ für keine Messfehler sind in Appendix B.1 dargestellt, allerdings gelten die Eigenschaften, wie z.B. konsistente Schätzer nicht mehr, wenn man die fehlerhafte Messung X^* in die Score Funktion $S()$ einsetzt. Ignoriert man die Tatsache dass Messfehler vorliegen, produziert sie verzerrte Schätzer.
Die zentrale Idee der korrigierten Score Funktion $S_c()$ ist, dass deren Erwartungswert, unter Berücksichtigung der Verteilung des Fehlers U und bedingt auf Y, Z, X, gleich der Score Funktion der wahren unabhängigen und unbeobachteten Variable X ist (Nakumara [1990, S.1]); kurz: die korrigierte Score Funktion liefert unverzerrte Schätzer. In Formeln:

$$E[S_c(Y,Z,X^*,\beta)|Y,Z,X] = S(Y,Z,X,\beta) \tag{11}$$

Das bedeutet insbesondere, dass sich die vorteilhaften Eigenschaften der Maximum -Likelihood-Schätzung (vgl. Appendix B.1) auf die korrigierte Score Funktion übertragen lassen können. Um den Maximum-Likelihood Schätzer zu erhalten löst man $S(Y,Z,X,\beta) = 0$ nach β auf (näheres Appendix B.1). Liegen Messfehler vor, so erhält man die korrigierten Schätzer durch das Auflösen von $S_c(Y,Z,X^*,\beta) = 0$ nach β. Den so erhaltenen Schätzer bezeichnet man als β^c. Konsistenz der β^c lässt sich dadurch erklären, dass die bedingte Verteilung der korrigierten Schätzer gegeben (Y,Z,X) um den Maximum-Likelihood Schätzer zentriert ist, das wiederrum bekanntlich um den wahren Parametern zentriert ist (Nakumara [1990, S.1]); formal:

$$E[E[S_c(Y,Z,X^*,\beta)|Y,Z,X]|Z,X] = E[S(Y,Z,X,\beta)|Z,X] = 0 \tag{12}$$

Die korrigierte Score Funktion sollte noch weiteren Bedingungen genügen, die in folgender Definition zu finden sind (Nakumara [1990, S.128]):

Definition 1 (Nakumara [1990, S.128]): Sei F eine konvexe Teilmenge eines Parameterraums, der auch β enthält.

- Eine Funktion $l_c(Y,Z,X^*,\beta)$ wird korrigierte log-likelihood genannt, falls

$$E[l_c(Y,Z,X^*,\beta)|Y,Z,X] = l(Y,Z,X,\beta) \tag{13}$$

für beliebiges β in F gilt.

- Wenn $l_c()$ differenzierbar in F ist, so ist $S_c(Y, Z, X^*, \beta) = \frac{\partial l_c}{\partial \beta}$ die korrigierte Score Funktion. Falls $E[\cdot|Y, Z, X]$ und $\partial\beta$ vertauschbar sind, dann gilt zusätzlich

$$E[S_c(Y, Z, X^*, \beta)|Y, Z, X] = S(Y, Z, X, \beta). \tag{14}$$

- Wenn $S_c()$ differenzierbar in F ist, so ist $I_c(Y, Z, X^*, \beta) = -\frac{\partial S_c}{\partial \beta}$ die korrigierte beobachtete Information. Falls $E[\cdot|Y, Z, X]$ und $\partial\beta$ austauschbar sind, dann gilt

$$E[I_c(Y, Z, X^*, \beta)|Y, Z, X] = I(Y, Z, X, \beta). \tag{15}$$

Der Parameter β^c, für den gilt, dass $S_c(Y, Z, X, \beta^c) = 0$ und $I_c(Y, Z, X, \beta^c)$ positiv definit, ist der korrigierter Schätzer.

4.2 Algorithmus der korrigierten Score Funktion

Um eine korrigierte Score Funktion zu erhalten, müssen die Bedingungen in Definition 1 erfüllt sein. Ein allgemeiner Algorithmus zu Findung solch einer Funktion lässt sich wie folgt aufstellen:

- Schritt 1: Stelle die Log-Likelihood Funktion für die wahren X auf:

$$l(Y, Z, X\beta)$$

- Schritt 2: Stelle die Log-Likelihood Funktion für die beobachteten bzw. fehlerhaft gemessenen X^* auf:

$$l(Y, Z, X^*, \beta) = l(Y, Z, X + U, \beta)$$

- Schritt 3: Bestimme $l_c()$,welche aus linear Kombinationen aus $l^*()$ und weiteren Termen bestehen kann, sodass gilt:

$$E[l_c(Y, Z, X + U, \beta)|Y, Z, X] = l(Y, Z, X, \beta)$$

  Allerdings ist es oft nicht möglich exakte korrigierte Score Funktionen aufzustellen, sodass man sich z.B. durch Monte-Carlo-Averaging eine korrigierte Score Funktion konstruieren kann (Caroll, R. J. and Ruppert, D. and Stefanski, L. A. and Crainiceanu, C. M. [2006, S.170 ff]). Dieses Verfahren wird genauer in Kapitel 4.5 erläutert.

- Schritt 3: Die korrigierte Score Funktion ergibt sich aus der Ableitung der $l_c()$:

$$\frac{\partial l_c}{\partial \beta} = S_c(Y, Z, X^*, \beta),$$

(wenn Bedingungen in Definition 1 erfüllt sind.)

- Schritt 4: Die korrigierten Parameterschätzer β^c ergeben sich als Lösung der Gleichung $S_c(Y, Z, X^*, \beta^c) = 0$, falls keine analytische Lösung für β^c möglich ist, so kann man die korrigierten Parameter durch eine Näherung mit Newton-Raphson und den Fisher-Rao-Algorithmus (Appendix B.2) erreichen.

4.3 Exakte korrigierte Score Funktionen

Nicht für alle Verteilungen lässt sich eine exakte korrigierte Score Funktion aufstellen. Wenn die wahre Score Funktion allerdings aus einer Linearkombination aus Produkten von Potenzfunktionen und Exponentialfunktionen besteht, so können exakte korrigierte Scores gefunden werden (Caroll, R. J. and Ruppert, D. and Stefanski, L. A. and Crainiceanu, C. M. [2006, S.172]). Beispiele hierfür sind die lineare und Poisson Regression.

4.3.1 Einfaches lineares Modell

Sei $Y = \underline{\underline{X}}\beta + \epsilon$ mit $\epsilon \sim N(0, \sigma^2)$. Gegeben den beobachteten Werten Y und X ist die Log- Likelihood Funktion:

$$\begin{aligned} l(Y, X, \beta) = &- \frac{1}{2} n \log(2\pi) - n \log(\sigma) \\ &- \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i^2 - 2y_i\beta_0 - 2y_i\beta_1 x_i + \beta_0^2 + 2\beta_0\beta_1 x_i + \beta_1^2 x_i^2) \end{aligned} \tag{16}$$

Die korrigierte Log-Likelihood Funktion ergibt sich zu:

$$\begin{aligned} l_c(Y, X^*, \beta) = &- \frac{1}{2} nlog(2\pi) - nlog(\sigma) \\ &- \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i^2 - 2y_i\beta_0 - 2y_i\beta_1 x_i^* + \beta_0^2 + 2\beta_0\beta_1 x_i^* + \beta_1^2 x_i^{2*} - \beta_1^2\sigma_u^2), \end{aligned} \tag{17}$$

daraus folgen die korrigierten Score Funktionen:

$$\frac{\partial l_c}{\partial \beta_0} = S_{c0}(Y, X^*, \beta) = \frac{1}{\sigma^2} (\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i^*) \tag{18}$$

$$\frac{\partial l_c}{\partial \beta_1} = S_{c1}(Y, X^*, \beta) = \frac{1}{\sigma^2} (\sum_{i=1}^{n} ((y_i - \beta_0 - \beta_1 x_i^*) x_i^*) + n\beta_1\sigma_u^2) \tag{19}$$

und es gilt $E[l_c(Y, X^*, \beta)|Y, X] = l(Y, X, \beta)$. Die korrigierten Schätzer β^c zu erhält man durch lösen der Gleichung $S_{c\cdot}(Y, X^*, \beta^c) = 0$:

$$\beta_0^c = \frac{1}{n}\sum_{i=1}^{n} y_i - \beta_1^c \frac{1}{n}\sum_{i=1}^{n} x_i^*$$
$$\beta_1^c = (\sum_{i=1}^{n} y_i x_i^* - \beta_0^c x_i^*)(x_i^{2*} - n\sigma_u^2)^{-1}$$

(Zwischenschritte und Herleitung im Appendix C.1.1)

4.3.2 Einfaches Poisson Modell

Sei $Y|X$ aus der Poissonverteilung P_λ mit Mittelwert $\lambda(X) = \exp(\underline{\underline{X}}\beta)$. Gegeben den beobachteten Werten Y und X ist die Log-Likelihood Funktion:

$$l(Y, X, \beta) = \sum_{i=1}^{n} -\exp(\beta_0 + \beta_1 x_i) + y_i\beta_0 + y_i x_i \beta_i - log(y_i!) \tag{20}$$

Die korrigierte Log- Likelihood Funktion ergibt sich zu:

$$l_c(Y, X^*, \beta) = \sum_{i=1}^{n} -\exp(\beta_0 + \beta_1 x_i^* - (\frac{1}{2}\beta_1^2\sigma_u^2)) + y_i\beta_0 + y_i\beta_1 x_i^* - log(y_i!) \tag{21}$$

daraus folgen die korrigierten Score Funktionen:

$$\frac{\partial l_c}{\partial \beta_0} = S_{c0}(Y, X^*, \beta) = \sum_{i=1}^{n} y_i - \exp(\beta_0 + \beta_1 x_i^* - \frac{1}{2}\beta_1^2\sigma_u^2) \tag{22}$$
$$\frac{\partial l_c}{\partial \beta_1} = S_{c1}(Y, X^*, \beta) = \sum_{i=1}^{n} y_i x_i^* - \exp(\beta_0 + \beta_1 x_i^* - \frac{1}{2}\beta_1^2\sigma_u^2)(x_i^* - \beta_1\sigma_u^2) \tag{23}$$

Die korrigierten Schätzer β^c erhält man durch lösen der Gleichung $S_{c\cdot}(Y, X^*, \beta^c) = 0$ nach β^c, allerdings gibt es hierfür keine geschlossene Lösung, sodass eine Näherung durch Fischer Rao Algorithmus und Newton Raphson durchzuführen ist (Appendix B.2). (Zwischenschritte und Herleitung im Appendix C.1.2)

4.4 Approximative korrigierte Score Funktion für ein einfaches logistisches Modell

Die korrigierte Score Funktion ist allgemein für ein Modell mit binärer Zielgröße nicht einfach aufstellbar. Aufgrund von Eigenschaften der logistischen Funktion $F(v) = \frac{1}{1+\exp(-v)}$ existieren nach Stefanski [1989, S.18] keine adäquate korrigierte Likelihood Score Funktion für die logistische Regression, wenn ein normalverteilter, additiver Messfehler vor-

liegt. Man kann stattdessen eine approximierte korrigierte Score aufstellen. Es handelt sich beim logistischen Modell also um einen Spezialfall. Daher entwickelten Novick and Stefanski [2002] eine approximative korrigierte Score Funktion für den logistischen Fall. Vorteile der approximativen korrigierten Score Funktion ist, dass die Approximation unabhängig von der Größe des Messfehlers ist (Buzas [2009, S.1]). Dennoch soll zunächst zur Veranschaulichung $l(), S()$ aufgestellt werden.

Sei $Y|X$ Bernoulliverteilt $Ber(p_i)$ mit $p_i = P(y_i = 1|x_i) = \frac{1}{1+\exp(-\underline{x}_i^t\beta)}$, mit Wahrscheinlichkeitsfunktion $f(y_i, x_i, \beta) = p_i^{y_i}(1-p_i)^{1-y_i}$. Gegeben die beobachteten Werte y_i und x_i ist die Log-Likelihood Funktion:

$$l(Y, X, \beta) = \sum_{i=1}^{n}(-\beta_0 - x_i\beta_1 - ln(1 + \exp(-\beta_0 - x_i\beta_1)) + y_i\beta_0 + y_i\beta_1 x_i) \tag{24}$$

und Score Funktionen:

$$\frac{\partial l}{\partial \beta_0} = S_0(Y, X, \beta) = \sum_{i=1}^{n} y_i - \frac{1}{1 + exp(-\beta_0 - \beta_1 x_i)} \tag{25}$$

$$\frac{\partial l}{\partial \beta_1} = S_1(Y, X, \beta) = \sum_{i=1}^{n} y_i x_i - \frac{x_i}{1 + exp(-\beta_0 - \beta_1 x_i)} \tag{26}$$

wobei die Verteilungsfunktion $F(v) = \frac{1}{1+\exp(-v)}$ ist.
Nach Buzas [2009, S.2352] lässt sich die Score auch darstellen als:

$$S(Y, X, \beta) = \underbrace{h(\beta_0 + \beta_1 X)}_{=1, für\ logistische\ Regression} (Y - F(\beta_0 + \beta_1 X)) \begin{pmatrix} 1 \\ X \end{pmatrix} \tag{27}$$

mit $h(v) = \frac{F'(v)}{F(v)(1-F(v))}$.

4.4.1 Approximative Score $S_A()$

Für die Aufstellung der Approximativen Likelihood Score $S_A()$ wird in Buzas [2009] die logistische $F(v)$ durch die Verteilungsfunktion der Standardnormalverteilung $\Phi(\lambda v)$ approximiert. Im Fall der Score Funktion für die logistische Regression erreicht man eine Approximation indem $F'(v)$ durch $\Phi'(\lambda v) = \lambda\Phi(\lambda v)$ bei der Berechnung von $h()$ ersetzt wird (Buzas [2009, S.2352]). Mit $\lambda = 1.7^{-1}$ wird der maximale Abstand zwischen den beiden Verteilungen minimiert (vgl. Buzas [2009, S.2352]) und Abbildung 1).
Alternativ kann λ selbst geschätzt werden (Buzas [2009, S.2532]). Der Wert von λ nimmt Einfluss auf die Varianz des resultierenden β-Schätzer, das aus dem approximativen Likelihood Score $S_A(Y, X, \beta) = 0$ berechnet werden kann. Im Rahmen dieser Arbeit wird in dem Simulationsteil $\lambda = 1.7^{-1}$ gesetzt.
Die approximative likelihood Score für die einfache logistische Regression ergibt sich

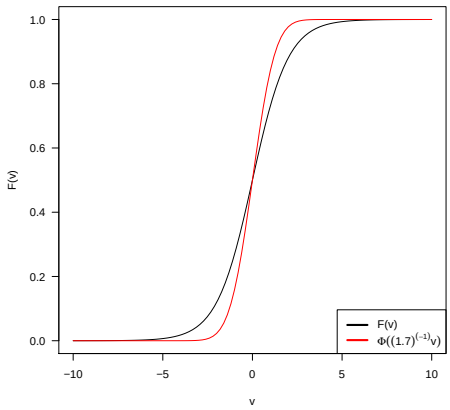

Abbildung 1: Vergleich der Verteilungsfunktion der logistischen Regression mit der Approximation durch die Standardnormalverteilung und $\lambda = 1.7^{-1}$

dann zu:

$$S_A(Y, X, \beta) = \frac{\lambda\Phi(\lambda(\beta_0 + \beta_1 X)}{F(\beta_0 + \beta_1 X)(1 - F(\beta_0 + \beta_1 X))}(Y - F(\beta_0 + \beta_1 X)) \begin{pmatrix} 1 \\ X \end{pmatrix} \tag{28}$$

wobei $F(v) = (1 + \exp(-v))^{-1}$ ist.
Beachtenswert ist, dass für den approximativen Likelihood Score für fehlerfrei gemessene X trotzdem $E[S_A(Y, X, \beta)|X] = 0$ gilt, falls das Ereignis $(y_i = 1)$ oft vorkommt (Buzas [2009, S.1]). Jedoch kann der approximative Likelihood Score nicht genauso effizient sein wie der Likelihood Score.

4.4.2 Korrigierter Score für approximativen Score $S_A()$

In diesem Kapitel stellen wir eine korrigierte Score Funktion $S_{Ac}()$ für den approximativen Score $S_A()$ der einfachen logistischen Regression auf, sodass gilt:

$$E[S_{Ac}(Y, X, \beta)|Y, X] = S_A(Y, X, \beta)$$

Dieser approximierter korrigierter Score eliminiert effektiv den Bias bei vorliegen von Messfehler (Buzas [2009, S.2]). Um solch einen Score zu konstruieren müssen zunächst folgende zwei Lemmas nach Buzas [2009, S.353 ff] eingeführt werden.

Lemma 1 (Buzas [2009, S.2353]):
Sei $k = \lambda\sqrt{1 - \beta_1^2\sigma_u^2\lambda^2}, \eta = \exp(-\frac{1}{2\lambda^2}k^2\beta_1^2\sigma_u^2)$ und $\gamma(X^*) = \exp(\frac{k^2}{\lambda}(\beta_0 + \beta_1 X^*))$. Definiere

- $g_1(X^*) = \frac{k}{\lambda}\Phi(k(\beta_0 + \beta_1 X^*))$
- $g_2(X^*) = \eta\gamma(\underline{X}^*)g_1(X^*)$
- $g_3(X^*) = g_2(X^*)/\gamma^2(X^*)$

vorausgesetzt $X^*|Y, X \sim N(X, \sigma_u^2)$ und, dass die definierten Funktionen folgende Eigenschaften haben:

- $E[g_1(X^*)|X] = \Phi(\lambda\beta_0 + \beta_1 X)$
- $E[g_2(X^*)|X] = \Phi(\lambda\beta_0 + \beta_1 X)\exp(\beta_0 + \beta_1 X)$
- $E[g_3(X^*)|X] = \Phi(\lambda\beta_0 + \beta_1 X)\exp(-(\beta_0 + \beta_1 X))$

Der Beweis hierzu lässt sich im Appendix in Buzas [2009] nachvollziehen.

Lemma 2 (Buzas [2009, S.2353]):
Sei $g_j(X^*), j = 1, 2, 3$ definiert wie in Lemma 1 und definiere
$g(X^*) = g_1(X^*)[2Y - 1] + g_2(X^*)[Y - 1] + g_3(X^*)Y$, dann gilt

$$E[g(X^*)|Y, X] = \tilde{h}(\beta_0 + \beta_1 X)[Y - F(\beta_0 + \beta_1 X)] \tag{29}$$

Einen ausführlichen Beweis für diese Gleichheit findet sich im Appendix C.2.1.

Die approximative korrigierte Likelihood Score ergibt sich aus folgender Proposition:

Proposition 1 (Buzas [2009, S.2354]): Sei $g(X^*)$ definiert wie in Lemma 2, dann definiere die approximative korrigierte Score als:

$$\begin{aligned} S_{AC}(Y, X^*, \beta) &= g(X^*)\begin{pmatrix} 1 \\ X^* - \sigma_u^2 \frac{g'(X^*)}{g(X^*)} \end{pmatrix} \\ &= \begin{pmatrix} g(X^*) \\ X^* g(X^*) - \sigma_u^2 g'(X^*) \end{pmatrix} \end{aligned} \tag{30}$$

wobei $g'(X^*) = \frac{\partial g(X^*)}{\partial X^*}$. Daraus folgt

$$E[S_{Ac}(Y, X^*, \beta)|Y, X] = S_A(Y, X, \beta) \tag{31}$$

Einen ausführlichen Beweis für diese Gleichheit findet sich im Appendix C.2.1.
Allerdings soll beachtet werden, dass die korrigierte approximative Score nach Lemma 1 nur angewendet werden kann, wenn $(1 - \beta_1^2\sigma_u^2\lambda^2) > 0$ gilt.

Ein typisches Problem bei Messfehlerkorrekturen ist, dass die Messfehlervarianz σ_u^2 bzw die Messfehlerkovarianz $\Sigma_{uu} = \begin{pmatrix} 0 & 0 \\ 0 & \sigma_u^2 \end{pmatrix}$ für die Korrektur benötigt wird. Daher sollte Σ_{uu} entweder bekannt sein oder geschätzt werden können. Die Möglichkeiten werden in Kapitel 4.6 erklärt.

4.5 Monte-Carlo korrigierte Score Funktionen

Der entscheidende Schritt ist es eine korrigierte Score Funktion zu finden, die Definition 1 genügt. Für einige Verteilungen ist eine exakte Lösung möglich (siehe Appendix C.1 und Caroll, R. J. and Ruppert, D. and Stefanski, L. A. and Crainiceanu, C. M. [2006, S.170ff]). Wenn jedoch keine exakte korrigierte Score Funktion aufgestellt werden kann, kann man nach Novick and Stefanski [2002] durch Monte Carlo Averaging diese alternativ konstruieren.

4.5.1 Algorithmus der Monte-Carlo korrigierten Score Funktion

Der Algorithmus zur Konstruktion eines Monte Carlo korrigierten Score benötigt die Verwendung von komplexen Zahlen. Caroll, R. J. and Ruppert, D. and Stefanski, L. A. and Crainiceanu, C. M. [2006, S.170] hat einen Algorithmus aufgestellt, der im folgenden für eine fehlerhafte Messvariable X vorgestellt wird.

- Generiere für $b = 1, ..., B$ zufällige Zahlen $u_{b,i}$, die normalverteilt $N(0, \sigma_u^2)$ sind.
- Konstruiere komplexe Zufallszahlen

$$\tilde{x}_{b,i}^* = x_i^* + \underbrace{\iota}_{\sqrt{-1}} u_{b,i} \tag{32}$$

  d.h. $\tilde{X}_b^* = (\tilde{x}_{b,1}, ..., \tilde{x}_{b,n})^t$ bzw. $\tilde{X}^* = \begin{pmatrix} \tilde{x}_{1,1} & ... & \tilde{x}_{b,1} \\ \cdot & & \\ \cdot & & \\ \tilde{x}_{1,n} & ... & \tilde{x}_{b,n} \end{pmatrix}$

- Die Monte Carlo korrigierte Score ergibt sich zu

$$S_{MCc,B}(Y, \tilde{X}^*, \beta) = B^{-1} \sum_{b=1}^{B} Re\{S(Y, \tilde{X}_b^*, \beta)\} \tag{33}$$

  wobei $Re(C)$ der Realteil der komplexen Zahl C ist.

- Es gilt

$$\lim_{B\to\infty} S_{MCc,B}(Y, \tilde{X}^*, \beta) \to S_c(Y, X^*, \beta) \tag{34}$$

  die Anzahl B der generierten Zufallszahlen in (32) sollte groß genug sein um einen approximativ korrekten Limes zu erhalten. Jedoch genügen in vielen Fällen auch kleinere Werte für B. (Caroll, R. J. and Ruppert, D. and Stefanski, L. A. and Crainiceanu, C. M. [2006, S.171] und Simulationsergebniss Kapitel 5.3.2)

- Die dadurch erhaltenen Schätzer sind M-Schätzer. M-Schätzer ist eine Klasse von Schätzfunktionen, die als Verallgemeinerung der Maximum-Likelihood-Methode angesehen werden können (siehe z.B.Czado and Schmidt [2011]).

Die Monte-Carlo korrigierte Scorefunktion konstruiert zunächst komplexe Zufallszahlen und benutzt anschließend wiederrum nur den Realteil. Was genau die Konstruktion von komplexen Zufallszahlen bewirkt soll in den folgenden beiden Kapitel am Beispiel von der linearen und Poisson Regression dargestellt werden. Da die logistische korrigierte Score Funktion nicht existiert und die Monte-Carlo korrigierte Score eine alternative Berechnung für die korrigierte Score darstellt, wird die Monte-Carlo korrigierte Score Funktion nicht aufgestellt und im Simulationsteil 5 nicht zum Vergleich mit anderen Methoden betrachtet. Kapitel 5 verdeutlicht, dass die Monte-Carlo korrigierte Score für logistische Regression keine bessere Schätzung als die naive Schätzung liefert.

4.5.2 Einfaches lineares Modell

Zunächst wird in die wahre Score Funktion ((55), (56)) $\tilde{X}_b$ eingesetzt:

$$S_0(Y, \tilde{X}_b, \beta) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i^* - \sqrt{-1}\beta_1 u_{b,i}) \tag{35}$$

$$S_1(Y, \tilde{Y}_b, \beta) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i x_i^* - x_i^*\beta_0 - x_i^{*2}\beta_1 + u_{b,i}^2\beta_1 - \sqrt{-1}u_{b,i}(x_i^*\beta_1 - y_i + \beta_0 + \beta_1 x_i^*) \tag{36}$$

der Realteil ist jeweils

$$Re(S_0(Y, \tilde{X}_b, \beta)) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i^*)$$

$$Re(S_1(Y, \tilde{X}_b, \beta)) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i x_i^* - x_i\beta_0 - x_i^{*2}\beta_1 + u_{b,i}^2\beta_1)$$

und die Summe über B bzw. die korrigierte Monte Carlo Scores lassen sich vereinfachen zu

$$S_{MCc0,B}(Y, \tilde{X}^*, \beta) = \frac{1}{B} \sum_{b=1}^{B} Re(S_0(Y, \tilde{X}_b, \beta)) = \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i^*) \tag{37}$$

und

$$\begin{aligned} S_{MCc1,B}(Y, \tilde{X}^*, \beta) &= \frac{1}{B} \sum_{b=1}^{B} Re(S_1(Y, \tilde{X}_b, \beta)) \\ &= \underbrace{\frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i x_i^* - \beta_0 x_i^* - x_i^{*2} \beta_1)}_{S_1(y_i, x_i^*, \beta)} + \frac{1}{\sigma^2} \sum_{i=1}^{n} \beta_1 \frac{1}{B} \sum_{b=1}^{B} u_{b,i}^2 \end{aligned} \tag{38}$$

Die Monte-Carlo korrigierte Scorefunktion für die einfache lineare Regression ergibt sich aus der Summe der Scorefunktion für die fehlerhafte Messung X^* und einen Korrekturterm, der vom arithmetischen Mittel über $u_{b,i}^2$ abhängt. Die Größe des Wertes von B bestimmt somit die Anzahl der Summanden des arithmetischen Mittels. Je höher B gewählt desto stabiler das arithmetische Mittel.
Im Fall der linearen Regression kann man verdeutlichen, dass die Monte-Carlo korrigierte Score Funktion tatsächlich eine alternative Berechnung der korrigierten Score Funktion darstellt. Es gilt nämlich:

$$E[\frac{1}{\sigma^2} \sum_{i=1}^{n} \beta_1 \frac{1}{B} \sum_{b=1}^{B} u_{b,i}^2 | y_i, x_i^*] = \frac{1}{\sigma^2} \sum_{i=1}^{n} \beta_1 \frac{1}{B} \sum_{b=1}^{B} E[u_{b,i}^2 | y_i, x_i^*] \underset{(58)}{=} \frac{1}{\sigma^2} n \beta_1 \sigma_u^2$$

sodass folgt (vgl.Caroll, R. J. and Ruppert, D. and Stefanski, L. A. and Crainiceanu, C. M. [2006, S.157]):

$$E[S_{MCc0,B}(Y, \tilde{X}^*, \beta) | y_i, x_i^*] = E[\frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i^*) | y_i, x_i^*] = E[S_{c0}(Y, \tilde{X}^*, \beta) | y_i, x_i^*] = S_{c0}(Y, \tilde{X}^*, \beta)$$

und

$$\begin{aligned} &E[S_{MCc1,B}(Y, \tilde{X}^*, \beta) | y_i, x_i^*] = E[S_1(Y, \tilde{X}^*, \beta) + \frac{1}{\sigma^2} \sum_{i=1}^{n} \beta_1 \frac{1}{B} \sum_{b=1}^{B} u_{b,i}^2 | y_i, x_i^*] \\ &= E[S_1(Y, \tilde{X}^*, \beta) + \frac{1}{\sigma^2} n \beta_1 \sigma_u^2 | y_i, x_i^*] \underset{(19)}{=} S_{c1}(Y, \tilde{X}^*, \beta) \end{aligned}$$

und aus Formel (12) folgt, dass die Monte-Carlo-korrigierte Score Funktion ebenfalls konsistent ist.
(Zwischenschritte und Herleitung im Appendix C.3.1).)

4.5.3 Einfache Poisson Modell

Zunächst wird in die wahre Score Funktionen ((66), (67)) $\tilde{x}_{b,i}$ eingesetzt:

$$\begin{aligned} S_0(Y, \tilde{X}_b, \beta) &= -\sum_{i=1}^{n}(\exp(\beta_0 + \beta_1 x_i^*)(cos(\beta_1 u_{b,i}) \\ &\quad + \sqrt{-1}\exp(\beta_0 + \beta_1 x_i^*)sin(\beta_1 u_{b,i}) - y_i) \qquad (39) \\ S_1(Y, \tilde{X}_b, \beta) &= -\sum_{i=1}^{n}(\exp(\beta_0 + \beta_1 x_i^*)(cos(\beta_1 u_{b,i})x_i^* - sin(\beta_1 u_{b,i})u_{b,i}) - y_i x_i^* \\ &\quad + \sqrt{-1}(\exp(\beta_0 + \beta_1 x_i^*)(sin(\beta_1 u_{b,i})x_i^* + cos(\beta_1 u_{b,i})u_{b,i}) - y_i u_{b,i})) \qquad (40) \end{aligned}$$

Der Realteil ist jeweils

$$\begin{aligned} Re(S_0(Y, \tilde{X}_b, \beta)) &= -\sum_{i=1}^{n}(\exp(\beta_0 + \beta_1 x_i^*)(cos(\beta_1 u_{b,i}) - y_i) \\ Re(S_1(Y, \tilde{X}_b, \beta)) &= -\sum_{i=1}^{n}(\exp(\beta_0 + \beta_1 x_i^*)(cos(\beta_1 u_{b,i})x_i^* - sin(\beta_1 u_{b,i})u_{b,i}) - y_i x_i^*) \end{aligned}$$

und die Summe über B bzw. die korrigierte Monte Carlo Scores lassen sich vereinfachen zu

$$\begin{aligned} S_{MCc0,B}(Y, \tilde{X}^*, \beta) &= \frac{1}{B}\sum_{b=1}^{B} Re(S_0(Y, \tilde{X}_b, \beta)) \\ &= -\sum_{i=1}^{n}(\exp(\beta_0 + \beta_1 x_i^*)(\frac{1}{B}\sum_{b=1}^{B} cos(\beta_1 u_{b,i})) - yi) \qquad (41) \\ S_{MCc1,B}(Y, \tilde{X}^*, \beta) &= \frac{1}{B}\sum_{b=1}^{B} Re(S_1(Y, \tilde{X}_b, \beta)) \\ &= -\sum_{i=1}^{n}(\exp(\beta_0 + \beta_1 x_i^*)\frac{1}{B}\sum_{b=1}^{B}(cos(\beta_1 u_{b,i})x_i^* - sin(\beta_1 u_{b,i})u_{b,i}) - y_i x_i^*) \qquad (42) \end{aligned}$$

Die Monte-Carlo korrigierte Scorefunktion für die einfache Poisson Regression entspricht der Scorefunktion für die fehlerhafte Messung X^*, nur dass zusätzlich der Summand, der unabhängig von Y ist durch einen Korrekturfaktor, das abhängig von $U_{b,i}$ ist multipliziert wird. Auch im Fall der Poisson Regression dient die Größe B zur Stabilisierung des arithmetischen Mittels des Summanden, die abhängig von $u_{b,i}$ ist.
(Zwischenschritte und Herleitung im Appendix C.3.2))

4.5.4 Logistische Regression

In Kapitel 4.4 wurde bereits geklärt, dass für die logistische Regression keine korrigierte Score existiert und somit auch nicht die Monte-Carlo korrigierte Score Funktion, da diese eine Alternative zur Konstruktion korrigierte Scorefunktionen ist. Denoch kann man nach Caroll, R. J. and Ruppert, D. and Stefanski, L. A. and Crainiceanu, C. M. [2006, S.158] trotzdem eine korrigierte Score mit der Monte-Carlo-Algorithmus herleiten und anwenden. Solange die Messfehlervarianz gering ist führt diese zu näherungsweise konsistente Schätzer, d.h. der Messfehler Bias wird reduziert, aber nicht beseitigt. Ein Nachteil, dass bei der approximativen Score nicht besteht (vgl. 4.4). Aus diesem Grund gilt Formel 12 nur approximativ. Wobei in der Umsetzung der resultierende Bias vernachlässigbar ist (Caroll, R. J. and Ruppert, D. and Stefanski, L. A. and Crainiceanu, C. M. [2006, S.158]).
Zunächst wird in die wahre Score Funktionen ((66), (67)) $\tilde{x}_{b,i}$ eingesetzt:

$$\begin{aligned} S_0(Y, \tilde{X}_b, \beta) = &- \sum_{i=1}^{n} (\exp(\beta_0 + \beta_1 x_i^*)(cos(\beta_1 u_{b,i}) \\ &+ \sqrt{-1} \exp(\beta_0 + \beta_1 x_i^*) sin(\beta_1 u_{b,i}) - y_i) \end{aligned} \tag{43}$$

$$\begin{aligned} S_1(Y, \tilde{X}_b, \beta) = &- \sum_{i=1}^{n} (\exp(\beta_0 + \beta_1 x_i^*)(cos(\beta_1 u_{b,i}) x_i^* - sin(\beta_1 u_{b,i}) u_{b,i}) - y_i x_i^* \\ &+ \sqrt{-1} (\exp(\beta_0 + \beta_1 x_i^*)(sin(\beta_1 u_{b,i}) x_i^* + cos(\beta_1 u_{b,i}) u_{b,i}) - y_i u_{b,i})) \end{aligned} \tag{44}$$

Der Realteil ist jeweils

$$Re(S_0(Y, \tilde{X}_b, \beta)) = - \sum_{i=1}^{n} (\exp(\beta_0 + \beta_1 x_i^*)(cos(\beta_1 u_{b,i}) - y_i)$$

$$Re(S_1(Y, \tilde{X}_b, \beta)) = - \sum_{i=1}^{n} (\exp(\beta_0 + \beta_1 x_i^*)(cos(\beta_1 u_{b,i}) x_i^* - sin(\beta_1 u_{b,i}) u_{b,i}) - y_i x_i^*)$$

und die Summe über B bzw. die korrigierte Monte Carlo Scores lassen sich vereinfachen zu

$$S_{MCc0,B}(Y, \tilde{X}^*, \beta) = - \sum_{i=1}^{n} (\exp(\beta_0 + \beta_1 x_i^*)(\frac{1}{B} \sum_{b=1}^{B} cos(\beta_1 u_{b,i})) - yi) \tag{45}$$

$$S_{MCc1,B}(Y, \tilde{X}^*, \beta) = - \sum_{i=1}^{n} (\exp(\beta_0 + \beta_1 x_i^*) \frac{1}{B} \sum_{b=1}^{B} (cos(\beta_1 u_{b,i}) x_i^* - sin(\beta_1 u_{b,i}) u_{b,i}) - y_i x_i^*) \tag{46}$$

Die Monte-Carlo korrigierte Scorefunktion für die einfache logistische Regression entspricht der Scorefunktion für die fehlerhafte Messung X^*, nur dass zusätzlich der Summand, der unabhängig von Y ist durch einen Korrekturfaktor, das abhängig von $U_{b,i}$

ist multipliziert wird. Auch im Fall der logistischen Regression dient die Größe B zur Stabilisierung des arithmetischen Mittels des Summanden, die abhängig von $u_{b,i}$ ist. (Zwischenschritte und Herleitung im Appendix C.3.3))

4.6 Schätzung nach verschiedenen Datentypen

Wie in der Regressionskalibrierung unterscheidet man auch bei der korrigierten Score Funktion und Monte-Carlos korrigierten Score Funktion zwischen verschiedenen Datentypen, unter anderem um die Fehlervarianz σ_{uu} zu schätzen. Die Fehlervarianz kann entweder durch externe Daten vorliegen oder muss geschätzt werden.

interne Validierungsdaten Wenn interne Validierungsdaten vorliegen, kann die Fehlervarianz σ_u^2 durch

$$\hat{\sigma}_u^2 = \frac{1}{n_v}\sum_{i=1}^{n_v}((x_i^* - x_i) - \overline{(x^* - x)})^t((x_i^* - x_i) - \overline{(x^* - x)}) = \frac{1}{n_v}\sum_{i=1}^{n_v}(u_i - \bar{u})^2 \qquad (47)$$

geschätzt werden, wobei nur die Beobachtungen berücksichtigt werden, für die die wahren x Messungen vorliegen und n_v gibt die Anzahl der Validierungsmessungen angibt.
Die korrigierte Score Funktion besteht in diesem Fall aus zwei Summanden,

$$\tilde{S}_c(Y, X, X^*, \beta) = S(Y, X, \beta) + S_c(Y, X^*, \beta)$$

d.h aus der wahren Score Funktion basierend auf den fehlerfrei gemessenen X und aus der korrigierten Score Funktion basierend auf fehlerbehaftete Messung X^*.

Wiederholungsdaten Bei vorliegen von Messwiederholungen wird zur Schätzung der Fehlervarianz die gleiche Schätzformel (1) wie in der Regressionakalibrierung benutzt (Caroll, R. J. and Ruppert, D. and Stefanski, L. A. and Crainiceanu, C. M. [2006, S.176]):

$$\widehat{\Sigma}_{uu} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{k_i}\left(X_{ij}^* - \overline{X_{i\cdot}^*}\right)\left(X_{ij}^* - \overline{X_{i\cdot}^*}\right)^t}{\sum_{i=1}^{n}(k_i - 1)}.$$

Nach Caroll, R. J. and Ruppert, D. and Stefanski, L. A. and Crainiceanu, C. M. [2006, S.176] sollte die Σ_{uu} bei der Schätzung durch Σ_{uu}/k_i bzw. $\hat{\Sigma}_{uu}/k_i$ ersetzt werden, wobei k_i die Anzahl der Messwiederholung der i-ten Beobachtung angibt. Es gilt wieder σ_u^2 wird geschätzt, wenn X eindimensional ist und Σ_{uu}, wenn X eine Matrix ist, mit jeder Spalte gleich eine Variable. Für den Fall, dass nur eine Messwiederholung vorliegt muss die Fehlervarianz aus externen Daten verwendet werden.

5 Simulationen

In der Simulationsstudie werden folgende Punkte untersucht:

- S1 Verhalten der korrigierten Schätzmethoden bei unterschiedlicher Größe der Messfehlervarianz.
- S2 Einen Vergleich der Monte-Carlo korrigierten Score Funktion mit verschiedenen Werten für B im Algorithmus.
- S3 Allgemeiner Vergleich der Schätzmethoden.
- S4 Das Verhalten der Schätzmethoden bei unterschiedlichem Anteil an Validierungsdaten.
- S5 Das Verhalten der Schätzmethoden bei unterschiedliche Anzahl an Messwiederholungen und Beobachtungen.
- S6 Das Verhalten der Schätzmethoden bei einer Messwiederholung und verschiedene Werte der übergebenen geschätzten Messfehlervarianz $\hat{\sigma}^2_{uu}$ aus externen Daten.
- S7 Das Verhalten bei Annahmeverletzungen:
  - AV1 Fehler U nicht rein additiv
  - AV2 Fehler U nicht normalverteilt
  - AV3 Fehler U nicht unabhängig von X
  - AV4 Fehler U unterschiedlich schief normalverteilt

Die Simulationsstudie beschränkt sich auf die Modellen und Methoden, die im Rahmen dieser Arbeit betrachten wurden. Um den Umfang der Arbeit einzuschränken, werden im Folgenden nur die β_1-Schätzer verglichen.

5.1 Datensätze

Für die Simulationsstudie wurden Datensätze bzw. Datensets, jeweils bestehend aus mindestens 5400 Datensätze generiert. Dabei können die Datensätze folgende Eigenschaften besitzen:

- E1 Regressionstyp: Lineare-, Logistische-, Poisson Regression
- E2 Anzahl Beobachtungen n: 20, 40, 60, 80, 100, 150, 500
- E3 Datentyp: Wiederholungsdaten\Validierungsdaten
- E4 Anteil an Validierungsdaten: 10%, 15%, 20%, 25%, 30%, 40%, 50%
- E5 Anzahl Messwiederholung: 1, 2, 4, 6, 8, 10, 15, 20

Die Anzahl der Datensätze wurden so bestimmt, dass jede Kombinationsmöglichkeit der Eigenschaften circa 100 mal vorkommt. Für die Regressionsmodelle wurden für die Parameter in allen Datensätzen folgende Werte festgelegt:

- Linearen Regression: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$
  mit $\beta_0 = 1, \beta_1 = 3$ und $\epsilon_i \sim N(0, 0.3^2)$ und $x_i \sim N(0, 1)$
- Logistischen Regression: $P(y_i = 1|x_i) = \frac{\exp(\beta_0+\beta_1 x_i)}{1+\exp(\beta_0+\beta_1 x_i)}$
  mit $\beta_0 = 0.5, \beta_1 = 1$
  $y_i \sim Ber(p_i)$, $p_i = P(y_i = 1|x_i)$
- Poisson Regression: $\mu(x_i) = \exp(\beta_0 + \beta_1 x_i)$
  mit $\beta_0 = 1, \beta_1 = 0.6$
  $y_i \sim P_\lambda$, $\lambda_i = \mu_i(x_i)$

Vor der Datengeneration werden vorher per Zufall die Kombination der Eigenschaften für jeden Datensatz festgelegt und abgespeichert. Für die Simulation der Daten werden die Eigenschaften an die datengenerierende Funktion übergeben. Dieser Aufbau ermöglicht es, zu jedem Datensatz die Merkmalsausprägungen genau erfassen zu können. Für diese Arbeit wurde drei Kombinationslisten erstellt.

- Kombinationsliste 1: Hat die Länge 5400 und Eigenschaften E1 bis E5, wobei E4 auf 1, 4 und 20 und E5 auf 15%, 25% und 50% und E2 auf 20, 100 und 500 eingeschränkt wurde.
- Kombinationsliste 2: Hat die Länge 16000 und Eigenschaften E1 bis E4.
- Kombinationsliste 3: Hat die Länge 16000 und Eigenschaften E1 bis E3 und E5.

Einen Überblick über die Verteilung der Eigenschaften sind in Tabellen 1 bis 8 dargestellt:

Tabelle 1: Kombinationsliste 1: Verteilung der Eigenschaften Regressionstyp und Anzahl Beobachtungen

Regressionstyp			Anzahl Beobachtungen n		
lin. Reg	log. Reg	Pois. Reg	n=20	n=100	n=500
1826	1731	1843	1766	1754	1880

Tabelle 2: Kombinationsliste 1: Verteilung der Eigenschaften Datentyp

Wiederholungsdaten			Validierungsdaten		
2769			2631		
Anzahl Wdh-Messungen			Anteil Valid-Messungen		
1 Wdh	4 Wdh	20 Wdh	15%	25%	50%
918	955	896	922	850	859

Tabelle 3: Kombinationsliste 2: Verteilung der Regressionstypen

lin. Reg	log. Reg	Pois. Reg
5680	5632	5488

Tabelle 4: Kombinationsliste 2: Verteilung der Anzahl der Beobachtungen

n=20	n=40	n=60	n=80	n=100	n=150	n=200
2134	2095	2095	2118	2077	2091	2147

Tabelle 5: Kombinationsliste 2: Verteilung der Anteil der Validierungsdaten

10%	15%	20%	25%	30%	40%	50%
2466	2331	2389	2391	2361	2401	2461

Tabelle 6: Kombinationsliste 3: Verteilung der Regressionstypen

lin. Reg	log. Reg	Pois. Reg
5602	5609	5589

Tabelle 7: Kombinationsliste 3: Verteilung der Anzahl der Beobachtungen

n=20	n=40	n=60	n=80	n=100	n=150	n=200
2117	2142	2111	2140	2131	2067	2075

Tabelle 8: Kombinationsliste 3: Verteilung der Anzahl der Messwiederholung

2 Wdh	4 Wdh	6 Wdh	8 Wdh	10 Wdh	15 Wdh	20 Wdh
2387	2397	2403	2410	2399	2424	2386

Jeder Datensatz enthält mindestens eine Variable mit fehlerbehaftete Messungen X^* von X, d.h.

$$X^* = X + U \text{ mit } U \sim N(0, 0.3^2) \tag{48}$$

und die wahren X Werte, die in den Datensätzen mit $x.true$ gekennzeichnet werden. Diese dienen dem Vergleich und liegen in realen Studien normalerweise nicht vor.
Alle Datensätze können folgende Formen haben, siehe Tabelle 9:

Tabelle 9: Beispiel Validierungsdaten, x enthält teilweise Validierungsmessungen. Wenn Validierungsmessungen vorliegen, dann hat Valid den Wert 1, sonst 0

y	x.true	x	xSt	Valid
-6.77	-2.41	NA	-2.59	0
-3.98	-1.69	NA	-1.16	0
-3.89	-1.57	-1.57	-1.65	1
...	...	...	...	...

Tabelle 10: Beispiel Wiederholungsdaten mit k Messwiederholung $(xSt1, ..., xSt_k)$

y	x.true	xSt1	...	xStk
0.45	-0.47	0.26	...	-0.70
3.32	0.76	1.08	...	1.31
-2.97	-1.10	-1.27	...	-1.13
...	...	...	...	...

Daten-Ziehungstypen: Für die Generierung der Daten gibt es zwei unterschiedliche Vorgehensweisen.

- Beim *Ziehungstyp 1* werden für jeden Datensatz die Variablenwerte neu gezogen, d.h. standardnormalverteilte X-Werte und je nach übergebenen Eigenschaften eine oder mehrere Wiederholungsmessungen bzw. Validierungsmessungen. Diese Vorgehensweise erzeugt unabhängige Datensätze, mit diesen Frage S1, S2, S5, S6 und S7 untersucht werden können.

- Beim *Ziehungstyp 2* werden vorausgehend drei Basisdatensätze simuliert aus je 5000 Beobachtungen, standardnormalverteilten X-Werte und 60 Wiederholungsmessungen (fehlerhafte Messung von X). Die drei Basisdatensätze unterscheiden sich durch den Regressionstyp. Ein Basisdatensatz soll die Grundgesamtheit darstellen, in realen Studien können nicht die Grundgesamtheit erhoben werden sondern immer nur Teile davon. Daher wird bei diesem Ziehungstypen Teildatensätze aus den Basisdatensätzen gezogen. Es stellt sich die Frage, wie am sinnvollsten Teildatensätze gezogen werden sollen um gute Schätzer zu erhalten, wie in Frage S3 und S4.

Für den Fall von Wiederholungsdaten mit einer Messung, ist für die Anwendung der Methoden die Varianz des Fehlers aus externen Daten nötig. Um künstlich eine externe Fehlervarianz zu erzeugen, wurde jeweils aus 1/3 der wahren Daten, die immer zusätzlich als $x.true$ in den Datensätzen enthalten sind, die Fehlervarianz geschätzt, nach

$$\hat{\sigma}_u^2 = \frac{1}{n_v}\sum_{i=1}^{n_v}((x_i^* - x_i) - \overline{(x^* - x)})^2 = \frac{1}{n_v}\sum_{i=1}^{n_v}(u_i - \overline{u})^2 \tag{49}$$

und als externes Wissen angenommen.

5.2 Kennzahlen

Um die Schätzmethoden zu vergleichen wird die relative Bias berechnet,

$$Bias_{\beta_1} = \frac{\hat{\beta}_1 - \hat{\beta}_{1bench}}{\hat{\beta}_{1bench}} \tag{50}$$

Wobei $\hat{\beta}_1$ der Schätzung aus der Messfehlerkorrekturmethode entstammt und $\hat{\beta}_{1bench}$ die Schätzung anhand der vorliegenden fehlerfreien X ($x.true$) entspricht. Analog $Bias_0$ für β_0. Diese Kennzahl ist zum einen sinnvoll, da die Schätzung anhand der wahren Daten möglicherweise nicht die gesetzten β-Parameter für die Simulation entspricht, daher wird als Bezugsparameter $\hat{\beta}_{1bench}$ herangezogen. Und zum anderen, da für unterschiedliche Regressionen die βs unterschiedlich groß gewählt worden sind (siehe Kapitel 5.1) und dadurch die Abweichung vom Benchmark-Parameter auch unterschiedliche Wertebereiche haben, durch Division durch $\hat{\beta}_{1bench}$ wird diesem Problem entgegengewirkt. Ein negativer relativer Bias bedeutet einen Unterschätzung, ein positiver Wert Überschätzung des Schätzers im Vergleich zum Benchmark Schätzer. Einen relativen Bias von Null bedeutet, dass die Schätzung dem Benchmark Schätzer entstpricht und somit erwartungstreu ist.
Eine weitere Kennzahl ist der Median des relativen Bias oder der Parameter Schätzer. Es wurde der Median gewählt, da dieser robuster als der Mittelwert gegen Ausreißer ist. Durch diese Kennzahl können die Schätzergebnisse aus mehreren Datensätzen zusammengefasst werden.

Auch wurde hauptsächlich Boxplots über die relativen Bias zur Darstellung verwendet, da diese die Verteilung des relativen Bias am anschaulichsten darstellt.
In den Grafiken sind die Wertebereich der Y-Achsen zur besseren Lesbarkeit der Grafiken eingeschränkt, da einige wenige Ausreißer vorliegen die in der Nähe des Tausenderbereichs liegen. Eine Abbildung der kompletten Y-Achse schränkt die Anschaulichkeit ein. Die Ausreißer sind wohl meist darauf zurückzuführen, dass in der korrigierten Score Funktion und der Monte-Carlo korrigierten Score Funktion mit dem Befehl “multiroot()“ die Nullstellen der Score Funktion gesucht worden sind (vgl. Appendix D), dieser Befehl konvergiert zum einen nicht immer und zum anderen muss diese auch nicht unbedingt gegen die Nullstelle konvergieren (roo, S.15). Außerdem benutzt “multiroot()“ die Newton Raphson Methode (Kapitel B.2), die einen vollen Rang der Jacobi Matrix annimmt, das in manchen Datensätzen nicht gegeben ist.

5.3 Ergebnisse

5.3.1 Unterschiedlicher Messfehlervarianz

Bevor weitere Datensätze und Analysen berechnet werden, wurde das Verhalten der Methoden bei unterschiedlicher Größe der Messfehlervarianz untersucht. Generiert wurden fünf Datensets mit je 5400 Datensätzen mit Daten-Ziehungstyp 1 und Kombinationsliste 1. Die Datensets unterscheiden sich in der Messfehlervarianz bzw. Standardabweichung:

- $\sigma \in \{0.1, 0.3, 0.5, 0.7, 1\}$

In Abbildung 2 deutlich zu erkennen ist, dass die Methoden bei steigender Messfehlervarianz mehr Unsicherheiten bei der Schätzung aufweisen. Dennoch verhalten sie sich robuster als in der naiven Regression. Wie bereits in Kapitel3.2 erwähnt, reduziert die Regressionskalibrierung den Bias. Diese Eigenschaft kann ebenfalls aus Abbildung 2 abgelesen werden, während der Median der relativen Bias für $\hat{\beta}_1$ aus der korrigierten Score Funktion unabhängig von der Standardabweichung bei Null liegt (siehe auch Kapitel 4), wird bei der Regressionskalibrierung die Schätzer ab einer Standardabweichung von 0.5 tendenziell eher unterschätzt. Bei der approximativ korrigierten Scorefunktion bei der logistischen Regression zeigt sich, anders als sonst, jedoch keine Mediantreue. Bemerkenswert ist, dass die Monte-Carlo korrigierten Score Funktion für die logistische besser zu funktionieren scheint, als die korrigierte Score, obwohl nach Kapitel 4.5.4 keine korrigierte Score Funktion für den logistischen Fall existiert.

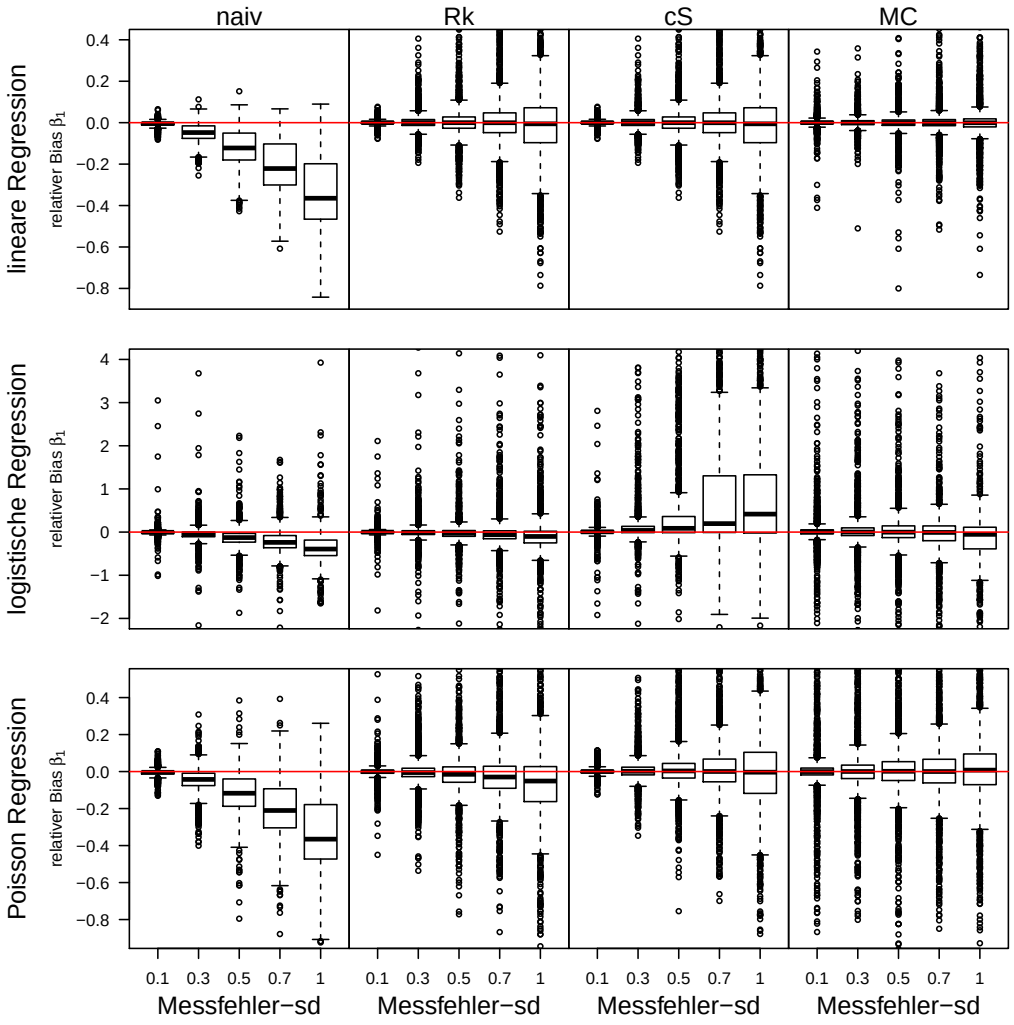

Abbildung 2: Vergleich der Methoden nach unterschiedlichen Regressionsmodellen bei unterschiedlicher Messfehlerstandardabweichung. Die Y-Achsen sind eingeschränkt.

Für weitere Analysen wird für die Simulation von Datensätzen die Standardabweichung des Messfehlers auf 0.3 gesetzt, da der relative Bias bei Regressionskalibrierung ab $\sigma = 0.5$ im Median nicht mehr Null ist. Außerdem gilt für die Anwendung der Monte-Carlo korrigierten Score Funktion, dass $\beta_x < \frac{1}{\sigma_u (1.7)^{-1}}$ gelten sollte (vgl. Kapitel 4.4). Nach Tabelle 11 und den bereits festgelegten β_1 Werten der Regression in Kapitel 5.1, in der $\beta_1 = 3$ für die lineare Regression gewählt wurde, ist $\sigma_u \geq 0.7$ ohnehin ausgeschlossen. Da der Wert 3.4 für $\sigma_u = 0.5$ dicht bei 3 liegt, ist $\sigma_u = 0.3$ zu bevorzugen.

Tabelle 11: Mögliche Werte für β_1 in Abhängigkeit von σ_u für die Anwendung der approximierten Score Funktion im logistischen Fall.

σ_u	0.1	0.3	0.5	0.7	1
$\beta_1 <$	17	5,7	3,4	2,4	1,7

5.3.2 Schätzung durch Monte-Carlo korrigierte Score und unterschiedliche Werte von B

Wie in Kapitel 4.5 beschrieben, wird zur Konstruktion des Monte-Carlo korrigierten Scores B Zufallszahlen $u_{b,i}$ aus $N(0, \sigma_u^2)$ gezogen. In diesem Kapitel wird untersucht inwiefern die Anzahl der gezogenen Zufallszahlen $u_{b,i}$ Einfluss auf die Schätzung nimmt. Hierfür wurde die Monte-Carlo korrigierte Score fünfmal auf den gleichen Datensatz angewendet, mit je B als 1, 10, 50 und 100 gewählt. Nach den Boxplots der relativen Bias in Abbildung 3 sind keine Unterschiede zu erkennen.

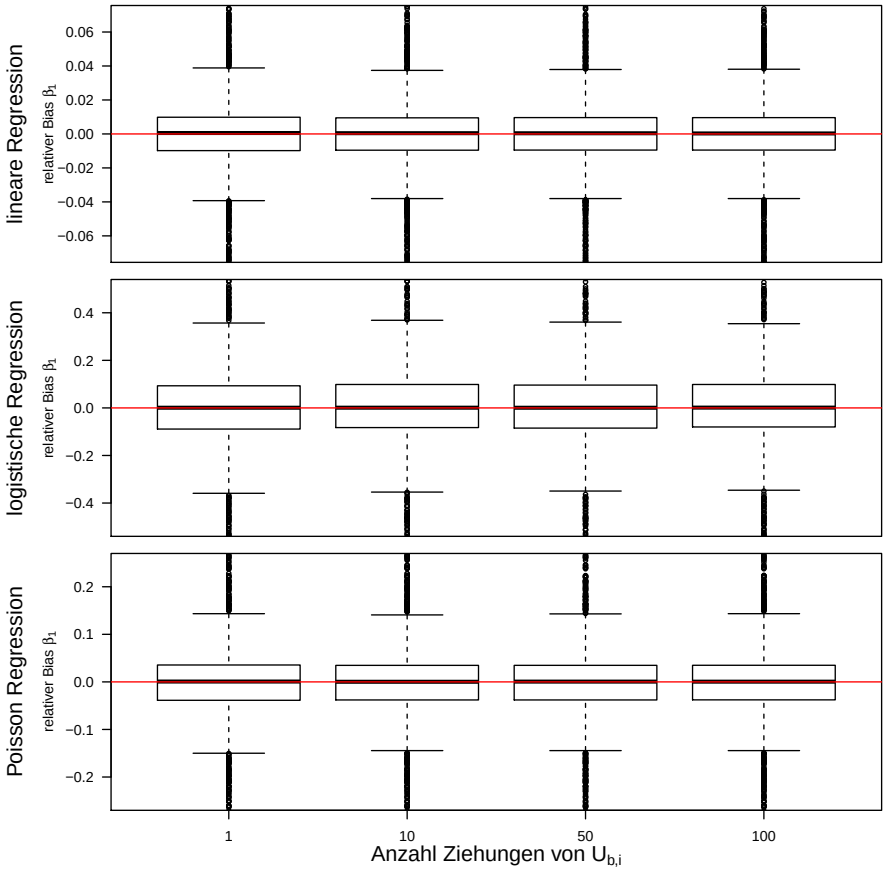

Abbildung 3: Vergleich der Schätzung des Monte-Carlo korrigierten Scores mit verschiedenen Werten für B. Ausreißer wurden in der Abbildung abgeschnitten.

Wie in Kapitel 4.5 beschrieben genügt bereits ein kleiner Werte für B aus, damit die Monte-Carlo korrigierte Score Funktion approximativ der korrigierten Score Funktion entspricht. Es wurde eine weitere Analyse durchgeführt um näheres über die Wirkung von B zu erfahren. Wie in Kapiteln 4.5.2 und 4.5.3 beschrieben, dient die Größe B dazu, dass das arithmetische Mittel, das abhängig von $u_{b,i}$ ist zu stabilisieren. Außerdem ist $u_{b,i}$ abhängig von der Messfehlervarianz. Aus diesem Grund wurde in einem weiteren Schritt eine große externe Messfehlervarianz $4\sigma_u^2$ übergeben, um größere Schwankungen für $u_{b,i}$ zu erreichen und schließlich mit B gleich 1, 1000 geschätzt. Abbildung 4 verdeutlicht, dass auch in diesem Experiment minimale Unterschiede zu erkennen sind. Der Interquartilsabstand und die Whisker werden betragsmäßig minimal kleiner, am deutlichsten noch zu erkennen an der logistischen Regression. Daher wurde im Weiteren für die restlichen Analysen B=1 gewählt um die Laufzeit zu verkürzen.

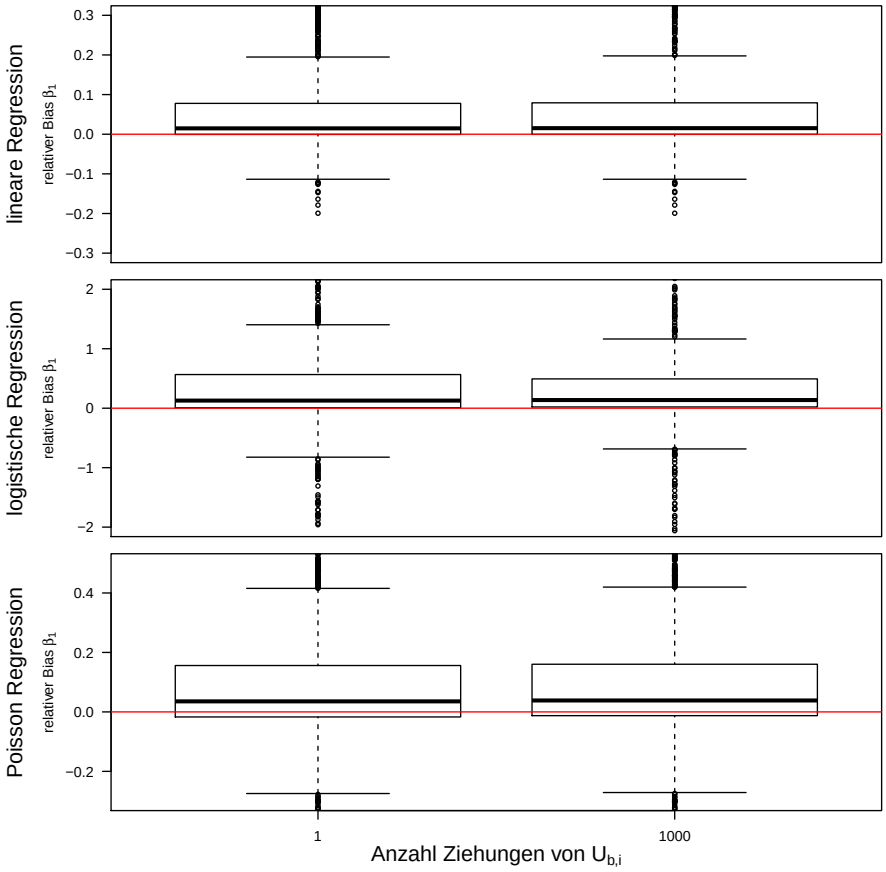

Abbildung 4: Vergleich der Schätzung des Monte-Carlo korrigierten Scores mit verschiedenen Werten für B und großer geschätzter Messfehlervarianz ($\hat{\sigma}_u^2 = 4 * 0.3^2$) aus externen Daten. Ausreißer wurden in der Abbildung abgeschnitten.

5.3.3 Unterschiedlichen Regressionsmodellen

Wenn man die Korrekturmethoden nach Regressionsmodellen vergleicht wie in Abbildung 5, erkennt man dass alle Methoden mediantreu sind, außer die korrigierte Score Funktion im logistischen Fall. Dass die Schätzung bei der korrigierten Score Funktion für die logistische Regression abweicht, lässt sich darauf zurückführen, dass man für die logistische Regression nur eine approximative korrigierte Score Funktion aufstellen kann (siehe Kapitel 4.4). Die korrigierte Monte-Carlo Funktion scheint eine gute Alternative Berechnung zur exakt korrigierten Score Funktion zu sein, für den Fall der linearen und logistischen Regression sogar tendenziell besser. Natürlich kann man nicht ausschließen, dass bei Regressionen für die man keine exakte korrigierte Score Funktion aufstellen kann, die Monte-Carlo Funktion sich genauso verhält wie in diesen Fällen.

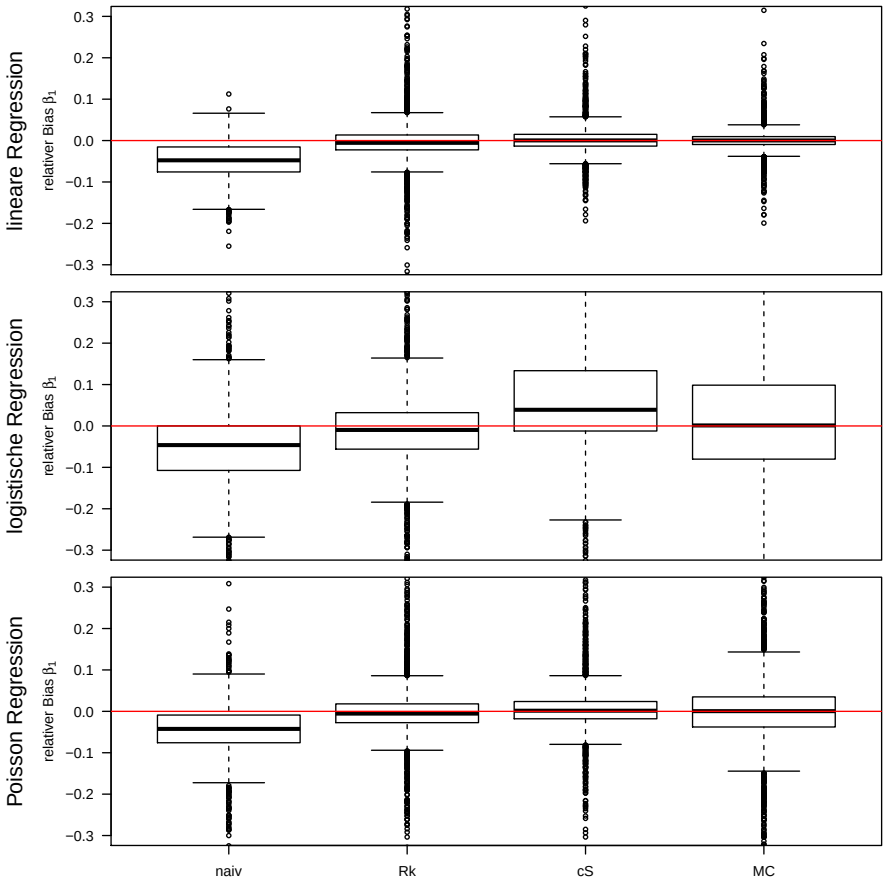

Abbildung 5: Vergleich der Methoden nach unterschiedlichen Regressionsmodellen. Die Y-Achse ist $\pm$ 0.5 eingeschränkt.

Abbildung 6 soll nochmals die Korrektureffekte visualisieren. Dafür wurde der Median über die bench, naiven und korrigierten β Parameter über alle Datensätze gebildet.

Es wird verdeutlicht, dass das die Anwendung der Korrekturmethoden die Schätzung gegenüber der naiven Regression verbessert.

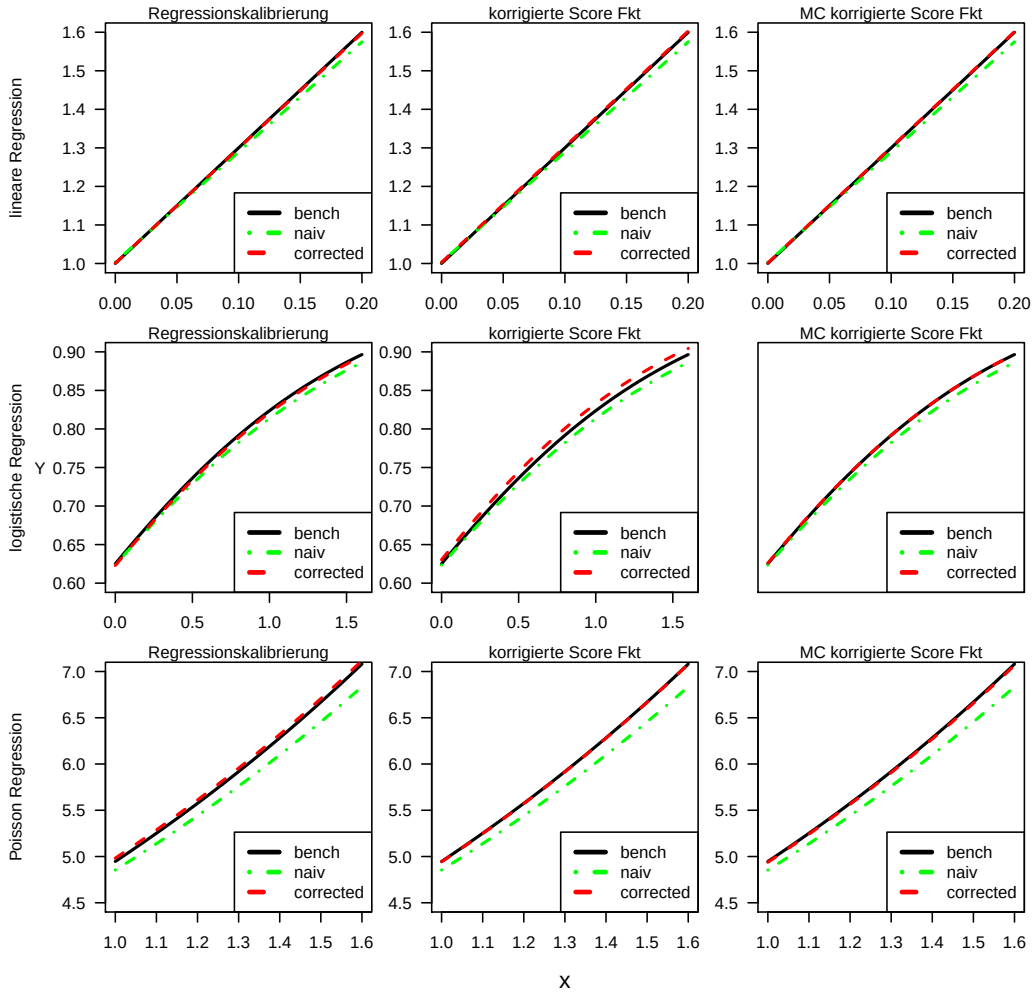

Abbildung 6: Visualisierung der drei Methoden nach unterschiedlichen Regressionsmodellen. Es wurde der Median für die geschätzten Benchparameter, naiv geschätzte Paramater und die korrigiert geschätzten Parameter berechnet. Schwarz durchgezogenen Linie ist die Regressionsgerade der Bench Schätzung, die grün gestrichelte und gepunktete Linie ist die Regressionsgerade aus der naiven Regression und die rot gestrichelte Linie die korrigierte Schätzung.

5.3.4 Unterschiedliche Anzahl an Messwiederholungen

Für diese Fragestellung wurden die Daten mit Daten-Ziehungstype 2 generiert. Der Grund kann durch folgendes Beispiel erklärt werden.
Beispiel 1: Von Interesse sei ob die Ernährung (pflanzliche und tierische Proteine) einen Einfluss auf Herz-Kreislauf-Erkrankungen hat. Die Grundgesamtheit der infrage kommenden Befragten hat einen Umfang von 5000 (entspricht Basisdatensatz aus Daten 3). Da die Aufnahme von pflanzlichen und tierischen Proteinen nicht direkt gemessen werden kann müssen die Informationen z.B. aus detaillierten Ernährungstagebüchern extrahiert werden. Exakte Messungen sind also nicht möglich. Aus finanziellen Gründen können zum einen nicht alle befragt werden und sind zum anderen Einschränkungen bei der Durchführung zu beachten. Man stellt sich folglich die Fragen: Inwiefern verbessert sich meine Messung wenn ich meine Anzahl an Messwiederholungen erhöhe?

Die Methoden wurden für zwei, vier, sechs, acht, zehn und 20 Messwiederholungen verglichen. Abbildung 11 zeigt exemplarisch die Schätzungen für zwei, acht und 20 Messwiederholungen.

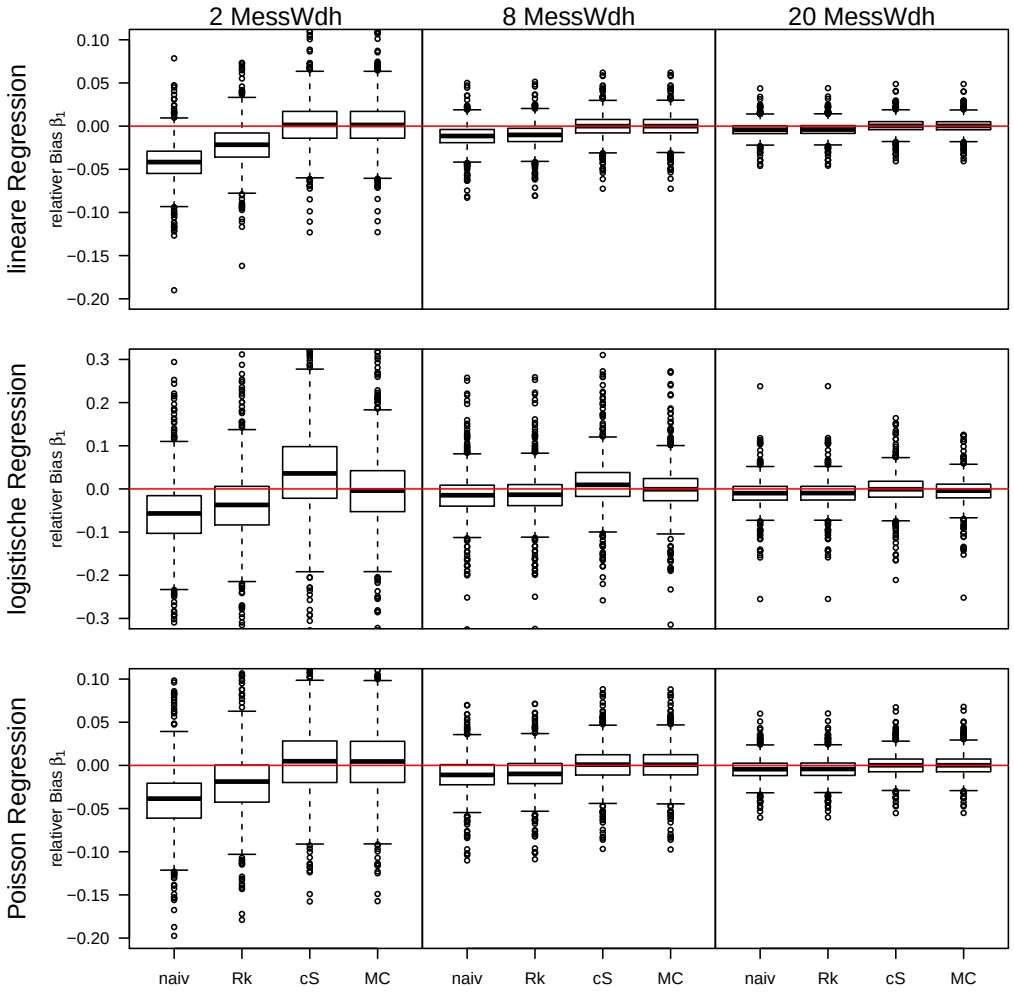

Abbildung 7: Vergleich der Methoden nach unterschiedlichen Anzahl an Messwiederholungen. Die Y-Achse ist eingeschränkt.

Im Allgemeinen führen eine höhere Anzahl an Messwiederholungen zu mediantreuen Schätzern. Die korrigierte und Monte-Carlo korrigierten Score scheinen bei bereits zwei Messwiederholungen mediantreue Schätzer zu liefern, ausgenommen den logistischen Fall für die korrigierte Score Funktion.

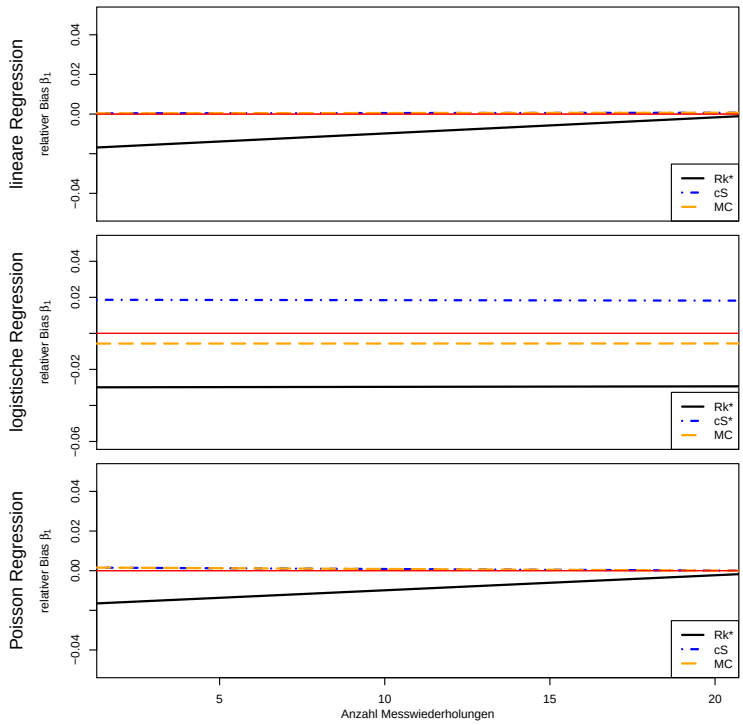

Abbildung 8: Vergleich der Methoden nach unterschiedliche Regressionsmodellen und nach unterschiedlichen Anzahl an Messwiederholungen durch robuste Regression von relativer Bias auf die Anzahl der Messwiederholungen. Signifikanz (0.05-Niveau) des Steigungsparameters werden mit einem * in der Legende markiert.

Zusätzlich wurde für jede Methode eine robuste lineare Regressionsgerade geschätzt, mit dem relativen Bias von β_1 als Zielgröße und die Anzahl der Messwiederholung als Einflussgröße. Es wurde eine robuste Regression aufgrund der vielen Ausreißer gewählt. Die berechneten robuste lineare Regressionen zeigen allerdings, dass die Anzahl an Messwiederholungen keine Auswirkung auf den relativen Bias haben, außer die Regressionskalibrierung im Fall von linearen und Poisson Regression. Wenn man aber die Y-Skala beachtet, ist die Steigung als gering zu bewerten.

5.3.5 Unterschiedliche Anteile an Validierungsdaten

In diesem Abschnitt wurden die Methoden für unterschiedliche Anteile an Validierungsdaten untersucht. Betrachtet wurden dabei 10%,15%,20%,25%,30%,40% und 50% an Validierungsdaten. Abbildung 11 zeigt exemplarisch die Schätzungen für 10%, 25% und 50% vorliegende Validierungsdaten.

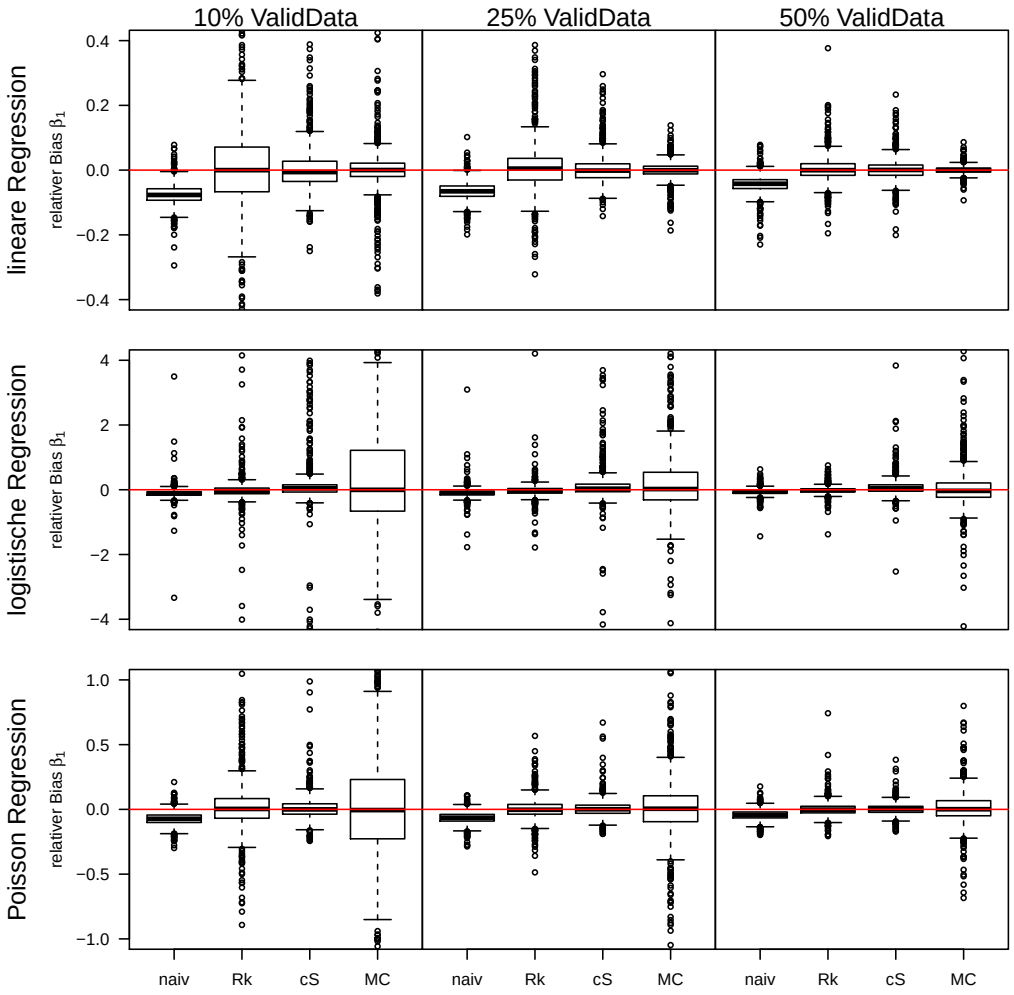

Abbildung 9: Vergleich der Methoden nach unterschiedlichem Anteil von Validierungsdaten. Die Y-Achsen sind eingeschränkt.

Im Gegensatz zu Messwiederholungen (Kapitel 5.3.5) scheint die korrigierte Score Funktion tendenziell das vorliegen von Validierungsdaten besser zu nutzen. Aus Abbildung 10 kann man erkennen, dass diese beinahe unabhängig vom Anteil der Validierungsdaten ist, während die Schätzung aus der Regressionskalibrierung und der Monte-Carlo

korrigierten Score Funktion bei steigendem Anteil verbessert. Alle drei Korrekturmethoden scheinen mediantreu zu sein. Betrachtet man die Ergebnisse der robusten Regression in Abbildung 10 scheint nur für die Regressionskalibrierung die Schätzung des Steigungsparameter signifikant (zum 0.05-Niveau) zu sein. Man sollte hier aber anmerken, dass eine Abweichung von 0.01 vom Nullpunkt ein geringer Wert ist, was für alle drei Korrekturmethoden spricht.

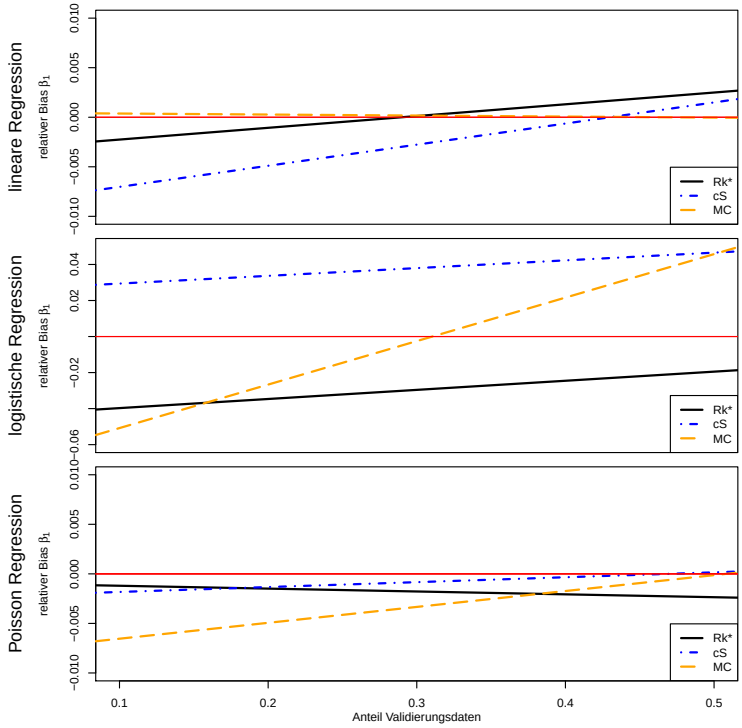

Abbildung 10: Vergleich der Methoden nach unterschiedliche Regressionsmodellen und Anteil an Validierungsdaten, durch robuste lineare Regressions von relativer Bias auf den Anteil an Validierungsdaten. Signifikanz (0.05-Niveau) des Steigungsparameters werden mit einem * in der Legende markiert. Beachte die unterschiedliche Y-Skalen.

5.3.6 Unterschiedliche Werte der geschätzten Messfehlervarianz aus externen Daten

Die Regressionskalibrierungsmethode bei Messwiederholungen und die (Monte Carlo) korrigierte Score Funktion bei Messwiederholungen und Validierungsdaten benötigen für die Parameterschätzung die Messfehlervarianz. Diese kann aus vorliegenden Daten geschätzt werden oder falls aus externen Daten bekannt auch übergeben werden. Für den Fall einer Messung muss dieser Wert aus externen Daten vorliegen. In diesem Kapitel wird untersucht wie unterschiedliche Genauigkeit der geschätzte Messfehlervarianz aus externen Daten auf die Parameterschätzung wirkt, dabei wird nur der Fall von Messwiederholungen betrachtet, da die Regressionskalibrierung bei Validierungsdaten keine Schätzung der Messfehlervarianz benötigt. Für die Messfehlervarianz wurden vier Fälle betrachtet

- $\hat{\sigma}_u^2 = \sigma_u^2$ wahre Messfehlervarianz
- $\hat{\sigma}_u^2 = 1.1 * \sigma_u^2$ kleine Abweichung von der wahren Messfehlervarianz
- $\hat{\sigma}_u^2 = 1.5 * \sigma_u^2$ mittlere Abweichung von der wahren Messfehlervarianz
- $\hat{\sigma}_u^2 = 2 * \sigma_u^2$ große Abweichung von der wahren Messfehlervarianz

Je größer die Abweichung des übergebenen Messfehlervarianz von der wahren Messfehlervarianz ist, desto schlechter wird die Schätzung (siehe Abbildung **??**). Die korrigierte Score Funktion scheint bei der logistischen Regression am empfindlichsten zu reagieren. Die Regressionskalibrierung ist auch bei große Abweichung von der wahren Messfehlervarianz noch mediantreu im Vergleich zu den anderen Methoden. Es ist dennoch bemerkenswert, das selbst wenn man die Messfehlervarianz 1.5 mal so groß schätzt wie sie tatsächlich ist, trotzdem mit den drei Methoden einen relativen Bias nahe Null erreichen kann, mit Ausnahme des logistischen Falles bei der korrigierten Score Funktion.

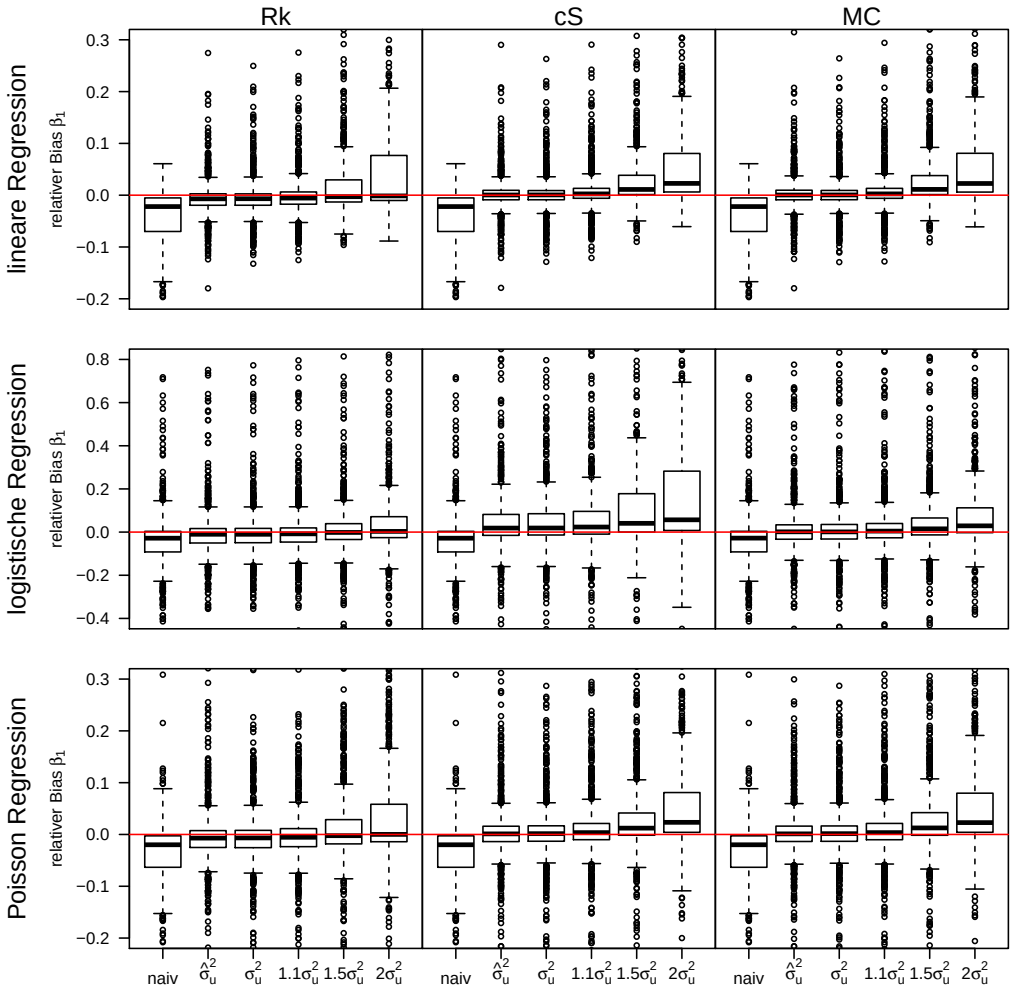

Abbildung 11: Vergleich der Methoden nach unterschiedlichen Werte für die extern geschätzte Messfehlervarianz. Verglichen werden die Fälle: Messfehlervarianz-Schätzung aus den vorliegenden Daten ($\hat{\sigma}_u^2$) und übergebene externe Messfehlervarianz, wobei $\sigma_u^2 = 0.3^2$ die wahre Messfehlervarianz ist . Ausreißer sind abgeschnitten.

5.3.7 Bei Annahmeverletzungen

Bis zu diesem Kapitel haben sich alle drei Methoden gut bewährt. Interessant wird es, wenn bestimmte Annahmen zum Messfehler U verletzt werden, da in realen Situationen das Erfüllen von Annahmen nicht immer gewährt werden kann. Daher werden Daten mit Kombinationsliste 1 und Daten-Ziehungstyp 1 simuliert und zusätzlich Annahmeverletzung eingebaut. Im folgenden steht die Abkürzung KA für keine Annahmeverletzung und AV für Annahmeverletzung.

5.3.7.1 Messfehler nicht rein additiv

Für die Annahmeverletzung AV1 geht man davon aus, dass kein rein additiver Fehler vorliegt, sondern ein zusätzlicher multiplikativer Fehler. Die fehlerhafte Messung wird wie folgt generiert:

$$X = X * \tilde{U} + U$$

mit U und $\tilde{U} \sim N(1, 0.3^2)$.
Trotz Annahmeverletzung bleiben die Schätzung der drei Korrekturmethoden mediantreu. Die Annahmeverletzung beeinflusst die naive Regression im negativem Sinne. Im Allgemeinen scheint aber ein multiplikativer Fehler mit gleich großer Messfehlervarianz wie der additiver Fehler die Güte der Schätzung bei Anwendung von Korrekturmethoden nicht ausschlaggebend zu verschlechtern, abgesehen von dem Fall der logistischen Regression für die korrigierte Score.

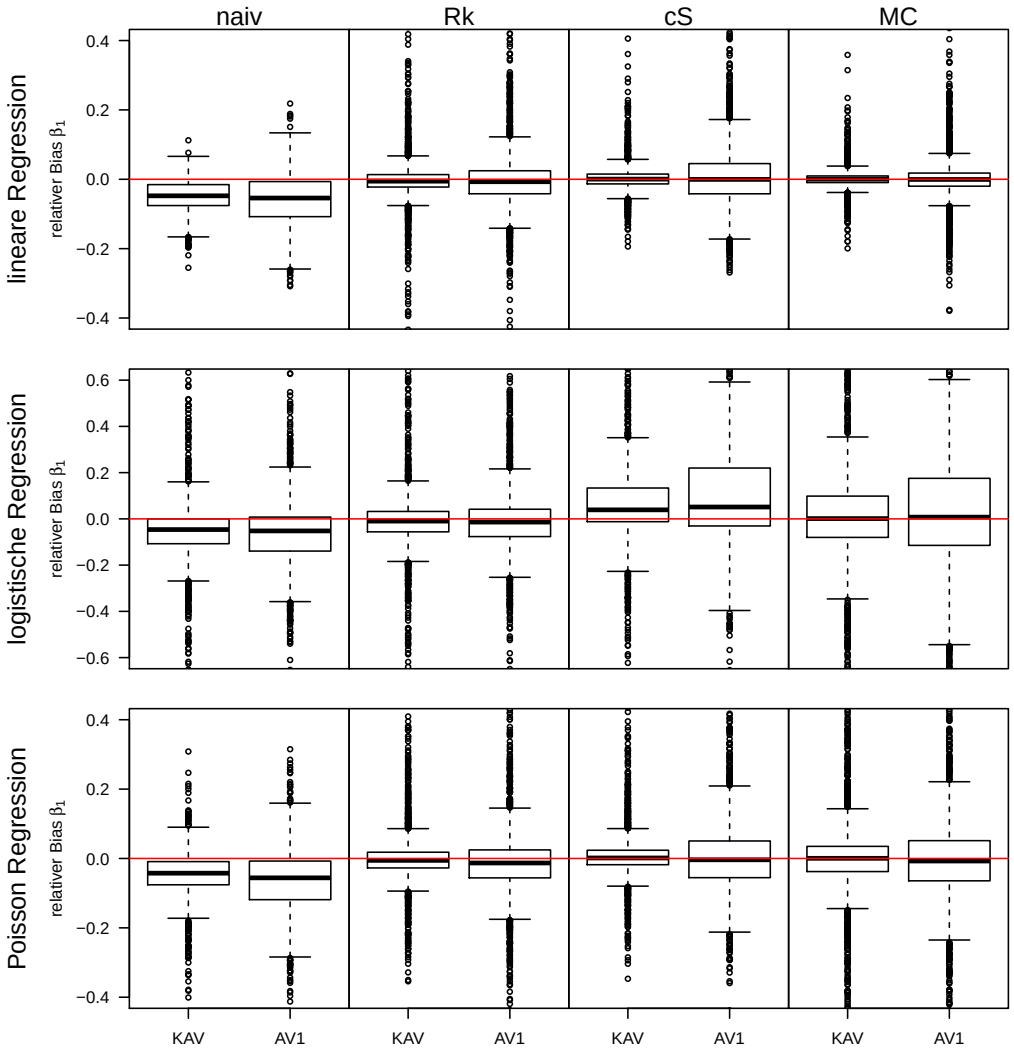

Abbildung 12: Vergleich der Methoden bei Annahmeverletzung, dass nicht nur ein additiver Fehler vorliegt, sondern zusätzlich einen multiplikativen Fehler.

5.3.7.2 Fehler U nicht Normalverteilt

Für die Annahmeverletzung AV2 geht man davon aus, dass kein normalverteilter Fehler vorliegt. Die fehlerhafte Messung U folgt nicht mehr der Normalverteilung sondern der Laplace Verteilung $Lp(\mu = 0, \sigma = 0.3)$, d.h. mit Mittelwert 0 und Varianz $2\sigma^2$. Die Laplace Verteilung ist eine heavy tail Verteilung und ist gut geeignet zur Simulation von Ausreißern. Plot 13 zeigt einen Vergleich der Normalverteilung und der Laplace Verteilung.

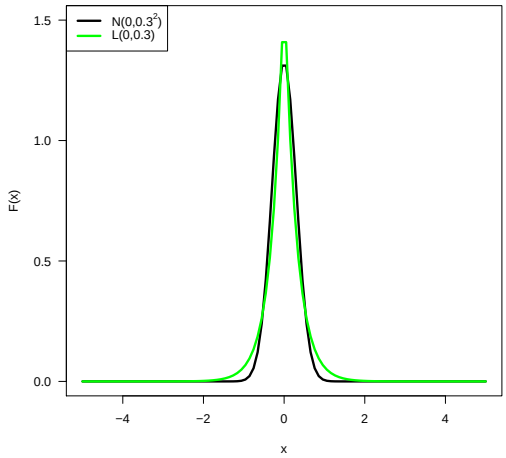

Abbildung 13: Vergleich der Normalverteilung und der Laplaceverteilung.

Auch bei dieser Annahmeverletzung scheint nach Abbildung 14 keine maßgebliche Verschlechterung vorzuliegen. Die Korrektur durch die Regressionskalibrierung verschlechtert sich etwas, aber bei der gewählten Y-Skala ist der Effekt vernachlässigbar. Und es gilt wieder dass die korrigierte Score für die logistische Regression von den restlichen Methoden und Fälle unterscheidet und nicht mediantreu ist.

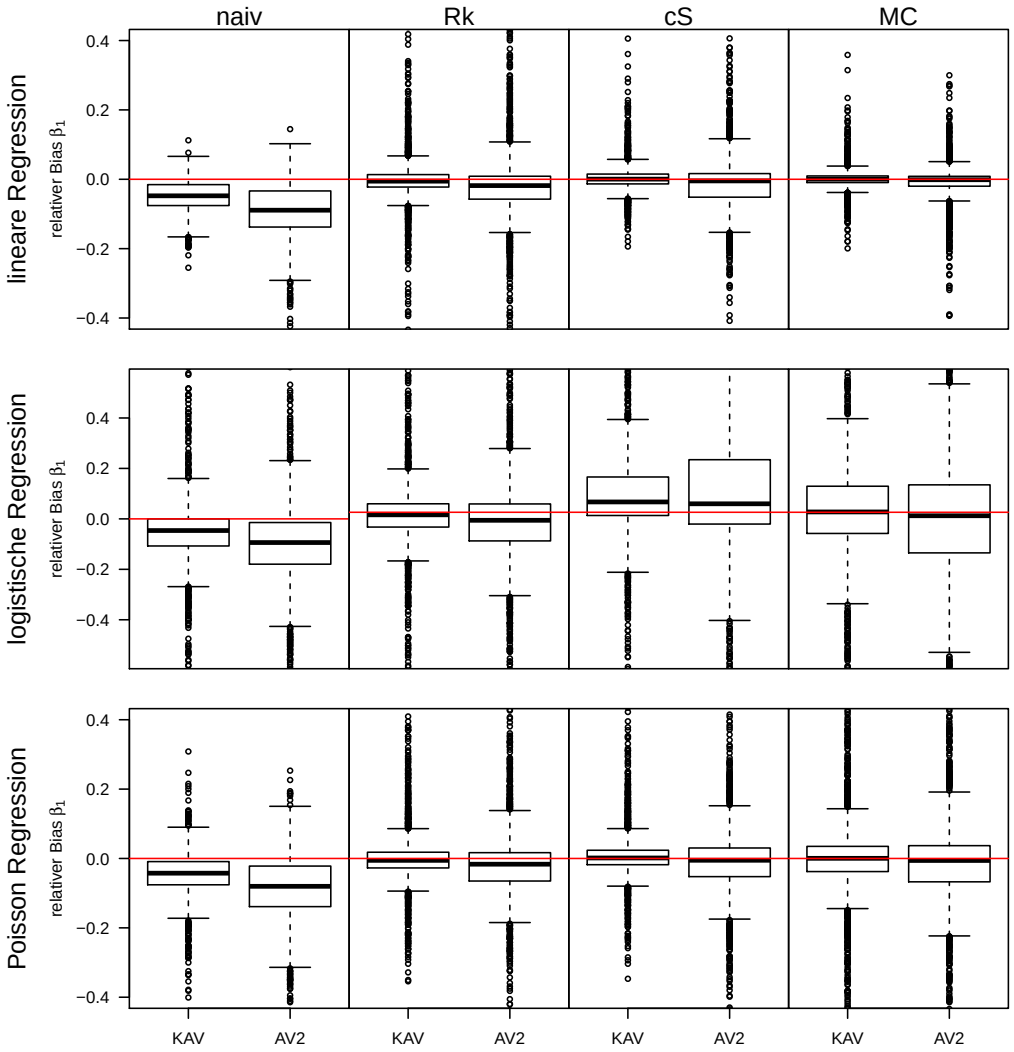

Abbildung 14: Vergleich der Methoden für die Annahmeverletzung, dass eine Laplaceverteilung statt eine Normalverteilung für Fehler U vorliegt.

5.3.7.3 Fehler U nicht unabhängig von X

Für die Annahmeverletzung AV3 geht man davon aus, dass der Fehler U nicht unabhängig ist von X. Dazu gilt für die Verteilung des Fehlers U

$$U \sim N(0, 0.3^2) \ für \ x < 1$$
$$U \sim N(0, (2 * 0.3)^2) \ für \ x \geq 1.$$

D.h. für größere X Werte erwartet man einen größere Messfehlervarianz.

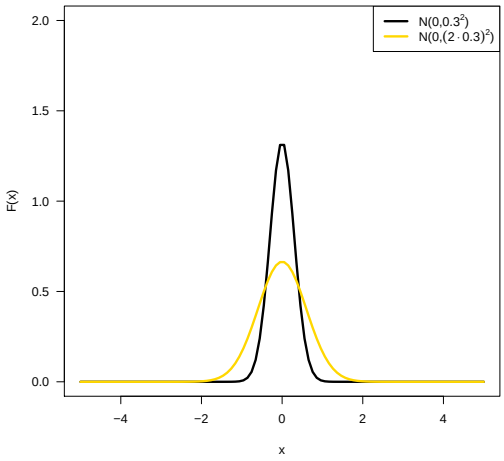

Abbildung 15: Vergleich der Normalverteilung mit unterschiedlicher Varianzen.

Im Allgemeinen scheint die Abhängigkeit von U keine ausschlagebenen Einfluss auf die Korrektur zu haben. Aber wieder erkennt man, dass die korrigierte Score Funktion für den logistischen Fall unsichere Schätzungen aufweist (vgl. Abbildung 16). Die Regressionskalibrierung verschlechtert sich bei Annahmeverletzung AV3 im Median ein wenig. Unter Berücksichtigung der Y-Skala, kann man dies vernachlässigen.

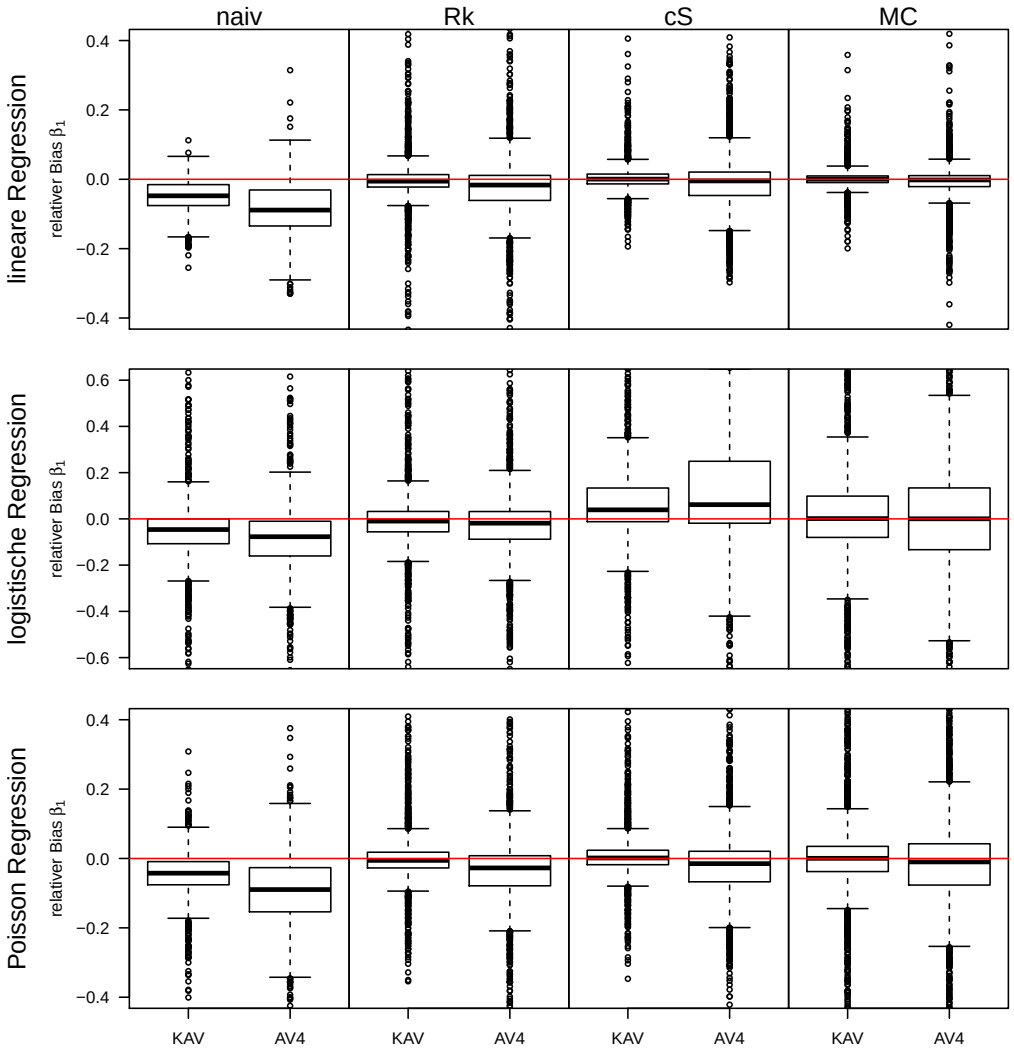

Abbildung 16: Vergleich der Methoden bei Annahmeverletzung, dass Messfehlervarianz U von X abhängig ist.

5.3.7.4 Fehler U unterschiedlich schief Normalverteilt

Für die Annahmeverletzung AV4 geht man zwar davon aus, dass ein normalverteilter Fehler vorliegt. Dafür ist der Fehler U einmal mit $N(0, 0.3^2)$ und Schiefeparameter 1.5 und einmal mit Schiefeparameter 3 verteilt; d.h. $U \sim SN(0, 0.3^2, \tilde{\lambda})$ mit $\tilde{\lambda}$=Schiefeparameter. Diese Fälle werden mit der symmetrischen Verteilung verglichen, in der der Fehler U $N(0, 0.3^2)$ verteilt ist mit Schiefe 1. Plot 17 visualiert die unterschiedliche Grade an Schiefe.

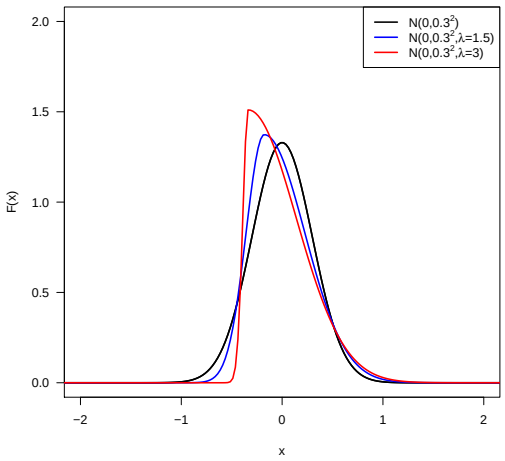

Abbildung 17: Vergleich der Normalverteilung mit unterschiedlichem Schiefeparameter

Die drei Messfehlerkorrekturmethoden scheinen auf diese Annahmeverletzung, dass kein symmetrische Normalverteilung vorliegt am robustesten zu reagieren. Man erkennt in Abbildung 18 kaum Unterschiede in der Güte der Schätzung, außer dass die korrigierte Score für die logistische Regression nicht mediantreu ist

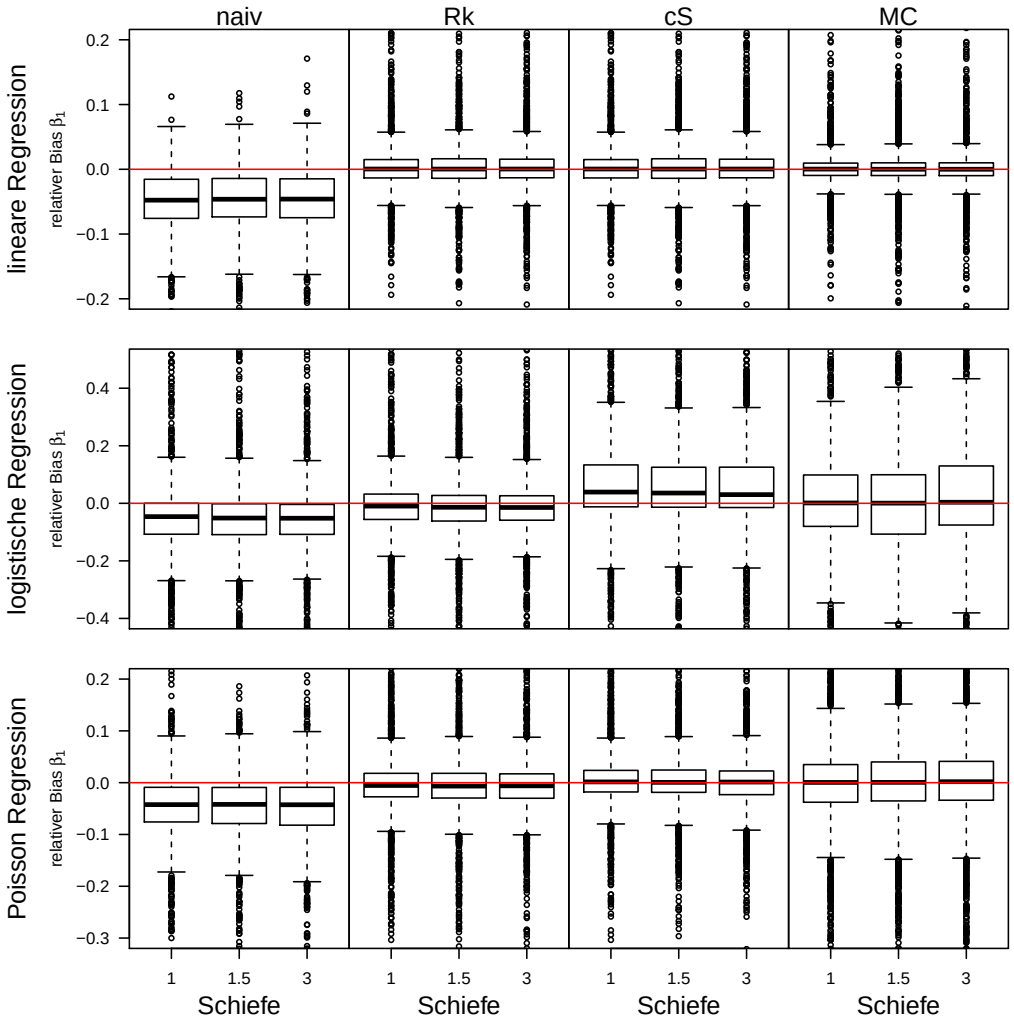

Abbildung 18: Vergleich der Methoden bei Annahmeverletzung, dass keine symmetrische Normalverteilung für den Messfehler vorliegt, sondern schiefe Normalverteilungen.

Abgesehen davon, dass diese Annahmeverletzung keinen Effekt auf die Schätzung zu haben scheint, hat man auch die Möglichkeit die korrigierte Score Funktion direkt für schief Normalverteilte Messfehler für die lineare und Poisson Regression herzuleiten.
Für die lineare Regression ist die Herleitung analog zu Kapitel C.1.1 bis zu Formel 60; Für die schiefe Normalverteilung $\tilde{U} \sim SN(\tilde{\lambda})$ gilt:

$$E[\tilde{U}] = \frac{\tilde{\lambda}}{\sqrt{1+\tilde{\lambda}^2}}\sqrt{\frac{2}{\pi}}$$
$$V[\tilde{U}] = 1 - E[\tilde{U}]^2 = 1 - \frac{\tilde{\lambda}^2}{1+\tilde{\lambda}^2}\frac{2}{\pi}$$

nach Werner [2013, S.23]. Weiterhin gilt nach dem Verschiebungssatz

$$E[\tilde{U}^2] = V[\tilde{U}] + E[\tilde{U}]^2 = 1 - E[\tilde{U}]^2 + E[\tilde{U}]^2 = 1$$

sodass statt Formel 61 Folgendes ergibt:

$$= -\frac{1}{2}nlog(2\pi) - nlog(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i^2 - 2\beta_0 y_i - 2\beta_1 E[y_i x_i|y_i, x_i] + 2\beta_1 E[y_i u_i|y_i, x_i] + \beta_0^2$$
$$+ 2\beta_0\beta_1 E[x_i|y_i, x_i] + 2\beta_0\beta_1 E[u_i|y_i, x_i] + \beta_1^2 E[x_i^2|y_i, x_i] + 2\beta_1^2 E[x_i u_i|y_i, x_i] + \beta_1^2 E[u_i^2|y_i, x_i])$$
$$= -\frac{1}{2}nlog(2\pi) - nlog(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i^2 - 2\beta_0 y_i - 2\beta_1 y_i x_i + 2\beta_1 y_i \frac{\tilde{\lambda}_i}{\sqrt{1+\tilde{\lambda}_i^2}}\sqrt{\frac{2}{\pi}}$$
$$+ \beta_0^2 + 2\beta_0\beta_1 x_i + 2\beta_0\beta_1 \frac{\tilde{\lambda}_i}{\sqrt{1+\tilde{\lambda}_i^2}}\sqrt{\frac{2}{\pi}} + \beta_1^2 x_i^2 + 2\beta_1^2 x_i \frac{\tilde{\lambda}_i}{\sqrt{1+\tilde{\lambda}_i^2}}\sqrt{\frac{2}{\pi}} + \beta_1^2)$$

die korrigierte Likelihood Funktion ergibt sich zu

$$l_c(Y, X^*, \beta) = -\frac{1}{2}nlog(2\pi) - nlog(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i^2 - 2\beta_0 y_i - 2\beta_1 y_i x_i^* + \beta_0^2 + 2\beta_0\beta_1 x_i^*$$
$$+ \beta_1^2 x_i^{*2} - 2\beta_1 y_i \frac{\tilde{\lambda}_i}{\sqrt{1+\tilde{\lambda}_i^2}}\sqrt{\frac{2}{\pi}} - 2\beta_0\beta_1 \frac{\tilde{\lambda}_i}{\sqrt{1+\tilde{\lambda}_i^2}}\sqrt{\frac{2}{\pi}} - 2\beta_1^2 x_i^* \frac{\tilde{\lambda}_i}{\sqrt{1+\tilde{\lambda}_i^2}}\sqrt{\frac{2}{\pi}} - \beta_1^2) \quad (51)$$

und somit ist die korrigierte Score Funktion für die lineare Regression mit schief normalverteilten Messfehler

$$\frac{\partial l_c}{\partial \beta_0} = S_{c0}(Y, X^*, \beta) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i^* + \beta_1 \frac{\tilde{\lambda}_i}{\sqrt{1+\tilde{\lambda}_i^2}}\sqrt{\frac{2}{\pi}})$$
$$\frac{\partial l_c}{\partial \beta_1} = S_{c1}(Y, X^*, \beta) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i x_i^* - \beta_0 x_i^* - \beta_1 x_i^{*2}$$
$$+ y_i \frac{\tilde{\lambda}_i}{\sqrt{1+\tilde{\lambda}_i^2}}\sqrt{\frac{2}{\pi}} + \beta_0 \frac{\tilde{\lambda}_i}{\sqrt{1+\tilde{\lambda}_i^2}}\sqrt{\frac{2}{\pi}} + 2\beta_1 x_i^* \frac{\tilde{\lambda}_i}{\sqrt{1+\tilde{\lambda}_i^2}}\sqrt{\frac{2}{\pi}} + \beta_1)$$

Für die Poisson Regression ist die Herleitung analog zu Kapitel C.1.2 bis zu Formel 70; Für die schiefe Normalverteilung $\tilde{U} \sim SN(\tilde{\lambda})$ ist die Momenterzeugende Funktion nach

Werner [2013, S.22]:

$$E[\exp(t\tilde{U})] = 2\exp(\frac{t^2}{2})\tilde{\Phi}(\frac{\tilde{\lambda}}{\sqrt{(}1+\tilde{\lambda}^2)}t), t \in \mathbb{R}$$

wobei $\tilde{\Phi}$ die Verteilungsfunktion von $\tilde{U}$ ist. Somit ergibt sich statt der Formel 71 Folgende Formel

$$\begin{aligned}
&= \sum_{i=1}^{n} E[-\exp(\beta_0 + \beta_1 x_i)|y_i, x_i]E[\exp(\beta_1 u_i)|y_i, x_i] + E[y_i\beta_0|y_i, x_i] \\
&+ E[y_i x_i \beta_1|y_i, x_i] + E[y_i u_i \beta_1|y_i, x_i] - E[log(y_i!)|y_i, x_i] \\
&= \sum_{i=1}^{n} -\exp(\beta_0 + \beta_1 x_i)2\exp(\frac{\beta_1^2}{2})\tilde{\Phi}(\frac{\tilde{\lambda}}{\sqrt{(}1+\tilde{\lambda}^2)}\beta_1) + y_i\beta_0 \\
&+ y_i x_i \beta_1 + y_i \beta_1 \frac{\tilde{\lambda_i}}{\sqrt{1+\tilde{\lambda_i}^2}}\sqrt{\frac{2}{\pi}} - log(y_i!) \\
&= \sum_{i=1}^{n} -\exp(\beta_0 + \beta_1 x_i + \frac{\beta_1^2}{2})2\tilde{\Phi}(\frac{\tilde{\lambda}}{\sqrt{(}1+\tilde{\lambda}^2)}\beta_1) + y_i\beta_0 \\
&+ y_i x_i \beta_1 + y_i \beta_1 \frac{\tilde{\lambda_i}}{\sqrt{1+\tilde{\lambda_i}^2}}\sqrt{\frac{2}{\pi}} - log(y_i!) \qquad (52)
\end{aligned}$$

die korrigierte Likelihood Funktion ergibt sich zu

$$\begin{aligned}
l_c(Y, X^*, \beta) = \sum_{i=1}^{n} &-\exp(\beta_0 + \beta_1 x_i - \frac{\beta_1^2}{2})\frac{1}{2}\tilde{\Phi}(\frac{\tilde{\lambda}}{\sqrt{(}1+\tilde{\lambda}^2)}\beta_1)^{-1} + y_i\beta_0 \\
&+ y_i x_i \beta_1 - y_i \beta_1 \frac{\tilde{\lambda_i}}{\sqrt{1+\tilde{\lambda_i}^2}}\sqrt{\frac{2}{\pi}} - log(y_i!) \qquad (53)
\end{aligned}$$

für die korrigierte Score Funktion für die Poisson Regression mit schief normalverteilten Messfehler ergibt:

$$\begin{aligned}
\frac{\partial l_c}{\partial \beta_0} &= S_{c0}(Y, X^*, \beta) = \sum_{i=1}^{n} -\exp(\beta_0 + \beta_1 x_i - \frac{\beta_1^2}{2})\frac{1}{2}\tilde{\Phi}(\frac{\tilde{\lambda}}{\sqrt{(}1+\tilde{\lambda}^2)}\beta_1)^{-1} + y_i \\
\frac{\partial l_c}{\partial \beta_1} &= S_{c1}(Y, X^*, \beta) = \sum_{i=1}^{n} -\exp(\beta_0 + \beta_1 x_i - \frac{\beta_1^2}{2})\frac{1}{2}\tilde{\Phi}(\frac{\tilde{\lambda}}{\sqrt{(}1+\tilde{\lambda}^2)}\beta_1)^{-1}(x_i - \beta_1) \\
&- \exp(\beta_0 + \beta_1 x_i - \frac{\beta_1^2}{2})\frac{1}{2}\frac{\partial\tilde{\Phi}(\frac{\tilde{\lambda}}{\sqrt{(1+\tilde{\lambda}^2)}}\beta_1)^{-1}}{\partial\beta_1} + y_i x_i - y_i \frac{\tilde{\lambda_i}}{\sqrt{1+\tilde{\lambda_i}^2}}\sqrt{\frac{2}{\pi}} - log(y_i!)
\end{aligned}$$

6 Fazit

Die vorgestellten Korrekturmethoden stellen eine effektive Möglichkeit dar mit Messfehlern in der Einflussgröße umzugehen. Große Vorteile aller drei Methoden sind, dass sie keine Annahmen über die Verteilung der nicht beobachtbaren wahren Variable X benötigen, auf viele (GLM) Modelle anwendbar sind und den Bias reduzieren, der durch Messfehler entsteht (Augustin et al. [2008, S.257], Caroll, R. J. and Ruppert, D. and Stefanski, L. A. and Crainiceanu, C. M. [2006, S.151], Buzas [2009, S.1]). Es werden zwar Annahmen über den Fehler U gemacht, allerdings scheinen Annahmeverletzungen keinen auschlaggebenden Einfluss auf die Schätzung zu haben, zumindest für die drei betrachteten Modelle der linearen, logistischen und Poisson Regression (vgl. Simulationsteil 5). Wobei die Regressionskalibrierung bei Annahmeverletzungen ein robusteres Verhalten aufweist also die Monte-Carlo korrigierte Score und korrigierte Score Funktion, dafür Verhalten diese sich tendenziell besser wenn Annahmen erfüllt sind.
Weitere Vor- und Nachteile der korrigierten Score Funktion und der Monte-Carlo korrigierten Score Funktion sind zum einen, dass sie zu konsistente valide Schätzer führen, gegeben dem Fehlermodell und den wahren Daten (Caroll, R. J. and Ruppert, D. and Stefanski, L. A. and Crainiceanu, C. M. [2006, S.151]) und somit ebenfalls valide Inferenz (Caroll, R. J. and Ruppert, D. and Stefanski, L. A. and Crainiceanu, C. M. [2006, S.151]). Die korrigierte Score Funktion ist zudem unverzerrt (Nakumara [1990, S.128]). Außerdem sind die beiden Methoden unabhängig von der Größe des Messfehlers (Buzas [2009, S.1]) ist, ausgenommen die approximative Score Funktion für die logistische Regression. Die Nachteile der exakt korrigierten Score Funktion ist, dass diese nicht immer aufstellbar ist, jedoch kann die Monte-Carlo korrigierte Score Funktion die korrigierte Score Funktion präzise schätzen bedingt auf den wahren Daten und dem behandelten Fehlermodell (Caroll, R. J. and Ruppert, D. and Stefanski, L. A. and Crainiceanu, C. M. [2006, S.151]). Allerdings benötigt die Monte-Carlo korrigierte Score Funktion Software die mit komplexen Zahlen umgehen können. Außerdem ist für jedes Modell die wahre Score Funktion aufzustellen, die oftmals schwierig zu berechnen ist. Auch sind sowohl für die waher Score Funktion als auch für die korrgierte Score Funktion die Nullstellen nicht immer auffindbar.
Und weitere Vor- und Nachteile der Regressionskalibrierung ist zum einen, dass der Methode einfache Berechnungen unterliegen, die ohne extra Implementierung durchführbar sind und dass eine anschließende Standardanalysen möglich ist, als hätte man die wahren X beobachtet (Caroll, R. J. and Ruppert, D. and Stefanski, L. A. and Crainiceanu, C. M. [2006, S.65]). Allerdings ist die Regressionskalibrierung nur ein nur approximatives Verfahren (Augustin et al. [2008, S.257]), jedoch in vielem GLM Modellen exakt (Caroll, R. J. and Ruppert, D. and Stefanski, L. A. and Crainiceanu, C. M. [2006, S.65]), d.h in manchen wichtigen Fällen konsistent (wie lineare Regression oder loglinear mean models (z.B. Poisson Modell) (Caroll, R. J. and Ruppert, D. and Stefanski, L. A. and Crainiceanu, C. M. [2006, S.151]). Außerdem stellt die Berechnung der Regression von X auf X^* und Z stellt eine Herausforderung dar, da X nicht beobachtbar ist (Caroll, R. J. and Ruppert, D. and Stefanski, L. A. and Crainiceanu, C. M. [2006, S.65]).

Grundsätzlich kann man sagen, dass alle Methoden eine gleich gute Wahl zur Messfehlerkorrektur sind. Vor allem im Hinblick auf den Simulationsteil dieser Arbeit gelten für die lineare, Poisson Regression und teilweise für die logistische Regression, dass die Anwendung zu besseren Schätzern im Vergleich zu naiven Regression führt. Zusätzlich muss man ergänzen, dass die Monte-Carlo korrigierte Score für den logistische Regression der approximierten korrigierten Score vorzuziehen ist.
Im Rahmen dieser Arbeit wurden die Methoden zur Anwendung von additiven Fehler vorgestellt. Diese Methoden können auch auf multiplikative Fehler (vgl. Caroll, R. J. and Ruppert, D. and Stefanski, L. A. and Crainiceanu, C. M. [2006, Kapitel 4.5], Nakumara [1990, S.132]), sowie auf nicht lineare Modelle (vgl. Caroll, R. J. and Ruppert, D. and Stefanski, L. A. and Crainiceanu, C. M. [2006, Kapitel 4.7], **??**), Zucker et al. [2013]) angewendet werden.
In der Praxis sind Messfehler üblich z.B. in der Medizin, jedoch nicht die Anwendung von Messfehlerkorrekturverfahren. In der Theorie sind die Methoden fortgeschritten entwickelt, jedoch ist es für Laien schwierig den Zugang zu der Anwendung dieser Methoden zu finden. Da das Ignorieren von Messfehlern zu verzerrter Schätzung führt sollte man in Zukunft versuchen auf anderen Gebieten außerhalb der Statistik das Problem von Messfehlern zu thematisieren und evtl. statistische Softwarepackages zu Verfügung stellen, die die Handhabung von Messfehlern erleichtern. Es gibt bereits Projekte wie das Stratos Projekt STRATOS (STRengthening Analytical Thinking for Observational Studies) Str, dass versucht für Laien einen Leitfaden aufzustellen wie man z.B. mit Messfehlern umgeht (in Stratos Topic 4). Mit solchen Projekten kann man erreichen, dass Ergebnisse aus Beobachtungsstudien, die durch statistische Analysen gewonnen werden, durch die richtige Anwendung der Methoden, nicht fundiert werden.

A Notationsübersicht

Übersicht über die in dieser Arbeit verwendeten Notationen.

Y, y_i	fehlerfrei gemessene Zielgröße
Z, z_i	fehlerfrei gemessene Einflussgröße
$\underline{\underline{X}}, X, \underline{x}_i, x_i$	fehlerfrei nicht beobachtbare/beobachtete Einflussgröße
U, u_i	additiver differentieller, klassischer Fehler mit $U \sim N(0, \Sigma_{uu})$
$\underline{\underline{X^*}}, X^*, \underline{x}_i^*, x_i^*$	fehlerbehaftet gemessene Einflussgröße, mit $X^* = X + U$
ϵ, ϵ_i	Residuen der linearen Regression $E[Y\|X, Z]$
$m_Y(X, Z, \beta)$	$E[Y\|X, Z]$
$m_X(X^*, Z, \gamma)$	$E[X\|X^*, Z]$
β_0	Intercept im vorliegenden Regressionsmodell
β_1	Koeffizient von X im vorliegenden Regressionsmodell
β_2	Koeffizient von Z im vorliegenden Regressionsmodell
β	Vektor aus $\beta_0, \beta_1, \beta_2$
β^*	Superscript $*$ kennzeichnet die Parameterschätzung aus der naiven Regression
β_{Rk}	Index Rk kennzeichnet die Parameterschätzung aus der Regressionskalibrierung
β^c	Superscript c kennzeichnet die korrigierte Parameterschätzung aus der korrigierten Score Funktion oder Monte-Carlo korrigierten Score Funktion
γ	Vektor aus γ_0, γ_1 zur Schätzung von $E[X\|X^*, Z]$
Σ_{uu}, σ_u^2	Messfehlerkovarianzmatrix bzw. Messfehlervarianz
σ^2	Varianz der Residuen im linearen Regressionsmodel $E[Y\|X, Z]$
$\tilde{\sigma}^2$	Varianz der Residuen im linearen Regressionsmodel $E[X\|X^*, Z]$
$l()$	wahre Log-Likelihood abhängig von X
$l_c()$	korrigierte Log-Likelihood abhängig von X^*
$S()$	Score Funktion bedingt auf X
$S_c()$	korrigierte Score Funktion abhängig von X^*
$S_{MCcS,B}()$	Monte-Carlo korrigierte Score Funktion abhängig von X^*

B Hintergrundmaterial

B.1 Maximum-Likelihood-Methode

In diesem Teil wird das Verfahren zur Bestimmung des Maximum-Likelihood- Schätzers kurz erläutert, da das die Grundlage der korrigierten Score Funktion ist. Angenommen es liegen $i \in \{1, ..., n\}$ Beobachtungen, die Zielgröße $Y = (y_1, y_2, ...y_n)^t$ aus n unabhängigen und identisch verteilten Realisierungen, und k Einflussgrößen $X = (x_{1i}, x_{2i}, ..., x_{ki})$ und somit k Parameter $\beta = (\beta_1, ..., \beta_k)$, vor. Sei weiterhin $f(y_i, x_i, \beta)$ die Wahrscheinlichkeitsfunktion der Beobachtung y_i. Das Faktorisieren der Wahrscheinlichkeitsfunktion führt zur Likelihood bzw. zur gemeinsamen Dichte der beobachteten Daten.

$$L(Y, Y, \beta) = \prod_{i=1}^{n} f(y_i, x_i, \beta)$$

und Log-likelihood

$$l(Y, X, \beta) = log(L(Y, X, \beta)) = \sum_{i=1}^{n} log(f(y_i, x_i, \beta)),$$

die das Aufstellen der Score Funktion erleichtert. Diese ergibt sich aus der Ableitung der Likelihood bzw. Log-Likelihood Funktion:

$$S(Y, X, \beta) = \frac{\partial l(Y, X, \beta)}{\partial \beta}$$

Das maximieren der $L(Y, X, \beta)$ oder der $l(Y, X, \beta)$ bzw. das Auflösen der Score Funktion $S(Y, X, \beta)$ nach β führt zur Maximum-Likelihood Schätzung $\hat{\beta}_{ML}$, vorausgesetzt die zweite Ableitung an dieser Stelle ist positiv.
Die besonderen Eigenschaften des Maximum-Likelihood- Schätzung sind: ML-Schätzer sind ...

- ...konsistent (für großen Stichprobenumfang)
- ...asymptotisch normalverteilt
- ...asymptotisch erwartungstreu
- ...asymptotisch effizient
- ...invariant

Aufgrund dieser Eigenschaften ist die Maximum- Likelihood Methode weit verbreitet, allerdings bringt diese Methode auch Nachteile mit sich. Zum einen sind die Schätzer oft eine komplizierte Funktion der unbekannten Parameter für die keine analytische Lösungen gefunden werden können, so muss man in vielen Fällen auf iterative nummerische Verfahren zurückgreifen, z.B. Newton-Raphson Methode (Appendix B.2). Zum anderen

setzt die Nutzung der ML-Schätzer Verteilungsannahmen voraus, wie etwa, dass alle x_i unabhängig identisch verteilt sind. Werden diese verletzt so können inkonsistente Schätzer resultieren.
(vgl. Fahrmeir et al. [2009, S.224, S.467])

B.2 Newton-Raphson und Fisher Scoring

Wie in Appendix B.1 beschrieben, ist die Ableitung der Log-Likelihood Funktion $S(Y, X, Z, \beta) = \frac{\partial l(Y,X,Z,\beta)}{\partial \beta}$ die Score Funktion. Wenn man annimmt dass die Log-Likelihood $l(Y, X, Z, \beta)$ zweimal differenzierbar ist, so ist die beobachtete Information als

$$I(y, \beta) = -\frac{\partial S(Y, X, Z, \beta)}{\partial \beta}$$

definiert, die erwartete Information bzw. Fisher Matrix als

$$J(Y, X, Z, \beta) = E[I(Y, X, Z, \beta)] = E[S(Y, X, Z, \beta)S'(Y, X, Z, \beta)]$$

mit $S'(Y, X, Z, \beta) = \frac{\partial S(Y,X,Z,\beta)}{\partial \beta}$
Unter der Annahme, dass die Likelihood den Maximum-Likelihood einen globalen Maximum hat, so erhält man den Schätzer durch lösen der Gleichung $S(Y, X, Z, \beta) = 0$. Jedoch ist es nicht immer möglich eine geschlossene Lösung für genannte Gleichung zu finden, daher ist es nötig numerische Iterationsverfahren einzusetzen. Weit verbreitet ist das Newton Raphson Iterationsverfahren, die Iterationsgleichung ist

$$\beta^{k+1} = \beta^k + \frac{S(Y, X, Z, \beta^k)}{I(Y, X, Z, \beta^k)}$$

wobei mit einem Startwert β^0 begonnen wird und gegen $\hat{\beta}_{ML}$ konvergiert (Qaqish, S.1). Da $J(Y, X, Z, \beta)$ einfacher zu berechnen ist als $I(Y, X, Z, \beta)$, ergibt sich

$$\beta^{k+1} = \beta^k + \frac{S(Y, X, Z, \beta^k)}{J(Y, X, Z, \beta^k)},$$

das man den Fisher Scoring nennt (Qaqish, S.1).
Für Messfehlerkorrekturverfahren gilt schließlich

$$\beta^{k+1} = \beta^k + \frac{S_c(Y, X, Z, \beta^k)}{J_c(Y, X, Z, \beta^k)},$$

sinnvollerweise wählt man den Startwert β_0 als die Parameterwerte aus der naive Schätzung. Voraussetzung ist, dass die Fischer Scoring Iteration gegen die Lösung β konvergiert, muss die Fischer Matrix $F(Y, X, Z, \beta)$ für alle β invertierbar sein. (vgl.Fahrmeir et al. [2009, S.202, S.473])

B.3 Bootstrapping für die Varianzschätzung in der Regressionskalibrierung

Wenn man die Regressionskalibrierung als Korrekturmethode auswählt, sollte man sich im Schritt 2 im Klaren sein, dass für eine Schätzung (von Y) eine Schätzung (von X) verwendet wird, daher sind die resultierenden Varianzen der Paramterschätzer nur als approximativen Wert anzusehen (Augustin et al. [2008, S.257 ff]). Neben dem in dieser Arbeit vorgestellten Bootstrapmethoden kann auch die Sandwichmethode zur Anpassung des Standardfehlers angewendet werden (Caroll, R. J. and Ruppert, D. and Stefanski, L. A. and Crainiceanu, C. M. [2006, S.66 ff, S.369 ff]).
Es gibt das "Resampling Pairs"und das "Resampling Residuals", das zu den nonparametrischen Bootstrap Verfahren zählt (Caroll, R. J. and Ruppert, D. and Stefanski, L. A. and Crainiceanu, C. M. [2006, S.378 ff]).
Vor Ziehung von Bootstrap-Stichproben sollten die unterschiedlichen Datenstrukturen berücksichtigt werden. Ziel ist es Bootstrapping aus einer homogenen Umgebung zu ziehen, oft liegen allerdings heterogenen Strukturen vor, d.h. beispielsweise für einige Beobachtungen können Validierungsdaten beobachtet werden, für andere wiederrum nur eine, oder mehre Messwiederholungen. Der Unterschied ist, dass dadurch unterschiedliche Informationen vorliegen. Das ignorieren dieser Strukturen führt zur erhöhten Varianz in der Bootstrapstichprobe (Caroll, R. J. and Ruppert, D. and Stefanski, L. A. and Crainiceanu, C. M. [2006, S.381 ff]), daher empfiehlt es sich vor der Ziehung nach den unterschiedlichen Datenstrukturen zu gruppieren.
Zur besseren Verständnis der Erklärungen werden im folgenden von Validierungsdaten und der Annahme $Y = m_Y(X, X^*, \alpha) + \epsilon$ ausgegangen, wobei X, X^* im Modell nicht unbedingt vorliegen müssen und α der Schätzer ist.
In diesem Abschnitt können X,X^*,Z können Matrizen sein, die Werte für mehrere Variablen beinhalten.

B.3.1 Resampling Vectors im Messfehlermodell

Wie in Caroll, R. J. and Ruppert, D. and Stefanski, L. A. and Crainiceanu, C. M. [2006, S.378 ff] und Le [2015, S.9] beschrieben, bedeutet das Resampling Vectors- Verfahren, dass vektorweise bzw. zeilenweise gezogen wird. Handelt es sich beispielsweise um Validierungsdaten, so wird Resampling Vectors Bootstrapping aus $\{(Y_i, X_i, X_i^*, Z_i)\}_{i=1}^{n_1}$ und aus $\{(Y_i, X_i^*, Z_i)\}_{i=1}^{n_2}$ gezogen, je nachdem ob wahre X Werte vorliegen oder nicht. Dabei gibt n_1 die Größe der Menge der Vektoren an mit wahren Messungen, n_2 entsprechend ohne wahre Messungen. Da zu jeder gezogenen Beobachtungen i alle zugehörigen Variablen mitgezogen werden, bleiben besondere Beziehungen z.B. Abhängigkeitsbeziehung zwischen den Variablen erhalten.

B.3.2 Resampling Residuals im Messfehlermodell

Die Basis des Resampling Residuals werden in Caroll, R. J. and Ruppert, D. and Stefanski, L. A. and Crainiceanu, C. M. [2006, S.379] beschrieben, die Anwendung auf die Regressionskalibrierung findet sich in Le [2015, S.9 ff] und in diesem Teil wieder. Um

resampling Residuals anwenden zu können müssen ϵ_i und ϵ_i^*, mit $\epsilon_i = Y_i - m_Y(Z_i, X_i, \hat{\alpha})$ und $\epsilon_i^* = Y_i - m_{X^*}(Z_i, X_i, \hat{\alpha})$, wobei α der nichtlineare KQ-Schätzer ist und die Fehler jeweils unabhängig identisch verteilt sein und annähernd der Homoskedastizitätsbedingungen genügen (Caroll, R. J. and Ruppert, D. and Stefanski, L. A. and Crainiceanu, C. M. [2006, S.379] und Le [2015, S.10]). Wenn man wieder von Validierungsdaten ausgehen so hat einen Teil der Daten (Y_i, X_i, X_i^*, Z_i) und einen Teil (Y_i, X_i^*, Z_i) als Beobachtung. Für den ersten Teil kann Resampling Residuals angewendet werden, für den zweiten Teil kann man auf Resampling Vectors zurückgreifen. Grundlage dieses Verfahren ist das vorliegen von den Regressionsmodellen $m_Y(Z_i, X_i)$ und $m_{X^*}(Z_i, X_i)$. Bei diesem Verfahren werden schließlich Residuen (nach Le [2015, S.9 ff]) wie folgt gezogen:
Sei M die Anzahl der Bootstrapstichprobe, für die m-te Ziehung von $Y_i^{(m)}$ gelten:

- 1. $\epsilon_i = Y_i - m_Y(Z_i, X_i, \widehat{\alpha})$ für i $\in$ {1,..,k}
- 2. $B = \{(\epsilon_i - \bar{\epsilon})\}_i^k$
- 3. k mal Ziehen mit zurücklegen aus B, man erhält die Menge $\{\epsilon_i^{(m)}\}_i^k$
- 4. Nun lässt sich $Y_i^{(m)} = m_Y(Z_i, X_i, \widehat{\mathcal{B}}) + \epsilon_i^{(m)}$ für i $\in$ {1,..,k} berechnen

Die Bootstrappziehung für $X_i^{*(m)}$ folgt analog. Die auf diese Art erhaltenen Bootstrapstichproben bedingen auf die wahren (Z_i, X_i).

B.3.3 Algorithmus des Bootstrappings in der Regressionskalibrierung

Für das Bootstrappping in der Regressionskalibrierung lässt sich auch ein allgemeiner Algorithmus aufstellen wie in Le [2015, S.10]:

- Schritt 1: Ziehe M Bootstrap-Stichproben, mithilfe von einem der oben genannten Verfahren.
- Schritt 2: Wende Schritt 1 und Schritt 2 des Regressionskalibrierungsalgorithmus auf jede Stichprobe an.
  - Nach *Resampling Vectors* auf $\{(Y_i, X_i, X_i^*, Z_i)\}_{i=1}^k{}^{(m)} \cup \{(Y_i, X_i^*, Z_i)\}_{i=k+1}^n{}^{(m)}$.
  - Nach *Resampling Residuals* auf $\{(Y_i^{(m)}, X_i^{*(m)}, Z_i, X_i)\}_{i=1}^k \cup \{(Y_i, X_i^*, Z_i)\}_{i=k+1}^n{}^{(m)}$.

  Man erhält somit nach M Durchläufen die Parameter $\widehat{\beta}_{RK_k}^{(1)}$,...,$\widehat{\beta}_{RK_k}^{(M)}$.
- Schritt 3: Aus den vorliegenden $\widehat{\beta}_{RK_k}$s kann nun die Standardabweichung $\widehat{\sigma}_{\beta_{RK_k}}$ geschätzt werden

$$\widehat{var(\widehat{\beta}_{RK_k})} = \frac{1}{M-1} \sum_{m=1}^{M} (\widehat{\beta}_{RK_k}^{(m)} - \overline{\widehat{\beta}_{RK_k}})(\widehat{\beta}_{RK_k}^{(m)} - \overline{\widehat{\beta}_{RK_k}})^t.$$

Mit den resultierenden Schätzungen $\widehat{\sigma}_{\beta_{RK_k}}$ sind nun Signifikanztests möglich. Beispielsweise der T-Test mit der Teststatistik $T = \frac{\widehat{\beta}_{RK_k} - \beta_{RK_k}}{\widehat{\sigma}_{\beta_{RK_k}}}$ für die Nullhypothese
H0: $\widehat{\beta}_{RK_k} = \beta_{RK_k}$.

C Technische Details

C.1 Herleitung exakter korrigierten Score Funktion

Im folgenden werden, die in Kapitel 4.3 aufgestellten korrigierten Score Funktion für die lineare und Poisson Regression, die Herleitung beschrieben. Allgemein gelten folgende Gleichungen:

$E[x_i|y_i, x_i] = x_i$
$E[x_i^2|y_i, x_i] = x_i^2$
$E[y_i|y_i, x_i] = y_i$
$E[y_i^2|y_i, x_i] = y_i^2$
$E[u_i|y_i, x_i] = 0$
$E[exp(\beta_0 + \beta_1 x_i)|y_i, x_i] = exp(\beta_0 + \beta_1 x_i)$
$E[y_i x_i|y_i, x_i] = E[x_i|y_i, x_i]E[y_i|y_i, x_i] = x_i y_i,$ wegen Unabhängigkeit
$E[x_i u_i|y_i, x_i] = E[x_i|y_i, x_i]E[u_i|y_i, x_i] = 0,$ wegen Unabhängigkeit
$E[y_i u_i|y_i, x_i] = E[y_i|y_i, x_i]E[u_i|y_i, x_i] = 0,$ wegen Unabhängigkeit
$E[exp(\beta_0 + \beta_1 x_i)exp(\beta_1 u_i)|y_i, x_i] = E[exp(\beta_0 + \beta_1 x_i)|y_i, x_i]E[exp(\beta_1 u_i)|y_i, x_i],$ wegen Unabhängigkeit
Konstanten können aus dem Erwartungswert herausgezogen werden

C.1.1 Einfache lineare Regression

Modelgleichung: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \underline{x}_i^t \beta + \epsilon_i$ mit $\epsilon_i \sim N(0, \sigma^2)$
Wahrscheinlichkeitsfunktion: $f(y_i, \underline{x}_i, \beta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{\frac{-1}{2\sigma^2}(y_i - \underline{x}_i^t \beta)^2\}$
Schritt 1: Aufstellen der Log-Likelihood Funktion für wahre $\underline{x}_i$

$$
\begin{aligned}
l(Y, X, \beta) &= -\frac{1}{2} nlog(2\pi) - nlog(\sigma) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \underline{x}_i^t \beta)^2 \\
&= -\frac{1}{2} nlog(2\pi) - nlog(\sigma) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i^2 - 2y_i \underline{x}_i^t \beta + \underline{x}_i^t \beta \underline{x}_i^t \beta) \\
&= -\frac{1}{2} nlog(2\pi) - nlog(\sigma) \\
&\quad - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i^2 - 2y_i \beta_0 - 2y_i \beta_1 x_i + \beta_0^2 + 2\beta_0 \beta_1 x_i + \beta_1^2 x_i^2) \qquad (54)
\end{aligned}
$$

Daraus lassen sich die wahren Score Funktionen ableiten zu:

$$\frac{\partial l}{\partial \beta_0} = S_0(Y, X, \beta) = \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) \tag{55}$$

$$\frac{\partial l}{\partial \beta_1} = S_1(Y, X, \beta) = \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) x_i \tag{56}$$

Schritt 2: Aufstellen der log-likelihood Funktion für beobachtete bzw. fehlerhaft gemessenen $\underline{x}_i^*$:

$$\begin{aligned} l(Y, X^*, \beta) = & -\frac{1}{2} nlog(2\pi) - nlog(\sigma) \\ & - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i^2 - 2y_i\beta_0 - 2y_i\beta_1 x_i^* + \beta_0^2 + 2\beta_0\beta_1 x_i^* + \beta_1^2 x_i^{*2}) \\ = & -\frac{1}{2} nlog(2\pi) - nlog(\sigma) \\ & - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i^2 - 2y_i\beta_0 - 2y_i\beta_1 (x_i + u_i) + \beta_0^2 + 2\beta_0\beta_1 (x_i + u_i) + \beta_1^2 (x_i + u_i)^2) \end{aligned} \tag{57}$$

Schritt 3: Berechnen den Erwartungswert $E[l(Y, X^*, \beta)|Y, X]$, wobei in der linearen Regression gilt:

$$E[u_i^2|y_i, x_i] = V[u_i|y_i, x_i] + E[u_i|y_i, x_i] = \sigma_u^2 \text{ Verschiebungssatz} \tag{58}$$

$$\begin{aligned} & E[l(Y, X, \beta)|Y, X)] = -\frac{1}{2} nlog(2\pi) - nlog(\sigma) \\ & - E[\sum_{i=1}^{n} (y_i^2 - 2y_i\beta_0 - 2y_i\beta_1 (x_i + u_i) + \beta_0^2 + 2\beta_0\beta_1 (x_i + u_i) + \beta_1^2 (x_i + u_i)^2)|Y, X] \end{aligned} \tag{59}$$

und der zweite Teil der Formel (59) ergibt sich zu

$$
\begin{aligned}
&= -\frac{1}{2\sigma^2}\sum_{i=1}^{n} E[(y_i^2 - 2y_i\beta_0 - 2y_i\beta_1(x_i+u_i) + \beta_0^2 + 2\beta_0\beta_1(x_i+u_i) + \beta_1^2(x_i+u_i)^2)|y_i, x_i] \\
&= -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i^2 - 2\beta_0 E[y_i|y_i,x_i] - 2\beta_1 E[y_i x_i + y_i u_i)|y_i, x_i] \\
&+ \beta_0^2 + 2\beta_0\beta_1 E[x_i + u_i|y_i, x_i] + \beta_1^2 E[x_i^2 + 2x_i u_i + u_i^2|y_i, x_i]) \\
&= -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i^2 - 2\beta_0 y_i - 2\beta_1 E[y_i x_i|y_i,x_i] + 2\beta_1 E[y_i u_i|y_i,x_i] + \beta_0^2 + 2\beta_0\beta_1 E[x_i|y_i,x_i] \\
&+ 2\beta_0\beta_1 E[u_i|y_i,x_i] + \beta_1^2 E[x_i^2|y_i,x_i] + 2\beta_1^2 E[x_i u_i|y_i,x_i] + \beta_1^2 E[u_i^2|y_i,x_i]) \qquad (60) \\
&= -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i^2 - 2\beta_0 y_i - 2\beta_1 y_i x_i + \beta_0^2 + 2\beta_0\beta_1 x_i + \beta_1^2 x_i^2 + \beta_1^2\sigma_u^2)
\end{aligned}
$$

$$
\begin{aligned}
&\Rightarrow E[l(Y, X^*, \beta|Y, X)|y_i, x_i] = -\frac{1}{2}nlog(2\pi) - nlog(\sigma) \\
&- \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i^2 - 2\beta_0 y_i - 2\beta_1 y_i x_i + \beta_0^2 + 2\beta_0\beta_1 x_i + \beta_1^2 x_i^2 + \beta_1^2\sigma_u^2) \qquad (61)
\end{aligned}
$$

Durch den Vergleich der Formel 54 und 61 lässt sich der Korrekturfaktor auf $\xi = -\beta_1^2\sigma_u^2$ bestimmen.
Schritt 4: Die korrigierte Likelihood ergibt sich also zu:

$$
\begin{aligned}
l_c(Y, X^*, \beta) = &-\frac{1}{2}nlog(2\pi) - nlog(\sigma) \\
&- \frac{1}{2\sigma^2}\sum_{i=1}^{n} y_i^2 - 2\beta_0 y_i - 2\beta_1 y_i x_i^* + \beta_0^2 + 2\beta_0\beta_1 x_i^* + \beta_1^2 x_i^{*2} + \underbrace{(-\beta_1^2\sigma_u^2)}_{\xi} \qquad (62)
\end{aligned}
$$

und die korrigierten Score Funktionen zu:

$$
\frac{\partial l_c}{\partial \beta_0} = S_{c0}(Y, X^*, \beta) = -\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i^*) \qquad (63)
$$

$$
\frac{\partial l_c}{\partial \beta_1} = S_{c1}(Y, X^*, \beta) = -\frac{1}{\sigma^2}(\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i^*)x_i^* + n\beta_1\sigma_u^2) \qquad (64)
$$

C.1.2 Einfache Poisson Regression

Modellgleichung: $\mu_i = \lambda_i = \exp(\beta_0 + \beta_1 x_i) = \exp(\underline{x_i}^t\beta)$ mit $E[y_i|x_i] = \lambda_i$ und $Y|\lambda \sim Po(\lambda)$
Wahrscheinlichkeitsfunktion: $f(y_i, \underline{x_i}, \beta) = \frac{\lambda_i^{y_i}}{y_i!}\exp(-\lambda_i)$ Schritt 1: Aufstellen der Log-Likelihood Funktion für wahre $\underline{x_i}$

$$
\begin{aligned}
l(Y, X, \beta) &= \sum_{i=1}^{n} -\exp(\underline{x_i}^t \beta) + y_i \underline{x_i}^t \beta - log(y_i!) \\
&= \sum_{i=1}^{n} -\exp(\beta_0 + \beta_1 x_i) + y_i \beta_0 + y_i x_i \beta_1 - log(y_i!)
\end{aligned} \tag{65}
$$

Daraus lassen sich die wahren Score Funktionen ableiten zu:

$$
\frac{\partial l}{\partial \beta_0} = S_0(Y, X, \beta) = \sum_{i=1}^{n} (y_i - \exp(\beta_0 + \beta_1 x_i)) \tag{66}
$$

$$
\frac{\partial l}{\partial \beta_1} = S_1(Y, X, \beta) = \sum_{i=1}^{n} (y_i - \exp(\beta_0 + \beta_1 x_i)) x_i \tag{67}
$$

Schritt 2: Aufstellen der Log-Likelihood Funktion für beobachtete bzw. fehlerhaft gemessenen $\underline{x}_i^*$ auf:

$$
\begin{aligned}
l(Y, X^*, \beta) &= \sum_{i=1}^{n} -\exp(\beta_0 + \beta_1 x_i^*) + y_i \beta_0 + y_i x_i^* \beta_1 - log(y_i!) \\
&= \sum_{i=1}^{n} -\exp(\beta_0 + \beta_1 (x_i + u_i)) + y_i \beta_0 + y_i (x_i + u_i) \beta_1 - log(y_i!) \\
&= \sum_{i=1}^{n} -\exp(\beta_0 + \beta_1 x_i + \beta_1 u_i) + y_i \beta_0 + y_i x_i \beta_1 + y_i u_i \beta_1 - log(y_i!)
\end{aligned} \tag{68}
$$

Schritt 3: Berechnen den Erwartungswert $E[l(Y, X^*, \beta)|y_i, x_i]$, wobei folgende Transformationen gelten.

$$
\begin{aligned}
&U \sim N(0, \sigma_u^2) \\
&(\beta_1 U) \sim N(0, \beta_1^2 \sigma_u^2) \\
&\tilde{U} = exp(\beta_1 U) \sim LN(0, \beta_1^2 \sigma_u^2) \\
&E[\tilde{U}] = exp(0 + \frac{1}{2} \beta_1^2 \sigma^2)
\end{aligned} \tag{69}
$$

$$
\begin{aligned}
&E[l(Y, X^*, \beta)] \\
&= E[\sum_{i=1}^{n} -\exp(\beta_0 + \beta_1 x_i + \beta_1 u_i) + y_i\beta_0 + y_i x_i \beta_1 + y_i u_i \beta_1 - log(y_i!)|y_i, x_i] \\
&= \sum_{i=1}^{n} E[-\exp(\beta_0 + \beta_1 x_i)|y_i, x_i] E[\exp(\beta_1 u_i)|y_i, x_i] + E[y_i\beta_0|y_i, x_i] \\
&+ E[y_i x_i \beta_1|y_i, x_i] + E[y_i u_i \beta_1|y_i, x_i] - E[log(y_i!)|y_i, x_i] \qquad (70) \\
&= \sum_{i=1}^{n} -\exp(\beta_0 + \beta_1 x_i) \exp(\frac{1}{2}\beta_1^2\sigma_u^2) + y_i\beta_0 + y_i x_i \beta_1 - log(y_i!) \\
&= \sum_{i=1}^{n} -\exp(\beta_0 + \beta_1 x_i + \frac{1}{2}\beta_1^2\sigma_u^2) + y_i\beta_0 + y_i x_i \beta_1 - log(y_i!) \qquad (71)
\end{aligned}
$$

Durch den Vergleich der Formel 65 und 71 lässt sich der Korrekturfaktor auf $\xi = -\frac{\beta_1^2}{2}\sigma_u^2$ bestimmen.
Schritt 4: Die korrigierte Likelihood ergibt sich also zu:

$$
l_c(Y, X^*, \beta) = \sum_{i=1}^{n} -\exp(\beta_0 + \beta_1 x_i^* + \underbrace{(-\frac{1}{2}\beta_1^2\sigma_u^2)}_{\xi}) + y_i\beta_0 + y_i x_i^* \beta_1 - log(y_i!) \qquad (72)
$$

und die korrigierten Score Funktionen zu:

$$
\frac{\partial l_c}{\partial \beta_0} = S_{c0}(Y, X^*, \beta) = \sum_{i=1}^{n} y_i - \exp(\beta_0 + \beta_1 x_i^* - \frac{1}{2}\beta_1^2\sigma_u^2) \qquad (73)
$$

$$
\frac{\partial l_c}{\partial \beta_1} = S_{c1}(Y, X^*, \beta) = \sum_{i=1}^{n} (y_i x_i^* - \exp(\beta_0 + \beta_1 x_i^* - \frac{1}{2}\beta_1^2\sigma_u^2)(x_i^* - \beta_1\sigma_u^2)) \qquad (74)
$$

C.2 Details zur approximativen korrigierten Score Funktion

C.2.1 Beweise zur approximativen korrigierten Score Funktion

Der Beweis zu Lemma 1 ist in ausführlicher Form in Buzas [2009] im Appendix. Die Beweise zu Lemma 2 und Proposition 1 sind in Buzas [2009, S.2353 ff] in Kurzform dargestellt, ausführlicher werden die Beweise in diesem Kapitel dargelegt.

Beweis zu Lemma 2:

setze $v = (\beta_0 + \beta_z^t Z + \beta_x X)$ Umformung der linken Seite:

$$
\begin{aligned}
&E[g(X^*)|Y,X,Z] \\
&= E[g_1(X^*)[2Y-1] + g_2(X^*)[Y-1] + g_3(X^*)Y|Y,X,Z] \\
&= E[2Yg_1(X^*) - g_1(X^*) + Yg_2(X^*) - g_2(X^*) + Yg_3(X^*)|Y,X,Z] \\
&= E[2Yg_1(X^*)|Y,X,Z] - E[g_1(X^*)|Y,X,Z] + E[Yg_2(X^*)|Y,X,Z] \\
&- E[g_2(X^*)|Y,X,Z] + E[Yg_3(X^*)|Y,X,Z] \\
&= 2YE[g_1(X^*)|Y,X,Z] - E[g_1(X^*)|Y,X,Z] + YE[g_2(X^*)|Y,X,Z] \\
&- E[g_2(X^*)|Y,X,Z] + YE[g_3(X^*)|Y,X,Z] \\
&= 2Y\Phi(\lambda(\beta_0 + \beta_z^t Z + \beta_x X)) - \Phi(\lambda(\beta_0 + \beta_z^t Z + \beta_x X)) \\
&+ Y\Phi(\lambda v)\exp(v) - \Phi(\lambda v)\exp(v) + Y\Phi(\lambda v)\exp(-v) \\
&= \Phi(\lambda v)\Big[2Y - 1 + Y - \exp(v) + Y\exp(-v)\Big] \\
&= \Phi(\lambda v)\Big[Y(2 + \exp(v) + \exp(-v)) - (1 + \exp(v))\Big] \\
&= \Phi(\lambda v)\Big[Y(2 + \exp(v) + \exp(-v))\Big] - \Phi(\lambda v)\Big[(1 + \exp(v))\Big] \qquad (75)
\end{aligned}
$$

Umformung der rechten Seite:

$$
\begin{aligned}
&\tilde{h}(v)[Y - F(v)] \\
&\text{es gilt in Allgemeinen } F'(v) \approx \lambda\Phi(\lambda v) \\
&= \frac{\Phi(\lambda v)}{F(v)(1-F(v))}[Y - F(v)] \\
&= Y\frac{\Phi(\lambda v)}{F(v)(1-F(v))} - \frac{\Phi(\lambda v)}{(1-F(v))} \\
&= Y\frac{\Phi(\lambda v)}{\frac{1}{1+\exp(-v)}(1 - \frac{1}{1+\exp(-v)})} - \frac{\Phi(\lambda v)}{(1 - \frac{1}{1+\exp(-v)})} \\
&= Y\frac{\Phi(\lambda v)}{\frac{1}{1+\exp(-v)}(\frac{\exp(-v)}{1+\exp(-v)})} - \frac{\Phi(\lambda v)}{(\frac{\exp(-v)}{1+\exp(-v)})} \\
&= Y\Phi(\lambda v)(1+\exp(-v))^2\exp(v) - \Phi(\lambda v)\exp(v)(1+\exp(-v)) \\
&= Y\Phi(\lambda v)(1 + 2\exp(-v) + \exp(-v)^2)\exp(v) - \Phi(\lambda v)(\exp(v)+1) \\
&= Y\Phi(\lambda v)(\exp(v) + 2 + \exp(-v)) - \Phi(\lambda v)(\exp(v)+1) \qquad (76)
\end{aligned}
$$

Einen vergleich von (75) und (76) zeigt die Gleichheit.

Beweis zu Proportion 1:
Für den Beweis der Erwartungstreue wird die Identität $E[g(X^*)(X^* - X)|Y,X,Z] =$

$\sigma_u^2 E[g'(X^*)|Y,X,Z]$ aus Stein,1981, pp 1136-37 zur Hilfe genommen. Die Umformung

$$\begin{aligned}
E[g(X^*)(X^*-X)|Y,X,Z] &= \sigma_u^2 E[g'(X^*)|Y,X,Z] \\
E[g(X^*)X^* - g(X^*)X|Y,X,Z] &= E[\sigma_u^2 g'(X^*)|Y,X,Z] \\
E[g(X^*)X^*|Y,X,Z] - E[g(X^*)X|Y,X,Z] &= E[\sigma_u^2 g'(X^*)|Y,X,Z] \\
E[g(X^*)X|Y,X,Z] &= E[g(X^*)X^*|Y,X,Z] - E[\sigma_u^2 g'(X^*)|Y,X,Z] \\
E[g(X^*)X|Y,X,Z] &= E[g(X^*)X^* - \sigma_u^2 g'(X^*)|Y,X,Z] \qquad (77)
\end{aligned}$$

liefert die entscheidende Formel (77).

$$\begin{aligned}
E[S_{Ac}(Y,X^*,Z,\beta)|Y,X,Z] &= E\left[g(X^*) \begin{pmatrix} 1 \\ Z \\ X^* - \sigma_u^2 \frac{g'(X^*)}{g(X^*)} \end{pmatrix} |Y,X,Z \right] \\
&= E\left[\begin{pmatrix} g(X^*) \\ Zg(X^*) \\ X^*g(X^*) - \sigma_u^2 g'(X^*) \end{pmatrix} |Y,X,Z \right] \\
&= \begin{pmatrix} E[g(X^*)|Y,X,Z] \\ E[Zg(X^*)|Y,X,Z] \\ E[X^*g(X^*) - \sigma_u^2 g'(X^*)|Y,X,Z] \end{pmatrix} \\
&= \begin{pmatrix} E[g(X^*)|Y,X,Z] \\ ZE[g(X^*)|Y,X,Z] \\ E[g(X^*)X|Y,X,Z] \end{pmatrix} \\
&= \begin{pmatrix} \tilde{h}(\beta_0+\beta_z Z+\beta_x X)[Y - F(\beta_0+\beta_z Z+\beta_x X)] \\ Z * \tilde{h}(\beta_0+\beta_z Z+\beta_x X)[Y - F(\beta_0+\beta_z Z+\beta_x X)] \\ X * \tilde{h}(\beta_0+\beta_z Z+\beta_x X)[Y - F(\beta_0+\beta_z Z+\beta_x X)] \end{pmatrix} \\
&= \tilde{h}(\beta_0+\beta_z Z+\beta_x X)[Y - F(\beta_0+\beta_z Z+\beta_x X)] \begin{pmatrix} 1 \\ Z \\ X \end{pmatrix} \\
&= S_A(Y,X,Z,\beta)
\end{aligned}$$

C.2.2 Formeln zur approximativ korrigierten Score Funktion

Für die Aufstellung der approximativen korrigieren Score Funktion der logistischen Regression werden Lemma 1 und Lemma 2 benötigt, eine detaillierte Form dieser Lemmas werden in diesem Teil dargestellt, wobei Notationen aus dem Buzas [2009] übernommen werden:

$$
\begin{aligned}
k &= \lambda/\sqrt{1-\beta_x^2\sigma^2\lambda^2} \\
g'(X^*) &= \frac{d}{dX^*}g(X^*) \\
\eta &= \exp(-0.5\frac{k}{\lambda}\beta_x^2\sigma^2) \\
g_1(X^*) &= \frac{k}{\lambda}\Phi(k(\beta_0+\beta_z^tZ+\beta_xX^*)) \\
&= \frac{k}{\lambda} * \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{1}{2}(k^2(\beta_0+\beta_z^tZ+\beta_xX^*)^2) \\
g_1'(X^*) &= \frac{-k^3}{\lambda\sqrt{2\pi}\sigma}\exp(\frac{-1}{2}k^2(\beta_0+\beta_z^tZ+\beta_xX^*)^2)(\beta_0\beta_x+\beta_z^t\beta_xZ+\beta_x^2X^*) \\
\gamma(X^*) &= \exp(\frac{k^2}{\lambda^2}(\beta_0+\beta_z^tZ+\beta_xX^*)) \\
\gamma'(X^*) &= \exp(\frac{k^2}{\lambda^2}(\beta_0+\beta_z^tZ+\beta_xX^*))(\frac{k^2}{\lambda^2}\beta_x) \\
g_2(X^*) &= \eta\gamma(X^*)g_1(X^*) \\
&= \exp(-0.5\frac{k^2}{\lambda^2}\beta_x^2\sigma^2) * \exp(\frac{k^2}{\lambda^2}(\beta_0+\beta_z^tZ+\beta_xX^*)) \\
&* \frac{k}{\lambda}\frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{1}{2}(k^2(\beta_0+\beta_z^tZ+\beta_xX^*)^2) \\
g_2'(X^*) &= (\eta\gamma(X^*)g_1(X^*))' = \eta(\gamma(X^*)g_1(X^*))' = \eta(\gamma'(X^*)g_1(X^*)+\gamma(X^*)g_1'(X^*)) \\
&= \exp(-0.5\frac{k^2}{\lambda^2}\beta_x^2\sigma^2)\Big[\exp(\frac{k^2}{\lambda^2}(\beta_0+\beta_z^tZ+\beta_xX^*))(\frac{k^2}{\lambda^2}\beta_x) \\
&* \frac{k}{\lambda\sqrt{2\pi}\sigma}\exp(-\frac{1}{2}(k^2(\beta_0+\beta_z^tZ+\beta_xX^*)^2)+\exp(\frac{k^2}{\lambda^2}(\beta_0+\beta_z^tZ+\beta_xX^*)) \\
&+ \frac{-k^3}{\lambda\sqrt{2\pi}\sigma}\exp(\frac{-1}{2}k^2(\beta_0+\beta_z^tZ+\beta_xX^*)^2)(\beta_0\beta_x+\beta_z^t\beta_xZ+\beta_x^2X^*)\Big] \\
g_3(X^*) &= \frac{g_2(X^*)}{\gamma(X^*)} = \frac{\eta\gamma(X^*)g_1(X^*)}{\gamma^2(X^*)} = \eta g_1(X^*)\gamma(X^*)^{-1} \\
&= \exp(-0.5\frac{k}{\lambda}\beta_x^2\sigma^2) * \frac{k}{\lambda}\frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{1}{2}(k^2(\beta_0+\beta_z^tZ+\beta_xX^*)^2) \\
&* \exp(-\frac{k^2}{\lambda^2}(\beta_0+\beta_z^tZ+\beta_xX^*))
\end{aligned}
$$

$$
\begin{aligned}
g_3'(X^*) &= (\frac{g_2(X^*)}{\gamma^2(X^*)})' = (\frac{\eta\gamma(X^*)g_1(X^*)}{\gamma^2(X^*)})' = \eta\frac{(g_1'(X^*)\gamma(X^*) - g_1(X^*)\gamma'(X^*))}{\gamma^2(X^*)} \\
&= \exp(-0.5 * \frac{k^2}{\lambda^2}\beta_x^2\sigma^2) * \frac{-k^3}{\lambda\sqrt{2\pi}\sigma}\exp(\frac{-1}{2}k^2(\beta_0 + \beta_z^t Z + \beta_x X^*)^2) \\
&\quad (\beta_0\beta_x + \beta_z^t\beta_x Z + \beta_x^2 X^*) * \\
g(X^*) &= g_1(X^*)[2Y-1] + g_2(X^*)[Y-1] + g_3(X^*)Y \\
&= g_1(X^*)[2Y-1] + \eta\gamma(X^*)g_1(X^*)[Y-1] + \eta Y g_1(X^*)\gamma(X^*)^{-1} \\
&= g_1(X^*)2Y - g_1(X^*) + \eta\gamma(X^*)g_1(X^*)Y - \eta\gamma(X^*)g_1(X^*) + \eta Y g_1(X^*)\gamma(X^*)^{-1} \\
&= g_1(X^*)\Big[Y(2 + \eta\gamma(X^*)^{-1}) - 1 + \eta\gamma(X^*)(Y-1)\Big] \\
&= \frac{k}{\lambda}\frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{1}{2}(k^2(\beta_0 + \beta_z^t Z + \beta_x X^*)^2)[Y(2 + \frac{\exp(-0.5\frac{k}{\lambda}\beta_x^2\sigma^2)}{\exp(\frac{k^2}{\lambda^2}(\beta_0 + \beta_z^t Z + \beta_x X^*))}) - 1 \\
&\quad + \exp(-0.5\frac{k}{\lambda}\beta_x^2\sigma) * \exp(\frac{k^2}{\lambda^2}(\beta_0 + \beta_z^t Z + \beta_x X^*)) * (Y-1)] \\
g'(X^*) &= g_1'(X^*)2Y - g_1'(X^*) + \eta(\gamma(X^*)g_1(X^*))'Y - \eta(\gamma(X^*)g_1(X^*))' + \eta Y(\frac{g_1(X^*)}{\gamma(X^*)})' \\
&= g_1'(X^*)2Y - g_1'(X^*) + \eta Y(\gamma'(X^*)g_1(X^*) + \gamma(X^*)g_1'(X^*)) \\
&\quad - \eta(\gamma'(X^*)g_1(X^*) + \gamma(X^*)g_1'(X^*)) + \eta Y(\frac{\gamma(X^*)g_1(X^*)' - g_1(X^*)\gamma(X^*)'}{\gamma(X^*)^2}) \\
&= g_1'(X^*)2Y - g_1'(X^*) + \eta Y\gamma'(X^*)g_1(X^*) + \eta Y\gamma(X^*)g_1'(X^*) \\
&\quad - \eta\gamma'(X^*)g_1(X^*) + \eta\gamma(X^*)g_1'(X^*)) + \eta Y\frac{g_1'(X^*)}{\gamma(X^*)} - \eta Y\frac{g_1'(X^*)\gamma'(X^*)}{\gamma^2(X^*)} \\
&= g_1'(X^*)[2Y - 1 + \eta Y\gamma(X^*) + \eta\gamma(X^*) + \frac{\eta Y}{\gamma(X^*)})] + g_1(X^*)[\eta\gamma'(X^*)(Y-1) - \frac{Y}{\gamma^2(X^*)}] \\
&= g_1'(X^*)[2Y - 1 + \eta(Y\gamma(X^*) + \gamma(X^*) + \frac{Y}{\gamma(X^*)})] + g_1(X^*)[\eta\gamma'(X^*)(Y-1) - \frac{Y}{\gamma^2(X^*)}] \\
&= \frac{-k^3}{\lambda\sqrt{2\pi}\sigma}\exp(\frac{-1}{2}k^2(\beta_0 + \beta_z^t Z + \beta_x X^*)^2)(\beta_0\beta_x + \beta_z^t\beta_x Z + \beta_x^2 X^*) \\
&\quad \Big[2Y - 1 + \exp(-0.5\frac{k}{\lambda}\beta_x^2\sigma^2)(Y * \exp(\frac{k^2}{\lambda^2}(\beta_0 + \beta_z^t Z + \beta_x X^*)) + \exp(\frac{k^2}{\lambda^2}(\beta_0 + \beta_z^t Z + \beta_x X^*)) \\
&\quad + \frac{Y}{\exp(\frac{k^2}{\lambda^2}(\beta_0 + \beta_z^t Z + \beta_x X^*))})\Big] + \frac{k}{\lambda} * \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{1}{2}(k^2(\beta_0 + \beta_z^t Z + \beta_x X^*)^2) \\
&\quad \Big[\exp(-0.5\frac{k}{\lambda}\beta_x^2\sigma^2) * \exp(\frac{k^2}{\lambda^2}(\beta_0 + \beta_z^t Z + \beta_x X^*))(\frac{k^2}{\lambda^2}\beta_x) * (Y-1) - \frac{Y}{\exp(\frac{k^2}{\lambda^2}(\beta_0 + \beta_z^t Z + \beta_x X^*))^2}\Big]
\end{aligned}
$$

C.3 Herleitung exakter korrigierte Monte-Carlo Score Funktion

C.3.1 Lineare Regression

Zunächst wird in die wahre Score Funktionen ((55), (56)) $\tilde{X}_b$ eingesetzt:
Für β_0:

$$\begin{aligned} & S_0(Y, \tilde{X}_b, \beta) \\ &= \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 \tilde{x}_{b,i}) \\ &= \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i^* - \sqrt{-1}\beta_1 u_{b,i}) \end{aligned} \tag{78}$$

Für β_1:

$$\begin{aligned} & S_1(Y, \tilde{X}_b, \beta) \\ &= \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - (x_i^* + \sqrt{-1}u_{b,i})\beta_1)(x_i^* + \sqrt{-1}u_{b,i}) \\ &= \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - x_i^*\beta_1 - \sqrt{-1}u_{b,i}\beta_1)(x_i^* + \sqrt{-1}u_{b,i}) \\ &= \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i x_i^* - x_i^*\beta_0 - x_i^{*2}\beta_1 - \sqrt{-1}u_{b,i}(x_i^*\beta_1 - y_i + \beta_0 + x_i^*\beta_1) - \sqrt{-1}^2 u_{b,i}^2\beta_1) \\ &= \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i x_i^* - x_i^*\beta_0 - x_i^{*2}\beta_1 + u_{b,i}^2\beta_1 - \sqrt{-1}u_{b,i}(x_i^*\beta_1 - y_i + \beta_0 + \beta_1 x_i^*) \end{aligned} \tag{79}$$

der Realteil ist jeweils

$$\begin{aligned} Re(S_0(Y, \tilde{X}_b, \beta)) &= \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i^*) \\ Re(S_1(Y, \tilde{X}_b, \beta)) &= \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i x_i^* - x_i\beta_0 - x_i^{*2}\beta_1 + u_{b,i}^2\beta_1) \end{aligned}$$

und die Summe über B bzw. die korrigierte Monte Carlo Scores lassen sich vereinfachen zu

$$S_{MCc0,B}(Y, \tilde{X}^*, \beta) = \frac{1}{B} \sum_{b=1}^{B} Re(S_0(Y, \tilde{X}_b, \beta)) = \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i^*) \tag{80}$$

und

$$
\begin{aligned}
S_{MCc1,B}(Y, \tilde{X}^*, \beta) &= \frac{1}{B}\sum_{b=1}^{B} Re(S_1(Y, \tilde{X}_b, \beta)) \\
&= \frac{1}{B}\sum_{b=1}^{B}(\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i x_i^* - \beta_0 x_i^* - x_i^{*2}\beta_1) + u_{b,i}^2\beta_1) \\
&= \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i x_i^* - \beta_0 x_i^* - x_i^{*2}\beta_1) + \frac{1}{\sigma^2}\sum_{i=1}^{n}\frac{1}{B}\sum_{b=1}^{B} u_{b,i}^2\beta_1 \\
&= \underbrace{\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i x_i^* - \beta_0 x_i^* - x_i^{*2}\beta_1)}_{S_1(y_i, x_i^*, \beta)} + \frac{1}{\sigma^2}\sum_{i=1}^{n}\beta_1\frac{1}{B}\sum_{b=1}^{B} u_{b,i}^2
\end{aligned} \tag{81}
$$

C.3.2 Poisson Regression

Zunächst wird in die wahre Score Funktionen ((66), (67)) $\tilde{X}_b$ eingesetzt:
Für β_0:

$$
\begin{aligned}
S_0(Y, \tilde{X}_b, \beta) &= -\sum_{i=1}^{n}(\exp(\beta_0 + \beta_1\tilde{X}_b) - y_i) \\
&= -\sum_{i=1}^{n}(\exp(\beta_0 + \beta_1(x_i^* + \sqrt{-1}u_{b,i})) - y_i) \\
&= -\sum_{i=1}^{n}(\exp(\beta_0 + \beta_1 x_i^* + \beta_1\sqrt{-1}u_{b,i}) - y_i) \\
&= -\sum_{i=1}^{n}(\exp(\beta_0 + \beta_1 x_i^*)\exp(\beta_1\sqrt{-1}u_{b,i}) - y_i) \\
&= -\sum_{i=1}^{n}(\exp(\beta_0 + \beta_1 x_i^*)(cos(\beta_1 u_{b,i}) + \sqrt{-1}sin(\beta_1 u_{b,i})) - y_i) \\
&= -\sum_{i=1}^{n}(\exp(\beta_0 + \beta_1 x_i^*)(cos(\beta_1 u_{b,i}) + \sqrt{-1}\exp(\beta_0 + \beta_1 x_i^*)sin(\beta_1 u_{b,i}) - y_i)
\end{aligned} \tag{82}
$$

Für β_1:

$$
\begin{aligned}
S_1(Y, \tilde{X}_b, \beta) &= -\sum_{i=1}^{n}(\exp(\beta_0 + \beta_1 \tilde{x}_{b,i})\tilde{x}_{b,i} - y_i \tilde{x}_{b,i}) \\
&= -\sum_{i=1}^{n}(\exp(\beta_0 + \beta_1 (x_i^* + \sqrt{-1}u_{b,i}))(x_i^* + \sqrt{-1}u_{b,i}) - y_i(x_i^* + \sqrt{-1}u_{b,i})) \\
&= -\sum_{i=1}^{n}(\exp(\beta_0 + \beta_1 x_i^* + \sqrt{-1}\beta_1 u_{b,i})(x_i^* + \sqrt{-1}u_{b,i}) - y_i x_i^* - \sqrt{-1}y_i u_{b,i}) \\
&= -\sum_{i=1}^{n}(\exp(\beta_0 + \beta_1 x_i^*)\exp(\sqrt{-1}\beta_1 u_{b,i})(x_i^* + \sqrt{-1}u_{b,i}) - y_i x_i^* - \sqrt{-1}y_i u_{b,i}) \\
&= -\sum_{i=1}^{n}(\exp(\beta_0 + \beta_1 x_i^*)\exp(\sqrt{-1}\beta_1 u_{b,i})x_i^* \\
&+ \sqrt{-1}\exp(\beta_0 + \beta_1 x_i^*)\exp(\sqrt{-1}\beta_1 u_{b,i})u_{b,i} - y_i x_i^* - \sqrt{-1}y_i u_{b,i}) \\
&= -\sum_{i=1}^{n}(\exp(\beta_0 + \beta_1 x_i^*)(cos(\beta_1 u_{b,i}) + \sqrt{-1}sin(\beta_1 u_{b,i}))x_i^* \\
&+ \sqrt{-1}\exp(\beta_0 + \beta_1 x_i^*)(cos(\beta_1 u_{b,i}) + \sqrt{-1}sin(\beta_1 u_{b,i}))u_{b,i} - y_i x_i^* - \sqrt{-1}y_i u_{b,i}) \\
&= -\sum_{i=1}^{n}(\exp(\beta_0 + \beta_1 x_i^*)cos(\beta_1 u_{b,i})x_i^* + \sqrt{-1}\exp(\beta_0 + \beta_1 x_i^*)sin(\beta_1 u_{b,i})x_i^* \\
&+ \sqrt{-1}\exp(\beta_0 + \beta_1 x_i^*)cos(\beta_1 u_{b,i})u_{b,i} \\
&+ \sqrt{-1}^2\exp(\beta_0 + \beta_1 x_i^*)sin(\beta_1 u_{b,i}))u_{b,i} - y_i x_i^* - \sqrt{-1}y_i u_{b,i}) \\
&= -\sum_{i=1}^{n}(\exp(\beta_0 + \beta_1 x_i^*)(cos(\beta_1 u_{b,i})x_i^* - sin(\beta_1 u_{b,i})u_{b,i}) - y_i x_i^* \\
&+ \sqrt{-1}(\exp(\beta_0 + \beta_1 x_i^*)(sin(\beta_1 u_{b,i})x_i^* + cos(\beta_1 u_{b,i})u_{b,i}) - y_i u_{b,i})) \qquad (83)
\end{aligned}
$$

Der Realteil ist jeweils

$$Re(S_0(Y, \tilde{X}_b, \beta)) = -\sum_{i=1}^{n}(\exp(\beta_0 + \beta_1 x_i^*)(cos(\beta_1 u_{b,i}) - y_i) \qquad (84)$$

$$Re(S_1(Y, \tilde{X}_b, \beta)) = -\sum_{i=1}^{n}(\exp(\beta_0 + \beta_1 x_i^*)(cos(\beta_1 u_{b,i})x_i^* - sin(\beta_1 u_{b,i})u_{b,i}) - y_i x_i^*) \qquad (85)$$

und die Summe über B bzw. die korrigierte Monte Carlo Scores lassen sich vereinfachen zu

$$\begin{aligned}
S_{MCc0,B}(Y, \tilde{X}^*, \beta) &= \frac{1}{B}\sum_{b=1}^{B} Re(S_0(Y, \tilde{X}_b, \beta)) \\
&= -\frac{1}{B}\sum_{b=1}^{B}\sum_{i=1}^{n}(\exp(\beta_0 + \beta_1 x_i^*)(cos(\beta_1 u_{b,i}) - yi) \\
&= -\sum_{i=1}^{n}(\exp(\beta_0 + \beta_1 x_i^*)(\frac{1}{B}\sum_{b=1}^{B} cos(\beta_1 u_{b,i})) - yi) \\
S_{MCc1,B}(Y, \tilde{X}^*, \beta) &= \frac{1}{B}\sum_{b=1}^{B} Re(S_1(Y, \tilde{X}_b, \beta)) \\
&= -\frac{1}{B}\sum_{b=1}^{B}\sum_{i=1}^{n}(\exp(\beta_0 + \beta_1 x_i^*)(cos(\beta_1 u_{b,i})x_i^* - sin(\beta_1 u_{b,i})u_{b,i}) - y_i x_i^*) \qquad (86) \\
&= -\sum_{i=1}^{n}(\exp(\beta_0 + \beta_1 x_i^*)\frac{1}{B}\sum_{b=1}^{B}(cos(\beta_1 u_{b,i})x_i^* - sin(\beta_1 u_{b,i})u_{b,i}) - y_i x_i^*) \qquad (87)
\end{aligned}$$

C.3.3 Logistische Regression

Zunächst wird in die wahre Score Funktionen ((25), (26)) $\tilde{x}_{b,i}$ eingesetzt:
Für β_0:

$$\begin{aligned}
S_0(Y, \tilde{X}_b, \beta) &= -\sum_{i=1}^{n}(\exp(\beta_0 + \beta_1 \tilde{x}_{b,i}) - y_i) \\
&= -\sum_{i=1}^{n}(\exp(\beta_0 + \beta_1 (x_i^* + \sqrt{-1}u_{b,i})) - y_i) \\
&= -\sum_{i=1}^{n}(\exp(\beta_0 + \beta_1 x_i^* + \beta_1\sqrt{-1}u_{b,i}) - y_i) \\
&= -\sum_{i=1}^{n}(\exp(\beta_0 + \beta_1 x_i^*)\exp(\beta_1\sqrt{-1}u_{b,i}) - y_i) \\
&= -\sum_{i=1}^{n}(\exp(\beta_0 + \beta_1 x_i^*)(cos(\beta_1 u_{b,i}) + \sqrt{-1}sin(\beta_1 u_{b,i})) - y_i) \\
&= -\sum_{i=1}^{n}(\exp(\beta_0 + \beta_1 x_i^*)(cos(\beta_1 u_{b,i}) + \sqrt{-1}\exp(\beta_0 + \beta_1 x_i^*)sin(\beta_1 u_{b,i}) - y_i) \qquad (88)
\end{aligned}$$

Für β_1:

$$
\begin{aligned}
S_1(Y, \tilde{X}_b, \beta) &= -\sum_{i=1}^{n}(\exp(\beta_0 + \beta_1 \tilde{x}_{b,i})\tilde{x}_{b,i} - y_i \tilde{x}_{b,i}) \\
&= -\sum_{i=1}^{n}(\exp(\beta_0 + \beta_1 (x_i^* + \sqrt{-1}u_{b,i}))(x_i^* + \sqrt{-1}u_{b,i}) - y_i(x_i^* + \sqrt{-1}u_{b,i})) \\
&= -\sum_{i=1}^{n}(\exp(\beta_0 + \beta_1 x_i^* + \sqrt{-1}\beta_1 u_{b,i})(x_i^* + \sqrt{-1}u_{b,i}) - y_i x_i^* - \sqrt{-1}y_i u_{b,i}) \\
&= -\sum_{i=1}^{n}(\exp(\beta_0 + \beta_1 x_i^*)\exp(\sqrt{-1}\beta_1 u_{b,i})(x_i^* + \sqrt{-1}u_{b,i}) - y_i x_i^* - \sqrt{-1}y_i u_{b,i}) \\
&= -\sum_{i=1}^{n}(\exp(\beta_0 + \beta_1 x_i^*)\exp(\sqrt{-1}\beta_1 u_{b,i})x_i^* \\
&+ \sqrt{-1}\exp(\beta_0 + \beta_1 x_i^*)\exp(\sqrt{-1}\beta_1 u_{b,i})u_{b,i} - y_i x_i^* - \sqrt{-1}y_i u_{b,i}) \\
&= -\sum_{i=1}^{n}(\exp(\beta_0 + \beta_1 x_i^*)(cos(\beta_1 u_{b,i}) + \sqrt{-1}sin(\beta_1 u_{b,i}))x_i^* \\
&+ \sqrt{-1}\exp(\beta_0 + \beta_1 x_i^*)(cos(\beta_1 u_{b,i}) + \sqrt{-1}sin(\beta_1 u_{b,i}))u_{b,i} - y_i x_i^* - \sqrt{-1}y_i u_{b,i}) \\
&= -\sum_{i=1}^{n}(\exp(\beta_0 + \beta_1 x_i^*)cos(\beta_1 u_{b,i})x_i^* + \sqrt{-1}\exp(\beta_0 + \beta_1 x_i^*)sin(\beta_1 u_{b,i})x_i^* \\
&+ \sqrt{-1}\exp(\beta_0 + \beta_1 x_i^*)cos(\beta_1 u_{b,i})u_{b,i} \\
&+ \sqrt{-1}^2\exp(\beta_0 + \beta_1 x_i^*)sin(\beta_1 u_{b,i}))u_{b,i} - y_i x_i^* - \sqrt{-1}y_i u_{b,i}) \\
&= -\sum_{i=1}^{n}(\exp(\beta_0 + \beta_1 x_i^*)(cos(\beta_1 u_{b,i})x_i^* - sin(\beta_1 u_{b,i})u_{b,i}) - y_i x_i^* \\
&+ \sqrt{-1}(\exp(\beta_0 + \beta_1 x_i^*)(sin(\beta_1 u_{b,i})x_i^* + cos(\beta_1 u_{b,i})u_{b,i}) - y_i u_{b,i})) \qquad (89)
\end{aligned}
$$

Der Realteil ist jeweils

$$Re(S_0(Y, \tilde{X}_b, \beta)) = -\sum_{i=1}^{n}(\exp(\beta_0 + \beta_1 x_i^*)(cos(\beta_1 u_{b,i}) - y_i) \qquad (90)$$

$$Re(S_1(Y, \tilde{x}_b, \beta)) = -\sum_{i=1}^{n}(\exp(\beta_0 + \beta_1 x_i^*)(cos(\beta_1 u_{b,i})x_i^* - sin(\beta_1 u_{b,i})u_{b,i}) - y_i x_i^*) \qquad (91)$$

und die Summe über B bzw. die korrigierte Monte Carlo Scores lassen sich vereinfachen zu

$$
\begin{aligned}
S_{MCc0,B}(Y, \tilde{X}^*, \beta) &= \frac{1}{B} \sum_{b=1}^{B} Re(S_0(Y, \tilde{X}_b, \beta)) \\
&= -\frac{1}{B} \sum_{b=1}^{B} \sum_{i=1}^{n} (\exp(\beta_0 + \beta_1 x_i^*)(cos(\beta_1 u_{b,i}) - yi) \\
&= -\sum_{i=1}^{n} (\exp(\beta_0 + \beta_1 x_i^*)(\frac{1}{B} \sum_{b=1}^{B} cos(\beta_1 u_{b,i})) - yi) \\
S_{MCc1,B}(Y, \tilde{X}^*, \beta) &= \frac{1}{B} \sum_{b=1}^{B} Re(S_1(Y, \tilde{X}_b, \beta)) \\
&= -\frac{1}{B} \sum_{b=1}^{B} \sum_{i=1}^{n} (\exp(\beta_0 + \beta_1 x_i^*)(cos(\beta_1 u_{b,i}) x_i^* - sin(\beta_1 u_{b,i}) u_{b,i}) - y_i x_i^*) \qquad (92) \\
&= -\sum_{i=1}^{n} (\exp(\beta_0 + \beta_1 x_i^*) \frac{1}{B} \sum_{b=1}^{B} (cos(\beta_1 u_{b,i}) x_i^* - sin(\beta_1 u_{b,i}) u_{b,i}) - y_i x_i^*) \qquad (93)
\end{aligned}
$$

D Rcodes

Die für den Simulationsteil erstellten Rcodes befindet sich im elektronischen Anhang.

D.1 Regressionskalibrierung

Die kommentierte Regressionskalibrierungsfunktion befindet sich in der Datei Rk_fkt_source.R
Beschreibung
Gegeben beobachtete Y und fehlerhafte Messungen X^* schätzt die Rk.fkt() korrigierte β^c Schätzer und die zugehörige Schätzer-Varianz für das einfache lineare, logistische und Poisson Regression durch die drei Schritte des Regressionskalibrierungs-Algorithmus im Kapitel 3.1. Wenn Validierungsdaten vorliegen wird Schritt 1 des Algorithmus direkt mit lm() aus dem 'stats' Package geschätzt, bei vorliegen von Wiederholungsdaten durch die Formeln in Kapitel 3.3, d.h. insbesondere die Messfehlervarianz σ_u^2 wird geschätzt. Für die Varianz der resultierenden Schätzer wird das Resampling Vectors-Verfahren aus Appendix B.3.1 angewendet.

Tabelle 12: Input des Regressionskalibrierungsfunktion

Input	Erläuterung
datatype	character, "validData"falls Validierungsdaten vorliegen bzw. "wdhData"falls Wiederholungsdaten vorliegen
regtype	character, "linReg"falls lineare Regression gerechnet werden soll, "logReg"falls logistische bzw. "poisReg"falls eine poisson Regression berechnet werden soll
y	vector, beobachtete y-Werte
x.true	vector, falls vorhanden die fehlerfrei gemessenen x-Werten, default ist NULL
x.valid	vector, falls Validierungsdaten vorliegen, dann Vektor aus NA und x-Werten, je nachdem, wenn keine oder eine Validierungsmessung vorliegt, default ist NULL
xSt	matrix, beobachtete fehlerhafte Messungen x^*. Anzahl der Spalten entspricht Anzahl der Wiederholungsmessungen
varU	numeric, falls extern geschätzte Messfehlervarianz vorliegt, default NULL
bootstrap	numeric, Anzahl der Bootstrapp-Ziehung zur Schätzung der Varianz der geschätzten $\hat{\beta}$, default 100
showplot	logical, falls TRUE dann wird eine Graphik ausgegeben mit eingezeichnetem y, xSt, der naiven und der geschätzten Regressionsgerade und falls x.true vorliegt, zusätzlich x.true und die wahre Regressionsgerade

Die RK.fkt() gibt einen numerischen Vektor der geschätzten Parameter zurück

Tabelle 13: Output des Regressionskalibrierungsfunktion

Output	Erläuterung
beta0.rk	geschätztes $\hat{\beta}_0$ aus der Regressionskalibrierung
beta1.rk	geschätztes $\hat{\beta}_1$ aus der Regressionskalibrierung
var.beta0.rk	geschätztes $\hat{\sigma}_{\beta_0}$ aus dem Bootstep-Schritt der Regressionskalibrierung
var.beta1.rk	geschätztes $\hat{\sigma}_{\beta_1}$ aus dem Bootstep-Schritt der Regressionskalibrierung
beta0.bench	falls *x.true* vorliegt: geschätztes $\hat{\beta}_0$ aus lm() oder glm()
beta1.bench	falls *x.true* vorliegt: geschätztes $\hat{\beta}_1$ aus lm() oder glm()
beta0.naiv	mit der fehlerhaften Messung X^* geschätztem $\hat{\beta}_0$ aus lm() oder glm()
beta0.naiv	mit der fehlerhaften Messung X^* geschätztem $\hat{\beta}_1$ aus lm() oder glm()

D.2 Korrigierte Score Funktion

Die kommentierte korrigierte Score Funktion befindet sich in der Datei corScore_fkt_source.R
Beschreibung
Gegeben beobachtete Y und fehlerhafte Messungen Y^* schätzt die corScore.fkt() korrigierte β^c Schätzer für das einfache lineare, logistische und Poisson Regression. Für die lineare und Poisson Regression wird die Nullstellen der exakt korrigierte Score Funktion (Formel aus Kapitel 4.3) mithilfe von multiroot() aus dem "rootSolve"geschätzt, das die Newton-Raphson Methode benutzt. Für die logistische Regression wird die Nullstellen der approximativen korrigierten Score Funktion (Formeln aus Kapitel 4.4.2) ebenfalls mit multiroot() geschätzt. Konvergiert multiroot() nicht, so werden NA ausgegeben. Für beide Regressionen wird die Messfehlervarianz benötigt, die innerhalb der Funktion geschätzt wird oder extern übergeben werden kann, bzw. für vorliegen von nur einer Messwiederholung muss diese aus externen Daten geschätzt werden. Für die Schätzung werden je nach Datentyp die Formeln aus 4.6 herangezogen.

Tabelle 14: Input des korrigierten Score Funktion

Input	Erläuterung
datatype	character, "validData"falls Validierungsdaten vorliegen bzw. "wdhData"falls Wiederholungsdaten vorliegen
regtype	character, "linReg"falls lineare Regression gerechnet werden soll, "logReg"falls logistische bzw. "poisReg"falls eine Poisson Regression berechnet werden soll
y	vector, beobachtete y-Werte
x.true	vector, falls vorhanden die fehlerfrei gemessenen x-Werten, default ist NULL
x.valid	vector, falls Validierungsdaten vorliegen, dann Vektor aus NA und x-Werten, je nachdem, wenn keine oder eine Validierungsmessung vorliegt, default ist NULL
xSt	matrix, beobachtete fehlerhafte Messungen x^*. Anzahl der Spalten entspricht Anzahl der Wiederholungsmessungen
varU	numeric, falls extern geschätzte Messfehlervarianz vorliegt, default NULL
bootstrap	numeric, Anzahl der Bootstrapp-Ziehung zur Schätzung der Varianz der geschätzten $\hat{\beta}$, default 100
showplot	logical, falls TRUE dann wird eine Graphik ausgegeben mit eingezeichnetem y, xSt, der naiven und der geschätzten Regressionsgerade und falls x.true vorliegt, zusätzlich x.true und die wahre Regressionsgerade

Die corScore.fkt() gibt einen numerischen Vektor der geschätzten Parameter zurück

Tabelle 15: Output des korrigierten Score Funktion

Output	Erläuterung
beta0.rk	geschätztes $\hat{\beta}_0$ aus der korrigierten Score Funktion bzw. approximierten korrigierten Score Funktion für das logistische Modell
beta1.rk	geschätztes $\hat{\beta}_1$ aus der korrigierten Score Funktion bzw. approximierten korrigierten Score Funktion für das logistische Modell
beta0.bench	falls *x.true* vorliegt: geschätztes $\hat{\beta}_0$ aus lm() oder glm()
beta1.bench	falls *x.true* vorliegt: geschätztes $\hat{\beta}_1$ aus lm() oder glm()
beta0.naiv	mit der fehlerhaften Messung X^* geschätztem $\hat{\beta}_0$ aus lm() oder glm()
beta0.naiv	mit der fehlerhaften Messung X^* geschätztem $\hat{\beta}_1$ aus lm() oder glm()

D.3 Monte-Carlo korrigierte Score Funktion

Die kommentierte Monte-Carlo korrigierte Score Funktion befindet sich in der Datei MCcorScore_fkt_source.R

Beschreibung

Gegeben beobachtete Y und fehlerhafte Messungen X^* schätzt die MCcorScore.fkt() korrigierte β Schätzer für das einfache lineare und Poisson Regression. Mithilfe der Generation von komplexen Zufallszahlen wird eine korrigierte Score Funktion nach dem Algorithmus in 4.5.1 aufgestellt. Die Nullstellen dieser korrigierten Score Funktion wird mit der Funktion multiroot() aus dem "rootSolve"Packages geschätzt, das die Newton-Raphson Methode benutzt. Konvergiert multiroot() nicht, so werden NA ausgegeben. Für beide Regressionen wird die Messfehlervarianz benötigt, die innerhalb der Funktion geschätzt wird oder extern übergeben werden kann, bzw. für vorliegen von nur einer Messwiederholung muss diese aus externen Daten geschätzt werden. Für die Schätzung werden je nach Datentyp die Formeln aus Kapitel 4.6 herangezogen.

Tabelle 16: Input des Monte-Carlo korrigierte Score Funktion

Input	Erläuterung
datatype	character, "validData"falls Validierungsdaten vorliegen bzw. "wdhData"falls Wiederholungsdaten vorliegen
regtype	character, "linReg"falls lineare Regression gerechnet werden soll, "logReg"falls logistische bzw. "poisReg"falls eine Poisson Regression berechnet werden soll
y	vector, beobachtete y-Werte
x.true	vector, falls vorhanden die fehlerfrei gemessenen x-Werten, default ist NULL
x.valid	vector, falls Validierungsdaten vorliegen, dann Vektor aus NA und x-Werten, je nachdem, wenn keine oder eine Validierungsmessung vorliegt, default ist NULL
xSt	matrix, beobachtete fehlerhafte Messungen x^*. Anzahl der Spalten entspricht Anzahl der Wiederholungsmessungen
varU	numeric, falls extern geschätzte Messfehlervarianz vorliegt, default NULL
b	numeric, Anzahl der Bootstrapp-Ziehung zur Schätzung der Varianz der geschätzten $\hat{\beta}$, default 100
showplot	logical, falls TRUE dann wird eine Graphik ausgegeben mit eingezeichnetem y, xSt, der naiven und der geschätzten Regressionsgerade und falls x.true vorliegt, zusätzlich x.true und die wahre Regressionsgerade

Die MCcorrScore.fkt() gibt einen numerischen Vektor der geschätzten Parameter zurück

Tabelle 17: Output des Monte-Carlo korrigierte Score Funktion

Output	Erläuterung
beta0.MCcor	geschätztes $\hat{\beta}_0$ aus der Monte-Carlo korrigierten Score Funktion
beta1.MCcor	geschätztes $\hat{\beta}_1$ aus der Monte-Carlo korrigierten Score Funktion
beta0.bench	falls *x.true* vorliegt: geschätztes $\hat{\beta}_0$ aus lm() oder glm()
beta1.bench	falls *x.true* vorliegt: geschätztes $\hat{\beta}_1$ aus lm() oder glm()
beta0.naiv	mit der fehlerhaften Messung X^* geschätztem $\hat{\beta}_0$ aus lm() oder glm()
beta0.naiv	mit der fehlerhaften Messung X^* geschätztem $\hat{\beta}_1$ aus lm() oder glm()

E Zusatzmaterial

Auflistung der R Dateien, die im elektronischen Anhang enthalten sind. Da “set.seed()“ gesezt wurden können die Datensätze und Ergebnisse reproduziert werden und sind daher nicht im elektronischen Anhang enthalten.

Tabelle 18: Datengeneration

Dateiname	Erläuterung
Data_1.R	Datengenerierende Funktion zur Erstellung der Datensätze zu Kapitel 5.3.1
Dara_2_3.R	Datengenerierende Funktion zur Erstellung der Datensätze zu Kapitel 5.3.4 und 5.3.5
Data_AV.R	Datengenerierende Funktion zur Erstellung der Datensätze mit Annahmeverletzungen zu Kapitel 5.3.7

Tabelle 19: Funktion der Methoden

Dateiname	Erläuterung
Rk_fkt_source.R	Funktion zur Anwendung der Regressionskalibrierung
corScore_fkt_source.R	Funktion zur Anwendung der Methode der korrigierte Score Funktion
MCcorScore_fkt_source.R	Funktion zur Anwendung der Methode der Monte-Carlo korrigierte Score Funktion

Tabelle 20: Anwendung der Regressionskalibrierung. Aufgrund der Laufzeit in mehrere Dateien aufgespalten.

Dateiname	Erläuterung
RkResults1_01.R	Anwendung der Regressionskalibrierungsfunktion auf den Datensatz mit Messfehlervarianz 0.1^2 (Kapitel 5.3.1)
RkResults1_03.R	Anwendung der Regressionskalibrierungsfunktion auf den Datensatz mit Messfehlervarianz 0.3^2 (Kapitel 5.3.1)
RkResults1_05.R	Anwendung der Regressionskalibrierungsfunktion auf den Datensatz mit Messfehlervarianz 0.5^2 (Kapitel 5.3.1)
RkResults1_07.R	Anwendung der Regressionskalibrierungsfunktion auf den Datensatz mit Messfehlervarianz 0.7^2
RkResults1_1.R	Anwendung der Regressionskalibrierungsfunktion auf den Datensatz mit Messfehlervarianz 1^2
RkResults2.R	Anwendung der Regressionskalibrierungsfunktion auf den Datensatz zu Kapitel 5.3.4
RkResults3.R	Anwendung der Regressionskalibrierungsfunktion auf den Datensatz zu Kapitel 5.3.5
RkResultsAV1.R	Anwendung der Regressionskalibrierungsfunktion auf den Datensatz zu Kapitel 5.3.7.1
RkResultsAV2.R	Anwendung der Regressionskalibrierungsfunktion auf den Datensatz zu Kapitel 5.3.7.2
RkResultsAV3.R	Anwendung der Regressionskalibrierungsfunktion auf den Datensatz zu Kapitel 5.3.7.4
RkResultsAV4.R	Anwendung der Regressionskalibrierungsfunktion auf den Datensatz zu Kapitel 5.3.7.3
RkvarU1.R	Anwendung der Regressionskalibrierungsfunktion auf den Datensatz mit Messfehlervarianz $\sigma_u^2 = 0.3^2$ wobei eine externe Messfehlervarianz von $1\sigma_u^2$ übergeben wird (Kapitel 5.3.6)
RkvarU1.1.R	Anwendung der Regressionskalibrierungsfunktion auf den Datensatz mit Messfehlervarianz $\sigma_u^2 = 0.3^2$ wobei eine externe Messfehlervarianz von $1.1\sigma_u^2$ übergeben wird (Kapitel 5.3.6)
RkvarU1.5.R	Anwendung der Regressionskalibrierungsfunktion auf den Datensatz mit Messfehlervarianz $\sigma_u^2 = 0.3^2$ wobei eine externe Messfehlervarianz von $1.5\sigma_u^2$ übergeben wird (Kapitel 5.3.6)
RkvarU2.R	Anwendung der Regressionskalibrierungsfunktion auf den Datensatz mit Messfehlervarianz $\sigma_u^2 = 0.3^2$ wobei eine externe Messfehlervarianz von $2\sigma_u^2$ übergeben wird (Kapitel 5.3.6)

Tabelle 21: Anwendung der korrigierten Score Funktion und der Monte-Carlo korrigierten Funktion. Aufgrund der Laufzeit wurde letzteres in mehrere Dateien aufgespalten.

Dateiname	Erläuterung
corScore_all.R	Anwendung der korrigierten Score Funktion auf alle Datensätze und Fälle aus Kapitel 5.3
MC_B_variation.R	Anwendung der Monte-Carlo korrigierten Score Funktion auf Datensatz mit Messfehlervarianz 0.3^2, wobei unterschiedliche Werte für B übergeben worden ist (Kapitel 5.3.2)
MC_01_05.R	Anwendung der Monte-Carlo korrigierten Score Funktion auf Datensatz mit Messfehlervarianz 0.1^2 und 0.5^2 (Kapitel 5.3.1)
MC_07_1.R	Anwendung der Monte-Carlo korrigierten Score Funktion auf Datensatz mit Messfehlervarianz 0.7^2 und 1^2 (Kapitel 5.3.1)
MC_AV1_AV2.R	Anwendung der Monte-Carlo korrigierten Score Funktion auf Datensatz zu Kapitel 5.3.7.1 und 5.3.7.2
MC_AV3_1_AV_3_2R	Anwendung der Monte-Carlo korrigierten Score Funktion auf Datensatz zu Kapitel 5.3.7.4
MC_AV4.R	Anwendung der Monte-Carlo korrigierten Score Funktion auf Datensatz zu Kapitel 5.3.7.3
MC_data2.R	Anwendung der Monte-Carlo korrigierten Score Funktion auf Datensatz zu Kapitel 5.3.4
MC_data3.R	Anwendung der Monte-Carlo korrigierten Score Funktion auf Datensatz zu Kapitel 5.3.5
MC_Var_Variation.R	Anwendung der Monte-Carlo korrigierten Score Funktion auf Datensatz mit Messfehlervarianz $\sigma_u^2 = 0.3^2$ wobei verschiedene externe Messfehlervarianz übergeben wird (Kapitel 5.3.6)
MC_grVar_B1	Anwendung der Monte-Carlo korrigierten Score Funktion auf Datensatz mit Messfehlervarianz $\sigma_u^2 = 0.3^2$ wobei große externe Messfehlervarianz $4 * \sigma_u^2$ übergeben wird und B=1 ist (Kapitel 5.3.2)
MC_grVar_B1000	Anwendung der Monte-Carlo korrigierten Score Funktion auf Datensatz mit Messfehlervarianz $\sigma_u^2 = 0.3^2$ wobei große externe Messfehlervarianz $4 * \sigma_u^2$ übergeben wird und B=1000 ist (Kapitel 5.3.2)

Tabelle 22: Rcodes zur Erstellung der Grafiken in dieser Arbeit.

Dateiname	Erläuterung
RCodeSimulationAuswertung.R	sämtliche Plots zur Auswertung der Ergebnisse
VerteilungsPlot.R	sämtliche Plots die keine Auswertung der Ergebnisse sind

Literatur

R-Package 'rootSolve'. https://cran.r-project.org/web/packages/rootSolve/rootSolve.pdf; abgerufen am 16.09.2015.

STRATOS (STRengthening Analytical Thinking for Observational Studies). http://www.stratos-initiative.org; abgerufen am 15.09.2015.

R-Package 'rootSolve'. https://cran.r-project.org/web/packages/rootSolve/rootSolve.pdf; abgerufen am 16.09.2015.

T. Augustin, A. Döring, and D. Rummel. Regression calibration for Cox regression under heteroscedastic measurement error — Determining risk factors of cardiovascular diseases from error-prone nutritional replication data. In C. Heumann and Shalabh, editors, Recent Advances in Linear Models and Related Areas, Essays in Honour of Helge Toutenburg , pages 253–278. Physika Verlag, Heidelberg, 2008. URL `http://dx.doi.org/10.1007/978-3-7908-2064-5_13`.

J. P. Buonaccors. Measurement Error. Chapman & Hall/CRC, Boca Raton, 1986.

Jeffrey S. Buzas. A note on corrected scores for logistic regression. Statistics & Probability Letters, 79(22):2351–2358, 2009.

Caroll, R. J. and Ruppert, D. and Stefanski, L. A. and Crainiceanu, C. M. Measurment Error in Nonlinear Models- A Modern Perspective. Chapman & Hall/CRC-Taylor & Francis Group, Boca Raton, 2006.

C. Czado and T. Schmidt. Mathematische Statistik (Statistik und ihre Anwendungen) . Springer- Verlag, Berlin Heidelberg, 2011.

L. Fahrmeir, L. Kneib, and T.S. Lang. Regression, Modelle, Methoden und Anwednungen. Springer-Verlag Berlin Heidelberg, Boca Raton, 2009.

P. Gustafson. Measurement Error and Missclassifictaion in Statistics and Epidemiology-Impacts and Ba
Chapman & Hall/CRC, Boca Raton, 2004.

J. W. Hardin, H. Schmeidiche, and R. J. Carroll. The regression-calibration method for fitting generalized linear models with additive measurement error. Stata Journal, 3(4):373–385, December 2003.

Yijian Huang and CY Wang. Nonparametric correction to errors in covariates. Unpublished technical report, 1999.

M.A. Le. Regressionskalibrierung, 2015. Seminararbeit an der Ludwig-Maximilians-Universität München.

T. Nakumara. Corrected score function for errors-in-variables models: Methodology and application to generalized linear models. Biometrika, 11(1):127–137, 1990.

Steven J. Novick and Leonard A. Stefanski. Corrected Score Estimation via Complex Variable Simulation Extrapolation. Journal of the American Statistical Association, 97(458):pp. 472–481, 2002. ISSN 01621459. URL http://www.jstor.org/stable/3085663.

B.F. Qaqish. Newton-Raphson Method and Fisher Scoring. http://www.bios.unc.edu/ qaqish/bios767/mle.pdf; abgerufen am 12.08.2015.

H. Schneeweiß and H. J. Mittag. Lineare Modelle mit fehlerbehaftete Daten. Phsyica-Verlag Heidelberg Wien, Wien, 1986.

L.A. Stefanski. Unbiased estimation of a nonlinear function of a normal mean with application to measurement error models. Communications in Statistics, Series A, pages 4335–4358, 1989.

H Stocker. Methoden der Empirischen Wirtschaftsforschung. http://www.uibk.ac.at/econometrics/einf/20.pdf; abgerufen am 12.08.2015.

H. Werner. Die schiefe Normalverteilung und ausgewählte Eigenschaften, 2013. Bachelorarbeit an der Philipps Universität Marburg; Online erhältlich unter http://www.uni-marburg.de/fb12/stoch/files/theses/bachelorarbeit-werner.pdf; abgerufen am 2.09.2015.

D.M. Zucker, M. Gorfine, Y. Li, M. Tadesse, and D. Spiegelman. A Regularization Corrected Score Method for Nonlinear Regression Models with Covariate Error. Biometrics, 69(1):80–90, 2013.

Ehrenwörtliche Erklärung

Ich erkläre hiermit ehrenwörtlich, dass ich die vorliegende Arbeit selbständig angefertigt habe; die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht.

Die Arbeit wurde bisher keiner anderen Prüfungsbehörde vorgelegt und auch noch nicht veröffentlicht.

München, den 23.09.2015

Minh-Anh Le

(Unterschrift)