Ludwig-Maximilians University

Seminararbeit

Regressionskalibrierung

Author:
Minh-Anh Le

Supervisor:
Prof. Dr. Thomas Augustin

15. März 2015

Inhaltsverzeichnis

1. Einleitung

Der Umgang mit Messfehlern in der Regression ist eine wichtige Problemstellung der Statistik. Würde man das Vorhandensein von Messfehlern bei Regressionsanalysen ignorieren, so würden die Parameter oft verzerrt geschätzt werden.

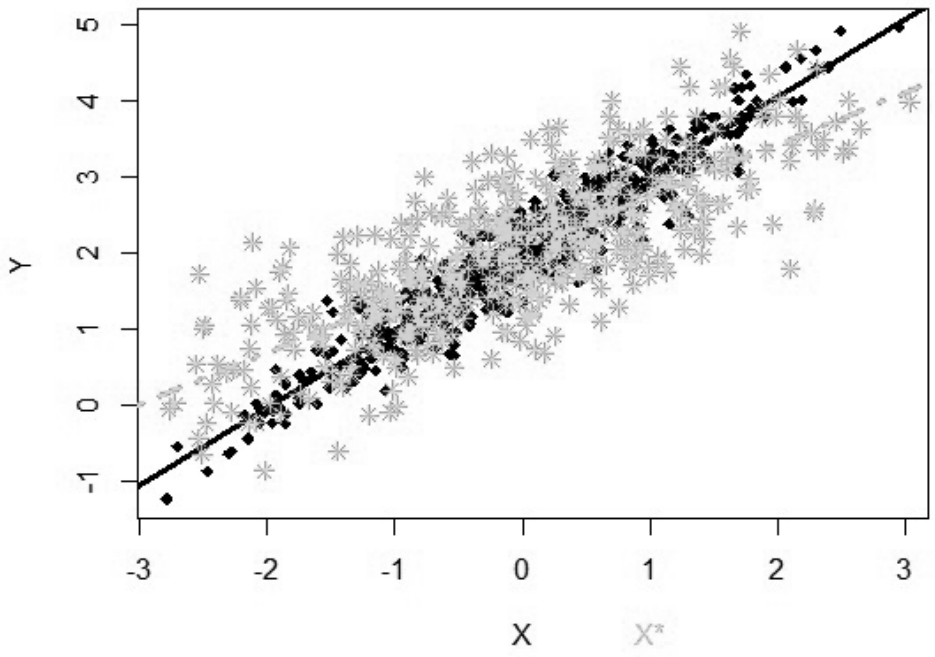

Abbildung 1: *Schwarze durchgezogene Linie* ist die wahre lineare Regressionsgerade. *Grün gepunktet und gestrichelte Linie* entspricht der linearen Regressionsgerade aus der naiven Schätzung bei vorliegen eines Messfehlers in der Einflussgröße (vgl. Rcode Anhang E.4).

Um dem verzerrten Effekt der Messfehler entgegenzuwirken, wurde im Laufe der Zeit eine Reihe von Methoden entwickelt. Ein einfaches und bewährtes Handwerkzeug, um eine Regression mit klassischem Fehler in der stetigen Einflussvariable anzupassen, ist die Regressionskalibrierung. Auch Messfehler in der Zielgröße sind denkbar vergleiche dazu die Arbeit von Markovic [2015].
Grundlage der Regressionskalibrierung ist es, im Hauptmodell, in dem die Zielgröße $\mathbf{Y}$ auf die Einflussgrößen ($\mathbf{Z}$, $\mathbf{X}$) regressiert wird, die Kovariable $\mathbf{X}$, die fehlerhaft gemessen wurde, durch eine Regression zu ersetzen, nämlich der Regression von $\mathbf{X}$ auf ($\mathbf{Z}$, $\mathbf{X}^*$). Dabei soll $\mathbf{Z}$ ohne Fehler gemessen werden können und die Hilfsgröße $\mathbf{X}^*$ sollte in Beziehung zu der eigentlich interessierenden Variable $\mathbf{X}$ stehen. $\mathbf{X}^*$ kann z.B. die fehlerhafte Messung von $\mathbf{X}$ sein. Nach diesem Vorgang kann ohne Weiteres eine gewöhnliche Stan-

dardanalyse durchgeführt werden.
Bei den verschiedenen Methoden zum Ausgleich von Messfehlern, unterscheidet man nach Caroll et al. [2006] (Kapitel 2.1, S.25) grundsätzlich zwischen *funktioneller Methode* und *struktureller Methode*. Erstere geht davon aus, dass $\mathbf{X}$ entweder konstant oder stochastisch ist und selbst im stochastischen Fall werden keine bzw. höchstens minimale Annahmen über die Verteilung von $\mathbf{X}$ getroffen, während in der *strukturellen Modellierung* eine parametrische Verteilung für die stochastische, unbeobachtete Variable $\mathbf{X}$ angenommen wird. Allerdings ist zu beachten, dass die resultierenden Schätzungen von der Wahl der Verteilung abhängen. Zu den strukturellen Verfahren zählen die Likelihood-, Quasilikelihood- und Bayes-Methode. Zu den funktionellen Verfahren gehören die Simulation Extrapolation (SIMEX) (vgl. Anhang F), die korrigierte Scorefunktion und die Regressionskalibrierung. Das bedeutet, dass die Verteilung von $\mathbf{X}$ bedingt auf die restlichen Kovariablen unbedingt nötig ist, um die Regressionskalibrierung anzuwenden.
Weiterhin unterscheidet Caroll et al. [2006] (Kapitel 2.5, S 36 ff.) zwischen *differentiellen* und *nicht differentiellen* Messfehlern. *Nicht differentielle* Messfehler treten auf, wenn die Verteilung von $\mathbf{Y}$ gegeben ($\mathbf{X}$, $\mathbf{Z}$, $\mathbf{X}^*$) allein von ($\mathbf{Z}$, $\mathbf{X}$) abhängt. In anderen Worten, wenn $\mathbf{X}^*$ ein Surrogat ist und somit, anders als $\mathbf{X}$ und $\mathbf{Z}$, keine zusätzlichen Informationen über $\mathbf{Y}$ enthält (näheres im Anhang A und Marshalava [2015]). Für die Anwendung der Regressionskalibrierung ist das Vorliegen von *nicht differentiellen* Fehlern erforderlich [Caroll et al., 2006] (Kapitel 2.5, S.37). Die Regressionskalibrierung kann sowohl bei Vorliegen von systematische als auch stochastische Fehler angewendet werden. Systematische Fehler führen zur verzerrten Schätzung des Intercepts, während stochastische Fehler zusätzlich die Schätzung der Slopes beeinflusst. Nach Caroll et al. [2006] und in dieser Arbeit wird der Fokus auf stochastische Fehler gelegt o.B.d.A, dass systematische Fehler korrigiert werden können.
Abschließend seien noch ein paar Bemerkungen zu den verschiedenen Fehlermodellen hinzuzufügen: Für die Regressionskalibrierung ist an dieser Stelle das *klassische Fehlermodell* und das *Berkson- Fehlermodell* zu nennen. Wenn das *klassische additive Fehlermodell* $\boldsymbol{X}_i^* = \boldsymbol{X}_i + \boldsymbol{U}_i$ vorliegt, d.h. wenn fehlerhafte Werte $\boldsymbol{X}_i^*$ vorliegen, die aus dem unbekannten Wert der wahren Variable $\boldsymbol{X}_i$ und dem additiven Fehler $\boldsymbol{U}_i$ bestehen, dann kann diese durch die Regressionskalibrierung in einen Berkson- Fehler $\boldsymbol{X}_i = \boldsymbol{X}_i^* + \boldsymbol{U}_i$ überführt werden [Caroll et al., 2006] (Kapitel 2.2.3, S.29). Das bedeutet, dass das wahre $\boldsymbol{X}_i$ bedingt auf $\boldsymbol{X}_i^*$ und $\boldsymbol{U}_i$ dargestellt werden kann, wobei für $\boldsymbol{U}_i$ Mittelwert Null und unabhängig identische Verteilung angenommen werden. Eine kurze Übersicht über die Unterschiede/Vorteile des Berkson- Fehlers gegenüber dem klassischen Fehler wird im Anhang B dieser Arbeit zusammengefasst. Weitere Messfehlerarten und deren Auswirkungen können in der Arbeit von Marshalava [2015] nachgelesen werden.

Zweck dieser Arbeit ist es, einen Überblick über die Methode der Regressionskalibrierung für additive Messfehler in der Einflussvariable zu geben. Dazu werden in Kapitel 2 die zentrale Idee und der Algorithmus der Regressionskalibrierung bereitgestellt. Im dritten Kapitel wird die entscheidende Parameterschätzung der Regressionskalibrierungsmethode in Abhängigkeit von der Art der vorliegenden Daten besprochen. Im vierten Kapitel

wird die Methode des Bootstrappings im Rahmen der Regressionskalibrierung erklärt. An einem theoretischen Beispiel der einfachen linearen Regression in Kapitel 5, einem praktischen Beispiel der MONICA- Studie (Kapitel 6) und anhand Simulationen in R (Anhang E) soll die Regressionskalibrierung veranschaulicht werden.

2. Die zentrale Idee und der Algorithmus der Regressionskalibrierung

2.1. Zentrale Idee

Wie eingangs bereits zusammengefasst, ermöglicht die Regressionskalibrierung, den Effekt von additiven Messfehlern auf die Schätzung der Kovariablen in einem generalisierten linearen Modell zu korrigieren. Die zentrale Idee der Regressionskalibrierung ist relativ einfach (vgl. Caroll et al. [2006], Kapitel 4.1, S.65 ff.). Angenommen die tatsächliche Einflussgröße $\mathbf{X}$ wurde nicht erhoben oder sie ist nicht beobachtbar, sondern nur fehlerbehaftet messbar als $\boldsymbol{X}^*$, aber es liegen die Zielgröße $\mathbf{Y}$ und weitere Einflussgrößen $\mathbf{Z}$, die ohne Fehler gemessen wurden, vor. Eine gewöhnliche Regression von $\mathbf{Y}$ auf ($\mathbf{Z}$, $\mathbf{X}$) zu rechnen, gestaltet sich im vorliegenden Fall als schwierig. Da aber $\mathbf{X}^*$, nach Caroll et al. [2006] (Kapitel 2.5.5, S.36) ein Surrogat, zur Verfügung steht, die als Ersatz für $\mathbf{X}$ dient, kann die Regressionskalibrierung angewendet werden. Wendet man die Regressionskalibrierung nicht an, sondern rechnet eine naive Regression von $\mathbf{Y}$ auf $\boldsymbol{(Z, X^*)}$ anstelle der eigentlich interessierenden Regression auf $\mathbf{(X, Z)}$, so hat dies zur Folge, dass die zugehörigen Schätzungen verzerrt sind. Die Anwendung der Regressionskalibrierung wirkt dem entgegen und ermöglicht näherungsweise unverzerrte Inferrenz. Anstatt $\mathbf{(Z,X)}$ durch $(\boldsymbol{Z, X^*})$ zu ersetzen, soll also zunächst eine Regression von $\mathbf{X}$ auf $(\boldsymbol{Z, X^*})$ berechnet werden. Man schreibt $\widehat{\boldsymbol{m}}_{\boldsymbol{x}}\boldsymbol{(Z, X^*, \widehat{\gamma})} = \widehat{\boldsymbol{X}}$ mit Koeffizientenschätzer $\widehat{\boldsymbol{\gamma}}$, als Schätzung von $\boldsymbol{m_X(Z, X^*, \gamma)} = \boldsymbol{E}(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{X^*})$, sodass die erklärenden Variablen für das Hauptmodell nun $\{\boldsymbol{m_X(Z, X^*, \widehat{\gamma}), Z}\}$ sind. Der Regressionskalibrierungs-Algorithmus besteht, angelehnt an Caroll et al. [2006] (Kapitel 4.2, S.66), im Wesentlichen aus drei Schritten, die im folgenden Kapitel dargestellt werden.

2.2. Algorithmus nach Caroll

Beachte, dass man ursprünglich an $E[\boldsymbol{Y}|\boldsymbol{Z}, \boldsymbol{X}] = \boldsymbol{m_Y(Z, X, \beta)}$ interessiert ist. da Der Algorithmus der Regressionskalibrierung nach Caroll et al. [2006] (Kapitel 4.2, S.66) hat folgende Form:

- Schritt 1: Schätze die unbeobachtete Variable $\mathbf{X}$ durch eine Regression von $\mathbf{X}$ auf ($\mathbf{Z}$, $\mathbf{X}^*$), also $E[\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{X^*}] = \boldsymbol{m_X}(\boldsymbol{Z}, \boldsymbol{X^*}, \boldsymbol{\gamma})$. Die Schätzung ist von $\boldsymbol{\gamma}$, genauer von der Schätzung $\widehat{\boldsymbol{\gamma}}$ abhängig (Kapitel 3).
- Schritt 2: Ersetze die nicht beobachtete Variable $\mathbf{X}$ durch die im vorherigen Schritt durchgeführte Schätzung, d.h. ersetze im Hauptmodell $\mathbf{X}$ durch $\boldsymbol{m_X}(\boldsymbol{Z}, \boldsymbol{X^*}, \widehat{\boldsymbol{\gamma}})$.

Führe anschließend eine Standardanalyse durch, um die Parameterschätzer zu erhalten. Somit erhält man:

$$E[\boldsymbol{Y}|\boldsymbol{Z}, \boldsymbol{X}^*] \approx \boldsymbol{m_Y}(\boldsymbol{Z}, \underbrace{\boldsymbol{m_X}(\boldsymbol{Z}, \boldsymbol{X}^*, \widehat{\gamma})}_{\widehat{\boldsymbol{X}}}, \boldsymbol{\beta_{RK}}) \tag{1}$$

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 Bootsrap-Verfahren oder der Sandwich-Methode die resultierenden Standardfehler (Kapitel 4).

Trotz der Einfachheit des Verfahrens sollte man aber die Leistungsfähigkeit nicht unterschätzen (vgl. Abbildung 2).

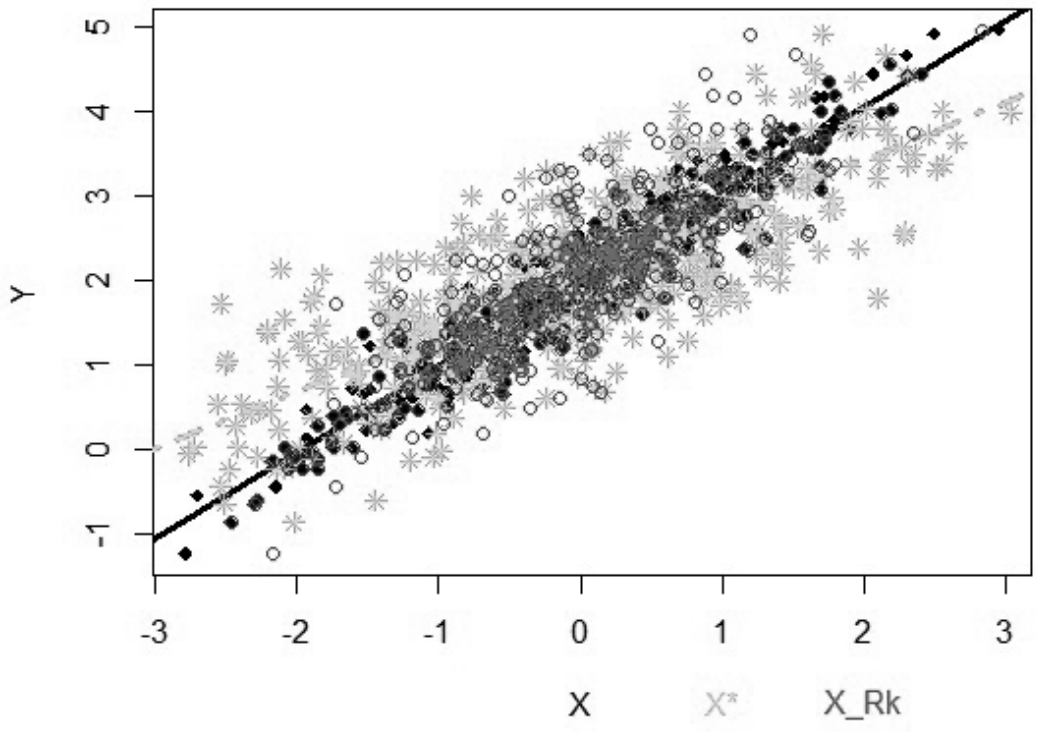

Abbildung 2: *Schwarze durchgezogene Linie* ist die wahre lineare Regressionsgerade. *Grün gepunktet und gestrichelte Linie* entspricht der linearen Regressionsgerade aus der naiven Schätzung bei vorliegen eines Messfehlers in der Einflussgröße.*rot gestrichelte Linie* stellt die Schätzung nach Anwendung der Regressionskalibrierung auf Validierungsdaten dar (vgl. Rcode).

Im Allgemeinen gilt, dass der resultierende Mittelwert zwar nicht notwendigerweise mit dem tatsächlichen Mittelwert der beobachteten Daten übereinstimmen muss, aber in den meisten Fällen kann man nur eine geringe Differenz verzeichnen. Diese approximative Korrekturmethode reduziert den Bias, aber produziert, abhängig vom Modell, nicht notwendigerweise konsistente Schätzer (vgl. Augustin et al. [2008],Kapitel 1, S.257).
Welche Möglichkeiten es für die Schätzung im ersten Schritt gibt, ist abhängig von den vorliegenden Daten bzw. zusätzlichen Informationen. Im nächsten Kapitel werden einige Fälle erläutert, insbesondere wird näher auf den Umgang mit Wiederholungsdaten eingegangen.

3. Parameterschätzung in der Kalibrierungsfunktion

3.1. Überblick

Ein großer Vorteil der Regressionskalibrierung besteht darin, dass man für die Messfehlerkorrektur mit jeder statistischen Standardsoftware ohne zusätzliche Implementierung arbeiten kann, indem man $\mathbf{X}$ durch eine Regression auf $\mathbf{(Z, X^*)}$ ersetzt. In diesem Abschnitt soll, abhängig von zusätzlich vorliegenden Informationen, aufgezeigt werden, wie eine geeignete Regression in Schritt 1 durchgeführt werden kann. Caroll et al. [2006] stellt folgende Szenarien vor: Es liegen interne Validierungsdaten, Instrumentaldaten oder Wiederholungsdaten vor. In den ersten beiden Fällen kann das Model für $\mathrm{E}[\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{X^*}]$ durch gewöhnliche Regressionsdiagnosen, bspw. Residuenplot geprüft werden. Wenn zusätzliche partielle Wiederholungsdaten vorliegen, kann auch das Modell aus dem Fall von Wiederholungsdaten überprüft werden; Genaueres wird in Kapitel 3.4.1 und Anhang D erläutert.

3.2. Interne Validierungsdaten

Wenn interne Validierungsdaten vorliegen, d.h. wenn für einen Teil der Daten die wahre Messung von $\mathbf{X}$ vorliegt, dann ist die einfachste Vorgehensweise eine Regression von $\mathbf{X}$ auf die Kovariablen ($\mathbf{Z}$, $\boldsymbol{X^*}$) aus den internen Validierungsdaten durchzuführen. Allerdings ist dieser Fall eigentlich ein Missing-Data Problem und man würde normalerweise eher Missing-Data-Methoden, statt der Regressionskalibrierung anwenden (vgl. Caroll et al. [2006],Kapitel 4.4.1 und die Arbeit von Pokatilo [2015]). Aus praktischer Sicht empfiehlt es sich für eine schnelle Analyse, die $\mathbf{X}$- Werte zu verwenden, die vorhanden sind; für die fehlenden $\mathbf{X}$- Werte soll die Regressionskalibrierungs-Methode angewendet werden. Caroll et al. [2006] schlägt in Kapitel 4.4.1 vor, an dieser Stelle zusätzlich eine Dummyvariable einzuführen, die angibt, ob es sich um wahre $\mathbf{X}$- Werte handelt oder

um die geschätzten Daten aus der Regressionskalibrierung. Beispiel dazu im Rcode im Anhang E.4

3.3. Instrumentaldaten

In manchen Situationen steht für einen Teil der Befragten zusätzlich zu $\mathbf{X}^*$ eine externe, fehlerfrei gemessene Variable $\mathbf{T}$, die mit $\mathbf{X}$ zusammenhängt, zur Verfügung, sogenannte Instrumentaldaten (vgl. Kapitel C). Dabei kann die Regressionskalibrierung nur angewendet werden, wenn $\mathbf{T}$ unverzerrt für $\mathbf{X}$ ist (Caroll et al. [2006], Kapitel 4.4, S.70). Anhand dieser zusätzlichen Information kann nun eine Regression von $\mathbf{T}$ auf ($\mathbf{Z}$, $\mathbf{X}^*$) berechnet werden, die genau der Regression von $\mathbf{X}$ auf ($\mathbf{Z}$, $\mathbf{X}^*$) entspricht. Anschließend funktioniert die Regressionskalibrierung analog zu dem Vorgehen bei Vorliegen von internen Validierungsdaten (Caroll et al. [2006], Kapitel 4.4, S.70).

3.4. Wiederholungsdaten

In diesem Abschnitt wird,entsprechend Caroll et al. [2006] (Kapitel 4.4.2,S.70), ein klassisches additives Fehlermodell $X^* = X + U$ betrachtet, wobei die Fehlerterme, bedingt auf ($\mathbf{Z}$, $\mathbf{X}$), einen Mittelwert von Null und eine konstante Kovarianzmatrix $\sum_{\boldsymbol{uu}}$ haben, d.h. es wird Homoskedastizität angenommen ($\boldsymbol{\Sigma_i \equiv \Sigma}$). Da in der Literatur der Umgang mit Wiederholungsdaten in der Regressionskalibrierung hervorgehoben wird, nimmt dieses Unterkapitel einen großen Teil ein. Im Folgenden beschreiben wir einen Algorithmus, der nach Caroll et al. [2006] (Kapitel 4.4.2, S.70 ff.) eine lineare Approximation der Regressionskalibrierungsfunktion liefert. Der Algorithmus ist anwendbar, wenn $\sum_{\boldsymbol{uu}}$ durch externe Daten oder interne Wiederholungsmessungen geschätzt wurde. Diese Methode wurde von Caroll und Stefanski (1990) und von Gleser (1990) entwickelt und wurde unter anderem von Liu und Liang (1992) und Wang, Caroll und Liang (1996) angewendet. In diesem Abschnitt befassen wir uns mit der Nutzung von Wiederholungsmessungen von $\mathbf{X}$ bzw. mit Wiederholungsmessungen von $\mathbf{X}^*$, die das gleiche $\mathbf{X}$ messen.
Angenommen es liegen k_i Wiederholungsmessungen vor, $\boldsymbol{X}^*_{i1}$...,$\boldsymbol{X}^*_{ik_\mathrm{i}}$ von $\mathbf{X_i}$ und $\overline{\boldsymbol{X}^*_{i\cdot}}$ ist der zugehörige Mittelwert. Der Index i gibt an, um welche Beobachtung es sich handelt, k_i gibt an wie viele Messungen für diese vorliegen.
Die beste lineare Approximation von $\mathbf{X}$ gegeben $\boldsymbol{(Z, \overline{X^*})}$ ist

$$E[X|Z,\overline{X^*}] \approx \mu_x + (\Sigma_{xx}, \Sigma_{zx}) \begin{bmatrix} \Sigma_{xx} + \Sigma_{uu}/k & \Sigma_{xz} \\ {\Sigma_{xz}}^t & \Sigma_{zz} \end{bmatrix}^{-1} \begin{pmatrix} \overline{\boldsymbol{X^*}} - \mu_{x^*} \\ \boldsymbol{Z} - \mu_z \end{pmatrix}. \tag{2}$$

Wiederholungsmessungen ermöglichen die Schätzung der Kovarianzmatrix der Messfehler $\boldsymbol{U_i}$

$$\widehat{\Sigma}_{uu} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{k_i} \left(\boldsymbol{X}^*_{ij} - \overline{\boldsymbol{X}^*_{i\cdot}}\right)\left(\boldsymbol{X}^*_{ij} - \overline{\boldsymbol{X}^*_{i\cdot}}\right)^t}{\sum_{i=1}^{n} (k_i - 1)}. \tag{3}$$

Nach Caroll et al. [2006] (Kapitel 4.4.2, S.71 ff.) gilt außerdem: Wenn in der linearen Regression nur eine Messwiederholung ($k_i = 1$) vorhanden ist, kann man auf die geschätzte Kovarianzmatrix $\widehat{\Sigma}_{uu}$ aus externen Daten zurückgreifen, falls diese vorliegt.
Weiterhin folgt auf Basis folgender Schätzungen

$$\widehat{\mu}_z = \overline{\boldsymbol{Z}}, \tag{4}$$

$$\widehat{\mu}_x = \widehat{\mu}_{x^*} = \sum_{i=1}^{n} k_i \overline{\boldsymbol{X}^*_{i\cdot}} / \sum_{i=1}^{n} k_i, \tag{5}$$

$$\nu = \sum_{i=1}^{n} k_i - \sum_{i=1}^{n} k_i^2 / \sum_{i=1}^{n} k_i, \tag{6}$$

$$\widehat{\Sigma}_{zz} = (n-1)^{-1} \sum_{i=1}^{n} (\boldsymbol{Z}_i - \overline{\boldsymbol{Z}_\cdot})(\boldsymbol{Z}_i - \overline{\boldsymbol{Z}_\cdot})^t, \tag{7}$$

$$\widehat{\Sigma}_{xz} = \sum_{i=1}^{n} k_i (\overline{\boldsymbol{X}^*_{i\cdot}} - \widehat{\mu}_{x^*})(\boldsymbol{Z}_i - \overline{\boldsymbol{Z}_\cdot})^t / \nu, \tag{8}$$

$$\widehat{\Sigma}_{xx} = \left[\left\{ \Sigma_{i=1}^{n} k_i (\overline{\boldsymbol{X}^*_{i\cdot}} - \widehat{\mu}_{x^*})(\overline{\boldsymbol{X}^*_{i\cdot}} - \widehat{\mu}_{x^*})^t \right\} - (n-1)\widehat{\Sigma}_{uu} \right] / \nu \tag{9}$$

die Gleichung der geschätzten Regressionskalibrierungsfunktion

$$\begin{aligned} E[\boldsymbol{X}_i | \widehat{\boldsymbol{Z}_i}, \overline{\boldsymbol{X}^*_{i\cdot}}] &\approx \widehat{\mu}_{x^*} + (\widehat{\Sigma}_{xx}, \widehat{\Sigma}_{zx}) \begin{bmatrix} \widehat{\Sigma}_{xx} + \widehat{\Sigma}_{uu}/k_i & \widehat{\Sigma}_{xz} \\ \widehat{\Sigma}_{xz}^t & \widehat{\Sigma}_{zz} \end{bmatrix}^{-1} \begin{pmatrix} \overline{\boldsymbol{X}^*_{i\cdot}} - \widehat{\mu}_{x^*} \\ \boldsymbol{Z}_i - \overline{\boldsymbol{Z}_\cdot} \end{pmatrix} \\ &\approx \boldsymbol{m_{X_i}(Z_i, X_i, \hat{\gamma})} = \widehat{\boldsymbol{X}_i}. \end{aligned} \tag{10}$$

Selbst wenn die Anzahl der Wiederholungen für jedes Beobachtungseinheit i nicht konstant ist, können mit dem vorgestellten Algorithmus in der linearen Regression konsistente Schätzer und in der logistischen Regression approximativ konsistente Schätzer erreicht werden [Caroll et al., 2006](Kapitel 4.4.2, S.72). Auf Basis dieser Approximation wird die allgemeine Formel in Kapitel 5 auf den Fall für eine einfach lineare Regression mit konstanten Wiederholungen reduziert. Im Anhang E.5 & E.6 finden sich beispielhafte R Anwendungen.

3.4.1. Überprüfung durch partielle Wiederholungsdaten

Wie bereits betont, sind die dargelegten Approximationen nur Näherungen, aber diese können durch die Wiederholungsmessungen selbst wiederum geprüft werden. Üblicherweise liegen nur für einen Teil der Daten interne, partielle Wiederholungsdaten vor (k_i=2), während die meisten Daten nicht wiederholbar sind (k_i=1). Die partiellen Wiederholungsdaten können dazu benutzt werden, um die beste lineare Approximation von E[$\mathbf{X}$|$\mathbf{Z}$,$\boldsymbol{X}^*$] aus Formel (10) zu bestimmen, indem man vorab eine Regression von $\boldsymbol{X}^*_{i2}$ auf $(\boldsymbol{Z_i}, \boldsymbol{X}^*_{i1})$ berechnet. Nach näherer Betrachtung gilt nämlich

$$\boldsymbol{X}^*_{i1} = \boldsymbol{X}_i + \boldsymbol{U}_{i1} \tag{11}$$
$$\boldsymbol{X}^*_{i2} = \boldsymbol{X}_i + \boldsymbol{U}_{i2} \tag{12}$$

und somit

$$E[\boldsymbol{X}^*_{i2}|\boldsymbol{Z_i}, \boldsymbol{X^*}_{i1}] = E[\boldsymbol{X}_i + \boldsymbol{U}_{i2}|\boldsymbol{Z}_i, \boldsymbol{X^*}_{i1}] \tag{13}$$
$$= \underbrace{E[\boldsymbol{X}_i|\boldsymbol{Z}_i, \boldsymbol{X^*}_{i1}]}_{Schritt\ 1\ der\ RK} + \underbrace{E[\boldsymbol{U}_{i2}|\boldsymbol{Z}_i, \boldsymbol{X^*}_{i1}]}_{V_i}. \tag{14}$$

Im klassischem Fehlermodell gilt für die bedingte Zufallsvariable V_i E[V_i]=0. Aus der Herleitung zeigt sich, dass eine Regression von $\boldsymbol{X^*}_{i2}$ auf $(\boldsymbol{Z}_i, \boldsymbol{X^*}_{i1})$ eine gute Schätzung für $E[\boldsymbol{X}_i|\boldsymbol{Z}_i, \boldsymbol{X^*}_{i1}]$ ist, das im ersten Schritt der Regressionakalibrierung geschätzt werden sollte. Weiteres im Anhang D.

4. Bootstrapping in der Regressionskalibrierung

In Kapitel 3 wurde der erste Schritt der Regressionskalibrierung ausführlich behandelt. Im zweiten Schritt wird eine einfache Regressionsanalyse geschätzt, mit den Schätzungen aus Schritt eins, also $E[\boldsymbol{Y}|\boldsymbol{Z}, \boldsymbol{X}^*] \approx \boldsymbol{m_Y}(\boldsymbol{Z}, \widehat{\boldsymbol{X}}, \boldsymbol{\beta_{RK}})$. Für dieses resultierende $\boldsymbol{\beta_{RK}}$ soll im dritten Schritt der Standardfehler bzw. die Varianz angepasst werden. An dieser Stelle sollte erwähnt werden, dass die ausgegebenen Standardfehler und p-Werte in statistische Softwares, nach dem ersten und zweiten Regressionskalibrierungsschritt, nur approximativ gelten und eher als ersten Eindruck für die Signifikanz der Schätzung dienen sollen, da die Schätzung im zweiten Schritt bereits auf einer Schätzung $\widehat{X}$ basiert (vgl. Augustin et al. [2008], Kapitel 1, S.257 ff.).
Die Anpassung der wahren Standardfehler können mit Bootstrapping oder der Sandwich-Methode erreicht werden (Caroll et al. [2006], Kapitel 4.2). Im Folgenden wird die Idee des Bootstrappings im Kontext der Regressionskalibrierung näher erläutert.
Die zentrale Idee des Bootstrapping ist es, Bootstrap-Daten zu simulieren, deren Verteilung der geschätzten Verteilung der tatsächlichen Daten entspricht (Caroll et al. [2006], Kapitel A.9.1, S.377 ff.). Dass statistische Verfahren auch auf Bootstrap-Stichproben angewendet werden können, ist ein Vorteil, der in der Regressionskalibrierung zum Einsatz kommt. Es gibt verschiedene Vorgehensweisen des Bootstrappings. Nach Efron and Tibshirani (1993) lauten diese: *resampling Pairs*, *resampling Residuals* und *parametric Bootstrap*, wobei die ersten beiden zu den *nonparametric Bootstrap-Verfahren* zählen (Caroll et al. [2006], Kapitel A.9.2, S.378 ff.) . Der Unterschied zwischen parametrischen Verfahren und nonparametrischen Verfahren ist, dass bei ersterem aus einer angenommenen Verteilung der Daten gezogen wird, bei letzterem wird mit Zurücklegen aus den vorliegenden Daten gezogen (Caroll et al. [2006] , Kapitel A.9.2, S.378 ff.). Im Folgenden wird kurz zusammengefasst, wie nonparametrische Bootstrapping-Verfahren in der

Regressionskalibrierung genutzt werden können.
Allgemein zu beachten ist, dass Daten in verschiedenen Formen vorliegen können. Beispielsweise können für einige Beobachtungen Validierungsdaten vorliegen, für einige andere nur Instrumentaldaten, für weitere nur Wiederholungsdaten. Diese unterschiedlichen Strukturen sollten beim ziehen der Bootstap-Stichprobe bei der Messfehlerkorrektur berücksichtigt werden, indem man die vorliegenden Daten aufteilt (nach der Strukturart) und anschließend aus jedem Teildatensatz Bootstrap-Stichproben zieht (vgl. Caroll et al. [2006], Kapitel A.9.5, S.381). Im klassischen Bootstrapping ist dieser Schritt nicht nötig, bei der Messfehlerkorrektur schon, da unterschiedliche Datenstrukturen unterschiedliche Mengen an Informationen haben. Ziel dieser Gruppierung der Daten ist es, dass jede Bootstrap-Stichprobe aus einer homogenen Umgebung gezogen wird. Gruppiert man vorher nicht, so wird Bootstrapping auf heterogene Daten angewendet und somit wird die Varianz erhöht.
Für die Erklärungen zu den verschiedenen Bootstrap-Ziehungs-Verfahren, wird im Folgenden das Vorliegen von Validierugsdaten angenommen.

4.1. Resampling Vectors im Messfehlermodell

Das *Resampling Vectors*- Verfahren ist eine Erweiterung des *Resampling Pairs*- Verfahren, in dem nicht mehr paarweise mit Zurücklegen gezogen wird, sondern vektorweise aus $\{(\boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{X}_i^*, \boldsymbol{Z}_i)\}_{i=1}^k$ bzw. aus $\{(\boldsymbol{Y}_i, \boldsymbol{X}_i^*, \boldsymbol{Z}_i)\}_{i=1}^n$, wenn keine wahren X-Werte vorliegen, mit k bzw. n als Umfang der Bootstrap-Stichprobe (vgl. Caroll et al. [2006] Kapitel A.9.2.1, S.378 ff.). Der Vorteil dieser Methode liegt darin, dass kaum Annahmen getroffen werden müssen, denn dadurch, dass vektorweise gezogen wird, müssen besondere Beziehungen zwischen den Variablen nicht beachtet werden. Beispielsweise kann ein Residuum $\boldsymbol{\epsilon_i}$ abhängig von $\boldsymbol{Z}_i$ sein, diese Abhängigkeitsbeziehung wird durch das vektorweise Ziehen berücksichtigt, da die Variablen $(\boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{X}_i^*, \boldsymbol{Z}_i)$ bzw. $(\boldsymbol{Y}_i, \boldsymbol{X}_i^*, \boldsymbol{Z}_i)$ immer zusammen gezogen werden. Ein Nachteil ist, dass die Bootstrap-Stichprobe nicht die gleiche Variablenmenge enthält wie die ursprünglichen Daten; das ist vor allem dann problematisch, wenn High- Leverage- Punkte vorliegen und diese häufiger oder kein einziges Mal in die Bootstrap-Stichprobe gezogen werden. Die Art von Stichprobenziehung erzeugt somit keine Stichprobenverteilung, die auf $(\boldsymbol{X}_i, \boldsymbol{X}_i^*, \boldsymbol{Z}_i)$ bedingt. Beachte, dass man eigentlich aus repräsentativen Gründen eine Stichprobenverteilung bedingt auf $\boldsymbol{Z}_i$ erreichen möchte Caroll et al. [2006] (Kapitel A.9.2.1, S. 378 ff.).

4.2. Resampling Residuals im Messfehlermodell

Anders als beim *Resampling Vectors* liefert *Resampling Residuals* eine Stichprobenverteilung, die auf $(\boldsymbol{Z_i}, \boldsymbol{X_i})$ bedingt. Diese Methode ist auf Validierungsdaten anwendbar, wenn zwei Regressionsmodelle vorliegen, nämlich $\boldsymbol{Y}_i$ gegeben $(\boldsymbol{Z_i}, \boldsymbol{X_i})$ und $\boldsymbol{X}_i^*$ gegeben $(\boldsymbol{Z_i}, \boldsymbol{X_i})$. Auf Grundlage des Kapitels A.9.2.2 in Caroll et al. [2006] kann man *Resampling Vectors* wie folgt anwenden:
Angenommen für i=1,...,k liegen wahre X-Werte vor und für die i=k+1...n fehlen diese X-Werte. Mit *Resampling Vectors* werden Bootsraps aus den Beobachtungen i=1,...k gezo-

gen. Um fiktive Daten $\boldsymbol{Y}_i^{(m)}$ für die Zielvariable zu erhalten, werden zunächst die Residuen $\boldsymbol{\epsilon_i} = \boldsymbol{Y_i} - m_{\boldsymbol{Y}}(\boldsymbol{Z}_i, \boldsymbol{X}_i, \hat{\mathcal{B}})$ berechnet, wobei $\hat{\mathcal{B}}$ der nichtlineare KQ-Schätzer ist. Um die m^{te} Bootstrap-Stichprobe zu erhalten, wird aus der Menge $\{(\boldsymbol{\epsilon}_i - \bar{\boldsymbol{\epsilon}})\}_{i=1}^k$ zufällig mit Zurücklegen $\{\boldsymbol{\epsilon}_i^{(m)}\}_{i=1}^k$ gezogen. Daraus lassen sich die Bootstraps $\boldsymbol{Y_i}^{(m)} = m_{\boldsymbol{Y}}(\boldsymbol{Z}_i, \boldsymbol{X}_i, \hat{\mathcal{B}}) + \boldsymbol{\epsilon}_i^{(m)}$ berechnen. Die m^{te} Bootstrap-Stichprobe besteht also aus $\{(\boldsymbol{Y}_i^{(m)}, \boldsymbol{Z}_i, \boldsymbol{X}_i)\}_{i=1}^k$. Analog, mithilfe der zweiten Regression, erhält man $\{(\boldsymbol{X}_i^{*(m)}, \boldsymbol{Z}_i, \boldsymbol{X}_i)\}_{i=1}^k$. Beachte, dass die Bootstrap-Stichprobe die tatsächlichen Werte von $(\boldsymbol{Z_i}, \boldsymbol{X_i})$ enthält, sodass die Verteilung der Bootstrap-Stichprobe auf $(\boldsymbol{Z_i}, \boldsymbol{X_i})$ bedingt. Dieses Verfahren ist nur dann anwendbar, wenn die Residuen $\boldsymbol{\epsilon}_i = \boldsymbol{Y_i} - m_{\boldsymbol{Y}}(\boldsymbol{Z}_i, \boldsymbol{X}_i, \hat{\mathcal{B}})$ und $\boldsymbol{\epsilon}_i^* = \boldsymbol{x_i^*} - m_{\boldsymbol{X^*}}(\boldsymbol{Z}_i, \boldsymbol{X}_i, \hat{\mathcal{B}^*})$ jeweils unabhängig identisch verteilt sind und annähernd der Homoskedastizitätsbedingung genügen. Für die Beobachtungen i=k+1,...,n können beispielsweise mit der *Resampling Vectors* Stichproben gezogen werden.

4.3. Algorithmus des Bootstrappings in der Regressionskalibrierung

Unabhängig davon welches Bootstrap-Verfahren angewendet wird, hat der Algorithmus des Bootsrappings in der Regressionskalibrierung nach eigener Zusammenfassung folgende Form:

- 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 $\{(\boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{X}_i^*, \boldsymbol{Z}_i)\}_{i=1}^k{}^{(m)} \cup \{(\boldsymbol{Y}_i, \boldsymbol{X}_i^*, \boldsymbol{Z}_i)\}_{i=k+1}^n{}^{(m)}$.
  - Nach *Resampling Residuals* auf $\{(\boldsymbol{Y}_i^{(m)}, \boldsymbol{X}_i^{*(m)}, \boldsymbol{Z}_i, \boldsymbol{X}_i)\}_{i=1}^k \cup \{(\boldsymbol{Y}_i, \boldsymbol{X}_i^*, \boldsymbol{Z}_i)\}_{i=k+1}^n{}^{(m)}$.

  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

$$\widehat{var(\widehat{\beta}_{RK_k})} = \frac{1}{M-1} \sum_{m=1}^{M} (\widehat{\beta}_{RK_k}^{(m)} - \overline{\widehat{\beta}_{RK_k}})(\widehat{\beta}_{RK_k}^{(m)} - \overline{\widehat{\beta}_{RK_k}})^t. \tag{15}$$

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: $\widehat{\beta}_{RK_k} = \beta_{RK_k}$.

5. Theoretisches Beispiel an einer einfachen linearen Regression

Nachdem alle drei Schritte des Regressionskalibrierungsalgorithmus dargestellt wurden, soll an einem einfachen linearen Regressionsmodell die Anwendung der Regressionskalibrierung gezeigt werden. Es soll davon ausgegangen werden, dass zusätzliche Informationen anhand von Wiederholungsdaten vorliegen.
Für die lineare Regression soll $E(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z}) = \beta_0 + \beta_1\boldsymbol{X} + \beta_2\boldsymbol{Z}$ gelten. Angenommen in diesem Beispiel ist $\mathbf{X}$ die einzige Einflussgröße, und statt des wahren Wertes von $\mathbf{X}$ liegen je Beobachtungseinheit eine Messwiederholung $\boldsymbol{X}^*_{\boldsymbol{ik_i}} = \boldsymbol{X}^*_{\boldsymbol{i1}} := \boldsymbol{X}^*_{\boldsymbol{i}}$ für alle i (also $k_i = 1$) vor, wobei $\boldsymbol{X}^*_i$ bedingt unabhängig von $(\boldsymbol{Y}_i, \boldsymbol{Z}_i)$ gegeben $\boldsymbol{X}_{\boldsymbol{i}}$ ist. Dann gilt vereinfacht für die interessierende Größe $\mathbf{Y}$ die Gleichung $E(\boldsymbol{Y}|\boldsymbol{X}) = \beta_0 + \beta_1\boldsymbol{X}$. Die Schritte der Regressionskalibrierung seien wie folgt:

- Schritt 1: Aus den getroffenen Annahmen folgt $\boldsymbol{Z}_i = \overline{\boldsymbol{Z}_\cdot} = 0$, $\hat{\Sigma}_{zx} = 0$, $\hat{\Sigma}_{zz} = 0$, $\hat{\Sigma}_{xx} = \sigma_x^2$ und $\hat{\Sigma}_{xx} + \hat{\Sigma}_{uu} = \hat{\Sigma}_{x^*x^*} = \sigma^2_{x^*}$. Und somit vereinfacht sich die Formel (10) zur Schätzung der unbeobachteten Variable $\mathbf{X}$ zu

$$E[\widehat{\boldsymbol{X}|\overline{\boldsymbol{X}}^*}] \approx \underbrace{\frac{\widehat{\sigma}_x^2}{\widehat{\sigma}^2_{x^*}}}_{\widehat{\gamma}_1}\boldsymbol{X}^* + \underbrace{\hat{\mu}_{x^*}(1 - \frac{\widehat{\sigma}_x^2}{\widehat{\sigma}_{x^*}})}_{\widehat{\gamma}_0} = \widehat{\boldsymbol{X}} \tag{16}$$

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

$$\hat{\mu}_x = \hat{\mu}_{x^*} = \frac{\sum_{i=1}^n \boldsymbol{X}^*_i}{n}, \tag{17}$$

$$\hat{\Sigma}_{xx} = \widehat{\sigma}_x^2 = \frac{\Sigma_{i=1}^n(\overline{\boldsymbol{X}^*_i} - \hat{\mu}_{x^*})(\boldsymbol{X}^*_i - \hat{\mu}_{x^*})^t}{(n-1)} - \hat{\Sigma}_{uu}. \tag{18}$$

- Schritt 2: Ersetzte die unbeobachtete Variable $\mathbf{X}$ durch die Schätzungen im letzten Schritt, mit dem Wissen aus Caroll et al. [2006] (Kapitel 2.5.0.1, S.38) und

Gustafson [2004] (Kapitel 4.9, S.90) und Anhang B folgt:

$$
\begin{aligned}
E(Y|X^*) &=^1 E(\{E(Y|X, X^*)\}|X^*) && (19)\\
&=^2 E(\{E(Y|X)\}|X^*) && (20)\\
&=^3 E(\{\beta_0 + \beta_1 X\}|X^*) && (21)\\
&= \beta_0 + \beta_1 E(X|X^*) && (22)\\
&\approx \beta_{Rk_0} + \beta_{Rk_1} E(\widehat{X|X^*}) && (23)\\
&\approx \beta_{Rk_0} + \beta_{Rk_1}(\frac{\hat{\sigma}_x^2}{\hat{\sigma}_{x^*}^2}\boldsymbol{X}^* + \hat{\mu}_{x^*}(1 - \frac{\hat{\sigma}_x^2}{\hat{\sigma}_{x^*}})) && (24)\\
&\approx \beta_{Rk_0}\beta_{Rk_1}\hat{\mu}_{x^*}(1 - \frac{\hat{\sigma}_x^2}{\hat{\sigma}_{x^*}}) + (\beta_{Rk_i}\frac{\hat{\sigma}_x^2}{\hat{\sigma}_{x^*}^2})\boldsymbol{X}^* && (25)\\
&\approx \underbrace{\beta_{Rk_0}\beta_{Rk_1}\hat{\gamma}_0}_{\beta_{naiv_0}} + \underbrace{\beta_{Rk_1}\hat{\gamma}_1}_{\beta_{naiv_1}}\boldsymbol{X}^*. && (26)
\end{aligned}
$$

[1] iterierter Erwartungswert, Satz der totalen Wahrscheinlichkeit
[2] X^* differentieller Fehler (Anhang A)
[3] $E(\boldsymbol{Y}|\boldsymbol{X}) = \beta_0 + \beta_1\boldsymbol{X}$

Beachte, β_{naiv_0} und β_{naiv_1} sind die Parameter einer naiven Regression von $\mathbf{Y}$ auf $\boldsymbol{X}^*$, d.h. wenn nicht berücksichtigt wird, dass Messfehler vorliegen. Unter der Bedingung, dass der Ausdruck (16) gilt, kann man die Schätzer der wahren Effekte $\boldsymbol{\beta_0, \beta_1}$ wie folgt extrahieren:

$$
\widehat{\beta}_1 \approx \widehat{\beta}_{Rk_1} = \frac{\widehat{\beta}_{naiv_1}}{\widehat{\gamma}_1}, \qquad \widehat{\beta}_0 \approx \widehat{\beta}_{Rk_0} = \frac{\widehat{\beta}_{naiv_0}}{\widehat{\gamma}_0\widehat{\beta}_1}, \tag{27}
$$

sodass im Grunde genommen die korrigierten Schätzer von $\boldsymbol{\beta_0, \beta_1}$ durch eine Regression von $\mathbf{Y}$ auf $\mathbf{X}^*$ statt auf $\mathbf{X}$ geschätzt werden können.

- Schritt 3: Anhand von Bootstrapping sollen die wahren $\sigma^2_{\beta_{RK_0}}$ und $\sigma^2_{\beta_{RK_1}}$ geschätzt werden. Dazu werden M Bootstrapping-Stichproben gezogen. Durch Wiederholen von Schritt 1 und Schritt 2 in jeder Stichprobe erhält man $\hat{\beta}^{(1)}_{RK_0}$,...,$\hat{\beta}^{(M)}_{RK_0}$ und $\hat{\beta}^{(1)}_{RK_1}$,...,$\hat{\beta}^{(M)}_{RK_1}$.
  Sei j $\in \{0,1\}$, so lässt sich die Varianz der Parameterschätzer wie folgt schätzen:

$$
\hat{\sigma}^2_{\beta_{RK_j}} = \frac{1}{M-1}\sum_{m=1}^{M}(\hat{\beta}^{(m)}_{RK_0} - \overline{\hat{\beta}_{RK_0}})^2.
$$

Die Regressionskalibrierung ist nicht nur in der Theorie anwendbar, sondern auch in der Praxis. Zur Veranschaulichung der praktischen Anwendung sind im Anhang E Rcodes

zur Simulation von Beispielen für die Anwendung der Regressionskalibrierun für $k_i = 1$ und $k_i = 4$ zu finden. Auch in der Praxis wurde die Methode bei Messfehlerproblemen zur Hilfe genommen, was am folgenden Kapitel verdeutlicht werden soll.

6. Beispiel-The Monica Study

In diesem Kapitel wird nach Augustin et al. [2008] die Anwendung der Regressionskalibrierung am realen Beispiel der WHO MONICA (MONitoring of trends and determinants on CArdiovascular disease) illustiert. Dazu werden zunächst die vorliegenden Daten vorgestellt und anschließend die Ergebnisse der Studie nach Anwendung der Regressionskalibrierung präsentiert.

6.1. Die Daten

Die WHO MONICA-Studie wurde 1984/1985 in Süddeutschland durchgeführt.

> Ziel des Gesamtprojektes war es, unter standardisierten Erhebungsbedingungen kardiovaskulär bedingte Erkrankungs- und Todesfälle in definierten Studienregionen vollständig zu erfassen, diese im zeitlichen Verlauf abzubilden, und in Beziehung zu verschiedenen Einflussfaktoren zu setzen

[Mon, accessed:19.10.2014]. Unter anderem hat man sich dafür interessiert, ob die Ernährung einen Einfluss auf Herz-Kreislauf-Erkrankungen hat. In dieser Studie wurde eine Stichprobe von 899 männlichen Personen zwischen 45 und 65 Jahren aufgefordert, ein umfangreiches Tagebuch zu führen, das für sieben aufeinanderfolgende Tage jede Mahlzeit detailiert auflistet. Anhand von bekannten Ernährungsdaten konnten interessierende Ernährungsvariablen, u.a. individuelle Aufnahme von pflanzlichen (PLANT) und tierischen Proteinen (ANIMAL), hergeleitet werden. Dabei bestehen zwei grundlegende Probleme in den Daten, weshalb man vorsichtig mit den Resultaten umgehen sollte. Zum einen sind aus Sicht einiger Epidemiologisten die Messungen basierend auf einem einwöchigen Ernährungsplan nicht repräsentativ für die individuelle langfristige Ernährung, zum anderen sind substantielle Fehlermessungen nicht vermeidbar, obwohl die Werte der Proteinaufnahme mit großer Sorgfalt extrahiert wurden. Aufgrund dessen wurde die Methode der Regressionskalibrierung angewandt. Zu den fehlerhaft gemessenen Variablen wurden noch Störvariablen wie Cholesterin (CHOL) und täglicher Alkoholkonsum (ALC) als metrische Variablen und Bluthochdruck (HYPER) und Raucher (SMOKER) als kategoriale Variablen berücksichtigt. Die Messfehler in diesen Variablen können im Vergleich zu denen in der Proteinaufnahme eher als gering eingeschätzt werden. Neben diesen Einflussgrößen wurden die Zielgrößen „Auftreten von Herzinfakt “und „Tod “durch Sterbe- und Krankheitsfolgeuntersuchungen über mehr als zehn Jahre registriert. Eine detaillierte Beschreibung der Schätzung kann in [Augustin et al., 2008] nachgelesen werden, im Nachfolgenden werden nur Ansatz und Ergebenis präsentiert.

6.2. Ergebnisse

Um die Fragestellung, ob die Ernährung einen Einfluss auf Herz-Kreislauf-Erkrankungen hat, wurde in Augustin et al. [2008] ein Cox-Modell zur Modellierung von Überlebenszeiten angewendet. Wie eingangs bereits erwähnt, waren tierische Proteine (ANIMAL), pflanzliche Proteine (PLANT), Cholesterinspiegel (CHOL), täglicher Alkoholkonsum (ALC), Bluthochdruck (HYPER) und Raucher (SMOKER) Einflussvariablen für die Responsevariablen Erkrankung (MORBILITY) und Sterblichkeit (MORTALITY). Man hat die Möglichkeit, entweder eine naive Schätzung durchzuführen oder eine Schätzung unter Berücksichtigung der Messfehler.Augustin et al. [2008] zufolge wird die Regressionskalibrierung für homoskedastischen Fehlern und mit heteroskedastischen Fehlern anhand von Messwiederholungen vogeschlagen. Die Theorie für die erste Vorgehensweise wurde im Kapitel 3.4. dieser Arbeit dargelegt, zur zweiten wird auf Augustin et al. [2008] (Kapitel 3.3, S.361 ff.) verwiesen.
Für alle Fälle wird angenommen, dass das *Cox's proportional hazard model* den Zusammenhang zwischen der Überlebenszeit und den Einflussgrößen beschreibt. Alle fehlerbehafteten Messungen werden als $\boldsymbol{X}_{ij}$, während alle fehlerfreien Messungen als $\boldsymbol{Z}_{ij}$ bezeichnet werden. Wobei j angibt um welche Kovariable es sich handelt, i um welche Beobachtung und $\underline{\boldsymbol{X}}_{\boldsymbol{i}}^{'} = (\boldsymbol{X_{i1}, ..., X_{ip}})$ bzw. $\underline{\boldsymbol{Z}}_{\boldsymbol{i}}^{'} = (\boldsymbol{Z_{i1}, ..., Z_{ik}})$ ein Vektor der i-ten Beobachtung mit p bzw. k Einflussgrößen darstellt. $\underline{\boldsymbol{X}}_{\boldsymbol{i}}^{\boldsymbol{2}}$ ist ein Vektor der die quadratischen Komponenten von $\underline{\boldsymbol{X}}_{\boldsymbol{i}}^{'}$ enthält. Die individuelle Hazard Rate $\lambda(t|\underline{X_i}, \underline{Z_i})$ hat folgende Form:

$$\lambda(t|\underline{X_i}, \underline{Z_i}) = \lambda_0(t) exp(\underline{\beta_1}^{'} \ \underline{X_i} + \underline{\beta_2}^{'} \ \underline{X_i^2} + \underline{\beta_z}^{'} \ \underline{Z_i}) \tag{28}$$

mit der nicht spezifizierten baseline Hazard Rate $\lambda_0(t)$ und den Paramtern $\beta_0^{'}, \beta_1^{'}, \beta_z^{'}$ (vgl.Augustin et al. [2008], Kapitel 2.2, S.259). Statt $\underline{\boldsymbol{X}}_{\boldsymbol{i}}$ wurden sieben Messwiederholungen (k=7) für jede Beobachtung erhoben, wobei diese Messungen dem klassischen Fehlermodel unterliegen. Um den geschätzten Effekt von tierischen bzw. pflanzlichen Proteinen darstellen zu können, werden folgende quadratische Funktionen zur Hilfe genommen:

$$f(x_{ANIMAL}) = \widehat{\beta}_{ANIMAL} \ x_{ANIMAL} + \widehat{\beta}_{ANIMAL^2} \ x_{ANIMAL}^2, \tag{29}$$

$$g(x_{PLANT}) = \widehat{\beta}_{PLANT} \ x_{PLANT} + \widehat{\beta}_{PLANT^2} \ x_{PLANT}^2, \tag{30}$$

diese sind monoton zum Risikoeffekt der tierischen bzw. pflanzlichen Fette im Cox-Modell (28) (vgl. Augustin et al. [2008], Kapitel 4.2, S.263). Grafisch sind die quadratischen Kurven in Abbildung 1 dargestellt.

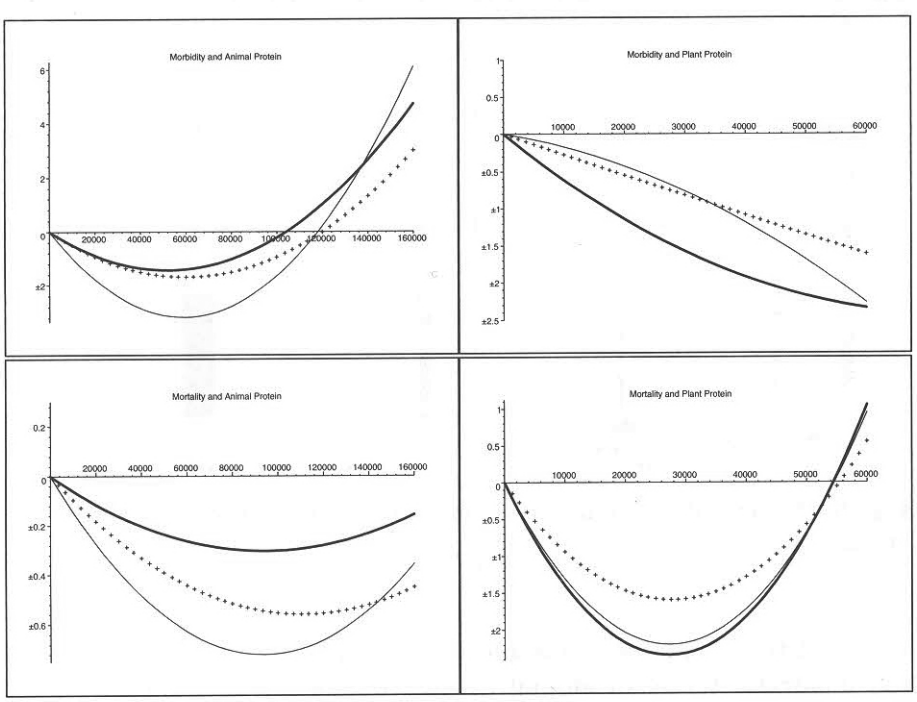

Abbildung 3: Geschätzter Gesamteinfluss der tierischen (*linke Grafiken*) und pfanzlichen Proteine (*rechte Grafiken*) auf die Herzinfakt-Erkrankung (*obere Grafiken*) und die Sterblichkeit (*untere Grafiken*), geschätzt durch naive Schätzung (*gepunktete Linie*), nach Messfehlerkorrektur für homoskedastische Fehler (*dünne, durchgezogene Linie*), nach Messfehlerkorrektur für heteroskedastische Fehler (*dicke, durchgezogene Linie*).(Aus Augustin et al. [2008], S.268)

Die abgebildeten Kurven sind abhängig von den β-Schätzern aus der naiven Schätzung und der Schätzung aus den verschiedenen Regressionskalibrierungsansätzen, eingesetzt in Formel (29), (30). Auffällig ist, dass alle drei Kurven sich stark unterscheiden. Besonders die rechte obere Grafik verdeutlicht, dass die Nichtberücksichtigung von Messfehlern starken Einfluss auf die Schätzung hat, in der naiven Schätzung würde man von beinahe linearem Einfluss ausgehen, während nach der Korrektur nicht lineare Einflüsse deutlich werden. Zusammengefasst führen im Allgemeinen zu hohe oder zu niedrige Aufnahme von Proteinen zu einer Risikoerhöhung.
An diesem Beispiel wird nochmal ein Vorteil der Regressionskalibrierung verdeutlicht; diese Methode greift nämlich auch dort, wo andere Messfehlerkorrekturverfahren scheitern, nämlich bei der Anwendung von quadratischen Kovariablen bzw. Kovariablen höheren Grades.

7. Fazit

Regressionskalibrierung stellt eine sehr effektive Methode im Umgang mit fehlerhaft gemessenen Einflussgrößen dar. Im Folgenden soll eine Übersicht über die Vor- und Nachteile gegeben werden (vgl. Caroll et al. [2006], Kapitel 4.1, S.65 ff.):

Vorteile

- effektive Methode im Umgang mit fehlerhaften Einflussgrößen
- auf viele Modelle anwendbar (GLM)
- einfache Berechnung
- anschließende Standardanalysen noch möglich
- Reduzierung der Bias
- wandelt klassische Fehler in Berkson um
- Regression möglich, obwohl wahres $\mathbf{X}$ nicht beobachtet
- keine extra Implementierung in statistische Programme nötig, Stata besitzt aber eine eigene Prozedur, die in Hardin et al. [2003] näher beschrieben wird.

Nachteile

- nur approximatives Verfahren
- Schätzer sind nicht unbedingt konsistent
- die Berechnung der Regression von $\mathbf{X}$ auf ($\mathbf{X}^*$,$\mathbf{Z}$) stellt eine Herausforderung dar, da $\mathbf{X}$ nicht beobachtbar ist

Im Rahmen dieser Arbeit haben wir die drei Schritte der Regressionskalibrierung kennengelernt und für klassische additive Fehlermodelle näher erläutert. Anzumerken ist aber, dass noch weitere Szenarien denkbar sind, die den Umfang dieser Arbeit übersteigen. So ist beispielsweise die Regressionskalibrierung auch auf multiplikative Fehlermodelle anwendbar, indem man diese durch eine log-Transformation in addtive Fehler umwandelt und das Regressionskalibrierungverfahren anpasst (näheres Caroll et al. [2006], Kapitel 4.5). Auch ist eine Anwendung auf heteroskedastische Regression möglich, die Schätzung von $\Sigma_{z,z}$ und Σ_{xz} entspricht jener aus Kapitel 3.4, für die Schätzung Σ_{xx} und Σ_{uu} muss die Abhängigkeit zwischen den Individuen berücksichtigt werden (weiteres Augustin et al. [2008], Kapitel 4, S.266 ff.). Die vorgestellten Regressionskalibrierungsverfahren funktionieren für Modelle der generalisierten linearen Regression, für höhere nichtlineare Probleme sind Transformationen der Regressionskalibrierungsverfahren erforderlich n (näheres Caroll et al. [2006], Kapitel 4.7). Abgesehen von der Regressionskalibrierung

gibt es weitere Korrekturverfahren für Messfehler in den Variablen, z.B. die Simulation-Extrapolation-Methode (Simex) (siehe Anhang F) oder die korrigierten Scorefunktion. Die letzten beiden genannten Methoden sind zwar schwierig zu berechnen, aber in vielen Fällen liefern sie bessere Schätzer.

A. Nicht-differentieller Fehler

Man unterscheidet zwischen *differeniellem Fehler* und *nicht-differentiellem Fehler*. Nach diesen Fehlertypen zu unterscheiden gilt als essentiell beim Umgang mit Messfehlern. Ein *nicht-differentieller Fehler* liegt nach Caroll et al. [2006] (Kapitel 2.5, S.30) vor, wenn der Messfehler $\boldsymbol{X}^*$ keine zusätzlichen Informationen, außer den Informationen die ohnehin durch $(\boldsymbol{X}, \boldsymbol{Z})$ gegeben sind, zu der Responsevariable $\mathbf{Y}$ enthält. Wenn also die Verteilung von $\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{X}^*, \boldsymbol{Z}$ gleich der Verteilung von $\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z}$ ist, spricht man von einem *nicht-differentiellen* Fehlermodell von X^* bzw. X^* ist ein Surrogat. Beispielsweise gilt dann:

$$E[Y|X, X^*, Z] = E[Y|X, Z]. \tag{31}$$

Diese Umformung stellt eine Grundlage der Regressionskalibrierungsmethode dar. Der Vorteil von einem *nicht-differentiellen Fehler* ist, dass, selbst wenn die wahren Einflussgrößen nicht beobachtbar sind, eine Parameterschätzung für die Responsevariable, gegeben den wahren Einflussgrößen, möglich ist. Abgesehen von einigen Ausnahmen ist das für den *differentiellen Fehler* nicht möglich. Warum das Vorliegen einer *nicht-differentiellen Fehlers* nützlich ist, kann anhand folgender Umformungen bei einer linearen Regression nachvollzogen werden.

$$\begin{aligned}
E(Y|X^*, Z) &=^1 E(\{E(Y|X, X^*, Z)\}|X^*, Z) && (32)\\
&=^2 E(\{E(Y|X, Z)\}|X^*, Z) && (33)\\
&=^3 E(\{\beta_0 + \beta_1 X + \beta_2 Z\}|X^*, Z) && (34)\\
&= \beta_0 + \beta_1 E(X|X^*, Z) + \beta_2 E[Z|X^*, Z] && (35)\\
&= \beta_0 + \beta_1 E(X|X^*, Z) + \beta_2 Z && (36)
\end{aligned}$$

[1] iterierter Erwartungswert, Satz der totalen Wahrscheinlichkeit
[2] X^* differentieller Fehler
[3] $E(\boldsymbol{Y}|\boldsymbol{X}) = \beta_0 + \beta_1 \boldsymbol{X}$
Eine lineare Regression auf den beobachteten Daten $(\boldsymbol{Y}, \boldsymbol{X^*}, \boldsymbol{Z})$ kann zu einer linearen Regression von $\mathbf{Y}$ auf $E[\boldsymbol{X}|\boldsymbol{X^*}, \boldsymbol{Z}]$ und Z vereinfacht werden. D.h. man beginnt zwar mit einer Regression auf den beobachteten Daten, aber die Umformung zeigt, dass diese in Beziehung zum wahren $\mathbf{X}$ bedingt $(\boldsymbol{X^*}, \boldsymbol{Z})$ gesetzt werden kann. An dieser Stelle knüpft Schritt 2 in Kapitel 5 an. Für genaueres wird auf Caroll et al. [2006] (Kapitel 2.5, S.36 ff.) verwiesen.

B. Berkson- Fehler

In diesem Teil der Arbeit werden die Unterschiede vom Berkson- Fehler gegenüber klassischen Fehlern nach Caroll et al. [2006] (Kapitel 2.2) erläutert. Um mit Messfehlern zu arbeiten, ist es zwingend erforderlich ein Fehlermodell zu spezifizieren. Dabei ist es von Interessse, ob eine Annahme über die beobachtete Variable $\boldsymbol{X}^*$ gegeben $\mathbf{X}$ gemacht wird oder andersherum. Es gibt also verschiedene Fehlermodelle, die zu erst genannte Schlussichtung entspricht dem klassischen Fehlermodell und die andere Richtung das Berkson- Fehlermodell (vgl. Buonaccors [1986], Kapitel 1.4.1, S.6ff). Auf Grundlage von Caroll et al. [2006] (Kapitel 2.2.2, S.27) bietet folgende Tabelle eine Übersicht:

	Klassisches Fehlermodell	Berkson- Fehlermodell
Gegebene Verteilung für	$X^*\|X$	$X\|X^*$
Modelliert durch	$X^* = X + U$	$X = X^* + U$
Verteilung des Fehlers	U|X $\sim N(0, \sigma^2)$	U|X $\sim N(0, \sigma^2)$
Stochastisch unabhängig	U , X	U , X^*
Erwartungswert von U	$E[U\|X] = 0$	$E[U\|X^*] = 0$
Außerdem gilt noch:		
	$E[X^*\|X] = X$	$E[X\|X^*] = X^*$
	$V(X^*) = V(X + U) =$ $V(X) + V(U) > V(X)$	$V(X) = V(X^* + U) =$ $V(X^*) + V(U) > V(X^*)$

Insbesondere erhält man in der linearen Regression, beim vorliegen eines Berkson-Fehlers, unverzerrte KQ-Schätzungen [Schneeweiß and Mittag, 1986] (Kapitel 1.3.1, S.32 ff.), weshalb das Vorliegen eines Berkson- Modell bevorzugt wird. Daher ist ein weiterer Vorteil der Regressionskalibrierung, dass diese Methode einen klassischen Fehler in einen Berkson- Fehler umwandelt (Caroll et al. [2006], Kapitel 2.2.3, S.28 ff.). Einen Beweis hierfür findet man in Caroll et al. [2006] (Kapitel 3.2.2, S.44 ff.).

C. Instrumentaldatan

Grundsätzlich gelten für Instrumentaldaten **T** nach Caroll et al. [2006] (Kapitel 6, S.129 ff.) folgende Eigenschaften:

- $\boldsymbol{T}$ ist abhängig von $\boldsymbol{X}$
- $\boldsymbol{T}$ ist unkorreliert mit Fehler $\boldsymbol{U} = \boldsymbol{X}^* - \boldsymbol{X}$
- $\boldsymbol{T}$ unkorreliert mit $\boldsymbol{\epsilon} = \boldsymbol{Y} - E[\boldsymbol{Y}|\boldsymbol{Z}, \boldsymbol{X}]$

Außerdem soll gelten $\boldsymbol{T}$ ist *unverzerrt* für $\boldsymbol{X}$ d.h. eine
Regression von T $\sim$ Z+X^* entspricht einer Regression von X $\sim$ Z+X^*

$$E[\boldsymbol{T}|\boldsymbol{X}^*, \boldsymbol{Z}] = E[\boldsymbol{X}|\boldsymbol{X}^*, \boldsymbol{Z}]$$

Die Variable **T** ist somit ein besonders gutes Surrogat für die Variable **X**.

D. Überprüfen der Schätzung in Schritt 1 der RK

Da Schritt 2 der Regressionskalibrierung nur sinnvoll ist, wenn die Schätzung in Schritt 1 gültig ist, ist es naheliegend bereits Schritt 1 zu prüfen. Im Fall von Validierungsdaten und Instrumentaldaten können gewöhnliche Regressionsdiagnosen durchgeführt werden. Beispielsweise Residuenplot mit $\mathbf{X}$ bzw. $\mathbf{T}$ als X-Achse und $\boldsymbol{\epsilon} = \boldsymbol{X} - \boldsymbol{E}[\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{X}^*]$ bzw. $\boldsymbol{\epsilon} = \boldsymbol{T} - \boldsymbol{E}[\boldsymbol{T}|\boldsymbol{Z}, \boldsymbol{X}^*]$ als Y-Achse.
Im Falle von Validierungsdaten liegen keine wahren $\mathbf{X}$-Werte vor, sodass einige vorüberlegungen nötig sind:
Für gewöhnlich liegen nicht für alle Beobachtungen gleich viele Wiederholungsmessungen vor, d.h. es liegen partielle Wiederholungsdaten vor. Dies reicht aber aus um die Schätzung im ersten Schritt der Regressionskalibrierung zu überprüfen. Basierend auf den Annahmen in Kapitel 3.4.1 kann man $\boldsymbol{\epsilon}_i^* = \boldsymbol{X^*}_{i2} - E[\boldsymbol{X}_{i2}^*|\boldsymbol{Z_i}, \boldsymbol{X^*}_{i1}]$ durch einsetzen von Formel (12) in zwei Bestandteile trennen. Der erste Teil ist das Residuum aus der Schätzung von dem wahren X-Wert, der zweite Teil das Residium aus der Schätzung des Fehlers.

$$
\begin{aligned}
\boldsymbol{\epsilon}_i^* &= \boldsymbol{X}_{i2}^* - E[\boldsymbol{X}_{i2}^*|\boldsymbol{Z_i}, \boldsymbol{X^*}_{i1}] \\
&= (\boldsymbol{X}_i + \boldsymbol{U}_{i2}) - (E[\boldsymbol{X}_{i2}|\boldsymbol{Z}_i, \boldsymbol{X^*}_{i1}] + E[\boldsymbol{U}_{i2}|\boldsymbol{Z}_i, \boldsymbol{X}_{i1}^*]) \\
&= (\boldsymbol{X}_i - E[\boldsymbol{X}_i|\boldsymbol{Z}_i, \boldsymbol{X}_{i1}^*]) \ + \ (\boldsymbol{U}_{i1} - E[\boldsymbol{U}_{i2}|\boldsymbol{Z}_i, \boldsymbol{X}_{i1}^*]) \\
&= \boldsymbol{\epsilon}_i + \tilde{\boldsymbol{\epsilon}}_i
\end{aligned}
$$

Auf der Y-Achse des Residuenplots können nun $\boldsymbol{\epsilon}_i^* = \boldsymbol{\epsilon}_i + \tilde{\boldsymbol{\epsilon}}_i$ abgetragen werden, die Werte der X-Achse sind die Werte von $\boldsymbol{X}_{i2}^*$. Auch diese Werte lassen sich sozusagen in einen wahren Teil $\boldsymbol{X}_i$ und einen Fehlerteil $\boldsymbol{U}_i$ zerlegen, sodass es nun aus den dargelegten Residuenplot Tendenzen ersichtlich sind, ob die Schätzung von $\mathbf{X}$ in Schritt 1 hinreichen gut ist.

E. Beispiel: Regressionskalibrierung Rcode

In diesem Anhang werden die Rcodes bereitgestellt, die im laufe der Arbeit entstanden sind. An einigen Stellen sind zusätzlich R-Outputs enthalten, die zum Verständnis beitragen sollen.
Zu Beginn (E.1) werden Daten generiert mit diesen im weiteren gearbeitet wird. E.2 und E.3 untersucht die Aufnahme von einer Dummyvariablen. Alle Schritte der Regressionskalibrierung, angewendet auf verschiedene Datentypen, sind im Anhang E.4 bis E.6 vollständig in R dargestellt und im Anhang E.7 wird die Güte der Messfehlerkorrekturen verglichen.

E.1. Daten generieren

Zu Beginn wurden Daten zufällig generiert. Die Daten enthalten insgesamt 500 Beobachtungen mit den Spalten $\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{X^*}, \boldsymbol{X_2^*}, \boldsymbol{X_3^*}, \boldsymbol{X_4^*}$, wobei $\boldsymbol{X_k^*}$ Messwiederholungen von $\boldsymbol{X}$ sind. Auf Basis dieser Daten werden in den folgenden Kapiteln Validierungsdaten und Wiederholungsdaten modelliert.

DatenGenerieren.R

```
1  ###############Regressionskalibrierung mit Validierungsdaten###############
2  ####################Vorbereitung Daten generieren#########################
3  set.seed(123)
4  #wahre X-Werte
5  x<-rnorm(500, mean=0, sd=1)
6
7  #wahre Parameter
8  intercept <- 2
9  beta <- 1
10
11 #wahre Y-Werte
12 y<-intercept+beta*x+0.3*rnorm(500)
13
14 #Variable Z wegelassen (Z=0) (zur Vereinfachung)
15
16 #X* fehlerhafte Messung von X
17 xSt<-x+0.7*rnorm(500) #X + fehler
18
19 #zweite fehlerhafte Messung (wdh)
20 xSt2<-x+0.7*rnorm(500) #X + fehler
21
22 #dritte fehlerhafte Messung (wdh)
23 xSt3<-x+0.7*rnorm(500) #X + fehler
24
25 #vierte fehlerhafte Messung (wdh)
26 xSt4<-x+0.7*rnorm(500) #X + fehler
27
28 #zusammengefasste Daten (ACHTUNG: Werte fuer X eig. nicht fuer alle
29 #500 Beobachtungen vorhanden)
30 dataOrigin<-data.frame(y, x, xSt, xSt2, xSt3, xSt4) #500 Beobachtungen (
       Vektoren)
31 head(dataOrigin)
32
33 ##       y           x         xSt         xSt2         xSt3        xSt4
34 ## 1 1.8892585  0.00945713  0.9162795  1.1296218 -0.48160290 -0.5732710
35 ## 2 1.7286159 -0.39682557 -1.1995464 -1.7046212  0.09411859  0.8066664
36 ## 3 3.5898259  1.00002565 -0.4112711  2.6786101  2.96909947  0.7918497
37 ## 4 1.8310851 -0.45995936 -0.7277402 -0.2497571 -1.51189905 -0.4932013
38 ## 5 1.3305249 -0.58564377 -0.2886400  0.3472726 -0.91507069  0.1443980
39 ## 6 0.1866607 -1.58644314 -1.4430657 -1.1518545 -1.20613112 -1.4235559
```

E.2. Vergleich RK mit/ohne Dummyvariable in Validierungsdaten

In Kapitel 3.2 wurde erwähnt, dass Caroll et al. [2006] (Kapitel 4.4, S.70) die Einführung einer Dummyvariable, bei einer Regressionskalibrierung auf Validierungsdaten, empfiehlt, die angibt ob es sich bei der Schätzung von $\boldsymbol{Y}$ in Schritt 2 um die wahren $\boldsymbol{X}$ oder um die geschätzten $\widehat{\boldsymbol{X}}$ aus Schritt 1 handelt. Folgende Simulation vergleicht die Schätzung bei Aufnahme einer Dummyvariable gegenüber der nicht Aufnahme einer Dummyvariable.

regcalDummyVergleich.R

```
1  ###############Regressionskalibrierung mit Validierungsdaten################
2  ###generieren von Daten aus dataOrigin, mit diesen im folgenden gearbeitet wird##
3  #####angenommen es liegen nur fuer die Haelfte(250) Validierungsdaten vor#######
4
5  #Nur fuer Validierungsdaten liegen Y, X* UND wahre X-Werte vor
6  validierung <- sample(500,250, replace=FALSE)
7  validierungSub <- dataOrigin[validierung, ][c(1,2,3)]
8  validierungSub$Valid <- 1 #Dummyvariable
9
10 #Fuer die restlichen Daten liegen keine wahren X-Werte vor
11 NotvalidierungSub <- dataOrigin[setdiff(seq(1:500), validierung), ][c(1,2,3)]
12 NotvalidierungSub[ ,"x"] <- NA
13 NotvalidierungSub$Valid <- 0 #Dummyvariable
14
15 #Daten auf die die Regressionskalibierung spaeter angewendet wird.
16 dataValid<-merge(validierungSub, NotvalidierungSub, all=T)
17 head(dataValid)
18
19 ##         y           x        xSt      Valid
20 ## 1 -0.4952122         NA -1.214837      0
21 ## 2 -0.4139702         NA -2.902119      0
22 ## 3 -0.3307201 -1.533457 -1.305155      1
23 ## 4 -0.3231430         NA -2.346582      0
24 ## 5 -0.3101120 -2.525401 -2.347733      1
25 ## 6 -0.3087614 -2.153913 -2.736493      1
26
27
28 ############Bootstrap zur Kontrolle mit oder ohne Dummy ###################
29 #100 Bootstrapstichproben
30 B <- 100
31 i <- 1
32 beta_0 <- c()
33 beta_1 <- c()
34 beta_0_Dummy <- c()
35 beta_1_Dummy <- c()
36 dataValid$x_hat <- NA
37
38 while( i <= B){
39
40   #Bootstrap ziehen
41   #dataValid besitzt zwei verschiedene Datenstrukturen (Validierungsdaten
42   #und Daten ohne wahren X-Werte)
43   #Daten muessen vor der Bootstrap-Ziehung nach den
44   #Strukturen getrennt werden (siehe Kapitel 4)
45
46   #Bootstrap von Vektoren aus Validierungsdaten
47   beob1 <- rownames(dataValid[dataValid$Valid==1, ])
48   #Ziehen Vektoren mit zuruecklegen aus Validierungsdaten
49   which1 <- sample(beob1, length(beob1), replace=TRUE)
```

```
  bootValid1 <- dataValid[dataValid$Valid==1, ][which1, ]

  #Bootstrap von Vektoren aus Daten ohne wahren X-Werte
  beob0 <- rownames(dataValid[dataValid$Valid==0, ])
  #Ziehen Vektoren mit zuruecklegen aus Validierungsdaten
  which0 <- sample(beob0, length(beob0), replace=TRUE)
  bootValid0 <- dataValid[dataValid$Valid==0, ][which0, ]

  #Bootstrap-Daten
  dataBoot <- merge(bootValid1, bootValid0, all=T)

###Schritt 1: Eine Regression von X auf (X*,Z=0) rechnen aus Validierungsdaten

#Beachte: Zeile, wo x=NA werden automatisch im lm() nicht beruecksichtigt
lm_xSt <- lm(x~xSt, data=dataBoot)

#Schaetzer anwenden auf alle X* fuer die keine wahren X vorliegen und
#in neue Variable x_hat speichern
dataBoot[(dataBoot$Valid==0),]$x_hat <- coef(lm_xSt)[1]+coef(lm_xSt)[2]*dataBoot[(
    dataBoot$Valid==0),]$xSt

###Schritt 2: nicht vorhandene wahre X durch die Schaetzung x_hat aus
###Regression in Schritt 1 ersetzen.
#x_reg soll die wahren X-Werte enthalten und wo die wahren X-Werte fehlen
#enthaelt sie die geschaetzten X-Werte
dataBoot$x_reg <- dataBoot[,"x"]
dataBoot[dataBoot$Valid==0,]$x_reg <- dataBoot[dataBoot$Valid==0,]$x_hat

##############################Ohne Dummy ###################################

lm_rk <- lm(y~x_reg, data=dataBoot)


beta_0[i] <- lm_rk$coef[1]
beta_1[i] <- lm_rk$coef[2]


##############################mit Dummy#####################################

lm_rkD <- lm(y~x_reg+Valid, data=dataBoot)

beta_0_Dummy[i] <- lm_rkD$coef[1]
beta_1_Dummy[i] <- lm_rkD$coef[2]

i <- i+1
}


########################Vergleich Kennzahlen##############################

mean_b_0 <- mean(beta_0)
mean_b_1 <- mean(beta_1)
mean_b_0_Dummy <- mean(beta_0_Dummy)
mean_b_1_Dummy <- mean(beta_1_Dummy)
var_b_0 <- var(beta_0)
var_b_1 <- var(beta_1)
var_b_0_Dummy <- var(beta_0_Dummy)
var_b_1_Dummy <- var(beta_1_Dummy)
bias_b_0 <- mean(beta_0-intercept)
bias_b_1 <- mean(beta_1-intercept)
bias_b_0_Dummy <- mean(beta_0_Dummy-beta)
bias_b_1_Dummy <- mean(beta_1_Dummy-beta)
```

```
111 MSE_b_0 <- bias_b_0^2+var_b_0
112 MSE_b_1 <- bias_b_1^2+var_b_1
113 MSE_b_0_Dummy <- bias_b_0_Dummy^2+var_b_0_Dummy
114 MSE_b_1_Dummy <- bias_b_1_Dummy^2+var_b_1_Dummy
115
116 Ohnedummy <- c(intercept,beta,mean_b_0,mean_b_1,var_b_0,var_b_1,bias_b_0,
117                bias_b_1,MSE_b_0,MSE_b_1 )
118 names(Ohnedummy) <- c("Wahr Beta0","wahr Beta1","Mean(beta0)","Mean(beta1)","V(
        beta0)","V(beta1)",
119                       "Bias(Beta0)","Bias(Beta1)","MSE(Beta0)","MSE(Beta1)")
120 Mitdummy <- c(intercept,beta,mean_b_0_Dummy,mean_b_1_Dummy,var_b_0_Dummy,
121               var_b_1_Dummy,bias_b_0_Dummy,bias_b_1_Dummy,MSE_b_0_Dummy,
122               MSE_b_1_Dummy)
123 names(Mitdummy) <- c("Wahr Beta0","wahr Beta1","Mean(beta0)","Mean(beta1)",
124                      "V(beta0)","V(beta1)","Bias(Beta0)","Bias(Beta1)","MSE(Beta0)
                           ","MSE(Beta1)")
125
126 tabelle<-cbind(Ohnedummy,Mitdummy)
127 tabelle
128 ##                 Ohnedummy      Mitdummy
129 ## Wahr Beta0    2.0000000000  2.0000000000
130 ## wahr Beta1    1.0000000000  1.0000000000
131 ## Mean(beta0)   2.0067515406  1.9953005075
132 ## Mean(beta1)   0.9741947479  0.9747533306
133 ## V(beta0)      0.0009687363  0.0033303455
134 ## V(beta1)      0.0006689308  0.0006676021
135 ## Bias(Beta0)   0.0067515406  0.9953005075
136 ## Bias(Beta1) -1.0258052521 -0.0252466694
137 ## MSE(Beta0)    0.0010143196  0.9939534458
138 ## MSE(Beta1)    1.0529453460  0.0013049964
```

Zeile 127-138 zeigt eine Vergleichstabelle, vor allem am MSE von $\widehat{\beta_1}$ wird deutlich, dass eine Modellierung mit Dummyvariable zur besseren Schätzung führt, wobei man allerdings anmerken muss, dass die Anpassung eines Modell mit steigendem Paramter steigt.

E.3. Vergleich RK mit/ohne Dummyvariable in Validierungsdaten-Plot

regcalDummyPlot.R

```
1  ###############Regressionskalibrierung mit Validierungsdaten################
2  ###generieren von Daten aus dataOrigin, mit diesen im folgenden gearbeitet wird##
3  #####angenommen es liegen nur fuer die Haelfte(250) Validierungsdaten vor#######
4
5  #Nur fuer Validierungsdaten liegen Y, X* UND wahre X-Werte vor
6  validierung <- sample(500,250, replace=FALSE)
7  validierungSub <- dataOrigin[validierung, ][c(1,2,3)]
8  validierungSub$Valid <- 1
9
10 #Fuer die restlichen Daten liegen keine wahren X-Werte vor
11 NotvalidierungSub <- dataOrigin[setdiff(seq(1:500), validierung), ][c(1,2,3)]
12 NotvalidierungSub[ ,"x"] <- NA
13 NotvalidierungSub$Valid <- 0
14
15 #Daten auf die die Regressionskalibierung spaeter angewendet wird.
16 dataValid <- merge(validierungSub, NotvalidierungSub, all=T)
17 head(dataValid)
18
19 ##       y              x     xSt     Valid
20 ## 1 -0.4952122         NA -1.214837     0
21 ## 2 -0.4139702         NA -2.902119     0
22 ## 3 -0.3307201 -1.533457 -1.305155     1
23 ## 4 -0.3231430         NA -2.346582     0
24 ## 5 -0.3101120 -2.525401 -2.347733     1
25 ## 6 -0.3087614 -2.153913 -2.736493     1
26
27 ####interessiert an Regression von Y auf (X,Z=0), aber eigentlich X nicht
28 ####beobachtet
29 ##visualisiern der Effekte der naiven Schaetzung und der korrigierten
30 ##Schaetzung mit RK
31
32 #plot mit wahren Werten (aus dataOrigin)
33 par(xpd=F)
34 plot(dataOrigin$x, dataOrigin$y, col="black", pch=18, ylab="Y", xlab="")
35 abline(lm(y~x, dataOrigin), col="black", lwd=3)
36 par(xpd=T)
37 text(-2,-3, labels="X", col="black")
38
39 #eine Regression von Y auf (X*,Z=0) statt auf (X,Z) fuehrt zur Verzerrung
40 par(xpd=F)
41 points(dataValid$xSt, dataValid$y, col="green", pch=8)
42 abline(lm(y~xSt, data=dataValid), col="green", lwd=3, lty=4)
43 par(xpd=T)
44 text(-1.3, -3, labels="X*", col="green")
45
46 ####Regressionskalibrierung Verfahren (siehe Kapitel 2.2)
47
48 ###Schritt 1: Eine Regression von X auf (X*,Z=0) rechnen aus Validierungsdaten
49
50 #Beachte: Zeile, wo x=NA werden automatisch in lm() nicht beruecksichtigt
51 lm_xSt <- lm(x~xSt, data=dataValid)
52 summary(lm_xSt)
53
54 #Schaetzer anwenden auf alle X* fuer die keine wahren X vorliegen und
55 #in neue Variable x_hat speichern
56 dataValid$x_hat <- NA
57 dataValid[(dataValid$Valid==0),]$x_hat<-coef(lm_xSt)[1]+coef(lm_xSt)[2]*
```

```
  dataValid[(dataValid$Valid==0),]$xSt

##############################Ohne Dummy ##################################

###Schritt 2: nicht vorhandene wahre X-Werte durch die Schaetzung x_hat aus
###Regression in Schritt 1 ersetzen.
#x_reg enthaelt die wahren X-Werte und wo die wahren X-Werte fehlen
#enthaelt sie die geschaetzten X-Werte
dataValid$x_reg<-dataValid[,"x"]
dataValid[dataValid$Valid==0,]$x_reg <- dataValid[dataValid$Valid==0,]$x_hat

#plot werte nach RK
par(xpd=F)
points(dataValid$x_reg, dataValid$y, col="blue", pch=1)

lm_rk <- lm(y~x_reg, data=dataValid)
summary(lm_rk)
#RK RegGerade nahe der wahren RegGerade
abline(lm_rk, col="blue", lwd=3, lty=3)
par(xpd=T)
text(0, -3,labels="X_Rk_oDummy", col="blue")
par(xpd=F)


##############################mit Dummy##################################

###Schritt 2: nicht vorhandene wahre X-Werte durch die Schaetzung x_hat aus
###Regression in Schritt 1 ersetzen.
#x_reg enthaelt die wahren X-Werte und wo die wahren X-Werte fehlen
#enthaelt sie die geschaetzten X-Werte
dataValid$x_reg <- dataValid[,"x"]
dataValid[dataValid$Valid==0,]$x_reg <- dataValid[dataValid$Valid==0,]$x_hat

#plot Werte nach RK
par(xpd=F)
points(dataValid$x_reg, dataValid$y, col="red", pch=1)

lm_rkVD <- lm(y~x_reg+Valid, data=dataValid)
summary(lm_rkVD)
#RK RegGerade fuer Valid=0
abline(coef(lm_rkVD)[1],coef(lm_rkVD)[2],col="red", lwd=3, lty=2)
#RK RegGerade fuer Valid=1
abline(coef(lm_rkVD)[1]+coef(lm_rkVD)[3],coef(lm_rkVD)[2],col="orange",
       lwd=3, lty=2)
par(xpd=T)
text(2, -3,labels="X_Rk_Dummy", col="red")
par(xpd=F)
```

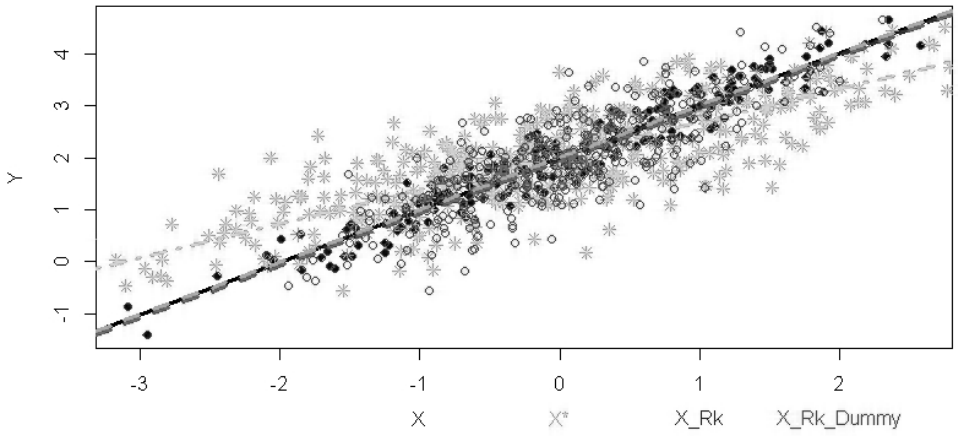

Abbildung 4: *Schwarze durchgezogene Linie* ist die wahre Regressionsgerade. *Grün gepunktet und gestrichelte Linie* entspricht der Regressionsgerade aus der naiven Schätzung. Die restlichen enstammen aus der Schätzung nach Anwendung der Regressionskalibrierung. Die *blaue gepunktete Linie* ist die Gerade aus der Schätzung ohne Dummyvariable, diese wird allerdings von der *gestrichelten roten* und der *gestrichelten orangen* Gerade verdeckt. Ersteres ist die Schätzung, wenn die Dummyvariable **Valid**=0 ist und zweiteres, wenn die Dummyvariable **Valid**=1 ist.

E.4. RK mit Validierungsdaten

In diesem Unterkapitel befindet sich der Rcode zur Regressionskalibrierung angewendet auf Validierungsdaten. In diesem Fall wird angenommen, dass für die Hälfte der Beobachtungen Validierungsdaten vorliegen. Für das Bootstrap-Verfahren im dritten Schritt der Regressionskalibrierung wurde das Resampling Vectors- Verfahren angewendet.

regcalValid.R

```
#####generieren von Daten aus dataOrigin, mit diesen im folgenden gearbeitet wird
#angenommen es liegen nur fuer die Haelfte (250) Validierungsdaten vor

#Nur fuer Validierungsdaten liegen Y, X* UND wahre X vor
validierung<-sample(500,250, replace=FALSE)
validierungSub<-dataOrigin[validierung, ][c(1,2,3)]
validierungSub$Valid<-1

#Fuer die restlichen Daten liegen keine wahren X-Werte vor
NotvalidierungSub<-dataOrigin[setdiff(seq(1:500), validierung), ][c(1,2,3)]
NotvalidierungSub[ ,"x"]<-NA
NotvalidierungSub$Valid<-0

#Daten auf die die Regressionskalibierung spaeter angewendet wird.
dataValid<-merge(validierungSub, NotvalidierungSub, all=T)
head(dataValid)
#         y          x         xSt     Valid
# 1 -1.1167461 -3.295136 -3.634460     1
# 2 -0.7145003         NA -1.903810     0
# 3 -0.5470220 -2.103604 -2.585469     1
# 4 -0.5008263         NA -2.003854     0
# 5 -0.4592409 -2.100544 -2.134435     1
# 6 -0.4576016         NA -2.202306     0

####interessiert an Regression von Y auf (X,Z=0), aber eigentlich X nicht
####beobachtet
##visualisiern der Effekte der naiven Schaetzung und der korrigierten
##Schaetzung mit RK

#plot der wahren Werte (aus dataOrigin)
par(xpd=F)
plot(dataOrigin$x, dataOrigin$y, col="black", pch=18, ylab="Y", xlab="")
abline(lm(y~x, dataOrigin), col="black", lwd=3)
par(xpd=T)
text(0,-3, labels="X", col="black")

#eine Regression von Y auf (X*,Z=0) statt auf (X,Z) fuehrt zur Verzerrung
#(aus dataValid)
par(xpd=F)
points(dataValid$xSt, dataValid$y, col="green", pch=8)
abline(lm(y~xSt, data=dataValid), col="green", lwd=3, lty=4)
par(xpd=T)
text(1, -3, labels="X*", col="green")


#########################################################################
####Regressionskalibrierung Verfahren (siehe Kapitel 2.2)

###Schritt 1: Eine Regression von X auf (X*,Z=0) rechnen aus
###Validierungsdaten

#Beachte: Zeile, wo x=NA werden automatisch nicht beruecksichtigt
```

```
lm_xSt<-lm(x~xSt, data=dataValid)

#Schaetzer anwenden auf alle X* fuer die keine wahren X vorliegen uns
#in neue Variable x_hat speichern
dataValid$x_hat<-NA
dataValid[(dataValid$Valid==0),]$x_hat<-coef(lm_xSt)[1]+coef(lm_xSt)[2]*
  dataValid[(dataValid$Valid==0),]$xSt
head(dataValid)

##         y         x          xSt     Valid     x_hat
## 1 -1.1167461 -3.295136 -3.634460     1         NA
## 2 -0.7145003        NA -1.903810     0 -1.216047
## 3 -0.5470220 -2.103604 -2.585469     1         NA
## 4 -0.5008263        NA -2.003854     0 -1.280737
## 5 -0.4592409 -2.100544 -2.134435     1         NA
## 6 -0.4576016        NA -2.202306     0 -1.409060

###Schritt 2: nicht vorhandene wahre X durch die Schaetzung x_hat aus
###Regression in Schritt 1 ersetzen.
#x_reg enthaelt die wahren X-Werte und wo die wahren X-Werte fehlen
#enthaelt sie die geschaetzten X-Werte, speichern in neue Variable x_reg
dataValid$x_reg<-dataValid[,"x"]
dataValid[dataValid$Valid==0,]$x_reg<-dataValid[dataValid$Valid==0,]$x_hat
head(dataValid)

##         y          x       xSt     Valid     x_hat    x_reg
## 1 -1.1167461 -3.295136 -3.634460     1         NA -3.295136
## 2 -0.7145003        NA -1.903810     0 -1.216047 -1.216047
## 3 -0.5470220 -2.103604 -2.585469     1         NA -2.103604
## 4 -0.5008263        NA -2.003854     0 -1.280737 -1.280737
## 5 -0.4592409 -2.100544 -2.134435     1         NA -2.100544
## 6 -0.4576016        NA -2.202306     0 -1.409060 -1.409060

#plot werte nach RK
par(xpd=F)
points(dataValid$x_reg, dataValid$y, col="red", pch=1)

lm_rk_valid<-lm(y~x_reg, data=dataValid)
abline(lm_rk_valid, col="red", lwd=3, lty=2) #RK RegGerade nahe der wahren
    RegGerade
par(xpd=T)
text(2, -3,labels="X_Rk", col="red")
par(xpd=F)


#Schritt 3: Anpassen der Standardfehler
#Bootstrapping (Resampling vectors im Messfehlermodell (Kaptitel 4.1))

#10 Bootstrapstichproben- Resampling vectros Verfahren
B <- 100
i<-1
beta_rk_0_valid <- numeric(B)
beta_rk_1_valid <- numeric(B)
while( i <= B){
  #dataValid besitzt zwei verschiedene Datenstrukturen (Validierungsdaten
  #und Daten ohne wahren X-Werte)
  #Daten muessen vor der Bootstrap-Ziehung getrennt werden nach den
  #Strukturen getrennt werden (siehe Kapitel 4)

  #Bootstrap von Vektoren aus Validierungsdaten
  beob1 <- rownames(dataValid[dataValid$Valid==1, ])
  #Ziehen Vektoren mit zuruecklegen aus Validierungsdaten
```

```
  which1 <- sample(beob1, length(beob1), replace=TRUE)
  bootValid1 <- dataValid[dataValid$Valid==1, ][which1, ]

  #Bootstrap von Vektoren aus Daten ohne wahren X-Werte
  beob0 <- rownames(dataValid[dataValid$Valid==0, ])
  #Ziehen Vektoren mit zuruecklegen aus Validierungsdaten
  which0 <- sample(beob0, length(beob0), replace=TRUE)
  bootValid0 <- dataValid[dataValid$Valid==0, ][which0, ]

  #Bootstrap-Daten
  dataBoot <- merge(bootValid1, bootValid0, all=T)
  head(dataBoot)

  #Regressionskalibrierung
  #Schritt 1:
  #Zeilen mit Na in x werden automatisch geloescht
  lm_xSt <- lm(x~xSt, data=dataBoot)

  #Anwendung der Schaetzung auf diejenigen X*, fuer die die wahren
  #X nicht vorliegen
  x_hat_boot <- coef(lm_xSt)[1]+coef(lm_xSt)[2]*
    (dataBoot[(dataBoot$Valid==0), ])$xSt

  ##x_reg_boot enthaelt die wahren X-Werte und wo die wahren X-Werte
  #fehlen enthaelt sie die geschaetzten X-Werte
  dataBoot$x_reg_boot<-dataBoot[ , "x"]
  dataBoot[dataBoot$Valid==0, ]$x_reg_boot <- x_hat_boot

  #Schritt 2: wahre X durch die Schaetzung aus Regression in Schritt 1
  #ersetzen
  lm_rk<-lm(y~x_reg_boot, data=dataBoot)

  #speichern der Paramter
  beta_rk_0_valid[i] <- coef(lm_rk)[1]
  beta_rk_1_valid[i] <- coef(lm_rk)[2]
  i<-i+1
}

#Parameterschaetzer aus den Bootstraps
beta_rk_0_valid
beta_rk_1_valid

#Varianz &Standardabweichung
var(beta_rk_0_valid)
sd(beta_rk_0_valid)

var(beta_rk_1_valid)
sd(beta_rk_1_valid)

summary(lm_rk_valid) #aus Schritt 2 zum vergleich
```

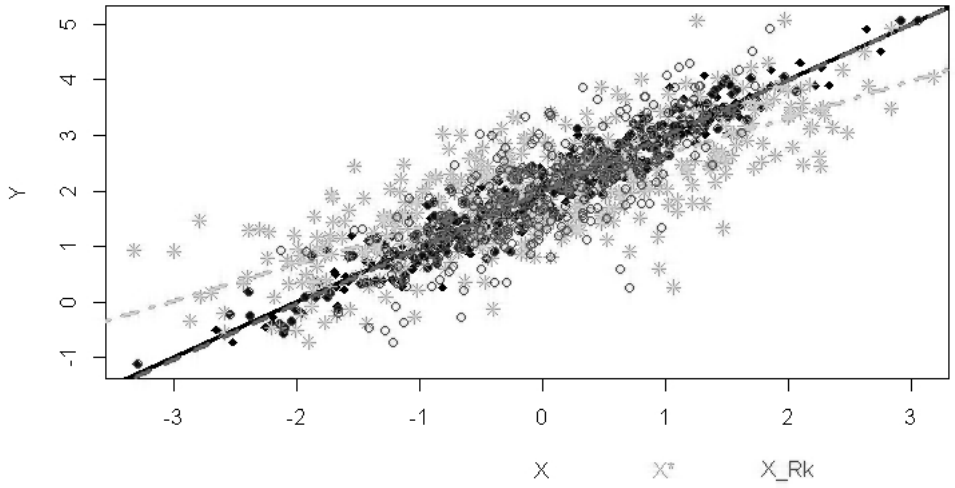

Abbildung 5: *Schwarze durchgezogene Linie* ist die wahre Regressionsgerade. *Grün gepunktet und gestrichelte Linie* entspricht der Regressionsgerade aus der naiven Schätzung. Die *rot gestrichelte Linie* stellt die Schätzung nach Anwendung der Regressionskalibrierung auf Validierungsdaten dar.

E.5. RK mit Wiederholungsdaten (k=1)

In Kapitel 5 wurden die Formeln für den Fall einer Messwiederholung aufbereitet. Im folgenden werden die Formel in R angewendet.

regcalWdh1.R

```
1  ##############Regressionskalibrierung mit Wiederholungsdaten ki=1#########
2  ####Regressionskalibrierung Vefahren (Kapitel 3.4/5)
3
4  #Werte wie oben
5
6  #Annahme: Es liegen Daten mit einer Messwiederholung X* (k=1)
7  dataWdh1 <- dataOrigin[c(1,3)]
8  head(dataWdh1)
9
10 #     y          xSt
11 # 1 1.177596 -0.5737402
12 # 2 2.692329 -0.6047632
13 # 3 2.012540  2.7118439
14 # 4 1.412986 -0.3814698
15 # 5 2.542520  0.8819578
16 # 6 0.215304 -2.4429059
17
18 #Plot wahre Werte
19 par(xpd=F)
20 plot(x, y, col="black", pch=18, xlab="")
21 abline(lm(y~x), col="black",lwd=3)
22 par(xpd=T)
23 text(0,-3, labels="X", col="black")
24
25 #Plot Werte mit Fehler 1.Messung
26 par(xpd=F)
27 points(dataWdh1$xSt,dataWdh1$y, col="green", pch=8)
28 abline(lm(y~xSt, data=dataWdh1), col="green", lwd=3, lty=4)
29 par(xpd=T)
30 text(1,-3, labels="X1*", col="green")
31
32 #Plot Werte nach RK
33
34 ##Schritt 1
35 #Formel von EW von X bedingt auf X* (Kapitel 5 Formel (16))
36
37 #Kapitel 3.4. besagt, dass die Varianz von u aus externen Daten verwendet werden
       kann
38 #extern geschaetzt (aus Validierungsdaten)
39 varU_hat <- var(dataValid[dataValid$Valid==1,]$xSt-dataValid[dataValid$Valid==1,]$
       x)
40
41 #Restliche Groessen aus den vorliegenden Daten schaetzen
42 n<-dim(dataWdh1)[1]
43 muXStern_hat <- sum(dataWdh1$xSt)/n #(Formel 17)
44 varX_hat<-sum((dataWdh1$xSt-muXStern_hat)^2)/(n-1)-varU_hat #(Formel 18)
45 varXStern_hat <- varX_hat+varU_hat #v(x*)=v(x)+v(u)
46
47 #gefittete X
48 dataWdh1$ewX_Xst1 <- (varX_hat/varXStern_hat)*dataWdh1$xSt+
49   muXStern_hat*(1-(varX_hat/varXStern_hat)) #(Formel 16)
50 points(dataWdh1$ewX_Xst1, dataWdh1$y,col="red", pch=1)
51
52 ##Schritt 2
```

```
53 par(xpd=F)
54 lm_rk_wdh1 <- lm(y~ewX_Xst1, data=dataWdh1)
55 abline(lm_rk_wdh1, col="red", lwd=3, lty=2)
56 par(xpd=T)
57 text(2,-3, labels="X_Rk*", col="red")
58
59 ##Schritt 3
60 #300 Bootstrapstichproben- Resampling vectors Verfahren
61 i <- 1
62 B <- 300
63 beta_rk_0_wdh1 <- numeric(B)
64 beta_rk_1_wdh1 <- numeric(B)
65 dataWdh1$x_hat_boot<-c()
66 for( i in 1:B){
67
68   beob <- rownames(dataWdh1)
69   #Ziehen Vektoren mit zuruecklegen aus Wdh-Daten
70   which <- sample(beob, length(beob), replace=TRUE)
71   bootWdh1 <- dataWdh1[which, ]
72
73   ##Schritt 1
74   #Formel von EW von X bedingt auf X* (Kapitel 5 Formel (2))
75
76   #Kapitel 3.4. besagt, dass die Varianz von u aus externen Daten verwendet
77   #werden kann extern geschaetzt (aus Validierungsdaten aus Kapitel E.4.)
78   varU_hat <- var(dataValid[dataValid$Valid==1,]$xSt-dataValid[dataValid$Valid
         ==1,]$x)
79
80   #Restliche Groessen aus den vorliegenden Daten schaetzen
81   n<-dim(bootWdh1)[1]
82   muXStern_hat <- sum(bootWdh1$xSt)/n #(Formel 17)
83   varX_hat<-sum((bootWdh1$xSt-muXStern_hat)^2)/(n-1)-varU_hat #(Formel 18)
84   varXStern_hat <- varX_hat+varU_hat #v(x*)=v(x)+v(u)
85
86   #gefittete X
87   bootWdh1$ewX_Xst1 <- (varX_hat/varXStern_hat)*bootWdh1$xSt+
88     muXStern_hat*(1-(varX_hat/varXStern_hat)) #(Formel 2)
89
90   ##Schritt 2
91   lm_rk_wdh1_boot <- lm(y~ewX_Xst1, data=bootWdh1)
92   beta_rk_0_wdh1[i] <- coef(lm_rk_wdh1_boot)[1]
93   beta_rk_1_wdh1[i] <- coef(lm_rk_wdh1_boot)[2]
94   i<-i+1
95 }
96
97 #Parameterschaetzer aus den Bootstraps
98 beta_rk_0_wdh1
99 beta_rk_1_wdh1
100
101 #Varianz &Standardabweichung
102 var(beta_rk_0_wdh1)
103 sd(beta_rk_0_wdh1)
104
105 var(beta_rk_1_wdh1)
106 sd(beta_rk_1_wdh1)
```

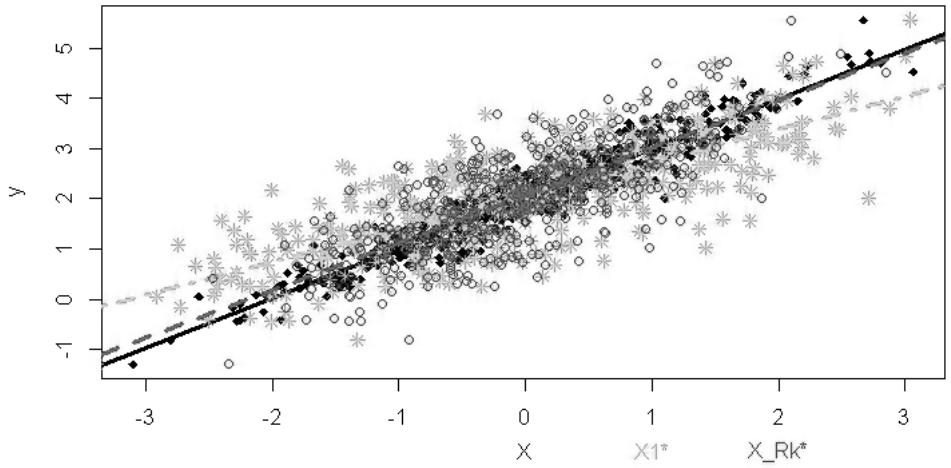

Abbildung 6: *Schwarze durchgezogene Linie* ist die wahre Regressionsgerade. *Grün gepunktet und gestrichelte Linie* entspricht der Regressionsgerade aus der naiven Schätzung. Die *rot gestrichelte Linie* stellt die Schätzung nach Anwendung der Regressionskalibrierung auf Wiederholungsdaten mit einer Messwiederholung dar.

E.6. RK mit Wiederholungsdaten (k=4)

In Kapitel 5 wurde die Regressionskalibrierung an einem theoretischem Beispiel anhand von Daten mit einer Messwiederholung dargestellt. In diesem Anhang wird die Regressionskalibrierung auf Daten mit vier Messwiederholungen angewendet.
Die Formeln aus Kapitel 3.4 vereinfachen sich zu:

$$\overline{X}^*_{i\cdot} = \frac{X^*_{i1} + X^*_{i2} + X^*_{i3} + X^*_{i4}}{4} \tag{37}$$

$$\widehat{\mu}_x = \widehat{\mu}_{x^*} = \frac{\sum_{i=1}^{n} \overline{\boldsymbol{X}^*_{i\cdot}}}{n} \tag{38}$$

$$\widehat{\Sigma}_{uu} = \widehat{\sigma}^2_u = \frac{\sum_{i=1}^{n}\sum_{j=1}^{4}\left(\boldsymbol{X}^*_{ij} - \overline{\boldsymbol{X}^*_{i\cdot}}\right)\left(\boldsymbol{X}^*_{ij} - \overline{\boldsymbol{X}^*_{i\cdot}}\right)^t}{3n} \tag{39}$$

$$\widehat{\Sigma}_{xx} = \widehat{\sigma}^2_x = \frac{\sum_{i=1}^{n}(\overline{\boldsymbol{X}^*_{i\cdot}} - \widehat{\mu}_{x^*})(\overline{\boldsymbol{X}^*_{i\cdot}} - \widehat{\mu}_{x^*})^t}{(n-1)} - \frac{\widehat{\Sigma}_{uu}}{4} \tag{40}$$

$$\widehat{E[\boldsymbol{X}|\overline{\boldsymbol{X}}^*]} \approx \frac{4\widehat{\sigma}^2_x}{4\widehat{\sigma}^2_x + \widehat{\sigma}^2_u}\overline{\boldsymbol{X}}^* + \hat{\mu}_{x^*}(1 - \frac{4\widehat{\sigma}^2_x}{4\widehat{\sigma}^2_x + \widehat{\sigma}^2_u}) = \widehat{\boldsymbol{X}} \tag{41}$$

regcalWdh4.R

```
1  ##############Regressionskalibrierung mit Wiederholungsdaten ki=4#########
2  ####Regressionskalibrierung Vefahren (Kapitel 3.4/5)
3
4  #Werte wie oben
5
6  #Annahme es liegen Daten vier Messwiederholungen X1*, x2*, x3*, x4* (k=4)
7  dataWdh4<-dataOrigin[c(1,3,4,5,6)]
8  head(dataWdh4)
9
10 ##        y         xSt        xSt2        xSt3        xSt4
11 ## 1 1.177596 -0.5737402 -0.9237768 -0.3431826 -1.9795782
12 ## 2 2.692329 -0.6047632  0.5024174  1.1258265  0.9936227
13 ## 3 2.012540  2.7118439  0.9781640  1.9664301  1.1731840
14 ## 4 1.412986 -0.3814698 -0.1526708  0.1482332  0.1532670
15 ## 5 2.542520  0.8819578  0.9571847  1.9208449 -0.3566576
16 ## 6 0.215304 -2.4429059 -1.5093773 -2.1616485 -1.4724310
17
18 #Plot wahre Werte
19 par(xpd=F)
20 plot(x, y, col="black", pch=18, xlab="")
21 abline(lm(y~x), col="black",lwd=3)
22 par(xpd=T)
23 text(-2,-3, labels="X", col="black")
24
25 #Plot Werte mit Fehler 1.Messung
26 par(xpd=F)
27 points(dataWdh4$xSt,dataWdh4$y, col="green", pch=8)
28 abline(lm(y~xSt, data=dataWdh4), col="green", lwd=3, lty=4)
29 par(xpd=T)
30 text(-1,-3, labels="X1*", col="green")
31
32 #Plot Werte mit Fehler 2.Messung
```

```
par(xpd=F)
points(dataWdh4$xSt2,dataWdh4$y, col="blue", pch=4)
abline(lm(y~xSt2, data=dataWdh4), col="blue", lwd=3, lty=3)
par(xpd=T)
text(0,-3, labels="X2*", col="blue")

#Plot Werte mit Fehler 3.Messung
par(xpd=F)
points(dataWdh4$xSt3,dataWdh4$y, col="darkorchid3", pch=6)
abline(lm(y~xSt3, data=dataWdh4), col="darkorchid3", lwd=3, lty=4)
par(xpd=T)
text(1,-3, labels="X3*", col="darkorchid3")

#Plot Werte mit Fehler 4.Messung
par(xpd=F)
points(dataWdh4$xSt4,dataWdh4$y, col="orange", pch=3)
abline(lm(y~xSt4, data=dataWdh4), col="orange", lwd=3, lty=3)
par(xpd=T)
text(2,-3, labels="X4*", col="orange")


#Plot Werte nach RK

##Schritt 1
#Formel von EW von X bedingt auf X* (Kapitel 3.4 Formel (2))
#Kapitel 3.4. stelle eine Formel bereit, wie man die Varianz von X
#berechnen kann, wenn X nicht vorliegt
n<-dim(dataWdh4)[1]
xStern_quer <- (dataWdh4$xSt+dataWdh4$xSt2+dataWdh4$xSt3+dataWdh4$xSt4)/4 #(Formel
     37)
varU_hat<-sum((dataWdh4$xSt-xStern_quer)^2+(dataWdh4$xSt2-xStern_quer)^2+
                (dataWdh4$xSt3-xStern_quer)^2+(dataWdh4$xSt4-xStern_quer)^2)/(3*n)
                     #(Formel 39)

#Restliche Groessen aus Daten schaetzen
muXStern_hat <- sum(xStern_quer)/n #(Formel 38)
varX_hat<-sum((xStern_quer-muXStern_hat)^2)/(n-1)-varU_hat/4 #(Formel 40)
varXStern_hat <- varX_hat+varU_hat #v(x*)=v(x)+v(u)

#gefittete X
dataWdh4$ewX_Xst4 <- ((4*varX_hat)/(3*varX_hat+varXStern_hat))*xStern_quer+
  muXStern_hat*(1-((4*varX_hat)/(3*varX_hat+varXStern_hat))) #(Formel 41)
points(dataWdh4$ewX_Xst4, dataWdh4$y,col="red", pch=1)

##Schritt 2
par(xpd=F)
lm_rk_wdh4<-lm(y~ewX_Xst4, data=dataWdh4)
abline(lm_rk_wdh4, col="red", lwd=3, lty=2)
par(xpd=T)
text(3,-3, labels="X_Rk*", col="red")

##Schritt 3
#300 Bootstrapstichproben- Resampling vectros Verfahren
B <- 300
beta_rk_0_wdh4 <- numeric(B)
beta_rk_1_wdh4 <- numeric(B)
dataWdh4$x_hat_boot<-c()
for( i in 1:B){

  beob <- rownames(dataWdh4)
  #Ziehen Vektoren mit zuruecklegen aus Wdh-Daten
  which <- sample(beob, length(beob), replace=TRUE)
```

```
  bootWdh4 <- dataWdh4[which, ]

  ##Schritt 1
  #Formel von EW von X bedingt auf X* (Kapitel 3.4 Formel (2))

  #Kapitel 3.4. stelle eine Formel bereit, wie man die Varianz von X
  #berechnen kann, wenn X nicht vorliegt
  n<-dim(bootWdh4)[1]
  xStern_quer <- (bootWdh4$xSt+bootWdh4$xSt2+bootWdh4$xSt3+bootWdh4$xSt4)/4
  varU_hat<-sum((bootWdh4$xSt-xStern_quer)^2+(bootWdh4$xSt2-xStern_quer)^2+
                  (bootWdh4$xSt3-xStern_quer)^2+
                  (bootWdh4$xSt4-xStern_quer)^2)/(3*n) #(Formel 39)

  #Restliche Groessen aus Daten schaetzen
  muXStern_hat <- sum(xStern_quer)/n #(Formel 48 )
  varX_hat<-sum((xStern_quer-muXStern_hat)^2)/(n-1)-varU_hat/4 #(Formel 40 )
  varXStern_hat <- varX_hat+varU_hat #v(x*)=v(x)+v(u)

  #gefittete X
  bootWdh4$ewX_Xst4 <- ((4*varX_hat)/(3*varX_hat+varXStern_hat))*xStern_quer+
    muXStern_hat*(1-((4*varX_hat)/(3*varX_hat+varXStern_hat))) #(Formel 41)

  ##Schritt 2
  lm_rk_wdh4_boot<-lm(y~ewX_Xst4, data=bootWdh4)

  beta_rk_0_wdh4[i] <- coef(lm_rk_wdh4_boot)[1]
  beta_rk_1_wdh4[i] <- coef(lm_rk_wdh4_boot)[2]
}

#Parameterschaetzer aus den Bootstraps
beta_rk_0_wdh4
beta_rk_1_wdh4

#Varianz &Standardabweichung
var(beta_rk_0_wdh4)
sd(beta_rk_0_wdh4)

var(beta_rk_1_wdh4)
sd(beta_rk_1_wdh4)
```

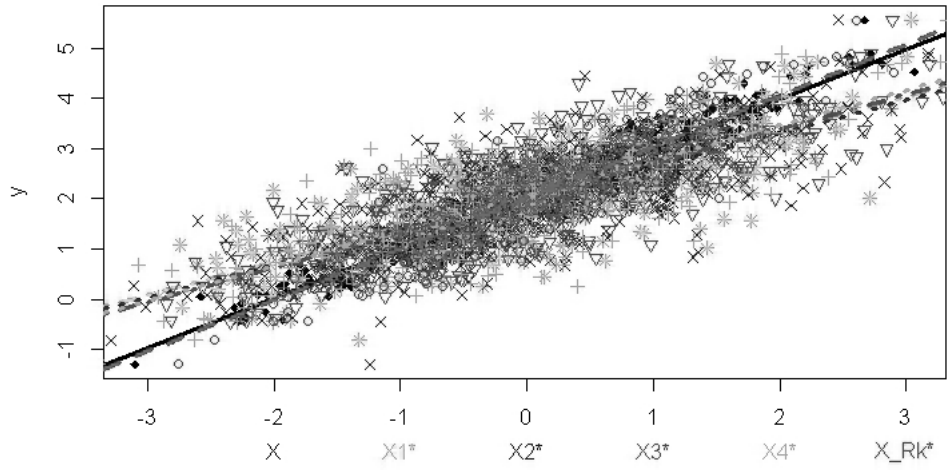

Abbildung 7: *Schwarze durchgezogene Linie* ist die wahre Regressionsgerade. Die *Grün,blau,lila und orange gepunktet und gestrichelte Linien* entsprechen der Regressionsgerade aus der naiven Schätzung. Die *rot gestrichelte Linie* stellt die Schätzung nach Anwendung der Regressionskalibrierung auf Wiederholungsdaten mit einer Messwiederholung dar.

E.7. Vergleich

In diesem Unterkapitel wird die Anwendung der Regressionskalibrierung in E.4, E.5, E.6 grafisch und anhand von Kennzahlen gegenübergestellt.

regcalVergleich.R

```
##########Vergleich der drei Vorgehen###############
#Wahre Warte
#plot der wahren Werte (aus dataOrigin)
par(xpd=F)
plot(dataOrigin$x, dataOrigin$y, col="black", pch=18, ylab="Y", xlab="")
abline(lm(y~x, dataOrigin), col="black", lwd=3)
par(xpd=T)
text(-1.5,-3, labels="X", col="black")

#Verfahren mit Validierungsdaten
par(xpd=F)
points(dataValid$x_reg, dataValid$y, col="red", pch=8)
#lm_rk_valid<-lm(y~x_reg, data=dataValid)
abline(lm_rk_valid, col="red", lwd=3,lty=3)
par(xpd=T)
text(-0.5,-3, labels="X_Rk_valid", col="red")

#Verfahren mit einer Messwdh
par(xpd=F)
points(dataWdh1$ewX_Xst1, dataWdh1$y,col="blue", pch=1)
#lm_rk_wdh1<-lm(dataWdh1$y~dataWdh1$ewX_Xst)
abline(lm_rk_wdh1, col="blue", lwd=3, lty=2)
par(xpd=T)
text(1,-3, labels="X_Rk_wdh1", col="blue")

#Verfahren mit vier Messwdh
par(xpd=F)
points(dataWdh4$ewX_Xst, dataWdh4$y, col="green")
#lm_rk_wdh4<-lm(dataWdh4$y~dataWdh4$ewX_Xst)
abline(lm_rk_wdh4, col="green", pch=8, lwd=4,lty=2)
par(xpd=T)
text(2.5,-3, labels="X_Rk_wdh4", col="green")

#Tabelle
mean_b0_valid<-mean(beta_rk_0_valid)
mean_b1_valid<-mean(beta_rk_1_valid)

mean_b0_wdh1<-mean(beta_rk_0_wdh1)
mean_b1_wdh1<-mean(beta_rk_1_wdh1)

mean_b0_wdh4<-mean(beta_rk_0_wdh4)
mean_b1_wdh4<-mean(beta_rk_1_wdh4)

var_b0_valid<-var(beta_rk_0_valid)
var_b1_valid<-var(beta_rk_1_valid)

var_b0_wdh1<-var(beta_rk_0_wdh1)
var_b1_wdh1<-var(beta_rk_1_wdh1)

var_b0_wdh4<-var(beta_rk_0_wdh4)
var_b1_wdh4<-var(beta_rk_1_wdh4)

bias_b0_valid<-mean(beta_rk_0_valid-intercept)
bias_b1_valid<-mean(beta_rk_1_valid-beta)
```

```
55
56 bias_b0_wdh1<-mean(beta_rk_0_wdh1-beta)
57 bias_b1_wdh1<-mean(beta_rk_1_wdh1-beta)
58
59 bias_b0_wdh4<-mean(beta_rk_0_wdh4-beta)
60 bias_b1_wdh4<-mean(beta_rk_1_wdh4-beta)
61
62 MSE_b0_valid<-bias_b0_valid^2+var_b0_valid
63 MSE_b1_valid<-bias_b1_valid^2+var_b1_valid
64
65 MSE_b0_wdh1<-bias_b0_wdh1^2+var_b0_wdh1
66 MSE_b1_wdh1<-bias_b1_wdh1^2+var_b1_wdh1
67
68 MSE_b0_wdh4<-bias_b0_wdh4^2+var_b0_wdh4
69 MSE_b1_wdh4<-bias_b1_wdh4^2+var_b1_wdh4
70
71 summary(lm_rk_valid)
72 summary(lm_rk_wdh1)
73 summary(lm_rk_wdh4)
74 valid<-c(var_b0_valid,var_b1_valid,MSE_b0_valid,MSE_b1_valid)
75 wdh1<-c(var_b0_wdh1,var_b1_wdh1,MSE_b0_wdh1,MSE_b1_wdh1)
76 wdh4<-c(var_b0_wdh4,var_b1_wdh4,MSE_b0_wdh4,MSE_b1_wdh4)
77
78 tab<-cbind(valid,wdh1,wdh4)
79 rownames(tab)<-c("hat_v(b0)","hat_v(b1)","MSE(b0)","MSE(b1)")
80 tab
81
82 #               valid          wdh1          wdh4
83 # hat_v(b0) 0.0029921079 0.0009294705 0.0003816482
84 # hat_v(b1) 0.0009590951 0.0026153208 0.0005811254
85 # MSE(b0)   0.0096390930 0.9549285336 0.9598549046
86 # MSE(b1)   0.0012459516 0.0027702883 0.0017095357
```

Wenn man den MSE der β_1 vergleicht (Codezeile 82-86), zeigt sich, dass die Schätzung anhand von vier Messwiederholungen zu besseren Schätzern führt als wenn man die Schätzung mit einer Messwiederholung vornimmt. In diesem Beispiel führen insgesamt die Schätzung an Validierungsdaten zu besseren Schätzern. Allerdings kann diese Aussage nicht verallgemeinert werden. Auch in der Praxis bestehen Diskussionen darüber ob es vorteilhafter wäre in Messwiederholungen zu investieren oder darin für einen Teil der Daten wahre Werten zu messen.

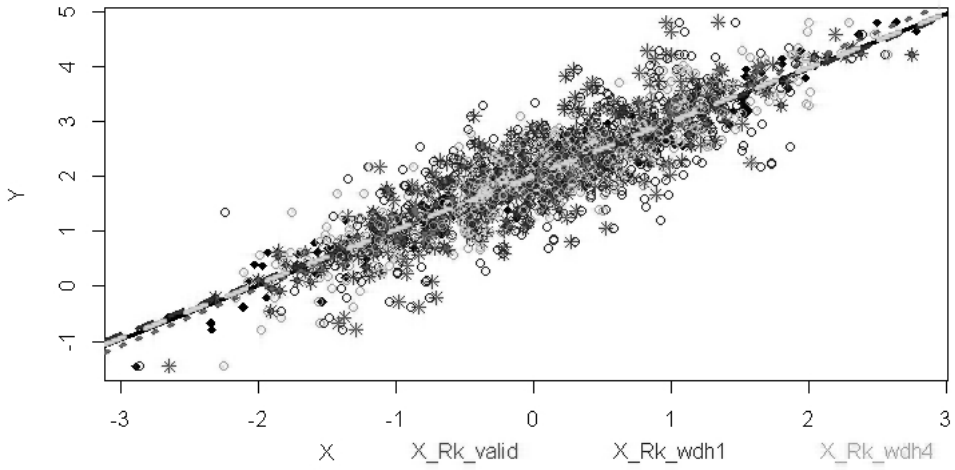

Abbildung 8: *Schwarze durchgezogene Linie* ist die wahre Regressionsgerade. Die restlichen Geraden entstammen aus der Schätzung nach Anwendung der Regressionskalibrierung. Die *rot gepunktete Linie* entspricht der Schätzung auf Validierungsdaten. Die *blau gestrichelte und gepunktete Linie* stellt die Schätzung auf Wiederholungsdaten mit einer Messwiederholung dar und die *grün gestrichelte Linie* auf Wiederholungsdaten mit vier Messwiederholungen (vgl. Rcode Anhang E.4.

F. Simex

Zur Messfehlerkorrektur kann nicht nur die Regressionskalibrierung angewendet werden, sondern Beispielsweise auch die Simulation Extrapolation Verfahren (SIMEX).
Am Beispiel einer einfachen linearen Regression $\boldsymbol{Y_i = \beta_0 + \beta_1 X_i + \epsilon_i}$, wobei $\boldsymbol{\epsilon_i}$ Normalverteilt ist, soll die Idee von Simex erläutert werden. Allerdings kann das Verfahren auf viele weitere Modelle angewendet werden. Weiterhin soll angenommen werden, dass ein additiver Messfehler vorliegt $\boldsymbol{X^* = X + U}$. Wie bereits in Abbildung 1 zu sehen ist, wird der Parameter $\boldsymbol{\beta_1}$ systematisch unterschätzt.
Bei SIMEX wird zu den bereits fehlerhaft gemessenen unabhängige Variable $\boldsymbol{X^*}$ einen weiteren Messfehler hinzugefügt, sodass die neue Standardabweichung der neu erzeugten Daten $X^*(\lambda_k)$ sich wie folgt zusammensetzt: $\sigma^2_{X^*}(\lambda_k) = \sigma^2_{X^*}(1 + \lambda_k)$, wobei $\sigma^2_{X^*}$ die Messfehlerstandardabweichung der ursprünglichen Daten ist (in Hölzl [2015] finden sich weitere Informationen zur Bestimmung dieser Kennzahl) und λ_k beliebig gewählt werden kann (vgl. Hölzl [2015]). Für $\lambda_1 = 0,\ \lambda_2 = 1,\ \lambda_3 = 2$ ergibt sich folgende Regressionsgeraden in Abbildung 9

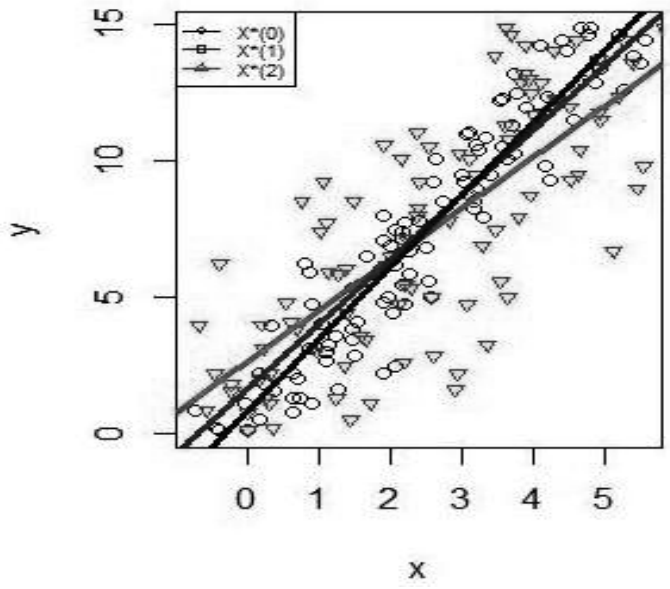

Abbildung 9: Die *schwarze Gerade* ist die geschätzte Regressionsgerade der ursprünglichen Daten $\boldsymbol{X^*(0)}$ mit einer Standardabweichung von $\sigma^2_{X^*}$, die *rote Gerade* die der Daten $\boldsymbol{X^*(1)}$ mit einer Standardabweichung von $2\sigma^2_{X^*}$ und die *blaue Gerade* die der Daten $\boldsymbol{X^*(2)}$ mit einer Standardabweichung von $3\sigma^2_{X^*}$. (Aus Hölzl [2015])

Die zugehörigen β_1- Schätzer sind in Abbildung 10 dargestellt, wobei für die Messfehlerstandardabweichung der erzeugten Daten gilt:

- $X^*(0) : \sigma^2_{x^*}(\lambda_1) = \sigma^2_{x^*}$ (ursprüngliche Daten mit Messfehler)

- $X^*(1) : \sigma^2_{x^*}(\lambda_2) = 2\sigma^2_{x^*}$ (neu erzeugten Daten mit zusätzlichen Messfehler $\sigma^2_{x^*}$)
- $X^*(2) : \sigma^2_{x^*}(\lambda_3) = 3\sigma^2_{x^*}$ (neu erzeugten Daten mit zusätzlichen Messfehler $2\sigma^2_{x^*}$)

Aus diesen Daten werden β-Schätzer geschätzt, die β_1-Schätzer sind in Abbildung 10 abgebildet. Es ist nun möglich ein lineares Modell für die Parameter zu fitten (siehe Gerade in Abbildung 10). Mithilfe des Modells kann der β_1-Schätzer für den Fall $\lambda_k = -1$ geschätzt werden. Für $\lambda_k = -1$ gilt nämlich $\sigma^2_{X^*}(\lambda_k) = \sigma^2_{X^*}(1-1) = 0$. Inhaltlich bedeutet es, dass kein Messfehlerstandardabweichung vorliegt. Somit ist der resultierende β_1-Schätzer eine Schätzung für den wahren Einfluss von X_i auf Y_i.

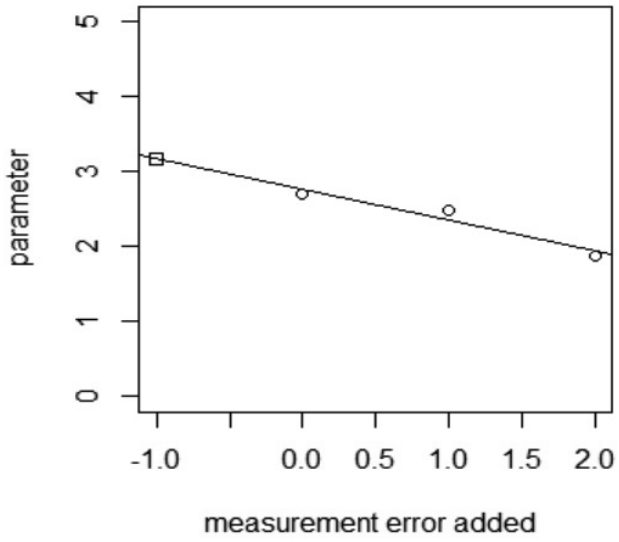

Abbildung 10: Geschätzte Parameter für *measurement error added* bzw. λ_k= -1, 0, 1, 2. (Aus Hölzl [2015])

Detailiertere Beschreibungen können in der Arbeit von Hölzl [2015] oder Caroll et al. [2006] (Kapitel 5) nachgeschlagen werden. Wie in der Regressionskalibrierung lässt sich auch für Simex ein allgemeiner Algorithmus aufstellen (vgl. Hölzl [2015] und Caroll et al. [2006] (Kapitel 5), S.100):

- **Simulationsschritt**
  - 1. Simuliere Pseudo-Daten $X^*_{b,i}(\lambda_k) = X^*_i + \sqrt{\lambda_k \sigma^2} U_{b,i}$, wobei $U_{b,i}$ aus einer angenommener Messfehlerverteilung gezogen wird mit i=1,..,n und b=1,...,B und k $\in$ N.
  - 2. Wiederhole Schritt 1 B mal, man erhält somit die Datenmengen $X^*_{1,i}(\lambda_k)$,...,$X^*_{B,i}(\lambda_k)$
  - Berechne die Mittelwertschätzer: $\hat{\beta}(\lambda_k) = \frac{1}{B} \sum_{b=0}^{B} \hat{\beta}_{naive}(Y, X_b(\lambda_k))$

- **Exploration**
- 1. Finde ein Modell, dass den Zusammenhang zwischen $(\lambda_k, \hat{\beta}_k(\lambda_k))$ modelliert
- 2. Treffe anhand des Modells eine vorhersage für $\hat{\beta}(-1)$

SIMEX ist, im Gegensatz zur Regressionskalibrierung, nur anwendbar, wenn der Messfehler $\sigma^2_{X^*}$ bekannt ist, andernfalls kann keine Messfehlerbehebung durchgeführt werden. Sowohl SIMEX als auch die Regressionskalibierung können auf komplexere Modelle angewendet werden, wobei Letztere auch dann funktioniert, wenn SIMEX scheitert, beispielsweise in einem COX-Modell. Außerdem kommt hinzu, dass Simex rechenaufwändiger als die Regressionskalibrierung ist. Wenn das untersuchte Modell allerdings für SIMEX geeignet ist, kann SIMEX unter Umständen bessere Parameter-Schätzer liefern als die Regressionskalibrierung.

Literatur

Helmholz zentrum münchen, kora - kooperative gesundheitsforschung in der region augsburg, accessed:19.10.2014. http://www.helmholtz-muenchen.de/kora/ueber-kora/historisches-zu-kora-und-monica/index.html.

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.

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

P. Gustafson. *Measurement Error and Missclassifictaion in Statistics and Epidemiology-Impacts and Bayesian Adjustments*. 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.

A. Hölzl. Seminararbeit, 2015. Simulation Extrapolation.

N. Markovic. Seminararbeit, 2015. Fehler in der abhängigen Variable.

H. Marshalava. Seminararbeit, 2015. Überblick über Messfehler und ihre Auswirkungen in der linearen Regression.

A. Pokatilo. Seminararbeit, 2015. Übersicht zu fehlenden Daten.

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