



LUDWIG-MAXIMILIANS UNIVERSITY

SEMINARARBEIT

Regressionskalibrierung

Author:
Minh-Anh LE

Supervisor:
Prof. Dr. Thomas AUGUSTIN

28. November 2014

Inhaltsverzeichnis

1. Einleitung	1
2. Die zentrale Idee und der Algorithmus der Regressionskalibrierung	2
2.1. Zentrale Idee	2
2.2. Algorithmus nach Carroll	3
3. Parameterschätzung in der Kalibrierungsfunktion	3
3.1. Überblick	3
3.2. Interne Validierungsdaten	4
3.3. Instrumentaldaten	4
3.4. Wiederholungsdaten	4
3.4.1. Überprüfung durch partielle Wiederholungsdaten	6
4. Bootstrapping in der Regressionskalibrierung	6
4.1. Resampling Vectors im Messfehlermodell	7
4.2. Resampling Residuals im Messfehlermodell	7
4.3. Algorithmus des Bootstrappings in der Regressionskalibrierung	8
5. Theoretisches Beispiel an einer einfachen linearen Regression	9
6. Beispiel-The Monica Study	11
6.1. Die Daten	11
6.2. Ergebnisse	11
7. Fazit	14
Anhang	15
A. Nicht-differentieller Fehler	15
B. Berkson- Fehler	16
C. Beispiel: Regressionskalibrierung Rcode	17
C.1. Daten generieren	17
C.2. Vergleich RK mit/ohne Dummyvariable in Validierungsdaten	18
C.3. Vergleich RK mit/ohne Dummyvariable in Validierungsdaten-Plot	21
C.4. RK mit Validierungsdaten	24
C.5. RK mit Wiederholungsdaten ($k=1$)	28
C.6. RK mit Wiederholungsdaten ($k=4$)	31
C.7. Vergleich	35

1. Einleitung

Der Umgang mit Messfehlern in stetigen Einflussvariablen 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. 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 Fehlermodell anzupassen, ist die Regressionskalibrierung.

Grundlage der Regressionskalibrierung ist es, im Hauptmodell, in dem die Zielgröße \mathbf{Y} auf die Einflussgrößen (\mathbf{X}, \mathbf{Z}) 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 Standardanalyse durchgeführt werden.

Bei den verschiedenen Methoden zum Ausgleich von Messfehlern, unterscheidet man nach Carroll et al. [2006] (Kapitel 2.1) grundsätzlich zwischen *funktionaler 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), die korrigierte Scorefunktion und, auch wenn eine Zuordnung nicht eindeutig ist, die Regressionskalibrierung. Das bedeutet, dass die Verteilung von \mathbf{X} bedingt auf die restlichen Kovariablen unbedingt nötig ist, um die Regressionskalibrierung anzuwenden. Weiterhin unterscheidet man 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{X}, \mathbf{Z}) 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. Für die Anwendung der Regressionskalibrierung ist das Vorliegen von *nicht differentiellen* Fehlern erforderlich. Näheres in (Anhang A). 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 [Carroll 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* $\mathbf{X}_i^* = \mathbf{X}_i + \mathbf{U}_i$ vorliegt, d.h. wenn fehlerhafte Werte \mathbf{X}_i^* vorliegen, die aus dem unbekanntem Wert der wahren Variable \mathbf{X}_i und dem additiven Fehler \mathbf{U}_i bestehen, dann

kann diese durch die Regressionskalibrierung in einen Berkson- Fehler $\mathbf{X}_i = \mathbf{X}_i^* + \mathbf{U}_i$ überführt werden. Das bedeutet, dass das wahre \mathbf{X}_i bedingt auf \mathbf{X}_i^* und \mathbf{U}_i dargestellt werden kann, wobei für \mathbf{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.

Zweck dieser Arbeit ist es, einen Überblick über die Methode der Regressionskalibrierung für additive Messfehler 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 einer Simulation in R (Anhang C) 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. Angenommen die tatsächliche Einflussgröße \mathbf{X} wurde nicht erhoben oder sie ist nicht beobachtbar, sondern nur fehlerbehaftet messbar als \mathbf{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{X}, \mathbf{Z}) zu rechnen, gestaltet sich im vorliegenden Fall als schwierig. Da aber \mathbf{X}^* , nach [Carroll et al., 2006] ein Surrogat zur Verfügung steht, die als Ersatz für \mathbf{X} dient, so kann die Regressionskalibrierung angewendet werden. Wendet man die Regressionskalibrierung nicht an, sondern rechnet eine naive Regression von \mathbf{Y} auf $(\mathbf{X}^*, \mathbf{Z})$ anstelle der eigentlich interessierenden Regression auf (\mathbf{X}, \mathbf{Z}) , so hat dies zur Folge, dass die zugehörigen Schätzungen verzerrt sind. Die Anwendung der Regressionskalibrierung wirkt dem entgegen und ermöglicht näherungsweise unverzerrte Inferenz. Anstatt (\mathbf{X}, \mathbf{Z}) durch $(\mathbf{X}^*, \mathbf{Z})$ zu ersetzen, soll also zunächst eine Regression von \mathbf{X} auf $(\mathbf{X}^*, \mathbf{Z})$ berechnet werden. Man schreibt $\widehat{\mathbf{m}}_x(\mathbf{X}^*, \mathbf{Z}, \hat{\gamma}) = \widehat{\mathbf{X}}$ mit Koeffizientenschätzer $\hat{\gamma}$, als Schätzung von $\mathbf{m}_x(\mathbf{X}^*, \mathbf{Z}, \gamma) = \mathbf{E}(\mathbf{X} | \mathbf{X}^*, \mathbf{Z})$, sodass die erklärenden Variablen für das Hauptmodell nun $\{\widehat{\mathbf{m}}_x(\mathbf{X}^*, \mathbf{Z}, \hat{\gamma}), \mathbf{Z}\}$ sind. Der Regressionskalibrierungsalgorithmus besteht, angelehnt an [Carroll et al., 2006], im Wesentlichen aus drei Schritten, die im folgenden Kapitel dargestellt werden.

2.2. Algorithmus nach Caroll

Beachte, dass man ursprünglich an $E[\mathbf{Y}|\mathbf{Z}, \mathbf{X}] = \mathbf{m}_y(\mathbf{X}, \mathbf{Z}, \beta)$ interessiert ist. Der Algorithmus der Regressionskalibrierung hat folgende Form:

- Schritt 1: Schätze die unbeobachtete Variable \mathbf{X} durch eine Regression von \mathbf{X} auf $(\mathbf{Z}, \mathbf{X}^*)$, also $E[\mathbf{X}|\mathbf{Z}, \mathbf{X}^*] = m_x(\mathbf{Z}, \mathbf{X}^*, \gamma)$. Die Schätzung ist von γ , genauer von der Schätzung $\hat{\gamma}$ 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 $\hat{m}_x(\mathbf{Z}, \mathbf{X}^*, \hat{\gamma})$. Führe anschließend eine Standardanalyse durch, um die Parameterschätzer zu erhalten. Somit erhält man:

$$E[\mathbf{Y}|\mathbf{Z}, \mathbf{X}^*] \approx m_y(\mathbf{Z}, \underbrace{m_x(\mathbf{Z}, \mathbf{X}^*, \hat{\gamma})}_{\hat{\mathbf{X}}}, \beta_{RK}) \quad (1)$$

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

- Schritt 3: Korrigiere mit dem Bootstrap-Verfahren oder der Sandwich-Methode die resultierenden Standardfehler (Kapitel 4).

Trotz der Einfachheit des Verfahrens sollte man aber die Leistungsfähigkeit nicht unterschätzen. 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 [Augustin et al., 2008].

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 $(\mathbf{Z}, \mathbf{X}^*)$ 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, Instrumental-
daten oder Wiederholungsdaten vor. In den ersten beiden Fällen kann das Modell für
 $E[\mathbf{X}|\mathbf{Z},\mathbf{X}^*]$ durch gewöhnliche Regressionsdiagnosen, bspw. Residuenplot, geprüft wer-
den. Wenn zusätzliche partielle Wiederholungsdaten vorliegen, kann das letzte Modell
ebenfalls überprüft werden; Genaueres wird in 3.4.1 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 $(\mathbf{Z},\mathbf{X}^*)$ aus den internen Validierungsdaten durchzuführen. Aller-
dings ist dieser Fall eigentlich ein Missing-Data Problem und man würde normalerweise
eher Missing-Data-Methoden, statt der Regressionskalibrierung anwenden. Aus prakti-
scher 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 Regressionsmethode angewendet
werden. [Carroll et al., 2006] schlägt 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 im Rcode im AnhangC.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, sogenan-
nte Instrumentaldaten. Dabei kann die Regressionskalibrierung nur angewendet werden,
wenn \mathbf{T} unverzerrt für \mathbf{X} ist. 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.

3.4. Wiederholungsdaten

In diesem Abschnitt wird 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 Σ_{uu} haben, d.h. es wird Homoskedastizität angenommen ($\Sigma_i \equiv \Sigma$).
Da in der Literatur der Umgang mit Wiederholungsdaten in der Regressionskalibrierung
hervorgehoben wird, nimmt dieses Unterkapitel einen großen Teil ein. Im Folgenden be-
schreiben wir einen Algorithmus, der eine lineare Approximation der Regressionskalibri-
erungsfunktion liefert. Der Algorithmus ist anwendbar, wenn Σ_{uu} durch externe Daten
oder interne Wiederholungsmessungen geschätzt wurde. Diese Methode wurde von Car-
roll und Stefanski (1990) und von Gleser (1990) entwickelt und wurde unter anderem
von Liu und Liang (1992) und Wang, Carroll 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, $\mathbf{X}_{i1}^*, \dots, \mathbf{X}_{ik_i}^*$ von \mathbf{X}_i und $\overline{\mathbf{X}_i^*}$ 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 $(\mathbf{Z}, \overline{\mathbf{X}^*})$ ist

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

Wiederholungsmessungen ermöglichen die Schätzung der Kovarianzmatrix der Messfehler \mathbf{U}_i

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

Nach [Carroll et al., 2006] gilt: 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 [Carroll et al., 2006]

Weiterhin folgt auf Basis folgender Schätzungen

$$\widehat{\mu}_z = \overline{\mathbf{Z}}; \quad (4)$$

$$\widehat{\mu}_x = \widehat{\mu}_{x^*} = \sum_{i=1}^n k_i \overline{\mathbf{X}_{i\cdot}^*} / \sum_{i=1}^n k_i; \quad (5)$$

$$\nu = \sum_{i=1}^n k_i - \sum_{i=1}^n k_i^2 / \sum_{i=1}^n k_i; \quad (6)$$

$$\widehat{\Sigma}_{zz} = (n-1)^{-1} \sum_{i=1}^n (\mathbf{Z}_i - \overline{\mathbf{Z}})(\mathbf{Z}_i - \overline{\mathbf{Z}})^t; \quad (7)$$

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

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

die Gleichung der geschätzten Regressionskalibrierungsfunktion

$$E[\widehat{\mathbf{X}}_i | \widehat{\mathbf{Z}}_i, \overline{\mathbf{X}_{i\cdot}^*}] \approx \widehat{\mu}_{x^*} + (\widehat{\Sigma}_{xx}, \widehat{\Sigma}_{zx}) \begin{bmatrix} \widehat{\Sigma}_{xx} + \widehat{\Sigma}_{uu}/k_i & \widehat{\Sigma}_{xz} \\ \widehat{\Sigma}_{xz}^t & \widehat{\Sigma}_{zz} \end{bmatrix}^{-1} \begin{pmatrix} \overline{\mathbf{X}_{i\cdot}^*} - \widehat{\mu}_{x^*} \\ \mathbf{Z}_i - \overline{\mathbf{Z}} \end{pmatrix} \quad (10)$$

$$\approx m_{x_i}(\mathbf{Z}_i, X_i, \widehat{\gamma}) = \widehat{\mathbf{X}}_i.$$

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 [Carroll et al., 2006](Kapitel 4.4.2.). 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 C.5 findet sich eine beispielhafte R Anwendung.

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 $E[\mathbf{X}|\mathbf{Z}, \mathbf{X}^*]$ aus Formel (10) zu bestimmen, indem man vorab eine Regression von \mathbf{X}_{i2}^* auf $(\mathbf{Z}_i, \mathbf{X}_{i1}^*)$ berechnet. Es gilt nämlich

$$\mathbf{X}_{i1}^* = \mathbf{X}_i + \mathbf{U}_{i1} \quad (11)$$

$$\mathbf{X}_{i2}^* = \mathbf{X}_i + \mathbf{U}_{i2} \quad (12)$$

$$(13)$$

und somit

$$E[\mathbf{X}_{i2}^*|\mathbf{Z}_i, \mathbf{X}_{i1}^*] = E[\mathbf{X}_i + \widehat{\mathbf{U}}_{i2}|\mathbf{Z}_i, \mathbf{X}_{i1}^*] \quad (14)$$

$$= \underbrace{E[\mathbf{X}_i|\mathbf{Z}_i, \mathbf{X}_{i1}^*]}_{\text{Schritt 1 der RK}} + \underbrace{E[\mathbf{U}_{i2}|\mathbf{Z}_i, \mathbf{X}_{i1}^*]}_{V_i}. \quad (15)$$

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

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 $E[\mathbf{Y}|\mathbf{Z}, \mathbf{X}^*] \approx m_y(\mathbf{Z}, \widehat{\mathbf{X}}, \beta_{RK}^*)$. Für dieses resultierende β_{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 Rstudio, 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 $\widehat{\mathbf{X}}$ basiert.

Die Anpassung der wahren Standardfehler können mit Bootstrapping oder der Sandwich-Methode erreicht werden. 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. 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. 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. 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 Bootstrap-Stichprobe bei der Messfehlerkorrektur berücksichtigt werden, indem man die vorliegenden Daten aufteilt (nach der Strukturart) und anschließend aus jedem Teildatensatz Bootstrap-Stichproben zieht. Im klassischen Bootstrapping ist dieser Schritt nicht nötig, bei der Messfehlerkorrektur schon, da unterschiedliche Datenstrukturen unterschiedliche Mengen an Informationen hatten. 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 Bootstrapp-Ziehungs-Verfahren, wird im Folgenden das Vorliegen von Validierungsdaten angenommen.

4.1. Resampling Vectors im Messfehlermodell

Das *Resampling Vectors*- Verfahren ist eine Erweiterung des *Resampling Pairs*- Verfahrens, in dem nicht mehr paarweise mit Zurücklegen gezogen wird, sondern vektorweise aus $\{(Y_i, X_i, X_i^*, Z_i)\}_{i=1}^n$, mit n als Umfang der Bootstrap-Stichprobe. 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 ϵ_i abhängig von Z_i sein, diese Abhängigkeitsbeziehung wird durch das vektorweise Ziehen berücksichtigt, da die Variablen (Y_i, X_i, X_i^*, 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 (X_i, X_i^*, Z_i) bedingt. Beachte, dass man eigentlich aus repräsentativen Gründen eine Stichprobenverteilung bedingt auf Z_i erreichen möchte.

4.2. Resampling Residuals im Messfehlermodell

Anders als beim *Resampling Vectors* liefert *Resampling Residuals* eine Stichprobenverteilung, die auf (Z_i, X_i) bedingt. Diese Methode ist auf Validierungsdaten anwendbar, wenn zwei Regressionsmodelle vorliegen, nämlich Y_i gegeben (Z_i, X_i) und X_i^* gegeben

$(\mathbf{Z}_i, \mathbf{X}_i)$.

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

4.3. Algorithmus des Bootstrappings in der Regressionskalibrierung

Unabhängig davon welches Bootstrapverfahren angewendet wird, hat der Algorithmus des Bootstrappings in der Regressionskalibrierung 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 $\{(\mathbf{Y}_i, \mathbf{X}_i, \mathbf{X}_i^*, \mathbf{Z}_i)\}_i^{n(m)}$.
 - Nach *Resampling Residuals* auf $\{(\mathbf{Y}_i^{(m)}, \mathbf{X}_i^{*(m)}, \mathbf{Z}_i, \mathbf{X}_i)\}_i^n$.

Man erhält somit nach M Durchläufen die Parameter $\hat{\beta}_{RK_k}^{(1)}, \dots, \hat{\beta}_{RK_k}^{(M)}$.

- Schritt 3: Aus den vorliegenden $\hat{\beta}_{RK_k}$ s kann nun die Standardabweichung $\hat{\sigma}_{\beta_{RK_k}}$ geschätzt werden

$$\text{var}(\widehat{\beta}_{RK_k}) = \frac{1}{M-1} \sum_{m=1}^M (\hat{\beta}_{RK_k}^{(m)} - \overline{\hat{\beta}_{RK_k}})(\hat{\beta}_{RK_k}^{(m)} - \overline{\hat{\beta}_{RK_k}})^t. \quad (16)$$

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

$H_0: \hat{\beta}_{RK_k} = \beta_{RK_k}$.

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 $E(\mathbf{Y}|\mathbf{X}, \mathbf{Z}) = \beta_0 + \beta_1\mathbf{X} + \beta_2\mathbf{Z}$ gelten. Angenommen in diesem Beispiel ist \mathbf{X} die einzige Einflussgröße, und statt des wahren Wertes von \mathbf{X} liegen je Beobachtungseinheit eine Messwiederholung $\mathbf{X}_{ik_i}^* = \mathbf{X}_i^*$ für alle i (also $k_i = 1$) vor, wobei \mathbf{X}_i^* bedingt unabhängig von $(\mathbf{Y}_i, \mathbf{Z}_i)$ gegeben \mathbf{X}_i ist. Dann gilt vereinfacht für die interessierende Größe \mathbf{Y} die Gleichung $E(\mathbf{Y}|\mathbf{X}) = \beta_0 + \beta_1\mathbf{X}$. Die Schritte der Regressionskalibrierung seien wie folgt:

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

$$E[\widehat{\mathbf{X}}|\widehat{\mathbf{X}}^*] \approx \underbrace{\frac{\hat{\sigma}_x^2}{\hat{\sigma}_{x^*}^2}}_{\hat{\gamma}_1} \mathbf{X}^* + \underbrace{\hat{\mu}_{x^*} \left(1 - \frac{\hat{\sigma}_x^2}{\hat{\sigma}_{x^*}^2}\right)}_{\hat{\gamma}_0} = \widehat{\mathbf{X}} \quad (17)$$

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

$$\hat{\mu}_x = \hat{\mu}_{x^*} = \frac{\sum_{i=1}^n \mathbf{X}_i^*}{n}; \quad (18)$$

$$\hat{\Sigma}_{xx} = \hat{\sigma}_x^2 = \frac{\sum_{i=1}^n (\mathbf{X}_i^* - \hat{\mu}_{x^*})(\mathbf{X}_i^* - \hat{\mu}_{x^*})^t}{(n-1)} - \hat{\Sigma}_{uu}. \quad (19)$$

- Schritt 2: Ersetze die unbeobachtete Variable \mathbf{X} durch die Schätzungen im letzten

Schritt

$$E(Y|X^*) =^1 E(\{E(Y|X, X^*)\}|X^*) \quad (20)$$

$$=^2 E(\{E(Y|X)\}|X^*) \quad (21)$$

$$=^3 E(\{\beta_0 + \beta_1 X\}|X^*) \quad (22)$$

$$= \beta_0 + \beta_1 E(X|X^*) \quad (23)$$

$$\approx \beta_{Rk_0} + \beta_{Rk_1} E(\widehat{X}|X^*) \quad (24)$$

$$\approx \beta_{Rk_0} + \beta_{Rk_1} \left(\frac{\hat{\sigma}_x^2}{\hat{\sigma}_{x^*}^2} \mathbf{X}^* + \hat{\mu}_{x^*} \left(1 - \frac{\hat{\sigma}_x^2}{\hat{\sigma}_{x^*}^2}\right) \right) \quad (25)$$

$$\approx \beta_{Rk_0} \beta_{Rk_1} \hat{\mu}_{x^*} \left(1 - \frac{\hat{\sigma}_x^2}{\hat{\sigma}_{x^*}^2}\right) + \left(\beta_{Rk_1} \frac{\hat{\sigma}_x^2}{\hat{\sigma}_{x^*}^2}\right) \mathbf{X}^* \quad (26)$$

$$\approx \underbrace{\beta_{Rk_0} \beta_{Rk_1} \hat{\gamma}_0}_{\beta_{naiv_0}} + \underbrace{\beta_{Rk_1} \hat{\gamma}_1}_{\beta_{naiv_1}} \mathbf{X}^*. \quad (27)$$

¹ iterierter Erwartungswert, Satz der totalen Wahrscheinlichkeit

² X^* differentieller Fehler (A)

³ $E(\mathbf{Y}|\mathbf{X}) = \beta_0 + \beta_1 \mathbf{X}$

Beachte, β_{naiv_0} und β_{naiv_1} sind die Parameter einer naiven Regression von \mathbf{Y} auf \mathbf{X}^* , d.h. wenn nicht berücksichtigt wird, dass Messfehler vorliegen. Unter der Bedingung, dass (17) gilt, kann man die Schätzer der wahren Effekte β_0, β_1 wie folgt extrahieren:

$$\widehat{\beta}_1 \approx \widehat{\beta}_{Rk_1} = \frac{\widehat{\beta}_{naiv_1}}{\widehat{\gamma}_1}, \quad \widehat{\beta}_0 \approx \widehat{\beta}_{Rk_0} = \frac{\widehat{\beta}_{naiv_0}}{\widehat{\gamma}_0 \widehat{\beta}_1}, \quad (28)$$

sodass im Grunde genommen die korrigierten Schätzer von β_0, β_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 $\sigma_{\beta_{Rk_0}}^2$ und $\sigma_{\beta_{Rk_1}}^2$ geschätzt werden. Dazu werden M Bootstrapping-Stichproben gezogen. Durch Wiederholen von Schritt 1 und Schritt 2 in jeder Stichprobe erhält man $\widehat{\beta}_{Rk_0}^{(1)}, \dots, \widehat{\beta}_{Rk_0}^{(M)}$ und $\widehat{\beta}_{Rk_1}^{(1)}, \dots, \widehat{\beta}_{Rk_1}^{(M)}$.

Sei $j \in \{0, 1\}$, lässt sich die Varianz der Parameterschätzer wie folgt schätzen:

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

Die Regressionskalibrierung ist nicht nur in der Theorie anwendbar, sondern auch in der Praxis. Zur Veranschaulichung der praktischen Anwendung sind im Anhang Rcodes (C) zur Simulation von Beispielen für die Anwendung der Regressionskalibrierung 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 die Anwendung der Regressionskalibrierung am realen Beispiel der WHO MONICA (MONitoring of trends and determinants on Cardiovascular disease) illustriert. 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 detailliert 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 Epidemiologen 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 „Aufreten von Herzinfarkt“ 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 Ergebnis präsentiert.

6.2. Ergebnisse

Um die Fragestellung, ob die Ernährung einen Einfluss auf Herz-Kreislauf-Erkrankungen hat, wurde ein Cox-Modell zur Modellierung von Überlebenszeiten angewendet. Wie ein-

gangs 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 mit homoskedastischen Fehlern und mit heteroskedastischen Fehlern anhand von Messwiederholungen vorgeschlagen. Die Theorie für die erste Vorgehensweise wurde im Kapitel 3.4. dieser Arbeit dargelegt, zur zweiten wird auf [Augustin et al., 2008] 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 \mathbf{X}_{ij} , während alle fehlerfreien Messungen als \mathbf{Z}_{ij} bezeichnet werden. Wobei j angibt um welche Kovariable es sich handelt, i um welche Beobachtung und $\underline{\mathbf{X}}'_i = (\mathbf{X}_{i1}, \dots, \mathbf{X}_{ip})$ bzw. $\underline{\mathbf{Z}}'_i = (\mathbf{Z}_{i1}, \dots, \mathbf{Z}_{ik})$ ein Vektor der i-ten Beobachtung mit p bzw. k Einflussgrößen darstellt. $\underline{\mathbf{X}}_i^2$ ist ein Vektor der die quadratischen Komponenten von $\underline{\mathbf{X}}'_i$ enthält. Die individuelle Hazard Rate $\lambda(t|\underline{X}_i, \underline{Z}_i)$ hat folgende Form:

$$\lambda(t|\underline{X}_i, \underline{Z}_i) = \lambda_0(t) \exp(\beta'_1 \underline{X}_i + \beta'_2 \underline{X}_i^2 + \beta'_z \underline{Z}_i) \quad (29)$$

mit der nicht spezifizierten baseline Hazard Rate $\lambda_0(t)$ und den Paramtern $\beta'_0, \beta'_1, \beta'_z$. Statt $\underline{\mathbf{X}}_i$ wurden sieben Messwiederholungen (k=7) für jede Beobachtung erhoben, wobei diese Messungen dem klassischen Fehlermodell unterliegen. Um den geschätzten Effekt von tierischen bzw. pflanzlichen Proteinen darstellen zu können, werden folgende quadratische Funktionen zur Hilfe genommen:

$$f(x_{ANIMAL}) = \hat{\beta}_{ANIMAL} x_{ANIMAL} + \hat{\beta}_{ANIMAL^2} x_{ANIMAL}^2, \quad (30)$$

$$g(x_{PLANT}) = \hat{\beta}_{PLANT} x_2 + \hat{\beta}_{PLANT^2} x_2^2, \quad (31)$$

diese sind monoton zum Risikoeffekt der tierischen bzw. pflanzlichen Fette im Cox-Modell (29). Grafisch sind die quadratischen Kurven in Abbildung 1 dargestellt.

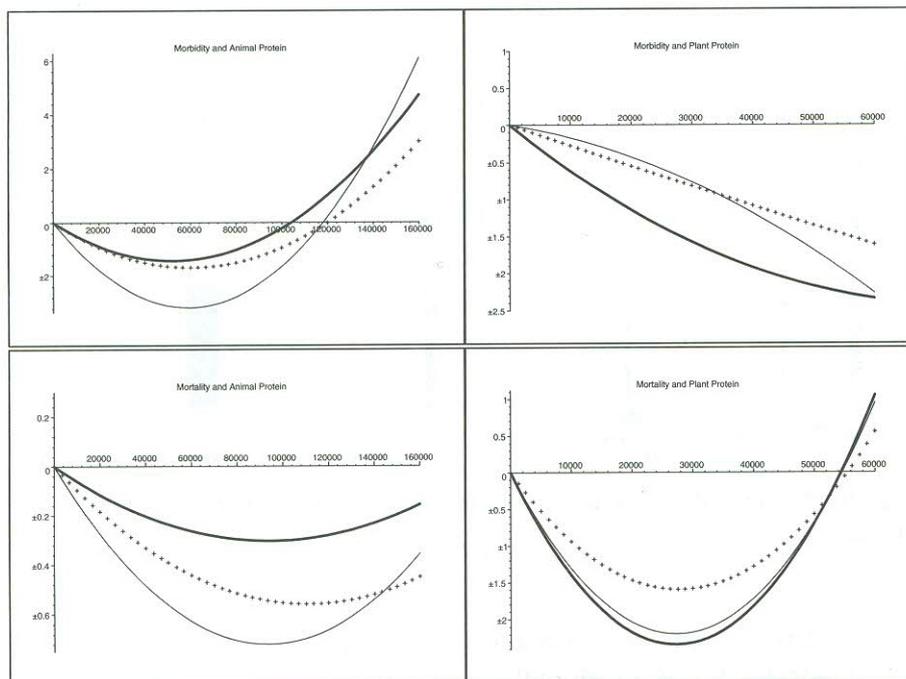


Abbildung 1: Geschätzter Gesamteinfluss der tierischen (*linke Grafiken*) und pflanzlichen Proteine (*rechte Grafiken*) auf die Herzinfakt-Erkrankung (*obere Grafiken*) und die Sterblichkeit (*untere Grafiken*), geschätzt durch naive Schätzung (*gepunktete Linie*), nach Messfehlerkorrektur für homoskedastische Fehler (*dünne, durchgezogene Linie*), nach Messfehlerkorrektur für heteroskedastische Fehler (*dicke, durchgezogene Linie*).

Die abgebildeten Kurven sind abhängig von den β -Schätzern aus der naiven Schätzung und der Schätzung aus den verschiedenen Regressionskalibrierungsansätzen, eingesetzt in (30), (31). 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:

Vorteile

- effektive Methode im Umgang mit fehlerhaften gemeinsam 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 außerdem 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 additive Fehler umwandelt und das Regressionskalibrierungsverfahren anpasst (Näheres [Carroll 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 (Näheres [Augustin et al., 2008]). Die vorgestellten Regressionskalibrierungsverfahren funktionieren für Modelle der generalisierten linearen Regression, für höhere nichtlineare Probleme sind Transformationen der Regressionskalibrierungsverfahren erforderlich (Näheres [Carroll 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) oder die der korrigierten Scorefunktion, die zwar schwierig zu berechnen sind, aber in vielen Fällen bessere Schätzer liefern.

A. Nicht-differentieller Fehler

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

$$E[Y|X, X^*, Z] = E[Y|X, Z]. \quad (32)$$

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.

$$E(Y|X^*, Z) = E(\{E(Y|X, X^*, Z)\}|X^*, Z) \quad (33)$$

$$= E(\{E(Y|X, Z)\}|X^*, Z) \quad (34)$$

$$= E(\{\beta_0 + \beta_1 X + \beta_2 Z\}|X^*, Z) \quad (35)$$

$$= \beta_0 + \beta_1 E(X|X^*, Z) + \beta_2 E(Z|X^*, Z) \quad (36)$$

$$= \beta_0 + \beta_1 E(X|X^*, Z) + \beta_2 Z \quad (37)$$

Eine lineare Regression auf den beobachteten Daten $(\mathbf{Y}, \mathbf{X}^*, \mathbf{Z})$ kann zu einer linearen Regression von \mathbf{Y} auf $E[\mathbf{X}|\mathbf{X}^*, \mathbf{Z}]$ 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 $(\mathbf{X}^*, \mathbf{Z})$ gesetzt werden kann. An dieser Stelle knüpft Schritt 2 in (5) an.

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 Interesse, ob eine Annahme über die beobachtete Variable \mathbf{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 [Buonaccors, 1986]. Eine Übersicht bietet folgende Tabelle:

	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:		
	$E[X^* X] = X$	$E[X X^*] = X^*$
	$V(X^*) = V(X + U) =$	$V(X) = V(X^* + U) =$
	$V(X) + V(U) > V(X)$	$V(X^*) + V(U) > V(X^*)$

Insbesondere erhält man in der linearen Regression, beim vorliegen eines Berkson-Fehlers, unverzerrte KQ-Schätzungen [Schneeweiß and Mittag, 1986], 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). Einen Beweis hierfür findet man in [Caroll et al., 2006] (Kapitel 3.2.2.)

C. Beispiel: Regressionskalibrierung Rcode

In diesem Anhang werden die Rcodes bereitgestellt, die im Laufe der Arbeit entstanden sind. Teilweise sind Outputs enthalten, die zum Verständnis beitragen sollen.

C.1. Daten generieren

Zu Beginn wurden Daten zufällig generiert. Die Daten enthalten insgesamt 500 Beobachtungen mit den Spalten $Y, X, X^*, X_2^*, X_3^*, X_4^*$. Auf Basis dieser Daten werden in den folgenden Kapiteln Validierungsdaten und Wiederholungsdaten modelliert.

DatenGenerieren.R

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

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

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

regcalDummyVergleich.R

```
1 ###generieren von Daten aus dataOrigin, mit diesen im folgenden gearbeitet wird##
2 #####angenommen es liegen nur fuer die Haelfte(250) Validierungsdaten vor#####
3
4 #Nur fuer Validierungsdaten liegen Y, X* UND wahre X vor
5 validierung <- sample(500,250, replace=FALSE)
6 validierungSub <- dataOrigin[validierung, ][c(1,2,3)]
7 validierungSub$Valid <- 1
8
9 #Fuer die restlichen Daten liegen keine wahren X-Werte vor
10 NotvalidierungSub <- dataOrigin[setdiff(seq(1:500), validierung), ][c(1,2,3)]
11 NotvalidierungSub[,"x"] <- NA
12 NotvalidierungSub$Valid <- 0
13
14 #Daten auf die die Regressionskalibrierung spaeter angewendet wird.
15 dataValid<-merge(validierungSub, NotvalidierungSub, all=T)
16 head(dataValid)
17
18 ##           y           x           xSt           Valid
19 ## 1 -0.4952122          NA -1.214837           0
20 ## 2 -0.4139702          NA -2.902119           0
21 ## 3 -0.3307201 -1.533457 -1.305155           1
22 ## 4 -0.3231430          NA -2.346582           0
23 ## 5 -0.3101120 -2.525401 -2.347733           1
24 ## 6 -0.3087614 -2.153913 -2.736493           1
25
26
27 #####Bootstrap zur Kontrolle mit oder ohne Dummy #####
28 #100 Bootstrapstichproben
29 B <- 100
30 i <- 1
31 beta_0 <- c()
32 beta_1 <- c()
33 beta_0_Dummy <- c()
34 beta_1_Dummy <- c()
35 dataValid$x_hat <- NA
36
37 while( i <= B){
38
39   #Bootstrap ziehen
40   #dataValid besitzt zwei verschiedene Datenstrukturen (Validierungsdaten
41   #und Daten ohne wahren X-Werte)
42   #Daten muessen vor der Bootstrap-Ziehung nach den
43   #Strukturen getrennt werden (siehe Kapitel 4)
44
45   #Bootstrap von Vektoren aus Validierungsdaten
46   beob1 <- rownames(dataValid[dataValid$Valid==1, ])
47   #Ziehen Vektoren mit zuruecklegen aus Validierungsdaten
48   which1 <- sample(beob1, length(beob1), replace=TRUE)
49   bootValid1 <- dataValid[dataValid$Valid==1, ][which1, ]
50
51   #Bootstrap von Vektoren aus Daten ohne wahren X-Werte
```

```

52 beob0 <- rownames(dataValid[dataValid$Valid==0, ])
53 #Ziehen Vektoren mit zuruecklegen aus Validierungsdaten
54 which0 <- sample(beob0, length(beob0), replace=TRUE)
55 bootValid0 <- dataValid[dataValid$Valid==0, ][which0, ]
56
57 #Bootstrap-Daten
58 dataBoot <- merge(bootValid1, bootValid0, all=T)
59
60 ###Schritt 1: Eine Regression von X auf (X*,Z=0) rechnen aus Validierungsdaten
61
62 #Beachte: Zeile, wo x=NA werden automatisch nicht beruecksichtigt
63 lm_xSt <- lm(x~xSt, data=dataBoot)
64
65 #Schaetzer anwenden auf alle X* fuer die keine wahren X vorliegen und
66 #in neue Variable x_hat speichern
67 dataBoot[(dataBoot$Valid==0),]$x_hat <- coef(lm_xSt)[1]+coef(lm_xSt)[2]*dataBoot[(
  dataBoot$Valid==0),]$xSt
68
69 ###Schritt 2: nicht vorhandene wahre X durch die Schaetzung x_hat aus
70 ###Regression in Schritt 1 ersetzen.
71 #x_reg soll die wahren X-Werte enthalten und wo die wahren X-Werte fehlen
72 #enthaelt sie die geschaezten X-Werte
73 dataBoot$x_reg <- dataBoot[, "x"]
74 dataBoot[dataBoot$Valid==0,]$x_reg <- dataBoot[dataBoot$Valid==0,]$x_hat
75
76 #####Ohne Dummy #####
77
78 lm_rk <- lm(y~x_reg, data=dataBoot)
79
80
81 beta_0[i] <- lm_rk$coef[1]
82 beta_1[i] <- lm_rk$coef[2]
83
84
85 #####mit Dummy#####
86
87 lm_rkD <- lm(y~x_reg+Valid, data=dataBoot)
88
89 beta_0_Dummy[i] <- lm_rkD$coef[1]
90 beta_1_Dummy[i] <- lm_rkD$coef[2]
91
92 i <- i+1
93 }
94
95
96 #####Vergleich Kennzahlen#####
97
98 mean_b_0 <- mean(beta_0)
99 mean_b_1 <- mean(beta_1)
100 mean_b_0_Dummy <- mean(beta_0_Dummy)
101 mean_b_1_Dummy <- mean(beta_1_Dummy)
102 var_b_0 <- var(beta_0)
103 var_b_1 <- var(beta_1)
104 var_b_0_Dummy <- var(beta_0_Dummy)
105 var_b_1_Dummy <- var(beta_1_Dummy)
106 bias_b_0 <- mean(beta_0-intercept)
107 bias_b_1 <- mean(beta_1-intercept)
108 bias_b_0_Dummy <- mean(beta_0_Dummy-beta)
109 bias_b_1_Dummy <- mean(beta_1_Dummy-beta)
110 MSE_b_0 <- bias_b_0^2+var_b_0
111 MSE_b_1 <- bias_b_1^2+var_b_1
112 MSE_b_0_Dummy <- bias_b_0_Dummy^2+var_b_0_Dummy

```

```

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

```

Zeile 128-138 zeigt einen Tabelle, vor allem am MSE von $\widehat{\beta}_1$ wird deutlich, dass eine Modellierung mit Dummyvariable zur besseren Schätzung führt.

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

regcalDummyPlot.R

```
1 ###generieren von Daten aus dataOrigin, mit diesen im folgenden gearbeitet wird##
2 #####angenommen es liegen nur fuer die Haelfte(250) Validierungsdaten vor#####
3
4 #Nur fuer Validierungsdaten liegen Y, X* UND wahre X vor
5 validierung <- sample(500,250, replace=FALSE)
6 validierungSub <- dataOrigin[validierung, ][c(1,2,3)]
7 validierungSub$Valid <- 1
8
9 #Fuer die restlichen Daten liegen keine wahren X-Werte vor
10 NotvalidierungSub <- dataOrigin[setdiff(seq(1:500), validierung), ][c(1,2,3)]
11 NotvalidierungSub[,"x"] <- NA
12 NotvalidierungSub$Valid <- 0
13
14 #Daten auf die die Regressionskalibrierung spaeter angewendet wird.
15 dataValid <- merge(validierungSub, NotvalidierungSub, all=T)
16 head(dataValid)
17
18 ##          y          x          xSt          Valid
19 ## 1 -0.4952122          NA -1.214837          0
20 ## 2 -0.4139702          NA -2.902119          0
21 ## 3 -0.3307201 -1.533457 -1.305155          1
22 ## 4 -0.3231430          NA -2.346582          0
23 ## 5 -0.3101120 -2.525401 -2.347733          1
24 ## 6 -0.3087614 -2.153913 -2.736493          1
25
26 #####interessiert an Regression von Y auf (X,Z=0), aber eigentlich X nicht
27 #####beobachtet
28 ##visualisierern der Effekte der naiven Schaetzung und der korrigierten
29 ##Schaetzung mit RK
30
31 #plot der wahren Werte (aus dataOrigin)
32 par(xpd=F)
33 plot(dataOrigin$x, dataOrigin$y, col="black", pch=18, ylab="Y", xlab="")
34 abline(lm(y~x, dataOrigin), col="black", lwd=3)
35 par(xpd=T)
36 text(-1,-3, labels="X", col="black")
37
38 #eine Regression von Y auf (X*,Z=0) statt auf (X,Z) fuehrt zur Verzerrung
39
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(0, -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 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),]$x_hat<-coef(lm_xSt)[1]+coef(lm_xSt)[2]*
58 dataValid[(dataValid$Valid==0),]$xSt
```

```

59
60 #####Ohne Dummy #####
61
62 ###Schritt 2: nicht vorhandene wahre X durch die Schaetzung x_hat aus
63 ###Regression in Schritt 1 ersetzen.
64 #x_reg enthaelt die wahren X-Werte und wo die wahren X-Werte fehlen
65 #enthaelt sie die geschaeztzten X-Werte
66 dataValid$x_reg<-dataValid[,"x"]
67 dataValid[dataValid$Valid==0,]$x_reg <- dataValid[dataValid$Valid==0,]$x_hat
68
69 #plot werte nach RK
70 par(xpd=F)
71 points(dataValid$x_reg, dataValid$y, col="blue", pch=1)
72
73 lm_rk <- lm(y~x_reg, data=dataValid)
74 summary(lm_rk)
75 #RK RegGerade nahe der wahren RegGerade
76 abline(lm_rk, col="blue", lwd=3, lty=3)
77 par(xpd=T)
78 text(1, -3,labels="X_Rk", col="blue")
79 par(xpd=F)
80
81
82 #####mit Dummy#####
83
84 ###Schritt 2: nicht vorhandene wahre X durch die Schaetzung x_hat aus
85 ###Regression in Schritt 1 ersetzen.
86 #x_reg enthaelt die wahren X-Werte und wo die wahren X-Werte fehlen
87 #enthaelt sie die geschaeztzten X-Werte
88 dataValid$x_reg <- dataValid[,"x"]
89 dataValid[dataValid$Valid==0,]$x_reg <- dataValid[dataValid$Valid==0,]$x_hat
90
91 #plot Werte nach RK
92 par(xpd=F)
93 points(dataValid$x_reg, dataValid$y, col="red", pch=1)
94
95 lm_rkVD <- lm(y~x_reg+Valid, data=dataValid)
96 summary(lm_rkVD)
97 #RK RegGerade fuer Valid=0
98 abline(coef(lm_rkVD)[1],coef(lm_rkVD)[2],col="red", lwd=3, lty=2)
99 #RK RegGerade fuer Valid=1
100 abline(coef(lm_rkVD)[1]+coef(lm_rkVD)[3],coef(lm_rkVD)[2],col="orange",
101        lwd=3, lty=2)
102 par(xpd=T)
103 text(2, -3,labels="X_Rk_Dummy", col="red")
104 par(xpd=F)

```

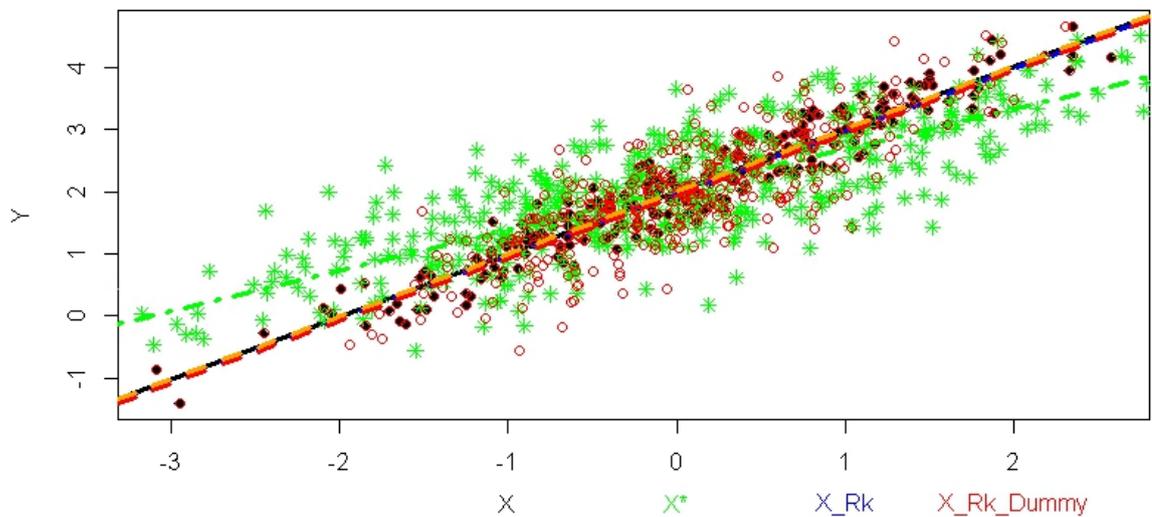


Abbildung 2: *Schwarze durchgezogene Linie* ist die wahre Regressionsgerade. *Grün gepunktet und gestrichelte Linie* entspricht der Regressionsgerade aus der naiven Schätzung. Die restlichen entstammen aus der Schätzung nach Anwendung der Regressionskalibrierung. Die *blaue gepunktete Linie* ist die 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.

C.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.

regcalValid.R

```
1 #####generieren von Daten aus dataOrigin, mit diesen im folgenden gearbeitet wird
2 #angenommen es liegen nur fuer die Haelfte (250) Validierungsdaten vor
3
4 #Nur fuer Validierungsdaten liegen Y, X* UND wahre X vor
5 validierung<-sample(500,250, replace=FALSE)
6 validierungSub<-dataOrigin[validierung, ][c(1,2,3)]
7 validierungSub$Valid<-1
8
9 #Fuer die restlichen Daten liegen keine wahren X-Werte vor
10 NotvalidierungSub<-dataOrigin[setdiff(seq(1:500), validierung), ][c(1,2,3)]
11 NotvalidierungSub[, "x"]<-NA
12 NotvalidierungSub$Valid<-0
13
14 #Daten auf die die Regressionskalibrierung spaeter angewendet wird.
15 dataValid<-merge(validierungSub, NotvalidierungSub, all=T)
16 head(dataValid)
17 #      y          x      xSt      Valid
18 # 1 -1.1167461 -3.295136 -3.634460      1
19 # 2 -0.7145003          NA -1.903810      0
20 # 3 -0.5470220 -2.103604 -2.585469      1
21 # 4 -0.5008263          NA -2.003854      0
22 # 5 -0.4592409 -2.100544 -2.134435      1
23 # 6 -0.4576016          NA -2.202306      0
24
25 #####interessiert an Regression von Y auf (X,Z=0), aber eigentlich X nicht
26 #####beobachtet
27 ##visualisieren der Effekte der naiven Schaetzung und der korrigierten
28 ##Schaetzung mit RK
29
30 #plot der wahren Werte (aus dataOrigin)
31 par(xpd=F)
32 plot(dataOrigin$x, dataOrigin$y, col="black", pch=18, ylab="Y", xlab="")
33 abline(lm(y~x, dataOrigin), col="black", lwd=3)
34 par(xpd=T)
35 text(0,-3, labels="X", col="black")
36
37 #eine Regression von Y auf (X*,Z=0) statt auf (X,Z) fuehrt zur Verzerrung
38 #(aus dataValid)
39 par(xpd=F)
40 points(dataValid$xSt, dataValid$y, col="green", pch=8)
41 abline(lm(y~xSt, data=dataValid), col="green", lwd=3, lty=4)
42 par(xpd=T)
43 text(1, -3, labels="X*", col="green")
44
45
46 #####
47 #####Regressionskalibrierung Verfahren (siehe Kapitel 2.2)
48
49 ###Schritt 1: Eine Regression von X auf (X*,Z=0) rechnen aus
50 ###Validierungsdaten
51
52 #Beachte: Zeile, wo x=NA werden automatisch nicht beruecksichtigt
53 lm_xSt<-lm(x~xSt, data=dataValid)
```

```

54
55 #Schaetzer anwenden auf alle X* fuer die keine wahren X vorliegen uns
56 #in neue Variable x_hat speichern
57 dataValid$x_hat<-NA
58 dataValid[(dataValid$Valid==0),]$x_hat<-coef(lm_xSt)[1]+coef(lm_xSt)[2]*
59   dataValid[(dataValid$Valid==0),]$xSt
60 head(dataValid)
61
62 ##          y          x          xSt      Valid      x_hat
63 ## 1 -1.1167461 -3.295136 -3.634460      1      NA
64 ## 2 -0.7145003          NA -1.903810      0 -1.216047
65 ## 3 -0.5470220 -2.103604 -2.585469      1      NA
66 ## 4 -0.5008263          NA -2.003854      0 -1.280737
67 ## 5 -0.4592409 -2.100544 -2.134435      1      NA
68 ## 6 -0.4576016          NA -2.202306      0 -1.409060
69
70 ###Schritt 2: nicht vorhandene wahre X durch die Schaetzung x_hat aus
71 ###Regression in Schritt 1 ersetzen.
72 #x_reg enthaelt die wahren X-Werte und wo die wahren X-Werte fehlen
73 #enthaelt sie die geschaetzten X-Werte, speichern in neue Variable x_reg
74 dataValid$x_reg<-dataValid[,"x"]
75 dataValid[dataValid$Valid==0,$x_reg<-dataValid[dataValid$Valid==0,$x_hat]
76 head(dataValid)
77
78 ##          y          x          xSt      Valid      x_hat      x_reg
79 ## 1 -1.1167461 -3.295136 -3.634460      1      NA -3.295136
80 ## 2 -0.7145003          NA -1.903810      0 -1.216047 -1.216047
81 ## 3 -0.5470220 -2.103604 -2.585469      1      NA -2.103604
82 ## 4 -0.5008263          NA -2.003854      0 -1.280737 -1.280737
83 ## 5 -0.4592409 -2.100544 -2.134435      1      NA -2.100544
84 ## 6 -0.4576016          NA -2.202306      0 -1.409060 -1.409060
85
86 #plot werte nach RK
87 par(xpd=F)
88 points(dataValid$x_reg, dataValid$y, col="red", pch=1)
89
90 lm_rk_valid<-lm(y~x_reg, data=dataValid)
91 abline(lm_rk_valid, col="red", lwd=3, lty=2) #RK RegGerade nahe der wahren
  RegGerade
92 par(xpd=T)
93 text(2, -3, labels="X_Rk", col="red")
94 par(xpd=F)
95
96
97 #Schritt 3: Anpassen der Standardfehler
98 #Bootstrapping (Resampling vectors im Messfehlermodell (Kaptitel 4.1))
99
100 #10 Bootstraptichproben- Resampling vectros Verfahren
101 B <- 100
102 i<-1
103 beta_rk_0_valid <- numeric(B)
104 beta_rk_1_valid <- numeric(B)
105 while( i <= B){
106   #dataValid besitzt zwei verschiedene Datenstrukturen (Validierungsdaten
107   #und Daten ohne wahren X-Werte)
108   #Daten muessen vor der Bootstrap-Ziehung getrennt werden nach den
109   #Strukturen getrennt werden (siehe Kapitel 4)
110
111   #Bootstrap von Vektoren aus Validierungsdaten
112   beob1 <- rownames(dataValid[dataValid$Valid==1, ])
113   #Ziehen Vektoren mit zuruecklegen aus Validierungsdaten
114   which1 <- sample(beob1, length(beob1), replace=TRUE)

```

```

115 bootValid1 <- dataValid[dataValid$Valid==1, ][which1, ]
116
117 #Bootstrap von Vektoren aus Daten ohne wahren X-Werte
118 beob0 <- rownames(dataValid[dataValid$Valid==0, ])
119 #Ziehen Vektoren mit zuruecklegen aus Validierungsdaten
120 which0 <- sample(beob0, length(beob0), replace=TRUE)
121 bootValid0 <- dataValid[dataValid$Valid==0, ][which0, ]
122
123 #Bootstrap-Daten
124 dataBoot <- merge(bootValid1, bootValid0, all=T)
125 head(dataBoot)
126
127 #Regressionskalibrierung
128 #Schritt 1:
129 #Zeilen mit Na in x werden automatisch geloescht
130 lm_xSt <- lm(x~xSt, data=dataBoot)
131
132 #Anwendung der Schaetzung auf diejenigen X*, fuer die die wahren
133 #X nicht vorliegen
134 x_hat_boot <- coef(lm_xSt)[1]+coef(lm_xSt)[2]*
135   (dataBoot[(dataBoot$Valid==0), ])$xSt
136
137 ##x_reg_boot enthaelt die wahren X-Werte und wo die wahren X-Werte
138 #fehlen enthaelt sie die geschaetzten X-Werte
139 dataBoot$x_reg_boot<-dataBoot[ , "x"]
140 dataBoot[dataBoot$Valid==0, ]$x_reg_boot <- x_hat_boot
141
142 #Schritt 2: wahre X durch die Schaetzung aus Regression in Schritt 1
143 #ersetzen
144 lm_rk<-lm(y~x_reg_boot, data=dataBoot)
145
146 #speichern der Paramter
147 beta_rk_0_valid[i] <- coef(lm_rk)[1]
148 beta_rk_1_valid[i] <- coef(lm_rk)[2]
149 i<-i+1
150 }
151
152 #Parameterschaetzer aus den Bootstraps
153 beta_rk_0_valid
154 beta_rk_1_valid
155
156 #Varianz &Standardabweichung
157 var(beta_rk_0_valid)
158 sd(beta_rk_0_valid)
159
160 var(beta_rk_1_valid)
161 sd(beta_rk_1_valid)
162
163 summary(lm_rk_valid) #aus Schritt 2 zum vergleich

```

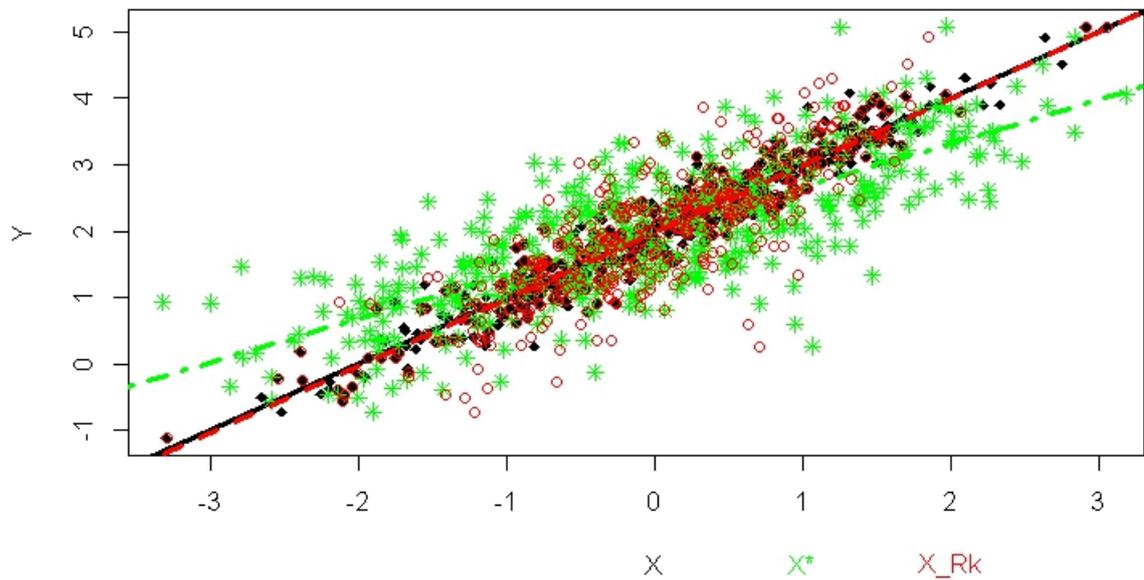


Abbildung 3: *Schwarze durchgezogene Linie* ist die wahre Regressionsgerade. *Grün gepunktet und gestrichelte Linie* entspricht der Regressionsgerade aus der naiven Schätzung. Die *rot gestrichelte Linie* stellt die Schätzung nach Anwendung der Regressionskalibrierung auf Validierungsdaten dar.

C.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 (17))
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 schaaetzen
42 n<-dim(dataWdh1)[1]
43 muXStern_hat <- sum(dataWdh1$xSt)/n #(Formel 18)
44 varX_hat<-sum((dataWdh1$xSt-muXStern_hat)^2)/(n-1)-varU_hat #(Formel 19)
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+
   muXStern_hat*(1-(varX_hat/varXStern_hat)) #(Formel 3)
49 points(dataWdh1$ewX_Xst1, dataWdh1$y,col="red", pch=1)
50
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 #100 Bootstrapstichproben- Resampling vectors Verfahren
61 i <- 1
62 B <- 100
63 beta_rk_0_wdh1 <- numeric(B)
64 beta_rk_1_wdh1 <- numeric(B)
65 dataWdh1$x_hat_boot<-c()
66 for( i in 1:B){
67
68   beob <- rownames(dataWdh1)
69   #Ziehen Vektoren mit zuruecklegen aus Wdh-Daten
70   which <- sample(beob, length(beob), replace=TRUE)
71   bootWdh1 <- dataWdh1[which, ]
72
73   ##Schritt 1
74   #Formel von EW von X bedingt auf X* (Kapitel 5 Formel (2))
75
76   #Kapitel 3.4. besagt, dass die Varianz von u aus externen Daten verwendet
77   #werden kann extern geschaetzt (aus Validierungsdaten) (Kapitel C.4.)
78   varU_hat <- var(dataValid[dataValid$Valid==1,]$xSt-dataValid[dataValid$Valid
       ==1,]$x)
79
80   #Restliche Groessen aus den vorliegenden Daten schaeetzen
81   n<-dim(bootWdh1)[1]
82   muXStern_hat <- sum(bootWdh1$xSt)/n #(Formel 18)
83   varX_hat<-sum((bootWdh1$xSt-muXStern_hat)^2)/(n-1)-varU_hat #(Formel 19)
84   varXStern_hat <- varX_hat+varU_hat #v(x*)=v(x)+v(u)
85
86   #gefittete X
87   bootWdh1$ewX_Xst1 <- (varX_hat/varXStern_hat)*bootWdh1$xSt+
88     muXStern_hat*(1-(varX_hat/varXStern_hat)) #(Formel 3)
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)

```

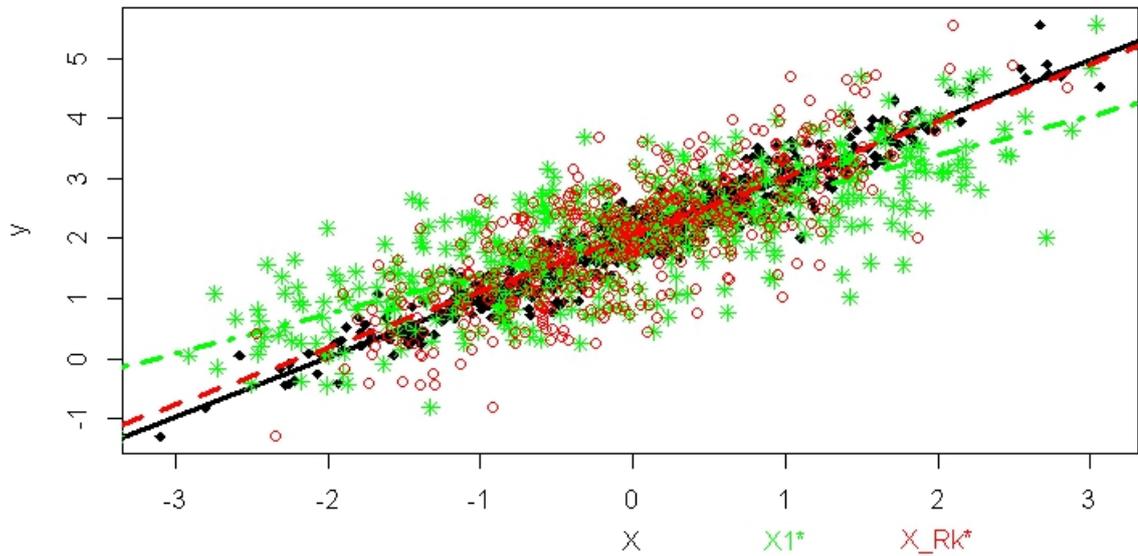


Abbildung 4: *Schwarze durchgezogene Linie* ist die wahre Regressionsgerade. *Grün gepunktet und gestrichelte Linie* entspricht der Regressionsgerade aus der naiven Schätzung. Die *rot gestrichelte Linie* stellt die Schätzung nach Anwendung der Regressionskalibrierung auf Wiederholungsdaten mit einer Messwiederholung dar.

C.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:

$$\hat{\Sigma}_{uu} = \hat{\sigma}_u^2 = \frac{\sum_{i=1}^n \sum_{j=1}^4 (\mathbf{X}_{ij}^* - \bar{\mathbf{X}}_{i\cdot}^*) (\mathbf{X}_{ij}^* - \bar{\mathbf{X}}_{i\cdot}^*)^t}{3n} \quad (38)$$

$$\bar{\mathbf{X}}_{i\cdot}^* = \frac{X_{i1}^* + X_{i2}^* + X_{i3}^* + X_{i4}^*}{4} \quad (39)$$

$$\hat{\mu}_x = \hat{\mu}_{x^*} = \frac{\sum_{i=1}^n \bar{\mathbf{X}}_{i\cdot}^*}{n} \quad (40)$$

$$\hat{\Sigma}_{xx} = \hat{\sigma}_x^2 = \frac{\sum_{i=1}^n (\bar{\mathbf{X}}_{i\cdot}^* - \hat{\mu}_{x^*}) (\bar{\mathbf{X}}_{i\cdot}^* - \hat{\mu}_{x^*})^t}{(n-1)} - \frac{\hat{\Sigma}_{uu}}{4} \quad (41)$$

$$E[\widehat{\mathbf{X}}|\bar{\mathbf{X}}^*] \approx \frac{4\hat{\sigma}_x^2}{4\hat{\sigma}_x^2 + \hat{\sigma}_u^2} \mathbf{X}^* + \hat{\mu}_{x^*} \left(1 - \frac{4\hat{\sigma}_x^2}{4\hat{\sigma}_x^2 + \hat{\sigma}_u^2}\right) = \widehat{\mathbf{X}} \quad (42)$$

regcalWdh4.R

```

1 #####Regressionskalibrierung mit Wiederholungsdaten ki=4#####
2 #####Regressionskalibrierung Verfahren (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
33 par(xpd=F)

```

```

34 points(dataWdh4$xSt2,dataWdh4$y, col="blue", pch=4)
35 abline(lm(y~xSt2, data=dataWdh4), col="blue", lwd=3, lty=3)
36 par(xpd=T)
37 text(0,-3, labels="X2*", col="blue")
38
39 #Plot Werte mit Fehler 3.Messung
40 par(xpd=F)
41 points(dataWdh4$xSt3,dataWdh4$y, col="darkorchid3", pch=6)
42 abline(lm(y~xSt3, data=dataWdh4), col="darkorchid3", lwd=3, lty=4)
43 par(xpd=T)
44 text(1,-3, labels="X3*", col="darkorchid3")
45
46 #Plot Werte mit Fehler 4.Messung
47 par(xpd=F)
48 points(dataWdh4$xSt4,dataWdh4$y, col="orange", pch=3)
49 abline(lm(y~xSt4, data=dataWdh4), col="orange", lwd=3, lty=3)
50 par(xpd=T)
51 text(2,-3, labels="X4*", col="orange")
52
53
54 #Plot Werte nach RK
55
56 ##Schritt 1
57 #Formel von EW von X bedingt auf X* (Kapitel 3.4 Formel (2))
58 #Kapitel 3.4. stelle eine Formel bereit, wie man die Varianz von X
59 #berechnen kann, wenn X nicht vorliegt
60 n<-dim(dataWdh4)[1]
61 xStern_quer <- (dataWdh4$xSt+dataWdh4$xSt2+dataWdh4$xSt3+dataWdh4$xSt4)/4 #(Formel
39)
62 varU_hat<-sum((dataWdh4$xSt-xStern_quer)^2+(dataWdh4$xSt2-xStern_quer)^2+
63 (dataWdh4$xSt3-xStern_quer)^2+(dataWdh4$xSt4-xStern_quer)^2)/(3*n)
#(Formel 38)
64
65 #Restliche Groessen aus Daten schaeetzen
66 muXStern_hat <- sum(xStern_quer)/n #(Formel 40)
67 varX_hat<-sum((xStern_quer-muXStern_hat)^2)/(n-1)-varU_hat/4 #(Formel 41)
68 varXStern_hat <- varX_hat+varU_hat #v(x*)=v(x)+v(u)
69
70 #gefittete X
71 dataWdh4$ewX_Xst4 <- ((4*varX_hat)/(3*varX_hat+varXStern_hat))*xStern_quer+
72 muXStern_hat*(1-((4*varX_hat)/(3*varX_hat+varXStern_hat))) #(Formel 42)
73 points(dataWdh4$ewX_Xst4, dataWdh4$y,col="red", pch=1)
74
75 ##Schritt 2
76 par(xpd=F)
77 lm_rk_wdh4<-lm(y~ewX_Xst4, data=dataWdh4)
78 abline(lm_rk_wdh4, col="red", lwd=3, lty=2)
79 par(xpd=T)
80 text(3,-3, labels="X_Rk*", col="red")
81
82 ##Schritt 3
83 #100 Bootstrapstichproben- Resampling vectros Verfahren
84 B <- 100
85 beta_rk_0_wdh4 <- numeric(B)
86 beta_rk_1_wdh4 <- numeric(B)
87 dataWdh4$x_hat_boot<-c()
88 for( i in 1:B){
89
90 beob <- rownames(dataWdh4)
91 #Ziehen Vektoren mit zuruecklegen aus Wdh-Daten
92 which <- sample(beob, length(beob), replace=TRUE)
93 bootWdh4 <- dataWdh4[which, ]

```

```

94
95 ##Schritt 1
96 #Formel von EW von X bedingt auf X* (Kapitel 3.4 Formel (2))
97
98 #Kapitel 3.4. stelle eine Formel bereit, wie man die Varianz von X
99 #berechnen kann, wenn X nicht vorliegt
100 n<-dim(bootWdh4)[1]
101 xStern_quer <- (bootWdh4$xSt+bootWdh4$xSt2+bootWdh4$xSt3+bootWdh4$xSt4)/4
102 varU_hat<-sum((bootWdh4$xSt-xStern_quer)^2+(bootWdh4$xSt2-xStern_quer)^2+
103              (bootWdh4$xSt3-xStern_quer)^2+
104              (bootWdh4$xSt4-xStern_quer)^2)/(3*n) #(Formel 38)
105
106 #Restliche Groessen aus Daten schaeetzen
107 muXStern_hat <- sum(xStern_quer)/n #(Formel 40 )
108 varX_hat<-sum((xStern_quer-muXStern_hat)^2)/(n-1)-varU_hat/4 #(Formel 41 )
109 varXStern_hat <- varX_hat+varU_hat #v(x*)=v(x)+v(u)
110
111 #gefittete X
112 bootWdh4$ewX_Xst4 <- ((4*varX_hat)/(3*varX_hat+varXStern_hat))*xStern_quer+
113 muXStern_hat*(1-((4*varX_hat)/(3*varX_hat+varXStern_hat))) #(Formel 42)
114
115 ##Schritt 2
116 lm_rk_wdh4_boot<-lm(y~ewX_Xst4, data=bootWdh4)
117
118 beta_rk_0_wdh4[i] <- coef(lm_rk_wdh4_boot)[1]
119 beta_rk_1_wdh4[i] <- coef(lm_rk_wdh4_boot)[2]
120 }
121
122 #Parameterschaetzer aus den Bootstraps
123 beta_rk_0_wdh4
124 beta_rk_1_wdh4
125
126 #Varianz &Standardabweichung
127 var(beta_rk_0_wdh4)
128 sd(beta_rk_0_wdh4)
129
130 var(beta_rk_1_wdh4)
131 sd(beta_rk_1_wdh4)

```

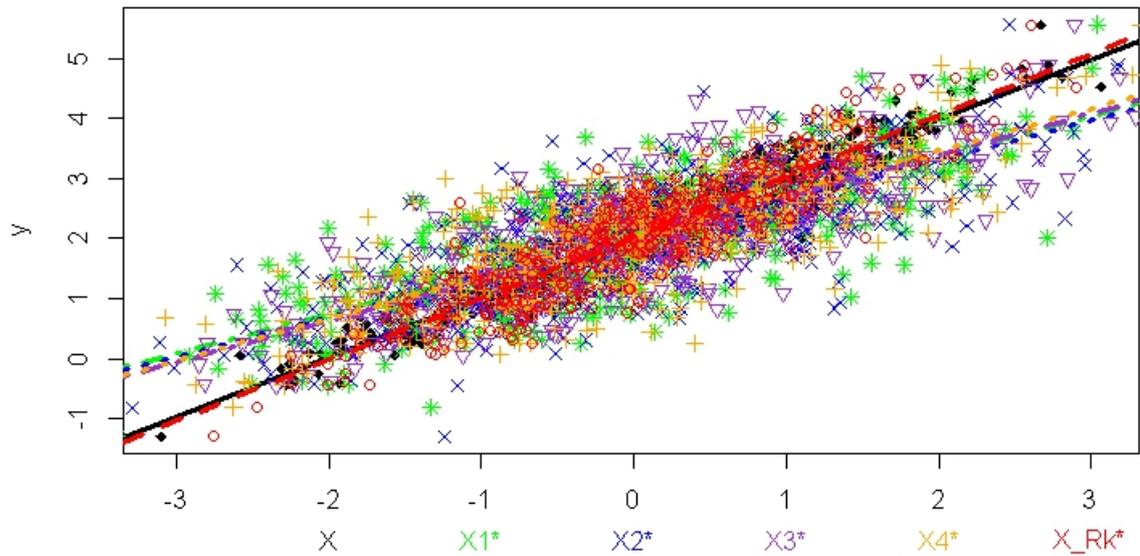


Abbildung 5: *Schwarze durchgezogene Linie* ist die wahre Regressionsgerade. Die *Grün,blau,lila und orange gepunktet und gestrichelte Linien* entsprechen der Regressionsgerade aus der naiven Schätzung. Die *rot gestrichelte Linie* stellt die Schätzung nach Anwendung der Regressionskalibrierung auf Wiederholungsdaten mit einer Messwiederholung dar.

C.7. Vergleich

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

regcalVergleich.R

```
1 #####Vergleich der drei Vorgehen#####
2 #Wahre Werte
3 #plot der wahren Werte (aus dataOrigin)
4 par(xpd=F)
5 plot(dataOrigin$x, dataOrigin$y, col="black", pch=18, ylab="Y", xlab="")
6 abline(lm(y~x, dataOrigin), col="black", lwd=3)
7 par(xpd=T)
8 text(-1.5,-3, labels="X", col="black")
9
10 #Verfahren mit Validierungsdaten
11 par(xpd=F)
12 points(dataValid$x_reg, dataValid$y, col="red", pch=8)
13 #lm_rk_valid<-lm(y~x_reg, data=dataValid)
14 abline(lm_rk_valid, col="red", lwd=3,lty=3)
15 par(xpd=T)
16 text(-0.5,-3, labels="X_Rk_valid", col="red")
17
18 #Verfahren mit einer Messwdh
19 par(xpd=F)
20 points(dataWdh1$ewX_Xst1, dataWdh1$y,col="blue", pch=1)
21 #lm_rk_wdh1<-lm(dataWdh1$y~dataWdh1$ewX_Xst)
22 abline(lm_rk_wdh1, col="blue", lwd=3, lty=2)
23 par(xpd=T)
24 text(1,-3, labels="X_Rk_wdh1", col="blue")
25
26 #Verfahren mit vier Messwdh
27 par(xpd=F)
28 points(dataWdh4$ewX_Xst, dataWdh4$y, col="green")
29 #lm_rk_wdh4<-lm(dataWdh4$y~dataWdh4$ewX_Xst)
30 abline(lm_rk_wdh4, col="green", pch=8, lwd=4,lty=2)
31 par(xpd=T)
32 text(2.5,-3, labels="X_Rk_wdh4", col="green")
33
34 #Tabelle
35 mean_b0_valid<-mean(beta_rk_0_valid)
36 mean_b1_valid<-mean(beta_rk_1_valid)
37
38 mean_b0_wdh1<-mean(beta_rk_0_wdh1)
39 mean_b1_wdh1<-mean(beta_rk_1_wdh1)
40
41 mean_b0_wdh4<-mean(beta_rk_0_wdh4)
42 mean_b1_wdh4<-mean(beta_rk_1_wdh4)
43
44 var_b0_valid<-var(beta_rk_0_valid)
45 var_b1_valid<-var(beta_rk_1_valid)
46
47 var_b0_wdh1<-var(beta_rk_0_wdh1)
48 var_b1_wdh1<-var(beta_rk_1_wdh1)
49
50 var_b0_wdh4<-var(beta_rk_0_wdh4)
51 var_b1_wdh4<-var(beta_rk_1_wdh4)
52
53 bias_b0_valid<-mean(beta_rk_0_valid-intercept)
54 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
72 valid<-c(var_b0_valid, var_b1_valid, MSE_b0_valid, MSE_b1_valid)
73 wd1<-c(var_b0_wdh1, var_b1_wdh1, MSE_b0_wdh1, MSE_b1_wdh1)
74 wd4<-c(var_b0_wdh4, var_b1_wdh4, MSE_b0_wdh4, MSE_b1_wdh4)
75
76 tab<-cbind(valid, wd1, wd4)
77 rownames(tab)<-c("hat_v(b0)", "hat_v(b1)", "MSE(b0)", "MSE(b1)")
78 tab
79
80 ## valid      wd1      wd4
81 ## hat_v(b0) 0.0007925223 0.001135910 0.0003691558
82 ## hat_v(b1) 0.0010264269 0.001884203 0.0005533111
83 ## MSE(b0)   0.0008346242 1.097682966 1.0386770648
84 ## MSE(b1)   0.0011584559 0.001970838 0.0014793844

```

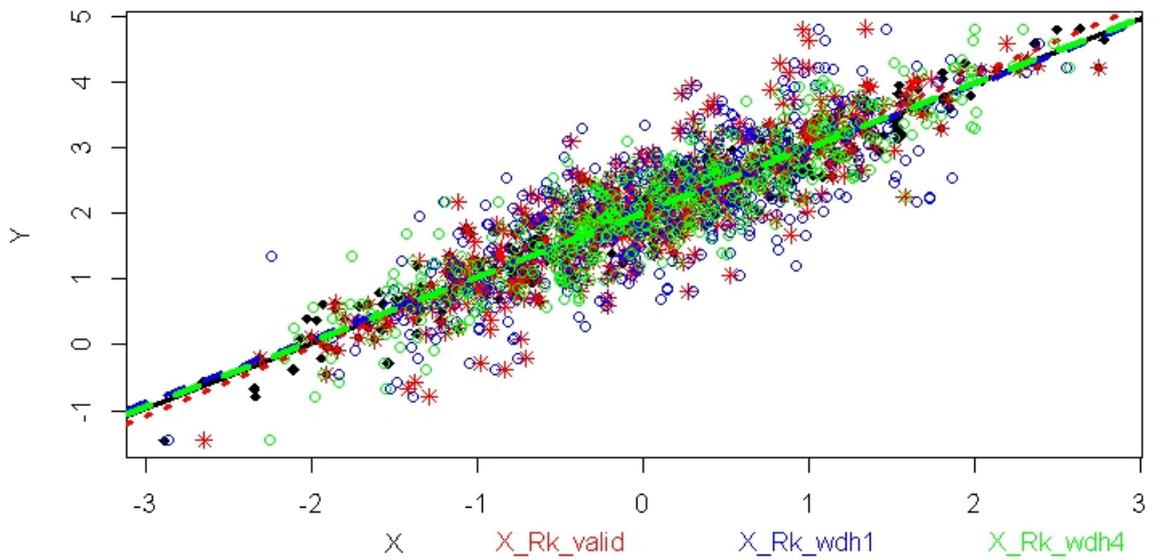


Abbildung 6: *Schwarze durchgezogene Linie* ist die wahre Regressionsgerade. Die restlichen Geraden entstammen aus der Schätzung nach Anwendung der Regressionskalibrierung. Die *rot gepunktete Linie* entspricht der Schätzung auf Validierungsdaten. Die *blau gestrichelte und gepunktete Linie* stellt die Schätzung auf Wiederholungsdaten mit einer Messwiederholung dar und die *grün gestrichelte Linie* auf Wiederholungsdaten mit vier Messwiederholungen.

Literatur

- Helmholtz 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. Carroll, D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu. *Measurement Error in Nonlinear Models- A Modern Perspective*. Chapman & Hall/CRC-Taylor & Francis Group, Boca Raton, 2006.
- P. Gustafson. *Measurement Error and Missclassification 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.
- H. Küchenhoff. Foliensatz (2014). Missclassification and Measurement Error in Regression Models Part 3.
- H. Schneeweiß and H. J. Mittag. *Lineare Modelle mit fehlerbehaftete Daten*. Physica-Verlag Heidelberg Wien, Wien, 1986.