Simulation Extrapolation
Seminarbericht

Andreas Hölzl
Betreuung : Eva Endres, Prof. Dr. Thomas Augustin

15. März 2015

Inhaltsverzeichnis

Kapitel 1

Einleitung

Das Vorliegen von Messfehlern in der unabhängigen Variable ist in der statistischen Anwendung oft ein Problem, da es zu verzerrten Parameterschätzerun führen kann. Es gibt allerdings Methoden, deren Ziel es ist, diese Verzerrung zu beheben und somit eine unverzerrte Parameterschätzung auch bei Vorliegen von Messfehlern in der unabhängigen Variable zu ermöglichen. Neben der Regressionskalibrierung ist dabei die SIMEX-Methode eine der bekanntesten und am meisten verwendeten Methoden zur Messfehlerbehebung. Der Begriff Messfehler bezieht sich im Folgenden ohne weitere Angaben immer auf stochastische, additive und klassische Messfehler, wie sie im Vortrag Frau Marshalava klassifiziert worden sind. Die SIMEX-methode zeichnet sich vor Allem durch Ihre Flexibilität aus, sie kann ohne viel Mehraufwand auf eine Vielzahl von Modellen angewandt werden.

In Kapitel 1 wird die SIMEX Methode allgemein erklärt, auf eine zuerst eher intuitive Erklärung folgt die formale Niederschrift der Methode. Zwei kurze Unterkapitel zur Theorie hinter SIMEX und Simulationsstudien runden das erste Kapitel ab. Das zweite Kapitel beschäftigt sich damit, wie der Messfehler bestimmt werden kann, damit die SIMEX-methode durchgeführt werden kann. Im dritten Kapitel wird ein SIMEX für eine spezielle Modellart, nämlich semiparametrischer Regression, vorgestellt. Im Schlusskapitel werden die wichtigsten Ergebnisse noch einmal zusammengefasst und ein Ausblick auf weitere Forschungsmöglichkeiten gegeben.

Kapitel 2

SIMEX

2.1 Erklärung von SIMEX

Im Folgenden wird versucht, die Idee hinter SIMEX zuerst intuitiv zu erklären, bevor das ganze Vorgehen im nächsten Kapitel formalisiert wird. Betrachten wir ein lineares Regressionsmodell

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

wobei ϵ_i normalverteilt sei. Ein derartiges Modell ist in der Grafik 2.1 visualisiert. Wie man sehen kann, sind die theoretische Regressionsgerade und die aus den Daten gefundene Regressionsgerade sehr ähnlich.

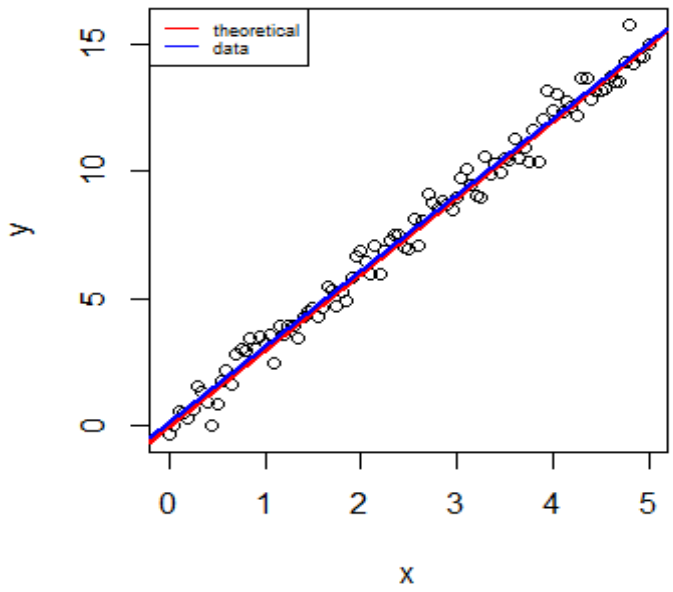

Abbildung 2.1: Schätzung der Regressionsgleichung bei Fehlern in der abhängigen Variable

Bei Vorliegen eines additiven Messfehlers wird die unabhängige Variable nicht genau gemessen. Unter der Annahme einer Normalverteilung für den Messfehler und einer festen Standardabweichung für diesen kann dieser Messfehler simuliert werden, was in Abbildung 2.2 abgebildet ist. Im Weiteren werde davon ausgegangen, dass die Messfehlerstandardabweichung σ_x dieser Normalverteilung bekannt sei. In einem weiteren Kapitel wird darauf eingegangen, wann und wie diese Messfehlerstandardabweichungen bestimmt werden können.

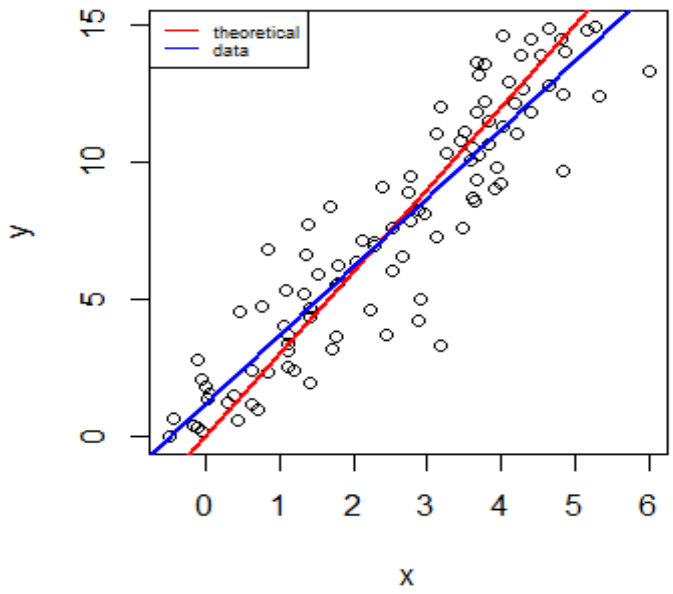

Abbildung 2.2: Schätzung der Regressionsgleichung bei Fehlern in der unabhängigen Variable

Das Vorliegen des Messfehlers in der unabhängigen Variable führt zu einem Bias der Regressionsschätzung, was im Vortrag von Frau Marshalava auch theoretisch hergeleitet wurde. Es gilt dabei

$$\beta_{1,est} = \frac{\beta_{1,true}}{1 + \frac{\sigma_{ges}}{\sigma_x}}$$

, der Parameter β_1 wird also systematisch unterschätzt. Eine Fehlerkorrektur im linearen Modell wäre also direkt analytisch machbar, allerdings wollen wir hier nur das lineare Modell als Beispiel nehmen um ein Verfahren herzuleiten, dass auch für viel allgemeinere Modelle angewendet werden kann.

Wir wollen im Folgenden betrachten, wie sich das Modell verändert, wenn man einen Messfehler zu der unabhängigen Variable hinzufügt. Hierfür fügen wir zu den bereits mit Messfehlern gemessenen Daten einen weiteren aus der gleichen Verteilung wie der ursprüngliche Messfehler stammenden

Messfehler hinzu. Die Standardabweichung dieses neuen Messfehlers entspricht dabei $\lambda_1\sigma_x$, wobei λ_1 zunächst frei wählbar ist, hier beispielsweise gleich 1.

Da der Messfehler $\lambda_1\sigma_x$ zu den bereits mit Messfehler σ_x gemessen Daten hinzugefügt wurde, gilt für den gesamten Messfehler der so erzeugten Daten $X^*(\lambda_1) : \sigma^2_{x*(\lambda_1)} = \sigma^2_x * (1 + \lambda_1)$.

In der Grafik 2.3 sind in schwarz die mit Messfehler gemessenden Punkte und die dazu gefittete Regressionsgerade abgebildet und in blau die Punkte, die entstehen, wenn ein weiterer Messfehler $\lambda_1\sigma_x$ zu diesen Punkten hinzugefügt wird. Die blaue Regressionsgerade zeigt das dazu gefittete Regressionsmodell an.

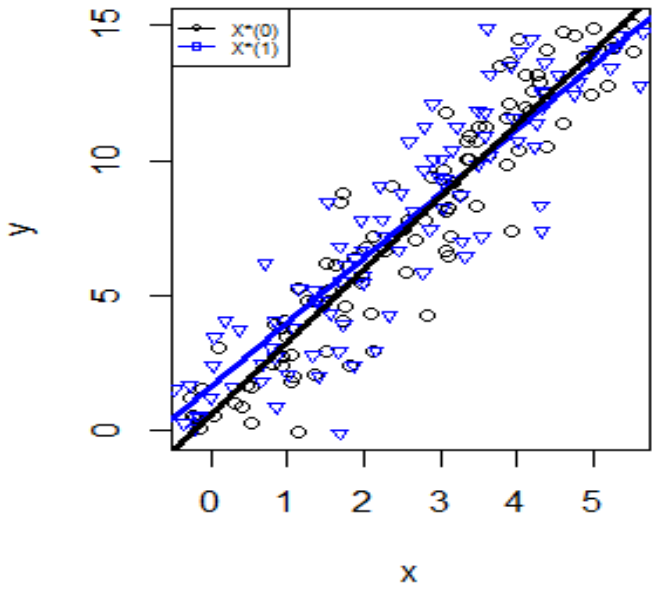

Abbildung 2.3: Hinzufügen eines zusätzlichen Messfehlers

Dies wird noch einmal gemacht, dieses Mal aber mit einer anderen Standardabweichung $\lambda_2 = 2$ und es wird wieder beobachtet, wie sich das lineare

Modell verändert. In Grafik 2.3 sind die roten Punkte die Punkte, die entstehen, wenn der Messfehler $\lambda_2\sigma_x$ zusätzlich hinzugefügt wird.

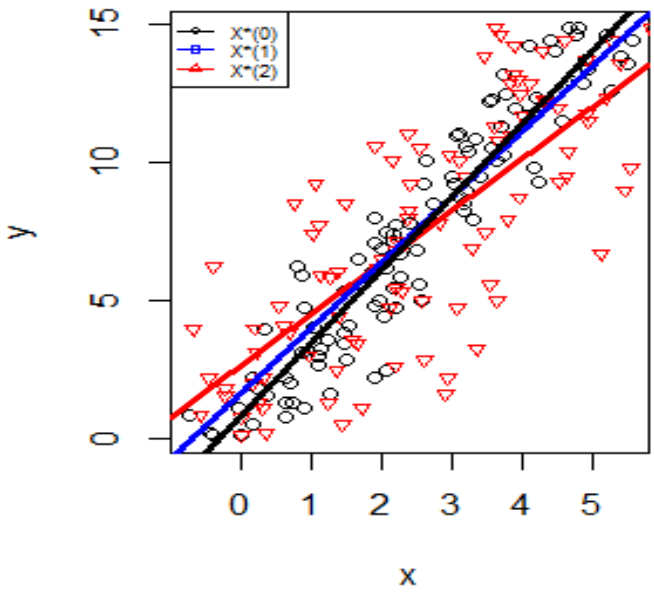

Abbildung 2.4: Hinzufügen eines weiteren zusätzlichen Messfehlers

Visuell wird schnell klar, dass der Intercept β_0 immer größer wird, während die Steigung β_1 immer kleiner wird. Man kann sich für alle dabei herauskommenden Parameter in einer Tabelle eintragen. Hier ergeben sich beispielsweise die folgenden Werte:

Parameter/λ	0	1	2
Intercept	0.9976851	1.969627	3.077275
Steigung	2.6683550	2.394560	1.859049

Man betrachtet die sich ergebenden Werte dann in einem Plot, um zu betrachten, welcher Zusammenhang zwischen λ und den Parametern vorliegen könnte. In 2.5 ist der Plot für den Parameter β_1 aufgeführt.

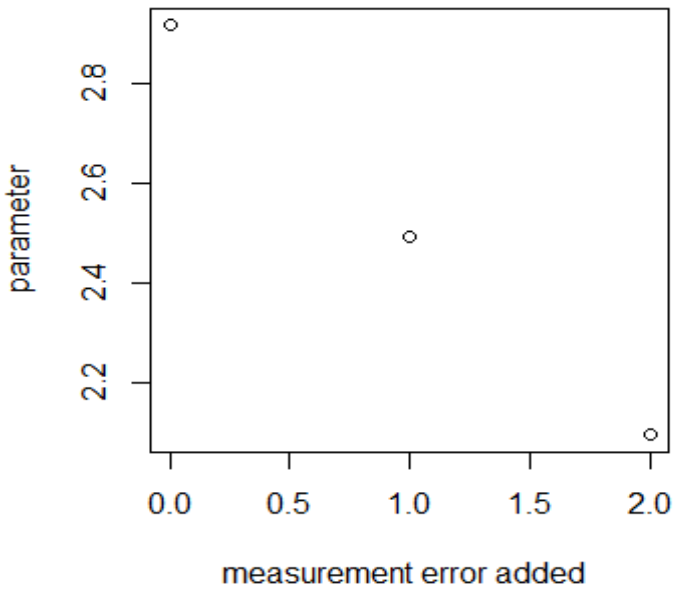

Abbildung 2.5: Parameterschätzer in Abhängigkeit von λ

Es scheint, dass hier ein lineares Modell gefittet werden kann, und genau das machen wir nun auch, um den Parameterschätzer für $\lambda = -1$ abzuschätzen, was dem Vorliegen von gar keinem Messfehler entspricht. Der Parameterschätzer für $\lambda = -1$ entspricht dabei nämlich laut unserer Interpretation von λ, dass ein der Messfehler $-\sigma_x$ zum ursprünglichen Messfehler σ_x hinzugefügt wird, was dem Messfehler 0, also dem Fehlen eines Messfehlers entspricht.

Genau dies ist die Idee hinter SIMEX. Man kann Messfehler nur hinzufügen und nicht wieder wegnehmen, aber man kann sich anschauen, wie das Hinzufügen von Messfehlern die Parameter verändert und über diesen Zusammenhang dann ein Modell erstellen, das dann vorhersagen kann, was passieren würde, wenn kein Messfehler vorliegen würde.

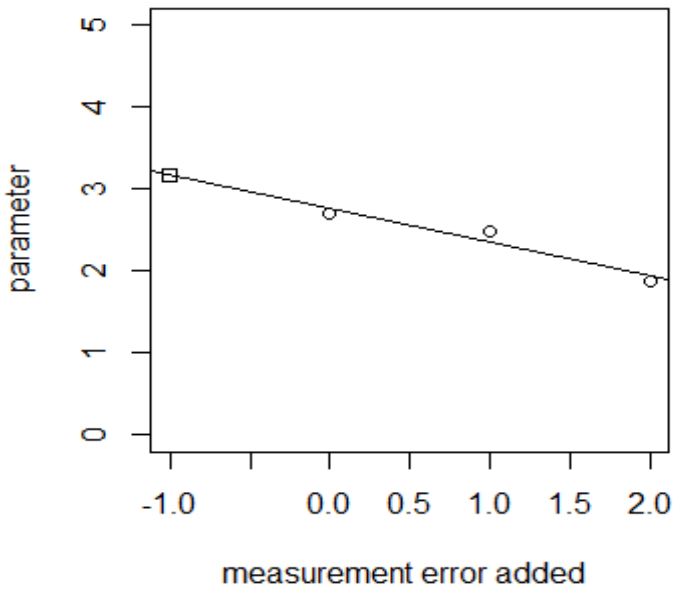

Abbildung 2.6: Schätzung des Parameters bei Vorliegen von keinem Messfehler

In der SIMEX-Funktion des R-Packages zu SIMEX wird das Ganze dann auch ähnlich visualisiert:

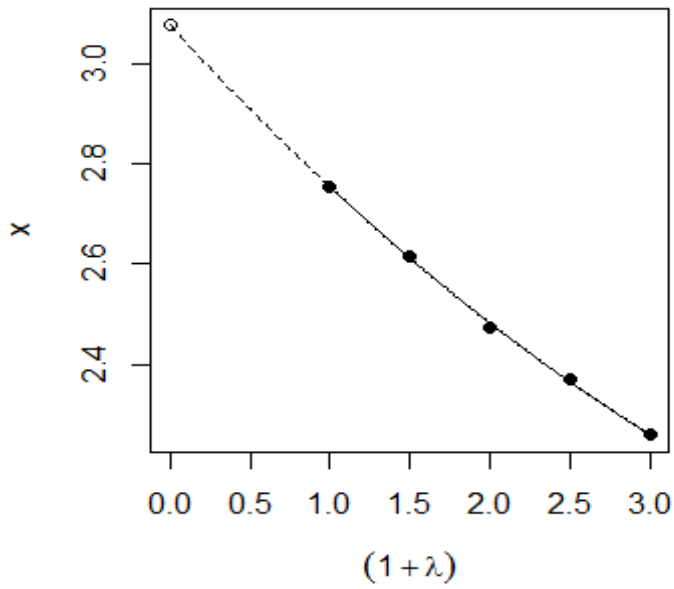

Abbildung 2.7: Schätzung des Parameters bei Vorliegen von keinem Messfehler im SIMEX-package

In diesem Kapitel wurde also die Idee für SIMEX intuitiv an einem kleinen Beispiel erläutert. Es gibt offensichtlich noch Verbesserungsschläge. So ist die Schätzung der Parameter eines Modells mit dazuaddiertem Messfehler $\lambda\sigma_x$ mit Varianz verbunden, weil der gefundene Parameter verschieden ist, je nachdem wie die Ausprägung der Messfehler tatsächlich gezogen wurde. Dieser Effekt kann aber dadurch verkleinert werden, dass man mehrere Ziehungen durchführt und über die dann jeweils gefundenen Parameter mittelt.

2.2 SIMEX-Algorithmus

Simulationsschritt

1. Simuliere Pseudo-daten $X_i^*(\lambda) = X_i^* + \sqrt{\lambda} U_i$ mit U_i aus angenommener Messfehlerverteilung gezogen
2. Mache dies B mal
3. Berechne Mittelwertschätzer für alle λ : $\hat{\beta}(\lambda) = \frac{1}{B} \sum_0^B \hat{\beta}_{naive}(Y_i, X_{b,i}(\lambda))$

Extrapolationsschritt

1. Finde ein Modell, dass den Zusammenhang zwischen $(\lambda_k, \hat{\beta}_k(\lambda_k))$ modelliert
2. Mache die Vorhersage dieses Modells für $\hat{\beta}(-1)$

Häufige Modelle sind dabei das lineare Modell und das quadratische Modell. Es gibt hier aber keine klaren Vorschriften, man sollte nur ein Modell finden, dass die Daten möglichst gut beschreibt.

SIMEX kann also angwendet werden, sobald es möglich ist, zu den bekannten Datenpunkten eine Ziehung aus der angenommenen Messfehlerverteilung hinzuzufügen. Im Vortrag von Frau Marshalava wurden also weitere Messfehlerarten auch multiplikative Messfehler vorgstellt, die auch mit SIMEX behandelt werden können, da auch das Hinzufügen von multiplikativen Messfehlern simuliert werden kann. Auch Fehler in der abhängigen Variable sind, was im Vortrag von Frau Markovic behandelt wurde, können simuliert und somit mit SIMEX behoben werden.

Auch die von Frau Le vorgestelle Methode der Regressionskalibrierung ist sehr flexibel und kann auf eine Vielzahl von Modellen angewendet werden. Simex ist rechenaufwändiger als Regressionskalibrierung und funktioniert auf einigen Modellen auch nicht, dazu mehr im Kapitel zu „Simulationsstudien zu SIMEX“. Wenn dieses Modell allerdings für SIMEX geeignet ist, kann SIMEX den Parameter unter Umständen besser besimmen als die Regressionskalibrierung.

2.3 Theorie zu SIMEX

SIMEX wurde zunächst von Cook und Stefanksi in (Cook&Stefanksi, 1994) vorgestellt, zunächst ohne formalen Beweis, warum die Methode funktioniert. In (Carroll&Küchenhoff et al. 1996) wird dann aber der Beweis erbracht, dass die SIMEX Methode unter sehr allgemeinen Bedingungen, nämlich dass Schätzgleichungen vorliegen, den messfehlerfreien Parameter asymptotisch richtig schätzt. Diese Bedingung bedeutet, dass es zu jedem möglichen hinzuaddierten Fehler mit der gleichen Gleichung die Parameter bestimmt werden können. Im linearen Modell ist das beispielsweise die Schätzgleichung

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

Somit können auch viele andere Modelle, beispielsweise auch generallisierte lineare Modelle mit SIMEX behandelt werden.

2.4 Simulationsstudien zu SIMEX

Auch praktisch wurde SIMEX in verschiedenen Modellen mit Hilfe von Simulationsstudien getestet. In (Shang, 2012) wurde SIMEX beispielsweise zur Messfehlerbehebung in der Quantilsregression verwendet und der Autor kommt zum Schluss „A simulation study is conducted to demonstrate the performance of the SIMEX method in reducing bias and mean squared error in quantile regression with a mismeasured predictor“und „the SIMEX method corrects such biases and closely reproduces conditional distributions of current test scores given past true scores“. Zu einem ähnlichen Schluss kommt der Artikel (Hu&Liang, 2012), bei dem die Anwendung von SIMEX auf longitudinale multi-state Modelle untersucht wird: „We present theoretical justification of the estimation procedure along with a simulation study to demonstrate finite sample performance“.

Aber SIMEX funktioniert nicht für alle Modelle. So kommen Bo Hu, Liang Li1 et al in (Mallick&Fung, 2002) zum Schluss „SIMEX, on the other hand, failed to adequately adjust for the effects of random measurement error in the Cox model, even in the presence of a moderate degree of measurement error.“

SIMEX hat also unter allgemeinen Bedingungen theoretisch bewiesene asymptotische Eigenschaften und es wurde auch in Simulationsstudien gezeigt, dass es für eine Vielzahl von Modellen gültig ist. Trotzdem wird es

im Vergleich dazu, wie oft es sinnvoll wäre, es zu verwenden, realtiv selten verwendet. Der Hauptgrund liegt darin, dass man für die SIMEX-Methode zuerst den Messfehler σ_x wissen muss. Diesen haben wir bisher als bekannt angenommen, er ist es aber im Normalfalls nicht. Das nächste Kapitel beschäftigt sich also damit, wie $\hat{\sigma}_x$ geschätzt werden kann.

Kapitel 3

Bestimmung des Messfehlers

Hier werden drei Möglichkeiten aufgeführt, wie der Messfehler bestimmt werden kann.

3.1 Kalibrierung

Wenn der wahre Wert von bestimmten Messungen bekannt ist, kann dies dazu verwendet werden, um die Messfehlerverteilung zu schätzen. Die Messung zu diesem Messwert wird dann mehrfach gemacht und die Verteilung der Messpunkte dann parametrisch geschätzt, beispielswiese durch eine Normalverteilung mit bestimmter Standardabweichung. Ein Beispiel wäre ein Gerät, dass den ph-Wert misst. Man weiß, dass steriles Wasser bei 25 Grad einen ph-Wert von 7 hat. Man misst also steriles Waser bei 25 Grad mehrfach mit dem Testgerät und kann dann die Verteilung schätzen, aus der die Messfehler hinzukommen.

3.2 Messwiederholung

Ist kein wahrer Wert gegeben, aber es werden mehrere Messungen gemacht, so kann unter der Annahme, dass kein Bias vorliegt, auch die Verteilung der Messung bestimmt werden. Hierfür wird angenommen, dass der Mittelwert der mehrfachen Messungen der wahre Wert ist und dann kann man wie im Kapitel zu Kalibrierung verfahren.

In (Devanarayana&Stefanski, 2002) wurde eine Methode vorgestellt, mit der im linearen Modell bei mehreren Messungen direkt SIMEX angewednet werden kann, ohne vorher die Messfehlerverteilung explizit zu schätzen.

Man hat also nicht nur einen mit Messfehlern behafteten Messwert X_i^* für die Messung mit dem Index i vorliegen, sondern mehrere mit gleichen Messfehler behaftete Messungen ${}^{X}*_{ij}$.

Es sollen neue Pseudodaten generiert werden, für die, damit SIMEX ausgeführt werden kann, gilt:

$$Var(X^*) = \sigma_x^2 * (1 + \lambda) \tag{3.1}$$

Dies soll gemacht werden, in dem Linearkombinationen der gegebenen Messwiederholungen X_{ij}^* gebildet werden.

$$X_i^*(\lambda) = \sum_{j=1}^{m_i} A_i X_{ij}^*$$

m_i beschreibt dabei die Anzahl der Messwiederholungen, die für Messung i vorliegen.

Wenn dabei für A_i dabei noch die zwei Bedinungen erfüllt sind

$$\sum_{1}^{m} A_i = 1$$

und

$$\sum_{1}^{m} A_i^2 = \frac{1+\lambda}{m}$$

so gilt 3.1 und Pseudodaten mit zusätzlichem Messfehler λ können mit diesem Verfahren erzeugt werden.

Der Beweis findet sich in (Devanarayana&Stefanski, 2002) und dort ist auch aufgeführt, wie die A_i computational effizient mit Normalverteilungen simuliert werden können.

3.3 Modellbasiert

Für bestimmte Modelle kann auch implizit ein Messfehler mitgeschätzt werden. Dies ist beispielsweise in der Faktoranalyse der Fall. Am Beispiel einer psychologischen Untersuchung soll die einfaktorielle Faktoranalyse hier kurz erklärt werden. Es soll in einer psychologischen Untersuchung untersucht werden, welchen Einfluss die Angst vor Mathematik auf die Note hat. Die Emotion Angst kann aber nicht direkt gemessen werden, sondern nur indirekt über Fragebögen. Jeder der n Testpersonen füllt also m Fragen über

diese Emotion aus. Diese seien so gestellt, dass sie alle Fragen über Angst und mit möglichst wenig Fremdeinfluss von anderen Gegebenheiten sind. Jede Frage wird auf einer Skala von 1 (stimme gar nicht zu) bis 5 (stimme sehr zu) beantwortet. Für das Modell der einfaktoriellen Faktoranalyse gilt dann

$$x_{ij} = t_i + u_{ij}$$

t_i ist dabei der wahre Wert der Angst und x_{ij} der gemessene Wert, u_{ij} beschreibt die Abweichung der zwei Werte für Person i und Item j Im Modell werden nun die t_i und u_{ij} mit Maximum Likelihood unter Annahme einer Normalverteilung der u_{ij} geschätzt.

Als unabhängige Variable für weitere Fragestellungen werde dann der Summenscore betrachtet, also für jede Person wird als Variable die Summe über die Items dieser Person berechnet. Es sei also

$$Y_i^2 = \sum_1^m X_{ij}$$

der Summenscore.

Gesucht ist jetzt die Varianz dieses Summenscores, im folgenden als $\hat{\sigma}_Y^2$ bezeichnet. Diese kann folgendermaßen abgeschätzt werden:

$$\hat{\sigma}_Y^2 = \hat{\sigma}^2_{\sum_1^m X_j} = \sum \hat{\sigma}_(X_j)^2 + 2 * \sum_{l>j}^{K} \hat{\sigma}_{X_l, X_j} = \sum \hat{\sigma}(X_i)^2 + 2 * \sum_{i>j}^{K} \hat{\sigma}_{t+u_l, t+u_j} =$$

$$\sum \hat{\sigma}(X_i)^2 + K * (K-1)\hat{\sigma}_t^2$$

$\hat{\sigma}(X_i)^2$ wird dabei aus den Daten geschätzt und $\hat{\sigma}_t^2$ ist auch bekannt, da t_i für alle Personen im Modell geschätzt wurde.

Die Abschätzung $\hat{\sigma}_{t+\epsilon_i, t+\epsilon_j} = \hat{\sigma}_t^2$ ist dabei eine nicht erwartungstreue Abschätzung.Diese Abschätzung wird in der Psychometrie aber trotzdem sehr oft als wahr angenommen, es basiert beispielsweise die Maßzahl Cronbach's Alpha für die Reliabilität auf dieser Herleitung(Cronbach, L.J. 1951).

Kapitel 4

SIMEX bei semiparametrischer Regression

4.1 Splines

Splines beschreiben Funktionen der Art

$$f(x) = \sum_{1}^{n} \gamma_i B(X)$$

Grundlage der semiparametrischen Regression ist eine Basifunktion B(X) mit einem bestimmten Knotenpunkt. Zu jedem der Werte X kann nun ein Wert berechnet werden, wie groß die Ausprägung dieses Punktes an der Basisfunktion mit dem entsprechenden Knotenpunkt ist. Die Designmatrix $Z = B_j(x_i)$ beschreibt dann die Ausprägung jedes Punkte x_i and der Basisfunktion mit Knoten j und ist Grundlage der Regression. Das Ergebnis der Regression ist dann ein Parameter γ pro Knotenpunkt und um die geschätzte Regressionsgleichung zu betrachten, wird jede Basisfunktion mit diesem Faktor multpliziert und dann alle Basisfunktionen zusammenaddiert. In 4.1 ist dies visualisiert.

Für die Schätzung von γ geht oft auch ein Glättungsparameter λ ein, der verhindern soll, dass die Schätzung sich zu sehr auf die Daten konzentriert und weniger den allgemeinen Trend findet als die speziellen Auspägungen dieser Daten.

Es soll dann

$$(y - Z\gamma)(y - Z\gamma) + \lambda\gamma K\gamma$$

minimert werden.

Dies geht bei Normalverteilungsanname der Residuen über

$$\hat{\gamma} = (Z^T Z + \lambda K)^{-1} Z^T y$$

Es wird also jeder einzelne Regressionsparameter bestimmt. Für jede Z und y kann für einen festen Glättungsparameter λ eindeutig der Gewichtsvektor $\hat{\gamma}$ gefunden, wie in Grafik 4.1 betrachtet werden kann. Die schwarzen Punkte sind die gegebenen Punkte und die blaue Kurve die darauf gefittete Kurve, die sich als Summe der dünner gezeichneten blauen Kurven ergibt.

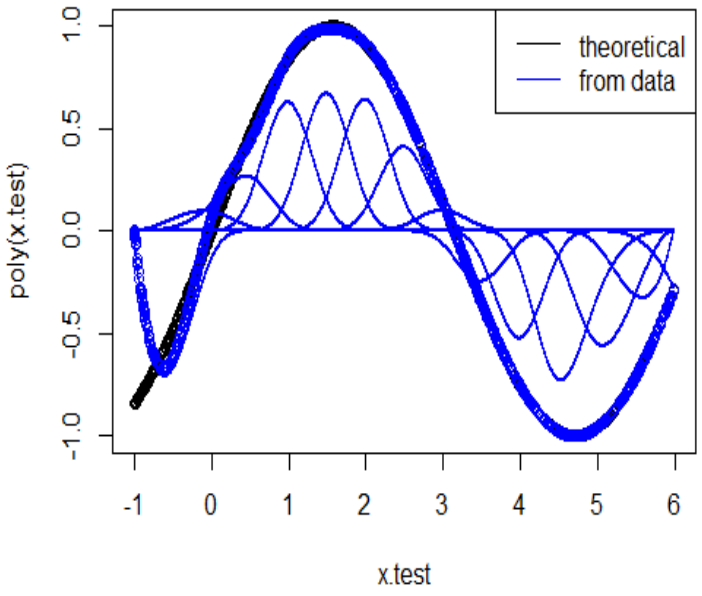

Abbildung 4.1: Visualisierung von Splines

Wir betrachten hier allerdings mit Messfehler gemessene Daten und diese könnten beispielsweise so aussehen wie in Grafik 4.2 visualisiert.

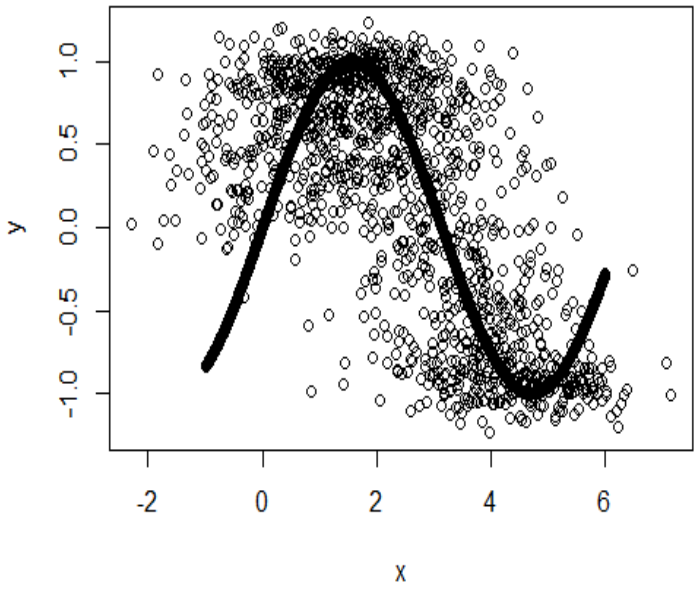

Abbildung 4.2: Hinzufügen eines zusätzlichen Messfehlers

Dabei sieht die geschätzte semiparametrische Funktion beispielsweise folgendermaßen aus:

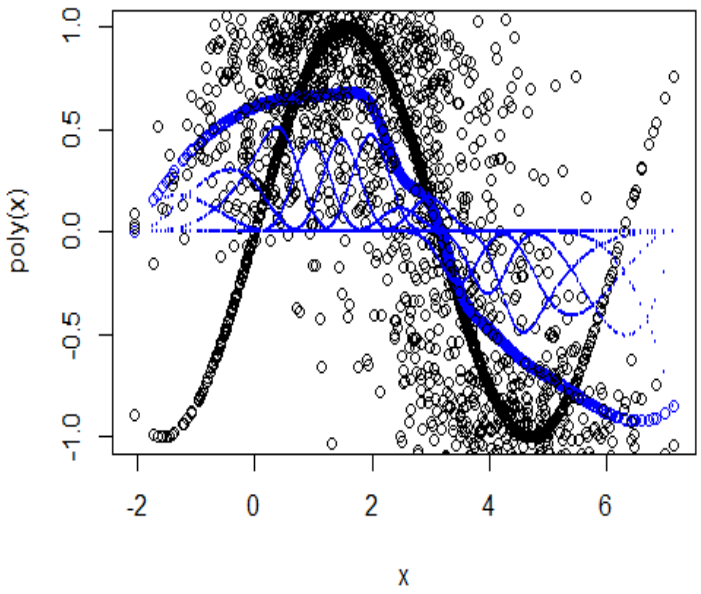

Abbildung 4.3: Splines bei Messfehlern

Wie beobachtet werden kann, wird die wahre Funktion nicht ganz richtig geschätzt, die geschätzte semiparametrische Funktion ist zu flach. Dies versuchen wir nun mit SIMEX zu korrigieren.

Die Idee für SIMEX lautet dabei:
1. Simulationsschritt : Für verschiedene λ füge aus der angenommenen Messfehlerverteilung mit der entsprechenden Standardabweichung Messfehler zur x-Variable hinzu und berechne die Designmatrix Z. Bereche jedes mal den Parametervektor $\hat{\gamma}$ mit

$$\hat{\gamma} = (Z^T Z + \lambda K)^{-1} Z^T y$$

und mittle für jedes einzelne λ die $\hat{\gamma}$ über B Ziehungen.

2. Extrapolationsschritt: Stelle eine Regressionsschätzung für jeden einzelnen Spline auf und extrapoliere damit zu $\lambda = -1$ zurück. Die dabei

herauskommenden Werte sind dann die von SIMEX geschätzten Werte.

Es gibt allerings noch zwei Probleme, die bei normalem SIMEX nicht auftreten:

4.1.1 Wahl der Knoten

Die Knoten müssen, damit SIMEX funktioniert, auf alle Fälle immer gleich gewählt sein. Deshalb wird es aktuell so gemacht, dass die Knoten bei den Daten der naiven Regression mit Messfehlern äquidistant gewählt werden.

4.1.2 Problematik des Glättungsparameters

Es gibt allerdings noch das Problem der Wahl des Glättungsparameters. In der Literatur wird dies für normale P-Splines zumeist durch Kreuzvalidierung gemacht. Das Problem bei SIMEX ist allerdings, dass für verschiedene Daten, also für die verschiedenen λ mit den verschiedenen hinzugefügten Messfehlern, verschiedene Glättungsparameter geschätzt werden würden. Wählt man also den Glättungsparameter als den Glättungsparameter, der bei den gemessenen Daten berechnet wird, so ist dieser Glättungsparameter aber nicht der optimale Glättungsparameter für die Daten ohne Messfehler oder mit mehr Messfehler. Wenn man den Glättungsparametern allerdings jedes mal neu wählt, gelten die in (Carroll&Küchenhoff et al. 1996) gezeigten asymptotischen Eigenschaften nicht mehr. Es gilt dann nämlich nicht mehr die gleiche Schätzgleichung für die verschiedenen Daten, die sich ergeben, wenn man verschiedene Messfehler hinzufügt.

In (Carroll&Ruppert, 2004) wurde beide Möglichkeiten, enweder den Glättungsparameter einmal auf den gegebenen Daten zu berechnen und dann immer zu verwenden oder den Glättungsparameter bei jedem Simulationsschritt neu zu berechnen, beschrieben, allerdings nicht miteinander verglichen. In (Carroll&Maca, 1999) wurde eine andere Methode zur Schätzung des Glättungsparameters für SIMEX vorgeschlagen, die darauf beruht, den MSE für verschiedene Glättungsparameter abzuschätzen.

Zwischen den möglichen Verfahren - Bestimmung des Glättungsparameters einmal fest über Kreuzvalidierung, Glättungsparameter, in jedem Schritt neu bestimmen und Glättungsparameter über das Verfahren in (Carroll&Maca, 1999) abzuschätzen, gab es allerdings noch keine vergleichende Simulationsstudie und für keines dieser Verfahren gibt es Theorie, die asymptotische

Eigenschaften beweisen würde. In dieser Arbeit wurde der Glättungsparameter jedes mal neu über Kreuzvalidierung bestimmt.

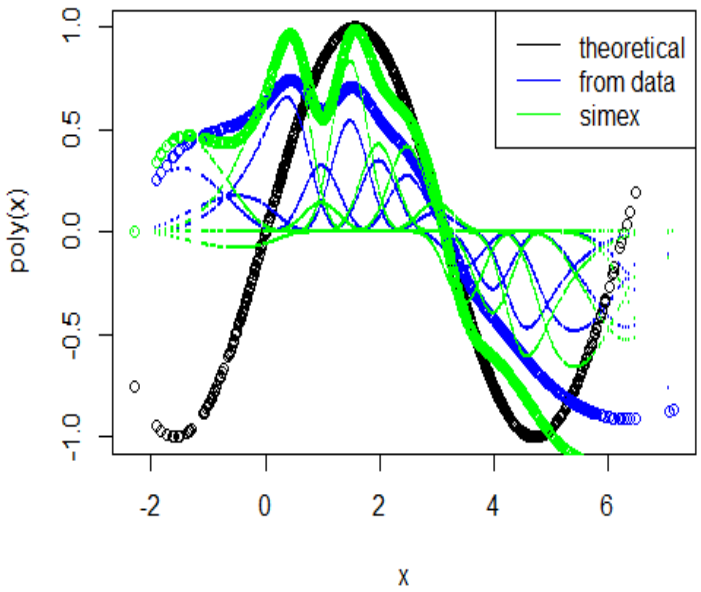

Abbildung 4.4: Splines bei Messfehlern

In 4.4 kann nun das Ergebnis der Anwendung der SIMEX-methode beobachtet werden. Die gefundene semiparametrische Funktion ist näher an der theoretisch angenommenen Funktion, Man muss aber trotzdem vorsichtig sein, da wir in diesem Fall keine aus Theorie hervorgehenden asymptotischen Eigenschaften der SIMEX-methode haben.

Man kann auch deutlich sehen, dass die Schätzung an den Rändern ungenauer wird. Die Abweichung Schätzung der Funktion außerhalb des tatsächlichen Definitionsbereichs von 1 bis 5 ist allerdings wieder unerheblich, da die Daten ohne Messfehler nur im Definitionsbereich simuliert wurden.

4.2 Anwendung auf reale Daten

Nun soll die SIMEX-methode für semiparametrische Daten auch angewendet werden. Hierfür werden Daten verwendet, die den Zusammenhang zwischen Angst vor Mathematik und den Noten in Mathematik untersuchen. Da Emotionen nicht direkt gemessen werden können, werden Items erhoben, auf jedes Item kann von 1 (stimme gar nicht zu) bis 5 (stimme sehr zu) beantwortet werden. Als unabhängige Variable für die nichtparametrische Regression wird dann der Summenscore der Items verwendet. Für die Abschätzung des Messfehlers wird die im Kapitel zu "Mesffehler"vorgeschlagene modellbasierte Abschätzung bei Itemdaten verwendet. Somit kann also die SIMEX-methode durchgeführt werden und es wird der in diesem Kapitel vorgestellte nichtparametrische SIMEX durchgeführt.

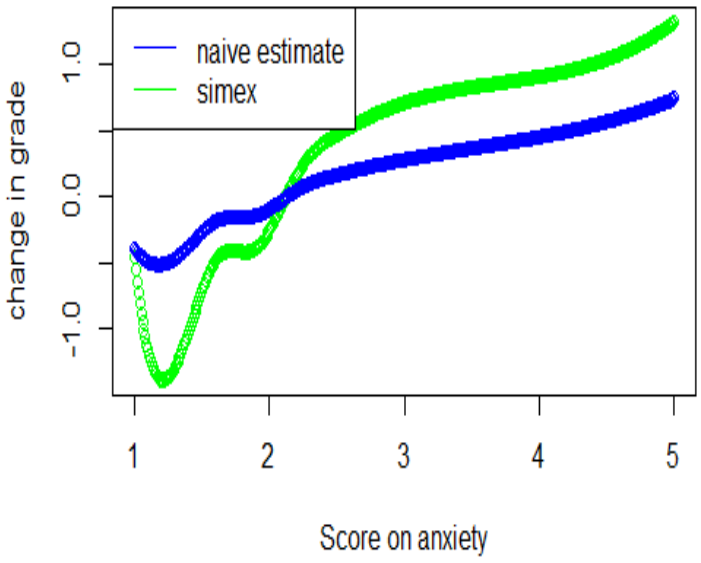

Abbildung 4.5: Anwendung der SIMEX-methode für Splines an echten Daten

In Grafik 4.5 werden die Ergebnisse dargestellt. Es scheint auf den ersten

Blick so zu sein, dass die sonst übliche lineare Modellierung nicht ganz angebracht ist, weil die Steigung zwischen 1 und 2 anders ist als im restlichen Wertebereich. Bei der Interpretation der Ergebnisse sollte man dennoch noch vorsichtig sein, da die vorgestellte Methode noch keine bewiesenen asymptotischen Eigenschaften besitzt.

Kapitel 5

Schluss

Die SIMEX-methode ist also eine sehr flexible Methode zur Behebung von Messfehlern, die sowohl theoretisch untermauert ist als auch in vielen praktischen Simulationsstudien getestet worden ist. Eins der Hauptprobleme dieser Methode ist es allerdings, dass der Messfehler bekannt sein muss, damit sie angewendet werden kann. Dies ist oft nicht der Fall und dann kann keine Messfehlerbehebung durchgeführt werden. Dieses Problem hat aber nicht nur die SIMEX-methode, sondern auch andere Methoden zur Messfehlerbehebung wie beispielsweise die Regressionskalibrierung. Für eine sehr häufige Art von Daten, nämlich Itemdaten, wurde in dieser Arbeit eine Möglichkeit vorgestellst, wie der Messfehler abgeschätzt werden könnte. Auch wurde SIMEX auf semiparametrische P-Spline-Regression angewandt. In Kombination würde diese zwei Erkentnisse die Möglichkeit bieten, latente Variablen Modelle nichtparametrisch modellieren zu können. Dies könnte eine Alternative sein zu den bisher verwendeten computational sehr aufwändigen nichtparametrischen latente Variablen Modellen (Fahrmeir, L.; Raach, A., 2006) Dafür sind allerdings noch weitere Nachforschungen und Simulationsstudien nötig.

Literatur

[Cook&Stefanksi, 1994] Cook, J.R. and Stefanski, L.A. (1994) Simulation-extrapolation estimation in parametric measurement error models. Journal of American Statistical Association, 89, 1314 - 1328

[Lederer&Küchenhoff,2006] Lederer, W. and Küchenhoff, H. (2006) A short introduction to the SIMEX and MCSIMEX. RNews, 6(4), 26 - 31

[Carroll&Küchenhoff et al. 1996] Carroll R.,Küchenhoff, H. et al. Asymptotics for the SIMEX estimatorin Nonlinear Measurement Error Model, Journal of the American Statistical Association; Mar 1996; 91, 433; ABI/INFORM Global pg. 242

[Shang, 2012] Shang Y. (2012), Measurement Error Adjustment Using the SIMEX Method: An Application to Student Growth Percentiles, Journal of Educational Measurement Winter 2012, Vol. 49, No. 4, pp. 446–465

[Hu&Liang, 2012] Hu B, Liang L. (2012), Nonparametric Multi-state Representations of Survival and Longitudinal Data with Measurement Error, Stat Med. 2012 September 20; 31(21)

[Mallick&Fung, 2002] Mallick R, Fung K (2012) Adjusting for measurement error in the Cox proportional hazards regression model., J Cancer Epidemiol Prev. 2002;7(4):155-64

[Devanarayana&Stefanski, 2002] Devanarayana V.; Stefanski L.(2002) Empirical simulation extrapolation for measurement error models with replicate measurements, Statistics and Probability Letters 59 (2002) 219–225

[Carroll&Ruppert, 2004] Carroll R.; Ruppert D.; Crainiceanu C..(2004), Nonlinear and Nonparametric Regression and Instrumental Variables,

Journal of the American Statistical Association September 2004, Vol. 99, No. 467

[Carroll&Maca, 1999] Carroll R.; Maca C.; Ruppert D. (1999), Nonparametric regression in the presence of measurement error , Biometrika 86,3 pp. 541 - 554

[Fahrmeir, L.; Raach, A., 2006] Fahrmeir, Ludwig; Raach, Alexander (2006) : A Bayesian semiparametric latent variable model for mixed responses, Discussion paper // Sonderforschungsbereich 386 der Ludwig-Maximilians-Universität München, No. 471, http://nbn-resolving.de/urn:nbn:de:bvb:19-epub-1839-8