Überblick über Messfehler und ihre Auswirkungen in der linearen Regression

Seminararbeit

Seminar "Statistische Herausforderungen im Umgang mit fehlenden bzw. fehlerbehafteten Daten"

Institut für Statistik

Autor: Hanna Marshalava

Leitung: Prof. Dr. Augustin
Eva Endres

Inhaltsverzeichnis

1 Einleitung

Ungenauigkeiten sind bei Messungen jeder Art nicht zu vermeiden. Dabei ist es wichtig denkbare Fehlerart und deren Größe zu kennen, um für jeden Messergebnis entsprechenden Loyalität angeben zu können. Das ist insbesondere dann unerlässlich, wenn die gesuchten Größen nicht inspiziert werden können.

Nimmt man eine Messung vor, z.B. physikalische oder chemische, so ist diese niemals exakt. Das bemerkt man, wenn die Messung wiederholt wird oder wenn verschiedene Personen das Gleiche messen. Es treten dann Schwankungen oder Ungenauigkeiten in den erhaltenen Messwerten auf. (Höpcke (1980), s. 41)

Zur Verdeutlichung der Problematik betrachten wir die Messergebnisse der Astronomischen Einheit - AE (große Halbachse der Erdbahnellipse = mittlerer Abstand Sonne-Erde):

Tabelle 1: Messergebnisse der Astronomischen Einheit (J. Hartung (2009), s. 321)

Ort bzw. Durchführende und Jahr der Messung	AE in Millionen Meilen	Schätzung der Schwankung durch Experimentator
Necomb, 1895	93.28	93.20 - 93.35
Hinks, 1901	92.83	92.79 - 92.87
Noteboom, 1921	92.91	92.90 - 92.92
Spencer Jones, 1928	92.87	92.82 - 92.91
Spencer Jones, 1931	93.99	92.99 - 93.01
Witt, 1933	92.91	92.90 - 92.92
Adams, 1941	92.84	92.77 - 92.92
Brouwer, 1950	92.977	92.945 - 93.008
Rabe, 1950	92.9148	92.9107 - 92.9190
Millstone Holl, 1958	92.874	92.873 - 92.875
Jodrell Bank, 1959	92.876	92.871 - 92.882
S.T.L., 1960	92.9251	92.9166 - 92.9335
Jodrell Bank, 1961	92.960	92.958 - 92.962
Cal.Tech, 1961	92.956	92.955 - 92.957
Soviets, 1961	92.813	92.819 - 92.816

Man weißt, dass die Astronomische Einheit AE in Beobachtungszeitraum konstant ist, so dass die unterschiedlichen Ergebnisse nur durch die Messfehler entstanden sein können.

Die graphische Visualisierung der Ergebnisse zeigt die Schwankungen in den Messwerten: Messfehler sind jeweils einer einzelnen Beobachtung zugeordnet. Sie werden unterschieden und nach ihren Ursachen oder charakteristischen Eigenschaften nähe bezeichnet.

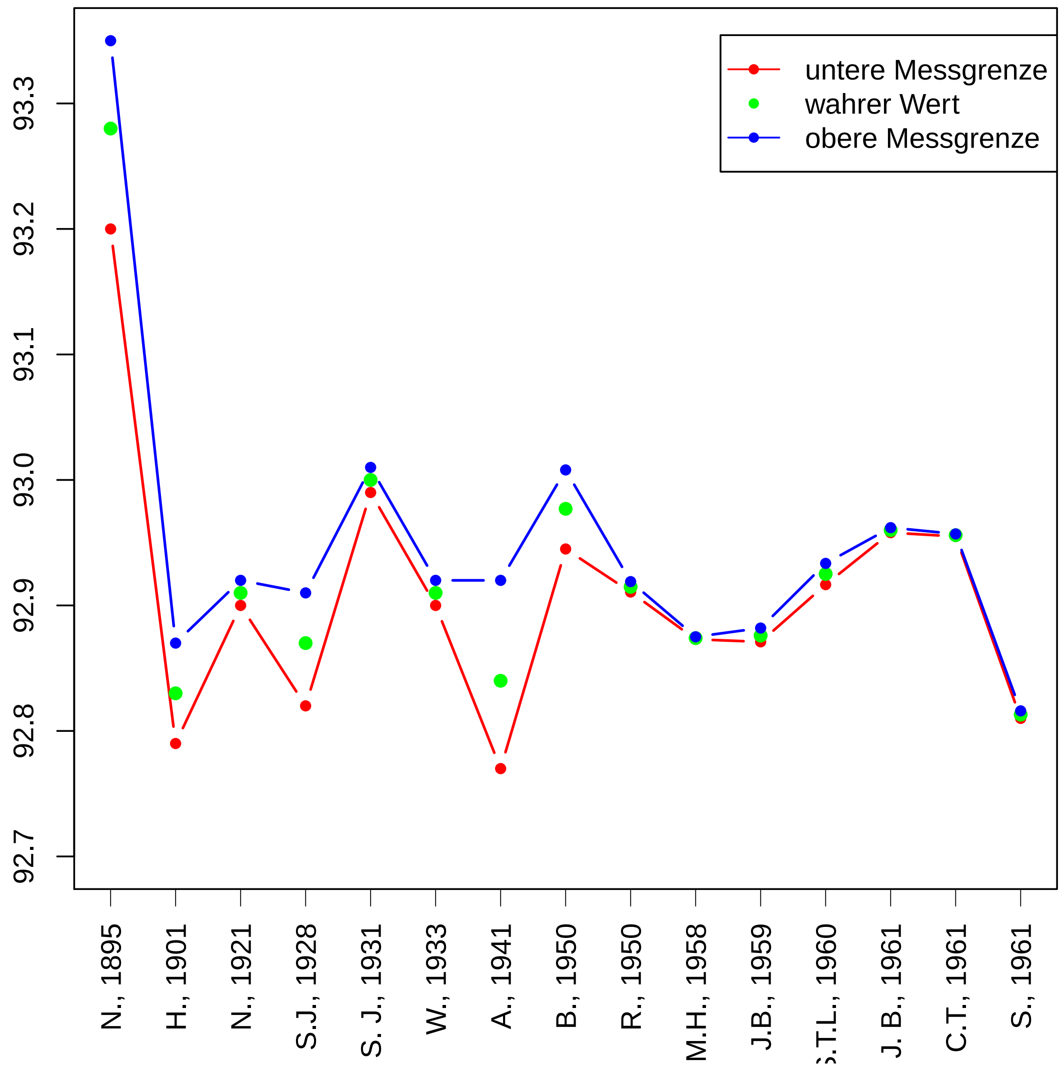

Abbildung 1: Übersicht der Messungen von der Astronomischen Einheit über die Jahre

Messfehler in Variablen haben folgende Auswirkungen:

- sie verursachen Abweichungen der Parameterschätzer in den linearen Modellen
- sie führen zu dem (manchmal hochgradigen) Potentialverlust (loss of power) bei der Erfassung der Zusammenhänge zwischen Variablen
- sie verdecken die Dateneigenschaften, was die graphische Darstellung der Datenanalyse erschwert.

Die ersten zwei Punkte werden "doppelseitiger Schlag"(double whammy) des Messfehlers genannt. Wenn man den dritten Punkt hinzunimmt, werden sie "dreifacher Whammy"genannt.(R.J. Carrol and C.Crainiceanu (2006), s. 1)

In der Literatur werden meistens sogenannte klassische Messfehler beschrieben. Dabei wird der wahrer Wert mit additivem Fehler und meistens konstanten Varianz gemessen.

2 Übersicht der Messfehler

2.1 Grundlagen Messfehlertheorie

Bei der Informationsgewinnung kommt es darauf an, dass die bestimmten Größen genau erfasst und durch einen Zahlenwert oder eine Grafik dargestellt werden können. Das Messergebnis ist der Schätzwert der Messgröße, der durch Auswertung einer Messung erhalten wird. Es stellt sich dabei die Frage nach der Vertrauenswürdigkeit derartiger Zahlenwerte, d.h. Messunsicherheit und Messungenauigkeit. Die Unsicherheit bzw. Ungenauigkeit des Messwertes wird von Fehlern negativ beeinflusst. Man kann dabei Fehler nach der Art ihres Auftretens in systematische und zufällige (statistische) Fehler, aber auch nach der Art ihrer Entstehung in subjektive und objektive Fehler einteilen.(Gräber (2009), Kapitel 6.1, s. 2)

Folgende Abbildung liefert eine Übersicht der 'natürlichen' statistischen Fehlern:

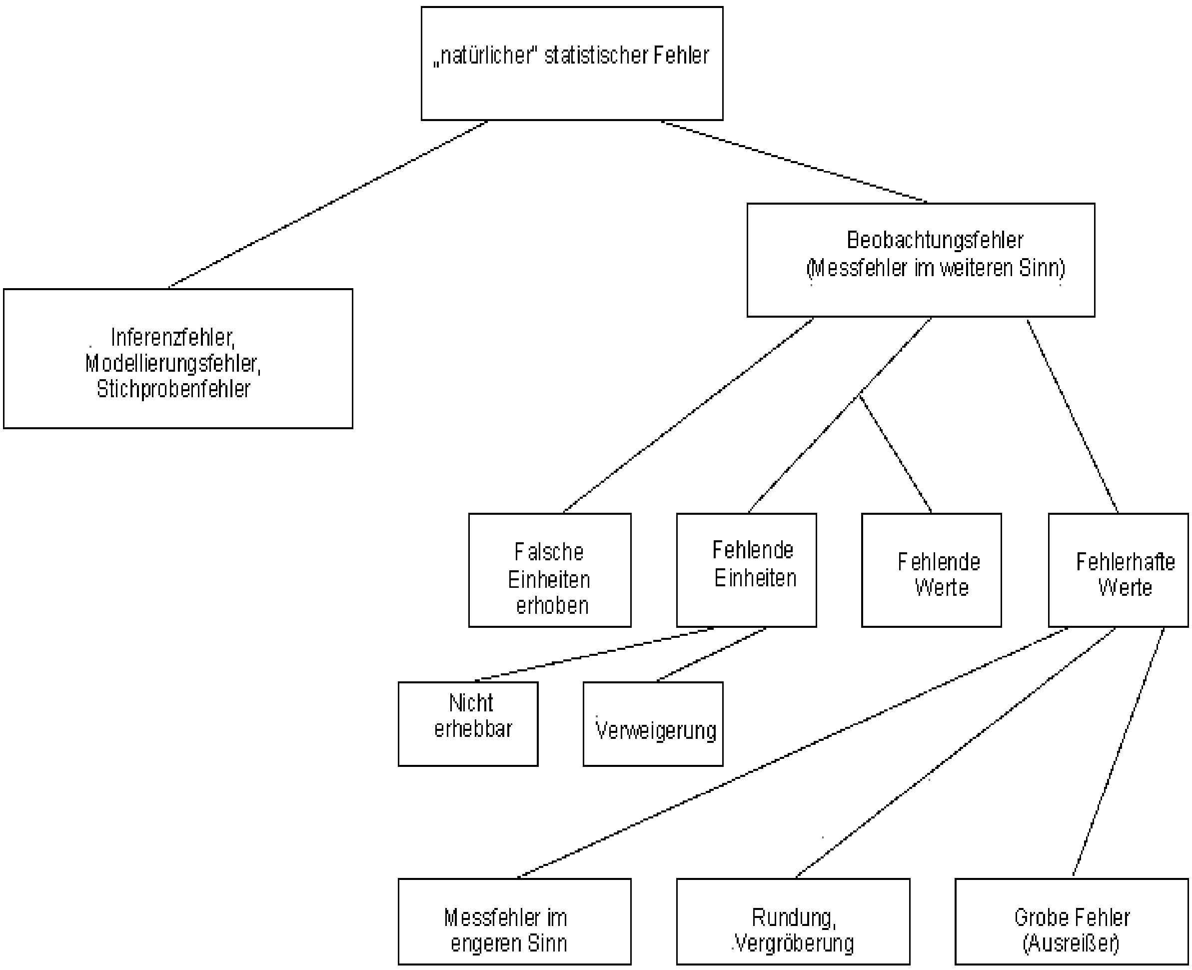

Abbildung 2: Arte der statistischen Fehlern T. Augustin (2013)

Ursachen für einen systematischen Messfehler können personenspezifisch, itemspezifisch, itemkontextbezogen oder erhebungskontextbezogen sein. Beispielsweise sind personenspezifische Effekte insbesondere dann zu erwarten, wenn die unabhängigen und die abhängigen Variablen bei denselben Probanden gemessen werden.

Bei jeder Untersuchung stellt sich die Frage nach der Qualität des Messvorgangs, die den Untersuchungserfolg und die Aussagefähigkeit der Ergebnisse entscheidend beeinflusst. Messfehler sind bei jedem Messvorgang, z.B. bei Persönlichkeitstests in der Psychologie oder Einstellungsfragebögen im Marketing, unvermeidbar. Daher muss es bei Messungen das Ziel sein, die Messqualität zu beurteilen und Messfehler zu minimieren. Traditionell werden in diesem Zusammenhang in der klassischen Testtheorie Haupt- und Nebengütekriterien betrachtet.

Der Schwerpunkt des Modells der klassischen Testtheorie liegt auf der Genauigkeit einer Messung bzw. auf der Größe des jeweiligen Messfehlers. Daher wird sie oft auch als Messfehlertheorie bezeichnet.

Es werden folgende Hauptgütekriterien eines Messinstruments unterschieden:

1. **Objektivität**
   Objektive Messergebnisse liegen vor, wenn verschiedene Personen, die die Messungen unabhängig voneinander vornehmen, zu den gleichen Messergebnissen gelangen. Die Objektivität wird über drei verschiedene Aspekte weiter differenziert: Durchführungsobjektivität ist gegeben, wenn der Untersuchungsleiter die Probanden nicht durch seine eigenen Vorstellungen und sein Untersuchungsziel beeinflusst. Auswertungsobjektivität ist dadurch gekennzeichnet, dass es bei der Auswertung der Messergebnisse keine Freiheitsgrade gibt. Schließlich betrifft die Interpretationsobjektivität den Spielraum bei der Interpretation der Messergebnisse. Interpretationsobjektivität ist dann vorhanden, wenn aus gleichen Ergebnissen gleiche Schlussfolgerungen gezogen werden.

2. **Reliabilität (Zuverlässigkeit)**
   Die Reliabilität betrifft die Zuverlässigkeit und Stabilität eines Messinstruments. Das Kriterium bezieht sich auf die Frage, wie gemessen wird, und fordert, dass die Messergebnisse bei wiederholter Messung reproduzierbar sein sollten.

3. **Validität (Gültigkeit)**
   Die Validität bezieht sich auf die Gültigkeit und materielle Genauigkeit eines Messinstruments. Im Rahmen der Validitätsprüfung ist zu fragen, ob mit einem Messinstrument das gemessen wird, was gemessen werden soll.

Die Reliabilität (Zuverlässigkeit) lässt sich anhand der Zerlegung des beobachteten Messwertes („observed score“, X_O) verdeutlichen : $X_O = X_T + X_S + X_R$. Der beobachtete Messwert setzt sich demnach aus dem „wahren“ Wert („true score“, X_T), einem systematischen Fehler („systematic error“, X_S) und einem zufälligen Fehler zusammen („random error“, X_R). Die Reliabilität bezieht sich auf den zufälligen, unsystematischen Fehler: Bei einer vollkommen reliablen Messung gibt es keine Zufallsfehler ($X_R = 0$). Dieser Zufallsfehler ist ein variabler Fehler, der alle die Einflussfaktoren enthält, die bei jeder Messung die Messergebnisse mit anderer Stärke ohne erkennbare Systematik beeinflussen. Alternativ zu dieser Betrachtungsweise wird in der klassischen Testtheorie die Reliabilität ausgedrückt als Verhältnis der Varianz der wahren Messwerte zur Varianz der beobachteten Messwerte. Da die Varianz der wahren Werte nicht beobachtbar ist, wird der Reliabilitätskoeffizient über Korrelationen geschätzt. Der zufällige Fehler, der für den Unterschied zwischen der Varianz der wahren und beobachteten Messwerte sorgt, kann auf verschiedene Ursachen zurückgeführt werden (z.B. Messfehler aufgrund unpräziser Fragestellungen, Einflüsse unterschiedlicher Interviewer, situative Unterschiede). Angesichts dieser Ursachen wurden unterschiedliche Methoden der Reliabilitätsprüfung entwickelt. (S.Albers (2007), s.375)

2.2 Stochastische und systematische Fehler

Führt man eine Messung durch, so wird der gemessene Wert fast immer vom wahren Wert abweichen; die Messung ist also mit einem Fehler behaftet. Dieser Fehler wird noch einmal in den systematischen und in den zufälligen Fehler aufgeteilt.

Zufällige bzw. Stochastische Fehler werden von messtechnisch nicht erfassbaren Änderungen der Maßkörper, der Messgeräte, des Messgegenstandes, der Umwelt und der Beobachter hervorgerufen. Auch numerische Berechnungen, z.B. mit einem Computer sind in der Regel nicht beliebig genau, sondern fehlerhaft. So führen Rundungen und endliche Tiefe jedes Iterationsverfahrens zu stochastische Fehlern. Die zufälligen Fehler können bei einer Einzelmessung weder nach ihrem Betrag noch nach ihrem Vorzeichen bestimmt werden; sie sind daher nicht zu korrigieren und machen das Ergebnis unsicher. Wiederholt man eine Messung unter gleichen Bedingungen, so weichen die einzelnen Messwerte voneinander ab, sie „streuen“. Sie können nach der Regeln der mathematischen Statistik in ihrer Gesamtheit zahlenmäßig erfasst und durch eine statistische Kenngröße charakterisiert werden.

Der Zufallsfehler ist also die zufällige Abweichung einer Beobachtung vom theoretisch wahren Wert, durch die die Reliabilität einer Messung beeinflusst wird. Reliabilität kann als der Grad der Messgenauigkeit eines Instruments definiert werden. Sie ist umso höher, je kleiner der zu einem Messwert gehörende Fehleranteil ist. Perfekte Reliabilität würde bedeuten, dass ein Instrument in der Lage ist, den wahren Wert ohne jeden Messfehler zu erfassen. Somit müsste eine vollständig reliable Messung bei wiederholter Befragung derselben Respondenten immer dieselben Ergebnisse liefern, sofern sich der wahre Wert nicht verändert. Eine perfekte Korrelation der Messergebnisse beider Messreihen wäre die Folge. Es gibt verschiedene statistische Tests auf Reliabilität (z.B. Retest-Reliabilität, Paralleltest- Reliabilität usw.), auf die an dieser Stelle aber nicht weiter eingegangen wird. Reliabilität ist eine notwendige, aber keine hinreichende Bedingung für die Validität einer Messung. Im Gegensatz zum Zufallsfehler beeinflusst der methodische Fehler die Validität einer Messung. Die Validität (Gültigkeit) ist das wichtigste Gütekriterium. Sie beschreibt, ob ein Test oder ein Konstrukt in der Lage ist, den Sachverhalt zu messen, der auch gemessen werden soll. Auch wenn eine hohe Reliabilität vorliegt, kann ein Test oder eine Befragung nutzlos sein, wenn nicht der richtige Sachverhalt gemessen wird.(Gräber (2009), Kapitel 6.1, s.1)

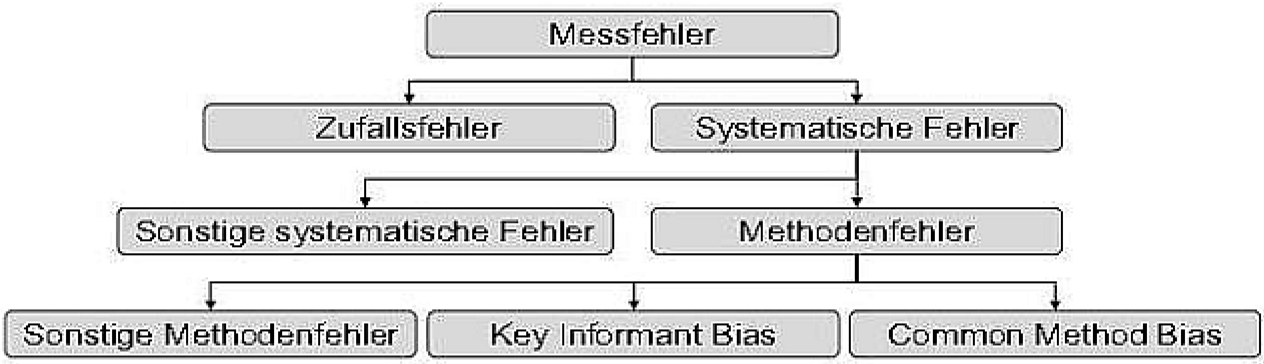

Abbildung 3: Systematisierung möglicher Messfehler (S.Albers (2007), s. 136)

Systematische bzw. statistische Fehler werden hauptsächlich durch Unvollkommenheiten der Maßverkörperung der Messgeräte, der Messverfahren und des Messgegenstandes sowie von messtechnisch erfassbaren Einflüssen der Umwelt und persönlichen Einflüssen der Beobachter hervorgerufen. Sie haben:

- ein bestimmtes Vorzeichen (+ oder -)
- unter gleichen Bedingungen den gleichen Betrag, d.h. sie können durch Wiederholung der Messung nicht festgestellt werden, sondern nur durch ein anderes (genaueres) Messgerät oder Messverfahren

Der systematische Fehler besitzt bei gleichen Messvorgängen die gleiche Struktur, z.B. eine Uhr geht um 1 Promille vor, so ergibt dies bei jeder Messung von 10 Sekunden einen systematischen Fehler von 0.01 Sekunden.

Der statistische Fehler beeinflusst die Messergebnisse in unterschiedlicher, unvorhersehbarer Weise. Die Größe des Fehlers lässt sich jedoch abschätzen, wenn Ergebnisse mehrerer gleichartiger, etwas wiederholter Messungen, zur Verfügung stehen. So stellt sich der zufällige Fehler oft als Summe vieler Elementarfehler (z.B. Ablesefehler, Umgebungsveränderung) dar und kann damit - wegen der Gültigkeit des zentralen Grenzwertsatzes - als normal verteilt angenommen werden.(Gräber (2009), Kapitel 6.1, s.2)

Die Ergebnisse $x_1, ..., x_n$ wiederholter Messungen eines wahren Wertes μ können dann als Realisationen einer $N(\mu^*, \sigma^2)$-verteilten Zufallsgröße X aufgefasst werden. Hier ist die zufällige Fehler bei der i-ten Messung gerade

$$\epsilon_i = x_i - \mu^*,$$

während der systematische Fehler, den man auch **Bias** nennt, gegeben ist durch

$$b = \mu^* - \mu$$

Damit ergibt sich nun:

$$\begin{aligned} x_i &= \mu + b + \epsilon_i \\ &= \text{wahrer Wert} + \text{systematischer Fehler} + \text{zufälliger Fehler}. \end{aligned}$$

Mit den Messwerten $x_1, ...x_n$ lassen sich $\mu^* = \mu + b$ und σ z.B. durch

$$\overline{x} = \frac{1}{n} \sum x_i \text{ und } s = \sqrt{\frac{1}{n-1} \sum (x_i - \overline{x})^2} \text{ für } i \in [1, n]$$

schätzen; weiterhin können Konfidenzintervalle für μ^* und σ^2 angegeben werden. Dadurch lässt sich ein Eindruck von der Größe und der Struktur des zufälligen Fehlers gewinnen.

Über die Größe des systematischen Fehlers b hingegen lässt sich aus den Messwerten $x_1, ..., x_n$ allein keine Informationen ziehen, da dieser für alle n Werte gleich ist. Es muss also gesorgt werden, dass ein systematischer Fehler weitgehend vermieden wird. Dabei hilft z.B. die Kalibration bzw. Justierung der verwandten Messgeräte. Kann man annehmen, dass kein systematischer Fehler vorhanden ist, so liefert $\overline{x}$ eine Schätzung für den wahren Wert μ, und ein Konfidenzintervall für μ entspricht dann gerade eine Konfidenzintervall für den Parameter μ^*. (J. Hartung (2009), s. 322)

2.3 Additive und multiplikative Fehler

Wie im vorherigen Kapitel bereits erläutert wurde, die systematische Messfehler stellen eine (deterministische) Funktion der systematischen Variablen dar. Das bedeutet:

$$U = f(X)$$

wobei U systematischer Messfehler und X wahrer (latenter) Werte sind.
Im einfachsten Fall sind die Funktionswerte konstant.

Für die beobachteten Variable gilt dann

$$W = X + U,$$

wobei U systematischer Messfehler ist. In dem Fall wird über einem **additiven** oder **konstanten systematischen Fehler** gesprochen.

Eine andere Möglichkeit ist der Fall eines **multiplikativen** oder **proportionalen systematischen Fehlers**. Hier ist der Fehler proportional zur systematischen Variablen:

$$U = pX$$

Daraus ergibt sich, dass die beobachtete Variable ebenfalls proportional zu der systematischen Variable ist. Mit $a := 1 + p$ gilt

$$W = X + pX = aX$$

Beide Fälle verallgemeinert werden als **lineare systematische Fehler** bezeichnet, wenn gilt :

$$W = aX + U$$

(H.Schneeweiß (1986), s. 21)

2.4 Differentieller und nicht-differentieller Fehler

Es ist wichtig, zwischen differentiellen und nicht-differentiellen Messfehler zu unterscheiden.

Wenn die interessierende Variable X nicht beobachtbar oder wegen fehlender Erhebungen nicht vorhanden ist und eine ähnliche Variable W keine andere Informationen über die Response Y als X und die fehlerfrei gemessene Variable Z enthält, ist W als nicht-differentieller Fehler äquivalent X. Formal hängt die Verteilung von $Y|(X, W, Z)$ nur von (X, Z) ab. In dem Fall wird W als Surrogat bezeichnet. Falls das nicht der Fall ist, dann ist der Messfehler differentiell.

Viele Probleme können bei der nicht-differentialen Messfehler auftreten. Z.B., wenn die wahre und beobachtete Variablen im bestimmten Zeitpunkt auftreten und Response erst später gemessen wird. So in der Brustkrebs-Studie wahre Variable ist

Ernährungsverhalten vor der Diagnose. Aber in den Fallkontrollstudien das berichtete Diät ist nur nach dem Befund beobachtbar. Eine Frau, bei der die Brustkrebs diagnostiziert wurde, stellt möglicherweise ihre Ernährung um. Daher die Messung des berichteten Ernährungsverhaltens nach dem Befund ist mit Krebsauswirkungen korreliert. Anders wäre bei der Betrachtung der Diät vor dem Befund.(R.J. Carrol and C.Crainiceanu (2006), s. 36)

Der Grund, warum die nicht-differentielle Messfehler wichtig sind, ist: Man kann den Response-Parameter bei der gegebenen wahren Einflussvariablen in einem Modell schätzen, selbst, wenn die wahre Variable X nicht beobachtbar sind. Anders aber bei den differentiellen Messfehlern. Außer ein paar speziellen Fällen, man muss wahre Variable beobachten können.

Differentielle Fehler können beispielsweise in biometrischen Untersuchungen relevant sein, wenn kranke Individuen ihr Verhalten (etwa Ernährungs- und Schlafgewohnheiten, Medikamentenkonsum, Rauch- und Suchtverhalten) in der Vergangenheit im Rückblick anders betrachten als gesunde Individuen. Bei selbst gemessenen oder selbst berichteten Variablen führt diese verschobene Wahrnehmung zu differentiellen Fehlern, die vom Krankheitsverlauf und damit der Response abhängen. Das Vorliegen eines nicht-differentiellen Fehler ist folglich kritisch zu prüfen.

Angenommen ein differentieller Messfehler beeinflusst die Zielvariable, d.h.

$$[Y|X;W] = [Y|X]$$

Für Y gibt es keine weitere Information in W oder X, wenn X bekannt ist. Dann kann das Fehler- und Haupt-Modell aufgesplittet werden:

$$[Y|W;X] = [Y|X]\,[W|X]\,[X] \mapsto f(Y|X)$$

Z.B. Y ist die Wahrscheinlichkeit einen Herzinfarkt zu bekommen, wenn ein bestimmter (bekannter!) Blutdruck X vorliegt.

Aus der substanziellen Sicht:

- Messprozess und Y sind unabhängig
- Blutdruck an einem bestimmten Tag ist irrelevant für das Herzinfarktrisiko, wenn ein Langzeit-Mittel bekannt ist.
- Durchschnittsexposition ist irrelevant, wenn die individuelle Exposition bekannt ist.
- Aber Menschen können ihr Ernährungsverhalten anders betrachten, wenn sie be-

reits einen Herzinfarkt hatten.

Nicht-differentielle Fehler haben den Vorteil, dass sich die Untersuchung der einfachen linearen Regression vereinfacht und die Bestimmung der Regressionsparameter grundsätzlich auch mit fehlerbehafteten Variable W möglich ist, schließlich ist

$$\begin{aligned} E(Y|W) &= E\{E(Y|X,W)|W\} \\ &= E\{E(Y|X)|W\} \\ &= E(\beta_o + \beta_x X|W) \\ &= \beta_0 + \beta_x E(X|W). \end{aligned}$$

(R.J. Carrol and C.Crainiceanu (2006), s. 38)

2.5 Klassische und Berkson Fehler

Zufällige Messfehler sind auf zwei unterschiedlichen Arten darstellbar, welche trotz ihrer technischen Ähnlichkeiten unterschiedliche Auswirkungen auf die Regression Untersuchungen haben und nachfolgend vorgestellt werden.

Im klassischen Fehlermodell ist W die fehlerbehaftete Messung des wahren latenten Wertes X, modelliert durch das Zusammenhang $W = X + U$ mit $U_{ij}|X_i \sim N(0, \sigma^2)$, wobei U und X stochastisch unabhängig sind. In diesem Fall gilt $E(W|X) = X$ und W ist folglich eine unverzerrte Messung für X. (R.J. Carrol and C.Crainiceanu (2006), s. 26)

Betrachtet man stattdessen $X = W + U$ mit $U_{ij}|X_i \sim N(0, \sigma^2)$, wobei U und W stochastisch unabhängig sind, so gilt $E(X|W) = W$. Man erhält trotz der Ähnlichkeit ein anderes Modell. Dieser Fehler wird Berkson-Fehler genannt und spiegelt eine andere Ausgangslage wieder. Während beim klassischen Fehler der exakte Wert X durch zufällige Messfehler verdeckt wird, stellt sich beim Berkson-Fehler der wahre nicht beobachtbare Wert X erst als Ergebnis einer (kontrollierbaren) Variablen W ein. Berkson-Modelle finden bei epidemiologischen Studien ihre Anwendung. In der Messfehler-Literatur wird hauptsächlich über die klassische Messfehler geschrieben. Diese messen den wahren Wert mit additivem Fehler und normalerweise konstanten Varianz.

Messfehler müssen mit erheblichen Vorsicht behandelt werden. Der Unterschied zwischen Berkson und klassischem Messfehler ist groß, wenn man vor hat, a priori zu erforschen, vor allem versucht den Einfluss zu kalkulieren. Es gibt einige technische Ähnlichkeiten zwischen klassischen und Berkson Fehler, aber in den Einflussberechnungen treten die unterschiedliche Aspekte auf. Für die gegebene Messfehlervarianz, wenn man sich überzeugen möchte, dass es trotz Messfehlern einen großen statistischen Einfluss gibt, muss angenommen werde, dass der Messfehler Berkson und nicht klassisch ist.

Wie unterscheiden wir Berkson und klassischen Fehler? Grundsätzlich, wenn die Auswahl zwischen den beiden Fehler gibt, handelt sich um einen klassischen Fehler, wenn die fehleranfällige Kovariable mithilfe von nötigen Messmittel eindeutig bei einer Person gemessen wurde, besonders bei Wiederholungsmöglichkeit der Messung. Zum Beispiel, wenn die Menschen die Fragen zum Essverhalten beantworten oder bei der Messung ihres Blutdrucks, handelt es sich um einen klassischen Messfehler.

Anders ist es, wenn alle Personen einer kleinen Gruppe oder einer sozialen Schicht Angaben zu einer fehleranfälligen Variable machen. Z.B. Textil- oder Bergarbeiter bei der gleichen Beschäftigungsdauer zeigen das gleiche Staubbelastungsbild, aber das wahre Bild ist bei jedem einzelnen Individuum jedoch anders. In diesem Fall handelt sich mit hoher Wahrscheinlichkeit um Berkson-Messfehler.(R.J. Carrol and C.Crainiceanu (2006), s. 27)

Weiterer Unterschied zwischen Berkson und klassischen Fehler ist:

- $var(W) \succ var(X)$ für klassischen Fehler und $var(X) \succ var(W)$ für Berkson-Fehler.

Im Praxis die Wahl besteht nicht immer nur zwischen klassischen additiven Messfehler-Modell ($W = X + U$) und klassischen Berkson-Fehler-Modell ($X = W + U$). Für die Messungen, die auf den Angaben der Befragten basieren, sind komplexere Modelle, die solche Abweichungen berücksichtigen, notwendig. Die individuellen Größen werden weiterhin gemessen und die Messungen wiederholt, aber die Abweichungen müssen gezeigt und berücksichtigt werden. Dabei ist das allgemeine klassische Fehler-Modell ($W = \gamma_0 + \gamma_x^t X + \gamma_z^t Z + U$, $E(U|X_Z) = 0$) geeignet.

Anstatt Berkson Modell wird oft für die ad hoc Zwecke die allgemeine Regressions Kalibrierung angewendet.(R.J. Carrol and C.Crainiceanu (2006), s. 26-28)

3 Messfehler in der linearen Regression

3.1 Übersicht der linearen Regressionsmodelle

Das zentrale Problem in einem linearen Modell mit fehlerbehafteten Daten ist die Parameterschätzung. Da die fehlerfreien systematische Variable X nicht direkt beobachtbar ist, muss man auf die fehlerbehaftete beobachtbare Variablen W zurückgreifen und mit ihrer Hilfe möglichst genaue Schätzungen konstruieren.

Man versucht zunächst eine Schätzung mit der Methode der kleinsten Quadrate (KQ-Methode) auf die Regression (Y, X) anzuwenden. Die so gewonnene Kleinstquadrat-Schätzer (KQ-Schätzer) sind i.d.R. inkonsistent. (H.Schneeweiß (1986), s.32)

In der Literatur wird oft über die Messfehler in der linearen Regression geschrieben und der Schluss gefasst, dass der Einfluss von den Messfehlern nur eine geringe Abweichung der Schätzung verursacht. Diese Schlussfolgerung muss mit Vorsicht angenommen werden. Die Messung, W, der wahre Wert, X, und möglicherweise anderen Variablen in einem Regressionsmodell können abhängig sein. Genauer gesagt, der Effekt des Messfehlers in einem linearen Modell muss unter Berücksichtigung anderer Variablen und deren Verteilung dargestellt werden.

Der Einfluss der Messfehler in der linearen Regression variiert in Abhängigkeit von den folgenden Merkmalen:

- ist das Regressionsmodell einfach oder multivariat
- ob die fehlerbehaftete Einflussgröße uni- oder multivariat ist
- ob die Messung eine Abweichung enthält

Die Verteilung des Messfehlers ist für die Auswirkung auf ein lineares Modell entscheidend. Davon hängt das zu wählende Verfahren für die Messfehlerkorrektur ab. (R.J. Carrol and C.Crainiceanu (2006), s.41)

Die grundlegende Parameter der Fehler-Modelle und linearen Regression für (Y, X, W) sind in der unterliegende Tabelle 2 gezeigt. Für alle Fälle wurde folgendes Regressionsmodell betrachtet:

$$Y = \beta_0 + \beta_x X + \epsilon$$

wobei X und ϵ sind unabhängig und $E(\epsilon) = 0$ und $var(\epsilon) = \sigma^2$.

Error Modell	ρ^2_{xw}	**Intercept**	**Slope**	**residual Variance**
Differential	ρ^2_{xw}	$\beta_0 + \beta_x\mu_x - \frac{\beta_x\sigma_{xw}+\sigma_{ew}}{\sigma^2_w}\mu_w$	$\beta_x(\sigma_{xw}\sigma^2_w) + \frac{\sigma_{ew}}{\sigma^2_w}$	$\sigma^2_e + \beta^2_x\sigma^2_x - \frac{(\sigma_{xw}\beta_x+\sigma_{ew})^2}{{\sigma_w}^2}$
Surrogate	ρ^2_{xw}	$\beta_0 + \beta_x\mu_x - \frac{\beta_x\sigma_{xw}}{\sigma^2_w}\mu_w$	$\beta_x(\frac{\sigma_{xw}}{\sigma^2_w})$	$\sigma^2_e + \beta^2_x\sigma^2_x(1-\rho^2_{xw})$
Classical	$\frac{\sigma^2_x}{\sigma^2_x+\sigma^2_{ue}}$	$\beta_0 + \beta_x\mu_x(1-\rho^2_{xw})$	$\beta_x(\frac{\sigma^2_x}{\sigma^2_x+\sigma^2_{ue}})$	$\sigma^2_e + \beta^2_x\sigma^2_x(1-\rho^2_{xw})$
B/C mixture	$\frac{\sigma^4_L(\sigma^2_L+\sigma^2_{u_b})^-1}{(\sigma^2_L+\sigma^2_{ue})}$	$\beta_0 + \beta_x\mu_x(1-\sigma^2_L\sigma^2_l + \sigma^2_{ue})$	$\beta_x(\frac{\sigma^2_L}{\sigma^2_l+\sigma^2_{ue}})$	$\sigma^2_e + \beta^2_x\sigma^2_x(1-\rho^2_{xw})$
Berkson	$\frac{\sigma^2_x-\sigma^2_{ub}}{\sigma^2_x}$	β_0	β_x	$\sigma^2_e + \beta^2_x\sigma^2_x(1-\rho^2_{xw})$
No error	1	β_0	β_x	σ^2_e

Tabelle 2: Übersicht zur Korrektur der Parameter: quadrierte Korrelation, Intercept, Steigung und Residualvarianzen in einem differentiellen, Surrogat, klassischen, klassisch-Berkson-gemischten, Berkson und fehlerfreien Regressionsmodell. (R.J. Carrol and C.Crainiceanu (2006), s.50)

3.2 Systematische Messfehler in der Regression

Wir betrachten eine einfache lineare Regression, die folgende Form hat:

$$Y = \beta_0 + \beta_x X + \epsilon, \text{ wobei } \epsilon \sim NV(0, \sigma_e^2)$$

Eine fehlerbehaftete Variable W ist über eine feste funktionale Abbildung $W = f(X)$ mit der fehlerfreien Variable X verknüpft. Wir beschränken hier nur auf die lineare Verknüpfungen und damit das Verhalten bei konstanten und proportionalen Fehlern.

Zuerst stellen wir eine fehlerbehaftete Variable $W = X + d$, die um einen konstanten Messfehler d von der wahre Variable abweicht, dar. In diesem Fall verändert sich die angegebenen Modellgleichung, und es gilt

$$Y = \beta_0 + \beta_X^*(X + d) + \epsilon$$

Durch Umformung ergibt sich mit $\beta_0^* + \beta_X^* X + \epsilon$

$$Y = \beta_0^* + \beta_X^* X + \epsilon$$

Als Parameterschätzer erhalten wir einen verzerrten, inkonsistenten Wert $\beta_0^* = \beta_0 + \beta_X d$ sowie konsistent geschätzten $\beta_X^* = \beta_X$. Die beobachteten Werte sind um einen konstanten Wert d nach rechts und in der Folge um den Wert $\beta_X d$ nach oben verschoben.

Betrachte eine mit einem proportionalen Fehler behaftete Variable $W = cX$. Die Modellgleichung an der Stelle sieht folgendermaßen aus:

$$Y = \beta_0^* + \beta_X^* cX + \epsilon$$

und die verzerrten KQ-Schätzer ergeben sich als β_0^* und $\beta_X^* = \frac{1}{c}\beta_X$.

In beiden Fällen erhalten wir durch KQ-Schätzung verzerrte Schätzer, die wegen bekannten konstanten Fehler d und proportionalen Fehler c korrigiert werden können. Ist der systematische Messfehler bekannt, so können auch die gemessenen, fehlerbehafteten Werte W vor Durchführung der Regression korrigiert werden. (H.Schneeweiß (1986), s. 32-37)

3.3 Stochastische Messfehler in der Regression

3.3.1 Klassische Messfehler in der Regression

Wie beschäftigen uns weiter mit dem linearen Regressionsmodell, das folgende Struktur hat:

$$Y = \beta_0 + \beta_X X + \epsilon, \text{ wobei } \epsilon \sim NV(0, \sigma_\epsilon^2)$$

Wenn X fehlerbehaftet ist, also $W = X + U$ mit $U \sim NV(0, \sigma_U^2)$, dann sieht das Modell folgend aus:

$$Y = \beta_0^* + \beta_W(X + U) + \epsilon.$$

(R.J. Carrol and C.Crainiceanu (2006), s. 42-43)

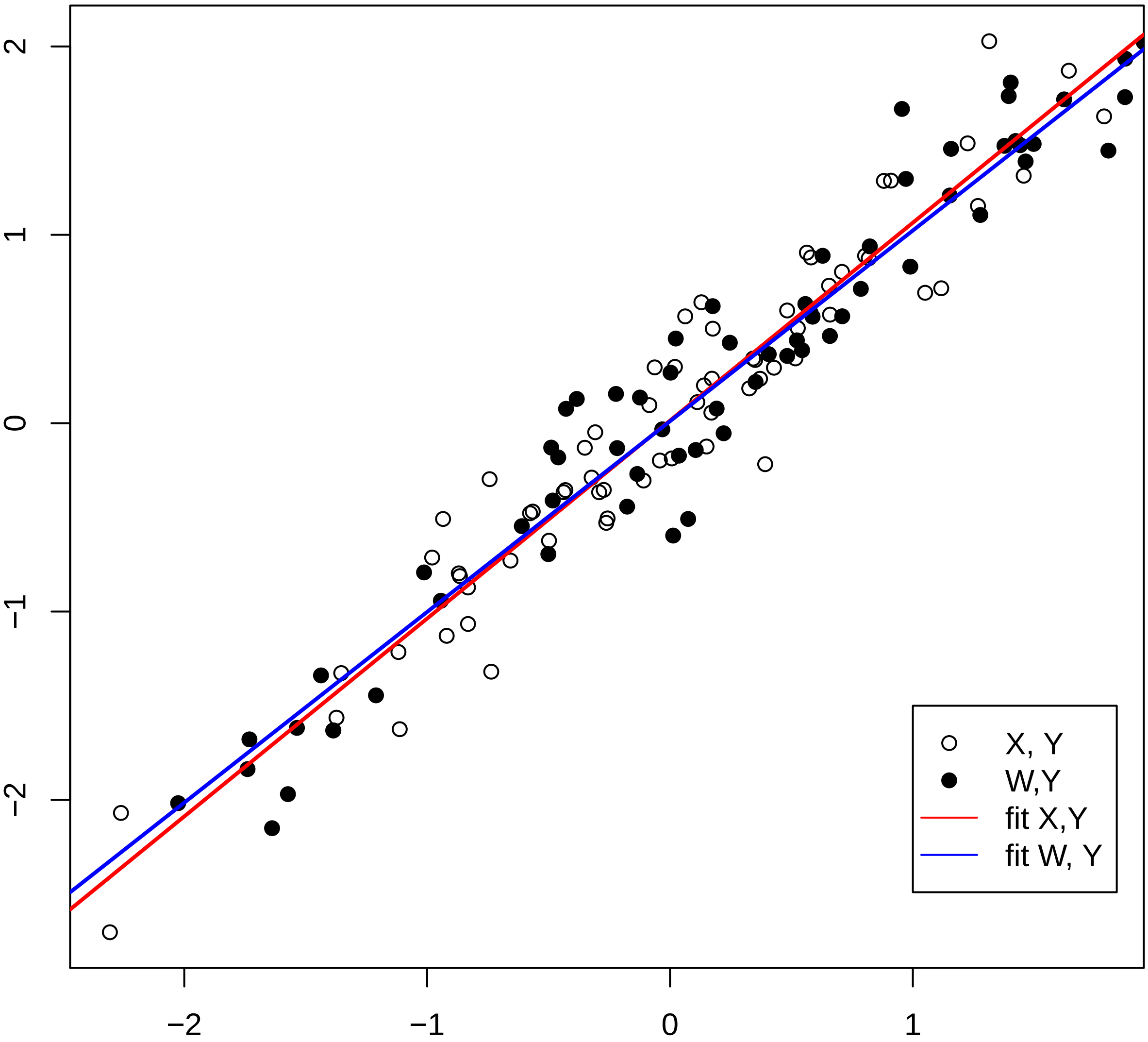

Abbildung 4: Einfache lineare Regression beim klassischen Fehler

Die Abbildung zeigt die Folgen der klassischen Messfehler in der linearen Regression. Die rote Regressionsgerade gibt den wahren Zusammenhang zwischen Einflussvariable und Response an. Klassischer Messfehler ist hier um einen konstanten Term nach links bzw. rechts modelliert. Die blaue Regressionsgerade bildet die durch Kleinste-Quadrate-Schätzung verursachte Veränderung ab. Es kommt zur einer systematischen Unterschätzung der Steigung.

Die gewöhnliche KQ-Schätzung von X auf Y ist kein konsistenter Schätzer von β_X. Anstatt β_x bekommt man $\beta_{x*} = \lambda\beta_x$, wobei

$$\lambda = \frac{\sigma_x^2}{\sigma_x^2+\sigma_u^2} < 1$$

In dem Fall wird λ *Reliabilitätsfaktor* genannt. Da W ein fehlerbehafteter Prädiktor ist, besitzt er einen schwächeren Einfluss auf die Response als X. Darüber hinaus vergrößert sich die Varianz der Beobachtungen. Anstelle von $Var(Y|X) = \sigma_i^2$ erhalten wir

$$var(Y|W) = \sigma_\epsilon^2 + \frac{\beta_x^2\sigma_u^2\sigma_x^2}{\sigma_x^2+\sigma_u^2} = \sigma_\epsilon^2 + \lambda\beta_x^2\sigma_u^2 > Var(Y|X).$$

(R.J. Carrol and C.Crainiceanu (2006), s. 44)

3.3.2 Berkson-Fehler in der Regression

Die Auswirkung des Berkson-Fehlers in der linearen Regression ist von großer Bedeutung. Stelle das folgende lineare Modell auf

$$Y = \beta_0 + \beta_X X + \epsilon,\ \epsilon \sim NV(0, \sigma_\epsilon^2)$$

mit $E(X|W) = W$ und $X = W + U_b$, wobei U_b ein Berkson-Fehler ist. In der Tabelle 2 ist zu sehen, dass die Regressionsparameter unverzerrt zum Berkson-Messfehler sind.

Da W ein Surrogate ist, betrachten wir die Auswirkungen der wahren Werte X und dazu gehörenden Werte der Response.

Die Kleinste-Quadrate-Regressionsgleichungen sind im beobachteten und wahren Fall identisch, wenn wir wie im klassischen Fall annehmen, dass alle beobachteten Punkte gleichermaßen realisiert werden. Es entsteht keine Verzerrung der Regressionskoeffizienten im Berkson-Modell. $X = W + U$ ergibt sich nach Festlegung von W als zufälliges Ergebnis. Betrachten wir folgendes Regressionsmodell:

$$Y = \beta_0^* + \beta_W W + \epsilon, \text{ wobei } \epsilon \sim NV(0, \sigma_\epsilon^2)$$

Da $E(X|W) = W$ (nicht-differentieller Fehler), $E(Y|W) = \beta_0 + \beta_X E(X|W) = \beta_0 + \beta_x W$. Daraus folgt, dass die Koeffizientenschätzer für β_0 und $\beta_X = \beta_W$ unverzerrt sind. Statt

$Var(Y|X)$ erhalten wir folgende

$$Var(Y|W) = Var(Y) + Var(E(Y|W)) = \sigma_\epsilon^2 + \beta_X^2\sigma_U^2 > Var(Y|X).$$

(R.J. Carrol and C.Crainiceanu (2006), s. 45)

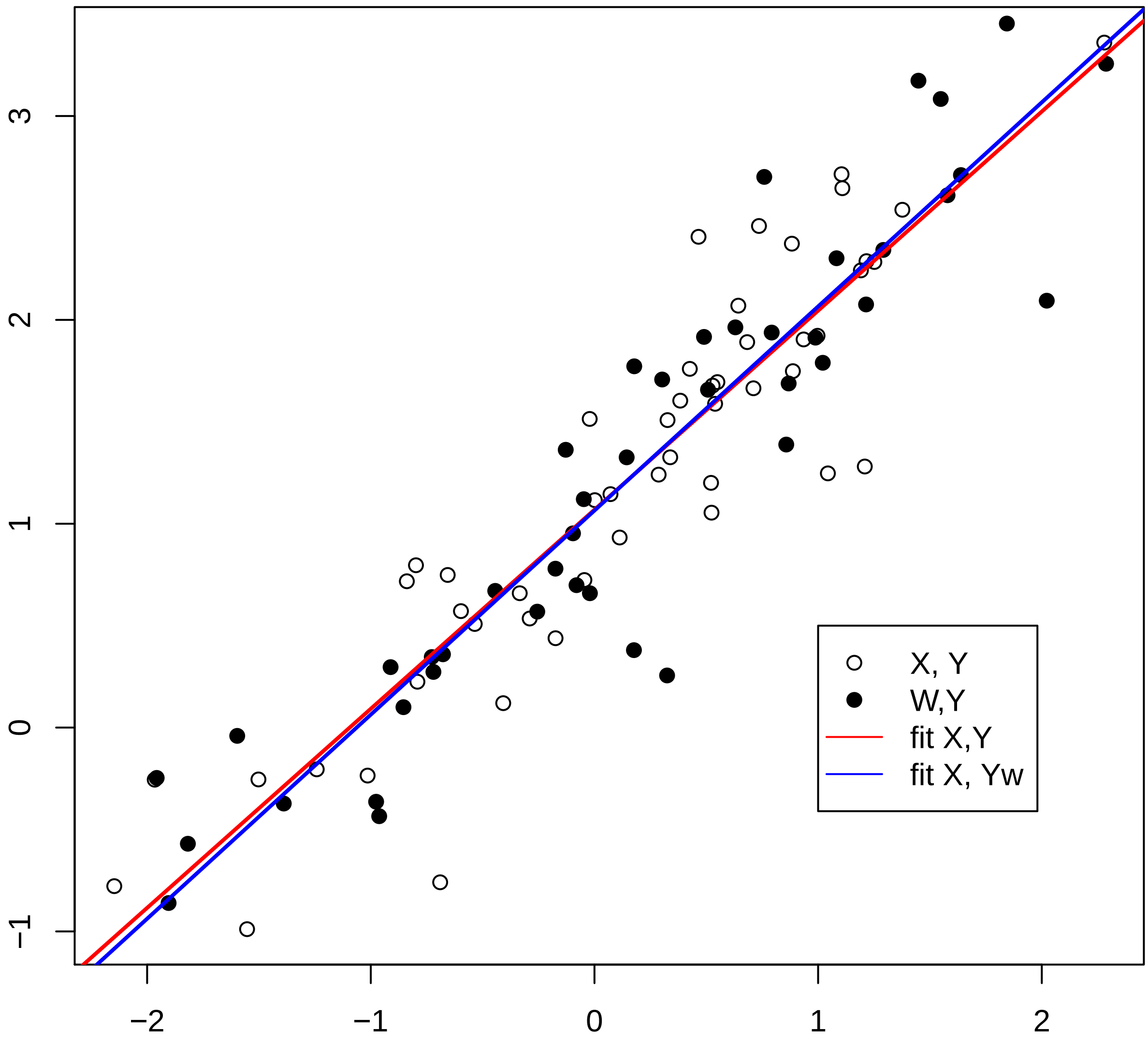

Abbildung 5: Auswirkung des Berkson-Fehlers in der linearen Regression: KQ-Regressionsgleichungen sind fast identisch und zur Folge keine Verzerrung der Regressionskoeffizienten.

3.3.3 Kombination von klassischen und Berkson-Fehlern

Wir betrachten ein Regressionsmodell, bei dem die Einflussvariable sowohl klassische als auch Berkson Komponente beinhaltet. Aus dem klassischen Fehlermodell ist bekannt, dass $W = X + U$ und der bester linearer Prädiktor für $(X|W)$

$$X = \lambda W + (1-\lambda)E(U) + U^*$$

wobei $U^* = (1-\lambda)(X - E(X)) - \lambda U$ und der Reliabilitätsfaktor $\lambda = \frac{\sigma_X^2}{\sigma_X^2 + \sigma_U^2} < 1$. Sie zeigen, dass U^* und W unkorreliert sind, also überführen folglich den klassischen Fehler in ein Berkson-Fehler.

Wie schon gezeigt wurde, ist das Berkson-Modell unverzerrt. Also die Fehlerstruktur wird durch Berkson-Modell als stochastischer Fehler dargestellt, verzerrt durch den systematischen, proportionalen Fehler λ. (R.J. Carrol and C.Crainiceanu (2006), s. 29)

3.3.4 Multiple lineare Regression

In der multiple lineare Regression sind die Auswirkungen der Messfehler komplexer, auch für das klassische additive Fehlermodell.

Wir betrachten skalare X und zusätzlich eine Kovariable Z, die fehlerfrei gemessen wurde. Das lineare Modell sieht jetzt folgendermaßen aus:

$$Y = \beta_0 + \beta_x X + \beta_z^t Z + \epsilon,$$

wobei Z und β_z Spaltenvektoren und β_z^t Zeilenvektor sind. Die Kleinste-Quadrate-Schätzer für die Regressionskoeffizienten sind im Falle multipler Regression wie bereits im Fall einfacher linearer Regression verzerrt. Statt β_x wird $\lambda_1 \beta_x$ mit $\lambda_1 = \frac{\sigma_{x|z}^2}{\sigma_{w|z}^2} = \frac{\sigma_{x|z}^2}{\sigma_{x|z}^2 + \sigma_u^2}$, und statt β_z wird $\beta_z + \beta_x(1-\lambda_1)\Gamma_z$ geschätzt. Γ_z ist der Regressionskoeffizient von X auf Z, d.h. $E(X|Z) = \Gamma_0 + \Gamma_z^t Z$. (R.J. Carrol and C.Crainiceanu (2006), s. 52)

Betrachten wir eine Variable X, die z.B. in zwei Gruppen unterteilt ist. Angenommen, Werte einer Gruppe sind im Durchschnitt höher als die Werte der zweiten Gruppe. Sei W eine fehlerbehaftete Variable und $W = X + U$, wo U ein Messfehler ist.

In der Regression bei der theoretisch wahren Variable X ist der Zusammenhang zur Response rein linear. Somit sind keine ausgeprägte Gruppenunterschiede festzustellen. Bei der Regression mit fehlerbehafteter Variable W sind deutliche Unterschiede der Gruppeneffekte und eine Abflachung der Regressionsgerade zu bemerken. Diese Auswirkung betrifft nur die fehlerbehaftete Daten (z.B. Geschlecht- und Placebo-Effekte, Zeitabschnitte in der Ökonometrie).

Somit wird an der Stelle betont, dass Messfehler auch auf die Koeffizientenschätzer der anderen fehlerfreien Variablen Auswirkungen haben können und deshalb nicht vernachlässigt werden dürfen.

3.3.5 Korrektur der Abweichung

Wie wir in vorherigen Kapiteln gesehen haben, KQ-Schätzer ist bei fehlerbehafteten Daten verzerrt. Die Richtung und Umfang der Verzerrung ist vom Regressionsmodell, Verteilung des Messfehlers und Korrelation zwischen geschätzten Variablen abhängig.

In der einfachen linearen Regression mit klassischen additiven Messfehlermodell die KQ-Schätzer ist als $\lambda\beta_x$ geschätzt, wobei $\lambda = \frac{\sigma_x^2}{\sigma_x^2+\sigma_u^2}$ Attenuation-Koeffizient ist. Wenn dieser bekannt ist, dann erhält man einen unverzerrten geschätzten β_x durch Multiplikation mit $\frac{1}{\lambda}$.

Der Attenuation-Koeffizient ist jedoch selten bekannt und muss geschätzt werden. Bei den systematischen Messfehlern ist das nur dann möglich, wenn die Fehlerparameter bekannt sind. Damit kann man die Daten reinigen oder die geschätzten Parameter korrigieren. Wenn $\widehat{\sigma_u^2}$ Schätzung der Messfehlervarianz und $\widehat{\sigma_w^2}$ Stichprobenvarianz von W ist , dann die konsistente Schätzung des Attenuation-Koeffizienten ist

$$\widehat{\lambda} = (\widehat{\sigma_w^2} - \widehat{\sigma_u^2})/\widehat{\sigma_w^2}.$$

Die resultierende Schätzung ist $\widehat{\beta_{x*}}/\widehat{\lambda}$

In der kleinen Stichproben, die Stichprobenverteilung von $\widehat{\beta_{x*}}/\widehat{\lambda}$ ist stark verzerrt. In diesen Fällen ist modifizierte Version von der Momentenmethode Schätzer ratsam. (R.J. Carrol and C.Crainiceanu (2006), s.55)

Orthogonale Regression

Sei $Y = \beta_0 + \beta_x X + \epsilon$ und $W = X + U$, wobei ϵ und U sind unkorreliert. Während Momentenmethode Schätzer verlangt, dass Messfehlervarianz σ_i^2 bekannt ist oder geschätzt werden kann, orthogonale Regression verlangt dasselbe für den Parameter $\eta = \sigma_e^2/\sigma_u^2$.

Orthogonaler Regressionsschätzer minimiert orthogonalen Abstand von (Y, W) zu der Gerade $y = \beta_0 + \beta_y x$, wenn $\eta = 1$.

Orthogonaler Regressionsschätzer ist der funktionale Maximum-Likelihood-Schätzer unter Annahme, dass $(X_1, \dots, X_n)$ unbekannte feste Konstanten und die Fehler (ϵ, U) unabhängig und normalverteilt sind.

Orthogonale Regression wird öfter als Momentenmethode Schätzung verwendet, in der nur der Kennzahl η von Fehlervarianzen bekannt oder geschätzt sein muss. Das Problem ist, dass η nicht richtig spezifiziert oder geschätzt werden kann. Verwendung von der orthogonalen Regression mit falsch spezifiziertem Wert η ergibt oft eine unakzeptable große Überkorrektur, die Abschwächung des Messfehlers verursacht.

Also das Verfahren ist trotz seiner Einfachheit mit Problemen verbunden. (R.J. Carrol and C.Crainiceanu (2006), s. 57)

4 Schlussfolgerung

In der Seminararbeit wurde versucht die am meisten bekannten und in der Literatur beschriebene Messfehler zu erläutern. Weiterhin wurde die Auswirkungen von diesen Messfehlern in der linearen Regression beschrieben. Dabei wurde nur ein Teil der Folgen der fehlerbehafteten Variablen in der linearen Regression dargestellt. Dabei wurde nur auf diese Art der Regression beschränkt. Das Ziel war die Grundlagen zu ermitteln und zu zeigen, dass die Berücksichtigung von Abweichungen in beobachteten Daten von großer Bedeutung sind. Die Auswirkungen von fehlerbehafteten Variablen in der linearen Regression hängt von der Fehlerstruktur, dessen Verteilung und Zusammenhang mit den anderen Kovariablen.

Literatur

Augustin, T. and Wiencierz, A. (2013). Wirtschafts- und Sozialstatistik Foliensatz WiSe 13/14. 14.10.2014http://www.statistik.lmu.de/institut/ag/agmg/lehre/2013_WiSe/Wiso/WiSo_folien_kap_1.pdf.

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

Gräber, P.-W. (2009). Systemanalyse. Foliensatz Automatisierungstechnik in der Wasserwirtschaft. 14.11.2014 http://tu-dresden.de/die_tu_dresden/fakultaeten/fakultaet_forst_geo_und_hydrowissenschaften/fachrichtung_wasserwesen/iaa/systemanalyse/studium/folder.2009-01-29.lehre/folder.2009-04-03.at/AT%206.pdf.

Hartung, J., Elpert, B. and Klösener, K. (2009). *Statistik: Lehr- und Handbuch der angewandten Statistik*, 15. edn, Oldenbourg Wissenschaftsverlag GmbH, München.

Höpcke, W. (1980). *Fehlerlehre und Ausgleichsrechnung*, de Gruyter, Berlin.

H.Schneeweiß and Mittag, H.-J. (1986). *Lineare Modelle mit fehlerbehafteten Daten*, Physica-Verlag Heidelberg Wien, Würzburg.

S.Albers, D.Klapper, U.Konradt, Walter, A. and J.Wolf (2007). *Methodik der empirischen Forschung*, 2. edn, GWV Fachverlage GmbH, Wiesbaden.