Messfehler und fehlende Daten

Vorbereitungsmaterial zum Seminar
„nationale und internationale Berichterstattung„

Autorin: Veronika Gschwilm
Betreuer: Georg Schollmeyer

13. Juni 2013

0.0.1 Zusammenfassung

Inhaltsverzeichnis

1 Einführung

Der Umgang mit Messfehlern und fehlenden Werten ist auch in den Sozialwissenschaften von enormer Wichtigkeit. Allein das Beispiel der heiklen Frage nach dem Einkommen einer Person in einer statistischen Untersuchung zeigt, wie alltäglich diese beiden Probleme im statistischen Alltag auftauchen:
Diese Frage stellt einen großen Eingriff in die Privatsphäre der Probanden einer Befragung dar. Zusätzlich befürchten viele Personen, dass ihre Daten an Dritte weitergegeben werden.
Daher verweigern viele Probanden die Antwort, so dass einige fehlende Werte auftreten.
Auch Messfehler entstehen häufig, wenn Leute sich nicht an ihre genaue Einkommenshöhe erinnern oder es absichtlich über- oder unterschätzen.
Aus diesen fehlenden Werten und Messfehlern können sich große Probleme für die statistische Analyse ergeben.
In dieser Seminararbeit soll daher die Problematik von fehlenden Daten und Messfehlern genauer erläutert werden und Methoden im Umgang aufgezeigt werden. Der erste Teil der Seminararbeit beschäftigt sich mit den Messfehlern; im zweiten Teil wird genauer auf fehlende Daten eingegangen.

2 Messfehler

Ziel jeder Untersuchung oder Studie sind exakte, fehlerfreie Messergebnisse, um mit diesen fehlerfreien Daten auch exakte Analysen durchführen zu können. In der Praxis jedoch sind Messfehler unvermeidlich, da es grundsätzlich nicht möglich ist ohne jeglichen Fehler zu messen.
Zunächst wird in dieser Arbeit geklärt, was überhaupt die Ursachen für Messfehler sind.
Anschließend wird eine Motivation zur Beschäftigung mit Messfehlern gegeben, indem die direkten negativen Auswirkungen dargestellt werden.
Darauf aufbauend werden die verschiedenen Messfehler in einzelne Kategorien aufgeteilt. Diese Unterscheidungen sind in der Praxis äußerst wichtig, da nur so der richtige Umgang mit Messfehlern gewährleistet werden kann.

2.1 Ursachen

Für Messfehler können mehrere Gründe aufgeführt werden. Da das zentrale Mittel zur Datenerhebung in den Sozialwissenschaften das Interview ist, wird an dieser Stelle auf die Messfehler in Interviews eingegangen. Diese lassen sich unterteilen in interviewerabhängige Messfehler und Fehler durch die Befragten.
Interviewerabhängige Messfehler sind einzuteilen in absichtliches Fehlhandeln, äußere Merkmale des Interviewers (wenn bestimmte demographische und sozio-ökonomische Merkmale des Interviewers das Verhalten des Befragten beeinflussen), unterschiedliche Bearbeitung des Fragebogens (z.B. Übersehen von Fragen, Umformulierungen) und unterschiedliche Assistenzleistungen(z.B. Hilfestellungen bei Nicht-Verstehen einer Frage).
Messfehler durch die Befragten entstehen beispielsweise durch Anpassung an die vermutete Meinung des Interviewers („Gefälligkeitsantworten“), soziale Erwünschtheit der Antwort oder der Drang nach positiver Selbstdarstellung (3, S.5-20).

2.2 Motivation

Messfehler in Kovariaten führen zu drei Problemen, welche die statistische Analyse erschweren bzw. behindern. Um diese Probleme aufzuzeigen ist in Abbildung 1 ein Beispiel aufgeführt:
Es wurde eine Regression von X auf Y durchgeführt. Dafür wurden 200 Daten verwendet, die über ein Intervall von $[-2, 2]$ verteilt sind . Der Mittelwert beträgt $sin(2X)$.
In der oberen Grafik wurden die Daten in Form einer Sinuskurve simuliert. Man kann klar den sinusförmigen Verlauf der Regressionsfunktion erkennen.
Im unteren Teil der Grafik wurden die Daten mit Messfehler simuliert. Das heißt die jetzt verwendeten Daten W sind jetzt eine erwartungstreue Schätzung der eigentlichen, wahren Daten.
Dies führt dazu, dass die Sinuskurve nun nicht mehr erkennbar ist und die Haupteigenschaften der Daten verdeckt werden.

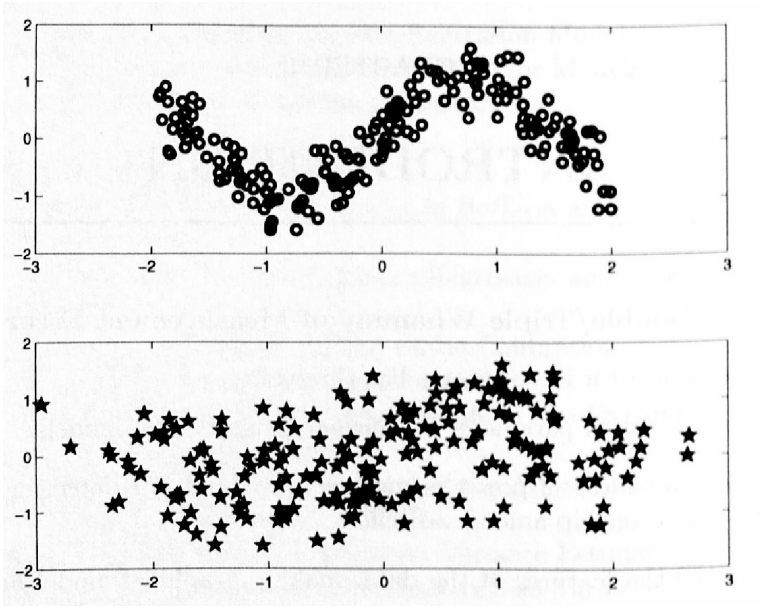

Abbildung 1: Die obere Abbildung zeigt eine Regression auf die wahren Werte von X. Auf der unteren Abbildung wird eine Regression auf die gemessene Werte W mit Messfehler veranschaulicht. (Quelle: (1, S.2)

So lassen sich die Probleme, die von Messfehlern verursacht werden, in 3 Bereiche aufteilen:
Zum einen entstehen dadurch verzerrte Parameterschätzer in statistischen Modellen. Zum anderen wird die Aussagekraft statistischer Tests (teilweise stark) reduziert, so dass die interessierenden Abhängigkeiten der Variablen schwerer oder gar nicht mehr entdeckt werden können. Diese zwei Probleme werden auch das „double whammy“ der Messfehler genannt, da sie bei der Analyse von Daten relativ große Probleme bereiten. Der dritte Effekt sagt aus, dass durch Messfehler Eigenschaften der Daten verborgen werden und die Analyse über graphische Modelle dadurch erschwert wird (1, S.1)

2.3 Notation

Zur Erleichterung der Darstellungen werden im Folgenden bestimmte Annahmen gemacht:

Y bezeichnet die Responsevariable hinsichlich der Prädiktoren.
Dabei unterscheidet man zwischen zwei verschiedenen Prädiktoren: Z bezeichnet jene Kovariablen, welche für alle Untersuchungsobjekte fehlerfrei gemessen werden können.

X hingegen bezeichnet eine Variable, welche nicht für alle Untersuchungsobjekte exakt gemessen werden kann.
Stattdessen wird nun eine ähnliche Variable W gemessen. Diese bezieht sich auf die unmessbare Variable X. (1, S.1-4)
X könnte beispielsweise das wahre Einkommen einer Person sein. Gemessen werden kann aber nur das beobachtete bzw. erfragte Einkommen W dieser Person.
Wahres und untersuchtes Einkommen unterscheiden sich durch einen additiven Fehler. Dieser kann beispielsweise durch Erinnerungsfehler oder absichtliche Unter- bzw. Überschätzungen auftreten.

Die Parameter in einem Modell für Y können dann logischerweise nicht direkt mit (Z, X) modelliert werden, da X ja nicht gemessen werden konnte. Stattdessen muss der „Umweg“ über (Z, W) gegangen werden. Denn würde man einfach X mit den Werten W ersetzen, würde man verzerrte Schätzer erhalten.
Das Ziel von Modellierungen mit Messfehlern ist es also eine erwartungstreue Schätzung für diese Parameter zu erhalten, indem man ein Modell für Y mit (Z, W) schätzt.
Das heißt aber auch, dass hier logischerweise Anpassungen im Modell nötig sind, da die Parameter der Regression von Y auf (Z, X) und von Y auf (Z, W) sich unterscheiden.
Wichtig ist aber, dass zunächst der Fehlertyp genau analysiert wird, um den bestmöglichen Umgang mit fehlerbehafteten Daten zu finden. (1, S.1-4)

2.4 Fehlerarten

Im diesem Gliederungspunkt sollen zunächst die verschiedenen Arten von Messfehlern vorgestellt und unterschieden werden. Dies ist nötig, da sich die einzelnen Fehlerarten verschieden auf die Daten auswirken und so unterschiedliche Behandlungsmethoden gewählt werden müssen.
Im Großen und Ganzen gibt es drei verschiedene Unterteilungen der Messfehler: Es wird unterschieden zwischen zufälligem und systematischem Messfehler und zwischen differentiellem und nicht-differentiellem Messfehler. Außerdem erfolgt noch die Trennung von Berkson-Fehler und klassischem Fehler.

2.4.1 zufälliger und systematischer Messfehler

Zunächst erfolgt eine Unterscheidung zwischen zufälligem und systematischem Messfehler.
Man spricht von einem systematischen Messfehler, wenn die Fehlervariable eine Funktion der systematischen Variablen ist.

Im einfachsten Fall spricht man von einem additiven systematischen Messfehler, dargestellt durch

$$W_i = X_i + b$$

Der gemessene Wert W_i ergibt sich aus dem wahren Wert X sowie der Konstanten b. Der für den Statistiker interessante Fall tritt ein, wenn der Messfehler stochastisch ist, also von Messung zu Messung variiert. Mathematisch wird dies wie folgt dargestellt:

$$W_i = X_i + U_i$$

Wichtig ist, dass für den additive Fehler U_i gilt $E(U) = 0$.
Zudem wird angenommen, dass die Höhe von U_i unabhängig von X_i ist. Das heißt, dass besonders hohe Ausprägungen von X keine Korrelation zu hohen additiven Fehlern U_i aufweisen.
Während der zufällige Messfehler durch die Zusammenfassung der Messergebnisse ausgeglichen wird, ist dies bei der Zusammenfassung von systematisch mit Messfehlern behafteten Messergebnissen nicht der Fall.
In der Praxis kommen in der Regel beide Messfehler gleichzeitig vor.
Der Unterschied zwischen systematischem und zufälligem Messfehler ist in Abbildung 2 veranschaulicht (6, S.21-29).

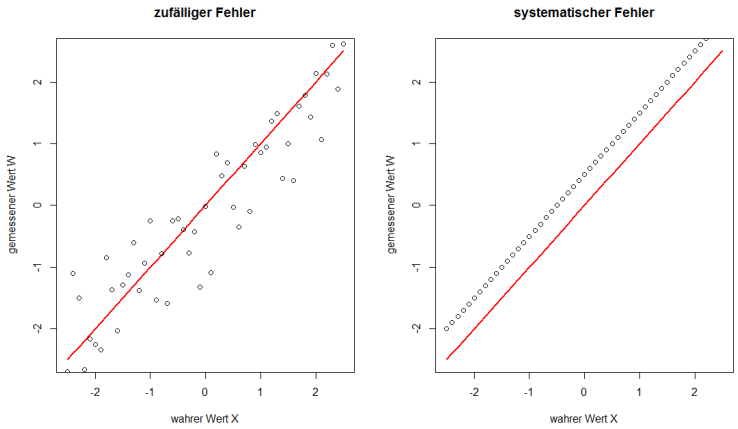

Abbildung 2: Zufälliger und theoretischer Messwert veranschaulicht anhand von theoretischen Messwerten

Als sozialwissenschaftliches Beispiel sei hier wieder die Befragung nach dem Einkommen der Personen aufgeführt.
Ein Teil der Abweichungen vom wahren Einkommen kann z.B. durch Erinnerungsfehler verursacht werden. Diese werden als rein zufällig aufgefasst.
Ist es den Personen jedoch unangenehm mit einem sehr hohen oder sehr niedrigen Einkommen aufzufallen, kann es vorkommen, dass diese ihr Einkommen absichtlich unter- bzw. überschätzen. Das heißt die Werte werden systematisch unter- oder überschätzt (5).

2.4.2 differentieller und nicht-differentieller Fehler

Desweiteren ist auch eine Trennung von differentiellem und nicht-differentiellem Fehler von großer Bedeutung.
Ein nicht-differentieller Fehler tritt ein, falls die Verteilung von Y gegeben (X, Z, W) nur von (X, Z) abhängt. W hat also keine zusätzliche Information über Y.
Dies bedeutet also, dass W bedingt unabhängig von Y ist, welches ja bereits durch (X, Z) gegeben ist.
In diesem Fall ist W ein Surrogat. Z.B. ist Angst nicht direkt messbar. Deshalb wird als Surrogat die Steigerung der Herzfrequenz in einer Angstsituation erfasst.
Tritt dieser Fall nicht ein, bezeichnet man den Messfehler als differentiellen Fehler.

In der Praxis hat der nicht-differentielle Fehler einen entscheidenden Vorteil: Damit können nämlich auch Modelle für den Response gegeben X geschätzt werden, wenn X gar nicht vorhanden ist. Bei einem differentiellen Messfehler müsste X zusätzlich erhoben werden.
Somit kann zum Beispiel eine einfache lineare Regression der untersuchten Daten mit einer Regression von Y über $E(X|W)$ durchgeführt werden.
Dies ist möglich da

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

(1, S.36 ff.)

2.4.3 klassischer Fehler und Berkson-Fehler

Zu guter Letzt erfolgt nun die Veranschaulichung des klassischen Fehlers und des Berkson-Fehlers. Diese müssen ebenfalls unterschieden werden.

Liegt ein klassischer Messfehler vor, so wird die wahre Variable X mit einem additiven Fehler U gemessen. Daraus ergibt sich dann der gemessene Wert W
Die mathematische Darstellung lautet:

$$W_i = X_i + U_i$$

Für den additiven Fehler U_{ij} muss $E(U_{ij}|X_i) = 0$ vorliegen, der Mittelwert ist also 0. Zudem ist der Fehler U unahängig von X. Die Fehlerstruktur von U_{ij} kann dabei homoskedastisch oder heteroskedastisch sein.

Der Berkson-Fehler besagt, dass der wahre Wert X sich aus dem gemessenen Wert W plus dem additiven Fehler U_i ergibt.
Der wahre Wert X hat hier eine größere Variabilität als beim klassischen Fehler. Der Unterschied ist, dass der Messfehler U hier unabhängig von W ist.
Es gilt also:

$$X_i = W_i + U_i$$

mit

$$E(U_{ij}|W_i) = 0$$

Ein Berkson-Fehler liegt beispielsweise vor, wenn in einer Textilfabrik alle Arbeiter, die gleichlang im selben Tätigkeitsbereich arbeiten, die selbe Staubbelastung zugewiesen bekommen. Die wahre Staubbelastung aber ist je nach Individuum verschieden.

Es ist wichtig, den Unterschied zwischen Berkson-Fehler und klassischem Fehler zu verstehen, da ein falsch gewähltes Modell oft zu fehlerhaften Rückschlüssen führt.
So berechnet sich die Varianz der wahren Messungen σ_x^2 für jeden der beiden Fehler unterschiedlich.
Es gilt für den klassischen Fehler, $\sigma_W^2 > \sigma_X^2$, während dies beim Berkson-Fehler genau umgekehrt gilt, nämlich $\sigma_X^2 > \sigma_W^2$.

Dies ist auch in Abbildung 3 zu sehen, für die in einer Studie die radioaktive Dosis W der Probanden untersucht wurde. Hier wird die wahre Dosis X dargestellt, für den Fall, das ein Berkson-Fehler angenommen wird (oben) und für den Fall, das der Fehler klassisch ist (unten).
Die geringere Variabilität der wahren Dosis beim klassischen Fehler lässt darauf schließen, dass die statistische Aussagekraft der Analysen geringer sein wird als bei der Annahme eines Berkson-Fehlers.

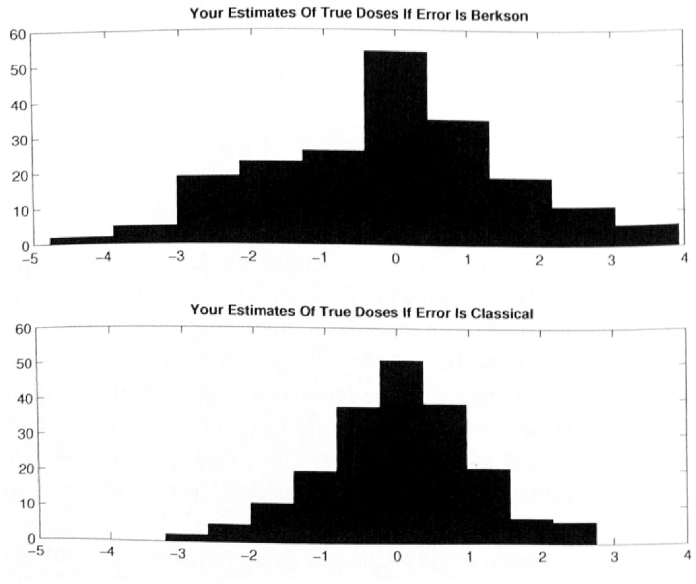

Abbildung 3: Darstellung des Unterschieds in der Variabilität bei Berkson-Fehler und klassischem Fehler. Quelle: (1, S.6)

2.5 Einfache Regression mit Messfehlern

Abschließend wird nun noch die Auswirkung der Messfehler auf die einfache Regression gezeigt. Zusätzlich werden statistische Methoden beschrieben, wie diese Effekte der Messfehler korrigiert werden können.

Prinzipiell können die Effekte, die Messfehler auf lineare Modelle haben, stark variieren. Sie reichen von einer Dämpfung der Messung über das Verbergen realer Zusammenhänge bis hin zum schlimmsten Fall: der Umkehrung der Vorzeichen der Koeffizienten.
Zusätzlich können fälschlicherweise Zusammenhänge aufgedeckt werden, die in den realen Daten ohne Messfehler überhaupt nicht existieren würden.

2.5.1 Einfache lineare Regression mit additivem Fehler

Diese Regression weißt genau den „double whammy“der Messfehler auf, welcher in der Motivation bereits angedeutet wurde, nämlich den Verlust statistischer Aussagekraft beim Testen und Verzerrungen in der Parameterschätzung.
In Abbildung 4 sind diese Effekte zu erkennen. In der linken Grafik werden die fehlerfreien Variablen (Y, X) abgebildet, welche durch das lineare Regressionsmodell

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

berechnet werden.
Die rechte Grafik bildet die fehlerhaften Daten (Y, W) ab mit $W = X + U$.
Dies wird als klassisches additives Messfehlermodell bezeichnet.

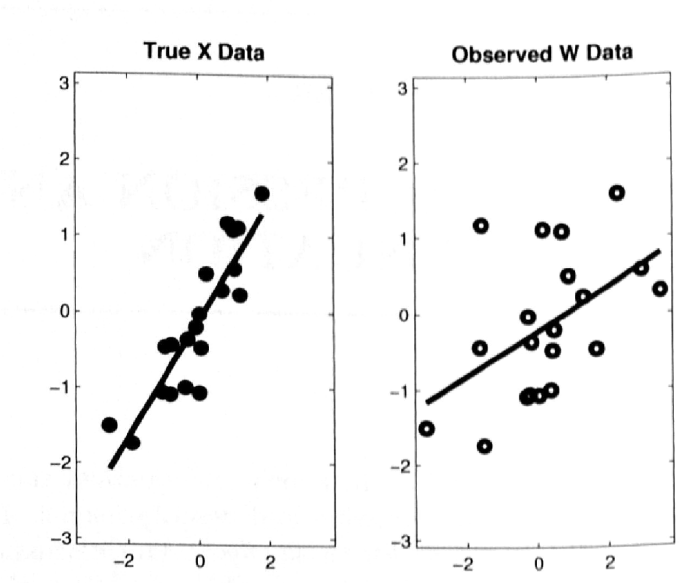

Abbildung 4: Unterschiede der Regression auf Y mit fehlerfreiem Wert X (links) und klassischem additivem Messfehler W (rechts). Quelle: (1, S.42)

Die Grafik zeigt, dass in der rechten Abbildung eine größere Variabilität der Daten vorhanden ist und auch die Steigung der Regressionsgeraden niedriger ausfällt als in der linken Abbildung mit den fehlerfreien Daten.
Dies zeigt den Verlust statistischer Aussagekraft durch additive Variabilität.
Auch zeigt sich die Verzerrung der KQ-Geraden durch den klassischen Messfehler.
Diese Probleme des einfachen linearen Regressionsmodels mit additivem klassischem

Fehler lassen sich wie folgt begründen:
Bei einer Regression von W auf Y wird nicht β_X geschätzt, sondern $\beta_{X^*} = \lambda\beta_X$.
Dieses λ, welches auch „reliability ratio“genannt wird, ergibt sich aus

$$\lambda = \frac{\sigma_X^2}{\sigma_X^2 + \sigma_U^2}$$

.
Kann λ geschätzt werden, so kann man auch β_X berechnen. (1, S.42 ff.)

2.5.2 Einfache Lineare Regression mit Berkson-Fehler

Auch bei einer einfachen linearen Regression mit erwartungstreuen Berkson-Fehler wird die Regressionsgleichung

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

verwendet.
Der Unterschied liegt hier darin, dass $X_i = W_i + U_i$ ist.
Daraus folgt, dass $E(X_i|W_i) = W_i$, womit sich ergibt, dass

$$E(Y_i|W_i) = \beta_0 + \beta_X W_i$$

.
Die Konsequenz daraus ist, dass der Schätzer, der Y_i über W_i bezieht, erwartungstreu für β_0 und β_X ist.
Dies stellt den großen Unterschied zur einfachen linearen Regression mit klassischem additivem Fehler dar.
Diese Erwartungstreue kann man auch in Abbildung 5 erkennen:

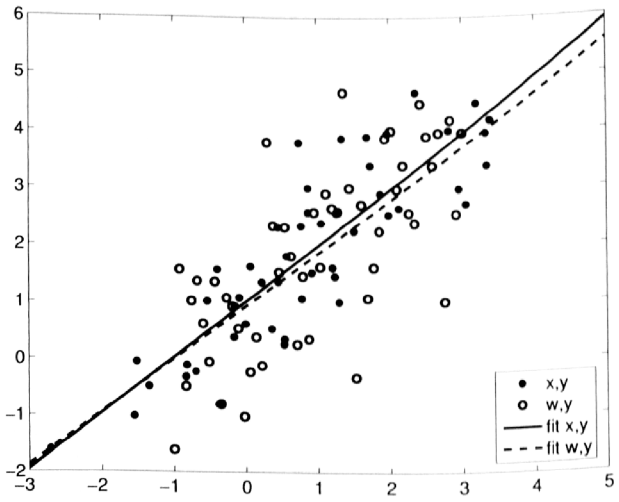

Abbildung 5: Unterschied der einfachen linearen Regression auf Y mit fehlerfreier Variablen X und mit Berkson-Fehler gemessener Variablen W. Die fehlerfreien Werte sind hier schwarz dargestellt. Die wahre Regressionsgerade ist durchgezogen. Quelle: (1, S.46)

Die Werte, die durch die schwarzen Punkte dargestellt sind die fehlerfreien Daten (Y, X). Wohingegen die weißen Punkte (Y, W) einen Berkson-Fehler aufweisen.
Die gestrichelte Regressionsgerade wurde mit Berkson-Fehler geschätzt. Die zwei Anpassungen sind sich also ähnlich. (1, S.45 f.)

3 fehlende Daten

In den Sozialwissenschaften, in denen die Daten, wie erwähnt, meist durch Interviews in Erfahrung gebracht werden, entstehen fehlende Daten meist aus dem Unvermögen eine Antwort aller Individuen einer Studie zu erhalten. Man geht davon aus, dass zwischen 25 - 30% fehlender Werte in einer sozialwissenschaftlichen Befragung typisch sind (5).
Diese fehlenden Daten können ein Problem bei der weiteren Analyse darstellen, da unter Umständen verzerrte Schätzer berechnet werden. Im folgenden Teil dieser Seminararbeit wird daher kurz auf die Ursachen fehlender Daten eingegangen, sowie auf verschiedene Möglichkeiten im Umgang mit fehlenden Daten aufmerksam gemacht.

3.1 Ursachen

Zunächst werden die verschiedenen Ursachen der fehlenden Werte bzw. Antwortausfälle dargestellt. Diese Antwortausfälle werden differenziert in „Unit-Nonresponse“und „Item-Nonresponse“.
Unit-Nonresponse liegt vor, wenn eine Untersuchungseinheit komplett ausfällt und für diesen Fall keinerlei Daten zur Verfügung stehen. Dies tritt zum Beispiel auf, wenn Adressen nicht aufgefunden werden, Personen durch eine Krankheit an der Teilnahme gehindert werden oder die Teilnahme komplett verweigert wird. Im Normallfall werden diese Ausfälle dann durch eine neue zusätzliche Untersuchungseinheit mit ähnlichen Quotenmerkmalen ersetzt.
Im Fall des Item-Nonresponse fehlen nur eine oder mehrere Variablen einer Untersuchungseinheit, wenn z.B. der Interviewer versehentlich eine Frage überspringt bzw. die teilnehmende Person eine Frage im Fragebogen übersieht.
Problematischer gestalten sich die Fälle, in denen die Teilnehmer eine Frage mit „Weiß nicht“beantworteten oder die Antwort explizit verweigerten (5).
Hängen diese „Antworten“explizit mit dem zu untersuchenden Merkmal zusammen, wenn beispielsweise vermehrt Leute mit hohem Einkommen die Antwort verweigern, so würde eine einfacher Ausschluss dieser Verweigerungen zu verzerrten Schätzern führen.
Das Hauptproblem bei Analyse von Daten mit fehlenden Werten liegt im häufig auftretenden systematischen Zusammenhang zwischen den Ursachen für das Fehlen und den inhaltlichen Aspekten der Untersuchung (7, S.468) .

3.2 Missing Data Mechanismen

Diesem Problem hat sich u.a. Rubin 1976 in seinen Forschungen angenommen und entwickelte, um solche Fehlschlüsse zu vermeiden, verschiedene Missing Data Mecha-

nismen. Diesen Ausfallmechanismen sollen überprüfen, wodurch fehlende Daten zu erklären sind, so dass anschließend eine geeignete Behandlungsmethode für die Daten gewählt werden kann.
Um diese Missing Data Mechanismen richtig erklären zu können, wurde sich auf folgende Bezeichnungen geeinigt.

X: vollständig gemessene Kovariable
Y: unvollständige nonresponse-Variable

Die Einteilung in die Missing Data Mechanismen erfolgt nun folgendermaßen:

Die Daten werden als Missing Completely at Random (MCAR) bezeichnet, wenn das Auftreten eines fehlenden Werts in Y unabhängig von der Ausprägung von y_i selbst und den restlichen Variablen $X_1 - X_n$ im Datensatz ist. Die Daten fehlen hier also rein zufällig. Es gibt keine bekannte Variable von der die fehlenden Daten abhängen.

Desweiteren sind Daten Missing at random (MAR), wenn das Auftreten eines fehlenden Wertes in Y unabhängig von der Ausprägung von y_i selbst ist, aber abhängig von den restlichen Variablen $X_1 - X_n$ im Datensatz.
Dies bedeutet, dass das Fehlen von Daten durch eine andere Variable erklärt werden kann.
Der letzte Missing Data Mechanismus, auf den an dieser Stelle eingegangen wird, heißt Missing Not at Random (MNAR).
Bei diesem Mechanismus ist das Auftreten eines fehlenden Wertes in Y abhängig von der Ausprägung von y_i und unabhängig von $X_1 - X_n$.
Bei diesem Mechanismus kann das Fehlen der Variablen von keiner anderen Variablen vorhergesagt werden, außer von der Variablen selbst.Die Verzerrung der Analyseergebnisse ist hier am größten.

Statistische Analysen benötigen bei Vorliegen von MNAR sehr starke theoretische annahmen, bei MAR hinfallen diese.

Beispiel: Einkommen als nonresponse-Variable

Um diese essentielle Unterscheidung der Missing Data Mechanismen besser zu nachvollziehen zu können, wird diese anhand eines Beispiels genauer erkläutert.

In einer Umfrage wurden in der vollständigen Variablen X das Alter und in der teilweise unvollständigen Variable Y das Einkommen der untersuchten Personen erhoben.

Werden die Daten als MCAR bezeichnet, dann hängt die Ausfallwahrscheinlichkeit weder von der Höhe des Einkommens ab, noch von der Variablen Alter, sondern sie fehlen zufällig.
Liegt nun MAR vor, so hängt die Ausfallwahrscheinlichkeit ebenfalls nicht von der Höhe des Einkommens ab. Sie hängt aber von der Höhe des Alters ab. Dies tritt ein, wenn z.B. die das Einkommen verhältnismäßig oft bei älteren Personen fehlt.
Klassifiziert man die vorliegenden Daten als NMAR, so ist die Ausfallwahrscheinlichkeit zwar unabhängig von der Variablen Alter, jedoch abhängig von der Höhe des Einkommens. Personen verweigern ab einem gewissen Schwellenwert mit höherer Wahrscheinlichkeit die Antwort. Dies ist jedoch sehr schwer nachzuprüfen (7, S.469).

Es ist unbedingt notwendig, diese Einordnungen in die verschiedenen Missing-Data-Mechanismen vor der Behandlung der fehlenden Daten zu treffen, da je nach MD-Mechanismus verschiedene Behandlungsmöglichkeiten für die fehlenden Daten zu treffen sind.
Würden die MD-Mechanismen ignoriert werden und die fehlenden Daten womöglich falsch behandelt, kann es zu einer gravierenden Verzerrung der Daten bzw. der späteren Analyseergebnisse kommen.

3.3 Umgang mit fehlenden Daten

Bisher haben wir uns mit der Klassifizierung der Missing-Data-Mechanismen beschäftigt. Diese sind eine Vorbedingung für den nun folgenden Teil, in dem verschiedene Möglichkeiten im Umgang mit fehlenden Daten dargestellt werden (4, S.13-17).

3.3.1 Eliminierungsverfahren

Eine einfache Möglichkeit, mit fehlenden Daten umzugehen, stellen die verschiedenen Eliminierungsverfahren dar. Hierbei werden fehlende Merkmale oder Objekte einfach aus der weiteren Analyse ausgeschlossen.
Wichtig bei diesen Verfahren ist, dass die fehlenden Werte unbedingt nach dem Mechanismus MCAR fehlen müssen. Denn nur wenn MCAR vorliegt, wird gewährleistet, dass durch diese Methode lediglich die Stichprobe kleiner wird.
Beachtet man diese Vorgabe nicht, so können erhebliche Verzerrungen in den Ergebnissen bzw. der Datenmatrix entstehen (4, S.39 f.).

complete-case-analysis

Bei der „complete-case-analysis“ werden nur diejenigen Untersuchungsobjekte bzw. Merkmale weiterverwendet, die vollständig vorliegen. Das heißt, die anschließenden Analysen können mit einem vollständigen Datensatz durchgeführt werden.
Daraus ergibt sich auch ein entscheidender Vorteil dieser Methode, denn nun können auch alle auf vollständigen Daten basierenden multivariaten Standardverfahren ohne Modifizierungen durchgeführt werden.
Zusätzlich bietet diese Methode auch eine Vergleichbarkeit univariater Statistiken, da diese auf der selben Stichprobengröße basieren.
Nachteilig wirkt sich bei der „complete-case-analysis“ aus, dass ein großer Teil der Informationen verloren geht. Sind in der Datenmatrix viele fehlende Werte oder ist die Stichprobe relativ klein, so kann dieser Informationsverlust teilweise erheblich ausfallen. (4, S.40 f.).

available-case-analysis

Um diesen Nachteil „abzuschwächen“ wird meistens der Ansatz der „available-case-analysis“ gewählt.
Hierbei werden für jede Auswertung einzeln all diejenigen Fälle ausgewählt, für welche der Messwert der interessierenden Variablen vorhanden ist.
Das heißt aber auch, dass nun bei nahezu jeder Berechnung unterschiedliche Stichprobengrößen hinzugezogen werden, so dass es nun schwerer ist, einen Vergleich der mit „available-case-analysis“ berechneten Ergebnisse durchzuführen.
Auch gestaltet sich die sinnvolle Auswertung der Daten nun ungemein schwerer, wenn Variablen mit Klasseneinteilungen betrachtet werden sollen. Hier ist es oft notwendig, jeweils eine feste Anzahl von Fällen in den einzelnen Klassen zu betrachten.
Nichtsdestotrotz wird die „available-case-analysis“ oft zur Schätzung von Mittelwert und Varianz verwendet. Für Kovarianzen oder Korrelationen wird die paarweise „available-case-analysis“ verwendet, bei der nur paarweise vorliegende Werte benutzt werden (4, S.41 f.).

3.3.2 Imputationsverfahren

Imputationsverfahren werden verwendet, um jeden fehlenden Wert im Datensatz mit einem Schätzwert zu ersetzen. Der entscheidende Vorteil hierbei ist, dass man anschließend mit einem vollständigen Datensatz weiterarbeiten kann.
Dabei wirkt sich positiv aus, dass auch die Informationen der fehlenden Daten für die weitere Analyse genutzt werden können. Dies ist bei der bloßen Eliminierung der

Daten nicht möglich (2, S.42).

einfache Imputationsmethoden

Bei den einfachen Imputationsmethoden wird jeder fehlende Wert mit genau einem Schätzwert ersetzt. Dies stellt also eine relativ einfache Methode dar, mit fehlenden Werten umzugehen.
Im Folgenden werden hier kurz die Imputation durch das arithmetische Mittel, sowie die Imputation durch Regression skizziert:

Imputation durch arithmetisches Mittel

Die Imputation durch das arithmetische Mittel ersetzt die fehlenden Werte durch das arithmetische Mittel der verfügbaren Fälle.
Dadurch wird natürlich auch die Variabilität der Daten verringert, was zu einer Unterschätzung der Varianz bzw. Standardabweichung führt.
Zudem verringert die Mittelwertsimputation auch die Höhe Zusammenhangsmaße, wie Korrelationen und Kovarianzen, da sie dem Datensatz Werte hinzufügt, die unkorreliert mit anderen Variablen im Datensatz sind.

Als Beispiel ist in Abbildung 6 die Regression durch das arithmetische Mittel dargestellt. Dabei wurden alle Mitarbeiter eines Büros einem IQ-Test unterzogen. Die Arbeitsleistung wurde nicht bei allen Arbeitern erhoben. Die fehlenden Werte werden durch das arithmetische Mittel der Arbeitsleistung ersetzt. (2, S.42 f.).

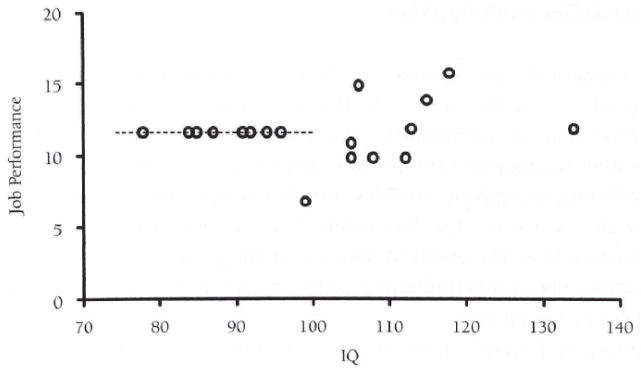

Abbildung 6: Imputation der Variable Arbeitsleistung durch das arithmetische Mittel. Quelle: (2, S.43)

Imputation durch Regression

Eine andere Methode der einfachen Imputation ist die Imputation durch Regression. Die Idee hinter dieser Methode ist, dass die Informationen des kompletten Datensatzes genutzt werden, um die fehlenden Daten zu schätzen.
Es wird sich also der Annahme bedient, dass die Variablen eines Datensatzes einen Hang zur Korrelation haben. Es wird also eine Regression mit dem fehlenden Wert als Responsevariable durchgeführt.
Bei einer Kovariablen X lautet die Formel für die Imputation fehlender Werte der Variablen Y folgendermaßen:

$$y_i = \beta_0 + \beta_1 x_i$$

Aus den vorhandenen Werten von X lassen sich so die Regressionskoeffizienten β_0 und β_1 berechnen.
Bei der stochastischen Regression Imputation wird zusätzlich noch ein normalverteiltes Residuum ϵ berechnet und auf die Regressionsgleichung addiert. Dadurch wird die Variabilität der Daten vergrößert bzw. die Verzerrung vermindert.

Zur Veranschaulichung wird nun das Beispiel der Mittelwertsimputation wieder aufgegriffen, bei dem die Variable Arbeitsleistung imputiert wurde. Dies geschieht nun mithilfe der Imputation durch Regression (2, S.44-48).

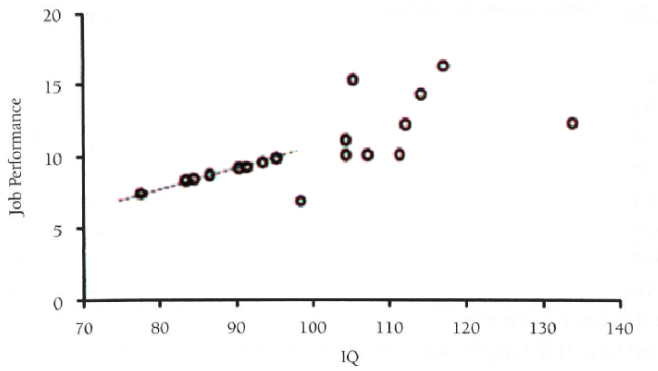

Abbildung 7: Imputation der Variable Arbeitsleistung durch Regression. Quelle: (2, S.46)

Alle beide dieser einfachen Imputationsmethoden erzielen sie jedoch oftmals verzerrte Parameterschätzungen, sogar wenn die Daten nach MCAR fehlen.
Deshalb ist die Verwendung dieser Methoden meist nicht ganz ideal und es wird eher zur Verwendung der multiplen Imputation geraten, bei der dieser Nachteil nicht auftritt.

multiple Imputation

Im Unterschied zu den einfachen Imputationsmethoden werden die fehlenden Werte bei der multiplen Imputation durch einen Vektor der Länge $m \geq 2$ geschätzt. Aus diesen Vektoren entstehen dann m vervollständigte Datensätze.
Zur weiteren Analyse der Datenmatrix können diese Datensätze kombiniert werden (4, S.255). Diese Imputationsmethode wird durchgeführt, wenn die Daten MAR und multivariat normalverteilt sind.
Die multiple Imputation besteht aus 3 verschiedenen Phasen: der imputation phase, der analysis phase und der pooling phase.
Die imputation phase erstellt m multiple Kopien des Datensatzes, von denen jede verschiedene Schätzungen der fehlenden Werte enthält.
Das heißt, es werden für jeden fehlenden Wert genau m Werte geschätzt, die in der Regel verschiedene Werte annehmen. Dadurch wird die Variabilität, die die Daten aufweisen, entsprechend dargestellt.
In der analysis phase werden dieselben Analysemethoden angewendet, die man auch für einen vollständigen Datensatz anwenden würde. Der Unterschied hierbei ist lediglich, dass jede dieser Methoden m Mal, sprich einmal für jeden vervollständigten Datensatz durchgeführt wird.
In der pooling phase werden dann anschließend die m Ergebnisse bspw. der Parameterschätzer oder der Standardfehler zu jeweils einem Ergebnis kombiniert.
Dieser Vorgang ist in Abbildung 8 dargestellt:

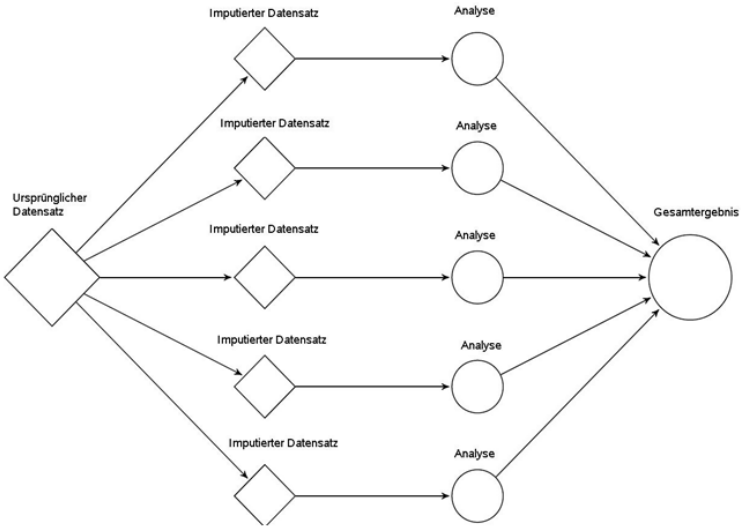

Abbildung 8: Multiple Imputation mit $\mathtt{m} = 5$

Die Vorteile dieser Imputationsmethode sind zum Einen, dass der Bias der Schätzungen im Vergleich zu den singulären Imputationsmethoden deutlich reduziert ist.
Zusätzlich wird die Unsicherheit der Imputation durch die Variabilität in den m vervollständigten Datensätzen widergespiegelt.
Als Nachteil ist aufzuführen, dass bei der multiplen Imputation eine größere Rechenzeit von Nöten ist, da mehrere Imputationen durchgeführt werden. Im Zeitalter schneller Rechner ist dies aber eher zu vernachlässigen (2, S.188 ff.)

Literatur

[1] Raymond J. Carroll. Measurement error in nonlinear models: A modern perspective. Chapman & Hall, Taylor & Francis, Boca Raton [etc.], 2nd ed., [rev.] edition, op. 2006. ISBN 1584886331.

[2] Craig K. Enders. Applied missing data analysis. Guilford Press, New York, 2010. ISBN 1606236393.

[3] Richard Költringer. Gültigkeit von Umfragedaten. Böhlau, Wien, 1993. ISBN 3-205-98114-6.

[4] Little, Roderick J. A and Donald B. Rubin. Statistical analysis with missing data. Wiley, New York, 1987. ISBN 0-471-80254-9.

[5] Nicoletti Cheti, Peracchi Franco, and Foliano Francesca. Estimating Income Poverty in the Presence of Missing Data and Measurement Error. Journal of Business & Economic Statistics, (29):61–72, 2011.

[6] H. Schneeweiss and Hans-Joachim Mittag. Lineare Modelle mit fehlerbehafteten Daten. Physica-Verlag, Heidelberg, 1986. ISBN 3-7908-0320-0.

[7] Rainer Schnell, Paul Bernhard Hill, and Elke Esser. Methoden der empirischen Sozialforschung. Oldenbourg, München [u.a.], 8., unveränd. aufl edition, 2008. ISBN 978-3-486-58708-1.