Überblick über Messfehler und ihre Auswirkungen in der linearen Regression

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

Hanna Marshalava

Institut für Statistik, LMU

November 21, 2014

Übersicht

Einleitung

Arten der Messfehler
- Stochastische und systematische Messfehler
- Additive und multiplikative Fehler
- Differentielle und nicht-differentielle Fehler
- Klassische und Berkson Fehler

Messfehler in der linearen Regression
- Systematische Messfehler in der linearen Regression
- Stochastische Messfehler in der Regression
  - Klassische Messfehler in der Regression
  - Berkson-Fehler in der Regression
- Kombination von klassischen und Berkson-Fehlern

Korrektur der Abweichung

Einleitung

Arten der Messfehler
- Stochastische und systematische Messfehler
- Additive und multiplikative Fehler
- Differentielle und nicht-differentielle Fehler
- Klassische und Berkson Fehler

Messfehler in der linearen Regression
- Systematische Messfehler in der linearen Regression
- Stochastische Messfehler in der Regression
  - Klassische Messfehler in der Regression
  - Berkson-Fehler in der Regression
- Kombination von klassischen und Berkson-Fehlern

Korrektur der Abweichung

Messfehler

Definition

Ein Messfehler oder **eine Messabweichung** ist die Abweichung eines aus Messungen gewonnenen Wertes vom wahren Wert der Messgröße.

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.

Beispiel: Übersicht der Messungen von der Astronomischen Einheit über die Jahre

Ursachen des Auftretens von Messfehlern

- Messgeräteabweichungen als Folge der Unvollkommenheit der Konstruktion, Fertigung, Justierung (z. B. durch Werkstoffe, Fertigungstoleranzen)
- Umwelteinflüsse als Folge von Änderungen der Einwirkungen aus der Umgebung (z. B. Temperatur, äußere elektrische oder magnetische Felder, Lage, Erschütterungen)
- Instabilitäten des Wertes der Messgröße oder des Trägers der Messgröße (z. B. statistische Vorgänge, Rauschen)
- Beobachtereinflüsse infolge unterschiedlicher Eigenschaften und Fähigkeiten des Menschen (z. B. Aufmerksamkeit, Übung, Sehschärfe, Schätzvermögen)

Einleitung

Arten der Messfehler
- Stochastische und systematische Messfehler
- Additive und multiplikative Fehler
- Differentielle und nicht-differentielle Fehler
- Klassische und Berkson Fehler

Messfehler in der linearen Regression
- Systematische Messfehler in der linearen Regression
- Stochastische Messfehler in der Regression
  - Klassische Messfehler in der Regression
  - Berkson-Fehler in der Regression
- Kombination von klassischen und Berkson-Fehlern

Korrektur der Abweichung

Stochastische und systematische Messfehler

Zufällige bzw. Stochastische Fehler

- hervorgerufen von messtechnisch nicht erfassbaren Änderungen der Messgeräte, des Messgegenstandes, der Umwelt und der Beobachter
- können bei einer Einzelmessung weder nach Betrag noch nach Vorzeichen bestimmt werden
- sind nicht zu korrigieren und machen das Ergebnis unsicher

Systematische bzw. statistische Fehler

- hervorgerufen durch Unvollkommenheiten der Messgeräte, der Messverfahren und des Messgegenstandes, messtechnisch erfassbaren Einflüssen der Umwelt und persönlichen Einflüssen der Beobachter
- haben ein bestimmtes Vorzeichen ($+$ oder $-$)
- unter gleichen Bedingungen den gleichen Betrag
- besitzt bei gleichen Messvorgängen die gleiche Struktur
- lässt sich abschätzen, wenn Ergebnisse mehrerer wiederholter Messungen vorhanden sind

$x_1, ..., x_n$ Ergebnisse wiederholter Messungen einer wahren Variable μ sind als Realisationen einer $N(\mu^*, \sigma^2)$-verteilten Zufallsgröße X aufgefasst
Dann ist der **zufällige Fehler** der i-ten Messung

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

und der **systematische Fehler (Bias)**:

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

Daraus folgt

$$x_i = \mu + b + \epsilon_i$$

= wahrer Wert + systematischer Fehler + zufälliger Fehler.

Mit den Messwerten $x_1, ..., x_n$ können

- $\mu^* = \mu + b$ und σ geschätzt werden

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

- Konfidenzintervalle für μ^* und σ^2 angegeben werden, die einen Eindruck von der Größe und Struktur des zufälligen Fehlers vermitteln
- keine Informationen über die Größe des systematischen Fehlers b ermittelt werden, da dieser für alle n Werte gleich ist⇒**Vermeidung des systematischen Fehlers!**

Additive und multiplikative Fehler

Additiver oder konstanter systematische Fehler

Es seien U der systematische Messfehler und X der wahre (latente) Wert, dann gilt

$$U = f(X)$$

Bei den konstanten Funktionswerten für die beobachtete Variable gilt

$$W = X + U$$

Multiplikativer oder proportionaler systematischer Fehler

Der Fehler ist proportional zur systematischen Variable:

$$U = pX$$

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

$$W = X + pX = aX$$

Verallgemeinernd werden die beiden Fälle als **lineare systematische Fehler** bezeichnet, wenn gilt :

$$W = aX + U$$

Differentielle und nicht-differentielle Fehler

Nicht-differentielle Fehler

Wenn

- die interessierende Variable X nicht beobachtbar oder wegen fehlender Erhebungen nicht vorhanden ist
- eine ähnliche Variable W keine anderen Informationen über den Response Y als X und eine fehlerfrei gemessene Kovariable Z besitzt

dann ist W mit einem nicht-differentiellen Fehler behaftet und äquivalent zu X

$$\Rightarrow W \text{ ist das Surrogat}$$

Formal: die Verteilung von $Y|(X, W, Z)$ ist nur von (X, Z) abhängig

Differentielle Fehler

In allen anderen Fällen

Warum sind nicht-differentielle Fehler wichtig?

- Schätzung des Response-Parameters bei gegebenem wahren Wert der Einflussvariable ist auch möglich, wenn die wahre Variable X nicht beobachtbar ist.
  Z.B. Angst kann nicht direkt gemessen werden. Man misst die Steigung der Herzfrequenz (Surrogat).
- Vereinfachung der Untersuchung der linearen Regression und die Bestimmung der Regressionsparameter auch mit der fehlerbehafteten Variable W ist möglich

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

Klassische und Berkson Fehler

Klassischer Messfehler

W sei die fehlerbehaftete Messung des wahren latenten Wertes X und $W = X + U$ mit $U|X \sim N(0, \sigma^2)$, wobei U und X stochastisch unabhängig sind.

$$E(W|X) = X$$

W ist folglich eine unverzerrte Messung für X.

Berkson Fehler

Man betrachte $X = W + U$ mit $U|X \sim N(0, \sigma^2)$, wobei U und W stochastisch unabhängig sind.

$$E(X|W) = W$$

Unterschied zwischen klassischen und Berkson Messfehler

Klassischer Fehler

- fehleranfällige Variable wird mit Hilfe eines geeigneten Messmittels eindeutig bei einer Person gemessen
- Wiederholungsmöglichkeit der Messung besteht
- Z.B. Fragen zum Essverhalten, Messung des Blutdrucks

Berkson Fehler

- Mitglieder einer kleinen Gruppe machen Angaben zu einer fehleranfälligen Variablen
- Z.B. Bergarbeiter bei gleicher Beschäftigungsdauer zeigen das gleiche Staubbelastungsbild. Jedoch ist das wahre Bild bei jedem einzelnen Individuum anders.

Einleitung

Arten der Messfehler
- Stochastische und systematische Messfehler
- Additive und multiplikative Fehler
- Differentielle und nicht-differentielle Fehler
- Klassische und Berkson Fehler

Messfehler in der linearen Regression
- Systematische Messfehler in der linearen Regression
- Stochastische Messfehler in der Regression
  - Klassische Messfehler in der Regression
  - Berkson-Fehler in der Regression
- Kombination von klassischen und Berkson-Fehlern

Korrektur der Abweichung

Messfehler in der linearen Regression

- Möglichst genaue Schätzungen mit der fehlerbehafteten beobachtbaren Variablen W, da fehlerfreie Variable X latent
- Darstellung des Effekts des Messfehlers in einem linearen Modell unter Berücksichtigung anderer Variablen und deren Verteilung
- Berücksichtigung der Verteilung des Messfehlers bei der Erforschung seiner Auswirkung auf ein lineares Modell

Systematische Messfehler in der linearen Regression

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

W sei die fehlerbehaftete Variable und $W = X + d$, wobei X die wahre Variable und d ein konstanter Messfehler ist. Dann gilt

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

Durch Umformung ergibt sich

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

Parameterschätzer

- $\beta_0^* = \beta_0 + \beta_X d$ verzerrt, inkonsistent
- $\beta_X^* = \beta_X$ konsistent geschätzt
- die beobachteten Werte sind um einen konstanten Wert d nach rechts und in Folge um den Wert $\beta_X d$ nach oben verschoben

Proportionaler Messfehler in der linearen Regression

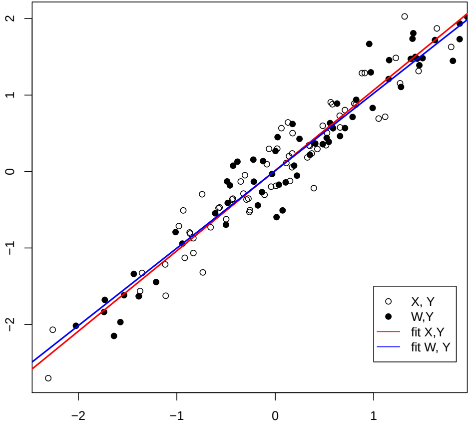
$W = cX$ sei die fehlerbehaftete Variable, es gilt

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

- β_0^* und $\beta_X^* = \frac{1}{c}\beta_X$ verzerrte KQ-Schätzer

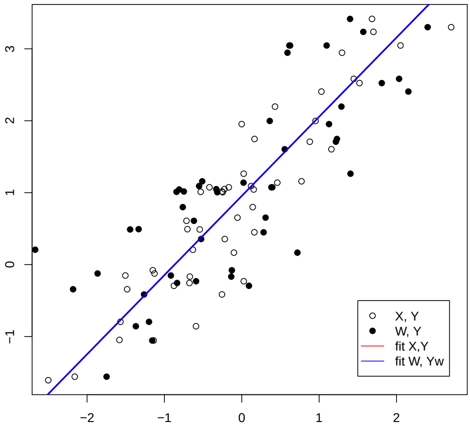
- Durch KQ-Schätzung verzerrte Schätzer können wegen dem bekannten konstanten Fehler d und dem proportionalen Fehler c korrigiert werden.
- Falls die systematischen Fehler bekannt sind, können die gemessenen, fehlerbehafteten Werte W vor Durchführung der Regression korrigiert werden.

Klassische Messfehler in der Regression



- $Y = \beta_0^* + \beta_W(X + U) + \epsilon$
- systematische Unterschätzung der Steigung
- KQ-Schätzung von X auf Y inkonsistent für β_X
- $\beta_{x*} = \lambda\beta_x$, wobei
$$\lambda = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2} < 1$$
- W besitzt schwächeren Einfluss auf die Response als X
- größere Varianz der Beobachtungen $var(Y|W) = \sigma_\epsilon^2 + \lambda\beta_x^2\sigma_u^2$

Berkson-Fehler in der Regression



- W ist das Surrogat
- $X = W + U$ und $E(X|W) = W$
- $Y = \beta_0^* + \beta_W W + \epsilon$
- $E(Y|W) = \beta_0 + \beta_X E(X|W) = \beta_0 + \beta_x W$
- Koeffizientenschätzer für β_0 und $\beta_X = \beta_W$ sind unverzerrt
- $Var(Y|W) = \sigma_\epsilon^2 + \beta_X^2 \sigma_U^2$

Kombination von klassischen und Berkson-Fehlern

Man betrachte ein Regressionsmodell, bei dem die Einflussvariable sowohl eine klassische als auch eine Berkson Komponente beinhaltet

- klassisches Fehlermodell mit $W = X + U$
- $X = \lambda W + (1-\lambda)E(U) + U^*$ Prädiktor für $(X|W)$ mit $U^* = (1-\lambda)(X - E(X)) - \lambda U$ und $\lambda = \frac{\sigma_X^2}{\sigma_X^2 + \sigma_U^2} < 1$
- U^* und W unkorreliert $\Rightarrow$ überführen folglich den klassischen Fehler in ein Berkson-Fehler
- Darstellung der Fehlerstruktur durch Berkson-Modell als stochastischer Fehler, verzerrt durch den systematischen, proportionalen Fehler λ

Einleitung

Arten der Messfehler
- Stochastische und systematische Messfehler
- Additive und multiplikative Fehler
- Differentielle und nicht-differentielle Fehler
- Klassische und Berkson Fehler

Messfehler in der linearen Regression
- Systematische Messfehler in der linearen Regression
- Stochastische Messfehler in der Regression
  - Klassische Messfehler in der Regression
  - Berkson-Fehler in der Regression
- Kombination von klassischen und Berkson-Fehlern

Korrektur der Abweichung

Korrektur der Abweichung

- KQ-Schätzer in der einfachen linearen Regression bei dem klassischen additiven Messfehlermodell: $\lambda\beta_x$
- $\lambda = \frac{\sigma_x^2}{\sigma_x^2+\sigma_u^2}$ Attenuation-Koeffizient
- ist λ bekannt, dann ist der unverzerrte geschätzter Wert β_x durch Multiplikation mit $\frac{1}{\lambda}$ ermittelbar
- Schätzung bei unbekanntem λ
- ist $\widehat{\sigma_u^2}$ die Schätzung der Messfehlervarianz und $\widehat{\sigma_w^2}$ die Stichprobenvarianz von W, dann ist die konsistente Schätzung des Attenuation-Koeffizienten
$$\widehat{\lambda} = (\widehat{\sigma_w^2} - \widehat{\sigma_u^2})/\widehat{\sigma_w^2}.$$
- resultierende Schätzung ist $\widehat{\beta_{x*}}/\widehat{\lambda}$

Orthogonale Regression

- Man betrachte das lineare Modell $Y = \beta_0 + \beta_x X + \epsilon$ und $W = X + U$, wobei ϵ und U unkorreliert sind.
  Bei der orthogonalen Regression werden die kürzesten Abstände zur Regressionsgerade gewichtet mit eine Faktor $\eta = \sigma_e^2/\sigma_u^2$ **minimiert**, d.h.
  $$\sum (Y_i - \beta_0 - \beta_1 X_i)^2 + \eta (W_i - X_i)^2$$
- Für die Orthogonale Regression muss der Parameter $\eta = \sigma_e^2/\sigma_u^2$ bekannt sein oder geschätzt werden
- Wenn $\eta = 1$, dann minimiert der Regressionsschätzer den orthogonalen Abstand von (Y, W) zur Geraden $y = \beta_0 + \beta_y x$.
- Unter der Annahme, dass $(X_1, \ldots, X_n)$ unbekannte feste Konstanten und die Fehler (ϵ, U) unabhängig und normalverteilt sind, ist der Orthogonale Regressionsschätzer der funktionale Maximum-Likelihood-Schätzer.
- η kann nicht richtig spezifiziert oder geschätzt werden.
- ein falsch spezifizierter Wert η in der orthogonalen Regression ergibt oft eine unakzeptabel große Überkorrektur, die eine Abschwächung des Messfehlers verursacht.

Vielen Dank für Ihre Aufmerksamkeit!

Quellenverzeichnis

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

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

Carrol, R., Ruppert, D., Stefanski, L. and Crainiceanu, C. (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.

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

Anhang

Orthogonale Regression