Fehler in der abhängigen Variablen

Bachelor- Seminar
im Wintersemester 2014/ 2015

Nina Markovic

Institut für Statistik, LMU

19. Dezember 2014

Übersicht

Messfehler in der Response Varaible

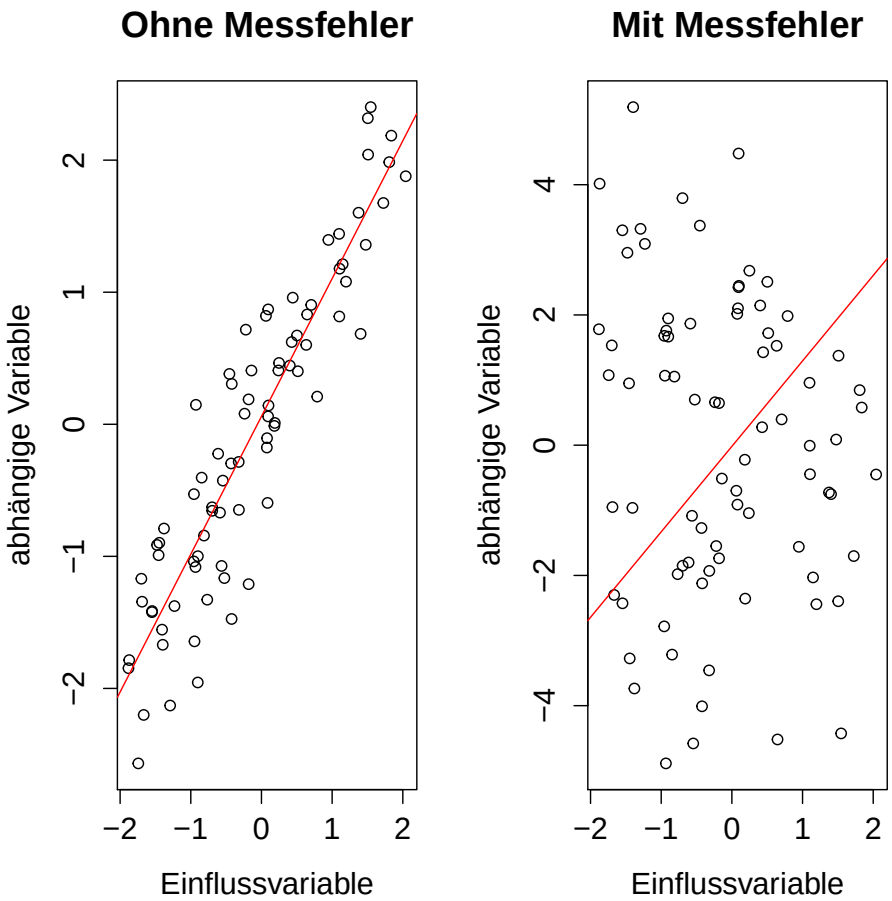

Abbildung: Eine Simulation von einem additiven Messfehler

Messfehler in der Response Varaible

- größere Streuung der Punkte im Modell mit Messfehler
- Regressionsgerade nicht mehr identisch

Messfehler in der Response Varaible

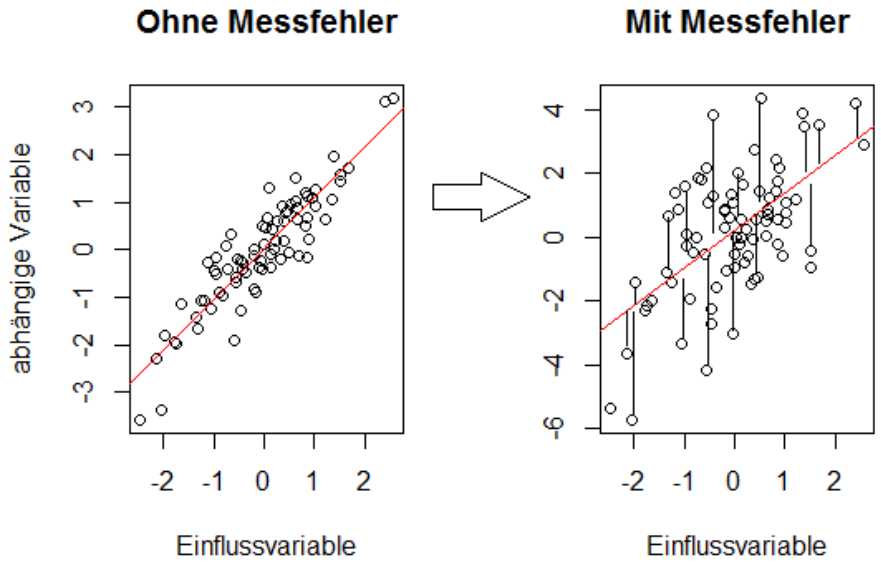

Abbildung: Vergleich der Modelle mit und ohne Messfehler

Nach Abrevaya und Hausman (2004) werden Messfehler in der Response Variablen über Residuen absorbiert

Messfehler in der Response Varaible

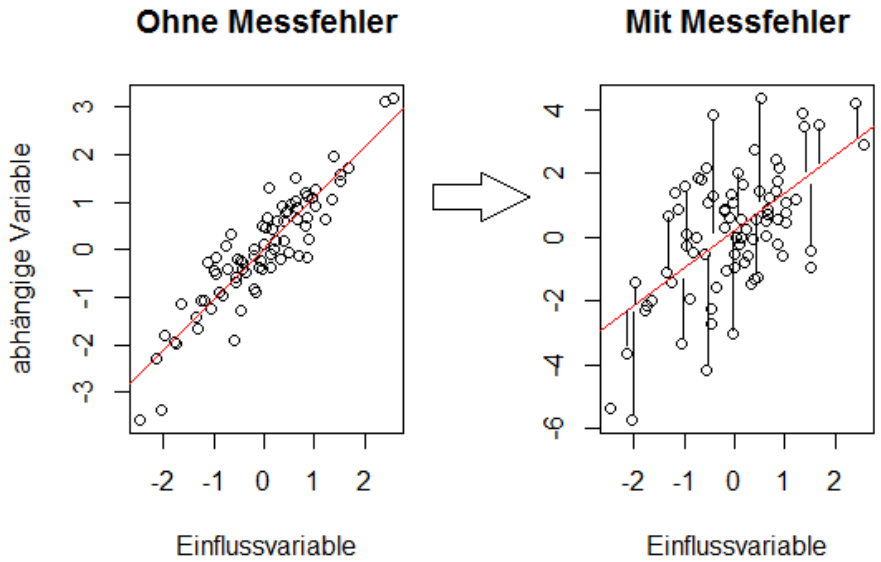

Abbildung: Vergleich der Modelle mit und ohne Messfehler

Nach Abrevaya und Hausman (2004) werden Messfehler in der Response Variablen über Residuen absorbiert

$\Rightarrow$ Messfehler darf ignoriert werden!?

Auswirkungen beim linearen Zusammenhang

Wie ist passt sich die neue Regressionsgerade an die Daten an?

Auswirkungen beim linearen Zusammenhang

Wie ist passt sich die neue Regressionsgerade an die Daten an?
$\Rightarrow$ Vergleich erfolgt am R^2 der beiden Modelle

R^2 eines Models ohne Messfehler	R^2 eines Models mit Messfehler
0.8032	0.3102

Tabelle: Vergleich am R^2

$\Rightarrow$ R^2 ist bei dem Modell ohne Messfehler deutlich höher

Auswirkungen beim linearen Zusammenhang

Inwieweit verändern sich die Regressionsgeraden?

Auswirkungen beim linearen Zusammenhang

Inwieweit verändern sich die Regressionsgeraden?

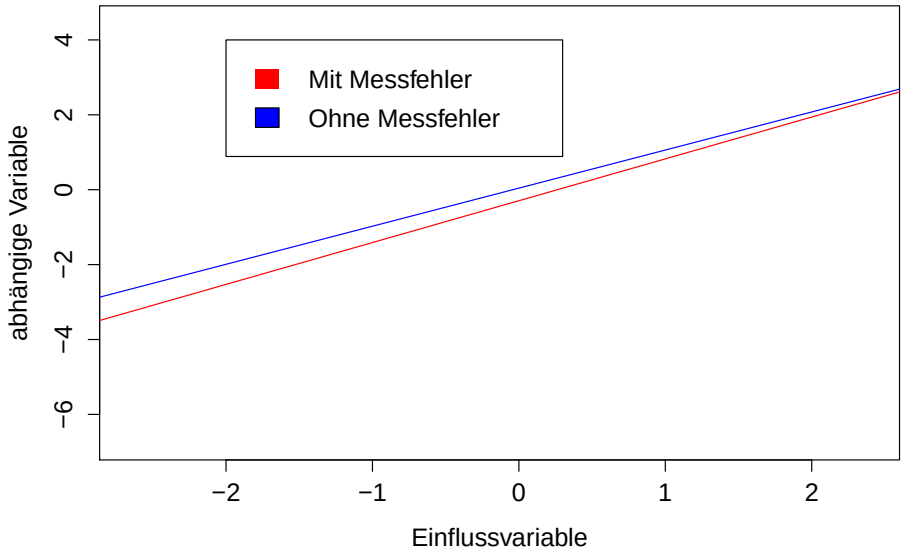

Abbildung: Vergleich der Regressionsgeraden

$\Rightarrow$ Regressionsgeraden stimmen nicht mehr überein

Auswirkungen beim linearen Zusammenhang

Inwieweit verändern sich die Regressionsgeraden?

Auswirkungen beim linearen Zusammenhang

Inwieweit verändern sich die Regressionsgeraden?

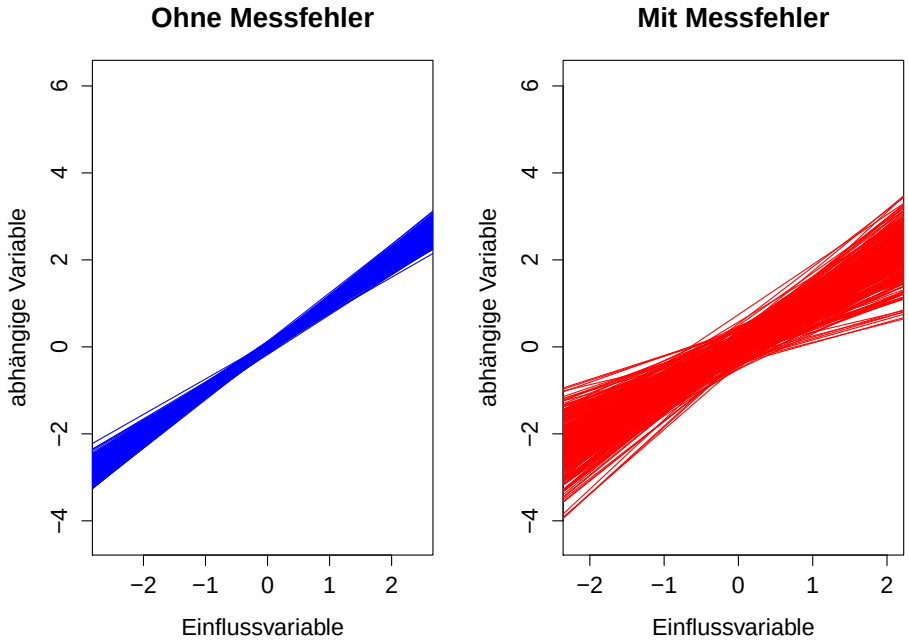

Abbildung: 350 simulierte Regressionsgeraden für jeweils ein Modell mit und ohne Messfehler.

$\Rightarrow$ Regressionsgeraden im Modell mit Messfehler variieren stärker

Auswirkungen beim linearen Zusammenhang

Hat ein Messfehler Modellverletzungen zu Folgen?

Auswirkungen beim linearen Zusammenhang

Hat ein Messfehler Modellverletzungen zu Folgen?

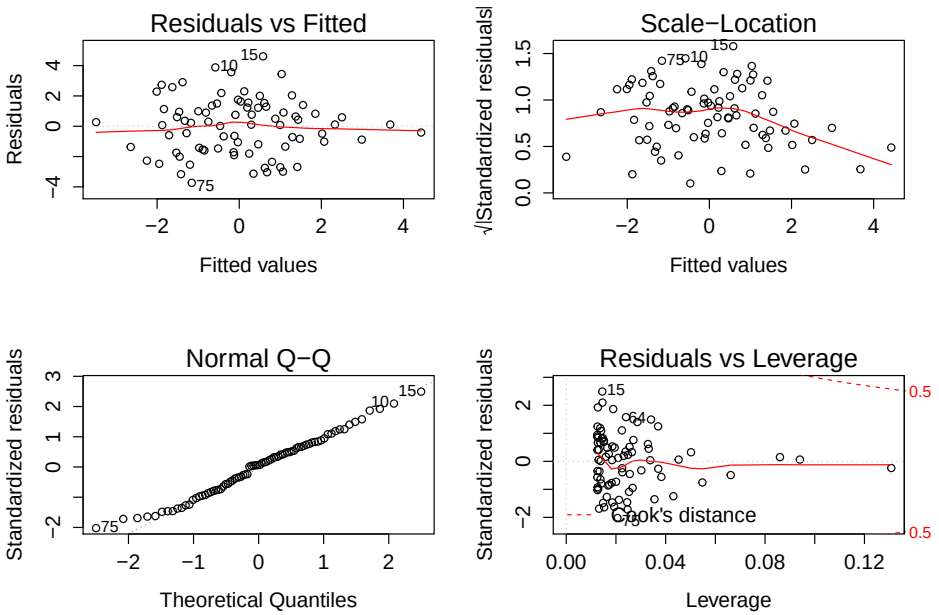

Abbildung: Diagnostik Plots eines Models mit Response- Messfehler.

$\Rightarrow$ grobe Modellverletzungen, wie Heteroskedastizität oder Nicht-Normalverteilung der Residuen nicht erkennbar

Auswirkungen beim nicht linearen Zusammenhang

Zusammenhang zwischen Einfluss- und Zielvariable beispielweise quadratisch:

$$Y_i = \beta_0 + \beta_1 Z_i + \beta_2 Z_i^2 + \epsilon_i$$

Auswirkungen beim nicht linearen Zusammenhang

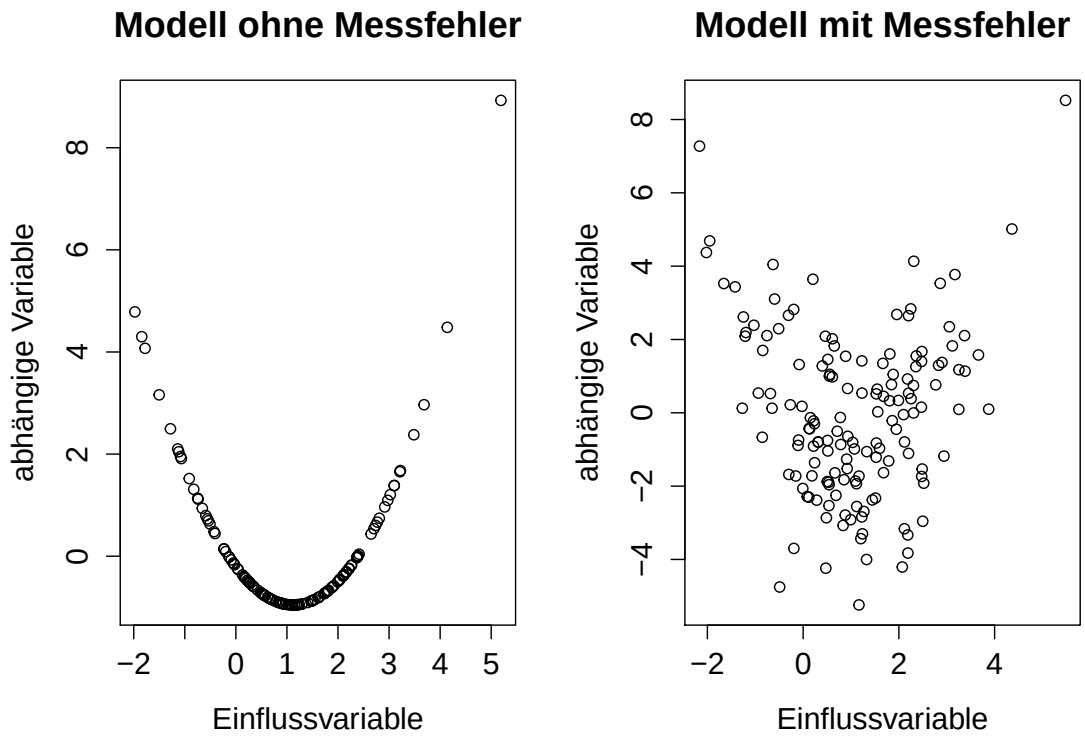

Abbildung: quadratischer Zusammenhang mit und ohne Messfehler

$\Rightarrow$ In dem Modell mit Messfehler ist der quadratischer Zusammenhang $(Y = \beta_0 + \beta_1 Z + \beta_2 Z^2)$ nicht mehr erkennbar

Auswirkungen beim nicht linearen Zusammenhang

Folgen:

- Vermutung über einen linearen Zusammenhang
- quadratische Einflussgröße Z^2 wird nicht mehr in das Modell mitaufgenommen
- Modellverletzungen schwer erkennbar

Auswirkungen beim nicht linearen Zusammenhang

Beispiel am Normal Q- Q Plot:

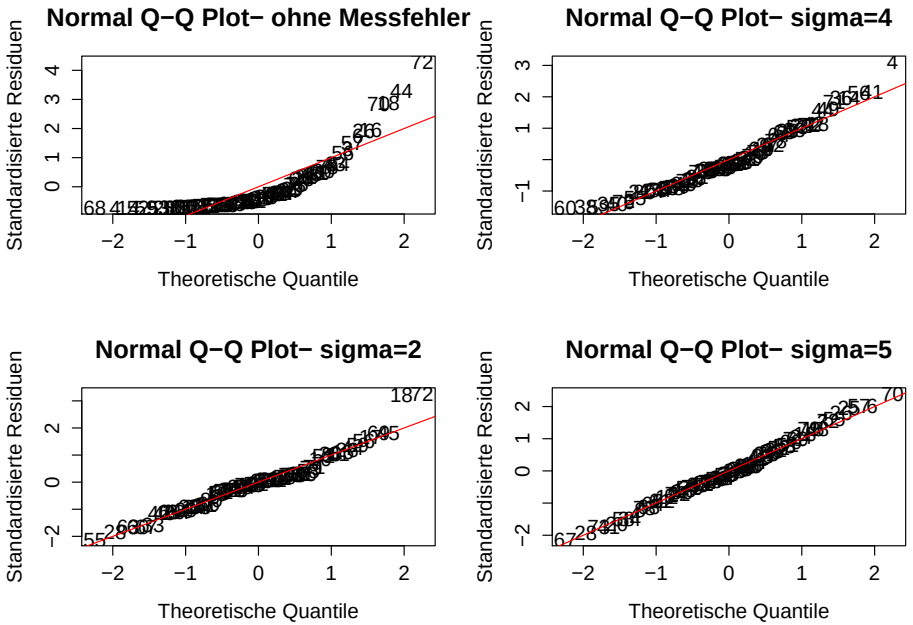

Abbildung: Normal Q-Q Plot mit verschiedenen Varianzen des Messfehlers

$\Rightarrow$ Je höher die Varianz des Messfehlers, desto weniger sind Modellverletzungen zu erkennen

Auswirkungen beim nicht linearen Zusammenhang

weitere Folgen nach Carroll et al. (2006, S.341):

- Inferenz bei strengen nicht- linearen Regressionsmodellen basiert auf einer Approximation durch ein lineares Modell: z.B. durch Entwicklung einer Taylor Reihe für β um den wahren Wert β_0:
  $Y_i = m_Y(Z_i, \beta) + \epsilon_i \approx m_Y(Z_i, \beta_0) + f'(Z_i, \beta_0)(\beta - \beta_0) + \epsilon_i$

Auswirkungen beim nicht linearen Zusammenhang

weitere Folgen nach Carroll et al. (2006, S.341):

- Inferenz bei strengen nicht- linearen Regressionsmodellen basiert auf einer Approximation durch ein lineares Modell: z.B. durch Entwicklung einer Taylor Reihe für β um den wahren Wert β_0:
  $Y_i = m_Y(Z_i, \beta) + \epsilon_i \approx m_Y(Z_i, \beta_0) + f'(Z_i, \beta_0)(\beta - \beta_0) + \epsilon_i$
- Fehler in der Taylor Approximation geht gegen Null, wenn sich β an β_0 nähert

Auswirkungen beim nicht linearen Zusammenhang

weitere Folgen nach Carroll et al. (2006, S.341):

- Inferenz bei strengen nicht- linearen Regressionsmodellen basiert auf einer Approximation durch ein lineares Modell: z.B. durch Entwicklung einer Taylor Reihe für β um den wahren Wert β_0:
  $Y_i = m_Y(Z_i, \beta) + \epsilon_i \approx m_Y(Z_i, \beta_0) + f'(Z_i, \beta_0)(\beta - \beta_0) + \epsilon_i$
- Fehler in der Taylor Approximation geht gegen Null, wenn sich β an β_0 nähert
- tendenzielle steigende Varianz des Messfehlers führt zu einer höheren Variation von $\hat{\beta}$ um β_0 und hat folglich Einfluss auf die Schiefe von $\hat{\beta}$

Additiver Messfehler

- Z: Einflussgrösse Y: Zielgröße
- Zusammenhang zwischen Einfluss- (Z) und Zielvariable (Y) linear
- S: beobachtete fehlerbehaftete Zielvariable

$$S_i = Y_i + V_i \text{ wobei } V \sim \mathbb{N}(0; 3.0)$$

(Vgl. (Carroll et al.; 2006, S.340))

Additiver Messfehler

Warum führt ein additiver Messfehler zu einer höheren Varianz der beobachteten Zielvariablen?

Additiver Messfehler

Warum führt ein additiver Messfehler zu einer höheren Varianz der beobachteten Zielvariablen?

$$\mathbb{V}(S|Z) = \mathbb{V}(Y + V|Z) = \mathbb{V}(Y|Z) + \mathbb{V}(V) + 2Cov(Y, V)$$

$$= \mathbb{V}(Y|Z) + \mathbb{V}(V)$$

,da Y und V voneinander unabhängig sind.

Additiver Messfehler

Warum führt ein additiver Messfehler zu einer höheren Varianz der beobachteten Zielvariablen?

$$\mathbb{V}(S|Z) = \mathbb{V}(Y + V|Z) = \mathbb{V}(Y|Z) + \mathbb{V}(V) + 2Cov(Y, V)$$

$$= \mathbb{V}(Y|Z) + \mathbb{V}(V)$$

,da Y und V voneinander unabhängig sind.
$\Rightarrow \mathbb{V}(S|Z)$ enthält zusätzlich die Varianz des Messfehlers V
d.h. $\sigma^2_{new} = \sigma^2 + \sigma^2_v$
$\Rightarrow$ Güte einiger Tests, wie z.B. Konfidenzintervalle, wird reduziert

Additiver Messfehler

Warum sind die Schätzungen bezüglich der Regressionskeoffizienten im Falle eines additiven Messfehlers unverzerrt?

Additiver Messfehler

Warum sind die Schätzungen bezüglich der Regressionskeoffizienten im Falle eines additiven Messfehlers unverzerrt?

$$\mathbb{E}(S|Z) = \mathbb{E}(Y + V|Z) = \mathbb{E}(Y|Z) + \mathbb{E}(V) = \mathbb{E}(Y|Z)$$

Additiver Messfehler

Warum sind die Schätzungen bezüglich der Regressionskeoffizienten im Falle eines additiven Messfehlers unverzerrt?

$$\mathbb{E}(S|Z) = \mathbb{E}(Y + V|Z) = \mathbb{E}(Y|Z) + \mathbb{E}(V) = \mathbb{E}(Y|Z)$$

$\Rightarrow$ Erwartungswerte $\mathbb{E}(\hat{\beta}_{messf}|Z) = \mathbb{E}(\hat{\beta}_{wahr}|Z)$ unterscheiden sich nicht

Lineare Messfehler

wie bisher:

- Z: Einflussgrösse Y: Zielgröße
- Zusammenhang zwischen Einfluss- (Z) und Zielvariable (Y) linear
- S: beobachtete fehlerbehaftete Zielvariable

$$S_i = \gamma_0 + \gamma_1 Y_i + \varepsilon$$

Lineare Messfehler

wie bisher:

- Z: Einflussgrösse Y: Zielgröße
- Zusammenhang zwischen Einfluss- (Z) und Zielvariable (Y) linear
- S: beobachtete fehlerbehaftete Zielvariable

$$S_i = \gamma_0 + \gamma_1 Y_i + \varepsilon$$

neue lineare Zusammenhang zwischen Einflussvariable Z und beobachtete Zielvariable S:

$$S_i = \gamma_0 + \beta_0\gamma_1 + \gamma_1\beta_1 Z_i + \varepsilon$$

$\Rightarrow$ zusätzliche Parameter γ_0, γ_1

Linearer Messfehler

Warum sind die Schätzungen bezüglich der Regressionskeoffizienten im Falle eines linearen Messfehlers verzerrt?

Linearer Messfehler

Warum sind die Schätzungen bezüglich der Regressionskeoffizienten im Falle eines linearen Messfehlers verzerrt?

$$\mathbb{E}(S|Z) = \mathbb{E}(\gamma_0 + \gamma_1 Y|Z) \neq \mathbb{E}(Y|Z)$$

Linearer Messfehler

Warum sind die Schätzungen bezüglich der Regressionskeoffizienten im Falle eines linearen Messfehlers verzerrt?

$$\mathbb{E}(S|Z) = \mathbb{E}(\gamma_0 + \gamma_1 Y|Z) \neq \mathbb{E}(Y|Z)$$

$$\Rightarrow \mathbb{E}(\hat{\beta}_{messf}|Z) \neq \mathbb{E}(\hat{\beta}_{wahr}|Z)$$

Allgemeine Likelihood Methoden

Ausgangssituation: ein Modell wurde nur für Y spezifiziert, jedoch wurde S beobachtet und nicht Y
Ziel der Likelihood Methode: Beide Punkte zu berücksichtigen, um bestmöglich Schätzung der Regressionsparameter zu gewährleisten

Allgemeine Likelihood Methoden

Ausgangssituation: ein Modell wurde nur für Y spezifiziert, jedoch wurde S beobachtet und nicht Y
Ziel der Likelihood Methode: Beide Punkte zu berücksichtigen, um bestmöglich Schätzung der Regressionsparameter zu gewährleisten

Erwartungswert $(S|Z)$:

$$\mathbb{E}_{\beta,\gamma}(S|Z) = \int \mathbb{E}_{\beta}(Y|Z) dP_{\gamma}(S|Y,Z)$$

(Vgl. (Pepe und Fleming; 1991, S.109))

Likelihood Methoden für diskrete Response- Variable

- S,Y diskret
- P ist ein Wahrscheinlichkeitsmaß
- **Likelihood für** $(S|Z)$**:**

$$f_{S|Z}(s|z, B, \gamma) = \sum_y f_{Y|Z}(y|z, B) \cdot f_{S|Y,Z}(s|y, z, \gamma)$$

(Vgl. Carroll et al. (2006, S.353))

Likelihood Methoden für diskrete Response- Variable

Erläuterung anhand von bedingten Wahrscheinlichkeiten:

$$P(S = s|Z = z) = \underbrace{\sum_{y} \underbrace{P(S = s|Y = y, Z = z)}_{:=1} \cdot \underbrace{P(Y = y|Z = z)}_{:=2}}_{:=3}$$

(1): die Wahrscheinlichkeit von S gegeben Y,Z
(2): die Wahrscheinlichkeit von Y gegeben Z
(3): summiert über alle beobachteten y

Einschub: Missklassifikation

Missklassifikation: Messfehler in einer binären Zielvariablen d.h. Zuordnungen in falsche Klassen

Einschub: Missklassifikation

Missklassifikation: Messfehler in einer binären Zielvariablen d.h. Zuordnungen in falsche Klassen

Missklassifikations- Wahrscheinlichkeiten:

- $\pi_0 = P(S = 1|Y = 0)$: Die Zuordnung einer Beobachtung zu Klasse 1, obwohl diese in Wahrheit der Klasse 0 entspricht z.B. der Patient wird anhand eines Testergebnisses als krank eingestuft, obwohl dieser in Wahrheit gesund ist (falsch positiv)
- $\pi_1 = P(S = 0|Y = 1)$: Die Zuordnung einer Beobachtung zu Klasse 0, obwohl diese in Wahrheit der Klasse 1 entspricht z.B. der Patient wird als gesund eingestuft, obwohl der Patient tatsächlich krank ist (falsch negativ)

Einschub: Missklassifikation

- in einem logistischem Regressionsmodell wird der Erwartungswert $\mathbb{E}(Y|Z)$ gebildet durch: $\mathbb{E}(Y|Z) = P(Y = 1|Z) = H(\beta_0 + \beta_1 Z)$

Einschub: Missklassifikation

- in einem logistischem Regressionsmodell wird der Erwartungswert $\mathbb{E}(Y|Z)$ gebildet durch: $\mathbb{E}(Y|Z) = P(Y = 1|Z) = H(\beta_0 + \beta_1 Z)$
- Sind Missklassifikations- Wahrscheinlichkeiten vorhanden und nur die Variable S beobachtbar, so ist $\mathbb{E}(S|Z)$:
$P(S = 1|Z) = \underbrace{\pi_0}_{(1)} + \underbrace{(1 - \pi_0 - \pi_1) H(\beta_0 + \beta_1 Z)}_{(2)}$
(Vgl. J. Abrevaya und Scott-Morton (1998, S.241))

(1): Dieser Term entspricht der Missklassifikations- Wahrscheinlichkeit, dass bei $S = 1$ falsch zugeordnet wird.

(2): Entspricht der Gegenwahrscheinlichkeit von (1)

Einschub: Missklassifikation

- in einem logistischem Regressionsmodell wird der Erwartungswert $\mathbb{E}(Y|Z)$ gebildet durch: $\mathbb{E}(Y|Z) = P(Y = 1|Z) = H(\beta_0 + \beta_1 Z)$
- Sind Missklassifikations- Wahrscheinlichkeiten vorhanden und nur die Variable S beobachtbar, so ist $\mathbb{E}(S|Z)$:
$P(S = 1|Z) = \underbrace{\pi_0}_{(1)} + \underbrace{(1 - \pi_0 - \pi_1) H(\beta_0 + \beta_1 Z)}_{(2)}$
(Vgl. J. Abrevaya und Scott-Morton (1998, S.241))

  (1): Dieser Term entspricht der Missklassifikations- Wahrscheinlichkeit, dass bei $S = 1$ falsch zugeordnet wird.

  (2): Entspricht der Gegenwahrscheinlichkeit von (1)
- Hinsichtlich Parameterschätzungen ergeben sich nun Verzerrungen, da $\mathbb{E}(S|Z) \neq \mathbb{E}(Y|Z)$

Likelihood Methoden für binäre Response- Variable

Schätzung aller unbekannten Parameter π_0, π_1, β_0 und β_1: $\Rightarrow$ ML Schätzung:

1. $$L(\pi_0, \pi_1, \beta_0, \beta_1) = \prod_{i=1}^{n} \left[\pi_0 + (1 - \pi_0 - \pi_1) H(\beta_0 \dot{+} \beta_1 Z)\right]^{S_i}$$
$$+ \left[1 - \pi_0 - (1 - \pi_0 - \pi_1) H(\beta_0 \dot{+} \beta_1 Z)\right]^{(1-S_i)}$$

Likelihood Methoden für binäre Response- Variable

Schätzung aller unbekannten Parameter π_0, π_1, β_0 und β_1: $\Rightarrow$ ML Schätzung:

1. $$L(\pi_0, \pi_1, \beta_0, \beta_1) = \prod_{i=1}^{n} \left[\pi_0 + (1 - \pi_0 - \pi_1) H(\beta_0 \dot{+} \beta_1 Z)\right]^{S_i}$$
$$+ \left[1 - \pi_0 - (1 - \pi_0 - \pi_1) H(\beta_0 \dot{+} \beta_1 Z)\right]^{(1-S_i)}$$

2. $$l(\pi_0, \pi_1, \beta_0, \beta_1) = \sum_{i=1}^{n} \left[S_i ln\left(\pi_0 + (1 - \pi_0 - \pi_1) H(\beta_0 \dot{+} \beta_1 Z)\right)\right]$$
$$[(1 - S_i) ln(1 - \pi_0 - (1 - \pi_0 - \pi_1) \cdot H(\beta_0 + \beta_1 Z))]$$

(Vgl. J. Abrevaya und Scott-Morton (1998)[S.242])

Likelihood Methoden für stetige Response- Variable

- S stetige Response- Variable mit linearen Messfehler
- P aus der allgemeinen Likelihood- Funktion ein Wahrscheinlichkeitsmaß
- **Likelihood für** $(S|Z)$**:**

$$f_{S|Z}(s|z, B, \gamma) = \int f_{Y|Z}(y|z, B) \cdot f_{S|Y,Z}(s|y, z, \gamma) d\lambda$$

wobei λ das Lebesgue- Maß ist.

Nachteile der Likelihood Methoden

- mühsame Berechnung
- starke Sensibilität gegenüber der Annahme über die Verteilung der Variablen S

Validierungsdaten im Fall einer stetigen Response

Validierungsdaten: Teildatensatz, bei denen ein Teil wahre Beobachtungen enthält und der andere Teil sich aus fehlerbehaftet Daten zusammensetzt

- Zuordnung der Beobachtungen zu den Validierungsdaten per Zufall
- Ziel: Verzerrung in den Parameterschätzungen verursacht durch einen linearen Messfehler zu eliminieren

Validierungsdaten im Fall einer stetigen Response

Methode nach Carroll et al. (2006, S.343):

1. Modellgleichung wird zwischen wahrer Response- Variable Y und Einflussvariable Z aufgestellt $\Rightarrow$ Schätzung von β_0, β_1
2. Zusammenhang zwischen S und Y wird untersucht $\Rightarrow$ Schätzung von γ_0, γ_1
3. Zuordnung der Schätzer zu der Menge $\hat{B}_1$
4. neue Variable S' wird erzeugt durch: $(S - \gamma_0)/\gamma_1$
5. Zusammenhang zwischen S' und Y $\Rightarrow$ Schätzung von $\gamma_{0,neu}, \gamma_{1,neu}$
6. Zuordnung der Schätzer zu der Menge $\hat{B}_2$
7. Mengen $\hat{B}_1$ und $\hat{B}_2$ multipliziert mit gemeinsamer Kovarianzmatrix (gebildet mittels Bootstrap) zu $\hat{B}$
8. Ergebnismatrix $\hat{B}$ enthält unverzerrte Schätzer

Validierungsdaten im Fall einer diskreten Response

Ziel: Missklassifikations- Wahrscheinlichkeiten anhand von Validierungsdaten zu schätzen

Validierungsdaten im Fall einer diskreten Response

Ziel: Missklassifikations- Wahrscheinlichkeiten anhand von Validierungsdaten zu schätzen

Ansatz zur Schätzung von π_0, π_1: Anteilsschätzer zwischen den Beobachtungen, die korrekt klassifiziert worden sind und den Beobachtungen dessen wahrer Wert (hier Y=1) beträgt
$\Rightarrow$ Schätzung von β_0, β_1 durch Einsetzten von $\hat{\pi_0}, \hat{\pi_1}$ in die Likelihood (Pseudo- Likelihood)

Validierungsdaten im Fall einer diskreten Response

Ziel: Missklassifikations- Wahrscheinlichkeiten anhand von Validierungsdaten zu schätzen

Ansatz zur Schätzung von π_0, π_1: Anteilsschätzer zwischen den Beobachtungen, die korrekt klassifiziert worden sind und den Beobachtungen dessen wahrer Wert (hier Y=1) beträgt
$\Rightarrow$ Schätzung von β_0, β_1 durch Einsetzten von $\hat{\pi_0}, \hat{\pi_1}$ in die Likelihood (Pseudo- Likelihood)

Nachteil: genaue und konsistente Schätzung nur unter hohem Stichprobenumfang möglich (Vgl. Copas (1988))

Validierungsdaten im Bezug auf die allgemeinen Likelihood Methoden

Ziel: Mit Hilfe der Validierungsdaten einfachere Form der Likelihood-Funktion

Validierungsdaten im Bezug auf die allgemeinen Likelihood Methoden

Ziel: Mit Hilfe der Validierungsdaten einfachere Form der Likelihood-Funktion Ansatz: Aufsplitten in zwei Produkte

$$\prod_{i=1}^{n} \left[f_{Y|Z}\left(y_i|z_i, B\right) f_{S|Y,Z}\left(s_i|y_i, z_i, \gamma\right) \right]^{1-\Delta_i} \left[f_{Y|Z}\left(y_i|z_i, B\right) f_{S|Y,Z}\left(s_i|y_i, z_i, \gamma\right) \right]^{\Delta_i}$$

mit

$$\Delta_i = \begin{cases} 1 & \text{wenn Beobachtung } i \in \text{Validierungsdaten} \\ 0 & \textit{sonst} \end{cases}$$

(Vgl. (Carroll et al.; 2006, S.354))

Complete Data Methoden

Complete Data Methoden (oder auch Complete- Cases):

- bezieht sich auf die Methode der Validierungsdaten
- stärkere Anforderungen an die Beobachtungen in den Validierungsdaten: nur wahre Beobachtungen Y und Z
- Modell nur auf Basis der Validierungsdaten

Complete Data Methoden

Complete Data Methoden (oder auch Complete- Cases):

- bezieht sich auf die Methode der Validierungsdaten
- stärkere Anforderungen an die Beobachtungen in den Validierungsdaten: nur wahre Beobachtungen Y und Z
- Modell nur auf Basis der Validierungsdaten

Schätzung der Regressionskoeffizienten:

- KQ- Methode, falls $(Y|Z)$ normalverteilt und linearer Zusammenhang
- ML Schätzung

Complete Data Methoden

Beispiel:

- Y als das verifizierte Einkommen betrachten (z.B. mit Nachweis einer Lohnabrechnung)
- S als das berichtete Einkommen, bzw. ohne jeglichen Nachweis über die tatsächliche Höhe des Einkommens

$\Rightarrow$ zwei unabhängigen Datensätzen aufgrund Datenschutz

Likelihood für alle Beobachtungen i=1,...,n:

$$\prod_{i=n}^{n} \{f(Y_i|Z_i, B)\}^{\Delta_i} \{f(S_i|Z_i, B, \gamma)\}^{1-\Delta_i}$$

(Vgl.Carroll et al. (2006)[S.356])

Nachteile der Complete Data Methoden

- Konkurrenz zu den Missklassifikations- Wahrscheinlichkeiten, wenn binäre Fall vorliegt
- Informationsverlust
- Verzerrung, wenn Auswahl der Validierungsdaten von S und Z abhängig

Vergleich der Methoden

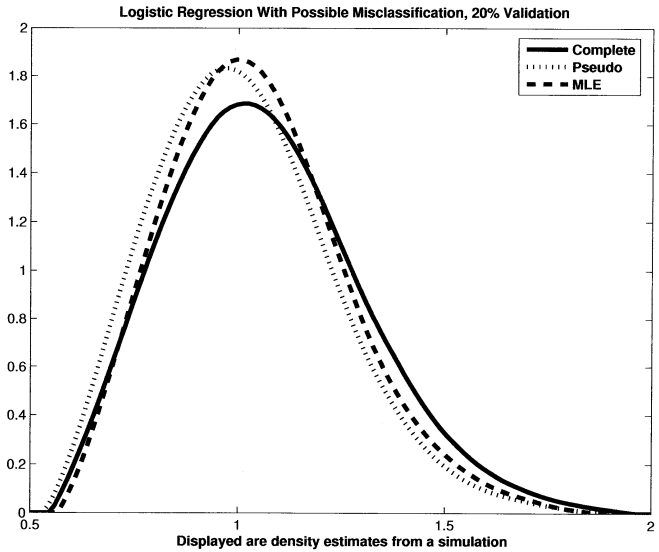

Abbildung: Vergleich der Methoden mit Auswahl an Valdierungsdaten per Zufall (Carroll et al. (2006, S.349))

Vergleich der Methoden

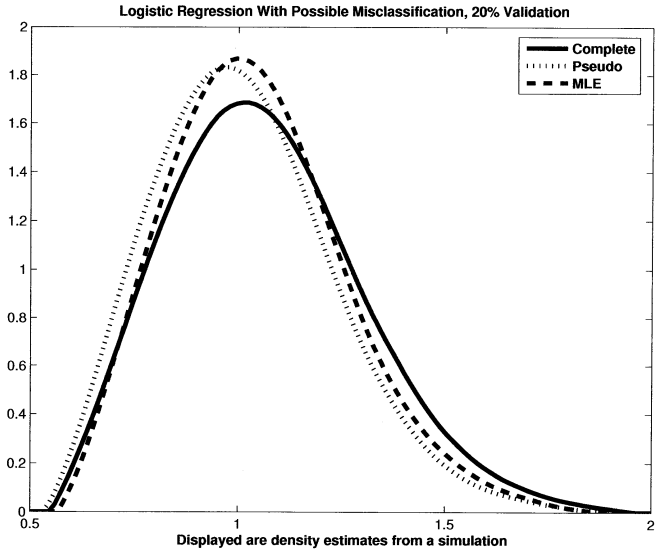

Abbildung: Vergleich der Methoden mit Auswahl an Valdierungsdaten per Zufall (Carroll et al. (2006, S.349))

$\Rightarrow$ Schätzer sind fast identisch

$\Rightarrow$ Problem bei Complete- Data Methode: Wahrscheinlichkeit in den Valiedierungsdaten nur gültige Werte zu erhalten sehr gering

Vergleich der Methoden

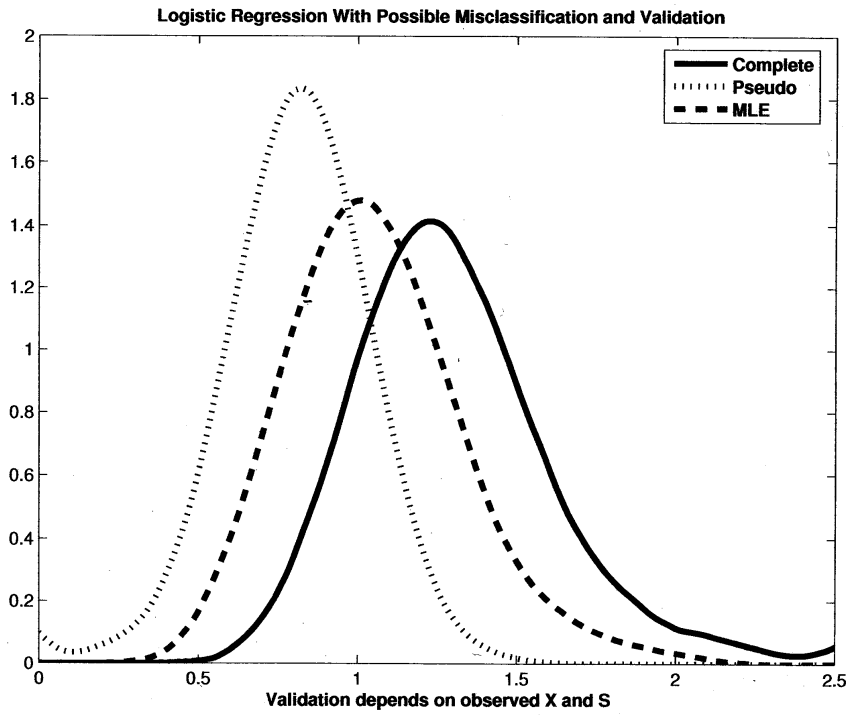

Abbildung: Vergleich der Methoden mit Auswahl an Validierungsdaten von S und Z abhängig (Carroll et al. (2006, S.350))

Vergleich der Methoden

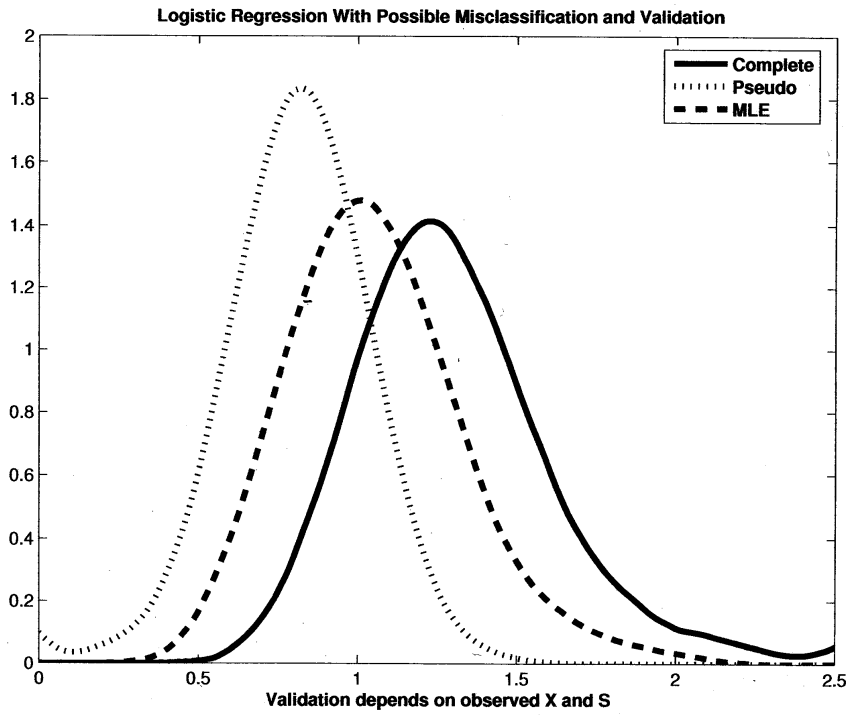

Abbildung: Vergleich der Methoden mit Auswahl an Validierungsdaten von S und Z abhängig (Carroll et al. (2006, S.350))

$\Rightarrow$ Schätzer von Complete Data Methode und Pseudolikelihood verzerrt $\Rightarrow$ ML Schätzer bleibt unverändert

Vergleich der Methoden: Beispiel

GVHR (Graft-versus-Host-Reaktion): immunologische Reaktion, die in der Folge einer allogenen Knochenmark- oder Stammzelltransplantation auftreten kann

- Ziel der Studie: Zusammenhang zwischen Alter eines Patienten und Auftreten einer GHVR untersuchen
- Y chronische GVHR, S akute GVHR
- Z jünger/ älter als 20

Vergleich der Methoden: Beispiel

Validierungsdaten:

Validation Data			
Z	**S**	**Y**	Count
0	0	0	19
0	0	1	5
0	1	0	7
0	1	1	14
1	0	0	28
1	0	1	27
1	1	0	8
1	1	1	24
Nonvalidation Data			
0	0	–	47

Abbildung: Validierungsdaten für das GHVR Beispiel mit festen $\pi(S, Z)$ (Carroll et al. (2006, S.351))

Vergleich der Methoden: Beispiel

Vergleich von Complete Data- Schätzung und ML- Schätzung:

	Validation Data	MLE
$\widehat{\beta}_z$	0.66	1.13
Standard Error	0.37	0.38
p-value	0.078	0.004

Abbildung: Vergleich Complete Data Methode und ML Schätzung (Carroll et al. (2006, S.352))

Vergleich der Methoden: Beispiel

Vergleich von Complete Data- Schätzung und ML- Schätzung:

	Validation Data	MLE
$\widehat{\beta}_z$	0.66	1.13
Standard Error	0.37	0.38
p-value	0.078	0.004

Abbildung: Vergleich Complete Data Methode und ML Schätzung (Carroll et al. (2006, S.352))

$\Rightarrow$ starke Verzerrungen der Schätzungen von β $\Rightarrow$ bei Signifikanzniveau von $\alpha = 0.05$ unterschiedliche Ergebnisse bezüglich des signifikanten Einflusses der Variable Z

Semiparametrische Methoden

Ziel: Empfindlichkeit der Verteilungsannahmen der beobachteten Response S in der Likelihood zu reduzieren

- z.B. $S = \gamma_0 + \gamma_1 Y \Rightarrow$ Dichtefunktion $f_{S|Y,Z}$ von γ_0, γ_1 unabhängig sein
- Problem: Variable S liefert in den Validierungsdaten keine Information über die Verteilung von Y
- Idee: neue Variable K zu erheben, die informative Komponente von S widerspiegelt (Surrogat) (Vgl. Pepe und Fleming (1991))

Semiparametrische Methoden: Beispiel

- Y die Konzentration eines Medikamentes im Blut
- K die Dosierung eines Medikamentes im Blut (Surrogat für Y) - ist einfacher und billiger zu erheben als Y

Semiparametrische Methoden: Beispiel

- Y die Konzentration eines Medikamentes im Blut
- K die Dosierung eines Medikamentes im Blut (Surrogat für Y) - ist einfacher und billiger zu erheben als Y

$\Rightarrow f_{K|Y}$ als empirischer Schätzer für $\hat{f}_{S|Y,Z}$

Semiparametrische Methoden: Beispiel

Likelihood- Ansatz:

1. Einsetzten von $f_{K|Y}$ anstatt $f_{S|Y,Z}$ in die allgemeine Likelihood

$$\hat{f}_{S|Z}(s|z,B) = \int f_{Y|Z}(y|z,B) \cdot f_{K|Y}(s|y) d\lambda$$

$\Rightarrow$ liefert einen Schätzer für $f_{S|Z}$

Semiparametrische Methoden: Beispiel

Likelihood- Ansatz:

1. Einsetzten von $f_{K|Y}$ anstatt $f_{S|Y,Z}$ in die allgemeine Likelihood

$$\hat{f}_{S|Z}(s|z,B) = \int f_{Y|Z}(y|z,B) \cdot f_{K|Y}(s|y) d\lambda$$

   $\Rightarrow$ liefert einen Schätzer für $f_{S|Z}$

2. Maximieren von

$$\prod_{i=n}^{n} \{f(Y_i|Z_i,B)\}^{\Delta_i} \left\{\hat{f}(S_i|Z_i,B)\right\}^{1-\Delta_i}$$

   $\Rightarrow$ liefert einen Schätzer für β_0, β_1

(Vgl. Carroll et al. (2006))

Fazit

- alle Methoden Verbesserungen anderer Methoden sind
- großes Potential zur Ausweitung und Verbesserung z.B. Kombinationen verschiedener Methoden
- schwer durchführbar wenn Teildatensatz mit wahren Beobachtungen oder eine Surrogat nicht vorhanden

Vielen Dank für die Aufmerksamkeit!

Anhang

Anhang: Auswirkungen

Was ist mit Schiefe gemeint?

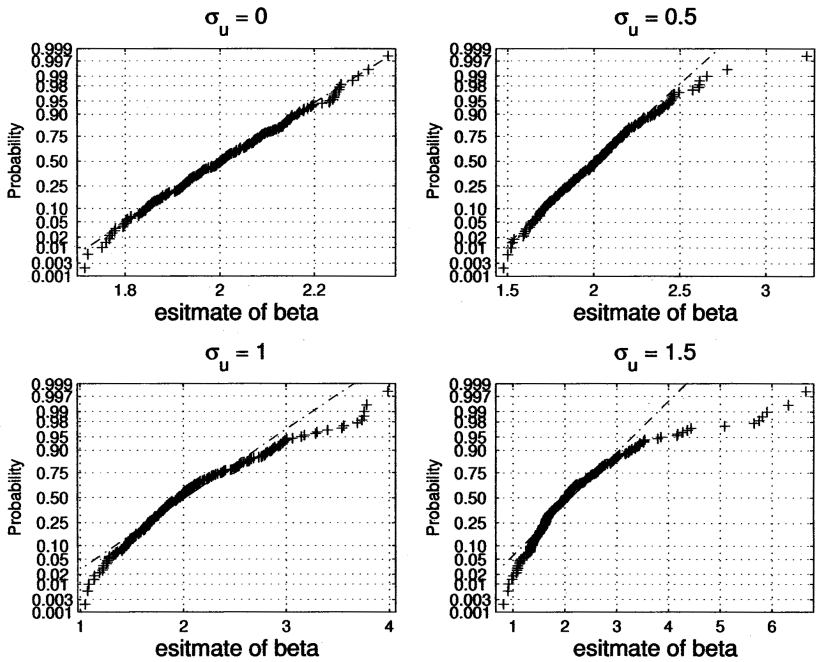

Abbildung: Abbild von $\hat{\beta}$ mit 250 Simulationen eines Exponentiellen Regressionsmodells (Carroll et al. (2006, S.341))

Anhang: Schätzung von $f_{S|Y,Z}(s|y,z,\gamma)$

Welche Möglichkeiten bestehen für die Schätzung von $f_{S|Y,Z}(s|y,z,\gamma)$?

- Anwenden eines multinomialen Regressionsmodels: Angenommen S ordinal, diskrete Größe mit Ausprägungen: 1,...,S
  $\Rightarrow$ Wahrscheinlichkeit von $S|Y,Z$:

$$P(S \geq s|Y,Z) = H(\gamma_{0s} + \gamma_1 Y + \gamma_2 Z), s = 1, ..., S$$

  (Vgl. Carroll et al. (2006))
  Falls S stetig $\Rightarrow$ Bilden von Schwellenwerten für die Variable S, sodass S ordinal ist

- Falls Surrogate vorhanden, anwenden der Semiparametrischen Methoden

Anhang: Surrogate

Surrogat: Ein Surrogat K spiegelt die informative Komponente einer Variablen S wider

- Verteilung von K hängt nur von der Variablen Y ab- nicht von den Einflussgrößen
- speziell wenn S Surrogat für Y und Z Einflussgröße (z.B. wenn Y latente Variable):
  $f_{S|Y,Z}(s|y,z,\gamma)$ vereinfacht sich zu $f_{S|Y}(s|y,\gamma)$ (Vgl. Carroll et al. (2006))

Anhang: Validierungsdaten

Wie wird $\hat{B}$ gewichtet im letzte Schritt von S.34 nach Methode von Carroll et al. (2006, S.343) ?

$$\hat{B} = \left(J^t \Sigma^{-1} J\right)^{-1} J^t \Sigma^{-1} \left(\hat{B}_1^t, \hat{B}_2^t\right)^t$$

wobei $J = (I, I)$ und I die Identitätsmatrix sei

Abrevaya, J. und Hausman, J. A. (2004). Response error in a transformation model with an application to earnings-equation estimation, *The Econometrics Journal* **7**: 366–388.

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

Copas, J. B. (1988). Binary regression models for contaminated data, *Journal of the Royal Statistical Society* **50**: 225–265.

J. Abrevaya, J. A. H. und Scott-Morton, F. (1998). Misclassification of the dependent variable in a discrete-response setting, *Journal of Econometrics* **87**: 239–269.

Pepe, M. S. und Fleming, T. R. (1991). A nonparametric method for dealing with mismeasured covariate data, *Journal of the American Statistical Associationt* **86**: 108–113.

Rügamer, D. (2014). Modelldiagnose. Folien zum Tutorium Lineare Modelle von Prof. Dr. Helmut Kchenhoff.