Schneeweiss:

Korrigierte Schaetzgleichungen fuer allgemeine Regressionsmodelle mit Fehlern in den Variablen

Sonderforschungsbereich 386, Paper 93 (1997)

Online unter: http://epub.ub.uni-muenchen.de/

Korrigierte Schätzgleichungen für allgemeine Regressionsmodelle mit Fehlern in den Variablen

Corrected Estimating Equations for General Regression Models with Errors in the Variables

Hans Schneeweiss, München

Nach einer kurzen Einführung in die Theorie der erwartungstreuen Schätzgleichungen für allgemeine Regressionsmodelle und der korrigierten Schätzgleichungen für Regressionsmodelle mit fehlerbehafteten Kovariablen wird die Approximationsgüte eines auf Reihenentwicklung basierenden Ansatzes von Stefanski diskutiert.

After a short introduction into the theory of unbiased estimating equations for general regression models and of corrected estimating equations for regression models with error ridden covariables, the precision of an approximative approach due to Stefanski is discussed.

1 Einleitung

Bekanntlich führt die Anwendung der Kleinst-Quadrat-Schätzung auf eine lineare Regression, deren Regressorvariablen mit zufälligen Fehlern behaftet sind, zu inkonsistenten Schätzungen der Regressionsparameter, vgl. z.B. Schneeweiß und Mittag (1986). Das gleiche gilt für allgemeinere, nichtlineare Regressionsmodelle. Aber während für das lineare Modell eine ausgefeilte und praktikable Schätztheorie existiert - vgl. z. B. Fuller (1987) -, gibt es im allgemeineren nichtlinearen Fall zwar eine große Anzahl verschiedener - oft nur approximativ einsetzbarer - Schätzansätze, doch keine geschlossene Theorie, vgl. Carroll et al. (1995). Das ist auch kaum verwunderlich, da es sehr unterschiedliche nichtlineare Modelle gibt - auch solche mit diskreten abhängigen Variablen - und da die restlose Ausschaltung einer Verzerrung bei nichtlinearen Modellen nur ausnahmsweise gelingt.

An dem Ansatz der sogenannten (fehler-) korrigierten Schätzgleichungen soll diese Schwierigkeit verdeutlicht werden. Im Gegensatz zu anderen verbreiteteren Schätzverfahren geht dieser Ansatz von der sogenannten Funktionalvariante des Fehler - in - den - Variablen - Modells aus, d.h. die nicht direkt beobachtbaren vom Fehler befreiten, wahren Regressorvariablen werden - wie üblicherweise auch im klassischen linearen Modell - als nichtstochastisch aufgefaßt. Damit entfällt die Spezifizierung einer Verteilung für diese latenten Regressoren, wie sie im entgegengesetzten Fall der sogenannten Strukturvariante vonnöten wäre. Erkauft wird diese Befreiung vom Spezifizierungszwang mit der Schwierigkeit, eine Korrektur der Schätzgleichungen so zu finden, daß dann auch wirklich konsistente Schätzungen resultieren, vgl. Stefanski (1989), Nakamura (1990), Carroll et al. (1995) und Buonaccorsi (1996).

In Abschnitt 2 wird eine kurze Einführung in die Theorie der erwartungstreuen Schätzgleichungen gegeben. Abschnitt 3 wendet die Theorie auf Modelle mit fehlerbehafteten Regressoren an und schildert, wie man zu konsistenten Schätzungen kommt, wenn man korrigierte Schätzgleichungen verwendet. In Abschnitt 4 wird schließlich ein Ansatz von Stefanski zur Konstruktion korrigierter Schätzgleichungen diskutiert. Dabei kann eine von Stefanski eingeführte Approximation in einen verbesserten Anwendungsbezug gebracht werden.

2 Erwartungstreue Schätzgleichungen und Quasi-Score-Funktionen

Ein allgemeines Regressionsmodell spezifiziert die Abhängigkeit einer Variablen y von einem Vektor gegebener Regressoren (oder Kovariablen) ξ *im Mittel*. Es wird also (zunächst) nur der bedingte Erwartungswert von y, gegeben ξ, als Funktion von ξ ausgedrückt, wobei noch unbekannte und zu schätzende Parameter β den funktionalen Zusammenhang näher bestimmen:

$$E(y_i \mid \xi_i; \beta) = \mu(\xi_i, \beta). \tag{1}$$

Die i.a. nichtlineare Funktion μ wird als bekannt unterstellt, der p-dimensionale Vektor β als unbekannt. Der Index i kennzeichnet die Beobachtung in einer Stichprobe $i = 1, \ldots, n$. Er wird im folgenden zumeist weggelassen. Es wird angenommen, daß die y_i, gegeben $\xi_1 \ldots \xi_n$, stochastisch unabhängig sind. Die Variablen ξ_i werden zunächst als fehlerfrei angenommen. Versteht man sie als feste, nichtstochastische Größen - was wir hier tun wollen - dann kann in (1) statt des bedingten Erwartungswertes auch der unbedingte stehen, so daß man kürzer schreiben kann:

$$E(y_i) = \mu(\xi_i, \beta).$$

In der Regel sollte neben (1) auch die Verteilung von y (gegeben ξ) bekannt sein. Das ist aber nicht unbedingt erforderlich. Wichtig ist allerdings, daß eine sogenannte erwartungstreue (vektorielle) Schätzgleichung gegeben ist. Ausgangspunkt ist eine p-dimensionale Funktion $\psi(y, \xi, \beta)$ mit

$$E\psi(y, \xi, \beta) = 0. \tag{2}$$

Unter sehr allgemeinen Bedingungen ist dann die Lösung $\hat{\beta}$ der Gleichung

$$\sum_{i=1}^{n} \psi(y_i, \xi_i, \hat{\beta}) = 0 \tag{3}$$

eine konsistente und asymptotisch normalverteilte Schätzung für β, vgl. Huber (1967). Die asymptotische Kovarianzmatrix von $\hat{\beta}$ läßt sich mit Hilfe der

Funktion ψ angeben. Das System (3) wird als *erwartungstreue Schätzgleichung* bezeichnet.

In vielen Fällen wird ψ die Score-Funktion sein, also

$$\psi(y, \xi, \beta) = \frac{\partial}{\partial \beta} l(\beta; y, \xi)$$

mit der Loglikelihood $l(\beta; y, \xi)$ für die Beobachtung (y, ξ), wobei freilich oft noch Nuisance-Parameter mit im Spiel sind, die hier der Einfachheit halber weggelassen sind. Allgemeiner kann aber auch ψ eine sogenannte Quasi-Score-Funktion sein, die aus der Likelihoodfunktion eines u. U. fehlspezifizierten Modells hervorgeht. Noch allgemeiner soll jede Funktion ψ, für die (2) gilt, als *Quasi-Score-Funktion* (QSF) bezeichnet werden, auch wenn sie nicht aus einer Likelihoodfunktion hervorgeht.

Zwei Beispiele sollen die Bildung von Quasi-Score-Funktionen veranschaulichen.

1. *Nichtlineare Regression:* Es sei

   $$y_i = \mu(\xi_i, \beta) + \varepsilon_i = \mu_i + \varepsilon_i$$

   mit unabhängig normalverteilten $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$. Die Loglikelihood der i-ten Beobachtung ist dann gegeben durch $l_i = -\frac{1}{2}(\varepsilon_i^2/\sigma_\varepsilon^2 + \log \sigma_\varepsilon^2)$ und die Score-Funktion durch

   $$\psi_i = \frac{1}{\sigma_\varepsilon^2}\{y_i - \mu(\xi_i, \beta)\}\frac{\partial \mu(\xi_i, \beta)}{\partial \beta} \tag{4}$$

   Dabei kann der Faktor $1/\sigma_\varepsilon^2$ auch weggelassen werden. Speziell für eine polynomiale Regression

   $$y_i = \beta_0 + \beta_1 \xi_i + \ldots + \beta_p \xi_i^p + \varepsilon_i$$

   erhält man die $(p+1)$-dimensionale Score-Funktion

   $$\psi_i = \frac{1}{\sigma_\varepsilon^2}(y_i - \beta_0 - \beta_1 \xi_i - \ldots - \beta_p \xi_i^p)(1, \xi_i, \ldots, \xi_i^p)'. \tag{5}$$

   Ist ε_i nicht normalverteilt, aber immer noch $E\varepsilon_i = 0$, dann ist ψ_i eine QSF, die immer noch zu konsistenten Schätzungen führt.

2. *Verallgemeinertes lineares Modell:* Die Verteilung von y folge einer einparametrigen Exponentialfamilie. Für den Erwartungswert $\mu = \mu(\xi, \beta)$ von y wird eine über eine umkehrbare Linkfunktion g vermittelte Abhängigkeit von der Linearkombination $\beta'\xi$ des p-dimensionalen Regressorvariablenvektors ξ postuliert:

   $$g(\mu) = \beta'\xi \tag{6}$$

und damit $\mu = g^{-1}(\beta'\xi)$. Auch die Varianz σ_y^2 von y hängt von μ und damit von β ab. Die Score-Funktion ergibt sich dann - vgl. Fahrmeir u. Tutz (1994) - zu

$$\psi(y,\xi,\beta) = \{\sigma_y^2 \frac{\partial g(\mu)}{\partial \mu}\}^{-1}(y-\mu)\xi. \tag{7}$$

Auch hier kann die zugrundeliegende Exponentialfamilie (d.h. insbesondere σ_y^2) fehlspezifiziert sein, ohne daß die Schätzgleichung (3) ihre Eigenschaft verliert, zu konsistenten Schätzungen für β zu führen, sofern nur die Mittelwertsspezifikation (6) korrekt ist. Die Funktion ψ ist dann wieder eine QSF.

Ein Beispiel ist das Modell einer exponentiell verteilten Variablen y mit der Mittelwertsfunktion

$$\mu = \exp(\beta'\xi).$$

Es ist also hier $g(\mu) = \log \mu$ und daher gemäß (7) (wegen $\sigma_y = \mu$)

$$\psi(y,\xi,\beta) = (y-\mu)\mu^{-1}\xi. \tag{8}$$

Die beiden Beispiele zeigen, daß eine QSF häufig die Form

$$\psi(y,\xi,\beta) = \{y - \mu(\xi,\beta)\}d(\xi,\beta)$$

hat mit einer von Fall zu Fall verschiedenen Funktion d, vgl. Buonaccorsi (1996). Für alle Funktionen ψ dieser Form ist offensichtlich (2) erfüllt.

3 Fehlerbehaftete Regressoren und Fehlerkorrektur

Lassen wir jetzt die Annahme fehlerfreier Regressoren fallen und nehmen im Gegenteil an, daß statt ξ die (i.a. vektorielle) Variable x beobachtet wird mit

$$x = \xi + v,$$

worin v eine (vektorielle) Meßfehlervariable darstellt mit $Ev = 0$, die von ξ, aber auch von y unabhängig sei. Die Verteilung von v soll bekannt sein; z.B. $v \sim N(0,\Sigma)$ mit bekannter Kovarianzmatrix Σ.

Das Regressionsmodell $\mu(\xi,\beta)$ sei wie bisher bei gegebenen $\xi_1,\ldots,\xi_n$ konsistent schätzbar über eine erwartungstreue Schätzgleichung (3) mit der QSF $\psi(y,\xi,\beta)$. Die ξ_i sind aber in Wirklichkeit nicht gegeben, sondern nur die fehlerbehafteten x_i. Unter einer korrigierten Quasi-Score-Funktion (KQSF) versteht man eine Funktion $\psi_K(y,x,\beta)$ in den *beobachtbaren* Variablen y, x und dem Parametervektor β, für die

$$E[\psi_K(y,x,\beta) \mid y] = \psi(y,\xi,\beta) \tag{9}$$

gilt und die daher an Stelle von ψ zur Konstruktion einer erwartungstreuen Schätzgleichung herangezogen werden kann. In der Tat ist

$$E\psi_K(y, x, \beta) = 0,$$

also ψ_K eine QSF, und die Lösung von

$$\sum_{i=1}^{n} \psi_K(y_i, x_i, \hat{\beta}) = 0 \tag{10}$$

liefert wieder unter allgemeinen Bedingungen eine konsistente, asymptotisch normalverteilte Schätzung $\hat{\beta}$ für β.

Der schwache Punkt dieses Ansatzes ist natürlich das Problem der Konstruktion einer KQSF. Der Ansatz steht und fällt mit der Möglichkeit, zu gegebener QSF eine KQSF zu finden. Stefanski (1989), Nakamura (1990) und Buonaccorsi (1996) geben einige Beispiele.

Hier sei das oben angeführte Beispiel einer exponentiell verteilten Variablen y noch einmal aufgegriffen, Stefanski (1989). Für den speziellen Fall, daß nur eine ξ-Variable vorliegt und die Linearkombination $\beta'\xi$ die Form $\beta_0 + \beta_1\xi$ hat (ξ skalar), wird aus (8) die Score-Funktion

$$\psi(y, \xi, \beta) = \{y \exp(-\beta_0 - \beta_1\xi) - 1\}(1, \xi)'.$$

Ferner sei $v \sim N(0, \sigma^2)$. Unter Verwendung der Gleichungen

$$E(e^x) = e^{\xi} e^{\sigma^2/2}, \;\; E(xe^x) = (\xi + \sigma^2)E(e^x)$$

läßt sich die KQSF wie folgt konstruieren:

$$\psi_K(y, x, \beta) = (ya - 1)\binom{1}{x} + \begin{pmatrix} 0 \\ y\beta_1\sigma^2 a \end{pmatrix}$$

mit $a = \exp(-\beta_0 - \beta_1 x - \frac{1}{2}\beta_1^2\sigma^2)$.

4 Stefanskis Reihenentwicklung

Um zu einem allgemeineren, von den diversen Einzelresultaten abgehobenen Schätzverfahren zu kommen, auch wenn dieses u.U. nur approximative Gültigkeit besitzt, geht Stefanski von einer Taylorreihenentwicklung in ξ für die QSF ψ aus. Der Einfachheit halber wird dabei ξ als skalar angenommen. Außerdem wird für v Normalverteilung unterstellt: $v \sim N(0, \sigma^2)$.

Die fundamentale Beziehung (9) zwischen KQSF und QSF läßt sich unter Konzentration auf die relevanten Variablen ξ und x kürzer wie folgt fassen:

$$E\psi_K(x) = \psi(\xi). \tag{11}$$

Der Erwartungswert ist dabei bezüglich der normalverteilten Variablen v zu bilden, die in $x = \xi + v$ enthalten ist. Gesucht ist die KQSF ψ_K zu gegebener QSF ψ.

Stefanski zeigt, daß ψ_K allenfalls dann existiert, wenn ψ eine überall konvergierende Taylorreihenentwicklung

$$\psi(\xi) = \sum_{k=0}^{\infty} c_k \xi^k \tag{12}$$

mit $c_k = \psi^{(k)}(0)/k!$ besitzt, was wir hier annehmen wollen. Dann findet man eine KQSF wie folgt. Sei

$$H_k(z) = \sum_{i=0}^{k} a_{ki} z^i$$

das Hermitesche Polynom k-ter Ordnung und sei

$$P_k(x) = \sigma^k H_k(x/\sigma) = \sum_{i=0}^{k} a_{ki} \sigma^{k-i} x^i. \tag{13}$$

Dann gilt (für $x \sim N(\xi, \sigma^2)$)

$$EP_k(x) = \xi^k.$$

Alsdann ist offenbar - absolute Konvergenz der Reihe vorausgesetzt -

$$\psi_K(x) = \sum_{k=0}^{\infty} c_k P_k(x) \tag{14}$$

die gesuchte KQSF.

Im Falle eines polynomialen Regressionsmodells ist, wie (5) zeigt, die QSF von vornherein in der Form (12), und Ersetzen der Potenzen ξ^k in (5) durch die entsprechenden $P_k(x)$ führt unmittelbar zur gesuchten KQSF.

In der Regel wird man keinen geschlossenen Ausdruck für die Reihe (14) finden. Man ist dann auf Approximationen angewiesen, die man durch Abbruch der Reihe an einer Stelle m erhält:

$$\psi_{Km}(x) = \sum_{k=0}^{m} c_k P_k(x). \tag{15}$$

Für den Fehler dieser Approximation gilt

$$F_m(x) = \psi_K(x) - \psi_{Km}(x) = \sum_{k=m+1}^{\infty} c_k P_k(x).$$

Stefanski untersucht diesen Fehler in Abhängigkeit von σ und glaubt, feststellen zu können, daß er von der Größenordnung $O(\sigma)$ ist. Das ist sicher nicht richtig. Wie ein Blick auf (13) lehrt, ist $P_k(x) = O(1)$, und nicht $= O(\sigma)$, und gleiches gilt daher für $F_m(x)$. Das heißt aber, daß man bei der Diskussion eines eventuellen Bias von $\hat{\beta}$ nicht mit einer Klein-σ-Asymptotik argumentieren kann. Letzteres hieße, daß bei Verwendung von ψ_{Km} anstelle von ψ_K in der Schätzgleichung (10) die resultierende Verzerrung von $\hat{\beta}$ mit kleiner werdendem σ gegen Null geht. Davon kann aber keine Rede sein. Der Bias bleibt, auch wenn σ verschwindet und damit die Regressorvariable fehlerfrei wird. Der Bias kann nur dadurch verkleinert werden, daß die Abbruchstelle m weiter hinausgeschoben wird. Auch ist i. a. dann mit einem kleinen Bias zu rechnen, wenn die $\mid \xi_i \mid$ klein sind und damit der erwartete Approximationsfehler

$$EF_m(x_i) = \sum_{k=m+1}^{\infty} c_k \xi_i^k$$

klein ausfällt. So gesehen, liegt es nahe, die Taylorreihe nicht um den Nullpunkt zu entwickeln wie bei Stefanski, sondern um eine Stelle im Zentrum der ξ_i, also etwa um $\bar{x}$. Das würde den erwarteten Approximationsfehler insgesamt verkleinern. Von σ ist dieser aber auf jeden Fall unabhängig.

Stefanski schlägt noch eine ganz andere Reihenentwicklung für ψ_K vor:

$$\psi_K(x) = \sum_{k=0}^{\infty} \frac{(-\sigma^2)^k}{2^k k!} \psi^{(2k)}(x)$$

mit einer entsprechenden Approximation $\tilde{\psi}_{Km}(x)$, wenn die Summe bei $k = m$ abgebrochen wird. Der dazugehörige Approximationsfehler

$$\tilde{F}_m(x) = \psi_K(x) - \tilde{\psi}_{Km}(x) = \sum_{k=m+1}^{\infty} \frac{(-\sigma^2)^k}{2^k k!} \psi^{(2k)}(x)$$

ist von der Ordnung $O(\sigma^{2m+2})$, so daß hier von einer Klein-σ-Asymptotik gesprochen werden kann.

Man könnte daher vorschlagen, daß man von der Approximation $\tilde{\psi}_{Km}$ dann Gebrauch machen sollte, wenn σ klein ist, wobei man dann u.U. sogar mit einem kleinen m auskommen mag. Die Approximation ψ_{Km} (eventuell mit einem anderen Entwicklungszentrum) sollte dagegen dann (u.U. mit großem m) verwendet werden, wenn σ groß ist. Das ist freilich nur ein grobes, qualitatives Kriterium, das von Fall zu Fall auf seine Gültigkeit hin überprüft werden müßte.

Natürlich ist es das beste, wenn es gelingt, die KQSF ψ_K exakt anzugeben, wie etwa in dem oben angeführten Beispiel der Exponentialverteilung oder wie im Falle einer polynomialen Regression, vgl. Stefanski (1989) und Cheng und Schneeweiß (1996). Wenn das aber nicht möglich ist, dann könnte das genannte Kriterium einen Hinweis darauf geben, welche Art der Approximation am geeignetsten ist, um die von den Fehlern in den Variablen herrührende Verzerrung der Schätzung $\hat{\beta}$ möglichst klein zu halten.

LITERATUR

BUONACCORSI, J. P. (1996) A modified estimating equation approach to correcting for measurement error in regression. *Biometrika* **83**, 433-440.

CARROLL, R. J., RUPPERT, D. and STEFANSIK, L.A. (1995) Measurement Error in Nonlinear Models. *London: Chapman and Hall.*

CHENG, C-L. and SCHNEEWEISS, H. (1996) The polynomial regression with errors in the variables. *Discussion Paper* 42, *Sonderforschungsbereich* 386, *Universität München.*

FAHRMEIR, L. and TUTZ, G. (1994) Multivariate Statistical Modelling based on Generalized Linear Models. *New York: Springer.*

FULLER, W. A. (1987) Measurement Error Models. *New York: Wiley.*

HUBER, P. J. (1967) The behavior of maximum likelihood estimates under nonstandard condition. *Proc. Fifth Berkeley Symposium on Mathematical Statistics and Probability* **1**, 221-233.

NAKAMURA, T. (1990) Corrected score function for errors-in-variables models. *Biometrika* **77**, 127-137.

SCHNEEWEISS, H. and MITTAG, H. J. (1986) Lineare Modelle mit fehlerbehafteten Daten. *Heidelberg: Physica-Verlag.*

STEFANSKI, L.A. (1989) Unbiased estimation of a nonlinear function of a normal mean with application to measurement error models. *Commun. Statist.-Theory Methods* **18**, 4335-4358.