Semi-Nichtparametrische Maximum-Likelihood Schätzung im binären Regressionsmodell

Gabler, Siegfried; Laisney, Francois; Lechner, Michael

Semi-Nichtparametrische Maximum-Likelihood Schätzung im binären Regressionsmodell

Von Siegfried Gabler, François Laisney und Michael Lechner

Die in den Sozialwissenschaften, den Wirtschaftswissenschaften und der Biometrie bekanntesten Modelle für binäre abhängige Variablen sind das Probit- und Logitmodell. Als Verteilung der Fehlervariablen verwendet man dabei die Normalverteilung beziehungsweise die Logistische Verteilung. Beide Modelle liefern in der Regel ähnliche Schätzungen. Ist die Verteilung der Fehlervariablen schief, so führen Tests schnell zur Ablehnung beider Modelle. Beim Gallant-Nychka Ansatz wird die Verteilung geeignet approximiert. Neben den Parametern des Modells sind simultan die Parameter der Verteilung zu schätzen. Simulationen zeigen, daß die Schätzung bei normalverteilten Fehlervariablen fast genauso effizient wie im Probitmodell, jedoch viel besser bei Abweichungen von der Normalverteilung ist.

1. Das Modell

Binäre Regressionsmodelle beschreiben den Zusammenhang zwischen der Wahrscheinlichkeit für das Eintreten eines Ereignisses und möglichen Einflußgrößen. Die Beschreibung erfolgt über das lineare Modell

$$y_i^* = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + u_i \qquad i = 1, \ldots, N.$$

$\beta = (\beta_0, \beta_1, \ldots, \beta_p)'$ ist der unbekannte Parametervektor und u_i sind nichtbeobachtbare Fehlervariablen. Wir fassen die Werte der exogenen Variablen im Vektor $x_i = (1, x_{i1}, \ldots, x_{ip})'$ zusammen. Das obige Modell läßt sich dann auch in der Form

$$y_i^* = \beta' x_i + u_i \qquad i = 1, \ldots, N$$

schreiben. Allgemeiner kann man

$$y_i^* = f(x_i, \beta) + u_i \qquad i = 1, \ldots, N$$

betrachten, wobei f eine bekannte Funktion ist. Beobachten wir nur $y_i = I(y_i^*)$ mit $I(v) = 1$, falls $v > 0$ und $I(v) = 0$, falls $v \leq 0$, so erhalten wir ein Modell mit binären Antwortvariablen. Beispiele sind überall dort zu finden, wo man nur feststellt, ob ein bestimmtes Ereignis vorliegt oder nicht. In einer Partie Waren werden Stücke danach klassifiziert, ob sie gut oder schlecht sind; die Behandlung von Kranken mit bestimmten Mitteln ist erfolgreich oder nicht; bei einer Meinungsumfrage ist man für oder gegen eine bestimmte Maßnahme. Dabei hängt es von bestimmten Faktoren ab, die eher zu einem Wert $y_i = 1$ als zu $y_i = 0$

führen: Ob ein Stück gut oder schlecht wird, hängt beispielsweise von der Einstellung von Maschinen ab; eine erfolgreiche Behandlung eines Patienten ist von der verabreichten Dosis eines Medikamentes abhängig; ob die Partei A gewählt wird oder nicht, hängt nicht zuletzt davon ab, welches Einkommen der Befragte besitzt, wie alt er ist usw.
Üblicherweise setzt man voraus, daß die u_i voneinander unabhängig und identisch verteilt sind. Weiter sei auch u_i von x_i unabhängig. Dies sichert unter schwachen Zusatzvoraussetzungen bis auf Skalierung die Identifizierbarkeit des Modells (Manski 1988). Ist F(u) die Verteilungsfunktion von u_i , so läßt sich die Wahrscheinlichkeit, daß y_i den Wert 1 annimmt, berechnen als

$$P(y_i = 1 \mid x_i) = P(y_i^* > 0 \mid x_i) = P(u_i > -\beta' x_i \mid x_i) = 1 - F(-\beta' x_i).$$

$P(y_i = 1 \mid x_i)$ ist gleich dem Erwartungswert $E(y_i \mid x_i)$, da y_i nur die Werte Eins und Null annimmt. Die entsprechende Likelihoodfunktion bei N Beobachtungen läßt sich in der Form

$$L = \prod_{k=1}^{N} [1 - F(-\beta' x_k)]^{y_k} [F(-\beta' x)]^{1 - y_k}$$

schreiben. In der Literatur am bekanntesten sind die Fälle, in denen F(u) die Normalverteilung beziehungsweise die logistische Verteilung ist. Im Falle der Normalverteilung erhalten wir das Probitmodell, im Falle der logistischen Verteilung das Logitmodell. Beide Modelle führen in der Regel zu ähnlichen Schätzergebnissen für den unbekannten Parameter β und sind in der Literatur ausgiebig behandelt. Wir verweisen auf Finney (1971), Amemiya (1981), Aldrich/Nelson (1984) und Cox/Snell (1989). Andere Verteilungen wurden in der Vergangenheit wiederholt vorgeschlagen und fanden mehr oder weniger Verbreitung. In den üblichen Programmpaketen findet man meist jedoch nur die Logit- und Probitanalyse. Ein Test der Verteilungsannahme, wie der Bera Jarque Lee Test, führt häufig zur Ablehnung dieser Modelle, wie Lechner (1990) gezeigt hat. Ist die tatsächliche Verteilung schief, so führen Logit- und Probitanalyse zu schlechten Ergebnissen. Über Heteroskedastie, Nichtlinearität und fehlende Variablen muß man sich ebenfalls Gedanken machen. Stukel (1988) führte eine Klasse von Modellen indiziert durch zwei Parameter ein, die das Logitmodell als Spezialfall enthält. Unsere Überlegungen gehen dahin, die Daten selbst sprechen zu lassen und neben dem Parameter β auch die Verteilung F(u) aus den Daten schätzen zu lassen. Das Hilfsmittel dazu gibt uns der im nächsten Kapitel vorgestellte Ansatz von Gallant/Nychka (1987).

2. Der Ansatz von Gallant und Nychka

Gallant/Nychka (1987) schlagen vor, irgendeine glatte Dichte (mit Erwartungswert Null) durch

$$h(u) = \sum_{i,j=0}^{K} \alpha_i \alpha_j u^{i+j} e^{-(u/d)^2}$$

zu approximieren. Dabei sind zwei Nebenbedingungen zu berücksichtigen. Die Funktion h(u) muß eine Dichte sein, das heißt das Integral über die reelle Zahlengerade muß

Eins sein, und der Erwartungswert einer gemäß h(u) verteilten Zufallsvariablen muß Null sein. Mit wachsendem K wird die Approximation immer besser. Es sei bemerkt, daß die Koeffizienten mit geraden Indizes symmetrische Abweichungen von der Normalverteilung erzeugen, während die Koeffizienten mit ungeraden Indizes asymmetrische Abweichungen verursachen. Verwenden wir h(u) beziehungsweise die entsprechende Verteilungsfunktion H(u) als Näherung für die wahre Verteilungsfunktion F(u) in der Likelihoodfunktion L, so erhalten wir eine Pseudo–Likelihoodfunktion

$$PL = \prod_{k=1}^{N}[1 - H(-\beta' x_k)]^{y_k}[H(-\beta' x_k)]^{1 - y_k}.$$

Diese wird als Funktion der vorkommenden Parameter $\alpha = (\alpha_0, ..., \alpha_K)'$ und β maximiert. Pseudo-Maximum-Likelihoodschätzungen und ihre Eigenschaften werden etwa in White (1982) untersucht. Da die Multiplikation von

$$y_i^* = \beta' x_i + u_i, \qquad i = 1, ..., N$$

mit einer positiven Konstanten den Wert von $y_i = I(y_i^*)$ nicht ändert, können wir d=$\sqrt{2}$ setzen. Bezeichnet weiter M(u) die Matrix mit $u^{i+j} exp\{-u^2/2\}$ als Element in der i+1-ten Zeile und j+1-ten Spalte (i, j=0,1,...,K), so läßt sich h(u) als quadratische Form

$$h(u) = \alpha' M(u)\alpha$$

schreiben. Da h(u) $\geq$ 0 ist, erhält man durch die Normierung

$$\frac{\alpha' M(u)\alpha}{\int_{-\infty}^{\infty} \alpha' M(v)\alpha dv} = \frac{\alpha' M(u)\alpha}{\alpha' M\alpha}$$

eine Dichte. Dabei steht in der Matrix M an (i+1, j+1)-ter Stelle (i, j=0,1,2,...K) gerade das zentrale Moment μ(i+j) der Standardnormalverteilung multipliziert mit $\sqrt{2\pi}$. Es ist μ(0)=1, μ(2k+1)=0 für alle natürlichen Zahlen k, und μ(2k+2) = (2k+1)$\mu(2k)$. Die Matrix M ist positiv definit. Da die Multiplikation von α mit einem Skalar den letzten Ausdruck unverändert läßt, wählen wir $\alpha_0 = (2\pi)^{-0.25}$ Diese Normierung führt im Falle K=0 zur Standardnormalverteilung. Mit der gewählten Normierung schließen wir zwar den theoretisch möglichen Fall $\alpha_0 = 0$ aus, was einem Wert von Null der Dichte am Erwartungswert entspricht. Derartige Dichten können praktisch aber auch durch Variieren der anderen α Werte erzeugt werden. Die zweite Nebenbedingung lautet

$$\int_{-\infty}^{\infty} uh(u)du = 0 \quad d.h. \qquad \alpha' Q\alpha = \sum_{i,j=0}^{K} \alpha_i\alpha_j\mu(i + j + 1) = 0.$$

Die symmetrische K+1 $\times$ K+1 Matrix Q hat für ungerades K vollen Rang, für gerades K den Rang K. Die Komponenten des Vektors $a = (a_0, a_1, ..., a_K)'$ mit Qa=0 sind gegeben durch

$$a_{2i+1} = 0 \quad und \quad a_{2i} = (-1)^i \frac{\binom{K/2}{i}}{\mu(2i + 2)} a_0 \qquad i = 0, 1, ..., K/2.$$

Die logarithmierte Pseudo-Likelihoodfunktion läßt sich schreiben als

$$ln\, PL(\alpha,\beta) = \sum_{k=1}^{N} ln \frac{\alpha' Z(-\beta' x_k)\alpha}{\alpha' M \alpha}$$

wobei $Z(-\beta' x_k)$ das Element

$$y_k \mu(i+j)\sqrt{2\pi} + (1-2y_k)\int_{-\infty}^{-\beta' x_k} u^{i+j} e^{-u^2/2} du$$

an der (i+1,j+1)-ten Stelle enthält (i,j=0,1,...,K). Die Maximierung von ln PL wird unter den Nebenbedingungen $\alpha_0 = (2\pi)^{-0.25}$ und $\alpha' Q \alpha = 0$ vorgenommen. Da die Koeffizienten vor α_k^2 alle verschwinden, können wir $\alpha' Q \alpha = 0$ nach α_1 auflösen, wenn der Koeffizient vor α_1 nicht gleich Null ist. (In diesem Falle ließe sich dann α_2 als Linearkombination der übrigen α_i schreiben.) Wir haben ein GAUSS(Version 2.0 R 40) Programm zur Lösung dieser Maximierungsaufgabe geschrieben, das auf die MAXLIK Prozedur von GAUSS zurückgreift. Analytische erste und zweite Ableitungen bezüglich der Parameter werden verwendet. Da einerseits größere K zu numerischen Schwierigkeiten und langen Rechenzeiten führen können, andererseits die Pseudo-Maximum-Likelihoodtheorie Pseudo-Score- und Pseudo-Wald-Tests ermöglicht, sollte etwa bei K=3 das Probit Modell getestet und K notfalls schrittweise erhöht werden. (Bei K=2 kann das Probit Modell nicht getestet werden, da die Pseudo-Likelihoodfunktion in diesem Fall ein lokales Extremum besitzt.) Relativ einfach ist ein Test auf Heteroskedastie, wenn der Fehler in der k−ten Beobachtung als

$$u_k = e_k exp\{z_k' \tau\}$$

geschrieben werden kann. Die e_k sind unabhängig und identisch verteilt. Der Vektor τ ist unbekannt und z_k beinhaltet solche Variablen, die die Varianz der Fehlervariablen der k-ten Beobachtung beeinflussen. Dabei spielt es keine Rolle, ob sie in den exogenen Variablen bereits vorkommen oder nicht. Da der multiplikative Term $exp\{z_k' \tau\}$ stets positiv ist, gilt

$$y_k = I(\beta' x_k + e_k exp\{z_k' \tau\}) = I(\beta' x_k / exp\{z_k' \tau\} + e_k).$$

Wenn τ der Nullvektor ist, haben wir das ursprüngliche Modell. Mit einem Pseudo-Score-Test kann die Nullbedingung getestet werden.

3. Eine Simulation

Um das Verhalten der Tests und Schätzer besser kennenzulernen, haben wir Daten nach dem folgenden Modell erzeugt:

$$y_i^* = \beta_0 + x_{ni}\beta_n + x_{ui}\beta_u + u_i.$$

Dabei ist $x_{ni} = v_{ni} - \bar{v}_n - 0.5$, wobei v_{ni} einer Standardnormalverteilung genügt, und $x_{ui} = w_{ui} - \bar{w}_u - 0.5$, wobei w_{ni} einer Gleichverteilung auf dem Intervall $[0;1]$ genügt. $\bar{v}_n$ beziehungsweise $\bar{w}_u$ bezeichnet das jeweilige arithmetische Mittel der Beobachtungen für die Variablen v_n und w_u. Setzt man $\beta_0 = \beta_u = \beta_n = 1$, so ergibt sich für die latente Variable y_i^* ein Erwartungswert von Null. Als Verteilung von u_i haben wir einmal die Standardnormalverteilung und zum andern eine Burr II Verteilung mit Parameter 0.05 gewählt. Die Dichte einer Burr II Verteilung ist durch

$$g(t \mid r) = \frac{r \cdot e^{-t}}{(1 + e^{-t})^{r+1}},$$

gegeben. r=1 ist die logistische Verteilung. r=0.05 ist eine schiefe Verteilung. Die Stichprobenumfänge waren 300 beziehungsweise 1000. Als K verwendeten wir K=3 und K=5. Die Simulation ergab, daß der Gallant-Nychka Ansatz im Falle der Normalverteilung der Fehlervariablen nur unwesentlich schlechtere Schätzer lieferte als die Probit Analyse. Im Falle der Burr II Verteilung zeigte sich die Anpassungsfähigkeit des Gallant-Nychka Ansatzes, während die Probit Analyse teilweise katastrophale Ergebnisse lieferte, besonders wenn die Güte nicht am mittleren quadratischen Fehler der Parameterschätzung, sondern an der Schätzung der Antwortwahrscheinlichkeiten, also der Struktur, gemessen wird.

Wir haben den Gallant-Nychka Ansatz auch auf reale Daten angewendet und dabei festgestellt, daß er nützliche und interessante Ergebnisse liefert. Heteroskedastie erzeugende Variablen konnten deutlicher herausgefunden werden als bei Verwendung des Probit Ansatzes.

Es soll nicht verschwiegen werden, daß der Gallant-Nychka Ansatz auch seine Nachteile hat. Die Pseudo-Loglikelihoodfunktion ist nicht mehr länger konkav, die Konvergenz des Maximierungsprozesses hängt von den Startwerten ab, die Rechenzeit auf dem Computer kann wesentlich länger sein als beim Probitmodell. Da unserer Meinung nach die Vorteile jedoch überwiegen, haben wir versucht, eine Lanze für diesen Ansatz zu brechen. Eine ausführliche Auflistung der Simulationsergebnisse und der Anwendung auf reale Daten würde den Rahmen dieser Arbeit sprengen. Interessierte Leser können sich jedoch gerne an die Autoren wenden.

Literatur

Aldrich, J.H./D.N.Nelson, 1984: Linear Probability, Logit and Probit Models, Beverly Hills:Sage University Press.

Amemyia, T., 1981: Qualitative Response Models: A Survey. Journal of Economic Literature 19:1483-1536.

Cox, D.R./E.J. Snell, 1989: Analysis of Binary Data, London: Chapman and Hall.

Finney, D.J., 1971: Probit Analysis, Cambridge: University Press.

Gallant, A.R./D.N. Nychka, 1987: Semi-Nonparametric Maximum Likelihood Estimation. Econometrica 55:363-390.

Lechner M., 1990: Testing Logit Models in Practice. Discussion Paper No.408-90. Institut für Volkswirtschaftslehre und Statistik der Universität Mannheim. (Erscheint in: Empirical Economics).

Manski, C.F., 1988: Identification of Binary Response Models. Journal of the American Statistical Association 83:729-738.

Stukel, T.A., 1988: Generalized Logistic Models. Journal of the American Statistical Association 83:426-431.

White, H., 1982: Maximum Likelihood Estimation of Misspecified Models. Econometrica, 50:1-25.