ТЕОРІЯ ОПТИМАЛЬНИХ РІШЕНЬ

УДК 519.21

З.В. НЕКРЫЛОВА

О МОДЕЛЯХ С ЗАВИСИМОЙ ПЕРЕМЕННОЙ НЕПРЕРЫВНОГО И ДИСКРЕТНОГО ВИДА

Изучаются модели, имеющие в зависимой переменной как дискретную, так и непрерывную части. Одна из них – Tobit-модель, которая является развитием probit-модели, но и одним из подходов к проблеме цензурования. Предлагаются два возможных метода состоятельного оценивания даже при условиях гетероскедастичности. Описываются класс проблем, включающих обработку данных и результат, и двухшаговый метод для их решения.

Введение. Для изучаемого класса моделей характерно, что зависимая переменная имеет как дискретную, так и непрерывную части. Одной из важнейших моделей такого типа является Tobit-модель (название происходит от фамилии автора – Tobin [1]). Модель, в частности, является развитием probit-модели [2], однако на самом деле – это один из подходов к проблеме цензурования данных.

Tobit-модель как развитие probit-модели. Рассмотрим, например, проблему покупки. Переменная y^* характеризует желание персоны купить. Определим переменную y, равную 1, если персона покупает, 0 – в противном случае. Формально описанную проблему можно сформулировать как probit-модель. Если же вместо простого констатирования о покупке возможно также сообщить и величину ее стоимости, то имеет место одно из природных развитий probit-модели, названное Tobit-моделью (Tobin's probit), которое задается в виде

$$y_i^* = X_i\beta + \varepsilon_i, \qquad (1)$$

где $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$, $y_i = \begin{cases} y_i^*, & y_i^* > 0, \\ 0, & y_i^* \le 0. \end{cases}$

Эта модель известна как цензурованная регрессионная модель (censored regression model), так как дает возможность рассматривать проблему как такую, в которой наблюдения y^* в нуле и ниже нуля подвергаются цензуре, т.е. модель можно записать в виде

$y_i = \max\{0,\ X_i\beta + \varepsilon_i\}$. Следует отличать такое представление от операции усечения, когда переменная X_i не наблюдается, но при этом отсутствует и отвечающая ей переменная y_i^*.

Оценим величины

$$P(y_i = 0) = P(y_i^* \le 0) = P(X_i\beta + \varepsilon_i \le 0) = P\left(\frac{\varepsilon_i}{\sigma} \le -\frac{X_i\beta}{\sigma}\right) = 1 - \Phi\left(\frac{X_i\beta}{\sigma}\right),$$

$$P(y_i = y_i^*) = P(X_i\beta + \varepsilon_i = y_i) = P\left(\frac{\varepsilon_i}{\sigma} = \frac{y_i - X_i\beta}{\sigma}\right) = \frac{1}{\sigma}\phi\left(\frac{y_i - X_i\beta}{\sigma}\right),$$

где Φ и ϕ – функция распределения и плотность нормального распределения, соответственно. Тогда функцию правдоподобия можно записать в виде

$$L = \prod_{y_i/y_i=0} P(y_i = 0) \prod_{y_i/y_i>0} P(y_i = y_i^*) = \prod_{y_i/y_i=0}\left(1 - \Phi\left(\frac{X_i\beta}{\sigma}\right)\right)\prod_{y_i/y_i>0}\frac{1}{\sigma}\phi\left(\frac{y_i - X_i\beta}{\sigma}\right). \quad (2)$$

Заметим, что, во-первых, вторая часть этой функции имеет сходство с функцией правдоподобия условного обыкновенного метода наименьших квадратов (ОНК-метод) для точек выборки, не подвергающихся цензуре (т.е. больших нуля). Первая часть встречается в функции правдоподобия probit-модели. Во-вторых, если в probit-модели нормализация $\sigma = 1$ является безвредной, то в данном случае это совсем не так, что создаёт серьёзные проблемы при наличии гетероскедастичности. Это следствие того, что в первой части (2) «идентифицируется» отношение β/σ (как в probit-модели), а во второй части σ и β идентифицируются порознь (как при ОНК-оценивании). В-третьих, возможно не всегда имеет смысл интерпретировать коэффициенты Tobit-модели, как это принято в нецензурованной линейной регрессионой модели. Рассмотрим следующие три производные относительно частной переменной X_k для наблюдения i.

Для условного математического ожидания линейной регрессии (1)

$$\mathrm{E}(y_i^*/X_i) = X_i\beta \text{ это будет } \partial\mathrm{E}(y_i^*/X_i)/\partial X_k = \beta_k.$$

Для условного математического ожидания probit-модели

$$\mathrm{E}(y_i/X_i) = P(y_i^* > 0) \text{ имеем } \partial P(y_i^* > 0)/\partial X_k = \beta_k\phi(X_i\beta/\sigma)/\sigma.$$

Для условного математического ожидания Tobit-модели

$$\mathrm{E}(y_i/X_i, y_i^* > 0) = \int_0^\infty z_i dP(z_i/X_i, y_i^* > 0) = X_i\beta + \frac{\sigma^2}{\beta}\frac{\phi(X_i\beta/\sigma)}{\Phi(X_i\beta/\sigma)},$$

$$\partial\mathrm{E}(y_i/X_i, y_i^* > 0)\Big/\partial X_k = \beta_k\left[1 - \frac{X_i\phi(X_i\beta/\sigma)}{\Phi(X_i\beta/\sigma)} - \frac{\sigma\phi^2(X_i\beta/\sigma)}{\beta\Phi^2(X_i\beta/\sigma)}\right].$$

Для Tobit-модели используют и условное математическое ожидание вида

$$\mathrm{E}\left(y_i/X_i\right) = P\left(y_i^* > 0\right)\cdot \mathrm{E}\left(y_i/X_i, y_i^* > 0\right) + 0\cdot P\left(y_i^* = 0\right),$$

тогда

$$\frac{\partial \mathrm{E}\left(y_i/X_i\right)}{\partial X_k} = \frac{\partial P\left(y_i^* > 0\right)}{\partial X_k}\mathrm{E}\left(y_i/X_i, y_i^* > 0\right) + P\left(y_i^* > 0\right)\ \frac{\partial \mathrm{E}\left(y_i/X_i, y_i^* > 0\right)}{\partial X_k}.$$

Такая декомпозиция была предложена в [3] и многие находят ее полезной. Любая из этих интерпретаций коэффициентов модели может быть интересной и используемой на практике в зависимости от изучаемой проблемы.

И, наконец, если истинная модель есть Tobit, то игнорирование цензурования и использование ОНК-оценивания приводит к некорректности. Действительно, для модели (1) такая оценка имеет вид

$$\hat{\beta}_{ОНК} = \left(X'X\right)^{-1}X'\ \ y\cdot P\left(y^* > 0\right) + 0\cdot P\left(y^* \le 0\right) = \left(X'X\right)^{-1}X'\ \ y\cdot P\left(y^* > 0\right).$$

Так как $P\left(y^* > 0\right) < 1$, то ОНК-оценка будет ослаблена. Если есть возможность определить состоятельную оценку $P\left(y^* > 0\right)$, например, в виде n_1/N, где N – общее число наблюдений, а n_1 – те, которые больше нуля, то состоятельной оценкой β будет $\hat{\beta}_{ОНК}\, N/n_1$.

Даже при оценивании Tobit-модели с использованием ее функции правдоподобия возможны случаи, когда Tobit-модель приводит к соотношению, которое не всегда соответствует истине. Один из подходов для понимания происходящего состоит в рассмотрении теста спецификации, основанного на следующем факте. При правильной спецификации Tobit-модели отношение оценок максимального правдоподобия для Tobit-модели, $\hat{\beta}_T/\hat{\sigma}_T$, должно быть таким же, как и для probit-модели, $\hat{\beta}_p/\hat{\sigma}_p$, на одних и тех же данных, если трактовать ненулевое значение как 1, а нулевое как 0. То есть для Tobit-модели требуется такое условие, чтобы соотношение, генерирующее нули и единицы, было бы таким же, как и процесс, производящий положительные значения. Для проверки может оказаться полезным тест Хаусмана [4], основанный на этом факте. В крайнем случае можно прибегнуть к визуальному сравнению величин $\hat{\beta}_T/\hat{\sigma}_T$ и $\hat{\beta}_p/\hat{\sigma}_p$, что предлагается в [5]. Сильное различие этих оценок будет свидетельствовать о неправильной спецификации Tobit-модели.

Возможные способы оценивания. Трудности при оценивании Tobit-моделей возникают из-за того, что y^* не наблюдаются, а ошибки наблюдения для y не являются симметрично распределенными. Ранее отмечалось, что и гетероскедастичность создаёт серьёзные проблемы при оценивании. Однако в [6], [7] предложены способы получения состоятельных оценок Tobit-моделей даже при наличии гетероскедастичности, что вкратце излагается далее. Один из таких

методов назван симметрично отсечённым методом наименьших квадратов (symmetrically trimmed least squares) [6]. Для каждой заданной независимой переменной предлагается наблюдать не только значение зависимой переменной, а целую область ее значений, лежащую правее 0 и левее $2X_i\beta$, не включая значения левее 0 и вычеркивая те, что правее $2X_i\beta$. Если таким образом «отсекать» данные, то распределение становится симметричным, и можно будет использовать ОНК-оценивание для функции цели следующего вида:

$$F(\beta) = \sum_{i=1}^{n}\left(y_i - \max\left(\frac{1}{2}y_i, X_i\beta\right)\right)^2 + \sum_{i=1}^{n} 1_{y_i > 2X_i\beta}\left(\left(\frac{1}{2}y_i\right)^2 - \max^2\left(0, X_i\beta\right)\right).$$

Для нахождения состоятельной оценки предлагается следующий алгоритм:

1. Находим оценку β, скажем, ОНК-методом на исходных данных.
2. Вычисляем предсказанную величину:
   если её значение отрицательное, то опускаем ее;
   если значение зависимой переменной больше удвоенного значения предсказанной величины, то считаем его равным $2X_i\beta$.
3. Применяем ОНК-оценивание к этим измененным данным.
4. Полученную оценку β используем для исходных данных как в п. 2.
5. Повторяем оценивание β, пока изменяется значение оценки.

Привлекательной особенностью этого метода является его устойчивость к гетероскедастчности ошибок наблюдения, но при условии симметричности и унимодальности их распределения. Он наиболее полезен, когда результат цензурования не является очень строгим. Однако отмечается, что процедуру не следует использовать для ограниченного количества данных, что может привести к потере эффективности оценки.

Вместо ОНК-оценивания в [7] предлагается использовать метод наименьших абсолютных величин (НАВ-метод), что соответствует регрессии для медианы и ослабляет требования на ошибки наблюдения. Снова объем наблюдений не должен быть ограниченным. В качестве функции цели берется функция

$$\Psi(\beta) = \sum_{i=1}^{n}\left|y_i - X_i\beta\right| \text{ или } \Psi(\beta) = \sum_{i=1}^{n}\left(y_i - X_i\beta\right)\operatorname{sgn}\left(y_i - X_i\beta\right),$$

где $\operatorname{sgn}(\cdot)$ принимает значения – 1, 1, 0 в зависимости от того, является ли аргумент отрицательной, положительной величиной или нулем, соответственно.

При НАВ-оценивании важен знак остатков, а не их значение. Следует отметить, что медиана распределения y^* не изменяется при переходе к y в отличие от математического ожидания, как было видно из вышеприведенных вычислений. Причем это сохраняется при самых общих формах распределения ошибок наблюдения. В частности, предположения о гомоскедастичности и нормальности уже не обязательно. Итеративная процедура остаётся такой же, как и при ОНК-оценивании, только применяется НАВ-метод и в пункте 2 отсутствует вто-

рой подпункт. Такой подход не гарантирует нахождения глобального минимума, что можно исправить, если начинать вычисление с различных начальных точек. Появляется проблема и с вычислением стандартных ошибок, для разрешения которой в [7] даются некоторые рекомендации.

Эффекты обработки данных и двухшаговые методы. Составление перечня моделей, сочетающих в себе ограниченные и непрерывные переменные, является довольно трудной задачей, однако в [8] можно найти достаточно хороший обзор таких моделей.

Далее будет кратко описан класс проблем, включающих переменные, представляющие обработку (обычно дихотомическая величина) и результат (обычно непрерывная величина). Терминология взята из биологии, медицины, когда дихотомическая переменная оценивания часто связана с использованием нового препарата или терапевтического режима, а результат – с мерой последствий обработки. В работе [9] Хекман описал общий вид модели такого типа:

$$y_{1i} = X_{1i}\beta_1 + \varepsilon_{1i}, \tag{3}$$

$$y_{2i} = X_{2i}\beta_2 + \varepsilon_{12i}, \tag{4}$$

$$T_i = 1\left(Z_i\gamma + \varepsilon_{0i} > 0\right), \tag{5}$$

$$y_i = T_i y_{1i} + \left(1 - T_i\right) y_{2i}, \tag{6}$$

где T_i – переменная, связанная с обработкой, которая принимает значение 1 или 0 в зависимости от того, является ли утверждение $1(\cdot)$ истинным или ложным, соответственно. Непрерывные меры y_{1i} и y_{2i} соответственно описывают связь между результатом и появившимися изменениями, связанными с тем, подвергался или нет индивид обработке.

В самой сути этой модели заложены два возможных исхода, которые названы (довольно произвольно):

– разнородность эффекта обработки, когда обработка меняется от индивида к индивиду в зависимости от их характеристик;

– наличие отбора, когда присутствует некоторая характеристика обработки, которая связана как со средством обработки, так и с получением результата, что приводит к приписыванию ложной причинной связи.

Хекман предложил в [10] простой двухшаговый метод для многих из моделей (3)–(6), чаще всего используемый в случае, когда имеет место отбор, т.е. Z и X могут включать общие переменные, а могут быть и идентичными. Для простоты ограничимся уравнениями (3) и (5). Чтобы понять следствие использования отбора, возьмём условное математическое ожидание от (3) при условии, что проведена обработка, т.е.

$$\mathrm{E}\left(y_{1i}/X_{1i}, T_i = 1\right) = X_{1i}\beta_1 + \mathrm{E}\left(\varepsilon_{1i}/\varepsilon_{0i} > -Z_i\gamma\right).$$

Величину $\mathrm{E}\left(\varepsilon_{1i}/\varepsilon_{0i} > -Z_i\gamma\right)$ можно определить, если знать вид зависимости ε_{1i} от ε_{0i}. При совместном нормальном распределении ε_{1i} и ε_{0i} из линейной среднеквадратической регрессии следует, что $\varepsilon_{1i} = \sigma_{0,1}^{(i)}\varepsilon_{0i}/\sigma_{0i}^2$, где σ_{0i}^2 – дисперсия ε_{0i}, $\sigma_{0,1}^{(i)}$ – корреляция между ε_{1i} и ε_{0i}, а $\mathrm{E}\varepsilon_{0i} = \mathrm{E}\varepsilon_{1i} = 0$. Тогда

$$\mathrm{E}\left(\varepsilon_{1i}/\varepsilon_{0i} > -Z_i\gamma\right) = \frac{\sigma_{0,1}^{(i)}}{\sigma_{0i}}\mathrm{E}\left(\frac{\varepsilon_{0i}}{\sigma_{0i}}\Big/\frac{\varepsilon_{0i}}{\sigma_{0i}} > \frac{Z_i\gamma}{\sigma_{0i}}\right) = \frac{\sigma_{0,1}^{(i)}}{\sigma_0}\frac{\phi\left(Z_i\gamma/\sigma_{0i}\right)}{\Phi\left(Z_i\gamma/\sigma_{0i}\right)},$$

где величина $\phi\left(Z_i\gamma/\sigma_{0i}\right)/\Phi\left(Z_i\gamma/\sigma_{0i}\right)$ известна как обратное отношение Миллса. Подход Хекмана заключается в том, чтобы при использовании ОНК-метода для (3), (5) оценивать регрессию для условного математического ожидания в виде

$$\mathrm{E}\left(y_{1i}/X_{1i}, T_i = 1\right) = X_{1i}\beta + \alpha_i\frac{\phi\left(Z_i\gamma/\sigma_{0i}\right)}{\Phi\left(Z_i\gamma/\sigma_{0i}\right)}, \qquad (7)$$

для чего предлагается использовать следующий двухшаговый метод:

– оценивается probit-модель обработки (5) для получения оценки γ/σ_{0i};

– она используется для оценивания обратного отношения Миллса;

– с помощью ОНК-метода оценивается регрссия (7).

Оценку $\sigma_{0,1}^{(i)}/\sigma_{0i}$ можно выразить как коэффициент α_i из обратного отношения Миллса, а получение стандартной ошибки можно найти в [8]. Аналогично оцениваются и уравнения (4), (6). Описанный метод очень чувствителен к нарушениям предположений, в которых он используется, поэтому в [11] предлагается сначала тестировать присутствие следствия отбора. Если обнаруживается что оно отсутствует, то рекомендуется использовать МП-оценивание.

Tobit-модель как специальный случай модели (3)–(6). Если $y_2 = 0,$ то изучаемую проблему можно интерпретировать с помощью двух латентных переменных [12] следующим образом:

$$\begin{cases} y_1^* = X\beta_1 + \varepsilon_1, \\ y_0^* = Z\gamma + \varepsilon_0, \end{cases} \quad y_i = \begin{cases} y_{1i}^*, \ y_{0i}^* > 0, \\ \text{не} \quad \text{наблюдается} \quad - \quad \text{иначе}. \end{cases}$$

Таким образом величина y_{0i}^* никогда не наблюдается, лишь отмечается ее знак, и при положительном его значении наблюдается величина y_{1i}^*, т.е. диапазон изменения y строго не ограничивается. Такие модели используются в проблемах страхования, рекламе, когда уравнение отбора (5) и уравнение результата (3)

обязательно являются различными. Если $y_2 = 0,\ X = Z,\ \gamma = \beta_1,\ \varepsilon_0 \equiv \varepsilon_1,$ то получим традиционную Tobit-модель.

Заключение. Рассмотрено несколько описаний моделей, зависимая переменная которых имеет как непрерывную, так и дискретную части. Такие модели можно использовать, например, для изучения проблем страхования, последствия действий администрации на предприятиях, рекламы и прочее. Предложены некоторые оригинальные способы их оценивания.

З.В. Некрилова

ПРО МОДЕЛІ ІЗ ЗАЛЕЖНОЮ ЗМІННОЮ НЕПЕРЕРВНОГО І ДИСКРЕТНОГО ВИГЛЯДУ

Вивчаються моделі, що мають в залежній змінній як дискретну, так і неперервну частини. Однією з них є Tobit-модель, що є розвитком probit-моделі, але й одним із підходів до проблеми цензурування. Пропонуються два можливі методи консистентного оцінювання навіть
в умовах гетероскедастичності. Наводяться модель для класу проблем, що містить обробку даних і результат, і двокроковий метод для її оцінювання.

Z.V Nekrylova

ABOUT THE MODELS WITH BOTH DISCRETE AND CONTINUOUS PARTS DEPENDENT VARIABLE

The models that have both discrete and continuous parts dependent variable are investigated. One important model is the Tobit. It is an extansion of the probit, but it is one approach the problem of censored data. Two possible solutions for consistent estimations even in the face of heteroscedasticity are discussed. A class of problems involving a treatment and an outcome and two-step method for them are discribed.

1. *Tobin J.* Estimation of Relationships for Limited Dependent Variables // Econometrica. – 1958. – 26. – P. 24–36.
2. *Некрылова З.В*. Об особенностях моделей с дискретными зависимыми переменными // Теорія оптимальних рішень. – К.: Ін-т кібернетики ім. В.М. Глушкова НАН України, 2008. – № 7 – С. 88–96.
3. *McDonald J., Moffit R.* The Uses of Tobit Analysis // Review of Economics and Statistics. – 1980. – 62. – P. 318–321.
4. *Hausman J.A.* Spesification Tests in Econometrics // Econometrica. – 1978. – 46.– P. 1251–1271.
5. *Fin T., Schmidt A.* A Test of the Tobit Specification against an Alternative Suggested by Cragg // Review of Economics and Statistics. – 1984. – 66. – P. 35–57.
6. *Powell J.* Symmetriccally Trimmed Least Squares Estimation for Tobit Models // Econometrica. – 1986. – 54. – P. 1435–1460.
7. *Powell J.* Least Absolute Deviations Estimation for the Censored Regression Models // Journal of Econometrics. –1984. – 25. – P. 303–325.
8. *Amemiya T.* Tobit-models: A Survey // Journal of Econometrics. – !984. – 24. – P. 3–63.

9. *Heckman J.* Varieties of Selection Bias // American Economic Review. – 1990. – 80. P. 313–318.
10. *Heckman J.* The Common Structure of Statistical Models of Truncation, Sample, Selection and Limited Variables and a Simple Estimator for such Models // Annals of Economic and Social Measurement. – 1976. – 5. – P. 475–492.
11. *Davidson, R.*, *MacKinnon, I. J.* Estimation and Inference in Econometrics. – Oxford: Oxford University Press. 1993. – 603 p.
12. *Hsiao C.* A Statistical Perspective on Insurance Rate–Making // Journal of Econometrics. – !990. – 44. – P. 5–24.

Получено 27.03.2009