Теорія оптимальних рішень

УДК 519.21

З.В. НЕКРЫЛОВА

ОБ ОСОБЕННОСТЯХ МОДЕЛЕЙ С ДИСКРЕТНЫМИ ЗАВИСИМЫМИ ПЕРЕМЕННЫМИ

Изучаются модели с дискретными зависимыми переменными, в частности, probit, logit и линейные модели вероятности. Описано формулирование этих моделей, обсуждается их ошибочная спецификация, особенности выбора типа модели и некоторые возможности их расширения.

Введение. В предлагаемой работе изучаются некоторые статистические модели, которые можно использовать в ситуациях, когда метод обыкновенных наименьших квадратов (ОНК) и двухшаговый метод наименьших квадратов являются непригодными. К тому же рассматриваемые модели являются нелинейными (по параметрам), поэтому в отличие от ОНК-моделей они часто не сохраняют желаемых асимптотических свойств, когда ошибки являются гетероскедастичными или не имеют нормального распределения. Таким образом, эти модели – менее устойчивые к ошибкам спецификации.

Современное развитие вычислительной техники даёт возможность использовать модели такого типа. Ограничимся моделями, в которых переменные наблюдения принимают дискретные значения или для переменных наблюдения используется разбиение на группы. Такие модели называют моделями с переменными дискретного вида или дискретными моделями [1, 2].

Типы моделей дискретного выбора. С помощью таких моделей делают попытки объяснить как дискретный выбор, так и дискретный результат. Существует, по меньшей мере, три типа дискретных переменных и для каждого из них нужна своя модель.

Дихотомические или бинарные переменные принимают два значения: 0 или 1 в зависимости от двух возможных исходов. Рассмотрим случаи, когда такие переменные находятся в левой части соотношения модели, т.е. являются эндогенными или зависимыми переменными, что порождает пробле-

му, не похожую на ту, которая возникает для случая, когда бинарными переменными являются экзогенные переменные.

Например, из биометрической литературы (в которой описаны модели с эндогенными бинарными переменными) можно привести пример оценивания действия инсектицида. Предположим, что терпимость y_i^* насекомого i к инсектициду имеет распределение $N(\mu, \sigma_\varepsilon^2)$ по всем насекомым. Насекомое гибнет, если его терпимость меньше дозы x_i инсектицида.

Проблема возникает из-за того, что невозможно наблюдать терпимость y_i^* насекомого, вместо этого можно только констатировать, живое оно или нет, что наводит на мысль ввести наблюдаемую переменную y_i:

$$\mathrm{y_i} = \begin{cases} 1, & \text{если насекомое гибнет,} \\ 0 - \text{в противном случае,} \end{cases}$$

и возвратиться к исследуемой проблеме в иной постановке: какова вероятность гибели насекомого, что сводится к вероятности того, что терпимость насекомого меньше дозы инсектицида, т.е.

$$P\{y_i = 1\} = P\{y_i^* < x_i\},$$

где наблюдения y_i генерируются по правилу

$$\mathrm{y_i} = \begin{cases} 1, & \mathrm{y_i^* < x_i}, \\ 0 - \text{иначе.} \end{cases}$$

Величину y_i^* называют латентной или показывающей переменной. Формулировка моделей в терминах латентных переменных часто аналитически удобна.

Полихотомические переменные основываются на дискретном количестве возможных значений исхода, превышающем два. Полихотомические переменные могут быть как упорядоченными, так и наоборот. Например, в экономике перевозок важно предвидеть, какой вид транспорта персона выбирает, чтобы добраться до работы и вернуться домой. В этом случае альтернативы исхода зависят от вариантов выбора персоной вида транспорта. Понятно, что данные варианты не подлежат упорядочению. Случай упорядочения переменных связан с возможностью ранжирования исходов. Например, при постановке диагноза могут быть три альтернативы: плохое, удовлетворительное и хорошее здоровье.

Специальным случаем упорядоченных переменных являются последовательные переменные при условии, что последующие исходы зависят от предыдущих. Например, в случае, когда альтернативами есть оценки степени образованности, то высшее образование включает в себя школьное, а ученая степень– и школьное, и высшее.

Возможны модели с переменными, которые могут принимать счётное число значений, например, переменная, учитывающая количество патентов в году. Такие типы дискретных моделей являются менее общепринятыми для использова-

ния в эконометрических исследованиях, и часто в этих случаях обходятся традиционными линейными моделями.

Далее будут рассматриваться только бинарные зависимые переменные, принимающие значения: 0 или 1.

Линейные модели вероятности. Пусть имеется выборка дискретных переменных вида

$$y_i = \begin{cases} 1, & \text{если персона является членом организации,} \\ 0 - \text{в противном случае,} \end{cases}$$

и рассмотрим следующую описательную модель:

$$y_i = Z_i\beta + \nu_i,$$

где Z_i – перечень мотивов, определяющих склонность персоны вступить в организацию; β –множество параметров. Бинарность переменной y_i порождает некоторые проблемы даже в интерпретации модели.

Рассмотрим оценивание этого уравнения с помощью ОНК-метода, назвав это линейной моделью вероятности (ЛВ-моделью). В случае непрерывной зависимой переменной данную регрессию обычно интерпретируют с помощью спецификации $E(y/Z)$, т. е. условного математического ожидания y при заданных значениях Z. В случае бинарной зависимой переменной недостаток ЛВ-модели состоит в том, что отсутствует ограничение, заставляющее предсказанные значения принимать значения между 0 и 1. В результате этого такие модели перестают быть интересными для применения, хотя иногда их используют в основном из-за простоты.

Формулирование модели вероятности. Полезный подход состоит в том, чтобы признать правильным намерение преобразовать $X\beta$ в вероятность, т. е. необходимость введения такой функции F, чтобы $P\{y_i = 1\} = F(X\beta)$. Естественным выбором в качестве F является функция распределения. Именно таким образом и можно определить модели с бинарным выбором.

Если F взять тождественной функцией, то получим $P\{y_i = 1\} = X\beta$, т. е. линейную модель вероятности, рассмотренную ранее, но ничего не появится дополнительного, чтобы заставить $X\beta$ принимать значения между 0 и 1.

Выбор F в виде функции распределения стандартного нормального распределения Φ является успешным подходом, приводящим к probit модели: $P\{y_i = 1\} = \Phi(X_i\beta)$, для которой $\lim_{z\to\infty} \Phi(z) = 1$ и $\lim_{z\to-\infty} \Phi(z) = 0$.

Выбрав F логистическим распределением, приходим к logit модели:

$$P\{y_i = 1\} = L(X_i\beta) = \frac{\exp X_i\beta}{1 + \exp X_i\beta}.$$

Однако не обязательно ограничиваться лишь этими двумя подходами, просто probit и logit модели – наиболее общие модели в практических ситуациях. Рассмотрим их детальнее.

Probit модель. Пусть переменная y может принимать одно из двух значений – 0 или 1. Определим латентную переменную

$$y_i^* = X_i\beta + \varepsilon_i . \tag{1}$$

Величина y^* не наблюдается, скорее это можно сказать об y, которое принимает значение 0 или 1 согласно следующему правилу:

$$y_i = \begin{cases} 1, & \text{если } y_i^* > 0, \\ 0 - \text{иначе}, \end{cases} \tag{2}$$

и пусть $\varepsilon_i \sim N(0, \sigma^2)$. В отличие от линейной модели вероятности величина y_i^* (условно относительно X_i) в модели (2) распределена нормально, а реализация y_i – нет. Чтобы показать, что (2) генерирует probit модель, отметим сначала следующее:

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

где $\varepsilon_i/\sigma \sim N(0,1)$, т. е. имеет стандартное нормальное распределение.

Для probit модели (как и logit модели) рассматриваемые распределения являются симметричными, поэтому предыдущую цепочку равенств можно продолжить:

$$P\{y_i = 1\} = P\left\{\frac{\varepsilon_i}{\sigma} > -X_i\frac{\beta}{\sigma}\right\} = P\left\{\frac{\varepsilon_i}{\sigma} < X_i\frac{\beta}{\sigma}\right\} = \Phi\left(X_i\frac{\beta}{\sigma}\right),$$

тогда $P\{y_i = 0\} = 1 - P\{y_i = 1\} = 1 - \Phi\left(X_i\frac{\beta}{\sigma}\right)$.

Если выборка состоит из n независимых одинаково распределённых наблюдений, то функцию правдоподобия для probit модели можно записать в виде

$$L = \prod_{i=1}^{m}\left(1 - \Phi\left(X_i\frac{\beta}{\sigma}\right)\right)\prod_{i=m+1}^{n}\Phi\left(X_i\frac{\beta}{\sigma}\right) = \prod_{i=1}^{n}\Phi^{y_i}\left(X_i\frac{\beta}{\sigma}\right)\left(1 - \Phi\left(X_i\frac{\beta}{\sigma}\right)\right)^{1-y_i},$$

где индексы $1,\ldots,m$ принадлежат тем наблюдениям, для которых $y_i = 0$, а индексы $m+1,\ldots,n$ – тем, для которых $y_i = 1$.

Обычно используют $\log L$, которая будет ограничена сверху нулём, так как $0 \le \Phi(\bullet) \le 1$. Другой важной особенностью функции правдоподобия является то, что параметры β и σ в ней всегда появляются вместе, в виде β/σ, т. е. они не являются идентифицируемыми порознь. Поэтому удобно нормировать σ так, чтобы она равнялась единице, тогда можно говорить только о параметре β.

Оценивание probit модели осуществляется непосредственно с помощью численных методов, даже если модель нелинейная и не существует таблиц для функции F. Причем функция правдоподобия для probit модели (и logit модели)

является глобально выпуклой. В качестве начального приближения можно взять оценку параметра, полученную из линейной модели вероятности.

Отметим, что вид знака коэффициентов в probit модели сохраняется таким же, как и в ЛВ-модели, а вот выражение для изменения вероятности в зависимости от изменения какой-либо независимой переменной модели будет уже не таким простым, как в ЛВ-модели. Если в ЛВ-модели β можно рассматривать как производную левой части относительно X, то в probit модели такая производная относительно какого-то X_i примет вид $\phi(X\beta)\beta_i$, где ϕ – плотность стандартного нормального распределения, а β_i – i-я компонента вектора β. То есть в выражении для изменения вероятности в зависимости от изменения X_i участвуют все значения X. Полученное различие можно использовать в практических целях при анализе полученных результатов в случае, если probit модель применяется не только для получения знака и оценок вектора коэффициентов β.

Logit модель. Построение logit модели идентично построению probit. Интерпретация латентной переменной осуществляется тем же способом, только в (1) ε_i подчинено так называемому распределению крайних значений (логистическому распределению) [3], что и отражает основное различие между нормальным распределение и логистическим.

Ошибочная спецификация в моделях с бинарными зависимыми переменными.

Гетероскедастичность. Обозначим регрессионную функцию модели $f(X)$ (обычно предполагается, что она линейная, т. е. $f(X)=X\beta$), а $F(\bullet)$ – соответствующую функцию распределения:

$$P(y=1)=F\left(\frac{f(X)}{\sigma}\right).$$

В случае гетероскедастичности для оценивания $k\times 1$-мерного вектора параметров β функция правдоподобия $L(X/\sigma_i)$ будет иметь $n+k$ переменных, $\sigma_1,\ldots,\sigma_n,\beta$ (n–число наблюдений), и оценивание становится невозможным без дополнительных условий. В случае гетероскедастичности известной параметрической формы её можно ввести в функцию правдоподобия [4]. Например, для случая выбора $\sigma_i=\sigma g(X_i)$ получим

$$P(y_i=1)=F\left(\frac{f(X_i)}{\sigma g(X_i)}\right). \qquad (3)$$

Присутствие гетероскедастичности приводит к несостоятельности оценки для β. Это хорошо видно на частном примере, когда $f(X)=X\beta$, $g(X)=X\beta/\gamma$. В этом случае получаем

$$P(y_i=1)=F\left(\frac{X_i\gamma}{\sigma}\right),$$

и понятно, что оценка будет несостоятельной для β (хотя состоятельной для параметра γ).

Подход (3) оправдан тем, что для разрешения проблемы важно идентифицировать эффект гетероскедастичности на вероятность, не заостряя внимание на том, чем он вызван. То есть вовсе не важно, осуществляется ли эффект X -в через регрессионную функцию f или «скедастичную» функцию g, конечно, если функция f не является объектом исследования. Однако не следует думать, что сведя всё к отношению $f(X)$ к $g(X)$, можно заключить, что ошибочная спецификация становится невозможной. Следует помнить о непредсказуемости, к которой может привести нелинейность модели.

Ошибка спецификации в probit и logit моделях. Вместо детального обсуждения рассмотрим эскиз некоторых возможных проблем. Трудность probit (logit) моделей состоит в том, что любая ошибка спецификации функции правдоподобия приведёт к несостоятельной оценке. Если же вспомнить о методе псевдомаксимума правдоподобия (ПМП-метод) [5, 6] и вычислении оценок стандартных ошибок, то можно воспользоваться следующим подходом. Используем то, что во многих прикладных случаях probit, logit и ЛВ модели дают сходные результаты. Один из путей объяснить такое сходство различных оценок состоит в том, чтобы отнести это за счёт «некорректной» спецификации и рассматривать каждую из них как ПМП-оценку истинной модели. Затем следует воспользоваться видом асимптотической дисперсионной матрицы для МП- и ПМП-оценок и видом оценок стандартных ошибок, полученных из них. Сходство или различие оценок стандартных ошибок, вычисленных для этих двух видов оценивания, покажет, отсутствует или имеет место ошибка спецификации для функции правдоподобия [7]. Обнаруженную ошибку спецификации можно ликвидировать путём добавления к независимым (объясняющим) переменным их значения более высоких степеней или специфицировать зависимую переменную кусочно-линейной функцией, чтобы сделать спецификацию более гибкой.

Выбор модели, наиболее правильной для применения. Рассмотрим модель вида $P\{y_i = 1\} = F(f(X)/\sigma)$. Предположив, что $f(X) = X\beta$, переходим к выбору вида функции F. Понятно, что выбор вида этих двух функций относится к проблемам одинакового порядка сложности. Что касается выбора F, то он отражает степень неосведомлённости исследователя о том, как независимые переменные входят в регрессионную функцию.

Однако ЛВ, probit и logit модели во многих практических ситуациях дают подобные ответы, поэтому, возможно, наилучшая стратегия заключается в том, чтобы в каждом отдельном случае делать наиболее подходящий выбор вида F.

Одним хорошим правилом, на котором следует остановиться, является сравнение попарно производных левых частей выбранных моделей относительно X_k для среднего значения X выборки. Обычно оценки производных должны быть приблизительно сходными.

Учитывая качественное подобие результатов рассматриваемых моделей, ЛВ-модель полезно использовать из-за её простоты в тех случаях, когда речь идёт об установленном результате исследования. ЛВ-модель является также наиболее подходящей, если эндогенные переменные присутствуют в правой части модели и требуется применение инструментальных переменных. Но ЛВ-модель не является совершенной, поэтому, если можно избежать её использования, то следует обращаться к probit или logit модели и оценивать степень зависимости их результатов от частной спецификации ЛВ-модели.

Расширение модели: сгруппированные данные. Часто встречаемым расширением является их применение к сгруппированным данным. То есть вместо индивидуального данного единица наблюдения представляет целый класс индивидов, находящихся в одном и том же состоянии.

Предположим, что сгруппированные данные включают J классов, и переменная X_i представляет i-й класс. Пусть y_i – бинарная переменная, тогда модель можно записать в виде $P\{y_i = 1\} = F(X_i\beta)$ с log-функцией правдоподобия:

$$l = \sum_{i=1}^{N} \{y_i \ln F(X_i\beta) + (1 - y_i)\ln(1 - F(X_i\beta))\},$$

которую можно переписать:

$$l = \sum_{i \in J} \{p_i \ln F(X_i\beta) + (1 - p_i)\ln(1 - F(X_i\beta))\} n_i, \qquad p_i = \frac{1}{n_i}\sum_i y_i,$$

где p_i – пропорция единиц, а n_i – количество наблюдений в i-м классе. Выбрав вид распределения F, продолжаем исследование подобно вышеприведенному.

Для сгруппированных данных можно рассматривать вполне насыщенные модели, если для каждого класса ввести свой параметр δ_j, $j = 1, \ldots, J$, не заостряя внимания на специфике влияния независимых переменных на зависимые. Тогда log-функция правдоподобия примет вид

$$l = \sum_{i \in J} \{p_i \ln \delta_i + (1 - p_i)\ln(1 - \delta_i)\} n_i,$$

а МП-оценивателем этой модели будет $\hat{\delta}_i = p_i$. Таким образом, получим наилучшую модель, которая возможна в терминах максимизации функции правдоподобия рассматриваемой модели.

Заключение. В работе изложены формулирование и особенности моделей, с помощью которых можно объяснить явления, когда дискретные переменные используются для описания как наблюдаемых величин, так и результатов исследования.

З.В. Некрилова

ПРО ОСОБЛИВОСТІ МОДЕЛЕЙ З ДИСКРЕТНИМИ ЗАЛЕЖНИМИ ЗМІННИМИ

Вивчаються моделі з дискретними залежними змінними, зокрема, probit, logit і лінійні моделі ймовірності. Описано формулювання цих моделей, їх помилкова специфікація, особливості вибору типу моделі та деякі можливості їх розширення.

Z.V. Nekrylova

ABOUT FEATURES OF THE DISCRETE DEPENDENT VARIABLE MODELS

Discrete dependent variable models, in particulare, probit, logit and the linear probability model are investigated. Formulating these models, their misspecification, feature of the model type use and their some extensions are discribed.

1. *Amemiya T.* Advanced Econometrics.– Harvard: Harvard University Press, 1985. – 340 p.
2. *Johnston J., DiNardo J*. Econometric Methods.– The McGraw-Hill Companies, Inc., 1997. – 531 p.
3. *McFadden D*. Econometric Analysis of Qualitative Choice Models. Chapter 24: Handbook of Econometrics / Eds. Z. Griliches, M.D. Intriligator. – North-Holland, 1984. – 562 p.
4. *Greene W. H.* Econometric Analysis: 2nd. ed. Macmillan, 1993.– 601 p.
5. *White H.* Maximum Likelihood Estimation of Misspecified Models // Econometrica. – 1982. – **50**. – P. 1–16.
6. *Gourierox C., Monfort A.* Statistics and Econometric Models. – Cambridge: Cambridge University Press, 1989. – **1**. – 491 p.
7. *Davidson R.*, *Mackinnon I.J.* Estimation and Inference in Econometrics. – Oxford: Oxford University Press, 1993. – 603 p.

Получено 27.03.2008