Теорія оптимальних рішень

УДК 519.21

З.В. НЕКРЫЛОВА, Г.А. ШУЛИНОК

ОБ ОСОБЕННОСТЯХ МОДЕЛЕЙ PANEL-ДАННЫХ

Приведены описание и две разновидности простейшей модели panel-данных. Для каждого из видов предлагаются способы оценивания и отмечаются свойства полученных оценивателей.

Введение. Обсуждаются различные подходы для оценивания моделей panel-данных. Под последними понимаются повторяемые наблюдения над одинаковыми множествами cross-section-единиц. Такими величинами называются элементы выборки n элементов из множества N объектов, состояние которых систематически документируется.

Источником panel-данных может быть, например, изучение динамики дохода населения. Для этого с постоянной периодичностью собирается информация с n хозяйств об изменениях их экономических, социальных, демографических и прочих характеристик. Другой тип множества panel-данных можно составить из повторяемых наблюдений над определенными показателями целого региона для изучения их взаимовлияния. Еще иной тип panel-данных использует предварительные группировки cross-section-данных в относительно однородные группы, например, по возрасту, полу, образованию. Если такой процесс повторяется для других временных периодов, то такие группы можно трактовать как группы постоянного действия, хотя и искусственно созданные.

Наипростейшая модель. Введем сначала обозначения:

$y_{it}, i = 1,\ldots,n, t = 1,\ldots,T$ – значение зависимой переменной для cross-section-единицы i в момент t;

$X_{it}^{j}, i = 1,\ldots,n, t = 1,\ldots,T$ – значение j-й объясняющей переменной для единицы i в

момент t; $j = 1,\ldots,K$.

Ограничимся оцениванием уравновешенных списков данных (balanced panels). То есть имеет место одинаковое число наблюдений над любой единицей, так что общее число наблюдений будет nT. Если $n = 1$, а T – большое, получаем временной ряд. Также, когда $T = 1$, n – большое, получаем cross-section-данные. Методы оценивания panel-данных относятся к случаям, когда $n > 1, T > 1$. Далее предполагаем, что n значительно больше T, т. е. возможно даже, что $n \to \infty$, а T остается фиксированным.

Наиболее общий способ представления данных осуществляется с помощью единиц, описывающих модель

$$y_i = \begin{pmatrix} y_{i1} \\ \vdots \\ y_{iT} \end{pmatrix}, \quad X_i = \begin{pmatrix} X_{i1}^1 & \ldots & X_{i1}^K \\ \vdots & \ldots & \vdots \\ X_{iT}^1 & \ldots & X_{iT}^K \end{pmatrix}, \quad \varepsilon_i = \begin{pmatrix} \varepsilon_{i1} \\ \vdots \\ \varepsilon_{iT} \end{pmatrix}, \tag{1}$$

где ε_{it} – возмущение i-й единицы в момент t. Часто данные записывают в виде, который известен как pool-данные

$$y' = (y_1', \ldots, y_n'), \quad X' = (X_1', \ldots, X_n'), \quad \varepsilon' = (\varepsilon_1', \ldots, \varepsilon_n'), \tag{2}$$

где $y - nT \times 1$-вектор, $X - nT \times K$-матрица, $\varepsilon - nT \times 1$-вектор. Стандартную линейную модель запишем в виде

$$y = X\beta + \varepsilon, \tag{3}$$

где $\beta' = (\beta_1, \ldots, \beta_K)$, т. е. коэффициенты не изменяются ни по времени, ни по индивидам. Лаги для зависимой переменной отсутствуют.

Для классической линейной модели (3) предполагается, что ε_{it} – независимые и одинаково распределенные, т. е. для каждого индивида наблюдения серийно некоррелируемые и гомоскедастичные по всем индивидам и моментам времени. Такие предположения существенно игнорируют структуру panel-данных, поэтому оценивание (3) методом обыкновенных наименьших квадратов (оНК) не даст результатов, ожидаемых исследователем.

Два развития наипростейшей модели. Считаем, что число индивидов большое, а величина временного периода мала и запишем модель в виде

$$y_{it} = X_{it}\beta + \varepsilon_{it}, \tag{4}$$

а возмущение определим, как

$$\varepsilon_{it} = \alpha_i + \eta_{it}, \tag{5}$$

и η_{it} не коррелирует с X_{it}. Такое представление свидетельствует о том, что наше неведение имеет две составляющие. Первая часть, α_i, изменяется в зависимости от индивида, но постоянна во времени. Ее называют индивидуальным воздействием. Вторая часть, η_{it}, изменяется несистематически, т. е. независимо ни от времени, ни от индивидов. Представление (5) – наипростейший способ

сформулировать то, что два наблюдения над одним и тем же индивидом будут более «подобны» друг другу, чем наблюдения над двумя разными индивидами.

При практическом применении модели (4)–(5) имеет место одно из двух предположений об индивидуальном воздействии, что влечет ее название :

– модель случайных воздействий (СВ-модель) (random effects model): α_i не коррелирует с X_{it} ;

– модель фиксированных воздействий (ФВ-модель) (fixed effects model): α_i коррелирует с X_{it} .

Данная терминология не единообразна в различных источниках, но она наиболее устоявшаяся среди исследователей [1, 2].

Модель случайных воздействий. Она имеет вид (4)–(5) и пусть

$$E\eta = 0,\ E\eta\eta' = \sigma_\eta^2 \mathrm{I}_{nT},\ E\alpha_i = 0,\ E\alpha_i\alpha_j = 0, i \neq j,\ E\alpha_i^2 = \sigma_\alpha^2,\ E\alpha_i\eta_{it} = 0, \qquad (6)$$

где E берется условно относительно X. Тогда $E\varepsilon_i\varepsilon_i' = \sigma_\eta^2 \mathrm{I}_T + \sigma_\alpha^2 ii'$, где $i - T \times 1$-вектор из единиц. Для данных, представленных в виде (2) корреляционная матрица, $E\varepsilon\varepsilon'$, будет иметь блочно-диагональный вид с $T \times T$ -матрицами $E\varepsilon_i\varepsilon_i'$ на диагонали.

Предположение о некоррелируемости α_i и X_{it} вместе с предположениями о η_{it} являются достаточными для асимптотической несмещенности оНК-оценивателя β [3]. Он будет состоятельным, но с заниженными стандартными ошибками. Кроме того, он будет неэффективным по сравнению с оценивателем обобщенного НК -метода (ОНК). Это является следствием того, что при оценивании не используется информация о гетероскедастичности из-за повторяющихся наблюдений одних и тех же cross-section-единиц. Таким образом, проблема оНК-оценивателя на pool-данных в том, что он взвешивает все наблюдения одинаково, не учитывая то, что маловероятно, чтобы повторное наблюдение над индивидом добавило бы столько же информации, сколько дополнительное наблюдение над новым индивидом. В сущности, СВ-модель иллюстрирует то, что T наблюдений над n индивидами не является тем же самым, что наблюдения над nT различными индивидами модели.

Для нахождения ОНК-оценивателя сначала надо найти оцениватель корреляционной матрицы возмущения ε, затем использовать его для получения оценивателя β. Для этого потребуются оценки для σ_η^2 и σ_α^2. Опишем способ их выведения с помощью специально построенных оНК-оценивателей β.

Случайные воздействия и специальные оцениватели. Рассмотрим два оценивателя, являющиеся состоятельными, но по сравнению с ОНК-оценивателем неэффективными. Вполне целесообразно вместо индивидуальных наблюдений рассмотреть их средние по времени, тогда (4) примет вид

$$\overline{y}_{i\cdot} = \overline{X}_{i\cdot}\beta + \text{возмущение}, \quad \overline{y}_{i\cdot} = \tfrac{1}{T}\sum_1^T y_{it}, \quad \overline{X}_{i\cdot} = \tfrac{1}{T}\sum_1^T X_{it}. \tag{7}$$

Такое преобразование исходных данных (1)– (2) можно получить, если ввести $nT \times n$-матрицу D. Каждый столбец этой матрицы состоит из T единиц и $nT - T$ нулей, т. е. является dummy-переменной, соответствующей каждой из cross-section-единице y_i. Введение симметричной и идемпотентной матрицы $P_D = D(D'D)^{-1}D'$ и предварительное умножение на нее трансформирует исходные данные в данные уравнения (7).

оНК-оцениватель этого уравнения,

$$\hat{\beta}_M = (X'P_D X)^{-1} X'P_D y, \tag{8}$$

называют между-оценивателем (between estimator). Он будет состоятельным, если оНК-оцениватель на pool-данных будет таким, однако неэффективным. В некоторых работах его называют оценивателем Вальда, так как при достаточно большом T этот оцениватель устойчив к классическим ошибкам измерения X (при условии, что некоррелируемость сохраняется для правильно измеренных данных) [4]. Самой простой для понимания интерпретацией (8) является то, что этот оцениватель соответствует двухшаговому методу НК, который использует dummy-переменные в качестве инструменталных переменных.

Для использования информации, упущенной при получении $\hat{\beta}_M$, определим матрицу $M_D = \mathrm{I}_{nT} - D(D'D)^{-1}D'$, которая также симметричная и идемпотентная. Если предварительно умножить все данные на M_D и применить оНК-метод к полученному преобразованному уравнению

$$y_{it} - \overline{y}_{i\cdot} = \left(X_{it} - \overline{X}_{i\cdot}\right)\beta + \text{возмущение}, \tag{9}$$

можно вывести внутри-оцениватель (within-estimator):

$$\hat{\beta}_B = (X'M_D X)^{-1} X'M_D y. \tag{10}$$

При тех же условиях этот оцениватель тоже состоятельный, но неэффективный (из-за включения в (9) n необязательных переменных). Название внутри-оцениватель – следствие того, что при оценивании (9) используется только вариация внутри каждой cross-section-единицы.

Отметим, что оНК-оцениватель pool-переменных это взвешенная сумма оценивателей (8) и (10), а именно:

$$\begin{aligned}\hat{\beta} = \left(X'X\right)^{-1}X'y = \left(X'X\right)^{-1}\left(X'M_D + X'P_D\right)y = \left(X'X\right)^{-1}X'M_D X\,\hat{\beta}_B + \\ + \left(X'X\right)^{-1}X'P_D X\,\hat{\beta}_M.\end{aligned} \tag{11}$$

С помощью (8), (10) можно найти следующие оценивателя:

$$\hat{\sigma}_\eta^2 = \frac{1}{nT - nk - n}\hat{u}_B'\hat{u}_B, \quad \hat{\sigma}_M^2 = \frac{1}{n-k}\hat{u}_M'\hat{u}_M, \quad \hat{\sigma}_\alpha^2 = \hat{\sigma}_M^2 - \frac{1}{T}\hat{\sigma}_\eta^2, \tag{12}$$

где $\hat{u}_B$ – остатки от вычисления регрессии (9), $\hat{u}_M$ – остатки регрессии (7), а $\hat{\sigma}_M^2$ – оценка дисперсии возмущения регрессии (7). Полученные оценивателя

будут асимптотически несмещенными [3]. Отметим, что при выведении $\hat{\sigma}_\eta$ учитывается то, что преобразование M_D является преобразованием отклонения от среднего, которое в возмущении оставляет только η. С помощью (12) можно получить оцениватель корреляционной матрицы ε и выполнить ОНК-оценивание.

Описанное преобразование является интуитивно привлекательным. А в случае отсутствия некоррелируемой компоненты индивидуальной спецификации ($\sigma_\alpha^2 = 0$) оцениватель СВ-модели (СВ-оцениватель) сводится к оНК-оценивателю pool-данных.

Модель фиксированных эффектов в двухпериодном случае. Пока что может показаться, что panel-данные могут появляться тогда, когда нет ничего заслуживающего особого внимания по сравнению с простыми cross-section-данными, что можно было бы предложить для оценивания. Действительно, до этого места panel-данные представлены как более сложная версия cross-section-данных в случае, если нет достаточной информации от n индивидов, наблюдаемых T раз по сравнению с nT наблюдениями над индивидами.

Наблюдается даже, что некоторые исследователи превратно отмечают, что преимущество panel-данных в том, что «они добавили работы эконометристам». Однако популярность методов, используемых при оценивании panel-данных, растет. И вызвано это тем, что появляется надежда с их помощью разрешить серьезную проблему, стоящую перед исследователями: отсутствие подхода для получения достаточного перечня независимых переменных для объяснения зависимой переменной. Чтобы проиллюстрировать это, начнем с оценивателя ФВ-модели (ФВ-оценивателя) одного фиксированного воздействия.

Рассмотрим простую двухпериодную модель ($t = 1,2$) вида

$$y_{it} = X_{it}\beta + Z_i\delta + \varepsilon_{it}, \qquad (13)$$

где X – матрица объясняющих переменных, меняющаяся по индивидам и времени; Z – матрица наблюдаемых переменных, меняющаяся по индивидам, но постоянная во времени. Введем снова (5), (6), где все E берутся условно относительно X и Z, обозначим $W_{it} = [X_{it} Z_i]$ и предположим, что

$$\mathrm{E}W_{it}'\varepsilon_{it} \neq 0, \qquad (14)$$

т. е. независимые переменные коррелируют с α. Из работы [3] известно, что это влечет смещенность оНК-оценивателя, которая зависит от точной природы связи между индивидуальным воздействием и другими объясняющими переменными.

Основная привлекательность panel-данных состоит в том, что если (13) справедливо для каждого $t = 1,2$, то любая линейная комбинация из таких соотношений множества данных будет также справедлива. В частности, можно записать, что имеет место

$$y_{i2} - y_{i1} = (X_{i2} - X_{i1})\beta + (Z_i - Z_i)\delta + (\varepsilon_{i2} - \varepsilon_{i1}) \text{ или } \Delta y = \Delta X\beta + \Delta Z\delta + \Delta\varepsilon, \qquad (15)$$

где Δ – оператор разности. Уравнение (15) эквивалентно следующему:

$$\Delta y = \Delta X\beta + \Delta\eta \tag{16}$$

и главное, что

$$\mathrm{E}[\Delta X \Delta\eta] = 0. \tag{17}$$

Поэтому оНК-оцениватель β будет уже несмещенным на преобразованных данных. Это является сущностью ФВ-модели. Из данного простого примера следуют еще три вывода, которые нагляднее в более общем случае ФВ-модели.

1. С помощью ФВ-оценивателя вообще-то нельзя восстановить оценки каких-либо постоянных во времени объясняющих переменных. (Если есть дополнительная априорная информация об элементах регрессоров, меняющихся во времени, то иногда возможно восстановить коэффициенты при регрессорах, постоянных во времени [5, 6]). Все воздействия, постоянные во времени, опускаются при разностном преобразовании данных.

2. Обратной стороной вывода 1 является то, что ФВ-оцениватель устойчив к тому, что опускаются уместные регрессоры, постоянные во времени. Это можно считать перспективной особенностью ФВ-оценивания. При таком оценивании значительно минимизируется информационное требование, необходимое для условия некоррелируемости. На приведенном примере это было видно.

3. Если СВ-модель является действительной, то ФВ-оцениватель будет еще состоятельно оценивать идентифицируемые параметры. То есть условие некоррелируемости (17) очевидно еще имеет место, если множество данных описывается СВ-моделью (хотя ФВ-оцениватель по сравнению с СВ-оценивателем не является эффективным).

Модель фиксированных воздействий в многопериодном случае. Напомним, что основным предположением для ФВ-модели является условие $\mathrm{cov}(X_{it}\alpha_i) \neq 0$, поэтому оценивать модель надо условно относительно присутствующих фиксированных воздействий, т. е. если модель переписать в виде

$$y_{it} = X_{it}\beta + \alpha_i + \eta_{it}, \tag{18}$$

то α_i трактуются как неизвестные параметры, подлежащие оцениванию. Однако в случае типичных panel-данных эти дополнительные параметры нельзя оценить состоятельно. Следует это из того, что для типичного случая T является малым, а n – большое, а для асимптотики n должно становиться все больше и больше. В рассматривамой постановке количество параметров возрастает с той же скоростью, как и объем выборки. Что касается остальных параметров (18), то их можно оценить состоятельно.

Действительно, перепишем (18) в матричном виде

$$y = X\beta + D\alpha + \eta, \tag{19}$$

где D ранее введенная матрица dummy-переменных. Из теоремы Frisch-Waugh-Lovell [7] известно, что оценивание (19) точно такое же, как оценивание регрессии y на остатки в X или оценивание остатков в y на остатки в X, когда ли-

нейное воздействие всех dummy-переменных удалено. Получить такие остатки можно с помощью уже знакомой матрицы $M_D = I - D(D'D)^{-1}D'$, а оценивание регрессии $M_D y$ на $M_D X$ приводит к ранее выведенному внутри-оценивателю, (10):

$$\hat{\beta}_B = (X'M_D X)^{-1} X'M_D y = \left((M_D X)'(M_D X)\right)^{-1} (M_D X)'M_D y, \qquad (20)$$

где последнее равенство следует из идемпотентности матрицы M_D. Этот оцениватель можно записать и в виде

$$\hat{\beta}_B = \left((M_D X)'(M_D X)\right)^{-1} (M_D X)'y. \qquad (21)$$

Эти две записи $\hat{\beta}_B$ иллюстрируют утверждение теоремы Frisch-Waugh-Lovell, так как оцениватель (20) является следствием выполнения регрессии $M_D y$ на $M_D X$, а (21) – регрессии y на $M_D X$.

Внутри-оцениватель это только один возможный оцениватель ФВ-модели. Любое преобразование, удаляющее элемент фиксированного воздействия из уравнения, будет приводить к оцениванию ФВ-модели. Например, $T \times (T-1)$ – матрица и последующее умножение на нее

$$F = \begin{bmatrix} -1 & 0 & 0 & \dots & 0 \\ 1 & -1 & 0 & \dots & 0 \\ 0 & 1 & -1 & \dots & 0 \\ 0 & 0 & 0 & \dots & -1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix} \qquad (22)$$

трансформирует $1 \times T$ -вектор из повторяемых наблюдений над одним и тем же индивидом в $1 \times (T-1)$ -вектор их первых разностей. Оно точно такое же, как приведенное в предыдущем разделе, поэтому удаляет элемент фиксированного воздействия из уравнения.

Можно использовать подход отклонения от среднего. В терминах средних уравнение (18) примет вид:

$$\bar{y}_{i\cdot} = \bar{X}_{i\cdot}\beta + \bar{\alpha}_i + \bar{\eta}_{i\cdot}. \qquad (23)$$

Так как среднее $\bar{\alpha}_i$ для индивида i есть просто α_i, то вычитание (23) из (18) приведет к

$$y_{it} - \bar{y}_{i\cdot} = \left(X_{it} - \bar{X}_{i\cdot}\right)\beta + \left(\eta_{it} - \bar{\eta}_{i\cdot}\right), \qquad (24)$$

и преобразование (24) удаляет элемент фиксированного воздействия из уравнения (18). оНК-оцениватели, полученные из переменных, преобразованных с помощью (22) или (24), не будут численно идентичными. Но при этом очень раз-

нящиеся ответы будут свидетельствовать о том, что предположение о фиксированных воздействиях не имеет места.

Во многих применениях самым легким путем оценивания ФВ-модели считается использование dummy-переменных, как в (19). Этот метод часто называют методом наименьших квадратов dummy-переменных (least squares dummy variables). Если n очень большое и вычислительно трудно подсчитать коэффициент для каждой cross-section-единицы, тогда прибегают к методу отклонения от среднего:

– для их получения используется матрица M_D;

– применяется оНК-метод на преобразованных переменных.

Этот оцениватель часто называют оценивателем внутри группы (within group estimator), т. е. таким, который использует только вариацию наблюдений внутри множества индивидов.

Заключение. Приведенные результаты дают представление о простейшей модели panel-данных, о способах задания возмущения модели и особенностях оценивания в зависимости от его структуры.

Предполагается введение и описание тестов для проверки качества оценок.

З.В. Некрилова, Г.О. Шулінок

ПРО ОСОБЛИВОСТІ МОДЕЛЕЙ PANEL-ДАНИХ

Наведено описання і два різновиди найпростішої моделі panel-даних. Для кожного з видів надаються способи оцінювання і зазначаються властивості оцінювачів.

Z.V. Nekrylova, G.A. Shulinok

ABOUT FEATURES OF THE PANEL DATE MODELS

The simplest panel date model and two its extensions are described. The astimate methods and the astimator properties are considered for them.

1. *Searle S., Cassella G., McCullouch C.* Variance Components. – New-York: Jonh Wiley-Sons. INC, 1992. – 348 p.
2. *Johnston J., DiNardo J.* Econometrica Methods. – New-York: The McGraw-Hill Companies. INC, 1997. – 531 p.
3. *Джонстон Дж.* Эконометриеские методы. – М.: Статистика, 1980. – 444 с.
4. *Wald A.* The Fitting of Straight Lines if Both Variables are Subject to Error // Annals of Mathematical Statistics. – 1940. – **2**. – P. 284–300.
5. *Hausman J., Taylor W.* Panel Data and Unobservable Individual Effects // Econometrica. – 1981. – **49**. – P. 1377–1389.
6. *Hsiao C.* Analysis of Panel Data. – Cambridge University Press, 1986. – Secnion 3.6.1.
7. *Davidson R., MacKinnon J.* Estimation and Inference in Econometrics. – Oxford University Press, 1993. – 520 p.