ЛЕКЦИОННЫЕ И МЕТОДИЧЕСКИЕ МАТЕРИАЛЫ

Введение в эконометрический анализ панельных данных

Ратникова Т.А.

В предыдущих номерах журнала были опубликованы восемь лекций из курса «Введение в эконометрический анализ панельных данных», где была изложена информация общего порядка о панельных данных, рассмотрены методы оценивания основных моделей, свойства полученных оценок, тесты на спецификацию. Также обсуждались проблемы оценивания регрессионных моделей панельных данных в условиях гетероскедастичности и автокоррелированности случайных ошибок и в условиях эндогенности, которая имеет место при коррелированности регрессоров с индивидуальными эффектами, при наличии ошибок измерения объясняющих переменных и при построении динамических моделей.

В этом выпуске вашему вниманию предлагаются две последние лекции, в первой из которых речь пойдет об оценивании моделей с дискретными и ограниченными зависимыми переменными, а в последней будут обсуждаться меры, позволяющие предотвращать или уменьшать последствия истощения выборки.

Лекция 9.

9. Модели с дискретными и ограниченными зависимыми переменными

Панельные данные часто используются для оценивания нелинейных моделей. Модели с дискретными или ограниченными зависимыми переменными – распространенное явление в этой области.

Помимо обычных вычислительных трудностей, связанных с оцениванием таких моделей, использование панельных данных порождает еще дополнительные проблемы. Дело в том, что модели дискретного или ограниченного выбора были разработаны первоначально для данных пространственного типа, где естественным ограничением является независимость наблюдений. В панельных данных же не предполагается, что наблюдения, относящиеся к одному и тому же индивидууму в разные моменты времени, должны быть независимы. Такое требование противоречило бы реальности. Наличие корреляций между различными компонен-

---

Ратникова Т.А. – к.ф.-м.н., доцент кафедры математической экономики и эконометрики ГУ ВШЭ.

тами случайной ошибки существенно усложняет вид функции правдоподобия и численные алгоритмы поиска ее максимума.

В этой части мы рассмотрим приемы оценивания логит, пробит и тобит-моделей.

9.1. Модели бинарного выбора

Как и в случае пространственных данных, модели бинарного выбора для панелей обычно формулируются в терминах латентной зависимой переменной

$$y_{it}^{*} = X_{it}'\beta + \alpha_i + \varepsilon_{it},$$

где реально наблюдаемая зависимая переменная

$$y_{it} = \begin{cases} 1 & \text{если} \quad y_{it}^{*} > 0 \\ 0 & \text{иначе.} \end{cases}$$

Например, y_{it} может означать, менял ли место работы i-й индивидуум в период времени t.

Предположим, что случайная ошибка ε_{it} имеет симметричное распределение с функцией распределения $F(\varepsilon)$, независимое и одинаковое по i и по t и независимое от X_{it}.

В отсутствие индивидуального эффекта α_i оценки коэффициентов получаются либо с помощью сквозной пробит-регрессии, либо с помощью сквозной логит-регрессии.

Присутствие слагаемого α_i существенно усложняет оценивание, причем неважно, рассматриваем ли мы его как ненаблюдаемый детерминированный индивидуальный эффект или как компоненту случайного возмущения.

9.1.1. Оценивание моделей с детерминированным индивидуальным эффектом

Если α_i трактуются как неизвестные детерминированные параметры, то этим самым в модель включаются N дамми-переменных. Тогда логарифм функции правдоподобия будет задаваться следующим выражением:

$$\ln L(\beta, \alpha_1, \ldots, \alpha_N) = \sum_{i,t} y_{it} \ln F(\alpha_i + X_{it}'\beta) + \sum_{i,t} (1 - y_{it}) \ln\left[1 - F(\alpha_i + X_{it}'\beta)\right].$$

Максимизация этого выражения по β и α_i (i = 1,...,N) приводит к состоятельным оценкам при условии, что число временных периодов T стремится к бесконечности. Для конечных значений T и $N \to \infty$ оценки будут несостоятельны. Причина кроется в том, что при конечных T число параметров растет с размером N, и происходит то, что называют «incidental parameter's»-проблемой, что можно перевести как проблему случайных параметров. Это означает следующее:

любое α_i может быть оценено состоятельно, только если мы имеем достаточно большое число наблюдений для каждого i-го объекта, т.е. когда $T \to \infty$. Если же число таких наблюдений мало, оценка α_i будет несостоятельна. В общем случае несостоятельность оценок α_i для фиксированных T будет перенесена и на оценки β.

Эта проблема, когда число параметров увеличивается с числом наблюдений, встречается в любой FE-модели (так для краткости в предыдущих лекциях было условлено называть модель с детерминированным эффектом), и линейной, и нелинейной. Но если в линейном случае у нас есть возможность исключить α_i из уравнения, так что β могут быть оценены состоятельно даже тогда, когда для всех α_i этого сделать нельзя, то для большинства нелинейных моделей такое исключение произвести невозможно, и несостоятельность оценок α_i влечет за собой несостоятельность всех остальных оценок регрессий. Также заметим, что с практической точки зрения оценивание более чем N параметров при $N \to \infty$ не слишком привлекательно.

Конечно, возможно преобразование модели с латентной зависимой переменной, элиминирующее индивидуальные эффекты, но в данном контексте это бесполезно, так как нельзя полагать, что разность $y_{it} - y_{it-1}$ является наблюдаемым аналогом разности $y^*_{it} - y^*_{it-1}$.

Поясним суть вышесказанного следующим примером.

Рассмотрим панель с $T = 2$. Пусть в регрессии присутствует только одна независимая переменная:

$$X_{it} = \begin{cases} 0 & \text{при} \quad t = 1 \\ 1 & \text{при} \quad t = 2. \end{cases}$$

В таком случае

$$X'_{it}\beta + \alpha_i = \begin{cases} \alpha_i & \text{при} \quad t = 1 \\ \beta + \alpha_i & \text{при} \quad t = 2. \end{cases}$$

Предположим, что исходы независимы и ошибка имеет логистическую функцию распределения

$$\varepsilon_{it} \sim F_L = \frac{1}{1 + \exp\left(- X'_{it}\beta - \alpha_i\right)}.$$

Тогда логарифмическая функция правдоподобия будет записываться следующим образом:

$$\ln L\left(\beta, \alpha_1, \ldots, \alpha_N\right) = \sum_{i=1}^{N} \ln L_i = \sum_{i=1}^{N} \left[\ln L_{i1} + \ln L_{i2}\right] =$$

$$= \sum_{i=1}^{N} \left[y_{i1} \ln \frac{\exp(\alpha_i)}{1+\exp(\alpha_i)} + (1-y_{i1}) \ln \frac{1}{1+\exp(\alpha_i)} + y_{i2} \ln \frac{\exp(\alpha_i+\beta)}{1+\exp(\alpha_i+\beta)} + (1-y_{i2}) \ln \frac{1}{1+\exp(\alpha_i+\beta)} \right].$$

Зафиксируем β и будем оптимизировать по α_i. Существует всего 4 комбинации возможных значений зависимой переменной:

$$y_{i1} = y_{i2} = 1;\ y_{i1} = y_{i2} = 0;\ y_{i1} = 0,\quad y_{i2} = 1;\ y_{i1} = 1,\quad y_{i2} = 0.$$

Для случая $y_{i1} = y_{i2} = 1$

$$\ln L_i = \ln\frac{\exp(\alpha_i)}{1+\exp(\alpha_i)} + \ln\frac{\exp(\alpha_i+\beta)}{1+\exp(\alpha_i+\beta)} = \alpha_i - \ln(1+\exp(\alpha_i)) + \alpha_i + \beta - \ln(1+\exp(\alpha_i+\beta));$$

и из условия первого порядка

$$\frac{\partial \ln L_i}{\partial \alpha_i} = 2 - \frac{\exp(\alpha_i)}{1+\exp(\alpha_i)} - \frac{\exp(\alpha_i+\beta)}{1+\exp(\alpha_i+\beta)} = 0$$

следует, что $\hat{\alpha}_i = \infty$, а $\ln L_i(\hat{\alpha}_i) = 0$.

Для случая $y_{i1} = y_{i2} = 0$ аналогичным образом получается, что $\hat{\alpha}_i = -\infty$, и $\ln L_i(\hat{\alpha}_i) = 0$.

Эти две ситуации оказываются неинформативны. В оставшихся двух ситуациях аналогичные выкладки приводят к результатам:

$$\hat{\alpha}_i = \frac{1}{2}\beta \text{ и } \ln L_i(\hat{\alpha}_i) = -2\ln(1+\exp(-\beta/2)) \text{ для } y_{i1} = 0,\quad y_{i2} = 1,$$

$$\hat{\alpha}_i = -\frac{1}{2}\beta \text{ и } \ln L_i(\hat{\alpha}_i) = -2\ln(1+\exp(\beta/2)) \text{ для } y_{i1} = 1,\quad y_{i2} = 0.$$

Теперь осталось максимизировать полученную функцию правдоподобия по параметру β:

$$\ln L(\beta) = -2n_{01}\ln(1+\exp(-\beta/2)) - 2n_{10}\ln(1+\exp(\beta/2)),$$

где n_{01} – это число наблюдений, для которых $y_{i1} = 0,\quad y_{i2} = 1$;

n_{10} – число наблюдений, для которых $y_{i1} = 1,\quad y_{i2} = 0$.

В итоге

$$\hat{\beta}^{FE}_{MMП} = 2\ln\frac{n_{01}}{n_{10}}.$$

Это редкий случай аналитического выражения для оценки ММП.

Выясним состоятельность этой оценки:

$$\frac{n_{01}}{n_{10}} = \frac{n_{01}/N}{n_{10}/N} \to \frac{P\{y_{i1}=0, y_{i2}=1\}}{P\{y_{i1}=1, y_{i2}=0\}} = \frac{1/(1+\exp(\alpha_i))\cdot\exp(\alpha_i+\beta)/(1+\exp(\alpha_i+\beta))}{\exp(\alpha_i)/(1+\exp(\alpha_i))/(1+\exp(\alpha_i+\beta))} = e^{\beta}.$$

Следовательно $E\left(\hat{\beta}^{FE}_{MMП}\right) = 2\ln e^{\beta} = 2\beta$. Оценка $\hat{\beta}^{FE}_{MMП}$ несостоятельна.

Существует альтернативный подход, предложенный в 1970 г. Андерсеном и развитый Чемберленом в работе 1980 г. [5].

Чемберлен предложил следующее:

- исключить из функции правдоподобия все слагаемые, для которых $y_{i1} = y_{i2}$;
- рассматривать условный максимум функции правдоподобия при условии $y_{i1} + y_{i2} = 1$.

Что это дает? Оказывается, с помощью такого приема можно элиминировать индивидуальный эффект из нелинейных моделей определенного типа и получить состоятельные оценки для параметров β.

Покажем это на нашем примере. Условная вероятность событий, для которых $y_{i1} = 0, \quad y_{i2} = 1,$ если ввести для упрощения выкладок следующие обозначения $Z_1 = \exp(X'_{i1}\beta + \alpha_i)$ и $Z_2 = \exp(X'_{i2}\beta + \alpha_i),$ может быть подсчитана следующим образом:

$$P\left[y_{i1} = 0, y_{i2} = 1 \mid y_{i1} + y_{i2} = 1, X'_{i1}, X'_{i2}\right] =$$

$$= \frac{P[y_{i1} = 0, y_{i2} = 1 \mid X'_{i1}, X'_{i2}]}{P[y_{i1} = 0, y_{i2} = 1 \mid X'_{i1}, X'_{i2}] + P[y_{i1} = 1, y_{i2} = 0 \mid X'_{i1}, X'_{i2}]} =$$

$$= \frac{(1+Z_1)^{-1} Z_2 (1+Z_2)^{-1}}{(1+Z_1)^{-1} Z_2 (1+Z_2)^{-1} + Z_1 (1+Z_1)^{-1} (1+Z_2)^{-1}} = \frac{Z_2}{Z_1 + Z_2} =$$

$$= \frac{\exp(X'_{i2}\beta)}{\exp(X'_{i2}\beta) + \exp(X'_{i1}\beta)} = \frac{\exp\left[(X_{i2} - X_{i1})'\beta\right]}{1 + \exp\left[(X_{i2} - X_{i1})'\beta\right]} = F_L\left[(X_{i2} - X_{i1})'\beta\right],$$

и тогда

$$P\left[y_{i1} = 1, y_{i2} = 0 \mid y_{i1} + y_{i2} = 1, X'_{i1}, X'_{i2}\right] = 1 - F_L\left[(X_{i2} - X_{i1})'\beta\right].$$

В самом деле, функция правдоподобия, которую можно построить на основании вычисленных вероятностей уже не будет содержать зависимости от α_i, и оценка максимального правдоподобия для β получается, как было доказано Чемберленом, состоятельной.

Аналогичным образом оцениваются модели для $T > 2$.

Следует подчеркнуть, что такой подход работает только для логит-моделей. Для пробит-моделей элиминировать индивидуальный эффект таким образом не удается.

Чтобы выяснить, какая спецификация регрессионной модели наиболее адекватна данным, с детерминированным индивидуальным эффектом α_i (FE) или без него, в нелинейных моделях используется тест Хаусмана. В нем проверяется основная гипотеза:

$$H_0 : \alpha_i = \alpha = const \text{ для всех } i,$$

при альтернативной гипотезе H_A, что существуют i, для которых это равенство нарушается. Тогда оценка сквозной логит-регрессии $\hat{\beta}_{Pooled}$ будет являться состоя-

тельной и асимптотически эффективной в случае справедливости основной гипотезы и несостоятельной в случае справедливости альтернативной, а оценка логит-регрессии с условным FE $\widehat{\beta}_{FE}$ будет состоятельна в любом случае. Тестовая статистика будет иметь вид

$$m = \left(\hat{\beta}_{FE} - \hat{\beta}_{Pooled}\right)' \left(\hat{V}\left(\hat{\beta}_{FE}\right) - \hat{V}\left(\hat{\beta}_{Pooled}\right)\right)^{-1} \left(\hat{\beta}_{FE} - \hat{\beta}_{Pooled}\right)$$

и подчиняться при условии справедливости H_0 χ^2-распределению с числом степеней свободы, соответствующим числу меняющихся со временем регрессоров.

Тест можно проводить для любого интересующего нас коэффициента регрессии $\beta\left(X_j\right)$, заменив тестовую статистику m на

$$t = \frac{\hat{\beta}_{FE}\left(X_j\right) - \hat{\beta}_{Pooled}\left(X_j\right)}{\sqrt{\hat{V}\left(\hat{\beta}_{FE}\left(X_j\right)\right) - \hat{V}\left(\hat{\beta}_{Pooled}\left(X_j\right)\right)}}.$$

9.1.2. Оценивание моделей со случайным индивидуальным эффектом

Модель со случайным индивидуальным эффектом может быть оценена как с помощью логит, так и с помощью пробит-регрессий.

Мы рассмотрим, на сей раз, процедуру пробит-оценивания.

Пусть имеется модель вида

$$y_{it}^* = X_{it}'\beta + \alpha_i + \varepsilon_{it}, \text{ где } y_{it} = \begin{cases} 1 & \text{если } y_{it}^* > 0 \\ 0 & \text{иначе.} \end{cases}.$$

Будем предполагать, что $\alpha_i \sim NID\left(0, \sigma_\alpha^2\right)$, $\varepsilon_{it} \sim NID(0,1)$ и ε_{it} независимы от α_i и X_{it}. Тогда наблюдения независимы по i, и логарифм функции правдоподобия можно представить следующим образом:

$$\ln L(\beta, \alpha_1, \ldots, \alpha_N) = \sum_{i=1}^{N} \ln L_i,$$

но из-за влияния α_i наблюдения не будут независимы по t и

$$L_i = P\left[y_{i1} = 0, y_{i2} = 1, \ldots, y_{iT} = 1 \mid X_{it}'\right] \neq P\left[y_{i1} = 0 \mid X_{i1}'\right] \cdot P\left[y_{i2} = 1 \mid X_{i2}'\right] \ldots \cdot P\left[y_{iT} = 1 \mid X_{iT}'\right],$$

где $P\left[y_{it} = 1 \mid X_{it}'\right] = P\left[X_{it}'\beta + \alpha_i + \varepsilon_{it} > 0 \mid X_{it}'\right] = P\left[\alpha_i + \varepsilon_{it} > -X_{it}'\beta \mid X_{it}'\right] = \Phi\left(\dfrac{X_{it}'\beta}{\sqrt{1-\sigma_\alpha^2}}\right)$, а

$P\left[y_{it} = 0 \mid X_{it}'\right] = 1 - \Phi\left(\dfrac{X_{it}'\beta}{\sqrt{1-\sigma_\alpha^2}}\right)$.

Выход из положения состоит в том, чтобы вычислить вероятности при различных α_i, а потом усреднить результат по i, воспользовавшись тем, что

$$E(z) = E\{E(z \mid \alpha)\},$$

а вероятности событий $A = \{\alpha_i + \varepsilon_{it} > -X'_{it}\beta\}$ – это математические ожидания индикаторных функций

$$P(A) = E\{P(A \mid \alpha)\}.$$

Учитывая все эти обстоятельства, мы можем переписать функцию правдоподобия для *i*-го объекта в следующем виде

$$L_i = P[y_{i1}, y_{i2}, \ldots, y_{iT} \mid X'_{it}] = E\{P[y_{i1}, y_{i2}, \ldots, y_{iT} \mid X'_{i1}, \ldots, X'_{iT}, \alpha_i]\} = E\left\{\prod_{t=1}^{T} P[y_{it} \mid X_{it}, \alpha_i]\right\} =$$

$$= \int_{-\infty}^{\infty} f(\alpha_i) \prod_{t=1}^{T} \left\{ \left[\Phi\left(\frac{X'_{it}\beta + \alpha_i}{\sqrt{1-\sigma_\alpha^2}} \right) \right]^{y_{it}} \cdot \left[1 - \Phi\left(\frac{X'_{it}\beta + \alpha_i}{\sqrt{1-\sigma_\alpha^2}} \right) \right]^{1-y_{it}} \right\} d\alpha_i,$$

где $f(\alpha_i) = \frac{1}{\sqrt{2\pi}\sigma_\alpha} \exp\left(-\frac{1}{2} \frac{\alpha_i^2}{\sigma_\alpha^2} \right)$, а интеграл берется численно.

В стандартных эконометрических пакетах, например в STATA, вся эта процедура запрограммирована.

Следует отметить, что состоятельные оценки параметров β можно получить и обычной пробит-регрессией, игнорирующей панельную природу данных, но эти оценки будут неэффективны, а их стандартные ошибки будут оценены некорректно.

9.1.3. Пример: выявление детерминант задолженности по заработной плате

В 1990-е гг. неплатежи и задержки заработной платы стали одним из вынужденных средств адаптации российской экономики к новым рыночным условиям. В предлагаемом примере на данных РМЭЗ исследуется характер зависимости долга по заработной плате, наличие которого отражает бинарная переменная *debt*, принимающая значение единица, если респондент имел задолженность по заработной плате, и ноль – в противном случае, от индивидуальных характеристик респондента.

Оцениваемое уравнение имеет вид:

$$Prob(debt_{it} = 1) = b_0 + b_1 educ_{it} + b_2 age_{it} + b_3 age2_{it} + b_4 stagna_{it} + b_5 gen_i + \\ + b_6 marst_{it} + b_7 city_{it} + b_8 isco\_1_{it} + b_9 isco\_2_{it} + \ldots + b_{14} isco\_7_{it} + b_{15} isco\_8_{it} + e_{it}.$$

В качестве начального приближения рассмотрим обычную процедуру пробит-оценивания, игнорирующую панельный характер данных.

```
Probit estimates                          Number of obs   =      10794
                                          LR chi2(18)     =    1583.96
                                          Prob > chi2     =     0.0000
Log likelihood = -6689.5837               Pseudo R2       =     0.1059
```

```
------------------------------------------------------
        debt |      Coef.   Std. Err.      z    P>|z|
-------------+----------------------------------------
        educ |  -.0097868   .0055185    -1.77   0.076
         age |    .035835   .0071306     5.03   0.000
        age2 |   -.000475   .0000844    -5.63   0.000
      stagna |   .0849118    .008532     9.95   0.000
         gen |  -.1353812    .030859    -4.39   0.000
       marst |  -.0017841   .0153579    -0.12   0.908
        city |  -.5223073   .0282049   -18.52   0.000
      isco_1 |  -.2816055   .0875224    -3.22   0.001
      isco_2 |  -.1270906   .0522407    -2.43   0.015
      isco_3 |  -.1715198   .0496234    -3.46   0.001
      isco_4 |  -.3662031    .060064    -6.10   0.000
      isco_5 |  -.4691111   .0608385    -7.71   0.000
      isco_6 |  -.3779202     .17376    -2.17   0.030
      isco_7 |  -.0362381   .0491966    -0.74   0.461
      isco_8 |    -.06042   .0474841    -1.27   0.203
         d96 |   .5011793    .034464    14.54   0.000
         d98 |   .5779015   .0352981    16.37   0.000
         d00 |  -.3716852   .0360992   -10.30   0.000
       _cons |  -.3901286   .1507308    -2.59   0.010
------------------------------------------------------
```

Очевидно, что вероятность задолженности растет с возрастом и стажем работы на данном месте, для женщин она ниже, чем для мужчин, для горожан ниже, чем для сельских жителей. Вероятность задолженности не зависит от семейного статуса и слабо и отрицательно зависит от уровня образования. Для всех профессиональных групп она ниже, чем для групп работников низкой квалификации. В 1996 и 1998 гг. вероятность задолженности значимо выше, чем в 1994 г., зато в 2000 г. она значимо ниже.

Как было отмечено выше, приведенные оценки могут быть менее эффективны, чем оценки панельной пробит-регрессии со случайным эффектом. Посмотрим, так ли это в нашем случае.

```
Random-effects probit                   Number of obs      =     10794
Group variable (i) : aid_i              Number of groups   =      3937
Random effects u_i ~ Gaussian           Obs per group: min =         1
                                                       avg =       2.7
                                                       max =         4
                                        Wald chi2(18)      =   1230.52
Log likelihood  = -6425.7036            Prob > chi2        =    0.0000
------------------------------------------------------
        debt |      Coef.   Std. Err.      z    P>|z|
-------------+----------------------------------------
        educ |  -.0107514   .0078893    -1.36   0.173
         age |    .042901   .0102503     4.19   0.000
        age2 |  -.0005653   .0001219    -4.64   0.000
      stagna |   .1077713    .011419     9.44   0.000
       marst |  -.0102822   .0203172    -0.51   0.613
```

```
------------------------------------------------------
        debt |      Coef.   Std. Err.      z    P>|z|
-------------+----------------------------------------
        city |  -.6828093   .0442328   -15.44   0.000
      isco_1 |   -.298468   .1133807    -2.63   0.008
      isco_2 |   -.120898   .0720817    -1.68   0.093
      isco_3 |  -.2164254   .0680787    -3.18   0.001
      isco_4 |  -.3981784   .0829853    -4.80   0.000
      isco_5 |  -.4991976   .0831599    -6.00   0.000
      isco_6 |  -.4947072   .2342688    -2.11   0.035
      isco_7 |  -.0075099   .0671357    -0.11   0.911
      isco_8 |  -.0346067   .0652791    -0.53   0.596
         d96 |    .649137   .0398622    16.28   0.000
         d98 |   .7430593   .0415524    17.88   0.000
         d00 |  -.4694101   .0419477   -11.19   0.000
       _cons |  -.4906535   .2151729    -2.28   0.023
-------------+----------------------------------------
    /lnsig2u |  -.4767715   .0745498
-------------+----------------------------------------
     sigma_u |   .7878987   .0293688
         rho |   .3830148   .0176172
------------------------------------------------------
Likelihood ratio test of rho=0: chibar2(01) = 527.76 Prob >=chibar2 = 0.000
```

Стандартные ошибки оценок коэффициентов несколько возрасли и сами оценки немного изменились, но в целом выводы не меняются.

Посмотрим теперь на логит-регрессию со случайными эффектами.

```
Random-effects logit                    Number of obs      =     10794
Group variable (i) : aid_i              Number of groups   =      3937
Random effects u_i ~ Gaussian           Obs per group: min =         1
                                                       avg =       2.7
                                                       max =         4
                                        Wald chi2(18)      =   1109.16
Log likelihood  = -6425.8727            Prob > chi2        =    0.0000
------------------------------------------------------
        debt |      Coef.   Std. Err.      z    P>|z|
-------------+----------------------------------------
        educ |  -.0186126   .0134519    -1.38   0.166
         age |   .0723137   .0174779     4.14   0.000
        age2 |  -.0009547   .0002078    -4.59   0.000
      stagna |   .1842564   .0195419     9.43   0.000
         gen |  -.2935808   .0788712    -3.72   0.000
       marst |  -.0173077   .0347788    -0.50   0.619
        city |  -1.160912   .0760577   -15.26   0.000
      isco_1 |  -.5184956   .1938356    -2.67   0.007
      isco_2 |  -.2039514   .1228729    -1.66   0.097
      isco_3 |  -.3659185   .1160045    -3.15   0.002
      isco_4 |  -.6815935   .1416501    -4.81   0.000
      isco_5 |  -.8516032   .1424288    -5.98   0.000
      isco_6 |  -.8538069   .4011953    -2.13   0.033
```

```
------------------------------------------------------
        debt |      Coef.   Std. Err.      z    P>|z|
-------------+----------------------------------------
      isco_7 |   -.014164   .1145148    -0.12   0.902
      isco_8 |  -.0583105   .1115143    -0.52   0.601
         d96 |   1.098994   .0682853    16.09   0.000
         d98 |   1.262838   .0715524    17.65   0.000
         d00 |  -.8017823    .071996   -11.14   0.000
       _cons |  -.8186575   .3665523    -2.23   0.026
-------------+----------------------------------------
    /lnsig2u |   .5872119   .0777082
-------------+----------------------------------------
     sigma_u |   1.341255   .0521133
         rho |   .6427252   .0178441
------------------------------------------------------
Likelihood ratio test of rho=0: chibar2(01)=526.12 Prob>=chibar2=0.000
```

Все выводы сохраняются, хотя значения оценок коэффициентов моделей пробит и логит нельзя сравнивать непосредственно.

И в завершении оценим логит-регрессию с детерминированным эффектом.

```
Conditional fixed-effects logit          Number of obs      =      6132
Group variable (i) : aid_i               Number of groups   =      1822
                                         Obs per group: min =         2
                                                        avg =       3.4
                                                        max =         4
                                         LR chi2(18)        =   1049.65
Log likelihood  =   -1780.37             Prob > chi2        =    0.0000
------------------------------------------------------
        debt |      Coef.   Std. Err.      z    P>|z|
-------------+----------------------------------------
        educ |   .0336732   .0350424     0.96   0.337
         age |   .0840919   .1181512     0.71   0.477
        age2 |   .0003831   .0005821     0.66   0.510
      stagna |   .2007279   .0276749     7.25   0.000
         gen |   (dropped)
       marst |   -.030725   .0507237    -0.61   0.545
        city |   (dropped)
      isco_1 |  -.0677049   .2691339    -0.25   0.801
      isco_2 |   .2799555   .2177781     1.29   0.199
      isco_3 |  -.1087625   .1934809    -0.56   0.574
      isco_4 |  -.0553374   .2380427    -0.23   0.816
      isco_5 |    -.10366   .2405099    -0.43   0.666
      isco_6 |  -1.195202   .6488973    -1.84   0.065
      isco_7 |   .1852537   .1909065     0.97   0.332
      isco_8 |   .0671342   .1903543     0.35   0.724
         d96 |    .940399   .2163371     4.35   0.000
         d98 |   .8469935   .4427995     1.91   0.056
         d00 |  -1.456035   .6564571    -2.22   0.027
------------------------------------------------------
```

В этой регрессии значительно сократилось число наблюдений, поскольку исключены все индивидуумы с неизменным значением зависимой переменной. Эф-

фективность оценок в связи с этим снизилась. Сохранилась значимая зависимость прежнего знака только от стажа, и по-прежнему значимыми остались временные эффекты.

Вообще же более уместно сравнение приведенных регрессий проводить с помощью предельных эффектов.

9.2. Модель тобит

Модель тобит используется, если зависимая переменная является количественной, но не все ее значения доступны для наблюдения, например, мы не можем наблюдать заработную плату индивидуума, если она меньше величины резервной заработной платы, или мы не можем наблюдать в данных РМЭЗ заработную плату индивидуумов из высокодоходных групп населения. Формулировка модели тобит со случайным эффектом отличается от формулировки модели пробит со случайным эффектом правилом отбора наблюдений [19]:

$$y_{it}^{*} = X_{it}'\beta + \alpha_i + \varepsilon_{it},$$

где $y_{it} = y_{it}^{*}$, если $y_{it}^{*} > 0$ (здесь для простоты взят ноль в качестве ограничения снизу);

$y_{it} = 0$, если $y_{it}^{*} \le 0$.

Относительно α_i и ε_{it} делаются обычные предположения о нормальности, независимости, одинаковой распределенности с нулевым математическим ожиданием и дисперсиями σ_α^2 и σ_ε^2 и независимости от $X_{i1},...,X_{iT}$. При этих предположениях функция правдоподобия запишется в виде

$$f(y_{i1},...,y_{iT} \mid X_{i1},...,X_{iT},\beta) = \int_{-\infty}^{\infty} \prod_t f(y_{it} \mid X_{it},\alpha_i,\beta) f(\alpha_i)\, d\alpha_i,$$

где $f(\alpha_i) = \dfrac{1}{\sqrt{2\pi}\sigma_\alpha} \exp\left(-\dfrac{1}{2}\dfrac{\alpha_i^2}{\sigma_\alpha^2}\right)$;

$$f(y_{it} \mid X_{it},\alpha_i,\beta) = \frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}} \exp\left(-\frac{1}{2}\frac{(y_{it} - X_{it}'\beta - \alpha_i)^2}{\sigma_\varepsilon^2}\right), \text{ если } y_{it} > 0$$

$$= 1 - \Phi\left(\frac{X_{it}'\beta + \alpha_i}{\sigma_\varepsilon}\right), \text{ если } y_{it} = 0.$$

Для создания полноты картины следует добавить, что можно использовать другие формы цензурирования, например, для оценивания упорядоченной пробит-модели со случайными эффектами. Во всех случаях интеграл по α_i будет браться численно.

Лекция 10.

10. Методы борьбы с истощением выборки

10.1. Анализ несбалансированных панелей

До сих пор мы имели дело только с полными или сбалансированными панелями, т.е. случаем, когда все объекты наблюдаются на протяжении одного и того же периода времени. Однако в типичных экономических эмпирических приложениях чаще приходится иметь дело с неполными панелями. Например, в процессе сбора данных относительно американских авиалиний через какое-то время обнаруживается, что некоторые фирмы ушли с рынка, в то время как новые участники появились в течение наблюдаемого периода. Точно так же при исследовании рабочей силы или анализе потребления на базе панелей домашних хозяйств можно обнаружить, что некоторые домашние хозяйства переместились или распались и больше не могут быть включены в панель.

Аналогичная ситуация имеет место при сборе данных относительно набора стран. Некоторые страны могут быть прослежены назад в прошлое дольше, чем другие.

Эти типичные сценарии ведут к несбалансированным или неполным панелям.

В этой главе будут изложены эконометрические проблемы, связанные с оцениванием таких неполных панелей, и проанализированы их отличия от случая полных данных. Мы будем предполагать в этой главе, и это очень существенное предположение, что панельные данные являются неполными из-за случайно пропущенных наблюдений.

10.1.1. Модель однокомпонентной ошибки с несбалансированными данными

Чтобы упростить изложение, мы будем анализировать панель, состоящую всего из двух объектов, с неравным числом временных наблюдений, т.е. два временных ряда неравной длины для двух различных индивидуумов (или стран, фирм и т.д.) и затем обобщим анализ для случая N единиц.

Пусть n_1 будет длина временного ряда, наблюдаемого для первого индивидуума ($i = 1$) и n_2 будет числом дополнительных наблюдений ряда времени, доступных для второго индивидуума ($i = 2$). В таком случае для второго индивидуума у нас в распоряжении имеется ($n_1 + n_2$) наблюдений. Тогда уравнение модели можно записать в виде

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \beta + \begin{pmatrix} u_1 \\ u_2 \end{pmatrix},$$

где ошибки u_i содержат случайную индивидуальную компоненту $(u_{it} = \mu_i + \varepsilon_{it})$; y_1 и y_2 – векторы размерности n_1 и $n_1 + n_2$ соответственно; X_1 и X_2 – матрицы размерности $n_1 \times K$ и $(n_1 + n_2) \times K$ соответственно. В этом случае $u_1' = \left(u_{11}, \ldots, u_{1n_1}\right)$ и

$u_2' = \left(u_{21},\ldots,u_{2n_1},\ldots,u_{2,n_1+n_2}\right)$, и ковариационная матрица вектора случайного возмущения имеет вид

$$\Omega = \begin{bmatrix} \sigma_\varepsilon^2 I_{n_1} + \sigma_\mu^2 J_{n_1} & 0 & 0 \\ 0 & \sigma_\varepsilon^2 I_{n_1} + \sigma_\mu^2 J_{n_1} & \sigma_\mu^2 J_{n_1 n_2} \\ 0 & \sigma_\mu^2 J_{n_1 n_2} & \sigma_\varepsilon^2 I_{n_2} + \sigma_\mu^2 J_{n_2} \end{bmatrix},$$

где I_{n_i} обозначает единичную матрицу порядка n_i; J_{n_i} – квадратная матрица из единиц порядка n_i; $J_{n_i n_j}$ – прямоугольная матрица из единиц размерности $n_i \times n_j$. Заметим, что все ненулевые внедиагональные элементы равны σ_μ^2. Кроме того, если положить $T_j = \sum_{i=1}^{j} n_i$ для j = 1,2, то очевидно, что Ω-блочно-диагональная матрица с j-ым блоком

$$\Omega_j = \left(T_j\sigma_\mu^2 + \sigma_\varepsilon^2\right)\frac{J_{T_j}}{T_j} + \sigma_\varepsilon^2\left(I_{T_j} - \frac{J_{T_j}}{T_j}\right) \text{ и}$$

$\sigma_\mu^2\Omega_j^{-1/2} = I_{T_j} - \left(1-\theta_j\right)\frac{J_{T_j}}{T_j}$, где $\theta_j^2 = \frac{\sigma_\varepsilon^2}{T_j\sigma_\mu^2 + \sigma_\varepsilon^2}$.

Вектор $\sigma_\mu^2\Omega_j^{-1/2}y_i$ будет состоять из элементов вида $y_{jt} - \theta_j y_{j\bullet}$, где $y_{j\bullet} = \frac{1}{T_j}\sum_{t=1}^{T_j} y_{jt}$. Заметим, что θ_j зависит для каждого j-го индивидуума от T_j. Таким образом, оценки ОМНК могут быть получены так же, как и в полных панелях. Основное различие состоит в том, что в неполных панельных данных веса существенно зависят от длины временного ряда, который имеется в нашем распоряжении для конкретного индивидуума.

Полученный выше результат обобщается в двух направлениях:

1) сделанные выводы остаются справедливыми независимо от длины периода, на протяжении которого наблюдаются оба объекта одновременно;

2) результаты распространяются от случая выборки из двух объектов до случая выборки из N объектов.

Доказательство просто. Так как недиагональные элементы матрицы ковариации равны нулю для наблюдений, принадлежащих различным фирмам, Ω остается блочно-диагональной, поскольку наблюдения упорядочены по фирмам. Также все ненулевые внедиагональные элементы равны σ_μ^2. Таким образом, матрица $\Omega_j^{-1/2}$ может быть получена тем же способом, что и выше.

В общем виде регрессионная модель с однокомпонентной случайной ошибкой для несбалансированных панелей задается в виде

$$y_{it} = \alpha + X'_{it}\beta + u_{it},\ i = 1,...,N,\ t = 1,...,T_i$$
$$u_{it} = \mu_i + \varepsilon_{it},$$

где X_{it} – вектор регрессоров; $\mu_i \sim iiN\left(0, \sigma_\mu^2\right)$ и не зависит от $\varepsilon_i \sim iiN\left(0, \sigma_\varepsilon^2\right)$. Эта панель несбалансирована, поскольку разные индивидуумы наблюдаются в течение различных временных периодов T_i для $i = 1,...,N$.

Переписав это уравнение в векторной форме, мы получаем

$$y = \alpha\vec{i}_n + X\beta + u = Z\delta + u$$
$$u = Z_\mu\mu + \varepsilon,$$

где y и Z размерностей (n, 1) и (n, K) соответственно; $Z = \left(\vec{i}_n, X\right)$, $\delta' = (\alpha', \beta')$, $n = \sum T_i$, $Z_\mu = diag\left(\vec{i}_{T_i}\right)$ и $\vec{i}_{T_i}, \vec{i}_n$ – вектора из единиц размерности T_i и n соответственно; $\mu = (\mu_1, \mu_2, ..., \mu_n)'$ и $\varepsilon = \left(\varepsilon_{11}, ..., \varepsilon_{1T_1}, ..., \varepsilon_{N1}, ..., \varepsilon_{NT_N}\right)'$.

Оценка МНК

$$\hat{\delta}_{МНК} = \left(Z'Z\right)^{-1} Z'y$$

будет наилучшей линейной несмещенной оценкой, если σ_μ^2 равна нулю. Даже когда σ_μ^2 положительна, МНК дает несмещенные и состоятельные оценки коэффициентов (если, конечно, нет корреляции между X и ошибкой), смещены будут лишь стандартные ошибки. Обозначим регрессионные остатки $\hat{u}_{МНК} = y - Z\hat{\delta}_{МНК}$.

Оценка «within» вектора коэффициентов в

$$\tilde{\beta} = \left(\tilde{X}'\tilde{X}\right)^{-1} \tilde{X}'\tilde{y}$$

может быть получена преобразованием $Q = diag\left(I_{T_i} - \frac{J_{T_i}}{T_i}\right)$ зависимой и независимых переменных: $\tilde{X} = QX$, $\tilde{y} = Qy$. Оценка свободного члена: $\tilde{\alpha} = y_{\bullet\bullet} - X_{\bullet\bullet}\tilde{\beta}$, где $y_{\bullet\bullet} = \sum\sum y_{it} / n$. Остатки «within» имеют вид $\tilde{u} = y - \tilde{\alpha}I_n - X\tilde{\beta}$.

Оценка «between» вектора коэффициентов:

$$\hat{\delta}_B = (Z'PZ)^{-1} Z'Py,$$

где $P = diag\left(\frac{J_{T_i}}{T_i}\right)$ и остатки «between» имеют вид $\hat{u}^B = y - Z\hat{\delta}_B$.

Оценка ОМНК в случае, если известна истинная ковариационная матрица, получается следующим образом:

$$\hat{\delta}_{ОМНК} = \left(Z'\Omega^{-1}Z\right)^{-1} Z'\Omega^{-1}y,$$

где $\Omega = \sigma_\varepsilon^2 \Sigma = E(uu')$ с $\Sigma = diag\left[\left(T_j \sigma_\mu^2 / \sigma_\varepsilon^2 + 1\right)\frac{J_{T_j}}{T_j}\right] + diag\left(I_{T_j} - \frac{J_{T_j}}{T_j}\right)$.

Теперь следует обратиться к более естественной ситуации с неизвестной ковариационной матрицей, компоненты которой подлежат оцениванию.

10.1.2. ANOVA-методы оценки ковариационных матриц

ANOVA-метод – один из самых популярных методов для оценки компонент дисперсии. Он представляет собой разновидность метода моментов, в котором приравнивают суммы квадратов остатков к их математическим ожиданиям и решают получившуюся таким образом линейную систему уравнений. Для сбалансированной модели ANOVA-оценки – лучшие квадратичные несмещенные (BQU – best quadratic unbiased) оценки компонент дисперсии [14]. При условии нормальности случайных возмущений эти ANOVA-оценки являются несмещенными и эффективными. Для несбалансированной модели с однокомпонентной ошибкой BQU-оценки компонент дисперсии есть функция самих этих компонент, но оптимальные свойства этих оценок, кроме несмещенности, утрачиваются [16].

ANOVA-методы для несбалансированных панелей являются обобщением случая сбалансированных панелей.

Рассмотрим две квадратичные формы, определяющие «within» и «between» суммы квадратов случайных ошибок:

$$q_1 = u'Qu \quad \text{и} \quad q_2 = u'Pu,$$

где $Q = diag\left(I_{T_i} - \frac{J_{T_i}}{T_i}\right)$ и $P = diag\left(\frac{J_{T_i}}{T_i}\right)$. Поскольку истинные случайные возмущения неизвестны, мы, следуя Уолласу и Хусейну [21], будем пользоваться остатками МНК и рассматривать математические ожидания квадратичных форм:

$$E(\hat{q}_1) = E(\hat{u}'_{МНК} Q \hat{u}_{МНК}) = \delta_{11}\sigma_\mu^2 + \delta_{12}\sigma_\varepsilon^2$$
$$E(\hat{q}_2) = E(\hat{u}'_{МНК} P \hat{u}_{МНК}) = \delta_{21}\sigma_\mu^2 + \delta_{22}\sigma_\varepsilon^2,$$

где

$$\delta_{11} = tr\left((Z'Z)^{-1} Z'Z_\mu Z'_\mu Z\right) - tr\left((Z'Z)^{-1} Z'PZ (Z'Z)^{-1} Z'Z_\mu Z'_\mu Z\right),$$
$$\delta_{12} = n - N - K + tr\left((Z'Z)^{-1} Z'PZ\right),$$
$$\delta_{21} = n - 2tr\left((Z'Z)^{-1} Z'Z_\mu Z'_\mu Z\right) + tr\left((Z'Z)^{-1} Z'PZ (Z'Z)^{-1} Z'Z_\mu Z'_\mu Z\right),$$
$$\delta_{22} = N - tr\left((Z'Z)^{-1} Z'PZ\right).$$

Подставляя $\hat{q}_i$ вместо их математических ожиданий и решая полученную систему уравнений, мы приходим к оценкам компонент дисперсии типа Уолласа и Хусейна [21].

Аналогично мы можем подставить остатки «within» в исходные квадратичные формы и получить $\tilde{q}_1 = \hat{u}'_W Q \hat{u}_W$ и $\tilde{q}_2 = \hat{u}'_W P \hat{u}_W$, что предлагал делать Амемия

[2] для сбалансированного случая. Математические ожидания этих квадратичных форм:

$$E(\tilde{q}_1) = (n-N-K+1)\sigma_\varepsilon^2,$$
$$E(\tilde{q}_2) = \left(N-1+tr\left[(X'QX)^{-1}X'PX\right]-tr\left[(X'QX)^{-1}X'\frac{J_n}{n}X\right]\right)\sigma_\varepsilon^2+\left[n-\frac{1}{n}\sum_{i=1}^{N}T_i^2\right]\sigma_\mu^2.$$

Приравнивая $\tilde{q}_i$ соответствующим математическим ожиданиям, мы получаем оценки компонент дисперсии, предложенные Амемией:

$$\hat{\sigma}_\varepsilon^2 = \hat{u}_W'Q\hat{u}_W/(n-N-K+1),$$
$$\hat{\sigma}_\mu^2 = \frac{\hat{u}_W'P\hat{u}_W-\left\{N-1+tr\left[(X'QX)^{-1}X'PX\right]-tr\left[(X'QX)^{-1}X'\frac{J_n}{n}X\right]\right\}\hat{\sigma}_\varepsilon^2}{n-\sum_{i=1}^{N}T_i/n}.$$

Можно получить оценки Свами и Арора [15], пользуясь суммами квадратов остатков регрессий «within» и «between» одновременно. Здесь приравниваются квадратичные формы $\tilde{q}_1 = \hat{u}_W'Q\hat{u}_W$ и $\hat{q}_2^B = \hat{u}_B'P\hat{u}_B$ и их математические ожидания:

$$E(\tilde{q}_1) = (n-N-K+1)\sigma_\varepsilon^2$$
$$E\left(\hat{q}_2^B\right) = \left[n-tr\left((Z'PZ)^{-1}Z'Z_\mu Z_\mu'Z\right)\right]\sigma_\mu^2+(N-K)\sigma_\varepsilon^2.$$

Оценка $\hat{\sigma}_\varepsilon^2 = \hat{u}_W'Q\hat{u}_W/(n-N-K+1)$ как у Амемии, а оценка параметра σ_μ^2 выглядит следующим образом:

$$\hat{\sigma}_\mu^2 = \frac{\hat{u}_B'P\hat{u}_B-(N-K)\hat{\sigma}_\varepsilon^2}{n-tr\left((Z'PZ)^{-1}Z'Z_\mu Z_\mu'Z\right)}.$$

Заметим, что первое слагаемое числителя – $\hat{u}_B'P\hat{u}_B$ – может быть получено как сумма квадратов остатков регрессии $\sqrt{T_i}y_{i\bullet}$ на $\sqrt{T_i}Z_{i\bullet}$.

В заключение можно привести еще один метод, который в литературе получил название метода Хендерсона – Фуллера – Баттес [7, 11]. Этот метод использует предсказанные значения констант:

$$\hat{\sigma}_\varepsilon^2 = \frac{y'y-R(\delta\,|\,\mu)-R(\mu)}{n-N-K+1}$$
$$\hat{\sigma}_\mu^2 = \frac{R(\mu\,|\,\delta)-(N-1)\hat{\sigma}_\varepsilon^2}{n-tr\left(Z_\mu'Z(Z'Z)^{-1}Z'Z_\mu\right)},$$

где $R(\mu) = y'Z_\mu\left(Z_\mu'Z_\mu\right)^{-1}Z_\mu'y = \sum_{i=1}^{N}\frac{y_{i\bullet}^2}{T_i}$, $R(\delta\,|\,\mu) = \tilde{y}'\tilde{X}\left(\tilde{X}'\tilde{X}\right)^{-1}\tilde{X}'\tilde{y}$, $R(\delta) = y'Z(Z'Z)^{-1}Z'y$ и $R(\mu\,|\,\delta) = R(\delta\,|\,\mu)+R(\mu)-R(\delta)$.

Помимо изложенных методов используется метод максимального правдоподобия.

10.2. Панели с замещением

Впервые панели с замещением рассмотрел в 1981 г. Бьерн [4]. Конструирование подобных панелей преследует цель поддержания одного и того же числа объектов, например домохозяйств, в выборке на протяжении всего периода наблюдения. Это достигается за счет того, что выбывающие к моменту нового этапа опроса домохозяйства замещаются таким же числом новых домохозяйств, не участвовавших в опросе ранее. Подобная мера призвана препятствовать постоянно происходящему истощению выборки, так как домохозяйства могут менять место проживания, распадаться, делиться, наконец, просто по каким-то причинам отказываться дальше участвовать в опросе. Так, в бюджетном обследовании норвежских домохозяйств, на основании которого написал свою работу Бьерн, половина выборки обновлялась на каждом этапе опроса.

Чтобы проиллюстрировать основные приемы работы с такими панелями, предположим для простоты, что $T = 2$ и половина выборки обновилась во втором периоде. В этом случае без потери общности домохозяйства 1, 2, ..., $N/2$ заменены домохозяйствами N+1, N+2, ..., N+$N/2$ в период 2. Очевидно, что только домохозяйства $N/2$+1, $N/2$+2, ..., N наблюдаются на протяжении обоих периодов. Первые же и последние $N/2$ домохозяйств наблюдаются только по одному периоду. В целом же наблюдение ведется над $3N/2$ домохозяйствами.

Теперь рассмотрим обычную модель со случайным индивидуальным эффектом

$$u_{it} = \mu_i + \varepsilon_{it}$$

с $\mu_i \sim iid(0,\sigma_\mu^2)$ и $\varepsilon_{it} \sim iid(0,\sigma_\varepsilon^2)$, независимыми друг от друга и X_{it}. Пронумеруем наблюдения как обычно так, чтобы первый индекс отвечал за номер домохозяйства, а последний за номер периода наблюдения, но упорядочим их немного иначе, чем ранее:

$$u' = \left(u_{11}, u_{21}, ..., u_{N1}, u_{N/2+1,2}, ..., u_{3N/2,2}\right) \text{ и}$$

$$E(uu') = \Omega = \begin{bmatrix} \sigma^2 I_{N/2} & 0 & 0 & 0 \\ 0 & \sigma^2 I_{N/2} & \sigma_\mu^2 I_{N/2} & 0 \\ 0 & \sigma_\mu^2 I_{N/2} & \sigma^2 I_{N/2} & 0 \\ 0 & 0 & 0 & \sigma^2 I_{N/2} \end{bmatrix}, \text{ где } \sigma^2 = \sigma_\mu^2 + \sigma_\varepsilon^2.$$

Легко заметить, что матрица Ω – блочно-диагональная и средний блок имеет вид, традиционный для модели однокомпонентной ошибки:

$$\sigma_\mu^2\left(J_2 \otimes I_{N/2}\right) + \sigma_\varepsilon^2\left(I_2 \otimes I_{N/2}\right).$$

Кроме того

$$\Omega^{-1/2}=\begin{bmatrix} \frac{1}{\sigma}I_{N/2} & 0 & 0 \\ 0 & \left(\frac{1}{\sigma_1^*}\frac{J_2}{2}+\frac{1}{\sigma_\varepsilon}\left(I_2-\frac{J_2}{2}\right)\right)\otimes I_{N/2} & 0 \\ 0 & 0 & \frac{1}{\sigma}I_{N/2} \end{bmatrix},\ \text{где}\ \left(\sigma_1^*\right)^2=2\sigma_\mu^2+\sigma_\varepsilon^2.$$

Домножив исходное регрессионное уравнение на $\Omega^{-1/2}$ и применив к преобразованной модели МНК, можно получить оценки ОМНК для панели с замещением. Для этого нужно просто поделить данные для первых $N/2$ и последних $N/2$ наблюдений на σ. Средние N наблюдений с номерами $i = N/2+1,\ N/2+2,...,N$ и $t = 1,\ 2$ нужно преобразовать следующим образом:

$$(1/\sigma_\varepsilon)\left(Y_{it}-\theta^*\overline{Y}_{i\bullet}\right),\ \text{где}\ \theta^*=1-\sigma_\varepsilon/\sigma_1^*\ \ \text{и}\ \ \overline{Y}_{i\bullet}=(Y_{i1}+Y_{i2})/2.$$

Аналогичным преобразованиям подвергаются и регрессоры.

Поскольку оценки ОМНК, как правило, бывают недоступны, можно заменить их оценками реализуемого ОМНК, т.е. РОМНК, оценив параметры σ_μ^2 и σ_ε^2. Одна состоятельная оценка σ_ε^2 может быть получена из средних N наблюдений (т.е. из наблюдений за домохозяйствами, которые присутствуют в выборке на протяжении всех периодов наблюдения). Для этих наблюдений σ_ε^2 состоятельно оценивается на основании остатков «within»:

$$\tilde{\sigma}_\varepsilon^2=\frac{1}{N}\sum_{t=1}^{2}\sum_{i=N/2+1}^{N}\left[\left(Y_{it}-\overline{Y}_{i\bullet}\right)-\left(X_{it}-\overline{X}_{i\bullet}\right)'\tilde{\beta}_W\right]^2,$$

в то время как полная дисперсия может быть вычислена состоятельно из МНК-остатков регрессии по всей выборке:

$$\tilde{\sigma}^2=\tilde{\sigma}_\varepsilon^2+\tilde{\sigma}_\mu^2=\frac{2}{3N}\sum_{t=1}^{2}\sum_{i=1}^{3N/2}\left(Y_{it}-X_{it}'\hat{\beta}_{МНК}\right)^2.$$

Мы можем переупорядочить наблюдения так, чтобы домохозяйства, наблюдаемые в течение одного периода, стояли в начале, а домохозяйства, наблюдаемые в течение двух периодов, стояли в конце. Такой способ расположения данных сводит задачу оценивания панели с замещением к задаче оценивания несбалансированной панели, в которой N домохозяйств наблюдаются один период и $N/2$ домохозяйств наблюдаются два периода.

Изложенный РОМНК-метод оценивания панелей с замещением легко распространяется на трехпериодные панели с ротацией $N/2$ домохозяйств, а также на трехпериодные панели с ротацией $N/3$ домохозяйств в каждый период и т.д. Для более общих схем ротации и более длинных панелей лучше использовать метод максимального правдоподобия.

Анализ панелей с замещением может быть также легко распространен на системы внешне не связанных уравнений, на системы одновременных уравнений и на динамические модели.

Оптимальный выбор периода ротации изучается в работе Ниймана, Вербика и Ван Суста [13] с помощью оценивания линейных комбинаций средних значений наблюдаемых величин по периоду.

Панели с замещением позволяют исследователю протестировать наличие специфических смещений, связанных с моментом структурных сдвигов, которые могут иметь место при пролонгированных обследованиях, т.е. выявить наличие значимых изменений в ответах на одни и те же вопросы у индивидуумов, интервьюируемых с начального периода обследования, и у индивидуумов, подключенных к опросу позже.

10.3. Псевдопанели

Для некоторых стран панели могут не существовать. Вместо них исследователь может располагать обширными данными ежегодных опросов домохозяйств, т.е. повторными пространственными выборками. Возможна ли качественная идентификация параметров регрессионных моделей на основании этих данных? В работах Ниймана и Вербика [20] показано, что в ряде случаев оценки некоторых параметров моделей, основанных на повторных пространственных выборках, оказываются более эффективны, чем оценки, полученные при анализе панелей. Тем не менее некоторые авторы (Хекман и Робб [10], Дитон и Моффитт [6, 12]) настаивают на том, что для оценивания многих общепринятых моделей могут быть использованы синтетические панельные данные, построенные на выборках когорт, а не индивидуумов.

В данном разделе будет рассмотрена модель с индивидуальным эффектом, коррелированным с регрессорами (модель с детерминированным эффектом), и проанализированы свойства оценок, полученных на основании панелей когорт, сконструированных из серий независимых пространственных данных. В этом подходе схожие по некоторым признакам индивидуумы группируются в когорты, после чего выборочные средние характеристики этих когорт рассматриваются в качестве наблюдений в синтетической панели. Поскольку наблюдаемые средние по когортам представляют собой как бы измеренные с ошибкой значения истинных теоретических характеристик когорты, Дитоном было предложено рассматривать модель с ошибками измерения переменных. Эта модель дает состоятельные оценки при достаточно слабых исходных предположениях.

Однако, если число наблюдений в когорте велико, проблему ошибок измерения можно игнорировать и применять к панели когорт те же методы анализа, что и к естественным панелям.

10.3.1. Оценивание по данным о когортах

Рассмотрим следующую линейную модель:

$$y_{it} = X'_{it}\beta + \mu_i + \varepsilon_{it},$$

где индекс i нумерует индивидуумов, индекс t нумерует временные периоды. Будем предполагать также, что $E\left[\varepsilon_{it} \mid X_{js}\right] = 0$ для любых i, j, t, s. В каждый период

времени доступны наблюдения над N независимыми индивидуумами, т.е. не предполагается, что в предыдущий период наблюдались те же самые респонденты, или не сохраняются их идентификационные номера, присвоенные в предыдущий период.

Во многих приложениях индивидуальный эффект μ_i коррелирован с объясняющими переменными X_{it}, так что модель со случайным индивидуальным эффектом дает несостоятельные оценки и следует пользоваться моделью с детерминированным индивидуальным эффектом. Когда доступны естественные панельные данные, такая модель оценивается с помощью преобразования «within», элиминирующего индивидуальные эффекты. Однако, очевидно, что эта стратегия неприемлема, когда вместо панели доступны повторные пространственные выборки.

В 1985 г. Дитон предложил конструировать для таких случаев псевдопанель или панель из когорт.

Пусть по некоторым общим признакам индивидуумы группируются в C когорт. Эти когорты определяются таким образом, чтобы каждый индивидуум был членом только одной когорты, и определение самой когорты не меняется в течение всего периода наблюдения. Например, когорта может состоять из мужчин, родившихся в 1945–1949 гг. Модель, записанная для созданных таким образом когорт, примет вид

$$y_{ct} = X'_{ct}\beta + \mu_{ct} + \varepsilon_{ct}, \text{ где } c = 1,...,C, \;\; t = 1,...,T.$$

Главная проблема при оценивании этой модели состоит в том, что индивидуальный эффект когорты зависит от времени и одновременно коррелирует с регрессорами. Пренебрежение этой корреляцией приводит к несостоятельным оценкам, а учет – к проблеме с идентификацией параметров, которая снимается только в случае, если зависимостью μ_{ct} от t можно пренебречь. Последнее возможно, если число индивидуумов в когорте велико.

Альтернативный путь решения проблемы предложил Дитон, который рассматривал регрессионную модель не для наблюдаемых выборочных реализаций когорт, а для когорт во всей их генеральной совокупности:

$$y^*_{ct} = X^{*'}_{ct}\beta + \mu^*_c + \varepsilon^*_{ct}, \text{ где } c = 1,...,C, \;\; t = 1,...,T,$$

где величины со звездочкой означают генеральные средние по когортам, и детерминированный индивидуальный эффект когорты теперь не зависит от времени, так как генеральные совокупности когорт содержат одних и тех же индивидуумов на протяжении всего периода наблюдения. Если вдруг окажется, что генеральные средние по когортам наблюдаемы, то последняя модель может быть оценена стандартными средствами. Но так как такая ситуация нереальна, естественнее рассматривать предыдущую модель с выборочными средними по когортам, которые трактуются как измеренные с ошибками генеральные средние. Дитон предположил, что ошибки измерения нормально распределены с нулевым средним и не зависят от истинных значений генеральных средних, т.е.

$$\begin{pmatrix} y_{ct} \\ X_{ct} \end{pmatrix} \sim N\left(\begin{pmatrix} y^*_{ct} \\ X^*_{ct} \end{pmatrix}; \;\; \begin{pmatrix} \sigma_{00} & \sigma' \\ \sigma & \Sigma \end{pmatrix} \right).$$

Один из путей оценивания параметра β лежит в рамках модели с ошибками измерения. Если обозначить вектор-строку дамми-переменных, отвечающих индивидуальному эффекту когорт, через d_c', а вектор столбец соответствующих им коэффициентов через $\mu^* = (\mu_1^*, \ldots, \mu_C^*)'$, то предложенная Дитоном оценка вектора всех коэффициентов будет выглядеть следующим образом:

$$\begin{pmatrix} \hat{\mu} \\ \hat{\beta} \end{pmatrix} = \left(\sum_{c=1}^{C} \sum_{t=1}^{T} \begin{pmatrix} d_c' d_c & d_c' X_{ct} \\ X_{ct}' d_c & X_{ct}' X_{ct} - \hat{\Sigma} \end{pmatrix} \right)^{-1} \left(\sum_{c=1}^{C} \sum_{t=1}^{T} \begin{pmatrix} d_c' y_{ct} \\ X_{ct}' y_{ct} - \hat{\sigma} \end{pmatrix} \right),$$

где $\hat{\Sigma}$ и $\hat{\sigma}$ – оценки, полученные на основании индивидуальных наблюдений. Если будут выполнены нижеследующие условия, то оценка $\hat{\beta}$ будет состоятельна, когда общее число наблюдений $CT \to \infty$, а оценка $\hat{\mu}$ будет состоятельна, когда $NT / C \to \infty$.

Утверждение 1. Матрица моментов теоретических средних объясняющих переменных $\underset{CT \to \infty}{p\lim} \frac{1}{CT} \sum_{c=1}^{C} \sum_{t=1}^{T} \begin{pmatrix} d_c' d_c & d_c' X_{ct} \\ X_{ct}' d_c & X_{ct}' X_{ct} - \hat{\Sigma} \end{pmatrix}$ не сингулярна.

Если число наблюдений в когорте не слишком мало, можно попытаться игнорировать проблему ошибок измерения и оценивать модель, предполагая, что генеральные и выборочные средние эквивалентны. В этом случае можно будет получить оценку $\hat{\beta}_W$:

$$\hat{\beta}_W = \left(\sum_{c=1}^{C} \sum_{t=1}^{T} (X_{ct} - X_c)' (X_{ct} - X_c) \right)^{-1} \left(\sum_{c=1}^{C} \sum_{t=1}^{T} (X_{ct} - X_c)' (y_{ct} - y_c) \right),$$

где $X_c = \frac{1}{T} \sum_{t=1}^{T} X_{ct}$, $y_c = \frac{1}{T} \sum_{t=1}^{T} y_{ct}$.

$\hat{\beta}_W$ будет несмещенной, если $E[\mu_{ct} - \mu_c \mid X_{ct} - X_c] = 0$, т.е. индивидуальный эффект когорт не коррелирует с регрессорами, и выполняется следующее утверждение.

Утверждение 2. Матрица моментов выборочных средних объясняющих переменных по когортам $\underset{CT \to \infty}{p\lim} \frac{1}{CT} \sum_{c=1}^{C} \sum_{t=1}^{T} (X_{ct} - X_c)' (X_{ct} - X_c)$ не сингулярна.

Если число наблюдений в когорте N/C велико, то условие $E[\mu_{ct} - \mu_c \mid X_{ct} - X_c] = 0$ выполняется. Но при этом следует отметить, что увеличение числа наблюдений в когорте приводит к уменьшению числа самих когорт в синтетической панели и, таким образом, к увеличению стандартных ошибок оценки $\hat{\beta}_W$. Выбор оптимального принципа разбиения на когорты должен учитывать влияние разбиения на величину смещения оценок и на величину дисперсии оценок.

10.3.2. Влияние выбора когорт на величину смещения

Первое, что нам предстоит выяснить, – справедливость утверждения, что большое число наблюдений в когорте избавляет от необходимости учитывать проблему ошибок измерения. Зафиксируем для простоты число наблюдений в когорте *N/C*. Чтобы упростить аналитические результаты, мы аппроксимируем смещение конечных выборок асимптотическим смещением при больших *C* и *N*. Как показали численные эксперименты, эта аппроксимация достаточно точна при *C* ~ 10–20. Будем рассматривать для простоты линейную модель с одним регрессором:

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

где, следуя Чемберлену, предположим, что выполняется следующее утверждение.

Предположение 1. *Регрессионная зависимость индивидуального эффекта с регрессором имеет вид*

$$\mu_i = \lambda X_{i\bullet} + \xi_i, \quad \text{где } E(\xi_i \mid X_{it}) = 0 \text{ для всех } t = 1,\ldots,T \text{ и } V(\xi_i) = \sigma_\xi^2.$$

Тогда $\lambda = 0$ – достаточное условие состоятельности $\hat{\beta}_W$. Принцип конструирования когорт формулируется следующим образом.

Предположение 2. *Когорты формируются на основании непрерывно распределенных, с дисперсией равной единице, независимых по индивидуумам величин z. Более того, когорты выбираются так, чтобы безусловная вероятность принадлежности к одной из них была одна и та же для всех когорт.*

В соответствии с этим предположением все когорты имеют примерно одно и то же число членов. На практике переменная *z* может основываться более чем на одной переменной, но выбор *z* ограничен следующими соображениями:

- z_i должна быть постоянной по времени для всех индивидуумов, так как индивидуумы не должны переходить из когорты в когорту;
- z_i должна быть наблюдаема для всех индивидуумов в выборке.

Последнее требование означает, что в качестве z_i не может быть использована переменная типа «заработная плата в 1988 г.» или «размер семьи на 1 января 1990 г.», так как подобные переменные, как правило, не наблюдаются для всех индивидуумов в выборке. Обычно на практике для выделения когорт используются такие переменные, как пол и дата рождения.

Чтобы оценки обладали хорошими свойствами, генеральные средние по когортам должны изменяться и по когортам, и по времени. Для моделирования этого обстоятельства будем использовать следующее предположение.

Предположение 3. *Регрессионная зависимость между* X_{it} *и* z_i *имеет вид* $X_{it} = \theta_t + \gamma_t z_i + v_{it}$, *где* θ_t *– детерминированный временной эффект, а* v_{it} *не коррелирует с* z_i, *имеет нулевое математическое ожидание и постоянную дисперсию* σ_v^2, *и* $E\{v_{it}v_{is}\} = \rho\sigma_v^2$ *для* $s \neq t$.

Можно показать, что при всех выше сделанных предположениях асимптотическое смещение оценки «within» имеет вид

$$p\lim_{C\to\infty}\left(\hat{\beta}_W-\beta\right)=\lambda\left[\frac{1+(T-1)\rho}{T}\right]\frac{\tau\omega_2}{\omega_1+\tau\omega_2}=\delta,$$

где $\tau=(T-1)/T$, ω_2 – дисперсия ошибки измерения X_{ct}, т.е.

$$\omega_2=p\lim_{C\to\infty}\frac{1}{CT}\sum_{c=1}^{C}\sum_{t=1}^{T}\left(X_{ct}-X_{ct}^{*}\right)^2=n_c^{-1}\sigma_v^2,$$

где n_c – число индивидуумов в каждой когорте (N/C), а ω_1 – истинная дисперсия «within» для когорт

$$\omega_1=p\lim_{C\to\infty}\frac{1}{CT}\sum_{c=1}^{C}\sum_{t=1}^{T}\left(X_{ct}^{*}-X_{c}^{*}\right)^2=\frac{1}{T}\sum_{t=1}^{T}\left(\theta_t-\frac{1}{T}\sum_{s=1}^{T}\theta_s\right)^2+\frac{1}{T}\sum_{t=1}^{T}\left(\gamma_t-\frac{1}{T}\sum_{s=1}^{T}\gamma_s\right)^2.$$

В рамках предположения 3 можно легко проверить, что утверждение 1 выполняется при $\omega_1>0$, в то время как утверждение 2 выполняется при $\omega_1+\tau\omega_2>0$, что возможно только тогда, когда θ_t или γ_t варьируются со временем. Если это не так, предел по вероятности для оценки Дитона не существует, а смещение оценки «within» будет максимальным и равным

$$p\lim_{C\to\infty}\left(\hat{\beta}_W-\beta\right)=\lambda\left[\frac{1+(T-1)\rho}{T}\right]=\delta_{\max},$$

что не зависит от размера когорт. Выбор больших когорт будет уменьшать смещение только при $\omega_1>0$. Поскольку ω_2 – убывающая функция n_c, смещение оценки «within» минимально, если число наблюдений в каждой когорте максимально велико.

Если отношение ω_1/σ_v^2 не слишком мало, то реальное смещение будет много меньше, чем максимальное смещение при больших n_c. Например, если $\omega_1/\sigma_v^2=0{,}5$, легко вычислить, что смещение будет меньше 2% от максимального смещения при наличии в когорте 100 или более членов. Если же $\omega_1/\sigma_v^2=0{,}05$, то смещение будет составлять более 17% .

Если говорить об эмпирических приложениях, то чаще всего ошибки измерения игнорируются, и используется стандартная оценка «within». Следует заметить, что размер когорт может быть выбран меньше, если переменная z_i – идентификатор когорты – выбрана таким способом, что истинная дисперсия «within» для когорт велика по сравнению с σ_v^2.

10.3.3. Влияние выбора когорт на дисперсию

В предыдущем подразделе было показано, что смещение оценки «within» для синтетической панели может быть мало, если число наблюдений в когорте

достаточно велико. Однако, увеличивая число наблюдений в когортах, мы тем самым уменьшаем число самих когорт, т.е. число наблюдений в синтетической когорте (CT), а следовательно, увеличиваем дисперсию оценки $\hat{\beta}_W$. В этом подразделе будет более детально проанализировано влияние выбора числа когорт на дисперсию $\hat{\beta}_W$. Будет показано, что разность между истинной дисперсией $\hat{\beta}_W$ и пределом по вероятности ее приближенного значения является исключительно функцией смещения.

Асимптотическая дисперсия $\hat{\beta}_W$ может быть записана следующим образом:

$$V\{\hat{\beta}_W\} = \frac{1}{CT}(\omega_1 + \tau\omega_2)^{-2} V^*,$$

где $V^* = \underset{C\to\infty}{p\lim} V\left\{\frac{1}{\sqrt{CT}}\sum_{c=1}^{C}\sum_{t=1}^{T}(X_{ct} - X_{c\bullet})(\mu_{ct} - \mu_{c\bullet} + \varepsilon_{ct} - \varepsilon_{c\bullet})\right\}$.

Следует заметить, что математическое ожидание выражения в фигурных скобках в последней формуле не равно нулю из-за несостоятельности оценки (если $\lambda \neq 0$). Более того, суммирование по c и t не является ни суммированием независимых, ни суммированием одинаково распределенных величин. Это усложняет дальнейшие выкладки. Но при дополнительных предположениях о том, что X_{ct}, μ_{ct} и ε_{ct} нормально распределены, дисперсию $\hat{\beta}_W$ можно записать в виде

$$V\{\hat{\beta}_W\} = \frac{1}{CT}\left[\left(\sigma_\mu^2 + \sigma_\varepsilon^2 n_c^{-1}\right)(\omega_1 + \tau\omega_2)^{-1} + \delta^2\right],$$

где δ – это асимптотическое смещение оценки $\hat{\beta}_W$ и

$$\sigma_\mu^2 = \sigma_\xi^2 n_c^{-1} + \lambda^2\left[\frac{1+(T-1)\rho}{T}\right]\omega_2.$$

Увеличение размера когорт n_c влияет на дисперсию $\hat{\beta}_W$ двояким образом:

- снижаются дисперсии ошибок измерения ω_2 и ошибки уравнения $\sigma_\mu^2 + \sigma_\varepsilon^2 n_c^{-1}$;
- снижается общее число наблюдений CT.

Первый эффект является доминирующим, так что, увеличивая n_c, можно вызвать уменьшение дисперсии $\hat{\beta}_W$ для синтетической панели.

В стандартных пакетах используется следующая оценка дисперсии:

$$\hat{V}\{\hat{\beta}_W\} = \hat{\sigma}^2\left[\sum_{c=1}^{C}\sum_{t=1}^{T}(X_{ct} - X_{c\bullet})^2\right]^{-1},$$

которая не является состоятельной, но в общем случае сходится по вероятности к

$$\check{V}\{\hat{\beta}_W\} = \check{\sigma}^2\frac{1}{CT}(\omega_1 + \tau\omega_2)^{-1},$$

где $\bar{\sigma}^2 = \underset{C\to\infty}{p\lim}\,\hat{\sigma}^2 = \sigma_\mu^2 + \sigma_\varepsilon^2 n_c^{-1} - \delta^2\left(\omega_1 + \tau\omega_2\right)^{-1}$ представляет собой недооцененную истинную дисперсию ошибки $\left(\sigma_\xi^2 + \sigma_\varepsilon^2\right)n_c^{-1}$. Используя этот предел по вероятности, можно записать оценку дисперсии в виде

$$\breve{V}\left\{\hat{\beta}_W\right\} = \frac{1}{CT}\left[\left(\sigma_\mu^2 + \sigma_\varepsilon^2 n_c^{-1}\right)\left(\omega_1 + \tau\omega_2\right)^{-1} - \delta^2\right].$$

Из этого выражения явствует, что разница между истинной дисперсией и пределом по вероятности оцененной дисперсии равна $2\delta^2 / CT$ и будет мала при небольших смещениях δ.

10.3.4. Приложение: оценивания кривой Энгеля

В качестве иллюстрации рассмотрим оценивание кривой Энгеля для расходов на питание в немецких домохозяйствах, проделанное в работе Вербика и Ниймана [20]. Работа выполнена на основании ежемесячных повторных пространственных выборок из 367 домохозяйств, наблюдаемых в течение 1986 г.

Оцениваемая модель имеет вид

$$w_{it} = \beta \ln X_{it} + \mu_i + \varepsilon_{it}, \quad t = 1,...,12,$$

где w_{it} – доля расходов на питание в общем бюджете; $\ln X_{it}$ – натуральный логарифм общих расходов на товары недлительного пользования. Индивидуальный эффект μ_i отражает влияние специфических характеристик домохозяйств (возраст, образование, размер семьи и т.д.), которые являются неизменными на протяжении периода наблюдения. Очевидно, что эти переменные коррелируют с общими расходами на товары недлительного пользования, и поэтому предпочтительнее использовать модель с детерминированным индивидуальным эффектом. Будем предполагать также, что выполнено предположение 1 предыдущего подраздела:

$$\mu_i = \lambda \ln X_{i\bullet} + \xi_i.$$

Конструирование когорты будет производиться на основании данных о дате рождения главы домохозяйства, как и во многих аналогичных исследованиях. Поскольку связь между возрастом и общими расходами скорее всего нелинейна, в качестве переменной-идентификатора когорты, z_i, выбирается квадрат отклонения индивидуальной даты рождения от средней даты рождения в выборке, выраженной в годах и месяцах. Далее z_i преобразована так, чтобы ее дисперсия была равна единице. Согласно предположению 3

$$\ln X_{it} = \theta_i + \gamma_t z_i + v_{it}.$$

Используя 367 наблюдений сбалансированной подпанели, легко получить состоятельные оценки параметров с помощью метода наименьших квадратов, представленные в таблице (в скобках приведены стандартные ошибки):

β	–0,188(0,006)	θ_1	12,235(0,041)	γ_1	–0,147(0,028)
λ	0,110(0,007)	θ_2	12,085(0,041)	γ_2	–0,132(0,028)
σ_ξ	0,105	θ_3	12,202(0,037)	γ_3	–0,164(0,026)
σ_ε	0,072	θ_4	12,238(0,041)	γ_4	–0,150(0,028)
σ_v^2	0,305	θ_5	12,270(0,043)	γ_5	–0,170(0,030)
ρ	0,634	θ_{61}	12,165(0,041)	γ_6	–0,156(0,028)
		θ_7	12,161(0,046)	γ_7	–0,156(0,022)
ω_1	0,00681	θ_8	12,152(0,042)	γ_8	–0,139(0,029)
		θ_9	12,180(0,039)	γ_9	–0,154(0,027)
		θ_{10}	12,328(0,042)	γ_{10}	–0,162(0,029)
		θ_{11}	12,224(0,043)	γ_{11}	–0,181(0,030)
		θ_{12}	12,385(0,048)	γ_{12}	–0,233(0,033)

Все оцененные значения γ_i отрицательны в предположении, что общие расходы на товары недлительного пользования максимальны в среднем возрасте 49,2 года. Хотя сами значения θ_i и γ_i значимо отличаются от нуля, их общая дисперсия, вычисленная как указано в подразделе 10.3.2 – $\omega_1 = 0{,}00681$, – мала по сравнению с $\sigma_v^2 = 0{,}305$. Хотя зависимость возраста от общих расходов существенна, оснований думать, что вид этой зависимости значимо меняется со временем, не обнаруживается. В частности, оценка, вычисленная методом Дитона с учетом ошибок измерения переменных, указывает на этот факт, поскольку дисперсия этой ошибки обратно пропорциональна ω_1.

Необходимо протестировать справедливость предположения 3 и структуру ковариационной матрицы v_{it}. LM-тест на автокорреляцию первого порядка дает значение тестовой статистики 0,057, что позволяет сделать вывод о том, что наша модель вполне согласуется с данными.

Из выражения $\underset{C\to\infty}{p\lim}\left(\hat{\beta}_W - \beta\right) = \lambda\left[\dfrac{1+(T-1)\rho}{T}\right] = \delta_{\max}$ мы немедленно получаем, что максимальное смещение оценки «within», основанное на данных по когортам за 12 периодов, равно 0,0731, что составляет 39% от (оцененной) истинной величины. Принимая во внимание, что выбор переменной-идентификатора когорт – в наших руках, мы можем элиминировать часть этого смещения, увеличив размер когорт. Это иллюстрируется табл. 1, где теоретическое смещение оценки «within» приведено для различных размеров когорт.

Заметим, что смещение медленно уменьшается с ростом размера когорт. Величины в последних столбцах вычислены в соответствии с выражениями

$$\breve{V}\left\{\hat{\beta}_W\right\} = \breve{\sigma}^2\,\frac{1}{CT}\left(\omega_1 + \tau\omega_2\right)^{-1} \quad \text{и} \quad V\left\{\hat{\beta}_W\right\} = \frac{1}{CT}\left[\left(\sigma_\mu^2 + \sigma_\varepsilon^2 n_c^{-1}\right)\left(\omega_1 + \tau\omega_2\right)^{-1} + \delta^2\right].$$

Несмотря на то, что смещение значительно, различия этих двух стандартных ошибок невелики. Обе стандартные ошибки растут с увеличением размера

когорт, что вызвано уменьшением общего числа наблюдений. На фоне этого роста эффект уменьшения ошибок измерения при увеличении размера когорт пренебрежимо мал.

Таблица 1.

n_c	Смещение (абсолютные значения)	Смещение, %	Предел оцененной стандартной ошибки/ $\sqrt{N}$	Истинная стандартная ошибка/ $\sqrt{N}$
2	0,0695	37,0	0,099	0,124
5	0,0650	34,6	0,152	0,171
10	0,0586	31,2	0,205	0,220
25	0,0453	24,1	0,287	0,298
50	0,0329	17,5	0,348	0,356
75	0,0258	13,7	0,379	0,386
100	0,0212	11,3	0,398	0,404
150	0,0157	8,3	0,420	0,424
200	0,0124	6,6	0,433	0,436

Таким образом, на основании всего вышесказанного можно сделать вывод о том, что в псевдопанелях, состоящих из больших когорт (100, 200 индивидуумов), искусственная природа наблюдений преодолевается.

10.4. Смещение самоотбора в неполных панелях

По различным причинам эмпирические панельные данные часто являются неполными. Причиной неполноты может быть и истощение, и работа с несбалансированными панелями. Иногда респонденты отвечают не на все предложенные им вопросы.

Последствий этой неполноты данных может быть несколько.

Первое последствие имеет вычислительный характер. Большая часть выражений, приведенных выше, не предполагает, что наблюдения могут быть пропущены. Самое простое решение проблемы – исключить из рассмотрения респондентов, по которым имеются пропущенные наблюдения, и рассматривать только тех, о которых имеется полная информация. В этом подходе мы будем оценивать зависимости только по сбалансированным подпанелям. Это удобно с вычислительной точки зрения, но крайне неэффективно: значительная часть информации будет просто отброшена. Использование несбалансированных панелей повышает эффективность оценок, но усложняет вычислительную процедуру.

Второе последствие – возможность смещения отбора – носит более серьезный характер. Если индивидуумы наблюдаются неполностью по эндогенным причинам, то использование сбалансированных подпанелей или несбалансированных панелей не помогут устранить смещение самоотбора оценок. Чтобы развить эту мысль, рассмотрим модель вида

$$y_{it} = X_{it}\beta + \alpha_i + \varepsilon_{it}.$$

Далее определим индикаторную величину r_{it} (ответ) так, что $r_{it} = 1$, если (X_{it}, y_{it}) наблюдаются, и $r_{it} = 0$, если нет. Наблюдения (X_{it}, y_{it}) являются пропущенными случайно, если r_{it} не зависит от α_i и ε_{it}. Это означает, что причины, обуславливающие процесс отбора, не влияют на условное распределение y_{it} при данных X_{it}. Если мы хотим сконцентрироваться на сбалансированных подпанелях, то $r_{i1} = ... = r_{iT} = 1$, и еще мы требуем, чтобы r_{it} не зависели от α_i и $\varepsilon_{i1},...,\varepsilon_{iT}$. В этом случае обычные свойства состоятельности оценок не нарушаются, если мы ограничиваем свое внимание только доступными или полными наблюдениями. Если отбор зависит от случайной ошибки уравнения, оценки МНК, модели со случайным и детерминированным эффектами могут страдать смещением самоотбора.

10.4.1. Оценивание при наличии случайно пропущенных данных

Если индикаторная переменная r_{it} не зависит от каких-бы то ни было ненаблюдаемых величин, то состоятельные оценки моделей с детерминированным и случайным эффектами для несбалансированных панелей можно переписать в виде

$$\hat{\beta}_W = \left(\sum_{i=1}^{N} \sum_{t=1}^{T} r_{it} (X_{it} - X_{i\bullet})(X_{it} - X_{i\bullet})' \right)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} r_{it} (X_{it} - X_{i\bullet})(y_{it} - y_{i\bullet}),$$

$$\hat{\beta}_{OMHK} = \left(\sum_{i=1}^{N} \sum_{t=1}^{T} r_{it} (X_{it} - X_{i\bullet})(X_{it} - X_{i\bullet})' + \sum_{i=1}^{N} \frac{1}{\theta_i^2} T_i (X_{i\bullet} - X_{\bullet\bullet})(X_{i\bullet} - X_{\bullet\bullet})' \right)^{-1} \cdot$$

$$\cdot \left(\sum_{i=1}^{N} \sum_{t=1}^{T} r_{it} (X_{it} - X_{i\bullet})(y_{it} - y_{i\bullet}) + \sum_{i=1}^{N} \frac{1}{\theta_i^2} T_i (X_{i\bullet} - X_{\bullet\bullet})(y_{i\bullet} - y_{\bullet\bullet}) \right),$$

где $\theta_i^2 = \dfrac{\sigma_\varepsilon^2 + T_i \sigma_\alpha^2}{\sigma_\varepsilon^2}$, $y_{i\bullet} = \dfrac{\sum_{t=1}^{T} r_{it} y_{it}}{\sum_{t=1}^{T} r_{it}}$, $X_{i\bullet} = \dfrac{\sum_{t=1}^{T} r_{it} X_{it}}{\sum_{t=1}^{T} r_{it}}$.

Состоятельные оценки для неизвестных параметров σ_ε^2 и σ_α^2 можно получить следующим образом:

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{\sum_{i=1}^{N} T_i - N} \sum_{i=1}^{N} \sum_{t=1}^{T} r_{it} \left(y_{it} - y_{i\bullet} - (X_{it} - X_{i\bullet})' \hat{\beta}_W \right)^2,$$

$$\hat{\sigma}_\alpha^2 = \frac{1}{N} \sum_{i=1}^{N} \left[\left(y_{i\bullet} - X_{i\bullet}' \hat{\beta}_B \right)^2 - \frac{1}{T_i} \hat{\sigma}_\varepsilon^2 \right].$$

10.4.2. Тестирование наличия смещения самоотбора

Однако предположение о независимой природе r_{it} может быть нереалистичным. Например, оценки регрессии, объясняющей доходность взаимных фондов,

часто страдают смещением из-за того, что часть фондов с низкой доходностью ликвидируется в процессе наблюдения. При изучении влияния уровня безработицы на величину индивидуальной заработной платы оценки могут быть смещены из-за того, что люди с относительно высокой заработной платой в случае повышения уровня безработицы с большей вероятностью покидают рынок труда.

Если r_{it} зависит от α_i и ε_{it}, смещение сомоотбора может отразиться на стандартных ошибках. Это означает, что распределение y при данных X в конкретной выборке и отличается от распределения y при данных X (которое, как правило, нас интересует). Для состоятельности оценок модели с детерминированным эффектом необходимо потребовать, чтобы

$$E\{(X_{it} - X_{i\bullet})\varepsilon_{it} \mid r_{i1},...,r_{iT}\} = 0.$$

Значит, если факт отсутствия наблюдения в выборке говорит нам нечто об ожидаемом значении случайной ошибки, связанной с X_{it}, то оценки будут несостоятельными. Однако отбор на основании ненаблюдаемых α_i не всегда ведет к несостоятельности оценок. Несостоятельности может и не быть, даже когда ε_{it} и r_{it} зависимы, лишь бы эта зависимость была инвариантной по времени.

Для состоятельности оценок модели со случайным эффектом необходимо выполнение условия

$$E\{X_{i\bullet}\alpha_i \mid r_{i1},...,r_{iT}\} = 0.$$

Если индивидуум с определенным значением ненаблюдаемой α_i с большой вероятностью будет отсутствовать в новой волне обследования, это приводит к смещению оценки модели со случайным эффектом, а если индивидуум с определенным значением шока ε_{it} с большой вероятностью будет отсутствовать в новой волне обследования, это приводит к несостоятельности оценок модели со случайным эффектом. Таким образом, оценки модели с детерминированным эффектом более робастны, чем оценки модели со случайным эффектом.

Еще одно важное наблюдение состоит в том, что оценки, полученные по несбалансированным панелям, меньше страдают от смещения отбора, чем оценки, полученные по сбалансированным подпанелям. Просто величина смещения будет иной.

Вербик и Нийман [18] предложили ряд простых тестов на предмет смещения самоотбора, основанных на изложенных выше соображениях.

Во-первых, поскольку условие состоятельности констатирует, что случайный член должен быть – в том или ином смысле – независимым от индикатора отбора, один из тестов может просто состоять во включении в модель каких-либо функций от $r_{i1},...,r_{iT}$ и проверки значимости этих включений. В качестве основной гипотезы будет выступать утверждение, что если некий индивидуум наблюдается все периоды с 1 до T, то это не дает никакой информации о его ненаблюдаемых характеристиках. Очевидно, что просто добавить в регрессию r_{it} нельзя, так как это ведет к мультиколлинеарности: у всех индивидуумов, попавших в выборку, $r_{it} = 1$. Вместо этого добавлять следует либо r_{it-1}, либо $c_i = \prod_{t=1}^{T} r_{it}$, либо $T_i = \sum_{t=1}^{T} r_{it}$.

Правда, это не подходит для сбалансированных подпанелей и работает только в модели со случайным эффектом. Поэтому, если основная гипотеза не отвергается, это еще не означает отсутствия смещения отбора из-за невысокой мощности теста.

Другая группа тестов основывается на идее сравнения четырех различных оценок, полученных по сбалансированной и несбалансированной панелям в моделях со случайным и детерминированным эффектами. Все они страдают от смещения отбора, но по-разному. Все оценки можно сравнивать попарно, но при этом иметь в виду, оценки моделей со случайным и детерминированным эффектами могут различаться не только из-за смещения отбора. Поэтому более естественно сравнивать оценки одноименных моделей, полученные в сбалансированном и несбалансированном случаях. Для сравнения удобно использовать статистику Хаусмана:

$$\xi_{RE} = \left(\hat{\beta}^{B}_{RE} - \hat{\beta}^{U}_{RE}\right)' \left[\hat{V}\left\{\hat{\beta}^{B}_{RE}\right\} - \hat{V}\left\{\hat{\beta}^{U}_{RE}\right\}\right]^{-1} \left(\hat{\beta}^{B}_{RE} - \hat{\beta}^{U}_{RE}\right),$$

где $\hat{V}$ означает оценку ковариационных матриц, а индексы B и U относятся к сбалансированной и несбалансированной панелям соответственно. Аналогично формулируется статистика для модели с детерминированными эффектами. Если верна основная гипотеза, тестовая статистика подчиняется χ^2-распределению с K степенями свободы. Но статистика может быть мала не только в случае справедливости основной гипотезы, но и в случае, когда оценки страдают от смещения отбора одинаково. Тест не способен различить эти ситуации. Его, так же как и обычный тест Хаусмана, можно проводить для подмножества коэффициентов.

10.4.3. Оценивание при наличии неслучайно пропущенных данных

Смещение самотбора является одной из разновидностей проблемы идентификации. Как следствие, невозможно состоятельно оценить параметры модели при наличии смещения отбора без дополнительных предположений. Рассмотрим в качестве иллюстрации пример, где индикатор отбора объясняется пробит-моделью со случайным эффектом:

$$r^{*}_{it} = z'_{it}\gamma + \xi_i + \eta_{it},$$

где $r_{it} = 1$, если $r^{*}_{it} > 0$ и $r_{it} = 0$ в противоположном случае, а z_{it} – это вектор экзогенных переменных, который включает X_{it}. Пусть модель, которую мы намереваемся оценивать, задана уравнением

$$y_{it} = X_{it}\beta + \alpha_i + \varepsilon_{it}.$$

Предположим, что случайные компоненты в обоих уравнениях имеют совместное нормальное распределение. Эта модель является обобщением модели Хекмана [9], построенной для случая пространственных выборок. Влияние принципа отбора отражается на математических ожиданиях ненаблюдаемых эффектов, условных по экзогенным переменным и индикаторам отбора:

$$E\{\alpha_i \mid z_{i1},...,z_{iT}, r_{i1},...,r_{iT}\} = 0,$$
$$E\{\varepsilon_{it} \mid z_{i1},...,z_{iT}, r_{i1},...,r_{iT}\} = 0.$$

Первое из приведенных выражений равно нулю, если $\mathrm{cov}\{\alpha_i, \xi_i\} = 0$, а если еще $\mathrm{cov}\{\varepsilon_{it}, \eta_{it}\} = 0$, то тогда оценки модели со случайным эффектом состоятельны. Может быть показано, что последнее выражение инвариантно по времени, если $\mathrm{cov}\{\varepsilon_{it}, \eta_{it}\} = 0$ или $z_{it}'\gamma$ инвариантно по времени. Это необходимые требования для состоятельности оценок модели с детерминированным эффектом.

Оценивание в более общем случае существенно затруднено. Хаусман и Вайс [8] рассмотрели случай панели из двух периодов, где истощение имело место только во втором периоде. В более общем случае одновременное оценивание двух уравнений требует двумерного численного интегрирования (по двум индивидуальным эффектам).

В настоящее время модель Хекмана на панельных данных реализована теоретически на основании непараметрического регрессионного анализа. Существуют также эмпирические работы, пока немногочисленные, где использованы эти методы. В частности, на кафедре математической экономики и эконометрики ГУ ВШЭ была защищена магистерская диссертация (Разин А.), посвященная исследованию предложения труда на панельных данных РМЭЗ, в которой был разработан программный модуль для панельного варианта модели Хекмана в среде STATA.

В скором времени будет издан учебник Марно Вербика в переводе на русский язык, где анализу панельных данных уделяется достаточно пристальное внимание. Избранные главы этого учебника печатаются в новом журнале «Прикладная эконометрика», и там содержится несколько более полная информация, связанная с проблемой смещения самоотбора.

* *
*

СПИСОК ЛИТЕРАТУРЫ

1. *Магнус Я.Р.*, *Катышев П.К.*, *Пересецкий А.А.* Эконометрика. Начальный курс: Учебник. 5-е изд., испр. М.: Дело, 2004.

2. *Amemiya T.* The Estimation of the Variances in a Variance-Components Model // International Economic Review. 1971. Vol. 12.

3. *Baltagi B.* Econometric Analysis of Panel Data. John Wiley & Sons, 1995.

4. *Biorn E.* Estimating Economic Relations From Incomplete Cross-Section/Time Series Data // Journal of Econometrics. 1981. Vol. 16.

5. *Chamberlain G.* Analysis of Covariance with Qualitative Data // Review of Economic Studies. 1980. Vol. 47.

6. *Deaton A.* Panel Data from Series of Cross-Sections // Journal of Econometrics. 1985. Vol. 30.

7. *Fuller W.A.*, *Battese G.E.* Estimation of Linear Models with Cross-Error Structure // Journal of Econometrics. 1974. Vol. 2.

8. *Hausman J.A.*, *Wise D.* Attrition Bias in Experimental and Panel Data: the Gary Income Maintenance Experiment // Journal of Econometrics. 1979. Vol. 47.

9. *Heckman J.J.* Sample Selection Bias as a Specification Error // Journal of Econometrics. 1979. Vol. 47.

10. *Heckman J.J., Robb R.* Alternative Models for Evaluating the Impact of Interventions: an Overview // Journal of Econometrics. 1985. Vol. 89.

11. *Henderson C.R.* Estimation of Variance Components // Biometrics. 1953. Vol. 9.

12. *Moffitt R.* Identification and Estimation of Dynamic Models with Time Series of Repeated Cross Sections. Brown University, Providence RI. Forthcoming in Journal of Econometrics. 1991. [Mimeo.]

13. *Nijman Th.E., Verbeek M., van Soest A.* The Efficiency of Rotating Panel Designs in Analysis of Variance Model // Journal of Econometrics. 1991. Vol. 49.

14. *Searl S.R.* Linear Models. N.Y.: Wiley, 1971.

15. *Swamy P.A.V.B., Arora S.S.* The Exact Finite Sample Properties of the Estimators of Coefficients in the Error Components Regression Models// Journal of Econometrics. 1972. Vol. 40.

16. *Townsend E.C., Searl S.R.* Best Quadratic Unbiased Estimation of Variance Components from Unbalanced Data in the One-Way Classification // Biometrics. 1971. Vol. 27.

17. *Verbeek M.* On the Estimation of a Fixed Effects Model with Selectivity Bias // Economics Letters. 1990. Vol. 34.

18. *Verbeek M., Nijman Th.E.* Testing for Selectivity Bias in Panel Data Models // International Economic Review. 1992. Vol. 33.

19. *Verbeek M.* A Guide to Modern Econometrics. John Wiley & Sons, 2003.

20. *Verbeek M., Nijman Th.E.* Can Cohort Data Be Treated As Genuine Panel Data? // Empirical Economics. 1992. Vol. 17.

21. Wallace T.D., Hussain A. The Use of Error Components Models in Combining Cross-Section and Time Series Data // Journal of Econometrics. 1969. Vol. 37.