Прикладной статистический анализ данных. 13. Анализ панельных данных.

Рябенко Евгений
riabenko.e@gmail.com

I/2015

Панельные данные

Панельные данные (panel/longitudinal/cross-sectional time series data) состоят из наблюдений над одними и теми же объектами в последовательные периоды времени.

Обычная регрессия:

$$y_i = \alpha + x_i^T \beta + \varepsilon_i, \quad i = 1, \ldots, N.$$

Панельная регрессия (базовый вариант):

$$y_{it} = \alpha + x_{it}^T \beta + u_{it}, \quad i = 1, \ldots, N, \; t = 1, \ldots, T.$$

О структуре u_{it} могут делаться дополнительные предположения.

Панельные данные

Пример (Магнус, Ben-Porah): предположим, что ежегодное исследование рынка труда показало, что 50% замужних женщин работает. Варианты интерпретации:

- каждая замужняя женщина имеет шанс 50% работать в течение года;
- 50% всех замужних женщин работают полный рабочий день, а остальные 50% вообще не работают.

Более адекватное представление можно получить, если проследить историю некоторого числа индивидуумов в течение определенного периода времени, т. е. в рамках панельных данных.

Виды моделей для панельных данных

1. Обычная линейная (объединённая) регрессия:
$$y_{it} = \alpha + \sum_{k=1}^{K} x_{itk}\beta_k + \varepsilon_{it}.$$
2. Свободный член варьируется по i:
$$y_{it} = \alpha_i + \sum_{k=1}^{K} x_{itk}\beta_k + \varepsilon_{it}.$$
3. Свободный член варьируется по i и t:
$$y_{it} = \alpha_{it} + \sum_{k=1}^{K} x_{itk}\beta_k + \varepsilon_{it}.$$
4. Все коэффициенты варьируются по i:
$$y_{it} = \alpha_i + \sum_{k=1}^{K} x_{itk}\beta_{ik} + \varepsilon_{it}.$$
5. Все коэффициенты варьируются по i и t:
$$y_{it} = \alpha_{it} + \sum_{k=1}^{K} x_{itk}\beta_{itk} + \varepsilon_{it}.$$

Виды моделей для панельных данных

Объединённая регрессия (1) предполагает, что все ошибки ε_{it} некоррелированы между собой как по i, так и по t, и некоррелированы со всеми объясняющими переменными x_{it}.

Наиболее полная модель (5) не может быть ни оценена, ни использована для предсказаний, поскольку на NT наблюдений параметров в ней приходится $NT\,(K+1) + (\text{число параметров распределения } \varepsilon_{it})$.

Модель объединённой регрессии (1)

Модель с неизменными коэффициентами (pooled model):

$$y_{it} = \alpha + \sum_{k=1}^{K} x_{itk}\beta_k + \varepsilon_{it}.$$

МНК-оценки параметров:

$$\bar{y} = \frac{1}{TN}\sum_{t=1}^{T}\sum_{i=1}^{N} y_{it}, \qquad \bar{x} = \frac{1}{TN}\sum_{t=1}^{T}\sum_{i=1}^{N} x_{it};$$

$$\hat{\beta} = T_{xx}^{-1}T_{xy}, \qquad \hat{\alpha} = \bar{y} - \hat{\beta}^T\bar{x},$$

$$T_{xx} = \sum_{t=1}^{T}\sum_{i=1}^{N}(x_{it}-\bar{x})(x_{it}-\bar{x})^T, \qquad T_{xy} = \sum_{t=1}^{T}\sum_{i=1}^{N}(x_{it}-\bar{x})(y_{it}-\bar{y}),$$

$$T_{yy} = \sum_{t=1}^{T}\sum_{i=1}^{N}(y_{it}-\bar{y})^2.$$

Остаточная сумма квадратов:

$$S_1 = T_{yy} - T_{xy}^T T_{xx}^{-1} T_{xy}.$$

Модели с переменным свободным членом (2), (3)

При построении регрессии часть предикторов может отсутствовать.
Среди невключённых факторов могут быть:

- индивидуальные постоянные (individual time-invariant) — не меняются со временем для каждого объекта (пол, качество управления фирмы, способности, социально-экономический фон);
- временные постоянные (period individual-invariant) — одинаковы для всех объектов в каждый фиксированный момент времени (макроэкономические показатели);
- индивидуальные нестационарные (individual time-varying) — варьируются в двух направлениях (доход фирмы, основной капитал, продажи).

Модели с переменным свободным членом предполагают, что эффект каждого невключённого индивидуального нестационарного признака в отдельности пренебрежимо мал, но в сумме значим, и имеет форму случайной величины, некоррелированной с остальными переменными. При этом эффекты остальных невключённых предикторов могут быть учтены с помощью константы.

Модель с однокомпонентной ошибкой (2)

Регрессионная модель с однокомпонентной ошибкой:

$$y_{it} = \alpha + x_{it}^T \beta + u_{it},$$
$$u_{it} = \mu_i + \nu_{it}.$$

μ_i — ненаблюдаемый эффект i-го объекта, ν_{it} — остаточная ошибка.

Пример:
y_{it} — заработок участника панели,
x_t — вектор, содержащий объективные измеримые показатели, влияющие на заработок (опыт работы, уровень образования, членство в профсоюзе),
μ_i — ненаблюдаемая мера предпринимательских способностей.

Модель (2) с фиксированным эффектом

$$y_{it} = \alpha + \mu_i + x_{it}^T \beta + \nu_{it}$$

Пусть μ_i — фиксированные параметры, а $\nu_{it} \sim N\left(0, \sigma_\nu^2\right)$
(мы рассматриваем фиксированное множество объектов и хотим делать выводы только о них).

Усредним уравнения по времени:

$$\bar{y}_{i\cdot} = \alpha + \bar{x}_{i\cdot}^T \beta + \mu_i + \bar{\nu}_{i\cdot}$$

и вычтем из исходного:

$$(y_{it} - \bar{y}_{i\cdot}) = (x_{it} - \bar{x}_{i\cdot})^T \beta + (\nu_{it} - \bar{\nu}_{i\cdot})$$

(внутренняя регрессия).

Отсюда получается оценка $\hat{\beta}_{Within}$.

Модель (2) с фиксированным эффектом

Можно оценить только $\alpha + \mu_i$; чтобы разделить эти слагаемые, нужно добавить ограничение на μ_i, например,

$$\sum_{i=1}^{N} \mu_i = 0.$$

Тогда

$$\hat{\alpha} = \bar{y}_{\cdot\cdot} - \bar{x}_{\cdot\cdot}^T \hat{\beta}_{Within},$$
$$\hat{\mu}_i = \bar{y}_{i\cdot} - \hat{\alpha} - \bar{x}_{i\cdot}^T \hat{\beta}_{Within}.$$

Недостатки модели:

- при больших N число параметров модели слишком велико, и получающиеся оценки неустойчивы;
- нельзя оценить эффект явно заданных индивидуальных постоянных — усреднение по времени их убивает;
- при фиксированном T и $N \to \infty$ состоятельна только оценка $\hat{\beta}_{Within}$.

Модель (2) со случайным эффектом

$$y_{it} = \alpha + \mu_i + x_{it}^T \beta + \nu_{it}$$

Пусть $\mu_i \sim N\left(0, \sigma_\mu^2\right)$ и $\nu_{it} \sim N\left(0, \sigma_\nu^2\right)$
(мы рассматриваем случайно выбранное множество объектов и хотим делать выводы о совокупности, из которой они извлечены).

$\hat{\beta}$ строится обобщённым методом наименьших квадратов; она является взвешенным средним оценок, получаемых из внутренней регрессии и внешней регрессии:

$$\bar{y}_{i\cdot} = \alpha + \bar{x}_{i\cdot}^T \beta + \bar{u}_{i\cdot}.$$

$$\hat{\beta}_{GLS} = W \hat{\beta}_{Within} + (I - W) \hat{\beta}_{Between},$$

W — некоторая вычислимая в явном виде матрица.

Случайный или фиксированный эффект?

При больших T различия между двумя моделями пропадают, но при маленьких T и больших N они могут быть значительными.

Пример (Hausman, 1978): y — доход выпускников школ в Мичигане, $N = 629$, $T = 6$; в модели с фиксированным эффектом жители сельской местности получают на 1% меньше, доверительный интервал $(-7\%, 5\%)$, а в модели со случайным эффектом — на 12% меньше, доверительный интервал $(-17\%, -7\%)$.

Если объекты случайно выбраны из популяции и нужно сделать вывод только о ней, используется модель со случайным эффектом. Если же они обладают выраженными индивидуальными различиями, которые нельзя объяснить случайными факторами, используется модель с фиксированным эффектом.

Модель с двухкомпонентной ошибкой (3)

Регрессионная модель с двухкомпонентной ошибкой:

$$y_{it} = \alpha + x_{it}^T \beta + u_{it},$$
$$u_{it} = \mu_i + \lambda_t + \nu_{it}.$$

μ_i — ненаблюдаемый эффект i-го объекта, λ_t — ненаблюдаемый временной эффект, ν_{it} — остаточная ошибка.

Пример:
y_{it} — заработок участника панели,
x_t — вектор, содержащий объективные измеримые показатели, влияющие на заработок (опыт работы, уровень образования, членство в профсоюзе),
μ_i — ненаблюдаемая мера предпринимательских способностей,
λ_t — интегральный показатель состояния экономики.

Модель (3) с фиксированными эффектами

$$y_{it} = \alpha + \mu_i + \lambda_t + x_{it}^T \beta + \nu_{it}$$

С ограничениями $\sum_i \mu_i = 0$ и $\sum_t \lambda_t = 0$ можно усреднить по i и t и получить внутреннюю регрессию:

$$(y_{it} - \bar{y}_{i\cdot} - \bar{y}_{\cdot t} + \bar{y}_{\cdot\cdot}) = (x_{it} - \bar{x}_{i\cdot} - \bar{x}_{\cdot t} + \bar{x}_{\cdot\cdot})^t \beta + (\nu_{it} - \bar{\nu}_{i\cdot} - \bar{\nu}_{\cdot t} + \bar{\nu}_{\cdot\cdot}),$$

откуда получается оценка $\hat{\beta}_{Within}$; по аналогии,

$$\begin{aligned}
\hat{\alpha} &= \bar{y}_{\cdot\cdot} - \bar{x}_{\cdot\cdot}^T \hat{\beta}_{Within}, \\
\hat{\mu}_i &= (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}) - (\bar{x}_{i\cdot} - \bar{x}_{\cdot\cdot})^T \hat{\beta}_{Within}, \\
\hat{\lambda}_t &= (\bar{y}_{\cdot t} - \bar{y}_{\cdot\cdot}) - (\bar{x}_{\cdot t} - \bar{x}_{\cdot\cdot})^T \hat{\beta}_{Within}.
\end{aligned}$$

Недостатки модели:

- нельзя оценить эффект явно заданных индивидуальных и временных постоянных;
- при фиксированном T и $N \to \infty$ состоятельна только оценка $\hat{\beta}$.

Модель (3) со случайными эффектами

$$y = \alpha + \mu_i + \lambda_t + x_{it}^T\beta + \nu_{it}$$

Пусть $\mu_i \sim N\left(0, \sigma_\mu^2\right)$, $\lambda_t \sim N\left(0, \sigma_\lambda^2\right)$ и $\nu_{it} \sim N\left(0, \sigma_\nu^2\right)$
(мы рассматриваем случайно выбранное множество объектов и моментов времени и хотим делать выводы о совокупности, из которой они извлечены).

Решение обобщённым методом наименьших квадратов даёт взвешенное среднее внутренней регрессии и двух внешних регрессий:

$$\bar{y}_{i\cdot} = \alpha + \bar{x}_{i\cdot}^T\beta + \bar{u}_{i\cdot},$$
$$\bar{y}_{\cdot t} = \alpha + \bar{x}_{\cdot t}^T\beta + \bar{u}_{\cdot t}.$$

$$\hat{\beta}_{GLS} = W_1\hat{\beta}_{Within} + W_2\hat{\beta}_{Between\ i} + W_3\hat{\beta}_{Between\ t},$$

W_1, W_2, W_3 — некоторые вычислимые в явном виде матрицы.

Полная статическая модель (4)

Наиболее полная статическая модель (unrestricted static model):

$$y_{it} = \alpha_i + \sum_{k=1}^{K} x_{itk}\beta_{ik} + \varepsilon_{it}.$$

МНК-оценки параметров (within-group estimates):

$$\bar{y}_i = \frac{1}{T}\sum_{t=1}^{T} y_{it}, \qquad \bar{x}_i = \frac{1}{T}\sum_{t=1}^{T} x_{it};$$

$$\hat{\beta}_i = W_{xx,i}^{-1} W_{xy,i}, \qquad \hat{\alpha}_i = \bar{y}_i - \hat{\beta}_i^T \bar{x}_i,$$

$$W_{xx,i} = \sum_{t=1}^{T} (x_{it} - \bar{x}_i)(x_{it} - \bar{x}_i)^T, \quad W_{xy,i} = \sum_{t=1}^{T} (x_{it} - \bar{x}_i)(y_{it} - \bar{y}_i),$$

$$W_{yy,i} = \sum_{t=1}^{T} (y_{it} - \bar{y}_i)^2, \qquad RSS_i = W_{yy,i} - W_{xy,i}^T W_{xx,i}^{-1} W_{xy,i}.$$

Остаточная сумма квадратов:

$$S_4 = \sum_{i=1}^{N} RSS_i.$$

Сравнение моделей

Как проверить возможность упрощения модели (4)?

H_1: ни свободный член, ни угловые коэффициенты не зависят от i, верна модель (1) $\Leftrightarrow \beta_1 = \cdots = \beta_N,\ \alpha_1 = \cdots = \alpha_N$.
H_2: угловые коэффициенты не зависят от i, верна модель (2) $\Leftrightarrow \beta_1 = \cdots = \beta_N$.

Пусть S_1 и S_2 — остаточные суммы квадратов для моделей (1) и (2). Если ε_{it} — независимые по i и t нормально распределённые с нулевым средним и общей дисперсией σ^2, для проверки H_0 можно использовать критерий Фишера.

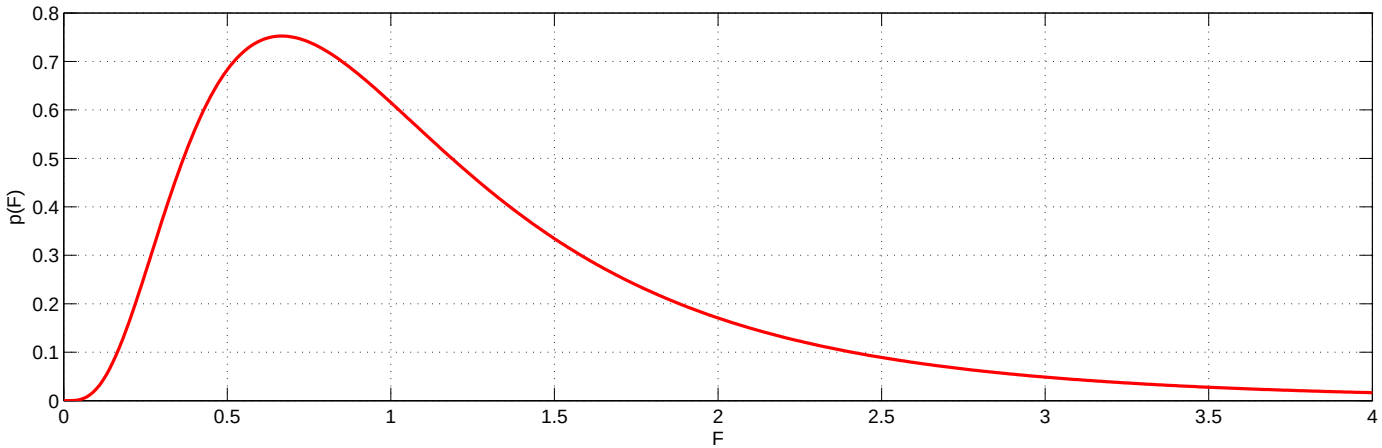
Критерий Фишера, (4) против (1)

нулевая гипотеза: $H_1 : \beta_1 = \cdots = \beta_N,\ \alpha_1 = \cdots = \alpha_N;$

альтернатива: $H_1' : H_1$ неверна;

статистика: $F_1 = \frac{(S_1 - S_4)/((N-1)(K+1))}{S_4/(NT - N(K+1))};$

$F_1 \sim F((N-1)(K+1), N(T-K-1))$ при H_1.



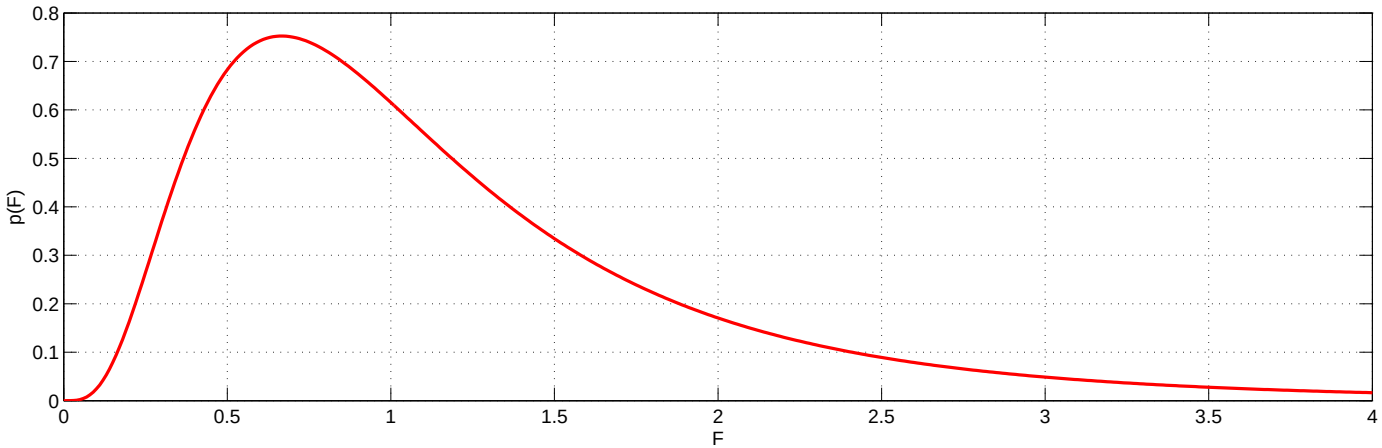
Критерий Фишера, (4) против (2)

нулевая гипотеза: $H_2\colon \beta_1 = \cdots = \beta_N$;

альтернатива: $H_2'\colon H_2$ неверна;

статистика: $F_2 = \frac{(S_4 - S_2)/(N-1)K}{S_4/(NT-N(K+1))}$;

$F_2 \sim F((N-1)K, N(T-K-1))$ при H_2.



Условный критерий Фишера, (2) против (1)

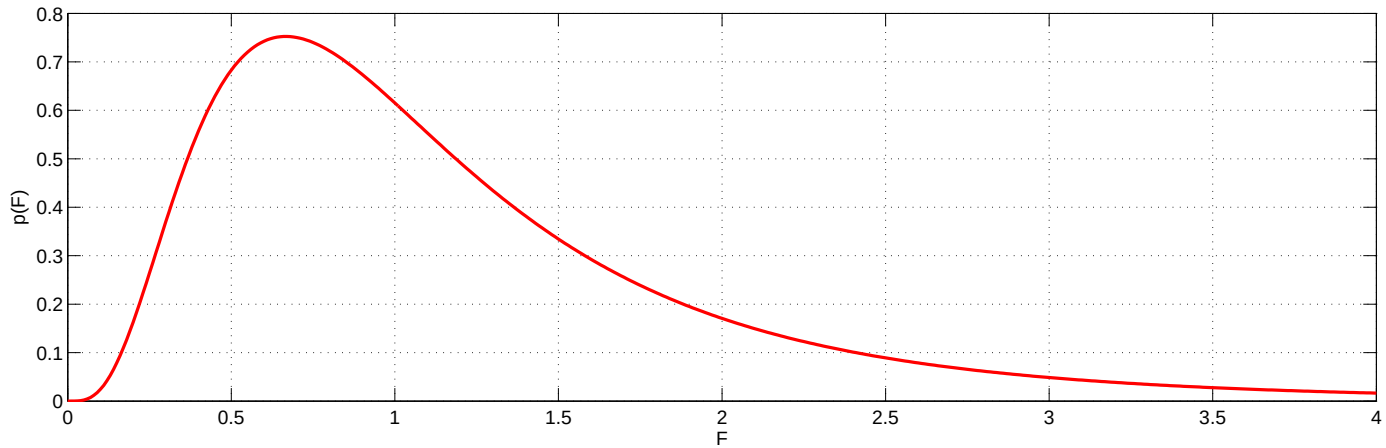
Если H_2 принята, можно воспользоваться условным критерием:

нулевая гипотеза: $H_3\colon \alpha_1 = \cdots = \alpha_N,\ \beta_1 = \cdots = \beta_N;$

альтернатива: $H_3'\colon H_3$ неверна, но $\beta_1 = \cdots = \beta_N;$

статистика: $F_3 = \frac{(S_1 - S_2)/(N-1)}{S_2/(N(T-1)-K)};$

$F_3 \sim F(N-1, N(T-1)-K)$ при H_3.



Совместное использование критериев

Критерии могут давать противоречивые результаты:

- может получиться, что F_1 отвергает общую гипотезу однородности всех коэффициентов, а F_2 и F_3 не отвергают своих нулевых гипотез (то есть, присутствует неоднородность коэффициентов, но мы не можем сказать, в α или β);
- может получиться, что F_1 не отвергает общую гипотезу однородности всех коэффициентов, а F_2 или F_3 отвергают свою нулевую гипотезу.

Динамические модели без индивидуального эффекта

Аналогичным образом можно настроить модель с учётом времени, предполагая отсутствие различий между объектами:

$$y_{it} = \alpha_t + \sum_{k=1}^{K} x_{itk}\beta_{tk} + \varepsilon_{it}$$

и проверить гипотезы о сводимости её к упрощённым моделям:

$$y_{it} = \alpha_t + \sum_{k=1}^{K} x_{itk}\beta_{k} + \varepsilon_{it}$$

и

$$y_{it} = \alpha + \sum_{k=1}^{K} x_{itk}\beta_{k} + \varepsilon_{it}.$$

Литература

- наиболее полная версия — Baltagi;
- кратко и по-русски — Магнус;
- использование в R — Croissant.

Baltagi B.H. *Econometric analysis of panel data*. — Chichester: John Wiley & Sons, 2005.
Croissant Y., Millo G. (2008). *Panel data econometrics in R: The plm package*. Journal of Statistical Software, 27(2), 1–43.
Магнус Я.Р., Катышев П.К., Пересецкий А.А. *Эконометрика. Начальный курс*. Москва: Дело, 2005.