Регрессионный анализ: панельные данные и каузальность

Лекция 1
Модели с фиксированными эффектами и модель со случайными эффектами: введение

Структуры данных

- Пространственные данные:
  в таких данных представлены наблюдения нескольких (N) пространственных единиц анализа за один и тот же временной период

Структуры данных

- Пространственные данные:
  в таких данных представлены наблюдения нескольких (N) пространственных единиц анализа за один и тот же временной период
- Временные ряды:
  данные одной пространственной единицы анализа представлены за несколько (T) временных периодов

Структуры данных

- Пространственные данные:
  в таких данных представлены наблюдения нескольких (N) пространственных единиц анализа за один и тот же временной период
- Временные ряды:
  данные одной пространственной единицы анализа представлены за несколько (T) временных периодов
- Панельные данные:
  данные нескольких (N) пространственных единиц представлены за несколько (T) временных периодов (есть и пространственная, и временная перспектива)

Пространственные данные: пример

id	age	spending_score	city	is_employed
1	37	42	London	True
2	51	78	Paris	False
3	22	15	Moscow	True

Временные ряды: пример

Quarter	GDP Growth (%)	Inflation (%)	Interest Rate (%)
2023 Q1	2.1	4.2	1.5
2023 Q2	1.8	4.5	1.8
2023 Q3	2.3	4.8	2.0
2023 Q4	2.0	5.2	2.2

Панельные данные: пример

Company	Year	Revenue ($K)	Profit Margin (%)	Country
1	2021	987.23	18.34	USA
1	2022	1054.67	12.56	USA
1	2023	912.45	21.78	USA
2	2021	1201.89	8.91	Germany
2	2022	1156.34	14.23	Germany
2	2023	1245.67	9.45	Germany
3	2021	845.12	19.67	Japan
3	2022	901.45	16.89	Japan
3	2023	878.90	22.34	Japan

TSCS-данные

Кроме того, в англоязычной литературе Вы можете встретить такое понятие, как Time-Series Cross-Sectional Data. Здесь имеет место

- специфика обозначения панельных данных в разных дисциплинах
- длительность временного периода: как правило, TSCS-данными называют данные с относительно небольшим числом пространственных единиц и продолжительным временным рядом. Здесь в большей степени фокусируемся на таких сюжетах, как (не)стационарность данных и моделирование автокорреляции в явном виде

Последствия игнорирования панельной структуры данных

Вопрос

К чему приводит оценивание «объединенной» (pooled – без поправок на подгруппы) применительно к панельным данным?

Последствия игнорирования панельной структуры данных

Вопрос

К чему приводит оценивание «объединенной» (pooled – без поправок на подгруппы) применительно к панельным данным?

Ответ

- смещенные оценки

Последствия игнорирования панельной структуры данных

Вопрос

К чему приводит оценивание «объединенной» (pooled – без поправок на подгруппы) применительно к панельным данным?

Ответ

- смещенные оценки
- некорректная значимость оценок: заниженные стандартные ошибки

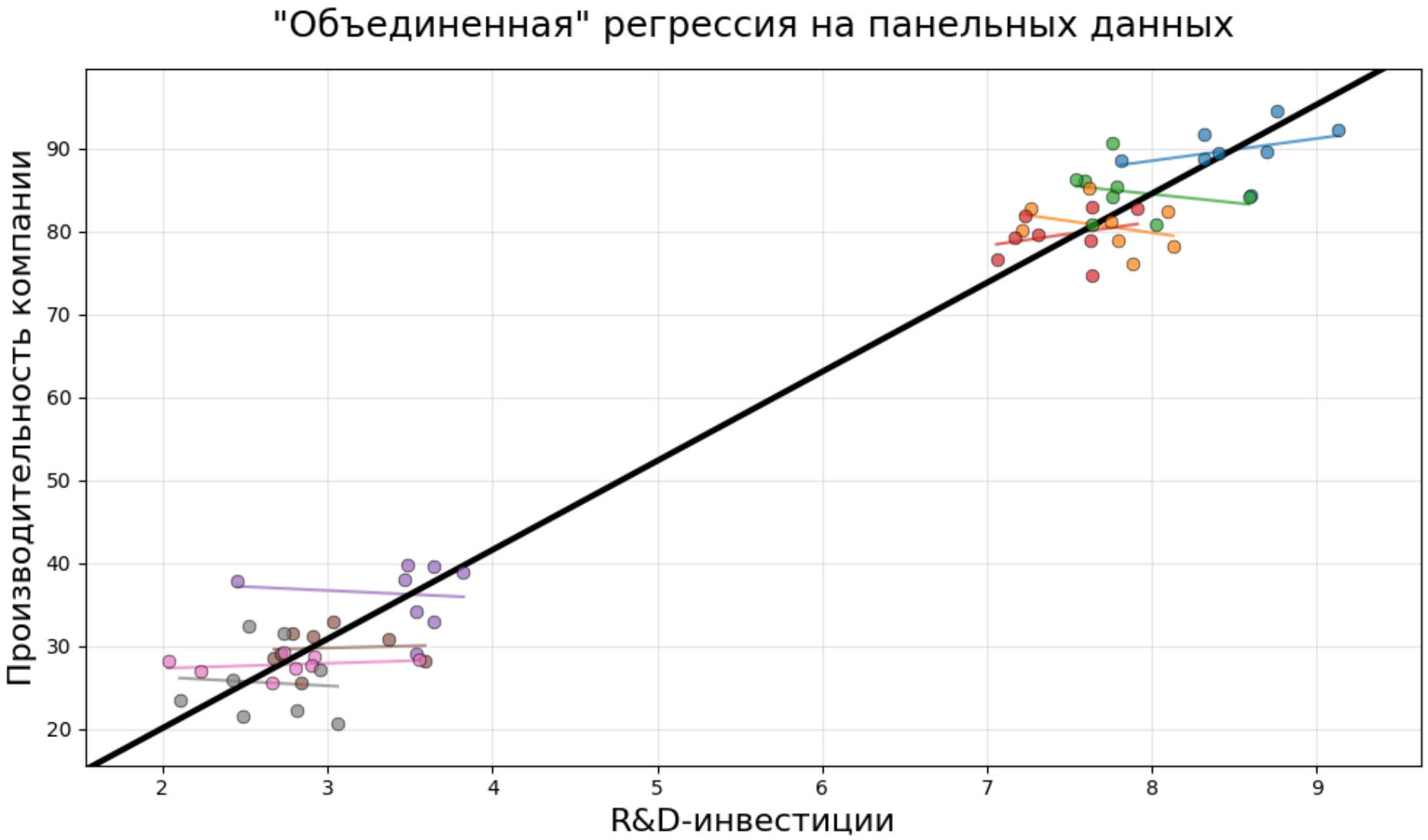
Последствия игнорирования панельной структуры данных

Вопрос

К чему приводит оценивание «объединенной» (pooled – без поправок на подгруппы) применительно к панельным данным?

Ответ

- смещенные оценки
- некорректная значимость оценок: заниженные стандартные ошибки
- не разделяем внутригрупповую (изменения во времени) и межгрупповую дисперсию (между пространственными единицами)


"Объединенная" регрессия на панельных данных
Производительность компании
R&D-инвестиции
20
30
40
50
60
70
80
90
2
3
4
5
6
7
8
9

Альтернативы*:

- Модели с фиксированными эффектами
  1. на пространственные единицы
  2. на временные периоды
  3. и на пространственные единицы, и на временные периоды
- Модель со случайными эффектами на пространственные единицы

* Список альтернатив будет постепенно пополняться по мере изучения материала

Англоязычные аббревиатуры для удобства

- FE-модель – модель с фиксированными эффектами (fixed-effects model)
- RE-модель – модель со случайными эффектами (random-effects model)
- LSDV-модель – модель с фиксированными эффектами в формате набора дамми-переменных (least-squares dummy-variables model)
- Within- – модель / оценки модели с внутригрупповым преобразованием

FE-модель на пространственные единицы

Представим спецификацию модели с фиксированными эффектами в формате модели с набором дамми-переменных (LSDV)

FE-модель на пространственные единицы

Представим спецификацию модели с фиксированными эффектами в формате модели с набором дамми-переменных (LSDV)

$$\hat{y}_{it} = \hat{\beta}_0 + \hat{\gamma}_1 D_{1i} + \ldots \hat{\gamma}_{n-1} D_{(n-1)i} + \hat{\beta}_1 x_{it}$$

FE-модель на пространственные единицы

Представим спецификацию модели с фиксированными эффектами в формате модели с набором дамми-переменных (LSDV)

$$\hat{y}_{it} = \hat{\beta}_0 + \hat{\gamma}_1 D_{1i} + \ldots \hat{\gamma}_{n-1} D_{(n-1)i} + \hat{\beta}_1 x_{it}$$

- $\hat{\beta}_0$ – чему в среднем равно значение зависимой переменной в базовой категории при равенстве предикторов 0

FE-модель на пространственные единицы

Представим спецификацию модели с фиксированными эффектами в формате модели с набором дамми-переменных (LSDV)

$\hat{y}_{it} = \hat{\beta}_0 + \hat{\gamma}_1 D_{1i} + ... \hat{\gamma}_{n-1} D_{(n-1)i} + \hat{\beta}_1 x_{it}$

- $\hat{\beta}_0$ – чему в среднем равно значение зависимой переменной в базовой категории при равенстве предикторов 0
- $\hat{\gamma}_i$ – на сколько в среднем отличается значение зависимой переменной в i-ой пространственной единице от базовой категории при прочих равных
- $\hat{\beta}_0 + \hat{\gamma}_i$ – индивидуальная константа (фиксированный эффект)

FE на пространственные единицы

Содержательно охватывают весь набор неизменяющихся во времени характеристик пространственных единиц

FE на пространственные единицы

Содержательно охватывают весь набор неизменяющихся во времени характеристик пространственных единиц

Примеры:

- географические характеристики (географическое положение, расстояние между пространственными единицами, климатическая зона и т.д.)
- членство в ЕС
- правовая система
- культурные нормы

FE на пространственные единицы

Содержательно охватывают весь набор неизменяющихся во времени характеристик пространственных единиц

Примеры:

- географические характеристики (географическое положение, расстояние между пространственными единицами, климатическая зона и т.д.)
- членство в ЕС
- правовая система
- культурные нормы

То есть, включение фиксированных эффектов не позволяет полностью избавиться от эндогенности, так как мы можем пропустить существенные изменяющиеся во времени характеристики (то есть, $Cov(e_{it}, x_{it}) \neq 0$)

Модель с внутригрупповым преобразованием

Мы можем представить FE-модель в более экономном виде (то есть, с меньшим количеством параметров), при этом сохранив скорректированную на панельную структуру данных оценку коэффициента наклона
Алгоритм:

1. рассчитываем центрированный y, при этом считаем средние по пространственным подгруппам: $y^*_{it} = y_{it} - \overline{y}_i$.
2. аналогичным образом преобразуем предикторы: $x^*_{it} = x_{it} - \overline{x}_i$.
3. оцениваем регрессию y^*_{it} на x^*_{it}:

$$\hat{y}^*_{it} = \hat{\beta}_1 x^*_{it}$$

Куда исчезла константа?

$$\hat{y}^*_{it} = \hat{\beta}_0 + \hat{\beta}_1 x^*_{it}$$

Вспоминаем, что для парной модели будет справедливо, что $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$

В данном случае:

$$\hat{\beta}_0 = \overline{y^*} - \hat{\beta}_1 \overline{x^*} = \frac{\sum_{i=1}^{N} (y_{it} - \overline{y_{i\cdot}})}{N} - \hat{\beta}_1 \frac{\sum_{i=1}^{N} (x_{it} - \overline{x_{i\cdot}})}{N} = 0$$

Как рассчитывается $\hat{\beta}_1$?

Оценку коэффициента наклона в FE-модели можно получить на основании соответствующих коэффициентов регрессий, оцененных на отдельных N подвыборках

Нас интересует оценка коэффициента при предикторе

$$\hat{y}^*_{it} = \hat{\beta}_1 x^*_{it}$$

1. Для каждой из N подвыборок оценим регрессию $\hat{y}_t = \hat{a}_0 + \hat{a}_1 x_t$ и сохраним $\hat{a}_1$ для каждой i-ой единицы
2. Суммируем взвешенные значения $\hat{a}_1$:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N} w_i \hat{\alpha}_{1i}}{\sum_{i=1}^{N} w_i}$$

$$w_i = \sum_{t=1}^{T_i} (x_{it} - \bar{x}_{i\cdot})^2$$

Выведение формулы $\hat{\beta}_1$ для FE-модели

$$\hat{\beta}_1 = \frac{\widehat{Cov}(x^*, y^*)}{\widehat{Var}(x^*)} = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it} - \bar{x}_{i\cdot})(y_{it} - \bar{y}_{i\cdot})}{\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it} - \bar{x}_{i\cdot})^2}$$

Домножим и разделим дробь в числителе на $\sum_{t=1}^{T}(x_{it} - \bar{x}_{i\cdot})^2$:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N}\left(\sum_{t=1}^{T}(x_{it} - \bar{x}_{i\cdot})^2 \cdot \frac{\sum_{t=1}^{T}(x_{it} - \bar{x}_{i\cdot})(y_{it} - \bar{y}_{i\cdot})}{\sum_{t=1}^{T}(x_{it} - \bar{x}_{i\cdot})^2}\right)}{\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it} - \bar{x}_{i\cdot})^2} = \frac{\sum_{i=1}^{N} w_i \hat{\alpha}_{1i}}{\sum_{i=1}^{N} w_i}$$

Алгоритм получения $\hat{\beta}_1$ в FE-модели при наличии контрольных переменных

1. Очистим y_{it} от эффекта контрольных переменных z_{it}. Для этого нужно сохранить остатки регрессии y_{it} на z_{it} (в этой модели учитываем FE)
2. По такому же принципу очищаем x_{it} от эффекта z_{it}
3. Далее повторяем уже знакомую процедуру, однако вместо y_{it} и x_{it} используем сохраненные остатки (очищенный y_{it} и очищенный x_{it})

Выводы по формуле $\hat{\beta}_1$ в FE-модели на пространственные единицы:

- Пространственные единицы с большей внутригрупповой изменчивостью предиктора (то есть, изменчивостью предиктора во времени) вносят больший вклад в оценку $\hat{\beta}_1$
- Те единицы анализа, у которых предиктор вообще не изменяется во времени, не участвуют в формировании оценки коэффициента $\hat{\beta}_1$ в FE-модели на пространственные единицы
- $\hat{\beta}_1$ формируется на основании внутригрупповой изменчивости предиктора. Поэтому интерпретация $\hat{\beta}_1$ будет в общем виде следующей: При увеличении предиктора на одну единицу, зависимая переменная в среднем растет на $\hat{\beta}_1$ для объектов внутри одной пространственной единицы при прочих равных условиях

Модель со случайными эффектами

Для моделирования различий в стартовых условиях можно воспользоваться моделью со случайными эффектами (RE-модель). Здесь различия в стартовых условиях представлены как случайная величина:

$$y_{it} = \beta_0 + \beta_1 x_{it} + \alpha_i + e_{it}$$

α_i – межгрупповая изменчивость (отклонения группового среднего от общего среднего)
e_{it} – внутригрупповая изменчивость (отклонения наблюдений за t-ый временной период от группового среднего)

Допущения RE-модели

- $Cov(\alpha_i, e_{it}) = 0$
- $Cov(x_{it}, e_{it}) = 0$
- $Cov(x_{it}, \alpha_i) = 0$
- $Cov(e_{it}, e_{is}) = 0$, при этом $t \neq s$
- $Cov(\alpha_i, \alpha_j) = 0$, при этом $i \neq j$
- $e_{it} \sim i.i.d.(0, \sigma_e^2)$; $\alpha_i \sim i.i.d.(0, \sigma_\alpha^2)$

Тест Хаусмана

$$H_0 : Cov(x_{it}, \alpha_i) = 0$$

$$\beta_{FE} = \beta_{RE}$$

$$H_1 : Cov(x_{it}, \alpha_i) \neq 0$$

$$\beta_{FE} \neq \beta_{RE}$$

Статистика критерия:

$$H = (\hat{\beta}_{FE} - \hat{\beta}_{RE})'[\widehat{Var}(\hat{\beta}_{FE}) - \widehat{Var}(\hat{\beta}_{RE})]^{-1}(\hat{\beta}_{FE} - \hat{\beta}_{RE}) \sim \chi^2_{df=k}$$

k – количество изменяющихся во времени предикторов

Если H_0 отвергается, значит модель с фиксированными эффектами более предпочтительна. Оценки модели со случайными эффектами оказываются смещенными и несостоятельными