

# Эндогенность

Эконометрика. Openedu. Неделя 13

# Разложение в сумму неоднозначно

$$4 = 3 + 1$$

$$4 = 2 + 2$$

# Несколько верных форм одной модели

Модель А:

$$y_i = 2x_i + \varepsilon_i$$

Модель Б:

$$y_i = 3x_i + u_i$$

Модели А и Б эквивалентны, если  $\varepsilon_i = x_i + u_i$

Если:

модель представлена в форме

$$y_i = \beta_1 + \beta_2 x_i + \beta_3 z_i + \varepsilon_i$$

где  $E(\varepsilon_i|X) = 0$  и [другие предпосылки]

То:

Оценки МНК состоятельны

$$\hat{\beta} \rightarrow \beta$$

и несмещены

$$E(\hat{\beta}|X) = \beta, \quad E(\hat{\beta}) = \beta$$

## Смысл предпосылки $E(\varepsilon_i|X) = 0$

Среднее значение  $\varepsilon_i$  не зависит от значений объясняющих переменных и равно нулю.

В частности,

$$E(\varepsilon_i|X) = 0 \Rightarrow \begin{cases} E(\varepsilon_i) = 0 \\ Cov(x_i, \varepsilon_i) = 0 \end{cases}$$

# Последствия нарушения предпосылки

Если  $Cov(x_i, \varepsilon_i) \neq 0$ , то оценки МНК несостоятельны:

$$\hat{\beta} \nrightarrow \beta$$

и смещены

$$E(\hat{\beta}|X) \neq \beta, E(\hat{\beta}) \neq \beta$$

## Пример у неоновой доски

$$y_i = 2 + 3x_i + \varepsilon_i$$

где  $Var(x_i) = 4$ ,  $Var(\varepsilon_i) = 3$ ,  $Cov(x_i, \varepsilon_i) = -2$

Оцениваем параметр  $\beta_2$  с помощью МНК, получаем  $\hat{\beta}_2$ .

Найдите  $\text{plim } \hat{\beta}_2$  (предел по вероятности)



Выборочная ковариация

$$sCov(x, y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

Выборочная дисперсия

$$sVar(x) = \frac{\sum (x_i - \bar{x})^2}{n - 1}$$

Следствие закона больших чисел:

Если выборка  $(x_i, y_i)$  случайна, то

$$\text{plim}_{n \rightarrow \infty} sCov(x, y) = Cov(x_i, y_i)$$

$$\text{plim}_{n \rightarrow \infty} sVar(x) = Var(x_i)$$

Коррелированность регрессоров и случайных ошибок,  $Cov(x_i, \varepsilon_i) \neq 0$ , называется эндогенностью

# Зачем возиться с эндогенностью?

У любой модели есть форма записи, в которой  $E(\varepsilon_i|X) = 0$ .  
Зачем нужны те формы записи, в которых  $E(\varepsilon_i|X) \neq 0$ ?

- Если модель используется для прогнозирования, то формы записи с эндогенностью возможно не нужны.
- В некоторых случаях форма записи с эндогенностью легче интерпретируется

# Некоторые причины эндогенности в перекрёстных выборках

- Ошибка измерения регрессора
- Пропущенный регрессор
- Одновременность определения значения переменных

# Ошибка измерения регрессора. Исходная форма модели.

Модель в форме А:

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$$

и  $\text{Cov}(x_i, \varepsilon_i) = 0$ .

Наблюдаем  $y_i$  и  $x_i^* = x_i + u_i$ , где  $u_i$ , ошибка измерения регрессора  $x_i$ , не зависит от  $x_i$  и  $\varepsilon_i$

# Ошибка измерения регрессора. Вывод другой формы модели.

Подставим  $x_i = x_i^* - u_i$  в форму А и получим:

$$y_i = \beta_1 + \beta_2(x_i^* - u_i) + \varepsilon_i$$

и модель в форме Б:

$$y_i = \beta_1 + \beta_2 x_i^* + w_i, \quad w_i = \varepsilon_i - \beta_2 u_i$$



## Эндогенность в форме Б:

$$y_i = \beta_1 + \beta_2 x_i^* + w_i, \quad w_i = \varepsilon_i - \beta_2 u_i$$

В форме Б:

$$\text{Cov}(x_i^*, w_i) = \text{Cov}(x_i + u_i, \varepsilon_i - \beta_2 u_i) = -\beta_2 \text{Var}(u_i) \neq 0$$

МНК оценки для формы Б несостоятельны

## Пример у неоновой доски

$$y_i = 2 + 3x_i + \varepsilon_i$$

Регрессор  $x_i$  ненаблюдаем

Наблюдаем  $x_i^* = x_i + u_i$ ,  $Var(x_i) = 9$ ,  $Var(u_i) = 4$ ,  $Var(\varepsilon_i) = 1$ .

К чему стремится МНК оценка модели  $\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i^*$ ?

Модель с ошибкой измерения регрессора:

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i, \text{ где наблюдаем } x_i^* = x_i + u_i$$

- Хотим оценить  $\beta_2$ , т.е. на сколько растёт  $y_i$  при росте настоящего  $x_i$  на единицу

# Мораль. МНК для нашей цели не состоятелен.

При МНК оценивании регрессии

$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i^*$$

получаем оценку  $\hat{\beta}_2$  несостоятельную для  $\beta_2$

- МНК оценивает на сколько растёт  $y_i$  при росте наблюдаемого  $x_i^*$  (включающего ошибку) на единицу

Хотим оценить форму записи А:

$$y_i = \beta_1 + \beta_2 x_i + \beta_3 d_i + \varepsilon_i$$

где  $\text{Cov}(x_i, d_i) \neq 0$ ,  $\text{Cov}(x_i, \varepsilon_i) = 0$ ,  $\text{Cov}(d_i, \varepsilon_i) = 0$ .

Не наблюдаем  $d_i$ .

# Пропущенная объясняющая переменная.

Форма записи Б:

$$y_i = \beta_1 + \beta_2 x_i + u_i \quad u_i = \beta_3 d_i + \varepsilon_i$$

Эндогенность:

$$\text{Cov}(x_i, u_i) = \text{Cov}(x_i, \beta_3 d_i + \varepsilon_i) = \beta_3 \text{Cov}(x_i, d_i)$$

## Пример у неоновой доски

$$y_i = 2 + 3x_i - 2d_i + \varepsilon_i$$

Регрессор  $d_i$  ненаблюдаем.

$$\text{Var}(x_i) = \text{Var}(d_i) = 9, \text{Var}(\varepsilon_i) = 1, \text{Cov}(x_i, d_i) = -6.$$

К чему стремится МНК оценка модели  $\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i$ ?

Модель пропущенным регрессором:

$$y_i = \beta_1 + \beta_2 x_i + \beta_3 d_i + \varepsilon_i, \text{ регрессор } d_i \text{ не наблюдаем}$$

- Хотим оценить  $\beta_2$ , т.е. на сколько растёт  $y_i$  при росте  $x_i$  на единицу и фиксированном  $d_i$



# Мораль. МНК для нашей цели не состоятелен.

При МНК оценивании регрессии

$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i$$

получаем оценку  $\hat{\beta}_2$  несостоятельную для  $\beta_2$

- МНК оценивает на сколько растёт  $y_i$  при росте  $x_i$  на единицу (и сопряженных с этим изменений в  $d_i$ )

## Пример у неоновой доски

Равновесная цена и объем продаж определяются из системы:

$$\begin{cases} q_i = 3 - p_i + \varepsilon_i, & \text{логарифм спроса} \\ q_i = 3 + 2p_i + u_i, & \text{логарифм предложения} \end{cases}$$

Ошибки  $u_i$  и  $\varepsilon_i$  независимы и нормальны  $N(0, 1)$

К чему стремится оценка коэффициента при цене при оценке уравнения спроса с помощью МНК?

Хотим состоятельно оценить  $\beta_2$  в форме записи:

$$y_i = \beta_1 + \beta_2 x_i + \beta_3 d_i + \varepsilon_i, \text{Cov}(x_i, \varepsilon_i) \neq 0$$

Возможный выход: найти “инструментальные переменные”  $z_i$ :

- $\text{Cov}(z_i, \varepsilon_i) = 0$
- $\text{Cov}(z_i, x_i) \neq 0$

# Как использовать инструментальные переменные?

Модель:

$$y_i = \beta_1 + \beta_2 x_i + \beta_3 d_i + \varepsilon_i$$

где  $Cov(x_i, \varepsilon_i) \neq 0$  и  $Cov(d_i, \varepsilon_i) = 0$ .

- нельзя просто заменить проблемный регрессор на инструментальную переменную

## Двухшаговый МНК:

Шаг 1. Построить регрессию каждого  $x_i$  коррелированного с  $\varepsilon_i$  на инструментальные переменные. Получить прогнозы  $\hat{x}_i$ .

Шаг 2. Оценить исходную модель, заменив  $x_i$  на  $\hat{x}_i$

$$y_i = \beta_1 + \beta_2 \hat{x}_i + \beta_3 d_i + u_i$$

Получаем  $\hat{\beta}_1^{IV}$ ,  $\hat{\beta}_2^{IV}$  и  $\hat{\beta}_3^{IV}$

Метод двухшагового МНК также называют методом инструментальных переменных:

$$\hat{\beta}^{2OLS} = \hat{\beta}^{IV}$$

# Простейший случай двухшагового МНК

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$$

МНК:

$$\hat{\beta}_2^{OLS} = \frac{sCov(x, y)}{sVar(x)}$$

Метод инструментальных переменных:

$$\hat{\beta}_2^{IV} = \frac{sCov(z, y)}{sCov(z, x)}$$

## Пример у неоновой доски. Спасение.

$$y_i = \beta_1 + \beta_2 x_i + \beta_3 d_i + \varepsilon_i$$

Регрессор  $d_i$  ненаблюдаем.

$$\text{Var}(x_i) = \text{Var}(d_i) = 9, \text{Var}(\varepsilon_i) = 1, \text{Cov}(x_i, d_i) = -6.$$

К чему стремится IV оценка модели  $\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i$ ?

Есть инструментальная переменная  $z_i$ ,  $\text{Cov}(x_i, z_i) = 1$ .



# Как найти инструментальную переменную?

Инструментальная переменная  $z_i$  для регрессора  $x_i$  может влиять на  $y_i$  через регрессор  $x_i$ , но не через ошибку  $\varepsilon_i$ .

# Связи инструментальной переменной

Модель с эндогенностью:

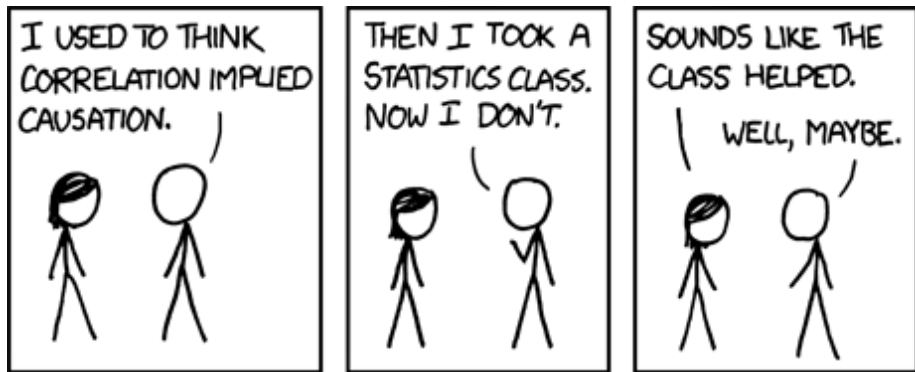
$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$$



$\varepsilon_i$

Стрелочки показывают направления влияния.

# Статистическая связь не означает причинно-следственной



Randall Munroe, <https://xkcd.com/552/>

- Данные наблюдений
- Данные экспериментов

# Данные наблюдений

Каждое утро выхожу на балкон и записываю, вижу ли я людей с зонтами и идет ли дождь

Утро	Люди с зонтами	Дождь
1	0	1
2	1	1
3	0	0
4	1	1

- По данным наблюдений не возможно определить направление причинно-следственной связи

# Данные экспериментов

Каждое утро подбрасываю монетку и в зависимости от монетки, либо беру зонт, либо не беру

Утро	Монетка	Я с зонтом	Дождь
1	Орёл	0	1
2	Решка	1	1
3	Решка	1	0
4	Орёл	0	1

- Искусственные. Проводятся человеком.
- Естественные. Сами собой возникают в природе.

# Стратегия идентификации причинно-следственных связей

- Придумать идеальный эксперимент
- Найти похожий естественный эксперимент



# Три маленьких зарисовки к данным наблюдений

- Публикационное смещение
- Выборочное исправление ошибок
- Байка про Абрахама Вальда

- У сенсационного результата больше шансов быть опубликованным

Исследователь Вениамин верит в  $H_0$ , но проводит честное исследование

- Нет ошибок. Вениамин честно опубликует исследование.
- Есть ошибка, смещающая результат в пользу  $H_0$ . Вениамин обрадуется результату и, вероятно, не заметит ошибку.
- Есть ошибка, смещающая результат в пользу  $H_a$ . Вениамин будет удивлен, трижды перепроверит работу и найдёт ошибку.

Было проведено исследование повреждений полученных вернувшимися с вылета самолетами. И предполагалось укрепить их там, где имеется больше всего повреждений.

Абрахам Вальд обратил внимание, что статистика собирается именно по вернувшимся с вылета самолётам. И, следовательно, поврежденные участки не мешают самолёту вернуться. А значит увеличивать броню надо на тех участках, где нет попаданий.

- Эндогенность — коррелированность случайной ошибки с регрессором
- Метод инструментальных переменных позволяет оценить модель в желаемой форме
- Статистическая взаимосвязь не означает причинно-следственной
- Необходимо помнить об отличии экспериментальных данных от данных наблюдений

- Борзых Д.А., Демешев Б.Б. Эконометрика в задачах и упражнениях: глава 10
- Катышев П.К., Пересецкий А. А. Эконометрика. Начальный курс: главы 8.1, 8.2, 8.3, 9.2
- Носко В.П., Эконометрика для начинающих, дополнительные главы ([www.iep.ru/files/persona/nosko/Book.pdf](http://www.iep.ru/files/persona/nosko/Book.pdf)): глава 2.1, 2.2, 2.3, 2.6