

**Эндогенность.  
Инструментальные  
переменные.**

# Эндогенность

- Эндогенный регрессор — регрессор, который коррелирован со случайными ошибками модели:  $Cov(x_i, \varepsilon_i) \neq 0$ .
- Экзогенный регрессор — регрессор, который **не** коррелирован со случайными ошибками модели:  $Cov(x_i, \varepsilon_i) = 0$ .

# Эндогенность

- Для использования МНК необходимо, чтобы все регрессоры были экзогенны (иначе получим смещенные и несостоятельные оценки)

# Эндогенность

- $y_i = \beta_1 + \beta_2 * x_i + \varepsilon_i$
- $x$  — эндогенный регрессор:  $Cov(x_i, \varepsilon_i) \neq 0$
- В этом случае МНК-оценка  $\widehat{\beta}_2$  несостоятельна и смещена

# Пример (доска)

# Когда возникает ЭНДОГЕННОСТЬ

1. Пропуск существенных переменных
2. Ошибки измерения регрессоров
3. Самоотбор
4. Одновременность

# Ошибка изменения регрессора

Замечание: если с ошибкой изменяется  $y$ , эндогенности не будет, будет просто потеря точности.

Модель в форме А:

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

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:

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

# Пропущенная переменная

- Проблема смещения из-за пропуска пропущенных переменных (если переменная пропущена, то она косвенно присутствует в ошибках уравнения)
- Инструментальные переменные можно использовать, если нет данных по пропущенной переменной и нет данных по проху – переменной.



# Пропущенная переменная

Хотим оценить форму записи А:

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

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

Не наблюдаем  $d_i$

Модель с пропущенным регрессором:

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

регрессор  $d_i$  не наблюдаем

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

Форма записи Б:

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

Эндогенность:

$$\begin{aligned} Cov(x_i, u_i) &= Cov(x_i, \beta_3 d_i + \varepsilon_i) = \\ &= \beta_3 Cov(x_i, d_i) \end{aligned}$$

# Пример (доска)

# Вывод

- При наличии эндогенности мы не можем использовать МНК

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 изменениях в  $d_i$ )

# Инструментальные переменные (IV)

- Пусть в нашем распоряжении есть переменная  $z$ , которая удовлетворяет двум свойствам:
  - Экзогенность: переменная не коррелирована со случайными ошибками  $Cov(z_i, \varepsilon_i) = 0$
  - Релевантность: переменная коррелирована с регрессором  $Cov(x_i, z_i) \neq 0$
- Тогда можно получить состоятельную оценку параметра  $\beta_2$ , используя **двухшаговый МНК**

# Инструментальные переменные

Какая из следующих пар представляет собой удачную пару переменная и инструментальная переменная к ней соответственно?

Рост и образование родителей в регрессии среднего количества очков за матч баскетболиста на его рост.

Производство самогона и покупка сахара домохозяйством в регрессии потребления самогона на его производство.

# Инструментальные переменные

Какая из следующих пар представляет собой удачную пару переменная и инструментальная переменная к ней соответственно?

Рост и образование родителей в регрессии среднего количества очков за матч баскетболиста на его рост.

Рост слабо коррелирует с образованием родителей, лучше подобрать какой-нибудь более явный инструмент.

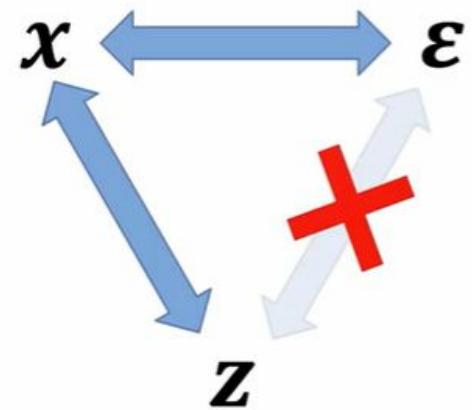
Производство самогона и покупка сахара домохозяйством в регрессии потребления самогона на его производство.

Верно! Потребление самогона зависит от склонности человека к потреблению спиртного, что не коррелирует с объёмами покупок сахара. А так как сахар является ингредиентом при производстве самогона, то эти две переменные скоррелированы.

Переменная  $z$  может использоваться как **инструментальная переменная** (кратко — **инструмент**), если она является **валидной**, то есть обладает двумя свойствами:

- **Экзогенность**: переменная не коррелирована со случайными ошибками  $Cov(z_i, \varepsilon_i) = 0$
- **Релевантность**: переменная коррелирована с регрессором  $Cov(x_i, z_i) \neq 0$

### Требования к инструментам



# Двухшаговый МНК (метод инструментальных переменных)

## Первый шаг

Оцениваем регрессию:  $x_i = \theta_1 + \theta_2 z_i + v_i$

Получаем прогнозные значения  $\hat{x}_i = \hat{\theta}_1 + \hat{\theta}_2 z_i$ .

## Второй шаг

Оцениваем регрессию:  $y_i = \beta_1 + \beta_2 \hat{x}_i + \varepsilon_i$ .

$$\left. \begin{array}{l} Cov(z_i, \varepsilon_i) = 0 \\ \hat{x}_i - \text{линейно выражено через } z_i \end{array} \right\} \Rightarrow Cov(\hat{x}_i, \varepsilon_i) = 0$$

=> Проблема эндогенности регрессора решена



# Двухшаговый МНК

Указанная процедура приводит к следующей формуле оценки коэффициента:

$$\widehat{\beta}_2^{TSLS} = \frac{\frac{1}{n} \sum (y_i - \bar{y})(z_i - \bar{z})}{\frac{1}{n} \sum (x_i - \bar{x})(z_i - \bar{z})}$$

$$\widehat{\beta}_2^{TSLS} = \frac{\frac{1}{n} \sum (y_i - \bar{y})(z_i - \bar{z})}{\frac{1}{n} \sum (x_i - \bar{x})(z_i - \bar{z})} \xrightarrow{p} \frac{Cov(y_i, z_i)}{Cov(x_i, z_i)} = \beta_2$$

Мы доказали, что двухшаговый МНК дает состоятельную оценку.

# Двухшаговый МНК

При  $n \rightarrow \infty$  выборочные ковариации сходятся к своим теоретическим аналогам =>

$$\begin{aligned}\widehat{\beta}_2^{TSLS} &= \frac{\frac{1}{n} \sum (y_i - \bar{y})(z_i - \bar{z})}{\frac{1}{n} \sum (x_i - \bar{x})(z_i - \bar{z})} \xrightarrow{p} \frac{Cov(y_i, z_i)}{Cov(x_i, z_i)} = \\ &= \frac{Cov(y_i, z_i)}{Cov(x_i, z_i)} = \frac{Cov(\beta_1 + \beta_2 * x_i + \varepsilon_i, z_i)}{Cov(x_i, z_i)} = \\ &= \frac{\beta_2 Cov(x_i, z_i) + Cov(\varepsilon_i, z_i)}{Cov(x_i, z_i)} = \beta_2\end{aligned}$$

Метод двухшагового МНК также называют методом инструментальных переменных:

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

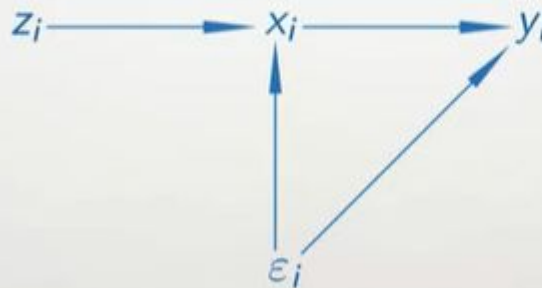
# Как найти инструментальную переменную?

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:

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



- Проблема смещения из-за пропуска пропущенных переменных (если переменная пропущена, то она косвенно присутствует в ошибках уравнения)
- Инструментальные переменные можно использовать, если нет данных по пропущенной переменной и нет данных по проху – переменной.

# Пример

Пусть истинная модель имеет вид

$$y = \beta_1 + \beta_2 x^{(2)} + \beta_3 x^{(3)} + u$$

При этом переменная  $x^{(3)}$  ненаблюдаема: у нас нет данных о ней.

**Пример:** на уровень дохода работника ( $y$ ) влияет его талант ( $x^{(3)}$ ), но у нас нет статистических данных об уровне таланта.  $x^{(2)}$  — число лет обучения — переменная, эффект которой нас интересует

# Пример

Пусть истинная модель имеет вид

$$(1) y = \beta_1 + \beta_2 x^{(2)} + \underbrace{\beta_3 x^{(3)} + u}_{\varepsilon}$$

$$(2) y = \beta_1 + \beta_2 x^{(2)} + \varepsilon$$

где  $\varepsilon = \beta_3 x^{(3)} + u$

Если  $\text{Cov}(x^{(2)}, x^{(3)}) \neq 0$ ,

то и  $\text{Cov}(x^{(2)}, \varepsilon) \neq 0$ , то есть

регрессор в модели (2) — эндогенный

# Пример

- Нам нужно придумать инструментальную переменную, которая коррелирована с уровнем образования, но некоррелирована с уровнем таланта. Это сложная задача.
- Например: насколько далеко индивид живет от колледжа.

# Пример в R-studio

```
# ЭНДОГЕННОСТЬ

data("cigarettesSW")
h <- cigarettesSW
help("cigarettesSW")

h2 <- mutate(h, rprice = price/cpi,
              rincome=income/cpi/population, rtax=tax/cpi)
h3 <- filter(h2, year=="1995")

model_0 <- lm(data=h3, log(packs)~log(rprice))
summary(model_0)
```

```
coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  10.3389      1.0353   9.986 4.25e-13 ***
log(rprice)  -1.2131      0.2164  -5.604 1.13e-06 ***
---
signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



# Пример в R-studio

- Как ценовые меры, используемые государством, повлияют на потребление сигарет?
- При оценке модели `log(packs) ~ log(rprice)` МНК мы предполагаем, что регрессор не коррелирует с ошибкой.
- Но если это не так, нужно использовать двухшаговый МНК

# Двухшаговый МНК

- Выберем инструментальную переменную, которая влияет на цену, но не влияет на спрос (акцизные сборы)
- Можно ожидать, что акцизы будут хорошей инструментальной переменной (доказать это мы не сможем)

# Двухшаговый МНК

```
# two stage OLS
# Step 1
st_1 <- lm(data=h3, log(rprice)~rtax)
h3$log_price_hat <- fitted(st_1)
# Step 2
st_2 <- lm(data=h3, log(packs)~log_price_hat)
summary(st_2)
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   10.0385     1.1651   8.616 3.72e-11 ***
log_price_hat  -1.1502     0.2436  -4.722 2.22e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Вот этот коэффициент уже можно интерпретировать как причинный, т.е. можно сказать, что при увеличении цены на 1% потребление сигарет снизится на 1,5%.

# Двухшаговый МНК

```
model_iv <- ivreg(data=h3, log(packs)~log(rprice)|rtax)
summary(model_iv)
```

coefficients:

```
      Estimate Std. Error t value Pr(>|t|)
(Intercept)  10.039      1.095   9.164 6.06e-12 ***
log(rprice)  -1.150      0.229  -5.022 8.16e-06 ***
---
```

```
mtable(model_0, st_2, model_iv)
```

```
=====
              model_0   st_2   model_iv
-----
(Intercept)  10.339*** 10.039*** 10.039***
              (1.035) (1.165)  (1.095)
log(rprice)  -1.213***           -1.150***
              (0.216)           (0.229)
log_price_hat           -1.150***
                       (0.244)
```

# Тест Хаусмана