



*Эконометрика*

# **Лекция 10**

## **Ошибки спецификации модели регрессии**

**Вакуленко Е.С.**

*д.э.н., доцент департамента прикладной экономики*

[evakulenko@hse.ru](mailto:evakulenko@hse.ru)

**Москва, 2022**



*«Регрессионный анализ – это своего рода водородная бомба в арсенале статистики».*



**Ч. Уилан**



# План

- Ошибки спецификации
- Пропущенные регрессоры в модели
- Лишние регрессоры в модели
- Приложение: модель Минцера для заработной платы
- Квадратичные модели и модели с перекрестными переменными
- Тест Рамсея



# Пропущенные регрессоры в модели



# Теорема Гаусса-Маркова

- **Если** выполнены предположения для случайного члена ( $E(\varepsilon_i) = 0$ ,  $Var(\varepsilon_i) = \sigma^2$ ,  $Cov(\varepsilon_i, \varepsilon_j) = 0$ , случайный член **независим** от объясняющих переменных)
- Модель регрессии **правильно специфицирована**
  - Нет **пропущенных** или **лишних** переменных
  - Выбрана правильная **функциональная форма**
- $X_i$  **детерминированы и не все равны** между собой

**То** оценки метода наименьших квадратов эффективны в классе линейных несмещённых оценок.



# Теорема Гаусса-Маркова

- Если выполнены предположения для случайного члена ( $E(\varepsilon_i) = 0$ ,  $Var(\varepsilon_i) = \sigma^2$ ,  $Cov(\varepsilon_i, \varepsilon_j) = 0$ , случайный член **независим** от объясняющих переменных)
- Модель регрессии правильно специфицирована
  - Нет пропущенных или лишних переменных
  - Выбрана правильная функциональная форма
- $X_i$  детерминированы и не все равны между собой

То оценки метода наименьших квадратов эффективны в классе линейных

Что произойдет с оценками МНК, если будет нарушено требование о правильной спецификации?



# Ошибки спецификации

- Пропуск важной переменной
- Включение лишней переменной
- Выбор неправильной функциональной формы



# Пример про гольф

- Игроки в гольф чаще болеют сердечно-сосудистыми заболеваниями, раком и артритом.



Источник: Ч. Уилан / forfun.com





# Пример про гольф

- Игроки в гольф чаще болеют сердечно-сосудистыми заболеваниями, раком и артритом.
- **Пропущена важная переменная – возраст!**
- Не гольф убивает людей, а старость.
- Включили возраст и получили:
- Для людей одного и того же возраста игра в гольф может стать профилактикой серьезных заболеваний.



Источник: Ч. Уилан / forfun.com



# Пример со школами

- **Задача:** объяснить качество школ
- Зависимая переменная: результаты экзаменов
- Объясняющая переменная: расходы школы
- **Результат:** положительная корреляция.



# Пример со школами

- **Задача:** объяснить качество школ
- Зависимая переменная: результаты экзаменов
- Объясняющая переменная: расходы школы
- **Результат:** положительная корреляция.
  
- **Что пропущено?**
- Способности учеников (уровень образования родителей)
- Социально-экономическое положение учащихся



# Пропуск важной переменной

- Оценки коэффициентов регрессии **смещены!**
- Оценки дисперсий коэффициентов регрессии также **смещены!**
- $t$  и  $F$  статистики рассчитываются неправильно
- Диагностика: **тест Рамсея**



# Ошибки спецификации: невключение существенной переменной

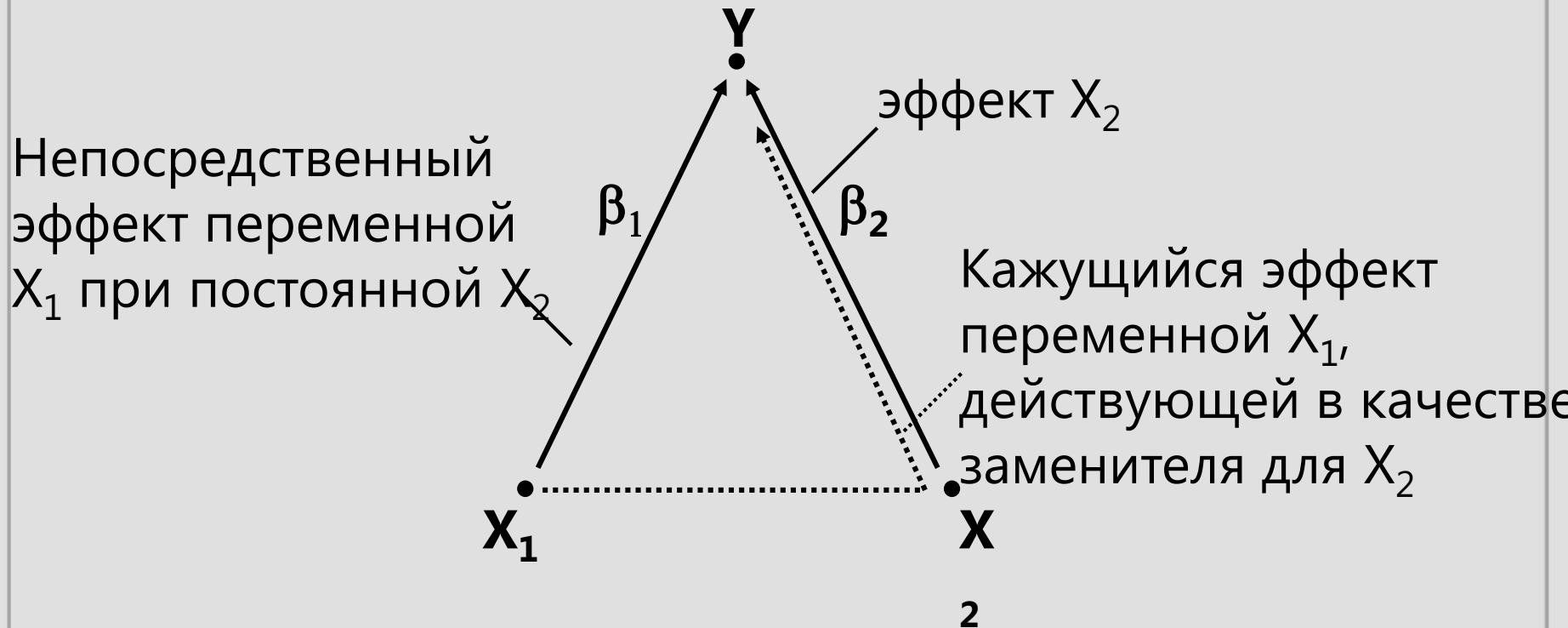
$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \varepsilon \quad \hat{Y} = \hat{\alpha} + \hat{\beta}_1 X_1$$

$$E(\hat{\beta}_1) = \beta_1 + \beta_2 \frac{\sum (X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2)}{\sum (X_{1i} - \bar{X}_1)^2}$$

Смещение в коэффициенте



# Ошибки спецификации: невключение существенной переменной



# Пропуск существенной переменной

Истинная модель (1)

$$Y = X\beta + Z\gamma + \varepsilon$$

Оцениваемая модель (2)

$$Y = X\beta + \varepsilon$$



# Оценки «длинной» модели (1)

Оценки коэффициентов

$$\hat{\beta}^{(1)} = (X'X)^{-1} X'Y - LM_Z^{-1} Z'M_X Y$$

$$\hat{\gamma}^{(1)} = M_Z^{-1} Z'M_X Y$$

Ковариационная матрица оценок коэффициентов

$$V \begin{pmatrix} \hat{\beta}^{(1)} \\ \hat{\gamma}^{(1)} \end{pmatrix} = \sigma_\varepsilon^2 \begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z \end{bmatrix}^{-1} = \sigma_\varepsilon^2 \begin{bmatrix} (X'X)^{-1} + LM_Z^{-1}L' & -LM_Z^{-1} \\ -M_Z^{-1}L' & M_Z^{-1} \end{bmatrix}$$

$$M_Z = I - Z(Z'Z)^{-1} Z'$$

где

$$M_X = I - X(X'X)^{-1} X'$$

$$L = (X'X)^{-1} X'Z$$





# На заметку: обращение блочных матриц

**Т** Теорема [Фробениус]. Пусть имеется блочная квадратная матрица вида

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix},$$

где матрица  $A$  — квадратная порядка  $k$ , а матрица  $D$  — квадратная порядка  $\ell$ . Тогда

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} A^{-1} + A^{-1}BK^{-1}CA^{-1} & -A^{-1}BK^{-1} \\ -K^{-1}CA^{-1} & K^{-1} \end{pmatrix},$$

где матрица

$$K = D - CA^{-1}B$$

называется **шуровским дополнением** к подматрице  $A$ . Здесь предполагается, что матрицы  $A$  и  $K$  — неособенные.

**=>** При  $B = \mathbb{O}$  имеем:

$$\begin{pmatrix} A & \mathbb{O} \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} A^{-1} & \mathbb{O} \\ -D^{-1}CA^{-1} & D^{-1} \end{pmatrix},$$

если матрицы  $A$  и  $D$  — неособенные.

**Доказательство.** Будем искать



## Оценки «короткой» модели (2)

Оценки коэффициентов

$$\hat{\beta}^{(2)} = (X'X)^{-1} X'Y$$

Ковариационная матрица оценок  
коэффициентов

$$V\left(\hat{\beta}^{(2)}\right) = \sigma_{\varepsilon}^2 (X'X)^{-1}$$



# Свойства оценок коэффициентов

Исследование свойств оценок коэффициентов модели (2):

$$\hat{\beta}^{(2)} = (X'X)^{-1} X'Y \stackrel{Y=X\beta+Z\gamma+\varepsilon}{=} \beta + (X'X)^{-1} X'Z\gamma + (X'X)^{-1} X'\varepsilon$$

Оценка оказывается смещенной:

$$E(\hat{\beta}^{(2)}) = \beta + (X'X)^{-1} X'Z\gamma + (X'X)^{-1} X'E(\varepsilon) = \beta + (X'X)^{-1} X'Z\gamma \neq \beta$$

Смещение  $(X'X)^{-1} X'Z\gamma$  не исчезает при увеличении объема выборки, оценка становится несостоятельной  
Интерпретация смещения: произведение коэффициентов регрессий  $Z$  на  $X$  и истинного значения коэффициентов  $\gamma$ .

Всегда ли будет смещение?  $X'Z = 0$

Если  $X$  и  $Z$  ортогональны, то есть  $X'Z = 0$ , в этом случае смещения не будет, но чем значительнее корреляция между  $X$  и  $Z$  тем серьезнее смещение



# Свойства ковариационной матрицы оценок коэффициентов

Исследование свойств ковариационной матрицы оценок коэффициентов:

$$\begin{aligned} V(\hat{\beta}^{(2)}) &= \sigma_{\varepsilon}^2 (X'X)^{-1} \neq V(\hat{\beta}^{(1)}) = \sigma_{\varepsilon}^2 \left[ (X'X)^{-1} + LM_Z^{-1}L' \right] = \\ &= \sigma_{\varepsilon}^2 \left[ (X'X)^{-1} + (X'X)^{-1} X'Z \left( I - Z(Z'Z)^{-1}Z' \right)^{-1} Z'X (X'X)^{-1} \right] \end{aligned}$$

Ковариационная матрица вычисляется неверно:  
ее диагональные элементы занижены по сравнению с теоретическими значениями

Всегда ли будет смещение?

Если  $X'Z = 0$ , в этом случае смещения не будет,  
но чем значительнее корреляция между  $X$  и  $Z$ , тем серьезнее смещение



# Оценка дисперсии регрессии

## Оценка дисперсии регрессии

$$\left(\hat{\sigma}_{\varepsilon}^2\right)^{(2)} = \frac{RSS^{(2)}}{n-k} = \frac{e^{(2)'}e^{(2)}}{n-k} = \frac{\left(Y - X\hat{\beta}^{(2)}\right)' \left(Y - X\hat{\beta}^{(2)}\right)}{n-k} = \frac{Y'M_x Y}{n-k}$$

Оценка оказывается **смещенной**:

$$\begin{aligned} E\left\{\left(\hat{\sigma}_{\varepsilon}^2\right)^{(2)}\right\} &= \frac{E\left\{\left(X\beta + Z\gamma + \varepsilon\right)' M_x \left(X\beta + Z\gamma + \varepsilon\right)\right\}_{M_x X\beta=0}}{n-k} = \\ &= \frac{E\left\{\left(Z\gamma + \varepsilon\right)' M_x \left(Z\gamma + \varepsilon\right)\right\}_{E(M_x \varepsilon)=0}}{n-k} = \frac{E\left(\varepsilon' M_x \varepsilon\right) + \gamma' Z' M_x Z \gamma}{n-k} = \\ &= \frac{\sigma_{\varepsilon}^2 \text{tr} M_x + \gamma' Z' M_x Z \gamma}{n-k} = \sigma_{\varepsilon}^2 + \frac{\gamma' Z' M_x Z \gamma}{n-k} \neq \sigma_{\varepsilon}^2 \end{aligned}$$



# Модель Минцера

- Зависимость заработной платы от индивидуальных характеристик работника.
- Расширение модели: характеристики предприятия, отрасли, макро переменные и т.д.
- Базовая модель:

$$\ln wage_i = \alpha + \beta_1 S_i + \beta_2 EXP_i + \beta_3 EXP_i^2 + \epsilon_i$$

$\ln wage$  – логарифм почасовой заработной платы;

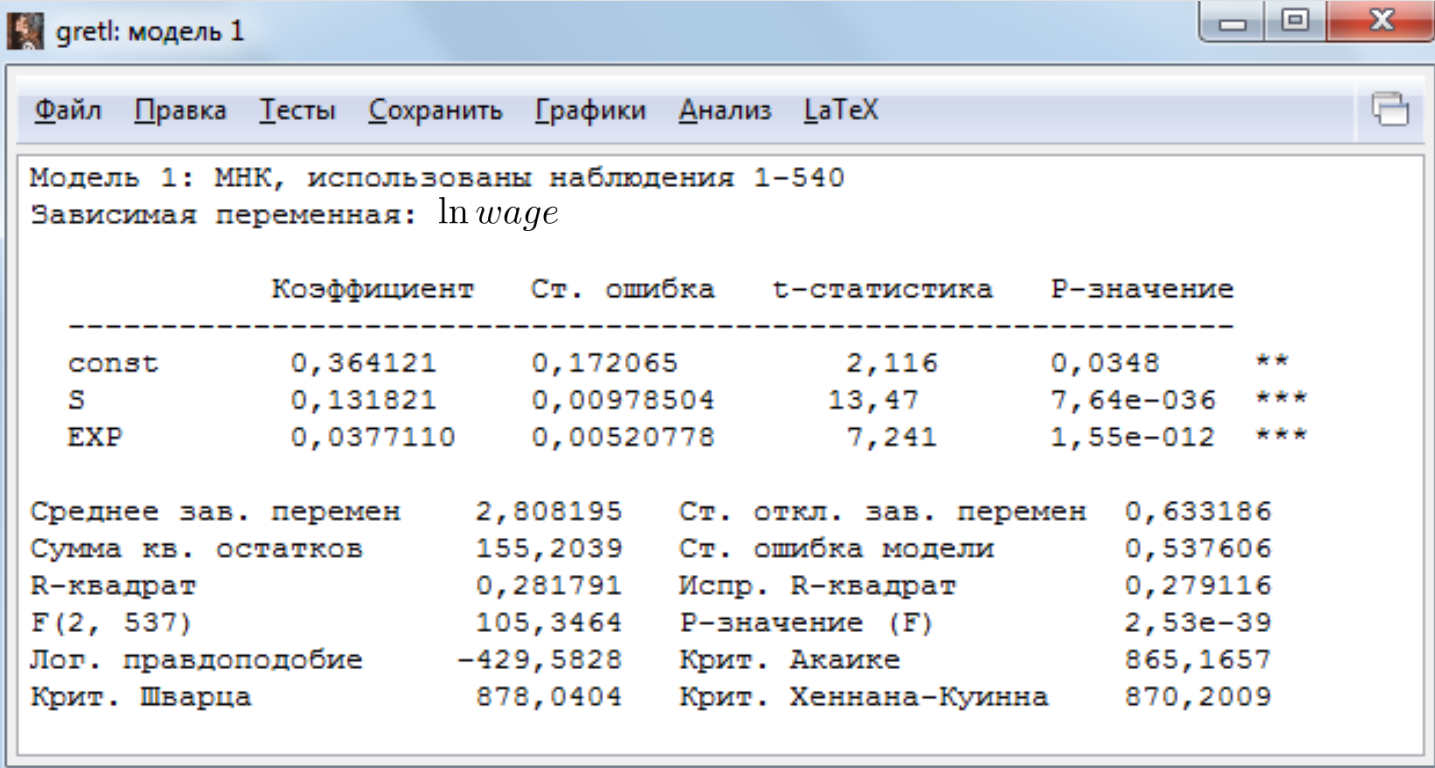
$S$  – число лет обучения

$EXP$  – опыт работы.



# Пример. Модель заработной платы

$$\ln \widehat{wage}_i = 0.36 + 0.13S_i + 0.04EXP_i$$



The screenshot shows the 'gretl: модель 1' window. The menu bar includes 'Файл', 'Правка', 'Тесты', 'Сохранить', 'Графики', 'Анализ', and 'LaTeX'. The main text area displays the model specification: 'Модель 1: МНК, использованы наблюдения 1-540' and 'Зависимая переменная:  $\ln wage$ '. Below this is a table of coefficients and statistics.

|                      | Коэффициент | Ст. ошибка             | t-статистика | P-значение |     |
|----------------------|-------------|------------------------|--------------|------------|-----|
| const                | 0,364121    | 0,172065               | 2,116        | 0,0348     | **  |
| S                    | 0,131821    | 0,00978504             | 13,47        | 7,64e-036  | *** |
| EXP                  | 0,0377110   | 0,00520778             | 7,241        | 1,55e-012  | *** |
| -----                |             |                        |              |            |     |
| Среднее зав. перемен | 2,808195    | Ст. откл. зав. перемен | 0,633186     |            |     |
| Сумма кв. остатков   | 155,2039    | Ст. ошибка модели      | 0,537606     |            |     |
| R-квадрат            | 0,281791    | Испр. R-квадрат        | 0,279116     |            |     |
| F(2, 537)            | 105,3464    | P-значение (F)         | 2,53e-39     |            |     |
| Лог. правдоподобие   | -429,5828   | Крит. Акаике           | 865,1657     |            |     |
| Крит. Шварца         | 878,0404    | Крит. Хеннана-Куинна   | 870,2009     |            |     |

$\ln wage$  – логарифм почасовой заработной платы,  
S – число лет обучения, EXP – опыт работы



# Интерпретация полулогарифмической модели

$$\ln \widehat{wage}_i = 0.36 + 0.13S_i + 0.04EXP_i$$

- При увеличении  $X$  на 1 ед. измерения  $Y$  изменится на  $(e^\beta - 1) \cdot 100\%$
- Увеличение числа лет обучения на 1 год приводит к росту заработной платы на 13.9%
- Увеличение опыта работы на 1 год приводит к увеличению заработной платы на 4%





# Пример

$$\ln \widehat{wage}_i = 0.36 + 0.13S_i + 0.04EXP_i$$

EXP

Корреляция между S и

|     | S       | EXP    |
|-----|---------|--------|
| S   | 1.0000  |        |
| EXP | -0.2179 | 1.0000 |

$$E(\hat{\beta}_1) = \beta_1 + \beta_2 \frac{\sum (S_i - \bar{S})(EXP_i - \overline{EXP})}{\sum (S_i - \bar{S})^2}$$

Если опущена переменная  $EXP$ , то смещение коэффициента перед переменной  $S$  будет отрицательным, т.к. оценка коэффициента  $\beta_2$  положительная, а коэффициент корреляции  $S$  и  $EXP$  отрицательный.



# Пример

$$\ln \widehat{wage}_i = 0.36 + 0.13S_i + 0.04EXP_i$$

EXP

Корреляция между S и

|     | S       | EXP    |
|-----|---------|--------|
| S   | 1.0000  |        |
| EXP | -0.2179 | 1.0000 |

$$E(\hat{\beta}_2) = \beta_2 + \beta_1 \frac{\sum (EXP_i - \overline{EXP})(S_i - \bar{S})}{\sum (EXP_i - \overline{EXP})^2}$$

Аналогично, если пропущена переменная S, то оценка коэффициента перед переменной EXP будет смещена вниз.



## Пример 2

Оценка полной модели

$$\ln \widehat{wage}_i = 0.36 + 0.13S_i + 0.04EXP_i$$

(0.17)   (0.01)   (0.01)

Оценка модели без опыта работы

$$\ln \widehat{wage}_i = 1.29 + 0.11S_i$$

(0.13)   (0.01)

Оценка модели без образования

$$\ln \widehat{wage}_i = 2.45 + 0.02EXP_i$$

(0.1)   (0.01)

Смещение в случае невключения одной из переменных  $S$  или  $EXP$  действительно является отрицательным.



# Включение лишней переменной



# Важное предостережение

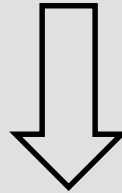


- Боясь пропустить важную переменную, не следует наращивать число регрессоров, включая их без особого основания.
- Какая-то из них по чистой случайности может оказаться значимой, хотя на самом деле это не так.
- Сложно распознать, какая из них действительно лишняя.
- Нужна экономическая теория!



# Включение лишней переменной

- Оценки коэффициентов регрессии **несмещенные**
- Оценки дисперсий коэффициентов **смещены**



- Оценки коэффициентов **неэффективные**
- Диагностика: **F-тест** на группу незначимых переменных



# Ошибки спецификации: включение лишней переменной

$$Y = \alpha + \beta_1 X_1 + \varepsilon$$

Истинная модель

$$\hat{Y} = \hat{\alpha} + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$$

модель

Оценивается такая

$$Y = \alpha + \beta_1 X_1 + 0X_2 + \varepsilon$$

Но в истинной модели  
нет фактора  $X_2$



# Ошибки спецификации: включение лишней переменной

При включении лишней переменной  $X_2$  увеличивается оценка дисперсии коэффициента перед переменной  $X_1$ . Добавляется множитель  $1 / (1 - r^2)$ , где  $r$  – коэффициент корреляции между  $X_1$  и  $X_2$ . Чем больше корреляция, тем больше дисперсия.

$$\sigma_{\hat{\beta}_1}^2 = \frac{\sigma_\varepsilon^2}{\sum (X_{1i} - \bar{X}_1)^2} \times \frac{1}{1 - r_{X_1, X_2}^2}$$







По данным 1995 г. US Consumer Expenditure Survey  
для 868 домохозяйств:

$$\widehat{LNFDHO} = 4.72 + 0.29LNEXP + 0.49LNSIZE$$

(0.22)    (0.02)                      (0.03)

$$R^2 = 0.52$$

$LNFDHO$  – логарифм ежегодных расходов домохозяйств на продукты домашнего потребления;  
 $LNEXP$  – логарифм общих годовых расходов домохозяйств;  $LNSIZE$  – логарифм числа потребителей в домохозяйстве.



# Пример

Добавим в модель лишнюю переменную LNHOUS (логарифм расходов на жилье).

Корреляция между LNHOUS, LNEXP, LNSIZE

|        | LNHOUS | LNEXP  | LNSIZE |
|--------|--------|--------|--------|
| LNHOUS | 1.0000 |        |        |
| LNEXP  | 0.8137 | 1.0000 |        |
| LNSIZE | 0.3256 | 0.4491 | 1.0000 |

$$\widehat{LNFDHO} = 4.71 + 0.27LNEXP + 0.49LNSIZE + 0.02LNHOUS$$

(0.22)   (0.04)                      (0.03)                      (0.03)

LNHOUS – логарифм годовых расходов на жилье.

Переменная незначима, т.е. лишняя в этой регрессии.



# Обратим внимание на коэффициенты регрессии

Первоначальное уравнение регрессии

$$\widehat{LNFDHO} = 4.72 + 0.29LNEXP + 0.49LN\text{SIZE}$$

(0.22)   (0.02)                      (0.03)

Оценка с дополнительным факторов – логарифм расходов на жилье.

$$\widehat{LNFDHO} = 4.71 + 0.27LNEXP + 0.49LN\text{SIZE} + 0.02LN\text{HOUS}$$

(0.22)   (0.04)                      (0.03)                      (0.03)

**Коэффициенты регрессии практически не меняются!**



# Обратим внимание на стандартные отклонения коэффициентов

Первоначальное уравнение регрессии

| LNFDHO | Coef.    | Std. Err. | t      | P> t  | [95% Conf. Interval] |          |
|--------|----------|-----------|--------|-------|----------------------|----------|
| LNEXP  | .2866813 | .0226824  | 12.639 | 0.000 | .2421622             | .3312003 |
| LNSIZE | .4854698 | .0255476  | 19.003 | 0.000 | .4353272             | .5356124 |
| _cons  | 4.720269 | .2209996  | 21.359 | 0.000 | 4.286511             | 5.154027 |

Оценка с дополнительным факторов – логарифм расходов на жилье.

| LNFDHO | Coef.    | Std. Err. | t      | P> t  | [95% Conf. Interval] |          |
|--------|----------|-----------|--------|-------|----------------------|----------|
| LNEXP  | .2673552 | .0370782  | 7.211  | 0.000 | .1945813             | .340129  |
| LNSIZE | .4868228 | .0256383  | 18.988 | 0.000 | .4365021             | .5371434 |
| LNHOUS | .0229611 | .0348408  | 0.659  | 0.510 | -.0454214            | .0913436 |
| _cons  | 4.708772 | .2217592  | 21.234 | 0.000 | 4.273522             | 5.144022 |

Стандартные отклонения коэффициентов регрессии растут!



## Ошибки спецификации

|                  |  | Истинная модель                          |  |
|------------------|--|--|--|
|                  |  | $Y = \alpha + \beta_1 X_1 + \varepsilon$ | $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$ |
| Оцененная модель | $\hat{Y} = \hat{\alpha} + \hat{\beta}_1 X_1$                     | Правильная спецификация, все в порядке   |  |
|                  | $\hat{Y} = \hat{\alpha} + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$ |  |  |



## Ошибки спецификации

|                  |  | Истинная модель                          |  |
|------------------|--|--|--|
|                  |  | $Y = \alpha + \beta_1 X_1 + \varepsilon$ | $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$ |
| Оцененная модель | $\hat{Y} = \hat{\alpha} + \hat{\beta}_1 X_1$                     | Правильная спецификация, все в порядке   |  |
|                  | $\hat{Y} = \hat{\alpha} + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$ |  | Правильная спецификация, все в порядке                 |



## Ошибки спецификации

|                  |  | Истинная модель   |  |
|------------------|--|---|--|
|                  |  | $Y = \alpha + \beta_1 X_1 + \varepsilon$                      | $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$ |
| Оцененная модель | $\hat{Y} = \hat{\alpha} + \hat{\beta}_1 X_1$                     | Правильная спецификация, все в порядке                        |  |
|                  | $\hat{Y} = \hat{\alpha} + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$ | Оценки коэффициентов являются несмещенными, но неэффективными | Правильная спецификация, все в порядке                 |



## Ошибки спецификации

|                  |  | Истинная модель   |  |
|------------------|--|---|--|
|                  |  | $Y = \alpha + \beta_1 X_1 + \varepsilon$                      | $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$ |
| Оцененная модель | $\hat{Y} = \hat{\alpha} + \hat{\beta}_1 X_1$                     | Правильная спецификация, все в порядке                        | <u>Оценки коэффициентов будут смещены</u>              |
|                  | $\hat{Y} = \hat{\alpha} + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$ | Оценки коэффициентов являются несмещенными, но неэффективными | Правильная спецификация, все в порядке                 |

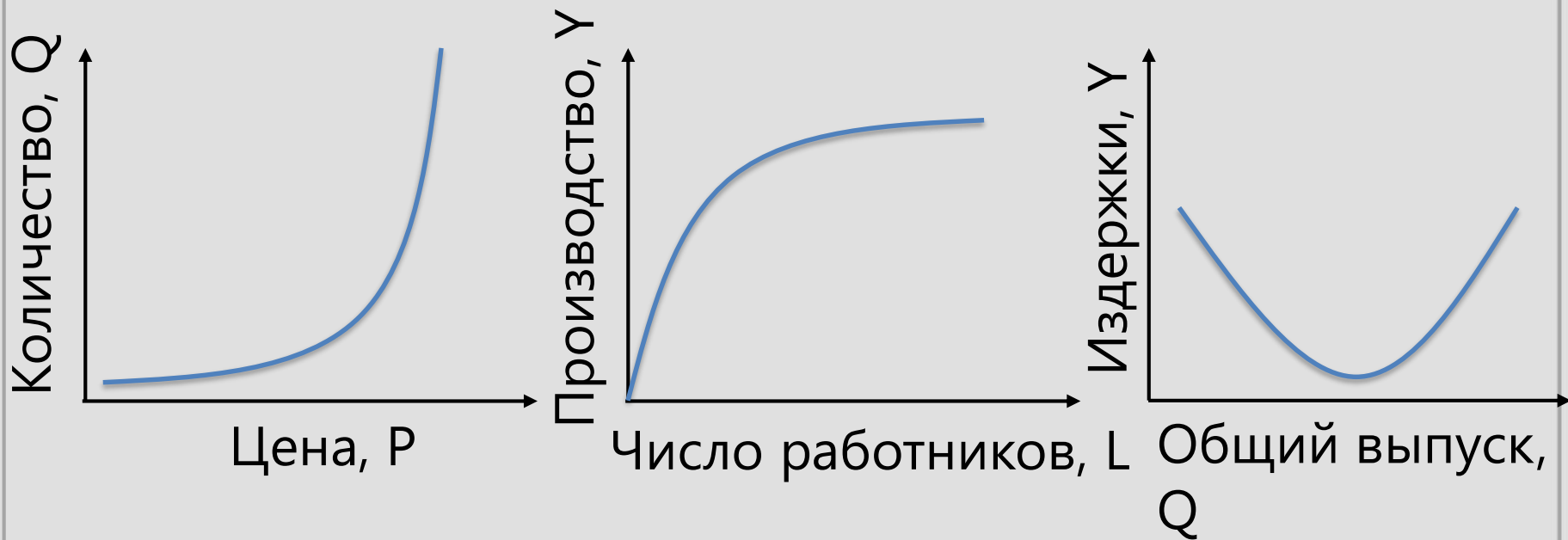




# **Квадратичные модели и модели с перекрестными переменными**



# Квадратичные модели и модели с перекрестными переменными



# Трансформация квадратичной модели

Рассмотрим модель, в которой объясняемая переменная  $y$  зависит от одного фактора  $x$ , но зависимость является квадратичной:

$$y = \alpha + \beta_1 x + \beta_2 x^2 + \varepsilon$$

Можно считать, что объясняемая переменная  $y$  – это заработная плата, а  $x$  – опыт работы.

Заметим, что в данном случае  $\beta_1$  не отражает изменение  $y$  при изменении  $x$ . Поскольку, изменяя  $x$ , меняется также  $x^2$ .



# Трансформация квадратичной модели

Рассмотрим оценку модели:

$$\hat{y} = \hat{\alpha} + \hat{\beta}_1 x + \hat{\beta}_2 x^2$$

тогда можно оценить изменение  $y$  как:

$$\Delta \hat{y} = (\hat{\beta}_1 + 2\hat{\beta}_2 x) \Delta x, \quad \Delta \hat{y} / \Delta x = \hat{\beta}_1 + 2\hat{\beta}_2 x$$

Таким образом, эффект изменения  $x$  на  $y$   
в данном случае зависит от конкретного значения  $x$ .



# Пример. Квадратичная модель

Рассмотрим взаимосвязь между заработной платой и опытом работы.

$$n = 526, R^2 = 0.093$$

$$\widehat{wage}_i = 3.73 + 0.298EXP_i - 0.0061EXP_i^2$$

(0.35)    (0.04)                    (0.001)



# Расчет вершины

Вершина параболы рассчитывается по следующей формуле:

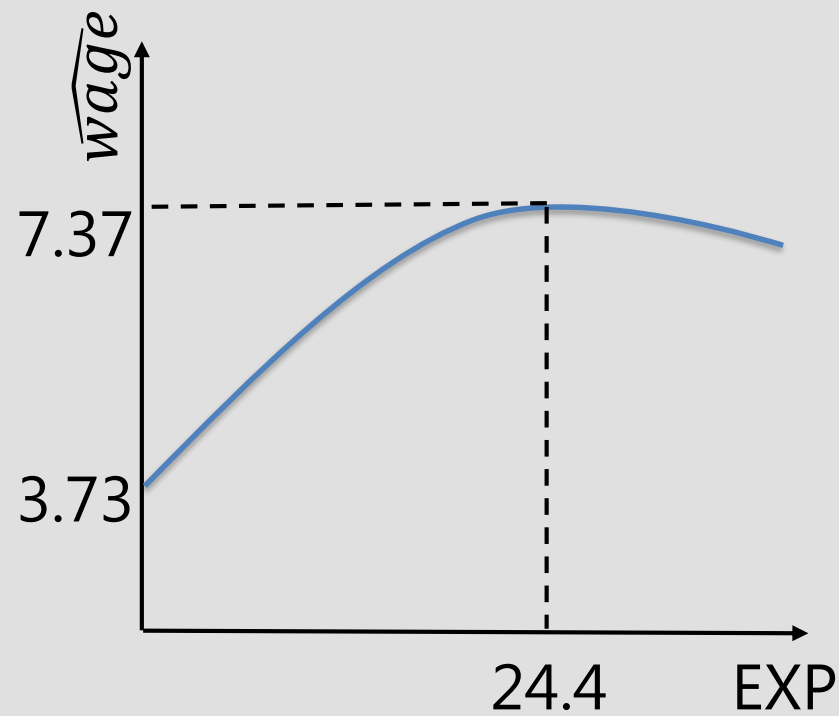
$$x^* = \left| \hat{\beta}_1 / (2\hat{\beta}_2) \right|$$

В примере с заработной платой:

$$x^* = \text{EXP}^* \quad 0.298 / [2(0.0061)] \approx 24.4$$



# Пример. Иллюстрация



# Пример. Модель с перекрестными переменными

$$price_i = \alpha + \beta_1 sqrf t_i + \beta_2 bdrms_i + \beta_3 sqrf t_i \cdot bdrms_i + \beta_4 bthrms_i + \varepsilon_i$$

- price — цена квадратного метра жилья;
- sqrf t — площадь жилья в футах;
- bdrms — число спальных комнат;
- bthrms — число ваннных комнат.





# Интерпретация

$$price_i = \alpha + \beta_1 sqrft_i + \beta_2 bdrms_i + \beta_3 sqrft_i \cdot bdrms_i + \\ + \beta_4 bthrms_i + \varepsilon_i$$

Если  $\beta_3 > 0$ , то дополнительная спальня дает более высокий рост цен на жилье для больших домов. Другими словами, существует перекрестный эффект между площадью дома и количеством спален.



# Интерпретация

$$price_i = \alpha + \beta_1 sqrf t_i + \beta_2 bdrms_i + \beta_3 sqrf t_i \cdot bdrms_i + \\ + \beta_4 bthrms_i + \varepsilon_i$$

Исходные параметры сложнее интерпретировать, когда мы включаем в уравнение перекрестный член.

Например,  $\beta_2$  - это влияние  $bdrms$  на цену дома с нулевой площадью!

Этот параметр не представляет особого интереса.

Вместо этого мы должны подставить такие значения  $sqrf t$ , как среднее или медианное значение выборки в уравнение.



# Перепараметризация

Рассмотрим модель с двумя объясняющими переменными и перекрестным членом:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$$

$\beta_2$  - это влияние  $x_2$  на  $y$ , когда  $x_1 = 0$ . Часто это не то, что интересует исследователя. Вместо этого мы можем перепараметризовать модель следующим образом:

где  $\mu_1$  - выборочное среднее  $x_1$ ,  $\mu_2$  - выборочное среднее  $x_2$ .

$$y = \alpha + \delta_1 x_1 + \delta_2 x_2 + \beta_3 (x_1 - \mu_1)(x_2 - \mu_2) + \varepsilon,$$



# Перепараметризация

$$y = \alpha + \delta_1 x_1 + \delta_2 x_2 + \beta_3 (x_1 - \mu_1) (x_2 - \mu_2) + \varepsilon,$$

Теперь  $\delta_2$  - это влияние  $x_2$  на  $y$ , когда  $x_1 = \mu_1$ .

Кроме того, мы сразу же получаем стандартные ошибки для предельных эффектов при средних значениях.



# Тест Рамсея



# RESET – regression specification error test

RESET – тест Рамсея отвечает на вопрос, надо ли включать в регрессию **степени независимых переменных** (регрессоров).



# RESET – тест Рамсея

$$Y = \alpha + \beta_1 X_1 + \dots + \beta_k X_k + \varepsilon \quad (*)$$

$H_0$ : спецификация модели (\*) является правильной;

$H_1$ : спецификация модели (\*) является неправильной.



# Процедура теста Рамсея

1. Оцениваем коэффициенты функции регрессии (\*)

$$\hat{Y} = \hat{\alpha}_1 + \hat{\beta}_1 X_1 + \dots + \hat{\beta}_k X_k$$

2. Сохраняем столбец оцененных значений  $\hat{Y}$
3. Оцениваем коэффициенты вспомогательной регрессии (\*\*)

$$Y = \alpha + \beta_1 X_1 + \dots + \beta_k X_k + \alpha_2 \hat{Y}^2 + \dots + \alpha_m \hat{Y}^m + \varepsilon$$





# Процедура теста Рамсея

4. Тогда проверка гипотезы о правильной спецификации равносильна проверке гипотезы (коэффициенты при  $\hat{Y}$ ):

$H_0$ :  $\alpha_2 = \dots = \alpha_m = 0$  модель правильно специфицирована

$H_1$ :  $\exists \alpha_i \neq 0, i = 2, \dots, m$

5. Вычисляем значение тестовой статистики

$$F = \frac{(RSS_R - RSS_{UR}) / (m - 1)}{RSS_{UR} / (n - (k + m))}$$

где  $RSS_R$  - это сумма квадратов остатков модели (\*),  
а  $RSS_{UR}$  - это сумма квадратов остатков модели (\*\*)



# Процедура теста Рамсея

4. Тогда проверка гипотезы о правильной спецификации равносильна проверке гипотезы (коэффициенты при  $\hat{Y}$ ):

$H_0$ :  $\alpha_2 = \dots = \alpha_m = 0$  модель правильно специфицирована

$H_1$ :  $\exists \alpha_i \neq 0, i = 2, \dots, m$

5. Вычисляем значение тестовой статистики

$$F = \frac{(RSS_R - RSS_{UR}) / (m - 1)}{RSS_{UR} / (n - (k + m))}$$

$$F > F_{m-1, n-(k+m)}$$

6. Если \_\_\_\_\_ для заданного уровня значимости  $\alpha$ , то гипотеза  $H_0$  отвергается.
7. Или если  $p - value < 0.05 \Rightarrow H_0$  отвергается на 5% уровне значимости



# Пример. Модель заработной платы



Имеется выборка результатов опросов населения РМЭЗ НИУ ВШЭ, XVI волна, 2007 г. Отобраны данные только по трудоспособному населению Центрального и Центрально–Черноземного экономического района.

## Список переменных:

- wage — заработная плата, полученная за последние 30 дней по основному месту работы (в рублях);
- high — 1, если высшее образование, 0 иначе;
- male — пол, 1 для мужчин, 0 для женщин;
- EXP — число лет общего трудового стажа респондента.



# Пример. Модель заработной платы



- Оцените зависимость заработной платы от опыта работы, образования и пола респондента в виде линейной регрессии:

$$\ln wage_i = \alpha + \beta_1 EXP_i + \beta_2 high_i + \beta_3 male_i + \varepsilon_i$$

- Прodelайте тест Рамсея на ошибки спецификации.
- Включите в модель дополнительную переменную — квадрат опыта работы. Оцените регрессию:
$$\ln wage_i = \alpha + \beta_1 EXP_i + \beta_4 EXP_i^2 + \beta_2 high_i + \beta_3 male_i + \varepsilon_i$$
- Прodelайте тест Рамсея на ошибки спецификации для расширенной модели. Сделайте выводы.



# Оценка модели в Gretl

$$\widehat{\ln wage_i} = 8.18 + 0.002EXP_i + 0.36high_i + 0.48male_i$$

gretl: модель 1

Файл Правка Тесты Сохранить Графики Анализ LaTeX

Модель 1: МНК, использованы наблюдения 1-1574  
Зависимая переменная: lnwage

|                      | Коэффициент | Ст. ошибка             | t-статистика | Р-значение |     |
|----------------------|-------------|------------------------|--------------|------------|-----|
| const                | 8,18381     | 0,0392599              | 208,5        | 0,0000     | *** |
| EXP                  | 0,00151431  | 0,00164966             | 0,9180       | 0,3588     |     |
| high                 | 0,358090    | 0,0434703              | 8,238        | 3,66e-016  | *** |
| male                 | 0,478253    | 0,0356629              | 13,41        | 6,66e-039  | *** |
| Среднее зав. перемен | 8,505909    | Ст. откл. зав. перемен | 0,751774     |            |     |
| Сумма кв. остатков   | 774,5999    | Ст. ошибка модели      | 0,702407     |            |     |
| R-квадрат            | 0,128688    | Испр. R-квадрат        | 0,127023     |            |     |
| F(3, 1570)           | 77,29365    | Р-значение (F)         | 1,24e-46     |            |     |
| Лог. правдоподобие   | -1675,404   | Крит. Акаике           | 3358,807     |            |     |
| Крит. Шварца         | 3380,253    | Крит. Хеннана-Куинна   | 3366,777     |            |     |

Исключая константу, наибольшее р-значение получено для переменной 3 (EXP)

Переменная EXP (опыт работы) оказалась незначимая.



# Результаты теста Рамсея (только квадраты)

gretl: тест Рамсея (RESET)

Вспомогательная регрессия для теста Рамсея  
МНК, использованы наблюдения 1-1574  
Зависимая переменная: lnwage

|        | Коэффициент | Ст. ошибка | t-статистика | P-значение |     |
|--------|-------------|------------|--------------|------------|-----|
| const  | 49,1480     | 17,2370    | 2,851        | 0,0044     | *** |
| EXP    | 0,0173411   | 0,00686032 | 2,528        | 0,0116     | **  |
| high   | 4,11905     | 1,58314    | 2,602        | 0,0094     | *** |
| male   | 5,46814     | 2,09996    | 2,604        | 0,0093     | *** |
| yhat^2 | -0,612105   | 0,257563   | -2,377       | 0,0176     | **  |

Тестовая статистика:  $F = 5,647880$ ,  
p-значение =  $P(F(1,1569) > 5,64788) = 0,0176$

$P - value = 0.0176 < 0.05$ . Следовательно, модель неправильно специфицирована, есть пропущенные степени регрессоров!



# Добавим опыт в квадрате

$$\ln \widehat{wage}_i = 8.1 + 0.01EXP_i - 0.0003EXP_i^2 + 0.36high_i +$$

gretl: модель 2

Файл П\_равка Т\_есты С\_охранить Г\_рафики А\_нализ L\_aT\_eX

Модель 2: МНК, использованы наблюдения 1-1574  
Зависимая переменная: lnwage

|                      | Коэффициент  | Ст. ошибка             | t-статистика | P-значение |     |
|----------------------|--------------|------------------------|--------------|------------|-----|
| const                | 8,10444      | 0,0534235              | 151,7        | 0,0000     | *** |
| EXP                  | 0,0142626    | 0,00605603             | 2,355        | 0,0186     | **  |
| EXP2                 | -0,000347773 | 0,000158976            | -2,188       | 0,0288     | **  |
| male                 | 0,486219     | 0,0358057              | 13,58        | 8,56e-040  | *** |
| high                 | 0,356150     | 0,0434271              | 8,201        | 4,90e-016  | *** |
| -----                |              |                        |              |            |     |
| Среднее зав. перемен | 8,505909     | Ст. откл. зав. перемен | 0,751774     |            |     |
| Сумма кв. остатков   | 772,2445     | Ст. ошибка модели      | 0,701562     |            |     |
| R-квадрат            | 0,131338     | Испр. R-квадрат        | 0,129123     |            |     |
| F(4, 1569)           | 59,30640     | P-значение (F)         | 1,11e-46     |            |     |
| Лог. правдоподобие   | -1673,007    | Крит. Акаике           | 3356,014     |            |     |
| Крит. Шварца         | 3382,821     | Крит. Хеннана-Куинна   | 3365,976     |            |     |

Переменная EXP (опыт работы) оказалась значимая, также как и EXP2 (квадрат опыта).



# Результаты теста Рамсея для модели с квадратами

gretl: тест Рамсея (RESET)

Вспомогательная регрессия для теста Рамсея  
МНК, использованы наблюдения 1-1574  
Зависимая переменная: lnwage

|        | Коэффициент | Ст. ошибка | t-статистика | Р-значение |    |
|--------|-------------|------------|--------------|------------|----|
| const  | 38,5041     | 16,1171    | 2,389        | 0,0170     | ** |
| EXP    | 0,127173    | 0,0601671  | 2,114        | 0,0347     | ** |
| EXP2   | -0,00310355 | 0,00146964 | -2,112       | 0,0349     | ** |
| male   | 4,32919     | 2,03775    | 2,124        | 0,0338     | ** |
| high   | 3,18968     | 1,50288    | 2,122        | 0,0340     | ** |
| yhat^2 | -0,463612   | 0,245793   | -1,886       | 0,0595     | *  |

Тестовая статистика:  $F = 3,557690$ ,  
р-значение =  $P(F(1,1568) > 3,55769) = 0,0595$

$P - value = 0.0595 > 0.05$ . Следовательно, модель правильно специфицирована, нет пропущенных степеней.





# Литература

Доугерти К. (1992). Введение в эконометрику. М. Инфра-М. Глава 6.

Демидова О. А., Малахов Д. И. (2016). ЭКОНОМЕТРИКА. Учебник и практикум для прикладного бакалавриата. М. : Юрайт. Глава 9.1, 9.2, 9.4.

Вербик М. Путеводитель по современной эконометрике. Научная книга, 2008. Глава 3.2.

Берндт, Э. Р. Практика эконометрики: классика и современность. М.: ЮНИТИ-ДАНА, 2005. - 863 с. Глава 4 (4.5С).

Борzych Д. А., Вакуленко Е. С., Фурманов К. К. Эконометрика: работа с данными на компьютере. Практикум: Элементы теории. Практические задания. Ответы и решения. Издательская группа URSS, 2021. Глава 3.

Вакуленко Е. С., Ратникова Т. А., Фурманов К. К. Эконометрика (продвинутый курс). Применение пакета Stata. М. : Юрайт, 2020. Глава 8.

