



Экономико- математическое моделирование

Магомед-Рамзан Бувайсарович,
кандидат экономических наук, доцент

Лекция 3. Некоторые обобщения множественной регрессии



План

1. Метод наименьших квадратов
2. Мультиколлинеарность. Фиктивные переменные

1. Метод наименьших квадратов (2)



Будем считать, что функция линейная: $\hat{y}_i = a + bx_i$

$e_i = y_i - \hat{y}_i = y_i - a - bx_i$ — отклонения значений этой функции ($\hat{y}_i$) от фактических наблюдений (y_i).

Идея метода наименьших квадратов: подбор параметров a и b таким образом, чтобы сумма квадратов отклонений

$$e_1^2 + e_2^2 + \dots + e_n^2 = \sum_{i=1}^n e_i^2$$

была наименьшей

МНК (продолжение)

$$S = \sum_{i=1}^n (y_i - \hat{y}_{x_i})^2 = \sum_{i=1}^n (y_i - a - b \cdot x_i)^2 \rightarrow \min$$

Необходимые условия экстремума:

$$\begin{cases} \frac{\partial S}{\partial a} = 0 \\ \frac{\partial S}{\partial b} = 0 \end{cases} \Rightarrow \begin{cases} \sum_{i=1}^n 2(y_i - a - bx_i) = 0 \\ \sum_{i=1}^n 2x_i(y_i - a - bx_i) = 0 \end{cases}$$

МНК (продолжение)



$$\left\{ \begin{array}{l} \sum_{i=1}^n y_i - na - b \sum_{i=1}^n x_i = 0 \\ \sum_{i=1}^n x_i y_i - a \sum_{i=1}^n x_i - b \sum_{i=1}^n x_i^2 = 0 \end{array} \right. \Rightarrow$$

$$\left\{ \begin{array}{l} \left(\frac{1}{n} \sum_{i=1}^n y_i \right) - a - b \left(\frac{1}{n} \sum_{i=1}^n x_i \right) = 0 \\ \left(\frac{1}{n} \sum_{i=1}^n x_i y_i \right) - a \left(\frac{1}{n} \sum_{i=1}^n x_i \right) - b \left(\frac{1}{n} \sum_{i=1}^n x_i^2 \right) = 0 \end{array} \right. \Rightarrow$$

МНК (продолжение)

Получаем систему нормальных уравнений:

$$\begin{cases} \bar{y} - a - b\bar{x} = 0 \\ \overline{xy} - a\bar{x} - b\overline{x^2} = 0 \end{cases} \Rightarrow a = \bar{y} - b\bar{x} \quad b = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{\text{cov}(x, y)}{\text{var}(x)}$$

- где $\text{cov}(x, y)$ ковариация признаков,
 - $\text{Var}(x)$ дисперсия признака x
- $$\bar{x} = \frac{1}{n} \sum x \quad \bar{y} = \frac{1}{n} \sum y \quad \overline{y \cdot x} = \frac{1}{n} \sum y \cdot x \quad \overline{x^2} = \frac{1}{n} \sum x^2$$

Интерпретация уравнения регрессии

b – коэффициент регрессии

Показывает среднее изменение результата с изменением фактора на одну единицу

a – может не иметь

экономического смысла

Остатки регрессии

y_i — фактические значения

$\hat{y}_i$ — предсказанные значения (лежат на линии регрессии)

$e_i = y_i - \hat{y}_i$ — остатки регрессии

Отметим несколько полезных свойств, связанных с остатками полученной парной регрессии:

$$(1) \sum e_i = 0 \qquad (2) \sum x_i e_i = 0$$

$$(3) \sum y_i = \sum \hat{y}_i \qquad (4) Cov(e, \hat{y}) = 0$$



Качество оценивания (1)



$$\hat{y}_i = a + bx_i, \quad e_i = y_i - \hat{y}_i \Rightarrow$$

$$y_i = e_i + \hat{y}_i$$

По правилам вычисления дисперсии:

$$Var(y) = Var(e + \hat{y}) = Var(e) + Var(\hat{y}) + 2Cov(e, \hat{y}),$$

Так как $Cov(e, \hat{y}) = 0$, то $Var(y) = Var(e) + Var(\hat{y})$

Качество оценивания (2)

$$\sum (y - \bar{y})^2 = \sum (\hat{y}_x - \bar{y})^2 + \sum (y - \hat{y}_x)^2$$

Общая сумма квадратов отклонений	=	Сумма квадратов отклонений объясненная регрессией	+	Остаточная сумма квадратов отклонений
TSS		RSS		ESS

TSS – total sum of squares,
RSS – regression sum of squares,
ESS – error sum of squares.



Коэффициент детерминации R^2

Коэффициентом детерминации или **долей**, «объясненной» уравнением дисперсии называется величина:

$$R^2 = \frac{RSS}{TSS} = 1 - \frac{ESS}{TSS}.$$

$$0 \leq R^2 \leq 1$$

О чем говорит и о чем не говорит R^2



- Высокий R^2 сам по себе не гарантирует, что модель является хорошей.
Остается риск *ложной регрессии*
- Низкий R^2 говорит о том, что существуют важные факторы, которые мы не учли в нашей модели

Вывод по F-критерию

- $F_{\text{факт}} > F_{\text{табл}}$ - H_0 отклоняется
- $F_{\text{факт}} < F_{\text{табл}}$ - уравнение регрессии считается статистически незначимым и H_0 не отклоняется.
- Величина F-критерия связана с коэффициентом детерминации

$$F = \frac{R^2}{1 - R^2} \cdot (n - 2)$$

Средняя ошибка аппроксимации

- Средняя ошибка аппроксимации - среднее отклонение расчетных значений от фактических:

$$\bar{A} = \frac{1}{n} \sum \left| \frac{y - \hat{y}}{y} \right| \cdot 100\%$$

- Допустимый предел значений - не более 8-10%.

Пример 1: По группе предприятий, выпускающих один и тот же вид продукции, рассматривается функция издержек.

№ Предприят ия	Выпуск продукции, тыс. ед. (x)	Затраты на производство, млн руб. (y)	yx	x ²	y ²	$\sum y_x$
1	1	30	30	1	900	31,1
2	2	70	140	4	4900	67,9
3	4	150	600	16	22500	141,6
4	3	100	300	9	10000	104,7
5	5	170	850	25	28900	178,4
6	3	100	300	9	10000	104,7
7	4	150	600	16	22500	141,6
Итого	22	770	2820	80	99700	770,0

Пример (продолжение)

- Система нормальных уравнений будет иметь вид:

$$\begin{cases} 7a + 22b = 770 \\ 22a + 80b = 2820 \end{cases}$$

$$b = \frac{\overline{yx} - \bar{y} \cdot \bar{x}}{\overline{x^2} - \bar{x}^2} = \frac{402,86 - 110 \cdot 3,14285}{11,43 - 3,14285^2} = 36,84$$

$$a = \bar{y} - b \cdot \bar{x} = 110 - 36,84 \cdot 3,14285 = -5,79$$

- Тогда: **$a = -5,79$; $b = 36,84$.**

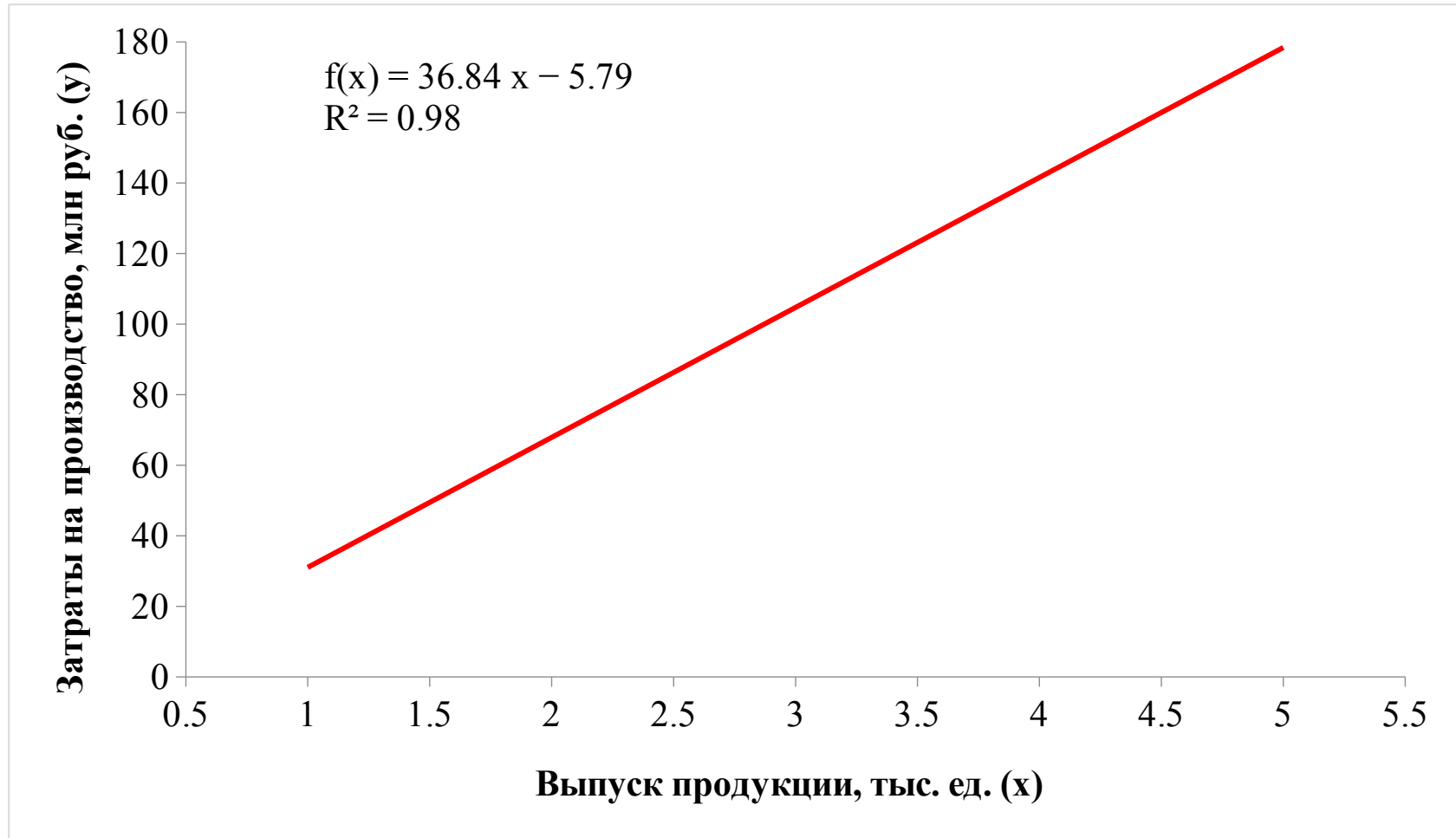
Пример (продолжение)

Уравнение регрессии:

$$\hat{y}_x = -5,79 + 36,84x$$

С увеличением выпуска продукции на 1 тыс. ед. затраты на производство возрастает в среднем на 36,84 млн. руб.

Общий чертёж диаграммы рассеяния и графика уравнения регрессии



Вычислим коэффициенты линейной парной корреляции

$$r_{xy} = \frac{Cov(x, y)}{\sqrt{Var(x) \cdot Var(y)}} = \frac{57,147}{\sqrt{1,55102 \cdot 2142,8571}} = 0,991$$

Следовательно, $R^2 = r_{yx}^2 = 0,9825$

Это означает, что 98% вариации затраты на производство (y) объясняется вариацией фактора x – выпуска продукции.

Качество модели определяет средняя ошибка аппроксимации

$$\bar{A} = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_x}{y} \right| * 100\% = \frac{0,3247}{7} 100\% \approx 4,5959\%$$

Качество построенной модели оценивается как хорошее, так как не $\bar{A}$ превышает 8-10%.

Оценку значимости уравнения регрессии в целом проведем с помощью F-критерия Фишера

- общая сумма квадратов

$$\sum (y - \bar{y})^2 = \sum y^2 - n \cdot \bar{y}^2 = 99700 - 7 \cdot 110^2 = 15000$$

- факторная сумма квадратов

$$\sum (\hat{y}_x - \bar{y})^2 = b^2 \cdot \sum (x - \bar{x})^2 = 36,84^2 \cdot (80 - 7(22/7)^2) = 14735$$

- остаточная сумма квадратов

$$\sum (y - \hat{y}_x)^2 = 15000 - 14735 = 265$$

Пример (продолжение)

$$D_{\text{факт}} = 14735$$

$$D_{\text{остат}} = 265 / 5 = 53$$

$$F = 14735 / 53 = 278$$

$$F_{\alpha=0,05} = 6,61; F_{\alpha=0,01} = 16,26$$

Вывод: уравнение регрессии значимо, так как

$$F_{\text{факт}} = 280 > F_{\text{табл}} = 6,6079$$

Случайные ошибки параметров a и b и коэффициента корреляции r_{xy}

Определим случайные ошибки m_a m_b $m_{r_{xy}}$

$$m_a = \sqrt{\frac{1}{\sum (x_i - \bar{x}_i)^2} \cdot \frac{\sum (y_i - \hat{y}_i)^2}{n - 2}} = \sqrt{\frac{1}{10,8571} \cdot \frac{263,1579}{7 - 2}} \approx 2,2017$$

$$m_b = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n - 2} \cdot \frac{\sum x_i^2}{n \sum (x_i - \bar{x}_i)^2}} = \sqrt{\frac{263,1579}{7 - 2} \cdot \frac{80}{7 \cdot 10,8571}} \approx 7,4432$$

$$m_{r_{xy}} = \sqrt{\frac{1 - r_{xy}^2}{n - 2}} = \sqrt{\frac{1 - 0,9912^2}{7 - 2}} \approx 0,0592$$

t-статистики Стьюдента

- Табличное значение t -критерия для числа степеней свободы
- $t_{\text{табл}} \approx 2,5607$, так как $df = n - 2 = 7 - 2 = 5$ и $\alpha = 0,05$.
- Тогда

$$t_a = \frac{a}{m_a} = \frac{36,8421}{2,2017} \approx 16,7332$$

$$t_b = \frac{b}{m_b} = \frac{- 5,7895}{7,4432} \approx - 0,7778$$

$$t_{r_{xy}} = \frac{r_{xy}}{m_{r_{xy}}} = \frac{0,9912}{0,0591} \approx 16,7332$$

Пример (продолжение)

Фактические и табличные значения t -статистики:

$$t_a = 16,7332 > t_{табл} = 2,5607$$

$$t_b = -0,7778 < t_{табл} = 2,5607$$

$$t_{r_{xy}} = 16,7332 > t_{табл} = 2,5607$$

Следовательно, параметры: a , b и r_{xy} не случайно отличаются от нуля, а статистически значимы.

2. Мультиколлинеарность. Фиктивные переменные



План

1. Мультиколлинеарность
2. Фиктивные переменные

Мультиколлинеарность

Виды мультиколлинеарности

Строгая (полная)

Частичная

Мультиколлинеарность

Строгая мультиколлинеарность — наличие линейной функциональной связи между объясняющими переменными.

Оценка коэффициентов модели при помощи метода наименьших квадратов **невозможна!**

Мультиколлинеарность

Возникновение строгой мультиколлинеарности означает, что на начальном этапе набор независимых переменных был сформирован некорректно.

Пример: включение в правую часть одного уравнения экспорта, импорта и чистого экспорта.

Мультиколлинеарность



Частичная мультиколлинеарность — наличие сильной линейной корреляционной связи между регрессорами.

Частичная мультиколлинеарность:

— Основное негативное последствие — стандартные ошибки оценок коэффициентов оказываются высокими.

Точность оценивания **оказывается низкой**

— Оценки коэффициентов остаются **несмещенными**

— Та или иная степень корреляции между регрессорами существует всегда. Проблема возникает только когда эта линейная связь проявляется слишком сильно.

Выявление м/к на начальном этапе моделирования (до регрессии)



- (1) Большие (по абсолютной величине) парные коэффициенты корреляции между независимыми переменными
- (2) Высокие (>10) значения коэффициента VIF

Коэффициент **VIF (variance inflation factor)** характеризует силу м/к. Вычисляется на основе значений R^2 во вспомогательных регрессиях одного регрессора на другие:

$$x_i^{(k)} = \beta_1 + \beta_2 x_i^{(2)} + \dots + \beta_{k-1} x_i^{(k-1)} + u_i$$

$$VIF = \frac{1}{1 - R^2}$$

Выявление м/к в построенной модели («симптомы» м/к)

- Небольшое изменение исходных данных, приводит к существенному изменению оценок коэффициентов
- Каждая переменная в отдельности является незначимой, а уравнение в целом имеет высокий R^2 и является значимым
- Оценки коэффициентов имеют неправильные с точки зрения экономической теории знаки или неоправданно большие значения

Устранение мультиколлинеарности

- Увеличить число наблюдений
- Исключить переменную, с которой связана мультиколлинеарность.
Следует помнить, что иногда это может привести к более серьезным проблемам
- Использовать нелинейные формы зависимостей
- Использовать агрегаты: линейные комбинации переменных

Фиктивные переменные

Фиктивные переменные — бинарные переменные (принимают значения 0 или 1).

Используются для моделирования качественных признаков.

Фиктивные переменные. Пример

Существует ли дискриминация на рынке труда?

- зарплата, долларов в час
- стаж работы, лет

Фиктивные переменные. Пример



(все переменные оказались значимы)

Мужчины ():

Женщины ():

Интерпретация:

При прочих равных условиях (в данном случае при равном стаже работы) женщины получают на 3,5 доллара в час меньше, чем мужчины.

Литература к лекции

- **Магнус Я.Р., Катышев П.К., Пересецкий А.А.** Эконометрика. Начальный курс: Учеб. — 6-е изд., перераб. и доп. — М.: Дело, 2004. **Глава 4.**
- **Доугерти К.** Введение в эконометрику: Учебник. 3-е изд. / Пер. с англ. — М.: ИНФРА-М, 2009. **Главы 4–5.**