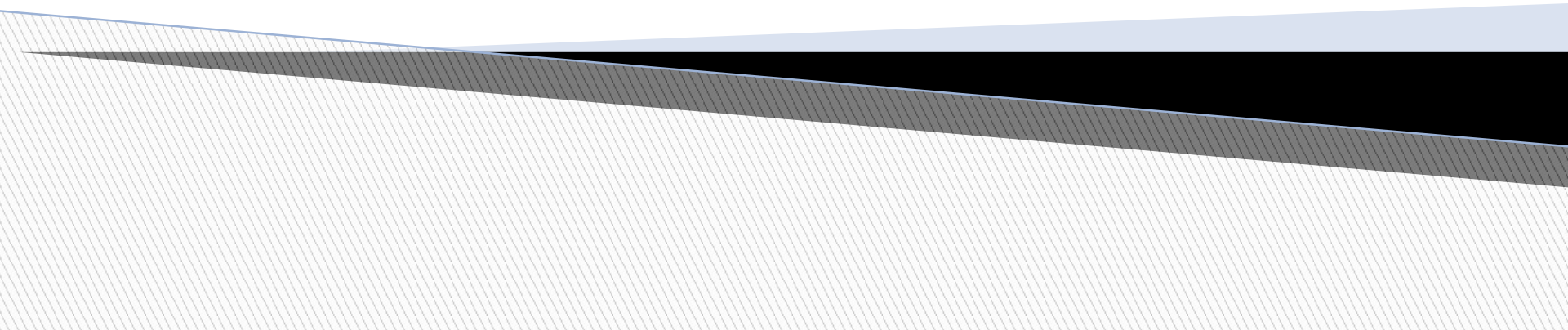
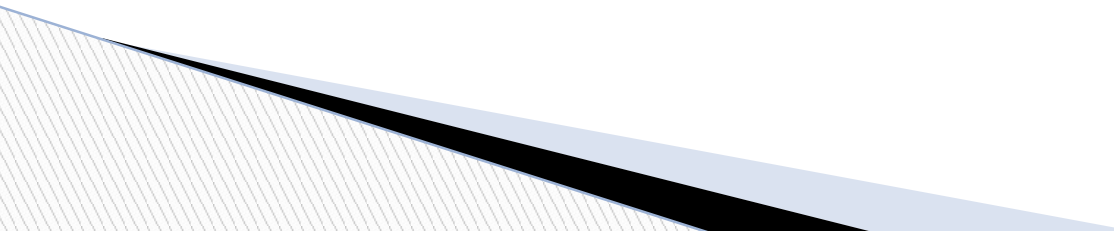


# Эконометрика

Семинар 7



# Задачи регрессионного анализа

1. установление формы зависимости между переменными;
  2. оценка модельной функции (модельного уравнения) регрессии;
  3. оценка неизвестных значений (прогноз значений) зависимой переменной.
- 

# Общий вид регрессионной модели

Здесь и далее строчными латинскими буквами с подстрочными индексами будем обозначать скалярные величины (переменные), а заглавными латинскими или строчными греческими без индексов — векторы и матрицы.

$$y = f(X) + \varepsilon$$

$y$  — функция отклика, результативный признак, предсказываемая, объясняемая переменная.

$X$  — одна или несколько объясняющих, предсказывающих переменных, факторных признаков, регрессоров.

$\varepsilon$  — возмущение, случайная ошибка модели.

# Метод наименьших квадратов (МНК).

## Теорема Гаусса-Маркова

Рассмотрим обобщенную модель линейной регрессии с  $p$  объясняющими переменными:

$$Y = \alpha X + \varepsilon.$$

$\alpha = (\alpha_0, \alpha_1, \dots, \alpha_p)^T$  — вектор параметров модели.

$\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)$  — вектор случайных ошибок модели.

$X$  — матрица размерности  $n \times (p + 1)$  наблюдений над объясняющими переменными.

$Y$  — вектор наблюдений над зависимой переменной.

Для оценки модели требуется найти вектор  $A = (a_0, a_1, \dots, a_n)$  оценок параметров  $\alpha_0, \alpha_1, \dots, \alpha_n$ .

## Предпосылки регрессионного анализа.

1. Зависимая переменная  $y_i$  есть величина случайная, а объясняющая переменная  $x_i$  есть величина неслучайная.
2. Математическое ожидание возмущения  $\varepsilon_i$  равно нулю:  
 $M\varepsilon_i = 0$ .
3. Дисперсия возмущений  $\varepsilon_i$  постоянна для любого  $i$ :  
 $D\varepsilon_i = \sigma_\varepsilon^2$ .
4. Возмущения  $\varepsilon_i$  и  $\varepsilon_j$  не коррелированы:  $\text{cov}(\varepsilon_i, \varepsilon_j) = 0$ .
5. Зависимая переменная  $y_i$  есть нормально распределенная случайная величина.

# Теорема Гаусса-Маркова

*Если условия 1—5 выполняются, то наилучшей линейной процедурой оценки линейной регрессионной модели является процедура*

$$A = (X^T X)^{-1} X^T Y,$$

*удовлетворяющая методу наименьших квадратов.*

# Суть метода наименьших квадратов

Вектор  $A$  минимизирует сумму квадратов отклонений наблюдаемых значений  $y_i$  от модельных значений  $\hat{y}_i = a_0 + a_1x_{i1} + \dots + a_px_{ip}$ , т. е. квадратичную форму:

$$Q = (Y - XA)^T (Y - XA) = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \rightarrow \min.$$

Решая систему

$$\begin{cases} Q'_{a_0} = 0, \\ Q'_{a_1} = 0, \\ \dots, \\ Q'_{a_p} = 0 \end{cases}$$

получают вектор

$$A = (X^T X)^{-1} X^T Y.$$

# Модель парной линейной регрессии

Рассмотрим простейшую модель регрессионного анализа, когда функция  $f(x)$  линейна как по параметрам, так и по переменным  $x_i$  ( $i = \overline{1, n}$ ):

$$y_i = \alpha_0 + \alpha_1 x_i + \varepsilon_i.$$

Оценкой линейной модели по выборке является уравнение регрессии  $\hat{y}_i = a_0 + a_1 x_i$ .

Здесь и далее параметры моделей будем обозначать греческими буквами ( $\alpha, \beta, \gamma$  и т. д.), а их оценки — соответствующими латинскими ( $a, b, c$  и т. д.).

Параметры  $a_0$  и  $a_1$  определяются при помощи метода наименьших квадратов.



## Оценивание модели парной линейной регрессии (минимизация квадратичной формы)

Минимизируем квадратичную форму  $Q$ :

$$Q = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - a_0 - a_1 x_i)^2 \rightarrow \min,$$

для чего решим следующую систему методом Крамера.

$$\begin{cases} Q'_{a_0} = -2 \sum_{i=1}^n (y_i - a_0 - a_1 x_i) = 0; \\ Q'_{a_1} = -2 \sum_{i=1}^n ((y_i - a_0 - a_1 x_i) x_i) = 0. \end{cases}$$

## Оценивание модели парной линейной регрессии (нормальная система уравнений)

Раскрыв скобки, получаем *нормальную систему уравнений* для модели парной линейной регрессии:

$$\begin{cases} \sum_{i=1}^n y_i = na_0 + a_1 \sum_{i=1}^n x_i, \\ \sum_{i=1}^n x_i y_i = a_0 \sum_{i=1}^n x_i + a_1 \sum_{i=1}^n x_i^2. \end{cases}$$

Повторите метод Крамера!

# Оценивание модели парной линейной регрессии (применение метода Крамера)

$$\theta = \begin{vmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{vmatrix},$$

$$\theta_0 = \begin{vmatrix} \sum_{i=1}^n y_i & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i y_i & \sum_{i=1}^n x_i^2 \end{vmatrix},$$

$$\theta_1 = \begin{vmatrix} n & \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i y_i \end{vmatrix}.$$

# Оценивание модели парной линейной регрессии (нахождение оценок модели)

Итак,

$$a_0 = \frac{\theta_0}{\theta} = \frac{\sum_{i=1}^n y_i \sum_{i=1}^n x_i^2 - \sum_{i=1}^n x_i \sum_{i=1}^n x_i y_i}{n \sum_{i=1}^n x_i^2 - \left( \sum_{i=1}^n x_i \right)^2};$$
$$a_1 = \frac{\theta_1}{\theta} = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - \left( \sum_{i=1}^n x_i \right)^2}.$$

# Пример

Вернемся к примеру, который мы детально анализировали при освоении темы «Корреляционный анализ». Пусть имеются наблюдения над валовыми выбросами загрязняющих веществ, отходящих от стационарных источников

**Задача.** Имеются наблюдения над следующими показателями экономического развития субъектов СФО в 2007 г.:

- $x_1$ : Выбросы загрязняющих веществ, отходящих от стационарных источников, тыс. т.
- $x_2$  : Добыча полезных ископаемых, т.
- $x_3$  : Производство электроэнергии, млрд кВт · ч.

# Пример

Построить модель парной линейной регрессии влияния объема производства электроэнергии на объемы выбросов загрязняющих веществ, отходящих от стационарных источников, если известно, что

$$\sum_i x_{1i} = 5\,808;$$

$$\sum_i x_{3i} = 205;$$

$$\sum_i x_{1i}x_{3i} = 223\,451;$$

$$\sum_i x_{1i}^2 = 8\,938\,184;$$

$$\sum_i x_{3i}^2 = 8\,311.$$

# Решение

Решение. Задача оценки модели парной линейной регрессии сводится к вычислению оценок  $a_0 = \theta_0/\theta$  и  $a_1 = \theta_1/\theta$ .

Вычислим сначала знаменатель  $\theta$ , единый для обеих оценок.

$$\theta = 14 \cdot 8\,311 - (205)^2 = 74\,329.$$

Теперь вычислим  $a_0$  и  $a_1$ :

$$\begin{aligned} a_0 &= \frac{\sum x_{1i} \sum x_{3i}^2 - \sum x_{3i} \cdot \sum x_{1i} x_{3i}}{\theta} = \\ &= \frac{5\,808 \cdot 8\,311 - 205 \cdot 223\,451}{74\,329} = 33,13; \end{aligned}$$

$$\begin{aligned} a_1 &= \frac{n \sum x_{1i} x_{3i} - \sum x_{3i} \cdot \sum x_{1i}}{\theta} = \\ &= \frac{14 \cdot 223\,451 - 205 \cdot 5\,808}{74\,329} = 26,07; \end{aligned}$$

# Решение

Таким образом, оценкой модели парной линейной регрессии влияния объема производства электроэнергии на объемы выбросов загрязняющих веществ, отходящих от стационарных источников, в данном случае является уравнение

$$\hat{x}_1 = 33,13 + 26,07x_3.$$



# Оценивание модели множественной линейной регрессии

Как мы говорили ранее, в общем случае, т. е. для множественной линейной регрессии, оценка запишется так:

$$A = (X^T X)^{-1} X^T Y.$$

Вычисление данной матрицы не представляет трудности, если хорошо владеть приемами вычисления обратной матрицы.

# Алгоритм нахождения обратной матрицы

Важно помнить, что матрица является обратимой тогда и только тогда, когда она является невырожденной, то есть ее определитель не равняется нулю.

$$\text{Пусть } Z = \begin{pmatrix} z_{11} & z_{12} & \dots & z_{1n} \\ z_{21} & z_{22} & \dots & z_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ z_{n1} & z_{n2} & \dots & z_{nn} \end{pmatrix}.$$

Тогда

$$Z^{-1} = \frac{1}{|Z|} \cdot \begin{pmatrix} c_{11} & c_{12} & \dots & c_{1n} \\ c_{21} & c_{22} & \dots & c_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ c_{n1} & c_{n2} & \dots & c_{nn} \end{pmatrix}^T,$$

где  $c_{ij}$  — алгебраическое дополнение к  $(i, j)$ -му элементу матрицы  $Z$ .

# Задача (самостоятельно!)

Задача. Вычислить обратные матрицы к нижеследующим:

$$\begin{pmatrix} 1 & -1 & 0 \\ 6 & 3 & 2 \\ 1 & 7 & 2 \end{pmatrix},$$

$$\begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \end{pmatrix}.$$

# Оценивание модели множественной линейной регрессии. Пример

Задача. Оценить модель множественной линейной регрессии по матрице

$$X = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 5 & 2 \\ 1 & 6 & 1 \end{pmatrix}$$

и вектору

$$Y = (1, 2, 3)^T.$$

# Решение

**Решение.** Напомним, что оценка модели множественной линейной регрессии записывается в матричной форме следующим образом:

$$A = (X^T X)^{-1} X^T Y.$$

Транспонируем матрицу  $X$ :

$$X^T = \begin{pmatrix} 1 & 1 & 1 \\ 2 & 5 & 6 \\ 3 & 2 & 1 \end{pmatrix}.$$

Вычислим произведение матрицы  $X^T$  на матрицу  $X$ :

$$X^T X = \begin{pmatrix} 1 & 1 & 1 \\ 2 & 5 & 6 \\ 3 & 2 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 2 & 3 \\ 1 & 5 & 2 \\ 1 & 6 & 1 \end{pmatrix} = \begin{pmatrix} 3 & 13 & 6 \\ 13 & 65 & 22 \\ 6 & 22 & 14 \end{pmatrix}.$$

# Решение

Теперь вычислим обратную матрицу (множитель  $1/2$  для удобства восприятия вынесен за матрицу):

$$(X^T X)^{-1} = \frac{1}{2} \cdot \begin{pmatrix} 213 & -25 & -52 \\ -25 & 3 & 6 \\ -52 & 6 & 13 \end{pmatrix}.$$

Вычислим произведение матрицы  $X^T$  на вектор  $Y$ :

$$X^T Y = \begin{pmatrix} 1 & 1 & 1 \\ 2 & 5 & 6 \\ 3 & 2 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} = \begin{pmatrix} 6 \\ 30 \\ 10 \end{pmatrix}.$$

# Решение

Итак,

$$\begin{aligned} A = (X^T X)^{-1} X^T Y &= \frac{1}{2} \cdot \begin{pmatrix} 213 & -25 & -52 \\ -25 & 3 & 6 \\ -52 & 6 & 13 \end{pmatrix} \cdot \begin{pmatrix} 6 \\ 30 \\ 10 \end{pmatrix} = \\ &= \begin{pmatrix} 4 \\ 0 \\ -1 \end{pmatrix}. \end{aligned}$$

Таким образом, итоговая оценка выглядит следующим образом:

$$\hat{y} = 4 + 0 \cdot x_1 - x_2 = 4 - x_2.$$