

Эконометрика. Лекция 1

12 ноября 2014 г.

Эконометрика на одном слайде :)

Вопросы:

- Как устроен мир? Как переменная x влияет на переменную y ?
- Что будет завтра? Как спрогнозировать переменную y ?

Ответ:

Модель — формула для объясняемой переменной

Например:

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

Основные типы данных:

- Временные ряды
Год Население Безработица
2010 142962 7.4
2011 142914 6.5
2012 143103 5.5
2013 143395 5.5
- Перекрестная выборка
Страна Золото Серебро Бронза
Россия 13 11 9
Норвегия 11 5 10
Канада 10 10 5
США 9 7 12
- Панельные данные сочетание первых двух
- ...

Данные — обозначения

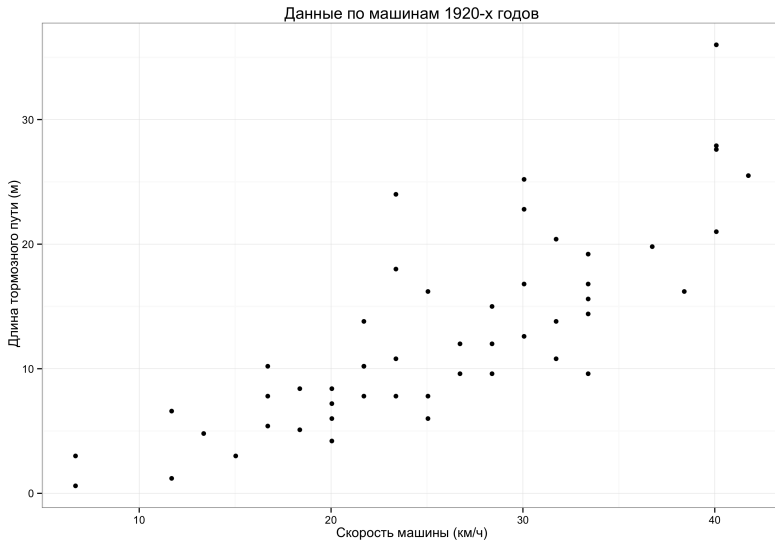
- Одна зависимая, объясняемая, переменная: y
- Несколько регрессоров, объясняющих, переменных: x, z, \dots
- По каждой переменной n наблюдений: y_1, y_2, \dots, y_n

Данные — пример

Исторические данные 1920-х годов :)

| Длина тормозного пути (м), y_i | Скорость машины (км/ч), x_i |
|----------------------------------|-------------------------------|
| 0.6 | 6.68 |
| 3.0 | 6.68 |
| 1.2 | 11.69 |
| ... | ... |

Всегда изображайте данные!



Модель:

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

- Наблюдаемые переменные: y , x
- Неизвестные параметры: β_1 , β_2
- Случайная составляющая, ошибка: ε

План действий

- придумать адекватную модель
- получить оценки неизвестных параметров: $\hat{\beta}_1$, $\hat{\beta}_2$
- прогнозировать, заменив неизвестные параметры на оценки:

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

Метод наименьших квадратов

- Способ получить оценки неизвестных параметров модели исходя из реальных данных.

Ошибка прогноза: $\hat{\varepsilon}_i = y_i - \hat{y}_i$.

Сумма квадратов ошибок прогноза:

$$Q(\hat{\beta}_1, \hat{\beta}_2) = \sum_{i=1}^n \hat{\varepsilon}_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Суть МНК: В качестве оценок взять такие $\hat{\beta}_1, \hat{\beta}_2$, при которых сумма квадратов ошибок прогноза, Q , минимальна.

Пример с машинами:

Фактические данные:

$$x_1 = 6.68, x_2 = 6.68, \dots,$$

$$y_1 = 0.6, y_2 = 3, \dots$$

Модель: $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$. Формула для прогнозов: $\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i$

Сумма квадратов ошибок прогнозов: $Q = \sum_{i=1}^n (y_i - \hat{y}_i)^2$

$$Q = (0.6 - \hat{\beta}_1 - \hat{\beta}_2 6.68)^2 + (3 - \hat{\beta}_1 - \hat{\beta}_2 6.68)^2 + \dots$$

Точка минимума, найдена в R: $\hat{\beta}_1 = -5.3, \hat{\beta}_2 = 0.7$:

Формула для прогнозов: $\hat{y}_i = -5.3 + 0.7x_i$

Простой пример (чудо-доска)

| Имя | Вес (кг), y_i | Рост (см), x_i |
|------|-----------------|------------------|
| Вася | 60 | 170 |
| Коля | 70 | 170 |
| Петя | 80 | 181 |

Оцените модели:

$$y_i = \beta x_i + \varepsilon_i$$

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

Маленькая подготовка $n\bar{x} = \sum_i x_i = \sum_i \bar{x}$

Готовые формулы МНК

В модели $y_i = \beta + \varepsilon_i$

$$\hat{\beta} = \bar{y}$$

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

$$\hat{\beta}_2 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x}$$

точка (\bar{x}, \bar{y}) лежит на прямой $\hat{y} = \hat{\beta}_1 + \hat{\beta}_2 x$

Терминология и обозначения:

y_i — зависимая, объясняемая, переменная

x_i — регрессор, объясняющая переменная

ε_i — ошибка, ошибка модели, случайная составляющая

\hat{y}_i — прогноз, прогнозное значение

$\hat{\varepsilon}_i = y_i - \hat{y}_i$ — остаток, ошибка прогноза

$RSS = \sum_{i=1}^n \hat{\varepsilon}_i^2$ — сумма квадратов остатков

Графическая иллюстрация (чудо-доска)

!показать, что регрессия проходит через среднюю точку

Много объясняющих переменных

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

Выпишем систему уравнений для оценок $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$

Чудо доска

Оценки находятся из системы

$$\begin{cases} \sum \hat{\varepsilon}_i \cdot 1 = 0 \\ \sum \hat{\varepsilon}_i \cdot x_i = 0 \\ \sum \hat{\varepsilon}_i \cdot z_i = 0 \end{cases}$$

Суммы квадратов

сумма квадратов остатков, $RSS = \sum \hat{\varepsilon}_i^2$

общая сумма квадратов, $TSS = \sum (y_i - \bar{y})^2$

объясненная сумма квадратов, $ESS = \sum (\hat{y}_i - \bar{y})^2$

Абсолютный ликбез по линейной алгебре

Обозначения:

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \quad \hat{\varepsilon} = \begin{pmatrix} \hat{\varepsilon}_1 \\ \hat{\varepsilon}_2 \\ \vdots \\ \hat{\varepsilon}_n \end{pmatrix} \quad \vec{1} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$$

В нашей модели: $\hat{y} = \hat{\beta}_1 \cdot \vec{1} + \hat{\beta}_2 \cdot x + \hat{\beta}_3 \cdot z$

$$X = \begin{pmatrix} 1 & x_1 & z_1 \\ 1 & x_2 & z_2 \\ \vdots & & \\ 1 & x_n & z_n \end{pmatrix}$$

Длина вектора

Длина вектора, $|y| = \sqrt{y_1^2 + y_2^2 + \dots + y_n^2}$

Квадрат длины вектора, $|y|^2 = y_1^2 + y_2^2 + \dots + y_n^2 = \sum_i y_i^2$

Примеры:

$RSS = \sum \hat{\varepsilon}_i^2$ — квадрат длины вектора $\hat{\varepsilon}$

$TSS = \sum (y_i - \bar{y})^2$ — квадрат длины вектора $(y - \bar{y} \cdot \vec{1})$

$$\begin{pmatrix} y_1 - \bar{y} \\ y_2 - \bar{y} \\ \vdots \\ y_n - \bar{y} \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} - \bar{y} \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} = y - \bar{y} \cdot \vec{1}$$

Скалярное произведение двух векторов:

$$(x, y) = |x| \cdot |y| \cdot \cos(\angle x, y)$$

$$(x, y) = x_1y_1 + x_2y_2 + \dots + x_ny_n = \sum_i x_iy_i$$

Условие перпендикулярности: $x \perp y$, $\sum_i x_iy_i = 0$, т.к. $\cos(90^\circ) = 0$.

Картинка для модели $y_i = \beta + \varepsilon_i$

Теорема Пифагора

Геометрическая интерпретация условий первого порядка

$$\begin{cases} \sum \hat{\varepsilon}_i \cdot 1 = 0 \\ \sum \hat{\varepsilon}_i \cdot x_i = 0 \\ \sum \hat{\varepsilon}_i \cdot z_i = 0 \end{cases} \Leftrightarrow \begin{cases} \hat{\varepsilon} \perp \vec{1} \\ \hat{\varepsilon} \perp x \\ \hat{\varepsilon} \perp z \end{cases}$$

Чудо-доска. Картинка для множественной регрессии (с TSS и RSS)

Если в регрессию включён свободный член (β_1)

Если в регрессию включён свободный член ($y_i = \beta_1 + \dots$) и оценки МНК единственны, то:

- $\sum \hat{\varepsilon}_i = 0$
- $\sum y_i = \sum \hat{y}_i$
- $\bar{y} = \bar{\hat{y}}$
- $TSS = RSS + ESS$

Коэффициент детерминации — простой показатель качества

В моделях со свободным членом $R^2 = ESS/TSS$

TSS — общий разброс y

ESS — объясненный регрессорами разброс

R^2 — доля объясненного разброса в общем разбросе

Теорема. Если в регрессию включён свободный член ($y_i = \beta_1 + \dots$) и оценки МНК единственны, то R^2 равен выборочной корреляции между y и \hat{y} , т.е.

$$R^2 = \left(\frac{\sum (y_i - \bar{y})(\hat{y}_i - \bar{y})}{\sqrt{\sum (y_i - \bar{y})^2} \sqrt{\sum (\hat{y}_i - \bar{y})^2}} \right)^2$$

Чудо-доска (доказательство)

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

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad X = \begin{pmatrix} 1 & x_1 & z_1 \\ 1 & x_2 & z_2 \\ \vdots & & \\ 1 & x_n & z_n \end{pmatrix}$$

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

Чудо-доска вывод формулы

Если слишком много видео, то уберём вывод формулы из видео.

УРА!!! МНК позволяет оценивать модели!!!

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

Получаем $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$

Вопросы

- Как выбрать форму модели?
- А будет ли решение задачи минимизации единственным?
- А будет ли решение задачи минимизации вообще существовать?
- А почему сумма квадратов остатков, а не, скажем, модулей?
- А насколько точны полученные оценки?
- ...