Эконометрика. Лекция 1

7 октября 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:

- $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$
- $y_t = y_{t-1} + \varepsilon_t$

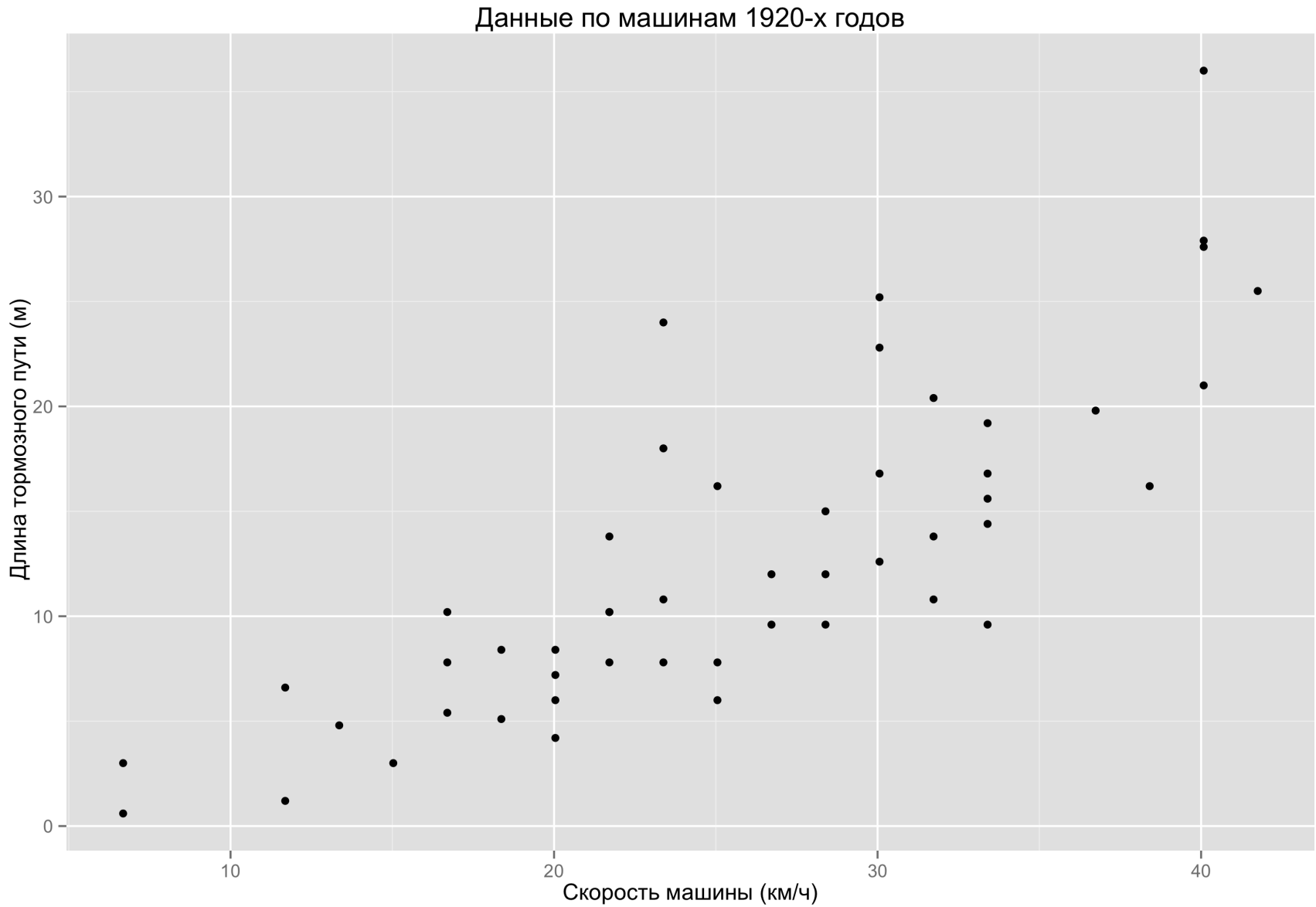
Данные:

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

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

Исторические данные 1920-х годов :)

Всегда изображайте данные!



Модель:

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

- Наблюдаемые переменные: y, x
- Неизвестные параметры: β_2, 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: $e_i = y_i - \hat{y}_i$.

Сумма квадратов ошибок прогноза:

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

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:

$x_1 = 6.68$, $x_2 = 6.68$, . . . ,

$y_1 = 0.6$, $y_2 = 3$, . . .

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

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

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

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

Простой пример

Имя	Вес (кг), y_i	Рост (см), x_i
Вася	60	170
Коля	70	170
Петя	80	180

Модель: $y_i = \beta x_i + \varepsilon_i$, Прогнозы: $\hat{y}_i = \hat{\beta} x_i$.

Сумма квадратов ошибок:

$$Q(\hat{\beta}) = (60 - \hat{\beta}170)^2 + (70 - \hat{\beta}170)^2 + (80 - \hat{\beta}180)^2$$

Решение задачи минимизации

Сумма квадратов ошибок:

$$Q(\hat{\beta}) = (60 - \hat{\beta}170)^2 + (70 - \hat{\beta}170)^2 + (80 - \hat{\beta}180)^2$$

Производная:

$$Q'(\hat{\beta}) = -2 \cdot 170 \cdot (60 - \hat{\beta}170) - 2 \cdot 170 \cdot (70 - \hat{\beta}170) - 2 \cdot 180 \cdot (80 - \hat{\beta}180) \quad (1)$$

Приравняв производную к нулю получаем:

$$\hat{\beta} = 0.4047$$

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

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

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

Три простых случая в явном виде

$$y_i = \beta + \varepsilon_i$$

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

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

Случай $y_i = \beta + \varepsilon_i$. Нет объясняющей переменной.

- Сколько лет Васе?

Анна: 35. Белла: 27. Вика: 34.

Спрогнозируем Васин возраст с помощью МНК!

Наблюдения: $y_1, y_2, \ldots, y_n$

Модель: $y_i = \beta + \varepsilon_i$. Прогнозы: $\hat{y}_i = \hat{\beta}$

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

Находим производную: $Q'(\hat{\beta}) = \sum_{i=1}^{n} -2(y_i - \hat{\beta})$

Случай $y_i = \beta + \varepsilon_i$. Решение.

Упрощаем производную:

$$Q'(\hat{\beta}) = \sum_{i=1}^{n} -2(y_i - \hat{\beta}) = -2\sum_{i=1}^{n}(y_i - \hat{\beta}) =$$

$$- 2(\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \hat{\beta}) = -2(\sum_{i=1}^{n} y_i - n\hat{\beta}) \quad (2)$$

Приравняв к нулю получаем: $\sum_{i=1}^{n} y_i = n\hat{\beta}$ или

$\hat{\beta} = \sum_{i=1}^{n} y_i / n = (y_1 + y_2 + ... + y_n)/n = \bar{y}$

МНК-прогноз возраста Васи: $\hat{\beta} = (35 + 27 + 34)/3 = 32$

Случай $y_i = \beta x_i + \varepsilon_i$. Пропорциональность.

Случай $ $. Парная регрессия.

Предварительные замечания:

$\bar{x} = \sum_{i=1}^{n} x_i$, поэтому:

$n\bar{x} = \sum_{i=1}^{n} x_i$ или $\sum_{i=1}^{n} \bar{x} = \sum_{i=1}^{n} x_i$.

Первая регрессия в R

(!) Установите R, Rstudio и дополнительные пакеты

Три режима работы с Rstudio:

- диалоговый, консольный
- написание скрипта или программы
- написание документа, “грамотное программирование”

Консольный режим

. . . тут скринкаст

Консольный режим. Резюме...

- R отличает заглавные и прописные буквы.

```
a <- 5
A <- 4
a + A
```

```
## [1] 9
```

- присваивания a <- 5 и a = 5 абсолютно равнозначны
- знак + в командной строке означает неоконченную команду
  - может означать забытую незакрытую скобку
  - избавиться можно нажатием клавиши Esc
- tab облегачает жизнь, дописывая длинные названия

Написание скрипта

(тут скринкаст)

Написание скрипта. Резюме...

- ctrl+Enter (cmd+Enter на Маке) исполняет текущую строчку или несколько строк
- два основных объекта: вектор и табличка с данными

```
x <- c(5,2,1)
d <- data.frame(rost=c(170,170,180),ves=c(60,70,80))
```

- любой реальный скрипт начинается с загрузки дополнительных пакетов

```
library("dplyr")
library("ggplot2")
```

Резюме. Загрузка данных.

Загрузка данных из электронной таблицы (Excel, Libre Office Calc, Gnumeric ...)

1. Причесать данные
2. Сохранить данные в формате csv
3. Прочитать данные в R командой

```
d <- read.table("mydata.csv")
```

Резюме. Поглядеть на табличку

Данные в табличке d.

- начало и конец таблички: head(d), tail(d)
- описание таблички: str(d)
- описательные статистики: summary(d)
- достать переменную speed из таблички: d$speed
- достать вторую строку из таблички: d[2,]
- достать второй столбец из таблички: d[,2]
- преобразовать или создать новую переменную: d <- mutate(d,speed2=speed^2)

Резюме. Два базовых графика

- Гистограмма
- Диаграмма рассеяния
- Не забывайте подписи!!!

Резюме. Простой пример регрессии

Вопросы

- 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?
- ...

Написание документа

```
library("dplyr")
library("ggplot2")
d <- cars
head(cars)
```

```
##   speed dist
## 1     4    2
## 2     4   10
## 3     7    4
## 4     7   22
## 5     8   16
## 6     9   10
```

```
# %>% mutate(dist=0.3*dist,speed=1.67*speed)
```

$\bar{y} = (y_1 + y_2 + ... + y_n)/n$ — среднее значение y

$TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$ — общая сумма квадратов

$ESS = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ — объясненная сумма квадратов