Лекция 6

Линейная регрессия

Курс: Введение в DS на УБ и МиРА (весна, 2022)
Преподаватель: Владимир Омелюсик

25 апреля 2022 г.

Моделирование

$X_1 \dots X_n \sim F_x(x)$ — модель до этого

Постановка задачи

$$\rightarrow y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + \varepsilon_i$$

- y_i — зависимая перем. (target, целев.)
- $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ — коэф-ты (константы)
- $x_{1i}, x_{2i}, \ldots, x_{ki}$ — независ. перем. (признаки, регрессоры)
- ε_i — случ. ошибка

	x_1	x_2	...	x_k	y
1	.	.	.	.	
2					
3					
⋮					

Пусть x_{ij} – константы

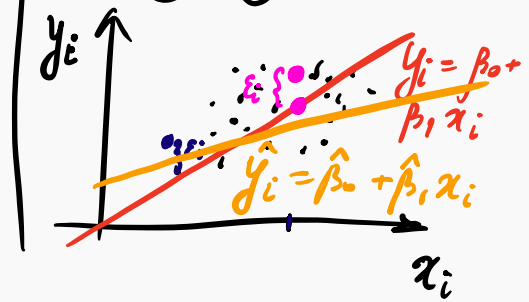
β_j – всегда константы

ε_i – сл. велич.

y_i – сл. велич.

Закон генер. данных

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$



Хотим: по выборке получить $\hat{\beta}_0$ и $\hat{\beta}_1$ – оценки β_0 и β_1

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

$\hat{\beta}_0, \hat{\beta}_1$ — сл. величины

По чему линейная линейная регрессия?

Пример:

Природа: $y_i = 3 + 4 x_i + \varepsilon_i$

- y_i — Доход
- $3 = \beta_0$
- $4 = \beta_1$
- x_i — Стаж работы (у.е.)
- ε_i — Сл. ошибка

Доход	vs	Стаж работы
300		2.5
⋮		⋮

2000 наблюд.

$\hat{\beta}_0 = 2.9$, $\hat{\beta}_1 = 4.3$

$$\hat{y}_i = 2.9 + 4.3 x_i$$

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki} + \varepsilon_i$$

1. y_i
2. β_0
3. x_{1i}
4. ε_i

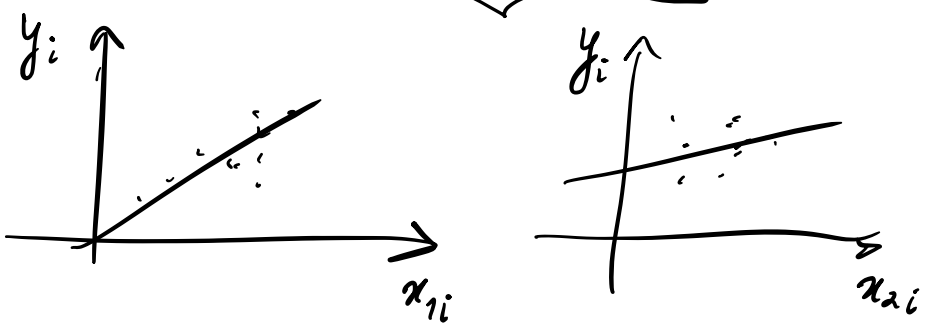
Лин. регр. лин-а по коэф-м

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad - \text{ЛР}$$

$$\ln y_i = \beta_0 + \beta_1 x_i^2 + \varepsilon_i^3 \quad - \text{ЛР}$$

$$y_i = \beta_0^2 + (\ln \beta_1) x_i + \varepsilon_i \quad - \text{не ЛР}$$

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i$$



ЛР

парная

$y_i = \beta_0^1 + \beta_1 x_i + \varepsilon_i$

2 пар.

множ

$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki} + \varepsilon_i$

≥ 2 пар.

	1	X_1	X_2	...	y
→	1				
→	1				
→	1				

Метод наименьших квадратов

$$\sum_i (y_i - \hat{y}_i)^2 \to \min_{\hat{\beta}_0, \dots, \hat{\beta}_k} \qquad \text{(МНК)}$$