ТВиМС
Основы регрессионного анализа. Занятие 1

15 января 2021

Рабочее поле на 2 курс

Рассматриваем

- cross-sectional data (пространственные)

Рабочее поле на 2 курс

Рассматриваем

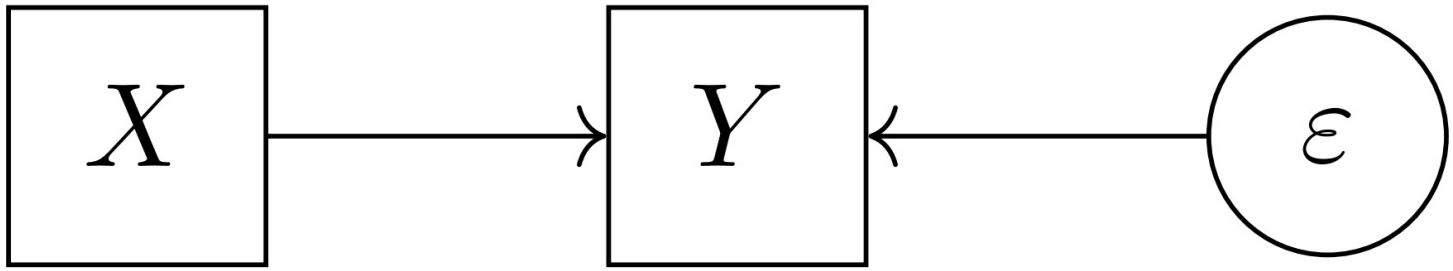
- cross-sectional data (пространственные)
- модели линейные с точки зрения коэффициентов, но при этом можем учитывать нелинейную связь X и Y

Рабочее поле на 2 курс

Рассматриваем

- cross-sectional data (пространственные)
- модели линейные с точки зрения коэффициентов, но при этом можем учитывать нелинейную связь X и Y
- интерпретация в терминах взаимосвязи зависимой переменной и предиктора, о выявлении эффекта (treatment effect) поговорим на 3-ем курсе

Путевая диаграмма: регрессия



Y – зависимая переменная (отклик);
X – независимая переменная (объясняющая переменная / предиктор);
ε – ошибка

Классическая линейная регрессия

Вопрос

Запишем спецификацию парной регрессии в общем виде.

Классическая линейная регрессия

Вопрос

Запишем спецификацию парной регрессии в общем виде.

Ответ

$y_i = b_0 + b_1 x_i + e_i$,
где y_i – зависимая переменная (отклик),
b_0 – константа (intercept),
b_1 – коэффициент при предикторе (slope coefficient),
x_i – независимая переменная (предиктор),
e_i – ошибка.

Классическая линейная регрессия

Вопрос

Запишем спецификацию парной регрессии в общем виде.

Ответ

$y_i = b_0 + b_1 x_i + e_i$,
где y_i – зависимая переменная (отклик),
b_0 – константа (intercept),
b_1 – коэффициент при предикторе (slope coefficient),
x_i – независимая переменная (предиктор),
e_i – ошибка.

$\hat{y}_i = \hat{b}_0 + \hat{b}_1 x_i$ – это предсказанное значение зависимой переменной;
$\hat{e}_i = y_i - \hat{y}_i$, где $\hat{e}_i$ – это остаток (оценка ошибки).

Классическая линейная регрессия

Вопрос

Метод наименьших квадратов (МНК) – один из методов оценивания параметров в регрессии. Покажем основной принцип этого метода.

Классическая линейная регрессия

Вопрос

Метод наименьших квадратов (МНК) – один из методов оценивания параметров в регрессии. Покажем основной принцип этого метода.

Ответ

В соответствии с МНК выбираем такие оценки коэффициентов, при которых линия предсказания наиболее близка к наблюдениям. Математически происходит минимизация суммы квадратов остатков:

$$\min \sum_{i=1}^{n} (y_i - (\hat{b}_0 + \hat{b}_1 x_i))^2.$$

Классическая линейная регрессия

Вопрос

Метод наименьших квадратов (МНК) – один из методов оценивания параметров в регрессии. Покажем основной принцип этого метода.

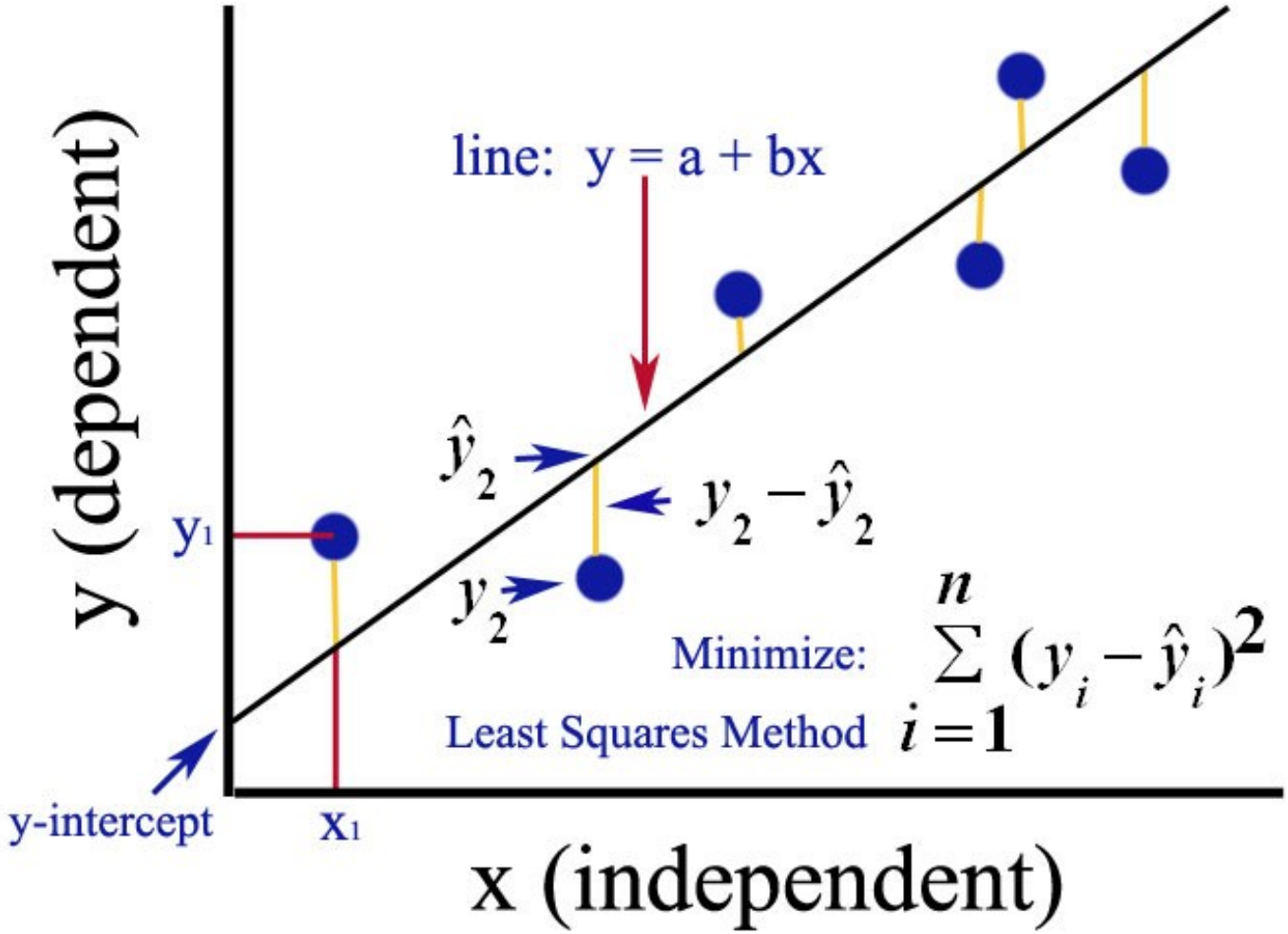
Ответ

В соответствии с МНК выбираем такие оценки коэффициентов, при которых линия предсказания наиболее близка к наблюдениям. Математически происходит минимизация суммы квадратов остатков:

$$\min \sum_{i=1}^{n} (y_i - (\hat{b}_0 + \hat{b}_1 x_i))^2.$$

Или можем переписать это в таком виде: $\min \sum^{n} (y_i - \hat{y}_i)^2$

Иллюстрация принципа МНК



Источник картинки: ссылка

Оценки в соответствии с МНК

Модель на константу

$$y_i = \beta_0 + e_i$$

$$\hat{\beta}_0 = \bar{y}$$

Оценки в соответствии с МНК

Модель на константу

$$y_i = \beta_0 + e_i$$

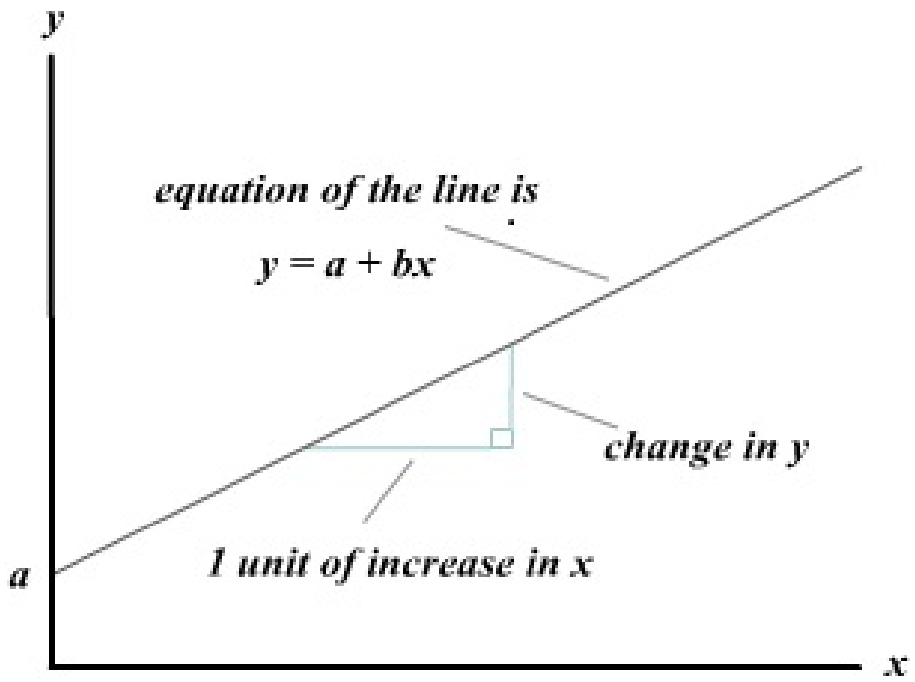
$$\hat{\beta}_0 = \bar{y}$$

Модель парной регрессии

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i(y_i - \bar{y})}{\sum_{i=1}^{n} x_i(x_i - \bar{x})} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{\widehat{Cov}(x, y)}{\widehat{Var}(x)}$$

Интерпретация оценок коэффициентов



$\hat{b}_0$ (также обозначается как a) – среднее значение отклика при условии равенства предикторов 0.
$\hat{b}_1$ – на сколько в среднем изменяется отклик при увеличении предиктора на единицу измерения при прочих равных.