Эконометрика 1
осень 2016

Лекция 2
14.09.2016

Пример: данные по результатам тестов в Калифорнии (продолжение)

$$\beta_{ClassSize} = \frac{\text{изменение } TestScore}{\text{изменение } ClassSize} = \frac{\Delta TestScore}{\Delta ClassSize} = \frac{\Delta Y}{\Delta X} \qquad (1)$$

или

$$\Delta TestScore = \beta_{ClassSize}\, \Delta ClassSize \qquad (2)$$

Пример: данные по результатам тестов в Калифорнии (продолжение)

Уравнение (1) – определение коэффициента наклона прямой, которая может быть записана

$$TestScore = \beta_0 + \beta_{ClassSize} \times ClassSize \qquad (3)$$

β_0 – константа, свободный член

$\beta_{ClassSize}$ –коэффициент наклона

Но!

$$TestScore = \beta_0 + \beta_{ClassSize} \times ClassSize + \text{другие факторы} \qquad (4)$$

Формальная модель (1)

Пусть

Y_i – среднее значение за тест в i-м школьном округе

X_i – среднее значение размера класса в i-м школьном округе

u_i – прочие факторы, влияющие на результаты обучения в i-м школьном округе

$$Y_i = \beta_0 + \beta_1 X_i + u_i \qquad (5)$$

Формальная модель (2)

Уравнение (5) – модель парной линейной регрессии или линейная модель наблюдений или линейная эконометрическая модель или линейная регрессионная модель

i – номер наблюдения ($i = 1, \dots, n$)

Y_i – зависимая переменная

X_i – независимая переменная или регрессор

u_i – случайная ошибка регрессии или ошибка i – го наблюдения

$Y = \beta_0 + \beta_1 X$ – линия (функция) теоретической регрессии (регрессии генеральной совокупности) или линейная модель связи

β_0 - свободный член (константа) линии теоретической регрессии

β_1 - коэффициент наклона линии теоретической регрессии

Пример: гипотетические данные по результатам тестов в Калифорнии

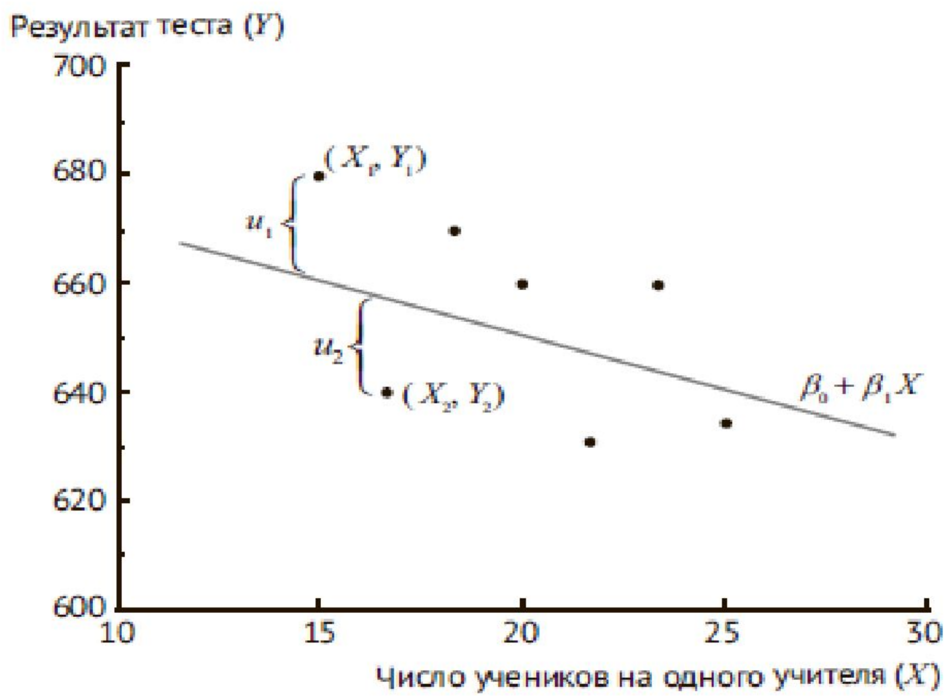


Рисунок 4.1. Диаграмма рассеяния результатов тестов относительно числа учеников в классе (гипотетические данные)

Оценка коэффициентов в модели парной линейной регрессии

Как оценить

$\beta_{ClassSize}$

или

β_1

в более общей постановке?

Что такое оценка?

Оценка (an estimator) – функция от результатов наблюдения (выборки), выбранных случайным образом из генеральной совокупности.

Оценка (an estimate) – численное значение оценки, полученной по данным из конкретной случайной выборки.

Какие бывают оценки: примеры

Пусть μ_Y - математическое ожидание Y в генеральной совокупности (обозначаем $E(Y)$).

Пусть $Y_1, Y_2, \dots, Y_n$ - выборка n независимых одинаково распределенных случайных величин (i.i.d) из рассматриваемой генеральной совокупности. Как мы можем оценить μ_Y?

Оценка 1: $\hat{\mu}_Y = \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$

Оценка 2: $\tilde{\mu}_Y = Y_1$

Оценка 3: $\bar{\bar{\mu}}_Y = \frac{1}{n}\left(\frac{1}{2}Y_1 + \frac{3}{2}Y_2 + \frac{1}{2}Y_3 + \frac{3}{2}Y_4 + \cdots + \frac{1}{2}Y_{n-1} + \frac{3}{2}Y_n\right)$, при четном n (для удобства)

Свойства оценок

Смещенность (несмещенность)

$\hat{\mu}_Y$ - несмещенная оценка μ_Y, если $E(\hat{\mu}_Y) = \mu_Y$

Смещением $\hat{\mu}_Y$ называется величина $E(\hat{\mu}_Y) - \mu_Y$

Состоятельность

$\hat{\mu}_Y$ - состоятельная оценка μ_Y, если $\hat{\mu}_Y \xrightarrow{p} \mu_Y$

Эффективность

$\hat{\mu}_Y$ и $\tilde{\mu}_Y$ - несмещенные оценки μ_Y. Тогда $\hat{\mu}_Y$ (более) эффективная чем $\tilde{\mu}_Y$, если

$$\mathrm{var}(\hat{\mu}_Y) < \mathrm{var}(\tilde{\mu}_Y)$$

МНК оценка

$$\min_{m} \sum_{i=1}^{n} (Y_i - m)^2$$

Пример: выборочное среднее – МНК оценка математического ожидания

МНК оценка коэффициентов парной линейной регрессии

$$Y_i = \beta_0 + \beta_1 X_i + u_i \qquad (5)$$

$$\min_{b_0, b_1} \sum_{i=1}^{n} \left(Y_i - b_0 - b_1 X_i\right)^2 \qquad (6)$$

$$\rightarrow \hat{\beta}_0,\ \hat{\beta}_1$$

$\hat{\beta}_0,\ \hat{\beta}_1$ - <u>МНК оценки</u> коэффициентов β_0 и β_1

$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + u_i$ - <u>МНК оценка линии регрессии</u>, (линия выборочной регрессии или функция выборочной регрессии);

$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ - <u>предсказанное</u> значение Y_i;

$\hat{u}_i = Y_i - \hat{Y}_i$ - остаток МНК регрессии

МНК оценка коэффициентов парной линейной регрессии

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X}$$

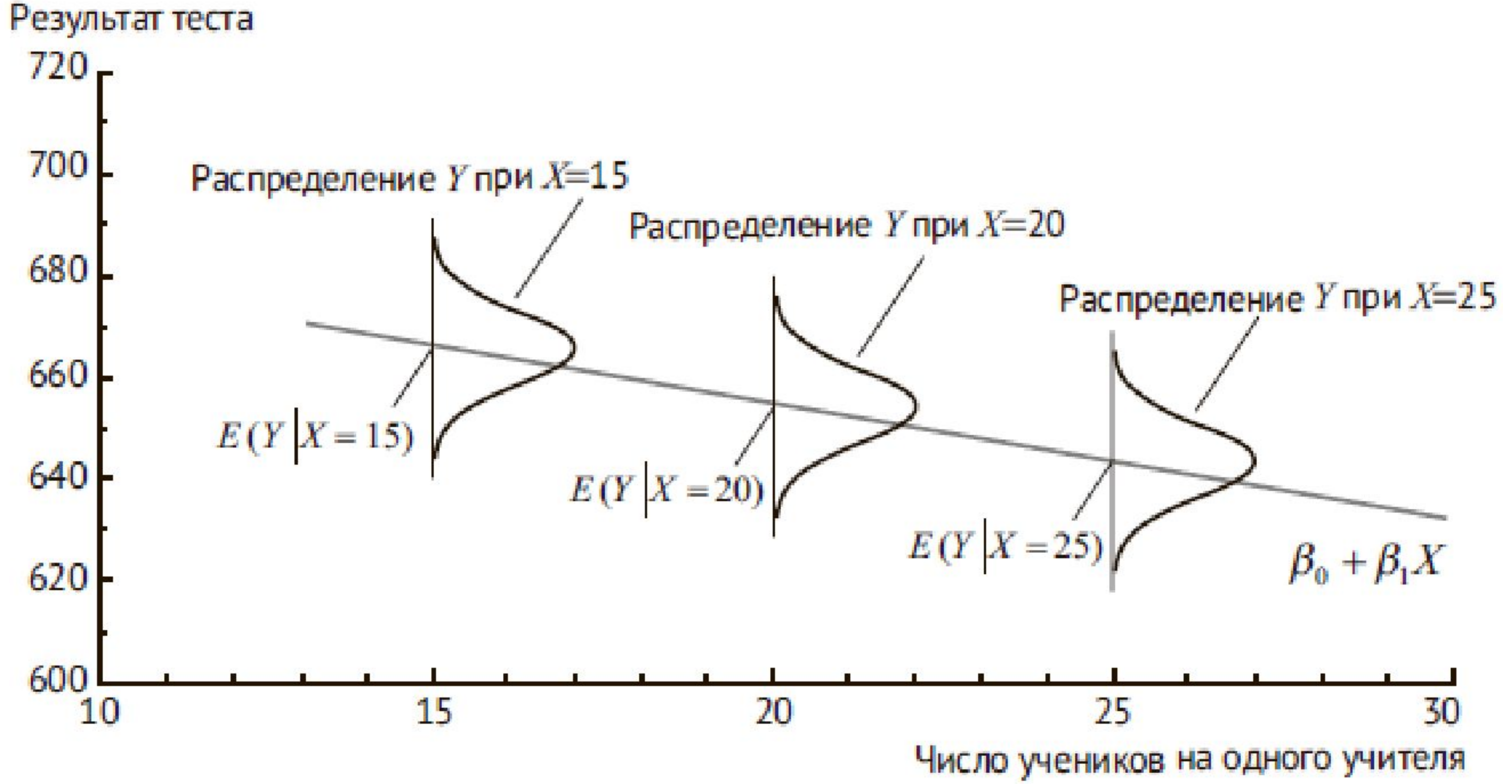
Предположения МНК

Предположение №1: условное распределение u_i относительно X_i имеет нулевое среднее: $E(u_i|X_i) = 0$

Предположение №2: $(X_i, Y_i), i = 1, \dots, n,$ независимы и одинаково распределены $(\mathrm{i.i.d.})$

Предположение №3: большие выбросы маловероятны: X_i и Y_i имеют ненулевые конечные четвертые моменты

Предположение №1: $E(u_i|X_i) = 0$



Предположение №2: (X_i, Y_i), $i = 1, \dots, n$, - (i.i.d.)

Это утверждение о способе формирования выборки – простым случайным образом из одной генеральной совокупности

Предположение №3: большие выбросы маловероятны

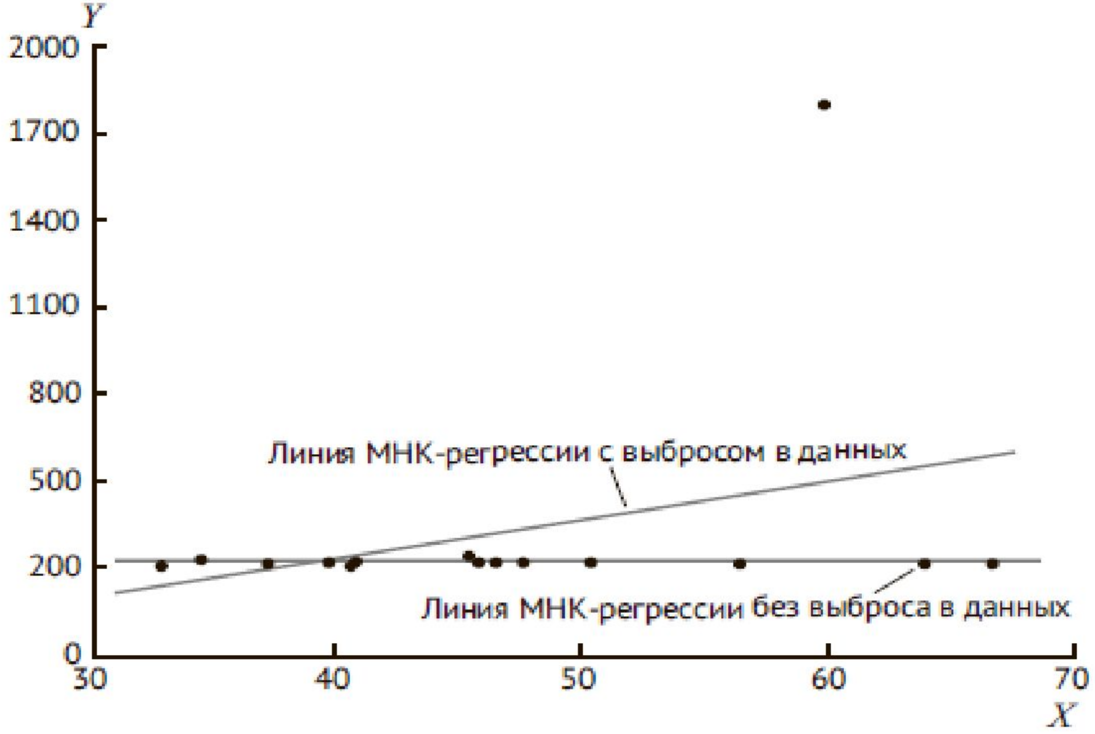


Рисунок 4.5. Чувствительность МНК к большим выбросам

Зачем нужны эти предположения?

- Математическая роль: при их выполнении МНК оценка имеет некоторые хорошие свойства
- Позволяют понять проблемы, возникающие при оценке МНК регрессии, если они нарушаются

Свойства МНК оценок

Если выполняются предположения 1-3 метода наименьших квадратов, то в больших выборках

1. $\hat{\beta}_0$ и $\hat{\beta}_1$ имеют совместное нормальное распределение;

2. $\hat{\beta}_1 \sim N\left(\beta_1, \sigma^2_{\beta_1}\right)$ (асимптотически), где

$$\sigma^2_{\beta_1} = \frac{1}{n}\frac{\mathrm{var}[(X_i - \mu_X)u_i]}{[\mathrm{var}(X_i)]^2}$$

3. $\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2_{\beta_0}\right)$ (асимптотически), где

$$\sigma^2_{\beta_0} = \frac{1}{n}\frac{\mathrm{var}[H_i u_i]}{[\mathrm{E}(H_i^2)]^2}, \text{ где } H_i = 1 - \left[\frac{\mu_X}{E(X_i^2)}\right] X_i$$

Свойства МНК оценок

- $\hat{\beta}_0$ и $\hat{\beta}_1$ - несмещенные оценки;
- $\frac{1}{n}\sum_{i=1}^{n}\hat{u}_i = 0$;
- $\frac{1}{n}\sum_{i=1}^{n}\hat{Y}_i = \bar{Y}$;
- $\sum_{i=1}^{n}\hat{u}_i X_i = 0$