Эконометрика 1
осень 2021

Лекция 2

14.09.2021

Маленькие классы → лучшие результаты? Способ 1

1. Сравнить средние значения *STR* в округах с маленькими классами и в округах с большими классами (“***оценка***”)
2. Проверить “нулевую” гипотезу о том, что средние в округах двух типов совпадают против “альтернативной” гипотезы о том, что они различаются (“***тестирование гипотезы***”)
3. Оценить интервал для разности средних в районах с маленькими и большими классами (“***доверительный интервал***”)

Сравним округа с «маленькими» (STR < 20) и «большими» (STR ≥ 20) классами:

Размер класса	Средний балл ($\bar{Y}$)	Стандартное отклонение (s_Y)	n
Маленький	657,4	19,4	238
Большой	650,0	17,9	182

1. ***Оценка*** Δ = разность групповых средних
2. ***Тестирование гипотезы*** Δ=0
3. ***Построение доверительного интервала*** для Δ

1. Оценка

$$\bar{Y}_{small} - \bar{Y}_{large} = \frac{1}{n_{small}} \sum_{i=1}^{n_{small}} Y_i - \frac{1}{large} \sum_{i=1}^{n_{large}} Y_i =$$
$$= 657{,}4 - 650{,}0 = 7{,}4$$

Велико ли это различие?

Стандартная ошибка по всем округам= 19,1

Разность между 60% и 75% процентилями результатов тестов равна 667,6 – 659,4 = 8,2

Является ли эта разность достаточно большой, чтобы принимать ее по внимание при реформировании системы образования?

2. Тестирование гипотезы

Тест на различие в средних: вычисляем t-статистику,

$$t = \frac{\bar{Y}_s - \bar{Y}_l}{\sqrt{\frac{s^2}{n_s}_s + \frac{s^2}{n_l}_l}} = \frac{\bar{Y}_s - \bar{Y}_l}{SE(\bar{Y}_s - \bar{Y}_l)}$$

где $SE(\bar{Y}_s - \bar{Y}_l)$ – "стандартная ошибка" разности средних $(\bar{Y}_s - \bar{Y}_l)$ и $s^2{}_s = \frac{1}{n_s - 1}\sum_{i=1}^{n_s}(Y_i - \bar{Y}_s)^2$ и т.д.

2. **Тестирование гипотезы:** вычисление статистики

Размер класса	Средний балл ($\bar{Y}$)	Стандартное отклонение (s_Y)	n
Маленький	657,4	19,4	238
Большой	650,0	17,9	182

$$t = \frac{\bar{Y}_s - \bar{Y}_l}{\sqrt{\frac{s_s^2}{n_s} + \frac{s_l^2}{n_l}}} = \frac{657{,}4 - 650{,}0}{\sqrt{\frac{19{,}4^2}{238} + \frac{17{,}9^2}{182}}} = \frac{7{,}4}{1{,}83} = 4{,}05$$

$|t| > 1{,}96 \rightarrow$ отвергаем нулевую гипотезу на 5%-м уровне значимости

3. Доверительный интервал

95%-й доверительный интервал для разности средних имеет вид:

$$(\bar{Y}_s - \bar{Y}_l) \pm 1{,}96 \times SE(\bar{Y}_s - \bar{Y}_l) =$$
$$= 7{,}4 \pm 1{,}96 \times 1{,}83 = (3{,}8; 11{,}0)$$

Два эквивалентных утверждения:

1. 95%-й доверительный интервал Δ не включает 0;
2. Нулевая гипотеза $\Delta=0$ отвергается на уровне значимости 5%.

Какое это имеет отношение к эконометрике?

- Рассмотренная концепция может быть распространена на регрессионный анализ
- Схема действий аналогична (оценка, тестирование нулевой гипотезы, построение доверительного интервала)

Еще способы: идеальный – эксперимент (экспериментальные данные)

Случайным образом поделить (всех) детей (и обучающих их учителей!) на группы и обучать одних в больших классах, а других – в маленьких

Проводить регулярное тестирование и сравнивать различия (как?)

Проблемы:

1. Технические сложности
2. Очень дорого (4-летний проект STAR, вторая половина 80-х - $12 млн (см. СУ, раздел 13.3))
3. Родителя хотят, чтобы их ребенок учился в маленьком классе → мешают эксперименту

Реалистичный способ (3) – наблюдаемые данные

Эконометристы работают с наблюдаемыми данными и моделями

Наша задача: что мы хотим понять?

→ Как меняются результаты обучения (баллы за тесты) при изменении размера класса?

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

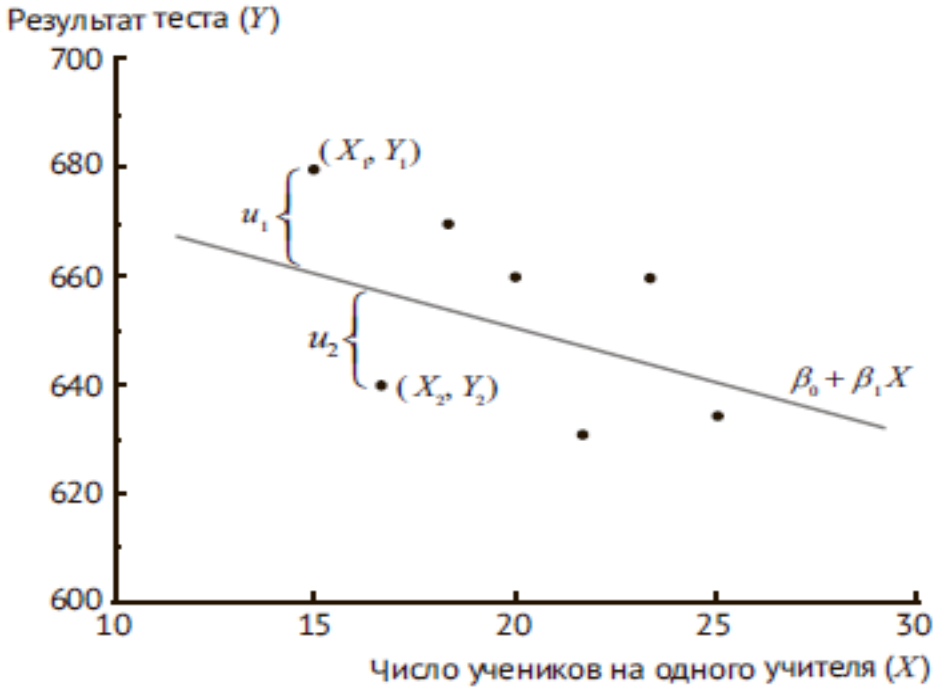


Рисунок 4.1. Диаграмма рассеяния результатов тестов относительно числа учеников в классе (гипотетические данные)

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

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

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

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

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

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

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

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

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

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

МНК оценка коэффициентов парной линейной регрессии

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

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

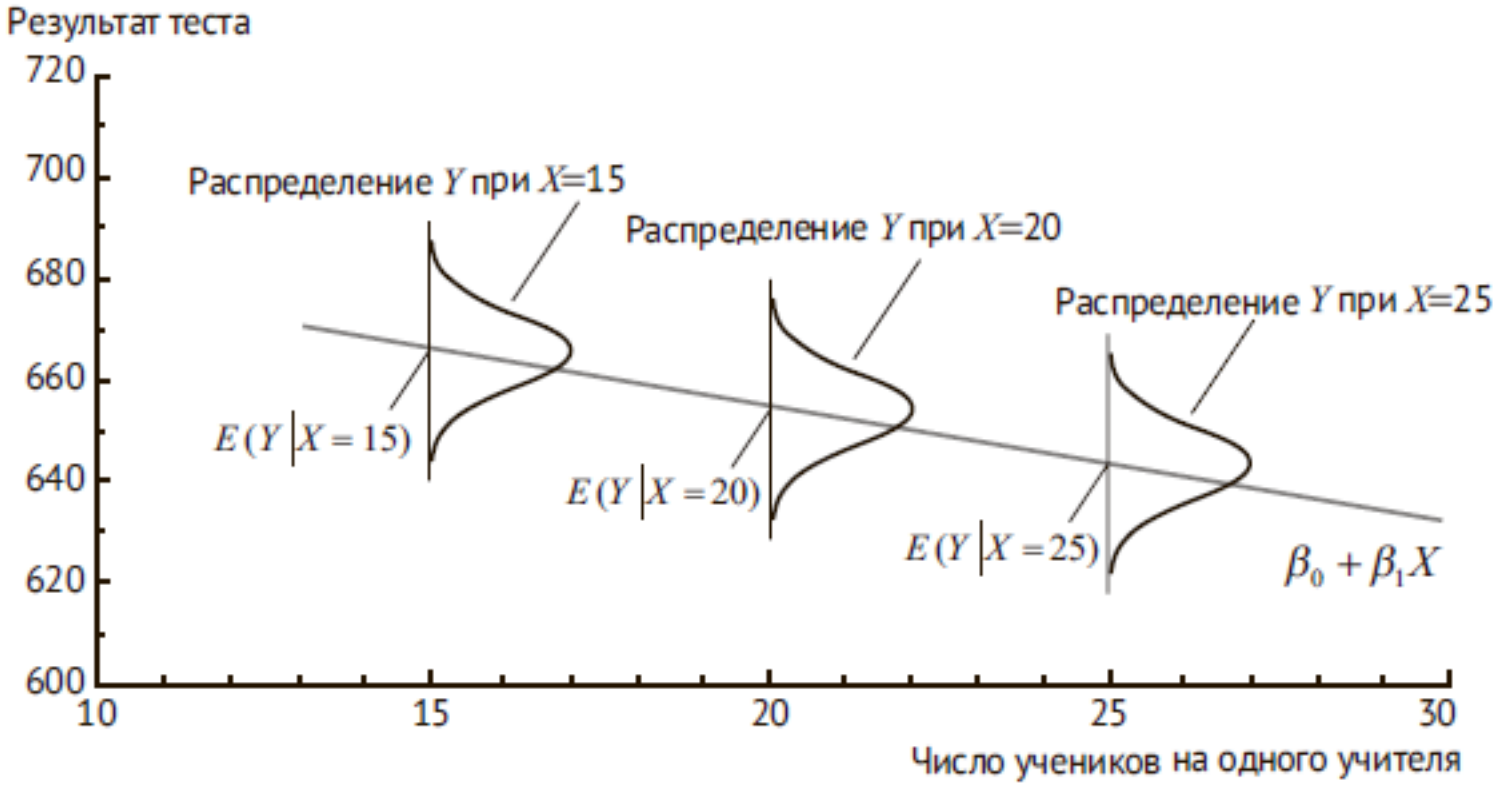
Предположения МНК

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 (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

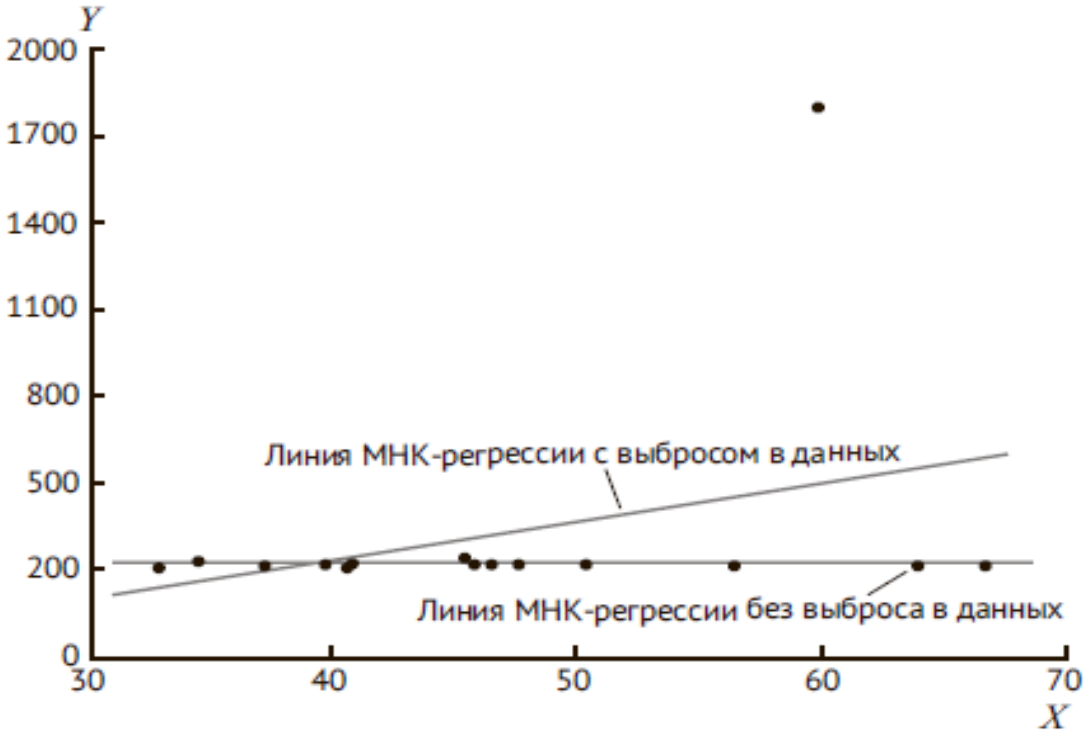


Рисунок 4.5. Чувствительность МНК к большим выбросам

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