



Эконометрика

Лекция 6

Неоднородность выборок. Дамми переменные

Вакуленко Е.С.

д.э.н., доцент департамента прикладной экономики

evakulenka@hse.ru

Москва, 2022



Измерение дискриминации в оплате труда: фактивные переменные в моделях регрессии



План

- Неоднородность выборок
- Фиктивные (дамми) переменные
- Дамми переменные на константу и угловые коэффициенты регрессии
- Тест Чоу на неоднородность
- Эквивалентность теста Чоу и теста на значимость дамми переменных на все коэффициенты модели



План

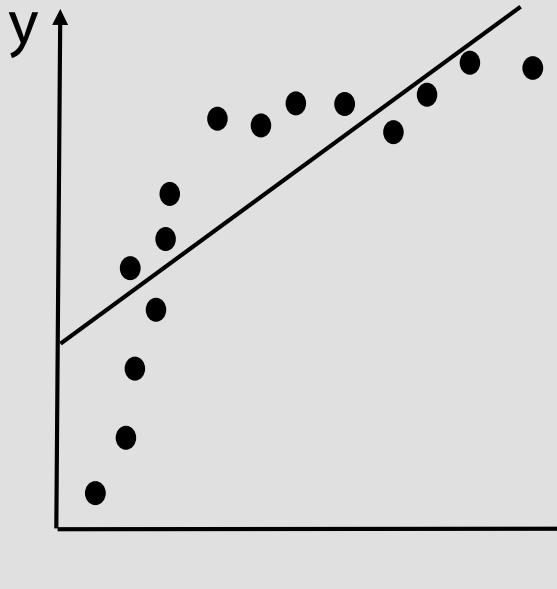
- Тест Чоу на прогнозную силу
- Иные применения дамми переменных
 - Сезонные дамми переменные;
 - Структурные сдвиги;
 - Выбросы и влиятельные наблюдения.



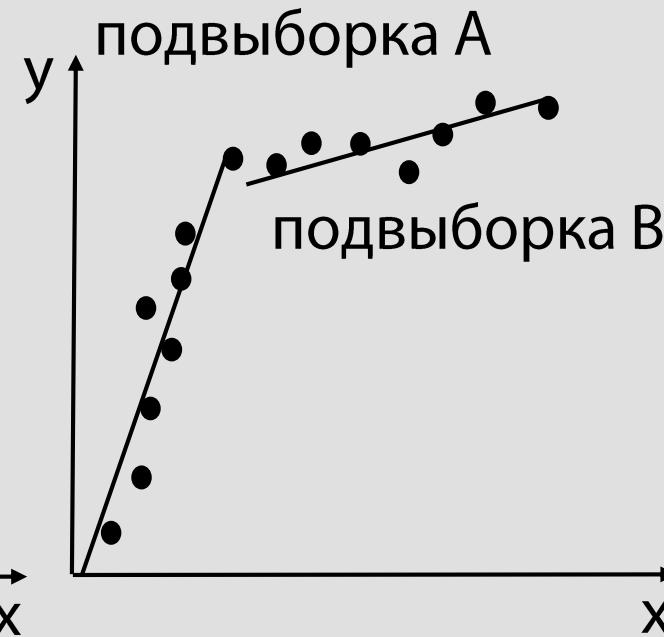
Неоднородность выборок



Пример неоднородности выборки



Объединенная регрессия



Регрессии подвыборок



Применение статистики

- По статистике женщины зарабатывают меньше мужчин.
- Почему?
- Можно ли объяснить эти различия с помощью регрессионной модели?



Международная статистика

- Во всем мире женщины зарабатывают гораздо меньше, чем мужчины. Такие данные были представлены в докладе «Глобальный гендерный разрыв» (The Global Gender Gap Report 2015) Всемирного экономического форума.
- В 2015 Россия заняла 53-е место из 145 стран-участниц по уровню зарплатного неравенства между мужчинами и женщинами.



Дискриминация по заработной плате

Более низкая оплата труда при прочих равных:

- Образование
- Опыт работы
- Возраст
- Отрасль занятости
- Профессия
- Знание языков
- И многое другое



Достойны лучшего

- По данным отчета Global Wage Report за 2014–2015 гг., подготовленного Международной организацией труда (МОТ), российские женщины действительно получают на 30% меньше, нежели мужчины.
- Представители МОТ подсчитали так называемую объяснимую разницу в заработной плате, которая учитывает, в частности, уровень образования.
- Исследователи обнаружили, что на самом деле зарплата российских женщин должна быть на 11.1% выше, чем у мужчин.





У вас степень по экономике,
вы чемпионка мира по конному спорту
и у вас медаль «За отвагу на пожаре»?
Прекрасное резюме!
Опубликуйте его
на сайте знакомств!

Дамми переменная на константу



Дамми (фиктивные) переменные



Для исследования влияния качественных признаков в модель можно вводить бинарные **(дамми) переменные**, которые, как правило, принимают два значения:

- 1, если данный качественный признак присутствует в наблюдении;
- 0 при его отсутствии.



Пример. Модель заработной платы

- Рассмотрим зависимость заработной платы от уровня образования, т.е. модель вида:

$$wage_i = \beta_1 + \beta_2 educ_i + \varepsilon_i,$$

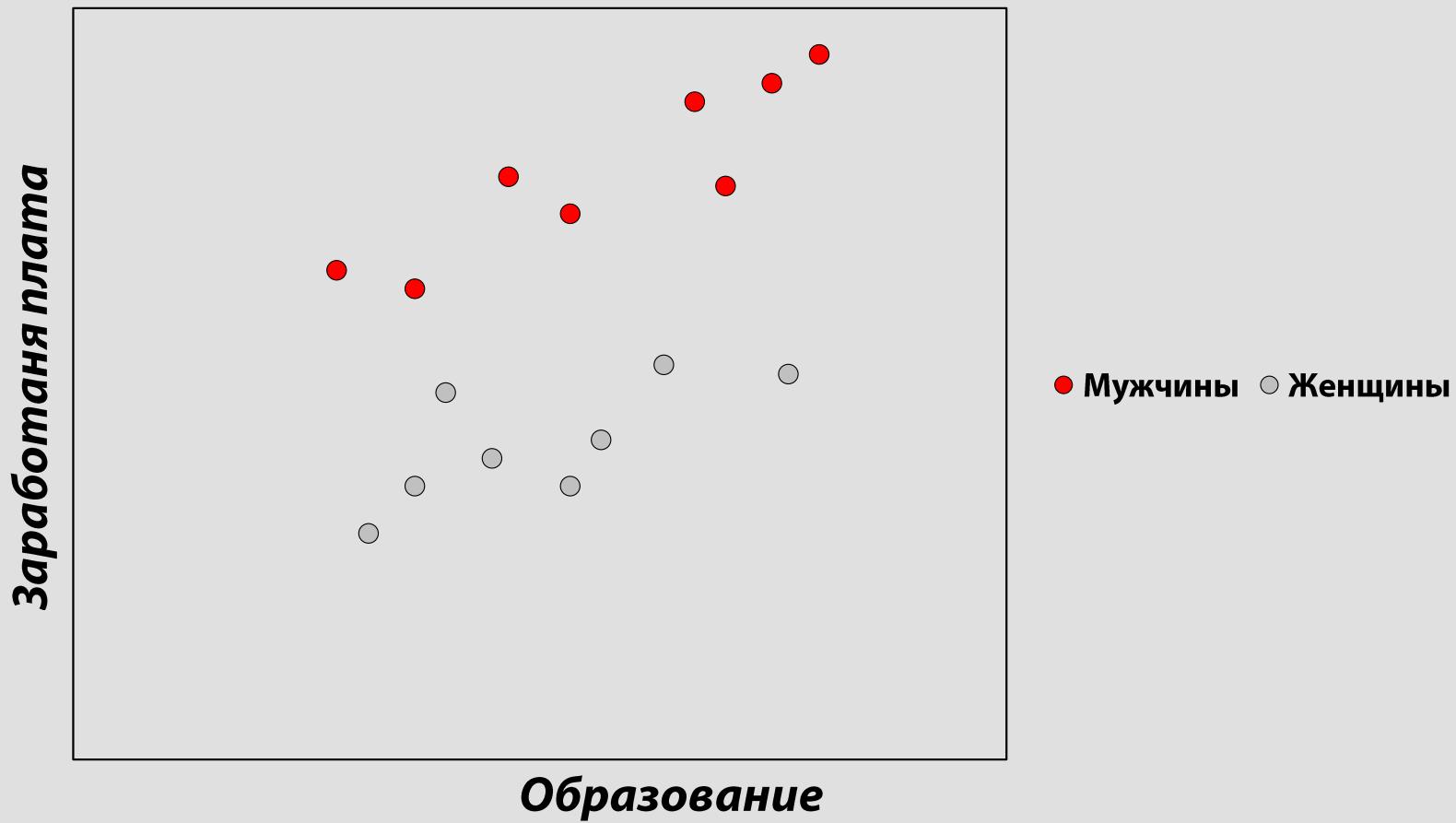
wage – заработная плата,

educ – уровень образования (в годах).

- В данных содержится информация о поле респондентов.



Графическое представление



Как учесть неоднородность?

Заработные платы мужчин больше, чем у женщин.

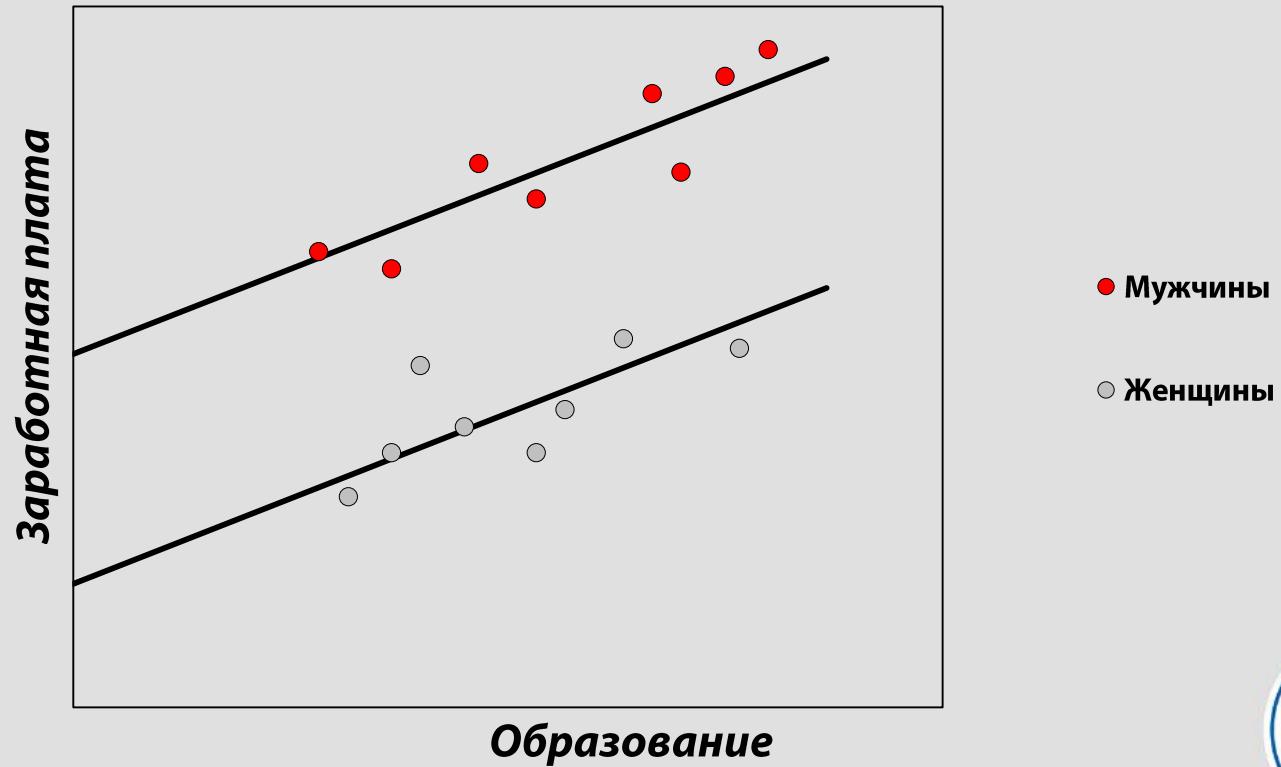
Если оценивать регрессии отдельно для мужчин и женщин, то размеры выборок уменьшаются, что снизит точность оценивания.

! На помощь приходят дамми переменные!



Пример дамми переменной при наличии двух категорий

Предположим, что коэффициенты наклона в регрессиях для мужчин и женщин одинаковы, а свободные члены различаются.



Пример дамми переменной при наличии двух категорий

Мы предполагаем, что базовые заработные платы для мужчин и женщин различаются, а отдача на образование у них одинакова.

Для женщин:

$$wage = \beta_1 + \beta_2 educ + \varepsilon_1$$

Для мужчин:

$$wage = \beta'_1 + \beta_2 educ + \varepsilon_2,$$

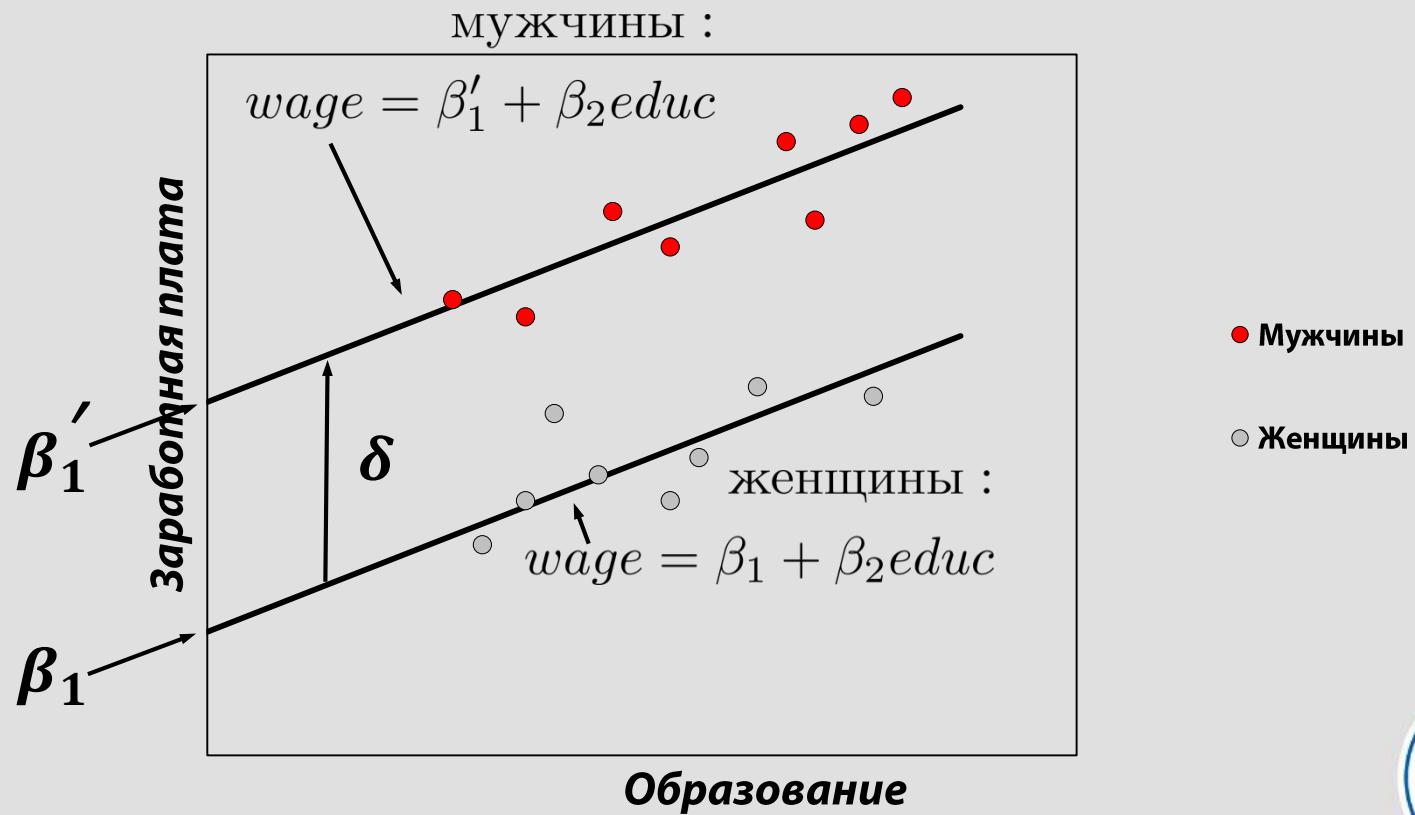
wage – заработная плата(\$ в час),

educ – уровень образования (в годах).



Пример дамми переменной при наличии двух категорий

Обозначим δ разность свободных членов: $\delta = \beta_1' - \beta_1$



Пример дамми переменной при наличии двух категорий

Тогда $\beta_1' = \beta_1 + \delta$, и мы можем переписать регрессию для мужчин.

Для женщин:

$$wage = \beta_1 + \beta_2 educ + \varepsilon_1$$

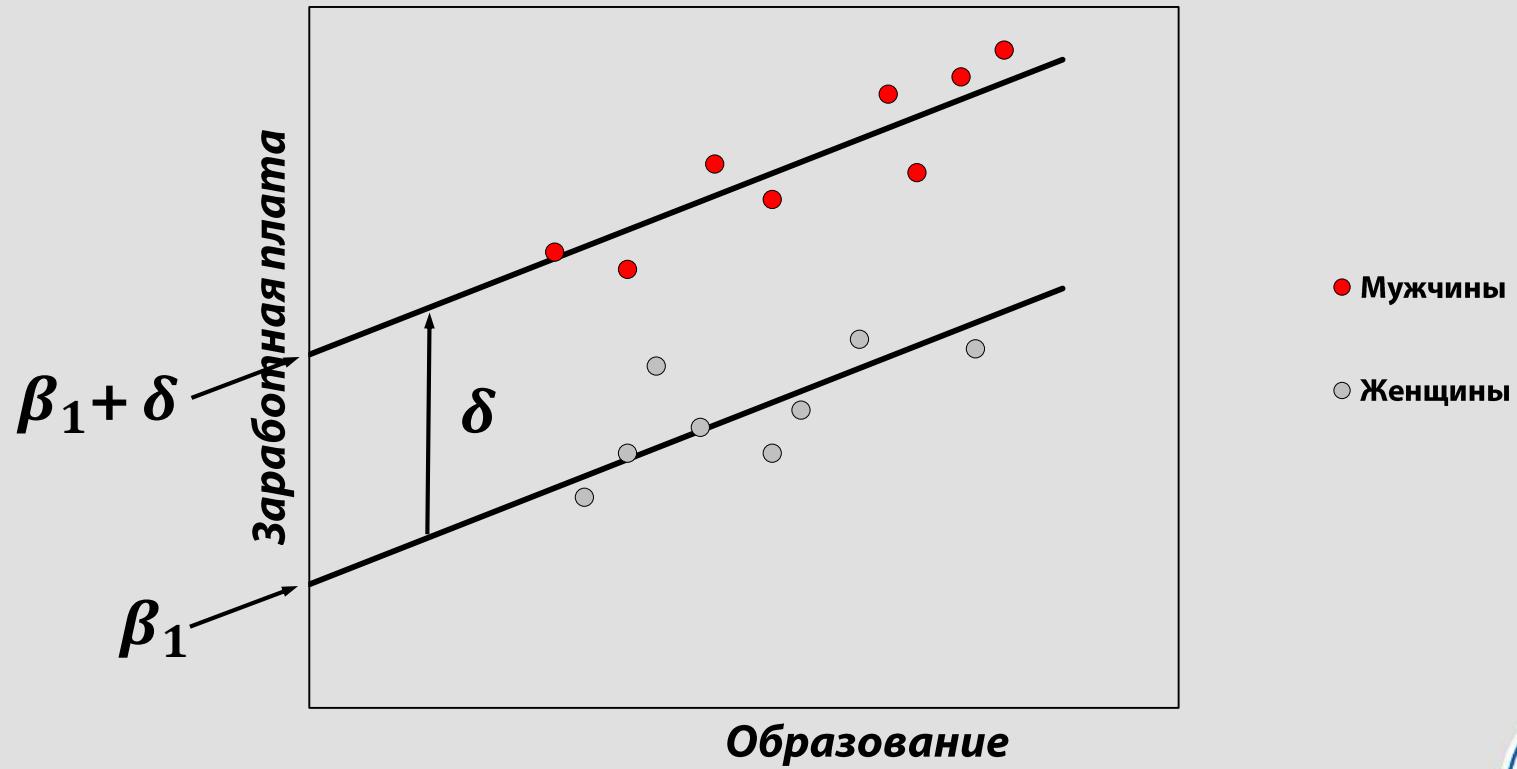
Для мужчин:

$$wage = \beta_1 + \delta + \beta_2 educ + \varepsilon_2$$



Пример дамми переменной при наличии двух категорий

δ разность свободных членов: $\delta = \beta_1' - \beta_1$



Пример дамми переменной при наличии двух категорий

Введем дамми переменную $male$, которая равна 0 для женщин и 1 для мужчин.

Дамми переменная всегда принимает только два значения, обычно 0 и 1.

Общее уравнение:

$$wage = \beta_1 + \delta male + \beta_2 educ + \varepsilon$$

Женщины, $male = 0$:

$$wage = \beta_1 + \beta_2 educ + \varepsilon$$

Мужчины, $male = 1$:

$$wage = \beta_1 + \delta + \beta_2 educ + \varepsilon$$



Пример использования дамми переменной

В последней колонке сформирована дамми переменная.
В приведенной таблице указаны данные лишь для 5 человек.
Всего в выборке 526 наблюдений.

Данные о заработных платах в США в 1976 г.

Респондент	Пол	wage	educ	male
1	Мужчина	3	11	1
2	Мужчина	6	8	1
3	Женщина	3.1	11	0
4	Мужчина	18	17	1
5	Женщина	3.2	12	0



Описательные статистики

gretl: статистика

	Среднее	Медиана	S.D.	Min	Max
wage	5,896	4,650	3,693	0,5300	24,98
educ	12,56	12,00	2,769	0,0000	18,00
male	0,5209	1,000	0,5000	0,0000	1,000

Источник: wage1.dta. Wooldridge (2016)



Пример использования дамми переменной

В таблице приведены результаты оценивания регрессии:

$$wage = \beta_1 + \delta male + \beta_2 educ + \varepsilon$$

gretl: модель 1

Файл Правка Тесты Сохранить Графики Анализ LaTeX

Модель 1: МНК, использованы наблюдения 1-526
Зависимая переменная: wage

	Коэффициент	Ст. ошибка	t-статистика	P-значение	
const	-1,65055	0,652317	-2,530	0,0117	**
male	2,27336	0,279044	8,147	2,76e-015	***
educ	0,506452	0,0503906	10,05	7,56e-022	***
Среднее зав. перемен	5,896103	Ст. откл. зав. перемен	3,693086		
Сумма кв. остатков	5307,161	Ст. ошибка модели	3,185520		
R-квадрат	0,258819	Испр. R-квадрат	0,255985		
F (2, 523)	91,31542	P-значение (F)	9,66e-35		
Лог. правдоподобие	-1354,289	Крит. Акаике	2714,578		
Крит. Шварца	2727,374	Крит. Хеннана-Куинна	2719,588		

Источник: wage1.dta. Wooldridge (2016)



Пример использования дамми переменной

$$\widehat{wage} = -1.65 + 2.27male + 0.51educ$$

gretl: модель 1

Файл Правка Тесты Сохранить Графики Анализ LaTeX				
Модель 1: МНК, использованы наблюдения 1-526				
Зависимая переменная: wage				
Коэффициент	Ст. ошибка	t-статистика	Р-значение	
const	-1,65055	0,652317	-2,530	0,0117
male	2,27336	0,279044	8,147	2,76e-015
educ	0,506452	0,0503906	10,05	7,56e-022
Среднее зав. перемен	5,896103	Ст. откл. зав. перемен	3,693086	
Сумма кв. остатков	5307,161	Ст. ошибка модели	3,185520	
R-квадрат	0,258819	Испр. R-квадрат	0,255985	
F (2, 523)	91,31542	Р-значение (F)	9,66e-35	
Лог. правдоподобие	-1354,289	Крит. Акаике	2714,578	
Крит. Шварца	2727,374	Крит. Хеннана-Куинна	2719,588	

Все коэффициенты значимы на 5% уровне значимости.



Результаты оценивания

Коэффициент при **male** значим, базовая заработка плата мужчин на 2.27\$ в час больше.

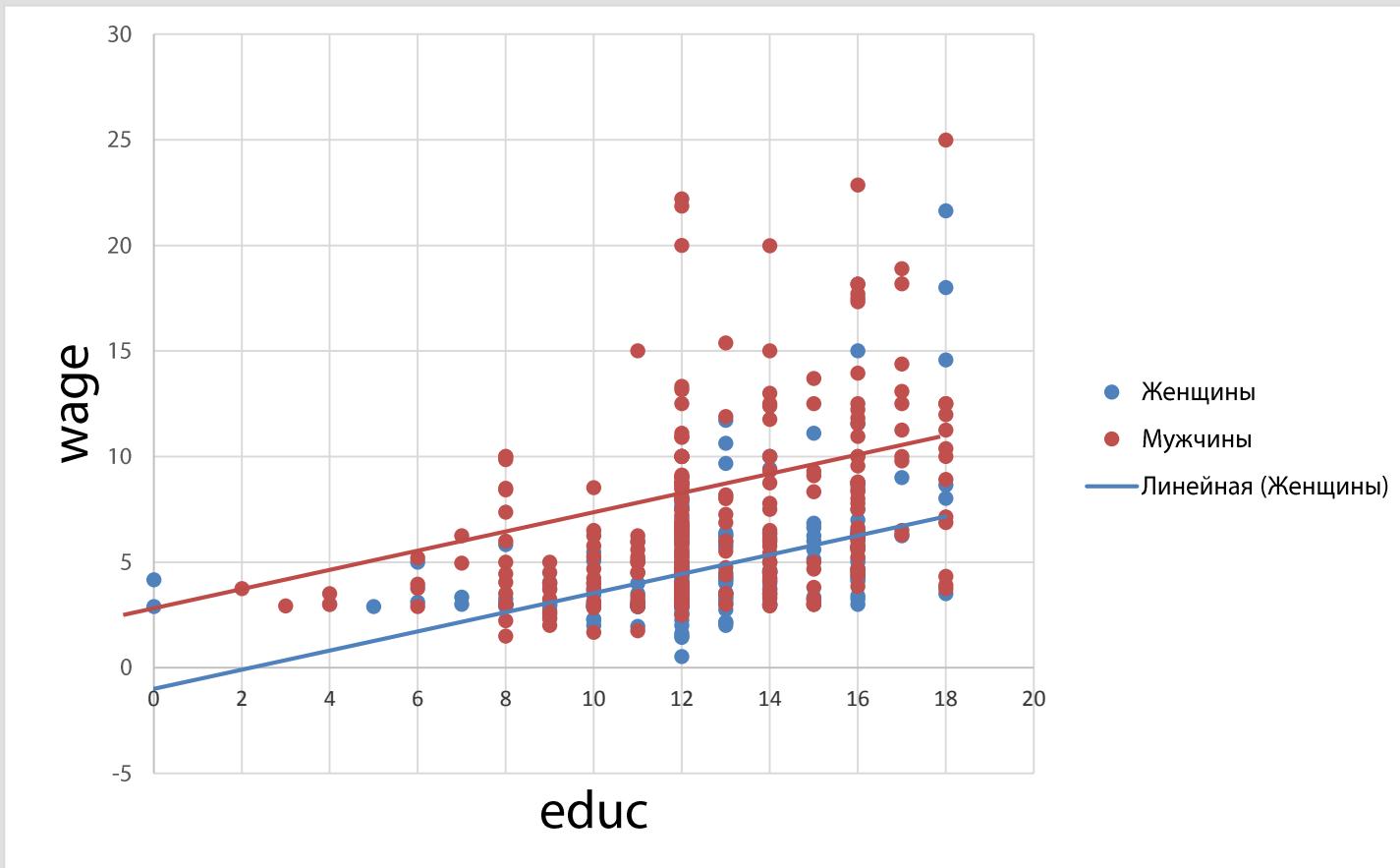
Коэффициент при **educ** значим. Каждый дополнительный год образования увеличивает почасовую заработную плату на 0.5\$.

Свободный член значим и отрицательный (-1.65). Не имеет экономического смысла.



Пример использования дамми переменной

А можете, пожалуйста, убрать из легенды надпись «линейная (женщины)»?



Регрессия на дамми переменную

$n = 526, R^2 = 0.116$

$$\widehat{wage} = 4.59 + 2.51male$$
$$(0.22) \quad (0.30)$$

Свободный член показывает среднюю заработную плату женщин в выборке ($male = 0$), таким образом, женщины зарабатывают в среднем 4.59\$ в час.

Коэффициент при $male$ - это разница в средней заработной плате между женщинами и мужчинами. Таким образом, средняя заработная плата мужчин в выборке составляет $4.59 + 2.51 = 7.10$ \$ в час.



**Дамми переменная для
коэффициента наклона**



Дамми переменные для коэффициента наклона

Ослабим требование об одинаковых отдачах на образование (коэффициентах наклона) для женщин и мужчин.

Введем переменную **meduc**, как произведение **educ** и **male**.

Для женщин переменная **male** равна 0 и, следовательно, **meduc** также равна 0.

Для мужчин переменная **male** равна 1, следовательно, переменная **meduc** равна **educ**.



Дамми переменные для коэффициента наклона

$wage = \beta_1 + \delta male + \beta_2 educ + \lambda meduc + \varepsilon$	
Женщины (male = meduc = 0)	$wage = \beta_1 + \beta_2 educ + \varepsilon$
Мужчины (male = 1; meduc = educ)	$wage = (\beta_1 + \delta) + (\beta_2 + \lambda)educ + \varepsilon$



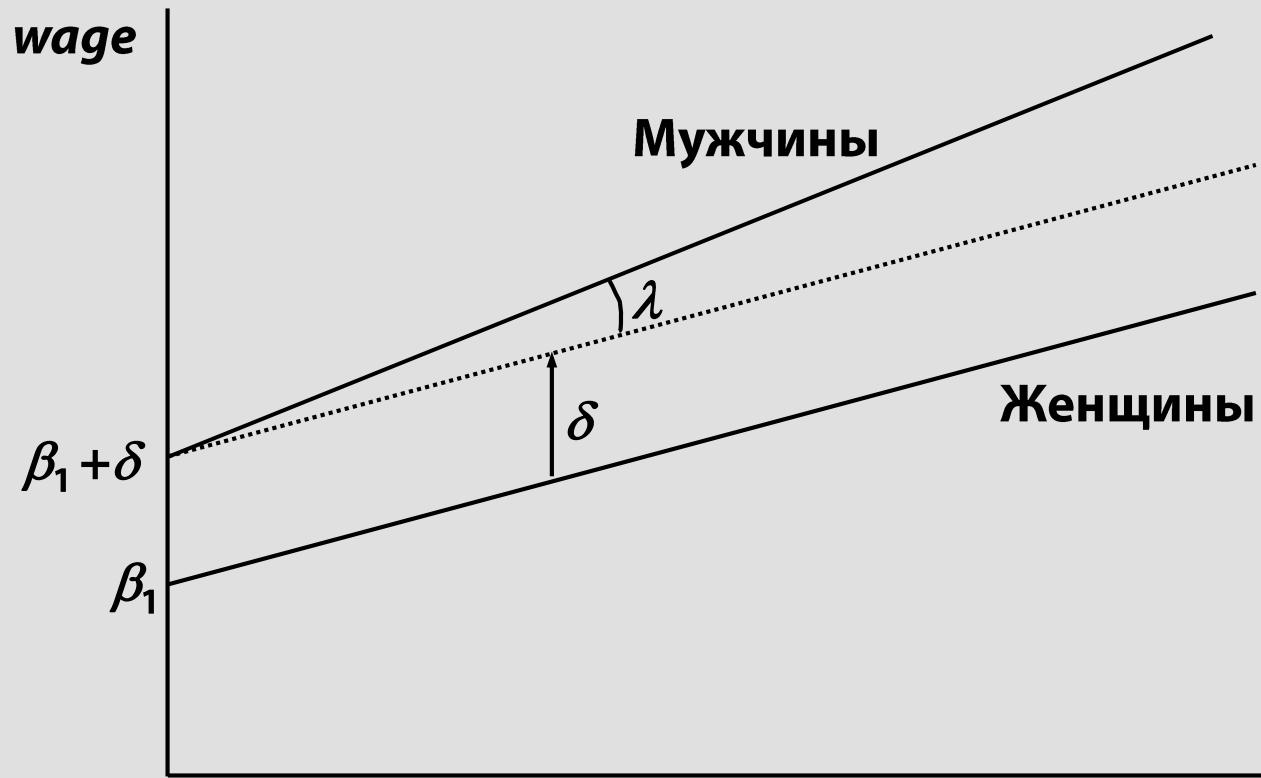
Дамми переменные для коэффициента наклона

Отдача на образование мужчин больше на λ по сравнению с отдачей на образование женщин, базовые заработные платы различаются на δ .



Дамми переменные для коэффициента наклона

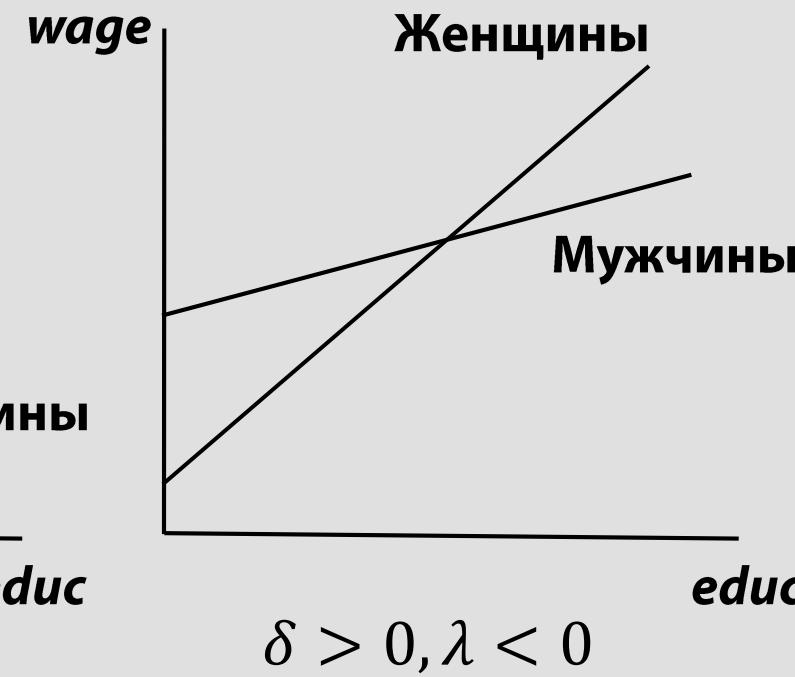
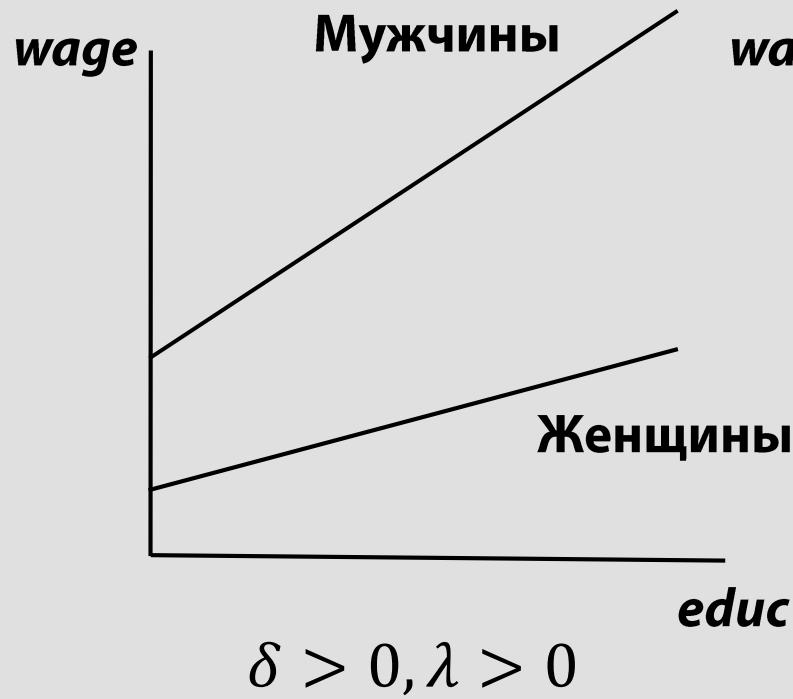
Диаграмма иллюстрирует разницу в коэффициентах наклона графически.



edus



Дамми переменные для коэффициента наклона: различные случаи



Оценка модели с дамми переменными на наконстанту и коэффициент наклона

$$\widehat{wage} = -1 + 1.2male + 0.45educ + 0.09meduc$$

gretl: модель 2

Файл Правка Тесты Сохранить Графики Анализ LaTeX

Модель 2: МНК, использованы наблюдения 1-526
Зависимая переменная: wage

	Коэффициент	Ст. ошибка	t-статистика	P-значение
const	-0,998027	1,02183	-0,9767	0,3292
male	1,19852	1,32504	0,9045	0,3661
educ	0,453477	0,0813414	5,575	3,98e-08 ***
meduc	0,0859990	0,103639	0,8298	0,4070

Среднее зав. перемен 5,896103 Ст. откл. зав. перемен 3,693086
Сумма кв. остатков 5300,170 Ст. ошибка модели 3,186469
R-квадрат 0,259796 Испр. R-квадрат 0,255542
F(3, 522) 61,07022 P-значение (F) 7,44e-34
Лог. правдоподобие -1353,942 Крит. Акаике 2715,885
Крит. Шварца 2732,946 Крит. Хеннана-Куинна 2722,565

Значима только переменная образование.



Выпишем модели для женщин и мужчин

$$\widehat{wage} = -1 + 1.2 male + 0.45educ + 0.09meduc$$

Женщины
(male = meduc = 0)

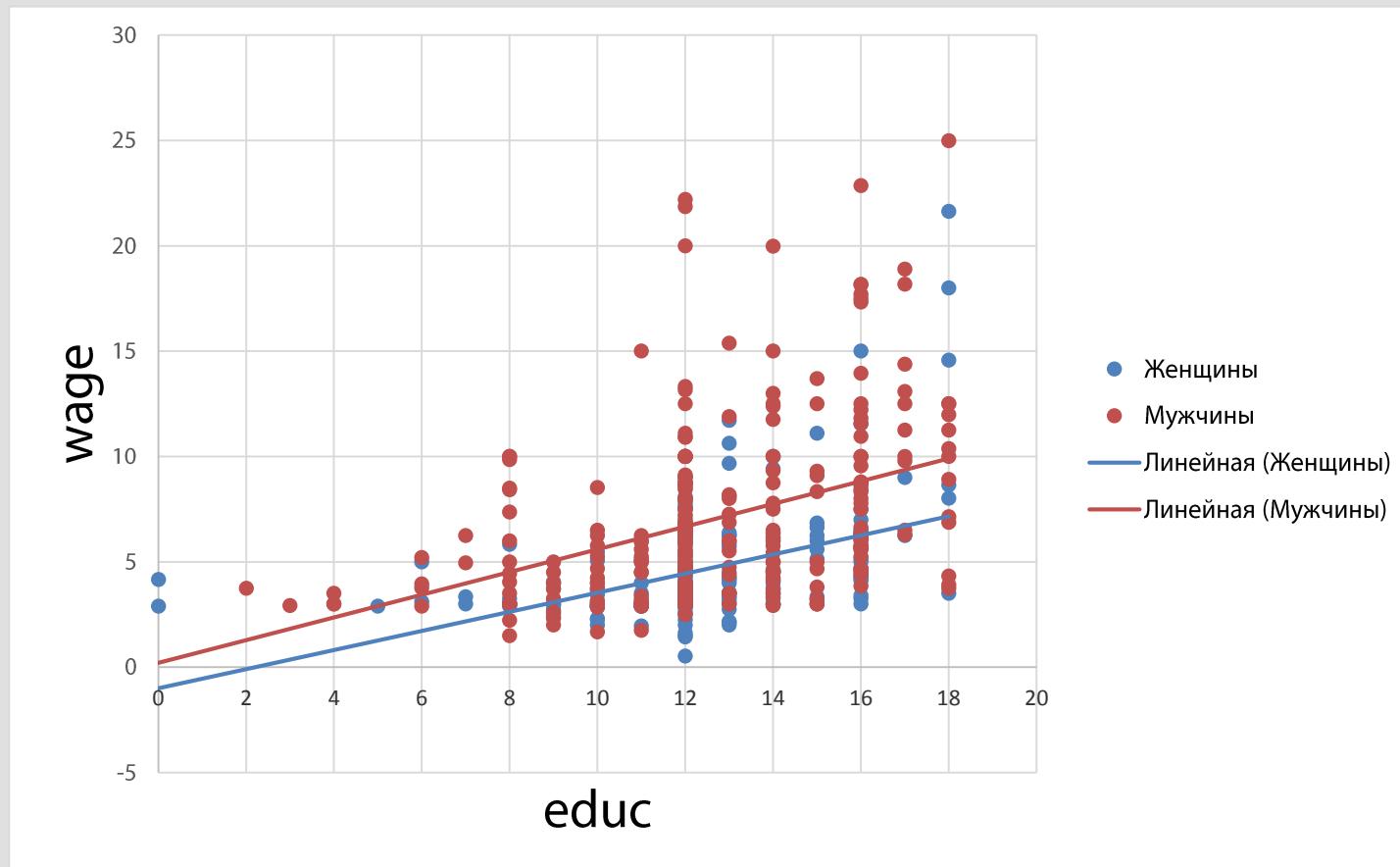
$$\widehat{wage} = -1 + 0.45educ$$

Мужчины
(male = 1; meduc = educ)

$$\begin{aligned}\widehat{wage} &= -1 + 1.2 + 0.45educ \\ &+ 0.09educ = 0.2 + 0.54educ\end{aligned}$$



Графическое представление полученных оценок



Результаты

Коэффициент при **medic** незначим, следовательно, отдача от образования мужчин и женщин не отличается.

Коэффициент при переменной **male** незначим, следовательно, базовые заработные платы также не различаются.



Проверка гипотез о значимости дамми переменных

Нулевая гипотеза состоит в том, что коэффициенты перед переменными male и meduc одновременно равны 0.

Альтернативной является гипотеза о том, что один или оба коэффициента неравны нулю.

$$H_0 : \begin{cases} \delta = 0 \\ \lambda = 0 \end{cases}$$

$$H_1 : \begin{bmatrix} \delta \neq 0 \\ \lambda \neq 0 \end{bmatrix}$$



Статистика для проверки гипотезы

$$F = \frac{(RSS_r - RSS_{ur}) / (2)}{RSS_{ur} / (n - k - 1)},$$

где k – количество регрессоров без константы;
 RSS_{ur} – RSS для модели с дамми переменными;
 RSS_r – RSS для модели без дамми переменных.

Как и ранее, сравниваем наблюдаемое и критическое значение статистики.

Если $F > F_{2,n-k-1}$ для заданного уровня значимости α ,
то гипотеза H_0 отвергается.



Проверка гипотез о значимости дамми переменных

Находим значение F-статистики и сравниваем его с критическим.

Для модели с дамми переменными:

Сумма кв. остатков (RSS_{ur}) = 5300.17

Для модели без дамми переменных:

Сумма кв. остатков (RSS_r) = 5980.68

$$F = \frac{(5980.68 - 5300.17)/2}{5300.17/522} \approx 33.51$$

$$F_{2,522,0.01} = 4.65$$

Поскольку $F = 33.51 > 4.65$, то гипотеза H_0 отвергается для уровня значимости 1%.



Расчет в Gretl

gretl: модель 4

Файл Правка Тесты Сохранить Графики Анализ LaTeX

Тестирование модели 2:

Нулевая гипотеза: параметры регрессии нулевые
male, meduc

Тестовая статистика: F(2, 522) = 33,5109, Р-значение 2,03093e-014
Omitting variables improved 0 of 3 information criteria.

Модель 4: МНК, использованы наблюдения 1-526
Зависимая переменная: wage

	Коэффициент	Ст. ошибка	t-статистика	Р-значение
const	-0,904852	0,684968	-1,321	0,1871
educ	0,541359	0,0532480	10,17	2,78e-022 ***

Среднее зав. перемен 5,896103 Ст. откл. зав. перемен 3,693086
Сумма кв. остатков 5980,682 Ст. ошибка модели 3,378390
R-квадрат 0,164758 Испр. R-квадрат 0,163164
F(1, 524) 103,3627 Р-значение (F) 2,78e-22
Лог. правдоподобие -1385,712 Крит. Акаике 2775,423
Крит. Шварца 2783,954 Крит. Хеннана-Куинна 2778,764

P – значение = 0, нулевая гипотеза отвергается.



Проверка гипотез о значимости дамми переменных

Поскольку значение F-статистики больше критического (при любом разумном уровне значимости), то нулевая гипотеза отвергается на 1% уровне значимости.

Следовательно, есть различия между базовыми заработными платами и отдачами на образование между мужчинами и женщинами.



Тест Чоу



Тест Чоу

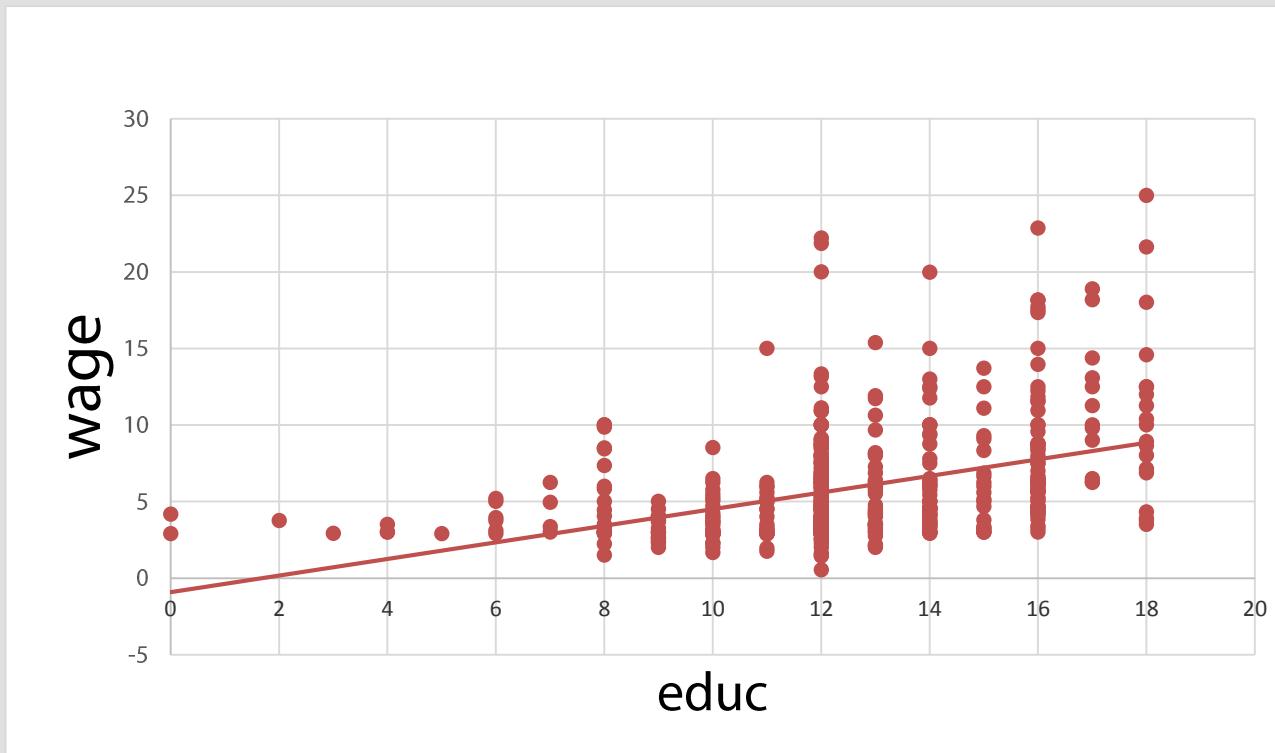
Тест Chow дает ответ на вопрос, можно ли считать что две выборки принадлежат одной генеральной совокупности, т.е. лучше оценивать одну регрессию, или к разным, тогда лучше оценивать две отдельные регрессии.

Тест может рассматриваться для более чем двух выборок.



Тест Чоу. Пример

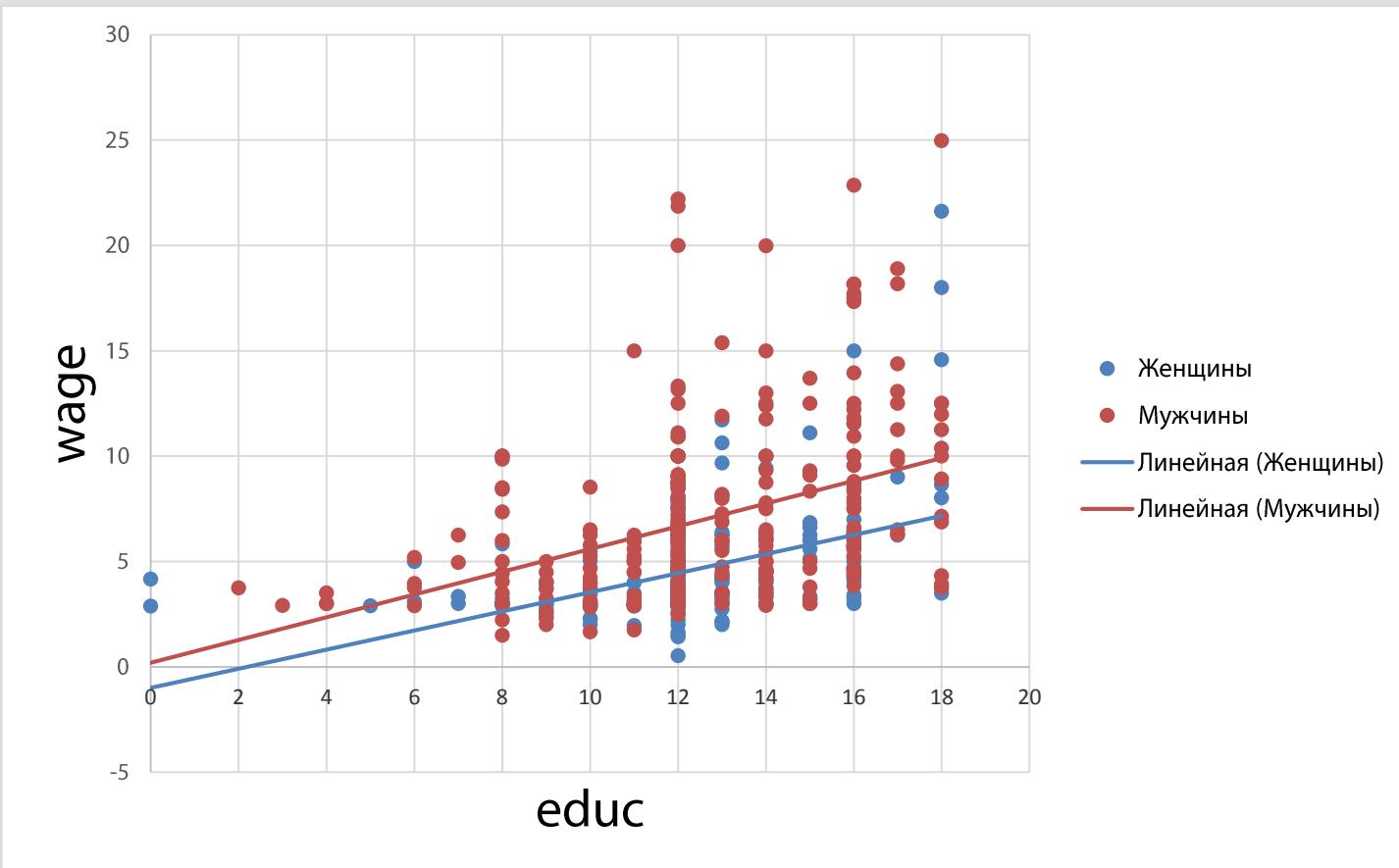
В тесте Чоу мы начинаем с оценки параметров регрессии по всем наблюдениям.



Тест Чоу. Пример

А можете, пожалуйста, убрать из легенды надпись
«линейная (женщины)» и «линейная (мужчины)» ?

Линии оцененных по двум выборкам функций регрессии.



Основная и альтернативная гипотезы в teste Чоу

Модель для первого набора наблюдений:

$$Y = \beta'_1 + \beta'_2 X_2 + \dots + \beta'_k X_k + \varepsilon'$$

Модель для второго набора наблюдений:

$$Y = \beta''_1 + \beta''_2 X_2 + \dots + \beta''_k X_k + \varepsilon''$$

$$H_0 : \beta'_1 = \beta''_1, \dots, \beta'_k = \beta''_k, \sigma_{\varepsilon'}^2 = \sigma_{\varepsilon''}^2$$

$$H_1 : \exists i : \beta'_i \neq \beta''_i$$



Тестовая статистика в teste Чоу

$$F = \frac{(RSS_p - [RSS_1 + RSS_2]) / (k + 1)}{(RSS_1 + RSS_2) / (n - 2k - 2)} \sim F_{k+1, n-2k-2},$$

Где k – это количество регрессоров без константы;

RSS_p – это сумма квадратов остатков для всей выборки;

RSS_1 – это сумма квадратов остатков для выборки 1;

RSS_2 – это сумма квадратов остатков для выборки 2.

Если $F > F_{k+1, n-2k-2}$ для заданного уровня значимости α ,
то основная гипотеза отвергается и нужно оценивать две
отдельные регрессии.



Тест Чоу. Пример

Вернемся к рассматриваемому примеру. Сравним RSS по всей выборке и отдельно по мужчинам и женщинам.

$$F = \frac{(RSS_p - [RSS_1 + RSS_2]) / (k + 1)}{(RSS_1 + RSS_2) / (n - 2k - 2)}$$

$$F = \frac{(5980.68 - [4009.93 + 1290.24]) / 2}{(4009.93 + 1290.24) / 522} \approx 33.51$$

$$RSS_p = 5980.68$$

$$RSS_1 \text{ (мужчины)} = 4009.93$$

$$RSS_2 \text{ (женщины)} = 1290.24$$



Тест Чоу. Пример

$$F = \frac{(5980.68 - [4009.93 + 1290.24])/2}{(4009.93 + 1290.24)/522} \approx 33.51$$

$$F_{2,522,0.01} = 4.65$$

Поскольку $F = 33.51 > 4.65$, то гипотеза H_0 отвергается для уровня значимости 1%, следовательно, для мужчин и для женщин имеет место разная зависимость. Нужно оценивать отдельные регрессии.



Тест Чоу. Gretl

gretl: результаты теста Чоу

Расширенная регрессия для теста Чоу
МНК, использованы наблюдения 1-526
Зависимая переменная: wage

	Коэффициент	Ст. ошибка	t-статистика	P-значение
const	-0,998027	1,02183	-0,9767	0,3292
educ	0,453477	0,0813414	5,575	3,98e-08 ***
male	1,19852	1,32504	0,9045	0,3661
ma_educ	0,0859990	0,103639	0,8298	0,4070

Среднее зав. перемен	5,896103	Ст. откл. зав. перемен	3,693086
Сумма кв. остатков	5300,170	Ст. ошибка модели	3,186469
R-квадрат	0,259796	Испр. R-квадрат	0,255542
F(3, 522)	61,07022	P-значение (F)	7,44e-34
Лог. правдоподобие	-1353,942	Крит. Акаике	2715,885
Крит. Шварца	2732,946	Крит. Хеннана-Куинна	2722,565

Тест Чоу для структурных изменений в точке male
 $F(2, 522) = 33,5109$ р-значение 0,0000

P – значение = 0. Следовательно H_0 отвергается,
есть различия в моделях для мужчин и для женщин.



Тест Чоу на прогнозную силу

С помощью теста Чоу можно проверять прогнозную силу модели. Для этого оценивается модель по так называемой обучающей выборке, это какая-то часть основной выборки (зачастую 70%). А затем оценивается модель по всей выборке и сравнивается сумма квадратов остатков в этих двух моделях.



Тест Чоу на прогнозную силу

Для проверки такой гипотезы, как и ранее, используется F-статистика Фишера. Расчетная статистика при этом имеет вид:

$$F = \frac{(RSS_p - RSS_f) / (n - n_f)}{(RSS_f) / (n_f - k - 1)} \stackrel{H_0}{\sim} F_{n-n_f, n_f-k-1},$$

RSS_p – это сумма квадратов остатков полной модели;

RSS_f – это сумма квадратов остатков модели, оцененной по обучающей выборке;

k – это число регрессоров без константы;

n – число наблюдений в полной выборке;

n_f – число наблюдений в обучающей выборке.



Эквивалентность теста Чоу и дамми переменных



Эквивалентность

! Тест Чоу эквивалентен тесту о значимости группы дамми переменных.

Если тест Чоу показывает, что есть различия в коэффициентах модели по двум наборам выборок, то можно оценить одну модель, но с дамми переменными.

Оценив модель с набором дамми переменных и проверив их совместную значимость, можно проверить ту же гипотезу, что и в teste Чоу.



Тест на незначимость группы переменных

Полная выборка

$$\widehat{wage} = -0.9 + 0.54educ$$

$$RSS_r = 5980.68$$

Полная выборка

$$\widehat{wage} = -1 + 1.2male + 0.45educ + 0.06meduc$$

$$RSS_{ur} = 5300.17$$

$$F = \frac{(5980.68 - 5300.17)/2}{5300.17/522} \approx 33.51$$

F-статистика для проверки значимости группы
дамми переменных.



Тест на незначимость группы переменных

Полная выборка

$$\widehat{wage} = -0.9 + 0.54educ$$

$$RSS_r = 5980.68$$

Полная выборка

$$\widehat{wage} = -1 + 1.2male + 0.45educ + 0.06meduc$$

$$RSS_{ur} = 5300.17$$

$$F = \frac{(5980.68 - 5300.17)/2}{5300.17/522} \approx 33.51$$

Сравнивая значения тестовой F-статистики с критическим, отвергаем гипотезу о незначимости группы дамми переменных.



Модель с дамми переменными и отдельно для каждой группы

Мужчины

$$\widehat{wage} = 0.2 + 0.54educ$$

Женщины

$$\widehat{wage} = -1 + 0.45educ$$

Вся выборка с дамми переменными

$$\widehat{wage} = -1 + 0.45educ + 1.2male + 0.09meduc$$

Если $male = 0$, то получаем уравнение для женщин:

$$\widehat{wage} = -1 + 0.45educ$$

Если $male = 1$, то получаем уравнение для мужчин:

$$\widehat{wage} = 0.2 + 0.54educ$$



Тест Чоу и тест на незначимость группы переменных

Мужчины: $RSS_1 = 4009.93$

Женщины: $RSS_2 = 1290.24$

Вся выборка с дамми переменными: $RSS_p = 5300.17$

$$F = \frac{(5980.68 - [4009.93 + 1290.24])/2}{(4009.93 + 1290.24)/522} \approx 33.51$$

$$F = \frac{(5980.68 - 5300.17)/2}{5300.17/522} \approx 33.51$$

F-статистики в teste о значимости группы дамми переменных и teste Чоу совпадают.



Тест Чоу и тест на незначимость группы переменных

$$F = \frac{(5980.68 - [4009.93 + 1290.24])/2}{(4009.93 + 1290.24)/522} \approx 33.51$$

$$F_{2,522,0.01} = 4.65$$

В каждом случае нулевая гипотеза отвергается при 1% уровне значимости.



Примеры применения дамми переменных



Дамми переменные для более двух категорий

- Например, в модели заработной платы уровень образования обозначен так:
 - Начальное;
 - Среднее;
 - Среднее специальное;
 - Высшее;
 - Аспирантура, докторантуре (кандидат/доктор наук).
- В таком случае в модель включается набор дамми переменных, без учета базовой категории.
- Иначе возникает точная линейная связь с константой!



Дамми переменные для более двух категорий

- Пусть начальное образование будет взято за базовую категорию, тогда набор дамми переменных будет таким:
- Среднее=1, если респондент имеет среднее образование, 0 – иначе;
- Высшее=1, если респондент имеет высшее образование, 0 – иначе;
- И т.д.

$$wage = \beta_1 + \beta_2 \text{Среднее} + \beta_3 \text{Специальное} + \\ + \beta_4 \text{Высшее} + \beta_5 \text{Аспирантура} + \varepsilon$$



Уравнения заработных плат для каждого уровня образования

$$wage = \beta_1 + \beta_2 \text{Среднее} + \beta_3 \text{Специальное} + \\ + \beta_4 \text{Высшее} + \beta_5 \text{Аспирантура} + \varepsilon$$

Начальное: $wage = \beta_1 + \varepsilon$

Среднее: $wage = \beta_1 + \beta_2 + \varepsilon$

Специальное: $wage = \beta_1 + \beta_3 + \varepsilon$

Высшее: $wage = \beta_1 + \beta_4 + \varepsilon$

Аспирантура: $wage = \beta_1 + \beta_5 + \varepsilon$



Сезонные дамми переменные

Часто в распоряжении исследователя имеются недельные, месячные или квартальные данные.

Если качественная переменная имеет k градаций, то в модель надо ввести $k - 1$ фиктивных переменных.

Если данные квартальные, то

$D1 = 1$, если наблюдение относится к 1-му кварталу
и 0, если не относится;

$D2 = 1$, если наблюдение относится к 2-му кварталу
и 0, если не относится;

$D3 = 1$, если наблюдение относится к 3-му кварталу
и 0, если не относится.



Сезонные дамми переменные. Пример

Рассмотрим квартальные данные.

В качестве базы выберем 4-ый квартал, тогда:

Модель: $Y = \alpha + \beta_1 D1 + \beta_2 D2 + \beta_3 D3 + \beta_4 X + \varepsilon$.

Оцененное уравнение регрессии:

$\hat{Y} = \hat{\alpha} + \hat{\beta}_1 D1 + \hat{\beta}_2 D2 + \hat{\beta}_3 D3 + \hat{\beta}_4 X,$

где $\hat{\alpha}, \dots, \hat{\beta}_4$ – оценки коэффициентов регрессии.



Сезонные дамми переменные

$$\hat{Y} = \hat{\alpha} + \hat{\beta}_1 D1 + \hat{\beta}_2 D2 + \hat{\beta}_3 D3 + \hat{\beta}_4 X$$

Поквартальные зависимости:

$$\hat{Y} = \hat{\alpha} + \hat{\beta}_1 + \hat{\beta}_4 X \text{ - для 1-го квартала;}$$

$$\hat{Y} = \hat{\alpha} + \hat{\beta}_2 + \hat{\beta}_4 X \text{ - для 2-го квартала;}$$

$$\hat{Y} = \hat{\alpha} + \hat{\beta}_3 + \hat{\beta}_4 X \text{ - для 3-го квартала;}$$

$$\hat{Y} = \hat{\alpha} + \hat{\beta}_4 X \text{ - для 4-го квартала (базового).}$$





Пример

- Файл «Chow_2.xls» содержит квартальные данные по экономике Баккардии с I квартала 2015 года по IV квартал 2022 года (всего 32 наблюдения):
- C_t – конечное потребление в период t ;
- Y_t – располагаемый доход.
- Показатели выражены в миллиардах баккардийских крон 2015 года.
- Также в таблице есть индикаторы для каждого квартала:
- $D1 = 1$, если наблюдение относится к 1–му кварталу какого-либо года и 0, если не относится;

...

- $D4 = 1$, если наблюдение относится к 4–му кварталу какого-либо года и 0, если не относится.

Источник: Борзых, Вакуленко, Фурманов (2021)



Пример. Вопросы



- Требуется проверить, есть ли в динамике потребления сезонные колебания, не обусловленные изменениями располагаемого дохода.
- Оцените модель с учетом сезонности и проверьте совместную значимость сезонных дамми переменных:

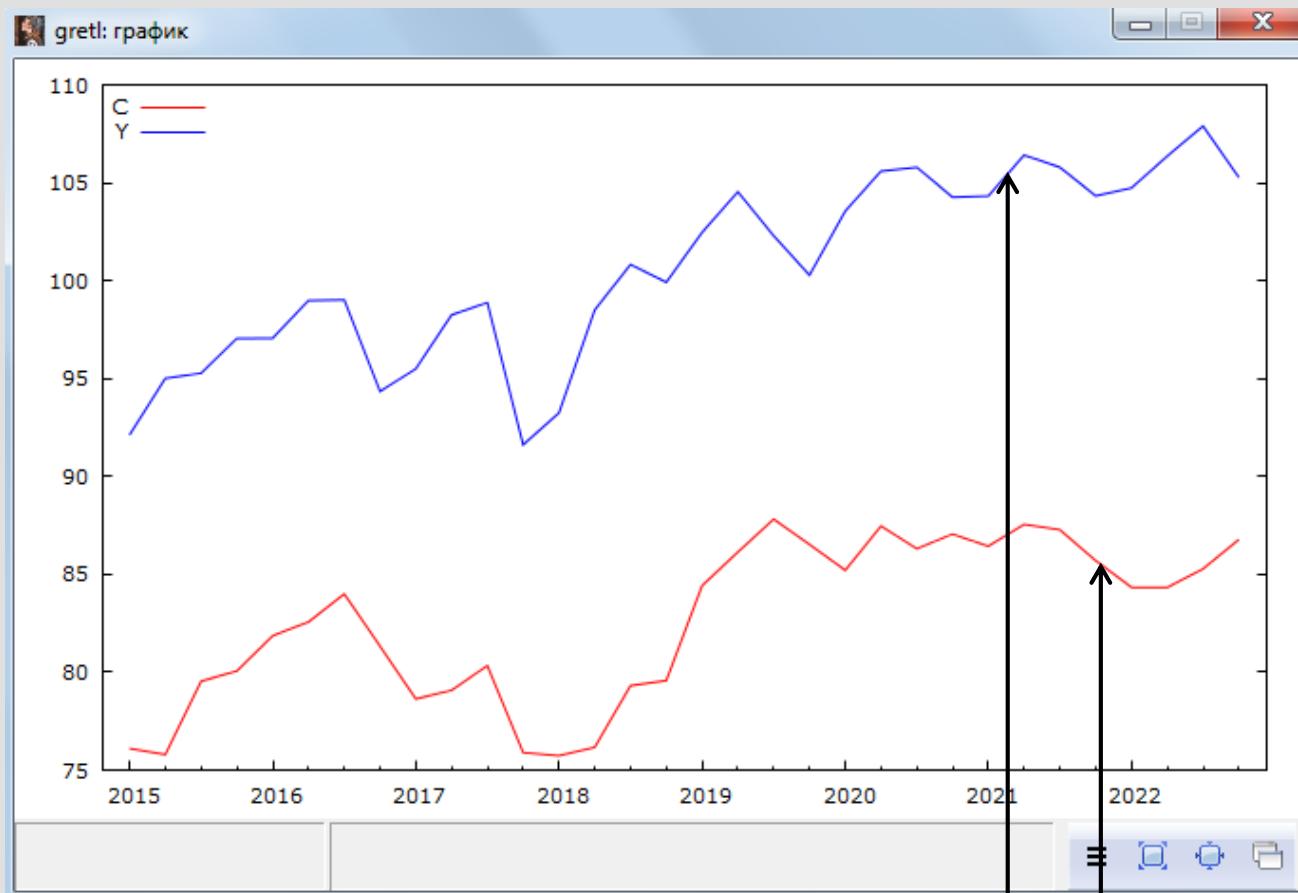
$$C_t = \gamma_0 + \gamma_1 Y_t + \gamma_2 D2_t + \gamma_3 D3_t + \gamma_4 D4_t + \varepsilon_t$$

- Оцените модель без константы, но со всеми категориями для дамми переменных.



Рассмотрим графики

Стрелки можно подвинуть довольно произвольно.



Красная линия – это конечное потребление (C),
синяя линия – это располагаемый доход (Y).



Решение в Gretl

$$\hat{C} = 4.83 + 0.77Y - 1.21D2 - 0.06D3 + 0.88D4$$

gretl: модель 1

Файл Правка Тесты Сохранить Графики Анализ LaTeX

Модель 1: МНК, использованы наблюдения 1-32
Зависимая переменная: C

	Коэффициент	Ст. ошибка	t-статистика	P-значение
const	4,82551	8,27462	0,5832	0,5646
Y	0,774491	0,0831436	9,315	6,36e-010 ***
D2	-1,20734	1,06937	-1,129	0,2688
D3	-0,0631452	1,07397	-0,05880	0,9535
D4	0,877791	1,04843	0,8372	0,4098
Среднее зав. перемен	82,64719	Ст. откл. зав. перемен	4,090451	
Сумма кв. остатков	118,5200	Ст. ошибка модели	2,095145	
R-квадрат	0,771499	Испр. R-квадрат	0,737647	
F(4, 27)	22,79038	P-значение (F)	2,53e-08	
Лог. правдоподобие	-66,35557	Крит. Акаике	142,7111	
Крит. Шварца	150,0398	Крит. Хеннана-Куинна	145,1404	

Все сезонные дамми переменные по отдельности
незначимы на любом разумном уровне значимости.



Уравнения регрессий для каждого квартала

Поквартальные зависимости:

$$\hat{C} = 4.83 + 0.77Y \text{ - для 1-го квартала (базового);}$$

$$\hat{C} = 4.83 - 1.21 + 0.77Y = 3.62 + 0.77Y \text{ - для 2-го квартала;}$$

$$\hat{C} = 4.83 - 0.06 + 0.77Y = 4.76 + 0.77Y \text{ - для 3-го квартала;}$$

$$\hat{C} = 4.83 + 0.88 + 0.77Y = 5.71 + 0.77Y \text{ - для 4-го квартала.}$$



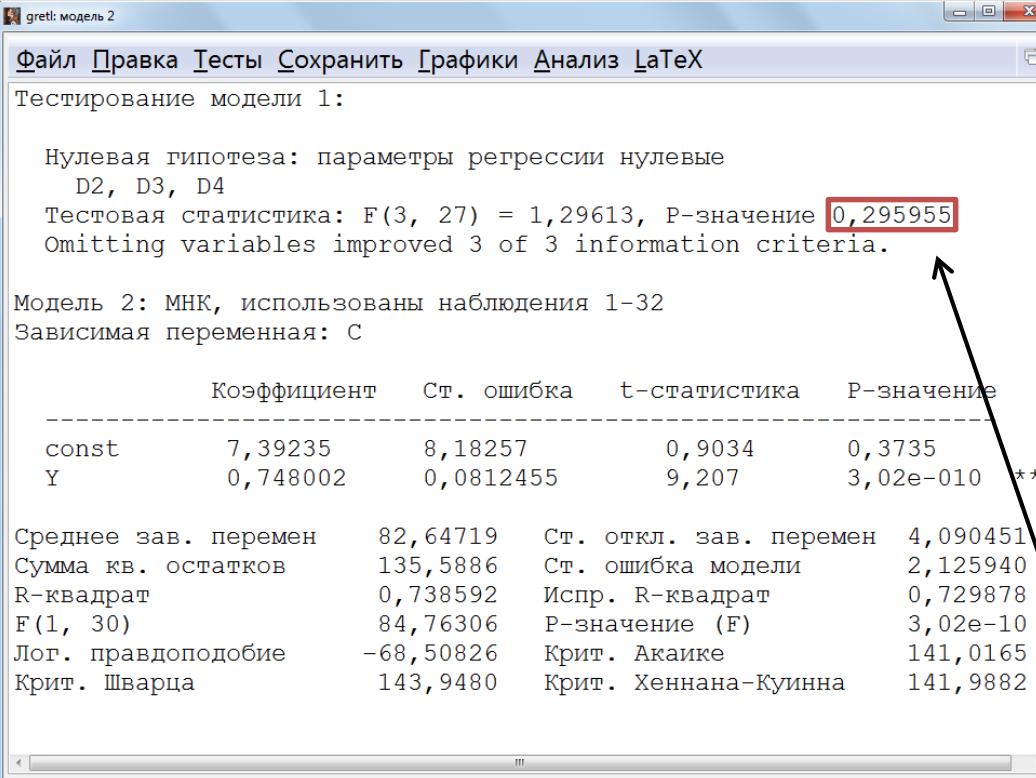
Проверка гипотезы

- Для проверки гипотезы о сезонности необходимо проверить:
- $H_0: \gamma_2 = \gamma_3 = \gamma_4 = 0$ (нет сезонности)
- $H_1: \gamma_2^2 + \gamma_3^2 + \gamma_4^2 > 0$ (есть сезонность)



Проверка совместной значимости дамми переменных

$$\hat{C} = 7.39 + 0.75Y$$



gretl: модель 2

Файл Правка Тесты Сохранить Графики Анализ LaTeX

Тестирование модели 1:

Нулевая гипотеза: параметры регрессии нулевые
D2, D3, D4

Тестовая статистика: F(3, 27) = 1,29613, Р-значение 0,295955
Omitting variables improved 3 of 3 information criteria.

Модель 2: МНК, использованы наблюдения 1-32
Зависимая переменная: C

	Коэффициент	Ст. ошибка	t-статистика	Р-значение
const	7,39235	8,18257	0,9034	0,3735
Y	0,748002	0,0812455	9,207	3,02e-010 **

Среднее зав. перемен 82,64719 Ст. откл. зав. перемен 4,090451
Сумма кв. остатков 135,5886 Ст. ошибка модели 2,125940
R-квадрат 0,738592 Испр. R-квадрат 0,729878
F(1, 30) 84,76306 Р-значение (F) 3,02e-10
Лог. правдоподобие -68,50826 Крит. Акаике 141,0165
Крит. Шварца 143,9480 Крит. Хеннана-Куинна 141,9882

Все сезонные дамми переменные совместно незначимы.
Не выявлено сезонных колебаний конечного
потребления, не связанных с располагаемым доходом.



Модель без константы, но с дамми переменными для всех кварталов

$$\hat{C} = 0.77Y + 4.83D1 + 3.62D2 + 4.76D3 + 5.70D4$$

Файл Правка Тесты Сохранить Графики Анализ LaTeX				
Модель 3: МНК, использованы наблюдения 1-32				
Зависимая переменная: C				
	Коэффициент	Ст. ошибка	t-статистика	P-значение
---	---	---	---	---
Y	0,774491	0,0831436	9,315	6,36e-010 ***
D1	4,82551	8,27462	0,5832	0,5646
D2	3,61817	8,48861	0,4262	0,6733
D3	4,76236	8,51035	0,5596	0,5804
D4	5,70330	8,31696	0,6857	0,4987
Среднее зав. перемен	82,64719	Ст. откл. зав. перемен	4,090451	
Сумма кв. остатков	118,5200	Ст. ошибка модели	2,095145	
R-квадрат	0,771499	Испр. R-квадрат	0,737647	
F (4, 27)	22,79038	P-значение (F)	2,53e-08	
Лог. правдоподобие	-66,35557	Крит. Акаике	142,7111	
Крит. Шварца	150,0398	Крит. Хеннана-Куинна	145,1404	



Уравнения регрессий для каждого квартала

$$\hat{C} = 0.77Y + 4.83D1 + 3.62D2 + 4.76D3 + 5.70D4$$

Поквартальные зависимости:

$$\hat{C} = 0.77Y + 4.83 \text{ - для 1-го квартала;}$$

$$\hat{C} = 0.77Y + 3.62 \text{ - для 2-го квартала;}$$

$$\hat{C} = 0.77Y + 4.76 \text{ - для 3-го квартала;}$$

$$\hat{C} = 0.77Y + 5.70 \text{ - для 4-го квартала.}$$

- Результаты оказываются такими же, как и для случая с константой.



Влиятельные наблюдения и выбросы

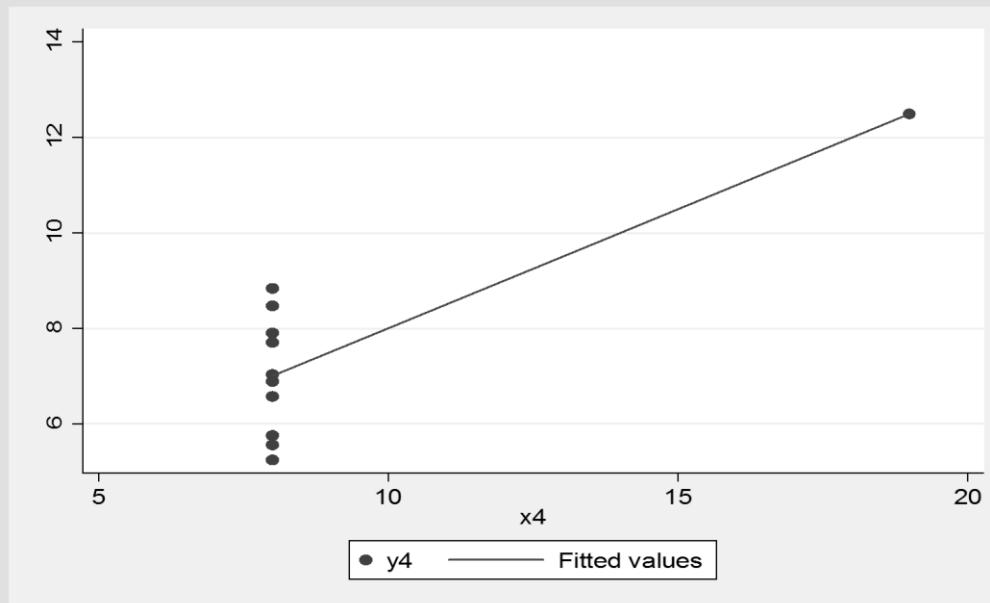


Влиятельные наблюдения

Влиятельными наблюдениями называются те наблюдения, присутствие которых в выборке привносит значительные изменения в оценки коэффициентов.



Пример влиятельного наблюдения (Anscombe, 1973)



Оценённая регрессия оказывается значимой и имеет коэффициент детерминации, равный 0.67. При этом после удаления точки в правом верхнем углу вообще невозможно говорить о зависимости между признаками, потому что среди оставшихся точек нет никакого разброса по горизонтальной оси.



DFBETA

Чтобы измерить чувствительность оценок к удалению наблюдений можно использовать показатель $DFBETA_{ij}$, характеризующий изменение оценки коэффициента β_j после удаления наблюдения i в выборке.

$DFBETA_{ij}$ – это разница между оценками коэффициента β_j в полной выборке и в выборке после удаления наблюдения i .



Выбросы

Интерес для исследователя могут представлять не только влиятельные наблюдения, но и **выбросы (outliers)** – наблюдения, плохо вписывающиеся в оценённую зависимость (иначе говоря, далеко отстоящие от линии регрессии).

Для их обнаружения можно пользоваться **стьюдентизированными остатками**.



Стьюдентизированный остаток

Стьюдентизированный остаток — это остаток, деленный на свое стандартное отклонение при условии исключения данного наблюдения, т.е.

$$e'_i = \frac{e_i}{S(i)\sqrt{1 - h_i}}$$

e_i — остаток для конкретного наблюдения, полученный по уравнению регрессии, построенному с учетом всех наблюдений;

$S(i)$ — стандартное отклонение остатков, полученное по уравнению регрессии, построенному по тому же набору наблюдений, но без учета наблюдения i .

h_i — это диагональный элемент матрицы проектора $X(X'X)^{-1}X'$.



Решающее правило

Стьюдентизированные остатки имеют t -распределение с $n - p - 1$ степенями свободы. Соответственно, мы можем использовать квантили t -распределения для проверки того, насколько статистически значимо определенное наблюдение является выбросом.



А можно и проще

Большие значение таких остатков (например, большие трёх, если руководствоваться **правилом трёх сигм**) можно воспринимать как сигнал, призывающий обратить внимание на нетипичное наблюдение.



Значимость наблюдений. Gretl

- Леверидж
- Воздействие
- DFFITS



Леверидж

Леверидж: h_i – это диагональный элемент матрицы проектора $X(X'X)^{-1}X'$.

Точки левериджа (“Leverage points”):

$$h_i > 2k/n,$$

где k – это число регрессоров;

n – число наблюдений.

В Gretl эти точки помечаются звездочкой (*).



Воздействие

$$e_i \frac{h_i}{1 - h_i},$$

где e_i – остаток модели;

h_i – это диагональный элемент матрицы проектор $X(X'X)^{-1}X'$.



$$DFFIT_i = \widehat{Y}_i - \widehat{Y}_{i(i)},$$

где \widehat{Y}_i и $\widehat{Y}_{i(i)}$ – это предсказанные по модели значения с учетом и без учета наблюдения i .

$$DFFITS_i = e'_i \sqrt{\frac{h_i}{1 - h_i}},$$

где e'_i – стьюдентизированный остаток.

Если $DFFITS_i > 2\sqrt{\frac{k}{n}}$, то i -е наблюдение может быть выбросом,
где k – это число регрессоров, а n – это число наблюдений.



Пример. Экономика Баккардии

Вернемся к примеру об экономики Баккардии

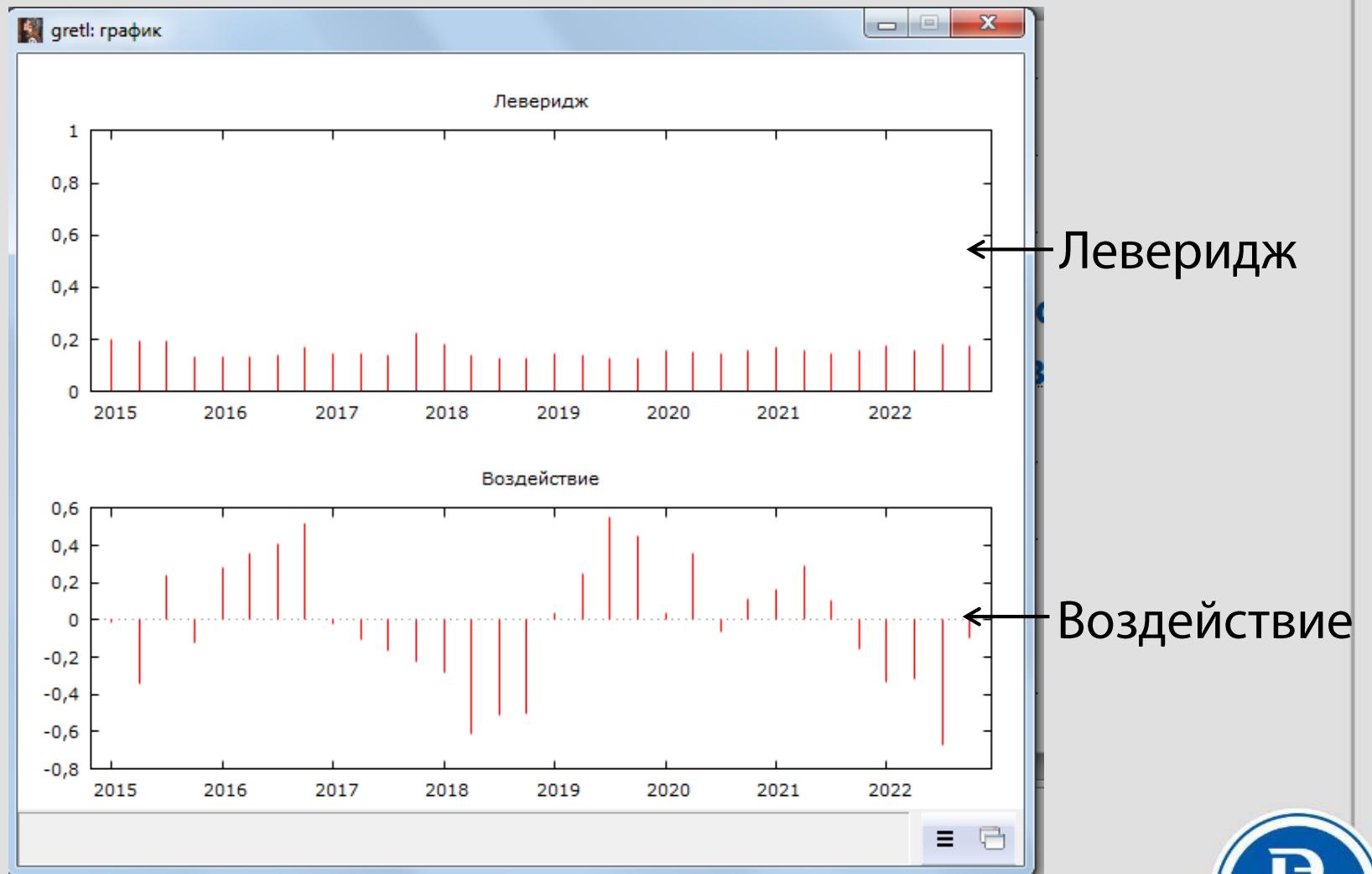
$$C_t = \gamma_0 + \gamma_1 Y_t + \gamma_2 D2_t + \gamma_3 D3_t + \gamma_4 D4_t + \varepsilon_t,$$

C_t – конечное потребление в период t ;

Y_t – располагаемый доход.



Пример. Экономика Баккардии



Пример. Gretl

	Остатки u	Леверидж $0 \leq h \leq 1$	Воздействие $u \cdot h / (1-h)$	DFFITS
2015:1	-0,061626	0,202	-0,015621	-0,016
2015:2	-1,4026	0,196	-0,34108	-0,365
2015:3	0,99187	0,196	0,24129	0,257
2015:4	-0,79766	0,136	-0,12504	-0,160
2016:1	1,8624	0,132	0,28248	0,371
2016:2	2,285	0,137	0,36157	0,470
2016:3	2,5475	0,139	0,41022	0,533
2016:4	2,5612	0,169	0,52136	0,615

	Остатки u	Леверидж $0 \leq h \leq 1$	Воздействие $u \cdot h / (1-h)$	DFFITS
2021:3	0,58423	0,148	0,10148	0,124
2021:4	-0,80595	0,160	-0,15321	-0,180
2022:1	-1,598	0,175	-0,33783	-0,384
2022:2	-1,663	0,159	-0,31467	-0,375
2022:3	-3,0422	0,180	-0,66877	-0,776
2022:4	-0,45398	0,175	-0,096316	-0,108

Точки левериджа не обнаружены

Cross-validation criterion = 161,717

Точки левериджа не обнаружены.



Учет выбросов

Выбросы нужно обнаруживать и либо исключать, либо особым образом моделировать.

В этом могут помочь [дамми переменные](#).

Задавать дамми переменные можно так: 1, если это наблюдение выброс, 0, иначе.



Основные выводы

Фиктивные (дамми) переменные – бинарные переменные, кодирующие качественные признаки.

С помощью дамми переменных можно моделировать:

- Неоднородность выборки по каким-либо признакам;
- Сезонность данных;
- Выбросы и влиятельные наблюдения;
- Структурные сдвиги.

Тест на структурные изменения в модели
или неоднородность данных: тест Чоу или значимость группы
дамми переменных.



Литература

Катышев, Магнус, Пересецкий (2005). Эконометрика.
Начальный курс. Глава 3.4, 4.2.

К. Доугерти (1999). Введение в эконометрику. М. Инфра-М.
Глава 9.

Борзых Д. А., Вакуленко Е. С., Фурманов К. К. Эконометрика:
работа с данными на компьютере. Практикум: Элементы
теории. Практические задания. Ответы и решения.
Издательская группа URSS, 2021. Глава 2.

Вакуленко Е. С., Ратникова Т. А., Фурманов К.
К. Эконометрика (продвинутый курс). Применение пакета
Stata. М. : Юрайт, 2020. Глава 5.

