



NATIONAL RESEARCH  
UNIVERSITY

## **Лекция 12.11.21, 2 модуль часть 1**

### **Фиктивные (дамми) переменные и тест Чоу**

**Демидова О.А.**

**E-mail: [demidova@hse.ru](mailto:demidova@hse.ru)**

# План лекции

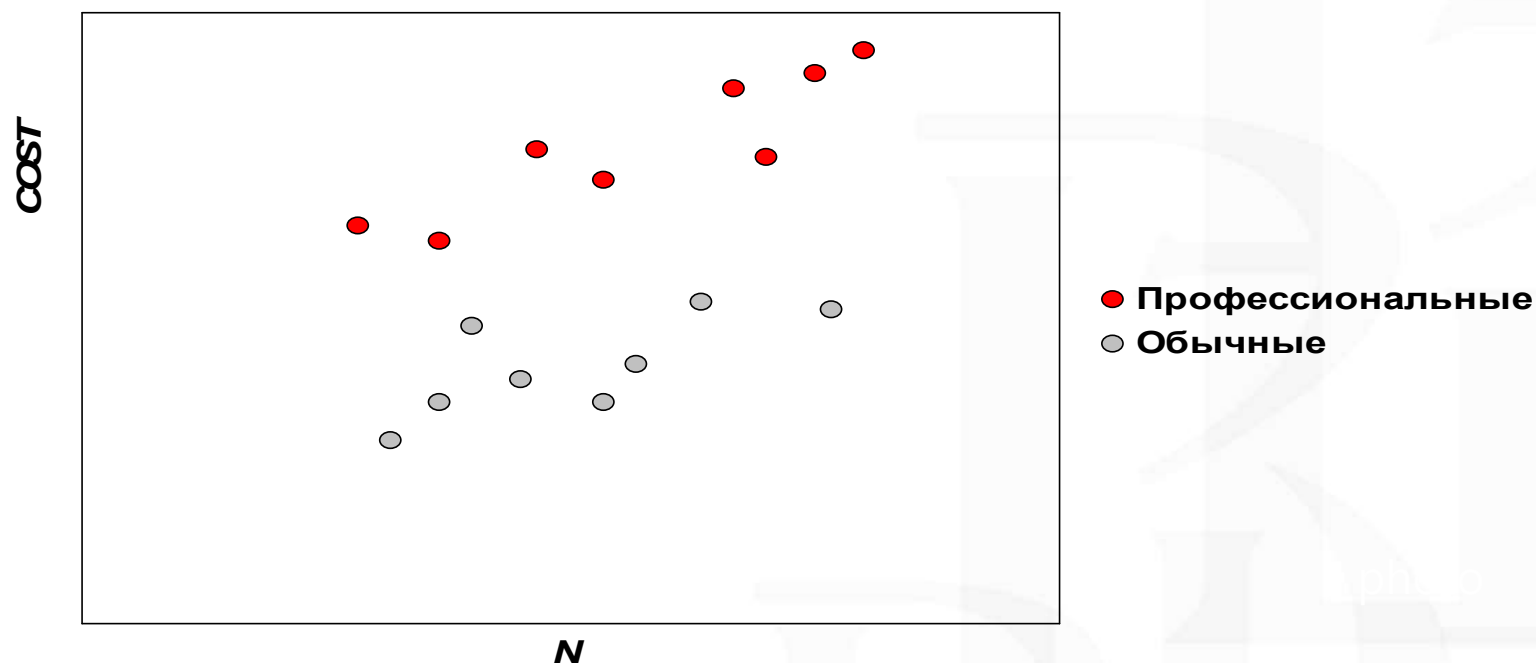
- Фиктивные (дамми) переменные
- Тест Чоу
- Моделирование сезонности с помощью дамми переменных

# Фиктивные (дамми) переменные

## Определение

Для исследования влияния качественных признаков в модель можно вводить бинарные (дамми) переменные, которые, как правило, принимают значение 1, если данный качественный признак присутствует в наблюдении, и значение 0 при его отсутствии.

# Пример использования дамми переменной



$COST$  – годовые издержки 74 средних школ в Шанхае в середине 1980-х годов,  $N$  – количество обучавшихся в них учеников.

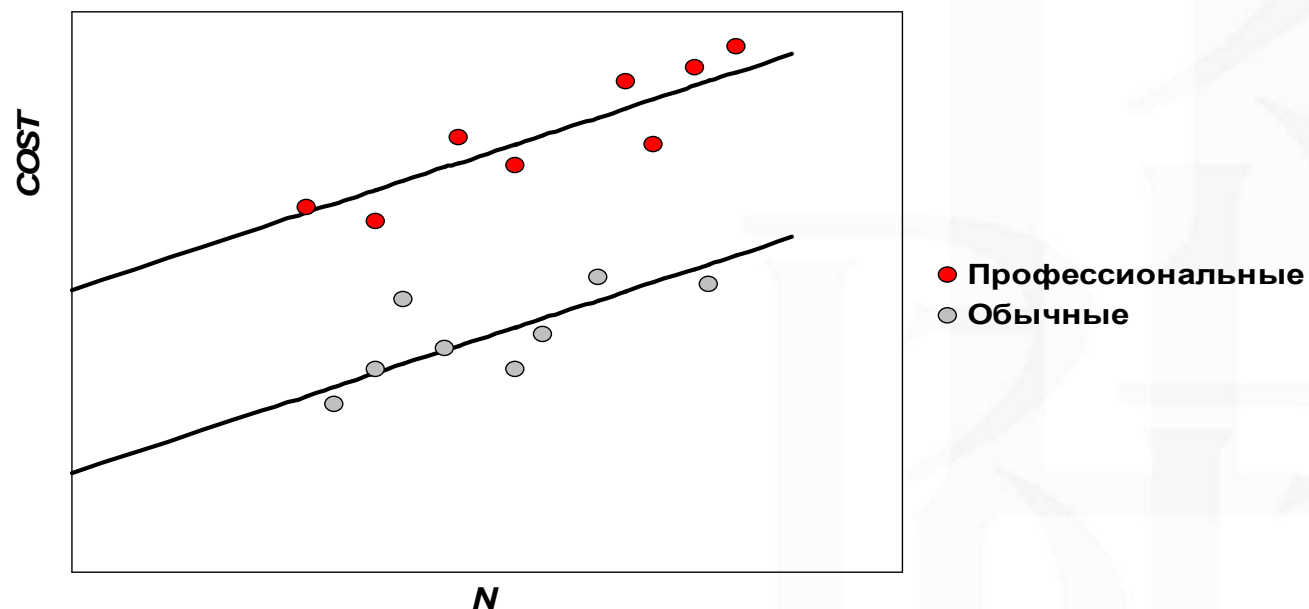
# Пример использования дамми переменной

Затраты в профессиональных школах больше, т.к. для обучения там используется специальное оборудование.

Если оценивать регрессии отдельно для профессиональных и обычных школ, то размеры выборок уменьшатся, что снизит точность оценивания.

# Пример использования дамми переменной

Предположим, что коэффициенты наклона в регрессиях для профессиональных и обычных школах одинаковы, а свободные члены различаются.



# Пример использования дамми переменной

Мы предполагаем, что постоянные затраты для двух типов школ различаются, а предельные затраты у них одинаковы.

Обычные школы

$$\text{COST} = \beta_0 + \beta_1 N + \varepsilon$$

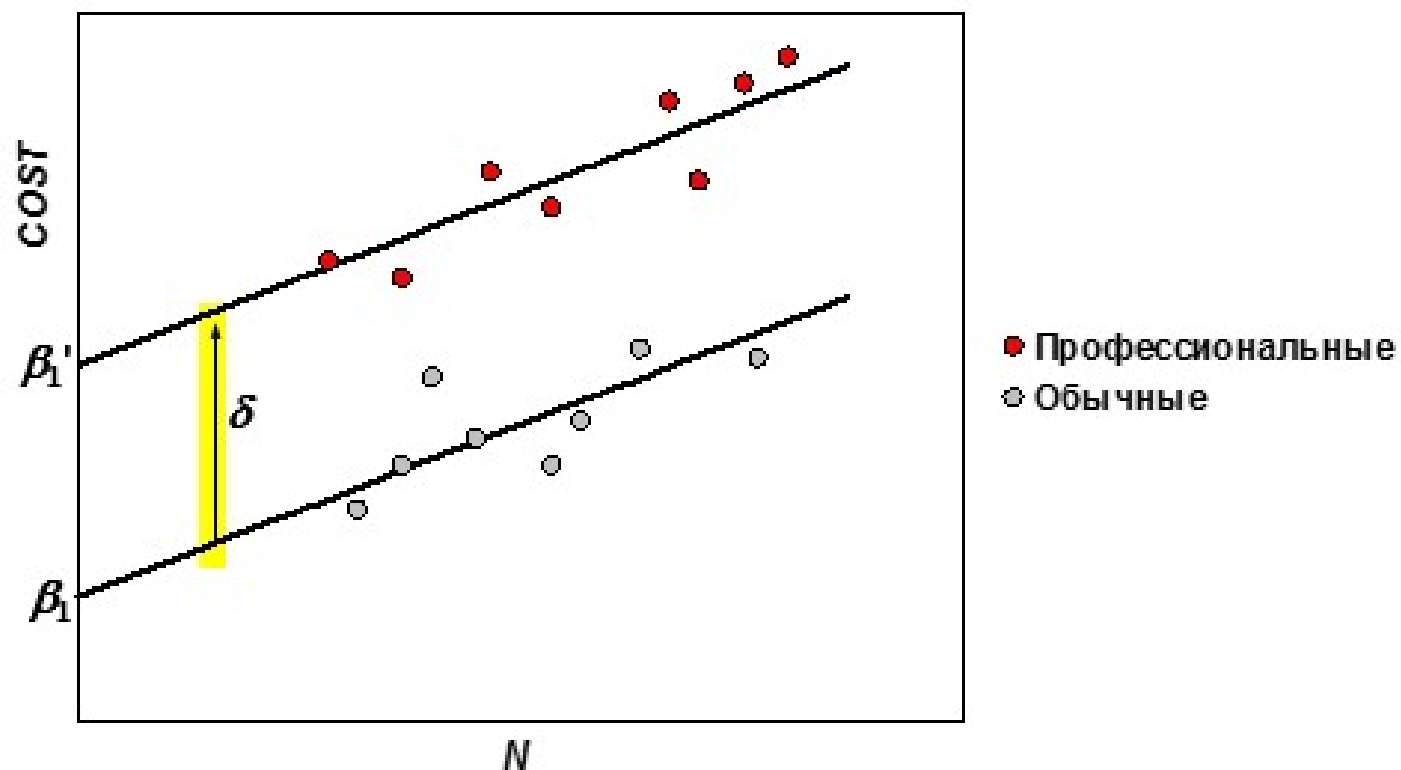
Профессиональные школы

$$\text{COST} = \beta_0 + \beta'_1 N + \varepsilon$$

# Пример использования дамми переменной

Обозначим  $\delta$  разность свободных членов:

$$\delta = \beta'_0 - \beta_0$$





# Пример использования дамми переменной

Тогда  $\beta'_0 = \beta_0 + \delta$  и мы можем переписать регрессию для профессиональных школ.

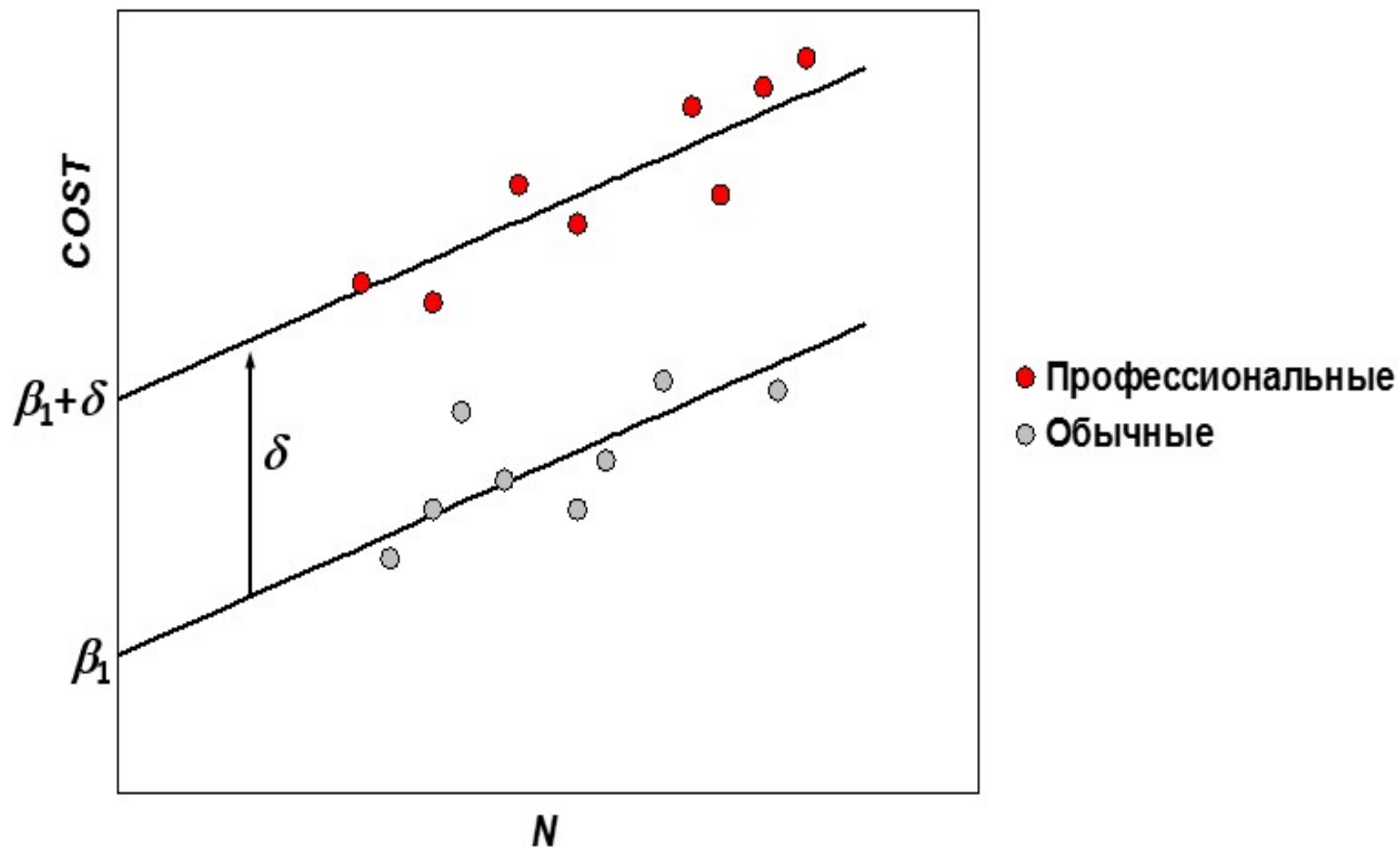
Обычные школы

$$\text{COST} = \beta_0 + \beta_1 N + \varepsilon$$

Профессиональные школы

$$\text{COST} = \beta_0 + \delta + \beta_1 N + \varepsilon$$

# Пример использования дамми переменной



## Пример использования дамми-переменной

Введем дамми- переменную OCC, которая равна 0 для обычных школ и 1 для профессиональных.

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

Общее уравнение

$$\text{COST} = \beta_0 + \delta \text{OCC} + \beta_1 N + \varepsilon$$

Обычные школы,  $\text{OCC} = 0$

$$\text{COST} = \beta_0 + \beta_1 N + \varepsilon$$

Профессиональные школы,  $\text{OCC} = 1$

$$\text{COST} = \beta_0 + \delta + \beta_1 N + \varepsilon$$

## Пример использования дамми-переменной

В последней колонке сформирована дамми - переменная. В приведенной таблице указаны данные лишь для 5 школ. В последней колонке сформирована дамми - переменная.

School	Type	COST	N	OCC
1	Occupational	345,000	623	1
2	Occupational	537,000	653	1
3	Regular	170,000	400	0
4	Occupational	526,000	663	1
5	Regular	100,000	563	0

# Пример использования дамми переменной

В таблице приведены результаты оценивания регрессии COST на N и OCC.

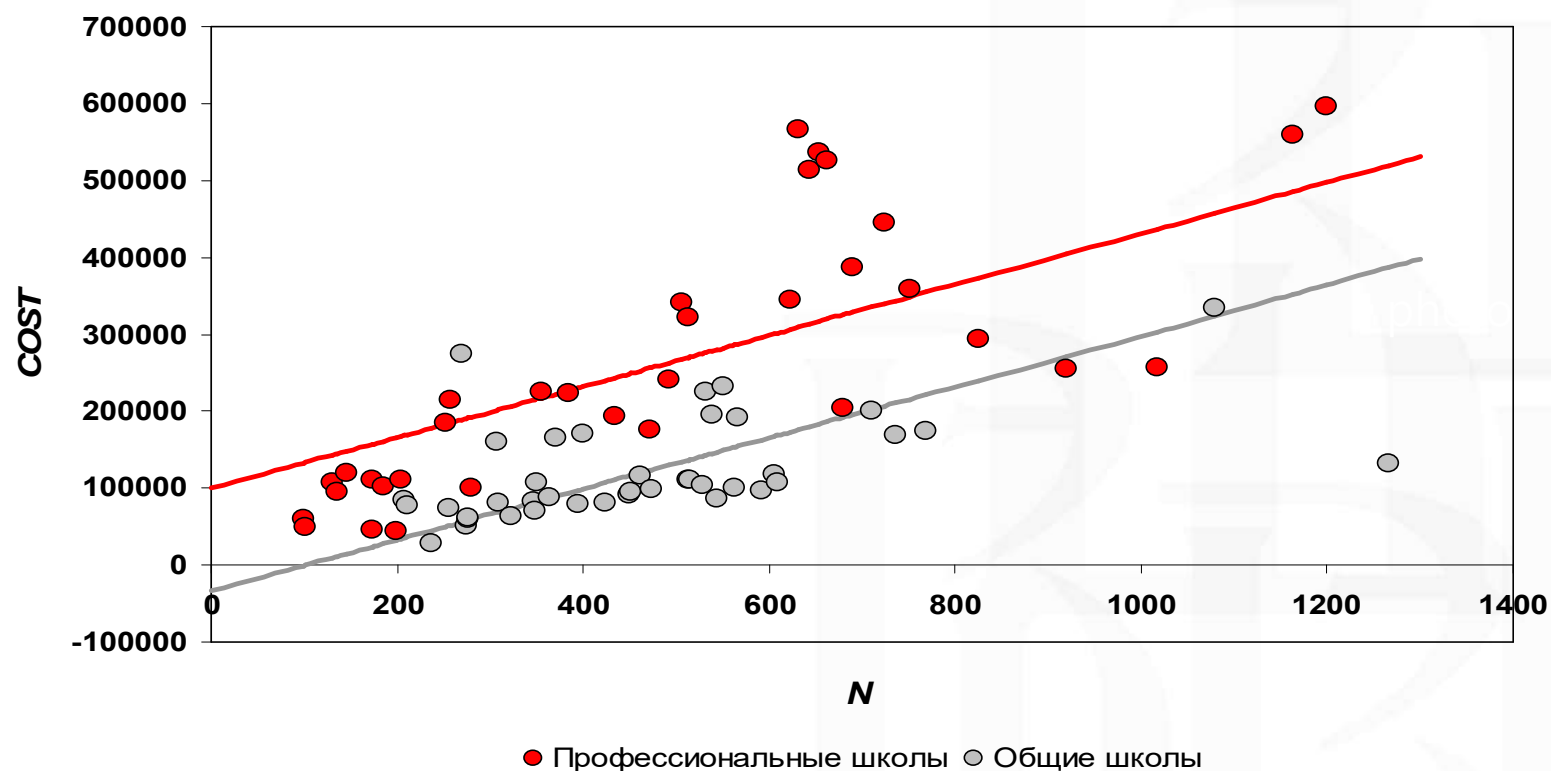
Source	SS	df	MS
Model	9.0582e+11	2	4.5291e+11
Residual	5.6553e+11	71	7.9652e+09
Total	1.4713e+12	73	2.0155e+10

Number of obs = 74  
 F( 2, 71) = 56.86  
 Prob > F = 0.0000  
 R-squared = 0.6156  
 Adj R-squared = 0.6048  
 Root MSE = 89248

COST	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
N	331.4493	39.75844	8.337	0.000	252.1732	410.7254
OCC	133259.1	20827.59	6.398	0.000	91730.06	174788.1
_cons	-33612.55	23573.47	-1.426	0.158	-80616.71	13391.61

# Пример использования дамми переменной

На диаграмме изображены наблюдения для 74 школ и проведены линии регрессии для одинаковых коэффициентов наклона двух типов школ.



# Дамми переменные для моделирования разницы коэффициентов наклона

Ослабим требование об одинаковых предельных издержках (коэффициентах наклона) для обычных и профессиональных школ.

Введем переменную NOCC, произведение  $N$  и OCC.

Для обычных школ переменная OCC равна 0 и, следовательно, NOCC также равна 0.

Для профессиональных школ переменная OCC равна 1, следовательно, переменная NOCC равна  $N$ .

# Дамми переменные для моделирования разницы коэффициентов наклона

$$COST = \beta_0 + \delta OCC + \beta_1 N + \lambda NOCC + \varepsilon$$

Общие школы  
( $OCC = NOCC = 0$ )

$$COST = \beta_0 + \beta_1 N + \varepsilon$$

Профессиональные школы  
( $OCC = 1; NOCC = N$ )

$$COST = (\beta_0 + \delta) + (\beta_1 + \lambda)N + \varepsilon$$

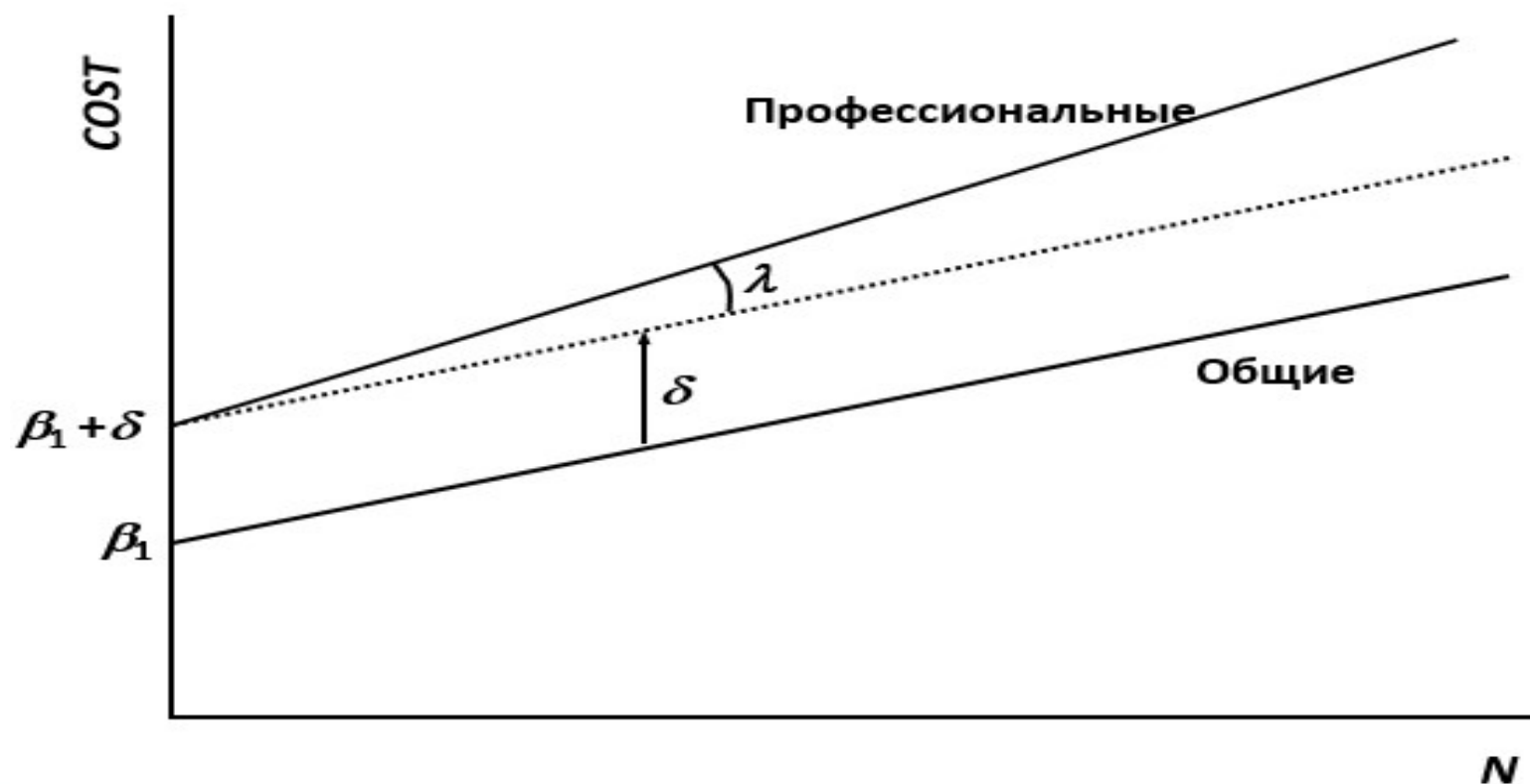


# Дамми переменные для моделирования разницы коэффициентов наклона

Предельные издержки на одного студента профессиональной школы больше на  $\lambda$  по сравнению с расходами на одного студента обыкновенной школы, постоянные издержки различаются на  $\delta$ .

# Дамми переменные для моделирования разницы коэффициентов наклона

Диаграмма иллюстрирует разницу в коэффициентах наклона графически.



# Дамми переменные для моделирования разницы коэффициентов наклона

## Результаты оценивания модели

```
. reg COST N OCC NOCC
```

Source	SS	df	MS
Model	1.0009e+12	3	3.3363e+11
Residual	4.7045e+11	70	6.7207e+09
Total	1.4713e+12	73	2.0155e+10

Number of obs = 74  
 F( 3, 70) = 49.64  
 Prob > F = 0.0000  
 R-squared = 0.6803  
 Adj R-squared = 0.6666  
 Root MSE = 81980

COST	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
N	152.2982	60.01932	2.537	0.013	32.59349	272.003
OCC	-3501.177	41085.46	-0.085	0.932	-85443.55	78441.19
NOCC	284.4786	75.63211	3.761	0.000	133.6351	435.3221
_cons	51475.25	31314.84	1.644	0.105	-10980.24	113930.7

# Дамми переменные для моделирования разницы коэффициентов наклона

$$\widehat{COST} = 51000 - 4000 OCC + 152N + 284NOCC$$

Общие школы  
( $OCC = NOCC = 0$ )

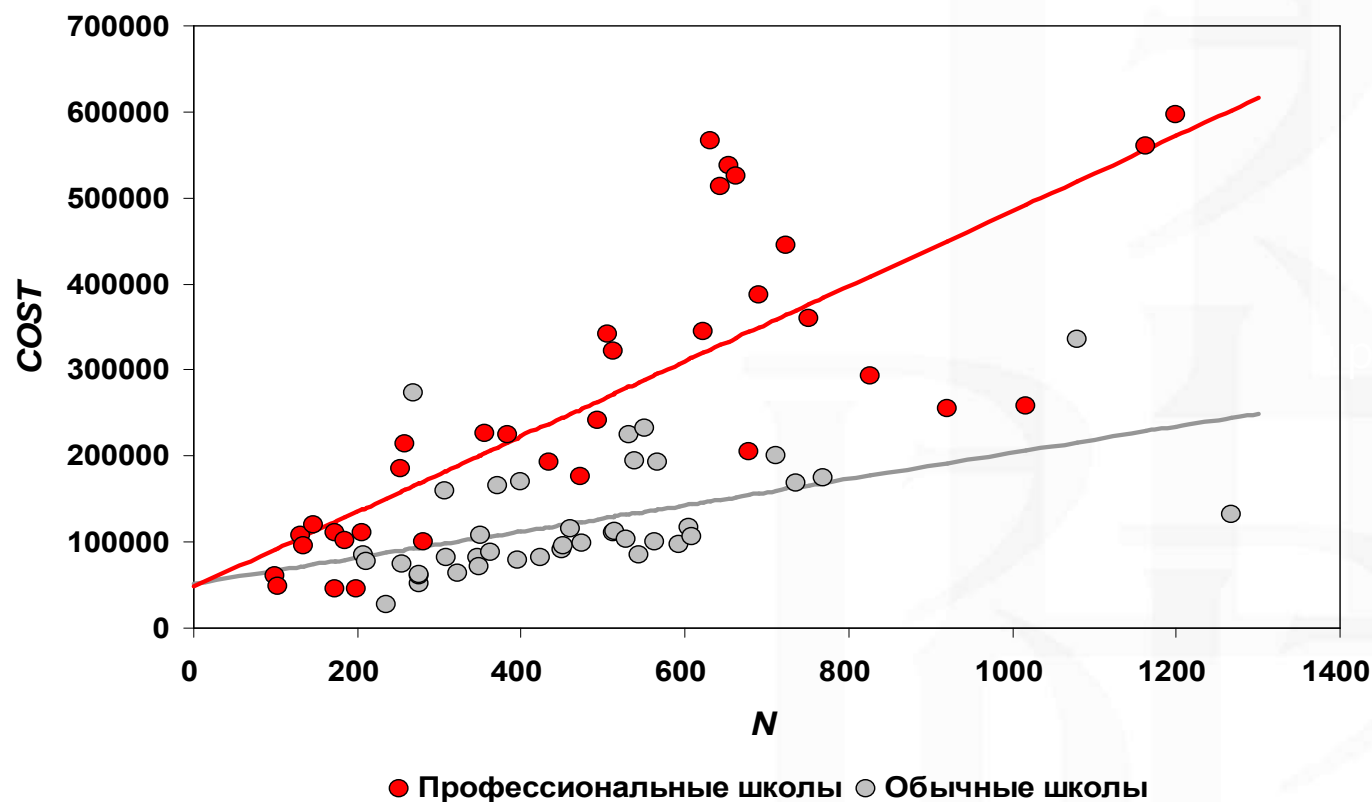
$$\widehat{COST} = 51000 + 152N$$

Профессиональные школы  
( $OCC = 1; NOCC = N$ )

$$\widehat{COST} = 51000 - 4000 + 152N + 284N = 47000 + 436N$$

# Дамми переменные для моделирования разницы коэффициентов наклона

На рисунке приведены графики оцененных регрессий для профессиональных и обычных школ.



# Дамми переменные для моделирования разницы коэффициентов наклона

$t$  – статистика переменной NOCC равна 3.76,  $p\text{-value} < 0.05$ , этот коэффициент значим, следовательно, предельные расходы для студентов обычных и профессиональных школ различаются.

Коэффициент при переменной OCC незначим, следовательно, постоянные расходы не различаются.

# Проверка гипотезы о совместной значимости переменных

$$\text{COST} = \beta_0 + \delta \text{OCC} + \beta_1 \text{N} + \lambda \text{NOCC} + \varepsilon$$

$$H_0: \delta = \lambda = 0$$

$$H_1: \delta \neq 0 \text{ или } \lambda \neq 0$$

```
. reg COST N OCC NOCC
```

Source	SS	df	MS
Model	1.0009e+12	3	3.3363e+11
Residual	4.7045e+11	70	6.7207e+09
Total	1.4713e+12	73	2.0155e+10

```
Number of obs =      74
F(   3,      70) =    49.64
Prob > F       =    0.0000
R-squared      =    0.6803
Adj R-squared  =    0.6666
Root MSE      =    81980
```

```
. reg COST N
```

Source	SS	df	MS
Model	5.7974e+11	1	5.7974e+11
Residual	8.9160e+11	72	1.2383e+10
Total	1.4713e+12	73	2.0155e+10

```
Number of obs =      74
F(   1,      72) =    46.82
Prob > F       =    0.0000
R-squared      =    0.3940
Adj R-squared  =    0.3856
Root MSE      =    1.1e+05
```

## Проверка гипотезы о совместной значимости переменных

$$\text{COST} = \beta_0 + \delta \text{OCC} + \beta_1 N + \lambda \text{NOCC} + \varepsilon$$

$$H_0: \delta = \lambda = 0$$

$$H_1: \delta \neq 0 \text{ или } \lambda \neq 0$$

$$F(2,70) = \frac{(8.92 \times 10^{11} - 4.70 \times 10^{11})/2}{4.70 \times 10^{11}/70} = 31.4$$

$$F(2,70)_{\text{crit}, 0.1\%} = 7.6$$

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

Следовательно, есть различия между школами в предельных или постоянных издержках.

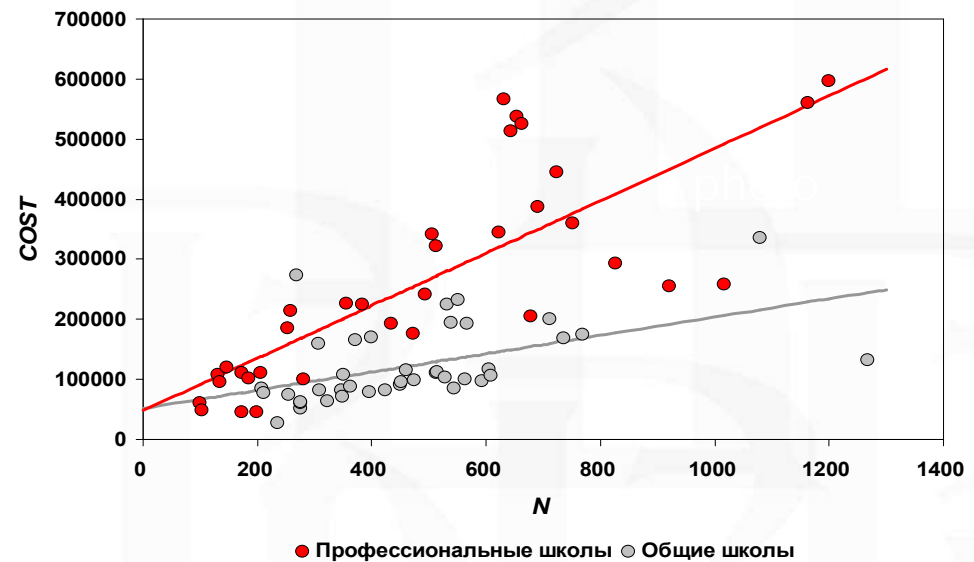
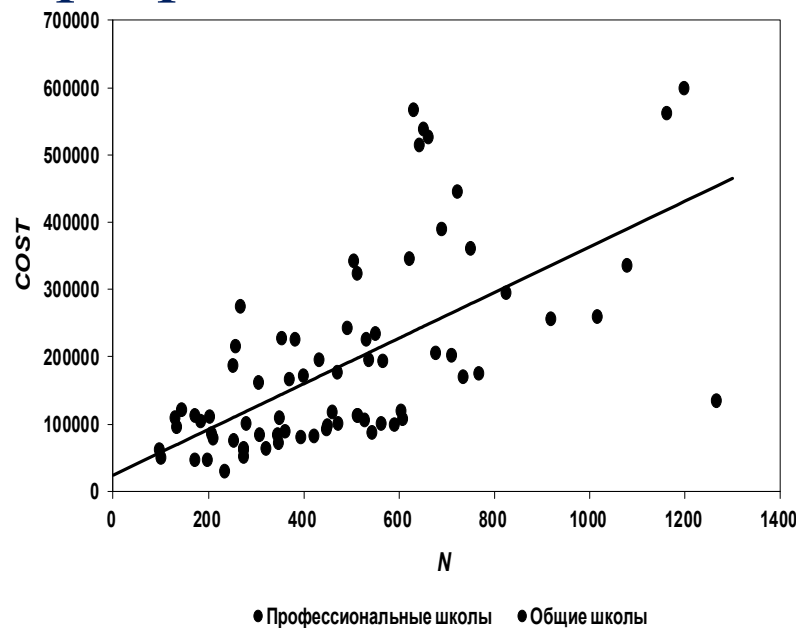


## Пример неоднородности выборки



# Тест Чоу

Тест Чоу 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 Чоу

Модель для первого набора наблюдений:

$$Y = \beta'_0 + \beta'_1 X_1 + \dots + \beta'_k X_k + \varepsilon'$$

Модель для второго набора наблюдений:

$$Y = \beta''_0 + \beta''_1 X_1 + \dots + \beta''_k X_k + \varepsilon''$$

$$H_0: \beta'_0 = \beta''_0, \dots, \beta'_k = \beta''_k, \sigma_{\varepsilon'}^2 = \sigma_{\varepsilon''}^2$$

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

## Тест Чоу

Модель для первого набора наблюдений:

$$Y = \beta'_0 + \beta'_1 X_1 + \dots + \beta'_k X_k + \varepsilon'$$

Модель для второго набора наблюдений:

$$Y = (\beta'_0 + \delta_0) + (\beta'_1 + \delta_1) X_1 + \dots + (\beta'_k + \delta_k) X_k + \varepsilon''$$

$$H_0: \beta'_0 = \beta''_0, \dots, \beta'_k = \beta''_k, \sigma_{\varepsilon'}^2 = \sigma_{\varepsilon''}^2$$

$$\Leftrightarrow H_0: \delta_0 = \delta_1 = \dots = \delta_k = 0$$

$$H_1: \delta_0^2 + \delta_1^2 + \dots + \delta_k^2 > 0$$

Общая модель:

$$Y = (\beta'_0 + \delta_0 D) + (\beta'_1 + \delta_1 D)X_1 + \dots + (\beta'_k + \delta_k D)X_k + \varepsilon''$$

$D = 0$  для наблюдений из первого набора,

$D = 1$  для наблюдений из второго набора

$$H_0: \delta_0 = \delta_1 = \dots = \delta_k = 0$$

$$H_1: \delta_0^2 + \delta_1^2 + \dots + \delta_k^2 > 0$$

## Тестовая статистика в тесте Чоу

$$F = \frac{(RSS_R - RSS_{UR})/(k+1)}{(RSS_1 + RSS_2)/(n - 2(k+1))} =$$

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

где  $RSS_p$  – 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  $F > F_{\text{критическое}}$  (при выбранном уровне значимости), то основная гипотеза отвергается и нужно оценивать две отдельные регрессии.

## Пример применения теста Чоу

Вернемся к рассматриваемому примеру. Сравним RSS по всей выборке и отдельно по двум группам школ.

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

$$\begin{aligned} F(2,70) &= \frac{(8.92 \times 10^{11} - [3.49 \times 10^{11} + 1.22 \times 10^{11}]) / 2}{(3.49 \times 10^{11} + 1.22 \times 10^{11}) / 70} \\ &= 31.4 \end{aligned}$$

Тест Чоу эквивалентен проверке гипотезы:

$$H_0: \delta_0 = \delta_1 = \dots = \delta_k = 0 \text{ (слайд 24)}$$

## Пример применения теста Чоу

$$F(2,70)_{\text{crit}, 0.1\%} = 7.6$$

Полученное значение F- статистики превышает критическое при любом разумном уровне значимости, следовательно, нулевая гипотеза отвергается, для профессиональных и обычных школ имеет место разная зависимость. Нужно оценивать отдельные регрессии.



Если качественная переменная имеет  $m$  градаций, то в модель надо ввести  $m - 1$  фиктивных переменных, если в уравнение регрессии включена константа (иначе мы попадем в ловушку дамми (dummy trap)), между столбцами матрицы  $X$  в модели  $Y = X\beta + \varepsilon$  будет линейная зависимость и мы не сможем однозначно оценить параметры этой модели с помощью МНК.

# Дамми переменные для моделирования сезонности

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

Если данные квартальные, то

$D_1 = 1$ , если наблюдение относится к 1 – му кварталу и 0, если не относится;

$D_2 = 1$ , если наблюдение относится к 2 – му кварталу и 0, если не относится;

$D_3 = 1$ , если наблюдение относится к 3 – му кварталу и 0, если не относится.

# Дамми переменные для моделирования сезонности

Рассмотрим квартальные данные. В качестве базы выберем 4-ый квартал, тогда:

$$\text{Модель: } Y = \beta_0 + \beta_1 D_1 + \beta_2 D_2 + \beta_3 D_3 + \beta_4 X + \varepsilon$$

Оцененное уравнение регрессии:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 D_1 + \hat{\beta}_2 D_2 + \hat{\beta}_3 D_3 + \hat{\beta}_4 X$$

# Дамми переменные для моделирования сезонности

Оцененное уравнение регрессии:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 D_1 + \hat{\beta}_2 D_2 + \hat{\beta}_3 D_3 + \hat{\beta}_4 X$$

Поквартальные зависимости:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 D_1 + \hat{\beta}_4 X \quad - \text{ для 1-го квартала,}$$

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_2 D_2 + \hat{\beta}_4 X \quad - \text{ для 2-го квартала,}$$

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_3 D_3 + \hat{\beta}_4 X \quad - \text{ для 3-го квартала,}$$

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_4 X \quad - \text{ для 4-го квартала (базового)}$$



NATIONAL RESEARCH  
UNIVERSITY

# Thank you for your attention!

20, Myasnitskaya str., Moscow, Russia, 101000  
Tel.: +7 (495) 628-8829, Fax: +7 (495) 628-7931  
[www.hse.ru](http://www.hse.ru)