

Семинар 5.

Фиктивные (dummy) переменные

Задача 1

Оцененная зависимость почасовой оплаты труда индивида Y (измеряется в долларах в час) от результатов выпускного теста X (измеряется в баллах) и пола (D – фиктивная переменная, равная 1 для мужчин и 0 для женщин) имеет вид:

$$\hat{Y} = 2 + 3.7X + 2.4D.$$

Все коэффициенты являются значимыми при уровне значимости 1%. При одинаковых результатах теста почасовая оплата мужчин выше почасовой оплаты женщин на

- 1) 0.024 \$ 2) 2.4 \$ 3) 0.024 % 4) 2.4%

Решение:

Уравнение для мужчин имеет вид: $\hat{Y}_m = 2 + 3.7X + 2.4, D = 1$

Уравнение для женщин: $\hat{Y}_f = 2 + 3.7X, D = 0$

Т.е. при равных результатах теста X почасовая оплата мужчин на 2.4\$ выше:

$$\hat{Y}_m - \hat{Y}_f = 2 + 3.7X + 2.4 - 2 - 3.7X = 2.4\$.$$

Задача 2

Оцененная зависимость почасовой оплаты труда американцев Y (измеряется в долларах) от стажа их работы X (измеряется в годах); пола, описываемого с помощью фиктивной переменной D_1 , равной 1 для мужчин и 0 для женщин; расовой принадлежности, описываемой с помощью фиктивной переменной D_2 , равной 1 для светлокожих и 0 для темнокожих американцев, имеет вид:

$$\hat{Y} = 4 + 0.8X + 0.04D_1 - 0.01D_2$$

Все коэффициенты являются значимыми при уровне значимости 1%.

Чему равна почасовая оплата труда темнокожих американцев при пятилетнем стаже работы?

Решение:

Для подгруппы светлокожих американцев $D_2 = 1$

При пятилетнем стаже работы $X = 5$: $\hat{Y} = 4 + 0.8 \cdot 5 + 0.04D_1 - 0.01 = 7.99 + 0.04D_1$

- для мужчин $D_1 = 1$ почасовая оплата труда составляет $\hat{Y}_m = 8.03 \$$
- для женщин $D_1 = 0$ $\hat{Y}_f = 7.99 \$$, т.е. на \$0.04 ниже, чем для мужчин.

Задача 3

Зависимость расходов на продукты питания от располагаемого дохода X имеет вид:

$$\hat{Y} = 2 + 0.6X + 0.07D_1X,$$

где D_1 – фиктивная переменная, равная 1 для городских и 0 для сельских жителей.

а) Коэффициент наклона в линейной зависимости для сельских жителей равен

- 1) 0,67 2) 0,6 3) 0,53 4) 2

б) Если вместо D_1 использовать переменную D_2 , равную 0 для городских и 1 для сельских жителей, то зависимость примет вид:

$$1) \hat{Y} = 2 + 0.67X - 0.07D_2X$$

$$2) \hat{Y} = 2 + 0.67X + 0.07D_2X$$

$$3) \hat{Y} = 2 + 0.6X - 0.07D_2X$$

$$4) \hat{Y} = 2.07 + 0.6X - 0.07D_2X.$$

Решение:

(а) Уравнение регрессии $\hat{Y} = 2 + 0.6X + 0.07D_1X$ для сельских жителей, т.е. при $D_1 = 0$ примет вид:

$$\hat{Y}_{rural} = 2 + 0.6X \text{ Т.е. угол наклона в линейной зависимости для сельских жителей равен } 0.6.$$

(б) Для сельских жителей: $\widehat{Y}_{rural} = 2 + 0.6X$.

Для городских жителей: $\widehat{Y}_{town} = 2 + 0.6X + 0.07 \cdot 1 \cdot X = 2 + 0.67X$

Соответственно, если мы возьмем дамми D_2 , в которой, наоборот, 1 будет для сельских жителей, а 0 для городских, мы можем представить нашу новую переменную как $D_2 = 1 - D_1$.

Тип местности	Дамми-переменные	
	D_1	D_2
Городская	1	0
Сельская	0	1

Из этой зависимости подставим теперь $D_1 = 1 - D_2$ в оцененное уравнение:

$$\hat{Y} = 2 + 0.6X + 0.07D_1X = 2 + 0.6X + 0.07(1 - D_2)X = 2 + 0.67X - 0.07D_2X.$$

Задача 4

Оцененная зависимость Y - расходов потребителей на газ и электричество в США в 1977 – 1999 г.г. в постоянных ценах I квартала 1977г. от времени ($t = 1$ для 1977 г., $t = 2$ для 1978 г. и т.д.) с учетом сезонных факторов ($D_i = 1$, если наблюдение относится к i -му кварталу и 0 иначе, $i = 1, \dots, 4$) имеет вид:

Если в качестве выделенной категории выбран первый квартал, оцененное уравнение имеет вид:

$$\hat{Y} = 8 + 0.1t - 3D_2 - 2.6D_3 - 2D_4$$

Если в качестве выделенной категории будет выбран не первый квартал, а второй, то уравнение регрессии примет вид

$$1) \hat{Y} = 5 + 0.1t + 3D_1 + 0.4D_3 + D_4$$

$$2) \hat{Y} = 8 + 0.1t - 3D_1 - 2.6D_3 - 2D_4$$

$$3) \hat{Y} = 5 + 0.1t - 3D_1 - 2.6D_3 - 2D_4$$

$$4) \hat{Y} = 5 + 0.1t - 3D_2 - 0.4D_3 - D_4$$

Решение:

Рассмотрим исходное уравнение $\hat{Y} = 8 + 0.1t - 3D_2 - 2.6D_3 - 2D_4$. Первый квартал базовый, от него ведется «отсчет». Квартальные дамми-переменные на 2й, 3й и 4й кварталы выглядят

следующим образом: $D_2 = \begin{cases} 1, & \text{во 2 кв} \\ 0, & \text{иначе} \end{cases}$, $D_3 = \begin{cases} 1, & \text{в 3 кв} \\ 0, & \text{иначе} \end{cases}$, $D_4 = \begin{cases} 1, & \text{в 4 кв} \\ 0, & \text{иначе} \end{cases}$

Для 1го квартала оно принимает вид: $\hat{Y} = 8 + 0.1t$.

Для 2го квартала: $\hat{Y} = 8 + 0.1t - 3 = 5 + 0.1t$

Для 3го квартала: $\hat{Y} = 8 + 0.1t - 2.6 = 5 + 0.1t + 0.4$

Для 4го квартала: $\hat{Y} = 8 + 0.1t - 2 = 5 + 0.1t + 1$

Если мы теперь возьмем за базовый второй квартал, то должны получить те же самые квартальные зависимости, только теперь у нас будут дамми-переменные $D_1 = \begin{cases} 1, & \text{в 1 кв} \\ 0, & \text{иначе} \end{cases}$,

$D_3 = \begin{cases} 1, & \text{в 3 кв} \\ 0, & \text{иначе} \end{cases}$, $D_4 = \begin{cases} 1, & \text{в 4 кв} \\ 0, & \text{иначе} \end{cases}$. По определению сезонных дамми-переменных для квартальных данных выполнено: $D_1 + D_2 + D_3 + D_4 = 1$.

Квартал	Дамми-переменные			
	D_1	D_2	D_3	D_4
I	1	0	0	0
II	0	1	0	0
III	0	0	1	0
IV	0	0	0	1

Т.к. теперь второй квартал базовый, нам нужно выразить:

$$D_2 = 1 - D_1 - D_3 - D_4$$

и подставить в наше уравнение:

$$\hat{Y} = 8 + 0.1t - 3(1 - D_1 - D_3 - D_4) - 2.6D_3 - 2D_4 = 5 + 0.1t + 3D_1 + 0.4D_3 + D_4$$

Задача 5

Оцененная зависимость почасовой оплаты труда американцев Y (измеряется в долларах в час) от длительности обучения X (измеряется в годах) и расовой принадлежности, описываемой с помощью фиктивной переменной D , равной 1 для светлокожих и 0 для темнокожих американцев, имеет вид: $Y = 5 + 0.7X + 0.04DX$.

Все коэффициенты являются значимыми при уровне значимости 1%.

Каждый дополнительный год обучения приводит к увеличению почасовой оплаты труда темнокожих американцев на

- 1) 0.74 \$ 2) 0.7 \$ ($D=0$) 3) 0.66 \$ 4) 0.74 %

Тест Chow для диагностики структурной стабильности

Задача 1

По данным для 570 индивидуумов оценили зависимость почасовой заработной платы EARN от длительности обучения S и от способностей индивидуума, описываемых обобщенной переменной ASVABC:

- по общей выборке

$$EARN = -9.96 + 0.93S + 0.21ASVABC \quad RSS_1 = 32189.36$$

(2.02) (0.16) (0.04)

- а также отдельно для мужчин

$$EARN = -7.23 + 1.01S + 0.35ASVABC \quad RSS_2 = 15223.7$$

(2.63) (0.27) (0.06)

- и женщин

$$EARN = -11.4 + 0.81S + 0.14ASVABC \quad RSS_3 = 10231.24$$

(3.24) (0.19) (0.03)

Можно ли считать, что эта зависимость одинакова для мужчин и женщин?

Решение:

Нам необходимо проверить гипотезу, что коэффициенты регрессии, оцененные отдельно для мужчин и отдельно для женщин совпадают.

$$H_0: \beta_i^1 = \beta_i^2 \quad \forall i = 1, 2, 3$$

$$H_1: \exists i: \beta_i^1 \neq \beta_i^2$$

Гипотезу мы будем проверять с помощью теста Чоу. Статистика для теста Чоу имеет вид:

$$F = \frac{(RSS_p - RSS_1 - RSS_2) / k}{(RSS_1 + RSS_2) / (n - 2k)} \sim F(k, n - 2k) \quad (\text{имеет F-распределение при нулевой гипотезе}).$$

RSS_p - RSS по общей выборке.

$$F = \frac{(32189.36 - 15223.7 - 10231.24) / 3}{(15223.7 + 10231.24) / (570 - 6)} = 49.7$$

$F_{5\%}(3, 564) \approx 2.62$. Основная гипотеза отвергается на 5% уровне значимости. Т.е. считать, что зависимость одинакова для мужчин и женщин, нельзя.

Задача 2

Оценивалась зависимость расходов на питание в расчете на одного человека от относительного индекса цен на питание и располагаемого дохода:

$$\ln Q = \beta_0 + \beta_1 \ln P + \beta_2 \ln In + \varepsilon.$$

Были получены следующие результаты:

	1927-1941 г.г. (1)	1948-1962 г.г. (2)	Все наблюдения
$\hat{\beta}_0$	4.555	5.052	4.058
$\hat{\beta}_1$	-0.235	-0.237	-0.123
$\hat{\beta}_2$	0.243	0.141	0.242
$RSS * 100$	0.1151	0.0544	0.2866

Можно ли считать зависимость единой для довоенных и послевоенных лет?

Решение:

Снова воспользуемся тестом Чоу.

$$H_0: \beta_i^1 = \beta_i^2 \quad \forall i = 0, 1$$

$$H_1: \exists i: \beta_i^1 \neq \beta_i^2$$

$$F = \frac{100 \cdot (0.2866 - 0.1151 - 0.0544) / 3}{100 \cdot (0.1151 + 0.0544) / (15 + 15 - 6)} = 5.53, \quad F_{5\%}(3, 24) = 3.01.$$

Основная гипотеза отвергается на уровне значимости 5%. Зависимость нельзя считать единой для довоенных и послевоенных лет.

Задача 3

Исследователь оценил зависимость продолжительности жизни от концентрации вредных промышленных выбросов в атмосфере и ежегодных средних частных расходов на медицинскую помощь с помощью регрессий со свободным членом для 1) 300 жителей индустриальных центров, 2) 200 сельских жителей, 3) по общей выборке и получил в этих регрессиях соответственно суммы квадратов остатков $RSS1 = 204$, $RSS2 = 290$, $RSS3 = 902$

Значение F – статистики для проверки гипотезы о том, что зависимость одинакова для городских и сельских жителей равно

- 1) 136 2) 137 3) 138 4) 140 5) 142

Решение:

$$F = \frac{(902 - 204 - 290) / 3}{(204 + 290) / (300 + 200 - 6)} = 136$$

Интерпретация коэффициентов при различных функциональных формах уравнения

Задача 1

По месячным данным с 01.2001 по 06.2003 были оценены три регрессии:

$$Y = 18 - 0.427P + 0.000007I \quad (1)$$

(0.23) (0.006) (0.00001)

$$\ln Y = 4.7 - 0.096P + 0.000001I \quad (2)$$

(0.15) (0.004) (0.000007)

$$\ln Y = 12 - 3.11 \ln P + 0.0317 \ln I \quad (3)$$

(0.73) (0.15) (0.029)

где Y – агрегированные расходы на медицинские услуги (в млрд. руб.), P – индекс цен на медицинские услуги, I – средний среднемесячный доход россиян (руб.), в скобках указаны стандартные отклонения.

Дайте экономическую интерпретацию полученным результатам.

Решение:

Линейная модель

Проверим значимость коэффициентов:

Для всех коэффициентов критическое значение будет одно и то же: $t_{2,5\%} \approx 2$.

$$t_2 = \frac{-0.427}{0.006} = -71.167 \Rightarrow \text{коэффициент перед } P \text{ значим на уровне } 5\%;$$

$$t_3 = \frac{0.000007}{0.00001} = 0.7 \Rightarrow \text{коэффициент перед } I \text{ не значим на уровне } 5\%.$$

Интерпретация значимого коэффициента:

При увеличении индекса цен на медицинские услуги P на 1 единицу агрегированные расходы на медицинские услуги Y уменьшаются на 0,427 единиц.

Полулогарифмическая модель

Проверим значимость коэффициентов:

$$t_2 = \frac{-0.096}{0.004} = -24 \Rightarrow \text{коэффициент перед } P \text{ значим на уровне } 5\%;$$

$$t_3 = \frac{0,000001}{0,00007} = 0,0143 \Rightarrow \text{коэффициент перед I не значим на уровне 5\%}.$$

Интерпретация значимого коэффициента:

При увеличении индекса цен на медицинские услуги Р на 1 единицу агрегированные расходы на медицинские услуги Y уменьшаются на $0,096 * 100\% = 9,6\%$.

Логарифмическая модель

Проверим значимость коэффициентов:

$$t_2 = \frac{-3,11}{0,15} = -20,733 \Rightarrow \text{коэффициент перед Р значим на уровне 5\%;}$$

$$t_3 = \frac{0,0317}{0,029} = 1,093 \Rightarrow \text{коэффициент перед I не значим на уровне 5\%}.$$

Интерпретация значимого коэффициента:

При увеличении индекса цен на медицинские услуги Р на 1% агрегированные расходы на медицинские услуги Y уменьшаются на 3,11%. Отношение процентного изменения одной величины к процентному изменению другой величины называется эластичностью. В данном случае эластичность агрегированных расходов на мед. услуги Y по индексу цен на мед. услуги Р равна 3,11.