Эконометрика, 2020-2021, 2 модуль
Семинар 3
9.11.20
для
Группы Э_Б2018_Э_3
Семинарист О.А.Демидова

Дамми (фиктивные) переменные и тест Чоу

Материалы из учебника О.Демидовой и Д.Малахова «Эконометрика. Учебник и практикум»

Задача 7.1. 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%

Задача 7.2.

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?

Задача 7.3.

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

Задача. Оценена зависимость расходов потребителей на газ и электричество Y 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, $i = 1,...,4$):

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

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

Задача 7.5.

По данным для 570 индивидуумов оценили зависимость длительности обучения индивидуума S от способностей индивидуума, описываемых обобщенной переменной ASVABC и пола индивидуума, описываемого с помощью фиктивной переменной MALE (равной 1 для мужчин и 0 для женщин) с помощью двух регрессий:

$$\hat{S} = \underset{(0.44)}{6.12} + \underset{(0.0088)}{0.147} \cdot ASVAB,\ RSS_1 = 2099,9$$

$$\hat{S} = \underset{(0.73)}{6.72} + \underset{(0.014)}{0.137} \cdot ASVAB - \underset{(0.933)}{1.035} \cdot MALE + \underset{(0.018)}{0.0166} \cdot (\text{MALE} \cdot \text{ASVABC}),\ RSS_2 = 2090,98$$

Зависит ли длительность обучения от пола индивидуума и почему?

Задача 7.6.

По квартальным данным 1960-1976 г.г. была оценена модель с тремя объясняющими факторами:

$$\hat{Y} = 1.03 + 0.1X_1 - 4.45X_2 + 0.26X_3,\ ESS = 103.5,\ RSS = 17.48.$$

При добавлении в модель трех сезонных dummy – переменных значение ESS увеличилось до 107.3.

Проверить гипотезу о наличии сезонности.

Задача 7.7.

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

$$EARN = \underset{(2.02)}{-9.96} + \underset{(0.16)}{0.93} S + \underset{(0.04)}{0.21} ASVABC \quad RSS_1 = 32189.36$$

- а также отдельно для мужчин

$$EARN = \underset{(2.63)}{-7.23} + \underset{(0.27)}{1.01} S + \underset{(0.06)}{0.35} ASVABC \quad RSS_2 = 15223.7$$

- и женщин

$$EARN = \underset{(3.24)}{-11.4} + \underset{(0.19)}{0.81} S + \underset{(0.03)}{0.14} ASVABC \quad RSS_3 = 10231.24$$

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

Упражнение. Зависимость длительности обучения индивида от его способностей и длительности обучения родителей

Для выполнения приведенных ниже упражнений используются данные файла Dougherty.dta.

Переменные:

EARNINGS – почасовая заработная плата индивида,

S – длительность обучения индивида, SM - длительность обучения мамы индивида,

SF длительность обучения отца индивида,

ASVABC – обобщенный показатель способностей индивида, рассчитанный по результатам тестов,

MALE – дамми переменная, равная 1 для мужчин и 0 для женщин,

ETHBLACK - дамми переменная, равная 1 для афроамериканцев и 0 иначе,

ETHHISP - дамми переменная, равная 1 для испаноязычных и 0 иначе и др.

Используя указанные (и по желанию другие) пременные, определите, влияет ли пол и раса на заработную плату.

Задания для самостоятельного решения

Упражнение 7.2.
В статистическом пакете Stata 14, по данным файла flats.dta , используя переменные price_metr, livesp, kitsp, dist, metrdist, floors, walk, (где price_metr - стоимость квадратного метра однокомнатной квартиры, описание остальных переменных дано в приложении, определите, одинакова ли зависимость для двух групп квартир (для которых время пути от метро дано в минутах пешком и для которых время пути от метро дано в минутах езды на транспорте).

1) С помощью оценки регрессии вида (7.4),
2) С помощью теста Чоу.

Используйте 5% процентный уровень значимости.

Рекомендации.

1) Для оценки регрессии с варьирующимися коэффициентами наклона для двух групп переменных создадим новые переменные, которые являются перемножением (cross-terms) переменных livesp, kitsp, dist, metrdist, floors и переменной walk (объясните, для чего это нужно):
2)

```
. gen livesp_walk= livesp* walk

. gen kitsp_walk= kitsp* walk

. gen dist_walk= dist* walk

. gen metrdist_walk= metrdist* walk
```

```
. gen floors_walk= floors* walk
```

Оцените новую регрессию с включенными cross-terms переменными:

```
. reg price_meter livesp kitsp dist metrdist floors walk livesp_walk kitsp_walk
dist_walk metrdist_walk floors_walk,
```

Проверим одновременную значимость всех коэффициентов переменных, содержащих walk, воспользовавшись командой test:

```
test livesp_walk= kitsp_walk= dist_walk= metrdist_walk= floors_walk= walk=0,
```

3) Проведем тест Чоу, оцениваем одну и ту же форму модели а) для всех квартир, б) для квартир, до которых время пути от метро дано в минутах пешим шагом, и в) для квартир, для которых время в пути дано в минутах езды на автомобиле.

Модель по данным для всех квартир можно оценить с помощью команды:

```
reg price_meter livesp kitsp dist metrdist floors,
```

Сохраним RSS с помощью команды scalar rssp=e(rss).

Модель для квартир, до которых время пути от метро дано в минутах пешим шагом можно оценить с помощью команды:

```
. reg price_meter livesp kitsp dist metrdist floors if walk==1,
```

Сохраним RSS с помощью команды scalar rss1=e(rss).

Модель для квартир, до которых время пути от метро дано в минутах езды можно оценить с помощью команды:

```
. reg price_meter livesp kitsp dist metrdist floors if walk==0,
```

Сохраним RSS с помощью команды scalar rss2=e(rss).

Используя RSS из оцененных регрессий, рассчитаем тестовую F – статистику:

. scalar F=((rssp-rss1-rss2)/6)/((rss1+rss2)/(773-2*6))

. display F

20.060314

Для нахождения p-value для F-статистики используйте команду

di Ftail(6, 761, 20.060314)

Упражнение 7.3.

Используя статистический пакет Stata, по данным файла nlsw88.dta (эта база данных встроена в статистический пакет Stata, сделать ее активной можно выбрав File->Example Datasets…-> Example Datasets Installed in Stata, описание можно найти, нажав на describe (также описание переменных дано в Приложении 1),

1) Оцените модель $wage_i = \beta_0 + \beta_1 \cdot \text{hours}_i + \beta_2 \cdot \text{ttl\_exp}_i + \beta_3 \cdot tenure_i + \beta_4 \cdot union_i + u_i, i = 1,...,n$.

Проинтерпретируйте значение оценки коэффициента перед переменной *union*.

2) Проанализируйте, нужно ли оценивать модель

$$wage_i = \beta_0 + \beta_1 \cdot \text{hours}_i + \beta_2 \cdot \text{ttl\_exp}_i + \beta_3 \cdot tenure_i + u_i$$

отдельно для тех женщин, которые состоят в союзе и для тех, которые не состоят.

Решение.

1) Заметим, что в базе данных значение union переменной union соответствует 1, значение nonunion соответствует 0, пропуски в значениях переменных обозначаются как "." (соответствующие наблюдения выкидываются).

Оценим искомую модель с помощью команды:

```
reg wage hours ttl_exp tenure  union,
```

получим:

```
      Source |       SS       df       MS              Number of obs =    1867
-------------+------------------------------           F(  4,  1862) =   84.07
       Model |  4961.07427     4  1240.26857           Prob > F      =  0.0000
    Residual |  27470.4872  1862  14.7532155           R-squared     =  0.1530
-------------+------------------------------           Adj R-squared =  0.1512
       Total |  32431.5615  1866   17.380258           Root MSE      =   3.841

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       hours |   .0191089   .0092112     2.07   0.038     .0010436    .0371743
     ttl_exp |   .2785683   .0243063    11.46   0.000     .2308978    .3262389
      tenure |   .0468629   .0196121     2.39   0.017     .0083988    .0853269
       union |   1.190311   .2085296     5.71   0.000     .7813344    1.599287
       _cons |   2.691105    .392896     6.85   0.000     1.920542    3.461667
------------------------------------------------------------------------------
```

Из результатов оценки модели, можно заметить, что если респондент состоит в профсоюзе, то при прочих равных его почасовая зарплата выше на 1.190311 долл. (коэффициент при переменной union значим на любом адекватном уровне значимости).

2) Теперь проанализируем, нужно ли оценивать вышеуказанную модель отдельно для каждой подвыборки.

Для этого создадим переменные, которые являются перемножением регрессоров и дамми-переменной:

```
. gen hours_union=hours*union
(369 missing values generated)

. gen ttl_exp_union=ttl_exp*union
(368 missing values generated)

. gen tenure_union=tenure*union
(378 missing values generated)
```

Оценим регрессию, включив в нее вновь созданные переменные:

```
reg wage hours ttl_exp tenure  union hours_union ttl_exp_union tenure_union,
```

получим:

```
      Source |       SS       df       MS              Number of obs =    1867
-------------+------------------------------           F(  7,  1859) =   50.31
       Model |  5165.36882     7  737.909831            Prob > F      =  0.0000
    Residual |  27266.1926  1859  14.6671289            R-squared     =  0.1593
-------------+------------------------------           Adj R-squared =  0.1561
       Total |  32431.5615  1866   17.380258           Root MSE      =  3.8298

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       hours |   .0356629   .0103232     3.45   0.001     .0154165    .0559093
     ttl_exp |    .282476   .0273754    10.32   0.000     .2287863    .3361656
      tenure |   .0449328   .0228423     1.97   0.049     .0001335    .0897322
       union |   4.483446   .9768301     4.59   0.000     2.567647    6.399245
 hours_union |   -.080646   .0226316    -3.56   0.000     -.125032   -.0362599
ttl_exp_union |  -.0235813   .0588922    -0.40   0.689     -.139083    .0919204
tenure_union |   .0148574   .0446927     0.33   0.740    -.0727958    .1025106
       _cons |   2.036556   .4369703     4.66   0.000     1.179552     2.89356
------------------------------------------------------------------------------
```

Проверим совместную значимость переменных, содержащих union с помощью команды:

```
. test  union=hours_union=ttl_exp_union=tenure_union=0

 ( 1)  union - hours_union = 0
 ( 2)  union - ttl_exp_union = 0
 ( 3)  union - tenure_union = 0
 ( 4)  union = 0

       F(  4,  1859) =   11.68
            Prob > F =    0.0000
```

Так как p-value соответствующей тестовой статистики равно 0, то оценки всех коэффициентов совместно отличны от нуля и необходимо оценивать модели для двух подвыборок отдельно, что мы и сделаем ниже.

Модель для профсоюзных рабочих:

```
. reg wage hours tenure ttl_exp if union==1

      Source |       SS       df       MS              Number of obs =     460
-------------+------------------------------           F(  3,   456) =   19.63
       Model |  914.385516     3  304.795172            Prob > F      =  0.0000
```

```
    Residual |  7079.67078   456  15.5255938           R-squared     =  0.1144
-------------+----------------------------           Adj R-squared =  0.1086
       Total |  7994.05629   459  17.4162447           Root MSE      =  3.9403

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       hours |  -.0449831    .020721    -2.17   0.030    -.0857037   -.0042625
      tenure |   .0597902   .0395226     1.51   0.131    -.0178788    .1374593
     ttl_exp |   .2588947   .0536471     4.83   0.000     .1534685    .3643209
       _cons |   6.520002   .8988477     7.25   0.000     4.753605    8.286399
```

Модель для респондентов, не состоящих в профсоюзе:

```
. reg wage hours ttl_exp tenure if union==0

      Source |       SS       df       MS              Number of obs =    1407
-------------+------------------------------           F(  3,  1403) =   81.47
       Model |  3516.54058     3  1172.18019           Prob > F      =  0.0000
    Residual |  20186.5219  1403  14.3881125           R-squared     =  0.1484
-------------+------------------------------           Adj R-squared =  0.1465
       Total |  23703.0624  1406  16.8585081           Root MSE      =  3.7932

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       hours |   .0356629   .0102246     3.49   0.001     .0156058      .05572
     ttl_exp |    .282476   .0271137    10.42   0.000     .2292881    .3356638
      tenure |   .0449328    .022624     1.99   0.047     .0005522    .0893134
       _cons |   2.036556    .432794     4.71   0.000     1.187563    2.885549
------------------------------------------------------------------------------
```

Стоит сделать важное замечание. При проведении теста Чоу необходимо сначала определить оптимальную модель для всей выборки, а затем уже проводить сам тест.

Исходя из простого сопоставления результатов оценки этих моделей, можно заключить, что результаты сильно отличаются для двух подвыборок, что свидетельствует о корректности проведенного теста.

Задание 7.4.

В приведенных ниже таблицах содержатся результаты оценивания функции спроса на молоко (в таблице 1 по всем наблюдениям, в таблице 2 – по наблюдениям для сельской местности, в таблице 3 – для городской местности).
Переменные:
buymilk – стоимость молока, купленного семьей за последние 7 дней (в руб.),
income – доход семьи за месяц,
prmilk - цена 1 л молока (в руб.),
status – тип населенного пункта (1 – областной центр, 2 – город, 3 –поселок городского типа, 4 – село),

Таблица 1.

```
reg  buymilk_c inc pr_milk
```

```
      Source |       SS       df       MS         Number of obs = 2127
-------------+------------------------------      F(  2,   785) = 943.58
       Model |  7855703.78    2    3927851.89     Prob > F      = 0.0000
    Residual |  8841601.29   2124  4162.71247     R-squared     = 0.4705
-------------+------------------------------      Adj R-squared = 0.4700
       Total |  16697305.1   2126  7853.85939     Root MSE      = 64.519

------------------------------------------------------------------------------
   buymilk_c |      Coef.   Std. Err.    t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         inc |   .0002428   .0000762   3.19   0.001      .0000934   .0003922
     pr_milk |   .8768133     .02023  43.34   0.000       .837140   .9164859
       _cons |   32.96319   1.746953  18.87   0.000      29.53727   36.38911
```

Таблица 2.

```
reg  buymilk_c inc pr_milk if  status==4
```

```
      Source |       SS       df       MS         Number of obs = 348
-------------+------------------------------      F(  2,   785) = 319.33
       Model |  3184511.16   2    1592255.58      Prob > F      = 0.0000
    Residual |  1720236.56  345   4986.19293      R-squared     = 0.6493
-------------+------------------------------      Adj R-squared = 0.6472
       Total |  4904747.72  347   14134.7197      Root MSE      = 70.613

------------------------------------------------------------------------------
   buymilk_c |     Coef.    Std. Err.    t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         inc |   .0002418   .0002566   0.94   0.347     -.0002629  .0007465
     pr_milk |   .9387025   .0371628  25.26   0.000      .8656084  1.011797
       _cons |   32.57962   4.539265   7.18   0.000      23.65151  41.50774
```

Таблица 3.

```
reg  buymilk_c inc pr_milk if  status==1 status==2 status==3
```

```
      Source |       SS       df       MS            Number of obs = 1779
-------------+------------------------------         F(  2,   785) = 586.49
       Model |  4688916.24    2    2344458.12        Prob > F      = 0.0000
    Residual |  7099423.62   1776  3997.42321        R-squared     = 0.3978
-------------+------------------------------         Adj R-squared = 0.3971
       Total |  11788339.9   1778  6630.11241        Root MSE      = 63.225

------------------------------------------------------------------------------
   buymilk_c |      Coef.   Std. Err.    t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         inc |   .0002451   .0000793    3.09  0.002    .0000896  .0004006
     pr_milk |   .8425161   .0246925   34.12  0.000    .7940866  .8909456
       _cons |   33.35554   1.894483   17.61  0.000    29.63989  37.07119
```

Можно ли считать зависимость спроса на молоко от его цены и дохода единой для городской и сельской местности? Ответ обоснуйте подходящим тестом.