тесты на функциональную форму

лекция 14

ПРОЦЕДУРЫ ПОШАГОВОГО ОТБОРА

1. ПОСЛЕДОВАТЕЛЬНОГО ПРИСОЕДИНЕНИЯ
2. ПОСЛЕДОВАТЕЛЬНОГО УДАЛЕНИЯ
3. ПРИСОЕДИНЕНИЯ-УДАЛЕНИЯ
4. УДАЛЕНИЯ-ПРИСОЕДИНЕНИЯ

См.:

Магнус, Катышев, Пересецкий «Эконометрика. Начальный курс»

Айвазян, Мхитарян «Прикладная статистика и эконометрика»

ПРОЦЕДУРЫ ПОШАГОВОГО ОТБОРА

STATA:

sw reg var_y (varlist1) var_x1… (…),

[*pr(level) | pe(level)* | *pr(level) pe(level) | pr(level) pe(level) forward*]

[**hier**] [**lockterm1**] [**lr**] options

pr(#)	backward selection
pr(#) hier	backward hierarchical selection
pr(#) pe(#)	backward stepwise
pe(#)	forward selection
pe(#) hier	forward hierarchical selection
pr(#) pe(#) forward	forward stepwise

ПРОЦЕДУРЫ ПОШАГОВОГО ОТБОРА

. reg EARNINGS MALE ETHWHITE COLLBARG URBAN ASVABC AGE S

Source	SS	df	MS
Model	9852.78065	7	1407.54009
Residual	30041.8558	562	53.4552594
Total	39894.6364	569	70.1135965

Number of obs = 570
F(7, 562) = 26.33
Prob > F = 0.0000
R-squared = 0.2470
Adj R-squared = 0.2376
Root MSE = 7.3113

EARNINGS	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
MALE	3.45469	.623498	5.54	0.000	2.230019	4.679361
ETHWHITE	.8424168	.9654416	0.87	0.383	-1.053898	2.738731
COLLBARG	1.253548	.9858559	1.27	0.204	-.682864	3.18996
URBAN	1.033109	.772808	1.34	0.182	-.4848363	2.551054
ASVABC	.1732071	.0435075	3.98	0.000	.0877499	.2586642
AGE	.2590228	.1433373	1.81	0.071	-.0225194	.540565
S	.9598976	.1622137	5.92	0.000	.6412785	1.278517
_cons	-20.36929	4.880397	-4.17	0.000	-29.95534	-10.78325

ПРОЦЕДУРЫ ПОШАГОВОГО ОТБОРА

```
. sw reg  EARNINGS MALE ETHWHITE  COLLBARG URBAN ASVABC AGE S, pe(.15)
               begin with empty model
      p = 0.0000 <  0.1500  adding   S
      p = 0.0000 <  0.1500  adding   MALE
      p = 0.0000 <  0.1500  adding   ASVABC
      p = 0.0565 <  0.1500  adding   AGE

. sw reg  EARNINGS MALE ETHWHITE  COLLBARG URBAN ASVABC AGE S, pr(.15)
               begin with full model
p = 0.3833 >= 0.1500  removing ETHWHITE
p = 0.2300 >= 0.1500  removing URBAN
p = 0.2013 >= 0.1500  removing COLLBARG

•        Source |       SS       df       MS              Number of obs =     570
•   -------------+------------------------------           F(  4,   565) =   45.05
•         Model |  9647.38895     4  2411.84724           Prob > F      =  0.0000
       Residual |  30247.2475   565  53.5349513           R-squared     =  0.2418
•   -------------+------------------------------           Adj R-squared =  0.2365
•         Total |  39894.6364   569  70.1135965           Root MSE      =  7.3168

------------------------------------------------------------------------------
    EARNINGS |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        MALE |   3.542326   .6215357     5.70   0.000     2.321524    4.763129
           S |   .9885847   .1598943     6.18   0.000     .6745249    1.302645
      ASVABC |    .182249   .0415634     4.38   0.000     .1006114    .2638866
         AGE |   .2735131    .143115     1.91   0.056    -.0075893    .5546155
       _cons |  -20.05516   4.855838    -4.13   0.000    -29.59285   -10.51746
•   ------------------------------------------------------------------------------
```

ПРОЦЕДУРЫ ПОШАГОВОГО ОТБОРА

```
. sw reg  EARNINGS MALE ETHWHITE  COLLBARG URBAN ASVABC AGE S, pe(.05) hier
                 begin with empty model
p = 0.0000 <  0.0500  adding   MALE
p = 0.0057 <  0.0500  adding   ETHWHITE
p = 0.0897 >= 0.0500  testing  COLLBARG

      Source |       SS       df       MS              Number of obs =     570
-------------+------------------------------           F(  2,   567) =   19.08
       Model |  2515.78772     2  1257.89386           Prob > F      =  0.0000
    Residual |  37378.8487   567  65.9238954           R-squared     =  0.0631
-------------+------------------------------           Adj R-squared =  0.0598
       Total |  39894.6364   569  70.1135965           Root MSE      =  8.1194

------------------------------------------------------------------------------
    EARNINGS |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        MALE |   3.655179   .6887306     5.31   0.000     2.302404    5.007953
    ETHWHITE |   2.766774   .9974825     2.77   0.006     .8075625    4.725986
       _cons |   8.915434   .9826909     9.07   0.000     6.985275    10.84559
------------------------------------------------------------------------------
```

ПРОЦЕДУРЫ ПОШАГОВОГО ОТБОРА

```
. sw reg  EARNINGS AGE S ASVABC COLLBARG MALE  URBAN  ETHWHITE, pr(.15) hier
                begin with full model
p = 0.3833 >= 0.1500  removing ETHWHITE
p = 0.2300 >= 0.1500  removing URBAN
p = 0.0000 <  0.1500  keeping  MALE

      Source |       SS       df       MS              Number of obs =     570
-------------+------------------------------           F(  5,   564) =   36.41
       Model |  9734.90813     5  1946.98163           Prob > F      =  0.0000
    Residual |  30159.7283   564  53.4746955           R-squared     =  0.2440
-------------+------------------------------           Adj R-squared =  0.2373
       Total |  39894.6364   569  70.1135965           Root MSE      =  7.3126

------------------------------------------------------------------------------
    EARNINGS |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         AGE |   .2612717   .1433541     1.82   0.069    -.0203015    .5428448
           S |   .9839398   .1598455     6.16   0.000     .6699746    1.297905
      ASVABC |   .1830015   .0415442     4.40   0.000     .1014014    .2646017
    COLLBARG |   1.255857   .9816639     1.28   0.201    -.6723063    3.184021
        MALE |   3.487598   .6226571     5.60   0.000     2.264588    4.710608
       _cons |  -19.73568   4.859525    -4.06   0.000    -29.28065    -10.1907
------------------------------------------------------------------------------
```

ТЕСТЫ НА ФУНКЦИОНАЛЬНУЮ ФОРМУ

1. RESET тест Рамсея
2. J-тест Дэвидсона и МакКиннона
3. PE-тест МакКинона
4. Метод Зарембки
5. Тест Кокса-Бокса
6. PC-тест Амемии (Акаике)
7. Информационный критерий Шварца
8. Информационный критерий Акаике

RESET тест Рамсея

Regression Equation Specification Erorr Test

$$H_0: \quad y_t = \mathbf{x}_t \boldsymbol{\beta} + \varepsilon_t; \quad \mathbf{x}_t = \left(x_{1t}, \ldots, x_{kt} \right); \quad t = 1, \ldots, n$$

$$H_1: \quad y_t = \mathbf{x}_t \boldsymbol{\beta} + \alpha_2 \hat{y}_t^{\,2} + \ldots \alpha_m \hat{y}_t^{\,m} + \varepsilon_t$$

$$H_1: \quad \boldsymbol{\alpha} = \mathbf{0}$$

$$H_1': \quad y_t = \mathbf{x}_t \boldsymbol{\beta} + \gamma_{21} x_{1t}^2 + \ldots + \gamma_{m1} x_{1t}^m + \ldots + \delta_1 x_{1t} x_{2t} + \ldots \varepsilon_t \sim$$

$$H_1': \quad \boldsymbol{\gamma} = \mathbf{0}, \boldsymbol{\delta} = \mathbf{0}$$

RESET тест Рамсея

Regression Equation Specification Erorr Test

```
. reg  EARNINGS MALE ASVABC AGE S

      Source |       SS       df       MS              Number of obs =     570
-------------+------------------------------           F(  4,   565) =   26.01
       Model |  5971.29146     4  1492.82286           Prob > F      =  0.0000
    Residual |  32425.7456   565  57.3907002           R-squared     =  0.1555
-------------+------------------------------           Adj R-squared =  0.1495
       Total |  38397.0371   569  67.4816117           Root MSE      =  7.5757

------------------------------------------------------------------------------
    EARNINGS |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        MALE |   2.873136   .6445776     4.46   0.000     1.607075    4.139197
      ASVABC |   .1258984   .0434857     2.90   0.004      .040485    .2113119
         AGE |   .1794028   .1450898     1.24   0.217    -.1055785    .4643842
           S |   .8348173   .1601258     5.21   0.000     .5203028    1.149332
       _cons |  -12.03274   5.008224    -2.40   0.017    -21.86975   -2.195733
------------------------------------------------------------------------------
```

```
. ovtest

Ramsey RESET test using powers of the fitted values of EARNINGS
       Ho:  model has no omitted variables
                 F(3, 562) =      3.26
                  Prob > F =      0.0213
```

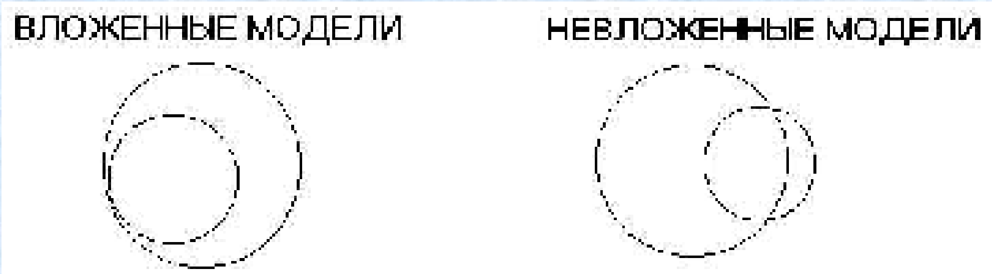
```
. ovtest,rhs

Ramsey RESET test using powers of the independent variables
       Ho:  model has no omitted variables
                 F(9, 556) =      2.50
                  Prob > F =      0.0082
```

Выбор переменных регрессии: вложенные и невложенные модели

Вложенная модель является частным случаем (ограниченной версией) более общей модели

Невложенные модели имеют разные наборы переменных



$$\hat{Y}_i = b_0 + b_1 X_{1i} + b_2 X_{2i}$$

$$\hat{Y}_i^{\text{влож}} = b_0' + b_1' X_{1i}$$

$$\hat{Y}_i^{A} = b_0 + b_1 X_{1i} + b_2 X_{2i}$$

$$\hat{Y}_i^{B} = b_0' + b_1' X_{1i} + b_2' Z_i$$

Вложенные модели непосредственно сравнимы.

Сравнение невложенных моделей возможно только с помощью специальных процедур.

J-тест Дэвидсона-МакКиннона

Сравнение невложенных моделей

$$A: \quad \mathbf{y} = \mathbf{x}\boldsymbol{\beta} + \boldsymbol{\varepsilon}; \quad \mathbf{x} = \left(\mathbf{x_1}, \ldots, \mathbf{x_k}\right)$$

$$B: \quad \mathbf{y} = \mathbf{z}\boldsymbol{\gamma} + \boldsymbol{\upsilon}; \quad \mathbf{z} = \left(\mathbf{z}_1, \ldots, \mathbf{z}_m\right)$$

$$C: \quad \mathbf{y} = \mathbf{x}\boldsymbol{\beta}\left(1-\delta\right) + \mathbf{z}\boldsymbol{\gamma}\delta + \mathbf{u}$$

$$\delta = 0: \; C = A; \quad \delta = 1: \; C = B$$

$\delta, \boldsymbol{\beta}, \boldsymbol{\gamma}$ не могут быть идентифицированы одновременно

J-тест Дэвидсона-МакКиннона

Сравнение невложенных моделей

$$A: \quad \mathbf{y} = \mathbf{x}\boldsymbol{\beta} + \boldsymbol{\varepsilon}; \quad \mathbf{x} = \left(\mathbf{x}_1, \ldots, \mathbf{x}_k\right)$$

$$B: \quad \mathbf{y} = \mathbf{z}\boldsymbol{\gamma} + \boldsymbol{\upsilon}; \quad \mathbf{z} = \left(\mathbf{z}_1, \ldots, \mathbf{z}_m\right) \Rightarrow \hat{\boldsymbol{\gamma}}; \hat{\mathbf{y}}_B$$

$$\mathbf{y} = \mathbf{x}\boldsymbol{\beta}\left(1-\delta\right) + \mathbf{z}\hat{\boldsymbol{\gamma}}\delta + \mathbf{u} = \mathbf{x}\boldsymbol{\beta}^* + \delta\hat{\mathbf{y}}_B, \quad \boldsymbol{\beta}^* = \boldsymbol{\beta}\left(1-\delta\right)$$

$$H_0: \quad \delta = 0$$

$$H_1: \quad \delta \neq 0$$

$$\delta = 0 \Rightarrow \operatorname{p\,lim}\hat{\delta} = 0, \quad t = \frac{\hat{\delta}}{s.e.\hat{\delta}} \sim N(0,1)$$

J-тест Дэвидсона-МакКиннона

Сравнение невложенных моделей

$$A: \quad \mathbf{y} = \mathbf{x\beta} + \boldsymbol{\varepsilon}; \quad \mathbf{x} = \left(\mathbf{x_1}, \ldots, \mathbf{x_k}\right) \Rightarrow \hat{\boldsymbol{\beta}}; \hat{\mathbf{y}}_A$$

$$B: \quad \mathbf{y} = \mathbf{z\gamma} + \boldsymbol{\upsilon}; \quad \mathbf{z} = \left(\mathbf{z}_1, \ldots, \mathbf{z}_m\right)$$

$$\mathbf{y} = \mathbf{x}\hat{\boldsymbol{\beta}}\delta + \mathbf{z\gamma}\left(1-\delta\right) + \mathbf{u} = \hat{\mathbf{y}}_A\delta + \mathbf{z\gamma}^*, \quad \boldsymbol{\gamma}^* = \boldsymbol{\gamma}\left(1-\delta\right)$$

$$H_0: \quad \delta = 0$$

$$H_1: \quad \delta \neq 0$$

$$\delta = 0 \Rightarrow \operatorname{p\,lim}\hat{\delta} = 0, \quad t = \frac{\hat{\delta}}{s.e.\hat{\delta}} \sim N(0,1)$$

J-тест ошибочной спецификации Дэвидсона-МакКиннона для невложенных моделей

$$\hat{Y}_i^A = b_0 + b_1 X_{1i} + b_2 X_{2i}$$

$$\hat{Y}_i^B = b_0' + b_1' X_{1i} + b_2' Z_i$$

1. Оценивается уравнение регрессии (*A*)
2. Расчетные значения зависимой переменной из модели (*A*) включаются в модель (*B*) в качестве дополнительной объясняющей переменной

$$Y_i^B = b_0'' + b_1'' X_{1i} + b_2'' Z_i + b_3'' \hat{Y}_i^A + \varepsilon_i$$

3. Проводится оценка улучшения модели по *F*-критерию
4. Делается симметричная процедура

J-тест ошибочной спецификации Дэвидсона-МакКиннона для невложенных моделей

В результате применения теста возможны четыре случая:

1. Модель (A) значимо улучшается, а (B) – нет.

Вывод: выбираем модель (B).

2. Модель (B) значимо улучшается, а (A) – нет.

Вывод: выбираем модель (A).

3. Обе модели значимо улучшаются.

Вывод: ни одна из них не пригодна.

4. Обе модели улучшаются не значимо.

Вывод: данных недостаточно, чтобы различить качество моделей.

PE-тест МакКиннона

сравнение невложенных моделей

$$A: \quad \mathbf{y} = \mathbf{x}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_1 \quad \Rightarrow \hat{\boldsymbol{\beta}}, \hat{\mathbf{y}}$$

$$B: \quad \ln(\mathbf{y}) = \mathbf{x}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}_2 \quad \Rightarrow \hat{\boldsymbol{\gamma}}, \widehat{\ln(\mathbf{y})}$$

$$C: \quad \mathbf{y} = \mathbf{x}\boldsymbol{\beta} + \delta\left(\ln\left(\hat{\mathbf{y}}\right) - \widehat{\ln(\mathbf{y})}\right) + \mathbf{u}$$

$$\delta = 0: \; C = A$$

$$H_{00}: \quad \delta = 0 \Rightarrow \text{верна линейная модель}$$

$$H_{01}: \quad \delta \neq 0 \Rightarrow \text{линейная модель не верна}$$

$$\delta = 0 \Rightarrow \; p\lim \hat{\delta} = 0, \quad t = \frac{\hat{\delta}}{s.e.\hat{\delta}} \sim N(0,1)$$

PE-тест МакКиннона

сравнение невложенных моделей

$$A: \quad \mathbf{y} = \mathbf{x}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_1 \quad \Rightarrow \hat{\boldsymbol{\beta}}, \hat{\mathbf{y}}$$

$$B: \quad \ln(\mathbf{y}) = \mathbf{x}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}_2 \quad \Rightarrow \hat{\boldsymbol{\gamma}}, \widehat{\ln(\mathbf{y})}$$

$$D: \quad \ln(\mathbf{y}) = \mathbf{x}\boldsymbol{\gamma} + \rho\left(\hat{\mathbf{y}} - \exp\left[\widehat{\ln(\mathbf{y})}\right]\right) + \mathbf{v}$$

$$\rho = 0: \; D = B$$

$H_{10}: \quad \rho = 0 \Rightarrow$ верна лог-линейная модель

$H_{11}: \quad \rho \neq 0 \Rightarrow$ лог-линейная модель не верна

$$\rho = 0 \Rightarrow \mathrm{p}\lim \hat{\delta} = 0, \quad t = \frac{\hat{\delta}}{s.e.\hat{\delta}} \sim N(0,1)$$

PE-тест МакКиннона
сравнение невложенных моделей

$$A: \quad \mathbf{y} = \mathbf{x\beta} + \boldsymbol{\varepsilon}_1$$

$$B: \quad \ln(\mathbf{y}) = \mathbf{x\gamma} + \boldsymbol{\varepsilon}_2$$

$$C: \quad \mathbf{y} = \mathbf{x\beta} + \delta\left(\ln(\hat{\mathbf{y}}) - \widehat{\ln(\mathbf{y})}\right) + \mathbf{u}$$

$$D: \quad \ln(\mathbf{y}) = \mathbf{x\gamma} + \rho\left(\hat{\mathbf{y}} - \exp\left[\widehat{\ln(\mathbf{y})}\right]\right) + \mathbf{v}$$

Исходы:

1. H_{00} , H_{11} $\Rightarrow$ $\delta = 0$ проходит, $\rho = 0$ не проходит $\Rightarrow$ выбираем модель A
2. H_{01} , H_{10} $\Rightarrow$ $\delta = 0$ не проходит, $\rho = 0$ проходит $\Rightarrow$ выбираем модель В
3. H_{00} , H_{10} $\Rightarrow$ $\delta = 0$ проходит, $\rho = 0$ проходит $\Rightarrow$ модели эквивалентны
4. H_{01} , H_{10} $\Rightarrow$ $\delta = 0$ не проходит, $\rho = 0$ не проходит $\Rightarrow$ модели не адекватны

PE-тест МакКиннона

сравнение невложенных моделей

```
. reg  EARNINGS S
     ---------------------------------------------------------------------------
EARNINGS |      Coef.          Std. Err.      t    P>|t|     [95% Conf. Interval]
----------  --+-------------------------------------------------------------
           S |   1.073055   .1324501     8.10   0.000     .8129028    1.333206
       _cons |  -1.391004   1.820305    -0.76   0.445    -4.966354    2.184347
----------------------------------------------------------------------------
. predict EARN_hat
. g lg_EARN_hat=log( EARN_hat)

. g lgEarn=log( EARNINGS)
. reg   lgEarn S
----------------------------------------------------------------------------
     lgEarn |      Coef.      Std. Err.                t    P>|t|     [95% Conf. Interval]
- -----------+-------------------------------------------------------------
           S |   .0792256   .0082061     9.65   0.000     .0631077    .0953435
       _cons |   1.358919   .1127785    12.05   0.000     1.137406    1.580433
----------------------------------------------------------------------------
. predict lgEARN_hat
. g EARN_exp_log=exp( lgEARN_hat)

. g dEARN= EARN_hat- EARN_exp_log
. g dLgEARN= lg_EARN_hat- lgEARN_hat
```

PE-тест МакКиннона

сравнение невложенных моделей

```
. reg  EARNINGS S   dLgEARN
------------------------------------------------------------------------------
EARNINGS |      Coef.        Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           S |   1.089211    .135206     8.06   0.000     .8236454    1.35477
dLgEARN |  -5.783632   9.591753    -0.60   0.547    -24.62334    13.05607
      _cons |  -.8961894   1.997658    -0.45   0.654    -4.819902    3.027523
------------------------------------------------------------------------------

. reg   lgEarn S  dEARN
------------------------------------------------------------------------------
     lgEarn |      Coef.        Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          S |   .0723465   .0112963     6.40   0.000     .0501589    .0945342
   dEARN |   .0582987   .0657774     0.89   0.376    -.0708984    .1874957
     _cons |   1.362541   .1128738    12.07   0.000      1.140839    1.584243
------------------------------------------------------------------------------
```

Метод Зарембки

Применим для выбора из двух форм моделей (несравнимых непосредственно), в одной из которых зависимая переменная входит с логарифмом, а в другой - нет

Метод позволяет сравнить линейную и логарифмическую регрессии и оценить значимость наблюдаемых различий

Сравнение различных моделей парной регрессии

Метод Зарембки

1. Вычисляем среднее геометрическое значений зависимой переменной и все ее значения делятся на это среднее

$$Y_i^* = Y_i / \sqrt[n]{Y_1 Y_2 \ldots Y_n} = Y_i / e^{\frac{1}{n}(\ln Y_1 + \ln Y_2 + \ldots + \ln Y_n)}$$

2. Рассчитываются линейная и логарифмическая регрессии и сравниваются значения их суммы квадратов остатков (SSR)

$$Y_i^* = \alpha_1 + \beta_1 X_i + u_i \; ; SSR_1$$

$$\ln Y_i^* = \alpha_2 + \beta_2 - - X_i + u_i \; ; SSR_2$$

3. Вычисляем χ^2-статистику для оценки значимости различий

$$\chi^2 = \left(\frac{n}{2}\right) \cdot \left| \ln \frac{SSR1}{SSR2} \right|$$

4. Сравниваем с критическим значением χ^2-распределения с одной степенью свободы, различия значимы, если $\chi^2 > \chi^2_{\text{крит.}}$

Метод Зарембки

$$Y = \beta_1 + \beta_2 X + u$$

$$\log Y = \beta_1 + \beta_2 X + u$$

$$Y^* = Y \, / \, \text{среднее геометрическое } Y$$

$$Y^* = \beta_1' + \beta_2' X + u$$

$$\log Y^* = \beta_1' + \beta_2' X + u$$

$$\frac{n}{2}\log_e \frac{\text{larger } RSS}{\text{smaller } RSS} \qquad \chi^2(1)$$

Метод Зарембки

$$Y^* = (Y_1 Y_2 \ldots Y_n)^{\frac{1}{n}}$$

```
. means EARNINGS

    Variable |    Type        Obs        Mean       [95% Conf. Interval]
    -------------+----------------------------------------------------------
EARNINGS | Arithmetic     570    13.11782        12.44201   13.79364
             |  Geometric     570    11.36039        10.88473   11.85684
             |   Harmonic     570    10.00975        9.617686   10.43514
----------------------------------------------------------------------

. g EARNSTAR= EARNINGS/11.36039

. g LGEARNSTAR=log( EARNSTAR)
```

Метод Зарембки

*EARN** = *EARNINGS* / среднее геометрическое *EARNINGS*

```
. reg  EARNSTAR S

     Source |       SS               df       MS                  Number of obs =     570
   -------------+------------------------------                   F(  1,   568) =   65.64
      Model |  30.818438        1     30.818438                   Prob > F      =  0.0000
   Residual |  266.698183      568  .469539054                    R-squared     =  0.1036
-------------+------------------------------                      Adj R-squared =  0.1020
      Total |  297.516621   569   .52287631                       Root MSE      =  .68523

------------------------------------------------------------------------------
    EARNSTAR |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           S |   .0944558   .0116589     8.10   0.000     .0715559    .1173557
       _cons |  -.1224433   .1602326    -0.76   0.445    -.4371641    .1922775
------------------------------------------------------------------------------

. reg  LGEARNSTAR S

     Source |       SS               df       MS                  Number of obs =     570
-------------+------------------------------                      F(  1,   568) =   93.21
      Model |  21.6812541       1     21.6812541                  Prob > F      =  0.0000
   Residual |  132.120642      568  .232606765                    R-squared     =  0.1410
-------------+------------------------------                      Adj R-squared =  0.1395
      Total |  153.801896   569  .270302103                       Root MSE      =  .48229

------------------------------------------------------------------------------
  LGEARNSTAR |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           S |   .0792256   .0082061     9.65   0.000     .0631077    .0953435
       _cons |  -1.071213   .1127785    -9.50   0.000    -1.292727   -.8496997
------------------------------------------------------------------------------
```

Метод Зарембки

$$\frac{n}{2}\log_e \frac{\text{larger } RSS}{\text{smaller } RSS} = \frac{570}{2}\log_e \frac{266.7}{132.1} = 200.2$$

$$\chi^2_{\text{crit}} = 10.83,\ 1\ \text{d.f},\ 0.1\%$$

Полулогарифмическая модель больше подходит на 0.1% уровне значимости

Метод Бокса-Кокса (Box-Cox)

Идея метода: переменная

$$\frac{Y^{\lambda}-1}{\lambda}$$

при λ=1 превращается в линейную функцию

$$\frac{Y^{1}-1}{1}$$

а при λ→0 переходит в логарифм

$$\frac{Y^{\lambda}-1}{\lambda} \to \ln Y$$

Плавно изменяя λ, можно постепенно перейти от линейной регрессии к логарифмической, все время сравнивая качество

Сравнение различных моделей парной регрессии

Метод Бокса-Кокса (Box-Cox)

1. Преобразуем зависимую переменную по методу Зарембки

$$Y_i^* = Y_i / \sqrt[n]{Y_1 Y_2 \dots Y_n} = Y_i / e^{\frac{1}{n}(\ln Y_1 + \ln Y_2 + \dots + \ln Y_n)}$$

2. Рассчитываем новые переменные (преобразование Бокса-Кокса) при значениях λ от 1 до 0.

$$Y_{i(B-C)} = (Y_i^{*\lambda} - 1)/\lambda$$

и

$$X_{i(B-C)} = (X_i^{\lambda} - 1)/\lambda$$

3. Рассчитываем регрессии для новых переменных при значениях λ от 1 до 0.

$$Y_{i\,(B-C)} = \alpha + \beta X_{i\,(B-C)} + u_i$$

4. Выбираем минимальное значение суммы квадратов остатков (SSR), выбираем одну из крайних регрессий, к которой ближе точка минимума

BOX-COX TESTS

$$EARN^* = EARNINGS \,/\, \text{geometric mean of } EARNINGS$$

```
. boxcox  EARNSTAR   AGE  S SEX  , model(theta)  nolog

                                          Number of obs   =       570
                                          LR chi2(4)      =     122.13
Log likelihood = -367.8914                  Prob > chi2     =     0.000
------------------------------------------------------------------------------
EARNSTAR |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  /lambda |   .6167905    .634108     0.97   0.331    -.6260383    1.859619
    /theta |  -.2561972   .0677446    -3.78   0.000    -.3889742   -.1234203
------------------------------------------------------------------------------
---------------------------------------------------------
  Test             Restricted
   H0:           log likelihood      chi2      Prob > chi2
---------------------------------------------------------
theta=lambda = -1     -423.15698      110.53         0.000
theta=lambda =  0     -376.24123       16.70         0.000
theta=lambda =  1     -579.52174      423.26         0.000
---------------------------------------------------------
```

```
Estimates of scale-variant parameters
----------------------------
                   |      Coef.
-------------------+--------------
     Notrans       |
            _cons  |  -1.921976
-------------------+--------------
Trans        AGE  |     .026491
                 S |   .2225095
       -------------+--------------
          /sigma |    .4613885
----------------------------
```

BOX-COX TESTS

$$EARN^* = EARNINGS \,/\, \text{geometric mean of } EARNINGS$$

```
. boxcox  EARNSTAR   AGE  S SEX  , model(rhsonly)  nolog
                                   Estimating full model
                                   Number of obs   =       570
                                   LR chi2(4)      =     88.29
Log likelihood = -579.35478                 Prob > chi2     =     0.000
------------------------------------------------------------------------------
   EARNSTAR |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    /lambda |   1.442164   .7869089     1.83   0.067    -.1001486    2.984477
------------------------------------------------------------------------------

---------------------------------------------------------
  Test         Restricted      LR statistic      P-Value
   H0:       log likelihood        chi2        Prob > chi2
---------------------------------------------------------
lambda = -1    -585.58984       12.47          0.000
lambda =  0    -581.35271        4.00          0.046
lambda =  1    -579.52174        0.33          0.563
---------------------------------------------------------
```

Estimates of scale-variant parameters

```
----------------------------
             |      Coef.
-------------+--------------
Notrans      |
       _cons |  -.4541757
-------------+--------------
Trans        |
         AGE |   .0052884
           S |    .030497
         SEX |   .2244905
-------------+--------------
      /sigma |   .6686285
----------------------------
```

BOX-COX TESTS

*EARN** = *EARNINGS* / geometric mean of *EARNINGS*

```
. boxcox  EARNSTAR   AGE  S SEX  , model(lhsonly)  nolog
                                                  Number of obs   =       570
                                                  LR chi2(3)      =    121.78
Log likelihood = -368.06337                       Prob > chi2     =     0.000
------------------------------------------------------------------------------
    EARNSTAR |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      /theta |   -.252988   .0674564    -3.75   0.000    -.3852001   -.1207759
------------------------------------------------------------------------------

---------------------------------------------------------
   Test         Restricted      LR statistic      P-Value
    H0:       log likelihood        chi2       Prob > chi2
---------------------------------------------------------
theta = -1     -422.73619        109.35          0.000
theta =  0     -375.39583         14.66          0.000
theta =  1     -579.52174        422.92          0.000
---------------------------------------------------------
```

Estimates of scale-variant parameters

```
----------------------------
             |      Coef.
-------------+--------------
Notrans      |
         AGE |   .0075793
           S |    .081562
         SEX |   .2278146
       _cons |  -1.743971
-------------+--------------
      /sigma |   .4615277
----------------------------
```

BOX-COX TESTS

$$EARN^* = EARNINGS \,/\, \text{geometric mean of } EARNINGS$$

```
. boxcox  lgEARNSTAR   AGE  S SEX  , model(rhsonly)  nolog
                                    Number of obs   =       570
                                    LR chi2(4)      =     120.20
Log likelihood = -375.35439                    Prob > chi2     =      0.000
------------------------------------------------------------------------------
 lgEARNSTAR |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    /lambda |    .811331   .6457532     1.26   0.209    -.454322    2.076984
------------------------------------------------------------------------------

---------------------------------------------------------
   Test         Restricted      LR statistic      P-Value
   H0:       log likelihood       chi2       Prob > chi2
---------------------------------------------------------
lambda = -1    -380.30004        9.89          0.002
lambda =  0    -376.24127        1.77          0.183
lambda =  1    -375.39587        0.08          0.773
---------------------------------------------------------
```

```
Estimates of scale-variant parameters
----------------------------
             |      Coef.
-------------+--------------
Notrans      |
       _cons |  -1.690962
-------------+--------------
Trans        |
         AGE |   .0182433
           S |   .1343269
         SEX |   .2448187
-------------+--------------
      /sigma |   .4674691
----------------------------
```

Тест ошибочной спецификации Амемии (Акаике)

Является вариантом скорректированного коэффициента детерминации и превосходит его

$$PC = \frac{RSS \cdot (n+m)}{n-m}$$

Выбирается уравнение с меньшим значением PC

Смысл теста Амемии в том, что он позволяет минимизировать среднюю ошибку оценки *b*

$$MSE = Var(b) + (смещение\ b)^2$$

ИНФОРМАЦИОННЫЕ КРИТЕРИИ

$t = 1, \ldots, n$

1. $y_t = f(\mathbf{X}_t, \boldsymbol{\varepsilon}_t); \quad \mathbf{X}_t = \left(X_{1t}, \ldots, X_{lt} \right)$

2. $y_t = g(\mathbf{Z}_t, \boldsymbol{\upsilon}_t); \quad \mathbf{Z}_t = \left(Z_{1t}, \ldots, Z_{mt} \right)$

AIC (Akaike information criterion):

$$AIC1 = \frac{2k}{n} \ln\left(\frac{RSS}{n} \right); \quad AIC2 = e^{\frac{2k}{n}} \frac{RSS}{n}$$

SIC (Schwarz information criterion):

$$SIC1 = \frac{k \ln(n)}{n} \ln\left(\frac{RSS}{n} \right); \quad SIC2 = n^{\frac{k}{n}} \frac{RSS}{n}$$

Первые множители - kind of degree of freedom penalty factor

Соотвественно:

ЧЕМ МЕНЬШЕ ЗНАЧЕНИЕ ИНФОРМАЦИОННОГО КРИТЕРЕИЯ,

ТЕМ МОДЕЛЬ ЛУЧШЕ

BOX-COX TESTS

$EARN^* = EARNINGS$ / geometric mean of $EARNINGS$

```
. reg  EARNSTAR S

      Source |       SS       df       MS              Number of obs =     570
-------------+------------------------------           F(  1,   568) =   65.64
       Model |   30.818438     1   30.818438           Prob > F      =  0.0000
    Residual |  266.698183   568  .469539054           R-squared     =  0.1036
-------------+------------------------------           Adj R-squared =  0.1020
       Total |  297.516621   569   .52287631           Root MSE      =  .68523

------------------------------------------------------------------------------
    EARNSTAR |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           S |   .0944558   .0116589     8.10   0.000     .0715559    .1173557
       _cons |  -.1224433   .1602326    -0.76   0.445    -.4371641    .1922775
------------------------------------------------------------------------------
```

```
. icomp

  Information criteria
AIC   = 1186.66424
SBIC  = 1191.00988
ICOMP = 1191.96900
```

```
. reg  LGEARNSTAR S

      Source |       SS       df       MS              Number of obs =     570
-------------+------------------------------           F(  1,   568) =   93.21
       Model |  21.6812541     1  21.6812541           Prob > F      =  0.0000
    Residual |  132.120642   568  .232606765           R-squared     =  0.1410
-------------+------------------------------           Adj R-squared =  0.1395
       Total |  153.801896   569  .270302103           Root MSE      =  .48229

------------------------------------------------------------------------------
  LGEARNSTAR |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           S |   .0792256   .0082061     9.65   0.000     .0631077    .0953435
       _cons |  -1.071213   .1127785    -9.50   0.000    -1.292727   -.8496997
------------------------------------------------------------------------------
```

```
. icomp

  Information criteria
AIC   = 786.29502
SBIC  = 790.64065
ICOMP = 791.59977
```

ИНФОРМАЦИОННЫЕ КРИТЕРИИ

$$EARN^* = EARNINGS \,/\, \text{geometric mean of } EARNINGS$$

```
. quietly reg  EARNSTAR S
. fitstat, saving(IC)

Measures of Fit for regress of EARNSTAR
Log-Lik Intercept Only:     -623.498     Log-Lik Full Model:        -592.332
D(568):                     1184.664     LR(1):                       62.331
                                         Prob > LR:                    0.000
R2:                            0.104     Adjusted R2:                  0.102
AIC:                           2.085     AIC*n:                     1188.664
BIC:                       -2419.657     BIC':                       -55.985
(Indices saved in matrix fs_IC)


. quietly reg  lgEARNSTAR S
. fitstat, force u(IC)

Measures of Fit for regress of lgEARNSTAR
                                 Current          Saved      Difference
Model:                           regress        regress
N:                                  570            570               0
Log-Lik Intercept Only:        -435.453       -623.498         188.044
Log-Lik Full Model:            -392.148       -592.332         200.185
D:                         784.295(568)  1184.664(568)    -400.369(0)
LR:                           86.611(1)      62.331(1)      24.281(0)
Prob > LR:                        0.000          0.000          -0.000
R2:                               0.141          0.104           0.037
Adjusted R2:                      0.139          0.102           0.037
AIC:                              1.383          2.085          -0.702
AIC*n:                          788.295       1188.664        -400.369
BIC:                          -2820.026      -2419.657        -400.369
BIC':                           -80.266        -55.985         -24.281
```

Конец лекции