Методический семинар 12.04.2012, часть 4

Метод пошагового включения переменных и метод пошагового исключения переменных

Демидова О.А.

Пошаговый метод предусматривает построение модели последовательно по шагам. Для пошагового метода включения (*Forwardstepwisemethod*) на каждом шаге оценивается вклад в регрессионную функцию не включённых в модель переменных. Переменная, обеспечивающая наибольший вклад, включается в модель, после чего производится переход к следующему шагу. Для пошагового метода исключения (*Backwardstepwisemethod*) характерно включение в модель на первом этапе всех переменных, после чего производится их последовательное исключение.

Пример

Файл CLOTHING содержит данные о годовых продажах и другие характеристики 400 голландских магазинов модной одежды в 1990 г.

Пример

Список переменных:

tsales	: Annual sales in Dutch guilders
sales	: sales per square meter
margin	: Gross-profit-margin
nown	: Number of owners (managers)
nfull	: Number of full-timers
npart	: Number of part-timers
naux	: Number of helpers (temporary workers)
hoursw	: Total number of hours worked
hourspw	: Number of hours worked per worker
inv1	: Investment in shop-premises
inv2	: Investment in automation.
ssize	: Sales floorspace of the store (in m2).
start	: year start of business

Метод последовательного исключения переменных

```
stepwise, pr(0.1): reg   sales margin nown nfull npart      naux hoursw hourspw inv1
              inv2        ssize       start
begin with full model
p = 0.9731 >= 0.1000  removing inv1
p = 0.7996 >= 0.1000  removing hoursw
p = 0.6921 >= 0.1000  removing inv2
p = 0.5229 >= 0.1000  removing start
Source       SS       df       MS Number of obs =     400
              F(  7,   392) =   41.65
Model   2.3795e+09     7   339924634    Prob > F      =  0.0000
Residual   3.1996e+09   392  8162299.11           R-squared     =  0.4265
              Adj R-squared =  0.4163
Total   5.5791e+09   399    13982691     Root MSE      =    2857

sales       Coef.          Std. Err.      t       P>t        [95% Conf. Interval]
margin     59.90574    29.54534     2.03    0.043       1.818588    117.9929
nown      672.7859     226.1869     2.97    0.003       228.0947    1117.477
nfull      1309.047    153.066       8.55    0.000       1008.114     1609.98
npart      1061.772    230.9229     4.60     0.000       607.7695    1515.774
naux        935.7677   376.4691     2.49     0.013       195.6167    1675.919
ssize       -24.31604  1.641706    -14.81    0.000      -27.54369    -21.08839
hourspw    223.7379    23.18619     9.65    0.000       178.1531    269.3228
_cons       -3091.438  1294.97      -2.39    0.017      -5637.394    -545.4823
```

Метод последовательного исключения переменных

```
reg   sales margin nown nfull npart naux hoursw hourspw inv1 inv2 ssize start

      Source |       SS       df       MS              Number of obs =     400
-------------+------------------------------           F( 11,   388) =   26.33
       Model |  2.3846e+09    11   216785544           Prob > F      =  0.0000
    Residual |  3.1945e+09   388  8233125.52           R-squared     =  0.4274
-------------+------------------------------           Adj R-squared =  0.4112
       Total |  5.5791e+09   399    13982691           Root MSE      =  2869.3

------------------------------------------------------------------------------
       sales |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      margin |   69.08252   32.96368     2.10   0.037     4.272738    133.8923
        nown |   740.8716   371.5816     1.99   0.047     10.30624    1471.437
       nfull |   1378.757   293.2506     4.70   0.000     802.1977    1955.316
       npart |   1111.709   267.8409     4.15   0.000     585.1077     1638.31
        naux |    996.577   449.8185     2.22   0.027     112.1902    1880.964
      hoursw |  -2.737927   10.73368    -0.26   0.799    -23.84139    18.36553
     hourspw |   243.4413   70.43075     3.46   0.001     104.9676     381.915
        inv1 |  -.0000524   .0015552    -0.03   0.973    -.0031101    .0030053
        inv2 |  -.0014312   .0039795    -0.36   0.719    -.0092553    .0063928
       ssize |  -24.17391    1.69074   -14.30   0.000    -27.49807   -20.84975
       start |  -8.224629    12.7344    -0.65   0.519     -33.2617    16.81244
       _cons |  -3509.928   1988.711    -1.76   0.078    -7419.928    400.0707
------------------------------------------------------------------------------
```

Метод последовательного исключения переменных

```
reg   sales margin nown nfull npart naux hoursw hourspw inv2 ssize start

      Source |       SS       df       MS              Number of obs =     400
-------------+------------------------------           F( 10,   389) =   29.04
       Model |  2.3846e+09    10   238463163           Prob > F      =  0.0000
    Residual |  3.1945e+09   389  8211984.72           R-squared     =  0.4274
-------------+------------------------------           Adj R-squared =  0.4127
       Total |  5.5791e+09   399    13982691           Root MSE      =  2865.7

------------------------------------------------------------------------------
       sales |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      margin |   69.07279   32.92006     2.10   0.037     4.349271    133.7963
        nown |   740.4297   370.8731     2.00   0.047     11.26306    1469.596
       nfull |   1378.033    292.088     4.72   0.000     803.7644    1952.302
       npart |   1110.099   263.2089     4.22   0.000      592.609    1627.589
        naux |   997.3463   448.6619     2.22   0.027     115.2406    1879.452
      hoursw |  -2.720658   10.70767    -0.25   0.800    -23.77281     18.3315
     hourspw |    243.301   70.21732     3.46   0.001     105.2481     381.354
        inv2 |  -.0014861   .0036262    -0.41   0.682    -.0086155    .0056433
       ssize |  -24.17753   1.685145   -14.35   0.000    -27.49066    -20.8644
       start |  -8.178698   12.64501    -0.65   0.518    -33.03982    16.68242
       _cons |  -3508.412   1985.648    -1.77   0.078    -7412.357    395.5331
------------------------------------------------------------------------------
```

Метод последовательного исключения переменных

```
reg       sales margin nown nfull npart naux hourspw inv2          ssize start

          Source       SS       df       MS Number of obs      =     400
                                             F(  9,   390)      =   32.34
          Model   2.3841e+09     9   264900164  Prob > F       =  0.0000
          Residual  3.1950e+09   390  8192287.73  R-squared          =  0.4273
                                             Adj R-squared      =  0.4141
          Total   5.5791e+09   399    13982691  Root MSE           =  2862.2


          sales       Coef.      Std. Err.        t      P>t        [95% Conf.Interval]

          margin      68.03328    32.62566     2.09   0.038    3.889102 132.1775
          nown        666.3008    228.7003     2.91   0.004    216.6611 1115.94
          nfull       1315.042    154.2658     8.52   0.000    1011.746 1618.339
          npart       1079.118    232.9839     4.63   0.000    621.0568 1537.18
          naux        936.1306    378.0323     2.48   0.014    192.8944 1679.367
          hourspw     226.493      23.51939    9.63   0.000    180.2523 272.7337
          inv2        -.0014329   .0036158    -0.40   0.692    -.0085418 .005676
          ssize       -24.24519   1.661976   -14.59   0.000   -27.51275 -20.97764
          start       -7.799186   12.54142    -0.62   0.534   -32.45643  16.85806
          _cons       -3127.498   1300.478    -2.40   0.017   -5684.322  -570.6728
```

И т.д.

Метод последовательного включения переменных

```
stepwise, pe(0.1): reg   sales margin nown nfull npart      naux hoursw hourspw inv1
              inv2        ssize       start
begin with empty model
p = 0.0000 <  0.1000  adding  ssize
p = 0.0000 <  0.1000  adding  hoursw
p = 0.0002 <  0.1000  adding  nfull
p = 0.0025 <  0.1000  adding  npart
p = 0.0063 <  0.1000  adding  hourspw
p = 0.0898 <  0.1000  adding  margin
Source       SS       df       MS Number of obs =     400
             F(  6,   393) =   46.98
Model   2.3303e+09     6   388376865    Prob > F      =  0.0000
Residual   3.2488e+09   393  8266749.37           R-squared     =  0.4177
                                                  Adj R-squared =  0.4088
Total   5.5791e+09   399    13982691     Root MSE      =  2875.2

sales       Coef.           Std. Err.      t         P>t   [95% Conf. Interval]
ssize      -24.24045     1.6359       -14.82   0.000   -27.45666   -21.02424
hoursw      18.15945      6.112656     2.97    0.003    6.141849    30.17704
nfull        897.7947     211.7212     4.24    0.000    481.5469    1314.042
npart        855.3101     242.858      3.52    0.000    377.8467    1332.773
hourspw     113.2123     45.44868      2.49    0.013    23.8594     202.5653
margin       50.11734    29.46866      1.70    0.090   -7.818595    108.0533
_cons        510.3715    1198.688      0.43    0.671   -1846.271    2867.014
```

Метод последовательного включения переменных

```
stepwise, pe(0.1): reg  start ssize inv2 inv1 hourspw hoursw naux npart nfull nown
margin
                  begin with empty model
p = 0.0000 <  0.1000  adding  margin
p = 0.0111 <  0.1000  adding  hourspw
p = 0.0211 <  0.1000  adding  npart
p = 0.0437 <  0.1000  adding  nown

      Source |       SS       df       MS              Number of obs =     400
-------------+------------------------------           F(  4,   395) =   34.54
       Model |  18277.8738     4  4569.46846            Prob > F      =  0.0000
    Residual |  52259.5368   395  132.302625            R-squared     =  0.2591
-------------+------------------------------           Adj R-squared =  0.2516
       Total |  70537.4106   399   176.78549            Root MSE      =  11.502

------------------------------------------------------------------------------
       start |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      margin |   1.076729   .1178685     9.13   0.000     .8450005    1.308457
    hourspw |   .2493196   .0830748     3.00   0.003     .0859955    .4126437
       npart |   2.035627   .8341378     2.44   0.015      .395722    3.675532
       nown |  -1.842024   .9102377    -2.02   0.044     -3.63154   -.0525076
       _cons |  -4.488794   4.443053    -1.01   0.313    -13.22378    4.246195
------------------------------------------------------------------------------
```

Теорема о “корне из r”

Пусть оценены коэффициенты регрессии

$$Y = \beta_1 + \beta_2 X_2 + \ldots + \beta_k X_k + u$$

Т.е. $\hat{Y} = b_1 + b_2 X_2 + \ldots + b_k X_k$

Если для r оценок коэффициентов при непостоянных факторах выполняется условие $|t| < \sqrt{r}$ **, то при удалении соответствующих факторов X качество подгонки регрессии может увеличиться, т.е. при удалении этой группы факторов** R^2_{adj} **может увеличиться.**

Замечание 1. Условие является необходимым, но не достаточным, т.е. при удалении соответствующей группы факторов R^2_{adj} **может не увеличиться.**

Теорема о “корне из r”

Замечание 2. Если r = 1, то условие является не только необходимым, но и достаточным, т.е. при удалении одного фактора с t – статистикой, меньше 1 (по модулю) R^2_{adj} увеличится.

Теорема о “корне из r”

Замечание 2. Если r = 1, то условие является не только необходимым, но и достаточным, т.е. при удалении одного фактора с t – статистикой, меньше 1 (по модулю) R^2_{adj} увеличится.

Пример 2

В файле data 6.6 содержатся данные об аренде жилья из базы НОБУС о месячной арендной плате за жилье.

Переменные:

a003pt – тип населенного пункта (1 – город с численностью более 1 млн.чел, …, 8 – село),

r203 – месячная арендная плата за жилье, если снимать,

r204 – год постройки жилья,

r205 – материал, из которого построены внешние стены жилья (1- кирпич, 2 – бетонные панели, 3- камень, 4 – дерево, 5 – другой материал),

r206 – этажность здания,

r207 – наличие лифта в доме,

r208m01 – общая площадь жилища в м2,

r208m02 – жилая площадь жилища в м2,

r209 – наличие электричества в д/х,

r210 – тип отопления в доме (1 – коллективное центральное отопление, 2 - индивидуальное отопление газом, 3 - индивидуальное отопление дровами, 4 – другое),

r211- откуда берется вода (1 – водопровод в квартире, 2 – колодец во дворе, 3 – общественная колонка, 4 – общественный колодец, 5 – из водоема, 6 – вода привозная, 7 – другое),

Пример 2

r212m1- наличие канализации (1 – да, 2 – нет),

r212m2 - наличие горячего водоснабжения (1 – да, 2 – нет),

r212m3 - наличие санузла внутри помещения (1 – да, 2 – нет),

r212m4 - наличие ванны, душевой (1 – да, 2 – нет),

r212m5 - наличие газа (1 – да, 2 – нет),

r212m6 - наличие электроплиты (1 – да, 2 – нет),

r212m7 - наличие телефона (1 – да, 2 – нет).

```
reg  r203 r204 r205 r206 r207 r208m01 r208m02 r209 r210 r211 r212m1 r212m2 r212m3
r212m4 r212m5 r212m6 r212m7

                                                                  Adj R-squared =  0.2860
------------------------------------------------------------------------------
      r203 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----------+------------------------------------------------------------------
      r204 |   .4397018   .1328318     3.31   0.001     .1793253    .7000784
      r205 |  -30.98764   11.19581    -2.77   0.006    -52.93364   -9.041638
      r206 |   202.0962   6.181453    32.69   0.000     189.9794    214.2131
      r207 |   3.509889   2.505189     1.40   0.161    -1.400775    8.420554
    r208m01 |   1.490023   .6665292     2.24   0.025     .1834942    2.796551
    r208m02 |   7.332909   1.073681     6.83   0.000     5.228282    9.437536
      r209 |   5.080075   13.77277     0.37   0.712    -21.91728    32.07743
      r210 |   21.17051   13.04943     1.62   0.105    -4.408944    46.74996
      r211 |  -30.15348   12.42321    -2.43   0.015    -54.50542   -5.801542
    r212m1 |  -123.8872   31.91985    -3.88   0.000    -186.4564   -61.31798
    r212m2 |  -242.4452   36.87398    -6.57   0.000    -314.7255    -170.165
    r212m3 |   254.495   49.92503     5.10   0.000     156.6321    352.3579
    r212m4 |  -371.7929   56.52312    -6.58   0.000    -482.5894   -260.9965
    r212m5 |   104.5059   32.15591     3.25   0.001     41.47394    167.5378
    r212m6 |  -380.7715   39.61693    -9.61   0.000    -458.4284   -303.1145
    r212m7 |  -2.487258   1.450154    -1.72   0.086    -5.329846    .355329
     _cons |   827.4628   252.4822     3.28   0.001      332.548    1322.378
```

r = 4

reg r203 r204 r205 r206 r208m01 r208m02 r211 r212m1 r212m2 r212m3 r212m4 r212m5 r212m6

Adj R-squared = 0.2833

r203	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
r204	.466184	.1128818	4.13	0.000	.2449134	.6874545
r205	-25.08236	10.94542	-2.29	0.022	-46.53754	-3.62718
r206	196.7912	5.54161	35.51	0.000	185.9286	207.6539
r208m01	1.667629	.6405464	2.60	0.009	.4120321	2.923226
r208m02	7.087622	1.054299	6.72	0.000	5.020988	9.154256
r211	-8.784775	4.749058	-1.85	0.064	-18.09386	.5243118
r212m1	-96.20204	28.61918	-3.36	0.001	-152.3013	-40.10282
r212m2	-267.7058	30.72414	-8.71	0.000	-327.9311	-207.4804
r212m3	277.6937	48.62443	5.71	0.000	182.3802	373.0071
r212m4	-408.502	53.80688	-7.59	0.000	-513.9741	-303.03
r212m5	105.3527	31.96179	3.30	0.001	42.70127	168.0041
r212m6	-376.0485	39.35726	-9.55	0.000	-453.1964	-298.9005
_cons	799.5115	215.8768	3.70	0.000	376.3505	1222.672

.

r = 1

reg r203 r204 r205 r206 r207 r208m01 r208m02 r210 r211 r212m1 r212m2 r212m3 r212m4 r212m5 r212m6 r212m7 Adj R-squared = 0.2861

r203	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
r204	.4392003	.1316196	3.34	0.001	.1811999	.6972007
r205	-30.34527	11.10232	-2.73	0.006	-52.10801	-8.582533
r206	202.8326	5.845733	34.70	0.000	191.3738	214.2914
r207	3.676723	2.440634	1.51	0.132	-1.107401	8.460847
r208m01	1.472691	.6652407	2.21	0.027	.1686886	2.776694
r208m02	7.358593	1.072067	6.86	0.000	5.25713	9.460056
r210	22.04013	12.8414	1.72	0.086	-3.131529	47.2118
r211	-30.87586	12.26581	-2.52	0.012	-54.91926	-6.832462
r212m1	-126.31	31.2354	-4.04	0.000	-187.5375	-65.08247
r212m2	-237.8919	33.00972	-7.21	0.000	-302.5975	-173.1864
r212m3	258.233	49.00863	5.27	0.000	162.1664	354.2995
r212m4	-375.8064	55.07122	-6.82	0.000	-483.7569	-267.856
r212m5	104.3208	32.14783	3.25	0.001	41.30472	167.3369
r212m6	-379.0669	39.40656	-9.62	0.000	-456.3115	-301.8223
r212m7	-2.522396	1.437339	-1.75	0.079	-5.339864	.2950725
_cons	823.0116	250.8121	3.28	0.001	331.3705	1314.653

```
 stepwise, pr(0.05): reg  r203 r204 r205 r206 r207 r208m01 r208m02 r209 r210 r211
r212m1 r212m2 r212m3 r212m4 r212m5 r212m6 r212m7
                  begin with full model
p = 0.7122 >= 0.0500  removing r209
p = 0.1288 >= 0.0500  removing r207
p = 0.2593 >= 0.0500  removing r212m7                          Adj R-squared =  0.2859

      r203 |     Coef.      Std. Err.       t     P>|t|     [95% Conf. Interval]
-----------+------------------------------------------------------------------
      r204 |    .410805   .1286401     3.19    0.001     .1586451    .6629649
      r205 |   -28.0391   10.99239    -2.55    0.011    -49.58634   -6.491858
      r206 |   204.1189   5.790608    35.25    0.000     192.7682    215.4696
    r212m6 |  -382.0085   39.37463    -9.70    0.000    -459.1905   -304.8265
   r208m01 |   1.529256   .6433419     2.38    0.017     .2681792    2.790333
   r208m02 |   7.338344    1.05609     6.95    0.000     5.268199    9.408489
    r212m5 |   99.87517   31.98271     3.12    0.002     37.18276    162.5676
      r210 |   24.66382   12.52983     1.97    0.049     .1028884    49.22476
      r211 |  -31.32608    12.2433    -2.56    0.011    -55.32534   -7.326805
    r212m1 |   -124.602   31.17879    -4.00    0.000    -185.7185   -63.48541
    r212m2 |   -232.177   33.21961    -6.99    0.000     -297.294    -167.06
    r212m3 |   259.1024   49.00936     5.29    0.000     163.0344    355.1704
    r212m4 |  -381.0481   54.99462    -6.93    0.000    -488.8484   -273.2478
     _cons |   872.6439   245.0481     3.56    0.000     392.3015    1352.986
------------------------------------------------------------------------------
```

```
 stepwise, pe(0.05): reg r203  r206 r212m2 r208m02 r212m6 r212m4 r205 r212m5 r212m3 r212m1
r204  r208m01
                  begin with empty model
p = 0.0000 <  0.0500  adding  r206
p = 0.0000 <  0.0500  adding  r212m2
p = 0.0000 <  0.0500  adding  r208m02
p = 0.0000 <  0.0500  adding  r212m6
p = 0.0000 <  0.0500  adding  r204
p = 0.0000 <  0.0500  adding  r205
p = 0.0000 <  0.0500  adding  r212m4
p = 0.0000 <  0.0500  adding  r212m3
p = 0.0020 <  0.0500  adding  r212m1
p = 0.0010 <  0.0500  adding  r212m5
p = 0.0185 <  0.0500  adding  r208m01                                  Adj R-squared =  0.2831
      r203 |     Coef.       Std. Err.      t        P>|t|     [95% Conf. Interval]
-------------+------------------------------------------------------------------
      r206 |   197.3864   5.532918    35.67   0.000     186.5408     208.232
    r212m2 |  -268.8213   30.72187    -8.75   0.000    -329.0422    -208.6004
   r208m02 |   7.303287   1.047957     6.97   0.000     5.249085     9.35749
    r212m6 |  -376.2769   39.36173    -9.56   0.000    -453.4336    -299.1202
      r204 |    .542568   .1050704     5.16   0.000     .3366094     .7485266
      r205 |  -25.92254   10.93729    -2.37   0.018    -47.36178    -4.483294
    r212m4 |  -408.7439   53.81311    -7.60   0.000    -514.2282    -303.2596
    r212m3 |   267.5707   48.32123     5.54   0.000     172.8516     362.2898
    r212m1 |  -90.83817   28.47528    -3.19   0.001    -146.6553    -35.02102
    r212m5 |   104.7584   31.96397     3.28   0.001     42.10273     167.4141
   r208m01 |   1.492547   .6335902     2.36   0.019     .2505861     2.734509
     _cons |   646.5848   199.4431     3.24   0.001     255.6373     1037.532
```