NATIONAL RESEARCH UNIVERSITY

Лекция 19.11.21, 2 модуль

Функциональные преобразования переменных в линейной регрессионной модели

Выбор между моделями

Демидова О.А.

E-mail:demidova@hse.ru

План лекции

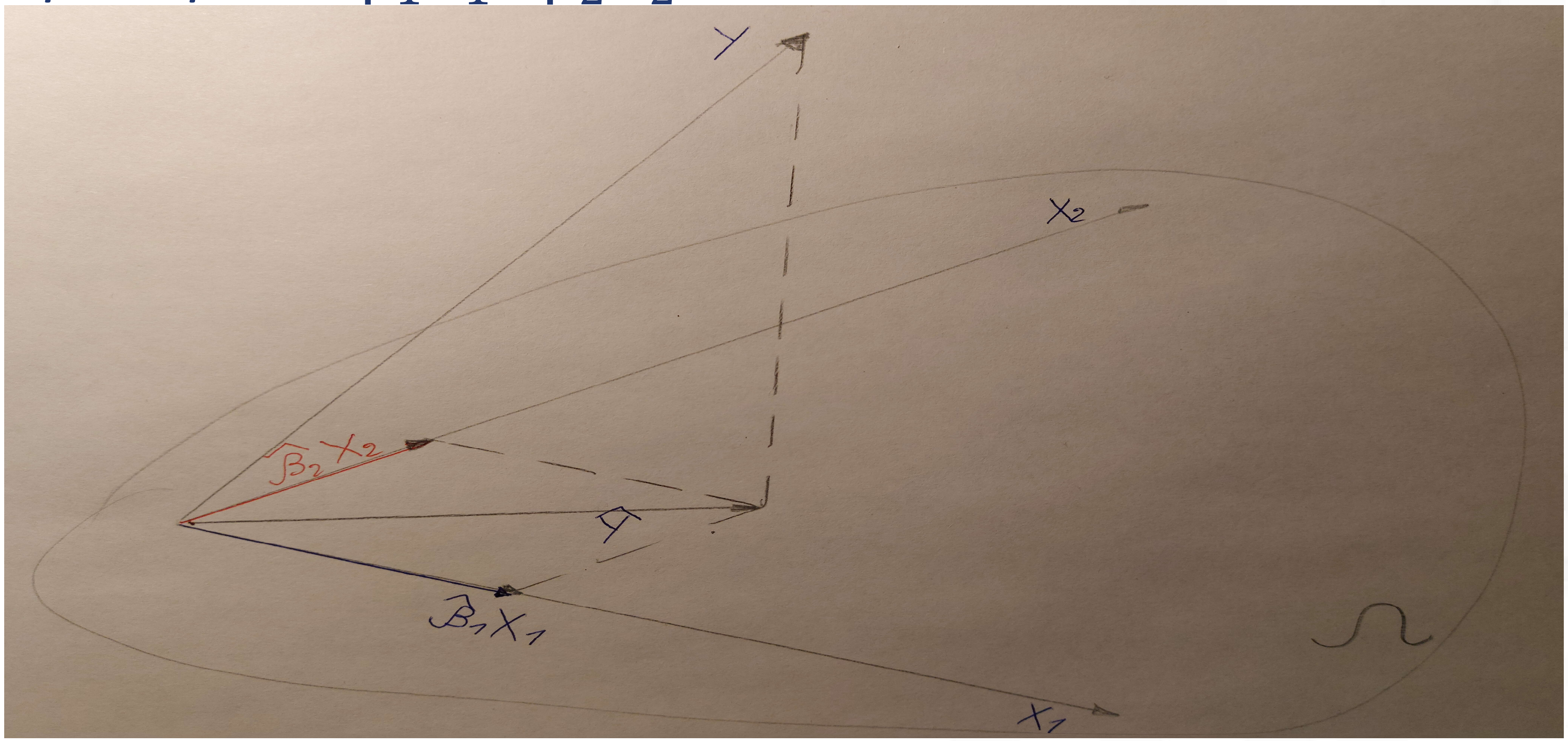
- Влияние изменения масштаба измерения переменных на оценки коэффициентов регрессии и их дисперсий
- Регрессия в центрированных и нормированных переменных
- Функциональные преобразования переменных в линейной регрессионной модели
- Линейная в логарифмах регрессия, как модель с постоянной эластичностью
- Модель с постоянными темпами роста (полулогарифмическая модель)
- Интерпретация оценок коэффициентов различных функциональных форм.
- Выбор между моделями. Тесты Бокса-Кокса, Бера и МакАлера, МакКиннона, Уайта и Дэвидсона

Влияние изменения масштаба

$Y = X\beta + \varepsilon,$

$\hat{\beta} = (X'X)^{-1}\,(X'Y), \quad var(\hat{\beta}) = \hat{\sigma}_{\varepsilon}^{2}(X'X)^{-1}$

Пример: $\hat{Y} = \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$



Влияние изменения масштаба. Пример

EARNINGS – почасовая з/п в $,
WEIGHT02 – вес в фунтах в 2002 (1 фунт = 0.454 кг),
HEIGHT – рост в дюймах (1 дюйм = 2.54 см),
ASVABC – общий результат тестов,
MALE – 1 для мужчин и 0 для женщин.

```
. sum EARNINGS WEIGHT02 HEIGHT ASVABC MALE
```

Variable	Obs	Mean	Std. Dev.	Min	Max
EARNINGS	540	19.71924	14.60151	2.13	127.9
WEIGHT02	540	180.0796	42.78456	87	400
HEIGHT	540	67.43333	4.04992	59	77
ASVABC	540	51.73125	9.114551	29.00336	66.07963
MALE	540	.5	.5004636	0	1

Влияние изменения масштаба. Пример

```
. reg EARNINGS WEIGHT02 HEIGHT ASVABC MALE

      Source |       SS           df       MS      Number of obs   =       540
-------------+----------------------------------   F(4, 535)       =     30.26
       Model |  21202.8353         4  5300.70882   Prob > F        =    0.0000
    Residual |  93714.1261       535  175.166591   R-squared       =    0.1845
-------------+----------------------------------   Adj R-squared   =    0.1784
       Total |  114916.961       539   213.20401   Root MSE        =    13.235

------------------------------------------------------------------------------
    EARNINGS |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    WEIGHT02 |  -.0387176   .0165291    -2.34   0.020    -.0711876   -.0062476
      HEIGHT |    .305907   .2240537     1.37   0.173    -.1342259    .7460399
      ASVABC |   .5281508   .0637492     8.28   0.000     .4029213    .6533803
        MALE |   6.127074    1.69514     3.61   0.000     2.797127    9.457022
       _cons |  -24.32227   13.61505    -1.79   0.075    -51.06777    2.423235
------------------------------------------------------------------------------
. reg EARNINGS WEIGHTkg HEIGHTm ASVABC MALE

      Source |       SS           df       MS      Number of obs   =       540
-------------+----------------------------------   F(4, 535)       =     30.26
       Model |  21202.8349         4  5300.70874   Prob > F        =    0.0000
    Residual |  93714.1264       535  175.166591   R-squared       =    0.1845
-------------+----------------------------------   Adj R-squared   =    0.1784
       Total |  114916.961       539   213.20401   Root MSE        =    13.235

------------------------------------------------------------------------------
    EARNINGS |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    WEIGHTkg |  -.0852811   .0364078    -2.34   0.020    -.1568009   -.0137613
     HEIGHTm |   12.04358   8.821012     1.37   0.173     -5.28449    29.37164
      ASVABC |   .5281508   .0637492     8.28   0.000     .4029213    .6533803
        MALE |   6.127075    1.69514     3.61   0.000     2.797127    9.457022
       _cons |  -24.32226   13.61504    -1.79   0.075    -51.06777    2.423239
------------------------------------------------------------------------------
```

Влияние изменения масштаба. Пример

```
. reg EARNINGS WEIGHTkg HEIGHTcm ASVABC MALE

      Source |       SS       df       MS              Number of obs =     540
-------------+------------------------------           F(4, 535)     =   30.26
       Model |  21202.8352     4  5300.70879           Prob > F      =  0.0000
    Residual |  93714.1262   535  175.166591           R-squared     =  0.1845
-------------+------------------------------           Adj R-squared =  0.1784
       Total |  114916.961   539  213.20401            Root MSE      =  13.235

------------------------------------------------------------------------------
    EARNINGS |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    WEIGHTkg |  -.0852811   .0364078    -2.34   0.020    -.1568009   -.0137613
    HEIGHTcm |   .1204358   .0882101     1.37   0.173    -.0528449    .2937165
      ASVABC |   .5281508   .0637492     8.28   0.000     .4029213    .6533803
        MALE |   6.127074    1.69514     3.61   0.000     2.797127    9.457021
       _cons |  -24.32227   13.61504    -1.79   0.075    -51.06777    2.423233
------------------------------------------------------------------------------
. reg EARNINGS WEIGHTkg HEIGHTm ASVABC MALE

      Source |       SS       df       MS              Number of obs =     540
-------------+------------------------------           F(4, 535)     =   30.26
       Model |  21202.8349     4  5300.70874           Prob > F      =  0.0000
    Residual |  93714.1264   535  175.166591           R-squared     =  0.1845
-------------+------------------------------           Adj R-squared =  0.1784
       Total |  114916.961   539  213.20401            Root MSE      =  13.235

------------------------------------------------------------------------------
    EARNINGS |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    WEIGHTkg |  -.0852811   .0364078    -2.34   0.020    -.1568009   -.0137613
     HEIGHTm |   12.04358   8.821012     1.37   0.173     -5.28449    29.37164
      ASVABC |   .5281508   .0637492     8.28   0.000     .4029213    .6533803
        MALE |   6.127075    1.69514     3.61   0.000     2.797127    9.457022
       _cons |  -24.32226   13.61504    -1.79   0.075    -51.06777    2.423239
------------------------------------------------------------------------------
```

3 формы системы нормальных уравнений

1) $Y = X\beta + \varepsilon,$

$X'X\hat{\beta} = X'Y - 1^{\text{ая}}$ форма

2) Уравнения в отклонениях:

$y = Y - \bar{Y}\bar{\imath},$

$x_j = X_j - \bar{X}_j\bar{\imath},\ j = 1, \dots, k$

$y = x\alpha + \varepsilon,$

$x'x\hat{\alpha} = x'y,$ можно переписать (: (n – 1)):

$C\hat{o}v[X]\hat{\alpha} = C\hat{o}v[X, Y] - 2^{\text{ая}}$ форма

Важный результат $\hat{\alpha} = \hat{\beta}_{-0}.$

3 формы системы нормальных уравнений

$\hat{Cov}[X]\hat{\alpha} = \hat{Cov}[X, Y]$ – $2^{\text{ая}}$ форма

Важный результат $\hat{\alpha} = \hat{\beta}_{-0}$.

Доказательство. $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_k X_k$

$\bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{X}_1 + \cdots + \hat{\beta}_k \bar{X}_k$

$\hat{y} = \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k \Rightarrow \ \hat{\alpha} = \hat{\beta}_{-0}$.

3 формы системы нормальных уравнений

3) Уравнения с центрированными и нормированными переменными:

$$\tilde{x}_j = \frac{x_j}{\hat{\sigma}_j},\ j = 1, \dots, k$$

$$\hat{\sigma}_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_{ji} - \bar{X}_j)^2,\ \tilde{y} = \frac{y}{\hat{\sigma}_y},$$

$$\tilde{x}\tilde{x}\hat{\tilde{\beta}} = \tilde{x}\tilde{y} \Leftrightarrow$$

$C\hat{o}r[X]\hat{\tilde{\beta}} = C\hat{o}r[X, Y] - 3^{\text{ья}}$ форма.

Важный результат: коэффициенты 3–ей формы можно сравнивать по абсолютному значению.

Линейная модель и ее интерпретация

$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \varepsilon$,

β_j значим=>

Если X_j увеличится на 1 единицу, то Y увеличится на $\hat{\beta}_j$ единиц (при прочих равных факторах).

Линейная модель. Пример

variable name	storage type	display format	value label	variable label
idcode	int	%8.0g		NLS id
age	byte	%8.0g		age in current year
race	byte	%8.0g	racelbl	race
married	byte	%8.0g	marlbl	married
never_married	byte	%8.0g		never married
grade	byte	%8.0g		current grade completed
collgrad	byte	%16.0g	gradlbl	college graduate
south	byte	%8.0g		lives in south
smsa	byte	%9.0g	smsalbl	lives in SMSA
c_city	byte	%8.0g		lives in central city
industry	byte	%23.0g	indlbl	industry
occupation	byte	%22.0g	occlbl	occupation
union	byte	%8.0g	unionlbl	union worker
wage	float	%9.0g		hourly wage
hours	byte	%8.0g		usual hours worked
ttl_exp	float	%9.0g		total work experience
tenure	float	%9.0g		job tenure (years)

Линейная модель. Пример

```
. sum wage ttl_exp tenure age married union
```

Variable	Obs	Mean	Std. Dev.	Min	Max
wage	2,246	7.766949	5.755523	1.004952	40.74659
ttl_exp	2,246	12.53498	4.610208	.1153846	28.88461
tenure	2,231	5.97785	5.510331	0	25.91667
age	2,246	39.15316	3.060002	34	46
married	2,246	.6420303	.4795099	0	1
union	1,878	.2454739	.4304825	0	1

```
. sum wage ttl_exp tenure age married union if union !=.
```

Variable	Obs	Mean	Std. Dev.	Min	Max
wage	1,878	7.565423	4.168369	1.151368	39.23074
ttl_exp	1,878	12.81837	4.606392	.1153846	28.88461
tenure	1,868	6.571065	5.640675	0	25.91667
age	1,878	39.22471	3.034623	34	46
married	1,878	.6506922	.4768783	0	1
union	1,878	.2454739	.4304825	0	1

Линейная модель. Пример

```
. reg wage ttl_exp tenure  married union

      Source |       SS           df       MS      Number of obs   =     1,868
-------------+----------------------------------   F(4, 1863)      =     82.84
       Model |  4898.98273         4  1224.74568   Prob > F        =    0.0000
    Residual |  27543.1739     1,863  14.7843124   R-squared       =    0.1510
-------------+----------------------------------   Adj R-squared   =    0.1492
       Total |  32442.1567     1,867  17.3766238   Root MSE        =     3.845

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     ttl_exp |   .2871015   .0239244    12.00   0.000       .24018    .3340229
      tenure |   .0467397   .0196266     2.38   0.017     .0082473     .085232
     married |  -.1464481    .187312    -0.78   0.434    -.5138116    .2209155
       union |   1.207615   .2087067     5.79   0.000     .7982918    1.616939
       _cons |   3.389744   .3073016    11.03   0.000     2.787053    3.992436
------------------------------------------------------------------------------
```

Линейная в логарифмах модель

$$\ln Y = \beta_0 + \beta_1 \ln X_1 + \ldots + \beta_j \ln X_j + \beta_k \ln X_k + \varepsilon$$

$$\widehat{\ln Y} = \hat{\beta}_0 + \hat{\beta}_1 \ln X_1 + \ldots + \hat{\beta}_j \ln X_j + \hat{\beta}_k \ln X_k$$

$\frac{\dot{Y}}{Y} = \hat{\beta}_j \frac{\dot{X}_j}{X_j}$, $\hat{\beta}_j = \frac{\dot{Y}/Y}{\dot{X}_j/X_j}$ – эластичность.

β_j значим=>

Если X_j увеличится на 1 %, то Y увеличится на $\hat{\beta}_j$ % (при прочих равных факторах).

Линейная в логарифмах модель. Пример

```
. reg lnwage lnexp tenure  married union

      Source |       SS           df       MS      Number of obs   =     1,868
-------------+----------------------------------   F(4, 1863)      =    128.50
       Model |  107.165264         4  26.791316   Prob > F        =    0.0000
    Residual |  388.422102     1,863  .208492808   R-squared       =    0.2162
-------------+----------------------------------   Adj R-squared   =    0.2146
       Total |  495.587366     1,867  .265445831   Root MSE        =    .45661

------------------------------------------------------------------------------
      lnwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       lnexp |   .3603689   .0249634    14.44   0.000     .3114097    .4093281
      tenure |   .0123301    .002189     5.63   0.000      .008037    .0166231
     married |  -.0001238   .0222199    -0.01   0.996    -.0437023    .0434547
       union |   .1789258   .0247734     7.22   0.000     .1303393    .2275123
       _cons |   .8815045    .059245    14.88   0.000     .7653111     .997698
------------------------------------------------------------------------------
```

Полулогарифмическая модель

$$\ln Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_j X_j + \beta_k X_k + \varepsilon$$

$$\widehat{\ln Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \ldots + \hat{\beta}_j X_j + \hat{\beta}_k X_k$$

$$\frac{\dot{Y}}{Y} = \hat{\beta}_j \dot{X}_j, \qquad \hat{\beta}_j = \frac{\dot{Y}/Y}{\dot{X}_j}$$

β_j значим=>

Если X_j увеличится на 1 единицу, то Y увеличится на

$\hat{\beta}_j * 100$ % (при прочих равных факторах).

Полулогарифмическая модель. Пример

```
. reg lnwage ttl_exp tenure  married union
```

Source	SS	df	MS
Model	105.101394	4	26.2753486
Residual	390.485972	1,863	.209600629
Total	495.587366	1,867	.265445831

Number of obs	=	1,868
F(4, 1863)	=	125.36
Prob > F	=	0.0000
R-squared	=	0.2121
Adj R-squared	=	0.2104
Root MSE	=	.45782

lnwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
ttl_exp	.0400278	.0028486	14.05	0.000	.0344409	.0456146
tenure	.0090427	.0023369	3.87	0.000	.0044595	.0136259
married	.0080798	.0223029	0.36	0.717	-.0356615	.0518211
union	.1847408	.0248503	7.43	0.000	.1360034	.2334782
_cons	1.269176	.0365898	34.69	0.000	1.197415	1.340937

Выбор между моделями

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \varepsilon \quad (1)$$

$$\ln Y = \beta_0 + \beta_1 \ln X_1 + \ldots + \beta_k \ln X_k + \varepsilon \quad (2)$$

$$\ln Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \varepsilon \quad (3)$$

(2) и (3) можно сравнить по R^2_{adj}.

Тест Бокса-Кокса

$$Y^{(\theta)} = \frac{Y^{\theta} - 1}{\theta}, \;\; \theta \neq 0$$

$$\lim_{\theta \mapsto 0} \frac{Y^{\theta} - 1}{\theta} = \lim_{\theta \mapsto 0} \frac{Y^{\theta} lnY}{1} = \ln Y$$

$$Y^{(\theta)} = \begin{cases} \dfrac{Y^{\theta} - 1}{\theta}, & \text{если } \theta \neq 0, \\ \ln Y, & \text{если } \theta = 0 \end{cases}$$

Тест Бокса-Кокса

$$Y^{(\theta)} = \beta_0 + \beta_1 X_1^{(\lambda)} + \ldots + \beta_k X_k^{(\lambda)} + \varepsilon,$$

$\hat{\beta}_{0ML}, \hat{\beta}_{1ML}, \ldots, \hat{\beta}_{kML}, \hat{\lambda}_{ML}, \hat{\theta}_{ML}$ – оценки максимального правдоподобия (тема будет позже)

$H_0: \lambda = \theta = 1$ – линейная модель

$H_0: \lambda = \theta = 0$ – линейная в логарифмах модель

Тест Бокса-Кокса

$$W^{(\theta)} = \beta_0 + \beta_1 H^{(\lambda)} + \varepsilon$$

```
                                   Number of obs   =        540
                                        LR chi2(2)  =     230.68
Log likelihood = -2659.5656            Prob > chi2  =     0.000

------------------------------------------------------------------------------
    WEIGHT02 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
------------------------------------------------------------------------------
     /lambda |   1.055498   1.892654     0.56   0.577    -2.654035     4.76503
      /theta |  -.0263371   .1471576    -0.18   0.858    -.3147607    .2620865
------------------------------------------------------------------------------

---------------------------------------------------------
                       Test           Restricted
   H0:            log likelihood        chi2   Prob > chi2
---------------------------------------------------------
theta=lambda = -1     -2680.8693        42.61     0.000
theta=lambda =  0     -2659.7618         0.39     0.531
theta=lambda =  1     -2685.5201        51.91     0.000
---------------------------------------------------------
```

Выбор между линейной и полулогарифмической моделями

$$Y = \beta_0 + \beta_1 X + \varepsilon \quad (1)$$

$$\ln Y = \beta_0 + \beta_1 X + \varepsilon \quad (2)$$

H_0 : качество подгонки моделей (1) и (2) одинаковое

H_1 : модель с меньшей RSS лучше

Выбор между линейной и полулогарифмической моделью можно осуществить с помощью теста Пола Зарембки (частный случай теста Бокса-Кокса).

Тест П.Зарембки

1) Вычисляется среднее геометрическое значение Y:

$\sqrt[n]{Y_1 \cdot Y_2 \cdot \ldots \cdot Y_n}$,

2) Вводится вспомогательная переменная

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

3) Оцениваются параметры вспомогательных регрессий:

$Y^* = \beta_0' + \beta_1' X + \varepsilon \qquad (3)$

$lnY^* = \beta_0' + \beta_1' X + \varepsilon \quad (4)$

Тест П.Зарембки

4) Вычисляется значение тестовой статистики:

$$\chi^2 = \frac{n}{2}\left|\ln\frac{RSS_3}{RSS_4}\right| \sim \chi^2(1),$$

где RSS_3, RSS_4 – суммы квадратов остатков в оцененных регрессиях (3) и (4).

5) Если при выбранном уровне значимости α Гипотеза H_0 отвергается, то между моделями (1) и (2) есть значимое различие. Лучше та модель, в которой меньше RSS.

ВМ тест для выбора между линейной и полулогарифмической моделями

Bera and McAleer test

H_0: $\ln Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \varepsilon$,

H_1: $Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \varepsilon$

Шаг1: $\widehat{\ln Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \ldots + \hat{\beta}_k X_k$,

$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \ldots + \hat{\beta}_k X_k$

Шаг 2: Оцениваются вспомогательные регрессии:

$\exp(\widehat{\ln Y}) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \nu_1$,

$\ln\hat{Y} = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \nu_2$.

Сохраняются $\hat{\nu}_1, \hat{\nu}_2$.

ВМ тест для выбора между линейной и полулогарифмической моделями

Шаг3: Оцениваются вспомогательные регрессии:

$$\ln Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \theta_1 \hat{v}_1 + \varepsilon_1,$$

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \theta_2 \hat{v}_2 + \varepsilon_2$$

Обычные $t - tests$ для проверки значимости θ_1 и θ_2.

Если $\theta_1 = 0$ не отвергается, а $\theta_2 = 0$ отвергается, то выбирается полулогарифмическая модель.

Если $\theta_2 = 0$неотвергается, а $\theta_1 = 0$ отвергается, то выбирается линейная модель.

Возникает проблема если обе гипотезы отвергаются или не отвергаются.

PE тест для выбора между линейной и полулогарифмической моделями

MacKinnon-White-Davidson PE test

H_0: $\ln Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \varepsilon,$
H_1: $Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \varepsilon$

Шаг1: $\widehat{\ln Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \ldots + \hat{\beta}_k X_k,$
$\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 X_2 + \ldots + \hat{\beta}_k X_k$
Шаг2: Оцениваются вспомогательные регрессии:
$\ln Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \theta_1[\hat{Y} - \exp(\widehat{\ln Y})] + \varepsilon_1,$
$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \theta_2[\widehat{\ln Y} - \ln \hat{Y}] + \varepsilon_2$

PE тест для выбора между линейной и полулогарифмической моделями

$\ln Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \theta_1[\hat{Y} - \exp(\widehat{\ln Y})] + \varepsilon_1,$

$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \theta_2[\widehat{\ln Y} - \ln\hat{Y}\,] + \varepsilon_2$

Обычные $\mathrm{t} - tests$ для проверки значимости θ_1 и θ_2.

Если $\theta_1 = 0$ не отвергается, a $\theta_2 = 0$ отвергается, то выбирается полулогарифмическая модель.

Если $\theta_2 = 0$ не отвергается, а $\theta_1 = 0$ отвергается, то выбирается линейная модель.

Возникает проблема, если обе гипотезы отвергаются или не отвергаются.

Thank you for your attention!

20, Myasnitskaya str., Moscow, Russia, 101000

Tel.: +7 (495) 628-8829, Fax: +7 (495) 628-7931

www.hse.ru