Методический семинар 12.04.2012, часть 1

Смещение в оценках коэффициентов, вызванное невключением существенных переменных

Демидова О.А.

Ошибки спецификации I: невключение существенной переменной

		Истинная модель	
		$Y = \beta_1 + \beta_2 X_2 + u$	$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u$
Оцененная модель	$\hat{Y} = b_1 + b_2 X_2$		
	$\hat{Y} = b_1 + b_2 X_2 + b_3 X_3$		

Ошибки спецификации I: невключение существенной переменной

		Истинная модель	
		$Y = \beta_1 + \beta_2 X_2 + u$	$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u$
Оцененная модель	$\hat{Y} = b_1 + b_2 X_2$	**Правильная спецификация, все в порядке**	
	$\hat{Y} = b_1 + b_2 X_2 + b_3 X_3$		

Ошибки спецификации I: невключение существенной переменной

		Истинная модель	
		$Y = \beta_1 + \beta_2 X_2 + u$	$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u$
Оцененная модель	$\hat{Y} = b_1 + b_2 X_2$	**Правильная спецификация, все в порядке**	
	$\hat{Y} = b_1 + b_2 X_2 + b_3 X_3$		**Правильная спецификация, все в порядке**

Ошибки спецификации I: невключение существенной переменной

		Истинная модель	
		$Y = \beta_1 + \beta_2 X_2 + u$	$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u$
Оцененная модель	$\hat{Y} = b_1 + b_2 X_2$		**Оценки коэффициентов будут смещенными**
	$\hat{Y} = b_1 + b_2 X_2 + b_3 X_3$		

Ошибки спецификации I: невключение существенной переменной

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u \qquad \hat{Y} = b_1 + b_2 X_2$$

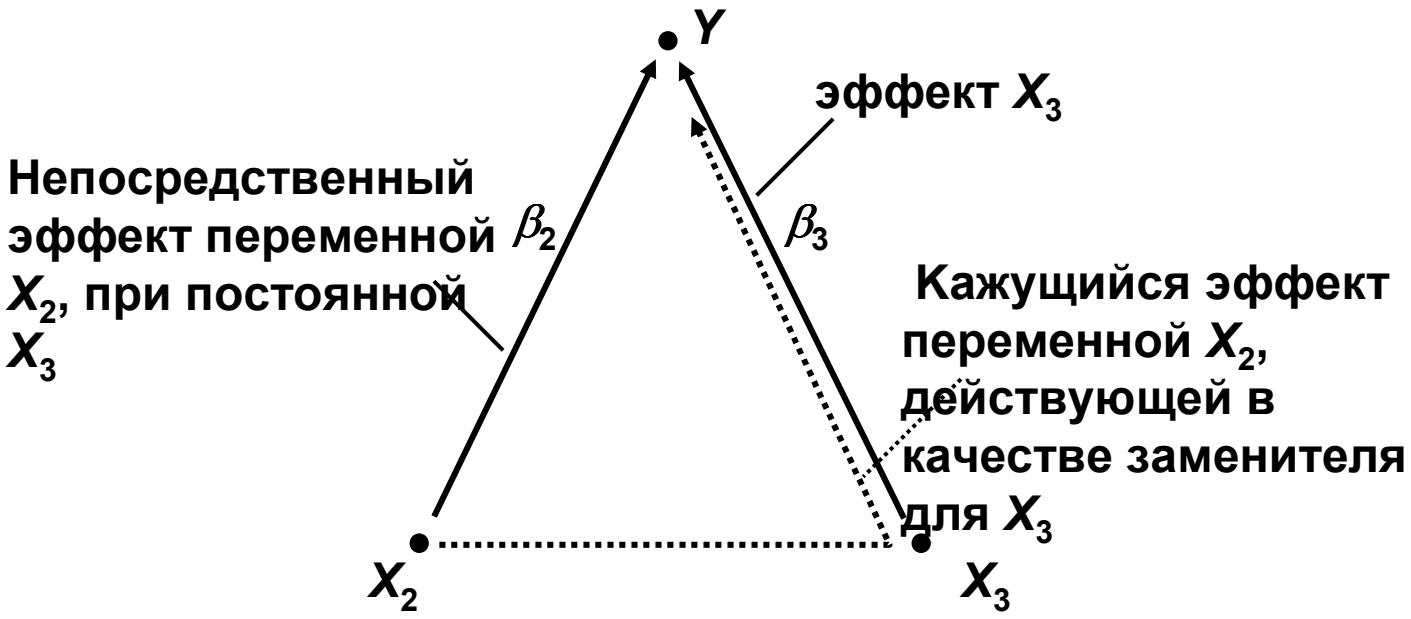
$$E(b_2) = \beta_2 + \beta_3 \frac{\sum (X_{2i} - \bar{X}_2)(X_{3i} - \bar{X}_3)}{\sum (X_{2i} - \bar{X}_2)^2}$$

Формула для смещения выделена желтым цветом.

Ошибки спецификации I: невключение существенной переменной

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u \qquad \hat{Y} = b_1 + b_2 X_2$$

$$E(b_2) = \beta_2 + \beta_3 \frac{\sum (X_{2i} - \overline{X}_2)(X_{3i} - \overline{X}_3)}{\sum (X_{2i} - \overline{X}_2)^2}$$



Вывод формулы для смещения

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u \qquad \hat{Y} = b_1 + b_2 X_2$$

$$\begin{aligned} Y_i - \bar{Y} &= \left(\beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i\right) - \left(\beta_1 + \beta_2 \bar{X}_2 + \beta_3 \bar{X}_3 + \bar{u}\right) \\ &= \beta_2 \left(X_{2i} - \bar{X}_2\right) + \beta_3 \left(X_{3i} - \bar{X}_3\right) + u_i - \bar{u} \end{aligned}$$

$$b_2 = \frac{\sum \left(X_{2i} - \bar{X}_2\right)\left(Y_i - \bar{Y}\right)}{\sum \left(X_{2i} - \bar{X}_2\right)^2}$$

$$= \frac{\sum \left[\beta_2 \left(X_{2i} - \bar{X}_2\right)^2 + \beta_3 \left(X_{2i} - \bar{X}_2\right)\left(X_{3i} - \bar{X}_3\right) + \left(X_{2i} - \bar{X}_2\right)\left(u_i - \bar{u}\right)\right]}{\sum \left(X_{2i} - \bar{X}_2\right)^2}$$

$$= \beta_2 + \beta_3 \frac{\sum \left(X_{2i} - \bar{X}_2\right)\left(X_{3i} - \bar{X}_3\right)}{\sum \left(X_{2i} - \bar{X}_2\right)^2} + \frac{\sum \left(X_{2i} - \bar{X}_2\right)\left(u_i - \bar{u}\right)}{\sum \left(X_{2i} - \bar{X}_2\right)^2}$$

Вывод формулы для смещения

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u \qquad \hat{Y} = b_1 + b_2 X_2$$

$$b_2 = \beta_2 + \beta_3 \frac{\sum (X_{2i} - \bar{X}_2)(X_{3i} - \bar{X}_3)}{\sum (X_{2i} - \bar{X}_2)^2} + \frac{\sum (X_{2i} - \bar{X}_2)(u_i - \bar{u})}{\sum (X_{2i} - \bar{X}_2)^2}$$

$$E(b_2) = \beta_2 + \beta_3 \frac{\sum (X_{2i} - \bar{X}_2)(X_{3i} - \bar{X}_3)}{\sum (X_{2i} - \bar{X}_2)^2} + E\left(\frac{\sum (X_{2i} - \bar{X}_2)(u_i - \bar{u})}{\sum (X_{2i} - \bar{X}_2)^2} \right)$$

Вывод формулы для смещения

$$E(b_2) = \beta_2 + \beta_3 \frac{\sum (X_{2i} - \bar{X}_2)(X_{3i} - \bar{X}_3)}{\sum (X_{2i} - \bar{X}_2)^2} + E\left(\frac{\sum (X_{2i} - \bar{X}_2)(u_i - \bar{u})}{\sum (X_{2i} - \bar{X}_2)^2} \right)$$

$$E\left(\frac{\sum (X_{2i} - \bar{X}_2)(u_i - \bar{u})}{\sum (X_{2i} - \bar{X}_2)^2} \right) = \frac{1}{\sum (X_{2i} - \bar{X}_2)^2} E\left(\sum (X_{2i} - \bar{X}_2)(u_i - \bar{u}) \right)$$

$$= \frac{1}{\sum (X_{2i} - \bar{X}_2)^2} \sum E\{(X_{2i} - \bar{X}_2)(u_i - \bar{u})\}$$

$$= \frac{1}{\sum (X_{2i} - \bar{X}_2)^2} \sum (X_{2i} - \bar{X}_2) E(u_i - \bar{u})$$

$$= 0$$

Вывод формулы для смещения

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u \qquad \hat{Y} = b_1 + b_2 X_2$$

$$E(b_2) = \beta_2 + \beta_3 \frac{\sum (X_{2i} - \bar{X}_2)(X_{3i} - \bar{X}_3)}{\sum (X_{2i} - \bar{X}_2)^2} + E\left(\frac{\sum (X_{2i} - \bar{X}_2)(u_i - \bar{u})}{\sum (X_{2i} - \bar{X}_2)^2} \right)$$

$$E(b_2) = \beta_2 + \beta_3 \frac{\sum (X_{2i} - \bar{X}_2)(X_{3i} - \bar{X}_3)}{\sum (X_{2i} - \bar{X}_2)^2}$$

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u \qquad \hat{Y} = b_1 + b_2 X_2$$

$$E(b_2) = \beta_2 + \beta_3 \frac{\sum (X_{2i} - \bar{X}_2)(X_{3i} - \bar{X}_3)}{\sum (X_{2i} - \bar{X}_2)^2} + E\left(\frac{\sum (X_{2i} - \bar{X}_2)(u_i - \bar{u})}{\sum (X_{2i} - \bar{X}_2)^2} \right)$$

$$E(b_2) = \beta_2 + \beta_3 \frac{\sum (X_{2i} - \bar{X}_2)(X_{3i} - \bar{X}_3)}{\sum (X_{2i} - \bar{X}_2)^2}$$

Оценки стандартных отклонений при невключении существенной переменной тоже являются смещенными, t и F – статистики рассчитываются неправильно.

Пример 1

$$S = \beta_1 + \beta_2 ASVABC + \beta_3 SM + u$$

```
. reg S ASVABC SM

      Source |       SS       df       MS              Number of obs =     540
-------------+------------------------------           F(  2,   537) =  147.36
       Model |  1135.67473     2  567.837363            Prob > F      =  0.0000
    Residual |  2069.30861   537  3.85346109            R-squared     =  0.3543
-------------+------------------------------           Adj R-squared =  0.3519
       Total |  3204.98333   539  5.94616574            Root MSE      =   1.963

------------------------------------------------------------------------------
           S |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      ASVABC |   .1328069   .0097389    13.64   0.000     .1136758     .151938
          SM |   .1235071   .0330837     3.73   0.000     .0585178    .1884963
       _cons |   5.420733   .4930224    10.99   0.000     4.452244    6.389222
------------------------------------------------------------------------------
```

Пример 1

$$S = \beta_1 + \beta_2 ASVABC + \beta_3 SM + u$$

$$E(b_2) = \beta_2 + \beta_3 \frac{\sum \left(ASVABC_i - \overline{ASVABC}\right)\left(SM_i - \overline{SM}\right)}{\sum \left(ASVABC_i - \overline{ASVABC}\right)^2}$$

```
. reg S ASVABC SM

      Source |       SS       df       MS              Number of obs =     540
-------------+------------------------------           F(  2,   537) =  147.36
       Model |  1135.67473     2  567.837363            Prob > F      =  0.0000
    Residual |  2069.30861   537  3.85346109            R-squared     =  0.3543
-------------+------------------------------           Adj R-squared =  0.3519
       Total |  3204.98333   539  5.94616574            Root MSE      =   1.963

------------------------------------------------------------------------------
           S |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      ASVABC |   .1328069   .0097389    13.64   0.000     .1136758     .151938
          SM |   .1235071   .0330837     3.73   0.000     .0585178    .1884963
       _cons |   5.420733   .4930224    10.99   0.000     4.452244    6.389222
------------------------------------------------------------------------------
```

Знак произведения зависит от двух множителей.

Пример 1

$$S = \beta_1 + \beta_2 ASVABC + \beta_3 SM + u$$

```
. reg S ASVABC SM
```

```
      Source |       SS       df       MS              Number of obs =     540
-------------+------------------------------           F(  2,   537) =  147.36
       Model |  1135.67473     2  567.837363            Prob > F      =  0.0000
    Residual |  2069.30861   537  3.85346109            R-squared     =  0.3543
-------------+------------------------------           Adj R-squared =  0.3519
       Total |  3204.98333   539  5.94616574            Root MSE      =   1.963

------------------------------------------------------------------------------
           S |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      ASVABC |   .1328069   .0097389    13.64   0.000     .1136758     .151938
          SM |   .1235071   .0330837     3.73   0.000     .0585178    .1884963
       _cons |   5.420733   .4930224    10.99   0.000     4.452244    6.389222
------------------------------------------------------------------------------
```

$$E(b_2) = \beta_2 + \beta_3 \frac{\sum \left(ASVABC_i - \overline{ASVABC}\right)\left(SM_i - \overline{SM}\right)}{\sum \left(ASVABC_i - \overline{ASVABC}\right)^2}$$

Оценка коэффициента β_3 положительна.

Пример 1

$$S = \beta_1 + \beta_2 ASVABC + \beta_3 SM + u$$

```
. reg S ASVABC SM
```

Source	SS	df	MS
Model	1135.67473	2	567.837363
Residual	2069.30861	537	3.85346109
Total	3204.98333	539	5.94616574

```
. cor SM ASVABC
(obs=540)
```

	SM	ASVABC
SM	1.0000	
ASVABC	0.4202	1.0000

S	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
ASVABC	.1328069	.0097389	13.64	0.000	.1136758	.151938
SM	.1235071	.0330837	3.73	0.000	.0585178	.1884963
_cons	5.420733	.4930224	10.99	0.000	4.452244	6.389222

$$E(b_2) = \beta_2 + \beta_3 \frac{\sum \left(ASVABC_i - \overline{ASVABC}\right)\left(SM_i - \overline{SM}\right)}{\sum \left(ASVABC_i - \overline{ASVABC}\right)^2}$$

Знак второго множителя в формуле для смещения совпадает со знаком коэффициента корреляции включенного и пропущенного фактора.

Пример 1

```
. reg S ASVABC SM
```

$$S = \beta_1 + \beta_2 ASVABC + \beta_3 SM + u$$

```
------------------------------------------------------------------------------
           S |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      ASVABC |   .1328069   .0097389    13.64   0.000     .1136758     .151938
          SM |   .1235071   .0330837     3.73   0.000     .0585178    .1884963
       _cons |   5.420733   .4930224    10.99   0.000     4.452244    6.389222
------------------------------------------------------------------------------
```

```
. reg S ASVABC
```

```
------------------------------------------------------------------------------
           S |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      ASVABC |    .148084   .0089431    16.56   0.000     .1305165    .1656516
       _cons |   6.066225   .4672261    12.98   0.000     5.148413    6.984036
------------------------------------------------------------------------------
```

Таким образом, знак смещения – положительный, что и наблюдается в реальности.

Пример 1

$$S = \beta_1 + \beta_2 ASVABC + \beta_3 SM + u$$

```
. reg S SM
```

```
      Source |       SS       df       MS              Number of obs =     540
-------------+------------------------------           F(  1,   538) =   80.93
       Model |  419.086251     1  419.086251           Prob > F      =  0.0000
    Residual |  2785.89708   538  5.17824736           R-squared     =  0.1308
-------------+------------------------------           Adj R-squared =  0.1291
       Total |  3204.98333   539  5.94616574           Root MSE      =  2.2756

------------------------------------------------------------------------------
           S |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          SM |   .3130793   .0348012     9.00   0.000     .2447165    .3814422
       _cons |   10.04688   .4147121    24.23   0.000     9.232226    10.86153
------------------------------------------------------------------------------
```

$$E(b_3) = \beta_3 + \beta_2 \frac{\sum \left(ASVABC_i - \overline{ASVABC}\right)\left(SM_i - \overline{SM}\right)}{\sum \left(SM_i - \overline{SM}\right)^2}$$

Предположим, что в уравнение регрессии не будет включена переменная *ASVABC.* Тогда коэффициент при переменной *SM* будет смещен. Как и в предыдущем случае, можно показать, что это смещение будет положительным, что и наблюдается.

Пример 2

$$LGEARN = \beta_1 + \beta_2 S + \beta_3 EXP + u$$

```
.reg LGEARN S EXP

      Source |       SS       df       MS              Number of obs =     540
-------------+------------------------------           F(  2,   537) =  100.86
       Model |  50.9842581     2   25.492129           Prob > F      =  0.0000
    Residual |  135.723385   537  .252743734           R-squared     =  0.2731
-------------+------------------------------           Adj R-squared =  0.2704
       Total |  186.707643   539   .34639637           Root MSE      =  .50274

------------------------------------------------------------------------------
      LGEARN |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           S |   .1235911   .0090989    13.58   0.000     .1057173     .141465
         EXP |   .0350826   .0050046     7.01   0.000     .0252515    .0449137
       _cons |   .5093196   .1663823     3.06   0.002     .1824796    .8361596
------------------------------------------------------------------------------
```

Пример 2

$$LGEARN = \beta_1 + \beta_2 S + \beta_3 EXP + u$$

```
. reg LGEARN S EXP

      Source |       SS       df       MS
-------------+------------------------------
       Model |  50.9842581     2   25.492129
    Residual |  135.723385   537  .252743734
-------------+------------------------------
       Total |  186.707643   539   .34639637

------------------------------------------------------------------------------
      LGEARN |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           S |   .1235911   .0090989    13.58   0.000     .1057173     .141465
         EXP |   .0350826   .0050046     7.01   0.000     .0252515    .0449137
       _cons |   .5093196   .1663823     3.06   0.002     .1824796    .8361596
------------------------------------------------------------------------------
```

```
. cor S EXP
(obs=540)

             |        S      EXP
-------------+------------------
           S|   1.0000
         EXP|  -0.2179   1.0000
```

$$E(b_2) = \beta_2 + \beta_3 \frac{\sum (S_i - \bar{S})(EXP_i - \overline{EXP})}{\sum (S_i - \bar{S})^2}$$

Если опущена переменная *EXP*, то смещение коэффициента перед переменной S будет отрицательным, т.к. оценка коэффициента β_2 положительная, а коэффициент корреляции S и EXP отрицательный.

Пример 2

$$LGEARN = \beta_1 + \beta_2 S + \beta_3 EXP + u$$

```
reg LGEARN S EXP
```

Source	SS	df	MS
Model	50.9842581	2	25.492129
Residual	135.723385	537	.252743734
Total	186.707643	539	.34639637

Adj R-squared = 0.2704
Root MSE = .50274

LGEARN	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
S	.1235911	.0090989	13.58	0.000	.1057173	.141465
EXP	.0350826	.0050046	7.01	0.000	.0252515	.0449137
_cons	.5093196	.1663823	3.06	0.002	.1824796	.8361596

```
. cor S EXP
(obs=540)
```

	S	EXP
S	1.0000	
EXP	-0.2179	1.0000

$$E(b_3) = \beta_3 + \beta_2 \frac{\sum \left(EXP_i - \overline{EXP}\right)\left(S_i - \overline{S}\right)}{\sum \left(EXP_i - \overline{EXP}\right)^2}$$

Аналогично, если опущена переменная S, то оценка коэффициента перед переменной *EXP* будет смещена вниз.

Пример 2

```
. reg LGEARN S EXP
------------------------------------------------------------------------------
      LGEARN |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           S |    .1235911   .0090989    13.58   0.000     .1057173     .141465
         EXP |    .0350826   .0050046     7.01   0.000     .0252515    .0449137
       _cons |    .5093196   .1663823     3.06   0.002     .1824796    .8361596

. reg LGEARN S
------------------------------------------------------------------------------
      LGEARN |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           S |    .1096934   .0092691    11.83   0.000     .0914853    .1279014
       _cons |    1.292241   .1287252    10.04   0.000     1.039376    1.545107

. reg LGEARN EXP
------------------------------------------------------------------------------
      LGEARN |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         EXP |    .0202708   .0056564     3.58   0.000     .0091595     .031382
       _cons |     2.44941   .0988233    24.79   0.000     2.255284    2.643537
```

Смещение в случае невключения одной из переменных S или EXP действительно является отрицательным.