

Лекция по эконометрике № 3, 2 модуль

Ошибки спецификации модели

Демидова
Ольга Анатольевна
https://www.hse.ru/staff/demidova_olga
E-mail: demidova@hse.ru
18.11.2019

План лекции № 3, 2 модуль

- Смещение в оценках коэффициентов, вызванное невключением существенных переменных
- Уменьшение эффективности оценок коэффициентов при включении в модель излишних переменных
- RESET- тест Рамсея для проверки гипотезы о существовании упущенных переменных

Ошибки спецификации I: невключение существенной переменной

		<i>Истинная модель</i>	
		$Y = \beta_0 + \beta_1 X_1 + \varepsilon$	$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$
<i>Оцененная модель</i>	$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1$		
	$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$		

Ошибки спецификации I: невключение существенной переменной

		<i>Истинная модель</i>	
		$Y = \beta_0 + \beta_1 X_1 + \varepsilon$	$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$
<i>Оцененная модель</i>	$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1$	Правильная спецификация, все в порядке	
	$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$		Правильная спецификация, все в порядке

Ошибки спецификации I: невключение существенной переменной

		<i>Истинная модель</i>	
		$Y = \beta_0 + \beta_1 X_1 + \varepsilon$	$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$
<i>Оцененная модель</i>	$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1$		Оценки коэффициентов будут смещенными
	$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$		

Ошибки спецификации I: невключение существенной переменной

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon \quad \hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1$$

$$E(\hat{\beta}_1) = \beta_1 + \beta_2 \frac{\sum (X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2)}{\sum (X_{1i} - \bar{X}_1)^2}$$

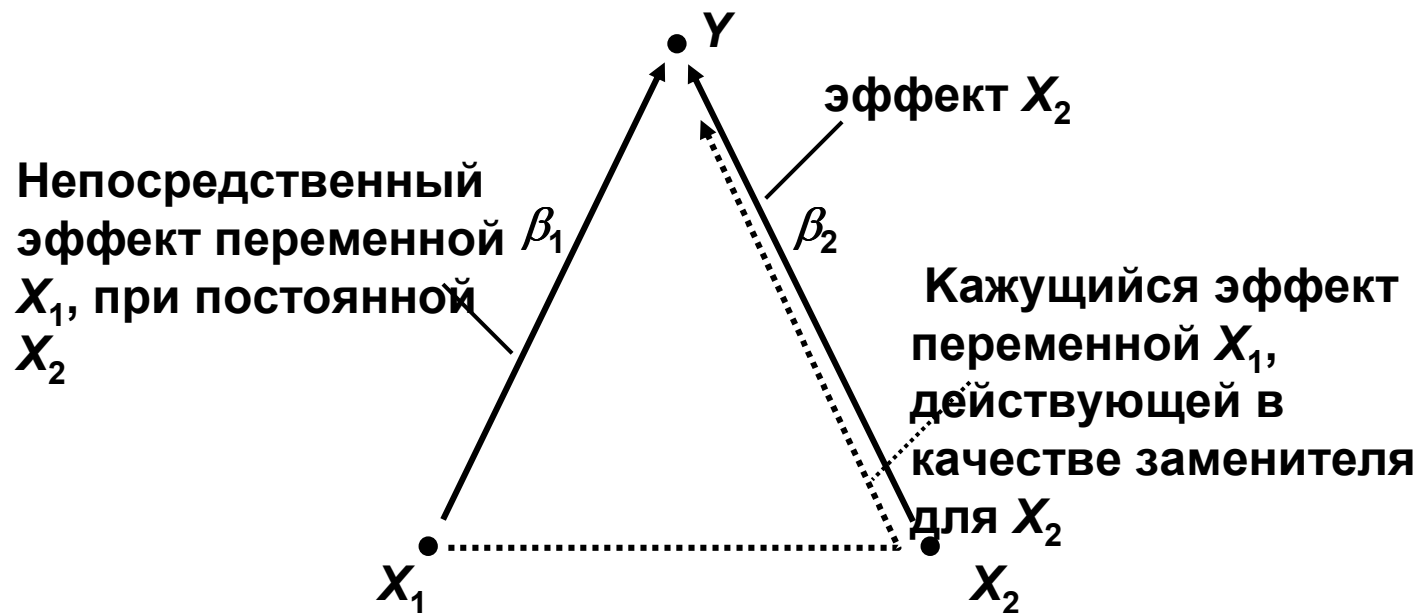
Формула для смещения выделена желтым цветом.

Ошибки спецификации I: невключение существенной переменной

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

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

$$E(\hat{\beta}_1) = \beta_1 + \beta_2 \frac{\sum (X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2)}{\sum (X_{1i} - \bar{X}_1)^2}$$



Вывод формулы для смещения

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

$$\begin{aligned} E\left(\frac{\sum (X_{1i} - \bar{X}_1)(\varepsilon_i - \bar{\varepsilon})}{\sum (X_{1i} - \bar{X}_1)^2}\right) &= \frac{1}{\sum (X_{1i} - \bar{X}_1)^2} E\left(\sum (X_{1i} - \bar{X}_1)(\varepsilon_i - \bar{\varepsilon})\right) \\ &= \frac{1}{\sum (X_{1i} - \bar{X}_1)^2} \sum E\{(X_{1i} - \bar{X}_1)(\varepsilon_i - \bar{\varepsilon})\} \\ &= \frac{1}{\sum (X_{1i} - \bar{X}_1)^2} \sum (X_{1i} - \bar{X}_1) E(\varepsilon_i - \bar{\varepsilon}) \\ &= 0 \end{aligned}$$

Пример 1

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

```
. reg S ASVABC SM
```

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

$$S = \beta_0 + \beta_1 ASVABC + \beta_2 SM + \varepsilon$$

$$E(\hat{\beta}_1) = \beta_1 + \beta_2 \frac{\sum (ASVABC_i - \overline{ASVABC})(SM_i - \overline{SM})}{\sum (ASVABC_i - \overline{ASVABC})^2}$$

```
. reg S ASVABC SM
```

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

```
. reg S ASVABC SM
```

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

$$E(\hat{\beta}_1) = \beta_1 + \beta_2 \frac{\sum (ASVABC_i - \overline{ASVABC})(SM_i - \overline{SM})}{\sum (ASVABC_i - \overline{ASVABC})^2}$$

Оценка коэффициента β_2 положительна.

Пример 1

```
. 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

$$E(\hat{\beta}_1) = \beta_1 + \beta_2 \frac{\sum (ASVABC_i - \overline{ASVABC})(SM_i - \overline{SM})}{\sum (ASVABC_i - \overline{ASVABC})^2}$$

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

```
. reg S SM
```

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

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

$$LGEARN = \beta_0 + \beta_1 S + \beta_2 EXP + \varepsilon$$

```
.reg LGEARN S EXP
```

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

Пример 2

```
. reg LG EARN S EXP
```

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

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

	S	EXP
S	1.0000	
EXP	-0.2179	1.0000

LG EARN	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

$$E(\hat{\beta}_1) = \beta_1 + \beta_2 \frac{\sum (S_i - \bar{S})(EXP_i - \overline{EXP})}{\sum (S_i - \bar{S})^2}$$

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

$$LGEARN = \beta_0 + \beta_1 S + \beta_2 EXP + \varepsilon$$

```
reg LGEARN S EXP
```

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

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

	S	EXP
S	1.0000	
EXP	-0.2179	1.0000

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

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.

Ошибки спецификации II: включение излишней переменной

		<i>Истинная модель</i>	
		$Y = \beta_0 + \beta_1 X_1 + \varepsilon$	$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$
Оцененная модель	$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1$	Правильная спецификация	Оценки коэффициентов являются смещенными
	$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$	Оценки коэффициентов являются несмещенными, но не эффективными	Правильная спецификация

Ошибки спецификации II: включение излишней переменной

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$$

$$Y = \beta_0 + \beta_1 X_1 + 0 X_2 + \varepsilon$$

Ошибки спецификации II: включение излишней переменной

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$$

$$Y = \beta_0 + \beta_1 X_1 + 0X_2 + \varepsilon$$

$$\sigma_{\hat{\beta}_1}^2 = \frac{\sigma_{\varepsilon}^2}{\sum (X_{1i} - \bar{X}_1)^2} \times \frac{1}{1 - r_{X_1, X_2}^2}$$

При включении излишней переменной X_2 увеличивается оценка дисперсии коэффициента перед переменной X_1 . Добавляется множитель $1 / (1 - r^2)$, где r – коэффициент корреляции между X_1 и X_2 .

Пример

```
. reg LGFDHO LGEXP LGSIZE
```

Source	SS	df	MS	Number of obs = 868		
Model	138.776549	2	69.3882747	F(2, 865)	=	460.92
Residual	130.219231	865	.150542464	Prob > F	=	0.0000
Total	268.995781	867	.310260416	R-squared	=	0.5159
				Adj R-squared	=	0.5148
				Root MSE	=	.388

LGFDHO	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
LGEXP	.2866813	.0226824	12.639	0.000	.2421622	.3312003
LGSIZE	.4854698	.0255476	19.003	0.000	.4353272	.5356124
_cons	4.720269	.2209996	21.359	0.000	4.286511	5.154027

LGFDHO – логарифм ежегодных расходов домохозяйств на продукты домашнего потребления, **LGEXP** – логарифм общих годовых расходов домохозяйств, **LGSIZE** – логарифм числа потребителей в домохозяйстве.

Пример

```
. reg LGFDHO LGEXP LGSIZE
```

Source	SS	df	MS	Number of obs = 868		
Model	138.776549	2	69.3882747	F(2, 865)	=	460.92
Residual	130.219231	865	.150542464	Prob > F	=	0.0000
Total	268.995781	867	.310260416	R-squared	=	0.5159
				Adj R-squared	=	0.5148
				Root MSE	=	.388

LGFDHO	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
LGEXP	.2866813	.0226824	12.639	0.000	.2421622	.3312003
LGSIZE	.4854698	.0255476	19.003	0.000	.4353272	.5356124
_cons	4.720269	.2209996	21.359	0.000	4.286511	5.154027

Данные 1995 г. US Consumer Expenditure Survey для 868 домохозяйств.

Пример

```
. reg LGFDHO LGEXP LGSIZE LGHOUS
```

Source	SS	df	MS	Number of obs = 868		
Model	138.841976	3	46.2806586	F(3, 864)	=	307.22
Residual	130.153805	864	.150640978	Prob > F	=	0.0000
Total	268.995781	867	.310260416	R-squared	=	0.5161
				Adj R-squared	=	0.5145
				Root MSE	=	.38812

LGFDHO	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
LGEXP	.2673552	.0370782	7.211	0.000	.1945813	.340129
LSIZE	.4868228	.0256383	18.988	0.000	.4365021	.5371434
LGHOUS	.0229611	.0348408	0.659	0.510	-.0454214	.0913436
_cons	4.708772	.2217592	21.234	0.000	4.273522	5.144022

LGHOUS – логарифм годовых расходов на жилье.

Пример

```
. reg LGFDHO LGEXP LGSIZE LGHOUS
```

Source	SS	df	MS
Model	138.841976	3	46.2806586
Residual	130.153805	864	.150640978
Total	268.995781	867	.310260416

LGFDHO	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
LGEXP	.2673552	.0370782	7.211	0.000	.1945813	.340129
LSIZE	.4868228	.0256383	18.988	0.000	.4365021	.5371434
LGHOU	.0229611	.0348408	0.659	0.510	-.0454214	.0913436
_cons	4.708772	.2217592	21.234	0.000	4.273522	5.144022

```
. cor LGHOUS LGEXP LGSIZE
(obs=869)
```

	LGHOU	LGEXP	LSIZE
LGHOU	1.0000		
LGEXP	0.8137	1.0000	
LSIZE	0.3256	0.4491	1.0000

Пример

```
. reg LGFDHO LGEXP LGSIZE
```

LGFDHO	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
LGEXP	.2866813	.0226824	12.639	0.000	.2421622	.3312003
LGSIZE	.4854698	.0255476	19.003	0.000	.4353272	.5356124
_cons	4.720269	.2209996	21.359	0.000	4.286511	5.154027

```
. reg LGFDHO LGEXP LGSIZE LGHOUS
```

LGFDHO	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
LGEXP	.2673552	.0370782	7.211	0.000	.1945813	.340129
LGSIZE	.4868228	.0256383	18.988	0.000	.4365021	.5371434
LGHOUS	.0229611	.0348408	0.659	0.510	-.0454214	.0913436
_cons	4.708772	.2217592	21.234	0.000	4.273522	5.144022

RESET – regression specification error test.

RESET – тест Рамсея отвечает на вопрос, надо ли включать в регрессию степени независимых переменных.

RESET – тест Рамсея

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k + u \quad (*)$$

H_0 : спецификация модели (*) является правильной

H_1 : спецификация модели (*) является неправильной,

RESET – тест Рамсея

Проведение RESET – теста Рамсея

- 1) Оцениваем коэффициенты функции регрессии (*)

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

- 2) Сохраняем столбец оцененных значений $\hat{Y}$

- 3) Оцениваем коэффициенты вспомогательной регрессии

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k + \alpha_2 \hat{Y}^2 + \dots + \alpha_m \hat{Y}^m + \varepsilon$$

4) Тогда проверка гипотезы о правильной спецификации равносильна проверке гипотезы

$$H_0: \alpha_2 = \dots = \alpha_m = 0$$

$$H_1: \exists \alpha_i \neq 0, i = 2, \dots, m$$

5) Вычисляем значение тестовой статистики

$$F = \frac{(RSS_R - RSS_{UR}) / (m - 1)}{RSS_{UR} / (n - (k + 1 + m - 1))}$$

6) Если $F > F_{\alpha}^{cr}(m-1, n-(k+m))$

то гипотеза H_0 отвергается.

Пример

reg earnings age educ sex

Source	SS	df	MS	Number of obs = 4893		
-----+-----				F(3, 4889) = 168.49		
Model	1.8575e+10	3	6.1918e+09	Prob > F = 0.0000		
Residual	1.7966e+11	4889	36747469.4	R-squared = 0.0937		
-----+-----				Adj R-squared = 0.0931		
Total	1.9823e+11	4892	40522005.4	Root MSE = 6062		

earnings	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
-----+-----						
age	-14.84304	7.18472	-2.07	0.039	-28.92832	-.7577593
educ	388.4089	25.7661	15.07	0.000	337.8958	438.9221
sex	-3307.229	176.4169	-18.75	0.000	-3653.085	-2961.372
_cons	5645.355	551.2111	10.24	0.000	4564.733	6725.976
-----+-----						

Пример

ovtest

**Ramsey RESET test using powers of the fitted
values of earnings**

Ho: model has no omitted variables

$F(3, 4886) = 4.90$

$\text{Prob} > F = 0.0021$

Пример

reg earnings age agesq educ sex

Source	SS	df	MS	Number of obs	= 4893
				F(4, 4888)	= 138.61
Model	2.0195e+10	4	5.0487e+09	Prob > F	= 0.0000
Residual	1.7804e+11	4888	36423634.2	R-squared	= 0.1019
				Adj R-squared	= 0.1011
Total	1.9823e+11	4892	40522005.4	Root MSE	= 6035.2

earnings	Coef.	Std. Err.	t	P>t	[95% Conf.	Interval]
age	272.1451	43.62763	6.24	0.000	186.6153	357.6748
agesq	-3.501215	.5250484	-6.67	0.000	-4.530546	-2.471884
educ	376.8115	25.7112	14.66	0.000	326.406	427.217
sex	-3343.731	175.7231	-19.0	0.000	-3688.228	-2999.235
_cons	542.6265	941.6523	0.58	0.564	-1303.435	2388.688