

## **F-тест точности подбора для всего уравнения**

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

$$H_0 : \beta_2 = \dots = \beta_k = 0$$

$$H_1 : \text{at least one } \beta \neq 0$$

Эта последовательность описывает два F-теста по точности подбора с множественной регрессией. Первый относится к точности подбора уравнения в целом.

## **F-тест точности подбора для всего уравнения**

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

$$H_0 : \beta_2 = \dots = \beta_k = 0$$

$$H_1 : \text{at least one } \beta \neq 0$$

Рассмотрим общий случай, когда имеются  $k - 1$  пояснительных переменных. Для F-критерия точности подбора уравнения в целом нулевая гипотеза, она состоит в том, что модель вообще не имеет объясняющей способности.

## **F-тест точности подбора для всего уравнения**

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

$$H_0 : \beta_2 = \dots = \beta_k = 0$$

$$H_1 : \text{at least one } \beta \neq 0$$

Конечно, мы надеемся опровергнуть это и сделать вывод, что модель имеет некоторую объяснительную силу.

## **F-тест точности подбора для всего уравнения**

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

$$H_0 : \beta_2 = \dots = \beta_k = 0$$

$$H_1 : \text{at least one } \beta \neq 0$$

**Модель не будет иметь объясняющей силы, если окажется, что  $Y$  не связано ни с одной из объясняющих переменных. Поэтому математически нулевая гипотеза состоит в том, что все коэффициенты  $\beta_2, \dots, \beta_k$  равны нулю.  $\beta_2, \dots, \beta_k$ .**

## **F-тест точности подбора для всего уравнения**

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

$$H_0 : \beta_2 = \dots = \beta_k = 0$$

$$H_1 : \text{at least one } \beta \neq 0$$

**Альтернативная гипотеза состоит в том, что хотя бы один из этих  $\beta$  коэффициентов отличен от нуля.**

## **F-тест точности подбора для всего уравнения**

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

$$H_0 : \beta_2 = \dots = \beta_k = 0$$

$$H_1 : \text{at least one } \beta \neq 0$$

**В модели множественной регрессии существует разница между ролями тестов F и t. Тест F проверяет общую объясняющую силу переменных, в то время как t-тесты проверяют их объясняющую силу отдельно.**

## **F-тест точности подбора для всего уравнения**

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

$$H_0 : \beta_2 = \dots = \beta_k = 0$$

$$H_1 : \text{at least one } \beta \neq 0$$

**В простой модели регрессии тест F был эквивалентен (двухстороннему) t-критерию по коэффициенту наклона, потому что «группа» состояла только из одной переменной.**

## **F-тест точности подбора для всего уравнения**

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

$$H_0 : \beta_2 = \dots = \beta_k = 0$$

$$H_1 : \text{at least one } \beta \neq 0$$

$$F(k-1, n-k) = \frac{ESS/(k-1)}{RSS/(n-k)}$$

Статистика F для теста была определена в последней последовательности в главе 2. ESS - это объясненная сумма квадратов, а RSS - остаточная сумма квадратов.



## F-тест точности подбора для всего уравнения

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

$$H_0 : \beta_2 = \dots = \beta_k = 0$$

$$H_1 : \text{at least one } \beta \neq 0$$

$$\begin{aligned} F(k-1, n-k) &= \frac{ESS/(k-1)}{RSS/(n-k)} \\ &= \frac{\frac{ESS}{TSS} / (k-1)}{\frac{RSS}{TSS} / (n-k)} \end{aligned}$$

Его можно выразить через R<sup>2</sup>, разделив числитель и знаменатель на TSS, общую сумму квадратов.

## F-тест точности подбора для всего уравнения

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

$$H_0 : \beta_2 = \dots = \beta_k = 0$$

$$H_1 : \text{at least one } \beta \neq 0$$

$$\begin{aligned} F(k-1, n-k) &= \frac{ESS/(k-1)}{RSS/(n-k)} \\ &= \frac{\frac{ESS}{TSS}/(k-1)}{\frac{RSS}{TSS}/(n-k)} = \frac{R^2/(k-1)}{(1-R^2)/(n-k)} \end{aligned}$$

ESS / TSS - это определение  $R^2$ . RSS / TSS равно  $(1 - R^2)$ . (См. Последнюю последовательность в главе 2.)

## ***F*-тест точности подбора для всего уравнения**

$$S = \beta_1 + \beta_2 ASVABC + \beta_3 SM + \beta_4 SF + u$$

**В качестве примера будет использована модель образовательного уровня. Мы будем предполагать, что *S* зависит от *ASVABC*, оценки способности и *SM*, и *SF*, высшего класса, завершеного матери и отцом респондента, соответственно.**

## **F-тест точности подбора для всего уравнения**

$$S = \beta_1 + \beta_2 ASVABC + \beta_3 SM + \beta_4 SF + u$$

$$H_0 : \beta_2 = \beta_3 = \beta_4 = 0, \quad H_1 : \text{at least one } \beta \neq 0$$

**Нулевой гипотезой для F-критерия точности подбора является то, что все три коэффициента наклона равны нулю. Альтернативная гипотеза состоит в том, что хотя бы одна из них отлична от нуля.**

## F-тест точности подбора для всего уравнения

$$S = \beta_1 + \beta_2 ASVABC + \beta_3 SM + \beta_4 SF + u$$

$$H_0 : \beta_2 = \beta_3 = \beta_4 = 0, \quad H_1 : \text{at least one } \beta \neq 0$$

```
. reg S ASVABC SM SF
```

Source	SS	df	MS	Number of obs = 500		
Model	1235.0519	3	411.683966	F( 3, 496)	=	81.06
Residual	2518.9701	496	5.07856875	Prob > F	=	0.0000
				R-squared	=	0.3290
				Adj R-squared	=	0.3249
				Root MSE	=	2.2536
S	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
ASVABC	1.242527	.123587	10.05	0.000	.999708	1.485345
SM	.091353	.0459299	1.99	0.047	.0011119	.1815941
SF	.2028911	.0425117	4.77	0.000	.1193658	.2864163
_cons	10.59674	.6142778	17.25	0.000	9.389834	11.80365

Вот результат регрессии с использованием набора данных 21.

## F-тест точности подбора для всего уравнения

$$S = \beta_1 + \beta_2 ASVABC + \beta_3 SM + \beta_4 SF + u$$

$$H_0 : \beta_2 = \beta_3 = \beta_4 = 0, \quad H_1 : \text{at least one } \beta \neq 0$$

```
. reg S ASVABC SM SF
```

Source	SS	df	MS	Number of obs	=	500
Model	1235.0519	3	411.683966	F( 3, 496)	=	81.06
Residual	2518.9701	496	5.07856875	Prob > F	=	0.0000
				R-squared	=	0.3290
				Adj R-squared	=	0.3249
Total	3754.022	499	7.52309018	Root MSE	=	2.2536

$$F(k-1, n-k) = \frac{ESS/(k-1)}{RSS/(n-k)} \quad F(3, 496)$$

В этом примере  $k - 1$ , количество объясняющих переменных, равно 3 и  $n - k$ , число степеней свободы, равно 496.

## F-тест точности подбора для всего уравнения

$$S = \beta_1 + \beta_2 ASVABC + \beta_3 SM + \beta_4 SF + u$$

$$H_0 : \beta_2 = \beta_3 = \beta_4 = 0, \quad H_1 : \text{at least one } \beta \neq 0$$

```
. reg S ASVABC SM SF
```

Source	SS	df	MS	Number of obs =	500
Model	1235.0519	3	411.683966	F( 3, 496) =	81.06
Residual	2518.9701	496	5.07856875	Prob > F =	0.0000
Total	3754.022	499	7.52309018	R-squared =	0.3290
				Adj R-squared =	0.3249
				Root MSE =	2.2536

$$F(k-1, n-k) = \frac{ESS/(k-1)}{RSS/(n-k)} \quad F(3, 496) = \frac{1235/3}{2519/496}$$

Числителем статистики F является объясненная сумма квадратов, деленная на k - 1. В выводе Stata эти числа приведены в строке model.

## F-тест точности подбора для всего уравнения

$$S = \beta_1 + \beta_2 ASVABC + \beta_3 SM + \beta_4 SF + u$$

$$H_0 : \beta_2 = \beta_3 = \beta_4 = 0, \quad H_1 : \text{at least one } \beta \neq 0$$

```
. reg S ASVABC SM SF
```

Source	SS	df	MS	Number of obs =	500
Model	1235.0519	3	411.683966	F( 3, 496) =	81.06
Residual	2518.9701	496	5.07856875	Prob > F =	0.0000
Total	3754.022	499	7.52309018	R-squared =	0.3290
				Adj R-squared =	0.3249
				Root MSE =	2.2536

$$F(k-1, n-k) = \frac{ESS/(k-1)}{RSS/(n-k)} \quad F(3, 496) = \frac{1235/3}{2519/496}$$

Знаменатель - это остаточная сумма квадратов, деленная на количество оставшихся степеней свободы.



## F-тест точности подбора для всего уравнения

$$S = \beta_1 + \beta_2 ASVABC + \beta_3 SM + \beta_4 SF + u$$

$$H_0 : \beta_2 = \beta_3 = \beta_4 = 0, \quad H_1 : \text{at least one } \beta \neq 0$$

```
. reg S ASVABC SM SF
```

Source	SS	df	MS	Number of obs =	500
Model	1235.0519	3	411.683966	F( 3, 496) =	81.06
Residual	2518.9701	496	5.07856875	Prob > F	= 0.0000
				R-squared	= 0.3290
				Adj R-squared	= 0.3249
Total	3754.022	499	7.52309018	Root MSE	= 2.2536

$$F(k-1, n-k) = \frac{ESS/(k-1)}{RSS/(n-k)} \quad F(3, 496) = \frac{1235/3}{2519/496} = 81.1$$

Следовательно, статистика F - 81,1. Все серьезные регрессионные пакеты вычисляют его как часть диагностики в регрессионном выпуске.

## F-тест точности подбора для всего уравнения

$$S = \beta_1 + \beta_2 ASVABC + \beta_3 SM + \beta_4 SF + u$$

$$H_0 : \beta_2 = \beta_3 = \beta_4 = 0, \quad H_1 : \text{at least one } \beta \neq 0$$

```
. reg S ASVABC SM SF
```

Source	SS	df	MS	Number of obs =	500
Model	1235.0519	3	411.683966	F( 3, 496) =	81.06
Residual	2518.9701	496	5.07856875	Prob > F	= 0.0000
				R-squared	= 0.3290
				Adj R-squared	= 0.3249
Total	3754.022	499	7.52309018	Root MSE	= 2.2536

$$F_{\text{crit},0.1\%}(3,500) = 5.51 \qquad F(3,496) = \frac{1235/3}{2519/496} = 81.1$$

Критическое значение для  $F(3,496)$  не указано в таблицах  $F$ , но оно должно быть очень близко к  $F(3500)$ . На уровне 0,1% это 5,51. Следовательно, мы легко отвергаем  $H_0$  на уровне 0,1%.

## F-тест точности подбора для всего уравнения

$$S = \beta_1 + \beta_2 ASVABC + \beta_3 SM + \beta_4 SF + u$$

$$H_0 : \beta_2 = \beta_3 = \beta_4 = 0, \quad H_1 : \text{at least one } \beta \neq 0$$

```
. reg S ASVABC SM SF
```

Source	SS	df	MS	Number of obs =	500
Model	1235.0519	3	411.683966	F( 3, 496) =	81.06
Residual	2518.9701	496	5.07856875	Prob > F	= 0.0000
				R-squared	= 0.3290
				Adj R-squared	= 0.3249
Total	3754.022	499	7.52309018	Root MSE	= 2.2536

$$F_{\text{crit},0.1\%}(3,500) = 5.51$$

$$F(3,496) = \frac{1235/3}{2519/496} = 81.1$$

Этот результат можно было бы ожидать, так как ASVABC и SF имеют очень значительную статистику t. Поэтому  $\beta_2$  и  $\beta_4$  не равны нулю.

## F-тест точности подбора для всего уравнения

$$S = \beta_1 + \beta_2 ASVABC + \beta_3 SM + \beta_4 SF + u$$

$$H_0 : \beta_2 = \beta_3 = \beta_4 = 0, \quad H_1 : \text{at least one } \beta \neq 0$$

```
. reg S ASVABC SM SF
```

Source	SS	df	MS	Number of obs =	500
Model	1235.0519	3	411.683966	F( 3, 496) =	81.06
Residual	2518.9701	496	5.07856875	Prob > F	= 0.0000
				R-squared	= 0.3290
				Adj R-squared	= 0.3249
Total	3754.022	499	7.52309018	Root MSE	= 2.2536

$$F_{\text{crit},0.1\%}(3,500) = 5.51$$

$$F(3,496) = \frac{1235/3}{2519/496} = 81.1$$

Необязательно, чтобы статистика F не была значительной, если некоторые статистические данные были значительными. Предположим, что мы выполнили регрессию с 40 объясняющими переменными, ни одна из которых не является истинным детерминантом зависимой переменной.

## F-тест точности подбора для всего уравнения

$$S = \beta_1 + \beta_2 ASVABC + \beta_3 SM + \beta_4 SF + u$$

$$H_0 : \beta_2 = \beta_3 = \beta_4 = 0, \quad H_1 : \text{at least one } \beta \neq 0$$

```
. reg S ASVABC SM SF
```

Source	SS	df	MS	Number of obs =	500
Model	1235.0519	3	411.683966	F( 3, 496) =	81.06
Residual	2518.9701	496	5.07856875	Prob > F	= 0.0000
				R-squared	= 0.3290
				Adj R-squared	= 0.3249
Total	3754.022	499	7.52309018	Root MSE	= 2.2536

$$F_{\text{crit},0.1\%}(3,500) = 5.51$$

$$F(3,496) = \frac{1235/3}{2519/496} = 81.1$$

Однако, если мы выполним t-тесты коэффициентов наклона на уровне 5% с 5% -ной вероятностью ошибки типа I, в среднем 2 из 40 переменных могут иметь «значимые» коэффициенты.

## F-тест точности подбора для всего уравнения

$$S = \beta_1 + \beta_2 ASVABC + \beta_3 SM + \beta_4 SF + u$$

$$H_0 : \beta_2 = \beta_3 = \beta_4 = 0, \quad H_1 : \text{at least one } \beta \neq 0$$

```
. reg S ASVABC SM SF
```

Source	SS	df	MS	Number of obs =	500
Model	1235.0519	3	411.683966	F( 3, 496) =	81.06
Residual	2518.9701	496	5.07856875	Prob > F	= 0.0000
				R-squared	= 0.3290
				Adj R-squared	= 0.3249
Total	3754.022	499	7.52309018	Root MSE	= 2.2536

$$F_{\text{crit},0.1\%}(3,500) = 5.51 \qquad F(3,496) = \frac{1235/3}{2519/496} = 81.1$$

С другой стороны, предположим, что у вас есть множественная регрессионная модель, которая правильно указана, а  $R^2$  высока. Вы ожидаете очень значительную статистику F.

## F-тест точности подбора для всего уравнения

$$S = \beta_1 + \beta_2 ASVABC + \beta_3 SM + \beta_4 SF + u$$

$$H_0 : \beta_2 = \beta_3 = \beta_4 = 0, \quad H_1 : \text{at least one } \beta \neq 0$$

```
. reg S ASVABC SM SF
```

Source	SS	df	MS	Number of obs =	500
Model	1235.0519	3	411.683966	F( 3, 496) =	81.06
Residual	2518.9701	496	5.07856875	Prob > F	= 0.0000
				R-squared	= 0.3290
				Adj R-squared	= 0.3249
Total	3754.022	499	7.52309018	Root MSE	= 2.2536

$$F_{\text{crit},0.1\%}(3,500) = 5.51 \qquad F(3,496) = \frac{1235/3}{2519/496} = 81.1$$

Однако, если объясняющие переменные сильно коррелированы и модель подвержена серьезной мультиколлинеарности, стандартные ошибки коэффициентов наклона могут быть настолько большими, что ни одна из статистических данных  $t$  не является значительной.

## F-тест точности подбора для всего уравнения

$$S = \beta_1 + \beta_2 ASVABC + \beta_3 SM + \beta_4 SF + u$$

$$H_0 : \beta_2 = \beta_3 = \beta_4 = 0, \quad H_1 : \text{at least one } \beta \neq 0$$

```
. reg S ASVABC SM SF
```

Source	SS	df	MS	Number of obs =	500
Model	1235.0519	3	411.683966	F( 3, 496) =	81.06
Residual	2518.9701	496	5.07856875	Prob > F	= 0.0000
				R-squared	= 0.3290
				Adj R-squared	= 0.3249
Total	3754.022	499	7.52309018	Root MSE	= 2.2536

$$F_{\text{crit},0.1\%}(3,500) = 5.51$$

$$F(3,496) = \frac{1235/3}{2519/496} = 81.1$$

В этой ситуации вы бы знали, что ваша модель хорошая, но вы не в состоянии точно определить вклад, создаваемый объясняющими переменными отдельно.