


STAT 501

Regression Methods

2.6 - The Analysis of Variance (ANOVA) table and the F-test

 [Printer-friendly version \(https://onlinecourses.science.psu.edu/stat501/print/book/export/html/266\)](https://onlinecourses.science.psu.edu/stat501/print/book/export/html/266)

We've covered quite a bit of ground. Let's review the analysis of variance table for the example concerning skin cancer mortality and latitude (skincancer.txt)

(<https://onlinecourses.science.psu.edu/stat501/sites/onlinecourses.science.psu.edu.stat501/files/data/skincancer.txt>).

The regression equation is Mort = 389 - 5.98 Lat

Predictor	Coef	SE Coef	T	P
Constant	389.19	23.81	16.34	0.000
Lat	-5.9776	0.5984	-9.99	0.000

S = 19.12 R-Sq = 68.0% R-Sq(adj) = 67.3%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	36464	36464	99.80	0.000
Residual Error	47	17173	365		
Total	48	53637			

Recall that there were 49 states in the data set.

- The degrees of freedom associated with SSR will always be 1 for the simple linear regression model. The degrees of freedom associated with $SSTO$ is $n-1 = 49-1 = 48$. The degrees of freedom associated with SSE is $n-2 = 49-2 = 47$. And the degrees of freedom add up: $1 + 47 = 48$.
- The sums of squares add up: $SSTO = SSR + SSE$. That is, here: $53637 = 36464 + 17173$.

Let's tackle a few more columns of the analysis of variance table, namely the "**mean square**" column, labeled **MS**, and the F -statistic column, labeled **F**.

Definitions of mean squares

We already know the "**mean square error (MSE)**" is defined as:

$$MSE = \frac{\sum (y_i - \hat{y}_i)^2}{n - 2} = \frac{SSE}{n - 2}.$$

That is, we obtain the mean square error by dividing the error sum of squares by its associated degrees of freedom $n-2$. Similarly, we obtain the "**regression mean square (MSR)**" by dividing the regression sum of squares by its degrees of freedom 1:

$$MSR = \frac{\sum(\hat{y}_i - \bar{y})^2}{1} = \frac{SSR}{1}.$$

Of course, that means the regression sum of squares (SSR) and the regression mean square (MSR) are always identical for the simple linear regression model.

Now, why do we care about mean squares? Because their expected values suggest how to test the null hypothesis $H_0: \beta_1 = 0$ against the alternative hypothesis $H_A: \beta_1 \neq 0$.

Expected mean squares

Imagine taking many, many random samples of size n from some population, and estimating the regression line and determining MSR and MSE for each data set obtained. It has been shown that the average (that is, the expected value) of all of the $MSRs$ you can obtain equals:

$$E(MSR) = \sigma^2 + \beta_1^2 \sum_{i=1}^n (X_i - \bar{X})^2$$

Similarly, it has been shown that the average (that is, the expected value) of all of the $MSEs$ you can obtain equals:

$$E(MSE) = \sigma^2$$

These expected values suggest how to test $H_0: \beta_1 = 0$ versus $H_A: \beta_1 \neq 0$:

- If $\beta_1 = 0$, then we'd expect the ratio MSR/MSE to equal 1.
- If $\beta_1 \neq 0$, then we'd expect the ratio MSR/MSE to be greater than 1.

These two facts suggest that we should use the ratio, MSR/MSE , to determine whether or not $\beta_1 = 0$.

Note that, because β_1 is squared in $E(MSR)$, we cannot use the ratio MSR/MSE :

- to test $H_0: \beta_1 = 0$ versus $H_A: \beta_1 < 0$
- or to test $H_0: \beta_1 = 0$ versus $H_A: \beta_1 > 0$.

We can only use MSR/MSE to test $H_0: \beta_1 = 0$ versus $H_A: \beta_1 \neq 0$.

We have now completed our investigation of all of the entries of a standard analysis of variance table. The formula for each entry is summarized for you in the following analysis of variance table:

Source of Variation	DF	SS	MS	F
Regression	1	$SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$	$MSR = \frac{SSR}{1}$	$F^* = \frac{MSR}{MSE}$
Residual error	$n-2$	$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$	$MSE = \frac{SSE}{n-2}$	
Total	$n-1$	$SSTO = \sum_{i=1}^n (y_i - \bar{y})^2$		

However, we will always let Minitab do the dirty work of calculating the values for us. Why is the ratio MSR/MSE labeled F^* in the analysis of variance table? That's because the ratio is known to follow an **F distribution** with 1 numerator degree of freedom and $n-2$ denominator degrees of freedom. For this reason, it is often referred to as the analysis of variance F -test. The following section summarizes the formal F -test.

The formal F -test for the slope parameter β_1

The null hypothesis is $H_0: \beta_1 = 0$.

The alternative hypothesis is $H_A: \beta_1 \neq 0$.

The test statistic is $F^* = \frac{MSR}{MSE}$.

As always, the P -value is obtained by answering the question: "What is the probability that we'd get an F^* statistic as large as we did, if the null hypothesis is true?"

The P -value is determined by comparing F^* to an F distribution with 1 numerator degree of freedom and $n-2$ denominator degrees of freedom.

In reality, we are going to let Minitab calculate the F^* statistic and the P -value for us. Let's try it out on a new example!

◀ 2.5 - Analysis of Variance: The Basic Idea (/stat501/node/265)

up (/stat501/node/260)

2.7 - Example: Are Men Getting Faster? › (/stat501/node/267)
