13.10: Computing F

The formula for the ANOVA F in regression is as follows:

$$F = \frac{RSS \div df_M}{SSE \div df_E}$$

The numerator is used to calculate the mean sum of squares for the regression model (MSSR). The denominator is used to calculate the mean sum of squares error (MSSE). The final F-value indicates the ratio of residuals reduced using the regression model relative the residuals remaining when using the regression model. The higher the F-value, the more the regression model improved predictions and, thus, the greater the chance of a significant result. The lower the F-value, the less the model improved predictions and, thus, the lesser the chance of a significant result.

For Data Set 12.1, the four parts needed to compute F are as follows:

1. SSR = 1,728.0234
2. SSE = 195.9766
3. df_M = 1
4. df_E = 8

These are plugged into the F-formula and used to solve for F as follows:

$$F = \frac{1,728.0234 \div 1}{195.9766 \div 8} = \frac{1,728.0234}{24.4971} = 70.5399$$

When rounded to the hundredths place, this result is:

$$F = 70.54$$

This is a very large result indicating that the regression model reduces a much greater proportion of error (i.e. residuals) than it leaves unexplained. Another way to say this is that the regression model provides much better predictions of Y than the alternative model.

When using hand-calculations to test a hypothesis, significance is assessed using a critical value (CV). We will see how to find the CV and use it to determine significance in a later section. For now, we will move on to reviewing the other analyses that are needed for a complete regression analysis.

Secondary Analyses for Regression: **t-Testing and Slopes**

When a regression ANOVA is significant, the slope of the regression line is computed and tested to see whether it is why the regression model significantly improved predictions. When the *t*-test for the slope is significant, it means that the slope significantly improved predictions of Y. When using a bivariate regression, it is redundant to check the significance of the slope because it is the only slope so it must be the one improving predictions. For this reason, if the ANOVA is significant, the *t*-test will also be significant in simple (bivariate) regression.

When must slope significance be checked?

When an advanced version of regression is used, there are multiple predictors and, thus, multiple slopes (one for each predictor variable). When those models are used, the *t*-tests function like post-hoc analyses in one-way ANOVA. Thus, when a regression model with multiple predictors is being tested and has a significant ANOVA result, *t*-tests are used to assess which slopes were significantly contributing to improving the predictions and which, if any, were not. However, the significance of the slope does not need to be checked in bivariate regression because there is only one predictor. Therefore, if the ANOVA is significant, the slope of that predictor is also significant and when the ANOVA is not significant, the slope of that predictor is also non-significant.

When the *t*-test is run using SPSS for Data Set 12.1, the results are shown by the software as follows:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	38.5529	5.177		7.447	<.001
	Hours of Sleep	6.3765	.759	.948	8.399	<.001
a. Dependent Variable: Quiz Scores						

The first row of the table shows the results for the y-intercept and the second row shows the results for the slope. Because the slope is the focus of the *t*-test, we will focus on the second row of results. It shows the name of the X-variable (Hours of Sleep), the slope of the line (6.3765), and the results of the *t*-test assessing the usefulness of that slope.

The obtained *t*-value for the slope, when rounded to the hundredths place, is 8.40. The "Sig." stands for "significance;" the value shown in that column is the *p*-value (which is the risk of a Type I Error). When $p < .05$, a result is significant. When $p \geq .05$, a result is *not* significant. The sig. value shown in the SPSS output is shown as "<.001" which means the risk of a Type I Error is less than 0.1%. This indicates that the slope of the regression line is significantly contributing to the improvement in predictions when using the regression model.

Interpreting Slopes

When a bivariate regression model is significant, the slope is often interpreted and reported as part of a complete results paragraph. The slope indicates the amount of change in the Y-variable that is predicted for every one unit increase in the X-variable. The slope, when rounded to the hundredths place, is 6.38. The X-variable is Hours of Sleep and the Y-Variable is Quiz Scores. Thus, the slope for Data Set 12.1 would be interpreted as follows:

> *For every one hour increase in sleep, there is a 6.38 unit increase predicted for quiz scores.*

Making Predictions

When a regression result is significant, the regression equation can be applied to make predictions. The formula used to make predictions was noted in an earlier section titled *Computing Residuals for the Regression Model: Sum of Squares Error (SSE)* so it will only be briefly reviewed here. The regression equation is written as follows:

$$\hat{Y} = b_0 + b_1 x$$

The necessary parts for this equation for Data Set 12.1 are as follows:

b_0: Y-intercept	38.5529
b_1: Slope of Hours of Sleep	6.3765
Note: Values continue but are shown to the fourth decimal place for space.	

Thus, the predicted Y-values for the regression model with Data Set 12.1 are computed using the following regression equation:

$$\hat{Y} = 38.5529 + 6.3765x$$

In the prior section, we used this equation to predict the Y-values using the X-values in Data Set 12.1. This was so we could compare the predictions to the actual Y-values in the data set. However, it can be used to predict using any X-value, not just those in the data set. To predict, take a given X-value, plug it into the equation, and compute $\hat{Y}$. Here is what it looks like to compute and interpret predicted Y given three different X-values:

Example 1	Example 2	Example 3

Example 1	Example 2	Example 3
Given X = 0.00 $\hat{Y}$ = 6.3765(0.00) + 38.5529 $\hat{Y}$ = 38.5529 The predicted quiz score when someone has gotten 0.00 hours of sleep is approximately 38.55 points.	Given X = 4.00 $\hat{Y}$ = 6.3765(4.00) + 38.5529 $\hat{Y}$ = 64.0589 The predicted quiz score when someone has gotten 0.00 hours of sleep is approximately 64.06 points.	Given X = 8.00 $\hat{Y}$ = 6.3765(8.00) + 38.5529 $\hat{Y}$ = 89.5649 The predicted quiz score when someone has gotten 0.00 hours of sleep is approximately 89.56 points.

Note: Slight error in being introduced by using rounded values for the slope and intercept.

Limitations

A major limitation of the regression, which it shares with correlation, is that it cannot be used to determine cause-effect relationships. Just because two things are mathematically related, does not mean that either is the cause of another. Therefore, though tempting, it is not generally appropriate to use causal language when interpreting the results of correlation nor regression (see Chapter 8 for a review of causal language). Some of the language can be misunderstood to be causal such as the use of the terms *predicted* and *explained.* However, these do not indicate that cause-effect has been determined. Instead, predictions are estimations of what is expected based on the current regression model. When we refer to the amount of variance that is "explained" we simply mean the amount that is accounted for based on the pattern among the data and corresponding regression line. When we say something is explained, we have not (and generally cannot) determine whether the relationship is causal and, if so, which variable is the cause. Therefore, it is important not to presume nor to indicate that a regression result is sufficient to determine that X caused Y.

1. What does SST represent?
2. How is SST calculated?
3. What does SSE represent?
4. How is SSE calculated?
5. What does SSR represent?
6. How is SSR calculated?
7. How is df_M calculated?
8. How is df_E calculated?
9. What is the formula for calculating F?