# 13.9: Testing a Regression Model

## Testing a Regression Model with ANOVA

ANOVA is used to test whether the regression model is a good fit to the data and, thus, is useful for predicting $Y$. To do so, three kinds of error are computed:

1. Residuals using the alternative model (i.e. the $\bar{Y}$ model),
2. Residual using the regression model (i.e. using $\hat{Y} = b_0 + b_1 x$ ), and
3. Residuals in the alternative model that are not accounted for using the regression model.

Residuals using the alternative model are known as the sum of squares total (SST). Think of SST as the error we would have if we used the simplest way of predicting. If the regression model is useful, it will produce residuals that are significantly lower than the SST. Thus, we must also compute the residuals using the regression model. The residuals that occur when using the regression model are known as the sum of squares errors (SSE); SSE summarizes the residuals that are not accounted for using the regression model. The difference between the residuals for the alternative model (SST) and the regression model (SSE) are the residuals that have been accounted for using the regression model; these are known as the regression sum of squares (a.k.a. sum of squares regression; SSR).

The three forms of residuals and what they stand for can be summarized as follows:

Error in the Alternative Model = Error Explained Using the Regression Model + Error Left Unexplained Using the Regression Model

SST = SSR + SSE

Think of SSR as the amount of error that is reduced when using the regression model. The higher the SSR, the lower the SSE and the better the regression model is at predicting $Y$-values.

### Computing Residuals for the Alternative Prediction Model: Sum of Squares Total (SST)

Residuals represent errors in predictions. The overall error observed is computed as the residuals when using the alternative model (SST). In this model, the mean of the $Y$-values is used as the prediction for all data points, regardless of their $X$-values.

Recall that residuals are differences between the predicted $Y$-values for a data set and the actual $Y$-values for the data set. The predicted $Y$-values in the alternative model are all the mean of $Y$. Therefore, a horizontal line at the $\bar{Y}$ is what is used to predict in the alternative model. As we can see in Graph 13.3, the $\bar{Y}$ prediction line is too high for some data points (i.e. the predicted $Y$ is higher than the actual $Y$ value) and is too low for other data points (i.e. the predicted $Y$ is lower than the actual $Y$ value). Thus, some residuals are negative while others are positive. Because the $\bar{Y}$ is being used, which balances deviations so that they sum to 0, it also causes residuals to balance so they sum to 0. To avoid this issue when computing total residuals, a squared version is used. This use of squaring to is just like we have saw when working with standard deviations in earlier chapters (such as Chapter 4).

The **sum of squares total (SST)** in regression refers to the deviations when using the alternative model. To compute SST, these four steps are followed:

1. Find $\bar{Y}$ (the mean of the $Y$-values)
2. Find the residuals by subtracting the $\bar{Y}$ from each known Y-value
3. Square the residuals
4. Sum the squared residuals to get the SST

Let's compute the SST using Data Set 12.1. In Table 13.1, we see the data and three additional columns:

1. The predicted quiz scores (i.e. the predicted $Y$-values),
2. The residuals, and
3. The squared residuals.

Notice that $\bar{Y}$ is used for all predicted scores in the alternative model. The sum of squared residuals (SST) is shown at the bottom of the table. The SST is 1,924.

Table 13.1: Computing Residuals for Data Set 12.1 Using the Alternative Model

| Sleep Hours | Actual Quiz Scores | Predicted Quiz Scores ($\bar{Y}$) | Residuals | Squared Residuals |
|---|---|---|---|---|
| 7 | 92 | 80 | 12 | 144 |
| 8 | 88 | 80 | 8 | 64 |
| 9 | 96 | 80 | 16 | 256 |
| 6 | 70 | 80 | -10 | 100 |
| 6 | 79 | 80 | -1 | 1 |

| Sleep Hours | Actual Quiz Scores | Predicted Quiz Scores ($\bar{Y}$) | Residuals | Squared Residuals |
|---|---|---|---|---|
| 4 | 64 | 80 | -16 | 256 |
| 5 | 75 | 80 | -5 | 25 |
| 10 | 98 | 80 | 18 | 324 |
| 3 | 53 | 80 | -27 | 729 |
| 7 | 85 | 80 | 5 | 25 |
| | | | | SST = 1,924 |

## Computing Residuals for the Regression Model: Sum of Squares Error (SSE)

The hypothesis when using regression is that the residuals will be significantly lower when using the regression model than the alternative model. Therefore, we must also compute the residuals when using the regression model. This can be thought of as the error that is left unaccounted for, unreduced, or unexplained when using the regression model. The error that occurs when using the regression model is known as the sum of squares error (SSE). It is important to note that SPSS software refers to SSE as "Sum of Squares Residual" rather that "Sum of Squares Error" (see Table 13.3 for a summary of the various names and symbols used for residuals computations in regression).

Before we can compute the residuals, we must make predictions using the regression model. The regression formula used to predict $Y$-values is:

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

To use this formula to make predictions, we must first know the $y$-intercept ($b_0$) and the slope ($b_1$) for the regression line. Computing these by hand is beyond the scope of this chapter. Instead, these are often generated using software such as SPSS. Thus, we will use the slope and y-intercept as computed in SPSS. For Data Set 12.1, the slope and intercept are as follows:

| | |
|---|---|
| $b_0$: $Y$-intercept | 38.5529 |
| $b_1$: Slope of Hours of Sleep | 6.3765 |
| *Note: Values continue but are rounded 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:

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

To use this equation, the X-value for each case is plugged in. It is then multiplied by the slope. Finally, it is added to the y-intercept. For example, in the first case for Data Set 12.1, the $X$-value is 7. Thus, the predicted $Y$-value for that case is computed as follows:

$$\hat{Y} = 38.5529 + 6.3765(7)$$
$$\hat{Y} = 38.5529 + 44.6355$$
$$\hat{Y} = 83.1884$$

This process is repeated to find the $\hat{Y}$ for each case. The $\hat{Y}$ for each case is shown in Table 13.2.

Now that the predicted $Y$ values are known, the sum of squares errors can be computed. **Sum of squares errors (SSE)** is computed by finding the residuals for each data point when using the regression model, squaring those residuals, and then summing them to get a total. Thus, the summary of the steps to compute SSE are as follows:

1. Find the predicted $Y$-values for each case using the regression equation, which is: $\hat{Y} = b_0 + b_1 x$
2. Find each residual by subtracting the $\hat{Y}$ from each known $Y$-value
3. Square the residuals
4. Sum the squared residuals to get the SSE

Let's compute the SSE using Data Set 12.1. Table 13.2 includes the data in the first two columns and three additional columns of computations:

1. The predicted quiz scores (i.e. the predicted $Y$-values),
2. The residuals for those predictions, and
3. The squared version of those residuals.

Notice that the predicted $Y$-values vary because each depends upon the corresponding $X$-value for the case. The sum of squared residuals (SSE) is shown at the bottom of the table. The SSE is 195.9766. This value represents the error that occurs when using the regression model to predict. Notice that this is much lower than the error we saw when using the alternative model to predict (i.e. the SSE of 195.9766 is noticeably lower than the SST of 1,924). This is a desirable outcome.

Table 13.2: Computing Residuals for Data Set 12.1 Using the Regression Model (SSE)

| Sleep Hours | Actual Quiz Scores | Predicted Quiz Scores ($\hat{Y}$) | Residuals | Squared Residuals |
|---|---|---|---|---|
| 7 | 92 | 83.1882 | 8.8118 | 77.6473 |
| 8 | 88 | 89.5647 | -1.5647 | 2.4483 |
| 9 | 96 | 95.9412 | 0.0588 | 0.0035 |
| 6 | 70 | 76.8118 | -6.8118 | 46.4001 |
| 6 | 79 | 76.8118 | 2.1882 | 4.7884 |
| 4 | 64 | 64.0588 | -0.0588 | 0.0035 |
| 5 | 75 | 70.4353 | 4.5647 | 20.8366 |
| 10 | 98 | 102.3176 | -4.3176 | 18.6420 |
| 3 | 53 | 57.6824 | -4.6824 | 21.9244 |
| 7 | 85 | 83.1882 | 1.8118 | 3.2825 |
| SSE = 195.9766 | | | | |

## Computing Residuals Explained using the Regression Model: Regression Sum of Squares (SSR)

We must also compute the amount of error that has been reduced or explained when using the regression model. The name isn't very intuitive but the regression sum of squares (SSR) refers to the amount of squared residuals that are reduced when using the regression model compared to the alternative model. Think of the SSR is the amount of improvement we get when using the regression model instead of the alternative model. In keeping, the greater the SSR, the better the predictions are.

Let's take a moment to consider how SSR, SSE, and SST are connected to each other. SSR and SSE are in opposition to one another. Recall that SSE refers to residuals left unexplained by the regression model. The lower the SSE, the better the predictions are. When SSR is higher, SSE is lower and when SSR is lower, SSE is higher. Recall also that the SST is the sum of the SSR and SSE. Therefore, we can find any one of these three if we already know the other two.

Here is a summary of the three forms of sum of squared residuals and how they are connected:

Error in the Alternative Model = Error Explained Using the Regression Model + Error Left Unexplained Using the Regression Model

SST = SSR + SSE

1,924 = SSR + 195.9766

SSR can, thus, be found by subtracting SSE from SST. For Data Set 12.1, these computations are as follows:

1,924 – 195.9766 = SSR

1,924 – 195.9766 = 1,728.0234

Thus, the SSR is easily and quickly computed using the SST and SSE.

Table 13.3. Symbols and Corresponding Formulas for $Y$

| Symbol | Meaning | Formula |
|---|---|---|
| $Y$ | Raw or observed score for a $Y$-variable | Given in a data set |
| $\bar{Y}$ | The mean of scores for a $Y$-variable | $\bar{Y} = \frac{\Sigma Y}{n}$ |
| $\hat{Y}$ | A predicted value of $Y$ | $\hat{Y} = b_1 x + b_0$ |
| $e$ | Residual; the difference between an observed value and its predicted value | $e = Y_i - \hat{Y}_i$ |

Table 13.4. Formulas for Computing Residuals

| Symbol | Name in SPSS | Meaning | Formula | Steps |
|---|---|---|---|---|
| SST | Sum | The total squared error when using the alternative model | $SST = \Sigma(Y - \bar{Y})^2$ | 1. Fi |

| Symbol | Name in SPSS | Meaning | or Formula | Steps |
|---|---|---|---|---|
| | of Squares Total | | SST = SSR + SSE | nd $\bar{Y}$ (the mean of the $Y$-values)<br>2. Find each residual by subtracting the $\bar{Y}$ from each known $Y$- |

| Symbol | Name in SPSS | Meaning | Formula | Steps |
|---|---|---|---|---|
| | | | | value<br>3. Square the residuals<br>4. Sum the squared residuals to get the SST<br>Alternately, it can be found as the sum of SSR and SSE |

| Symbol | Name in SPSS | Meaning | Formula | Steps when those two values are already known. |
|---|---|---|---|---|
| SSE | Sum of Squares Residuals | The total squared error when using the regression model. This can also be thought of as the amount of error that is not explained or reduced by using the regression model | $\text{SSE} = \Sigma(Y - \bar{Y})^2$ or $\text{SSE} = \Sigma(e)^2$ | 1. Find each predicted $Y(\hat{Y})$<br>2. Find the residual (error) for each case by subtracting pre di |

| Symbol | Name in SPSS | Meaning | Formula | Steps |
|---|---|---|---|---|
| | | | | cted $Y(\hat{Y})$ from observed $\boldsymbol{Y}$.<br>3. Square the residuals.<br>4. Sum the squared residuals. |
| SSR | Sum of Squares Regression | The total squared error that is reduced or explained when using the regression model. This can also be thought of as the amount of improvement when using the regression model. | $\text{SSR} = \Sigma(\hat{Y} - \bar{Y})^2$<br>or<br>SSR = SST - SSE | 1. Find each predict |

| Symbol | Name in SPSS | Meaning | Formula | Steps |
|---|---|---|---|---|
| | | | | ed $Y(\hat{Y})$ using $\hat{Y} =$<br>2. Find the mean of $Y(\hat{Y})$.<br>3. Find the residuals for the mean of $\boldsymbol{Y}$ by subtracting the mean of $Y(\hat{Y})$ fr |

| Symbol | Name in SPSS | Meaning | Formula | Steps |
|---|---|---|---|---|
| | | | | o m ea c h pr e di ct e d $Y(\hat{Y})$ . 4. S q u ar e th e re si d u al s. 5. S u m th e sq u ar e d re si d u al s. Alter natel y, it can be foun d as the differ ence betw een |

| Symbol | Name in SPSS | Meaning | Formula | Steps |
|---|---|---|---|---|
| | | | | SST and SSE when those two values are already known. |

*Note:* The subscript $i$ is used to denote individual scores.

### Testing Goodness-of-Fit with ANOVA

When a regression is significant, it means a substantial enough proportion of the variance is accounted for or reduced using the regression model compared to how much is left unaccounted for by the model. To test this, regression uses ANOVA. Recall from Chapter 10 that ANOVA computes sum of squares between to represent variation that is systematic between groups and sum of squares within which represents variation that is not systematic (and occurs within groups). This same concept is used for an ANOVA within regression but with different names. Specifically, when ANOVA is used in regression, the regression sum of squares (SSR) represents the systematic variation (i.e. residuals which are accounted for by the regression model) and the sum of squares error (SSE) represents non-systematic variation (i.e. residuals which are unaccounted for by the regression model). These are used to assess the goodness-of-fit of the regression model.

The four components used to find $F$ for a regression ANOVA are as follows:

1. SSR
2. SSE
3. $df_M$
4. $df_E$

We learned how to find SSR and SSE in the prior section. For this section, therefore, we will only add in how to compute the remaining two of these four components: $df_M$ and $df_E$.

### Degrees of Freedom for Regression

The $df_M$ refers to the degrees of freedom for the regression model. This is equal to the number of predictors being used. In a simple, bivariate regression $df_M$ = 1 because only one predictor variable is being used (the $X$-variable). Thus, for Data Set 12.1, $df_M$ = 1.

The $df_E$ refers to the degrees of freedom error. This is equal to the sample size ($n$) minus the number of predictors being used plus 1 like so:

$$df_E = n - (\text{number of predictors} + 1)$$

In a simple, bivariate regression $df_E = n - 2$ because only one predictor variable is being used (the $X$-variable) and if we add 1 to the number of predictors we get 2. Thus, for Data Set 12.1, we compute:

$$\begin{gathered} df_E = n - (\text{ number of predictors } + 1) \\ df_E = 10 - (1 + 1) \\ df_E = 10 - 2 \\ df_E = 8 \end{gathered}$$