Print

Comparing the Regression Model to a Baseline Model

To determine whether the predictor variable explains a significant amount of variability in the response variable, the simple linear regression model is compared to the baseline model. The fitted regression line in a baseline model is just a horizontal line across all values of the predictor variable. The slope of this line is 0, and the y-intercept is the sample mean of Y, which is $\overline{Y}$.

In a baseline model, the X and Y variables are assumed to have no relationship. This means that for predicting values of the response variable, the mean of the response, $\overline{Y}$, does not depend on the values of the X variable. To determine whether a simple linear regression model is better than the baseline model, you compare the explained variability to the unexplained variability similarly to ANOVA.

The explained variability is related to the difference between the regression line and the mean of the response variable. For each data point, you calculate this difference, which equals

$$\widehat{Y}_i - \overline{Y}$$

. To eliminate negative distances, you square each of these values. Then sum these squared values to obtain the model sum of squares, or SSM, which is the amount of variability that your model explains.

$$\Sigma(\widehat{Y}_i - \overline{Y})^2$$

The unexplained variability is the difference between the observed values and the regression line. For each data point, you calculate $Y_i - \widehat{Y}_i$ and square the difference. Then sum these squared values to find the error sum of squares, or SSE, which is the amount of variability that your model fails to explain.

$$\Sigma(Y_i - \widehat{Y}_i)^2$$

The total variability is the difference between the observed values and the mean of the response variable. For each data point, you calculate $Y_i - \overline{Y}$ and square the difference. Then sum these squared values to get the corrected total sum of squares, or SST, which is, of course, the sum of the model and error sum of squares.

$$\Sigma(Y_i - \overline{Y})^2$$

The SSM and SSE are divided by their corresponding degrees of freedom to produce the mean-square model (MSM) and mean-square error (MSE). The significance of the regression analysis is assessed the same way as ANOVA, that is, by computing the F ratio, the mean squared model divided by the mean squared error, and the corresponding p-value. In fact, you'll see an ANOVA table in your regression output as well.

Close