STAT 308 – Chapter 7

Background Information

We have discussed hypothesis testing for a simple linear regression for three different types of alternative hypotheses $H_a : \beta_1 > 0$, $H_a : \beta_1 < 0$, and $H_a : \beta_1 \neq 0$. We have also discussed r^2 and how it is used to determine the percent of variation in Y that can be explained by its *linear* relationship with X. Now, we will show a new way to obtain this information that can be easily extended to multiple linear regression

Important Definitions

Analysis of Variance (ANOVA): A table that breaks down the sources of the variation in the response variable, Y, when we include the explanatory variable, X, in our linear model

$$\begin{aligned}
\text{Total Sums of Squares} = \text{SSY} &= \sum_{i=1}^{n}(y_i - \bar{y})^2 \\
&= \sum_{i=1}^{n}(y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 \\
&= \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \\
&= \text{Model Sums of Squares} + \text{Sum of Squared Errors}
\end{aligned}$$

where $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$.

Model Sums of Squares (SSM): The amount of variation in Y explained by the linear model with X

$$\text{SSM} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

Sum of squared errors (SSE) is the same as chapter 5-6

$$\text{SSE} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Hypothesis Testing for Simple Linear Regression using ANOVA

Suppose, we are interested in testing whether or not two variables X and Y have a significant linear relationship. Recall, this is equivalent to testing

$$H_0 : \beta_1 = 0$$

$$H_a : \beta_1 \neq 0.$$

It turns out that MSM $= \frac{\text{SSM}}{1}$ and MSE $= \frac{\text{SSE}}{n-2}$ are statistically independent where MSM stands for Mean Squares from the Model and MSE stands for MSE. Then, we can say

$$\frac{\text{SSM}}{\text{SSE}/(n-2)} \sim F_{df1=1,df2=n-2}$$

Let $f = \frac{\text{SSM}}{\text{SSE}/(n-2)}$, where f is now the appropriate test statistic. Then,

$$p-value = Pr(F_{df1=1,df2=n-2} > f).$$

- If $p-value \leq \alpha$, reject H_0 and say there is significant evidence of a linear relationship between X and Y
- If $p-value > \alpha$, reject H_0 and say there is not significant evidence of a linear relationship between X and Y

ANOVA Table for Simple Linear Regression

	df	Sums of Squares	Mean Square	f Value	Pr(>f)
X(Model)	1	SSM	MSM	$\frac{\text{MSM}}{\text{MSE}}$	$Pr(F_{1,n-2} > \frac{\text{MSM}}{\text{MSE}})$
Error(Residuals)	$n-2$	SSE	MSE		
Total	$n-1$	SSY = SSM + SSE			

Example

Use the following incomplete ANOVA table to answer the following questions.

	df	Sums of Squares	Mean Square	f Value	Pr(>f)
Model	1	50.83			
Error	48				
Total		98.48			

a. What is the total degrees of freedom?

b. What is the sum of squared errors?

c. What is the mean squares of the model and the mean squared error?

d. What is the test statistic and p-value for testing for a significant linear relationship?

Example

Using the bloodpressure dataset, obtain an ANOVA table and perform a hypothesis test for a significant linear relationship.

What can we obtain from an ANOVA table?

We can find an estimate for the regression variance ($\sigma^2_{Y|X}$(.

$$s^2_{Y|X} = \text{MSE} = \frac{\text{SSE}}{n-2}$$

.

We can find the proportion of variance explained by the model (r^2).

$$r^2 = \frac{\text{SSM}}{\text{SST}}.$$

For simple linear regression, the result of the F-test is the same as the t-test for testing for a significant linear relationship.

This follows from the fact that $F_{1,n-2} = T^2_{n-2}$ and $f = t^2$ where $t = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}}$. It follows that

$$Pr(F_{1,n-2} > f)$$
$$= Pr(T^2_{1,n-2} > t^2)$$
$$= Pr(T_{1,n-2} > |t|),$$

which is the p-value we formulated in Chapter 5.

Let's show this relationship through a formal example.