Week Four Handout

Manu Navjeevan

October 20, 2019

1 Theory Overview

In the past three sections, we have hypothesized a theoretical model for the data of the form:

$$Y_i = \beta_1 + \beta_2 \cdot X_i + \epsilon_i$$

We have gone over how to estimate the parameters of this model $(\hat{\beta}_1, \hat{\beta}_2)$, obtain confidence intervals, and even extend this model to nonlinear functions of X. However, what we may be interested in is how *good* this model is. In other words, we may be interested in how well this theorized model describes the real relationship between X and Y.

A primary goal of linear regression is to use X to predcit Y. For this reason, we may be interested in estimating how confident we are in our predicted value of Y for any value of X. To do so, we can create prediction intervals for $\hat{y}_0$, where $\hat{y}_0$ is given:

$$\hat{y}_0 = \hat{\beta}_1 + \hat{\beta}_2 \cdot x_0$$

The construction of these is similar to the construction of confidence intervals for β. However, this time instead of using the standard error of β, we use the standard error of the model calculated

$$\hat{var}(f) = \hat{\sigma}^2[1 + \frac{1}{N} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}]$$
$$se(f) = \sqrt{\hat{var}(f)}$$

The prediction interval for y_0 (the real predicted value at x_0) is then

$$\hat{y} \pm t_{n-2,(1-\alpha/2)} se(f)$$

A wide prediction interval can give us an indication that our model is not doing a great job of predicting y, at least at a certain value of x_0. However, we may want to evaluate

the overall goodness of fit of our model. To do this, we go back to a concept from week 1, R^2. First, we calculate the (sample) correlation coeffecient:

$$r_{xy} = \frac{s_x y}{s_x s_y}$$

where

$$s_{xy} = \frac{1}{N-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{x_i})$$

$$s_x = \sqrt{\frac{1}{N-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$s_y = \sqrt{\frac{1}{N-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

To get R^2, we simply square r_{xy}

$$R^2 = r_{xy}^2$$

we can interpret this as the amount of variation in y explained by the linear model with x. This can also be obtained in more general models by using the SSE and SST. Formally:

$$SST = \sum (y_i - \bar{y})^2$$

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

Then

$$R^2 = 1 - \frac{SSE}{SST}$$

Finally, when we specified the error, we imposed some restrictions on the error terms. We can check these by looking at the residuals plots. For more info on this check the Week 1 handout. In addition, however, note that for confidence intervals, we have imposed that our errors are normally distributed. To do so, look at the histogram of your errors. In figure 1 we have plotted the histogram of errors for the regression line from week 1. We can see here that distribution looks roughly normal

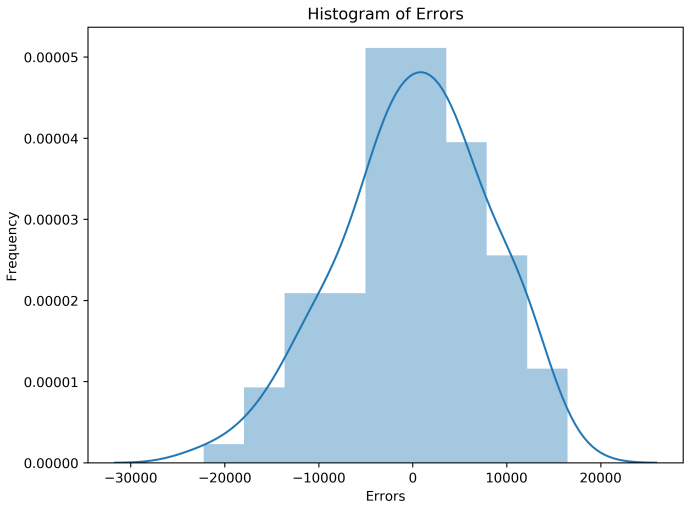

Figure 1: Errors look roughly normally distributed

2 Practice Problems

1. (a) Suppose there is a regression with quantities $\sum(y_i - \bar{y})^2 = 631.63$ and $\sum \hat{e}_i^2 = 182.85$. Find R^2.

   (b) Suppose that a simple regression has quantities $N = 20$, $\sum y_i^2 = 5930.94$, $\bar{y} = 16.035$, and $SSR = 666.72$. Find R^2.

   (c) Suppose that a simple regression has quantities $R^2 = 0.7911$, $SST = 552.36$ and $N = 20$. Find $\hat{\sigma}^2$.

2. Suppose that you are estimating a simple linear regression model.

   (a) If you multiply all the x values by 20, but not the y values, what happens to the parameter values β_1 and β_2. What about the least squares estimates b_1 and b_2? What about the variance of the error term?

   (b) If you multiply all the y values by 50, but not the x values, what happens to the parameter values β_1 and β_2? What happens to the least squares estimates b_1 and b_2? What happens to the variance of the error term?