Lab 14 - Bivariate Regression & Interpretation

Jackson Nahpi

November 28, 2017

Complete the following exercises below and include all code used to find the answers. Knit together the PDF document and commit both the Lab 14 RMD file and the PDF document to Git. Push the changes to GitHub so both documents are visible in your public GitHub repository.

1. Select the main focal relationship you're interested in exploring for your poster project.

a. Describe the response variable and the explanatory variable and the theoretical relationship you believe exists between these two variables.

The response variable I want to use is the respondents percieved probability of finishing college and my explanatory variable will be the level of rewards they recieved due to their grades. I believe that there may be a positive relationship between the two variables since rewards for grades may influence a respondents work ethic and therefore may feel more inclined to try harder in college due to knowing their may be a payout to doing so.

b. Conduct a simple (bivariate) linear regression on your focal relationship and save the model object. Print out the full results by calling `summary()` on your model object.

```
##
## Call:
## lm(formula = probfincol ~ rwrdfrgrdslyr, data = freshman_dataset_v4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.8271  0.1729  0.2145  0.2561  0.2978
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.660600   0.029433 328.227  < 2e-16 ***
## rwrdfrgrdslyr 0.041627   0.009611   4.331 1.52e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7522 on 3647 degrees of freedom
## Multiple R-squared:  0.005117,   Adjusted R-squared:  0.004844
## F-statistic: 18.76 on 1 and 3647 DF,  p-value: 1.524e-05
```

c. What is the direction, magnitude, and statistical significance of the bivariate association between the explanatory and response variables.

The direction of the relationship is positive, the magnitude is fairly small. However, the association is significant out to the 0 level.

d. What is the meaning of the model intercept?

The intercept represents the expected value of y when x is equal to zero. Therefore, when x is zero in this model the expected value of y is 9.66.

e. How well does the bivariate model fit the data? How is this information calculated?

The bivariate model in this situation does not fit the data well at all. The R-squared value 0.005117 which essentially shows how poor of a fit this model is since it is near 0. The information about fit is calculated

through the R-squared this is enough to where we don't even need to look at the adjusted R-squared to confirm how poor the fit is. R-squared can be calculated by finding the sum of squared residuals and dividing that number by the total sum of squares, once this is done you can subtract that value from 1 and you should receive the R-squared value.

f. Is the observed association between the independent variable and dependent variable consistent with your hypothesis? Why or why not?

Yes the association is consistent with my hypothesis and we can reject the null hypothesis. Instead we can say both of the variables are related to one another. However, the fit of the model is of worry since it is so low essentially invalidating our rejection of the hypothesis.

2. Select a different focal relationship related to your project. This could be:

- **A different response and a different explanatory variable**
- **A different response and the same explanatory variable**
- **The same response and a different explanatory variable**

a. Describe the response variable and the explanatory variable and the theoretical relationship you believe exists between these two variables.

The response variable I am using for this question is similar to the one used in question 1 but it is the percieved probability of acquiring a graduate degree. My explanatory variable will be the same as the one in question 1. The believed relationship is that due to more rewards that come from schooling, respondents may pursue further education due to potential to futher the rewards they reap.

b. Conduct a simple (bivariate) linear regression on your focal relationship and save the model object. Print out the full results by calling `summary()` on your model object.

```
##
## Call:
## lm(formula = probfingraddeg ~ rwrdfrgrdslyr, data = freshman_dataset_v4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.2378 -1.0797  0.2365  1.9203  2.3946
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    7.44733    0.08562  86.977  < 2e-16 ***
## rwrdfrgrdslyr  0.15809    0.02796   5.654 1.69e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.188 on 3647 degrees of freedom
## Multiple R-squared:  0.008689,   Adjusted R-squared:  0.008418
## F-statistic: 31.97 on 1 and 3647 DF,  p-value: 1.687e-08
```

c. What is the direction, magnitude, and statistical significance of the bivariate association between the explanatory and response variables.

The direction of the relationship is positive. The magnitude and stastical significance are both greater than the previous model in question 1. Just as in the first question the association of the model is significant at the 0 level but has a lower probability of being up to chance and a greater t value than the model used in question 1.

d. What is the meaning of the model intercept?

The model intercept is representive of what the estimated y would be if x is zero. In this question the estimated value of y when x is zero is 7.44733.

e. How well does the bivariate model fit the data? How is this information calculated?

This model has a similar issue as in question one where the R-squared is very low (0.008689), due to having a low R-squared it is hard to reject the null hypothesis. R-squared is calculated by dividing the sum of squared residuals by the total sum of squares and then subtracting the value you get from 1.

f. Is the observed association between the independent variable and dependent variable consistent with your hypothesis? Why or why not?

The significance is relevant among the relationship I hypothesized but it is hard to reject the null. Our low R-squared values devaluates the significance of the results since our model fit is not even capable of explaining 1 percent of the variance we see in the model. Therefore, we cannot safely reject the null hypothesis.