Lab 14 - Bivariate Regression & Interpretation

Ramin Jabbarialghanab

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 explanatory variable for the focal relationship which I am interested in is gender, and the reponse variable is value which represent unemployment rate in the data. The theoretical aspect of the relationship is the gender inequality in the society. Since there are less job opportunities in Iran and the society is a patriarchal society, men tend to take more job opportunitied than women.

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.

```
library(readxl)
joint_unemployment_total <- read_excel("~/Desktop/Autumn 2017/Statistics 321/unemployment rate/joint_un
model1 <-lm(value ~ gender, data = joint_unemployment_total)
```

```
summary(model1)
```

```
##
## Call:
## lm(formula = value ~ gender, data = joint_unemployment_total)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7111 -1.4917  0.2833  0.9639  2.8333
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  10.3667     0.5533   18.73 2.61e-12 ***
## genderwomen   8.5444     0.7825   10.92 7.98e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.66 on 16 degrees of freedom
## Multiple R-squared:  0.8817, Adjusted R-squared:  0.8743
## F-statistic: 119.2 on 1 and 16 DF,  p-value: 7.975e-09
```

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

since 8.5444 is positive the direction is positive, the magnitude is not small, and statistically is significant because of 7.98e-09 ***

Based on multiple R-squared 88% of the variation in umeployment rate can be explained by our gender difference in our model. the direction is positive and the association between gender and unemployment rate is strong. The statistical significance of the association based on p-values is 7.975e-09. It is not less then 5%, but still I think that this statistically somehow significant.

d. What is the meaning of the model intercept? That means that y intercept of the model is 10.3667. Intercept repserent the mean of Y when X= 0. Since y is the unemployment rate and X is the gender and it is not reseanable to assume gender equal zero, the intercept has no intrinsic meaning here.

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

This model explains 88% of variation. So the model fits. The model fits as well because root-mean-square error (RMSE) or residual standard error is low (1.66). RMSE is the standard deviations of the residuals which we divide by the degree of freedom which here is 16.

Another way, is computing SSE for Sum of Squared Errors.It can be computed as the variance of the residuals multiplied by 1 fewer than the number of observations.The SSE is a single number that captures how much our model missed by. and SST is a measure of the overall variability in the response variable and then using the forumula R^2 = 1- SSE/SST we can calcuated how the bivariate model fit.

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

Yes, the association consists with my hyphothesis that gender inequlaity leads to domination of men over women on job opportunities. This model explains 88% of the variation of unemployment rate. So, this model consist with my observation.

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

```
model2 <-lm(value ~ year, data = joint_unemployment_total)
```

```
summary(model2)
```

```
##
## Call:
## lm(formula = value ~ year, data = joint_unemployment_total)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.8178 -4.0682 -0.2906  5.1061  6.3616
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 135.77448  605.29153   0.224    0.825
## year         -0.06029    0.30124  -0.200    0.844
##
## Residual standard error: 4.82 on 16 degrees of freedom
## Multiple R-squared:  0.002497,   Adjusted R-squared:  -0.05985
## F-statistic: 0.04005 on 1 and 16 DF,  p-value: 0.8439
```

- **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 explanotory variable for here is year and the response variable is value which shows unemployment rate by gender. I picked these variables just to this practice. So, I would not is there any difference by year.

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.

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

Direction is negative and magnitude is small (-0.06029), and statistally is not significant (0.844).

d. What is the meaning of the model intercept?

The y intercept of the model is 135.77448. Intercept repserent the mean of Y when X= 0. Since y is the unemployment rate and X is the year and it is not reseanable to assume year equal zero in my data, the intercept has no intrinsic meaning.

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

One way to know that the model fits or not is to consider R-squrated. The R-squared is 0.002497 which is almost shows nothing. So that does not fit.

Alternative way, could be looking at RMSE. I think this model does not fit here because root-mean-square error (RMSE) or residual standard error is not low (4.82). RMSE is the standard deviations of the residuals which we divide by the degree of freedom which here is 16.

Another way, is computing SSE for Sum of Squared Errors.It can be computed as the variance of the residuals multiplied by 1 fewer than the number of observations.The SSE is a single number that captures how much our model missed by. and SST is a measure of the overall variability in the response variable and then using the forumula Rˆ2 = 1- SSE/SST we can calcuated how the bivariate model fit.

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

No, that does not consist because there is almost no assication between year and unemployment. The model explians almost no variation(0.002497).