DSC 520 Week 6 Linear Regression Assignment

Dan Wiltse

Jan 19th 2020

## 1st Research Question

“Is there a significant relationship between the number of siblings a survey respondent has and number of his or her children?”

## Part 1

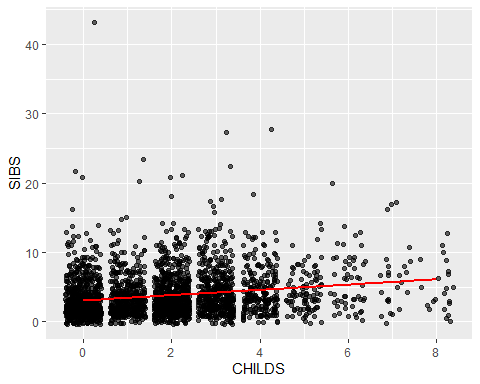
## a. Construct a scatterplot of these two variables in R studio and place the best-fit linear regression line on the scatterplot. Describe the relationship between the number of siblings a respondent has (SIBS) and the number of his or her children (CHILDS).

As the scatterplot below shows, there appears to be a slight positive relationship between the number of children a respondent has and the number of siblings they have, as shown by the best fit line.

ggplot(data = gss\_2016, aes(x = CHILDS, y = SIBS)) +  
 geom\_point(position = "jitter", alpha = 0.6) +  
 stat\_smooth(method = "lm", col = "red", se = FALSE)

## Warning: Removed 11 rows containing non-finite values (stat\_smooth).

## Warning: Removed 11 rows containing missing values (geom\_point).



## b. Use R to calculate the covariance of the two variables and provide an explanation of why you would use this calculation and what the results indicate.

As seen below, the covariance is 1.06. You would use this calculation to see how two variables vary together. You have to be careful with interpretations of covariance, as different scales for the variables will impact the results, but the positive covariance shows that there is at least a slight positive relationship between the two variables, as the best fit line also shows above.

cov(gss\_2016$CHILDS, gss\_2016$SIBS, use = "complete.obs")

## [1] 1.064853

## c. Choose the type of correlation test to perform, explain why you chose this test, and make a prediction if the test yields a positive or negative correlation?

I will use Pearson’s R correlation because the data is normally distributed and measured at interval ratio, which are the requirements of using Pearson’s R. I predict there will be a slightly positive correlation between number of children and number of siblings.

## d. Perform a correlation analysis of the two variables and describe what the calculations in the correlation matrix suggest about the relationship between the variables. Be specific with your explanation.

The correlation between the two variables is .198, with a t value of 10.84. As shown in the best fit line and covariance, there appears to be a slightly significant relationship between number of siblings a respondent had and the number of children they had.

cor.test(gss\_2016$CHILDS, gss\_2016$SIBS)

##   
## Pearson's product-moment correlation  
##   
## data: gss\_2016$CHILDS and gss\_2016$SIBS  
## t = 10.84, df = 2854, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.1633721 0.2338305  
## sample estimates:  
## cor   
## 0.1988582

## e. Calculate the correlation coefficient and the coefficient of determination, describe what you conclude about the results.

See f below

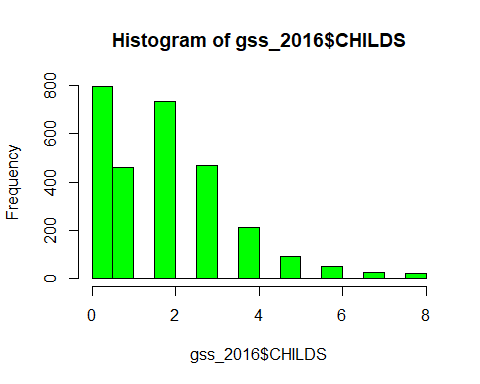
## f. Based on your analysis, what can you say about the relationship between the number of siblings and the number of his or her children?

The correlation coefficient is r = .198 and the coefficient of determination is .039, which is the squared value of the correlation coefficient. This shows that number of children is positively correlated with number of siblings, and the number of siblings accounts for approximaetely 4% of the variation in number of children a respondent had, so other factors made up 96% of the remaining variation in number of children.

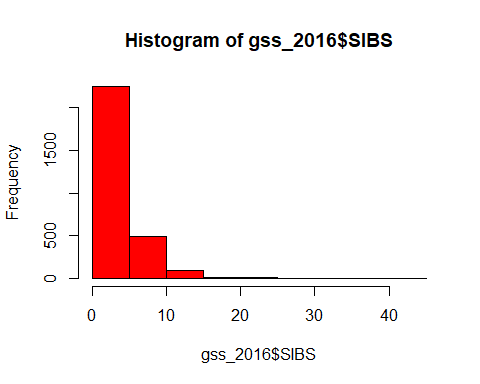
## g. Produce an appropriate graph for the variables. Report, critique and discuss the skewness and any significant scores found.

Looking at a historgram, you can see most of the data for siblings is skewed to the left end of the chart, with most people having less than 10 siblings, while the data for children is more spread out, as while the data is definitely skewed to the left of the histogram, there are also a wider range of children than siblings in the graphs.

hist(gss\_2016$CHILDS, col="green")



hist(gss\_2016$SIBS, col="red")



## h. Expand your analysis to include a third variable – Sex. Perform a partial correlation, “controlling” the Sex variable. Explain how this changes your interpretation and explanation of the results.

The partial correlation between children and siblings, while controlling for Sex, was .197. This has a R squared value of .038, which is means that number of siblings accounts for 3.8% of the number of children the respondent had, when controlling for gender. This is also very similar to the correlation between siblings and children without accounting for Sex.

library("ggm")  
gss\_pcorvars <- gss\_2016[,c("SIBS","CHILDS","SEX")]  
gss\_pcor\_NONA <- gss\_pcorvars[complete.cases(gss\_pcorvars),]  
pcgss<-pcor(c("CHILDS", "SIBS", "SEX"), var(gss\_pcor\_NONA))  
pcgss

## [1] 0.1970057

pcgss^2

## [1] 0.03881125

pcor.test(pcgss,1,2866) $pvalue

## [1] 1.844857e-26

# Part 2

## a. Run a regression analysis where SIBS predicts CHILDS.

See below

## b. What are the intercept and the slope? What are the coefficient of determination and the correlation coefficient?

The intercept is 1.4678 and the slope is .1036.

#Linear Regression  
lm(CHILDS~SIBS, data = gss\_2016)

##   
## Call:  
## lm(formula = CHILDS ~ SIBS, data = gss\_2016)  
##   
## Coefficients:  
## (Intercept) SIBS   
## 1.4678 0.1036

See description of correlations in part c

## c. For this model, how do you explain the variation in the number of children someone has? What is the amount of variation not explained by the number of siblings?

The intercept is 1.46, meaning that if the respondent has no silblings, the model predicts that they would have 1.46 children. For each additional sibling, it would predict an additional .10 children the respondent would have.

As described in Part 1, correlation coefficient is r = .198 and the coefficient of determination is .039, which is the squared value of the correlation coefficient. This shows that number of children is positively correlated with number of siblings, and the number of siblings accounts for approximaetely 4% of the variation in number of children a respondent had, so other factors made up 96% of the remaining variation in number of children.

## d. Based on the calculated F-Ratio does this regression model result in a better prediction of the number of children than if you had chosen to use the mean value of siblings?

The F value for this regresssion model of 117.5, which is significant at the .001 level. This tells us that there is less than a 0.1% chance that an F value this large would happen than if null hypothesis was true, meaning the prediction is signifcantly better than if we simply used the mean

#Linear regression with details  
gss\_lm <-lm(CHILDS~SIBS, data = gss\_2016)  
summary(gss\_lm)

##   
## Call:  
## lm(formula = CHILDS ~ SIBS, data = gss\_2016)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.9216 -1.5713 0.0143 1.0143 6.5322   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.467767 0.046889 31.30 <2e-16 \*\*\*  
## SIBS 0.103577 0.009555 10.84 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.637 on 2854 degrees of freedom  
## (11 observations deleted due to missingness)  
## Multiple R-squared: 0.03954, Adjusted R-squared: 0.03921   
## F-statistic: 117.5 on 1 and 2854 DF, p-value: < 2.2e-16

## e. Use the model to make a prediction: What is the predicted number of children for someone with three siblings?

A person with 3 siblings would be predicted to have 1.78 children, the interept of 1.47, which predicts the number of children if they had no siblings, plus .1036 \* 3 (.3108)

gss\_lm <-lm(CHILDS~SIBS, data = gss\_2016)  
new.df <- data.frame(SIBS = 3)   
predict(gss\_lm, new.df)

## 1   
## 1.778498

## f. Use the model to make a prediction: What is the predicted number of children for someone without any siblings?

A person with no siblings would be predicted to have the same number of children as the intercept, which is 1.46.