6.1 GSS 2016 Survey Data

Brandon Sams

1/19/2020

Overview

Work with your previously assigned team members on this assignment. Data for this assignment originated from the General Society Survey (GSS). The GSS gathers data on contemporary American society in order to monitor and explain trends and constants in attitudes, behaviors, and attributes. Hundreds of trends have been tracked since 1972. In addition, since the GSS adopted questions from earlier surveys, trends can be followed for up to 70 years. Only data from the 2016 GSS survey is included in this dataset – GSS2016.csv.

If you are interested in getting at a different year or a cumulative dataset you can visit http://www.gss.norc.org

For this assignment, you will need to load and activate the ggplot2 package. You are encouraged to complete the assigned reading and DataCamp exercises before starting their work. Each of you should first produce your own version of the following deliverables then share with your team members. Team members will then collaborate on their approach and insights, refining their work before submitting for a grade.

As a data science intern with newly learned knowledge in skills in statistical correlation, regression and R programming, you are interested in looking at the GSS 2016 survey data, specifically the Siblings and Childs variables have peaked your interest. A codebook for the GSS is available here:

GSS_Codebook_Index.pdf and contains all of the GSS variables and descriptions. The first question you are interested in answering is: **"Is there a significant relationship between the number of siblings a survey respondent has and number of his or her children?"**

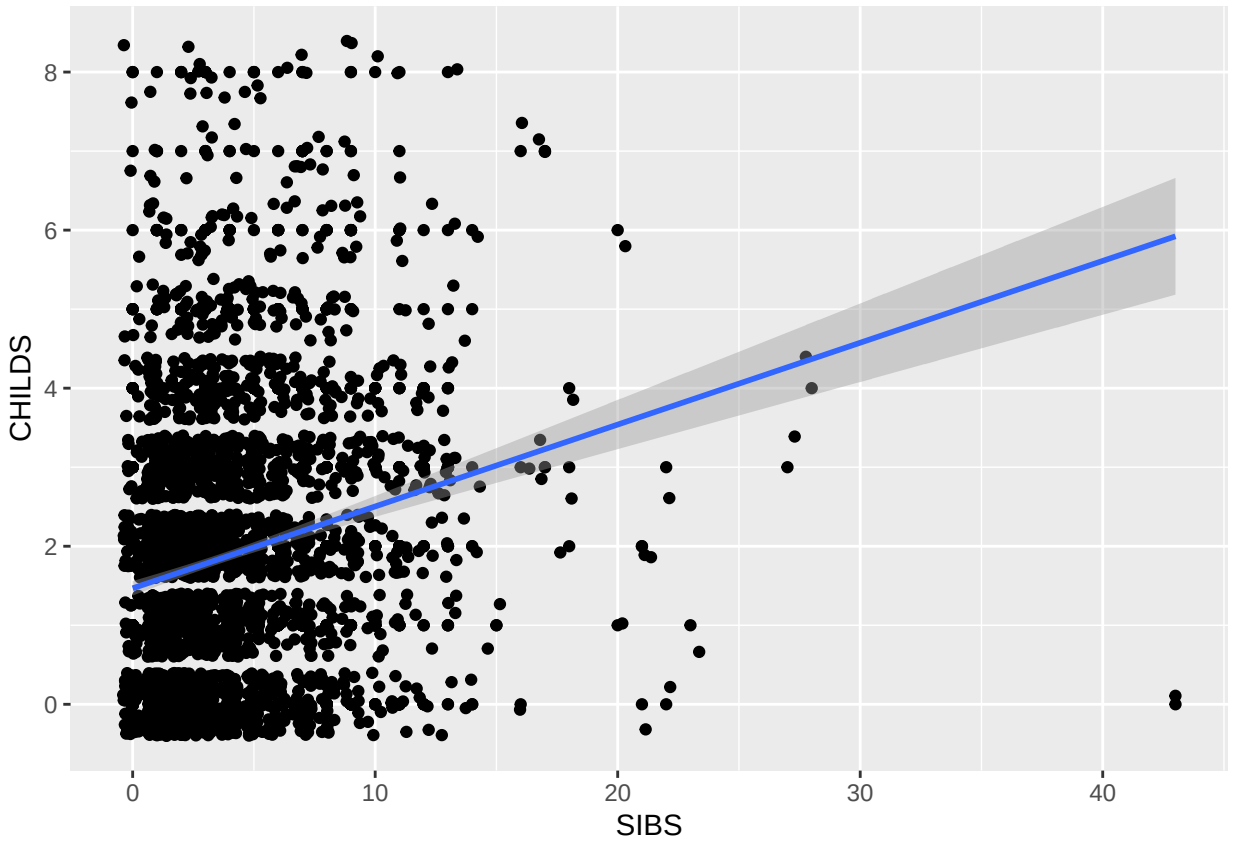
The following guidelines describe minimum deliverables needed for the assignment submission.

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).

```
gss_2016_limited <- na.omit(gss_2016[c("SIBS","CHILDS")])

ggplot(data = gss_2016_limited,aes(x = SIBS,y = CHILDS)) + geom_point() + geom_jitter() + geom_smooth(m
```



The number of siblings that a person has is positively correlated with the number of children that they have.

(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.

```
cov(gss_2016_limited)
```

```
##              SIBS   CHILDS
## SIBS   10.280759 1.064853
## CHILDS  1.064853 2.789121
```

The covariance calculation is used to see how, as one variable changes, other other variable in question changes with it. The results show that as you start to look at respondents that have more siblings, the likelyhood that they will also have more children increases as well. You can see this becuase the covariance between SIBS and CHILDS is a positive value.

(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?

This correlation test, which found Spearman's Rho, is a useful test when data is distributed in a non-normal way. There seems to be a significant amount of data skewed to the lower right section of the plot. Most people do not have a large number of siblings or children, just a few.

I expect that the correlation will be positive, based on the line of best fit. However, I think it will be small.

(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.

```
cor.test(gss_2016_limited$SIBS,gss_2016_limited$CHILDS,method = "spearman")
```

```
## Warning in cor.test.default(gss_2016_limited$SIBS, gss_2016_limited$CHILDS, :
```

```
## Cannot compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  gss_2016_limited$SIBS and gss_2016_limited$CHILDS
## S = 3047182852, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.2151707
```

```
cor(gss_2016_limited$SIBS,gss_2016_limited$CHILDS)
```

```
## [1] 0.1988582
```

The value for Spearman's Rho was small, but positive. The correlation matrix shows very similar behavior. This means that there is a relationship between the two variables, but it is a small relationship.

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

```
cov(gss_2016_limited$SIBS,gss_2016_limited$CHILDS)/(sd(gss_2016_limited$SIBS) * sd(gss_2016_limited$CHI
```

```
## [1] 0.1988582
```

```
summary(lm(CHILDS ~ SIBS, data = gss_2016_limited))$r.squared
```

```
## [1] 0.03954459
```

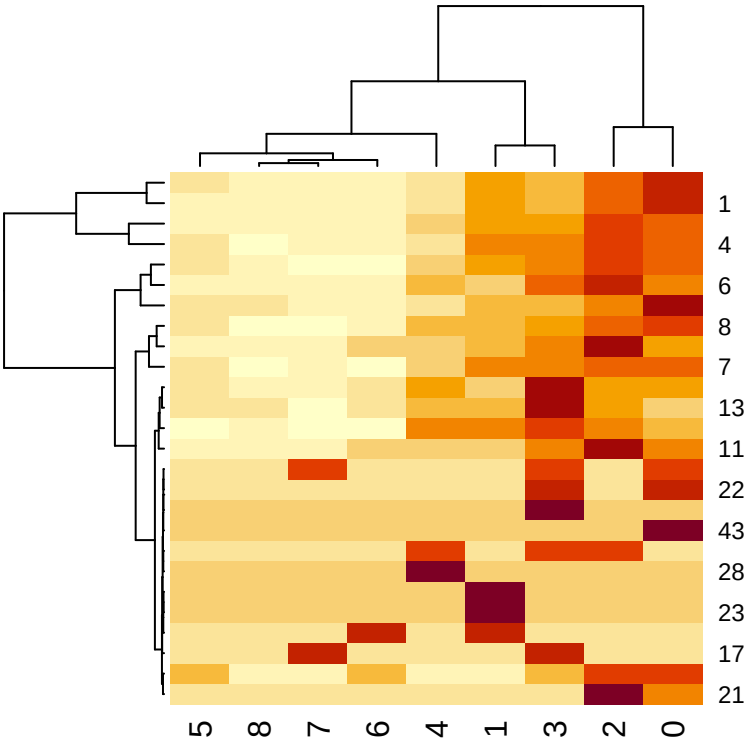
The correlation coefficient is small, but positive. This is in line with the correlation values observed in the previous section. The correlation of determination, however, is near zero. This tells me that the correlation is not very strong at all.

(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?

There is not a very strong correlation between the number of siblings and the number of children a person has. There is a positive correlation, but it is not a very strong correlation.

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

```
heatmap(xtabs(formula = ~SIBS+CHILDS,data = gss_2016_limited))
```



This graph shows that there is a signicant clustering of occurences where a respondent has only a few children and only a few siblings. Sure, there are some instances where a person has a lot of siblings, but they are very few.

(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.

```
library(ggm)
```

```
## Loading required package: igraph
```

```
##
## Attaching package: 'igraph'
```

```
## The following objects are masked from 'package:stats':
##
##     decompose, spectrum
```

```
## The following object is masked from 'package:base':
##
##     union
```

```
##
## Attaching package: 'ggm'
```

```
## The following object is masked from 'package:igraph':
##
##     pa
```

```
gss_2016_sex <- na.omit(gss_2016[c("SIBS","CHILDS","SEX")])

pcor(c("SIBS", "CHILDS", "SEX"), var(gss_2016_sex))
```

```
## [1] 0.1970057
```

PART 2

a Run a regression analysis where SIBS predicts CHILDS.

```
linearModel <- lm(CHILDS ~ SIBS, data = gss_2016_limited)
summary(linearModel)
```

```
##
## Call:
## lm(formula = CHILDS ~ SIBS, data = gss_2016_limited)
##
## 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
## Multiple R-squared:  0.03954,    Adjusted R-squared:  0.03921
## F-statistic: 117.5 on 1 and 2854 DF,  p-value: < 2.2e-16
```

b What are the intercept and the slope? What are the coefficient of determination and the correlation coefficient? The intercept is at 1.46, and the slope is 0.1036.

The coefficient of determination is 0.3954. The correlation coefficient is 1.637

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 variation is explained by the binned-ness of a person only being able to have a whole number of shildren and siblings.

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 Ratio does not result in a better prediciton of the number of children.

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

```
predict(linearModel,data.frame(SIBS = 3))
```

```
##        1
## 1.778498
```

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

```
predict(linearModel,data.frame(SIBS = 1))
```

```
##        1
## 1.571344
```