General Society Survey Data Analysis

Arindam Samanta

January 19 2020

## Part 1&2: Is there a significant relationship between the number of siblings a survey respondent has and number of his or her children?”

### General Society Survey data

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](http://content.bellevue.edu/cst/dsc/520/id/resources/gss-2016.csv).

## Preparations

We have used the following libraries for our regression analysis. Regression analysis is used in stats to find trends in data. For example, we are guessing that there is a connection between how many child a respondant has and how many siblings he or she has; regression analysis can help us quantify that.

library(dplyr)  
library(ggplot2)  
library(readr)  
library(corrplot)

## Loading data to include the SIBS,CHILDS and SEX variables from the actual data set

As we are interested here only in the above features in our data set.

#loading data into dataframe  
mycols <- cols\_only(  
 CHILDS = col\_double(),  
 SIBS = col\_double(),  
 SEX = col\_factor()  
 )  
wd <- getwd()  
fname <- "gss-2016.csv"  
path\_to\_file <- paste(wd,'/dataset/',fname, sep = "")  
path\_to\_file  
  
survey\_data <- read\_csv(path\_to\_file)

## Parsed with column specification:  
## cols(  
## .default = col\_double(),  
## MAR9 = col\_logical(),  
## MAR10 = col\_logical(),  
## MAR13 = col\_logical(),  
## MAR14 = col\_logical(),  
## AGE3 = col\_logical(),  
## AWAY13 = col\_logical(),  
## GENDER10 = col\_logical(),  
## GENDER13 = col\_logical(),  
## GENDER14 = col\_logical(),  
## OLD10 = col\_logical(),  
## OLD13 = col\_logical(),  
## OLD14 = col\_logical(),  
## RELATE10 = col\_logical(),  
## RELATE13 = col\_logical(),  
## RELATE14 = col\_logical(),  
## RELHH10 = col\_logical(),  
## RELHH13 = col\_logical(),  
## RELHH14 = col\_logical(),  
## RELHHD10 = col\_logical(),  
## RELHHD13 = col\_logical()  
## # ... with 4 more columns  
## )

## See spec(...) for full column specifications.

## Warning: 36 parsing failures.  
## row col expected actual file  
## 1280 MAR9 1/0/T/F/TRUE/FALSE 5 '/media/samantoz/storage/arindam/github-repo/bu-datascience/dataset/gss-2016.csv'  
## 1291 MAR13 1/0/T/F/TRUE/FALSE 5 '/media/samantoz/storage/arindam/github-repo/bu-datascience/dataset/gss-2016.csv'  
## 1291 OLD13 1/0/T/F/TRUE/FALSE 19 '/media/samantoz/storage/arindam/github-repo/bu-datascience/dataset/gss-2016.csv'  
## 1291 WHERE13 1/0/T/F/TRUE/FALSE 3 '/media/samantoz/storage/arindam/github-repo/bu-datascience/dataset/gss-2016.csv'  
## 1901 RELSP12 1/0/T/F/TRUE/FALSE 12 '/media/samantoz/storage/arindam/github-repo/bu-datascience/dataset/gss-2016.csv'  
## .... ....... .................. ...... .................................................................................  
## See problems(...) for more details.

gss2016.sub <- read\_csv(path\_to\_file,   
 col\_types = mycols)

For this analysis we are interested in looking at the GSS 2016 survey data, specifically the Siblings and Childs variables.

## # A tibble: 6 x 3  
## CHILDS SEX SIBS  
## <dbl> <fct> <dbl>  
## 1 3 1 2  
## 2 0 1 3  
## 3 2 1 3  
## 4 4 2 3  
## 5 2 2 2  
## 6 2 2 2

Summaize the data based on CHILDS to see how the distribution looks

## # A tibble: 10 x 2  
## CHILDS `n()`  
## <dbl> <int>  
## 1 0 797  
## 2 1 459  
## 3 2 733  
## 4 3 467  
## 5 4 213  
## 6 5 92  
## 7 6 51  
## 8 7 25  
## 9 8 22  
## 10 NA 8

Summaize the data based on SIBS to see how the distribution looks

## # A tibble: 27 x 2  
## SIBS `n()`  
## <dbl> <int>  
## 1 0 130  
## 2 1 550  
## 3 2 596  
## 4 3 452  
## 5 4 319  
## 6 5 206  
## 7 6 161  
## 8 7 128  
## 9 8 90  
## 10 9 72  
## # … with 17 more rows

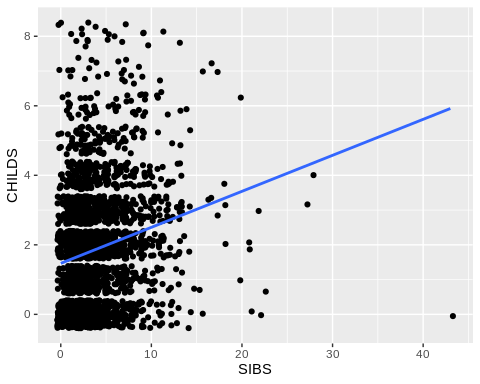
The summarized data shows that there are 8 rows that has Child as NA and 5 rows that has SIBS as NA.

### Correlation Coefficient (Pearson Product moment Correlation)

The covarience between the two variables SIBS and CHILD is 0.1988582

The relationship shows a weak positive trend,irrespective on how we measure the variables which means that No. of siblings do not have any relationship to the number of child a respondant has. This could also be seen in the plot below and fitting a line. ## Including Plots Now 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).

# Creating a scatter plot with the survey data  
ggplot(data = gss2016.sub, aes(x = SIBS, y = CHILDS)) +  
   
 geom\_point(position = "jitter") +  
 geom\_smooth(method = "lm", se = FALSE)



## Expanding the analysis to bring in the third variable

Generating a scatter plot with the third variable SEX of the respondant.

# Creating a scatter plot with the third variable SEX survey data  
ggplot(data = gss2016.sub, aes(x = SIBS, y = CHILDS, color = SEX)) +  
   
 geom\_point(position = "jitter") +  
 geom\_smooth(method = "lm", se = FALSE)



## Trying to fit a linear model

In this model we are trying to see if the # of SIBS could predict the # of CHILD.The intercept and the slope of the model is as below. Given below is the coefficient values and summary of the model.

## (Intercept) SIBS   
## 1.4677665 0.1035773

##   
## Call:  
## lm(formula = fmla, data = gss2016.sub)  
##   
## 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

Trying to predict using the model.

# Created a new dataframe with 3 SIBS  
new\_df3 <- data.frame("SIBS" = 3 )  
# Predicted no. of CHILDS  
pred3 <- predict(mod, newdata = new\_df3)  
pred3  
  
# Created a new dataframe with 0 SIBS  
new\_df0 <- data.frame("SIBS" = 0 )  
# Predicted no. of CHILDS  
pred0 <- predict(mod, newdata = new\_df0)  
pred0

Predicted number of children for someone with three siblings :1.7784984

Predicted number of children for someone with no siblings :1.4677665