QMSSGR5015- Lab 2- Yue Ma

Yue Ma

2018/9/29

```
setwd("/Users/mayue/Desktop/qmss/data analysis/lab 2")
d = read.csv("GSS.2006.csv")
```

1. Recode 1 *sort of* continuous variable into categories. Tell what you did and explain the variable(s).

A. The simplest way to make a dummy variable:

It is binary recode, where we make it 1 if "How old were you when your first child was born?" is age as old as 35 or more, 0 otherwise. People who are pregnant or have first child at age 35 or older are often referred to as "advanced maternal age". From the results, I can find that most people have their first child under advanced maternal age.

```
d$hi.agekdbrn = ifelse((d$agekdbrn>34), 1, 0)
table(d$hi.agekdbrn, d$agekdbrn)
```

```
##
##      13  14  15  16  17  18  19  20  21  22  23  24  25  26  27  28  29
##   0   2   8  22  62  89 134 156 201 216 148 149 120 132 122  91  80  84
##   1   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
##
##      30  31  32  33  34  35  36  37  38  39  40  41  42  43  44  45  46
##   0  79  42  49  31  21   0   0   0   0   0   0   0   0   0   0   0   0
##   1   0   0   0   0   0  21  23  15  13   6  10   4   5   3   2   1   1
##
##      47  48  50  51  53  56
##   0   0   0   0   0   0   0
##   1   1   1   1   1   1   1
```

B. Breaking a variable into categories:

To create a number of categories for first maternal age, people who have first child at age 35 or less are often reffered to as "young maternal age", people who have first child at age between 18 and 35 are often seen as normal maternal age, those who have first child at age 35 or older are often referred to as "advanced maternal age". For the results, I can conclude that most people have their first child at normal maternal age.

```
d$agekdbrn.cat = cut(d$agekdbrn, breaks = c(1, 18, 34, 98), label=c("young maternal age", "normal matern
table(d$agekdbrn.cat, d$agekdbrn)
```

```
##
##                          13  14  15  16  17  18  19  20  21  22  23  24
##   young maternal age      2   8  22  62  89 134   0   0   0   0   0   0
##   normal maternal age     0   0   0   0   0   0 156 201 216 148 149 120
##   advanced maternal age   0   0   0   0   0   0   0   0   0   0   0   0
##
##                          25  26  27  28  29  30  31  32  33  34  35  36
```

```
##   young maternal age      0   0   0   0   0   0   0   0   0   0   0   0
##   normal maternal age   132 122  91  80  84  79  42  49  31  21   0   0
##   advanced maternal age   0   0   0   0   0   0   0   0   0   0  21  23
##
##                          37  38  39  40  41  42  43  44  45  46  47  48
##   young maternal age      0   0   0   0   0   0   0   0   0   0   0   0
##   normal maternal age     0   0   0   0   0   0   0   0   0   0   0   0
##   advanced maternal age  15  13   6  10   4   5   3   2   1   1   1   1
##
##                          50  51  53  56
##   young maternal age      0   0   0   0
##   normal maternal age     0   0   0   0
##   advanced maternal age   1   1   1   1
```

C. Coding multiple conditions:

It is a table with multiple conditions at once, both female and currently married.

```
d$bothftw = ifelse((d$marital==1 & d$sex==2), 1, 0)
table(d$bothftw, d$marital, d$sex)
```

```
## , ,  = 1
##
##
##        1    2    3    4    5
##   0 1018   65  320   61  534
##   1    0    0    0    0    0
##
## , ,  = 2
##
##
##        1    2    3    4    5
##   0    0  301  412   95  546
##   1 1152    0    0    0    0
```

D. Another way to apply multiple labels. It is consistent with the original.

```
d$sex[d$marital==1] <- 0
d$sex[d$marital==2] <- 1
d$sex[d$marital==3] <- 1
d$sex[d$marital==4] <- 1
d$sex[d$marital==5] <- 1
table(d$sex, d$marital)
```

```
##
##        1    2    3    4    5
##   0 2170    0    0    0    0
##   1    0  366  732  156 1080
##   2    0    0    0    0    0
```

E. Changing to missing values

The answers to "Were both your parents born in this country??" and 7="Neither born in U.S.", so I want to make that a missing answer.

```
d$parborn.new = d$parborn
d$parborn.new[d$parborn==8] <- NA

table(d$parborn, d$parborn.new)
```

```
##
##        0    1    2    3    4    5    6    7
##   0 2360    0    0    0    0    0    0    0
##   1    0   91    0    0    0    0    0    0
##   2    0    0   55    0    0    0    0    0
##   3    0    0    0    6    0    0    0    0
##   4    0    0    0    0    5    0    0    0
##   5    0    0    0    0    0    1    0    0
##   6    0    0    0    0    0    0    1    0
##   7    0    0    0    0    0    0    0    1
##   8    0    0    0    0    0    0    0    0
```

2. Recode 1 other variable and attach value labels. Tell what you did and explain the variable(s).

A. Add labels to existing variables:

The results show that most people have their first baby in non-advanced maternal age.

```
d$hi.agekdbrn.lab <- ordered(d$hi.agekdbrn, levels = c(0,1), labels = c("non-advanced maternal age", "a
table(d$hi.agekdbrn.lab, d$hi.agekdbrn)
```

```
##
##                                0    1
##   non-advanced maternal age 2038    0
##   advanced maternal age        0  110
```

B. Reverse code a variable and then add labels and make it ordered:

The variable "natenvir" indicates the fact that how much time the respondents spend on improving and protecting the enviroment. (too much, too little or about the right amount time.)

I reverse code firstm, make the numeric variable into a factor, make the factor variable into an ORDERED factor with value labels, and get the mean.

```
d$rnatenvir = 4-d$natenvir

d$rnatenvir.fact = as.factor(d$rnatenvir)

d$lab.rnatenvir <- ordered(d$rnatenvir, levels = c(1,2,3), labels = c("too much", "all-right", "too lit

table(d$lab.rnatenvir, d$rnatenvir)
```

```
##
##                1   2   3
```

```
##   too much    89   0   0
##   all-right    0 365   0
##   too little   0   0 992
```

```
mean(d$natenvir, na.rm=T)
```

```
## [1] 1.375519
```

```
mean(as.numeric(d$lab.rnatenvir), na.rm=T)
```

```
## [1] 2.624481
```

3. Use one (or both) of your recoded variables to do a cross-tabulation (like last week, with prop.table, doBy, or ddply). Explain your results.

The variable "agekdbrn" indicates the age people have thier first baby.

The variable "childs" indicates the how many children have you ever had. People who have less than 3 children are coded 1, otherwise 0.

The results indicate that there is no obivious relation between the number of childs people have and the age that they have their first baby.

```
library(gmodels)
d$hi.agekdbrn = ifelse((d$agekdbrn>34), 1, 0)
d$hi.childs = ifelse((d$childs<3), 1, 0)
CrossTable(d$hi.agekdbrn, d$hi.childs, prop.r=F, prop.c=T, prop.t=F, prop.chisq=F, format="SPSS")
```

```
##
##    Cell Contents
## |-------------------------|
## |                   Count |
## |          Column Percent |
## |-------------------------|
##
## Total Observations in Table:  2148
##
##                | d$hi.childs
## d$hi.agekdbrn |        0  |        1  | Row Total |
## --------------|-----------|-----------|-----------|
##             0 |      913  |     1125  |     2038  |
##               |   98.596% |   92.062% |           |
## --------------|-----------|-----------|-----------|
##             1 |       13  |       97  |      110  |
##               |    1.404% |    7.938% |           |
## --------------|-----------|-----------|-----------|
##  Column Total |      926  |     1222  |     2148  |
##               |   43.110% |   56.890% |           |
## --------------|-----------|-----------|-----------|
##
##
```

4. Run a linear regression with 1 independent and 1 dependent variable; make all of the recodes necessary to make the model as easy to interpret as possible; and explain the results.

The variable "sibs" is about the question - How many brothers and sisters did you have?

The variable "childs" is about the question - How many children have you ever had?

From the description results, we can see that the average number of brothers and sisters people have is about 4, and the average number of the children people have is about 2.

I use a linear regression to see the relationship bewteen the number of one's sibling and children. From the results, a coefficient of 0.43514 indicates there is a positive correlation between the number of one's sibling and the number of one's children.

```
library(psych)
describe(d$sibs)
```

```
##    vars    n mean   sd median trimmed  mad min max range skew kurtosis
## X1    1 2988 3.76 3.18      3    3.29 2.97   0  34    34 2.19    10.26
##      se
## X1 0.06
```

```
describe(d$childs)
```

```
##    vars    n mean   sd median trimmed  mad min max range skew kurtosis
## X1    1 4497  1.9 1.68      2    1.71 1.48   0   8     8  0.9     0.86
##      se
## X1 0.03
```

```
lm1 = lm(sibs ~ childs, data=d)
summary(lm1) ## examine the results: a coefficient of 0.04662 indicates no obvious relationship between
```

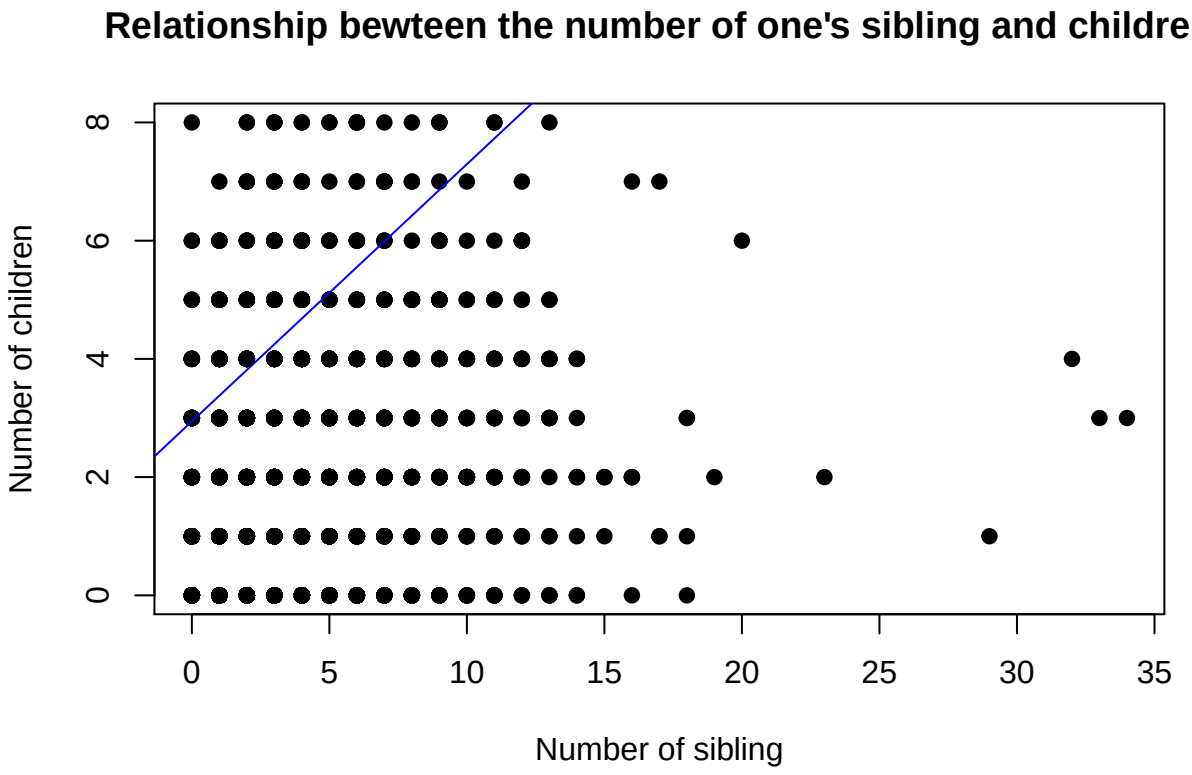
```
##
## Call:
## lm(formula = sibs ~ childs, data = d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.4234 -1.9423 -0.8126  1.1874 29.7523
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.94231    0.08498   34.62   <2e-16 ***
## childs       0.43514    0.03373   12.90   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.094 on 2981 degrees of freedom
##   (1527 observations deleted due to missingness)
## Multiple R-squared:  0.05289,    Adjusted R-squared:  0.05257
## F-statistic: 166.5 on 1 and 2981 DF,  p-value: < 2.2e-16
```

5. Plot two variables, either as a scatter plot or boxplot; add in trend/regression lines; and explain your results.

From the plot, there is a positive correlation between the number of one's sibling and the number of one's children.

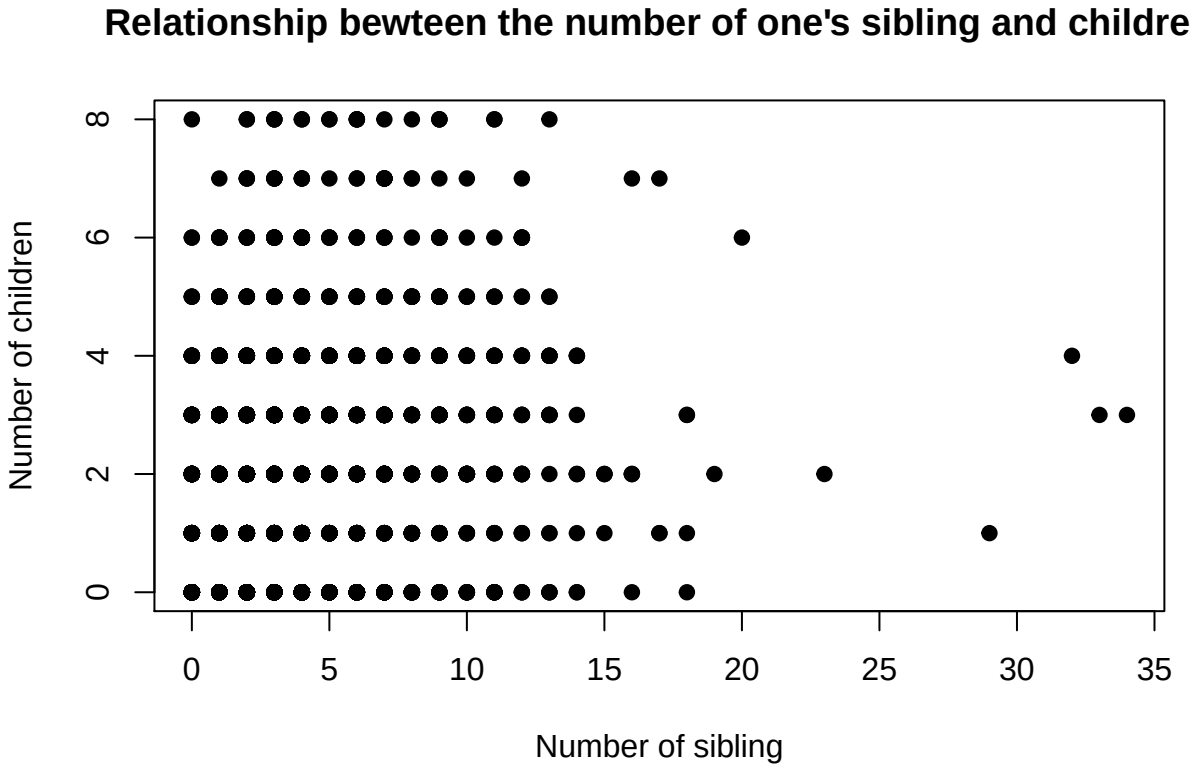
```
plot(d$sibs, d$childs, main="Relationship bewteen the number of one's sibling and children",
     xlab="Number of sibling", ylab="Number of children", pch=19)

abline(lm(sibs ~ childs, data=d), col="blue") ## add in a regression line ##
```

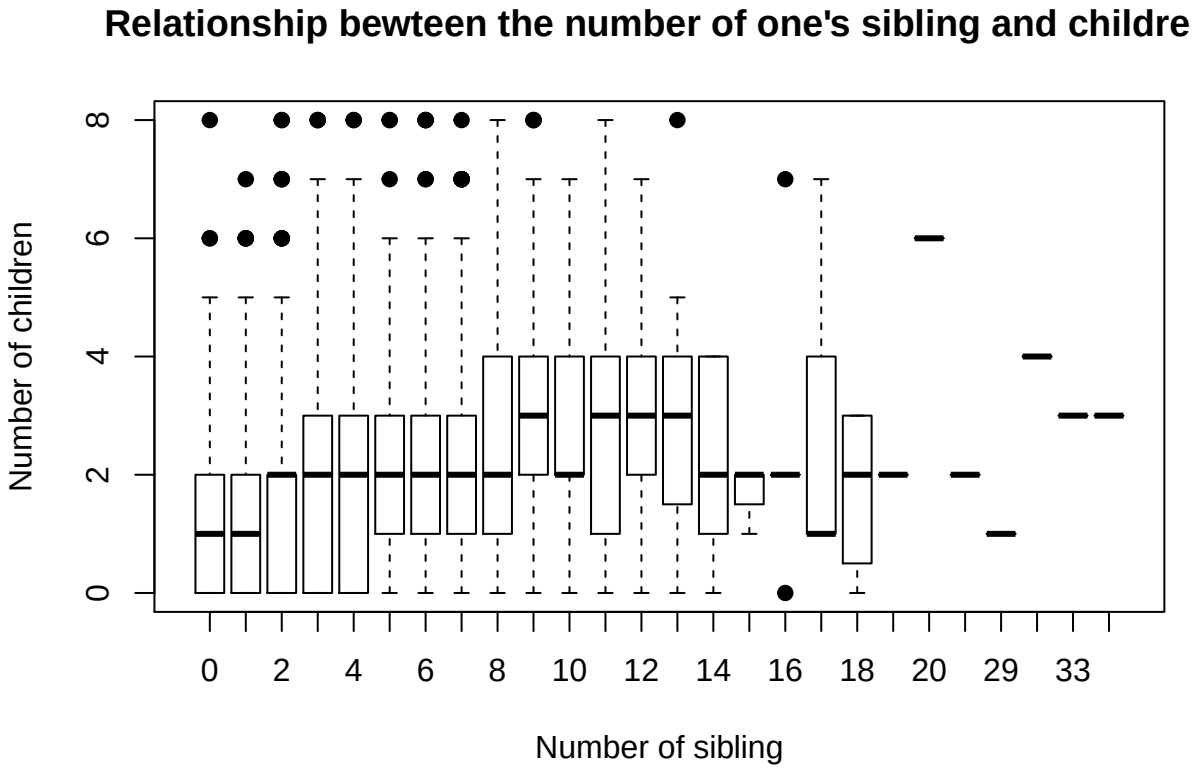


– or (for a boxplot) —

```
plot(d$sibs, d$childs, main="Relationship bewteen the number of one's sibling and children",
     xlab="Number of sibling", ylab="Number of children", pch=19)
```




```
plot(as.factor(d$sibs), d$childs, main="Relationship bewteen the number of one's sibling and children",
     xlab="Number of sibling", ylab="Number of children", pch=19) ## this creates a box plot ##
```



```r
mean(d[d$sibs == 1, 'childs'], na.rm=T)
```

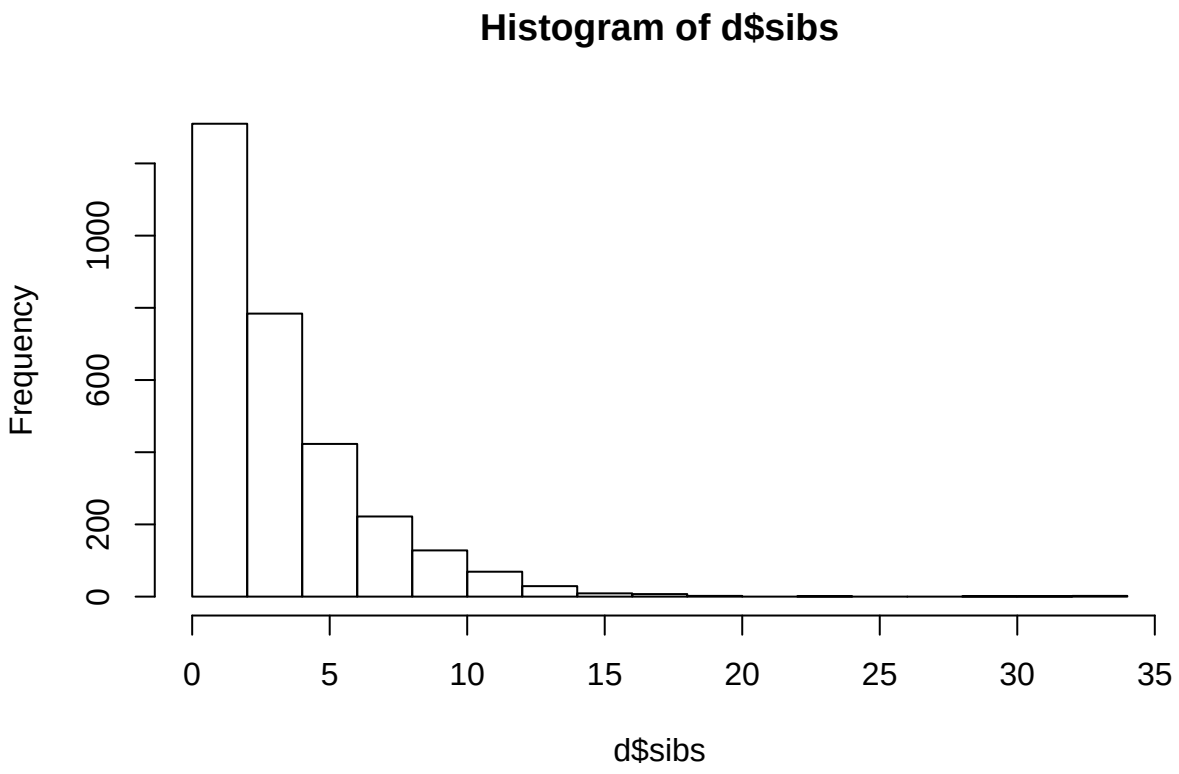
```
## [1] 1.462366
```

```r
describe(d[d$sibs == 1, 'childs']) ## respondent's work hours ##
```

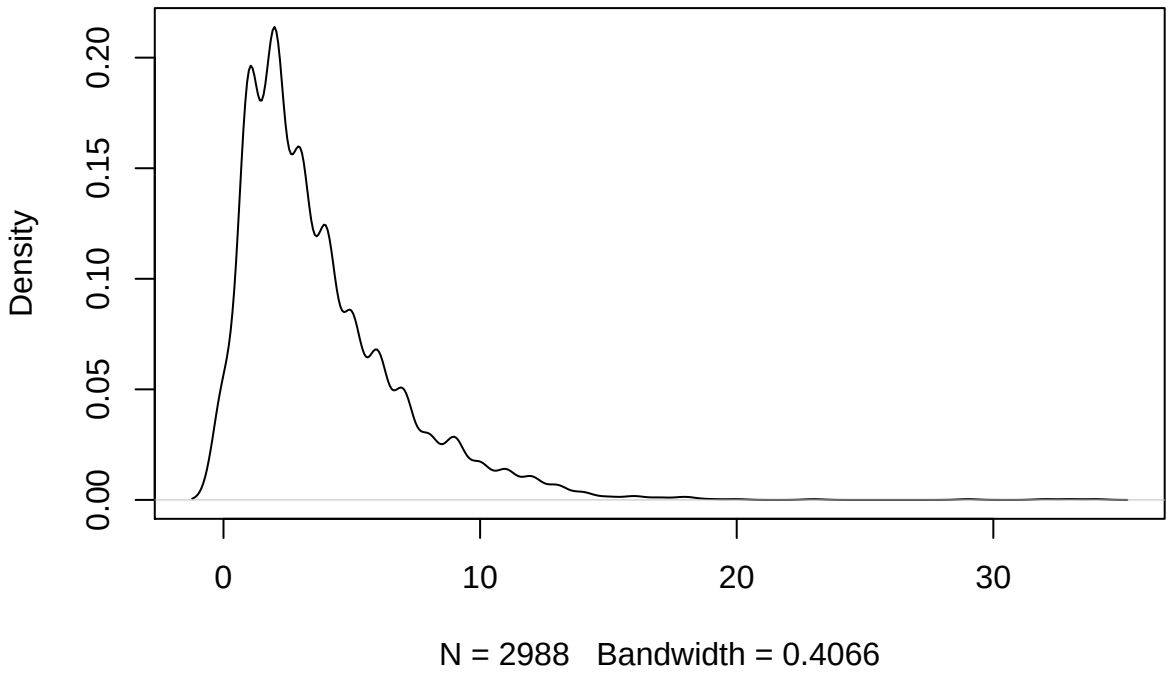
```
##    vars   n mean   sd median trimmed  mad min max range skew kurtosis   se
## X1    1 558 1.46 1.43      1    1.29 1.48   0   7     7 0.79     0.16 0.06
```

—— other useful graphing codes —-

```r
hist(d$sibs) ## draws a histogram ##
```




```r
dense <- density(d$sibs, na.rm=T) # returns the density data
plot(dense) # plots the results as a kernel density plot
```




```
library(ggplot2)
```

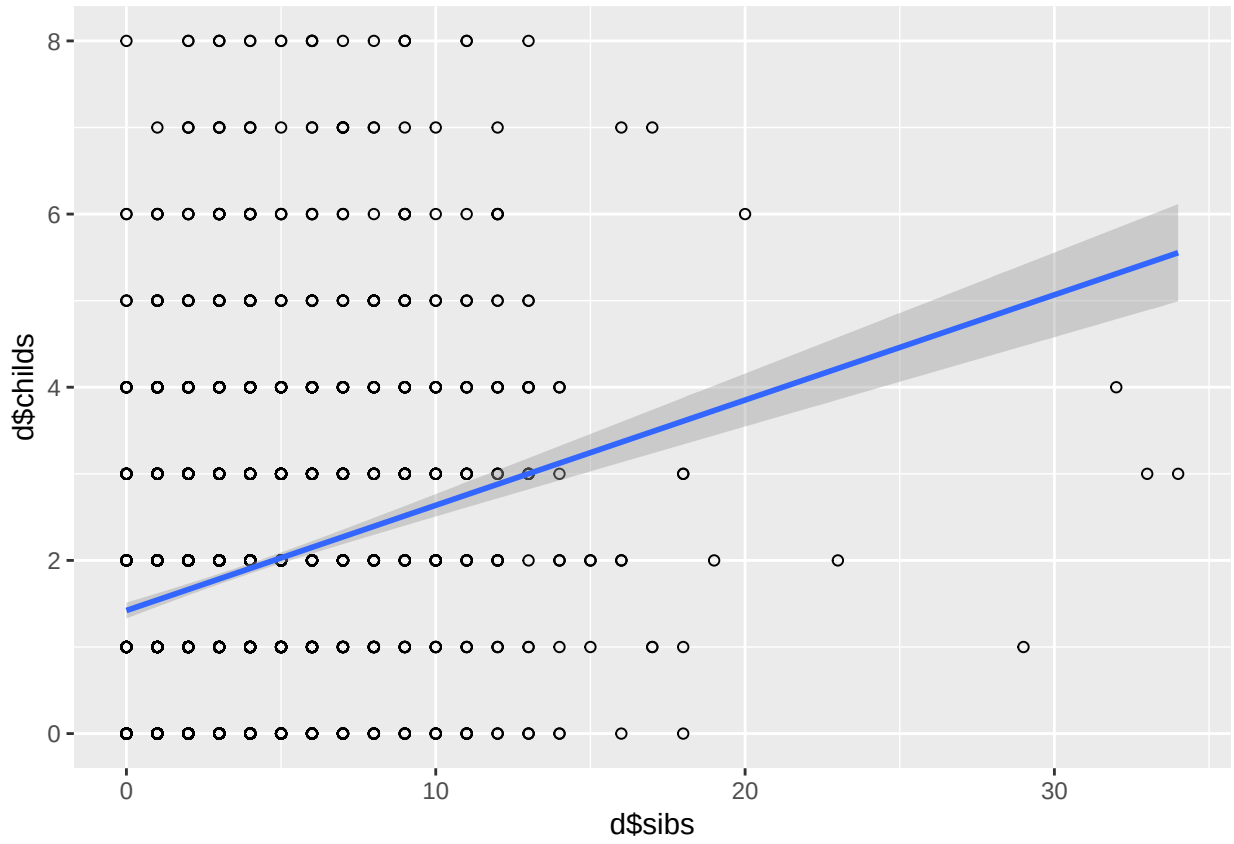
```
##
## Attaching package: 'ggplot2'
```

```
## The following objects are masked from 'package:psych':
##
##     %+%, alpha
```

```
ggplot(d, aes(x=d$sibs, y=d$childs)) + ## Another scatter plot
  geom_point(shape=1)      +    # Use hollow circles
  geom_smooth(method=lm)   # Add linear regression line
```

```
## Warning: Removed 1527 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 1527 rows containing missing values (geom_point).
```



6. Tell me two theories/ideas you might want to test in this course. Do you have a dataset for these ideas/theories already? Do you have it in R-readable format already? What is your main independent variable? What is your main dependent variable? Send me an email with the subject "Independent Project Ideas - [your name]" to gme2101@columbia.edu

See the email please, thank you.