An Exploratory Analysis of Fertility and Schooling

Paul Laskowski

August 15, 2016

Introduction

This analysis is motivated by the following research question:

> How is the amount of schooling received by a woman related to the number of children she has?

We will address this question using exploratory data analysis techniques. Our data comes from the General Social Survey (GSS), a large-scale sociodemographic study that has taken place since 1972. Nowadays, the GSS takes place every other year. The survey is conducted face-to-face and takes about 90 minutes to administer.

Our procedure is loosely based on an analysis by Sander.[1] Unlike Sander's study, we will not perform any statistical modeling, but we will dive deeper into an exploratory analysis of the data.

We will use a pooled cross section of data, including the years 2010, 2012, and 2014.

In line with Sander, we will restrict our analysis to women between the ages of 35 and 44. The reason for the lower bound is that most women age 35 and above have reached the point at which they will not have more children. This is convenient for our analysis, because we don't have to worry if participants with few years of education are simply so young that they haven't had all of the children they're going to have. Similarly, most women above age 34 have completed their education, which also simplifies our analysis.

The upper bound on age serves a different purpose. We want to compare women born in roughly the same time period. Over the last decades, we know that the average amount of education has grown, while the average number of children has decreased. As a result, women with a high amount of education tend to be born in more recent years, and for this reason alone, they will also tend to to have fewer children. By capping the age of women in our sample at 44, we reduce the possibility that age drives the relationships we observe.

Some caveats to keep in mind:

1. The GSS is a survey, and our sample is not representative of the U.S. population. Statisticians have specialized techniques for working with surveys. For example, the General Social Survey assigns a weight to each observation, which is meant to account for differences in how likely each individual is to appear in the sample. The current analysis is limited to descriptive statistics, so we will only be making claims about the sample, not the population.
2. Our analysis is not causal. The techniques we use here do not address the question of whether more education causes people to have more or less children. Note that Sander's paper claims to measure the causal effect of education, but there are good reasons to be suspicious of this claim. We will return to issues of causality later in the course.

Setup

First, we load the car library, which gives us a convenient scatterplotMatrix function.

```
library(car)
```

```
## Loading required package: carData
```

I used the GSS data explorer (https://gssdataexplorer.norc.org) to download a dataset with just the variables are are interested in. I performed some minimal editing to make sure the variable names were correct, and saved the data to the file Fertility.RData.

[1] Sander. The Effect of Women's Schooling on Fertility. *Economics Letters* 40, 229-233

When loading this file, we don't need to set the working directory, because we put the data file in the same directory as our Rmd file. A quirk of Knitr is that the working directory is reset to the place the Rmd file is located with every new chunk (we could override this with a command like opts_knit$set(root.dir = '~/Desktop/data_w203')). To keep things simple, we try to always keep our data together with our script files.

```
# Load the data
load("Fertility.Rdata")
```

Data Selection

We note that we have 6558 observations and 5 variables.

```
nrow(Fert_all)
```

```
## [1] 6558
```

```
str(Fert_all)
```

```
## 'data.frame':    6558 obs. of  5 variables:
##  $ year: num  2010 2010 2010 2010 2010 2010 2010 2010 2010 2010 ...
##  $ age : num  31 23 71 82 78 40 46 80 31 NA ...
##  $ kids: num  0 0 3 5 NA 2 1 1 3 2 ...
##  $ sex : Factor w/ 3 levels "","Female","Male": 3 2 2 2 2 3 2 2 2 2 ...
##  $ educ: num  16 16 8 10 0 6 16 15 14 14 ...
```

Looking at the age variable and the sex variable, we notice that we have both men and women and individuals as young as 18.

```
summary(Fert_all$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    18.0    34.0    47.0    48.1    61.0    88.0      73
```

```
summary(Fert_all$sex)
```

```
##        Female   Male
##      2   3639   2917
```

Our first step is to subset the data. To do this we will select women that are at least 35 years old using a logical vector.

```
subcases = 35 <= Fert_all$age & Fert_all$age < 45 & Fert_all$sex == "Female" &
  !is.na(Fert_all$age) & ! is.na(Fert_all$sex)
```

Note that we have to specifically check if age or sex are NA. If we don't do this, observations for which age or sex are missing will show up as NA's in our logical vector and we will end up keeping them. Since we want to ensure that all our individuals belong to the subgroup of women aged 35 and above, we use the is.na function to make sure such cases end up as False in our logical vector.

We use our logical vector to pull out just the rows we want and save them into a new data frame.

```
Fert = Fert_all[subcases, ]
nrow(Fert)
```

```
## [1] 683
```

We could have accomplished the same operation using the convenient subset command

```
Fert = subset(Fert_all, sex == "Female" & 35 <= age & age < 45)
nrow(Fert)
```

```
## [1] 683
```

Our new data frame has just 683 observations. We examine a summary.

```
summary(Fert)
```

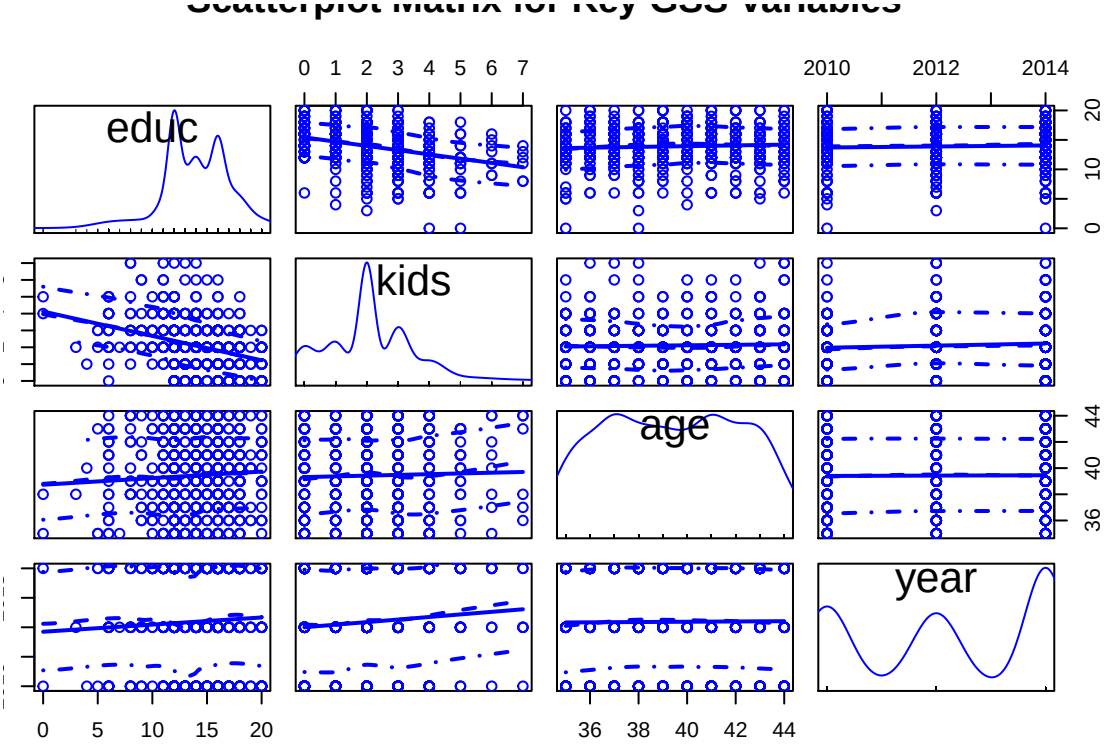
```
##       year           age             kids           sex
##  Min.   :2010   Min.   :35.00   Min.   :0.000         :  0
##  1st Qu.:2010   1st Qu.:37.00   1st Qu.:1.000   Female:683
##  Median :2012   Median :39.00   Median :2.000   Male  :  0
##  Mean   :2012   Mean   :39.41   Mean   :2.108
##  3rd Qu.:2014   3rd Qu.:42.00   3rd Qu.:3.000
##  Max.   :2014   Max.   :44.00   Max.   :7.000
##                                 NA's   :4
##       educ
##  Min.   : 0.00
##  1st Qu.:12.00
##  Median :14.00
##  Mean   :13.89
##  3rd Qu.:16.00
##  Max.   :20.00
##  NA's   :1
```

Looking at our subset of data, we note that there are some missing values. We don't know the number of kids for 4 individuals, and we don't have years of education for 1. This is a reasonably small fraction of our cases.

Exploratory Analysis

When we have a large number of variables, we often begin with a scatterplot matrix. This is helpful for getting a high-level overview of the relationships between our variables and can draw our attention to important features we want to investigate further.

```
scatterplotMatrix( ~ educ + kids + age + year, data = Fert,
                   main = "Scatterplot Matrix for Key GSS Variables")
```



We notice a couple of features in this matrix that can help guide our analysis.

1. There is a noticable negative relationship between educ and kids. This is the relationship we are most interested in.
2. We notice that age and year have much weaker relationships with the other variables. This is good, because these variables could confound our analysis. We can pick out some interesting features. The educ vs. age plot reveals a noticable downward tick for low levels of education.
3. Year does not seem to have a strong relationship with any other variable, which is good since we're pooling three years together. There seems to be an effect where women with the most kids are sampled in the later years.

This plot suggests that educ and kids are closely related and it also highlights that we should worry that age (and to a lesser extent year) could affect the bivariate relationships we see.

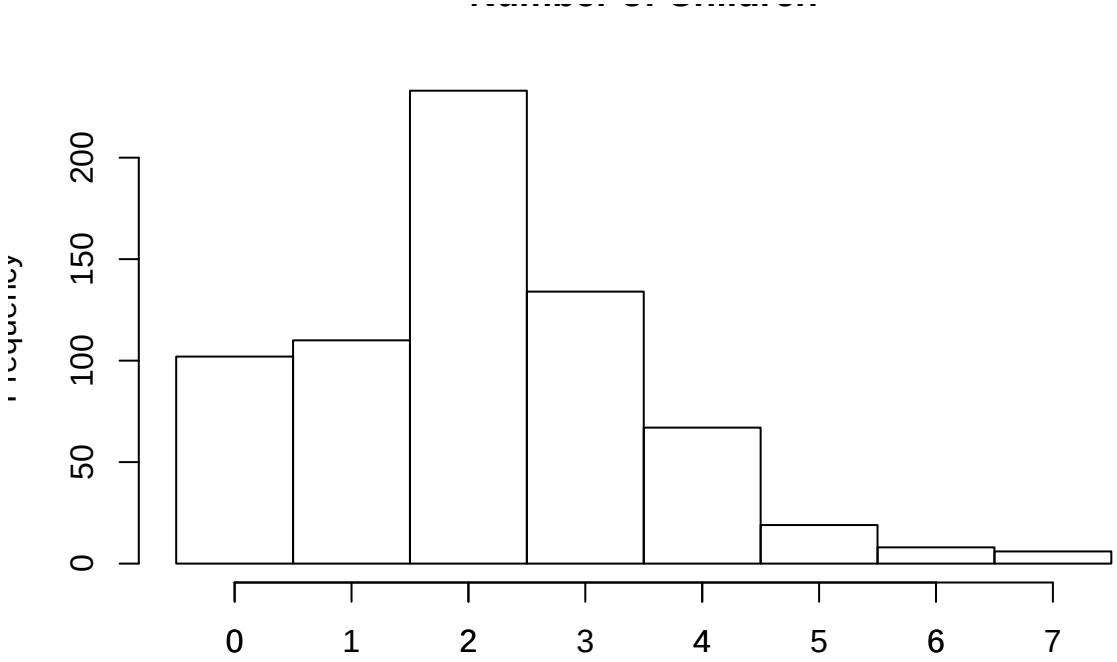
Though we will not do any modeling, we will think of kids as our outcome variable. We first summarize it.

```
summary(Fert$kids)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000   1.000   2.000   2.108   3.000   7.000       4
```

We notice that kids seems to take on integer, nonnegative values, as we would expect. Next, we create a histogram of kids, making sure that we set the cut points correctly using the breaks argument. We also supply tick marks on the x-axis manually.

```
hist(Fert$kids, breaks = 0:8 - 0.5, main = "Number of Children",
     xlab = NULL)
axis(1, at = 0:7)
```



We notice a few important features here:

1. kids takes on integer values. Moreover, the maximum is only 7, so integer effects may play an important role. That is, a model that assumes kids to be continuous may be distoring.
2. kids is also nonnegative, and this boundary seems important in the sense that there is a sharp dropoff at zero. It is as though the left end of a larger distribution is being chopped off. kids is what we would call a count variable.
3. Visually, the histogram seems to have a positive skew (right skew). This means that there are observations stretching further to the right of the bulk of the data. We note that the mean is greater than the median, which is typically what we see for positively skewed variables.

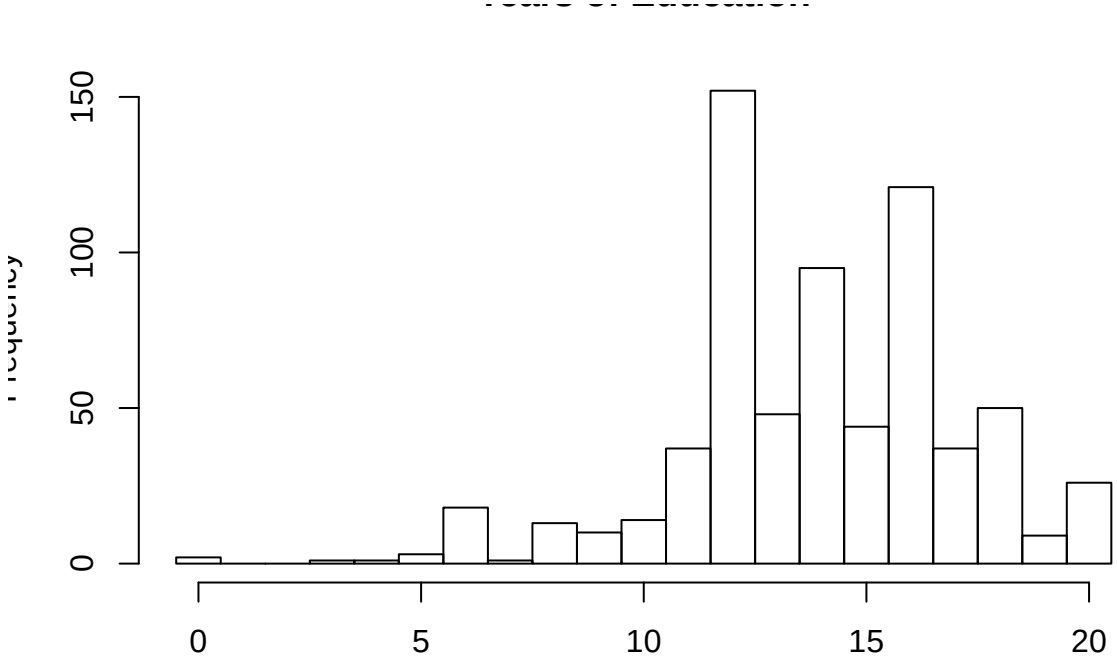
Because kids is a count variable, we may need to consider special techniques when we model it. For example, there is an advanced technique called Poisson regression that may be applicable here.

We next examine years of education, which we consider to be our main input variable. As before, we must manually set the cut points for our histogram.

```
summary(Fert$educ)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.00   12.00   14.00   13.89   16.00   20.00       1
```

```
hist(Fert$educ, breaks = -1:20 + 0.5, main = "Years of Education",
     xlab = NULL)
```



We notice several features from the summary and histogram:

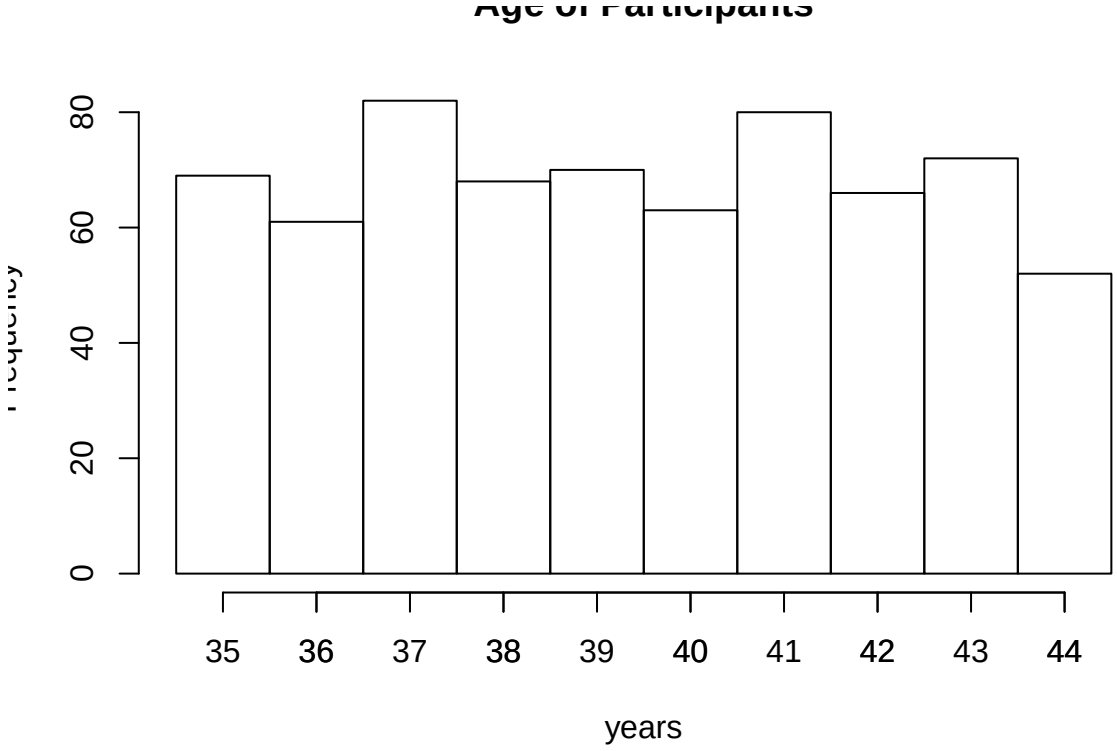
1. Like kids, education seems to take on integer values.
2. Unlike kids, years of education seems more dispersed, so integer effects are not as important. Additionally, the variable is non-negative, but the lower boundary does not seem important. The vast majority of the data is greater than zero.
3. There is an unusual spike at 12 years of education. This corresponds to individuals that finish high school but do not go on to college. Similarly, a smaller spike can be discerned at 16 years, which corresponds to individuals that finish a 4-year college. This is important to note because there may be something special about graduating high school, or graduating college, as opposed to getting an extra year of education that does not result in graduation. Later in the course, we will learn about index variables, which are a technique we can apply to model such graduation effects. For the current analysis, we will remember to check for graduation effects as we continue exploring.

We next examine our age variable.

```
summary(Fert$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   35.00   37.00   39.00   39.41   42.00   44.00
```

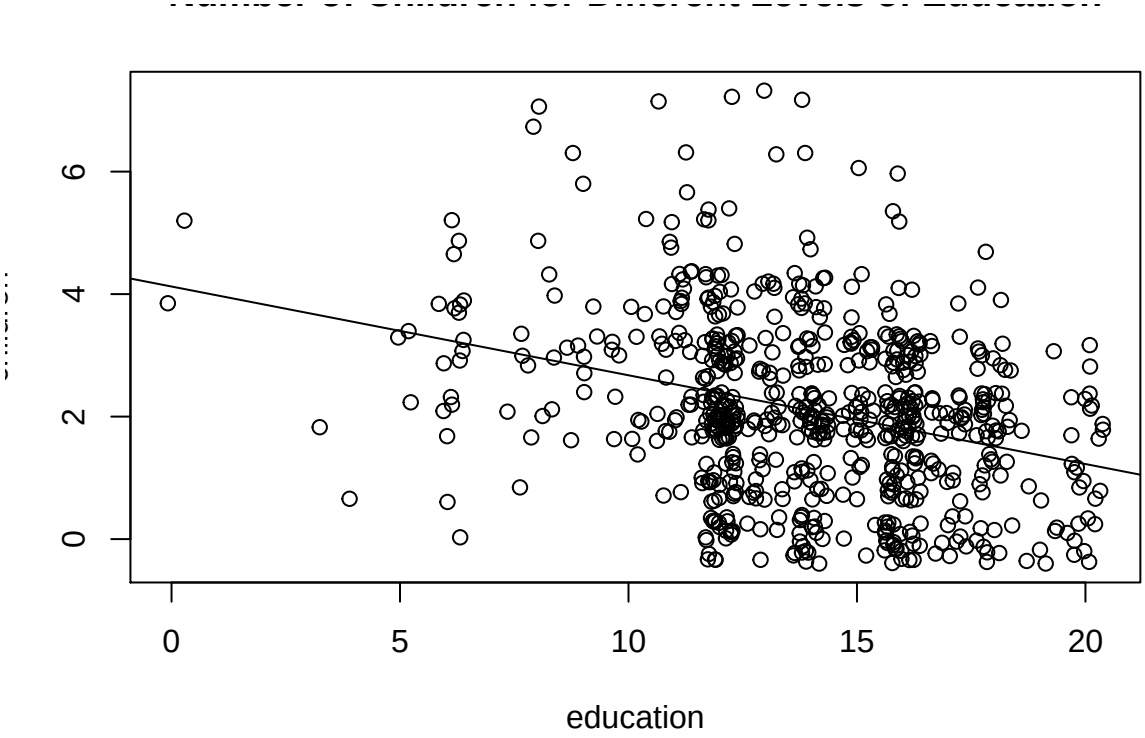
```
hist(Fert$age, breaks = 34:44 + 0.5, main = "Age of Participants",
     xlab = "years")
axis(1, at = 35:44)
```



Note that age ranges from 35 to 44 and the distribution seems uniform in this range.

We want to understand what bivariate relationship exists between our main variables of interest, kids and educ. We begin with a scatterplot. Because each variable takes on a rather limited set of values, we add jitter along each axis to make sure points don't overlap perfectly. We also add the ordinary least squares regression line, which is a way of visualizing what linear relationship exists in the data. We will have much more to say about this line later in the course.

```
plot(jitter(Fert$educ, factor=2), jitter(Fert$kids, factor=2),
     xlab = "education", ylab = "children",
     main = "Number of Children for Different Levels of Education")
abline(lm(Fert$kids ~ Fert$educ))
```



Without going into too much detail, this plot tells us that there is an overall negative linear relationship between our variables. More education is associated with less children, on the average. We can confirm this by checking the sample correlation

```
cor(Fert$kids, Fert$educ, use = "complete.obs")
```

```
## [1] -0.3234381
```

Note that the cor function does not accept an na.rm argument. instead, we set the use argument to "complete.obs", which tells the function to only look at rows for which all variables have a value.

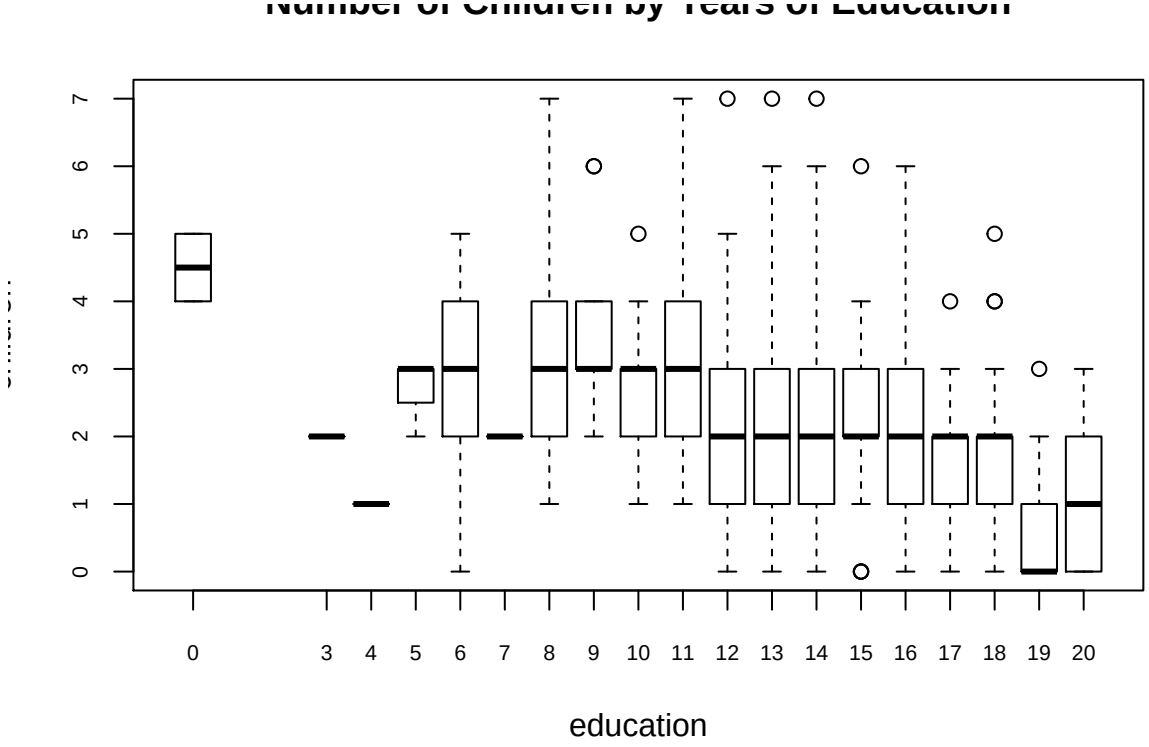
The correlation of -0.3234381 is indeed negative, though it is moderate in magnitude.

Regression is a useful tool, but this plot only tells us what the best fitting line looks like, it does not reveal more complicated relationships, or whether the relationship between our variables is actually linear.

To gain a more detailed view, we try using a boxplot. We use the at argument to specify where the boxes go along the x-axis. If we didn't do this, the box corresponding to 0 years of education would be right next to the box for 3 years, since there are no data points in between. By putting the unique values of education into the at argument, we ensure that the boxes are spaced linearly.

We also use the cex.axis argument to scale the numbers on the axes, to prevent them from overlapping and being cut off.

```
boxplot(kids ~ educ, data = Fert,
        at = sort(unique(Fert$educ)), cex.axis = .7,
        main = "Number of Children by Years of Education",
        xlab = "education", ylab = "children")
```



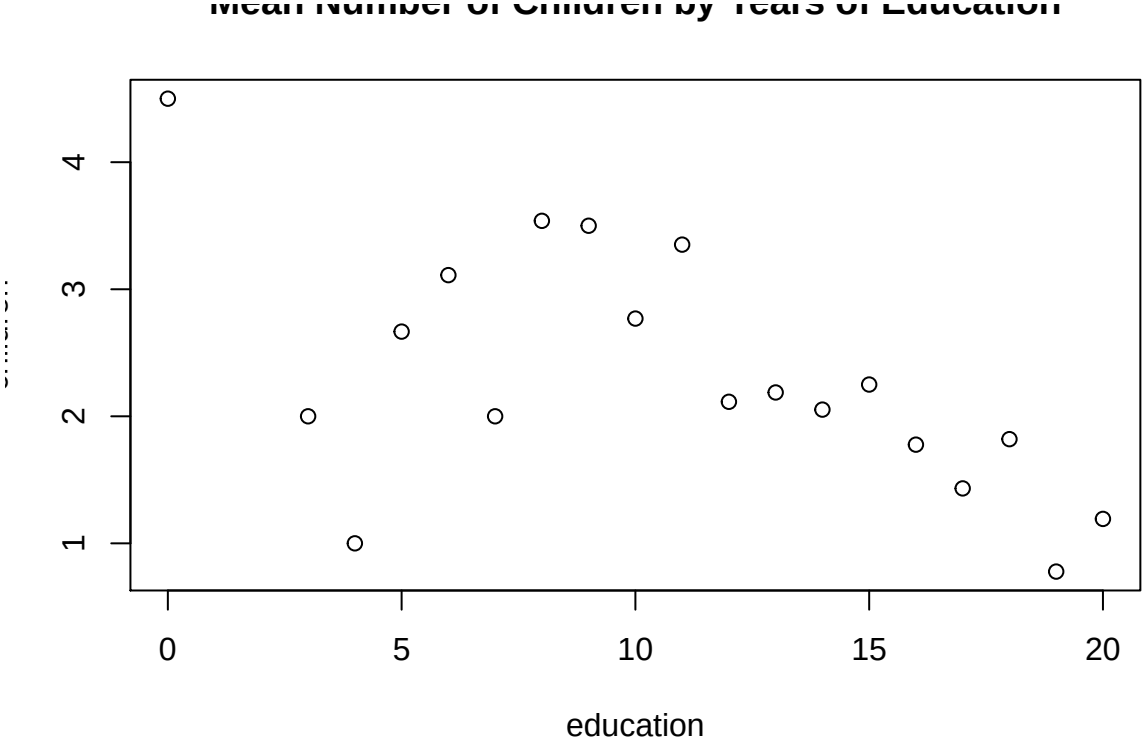
Overall, the relationship does not appear especially linear. We shouldn't pay too much attention to the left 3 or 4 bars of the graph, since there are so few data points there. Up to about 11 years, the trend seems to be flat, or perhaps even increasing. At 12 years, the number of kids drops and the trend to the right of this point seems decreasing.

To better assess this relationship, we will plot the mean number of kids for each level of education. We can use a by statement to get the means.

```
kid_means = by(Fert$kids, Fert$educ,  mean, na.rm = T)
```

To plot the means correctly, we must enter the x-values manually into the plot command. We can use the unique function to extract the x-values and the sort function to sort them so that they work with the plot command.

```
plot(sort(unique(Fert$educ)), kid_means,
     xlab = "education", ylab = "children",
     main = "Mean Number of Children by Years of Education")
```



There is a lot of noise on the left side of the graph, and 0 years of education appears as a notable outlier. We might wonder if respondents that entered 0 misunderstood the question, or if their answers were recorded incorrectly. Setting this point aside, the overall trend could be approximated by a parabola. Later in the course, we will discuss the use of quadratic terms to model parabolic relationships.

We mentioned the possibility of a graduation effect at 12 years and 16 years of education, and we can see some evidence for this in the box plot and the plot of means. The mean at 12 years of education seems noticably lower than the trend left of that point. A similar downward jog can be seen at 16 years of education.

We might speculate that participants that graduate from high school or college might be especially career-focused, resulting in the lower number of children. Given the variables we have in this dataset, we are not able to futher investigate this possibility.

As we mentioned before, we can use a technique called indicator variables to model these types of discontinuities. We will learn about indicator variables later in the course and discuss the tradeoffs inherent in choosing a more complicated model.

To focus our attention on graduation events, we may want to bin our educ variable into the intervals between graduations. This is easily done with the cut command.

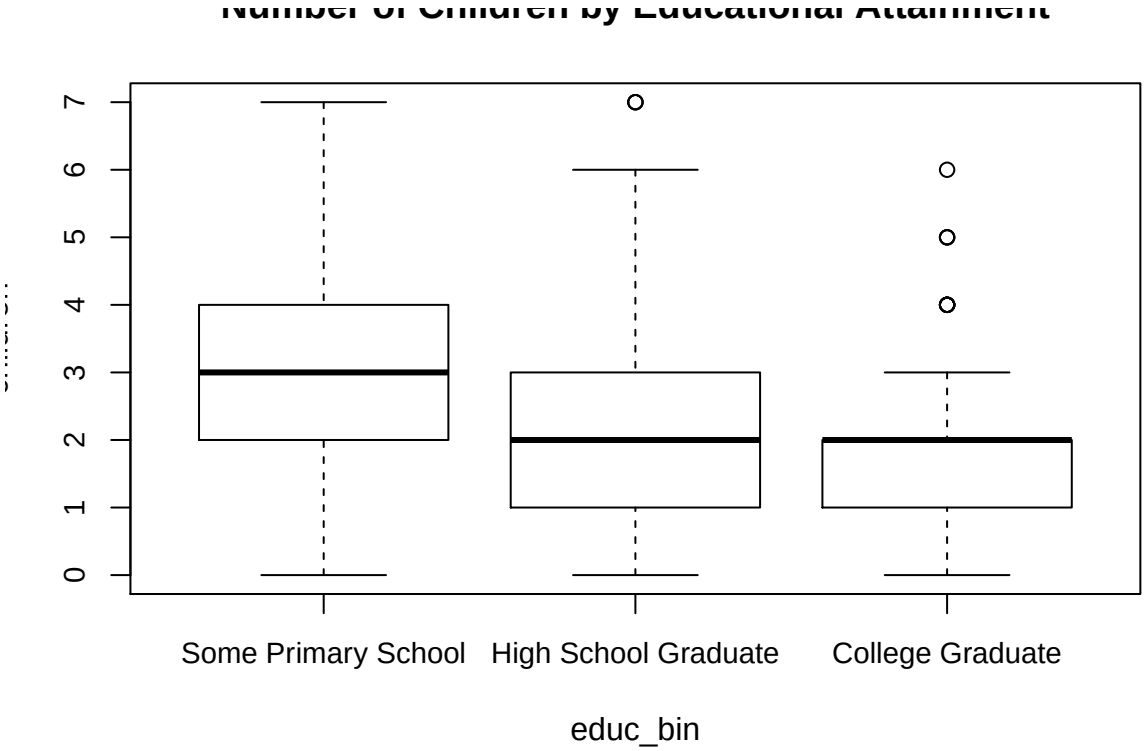
```
educ_bin = cut(Fert$educ, breaks = c(0,11,15, Inf), labels =
      c("Some Primary School", "High School Graduate", "College Graduate"))
summary(educ_bin)
```

```
##  Some Primary School High School Graduate     College Graduate
##                   98                  339                  243
##                 NA's
##                    3
```

The resulting boxplot then clearly shows the number of kids for each group, telling a different kind of story.

```
boxplot(kids ~ educ_bin, data = Fert, cex.axis = .9,
        main = "Number of Children by Educational Attainment",
```

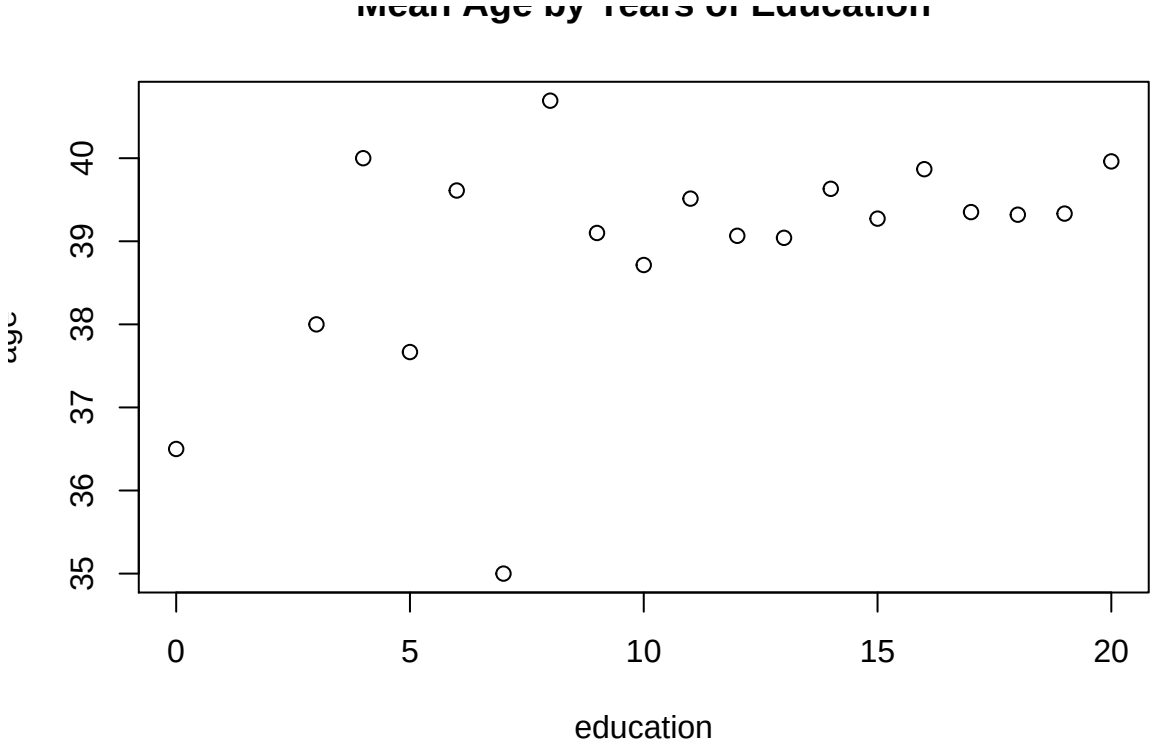
```
        ylab = "children")
```



Finally, we want to examine how our age variable relates to education and to kids. This will help us assess whether the effects we observe could be driven by age, and whether our strategy of limiting our analysis to women between the ages of 35 and 44 was successful.

First we examine the relationship between age and educ. We could use a boxplot or a plot of means. Here, we select a plot of means.

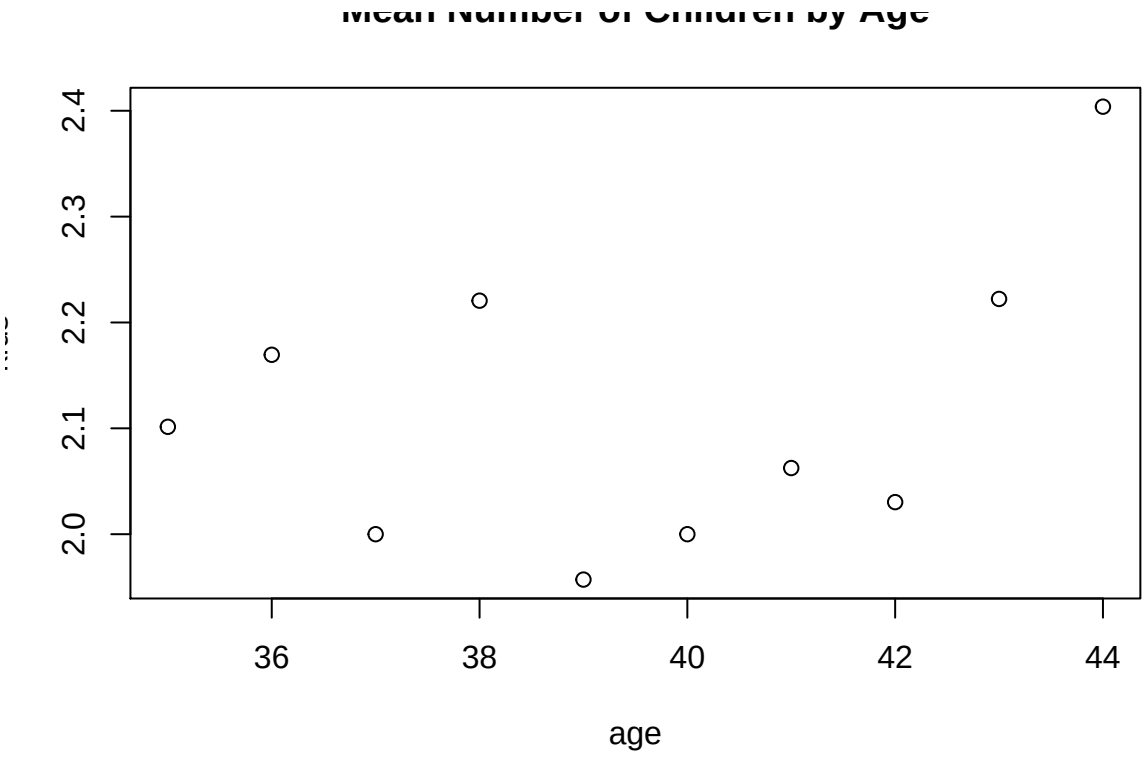
```
age_means = by(Fert$age, Fert$educ,  mean, na.rm = T)
plot(sort(unique(Fert$educ)), age_means,
     xlab = "education", ylab = "age",
     main = "Mean Age by Years of Education")
```



We observe that women with less than 8 years of education appear to be several years younger than women with at least 8 years. Above 8 years of education, women in each column seem to have roughly the same distribution of ages, as we hoped to achieve.

We next examine the relationship between age and kids. Again, we select a plot of means.

```
kids_by_age = by(Fert$kids, Fert$age,  mean, na.rm = T)
plot(sort(unique(Fert$age)), kids_by_age,
     xlab = "age", ylab = "kids",
     main = "Mean Number of Children by Age")
```



In this graph, we don't see a very strong relationship between age and number of kids. There is a noticable outlier at 44 years of age. In particular, we are curious to see if the younger women in the sample have fewer kids, which might explain why the number of kids we saw for low levels of education was so low. However, women on the left side of this graph do not seem to stand out. Note that the range on the y-axis is less than 0.5 children, so this plot does not really explain the larger differences we saw earlier.

Even though we don't see a clear story about how age could influence the results we see, the low age of women with 1-7 years of education does suggest that this group of women might be especially unique in some way. This is especially plausible given the small number of women in this group and the fact that the US school system typically requires students to remain in school until the age of 16, when most would have at least 10 years of education.

Above 7 years of education, we don't see any strong relationship between age and our other variables, which strengthens our belief that the women in this group can be meaningfully compared against each other.

Discussion

Our exploratory analysis has uncovered a number of features that would prove useful as we turn to statistical modeling. Among the important features we found, there were skewed distributions, attenuated / count variables, (possibly quadratic) non-linearities, outliers, and graduation effects. Later, we will learn more about statistical modeling, and we will use some of these same exploratory techniques to test the assumptions of our models.