STAT 22000 Summer 2020 Homework 2

Yibi Huang

June. 25, 2020

Beginners often have trouble loading data sets to R from a file. Please refer to Section 3 of Lab #1: http://www.stat.uchicago.edu/ yibi/s220/labs/lab01.html about how to change the working directory and load data set from a file.

You can refer to Lab #2: http://www.stat.uchicago.edu/ yibi/s220/labs/lab02.html to find the details for R commands you need to make the histograms, boxplots, and scatterplots required below.

The data file `NLSY.txt` posted on Canvas along with HW2 contains data come from the NationalLongitudinal Study of Youth (NLSY), U.S. Bureau of Labor Statistics https://www.bls.gov/nls/home.htm. The subjects are 1306 American men and 1278 American women between the ages of 14 and 22 in 1979. The variables include

- `Gender`
- `AQFT`: the percentile scores on the Armed Forces Qualifying Test, which is designed for evaluating the suitability of military recruits but which is also used by researchers as a general intelligence test
- `Edu2006`: years of education achieved by time of interview in 2006
- `Income2005`: annual income in thousands of dollars in 2005

The orignal data include more subjects but those with missing values of variables are omitted.

First, let's load the data set to R and load the mosaic library.

```
NLSY = read.table("NLSY.txt",header=T)
library(mosaic)
```

(a)

Make a histogram for the variable `Edu2006` using the R codes below. Explain the location of the modes of the histogram.

```
histogram(~Edu2006, data=NLSY, width=?, xlab="??")
```

Be sure to adjust the bin width by replacing the question mark in `width=?` with a number (What's a natural binwidth for `Educ2006`?) Also please label the axis properly by changing the ?? in `xlab="??"` with some description of the variable.

(b)

The R commands below make two histograms for `Edu2006` on the same horizontal axis report the favorite summaries (mean, SD, five-number summary) for `Edu2006` for males and females. (Please change the binwidth and label the axis yourself) From the histograms and the summary, do you think that males and females in the data set have different distributions in their education levels?

```
histogram(~Edu2006 | Gender, data=NLSY, width=?, xlab="??",layout=c(1,2))
favstats(Edu2006 ~ Gender, data=NLSY)
```

(c)

Make a side-by-side boxplot comparing the education levels of males and females using the R command below. What information is available in the histogram but missing in the boxplot?

```
bwplot(Edu2006 ~ Gender, data=NLSY)
```

(d)

Make a side-by-side boxplot and two histograms sharing the same horizontal axis comparing distribution of `Income2005`. Please always label the axes and adjust the binwidth of the histograms. Comment on the shape of the hisograms. Which gender had a higher income? Which gender had a higher variability in their distributions of incomes?

```
histogram(~Income2005 | Gender, data=NLSY, width=10, layout=c(1,2))
bwplot(Income2005 ~ Gender, data=NLSY)
favstats(Income2005 ~ Gender, data=NLSY)
```

(e)

Make the same side-by-side boxplot and histograms as in the previous part but for the **logarithm** of `Income2005`. Please always label the axes and adjust the binwidth of the histograms. Comment on the skewness of the distributions after the log transfomation.

(f)

When talking about gender pay gap, one should compare the income of men and women with the same qualification. However, we don't have a good measure of qualification in the data. As an indirect measurement of qualification is the education level, One can compare the income of men and women with the same education level. The R codes below split the data by the levels of `Edu2006`, and make a side-by-side boxplot comparing the income of men and women with the same years of education.

```
bwplot(Income2005 ~ Gender | Edu2006, data=NLSY)
```

Here is the same plot but with better labels for education levels.

```
NLSY$Edu2006.fac = factor(NLSY$Edu2006, labels = paste(6:20,"yr edu"))
bwplot(Income2005 ~ Gender | Edu2006.fac, data=NLSY)
```

Taking logarithm might make it easier to tell which gender had a higher income.

```
bwplot(log(Income2005) ~ Gender | Edu2006.fac, data=NLSY)
```

Did men earn more than women, even after adjusted for their education level?

(g)

Another indirct measre of qualification is the intelligence test score percentiles (`AFQT`). Make a scatter plot between `AQFT` and `Income2006` and another scatter plot with `Income2006` log-transformed using the R commands below. Please label the axes properly.

```
qplot(AFQT, Income2005, data=NLSY, xlab="??", ylab="??")
qplot(AFQT, log(Income2005), data=NLSY, xlab="??", ylab="??")
```

Please make the plots above answer the following questions based on the plots.

1. How did `Income2005` change with `AQFT`? If yes, how?
2. Did the variability of `Income2005` change with `AQFT`? If yes, how?
3. Did the variability of the **logarithm** of `Income2005` change with `AQFT`? If yes, how?

(h)

The R code below produce a color-coded scatterplot between `AFQT` and `Income2005` and the color of points represents the Gender of the subject.

```
qplot(AFQT, log(Income2005), data=NLSY, xlab="??", ylab="??",color=Gender)
```

The R code below produce separate scatterplots for men and women.

```
qplot(AFQT, log(Income2005), data=NLSY, xlab="??", ylab="??",facets = ~Gender)
```

Please make the plots above answer the following questions based on the plots.

1. For each gender, how did the logarithm of `Income2005` change with `AQFT`? How did the variability of the logarithm of `Income2005` change with `AQFT`?
2. Comparing men and women with similar intellegence test score percentiles, did men earn more than women in general?

(i)

Make color-coded scatterplots between `AFQT` and `Income2005` for each level of years of education, using the color of points to represent the Gender of the subjects using the R codes below. Comparing men and women with the same years of education and with similar intellegence test score percentiles, did men earn more than women in general?

```
qplot(AFQT, log(Income2005), data=NLSY, xlab="??", ylab="??",
      color=Gender, facets=~Edu2006.fac)
```