STATS 500 Homework 1 Yuan Yin

Problem 1

Summary of the data uswages

The data contains information about the wages of people with different race, living places, metropolitan or not, type of their job (part-time or full-time) and years of education and work experience.

Here as follows shows some interesting results found by R:

First, we need to abandon the missing values, we sorted the years of working experience and found that some of them are missing, it will affect the results a lot if we don't abandon them before we summarize the data. Also, we found that to the upper limit of the wages, some of them has been over than $3000. Comparing with what most people earn ([0,1000]), these data is in small numbers and large values. In this way, we also should abandon them previously.

Now we get the data that is valuable and deserve observation. In order to find how the different factors influence the wages. First we plot the frequency histogram and the box plot of wage:

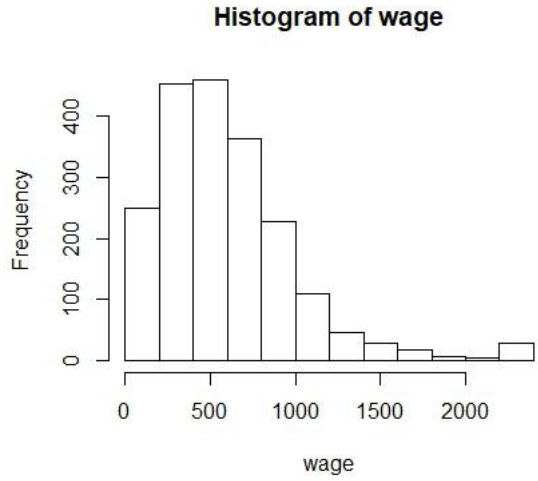

fig 1

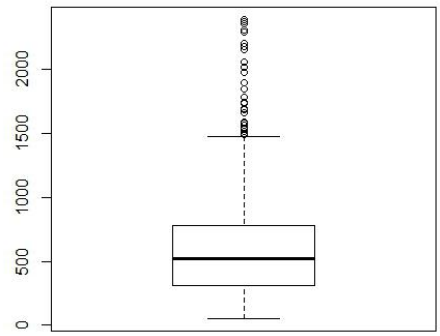

fig 2 box plot of wage

We can easily found that most people's wages are $500 more or less (±$300), when wage gets higher, there are less and less people. However, the interesting thing is, when the wage is higher than $2200, people gets a little more than the previous stage. If it isn't the reason of mistake or sample data not enough, then maybe it is because for high-income groups, $2200 is a popular usage that companies usually set more than $2200 as the lowest limit for high-income groups. And we can confirm the conclusion from figure 2.

For the second step, we plot the 4 types of relationship between wage and years of education, years of experience, people's race and whether in metropolitan statics area respectively, here as follows is what we know from R:

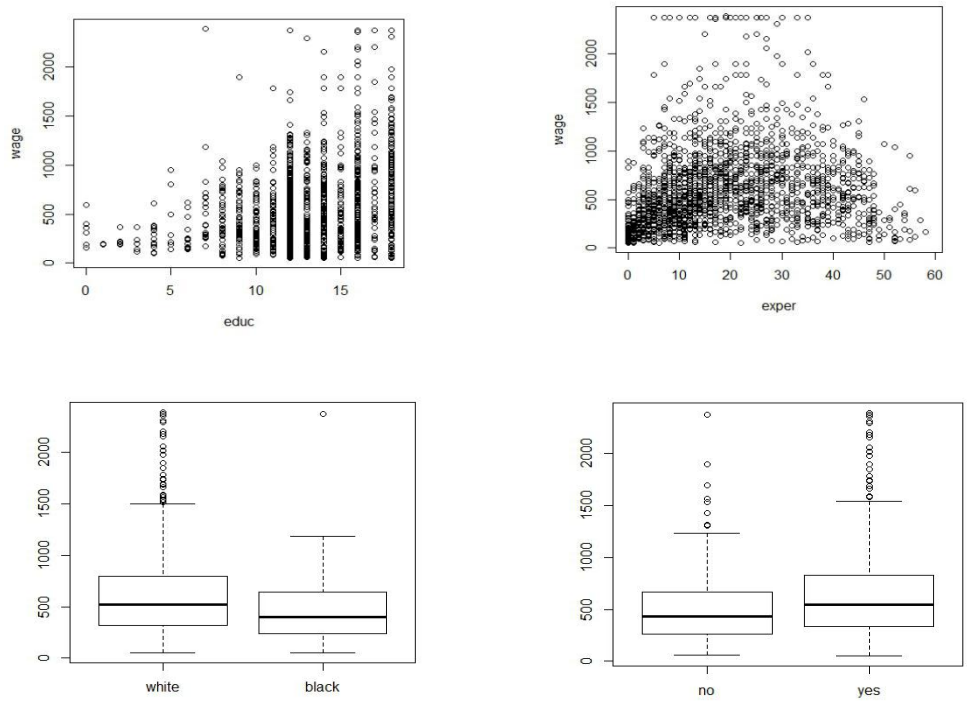

fig 3 relationship with wage

According to figure 3, as the years of education rise, the upper limit of wage rises, which means more people can earn much higher than people with less years of education. When the years are between 5 to 10, the upper limit is rising linearly, but when the years exceed 10, the distributions of wage almost stay the same. Most people earn between 0 to $1500.

What's more, with the rising of years of experience, wage doesn't rise as we thought. The highest income distributed evenly between 0 to 40 years. However, if a person has work experience more than 40 years, the highest wage they can earn declines with year rises, this maybe because those people are aged but companies prefer to hire younger workers who can learn fast and work more efficiently.

Also we found that white people generally earn more than black people. Especially evident in high-income groups. The same result for people in metropolitan area who can generally earn more than those who are not.

Finally, we compare the wages by different living regions, we calculate the mean of wage in each region, and here is the result:

region	wage
ne (northeast)	621.7852
mw (mid west)	585.3677
so (south)	569.4012
we (west)	635.0164

tab 1

The wage in the right column is the mean of all people living in the specific region on the left, we can found that people living in west and northeast commonly have higher wages than people living in south and mid west.

Above all is what I found and summarized from the data "uswages"

Appendix

```
## read in the data
data(uswages)
## dimension of the data
dim(uswages)
## numerical summaries
summary(uswages)
## missing values
uswages$exper[uswages$exper < 0] = NA
uswages$wage[uswages$wage > 3000] = NA
## new summary
summary(uswages)
## categorical variable
uswages$race = factor(uswages$race)
uswages$smsa = factor(uswages$smsa)
uswages$ne = factor(uswages$ne)
uswages$mw = factor(uswages$mw)
uswages$so = factor(uswages$so)
uswages$we = factor(uswages$we)
uswages$pt = factor(uswages$pt)
levels(uswages$race) = c("white", "black")
levels(uswages$smsa) = c("no", "yes")
levels(uswages$ne) = c("no", "yes")
levels(uswages$mw) = c("no", "yes")
levels(uswages$so) = c("no", "yes")
levels(uswages$we) = c("no", "yes")
levels(uswages$pt) = c("full", "part")
summary(uswages)
## graphical summaries
attach(uswages)
hist(wage)
##interesting of range[2200, 2400]
boxplot(wage)
reigon_to_data <- aggregate(wage ~ ne + mw + so + we, data = uswages, mean)
plot(educ,wage)
plot(exper,wage)
plot(race,wage)
plot(smsa,wage)
reigon_to_data
```