Project 1

Zhenjian Wang

PID: A15880200

1. choose two variables that you are interested to know effects of one on another (review the given data files - pick a data file with your two variables of interest).

Background Information:

I am interested in exploring the relationship between hourly wage in dollars and labor market experience in years. Take a worker in a factory for an example. If the worker works in the factory for years and knows the producing process and how to handle machines well, he is usually get paid more than a new worker. Thus, in my opinion, if a labor gets more skilled in a field, his value should go up. So my hypothesis is that if a person has more labor market experience, his hourly wage should be higher. I will use data to check if the hypothesis is true.

Definition of Variables:

The variable ‘wage’ means hourly wage in dollars, and it is my dependent variable.

The variable ‘exper’ means labor market experience in years, and it is the independent variable.

2. Work on descriptive statistics, data screening and data cleaning, as much as practiced in the course.

Import data

```
data <- read.csv("C:/Users/71778/Downloads/wage.csv")

head(data)
```

```
##   wage exper female tenure3 educ
## 1  3.1     2      1       0   11
## 2  3.2    22      1       0   12
## 3  3.0     2      0       0   11
## 4  6.0    44      0       1    8
## 5  5.3     7      0       0   12
## 6  8.8     9      0       1   16
```

Our variables are wage and exper which are explained as above.

We could see that wage is a column with numeric values and exper is a column with integers.

We could also see the minimum, mean, max of the ‘wage’ and ‘exper’ columns. The minimum of wage is 0.530 and the maximum is 25.000, with a mean of 5.909. The minimum of labor market experience is 1 year, and the maximum is 51 years, with a mean of 17.02 years.

```
summary(data)
```

```
##       wage            exper           female          tenure3
##  Min.   : 0.530   Min.   : 1.00   Min.   :0.0000   Min.   :0.0000
##  1st Qu.: 3.300   1st Qu.: 5.00   1st Qu.:0.0000   1st Qu.:0.0000
##  Median : 4.700   Median :13.50   Median :0.0000   Median :0.0000
##  Mean   : 5.909   Mean   :17.02   Mean   :0.4791   Mean   :0.4734
##  3rd Qu.: 6.900   3rd Qu.:26.00   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :25.000   Max.   :51.00   Max.   :1.0000   Max.   :1.0000
##       educ
##  Min.   : 0.00
##  1st Qu.:12.00
##  Median :12.00
##  Mean   :12.56
##  3rd Qu.:14.00
##  Max.   :18.00
```

```
str(data)
```

```
## 'data.frame':    526 obs. of  5 variables:
##  $ wage   : num  3.1 3.2 3 6 5.3 8.8 11 5 3.6 18 ...
##  $ exper  : int  2 22 2 44 7 9 15 5 26 22 ...
##  $ female : int  1 1 0 0 0 0 0 1 1 0 ...
##  $ tenure3: int  0 0 0 1 0 1 1 1 1 1 ...
##  $ educ   : int  11 12 11 8 12 16 18 12 12 17 ...
```

Data screening and cleaning:

First, we set all the negative values in ‘wage’ and ‘exper’ columns to NA because the wage and working experience in years should not be a negative value. Second, we get rid of the data that is NA.

```
data$wage[data$wage < 0]<-NA
data$exper[data$exper < 0] <- NA
data <- na.omit(data)
wage = data$wage
exper = data$exper
```

We plot the relationship between labor market experience in years and the hourly wage, but we could not see any clear pattern.

```
plot(exper, wage, xlab = 'labor market experience in years', ylab = 'hourly wage')
```

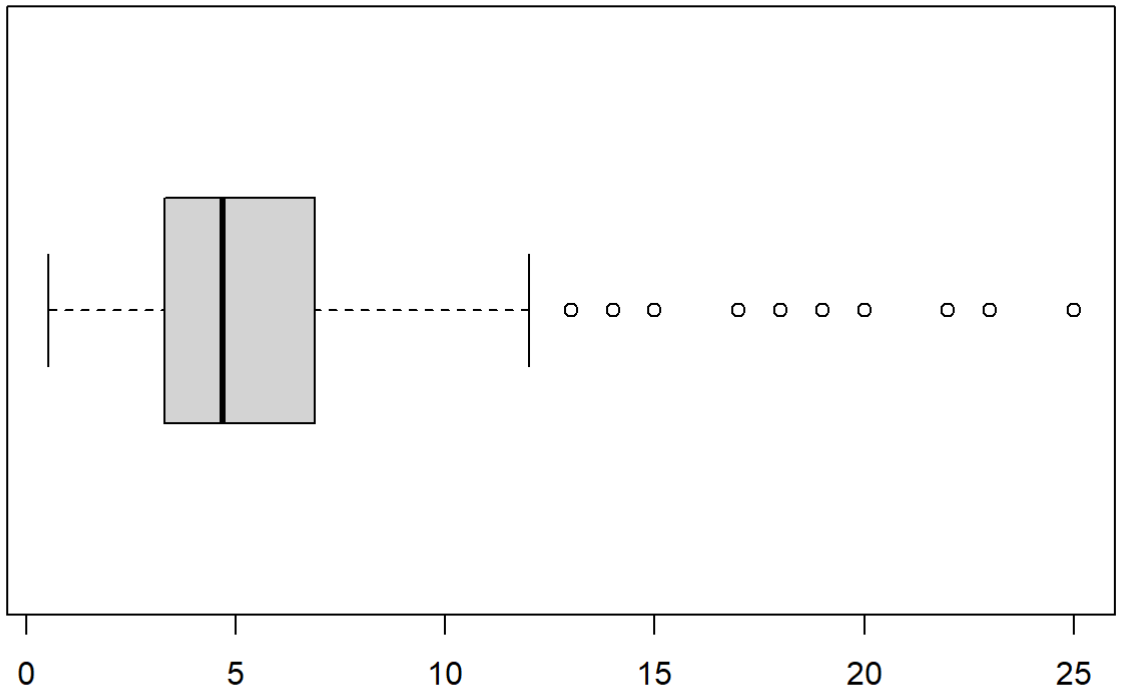
3. Clarify distribution of your sample observations, give some statistics and provide graphs.

Box plot for wage and exper

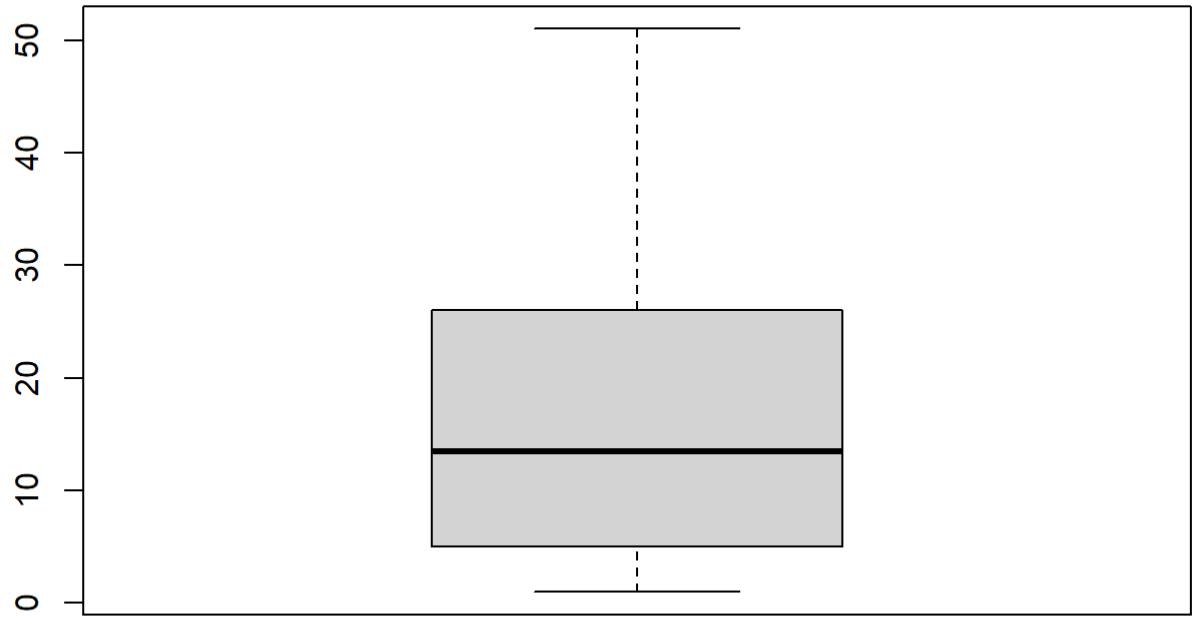
we could see that in our sample, the wage is mostly between 3 - 6 dollars per hour, which is a low number compared to the wage level in California, because the minimum wage in California is 12 USD per hour.

the labor market experience is mostly 5 -26 years.

```
boxplot(wage, horizontal = T)
```




```
boxplot(exper,horizontal = F)
```



histograms

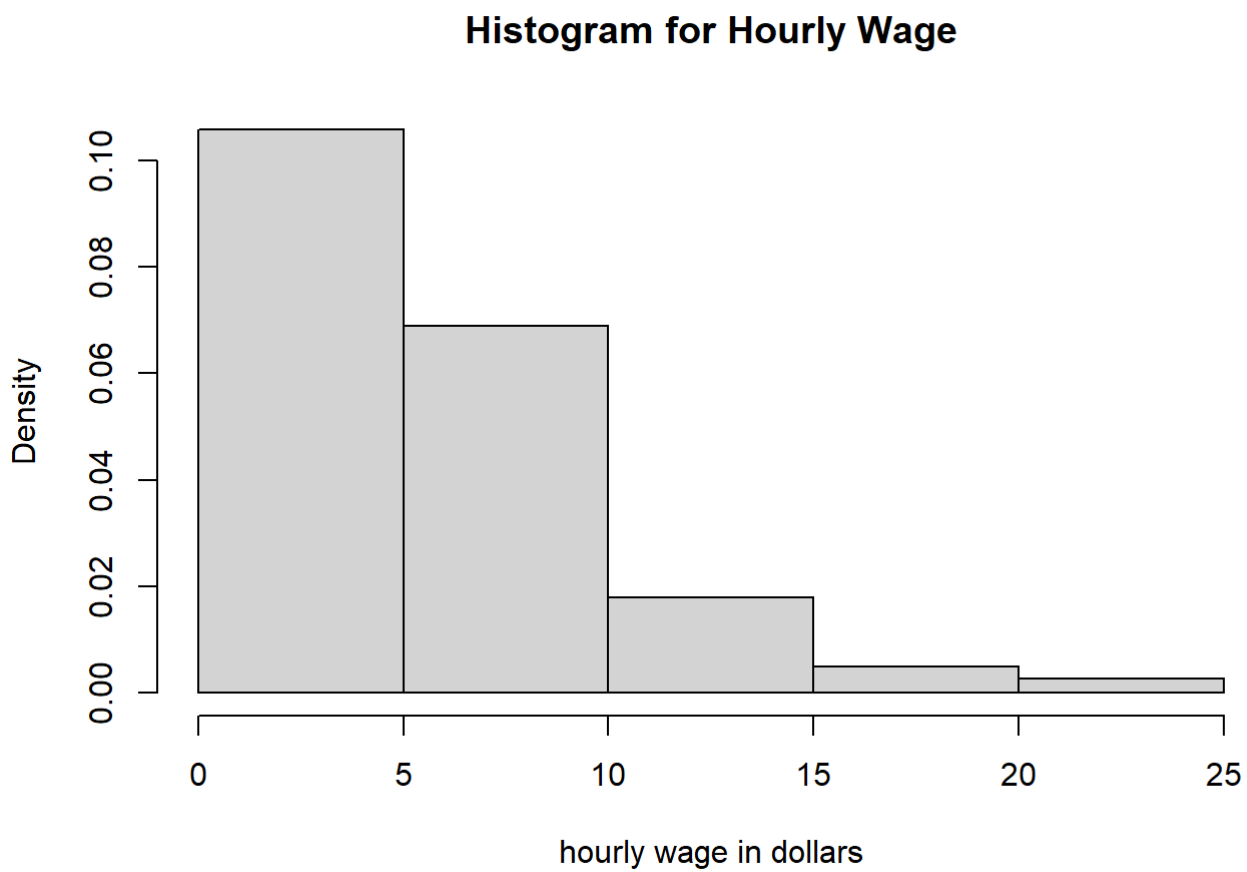
The histogram clearly shows that our data of wage and our data of labor market experience are right skewed. Besides, we could easily see that for wage and exper, the mean is greater than the median.

```
breaks=c(0,5,10,15,20,25)

table(cut(wage,breaks,right=F))
```

```
##
##   [0,5)  [5,10) [10,15) [15,20) [20,25)
##     278     181      47      13       6
```

```
hist(wage,br=breaks,freq=F, right=F, xlab="hourly wage in dollars",ylab="Density",main="Histogram for Hourly Wage")
```




```
breaks=c(0,5,10,20,30,55)

table(cut(exper,breaks,right=F))
```

```
##
##   [0,5)  [5,10) [10,20) [20,30) [30,55)
##     112      94     129      80     111
```

```
hist(wage,br=breaks,freq=F, right=F, xlab="labor market experience in years",ylab="Density",main="Histogram for Labor Market Experience")
```


```
summary(data)
```

```
##       wage            exper           female          tenure3
##  Min.   : 0.530   Min.   : 1.00   Min.   :0.0000   Min.   :0.0000
##  1st Qu.: 3.300   1st Qu.: 5.00   1st Qu.:0.0000   1st Qu.:0.0000
##  Median : 4.700   Median :13.50   Median :0.0000   Median :0.0000
##  Mean   : 5.909   Mean   :17.02   Mean   :0.4791   Mean   :0.4734
##  3rd Qu.: 6.900   3rd Qu.:26.00   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :25.000   Max.   :51.00   Max.   :1.0000   Max.   :1.0000
##       educ
##  Min.   : 0.00
##  1st Qu.:12.00
##  Median :12.00
##  Mean   :12.56
##  3rd Qu.:14.00
##  Max.   :18.00
```

4. Estimate your regression model using LS estimation method.

Simple Linear Regression Model

```
cor(wage, exper)
```

```
## [1] 0.1129904
```

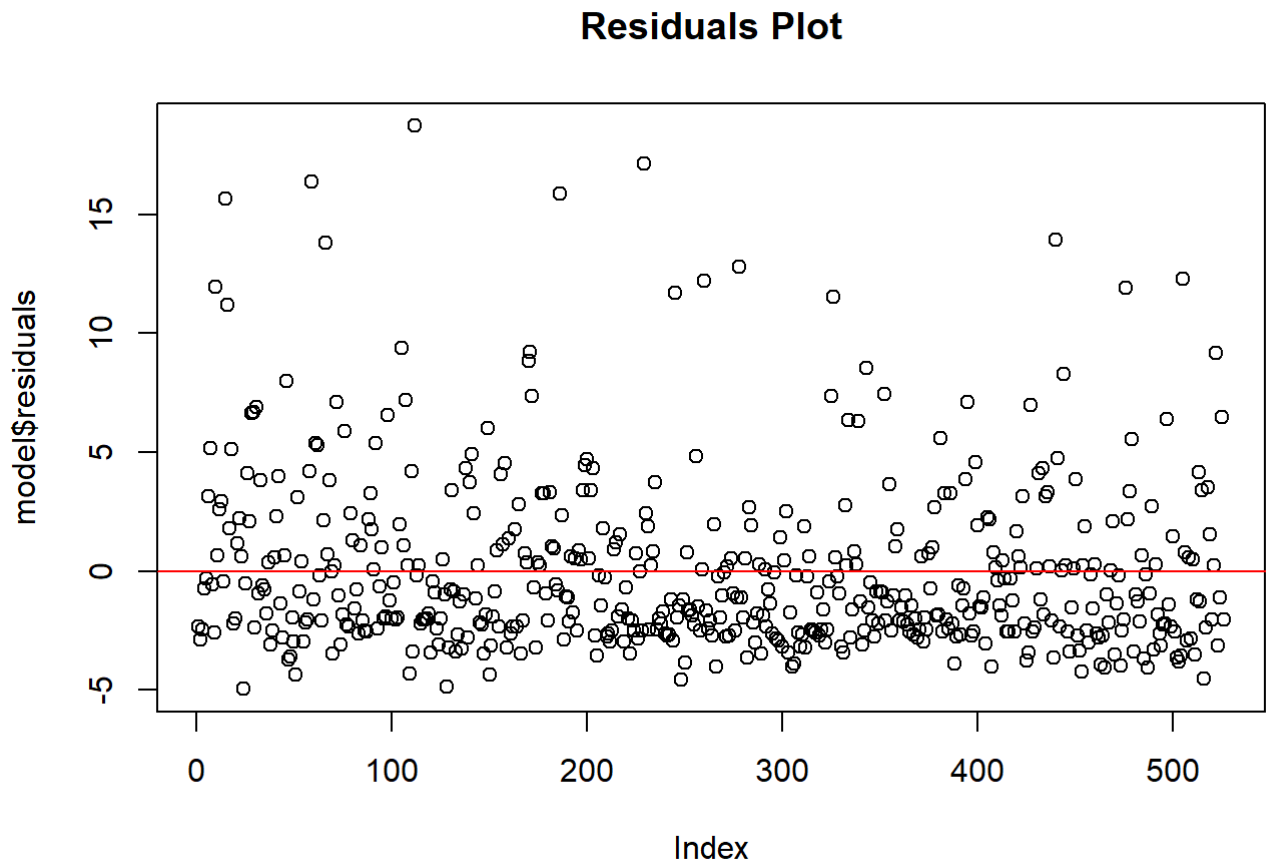
```
lm(wage~exper)
```

```
##
## Call:
## lm(formula = wage ~ exper)
##
## Coefficients:
## (Intercept)        exper
##     5.38352      0.03088
```

```
model <- lm(wage ~exper)

plot(model$residuals, main = 'Residuals Plot')

abline(h = 0, col = 'red')
```



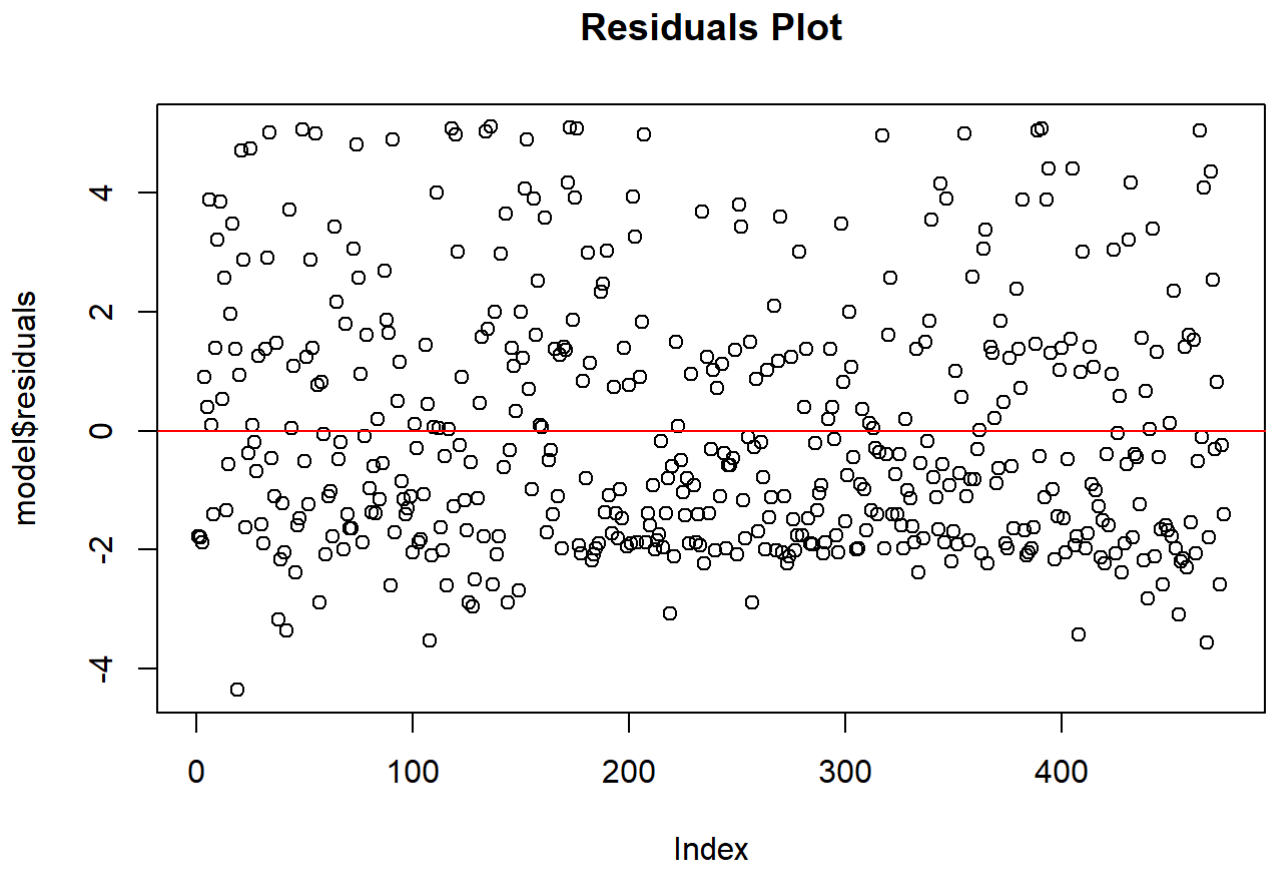
we plot the residuals, and we find out that there are some outliers. Since we plot the box plot before and we know the IQR and Q3, we take out the outliers and use wage that is under 11.

```
data2 <- data[data$wage < 11,]

wage <- data2$wage
exper <- data2$exper
model <- lm(wage ~exper)

plot(model$residuals, main = 'Residuals Plot')

abline(h = 0, col = 'red')
```



```
plot(exper, wage, xlab = 'labor market experience in years', ylab = 'hourly wage')
```


```
mean_exper = mean(exper)
mean_wage = mean(wage)
plot(exper, wage, xlab = 'labor market experience in years', ylab = 'hourly wage')

points(mean_exper,mean_wage, col = 'blue', pch = 19, cex = 2)

legend('topright', c('points of average(center)'),col = 'blue', pch = 19)
```

Regression Line

```
# y = ax + b

sd_exper = sd(exper)
sd_wage = sd(wage)

r = cor(wage,exper)
a = r * sd_wage / sd_exper
b= mean_wage - a * mean_exper

plot(exper, wage, xlab = 'labor market experience in years', ylab = 'hourly wage')

points(mean_exper,mean_wage, col = 'blue', pch = 19, cex = 2)# points of averages(center)

abline(b,a,col = 'red', lwd = 1)
abline(v = c(30,40), lty = 3)
```


```
c(r)
```

```
## [1] 0.0337019
```

5. Test for appropriateness of your estimated model, comment on the problems if any (suggest/apply a solution, if you have any ideas

Assumption 1: the regression model is linear in parameters.

```
c(a)
```

```
## [1] 0.005146165
```

```
c(b)
```

```
## [1] 4.871319
```

our regression line is y = ax + b which is 0.005x + 4.87, so x and y are linear.

Assumption 2: random sampling

our sample may not be a random sample from the population which is all the workers. As we could see in the graph below, most people in our sample has hourly wage below 10 dollars which is low, and has 0 - 10 market experience in years. The sample is biased because it contained too many low-income workers and new workers, lacking more diversed data.

```
breaks=c(0,5,10,15,20,25)

table(cut(wage,breaks,right=F))
```

```
##
##   [0,5)  [5,10) [10,15) [15,20) [20,25)
##     278     181      16       0       0
```

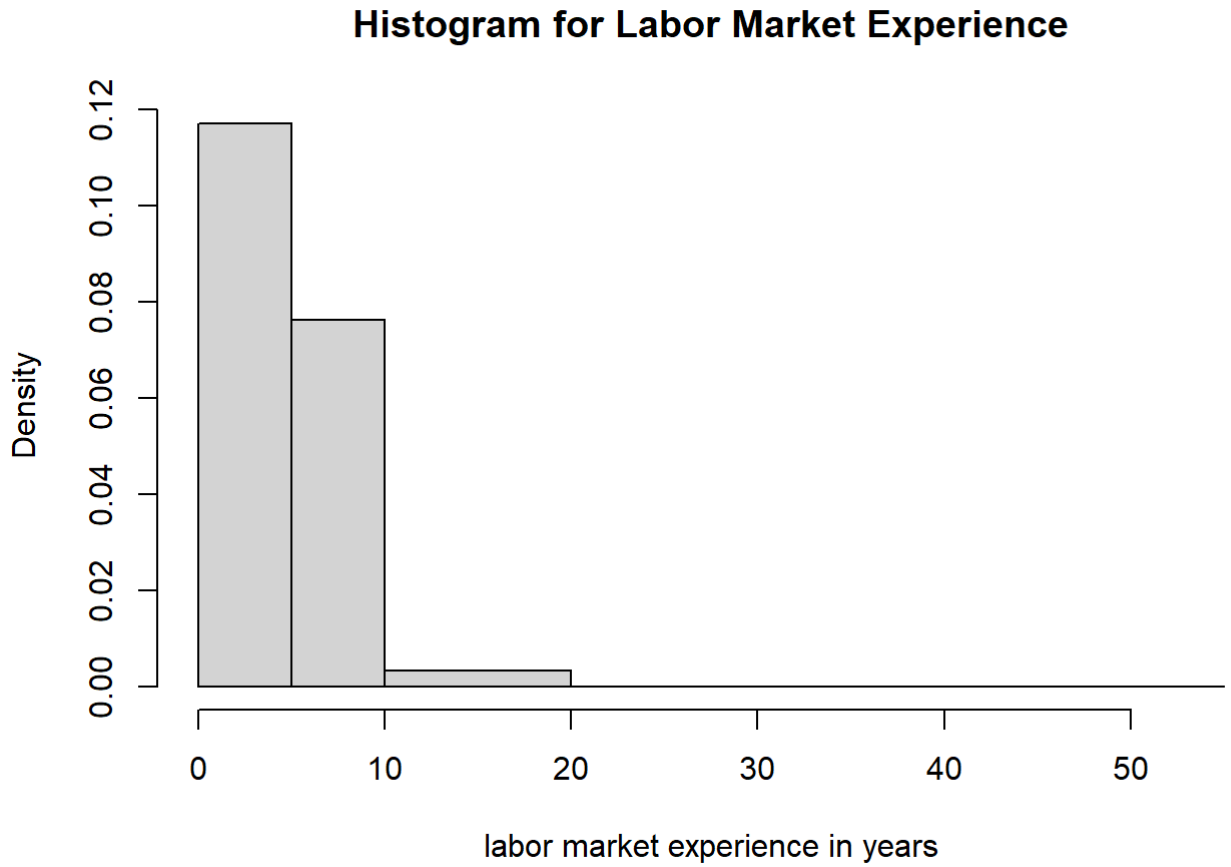
```
hist(wage,br=breaks,freq=F, right=F, xlab="hourly wage in dollars",ylab="Density",main="Histogram for Hourly Wage")
```


```
breaks=c(0,5,10,20,30,55)

table(cut(exper,breaks,right=F))
```

```
##
##   [0,5)  [5,10) [10,20) [20,30) [30,55)
##     111      88     116      63      97
```

```
hist(wage,br=breaks,freq=F, right=F, xlab="labor market experience in years",ylab="Density",main="Histogram for Labor Market Experience")
```



Assumption 3: check independent

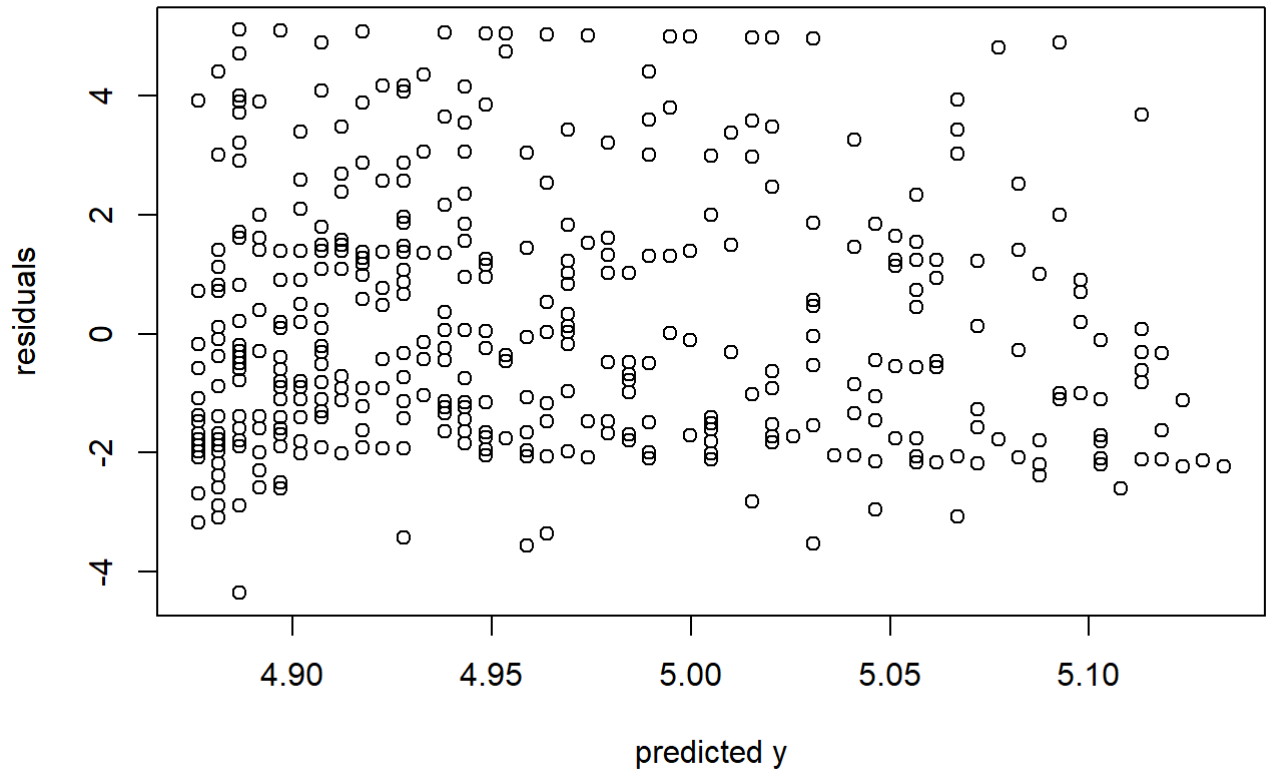
variable x is not a constant value

As shown in the Histogram for Hourly Wage, our independent variable, labor market experience in years, is not constant.

Assumption 4: check the error term is not correlated with independent variables.

As shown in the graph below, there is no pattern between residual and predicted y, indicating the zero condition mean.

```
model2 <- lm(wage ~exper)
resids <- residuals(model2)
predy <- predict(model2)
Y = resids
X = predy
plot(X,Y,xlab = 'predicted y', ylab = 'residuals')
```



Assumption 5: Homoscedasticity

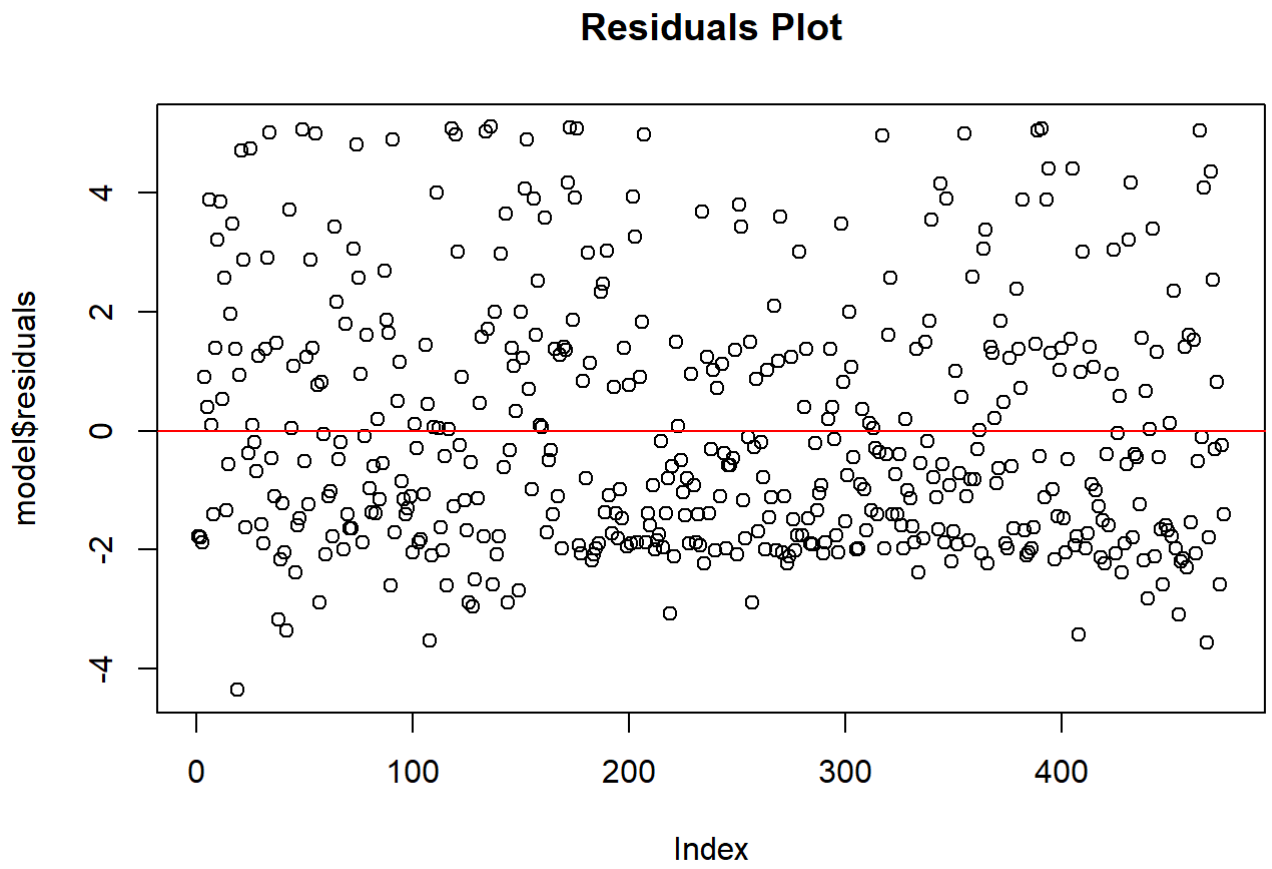
the ticness is not changing much, which satisfies the homoscedasticity.Also there is no pattern.

```
data2 <- data[data$wage < 11,]

wage <- data2$wage
exper <- data2$exper
model <- lm(wage ~exper)

plot(model$residuals, main = 'Residuals Plot')

abline(h = 0, col = 'red')
```



6. Estimate your final model and interpret your findings. Are your findings as you were expecting? If not, explain why.

The correlation coefficient is 0.033, which is a weak positive correlation because it is below 0.1. The 0.033 correlation shows that there is only a minor relationship between hourly wage and labor market experience in years.

Our finding is not what we expected. We had the hypothesis that if a person has more labor market experience, his hourly wage should be higher, which should be a strong positive correlation.

Here are the possible reasons why our finding is not what we expect. First, according to the histogram shown earlier, we could clearly see that a big part of our sample only has a wage between 0 - 10. This could cause problems because a low-paying job usually does not require many skills, which at the same time does not require any labor experience. Besides, not only do we lack the data that records wage over 10, most people in our sample only has 0 - 10 years labor experience, which means we are lacking data about people who work more than 10 years. Thus, while doing a linear regression model, if we gather more diverse data, we may have new findings.