Question 1:

1. The code in R:

m1<-lm(formula = SALNOW ~ 1, data = bank)

summary(m1)

The model equation is: SALNOW = 13767.8 = mean(SALNOW)

1. The residual sum of squares for this model = 22066639270

Code in R:

residual = bank$SALNOW - fitted(m1)

residual\_sq = sum(residual^2)

Or you can call

rss <- deviance(m1)

Or

sum(resid(m1)^2)

1. The residual standard error = 6830.265

Code in R:

residual\_std\_error = sqrt(residual\_sq / 473)

Or you can read the result from the model summary where the result is the same.

Question 2:

1. The code in R:

m2<-lm(SALNOW ~EDLEVEL, data=bank)

summary(m2)

The model equation is SALNOW = 1563.96\*EDLEVEL - 7332.47

1. The residual sum of squares for this model = 12438124428

Code in R:

residual1 = bank$SALNOW - fitted(m2)

residual\_sq\_1 = sum(residual1^2)

Or

rss <- deviance(m2)

1. The residual standard error = 5133.416

Code in R:

residual\_std\_error1 = sqrt(residual\_sq\_1 / 472)

Or you can read the result from the model summary where the result is the same.

Question 3:

1. The code in R:

m3<-lm(SALNOW ~EDLEVEL + SEX, data=bank)

summary(m3)

The model equation is SALNOW = 1356.67\*EDLEVEL + 3369.38\*SEX - 6369.78

1. The residual sum of squares for this model = 11272531174

Code in R:

residual2 = bank$SALNOW - fitted(m3)

residual\_sq\_2 = sum(residual2^2)

Or

rss <- deviance(m3)

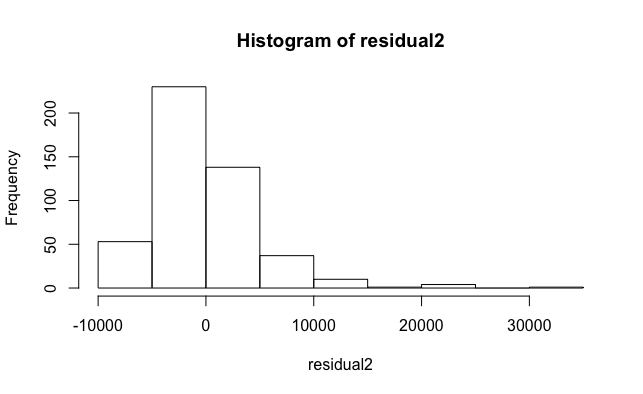
1. The residual standard error = 4892.156

residual\_std\_error2 = sqrt(residual\_sq\_2 / 471)

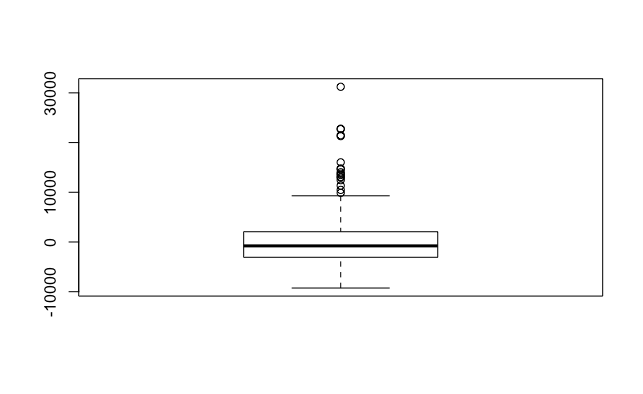
Or you can read the result from the model summary where the result is the same.

Question 4:

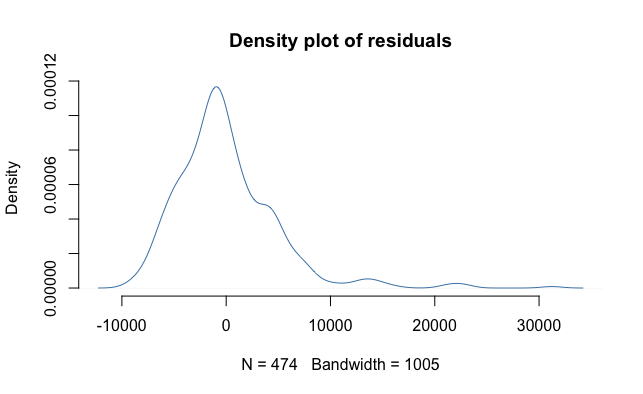
1. Histogram:



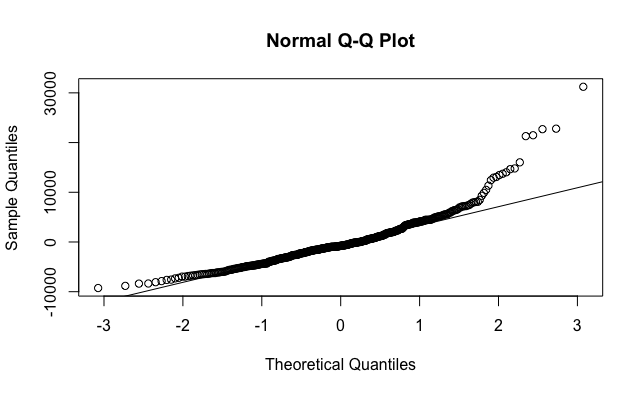
Box plot:



Density plot:



QQ plot:



All the graphs show that the residual data is right-skewed. The skewness indicates that the residual is not normal distributed!

1. Code in R:

ks.test(residual2,"pnorm",mean(residual2),sd(residual2))

One-sample Kolmogorov-Smirnov test

data: residual2

D = 0.10292, p-value = 8.706e-05

alternative hypothesis: two-sided

The p-value again reassure that the residual from linear regression is not normal distributed. There are big differences in the end of the value from the data set with the predicted value we have.

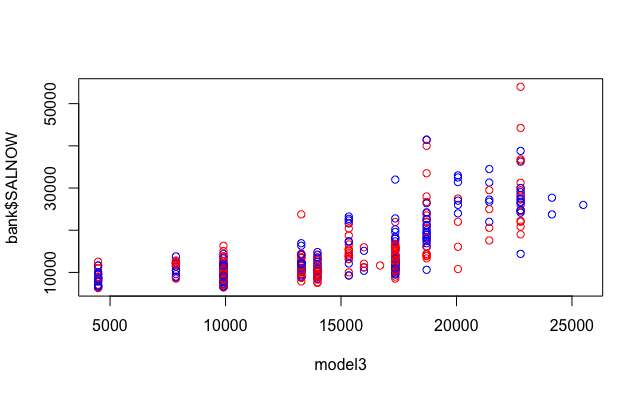
Question 5:

Code from R:

model3 = fitted(m3)

plot(model3,bank$SALNOW,col = c("red", "blue"))

The plot:



Code in R for Pearson correlation coefficient:

cor.test(model3,bank$SALNOW,method="pearson")

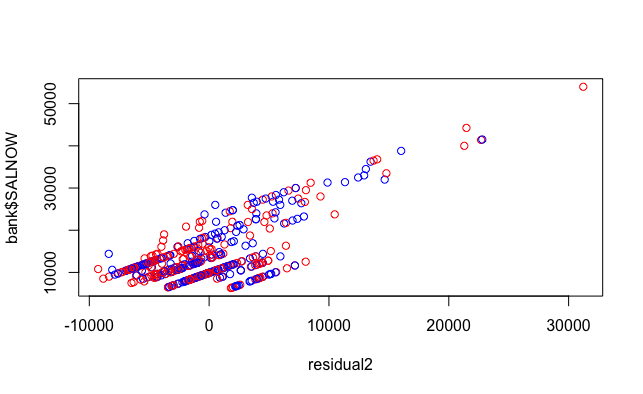
The Pearson correlation coefficient = 0.6993994.

Question 6:

Code from R:

plot(residual2,bank$SALNOW,col = c("red", "blue"))

The plot:



Code in R for Pearson correlation coefficient:

cor.test(residual2,bank$SALNOW,method="pearson")

Pearson's product-moment correlation

data: residual2 and bank$SALNOW

t = 22.202, df = 472, p-value < 2.2e-16

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

0.6676434 0.7561230

sample estimates:

cor

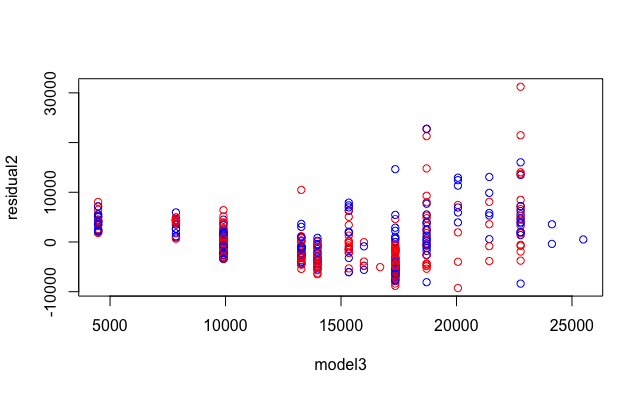
0.714731

The result is the same as I expected. We can see the model have the best prediction in the middle of the dataset and in the end the residual is getting bigger and bigger (the blue circle is not as close to the red circle as before). The Pearson correlation coefficient indicates that the correlation between the residual and the data value is 71% (the middle part of the dataset is quite close but the tail the linear model performance is not good)

Question 7:

The code from R:

plot(model3,residual2,col = c("red", "blue"))



The code from R for Pearson correlation coefficient:

cor.test(residual2,model3,method="pearson")

Pearson's product-moment correlation

data: residual2 and model3

t = -4.2348e-16, df = 472, p-value = 1

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

-0.09006565 0.09006565

sample estimates:

cor

-1.949243e-17

The plot does look like what i expected. The residuals are a bit further away from the predicted values. The Pearson correlation coefficient is -1.949243e-17 the correlation is very small indicates that’s the

Question 8:

1. The variables that are significant included: SALBEG, TIME, EDLEVEL, JOBCATCollegeTrainee, JOBCATExempt, JOBCATMBATrainee,JOBCATSecurity, JOBCATTechnical

The code from R:

m4<-lm(formula = SALNOW ~ SALBEG + SEX + TIME + AGE + EDLEVEL + WORK + JOBCAT + MINORITY , data = bank)

1. 84.36% variability in salaries is explained by this model.
2. There are no evidences indicate the relationship between sex and the salaries a person get.

Question 9:

1. The variables that are significant included: SALBEG, SEXMale, TIME,EDLEVEL,WORK, JOBCATCollegeTrainee, JOBCATExempt, JOBCATMBATrainee

The code from R:

m5<-lm(formula = SALNOW ~ SALBEG + SEX + TIME + EDLEVEL + WORK + JOBCAT + MINORITY , data = bank)

1. 84.31% variability in salaries is explained by this model.
2. There is an evidence indicates the relationship between sex and the salaries a person get

Question 10:

Analysis of Variance Table

Model 1: SALNOW ~ 1

Model 2: SALNOW ~ EDLEVEL

Model 3: SALNOW ~ EDLEVEL + SEX

Model 4: SALNOW ~ SALBEG + SEX + TIME + AGE + EDLEVEL + WORK + JOBCAT +

MINORITY

Model 5: SALNOW ~ SALBEG + SEX + TIME + EDLEVEL + WORK + JOBCAT + MINORITY

Res.Df RSS Df Sum of Sq F Pr(>F)

1 473 2.2067e+10

2 472 1.2438e+10 1 9628514842 1319.6883 <2e-16 \*\*\*

3 471 1.1273e+10 1 1165593254 159.7567 <2e-16 \*\*\*

4 460 3.3562e+09 11 7916346997 98.6380 <2e-16 \*\*\*

5 461 3.3738e+09 -1 -17610028 2.4136 0.121

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

>

The increase in the number of predictors decrease the residual sum square because you have more feature of the data ==> explainability of the model increase.

The increase in the number of predictors decrease the sum of square predictors. The sum of square predictors have the positive correlation with the F-value. The results show that there is no relation between sexism and salaries of employees in the dataset.