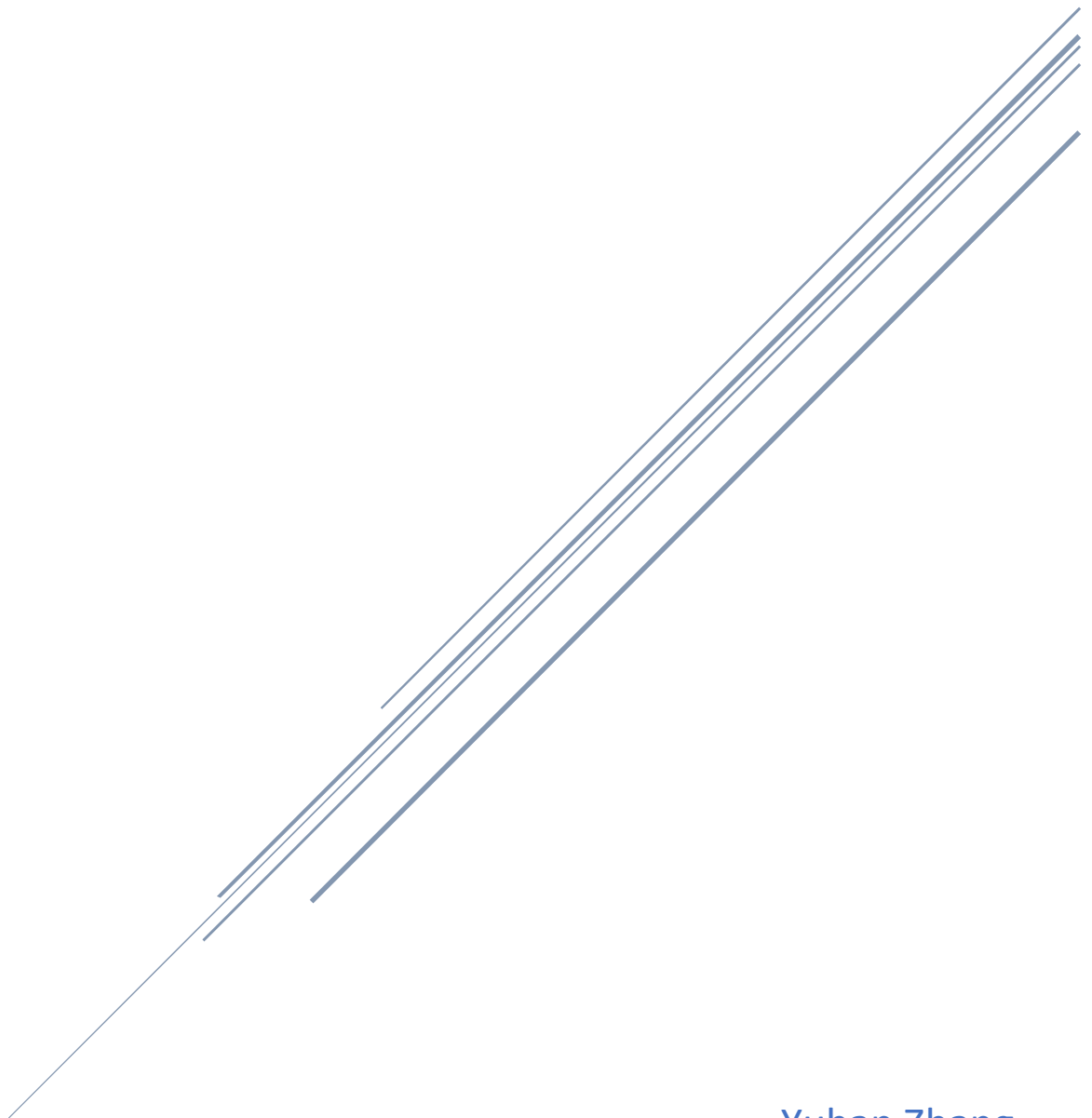# STAT 350

Assignment6

Yuhan Zhang
301345627

Q1

a)

```r
###### Q 1######
##a
```{r}
setwd("C:/Users/carol/Desktop/stat350")
prestige <- read.csv("prestige.csv")

lm <- lm(prestige ~ education+income+type, data=prestige)
summary(lm)
anova(lm)

reduce_type <- lm(prestige ~ education+income, data=prestige)
anova(reduce_type, lm) #p-value 1.208e-06 is small, "type" is good

reduce_income <- lm(prestige ~ education+type, data=prestige)
anova(reduce_income, lm) #p-value 5.124e-08 is small, "income" is good

reduce_education <- lm(prestige ~ income+type, data=prestige)
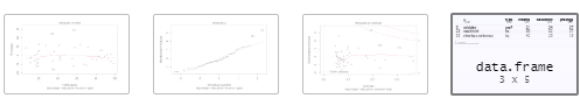anova(reduce_education,lm) #p-value 0.004164 is small, "education" is good
```
```

```
#Ho: B1=B2=B3=0
#Ha: at least one of those factors is not equal to 0.
#p-value: < 2.2e-16
#Conclution: reject Ho. all of three factors are significant.
```

b)

```r
##b
```{r}
plot(lm, which=c(1,2,5))
prestige[c(6,28,33),]
```
```

data.frame
3 x 5

	X <chr>	type <chr>	income <int>	education <int>	prestige <int>
6	minister	prof	21	84	87
28	machinist	bc	36	32	57
33	streetcar.motorman	bc	42	26	19

3 rows

```
##6(minister),28(machinist),33(motorman) are outliers. obviously differential
##17, 27 only have high leverage, but small coke distance.Therefore, we can classify them as not differential.
```

c)

```r
##c
```{r}
install.packages("MASS")
library(MASS)
rr.lm <- rlm(prestige~education+income+type,  data=prestige,psi=psi.huber)
summary(rr.lm)
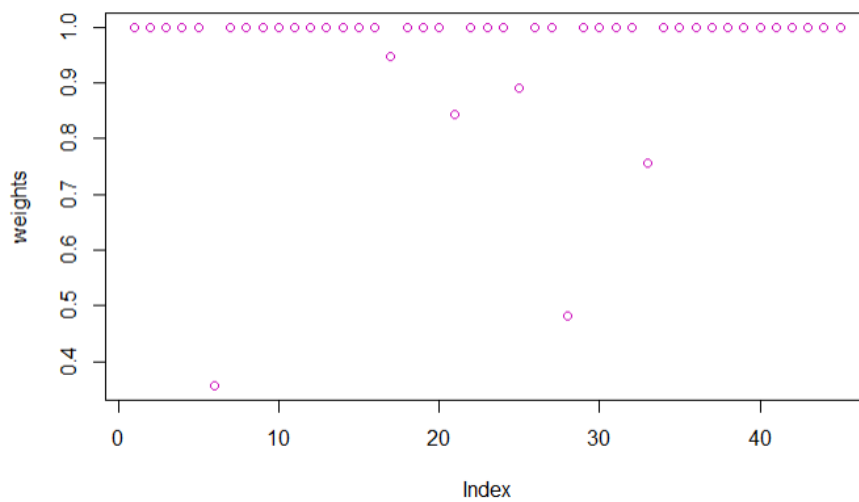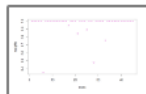rr.lm$w
```

#Yes. They are all significant.
```

d)

```r
77 - ##d
78 - ```{r}
79  rr.lm$w
80  plot(rr.lm$w, col=6,ylab="weights")
81
82  which(rr.lm$w == min(rr.lm$w))
83  prestige$X[6]
84 - ```
```

```
[1] "minister"
```

```r
85 - #Minister(6-th) is the smallest.
86
```

Q2

a)
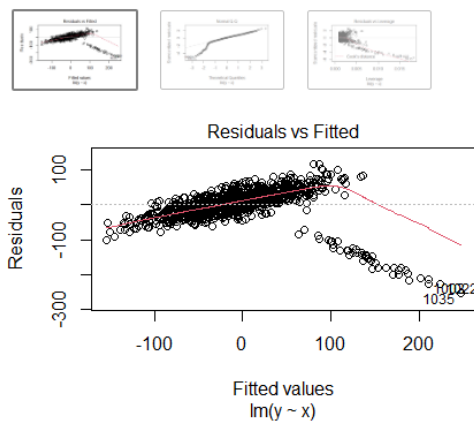
```r
###### Q2 ######

###a
```{r}
setwd("C:/Users/carol/Desktop/stat350")
assign6 <- read.csv("assignment6.csv")
lm_Q2 <- lm(y~x, data=assign6)
summary(lm_Q2)$coef
```
```

```
            Estimate Std. Error t value  Pr(>|t|)
(Intercept)    -9.07       1.46   -6.21  7.64e-10
x              49.85       1.23   40.46 2.41e-216
```

```r
#y= 49.85x -9.07
```

b)

```r
###b
```{r}
plot(lm_Q2, which=c(1,2,5))
```
```

Residuals vs Fitted
lm(y ~ x)

```r
##It is uncomfortable as there are many outliers shown by residual plot. The distribution is not normal shown by QQ plot, and
there are some variables with large coke's distance.
```

c)

```r
###c
```{r}
install.packages("MASS")
library(MASS)
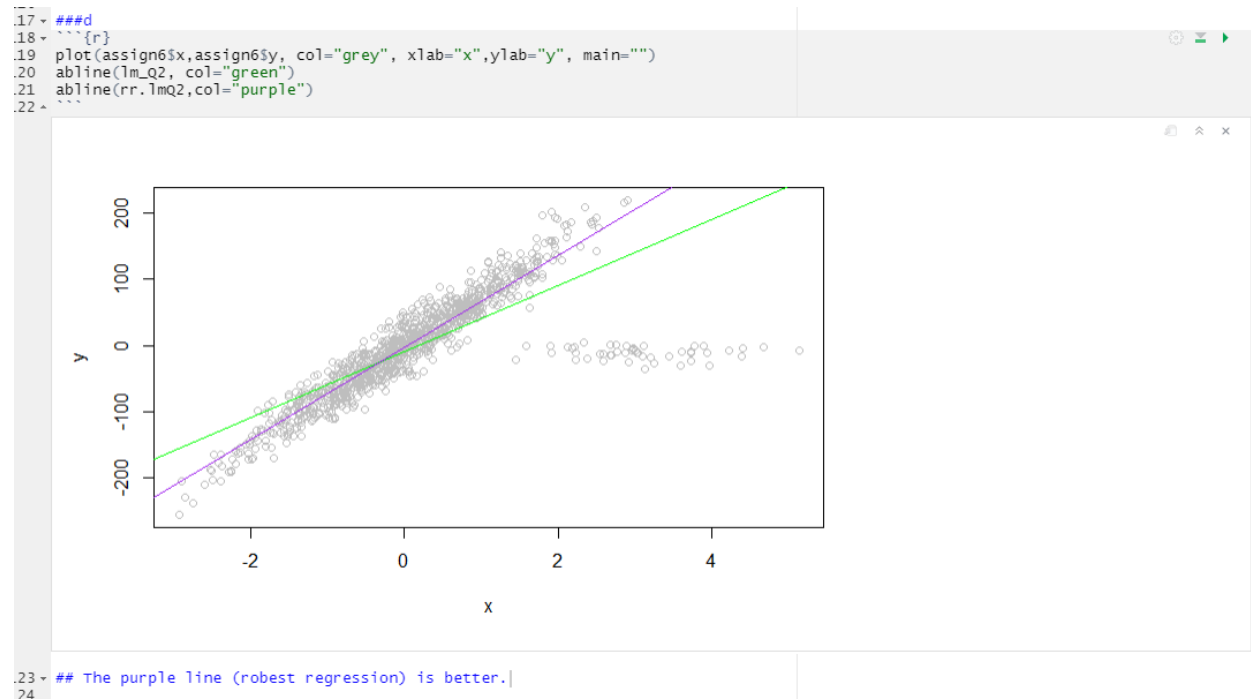rr.lmQ2 <- rlm(y~x, data=assign6, psi=psi.huber)
rr.lmQ2
```
```

```
Error in install.packages : Updating loaded packages
package �MASS� was built under R version 4.0.3Call:
rlm(formula = y ~ x, data = assign6, psi = psi.huber)
Converged in 8 iterations

Coefficients:
(Intercept)          x
      -2.95      69.74

Degrees of freedom: 1050 total; 1048 residual
Scale estimate: 21
```

```r
#y=69.74x -2.95
```

d)

```r
.17 ###d
.18 ```{r}
.19 plot(assign6$x,assign6$y, col="grey", xlab="x",ylab="y", main="")
.20 abline(lm_Q2, col="green")
.21 abline(rr.lmQ2,col="purple")
.22 ```
```



```r
.23 ## The purple line (robest regression) is better.
.24
```
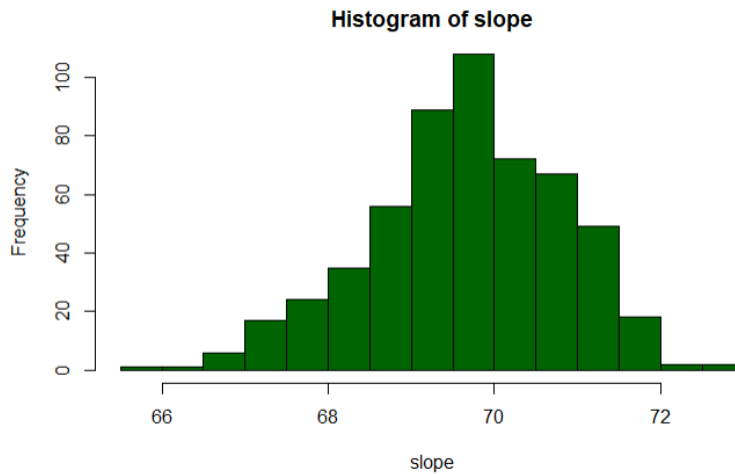
e)

```r
25
26 ###e
27 ```{r}
28 install.packages("MASS")
29 library(MASS)
30 index= sample.int(1050, size=1050, replace=TRUE)
31 rlm(y~x, psi=psi.huber, data=assign6[index,])$coef[2]
32
33 slope<- NULL
34
35 set.seed(666)
36 for(i in 1:1000){
37   index = sample.int(1050, size=1050, replace=TRUE)
38   slope[i]= rlm(y~x, psi=psi.huber,data=assign6[index,])$coef[2]
39 }
40
41 hist(slope, col="dark green")
42 ```
```


Histogram of slope

```r
126 ###e
127 ```{r}
128 install.packages("MASS")
129 library(MASS)
130 index= sample.int(1050, size=1050, replace=TRUE)
131 rlm(y~x, psi=psi.huber, data=assign6[index,])$coef[2]
132
133 slope<- NULL
134
135 set.seed(666)
136 for(i in 1:1000){
137   index = sample.int(1050, size=1050, replace=TRUE)
138   slope[i]= rlm(y~x, psi=psi.huber,data=assign6[index,])$coef[2]
139 }
140
141 hist(slope, col="dark green")
142 ```
```

```
Error in install.packages : Updating loaded packages
package ◆MASS◆ was built under R version 4.0.3 x
70

Restarting R session...
```

```r
143 #estimated slope is 70.|
144
```

f) &g)

```
144
145 ###f
146 ```{r}
147 quantile(slope, probs=c(0.025,0.975))
148 ```
```

```
    2.5% 97.5%
    67.3  71.7
```

```
149 #The CI is (67.3,71.7)
150
151 ###g
152 #H0:slope=0
153 #Ha: slope is not equal to 0
154 #95% CI(67.3, 71.7) does not contain zero, so reject H0. Therefore, we can conslude that the slope is different than zero at
    0.5% significant level.
```