HUDM 5126 Linear Models and Regression Analysis HW3

Yifei Dong

9/20/2020

0. Data Preparation

```r
setwd("/Users/yifei/Documents/Teachers College/Linear Models and Regression/Week 3")
getwd()
```

```
## [1] "/Users/yifei/Documents/Teachers College/Linear Models and Regression/Week 3"
```

```r
library(ggplot2)
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```r
library(hrbrthemes)
library(extrafont)
```

```
## Registering fonts with R
```

1. Grade Point Average

This question is adapted from Q3.3. Refer to Grade Point Average data in Problem 1.19.

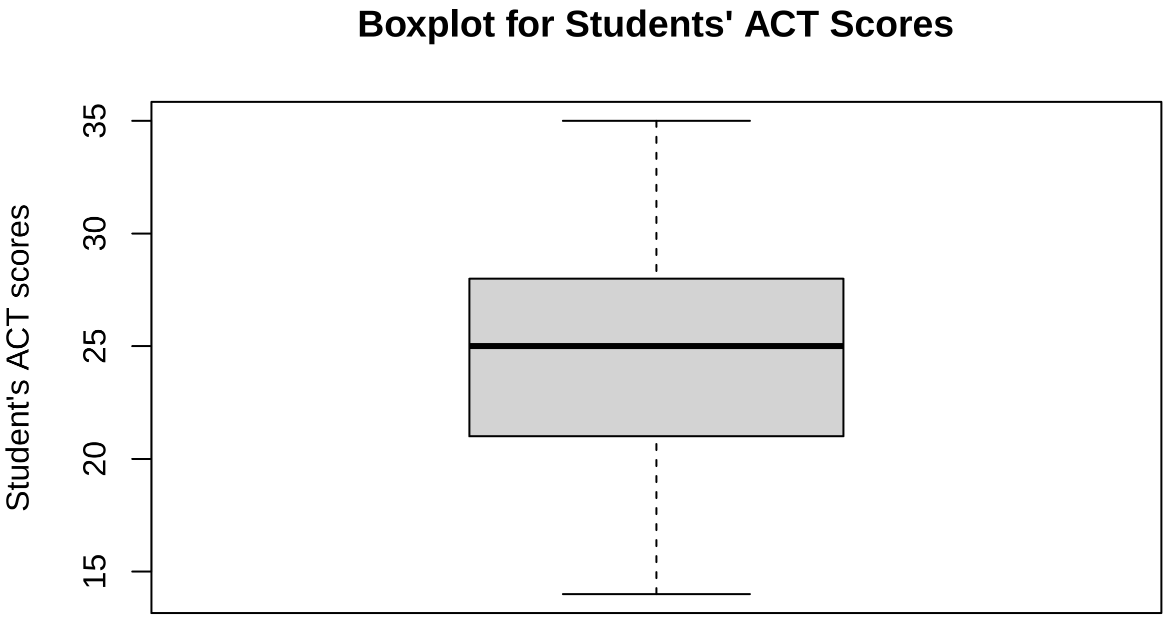
a. Prepare a boxplot of the ACT scores (X variable). Are there any noteworthy features on the plot?

```
# Load data
mydata<-read.table(
  "http://users.stat.ufl.edu/~rrandles/sta4210/Rclassnotes/data/textdatasets/KutnerData/Chapter%20%201%
names(mydata)=c("Y","X")
```

```
# Descriptive values for the variable X
summary(mydata$X)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   14.00   21.00   25.00   24.73   28.00   35.00
```

```
# Boxplot of the ACT Scores
boxplot(mydata$X,main="Boxplot for Students' ACT Scores",ylab="Student's ACT scores")
```



ANSWER: The boxplot looks “normal”. It seems symmetric, which most of data clustered around the middle but with no extreme outliers. It looks to be from a random sample.

b. Prepare a histogram of the residuals. What information does the plot provide?

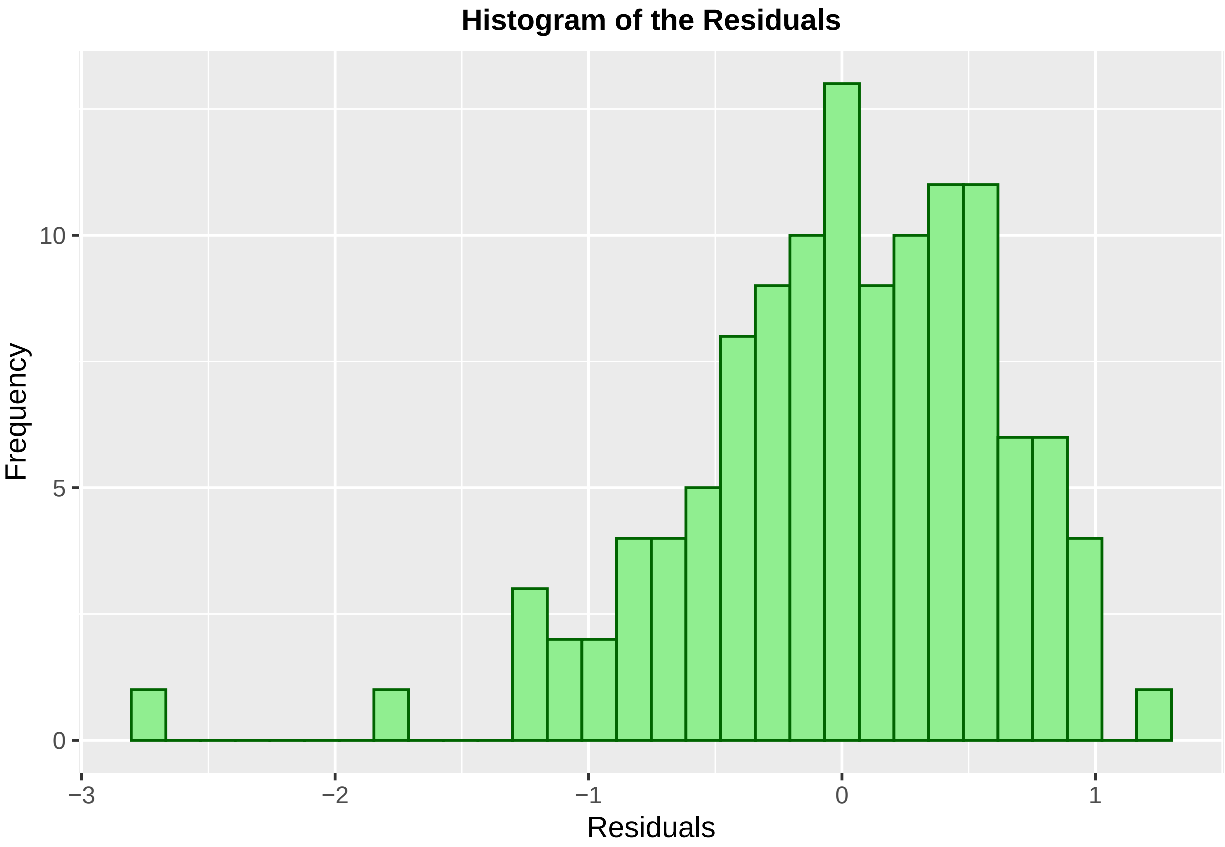
```
# Fitting the Simple Linear Regression Model
reg<-lm(mydata$Y~mydata$X)
summary(reg)
```

```
##
## Call:
## lm(formula = mydata$Y ~ mydata$X)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.74004 -0.33827  0.04062  0.44064  1.22737
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.11405    0.32089   6.588  1.3e-09 ***
## mydata$X     0.03883    0.01277   3.040  0.00292 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6231 on 118 degrees of freedom
## Multiple R-squared:  0.07262,    Adjusted R-squared:  0.06476
## F-statistic:  9.24 on 1 and 118 DF,  p-value: 0.002917
```

```
# Obtain the residuals
e<-resid(reg)
mydata<-cbind(mydata,e)
```

```
# Histogram of the residuals
p1<-ggplot(mydata,aes(x=e))+geom_histogram(color="darkgreen",fill="lightgreen")+
  labs(x="Residuals",y="Frequency",title = "Histogram of the Residuals")+
  theme(plot.title = element_text(color = "black",size=11,face = "bold",hjust = 0.5))
p1
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
mean(mydata$e)
```

```
## [1] -2.452894e-17
```

```
median(mydata$e)
```

```
## [1] 0.04061829
```

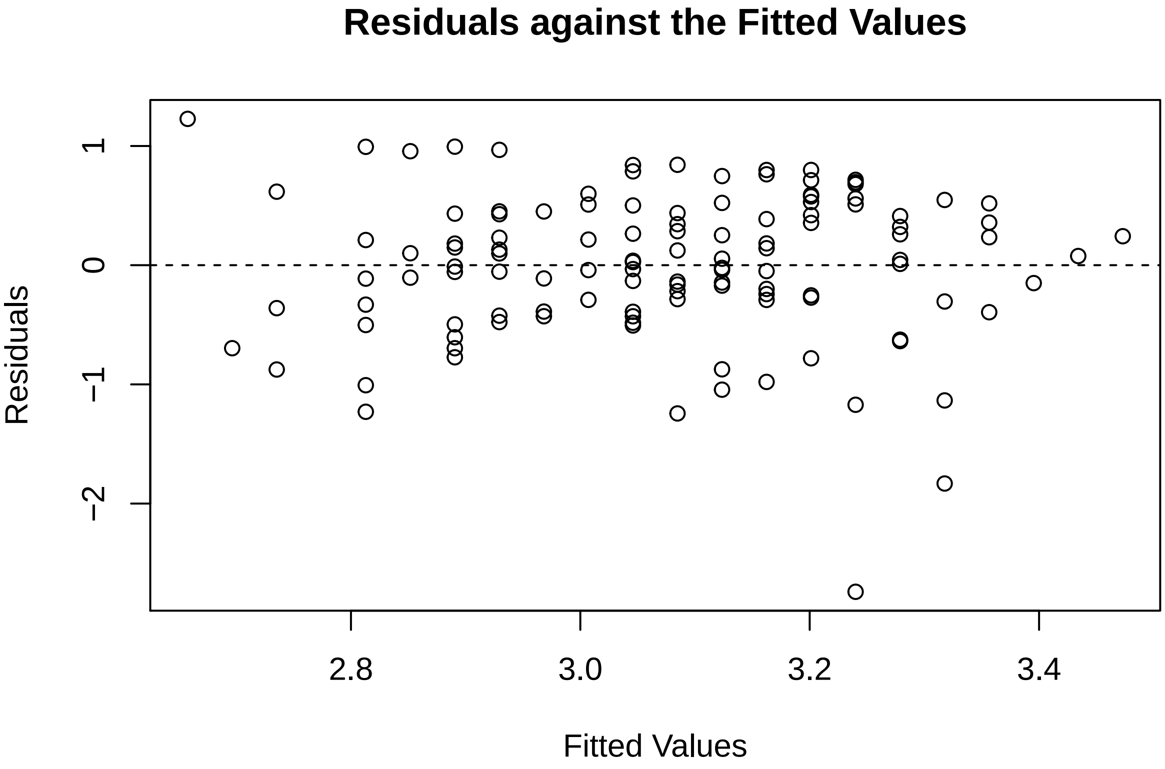
```
mean(mydata$e)<median(mydata$e)
```

```
## [1] TRUE
```

ANSWER: According to the histogram of the residuals, the residuals are left-skewed, which means the mean of residuals is smaller than the median of residuals.

c. Plot the residuals against the fitted value $\widehat{Y}$. What are your findings about departures from the regression assumption?

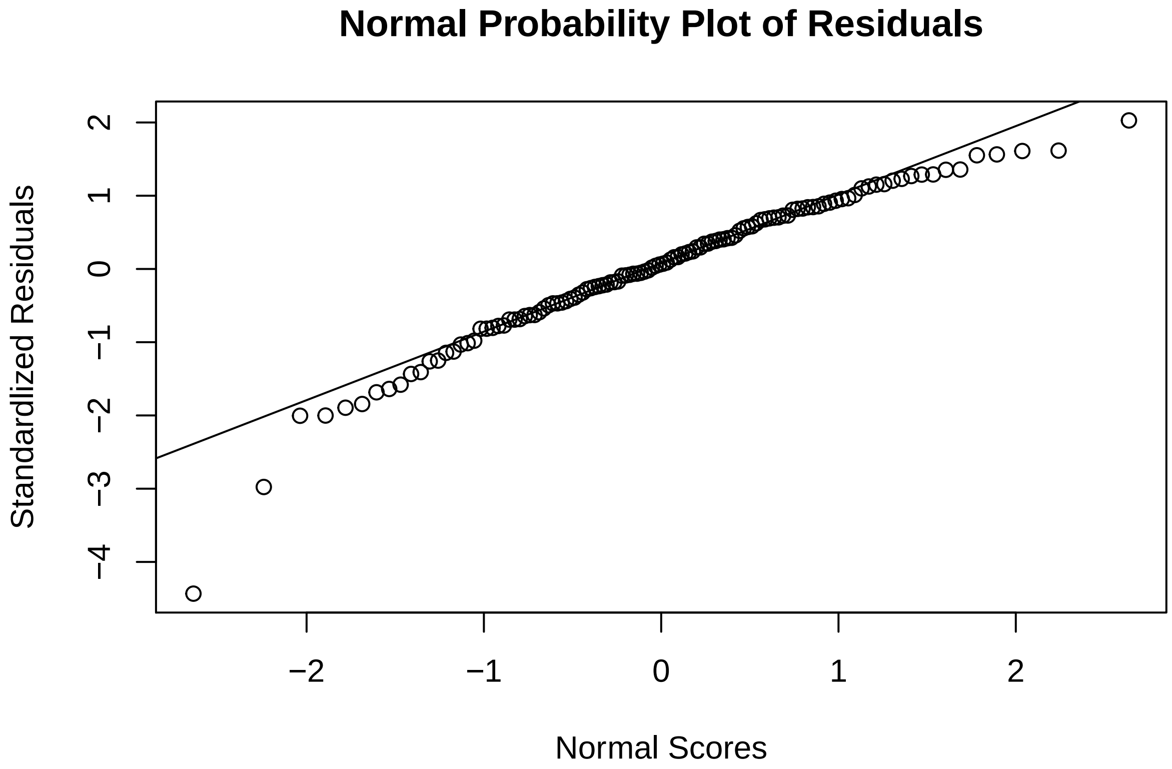
```
# Find the fitted values first
y.hat<-predict(reg)
mydata<-cbind(mydata,y.hat)
plot(mydata$y.hat,mydata$e,main = "Residuals against the Fitted Values",
     xlab="Fitted Values",ylab="Residuals")
abline(h=0,lty=2)
```



ANSWER: The residuals against the fitted value plot is relatively shapeless without clear patterns in the data and be generally symmetrically distributed around the 0 line, except for those outliers. Therefore, the constant variance assumption is hold or we can say there is no heteroskedasticity.

d. Prepare a normal probability plot of the residuals. Test the reasonableness of the normality assumption with the KS test using $\alpha = 0.05$. What do you conclude?

```
# A normal probability of the residuals
# We compute the standardlized residuals with the standard function first.
stdres<-rstandard(reg)
qqnorm(stdres,xlab="Normal Scores", ylab="Standardlized Residuals",
       main = "Normal Probability Plot of Residuals")
qqline(stdres)
```



```
# Test the reasonableness of the normality assumption: Kolmogorov-Smirnov test for normality
ks.test(rstandard(reg),"pnorm")
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  rstandard(reg)
## D = 0.057725, p-value = 0.8188
## alternative hypothesis: two-sided
```

ANSWER: Since P-value=0.8188, which is greater than 0.05, we cannot reject the null hypothesis. Therefore, normality assumption holds.

e. Conduct the BP test to determine if the variance varies with the level of X. Use $\alpha = 0.01$. State your conclusion. Does your conclusion support your preliminary findings in part c)?

```
# Breusch-Pagan test for constant variance
bptest(reg)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  reg
## BP = 0.29397, df = 1, p-value = 0.5877
```

ANSWER: P-value=0.5877 is not less than 0.01. Do not reject H_0. Conclude error variance constant. This conclusion support my preliminary findings in part c).

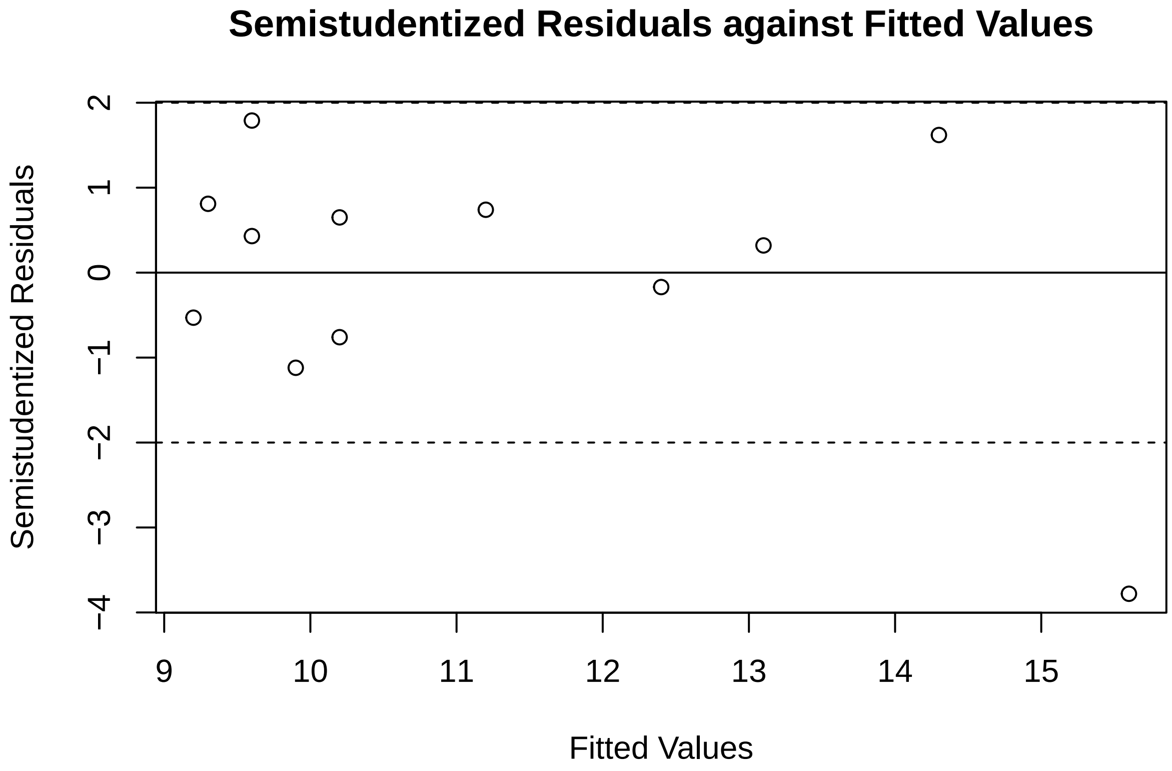
2. Per capita earnings Q3.10

```r
mydata2<-read.table("http://users.stat.ufl.edu/~rrandles/sta4210/Rclassnotes/data/textdatasets/KutnerDa
names(mydata2)=c("Fitted Values","Semistudentized Residuals")
mydata2
```

```
##    Fitted Values Semistudentized Residuals
## 1            9.9                     -1.12
## 2            9.3                      0.81
## 3           10.2                     -0.76
## 4            9.6                      0.43
## 5           10.2                      0.65
## 6           12.4                     -0.17
## 7           14.3                      1.62
## 8            9.6                      1.79
## 9            9.2                     -0.53
## 10          15.6                     -3.78
## 11          11.2                      0.74
## 12          13.1                      0.32
```

a. plot the semistudentized residuals against the fitted values. What does the plot suggest?

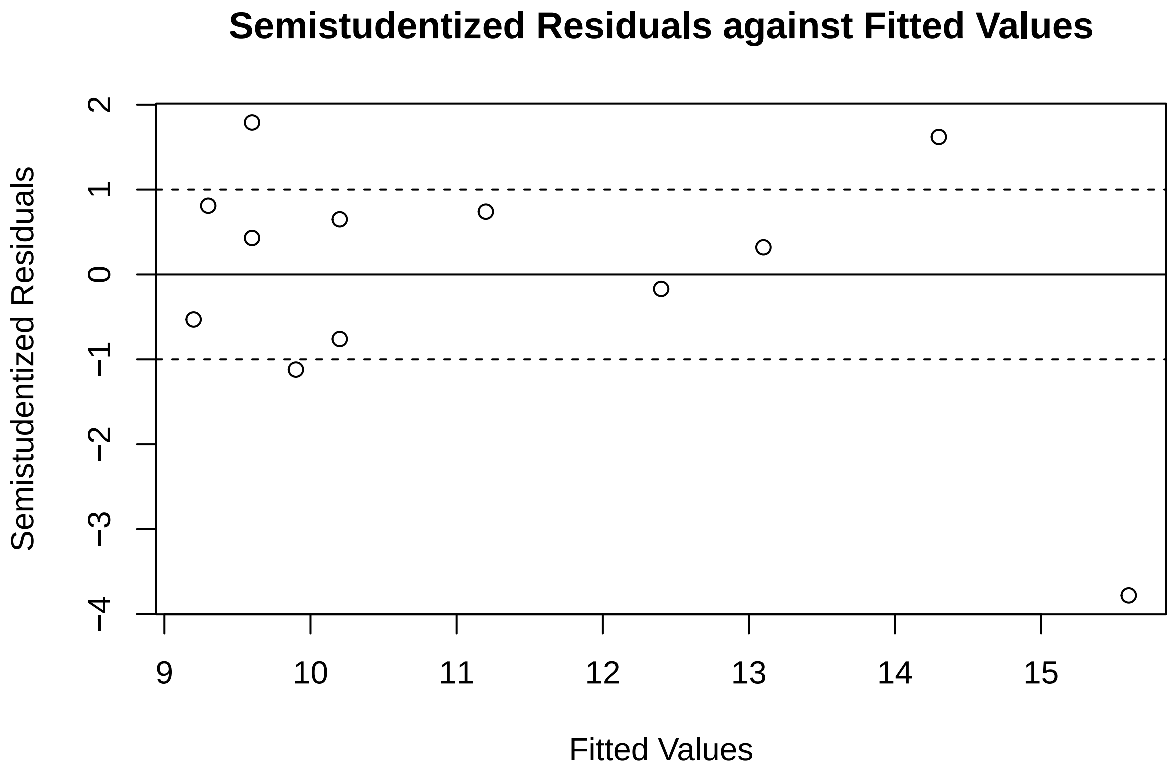
```r
plot(mydata2$`Fitted Values`,mydata2$`Semistudentized Residuals`,
     main = "Semistudentized Residuals against Fitted Values",
     xlab="Fitted Values",ylab="Semistudentized Residuals")
abline(h=0)
abline(h=-2,lty=2)
abline(h=2,lty=2)
```



ANSWER: Based on the plot, there is an outlier outside ±2 standard deviation in the observations. The outlier is at the largest predicted value, which is $\widehat{Y} = 15.6$ and it can be very influencial beacuse it is far away from the 0 line, which could change the slope coeffieicent significantly.

b. How many semistudentized residuals are outside ±1 standard deviation? Approximately how many would you expect to see if the normal error model is appropriate?

```r
plot(mydata2$`Fitted Values`,mydata2$`Semistudentized Residuals`,
     main = "Semistudentized Residuals against Fitted Values",
     xlab="Fitted Values",ylab="Semistudentized Residuals")
abline(h=0)
abline(h=-1,lty=2)
abline(h=1,lty=2)
```



ANSWER: 4 semistudentized residuals are outside ±1 standard deviation. If the normal error model is appropriate, there should be 4 points fall outside.

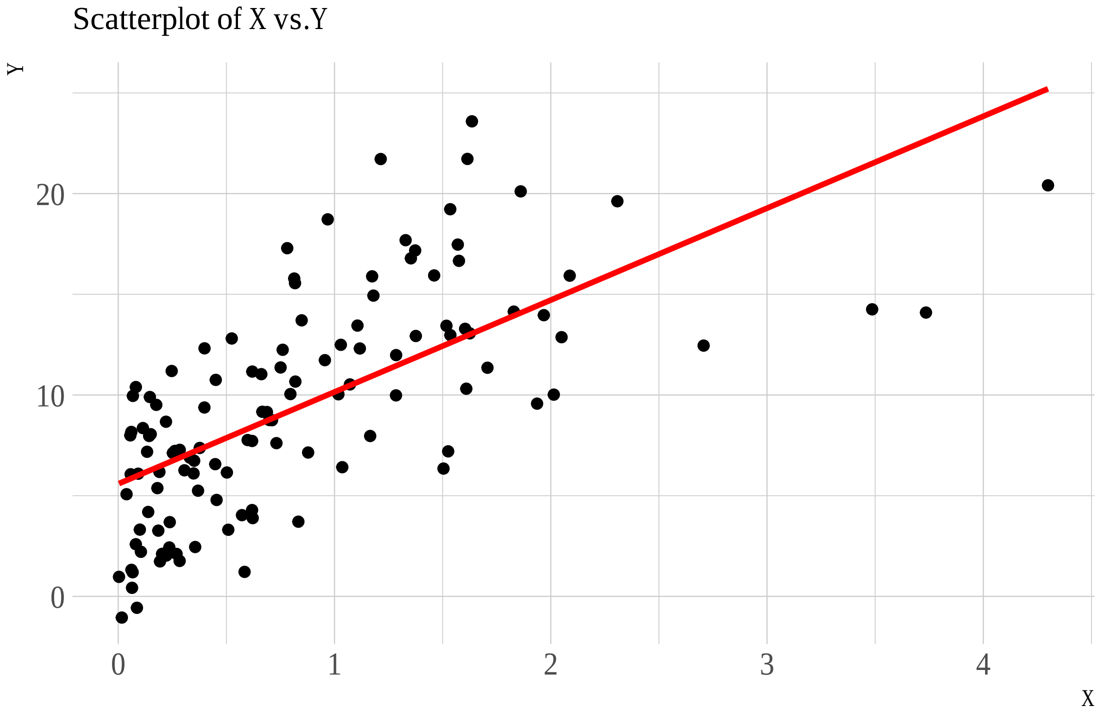
3. R and RStudio

```r
# Read the data frame from the file HW3.RData
load("HW3.RData")
mydata3<-data
```

a. Prepare a scatterplot of X vs. Y overlaid with the estimated regression line.

```r
p2<-ggplot(mydata3,aes(x=x,y=y))+geom_point(color="black")+
  xlab("X")+ylab("Y")+ggtitle("Scatterplot of X vs.Y")+
  theme_ipsum(plot_title_size = 12)+
  geom_smooth(method = lm,color="red",se=FALSE)
p2
```

```
## `geom_smooth()` using formula 'y ~ x'
```



b. Calculate the correlation coefficient between X and Y and comment on the strength of the linear association, using the standard cutoff point of ±0.7.

```r
cor(mydata3$x,mydata3$y)
```

```
## [1] 0.6642197
```

ANSWER: The correlation coefficient between X and Y is 0.664. Since it is smaller than standard cutoff point of 0.7, we can conclude that the variables are not strongly positively linearly related.

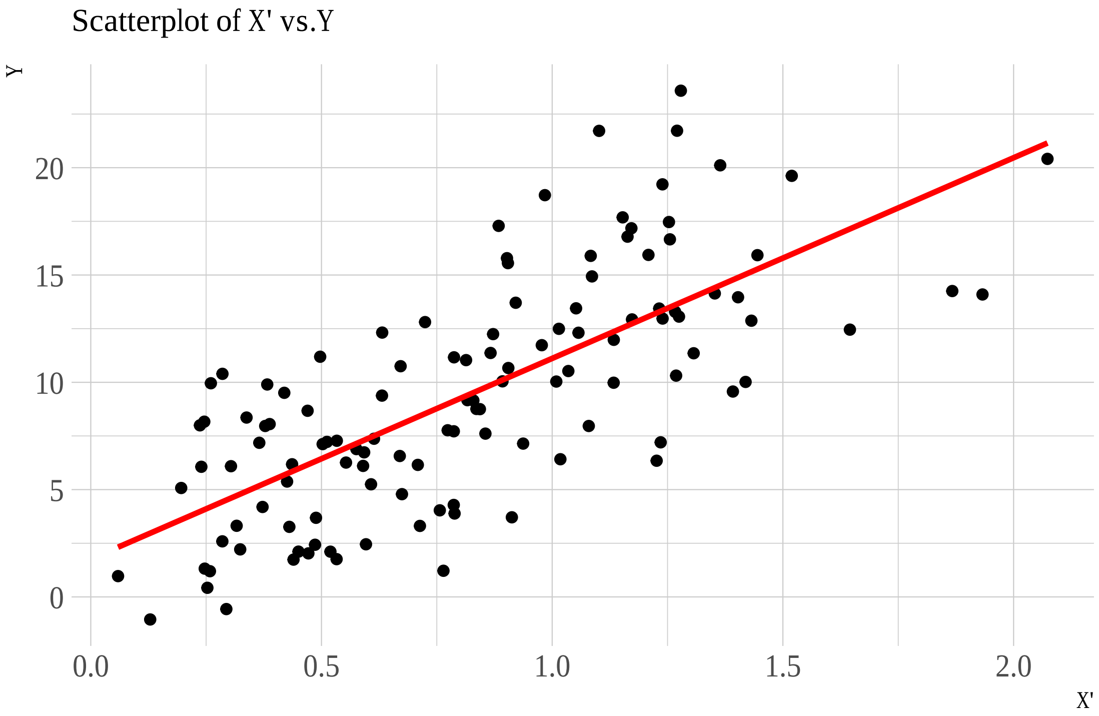
c. Create a new variable $X' = \sqrt{X}$.

```r
x2<-sqrt(mydata3$x)
mydata3<-cbind(mydata3,x2)
```

d. Prepare a scatterplot of X' vs. Y overlaid with the estimated regression line.

```
p3<-ggplot(mydata3,aes(x=x2,y=y))+geom_point(color="black")+
  xlab("X'")+ylab("Y")+ggtitle("Scatterplot of X' vs.Y")+
  theme_ipsum(plot_title_size = 12)+
  geom_smooth(method = lm,color="red",se=FALSE)
p3
```

```
## `geom_smooth()` using formula 'y ~ x'
```



e. Caculate the correlation coefficient between X' and Y and comment on the strength of the linear association.

```
cor(mydata3$x2,mydata3$y)
```

```
## [1] 0.7171237
```

ANSWER: The correlation coefficient between X' and Y is 0.717. Since it is larger than standard cutoff point of 0.7, we conclude that the variables are strongly positively linearly related.

f. Obtain the estimated linear regression function for the transformed data.

```
reg2<-lm(mydata3$y~mydata3$x2)
summary(reg2)
```

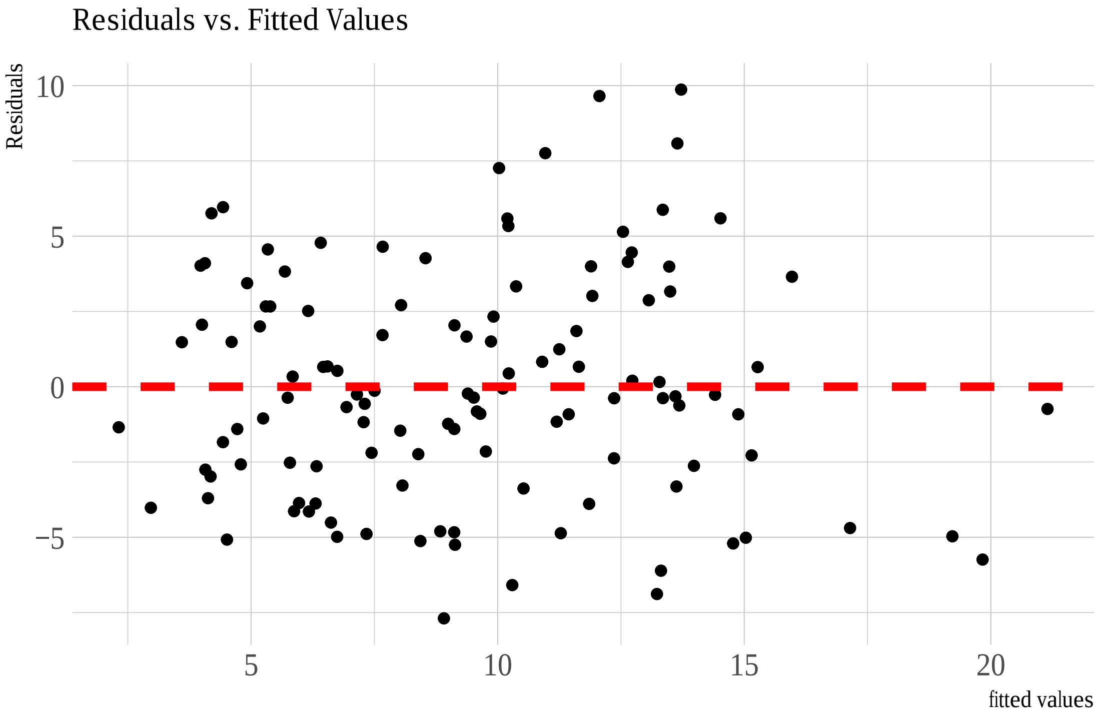
```
##
## Call:
## lm(formula = mydata3$y ~ mydata3$x2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.6928 -2.6700 -0.3411  2.6769  9.8673
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.7636     0.7692   2.293   0.0236 *
## mydata3$x2    9.3497     0.8365  11.177   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.785 on 118 degrees of freedom
## Multiple R-squared:  0.5143, Adjusted R-squared:  0.5102
## F-statistic: 124.9 on 1 and 118 DF,  p-value: < 2.2e-16
```

ANSWER: $\widehat{Y} = 1.7636 + 9.3497 * X$

g. Plot the residuals vs. fitted values.

```
y.hat2<-predict(reg2)
e2<-resid(reg2)
mydata3<-cbind(mydata3,y.hat2,e2)
# Draw the graph
p4<-ggplot(mydata3,aes(x=y.hat2,y=e2))+geom_point(color="black")+
  xlab("fitted values")+ylab("Residuals")+ggtitle("Residuals vs. Fitted Values")+
  theme_ipsum(plot_title_size = 12)+
  geom_hline(yintercept = 0,linetype="dashed",color="red",size=1.5)
p4
```



h. What does the plot from part g) show?

ANSWER: The plot from part g) shows that after the transformation of data, we observe a constant variance for the residuals.