Hw03ST430Yu

Haozhe (Jerry) Yu

2023-10-26

Question 1

```
educ <- as_tibble(read.table("https://users.stat.ufl.edu/~rrandles/sta4210/Rclassnotes/data/textdatasets
            #sep = "",
            strip.white=TRUE,
            col.name = c("Crime.Rate","High.School.Diploma")
            ))
educ
```

```
## # A tibble: 84 x 2
##    Crime.Rate High.School.Diploma
##         <int>               <int>
##  1       8487                  74
##  2       8179                  82
##  3       8362                  81
##  4       8220                  81
##  5       6246                  87
##  6       9100                  66
##  7       6561                  68
##  8       5873                  81
##  9       7993                  74
## 10       7932                  82
## # i 74 more rows
```

a. Find the least squares regression equation to predict the crime rate from the percent of individuals having at least a high school education. [Paste R or SAS output and then answer your question]

```
educm <- lm(Crime.Rate~High.School.Diploma,data=educ)
```

The equation to predict crime rate (per 100,000 residents) from the percent of individuals in a country with at least a high school diploma is

Crime Rate = 2.05176×10^4 + -170.5751886High School Percent

b. Give the ANOVA Table for this regression analysis. [Paste R or SAS output]

```
educma <- anova(educm)

educma
```

```
## Analysis of Variance Table
##
## Response: Crime.Rate
##                      Df    Sum Sq  Mean Sq F value    Pr(>F)
## High.School.Diploma  1  93462942 93462942  16.834 9.571e-05 ***
## Residuals           82 455273165  5552112
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

c. Find SSE and MSE for this model.

The SSE for this model is 4.5527317×10^8 and the MSE is 5.5521118×10^6.

d. What is the estimate of sigma from this analysis?

The estimate of σ for this analysis is 2356.2919539

e. What percent of the variation in crime rates can be explained by the percent of high school graduates?

The percent of variation in crime rates explained by the percent of high school grads is 0.170324

f. What is the correlation between crime rates and percent of high school graduates?

The correlation between crime rates and percent of high school graduates is -0.4127033

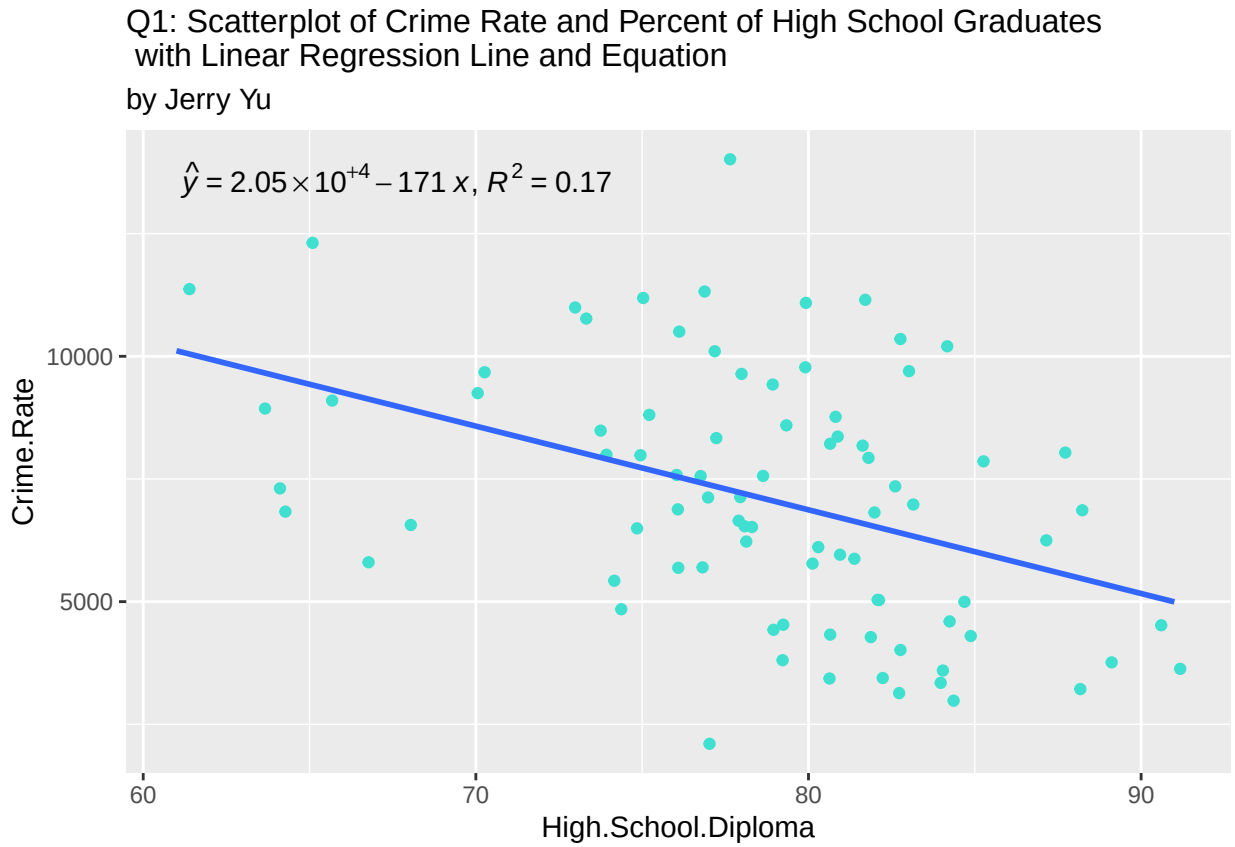
g. Based on your ANOVA table, is the linear relationship between X and Y statistically significant? Be sure to give an appropriate null and alternate hypothesis, test statistic, its associated degrees of freedom, and the p-value.

- H0: There is no linear relationship between crime rates and percent of high school graduates (β_1 =0)
- HA: There is a linear relationship between crime rates and percent of high school graduates (β_1 ne 0)
- Test Statistic (F value): 16.8337645
- Degrees of Freedom: 1 for the model (High School Diploma Percent), and 82 for the error.
- P value: 9.5713958×10^{-5}

As p < 0.05, we reject H0 at $\alpha = 0.05$ and conclude that there is statistically significant evidence for a linear relationship between crime rate and percent of high school graduates.

h. Give a scatter plot of crime rates vs. percent of high school graduates, with the regression line. Comment about linearity

```
ggplot(educ,aes(x=High.School.Diploma,y=Crime.Rate))+
  geom_jitter(color="turquoise")+
  geom_smooth(method='lm', formula= y~x,
              se=FALSE,
              show.legend=TRUE)+
    stat_poly_eq(eq.with.lhs = "italic(hat(y))~`=`~",
                 use_label(c("eq", "R2")))+
  labs(title = paste("Q1: Scatterplot of Crime Rate and Percent of High School Graduates \n with Linear
         subtitle = "by Jerry Yu")+
  theme(plot.title = element_text(size = 12))
```



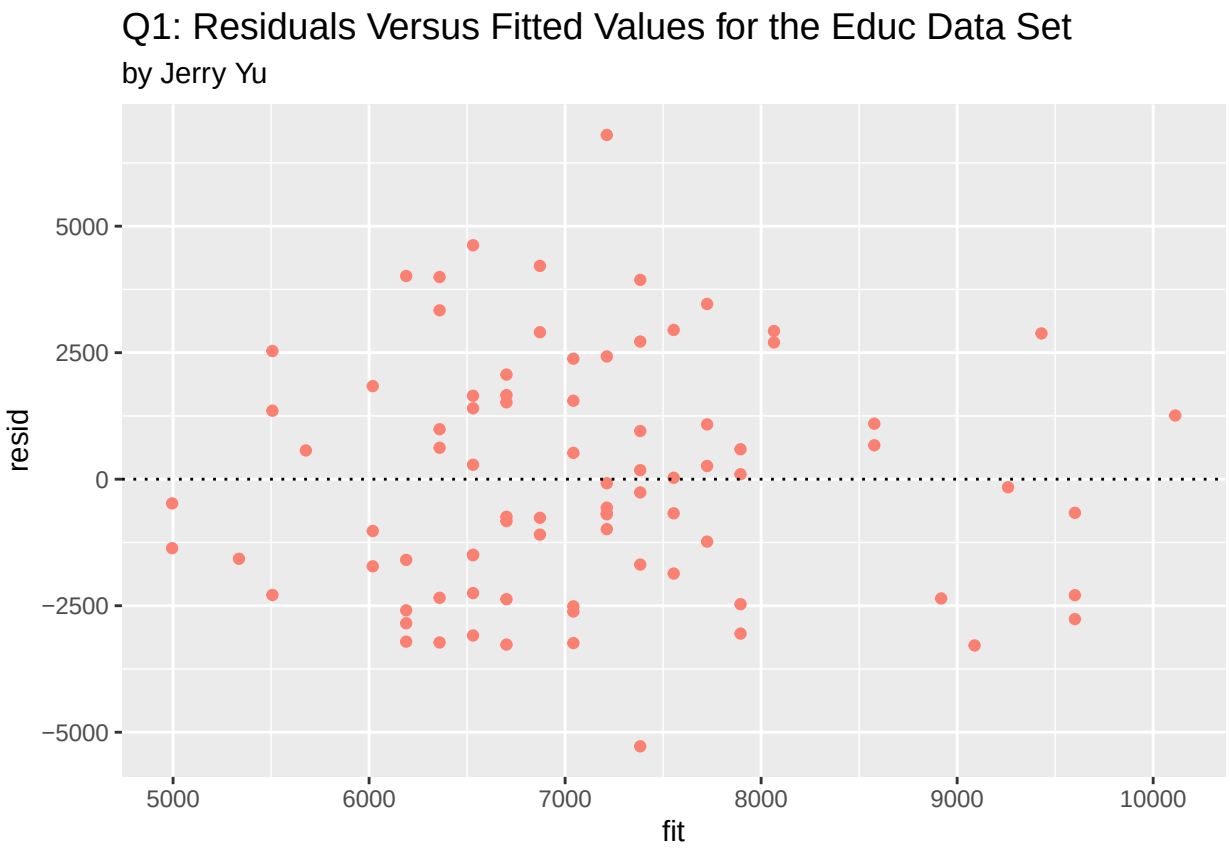
As there does not seem to be a nonlinear pattern in the scatterplot, and the regression line seems to slice across the residuals equally, leaving about 1/2 above and below. I would say that the data seems linear.

i. Give the Residual Plot (residuals vs. fitted values). Test for Non-Linear and Non-constant variance.

```
educmr <- tibble(
  "fit" = educm$fitted.values,
  "resid" = educm$residuals
```

```
)

ggplot(educmr,aes(x=fit,y=resid))+
  geom_jitter(color="salmon")+
  geom_hline(yintercept = 0, linetype="dotted")+
  labs(title = paste("Q1: Residuals Versus Fitted Values for the Educ Data Set"),
         subtitle = "by Jerry Yu")+
  theme(plot.title = element_text(size = 14))
```



As there do not seem to be patterns in the distribution of the residuals, nor any fan and funnel shapes, I conclude that the variance is likely linear and constant.

j. Conduct Breusch-Pagan Test for the constancy of the error variance. Be sure to give an appropriate null and alternate hypothesis, test statistic, its associated degrees of freedom, and the p-value.

```
educmbp <- bptest(educm,studentize = FALSE)
educmbp
```

```
##
##  Breusch-Pagan test
##
## data:  educm
```

```
## BP = 0.005045, df = 1, p-value = 0.9434
```

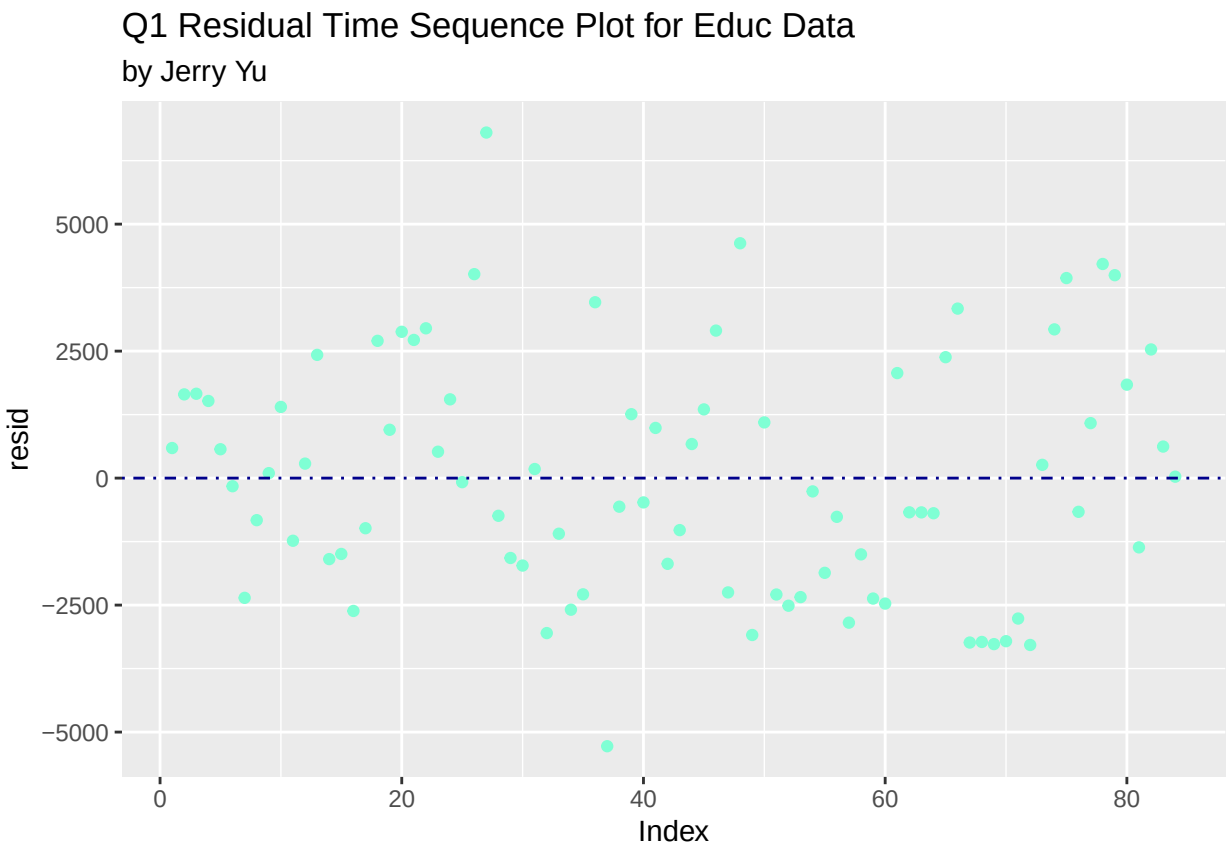
```
ncvTest(educm)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.005045022, Df = 1, p = 0.94338
```

- H0: Equal Variance Among Errors
- HA: Unequal Variance Among Errors
- Degree of Freedom: 1
- P Value: 0.9433752

k. Index Plot to test for Independence of errors.

```
ggplot(educmr, aes(x = 1:length(resid), y = resid)) +
  geom_point(color = "aquamarine") +
  labs(x = "Index",
       title = "Q1 Residual Time Sequence Plot for Educ Data",
       subtitle = "by Jerry Yu") +
  geom_hline(yintercept = 0,
             color = "darkblue",
             linetype = "dotdash")
```



l. Conduct Durbin-Watson Test. Be sure to give an appropriate null and alternate hypothesis, test statistic and the p-value.

```
dwtest(educm)
```

```
##
##  Durbin-Watson test
##
## data:  educm
## DW = 1.4951, p-value = 0.008696
## alternative hypothesis: true autocorrelation is greater than 0
```

```
educmw <-durbinWatsonTest(educm)
educmw
```

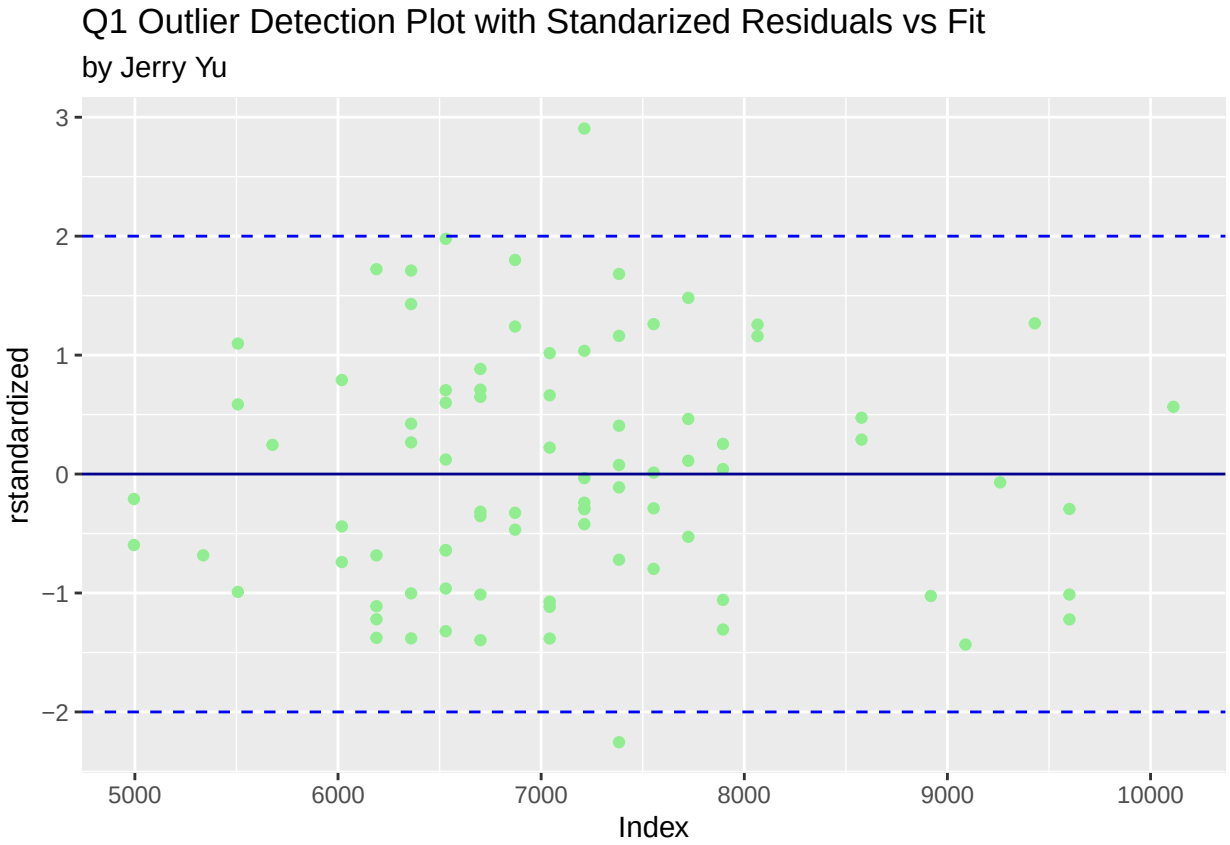
```
##  lag Autocorrelation D-W Statistic p-value
##    1         0.25204      1.495148   0.016
##  Alternative hypothesis: rho != 0
```

- H0: Errors are uncorrelated over time
- HA: Errors are correlated (either positive or negative). I used the `car` test where the alternative hypothesis is 2 sided
- Test Statistic: 1.4951485
- p value: 0.016

m. Outlier deduction test [Plot standardized Residuals versus fitted values]

```
educmrs <- add_column(educmr, "rstandardized"=rstandard(educm))

ggplot(educmrs, aes(x = fit, y = rstandardized)) +
  geom_point(color = "lightgreen") +
  labs(x = "Index",
       title = "Q1 Outlier Detection Plot with Standarized Residuals vs Fit",
       subtitle = "by Jerry Yu") +
  geom_hline(yintercept = 0,
             color = "darkblue",
             linetype = "solid") +
    geom_hline(yintercept = -2,
             color = "blue",
             linetype = "dashed")+
  geom_hline(yintercept = 2,
             color = "blue",
             linetype = "dashed")
```




```
educmro <- filter(educmrs,abs(rstandardized) >2)
educmro
```

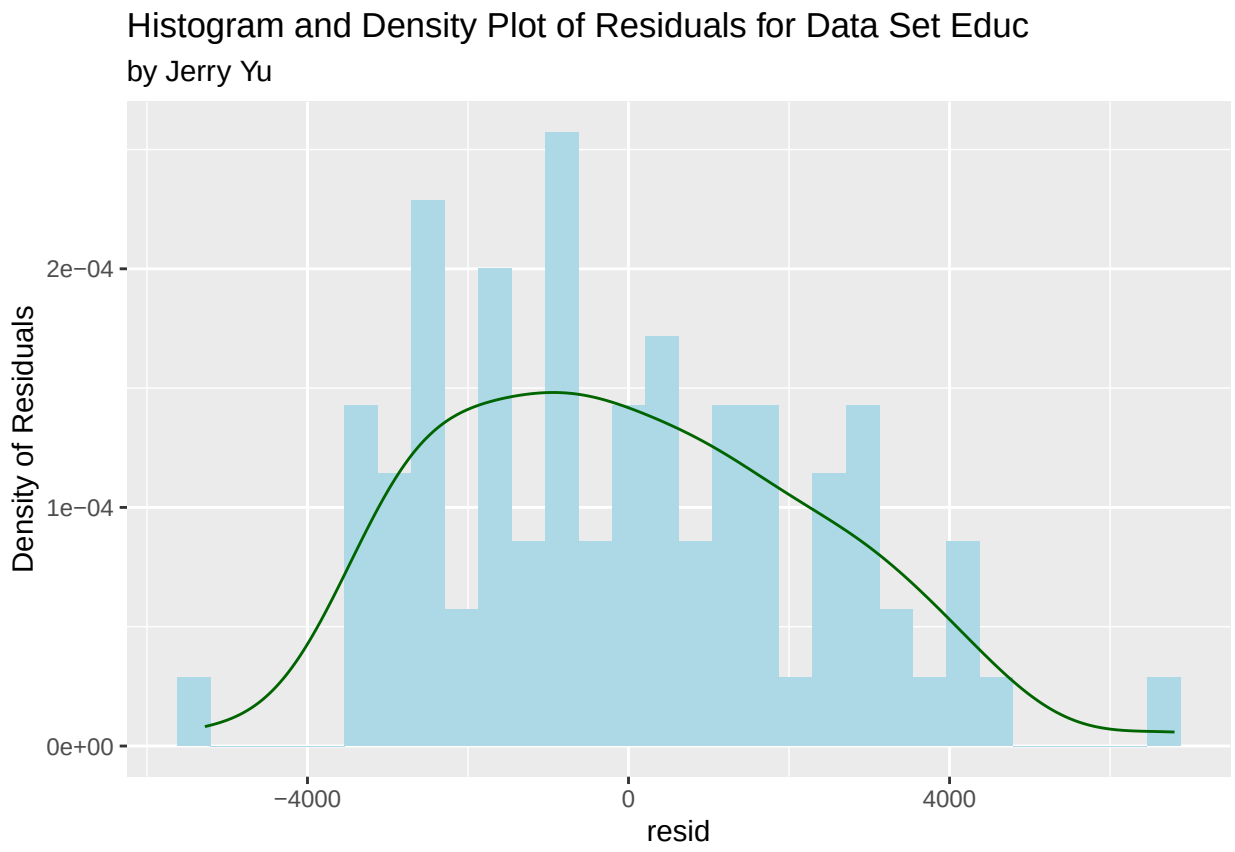
```
## # A tibble: 2 x 3
##     fit  resid rstandardized
##   <dbl>  <dbl>         <dbl>
## 1 7213.  6803.          2.90
## 2 7383. -5278.         -2.25
```

We have 2 outliers, one where the fit = 7212.7352313 and 7383.31042

n. Give a Histogram of the residuals and the density curve. Comment about the distribution of residuals.

```
ggplot(data = educmrs, aes(x = resid, y = after_stat(density))) +
  geom_histogram(fill = "lightblue") +
  geom_density(color = "darkgreen") +
  labs(
    title = paste("Histogram and Density Plot of Residuals for Data Set Educ"),
    subtitle = "by Jerry Yu"
  ) +
  ylab("Density of Residuals")
```

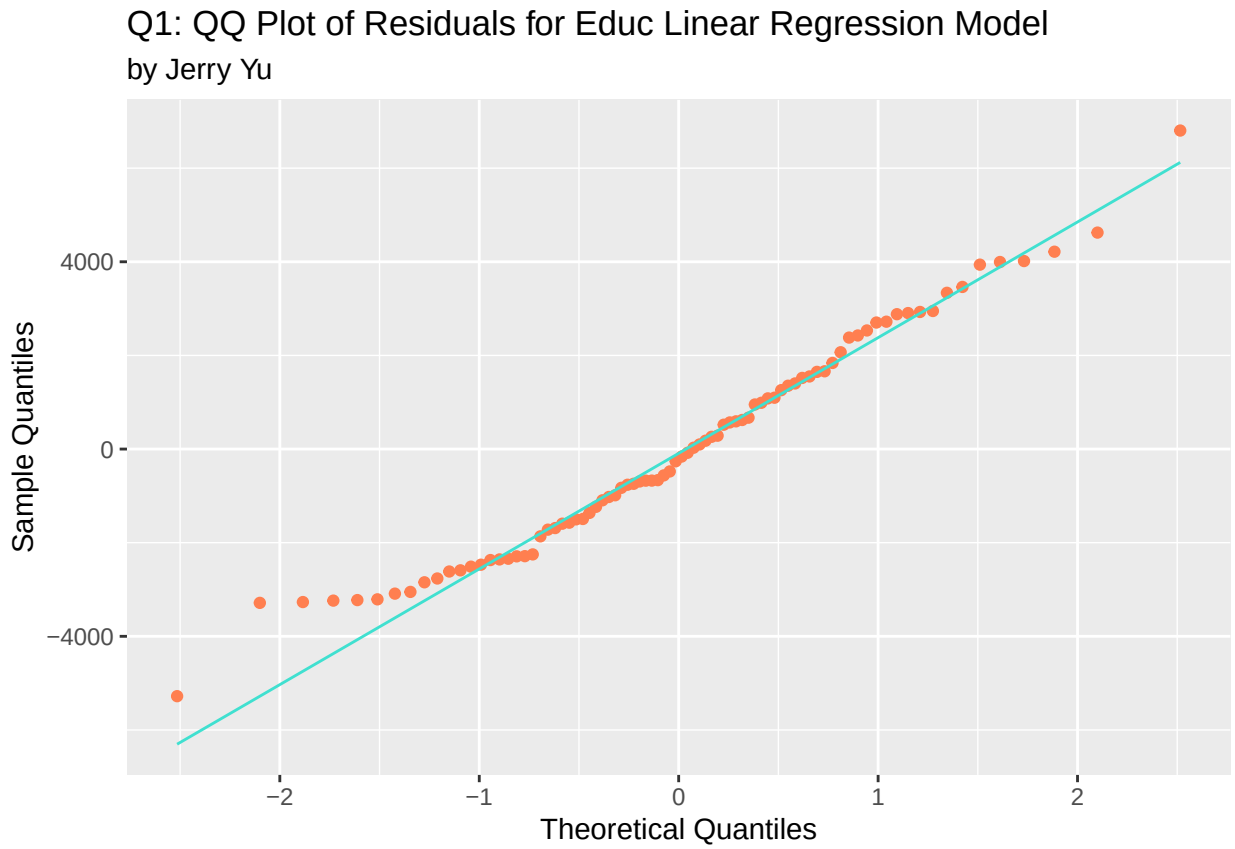
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



There seems to be a slight right skew in the data, The 2 outliers detected in part m are clearly visible.

o. Give a QQ-plot of the residuals to test for normality of error terms. Comment about the distribution of residuals.

```
ggplot(data = educmrs, aes(sample = resid))+
  geom_qq( color="coral")+
  geom_qq_line( color="turquoise")+
  labs(
    title = paste("Q1: QQ Plot of Residuals for Educ Linear Regression Model"),
    subtitle = "by Jerry Yu"
  ) +
  xlab("Theoretical Quantiles")+
  ylab("Sample Quantiles")
```



The data visually does not look normal, as the extreme Residuals both look flatter than the Theoretical Residuals (the line).

p. Conduct a Shapiro-Wilk Test on the residuals. Be sure to give an appropriate null and alternate hypothesis, test statistic and the p-value. Give the p-value for this test and explain what this means in terms of our model assumptions.

```
shap1 <- shapiro.test(educmrs$resid)
shap1
```

```
##
##  Shapiro-Wilk normality test
##
## data:  educmrs$resid
## W = 0.97763, p-value = 0.1515
```

- H0: The random error is normally distributed
- Ha: The random error is not normally distributed
- Test Statistic: 0.9776328
- p value: 0.1514916

As p > 0.05, we fail to reject H0 at $\alpha = 0.05$ and conclude that there is no statistically significant evidence that the random error is not normally distributed.

Question 2

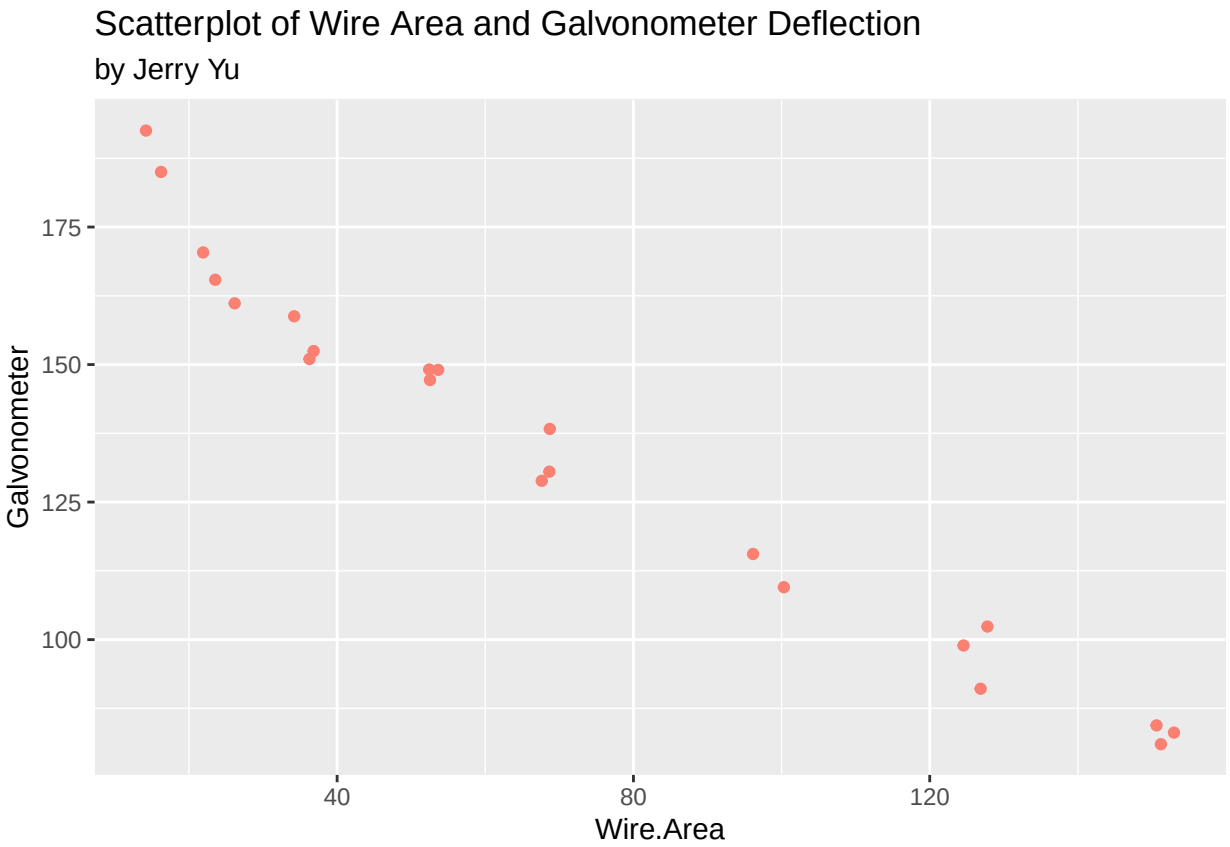
2.Download the "Explosives dataset" from Moodle. Fit a simple linear regression, relating the deflection of galvonometer (Y) to the area of the wires on the coupling (X). Complete the following parts.

```
xplode <- as_tibble(read.table("Datasets/explosives.txt",
            strip.white=TRUE,
            col.name = c("Coupling.Number","Wire.Area","Galvonometer")
            ))
xplode
```

```
## # A tibble: 22 x 3
##    Coupling.Number Wire.Area Galvonometer
##              <int>     <int>        <dbl>
##  1               1       152         85
##  2               1       152         81.5
##  3               1       152         83.5
##  4               2       125        102
##  5               2       125         90.5
##  6               2       125         98.5
##  7               3        99        109
##  8               3        99        116.
##  9               4        66        131
## 10               4        66        128.
## # i 12 more rows
```

a. Give a scatter plot

```
ggplot(xplode,aes(x=Wire.Area,y=Galvonometer))+
  geom_jitter(color="salmon")+
  labs(title = paste("Scatterplot of Wire Area and Galvonometer Deflection"),
         subtitle = "by Jerry Yu")
```



b. Find the least squares regression.

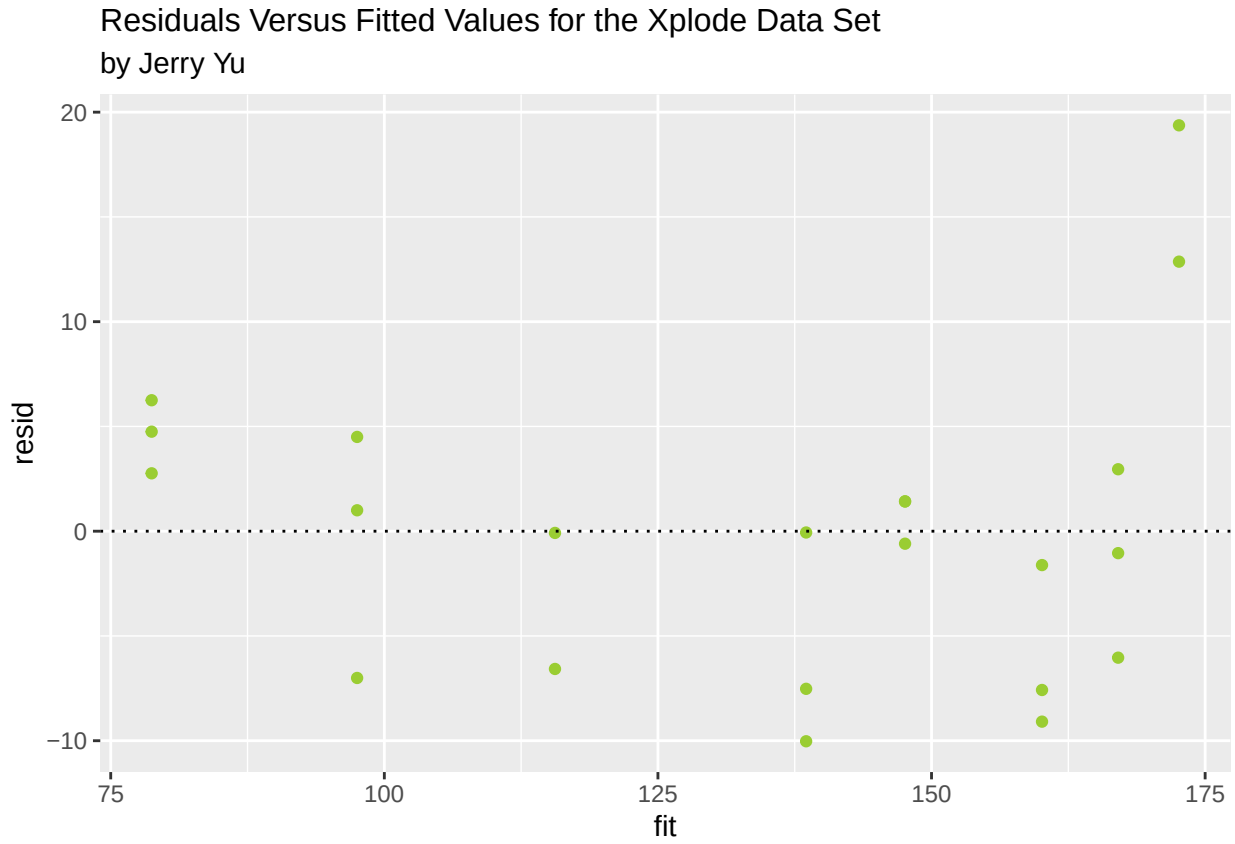
```
xplodem <- lm(Galvonometer~Wire.Area,xplode)
xplodem
```

```
##
## Call:
## lm(formula = Galvonometer ~ Wire.Area, data = xplode)
##
## Coefficients:
## (Intercept)    Wire.Area
##    184.4357      -0.6954
```

c. Give the Residual Plot (residuals vs. fitted values). Test for Non-Linear and Non-constant variance.

```
xplodemr <- tibble(
  "fit" = xplodem$fitted.values,
  "resid" = xplodem$residuals
)
```

```
ggplot(xplodemr,aes(x=fit,y=resid))+
  geom_jitter(color="yellowgreen")+
  geom_hline(yintercept = 0, linetype="dotted")+
  labs(title = paste("Residuals Versus Fitted Values for the Xplode Data Set"),
         subtitle = "by Jerry Yu")+
  theme(plot.title = element_text(size = 12))
```



There does seem to be a pattern the the distribution of residuals and fitted values, with mostly positive residuals between 75 and 100 and 160-175, and mostly negative between 100 and 160. This might indicate that the variance is not linear. However, the pattern of the variances does not assume the shape of a funnel, so there is no evidence that the variance is non constant. However, as there are relatively few data points, we cannot conclude anything from the residual plot, and will rely on the Breusch-Pagan Test in part d.

d. Conduct Breusch-Pagan Test for the constancy of the error variance.

```
xplodembp <- bptest(xplodem,studentize = FALSE)
xplodembp
```

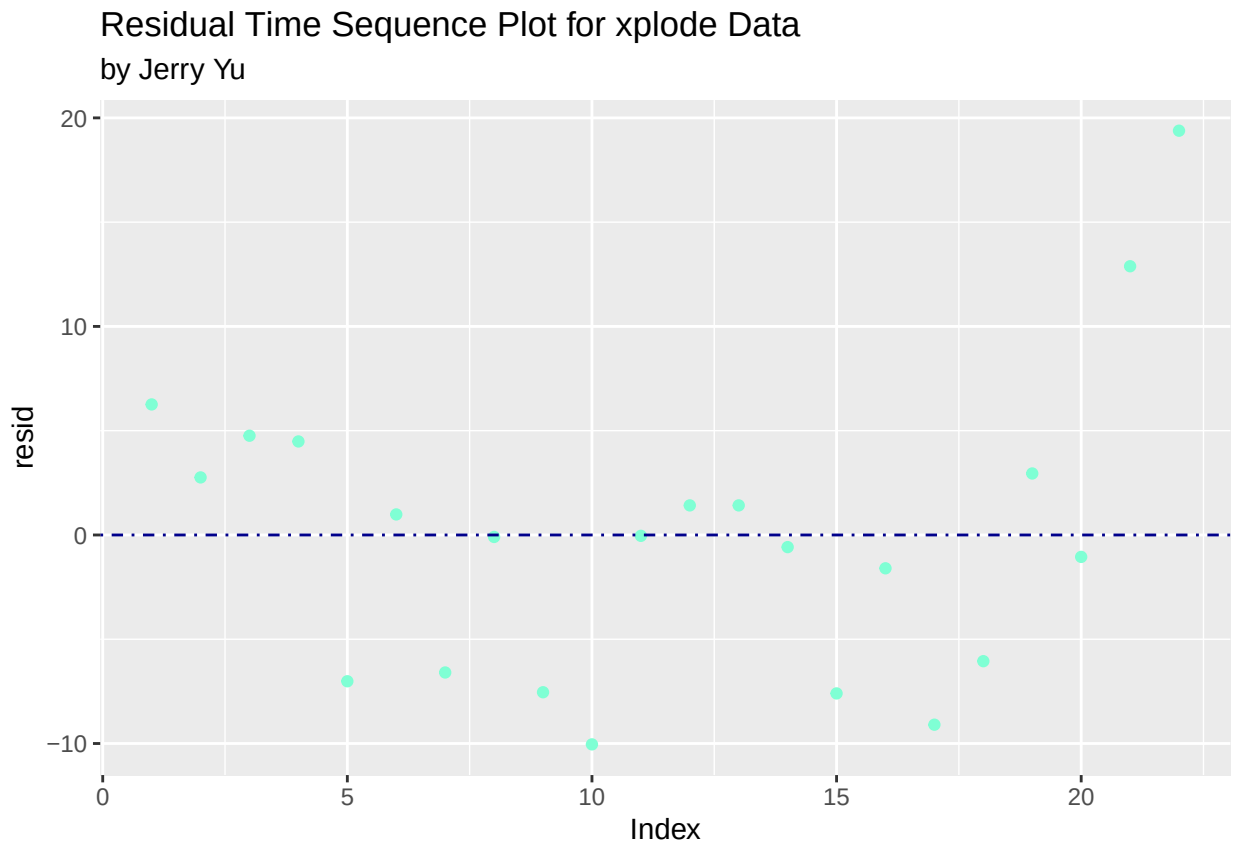
```
##
##  Breusch-Pagan test
##
## data:  xplodem
## BP = 3.6375, df = 1, p-value = 0.05649
```

```
ncvTest(xplodem)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 3.637537, Df = 1, p = 0.05649
```

e. Index Plot to test for Independence of errors.

```
ggplot(xplodemr, aes(x = 1:length(resid), y = resid)) +
  geom_point(color = "aquamarine") +
  labs(x = "Index",
       title = "Residual Time Sequence Plot for xplode Data",
       subtitle = "by Jerry Yu") +
  geom_hline(yintercept = 0,
             color = "darkblue",
             linetype = "dotdash")
```



f. Conduct Durbin-Watson Test.

```
dwtest(xplodem)
```

```
##
##  Durbin-Watson test
##
## data:  xplodem
## DW = 0.89573, p-value = 0.0008088
## alternative hypothesis: true autocorrelation is greater than 0
```

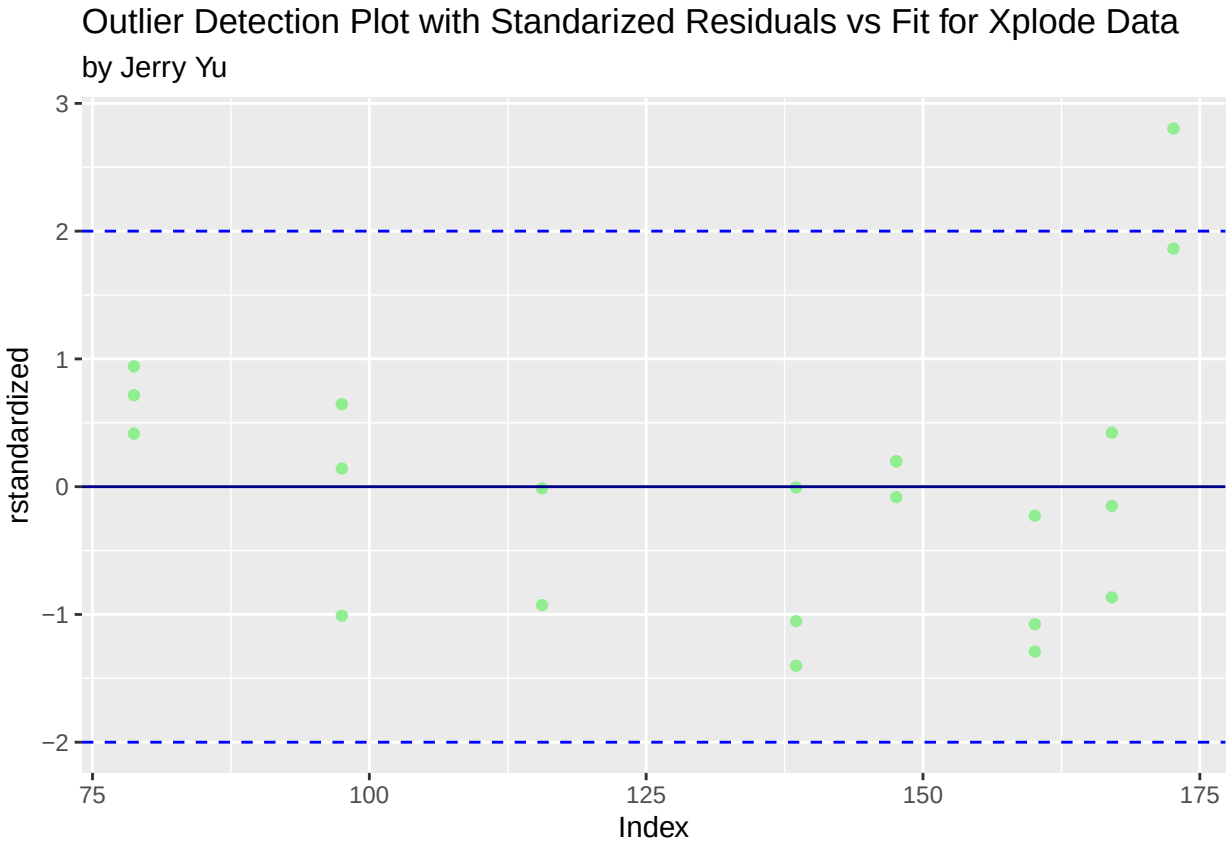
```
xplodemw <-durbinWatsonTest(xplodem)
xplodemw
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.3594126     0.8957278   0.002
##  Alternative hypothesis: rho != 0
```

g. outlier deduction test. [Plot standardized Residuals versus fitted values]

```
xplodemrs <- add_column(xplodemr, "rstandardized"=rstandard(xplodem))

ggplot(xplodemrs, aes(x = fit, y = rstandardized)) +
  geom_point(color = "lightgreen") +
  labs(x = "Index",
       title = "Outlier Detection Plot with Standarized Residuals vs Fit for Xplode Data",
       subtitle = "by Jerry Yu") +
  geom_hline(yintercept = 0,
             color = "darkblue",
             linetype = "solid") +
    geom_hline(yintercept = -2,
             color = "blue",
             linetype = "dashed")+
  geom_hline(yintercept = 2,
             color = "blue",
             linetype = "dashed")
```



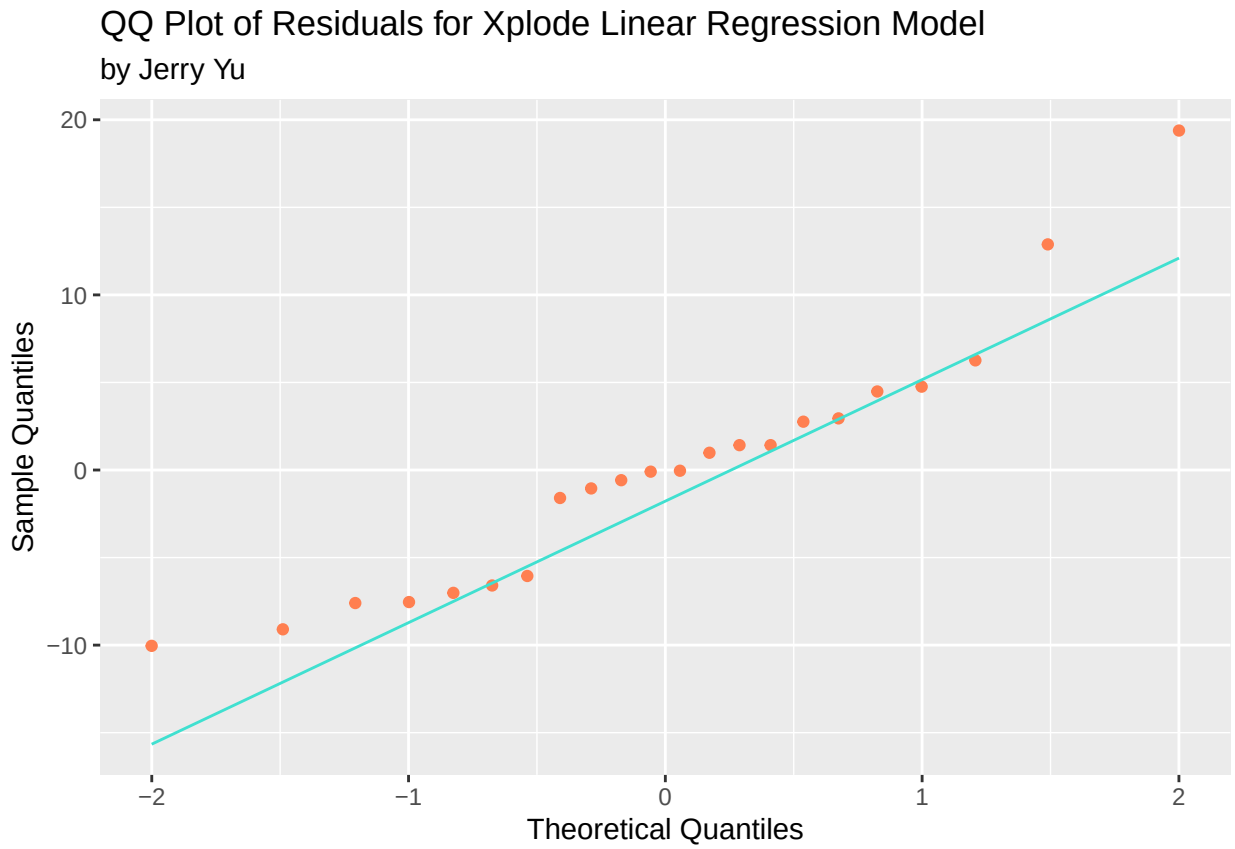

```
xplodemro <- filter(xplodemrs,abs(rstandardized) >2)
xplodemro
```

```
## # A tibble: 1 x 3
##     fit resid rstandardized
##   <dbl> <dbl>          <dbl>
## 1  173.  19.4           2.80
```

We have 1 outlier where the fit = `r xplodemro[[1,1]]`.

i. Give a QQ-plot of the residuals. Normality of error terms.

```
ggplot(data = xplodemrs, aes(sample = resid))+
  geom_qq( color="coral")+
  geom_qq_line( color="turquoise")+
  labs(
    title = paste("QQ Plot of Residuals for Xplode Linear Regression Model"),
    subtitle = "by Jerry Yu"
  ) +
  xlab("Theoretical Quantiles")+
  ylab("Sample Quantiles")
```



j. Conduct a Shapiro-Wilk Test on the residuals. Give the p-value for this test and explain what this means in terms of our model assumptions.

```
xplodeshap <- shapiro.test(xplodemrs$resid)
xplodeshap
```

```
##
##  Shapiro-Wilk normality test
##
## data:  xplodemrs$resid
## W = 0.92354, p-value = 0.09004
```

- H0: The random error is normally distributed
- Ha: The random error is not normally distributed
- p value: 0.0900395.

As p $>$ 0.05, we fail to reject H0 at $\alpha = 0.05$ and conclude that there is no statistically significant evidence that the random error is not normally distributed.

k. Give the ANOVA Table for this regression analysis. Based on your ANOVA table, is the linear relationship between X and Y statistically significant? Be sure to give an appropriate test statistic, its associated degrees of freedom, and the p-value.

```
xplodema <- anova(xplodem)

xplodema
```

```
## Analysis of Variance Table
##
## Response: Galvonometer
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## Wire.Area  1 22749.5 22749.5   422.6 6.386e-15 ***
## Residuals 20  1076.6    53.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- H0: There is no linear relationship between Wire Area (1/100,000 in) and Deflection of Galvonometer in mm (β_1 =0)
- H0: There is a linear relationship between Wire Area (1/100,000 in) and Deflection of Galvonometer in mm (β_1 ne 0)
- Test Statistic (F value): 422.6036859
- Degrees of Freedom: 1 for the model (Area of Wires (1/100,000 in)), and 20 for the error.
- P value: 6.386449×10^{-15}

As p < 0.05, we reject H0 at $\alpha = 0.05$ and conclude that there is statistically significant evidence for a linear relationship between Wire Area (1/100,000 in) and Deflection of Galvonometer in mm.

```
surg.data <- read.table(
  "Datasets/Surgical Unit.txt",
  header = FALSE,
  col.names = c(
    "clot",
    "PI",
    "enzy",
    "liver",
    "age",
    "gender",
    "mod_use",
    "heavy_use",
    "sur_time",
    "ln_sur_time"
  )
)

attach(surg.data)
gender = factor(gender)
```

```r
surg.datam <- lm(sur_time~clot,data=surg.data)

surg.datam
```

```
##
## Call:
## lm(formula = sur_time ~ clot, data = surg.data)
##
## Coefficients:
## (Intercept)         clot
##      205.26        85.91
```

```r
surg.dataa <- anova(surg.datam)

surg.dataa
```

```
## Analysis of Variance Table
##
## Response: sur_time
##           Df  Sum Sq Mean Sq F value  Pr(>F)
## clot       1 1005152 1005152  7.0974 0.01025 *
## Residuals 52 7364369  141622
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Question 3

a. Based on your ANOVA table, is the linear relationship between X and Y statistically significant? Be sure to give an appropriate null and alternate hypothesis, test statistic, its associated degrees of freedom, and the p-value.

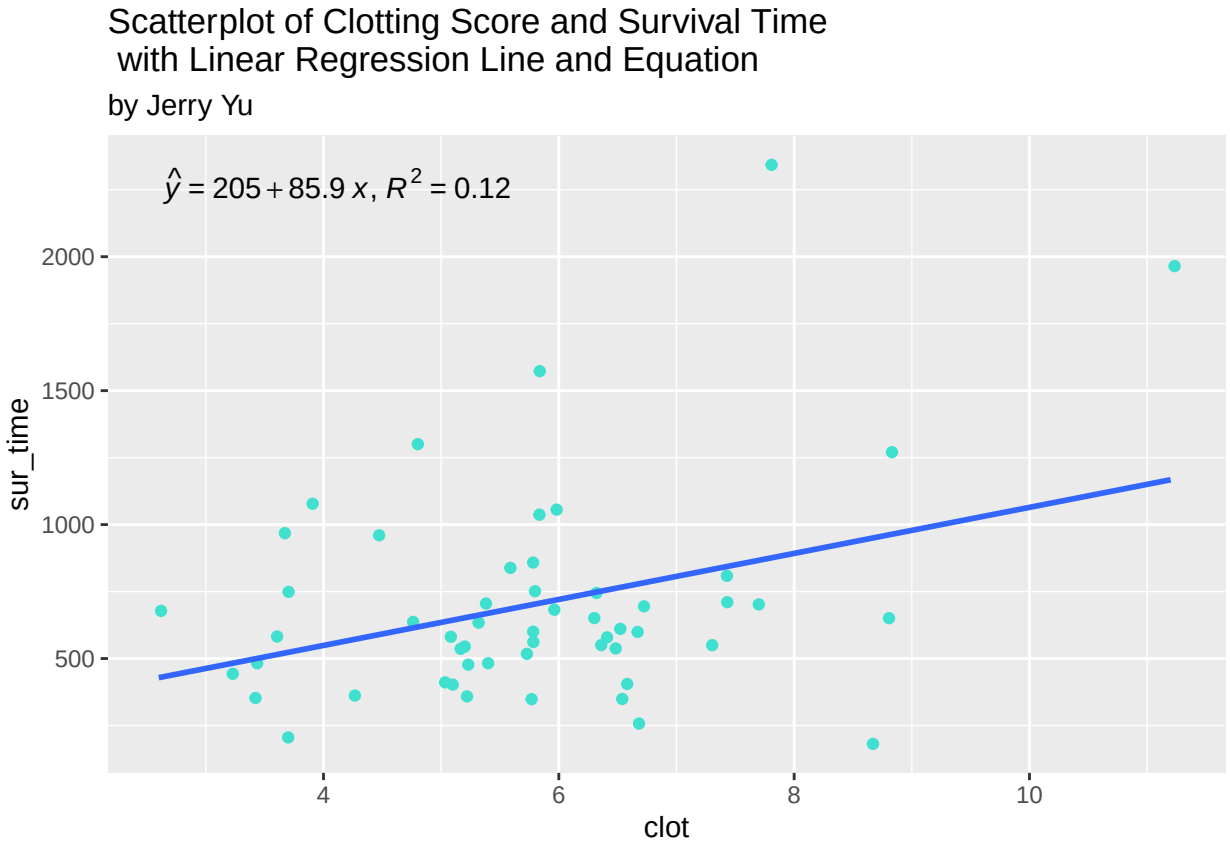
- H0: There is no linear relationship between survival time and blood clotting score (β_1 =0)
- HA: There is a linear relationship between survival time and blood clotting score (β_1 ne 0)
- Test Statistic (F value): 7.097402
- Degrees of Freedom: 1 for the model (Blood-clotting score), and 52 for the error.
- P value: 0.0102549

As p $<$ 0.05, we reject H0 at $\alpha = 0.05$ and conclude that there is statistically significant evidence for a linear relationship between blood clotting score and survival time.

b. Give a scatter plot of clot vs. sur_time, with the

regression line. Comment about linearity

```
ggplot(surg.data,aes(x=clot,y=sur_time))+
  geom_jitter(color="turquoise")+
  geom_smooth(method='lm', formula= y~x,
              se=FALSE,
              show.legend=TRUE)+
    stat_poly_eq(eq.with.lhs = "italic(hat(y))~`=`~",
                 use_label(c("eq", "R2")))+
  labs(title = paste("Scatterplot of Clotting Score and Survival Time \n with Linear Regression Line an
         subtitle = "by Jerry Yu")
```



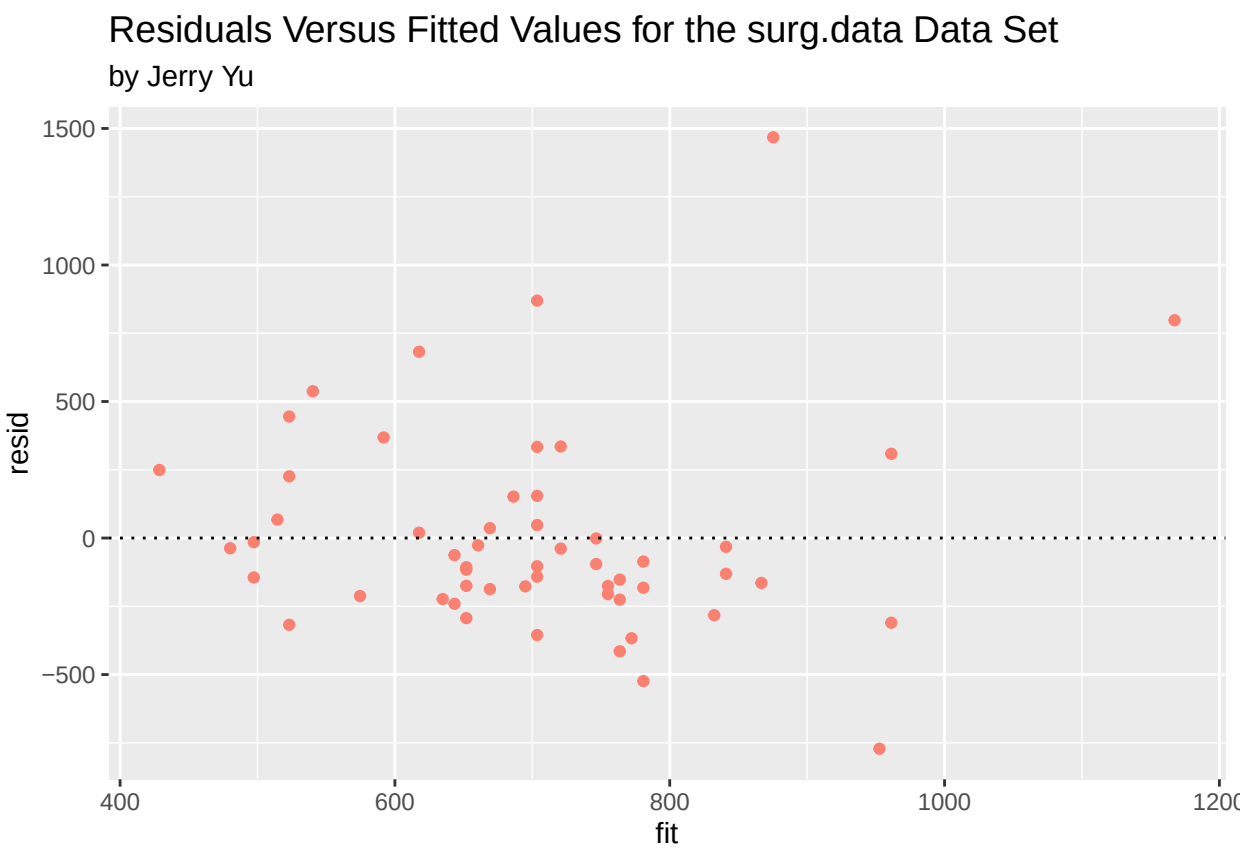
I would say that the distributions of residuals above and below the regression line look somewhat even. However there is a noticeable funnel pattern, but that is indicative of heteroscedasticity, not non-linearity.

c. Give the Residual Plot (residuals vs. fitted values). Test for Non-Linear and Non-constant variance.

```
surg.datamr <- tibble(
  "fit" = surg.datam$fitted.values,
  "resid" = surg.datam$residuals
)

ggplot(surg.datamr,aes(x=fit,y=resid))+
  geom_jitter(color="salmon")+
```

```
  geom_hline(yintercept = 0, linetype="dotted")+
  labs(title = paste("Residuals Versus Fitted Values for the surg.data Data Set"),
         subtitle = "by Jerry Yu")+
  theme(plot.title = element_text(size = 14))
```



d. Conduct Breusch-Pagan Test for the constancy of the error variance. Be sure to give an appropriate null and alternate hypothesis, test statistic, its associated degrees of freedom, and the p-value.

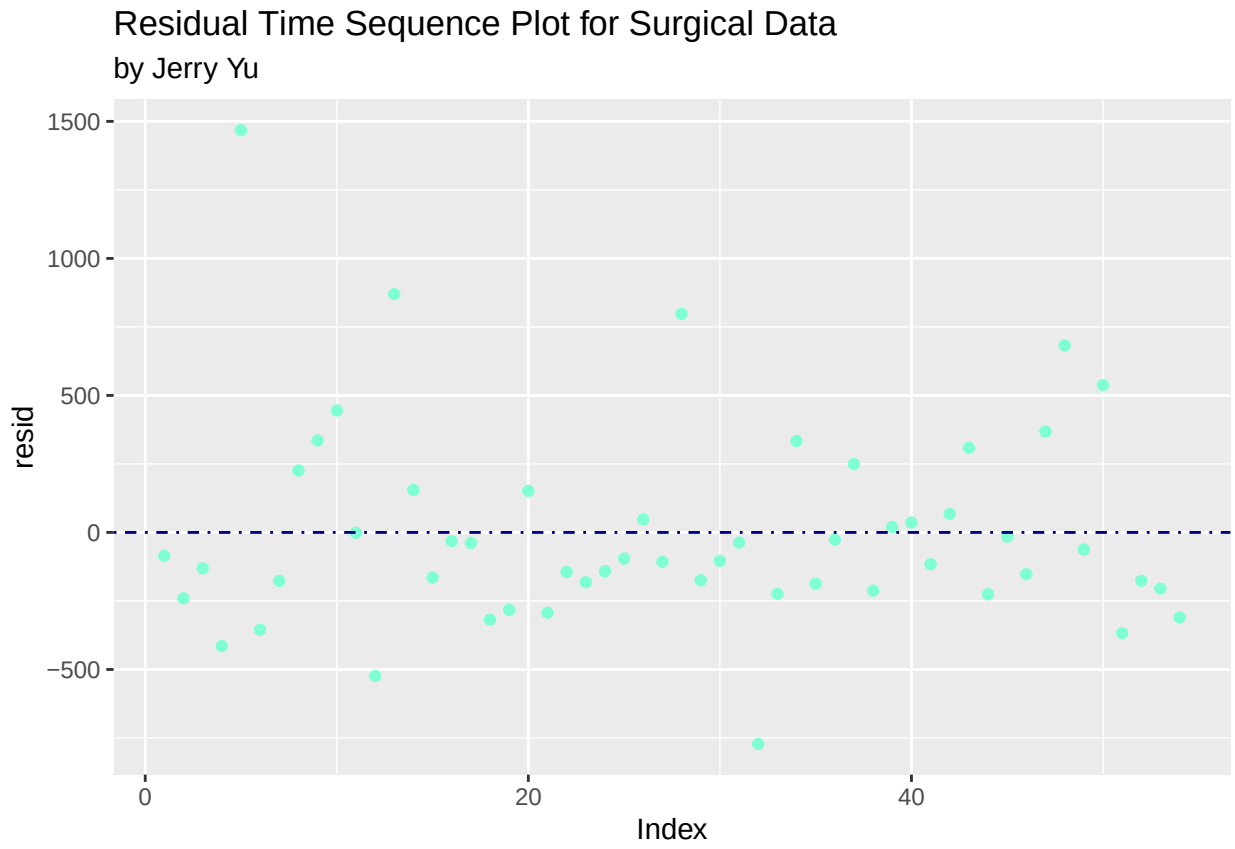
```
surg.datambp <- bptest(surg.datam,studentize = FALSE)
ncvTest(surg.datam)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 14.44279, Df = 1, p = 0.00014448
```

- H0: Equal Variance Among Errors
- HA: Unequal Variance Among Errors
- Degree of Freedom: 1
- P Value: 1.4448211×10^{-4}

e. Index Plot to test for Independence of errors.

```
ggplot(surg.datamr, aes(x = 1:length(resid), y = resid)) +
  geom_point(color = "aquamarine") +
  labs(x = "Index",
       title = "Residual Time Sequence Plot for Surgical Data",
       subtitle = "by Jerry Yu") +
  geom_hline(yintercept = 0,
             color = "darkblue",
             linetype = "dotdash")
```



f. Conduct Durbin-Watson Test. Be sure to give an appropriate null and alternate hypothesis, test statistic and the p-value.

```
dwtest(surg.datam)
```

```
##
##  Durbin-Watson test
##
## data:  surg.datam
## DW = 2.3034, p-value = 0.8724
## alternative hypothesis: true autocorrelation is greater than 0
```

```
surg.datamw <-durbinWatsonTest(surg.datam)
surg.datamw
```

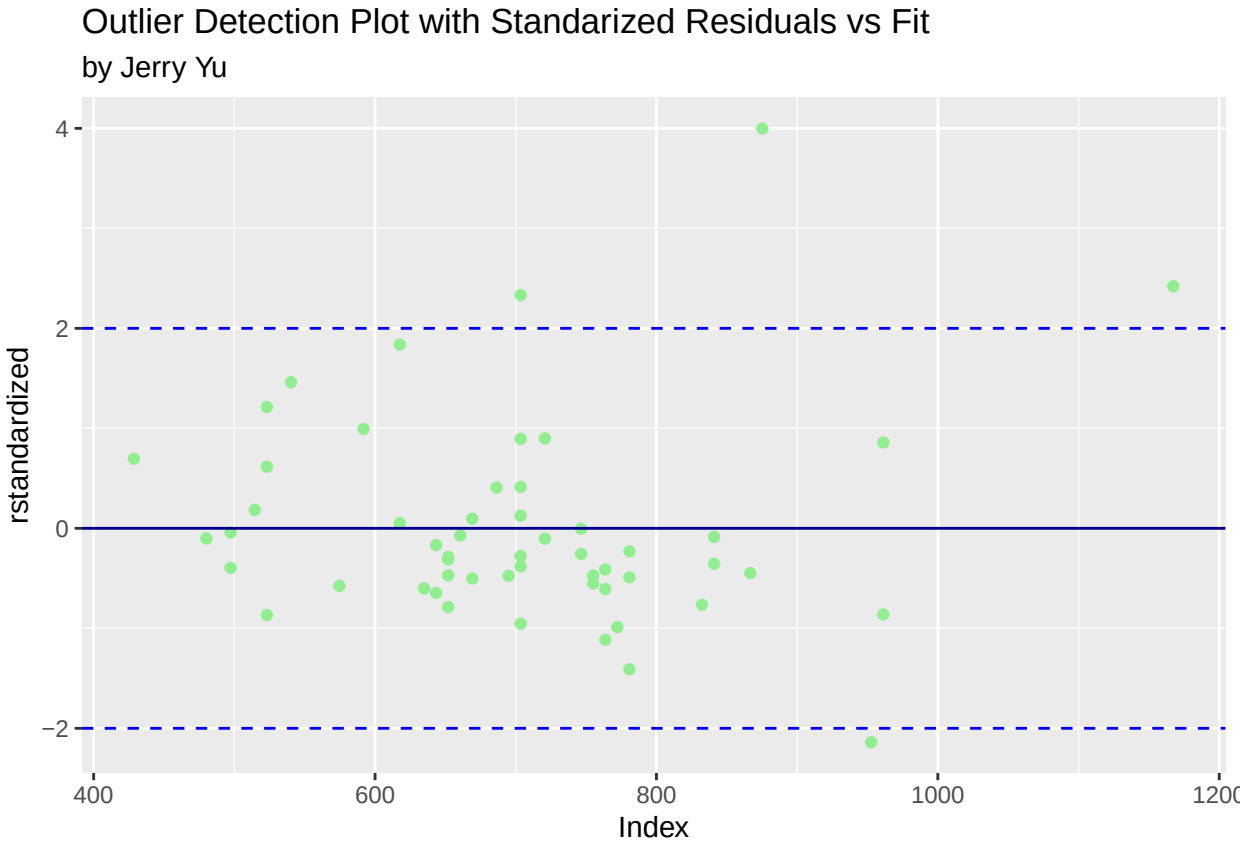
```
##  lag Autocorrelation D-W Statistic p-value
##    1      -0.1587264      2.303382    0.23
##  Alternative hypothesis: rho != 0
```

- H0: Errors are uncorrelated over time
- HA: Errors are correlated (either positive or negative). I used the `car` test where the alternative hypothesis is 2 sided
- Test Statistic: 2.3033819
- p value: 0.23

g. Outlier deduction test [Plot standardized Residuals versus fitted values]

```
surg.datamrs <- add_column(surg.datamr, "rstandardized"=rstandard(surg.datam))

ggplot(surg.datamrs, aes(x = fit, y = rstandardized)) +
  geom_point(color = "lightgreen") +
  labs(x = "Index",
       title = "Outlier Detection Plot with Standarized Residuals vs Fit",
       subtitle = "by Jerry Yu") +
  geom_hline(yintercept = 0,
             color = "darkblue",
             linetype = "solid") +
    geom_hline(yintercept = -2,
             color = "blue",
             linetype = "dashed")+
  geom_hline(yintercept = 2,
             color = "blue",
             linetype = "dashed")
```




```
surg.datamro <- filter(surg.datamrs,abs(rstandardized) >2)
surg.datamro
```

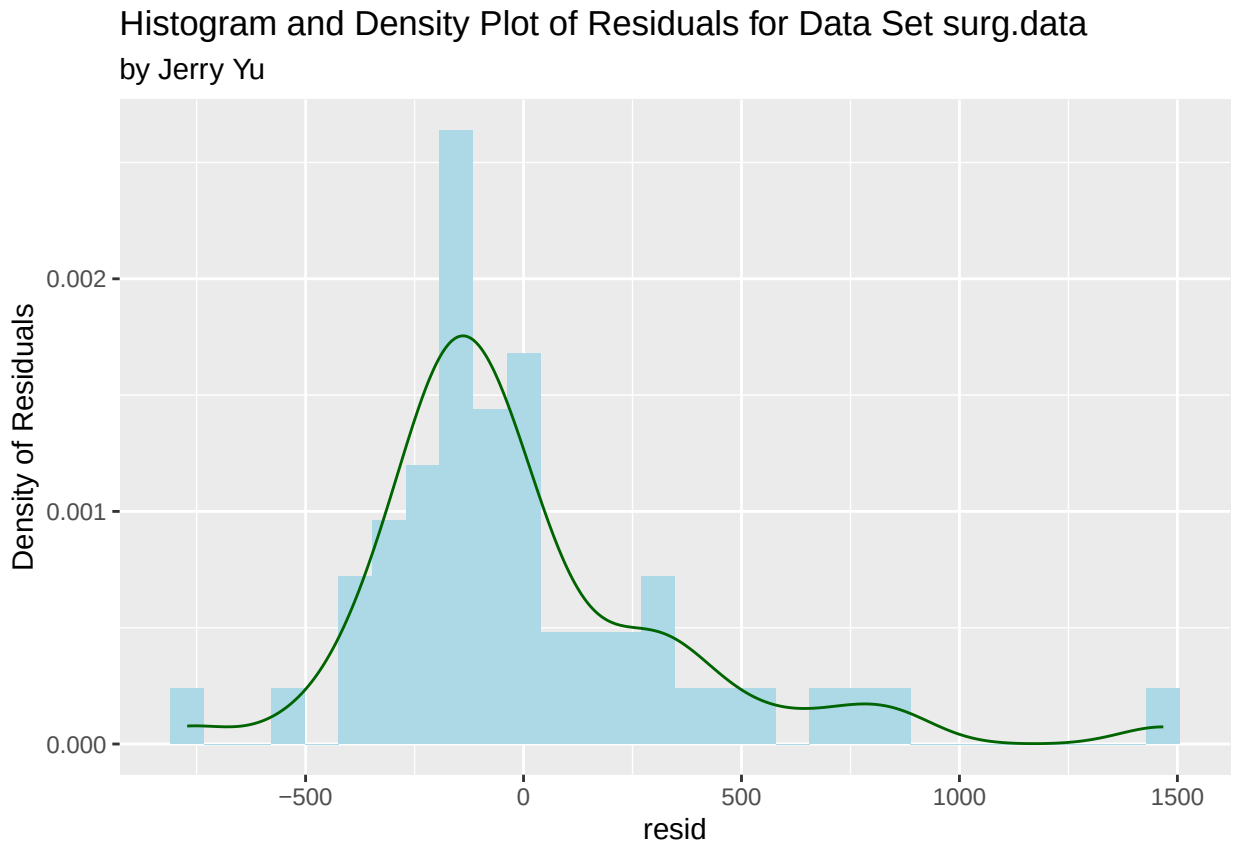
```
## # A tibble: 4 x 3
##     fit resid rstandardized
##   <dbl> <dbl>         <dbl>
## 1  875. 1468.          4.00
## 2  704.  869.          2.33
## 3 1167.  798.          2.42
## 4  953. -772.         -2.14
```

We have 4 outliers, shown in the table `surg.datamro`.

h. Give a Histogram of the residuals and the density curve. Comment about the distribution of residuals.

```
ggplot(data = surg.datamrs, aes(x = resid, y = after_stat(density))) +
  geom_histogram(fill = "lightblue") +
  geom_density(color = "darkgreen") +
  labs(
    title = paste("Histogram and Density Plot of Residuals for Data Set surg.data"),
    subtitle = "by Jerry Yu"
  ) +
  ylab("Density of Residuals")
```

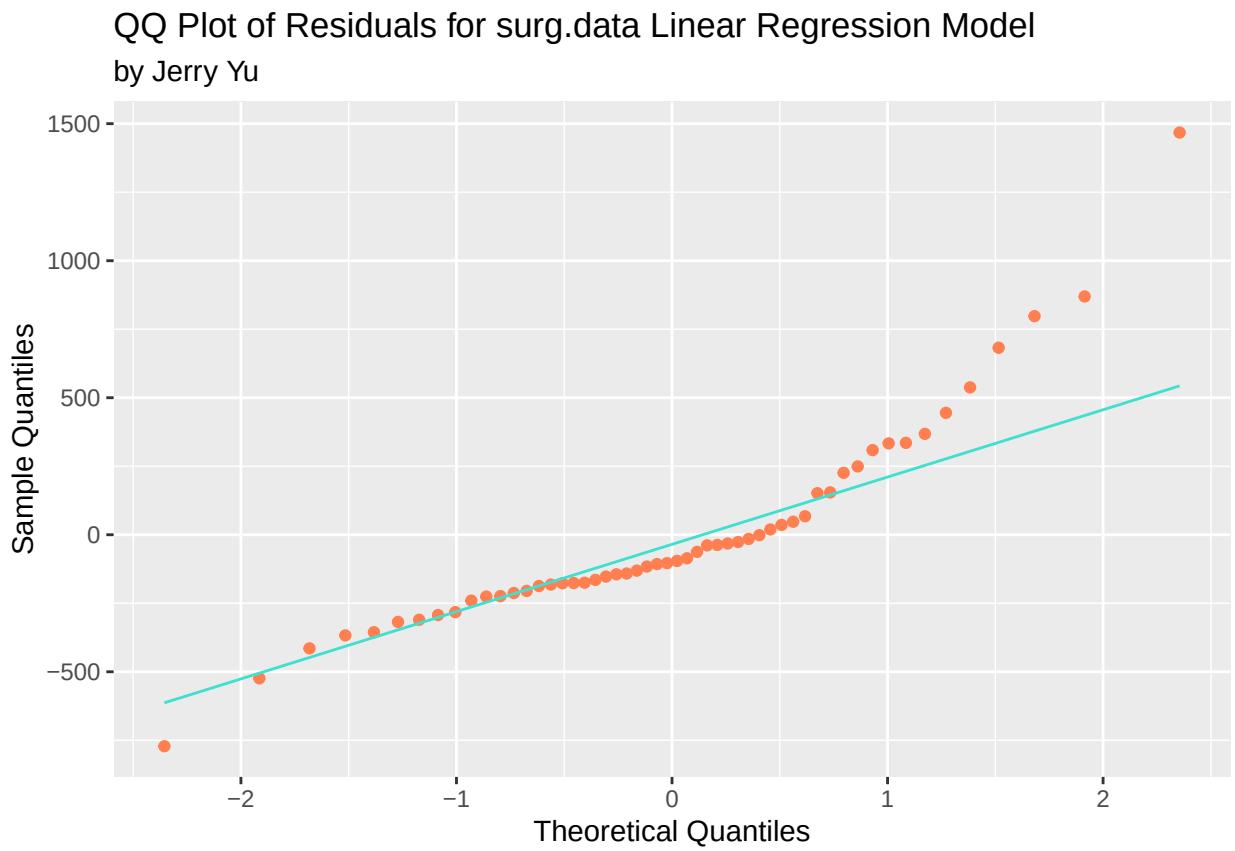
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



There seems to be a slight right skew in the data, 3 of the 4 outliers detected in part g are clearly visible.

i. Give a QQ-plot of the residuals to test for normality of error terms. Comment about the distribution of residuals.

```
ggplot(data = surg.datamrs, aes(sample = resid))+
  geom_qq( color="coral")+
  geom_qq_line( color="turquoise")+
  labs(
    title = paste("QQ Plot of Residuals for surg.data Linear Regression Model"),
    subtitle = "by Jerry Yu"
  ) +
  xlab("Theoretical Quantiles")+
  ylab("Sample Quantiles")
```



The data visually does not look normal, and demonstrates strong signs of heteroscedasticity.The residuals at the end seem to all deviate from the line.

j. Conduct a Shapiro-Wilk Test on the residuals. Be sure to give an appropriate null and alternate hypothesis, test statistic and the p-value. Give the p-value for this test and explain what this means in terms of our model assumptions

```
surg.datasp <- shapiro.test(surg.datamrs$resid)
surg.datasp
```

```
##
##  Shapiro-Wilk normality test
##
## data:  surg.datamrs$resid
## W = 0.87793, p-value = 5.348e-05
```

- H0: The random error is normally distributed
- Ha: The random error is not normally distributed
- Test Statistic: 0.8779272
- p value: 5.3476597×10^{-5}

As p < 0.05, we reject H0 at $\alpha = 0.05$ and conclude that there is statistically significant evidence that the random error is not normally distributed. This means that we cannot assume that the random error is normally distributed for our model, and thus a linear regression without transformation of the data is not advised.