STAT 511: HW #4

Rumil Legaspi

1 March 2021

Contents

Workspace Setup

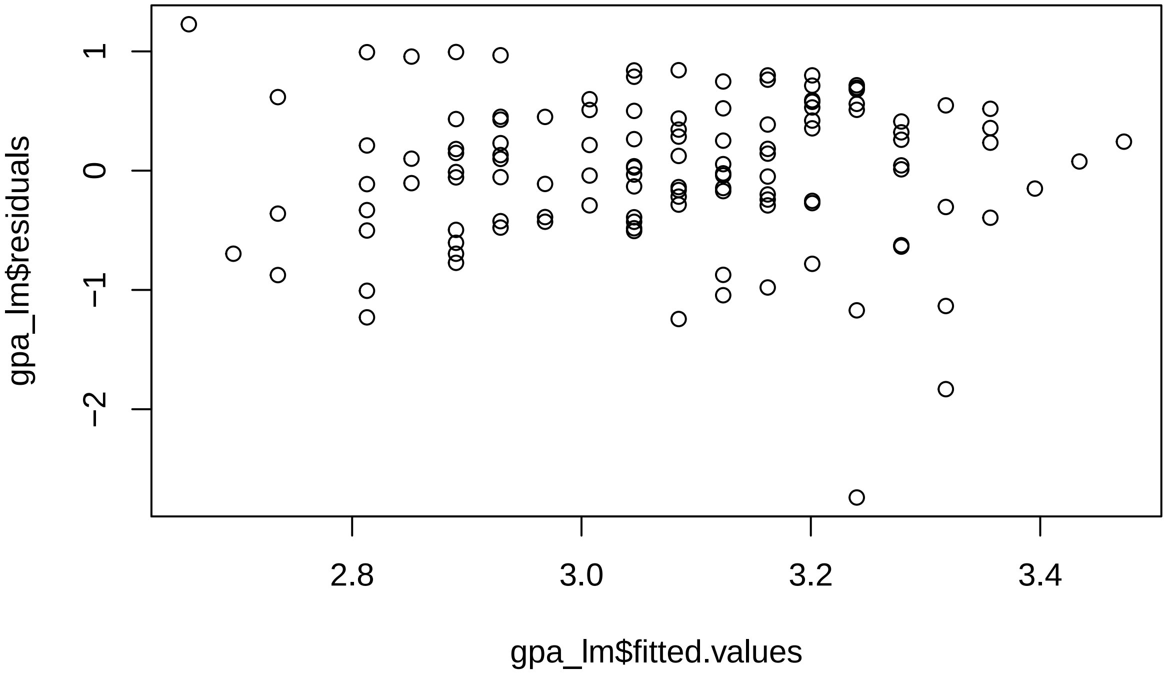
```r
library(nortest)
library(olsrr)
library(car)
library(lmtest)
library(MASS)
library(ggplot2)

setwd("C:/Users/RUMIL/Desktop/APU/STAT 511 - Millie Mao (Applied Regression Analysis)/Week 6/Hw 4")

gpa_data = read.table(file = "GPA.txt", header = FALSE, sep = "")

gpa_data_extended = read.table(file = "GPA_Extended.txt", header = FALSE, sep = "")

# #Adding headers
names(gpa_data) <- c("GPA", "ACT")
names(gpa_data_extended) <- c("GPA", "ACT", "IQ", "Rank")
# names(bank_data) <- c("", "")

#Defining dependent and independent vars
GPA = gpa_data$GPA #Y
ACT = gpa_data$ACT #X1
IQ = gpa_data_extended$IQ #X2
RANK = gpa_data_extended$Rank #X3

gpa_lm = lm(GPA ~ ACT, data = gpa_data)
summary(gpa_lm)
##
## Call:
## lm(formula = GPA ~ ACT, data = gpa_data)
##
## 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 ***
## ACT          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
```

Refer to the GPA problem (GPA.txt) for Questions 1-3

1. Diagnostic plots:

(a). Plot the regression residuals against the predicted values of the variable (residuals on the vertical axis). Check the linearity assumption visually.

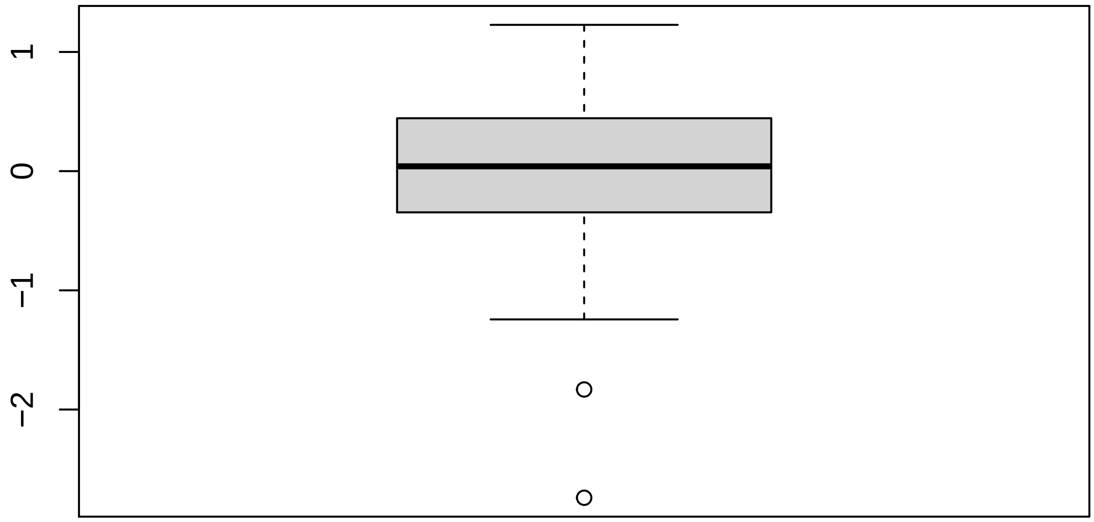
```
plot(gpa_lm$fitted.values, gpa_lm$residuals)
```



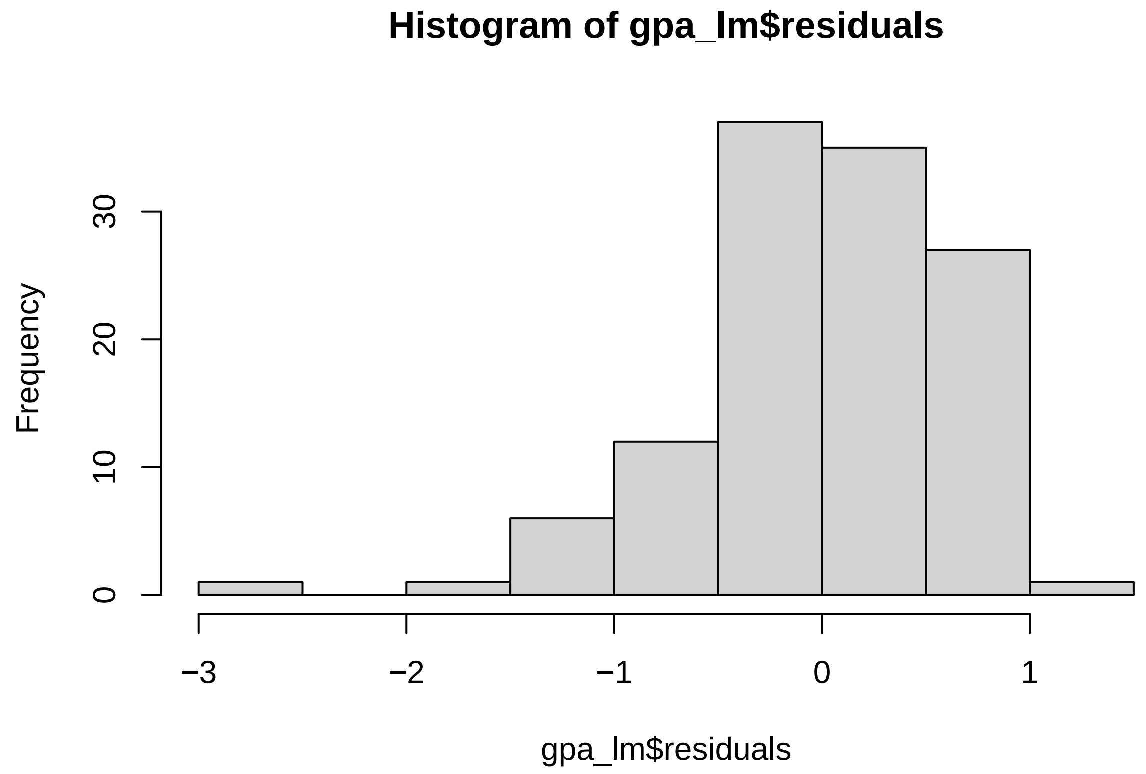
THe assumption of linearity is **not violated** because we are **not seeing** any systemic patterns in the plots.

(b). Draw the boxplot, histogram, and normal probability plot of the regression residuals. Check the normality assumption visually.

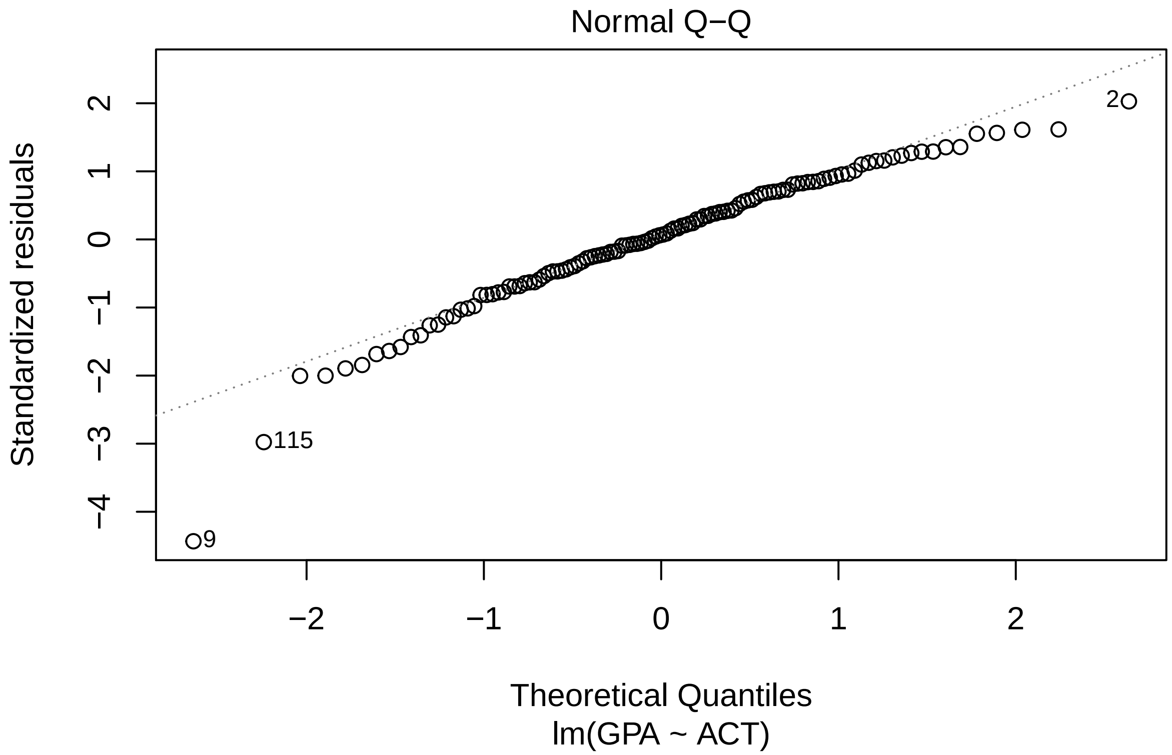
```
#boxplot
boxplot(gpa_lm$residuals)
```



```
#histogram
hist(gpa_lm$residuals)
```



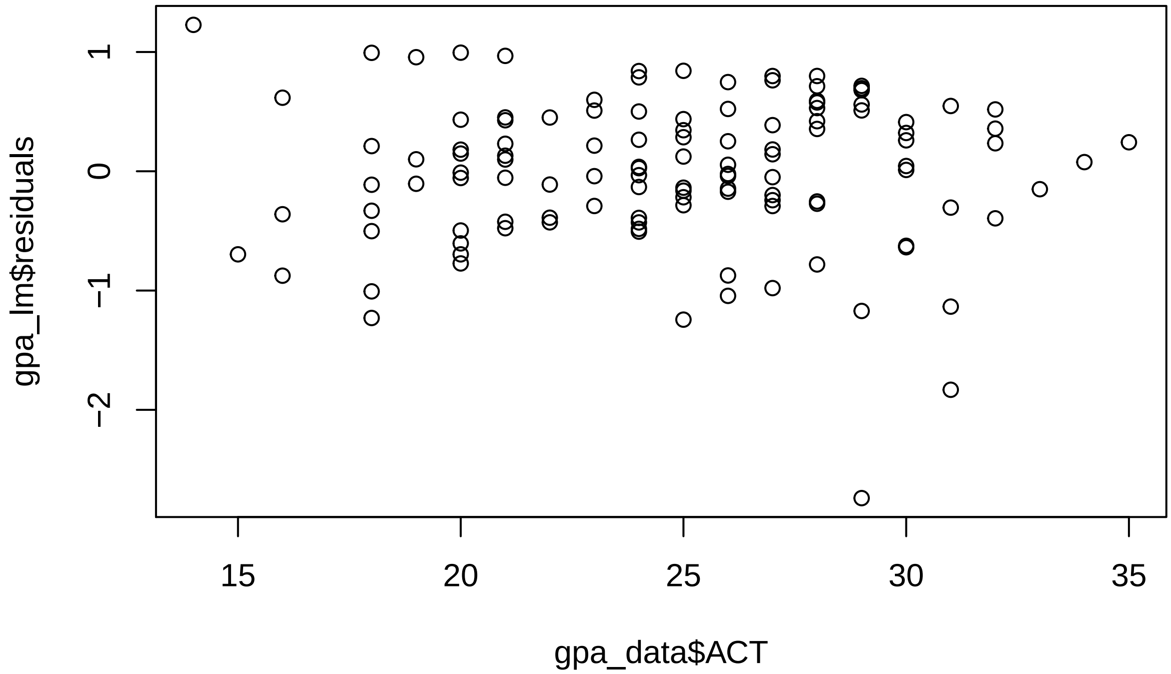
```r
#Plotting specifically for QQ Plot
plot(gpa_lm, c(2))
```



Based on the outputs, the box plot is asymmetrical as shown by the outliers, the histogram is left skewed similar to what the normal Q-Q plot is indicating. These plots show that the assumption of normality is violated.

(c). Plot the regression residuals against the variable (residuals on the vertical axis). Check the equal variance assumption visually.

```
plot(gpa_data$ACT, gpa_lm$residuals)
```



Based on the plot, we can see no systematic pattern and can therefore conclude visually that our assumption of equal variance is not violated.

2. Diagnostic Tests:

(a). Use normality tests to check the normality assumption and draw a conclusion.

Stating our Hypothesis

Null Hypothesis: H_0: The data **is** from a normal distribution

Alternative Hypothesis: H_1: The data is **NOT** from a normal distribution

Testing our Hypothesis

To test these we can use several normality tests using...

- Shapiro-Wilk normality test
- Shapiro-Francia normality test
- Anderson-Darling normality test

```
#Shapiro-Wilk normality test
shapiro.test(gpa_lm$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  gpa_lm$residuals
## W = 0.95249, p-value = 0.0003304
```

```r
#Shapiro-Francia normality test
nortest::sf.test(gpa_lm$residuals)
```

```
##
##  Shapiro-Francia normality test
##
## data:  gpa_lm$residuals
## W = 0.94815, p-value = 0.0003307
```

```r
#Anderson-Darling normality test
nortest::ad.test(gpa_lm$residuals)
```

```
##
##  Anderson-Darling normality test
##
## data:  gpa_lm$residuals
## A = 0.77141, p-value = 0.04384
```

Interpretation of Normality Tests

Looking at the results of these three tests we can see that the p-values are smaller than our alpha. Therefore we reject our NULL hypothesis and that **there is an issue and a violation of our normality assumption.**

(b). Use Modified Levene Test and Breusch-Pagan Test to check the equal variance assumption and draw a conclusion.

Testing Equal Variance Assumptions

Stating our Hypothesis

Null Hypothesis: H_0: The variances in the data **is** equal **Alternative Hypothesis**: H_1: The variances in the data are **NOT** equal

Testing our Hypothesis Using Breusch-Pagan test

```r
#Conducting Levene Test splitting into two groups

#obtaining median of X to use as a threshold
gpa_median = median(gpa_data$ACT)
#ifelse: spliting X into 2 groups
  #one group x < gpa_median, another with x >= gpa_median
  #ifelse( "if this equation is true", "then do this", "else do this")
gpa_group = ifelse(gpa_data$ACT < gpa_median,
                   "Group1",
                   "Group2")

#Levene "Modified" test using median (default in R)
leveneTest(gpa_lm$residuals, gpa_group)
```

```
## Warning in leveneTest.default(gpa_lm$residuals, gpa_group): gpa_group coerced to
## factor.
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   1  0.3997 0.5285
##       118
```

```
#bf.test(infection_lm, data = SENIC)

lmtest::bptest(gpa_lm, studentize = FALSE)
```

```
##
##  Breusch-Pagan test
##
## data:  gpa_lm
## BP = 0.63928, df = 1, p-value = 0.424
```

Both Modified Levene and Breusch-Pagan test gives us high p values indicating that we cannot **reject the null hypothesis** and conclude that there is no issue with our equal variance assumption.

(c). Conduct a lack-of-fit test for the regression model and conclude on the model fitness.

Stating our Hypothesis

Null Hypothesis: H_0: The regression line **IS adequate** in describing the relationship between ACT and GPA

Alternative Hypothesis: H_1: The regression line is **NOT adequate** in describing the relationship between ACT and GPA

Testing our Hypothesis by Conducting our Lack of Fit Test (F Test)

```
#Lack of Fit Test
ols_pure_error_anova(gpa_lm)
```

```
## Lack of Fit F Test
## -----------------
## Response :   GPA
## Predictor:   ACT
```

```
##
##                          Analysis of Variance Table
## ------------------------------------------------------------------------
##                  DF      Sum Sq      Mean Sq      F Value       Pr(>F)
## ------------------------------------------------------------------------
## ACT                1     3.587846     3.587846     9.030747    0.003243287
## Residual         118     45.81761    0.3882848
##  Lack of fit      19     6.485674    0.3413513    0.8591944      0.6324492
##  Pure Error       99     39.33193    0.3972923
## ------------------------------------------------------------------------
```

Based on the small p value 0.003243287, we reject the null hypothesis and conclude with the alternative hypothesis that our regression line is **NOT adequate** in describing the relationship between ACT and GPA.

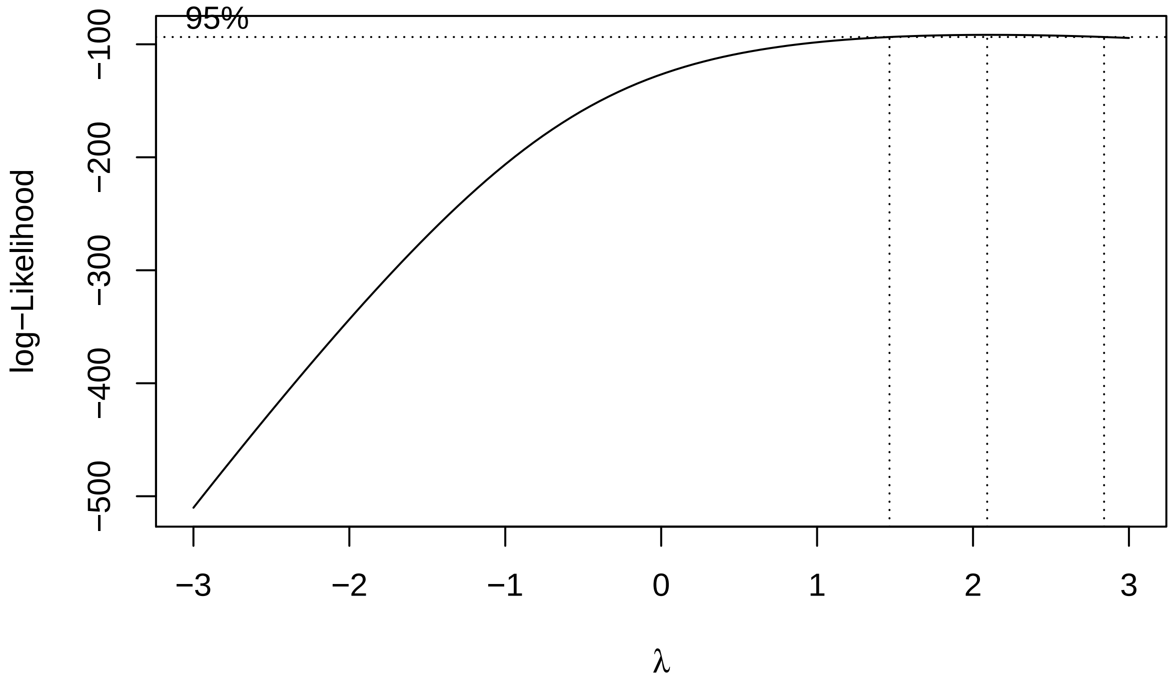
3. Remediation:

(a). Use Box-Cox method to find the best transformation of based on a range of [-3, 3], i.e., what is an approximate value of the optimal in ?

```
#Applying Box-Cox Method
MASS::boxcox(gpa_lm, lambda = seq(-3, 3, by = 0.1), plotit = FALSE)
```

```
## $x
##  [1] -3.0 -2.9 -2.8 -2.7 -2.6 -2.5 -2.4 -2.3 -2.2 -2.1 -2.0 -1.9 -1.8 -1.7 -1.6
## [16] -1.5 -1.4 -1.3 -1.2 -1.1 -1.0 -0.9 -0.8 -0.7 -0.6 -0.5 -0.4 -0.3 -0.2 -0.1
## [31]  0.0  0.1  0.2  0.3  0.4  0.5  0.6  0.7  0.8  0.9  1.0  1.1  1.2  1.3  1.4
## [46]  1.5  1.6  1.7  1.8  1.9  2.0  2.1  2.2  2.3  2.4  2.5  2.6  2.7  2.8  2.9
## [61]  3.0
##
## $y
##  [1] -510.19604 -492.76784 -475.47555 -458.32876 -441.33815 -424.51564
##  [7] -407.87457 -391.42991 -375.19840 -359.19888 -343.45240 -327.98248
## [13] -312.81524 -297.97951 -283.50680 -269.43114 -255.78868 -242.61714
## [19] -229.95486 -217.83967 -206.30744 -195.39047 -185.11583 -175.50376
## [25] -166.56641 -158.30692 -150.71918 -143.78812 -137.49056 -131.79654
## [31] -126.67096 -122.07523 -117.96895 -114.31140 -111.06269 -108.18472
## [37] -105.64182 -103.40108 -101.43256  -99.70933  -98.20731  -96.90513
## [43]  -95.78392  -94.82705  -94.01992  -93.34967  -92.80504  -92.37612
## [49]  -92.05418  -91.83152  -91.70132  -91.65754  -91.69481  -91.80831
## [55]  -91.99373  -92.24719  -92.56518  -92.94453  -93.38232  -93.87593
## [61]  -94.42291
```

```
MASS::boxcox(gpa_lm, lambda = seq(-3, 3, by = 0.1), plotit = TRUE)
```

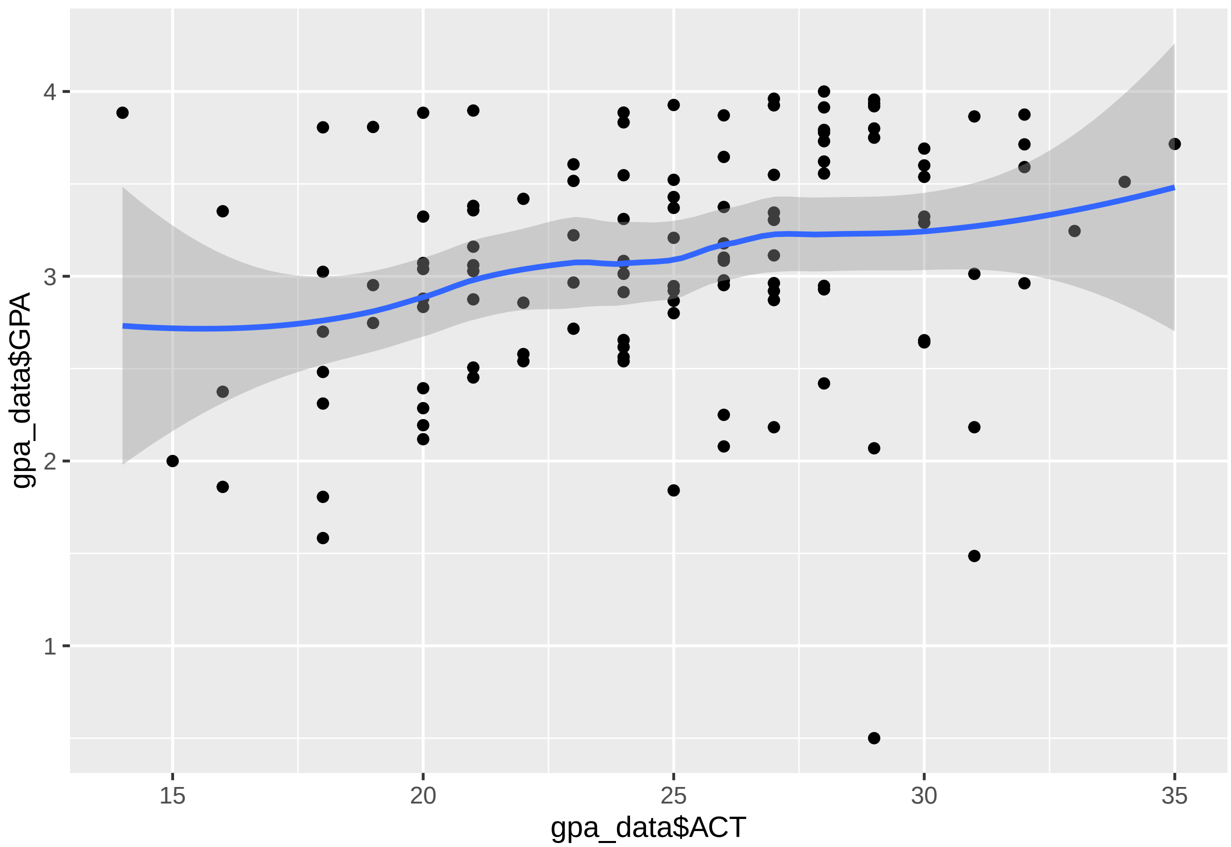


The graph shows us the most optimal value of λ is at roughly 2.1

(b). Plot a smooth curve that best fits the dataset using LOESS method. Is the fitted smooth curve close to linear?

```
#LOESS scatterplot and smoothed curve
smoothplot = qplot(gpa_data$ACT, gpa_data$GPA, geom=c("point", "smooth"))
smoothplot
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



***The fitted smooth curve is somewhat linear although the fit it is not a good fit.

4. Check for Omitted Predictors

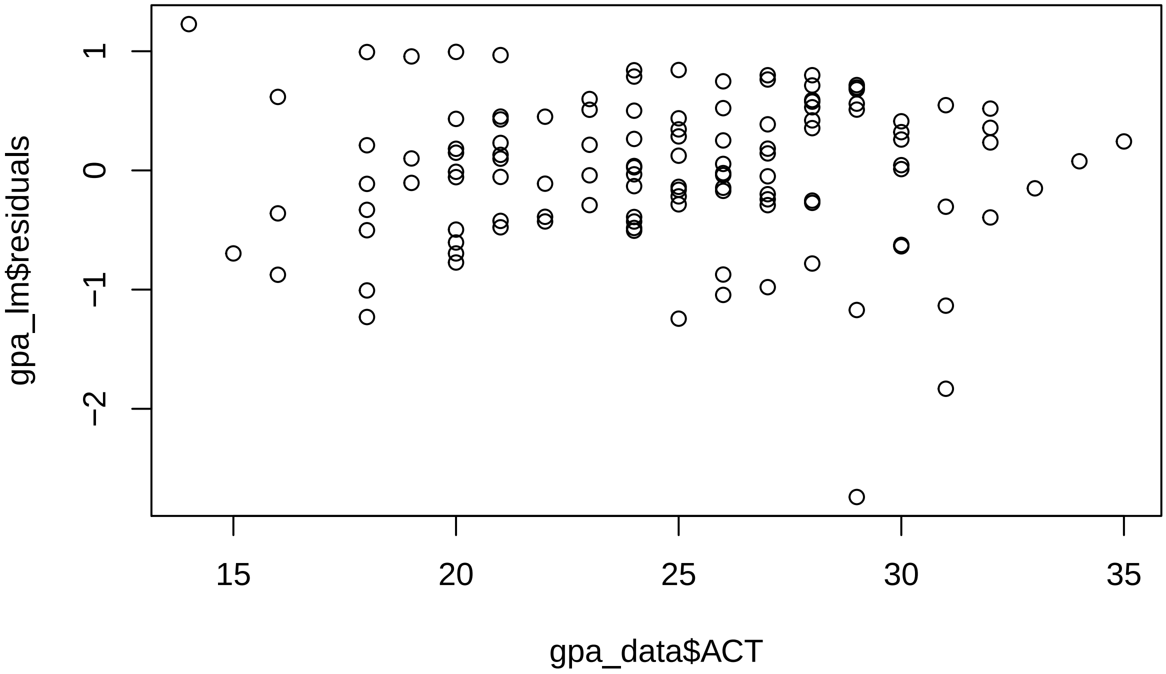
```
#Running a simple linear regression of GPA on ACT
lm(GPA ~ ACT, data = gpa_data)
```

```
##
## Call:
## lm(formula = GPA ~ ACT, data = gpa_data)
##
## Coefficients:
## (Intercept)          ACT
##     2.11405      0.03883
```

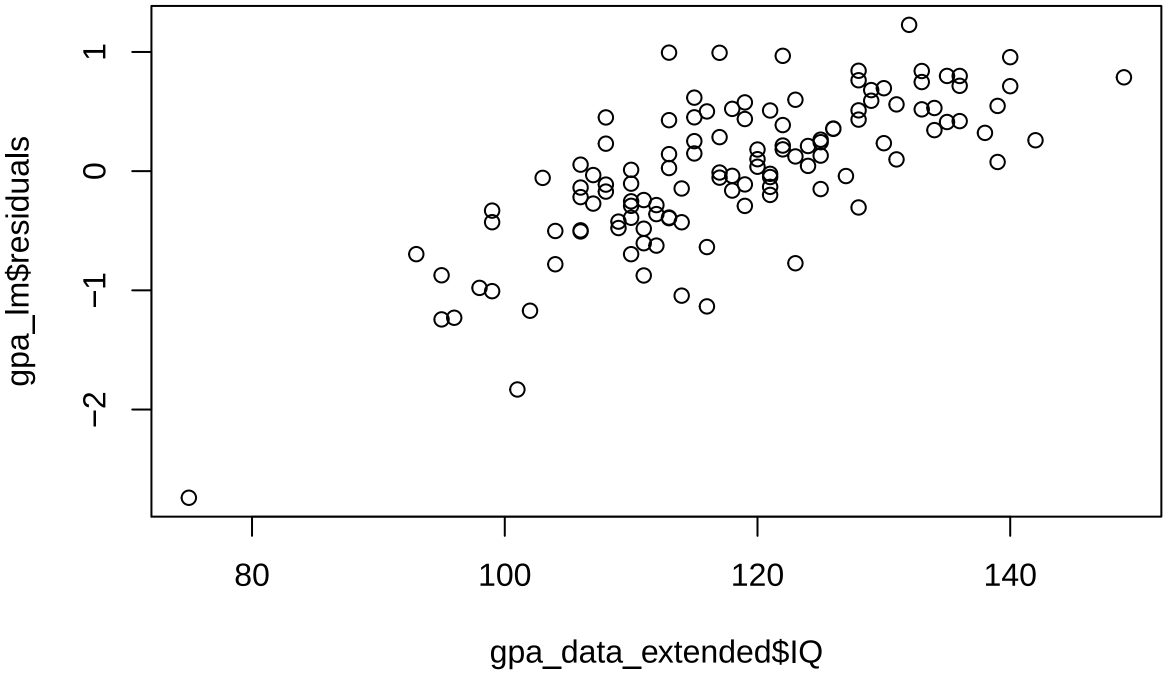
```
gpa_lm
```

```
##
## Call:
## lm(formula = GPA ~ ACT, data = gpa_data)
##
## Coefficients:
## (Intercept)          ACT
##     2.11405      0.03883
```

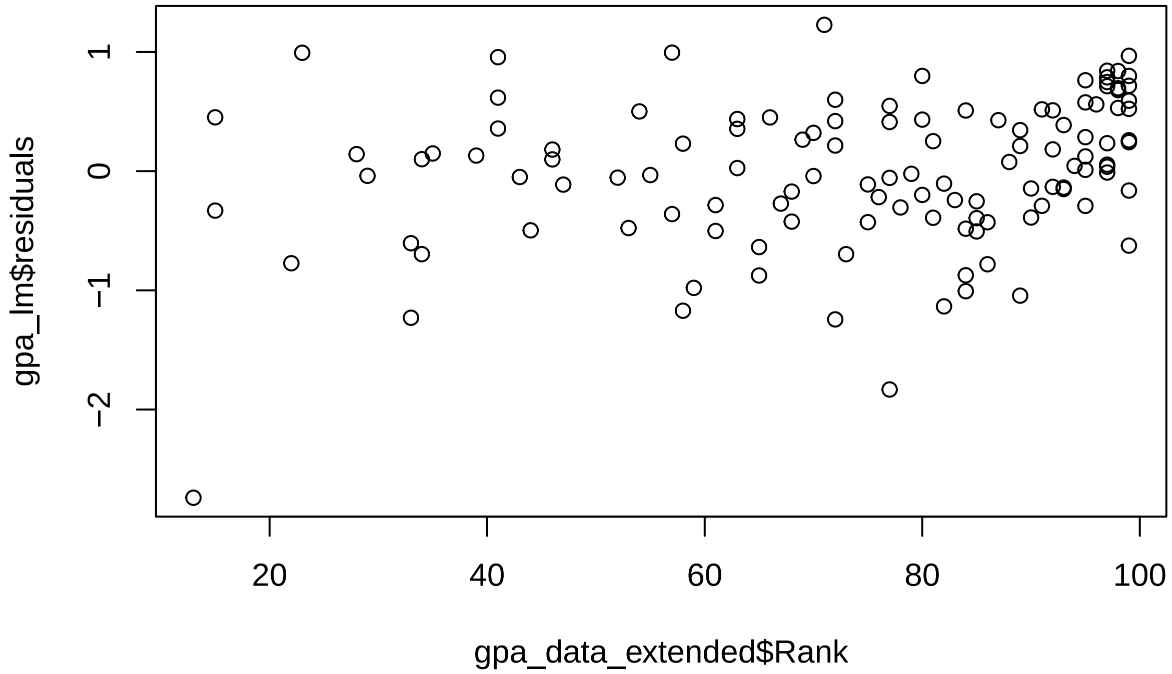
```
#Plotting gpa_lm
plot(gpa_data$ACT, gpa_lm$residuals)
```



```
#plotting regression residuals of gpa_lm against IQ (X2) and Rank(X3)
plot(gpa_data_extended$IQ, gpa_lm$residuals)
```



```
plot(gpa_data_extended$Rank, gpa_lm$residuals)
```



```
summary(lm(GPA ~ IQ, data = gpa_data_extended))
```

```
##
## Call:
## lm(formula = GPA ~ IQ, data = gpa_data_extended)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.1672 -0.2402 -0.0225  0.2977  1.0193
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.873921   0.345709  -5.421  3.2e-07 ***
## IQ           0.041944   0.002915  14.389  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3899 on 118 degrees of freedom
## Multiple R-squared:  0.637,  Adjusted R-squared:  0.6339
## F-statistic:   207 on 1 and 118 DF,  p-value: < 2.2e-16
```

```
summary(lm(GPA ~ Rank, data = gpa_data_extended))
```

```
##
```

```
## Call:
## lm(formula = GPA ~ Rank, data = gpa_data_extended)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.94233 -0.40879  0.05516  0.48679  1.25950
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.306901   0.185497  12.436  < 2e-16 ***
## Rank        0.010417   0.002406   4.329 3.15e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6011 on 118 degrees of freedom
## Multiple R-squared:  0.1371, Adjusted R-squared:  0.1298
## F-statistic: 18.74 on 1 and 118 DF,  p-value: 3.153e-05
```

Based on the outputs, there is a visible pattern when plotting the residuals of our GPA-ACT linear regression model against potentially omitted variable, IQ (X_2). As for Rank it the plots on the graph seem somewhat scattered but more concentration the higher the rank gets.

***What constitutes as a distinct and systematic visible pattern?

IQ is a potentially omitted variable in our GPA-ACT regression model, because the plot shows a linear pattern. Rank…