STAT 511: HW #4

Rumil Legaspi

1 March 2021

## **Workspace Setup**

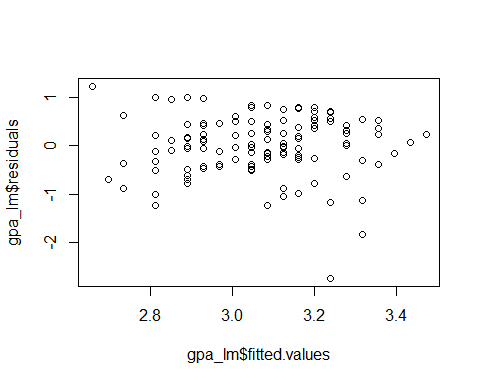
library(nortest)  
library(olsrr)  
library(car)  
library(lmtest)  
library(pagedreport)  
  
setwd("C:/Users/RUMIL/Desktop/APU/STAT 511 - Millie Mao (Applied Regression Analysis)/Week 6/Hw 4")  
  
gpa\_data = read.table(file = "GPA.txt", header = FALSE, sep = "")  
  
gpa\_data\_extended = read.table(file = "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 𝑌 variable (residuals on the vertical axis). Check the linearity assumption visually.**

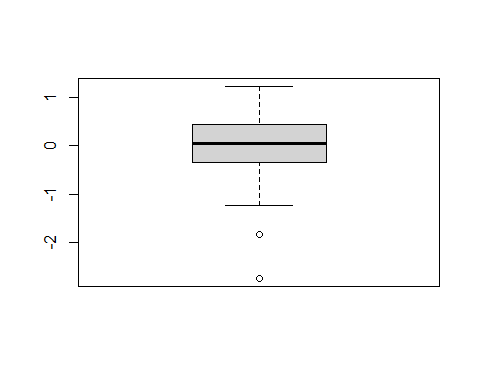
plot(gpa\_lm$fitted.values, gpa\_lm$residuals)



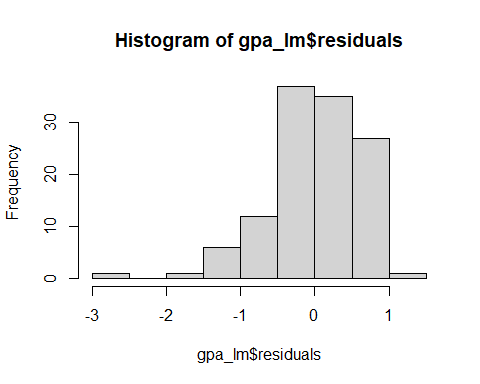
THe assumption of linearity is **not violated** because we are **not seeing** any systemic patterns in the plots.

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

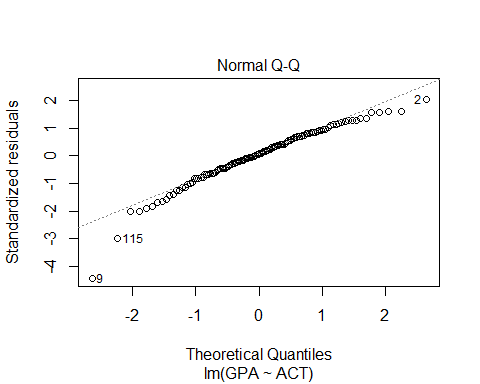
#boxplot  
boxplot(gpa\_lm$residuals)



#histogram  
hist(gpa\_lm$residuals)



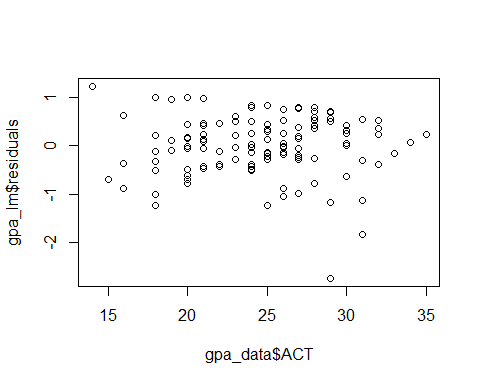
#Plotting specifically for QQ Plot  
plot(gpa\_lm, c(2))



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 𝑋 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**: : The data **is** from a normal distribution

**Alternative Hypothesis**: : 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

#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

#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**: : The variances in the data **is** equal **Alternative Hypothesis**: : The variances in the data are **NOT** equal

### Testing our Hypothesis Using Breusch-Pagan test

#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**: : The regression line **IS adequate** in describing the relationship between ACT and GPA

**Alternative Hypothesis**: : 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 𝑌 based on a range of 𝜆 ∈ [−3, 3], i.e., what is an approximate value of the optimal 𝜆 in 𝑌 𝜆 ?