POLYTECHNIC UNIVERSITY
OF THE PHILIPPINES
STA. MESA, MANILA

SEATWORK NO. 4

Name: Chauncey Miguel P. Francisco

Section: BSAM 4-2

The following methods/tests below can be used to confirm the listed violations on the assumptions of a linear regression model:

1) Auto correlation;

2) Model does not fit the Outliers; and

3) Residuals are not normally distributed.

Using the given excel file (sample_grades.xlsx), a linear regression model is constructed using the following lines of code in R Studio,

```
library(readxl)
sample_grades <- read_excel("C:/Users/Miguel/Documents/MIGUEL/lessons 4th yea
r 1st sem/REA/Seatwork_4/sample_grades.xlsx")

attach(sample_grades)

names(sample_grades)[2] <- 'Year_grad'

# Construct model

library(carData)
library(car)

model_L <- lm(Licensure~Year_grad+Major_grade,sample_grades)
```

If desired, the information summary can be presented in the output,

```
summary(model_L)

##
## Call:
## lm(formula = Licensure ~ Year_grad + Major_grade, data = sample_grades)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -24.126  -1.875   1.256   3.085  10.993
```

```
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3121.262    482.222  -6.473 4.19e-10 ***
## Year_grad       1.518      0.242   6.276 1.29e-09 ***
## Major_grade     1.572      0.223   7.050 1.35e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.303 on 285 degrees of freedom
## Multiple R-squared:  0.3099, Adjusted R-squared:  0.3051
## F-statistic:    64 on 2 and 285 DF,  p-value: < 2.2e-16
```

1 AUTO CORRELATION

Take note of the following hypotheses for autocorrelation assumption,

H_0: Autocorrelation does not exist in the data

H_a: Autocorrelation exists in the data

To test for Autocorrelation, Durbin-Watson Test can be used and is presented in an R studio output below,

```
durbinWatsonTest(model_L)

##  lag Autocorrelation D-W Statistic p-value
##    1       0.3350429      1.329772       0
##  Alternative hypothesis: rho != 0
```

The R output for the p-value is very small that it is rounded off the zero. Due to 0.05 ≥ p-value, the null hypothesis is rejected and therefore Autocorrelation exists in the data.

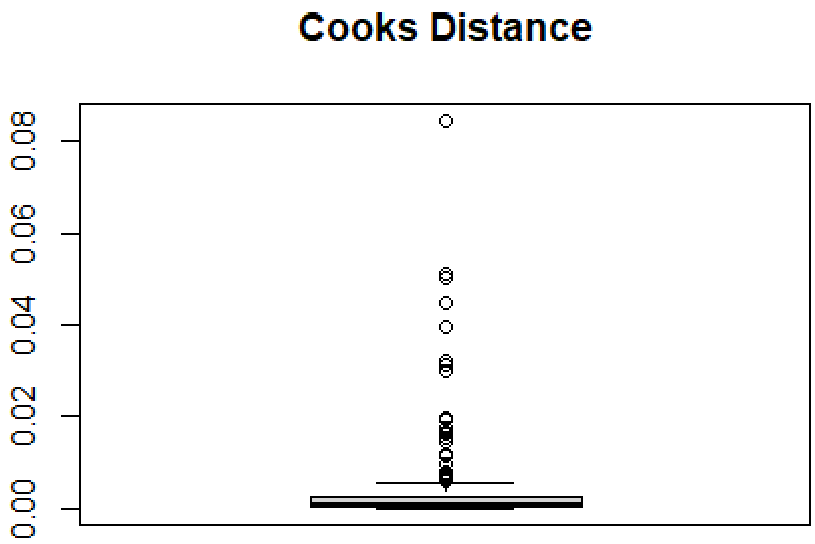
Furthermore, the statistic d=1.329772 is beyond the rule of thumb of 1.5 < d < 2.5, which additionally proves that Autocorrelation is present.

2 MODEL DOES NOT FIT THE OUTLIERS

To detect outliers in the data, Cook's Distance can be used in each data point. Outliers can be detected by visually finding data points with a large Cook's distance. This can be shown in a boxplot,

```
outlier_L <- cooks.distance(model_L)

boxplot(outlier_L, main='Cooks Distance')
```

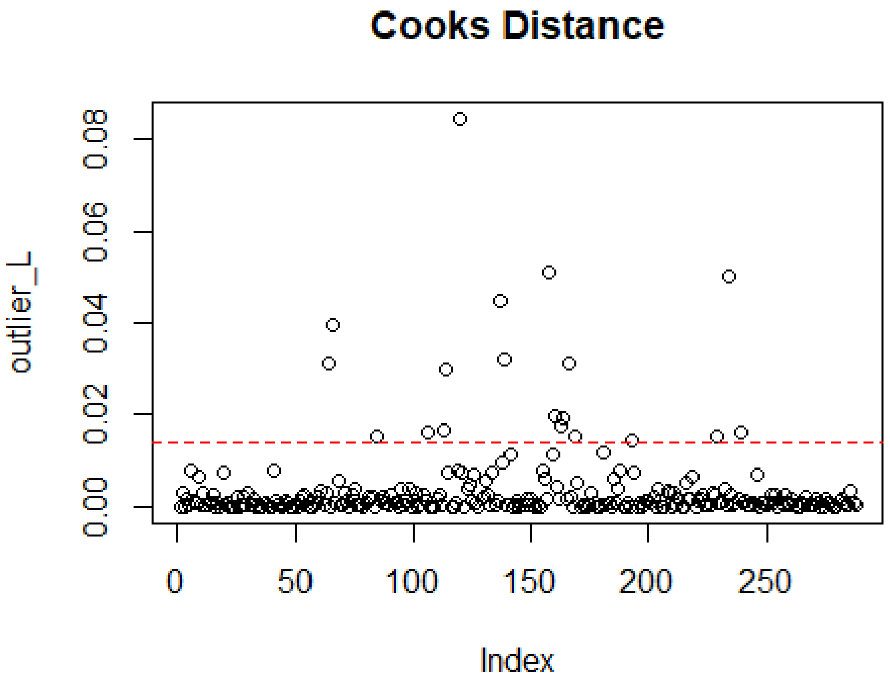


wherein, the output shows the big number of outliers present in the data. This is further showed by the plot() function where the acceptable distance of 4/n (n: number of observations) is plotted. Any point beyond this area is generally considered to be an outlier.

```
plot(outlier_L, main='Cooks Distance')

n <- nrow(sample_grades)
```

```
abline(h=4/n, lty=2, col='Red')
```



3 RESIDUALS ARE NOT NORMALLY DISTRIBUTED

Testing this statistically requires a normality test. The following tests are used: 1) Shapiro-Wilk Test 2) Kolmogorov-Smirnov Test 3) Cramer-von Mises Test 4) Anderson-Darling Test Also take note of the following hypotheses for normality assumption, (insert text box)

H_0: *Residuals are normally distributed*

H_a: *Residuals are not normally distributed*

```
resid_L <- resid(model_L)

library(olsrr)

##
## Attaching package: 'olsrr'

## The following object is masked from 'package:datasets':
##
##     rivers

olsrr::ols_test_normality(resid_L)

## Warning in ks.test.default(y, "pnorm", mean(y), sd(y)): ties should not be
## present for the Kolmogorov-Smirnov test

## -----------------------------------------------
##        Test             Statistic       pvalue
## -----------------------------------------------
## Shapiro-Wilk              0.8741         0.0000
## Kolmogorov-Smirnov        0.1417         0.0000
## Cramer-von Mises         22.5552         0.0000
## Anderson-Darling         10.6448         0.0000
## -----------------------------------------------
```

Just like in the first assumption, p-values are very small that it is approximated to be zero. Even with the Kolmogorov-Smirnov Test having ties in its test, it still gave a very small p-value. Hence, with 0.05 ≥ p-value, the null hypothesis is rejected and therefore be resulted in the Residuals being not normally distributed.