Probability and Statistical Inference
Continuous Assessment Part II

Student Number: D20123666

Name: Amit Kumar Gangwar

Programme Code: TU059

Section 1 - Research Question

Whether a student's ability of Critical Reading and English language skills affects their competency as a citizen, their ability to understand the social world.

Section 2 - Dataset

Dataset is about academic performance evolution for engineering students. The dataset contains the results in national assessments for secondary and university education in engineering students and contains academic, social, economic information for 12,411 students. The data was collected as part of the Master's Degree in Engineering project of the Technological University of Bolívar (UTB) titled Academic Efficiency Analysis in Engineering students

A full descriptor is available at:
https://www.sciencedirect.com/science/article/pii/S2352340920304315#utbl0001

Link To download dataset: https://data.mendeley.com/datasets/83tcx8psxv/1

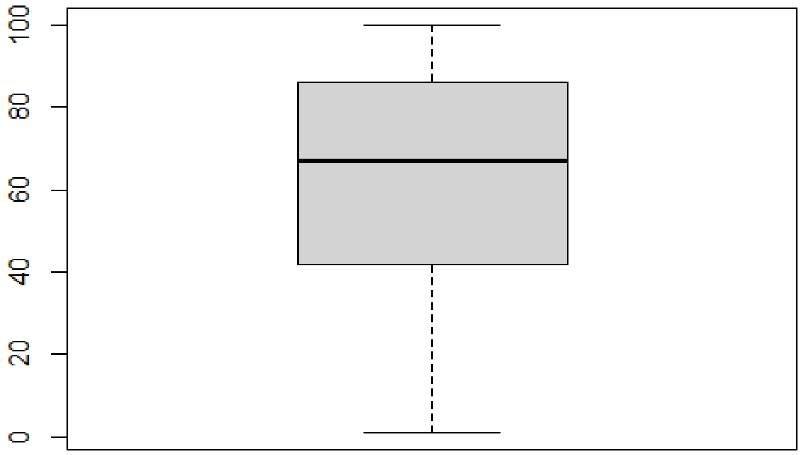
Reference: Delahoz-Dominguez, E., Zuluaga, R., & Fontalvo-Herrera, T. (2020). Dataset of academic performance evolution for engineering students. *Data in Brief*, *30*, 105537. https://doi.org/10.1016/j.dib.2020.105537

- Missingness: There are no missing values in the Dataset
- The variables of interest are **Nature of School (SCHOOL_NAT), Critical Reading (CR_PRO), Citizen Competencies SPRO(CC_PRO), English (ENG_PRO)**
- Relevant descriptive statistics and visuals:

1. **Critical Reading (CR_PRO)**

   The sample as a whole has above average scores in critical reading (M = 62.2, SD = 27.66)

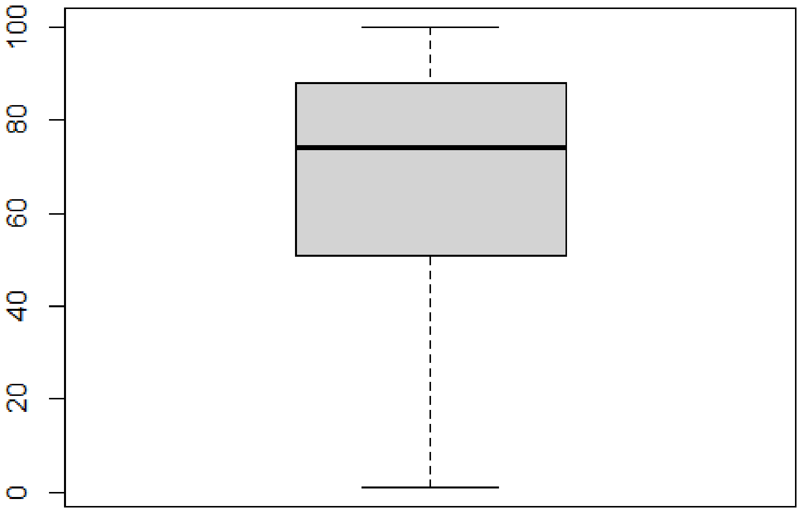
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Std. Dev.	Variance	Outliers
1.0	42.0	67.0	62.2	86.0	100.0	27.66656	765.4384	NA



2. English (ENG_PRO)

The sample as a whole has above average scores in critical reading (M = 67.5, SD = 25.5)

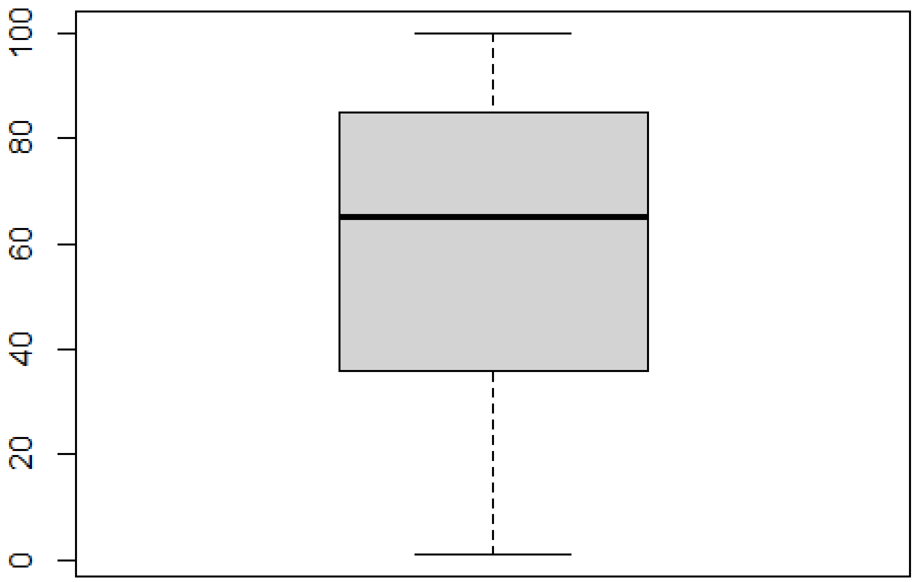
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Std. Dev.	Variance	outliers
1.0	51.0	74.0	67.5	88.0	100.0	25.4951	649.9999	NA



3. Citizen Competencies SPRO(CC_PRO)

The sample as a whole has average scores in critical reading (M = 59.2, SD = 28.99)

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Std. Dev.	Variance	Outliers
1.00	36.00	65.00	59.19	85.00	100.00	28.99184	840.5268	NA



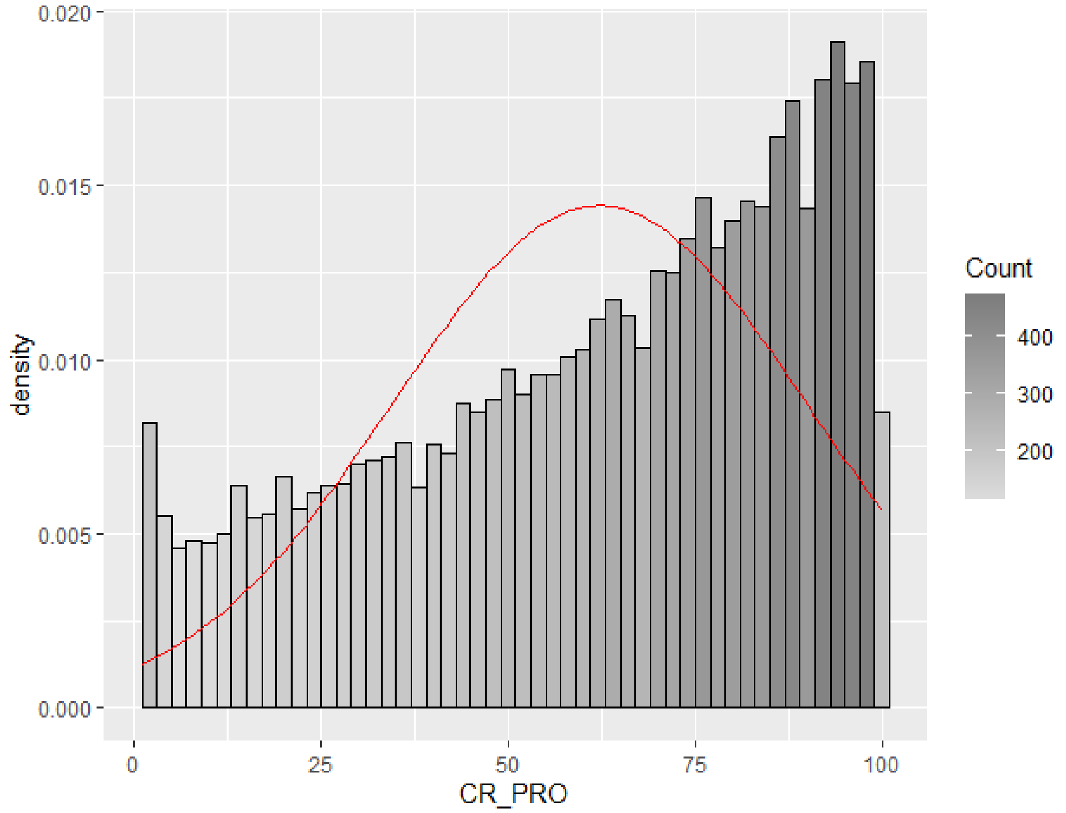
4. **Nature of School (SCHOOL_NAT)**
Nearly half of the students studied in private school (52.89%) and rest half studied in public school (47.1).

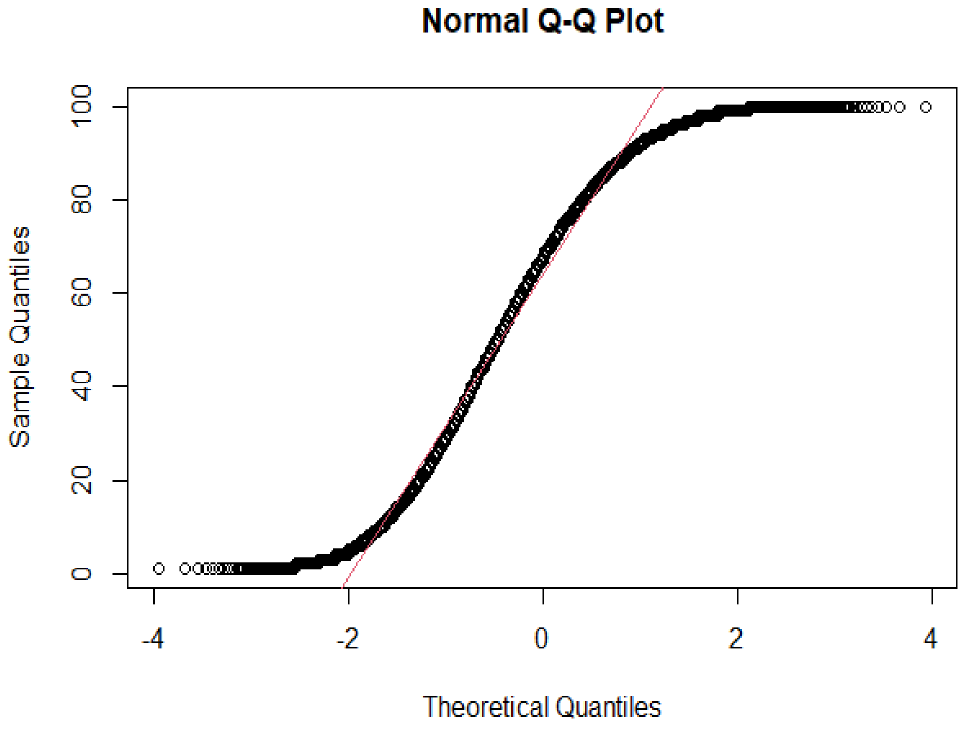
Total values: 12411

Distinct values:	PRIVATE	PUBLIC
	6565	5846

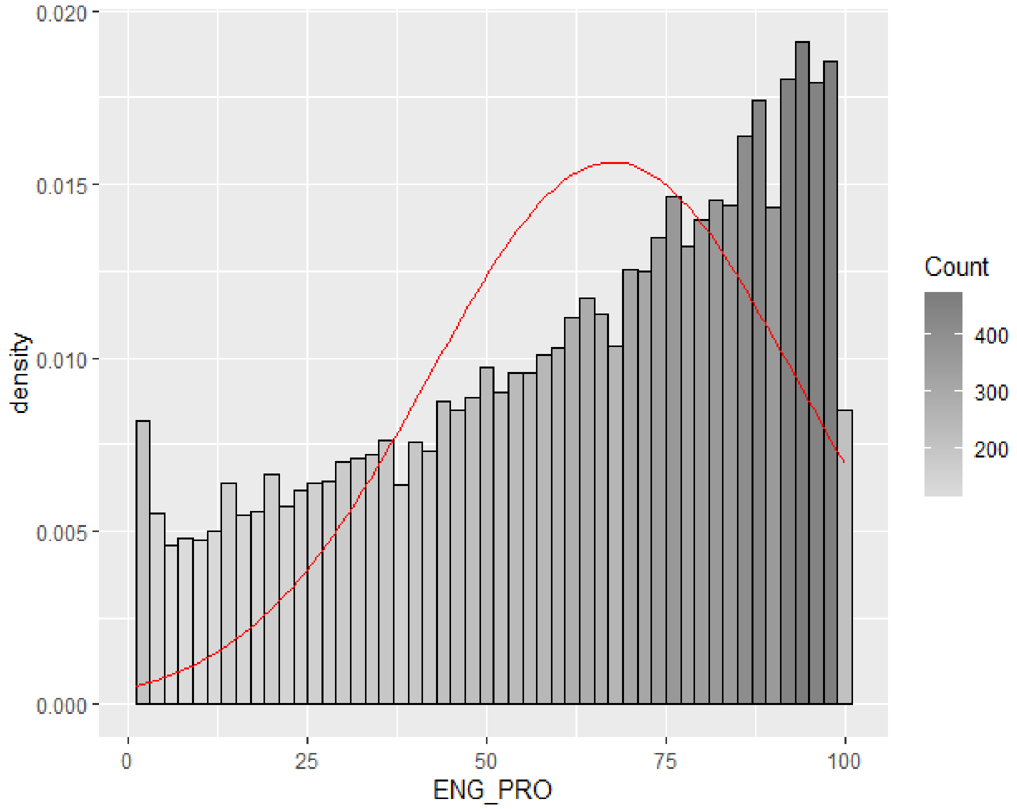
- Assessing normality:

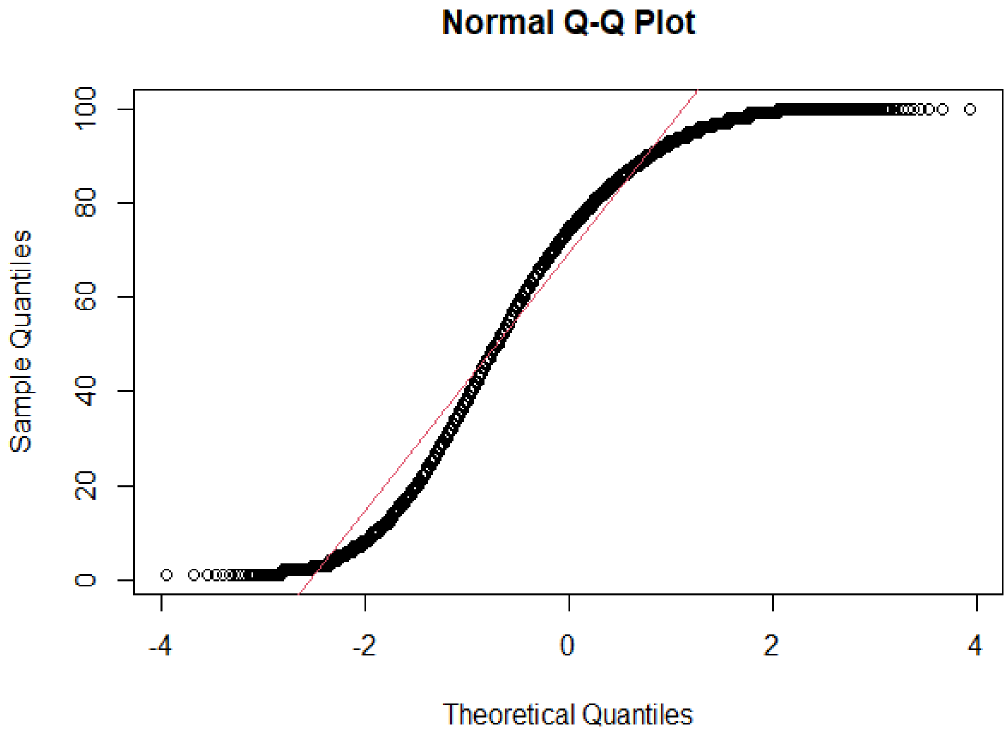
1. **Critical Reading (CR_PRO)**
The input variable (CR_PRO) is approximately normally distributed as 99.9% of z-scores lie between −3.29 and 3.29.



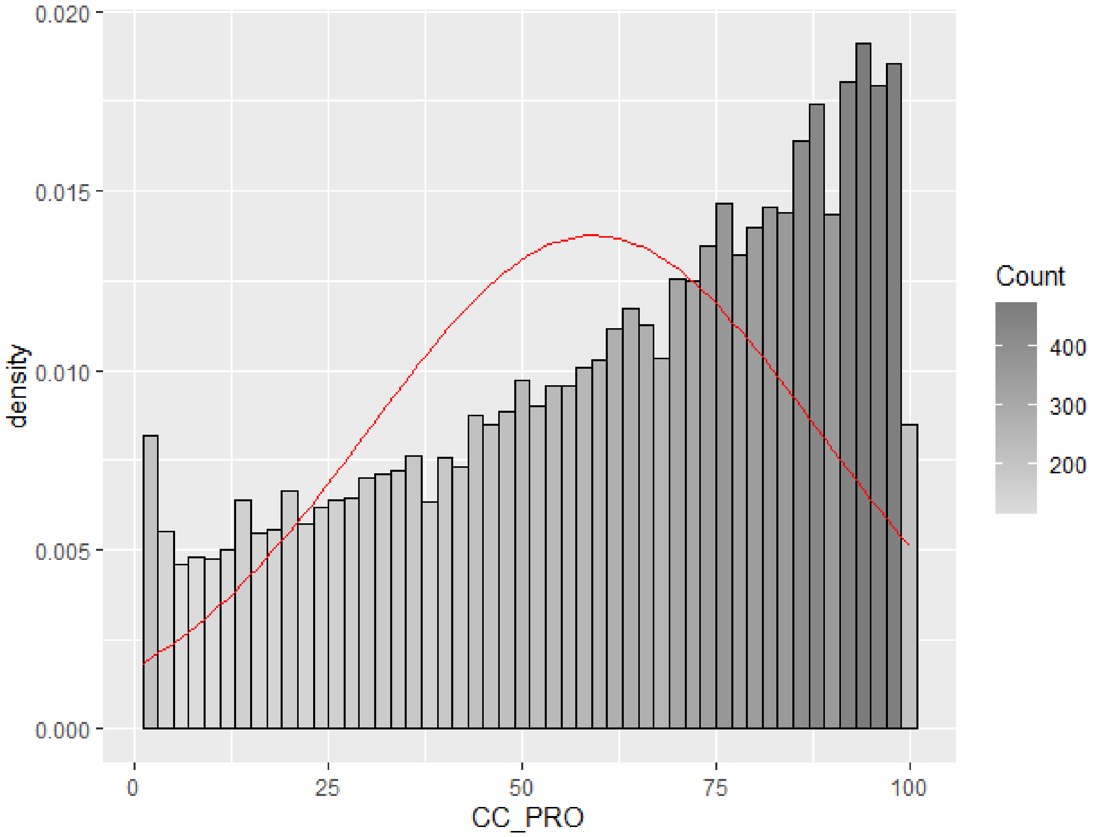


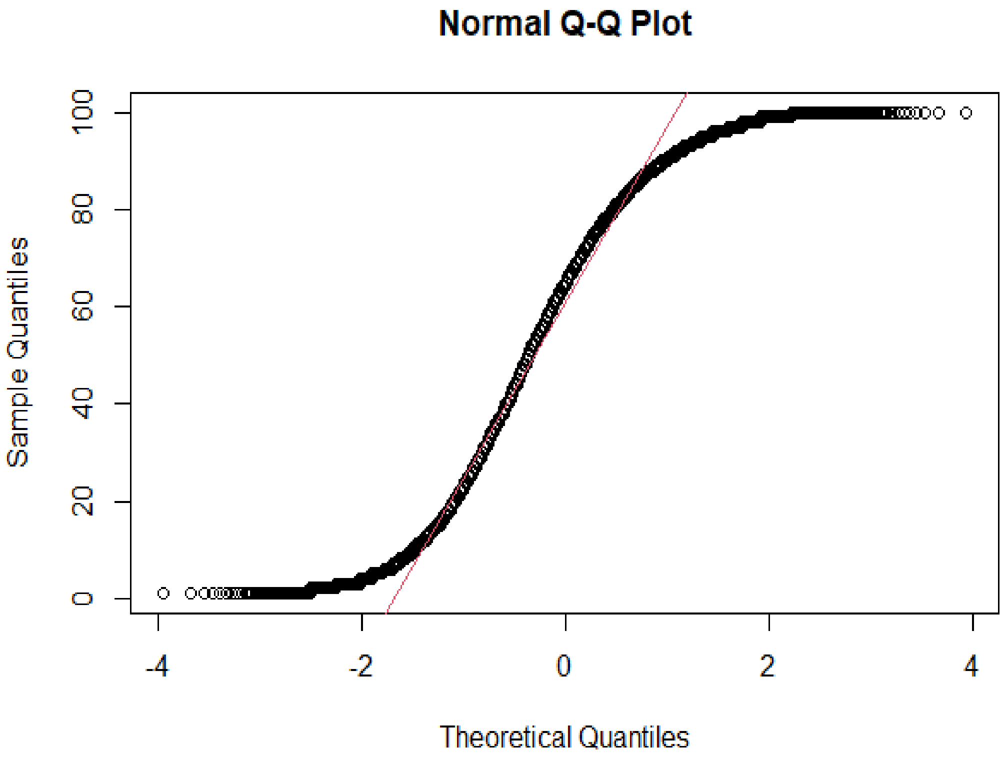
2. **English (ENG_PRO)**
The input variable (ENG_PRO) is approximately normally distributed as 99.9% of z-scores lie between −3.29 and 3.29.





3. **Citizen Competencies SPRO(CC_PRO)**
   The response variable (CC_PRO) is approximately normally distributed as 99.9% of z-scores lie between −3.29 and 3.29.


density
0.020
0.015
0.010
0.005
0.000
0
25
50
75
100
CC_PRO
Count
400
300
200

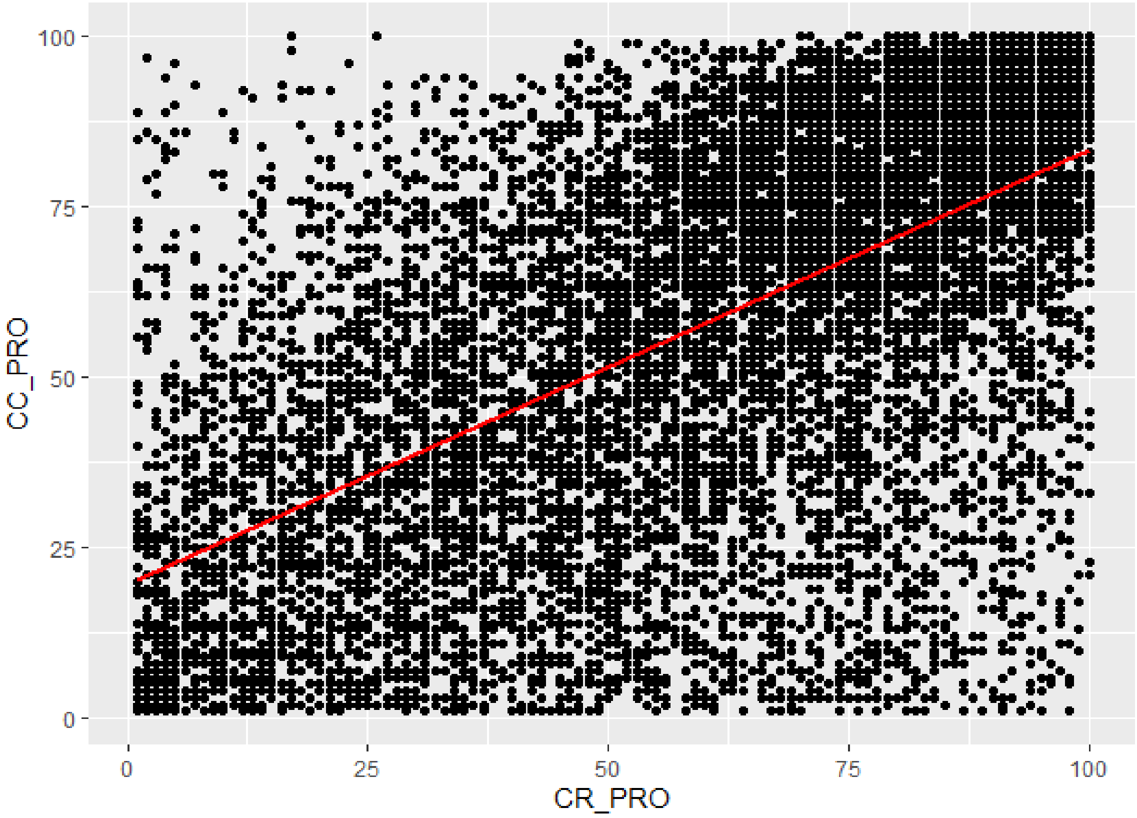

Normal Q-Q Plot
Sample Quantiles
0
20
40
60
80
100
-4
-2
0
2
4
Theoretical Quantiles

Section 3 - Results

Section 3.1 - Statistical Evidence

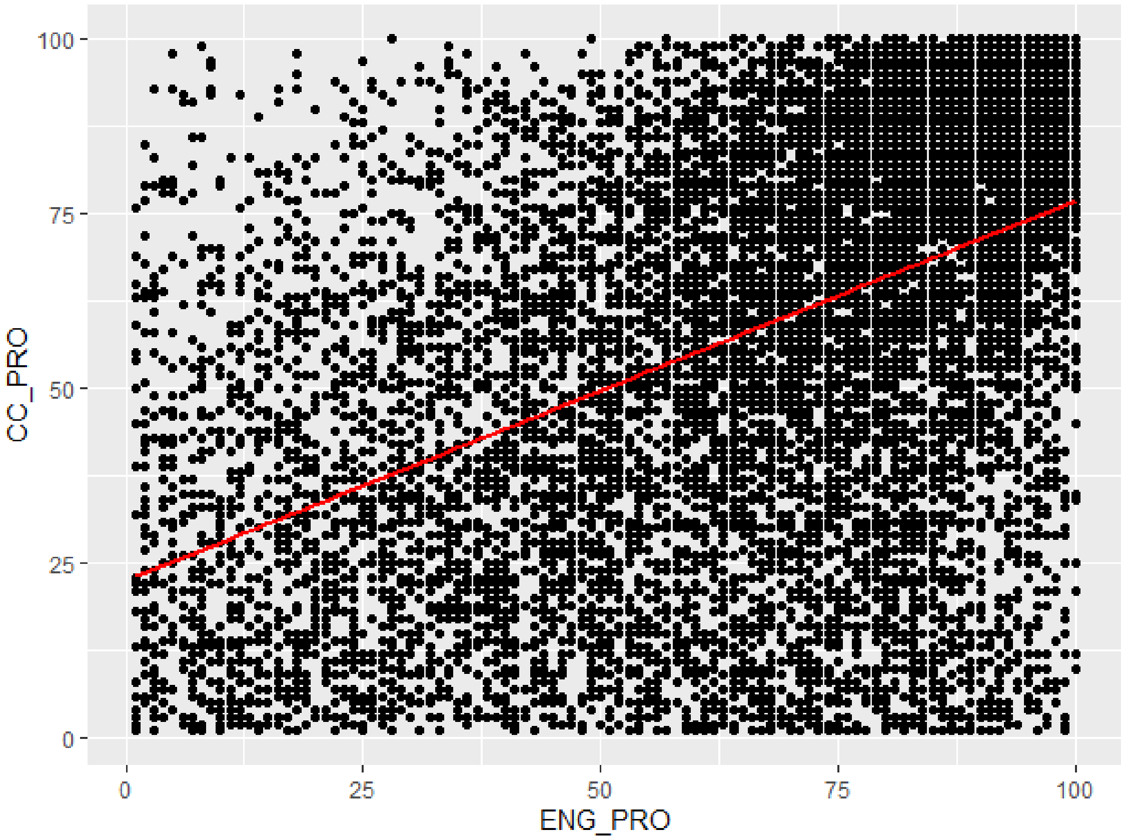
1. **Correction Test between CR_SPRO and CC_PRO**:
   Null Hypothesis: There is a significant difference between CR_SPRO scores and CC_PRO scores of students.
   Alternate Hypothesis: There is no significant difference between CR_SPRO scores and CC_PRO scores of students.

   Results:
   A Pearson correlation coefficient was computed to assess the relationship between Critical Reading (CR_PRO) and Citizen Competencies SPRO(CC_PRO) exam marks. There was a significant positive correlation between the two variables, r = 0.608, n = 12409, p < 2.2e-16. A scatterplot below summarizes the results.



2. **Correlation Test between ENG_PRO and CC_PRO:**
   Null Hypothesis: There is a significant difference between ENG_PRO scores and CC_PRO scores of students.
   Alternate Hypothesis: There is no significant difference between ENG_PRO scores and CC_PRO scores of students.

   Results:
   A Pearson correlation coefficient was computed to assess the relationship between English (ENG_PRO) and Citizen Competencies SPRO(CC_PRO) exam marks. There was a significant positive correlation between the two variables, r = 0.48, n = 12409, p < 2.2e-16. A scatterplot below summarizes the results.



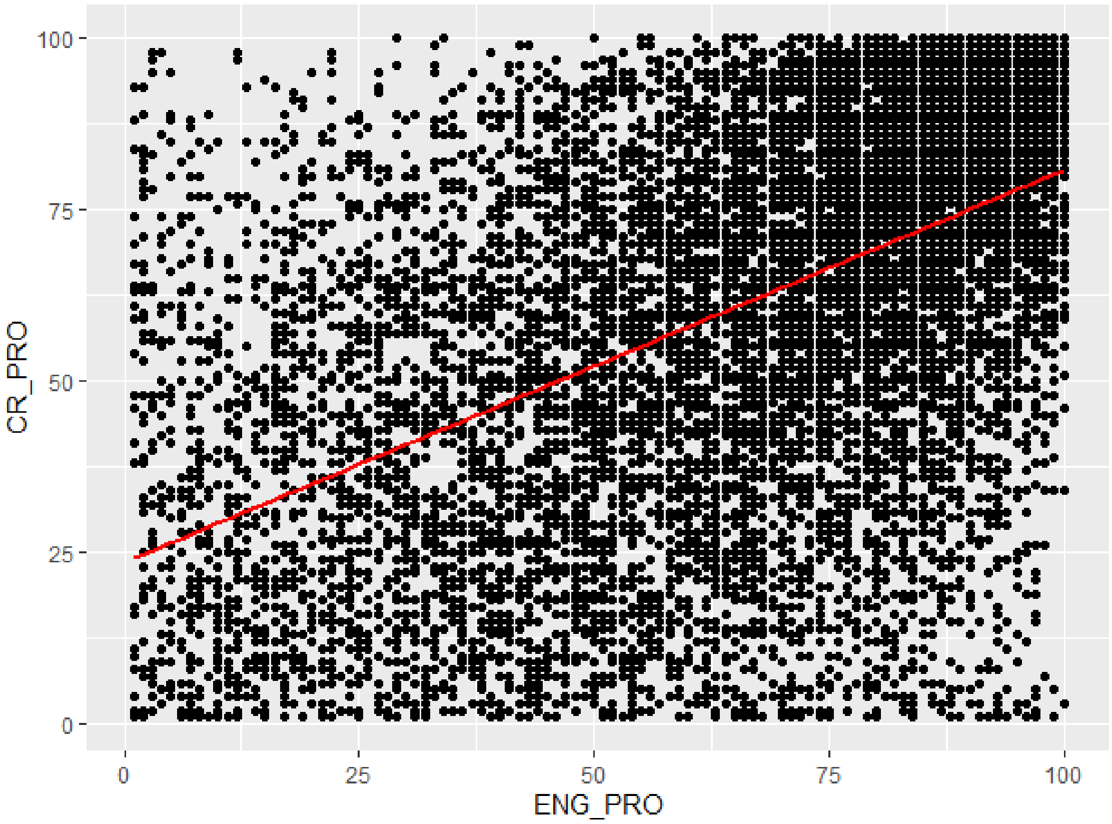
3. **Correction Test between ENG_PRO and CR_SPRO:**

Null Hypothesis: There is a significant difference between ENG_PRO scores and CR_SPRO scores of students.

Alternate Hypothesis: There is no significant difference between ENG_PRO scores and CR_SPRO scores of students.

Results:

A Pearson correlation coefficient was computed to assess the relationship between English (ENG_PRO) and Critical Reading (CR_PRO) exam marks. There was a significant positive correlation between the two variables, r = 0.53, n = 12409, p < 2.2e-16. A scatterplot below summarizes the results. The collinearity between these independent variables doesn't exist since correlation is not very strong (r <8) so these variables can be used as input variables in the prediction model.



Section 3.2 - Model1 Results

A multiple linear regression model was calculated to predict student’s competency as a citizen (denoted by CC_PRO) based on their ability of Critical Reading (denoted by CR_PRO) and their English language skills (denoted by ENG_PRO). A significant regression equation was found **(F (2, 12408) = 4,201.48, p<0.01),** with an **R2 of 0.404**. Participants predicted competency score is equal to **10.333704+ 0.247549 (ENG_PRO) + 0.516789 (CR_PRO)**, where ENG_PRO and CR_PRO are scores which can vary from 0 to 100. Participant's CC_PRO score increases by 0.247549 for each unit in ENG_PRO and by 0.516789 unit for each unit increase in CR_PRO. Both CR_PRO and ENG_PRO were significant predictors of CC_PRO.

CC_PRO = 10.333704+ 0.247549 (ENG_PRO) + 0.516789 (CR_PRO)

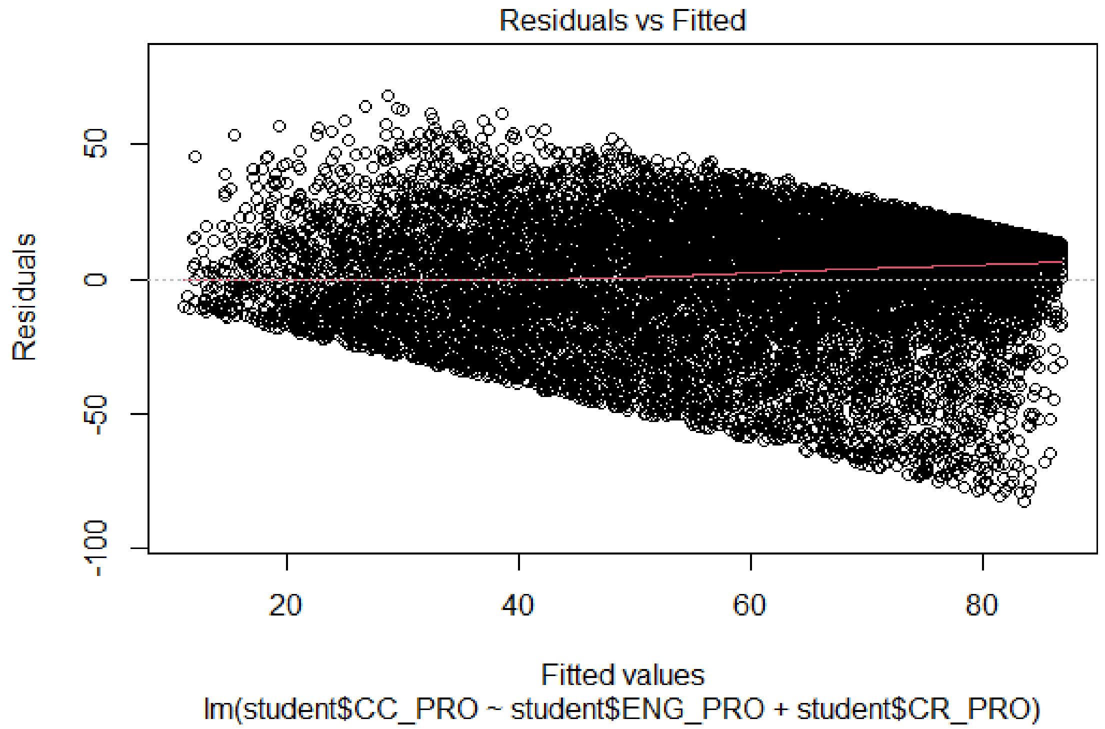
Let’s take mean values for ENG_PRO (67.5) and CR_PRO (62.2) and calculate CC_PRO

CC_PRO = 10.333704+ 0.247549 (67.5) + 0.516789 (62.2) = 59.09

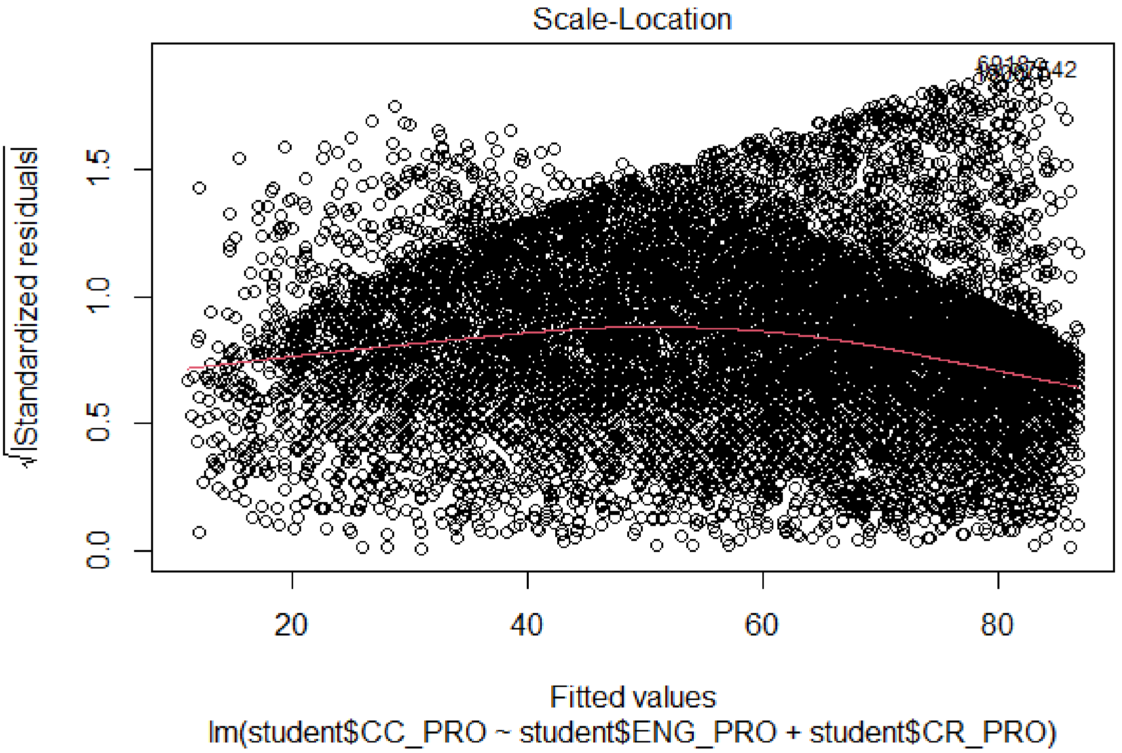
59.09, Which is very close to the mean value of CC_PRO (59.2), this proves that model is significant with very small error term.

Assessing the model against assumptions:

Homoscedasticity


Residuals vs Fitted
Residuals
50
0
-50
-100
20
40
60
80
Fitted values
lm(student$CC_PRO ~ student$ENG_PRO + student$CR_PRO)

PLOT 1


Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
20
40
60
80
Fitted values
lm(student$CC_PRO ~ student$ENG_PRO + student$CR_PRO)

PLOT 2

PLOT1 is the chart of residuals vs fitted values; in PLOT2 the standardised residuals are on the Y axis. If there is absolutely no heteroscedastity, we should see a completely random, equal distribution of points throughout the range of X axis and a flat red line. We really want to see that there is no pattern in the residuals and that they are equally spread around the y = 0 line - the dashed line.

As you can notice the red line is later slightly lifted upward on plot 1 but is not a big problem. Looking at the second plot we can see that while it is a problem it is not huge. Only a concern if there are definite patterns.

Collinearity

Tests to see if the data met the assumption of collinearity indicated that multicollinearity was not a concern (ENG_PRO, Tolerance = 0.7221794, VIF = 1.384697; CR_PRO, Tolerance = 0.7221794 VIF = 1.384697).

Fit and Goodness

The independent variables explain explains 40.4% (Adjusted R2 = 0.404) variations in the students' CC_PRO scores (competency as a citizen).

Section 3.3 - Model 2 Results

A multiple linear regression model was calculated to predict student's competency as a citizen (denoted by CC_PRO) based on their ability of Critical Reading (denoted by CR_PRO) and their English language skills (denoted by ENG_PRO) and the type of school they study in (denoted by SCHOOL_NAT). A significant regression equation was found **(F (3, 12407) = 2,804.05, p<0.01),** with an **R2 of 0.404**. Participants predicted competency score is equal to **9.38 + 0.255 (ENG_PRO) + 0.515 (CR_PRO) + 1.042568 (SCHOOL_NAT), where SCHOOL_NAT is coded as 'PUBLIC=1', which is the dummy variable of interest and 'PRIVATE=0'**, ENG_PRO and CR_PRO are scores which can vary from 0 to 100. Participant's CC_PRO score increases by 0.255 for each unit in ENG_PRO and by 0.515 unit for each unit increase in CR_PRO and 1.042568 for School being Public where it is 0 for private school, so **there is a positive differential effect for Public schools**. All variables were significant predictors of CC_PRO while SCHOOL_NAT was the least significant with p=0.0153 (p<0.05).

CC_PRO = 9.38 + 0.255 (ENG_PRO) + 0.515 (CR_PRO) + 1.042568 (SCHOOL_NAT) for public schools

CC_PRO = 9.38 + 0.255 (ENG_PRO) + 0.515 (CR_PRO) for private schools

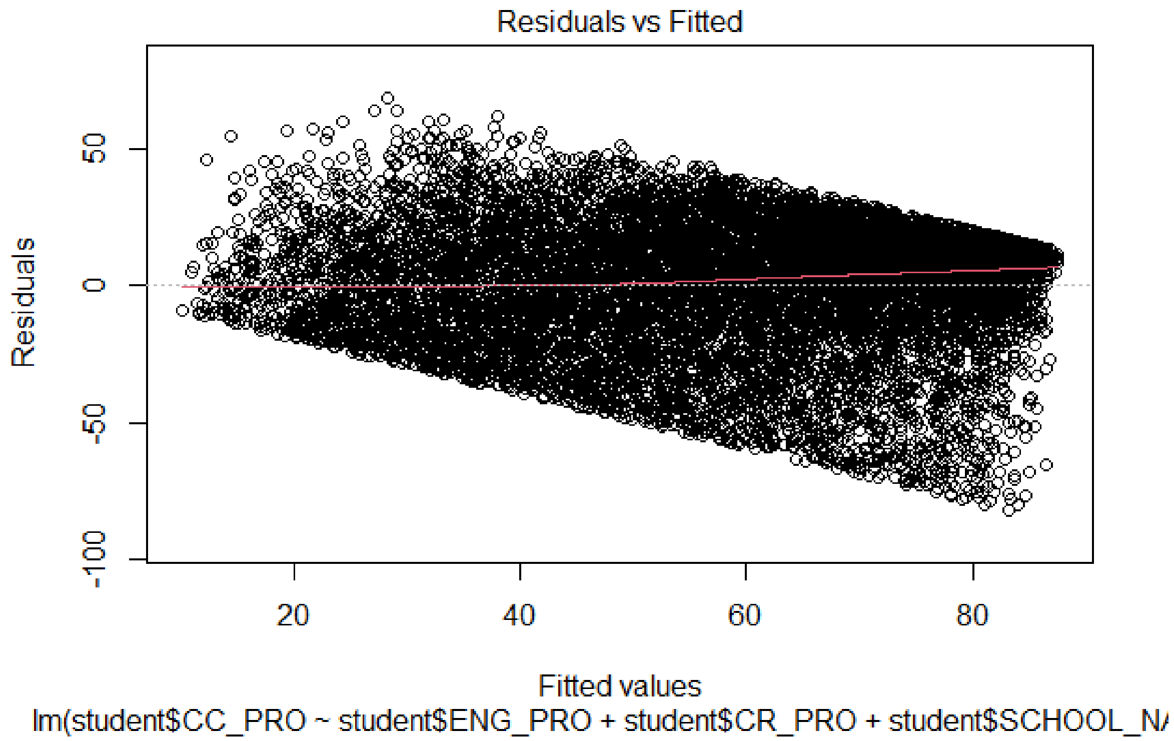
Let's take mean values for ENG_PRO (67.5) and CR_PRO (62.2) and calculate CC_PRO for Public school,

CC_PRO = 9.38 + 0.255 (67.5) + 0.515 (62.2) + 1.042568 (1) = 59.663

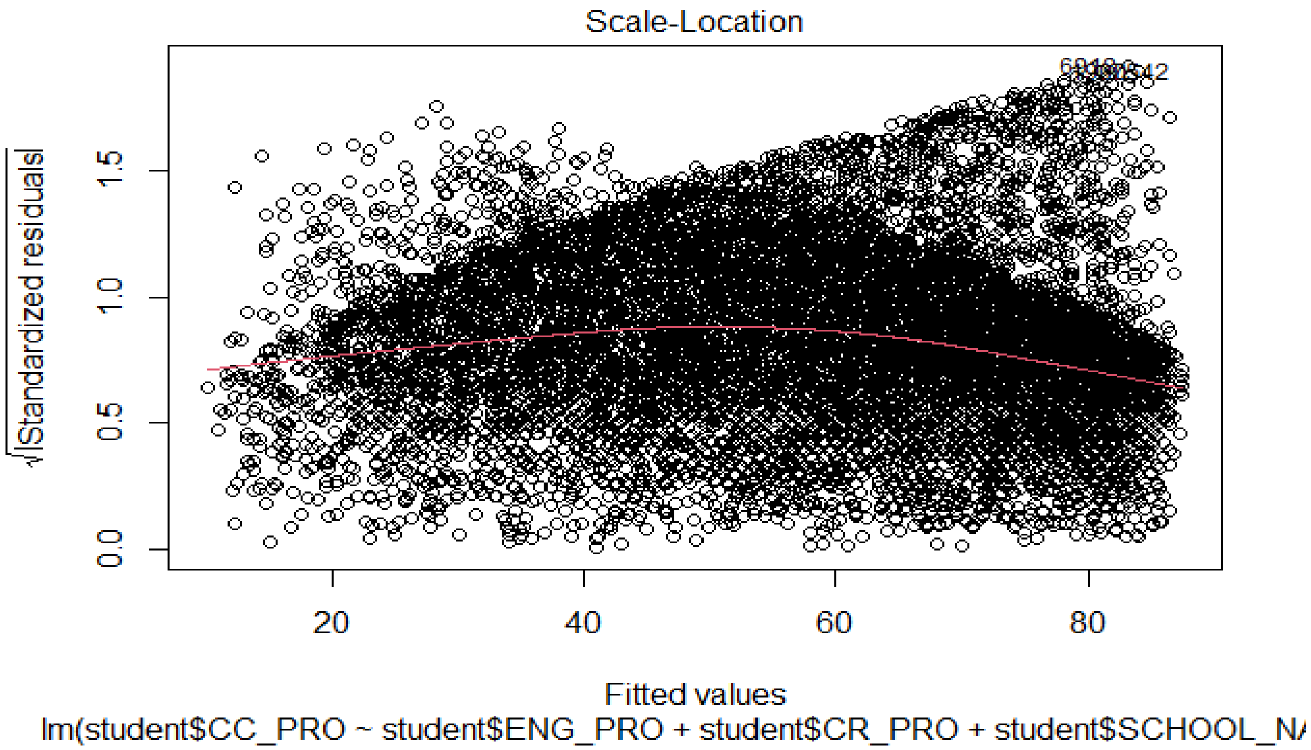
59.663, Which is very close to the mean value of CC_PRO (59.2), this proves that model is significant with very small error term.

Assessing the model against assumptions:

Homoscedasticity



PLOT1



PLOT 2

PLOT1 is the chart of residuals vs fitted values; in PLOT2 the standardised residuals are on the Y axis. If there is absolutely no heteroscedastity, we should see a completely random, equal distribution of points throughout the range of X axis and a flat red line. We really want to see that there is no pattern in the residuals and that they are equally spread around the y = 0 line - the dashed line.

As you can notice the red line is later slightly lifted upward on plot 1 but is not a big problem. Looking at the second plot we can see that while it is a problem it is not huge. Only a concern if there are definite patterns.

Collinearity

Tests to see if the data met the assumption of collinearity indicated that multicollinearity was not a concern (ENG_PRO, Tolerance = 0.6484227, VIF = 1.542204; CR_PRO, Tolerance = 0.7210604 VIF = 1.386846; SCHOOL_NAT, Tolerance = 0.8769922, VIF = 1.140261).

Fit and Goodness

The independent variables explain explains 40.4% (Adjusted R2 = 0.404) variations in the students' CC_PRO scores (competency as a citizen).

Section 3. Model Comparison

Though both the models (Model 1 and Model 2) have the same fit as both have R2=0.404.

The model comparison (ANOVA) results show a Df of 1 (indicating that the more complex model has one additional parameter), and a p-value (p=0.0153, i.e., $p < 0.05$). This means that adding the SCHOOL_NAT to the model2 did lead to a slightly improved fit over the model 1, but not very significant.

Code with Output:

```
> #First unzip the Dataset file and place the xlsx file in the Working directory
> #setting wroking directory
> setwd('C:/Users/toami/OneDrive/Desktop/Data Science/SEM1/Probability & Stats/CA2')
> par(mar=c(1,1,1,1))
> #clearing environment
> rm(list = ls())
> 
> #installing required packages
> needed_packages <- c("pastecs", "ggplot2", "semTools", "FSA","dplyr",
+                      "tidyr","outliers","ggplot2", "readxl")
> not_installed <- needed_packages[!(needed_packages %in% installed.packages()[ , "Package"])]
> if(length(not_installed)) install.packages(not_installed)
> library(pastecs)
> library(semTools)
> 
> #loading library
> library(dplyr)
> library(tidyr)
> library(outliers)
> library(ggplot2)
> library(lm.beta)
> library(stargazer)
> library(readxl)
> 
> #reading xlsx file
> student=read_excel("data_academic_performance.xlsx", sheet = "SABER11_SABERPRO")
New names:
* `` -> ...10
> 
> #DATA CLEANING
> #removing blank column
> student = subset(student, select = -c(...10) )
> 
> #checking for NA values
> student[is.na(student)==TRUE]
<unspecified> [0]
> 
> #Checking for duplicate rows
> which(duplicated(student))
integer(0)
> 
> #removing duplicates
> student=student[!duplicated(student),]
> 
> #variables of interest
> #Nature of School(SCHOOL_NAT)
> #Critical Reading(CR_PRO)
> #Citizen Competencies SPRO(CC_PRO)
> #English(ENG_PRO)
> 
> #summary (CR_PRO)
> summary(student$CR_PRO)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
    1.0    42.0    67.0    62.2    86.0   100.0 
> sd(student$CR_PRO)
[1] 27.66656
> var(student$CR_PRO)
[1] 765.4384
> Outlier = boxplot(student$CR_PRO)$out
> print(unique(Outlier))
numeric(0)
> 
> #summary (ENG_PRO)
> summary(student$ENG_PRO)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
    1.0    51.0    74.0    67.5    88.0   100.0 
> sd(student$ENG_PRO)
[1] 25.4951
> var(student$ENG_PRO)
[1] 649.9999
> Outlier = boxplot(student$ENG_PRO)$out
> print(unique(Outlier))
numeric(0)
> 
> #summary (CC_PRO)
> summary(student$CC_PRO)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   1.00   36.00   65.00   59.19   85.00  100.00 
> sd(student$CC_PRO)
[1] 28.99184
> var(student$CC_PRO)
[1] 840.5268
> Outlier = boxplot(student$CC_PRO)$out
> print(unique(Outlier))
numeric(0)
> 
> #summary SCHOOL_NAT
> summary(student$SCHOOL_NAT)
   Length     Class      Mode 
    12411 character character 
> table(student$SCHOOL_NAT)

```

```
PRIVATE  PUBLIC 
   6565    5846 
> 
> #===============================================================================
> #NORMALITY CHECK OF VARIABLES OF INTEREST
> #NORMALITY TEST of CR_PRO
> gg <- ggplot(student, aes(x=student$CR_PRO))
> gg <- gg + labs(x="CR_PRO")
> gg <- gg + geom_histogram(binwidth=2, colour="black", aes(y=..density.., fill=..count..))
> gg <- gg + scale_fill_gradient("Count", low="#DCDCDC", high="#7C7C7C")
> gg <- gg + stat_function(fun=dnorm, color="red",
+                           args=list(mean=mean(student$CR_PRO, na.rm=TRUE),
+                                     sd=sd(student$CR_PRO, na.rm=TRUE)))
> #Plotting Histogram
> gg
> 
> #Creating qqplot
> qqnorm(student$CR_PRO)
> qqline(student$CR_PRO, col=2)
> 
> #Summary Statistics
> pastecs::stat.desc(student$CR_PRO, basic=F)
      median         mean      SE.mean CI.mean.0.95          var      std.dev 
  67.0000000   62.1993393    0.2483429    0.4867906  765.4384234   27.6665579 
    coef.var 
   0.4448047 
> 
> #skew
> tpskew<-semTools::skew(student$CR_PRO)
> tpskew[1]/tpskew[2]
skew (g1) 
-23.52302 
> 
> #kurtosis
> tpkurt<-semTools::kurtosis(student$CR_PRO)
> tpkurt[1]/tpkurt[2]
Excess Kur (g2) 
      -19.24846 
> 
> ztpCR_PRO<- abs(scale(student$CR_PRO))
> 
> #how much data fall outside 95% of region
> FSA::perc(as.numeric(ztpCR_PRO), 1.96, "gt")
[1] 3.658045
> 
> #how much data fall outside 99.9% of region
> FSA::perc(as.numeric(ztpCR_PRO), 3.29, "gt")
[1] 0
> 
> #-------------------------------------------------------------------------------
> #NORMALITY TEST of ENG_PRO
> gg <- ggplot(student, aes(x=student$CR_PRO))
> gg <- gg + labs(x="ENG_PRO")
> gg <- gg + geom_histogram(binwidth=2, colour="black", aes(y=..density.., fill=..count..))
> gg <- gg + scale_fill_gradient("Count", low="#DCDCDC", high="#7C7C7C")
> gg <- gg + stat_function(fun=dnorm, color="red",
+                           args=list(mean=mean(student$ENG_PRO, na.rm=TRUE),
+                                     sd=sd(student$ENG_PRO, na.rm=TRUE)))
> #Plotting Histogram
> gg
> 
> #Creating qqplot
> qqnorm(student$ENG_PRO)
> qqline(student$ENG_PRO, col=2)
> 
> #Summary Statistics
> pastecs::stat.desc(student$ENG_PRO, basic=F)
      median         mean      SE.mean CI.mean.0.95          var      std.dev 
  74.0000000   67.4983482    0.2288512    0.4485839  649.9998965   25.4950955 
    coef.var 
   0.3777144
> 
> #skew
> tpskew<-semTools::skew(student$ENG_PRO)
> tpskew[1]/tpskew[2]
skew (g1) 
-36.31304 
> 
> #kurtosis
> tpkurt<-semTools::kurtosis(student$ENG_PRO)
> tpkurt[1]/tpkurt[2]
Excess Kur (g2) 
      -7.222944 
> 
> ztpENG_PRO<- abs(scale(student$ENG_PRO))
> 
> #how much data fall outside 95% of region
> FSA::perc(as.numeric(ztpENG_PRO), 1.96, "gt")
[1] 5.752961
> 
> #how much data fall outside 99.9% of region
> FSA::perc(as.numeric(ztpENG_PRO), 3.29, "gt")
[1] 0
> 
> #------------------------------------------------------------------------------
> #NORMALITY TEST of CC_PRO
> gg <- ggplot(student, aes(x=student$CR_PRO))
> gg <- gg + labs(x="CC_PRO")
> gg <- gg + geom_histogram(binwidth=2, colour="black", aes(y=..density.., fill=..count..))
> gg <- gg + scale_fill_gradient("Count", low="#DCDCDC", high="#7C7C7C")
> gg <- gg + stat_function(fun=dnorm, color="red",
+                           args=list(mean=mean(student$CC_PRO, na.rm=TRUE),
+                                     sd=sd(student$CC_PRO, na.rm=TRUE)))
> #Plotting Histogram
> gg
> 
> #Creating qqplot
> qqnorm(student$CC_PRO)
> qqline(student$CC_PRO, col=2)
> 
> #Summary Statistics
> pastecs::stat.desc(student$CC_PRO, basic=F)
      median         mean      SE.mean CI.mean.0.95          var      std.dev 
  65.0000000   59.1867698    0.2602390    0.5101088  840.5267581   28.9918395 
    coef.var 
   0.4898365 
> 
> #skew
> tpskew<-semTools::skew(student$CC_PRO)
> tpskew[1]/tpskew[2]
skew (g1) 
-19.08356 
> 
> #kurtosis
> tpkurt<-semTools::kurtosis(student$CC_PRO)
> tpkurt[1]/tpkurt[2]
Excess Kur (g2) 
      -24.02213 
> 
> ztpCC_PRO<- abs(scale(student$CC_PRO))
> 
> #how much data fall outside 95% of region
> FSA::perc(as.numeric(ztpCC_PRO), 1.96, "gt")
[1] 1.232777
> 
> #how much data fall outside 99.9% of region
> FSA::perc(as.numeric(ztpCC_PRO), 3.29, "gt")
[1] 0
> #============================================================================
> #CORRELATION CHECK AMONG VARIABLES OF INTEREST
> #correlation between CR_PRO and CC_PRO
> #Scatterplot
> scatter <- ggplot(student, aes(CR_PRO, CC_PRO))
> scatter + geom_point() + geom_smooth(method = "lm", colour = "Red", se = F) + labs(x = "CR_PRO", y = "CC_PRO")
`geom_smooth()` using formula 'y ~ x'
>
> #Pearson Correlation
> stats::cor.test(student$CR_PRO, student$CC_PRO, method='pearson')

          Pearson's product-moment correlation

data:  student$CR_PRO and student$CC_PRO
t = 85.287, df = 12409, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5966959 0.6188821
sample estimates:
      cor
0.6079076

>
> #correlation between ENG_PRO and CC_PRO
> #Scatterplot
> scatter <- ggplot(student, aes(ENG_PRO, CC_PRO))
> scatter + geom_point() + geom_smooth(method = "lm", colour = "Red", se = F) + labs(x = "ENG_PRO", y = "CC_PRO")
`geom_smooth()` using formula 'y ~ x'
>
> #Pearson Correlation
> stats::cor.test(student$ENG_PRO, student$CC_PRO, method='pearson')

          Pearson's product-moment correlation

data:  student$ENG_PRO and student$CC_PRO
t = 60.561, df = 12409, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.4639377 0.4910993
sample estimates:
      cor
0.4776327

>
> #correlation between ENG_PRO and CR_PRO
> #Scatterplot
> scatter <- ggplot(student, aes(ENG_PRO, CR_PRO))
> scatter + geom_point() + geom_smooth(method = "lm", colour = "Red", se = F) + labs(x = "ENG_PRO", y = "CR_PRO")
`geom_smooth()` using formula 'y ~ x'
>
> #Pearson Correlation
> stats::cor.test(student$ENG_PRO, student$CR_PRO, method='pearson')

          Pearson's product-moment correlation

data:  student$ENG_PRO and student$CR_PRO
t = 69.092, df = 12409, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5142623 0.5396758
sample estimates:
      cor
0.5270869

>
> #============================================================================================
> #MULTIPLE REGRESSION MODEL 1
> #building multiple regression model
> model1<-lm(student$CC_PRO~student$ENG_PRO+student$CR_PRO)
```

```
>
> #Analysis of Variance Table
> anova(model1)
Analysis of Variance Table

Response: student$CC_PRO
                   Df  Sum Sq Mean Sq F value    Pr(>F)
student$ENG_PRO     1 2379641 2379641  4747.7 < 2.2e-16 ***
student$CR_PRO      1 1832117 1832117  3655.3 < 2.2e-16 ***
Residuals       12408 6219179     501
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> #plots
> plot(model1,1)
> plot(model1,2)
> plot(model1,3)
>
> #Calculate Collinearity
> vifmodel1<-car::vif(model1)
> vifmodel1
student$ENG_PRO  student$CR_PRO
       1.384697        1.384697
>
> #Calculate tolerance
> 1/vifmodel1
student$ENG_PRO  student$CR_PRO
      0.7221794       0.7221794
>
> #summary
> summary(model1)

Call:
lm(formula = student$CC_PRO ~ student$ENG_PRO + student$CR_PRO)

Residuals:
    Min      1Q  Median      3Q     Max
-82.506 -13.001   4.772  15.257  68.304

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     10.333704   0.603445   17.12   <2e-16 ***
student$ENG_PRO  0.247549   0.009276   26.69   <2e-16 ***
student$CR_PRO   0.516789   0.008548   60.46   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 22.39 on 12408 degrees of freedom
Multiple R-squared:  0.4038, Adjusted R-squared:  0.4037
F-statistic:  4201 on 2 and 12408 DF,  p-value: < 2.2e-16

>
> #coefficients
>  #Will allow us to isolate the beta co-efficients
> lm.beta(model1)

Call:
lm(formula = student$CC_PRO ~ student$ENG_PRO + student$CR_PRO)

Standardized Coefficients::
    (Intercept) student$ENG_PRO  student$CR_PRO
      0.0000000       0.2176918       0.4931652

>
> #Tidy output of all the required stats
> #For formatting outputs/tables
> stargazer(model1, type="text")

================================================
                        Dependent variable:
```

```
                    ----------------------------
                               CC_PRO
------------------------------------------------
ENG_PRO                       0.248***
                              (0.009)

CR_PRO                        0.517***
                              (0.009)

Constant                     10.334***
                              (0.603)

------------------------------------------------
Observations                   12,411
R2                             0.404
Adjusted R2                    0.404
Residual Std. Error     22.388 (df = 12408)
F Statistic         4,201.478*** (df = 2; 12408)
================================================
Note:                *p<0.1; **p<0.05; ***p<0.01
>
> #==============================================================================
> #MULTIPLE REGRESSION MODEL 2
> #building multiple regression model
> model2<-lm(student$CC_PRO~student$ENG_PRO+student$CR_PRO+student$SCHOOL_NAT)
>
> #Analysis of Variance Table
> anova(model2)
Analysis of Variance Table

Response: student$CC_PRO
                      Df  Sum Sq Mean Sq   F value Pr(>F)
student$ENG_PRO        1 2379641 2379641 4749.5335 <2e-16 ***
student$CR_PRO         1 1832117 1832117 3656.7292 <2e-16 ***
student$SCHOOL_NAT     1    2948    2948    5.8834 0.0153 *
Residuals          12407 6216232     501
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> #plots
> plot(model2,1)
> plot(model2,2)
> plot(model2,3)
>
> #Calculate Collinearity
> vifmodel2<-car::vif(model2)
> vifmodel2
   student$ENG_PRO     student$CR_PRO student$SCHOOL_NAT
          1.542204           1.386846           1.140261
>
> #Calculate tolerance
> 1/vifmodel2
   student$ENG_PRO     student$CR_PRO student$SCHOOL_NAT
         0.6484227          0.7210604          0.8769922
>
> #summary
> summary(model2)

Call:
lm(formula = student$CC_PRO ~ student$ENG_PRO + student$CR_PRO +
    student$SCHOOL_NAT)

Residuals:
    Min      1Q  Median      3Q     Max
-82.164 -12.929   4.782  15.292  68.727

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)              9.381324   0.719839  13.033   <2e-16 ***
student$ENG_PRO          0.255136   0.009787  26.068   <2e-16 ***
```

```
student$CR_PRO           0.515972   0.008553  60.329   <2e-16 ***
student$SCHOOL_NATPUBLIC 1.042568   0.429822   2.426   0.0153 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 22.38 on 12407 degrees of freedom
Multiple R-squared:  0.4041,	Adjusted R-squared:  0.4039 
F-statistic:  2804 on 3 and 12407 DF,  p-value: < 2.2e-16

> 
> #coefficients
> lm.beta(model2)

Call:
lm(formula = student$CC_PRO ~ student$ENG_PRO + student$CR_PRO + 
    student$SCHOOL_NAT)

Standardized Coefficients::
             (Intercept)          student$ENG_PRO           student$CR_PRO 
              0.00000000               0.22436344               0.49238586 
student$SCHOOL_NATPUBLIC 
              0.01795089 

> 
> #Tidy output of all the required stats
> stargazer(model2, type="text")

================================================
                        Dependent variable:    
                    ---------------------------
                              CC_PRO           
------------------------------------------------
ENG_PRO                      0.255***          
                             (0.010)           
                                               
CR_PRO                       0.516***          
                             (0.009)           
                                               
SCHOOL_NATPUBLIC             1.043**           
                             (0.430)           
                                               
Constant                     9.381***          
                             (0.720)           
                                               
------------------------------------------------
Observations                  12,411           
R2                             0.404           
Adjusted R2                    0.404           
Residual Std. Error     22.384 (df = 12407)    
F Statistic         2,804.049*** (df = 3; 12407)
================================================
Note:                *p<0.1; **p<0.05; ***p<0.01
> #===============================================================================
> #MODEL COMPARISON
> anova(model1,model2)
Analysis of Variance Table

Model 1: student$CC_PRO ~ student$ENG_PRO + student$CR_PRO
Model 2: student$CC_PRO ~ student$ENG_PRO + student$CR_PRO + student$SCHOOL_NAT
  Res.Df     RSS Df Sum of Sq      F Pr(>F)  
1  12408 6219179                              
2  12407 6216232  1    2947.8 5.8834 0.0153 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> #END
> 
```