World University Ranking Analysis

By

Bhavya Kavuri

Dhruva Kiran Veeravalli

Aditya Gidda

Jaswanth Das Gunturu

**Table of Contents**

[Introduction 3](#_Toc500693944)

[Statistical Analysis 5](#_Toc500693945)

[Examining Distributions 5](#_Toc500693946)

[Histograms 6](#_Toc500693947)

[Relationships 7](#_Toc500693948)

[Multiple Linear Regression 8](#_Toc500693949)

[Missing data Analysis 9](#_Toc500693950)

[Correlation Plot for the newly imputed dataset 10](#_Toc500693951)

[Predictions for data 18](#_Toc500693952)

[Assumptions of Multiple Linear Regression: 19](#_Toc500693953)

[Linearity 19](#_Toc500693954)

[Equality 19](#_Toc500693955)

[Independence 20](#_Toc500693956)

[Durbin Watson Statistic 20](#_Toc500693957)

[Multi Collinearity 22](#_Toc500693958)

[Residuals 22](#_Toc500693959)

[Ordinal Regression: 24](#_Toc500693960)

# Introduction

The dataset considered for the project is World University Rankings by Times Higher Education (THE) as it is one of the most influential international university ranking systems. This dataset contains 800 rankings of universities for the year 2016. Since the dataset also contains rankings with groups, we consider the first 200 rankings for the analysis. The source has been obtained from timeshighereducation.com. The complete dataset ca be found at : <https://goo.gl/gDsdZL>

The dataset contains two types of variables: categorical and quantitative.

The categorical variables in our data are:

|  |  |  |
| --- | --- | --- |
| university\_name | name of university | String |
| country | country of each university | String |
| international\_students | Percentage of students who are international | String |
| female\_male\_ratio | Female student to Male student ratio | String |

Similarly, the quantitative variables are:

|  |  |  |
| --- | --- | --- |
| **Column** | **Metadata** | **Datatype** |
| world\_rank | world rank for the university. Contains rank ranges and equal ranks (eg. =94 and 201-250) | Numeric |
| Teaching | university score for teaching (the learning environment) | Numeric |
| international | university score international outlook (staff, students, research) | Numeric |
| research | university score for research (volume, income and reputation) | Numeric |
| citations | university score for citations (research influence) | Numeric |
| income | university score for industry income (knowledge transfer) | Numeric |
| total\_score | total score for university, used to determine rank | Numeric |
| num\_students | number of students at the university | Numeric |
| student\_staff\_ratio | Number of students divided by number of staff | Numeric |

In our analysis with regression, we try to answer the questions regarding the variables influencing the rank of each university in the world. We train the first 150 records for the regression analysis and predict ranks for the remaining 50 records.

# 

# Statistical Analysis

## Examining Distributions

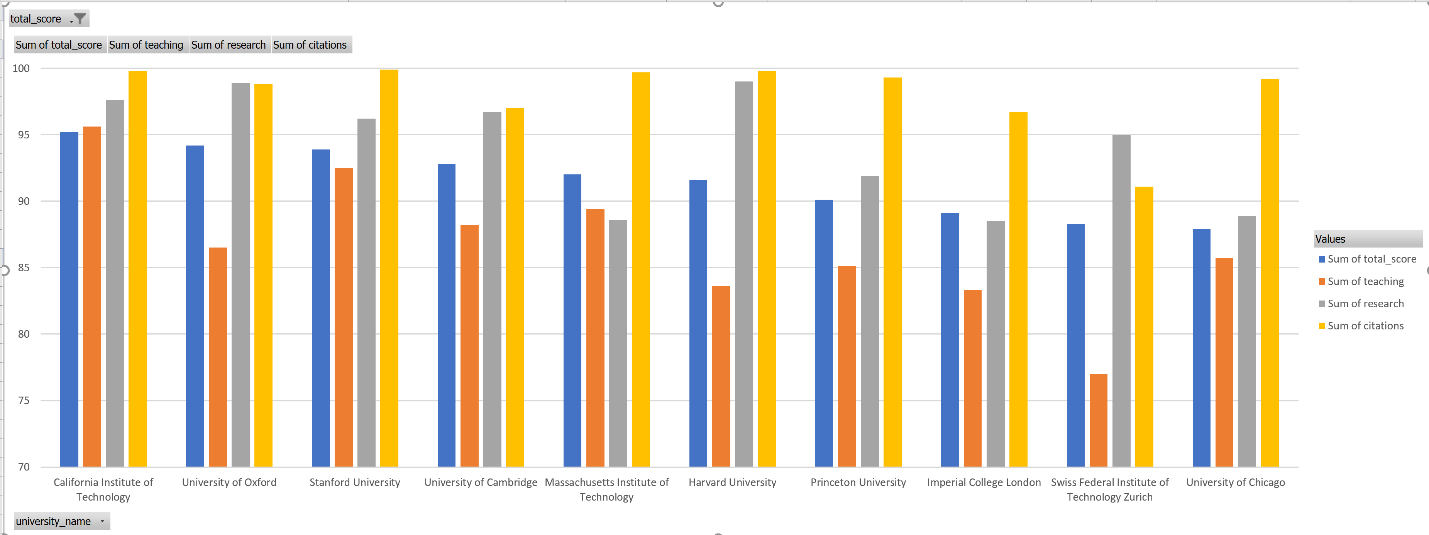
|  |  |  |
| --- | --- | --- |
| Description: C:\Users\Aditya chintu\Desktop\Boxplots\citations_dens_scatter_box.png | Description: C:\Users\Aditya chintu\Desktop\Boxplots\Female_Students_dens_scatter_box.png | Description: C:\Users\Aditya chintu\Desktop\Boxplots\income_dens_scatter_box.png |
| Description: C:\Users\Aditya chintu\Desktop\Boxplots\international_dens_scatter_box.png | Description: C:\Users\Aditya chintu\Desktop\Boxplots\international_students_dens_scatter_box.png | Description: C:\Users\Aditya chintu\Desktop\Boxplots\num_students_dens_scatter_box.png |
| Description: C:\Users\Aditya chintu\Desktop\Boxplots\research_dens_scatter_box.png | Description: C:\Users\Aditya chintu\Desktop\Boxplots\student_staff_ratio_dens_scatter_box.png | Description: C:\Users\Aditya chintu\Desktop\Boxplots\teaching_dens_scatter_box.png |
| Description: C:\Users\Aditya chintu\Desktop\Boxplots\total_score_dens_scatter_box.png | Description: C:\Users\Aditya chintu\Desktop\Boxplots\World_rank_dens_scatter_box.png |

## Histograms

A screenshot of a map

Description generated with very high confidence

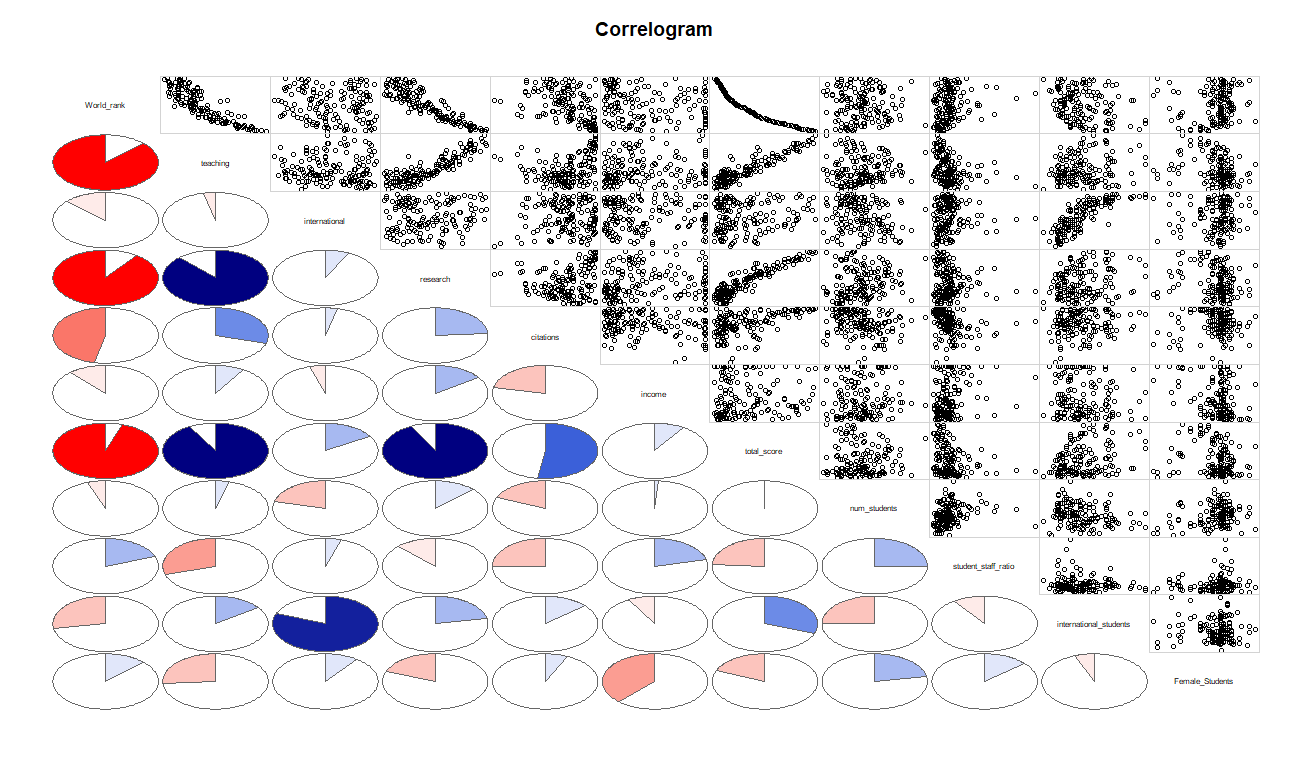
Analysis: The above graph has been plotted to analyze which of the world countries have the most number of universities in the top 200 universities. And we’ve concluded that United States of America has the highest number of Universities in the top 200 Universities of the world followed by United Kingdom then by Germany and Netherlands.



Observation: In the above graph, total score, teaching, research, citations have been plotted. If we take top 10 universities and see the top five and below five as two parts, we can see that some values in the below 5 universities have greater values when compared to the top 5 universities. This states that the total score doesn’t depend only on a single variable but depends on combination of variables.

## Relationships

We can observe from the below correlogram that,



* **international\_students** are highly correlated to **international outlook**.
* The highest correlation is between **teaching** and **research**.
* **World\_rank** and **Total\_score** is each highly correlated to Teaching, research and citations.

# Multiple Linear Regression

After the examination of the relationship between different variables, we could see that the variables are linearly related. Therefore, we would like to build a multiple linear regression model which predicts the ranks of the university based on the ratings and other variables.

Using R, we calculate the regression model as follows,

We will be installing the following packages

install.packages("mice")

install.packages("VIM")

install.packages("corrgram")

install.packages("corrplot")

install.packages("ggplot2")

install.packages(“outliers”)

install.packages("car")

Below are the libraries which are loaded

library("mice")

library("VIM")

library("corrgram")

library("corrplot")

library("outliers")

library("car")

library("ggplot2")

Uploading the data

getwd()

setwd("C:/Users/KAVURI/Desktop/Stats/project")

ur = read.csv("Rankings\_New.csv", header=T)

Change columns to numeric.

ur$income = sub('-',' ',ur$income)

ur$income = as.numeric(as.character(ur$income))

ur$num\_students = gsub(',','',ur$num\_students)

ur$num\_students = as.numeric(as.character(ur$num\_students))

ur$international\_students = as.numeric(as.character(gsub('%','',ur$international\_students)))

colnames(ur)[colnames(ur)=="X.\_Female\_Students"] <- "Female\_Students"

colnames(ur)[colnames(ur)=="ï..world\_rank"] <- "World\_rank"

## Missing data Analysis

pMiss <- function(x){sum(is.na(x))/length(x)\*100}

apply(ur,1,pMiss)

apply(ur,2,pMiss)

We can see in the below image that Female\_Students missing data is 12.5% which is above 5%.

A screenshot of a social media post

Description generated with very high confidence

The below picture gives a clear image of missing data.

md.pattern(ur)

A picture containing screenshot

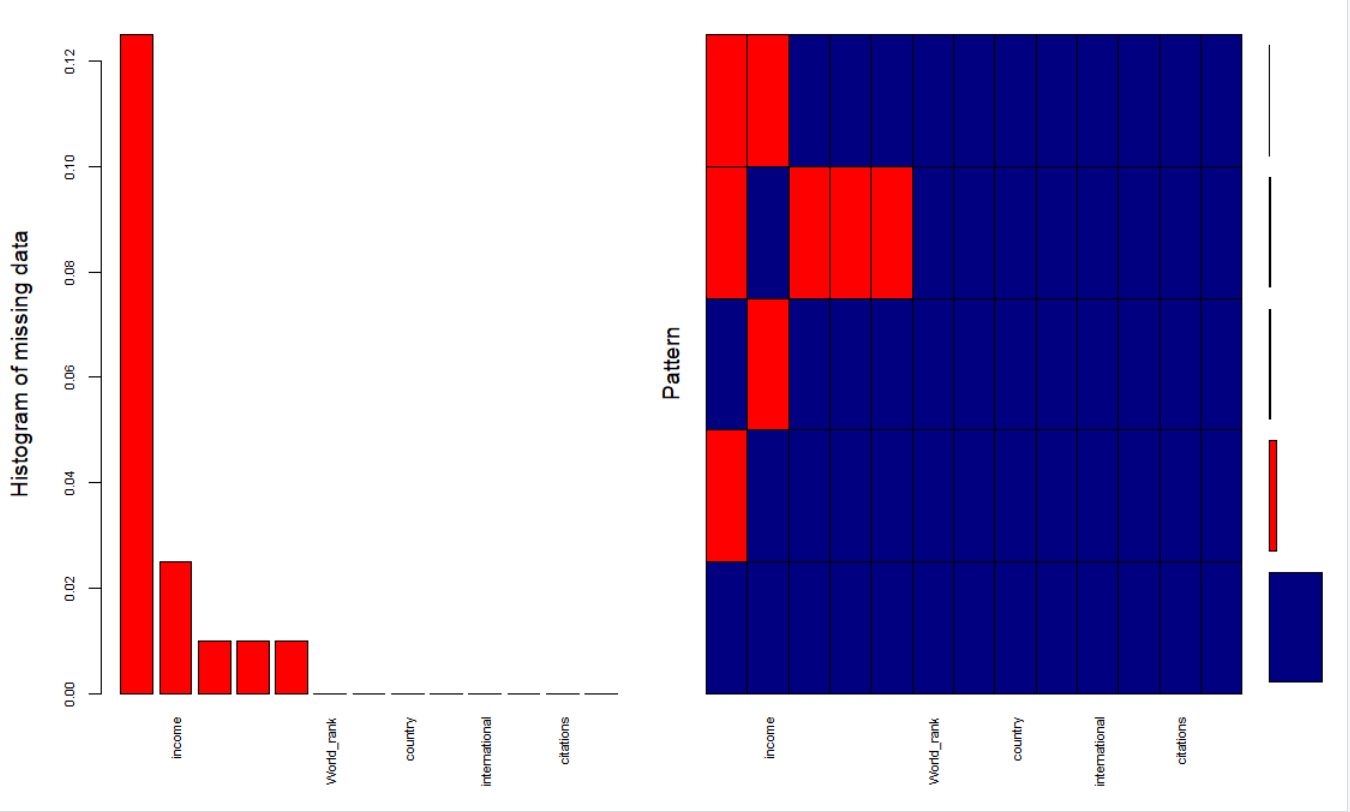
Description generated with very high confidence

Check missing values again

sapply(ur, function(x) sum(is.na(x)))

The graph below shows that majority of the values missing are of income which is above considerable level.

aggr\_plot <- aggr(ur, col=c('navyblue','red'), numbers=TRUE, sortVars=TRUE, labels=names(ur), cex.axis=.7, gap=3, ylab=c("Histogram of missing data","Pattern"))



The mice package in R helps to impute missing data in our analysis,

tempData <- mice(ur,m=5,meth='cart')

sapply(training\_set, function(x) sum(is.na(x)))

modelFit2 <- with(tempData, lm(total\_score~research+citations+international+income))

summary(modelFit2)

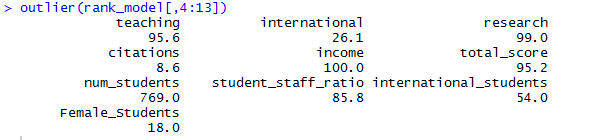
train\_complete <- complete(tempData,"long")

training\_set = train\_complete[(1:150),]

test\_set = train\_complete[(150:200),]

## Outlier Detection

outlier(rank\_model[,4:13])

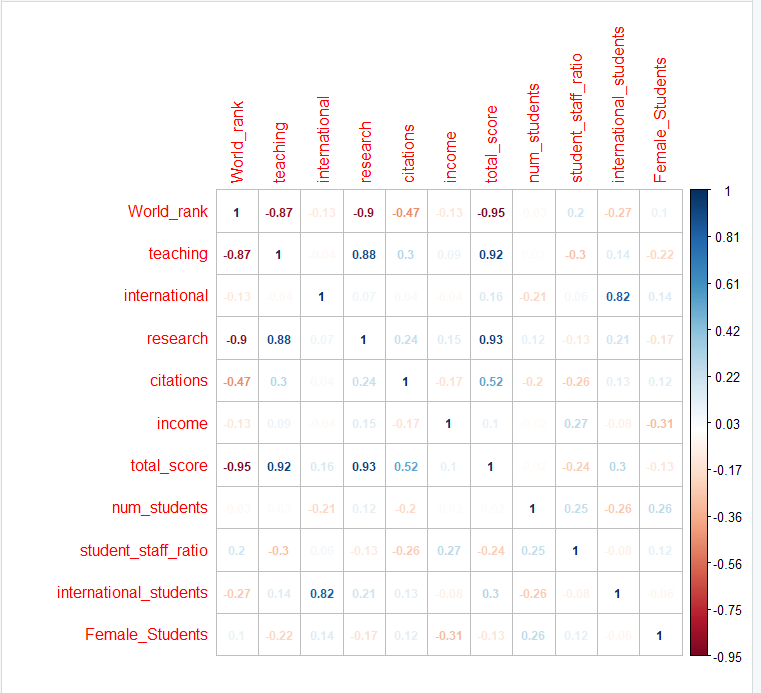


The outliers do not have significant influence over the model. Hence we do not delete the outliers.

## Correlation Plot for the newly imputed dataset

q = as.matrix(training\_set[,c(3,6:15)])

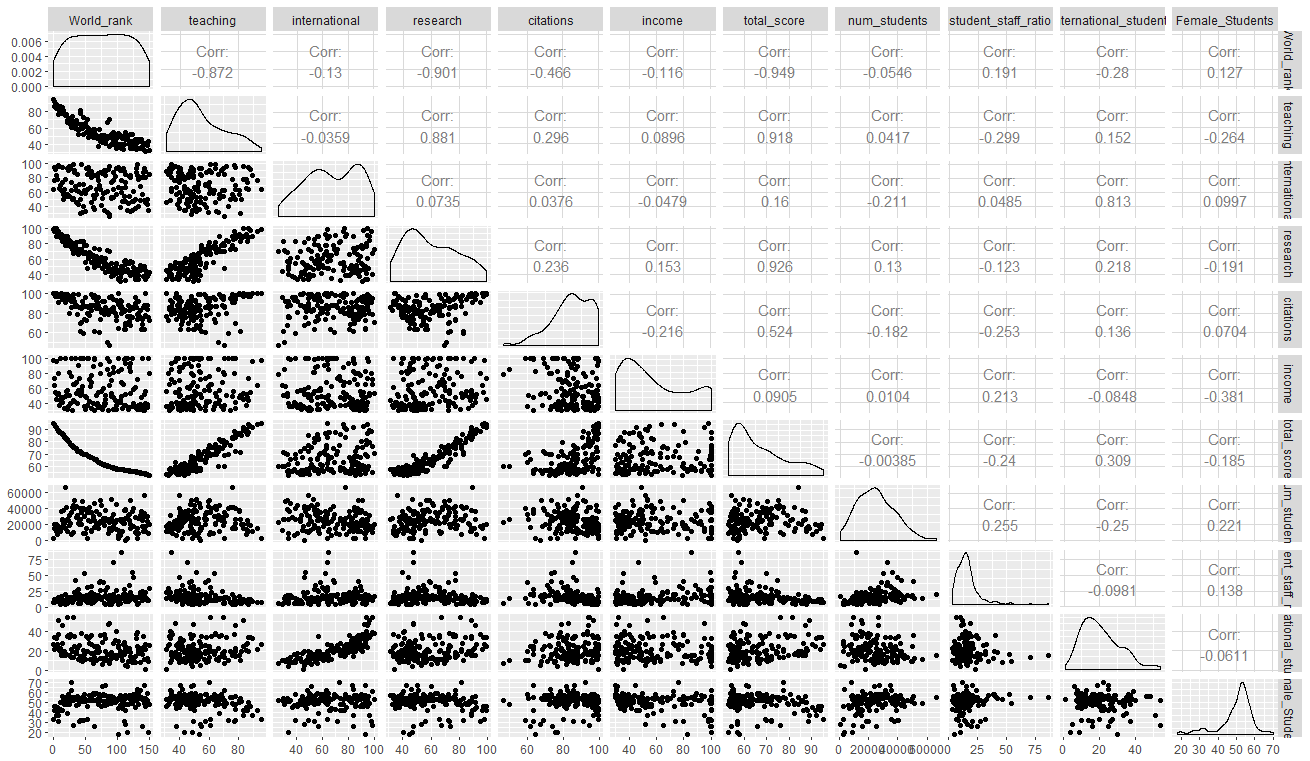
corrplot(cor(q), method = "number", number.cex=0.75, is.corr = FALSE)



par(mfrow=c(2,2))

plot(training\_set, col="blue", main="Matrix Scatterplot of all the independent variables")

ggpairs(training\_set[,c(3,6:15)])



rank\_model = lm(World\_rank~research+citations+international+teaching+total\_score+Female\_Students+student\_staff\_ratio+international\_students+num\_students+income, data=training\_set)

summary(rank\_model)

A screenshot of a cell phone

Description generated with very high confidence

For the above model, the adjusted R-square is 90% which is good but none of the variables are significant. Hence, we continue to remove the insignificant variables and only include relevant variables and the model is as follows,

rank\_model = lm(World\_rank~research+citations+international+income, data=training\_set)

summary(rank\_model)

A screenshot of a cell phone

Description generated with very high confidence

The adjusted R-square is 88% which is less than the previous model, but all the variable are significant in this model. But for ranks, we are unable to predict the ranks with the above model, we consider to do with the total\_score which is correlated to the world\_rank. Higher the total\_score, higher the rank.

The following model is with total\_score as independent variable and some of the independent variables.

rank\_model = lm(total\_score~research+international+Female\_Students+student\_staff\_ratio, data=training\_set)

summary(rank\_model)

A screenshot of a cell phone

Description generated with very high confidence

The adjusted R-square is about 88% and we try to improve the model by including the rankings of research, citation, international\_outlook and income.

rank\_model = lm(total\_score~research+citations+international+income, data=training\_set)

summary(rank\_model)

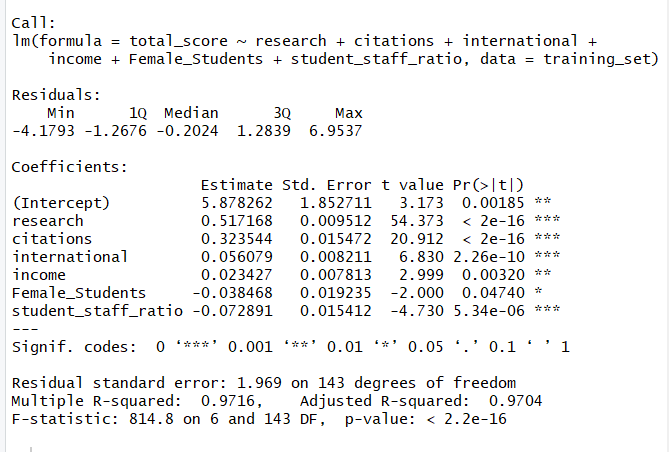
A screenshot of a cell phone

Description generated with very high confidence

The adjusted R-square is about 96.43% and we still try to improve the model.

rank\_model = lm(total\_score~research+citations+international+income+Female\_students+student\_staff\_ratio, data=training\_set)

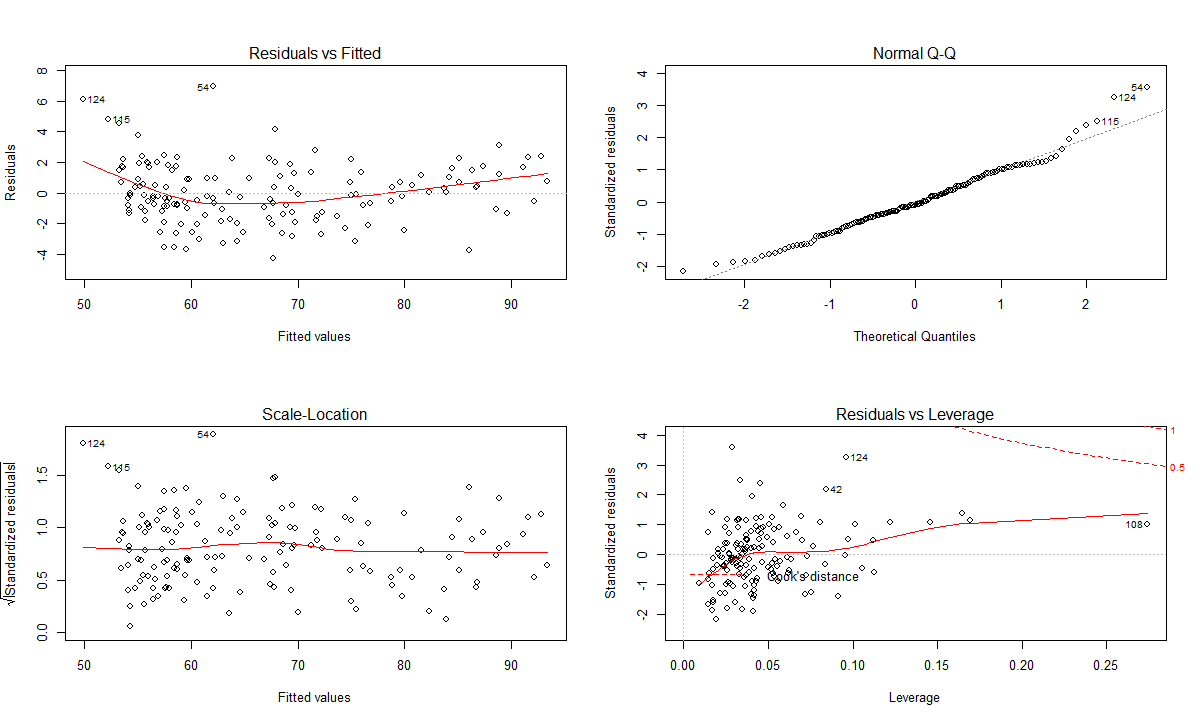
summary(rank\_model)



The adjusted R-square is 97% which is pretty good and all the variables are significant.

par(mfrow=c(2,2))

plot(rank\_model)



# Predictions for data

predict (rank\_model, test\_set)

A screenshot of a cell phone

Description generated with high confidence

# 

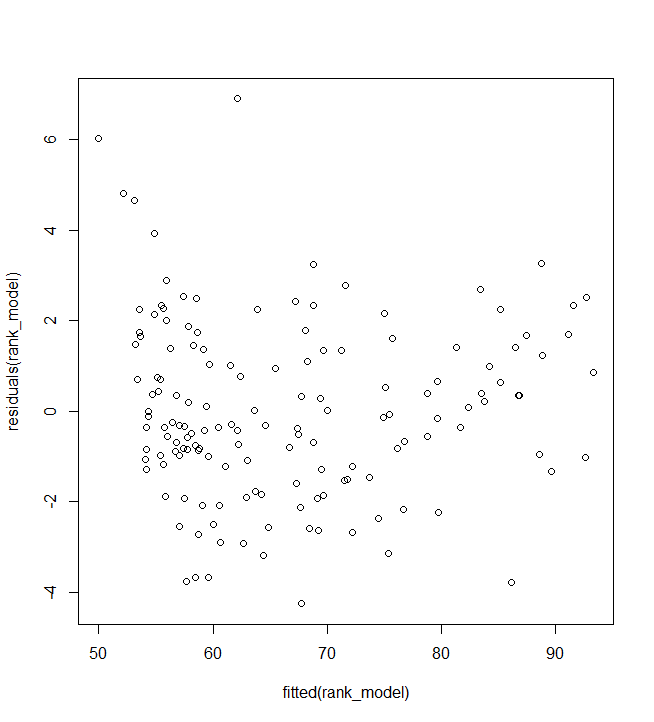
# The predicted values are close to the actual values. Hence, we conclude that our model is significant.

# Assumptions of Multiple Linear Regression:

## Linearity

plot(fitted(rank\_model), residuals(rank\_model))

abline(0,1, col="blue", lwd=2)



## Equality

plot(fitted(rank\_model), residuals(rank\_model))

abline(0,1, col="blue", lwd=2)

By executing the above commands, we know that our dataset that using the below plot that it is linear, independent and residuals have a constant variance.

A picture containing photo

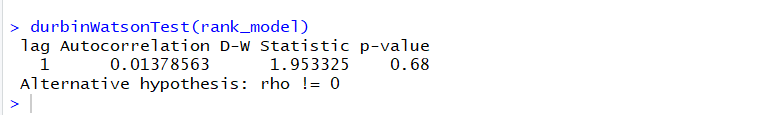
Description generated with high confidence

## Independence

### Durbin Watson Statistic

durbinWatsonTest(rank\_model)

The durbinwatson statistic value should be close to 2, for our data set it is around 1.95 which clearly states that autocorrelation does not exist.



##Normality

hist(rank\_model$residuals, main ="Histogram of residuals")

qqPlot(rank\_model, main="QQ Plot")

From the below graphs, we state that the standardized residuals approximately display in a straight line.

A screenshot of a cell phone

Description generated with high confidence

A close up of a map

Description generated with high confidence

## 

## Multi Collinearity

vif(rank\_model)

All the values exhibit below 5. By this, we can say there is no collinearity between predictor variables.

A close up of a logo

Description generated with high confidence

## Residuals

All the residuals plots are normal and there are no deviations in our model.

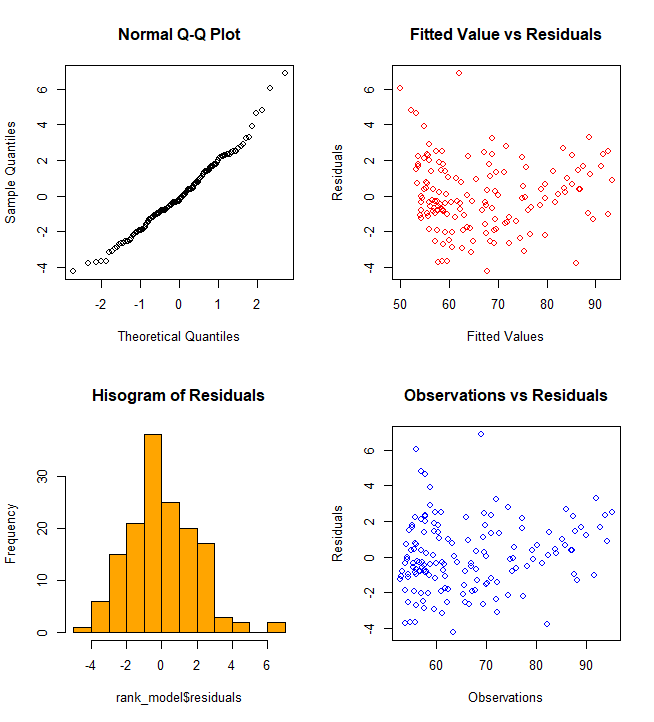
layout(matrix(c(1,2,3,4), 2, 2, byrow = TRUE))

qqnorm(rank\_model$residuals)

plot(rank\_model$fitted.values,rank\_model$residuals,main="Fitted Value vs Residuals",xlab="Fitted Values",ylab="Residuals",col="red")

hist(rank\_model$residuals,col="orange",main = "Histogram of Residuals")

plot(training\_set$total\_score,rank\_model$residuals,main="Observations vs Residuals",xlab="Observations",ylab="Residuals",col="blue")



# 

# Ordinal Regression:

Even though we can predict the rank of a university using the total score, we would like to predict the ranks using the ranks given with the help of ordinal regression.

We wanted to try with the first 30 ranks and check the regression model,

library(“ordinal”)

library(“MASS”)

str(ur)

training\_set = train\_complete[(1:30),]

The below code is to order the ranking variable which is originally a numeric variable and needed to be a factor variable in order to perform the ordinal regression.

training\_set = training\_set[,c(3,6:15)]

training\_set$World\_rank = as.factor (training\_set$World\_rank)

training\_set$World\_rank <- ordered (training\_set$World\_rank, levels = c(30:1))

model <- polr(World\_rank~research+citations+income,training\_set, Hess = TRUE)

summary(model)

> summary(model)

Call:

polr(formula = World\_rank ~ research + citations + income, data = training\_set,

Hess = TRUE)

Coefficients:

Value Std. Error t value

research 0.57934 0.10771 5.379

citations 0.36796 0.09944 3.700

income 0.02419 0.01599 1.513

Intercepts:

Value Std. Error t value

30|29 77.4651 14.9765 5.1724

29|28 78.8236 15.0610 5.2336

28|27 79.6216 15.1294 5.2627

27|26 80.2198 15.1634 5.2903

26|25 81.0124 15.3014 5.2944

25|24 81.7855 15.4520 5.2929

24|23 82.5162 15.5620 5.3024

23|22 83.5782 15.8263 5.2810

22|21 84.4474 16.0192 5.2716

21|20 84.9727 16.0816 5.2838

20|19 85.5279 16.1640 5.2913

19|18 86.1231 16.2751 5.2917

18|17 86.6697 16.3787 5.2916

17|16 87.1611 16.4585 5.2958

16|15 87.5576 16.5017 5.3060

15|14 87.9837 16.5496 5.3164

14|13 88.4896 16.6238 5.3231

13|12 88.9160 16.6736 5.3327

12|11 89.3297 16.7177 5.3434

11|10 89.7448 16.7602 5.3546

10|9 90.1490 16.7944 5.3678

9|8 90.5882 16.8390 5.3797

8|7 91.1363 16.9221 5.3856

7|6 91.8871 17.0666 5.3840

6|5 92.5096 17.1521 5.3935

5|4 93.1183 17.2250 5.4060

4|3 93.9793 17.3641 5.4123

3|2 95.0455 17.5277 5.4226

2|1 96.3316 17.6562 5.4560

Residual Deviance: 141.4078

AIC: 205.4078

The table below gives the coefficients of all independent variables and ranks.

(ctable = coef(summary(model)))

> (ctable = coef(summary(model)))

Value Std. Error t value

research 0.57934098 0.10770924 5.378749

citations 0.36795561 0.09944030 3.700266

income 0.02418672 0.01598601 1.512993

30|29 77.46514443 14.97649897 5.172447

29|28 78.82362978 15.06101656 5.233619

28|27 s79.62157446 15.12943055 5.262695

27|26 80.21979553 15.16343395 5.290345

26|25 81.01243409 15.30141477 5.294441

25|24 81.78553395 15.45195472 5.292892

24|23 82.51623923 15.56203714 5.302406

23|22 83.57823988 15.82633101 5.280961

22|21 84.44739770 16.01918024 5.271643

21|20 84.97270662 16.08164979 5.283830

20|19 85.52790256 16.16395976 5.291272

19|18 86.12305974 16.27513585 5.291695

18|17 86.66966907 16.37872935 5.291599

17|16 87.16105442 16.45854024 5.295795

16|15 87.55762898 16.50167524 5.305984

15|14 87.98367637 16.54957373 5.316371

14|13 88.48959244 16.62379244 5.323069

13|12 88.91602362 16.67358662 5.332747

12|11 89.32971497 16.71768829 5.343425

11|10 89.74476181 16.76022811 5.354627

10|9 90.14904569 16.79439402 5.367806

9|8 90.58824078 16.83900873 5.379666

8|7 91.13631326 16.92208495 5.385643

7|6 91.88706739 17.06656295 5.384041

6|5 92.50963670 17.15213447 5.393477

5|4 93.11830740 17.22500824 5.405995

4|3 93.97925603 17.36408265 5.412279

3|2 95.04545827 17.52773237 5.422576

2|1 96.33159020 17.65621660 5.455959

p <- pnorm(abs(ctable[, "t value"]), lower.tail = FALSE) \* 2

(ctable <- cbind(ctable, "p value" = p))

All the p-values seem to be normal and very less indicating that it is a good model.

Value Std. Error t value p value

research 0.57934098 0.10770924 5.378749 7.500503e-08

citations 0.36795561 0.09944030 3.700266 2.153734e-04

income 0.02418672 0.01598601 1.512993 1.302814e-01

30|29 77.46514443 14.97649897 5.172447 2.310483e-07

29|28 78.82362978 15.06101656 5.233619 1.662225e-07

28|27 79.62157446 15.12943055 5.262695 1.419592e-07

27|26 80.21979553 15.16343395 5.290345 1.220859e-07

26|25 81.01243409 15.30141477 5.294441 1.193812e-07

25|24 81.78553395 15.45195472 5.292892 1.203968e-07

24|23 82.51623923 15.56203714 5.302406 1.142863e-07

23|22 83.57823988 15.82633101 5.280961 1.285079e-07

22|21 84.44739770 16.01918024 5.271643 1.352080e-07

21|20 84.97270662 16.08164979 5.283830 1.265105e-07

20|19 85.52790256 16.16395976 5.291272 1.214688e-07

19|18 86.12305974 16.27513585 5.291695 1.211877e-07

18|17 86.66966907 16.37872935 5.291599 1.212515e-07

17|16 87.16105442 16.45854024 5.295795 1.184997e-07

16|15 87.55762898 16.50167524 5.305984 1.120666e-07

15|14 87.98367637 16.54957373 5.316371 1.058571e-07

14|13 88.48959244 16.62379244 5.323069 1.020311e-07

13|12 88.91602362 16.67358662 5.332747 9.673790e-08

12|11 89.32971497 16.71768829 5.343425 9.120653e-08

11|10 89.74476181 16.76022811 5.354627 8.573332e-08

10|9 90.14904569 16.79439402 5.367806 7.970030e-08

9|8 90.58824078 16.83900873 5.379666 7.462421e-08

8|7 91.13631326 16.92208495 5.385643 7.218600e-08

7|6 91.88706739 17.06656295 5.384041 7.283207e-08

6|5 92.50963670 17.15213447 5.393477 6.910731e-08

5|4 93.11830740 17.22500824 5.405995 6.444962e-08

4|3 93.97925603 17.36408265 5.412279 6.222771e-08

3|2 95.04545827 17.52773237 5.422576 5.874627e-08

2|1 96.33159020 17.65621660 5.455959 4.870929e-08

The confidence intervals for the independent variables can be calculated with the below code,

# confidence intervals

(ci <- confint(model))

> (ci <- confint(model))

Waiting for profiling to be done...

2.5 % 97.5 %

research 0.387488857 0.81277482

citations 0.186110025 0.58222650

income -0.007053605 0.05641264

(exp(coef(model)))

research citations income

1.784862 1.444778 1.024482

exp(cbind(OR = coef(model), ci))

OR 2.5 % 97.5 %

research 1.784862 1.4732765 2.254154

citations 1.444778 1.2045548 1.790019

income 1.024482 0.9929712 1.058034

pred <- predict(model, training\_set)

print(pred, digits = 3)

|  |
| --- |
|  |
| |  | | --- | | [1] 1 2 3 4 7 2 7 14 7 14 7 14 7 14 22 14 14 18 19 22 19 12  [23] 23 26 23 29 29 23 29 30  30 Levels: 30 29 28 27 26 25 24 23 22 21 20 19 18 17 16 15 14 ... 1 | | The predictions are almost close to the original ranks. | | |  | | --- | |  | | |