# APPLIED REGRESSION ANALYSIS

## REGRESSION ANALYSIS ON WORLD UNIVERSITY RANKINGS

**Islam Ehab Anwar**

**Student ID: 201900337**

**Faculty of Science & Innovation**

**The Universities of**

December 2021

## Contents

## ABSTRACT

Regression Analysis is used to find the relationships between the response variable and several independent variables. Furthermore, to understand the underlying architecture of our dataset to generate an effective linear regression model. The experiment was conducted on the Center for World University Rankings (CWUR) dataset. Various assumption violations were detected. Remedial actions such as power transformations, principal component analysis, and more techniques were performed to mitigate these violations. Results illustrate that our final model was a more effective model than the initial model as we adjusted many different aspects of our data such as multicollinearity and normality of the data. Furthermore, we will be able to use the model performance to measure the proportion of variation in the dependent variable that is given by the predictor variables.



***Keywords*** Regression Analysis · CWUR · Transformations · Principal Component Analysis

# 1 Problem Statement

## 1.1 Why We Perform Analysis?

Data Analysis is the process of inspecting, cleansing, transforming, and modeling data to discover useful information, suggesting conclusions, and supporting decision-making (2). We perform regression analysis because we want to identify relationships between certain variables in a set of data, discover hidden relationships and meanings within the data, as well as provide our insights concisely and clearly for our interpretation.

Some questions we can answer using data analysis:

- What is the relationship between the response & predictor variables?
- Is there a relationship between the predictor variables?
- Is there any underlying information we can uncover using regression analysis?
- What are the significant variables in predicting the university rank?

## 1.2 Literature Review

We will be performing Regression Analysis on the CWUR data set, analyzing its properties & graphs before fitting it into a model after fitting the data. Comparisons will be conducted in the post-fitting stage. However, this is an iterative process where we must keep appending until we reach an appropriate and accurate model that represents our data without any issues.

This report will serve as an example for Multiple Regression Analysis of various predictor variables and identifying relationships between them. Firstly, we will perform Exploratory Data Analysis, which is a term for certain kinds of initial analysis and findings done with data sets, usually early on in an analytical process (7). Once we've gathered significant information we will fit our regression model into the data set. In addition, we will conduct more analysis on how our model affected the data set using graphical methods.

# 2 Data Description

## 2.1 Complete Source of Data

| | score | country | national_rank | quality_of_education | alumni_employment | quality_of_faculty | publications | influence | citations | broad_impact | patents |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 100.00 | USA | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 3 |
| 2 | 98.66 | USA | 2 | 9 | 2 | 4 | 5 | 3 | 3 | 4 | 10 |
| 3 | 97.54 | USA | 3 | 3 | 11 | 2 | 15 | 2 | 2 | 2 | 1 |
| 4 | 96.81 | United Kingdom | 1 | 2 | 10 | 5 | 11 | 6 | 12 | 13 | 48 |
| 5 | 96.46 | United Kingdom | 2 | 7 | 13 | 10 | 7 | 12 | 7 | 9 | 15 |
| 6 | 96.14 | USA | 4 | 13 | 6 | 9 | 13 | 13 | 11 | 12 | 4 |
| 7 | 92.25 | USA | 5 | 5 | 21 | 6 | 10 | 4 | 4 | 7 | 29 |
| 8 | 90.70 | USA | 6 | 11 | 14 | 8 | 17 | 16 | 12 | 22 | 141 |
| 9 | 89.42 | USA | 7 | 4 | 15 | 3 | 72 | 25 | 24 | 33 | 225 |
| 10 | 86.79 | USA | 8 | 12 | 18 | 14 | 24 | 15 | 25 | 22 | 11 |
| 11 | 86.61 | USA | 9 | 10 | 26 | 11 | 18 | 8 | 35 | 20 | 49 |
| 12 | 84.40 | USA | 10 | 6 | 328 | 7 | 53 | 9 | 19 | 25 | 13 |
| 13 | 78.23 | Japan | 1 | 16 | 3 | 38 | 14 | 19 | 31 | 29 | 7 |
| 14 | 77.60 | USA | 11 | 20 | 4 | 28 | 8 | 18 | 14 | 9 | 14 |
| 15 | 76.91 | USA | 12 | 28 | 27 | 13 | 6 | 14 | 8 | 6 | 9 |
| 16 | 71.60 | USA | 13 | 18 | 84 | 16 | 4 | 11 | 5 | 3 | 2 |
| 17 | 68.60 | Japan | 2 | 22 | 16 | 24 | 30 | 42 | 88 | 60 | 16 |
| 18 | 68.39 | USA | 14 | 32 | 22 | 18 | 47 | 29 | 46 | 32 | 43 |
| 19 | 68.36 | USA | 15 | 24 | 17 | 140 | 3 | 20 | 6 | 14 | 12 |
| 20 | 66.93 | Switzerland | 1 | 17 | 64 | 17 | 44 | 27 | 39 | 79 | 57 |
| 21 | 66.59 | USA | 16 | 36 | 567 | 19 | 16 | 5 | 15 | 15 | 17 |
| 22 | 66.56 | USA | 17 | 163 | 12 | 104 | 27 | 34 | 27 | 26 | 42 |
| 23 | 65.71 | Israel | 1 | 15 | 164 | 15 | 120 | 97 | 368 | 143 | 35 |
| 24 | 64.82 | South Korea | 1 | 367 | 9 | 218 | 36 | 163 | 146 | 112 | 6 |
| 25 | 64.51 | USA | 18 | 29 | 29 | 34 | 21 | 24 | 22 | 27 | 27 |
| 26 | 63.69 | USA | 19 | 367 | 567 | 20 | 22 | 7 | 15 | 11 | 56 |
| 27 | 62.27 | United Kingdom | 3 | 21 | 447 | 27 | 12 | 22 | 15 | 18 | 67 |
| 28 | 61.55 | USA | 20 | 74 | 31 | 56 | 19 | 23 | 15 | 16 | 33 |
| 29 | 61.28 | USA | 21 | 49 | 567 | 12 | 372 | 28 | 115 | 70 | 289 |
| 30 | 61.14 | USA | 22 | 75 | 36 | 25 | 48 | 50 | 26 | 71 | 38 |

Figure 1: Sample Observations - CWUR Dataset

The Center for World University Rankings (CWUR) dataset is a collection of data that represents the top 100 universities across the globe. In this report, we adjusted the original dataset to set the score of each respective university as the response variable. Furthermore, we also removed certain variables such as the year of rankings to remove the aspect of time series. The purpose of this experiment was to conduct multiple linear regression without the factor of time series.

### 2.2 Types of Observations

Our CWUR dataset consists of observations that provide information on the ranking of the top universities around the world. It contains a multitude of variables such as the score of each university, what country they reside in, metrics that quantify the quality of education as well as other factors such as influence or broad impact.

### 2.3 Variable Types & Units of Measurement

Concerning our variables, Figure (2) below highlights all the variables involved in our dataset along with their respective data types & units of measurement. The information below illustrates that we have several numerical predictor variables with a single qualitative variable, with our response variable being a numeric variable representing university scores.

```
'data.frame':   1000 obs. of  11 variables:
 $ score               : num  100 98.7 97.5 96.8 96.5 ...
 $ country             : chr  "USA" "USA" "USA" "United Kingdom" ...
 $ national_rank       : num  1 2 3 1 2 4 5 6 7 8 ...
 $ quality_of_education: num  1 9 3 2 7 13 5 11 4 12 ...
 $ alumni_employment   : num  1 2 11 10 13 6 21 14 15 18 ...
 $ quality_of_faculty  : num  1 4 2 5 10 9 6 8 3 14 ...
 $ publications        : num  1 5 15 11 7 13 10 17 72 24 ...
 $ influence           : num  1 3 2 6 12 13 4 16 25 15 ...
 $ citations           : num  1 3 2 12 7 11 4 12 24 25 ...
 $ broad_impact        : num  1 4 2 13 9 12 7 22 33 22 ...
 $ patents             : num  3 10 1 48 15 4 29 141 225 11 ...
```

Figure 2: CWUR Variable Data Types

### 2.4 Response Variable

Our response variable is the score variable, which indicates the overall score of the universities based on the aforementioned predictors such as country of origin, number of publications, quality of education, and more. We will be performing regression analysis to identify underlying patterns between the response variable and several of our predictors, as well as mitigating any potential assumption violations.

### 2.5 Hypothesis of Predictors Influence on Response

I believe that predictors such as the country of origin, number of patents, and quality of education will have a strong influence on the value of the response variable because, regarding real-world data, universities are renowned for their quality of education as well as their received patents.

## 3 Data Analysis

This section of the report contributes to describing the series of steps performed on our CWUR dataset including graphs before fitting the model, any potential transformation of data, graphs after fitting the model, and analysis of the generated graph structures.

### 3.1 Graphs before Fitting Model to Data

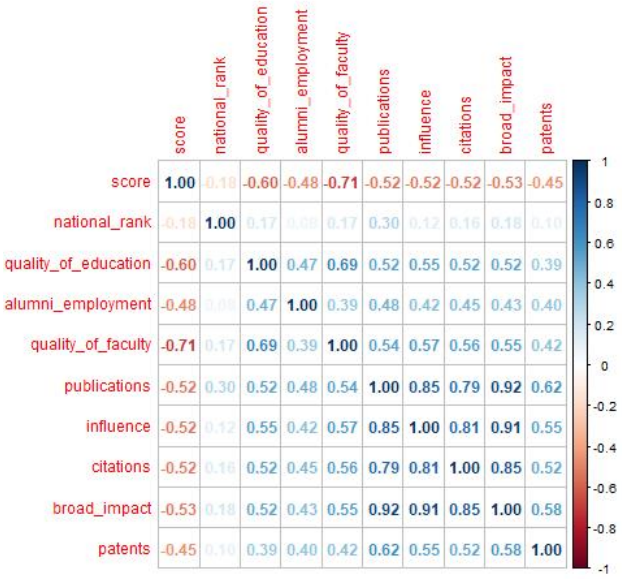



Figure 3: Correlation Matrix

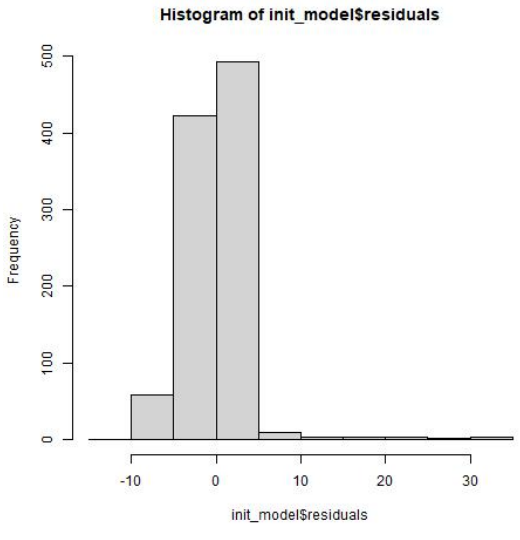



Figure 4: Histogram Plot before Fitting Model

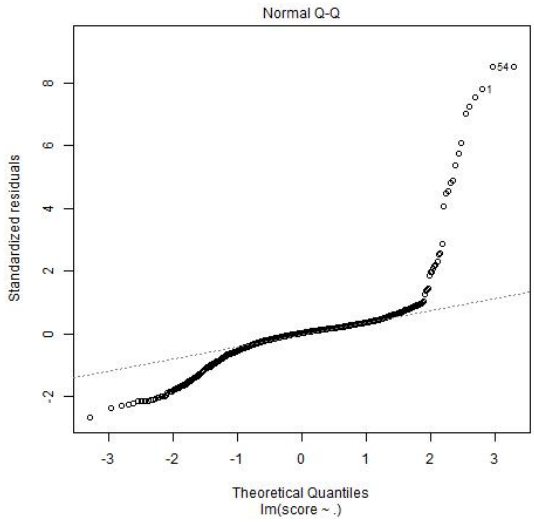



Figure 5: Q-Q Plot before Fitting Model

As shown from the above figures (3-5), there are violations across the board ranging from multicollinearity, normality to linearity. We will mitigate these violations as best as possible given the nature of the dataset provided.

### 3.2 Transformation of Variables

Data does not always arrive in the format we need to perform analytical measures. Therefore, we use a transformation of variables to work around this dilemma. Transformations are applied to accomplish certain objectives such as to ensure linearity, to achieve normality, or to stabilize the variance (O'Reilly). Evidently, from the last subsection, we identified that there was non-linearity in our dataset. Therefore, the remedial action to fix non-linearity is a transformation of variables. This may potentially change the structure of our variables to improve model accuracy and mitigate the violations of both normality and linearity.

Provided Figure (6) below, we concluded that the most appropriate value for our slider was '0' since it clustered the data leniently and other values of lambda caused larger spreads as well as less accurate model summaries. This could in theory end up violating the constant variance assumption. As a result, we ended up taking the logarithm of the base model to transform our variables.

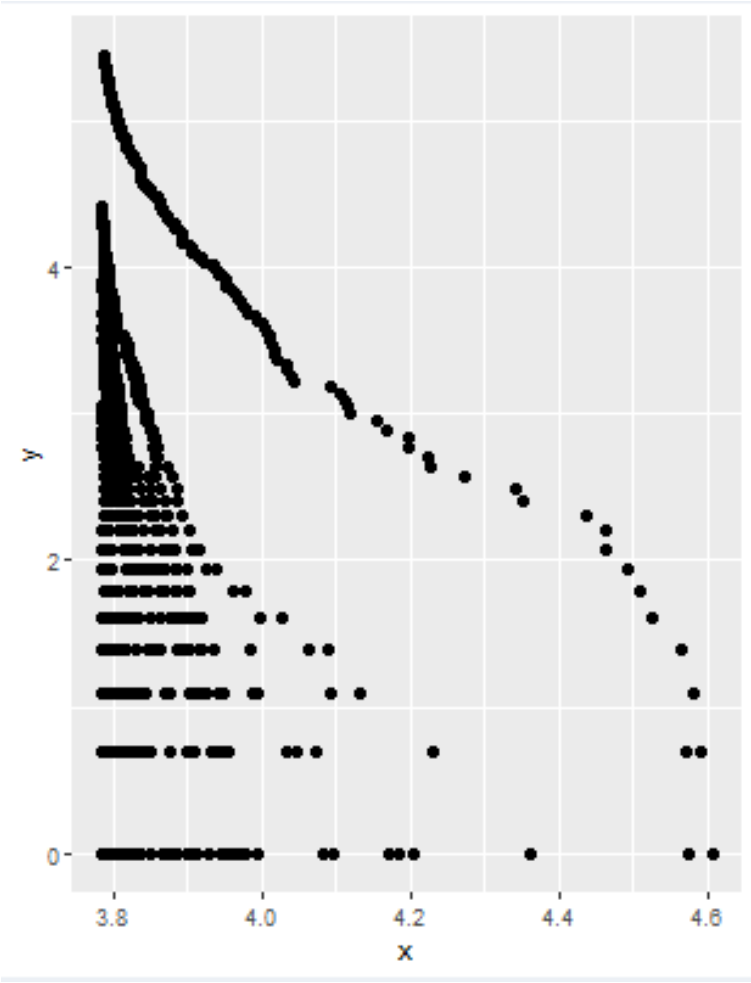



Figure 6: Transformation Slider at $\lambda = 0$

Upon transforming our dataset, we ended up fitting our model through the transformed dataset as shown in Figures (7-10) below. We can observe from figures (7 & 8) that some of the normality violation was fixed but we still have a right-sided skewness. In addition to that, our residual plot shows some improvements in the constant variance assumption. However, we still suffer from a severe case of multicollinearity, which we will attempt to solve very soon. In Figure (28) in the appendix, we see that our adjusted coefficient of determination is at 90%, which is very good. However, due to the high amount of correlation, the result is not very reliable concerning our modeling of the data.

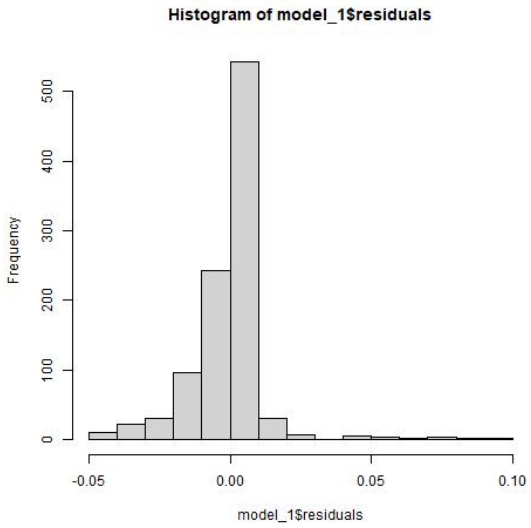



Figure 7: Histogram after Transformation

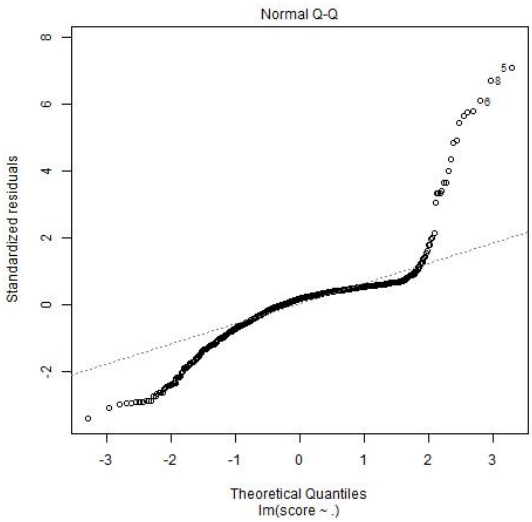



Figure 8: Q-Q after Transformation

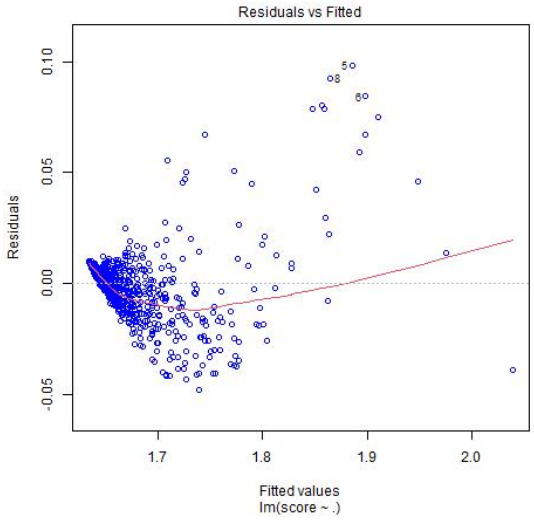



Figure 9: Residual after Transformation

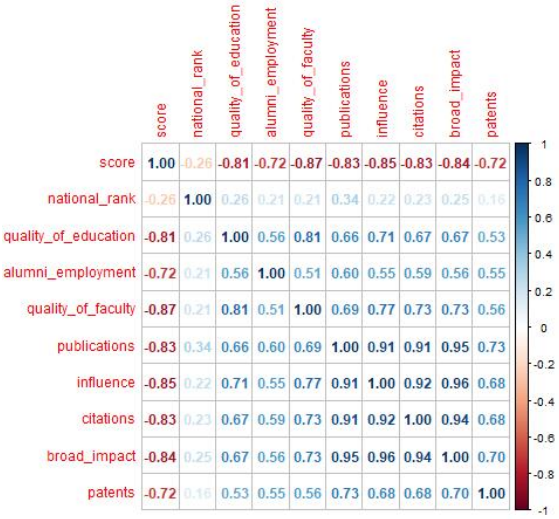



Figure 10: Correlation after Transformation

### 3.3 Initial Model Fit to Data

The initial model fit the dataset was conducted and generated a very poor model as shown in Figures (22 & 22) in the appendix. We measured the performance of this initial model by interpreting our adjusted coefficient of determination, with a value of 0.59 the model is only explaining the variability of the response variable by 60%. The adjusted coefficient of determination is a valid measure of performance for our model as it shows that the independent variables explain a portion of the variation of the response variable, taking into account the increasing number of variables in our dataset (mod). This could be due to many reasons such as many predictor variables being insignificant in this model, as well as a low F-test score is shown in the ANOVA table in Figure (24).

### 3.4 Graphs after Fitting Model to Data

Given the Figures (11-13) below, which are the plots after fitting the model to the dataset, we notice that the data becomes more linearly presented. Furthermore, the residual plot shows that our variance is steadily becoming more constant and does not illustrate any non-linear curves or shapes. Concerning normality, Figure (12) shows that we still have a histogram skewed to the right. However, it appears more normally distributed than the initial model.

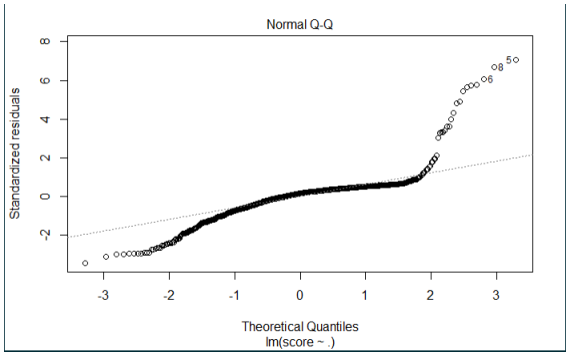



Figure 11: Q-Q Plot after fitting Model

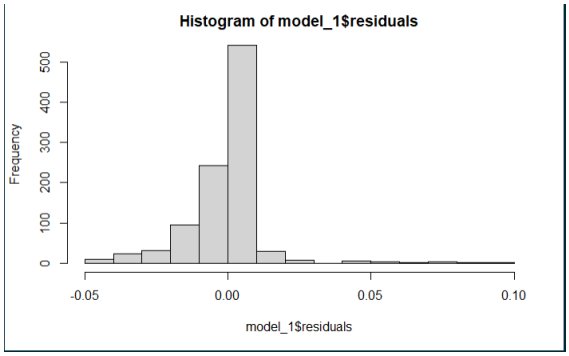



Figure 12: Histogram after Fitting Model

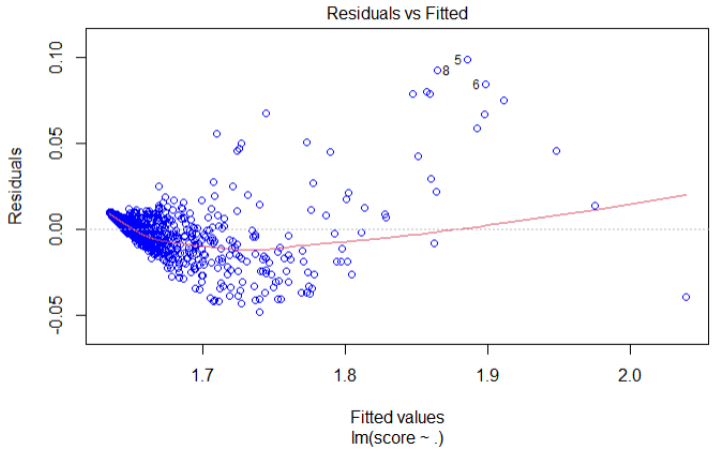



Figure 13: Residual after fitting Model

### 3.5 Regression Diagnostics & Remedial Actions

There is a multitude of graphs and metrics that need to be analyzed before fitting our data to a linear regression model. These visual plots computations help guide us in identifying violations within the data that could hinder the ability of our regression model's performance. We will attempt to mitigate each violation independently and observe the changes in the assumptions whether they were fixed or not. We already covered mitigating the linearity assumption using power transformations in the previous section. Therefore, we will cover the other assumptions and their remedial actions. The collection of assumptions we need to verify aren't violated by our data are listed below:

- Linearity
- Normality
- Autocorrelation
- Multicollinearity

- Constant Variance

1. Linearity:
Regarding the linearity assumption, it means that each predictor variable must have a linear relationship with the response variable (3). This suggests that a change in 'y' is due to a single unit change for 'x' is constant. We infer if the data violates the linearity assumption by plotting the residuals against the fitted values. If our data displays a pattern similar to a general function, then our data is non-linear. According to Figure (14), there is some non-linearity and non constant variance of our data, therefore we need to perform certain remedial actions.

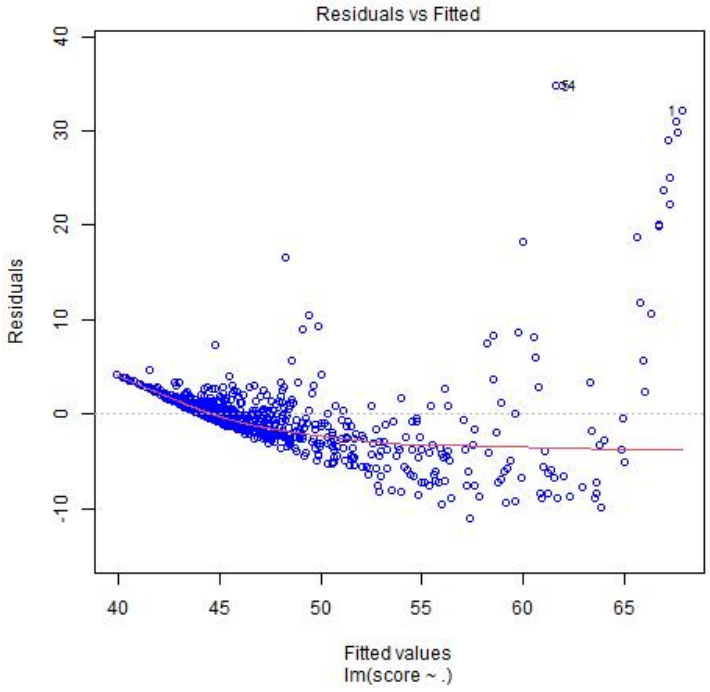



Figure 14: Residuals vs Fitted Values before Fitting Model

2. Normality:
Concerning normality, the errors of our dataset must be normally distributed (3). We interpret if the data is normal by plotting a Q-Q plot, if the data is a linear line with an intercept of 0, then our data is normally distributed. Another method is via the histogram to view normality. According to Figures (15 & 16), the data is not normally distributed and is skewed to the right. Thus, we perform remedial measures to fix this violation.

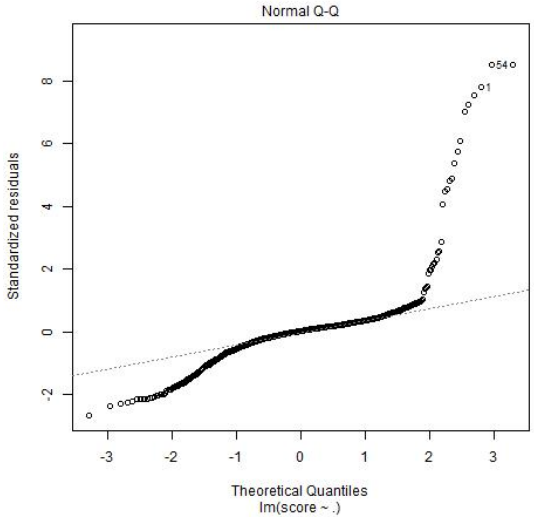



Figure 15: Q-Q Plot before fitting Model

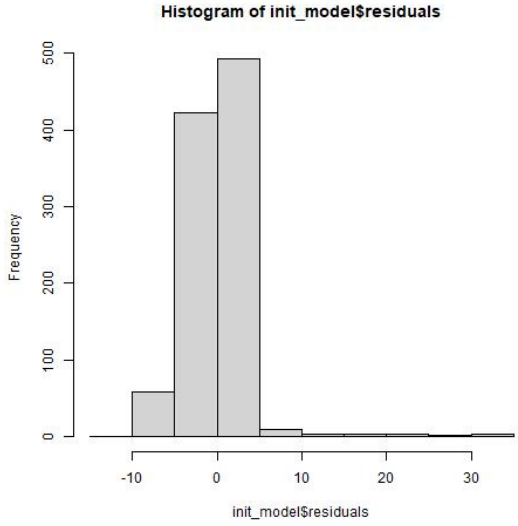



Figure 16: Histogram before Fitting Model

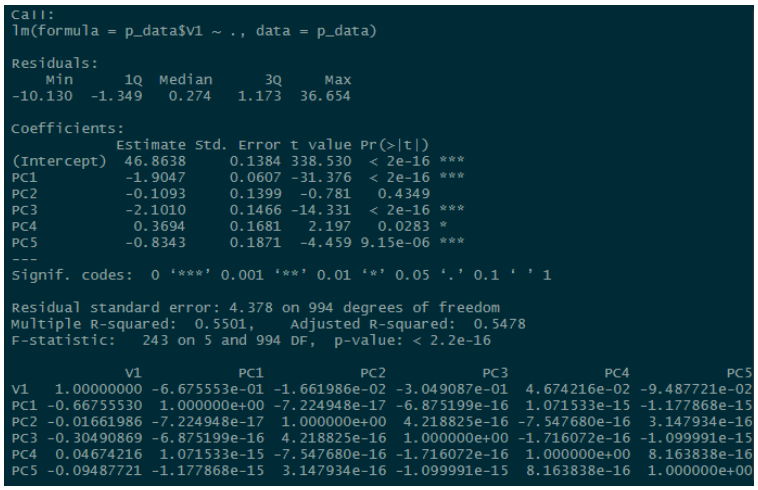
3. Autocorrelation:
For autocorrelation, this is discovered when the following instances or observations depend on the previous instances, which can drastically affect model accuracy (3). We can identify if autocorrelation exists by computing the Durbin-Watson (DW) statistic. If our statistic value is equal to 2 then we have no autocorrelation, which would be our ideal scenario. According to Figure (25), with a value of 0.6, we can conclude that we have a case of positive autocorrelation.

4. Multicollinearity:
In terms of collinearity, this is identified when there is a high correlation between independent variables (3). We check for multicollinearity by plotting a correlation matrix that shows the individual correlation coefficient values between all variables in our dataset. Another method is using the Variance Inflation Factor (VIF) to detect collinearity within our data. According to the output figure (24) shown in the appendix, we appear to have severe cases of collinearity with the 'country' variable as well as low-moderate collinearity with the 'publications' as well as 'broad impact'.

5. Constant Variance:
As for constant variance, this is when most or all of our data has very similar and non-extreme leverage point values (3). We can identify a non-constant variance when the residuals plot exhibits a funnel-shaped curve, which is evident in Figure (13). We can conclude that we have a violation of the constant variance and we must take remedial actions.

Regarding remedial actions, Principal Component Analysis (PCA) was performed to mitigate the violation of multicollinearity. PCA is an effective dimensionality reduction tool used in linear regression to handle problems of multicollinearity (6). This was performed after the transformation of the dataset to ensure we fixed the violation of the normality assumption, and that the independent variables had a linear relationship with the response variable.

Using the "tidyverse" library within R, we separated the numerical predictors from the response variable and scaled the predictors, then summarized the contribution of each principal component to the dataset as shown in figure (show variance contribution). After that, we standardized the predictors and fit the principal components to produce our final data as shown in figure (show final data). Upon computing the correlation, we ended up eliminating our multicollinearity and then ran a specific number of principal components to compile approximately 80% of the total variance contribution as shown in Figure (17).



```
Call:
lm(formula = p_data$V1 ~ ., data = p_data)

Residuals:
    Min      1Q  Median      3Q     Max
-10.130  -1.349   0.274   1.173  36.654

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  46.8638     0.1384 338.530  < 2e-16 ***
PC1          -1.9047     0.0607 -31.376  < 2e-16 ***
PC2          -0.1093     0.1399  -0.781   0.4349
PC3          -2.1010     0.1466 -14.331  < 2e-16 ***
PC4           0.3694     0.1681   2.197   0.0283 *
PC5          -0.8343     0.1871  -4.459 9.15e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.378 on 994 degrees of freedom
Multiple R-squared:  0.5501,    Adjusted R-squared:  0.5478
F-statistic:   243 on 5 and 994 DF,  p-value: < 2.2e-16

              V1           PC1           PC2           PC3           PC4           PC5
V1   1.00000000 -6.675553e-01 -1.661986e-02 -3.049087e-01  4.674216e-02 -9.487721e-02
PC1 -0.66755530  1.000000e+00 -7.224948e-17 -6.875199e-16  1.071533e-15 -1.177868e-15
PC2 -0.01661986 -7.224948e-17  1.000000e+00  4.218825e-16 -7.547680e-16  3.147934e-16
PC3 -0.30490869 -6.875199e-16  4.218825e-16  1.000000e+00 -1.716072e-16 -1.099991e-15
PC4  0.04674216  1.071533e-15 -7.547680e-16 -1.716072e-16  1.000000e+00  8.163838e-16
PC5 -0.09487721 -1.177868e-15  3.147934e-16 -1.099991e-15  8.163838e-16  1.000000e+00
```

Figure 17: Model after applying PCA

### 3.6 Identification of Outliers, Leverage Points, and Influential Observations

Concerning outliers, R is designed to detect outliers when plots of our models are displayed. According to Figure (18), observations 1, 2, and 3 are all potential outliers in our dataset. Moreover, the provided figures (19 & 20) display our calculation of leverage values in our CWUR dataset.

Leverage points are points that can be proven as outliers if a corresponding leverage value exceeds a certain threshold. It is evident that from our sorted sample leverage values that none of our leverage points are actual outliers but we're viewed as potential outliers awaiting further investigation.

Another measure of importance is Cook's Distance, which is used to measure the influence of each data point in our dataset on the overall model (5). As shown in Figure (34), we calculate the influence of each observation within our dataset then generate a plot showing the most influential points that exceed a threshold. In addition, Figure (34) illustrates that our first few observations exceed our threshold and adversely affect our model since they are influential data points to our model.

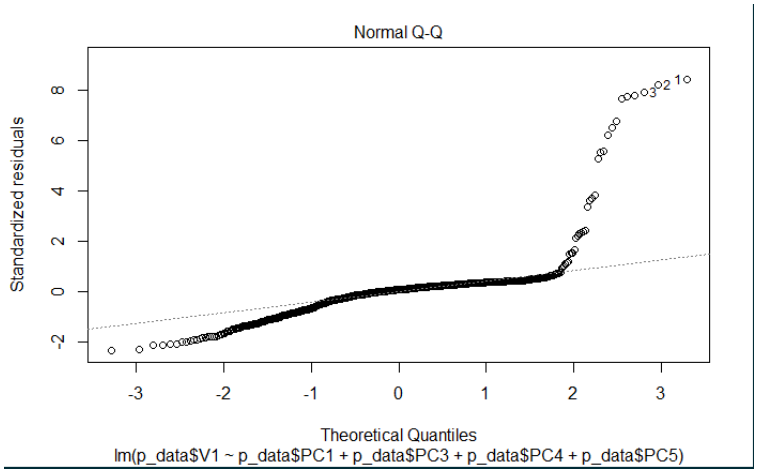



Figure 18: Model Outliers

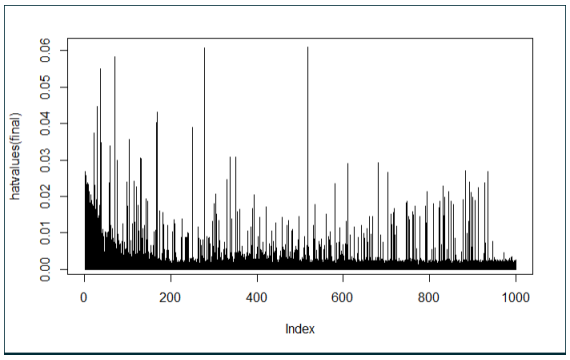



Figure 19: Leverage Index Plot

```r
hats <- as.data.frame(hatvalues(final))
head(hats)
```

Description: df [6 x 1]

| | hatvalues(final) <dbl> |
|---|---|
| 1 | 0.02683795 |
| 2 | 0.02508526 |
| 3 | 0.02568867 |
| 4 | 0.02457787 |
| 5 | 0.02339119 |
| 6 | 0.02323714 |

6 rows

```r
head(hats[order(-hats['hatvalues(final)']), ])
```

```
Warning in xtfrm.data.frame(x) : cannot xtfrm data frames
[1] 0.06095691 0.06063543 0.05818175 0.05494280 0.04469339 0.04314454
```

Figure 20: Sorted Sample Leverage Values

### 3.7 Necessary Repetitions

As observed by Figure (32), we notice that our second transformation resembles a more linear plot when $\lambda$ is equal to 2. As a result, we square our principal components as a compound step to then create another linear regression model that would produce more effective results. According to Figure (31), we notice an adjusted coefficient of determination of approximately 61% which is an improvement on our 1st iterative model at 54%.

# 4 Summary & Conclusions

## 4.1 Final Model

According to Figure (31), our final model produced a decent representation of our dataset after performing Principal Component Analysis and two power transformations to mitigate multicollinearity as well as solve the violations regarding normality and linearity respectively.

Furthermore, most of the chosen principal components were significant to our model, which shows that our values of the adjusted coefficient of determination are reliable and explain the variability of the response variable by approximately 60%. Another indicator that our model is a decent model is interpreting the minimal values of our leverage points. We found that almost all of our points leverage values we're extremely small so that was an effective indicator of our model performance.

## 4.2 Conclusions

In conclusion, the initial model we began with had serious assumption violations. We performed certain remedial actions such as power transformations, principal component analysis, as well as choosing the most effective components to produce a significant and accurate model.

Our model is not perfect, we still have some underlying violations of assumptions such as the linearity of our dataset. However, that could be due to the changes we performed in the source data by removing certain fields and choosing 'score' as our response variable for the usability of our regression analysis. We managed to improve the initial model but there will be many more iterations to be performed in the future to ensure the assumptions remain unviolated and to generate an accurate predictive linear regression model.

# 5 Appendix

## 5.1 Annotated Computer Output (Tables & Figures)

```r
init_model <- lm(score ~., data = cwur)
jpeg("initalmodel.png", quality = 75)
summary(init_model)
dev.off()
```

```
Call:
lm(formula = score ~ ., data = cwur)

Residuals:
    Min      1Q  Median      3Q     Max
-11.041  -1.138   0.076   0.922  34.869

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              59.9435309  2.6219979  22.862  < 2e-16 ***
countryAustralia          0.9719097  2.5636074   0.379 0.704687
countryAustria           -0.6544266  2.7350959  -0.239 0.810949
countryBelgium           -0.5271991  2.8010106  -0.188 0.850747
countryBrazil             0.8831326  2.6085474   0.339 0.735022
countryBulgaria          -0.3452953  4.8184597  -0.072 0.942887
countryCanada             1.5183152  2.5535128   0.595 0.552256
countryChile             -0.6491940  3.2208067  -0.202 0.840303
countryChina              2.6606121  2.4893797   1.069 0.285443
countryColombia          -1.3530537  3.8396815  -0.352 0.724628
countryCroatia            0.8137963  4.8232015   0.169 0.866049
countryCyprus             0.3599247  4.8141363   0.075 0.940418
countryCzech Republic    -0.4223762  3.0633690  -0.138 0.890365
countryDenmark            0.1488606  3.1074913   0.048 0.961803
countryEgypt             -0.6330937  3.1955837  -0.198 0.842998
countryEstonia            0.2842081  4.8498299   0.059 0.953282
countryFinland            0.2901844  2.8072471   0.103 0.917692
countryFrance             1.3598022  2.5128982   0.541 0.588548
countryGermany            1.2878441  2.5338827   0.508 0.611399
countryGreece             1.1730958  2.8983591   0.405 0.685758
countryHong Kong          0.5838851  3.0093180   0.194 0.846198
countryHungary           -1.0308366  2.9790681  -0.346 0.729401
countryIceland            0.8502278  4.8447264   0.175 0.860728
countryIndia             -0.1058253  2.6558559  -0.040 0.968224
countryIran               0.6015360  2.8353383   0.212 0.832031
countryIreland            0.0125234  2.8879825   0.004 0.996541
countryIsrael            -0.3390476  2.9525697  -0.115 0.908604
countryItaly              1.6829076  2.5169584   0.669 0.503899
countryJapan              2.1087689  2.5501176   0.827 0.408488
countryLebanon           -2.8912700  4.8457896  -0.597 0.550882
countryLithuania         -0.2096529  4.8269540  -0.043 0.965365
countryMalaysia          -0.0423385  3.4745417  -0.012 0.990280
countryMexico             0.9846401  3.8073615   0.259 0.795990
countryNetherlands        2.1106754  2.7192920   0.776 0.437836
countryNew Zealand       -0.0937111  2.9651788  -0.032 0.974795
countryNorway            -1.4324584  3.0677860  -0.467 0.640655
countryPoland            -0.2234933  2.7869733  -0.080 0.936102
countryPortugal          -0.5548502  2.8963815  -0.192 0.848123
countryPuerto Rico       -0.2171461  4.8274579  -0.045 0.964132
countryRomania           -0.2125152  3.8093312  -0.056 0.955523
countryRussia            -1.1971622  3.0652129  -0.391 0.696209
countrySaudi Arabia      -1.7623458  3.2296012  -0.546 0.585413
countrySerbia             0.8470170  4.8126048   0.176 0.860332
countrySingapore          6.7431848  3.8711561   1.742 0.081855 .
countrySlovak Republic    0.4511363  4.8234110   0.094 0.925502
countrySlovenia           0.7803401  3.8136731   0.205 0.837917
countrySouth Africa      -1.8295783  3.1029805  -0.590 0.555589
countrySouth Korea        2.0705186  2.5930055   0.799 0.424783
countrySpain              1.6325086  2.5338525   0.644 0.519553
countrySweden            -0.1011255  2.7450694  -0.037 0.970621
countrySwitzerland       -0.0672493  2.8426920  -0.024 0.981131
countryTaiwan             1.4716245  2.5818648   0.570 0.568825
countryThailand           0.5464802  3.4162187   0.160 0.872942
countryTurkey             0.2722672  2.7607176   0.099 0.921460
countryUganda            -0.0335890  4.8228791  -0.007 0.994445
countryUnited Arab Emirates -0.6649140  4.8403560  -0.137 0.890769
countryUnited Kingdom     2.4010511  2.5264558   0.950 0.342174
countryUruguay            3.6145366  4.8268227   0.749 0.454140
countryUSA                8.0315210  2.6606519   3.019 0.002608 **
national_rank            -0.0544074  0.0063621  -8.552  < 2e-16 ***
quality_of_education     -0.0068302  0.0018807  -3.632 0.000297 ***
alumni_employment        -0.0046063  0.0009366  -4.918 1.03e-06 ***
quality_of_faculty       -0.0506928  0.0038299 -13.236  < 2e-16 ***
publications              0.0029491  0.0016758   1.760 0.078769 .
influence                -0.0010775  0.0013749  -0.784 0.433424
citations                 0.0006162  0.0010859   0.567 0.570568
broad_impact             -0.0009903  0.0019407  -0.510 0.609955
patents                  -0.0011671  0.0008510  -1.371 0.170566
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.159 on 932 degrees of freedom
Multiple R-squared:  0.6191,    Adjusted R-squared:  0.5917
F-statistic: 22.61 on 67 and 932 DF,  p-value: < 2.2e-16
```

Figure 21: Initial Model Summary Part 1

Figure 22: Initial Model Summary Part 2

```
Analysis of Variance Table

Response: score
                     Df  Sum Sq Mean Sq  F value    Pr(>F)
country              58  4488.6    77.4   4.4730 < 2.2e-16 ***
national_rank         1 15167.6 15167.6 876.6772 < 2.2e-16 ***
quality_of_education  1  3117.8  3117.8 180.2051 < 2.2e-16 ***
alumni_employment     1   290.2   290.2  16.7742 4.576e-05 ***
quality_of_faculty    1  3059.4  3059.4 176.8325 < 2.2e-16 ***
publications          1    36.2    36.2   2.0952    0.1481
influence             1    13.5    13.5   0.7797    0.3774
citations             1     3.5     3.5   0.2033    0.6522
broad_impact          1     2.3     2.3   0.1340    0.7144
patents               1    32.5    32.5   1.8809    0.1706
Residuals           932 16124.8    17.3
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 23: Initial Model Anova

```r
vif(init_model)
```

```
                       GVIF Df GVIF^(1/(2*Df))
country               51.583208 58        1.034577
national_rank          6.678873  1        2.584352
quality_of_education   2.432565  1        1.559668
alumni_employment      1.767046  1        1.329303
quality_of_faculty     2.443173  1        1.563065
publications          13.521041  1        3.677097
influence              9.083174  1        3.013830
citations              4.327240  1        2.080202
broad_impact          17.907398  1        4.231713
patents                3.183148  1        1.784138
```

Figure 24: VIF before Fitting

```r
durbinWatsonTest(init_model)
```

```
 lag Autocorrelation D-W Statistic p-value
   1        0.661331     0.6130749       0
 Alternative hypothesis: rho != 0
```

Figure 25: Durbin-Watson before Fitting

```r
library(shiny)
library(shinydashboard)
library(ggplot2)

library(MASS)
df=cwur_num # name your data df

slider=c(-2,2,0.05)
data <-  data.frame(x=c(1,2,3,4),y=c(10,11,12,13))
ui <- dashboardPage(
  dashboardHeader(),
  dashboardSidebar(sliderInput("Lambda","Lambda", min=slider[1], max=slider[2], step=slider[3], value=1)),
  dashboardBody(
    fluidRow(column(6,plotOutput('waveplot')))
  ))

server <- function(input, output, session) {
  output$waveplot <- renderPlot({
#    yfxn<- function(df) { df^input$Lambda}
    yfxn<- function(df) { ifelse(input$Lambda!=0, return(df^input$Lambda), return(log(df)) ) }
    df1 <- yfxn(df)
    df1 <- data.frame(df1)
x=df1[,1]; y=df1[,2]
    ggplot(df1,aes_string(x=x,y=y))+geom_point(size=2)+
        scale_x_continuous()
  })
}
shinyApp(ui, server)
```

Figure 26: Power Transformation Slider Snippet Code

```r
model_1 <- lm(score ~., data = transformed_model)
summary(model_1)
```

```
Call:
lm(formula = score ~ ., data = transformed_model)

Residuals:
      Min        1Q    Median        3Q       Max
-0.048092 -0.005458  0.002349  0.005968  0.098421

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          2.0457276  0.0041520 492.711  < 2e-16 ***
national_rank       -0.0011614  0.0008092  -1.435 0.151517
quality_of_education -0.0173239  0.0024657  -7.026 3.95e-12 ***
alumni_employment    -0.0289091  0.0015703 -18.409  < 2e-16 ***
quality_of_faculty   -0.0693950  0.0033418 -20.766  < 2e-16 ***
publications         -0.0022303  0.0038338  -0.582 0.560870
influence            -0.0058236  0.0040170  -1.450 0.147448
citations             0.0006329  0.0033766   0.187 0.851351
broad_impact         -0.0201046  0.0052751  -3.811 0.000147 ***
patents              -0.0139389  0.0015875  -8.781  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01414 on 990 degrees of freedom
Multiple R-squared:  0.9099,    Adjusted R-squared:  0.909
F-statistic:  1110 on 9 and 990 DF,  p-value: < 2.2e-16
```

Figure 27: Summary of Model after Transformation

```r
predictors_pca$rotation
```

| | PC1 | PC2 | PC3 | PC4 | PC5 | PC6 | PC7 | PC8 | PC9 |
|---|---|---|---|---|---|---|---|---|---|
| national_rank | 0.1106127 | 0.96668737 | -0.01299575 | 0.15555944 | -0.001389596 | 0.01852019 | -0.05377589 | -0.15288367 | -0.047984880 |
| quality_of_education | 0.3126158 | 0.02908563 | 0.57324773 | -0.25123086 | 0.062370477 | -0.70184564 | -0.09368236 | 0.06130367 | -0.026145438 |
| alumni_employment | 0.2624376 | -0.15298708 | 0.43542232 | 0.71349032 | -0.415038105 | 0.16063971 | 0.06196273 | -0.07443548 | -0.040232406 |
| quality_of_faculty | 0.3190883 | 0.03669986 | 0.45592541 | -0.40487210 | 0.241932235 | 0.66755833 | 0.13836843 | 0.04161386 | 0.001873327 |
| publications | 0.4013063 | 0.08545577 | -0.25883192 | 0.06038489 | -0.080689394 | -0.06540829 | 0.33992334 | 0.63736036 | 0.477959348 |
| influence | 0.3952532 | -0.10299412 | -0.23885052 | -0.17161450 | -0.144786745 | -0.09068363 | 0.30712869 | -0.73088959 | 0.298247673 |
| citations | 0.3837421 | -0.06613860 | -0.19632720 | -0.11330679 | -0.224580099 | 0.13894552 | -0.84190671 | 0.03464647 | 0.131284170 |
| broad_impact | 0.4047609 | -0.04462399 | -0.30563309 | -0.09677622 | -0.143631694 | -0.03257204 | 0.18372070 | 0.12594039 | -0.812462194 |
| patents | 0.3016358 | -0.12506617 | -0.13325949 | 0.42940951 | 0.816565915 | -0.05255377 | -0.11362415 | -0.09003835 | -0.025778845 |

Figure 28: Eigenvectors of PCA

```r
summary(predictors_pca)
```

```
Importance of components:
                          PC1    PC2    PC3    PC4    PC5    PC6    PC7    PC8    PC9
Standard deviation     2.2816 0.9900 0.94478 0.8238 0.74029 0.55141 0.46227 0.35318 0.22913
Proportion of Variance 0.5784 0.1089 0.09918 0.0754 0.06089 0.03378 0.02374 0.01386 0.00583
Cumulative Proportion  0.5784 0.6873 0.78649 0.8619 0.92278 0.95656 0.98031 0.99417 1.00000
```

Figure 29: Influence of each PCA Component

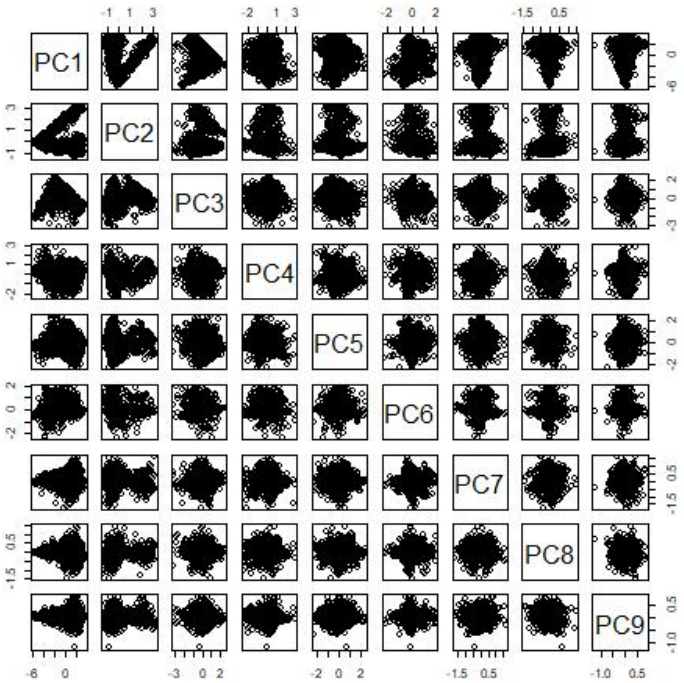



Figure 30: Scatter Plot of PCA Components

```r
transformed_p_data <- p_data^2
summary(transformed_p_data)
```

```
       V1              PC1               PC2               PC3               PC4               PC5          
 Min.   : 1938   Min.   : 0.0000   Min.   : 0.000001   Min.   : 0.000004   Min.   :0.00000   Min.   :0.000001  
 1st Qu.: 1962   1st Qu.: 0.8557   1st Qu.: 0.068688   1st Qu.: 0.094100   1st Qu.:0.06262   1st Qu.:0.040136  
 Median : 2005   Median : 3.1745   Median : 0.276675   Median : 0.392163   Median :0.26041   Median :0.216286  
 Mean   : 2239   Mean   : 5.2004   Mean   : 0.979215   Mean   : 0.891714   Mean   :0.67792   Mean   :0.547486  
 3rd Qu.: 2166   3rd Qu.: 6.7756   3rd Qu.: 0.818687   3rd Qu.: 1.167878   3rd Qu.:0.78686   3rd Qu.:0.701093  
 Max.   :10000   Max.   :35.4846   Max.   :10.373506   Max.   :10.117155   Max.   :8.58798   Max.   :5.754789  
```

```r
final <- lm(transformed_p_data$v1 ~., data = transformed_p_data)
summary(final)
```

```
Call:
lm(formula = transformed_p_data$v1 ~ ., data = transformed_p_data)

Residuals:
     Min       1Q   Median       3Q      Max 
-1244.3  -220.0    52.2   206.7  4511.6 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1567.904     30.657  51.144  < 2e-16 ***
PC1           88.914      2.671  33.291  < 2e-16 ***
PC2           -7.816      8.834  -0.885  0.37652    
PC3          158.183     13.489  11.727  < 2e-16 ***
PC4           67.354     15.563   4.328 1.66e-05 ***
PC5           53.348     20.275   2.631  0.00864 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 522.9 on 994 degrees of freedom
Multiple R-squared:  0.6136,    Adjusted R-squared:  0.6117 
F-statistic: 315.7 on 5 and 994 DF,  p-value: < 2.2e-16
```

Figure 31: Final Model

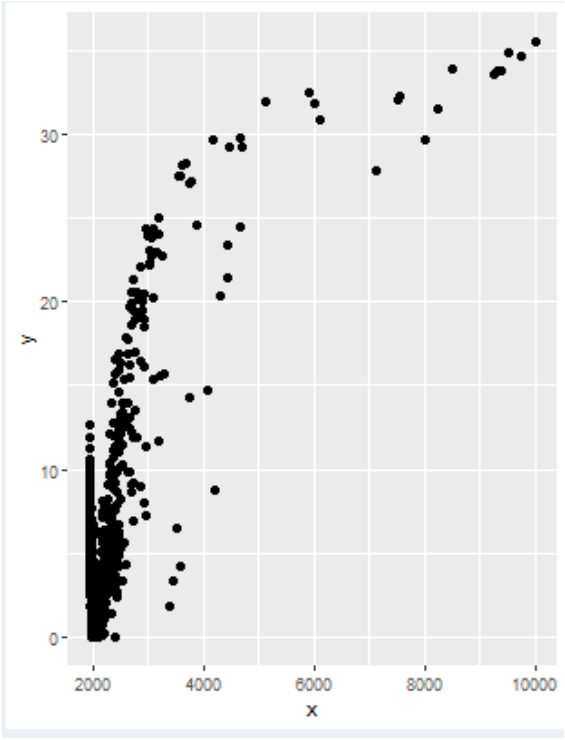



Figure 32: 2nd Transformation at $\lambda = 2$

```r
cooksD <- cooks.distance(final)
head(cooksD)
```

```
        1         2         3         4         5         6
0.3516451 0.3054579 0.2767400 0.2569519 0.2441279 0.2366720
```

```r
n <- nrow(transformed_p_data)
plot(cooksD, main = "Cooks Distance for Influential Obs")
abline(h = 4/n, lty = 2, col = "steelblue")
```

Figure 33: Leverage Index Plot

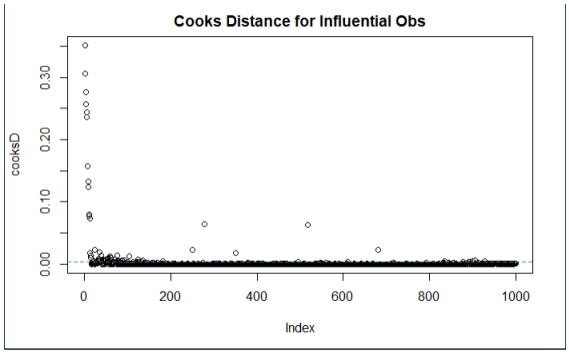



Figure 34: Sorted Sample Leverage Values

## References

[mod]

[2] Johnson, D. (2021). What is data analysis? research: Types: Methods: Techniques.

[3] Karthe (2020). Assumptions of regression analysis, plots amp; solutions.

[O'Reilly] O'Reilly. Regression analysis by example, 4th edition.

[5] Stephanie (2018). Cook's distance / cook's d: Definition, interpretation.

[6] Stephanie (2021). Principal component analysis (pca), regression amp; parafac.

[7] Techopedia (2017). What is exploratory data analysis (eda)? - definition from techopedia.