World Rankings for Universities

James Tsai – Spring 2016, MSDS6372 – Experimental Statistics II

# Introduction

From sports to education, there is a fascination with rankings. There is a desire for one’s favorite team to be the best, and for every parent to want their child to attend the very best college or university. While sports teams compete on the playing field to determine which team is better, it is not so clear-cut when it comes to ranking universities. The size and purpose of institutions and the methodologies used to rank them is a hotly debated issue. Unfortunately, despite the subjectivity of these methodologies, the appeal of university ranking publications has become more influential across the globe as students have become ever more mobile and are looking far beyond their borders. For universities, having a good reputation translates to money in the form of more applications, more tuition dollars, and greater levels of alumni giving.

# Problem Statement

The purpose of the study is to take one of the most influential and observed university measures, The Times Higher Education World University Ranking for 2016, and determine whether any of the nine available variables has the greatest influence on the total score. Understanding these variables may also give us some insight into the nature of the ranking methodology and whether specific criticisms levied against the ranking system are justified.

# Constraints and Limitations

Since the scope of this observational study is for The Times Higher Education World University Ranking for 2016, no casual inference can be made between the explanatory and response variables. It is possible that there are more important measures, which better explain the overall ranking of the universities, which were not published.

Furthermore, we must bear in mind that the methodologies and weightings for each criterion used by The Times Higher Education World Ranking maybe adjusted from year-to-year. We will weight the explanatory variables equally as we are investigating the influence of all variables equally. Finally, to get another perspective of this study, we note that seven schools have consistently taken one of top seven spots under this ranking methodology from 2011-2016. They are: California Institute of Technology, University of Oxford, Stanford University, University of Cambridge, Massachusetts Institute of Technology, Harvard University, and Princeton University. Such a small group of ‘winner take all’ in the ranking results may lead one to question the ranking methodology for the top universities as it may not have practical significance to the student applying for admission.

# Data Set Description

The analysis is based on publicly available dataset, available from the Kaggle website and also from The Times Higher Education World University Rankings website. Please refer to: <https://www.kaggle.com/mylesoneill/world-university-rankings> <https://www.timeshighereducation.com/world-university-rankings/2016/world-ranking#!/page/0/length/25>

In Figure 1, the data for the top 200 universities was sorted by the response variable total score, which determines the overall ranking of the university.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | **Usage** | **Description** | **Type** | **Range** |
| University | Identifier | University name | String | N/A |
| Total Score | Response | Total score, used to determine rank | Decimal | 0 to 100 |
| Teaching | Explanatory | University score for teaching (the learning environment) | Decimal | 0 to 100 |
| International | Explanatory | University score for international outlook (staff, students, research); ability to attract undergrads, postgraduates, and faculty from all over the planet is key to it’s success on the world stage | Decimal | 0 to 100 |
| Research | Explanatory | University score for research (volume, income, reputation) | Decimal | 0 to 100 |
| Citations | Explanatory | University score for citations (research influence); indicator at the role in spreading new knowledge and ideas | Decimal | 0 to 100 |
| Income | Explanatory | University score for industry income (knowledge transfer); this category seeks to capture the ability to help industry with innovations, inventions, and consultancy | Decimal | 0 to 100 |
| # Of Students | Explanatory | Total Number of Students; the number of full time equivalent students at the University | Integer | >0 |
| Student/Staff | Explanatory | # of Students to Staff; ratio of full time equivalent students to the number of academic staff, those involved in teaching or research | Decimal | >0 |
| International | Explanatory | Percentage of International Students; students originating from outside of the country of the University | Percent | 0 to 100 |
| Female/Male | Explanatory | # of Females to # of Males | Decimal | >0 |
| *Figure 1. List of all variables.* | | | | |

The next five variables, Teaching, International, Research, Citations, and Income were used by The Times Higher Education World University Rankings to determine total score. However, after some research on the methodology, we note that the non-equal weighting used in 2016 by The Times Higher Education World University Rankings to determine total score as shown in Figure 2.

|  |  |
| --- | --- |
| **Category** | **Weighting** |
| Teaching | 30% of overall score |
| International | 7.5% of overall score |
| Research | 30% of overall score |
| Citations | 30% of overall score |
| Income | 2.5% of overall score |
| *Figure 2. Weighting for each category.* | |

The next four variables called ‘Key Statistics’ are: # of Students, Student/Staff, International, and Female/Male, and are not part of the criterion used in the calculation of the total score. The Female/Male variable was converted from a ratio to a decimal in order to allow for analysis (i.e. a value of 1 would indicate equal number of females and males). The ‘Key Statistics’ were provided by the universities themselves and represents data from the 2013 academic year and may vary from earlier or subsequent years.

# Exploratory Data Analysis and Screening

We first begin by exploring the completeness of data. As shown in Figure 3, the summary statistics indicate of the five variables used to determine the ranking; there are five universities that did not contain observations for the variable ‘income’. The specific five universities with the missing values are listed in Figure 4. Also, we note that there are a handful of missing values for the variables number of students, student/staff ratio, international students, and female/male ratio. We can proceed, as the dataset is for the most part complete, understanding that even for a professional publication, data collection maybe difficult or unavailable given such a wide variety of institutions.

Examining the summary statistics, we notice that the ranges of values for explanatory variables num\_students and female\_male\_ratio differ by several factors, and thus we must standardize the data before conducting the PCA analysis to determine which variables have the highest influence.

Next, we perform a visual examination of the scatterplot matrix to determine if there is evidence of nonlinear relationships. Figure 5 confirms there is no evidence of nonlinear relationships.

Finally, we examine the Pearson correlation matrix to examine if the variables have high correlation with each other. From a high-level, PCA is about identifying variables that describe a similar construct; we can investigate the correlation matrix to get a glimpse of which explanatory variables maybe related. Judging from Figure 6, we find that there is a high correlation between income/num\_students and international/student\_staff\_ratio. These are actually the smallest correlations… I think you are looking at the pvalues. These correlations will help make sense of the PCs later.

|  |  |
| --- | --- |
| **Summary Statistics** | **Incomplete Data** |
|  | Columbia University  University of Bonn  Purdue University  University of Florida  Yeshiva University |
| *Figure 3. Summary statistics for response and explanatory variables.* | *Figure 4. Universities with missing ‘income’ variables.* |
| **Scatter Plot Matrix** | |
|  | |
| *Figure 5. Scatterplot matrix of response and explanatory variables. Visual confirmation shows some multicollinearity but no nonlinear relationships.* | |
| **Correlation Matrix** | |
|  | |
| *Figure 6. Pearson Correlation Matrix. The red boxes highlight the variables with high multicollinearity.* | |

# PCA Results

As we have fulfilled the suitability for PCA analysis, we can continue the analysis using the correlation matrix. We examine the Scree Plot as shown on Figure 7. Unfortunately, we do not see a steep drop, so there is not clear-cut answer at this point as to the number of principal components to retain. However, we do note that the first four principal components account for approximately 80% of the cumulative variance as shown in Figure 8 and 9. We note the close grouping of research/teaching in Principal Component 1 and international\_students/international in Principal Component 2. This is great!

|  |  |
| --- | --- |
| **Scree Plot** | **Cumulative Variance** |
|  |  |
| *Figure 7. Scree Plot showing no sharp drop off.* | *Figure 8. Cumulative variance as explained by each principal component.* |
| **Eigenvalues** | **Principal Components 1 & 2** |
|  |  |
| *Figure 9. Eigenvalues of the Correlation Matrix.* | *Figure 10. First two principal components.* |

|  |
| --- |
| **Eigenvectors** |
|  |
| *Figure 11. Eigenvectors representing the loadings for the nine principal components.* |

We are now at the point where we examine the Eigenvectors in Figure 11 to see if we can associate real-world meaning to the principal components.

The first principal component associates the traditional measurements of teaching and research. That is, both increase in teaching and research have a positive association with the overall score, which isn’t too surprising. The second principal component groups international outlook and international students, having a negative association with the total score. Somewhat surprising at first glance, but remember the weighting for international is only 7.5%, and the negative association in this principal component confirms that teaching, research, and citations are the main driving factors. The third principal component shows almost perfect contrast between citations and student to staff ratio. That is, citations decrease as the number of students to staff increases. The fourth principal component groups number of students and female to male ratio. That is the tendency to have a greater female to male population as the number of students increase.

Based on these rough interpretations, it would seem reasonable to proceed with fitting a regression model, which according to Figure 9, should account for 77.6% of the variation in the data.

# Model Selection and Regression Analysis

By definition, principal components are orthogonal, thus we can represent the regression equation as such:

Where are principal components.

|  |  |
| --- | --- |
| **Parameter Estimates** | **Analysis of Variance** |
|  |  |
| *Figure 12. Parameter estimates comprising of the first four principal components.* | *Figure 13. Results of fitting the first four principal components.* |

After fitting the first four principal components, we have a good overall fit (F=945.36, p<0.0001), indicating the first four principal components are significant as shown in Figure 13. The R2 value of 95.65% indicates about 4.35% of the variation in the data is not explained by the data. As show in Figure 12, there are significant effects for all four principal components (t=58.14, P<0.0001, t=11.88, P<0.0001, t=-9.56, P<0.0001, t=13.01, P<0.0001). As expected, the scatterplot matrix in Figure 14 shows no evidence of a nonlinear relationship to the response variable.

Now we move on to check the assumptions. Overall, the residual errors look good, as show in the diagnostic plots of Figure 15. There is no evidence to suggest interdependence for the predicted value plots. There is no evidence of non-normal distribution in the quantile and residual histogram plots. Finally, the residual plots for the principal components in Figure 16 show no evidence of non-constant variance.

Next, we examine the Outliers and Leverage points and note that we have four observations that are high Cook’s D value and Outliers and Leverage points, they are 108, 161, 175, and 189. These correspond to the Universities: University of Mannheim, Lomonosov Moscow State University, University of Konstanz, and Arizona State University. There is no reason to exclude these observations, as we have no reason to expect that these four universities are not part of the model.

|  |  |  |
| --- | --- | --- |
| **Scatter Plot Matrix** | **Fit Diagnostics** | **Residual by Regressors** |
|  |  |  |
| *Figure 14. Scatterplot matrix of the response variable and first four principal components.* | *Figure 15. Diagnostic plots for the fitted model using all four principal components.* | *Figure 16. Residuals for each principal component.* |

We have no good reason to simplify the model any further, and thus our regression equation for the final model remains with all four principal components intact:

# Conclusions

In the quest to find the variables that have the greatest influence of university rankings using principal component analysis, the findings must be framed in an appropriate manner. While a prediction model is not really useful, the discovery of how variables were grouped in the principal components yielded some interesting results.

The first principal component associates the variables teaching and research. It is not surprising as these two fundamental areas reinforce each other in an institution of higher learning and contribute to a total higher score. We note that both teaching and research were assigned a weight of 30% in the original methodology, but surprisingly citations is not part of the first principal component as it was also assigned 30% weighting. **I was surprised by the fact that it was not in the PC with the research. But after review of the correlation matrix in Fig 6 it shows that research and citation have relatively small correlation. So since 30% of the score is built from the citations but has relatively low correlation with the other variables, I am not surprised that is seems to have carved out it’s own PC. (PC 3)**

The second principal component groups international outlook and international students, having a negative association with the total score. At first glance, this may seem somewhat surprising given the international outlook one of the five published criteria as being essential for a university (“ability to attract undergrads, postgraduates, and faculty from all over the planet is key to it’s success on the world stage”). However, taking a closer look at the weighting methodology, we see that international outlook is given only 7.5% weighting. We confirm this and note that many of the top ranked schools have relatively low international outlook scores. For example, California Institute of Technology is ranked at the top in 2016, but only has an international outlook score of 64 out of 100, below the average of all universities at 66.79. Interestingly, this principal component seems to provide some circumstantial evidence to support the criticism levied against The Times Higher Education World University Rankings that they are in fact undermining non-English speaking institutions. Please refer to: <https://en.wikipedia.org/wiki/Times_Higher_Education_World_University_Rankings>

The highest non-English speaking ranked school at #9 is Swiss Federal Institute of Technology in Zurich, Switzerland with a score of 97.9 on international outlook. Given a higher weighting for international outlook would in fact put this school as a contender for the top seven spots.

The third principal component shows almost perfect contrast between citations and student to staff ratio. One interpretation of this principal component is that having fewer students per staff increases the quality of research and therefore leads to more citations. In fact, the top seven-ranked institutions have an average of 9.2 students to staff ratio and an average citations score of 99.19 out of 100. The student to staff ratio for all the universities is 17.2 with an average citations score of 82.915 (see Figure 2). We should also note here that having many citations might not have practical significance to the quality of undergraduate study.

The fourth principal component groups number of students and female to male ratio. As the number of students’ increase, so does the female to male ratio. While this principal component does not contain any variables that were used to calculate the original total score, it does provide evidence that the ranking system is including schools that teach majors that not only include traditionally men dominated fields such as engineering.

While principal components one, three, and four were interesting, principal component two revealed the most surprising result. It’s difficult to assess the how the weightings are determined, but there is definitely a bias towards teaching/research/citations (Because these make up 90% of the score … mathematically.) Finally, we should address some of the practical issues with these rankings as they are simply too broad when it comes to specific fields of study. A student that is interested in studying Music would have no reason to apply to California Institute of Technology, just as a student interested in studying Mathematics would not apply to Julliard. The best advice to an aspiring college student may well be to take these University rankings with a grain of salt and gain the perspective that these rankings are limited in their scope of analysis.

**Hi James,**

**I am torn between loving the analysis and then asking myself … am I surprised as we are essentially regressing the score against variables that we know are significant because we know the equation of the score. My intuition is that if you did normal MLR you would find very much the same result. However, maybe not. We know that the PCs are independent of one another while the original variables are not. So grouping teaching and research in PC 1 allows us to conduct a regression free of multicollinearity and inflated variances.**

**I do like that your analysis also considered variable that were not used to construct the score and that you were able to gain some inference from that analysis.**

**In conclusion, I think PCA is a good method to answer your QOIs. You really got me thinking.**

**Your analysis was very well organized and presented. In addition I was particularly impressed with your interpretation and explanation of what the inference on the PCs really meant.**

**I would have no problem giving your analysis a super high score and I hate to end on a negative but the interpretation of the correlation table above is the only glaring deficiency I can see in your paper. Please look back and make sure that makes sense and makes sense with respect to the loadings in the PC table (Figure 11).**

**So I won’t end on a negative note: Overall .. you did an excellent job!**

**95%**

# APPENDIX

/\* Original Data Load \*/

PROC IMPORT OUT=WORK.PROJECT2

DATAFILE="C:/Documents and Settings/james/My Documents/My SAS Files/9.4/MSDS6372/Project2/timesData3.csv"

DBMS=CSV REPLACE;

GETNAMES=YES;

DATAROW=2;

RUN;

PROC PRINT DATA=WORK.PROJECT2; RUN;

PROC MEANS DATA=WORK.PROJECT2 N MEAN MEDIAN STD MIN MAX;

VAR total\_score teaching international research citations income num\_students student\_staff\_ratio international\_students female\_male\_ratio;

RUN;

PROC SGSCATTER DATA=WORK.PROJECT2;

MATRIX total\_score teaching international research citations income num\_students student\_staff\_ratio international\_students female\_male\_ratio;

RUN;

PROC CORR DATA=WORK.PROJECT2 PLOTS(MAXPOINTS=NONE) = MATRIX(NVAR=ALL);

RUN;

PROC PRINCOMP PLOTS=ALL DATA=WORK.PROJECT2 OUT=PCA;

VAR teaching international research citations income num\_students student\_staff\_ratio international\_students female\_male\_ratio;

RUN;

PROC REG DATA=PCA outest=PCARESULT plots(label) = (rstudentbyleverage cooksd);;

MODEL TOTAL\_SCORE=PRIN1-PRIN4;

RUN;

PROC CORR DATA=PCA PLOTS=MATRIX(HISTOGRAM);

VAR TOTAL\_SCORE PRIN1-PRIN4;

RUN;