Qin Liu, Marisol Arita

Stat 311 – Regression Analysis

University case study

Analyzing Institutional and Demographic Factors Influencing College Applications Rates

Introduction

The objective of this study is to explore the factors influencing college graduation rates across various institutions in the United States. By examining a dataset of 777 universities and colleges, we aim to uncover relationships between acceptance rates and institutional, demographic, and financial characteristics. These findings will provide insights into how different attributes contribute to application rates.

We will also explore how variations in college affordability, represented by tuition and associated costs, influence application rates. Furthermore, we will assess the role of faculty qualifications—measured by the percentage of faculty with Ph.D.’s—in shaping number of applications.

This research leverages statistical methods to build models addressing these questions, enabling the identification of key predictors and their significance. Through this investigation, we aim to provide actionable insights for institutions seeking to enhance application rates and improve student experiences.

Data and Methods

DATA

We used a dataset of demographic characteristics and tuition information from 777 universities and colleges across the United States. The entries were provided in a csv file. We performed all statistical analysis using the regression analysis software JMP.

These are the first 24 data entries that show an example of what the csv file contains:

A table with numbers and text

Description automatically generated

VARIABLES

**Variables in the dataset:**

Apps = Number of applications received

UniversityName= Name of the university

Private = Public/private indicator

Accept = Number of applicants accepted

Enroll = Number of new students enrolled

Top10perc = New students from top 10 % of high school class

Top25perc = New students from top 25 % of high school class

F.Undergrad = Number of full-time undergraduates

P.Undergrad = Number of part-time undergraduates

Outstate = Out-of-state tuition

Room.Board = Room and board costs

Books = Estimated book costs

Personal = Estimated personal spending

PhD = Percent of faculty with Ph.D.’s

Terminal = Percent of faculty with terminal degree

S.F.Ratio = Student/faculty ratio

perc.alumni = Percent of alumni who donate

Expend = Instructional expenditure per student

Grad.Rate = Graduation rate

We used **apps** as our response variable (y) and the remaining variables as independent variables to be analyzed and filtered to find the best fit for our model and the appropriate predictor variables (x). Since the **UniversityName** column contains all unique row entries, we concluded that having so many unique levels adds unnecessary complexity, and thus it will not be included as an independent variable in the model. We also performed some data transformation to convert the Top10perc, TOP 25perc, PhD, Terminal, perc.alumni, Grad.Rate numbers to a decimal (50% -> .50 i.e.) and give new variables names as follows:

TOP10%, TOP 25%, PHD%, Terminal%, Alum%, Grad Rate% and saved in a new file titled ‘University Data\_updated.csv’.

Exploratory Data Analysis

We performed exploratory analysis to see different patterns within the data.

Examples:

Apps



Accept



There is one outlier for number of applications and acceptances on row 484 – Rutgers at New Brunswick, which has 48094 applications, and 26330 applicants accepted. These numbers seem high considering the upper end of applications received lie between 10,000 – 20,000 applications for Ivy League Universities such as Yale and University of Pennsylvania, and significantly less acceptances. It seems unlikely 26,000 applicants were admitted as freshman class. We foresee this data entry will be an influential data point but will perform some tests to confirm.

Initial Model:

We employed the All-Possible-Regression Selection Method to evaluate the 17 viable variables. This allows us to screen for the important independent variables.

A screenshot of a computer

Description automatically generated

A graph with numbers and symbols

Description automatically generated

A graph with numbers and symbols

Description automatically generated

A graph with numbers and symbols

Description automatically generated

We picked the model with one of the highest R2 values, with one of the lowest Cp values that we believe would create an uncomplicated, yet powerful model. We can see that 11 independent variables should be included in the group as most important variables. By analyzing metrics such as , RMSE, , AIC and BIC, we identified a set of potentially significant predictors prioritizing a balance of strong predictive power without adding unnecessary complexity or the risk of overfitting and developed the initial model as shown below:

We performed tests to assess the initial fit of the model. F-test to assess whether the regression model explains a significant portion of the variability in the response variable compared to the variability due to error, as well as individual T-tests to assess the validity of each predictor.

F-test

A screenshot of a calculator

Description automatically generated

A low p-value of < .0001 which is less than a .05 significance level indicates a strong and statistically significant model. A high F ratio also denotes that the model explains more than random error.

Individual T-tests

When performing a t-test on each of the variables we find that when using a significance level of .05, all predictor variables we chose have p-values less than a .05 significance level, so we reject the null hypothesis and thus we know they are statistically significant in the model and predicting y.

A table with numbers and text

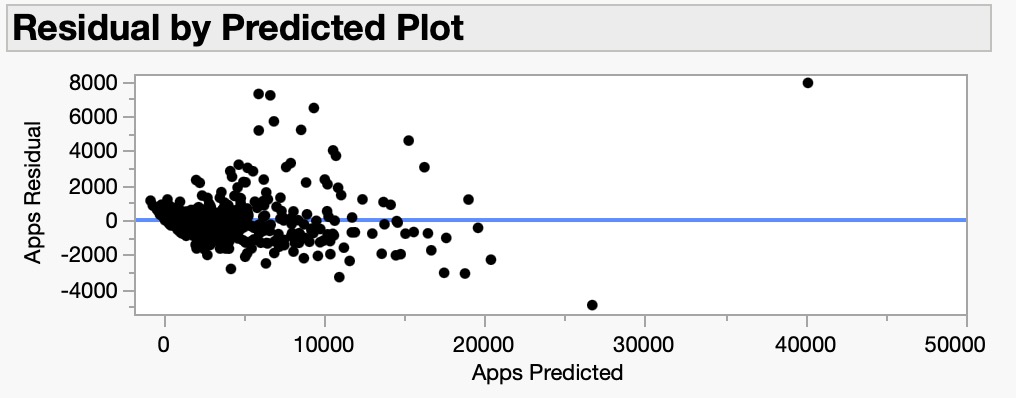
Description automatically generated

A white box with black text

Description automatically generated

R2 and R2 adjusted are 92.88% and 92.78% respectively, meaning 92% of the variability is explained by the model which suggests a strong fit.

RESIDUAL ANALYSIS



We check to make sure the assumptions are met. The residual plot however showed residuals aren’t fully random in pattern. Residual variability seems to increase as the predicted values increase. This suggests heteroscedasticity which violates one of the assumptions of linear regression. We can see some outliers that may also be influential in regards to the model and skewing results.

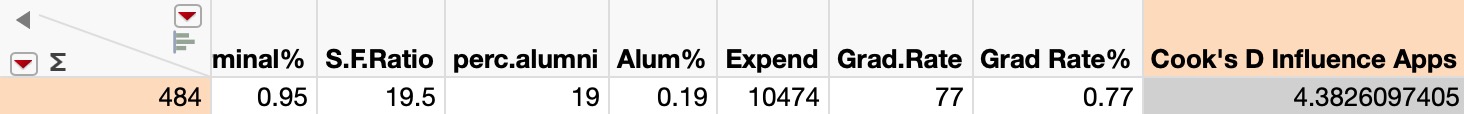
A screen shot of a graph

Description automatically generated



We can see in the last two graphs that there are potentially influential outliers. The residual **apps** vs. x’s also suggests there may be some interactions or non-linear relationships we didn’t capture with the initial model. In this normal quantile plot the residuals deviate from the line especially at the ends (or tails). This deviation indicates that the residuals are not evenly distributed, which violates an assumption in linear regression.

To fix the problem we excluded two data entries (row 484 and row 462) which are outliers in the model skewing results as can be observed in the residual by predicted plot, as well as row 484 being > 1 and having a very high number when looking at Cook’s D influence. We also performed a square root transformation on the response variable (y), to stabilize the model and variance.



Final model:

Our final model is a multiple linear regression model with a square root transformation applied to the response variable (y) which is **apps** or number of applications received by a respective university. We employ a combination of continuous and categorical variables to achieve the best fit model. We also employed interaction terms and quadratic terms to encapsulate the full complexity of the research question as a simpler regression model would not achieve on its own.

The square root transformation was chosen as a way to stabilize the variance and therefore improve the normality of the residuals as seen below:

= 11.53-0.0000006 +0.0002

A close-up of a list of fit

Description automatically generatedA screenshot of a calculator

Description automatically generated

A screenshot of a calculator function

Description automatically generated

In the summary of fit table, we observe R2 and R2 adjusted are both high and have improved since the initial model. At 94.3% and 94.2% when adjusting for number of predictors, they indicate that the variability is mostly explained by the predictors chosen for the model and have a better fit with the standardizing of response variable y. The model in this case has a strong fit. The root mean square error (RMSE) is 6.37 which is low and indicates strong accuracy when it comes to predicting.

When performing individual t-tests we see that Private, Accept, top 10%, Room and Board, Grad Rate are very impactful on the model.

When reading the ANOVA table we observe a low p-value which confirms that the model is statistically significant when predicting the number of applications (y). A higher F ratio also suggests that the predictors (x) have a strong relationship to the response variable y.

A graph of normal quantile plot

Description automatically generated

The updated residual normal quantile plot has also improved, and the data points are closer to the line and normal distribution.

Conclusion

This study investigated the institutional and demographic factors influencing the number of applications received by U.S. universities. Using a comprehensive dataset and advanced regression techniques, we identified key predictors of application rates, including institutional type, affordability measures (tuition and associated costs), and student-to-faculty ratios. Our final model demonstrated strong predictive power with an adjusted R2 of 94.2%, indicating that the variability in applications is largely explained by the selected predictors.

The analysis highlights key factors influencing college application trends, such as academic prestige (top 10% of high school class) and affordability (room and board costs). Surprisingly, a higher percentage of faculty with Ph.D.s appears to negatively impact applications, potentially reflecting applicant preferences or perceptions. These insights can help institutions strategically enhance enrollment by balancing costs, emphasizing academic achievements, and addressing perceptions of faculty composition. However, we noted that outliers, such as Rutgers University, significantly impacted the initial analysis, necessitating their exclusion and the use of a transformation to stabilize variance.

In summary, universities aiming to increase application numbers should prioritize affordability, target students comprising of the top 10% of graduating classes, and emphasize university graduation rates. By addressing these factors strategically, institutions can enhance their appeal to prospective students, ultimately driving growth and diversity in their applicant pools.