ST 704 Final Project Proposal

Carter Hall, Connor McNeill, and Miles Woollacott

March 22, 2024

1 Introduction

Data Overview

Our dataset is a conglomeration of resources related to college *tuition*, *diversity*, and potential *salary* after graduation across the United States over a 30-year period. The data[1] come from a number of resources including the **National Center for Education Statistics**. Per university, the following metrics are recorded:

1. Tuition, fees, and academic information (e.g., private/public institution, location)
2. Diversity metrics (demographic statistics)
3. Cost to students
4. Graduation rates and salary data

With almost $2 trillion in student-loan debt across all debt-holders in the United States[2], it is no secret that the cost of college is a burden on many American families. As more students apply to college, and with recent judicial results related to Affirmative Action legislature, understanding the shifting landscape of post-secondary education in the United States becomes quite the interesting pursuit. It is the focus of this project to analyze what factors make students of certain universities successful and how college education across America can be enhanced.

Research Question

Our primary research question is

> How do tuition and socioeconomic factors impact success post-graduation?

[1] https://www.kaggle.com/datasets/jessemostipak/college-tuition-diversity-and-pay/data
[2] https://www.forbes.com/advisor/student-loans/average-student-loan-debt-statistics/

With this, we hope to explore the relationship between predictors such as tuition, the type of school (private vs public), first-year retention rates, and socioeconomic factors such as percentage of students from underrepresented populations, family income, and the amount of grant/aid being provided to students. Success indicators to be explored may include graduation rates and income post-graduation and mid-career.

We will explore and consider multiple methods as part of this project. Models we are considering include multiple linear regression, penalized regression, nonlinear regression, and generalized linear models such as logistic regression. Our plan for the more advanced model is to explore utilizing a Bayesian Logistic Regression model, but we may modify the type of model considered in the Bayesian context based on what success indicators we choose to consider.

2 Project Timeline

Task	Target Completion	Deadline
Data Cleaning	3/28	
Exploratory Data Analysis	3/31	
Method Implementation in R	4/5	
Interim Report	4/5	4/7
Writeup & Analysis	4/13	
Final Report	4/16	4/21
Presentation	4/19	4/21

3 Description of Group Responsibilities

The following is a quasi-breakdown of the responsibilities of each group member, taking into account perceived individual strengths. By no means is this a strict partition of the project duties, as we *all* are statisticians and should therefore be 'experts' in the aforementioned responsibilities.

1. **Carter**: Implementing models (including the Bayesian method as needed), quality control of codebase, assistance on exploratory data analysis and writeup.
2. **Connor**: Data cleaning, implementation of Bayesian method, presentation preparation, assistance on writeup and analysis
3. **Miles**: EDA, model implementation (for the remaining models), assistance on presentation preparation.