STA302/1001 Methods of Data Analysis 1 (LEC0101)
Final Project - due June 25, 2020 at 11:59PM EST on Quercus

Note on the deadline:

The above is a **HARD** deadline – this means that there will be NO extensions granted under any circumstances, and no final project will be accepted once the submission deadline has passed. Please make sure that you start the submission process early so that your project is graded.

Goal of the Project:

Admission rates of colleges and universities in the United States can be drastically different from one another and this can be due to a variety of reasons, such as lack of facilities, cost, or even the demographics of the applicants themselves. The purpose of this project is to investigate which of the factors/variables in the provided dataset best explains the variation observed in admission rates for 1508 colleges and universities in the U.S. While the primary use of the model you will built is for description and understanding of the factors affecting admission rates, it may also be used to predict admission rates if certain changes are made to the admission practices of schools. Therefore, the model you decide is best for explaining this variation must be both simple enough for various stakeholders to understand, but also complicated enough and with all necessary properties required to make good predictions. You may use any and all techniques and methods learned throughout the term in this course (Weeks 1-6) to develop your model, but you must justify their use and use them correctly.

How to present your results:

Once you have decided upon the 'best' model to fulfill the goal of the project, you must write up a short scientific report. There should be 4 main sections of your report:

- Introduction section: where you introduce the purpose and relevance of the project
- Methods section: where you describe and explain the methods, tools and techniques used to arrive at your final model
- Results section: where you present a description of your study sample, important results that led you to make crucial decision in building your model, and the final model and any other important results
- Discussion section: where you interpret your final model and describe why it answers the research question and why it is important, as well as discuss any limitations that still exist based on your results.

You may use tables and plots to help present your results, but they must be relevant and well-thought out so as to convey as much information as possible without being too overwhelming or confusing. When explaining your methods and results, try to avoid just stating that you used a specific method, but add an explanation for why it is the correct tool for the job at hand. See the

rubric on the Quercus assignment page for more information regarding the various report components.

Technical Requirements of the Final Report:

Your report should be typed using whatever software you prefer but must be saved and submitted in one of these standard file types: .doc, .docx, .pdf. Your report must meet the following requirements:

- Font: 12-point font in a style similar to Times New Roman
- Spacing: single-spaced
- Word count: up to 1500 words in total (not including captions on figures and tables)
- Number of tables/plots/figures in main report: 5 in total, but you may use any combination of tables and figures
- Number of tables/plots/figures in an appendix: up to 3 additional tables/figures but they should only be included if they are relevant to the analysis and are referred to in the main text.

Some additional submission requirements are:

- A separate file containing your R code in case we need to verify your results – please make sure that it contains the code that led to your final results.
- A signed copy of the **Academic Integrity Acknowledgement Form**, attesting that you have completed the project individually and have not helped or been helped by another student or unauthorized person. This form can be found on the Quercus assignment page.

Description of the Dataset:

The dataset contains 30 variables (1 response and 29 other variables) on 1508 colleges and universities in the United States. This dataset was derived from a larger collection of measures on schools in the United States (https://collegescorecard.ed.gov/data/). The variables collected here can be grouped into three categories: school identifiers, school characteristics, and student population/applicant characteristics. Below is a summary of the variables in the data:

Variable Name	Characteristic	Detailed Description of Variable
UNITID	Unit ID for institution	Unique numerical identifier of institution
INSTNM	Name of institution	Actual name of institution (text)
STABBR	State Postcode	Two-character label for state in which institution is located (i.e. factor variable)
NUMBRANCH	Number of branch campuses	Numeric value for number of satellite campuses/affiliations • 1 indicates there is only one main campus

CONTROL	Control of institution	Factor indicating whether the institution is public (1), private non-profit (2), or private for-profit (3)
REGION	Regional location of institution	Factor variable with the following levels: • 0 = US Service Schools • 1 = New England (CT, ME, MA, NH, RI, VT) • 2 = Mid East (DE, DC, MD, NJ, NY, PA) • 3 = Great Lakes (IL, IN, MI, OH, WI) • 4 = Plains (IA, KS, MN, MO, NE, ND, SD) • 5 = Southeast (AL, AR, FL, GA, KY, LA, MS, NC, SC, TN, VA, WV) • 6 = Southwest (AZ, NM, OK, TX) • 7 = Rocky Mountains (CO, ID, MT, UT, WY) • 8 = Far West (AK, CA, HI, NV, OR, WA) • 9 = Outlying Areas (AS, FM, GU, MH, MP, PR, PW, VI)
HBCU	Historically Black College and University	Factor variable where levels are no (0) or yes (1)
PBI	Predominantly Black University	Factor variable indicating whether currently predominantly serves black community (1) or not (0)
TRIBAL	Tribal college and university	Factor variable indicating whether university for native American tribes (1) or not (0)
HSI	Hispanic-serving institution	Factor variable indicating whether university serves Hispanic communities (1) or not (0)
WOMENONLY	Women-only College	Factor variable indicating women-only institution (1) or not (0)
ADM_RATE	Admission rate (Response)	Defined as number of admitted students out of total number of undergraduate applications
COSTT4_A	Average cost of attendance per academic year	average annual total cost of attendance, including tuition and fees, books and supplies, and living expenses for all full-time, first-time, degree/certificate-seeking undergraduates
AVGFACSAL	Average faculty salary	Calculated as the total salary outlays divided by the number of months worked for all full-time nonmedical instructional staff.
PFTFAC	Proportion of full-time faculty members	Calculated as the number of full-time nonmedical faculty divided by the total number of nonmedical faculty. Proportions are expressed as decimals

PCTPELL	Percentage of undergraduates receiving Pell grant (financial aid)	calculated as the quotient of the number of Pell grant recipients divided by the count of all undergraduates for either a fall enrollment cohort. Proportions are expressed as decimals
UG25ABV	Percentage of undergraduates aged 25 and above	A numeric variable
INC_PCT_LO	Percentage of aided students whose family income is between $0-$30,000	A numeric variable indicating whether financial aid is awarded primarily to low income students
PAR_ED_PCT_1STGEN	Percentage of first-generation students	A numeric variable
FEMALE	Proportion of student body that is female	A numeric variable
MD_FAMINC	Median family income of students	A numeric variable
PCT_WHITE	Percent of the population from students' zip codes that is White, via Census data	Numeric variable describing demographics of student's home neighbourhood.
PCT_BLACK	Percent of the population from students' zip codes that is Black, via Census data	Numeric variable describing demographics of student's home neighbourhood.
PCT_ASIAN	Percent of the population from students' zip codes that is Asian, via Census data	Numeric variable describing demographics of student's home neighbourhood.
PCT_HISPANIC	Percent of the population from students' zip codes that is Hispanic, via Census data	Numeric variable describing demographics of student's home neighbourhood.
PCT_BA	Percent of the population from students' zip codes	Numeric variable describing education levels of student's home neighbourhood.

	with a bachelor's degree over the age 25, via Census data	
PCT_GRAD_PROF	Percent of the population from students' zip codes over 25 with a professional degree, via Census data	Numeric variable describing education levels of student's home neighbourhood.
PCT_BORN_US	Percent of the population from students' zip codes that was born in the US, via Census data	Numeric variable describing demographics of student's home neighbourhood.
POVERTY_RATE	Poverty rate, via Census data	Numeric variable describing income levels of student's home neighbourhood.
UNEMP_RATE	Unemployment rate, via Census data	Numeric variable describing income levels of student's home neighbourhood.