Statistical Methods in Education Research
Vanderbilt University
Prof. Sean P. Corcoran

Due December 6, 2022

Statistical Project

Instructions: This assignment is an opportunity for you to apply your knowledge of statistics to real-world data. You may choose to: (1) provide a written report on the "default" project described below, or (2) provide a written analysis using appropriate data you supply yourself. I will provide data and instructions for the former, and guidance on the latter. Doctoral students who would like to gain experience with a dataset relevant to their own research are encouraged (but not required) to choose the second option. If you choose option (2), you may craft your research question as you see fit—but the quantity and rigor of the analysis must match or exceed that required under option (1).

Your report should be no more than 15 pages, including tables and graphs. All statistical work should be completed using Stata. The paper should be double-spaced, with standard margins and font size. You may wish to include some tables and graphs in the main body of the paper, and relegate others to an appendix for readability. Appendix pages will not count toward the page limitation. Be sure to appropriately reference all figures and tables in the text itself. (E.g., "Table 2 shows that..." and "In Appendix Table 1, I report descriptive statistics for...").

Your report should read like a narrative supported by statistical analysis, and not like a list of problem set answers. You may use any techniques learned in class to address your research questions, unless otherwise specified (in option 1). Show some variety in your use of methods; don't use the same techniques to answer every question. Be selective about the tables, graphs, and statistics you include. Choose the method that most clearly communicates the result you are trying to convey, and do not overwhelm the paper with output.

Your grade will depend on the extent to which you: (1) provide an analysis that fully addresses your research questions (in option 1, this means including a complete and correct response to each of the questions below), (2) choose the appropriate statistical methodology to respond to each question, (3) write a *clear and concise* analysis that flows logically, interprets statistical results correctly, and integrates your statistical work into the prose. The end of this document includes a rubric that I will follow when grading the project.

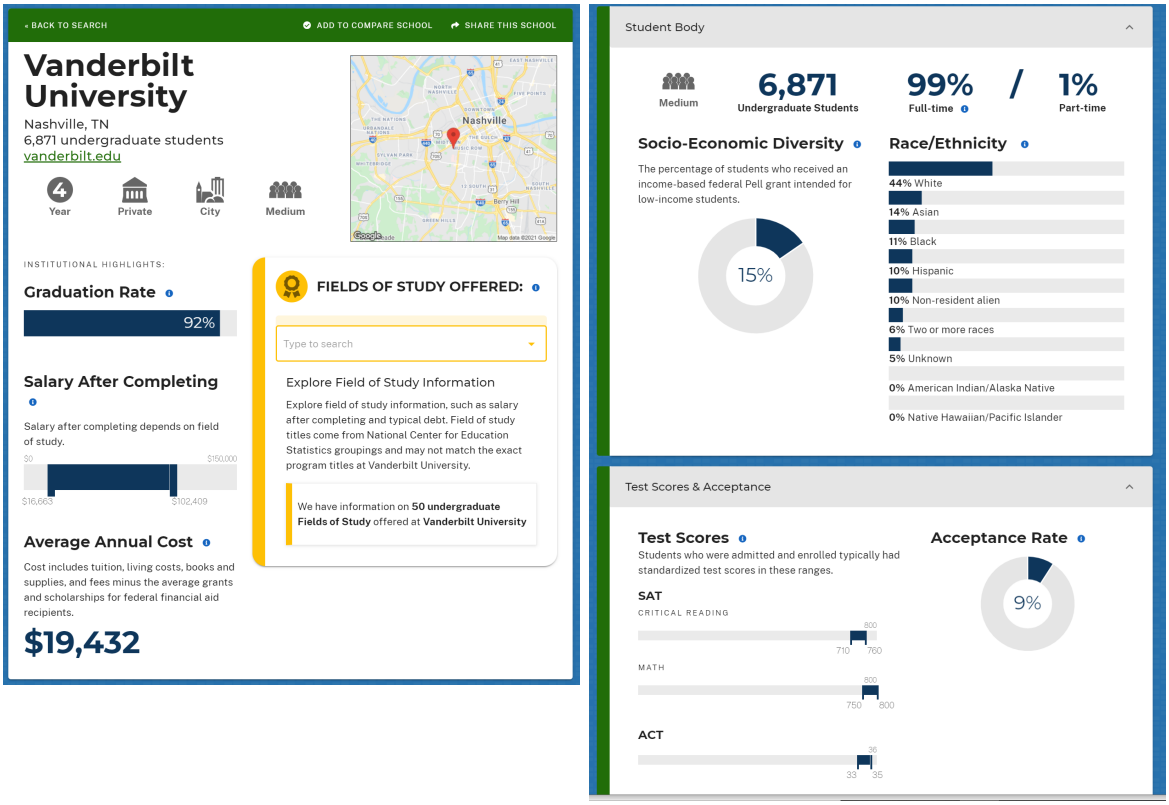
Sample projects will be made available on Github as examples of past work. Please submit your completed project as a PDF document via email to `sean.corcoran@vanderbilt.edu`. Use your last name and the words "Statistical Project" as the filename (e.g. *Smith Statistical Project.pdf*).

Option 1: Default Project

This project will use College Scorecard data to examine variation across public, private, and for-profit colleges and universities in selectivity, cost of attendance, and financial aid.

Background

The College Scorecard (`https://collegescorecard.ed.gov/`) was launched by the Obama administration in 2015 with the aim of making information about the cost and quality of colleges more accessible to families (Hurwitz & Smith, 2018). The underlying data consists of hundreds of elements gathered from a variety of sources. The portal linked above compiles these data into a few easy-to-understand dashboards, including cost, graduation and retention, financial aid and debt, salary by field of study, student body, and test scores / acceptance rates. A snippet of the scorecard for Vanderbilt is shown below:



For this project I created an extract of the College Scorecard data for 2019-20. The data and .do files used to create this dataset are posted on Github. (The .do files are just for your information).[1] The extract includes only currently-operating colleges and universities that predominantly award bachelor's degrees. Community and technical colleges are excluded,

[1]The raw data were downloaded from the following website: `https://data.ed.gov/dataset/college-scorecard-all-data-files-through-6-2020/resources`. Full documentation is available here: `https://collegescorecard.ed.gov/assets/FullDataDocumentation.pdf`

as are institutions that exclusively award graduate degrees (e.g., theological seminaries, law schools). Each observation is a college or university operating in 2019-20, with aggregate data for that institution reported. Some elements in the dataset (e.g., related to financial aid) pertain only to students at the institution who received federal aid such as grants or loans.

Most variables in this dataset will be self-explanatory based on their variable labels. However, on Github I have posted a data dictionary that provides detailed definitions of each variable. See the "Institution_Data_Dictionary" tab in this file. The variables are listed under VARIABLE_NAME. Value labels are provided for categorical variables. See the "notes" column for more details on how variables are defined.

Questions to Address

Your report will be a descriptive analysis of U.S. colleges and universities in the 2019-20 College Scorecard. You should address all of the below questions in your report. You do not necessarily have to cover these in the order listed, but I have attempted to order them in a sensible way. If you have another way you would like to tell the story, feel free—just be sure to address all of the assigned questions. Further, you are not confined to the questions listed here; if there are additional interesting questions you would like to address with the data, you are encouraged to do so (as long as you have space)! There are many more data elements in this extract—and in the original data—that are not addressed here.

1. Introduce your analysis by describing the main objective of your report and the data you use. (Refer to the text above and to any of the linked resources to get more information about the data). Provide a brief summary of what your analysis finds.

2. Provide some basic information about the colleges and universities in your dataset, including the following:

    - How many institutions are in the data, and where are they located (e.g., states and/or regions)?
    - How many students (in total) are served by the institutions in this data? Report separate totals for certificate/degree-seeking undergraduates, non-degree seeking undergraduates, graduate students, and all students combined.
    - What percentage of institutions are public, private not-for-profit, and private for-profit? Provide a cross-tabulation showing the percent of institutions by sector within each region.
    - Across institutions, what is the mean undergraduate enrollment (for institutions with at least 10 undergraduates)? What is the mean percent of undergraduates that are women?

3. Provide a descriptive analysis of institutional selectivity:

- Provide measures of central tendency and dispersion for the following variables: admissions rate (use *adm_rate_all*), midpoint of the SAT verbal for admitted students, midpoint of the SAT math for admitted students, and the average SAT equivalent score for admitted students (use *sat_avg_all*). Explain in words what you find; don't just report results in tables.
- Show the distribution of admissions rates overall (*adm_rate_all*), and separately by institutional control (public, private not-for-profit, and private for-profit). Use an appropriate graph and describe what you see. Are these distributions symmetric? Skewed? How do the distributions differ, if at all, by institutional control?
- Are public institutions less selective than private not-for-profit institutions? Conduct a t-test of the null hypothesis that these two types of institutions are equally selective (based on *adm_rate_all*), and report 95% confidence intervals for the mean admissions rate in these two sectors. Hint: you will probably need to create a new *control* variable that is only populated for these two sectors.
- Provide a scatter plot that shows the relationship between average SAT equivalent scores (on the vertical axis) and institutional selectivity (*adm_rate_all*, on the horizontal axis). How would you describe the relationship between these two variables? Is it linear? Visually, how much variability is there in SAT scores among highly selective colleges? How much variability is there in SAT scores among less selective colleges?

4. Provide a descriptive analysis of the cost of attending college.

   - The variable *costt4_a* represents the 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 who receive Title IV aid. Use a histogram and descriptive statistics to describe the distribution of this variable. (Report the N, mean, standard deviation, minimum, p25, median, p75, and maximum).
   - Prospective college students are often dissuaded from applying to colleges with a high "sticker price". The variable *npt4* contains the average *net* price of annual attendance. It is the total cost of attendance minus the average grant/scholarship aid for all full-time, first-time, degree/certificate undergraduates who receive Title IV aid. Use a histogram and descriptive statistics to describe the distribution of this variable. How does this distribution compare to that of the full cost above? (Report the N, mean, standard deviation, minimum, p25, median, p75, and maximum).
   - Calculate a new variable that equals the ratio of the net price *npt4* to the total cost of attendance *costt4_a*, and use a histogram and descriptive statistics to describe the distribution of this variable. (Report the N, mean, standard deviation, minimum, p25, median, p75, and maximum). At the "typical" college, how much do students pay as a proportion of the "sticker" price?

- Using a box plot, show the distribution of the ratio calculated above separately by sector (public, private non-profit, private for-profit). Describe what you see.
- Do more selective colleges offer a larger "discount" on the total cost of attendance? Provide a scatter plot showing the relationship between the ratio calculated above (on the vertical axis) and institutional selectivity (*adm_rate_all*, on the horizontal axis). Describe what you see, and report the Pearson correlation coefficient. Calculate a bivariate regression that corresponds to this scatter plot. Provide an interpretation of your intercept and slope coefficient. Is the slope coefficient statistically significant?
- Do colleges with a larger "discount" have higher completion rates? Calculate a bivariate regression between the completion rate for first-time, full-time students (*c150_4*) and the ratio calculated above. Interpret what you find (and think about what sign you think the slope should be). Explain why this relationship probably should not be considered *causal*.

5. Provide a descriptive analysis of college endowments, and how it relates to institutional selectivity and financial aid.

   - Create an endowment per student variable that equals the value of the institution's total endowment at the beginning of the year divided by total enrollment. (Replace this variable with a zero if the endowment value is missing but enrollment is not missing. We will assume these institutions do not have an endowment). Use a histogram and descriptive statistics to describe the distribution of institutional endowment per student. What percentage of colleges and universities have a positive endowment? As an aside, what are the 10 institutions with the largest endowment per student, and where does Vanderbilt fall in this distribution of this variable?
   - Do institutions with larger endowments per student offer a larger "discount" on the sticker price? Provide a scatter plot showing the relationship between the ratio created in part 4 (on the vertical axis) and endowment per student (on the horizontal axis). Describe what you see, and report the Pearson correlation coefficient. Do this only for institutions with a positive endowment per student.
   - What other institutional characteristics are associated with a larger endowment per student? Report bivariate correlation coefficients between endowment per student and the following variables: HBCU status, selectivity (admissions rate), SAT equivalent scores, percent of undergraduates who are Black, instructional expenditures per FTE student, average faculty salaries, and percent of faculty who are full time.

6. Finally, conclude your paper with a brief summary of your findings. Does your analysis suggest other research questions that researchers and policymakers should investigate in greater depth?

Grading Rubric: Statistical Project

Criteria	Points possible
Paper introduction	**10**
Introduction clearly describes the research question being investigated	4
Introduction adequately describes the data set being used in the analysis	4
Introduction provides a brief overview of the paper's findings	2
Analysis	**75**
Statistical tools employed are appropriate for the posed questions	23
Reported results are accurate and complete	23
Written descriptions and/or interpretations of results are clear and correct	23
The analysis uses variety in statistical methods employed	6
Presentation and style	**15**
Writing style is clear and concise, and flows logically	3
The paper reads like a report, not like a series of problem set answers	3
Graphs and tables are accurate, neat, and referenced in the text of the paper	3
The paper meets the length and formatting requirements	3
Conclusion provides a concise and useful summary of the paper's findings	3