Data Science in Education: Mini-Project #2

In this assignment, you will return to the South Carolina school report card data, again focusing only on traditional high schools. This time, you will conduct a longitudinal analysis to investigate how your response variable of interest varies by poverty category over the five academic years from 2017-2018 to 2021-2022. You should use R markdown to create a write-up of your analysis and submit your knitted HTML file to Canvas when finished. Use the headings below in your report. Include all R code used to generate your output and appropriate commentary throughout, but hide any messages and warnings and adjust font sizes in visual displays as necessary to create a polished report.

Introduction

10 pts

- In the Introduction section of your report, identify your selected response variable by stating the statistical research question your longitudinal analysis will be investigating.
- Write a few sentences to provide some insight into why the response variable you selected is of interest to you and how you anticipate this variable will change over the five years of the study by school poverty category (low, mid-low, mid-high, high). No references are needed, just some initial thoughts on the potential variable relationships based upon your expertise in the field of education.

Load Packages

5 pts

- In this section, you only need to provide the R code for installing and/or loading the packages you used to complete this assignment.
- At a minimum, this will include `readxl`, `tidyverse`, `nlme`, and `emmeans`.

Import Data

10 pts

- For each year, you will need to import the financial data page and bind the rows from the different years together. Chapter 10 of the textbook provides one strategy for doing such a task.
- If your response variable of interest is not on the financial data page, you will also need to import the sheet containing your response variable of interest from each year and bind the rows from the different years together.
- Note that with the `sheet` argument of the `read_excel` function, you can specify the Excel sheet to import either by position or by name, but some years contained a cover sheet and some did not.
- Note that page 4a was named "4a.CCRPage_ACT_SAT_DUAL_AP_IB" in the 2018 and 2019 files but the final "B" was left off the page name in the 2020-2022 files. There may be other inconsistencies, so be on the lookout.
- You will need to use the `na` argument of the `read_excel` function to specify the codes used for missing values. Note that these vary from sheet to sheet of the data file. Some sheets use more than one code and in this case be sure to use the combine function, `c()`, in the `na` argument.

- Note that the `read_excel` function tries to guess the type of variable (e.g., numeric or character) in each column by default. However, the way some variables are stored in Excel changes across the years, causing R to guess different variable types in different years which becomes problematic when attempting to bind the rows together. One strategy for handling this issue is to read all columns in as character variables by specifying `col_types = "text"` in the `read_excel` function, and then after binding the rows together changing columns to numeric variables as appropriate using the `as.numeric` function.
- If your response variable has dollar signs or commas in what should be a column of numeric data, R will think these are character variables. If you try to coerce them to numeric variables without removing these symbols, R will convert them to missing data. You can remove the symbols using the `mutate` function in combination with the `str_replace_all` function.

Process Data

20 pts

- If your response variable was not on the financial data page, use a `join` function to merge the different sheets of data together by ReportCardYear, SCHOOLID, and SCHOOLTYPECD.
- Use the filter function to reduce the observations to only high schools (H) based on the SCHOOLTYPECD variable.
- Use the `filter` function in combination with the `str_detect` function with the `negate = TRUE` option to reduce the observations to only traditional high schools. This involves removing any school name that includes the strings "Academy", "Charter", "College", "Magnet", "School Of", or "School For" as well as any district name that includes the strings "Charter", "Unified", "Governor's", "djj", or "Juvenile" (all schools in these districts have a SCHOOLID > 4700000).
- In the 2019 financial data page, information for Ridge Springs-Monetta High School was entered twice. You can easily remove the duplicate entry using the `unique` function.
- Use the `select` function to reduce your data frame down to include only the variables for report card year, district name, school name, school ID, StudentsinPoverty_PctCurrYr, and the response variable you selected for your analysis.
- Note that the school names and school IDs for a few schools are inconsistent from year to year, making it difficult to do an analysis grouped by school. Find the schools that have less than five years of observations using the `group_by`, `summarize`, `n()`, and `filter` functions. You will also want to request the school names for this set. Determine which schools had inconsistent school IDs and correct the typos.
- Next, use the `group_by` and `mutate` functions to create a variable giving the mean poverty rate of a school across the five years of the study. Use the mean poverty rate to classify each school as a low, mid-low, mid-high, or high poverty school using the Institute of Education Statistics definitions: low is 25% or less, mid-low is more than 25% up to and including 50%, mid-high is more than 50% up to and including 75%, and high is more than 75% in poverty.

Visualize Data

20 pts

- Produce a table displaying the mean of your response variable in each year for each poverty group.
- Construct a line plot to visualize the change in the mean of your response variable over time separately for each poverty group.
- Describe the pattern in your line plot.

Model Data

25 pts

- Fit a linear mixed-effects model for your response variable with a school-specific random effect and fixed effects for time, poverty group, and their interaction using `na.omit` for the `na.action` argument. Request an ANOVA table for the fitted model.
- Keeping in mind the principle of hierarchy, refine the model until all remaining terms are statistically significant at the 5% level (or there is only one variable left in the model).
- Use the contrast function of the `emmeans` package to determine which particular levels of the remaining variables are significantly different.
- Include commentary throughout the model refinement and provide an interpretation of your requested contrast.

Conclusion

10 pts

In the conclusion, summarize your methods and findings in a few sentences. Discuss whether the relationships you expected to see were present and also discuss any surprising relationships you found.