Center for Education Policy Research
HARVARD UNIVERSITY

STRATEGIC DATA PROJECT

STATISTICAL COMPUTING

Identifying School Performance with Regression

Jared Knowles
jknowles@gmail.com
Slack: @jknowles
Office Hours: Check schedule on Canvas

Scenario

Upon completion of the last module, your department was able to dissuade the agency from using a simple one-size-fits-all regression model to identify high performing schools and grades. Not only did your department demonstrate that the model itself was flawed, but you also suggested several alternative models that explained test score performance better.

Unfortunately, the project is not over. The desire to do this kind of analysis is very strong in the agency, and your supervisor has informed you that a new model has been proposed. This model fits a separate linear regression for each grade-subject-school year for test score growth. In the data file you have been working in this results in 50 separate regression models.

Your supervisor has asked you to pair up with another analyst and has asked you to independently develop alternative models and replicate and critique each other's results. The idea is that together your department will look in pairs for alternatives to this approach.

In the guided analysis, your supervisor gave you an overview of some of the approaches you may wish to use to critique the model and helped familiarize you with the dataset you now have available (IT finally provided the full joined dataset for all analysts to use). Use this as a starting point for your own analysis of the data.

Instructions

Identify two additional variables you would like to include in your regression analysis to try to improve the suitability of this model for the business purpose of identifying schools. Choose variables you did not use in the last module. Coordinate with your partner to ensure you both do not choose the same two additional variables. Part of the goal in your memo will be to justify your inclusion of these variables, and part of the challenge as an analyst is quickly identifying promising variables and testing them.

Below are a list of some approaches you might choose to take in your analysis – do not feel obligated to address all of them in your script and/or memo. Do feel free to focus on the parts you want to learn more about using or are most relevant in your current work:

- Measuring the improvement in model fit (pick your favorite approach) for each of the fifty regression models by adding additional variables.
- Identifying relationships between high residual observations from the nested models and other data in the dataset (e.g. are schools with high residual values more likely to be a certain size or type).
- Using formal model diagnostics to show model improvement.
- Evaluate the impact of outliers on the proposed model compared to your own model.
- Compare the 50 nested model approach to a more complex global model fit to all the data.

Write an R/Stata script for your analysis. This script should be fully reproducible with the dataset allowing your partner to recreate your analysis and step through it line by line on their own machine.

Submit your script online to GitHub as a "Gist" and give the link to your partner. To do this, you will need a GitHub account.[1] Reproduce your partner's analysis on your local machine, step by step. Review this script as a critical friend – identify places where the script is unclear, code snippets you are unfamiliar with and are unexplained, and any mistakes you identify. You should share this feedback with your partner as a comment on the Gist, or if your feedback is more extensive, download the script and insert your comments directly in the code. If you cannot reproduce the analysis successfully, try to work with your partner to identify the problem and modify the script until you can.

After replication, write a brief memo together with your partner identifying the approach you both took and making a recommendation to your supervisor (see below for details).

Submit both the memo and a link to your script (hosted on GitHub) on Canvas.

Deliverable

An individual R or Stata script that together with the dataset reproduces your analysis accurately on another machine.

A memo of 800 words or fewer, co-authored with your partner, describing the results of both of your models and issuing a final recommendation to do one of the following:

[1] Details on how to submit a Gist on GitHub are included in the Canvas post of this assignment. Remember: https://gist.github.com/

a) Keep the existing model without additional variables
b) Adopt either your model or your partner’s model, or a hybrid of the two
c) Abandon the modeling task altogether

If your recommendation is to abandon the modeling task you will need a lot of evidence. There is a strong desire to do this work and the high R-squared value of the current model is persuasive to leaders in your agency that the model explains school outcomes. Be prepared to refute these and other likely challenges to your recommendation.

Appendix

As a refresher here is a description of the additional data elements available to you.

Additional Dataset	Description
Attendance	Total school enrollment, days possible, days attended, attendance rate
Program Enrollment	Total School Enrollment, Percent FRL eligible, FRL eligible count, student with disability percentage, student with disabilities count, English Language Learner count, English Language learner percentage
Discipline	Total School Enrollment, suspension rate, suspension count, suspension rate for female students, suspension rate for male students, Student enrollment by race/ethnicity, student suspension counts by race/ethnicity
Staff	Total School Enrollment, Administrator FTE, Student-Admin ratio, Support Staff FTE, Student-Support Staff ratio, Licensed Teacher FTE, student-licensed teacher FTE, total staff, Student-total staff ratio
School Attributes	Locale, County, Athletic Conference, Type of School, Grade Group, Low Grade, High Grade, School Size, Charter Indicator, Title I program Type