**INTEGRATIVE STATISTICS**

Customized Statistical Analysis

and Survey Research

# Analyzing Student Retention - a Peculiar Case Study, with Regression

## I. Introduction

Presenting at the Best Practice Solutions higher-ed. Enrollment Management Symposium in Philadelphia (July 22, 2022) spurred me to solve a peculiar problem. The way financial aid related to attrition at a certain selective NY college had been stymying me. The solution took actual thought. It might be instructive, or entertaining, for you.

Techniques included

- Correlation
- Linear regression
- Logistic regression
- A variety of data graphics

## II. Data Characteristics

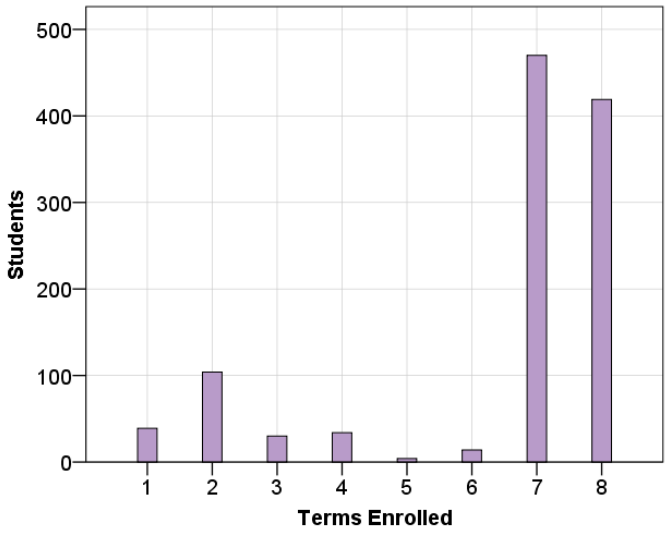
A private, selective college in New York state had provided records on 1,100 students. The goal was to understand as much as possible about the factors affecting retention/attrition – either in terms of

- a binary outcome of staying in school vs. dropping out (17% dropped out), or
- a more fine-grained outcome of how many semesters the student remained in school, from 1 to 8 (Terms Enrolled; mean = 6.5; a few values > 8 converted to 8)

For the purpose of this account, the predictors or independent variables being used to try to understand these outcomes included

- Grant money awarded by the college for the first year (First-Year Grants; mean = $7,100)
- Grant awarded for sophomore year (Sophomore-Year Grants; mean = $6,400)
- The product of the two (Grants' Product; based on centered and standardized original variables)

Terms Enrolled was distributed as follows:




**III. Methods and Results**

It was no small feat to fit a good model to the distribution graphed just above. It was so far from normal that correlation and regression coefficients and so on would be far from precise. For that reason, such results are rounded to just a single decimal place, as here:

Bivariate Correlations Among Main Variables

| | First-Year Grants | Sophomore-Year Grants | Terms Enrolled |
|---|---|---|---|
| Sophomore-Year Grants | .8 | | |
| Terms Enrolled | .1 | .4 | |
| Grants' Product | -.1 | .1 | .3 |

These results apply to a slightly larger sample than was used in listwise regression, which helps explain why, e.g., a bivariate *r* of .1 here became 0 below.

An early approach was to model the binary indicator -- of whether the student dropped out at any point -- in logistic regression. That proved somewhat informative. It led to the even more informative use of linear regression. The latter was preferred because it yielded zero-, partial-, and part correlations, and it allowed for partial regression plots to help with investigation of relationships. All of these helped shed light on the way a given predictor related to the outcome when other predictors were controlled.

Notably, First-Year Grants didn't seem to matter initially (zero-order $r \sim 0$). Then, keeping it in the model and controlling for Sophomore-Year Grants, First-Year Grants inexplicably became very important, but with a negative association:

| Model Stage | | Correlations | | | Adjusted $R^2$ |
|---|---|---|---|---|---|
| | | Zero-order | Partial | Part | |
| 1 | (Constant) | | | | -.001 |
| | First-Year Grants | 0 | 0 | 0 | |
| | Sophomore-Year Grants | | | | |
| 2 | (Constant) | | | | .47 |
| | First-Year Grants | 0 | -.6 | -.6 | |
| | Sophomore-Year Grants | .4 | .7 | .7 | |

Going in reverse, Sophomore-Year Grants, entered alone, showed a moderate connection, with $r \sim .4$ (and Adjusted $R^2 \sim .14$). These, too, ballooned – but now $r$ remained positive – once First-Year Grants was controlled.

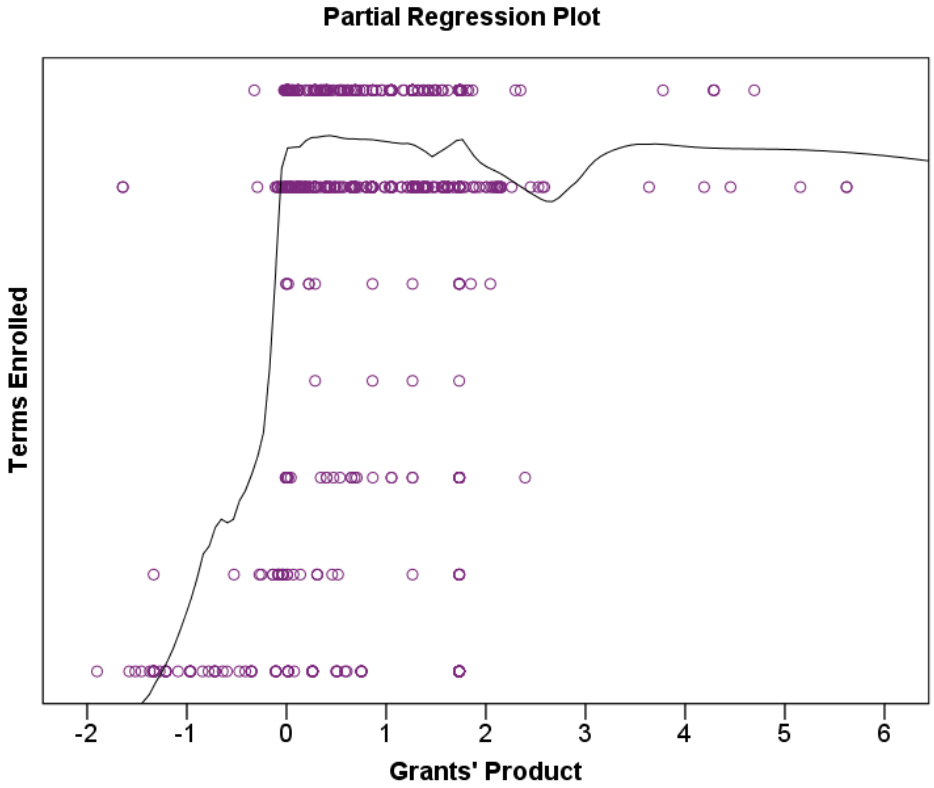
The statistical literature abounds with commentary on control, and more specifically on suppressor, mediation, and confounder effects. E.g., one might explore MacKinnon, Krull, and Lockwood (2000) or even a piece to which I contributed, Papadopoulos A and Stark RB (2019). But this situation seemed poorly covered by the usual explanations.

It seemed sensible to try an interaction for the two grant variables. Their product (centered and standardized) was entered in a new, third stage of a regression like the ones above.

The product had little role to play! In fact, its zero-order $r$ of .4 dropped to 0 in the presence of the other two predictors. With this interaction in the model, adjusted $R^2$ did not budge from its previous .47. Nor did the partial or part r of the other predictors change much.

A 3-D scatterplot seemed worth trying next. Sometimes it reveals patterns not visible with other methods. However, even with abundant turning and zooming to see the data from different vantage points, little was resolved.

However, one graphical tool that made an enormous difference was the partial regression plot (a.k.a. partial plot; a scatterplot showing an X's relationship with Y once each has been residualized from its connections with the other variables in the model). This plot, especially after being fit with a lowess or moving-average smoother, presented a big surprise:

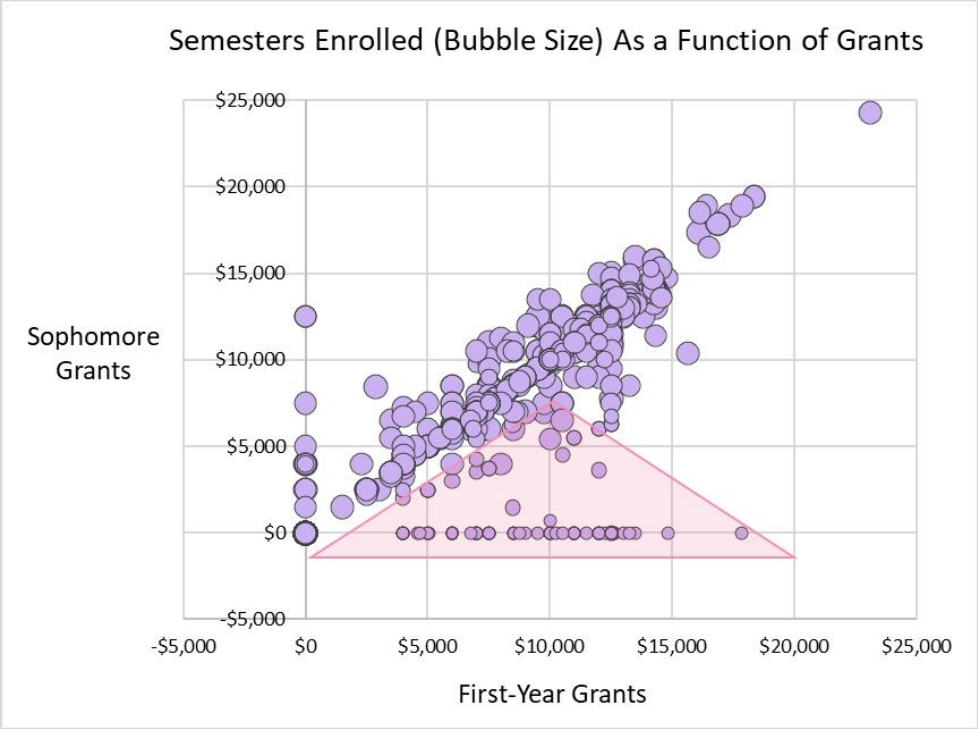



The numerical axis for Terms Enrolled is not labeled because through regression this variable has been altered via residualization with others in the model; the original scale values do not hold. For the Grants' Product this is true strictly speaking but in practice not meaningfully so.

This lowess line fits no typical equation; you will not capture it with a quadratic or cubic line, for instance. But does it ever show a distinct relationship! It's about as clear an example of a stair-step or threshold effect as you are likely to see in education or social-science research.

Within the range of -2 to 0, higher values of the Grants' Product were linked with dramatically higher values of Y, going from the worst to nearly the best. Further to the right, the relationship was flat.

In other words, to the extent that the product was negative, where one grant was above average and the other was below, Terms Enrolled was low. It turns out that almost every such situation entailed a *drop* from year 1 to year 2. This becomes clearer in the bubble plot.




Cases within the triangle are those for which the grant dropped substantially from the first to second year. The markedly smaller bubbles in that triangle indicate the much poorer retention results for those students. This is especially true at the bottom, where the drop was greatest. To be more quantitative about it, when the drop occurred – which was true for 19% of the students -- mean Terms Enrolled was 4.6, much lower than the 6.5 we saw for the whole group. And intuitively this explains a lot. Dissatisfaction with decreased second-year aid must have contributed to attrition.

To sum up: the connection between drop in aid and attrition didn't show up as a linear effect for the basic predictors in a regression. Nor would it show up as a quadratic or cubic effect. Nor as a sizeable interaction coefficient. But the drop was statistically very strongly connected to the outcome nevertheless, and this shows up if one is persistent enough (or, for some other researcher, smart enough at the outset!) to create the right plots.

**IV. References**

MacKinnon DP, Krull JL, Lockwood CM (2000). Equivalence of the Mediation, Confounding and Suppression Effect. *Prevention Science 1*:4, 173-181. https://www.public.asu.edu/~davidpm/classes/psy536/PrevScience2000.pdf

Papadopoulos A and Stark RB (2019). Does home health care increase the probability of 30-day hospital readmissions? Interpreting coefficient sign reversals, or their absence, in binary logistic regression analysis. *The American Statistician 75*(2), 1-32. DOI: 10.1080/00031305.2019.1704873.

*- Roland B. Stark, M.Ed.*

*Statistician and Research Consultant*

*July 2022*