Cost of Research and Education Activities in US Colleges - Scalability, Complementarity, and Heterogeneous Efficiency

Authors: Hajime Shima, Xiaoxiao Li, Michael Holton Price, Christopher P. Kempes
Affiliations: a. Santa Fe Institute, b. McGill University, c. Villanova University
Contact: hajime.fr@gmail.com

Introduction
Universities in the United States are remarkably diverse in their efficiency, both in terms of research output and educational achievement. In the existing literature, the heterogeneity is under-explored due to the lack of appropriate data and methodological limitations. In this paper, we address this by exploiting a newly consolidated dataset and adopting a neural-network-based method to infer cost functions for universities. Our analyses reveal that there are substantial efficiency differences across universities. Particularly, we show that while both research and education outputs generally exhibit an economy of scale, their scalability largely depends on the size and other institutional characteristics. Similarly, research and education activities are complementary to each other (economy of scope) only when the scales of productions are small to medium. Furthermore, the empirical cost isoclines of universities can be non-convex which leads to important policy implications, including diverse optimal portfolios and specialization. In short, our fully data-driven analysis suggests that model assumptions need empirical validation.

Methodology
Our framework combines modern deep neural network techniques with recent advances in “interpretable machine learning” algorithms. Specifically, we estimate cost as a nonlinear function encoded by a neural network without pre-specifying parametric assumptions; this allows complete flexibility in the functional form. Further, we utilize Bayesian optimization to find the architecture and hyperparameters of the neural network that properly model the non-linearity of the cost function without overfitting the data. We thus allow for heterogeneity from both observable characteristics and unexplained factors, such that the cost function captures the diversity exhibited by schools and can offer effective policy implications specific to each institution. In short, to model the total cost, we replace a translog function with the following flexible functional form:
\[\log_{10}(C) = f(y, x_1, x_2, \ldots, x_n) + \epsilon\]
where \(f(\cdot)\) represents the function represented by the neural network. This flexible cost function allows analyses of the scalability and complementarity of research and education that are not possible with restrictive specifications of the cost function, such as the translog cost function.

Data – Heterogeneity
We construct comprehensive measures of university outputs for both research and education.
UnivProd dataset: (1) Microsoft Academic Graph (MAG) which contains detailed information on academic publications; including author and institutional affiliation. (2) Mobility Report Card collected by Chetty et al. (2017) which reports income data for graduates and parents by university. (3) IPEDS data which provide budgetary information. We describe the generation of the dataset in a separate paper (Price et al. 2022). Here, we offer the first analysis that uses this dataset.

Results – Scalability
We would like to assess how much a school is paying for a unit of publication/degree. To this end, we compute the average incremental cost (AIC) of research and education outputs, respectively, as
\[\frac{AIC}{y} = \frac{C_i(y, y^*; \mathbf{X}) - C_i(0, y^*; \mathbf{X})}{y},\]  
\[\frac{AIC}{y^*} = \frac{C_i(y, y^*; \mathbf{X}) - C_i(y, 0; \mathbf{X})}{y^*},\]

Results – Complementarity
Is there complementarity between research and education production? What’s the efficiency gain in operating these two activities within the same organization? To explore this, we compute a metric of complementarity (i.e. economy of scope) between research and education productions as
\[SC = C_i(0, y^*; \mathbf{X}) - C_i(y, 0; \mathbf{X}) + C_i(y, y^*; \mathbf{X}) - C_i(0, 0; \mathbf{X}).\]

Results – Theory Implication
To further infer what institutional characteristics can explain the observed cost efficiency, we apply the SHAP analysis, an interpretable machine learning algorithm, to the obtained per-unit costs.

Results – SHAP Analysis
A: The distribution of the quantity-quality in research and educational outputs. The colors are based on the Barron’s selectivity measure for small dots (darker color indicates more selective), and large dots indicates observations for a set of arbitrarily selected schools.
B: The distribution of the unobserved cost shift. Most of the value range between -1 and 1, indicating that total cost may vary 2.196% based on factors not observed in the dataset.
C: The distribution of the observed cost shift. The most of the value range between -1 and 1, indicating that total cost may vary 2.196% based on factors not observed in the dataset.
D: The absolute importance of each feature on the total cost. The most important feature is research cost (C), and the fourth important feature reducing its educational cost (D).
E, F: SHAP analysis on the AICs of Caltech. Contrary to Stanford, its small-scale education reduces the research cost but increases the education cost.

Conclusion
In this paper, by applying modern machine learning techniques to our novel data, we attempt to minimize the assumptions imposed by researchers and our fully data-driven analysis reveals that questions of scalability and complementarity are highly conditioned on features of universities. Our resulting dataset, UnivProd, will be freely available soon. We describe the generation of the dataset in a separate paper (Price et al. 2022). Please contact the corresponding author, Hajime Shima (hajime.fr@gmail.com), for questions as well as access to the dataset and our manuscripts.