# Do College Characteristics Predict Alumni Earnings?*

Sai Pranav Sripathi

This paper investigates to identify observable institutional characteristics that are associated with higher alumni earnings at U.S. colleges and universities over a 10 year period. I studied how median earnings in ten years after entry vary with instructional expenditures per full-time equivalent student, admissions selectivity, student financial aid composition, institution size, and the mix of academic programs. I fit multiple linear regression models with a log earnings outcome as well as a LASSO model for variable selection, and I evaluated prediction accuracy using a train/test split. I also constructed prediction intervals to assess how well the models quantify uncertainty in earnings predictions. Results indicate that higher instructional spending and greater selectivity are positively associated with alumni earnings, while a larger share of Pell Grant recipients is associated with lower median earnings, even after controlling for sector and state effects. I conclude by discussing limitations of this observational, cross-sectional approach, including timing mismatches between current institutional characteristics and earnings measured ten years after entry.

## 1 Introduction

The economic outcomes of college graduates are a major focus for students, families, and lawmakers. In particular, the median earnings of former students are often used as one measure of the "value" that a college or university provides, especially while calculating the ROI before joining any institution. However, institutions differ greatly in their resources, student populations, and academic programs, which may all influence earnings, and early work in economics found that measures of college quality have a substantial impact on lifetime earnings (Solmon 1975) .The main question of this paper is: *How do instructional expenditures, selectivity, student composition, and program mix relate to an institution's median alumni earnings ten years after entry?*

*Project repository available at: https://github.com/SaipranavSripathi/Math261_Project2.

This is an important issue because popular rankings systems increasingly emphasize earnings measures without always explaining what underlying factors drive them. Institutions themselves also make strategic decisions about how much to spend on instruction, which programs to emphasize, and what types of students to recruit. Understanding which observable characteristics are most strongly associated with alumni earnings can inform both institutional planning and how the public interprets this data.

I used institutional-level data from the U.S. Department of Education's College Scorecard for 2020–21 (U.S. Department of Education 2023), which provides detailed information on Title IV–participating colleges and universities in the United States, including student outcomes such as median earnings ten years after entry. Title IV participating institutions are those that are eligible to offer federal aid like grants and loans to students. Recent work has used College Scorecard data and machine learning methods to predict post-collegiate earnings and debt (Agrawal, Ganesan, and Wyngarden 2015) . Using this dataset, I fit regression models where the outcome is the log of median earnings ten years after entry, and the predictors include log instructional expenditures per full-time equivalent (FTE) student, admission rates as measures of selectivity, the share of undergraduates receiving Pell Grants, the log of undergraduate enrollment, and the distribution of completions across broad program areas.

The analysis in this paper is descriptive and predictive rather than causal. I aim to quantify the associations between institutional characteristics and alumni earnings, compare the predictive performance of a standard multiple regression and a LASSO model, and evaluate the width of prediction intervals for new institutions. The remainder of this paper has the following structure, Section 2 describes the dataset and main variables, Section 3 explains the methods and statistical tests, Section 4 presents the results, and Section 5 discusses the broader implications and limitations.

## 2 Data and Variables

This analysis uses the 2020–21 College Scorecard institutional file (`MERGED2020_21_PP.csv`), published by the U.S. Department of Education (U.S. Department of Education 2023). Each record in this file corresponds to a single Title IV–participating post-secondary institution in the United States, including public universities, community colleges, private nonprofit colleges, and for-profit institutions. The unit of observation in all analysis is the institution.

The primary response variable is **median earnings ten years after entry**, reported in the Scorecard as `MD_EARN_WNE_P10`. This variable records, for each institution, the median annual earnings of former students ten years after they first enrolled, conditional on being employed and not enrolled in further education. Since, the earnings are right-skewed as seen in Figure 1 , I worked with the log of median earnings in the regression models. Let $Y_i = log(MD\_EARN\_WINE\_P10_i)$ denote the log median earnings for institution i.

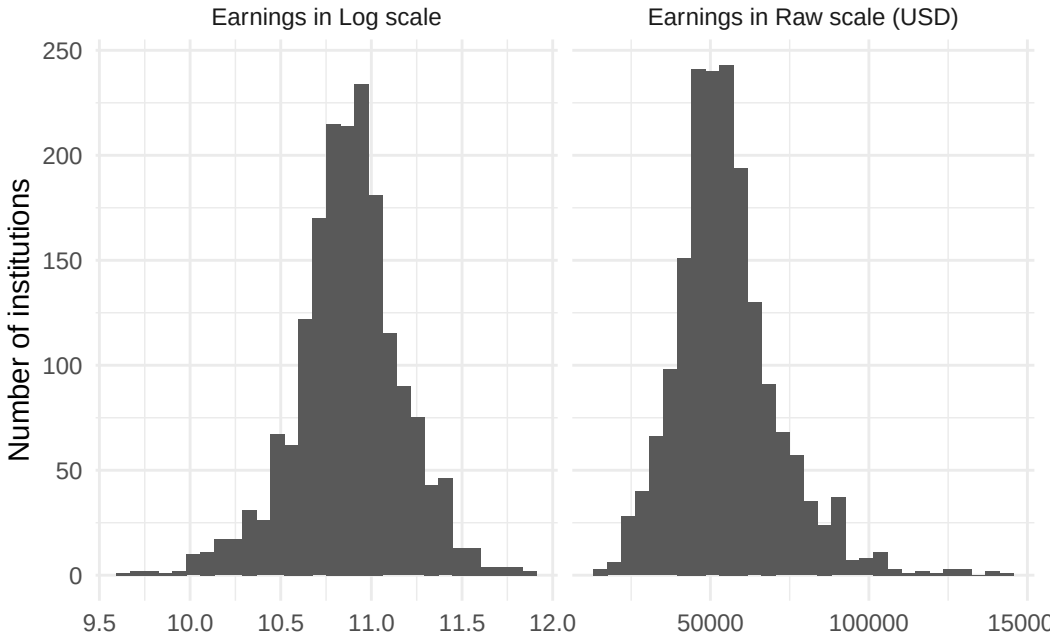


Figure 1: Distribution of median earnings (raw and log scale) ten years after entry

Several institutional characteristics in the Scorecard data are plausibly related to alumni earnings, and their measures are summarized in Figure 2 . Instructional expenditures per FTE (INEXPFTE) is measured by calculating total instructional spending divided by the number of full-time equivalent students and capture how much an institution invests in instruction per student. The box-plot for log(INEXPFTE) shows substantial variation, with most institutions clustered around a moderate spending level but a handful of high-spending outliers. I modeled the log of instructional expenditures so that changes can be interpreted approximately in percentage terms and so that these very high-spending institutions do not dominate the analysis.

PCIP refers to the Percentage of degrees awarded in a given program, where PCIP14 refers to share of completions in Engineering, PCIP11 for Computer Science, PCIP52 for Business, PCIP26 for Biological Sciences. Selectivity is summarized by the admission rate (ADM_RATE), the proportion of applicants who are admitted, and, when available, the average SAT score of enrolled students (SAT_AVG). The boxplot for ADM_RATE indicates that many institutions admit a large majority of applicants, but there is also a set of highly selective institutions with very low admission rates. The composition of students is described by the share of undergraduates receiving Pell Grants (PCTPELL) and by institution size. PCTPELL is the percentage of students who receive a Pell Grant, a need-based federal grant that provides financial aid to low-income undergraduate students. Its box-plot shows wide dispersion, with some institutions serving relatively few Pell recipients and others serving

very high shares. Institution size is measured by total undergraduate enrollment (UGDS), and the box-plot for log(UGDS) reveals a long upper tail of very large institutions, which motivates working with the logarithm rather than raw enrollment.

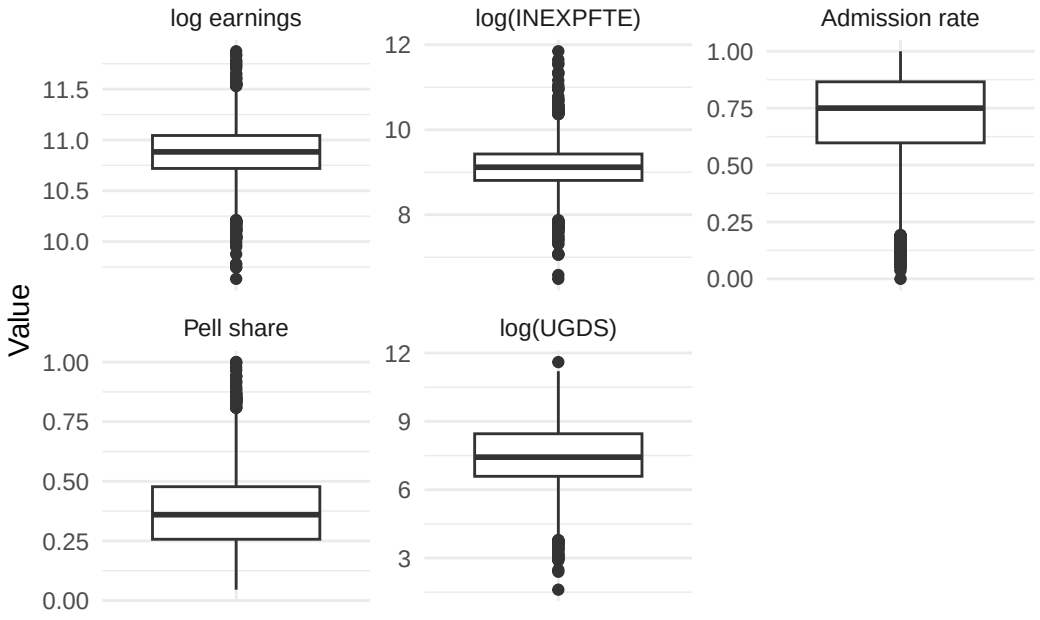


Figure 2: Boxplots of key continuous variables

Finally, for sector and state using the variables CONTROL and STABBR. CONTROL indicates whether an institution is public, private nonprofit, or private for-profit, and STABBR is the two-letter state abbreviation. In the regression models, both are included as sets of indicator variables (fixed effects), allowing for average differences across sectors and states that are not explained by the other predictors.

To form a reliable sample, I avoided institutions that have missing values for the outcome variable `MD_EARN_WNE_P10`, for instructional expenditures per FTE, and for the core predictors (`ADM_RATE`, `PCTPELL`, `UGDS`, and the selected PCIP shares). After applying these conditions, the number of institutions drops from 6681 in the raw file to about 1794 used for the regression analysis. This means that institutions with incomplete reporting are excluded from the analysis and the conclusions apply most directly to institutions with relatively complete data and may not generalize to all U.S. colleges and universities.

Table 1: Number of institutions in raw Scorecard file and in analytic sample.

| Sample | n |
|---|---|
| Number of Institutions | 6681 |
| Analytic sample (complete cases) | 1794 |

A fundamental limitation of the dataset is the **time lag** between institutional characteristics and the earnings outcome. The variable `MD_EARN_WNE_P10` reflects the earnings of students ten years after they first entered the institution, while the finance and admissions variables in the 2020–21 file describe a more recent snapshot of the institution. For example, if a university has substantially increased its instructional expenditures in the last few years, those changes may not yet be visible in the earnings of cohorts who entered a decade ago. Throughout the paper, I therefore interpret the estimated relationships as *contemporary associations with errors* between current institutional characteristics and alumni earnings rather than as causal effects.

# 3 Methods

All data cleaning, visualization, and modeling were conducted in R (R Core Team 2023). I used the `glmnet` package (Friedman, Hastie, and Tibshirani 2010) to fit the LASSO models and multiple standard R packages for data manipulation and graphics (like the `tidyverse` and related packages).

To study how institutional characteristics are related to alumni earnings, I fit multiple linear regression models (Gelman, Hill, and Vehtari 2021) and a LASSO model (Tibshirani 1996).

## 3.1 Multiple linear regression model

Let $Y_i$ denote the log median earnings for institution i. The main regression model is:

$$Y_i = \beta_0 + \beta_1 log(INEXPFTE_i) + \beta_2(ADMRATE_i) + \beta_3(PCTPELL_i) + \beta_4(log(UGDS_i)) + Y^T(PCIP_i) + \delta sector(i) + \delta state(i) + \epsilon_i$$

where $PCIP_i$ is the vector of selected program-mix shares for institution i. The PCIP variables record the shares of an institution's completions in broad fields of study (for example, engineering, computer science, business, and biological sciences), so they summarize the mix of majors offered at each college. Including these program-mix shares allows the model to capture the fact that institutions emphasizing higher-paying fields may have higher median earnings even if their spending, selectivity, and student composition are similar. $\delta sector(i)$ and $\delta state(i)$ are sector and state fixed effects, and $\epsilon_i$ is an error term capturing unexplained

variation in log earnings. The coefficients are estimated using ordinary least squares (OLS). Under this model, a coefficient on a continuous predictor represents the approximate percentage change in median earnings associated with a one-unit change in that predictor, holding other variables fixed.

One parameter of particular interest is $\beta_1$, the coefficient on log instructional expenditures per FTE. This coefficient measures the elasticity of median earnings with respect to instructional spending. I have performed the one-sided hypothesis test: $H_0 : \beta_1 <= 0 \;\; vs. \;\; H_1 : \beta_1 > 0$ which asks whether higher instructional spending is associated with higher alumni earnings. The test statistic is the usual t-statistic for $\beta_1$ from the OLS regression. I report the estimated coefficient, its standard error, the corresponding t-value, and p-value. I have also constructed a 95% confidence interval for $\beta_1$. In addition, I examine the estimated coefficients for the program-mix variables to assess whether they appear to be important predictors of earnings.

Standard linear regression assumptions are used: the errors $\epsilon_i$ are assumed to be independent with mean zero and constant variance, and approximately normally distributed. I assess these assumptions using residual diagnostics (residuals versus fitted values, Q–Q plots) and also compute heteroskedasticity-robust standard errors to reduce sensitivity to unequal variances across institutions.

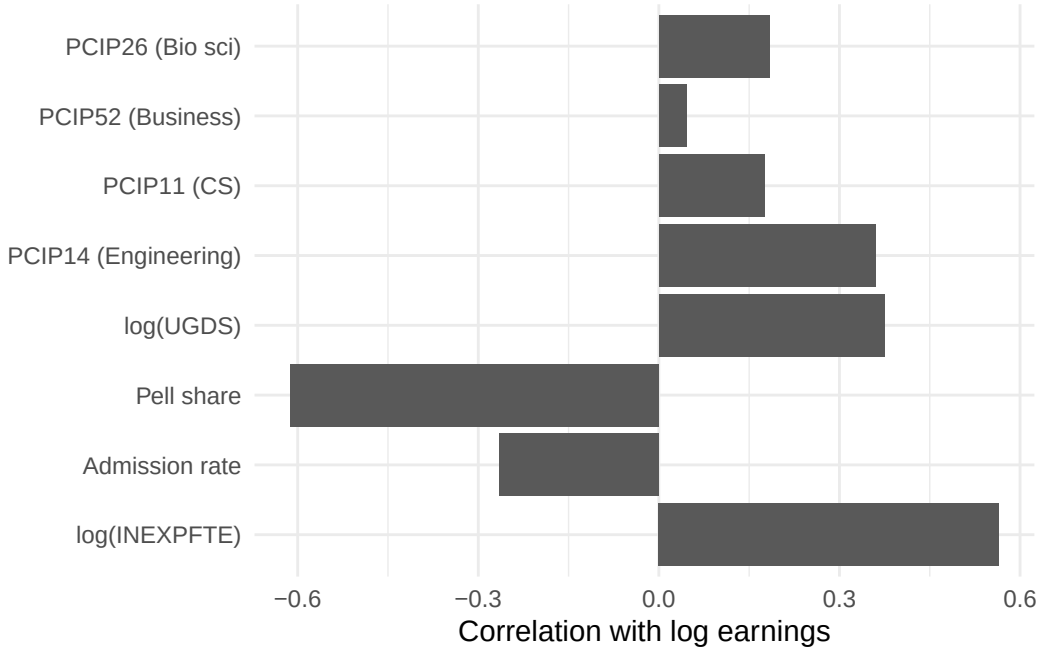


Figure 3: Correlations between log median earnings and key continuous predictors

### 3.2 LASSO model for variable selection

The design matrix includes several correlated predictors and many indicator variables for states and sectors, I also fit a LASSO regression model (Tibshirani 1996) . The LASSO estimates coefficients by minimizing the usual sum of squared residuals plus an $l_1$ penalty on the size of the coefficients. This penalty shrinks many coefficients toward zero and can set some of them exactly to zero, effectively selecting a smaller set of predictors.

I constructed a model matrix that includes all variables from the OLS model, standardized where appropriate, and fit the LASSO using K-fold cross-validation to choose the penalty parameter $\lambda$. I focus on the "one standard error" rule to obtain a more simple model which can explain the data comparably well. After identifying the set of predictors with non-zero coefficients at this value of $\lambda$, I refit an OLS model using only those variables to obtain interpretable coefficient estimates and standard errors. To evaluate predictive performance and construct prediction intervals, I split the analytic sample into a training set (80%) and a test set (20%), stratified by the outcome variable. All models are estimated using the training data only. I then used the fitted models to predict log median earnings for institutions in the test set.

For each model, I computed **root mean squared error (RMSE)**, **mean absolute error (MAE)**, and $R^2$ on the test sample. For the OLS model, I also compute 95% prediction intervals for the test institutions. Prediction intervals account for both the uncertainty in the estimated regression line and the residual variability around that line. I then exponentiate the interval endpoints to obtain prediction intervals on the dollar scale for median earnings from the log scale. Comparing these intervals across institutions and examining their width helps in assessing how precisely the model can predict earnings for new colleges.

## 4 Results

Descriptive plots relating log median earnings against each main predictor suggest positive associations with log instructional expenditures and selectivity, and a negative association with the Pell share and admission rate. Plots involving program-mix shares and the regression results shows that institutions with a greater emphasis on high paying fields such as Engineering, Computer Science, Business, Biological Science etc. tend to have higher earnings than other institutions, consistent with known differences in labor-market returns by major in general.

### 4.1 Multiple regression estimates

Table 2 reports the OLS estimates for the full model with log median earnings as the outcome. The coefficient on log instructional expenditures per FTE, $\beta_1$ is positive and statistically significant at the 5% level. Interpreting this coefficient, a 10% increase in instructional spending per FTE is associated with an approximate 0.10 x 0.113 = 0.0113 log units = 1.1% increase in

median alumni earnings, holding other variables constant. This provides evidence against the null hypothesis that higher instructional spending has no association with alumni earnings, in favor of the one-sided alternative that higher earnings. A 95% confidence interval for $\beta_1$ is approximately $(0.084, 0.142)$, implying that a 10% increase in instructional spending per FTE is associated with between about 0.8% and 1.4% higher median earnings.

Table 2: OLS regression of log median earnings on institutional characteristics.

| term | estimate | std.error | statistic | p.value |
|---|---|---|---|---|
| log_inexpfte | 0.113 | 0.015 | 7.52 | 0.00e+00 |
| ADM_RATE | -0.133 | 0.031 | -4.23 | 2.53e-05 |
| PCTPELL | -0.415 | 0.053 | -7.86 | 0.00e+00 |
| log_ugds | 0.030 | 0.006 | 5.32 | 1.00e-07 |
| PCIP14 | 0.591 | 0.061 | 9.76 | 0.00e+00 |
| PCIP11 | 0.344 | 0.134 | 2.57 | 1.02e-02 |
| PCIP52 | 0.159 | 0.054 | 2.96 | 3.13e-03 |
| PCIP26 | 0.276 | 0.102 | 2.70 | 7.10e-03 |

The estimates for selectivity variables also show strong relationships: institutions with lower admission rates or higher average SAT scores tend to have higher alumni earnings. The coefficient on `PCTPELL` is negative, indicating that institutions enrolling a larger share of Pell recipients generally report lower median earnings, even after controlling for spending, size, and program mix. For example, an increase of 10 percentage points in the Pell share is associated with about a 0.10 x (-0.415) = -0.0415 = 4.2% decrease in median earnings. Several program-mix (PCIP) variables are significant as well. Higher shares of completions in Engineering, Computer Science, Business, Biological Science etc. are associated with higher median earnings. Sector and state indicators capture remaining systematic differences across institution types and geographic locations, but the main patterns described above remain even after including these controls.

The diagnostic plots for the OLS model indicate that the residuals are nearly centered around zero in the residuals vs fitted plot across the entire range of fitted values and doesn't display strong curvature. This is a strong indication of linearity assumption being reasonably satisfied. But, since the spread of residuals is not constant, and has points with large residuals, it could mean a possible heteroskedasticity and potential outliers. The Q-Q plot compares the standardized residuals to a normal distribution, and clearly indicates that it follows the reference line fairly in the middle, but has large positive and negative residuals. This indicates heavier tails instead of a normal distribution of errors.
These are some of the observations based on the errors, but the overall conclusions about which predictors are important or how each predictor impacts the median earnings remains the same.

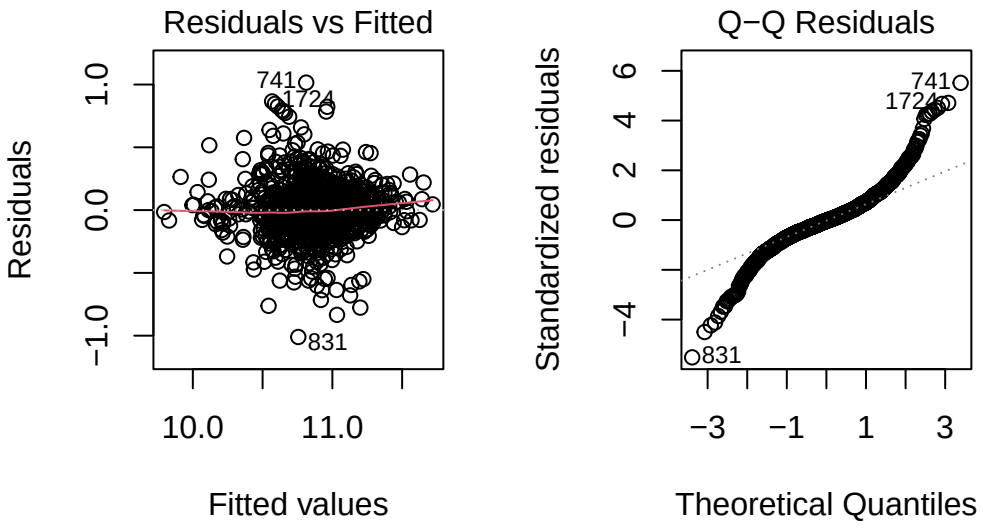


Figure 4: Regression diagnostics for the OLS model, residuals vs fitted and normal Q–Q plot.

### 4.2 LASSO selection and predictive performance

The LASSO regression selects a subset of predictors that includes log instructional expenditures, admission rate, Pell share, log enrollment, and several key PCIP variables, along with some of the sector and state indicators. Many weaker predictors are shrunk exactly to zero. When I refit OLS using only the selected predictors, the resulting model achieves predictive performance on the test set that is similar to the full OLS model: the coefficient on log instructional spending is about 0.13, coefficient on admission rate is about -0.12, and the coefficient on Pell share is about -0.44. Interpreted in the same way, these estimates imply that a 10% increase in instructional spending would roughly increase the median alumni earnings by 1.3%, similarly for Pell share it would indicate a 4.4% drop in median earnings, if we hold other predictors in each case constant.

Table 3: Test-set predictive performance for OLS and LASSO-based models.

| Model | RMSE | MAE | R2 |
|---|---|---|---|
| OLS (full) | 0.188 | 0.127 | 0.585 |
| OLS (LASSO-selected) | 0.195 | 0.137 | 0.549 |

To assess whether this simpler model loses predictive accuracy, I have compared the full OLS

model with the LASSO-selected model. The full OLS model achieves an RMSE of 0.188, an MAE of 0.127 and an $R^2$ of 0.585. The OLS model based on LASSO-selected predictors has a slightly higher RMSE of 0.195, and MAE of 0.137 with $R^2$ of 0.549. This indicates that the LASSO-selected model explains slightly less of the variation in log median earnings but does so with fewer predictors.

Several program-mix (PCIP) variables are significant as well. Higher shares of completions in Engineering, Computer Science, Business, Biological Science etc. are associated with higher median earnings. Sector and state indicators capture remaining systematic differences across institution types and geographic locations, but the main patterns described above remain even after including these controls.

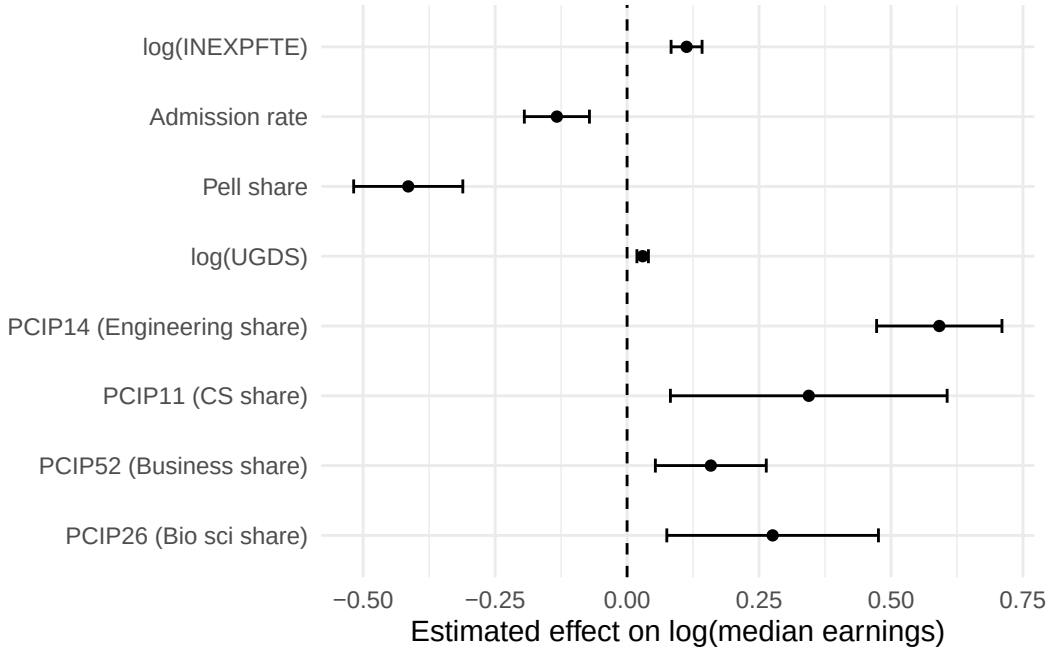


Figure 5: Estimated effects of key predictors on log median earnings

The scatter plots of predicted vs observed log earnings for these two models look very similar, with predictions clustering around 45-degree line and has relatively high errors for those institutions with higher earnings. Overall, LASSO confirms set of predictors that are most important for forecasting earnings and shows that spending, selectivity, Pell share, program mix in high-return fields and sector and state effects have most of the predictive power in the OLS model.

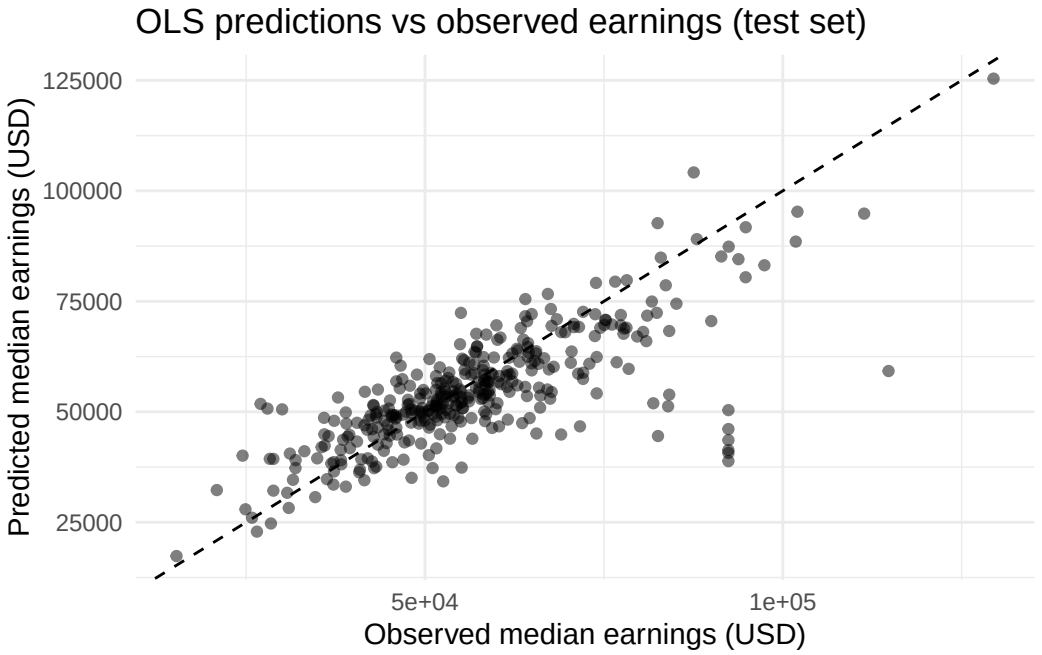


Figure 6: Predicted vs observed median earnings on test set

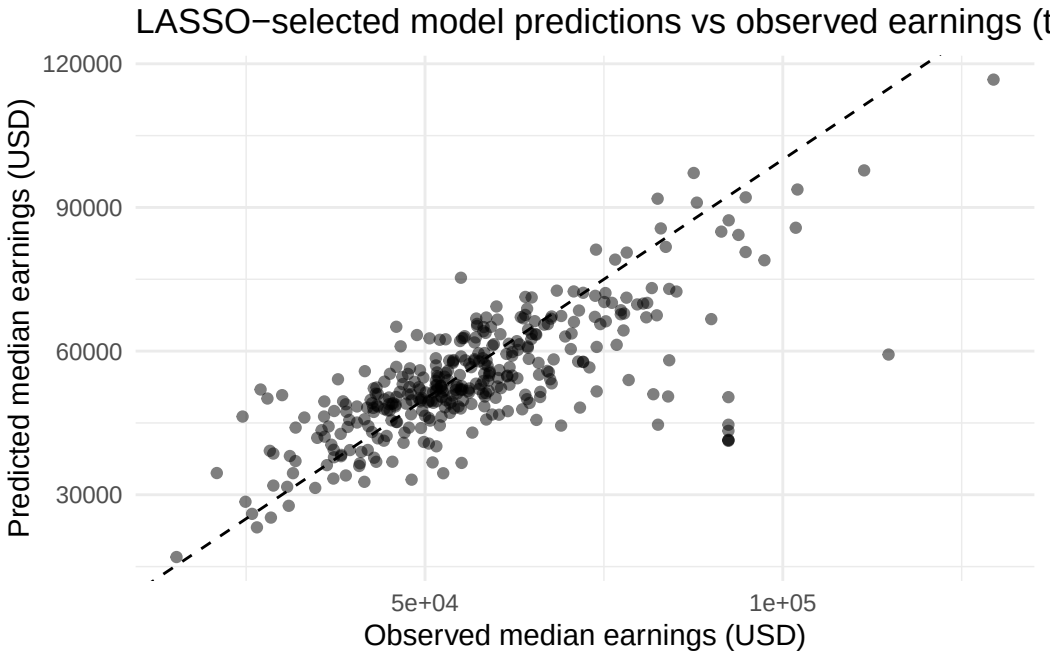


Figure 7: LASSO-selected model predictions vs observed median earnings on test set

### 4.3 Prediction intervals

For institutions in the test set, I have computed 95% prediction intervals for log earnings from the OLS model and transform them back to the dollar scale. These intervals are often fairly wide. For a representative public four-year institution with typical values of the predictors, the model predicts median earnings of about $58811, with a 95% prediction interval from roughly $40392 to $85627. This means that two institutions with very similar observed characteristics can still have substantially different alumni earnings, emphasizing the uncertainty involved in using such models for precise prediction.

Overall, the results show that the predictors considered especially instructional spending, selectivity, Pell share, and program mix are clearly associated with alumni earnings, but they do not fully determine them. Table 4 illustrates this by reporting observed earnings, point predictions, and prediction intervals for a subset of institutions: in many cases the intervals span several tens of thousands of dollars, even when the model's point prediction is close to the observed value.

Table 4: Observed and predicted median earnings (with 95% prediction intervals) for the top five institutions in the test set by observed earnings.

| Institution | State | Sector | Observed (USD) | Predicted (USD) | Lower 95% | Upper 95% |
|---|---|---|---|---|---|---|
| Franklin W Olin College of Engineering | MA | Private non-profit | 129455 | 125390 | 84835 | 185334 |
| Gnomon | CA | Private for-profit | 114785 | 59224 | 40882 | 85795 |
| University of Pennsylvania | PA | Private non-profit | 111371 | 94843 | 65497 | 137339 |
| Rensselaer Polytechnic Institute | NY | Private non-profit | 102051 | 95272 | 65614 | 138336 |
| Harvard University | MA | Private non-profit | 101817 | 88515 | 61080 | 128274 |

Table 5 shows the corresponding results for San Jose State University, a large public institution. For SJSU, the observed median earnings are about $78988, while the model predicts roughly $69002 with a 95% prediction interval from about $47716 to $99783. Again, the observed outcome lies well inside the interval, but the range is close to $50000 wide.

Table 5: Observed and predicted median earnings for San Jose State University (with 95% prediction interval).

| Institution | State | Sector | Observed (USD) | Predicted (USD) | Lower 95% | Upper 95% |
|---|---|---|---|---|---|---|
| San Jose State University | CA | Public | 78988 | 69002 | 47716 | 99783 |

## 5 Discussion

This study used multiple linear regression and LASSO regression to examine how institutional characteristics relate to median alumni earnings ten years after entry, using data from the 2020–21 College Scorecard. The main findings are that higher instructional expenditures per FTE, greater admissions selectivity, and a program mix emphasizing high-earning fields are associated with higher alumni earnings, while a higher share of Pell Grant recipients is associated with lower earnings. A LASSO model confirms the importance of these variables and yields similar predictive performance to the full regression model. However, prediction intervals are quite wide, indicating that even with detailed institutional data, there is substantial uncertainty in predicting earnings for any given college.

Several limitations qualify these conclusions. First, the analysis is purely observational and cross-sectional. Important factors such as students' prior academic preparation, family background, and local labor market conditions are not directly observed, and these could significantly impact the estimated relationships. Second, the timing mismatch between current institutional characteristics and earnings measured ten years after entry indicates that recent changes in spending, selectivity, or program mix may not yet be reflected in the outcome; institutions that have recently increased instructional expenditures, for example, may still appear similar to lower-spending institutions in this dataset. Third, missing data reduced the analytic sample and this may have introduced selection bias if institutions with incomplete reporting differ systematically from those with full data. Finally, all analyses are at the institution level, so they do not capture within-institution variation across student subgroups.

Even with these limitations, the results contribute to our understanding of how institutional resources and student and program composition are related to alumni earnings at the college level. They suggest that policies or strategies focused on instructional investment and on offering programs with strong labor-market returns may be associated with higher earnings, although causal claims cannot be made without stronger designs. Future work could build on this project by using panel data across multiple years of the Scorecard to align inputs and outcomes more carefully over time, by exploring nonlinear models or interaction effects, and inclusion of regional labor markets when available.

## References


Agrawal, Monica, Priya Ganesan, and Keith Wyngarden. 2015. "Prediction of Post-Collegiate Earnings and Debt." Stanford University. https://cs229.stanford.edu/proj2015/212_report.pdf.

Friedman, Jerome, Trevor Hastie, and Rob Tibshirani. 2010. "Regularization Paths for Generalized Linear Models via Coordinate Descent." *Journal of Statistical Software* 33 (1): 1–22. https://www.jstatsoft.org/article/view/v033i01.

Gelman, Andrew, Jennifer Hill, and Aki Vehtari. 2021. *Regression and Other Stories*. Cambridge University Press.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Solmon, Lewis C. 1975. "The Definition of College Quality and Its Impact on Earnings." In *Explorations in Economic Research, Volume 2, Number 4*, 537–87. New York: National Bureau of Economic Research. https://www.nber.org/books-and-chapters/explorations-economic-research-volume-2-number-4/definition-college-quality-and-its-impact-earnings.

Tibshirani, Robert. 1996. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society: Series B (Methodological)* 58 (1): 267–88.

U.S. Department of Education. 2023. "College Scorecard Institution-Level Data, 2020–21." https://collegescorecard.ed.gov/data/.