Understanding the Variation in Academic Testing Results

David Trobisch

Department of Economics

University of California, Davis

March 17, 2024

Abstract

This paper uses data from elementary schools to predict average STAR testing results. Particularly, we want to know how to explain the variation in academic testing results. We use the computer-to-student ratio, teacher-to-student ratio, and the average monthly family income as predictors, which collectively represent the resources of a school. After implementing multiple prediction methods, such as OLS and LASSO regression, we perform cross validation to determine which method has the best predictive power. Results indicate the best prediction method is the OLS regression with polynomials. After implementation, we predict that schools with lower resources will have lower average STAR testing results. This is consistent with what existing literature on this subject.

Introduction

This paper uses primary school data to explain the variation in academic testing results, particularly the California STAR examination scores. We utilize a dataset that contains elementary school STAR testing results and characteristics on school resources. The models include STAR results as the response variable, and computer-to-student ratio, teacher-to-student ratio, and monthly average family income as predictors. Each elementary school counts as one observational unit. Because each elementary school has many students and test takers, we try to

predict the average STAR results across students. The reason we choose such predictors is because they can be a proxy for the amount of resources a school has. When a school has more resources, they generally have more teachers, computers, and students from higher income families. Therefore, we are exploring the effect of resource level on academic testing results. Our research question is important because academic testing results are indicators of future student success, so it is important to observe how resources and school environments are impacting students. This research may be valuable for public policy and education reform.

For prediction, we implement numerous methods that take parametric and non-parametric forms. Ultimately, we want to use the prediction method with the best predictive ability outside of the sample – known as out-of-sample performance. The test mean squared error is a measurement for out-of-sample performance, and we want the prediction method that minimizes the test MSE. Although, we can never know the true test MSE because we don't know the true values of our test data. If we don't know the true responses, we cannot compute the prediction error. To solve this, we use cross-validation to estimate the test MSE. Throughout this paper, I use the split-half validation set approach to estimate the test MSE. This involves training the model using half of your data, then predicting using the other half of your data as test data. This allows us to estimate the mean squared error.

We implement 5 prediction methods: multiple OLS linear regression, OLS regression with polynomials, OLS regression with step functions, LASSO regression with polynomials, and classic regression tree. With each prediction method, we use cross-validation to estimate the test MSE. Before implementation, we run a principal component analysis on our three predictors. We find that we can use two out of the three predictors without damaging the test MSE. The advantage of this is better interpretability and simplicity. When we implement the first prediction

method, multiple OLS regression, we obtain a cross-validated estimate of the test MSE equal to 151.16. OLS regression fits the model by minimizing the training residual sum of squares. But because we care less about the training MSE and more about the test MSE, we don't necessarily want to minimize the training MSE, as this can lead to overfitting and worsen out-of-sample performance. The OLS regressions with polynomials and step-functions are fit the same way. Their respective estimated test MSEs are 148.83 and 173.07. LASSO regression is implemented differently. Rather than minimizing the training RSS, we minimize the training RSS plus some penalty term that penalizes model complexity. The purpose of LASSO is to avoid overfitting, ensuring out-of-sample performance is not sacrificed. We find that the LASSO estimated test MSE is 165.48. Lastly, we implement a regression tree on our data. This is a non-parametric prediction method because there are no parameters specified between our predictors and our response. The regression tree provides a cross-validated estimate of the test MSE equal to 179.60. The best prediction method to use in terms of out-of-sample performance is the OLS regression with third-degree polynomials. We predict with lower teacher-to-student ratios, computer-to-student ratios, and average family incomes, a school will have lower predicted STAR testing results. This implies there is a positive correlation between school resources and academic testing results.

My research adds to the large base of literature on the matter. There are many studies that have used similar prediction methods and obtained similar results. Namoun & Alshanqiti (2021) predicts testing results using a multiple OLS linear regression with data on afterschool programs. This is similar to our analysis because we use multiple OLS regression with proxy resource variables. Bramhe (2023) predicts academic testing performance using polynomial regressions and family resources. Our analyses are alike in that we both conclude polynomial regressions

provide the best out-of-sample performance when predicting testing performance. Dagdagui (2022) uses stepwise multiple regression to predict academic performance, which is similar to our analysis in that we use the same prediction method. We also both find that schools with a smaller number of computers and administrators (e.g. staff) have lower predicted academic testing results. Puah (2015) predicts academic student performance by using the LASSO method. Our analysis also uses the LASSO method, though we have a large difference in the number of predictors, so our LASSO model does not omit as many variables as theirs does. Puah predicts that schools with greater funding and programs have higher than average academic performance. Fernandez (2019) uses a decision tree to predict academic test performance. While we both p9u[we decision trees in our analyses, they utilize a classification tree. However, our results are similar. We both predict that students with greater financial support and school resources will perform higher.

Data

The data set is called "2014 STAR" and summarizes the 2014 California STAR testing results for elementary schools from different counties across the state. The STAR results comprise only of the STAR reading and math examinations, not the STAR Literacy Enterprise or STAR Reading Spanish examinations. The set also contains other variables about the schools like their teacher-to-student ratio, average monthly family income, and computer-to-student ratio, which will be the three base predictors in my model matrix. The set was provided and compiled by UCD affiliated econometrician, Matthew Reimer. The data is cross-sectional and there are 412 observed schools. Because we count each elementary school as one observational unit, our sample size is 412. The teacher-to-student and computer-to-student ratios are both

measured as percentages, the average monthly income is measured in thousands, and the Star results are measured on a scale from zero to 1400.

The sample statistics are the following. The STAR test scores have a mean of 654.33, standard deviation of 18.87, and a median of 654.70. We can conclude that because the mean is roughly equal to the median, the scores are not skewed. Therefore, they likely follow a symmetric distribution. The average monthly family incomes have a mean of 15.34, a standard deviation of 7.26, and a median of 13.73, meaning their distribution is likely right-skewed. This implies that the schools with high average monthly family incomes are pulling up the mean. The computer-to-student ratios have a mean of 0.14, a standard deviation of 0.06, and a median of 0.13. And lastly, the teacher-to-student ratios have a mean of 0.05, a standard deviation of 0.01, and a median of 0.05. The distributions for the previous ratios are also symmetric due to the alignment of the median and mean. More summary statistics can be found in the appendix table A1.

Models

There are three prediction models that this paper focuses on. It should be noted that different prediction methods are used on the same model. For example, we can run a simple OLS regression on a model, and then run a LASSO regression on the same model. Model 1 (M1) is **testscr = α + β1*tsr + β2*csr + β3*avginc + ε**, where testscr is the average STAR testing score, tsr is the school's teacher to student ratio, csr is the school's computer to student ratio, and avginc is the average monthly family income at a school. These three predictors are also included in the next two models, though polynomials or transformations of a variable are included as well. For example, model 2 (M2) is **testscr = α + β1*tsr + β2*csr + β3*avginc + β4*avginc^2 + β5*avginc^3 + ε**. Model 2 includes a third-degree polynomial of average family income. The

final model, model 3 (M3), is **testscr = α + β1*tsr + β2*csr + β3*C1(avginc) + β4*C2(avginc) + β5*C3(avginc) + β6*C4(avginc) + β7*C5(avginc) + β8*C6(avginc).** Model 3 includes six indicator variables that are constructed with step functions of the underlying average income predictor. The number of indicator variables were chosen with cross-validation.

Results

A principal component analysis constructs new variables from linear combinations of select predictors, with each component having their own distinct weights for each predictor. The aim is to reduce the number of predictors for better interpretability and visualization. This is possible because each component explains some portion of the predictor variance, though the variance explained decreases with each component added. Our PC analysis is on the model matrix resembling that of M1. The results show that using one principal component captures 50% of the predictor variation, and using a second component captures and additional 27.5% of the variance. This means we can use at least two variables to capture 77.5% of the original predictor variance. The scree-plot is shown in figure A2. It is important to choose the number of components with the best out-of-sample performance, and this can be done through cross-validation, which provides an estimate of the test MSE. After comparing the estimated test MSEs per number of components, it can be concluded that using two principal components dramatically decreases the estimated test MSE, while using a third principal component would not change the estimated test MSE. Therefore, we could lose a component for better interprability with no expense to the out-of-sample performance.

The first prediction method we run is an OLS multiple linear regression on M1. The response is elementary school STAR results, and the predictors are average monthly family income, teacher-to-student ratio, and computer-to-student ratio. To fit the model, we capture the

unknown parameters by minimizing the residual sum of squares. which is a summation of the squared prediction errors. After fitting the model, we use the split-half validation set to compute the estimated test MSE, of which we obtain an estimate of 151.16. This estimate on its own does not have interpretational value until we compare to other prediction methods' test MSEs.

The next prediction method we implement is another OLS multiple linear regression, though we run it on M2 now, which includes an additional third-degree polynomial of average monthly family income. This is because there can often be non-linear relationships between a predictor and response variable. After using the split-half validation set approach, we get a cross-validated test MSE estimate of 148.83. The results indicate the OLS model with added polynomial predictors has slightly better out-of-sample predicting ability.

We then run an OLS multiple linear regression on M3, which includes a step function of monthly average family income in replacement of the original predictor. The step function creates indicator variables that are equal to one when the average monthly family income is between some range. Each indicator variable has a different range criteria and the cuts are made in uniform quantiles. We choose the number of indicator variables through cross-validation. After using the split-half validation set approach, the smallest test MSE is associated with six additional variables comprising seven regions. The respective estimate is 173.07.

The next prediction method we implement is the LASSO regression on M2. The difference with LASSO is that it is a shrinkage method: its objective function includes the training residual sum of squares plus a penalty for complexity. We penalize complicated models because they can lead to overfitting, which can theoretically decrease the out-of-sample performance. The penalty term includes a tuning parameter, which represents the preference for a simple model. LASSO performs variable selection because it will eventually set some

coefficients to zero when the tuning parameter is large enough, thereby omitting the variables associated with such parameters. The tuning parameter is chosen through cross-validation, and if done properly, LASSO will have better out-of-sample performance than OLS regression. Our programming will automatically perform cross-validation to find the tuning parameter associated with the smallest test MSE. Results indicate the best tuning parameter to use is equal to 0.002. After fitting the LASSO model with the correct model simplicity, we perform the split-half set validation approach, obtaining an estimated test MSE equal to 165.48.

The final prediction method we implement is a regression tree on M1. This method is non-parametric and splits the data set into a select number of regions that have their own average response estimates. The predictions are made according to whether an observation belongs to some region based on their characteristics. If the observation belongs to a select region, their associated response is equal to the region's average. We choose a sequence of splits by minimizing the training RSS plus a penalty term. We choose the correct penalty based on cross-validation; this will result in the number of desired groups. First, we fit the tree independent of the number of terminal nodes. Then, we prune the tree using the best number of terminal nodes. Cross-validation indicates the best number of terminal nodes (groups) for out-of-sample performance is four. See the resulting tree plot in figure A3. Using cross validation, we find that a regression tree with four groups has an estimated test MSE equal to 179.60.

The prediction method with the best estimated out-of-sample performance is the OLS polynomial regression on M2. Our test data includes five observations, each with unique values of average monthly family income, teacher-to-student ratios, and computer-to-student ratios. If a test observation has a smallest average monthly family income, it will have low ratios as well. This enables an analysis of the aggregate effect of the levels of each variable. For example, we

can analyze how having less resources in general can impact the STAR testing results. We predict that when an elementary school has a teacher-to-student ratio of 0.05%, a computer-to-student ratio of 0.1%, and an average monthly family income of $5,000, they will have an average STAR score of 606.22. When a school has a teacher-to-student ratio of 0.5%, a computer-to-student ratio of 5%, and an average monthly family income of $10,000, we predict they will have an average STAR score of 627.86. And when a school has a teacher-to-student ratio of 2%, a computer-to-student ratio of 20%, and an average monthly family income of $20,000, we predict they will have an average STAR score of 662.84. This implies that having greater resources (e.g. more teachers) can have a positive impact on STAR testing results. See more prediction results in table A4.

Conclusion

This paper adds to the extensive research on variation in academic testing results. I gathered state elementary school data that contains the average STAR testing results for each school, as well as variables that summarize the amount of resources a school has, such as teacher-to-student ratio, computer-to-student ratio, and average monthly family income. The goal of this paper is to predict a school's STAR testing results using a function of three previously listed predictors. We first run a principal component analysis on our predictors and find that we can use one less variable without sacrificing out-of-sample performance. We then run numerous prediction methods on three models which contain our base predictors plus extra polynomials and transformations. After implementing OLS regressions with step functions and polynomials, we implement LASSO regression and then a classic regression tree. Upon each method's implementation, we calculate the estimated test MSE using cross-validation to evaluate the out-of-sample predicting performance of said method.

It is found that the prediction method with the lowest estimated test MSE is the multiple OLS regression with a third-degree polynomial of average monthly family income. We predict that when schools have low teacher and computer ratios, as well as low monthly family income, they will have low STAR testing results as well. While we do not interpret causal effects from any predictors, we can conclude there could be a strong association between resources (proxied by school ratios and family incomes) and average STAR testing results. This conclusion aligns with what previous research has found on the matter, that there is a positive correlation between resources and student achievement. Future research should strive to utilize larger datasets with greater number of base predictors, as we were working with a relatively limited dataset. Additionally, a dataset with counties from more than one state could provide greater prediction due to greater variation in the variables. Nevertheless, this paper attempts to explain the variation in academic testing results.

References

Bramhe, M.V. (2023). Student Performance Prediction Using Regression Analysis & Feature-Based Opinion Mining on Student Feedback. *Journal of Harbin Engineering University, 44*(7), https://harbinengineeringjournal.com/index.php/journal/article/view/377

Dagdagui, R. T. (2022). Predicting Students' Academic Performance Using Regression Analysis. *American Journal of Educational Research, 10*(11), 640-646, https://pubs.sciepub.com/education/10/11/2/index.html

Fernandes, E. (2019). Educational data mining: Predictive analysis of academic performance of public-school students in the capital of Brazil. *Journal of Business Research, 94*, 335–343. https://doi.org/10.1016/j.jbusres.2018.02.012

Namoun, A., & Alshanqiti, A. (2021). Predicting Student Performance Using Data Mining and Learning Analytics Techniques: A Systematic Literature Review. *Applied Sciences, 11*(1), 237-. https://doi.org/10.3390/app11010237

Puah, S. (2021). Predicting Students' Academic Performance: A Comparison between Traditiona l MLR and Machine Learning Methods with PISA 2015. *Journal of Engineering* (pp. 194 9-). https://search.library.ucdavis.edu/permalink/01UCD_INST/1hjlc2p/cdi_gale_infotrac academiconefile_A651754783

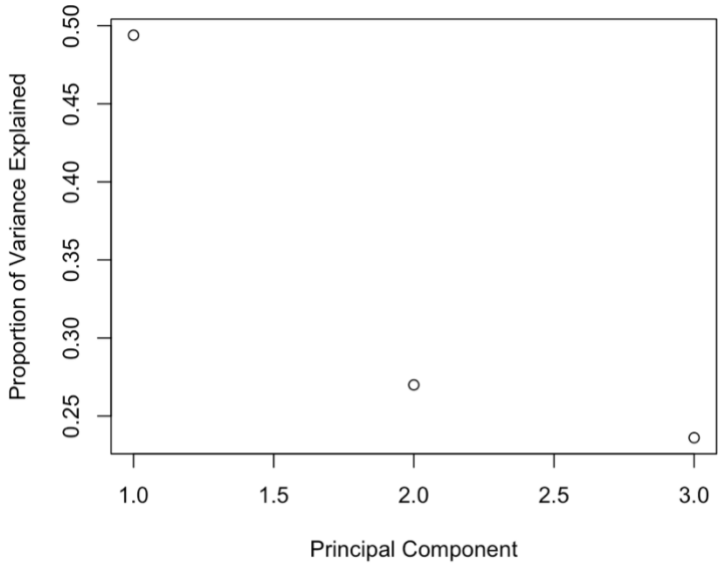
Appendix

Table A1

Variable	Mean	Std. Deviation	Median	Range
testscr	654.3303	18.87494	654.725	(605.55, 706.75)
csr	0.138566	0.06272774	0.1261233	(0.014, 0.421)
tsr	0.05149456	0.005128411	0.05081264	(0.040, 0.071)
Avg. monthly family income	15.33686	7.257288	13.7278	(5.335, 55.328)

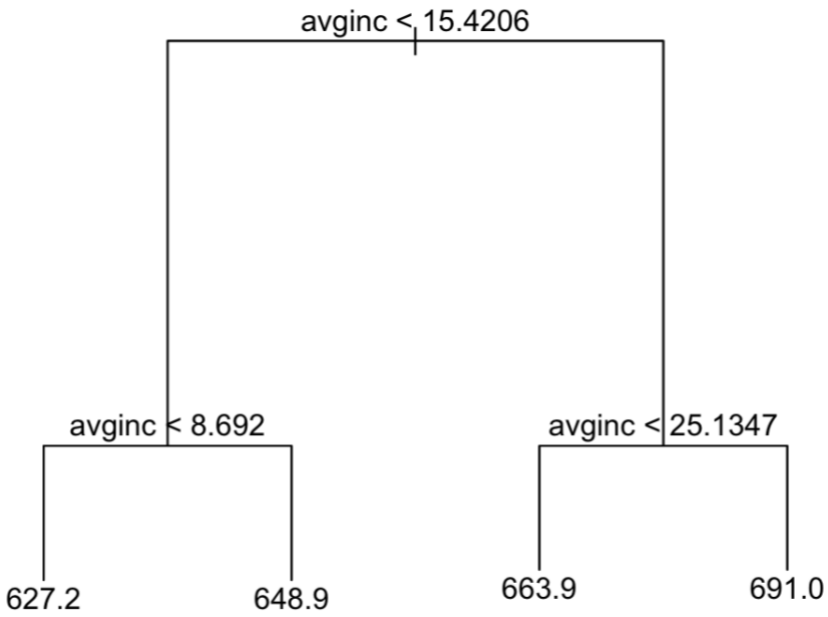
Above is a table with the summary statistics for the most important variables in our models.

Figure A2



Above is a scree plot with principal components plotted against proportion of variance explained. The variance explained decreases as components increase.

Figure A3



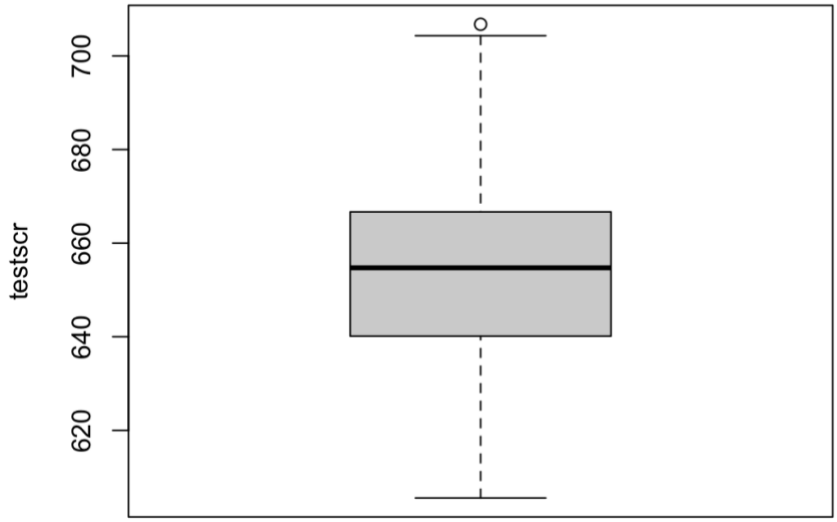
The above is a regression tree with four terminal nodes. We can interpret that when the average monthly family income is between zero and 8.692, the predicted average test score is 627.2.

Table A4

Prediction	tsr	csr	AMFI	Predicted testscr
1	0.0005	0.001	5	606.2221
2	0.001	0.01	8	618.3961
3	0.005	0.05	10	627.8648
4	0.01	0.1	14	642.9328
5	0.02	0.2	20	662.8384

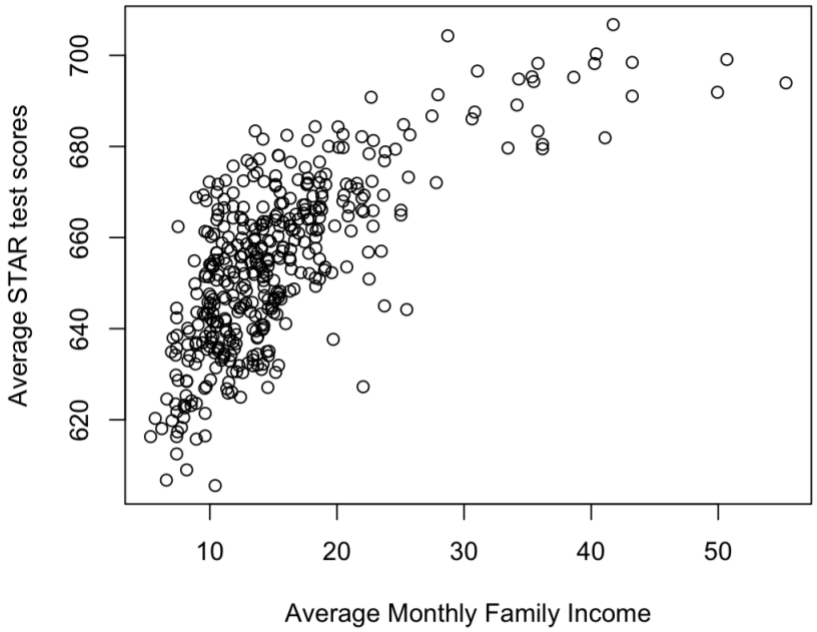
The above summarizes the prediction results for a range of test data. As the predictor values increase, the predicted test score also increases.

Figure A5



The above is a boxplot of the STAR test score variable. We can see that the median is around 650, the lower quantile is around 640, and the upper quantile is around 665.

Figure A6



The above is a scatterplot with average monthly family income on the x-axis and average STAR testing score on the y-axis. There appears to be a logarithmic relationship.