Case Study 3

Elementrary Education

Sam Lee, Ty Hawkes

Introduction

Research shows that strong academic performance during a child's elementary school years is a strong predictor of their successes later in life. Understanding which things are related to a student's academic performance in elementary school can help educators, administrators, and government leaders make informed decisions that positively affect rising generations. In this analysis, we hope to inform school officials and policy makers about elementary test scores and the factors that may affect them. We will study the statewide standardized test scores of various school districts in California and examine several factors associated with an increase or decrease in overall test scores.

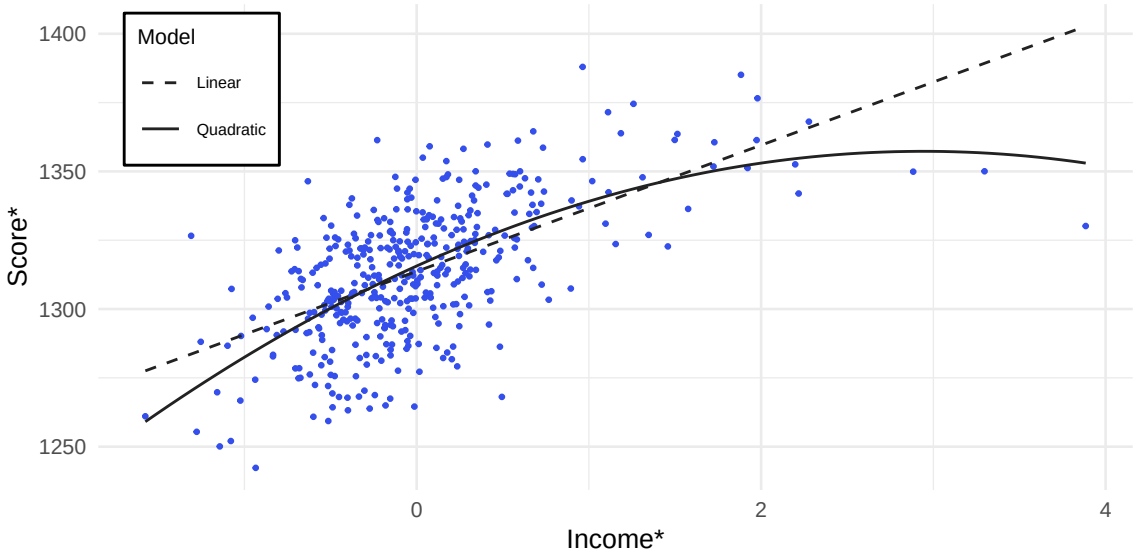

Figure 1: Non-linear Relationship between Income and Score

After conducting an exploratory data analysis, we found that there are two potential issues that could affect our primary analysis. Some of the factors in this study, like the district average income and the percentage of students who qualify for reduced-price lunch, are closely related to each other. If ignored, it can be difficult to determine the relationship between these factors and test scores. To avoid this, we will run tests to evaluate the severity of the issue, and remove factors if they pose a big enough problem. In addition to this, district average income appears to have a non-linear[1] (See Figure Figure 1) relationship with test scores.

[1]We visually model the relationship between *Income* and *Score* by partialling out the regressor (Income) and the regressand (Score). Thus, we compute $Income^*$ as $Income - Z((Z'Z)^{-1}Z'Income)$, where Z are the set of covariates excluding *Income*; the set of covariates are an $n \times (k-p)$ matrix, where p is the number of covariates that are "partialled out" (the *Income* in this case). Similarly, we compute *Score* as $Score - Z((Z'Z)^{-1}Z'Score$. Hence, Figure Figure 1 represents the non-linear effect that income has on Score, holding all else constant. Note that in *Income* represents the scaled *Income*.

This reduces the accuracy of our results if not accounted for. To solve this issue, we will add another factor to our analysis that will improve our model accuracy and provide additional insight, at the expense of some interpretability.

Methodology

To approach this problem, we will propose two methods and select one that best accomplishes the goals of this analysis.

Our first proposed model is a multiple linear regression model with an added second degree polynomial term for income and variable selection using LASSO regularization. This model is a good candidate because it will eliminate factors that reduce our accuracy, and it accounts for the non-linearity present in our data. More importantly, it will allow us to evaluate the relationships between test scores and various factors. Despite these strengths, this model may not be as predictive of student scores as other models. Additionally, this model will only fit well if the relationship between test scores and income is quadratic.

Our second proposed model is a Generalized Additive Model (henceforth known as GAM) regression with variable selection using LASSO regularization. This model is a good candidate because it will also eliminate factors that reduce our accuracy, and it will also account for the non-linearity of the data with smoothing techniques. One advantage to using this model over the linear regression model is that GAM regression can model complex, non-linear relationships that may not be captured well with polynomial expansions, possibly leading to a better fit. Due to this flexibility, however, GAM regression does not provide estimates for the effect size that each factor has on test scores. Still, we are able to accomplish the goals of this analysis with this type of model because GAM regression allows us to determine the statistical significance of these relationships, and visually interpret their direction.

Both models require that these assumptions are true if the models are to be accurate: the data must be linear, and the residual errors must be normally distributed, independent, and homoskedastic.

Model Evaluation

We evaluated both our GAM regression model and Linear regression model on their in-sample and out-of-sample performance measures. For our in-sample evaluation, we used adjusted R squared. Our linear regression model had an adjusted R squared value of [insert r squared] and our GAM regression model had an adjusted R squared value of [insert r squared]. Both models fit the data well, with the GAM regression model fitting slightly better.

For our out-of-sample evaluation, we chose to estimate the out-of-sample RMSE using k-folds cross validation. Our linear regression model had an estimated RMSE of [insert rmse], and our GAM regression model had an estimated RMSE of [insert rmse]. Both models performed exceptionally well when cross validated.

Because both models showed similar predictability, we ultimately chose to use our linear model in favor of it's superior interpretability. Our linear model[2] can be represented by the solution to the linear combination shown in Equation (1) below.

$$\begin{aligned}\hat{Y}_i = 1337.0 - 0.1393 \cdot \text{Lunch}_i + 0.005569 \cdot \text{Computer}_i - 0.001806 \cdot \text{Expenditure}_i \quad (1)\\ -0.5642 \cdot \text{English}_i + 549.5 \cdot \text{Income}_i - 155.2 \cdot \text{Income}_i^2\end{aligned}$$

As previously mentioned,

[2]This model assumes linear relationships between all co-variates and the response—e.g. $Y = X\beta + \epsilon$—and assumes the following distribution for the errors: $\epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$

Results

Conclusion

Teamwork