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 state-wide standardized test scores of various school districts in California and examine several factors associated with an increase or decrease in overall test scores.

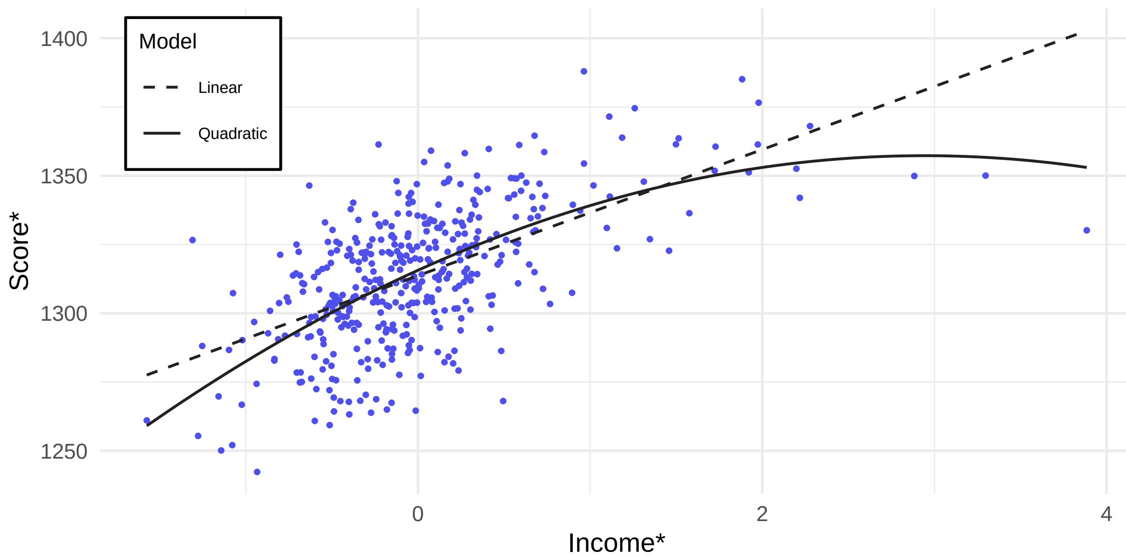

Figure 1: Non-linear Relationship between Income and Score

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

We first fit a LASSO regression model with a second-degree polynomial term included for *Income* to assess variable selection. Through this process, we eliminated using *STratio* (student-to-teacher ratio) as a predictor as the LASSO model shrinked *STratio* to zero. We used the LASSO-selected covariates (all of which have descriptions summarized in Table 1) for all models going forward.

We first propose a multiple linear regression model with an added second degree polynomial term for income. This model is a good candidate because it accounts for the non-linearity present in *Income*. More importantly, it will allow us to evaluate the relationships between test scores and various factors due to its parsimony. 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 and the other relationships between the factors and *Score* holds linearly.

Next, we propose a Generalized Additive Model (henceforth known as GAM). This model is a good candidate because 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 assume that the data are independent. Although GAM is somewhat flexible in the functional form of the data, we will model the data under the assumption that data are Normally distributed. Our linear model is a bit more restrictive in the sense that we model the data under (i) a homoskedastic variance, and (ii) the assumption that the response can be approximated through a *linear* combination of the covariates. Conversely, while GAM approximates the response in the form of an additive model, each additive model is in itself non-parametric, and maintains no assumption about the linearity of the data. Since we assume a Normal distribution of our response for both models, we implicitly assume a homoskedastic (constant) variance of the error terms (ε) within the GAM framework as well.

Model Evaluation

We first tuned our GAM model by selecting optimal hyperparameters for each covariate in the model. This process was achieved through a *randomized grid search* over the parameter space of interest. We chose an optimal basis function and the optimal number of "knots" for each factor and cross-validated each selection of hyperparameters through k-fold (with $k = 5$) validation. The final form of our GAM model[2] can be represented by Equation 1:

$$y_i = \beta_0 + \sum_{j=1}^{13} \alpha_j \phi_j(\text{Lunch}_{i1}) + \sum_{j=1}^{4} \beta_j \psi_j(\text{Computer}_{i2}) + \sum_{j=1}^{14} \gamma_j \psi_j(\text{Expenditures}_{i3}) + \sum_{j=1}^{19} \delta_j \phi_j(\text{Income}_{i4}) + \sum_{j=1}^{20} \zeta_j \phi_j(\text{English}_{i5}) + \varepsilon_i \quad (1)$$

$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

[2]Nomenclature: $\alpha_j, \beta_j, \gamma_j, \delta_j$, and ζ_j represent spline coefficients. $\phi_j(\cdot)$ represent thin-plate spline basis functions, and $\psi_j(\cdot)$ represent cubic basis functions.

We evaluated both our GAM regression model and Linear regression model on their in-sample and out-of-sample performance measures. The out-of-sample RMSE for each model was evaluated using k-fold cross-validation (using $k = 20$). These results are summarized in Table @tab-rmse. Both models performed exceptionally well when cross validated.

For our in-sample evaluation, we used adjusted R squared. Our linear regression model had an adjusted R squared value of 0.7877 and our GAM regression model had an adjusted R squared value of 0.7796. Both models fit the data well, with the linear regression achieving a higher adjusted R-squared due to its parsimonious fit.

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

$$
\begin{aligned}
y_i = \beta_0 + \beta_1 \text{Lunch}_i + \beta_2 \text{Computer}_i + \beta_3 \text{Expenditure}_i \quad (2) \\
+ \beta_4 \text{English}_i + \beta_5 \text{Income}_i + \beta_6 \text{Income}_i^2 + \epsilon_i \\
\varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)
\end{aligned}
$$

We present the estimated model below (note that these coefficients represent the estimated effects on the *unscaled/raw* factors).

$$
\begin{aligned}
\hat{y}_i = 1337.0 - 0.1393 \cdot \text{Lunch}_i + 0.005569 \cdot \text{Computer}_i - 0.001806 \cdot \text{Expenditure}_i \quad (3) \\
- 0.5642 \cdot \text{English}_i + 549.5 \cdot \text{Income}_i - 155.2 \cdot \text{Income}_i^2
\end{aligned}
$$

As previously mentioned,

Results

We present our results in terms of the estimation of Equation 2 by showing the estimates of coefficients on the scaled factors below in Table 1—in other words, each estimated coefficient represents a *relative* effect on *Score* as it relates to the other coefficients. Hence, we compare the magnitude in the coefficients (and their respective standard errors) to assess which factors contribute most significantly.

Table 1: Regression Results

Variable	Description	Estimate	95% Confidence Interval
Intercept		1313.666***	(1311.872, 1315.46)
Computer	Number of Computers	2.458*	(0.48, 4.436)
English	Percent of English learners	-10.317***	(-12.906, -7.728)
Expenditure	Expenditure per student	-1.145	(-3.121, 0.831)
Income	District average income (in USD 1,000)	52.568***	(43.849, 61.287)
Income^2	Income squared	-26.815***	(-34.285, -19.346)
Lunch	Percent qualifying for reduced-price lunch	-3.777*	(-7.461, -0.094)

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 18.7 on 413 degrees of freedom
Multiple R-squared: 0.7907, Adjusted R-squared: 0.7877
F-statistic: 260.1 on 6 and 413 DF, p-value: $< 2.2 \times 10^{-16}$

We return to Figure 1, and address the non-linearities in *Score* with respect to *Income*. We assess whether the quadratic term, $Income^2$ contributes significantly to explaining the variablity in *Score* by running an analysis of variance test (ANOVA) on two models: The first being the empirical model established by Equation 2; The second, like unto the first, omitting the quadratic term. The results of this test are summarized in Table 2. We reject the null hypothesis that there is no difference between the two models. In other words, the quadratic term cannot be omitted without sacrificing predictability in *Score*. Since the coefficient on $Income^2$ is significantly negative, this provides significant evidence to suggest that as income increases, the total effect that income has on increasing *Score*, decreases. This supports the *diminishing marginal returns* hypothesis.

Table 2: Analysis of Variance Table

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	413	144446				
2	414	161863	-1	-17417	49.799	7.238×10^{-12}

Model 1: $Y = X\beta + \epsilon$
Model 2: $Y = X_{-Income^2}\beta_{-Income^2_p} + \epsilon$

To assess predictability, we compare our linear model to GAM (as proposed earlier) and local linear regression with a quadratic term (LOESS). Similar to our tuned GAM model, we tuned our LOESS model through k-fold cross validation[3].

We performed k-fold cross validation using $k = 20$, and summarize the following results below as the out-of-sample predictive fit for each model:

Table 3: Out of Sample RMSE for each Predictive Model

Linear Model	LOESS	GAM
18.51708	18.23348	18.53671

We also summarize this visually in Figure Figure 2 by modeling how well each model predicts using *Income* as a predictor. We visually compare our quadratic model (Equation 2) to LOESS by fitting both on the partialled-out *Income* (since the LOESS was fitted on the dimensionally-reduced covariate matrix) and *Score*. Similarly, we compare our quadratic model to a more flexible GAM model by fitting both to the original scaled *Income*. On this note, we acknowledge that all three models perform similar predictively, but given that our quadratic model remains as the most interpretable, with only marginal improvements in RMSE shown by the local linear regression model, we maintain that the quadratic model proves as the best predictive and parsimonious model for this set of data.

We now turn to address the more precarious causal questions we have been presented with. Table 1 suggests that the factor, *English*, is significantly negative—in other words, for every one percent increase in the number of English learners within a school district, our model suggests that the average score for that district will decrease by -0.5642, on average. Hence, learning English as a second language may be a barrier to learning if[4] we do not suspect there are other unobserved covariates influencing the percent of English learners that are *also* correlated with *Score*. Unobserved factors such as school funding, parental education levels, teacher

[3] Our LOESS model was tuned through a randomized grid search algorithm, where we optimized over a space of *span* (the percentile of data used for each 'local' regression) parameters in the set $(0.1, 1)$. Through five-fold cross-validation, we found the optimal span to be 0.9919. We note that since this was optimized over a random selection over the parameter space from $(0.1, 1)$, the theoretical optimal may approach 1. To optimize algorithmic performance, we performed dimension reduction by partialling out the set of regressors (Z) on both (scaled) *Income* and *Score*.

[4] For a sufficient sample size n, if the endogeneity assumption holds, that is, if we have sufficient evidence to believe that $\mathbb{E}[\varepsilon|X] = 0$, then a causal claim may be warranted. However, we caution against this due to confounding factors in the data as we mentioned.

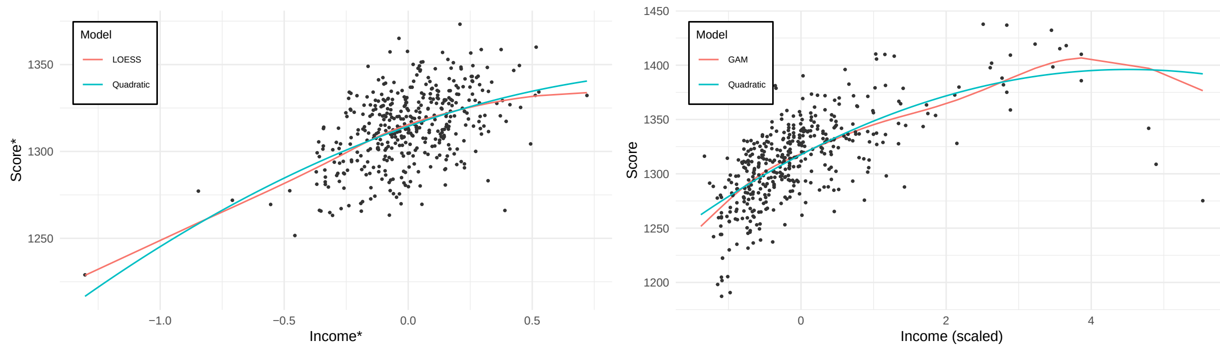

Figure 2: Model Prediction on *Score* with respect to *Income*

quality, and immigration status are all factors that confound the causal inference assumption. Thus, we caution from making any binding causal claims on this issue.

Additionally, the number of computers per district has the highest shift in *Score* as a positive predictor. Since this may be correlated with unobserved factors preventing any causal statements, we would recommend investigating these effects further: To establish a causal claim, we recommend parsing out the specific effect that *Computers* has on *Score* through a randomized experiment: Randomly assign a varying number of computers to different schools or classrooms within a district. This would ensure that the number of computers per school or classroom is not correlated with any other factors that could influence test scores (e.g., teacher quality, existing technology, or socioeconomic status). Then, establish an intervention group by providing additional computers or upgrade existing ones to some schools or classrooms. Finally, establish a control group by providing no additional computers or maintain the current number of computers in other schools or classrooms.

Conclusion

Teamwork