Elementary Education Analysis

Josh Christensen Matthew Morgan

Department of Statistics

BYU

Section 1

Problem Statement and Understanding

Background

- Research has shown that while in grades K-2, children learn at a greatly increased rate, compared to ages
- How well children learn in these formative years gives insight into how they will continue to progress in academics and life

SchoolResults.txt Dataset

- `Score`: Average cumulative Score on the Stanford 9 standardized test (out of 1600)
- `Lunch`: Percent qualifying for reduced-price lunch
- `Computer`: Number of Computers
- `Expenditure`: Expenditure per student
- `Income`: District average income (in USD 1,000)
- `English`: Percent of English learners
- `STratio`: Student-to-teacher ratio

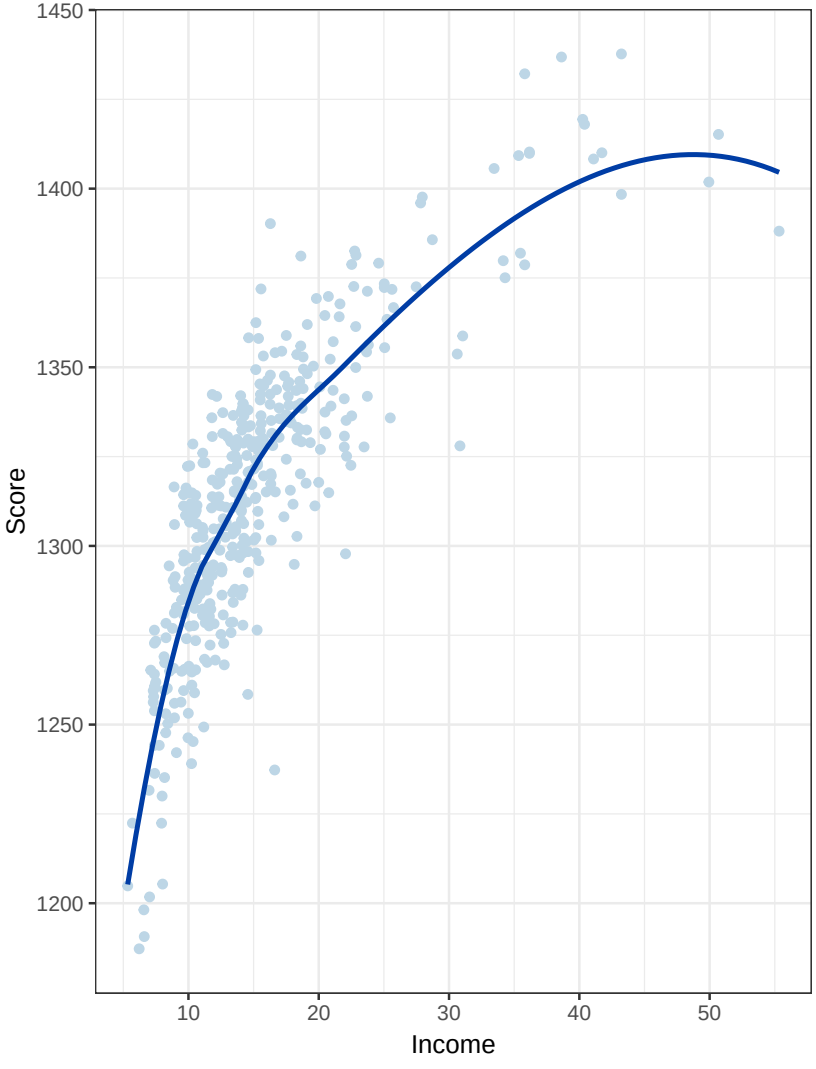
Analysis Goals

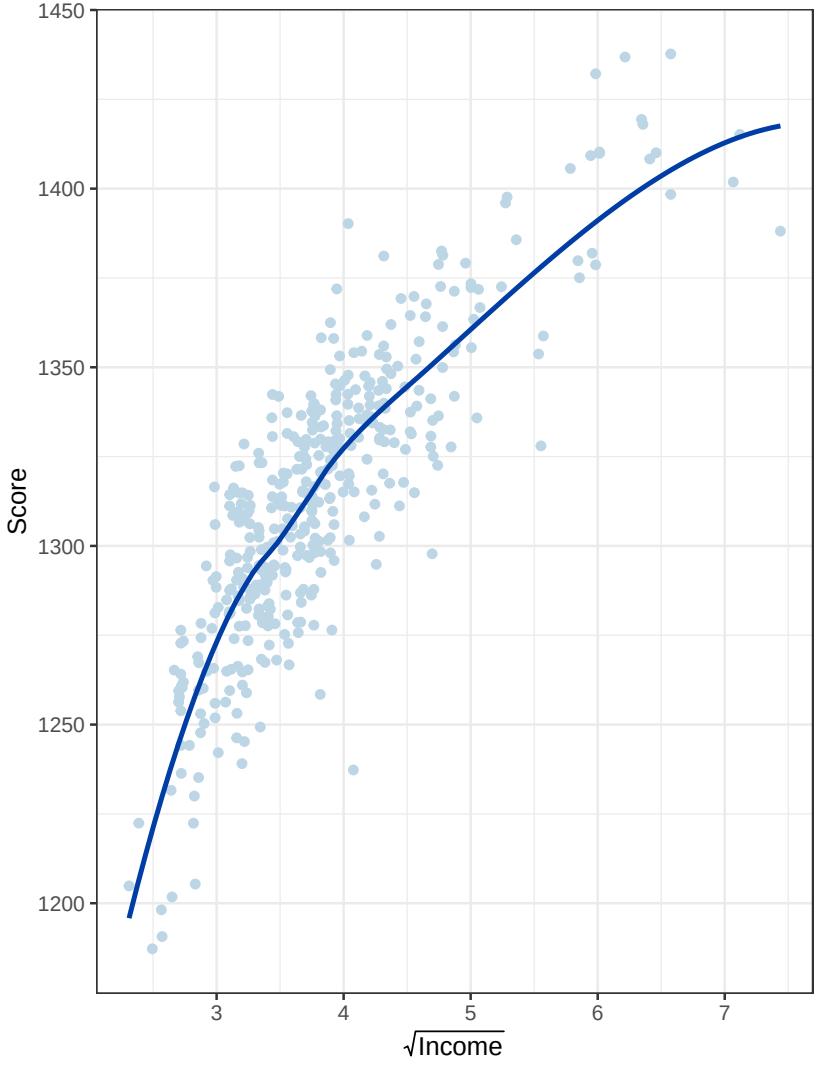
1. Is there evidence of diminishing returns on extracurricular activities in terms of student learning?
2. Is English as a second language a barrier to student learning?
3. What can be done to increase student learning?
4. How well does the model predict compared to alternatives?

Exploratory Data Analysis (EDA)

Relationship bewteen extracurricular spending and student learning

Is there evidence of diminishing returns?

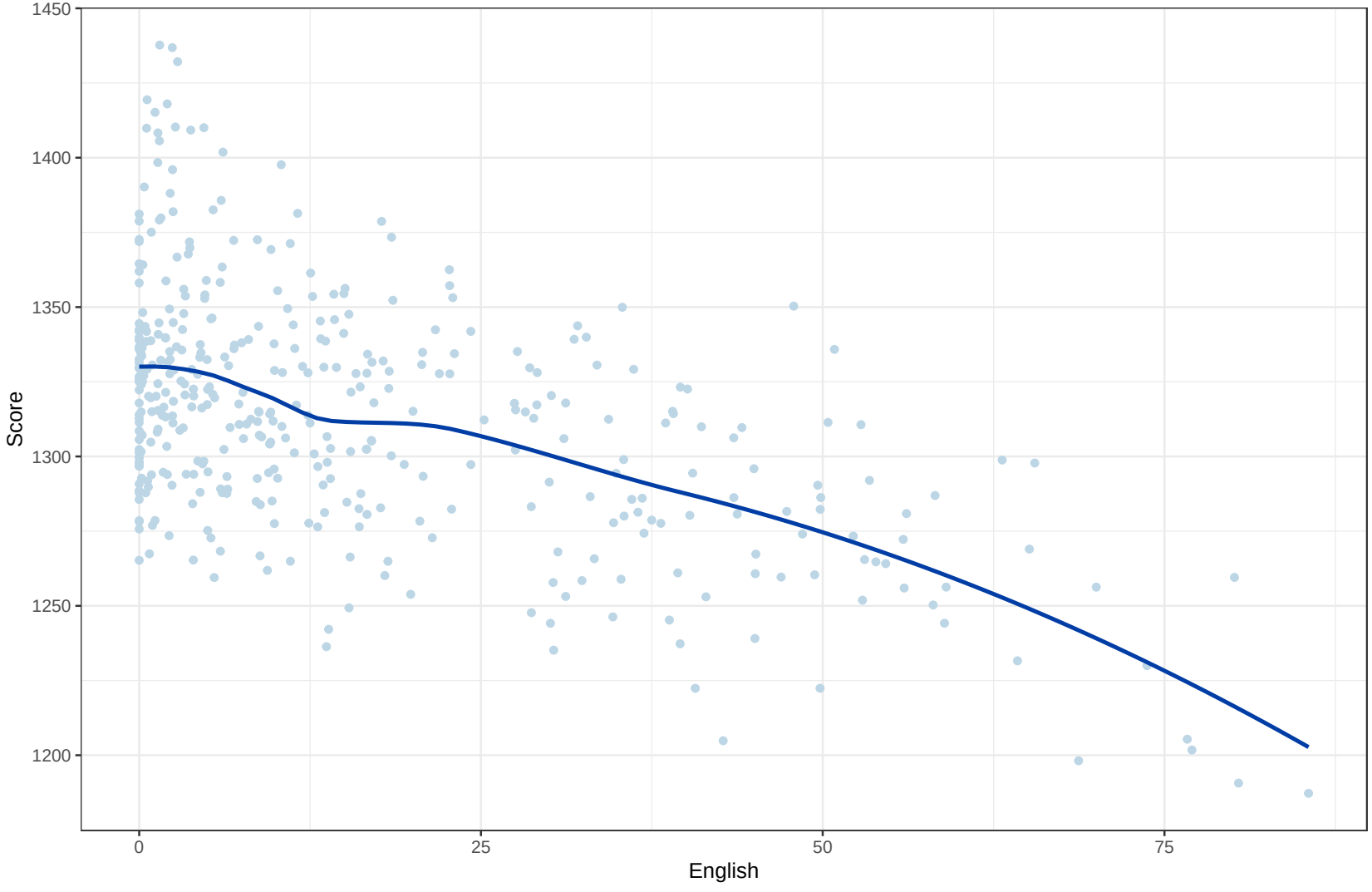




Exploratory Data Analysis (EDA)

Relationship English as a second language and student learning

Is it a barrier to student learning?



Section 2

Model Specification

Linear Regression

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6} + \beta_7 \sqrt{x_{i4}}$$

$$\epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

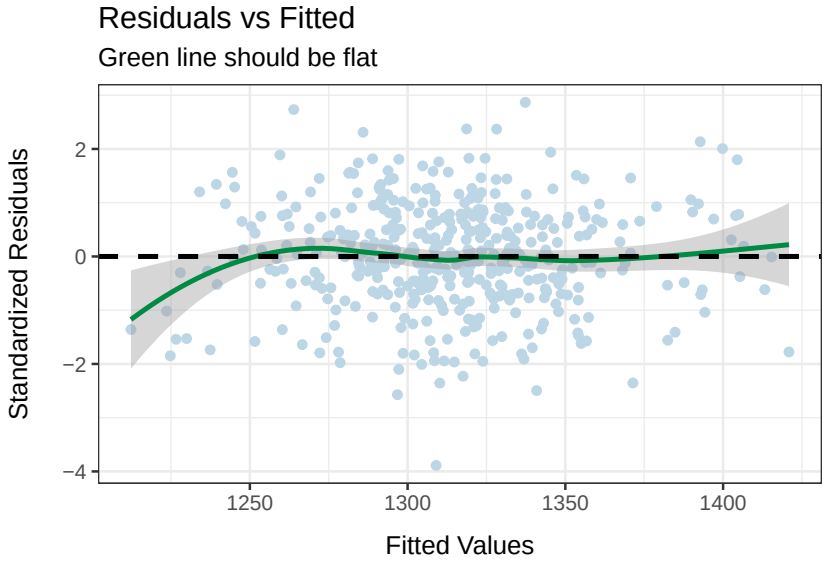
Parameter Interpretation

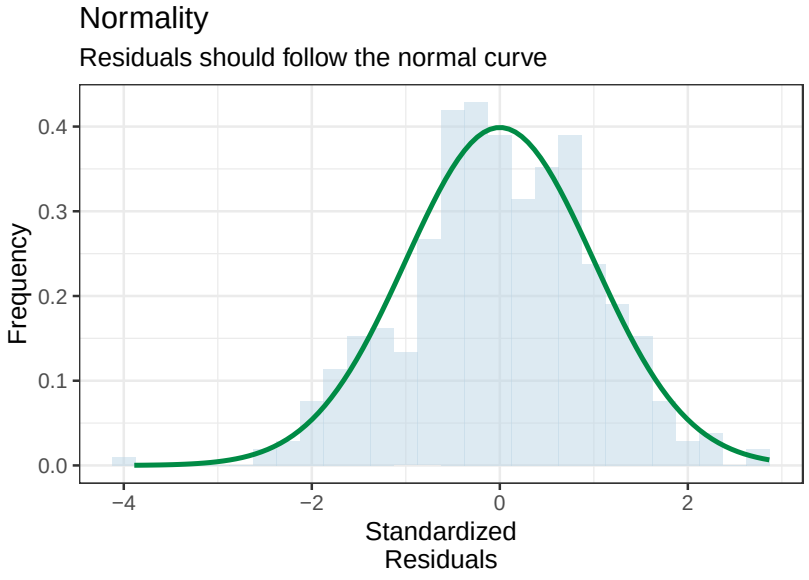
- y_i - Model will predict `Score` for the i^{th} school district
- x_{i1} - `Lunch` value of the i^{th} California school district
- x_{i2} - `Computer` value of the i^{th} California school district
- x_{i3} - `Expenditure` value of the i^{th} California school district
- x_{i4} - `Income` value of the i^{th} California school district
- x_{i5} - `English` value of the i^{th} California school district
- x_{i6} - `STratio` value of the i^{th} California school district
- β_0 - `Score` is β_0 on average
- β_5 - Holding all other x_{ip} constant, as `English` increases by 1 unit, `Score` is expected to change by β_5 points on average
- (β_4, β_7) - Holding all other x_{ip} constant, as `Income` increases by \$1,000, `Score` is expected to change by $\beta_4 + \beta_7$ points on average
- σ - For any x_{ip}, 99.7% of the `Score` values will be within 3σ of $\beta_0 + \beta_1 x_{i1} + \cdots + \beta_7 \sqrt{x_{i4}} + \epsilon_i$
- ϵ_i - Residuals will be normally distributed, with a mean of 0 and with a variance of σ^2

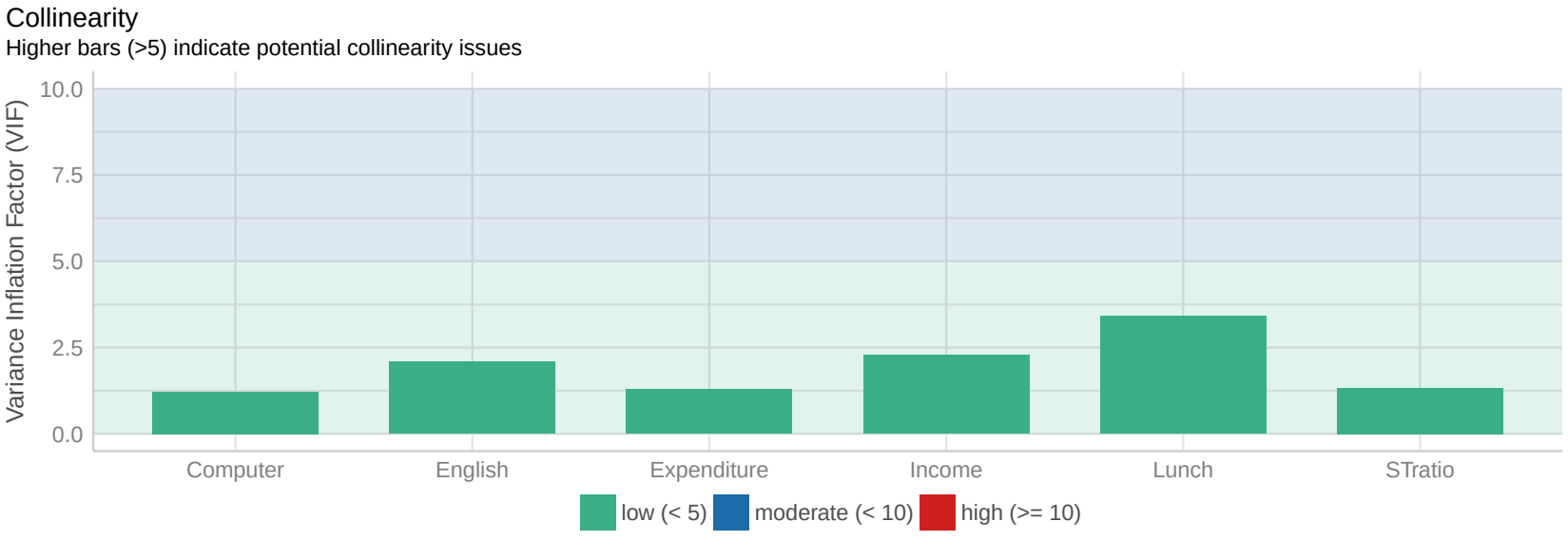
Section 3

Model Justification & Performance Evaluation

Model Assumptions







Model Metrics

Table 1: Prediction Evaluation

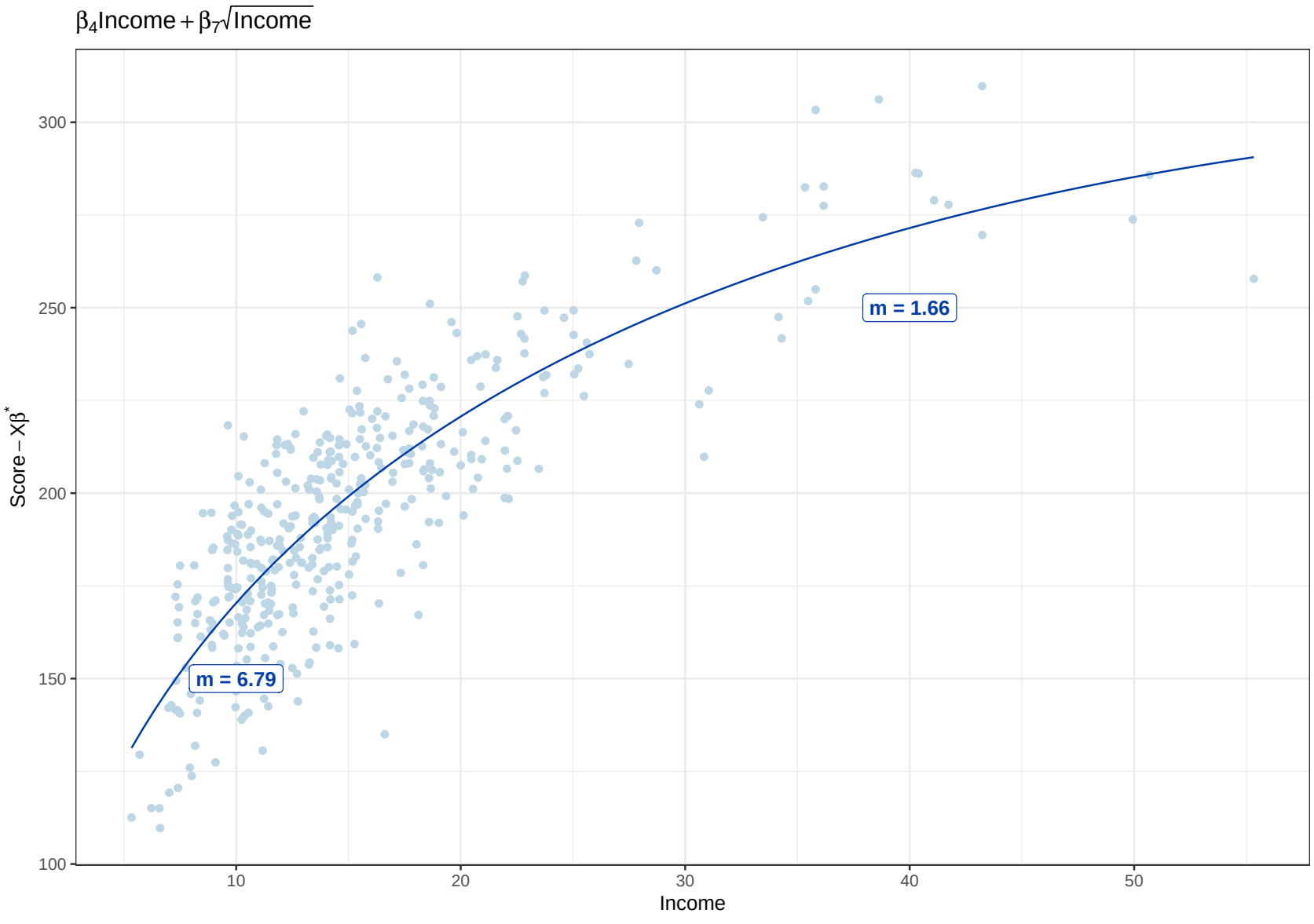
Method	R^2	RMSE	Bias
Linear Regression	0.7885	18.8330	0.0000
Random Forest	0.7805	18.9946	-0.0618
Neural Network	0.7765	19.4833	0.2073

- *All models were trained with a 5-fold cross validation procedure*
- The standard deviation of `Score` was 40.59 points

Section 4

Results

Diminishing returns with Income?



English second language barrier?

- Holding all other variables constant, as percent of students learning English (`English`) increases by 1 percentage point, `Score` is expected to change by -0.51 points on average.
- The 95% confidence interval for the effect of the `English` variable is (-0.66, -0.37)

Steps to increase learning

Table 2: Manipulable and Significant β's

	Coefficients
Number of Computers	0.0052
Percent ESL	-0.5135
Income	-3.4635
$\sqrt{\text{Income}}$	64.8291

- Suggested steps
  - Increase access to extracurricular activities and computers.
  - Increasing availability and quality of ESL programs.

Prediction comparison

- Linear regression model predicted on par with the other models considered in this analysis
  - Random Forest, Neural Network
- Given the comparable predictive capabilities of the linear model, the lack of interpretability inherent in the other models was not worth the minimal increase in predictive power
  - More data could see an increase in predictive power in non-interpretable models

Section 5

Conclusions

Revisiting goals

- The goals of our analysis were to:
  1. Find a model with good fit
  2. Answer the questions of interest
- Table 2 summarizes various models with good fit and predictive power, including our chosen linear regression model
- In the results section we showed characteristics of the relationships between the variables `English`, `Income` and `Score`, in addition to exploring possible methods of increasing student learning.

Shortcomings & Next Steps

- A potential shortcoming of our analysis is that it does not account for potential spatial correlation between districts.
- Linear models are also less scalable than other models. This same framework could not be applied to datasets with more variables and more observations as easily as a random forest or neural network.
- Variables such as teacher quality, tutoring programs, advanced learning programs, etc. are not included in our dataset, but may have an effect on student learning.
- Potential next steps for modeling could include exploring more interactions or fitting a spatially correlated model.
- Next steps for the schools would be to test the suggested changes and see if there is measurable improvement.

Section 6

Teamwork

Teamwork Summary

- Both of us fit neural nets and linear regressions in various different packages
- Both of us worked on the presentation graphics
- Both of us wrote the slide content
- Matt created the beamer template, including BYU colors and worked on the final linear regression metrics and fit
- Josh worked on getting the final neural nets to have consistent results and good fit