HW3 Proposal

RJ Cass

1. Understanding the Problem

Background:

We have a dataset of various factors of conditions in a school, and the corresponding test scores of that school. We want to use these data to understand the how the different factors influence test scores, as well as be able to predict test scores.

Goals:

In this analysis, we want to address several key questions:

i. Is there evidence of diminishing returns on extracurricular activities in terms of student learning?

- To answer this we will examine the relationship between Income (the factor that measures spending on extracurricular activities) and Score. If there are diminishing returns of the effect of Income on Score, we would expect the graph to show a relationship between the variables where the slope decreases as Income increases.

ii. Is English as a second language a barrier to student learning?

- To answer this, we will develop a model indicating the significance of each factor. If the English factor is indicated as being significant, then it will show that English as a second language is a barrier to student learning.

iii. In your opinion and based on the data, what can be done to increase student learning?

- To answer this, we will review the model to identify which factors were indicated as important. That will then drive our suggestions on what should be changed to improve the score (ie. student learning).

iv. How well does the model predict compared to alternatives?

- To answer this, we will examine 2 different models in our analysis and determine/compare their predictive capabilities.

2. Exploratory Data Analysis

In our initial analysis, we noticed that the output variable, score, does appear to be normally distributed. We did see that many of the of the explanantory variables do not appear to be linearly related to the output variable. This means we will not be able to use standard linear models when analyzing this data. Furthermore, several of the factors appear to be collinear. We will need to ensure our model is able to account for this or else it will not accurately predict.

3. Desired Attributes

Model attributes required from the research questions

The research goals require an model with the ability to identify important factors, and that can predict.

Model attributes required from the data

The data show the selected model will need to address collinearity amongst the explaining factors. The model also needs to work for non-linear relationships with the factors.

Any other anticipated problems

We do not anticipate any other problems in the data.

What goes wrong if the above are not accounted for?

If the above issues are not accoutned for, the model will not accurately predict school scores. Furthermore, the model will not correctly represent the relationship between the explanatory factors and the output variable, meaning it will not be useful to identify what can be done to improve score.

4. Proposed Method

Appropriate models

For this analysis, a model using a non-linear relationship, such as a polynomial spline or a kernel smoothing model will be appropriate.

Specific model proposal

We propose using a ploynomial spline model for this analysis, following the example format of:

$$Y = \beta_0 + \sum_{i=1}^{n} [\beta_{1_i} X_i + \beta_{2_i} X_i^2] + \epsilon : \epsilon \sim N(0, \sigma^2)$$

where n is the number of explanatory factors that end up being included in the model.

Method strengts/weakness

i. How this method accounts for issues in the dataset

We can use this modeling method to account for non-linearity in the data. This method also does not have any inherent assumptions about the data, so we do not need to concern ourselves with other factors such as equal variance, etc.

ii. How this method accomplishes research/analysis goals and yields appropriate estimators

This model accomplishes our research goals by allowing us to identify the relationship between the factors and the outputs, as well as provide predictions.

iii. How this method will answer the research questions

We will be able to review the relationships outlined in the model to see if there are diminishing returns on extracurricular activities, and identify if English as a second language is a barrier to learning. This will also allow us to see which factors are most impactful on score, and will indicate which actions we should suggest to improve score. Finally, this model does allow for prediction, and we will be able to compare it to other models which allow for prediction.

iv. What assumptions are needed to use the model adequately? Are they reasonable to assume and explained well?

One of the strengths of a polynomial spline model is that there are no inherent assumptions the data need to meet. Thus, we do not need to worry about any assumptions.