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 measure 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 secone 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 develop a model that can identify important factors in relation to score. From this we will be able to indicate what actions would most impact scores (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 their predictive capabilities.

2. Exploratory Data Analysis

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 non-normality of the output variable, and collinearity amongst the explaining factors.

Any other anticipated problems

The number of explaining factors is very large compared to the new number of available data points. We will need to account for this in our variable selection/importance, as basic variable selection methods will not perform well.

What goes wrong if the above are not accounted for?

Without meeting the requirements of the research questions, we will not be able to answer those questions. If we don't address collinearity, the calculated coefficient estimates won't be true. If we don't address non-normality, the standard models' likelihoods won't be appropriate. If we don't use an appropriate variable selection method, we will get a model that does not reflect the true effect of factors on output.

4. Proposed Method

Appropriate models

In this analysis, a linear regression model is appropriate, once having accounted for the issues identified previously.

Specific model proposal

I will use Oridnary Least Squares (OLS) regression using the format:

$$Y = \sum X_i \beta_i + \epsilon : \epsilon \sim N(0, \sigma^2)$$

Method strengts/weakness

i. How this method accounts for issues in the dataset

We can perform transformations in X to account for non-normality. We can use a vareity of methods (LASSO, PCR, PLS) variable selection/importance to account for collinearity and overcome the limitation of the ratio of data points to explaining factors.

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

For the goals of the analysis, regression can provide R^2 values, as well as predict, meeting the requirements of the research questions.

iii. How this method will answer the research questions

This method will answer the research questions by (for each question) 1) stating the included variables (or the relative significance of each variable), 2) state R^2 value and explain ideas for non-included variables that may be significant, 3) testing the predictive capabilities by cross validating on the data and providing a measure of prediction accuracy.

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

To properly use this model we muse assume linearity, independence, normality, and equal variance. Linearity appears safe to assume, and we have tools to address the other assumptions should they turn out not to be true.