HW2 Proposal

RJ Cass

1. Understanding the Problem

Background:

We have a record of various climatic, geographic, and biological factors around rivers in the Rocky mountain range. We want to use these data to understand what factors influence, and can be used to predict, waterflow through the rivers.

Goals:

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

i. What are the biggest climate / river network / human factors that impact overall river flow?

- We will perform variable selection/other fitting to determine which factors are most important to river flow. Given the quantity of factors compared the number of data points we will perform LASSO/Ridge, PCR, or PLS modeling to identify important factors.

ii. How well do the factors listed in #1 explain overall flow?

- To determine how well the included factors explain overall flow, I will calculate the R^2 value of the selected model.

iii. How predictive of overall flow are these identified factors?

- To determine how predictive the identified factors are of overall flow, I will perform cross validation and report the Root Mean Square Error.

2. Exploratory Data Analysis

In our initial exploration, the first thing that sticks out about the data is that flow is pretty heavily skewed with a long left tail. This violates the assumption of normality so we may need to perform a transformation. We also see indicators of strong collinearity between a large number of the factors and will need to account for that in our model. Also, there are a large number of factors when compared to available data points, so we will need to be careful in determining the coefficients of the factors. There are also 2 factors which just have constant values: we can effectively just ignore those columns in our analysis as they provide no useful information.

3. Desired Attributes

Model attributes required from the research questions

The research goals require an model with only significant factors (or the ability to identify important factors), that gives an R^2 value, and 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.