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? REVISIT ONCE LEARN STUFF

- We will perform variable selection to determine which factors are most important to river flow. Given the quantity of factors compared the number of data points we will use the LASSO method to select an appropriate model.

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.

3. Desired Attributes

Model attributes required from the research questions

The research goals require an model with only significant 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, 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.