HW2 Report

Brett Pedersen and RJ Cass

Abstract

We have a dataset of various factors at measurement sites in rivers in the Rocky Mountains, as well as the corresponding flow at that site. We used this data to create a linear model of 'Metric' (the measure showing flow rate at a site) to predict the flow rate of a river given various input factors. Using the modeling methods, we found that the factors that most impact the flow 'Metric' are X Y Z. We found that these factors explain X% of the variance in 'Metric'. We also identified that our predictive model has an RMSE of X.

1: Introduction

We have a collection of data measuring river flow rates of across the Rocky Mountains, with corresponding measurements indicating various human, river network, and climate factors. We want to use this data to create a model to 1) Understand which variables are most impactful in determining river flow rate 2) Identify how well our selected model explains the variance of flow rate 3) Quantify how successful our selected model is at predciting flow rate

Upon investigating the data, we see a few prominent issues that will need to be addressed in our analysis. As shown in Figure 1 - Left, the Flow Rate is not normally distirbuted. If not accounted for, the resulting model will not correctly relate the explanatory factors to the output. Another issue we identified is that there are many factors that exhibit strong collinearity (Figure 1 - Right). If we don't account for collinearity in our model, the estimated factor coefficients will not be reliable. The final primary issue we will need to address is that there are almost as many factors as there are data points. As such, we will need to be careful in our model selection to ensure we pick a model that can correctly identify which factors are important with limited data.

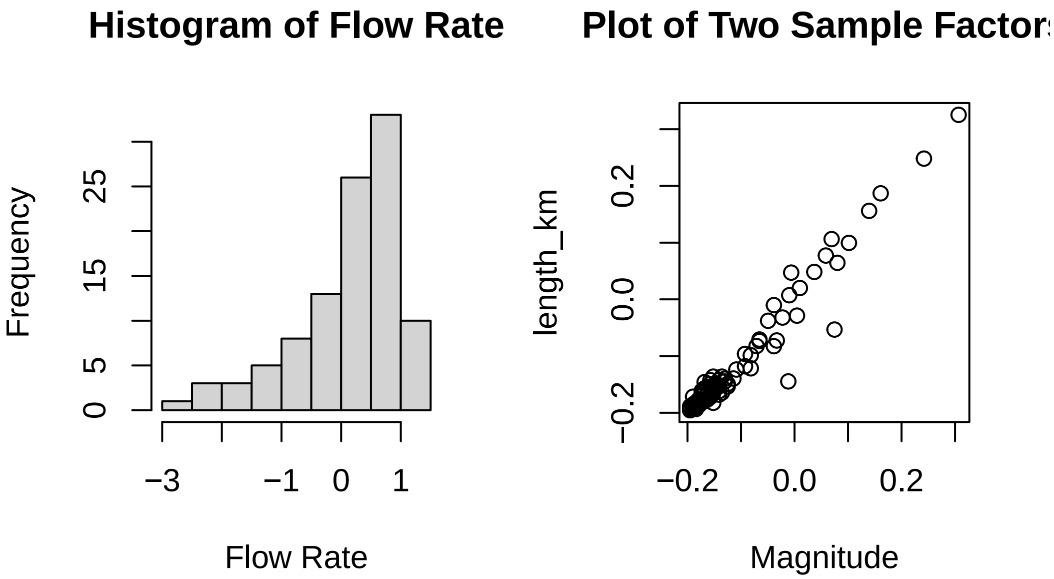

Figure 1: Left - histogram of Flow Rate. Note that the distribution is left skewed, showing non-normality. Right - a scatterplot of two sample factors (Magnitude and Length) showing collinearity.

2: Methodology

2.1 Proposed Models

Model 1

The first model we are proposing for this analysis is a Principal Component Regression (PCR) model. PCR models are particularly handy for working with datasets which have a large number of factors compared to data points. Also, they work well in handling collinear variables. However, this model does depend on the LINE assumptions, of which we have already indicated previously that normality is a concern. To account for this, we will need to perform a transform on the output variable to make it more closely represent a normal distirbution.

Model 2

The second model we are proposing is a LASSO model. LASSO models are useful in reducing the number of factors (including accounting for collinearity). They are also capable of performing well given a large number of factors compared to data points. This model in particular is useful because it does not require normality in the output variable so we will not need to perform any transformations on the data.

2.2 Model Evaluation

We created a model using both methods described above. To compare the models, we looked at the number of impactful factors, the R^2 value of the model, and the RMSE of the predcitions of the model.

Model	# Factors	R^2	RMSE
PCR	95	0.729	0.11
LASSO	21	0.805	0.387

Table 1: Comparison of the two models outlined previously. Note that LASSO has limited the selected factors to only those most important.

As shown in Table 1, LASSO has identified only the msot important factors (PCR does similar, but does not ‘zero-out’ the un-important factors). The R^2 for the LASSO model is better, but the RMSE for the PCR model is better. However, since we had to perform a transform on the PCR model, and the PCR model has such a larger number of factors, the LASSO model is much more explainable. Due to these factors, we are choosing to use the LASSO model to answer our research questions.

The selected model follows the format:

$$\begin{aligned} FlowRate = \beta_0 + \beta_1 Factor1 + \beta_2 Factor2 + \beta_3 Factor3 \\ + \beta_4 Factor4 + \beta_5 Factor5 + \beta_6 Factor6 + \beta_7 Factor7 \end{aligned} \tag{1}$$

3: Results

3.1 Estimation of Model Parameters