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 Flow Rate (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 10 of the original 97 factors most influence flow rate. We found that these factors explain 70.7% of the variance in Flow Rate. Our model achieved an out of sample RMSE of 0.519.

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.

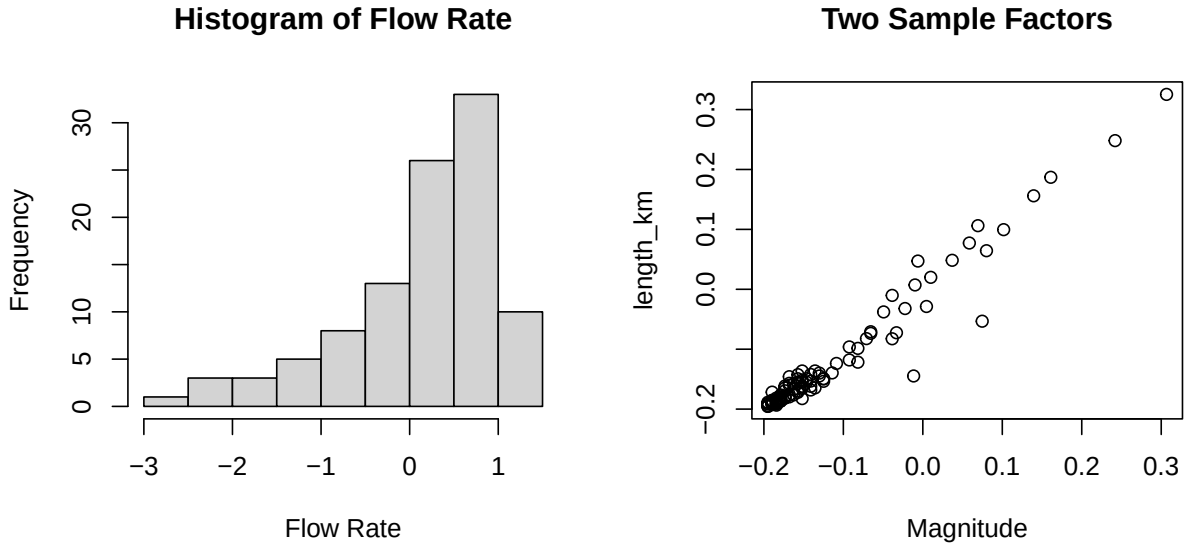

Figure 1: Left - histogram of Flow Rate. Note that the distribution is left skewed, showing non-normality. Middle - a scatterplot of two sample factors (Magnitude and Length) showing collinearity.

We also identified 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. Finally, we saw that while the Flow Rate metric does appear linearly distributed according to some of the factors (Figure 2 - Left), other factors do not appear to be linearly related to Flow Rate (Figure 2 - Right). If we do not account for this when constructing our model, then we will not be able to use a linear model.

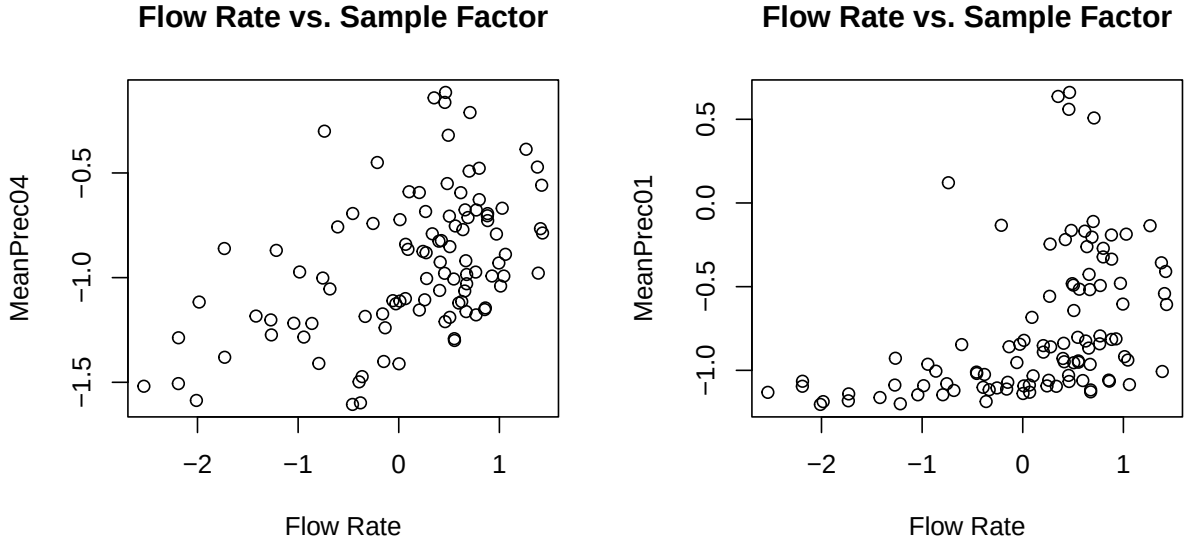

Figure 2: Left - a scatterplot showing linearity of Flow Rate vs. a sample factor. Right - a scatterplot showing non-linearity of Flow Rate vs. a different sample factor.

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.

This model will allow us to answer the research questions by providing coefficients for each of the factos which, if we scale them accordingly, will indicate relative importance in determining flow rate. We can also extract an R^2 value and a measure of the predictive power of the model.

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. One challenge with this model is that, while it is capable of handling colinear factors, the method through which it selects the impactful factors can be somewhat arbitrary.

This model will allow us to answer the research questions by reducing the factors to only those which have the most significant impact on the Flow Rate. We can also extract an R^2 value and a measure of the predictive power of the model.

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	0.758	0.106
LASSO	10	0.707	0.519

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 this format (for simplicity's sake, the names of the significant factors are excluded here and can be found in Table 2):

$$FlowRate = \beta_0 + \beta_1 Factor_1 + \beta_2 Factor_2 + ... + \beta_9 Factor_9 + \beta_{10} Factor_{10} \quad (1)$$

3: Results

The table below depicts the variables that were selected by lasso along with their associated beta coefficients. These coefficients have the typical linear regression interpretation (e.g. A 1 unit increase in gord corresponds to a 0.254 unit increase in Metric). Uncertainties are not given for the coefficients since lasso regression doesn't include distributional assumptions besides linearity.

	Variable	Coefficient
1	(Intercept)	16.150319241
2	gord	0.253666333
3	CumPrec03	1.389135105
4	CumPrec04	2.228612053
5	bio15	-0.273249801
6	cls1	0.010815459
7	cls2	55.961854567
8	cls5	-0.001884255
9	cls8	1.270039673
10	meanPercentDC_SomewhatExcessive	0.394794265
11	Lon	-0.039028936

We can use the lasso model to answer our research questions:

1. The most impactful factors for overall river flow are those that are included in the table above.
2. Model evaluation metrics were included in section 2.2. We can say that 70.7% of the variation in Metric is explained by the listed factors.
3. When fitting the model to a subset of the dataset and making predictions on data that was left out (out of sample), predictions for Metric typically differed from the true value by around 0.519.

4: Conclusions

By using a lasso regression model, we were able to answer the research questions given. We found 10 variables (out of the original 97) that were the most important for explaining river flow. We were able to explain 70.7% of the variation in river flow with our model, and achieved an out of sample root mean squared error of 0.519. One shortcoming of the chosen model is the lack of uncertainty quantification on the coefficient estimates. To address this, a good next step will be

to take repeated bootstrap samples of the data, fit a lasso model to each of them, and attain a distribution for each of the betas. Another next step will be to determine whether there are any meaningful interactions between variables in the data.