HW2 Report

Brett Pedersen and RJ Cass

Abstract

We have a dataset that contains the overall waterflow of various rivers in the U.S. Rocky Mountains along with nearly 100 potential explanatory factors. We used LASSO to create a linear model of the flow rate metric using a subset of these factors. With our model, we found that 10 of the original 97 factors could be used to explain 70.7% of the variation in flow rate with out of sample predictions being off by around 0.519.

1: Introduction

We have a collection of data measuring flow rates of 102 rivers across the U.S. Rocky Mountains along with measurements for various human, river network, and climate factors. With this data, our goal is to create a model to help us

1) Understand which variables are most impactful in determining river flow rate

2) Identify how well these variables explain the variation in the flow rate

3) Quantify how well we can predict the flow rate on unseen data

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 distributed. 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.

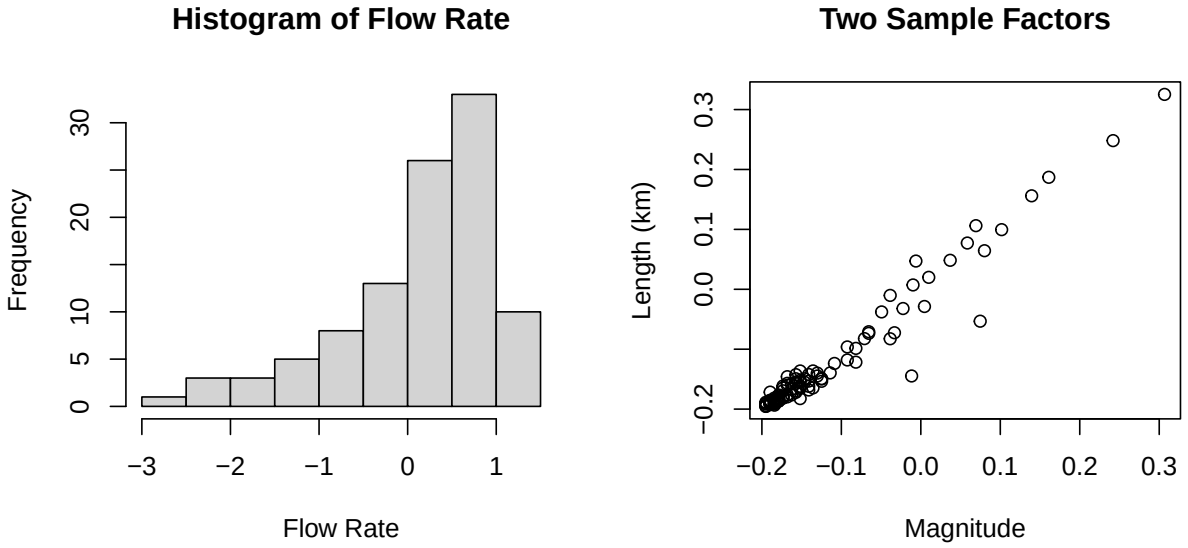

Figure 1: Left - histogram of Flow Rate. Note that the distribution is left skewed, showing non-normality. Right - a scatterplot of two sample factors (Magnitude and Length) showing strong collinearity.

Lastly, we noticed 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.

2: Methodology

2.1 Proposed Models

Model 1

The first model we are proposing for this analysis is Principal Component Regression (PCR). PCR models are particularly handy for working with datasets which have a large number of factors compared to data points. They also work well in handling collinear variables. However, this model does depend on all the LINE assumptions (linear, independence, normality, equal variance), of which we have already indicated that normality is a concern. To account for this, we will need to perform a transformation on Flow Rate to make it more closely follow a normal distribution.

This model will allow us to answer the research questions by providing coefficients for each of the factors 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 regression model. LASSO models are useful in reducing the number of factors (which also helps with collinearity). LASSO also only depends on the linearity assumption. We do not need to transform the output variable to satisfy the poor normality

issue. A disadvantage, however, is that this model potentially has less predictive power than other models. This is because it removes several variables that may have marginal relationships with the flow rate.

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 manually calculate an R^2 value, and the method automatically performs cross validation, outputting out of sample RMSE for us.

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 model's predictions.

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 most important factors (PCR does similar, but does not 'zero-out' the unimportant factors). The PCR has superior R^2 and RMSE metrics. However, since we had to perform a transformation on the response with the PCR model, and it has such a larger number of factors, the LASSO model is much more explainable. As a result, we are choosing to use the LASSO model to answer our research questions. The linearity assumption is validated in Figure 2 below.

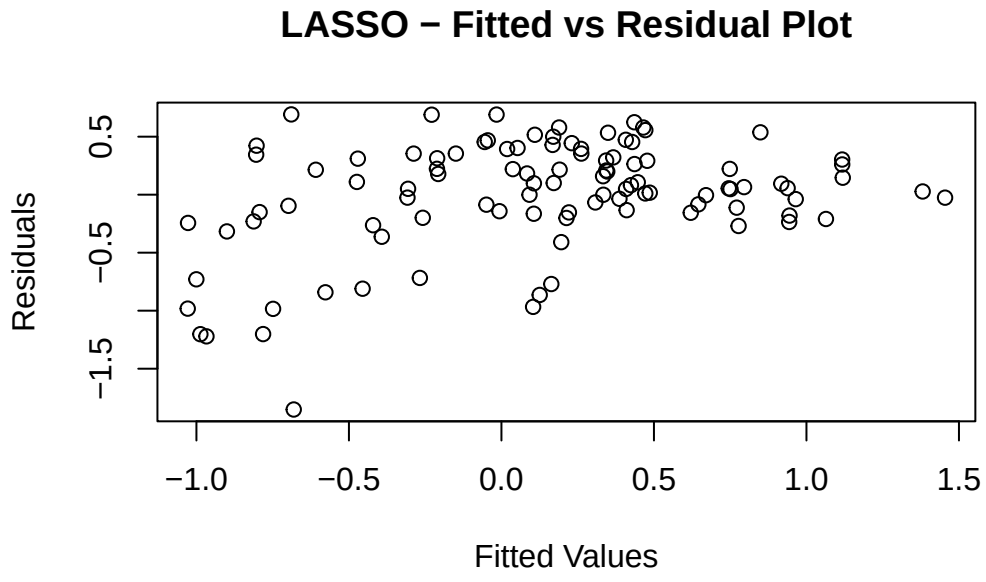

Figure 2 - Plot of the residuals vs the fitted values for our LASSO model. The linearity assumption seems reasonable, which is the only assumption necessary for this model choice. The abnormal behavior is more an indication of poor homoscedasticity.

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 Flow Rate). Uncertainties are not given for the coefficients since lasso regression doesn't include any distributional assumptions on top of linearity.

Variable	Coefficient
(Intercept)	16.150319241
gord	0.253666333
CumPrec03	1.389135105
CumPrec04	2.228612053
bio15	-0.273249801
cls1	0.010815459
cls2	55.961854567
cls5	-0.001884255
cls8	1.270039673
meanPercentDC_SomewhatExcessive	0.394794265
Lon	-0.039028936

Table 2: Output of the LASSO model with the 10 most important explanatory variables for Flow Rate and their coefficients.

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 Table 2.
2. Model evaluation metrics were included in Table 1. We can say that 70.7% of the variation in Flow Rate is explained by the chosen 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 Flow Rate 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.

RJ - PCR model, Abstract, Introduction, Methodology

Brett - LASSO model, Results, Conclusion, Final Proofreading/Editing