Global River Flow Analysis
Stat 536 - Section 1 - Dr. Heaton

Jacob Andros & Michael Shull

February 9, 2021

Abstract

Dihydrogen monoxide, commonly known as water, is an important resource for almost all forms of life on planet Earth. For this reason, it is important for us to understand the factors that lead to better (and worse) water flow. In this analysis we seek to understand what factors contribute to the water flow of rivers in the Rocky Mountains using the techniques of Least Absolute Shrinkage and Selection Operator (LASSO) and Principal Component Analysis (PCR).

Keywords: LASSO, PCR, Shrinkage, Regularization, Dimension Reduction

Introduction

In the Rocky Mountains, river flow accounts for the majority of the water that its inhabitants source for farming and for everyday life. To ensure that inhabitants of this region in the U.S. have consistent and ongoing access to enough water, it is important to understand what factors might contribute to river flow. In this analysis, we examine the effects that various factors from climate, humans, and the river network have on overall river flow.

By the end of the analysis, we wish to:

- Uncover which factors from the climate, human presence, and river networks have the most significant effect on water flow,
- Determine whether said factors explain water flow overall, or if different predictors are required, and
- Assess the capability of said factors in predicting water flow.

The data used for this analysis includes measurements for nearly 100 variables from different river locations across the mountain west states. Some of these variables are climate-related, such as average monthly temperatures, precipitation levels, and vegetation levels. Other variables are characteristics of the local river networks, such as elevation and drainage density. Many factors are the results of human interactions, like the population density. Regardless, there are too many covariates to be described in detail individually. The dataset also includes a standardized variable that indicates the overall water flow for each of the 102 river locations sampled. This response variable will be referred to as *Metric*, and has a slightly left skewed distribution centered near 0, as seen in Figure 1.

Having such a large number of covariates in the dataset creates a number of issues when trying to create a statistical model. First, when they are this many explanatory variables available, some of them are bound to be collinear. Using even just a few explanatory variables that are already correlated with each other is redundant in prediction, and leads to inflated standard errors in statistical inference. In this dataset, for example, there are 22 different variables dealing just with temperature, and 32 dealing with precipitation. More specifically, *bio7*, *CumPrecTotal*, *MeanPrecAnn*, and *MeanTempAnn* are all linear combinations of other variables. The collinearity between some of these covariates is illustrated in the scatterplot matrices in Figures 2 and 3.

Second, there are two variables in the dataset for which every river has the exact same measurement. These variables are the mean well-drained class and mean very poor-drained class (in reference to land cover).

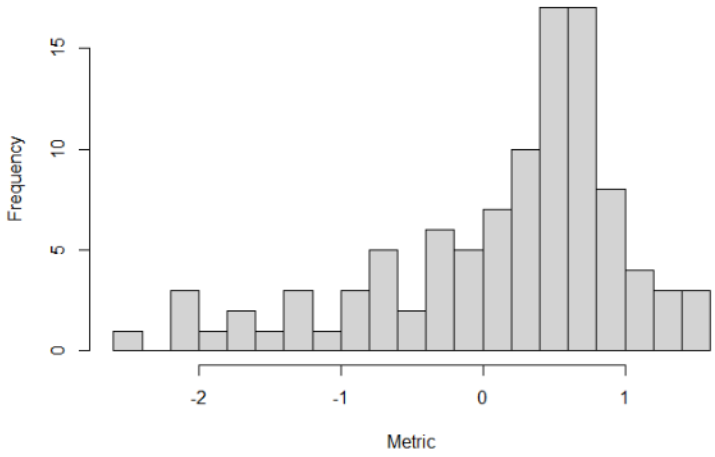

Figure 1: Distribution and Range of Metric Variable

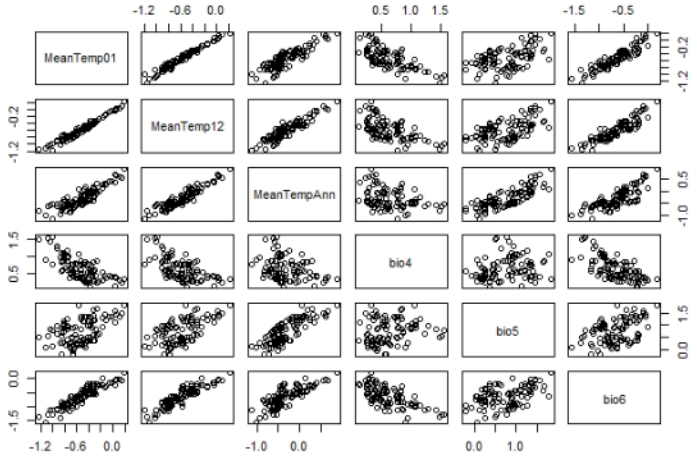

Figure 2: Temperature Variable Collinearity

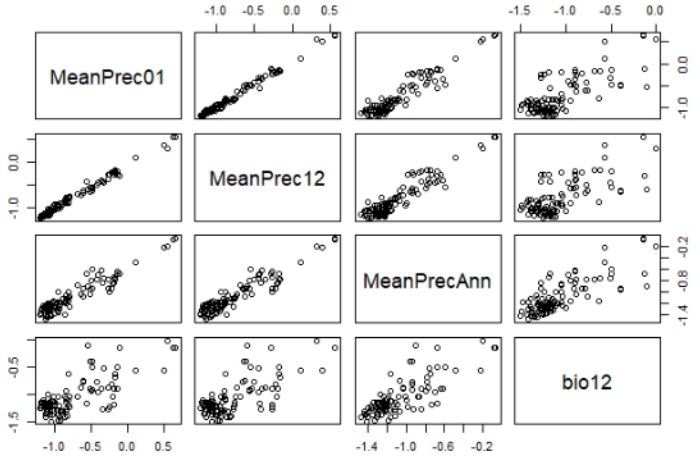

Figure 3: Precipitation Variable Collinearity

Attempting to include these variables in the model would create singularity in the matrix multiplication. It would also contribute nothing to the model - since every river has the same measurement for these variables, there is no way we can attribute a difference in *Metric* to either one.

Finally, even after accounting for collinearity and removing redundant predictors, it is likely that there will still be many explanatory variables remaining. Though this is not necessarily a bad thing, it would be difficult to pinpoint which variables have the greatest effect on *Metric* if there are too many in the model. As a result, we will use statistical techniques to further reduce the number of predictors used in the model.

Methods

For this analysis we used two different methods in order to reduce the dimensionality of the problem so that we could identify covariates that have an important effect upon our metric of water flow. The methods are Least Absolute Shrinkage and Selection Operator (LASSO) which is a regularization technique and Principal Component Regression (PCR) which is a dimension reduction technique. The benefit of these techniques is that we are able to use all of the data in the model while still accounting for collinearity and not overfitting to the data. The main method used in this analysis was LASSO which we will describe in detail. For a detailed explanation of PCR, we refer you to other papers on the subject.

LASSO is a Regularization technique that depends upon the idea of penalized least squares. Let y_i refer to the i^{th} observation in the data set, $\mathbf{x}'_i$ be a row vector of covariates associated with that observation, N be the number of observations, and P be the number of covariates in the model. Here we assume that the relationship between the covariates and response is that given in equation (1). The idea is to find a vector of constants, $\boldsymbol{\beta}$ (the partial slopes) that minimizes the equation found in equation (2). Here we call λ the shrinkage parameter and $\text{Size}(\beta_p)$ is a measure of how big $\boldsymbol{\beta}$ is, commonly called the shrinkage penalty. The selection of λ is normally through cross validation. The higher the value of λ, the more the model is penalized and the smaller the values of the coefficients in $\boldsymbol{\beta}$ are. One of the benefits of this model is that there is only one assumption to check which is that there is a linear relationship between each of the covariates and the response variable.

$$y_i = \mathbf{x}'_i\boldsymbol{\beta} \tag{1}$$

$$\sum_{i=1}^{N}(y_i - \mathbf{x}'_i\boldsymbol{\beta})^2 + \lambda\sum_{p=1}^{P}\text{Size}(\beta_p) \tag{2}$$

When using LASSO, we set $\text{Size}(\beta_p) = |\beta_p|$. This is quite nice in practice because this will cause many of our coefficients to be zeroed in the model. This is basically a built in way of letting us know which variables have the largest impact on the response. One of the things to note about LASSO is that we make no distributional assumptions. This means that we do not have sampling distributions that we can use in order to perform inference. However, we can always use bootstrap sampling to approximate sampling distributions and perform inference that way.

Model Justification

As mentioned in the introduction, there are two covariates that have a standard deviation of zero, meaning that they take on the same value for all rivers (mean well-drained class and mean very poor-drained class with reference to land cover). There is another variable, *cls2*, which refers to Broad-leaf Evergreen land coverage that has a standard deviation of approximately 0.001 with a range of about 0.008. This predictor varies so little that when we include it in the model gets rather large estimates of its coefficient. Because of this we dropped all three of these variables from the dataset before beginning the analysis.

One of the difficulties that we face when we use LASSO is deciding how to select the value of λ that we wish to use. As mentioned in the methods section cross-validation is the typical way to select this. As the dataset only has 102 observations, we opted to use leave one out cross-validation using mean squared error as the metric for deciding which value of λ to use. After performing our cross-validation we ended up using a value of $\lambda = 0.02245938$.

In fitting the LASSO model we have assumed that there is a linear relationship between the covariates and our response variable. In the large dimensional space this is very hard to check. If we were doing ordinary least square's regression we could easily construct added-variable plots. However in this penalized regression regime we cannot do such a thing. We seek to imitate this behavior for six of the covariates in our model: *gord*, *CumPrec05*, *cls1*, *cls8*, *meanPercentDC_Poor*, and *meanPercentDC_SomewhatExcessive*. We will use *gord*, or global stream order, as an example for how these plots were constructed. First we took our $\mathbf{X}$ matrix and deleted the column corresponding to *gord*, denoted $\mathbf{X}_{\text{-gord}}$, and we took our $\boldsymbol{\beta}$ vector and deleted the *gord* coefficient which we denoted $\boldsymbol{\beta}_{\text{-gord}}$. We then calculated the residual of *metric* given all covariates other than *gord* as $\mathbf{Y} - \mathbf{X}_{\text{-gord}}\boldsymbol{\beta}_{\text{-gord}}$. We plotted this against *gord* and added a regression line. You can see the plots for all of six of these covariates in Figure 4. Looking at these six variables we say that these look linear enough for this assumption to be met. Since there are so many variables to look at, we don't look at all of the others and consider the assumption of linearity met.

When fitting a model, it is important to get some sort of idea of how good the model is at explaining the response variable and how good it is at predicting. The LASSO model has an R^2 value of 0.8044668 which means that 80% of the variance in *metric* is explained by the model. We also performed leave one out cross-validation to access the prediction error. With a root mean square error of 0.4803828, we feel that the model does an adequate job of predicting. A comparison of R^2, RMSE, and Bias for the LASSO and PCR models is given in Table 1. Later in the analysis, these figures will be compared to the spread of water flow in the dataset, and a more detailed description of what they each mean will be given.

Statistic	**LASSO**	**PCR**
R^2	0.804	0.731
Bias	0.010	0.015
RMSE	0.48	0.51

Table 1: Comparison of LASSO and PCR

Model Performance and Results

We have now constructed two models using LASSO and principle component regression. It was shown previously that the LASSO regression model meets the necessary assumptions while also reducing the number

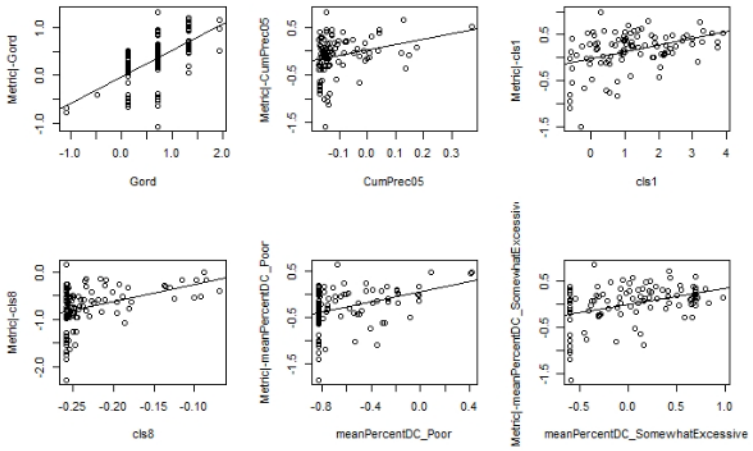

Figure 4: Marginal Effects of Selected Covariates

of coefficients to keep track of. The LASSO model also provides better overall fit and less prediction error than the PCR model. For that reason, we use the LASSO model for the remainder of the analysis.

We begin by examining the estimated effect of each coefficient on *Metric*. These effects, as well as their corresponding 95% confidence intervals, are given in Table 2. Note that these confidence intervals are built using a bootstrapping procedure. In bootstrapping, the river observations are sampled with replacement such that some observations are drawn multiple times, and some not at all. A LASSO model is fit using the same value of λ but with the new bootstrap sampled data. The coefficient estimates for each of the original nonzero coefficients are pulled from the model. This process was repeated $n = 1000$ times, and confidence intervals were taken as the 2.5% and 97.5% quantiles of each coefficient's sample data. LASSO regression has the tendency to zero out coefficients as much as possible, which means that many of the confidence intervals contain a zero as either the upper or lower bound.

These coefficients are estimates of the expected increases in *metric* for each unit increase in the corresponding covariate while holding all else constant. As an example, we take the *bio1* variable, which represents the standardized mean annual temperature. Based on this LASSO regression model, we expect that for every additional standardized unit of mean annual temperature, the average river flow (as measured by *metric*) would decrease by 0.09 if all other variables hold the same. Furthermore, for the mean annual temperature, we are 95% confident that this effect value is somewhere between -0.39 and 0. Since 0 is in the interval, we can't firmly conclude that mean annual temperature actually has a significant effect on *metric*. Many of the coefficients include 0 in their interval, which means that, although they are all part of predicting *metric* using this particular model, they are not all necessarily significant.

The only coefficient that was almost never zeroed out in the bootstrap sampling (besides the intercept) was *gord*, or the global stream order. It seems then that global stream order is the most significant predictor in determining water flow. The other variables that were selected in the LASSO model are mainly either individual monthly precipitation levels (April through September), mean annual temperature, measure of the presence of certain plants, or drain class. Again, these predictors may or may not have a statistically significant relationship with water flow, but are part of the model's predictions nonetheless.

Having established which variables have the most impact on water flow, we asses how much variation in *metric* is actually accounted for by the covariates in the model. The R-squared value from the LASSO model is 0.804. This means that 80.4% of the variation in *metric* is explained by the covariates described in the previous paragraphs. Though there is room for improvement, this is a reasonably high value for r-squared, and shows that the LASSO model fits the data quite well. R-squared would likely be higher if more covariates were included, but then there would be more coefficients to keep track of and model interpretability could suffer.

Finally, we wish to examine the predictive capabilities of this LASSO model. There are two key measures used to do this - bias and RMSE (root mean square error). Bias indicates whether the model tends toward

Coefficient	Estimate	Lower	Upper
Intercept	-3.60	-6.85	2.81
gelev_m	0.03	0	0.12
Gord	0.51	0.08	0.66
MeanTemp08	-0.01	-0.28	0
MeanPrec07	-0.44	-1.20	0
CumPrec04	1.06	0	3.90
CumPrec05	0.92	0	2.27
CumPrec09	0.58	0	2.64
bio1	-0.09	-0.39	0
bio15	-0.11	-0.35	0
bio16	-0.22	-0.55	0
bio18	-0.09	-0.80	0
cls1	0.13	0	0.32
cls5	-0.09	-0.26	0
cls8	3.15	0	4.46
HydroLakes Area	0.98	0	8.54
Percent DC Mod. Well	-0.16	-0.34	0
Percent DC Poor	0.48	0	0.82
Percent DC S.E.	0.30	0	0.71
Longitude	-0.04	-0.07	0

Table 2: Effect Estimates on *Metric* using LASSO Model

under-predicting or over-predicting the water flow of new rivers, and RMSE is the average error of the model's predictions. Both were obtained in a leave-one-out cross validation procedure. The mean bias was calculated to be 0.01, and the RMSE was 0.48. This means that the LASSO model tends towards just barely over-predicting water flow, and that its predictions for *metric* are off by about 0.48, on average. It is useful to compare the RMSE to the actual spread of *metric*. The response variable's standard deviation is 0.88 and the range is 3.96. On average, the LASSO model's prediction error is equal to approximately half of the standard deviation of *metric*, and equivalent to about one-ninth of the range. Though this could be better, it means that the model will usually be able to generate reasonable and fairly accurate predictions for new rivers' water flow given measurements for the covariates in the model.

Conclusion

In this analysis, we set out to determine which variables had the most impact on river flow while assessing the overall fit and predictive capabilities of those variables. We fit LASSO and PCR models and demonstrated that the LASSO model adequately met all necessary assumptions. The LASSO model included variables such as global stream order, mean annual temperature, presence of certain plant species, population density, and drain class and showed how each one could positively or negatively affect river flow. We showed that this LASSO model fit the data adequately using R-squared and that it had fairly accurate predictions using the bias and RMSE from cross-validation.

One of the biggest flaws of the LASSO and PCR models is that they assume there is independence between observations. Since many of the rivers sampled were in geographical proximity to each other, this assumption could be violated due to spatial correlation between river flows. In addition, there were some variables in this dataset that had zero variance or near-zero variance. We simply removed these variable before performing any regression, but it may be useful to study these variables more and find out if there are suitable alternatives.

To build upon this analysis, there are a few things that should be considered. To fix the independence issue just described, we could fit a spatial correlation model to account for the latitude and longitude. of each river in relation to each other. To do this, it would be helpful to avoid LASSO or PCR regression altogether, and to instead just select certain variables that are known to be impactful on water flow. We

could use the variables from the LASSO model in this analysis, or attempt other regression methods (ridge or partial least-squares regression) to further study which predictors are most helpful. Furthermore, expanding the analysis to include other areas of the United States may help explain how some of the predictors in the dataset interact with the regional conditions of the Rocky Mountains (low humidity, high elevation, etc.).

Appendix: Teamwork

Coding: Both.
Abstract: Michael
Introduction: Jacob
Methods: Michael
Model Justification: Michael
Model Performance and Evaluation: Jacob
Conclusion: Jacob
Figure 1: Jacob
Figure 2: Michael
Figure 3: Michael
Figure 4: Michael
Table 1: Jacob
Table 2: Jacob