Case Study 2

Rocky Mountain River Drainage

Sam Lee & Patric Platts

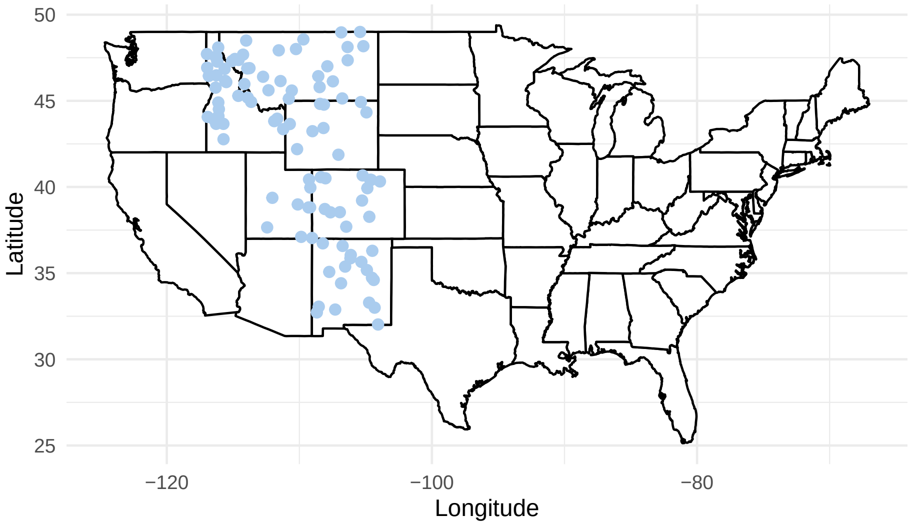

Figure 1: Scatter plot of spatial displacement of each observation of recorded river flow

1

After standardizing the numerical covariates, we will fit two variable selection models: We will fit a lasso penalized regression model and a partial component regression (PCR) model. After fitting our models, we will apply a level of k-fold cross-validation to select the best hyperparameters for the model of best fit. After selecting our chosen model, we would examine the covariates that are the largest in absolute value (since the model would fit on scaled vectors) and which are significant.

2

We would look at the adjusted R^2 statistic as an overall model fit adjusted for the number of covariates included. We would also compare how well our selected covariates from the lasso regression compare to the full model fit by running an ANOVA test of the two models (comparing their f-statistics of significance).

3

To assess the predictive power of the model, we would look at the (out-of-sample) RMSE through k-fold cross-validation. To specifically look at the predictive power of the covariates we included, for a model like PCR, we could assess the portion of variance explained by each component as well as the total cumulative explained variance[1]. For our lasso regression model, in addition to RMSE as assessed through cross-validation and the adjusted R^2, we would create (B # of) bootstraps of the data and fit the lasso model on each bootstrap, checking the frequency at which each covariate is selected[2]. We can also use bootstrap to obtain confidence intervals for each coefficient as see if these are statisticaly significant.

Methodology

To reduce potential colinearity between the different factors in the data set and arrive at an optimal parsimonious model, we propose two models to assess the overall water flow of water sources in the Rocky Mountains. In this section, we will discuss both candidate models and how these models can be used to answer the research questions at hand.

We first propose a Partial Component Regression (PCR) model. PCR combines Principal Component Analysis (PCA) with linear regression. Under the assumption that the parameters of interest (β) are linear—that is, assuming a one-unit increase in a pth factor (among those we consider) implies a β_p increase in the water flow metric—we leverage this by applying linear regression to the set of orthogonal components computed by PCA[3]. We used ten-fold cross-validation to select the most optimal number of components, k^*; through this process, we chose $k^* = 9$. The strengths with PCR come with its robustness to multicolinearity in the covariate matrix, X. Additionally, PCR performs dimensionality reduction by only selecting the top principal components (in our case, we selected 9) to achieve a parsimonious model.

The tradeoff that comes with using PCR, however, is its lack of interpretability. Since each component is a linear combination of all individual covariates in X, the coefficients derived from our PCR model are not directly interpretable. Additionally, PCR computes and therefore selects components based on the variance of the covariate matrix X, as opposed to each factor's relationship with our response variable, the metric of water flow. Hence, the components may not necessarily contribute to predicting the outcome of interest.

Secondly, we propose fitting a Lasso Linear Regression model to accomplish both dimension reduction through variable selection and interpretability. Similar to our PCR model, we will operate on the assume that each of our factors have a linear effect on the water flow metric. However,

[1] We would assess this using the criteria established by the following expression: $\sum_{j=1}^{k} \frac{\lambda_j^2}{\sum_{i=1}^{p} \lambda_i^2}$, where each λ_j corresponds to the singular value of the j-th principal component.

[2] We can assess the selection frequency of each j-th covariate by using this bootstrapping method as described through the following expression: $\frac{1}{B} \sum_{b=1}^{B} \mathbb{1}(\hat{\beta}_j^{\text{lasso, } b} \neq 0)$.

[3] We first orthogonalize the set of all factors of interest, X, through singular-value decomposition, where $X = U\Sigma V'$. We then compute $Z_k = XV_k$, for k number of components where V_k is a subset of V consisting of the first k columns of V. Each column of Z_k is then orthogonal to each other, that is, $Z_i' Z_j = 0 \; \forall i \neq j$. Then, performing linear regression, we compute the set of linear γ_k coefficient parameters (where γ_k is of dimension k) using Z_k as the new covariate matrix. Solving for γ_k, $\gamma_k = (Z_k' Z_k)^{-1} Z_k' Y$. $\hat{Y}$ is then computed as $\hat{Y} = Z_k \gamma_k$.

after standardization on the matrix X, Lasso Regression imposes an L_1 penalty[4] to both shrink the estimated coefficients and perform variable selection. Our Lasso Regression model is also suited to handle multicolinearity through the penalization parameter. However, unlike PCR, we can focus on predictive power since there is a direct relationship between the water flow metric and its covariates. Hence, we believe this model to be more interpretable.

When we introduce the penalty parameter, however, the coefficients on this model will be biased. We sacrifice this bias however for a decrease in the variance of the parameters. As a result, to accurately assess the standard error of each covariate effect, we perform bootstrapping methods to estimate 95% confidence intervals on β.

$$\arg\min_{\beta} \sum_{i=1}^{n} (y_i - x_i'\beta)^2 + \lambda \sum_{p=1}^{P} |\beta_p|$$

Results

Table 1: Lasso Coefficient Estimates

Covariate	Description	Estimate	95% CI	Inclusion Frequency
(Intercept)		0.125*	(0.017, 0.208)	1.000
bio10	Mean Temperature of Warmest Quarter (degrees Celsius)	0	(0, 0.102)	0.207
bio15	Precipitation Seasonality (Coefficient of Variation) (milimeter)	-0.188*	(-0.346, -0.079)	0.990
bio18	Precipitation of Warmest Quarter (milimeter)	-0.014	(-0.028, 0.113)	0.482
cls1	Evergreen_Dec_Needle_Trees (percent)	0.034	(-0.205, 0.068)	0.664
cls2	Evergreen_Broadleaf (percent)	0.082	(-0.126, 0.163)	0.934
cls5	Shrubs (percent)	-0.019	(-0.037, 0.16)	0.478
cls8	Regularly Flooded Vegetation (percent)	0.081*	(0.022, 0.161)	0.846
CumPrec03	Cumulative March Precipitation for the Watershed Upstream of Grdc Station (milimeter)	0.049	(-0.157, 0.097)	0.546
CumPrec04	Cumulative April Precipitation for the Watershed Upstream of Grdc Station (milimeter)	0.132*	(0.029, 0.264)	0.494
gord	Global Stream Order from Stream Dem (Predicted Relationship with Area) (categorical)	0.174*	(0.071, 0.348)	0.950
Lon	Longitude	-0.177*	(-0.354, -0.096)	0.750
meanPercentDC_Poor	Mean Poorly Drained Class (percent)	0.03	(-0.06, 0.059)	0.482
meanPercentDC_SomewhatExcessive	Mean Somewhat Excessive Drainage Class (percent)	0.18*	(0.046, 0.359)	0.946
MeanPrec07	Mean July Precipitation for the Watershed Upstream of Grdc Station (milimeter)	-0.002	(-0.003, 0.208)	0.536
MeanTemp05	Mean May Temperature (degrees Celsius)	-0.002	(-0.005, 0.083)	0.144

[4] Formally, the L_1 penalty is computed as a vector norm ($||\cdot||$), where, for a vector β with dimension P, $||\beta|| = \sum_{p=1}^{P} |\beta_p|$.

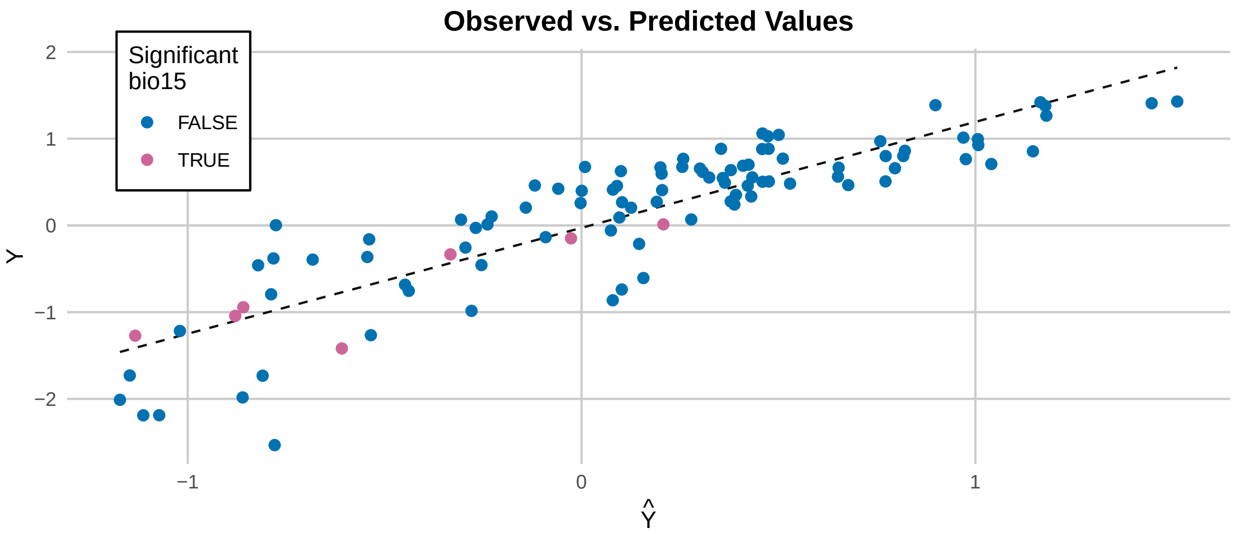

Figure 2: A comparison of actual and predicted values using Lasso Regression

Make a table with coefficients, description, effects, and confidence interval of bootstrap distributions

For in-sample fit, we computed an adjusted-R^2 of....

Using LOOCV, we computed an out-of-sample RMSE of....So on average the out-of-sample prediction is [LOOCV RMSE] away from the actual measurement of river flow.