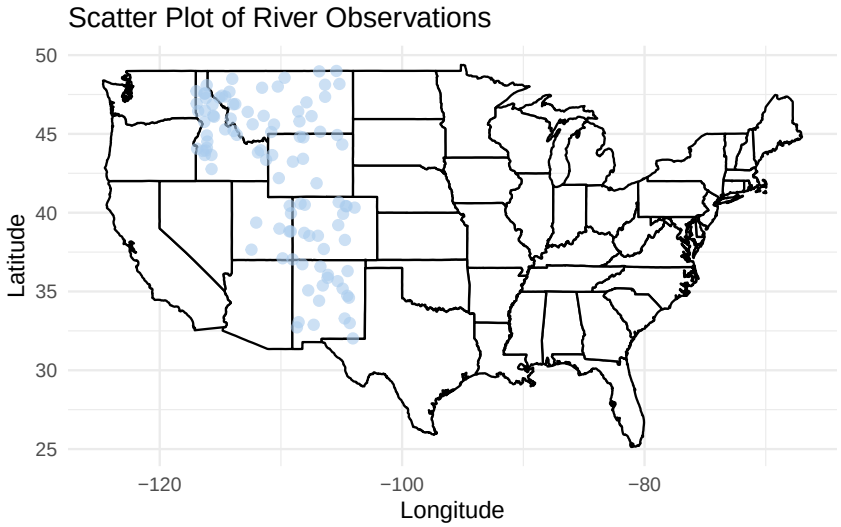
Case Study 2

Sam Lee & Patric Platts



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

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

Results

Table 1: Lasso Coefficient Estimates with Bootstrap Confidence Intervals, Descriptions, and Units

Covariate	Description	Estimate	95% CI	Inclusion Frequency
(Intercept)		0.127*	(0.02, 0.211)	1.000
bio10	Mean Temperature of Warmest Quarter (degrees Celsius)	-0.005	(-0.009, 0.093)	0.207
bio15	Precipitation Seasonality (Coefficient of Variation) (milimeter)	-0.186*	(-0.343, -0.076)	0.990
bio18	Precipitation of Warmest Quarter (milimeter)	-0.016	(-0.033, 0.109)	0.482
cls1	Evergreen_Dec_Needle_Trees (percent)	0.035	(-0.202, 0.071)	0.664
cls2	Evergreen_Broadleaf (percent)	0.08	(-0.129, 0.161)	0.934
cls5	Shrubs (percent)	-0.024	(-0.049, 0.149)	0.478
cls8	Regularly Flooded Vegetation (percent)	0.081*	(0.022, 0.162)	0.846
CumPrec03	Cumulative March Precipitation for the Watershed Upstream of Grdc Station (milimeter)	0.056	(-0.143, 0.111)	0.546
CumPrec04	Cumulative April Precipitation for the Watershed Upstream of Grdc Station (milimeter)	0.127*	(0.02, 0.254)	0.494
gord	Global Stream Order from Stream Dem (Predicted Relationship with Area) (categorical)	0.172*	(0.067, 0.344)	0.950
Lon	Longitude	-0.172*	(-0.343, -0.085)	0.750
meanPercentDC_Poor	Mean Poorly Drained Class (percent)	0.028	(-0.063, 0.057)	0.482
meanPercentDC_SomewhatExcessive	Mean Somewhat Excessive Drainage Class (percent)	0.175*	(0.037, 0.35)	0.946
MeanPrec07	Mean July Precipitation for the Watershed Upstream of Grdc Station (milimeter)	-0.002	(-0.004, 0.207)	0.536

[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: $\frac{1}{B} \sum_{b=1}^{B} \mathbb{1}(\hat{\beta}_j^{\text{lasso, } b} \neq 0)$.

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

```
[1] 0.738876
```

```
[1] 0.696856
```

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.