December 2, 2016

1 Methods

This project involved the use and implementation of 3 different linear models as well as hypothesis testing.

1.1 Ridge Regression

Ridge Regression is the first of the two shrinkage modelling methods we used. When using shrinkage methods the goal is to penalize certain parameters that should have a less significant effect on the model. To do so we use the tuning parameter, λ, times $\sum_{j=1}^{p} \beta_j^2$ to yield the shrinkage penalty. We then determine the coefficients that minimize the following equation:

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij}) + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{1}$$

To find these coefficients we used the function *cv.glmnet*() to determine through cross validation which λ value minimizes the above function (when λ was a given sequence of numbers under *grid*, and *alpha* was set to 0). From there we calculated the mean squared error (a measure of predicted power), by using the function:

$$MSE = \frac{1}{n} \sum_{i=1}^{n}(y_i - y_{di})^2 \tag{2}$$

1.2 Lasso Regression

Lasso regression is the second and final shrinkage modelling method used in this project. Although it is very similar to ridge regression there are a few key differences: the shrinkage penalty is now λ, times $\sum_{j=1}^{p} \beta_j^2$ (β is not squared), and lasso allows for the removal of certain variables (not just dampening their effect). To determine the coefficients we must look for the β's that minimize:

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij}) + \lambda \sum_{j=1}^{p} |\beta_j| \tag{3}$$

Again, to find the coefficient we used $cv.glmnet()$ and set λ to $grid$. However, now, $alpha$ was set to 1. Finally, we calculated MSE to determine the predictive power of our model.

1.3 Principle Components Regression

PCR is the first of two dimension reduction modelling methods we used. This method labors under the assumption that a subset of all of the predictor variables account for the vast majority of the variance. These more significant variables are referred to as principle components (M). PCR works by setting M equal to some reduced number of variables and running cross validation on, the model with the lowest cross validation error is selected.

To develop a model through PCR we used the $pcr()$ function and set $validation$= CV. We then found the model in which PRESS was larger to avoid overfitting. Finally, we calculated MSE to determine the predictive power of our model.

1.4 Hypothesis Testing

Hypothesis tests allows for comparison across groups. By separating the data points into categories based on how much funding the school received, we can then compare those categories based on unemployment rate and see how they differ.

To utilize this method we:

- Separated the data into groups by amount of funding the school received (low, low-mid, mid-high, high).
- Created hypothesis test, $W(u)$, to describe the difference being analyzed.
- Computed the estimate for effect of funding and variance of that effect.