Homework 2, STAT365

Girish Sastry

February 14, 2011

1 Exercise 1

Without loss of generality assume that $\beta_0 = 0$ (if not, shift the y_i terms such that it is; no coefficient $\beta_1, \beta_2, ..., \beta_p$ will be affected). Then, we rewrite:

$$\hat{\beta}^{lasso} = \arg\min_{\beta} = \left(\frac{1}{2}||\mathbf{y} - \mathbf{x}\beta||^2 + \lambda \sum_{j=1}^{p} (\beta_j) \right)$$

To calculate β^{lasso}, differentiate w.r.t. β:

$$\frac{\partial}{\partial \beta} \left[\frac{1}{2} (\mathbf{y} - \mathbf{x}\beta)^T (\mathbf{y} - \mathbf{x}\beta) + \lambda \sum_{j=1}^{p} |\beta_j| \right] = \frac{1}{2} \left(-2\mathbf{x}^{\mathbf{T}}\mathbf{y} + 2\mathbf{x}^{\mathbf{T}}\mathbf{x}\beta \right) + \lambda \mathbf{s}$$

where

$$\mathbf{s} = (s_1, s_2, ..., s_p)^T = (sign(\beta_1), sign(\beta_2), ..., sign(\beta_p)$$

Note that this is not differentiable at $\beta = 0$.

Thus we now have to minimize:

$$\begin{aligned}
& -\mathbf{x}^{\mathbf{T}}\mathbf{y} + \mathbf{x}^{\mathbf{T}}\mathbf{x}\beta + \lambda\mathbf{s} = 0 \\
& \Rightarrow \mathbf{x}^{\mathbf{T}}\mathbf{x}\beta = \mathbf{x}^{\mathbf{T}}\mathbf{y} - \lambda\mathbf{s} \\
& \Rightarrow \hat{\beta}^{lasso} = (\mathbf{x}^{\mathbf{T}}\mathbf{x})^{-1}\mathbf{x}^{\mathbf{T}}\mathbf{y} - (\mathbf{x}^{\mathbf{T}}\mathbf{x})^{-1}\lambda\mathbf{s} \\
& \Rightarrow \hat{\beta}^{lasso} = \beta^{\hat{OLS}} - (\mathbf{x}^{\mathbf{T}}\mathbf{x})^{-1}\lambda\mathbf{s}
\end{aligned}$$

Because the inputs are orthogonal, $\mathbf{x}^{\mathbf{T}}\mathbf{x}$ is diagonal and $(\mathbf{x}^{\mathbf{T}}\mathbf{x})_k = \sum_{i=1}^{n} x_{ik}^2$ Next, we proceed casewise. First suppose that $\hat{\beta_k}^{lasso} > 0$. Then $\hat{\beta_k}^{LS} - \gamma sign(\hat{\beta_k}^{lasso}) = \hat{\beta_k}^{LS} - \gamma > 0 \Rightarrow \hat{\beta_k}^{LS} > \gamma$; since $\gamma > 0$ we have $\hat{\beta_k}^{LS} > 0$. Next consider $\hat{\beta_k}^{lasso} < 0$. Then $\hat{\beta_k}^{LS} - \gamma sign(\hat{\beta_k}^{lasso}) = \hat{\beta_k}^{LS} + \gamma < 0$. $\Rightarrow \hat{\beta_k}^{LS} < -\gamma < 0$. Thus $sign(\hat{\beta_k}^{lasso}) = sign(\hat{\beta_k}^{LS})$ and so

$$\hat{\beta_k}^{lasso} = sign\left(\hat{\beta_k}^{LS}\right)\left(|\hat{\beta_k}^{LS}| - \gamma\right)$$

Note that $|\hat{\beta_k}^{LS}| \geq \gamma$; however, $\frac{1}{2}||\mathbf{y} - \mathbf{x}\beta||^2 + \sum_{j=1}^{p} |\beta_k|$ is not differentiable at $\beta_k = 0$; but when $\beta_k^{LS} = 0$ we have $\hat{\beta_k}^{lasso} = 0$ to minimize penalty. Thus we define

$$\hat{\beta_k}^{lasso} = sign\left(\hat{\beta_k}^{LS}\right)\left(|\hat{\beta_k}^{LS}| - \gamma\right)_+$$

2 Exercise 2

First, we preprocessed the data set. After noting that predictor variables that contained missing values contained a great proportion of missing values, we elected to throw away predictor variables with missing values. A simple count of missing values in R revealed that most columns with missing values had over 1000 of them. Next, backwise step selection was used with AIC to determine a subset of predictors to use. The mean squared error for a simple linear model was 0.1250 while the mean squared error after backwise step selection was 0.1227.

We next carry out a lasso regression procedure on the training data. To find the optimal choice of tuning parameter, we use 10-fold cross validation and attempt to minimize our mean square error over several choices of λ. (Please refer to Figure 1 in the appendix here). We find that our optimal choice of λ is 0.44. Thus we use $\lambda = 0.44$ in order to run the lasso procedure on the training set and obtain coefficients to fit the test set to. We obtain an error of 0.1224, which is slightly better than the error from the stepwise regression procedure.

Next, we carry out a ridge regression procedure using 10-fold cross validation to find the optimal choice of tuning parameter. We compute Mean Square Errors for $\lambda \in \{0, 0.1, 0.2, ..., 200\}$. Ridge regression suggests an optimal $\lambda = 51$ (see figure 2). With ridge regression, we obtain a MSE of 0.1227, which is essentially the same as the error from backwise step selection.

All three procedures of stepwise selection, lasso, and ridge regression give results with MSE all < 0.2. Even a simple linear model on all parameters gives a mean square error of 0.125. However, these methods of subset selection did give minor reductions in the MSE from a linear model regressed on all parameters. The lasso method gave us the best result, which may make sense for our data if regressing on the L-0 norm overfits and regressing on the L-2 norm underfits. One thing to note is that the violent crimes per population may depend on the interaction between the variables, rather than the variables themselves. All of these models ignore these potential interaction terms so they probably underestimate the effect of a change in one of these variables. We got decent MSEs with these methods, but subset selection and shrinkage methods do not tell us about the the interaction between variables.

3 Appendix

This section contains figures and R code.

Table 1: Mean Square Errors for Different Methods

Method	Test Error
Linear Regression (all parameters)	1.252%
Stepwise Regression	1.227%
Lasso	1.224%
Ridge Regression	1.227%

Figure 1: Cross Validation for Lasso

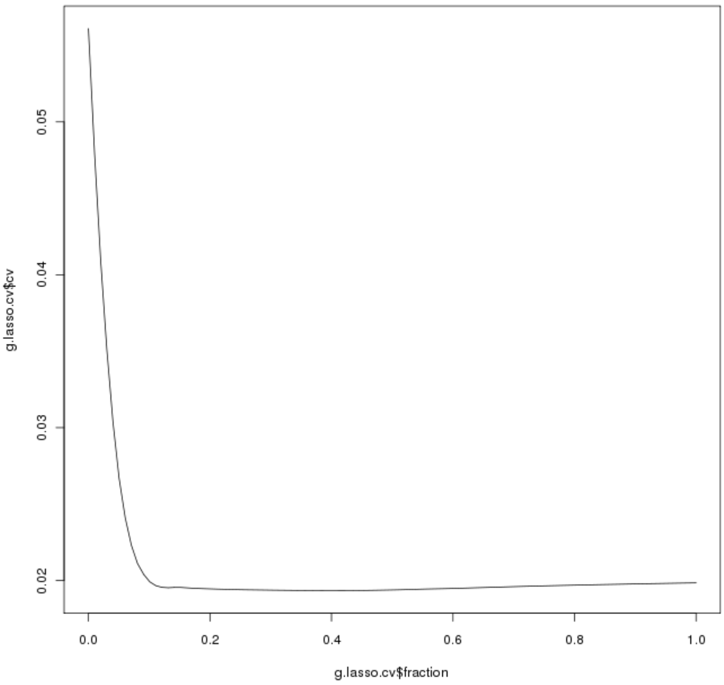
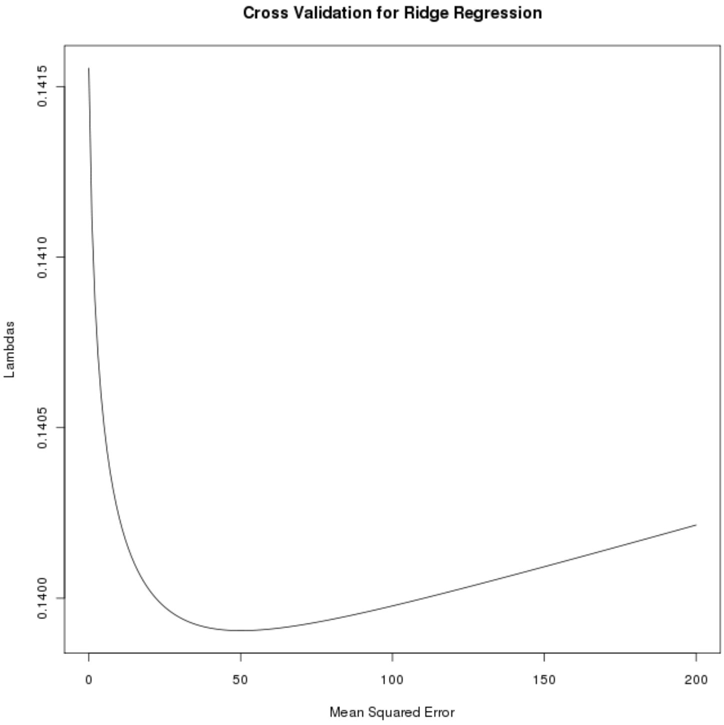


Figure 2: Cross Validation for Ridge Regression


Cross Validation for Ridge Regression
Lambdas
Mean Squared Error