Lecturer: Prof. Homrighausen Scribe: Yewon Kim

1 Recall LASSO

In short, $\mathcal{L}_1$ constraint is both convex and able to do model selection in the sense it forces some parameters to be zero. The estimator satisfies,

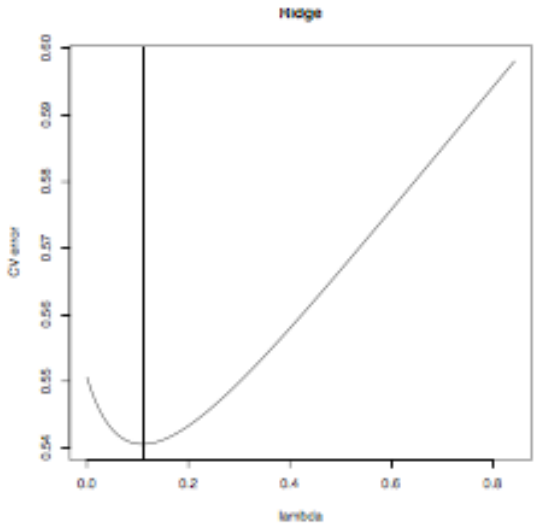
$\hat{\beta}_{lasso}(t) = argmin_{||\beta||_1 \leq t} ||Y - X\beta||_2^2$
In its corresponding Lagrangian dual form is $\hat{\beta}_{lasso}(\lambda) = argmin_\beta ||Y - X\beta||_2^2 + \lambda||\beta||_1$.

- We use glmnet uses gradient descent to quickly fit the lasso solution.
- The tuning parameter is often chosen by cross-validation.

2 Grids and Cross-validation

In this section's examples are in terms of Ridge regression because of simplicity.

Consider the below plot. There are many solutions have almost the same CV error and in fact, CV is a risk estimate which is random. The lower end point of the grid is somewhat arbitrary chosen.



The way that glmnet works is to

- form a grid of λ values,

- find the cross-validation error for each ridge solution on that grid
- compute the minimum cross-validation λ : $\hat{\lambda}$
- report $\hat{\beta}_{ridge}(\hat{\lambda})$ as the final solution

The important piece is that the final solution depends on which grid we choose. Although glmnet automatically allocates a grid, it is not necessary any good.

Example 2.1. *The grid does not allow small/ large enough λ values.*

3 Sparse matrices

All numbers in R take up the same space which means memory. So, if we can tell R in advance which entries are zero, it does not need to save that number. This can be accomplished in several ways in R, one is with the Matrix package.

```
library('Matrix')

Xspar = Matrix(X,sparse=T)
```

- Sparse matrices act like regular matrices (dense matrices)
- They just only keep track of which entries are non zero and perform the operation on these entries.
- For our purpose, glmnet (and other methods) automatically check to see if X is a sparse matrix object.
- This can be a substantial speed/storage saving for large, sparse matrices

4 SVD

The full SVD takes $O(min\{n^2p+p^3\})$ operations but often, we only need a few (q) singular values /vectors. For this we can use Krylov subspace techniques in $O(min\{npq\})$. The irlba(in R) function leverages the sparse matrix data structure.

The irlba function comes with additional choices :

- **adiust** : With irlba, we do not want to just compute q singular vectors if you need q, instead compute q+ adjust to enhance converge. (5 usually fine)
- **maxit** : irlba is iterative by nature. Check the output object iter to make sure the computation did not terminate based on iterations.

5 Elastic net

The ridge solution is always unique and does well when the covariates are highly related to each other. $\hat{\beta}_{rigde,\lambda} = argmin_\beta ||Y - X\beta||_2^2 + \lambda||\beta||_2^2 = (X^TX + \lambda I)^{-1}X^TY$.
The LASSO solution, $\hat{\beta}_{lasso,\lambda} = argmin_\beta ||Y - X\beta||_2^2 + \lambda||\beta||_1$ which is not necessarily unique, but it can do model selection. However, it can do poorly at model selection if the covariates are highly related to each other. The elastic net was introduced to combine both of these behaviors.

$\hat{\beta}_{\alpha,\lambda} = argmin_\beta\{||Y - X\beta||_2^2 + \lambda((1-\alpha)||\beta||_2^2 + \alpha||\beta||_1^1)\}$

We can do the elastic net in R with glmnet and the parameter α should be set. If there does not exist any convention for this, but CV can be used.

6 Refitted lasso

The lasso does both regularization and model selection, but it can produce a solution that produced too much bias. A common approach is to do the following two step:

- choose the λ via the one standard error rule
- refit the (unregularized) least squares solution on the selected covariates

The parameter values are estimates of the effect of that covariate however, the reported p-values are not valid.

one standard error rule : The one standard error rule is an alternative way of choosing θ from the CV curve. Let us start with the usual estimate $\hat{\theta} = argmin_{\theta\in\{\theta_1,\ldots\theta_m\}}CV(\theta)$ and we move θ in the direction of increasing regularization until it ceases to be true that $CV(\theta) \leq CV(\hat{\theta}) + SE(\hat{\theta})$. In other words, we take the simplest (the most regularized) model whose error is within one standard error of the minimal error.

7 Scaled-sparse regression

Theoretically, the optimal value for λ looks like : $\lambda = C\sigma$ for some constant C. Scaled sparse regression coefficients and noise level in a linear model. It alternates between, estimation σ via $\hat{\sigma} = \sqrt{\frac{||Y - X\hat{\beta}_{lasso}(\lambda)||_2^2}{n}}$.
Setting $\lambda = \hat{\sigma}\sqrt{\frac{n}{log(p)}}$