4.1: Penalized Regression

Dr. Bean - Stat 5100

1 Why Penalized Regression?

Recall linear regression model and predictive equation:

$$
\begin{aligned}
Y &= \beta_0 + \beta_1 X_1 + \ldots + \beta_{p-1} X_{p-1} + \varepsilon \\
\hat{Y} &= b_0 + b_1 X_1 + \ldots + b_{p-1} X_{p-1}
\end{aligned}
$$

IF the assumptions regarding residuals are satisfied, then ordinary least squares (OLS) provides the best (i.e. minimum variance) unbiased estimator for each β_k $(k = 1, \ldots, p-1)$ using the **loss function**

$$\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2.$$

However, when multicollinearity is present, the variance of the estimates for the β_k are inflated. What we would like is a way to shrink the variance of our estimated coefficients, perhaps forcing some coefficients all the way to zero (i.e. variable selection). This will allow us to **stabilize** our coefficient estimates while at the same time provide an alternative approach for variable selection.

However, nothing in statistics comes free. Like the "soul stone" from the avengers series, we must sacrifice something we love in order to obtain smaller variance and a new approach for variable selection.

Our Solution: Sacrifice **unbiased** estimates of the β coefficients in order to reduce their variance.

(Individual) What does it mean to be unbiased?

$$E(b_k) = \beta_k$$

In other words, if I were to use multiple *different* samples to fit my regression line, the estimated coefficients will all be different, but will all be centered around the true (and unknown) coefficients. This is important because it means that as my sample size increases, I expect to get estimates that are closer and closer to the "truth".

(Why might we be OK with giving up unbiasedness in order to minimize variance?

- Coefficients are biased to have smaller magnitude compared to the "truth" so we can still interpret the sign of each estimator.
- Biased, yet stable, estimates of the coefficients can often provide greater predictive accuracy than an OLS model.

2 Penalized Regression Approaches

Alternative Loss Functions:

- Ridge regression

$$\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2 \quad + \quad \lambda\sum_{k=0}^{p-1}(\beta_k)^2$$

- LASSO

$$\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2+\lambda\sum_{k=1}^{p-1}|\beta_k|$$

- Adaptive LASSO

$$\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2+\lambda\sum_{k=1}^{p-1}\frac{|\beta_k|}{\tilde{b}_k}$$

  - Where $\tilde{b}_k$ represents some initial estimate of the model coefficients (perhaps using OLS or traditional LASSO).

- Elastic Net

$$\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2+\lambda_1\sum_{k=0}^{p-1}(\beta_k)^2+\lambda_2\sum_{k=1}^{p-1}|\beta_k|$$

- Select values of λ that balances added bias with reduced variance.
- Our goal is impose the least amount of biasedness that we can in order to achieve an acceptable reduction in variance.
- One potential solution would be to select λ in such a way that minimizes the cross validation error.

Check out `https://ww2.amstat.org/meetings/csp/2014/onlineprogram/handouts/T3-Handouts.pdf` for additional info on these approaches.

Note that the explanatory variables MUST be standardized in order to use penalized regression techniques. Many functions perform this standardization automatically "under the hood."

2.1 Ridge Regression

Recall Linear Algebra Representation of OLS Regression:

$$Y = X\beta + \epsilon b \qquad\qquad = (X'X)^{-1}X'Yb \sim N(\beta, (X'X)^{-1}\sigma^2)$$

Recall also how we can standardize our X and Y variables producing:

$$\begin{aligned} Y^* &= X^*\beta^* + \varepsilon \\ b^* &= (X^{*\prime}X^*)^{-1}X^{*\prime}Y^* \\ &= (r_{XX})^{-1}r_{YX} \\ Cov(b^*) &= (r_{XX})^{-1}\sigma^2 \end{aligned} \qquad \begin{aligned} Y_i^* &= \frac{1}{\sqrt{n-1}} \cdot \frac{Y_i - \bar{Y}}{\text{SD of } Y} \\ X_{k,i}^* &= \frac{1}{\sqrt{n-1}} \cdot \frac{X_{k,i} - \bar{X}_k}{\text{SD of } X_k} \\ r_{XX} &= \text{correlation matrix of } X\text{'s} \\ r_{YX} &= \text{correlation vector between } Y \text{ and } X\text{'s} \end{aligned}$$

Ridge Regression introduces a small positive biasing constant $\lambda > 0$ so that

$$b^R = (r_{XX} + \lambda \cdot I)^{-1}r_{YX}$$

where I is the identity matrix (one's on the diagonal of the matrix and zeros elsewhere).

SAS Code:

```
proc reg data=<dataset> ridge=0 to <upper bound> by <step size>
     outvif outest=<named dataset of relevant ridge output>
     plots(only)=ridge(VIFaxis=log);
  model <model statement> / vif;
run;
```

Two graphical summaries to choose the "right" ridge parameter c:
(Note: these are guides; there is no "optimal" decision)

1. Ridge Trace Plot

   - (Need standardized data for this to be meaningful; SAS does internally)
   - Simultaneous plot of $b_1^R, \ldots, b_{p-1}^R$ (using standardized data) for different ridge parameters c (usually from 0 to 1 or 2)

   - As c increases from 0, the b_k^R may fluctuate wildly and even change signs
   - Eventually the b_k^R will move slowly toward 0

2. VIF Plot

   - Simultaneous plot of the variance inflation factor for the $p-1$ predictors for different ridge parameters

- As c increases from 0, the VIF drop toward 0

In general, choose smallest ridge parameter c:

1. where the b_k^R first become "stable" (their approach towards 0 has slowed)
2. and the VIF's have become "small enough" (close to 1 or less than 1)

2.1.1 Comments on Ridge Regression

- Choice of ridge parameter is somewhat subjective, but must be defendable (i.e. with a trace plot)
- given ridge parameter c, can get resulting parameter estimates b on the "unstandardized" (original data) scale
  - SAS gives these automatically, but need textbook equation 7.46b to get intercept b_0:

$$\beta_0 = \bar{Y} - \beta_1 \bar{X}_1 - \beta_2 \bar{X}_2 - \ldots - \beta_{p-1} \bar{X}_{p-1}$$

- ridge regression estimates b tend to be more robust against small changes to data than are OLS estimates
- predictors with very unstable ridge trace (tends toward zero without any plateau or slowing down) may be dropped from model, providing an alternative to stepwise variable selection techniques
- **major limitation**: traditional inference is not directly applicable to ridge regression estimates (part of our "soul stone" sacrifice)

2.2 LASSO (Least Absolute Shrinkage and Selection Operator)

Find b to minimize

$$\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i \right)^2 + \lambda \sum_{k=1}^{p-1} |\beta_k|$$

Switching from $\lambda \sum_{k=1}^{p-1} \beta_k^2$ in ridge regression to $\lambda \sum_{k=1}^{p-1} |\beta_k|$ in LASSO, may seem minor, but this change causes b_k values to now shrink all the way to zero.

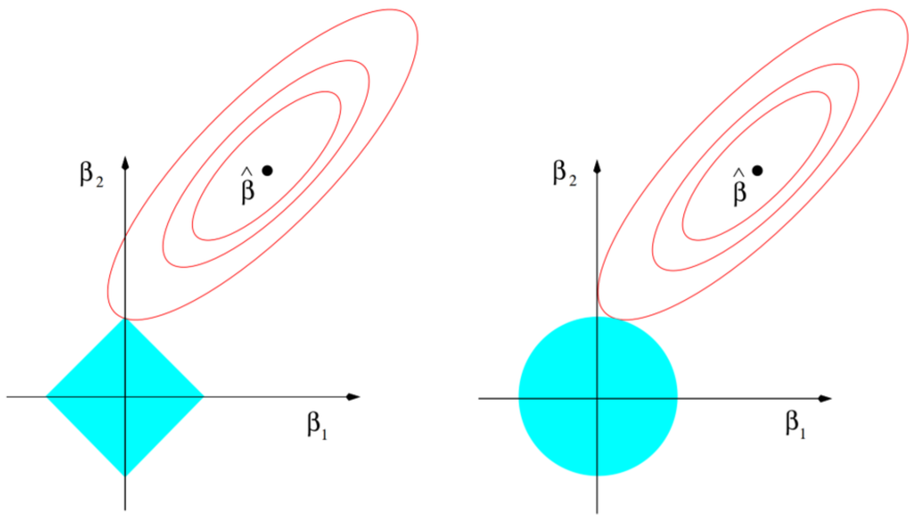

FIGURE 3.11. *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions $|\beta_1| + |\beta_2| \leq t$ and $\beta_1^2 + \beta_2^2 \leq t^2$, respectively, while the red ellipses are the contours of the least squares error function.*

Options exist for choosing λ We can use these because we now have models with different numbers of coefficients, not the case in ridge regression.

- likelihood function-based criteria (Adj. R^2, C_p, AIC, SBC, etc.)
- cross-validation
  - withhold some of the data, fit on the rest, then predict on withheld portion
  - select λ to minimize something like (others exist)

$$PRESS = \sum_{i=1}^{n} \left(\frac{Y_i - \hat{Y}_i}{1 - h_{ii}} \right)^2$$

SAS Code

```
proc glmselect data=<dataset> plots=(criterion <measure>);
 class <all qualitative variables>;
 model <your model>
        / selection=lasso(adaptive choose=<selection method> stop=none);
 output out=<output dataset> p=<lasso predictions>;
run;
```

One way to visualize progress of model is to show ASE as each variable is added

$$\text{ASE} = \frac{\text{SSE}}{n} \qquad MSE = \frac{\text{SSE}}{n-p}$$

.

2.3 Adaptive LASSO

- Problem: LASSO is known to give more biased estimates of nonzero coefficients
- Solution: Allow higher penalty for zero coefficients and lower penalty for nonzero coefficients

Find b to minimize

$$\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 + \lambda \sum_{k=1}^{p-1} \frac{|\beta_k|}{b_k}$$

"Adaptive" weights: $\frac{1}{b_k}$, where b_k is obtained from an initial model fit (using OLS or regular LASSO or something else)
– control shrinking of zero coefficients more than nonzero coefficients

2.4 Elastic Net

> *Similar to the lasso, the elastic net simultaneously does automatic variable selection and continuous shrinkage, and it can select groups of correlated variables. It is like a stretchable fishing net that retains 'all the big fish'"* - Zou and Hastie (2005)

Some limitations of LASSO:

- When number of predictors $(p-1)$ exceeds sample size (n), LASSO will select up to n predictor variables before it saturates.

- In the presence of high multicollinearity, LASSO tends to select only one variable from the group of correlated predictors.

- When sample size (n) exceeds number of predictors $(p-1)$ <u>and</u> there is high multicollinearity, LASSO is out-performed (prediction-wise) by ridge regression.

Elastic Net overcomes these limitations:

- can select more than n variables
- can select more than one variable from a group of highly collinear predictors
- can achieve better predictive performance

Find b to minimize

$$\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 + \lambda_1 \sum_{k=0}^{p-1} (\beta_k)^2 + \lambda_2 \sum_{k=1}^{p-1} |\beta_k|$$