Advanced Methods in Biostatistics II

Lecture 7

November 14, 2017

Linear model

- Consider the linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I})$.

- Today we will begin discussing regularization methods.
- These include ridge regression and the lasso.

Regularization methods

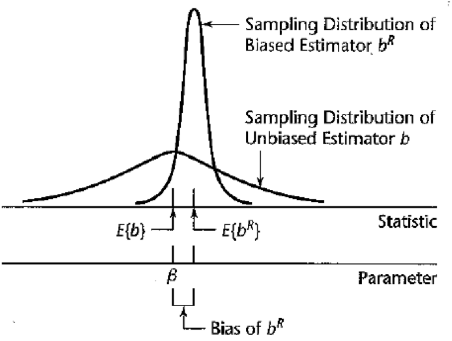
- Regularization imposes an upper threshold on the value coefficients can take, potentially providing more parsimonious solutions.
- Regularization methods are particularly useful when variables are highly correlated with one another.
- But they also have utility as a variable selection tool.

Least-squares solution

- So far we have been using the ordinary least-squares estimate:
$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$
- We have previously shown this to be the BLUE.
- However, is it still possible to improve upon it?

Bias-variance tradeoff

- If an estimate has only a small bias but is substantially more precise than an unbiased estimate it may be preferable.



Bias-variance tradeoff

- The quality of an estimator can be quantified using the mean square error:

$$MSE(\hat{\beta}) = E[(\hat{\beta} - \beta)'(\hat{\beta} - \beta)].$$

- The MSE can be written as the sum of the variance of the estimator and its squared bias, i.e.,

$$MSE(\hat{\beta}) = tr(Var(\hat{\beta})) + Bias(\hat{\beta}, \beta)^2.$$

- If $\hat{\beta}$ is unbiased, then $MSE(\hat{\beta}) = tr(Var(\hat{\beta}))$.

Penalized least squares

- Consider adding a constraint to the least squares equation:

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

subject to

$$\sum_{j=1}^{p} \beta_j^2 \leq s.$$

- This 'shrinks' the values of the coefficients by placing a constraint on their size.
- This is referred to as an L_2-penalty.

Ridge regression

- Using a Lagrange multiplier this can alternatively be expressed as follows:

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

- The term $\lambda \geq 0$ is a complexity parameter that controls the amount of shrinkage.
- There is a on-to-one correspondence between λ and s.

Ridge regression

- The addition of the penalty term is called “Tikhonov regularization” for the mathematician of that name.
- Since both the objective function and constraints are convex functions, this is a convex optimization problem.

Ridge regression

- The specific use of this regularization in the regression setting is called ridge regression.
- The ridge regression coefficients will generally be smaller in absolute magnitude than the standard OLS estimators.
- They are therefore often called shrinkage estimators.

Ridge regression

- The ridge solutions are not equivariant under re-scaling of the explanatory variables.
- Therefore, we typically study the problem after first mean-centering the data.
- Let $\mathbf{X}$ now represent a mean centered design matrix without an intercept term.

Ridge regression

- In matrix format we can write the penalized least squares criteria for ridge regression as follows:

$$f(\beta) \quad = \quad (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) + \lambda\beta'\beta$$

- To minimize $f(\beta)$, we begin by taking the derivative with respect to β and setting it equal to zero:

$$\frac{df}{d\beta} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\beta + 2\lambda\beta = 0.$$

Ridge regression

- This expression can be simplified as follows:

$$(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})\beta = \mathbf{X}'\mathbf{y}.$$

- The estimate is given by:

$$\hat{\beta}_{ridge} = (\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}'\mathbf{y}.$$

- Note that since we are adding a positive constant to the diagonal of $\mathbf{X}'\mathbf{X}$, the matrix $\mathbf{X}'\mathbf{X} + \lambda\mathbf{I}$ will be invertible even if $\mathbf{X}'\mathbf{X}$ is singular.

Ridge regression

- The ridge regression estimator is related to the standard OLS estimate as follows:

$$\hat{\beta}_{ridge} = (\mathbf{I} - \lambda(\mathbf{X}'\mathbf{X})^{-1})^{-1}\hat{\beta}$$

  assuming $\mathbf{X}'\mathbf{X}$ is non-singular.

- In addition, if $\mathbf{X}$ is orthogonal (i.e., $\mathbf{X}'\mathbf{X} = \mathbf{I}$), we have that:

$$\begin{aligned} \hat{\beta}_{ridge} &= (\mathbf{I} - \lambda\mathbf{I})^{-1}\hat{\beta} \\ &= \frac{1}{1-\lambda}\hat{\beta}. \end{aligned}$$

- This illustrates the shrinkage property.

Ridge regression

- The difficulty in performing ridge regression comes in choosing an appropriate coefficient λ for the penalty term.
- If we knew MSE as a function of λ then we could choose the value that minimizes MSE.
- A popular method for estimating MSE is cross-validation.

Ridge regression

- Frequently, a ridge trace is used to determine λ.
- This is a simultaneous plot of the estimated regression coefficients (which are functions of λ) against λ.
- The value of λ is chosen so that the regression coefficients change little for any larger values of λ.

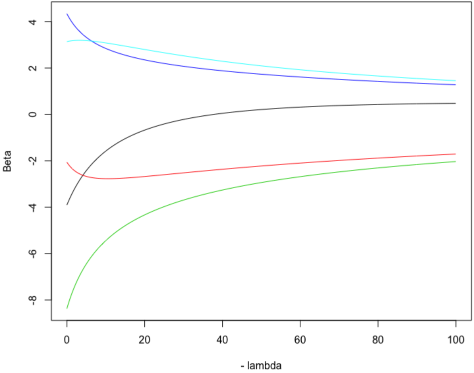
Coding example

- We use the `swiss` fertility data set to illustrate fitting ridge regression.
- In this example, penalization isn't really necessary, so the code is intended to simply illustrate the approach.

Coding example

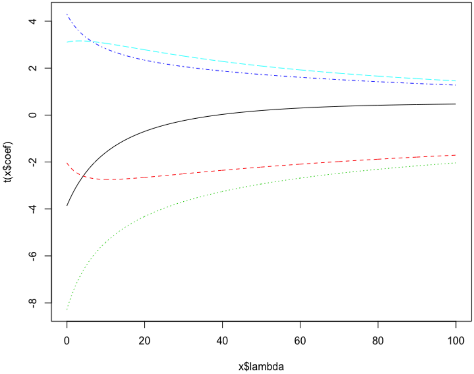
```
data(swiss)
y = swiss[,1]
x = swiss[,-1]
y = y - mean(y)
x = apply(x, 2, function(z) (z - mean(z)) / sd(z))
n = length(y); p = ncol(x)
lambdaSeq = seq(0, 100, by = .1)
betaSeq = sapply(lambdaSeq, function(l) solve(t(x) %*% x + l * diag(rep(1, p)), t(x) %*% y))
plot(range(lambdaSeq), range(betaSeq), type = "n", xlab = "- lambda", ylab = "Beta")
for (i in 1 : p) lines(lambdaSeq, betaSeq[i,],col=i)
```

Coding example



Coding example

```
library(MASS)
fit = lm.ridge(y ~ x, lambda = lambdaSeq)
plot(fit)
```



Ridge regression

- Assume $\mathbf{X}$ is mean-centered and of full column rank.
- Let us re-express the design matrix using its singular value decomposition (SVD):

$$\mathbf{X} = \mathbf{UDV}.$$

- We will use the result to study some of the properties of the hat matrix used in ridge regression.

Ridge regression

- First, consider the fitted values:

$$\begin{aligned}\hat{\mathbf{y}}_{ridge} &= \mathbf{X}(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}'\mathbf{y} \\ &= \mathbf{H}_\lambda \mathbf{y}.\end{aligned}$$

- Next, observe that:

$$\mathbf{X}'\mathbf{X} = \mathbf{V}\mathbf{D}^2\mathbf{V}'.$$

Ridge regression

- Since $\mathbf{X}$ is of full column rank, it follows that $\mathbf{V}$ is a $p \times p$ matrix of full rank with $\mathbf{V}'\mathbf{V} = \mathbf{V}\mathbf{V}' = \mathbf{I}$ and $\mathbf{V}^{-1} = \mathbf{V}'$.

- Thus, we can write

$$
\begin{aligned}
\mathbf{H}_\lambda &= \mathbf{X}(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}' \\
&= \mathbf{U}\mathbf{D}\mathbf{V}'(\mathbf{V}\mathbf{D}^2\mathbf{V}' + \lambda\mathbf{I})^{-1}\mathbf{V}\mathbf{D}\mathbf{U}' \\
&= \mathbf{U}\mathbf{D}\mathbf{V}'(\mathbf{V}\mathbf{D}^2\mathbf{V}' + \lambda\mathbf{V}\mathbf{V}')^{-1}\mathbf{V}\mathbf{D}\mathbf{U}' \\
&= \mathbf{U}\mathbf{D}\mathbf{V}'\{\mathbf{V}(\mathbf{D}^2 + \lambda\mathbf{I})\mathbf{V}'\}^{-1}\mathbf{V}\mathbf{D}\mathbf{U}' \\
&= \mathbf{U}\mathbf{D}\mathbf{V}'\mathbf{V}(\mathbf{D}^2 + \lambda\mathbf{I})^{-1}\mathbf{V}'\mathbf{V}\mathbf{D}\mathbf{U}' \\
&= \mathbf{U}\mathbf{D}(\mathbf{D}^2 + \lambda\mathbf{I})^{-1}\mathbf{D}\mathbf{U}' \\
&= \mathbf{U}\mathbf{W}\mathbf{U}'.
\end{aligned}
$$

Ridge regression

- Here $\mathbf{W}$ is a diagonal matrix whose elements are:

$$\frac{d_i^2}{d_i^2 + \lambda}$$

  where d_i^2 are the diagonal elements of $\mathbf{D}$ (i.e., the eigenvalues).

- Note that the smaller the value of d_i, the more the i^{th} coefficient is shrunk towards 0.

Ridge regression

- The trace of the hat matrix can be used to compute the degrees of freedom associated with the linear model.
- Here, we see that:

$$tr(\mathbf{H}_\lambda) = \sum_{j=1}^{p} \frac{d_i^2}{d_i^2 + \lambda}.$$

- This is referred to as the effective degrees of freedom.

Ridge regression

- If $\lambda \to 0$, then $tr(\mathbf{H}_\lambda) = p$, which is the standard OLS result.
- However, if $\lambda \to \infty$, then $tr(\mathbf{H}_\lambda) \to 0$.
- Thus, regularization reduces the parameters effective degrees of freedom.

Bias and Variance

- Let us study the bias and variance properties of the ridge estimator.

$$\begin{aligned} E(\hat{\beta}_{ridge}) &= (\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}'\mathbf{y} \\ &= (\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \\ &= (\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I} - \lambda\mathbf{I})\boldsymbol{\beta} \\ &= (\mathbf{I} - \lambda(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1})\boldsymbol{\beta} \\ &= \boldsymbol{\beta} - \lambda(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}\boldsymbol{\beta} \end{aligned}$$

- The bias increases as a function of λ.

Bias and Variance

- The variance-covariance matrix can be expressed as follows:

$$Var(\hat{\beta}_{ridge}) = \sigma^2(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}(\mathbf{X}'\mathbf{X})(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}$$

which can be simplified as follows:

$$Var(\hat{\beta}_{ridge}) = \sigma^2\mathbf{V}(\mathbf{D}^2 + \lambda\mathbf{I})^{-1}\mathbf{D}^2(\mathbf{D}^2 + \lambda\mathbf{I})^{-1}\mathbf{V}'.$$

Bias and Variance

- The total variability is represented by the trace of the variance-covariance matrix.
- Here we can write:

$$tr(Var(\hat{\beta}_{ridge})) = \sigma^2 \sum_{j=1}^{p} \frac{d_j^2}{(d_j^2 + \lambda)^2}.$$

Bias and Variance

- Compare this to the OLS solution:

$$tr(Var(\hat{\beta})) = \sigma^2 \sum_{j=1}^{p} \frac{1}{d_j^2}.$$

- Thus, one can show that the ridge estimator has systematically smaler total variation, i.e.

$$tr(Var(\hat{\beta}_{ridge})) \leq tr(Var(\hat{\beta})).$$

Multicollinearity

- Collinearity or multicollinearity refers to the problem when the columns of the design matrix are nearly linear dependent.
- If the columns of $\mathbf{X}$ are linearly dependent, there exists an infinite number of least squares estimates for the true regression coefficients.
- If $\mathbf{X}$ is nearly linearly dependent, the estimated regression coefficients may not be meaningful and may be highly variable.

Multicollinearity

- Ridge regression was originally proposed as a method to deal with issues related to multicollinearity.
- Now it is more commonly viewed as a form of penalized likelihood estimation:

$$(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) + \lambda p(\beta)$$

where $p(\beta)$ is a nonnegative penalty function.

Bayesian interpretation

- Another way to envision ridge regression is to think of it in the terms of a posterior mode on a regression model.
- Specifically consider the model where $\mathbf{y} \mid \beta \sim N(\mathbf{X}\beta, \sigma^2\mathbf{I})$ and $\beta \sim N(\mathbf{0}, \tau^2\mathbf{I})$.
- Then one obtains the posterior for β and σ by multiplying the two densities.

Bayesian interpretation

- Thus, one can show that the posterior mode can be obtained by minimizing:

$$||\mathbf{y} - \mathbf{X}\beta||^2/\sigma^2 + \beta'\beta\sigma^2/\tau^2.$$

- This is equivalent to ridge regression in terms of maximization for β, with $\lambda = \sigma^2/\tau^2$.