Chapter 1. Linear regression Model part 3

January 18, 2007

1 Ride regression

Note that the basic requirement for the Least squares (LS) estimation of a linear regression is

$$(\mathbf{X}^\top\mathbf{X})^{-1}\text{exists.}$$

There are two reasons that the inverse does not exits. (1) $p > n$ and (2) collinearity. The "badly conditioned linear regression problems" (Hoerl and Kennard, 1970) has long been an important problem in statistics and computer science. The problem is even intrict in high dimensional data as most of the genetics data are. The technique of ridge regression (RR) is one of the most popular and best performing (Frank and Friedman, 1993) alternatives to the ordinary least squares (LS) methods.

A simple way to guarantee the invertibility is adding a diagonal matrix to $\mathbf{X}^\top\mathbf{X}$, i.e. $\mathbf{X}^\top\mathbf{X} + \lambda I$, where I is a $(p+1) \times (p+1)$ identity matrix. The ridge regression estimator is then

$$\hat{\beta}_r = (\mathbf{X}^\top\mathbf{X} + \lambda I)^{-1}\mathbf{X}^\top\mathbf{Y}$$

where $\lambda > 0$ is a parameter needs to be chosen (HOW? any idea). To make the notation clear, denote the LS estimator by

$$\hat{\beta}_{LS} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$$

(providing it exists)

Mean of $\hat{\beta}_r$

$$E\hat{\beta}_r = E\{(\mathbf{X}^\top\mathbf{X} + \lambda I)^{-1}\mathbf{X}^\top(\mathbf{X}\beta + \mathcal{E})\} = (\mathbf{X}^\top\mathbf{X} + \lambda I)^{-1}\mathbf{X}^\top\mathbf{X}\beta = \beta + \lambda(\mathbf{X}^\top\mathbf{X} + \lambda I)^{-1}\beta$$

It is not unbiased. The bias is

$$bias(\hat{\beta}_r) = E\hat{\beta}_r - \beta = \lambda(\mathbf{X}^\top\mathbf{X} + \lambda I)^{-1}\beta$$

Recall that

$$E\hat{\beta}_{LS} = \beta.$$

which is unbiased

Variance-covariance matrix of $\hat{\beta}_r$: If $var(\mathcal{E}) = \sigma^2 I_n$, then

$$var(\hat{\beta}_r) = (\mathbf{X}^\top\mathbf{X} + \lambda I)^{-1}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X} + \lambda I)^{-1}\sigma^2$$

Recall that

$$var(\hat{\beta}_{LS}) = (\mathbf{X}^\top\mathbf{X})^{-1}\sigma^2$$

Mean squared error of $\hat{\beta}_r$: $E||\hat{\beta}_r - \beta||^2$

$$\begin{aligned}
MSE &= E||\hat{\beta}_r - \beta||^2 = E(\hat{\beta}_r - \beta)^\top(\hat{\beta}_r - \beta) \\
&= E(\hat{\beta}_r - E\hat{\beta}_r + E\hat{\beta}_r - \beta)^\top(\hat{\beta}_r - E\hat{\beta}_r + E\hat{\beta}_r - \beta) \\
&= E||\hat{\beta}_r - E\hat{\beta}_r||^2 + 2E\{(E\hat{\beta}_r - \beta)^\top(\hat{\beta}_r - E\hat{\beta}_r)\} + ||E\hat{\beta}_r - \beta||^2 \\
&= E||\hat{\beta}_r - E\hat{\beta}_r||^2 + ||E\hat{\beta}_r - \beta||^2 \\
&= tr(Variance) + ||bias||^2
\end{aligned}$$

The LS has no bias but with a bigger variance than the ridge regression estimator. People proved that we can always find a λ such that

$$MSE(\hat{\beta}_r) < MSE(\hat{\beta}_{LS}).$$

In other words, ridge regression can improve the estiamtion of β.

2 An alternative way of understanding ridge regression

The motivation of ridge regression is very simple, but it has good performance. Another way to understand it is that we dont expect an estimator with too large β. Thus, we penalize the value of β. Recall the LS estimation is to minimize

$$\sum_{i=1}^{n}(Y_i - X_i\beta)^2$$

or

$$\min_{\beta} \sum_{i=1}^{n} (Y_i - X_i \beta)^2$$

To penalize the value of β, we can consider estimate β by minimizing

$$\sum_{i=1}^{n} (Y_i - X_i \beta)^2 + \lambda ||\beta||^2.$$

or

$$\min_{\beta} \{\sum_{i=1}^{n} (Y_i - X_i \beta)^2 + \lambda ||\beta||^2\}.$$

It is not difficult to prove that to solution of β to the above problem is

$$\hat{\beta}_r = (\mathbf{X}^\top \mathbf{X} + \lambda I)^{-1} \mathbf{X}^\top Y.$$

Note that with larger λ, the penalty on β tends to be stronger; the solution of β will be smaller.

3 Another expression of the LS and ridge estimator

$$\hat{\beta}_{LS} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top Y = (\sum_{i=1}^{n} X_i^\top X_i)^{-1} \sum_{i=1}^{n} X_i^\top Y_i$$

Therefore, if observation (X_j, Y_j) is removed, then the delete-one-out LS estimator is

$$\hat{\beta}_{LS}^j = (\sum_{i=1, i \neq j}^{n} X_i^\top X_i)^{-1} \sum_{i=1, i \neq j}^{n} X_i^\top Y_i.$$

Similarly, the delete-one-out (delete j'th observation) ridge estimator is

$$\hat{\beta}_r^j = (\sum_{i=1, i \neq j}^{n} X_i^\top X_i + \lambda I)^{-1} \sum_{i=1}^{n} X_i^\top Y_i.$$

4 Selection of λ in ridge regression

Suppose we have observations, $(X_1, Y_1), ..., (X_n, Y_n)$ and consider linear regression model

$$Y_i = X_i \beta + \varepsilon_i, \quad i = 1, 2, ..., n.$$

For different λ, we have different ridge regression estimator for the model.

$$\hat{\beta}_r(\lambda) = (\mathbf{X}^\top \mathbf{X} + \lambda I)^{-1} \mathbf{X}^\top \mathbf{Y}.$$

4.1 selection of λ via CV

We select a large range for possible λ: $[0, c]$. For each fixed λ in $[0, c]$, consider the CV as follows. For each j,

$$\hat{\beta}_r^j(\lambda) = (\sum_{i \neq j} X_i^\top X_i + \lambda I)^{-1} \sum_{i \neq j} X_i^\top Y_i.$$

The prediction error for (X_j, Y_j) is

$$err^j(\lambda) = (Y_j - X_j \hat{\beta}_r^j(\lambda))^2$$

The CV value is then

$$CV(\lambda) = n^{-1} \sum_{j=1}^n err^j(\lambda)$$

The best λ is the minimum point of $CV(\lambda)$.

Example 4.1 (Near Infra-red Calibration for Protein, Fearn (1983) (dataA)) . *In the data, Y is protein percentage with 6 explanatory variables $X = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_6)$, which are log(1/reflectance) values at six wavelengths.*

The LS estimated model is

$$y = 29.372 - 0.1692\mathbf{x}_1 - 0.1536\mathbf{x}_2 + 0.5333\mathbf{x}_3 - 0.1362\mathbf{x}_4 - 0.008\mathbf{x}_5 - 0.0615\mathbf{x}_6$$

Using CV, the selected λ is 4.4. The estimated model is

$$y = 1.8843 + 0.0515\mathbf{x}_1 - 0.22783\mathbf{x}_2 + 0.4726\mathbf{x}_3 - 0.28769\mathbf{x}_4 + 0.0058\mathbf{x}_5 - 0.0154\mathbf{x}_6$$

To check the models, a new experiment was done and the data were collected **(dataB)**

The prediction errors for the new data set are respectively: Least square estimation: 0.09397779; Ridge regression: 0.07783629.

R code for the calculation **(code)**

Example 4.2 (cell classification based on gene) *For the leukemia gene expression data (***(training data)***. There are 38 cells with 250 genes (selected from about 7000 genes). they are from two types of cells.*

*To check the models, a new experiment was done and the data were collected (***(testing set)***.*

The prediction errors for the new data set based on Ridge regression: 3.676901e-08. (with ridge parameter $\lambda = 0.05$) From figure 1, we can see that we can have a very accurate classification for the new data.

R code for the calculation **(code)**

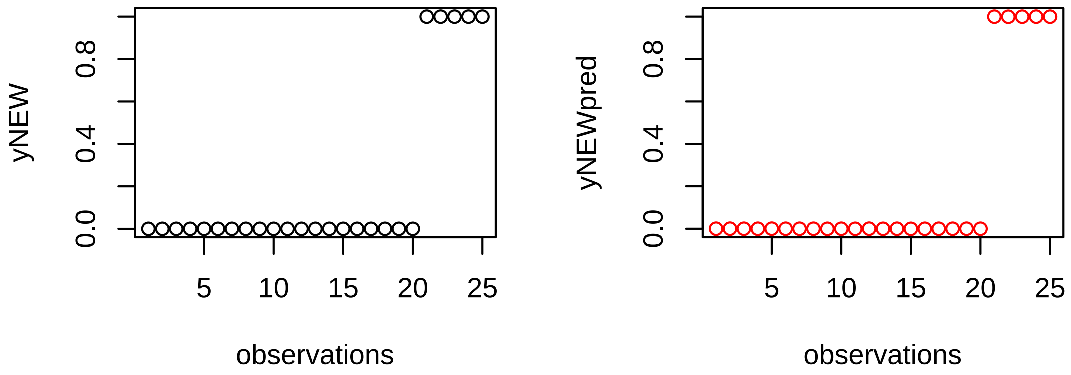

Figure 1: The left panel is the true classification of the 13 cells; the right panel is predicted clasifiction

4.2 selection of λ via GCV

with λ the ridge regression will give a fit to the observations as

$$\hat{Y} = \mathbf{X}\hat{\beta}_r = \mathbf{X}(\mathbf{X}^\top \mathbf{X} + \lambda I)^{-1}\mathbf{X}^\top Y = SY,$$

where $S = \mathbf{X}(\mathbf{X}^\top \mathbf{X} + \lambda I)^{-1}\mathbf{X}^\top$. The GCV is then

$$GCV(\lambda) = n^{-1}(\mathbf{Y} - \hat{\mathbf{Y}})^\top(\mathbf{Y} - \hat{\mathbf{Y}})/(1 - tr(S)/n)^2.$$

The best λ is the minimum point of $GCV(\lambda)$.

Example 4.3 (The same data sets above in Example 4.1.) *Using GCV, the selected λ is 3.3e-05. The estimated model is*

$$y = 28.0220 - 0.1584\mathbf{x}_1 - 0.1573\mathbf{x}_2 + 0.5304\mathbf{x}_3 - 0.1437\mathbf{x}_4 - 0.00755\mathbf{x}_5 - 0.0592\mathbf{x}_6$$

The prediction errors for the new data set are respectively: Least square estimation: 0.09397779; Ridge regression: 0.09073758

R code for the calculation **(code)**

Example 4.4 (The same data sets above in Example 4.2.) *The prediction errors for the new data set based on Ridge regression: 7.471459e-05. (with ridge parameter $\lambda = 2.45$)*

R code for the calculation **(code)**

4.3 Other selection of λ

Suppose we can obtain the least square estimator $\hat{\beta}$ and estimator of $\hat{\sigma}^2$. then

$$\lambda = \frac{(p+1)\hat{\sigma}^2}{||\hat{\beta}||^2}$$

If we cannot get the least square estimator, we can use a ridge regression with very small λ. And get a similar value of λ.

5 Extension of ridge regression

The bridge regression proposed by Frank and Friedman (1993) can be written as

$$\min_{\beta} \Big\{ n^{-1}(Y - X\beta)^\top (Y - X\beta) + \lambda \sum_{k=1}^{p} |\beta_k|^\gamma \Big\}, \tag{5.1}$$

where $\gamma > 0$. If $\gamma = 2$, it is the ridge regression; if $\gamma = 1$, it is an equivalent of the Lasso proposed by Tibishrani (1996).

6 Lasso: Least absolute shrinkage and selection operator

If we estimate β by

$$\min_{\beta} \Big\{ n^{-1}(Y - X\beta)^\top (Y - X\beta) + \lambda \sum_{k=1}^{p} |\beta_k| \Big\}, \tag{6.2}$$

the estimation procedure is called Lasso. Lasso simultaneously accomplish model estimation and variable selection.