LINEAR REGRESSION ANALYSIS

MODULE – IX

Lecture - 31

Multicollinearity

Dr. Shalabh

Department of Mathematics and Statistics

Indian Institute of Technology Kanpur

6. Ridge regression

The OLSE is the best linear unbiased estimator of regression coefficient in the sense that it has minimum variance in the class of linear and unbiased estimators. However if the condition of unbiasedness can be relaxed then it is possible to find a biased estimator of regression coefficient say $\hat{\beta}$ that has smaller variance them the unbiased OLSE *b*. The mean squared error (*MSE*) of $\hat{\beta}$ is

$$\begin{aligned} MSE(\hat{\beta}) &= E(\hat{\beta}-\beta)^2 \\ &= E\left[\left\{\hat{\beta}-E(\hat{\beta})\right\}+\left\{E(\hat{\beta})-\beta\right\}\right]^2 \\ &= Var(\hat{\beta})+\left[E(\hat{\beta})-\beta\right]^2 \\ &= Var(\hat{\beta})+\left[Bias(\hat{\beta})\right]^2. \end{aligned}$$

.

Thus $MSE(\hat{\beta})$ can be made smaller than $Var(\hat{\beta})$ by introducing small bias in $\hat{\beta}$. One of the approach to do so is the ridge regression. The ridge regression estimator is obtained by solving the normal equations of least squares estimation. The normal equations are modified as

$$\begin{aligned} &\left(X'X+\delta I\right)\hat{\beta}_{ridge} = X'y \\ \Rightarrow\ &\hat{\beta}_{ridge} = \left(X'X+\delta I\right)^{-1}X'y. \end{aligned}$$

$\hat{\beta}_{ridge}$ is the **ridge regression estimator** of β and $\delta \geq 0$ is any characterizing scalar termed as **biasing parameter**.

As

$$\delta \to 0,\ \hat{\beta}_{ridge} \to b(OLSE) \text{ and as } \delta \to \infty,\ \hat{\beta}_{ridge} \to 0.$$

So larger the value of δ , larger shrinkage towards zero. Note that the OLSE is inappropriate to use in the sense that it has very high variance when multicollinearity is present in the data. On the other hand, a very small value of $\hat{\beta}$ may tend to accept the null hypothesis $H_0 : \beta = 0$ indicating that the corresponding variables are not relevant. The value of biasing parameter controls the amount of shrinkage in the estimates.

Bias of ridge regression estimator

The bias of $\hat{\beta}_{ridge}$ is

$$\begin{aligned} Bias(\hat{\beta}_{ridge}) &= E(\hat{\beta}_{ridge}) - \beta \\ &= (X'X + \delta I)^{-1} X'E(y) - \beta \\ &= \left[(X'X + \delta I)^{-1} X'X - I \right] \beta \\ &= (X'X + \delta I)^{-1} \left[X'X - X'X - \delta I \right] \beta \\ &= -\delta (X'X + \delta I)^{-1} \beta. \end{aligned}$$

Thus the ridge regression estimator is a biased estimator of β .

Covariance matrix

The covariance matrix of $\hat{\beta}_{ridge}$ is defined as

$$V(\hat{\beta}_{ridge}) = E\left[\left\{\hat{\beta}_{ridge} - E(\hat{\beta}_{ridge})\right\}\left\{\hat{\beta}_{ridge} - E(\hat{\beta}_{ridge})\right\}'\right].$$

Since

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

so

$$\begin{aligned}V(\hat{\beta}_{ridge}) &= (X'X + \delta I)^{-1}X'V(\varepsilon)X(X'X + \delta I)^{-1} \\ &= \sigma^2(X'X + \delta I)^{-1}X'X(X'X + \delta I)^{-1}.\end{aligned}$$

Mean squared error

Writing $X'X = \Lambda = diag(\lambda_1, \lambda_2, ..., \lambda_k)$, the mean squared error of $\hat{\beta}_{ridge}$ is

$$MSE(\hat{\beta}_{ridge}) = Var(\hat{\beta}_{ridge}) + \left[bias(\hat{\beta}_{ridge})\right]^2$$

$$= tr\left[V(\hat{\beta}_{ridge})\right] + \left[bias(\hat{\beta}_{ridge})\right]^2$$

$$= \sigma^2 tr\left[(X'X + \delta I)^{-1} X'X (X'X + \delta I)^{-1}\right] + \delta^2 \beta' (X'X + \delta I)^{-2} \beta$$

$$= \sigma^2 \sum_{j=1}^{k} \frac{\lambda_j}{(\lambda_j + \delta)^2} + \delta^2 \sum_{j=1}^{k} \frac{\beta_j^2}{(\lambda_j + \delta)^2}$$

where $\lambda_1, \lambda_2, ..., \lambda_k$ are the eigenvalues of $X'X$.

Thus as δ increases, the bias in $\hat{\beta}_{ridge}$ increases but its variance decreases. The trade off between bias and variance hinges upon the value of δ. It can be shown that there exists a value of δ such that $MSE(\hat{\beta}_{ridge}) < Var(b)$ provided $\beta'\beta$ is bounded.

Choice of δ

The estimation of ridge regression estimator depends upon the value of δ. Various approaches have been suggested in the literature to determine the value of δ. The value of δ can be chosen on the basis of criteria like

- stability of estimators with respect to δ .
- reasonable signs.
- magnitude of residual sum of squares etc.

We consider here the determination of δ by the inspection of ridge trace.

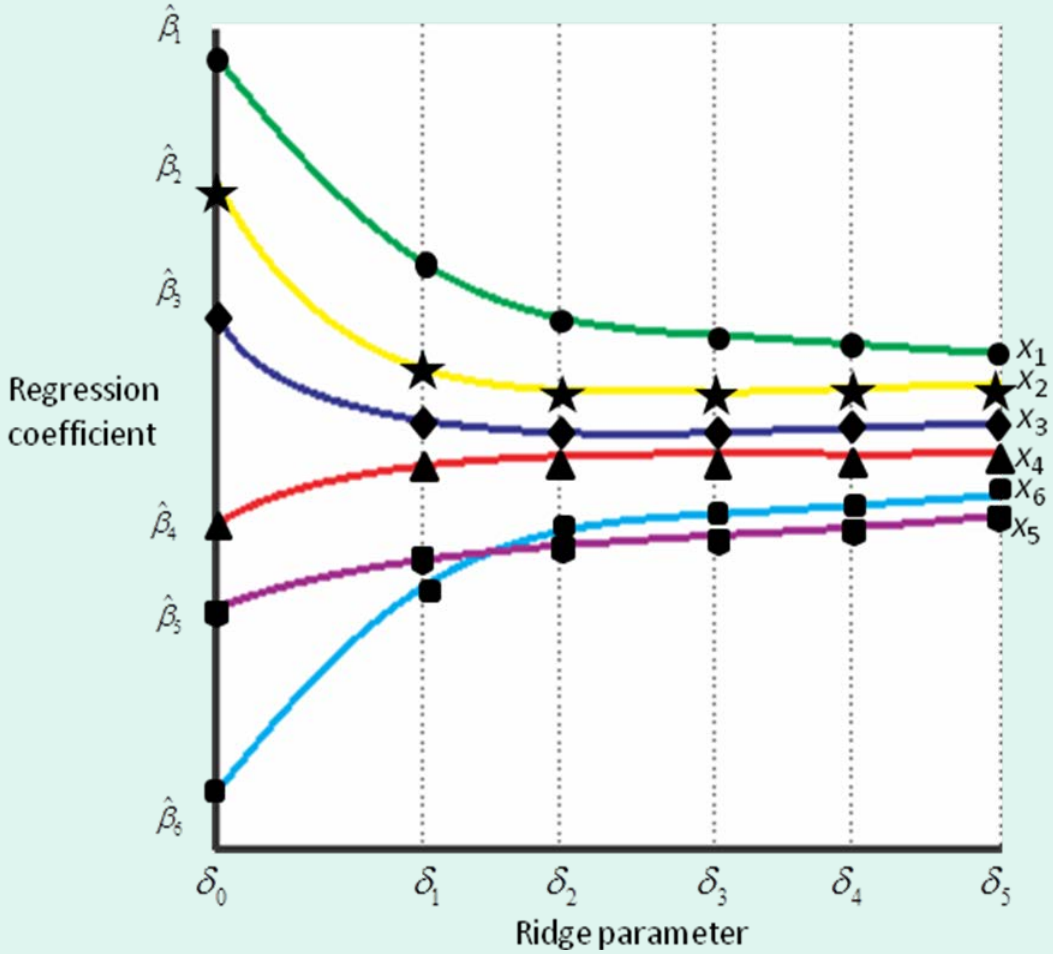
Ridge trace

Ridge trace is the graphical display of ridge regression estimator versus δ .

If multicollinearity is present and is severe, then the instability of regression coefficients is reflected in the ridge trace. As δ increases, some of the ridge estimates vary dramatically and they stabilizes at some value of δ. The objective in ridge trace is to inspect the trace (curve) and find the reasonably small value of δ at which the ridge regression estimators are stable. The ridge regression estimator with such a choice of δ will have smaller MSE than the variance of OLSE.

An example of ridge trace for a model with 6 parameters is as follows . In this ridge trace, the $\hat{\beta}_{ridge}$ is evaluated for various choices of δ and the corresponding values of all regression coefficients $\hat{\beta}_{j(ridge)}$'s, j = 1,2,…,6 are plotted versus δ . These values are denoted by different symbols and are joined by a smooth curve. This produces a ridge trace for respective parameter. Now choose the value of δ where all the curves stabilize and become nearly parallel. For example, the curves in following figure become nearly parallel starting from $\delta = \delta_4$ or so. Thus one possible choice of δ is $\delta = \delta_4$ and parameters can be estimated as

$$\hat{\beta}_{ridge} = \left(X'X + \delta_4 I\right)^{-1} X'y.$$



The figure drastically exposes the presence of multicollinearity in the data. The behaviour of $\hat{\beta}_{i(ridge)}$ at $\delta_0 \approx 0$ is very different than at other values of δ. For small values of δ, the estimates change rapidly. The estimates stabilize gradually as δ increases. The value of δ at which all the estimates stabilize gives the desired value of δ because moving away from such δ will not bring any appreciable reduction in the residual sum of squares. If multicollinearity is present, then the variation in ridge regression estimators is rapid around $\delta_0 \approx 0$. The optimal δ is chosen such that after that value of δ, almost all traces stabilize.

Limitations

1. The choice of δ is data dependent and therefore is a random variable. Using it as a random variable violates the assumption that δ is a constant. This will disturb the optimal properties derived under the assumption of constancy of δ .
2. The value of δ lies in the interval $(0,\infty)$. So large number of values are required for exploration. This results is wasting of time. However, this is not a big issue when working with software.
3. The choice of δ from graphical display may not be unique. Different people may choose different δ and consequently the values of ridge regression estimators will be changing. However δ is chosen so that all the estimators of all the coefficients stabilize. Hence small variation in choosing the value of δ may not produce much change in ridge estimators of the coefficients. Another choice of δ is

$$\delta = \frac{k\hat{\sigma}^2}{b'b}$$

where b and $\hat{\sigma}^2$ are obtained from the least squares estimation.

4. The stability of numerical estimates of $\hat{\beta}_i's$ is a rough way to determine δ. Different estimates may exhibit stability for different δ and it may often be hard to strike a compromise. In such situation, generalized ridge regression estimators are used.
5. There is no guidance available regarding the testing of hypothesis and for confidence interval estimation.

Idea behind ridge regression estimator

The problem of multicollinearity arises because some of the eigenvalues roots of $X'X$ are close to zero or are zero. So if $\lambda_1, \lambda_2, ..., \lambda_p$ are the characteristic roots, and if

$$X'X = \Lambda = diag(\lambda_1, \lambda_2, ..., \lambda_k)$$

then

$$\hat{\beta}_{ridge} = (\Lambda + \delta I)^{-1} X^1 y = (I + \delta \Lambda^{-1})^{-1} b$$

where b is the OLSE of β given by

$$b = (X'X)^{-1} X'y = \Lambda^{-1} X'y.$$

Thus a particular element in $\hat{\beta}_{ridge}$ will be of the form

$$\frac{1}{\lambda_i + \delta} x_i' y = \frac{\lambda_i}{\lambda_i + \delta} b_i.$$

So a small quantity δ is added to λ_i so that if $\lambda_i = 0,$ even then $\frac{1}{\lambda_i + \delta}$ remains meaningful.

Another interpretation of ridge regression estimator

In the model $y = X\beta + \varepsilon,$ obtain the least squares estimator of β when $\sum_{i=1}^{k} \beta_1^2 = C,$ where C is some constant. So minimize

$$\delta(\beta) = (y - X\beta)'(y - X\beta) + \delta(\beta'\beta - C)$$

where δ is the Lagrangian multiplier. Differentiating $S(\beta)$ with respect to β, the normal equations are obtained as

$$\frac{\partial S(\beta)}{\partial \beta} = 0 \Rightarrow -2X'y + 2X'X\beta + 2\delta\beta = 0$$

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

Note that if δ is very small, it may indicate that most of the regression coefficients are close to zero and if δ is large, then it may indicate that the regression coefficients are away from zero. So δ puts a sort of penalty on the regression coefficients to enable its estimation.