MULTICOLLINEARITY & RIDGE REGRESSION

ISMAIL SAVRUK, BILKENT UNIVERSITY

Abstract. In this project, we have investigated a model of multiple regression: **Ridge Regression.** We examine the causes of multicollinearity, some of its specific effects on inference, methods of detecting the presence of multicollinearity and we study a technique for dealing with the multicollinearity problem. We also tried to minimize mean squared error of ridge regression and understand the behaviour of ridge traces.

**The Project Adviser : **Dr. Dilek Güvenç

**The Course Coordinator : **Prof. Dr. Aurelian Gheondea

Introduction

Regression analysis is a statistical techique for investigating and modeling the relationship between variables. Applications of regression are numerous and occur in almost every field, including engineering, the physical scince, economics, management, the biological science and the social science. In fact, regression analysis may be the most widely used statistical techique.

Ridge regression was inroduced in 1962 by Arthur Hoerl in an article in a chemical engineering journal. In his experience in regression analysis he had found that when there were correlations among the explanatory variables, the least squares estimates often did not make sense when he put into the context of the process that generated the data. He proposed a method to obtain better estimates.[4]

1. A Review of Multiple Linear Regression

A regression model that involves more than one regressor variable is called a multiple regression model. Let y be the response variable, $x_1, x_2, ...x_k$ be regressor variables, $\beta_0, \beta_1, ..., \beta_k$ be regression coefficients and ε represents the random error component, then the following model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon$$

is called multiple linear regression model.

The errors are assumed to have zero mean and unknown constant variance σ^2. The errors are independent, moreover we are assumed that the errors are normally and identically distributed.

$$\varepsilon_i \sim NID(0, \sigma^2) \tag{1.1}$$

The regressors x_i's are controlled by the data analyst and measured with neglegible error, while the response y is a random variable. So, there is a probability distribution for y at each possible value for x. The mean of this distribution is

$$E(y|x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$$

and the variance is

$$V(y|x) = V(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon) = \sigma^2.$$

Thus the mean of y is a linear function of x although the variance of y does not depend on the value of x.

An important objective of the regression analysis is to estimate the unknown parametres in the regression model. If we have n observations to find unknown parameters, then the i^{th} observation will be

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \varepsilon_i.$$

It is more convinient to deal with multiple regression models in matrix notation. This allows a very compact display of the model, data and results. The model in terms of the observations may be written as

$$Y = X\beta + \epsilon, \tag{1.2}$$

where

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_{11} & x_{12} & ... & x_{1k} \\ 1 & x_{21} & x_{22} & ... & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & ... & x_{nk} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \epsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

In general, Y is an $n \times 1$ vector of the observations. X is an $n \times p$ matrix of the levels of the regressor variables(here $p = k + 1$), β is a $p \times 1$ vector of the regression coefficients, and ϵ is an $n \times 1$ vector of random errors.

2. Least Squares Estimation of the Regression Coefficients

The parameters $\beta_0, \beta_1, ..., \beta_k$ are unknown and must be estimated using sample data. The method of least squares is used to estimate these unknowns. So, we will estimate $\beta_0, \beta_1, ..., \beta_k$ so that the sum of the squares of the differences between the observations y_i and their true means is a minimum. Suppose that we have n observations in the data, say, $(y_1, x_{11}, \ldots, x_{1k})$, $(y_2, x_{21}, \ldots, x_{2k})$, $\ldots$,$(y_n, x_{n1}, \ldots, x_{nk})$.

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \varepsilon_i \\ &= \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \varepsilon_i, \qquad\qquad i = 1, 2, \ldots, n. \end{aligned}$$

The least squares function is

$$S(\beta_0, \beta_1, \ldots, \beta_k) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \quad y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij} \Bigg)^2.$$

To minimize S w.r.t. $\beta_0, \beta_1, \ldots, \beta_k$, the estimators $\widehat{\beta}_0, \widehat{\beta}_1, ..., \widehat{\beta}_k$ must satisfy the following;

$$\left.\frac{\partial S}{\partial \beta_0}\right|_{\hat\beta_0,\hat\beta_1,\ldots,\hat\beta_k} = -2\sum_{i=1}^{n} \quad y_i - \widehat{\beta}_0 - \sum_{j=1}^{k}\widehat{\beta}_j x_{ij}\Bigg) = 0 \tag{2.1}$$

and

$$\left.\frac{\partial S}{\partial \beta_j}\right|_{\hat\beta_0,\hat\beta_1,\ldots,\hat\beta_k} = -2\sum_{i=1}^{n} \quad y_i - \widehat{\beta}_0 - \sum_{j=1}^{k}\widehat{\beta}_j x_{ij}\Bigg) x_{ij} = 0, \tag{2.2}$$

where $j = 1, 2, \ldots, k$. Simplifing (2.1) and (2.2) we get

$$\sum_{i=1}^{n} Y_i = n\widehat{\beta}_0 + \widehat{\beta}_1\sum_{i=1}^{n} x_{i1} + \widehat{\beta}_2\sum_{i=1}^{n} x_{i2} + \ldots + \widehat{\beta}_k\sum_{i=1}^{n} x_{ik}$$

$$\sum_{i=1}^{n} Y_i x_{i1} = \widehat{\beta}_0\sum_{i=1}^{n} x_{i1} + \widehat{\beta}_1\sum_{i=1}^{n} x_{i1}x_{i1} + \widehat{\beta}_2\sum_{i=1}^{n} x_{i1}x_{i2} + \ldots + \widehat{\beta}_k\sum_{i=1}^{n} x_{i1}x_{ik}$$

$$\vdots$$

$$\sum_{i=1}^{n} Y_i x_{ik} = \widehat{\beta}_0\sum_{i=1}^{n} x_{ik} + \widehat{\beta}_1\sum_{i=1}^{n} x_{ik}x_{i1} + \widehat{\beta}_2\sum_{i=1}^{n} x_{ik}x_{i2} + \ldots + \widehat{\beta}_k\sum_{i=1}^{n} x_{ik}x_{ik}.$$

The above equations are called the **normal equations**. In order to solve $\widehat{\beta}_i$'s and to show them in the matrix notation we need equation (1.2);

$$\begin{aligned} S(\beta) = \sum_{i=1}^{n}\varepsilon_i^2 &= \epsilon^T\epsilon = (Y - X\beta)^T(Y - X\beta), \qquad \textit{since } \epsilon = Y - X\beta \\ &= Y^TY - \beta^TX^TY - Y^TX\beta + \beta^TX^TX\beta \\ &= Y^TY - 2\beta^TX^TY + \beta^TX^TX\beta. \end{aligned}$$

Since β^TX^TY is a 1 by 1 matrix so its transpose equals to itself;

$$(\beta^TX^TY)^T = Y^TX\beta = \beta^TX^TY.$$

Notation: For the sake of simplicity, from now on, we use X' instead of X^T.
Deriving S it w.r.t β we get

$$\left.\frac{\partial S}{\partial \beta}\right|_{\hat\beta} = -2X'Y + 2X'X\widehat{\beta} = 0$$

$$X'X\widehat{\beta} = X'Y. \tag{2.3}$$

So if we look at our normal equations, the left hand side of the equation gives us $X'Y$ and the right hand side is $X'XY$. Now assume that $X'X$ is a nonsingular matrix. Then mutliply (2.3),from the left, by $(X'X)^{-1}$,[5]

$$\begin{aligned} (X'X)^{-1}X'X\widehat{\beta} &= (X'X)^{-1}X'Y \\ \widehat{\beta} &= (X'X)^{-1}X'Y. \end{aligned} \tag{2.4}$$

The uniquness of normal equations: The unique solution of the normal equations exists if the inverse of the $X'X$ matrix exists. This requires that the matrix X be of full rank, there can be no linear dependencies among the independent variables.

3. Properties of Least Square Estimators

Expressing $\widehat{\beta}$ as $\widehat{\beta} = [(X'X)^{-1}X']Y$ shows that the estimates of the regression coefficients are linear functions of the dependent variable Y, with the coefficients being given by $[(X'X)^{-1}X']$. Assuming that X_i's are constants, the expectations of the estimated regression coefficients involve only the expectation of Y. Let us find the expectation and variance of the $Y, \widehat{\beta}$ and $\widehat{Y}$.

The Y vector: Since the elements of X and β are constants, the $X\beta$ term in the model is a set of constants being added to the vector of random errors. Using (1.1);

$$E(Y) = E(X\beta + \epsilon) = E(XB) + \underbrace{E(\epsilon)}_{0} = XB$$

and

$$V(Y) = V(X\beta + \epsilon) = V(\epsilon) = I\sigma^2.$$

Variance of Y is the same as variance of ϵ since adding a constant to a random variable does not change the variance. When ϵ is normally distributed, Y is also normally distributed. Thus

$$Y \sim N(X\beta, I\sigma^2).$$

The $\widehat{\beta}$ vector: To find expectation and variance of $\widehat{\beta}$ we need (2.4)

$$\begin{aligned} E(\widehat{\beta}) &= E((X'X)^{-1}X'Y) \\ &= ((X'X)^{-1}X')E(Y) \\ &= ((X'X)^{-1}X')(X\beta) \\ &= ((X'X)^{-1}X'X)\beta) \\ &= \beta \end{aligned}$$

and to find variance we need the following fact.

Fact: Variance of any distubition has the property that $V(cx) = c^2V(x)$ where c is constant. However if C and X are matrices and if C is matrix of constants, which means X is independent of C, then the above property becomes $V(CX) = CC'V(X)$[5]. Thus

$$\begin{aligned} V(\widehat{\beta}) &= V((X'X)^{-1}X'Y) \\ &= ((X'X)^{-1}X')((X'X)^{-1}X')'V(Y) \\ &= (X'X)^{-1}(X'X)(X'X)^{-1}\sigma^2 \\ &= (X'X)^{-1}\sigma^2. \end{aligned}$$

The $\widehat{Y}$ vector: The vector of estimated of the dependent variable Y for the values of the independent variables in the data set is computed as

$$\widehat{Y} = X\widehat{\beta} + \widehat{\epsilon} = X\widehat{\beta}.$$

$\widehat{\epsilon} = 0$ since $E(\epsilon) = 0$. It is useful to express $\widehat{Y}$ as a linear function of Y by substituting $(X'X)^{-1}X'Y$ for $\widehat{\beta}$. Thus

$$\begin{aligned} \widehat{Y} &= X\widehat{\beta} \\ &= X[(X'X)^{-1}X'Y] \\ &= [X(X'X)^{-1}X']Y \\ &= PY. \end{aligned} \tag{3.1}$$

Equation(3.1) defines the matrix P, an $n \times n$ matrix determined entirely by the $X's$, $P = X(X'X)^{-1}X'$. This matrix plays an important role in the regression analysis. It is a symmetric matrix ($P' = P$). Expectation of $\widehat{Y}$ is

$$E(\widehat{Y}) = E(PY) = PE(Y) = PX\beta = X\beta.$$

The variance of $\widehat{Y}$ can be calculated as follows;

$$V(\widehat{Y}) = V(PY) = PP'V(Y) = P\sigma^2.$$

The residuals vector e: The residuals vector e reflects the lack of agreement between the observed Y and the estimated $\widehat{Y}$

$$\begin{aligned} e &= Y - \widehat{Y} \\ &= Y - PY \\ &= (I - P)Y. \end{aligned}$$

The expectation of the residuals vector is

$$\begin{aligned} E(e) &= E((I - P)Y) \\ &= (I - P)E(Y) \\ &= (I - P)X\beta \\ &= (X - PX)\beta \\ &= (X - X)\beta \\ &= 0. \end{aligned}$$

The variance of the residuals vector is

$$\begin{aligned} V(e) &= V((I - P)Y) \\ &= (I - P)V(Y) \\ &= (I - P)\sigma^2. \end{aligned}$$

4. Covariance and Correlation Matrices

The **covariance matrix** is a matrix of covariances between variables. If there are k independent variables in the regression model, this will be a $k \times k$ square matrix. A dioganal element of this matrix shows the variance of that variable and the off-diagonal elements are covariances.[3] So, it is also called variance-covariance matrix. And clearly this matrix is symmetric because $Cov(x_i, x_j) = Cov(x_j, x_i)$. If $i = j$ then $Cov(x_i, x_i) = Var(x_i)$. Let W shows the covariance matrix. Then,

$$W = \begin{bmatrix} Var(x_1) & Cov(x_1, x_2) & ... & Cov(x_1, x_k) \\ Cov(x_2, x_1) & Var(x_2) & ... & Cov(x_2, x_k) \\ \vdots & \vdots & & \vdots \\ Cov(x_k, x_1) & Cov(x_k, x_2) & ... & Var(x_k) \end{bmatrix}.$$

The **correlation matrix** is a $k \times k$ square matrix of k random variables whose (i, j)-th entry shows the correlation between X_i and X_j. Correlation, (often measured as a correlation coefficient), indicates the strength and direction of a linear relationship between two random variables. In general statistical usage, correlation or co-relation refers to the departure of two variables from independence. The correlation matrix is also symmetric because the correlation between between X_i and X_j is the same as the correlation between X_j and X_i. The diagonal elements of correlation matrix shows the correlation between X_i and X_i which equals to 1. The correlation between X_i and X_j is always between -1 and $+1$. If the absolute value of the correlation between X_i and X_j is closed to 0, it means these 2 variables are not correlated. If it is closed to 1, these variables are highly correlated.[3] The correlation matrix is the same as the *covariance matrix* of the standardized (centered and scaled) variables. Let R shows the correlation matrix. Then,

$$(4.1) \quad R = \begin{bmatrix} 1 & r_{12} & ... & r_{1k} \\ r_{21} & 1 & ... & r_{2k} \\ \vdots & \vdots & & \vdots \\ r_{k1} & r_{k2} & ... & 1 \end{bmatrix} \quad where \quad r_{ij} = \frac{\sum_{l=1}^{n} (x_{il} - \overline{x_i})(x_{jl} - \overline{x_j})}{\sqrt{\sum_{l=1}^{n} (x_{il} - \overline{x_i})^2 \sum_{n=1}^{k} (x_{jl} - \overline{x_j})^2}}$$

Centering and Scaling the Data (Standardized Variables): It is usually difficult to directly compare regression coefficients because the magnitude of $\widehat{\beta_j}$ reflects the units of measurement of the regressor x_j. For instance, suppose in a regression model y is measured in liters, x_1 is measured in mililiters and x_2 is measured in liters. Although $\widehat{\beta_2}$ is considerably larger than $\widehat{\beta_1}$, the effect of both regressors on $\widehat{y}$ is identical, since a one liter change in either x_1 or x_2 when the other variables are held constant produces the same change in $\widehat{y}$. For this reason, it is sometimes helpful to work with centered and scaled regressor and response variables that produce dimensionless regression coefficients[2]. Centering and scaling the data transforms a data set by subtracting the column mean from each column and dividing each column by the square root of its variance(standart deviation). The standardized regression variables, represent the change in a dependent variable that result from a change of one standard deviation in an independent variable. Let our model be

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon.$$

We can rewrite it in the following form by substracting and adding the mean of each variable

$$y = \underbrace{\{\beta_0 + \beta_1 \overline{x_1} + \beta_2 \overline{x_2} + ... + \beta_k \overline{x_k}\}}_{\beta_0'} + \beta_1 \underbrace{(x_1 - \overline{x_1})}_{z_1} + \beta_2 \underbrace{(x_2 - \overline{x_2})}_{z_2} + ... + \beta_k \underbrace{(x_k - \overline{x_k})}_{z_k} + \varepsilon.$$

So,

$$y = \beta_0' + \beta_1 z_1 + \beta_2 z_2 + ... + \beta_k z_k + \varepsilon.$$

The i^{th} observation is

$$y_i = \beta_0' + \beta_1 z_{i1} + \beta_2 z_{i2} + ... + \beta_k z_{ik} + \varepsilon_i.$$

The residual sum of square will be

$$\sum_{i=1}^{n} \varepsilon_i^2 \;=\; \sum_{i=1}^{n} (y_i - \beta_0' - \beta_1 z_{i1} - \beta_2 z_{i2} - ... - \beta_k z_{ik})^2.$$

To find the normal equations take the derivative w.r.t. β_0', we get

$$-2 \sum_{i=1}^{n} (y_i - \beta_0' - \beta_1 z_{i1} - \beta_2 z_{i2} - ... - \beta_k z_{ik}) = 0$$

$$\sum_{i=1}^{n} y_i = n\beta_0' + \beta_1 \sum_{i=1}^{n} z_{i1} + \beta_2 \sum_{i=1}^{n} z_{i2} + ... + \beta_k \sum_{i=1}^{n} z_{ik} = n\overline{y}.$$

Then,

$$\overline{y} = \beta_0' + \beta_1 \overline{z_1} + \beta_2 \overline{z_2} + ... + \beta_k \overline{z_k} = \beta_0'.$$

Because $z_i = x_i - \overline{x_i}$ then $\overline{z_i} = \overline{x_i} - \overline{x_i} = 0$. Now we have the following **centered data**;

$$y - \overline{y} = \beta_1 (x_1 - \overline{x_1}) + \beta_2 (x_2 - \overline{x_2}) + ... + \beta_k (x_k - \overline{x_k}) + \varepsilon.$$

So, our centered X matrix is dentoded by C,

$$C = \begin{bmatrix} x_{11} - \overline{x_1} & x_{12} - \overline{x_2} & ... & x_{1k} - \overline{x_k} \\ x_{21} - \overline{x_1} & x_{22} - \overline{x_2} & ... & x_{2k} - \overline{x_k} \\ \vdots & \vdots & & \vdots \\ x_{n1} - \overline{x_1} & x_{n2} - \overline{x_2} & ... & x_{nk} - \overline{x_k} \end{bmatrix}.$$

Now in order to **scale** the centered data, we need to divide each column by standart deviation of its variable.

$$x_{ij}^* = \frac{x_{ij} - \overline{x_i}}{\sqrt{S_{ii}}} \;,\;\; y_i^* = \frac{y_i - \overline{y}}{\sqrt{S_{yy}}} \;,$$

where $\quad S_{yy} = \sum_{i=1}^{n} (y_i - \overline{y})^2 \quad$ and $\quad S_{ij} = \sum_{l=1}^{n} (x_{il} - \overline{x_i})(x_{jl} - \overline{x_j})$.

Let us produce our centered and scaled X and Y matrices, which we called Z and Y^* respectively;

$$Z = \begin{bmatrix} x_{11}^* & x_{12}^* & ... & x_{1k}^* \\ x_{21}^* & x_{22}^* & ... & x_{2k}^* \\ \vdots & \vdots & & \vdots \\ x_{n1}^* & x_{n2}^* & ... & x_{nk}^* \end{bmatrix} \quad Y^* = \begin{bmatrix} y_1^* \\ y_2^* \\ \vdots \\ y_n^* \end{bmatrix}.$$

If we compute $Z'Z$ matrix;

$$Z'Z = \begin{bmatrix} 1 & r_{12} & ... & r_{1k} \\ r_{21} & 1 & ... & r_{2k} \\ \vdots & \vdots & & \vdots \\ r_{k1} & r_{k2} & ... & 1 \end{bmatrix},$$

which is the form of correlation matrix R in (4.1). So our least squares regression coefficients will be in the following form;

$$\widehat{b} = (Z'Z)^{-1}Z'Y^*.$$

The new regression coefficients $\widehat{b}$ are called **standardized regression coefficients**. The relationship between the original and standardized regression coefficients is[4]

$$\widehat{\beta_j} = \widehat{b_j}\left(\frac{S_{yy}}{S_{jj}}\right)^{1/2}$$

and

$$\widehat{\beta_0} = \overline{y} - \sum_{j=1}^{k} \widehat{\beta_j}\overline{x_j}.$$

5. Multicollinearity

Multicollinearity occurs when the existance of high correlations among the independent variables in a regression model. In other words when there are near linear dependencies between the regressors, the problem of *multicollineariy* is said to exist. Two variables are collinear if there is an exact linear relationship between the two. A set of variables is collinear if there exists one or more linear relationships among the variables. Let us define multicollinearity in terms of the linear dependence of the columns of X. The vectors $X_1, X_2, \ldots, X_k$ are linearly dependent is there is a set of constants $c_1, c_2, \ldots, c_k$ not all zero, such that[3]

$$\sum_{i=1}^{k} c_i X_i = 0. \tag{5.1}$$

If (5.1) holds for a subset of the columns of X, then the rank of the $X'X$ matrix is less then k(Note that $X'X$ matrix is $k \times k$ square matrix). Hence $X'X$ matrix will be singular. Then $(X'X)^{-1}$ does not exist. It means we cannot find the regression coefficients $\widehat{\beta_j}$'s.(Since $\widehat{\beta} = (X'X)^{-1}X'Y$) However, if the equation (5.1) is *approximately true* for some subset of the columns of X. Then there will a be *near linear dependency* in $X'X$ matrix and the problem of multicollinearity again exists.

Effects of Multicollinearity: The presence of multicollinearity has a number of serious effects on the least squares estimates of the regression coefficients. For example, suppose that there are only two regressor variables, x_1 and x_2. The model, assuming that x_1, x_2 and y are standardized, is

$$y = \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

and the least squares normal equations are $(X'X)\widehat{\beta} = X'Y$. Then,

$$\begin{bmatrix} 1 & r_{12} \\ r_{21} & 1 \end{bmatrix} \begin{bmatrix} \widehat{\beta_1} \\ \widehat{\beta_2} \end{bmatrix} = \begin{bmatrix} r_1 y \\ r_2 y \end{bmatrix},$$

where r_{jy} is the simple correlation between x_j and y. Now the inverse of $X'X$ is

$$D = \begin{bmatrix} \frac{1}{(1-r_{12}^2)} & \frac{-r_{12}}{(1-r_{12}^2)} \\ \frac{-r_{12}}{(1-r_{12}^2)} & \frac{1}{(1-r_{12}^2)} \end{bmatrix}. \tag{5.2}$$

And the estimates of the regression coefficients are

$$\widehat{\beta_1} = \frac{r_{1y} - r_{12} r_{2y}}{(1 - r_{12}^2)} \quad , \qquad \widehat{\beta_2} = \frac{r_{2y} - r_{12} r_{1y}}{(1 - r_{12}^2)} \quad .$$

If there is strong multicollinearity between x_1 and x_2, then the correlation coefficient r_{12} will be large. From (5.2) we see that as $|r_{12}|\rightarrow 1$, $V(\widehat{\beta_j}) = C_{jj}\sigma^2 \rightarrow \infty$ and $Cov(\widehat{\beta_1}, \widehat{\beta_2}) = C_{12}\sigma^2 \rightarrow \infty$. Then strong multicollinearity between x_1 and x_2 results in large variances and covariances for the least squares estimators of the regression coefficients. This implies that different samples taken at the same x levels could lead to widely different estimates of the model parametres. When there are more than two regressor variables, multicollinearity produces similar effects.

6. Biased Estimation

The least squares estimators of the regression coefficients are the best linear unbiased estimators. That is, of all possible estimators that are both linear functions of the data and unbiased for the parameters being estimated, the least squares estimators have the smallest variance. In the presence of collinearity, however, this minimum variance may be unaccceptably large. Relaxing the least squares conditions that estimators be unbiased opens for consideration a much larger set of possile estimators from which one with better properties in the presence of collinearity might be found. **Biased regression** refers to this class of regression methods in which unbiasedness is no longer required. Such methods have been suggested as a possible solution to the collinearity problem.[1]

Mean Squared Error(MSE): The mean squared error or MSE of an estimator is a way to quantify the amount by which an estimator differs from the true value of the quantity being estimated. MSE measures the average of the square of the error. The error is the amount by which the estimator differs from the quantity to be estimated. Minimizing MSE is a key criterion in selection estimators. Among unbiased estimators, the minimal MSE is equivalent to minimizing the variance, however, a biased estimator

may have lower MSE.
If $\widehat{\theta}$ is the estimator of θ, then MSE can be defined as following;

$$MSE(\widehat{\theta}) = E(\widehat{\theta} - \theta)^2.$$

Let us find another way of interpretation of MSE. We know that

$$Var(\widehat{\theta}) = E[\widehat{\theta} - E(\widehat{\theta})]^2 = E(\widehat{\theta^2}) - (E(\widehat{\theta}))^2 \tag{6.1}$$

and bias is defined as

$$Bias(\widehat{\theta}) = |E(\widehat{\theta}) - \theta|. \tag{6.2}$$

So MSE can be copmuted using (6.1), (6.2) and the obvious fact $E(\theta) = \theta$.

$$\begin{aligned} MSE(\widehat{\theta}) = E(\widehat{\theta} - \theta)^2 &= E(\widehat{\theta^2}) - 2E(\widehat{\theta})E(\theta) + E(\theta^2) - E(\widehat{\theta})^2 + E(\widehat{\theta})^2 \\ &= \underbrace{E(\widehat{\theta^2}) - E(\widehat{\theta})^2}_{} + \underbrace{E(\widehat{\theta})^2 - 2E(\widehat{\theta})E(\theta) + E(\theta^2)}_{} \\ &= \qquad Var(\widehat{\theta}) \qquad + \qquad (Bias(\widehat{\theta}))^2. \end{aligned}$$

It is possible for the variance of a biased estimators to be sufficiently smaller than the variance of an unbiased estimator. In this case, the biased estimator is closed on the average to the parameter being estimated than is the unbiased estimator. As a simple example of this fact, suppose we have two estimators $\widehat{\theta}_1$ and $\widehat{\theta}_2$ for a parameter θ. Let $\widehat{\theta}_1$ is unbiased, i.e. $E(\widehat{\theta}_1) = \theta$, so $Bias(\widehat{\theta}_1) = 0$ and $\widehat{\theta}_2$ is biased and say the bias is 4. And suppose variance of the first parameter is 100 and the second one is 16. Now if we compute $MSE_1 = MSE(\widehat{\theta}_1)$ and $MSE_2 = MSE(\widehat{\theta}_2)$, we get

$$MSE_1 = 100 + 0 = 100$$
$$MSE_2 = 16 + 4^2 = 32.$$

Although θ_2 is unbiased, $MSE_2 < MSE_1$, thus, we choose θ_2 as an estimator of θ. To sum up, *biased estimation* can be suggested as a possible solution of collinearity problem in some certain conditions.

7. Ridge Regression

A number of procedures have been developed for obtaining biased estimators of regression coeffients. One of these procedures is ridge regression, originally proposed by Hoerl and Kennard[2]. The ridge estimator is found by solving a slightly modified version of the normal equations. In other words ridge regression builds on the fact that a singular square matrix can be made nonsingular by adding a constant to the diagonal of the matrix. That is, if $X'X$ is singular, than $(X'X + kI)$ is nonsingular, where k is some small positive constant,generally at most 1.$(0 \leq k \leq 1)$, and I represents the identity matrix. This concept is added to the diagonal of the nearly singular $X'X$. Specifically, we define the ridge estimator $\widehat{\beta_k}$ as the solution to

$$(X'X + kI)\widehat{\beta_k} = X'Y,$$

so

$$\widehat{\beta_k} = (X'X + kI)^{-1}X'Y.$$

When k=0, the ridge estimator is the least squares estimator. Ridge regression works with the centered and scaled independent variables Z so that the sum of squares and products matrix of the independent variables is the correlation matrix. Using Z, which does not include the vector of ones for the intercept, we can write the model as

$$Y = 1\beta_0 + Z\beta + \epsilon,$$

where **1** is the column vector of ones and β is the vector of all regression coefficients except β_0. The ridge estimator is a linear transformation of the least squares estimator, since

$$\begin{aligned}\widehat{\beta_k} &= (Z'Z + kI)^{-1}Z'Y^* \\ &= (Z'Z + kI)^{-1}(Z'Z)\widehat{\beta} \\ &= Z_k\widehat{\beta}.\end{aligned}$$

Consequently $\widehat{\beta_k}$ is a biased estimator of β, because $E(\widehat{\beta_k}) = E(Z_k\widehat{\beta}) = Z_k\beta$.

The MSE of Ridge Regression: To find the MSE, we need variance of $\widehat{\beta_k}$ and $Bias(\widehat{\beta_k})$. Hence,

$$\begin{aligned}Var(\widehat{\beta_k}) &= Var((Z'Z + kI)^{-1}Z'Y^*) \\ &= (Z'Z + kI)^{-1}Z'(Var(Y^*))[(Z'Z + kI)^{-1}Z']' \\ &= \sigma^2 I(Z'Z + kI)^{-1}Z'(Z')'((Z'Z + kI)')^{-1} \\ &= \sigma^2(Z'Z + kI)^{-1}(Z'Z)(Z'Z + kI)^{-1}\end{aligned}$$

and $Bias(\widehat{\beta_k})$ computed as follows[2]

$$Bias(\widehat{\beta_k})^2 = (E(Z_k\widehat{\beta}) - \beta)^2 = (Z_k\beta - \beta)^2 = k^2\beta'((Z'Z + kI)^{-2})\beta.$$

So,

$$\begin{aligned}MSE(\widehat{\beta_k}) &= Var(\widehat{\beta_k}) + Bias(\widehat{\beta_k}) \\ &= \sigma^2 Tr[(Z'Z + kI)^{-1}(Z'Z)(Z'Z + kI)^{-1}] + k^2\beta'((Z'Z + kI)^{-2})\beta \\ &= \sigma^2\sum_{i=1}^{p}\frac{\lambda_i}{(\lambda_i + k)^2} + k^2\beta'((Z'Z + kI)^{-2})\beta, \qquad (7.1)\end{aligned}$$

where λ_i's are the eigenvalues of $Z'Z$ matrix with size $p \times p$.[2]

The Choice of k: The goal is to choose k in order to minimize the MSE. One can observe from the equation (7.1) that the bias increases with k, when the variance decreases as k increases. The MSE for any particular regression coefficient is expected to decrease initially, because of initial rapid decreases in variance, and increase as bias begins to dominate. Therefore, the strategy is to choose the smallest k that appears to be producing stable estimates of the regression coefficients. Hoerl and Kennard have suggested that an appropriate value of k may be determinated by inspectation of the **ridge trace**. The ridge trace is a plot of the elements of $\widehat{\beta_k}$ versus k. If multicolinearity is severe, the instability in the regression coefficients will be obvious from the ridge trace. As k increased, some of the ridge estimates will vary enormously. At some value of k, the ridge estimates $\widehat{\beta_k}$'s will stabilize. The objective is to select a reasonably small k at which the ridge estimates $\widehat{\beta_k}$ are stable. Therefore, this will produce a set of estimates with smaller MSE than least square estimates. Hoerl and Kennard suggest the use of

$$k = \frac{ps^2}{\widehat{\beta_0}'\widehat{\beta_0}}\ , \tag{7.2}$$

where p is the number of parameters excluding β_0 and s^2 is the residual mean square estimated from the ordinary least squares regression$(k = 0)$[1] Given the chosen value of k, the ridge regression equation becomes

$$\widehat{Y} = 1\overline{Y} + Z\widehat{\beta_k}.$$

The key steps in ridge regression are summarized as follows:

(1) Center and scale the independent variables to obtain Z.
(2) Compute $Z'Z$ and $Z'Y$.
(3) Compute ordinary least squares results and compute the value of k.
(4) For a squence of constants,say $k_1, k_2, \ldots$, including the value from the equation (7.2), compute $\widehat{\beta_k}$ and $Var(\widehat{\beta_k})$.
(5) Plot ridge traces, $(\widehat{\beta_k})_j$ versus k.
(6) Choose the value of k by the equation (7.2) or where the ridge traces have stabilized.

The results for the choosen value of k are the ridge regression solution.

References

[1] J.O. Rawlings, *Applied Regression Analysis: A Research Tool*, Wadsworth and Brooks/Cole Advanced Books and Software, California, 1988.

[2] D.C. Montgomery, E.A. Peck, *Introduction to Linear Regression Analysis*, Wiley Series in Probability and Mathematical Statistics, New York, 1992.

[3] S. Weisberg, *Applied Linear Regression*, Wiley Series in Probability and Mathematical Statistics, New York, 1985.

[4] D. Birkes, Y. Dodge, *Alternative Methods of Regression*, Wiley Series in Probability and Mathematical Statistics, New York, 1993.

[5] N.R. Draper, H. Simith, *Applied Regression Analysis*, Wiley Series in Probability and Mathematical Statistics, New York, 1966.