PhUSE 2012

Paper SP07

Multicollinearity Diagnostics in Statistical Modeling & Remedies to deal with it using SAS

Harshada Joshi, Cytel Statistical Software & Services Pvt Ltd, Pune, India
Hrishikesh Kulkarni, Cytel Statistical Software & Services Pvt Ltd, Pune, India
Swapna Deshpande, Cytel Statistical Software & Services Pvt Ltd, Pune, India

ABSTRACT

Regression modeling is one of the most widely used statistical techniques in clinical trials. Many a times, when we fit a multiple regression model, results may seem paradoxical. For instance, the model may fit the data well, even though none of the predictors has a statistically significant impact on explaining the outcome variable. How is this possible? This happens when multicollinearity exists between two or more predictor variables. If the problem of multicollinearity is not addressed properly, it can have a significant impact on the quality and stability of the fitted regression model. The aim of the proposed paper is to explain the issue of multicollinearity, effects of multicollinearity, various techniques to detect multicollinearity and the remedial measures one should take to deal with it. The paper will focus on explaining it theoretically as well as using SAS® procedures such as PROC REG and PROC PRINCOMP.

INTRODUCTION

Recall that the multiple linear regression model is

$$Y = X\beta + \epsilon \, ;$$

Where Y is an n x 1 vector of responses, X is an n x p matrix of the regressor variables, β is a p x 1 vector of unknown constants, and ϵ is an n x 1 vector of random errors, with $\epsilon_i \sim NID(0, \sigma^2)$

Multicollinearity, or near-linear dependence, is a statistical phenomenon in which two or more predictor variables in a multiple regression model are highly correlated.

Let the J^{th} column of the matrix X be Xj, so that $X = [X_1, X_2, \ldots X_p]$. Thus, Xj contains the n levels of the J^{th} regressor variable. We can define multicollinearity in terms of the linear dependence of the columns of X. The vectors $X_1, X_2, \ldots X_p$ are linearly dependent if there is a set of constants $c_1, c_2, \ldots c_p$, not all zero, such that

$$\sum_{j=1}^{p} c_j X_j = 0 \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots..(1)$$

Here, we assume that the regressor variables and the response have been centered and scaled to unit length. Consequently X'X is a p x p matrix of correlations between the regressors and X'Y is a p x1 vector of correlations between the regressors and the response.

If eq. (1) holds exactly for a subset of the column of X, then the rank of the X'X matrix is less than p and inverse of (X'X) does not exist. However if eq. (1) is approximately true for some subset of the columns of X, then there will be a near-linear dependency in X'X and the problem of multicollinearity is said to exist which will make the X'X matrix ill-conditioned.

EFFECTS OF MULTICOLLINEARITY

The presence of multicollinearity has a number of potentially serious effects on the least-squares estimates of the regression coefficients. Suppose that there are only two regressor variables, X_1 and X_2. The model, assuming that x_1, x_2, and y, are scaled to unit length, is

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

and the least-squares normal equations are

$$(X'X)\ \hat{\beta}1 = X'y$$

$$\begin{bmatrix} 1 & r_{12} \\ r_{12} & 1 \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} r_{1y} \\ r_{2y} \end{bmatrix}$$

Where r_{12} is the simple correlation between x1 and x2 and r_{jy} is the simple correlation between xj and y, j=1,2.
Now, the inverse of (X'X) is

$$C = (X'X)^{-1} = \begin{bmatrix} 1/(1-r_{12}*r_{12}) & -r_{12}/(1-r_{12}*r_{12}) \\ -r_{12}/(1-r_{12}*r_{12}) & 1/(1-r_{12}*r_{12}) \end{bmatrix} \quad \ldots\ldots\ldots\ldots\ldots\ldots(2)$$

And the estimates of the regression coefficients are

$$\hat{\beta}_1 = (r_{1y}-r_{12}r_{2y})/(1-r^2_{12}) \ ,$$
$$\hat{\beta}_2 = (r_{2y}-r_{12}r_{1y})/(1-r^2_{12}) \quad \ldots\ldots\ldots\ldots\ldots\ldots(3)$$

If there is a strong multicollinearity between X_1 and X_2, then the correlation coefficient r_{12} will be large (close to ±1).
From eq. (2) we see that as $|r_{12}| \rightarrow 1$, Var($\hat{\beta}$j) = $C_{jj}\ \sigma^2 \rightarrow \infty$

And Cov ($\hat{\beta}_1$, $\hat{\beta}_2$) = $C_{12}\ \sigma^2 \rightarrow \pm\infty$ depending on whether $r_{12} \rightarrow +1$ or $r_{12} \rightarrow -1$.

Therefore, strong multicollinearity between X_1 and X_2 results in large variances and covariances for the least-squares estimators of the regression coefficients.

When there are more than two regressor variables, multicollinearity produces similar effects.
The diagonal elements of the C = $(X'X)^{-1}$ matrix are

$$C_{jj} = 1/\ (1-Rj^2),\ \ j=1,2,\ \ldots,p \quad \ldots\ldots\ldots\ldots\ldots\ldots(4)$$

Where Rj^2 is the coefficient of determination from the regression of Xj on the remaining p-1 regressor variables.

If there is strong multicollinearity between Xj and any subset of the other p-1 regressors, then the value of Rj^2 will be close to unity. Since var($\hat{\beta}_j$) = $C_{jj}\ \sigma^2 = (1-Rj^2)^{-1}\ \sigma^2$, strong multicollinearity implies that the variance of the least-squares estimate of the regression coefficient βj is very large. Generally, the covariance of $\hat{\beta}$i and $\hat{\beta}$j will also be large if the regressors Xi and Xj are involved in a multicollinear relationship.

Multicollinearity also tends to produce least-squares estimates $\hat{\beta}_j$ that are too large in absolute value.

To see this, consider the squared distance from $\hat{\beta}$ to the true parameter vector β.

For example, $$L_1^2 = (\hat{\beta}-\beta)'(\hat{\beta}-\beta) \qquad \text{.....................(5)}$$

The expected squared distance, $E(L_1^2)$, is

$$
\begin{aligned}
E(L_1^2) &= E(\hat{\beta}-\beta)'(\hat{\beta}-\beta) \\
&= \sum E(\hat{\beta}_j-\beta_j)^2 \quad \text{where } j=1 \text{ to } p \\
&= \sum Var(\hat{\beta}_j) \quad \text{where } j=1 \text{ to } p \\
&= \sigma^2 Tr(X'X)^{-1} \qquad \text{.........................(6)}
\end{aligned}
$$

Where, the trace of a matrix (Tr) is just the sum of the main diagonal elements. When multicollinearity is present, some of the eigenvalues of X'X will be small. Since the trace of a matrix is also equal to the sum of its Eigenvalues, eq (6) becomes

$$E(L_1^2) = \sigma^2 \sum 1/\lambda_j \quad \text{where } j=1 \text{ to } p \qquad \text{.........................(7)}$$

Where $\lambda_j > 0$, j=1 to p are the eigenvalues of X'X. Thus, if the X'X matrix is ill-conditioned because of multicollinearity, at least one of the λ_j will be small, and eq. (7) implies that the distance from least-square estimates $\hat{\beta}$ to the true parameters β may be large.

$$
\begin{aligned}
E(L_1^2) &= E(\hat{\beta}-\beta)'(\hat{\beta}-\beta) \\
&= E(\hat{\beta}'\hat{\beta}-2\hat{\beta}'\hat{\beta}+\beta'\beta)
\end{aligned}
$$

OR,

$$E(\hat{\beta}'\hat{\beta}) = \beta'\beta + \sigma^2 Tr(X'X)^{-1} \qquad \text{.........................(8)}$$

That is, the vector $\hat{\beta}$ is generally longer than the vector β.

This implies that the method of least squares produces estimated regression coefficients that are too large in absolute value.

DETECTION OF MULTICOLLINEARITY

Following are the indications of possible multicollinearity:

EXAMINATION OF THE CORRELATION MATRIX:

Large correlation coefficients in the correlation matrix of predictor variables indicate the possibility of multicollinearity. We can check this by examining the off-diagonal elements r_{ij} in X'X matrix. If regressors Xi and Xj are nearly linearly dependent, then $|r_{ij}|$ will be near to unity.

VARIANCE INFLATION FACTOR (VIF):

The variance inflation factor (VIF) quantifies the severity of multicollinearity in an ordinary least squares regression analysis.

Let R_j^2 denote the coefficient of determination when Xj is regressed on all other predictor variables in the model.

Let $VIF_j = 1/(1 - R_j^2)$ $j = 1,2, \ldots\ldots p-1$

$VIF_j = 1$ when $R_j^2 = 0$, i.e. when X_j is not linearly related to the other predictor variables.
$VIF_j \rightarrow \infty$ when $R_j^2 \rightarrow 1$, i.e. when X_j tends to have a perfect linear association with other predictor variables.

The VIF provides an index that measures how much the variance of an estimated regression coefficient is increased because of the multicollinearity.

For eg. If the VIF of a predictor variable X_j is 9, this means that the variance of the estimated β_j is 9 times as large as it would be if X_j is uncorrelated with other predictor variables.

As per practical experience, if any of the VIF values exceeds 5 or 10, it is an indication that the associated regression coefficients are poorly estimated because of multicollinearity (Montgomery, 2001).

EIGENSYSTEM ANALYSIS OF X'X:

The eigenvalues can also be used to measure the presence of multicollinearity. If there are one or more near-linear dependences in the predictor variables, then one or more of the eigenvalues will be small.

Let $\lambda_1 \ldots\ldots\ldots \lambda_p$ be the eigenvalues of X'X. The condition number of X'X is defined as

$$K = \lambda_{max} \,/\, \lambda_{min}.$$

A matrix with a low condition number is said to be well-conditioned, while a matrix with a high condition number is said to be ill-conditioned. The condition indices of X'X are defined as

$$K_j = \lambda_{max} \,/\, \lambda_j, \qquad j=1,2,\ldots\ldots,p.$$

The number of condition indices that are large is a useful measure of the number of near-linear dependences in X'X.

Generally, if the condition number is less than 100, there is no serious problem with multicollinearity and if a condition number is between 100 and 1000 implies a moderate to strong multicollinearity. Also, if the condition number exceeds 1000, severe multicollinearity is indicated (Montgomery, 2001).

Consider the data from a trial in which the primary endpoint is the change in the disability score in patients with a neurodegenerative disease.

The main aim of the trial is to find out the relation between change in the disability score and the following explanatory variables.

Explanatory variables:

- Age
- Duration of Disease
- Number of relapses within one year prior to study entry
- Disability score
- Total number of lesions
- Total volume of lesions

In order to detect the possible multicollinearity by examining the correlation matrix, PROC CORR procedure has to be performed.

```
proc corr data=one SPEARMAN;
var age dur nr_pse dscore num_l vol_l;
run;
```

The procedure generates the following SAS output:

Table 1:

Spearman Correlation Coefficients
Prob > |r| under H0: Rho=0

	age	dur	nr_pse	dscore	num_l	vol_l
age	1.00000	-0.16152	0.18276	-0.11073	-0.29810	-0.38682
age		0.3853	0.3251	0.5532	0.1033	0.0316
dur	-0.16152	1.00000	0.04097	0.00981	0.07541	0.14260
dur	0.3853		0.8268	0.9582	0.6868	0.4441
nr_pse	0.18276	0.04097	1.00000	0.43824	0.19606	0.13219
nr_pse	0.3251	0.8268		0.0137	0.2905	0.4784
dscore	-0.11073	0.00981	0.43824	1.00000	0.40581	0.35395
dscore	0.5532	0.9582	0.0137		0.0235	0.0508
num_l	-0.29810	0.07541	0.19606	0.40581	1.00000	0.93152
num_l	0.1033	0.6868	0.2905	0.0235		<.0001
vol_l	-0.38682	0.14260	0.13219	0.35395	0.93152	1.00000
vol_l	0.0316	0.4441	0.4784	0.0508	<.0001	

Strong correlation (r=0.93152) between "total number of lesions" and "total volume of lesions" from the above SAS output (Table 1) indicates the possibility of multicollinearity between these two variables.

This can also be checked by calculating variance inflation factor and eigenvalues. To calculate variance inflation factor and eigenvalues, PROC REG procedure with VIF and COLLIN option has to be used.

```
proc reg data=one;
model dcng=age dur nr_pse dscore num_l vol_l/VIF TOL COLLIN;
run;
```

The PROC REG procedure gives the following VIF and eigenvalues:
Table 2:

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	Intercept	1	108.97953	12.54286	8.69	<.0001	.	0
age	age	1	-0.25458	0.10044	-2.53	0.0182	0.67920	1.47231
dur	dur	1	-0.09890	0.05517	-1.79	0.0856	0.88137	1.13460
nr_pse	nr_pse	1	-2.46940	0.35866	-6.89	<.0001	0.70729	1.41385
dscore	dscore	1	-0.04471	0.06751	-0.66	0.5141	0.73413	1.36216
num_l	num_l	1	-0.42476	0.12104	-3.51	0.0018	0.12085	8.27501
vol_l	vol_l	1	0.33868	0.13884	2.44	0.0225	0.11495	8.69969

Collinearity Diagnostics

Number	Eigenvalue	Condition Index
1	6.94936	1.00000
2	0.01796	19.67336
3	0.01553	21.15034
4	0.00975	26.69145
5	0.00615	33.61704
6	0.00107	80.75820
7	0.00018093	195.98419

From Table 2, one can see that VIF values of num_l and vol_l are greater than 5 and eigenvalues of num_l and vol_l are 0.00107 and 0.00018093 which are very small (close to zero) and corresponding condition indices are large. This indicates that there is a multicollinearity between variables num_l and vol_l.

REMEDIAL MEASURES

Following remedial measures can be taken if presence of multicollinearity is detected.

- To drop one or several predictor variables in order to lessen the multicollinearity and thereby improve the estimation precision of the remaining regression coefficients.

- If none of the predictor variables can be dropped, alternative methods of estimation should be considered such as

    1. Ridge Regression
    2. Principle Component Regression.

RIDGE REGRESSION:

Ridge regression provides alternative estimation method that may be used where multicollinearity is suspected. Multicollinearity leads to small characteristic roots and when one or more of the λ's are small, the total mean square error of $\hat{\beta}$ is large suggesting imprecision in the least squares estimation method. The ridge regression approach is an attempt to construct an alternative estimator that has a smaller total mean square error value.

There are number of ways to define and compute ridge estimates. The method associated with ridge trace is explained below.

Consider the regression model:

$Y = X\beta + \epsilon$;

Where Y is an n x 1 vector of responses, X is an n x p matrix of the regressor variables, β is a p x 1 vector of unknown constants, and ϵ is an n x 1 vector of random errors, with $\epsilon_i \sim NID(0, \sigma^2)$

The least squares estimator for β is

$$\hat{\beta} = (X'X)^{-1}X'Y$$

It can be shown that

$$E[(\hat{\beta}-\beta)'(\hat{\beta}-\beta)] = \sigma^2 \sum 1/\lambda_j \quad , \; j=1 \text{ to } p \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(9)$$

Where $\lambda_1 \geq \lambda_2 \ldots\ldots \geq \lambda_p$ are the characteristic roots of X'X. The left hand side of the eq. (9) is called the total mean square error. It serves as a composite measure of the squared distance of the estimated regression coefficients from their true values.

Hoerl and Kennard (1970) suggest a class of estimators indexed by a parameter k > 0.

The estimator is (for a given value of k)

$$\hat{\beta}(k) = (X'X+kI)^{-1} X'Y = (X'X+kI)^{-1} X'X\,\hat{\beta}$$

The expected value of $\hat{\beta}(k)$ is

$$E[\hat{\beta}(k)] = (X'X+kI)^{-1} X'X\beta$$

And the variance-covariance matrix is
$\mathrm{Var}[\hat{\beta}(k)] = (X'X+kI)^{-1} X'X (X'X+kI)^{-1}\sigma^2$

The residual sum of squares can be written as
$(Y-X\hat{\beta}(k))' (Y-X\hat{\beta}(k)) = (Y-X\hat{\beta})' (Y-X\hat{\beta}) + (\hat{\beta}(k) - \hat{\beta})'X'X(\hat{\beta}(k) - \hat{\beta})$

The total mean square error is

$$E[(\hat{\beta}(k) - \beta)' (\hat{\beta}(k) - \beta)] = \sigma^2 \, \mathrm{trace}\, [(X'X+kI)^{-1}X'X (X'X+kI)^{-1}] + k^2\beta' (X'X+kI)^{-2} \beta$$
$$= \sigma^2 \sum_{j=1}^{p} \lambda_j (\lambda_j + k)^{-2} + k^2 \beta' (X'X+kI)^{-2} \beta \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(10)$$

The first term on the right-hand side of eq. (10) is the sum of the variances of the components of $\hat{\beta}(k)$ (total variance) and the second term is the square of the bias. For k>0, $\hat{\beta}(k)$ is biased and the bias increases with k. On the other hand, the total variance is a decreasing function of k. The idea of ridge regression is to pick a value of k for which the reduction in total variance is not exceeded by the increase in bias.

Hoerl and Kennard (1970) prove that there exists a value of k>0 such that

$$E[(\hat{\beta}(k) - \beta)' (\hat{\beta}(k) - \beta)] < E[(\hat{\beta} - \beta)' (\hat{\beta} - \beta)]$$

An appropriate value of k may be selected by observing the ridge trace for $\hat{\beta}(k)$. The ridge trace is a simultaneous plot of the components of $\hat{\beta}(k)$ against k, for k in the interval from zero to one. If multicollinearity is a serious problem, the ridge estimators will vary dramatically as k is slowly increased from zero. $\hat{\beta}(k)$ will eventually stabilize. The behaviour of $\hat{\beta}(k)$ as a function of k is easily observed from the ridge trace. The value of k selected is the smallest value for which $\hat{\beta}(k)$ is stable. Generally this will produce a set of estimates with smaller mean squared error (MSE) than the least-square estimates.

The ridge estimates are stable in the sense that they are not affected by slight variations in the estimation data. Because of the smaller mean square error property, values of the ridge estimated coefficients are expected to be closer than ordinary least square estimates to the true values of estimated coefficients.

This can be seen by taking an example and solving it using ridge regression.

Consider the data from a trial in which relation between birth weight and following explanatory variables is to be found out.
Explanatory variables are:

- Skeletal size
- Fat
- Gestational age

PROC REG procedure with VIF and COLLIN option will check for multicollinearity and it will also give parameter estimates using ordinary least squares method.

```
proc reg data=two;
model birthwt=sksize fat gage/VIF TOL COLLIN;
run;
```

Table 3:

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	Intercept	1	-9.73849	1.16991	-8.32	<.0001	.	0
sksize	sksize	1	-0.02851	0.06475	-0.44	0.6730	0.00593	168.65281
fat	fat	1	0.57585	0.08938	6.44	0.0004	0.96986	1.03108
gage	gage	1	0.25234	0.09462	2.67	0.0322	0.00592	168.91193

From Table 3, it is very clear that there exists multicollinearity between “Skeletal size” and “Gestational age” and the parameter estimates of skeletal size, fat and gestational age using ordinary least squares method are

$\hat{\beta}_1$ = -0.02851, $\hat{\beta}_2$= 0.57585 and $\hat{\beta}_3$= 0.25234 respectively.

To apply ridge regression, PROC REG procedure with RIDGE option can be used and RIDGEPLOT option will give the graph of ridge trace.

```
proc reg data=two outvif
outest=b ridge=0 to 0.05 by 0.002;
model birthwt=sksize fat gage;
plot / ridgeplot nomodel nostat;
run;
proc print data=b;
run;
```

How to choose K: Things to look for

- Get the variance inflation factors (VIF) close to 1
- Estimated coefficients should be “stable”
- Look for only “modest” change in $R^{2.}$

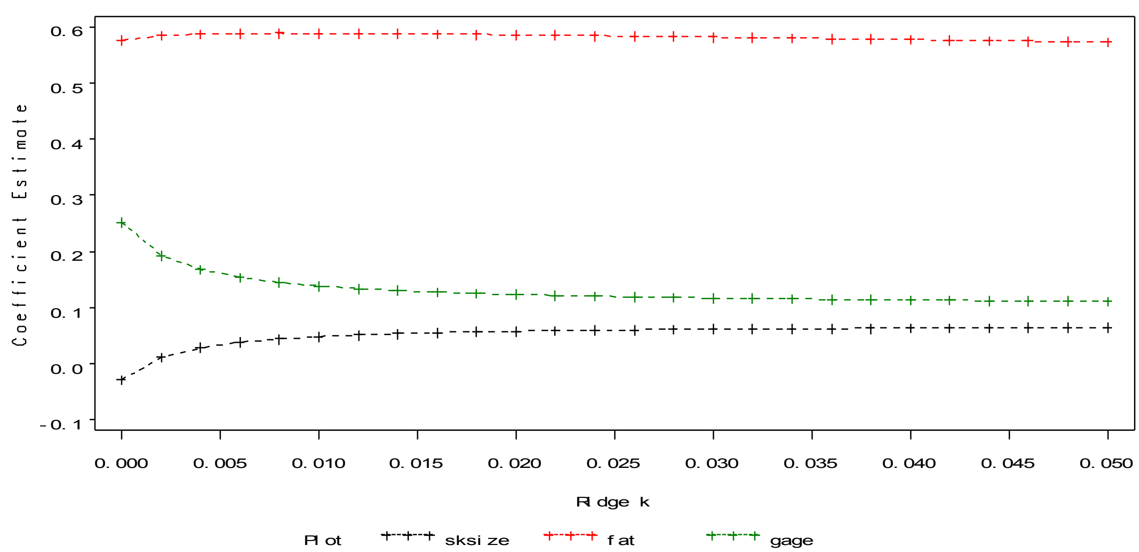
The variance inflation factors for regression coefficients and ridge trace are shown below:

Variance Inflation Factors & Parameter Estimates:
Table 4:

Obs	_MODEL_	_TYPE_	_DEPVAR_	_RIDGE_	_RMSE_	Intercept	sksize	fat	gage	birthwt
1	MODEL1	PARMS	birthwt	.	0.47508	-9.73849	-0.029	0.57585	0.252	-1
2	MODEL1	RIDGEVIF	birthwt	0.000	.	.	168.653	1.03108	168.912	-1
3	MODEL1	RIDGE	birthwt	0.000	0.47508	-9.73849	-0.029	0.57585	0.252	-1
4	MODEL1	RIDGEVIF	birthwt	0.002	.	.	60.333	1.00930	60.425	-1
5	MODEL1	RIDGE	birthwt	0.002	0.48819	-9.36350	0.012	0.58395	0.193	-1
6	MODEL1	RIDGEVIF	birthwt	0.004	.	.	30.787	1.00046	30.833	-1
7	MODEL1	RIDGE	birthwt	0.004	0.50143	-9.18575	0.029	0.58680	0.168	-1
8	MODEL1	RIDGEVIF	birthwt	0.006	.	.	18.684	0.99450	18.711	-1
9	MODEL1	RIDGE	birthwt	0.006	0.51065	-9.07331	0.038	0.58792	0.154	-1
10	MODEL1	RIDGEVIF	birthwt	0.008	.	.	12.573	0.98954	12.591	-1
11	MODEL1	RIDGE	birthwt	0.008	0.51724	-8.99047	0.044	0.58825	0.145	-1
12	MODEL1	RIDGEVIF	birthwt	0.010	.	.	9.064	0.98504	9.076	-1
13	MODEL1	RIDGE	birthwt	0.010	0.52219	-8.92358	0.049	0.58817	0.139	-1
14	MODEL1	RIDGEVIF	birthwt	0.012	.	.	6.865	0.98078	6.874	-1
15	MODEL1	RIDGE	birthwt	0.012	0.52608	-8.86626	0.052	0.58784	0.134	-1
16	MODEL1	RIDGEVIF	birthwt	0.014	.	.	5.397	0.97665	5.404	-1
17	MODEL1	RIDGE	birthwt	0.014	0.52926	-8.81515	0.054	0.58735	0.131	-1
18	MODEL1	RIDGEVIF	birthwt	0.016	.	.	4.368	0.97263	4.373	-1
19	MODEL1	RIDGE	birthwt	0.016	0.53194	-8.76831	0.056	0.58675	0.128	-1
20	MODEL1	RIDGEVIF	birthwt	0.018	.	.	3.619	0.96867	3.623	-1
21	MODEL1	RIDGE	birthwt	0.018	0.53428	-8.72453	0.057	0.58607	0.125	-1
22	MODEL1	RIDGEVIF	birthwt	0.020	.	.	3.057	0.96476	3.060	-1
23	MODEL1	RIDGE	birthwt	0.020	0.53637	-8.68304	0.058	0.58534	0.124	-1
24	MODEL1	RIDGEVIF	birthwt	0.022	.	.	2.624	0.96090	2.627	-1
25	MODEL1	RIDGE	birthwt	0.022	0.53828	-8.64329	0.059	0.58456	0.122	-1
26	MODEL1	RIDGEVIF	birthwt	0.024	.	.	2.284	0.95708	2.286	-1
27	MODEL1	RIDGE	birthwt	0.024	0.54005	-8.60492	0.060	0.58376	0.120	-1
28	MODEL1	RIDGEVIF	birthwt	0.026	.	.	2.012	0.95329	2.013	-1
29	MODEL1	RIDGE	birthwt	0.026	0.54173	-8.56765	0.061	0.58293	0.119	-1
30	MODEL1	RIDGEVIF	birthwt	0.028	.	.	1.790	0.94953	1.792	-1
31	MODEL1	RIDGE	birthwt	0.028	0.54333	-8.53128	0.061	0.58209	0.118	-1

32	MODEL1	RIDGEVIF	birthwt	0.030	.	.	1.608	0.94580	1.609	-1
33	MODEL1	RIDGE	birthwt	0.030	0.54489	-8.49565	0.062	0.58122	0.117	-1
34	MODEL1	RIDGEVIF	birthwt	0.032	.	.	1.456	0.94209	1.457	-1
35	MODEL1	RIDGE	birthwt	0.032	0.54641	-8.46066	0.062	0.58035	0.116	-1
36	MODEL1	RIDGEVIF	birthwt	0.034	.	.	1.328	0.93841	1.329	-1
37	MODEL1	RIDGE	birthwt	0.034	0.54791	-8.42621	0.063	0.57947	0.116	-1
38	MODEL1	RIDGEVIF	birthwt	0.036	.	.	1.219	0.93476	1.219	-1
39	MODEL1	RIDGE	birthwt	0.036	0.54939	-8.39222	0.063	0.57858	0.115	-1
40	MODEL1	RIDGEVIF	birthwt	0.038	.	.	1.126	0.93113	1.126	-1
41	MODEL1	RIDGE	birthwt	0.038	0.55088	-8.35864	0.063	0.57769	0.114	-1
42	MODEL1	RIDGEVIF	birthwt	0.040	.	.	1.045	0.92752	1.045	-1
43	MODEL1	RIDGE	birthwt	0.040	0.55237	-8.32541	0.064	0.57679	0.114	-1
44	MODEL1	RIDGEVIF	birthwt	0.042	.	.	0.974	0.92393	0.975	-1
45	MODEL1	RIDGE	birthwt	0.042	0.55386	-8.29250	0.064	0.57589	0.113	-1
46	MODEL1	RIDGEVIF	birthwt	0.044	.	.	0.913	0.92037	0.913	-1
47	MODEL1	RIDGE	birthwt	0.044	0.55537	-8.25988	0.064	0.57498	0.113	-1
48	MODEL1	RIDGEVIF	birthwt	0.046	.	.	0.859	0.91683	0.859	-1
49	MODEL1	RIDGE	birthwt	0.046	0.55690	-8.22752	0.064	0.57407	0.112	-1
50	MODEL1	RIDGEVIF	birthwt	0.048	.	.	0.811	0.91331	0.810	-1
51	MODEL1	RIDGE	birthwt	0.048	0.55844	-8.19539	0.064	0.57317	0.112	-1
52	MODEL1	RIDGEVIF	birthwt	0.050	.	.	0.768	0.90981	0.768	-1

Ridge Trace:



From table 4, we can see that at RIDGE = 0.040, the VIF's are close to 1 and root mean squared error (RMSE) has only increased from 0.47508 to 0.55237. New parameter estimates of skeletal size, fat and gestational age are $\hat{\beta}_1$ = 0.064, $\hat{\beta}_2$= 0.57679 and $\hat{\beta}_3$= 0.114 respectively. The improper negative sign on the estimate of $\hat{\beta}_1$ has disappeared and parameter estimates are better interpretable.

PRINCIPAL COMPONENT REGRESSION:

Principal component regression provides a unified way to handle multicollinearity which requires some calculations that are not usually included in standard regression analysis. The principle component analysis follows from the fact that every linear regression model can be restated in terms of a set of orthogonal explanatory variables. These new variables are obtained as linear combinations of the original explanatory variables. They are referred to as the principal components.

Consider the following model,

$$Y = X\beta + \epsilon;$$

Where Y is an n x 1 vector of responses, X is an n x p matrix of the regressor variables, β is a p x 1 vector of unknown constants, and ϵ is an n x 1 vector of random errors, with $\epsilon_i \sim NID(0, \sigma^2)$

There exists a matrix C, satisfying

$C'(X'X)C = \Lambda$ and $C'C=CC' = I$

Where Λ is a diagonal matrix with the ordered characteristic roots of X'X on the diagonal. The characteristic roots are denoted by $\lambda_1 \geq \lambda_2, \ldots, \geq \lambda_p$. C may be used to calculate a new set of explanatory variables, namely

$$(W_{(1)}, W_{(2)}, \ldots, W_{(p)}) = W = XC = (X_{(1)}, X_{(2)}, \ldots, X_{(p)})\, C$$

that are linear functions of the original explanatory variables. The W's are referred to as principal components.

Thus the regression model can be restated in terms of the principal components as

$$Y=W\alpha + \epsilon \text{, where } W=XC,\ \alpha = C'\beta$$

$$W'W = C'X'XC = C'C\ \Lambda\ C'C$$

The least square estimator of α is

$$\hat{\alpha} = (W'W)^{-1}W'Y = \Lambda^{-1}W'Y$$

and the variance-covariance matrix of $\hat{\alpha}$ is

$$V(\hat{\alpha}) = \sigma^2(W'W)^{-1} = \sigma^2\, \Lambda^{-1}$$

Thus a small eigenvalue of X'X implies that the variance of the corresponding regression coefficient will be large.

Since $W'W = C'X'XC = C'C\ \Lambda\ C'C = \Lambda$, we often refer to the eigenvalues λ_j as the variance of the jth principal component. If all λ_j equal to unity, the original regressors are orthogonal, while if a λ_j is exactly equal to zero, then it implies a perfect linear relationship between the original regressors. One or more λ_j near to zero implies that multicollinearity is present.

The principal component regression approach combats multicollinearity by using less than the full set of principal components in the model. To obtain the principal components estimators, assume that the regressors are arranged in order of decreasing eigenvalues, $\lambda_1 \geq \lambda_2 \ldots \geq \lambda_p > 0$. In principal components regression the principal components corresponding to near zero eigenvalues are removed from the analysis and least squares applied to the remaining components.

Consider the same example of birth weight. It can also be solved using principal component regression.

To apply principal component regression, PROC PRINCOMP procedure can be used.

```
PROC PRINCOMP DATA=one OUT=Result_1 N=3 PREFIX=Z OUTSTAT =Result_2;
VAR sksize fat gage;
RUN;
```

SAS output will provide the Correlation matrix, eigen values and eigenvectors.

Table 5:

Correlation Matrix

		sksize	fat	gage
sksize	sksize	1.0000	0.0538	0.9970
fat	fat	0.0538	1.0000	0.0665
gage	gage	0.9970	0.0665	1.0000

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	2.00415882	1.01128420	0.6681	0.6681
2	0.99287461	0.98990805	0.3310	0.9990
3	0.00296657		0.0010	1.0000

Eigenvectors

		Z1	Z2	Z3
sksize	sksize	0.704315	-.066090	0.706805
fat	fat	0.084416	0.996390	0.009050
gage	gage	0.704851	-.053292	-.707351

So, the principal components of the standardized explanatory variables are
Z1 = 0.7043 sksize + 0.0844 fat + 0.7085 gage
Z2 = -0.0660 sksize + 0.9964 fat – 0.0533 gage
Z3 = 0.7068 sksize + 0.0090 fat – 0.7074 gage

The Z's have sample variances λ_1 = 2.00415882, λ_2 = 0.99287461, λ_3 = 0.00296657

The model can be written in the form of principal components as

Birthwt = $\alpha_1 Z_1 + \alpha_2 Z_2 + \alpha_3 Z_3 + \epsilon$ ....(11)

The principal components technique can be used to reduce multicollinearity in the estimation data. The reduction is accomplished by using less than the full set of principal components to explain the variation in the response variable. When all three principal components are used, the ordinary least squares solution can be reproduced.

Since Z_3 has variance 0.0029, the linear function defining Z_3 is approximately equal to zero and is the source of multicollinearity in the data. We can exclude Z_3 and consider regression of birthwt against Z_1 and Z_2.

Thus, Birthwt = $\alpha_1 Z_1 + \alpha_2 Z_2 + \epsilon$ ....(12)

Estimated values of α's can be obtained using eq. (12). i.e. by regressing birthwt against Z_1 and Z_2

```
PROC reg DATA=Result_1;
model birthwt= Z1 Z2/ VIF;
RUN;
```

Table 6:

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	21.89091	0.15535	140.92	<.0001	0
Z1		1	3.14802	0.11509	27.35	<.0001	1.00000
Z2		1	0.75853	0.16351	4.64	0.0017	1.00000

Thus, selecting a model based on first two principal components Z_1 and Z_2 will remove the multicollinearity and hence will produce better results.

CONCLUSION

When multicollinearity is present in the data, ordinary least square estimators are imprecisely estimated. If the goal is to understand how the various X variables impact Y, then multicollinearity is a big problem. Thus, it is very essential to detect and solve the issue of multicolleariy before estimating the parameters based on fitted regression model.

We have seen various methods for detection of multicollinearity like examination of correlation matrix, calculating the variance inflation factor (VIF), by the eigensystem analysis and we have also seen that the remedial measures such as respecification of model, ridge regression, reduction technique like principal component analysis help to solve the problem of multicollinearity.

REFERENCES

Draper, N. R., Smith, H. (2003). Applied regression analysis, 3rd edition, Wiley, New York.

Montgomery, D. C., Peck, E. A., Vining, G. G. (2001). Introduction to linear regression analysis, 3rd edition, Wiley, New York.

Chatterjee, S., Price, B. Regression Analysis by Example, 3rd edition

CONTACT INFORMATION

Your comments and questions are valued and encouraged.

Contact the author at:

Harshada Joshi,
Senior Statistician
Cytel Statistical Software & Services Pvt. Ltd.
S.No.150,Lohia-Jain IT Park,
"A" Wing, 6th Floor, Paud Road,
Kothrud, Pune,
India 411038
Email Id : harshada.joshi@cytel.com