

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

Harshada Joshi

Session SP07 - PhUSE 2012

Agenda

- **What is Multicollinearity?**
- **How to Detect Multicollinearity?**
 - Examination of Correlation Matrix**
 - Variance Inflation Factor**
 - Eigensystem Analysis of Correlation Matrix**
- **Remedial Measures**
 - Ridge Regression**
 - Principal Component Regression**

What is Multicollinearity?

- **Multicollinearity is a statistical phenomenon in which there exists a perfect or exact relationship between the predictor variables.**
- **When there is a perfect or exact relationship between the predictor variables, it is difficult to come up with reliable estimates of their individual coefficients.**
- **It will result in incorrect conclusions about the relationship between outcome variable and predictor variables.**

What is Multicollinearity?

Recall that the multiple linear regression is

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

Where,

Y is an $n \times 1$ vector of responses,

X is an $n \times p$ matrix of the predictor variables,

β is a $p \times 1$ vector of unknown constants,

ϵ is an $n \times 1$ vector of random errors,

with $\epsilon_i \sim \text{NID}(0, \sigma^2)$

What is Multicollinearity?

- Multicollinearity inflates the variances of the parameter estimates and hence this may lead to lack of statistical significance of individual predictor variables even though the overall model may be significant.
- The presence of multicollinearity can cause serious problems with the estimation of β and the interpretation.

How to detect Multicollinearity?

1. **Examination of Correlation Matrix**
2. **Variance Inflation Factor (VIF)**
3. **Eigensystem Analysis of Correlation Matrix**

How to detect Multicollinearity?

1. Examination of Correlation Matrix:

- Large correlation coefficients in the correlation matrix of predictor variables indicate multicollinearity.
- If there is a multicollinearity between any two predictor variables, then the correlation coefficient between these two variables will be near to unity.

How to detect Multicollinearity?

2. Variance Inflation Factor:

- 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 X_j is regressed on all other predictor variables in the model.

$$\text{Let } VIF_j = 1 / (1 - R_j^2), \quad \text{for } j = 1, 2, \dots, p-1$$

- $VIF_j = 1$ when $R_j^2 = 0$, i.e. when j^{th} variable is not linearly related to the other predictor variables.
- $VIF_j \rightarrow \infty$ when $R_j^2 \rightarrow 1$, i.e. when j^{th} variable is linearly related to the other predictor variables.

How to detect Multicollinearity?

- The VIF is an index which measures how much variance of an estimated regression coefficient is increased because of multicollinearity.
- Rule of Thumb: If any of the VIF values exceeds 5 or 10, it implies that the associated regression coefficients are poorly estimated because of multicollinearity (Montgomery, 2001).

How to detect Multicollinearity?

3. Eigensystem Analysis of Correlation Matrix:

- The eigenvalues can also be used to measure the presence of multicollinearity.
- If multicollinearity is present in the predictor variables, one or more of the eigenvalues will be small (near to zero).
- Let $\lambda_1, \dots, \lambda_p$ be the eigenvalues of correlation matrix. The condition number of correlation matrix is defined as
$$K = \sqrt{(\lambda_{\max} / \lambda_{\min})}$$
 &
$$K_j = \sqrt{(\lambda_{\max} / \lambda_j)}, j=1,2,\dots,p.$$

How to detect Multicollinearity?

- **Rule of Thumb:** If one or more of the eigenvalues are small (close to zero) and the corresponding condition number is large, then it indicates multicollinearity (Montgomery, 2001).

Example

- **Primary endpoint is change in the disability score in patients with a neurodegenerative disease.**
- **Aim: To find the relation between change in the disability score and the following baseline 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

By examining the correlation matrix

```
proc corr data=one SPEARMAN;
var age dur nr_pse dscore num_1 vol_1;
run;
```

Spearman Correlation Coefficients						
Prob > r under H0: Rho=0						
	age	dur	nr_pse	dscore	num_1	vol_1
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				0.14260
dur		0.3853				0.4441
nr_pse	0.18276	0.04097	1.00000			0.13219
nr_pse		0.3251	0.8268			0.4784
dscore	-0.11073	0.00981	0.43824	1.00000		0.35395
dscore		0.5532	0.0137			0.0508
num_1	-0.29810	0.07541	0.19606	0.40581	1.00000	0.93152
num_1		0.1033	0.2905	0.0235		<.0001
vol_1	-0.38682	0.14260	0.13219	0.35395	0.93152	1.00000
vol_1		0.0316	0.4441	0.0508	<.0001	

**Strong Correlation
Multicollinearity**

0.93152
<.0001

Variance Inflation Factor & Eigenvalues

This can also be checked by calculating variance inflation factor and eigenvalues.

To calculate variance inflation factor and eigenvalues, we can use PROC REG procedure with VIF and COLLIN option.

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

Variance Inflation Factor & Eigenvalues

Variable	Label	DF	Parameter Estimate	Standard Error	t Statistic	Pr > t	Tolerance	Variance Inflation
Intercept	Intercept	1	108.9	0.00000				
age	age	1	-0.25458	0.10044	-2.53	0.0182	0.67920	1.47231
dur	dur	1	-0.09890	0.05517	-1.79	0.08137	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	0.73421	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

Number	Eigenvalue	Condition Index
1	6.94936	1.00000
2	0.01796	19.67336
3	0.01553	21.15034
4	0.000075	26.69145
5	0.00615	33.61704
6	0.00107	80.75820
7	0.00018093	195.98419

VIF for num_l = 8.27501
VIF for vol_l = 8.69969

Eigenvalues
num_l = 0.0010
Vol_l = 0.00018

Remedial Measures

- To drop one or several predictor variables in order to lessen the multicollinearity.
- If none of the predictor variables can be dropped, alternative methods of estimation are
 1. Ridge Regression
 2. Principal Component Regression

Ridge Regression

- Ridge regression provides an alternative estimation method that can be used where multicollinearity is suspected.
- Multicollinearity leads to small characteristic roots and when characteristic roots are small, the total mean square error of $\hat{\beta}$ is large which implies an imprecision in the least squares estimation method.
- Ridge regression gives an alternative estimator (k) that has a smaller total mean square error value.

Ridge Regression

- The value of k can be estimated by looking at the ridge trace plot.
- Ridge trace plot is a plot of parameter estimates vs k where k usually lies in the interval of $[0,1]$.
- Rule of Thumb:
 1. Pick the smallest value of k that produces a stable estimate of β .
 2. Get the variance inflation factors (VIF) close to 1.

Example

- 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

Example

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

SAS Output :

Parameter Estimates								
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Tolerance	Variance Inflation
Intercept	Intercept	1	-9.72849	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

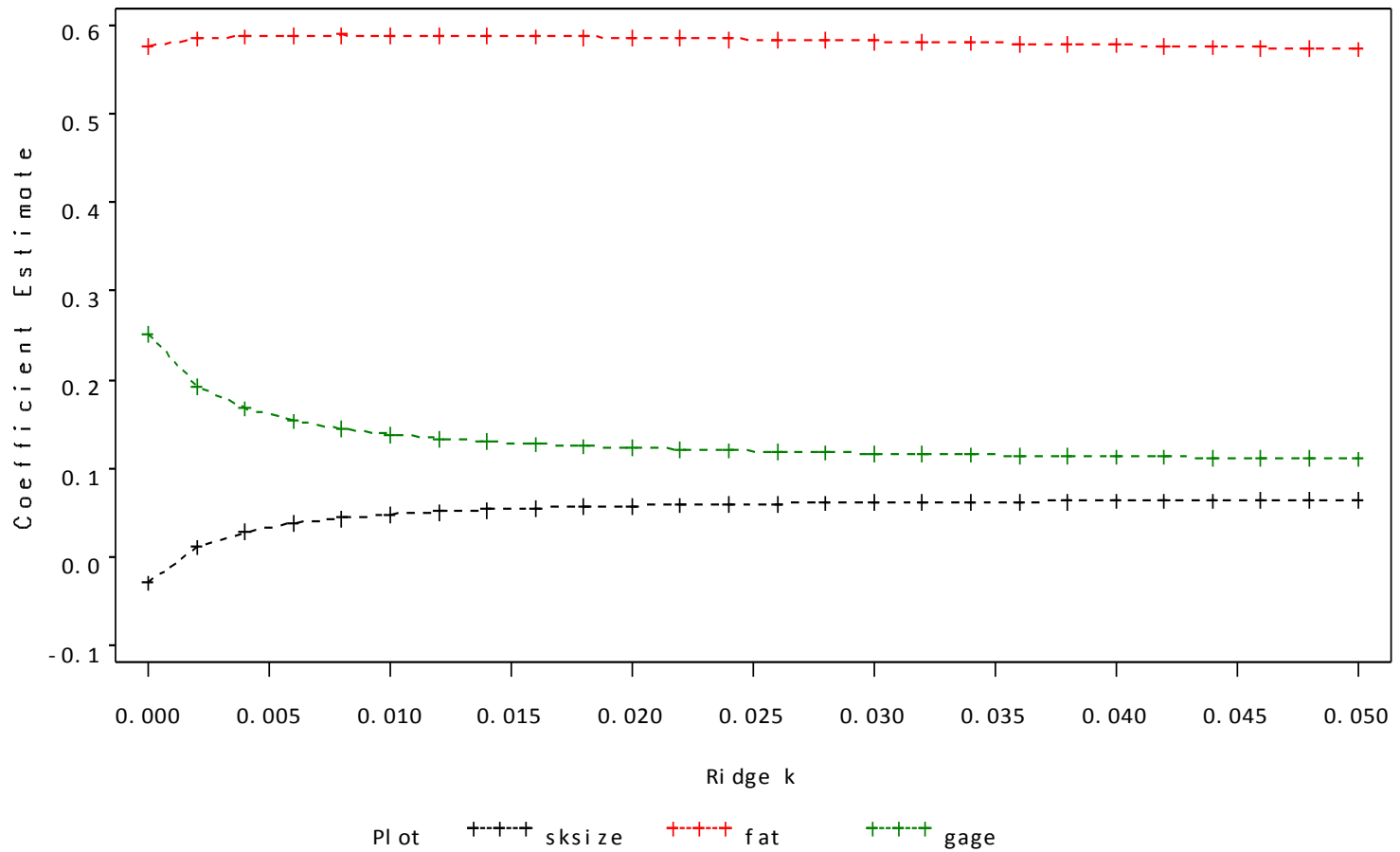
Example

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;
```

Example

Ridge Trace:



Variance Inflation Factors & Parameter Estimates:

Obs	_MODEL_	_TYPE_	_DEPVAR_	_RIDGE_	_RMSE_	Intercept	sksize	fat	gage
1	MODEL1	PARMS	birthwt	.	0.47508	-9.73849	-0.029	0.57585	0.252
2	MODEL1	RIDGEVIF	birthwt	0.000	.	.	168.653	1.03108	168.912
3	MODEL1	RIDGE	birthwt	0.000	0.47508	-9.73849	-0.029	0.57585	0.252
						.	60.333	1.00930	60.425
						19 -9.36350	0.012	0.58395	0.193
						.	30.787	1.00046	30.833
						43 -9.18575	0.029	0.58680	0.168
						.	.	.	.
						.	.	.	.
						.	.	.	.
						.	.	.	.
						.	.	.	.
						.	.	.	.
						.	.	.	.
						.	.	.	.
38	MODEL1	RIDGEVIF	birthwt	.	.	.	1.219	0.93476	1.219
39	MODEL1	RIDGE	birthwt	.	0.54939	-8.39222	0.063	0.57858	0.115
40	MODEL1	RIDGEVIF	birthwt	0.038	.	.	1.126	0.93113	1.126
41	MODEL1	RIDGE	birthwt	0.038	0.55337	-8.25988	0.064	0.57498	0.113
42	MODEL1	RIDGEVIF	birthwt	0.040	.	.	1.045	0.92752	1.045
43	MODEL1	RIDGE	birthwt	0.040	0.55237	-8.32541	0.064	0.57679	0.114
44	MODEL1	RIDGEVIF	birthwt	0.042	.	.	0.913	0.92037	0.913
45	MODEL1	RIDGE	birthwt	0.042	0.55386	-8.29250	0.064	0.57589	0.113
46	MODEL1	RIDGEVIF	birthwt	0.044	.	.	0.913	0.92037	0.913
47	MODEL1	RIDGE	birthwt	0.044	0.55537	-8.25988	0.064	0.57498	0.113
48	MODEL1	RIDGEVIF	birthwt	0.046	.	.	0.859	0.91683	0.859
49	MODEL1	RIDGE	birthwt	0.046	0.55690	-8.22752	0.064	0.57407	0.112
50	MODEL1	RIDGEVIF	birthwt	0.048	.	.	0.811	0.91331	0.810
51	MODEL1	RIDGE	birthwt	0.048	0.55844	-8.19539	0.064	0.57317	0.112
52	MODEL1	RIDGEVIF	birthwt	0.050	.	.	0.768	0.90981	0.768

At K=0.04,
Parameter estimate for
sksize = 0.064
fat = 0.67679
gage = 0.114

Principal Component Regression

- Every linear regression model can be restated in terms of a set of orthogonal explanatory variables.
- These new variables are obtained as a linear combinations of the original explanatory variables. They are referred to as the principal components.
- The principal component regression approach combats multicollinearity by using less than the full set of principal components in the model.

Principal Component Regression

- To obtain the principal components estimators, assume that the regressors are arranged in order of decreasing eigenvalues, $\lambda_1 \geq \lambda_2 \dots \geq \lambda_p > 0$.
- 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.

Principal Component Regression

- Consider the example of birth weight. It can also be solved by 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;
```

Correlation matrix, Eigen values, Eigen vectors

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

$$Z_1 = 0.7043 \text{ sksize} + 0.0844 \text{ fat} + 0.7085 \text{ gage}$$

$$Z_2 = -0.0660 \text{ sksize} + 0.9964 \text{ fat} - 0.0533 \text{ gage}$$

$$Z_3 = 0.7068 \text{ sksize} + 0.0090 \text{ fat} - 0.7074 \text{ gage}$$

Principal Component Regression

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

$$\text{Birthwt} = \alpha_1 Z_1 + \alpha_2 Z_2 + \alpha_3 Z_3 + \epsilon$$

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

Since Z_3 has variance 0.0029, the linear function defining Z_3 is approximately equal to zero.

We can exclude Z_3 and consider regression of birthwt against Z_1 and Z_2 .

Principal Component Regression

Thus,

$$\text{Birthwt} = \alpha_1 Z_1 + \alpha_2 Z_2 + \epsilon$$

Estimated values of α 's can be obtained by regressing birthwt against Z_1 and Z_2

```
PROC REG DATA=Result_1;  
MODEL birthwt= Z1 Z2/ VIF;  
RUN;
```

Principal Component Regression

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.

Summary

- When multicollinearity is present in the data, ordinary least square estimators are imprecisely estimated.
- If 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 multicollinearity before estimating the parameters based on fitted regression model.

Summary

- **Detection of multicollinearity can be done by examining the correlation matrix or by using VIF and eigenvalues.**
- **Remedial measures such as ridge regression using PROC REG with ridge option and principal component regression using PROC PRINCOMP help to solve the problem of multicollinearity.**

Conclusion

- **Detection of multicollinearity is very important.**
- **Once multicollinearity is detected, it is necessary to introduce appropriate changes in the model specification.**
- **Improper model specification may result in misleading and improper conclusions.**
- **Remedial measures such as ridge regression and principal component regression help to solve the problem of multicollinearity.**

Questions ??

Thank You!

harshada.joshi@cytel.com