MWSUG – PO05

Multicollinearity: What Is It and What Can We Do About It?

Deanna Naomi Schreiber-Gregory, MS, National University

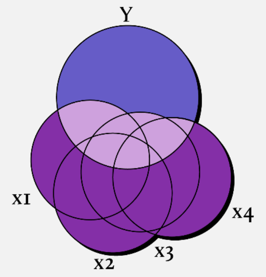
Definition

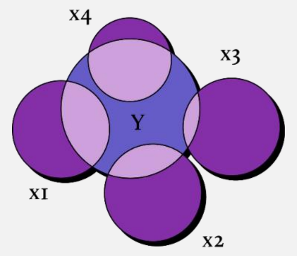
Definition
- A statistical phenomenon wherein there exists a perfect or exact relationship between predictor variables

From a Conventional Standpoint
- Occurs in regression when several predictors are high correlated
- Linear Dependence: Fit well into a straight line that passes through many data points
- Another way to look at collinearity is co-dependence

Consequence
- Creates difficulty in creating reliable estimates of individual coefficients for the predictor variables
- Results in incorrect conclusions about the relationship between outcome and predictor variables
- As degree of multicollinearity increases, regression model estimates of the coefficients become unstable

Consequence of Variance Inflation
- Multicollinearity inflates the variances of the parameter estimates
- Look at R-square = higher the value, better the model
- Collinearity results in inflation of variance, standard error, and parameter estimates
- Can lead you to an over-specified model
- Include predictor variables with low statistical significance
- The presence of multicollinearity can cause serious problems with the estimation of B and its interpretation





Explanatory vs Predictive Models
- Collinearity is a problem when a model's purpose is explanation and not prediction
- More difficult to achieve significance of collinear parameters
- Note: if estimates are statistically significant, they are as reliable as any other variable in the model
- If they are not significant, the sum of the coefficient is likely to be reliable
- In the case of a predictive model: just need to increase sample size
- In the case of an explanatory model: further measures are needed

Detection

Examination of the Correlation Matrix
- Large correlation coefficients in the correlation matrix of predictor variables indicate multicollinearity
- If there is multicollinearity between any two predictor variables, then the correlation coefficient between those two variables will be near to unity (1.0000)
- PROC CORR

Variance Inflation Factor
- Quantifies the severity of multicollinearity in an ordinary least-squares regression analysis
- Consider equation: VIFj= 1/(1-Rj^2), for j= 1,2,…..p-1
- Let Rj^2 denote the coefficient of determination when Xj is regressed on all other predictor variables in the model
  - VIFj= 1 when RJ^2 = 0 ……. When jth variable is not linearly related to the other variables
  - VIFj-> ∞ when Rj^2 -> 1 …….. When jth variable is linearly related to the other predictor variables
- The VIF is an index which measures how much an estimated regression coefficient's variance is increased due to multicollinearity
  - Example:
    - VIF for Xj is 5
    - Variance of estimated Bj is 5 times larger than if Xj was uncorrelated with other predictors
- Note: 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)

Tolerance
- Another way of looking at Variance Inflation Factor
- Represented by 1/VIF

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 λ1…….λp be the eigenvalues of correlation matrix. The condition number of correlation matrix is defined as follows: K = sqrt(λmax / λmin) &
  - Condition indices of correlation matrix are defined as: Kj= sqrt(λmax / λj), j=1,2,….,p
- Note: 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)

Control

Ways to Control for Multicollinearity
- Easiest to just 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 need to be employed:
  - Ridge Regression or Principal Component Regression
- For regression models with interactive terms, quadratic terms, or cubic terms:
  - Centered-score regression or Orthogonalization

Ridge Regression
- Logic: Multicollinearity leads to small characteristic roots
  - When characteristic roots are small, the total mean square error of 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
- Result:
  - Allows for better interpretation of regression coefficients by imposing some bias on regression coefficients and shrinking their variances
  - Consider Factor analysis: replaces inter-correlated predictors with principal components
- Calculation
  - The value of k can be estimated by looking at a ridge trace plot
  - Ridge trace plots are plots of parameter estimates vs k where k usually lies in the interval [0,1]
  - Pick the smallest value of k that produces a stable estimate of β
  - Get the variance inflation factors (VIF) close to 1
  - Want a "modest" change in R-square

Principal Component Regression
- Logic: Every linear regression model can be restated in terms of a set of orthogonal explanatory variables
  - New variables are obtained as linear combinations of the original explanatory variables: Principal Components
  - Uses less than the full set of principal components in the model
- Calculation:
  - Assume the regressor are arranged in order of decreasing eigenvalues, λ1 ≥ λ2 ………. ≥ λ p > 0
  - The principal components corresponding to near zero eigenvalues are removed from the analysis
  - Least squares is then applied to the remaining components

```
/* Examination of the Correlation Matrix */
Proc corr data=temp;
 Var hypertension aspirin hicholesterol anginachd smokingstatus obese_BMI exercise _AGE_G sex alcoholbinge; Run;

/* Multicollinearity Investigation: VIF TOL COLLIN */
Proc reg data=temp;
 Model stroke = hypertension aspirin hicholesterol anginachd smokingstatus obese_BMI exercise _AGE_G sex alcoholbinge / vif tol collin;
Run;     Quit;
```

Pearson Correlation Coefficients, N = 36345
Prob > |r| under H0: Rho=0

	hypertension	aspirin	hicholesterol	anginachd	smokingstatus	obese_BMI	exercise	_AGE_G	SEX	alcoholbinge
hypertension	1.00000	0.08742 <.0001	0.21641 <.0001	0.14202 <.0001	0.00978 0.0622	0.17714 <.0001	-0.09920 <.0001	0.15519 <.0001	-0.00693 0.1867	-0.00014 0.9782
aspirin	0.08742 <.0001	1.00000	0.07538 <.0001	0.04244 <.0001	0.00902 0.0853	0.02274 <.0001	0.01524 0.0037	0.08680 <.0001	0.00469 0.3716	-0.00798 0.1281
hicholesterol	0.21641 <.0001	0.07538 <.0001	1.00000	0.16461 <.0001	0.04842 <.0001	0.08058 <.0001	-0.03651 <.0001	0.10100 <.0001	-0.00336 0.5223	-0.00570 0.2774
anginachd	0.14202 <.0001	0.04244 <.0001	0.16461 <.0001	1.00000	0.08779 <.0001	0.03634 <.0001	-0.06577 <.0001	0.09145 <.0001	-0.08674 <.0001	-0.02660 <.0001
smokingstatus	0.00978 0.0622	0.00902 0.0853	0.04842 <.0001	0.08779 <.0001	1.00000	-0.03687 <.0001	-0.09587 <.0001	-0.07621 <.0001	-0.09950 <.0001	0.10007 <.0001
obese_BMI	0.17714 <.0001	0.02274 <.0001	0.08058 <.0001	0.03634 <.0001	-0.03687 <.0001	1.00000	-0.07686 <.0001	-0.07462 <.0001	-0.13354 <.0001	-0.03241 <.0001
exercise	-0.09920 <.0001	0.01524 0.0037	-0.03651 <.0001	-0.06577 <.0001	-0.09587 <.0001	-0.07686 <.0001	1.00000	-0.01925 0.0002	-0.06550 <.0001	0.02262 <.0001
_AGE_G IMPUTED AGE IN SIX GROUPS	0.15519 <.0001	0.08680 <.0001	0.10100 <.0001	0.09145 <.0001	-0.07621 <.0001	-0.07462 <.0001	-0.01925 0.0002	1.00000	0.05388 <.0001	-0.02970 <.0001
SEX RESPONDENTS SEX	-0.00693 0.1867	0.00469 0.3716	-0.00336 0.5223	-0.08674 <.0001	-0.09950 <.0001	-0.13354 <.0001	-0.06550 <.0001	0.05388 <.0001	1.00000	-0.02312 <.0001
alcoholbinge	-0.00014 0.9782	-0.00798 0.1281	-0.00570 0.2774	-0.02660 <.0001	0.10007 <.0001	-0.03241 <.0001	0.02262 <.0001	-0.02970 <.0001	-0.02312 <.0001	1.00000

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	Intercept	1	-0.01888	0.01146	-1.65	0.0993	.	0
hypertension		1	0.03944	0.00328	12.03	<.0001	0.88842	1.12559
aspirin		1	0.05142	0.00347	14.83	<.0001	0.98259	1.01772
hicholesterol		1	0.01179	0.00314	3.76	0.0002	0.92484	1.08127
anginachd		1	0.07910	0.00422	18.74	<.0001	0.93978	1.06408
smokingstatus		1	0.01990	0.00214	9.30	<.0001	0.95168	1.05078
obese_BMI		1	-0.01431	0.00341	-4.20	<.0001	0.92887	1.07658
exercise		1	-0.03434	0.00329	-10.44	<.0001	0.96555	1.03568
_AGE_G	IMPUTED AGE IN SIX GROUPS	1	0.00407	0.00169	2.40	0.0162	0.94089	1.06283
SEX	RESPONDENTS SEX	1	0.01690	0.00303	5.58	<.0001	0.95513	1.04698
alcoholbinge		1	-0.03391	0.00680	-4.99	<.0001	0.98622	1.01397

Number	Eigenvalue	Condition Index
1	7.26674	1.00000
2	0.96463	2.74467
3	0.82476	2.96829
4	0.51415	3.75945
5	0.38421	4.34895
6	0.31447	4.80710
7	0.25041	5.38694
8	0.24042	5.49773
9	0.17624	6.42124
10	0.05282	11.72959
11	0.01115	25.52670

Ridge Regression

```
PROC REG DATA=stroke;
MODEL stroke = genhealth bp chol/ VIF TOL COLLIN;
RUN;
```

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	Intercept	1	-9.75748	1.13256	-8.32	<.0001	.	0
Genhealth	Genhealth	1	-0.02367	0.06768	-0.44	0.6730	0.00463	168.63567
Bp	Bp	1	0.57658	0.08595	6.44	0.0004	0.96456	1.02894
Chol	Chol	1	0.22789	0.09457	2.67	0.0322	0.00436	168.93865

```
PROC REG DATA=stroke OUTVIF;
  OUTEST=rrstrokeRIDGE=0 to 0.05 by 0.002;
  MODEL stroke = genhealth bp chol;
  PLOT / RIDGEPLOT NOMODEL NOSTAT;    RUN;
PROC PRINT DATA=rrstroke;    RUN;
```

Obs	_Model_	_Type_	_Depvar_	_Ridge_	_RMSE_	Intercept	Genhealth	Bp	chol
1	Model1	Parms	Stroke	.	0.47508	-9.73849	-0.029	0.57585	0.252
2	Model1	RidgeVIF	Stroke	0.000	.	.	168.653	1.03108	168.912
3	Model1	Ridge	Stroke	0.000	0.47508	-9.73849	-0.029	0.57585	0.252
..	..	..	..	..	..	..	..	..	..
33	Model1	Ridge	Stroke	0.038	0.55088	-8.35864	0.063	0.57769	0.114
34	Model1	RidgeVIF	Stroke	0.040	.	.	1.045	0.92752	1.045
35	Model1	Ridge	Stroke	0.040	0.55237	-8.32541	0.064	0.57679	0.114
36	Model1	RidgeVIF	Stroke	0.042	.	.	0.974	0.92393	0.975
37	Model1	Ridge	Stroke	0.042	0.55386	-8.29250	0.064	0.57589	0.113

Principal Component Regression

```
PROC PRINCOMP DATA=stroke
    OUT=result_1 N=3 PREFIX=z OUTSTAT=result_2;
    VAR genhealth bp chol;     RUN;
```

Correlation Matrix

		Genhealth	Bp	Chol
Genhealth	Genhealth	1.0000	0.0538	0.9970
Bp	Bp	0.0538	1.0000	0.0665
Chol	Chol	0.9970	0.0665	1.0000
Eigenvalues of the Correlation Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	2.0042	1.0113	0.6681	0.6681
2	0.9928	0.9899	0.3310	0.9990
3	0.0029		0.0010	1.0000
Eigenvectors				
		Z1	Z2	Z3
Genhealth	Genhealth	0.704315	-0.066090	0.706805
Bp	Bp	0.084416	0.996390	0.009050
Chol	chol	0.704851	-0.053292	-0.707351

```
PROC REG DATA=result_1
    MODEL stroke = z1 z2 / VIF;    RUN;
```

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	1.0000
Z2		1	0.75853	0.16351	4.64	0.0017	1.0000

MWSUG – PO05

Multicollinearity: What Is It and What Can We Do About It?

Deanna Naomi Schreiber-Gregory, MS, National University

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

Joshi, H., Kulkarni, H., Deshpande, S. (2012). *Multicollinearity Diagnostics in Statistical Modeling and Remedies to deal with it using SAS*. PhUSE2012.

Acknowledgements

BRFSS –CDC: For providing the dataset

National University: For encouraging personal research

Neuropsychiatric Research Institute: For providing a unique research opportunity

Contact Information

Name: **Deanna (DeDe) Naomi Schreiber-Gregory**

Organization: National University, Peace-Work

Location: Moorhead, MN

E-mail: d.n.schreibergregory@gmail.com

Twitter: https://twitter.com/DN_SchGregory

LinkedIn: https://www.linkedin.com/in/deanna-dedeschreiber-gregory-a54a7b66

Thank You!