PharmaSUG 2018 - Paper AA-01

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

Deanna N Schreiber-Gregory, Henry M Jackson Foundation for the Advancement of Military Medicine

ABSTRACT

Multicollinearity can be briefly described as the phenomenon in which two or more identified predictor variables in a multiple regression model are highly correlated. The presence of this phenomenon can have a negative impact on the analysis as a whole and can severely limit the conclusions of the research study. This paper reviews and provides examples of the different ways in which multicollinearity can affect a research project, and tells how to detect multicollinearity and how to reduce it once it is found. In order to demonstrate the effects of multicollinearity and how to combat it, this paper explores the proposed techniques by using the Youth Risk Behavior Surveillance System data set. This paper is intended for any level of SAS® user. This paper is also written to an audience with a background in behavioral science or statistics.

INTRODUCTION

Multicollinearity is often described as the statistical phenomenon wherein there exists a perfect or exact relationship between predictor variables. From a conventional standpoint, this occurs in regression when several predictors are highly correlated. (As a disclaimer, variables do not need to be highly correlated to be collinear, though this is usually the case.) Another way to think of collinearity is “co-dependence” of variables.

Why is this important? Well, when things are related, we say that they are linearly dependent. In other words, they fit well into a straight regression line that passes through many data points. In the incidence of multicollinearity, it is difficult to come up with reliable estimates of individual coefficients for the predictor variables in a model which results in incorrect conclusions about the relationship between outcome and predictor variables. Therefore, in the consideration of a multiple regression model in which a series of predictor variables were chosen in order to test their impact on the outcome variable, it is essential that multicollinearity not be present!

Another way to look at this issue is by considering a basic multiple linear regression equation:

$$y = x\beta + \varepsilon$$

Where y is an nx1 vector of response, x is an nxp matrix of predictor variables, β is a px1 vector of unknown constants, and ε is an nx1 vector of random errors with $\varepsilon_i \sim NID(0,\sigma^2)$. Considering this equation, consider the fact that multicollinearity tends to inflate the variances of the parameter estimates, which would lead to a lack of statistical significance of the individual predictor variables even though the overall model itself remains significant. Therefore, the presence of multicollinearity can end up causing serious problems when estimating and interpreting β.

Why should we care? Consider this example: Your company has just undergone a major overhaul and it was decided that each department lead should choose an assistant lead to help with their workload. The assistant leads were chosen by each department lead after a series of rigorous interviews and discussions with each applicant’s references. It is now time for next year’s budget to be decided. An administrative meeting is held during which both department leads and their new assistant leads are present. It comes time to vote, by show of hands, on a major budget revision. Both the leads and their assistants (of whom they are also supervisors) will be voting. Do you think any of the assistants will vote against their leads? Probably not. This will end up resulting in a biased vote as the votes of the assistants would be dependent on the votes of their leads. A relationship such as this between two variables in a

model could lead to an even more biased outcome, thus leading to results that have been affected in a detrimental way.

Collinearity is especially problematic when a model's purpose is explanation rather than prediction. In the case of explanation, it is more difficult for a model containing collinear variables to achieve significance of the different parameters. In the case of prediction, if the estimates end up being statistically significant, they are still only as reliable as any other variable in the model, and if they are not significant, then the sum of the coefficient is likely to be reliable. In summary if collinearity is found in a model testing prediction, then one need only increase the sample size of the model. However, if collinearity is found in a model seeking to explain, then more intense measures are needed. The primary concern resulting from multicollinearity is that as the degree of collinearity increases, the regression model estimates of the coefficients become unstable and the standard errors for the coefficients become wildly inflated.

DETECTING MULTICOLLINEARITY

This first section will explain the different diagnostic strategies for detecting multicollinearity in a dataset. While reviewing this section, the author would like you to think logically about the model being explored. Try identifying possible multicollinearity issues before reviewing the results of the diagnostic tests.

INTRODUCTION TO THE FIRST DATSET

The Youth Risk Behavior Surveillance System (YRBSS) was developed as a tool to help monitor priority risk behaviors that contribute substantially to death, disability, and social issues among American youth and young adults today. The YRBSS has been conducted biennially since 1991 and contains survey data from national, state, and local levels. The national Youth Risk Behavior Survey (YRBS) provides the public with data representative of the United States high school students. On the other hand, the state and local surveys provide data representative of high school students in states and school districts who also receive funding from the CDC through specified cooperative agreements. The YRBSS serves a number of different purposes. The system was originally designed to measure the prevalence of health-risk behaviors among high school students. It was also designed to assess whether these behaviors would increase, decrease, or stay the same over time. An additional purpose for the YRBSS is to have it examine the co-occurrence of different health-risk behaviors.

The particular study used in this paper examines the co-occurrence of suicidal ideation as an indicator of psychological unrest with other health-risk behaviors. The purpose of this study is to serve as an exercise in examining multicollinearity in a sensitive population through the examination of several health-risk behaviors and their link to suicidal ideation. The outcome variable of interest in this study was suicidal ideation and the predictor variables of interest were lifetime substance abuse participation, age of participant, gender of participant, race of participant, identification of depression within last year, recent substance abuse participation, being a victim of violence, and being an active participant in violence.
As a first step in the examination of the question being asked – do target health-risk behaviors contribute to thoughts of suicide in America's youth – we must first identify which datasets will be used in the analysis, what differences arise between the datasets, and how to address those differences. In short, we must clean the data for our analysis. Most of you know this already, but it is a worthy note to make considering the type of analysis we are about to conduct. The exact method to cleaning the data will not be covered in this section, for the sake of space and time, but the author would like to note that YRBS years 1991 – 2015 were cleaned and prepped for the purposes of this analysis, with years 1999 – 2015 ending up in the final cut due to the variety of target variables available during these years. These years were then concatenated into one dataset and the contents procedure run to verify its contents:

```
/* Note: Years 1991, 1993, 1995, 1997 excluded due to lack of
Depression Variable */
proc contents data=YRBS_Total;
run;
```

Next, frequency procedures were performed in order to explore the descriptive and univariate statistics of our target predictor variables within the dataset:

```
/* Building of Table 1: Descriptive and Univariate Statistics */
proc freq data=YRBS_Total;
tables SubAbuseBin_Cat * SI_Cat;
run;

proc freq data=YRBS_Total;
tables (SubAbuse_Cat Age_Cat Sex_Cat Race_Cat Depression_Cat RecSubAbuse_Cat
VictimViol_Cat ActiveViol_Cat) * SI_Cat / chisq;
run;

data newYRBS_Total (keep = SubAbuse SubAbuse_Cat Age Age_Cat Sex
Sex_Cat Race Race_Cat Depression Depression_Cat RecSubAbuse
RecSubAbuse_Cat VictimViol VictimViol_Cat ActiveViol ActiveViol_Cat SI
SI_Cat SubAbuseBin_Cat);
set YRBS_Total (where= ( (SubAbuse in (0,1,2,3)) and (Age
in(12,13,14,15,16,17,18)) and (Sex in (1,2)) and (Race in
(1,2,3,4,5,6)) and (Depression in (0,1)) and (RecSubAbuse in (0,1)) and
(VictimViol in (0,1,2)) and (ActiveViol in (0,1,2)) and (SI in (0,1))
and (SubAbuseBin in (0,1)) ));
run;

proc freq data=newYRBS_Total;
tables ( Age_Cat Sex_Cat Race_Cat Depression_Cat RecSubAbuse_Cat
VictimViol_Cat ActiveViol_Cat ) * SubAbuse_Cat / chisq;
run;
```

After we have reviewed these results and obtained a good grasp on the relationships between each of the variables, we can then run the descriptive and univariate statistics on the predictor variables and the target outcome variable:

```
/* Building of Table 2: Descriptive and Univariate Statistics */
proc freq data=newYRBS_Total;
tables (SubAbuse_Cat Age_Cat Sex_Cat Race_Cat Depression_Cat
RecSubAbuse_Cat VictimViol_Cat ActiveViol_Cat) * SI_Cat / chisq;
run;
```

After another thorough review of these results, we can then run a preliminary multivariable logistic regression analysis to examine the multiplicative interaction of the chosen variables. An initial examination of the interactions can be made at this time through the results of the analysis:

```
proc logistic data = newYRBS_Total;
class SI_Cat (ref='No') SubAbuse_Cat (ref='1 None') / param=ref;
model SI_Cat = SubAbuse_Cat / lackfit rsq;
title 'Suicidal Ideation by Lifetime Substance Abuse Severity,
Unadjusted';
run;

proc logistic data = newYRBS_Total;
class SI_Cat(ref='No') SubAbuse_Cat (ref='1 None') Age_Cat (ref='12 or
younger') Sex_Cat (ref='Female') Race_Cat (ref='White') Depression_Cat
(ref='No') RecSubAbuse_Cat (ref='No') VictimViol_Cat (ref='None')
ActiveViol_Cat (ref='None') / param=ref;
model SI_Cat = SubAbuse_Cat Age_Cat Sex_Cat Race_Cat Depression_Cat
RecSubAbuse_Cat VictimViol_Cat ActiveViol_Cat / lackfit rsq;
```

```
title 'Suicidal Ideation by Lifetime Substance Abuse Severity, Adjusted
- Multivariable Logistic Regression';
run;
```

MULTICOLLINEARITY INVESTIGATION

Finally! We can begin to explore whether or not our chosen model is suffering the effects of multicollinearity! Given the analyses we conducted above, could you identify any possible variable interactions that could be ending in multicollinearity? Here's a hint: could being a victim of violence lead to depression? Could recent substance abuse be highly correlated with lifetime substance abuse? These are questions we will be able to answer through our multicollinearity analysis.

Our first step is to explore the correlation matrix. We can do this through implementation of the corr procedure:

```
/* Examination of the Correlation Matrix */
proc corr data=newYRBS_Total;
var SI SubAbuse Age Sex Race Depression RecSubAbuse VictimViol
ActiveViol;
title 'Suicidal Ideation Predictors - Examination of Correlation
Matrix';
run;
```

Pretty easy right? Now let's look at the results:

Pearson Correlation Coefficients, N = 119374
Prob > |r| under H0: Rho=0

	SI	SubAbuse	Age	Sex	Race	Depression	RecSubAbuse	VictimViol	ActiveViol
SI	1.00000	0.16274 <.0001	-0.02536 <.0001	-0.12442 <.0001	0.03251 <.0001	0.41170 <.0001	0.13484 <.0001	0.18064 <.0001	0.12845 <.0001
SubAbuse	0.16274 <.0001	1.00000	0.17483 <.0001	0.07054 <.0001	-0.01079 0.0002	0.16046 <.0001	0.67232 <.0001	0.09992 <.0001	0.31903 <.0001
Age	-0.02536 <.0001	0.17483 <.0001	1.00000	0.04411 <.0001	-0.02015 <.0001	0.00497 0.0863	0.12273 <.0001	-0.04538 <.0001	-0.02538 <.0001
Sex	-0.12442 <.0001	0.07054 <.0001	0.04411 <.0001	1.00000	-0.00597 0.0393	-0.16646 <.0001	0.02899 <.0001	0.00651 0.0245	0.26876 <.0001
Race	0.03251 <.0001	-0.01079 0.0002	-0.02015 <.0001	-0.00597 0.0393	1.00000	0.06307 <.0001	-0.01675 <.0001	0.02870 <.0001	0.01487 <.0001
Depression	0.41170 <.0001	0.16046 <.0001	0.00497 0.0863	-0.16646 <.0001	0.06307 <.0001	1.00000	0.13819 <.0001	0.20213 <.0001	0.11232 <.0001
RecSubAbuse	0.13484 <.0001	0.67232 <.0001	0.12273 <.0001	0.02899 <.0001	-0.01675 <.0001	0.13819 <.0001	1.00000	0.07573 <.0001	0.26472 <.0001
VictimViol	0.18064 <.0001	0.09992 <.0001	-0.04538 <.0001	0.00651 0.0245	0.02870 <.0001	0.20213 <.0001	0.07573 <.0001	1.00000	0.17718 <.0001
ActiveViol	0.12845 <.0001	0.31903 <.0001	-0.02538 <.0001	0.26876 <.0001	0.01487 <.0001	0.11232 <.0001	0.26472 <.0001	0.17718 <.0001	1.00000

Figure 1: Pearson Correlation Results

Keep in mind, while reviewing these results we want to check to see if any of the variables included have a high correlation – about 0.8 or higher – with any other variable. As we can see, upon review of this correlation matrix, there does not appear to be any variables with a particularly high correlation. We are not done yet, though. Next we will examine multicollinearity through the Variance Inflation Factor and Tolerance. This can be done by specifying the "vif", "tol", and "collin" options after the model statement:

```
/* Multicollinearity Investigation of VIF and Tolerance */
proc reg data=newYRBS_Total;
```

```
model SI = SubAbuse Age Sex Race Depression RecSubAbuse VictimViol
ActiveViol / vif tol collin;
title 'Suicidal Ideation Predictors - Multicollinearity Investigation
of VIF and Tol';
run;
quit;
```

First we will review the parameter estimates, tolerance, and variance inflation.

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	1	0.25112	0.01319	19.04	<.0001	.	0
SubAbuse	1	0.02387	0.00113	21.05	<.0001	0.51212	1.95266
Age	1	-0.00994	0.00080178	-12.40	<.0001	0.95617	1.04584
Sex	1	-0.06526	0.00205	-31.88	<.0001	0.88446	1.13064
Race	1	0.00175	0.00070814	2.47	0.0136	0.99460	1.00543
Depression	1	0.29035	0.00223	130.47	<.0001	0.89608	1.11597
RecSubAbuse	1	0.01239	0.00262	4.73	<.0001	0.54332	1.84053
VictimViol	1	0.03899	0.00121	32.25	<.0001	0.93201	1.07295
ActiveViol	1	0.03161	0.00137	23.07	<.0001	0.79769	1.25362

Figure 2: Tolerance and VIF Investigation Results

In reviewing tolerance, we want to make sure that no values fall below 0.1. In the above results, we can see that the lowest tolerance value is 0.51212, so there is no threat of multicollinearity indicated through our tolerance analysis. As for variance inflation, the magic number to look out for is anything above the value of 10. As we can see from the values indicated in this column, our highest value sits at 1.95266, indicating a lack of multicollinarity, according to these results. However, we are not done yet, we will now look at the collinearity diagnostics for an eigensystem analysis of covariance comparison:

Collinearity Diagnostics

Number	Eigenvalue	Condition Index	Proportion of Variation								
			Intercept	SubAbuse	Age	Sex	Race	Depression	RecSubAbuse	VictimViol	ActiveViol
1	6.15520	1.00000	0.00012813	0.00401	0.00013209	0.00202	0.00532	0.00674	0.00489	0.00724	0.00697
2	0.71214	2.93994	0.00017002	0.00454	0.00018935	0.00604	0.00805	0.44505	0.00874	0.30195	0.00008712
3	0.66895	3.03335	0.00053394	0.03582	0.00051801	0.00610	0.06081	0.00025086	0.12401	0.01577	0.20191
4	0.57755	3.26458	0.00000936	0.00648	0.00001596	0.00091695	0.00212	0.38056	0.02326	0.41327	0.17436
5	0.46314	3.64555	0.00000662	0.02879	0.00000214	0.00205	0.00678	0.09846	0.12903	0.25195	0.50663
6	0.22094	5.27823	0.00151	0.00347	0.00167	0.05590	0.86056	0.01125	0.04167	0.00335	0.01567
7	0.14138	6.59826	0.00026638	0.89472	0.00018248	0.01189	0.01069	0.00417	0.66683	1.773809E-7	0.00611
8	0.05795	10.30616	0.01681	0.00857	0.01889	0.91118	0.04010	0.05263	0.00150	0.00224	0.08397
9	0.00276	47.22351	0.98057	0.01361	0.97840	0.00389	0.00556	0.00088405	0.00005364	0.00424	0.00429

Figure 3: Collinearity Investigation Results

In review of these results, our focus is going to be on the relationship of the eigenvalue column to the condition index column. If one or more of the eigenvalues are small (close to zero) and the corresponding condition number large, then we have an indication of multicollinearity. As we can see from the above results, none of our eigenvalues and condition index associations match this description.

So what is our conclusion from this example? This example was covered in order to show you that multicollinearity can not be deduced from simply thinking about the data in a logical manner. Knowing your data and thinking about possible confounding interactions is certainly a best practices guideline, but

multicollinearity analyses should still be conducted to test your theory before taking measures to combat something that is not there.

COMBATING MULTICOLLINEARITY

Do you feel betrayed? Don't feel that way! Next we will cover a dataset that is flush with multicollinearity in order to appropriately show you how to combat it. This second section will explain the different strategies for combating multicollinearity in a dataset. While reviewing this section, the author would like you to, again, think logically about the model being explored. Try identifying possible multicollinearity issues before reviewing the results of the diagnostic tests, and then think critically about the different strategies used to combat the collinearity issue.

INTRODUCTION TO THE SECOND DATSET

This second dataset is easily accessible by anyone with access to SAS®. It is a sample dataset titled "lipids". The background to this sample dataset states that it is from a study to investigate the relationships between various factors and heart disease. In order to explore this relationship, blood lipid screenings were conducted on a group of patients. Three months after the initial screening, follow-up data was collected from a second screening that included additional information such as gender, age, weight, total cholesterol, and history of heart disease. The outcome variable of interest in this analysis is the reduction of cholesterol level between the initial and 3-month lipid panel or "cholesterolloss". The predictor variables of interest are age (age of participant), weight (weight at first screening), cholesterol (total cholesterol at first screening), triglycerides (triglycerides level at first screening), HDL (HDL level at first screening), LDL (LDL level at first screening), height (height of participant), skinfold (skinfold measurement), systolicbp (systolic blood pressure) diastolicbp (diastolic blood pressure), exercise (exercise level), and coffee (coffee consumption in cups per day).

To begin our analysis, we will first explore the dataset, just as we did in the earlier example:

```
/* Example of Multicollinearity Findings */
libname health
"C:\ProgramFiles\SASHome\SASEnterpriseGuide\7.1\Sample\Data";
data health;
set health.lipid;
run;

proc contents data=health;
title 'Health Dataset with High Multicollinearity';
run;
```

For the sake of time, we will skip the thorough investigation of the relationships between the different variables in the dataset. Instead, we want to concentrate on the impending incidence of multicollinearity and how to combat it. Therefore, we will then test for multicollinearity using the procedures outlined earlier and review the results:

```
/* Assess Pairwise Correlations of Continuous Variables */
proc corr data=health;
var age weight cholesterol triglycerides hdl ldl height skinfold
systolicbp diastolicbp exercise coffee cholesterolloss;
title 'Health Predictors - Examination of Correlation Matrix';
run;

proc reg data=health;
```

```
model cholesterolloss = age weight cholesterol triglycerides hdl ldl
height skinfold systolicbp diastolicbp exercise coffee / vif tol
collin;
title 'Health Predictors - Multicollinearity Investigation of VIF and
Tol';
run;
```

Below you will find a clip of the correlation procedure results:

Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	Age	Weight	Cholesterol	Triglycerides	HDL	LDL	Height	Skinfold	SystolicBP	DiastolicBP	Exercise	Coffee	CholesterolLoss
Age	1.00000 95	0.08935 0.3892 95	0.26282 0.0101 95	0.21167 0.0395 95	0.20310 0.0484 95	0.21588 0.0356 95	-0.02080 0.8414 95	0.10625 0.3055 95	0.02384 0.8186 95	-0.06384 0.5388 95	-0.12193 0.2392 95	0.25089 0.0142 95	0.09914 0.5270 43
Weight	0.08935 0.3892 95	1.00000 95	-0.02188 0.8333 95	0.10757 0.2994 95	-0.27555 0.0069 95	0.05743 0.5804 95	0.69794 <.0001 95	0.07427 0.4744 95	0.15740 0.1277 95	0.13627 0.1879 95	0.03254 0.7542 95	0.05720 0.5819 95	-0.24221 0.1176 43
Cholesterol	0.26282 0.0101 95	-0.02188 0.8333 95	1.00000 95	0.40081 <.0001 95	0.35246 0.0005 95	0.96170 <.0001 95	-0.07521 0.4688 95	0.07588 0.4649 95	-0.04103 0.6930 95	0.15969 0.1221 95	0.01305 0.9001 95	-0.01157 0.9114 95	0.40318 0.0073 43
Triglycerides	0.21167 0.0395 95	0.10757 0.2994 95	0.40081 <.0001 95	1.00000 95	-0.27838 0.0063 95	0.48904 <.0001 95	0.04071 0.6953 95	0.09292 0.3704 95	0.14545 0.1596 95	0.14073 0.1737 95	-0.11162 0.2815 95	-0.00350 0.9731 95	0.11396 0.4669 43
HDL	0.20310 0.0484 95	-0.27555 0.0069 95	0.35246 0.0005 95	-0.27838 0.0063 95	1.00000 95	0.08340 0.4217 95	-0.24465 0.0169 95	0.11116 0.2835 95	-0.06008 0.5630 95	0.02410 0.8167 95	-0.03055 0.7688 95	0.10955 0.2906 95	0.19099 0.2199 43
LDL	0.21588 0.0356 95	0.05743 0.5804 95	0.96170 <.0001 95	0.48904 <.0001 95	0.08340 0.4217 95	1.00000 95	-0.00777 0.9404 95	0.04547 0.6617 95	-0.03028 0.7708 95	0.16118 0.1187 95	0.02672 0.7972 95	-0.04585 0.6591 95	0.37389 0.0135 43
Height	-0.02080 0.8414 95	0.69794 <.0001 95	-0.07521 0.4688 95	0.04071 0.6953 95	-0.24465 0.0169 95	-0.00777 0.9404 95	1.00000 95	-0.13762 0.1835 95	0.08432 0.4166 95	0.06327 0.5424 95	0.00521 0.9600 95	0.07165 0.4902 95	-0.27042 0.0795 43
Skinfold	0.10625 0.3055 95	0.07427 0.4744 95	0.07588 0.4649 95	0.09292 0.3704 95	0.11116 0.2835 95	0.04547 0.6617 95	-0.13762 0.1835 95	1.00000 95	-0.09901 0.3398 95	-0.03817 0.7134 95	-0.26581 0.0092 95	0.07833 0.4505 95	-0.03538 0.8218 43
SystolicBP	0.02384 0.8186 95	0.15740 0.1277 95	-0.04103 0.6930 95	0.14545 0.1596 95	-0.06008 0.5630 95	-0.03028 0.7708 95	0.08432 0.4166 95	-0.09901 0.3398 95	1.00000 95	0.33476 0.0009 95	-0.05138 0.6209 95	-0.05048 0.6271 95	-0.07917 0.6138 43
DiastolicBP	-0.06384 0.5388 95	0.13627 0.1879 95	0.15969 0.1221 95	0.14073 0.1737 95	0.02410 0.8167 95	0.16118 0.1187 95	0.06327 0.5424 95	-0.03817 0.7134 95	0.33476 0.0009 95	1.00000 95	-0.03647 0.7257 95	0.03908 0.7069 95	0.13192 0.3991 43

Figure 4: Pearson Correlation Results

Upon inspection of these results, one is quickly drawn to the correlation coefficient of LDL and Cholesterol which values at a whopping 0.96170. We definitely have a case for further collinearity investigation here. This is further supported in our review of the parameter estimates results for VIF and Tol:

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	1	5.72484	108.12644	0.05	0.9581	.	0
Age	1	-0.67645	2.20644	-0.31	0.7613	0.32637	3.06405
Weight	1	-0.20743	0.27789	-0.75	0.4612	0.32763	3.05224
Cholesterol	1	-182.68577	170.82886	-1.07	0.2934	4.326797E-7	2311178
Triglycerides	1	2.91187	2.73231	1.07	0.2951	0.00034921	2863.60930
HDL	1	182.75031	170.71293	1.07	0.2929	0.00000516	193966
LDL	1	183.05303	170.82561	1.07	0.2925	5.113026E-7	1955789
Height	1	-0.18955	1.61295	-0.12	0.9072	0.43551	2.29616
Skinfold	1	-0.07347	0.53443	-0.14	0.8916	0.77820	1.28502
SystolicBP	1	0.07945	0.63738	0.12	0.9016	0.66694	1.49939
DiastolicBP	1	-0.08111	0.43028	-0.19	0.8518	0.66583	1.50190
Exercise	1	0.05167	0.05513	0.94	0.3562	0.77863	1.28430
Coffee	1	3.99259	3.68202	1.08	0.2868	0.44992	2.22261

Figure 5: Tolerance and VIF Investigation Results

In review of the Tolerance results, we can see several variables – namely cholesterol, triglycerides, HDL, and LDL – having values well below our 0.1 cutoff value. This finding is echoed in review of the Variance Inflation results, where these same variables reveal values far larger than our 10 cutoff for this column. For the sake of completeness, we will also review the collinearity diagnostics:

Collinearity Diagnostics														
			Proportion of Variation											
Number	Eigenvalue	Condition Index	Intercept	Age	Weight	Cholesterol	Triglycerides	HDL	LDL	Height	Skinfold	SystolicBP	DiastolicBP	Exercise	Coffee
1	11.29489	1.00000	0.00001138	0.00004637	0.00006086	1.18985E-10	6.589162E-7	2.117859E-9	2.001E-10	0.00001048	0.00096170	0.00002146	0.00011281	0.00154	0.00092480
2	0.68622	4.05704	0.00000331	0.00000701	0.00001442	1.19889E-10	1.143653E-8	2.18204E-10	4.01922E-10	0.00000233	0.00417	0.00000915	0.00000136	0.14262	0.28041
3	0.47052	4.89952	1.797612E-8	0.00000475	0.00006283	1.43701E-10	0.00007230	6.693999E-9	6.81584E-10	7.934101E-7	0.00922	0.00000181	0.00000201	0.35333	0.11353
4	0.27571	6.40053	0.00007350	0.00002441	0.00066563	4.12089E-15	0.00024185	6.082316E-8	3.70204E-10	0.00008019	0.05181	0.00011487	0.00038936	0.17610	0.07292
5	0.14667	8.77543	0.00009651	0.00053911	0.00028359	1.554489E-9	0.00000596	7.219312E-8	1.834574E-9	0.00014592	0.77592	0.00026960	0.00215	0.19373	0.00009012
6	0.06145	13.55776	0.00082045	0.00016295	0.02554	4.483083E-8	0.00000217	0.00000126	6.61196E-8	0.00162	0.01235	0.00304	0.00411	0.00501	0.00073618
7	0.02723	20.36502	0.00093349	0.00170	0.04781	4.702702E-8	0.00025889	0.00000293	2.825254E-7	0.00015520	0.00302	0.00381	0.02703	0.00765	0.08354
8	0.02089	23.25002	0.00003049	0.04483	0.02385	4.595332E-9	0.00004015	0.00000125	4.865568E-8	0.00044312	0.00851	0.00011118	0.44175	0.01589	0.00722
9	0.00826	36.97981	0.03667	0.00022313	0.32353	7.986649E-9	0.00012325	0.00000321	1.021936E-7	0.00897	0.00304	0.04829	0.21593	0.07217	0.04171
10	0.00535	45.93079	0.00836	0.74848	0.10288	1.801143E-8	0.00013411	0.00000190	9.625344E-9	0.02801	0.01629	0.00809	0.13539	0.00271	0.23169
11	0.00195	76.09944	0.07125	0.17866	0.11829	2.005126E-8	5.808795E-8	6.652202E-7	2.868043E-8	0.13026	0.00141	0.85602	0.14837	0.00064741	0.16488
12	0.00085064	115.23088	0.87069	0.01200	0.27559	5.802692E-9	0.00002858	1.490124E-7	2.525126E-8	0.70634	0.10713	0.06251	0.00076513	0.02725	0.00001376
13	9.448677E-9	34574	0.01106	0.01333	0.08142	1.00000	0.99909	0.99999	1.00000	0.12396	0.00617	0.01772	0.02400	0.00133	0.00234

Figure 6: Collinearity Investigation Results

In review of the eigenvalue and condition index association, we can see a large deviation in the final three factors, with the eigenvalue resulting very close to zero and the condition index resulting quite large in comparison.

So, we have found a prime case for multicollinearity. Now that we have identified it, what can we do about it?

COMBATING MULTICOLLINEARITY

Is there an easy way to combat multicollinearity? Yes! All you need to do is drop one of your problem variables, rerun your analysis to test for further multicollinearity, and if none exist, then you are good to go! Can we always do this? Of course not. There are just some variables, no matter how highly correlated they are, that we need to keep in the model for the sake of scientific advancement and model completeness. If you run into a case where dropping a variable is not an option, you are in luck! There are

at least two alternative methods of estimation that can be employed. The main two that will be discussed in this paper are ridge regression and principal component regression.

Ridge Regression

Ridge regression is a variant to least squares regression and is oftentimes used when a multicollinearity case is identified. The traditional ordinary least squares (OLS) regression produces unbiased estimates for the regression coefficients, however, if you introduce the confounding issue of highly correlated explanatory variables, your resulting OLS parameter estimates end up with large variance. Therefore, it could be beneficial to utilize a technique such as ridge regression in order to ensure a smaller variance in resulting parameter estimates. Unfortunately, the trade-off of this is that a method such as ridge regression results in biased estimates. A more thorough review into the assumptions and specifications of ridge regression would be appropriate if this is a route you would like to take, but for now, we will run through the example as though we have decided that this is the best course of action:

```
/* Ridge Regression Example */
proc reg data=health outvif plots(only)=ridge(unpack VIFaxis=log)
outest=rrhealth ridge=0 to 0.10 by .002;
model cholesterolloss = age weight cholesterol triglycerides hdl ldl
height skinfold systolicbp diastolicbp exercise coffee;
plot / ridgeplot nomodel nostat;
title 'Health - Ridge Regression Calculation';
run;

proc print data=rrhealth;
title 'Health - Ridge Regression Results';
run;
```

A clip of the results produced by this procedure are outlined below:

Health - Ridge Regression Results

Obs	_MODEL_	_TYPE_	_DEPVAR_	_RIDGE_	_PCOMIT_	_RMSE_	Intercept	Age	Weight	Cholesterol	Triglycerides	HDL	LDL	Height	Skinfold	SystolicBP	DiastolicBP	Exe
1	MODEL1	PARMS	CholesterolLoss	.	.	27.1752	5.7248	-0.67645	-0.20743	-182.69	2.91	182.75	183.05	-0.18955	-0.07347	0.07945	-0.08111	0.0
2	MODEL1	RIDGEVIF	CholesterolLoss	0.000	.	.	.	3.06405	3.05224	2311178.32	2863.61	193965.71	1955789.11	2.29616	1.28502	1.49939	1.50190	1.2
3	MODEL1	RIDGE	CholesterolLoss	0.000	.	27.1752	5.7248	-0.67645	-0.20743	-182.69	2.91	182.75	183.05	-0.18955	-0.07347	0.07945	-0.08111	0.0
4	MODEL1	RIDGEVIF	CholesterolLoss	0.002	.	.	.	2.95765	2.74482	0.53	2.55	2.17	0.94	1.98441	1.26699	1.45826	1.45013	1.2
5	MODEL1	RIDGE	CholesterolLoss	0.002	.	27.6892	18.0400	-0.93560	-0.12267	0.15	-0.01	0.04	0.22	-0.79704	-0.02847	0.16709	-0.00780	0.0
6	MODEL1	RIDGEVIF	CholesterolLoss	0.004	.	.	.	2.89451	2.68813	0.51	2.51	2.13	0.91	1.95803	1.25712	1.44402	1.43484	1.2
7	MODEL1	RIDGE	CholesterolLoss	0.004	.	27.6894	18.1792	-0.92255	-0.12276	0.16	-0.01	0.03	0.21	-0.79677	-0.02841	0.16401	-0.00589	0.0
8	MODEL1	RIDGEVIF	CholesterolLoss	0.006	.	.	.	2.83372	2.63353	0.50	2.46	2.09	0.89	1.93237	1.24746	1.43010	1.41993	1.2
9	MODEL1	RIDGE	CholesterolLoss	0.006	.	27.6896	18.3156	-0.90977	-0.12284	0.16	-0.01	0.03	0.20	-0.79650	-0.02837	0.16100	-0.00403	0.0
10	MODEL1	RIDGEVIF	CholesterolLoss	0.008	.	.	.	2.77514	2.58093	0.49	2.42	2.05	0.87	1.90738	1.23799	1.41649	1.40541	1.2
11	MODEL1	RIDGE	CholesterolLoss	0.008	.	27.6900	18.4497	-0.89727	-0.12292	0.16	-0.01	0.03	0.20	-0.79623	-0.02833	0.15805	-0.00221	0.0

Figure 7: Ridge Regression Results

From these results we want to derive the appropriate ridge parameter or “k” to include in the analysis. The ridge parameter column is labeled _RIDGE_ and the associated values under each variable column are the new parameter estimates. There are several schools of thought concerning how to choose the best value of “k”. I recommend reading Dorugade and Kashid’s 2010 paper for more information on this matter. For the sake of time and brevity, the current paper will simply look at the least increase in _RMSE_ and a decrease in ridge variable inflation factors for each variable. In order to achieve this, we need not look far, as the ridge parameter of .002 increases the _RMSE_ only slightly from 27.1752 to 27.6894 and drops the VIF for each of our problem variables to below our 10 cutoff. Therefore, this study will choose the ridge parameter of .002 for the resulting parameter adjustments which are completed in the following code:

```
proc reg data=health outvif plots(only)=ridge(unpack VIFaxis=log)
```

```
outest=rrhealth_final ridge=.002;
model cholesterolloss = age weight cholesterol triglycerides hdl ldl
height skinfold systolicbp diastolicbp exercise coffee;
plot / ridgeplot nomodel nostat;
title 'Health - Ridge Regression Calculation';
run;

proc print data=rrhealth_final;
title 'Health - Ridge Regression Results';
run;
```

The results of which are the final adjusted model with the multicollinearity issue controlled!

Principal Components Regression

Another way to combat multicollineairty is through Principal Components Regression.

```
/* Principal Component Regression Example */
proc princomp data=health
out=pchealth prefix=z outstat=PCRhealth;
var age weight cholesterol triglycerides hdl ldl height skinfold
systolicbp diastolicbp exercise coffee;
title 'Health - Principal Component Regression Calculation';
run;
```

The results from this procedure are as follows:

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	2.48956585	0.46788004	0.2075	0.2075
2	2.02168581	0.49008701	0.1685	0.3759
3	1.53159881	0.27585212	0.1276	0.5036
4	1.25574669	0.10608628	0.1046	0.6082
5	1.14966041	0.21116409	0.0958	0.7040
6	0.93849633	0.12548045	0.0782	0.7822
7	0.81301588	0.13686385	0.0678	0.8500
8	0.67615203	0.15358194	0.0563	0.9063
9	0.52257008	0.15598914	0.0435	0.9499
10	0.36658094	0.13165403	0.0305	0.9804
11	0.23492691	0.23492664	0.0196	1.0000
12	0.00000026		0.0000	1.0000

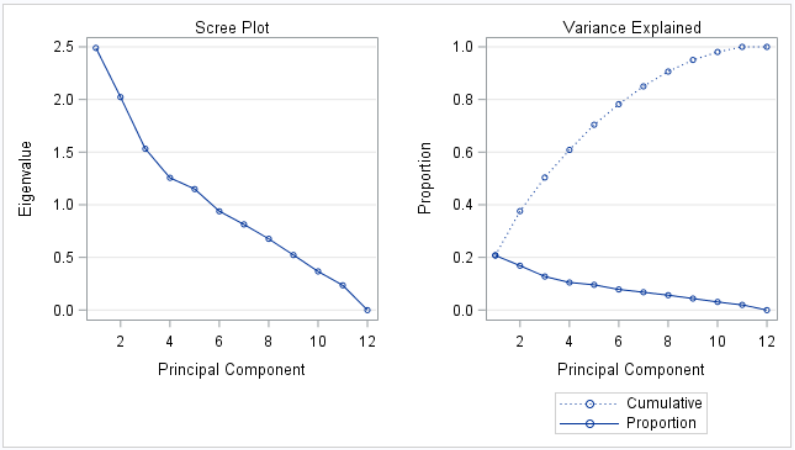

Figure 8 & 9: Principal Component Analysis Results

The main points we want to pull from these analyses are from the eigenvalues of the correlation matrix and a general review of the scree plot. Through these reviews we want to decide on the number of factors to keep for our final model. There are several schools of thought on how to choose this number, one school of thought being that any factor with an eigenvalue higher than 1.000 can remain in the model as it explains at least 1 variable's worth of information. If we were to go by this school of thought, our resulting model would include 5 factors and would be written like this:

```
/* With Eigenvalue Cutoff of 1.0000 */
proc reg data=pchealth;
model cholesterolloss = z1 z2 z3 z4 z5 / VIF;
title 'Health - Principal Component Regression Adjustment';
```

```
run;
```

However, there is another school of thought which utilizes a parallel analysis criterion in order to determine the appropriate number of factors to use in a model. The theory behind this school of thought explains that the eigenvalue obtained for the Nth factor should be larger than the associated eigenvalue computed analyzing a set of random data. The program associated with this school of thought is easy to obtain and available for public use via the site: https://people.ok.ubc.ca/brioconn/nfactors/parallel.sas. A sample of the user specifications indicated for this example are outlined below, though the program itself covers many more lines of code:

```
/****************** Parallel Analysis Program ***********************/
options nocenter nodate nonumber linesize=90; title;
proc iml;
reset noname;
/* enter your specifications here */
Ncases = 95;
Nvars = 12;
Ndatsets = 100;
percent = 95;
/* Specify the desired kind of parellel analysis, where:
1 = principal components analysis
2 = principal axis/common factor analysis */
kind = 1 ;
/* When seed = 0, the clock is used as the seed for the random
number generations. This produces different random numbers
on different runs of the program. To use the same random numbers on
different runs of the program, set seed to a value
other than 0 */
seed = 0;
/************* End of user specifications **********************/
```

The results of this analysis are available below and a copy of my previous analysis results are located to the right for comparison:

Random Data Eigenvalues

Root	Means	Prcntyle
1.000000	1.621344	1.750526
2.000000	1.451561	1.559002
3.000000	1.315314	1.414883
4.000000	1.197585	1.265765
5.000000	1.098348	1.165495
6.000000	1.010359	1.079955
7.000000	0.926823	1.004203
8.000000	0.844350	0.910322
9.000000	0.765708	0.825356
10.000000	0.677576	0.731967
11.000000	0.591609	0.666711
12.000000	0.499424	0.582079

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	2.48956585	0.46788004	0.2075	0.2075
2	2.02168581	0.49008701	0.1685	0.3759
3	1.53159881	0.27585212	0.1276	0.5036
4	1.25574669	0.10608628	0.1046	0.6082
5	1.14966041	0.21116409	0.0958	0.7040
6	0.93849633	0.12548045	0.0782	0.7822
7	0.81301588	0.13686385	0.0678	0.8500
8	0.67615203	0.15358194	0.0563	0.9063
9	0.52257008	0.15598914	0.0435	0.9499
10	0.36658094	0.13165403	0.0305	0.9804
11	0.23492691	0.23492664	0.0196	1.0000
12	0.00000026		0.0000	1.0000

Figure 10 & 11: Parallel Analysis Results for Comparison to PCA

In comparing these analyses, we want to look at the Prcntyle column of the Parallel Analysis Criterion results and the Eigenvalue Column of the target model results. In comparison of these results, we see that the eigenvalues for factor numbers 1-3 are all greater than the Prcntyle specified by the Parallel

Analysis Criterion, however, all factors after those first three fall short. Therefore, according to this school of thought, we should only keep 3 factors in our final model. This makes the final multicollinearity-adjusted model look like this:

```
/* After Parallel Analysis */
proc reg data=pchealth;
model cholesterolloss = z1 z2 z3 / VIF;
title 'Health - Principal Component Regression Adjustment';
run;
```

Either way you decide, you have combated multicollinearity in your final model!

CONCLUSION

Multicollinearity, if left untouched, can have a detrimental impact on the generalizability and accuracy of your model. If multicollinearity exists the traditional ordinary least squares estimators are imprecisely estimated, which leads to this inaccuracy in your judgment as to how each predictor variable impacts your target outcome variable. Given this information it is essential to detect and solve the issue of multicollinearity before estimating the parameters based on a fitted regression model.

Detecting multicollinearity is a fairly simple procedure involving the employment of VIF, Tol, and Collin model options. The CORR procedure is also useful in multicollinearity detection. After discovering the existence of multicollinearity, you can take one of three easily conducted roads: (1) drop a variable, (2) employ ridge regression, or (3) employ principal components regression. Through the steps outlined in this paper, one should be able to not only detect any issue of multicollinearity, but also resolve it in only a few short steps!

REFERENCES

Allison, P. (2012). “When Can You Safely Ignore Multicollinearity?” Statistical Horizons. September 10, 2012. http://statisticalhorizons.com/multicollinearity

Centers for Disease Control and Prevention (CDC). (2004). “Methodology of the Youth Risk Behavior Surveillance System”. Atlanta, Georgia: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention.

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

Columbia University Mailman School of Public Health. “Ridge Regression”. https://www.mailman.columbia.edu/research/populationhealthmethods/Ridgeregression

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

Dorugade, A. V., and Kashid, D. N. (2010). “Alternative Method for Choosing Ridge Parameter for Regression”. Applied Mathematical Sciences. 4(9): 447-456.

Joshi, H., Kulkarni, H., and Deshpande, S. (2012). “Multicollinearity Diagnostics in Statistical Modeling & Remedies to Deal With it Using SAS”. PhUSE 2012.

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

Parallel Analysis Criterion Program. https://people.ok.ubc.ca/brioconn/nfactors/parallel.sas

Unknown. “What is Multicollinearity?” https://onlinecourses.science.psu.edu/stat501/node/344

Wicklin, R. (2012). “Understanding Ridge Regression in SAS.” March 20, 2013. http://blogs.sas.com/content/iml/2013/03/20/compute-ridge-regression.html.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Deanna N Schreiber-Gregory, MS
Data Analyst / Research Associate
Contractor with Henry M Jackson Foundation for the Advancement of Military Medicine
Department of Internal Medicine
Uniformed Services University of the Health Sciences
E-mail: d.n.schreibergregory@gmail.com