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

Deanna N Schreiber-Gregory, MS

Henry M Jackson Foundation for the Advancement of Military Medicine

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

Presenter

Deanna N Schreiber-Gregory, Data Analyst II / Research Associate, Henry M Jackson Foundation for the Advancement of Military Medicine

Deanna is a Data Analyst and Research Associate through the Henry M Jackson Foundation. She is currently contracted to USUHS and Walter Reed National Military Medical Center in Bethesda, MD. Deanna has an MS in Health and Life Science Analytics, a BS in Statistics, and a BS in Psychology. Deanna has presented as a contributed and invited speaker at over 40 local, regional, national, and global SAS user group conferences since 2011.

@DN_SchGregory

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

Defining Multicollinearity

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

Defining Multicollinearity

What is Multicollinearity?

- Definition
  - A statistical phenomenon wherein there exists a perfect or exact relationship between predictor variables
- From a conventional standpoint:
  - Predictors are highly correlated
  - Predictors are co-dependent
- Notes
  - When things are related, we say they are linearly dependent
    - Fit well into a straight regression line that passes through many data points
  - Multicollinearity makes it difficult to come up with reliable estimates of individual coefficients for the predictor variables
    - Results in incorrect conclusions about the relationship between outcome and predictor variables

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

Defining Multicollinearity

What is Multicollinearity?

- Consider multiple linear regression equation:

$$Y = X\beta + \varepsilon$$

- Considering Equation:
  - Multicollinearity inflates the variances of the parameter estimates
    - (1) Lack of statistical significance of individual predictor variables, though overall model is still significant
    - (2) Biased outcome
  - The presence of multicollinearity can cause serious problems with the estimation of β and its interpretation

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

Defining Multicollinearity

Why Should We Care About Multicollinearity?

- Problems in Explanation vs Prediction Models
  - Explanation:
    - More difficult to achieve significance of collinear parameters
  - Prediction:
    - if estimates are statistically significant, they are only as reliable as any other variable in the model
    - If they are not significant, the sum of the coefficient is likely to be reliable
  - Corrections:
    - In the case of a predictive model: just need to increase sample size
    - In the case of an explanatory model: further measures are needed
- Primary concern: as the degree of multicollinearity increases…
  - Regression model estimates of the coefficients become unstable
  - Standard errors for the coefficients become wildly inflated

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

Detecting Multicollinearity

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

Detecting Multicollinearity

Ways to Detect Multicollinearity

- There are three ways to detect multicollinearity
  - Examination of the correlation matrix
  - Variance Inflation Factor (VIF)
  - Eigensystem Analysis of Correlation Matrix

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

Detecting Multicollinearity

Examination of the Correlation Matrix

- 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
- Proc Corr

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

Detecting Multicollinearity

Variance Inflation Factor / Tolerance

- Variance Inflation Factor
  - The Variance Inflation Factor (VIF) quantifies the severity of multicollinearity in an ordinary least-squares regression analysis
  - The VIF is an index which measures how much variance of an estimated regression coefficient is increased because of multicollinearity
  - Note: If any of the VIF values exceeds 5 or 10 it implies that the associated regression coefficients are poorly estimated because of multicollinearity
- Tolerance
  - Represented by 1/VIF

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

Detecting Multicollinearity

Eigensystem Analysis of Correlation Matrix

- 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)
  - Note: if one or more of the eigenvalues are small (close to zero) and a corresponding condition number is large, then it indicates multicollinearity

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

Detecting Multicollinearity

Example

- Model:
  - **Suicidal Ideation** = Lifetime Substance Use + Age + Gender + Racial Identification + Depression + Recent Substance Use + Victim of Violence + Participant in Violence
  - Suicidal Ideation = $\beta_0 + \beta_1$*(Lifetime Substance Use)* + β_2*(Age)* + β_3*(Gender)* + β_4*(Racial Identification)* + β_5*(Depression)* + β_6*(Recent Substance Use)* + β_7*(Victim of Violence)* + β_8*(Participant in Violence)*

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

Detecting Multicollinearity

Example

- Descriptive Statistics and Initial Examination

```
/* 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;


/* 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;


/* Building of Table 3: Multivariable Logistic Regression w/ Multiplicative Interaction */

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;
```


SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

Detecting Multicollinearity

Example

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	106966.38	84184.255
SC	106976.07	84407.125
-2 Log L	106964.38	84138.255

R-Square	0.1740	Max-rescaled R-Square	0.2941

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	22826.1273	22	<.0001
Score	23938.7349	22	<.0001
Wald	18179.0910	22	<.0001

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
SubAbuse_Cat	4	472.4316	<.0001
Age_Cat	6	112.6745	<.0001
Sex_Cat	1	972.4522	<.0001
Race_Cat	5	300.5116	<.0001
Depression_Cat	1	11300.3455	<.0001
RecSubAbuse_Cat	1	23.3679	<.0001
VictimViol_Cat	2	861.6686	<.0001
ActiveViol_Cat	2	527.6158	<.0001

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

Detecting Multicollinearity

Example

- Test: Examination of the Correlation Matrix

```
/* 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;
```


SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

Detecting Multicollinearity

Example

- Note: No highly correlated predictor variables

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

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

Detecting Multicollinearity

Example

- Tests:
  - Variance Inflation Factor
  - Eigensystem Analysis of Correlation Matrix

```
/* 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;
```

- Note:
  - Common cut point for VIF = 10 (higher indicates multicollinearity)
  - Common cut point for Tol = .1 (lower indicates multicollinearity)

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

Detecting Multicollinearity

Example

- Note: VIF cut point = 10, Tolerance cut point = .1

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

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

Detecting Multicollinearity

Example

- Note:
  - **Eigensystem Analysis of Covariance**: If one or more of the eigenvalues are small (close to zero) and the corresponding

			Collinearity Diagnostics								
			Proportion of Variation								
Number	Eigenvalue	Condition Index	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

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

Combating Multicollinearity

Introduction to Techniques

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

Detecting Multicollinearity
Example

- The dataset: SAS Sample Data

```
libname health "C:\Program Files\SASHome\SASEnterpriseGuide\7.1\Sample\Data";
data health;      set health.lipid;       run;

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

- The example:
  - **Outcome**: Cholesterol loss between baseline and check-up
  - **Predictors (Baseline)**: Age, Weight, Cholesterol, Triglycerides, HDL, LDL, Height

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

Detecting Multicollinearity

Example

- Test: Examination of the Correlation Matrix

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


SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

Detecting Multicollinearity

Example

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

	Age	Weight	Cholesterol	Triglycerides	HDL	LDL	Height	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	0.09914 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	-0.24221 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	0.40318 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	0.11396 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	0.19099 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	0.37389 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	-0.27042 0.0795 43
CholesterolLoss	0.09914 0.5270 43	-0.24221 0.1176 43	0.40318 0.0073 43	0.11396 0.4669 43	0.19099 0.2199 43	0.37389 0.0135 43	-0.27042 0.0795 43	1.00000 43

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

Detecting Multicollinearity

Example

- Tests:
  - Variance Inflation Factor
  - Eigensystem Analysis of Correlation Matrix

```
/* Multicollinearity Investigation of VIF and Tolerance */
proc reg data=health;
    model cholesterolloss = age weight cholesterol triglycerides hdl ldl height /
    vif tol collin;
    title 'Health Predictors - Multicollinearity Investigation of VIF and  Tol';
run;
```

- Note:
  - Common cut point for VIF = 10 (higher indicates multicollinearity)
  - Common cut point for Tol = .1 (lower indicates multicollinearity)

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

Detecting Multicollinearity

Example

- Note: VIF cut point = 10, Tolerance cut point = .1

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	1	18.38590	86.45275	0.21	0.8328	.	0
Age	1	0.63264	1.68351	0.38	0.7093	0.51425	1.94457
Weight	1	-0.29825	0.24873	-1.20	0.2385	0.37514	2.66571
Cholesterol	1	-169.20149	157.59569	-1.07	0.2903	4.663583E-7	2144274
Triglycerides	1	2.67536	2.51627	1.06	0.2950	0.00037770	2647.57331
HDL	1	169.19195	157.46718	1.07	0.2900	0.00000556	179909
LDL	1	169.52519	157.59200	1.08	0.2894	5.511058E-7	1814534
Height	1	-0.26426	1.45480	-0.18	0.8569	0.49108	2.03634

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

Detecting Multicollinearity

Example

- **Eigensystem Analysis of Covariance**: If one or more of the eigenvalues are small (close to zero) and the corresponding condition number is large, then it indicates multicollinearity

Collinearity Diagnostics										
			Proportion of Variation							
Number	Eigenvalue	Condition Index	Intercept	Age	Weight	Cholesterol	Triglycerides	HDL	LDL	Height
1	7.57480	1.00000	0.00003622	0.00016237	0.00015525	2.87683E-10	0.00000165	5.04002E-9	4.85942E-10	0.00002624
2	0.31551	4.89979	0.00014232	0.00018194	0.00043972	3.21062E-11	0.00033484	1.082107E-7	2.794E-10	0.00010102
3	0.05782	11.44595	0.00178	0.00184	0.05104	4.361274E-8	1.141859E-7	0.00000124	6.388516E-8	0.00275
4	0.03337	15.06626	0.00044517	0.01226	0.01308	5.377563E-8	0.00025542	0.00000323	3.193503E-7	0.00016967
5	0.01055	26.79431	0.06288	0.31489	0.12880	2.36137E-15	0.00001378	8.595756E-8	6.73401E-10	0.02608
6	0.00695	33.01681	0.02236	0.61435	0.40629	2.946854E-9	0.00023471	0.00000642	2.086847E-8	0.00031216
7	0.00100	86.86528	0.84879	0.02428	0.28558	5.400146E-9	0.00002137	1.778525E-7	2.419023E-8	0.85275
8	1.018426E-8	27272	0.06358	0.03202	0.11462	1.00000	0.99914	0.99999	1.00000	0.11780

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

Combating Multicollinearity

What Can We Do?

- Easist
  - 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:
  - Principal Component Regression
  - Regularization Techniques
    - L1: Lasso Regression
    - L2: Ridge Regression

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

Combating Multicollinearity

Principal Component Regression

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

Combating Multicollinearity

Principal Component Regression

- Logic:
  - 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
    - Often referred to as: Principal Components
  - The principal component regression approach combats multicollinearity by using less than the full set of principal components in the model
- Calculation:
  - To obtain the principal components estimators
    - Assume the regressors are arranged in order of decreasing eigenvalues, Λ1 ≥ Λ2 ……….. ≥ Λp > 0
  - In principal components regression, the principal components corresponding to near zero eigenvalues are removed from the analysis
    - Least squares is then applied to the remaining components

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

Combating Multicollinearity

Principal Component Regression Example

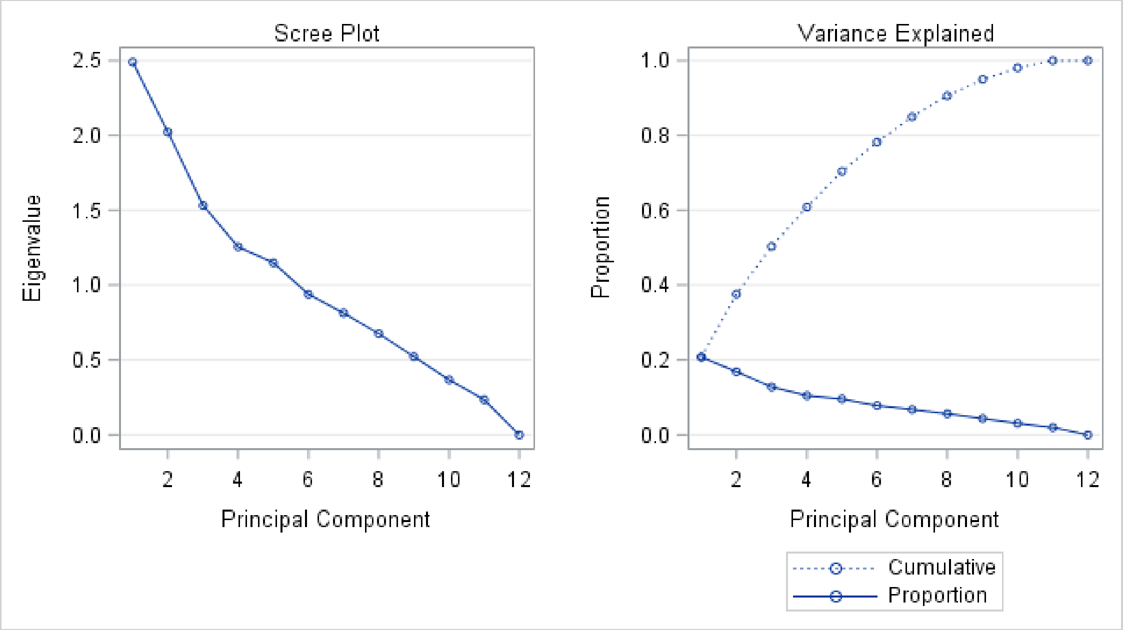
```
/* Principal Component Regression Example */
proc princomp data=health
   out=pchealth prefix=z outstat=PCRhealth;
   var age weight cholesterol triglycerides hdl ldl height skinfold
systolicbp diastolicbp exercise coffee;
   title 'Health - Principal Component Regression Calculation';
run;
```


SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

Combating Multicollinearity

Principal Component Regression Example

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



SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

Combating Multicollinearity

Principal Component Regression Example

Eigenvectors												
	z1	z2	z3	z4	z5	z6	z7	z8	z9	z10	z11	z12
Age	0.285637	-.044530	0.360129	-.107183	0.197629	-.424055	0.538446	0.083846	-.463623	-.148552	0.149668	-.000052
Weight	0.045070	0.576106	0.162040	-.206756	0.178609	0.253053	0.146115	0.050489	-.009628	-.125054	-.679337	0.000032
Cholesterol	0.589099	-.098840	-.155368	-.093205	0.134127	0.116176	-.103844	-.123895	0.130719	-.137223	0.016940	-.718708
Triglycerides	0.397049	0.208297	-.043878	0.015019	-.473430	-.287212	0.025209	0.121105	0.034811	0.670718	-.153047	0.019703
HDL	0.148467	-.411915	0.120623	0.128204	0.545978	0.268863	0.130258	-.188288	0.070864	0.521875	-.189061	0.203421
LDL	0.579902	0.012948	-.203531	-.140425	-.008074	0.051902	-.152852	-.079805	0.118651	-.328047	0.080786	0.664598
Height	-.028104	0.558856	0.109706	-.255942	0.267610	0.131273	-.096791	-.255442	0.011081	0.287750	0.602460	0.000040
Skinfold	0.118232	-.111201	0.403969	-.038559	-.357155	0.622241	0.275311	0.363410	0.157077	-.021458	0.247465	-.000122
SystolicBP	0.042721	0.270369	-.053289	0.638077	0.050025	-.122874	0.439890	-.165137	0.499658	-.135798	0.092847	-.000017
DiastolicBP	0.166446	0.211570	-.089449	0.607751	0.156683	0.217127	-.291874	0.350150	-.514758	0.011380	0.074679	-.000032
Exercise	-.075818	0.023719	-.525480	-.235626	0.321657	-.054307	0.250184	0.658040	0.196897	0.098325	0.107622	-.000014
Coffee	0.057609	-.003293	0.547553	0.040980	0.243041	-.339950	-.458832	0.371986	0.411539	-.046201	-.011938	-.000042

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

Combating Multicollinearity

Principal Components Regression Example

- Two ways to estimate the appropriate eigenvalue cut-off
  - Common: cut-off of 1
    - Explains at least 1 variable's worth of information
  - Parallel Analysis Criterion
    - Eigenvalue obtained for the Nth factor should be larger than the associated eigenvalue computed analyzing a set of random data

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

Combating Multicollinearity

Principal Component Regression Example

- First Example: Common method using eigenvalue of at least 1.0000

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

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

Combating Multicollinearity

Principal Component Regression Example

- Model is then rewritten in the form of principal components:
  - Cholesterol Loss = $\alpha 1z1 + \alpha 2z2 + \alpha 3z3 + \alpha 4z4 + \alpha 5z5 + \varepsilon$
    - Zn = Eigenvector(age) + Eigenvector(weight) + …….. + Eigenvector(coffee)
    - Estimated values of alphas can be obtained by regression cholesterol loss against z1, z2, z3, z4, & z5

```
/* 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;
```


SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

Combating Multicollinearity

Principal Component Regression Example

Health - Principal Component Regression Adjustment

The REG Procedure
Model: MODEL1
Dependent Variable: CholesterolLoss

Number of Observations Read	95
Number of Observations Used	43
Number of Observations with Missing Values	52

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	7389.99357	1477.99871	2.22	0.0732
Error	37	24668	666.69408		
Corrected Total	42	32058			

Root MSE	25.82042	R-Square	0.2305
Dependent Mean	9.76744	Adj R-Sq	0.1265
Coeff Var	264.35192		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	10.68090	4.05210	2.64	0.0122	0
z1	1	5.61148	2.23809	2.51	0.0167	1.01229
z2	1	-5.27905	2.95764	-1.78	0.0825	1.08885
z3	1	-3.59451	2.86619	-1.25	0.2177	1.00304
z4	1	0.98377	3.34227	0.29	0.7701	1.06419
z5	1	1.64616	3.21785	0.51	0.6120	1.06963

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

Combating Multicollinearity

Principal Components Regression Example

- Second Example: Parallel Analysis Criterion

```
/****************** Parallel Analysis Program ************************/
/* Location: https://people.ok.ubc.ca/brioconn/nfactors/parallel.sas */
```

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

Random Data Eigenvalues

Root	Means	Prcntyle
1.000000	1.636113	1.793214
2.000000	1.456865	1.566985
3.000000	1.309524	1.398354
4.000000	1.199819	1.275018
5.000000	1.105158	1.174932
6.000000	1.012580	1.066621
7.000000	0.926869	0.987542
8.000000	0.843625	0.919325
9.000000	0.754799	0.816677
10.000000	0.677048	0.756767
11.000000	0.586043	0.670146
12.000000	0.491558	0.588654

trademarks or trademarks ... SA registration. Other brand and product names are trademarks of their respective companies.

Combating Multicollinearity

Principal Component Regression Example

- Model is then rewritten in the form of principal components:
  - Cholesterol Loss = $\alpha 1z1 + \alpha 2z2 + \alpha 3z3 + \varepsilon$
    - Zn = Eigenvector(age) + Eigenvector(weight) + …….. + Eigenvector(coffee)
    - Estimated values of alphas can be obtained by regression cholesterol loss against z1, z2, & z3

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


SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

Combating Multicollinearity

Principal Component Regression Example

Health - Principal Component Regression Adjustment

The REG Procedure
Model: MODEL1
Dependent Variable: CholesterolLoss

Number of Observations Read	95
Number of Observations Used	43
Number of Observations with Missing Values	52

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	7114.05734	2371.35245	3.71	0.0194
Error	39	24944	639.57993		
Corrected Total	42	32058			

Root MSE	25.28992	R-Square	0.2219
Dependent Mean	9.76744	Adj R-Sq	0.1621
Coeff Var	258.92058		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	10.47191	3.94536	2.65	0.0114	0
z1	1	5.57863	2.18963	2.55	0.0149	1.01000
z2	1	-5.40367	2.79301	-1.93	0.0603	1.01218
z3	1	-3.54587	2.80624	-1.26	0.2139	1.00229

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

Combating Multicollinearity

Ridge Regression

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

Combating Multicollinearity

Regularization Methods

- Logic:
  - Regularization adds a penalty to model parameters (all except intercepts) so the model generalizes the data instead of overfitting (a side effect of multicollinearity)
  - Two main types:
    - L1 – Lasso Regression
    - L2 – Ridge Regression

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

Combating Multicollinearity

Regularization Methods

- Ridge Regression
  - Squared magnitude of the coefficient is added as penalty to loss function
  - $\sum_{i=1}^{n}(Y_i - \sum_{j=1}^{p} X_{ij}\beta_j)^2 + \Lambda\sum_{j=1}^{p}\beta_j^2$
- Lasso Regression
  - Absolute value of magnitude of the coefficient is added as penalty to loss function
  - $\sum_{i=1}^{n}(Y_i - \sum_{j=1}^{p} X_{ij}\beta_j)^2 + \Lambda\sum_{j=1}^{p}|\beta_j|$
- Result:
  - if Λ =0 then the equation will go back to OLS estimations
  - If Λ is very large, too much weight would be added = under-fitting
  - NOTE: need to be careful with choice of Λ

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

Combating Multicollinearity

Regularization Methods

- Key difference:
  - Lasso Regression is meant to shrink the coefficient of the less important variables to zero
    - This works well if feature selection is the goal
    - Not necessarily good for multicollinearity
  - Ridge Regression adjust weights of the variables
    - Goal is not to shrink the coefficients to zero, but to adjust for representation of all relevant variables

- Ridge Regression Trade-Off
  - We are still dealing with an adjustment
  - Naturally results in biased outcomes

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

Combating Multicollinearity

Ridge Regression

- Ridge regression provides an alternative estimation method that can be used where multicollinearity is suspected

- Logic:
  - Multicollinearity leads to small characteristic roots
    - 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

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

Combating Multicollinearity

Ridge Regression

- Ridge Regression for alternative estimator
  - 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]
  - Note:
    - Pick the smallest value of k that produces a stable estimate of β
    - Get the variance inflation factors (VIF) close to 1

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

Combating Multicollinearty

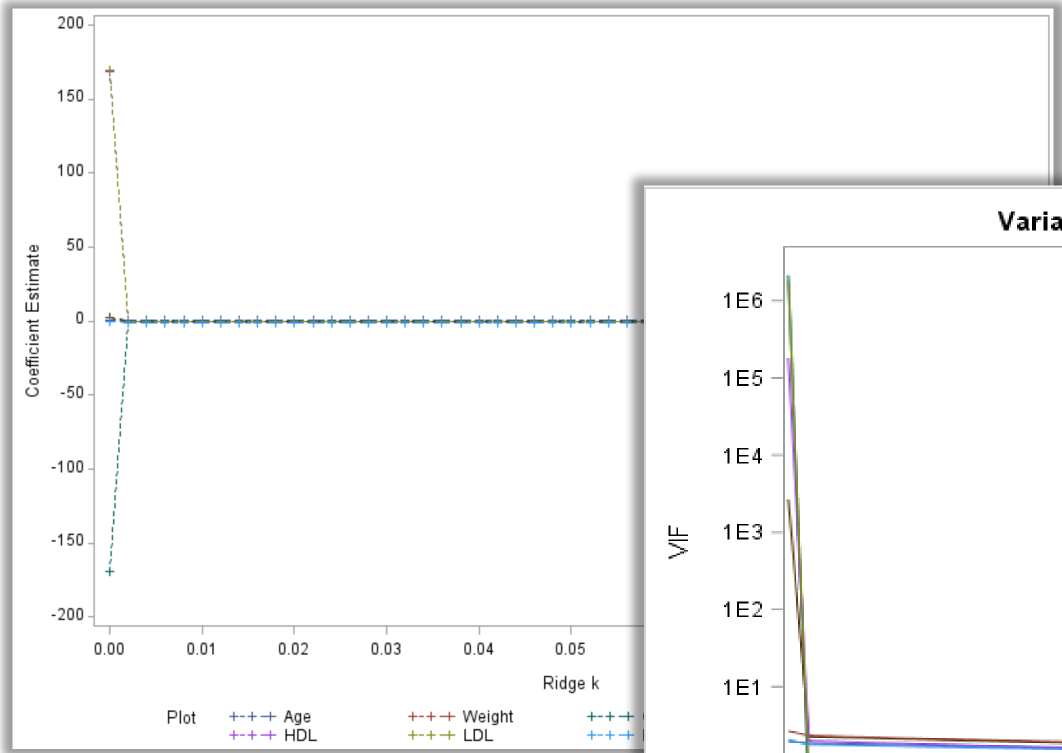
Ridge Regression Example

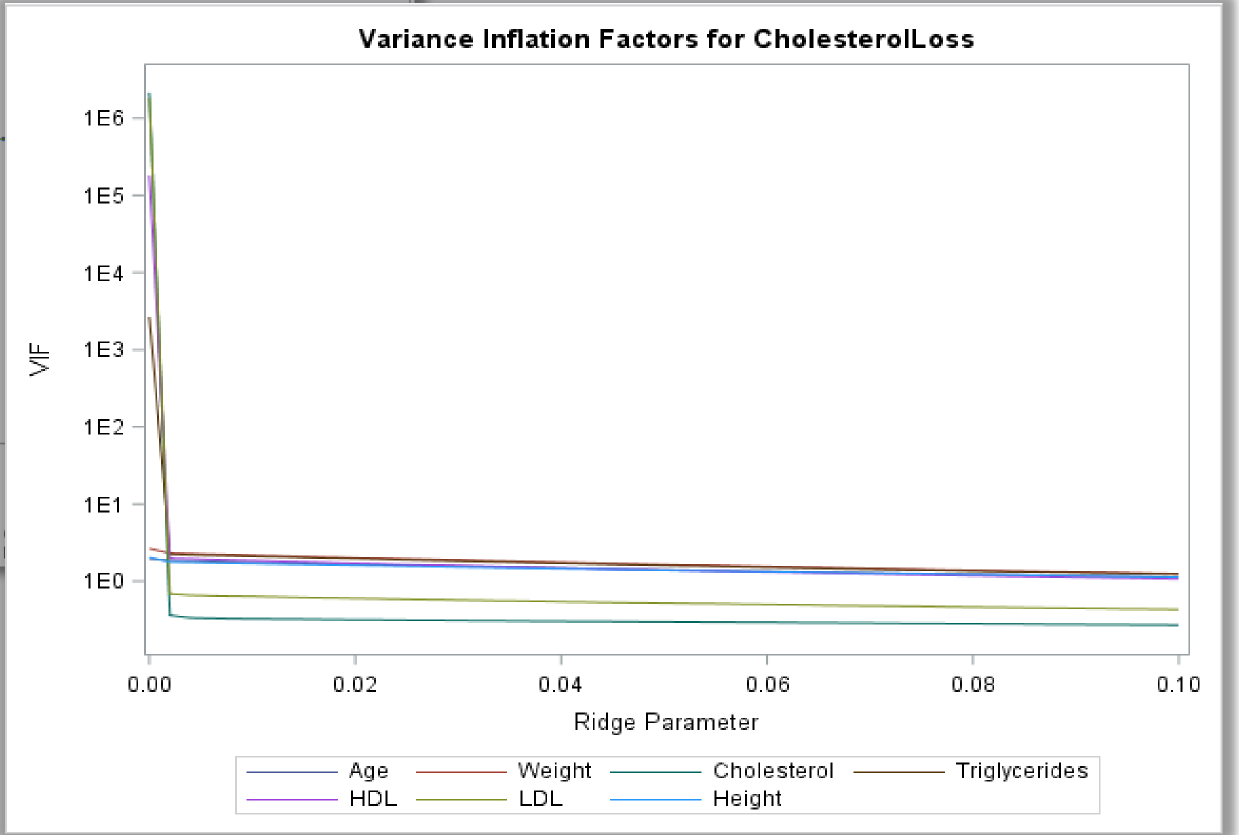
- Applying Ridge Regression:
  - Use PROC REG procedure with RIDGE option
  - RIDGEPLOT option will give graph of ridge trace

```
/* 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;
    plot / ridgeplot nomodel nostat;
    title 'Health - Ridge Regression Calculation';
run;

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


SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.





USERS PROGRAM

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

Combating Multicollinearity

Ridge Regression Example

Obs	_MODEL_	_TYPE_	_DEPVAR_	_RIDGE_	_PCOMIT_	_RMSE_	Intercept	Age	Weight	Cholesterol	Triglycerides	HDL	LDL	Height	CholesterolLoss
1	MODEL1	PARMS	CholesterolLoss	.	.	26.0275	18.3859	0.63264	-0.29825	-169.20	2.68	169.19	169.53	-0.26426	-1
2	MODEL1	RIDGEVIF	CholesterolLoss	0.000	.	.	.	1.94457	2.66571	2144274.02	2647.57	179909.00	1814533.58	2.03634	-1
3	MODEL1	RIDGE	CholesterolLoss	0.000	.	26.0275	18.3859	0.63264	-0.29825	-169.20	2.68	169.19	169.53	-0.26426	-1
4	MODEL1	RIDGEVIF	CholesterolLoss	0.002	.	.	.	1.85746	2.32171	0.36	2.25	1.98	0.69	1.77606	-1
5	MODEL1	RIDGE	CholesterolLoss	0.002	.	26.4533	41.8777	0.30397	-0.20670	0.13	-0.03	0.00	0.20	-0.80295	-1
6	MODEL1	RIDGEVIF	CholesterolLoss	0.004	.	.	.	1.83329	2.28437	0.34	2.21	1.94	0.66	1.75614	-1
7	MODEL1	RIDGE	CholesterolLoss	0.004	.	26.4534	41.9448	0.29907	-0.20563	0.14	-0.03	-0.00	0.19	-0.80508	-1
8	MODEL1	RIDGEVIF	CholesterolLoss	0.006	.	.	.	1.80977	2.24812	0.33	2.18	1.91	0.65	1.73665	-1
9	MODEL1	RIDGE	CholesterolLoss	0.006	.	26.4535	42.0080	0.29431	-0.20460	0.14	-0.03	-0.00	0.18	-0.80713	-1
10	MODEL1	RIDGEVIF	CholesterolLoss	0.008	.	.	.	1.78687	2.21290	0.33	2.14	1.88	0.64	1.71759	-1
11	MODEL1	RIDGE	CholesterolLoss	0.008	.	26.4536	42.0680	0.28969	-0.20359	0.14	-0.03	-0.00	0.18	-0.80909	-1

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

Combating Multicollinearity

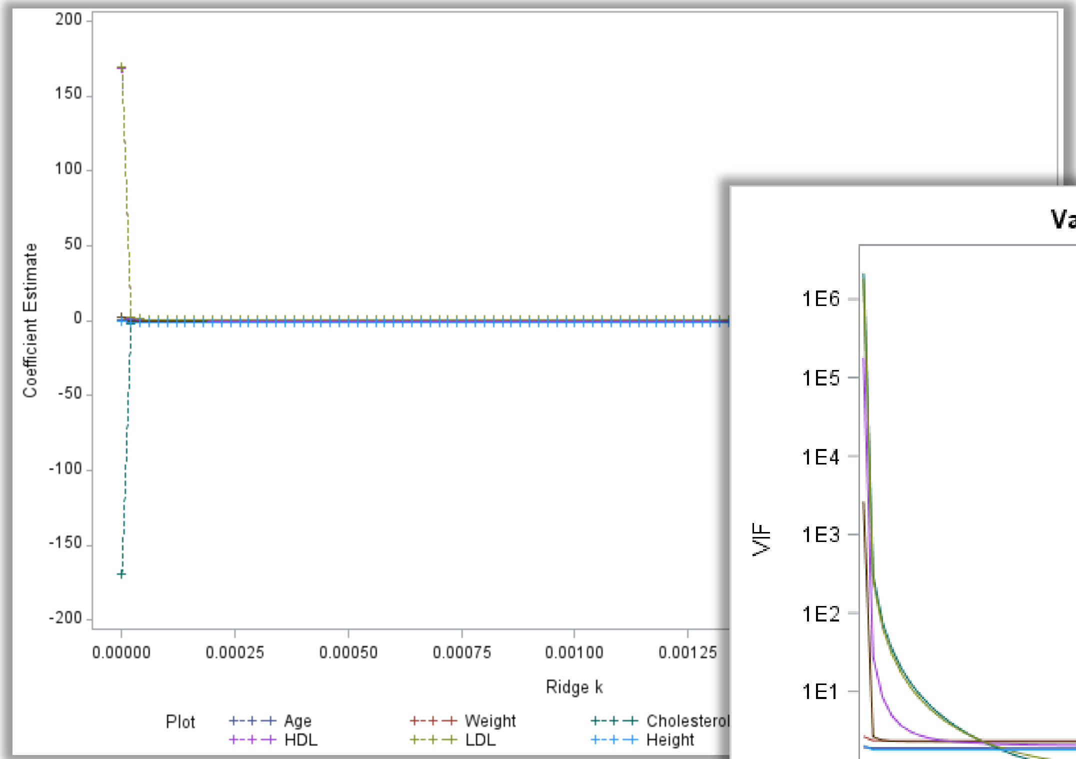
Ridge Regression Example

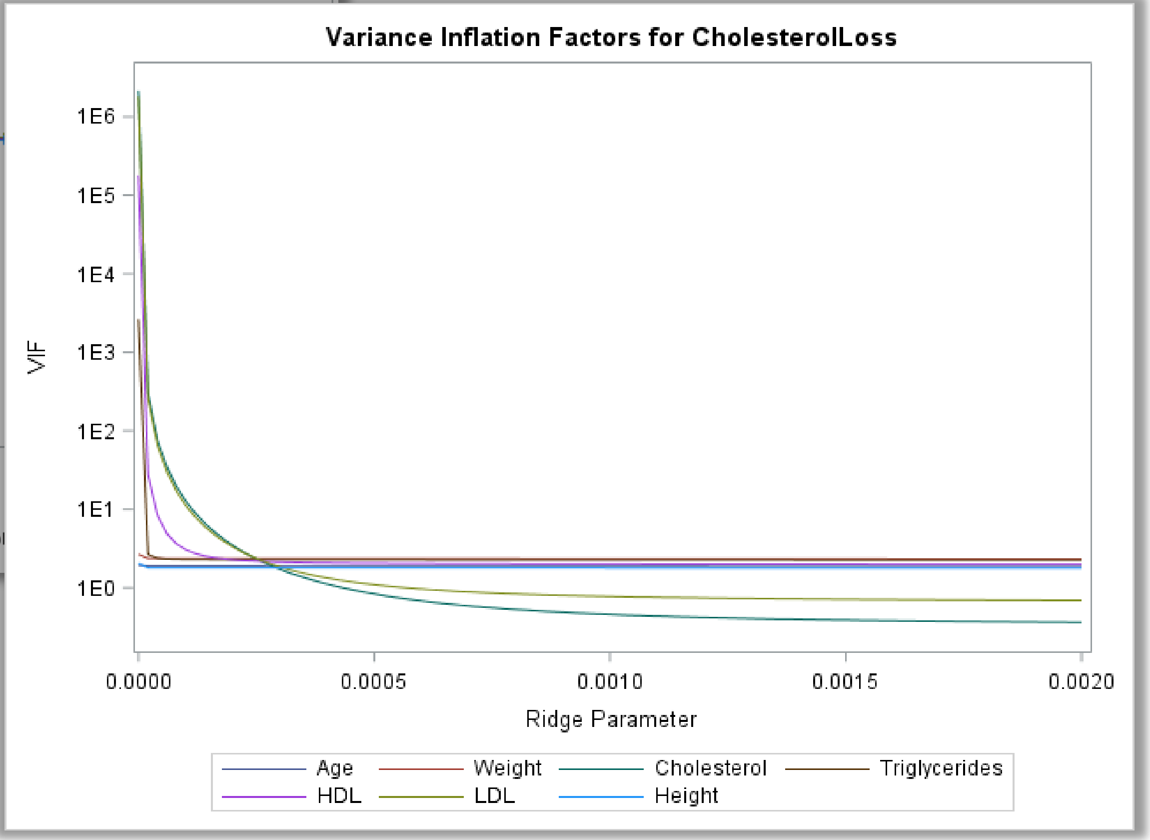
- Choose your alternative estimator
  - Pick the smallest value of k that process a stable estimate of β
  - Get the variance inflation factors (VIF) close to 1

```
proc reg data=health outvif plots(only)=ridge(unpack VIFaxis=log)
    outest=rrhealth_final ridge=0 to 0.002 by 0.00002;
    model cholesterolloss = age weight cholesterol triglycerides hdl ldl height;
    plot / ridgeplot nomodel nostat;
    title 'Health - Ridge Regression Calculation';
run;

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


SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.





SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

Combating Multicollinearity

Ridge Regression Example

Obs	_MODEL_	_TYPE_	_DEPVAR_	_RIDGE_	_PCOMIT_	_RMSE_	Intercept	Age	Weight	Cholesterol	Triglycerides	HDL	LDL	Height	CholesterolLoss
1	MODEL1	PARMS	CholesterolLoss	.	.	26.0275	18.3859	0.63264	-0.29825	-169.20	2.68	169.19	169.53	-0.26426	-1
2	MODEL1	RIDGEVIF	CholesterolLoss	.00000	.	.	.	1.94457	2.66571	2144274.02	2647.57	179909.00	1814533.58	2.03634	-1
3	MODEL1	RIDGE	CholesterolLoss	.00000	.	26.0275	18.3859	0.63264	-0.29825	-169.20	2.68	169.19	169.53	-0.26426	-1
4	MODEL1	RIDGEVIF	CholesterolLoss	.00002	.	.	.	1.88207	2.35983	305.48	2.66	27.61	258.89	1.79627	-1
5	MODEL1	RIDGE	CholesterolLoss	.00002	.	26.4434	41.5330	0.31276	-0.20883	-1.87	0.00	2.00	2.20	-0.79445	-1
6	MODEL1	RIDGEVIF	CholesterolLoss	.00004	.	.	.	1.88181	2.35940	77.54	2.38	8.49	66.00	1.79604	-1
7	MODEL1	RIDGE	CholesterolLoss	.00004	.	26.4483	41.6726	0.31079	-0.20829	-0.87	-0.01	1.00	1.20	-0.79765	-1
8	MODEL1	RIDGEVIF	CholesterolLoss	.00006	.	.	.	1.88156	2.35901	34.78	2.32	4.90	29.82	1.79583	-1
9	MODEL1	RIDGE	CholesterolLoss	.00006	.	26.4500	41.7200	0.31009	-0.20809	-0.53	-0.02	0.66	0.86	-0.79874	-1
10	MODEL1	RIDGEVIF	CholesterolLoss	.00008	.	.	.	1.88130	2.35861	19.75	2.30	3.64	17.10	1.79562	-1
11	MODEL1	RIDGE	CholesterolLoss	.00008	.	26.4508	41.7441	0.30972	-0.20799	-0.36	-0.02	0.49	0.69	-0.79930	-1
12	MODEL1	RIDGEVIF	CholesterolLoss	.00010	.	.	.	1.88105	2.35822	12.77	2.30	3.05	11.20	1.79542	-1
13	MODEL1	RIDGE	CholesterolLoss	.00010	.	26.4513	41.7589	0.30947	-0.20793	-0.26	-0.02	0.39	0.59	-0.79965	-1
14	MODEL1	RIDGEVIF	CholesterolLoss	.00012	.	.	.	1.88080	2.35783	8.98	2.29	2.73	7.99	1.79521	-1
15	MODEL1	RIDGE	CholesterolLoss	.00012	.	26.4517	41.7689	0.30929	-0.20788	-0.19	-0.02	0.32	0.52	-0.79988	-1
16	MODEL1	RIDGEVIF	CholesterolLoss	.00014	.	.	.	1.88055	2.35744	6.69	2.29	2.54	6.05	1.79500	-1
17	MODEL1	RIDGE	CholesterolLoss	.00014	.	26.4519	41.7764	0.30915	-0.20784	-0.14	-0.02	0.27	0.47	-0.80006	-1

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

Combating Multicollinearity

Ridge Regression Example

- Choose your alternative estimator
  - Pick the smallest value of k that process a stable estimate of β
  - Get the variance inflation factors (VIF) close to 1

```
proc reg data=health outvif plots(only)=ridge(unpack VIFaxis=log)
    outest=rrhealth_final ridge=0.00012;
    model cholesterolloss = age weight cholesterol triglycerides hdl ldl height;
    plot / ridgeplot nomodel nostat;
    title 'Health - Ridge Regression Calculation';
run;

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


SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

Combating Multicollinearity

Ridge Regression Example

Obs	_MODEL_	_TYPE_	_DEPVAR_	_RIDGE_	_PCOMIT_	_RMSE_	Intercept	Age	Weight	Cholesterol	Triglycerides	HDL	LDL	Height	CholesterolLoss
1	MODEL1	PARMS	CholesterolLoss	.	.	26.0275	18.3859	0.63264	-0.29825	-169.201	2.67536	169.192	169.525	-0.26426	-1
2	MODEL1	RIDGEVIF	CholesterolLoss	.00012	.	.	.	1.88080	2.35783	8.980	2.29088	2.734	7.988	1.79521	-1
3	MODEL1	RIDGE	CholesterolLoss	.00012	.	26.4517	41.7689	0.30929	-0.20788	-0.192	-0.02197	0.321	0.520	-0.79988	-1

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

Combating Multicollinearity

Ridge Regression Example

Health - Ridge Regression Calculation

The REG Procedure
Model: MODEL1
Dependent Variable: CholesterolLoss

Number of Observations Read	95
Number of Observations Used	43
Number of Observations with Missing Values	52

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	8347.58570	1192.51224	1.76	0.1270
Error	35	23710	677.43111		
Corrected Total	42	32058			

Root MSE	26.02751	R-Square	0.2604
Dependent Mean	9.76744	Adj R-Sq	0.1125
Coeff Var	266.47209		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	18.38590	86.45275	0.21	0.8328
Age	1	0.63264	1.68351	0.38	0.7093
Weight	1	-0.29825	0.24873	-1.20	0.2385
Cholesterol	1	-169.20149	157.59569	-1.07	0.2903
Triglycerides	1	2.67536	2.51627	1.06	0.2950
HDL	1	169.19195	157.46718	1.07	0.2900
LDL	1	169.52519	157.59200	1.08	0.2894
Height	1	-0.26426	1.45480	-0.18	0.8569

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

Combating Multicollinearity

Ridge Regression Example

- Modify Output for Interpretation
  - Standard errors (SEB)
  - Parameter Estimates

```
proc reg data=health outvif plots(only)=ridge(unpack VIFaxis=log)
    outest=rrhealth_final outseb ridge=0.00012;
    model cholesterolloss = age weight cholesterol triglycerides hdl ldl height;
    plot / ridgeplot nomodel nostat;
    title 'Health - Ridge Regression Calculation';
run;

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


SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

Combating Multicollinearity

Ridge Regression Example

Before `outseb`

Obs	_MODEL_	_TYPE_	_DEPVAR_	_RIDGE_	_PCOMIT_	_RMSE_	Intercept	Age	Weight	Cholesterol	Triglycerides	HDL	LDL	Height	CholesterolLoss
1	MODEL1	PARMS	CholesterolLoss	.	.	26.0275	18.3859	0.63264	-0.29825	-169.201	2.67536	169.192	169.525	-0.26426	-1
2	MODEL1	RIDGEVIF	CholesterolLoss	.00012	.	.	.	1.88080	2.35783	8.980	2.29088	2.734	7.988	1.79521	-1
3	MODEL1	RIDGE	CholesterolLoss	.00012	.	26.4517	41.7689	0.30929	-0.20788	-0.192	-0.02197	0.321	0.520	-0.79988	-1

After `outseb`

Obs	_MODEL_	_TYPE_	_DEPVAR_	_RIDGE_	_PCOMIT_	_RMSE_	Intercept	Age	Weight	Cholesterol	Triglycerides	HDL	LDL	Height	CholesterolLoss
1	MODEL1	PARMS	CholesterolLoss	.	.	26.0275	18.3859	0.63264	-0.29825	-169.201	2.67536	169.192	169.525	-0.26426	-1
2	MODEL1	SEB	CholesterolLoss	.	.	26.0275	86.4527	1.68351	0.24873	157.596	2.51627	157.467	157.592	1.45480	-1
3	MODEL1	RIDGEVIF	CholesterolLoss	.00012	.	.	.	1.88080	2.35783	8.980	2.29088	2.734	7.988	1.79521	-1
4	MODEL1	RIDGE	CholesterolLoss	.00012	.	26.4517	41.7689	0.30929	-0.20788	-0.192	-0.02197	0.321	0.520	-0.79988	-1
5	MODEL1	RIDGESEB	CholesterolLoss	.00012	.	26.4517	85.0039	1.68266	0.23774	0.328	0.07522	0.624	0.336	1.38822	-1

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

Conclusion

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

Summary

- When multicollinearity is present in data
  - Ordinary least squares estimators are imprecisely estimated
  - This could result in misleading or improper conclusions
- If your goal is to understand how your predictors impact your outcome
  - Then multicollinearity poses a problem
  - Therefore, it is essential to detect and solve this issue before estimating the parameters based on the fitted regression model
- The detection of multicollinearity is important

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

Conclusions

- Once multicollinearity is detected
  - Necessary to introduce appropriate changes in model specification to combat
- Remedial measures can help solve this problem
  - Removing a variable
  - Principal Component Regression
  - Regularization Techniques
    - L1: Lasso Regression
    - L2: Ridge Regression

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

Conclusions

- Remember the Trade-Off?
  - Ridge Regression is still an adjustment
  - Naturally results in biased outcomes
- Elastic Nets / Bootstrapping
  - Could help resolve L1/L2 debate
  - Could help address adjustment concerns

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

Thank You!!

Name: Deanna Schreiber-Gregory

Organization: Henry M Jackson Foundation

Title: Data Analyst, Research Associate

Location: Bethesda, MD

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

On LinkedIn

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.