SAS® GLOBAL FORUM 2019

USERS PROGRAM

APRIL 28 – MAY 1, 2019 | DALLAS, TX

Regularization Techniques for Multicollinearity: Lasso, Ridge, and Elastic Nets

Deanna (DeDe) N Schreiber-Gregory

Henry M Jackson Foundation for the Advancement of Military Medicine

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

Overview

- Definition of Multicollinearity
- The Dataset
- Detecting Multicollinearity
- Combating Multicollinearity
  - LASSO Regression
  - Ridge Regression
  - Elastic Net

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

The Dataset

The Dataset

- 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 Dataset

➢ The Example

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

Detecting Multicollinearity

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

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

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

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

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; run;
```

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

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;
run;
```

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

Detecting Multicollinearity

Example

- Note: VIF cut point = 10, Tolerance cut point = 0.1

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

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

Combating Multicollinearity

Overview

Combating Multicollinearity

What Can We Do?

- Easiest
  - 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
    - Elastic Net

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, $\lambda_1 \geq \lambda_2 \ldots\ldots\ldots \geq \lambda_p > 0$
  - 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

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
  - Elastic Nets

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 ʎ =0 then the equation will go back to OLS estimations
  - If ʎ is very large, too much weight would be added = under-fitting
  - NOTE: need to be careful with choice of ʎ

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 grouped selection
  - 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

- Some Trade-Offs
  - We are still dealing with an adjustment
  - Naturally results in biased outcomes

Combating Multicollinearity

Elastic Nets

- Elastic Net
  - Balances having parsimonious model
    - Borrows strength from correlated regressors
    - Constraints on sum of absolute value of magnitude of the coefficient
    - Constraints on sum of the squared coefficient

$$\sum_{i=1}^{n}\left(Y_i - \sum_{j=1}^{p} X_{ij}\beta_j\right)^2 + \lambda\left(\propto \sum_{j=1}^{p}|\beta_j| + (1-\propto)\sum_{j=1}^{p}\beta_j^2\right)$$

Combating Multicollinearity

LASSO Regression

Combating Multicollinearity

LASSO Regression Example

LASSO: Least Absolute Shrinkage and Selection Operator

- Logic
  - Constrained form of ordinary least squares regression
  - Sum of the absolute values of the regression coefficients is constrained to be smaller than a specified parameter
  - Does not punish high values of the coefficients β
    - Instead, figures out which values are irrelevant and sets them to zero
  - Results in fewer features being included in the final model

- LASSO Variants
  - Early implementations used quadratic programming techniques
    - LAR (Least Angle Regression)

Combating Multicollinearity

LASSO Regression Example

- Applying LASSO Regression
  - Can do through GLMSelect (or Proc Hpreg)
  - Specify criterion to choose among models at each step (CHOOSE =)
    - LASSO, LAR, LSCOEFFS
  - Can specify stopping criterion (STOP =)

```
/* Lasso Selection */
proc glmselect data=health plots=all;
  model cholesterolloss = age weight cholesterol
  triglycerides hdl ldl height skinfold systolicbp
  diastolicbp exercise coffee
   selection=lar (choose=cv stop=none) cvmethod=random(10);
   title 'Health - Lasso Regression Calculation';
run;
```

Combating Multicollinearity

LASSO Regression Example

Health - Lasso Regression Calculation

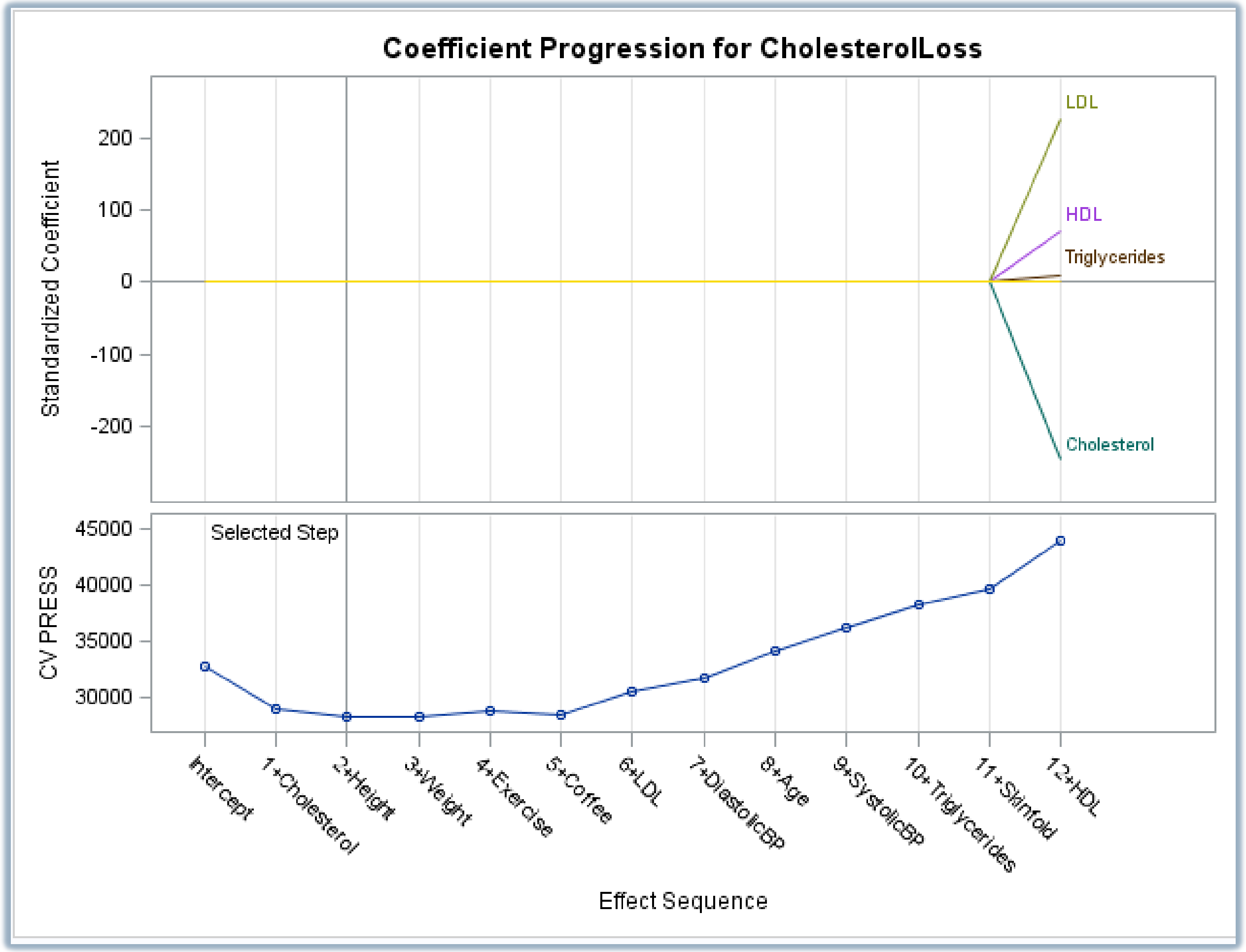
The GLMSELECT Procedure

LAR Selection Summary

Step	Effect Entered	Number Effects In	CV PRESS
0	Intercept	1	32892.3215
1	Cholesterol	2	29023.4028
2	Height	3	28345.4812*
3	Weight	4	28353.9319
4	Exercise	5	28918.5462
5	Coffee	6	28513.9674
6	LDL	7	30648.6581
7	DiastolicBP	8	31819.5480
8	Age	9	34142.8348
9	SystolicBP	10	36309.8315
10	Triglycerides	11	38241.8217
11	Skinfold	12	39651.2277
12	HDL	13	44039.3374

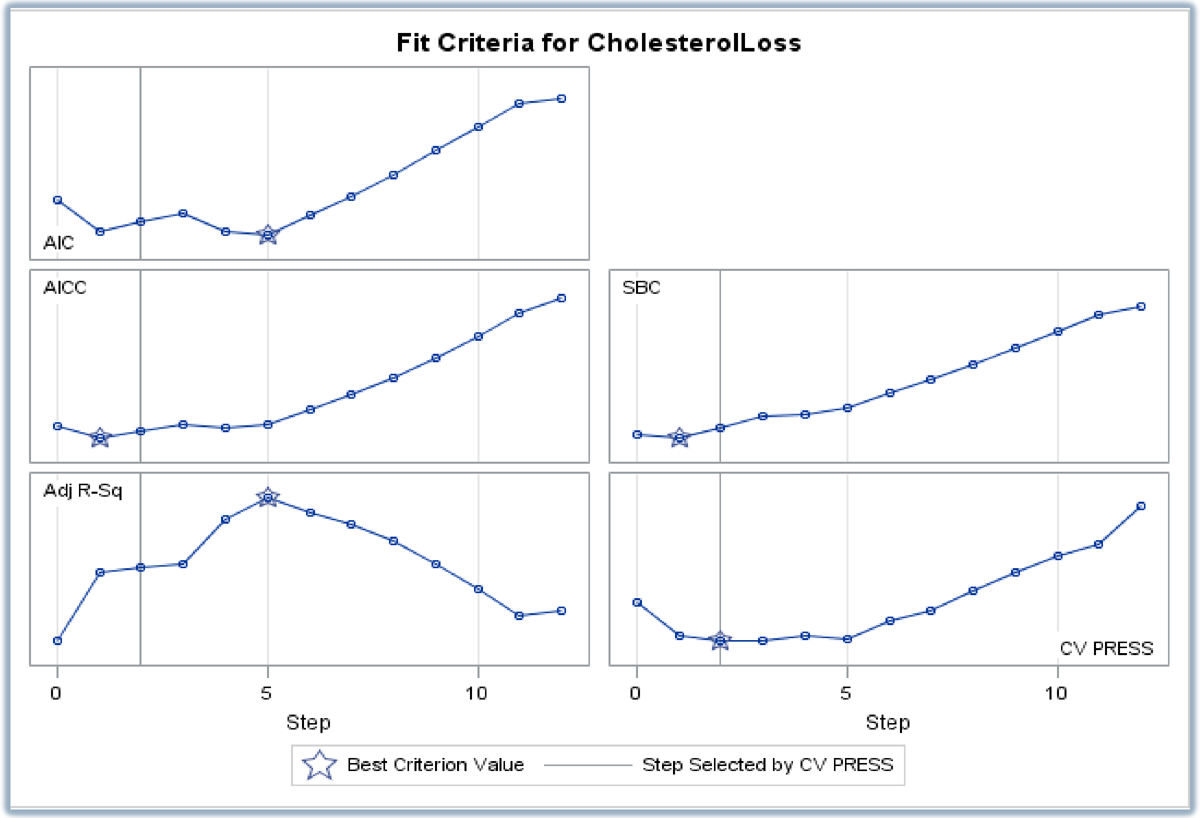
* Optimal Value of Criterion

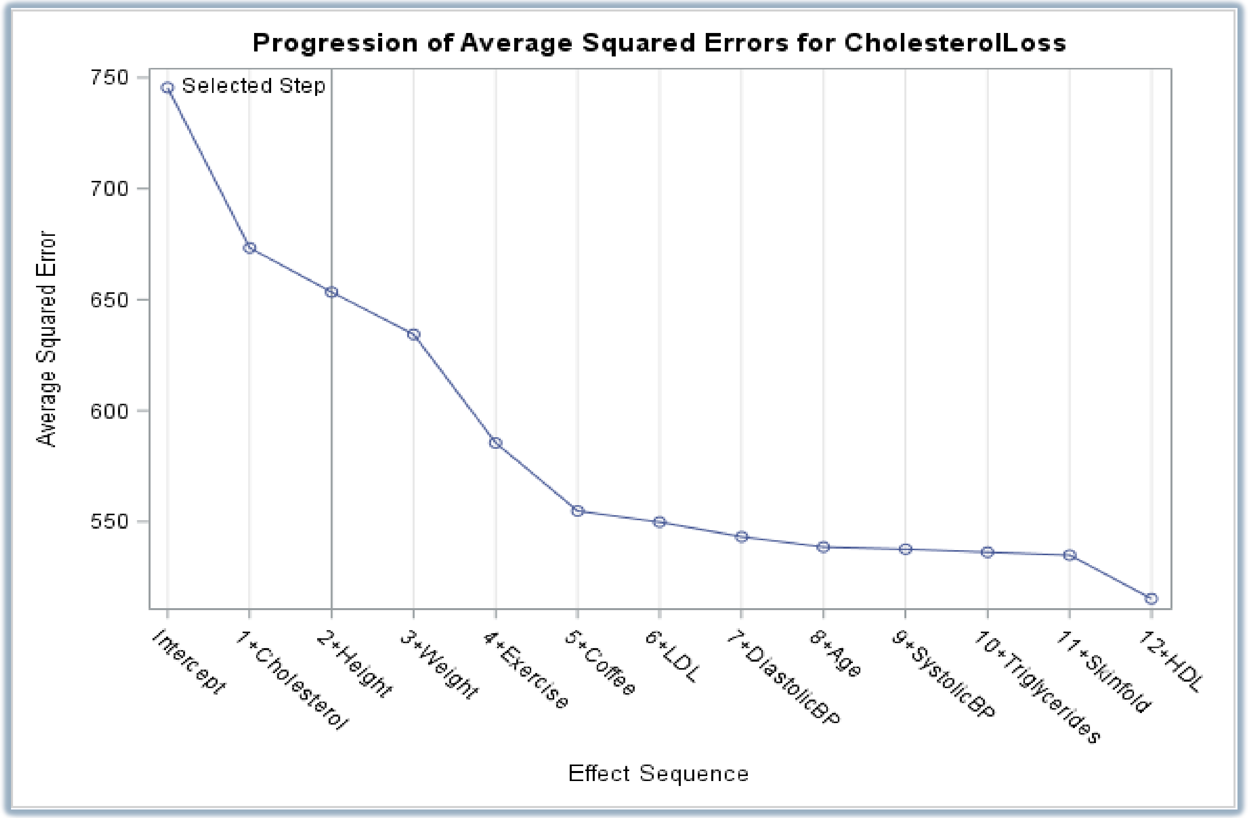
Selection stopped because all effects are in the final model.



Combating Multicollinearity

LASSO Regression Example





Combating Multicollinearity

LASSO Regression Example

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	2	3963.41962	1981.70981	2.82
Error	40	28094	702.35637	
Corrected Total	42	32058		

Root MSE	26.50201
Dependent Mean	9.76744
R-Square	0.1236
Adj R-Sq	0.0798
AIC	329.73117
AICC	330.78380
SBC	290.01477
CV PRESS	28345

Parameter Estimates		
Parameter	DF	Estimate
Intercept	1	-1.388985
Cholesterol	1	0.129281
Height	1	-0.194803

Combating Multicollinearity

Ridge Regression

Combating Multicollinearity

Ridge Regression

- 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

Combating Multicollinearity

Ridge Regression

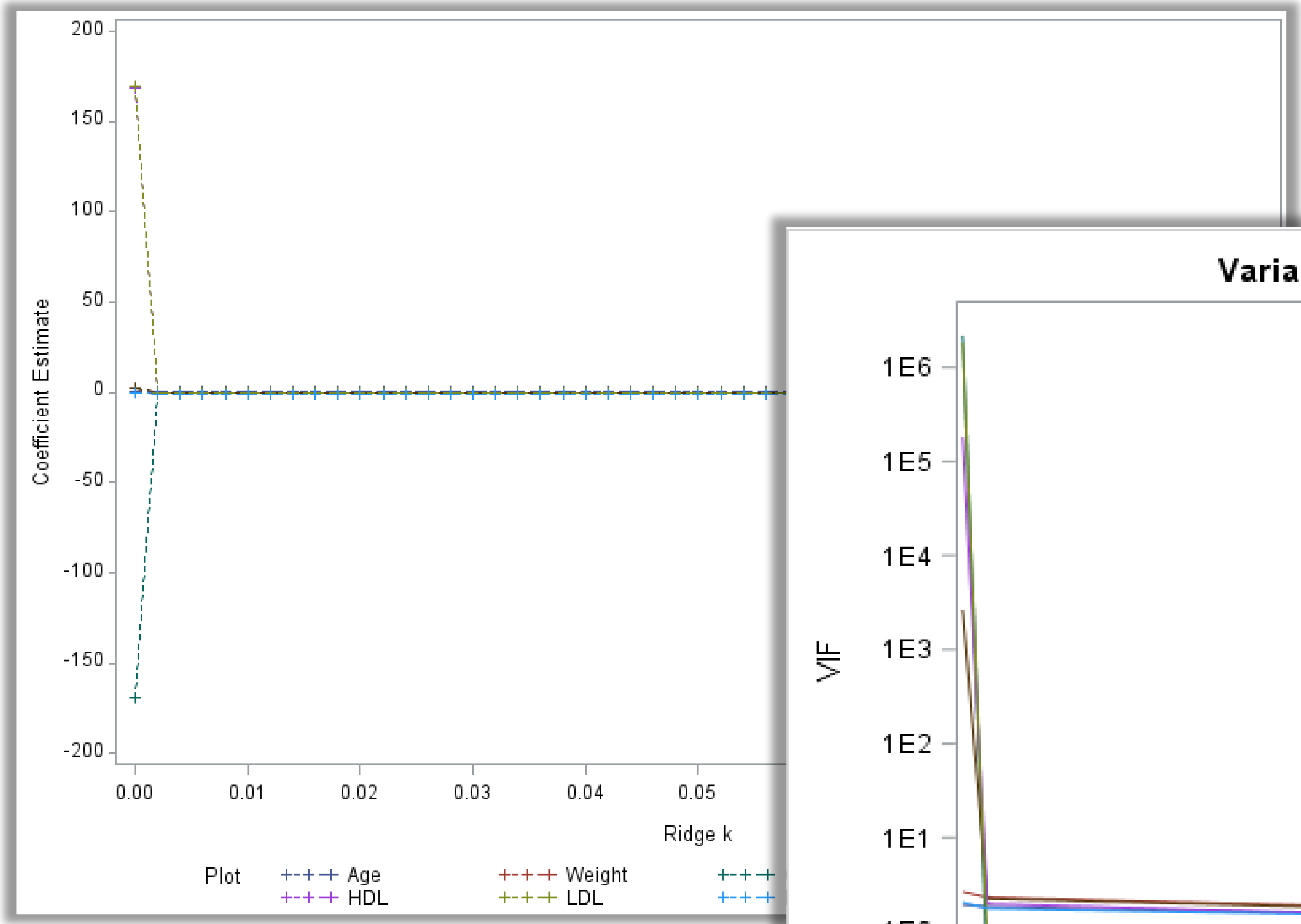
- Ridge Regression for alternative estimator
  - The value of k can be estimated by looking at a ridge trace plot
  - Ridge trace plots are plots of parameter estimates vs k where k usually lies in the interval [0,1]
  - Note:
    - Pick the smallest value of k that produces a stable estimate of β
    - Get the variance inflation factors (VIF) close to 1

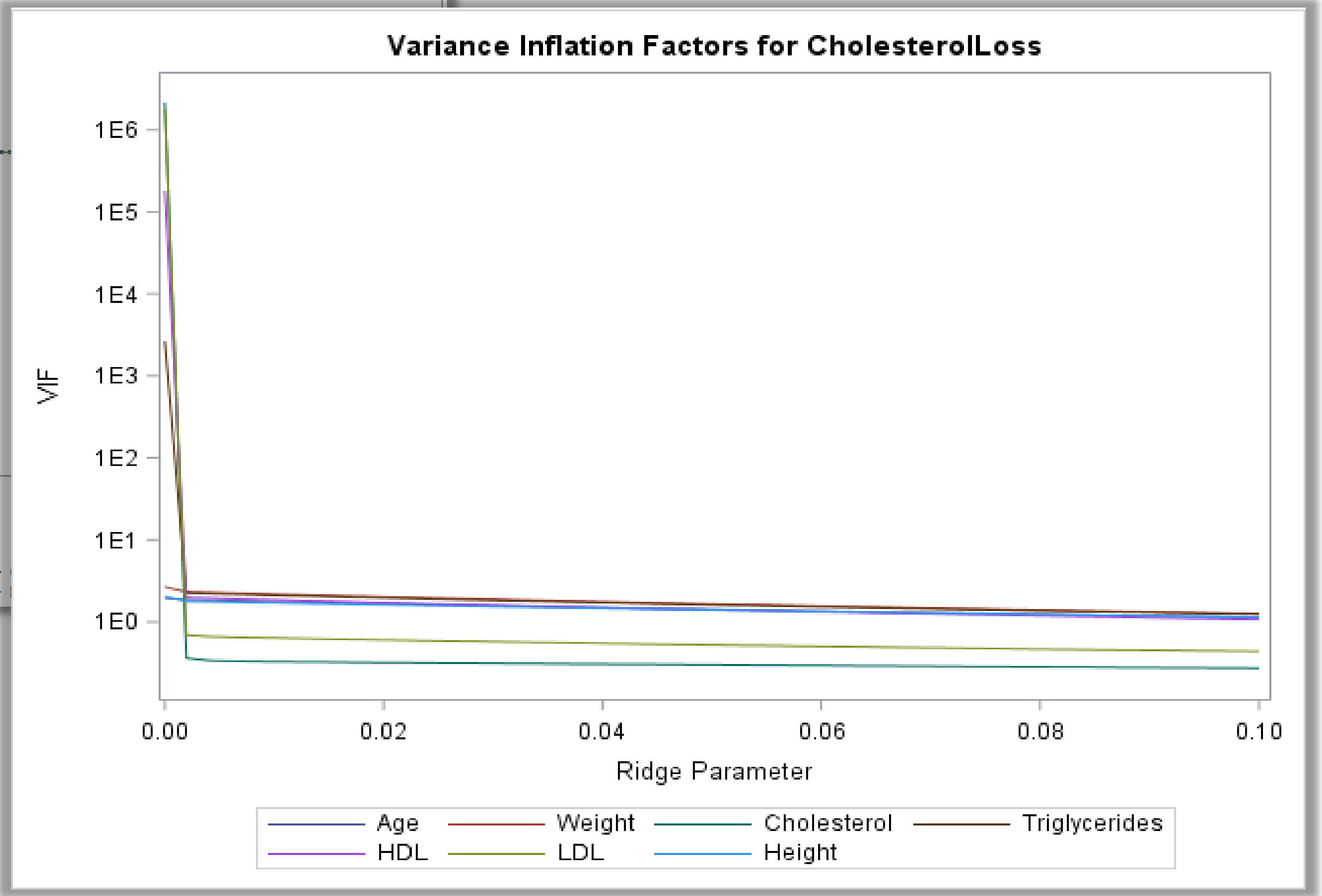
Combating Multicollinearity

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;
```





USERS PROGRAM

Combating Multicollinearity

Ridge Regression Example

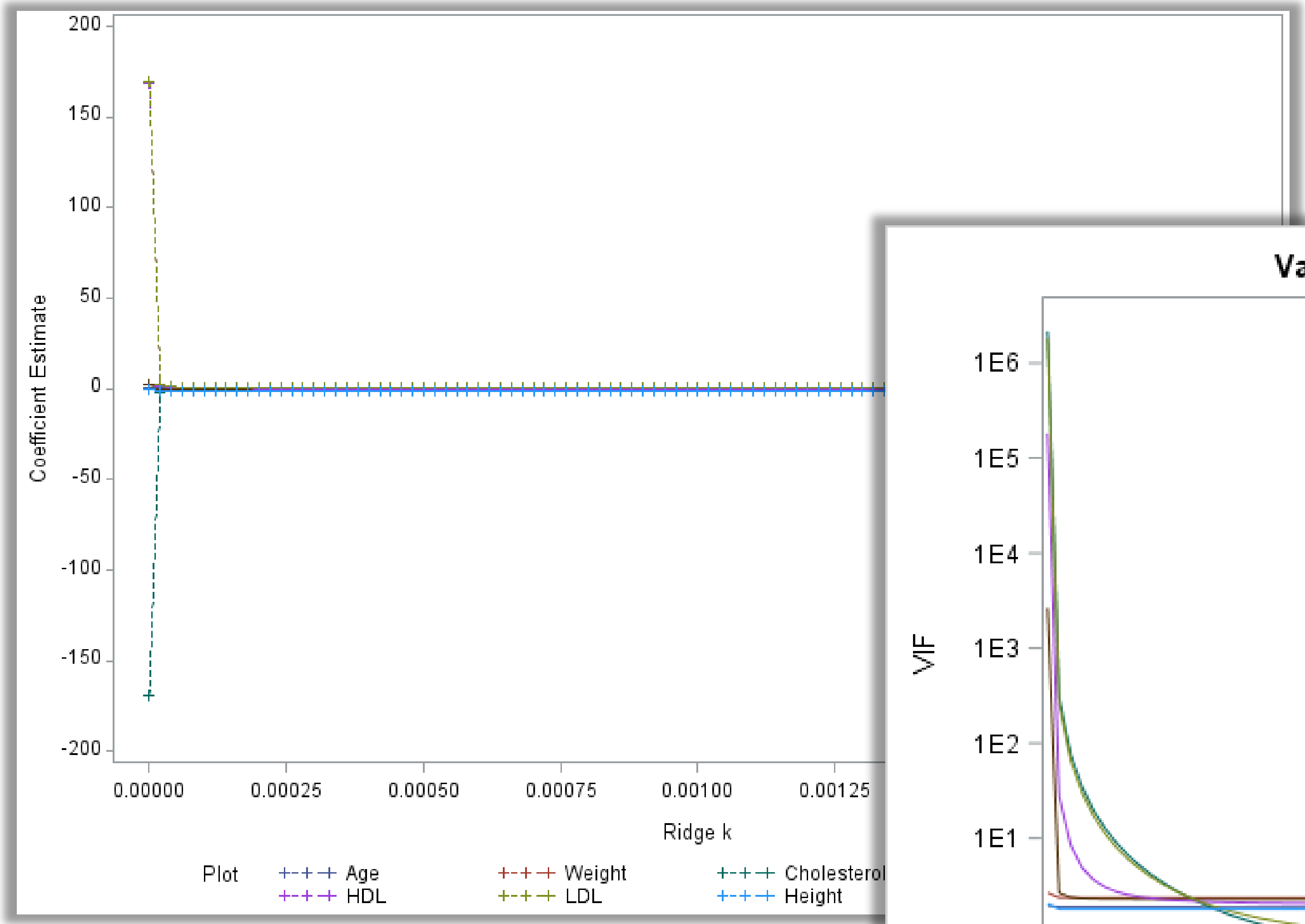
Obs	_MODEL_	_TYPE_	_DEPVAR_	_RIDGE_	_PCOMIT_	_RMSE_	Intercept	Age	Weight	Cholesterol	Triglycerides	HDL	LDL	Height	CholesterolLoss
1	MODEL1	PARMS	CholesterolLoss	.	.	26.0275	18.3859	0.63264	-0.29825	-169.20	2.68	169.19	169.53	-0.26426	-1
2	MODEL1	RIDGEVIF	CholesterolLoss	0.000	.	.	.	1.94457	2.66571	2144274.02	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

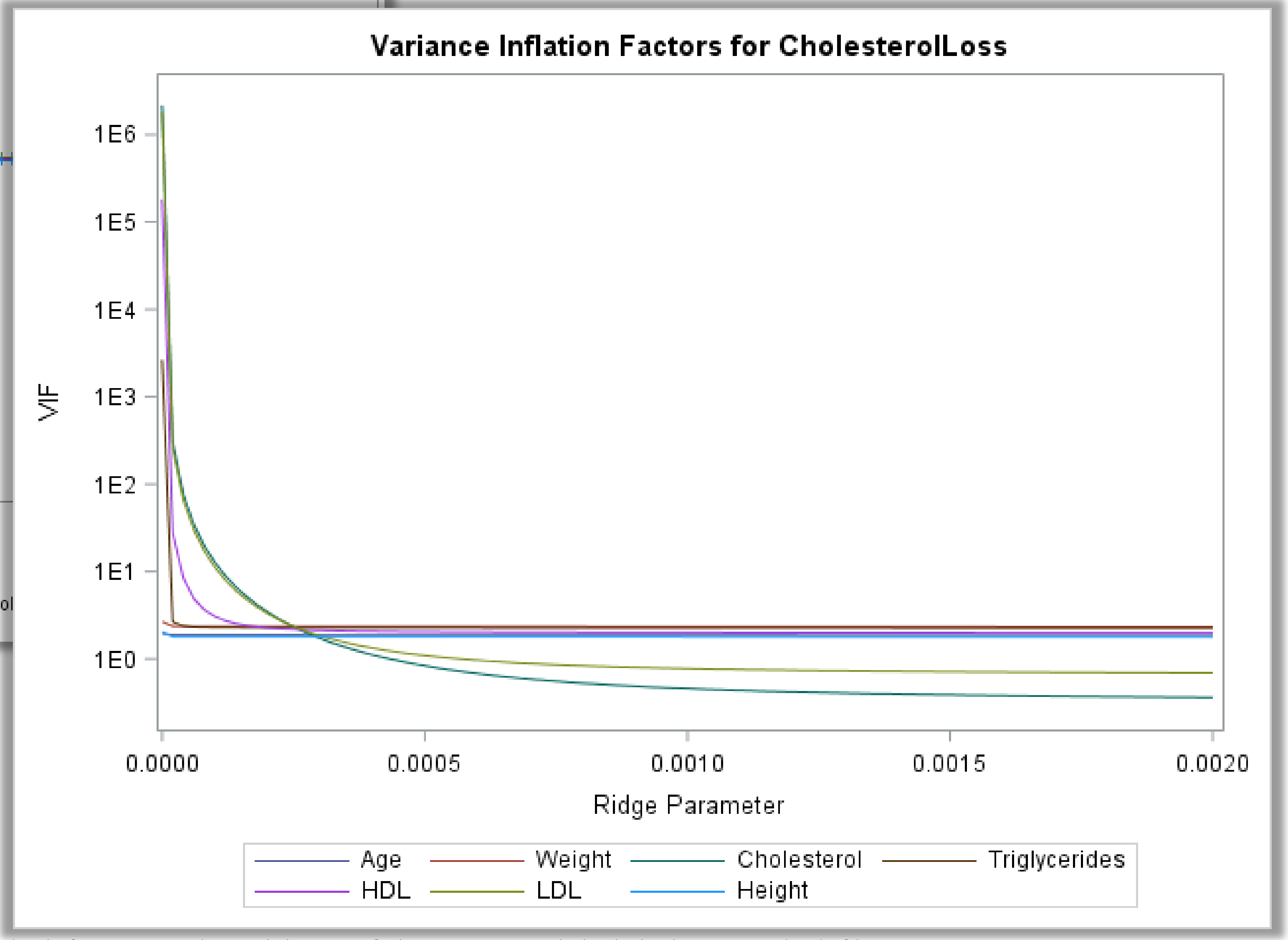
Combating Multicollinearity

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;
```





USERS PROGRAM

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

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;
```

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

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

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;
```

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

Combating Multicollinearity

Elastic Net

Combating Multicollinearity

Elastic Net Regression

- Logic
  - Both LASSO and Ridge pros and cons
  - Elastic Net attempts to take the best features of these two procedures and use them at the same time

Combating Multicollinearity

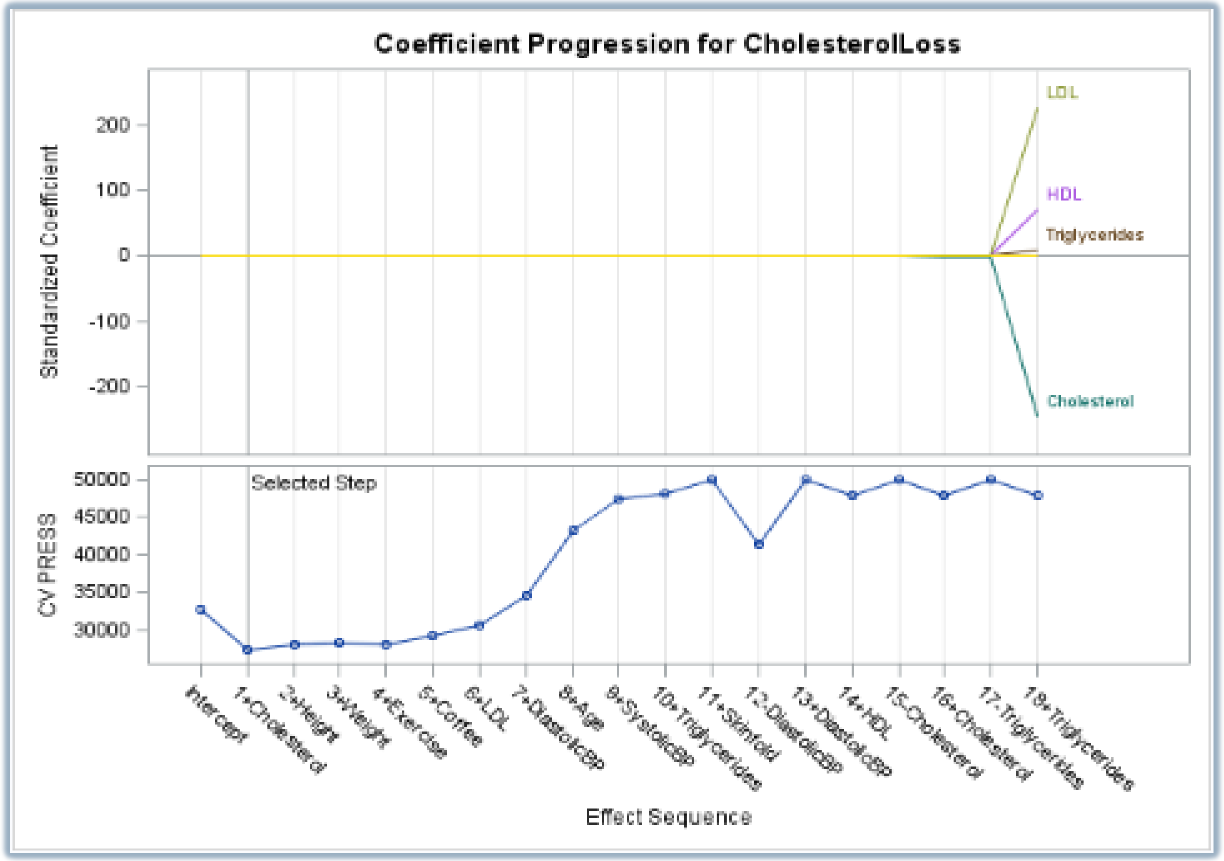
Elastic Net Example

- Similar options to LASSO
- STEPS = specifies number selection steps to be performed
- L2 = specifies value of ridge parameter

```
/* Elastic Net */
proc glmselect data=health plots=coefficients;
   model cholesterolloss = age weight cholesterol triglycerides
   hdl ldl height skinfold systolicbp diastolicbp exercise coffee
   / selection=elasticnet(steps=120 choose=cv) cvmethod=split(4);
  title 'Health - Elastic Net Regression Calculation';
run;
```

Combating Multicollinearity

Elastic Net Example



Elastic Net Selection Summary

Step	Effect Entered	Effect Removed	Number Effects In	CV PRESS
0	Intercept		1	32755.1101
1	Cholesterol		2	27348.8251*
2	Height		3	27971.7591
3	Weight		4	28235.2852
4	Exercise		5	28030.3185
5	Coffee		6	29198.0055
6	LDL		7	30535.9524
7	DiastolicBP		8	34489.6987
8	Age		9	43119.5067
9	SystolicBP		10	47414.7455
10	Triglycerides		11	48196.2793
11	Skinfold		12	49928.4795
12		DiastolicBP	11	41468.3373
13	DiastolicBP		12	49928.4795
14	HDL		13	47896.2578
15		Cholesterol	12	49908.5770
16	Cholesterol		13	47896.2619
17		Triglycerides	12	49870.0391
18	Triglycerides		13	47896.2116

* Optimal Value of Criterion

Selection stopped because all effects are in the final model.

Combating Multicollinearity

Elastic Net Example

The selected model, based on Cross Validation, is the model at Step 1.

Effects	Intercept Cholesterol

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	1	3104.61924	3104.61924	4.40
Error	41	28953	706.17208	
Corrected Total	42	32058		

Root MSE	26.57390
Dependent Mean	9.76744
R-Square	0.0968
Adj R-Sq	0.0748
AIC	329.02593
AICC	329.64131
SBC	287.54833
CV PRESS	27349

Parameter Estimates		
Parameter	DF	Estimate
Intercept	1	-11.027658
Cholesterol	1	0.108716

Combating Multicollinearity

Comparing LASSO, Ridge, and Elastic Net

Combating Multicollinearity

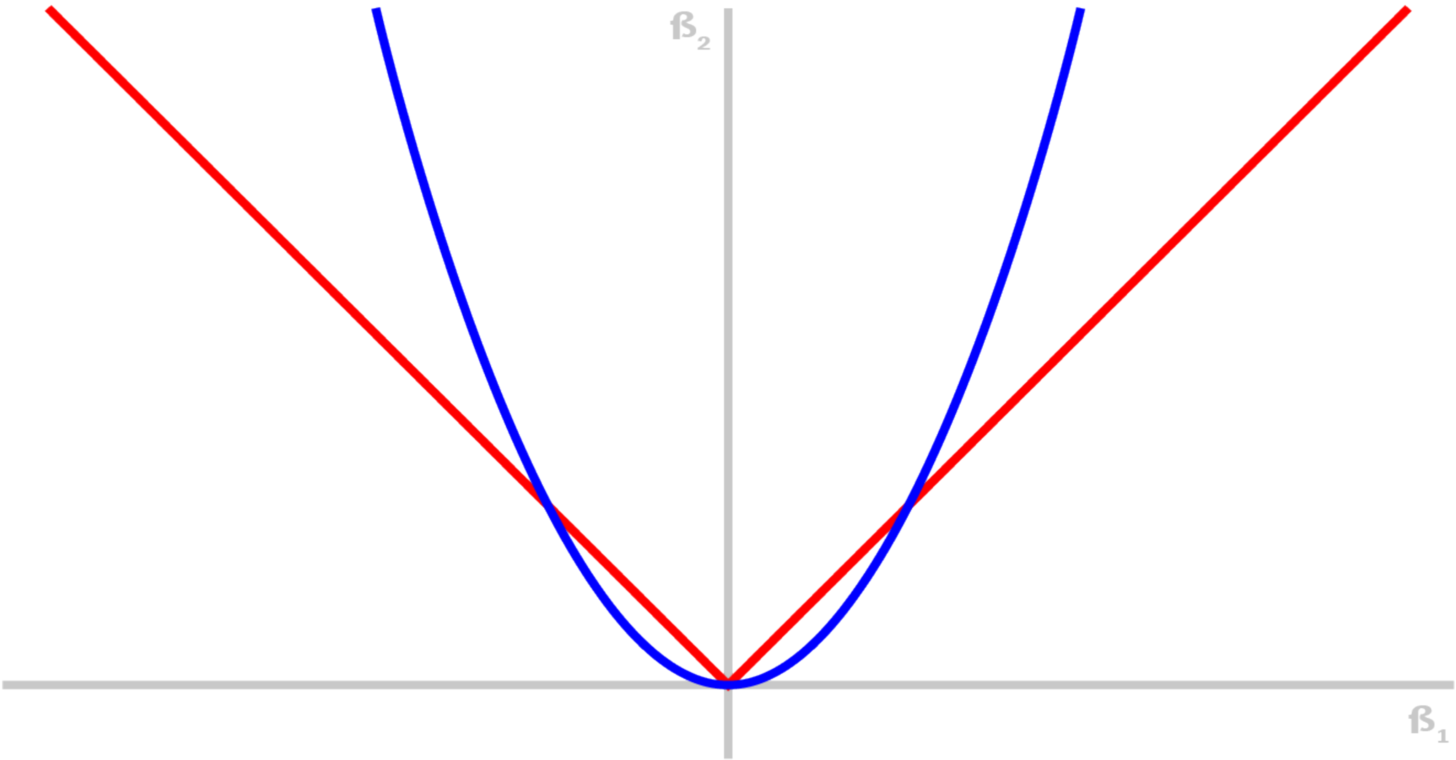
LASSO Regression Advantage/Disadvantage

- LASSO Advantages
  - Great if goal is to reduce the number of variables
  - It enforces sparcity in parameter selection and inclusion
  - Does have a quadratic programming problem, but can be solved through use of LAR solution or other approaches
- LASSO Disadvantages
  - If group of predictors are highly correlated, LASSO tends to pick only one of them and will shrink the others to zero
  - LASSO can not perform grouped selection

Combating Multicollinearity

LASSO Regression

- LASSO regression adjustment
- Linear regression



Combating Multicollinearity

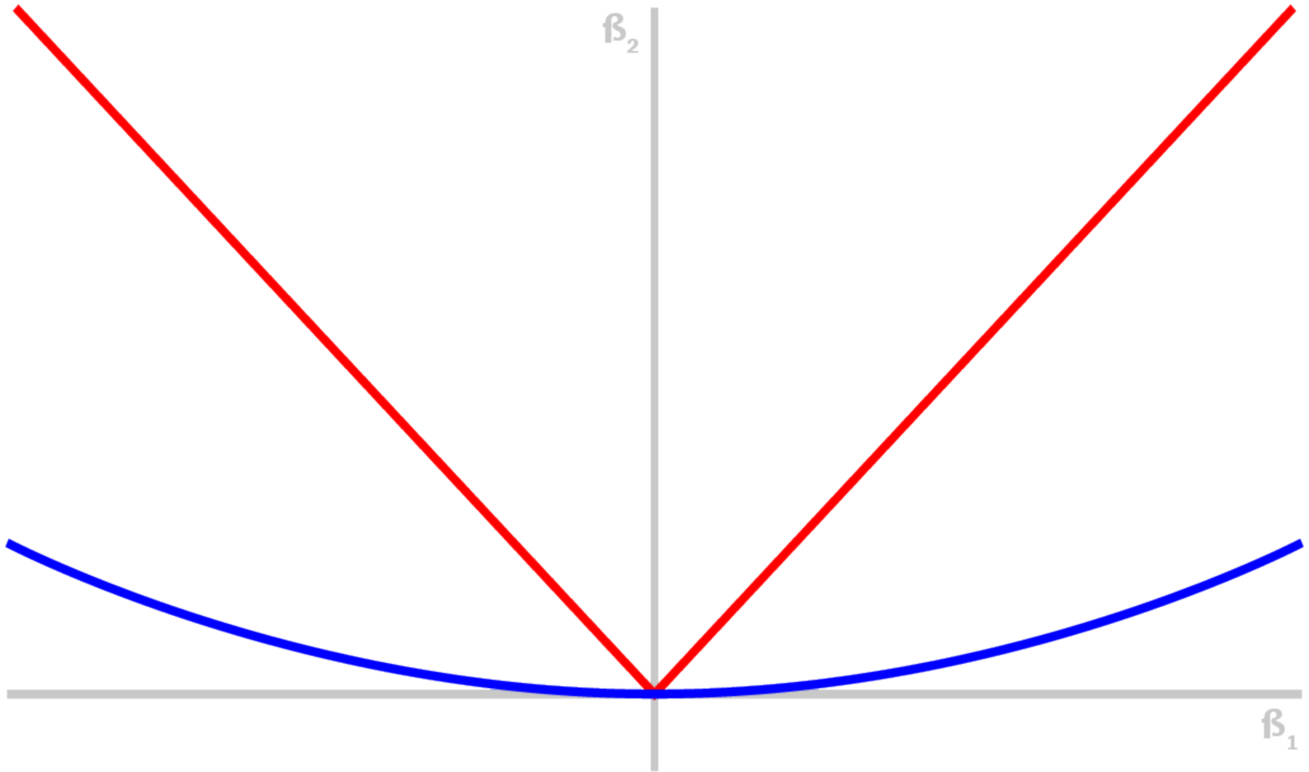
Ridge Regression Advantage/Disadvantage

- Ridge Advantages
  - It is great if your goal is to adjust for multicollinearity with grouped selections
  - Produces biased but smaller variance and smaller Mean Square Error (MSE)
  - Results in the explicit solution
- Ridge Disadvantages
  - Aforementioned biased results
  - Tends to shrink coefficients to near zero but can not produce a parsimonious model

Combating Multicollinearity

Ridge Regression

- Ridge regression adjustment
- Linear regression



Combating Multicollinearity

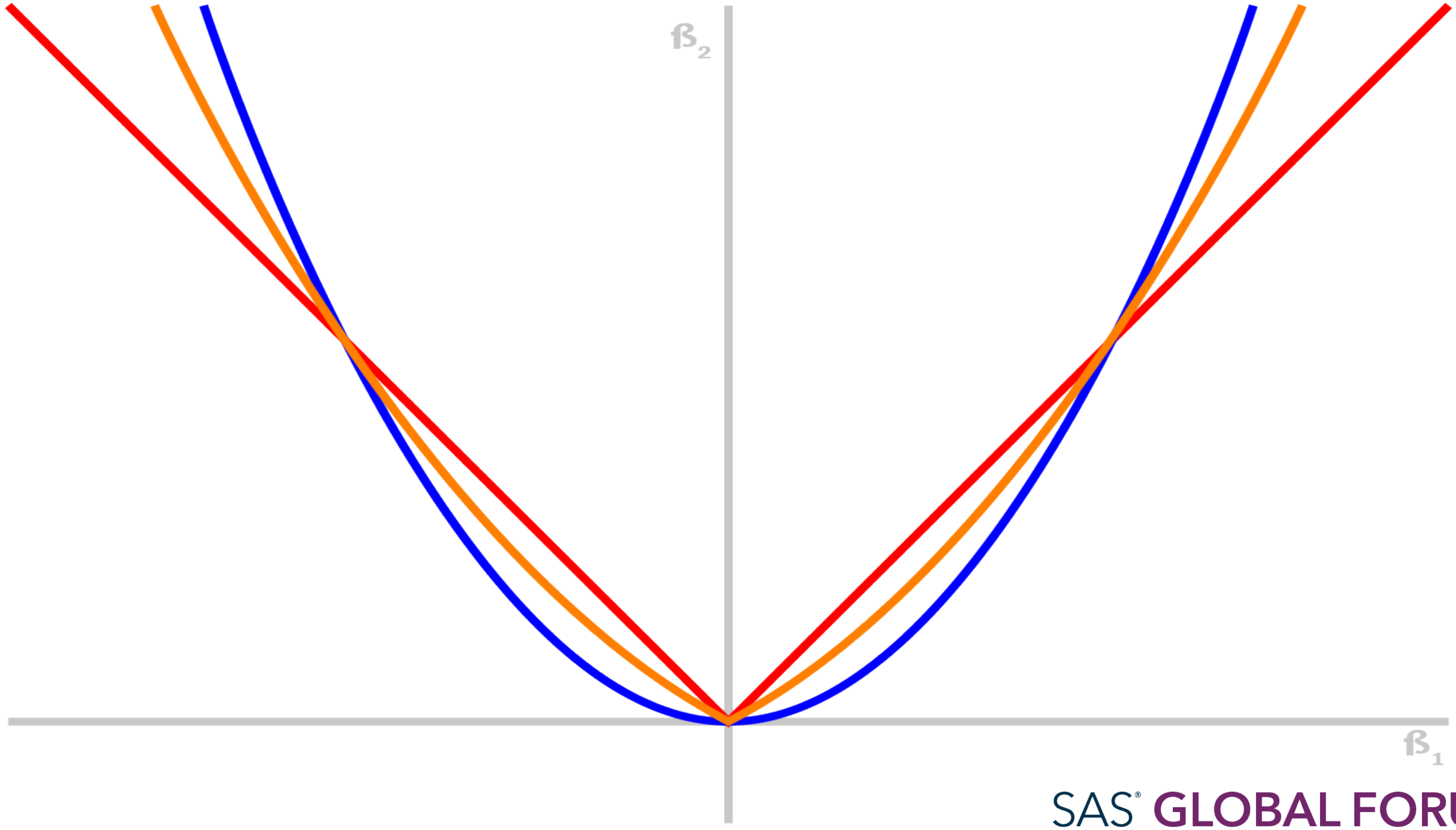
Elastic Net Advantage/Disadvantage

- Elastic Net Advantages
  - Enforce Sparsity
  - Has no limitation on the number of selected variables
  - Encourages a grouping effect in the presence of highly correlated predictors
- Elastic Net Disadvantages
  - Naïve elastic net can suffer from double shrinkage
    - Needs to be carefully employed

Combating Multicollinearity

LASSO / Ridge / Elastic Net

- Ridge regression adjustment
- LASSO regression
- Elastic Net



Conclusion

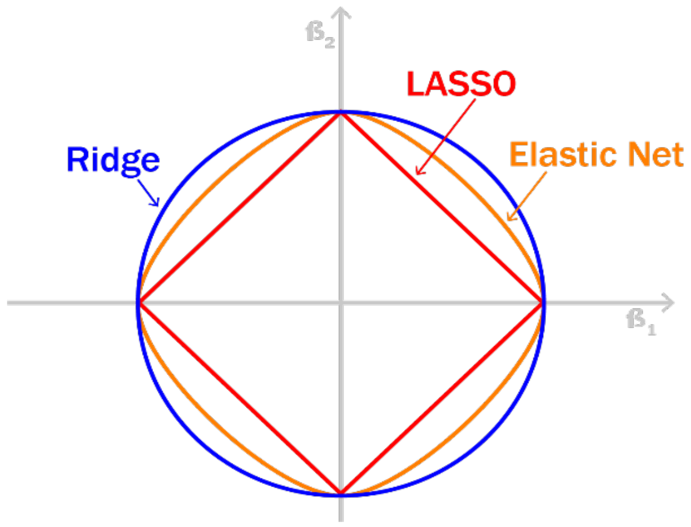
Summary

- 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

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
    - Elastic Net



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

Name: Karlen Bader
Organization: Henry M Jackson Foundation
Title: Research Assistant
Location: Bethesda, MD


#SASGF
SAS®
GLOBAL
FORUM
2019
APRIL 28 – MAY 1, 2019 | DALLAS, TX
Kay Bailey Hutchison Convention Center