Paper 131-2018

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

Deanna Schreiber-Gregory, Henry M Jackson Foundation

ABSTRACT

Multicollinearity can be briefly described as the phenomenon in which two or more identified predictor variables are linearly related, or codependent. The presence of this phenomenon can have a negative impact on an analysis as a whole and can severely limit the conclusions of a research study. In this paper, we will briefly review how to detect multicollinearity, and once it is detected, which regularization techniques would be the most appropriate to combat it. The nuances and assumptions of R1 (Lasso), R2 (Ridge Regression), and Elastic Nets will be covered in order to provide adequate background for appropriate analytic implementation. This paper is intended for any level of SAS® user. This paper is also written to an audience with a background in theoretical and applied statistics, though the information within will be presented in such a way that any level of statistics/mathematical knowledge will be able to understand the content.

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 can occur in regression when several predictors are highly correlated. (As a disclaimer, variables do not need to be highly correlated for multicollinearity to exist, though this is oftentimes the case.) Another way to think of collinearity is as a type of variable “co-dependence”.

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 the 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!

A LINEAR EXAMPLE

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

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

In this equation, *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 $\varepsilon i \sim NID(0,\sigma\text{^}2)$. In a model such as this, the presence of multicollinearity would inflate the variances of the parameter estimates, leading to a lack of statistical significance of the individual predictor variables even if the overall model itself remains significant. Considering this, we can see how the presence of multicollinearity can end up causing serious problems when estimating and interpreting β, even in the simplest of equations.

A LIVING EXAMPLE

Why should we care? Consider this example: your company has just undergone a major overhaul and it was decided that half of the department heads would choose an assistant lead to help with their workload. The assistant leads were chosen by the identified department heads 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 heads and their new assistant leads are present. Keep in mind that only half of the departments have two representatives, whereas the other half only has one representative per department. It comes time to vote, by show of hands, on a major budget revision. Both the leads and assistants 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, thus giving favor to the departments with two representatives. 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.

DIFFERENT MODELS, DIFFERENT CIRCUMSTANCES

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 coefficients 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

We will begin by exploring 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 DATSET

The dataset used for this paper 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).

DATA CLEANING AND PREPARATION

As a first step in the examination of our research question − do target health outcome variables contribute to the amount of cholesterol lost between baseline and a 3 month follow-up − we must first identify which variables will be used in the analysis, what these variables look like, and how these variables will interact with each other. In short, we must clean and prepare the data for our analysis. This may seem redundant, but it is a worthy note to make considering the type of analysis we are about to conduct. We will begin by identifying the dataset and making sure that it is appropriately imported into the SAS environment. At this time we will also use the CONTENTS procedure to check the structure and types of variables we will be working with:

```
/* 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;
```

Next, frequency, means, and univariate procedures were performed in order to explore the descriptive statistics, skewness, and kurtosis of our target outcome and predictor variables within the dataset and to identify any possible errors, outliers, and missing information that may exist.

```
/* Exploration of Skewness and Kurtosis */
proc univariate data= health;
var age weight cholesterol triglycerides hdl ldl height skinfold
systolicbp diastolicbp exercise coffee cholesterolloss;
probplot age weight cholesterol triglycerides hdl ldl height skinfold
systolicbp diastolicbp exercise coffee cholesterolloss / normal (mu=est
sigma=est) square;
run;
```

In the above code, the chisq option is indicated in the table statement of the FREQ procedure to receive chi-square test results in the output. In the UNIVARIATE procedure, the normal option is used to request tests for normality and goodness-of-fit, mu is used to indicate the value of the mean or location parameter for the normal curve, sigma is used to specify the standard deviation for the normal curve, and square is used to display a P-P plot in the square format.

If we need to correct for any errors, skewness, kurtosis, or control for missing values, we would complete those at this time before we construct our final data tables for descriptive and univariate analyses. Once we have corrected for our errors and missing data, we can then rerun these procedures (minus our outcome variable) with our corrected values to look at the univariate relationships between our scrubbed predictor variables.

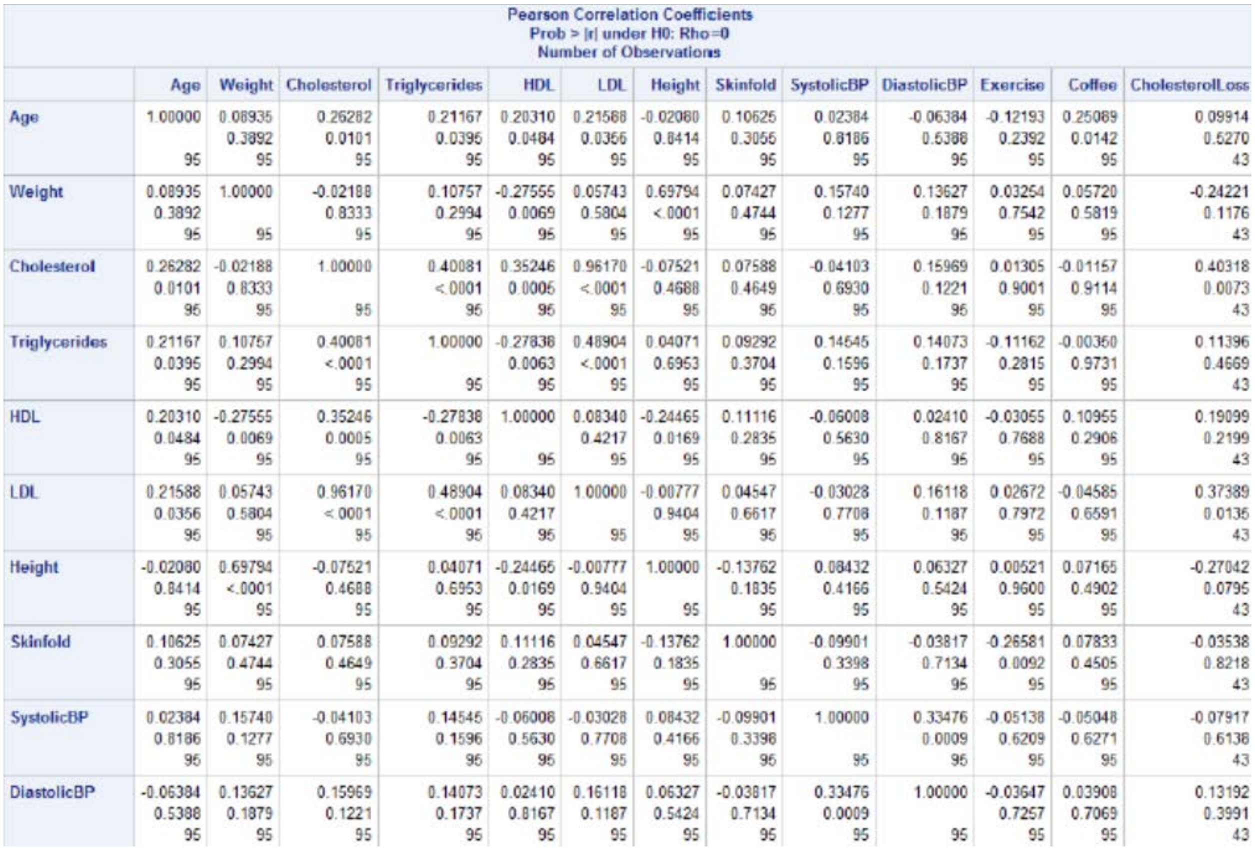
MULTICOLLINEARITY INVESTIGATION

Now 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 an increase in exercise help with a decrease in cholesterol loss? Could overall cholesterol be related to HDL and LDL levels? 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:

```
/* 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;
```

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



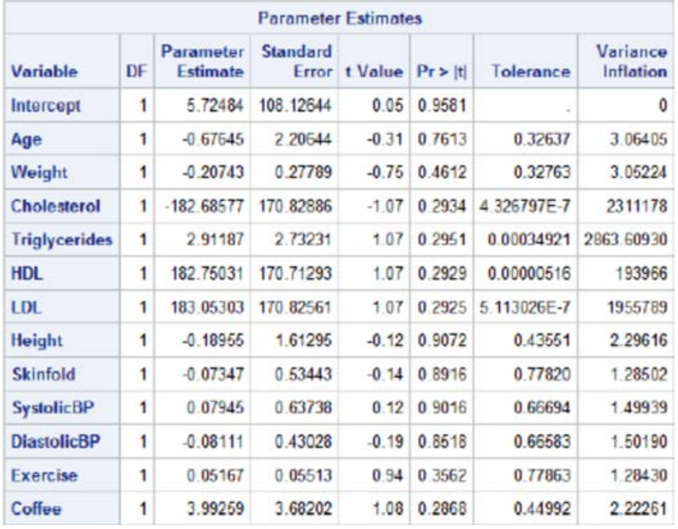
Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	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	0.25069 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 0.0136 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

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 seems to be some particularly high correlations between a few of the variables. Some relationships of note would be Cholesterol / LDL (0.96) and Weight / Height (0.70). Next we will examine multicollinearity through the Variance Inflation Factor, Tolerance, and Collinearity Diagnostics. This can be done by specifying the vif, tol, and collin options respectively after the model statement:

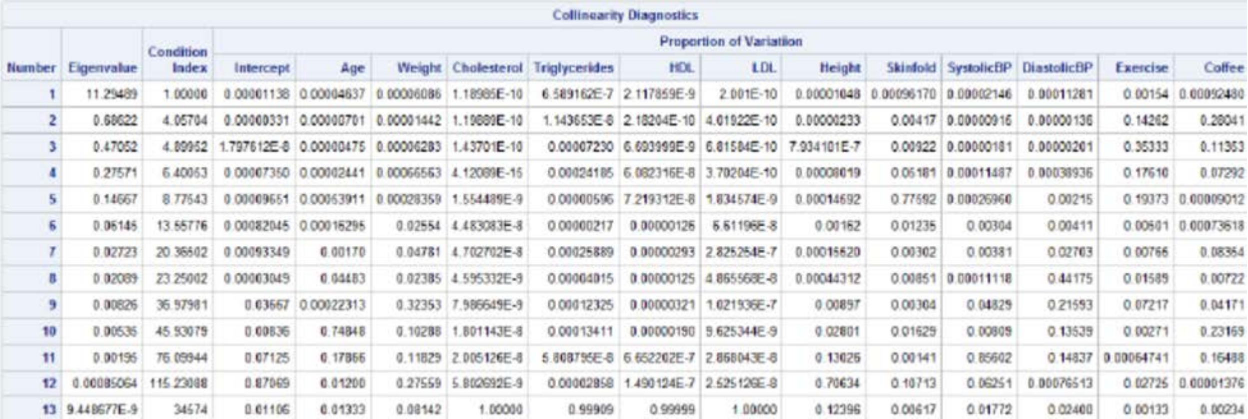
```
proc reg data=health;
model cholesterolloss = age weight cholesterol triglycerides hdl ldl
height skinfold systolicbp diastolicbp exercise coffee / vif tol
collin;
title 'Health Predictors - Multicollinearity Investigation of VIF and
Tol';
run;
```



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

Tolerance and VIF Investigation Results

When considering tolerance, we want to make sure that no values fall below 0.1. In reviewing our results, we can see several variables − namely cholesterol, triglycerides, HDL, and LDL − had values well below our 0.1 cutoff value. As for variance inflation, the magic number to look out for is anything above the value of 10. 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. Next, we will look at the collinearity diagnostics for an eigensystem analysis of covariance comparison:



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	0.00000916	0.00000136	0.14262	0.28041
3	0.47052	4.89962	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	6.40063	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	8.77643	0.00009651	0.00063911	0.00028359	1.554489E-9	0.00000596	7.219312E-8	1.834574E-9	0.00014692	0.77592	0.00026960	0.00215	0.19373	0.00009012
6	0.06146	13.65776	0.00082045	0.00016295	0.02554	4.483083E-8	0.00000217	0.00000126	6.61196E-8	0.00162	0.01235	0.00304	0.00411	0.00601	0.00073618
7	0.02723	20.36602	0.00093349	0.00170	0.04781	4.702702E-8	0.00025889	0.00000293	2.825254E-7	0.00015620	0.00302	0.00381	0.02703	0.00766	0.08364
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	0.01585	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	0.00536	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	0.00196	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

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 for our results, we can see a large deviation in the final three factors, with the eigenvalue landing very close to zero and the condition index being quite large in comparison.

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!

REGULARIZATION METHODS

Statistical theory and machine learning have made great strides in creating regularization techniques that are designed to help generalize models with highly complex relationships (such as multicollinearity). In its most simplistic form, regularization adds a penalty to model parameters (all except intercepts) so the model generalizes the data instead of overfitting (a side effect of multicollinearity).

There are two main types of regularization: L1 (Lasso Regression) and L2 (Ridge Regression). The key difference between these two types of regularization can be found in how they handle the penalty. Through Ridge regression, a squared magnitude of the coefficient is added as the penalty term to the loss function. Take the following cost function as an example:

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

Considering the above equation, if lambda (λ - the penalty) is zero then the equation will go back to ordinary least squares estimations, whereas a very large lambda would add too much weight to the model which will lead to under-fitting. Considering this, it is worthy to note the necessity in making sure we have reviewed exactly how lambda is chosen, as this could help avoid this issue of over-fitting.

Through Lasso Regression (Least Absolute Shrinkage and Selection Operator), the absolute value of magnitude of the coefficient is added as the penalty term to the loss function. As before, let us take the following cost function into consideration:

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

Considering the above equation like before, if lambda (λ - the penalty) is zero then the equation will again go back to ordinary least squares estimations, whereas a very large lambda would make the coefficients approach zero, thus resulting in an under-fit model like before.

The key difference between these two techniques lies in the fact that Lasso is intended to shrink the coefficient of the less important variables to zero, thus removing some of these features altogether, which works well if feature selection is the goal of a particular model trimming technique. However, if the correction of multicollinearity is your goal, then Lasso (L1 regulation) isn't the way to go. Therefore, L2 regulation techniques become our method of choice. Ridge Regression is a relatively simple process that can be employed to help correct for incidents of multicollinearity where the subtraction of a variable is not an option and feature selection is not a concern.

LASSO REGRESSION FOR LINEAR MODELS

LASSO selection arises from a constrained form of ordinary least squares regression where the sum of the absolute values of the regression coefficients is constrained to be smaller than a specified parameter. Instead of punishing the high values of the coefficients β (like in ridge regression), it figures out which values are irrelevant and sets them to zero. Therefore, this method results in fewer features being included in the final model, which can be an advantage in some situations.

More precisely, let $X = (x_1, x_2, \ldots.., x_m)$ denote the matrix of covariates and let y denote the response, where the x_is have been centered and scaled to have a single unit standard deviation and mean zero, and y also has mean zero. For a given parameter t, the LASSO regression coefficients $\beta = (\beta_1, \beta_2, \ldots.., \beta_m)$ are the solution to the constrained optimization problem

$$minimize \; \|y - X\beta\|^2 \qquad subject \; to \; \sum_{j=1}^{m} |\beta_j| \leq t$$

Provided that the LASSO parameter t is small enough, some of the regression coefficients will be exactly zero. Hence, you can view the LASSO as selecting a subset of the regression coefficients for each LASSO parameter. By increasing the LASSO parameter in discrete steps, you obtain a sequence of regression coefficients where the nonzero coefficients at each step correspond to selected parameters.

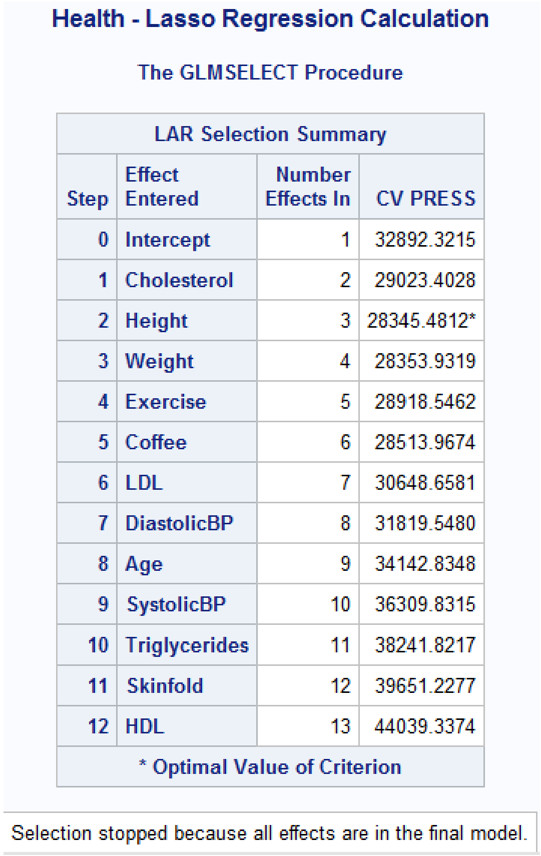
Early implementations (Tibshirani 1996) of LASSO selection used quadratic programming techniques to solve the constrained least squares problem for each LASSO parameter of interest. However, Osborne, Presnell, and Turlach (2000) later developed a "homotopy method" which generates the LASSO solutions for all values of t. Efron and colleauges (2004) then derived a variant of their algorithm for least angle regression (LAR) that can be used to obtain a sequence of LASSO solutions from which all other LASSO solutions can be obtained by linear interpolation. This algorithm for SELECTION=LASSO is used in PROC GLMSELECT. It can be viewed as a stepwise procedure with a single addition to or deletion from the set of nonzero regression coefficients at any step.

As with the other selection methods supported by PROC GLMSELECT, you can specify a criterion to choose among the models at each step of the LASSO algorithm with the CHOOSE= option. You can also specify a stopping criterion with the STOP= option. See the discussion in the section Forward Selection (FORWARD) for additional details. The model degrees of freedom PROC GLMSELECT uses at any step of the LASSO are simply the number of nonzero regression coefficients in the model at that step. Efron et al. (2004) cite empirical evidence for doing this but do not give any mathematical justification for this choice.

One issue with LASSO regression is that it has a quadratic programming problem, however, this can be solved by utilizing the LAR solution as detailed in Efron and colleagues (2004) work. This method utilizes a stepwise variable selection algorithm (Least Angel Regression and Shrinkage). This option is a less greedy version of traditional forward selection methods and is designed to efficiently solve LASSO's solution path issue. It has the same order of computational efforts as a single OLS fit $O(np^2)$. In SAS the LAR modification of LASSO selection uses the LASSO algorithm to select the set of covariates in the model at any step, but uses ordinary least squares regression with just these covariates to obtain the regression coefficients. You can request this hybrid method by specifying the LSCOEFFS suboption of SELECTION=LASSO.

This procedure can be used for both linear and logistic regression and is completed by employing either Proc GLMSelect or Proc HPreg.

```
/* 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;
```



Health - Lasso Regression Calculation

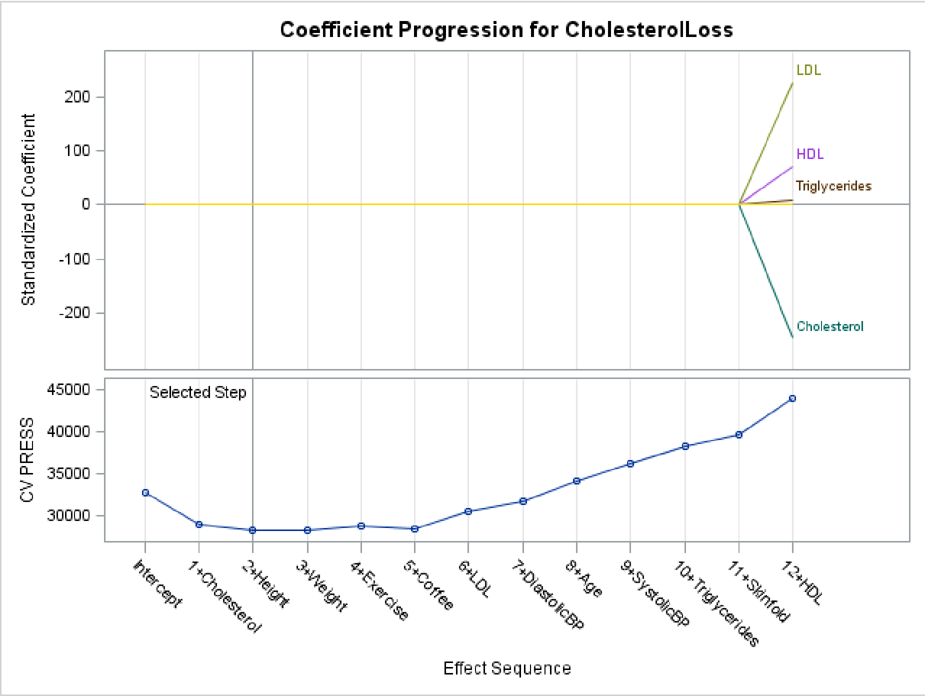
The GLMSELECT Procedure

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

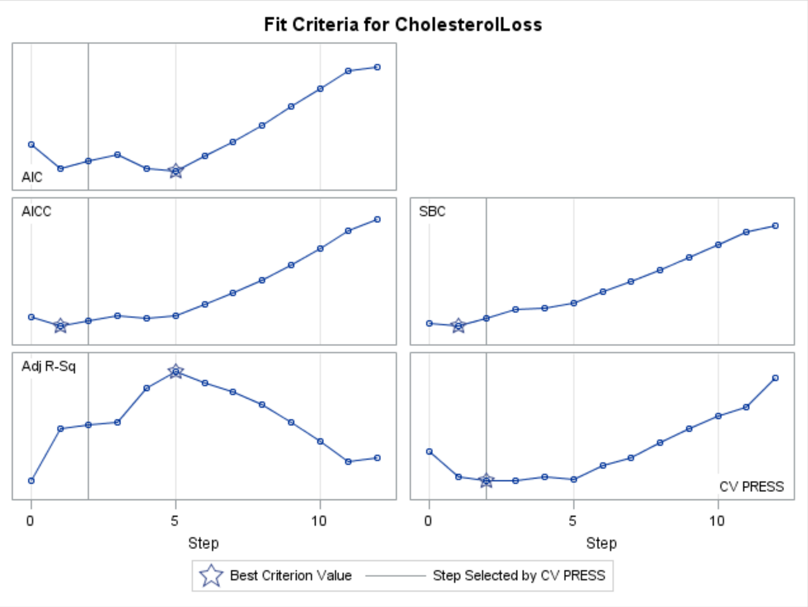
* Optimal Value of Criterion

Selection stopped because all effects are in the final model.

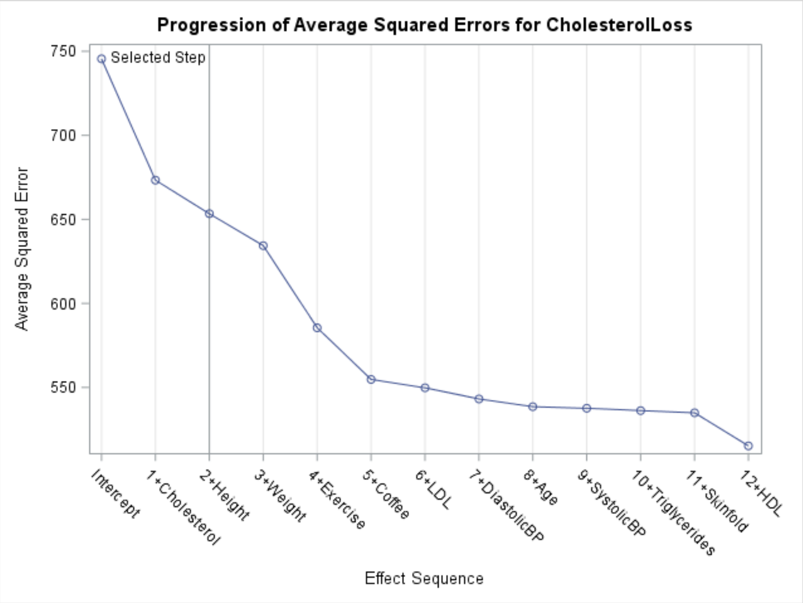
LAR Selection Summary for GLM Select



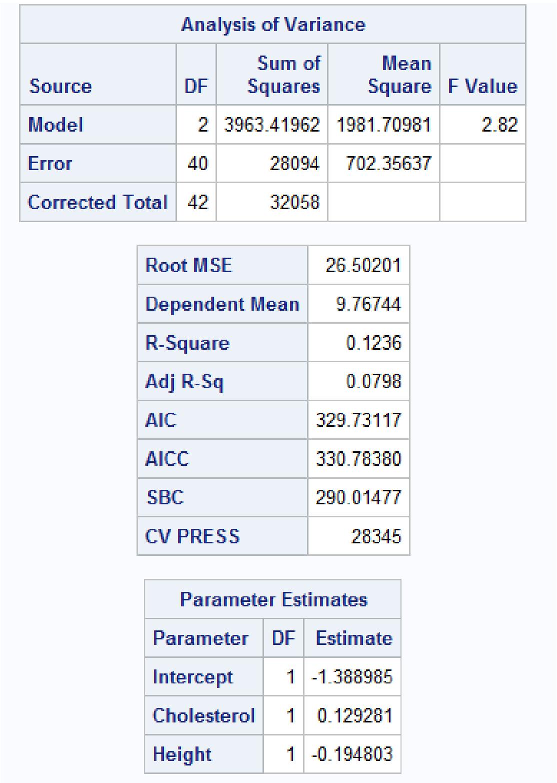
Coefficient Progression for Outcome Variable



Fit Criteria for Outcome Variable



Progression of Average Squared Errors for Outcome Variable



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

Final Selected Model by LASSO Regression and LAR Procedures

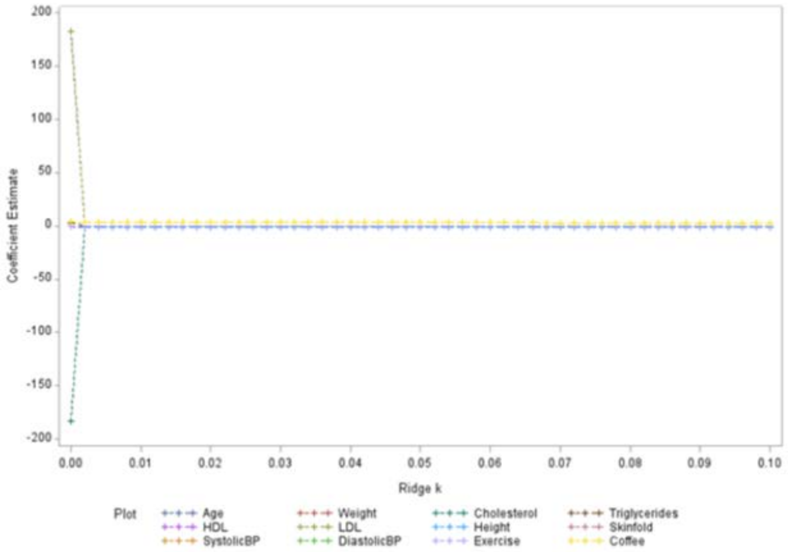
RIDGE REGRESSION FOR LINEAR MODELS

Ridge regression is a variant of least squares regression and is oftentimes used when multicollinearity cases are 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 (as discussed earlier). Therefore, it could be beneficial to utilize a technique such as ridge regression in order to ensure a smaller variance in resulting parameter estimates. The following code details a ridge regression application:

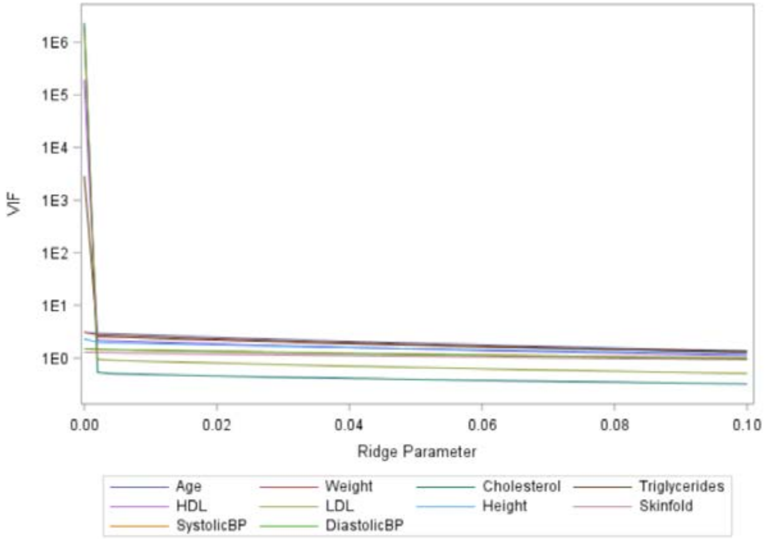
```
/* Ridge Regression Example */
proc reg data=health outvif plots(only)=ridge(unpack VIFaxis=log)
outest=rrhealth ridge=0 to 0.10 by .002;
model cholesterolloss = age weight cholesterol triglycerides hdl ldl
height skinfold systolicbp diastolicbp exercise coffee;
plot / ridgeplot nomodel nostat;
title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

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

The ridge= option requests the ridge regression technique in the REG procedure, the outvif option is indicated to ouput= the variance inflation factors, and the outset option displays the data table with our results. For this study, we also wanted to look at each of the individual plots for ridge traces and VIF traces, so the unpack suboption of the plots(only)=ridge option is designated. The plot statement is designated to display scatter plots of the y and x variables, ridgeplot to request the ridge trace for ridge regression, nomodel to suppress the display of the fitted model and lable, and the nostat suppresses the display of the default statistics.

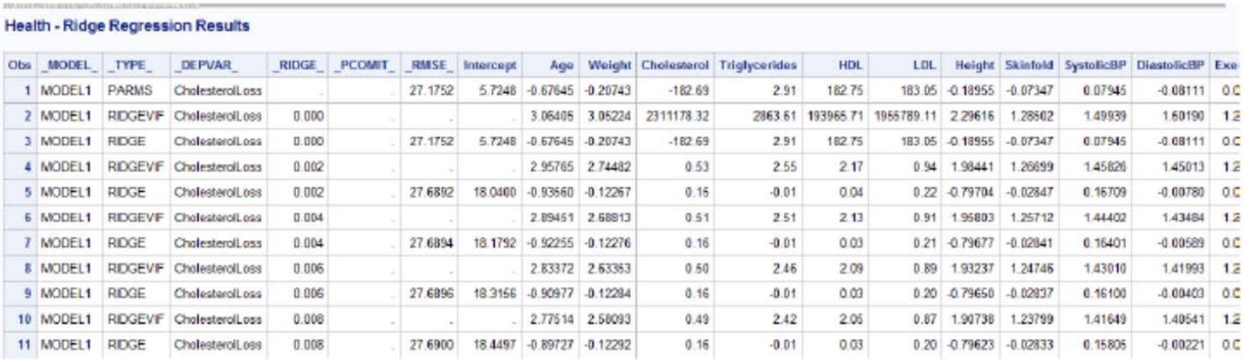
The results produced by this procedure can be seen below:



Ridge Trace Results



Variance Inflation Factors for CholesterolLoss



Health - Ridge Regression Results

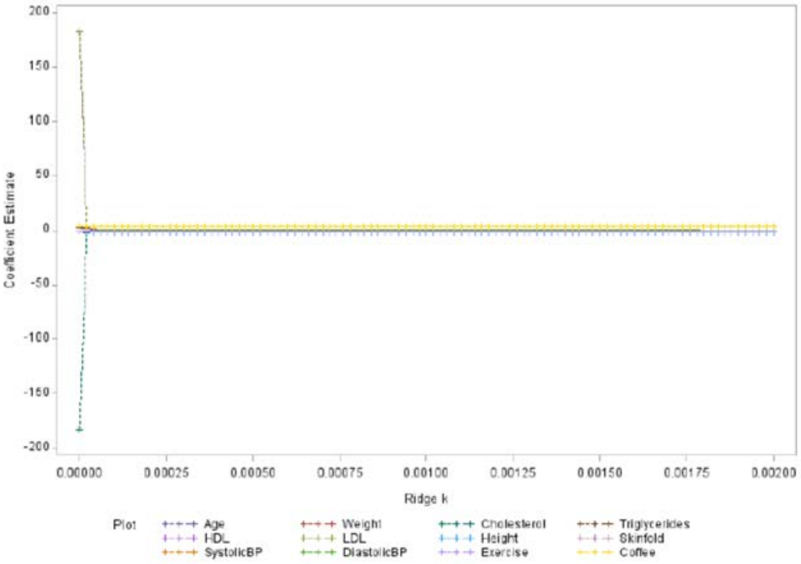
Obs	_MODEL_	_TYPE_	_DEPVAR_	_RIDGE_	_PCOMIT_	_RMSE_	Intercept	Age	Weight	Cholesterol	Triglycerides	HDL	LDL	Height	Skinfold	SystolicBP	DiastolicBP	Exe
1	MODEL1	PARMS	CholesterolLoss	.	.	27.1752	5.7248	-0.67645	-0.20743	-182.69	2.91	182.75	183.05	-0.18955	-0.07347	0.07945	-0.08111	0.C
2	MODEL1	RIDGEVIF	CholesterolLoss	0.000	.	.	.	3.06405	3.05224	2311178.32	2863.61	193965.71	1956789.11	2.29616	1.28502	1.49939	1.60190	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.C
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	0.16	-0.01	0.04	0.22	-0.79704	-0.02847	0.16709	-0.00780	0.C
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.C
8	MODEL1	RIDGEVIF	CholesterolLoss	0.006	.	.	.	2.83372	2.63363	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.C
10	MODEL1	RIDGEVIF	CholesterolLoss	0.008	.	.	.	2.77514	2.58093	0.49	2.42	2.06	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	0.15806	-0.00221	0.C

Ridge Regression Results

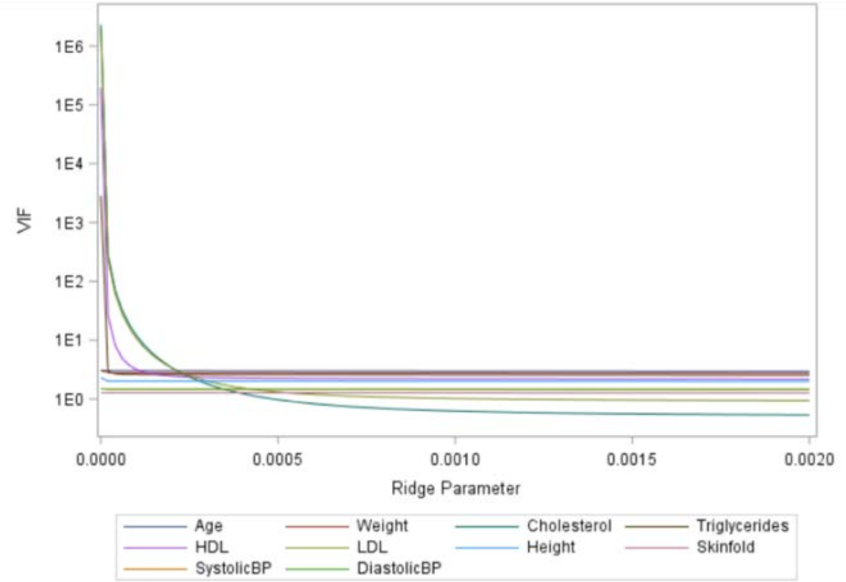
From these results we want to derive the appropriate ridge parameter or "k" to include in the analysis. The ridge parameter column is labeled _RIDGE_ and the associated values under each variable column are the new parameter estimates. There are several schools of thought concerning how to choose the best value of "k". I recommend reading Dorugade and Kashid's 2010 paper for more information on this matter. The current paper will simply look at the least increase in _RMSE_ and a decrease in ridge variable inflation factors for each variable. Given that our current range of "k" displayed an immediate correction (as can be seen visually in our ridge trace and VIF graphs), we will dig down further into the potential "k" values to find a more specific value for our use:

```
proc reg data=health outvif plots(only)=ridge(unpack VIFaxis=log)
outest=rrhealth ridge=0 to 0.002 by .00002;
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;
```

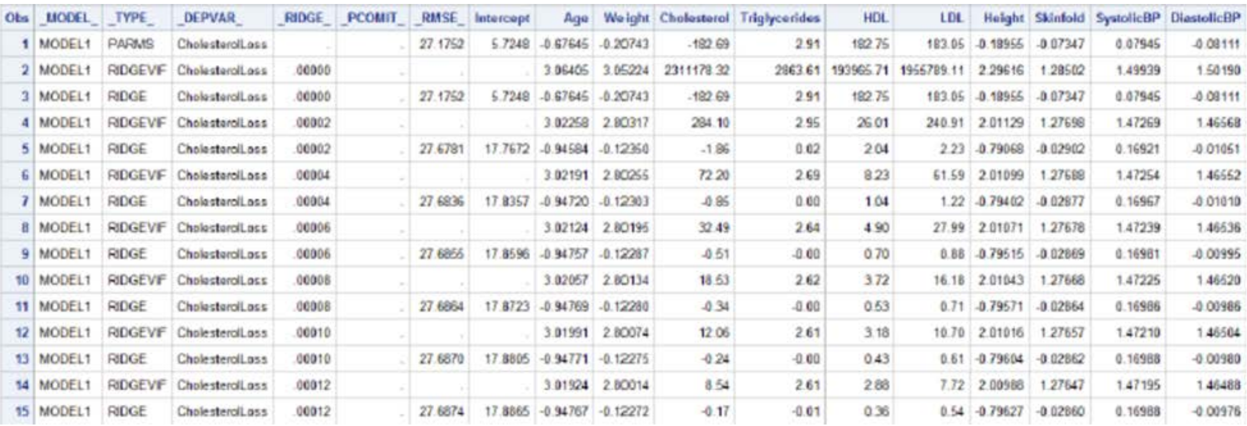
The results of this more detailed dig are as follows:



Ridge Trace Results



Variance Inflation Factors for CholesterolLoss



Obs	_MODEL_	_TYPE_	_DEPVAR_	_RIDGE_	_PCOMIT_	_RMSE_	Intercept	Age	Weight	Cholesterol	Triglycerides	HDL	LDL	Height	Skinfold	SystolicBP	DiastolicBP
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
2	MODEL1	RIDGEVIF	CholesterolLoss	.00000	.	.	.	3.06405	3.05224	2311178.32	2863.61	193965.71	1956789.11	2.29616	1.28502	1.49939	1.50190
3	MODEL1	RIDGE	CholesterolLoss	.00000	.	27.1752	5.7248	-0.67645	-0.20743	-182.69	2.91	182.75	183.05	-0.18955	-0.07347	0.07945	-0.08111
4	MODEL1	RIDGEVIF	CholesterolLoss	.00002	.	.	.	3.02258	2.80317	284.10	2.95	26.01	240.91	2.01129	1.27698	1.47269	1.46568
5	MODEL1	RIDGE	CholesterolLoss	.00002	.	27.6781	17.7672	-0.94584	-0.12350	-1.86	0.02	2.04	2.23	-0.79068	-0.02902	0.16921	-0.01051
6	MODEL1	RIDGEVIF	CholesterolLoss	.00004	.	.	.	3.02191	2.80255	72.20	2.69	8.23	61.59	2.01099	1.27688	1.47254	1.46552
7	MODEL1	RIDGE	CholesterolLoss	.00004	.	27.6836	17.8357	-0.94720	-0.12303	-0.85	0.00	1.04	1.22	-0.79402	-0.02877	0.16967	-0.01010
8	MODEL1	RIDGEVIF	CholesterolLoss	.00006	.	.	.	3.02124	2.80195	32.49	2.64	4.90	27.99	2.01071	1.27678	1.47239	1.46536
9	MODEL1	RIDGE	CholesterolLoss	.00006	.	27.6855	17.8596	-0.94757	-0.12287	-0.51	-0.00	0.70	0.88	-0.79515	-0.02869	0.16981	-0.00995
10	MODEL1	RIDGEVIF	CholesterolLoss	.00008	.	.	.	3.02057	2.80134	18.53	2.62	3.72	16.18	2.01043	1.27668	1.47225	1.46520
11	MODEL1	RIDGE	CholesterolLoss	.00008	.	27.6864	17.8723	-0.94769	-0.12280	-0.34	-0.00	0.53	0.71	-0.79571	-0.02864	0.16986	-0.00986
12	MODEL1	RIDGEVIF	CholesterolLoss	.00010	.	.	.	3.01991	2.80074	12.06	2.61	3.18	10.70	2.01016	1.27657	1.47210	1.46504
13	MODEL1	RIDGE	CholesterolLoss	.00010	.	27.6870	17.8805	-0.94771	-0.12275	-0.24	-0.00	0.43	0.61	-0.79604	-0.02862	0.16988	-0.00980
14	MODEL1	RIDGEVIF	CholesterolLoss	.00012	.	.	.	3.01924	2.80014	8.54	2.61	2.88	7.72	2.00988	1.27647	1.47195	1.46488
15	MODEL1	RIDGE	CholesterolLoss	.00012	.	27.6874	17.8865	-0.94767	-0.12272	-0.17	-0.01	0.36	0.54	-0.79627	-0.02860	0.16988	-0.00976

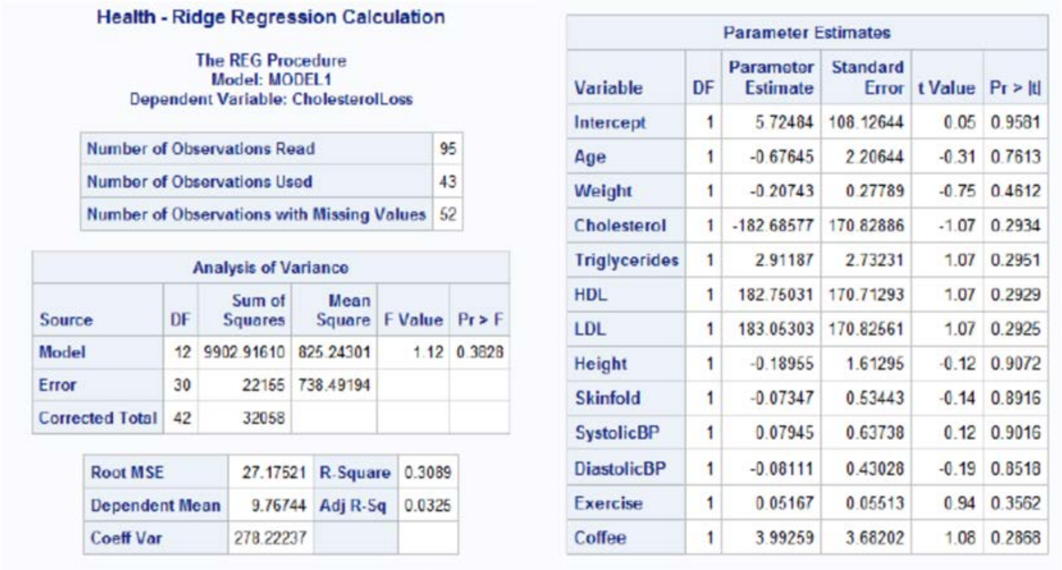
Ridge Regression Results

These results display a more gradual adjustment over several iterations of potential “k” values. Ultimately, it seems that the ridge parameter of 0.0001 may be our winner, as we see a slight increase in _RMSE_ from 27.1752 to 27.6864 and significant drop in the VIF for each of our problem variables to below our cutoff of 10. Therefore, this study will choose the ridge parameter of 0.0001 for the resulting parameter adjustments which are identified in the following code:

```
proc reg data=health outvif plots(only)=ridge(unpack VIFaxis=log)
outest=rrhealth_final ridge=.0001;
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;
```

These results can then be used as our final adjusted model with the multicollinearity issue controlled!



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

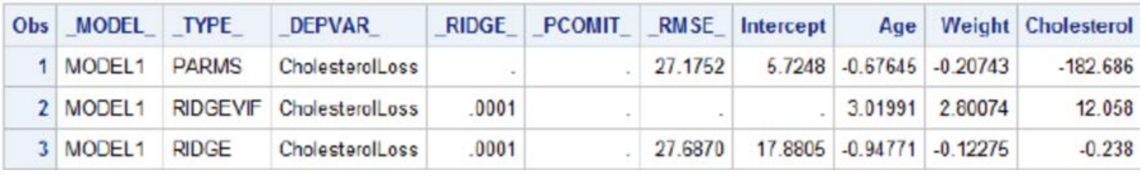
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	12	9902.91610	825.24301	1.12	0.3828
Error	30	22165	738.49194		
Corrected Total	42	32068			

Root MSE	27.17521	R-Square	0.3089
Dependent Mean	9.76744	Adj R-Sq	0.0325
Coeff Var	278.22237		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	5.72484	108.12644	0.05	0.9581
Age	1	-0.67645	2.20644	-0.31	0.7613
Weight	1	-0.20743	0.27789	-0.75	0.4612
Cholesterol	1	-182.68577	170.82886	-1.07	0.2934
Triglycerides	1	2.91187	2.73231	1.07	0.2951
HDL	1	182.75031	170.71293	1.07	0.2929
LDL	1	183.05303	170.82561	1.07	0.2925
Height	1	-0.18955	1.61295	-0.12	0.9072
Skinfold	1	-0.07347	0.53443	-0.14	0.8916
SystolicBP	1	0.07945	0.63738	0.12	0.9016
DiastolicBP	1	-0.08111	0.43028	-0.19	0.8518
Exercise	1	0.05167	0.05513	0.94	0.3562
Coffee	1	3.99259	3.68202	1.08	0.2868

Ridge Regression Results for Original Model



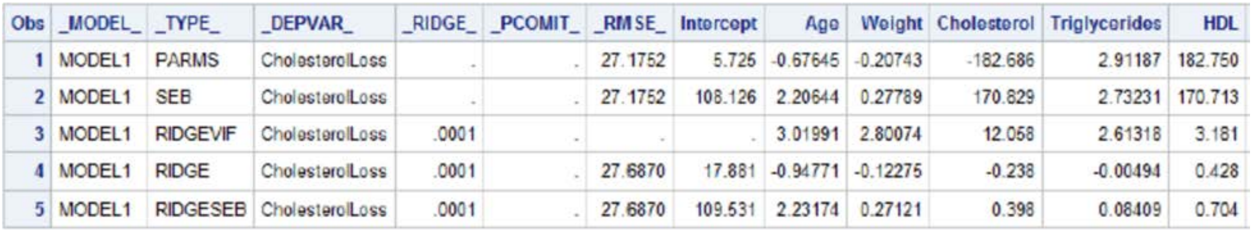
Obs	_MODEL_	_TYPE_	_DEPVAR_	_RIDGE_	_PCOMIT_	_RMSE_	Intercept	Age	Weight	Cholesterol
1	MODEL1	PARMS	CholesterolLoss	.	.	27.1752	5.7248	-0.67645	-0.20743	-182.686
2	MODEL1	RIDGEVIF	CholesterolLoss	.0001	.	.	.	3.01991	2.80074	12.058
3	MODEL1	RIDGE	CholesterolLoss	.0001	.	27.6870	17.8805	-0.94771	-0.12275	-0.238

Adjusted Ridge Regression Results

If we want to see standard errors and parameter estimates for our new model, we can designate outseb in our model statement when we rerun the model.

```
proc reg data=health outvif plots(only)=ridge(unpack VIFaxis=log)
outest=rrhealth_final outseb ridge=.0001;
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;
```

Our results will then look something like this:



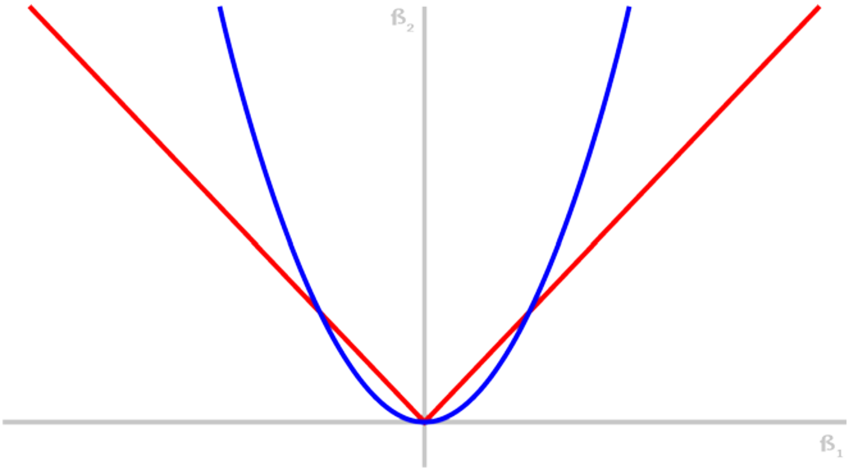
Obs	_MODEL_	_TYPE_	_DEPVAR_	_RIDGE_	_PCOMIT_	_RMSE_	Intercept	Age	Weight	Cholesterol	Triglycerides	HDL
1	MODEL1	PARMS	CholesterolLoss	.	.	27.1752	5.725	-0.67645	-0.20743	-182.686	2.91187	182.750
2	MODEL1	SEB	CholesterolLoss	.	.	27.1752	108.126	2.20644	0.27789	170.829	2.73231	170.713
3	MODEL1	RIDGEVIF	CholesterolLoss	.0001	.	.	.	3.01991	2.80074	12.058	2.61318	3.181
4	MODEL1	RIDGE	CholesterolLoss	.0001	.	27.6870	17.881	-0.94771	-0.12275	-0.238	-0.00494	0.428
5	MODEL1	RIDGESEB	CholesterolLoss	.0001	.	27.6870	109.531	2.23174	0.27121	0.398	0.08409	0.704

Ridge Regression Results With Outseb

The SEB and RIDGESEB rows (_TYPE_ column) gives us the standard errors and parameter estimates of our original and adjusted models respectively.

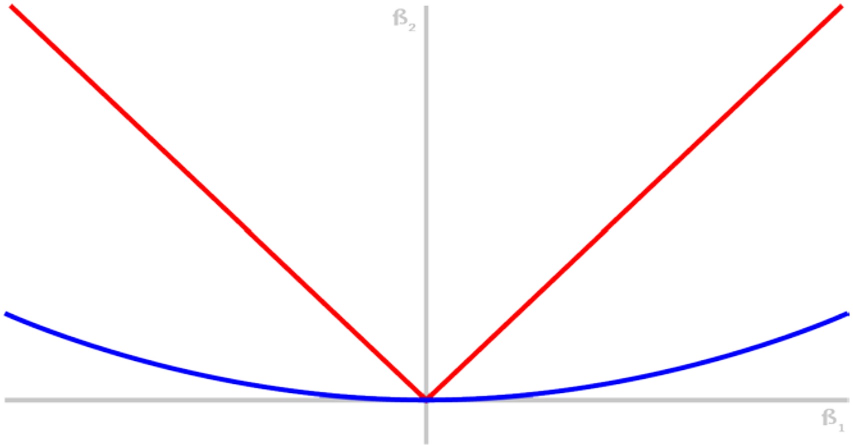
ELASTIC NETS FOR LINEAR MODELS

Some advantages of LASSO regression: (1) it is great if your goal is to reduce the number of variables in your model in order to control for multicollinearity, (2) it enforces sparcity in parameter selection and inclusion, (3) it has a quadratic programming problem that can be resolved through use of the LAR solution (Efron et al, 2004), (4) as λ goes to 0, t goes to ∞, and the model approaches the traditional OLS solution (reduces the amount of bias), (5) as λ goes to ∞, t goes to 0, and $\hat{\beta}$ is zeroed out. Some disadvantages of LASSO regression: (1) if a group of predictors are highly correlated among themselves, LASSO tends to pick only one of them and will shrink the others to zero, (2) LASSO can not perform grouped selection, as it tends to select only one variable.



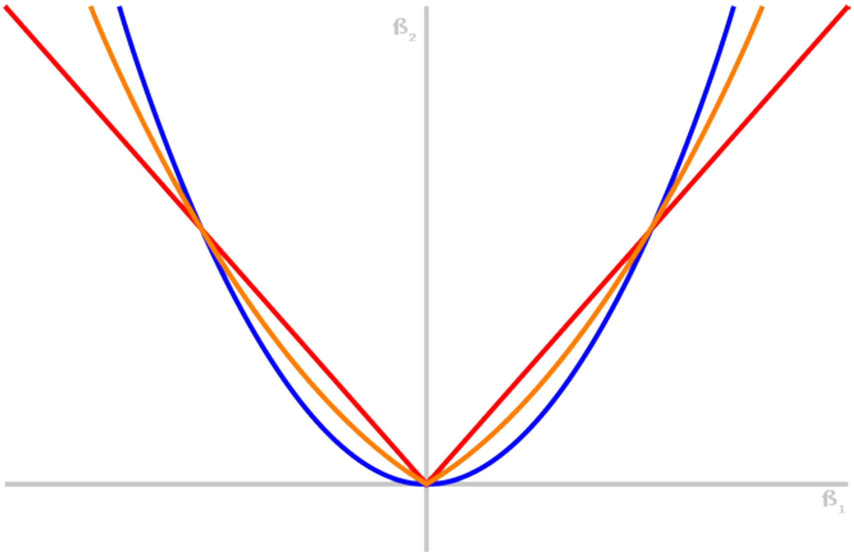
Lasso Regression Adjustment (Blue) to Linear Regression (Red)

Some advantages of Ridge regression: (2) it is great if your goal is to adjust for multicollinearity with grouped selections, rather than for variable reduction, (2) it produces biased but smaller variance and smaller Mean Square Error (MSE), (3) it results in an explicit solution. Some disadvantages of Ridge regression: (1) the aforementioned biased results, (2) it tends to shrink coefficients to near zero but can not produce a parsimonious model.



Ridge Regression Adjustment (Blue) to Linear Regression (Red)

Several researchers and data scientists have worked hard to explore the value of procedures like Elastic Nets to help resolve the L1/L2 debate to multicollinearity correction. Through this technique, we are able to combine the strengths of both Ridge and LASSO regression, while minimizing the negative impact of either of these procedures. Some advantages of Elastic Net is that it is able to (1) enforce sparsity, (2) it has no limitation on the number of selected variables, and (3) it encourages a grouping effect in the presence of highly correlated predictors. A main disadvantage of this technique is that a naïve elastic net can suffer from double shrinkage, therefore, one needs to be careful when employing this option. If a naïve elastic net is found, a correction does exist to help control for this.



Elastic Net (Orange) Compared to LASSO (Red) and Ridge Regression (Blue) Adjustments

The elastic net method bridges the LASSO method and ridge regression. It balances having a parsimonious model with borrowing strength from correlated regressors, by solving the least squares regression problem with constraints on both the sum of the absolute coefficients and the sum of the squared coefficients. More specifically, the elastic net coefficients $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_m)$ are the solution to the constrained optimization problem

$$\min ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2 \qquad \text{subject to} \qquad \sum_{j=1}^{m} |\beta_j| \leq t_1, \sum_{j=1}^{m} \beta_j^2 \leq t_2$$

The method can be written as the equivalent Lagrangian form

$$\min ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2 + \lambda_1 \sum_{j=1}^{m} |\beta_j| + \lambda_2 \sum_{j=1}^{m} \beta_j^2$$

Or as

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

If t_1 is set to a very large value or, equivalently, if λ_1 is set to 0, then the elastic net method reduces to ridge regression. If t_2 is set to a very large value or, equivalently, if λ_2 is set to 0, then the elastic net method reduces to LASSO. If t_1 and t_2 are both large or, equivalently, if λ_1 and λ_2 are both set to 0, then the elastic net method reduces to ordinary least squares regression.

Additionally, as stated by Zou and Hastie in their 2005 article, the elastic net method can overcome the limitations of variable selection through LASSO, in particular, in the following three scenarios:

- In the case where you have more parameters than observations, $m > n$, the LASSO method selects at most n variables before it saturates, because of the nature of the convex optimization problem. This can be a defect for a variable selection method. By contrast, the elastic net method can select more than n variables in this case because of the ridge regression regularization.
- If there is a group of variables that have high pairwise correlations, then whereas LASSO tends to select only one variable from that group, the elastic net method can select more than one variable.
- In the $n > m$ case, if there are high correlations between predictors, it has been empirically observed that the prediction performance of LASSO is dominated by ridge regression. In this case, the elastic net method can achieve better prediction performance by using ridge regression regularization.

An elastic net fit is achieved by building on LASSO estimation, in the following sense. Let $\tilde{\mathbf{X}}$ be a matrix obtained by augmenting $\mathbf{X}$ with a scaled identity matrix,

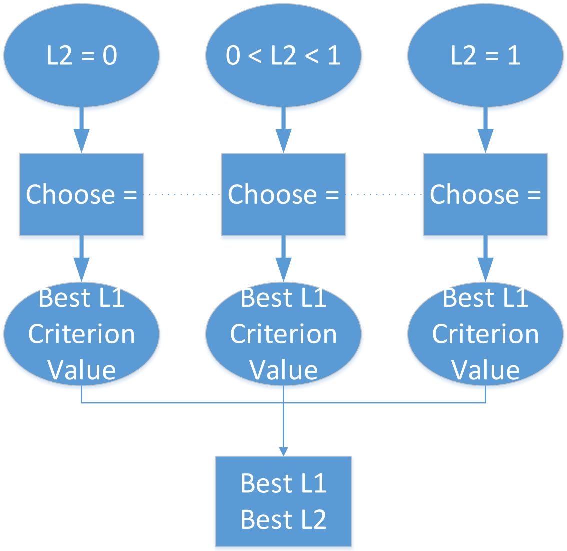
$$\tilde{\mathbf{X}} = [\mathbf{X}; \sqrt{\lambda_2} I]$$

Let $\tilde{\mathbf{y}}$ be a vector correspondingly obtained by augmenting the response $\mathbf{y}$ with m 0's, $\tilde{\mathbf{y}} = [\mathbf{y}; \mathbf{0}]$
Then the Lagrangian form of the elastic net optimization problem can be reformulated as

$$\min ||\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\boldsymbol{\beta}||^2 + \lambda_1 \sum_{j=1}^{m} |\beta_j|$$

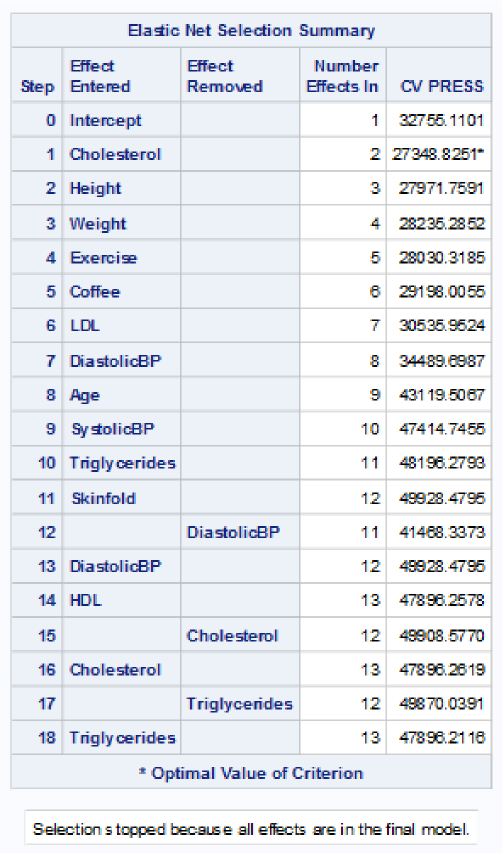
In other words, you can solve the elastic net method in the same way as LASSO by using this augmented design matrix $\tilde{\mathbf{X}}$ and response $\tilde{\mathbf{y}}$. Therefore, for given λ_2, the coefficients of the elastic net fit follow the same piecewise linear path as LASSO. Zou and Hastie (2005) suggest rescaling the coefficients by $1 + \lambda_2$ to deal with the double amount of shrinkage in the elastic net fit, and such rescaling is applied when you specify the ENSCALE option in the MODEL statement.

If you have a good estimate of λ_2, you can specify the value in the L2= option. If you do not specify a value for λ_2, then by default PROC GLMSELECT searches for a value between 0 and 1 that is optimal according to the current CHOOSE= criterion. Figure 48.12 illustrates the estimation of the ridge regression parameter λ_2 (L2). Meanwhile, if you do not specify the CHOOSE= option, then the model at the final step in the selection process is selected for each λ_2 (L2), and the criterion value shown in the below figure is the one at the final step that corresponds to the specified STOP= option (STOP=SBC by default). Additionally, it is worthwhile to note that when you specify the L2SEARCH=GOLDEN, it is assumed that the criterion curve that corresponds to the CHOOSE= option with respect to λ_2 is a smooth and bowl-shaped curve. However, this assumption is not checked and validated. Hence, the default value for the L2SEARCH= option is set to GRID.



Elastic Net can be used for both linear and logistic regression and is completed by employing either Proc GLMSelect or Proc HPreg.

```
/* 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;
```



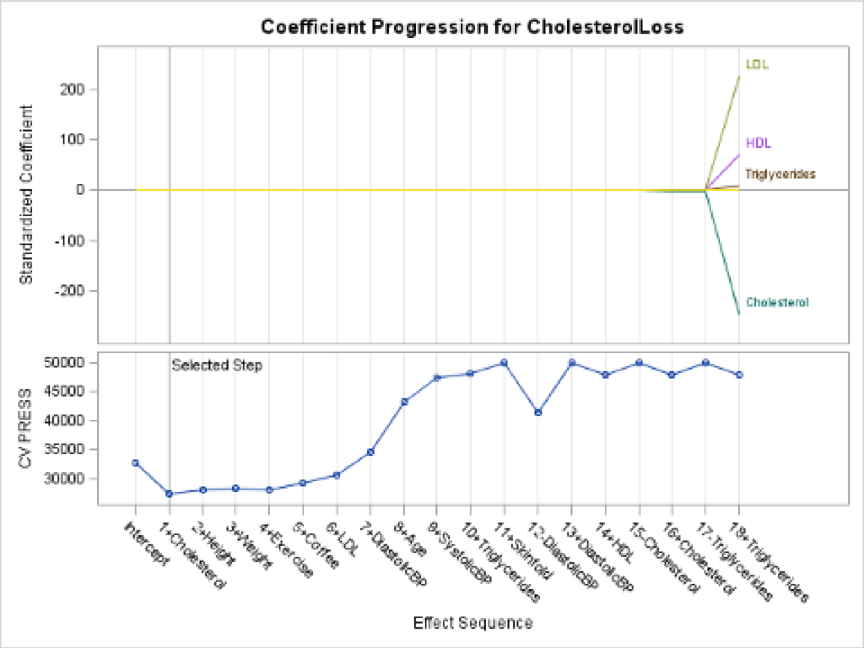
Elastic Net Selection Summary

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

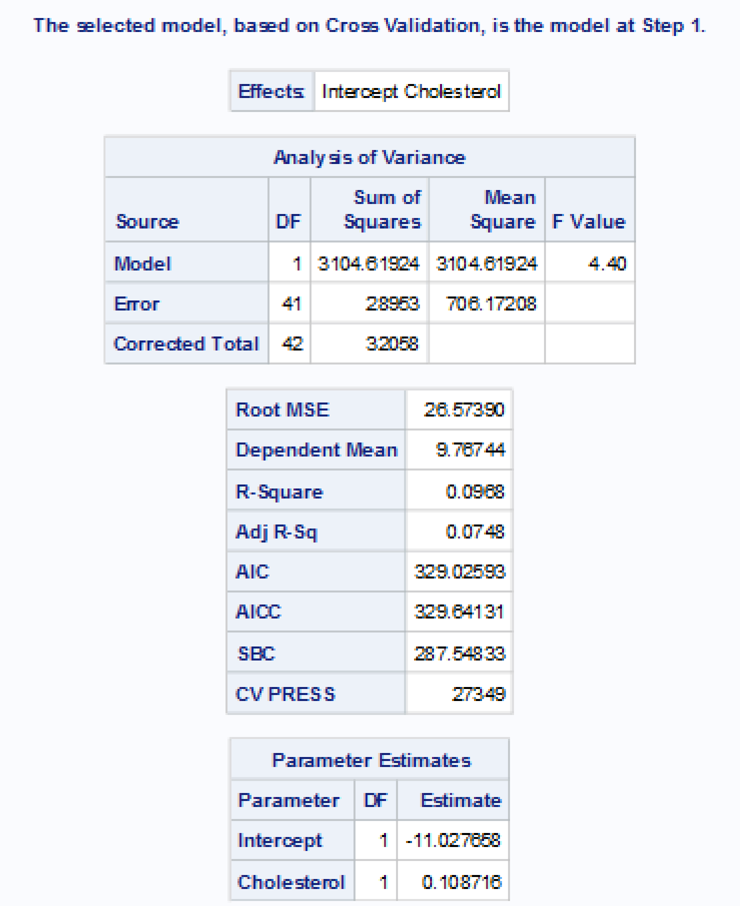
* Optimal Value of Criterion

Selection stopped because all effects are in the final model.

Elastic Net Selection Summary



Coefficient Progression for the Outcome Variable



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

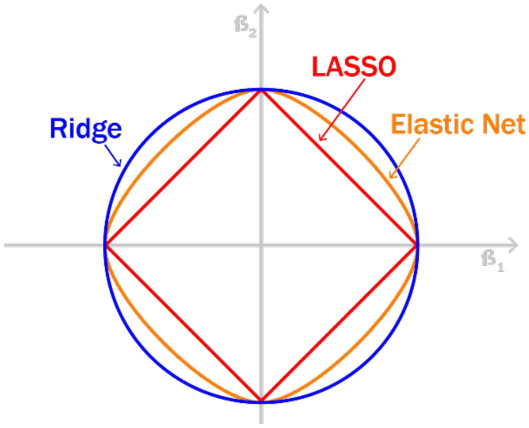
Parameter Estimates		
Parameter	DF	Estimate
Intercept	1	-11.027658
Cholesterol	1	0.108716

Final Model with Elastic Net Variable Selection

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 correct for this through the utilization of several different regularization and variable reduction techniques. A few ways in which to control for multicollinearity is through the implementation of techniques such as Ridge Regression, LASSO regression, and Elastic Nets.



Elastic Net Compared to LASSO and Ridge Regression Adjustments

Along with these procedures, there also exists substantive research into the cause and effect of multicollinearity in studies from fields across the research spectrum. For every issue that arises, there is a plethora of procedures that could be used to help control for and correct the effects that an issue such as multicollinearity can have on the integrity of a model. Given this, the author has included several references and recommended articles for your review to help further the understanding of all statisticians and programmers as to the effects of multicollinearity on research models.

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 AND RECOMMENDED READING

Afshartous, D., & Preston, R. A. (2011). Key Results of Interaction Models With Centering. Journal of Statistics Education. 19 (3). Retrieved from https://www.amstat.org/publications/jse/v19n3/afshartous.pdf

Allison, P. (2012, September 10). When Can You Safely Ignore Multicollinearity? Statistical Horizons.

Breiman, L. Better Subset Regression Using the Nonnegative Garrote Technometrics, 1995, 37, 373-384.

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.

Chakon, O. (2017). Practical Machine Learning: Ridge Regression Vs Lasso. Coding Startups: Coders With Entrepreneurial Mindset. Published August 3rd, 2017.

Chatterjee, S., Hadi, A.S. and Price, B. (2000). Regression Analysis by Examples. 3rd Edition, Wiley VCH, New York.

Coursera. (Accessed August 2018). Testing a Lasso Regression in SAS. Available at: https://www.coursera.org/lecture/machine-learning-data-analysis/testing-a-lasso-regression-with-sas-ntKNE

Cross Validated. (2015, November 28). What is elastic net regularization, and how does it solve the drawbacks of Ridge (L2) and Lasso (L1)? Retrieved from https://stats.stackexchange.com/questions/184029/what-is-elastic-net-regularization-and-how-does-itsolve-the-drawbacks-of-ridge/184031#184031

Dixon, P.M. (1993) The bootstrap and the jackknife: Describing the precision of ecological indices. Design and Analysis of Ecological Experiments, New York: Chapman & Hall, pp 290-318.

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.

Efron, B. and Tibshirani, R.J. (1993). An Introduction to the Bootstrap. New York: Chapman & Hall.
Hall, P. (1992). The Bootstrap and Edgeworth Expansion. New York: Springer-Verlag.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), "Least Angle Regression," The Annals of Statistics, 32, 407–499.

Hjorth, J.S.U. (1994) Computer Intensive Statistical Methods. London: Chapman & Hall.
Shao, J. and Tu, D. (1995). The Jackknife and Bootstrap. New York: Springer-Verlag.

Joshi, H., Kulkarni, H., and Deshpande, S. (2012). Multicollinearity Diagnostics in Statistical Modeling & Remedies to Deal With it Using SAS. Proceedings from PhUSE 2012. Retrieved from https://www.lexjansen.com/phuse/2012/sp/SP07.pdf.

Kyung, M.; Gill, J.; Ghosh, M. & Casella, G. Penalized Regression, Standard Errors, and Bayesian Lassos Bayesian Analysis, 2010, 5, 369-412.

Melkumova, L. E. & Shatskikh, S. Y. (2017). Comparing Ridge and LASSO estimators for data analysis. 3rd International Conference "Information Technology and Nanotechnology", April 2017, Samara, Russia.

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

Park, T. & Casella, G. The Bayesian Lasso Journal of the American Statistical Association, 2008, 103, 681-686.

Shtatland, E. S., Kleinman, K., Cain, E. M. (2004). A New Strategy of Model Building in PROC LOGISTIC With Automatic Variable Selection, Validation, Shrinkage and Model Averaging. Conference proceedings from SAS Users Group International Meeting 2004 (SUGI 29). Montreal, Canada. Retrieved from http://www2.sas.com/proceedings/sugi29/191-29.pdf

Stine, R. (1990). An introduction to bootstrap methods: Examples and ideas. Sociological Methods & Research, 18, 243-291.

Unknown. (2010, December 3). Sample 24982: Jackknife and Bootstrap Analyses. Retrieved from http://support.sas.com/kb/24/982.html#pur

Unknown. (2018). Elastic Net Selection (ELASTICNET). Retrieved from http://support.sas.com/documentation/cdl/en/statug/67523/HTML/default/viewer.htm#statug_glmselect_details12.htm

Unknown. (2018). What is Multicollinearity? [Lecture notes]. Retrieved from https://onlinecourses.science.psu.edu/stat501/node/344

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

Yuan, M. & Lin, Y. On the non-negative garrotte estimator Journal of the Royal Statistical Society (Series B), 2007, 69, 143-161.

Zou, H. & Hastie, T. Regularization and variable selection via the elastic net Journal of the Royal Statistical Society (Series B), 2005, 67, 301-320.

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