Jake Schinto
Math 312

Homework 9

1. Exploratory Data Analysis

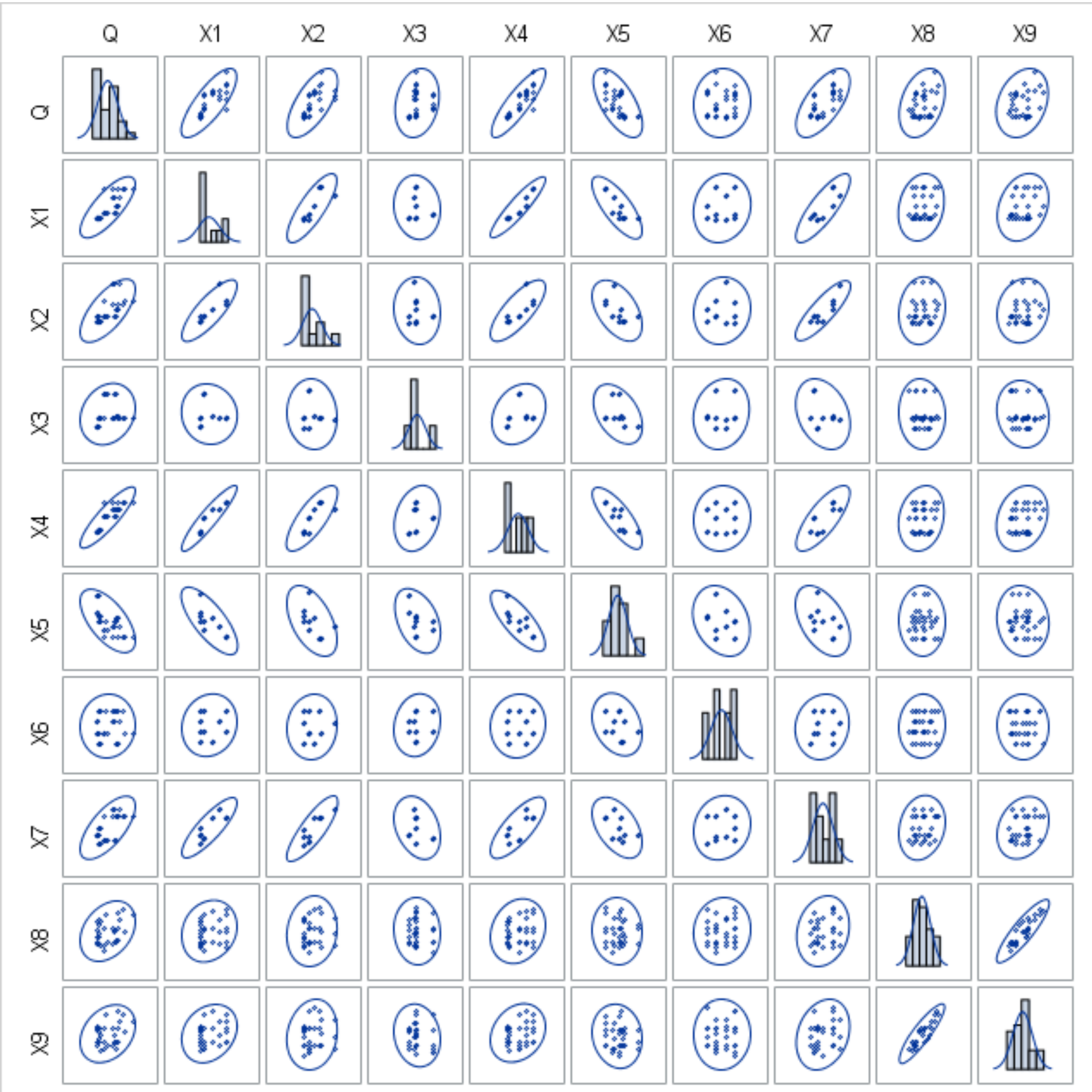


Figure 1.1 Scatter Matrix for Q and X1-X9 with histogram on the diagonal

By examining the scatterplot data, it looks like there are multiple possible correlations in the data. Q seems to correlate with X1, X2, X3, X4, X5, and X7. X1 appears to correlate with X2, X4, X5 and X7. X2 appears to correlate with X4 and X7. X4 appears to correlate with X5 and X7. X8 appears to correlate with X9. The others are too difficult to tell by only looking at the scatterplot. Looking at the histogram gives the impression that a couple of the variables seem to have a right skew: Q, X1, X2, X3, and X5. The other variables don't seem to have too noticeable of a skew.

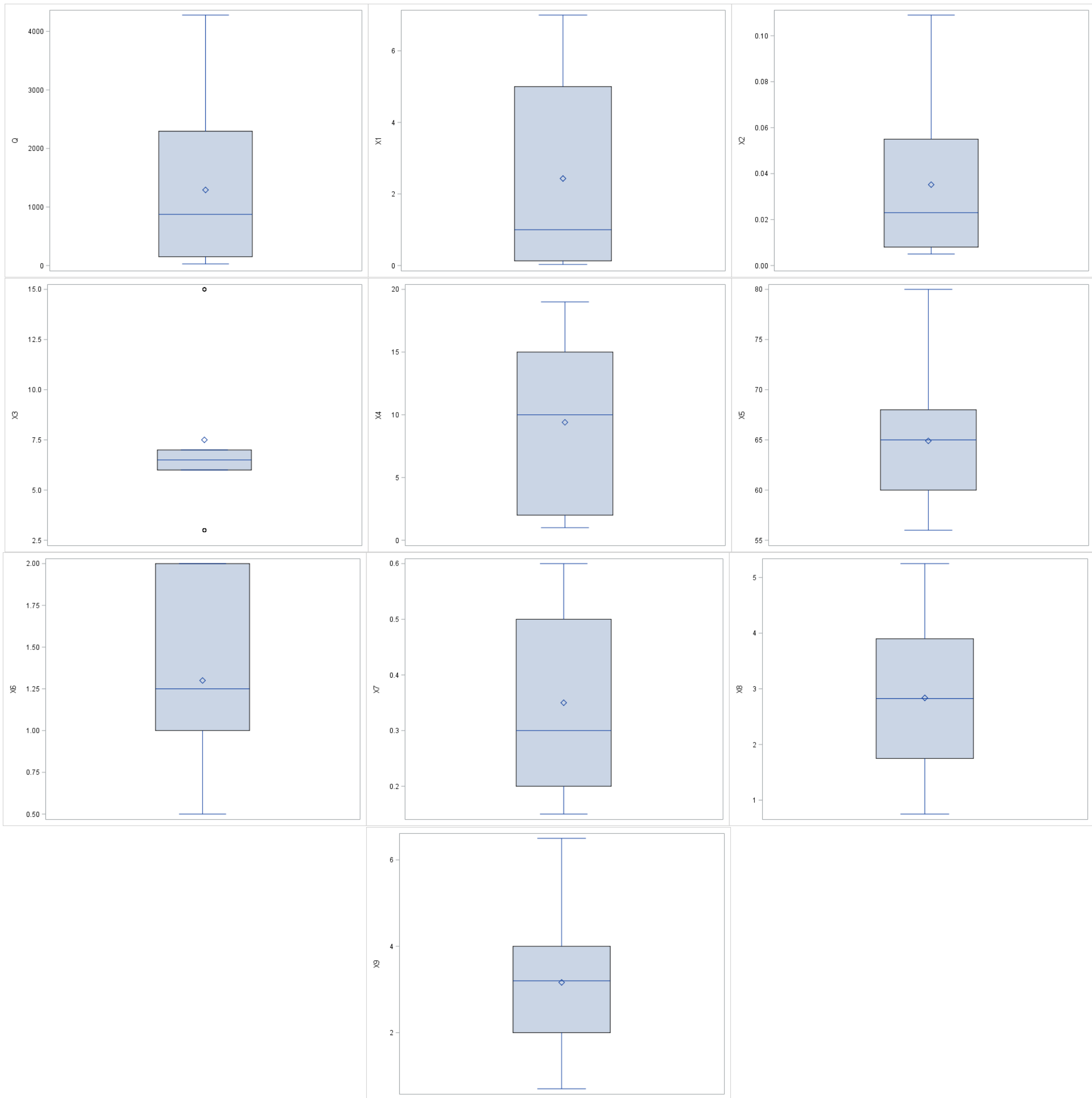


Figure 1.2 Boxplot for Q (1st row left), X1 (1st row center), X2 (1st row right), X3 (2nd row left), X4 (2nd row center), X5 (2nd row right), X6 (3rd row left), X7 (3rd row center), X8 (3rd row right), X9 (4th row center)

The boxplots appear to reinforce the skewness that the histograms showed except for X5 and X7. X5 does not appear to be very skewed and X7 has a right skew.

2. Correlation Analysis on Numerical Features

Spearman Correlation Coefficients, N = 30 Prob > |r| under H0: Rho=0										
	Q	X1	X2	X3	X4	X5	X6	X7	X8	X9
Q	1.00000	0.87753 <.0001	0.79751 <.0001	0.48664 0.0064	0.88503 <.0001	-0.69406 <.0001	0.03113 0.8703	0.52995 0.0026	0.31407 0.0910	0.20064 0.2877
X1	0.87753 <.0001	1.00000	0.88004 <.0001	0.37592 0.0406	0.99689 <.0001	-0.87039 <.0001	0.09495 0.6177	0.63751 0.0002	0.13998 0.4607	0.15678 0.4080
X2	0.79751 <.0001	0.88004 <.0001	1.00000	0.15782 0.4049	0.86426 <.0001	-0.65750 <.0001	-0.06272 0.7420	0.65643 <.0001	0.12186 0.5212	0.12855 0.4984
X3	0.48664 0.0064	0.37592 0.0406	0.15782 0.4049	1.00000	0.40905 0.0248	-0.43693 0.0158	-0.03247 0.8648	-0.26928 0.1502	0.01115 0.9534	-0.07374 0.6986
X4	0.88503 <.0001	0.99689 <.0001	0.86426 <.0001	0.40905 0.0248	1.00000	-0.85453 <.0001	0.06350 0.7389	0.62696 0.0002	0.14859 0.4332	0.16757 0.3761
X5	-0.69406 <.0001	-0.87039 <.0001	-0.65750 <.0001	-0.43693 0.0158	-0.85453 <.0001	1.00000	-0.31456 0.0905	-0.43171 0.0172	-0.02870 0.8803	-0.02824 0.8822
X6	0.03113 0.8703	0.09495 0.6177	-0.06272 0.7420	-0.03247 0.8648	0.06350 0.7389	-0.31456 0.0905	1.00000	0.17517 0.3545	0.09003 0.6361	-0.02791 0.8836
X7	0.52995 0.0026	0.63751 0.0002	0.65643 <.0001	-0.26928 0.1502	0.62696 0.0002	-0.43171 0.0172	0.17517 0.3545	1.00000	0.14667 0.4393	0.21047 0.2643
X8	0.31407 0.0910	0.13998 0.4607	0.12186 0.5212	0.01115 0.9534	0.14859 0.4332	-0.02870 0.8803	0.09003 0.6361	0.14667 0.4393	1.00000	0.87802 <.0001
X9	0.20064 0.2877	0.15678 0.4080	0.12855 0.4984	-0.07374 0.6986	0.16757 0.3761	-0.02824 0.8822	-0.02791 0.8836	0.21047 0.2643	0.87802 <.0001	1.00000

Figures 2.1 Correlation Analysis Data for Q and X1-X9

After reviewing the correlation data, it confirms all my initial predictions for correlations with the variable Q. Q and X1, X2, X3, X4, X5, and X7 all have a p-value < 0.05, so a correlation is likely. There appears to be a couple other correlations as well, but I will not be focusing on these going forward: X1 and X2, X3, X4, X5, X7; X2 and X4, X5, and X7; X3 and X4, and X5; X4 and X5, and X7; X5 and X7; X8 and X9.

Variable	DF	Parameter Estimate
Intercept	1	3.40226
X1	1	-0.01353
X2	1	-1.02366
X3	1	0.17797
X4	1	0.10879
X5	1	-0.00962
X6	1	-0.38947
X7	1	4.23348
X8	1	0.63007
X9	1	-0.46228

Figure 2.2 Projected values of model if to be implemented

3. Regression Analysis

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	34143008	3793668	10.22	<.0001
Error	20	7425127	371256		
Corrected Total	29	41568135			

Root MSE	609.30811	R-Square	0.8214
Dependent Mean	1291.23333	Adj R-Sq	0.7410
Coeff Var	47.18807		

Figure 3.1 ANOVA Table for model

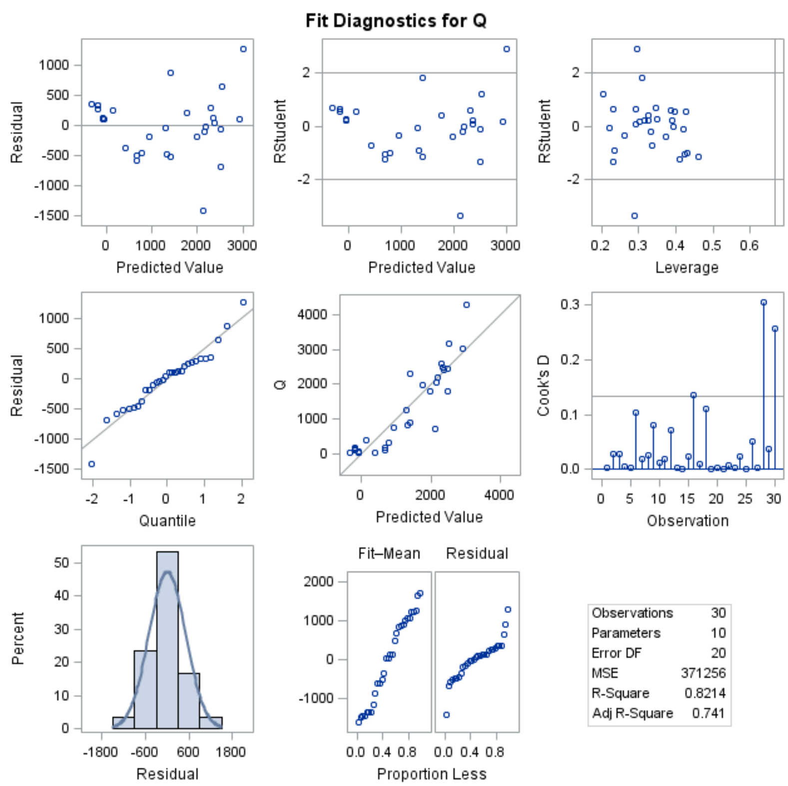
The F-test in the table shows a significant value of <.0001 returned. This is below .05, which suggests that the full model should be considered. The Adjusted R-Sq value is 0.7410.

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	292.56090	4428.61753	0.07	0.9480	0
X1	1	-203.14370	410.26785	-0.50	0.6259	101.85971
X2	1	1055.78221	9833.70009	0.11	0.9156	7.52471
X3	1	-49.23964	156.20014	-0.32	0.7558	31.44639
X4	1	209.76225	162.04557	1.29	0.2103	105.75471
X5	1	-10.19673	51.08770	-0.20	0.8438	9.67828
X6	1	-24.55815	303.52900	-0.08	0.9363	2.30786
X7	1	142.77797	3288.44271	0.04	0.9658	20.53505
X8	1	511.71277	209.74144	2.44	0.0241	5.50498
X9	1	-301.87185	171.99595	-1.76	0.0945	5.75225

Collinearity Diagnostics													
			Proportion of Variation										
Number	Eigenvalue	Condition Index	Intercept	X1	X2	X3	X4	X5	X6	X7	X8	X9
1	8.30837	1.00000	0.00000846	0.00005329	0.00063552	0.00007602	0.00004150	0.00001418	0.00081097	0.00010661	0.00036879	0.00041520
2	0.91465	3.01391	0.00003728	0.00195	0.01170	0.00057391	0.00046516	0.00010818	0.00355	0.00018172	0.00147	0.00118
3	0.31571	5.12998	0.00000696	0.00001602	0.00008258	0.00974	0.00047815	6.076062E-9	0.00818	0.00029280	0.01758	0.02918
4	0.20278	6.40097	0.00011636	0.00036917	0.02265	0.00699	0.00203	0.00042073	0.05941	0.00373	0.01108	0.01969
5	0.12835	8.04550	0.00017558	0.00521	0.09237	0.00098186	0.00006328	0.00076041	0.26928	0.00069137	0.00072462	0.00032617
6	0.08392	9.95002	0.00086533	0.00582	0.27683	0.00265	0.00171	0.00168	0.07217	0.00402	0.01666	0.00599
7	0.02446	18.43115	0.00123	0.03945	0.05574	0.00074067	0.00793	0.00333	0.00491	0.14851	0.09954	0.09952
8	0.01726	21.93742	0.00065755	0.00743	0.04086	0.00035541	0.00494	0.00023643	0.00987	0.03848	0.75719	0.74893
9	0.00415	44.71890	0.00706	0.15914	0.04939	0.17580	0.34454	0.03964	0.37942	0.24699	0.01936	0.00552
10	0.00033833	156.70609	0.98984	0.78056	0.44974	0.80209	0.63780	0.95381	0.19240	0.55700	0.07603	0.08924

Figure 3.2 Parameter Estimates and Eigenvalue for full model

Looking at the Variance Inflation field, you can see multiple values >10, which raises flags for multicollinearity. In addition, the Condition Index gets to very high values >30 so we must consider a fix.



Durbin-Watson D	2.090
Pr < DW	0.1823
Pr > DW	0.8177
Number of Observations	30
1st Order Autocorrelation	-0.155

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.931552	Pr < W	0.0540
Kolmogorov-Smirnov	D	0.15683	Pr > D	0.0587
Cramer-von Mises	W-Sq	0.115689	Pr > W-Sq	0.0691
Anderson-Darling	A-Sq	0.736679	Pr > A-Sq	0.0490

```
         D'AGOSTINO TEST OF NORMALITY FOR VARIABLE D, N=30
       G1=-0.29738  SQRTB1=-0.28230  Z=-0.73161     P=0.4644
       G2=2.97529   B2=5.30847       Z= 2.38594     P=0.0170
       K**2=CHISQ(2 DF)= 6.22794                    P=0.0444
```

Figure 3.3 Model Diagnostic Tests

From the figures above, there seems to be some evidence of heteroscedasticity violation since the data points in the top 2 scatterplot seem to slope downwards. There seems to be no autoregressive effects (Pr < DW and Pr > DW are both greater than .05). Looking at the normality charts, some of the values seem to be below .05, so there may be a violation to normality.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	67.63526	7.51503	40.00	<.0001
Error	20	3.75729	0.18786		
Corrected Total	29	71.39255			

Root MSE	0.43343	R-Square	0.9474
Dependent Mean	6.36677	Adj R-Sq	0.9237
Coeff Var	6.80775		

Figure 3.4 log(Q) ANOVA Table for model

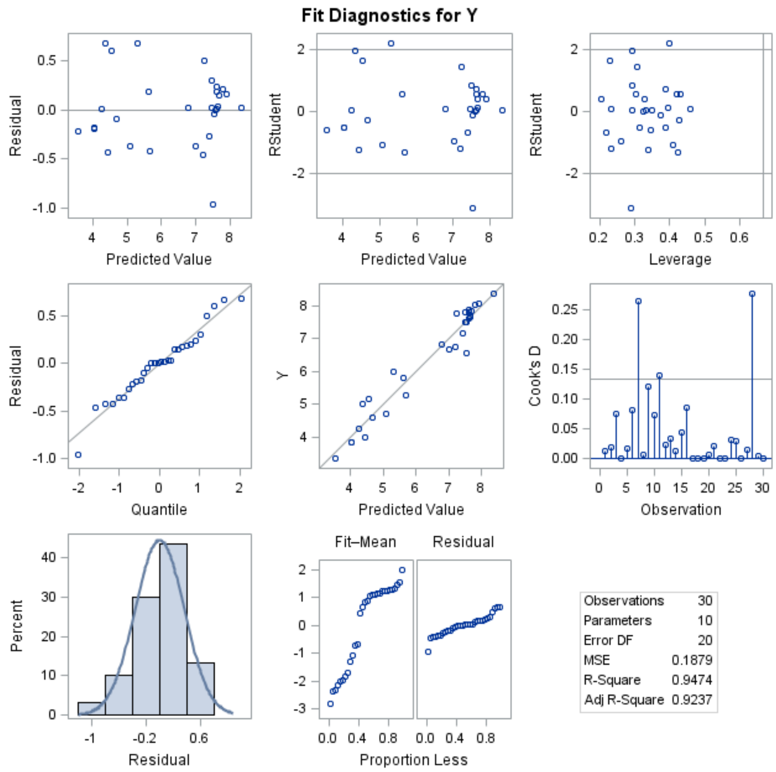
The F-test in the table shows a significant value of <.0001 returned. This is below .05, which suggests that the full model should be considered. The Adjusted R-Sq value is 0.9237.

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	3.40226	3.15031	1.08	0.2930	0
X1	1	-0.01353	0.29185	-0.05	0.9635	101.85971
X2	1	-1.02366	6.99524	-0.15	0.8851	7.52471
X3	1	0.17797	0.11111	1.60	0.1249	31.44639
X4	1	0.10879	0.11527	0.94	0.3566	105.75471
X5	1	-0.00962	0.03634	-0.26	0.7939	9.67828
X6	1	-0.38947	0.21592	-1.80	0.0863	2.30786
X7	1	4.23348	2.33924	1.81	0.0854	20.53505
X8	1	0.63007	0.14920	4.22	0.0004	5.50498
X9	1	-0.46228	0.12235	-3.78	0.0012	5.75225

Collinearity Diagnostics												
			Proportion of Variation									
Number	Eigenvalue	Condition Index	Intercept	X1	X2	X3	X4	X5	X6	X7	X8	X9
1	8.30837	1.00000	0.00000846	0.00005329	0.00063552	0.00007602	0.00004150	0.00001418	0.00081097	0.00010661	0.00036879	0.00041520
2	0.91465	3.01391	0.00003728	0.00195	0.01170	0.00057391	0.00046516	0.00010818	0.00355	0.00018172	0.00147	0.00118
3	0.31571	5.12998	0.00000696	0.00001602	0.00008258	0.00974	0.00047815	6.076062E-9	0.00818	0.00029280	0.01758	0.02918
4	0.20278	6.40097	0.00011636	0.00036917	0.02265	0.00699	0.00203	0.00042073	0.05941	0.00373	0.01108	0.01969
5	0.12835	8.04550	0.00017558	0.00521	0.09237	0.00098186	0.00006328	0.00076041	0.26928	0.00069137	0.00072462	0.00032617
6	0.08392	9.95002	0.00086533	0.00582	0.27683	0.00265	0.00171	0.00168	0.07217	0.00402	0.01666	0.00599
7	0.02446	18.43115	0.00123	0.03945	0.05574	0.00074067	0.00793	0.00333	0.00491	0.14851	0.09954	0.09952
8	0.01726	21.93742	0.00065755	0.00743	0.04086	0.00035541	0.00494	0.00023643	0.00987	0.03848	0.75719	0.74893
9	0.00415	44.71890	0.00706	0.15914	0.04939	0.17580	0.34454	0.03964	0.37942	0.24699	0.01936	0.00552
10	0.00033833	156.70609	0.98984	0.78056	0.44974	0.80209	0.63780	0.95381	0.19240	0.55700	0.07603	0.08924

Figure 3.5 Log(Q) Parameter Estimates and Eigenvalue for full model

Looking at the Variance Inflation field, you can see multiple values >10, which raises flags for multicollinearity. In addition, the Condition Index gets to very high values >30 so we must look into it.



Durbin-Watson D	2.053
Pr < DW	0.1563
Pr > DW	0.8437
Number of Observations	30
1st Order Autocorrelation	-0.031

Tests for Normality

Test	Statistic		p Value	
Shapiro-Wilk	W	0.958445	Pr < W	0.2825
Kolmogorov-Smirnov	D	0.101556	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.068521	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.438196	Pr > A-Sq	>0.2500

```
        D'AGOSTINO TEST OF NORMALITY FOR VARIABLE D, N=30
G1=-0.35607   SQRTB1=-0.33802   Z=-0.87220      P=0.3831
G2=1.51891    B2=4.08376        Z= 1.62280      P=0.1046
K**2=CHISQ(2 DF)= 3.39420                       P=0.1832
```

Figure 3.6 Log(Q) Model Diagnostic Tests

From the figures above, there seems to be no more evidence of heteroscedasticity since the data points in the scatterplot seem to be evenly distributed within the constraints. There seems to be no autoregressive effects (Pr < DW and Pr > DW are both greater than .05). Looking at the normality charts, all values appear to be above .05 so there are no major violations there. Overall, the log model seems like a better fit.

There does still appear to be multicollinearity, so we will try centering first.

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	6.36677	0.07913	80.46	<.0001	0
X1	1	-0.01353	0.29185	-0.05	0.9635	101.85971
X2	1	-1.02366	6.99524	-0.15	0.8851	7.52471
X3	1	0.17797	0.11111	1.60	0.1249	31.44639
X4	1	0.10879	0.11527	0.94	0.3566	105.75471
X5	1	-0.00962	0.03634	-0.26	0.7939	9.67828
X6	1	-0.38947	0.21592	-1.80	0.0863	2.30786
X7	1	4.23348	2.33924	1.81	0.0854	20.53505
X8	1	0.63007	0.14920	4.22	0.0004	5.50498
X9	1	-0.46228	0.12235	-3.78	0.0012	5.75225

Figure 3.7 Centering the Regressors

Looking at the results of centering, the Variance inflation value still appears to be >>10. For this reason, we must continue to look for another option.

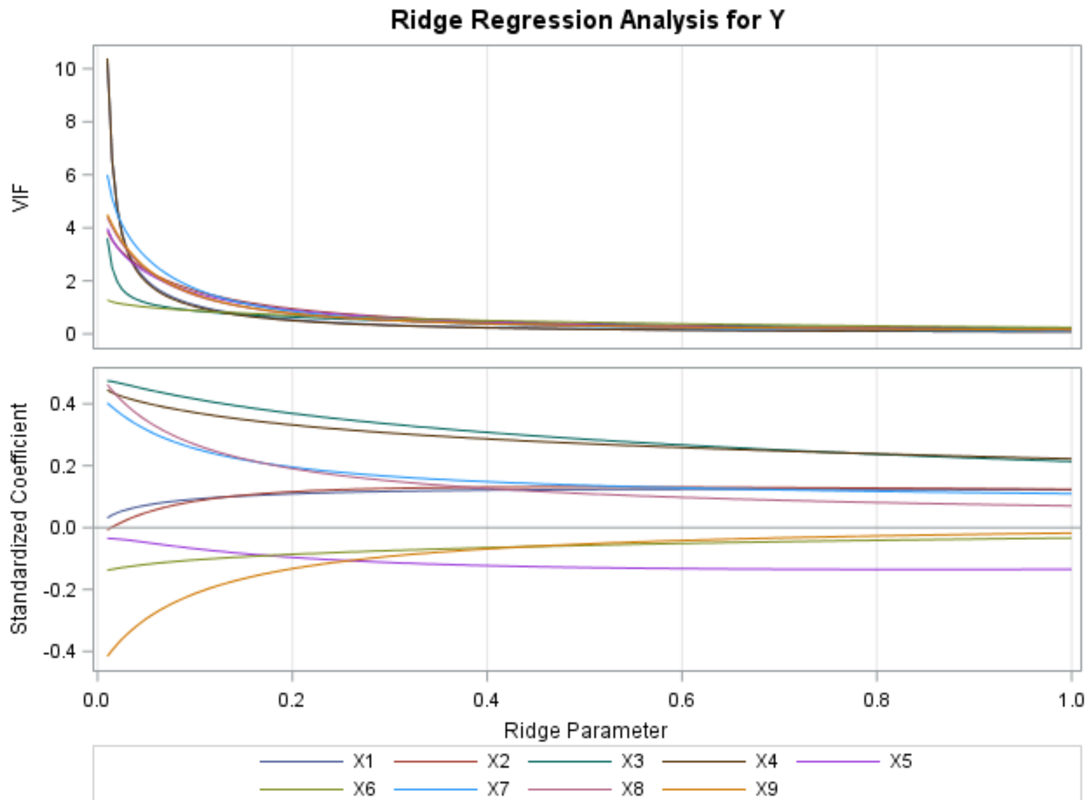


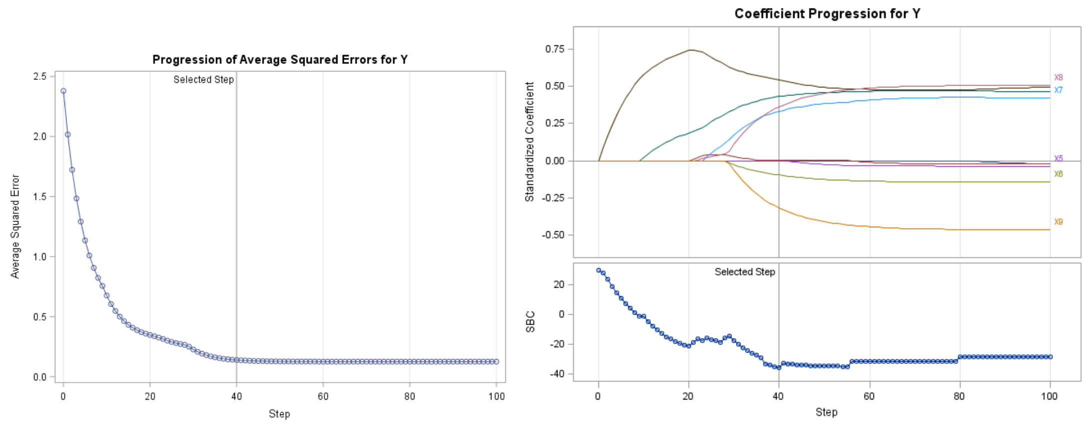
Figure 3.8 Ridge Regression Analysis

Looking at the Ridge Regression Analysis, we choose to delete X6 and X9 (yellow and green lines) because they happen to be the closest plots to zero.

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	1.97737	3.60999	0.55	0.5894	0
X1	1	-0.07273	0.28059	-0.26	0.7979	55.31273
X2	1	2.63273	8.12017	0.32	0.7488	5.95660
X3	1	0.15650	0.10200	1.53	0.1392	15.56745
X4	1	0.15194	0.10245	1.48	0.1522	49.07425
X5	1	0.00962	0.04251	0.23	0.8231	7.78127
X7	1	2.66582	2.30117	1.16	0.2591	11.67415
X8	1	0.11076	0.08592	1.29	0.2108	1.07250

Figure 3.9 Deleted X6 and X9

The variance inflation is a lot closer to 10 this time, so it had a definite improvement.



Parameter Estimates		Parameter Estimates		
Parameter	Estimate	Parameter	DF	Estimate
Intercept	2.901367	Intercept	1	2.883875
X3	0.166821	X3	1	0.168245
X4	0.118523	X4	1	0.117723
X6	-0.265762	X6	1	-0.274126
X7	3.347102	X7	1	3.408643
X8	0.447638	X8	1	0.460938
X9	-0.314800	X9	1	-0.325568

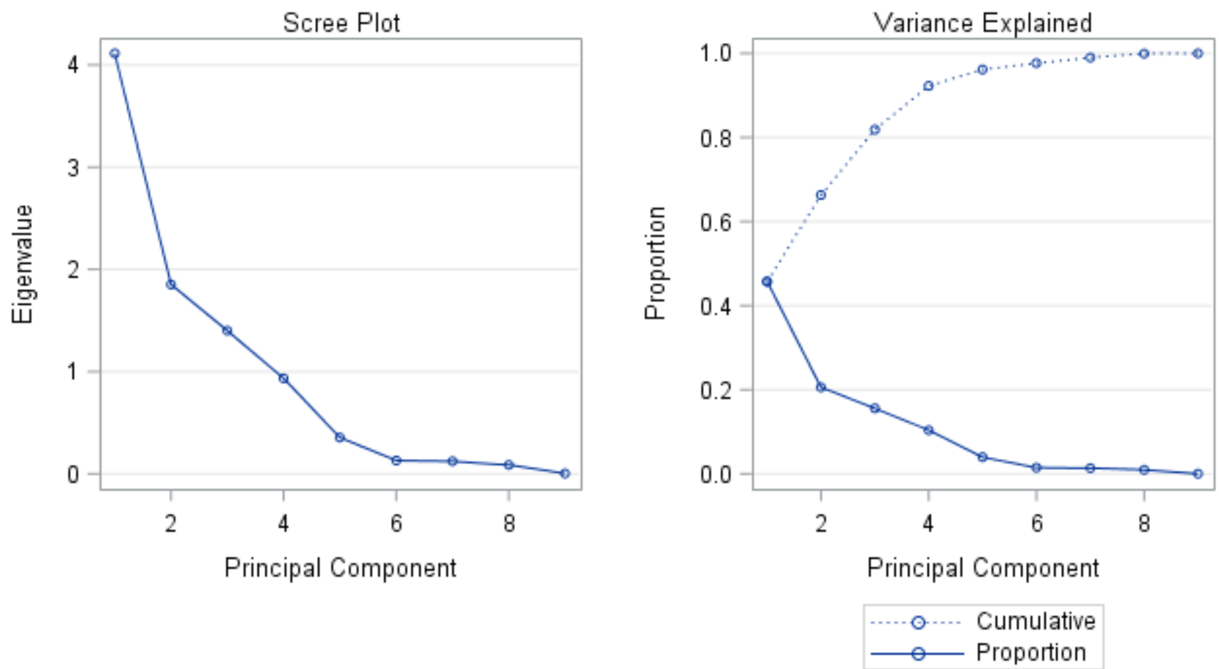
Figure 3.10 LASSO (Left) and Elastic Net (Right) Selection Summary

So the Model Selected by Lasso and Elastic Net is: y = 2.9 + 0.17(X3) + 0.12(X4) – 0.27(X6) + 3.4(X7) + 0.45(X8) – 0.32(X9)

Obs	_MODEL_	_TYPE_	_DEPVAR_	_RIDGE_	_PCOMIT_	_RMSE_	Intercept	X1	X2	X3	X4	X5	X6	X7	X8	X9	Y
3	MODEL1	RIDGE	Y	0.4	.	0.66834	5.79184	0.068243	6.48625	0.11888	0.062631	-0.027995	-0.17984	1.49190	0.15686	-0.068354	-1

Figure 3.11 Ridge Regression

The estimated model is Y = 5.79 + 0.07(X1) + 6.49(X2) + 0.12(X3) + 0.06(X4) – 0.03(X5) – 0.18(X6) + 1.49(X7) + 0.16(X8) - 0.07(X9)



Obs	_MODEL_	_TYPE_	_DEPVAR_	_RIDGE_	_PCOMIT_	_RMSE_	Intercept	X1	X2	X3	X4	X5	X6	X7	X8	X9	Y
2	MODEL1	IPC	Y	.	4	0.61141	5.10980	0.054405	15.9396	0.16807	0.062147	-0.023155	-0.27104	0.79470	0.088080	0.014513	-1

Figure 3.12 Principal Component Regression Results

For the Principal Component, I chose 4 as the principal component because it appears right before the elbow. The estimated model is therefore: Y = 5.11 + 0.05(X1) + 15.94(X2) + 0.17(X3) + 0.06(X4) – 0.02(X5) – 0.27(X6) + 0.79(X7) + 0.09(X8) + 0.015(X9)

In conclusion, I am picking the GroupLasso model and will drop X1, X2, X5

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	67.59120	11.26520	68.16	<.0001
Error	23	3.80135	0.16528		
Corrected Total	29	71.39255			

Root MSE	0.40654	R-Square	0.9468
Dependent Mean	6.36677	Adj R-Sq	0.9329
Coeff Var	6.38537		

Figure 3.13 ANOVA Table for final model

The F-test in the table shows a significant value of <.0001 returned. This is below .05, which suggests that the full model should be considered. The Adjusted R-Sq value is 0.9329.

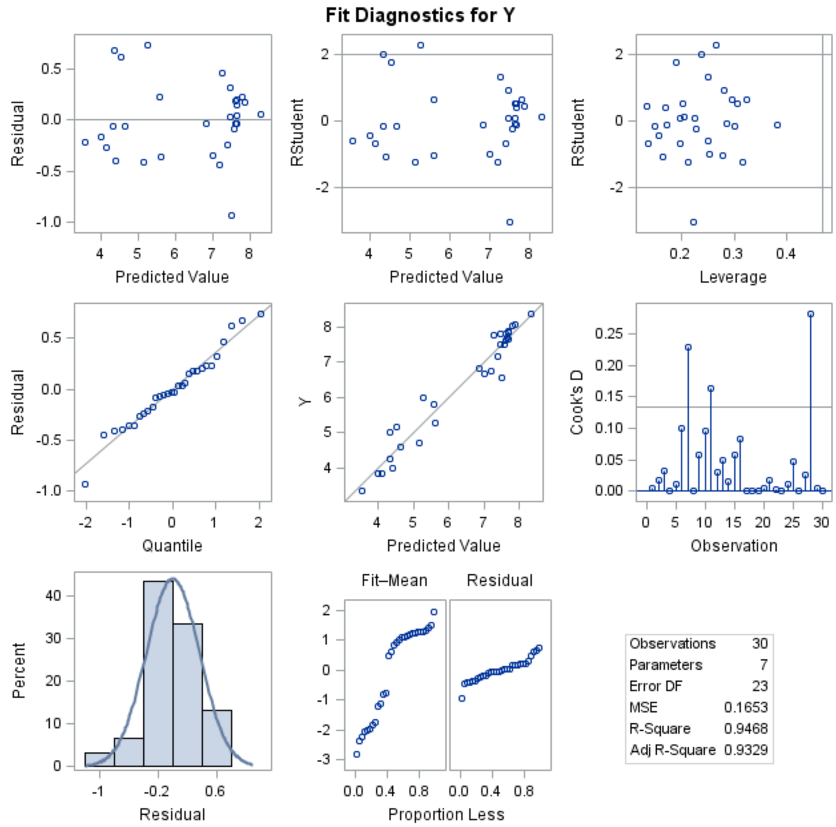
Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	2.69180	0.44521	6.05	<.0001	0
X3	1	0.18384	0.03227	5.70	<.0001	3.01431
X4	1	0.10905	0.02569	4.24	0.0003	5.97259
X6	1	-0.36752	0.14552	-2.53	0.0189	1.19160
X7	1	4.08497	1.21317	3.37	0.0027	6.27802
X8	1	0.61161	0.13256	4.61	0.0001	4.93953
X9	1	-0.44764	0.10826	-4.13	0.0004	5.11936

Collinearity Diagnostics									
			Proportion of Variation						
Number	Eigenvalue	Condition Index	Intercept	X3	X4	X6	X7	X8	X9
1	6.06604	1.00000	0.00071470	0.00158	0.00129	0.00305	0.00062554	0.00080150	0.00090662
2	0.33835	4.23420	0.00256	0.00579	0.10406	0.02244	0.01173	0.00840	0.00657
3	0.31443	4.39227	0.00078277	0.11754	0.00058539	0.02635	0.00325	0.01523	0.02814
4	0.18092	5.79041	0.00432	0.08724	0.01712	0.30810	0.01884	0.00996	0.01855
5	0.07065	9.26602	0.17673	0.00301	0.07178	0.54827	0.04521	0.01144	0.00899
6	0.01847	18.12129	0.00014492	0.00896	0.00534	0.03065	0.00000748	0.94564	0.93663
7	0.01114	23.33683	0.81475	0.77588	0.79982	0.06115	0.92034	0.00853	0.00020729

Figure 3.14 Parameter Estimates for final model

It would appear from these two tables that all Variance Inflation is now below 10 and the Condition Index is all below 30, so there are no more violations to multicollinearity.

Durbin-Watson D	2.035
Pr < DW	0.3226
Pr > DW	0.6774
Number of Observations	30
1st Order Autocorrelation	-0.022



Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	W	0.963115	Pr < W	0.3711
Kolmogorov-Smirnov	D	0.103326	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.053107	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.384493	Pr > A-Sq	>0.2500

```
        D'AGOSTINO TEST OF NORMALITY FOR VARIABLE D, N=30
      G1=-0.19046   SQRTB1=-0.18080   Z=-0.47141      P=0.6374
      G2=1.38692    B2=3.97276        Z= 1.53432      P=0.1250
      K**2=CHISQ(2 DF)= 2.57637                       P=0.2758
```

Figure 3.15 Final Model Diagnostic Tests

From the figures above, there seems to be no evidence of serious heteroscedasticity since the data points in the scatterplot are evenly and randomly distributed within the constraints. There seems to be no autoregressive effects (Pr < DW and Pr > DW are both greater than .05). Looking at the normality charts, all values appear to be above .05 so there are no major violations there. Overall, this final model looks like a really good fit.

Based on the numbers that this new model provided it appears that I no longer need to deal with the multicollinearity because the problem variables have all been dropped.

SAS Code:

```
%MACRO NORMTEST(VAR,DATA);
/*********************************************************************************/
/* Macro NORMTEST is revised from the code in D'Agostino's paper.              */
/* "A Suggestion for Using Powerful and Informative Tests of Normality"         */
/* Author(s): Ralph B. D'Agostino, Albert Belanger, and Ralph B. D'Agostino Jr.  */
/* Source: The American Statistician, Vol. 44, No. 4 (Nov., 1990), pp. 316-321   */

/* It provides five hypothesis tests                                  */
/* (1) Shapiro-Wilk test                                            */
/* (2) Kolmogorov-Smirnov test                                       */
/* (3) Cramer-von Mises test                                         */
/* (4) Anderson-Darling                                             */
/* (5) D'Agostino's K^2                                             */
/* For details about the first four tests, users are referred to SAS online doc  */
/* under UNIVARIATE procedure. As for D'Agostino's test, please refer to the art.*/
/* mentioned above.                                                 */
/* Revised by Ping-Shi Wu Dec. 2015 @ Lehigh University                        */
/*********************************************************************************/
 ODS NOPROCTITLE;
 ODS GRAPHICS /BORDER=OFF;
 ODS SELECT Moments Histogram QQPlot CDFPlot;
 TITLE "NORMAL-TEST";
 PROC UNIVARIATE DATA=&DATA NORMAL;
  VAR &VAR;
        HISTOGRAM &VAR/NORMAL(MU=EST SIGMA=EST) KERNEL;
  QQPLOT &VAR/NORMAL(MU=EST SIGMA=EST);
  CDFPLOT &VAR/NORMAL(MU=EST SIGMA=EST);
  OUTPUT OUT=XXSTAT N=N MEAN=XBAR STD=S SKEWNESS=G1 KURTOSIS=G2;
 RUN;
 ODS SELECT TestsForNormality;
 PROC UNIVARIATE DATA=&DATA NORMAL;
  VAR &VAR;
 RUN;
 TITLE;
 OPTIONS LS=80;
 DATA _NULL_;
  SET XXSTAT;
  SQRTB1=(N-2)/SQRT(N*(N-1))*G1;
  Y=SQRTB1*SQRT((N+1)*(N+3)/(6*(N-2)));
  BETA2=3*(N*N+27*N-70)*(N+1)*(N+3)/((N-2)*(N+5)*(N+7)*(N+9));
  W=SQRT(-1+SQRT(2*(BETA2-1)));
  DELTA=1/SQRT(LOG(W));
  ALPHA=SQRT(2/(W*W-1));
        Z_B1=DELTA*LOG(Y/ALPHA+SQRT((Y/ALPHA)**2+1));
  B2=3*(N-1)/(N+1)+(N-2)*(N-3)/((N+1)*(N-1))*G2;
  MEANB2=3*(N-1)/(N+1);
  VARB2= 24*N*(N-2)*(N-3)/((N+1)*(N+1)*(N+3)*(N+5));
  X=(B2-MEANB2)/SQRT(VARB2);
  MOMENT=6*(N*N-5*N+2)/((N+7)*(N+9))*SQRT(6*(N+3)*(N+5)/(N*(N-2)*(N-3)));
  A=6+8/MOMENT*(2/MOMENT+SQRT(1+4/(MOMENT**2)));
  Z_B2=(1-2/(9*A)-((1-2/A)/(1+ X*SQRT(2/(A-4))))**(1/3))/SQRT(2/(9*A));
  PRZB1=2*(1-PROBNORM(ABS(Z_B1)));
  PRZB2=2*(1-PROBNORM(ABS(Z_B2)));
  CHITEST=Z_B1*Z_B1 + Z_B2*Z_B2;
```

```
   PRCHI=1-PROBCHI(CHITEST,2);
   FILE PRINT;
   PUT @22 "D'AGOSTINO TEST OF NORMALITY FOR VARIABLE &VAR, "
   N = /@20 G1=8.5 @33 SQRTB1 =8.5 @50 "Z=" Z_B1 8.5 @65 "P=" PRZB1 6.4
      /@20 G2=8.5 @33 B2=8.5 @50 "Z=" Z_B2 8.5 @65 "P=" PRZB2 6.4
      /@20 "K**2=CHISQ(2 DF)=" CHITEST 8.5 @65 "P=" PRCHI 6.4;
 RUN;
 TITLE;
%MEND NORMTEST;
DATA FLOW;
 INPUT X1 X2 X3 X4 X5 X6 X7 X8 X9 Q ;
 Y=LOG(Q);
DATALINES;
.03 .006 3.0 1 70 1.5 .25 1.75 2.0 46
.03 .006 3.0 1 70 1.5 .25 2.25 3.7 28
.03 .006 3.0 1 70 1.5 .25 4.00 4.2 54
.03 .021 3.0 1 80 1.0 .25 1.60 1.5 70
.03 .021 3.0 1 80 1.0 .25 3.10 4.0 47
.03 .021 3.0 1 80 1.0 .25 3.60 2.4 112
.13 .005 6.5 2 65 2.0 .35 1.25 .7 398
.13 .005 6.5 2 65 2.0 .35 2.30 3.5 98
.13 .005 6.5 2 65 2.0 .35 4.25 4.0 191
.13 .008 6.5 2 68 .5 .15 1.45 2.0 171
.13 .008 6.5 2 68 .5 .15 2.60 4.0 150
.13 .008 6.5 2 68 .5 .15 3.90 3.0 331
1.00 .023 15.0 10 60 1.0 .20 .75 1.0 772
1.00 .023 15.0 10 60 1.0 .20 1.75 1.5 1268
1.00 .023 15.0 10 60 1.0 .20 3.25 4.0 849
1.00 .023 15.0 10 65 2.0 .20 1.80 1.0 2294
1.00 .023 15.0 10 65 2.0 .20 3.10 2.0 1984
1.00 .023 15.0 10 65 2.0 .20 4.75 6.0 900
3.00 .039 7.0 15 67 .5 .50 1.75 2.0 2181
3.00 .039 7.0 15 67 .5 .50 3.25 4.0 2484
3.00 .039 7.0 15 67 .5 .50 5.00 6.5 2450
5.00 .109 6.0 15 62 1.5 .60 1.50 1.5 1794
5.00 .109 6.0 15 62 1.5 .60 2.75 3.0 2067
5.00 .109 6.0 15 62 1.5 .60 4.20 5.0 2586
7.00 .055 6.5 19 56 2.0 .50 1.80 2.0 2410
7.00 .055 6.5 19 56 2.0 .50 3.25 4.0 1808
7.00 .055 6.5 19 56 2.0 .50 5.25 6.0 3024
7.00 .063 6.5 19 56 1.0 .50 1.25 2.0 710
7.00 .063 6.5 19 56 1.0 .50 2.90 3.4 3181
7.00 .063 6.5 19 56 1.0 .50 4.76 5.0 4279
;

PROC SGSCATTER DATA = FLOW;
 MATRIX Q X1-X9
             / ellipse
               diagonal = (histogram normal);
RUN;

/*Boxplots*/
PROC SGPLOT DATA=FLOW;
 VBOX Q;
RUN;
PROC SGPLOT DATA=FLOW;
 VBOX X1;
RUN;
PROC SGPLOT DATA=FLOW;
 VBOX X2;
RUN;
PROC SGPLOT DATA=FLOW;
 VBOX X3;
RUN;
PROC SGPLOT DATA=FLOW;
 VBOX X4;
RUN;
PROC SGPLOT DATA=FLOW;
 VBOX X5;
RUN;
PROC SGPLOT DATA=FLOW;
```

```
 VBOX X6;
RUN;
PROC SGPLOT DATA=FLOW;
 VBOX X7;
RUN;
PROC SGPLOT DATA=FLOW;
 VBOX X8;
RUN;
PROC SGPLOT DATA=FLOW;
 VBOX X9;
RUN;

/*Correlation Analysis*/
PROC CORR DATA=FLOW SPEARMAN FISHER(BIASADJ=NO);
 VAR Q X1-X9;
RUN;
QUIT;

/*Full Model Fit w/ model doagnostics*/
PROC REG DATA=FLOW;
 MODEL Q = X1-X9/DWPROB VIF COLLIN;
 OUTPUT OUT=SFM_FIT RSTUDENT=D;
RUN;
QUIT;

%NORMTEST(D,SFM_FIT)

/*Full Model Fit w/ model doagnostics*/
PROC REG DATA=FLOW;
 MODEL Y = X1-X9/DWPROB VIF COLLIN;
 OUTPUT OUT=SFM_FIT RSTUDENT=D;
RUN;
QUIT;

%NORMTEST(D,SFM_FIT)

/*TRY CENTERING FIRST*/
PROC STDIZE DATA=FLOW OUT=FLOW2 METHOD=MEAN;
 VAR X1-X9;
RUN;

PROC REG DATA=FLOW2 PLOTS=NONE;
 MODEL Y=X1-X9/COLLIN VIF;
RUN;
QUIT;

/*COLLINEAR STILL*/
/*TRY RIDGE TRACE NEXT TO SEE IF DELETION OF INSIGNIFICANT VARIABLE CAN HELP */
PROC REG DATA=FLOW OUTEST=EST_RIDGE RIDGE=0.01 TO 1 BY 0.005 OUTVIF;
 MODEL Y= X1-X9;
RUN;
QUIT;

/*TRY DELETING X3*/
PROC REG DATA=FLOW PLOTS=NONE;
 MODEL Y= X1-X5 X7 X8/VIF COLLIN;
RUN;
QUIT;

PROC GLMSELECT DATA=FLOW PLOTS=ALL;
  MODEL Y=X1-X9/SELECTION=GROUPLASSO(CHOOSE=SBC STOP=NONE) CVMETHOD=RANDOM(10);
RUN;
PROC GLMSELECT DATA=FLOW;
  MODEL Y=X1-X9/SELECTION=ELASTICNET(CHOOSE=SBC STOP=NONE) CVMETHOD=RANDOM(10);
RUN;

PROC REG DATA=FLOW OUTEST=EST_RIDGE RIDGE=0.4 OUTVIF;
 MODEL Y= X1-X9;
RUN;
QUIT;
PROC PRINT DATA=EST_RIDGE;
```

```
 WHERE _TYPE_='RIDGE';
RUN;
/*(2-A2) PC REGRESSION IF NO SELECTION OF FEATURES IS INTENDED*/
/*PCA TO DECIDE HOW MANY COMPONENTS*/
PROC PRINCOMP DATA=FLOW;
 VAR X1-X9;
RUN;
PROC REG DATA=FLOW OUTEST=EST_PCR PCOMIT=4;
 MODEL Y= X1-X9;
RUN;
QUIT;
PROC PRINT DATA=EST_PCR;
 WHERE _TYPE_='IPC';
RUN;

PROC REG DATA=FLOW;
 MODEL Y = X3 X4 X6-X9/DWPROB VIF COLLIN;
 OUTPUT OUT=SFM_FIT RSTUDENT=D;
RUN;
QUIT;
%NORMTEST(D,SFM_FIT)
```