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.

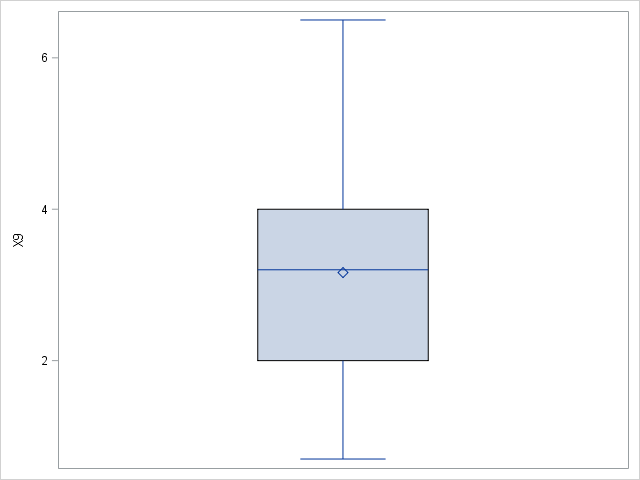
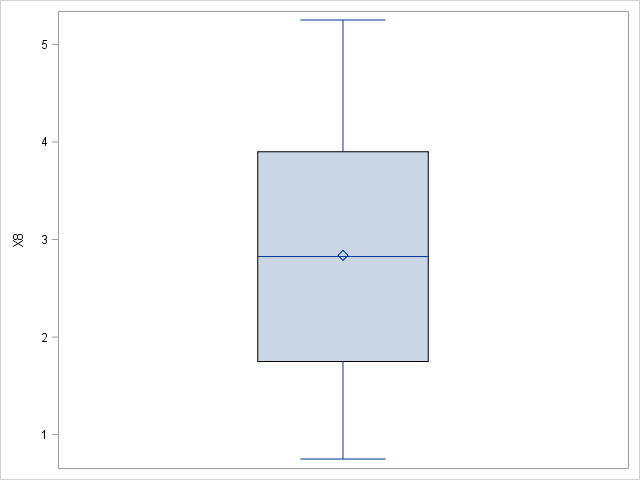
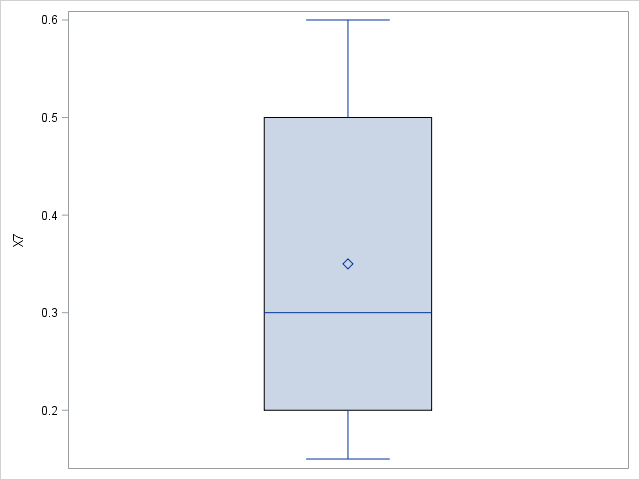
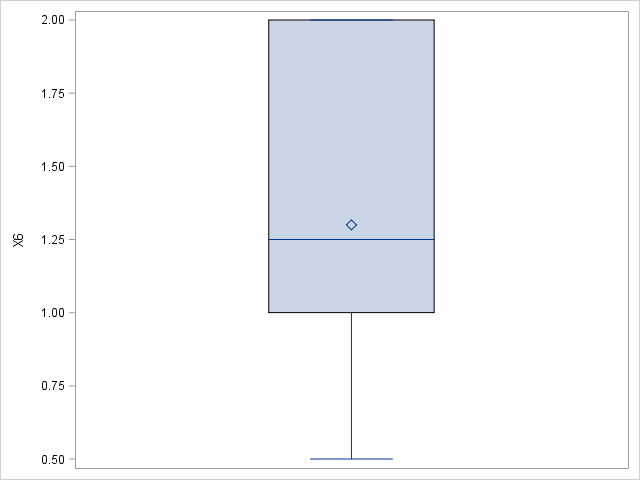
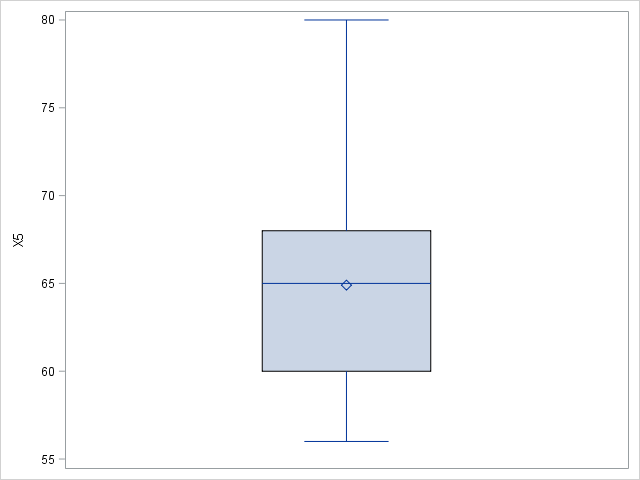
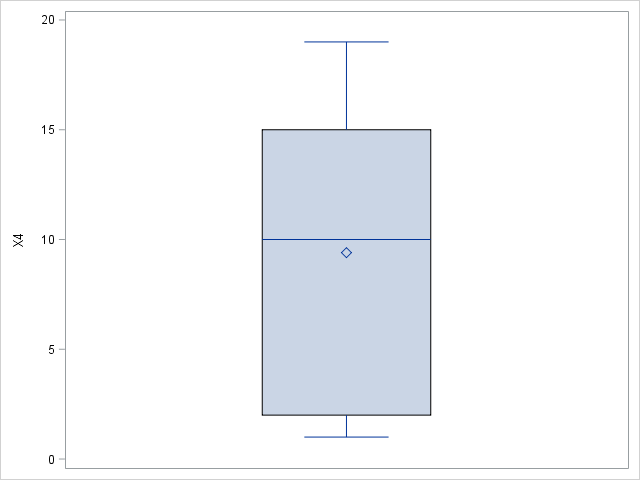
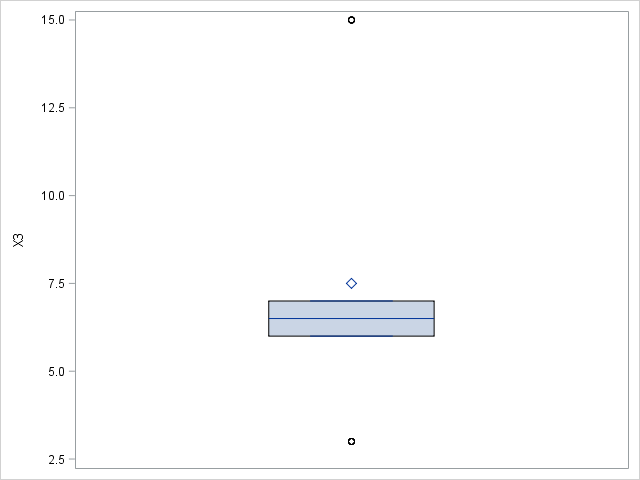
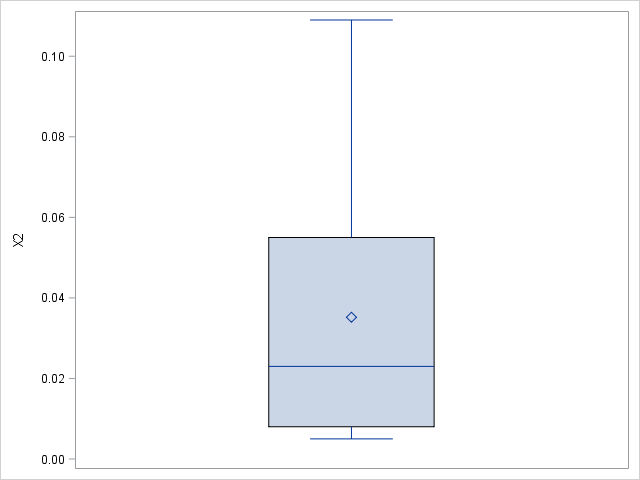
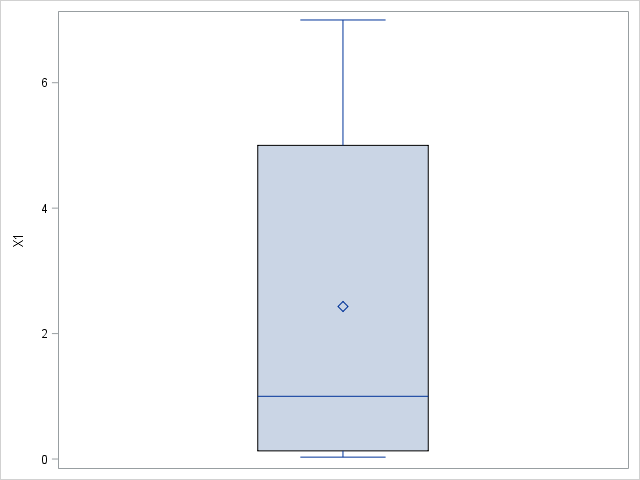
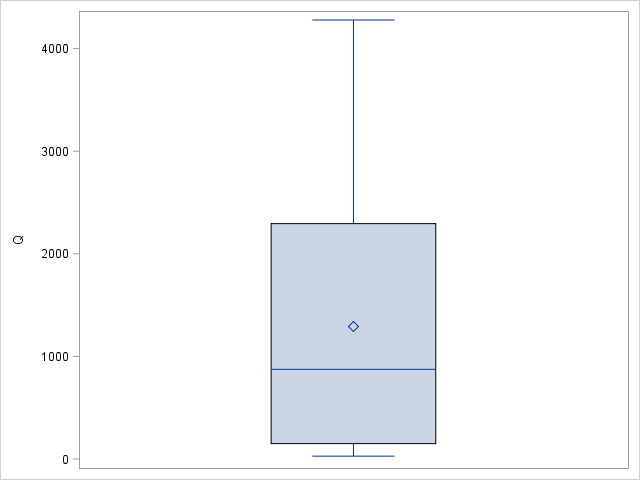
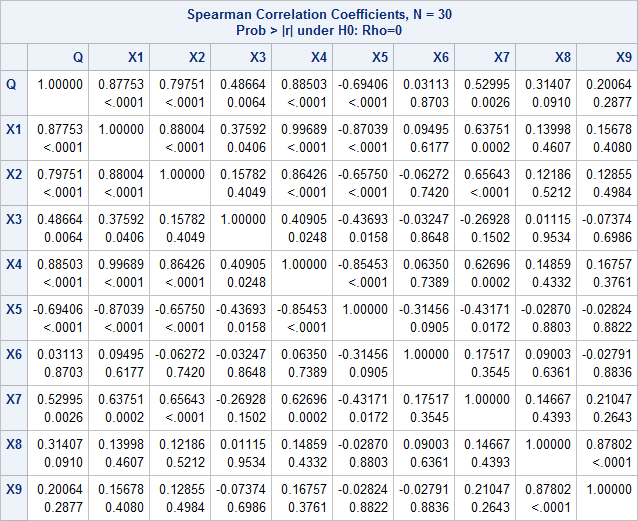


Figure 1.2 Boxplot for Q (1st row left), X1 (1st row center), X2 (1st row right), X3 (2nd row left), X4 (2nd row center), X5 (2nd row right), X6 (3rd row left), X7 (3rd row center), X8 (3rd row right), X9 (4th row center)

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



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.



Figure 2.2 Projected values of model if to be implemented

3. Regression Analysis

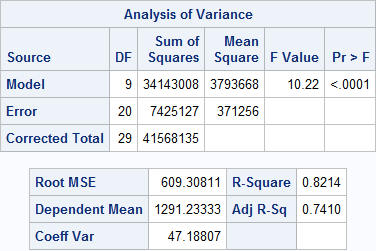
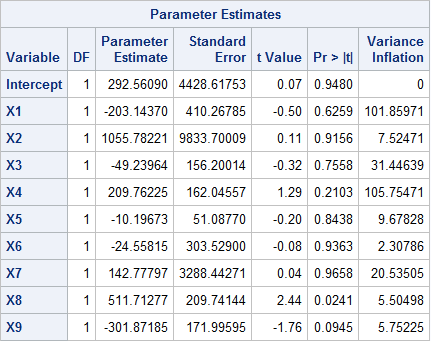


Figure 3.1 ANOVA Table for model

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.



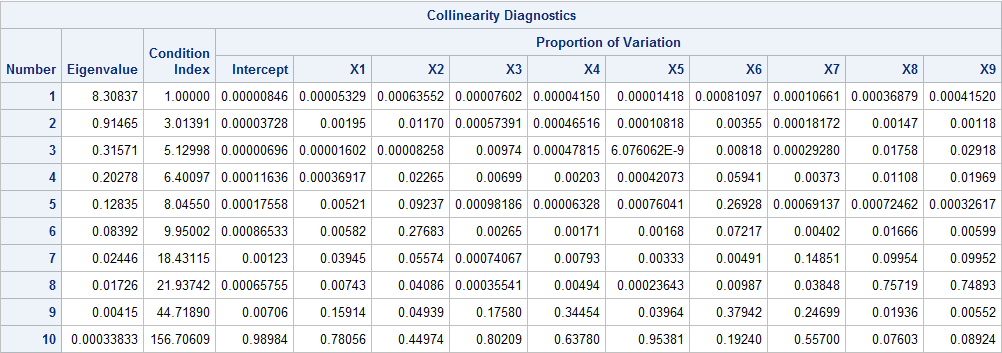
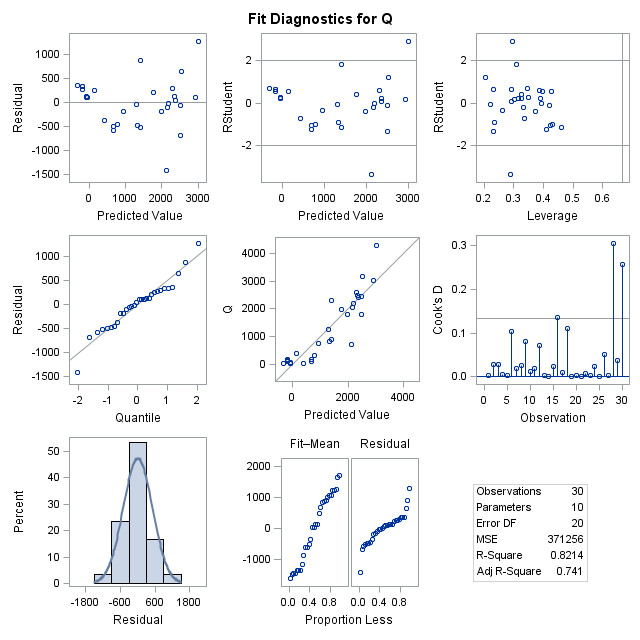


Figure 3.2 Parameter Estimates and Eigenvalue for full model

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.



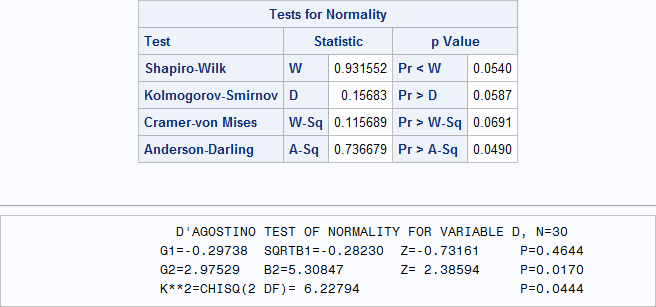


Figure 3.3 Model Diagnostic Tests

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.

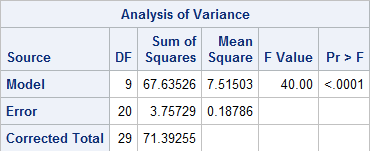


Figure 3.4 log(Q) ANOVA Table for model

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.

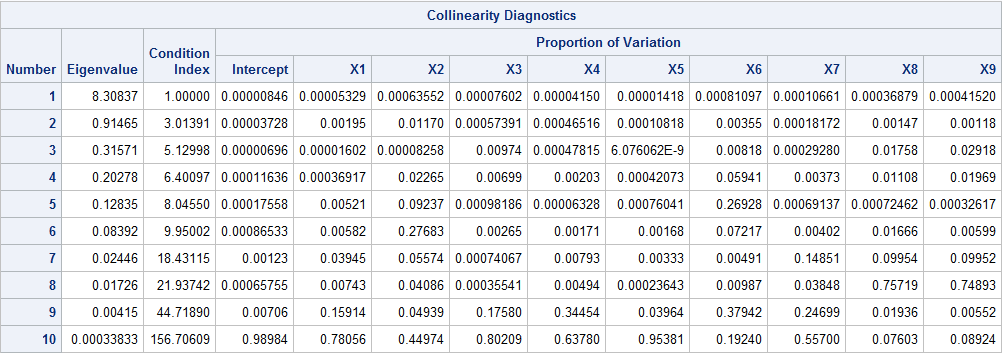
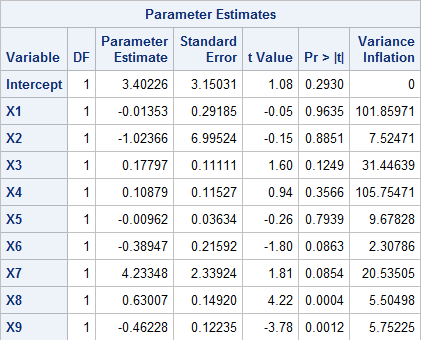
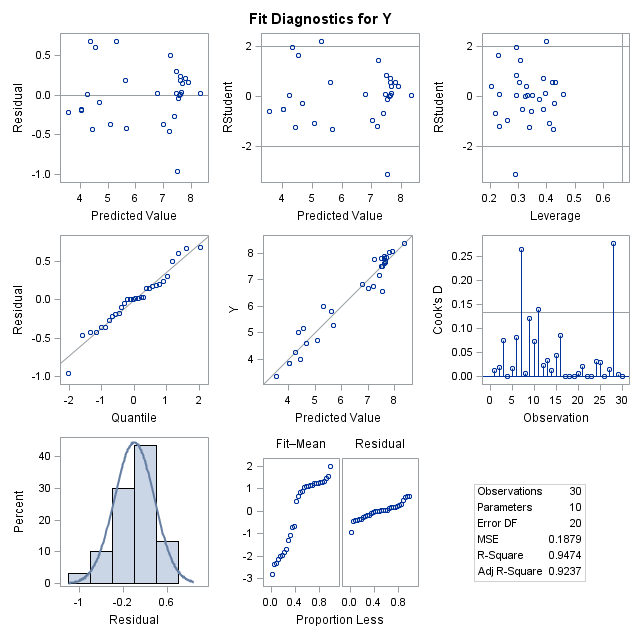


Figure 3.5 Log(Q) Parameter Estimates and Eigenvalue for full model

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.



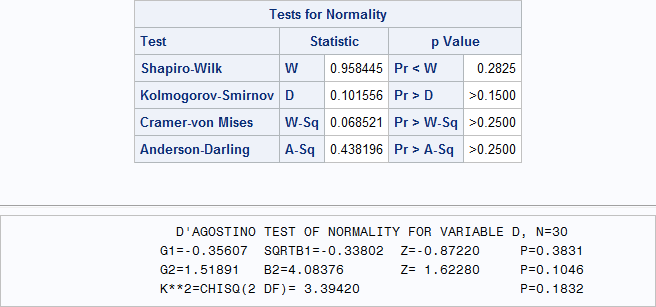
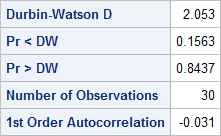


Figure 3.6 Log(Q) Model Diagnostic Tests

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.

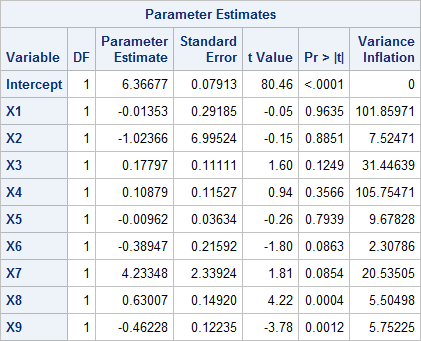


Figure 3.7 Centering the Regressors

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.

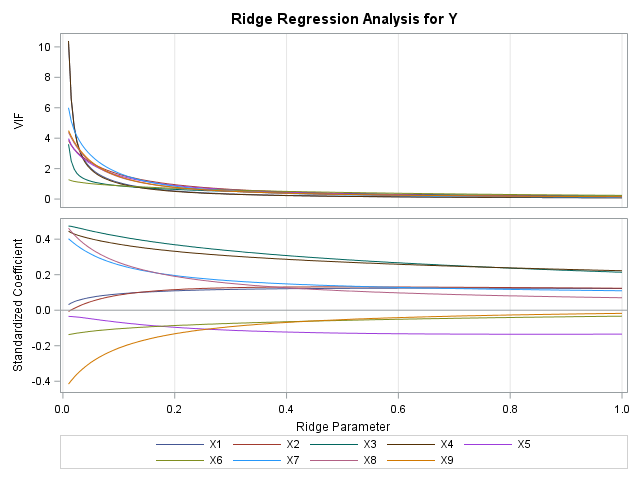


Figure 3.8 Ridge Regression Analysis

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.

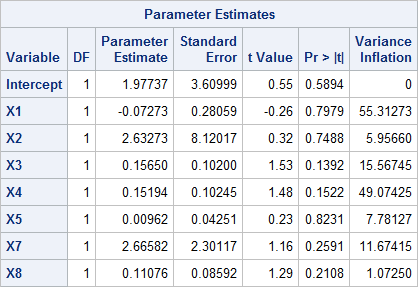


Figure 3.9 Deleted X6 and X9

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

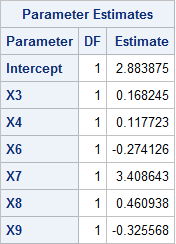
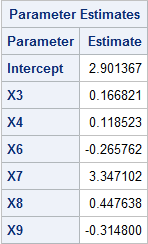
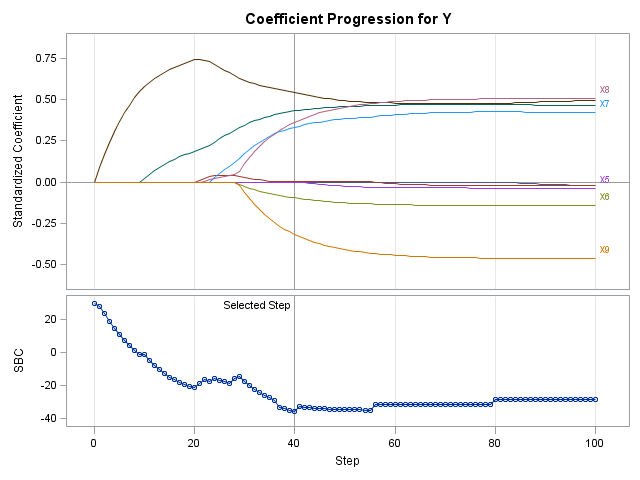
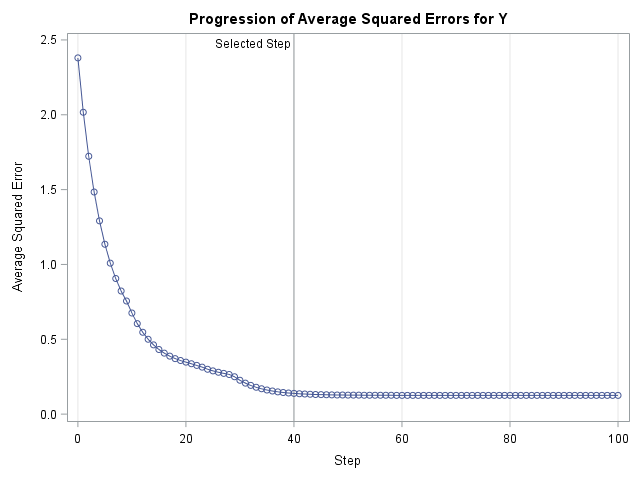


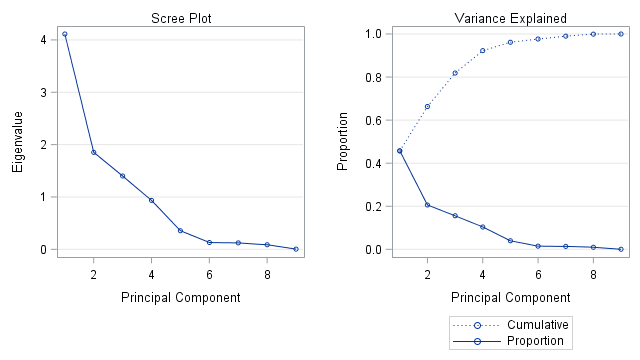
Figure 3.10 LASSO (Left) and Elastic Net (Right) Selection Summary

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)

/Users/jake/Desktop/Screen Shot 2020-05-01 at 10.00.55 PM.png

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)



/Users/jake/Desktop/Screen Shot 2020-05-01 at 10.11.02 PM.png

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

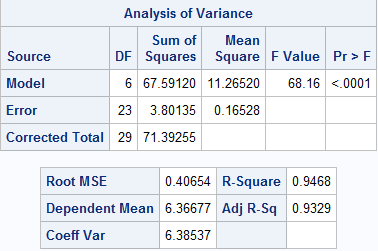


Figure 3.13 ANOVA Table for final model

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.

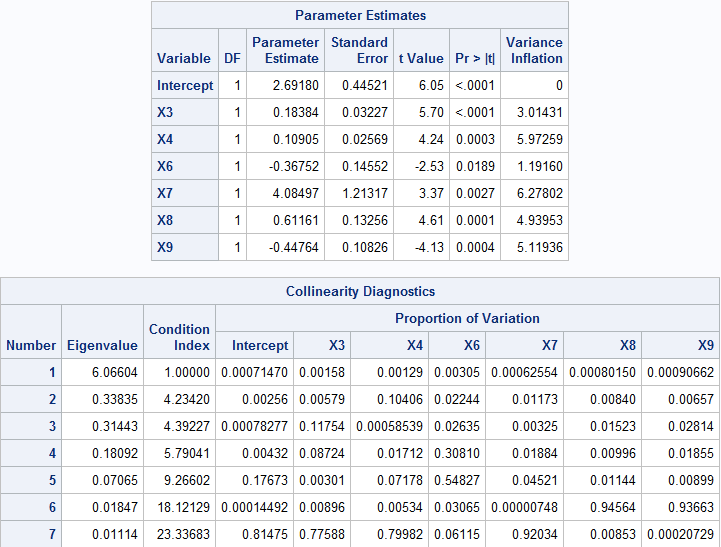


Figure 3.14 Parameter Estimates for final model

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.

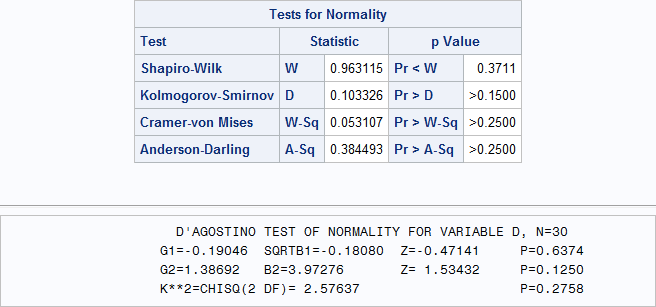
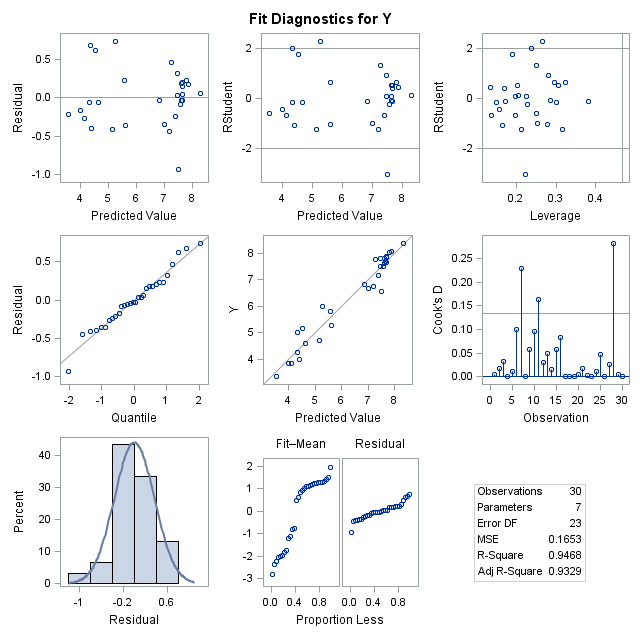


Figure 3.15 Final Model Diagnostic Tests

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)