Jake Schinto

Math 312

Homework 8

1. Exploratory Data Analysis

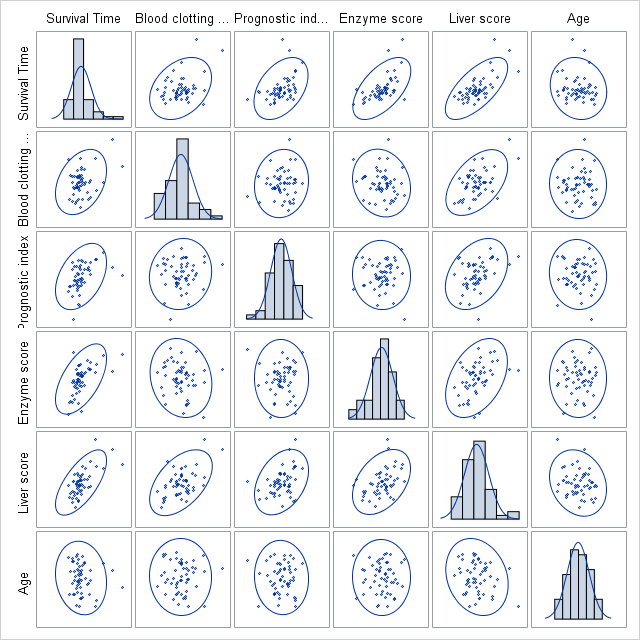


Figure 1.1 Scatter matrix for Y and X1-X5 with histogram on diagonal

By looking at the scatterplot data, it would appear that there is a possible correlation between Survival Time and Prognostic Index, Survival Time and Enzyme Score, Survival Time and Liver Score, and Blood Clotting and Liver Score. The rest seems too difficult to tell from the plot alone. Looking at the histograms, it looks like there is a right skew in Survival Time and Liver Score. On the other side, there appears to be a left skew on Enzyme Score. The Rest show no skew.

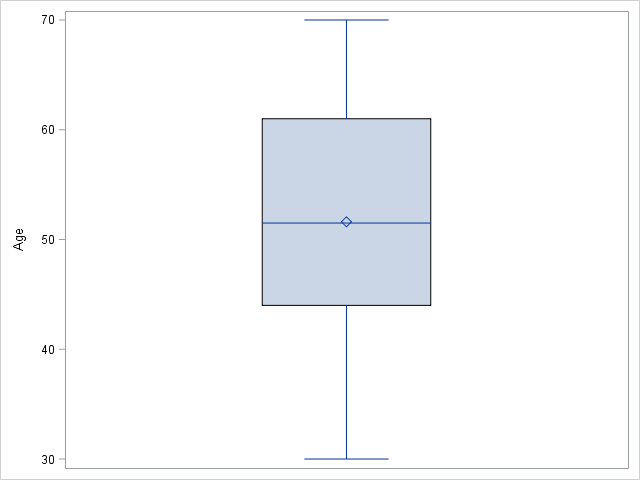
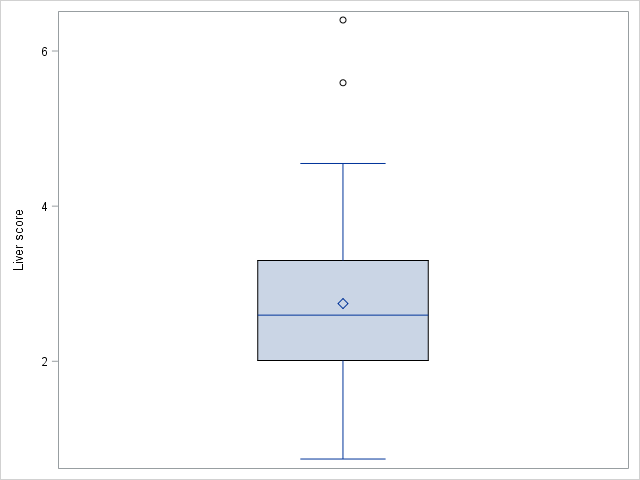
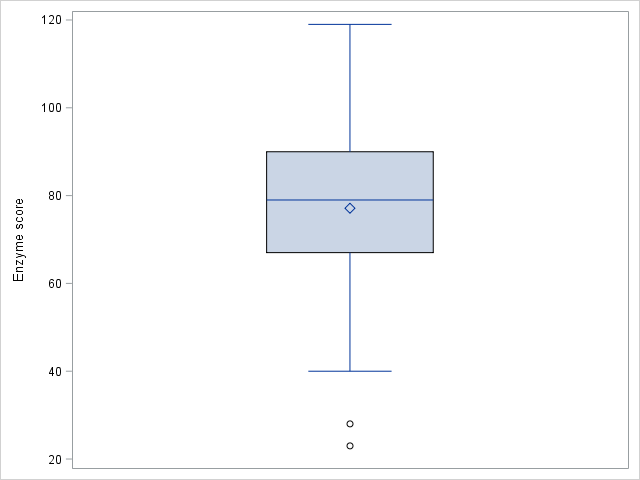
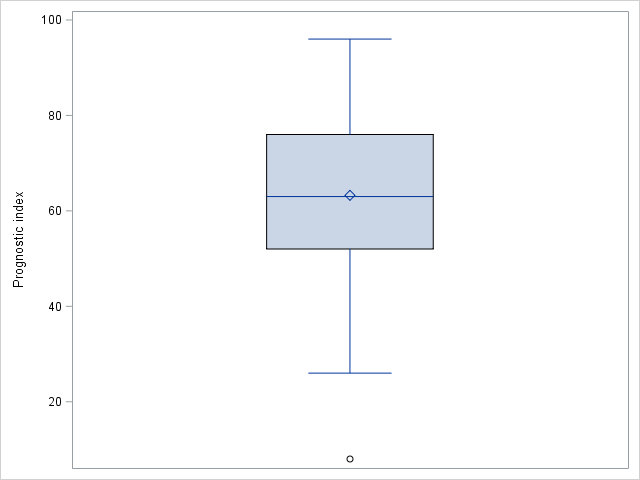
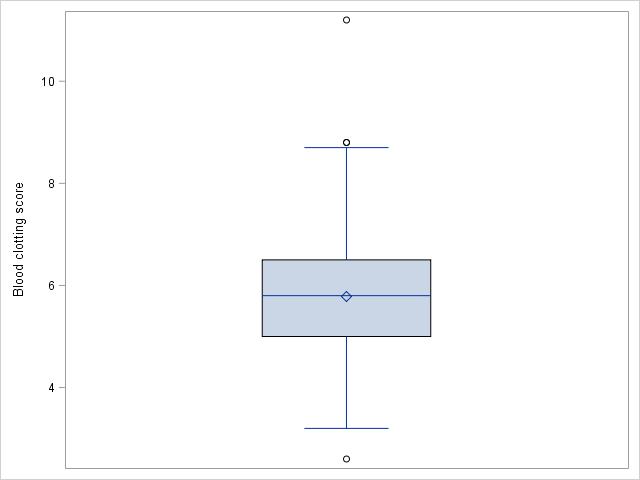
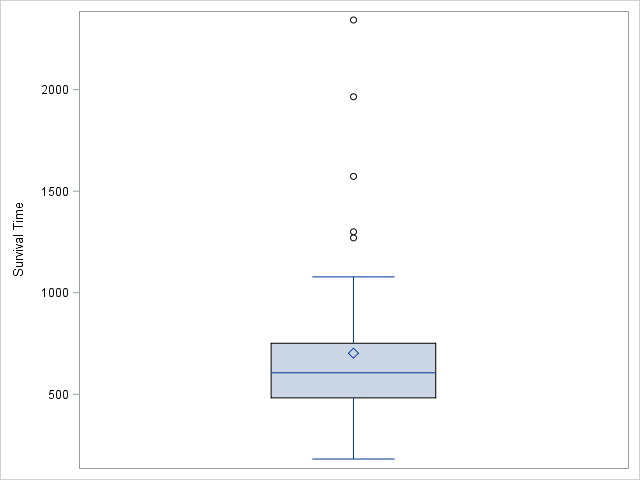


Figure 1.2 Boxplot for Y (top left), X1 (top center), X2 (top right), X3 (bottom left), X4 (bottom center), X5(bottom right)

The skewness seems consistent with what the boxplots show. Enzyme Score and Liver Score appear less Skewed in the box plots than the histogram, but the skew is still apparent.

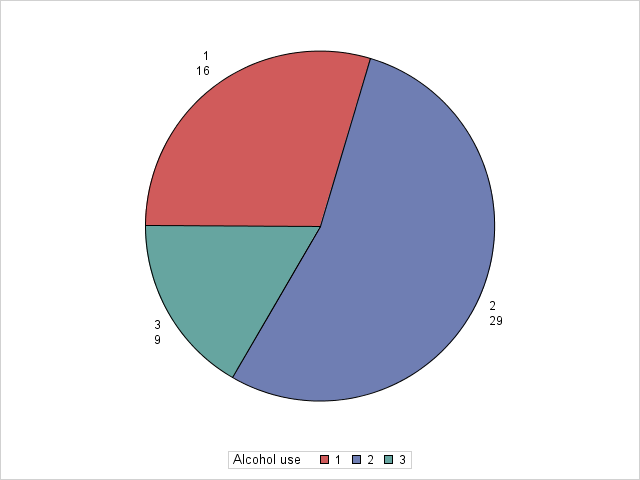
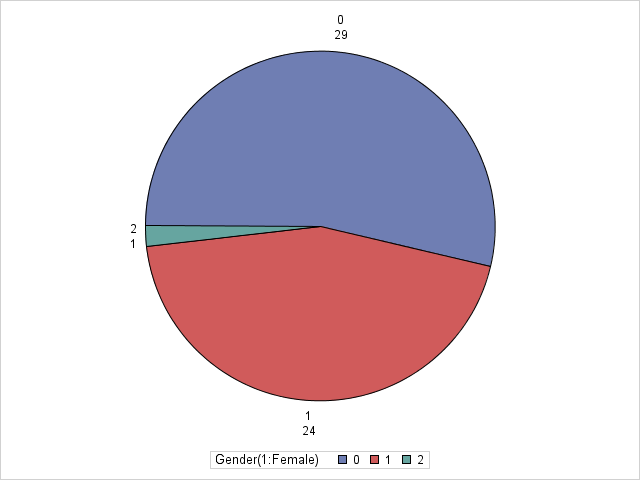


Figure 1.3 Pie Chart for Gender (left) and Alcohol use (right)

The gender “2” is most likely a mistake in the dataset, but I will continue with it present.

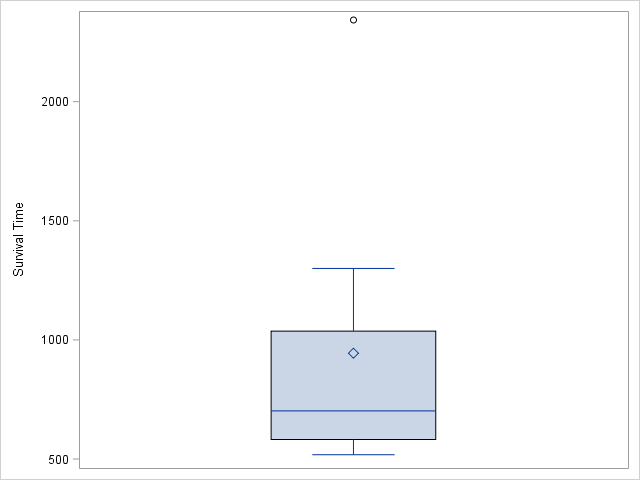
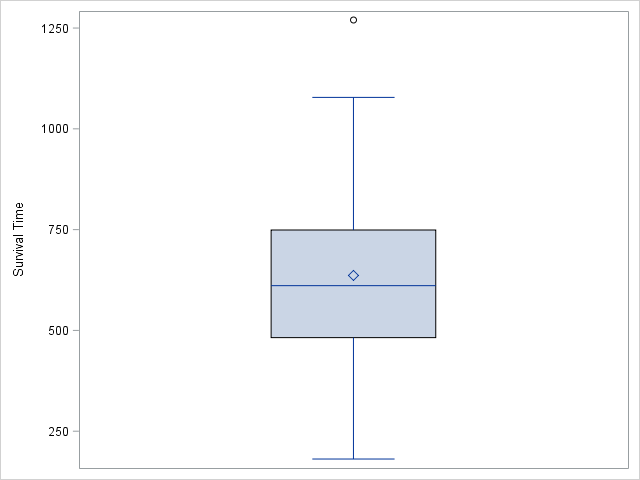
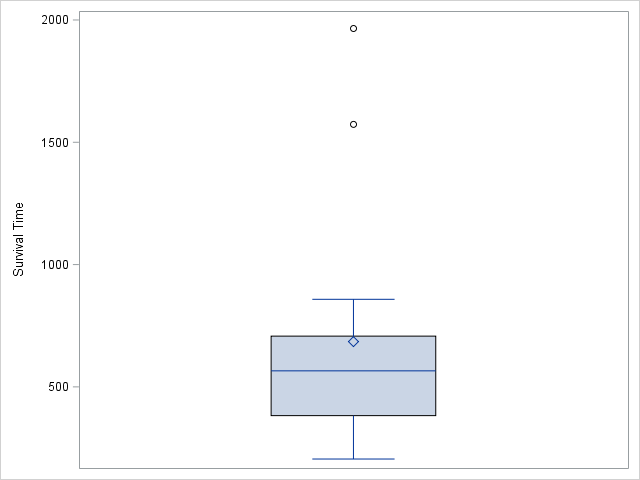
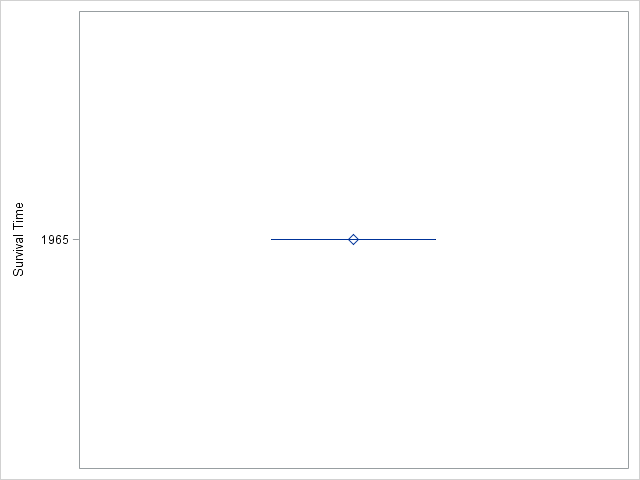
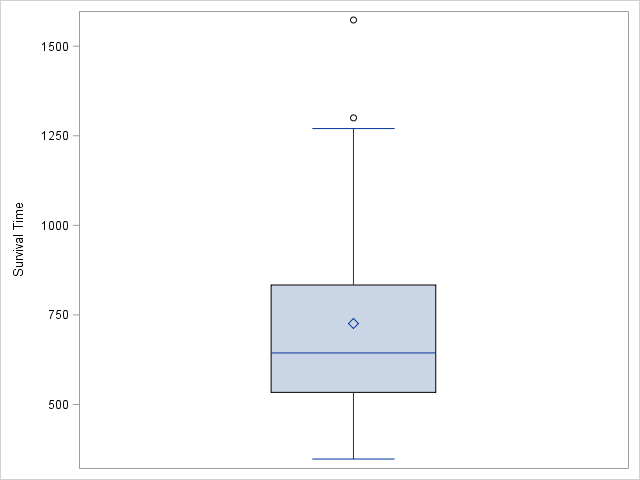
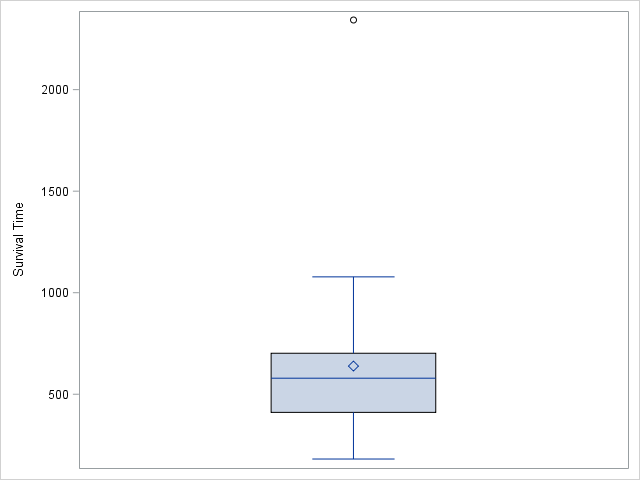


Figure 1.4 Boxplot for Y (Survival Time) across Gender = 0 (top left), Gender = 1 (top center), Gender = 2 (top right), Alcohol use = 1 (bottom left), Alcohol use = 2 (bottom center), Alcohol use = 3 (bottom right)

The means seem relatively consistent, however, the skewness in each graph does seem to increase in magnitude for certain groups. This might play a role different mean values.

2. Correlation Analysis on Numerical Features

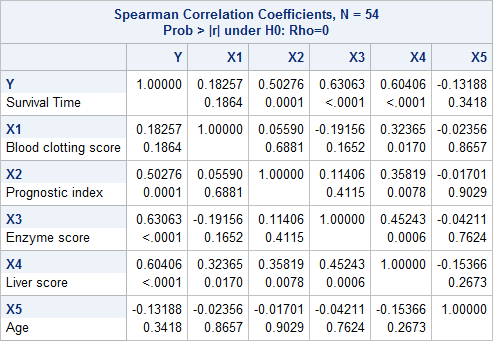


Figure 2.1 Correlation Analysis Data for Y and X1-X5

Looking at the Correlation Data reported from SAS, it shows that there are a few correlated values. The p-values confirm my predictions made earlier about the correlations with the Y variable. Survival Time and Prognostic Index, Survival Time and Enzyme Score, and Survival Time and Liver Score all show a p-value < .05. There are a couple of other correlations as well (Liver Score and X1-X3), but we do not need to focus on them going forward.

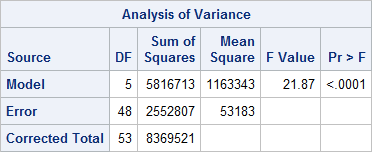
3. Frequency Analysis on Categorical Features



Figure 3.1 Correlation between Survival Time and Gender/Alcohol consumption

Looking through the results of the correlation analysis, it can easily be determined that there is in fact no correlation between gender and alcohol consumption within this dataset. The p-value is set at 0.41 which is significantly higher than the 0.05 threshold. In addition, when looking for a relationship with Survival Time, the p-values for gender and alcohol use are 0.076 and 0.136 respectively. While they are much closer to the 0.05 range, we must still reject all correlations.

4. Regression Analysis



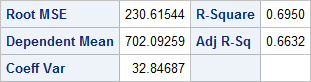


Figure 4.1 ANOVA table for data

The F-test in the table shows a significant value of <.0001 returned. This value is below .05 which suggests that the full model should be considered. The Adjusted R-Sq value is 0.6632.

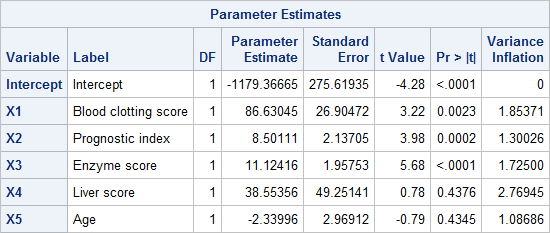


Figure 4.2 Parameter Estimates for the full model

Looking at the estimates provided and specifically the Variance Inflation value, shows a value under the threshold of 10. This allows us to proceed without having to worry about any serious multicollinearity.



Figure 4.3 Durbin-Watson test

The Pr<DW and Pr>DW values are both greater that 0.05, so there is no significant autoregressive effect.

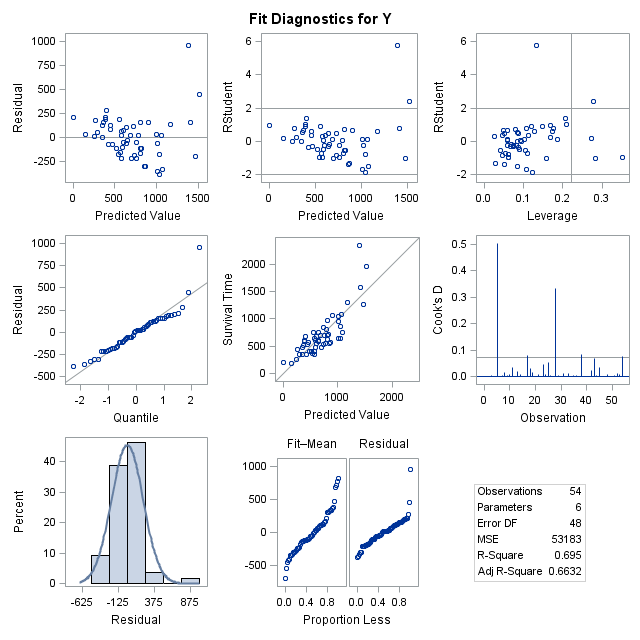


Figure 4.4 Model Diagnostics panel Survival Time on X1-X5

From the top left and top center panel, you can see that the majority of residuals are randomly distributed within the uniform band. Therefore, there are no violations against Homoscedasticity.

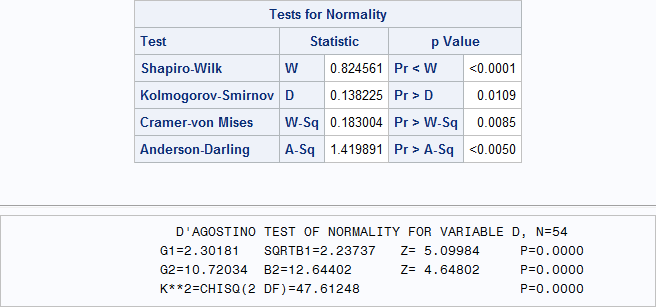


Figure 4.5 Normality Test for the full model

Looking at the normality information given here, it would appear to fail the normality test because all of the p-values are <0.05.

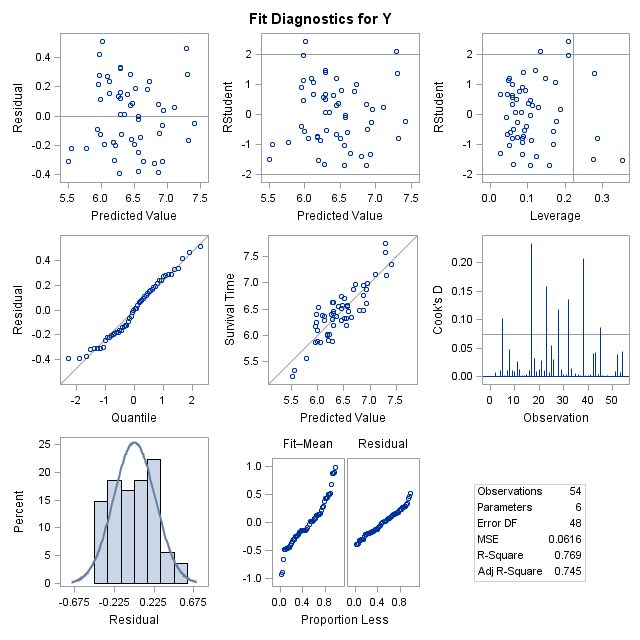
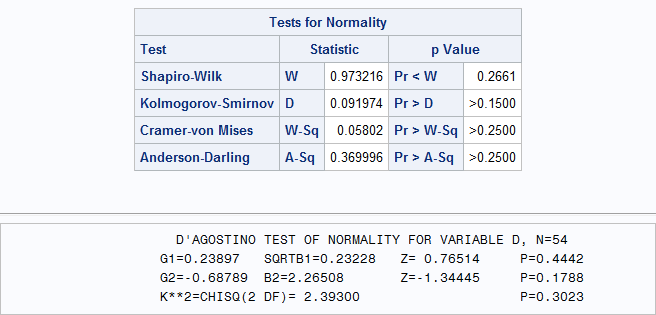
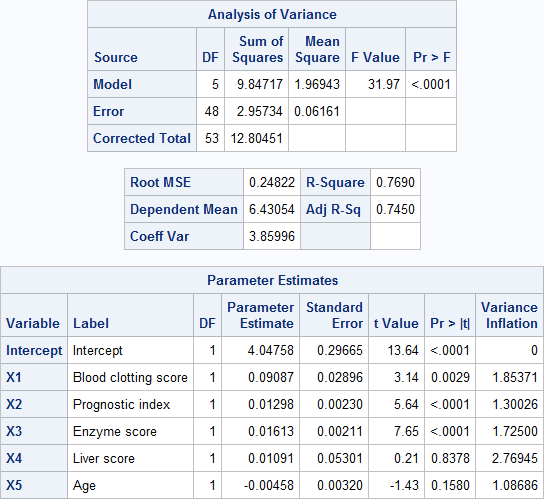
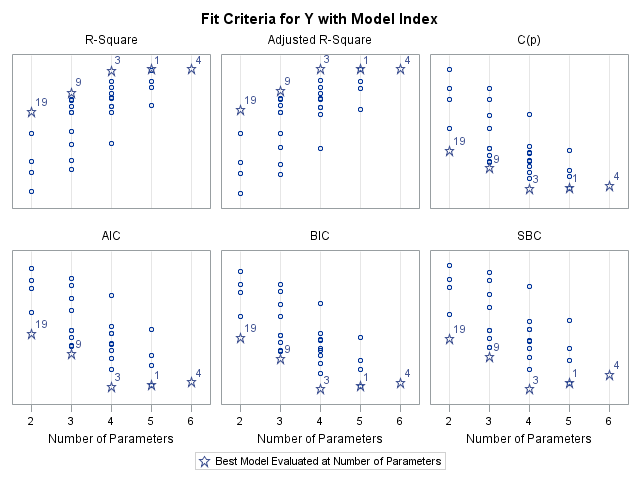
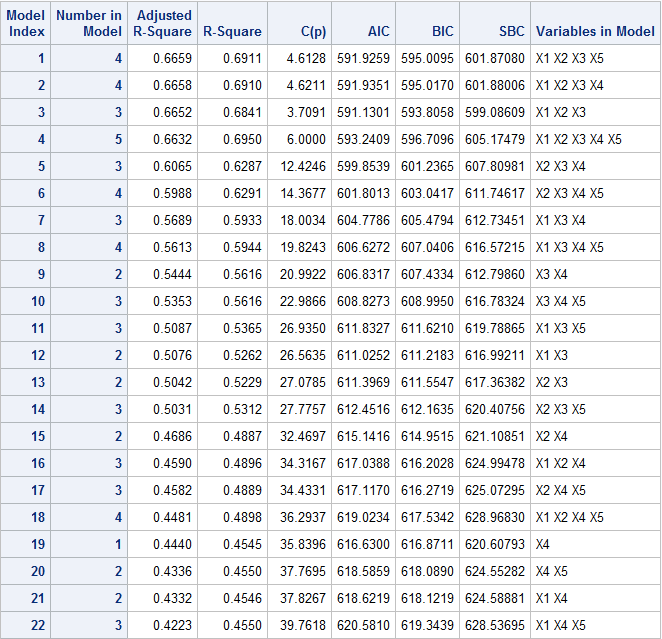


Figure 4.6 tests for log(Y)

The F-test in the table shows a significant value of <.0001 returned. This value is below .05 which suggests that the full model should be considered. The Adjusted R-Sq value is 0.745. This time all of the normality tests are >.05 and are therefore no violations against normality. All Variance inflation values are less than 10 as well, so no autocorrelation exists. Overall this model is a better fit.



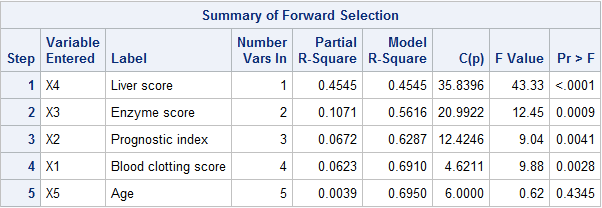
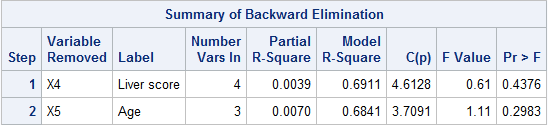
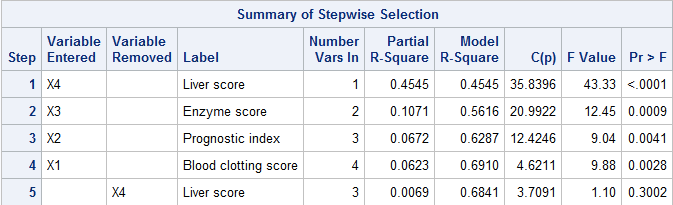
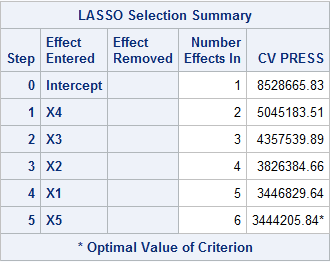
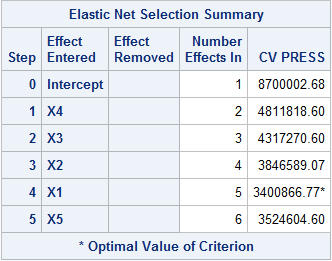
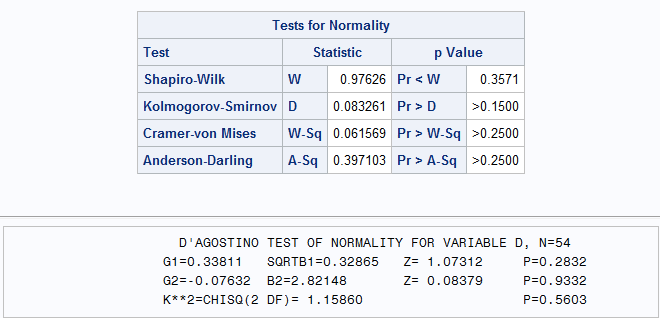
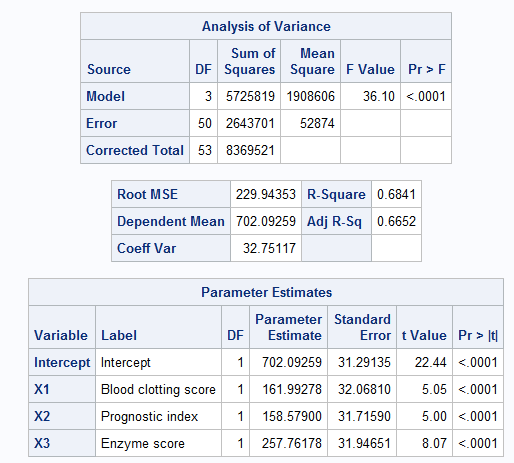


Figure 4.7 Model Selection

From the figures above, you can see the best of each criteria on the models and it would appear that the model that did the best overall in the selection process is model 4, where X4 and X5 are the variables removed. Model 3 showed the most efficiencies in the model selection graph.



The final model passes all of the tests and seems like a very good fit for the data since all p-values are low



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** SURGICAL;

  INPUT X1-X7 Y LY;

  LABEL  Y="Survival Time"

        X1="Blood clotting score"

        X2="Prognostic index"

        X3="Enzyme score"

        X4="Liver score"

        X5="Age"

        X6="Gender(1:Female)"

        X7="Alcohol use";

DATALINES;

6.7 62 81 2.59 50 0 2 695 6.544

5.1 59 66 1.70 39 0 1 403 5.999

7.4 57 83 2.16 55 0 1 710 6.565

6.5 73 41 2.01 48 0 1 349 5.854

7.8 65 115 4.30 45 0 3 2343 7.759

5.8 38 72 1.42 65 1 2 348 5.852

5.7 46 63 1.91 49 1 3 518 6.250

3.7 68 81 2.57 69 1 2 749 6.619

6.0 67 93 2.50 58 0 2 1056 6.962

3.7 76 94 2.40 48 0 2 968 6.875

6.3 84 83 4.13 37 0 2 745 6.613

6.7 51 43 1.86 57 0 2 257 5.549

5.8 96 114 3.95 63 1 1 1573 7.361

5.8 83 88 3.95 52 1 1 858 6.754

7.7 62 67 3.40 58 0 3 702 6.554

7.4 74 68 2.40 64 1 2 809 6.695

6.0 85 28 2.98 36 1 2 682 6.526

3.7 51 41 1.55 39 0 1 205 5.321

7.3 68 74 3.56 59 1 1 550 6.309

5.6 57 87 3.02 63 0 3 838 6.731

5.2 52 76 2.85 39 0 1 359 5.883

3.4 83 53 1.12 67 1 2 353 5.866

6.7 26 68 2.10 30 0 3 599 6.395

5.8 67 86 3.40 49 1 2 562 6.332

6.3 59 100 2.95 36 1 2 651 6.478

5.8 61 73 3.50 62 1 2 751 6.621

5.2 52 86 2.45 70 0 2 545 6.302

11.2 76 90 5.59 58 2 1 1965 7.583

5.2 54 56 2.71 44 1 1 477 6.167

5.8 76 59 2.58 61 1 2 600 6.396

3.2 64 65 0.74 53 0 2 443 6.094

8.7 45 23 2.52 68 0 2 181 5.198

5.0 59 73 3.50 57 0 2 411 6.019

5.8 72 93 3.30 39 1 3 1037 6.944

5.4 58 70 2.64 31 1 2 482 6.179

5.3 51 99 2.60 48 0 2 634 6.453

2.6 74 86 2.05 45 0 1 678 6.519

4.3 8 119 2.85 65 1 1 362 5.893

4.8 61 76 2.45 51 1 2 637 6.457

5.4 52 88 1.81 40 1 1 705 6.558

5.2 49 72 1.84 46 0 1 536 6.283

3.6 28 99 1.30 55 0 3 582 6.366

8.8 86 88 6.40 30 1 2 1270 7.147

6.5 56 77 2.85 41 0 2 538 6.288

3.4 77 93 1.48 69 0 2 482 6.178

6.5 40 84 3.00 54 1 2 611 6.416

4.5 73 106 3.05 47 1 2 960 6.867

4.8 86 101 4.10 35 1 3 1300 7.170

5.1 67 77 2.86 66 1 1 581 6.365

3.9 82 103 4.55 50 0 2 1078 6.983

6.6 77 46 1.95 50 0 2 405 6.005

6.4 85 40 1.21 58 0 3 579 6.361

6.4 59 85 2.33 63 0 2 550 6.310

8.8 78 72 3.20 56 0 1 651 6.478

;

**RUN**;

**PROC** **SGSCATTER** DATA = SURGICAL;

  MATRIX Y X1-X5

                / ellipse

                  diagonal = (histogram normal);

**RUN**;

/\*Boxplots\*/

**PROC** **SGPLOT** DATA=SURGICAL;

  VBOX Y;

**RUN**;

**PROC** **SGPLOT** DATA=SURGICAL;

  VBOX X1;

**RUN**;

**PROC** **SGPLOT** DATA=SURGICAL;

  VBOX X2;

**RUN**;

**PROC** **SGPLOT** DATA=SURGICAL;

  VBOX X3;

**RUN**;

**PROC** **SGPLOT** DATA=SURGICAL;

  VBOX X4;

**RUN**;

**PROC** **SGPLOT** DATA=SURGICAL;

  VBOX X5;

**RUN**;

/\*Pie Charts\*/

**PROC** **TEMPLATE**;

   DEFINE STATGRAPH pie;

      BEGINGRAPH;

         LAYOUT REGION;

            PIECHART CATEGORY = X6 /

            DATALABELLOCATION = OUTSIDE

            CATEGORYDIRECTION = CLOCKWISE

            START = **180** NAME = 'pie';

            DISCRETELEGEND 'pie' /

            TITLE = 'Gender(1:Female)';

         ENDLAYOUT;

      ENDGRAPH;

   END;

**RUN**;

**PROC** **TEMPLATE**;

   DEFINE STATGRAPH pie2;

      BEGINGRAPH;

         LAYOUT REGION;

            PIECHART CATEGORY = X7 /

            DATALABELLOCATION = OUTSIDE

            CATEGORYDIRECTION = CLOCKWISE

            START = **180** NAME = 'pie';

            DISCRETELEGEND 'pie' /

            TITLE = 'Alcohol use';

         ENDLAYOUT;

      ENDGRAPH;

   END;

**RUN**;

**PROC** **SGRENDER** DATA = SURGICAL

            TEMPLATE = pie;

**RUN**;

**PROC** **SGRENDER** DATA = SURGICAL

            TEMPLATE = pie2;

**RUN**;

/\*Boxplot split by gender and alcohol use\*/

**data** gender0;

    set SURGICAL;

     if X6=**0**;

**run**;

**data** gender1;

    set SURGICAL;

     if X6=**1**;

**run**;

**data** gender2;

    set SURGICAL;

     if X6=**2**;

**run**;

**data** alcohol1;

    set SURGICAL;

     if X7=**1**;

**run**;

**data** alcohol2;

    set SURGICAL;

     if X7=**2**;

**run**;

**data** alcohol3;

    set SURGICAL;

     if X7=**3**;

**run**;

**PROC** **SGPLOT** DATA=gender0;

  VBOX Y;

**RUN**;

**PROC** **SGPLOT** DATA=gender1;

  VBOX Y;

**RUN**;

**PROC** **SGPLOT** DATA=gender2;

  VBOX Y;

**RUN**;

**PROC** **SGPLOT** DATA=alcohol1;

  VBOX Y;

**RUN**;

**PROC** **SGPLOT** DATA=alcohol2;

  VBOX Y;

**RUN**;

**PROC** **SGPLOT** DATA=alcohol3;

  VBOX Y;

**RUN**;

/\*Correlation Analysis\*/

**PROC** **CORR** DATA=SURGICAL SPEARMAN FISHER(BIASADJ=NO);

  VAR Y X1-X5;

**RUN**;

**PROC** **CORR** DATA=SURGICAL SPEARMAN FISHER(BIASADJ=NO);

  VAR Y X6 X7;

**RUN**;

/\*Full Model Fit w/ model doagnostics\*/

**PROC** **REG** DATA=SURGICAL;

  MODEL Y = X1-X5/DWPROB VIF COLLIN;

  OUTPUT OUT=SFM\_FIT RSTUDENT=D;

**RUN**;

**QUIT**;

%***NORMTEST***(D,SFM\_FIT)

/\*Full Model Fit w/ model doagnostics log\*/

**data** SURGICAL2;

    set SURGICAL;

     Y = log(Y);

**run**;

**PROC** **REG** DATA=SURGICAL2;

  MODEL Y = X1-X5/DWPROB VIF COLLIN;

  OUTPUT OUT=SFM\_FIT RSTUDENT=D;

**RUN**;

**QUIT**;

%***NORMTEST***(D,SFM\_FIT)

/\*Model Selection\*/

**PROC** **REG** DATA=SURGICAL PLOTS(LABEL)=CRITERIA;

  MODEL Y = X1-X5/SELECTION=ADJRSQ CP AIC BIC SBC;

**RUN**;

**QUIT**;

**PROC** **REG** DATA=SURGICAL PLOTS(LABEL)=CRITERIA;

  MODEL Y = X1-X5/SELECTION=FORWARD;

**RUN**;

**QUIT**;

**PROC** **REG** DATA=SURGICAL PLOTS(LABEL)=CRITERIA;

  MODEL Y = X1-X5/SELECTION=BACKWARD;

**RUN**;

**QUIT**;

**PROC** **REG** DATA=SURGICAL PLOTS(LABEL)=CRITERIA;

  MODEL Y = X1-X5/SELECTION=STEPWISE;

**RUN**;

**QUIT**;

**PROC** **GLMSELECT** DATA=SURGICAL PLOTS=ALL;

   MODEL Y = X1-X5/SELECTION=LASSO(CHOOSE=CV STOP=NONE) CVMETHOD=RANDOM(**10**);

**RUN**;

**PROC** **GLMSELECT** DATA=SURGICAL PLOTS=ALL;

   MODEL Y=X1-X5/SELECTION=ELASTICNET(CHOOSE=CV STOP=NONE) CVMETHOD=RANDOM(**10**);

**RUN**;

**PROC** **REG** DATA=SURGICAL PLOTS(LABEL)=(COOKSD DIAGNOSTICS RESIDUALS(SMOOTH));

  MODEL Y = X1 X2 X3/DWPROB INFLUENCE;

  OUTPUT OUT=SRM\_FIT RSTUDENT=D;

**RUN**;

**QUIT**;

%***NORMTEST***(D,SRM\_FIT)

/\*Standardize X1 X3 X6 to compare the impact\*/

**PROC** **STDIZE** DATA=SURGICAL OUT=STDSURGICAL;

  VAR X1 X2 X3;

**RUN**;

**PROC** **REG** DATA=STDSURGICAL OUTEST=SRM\_EST PLOTS=NONE;

  MODEL Y = X1 X2 X3;

**RUN**;

**QUIT**;

ODS RTF CLOSE;