Jake Schinto
Math 312

Homework 8

1. Exploratory Data Analysis

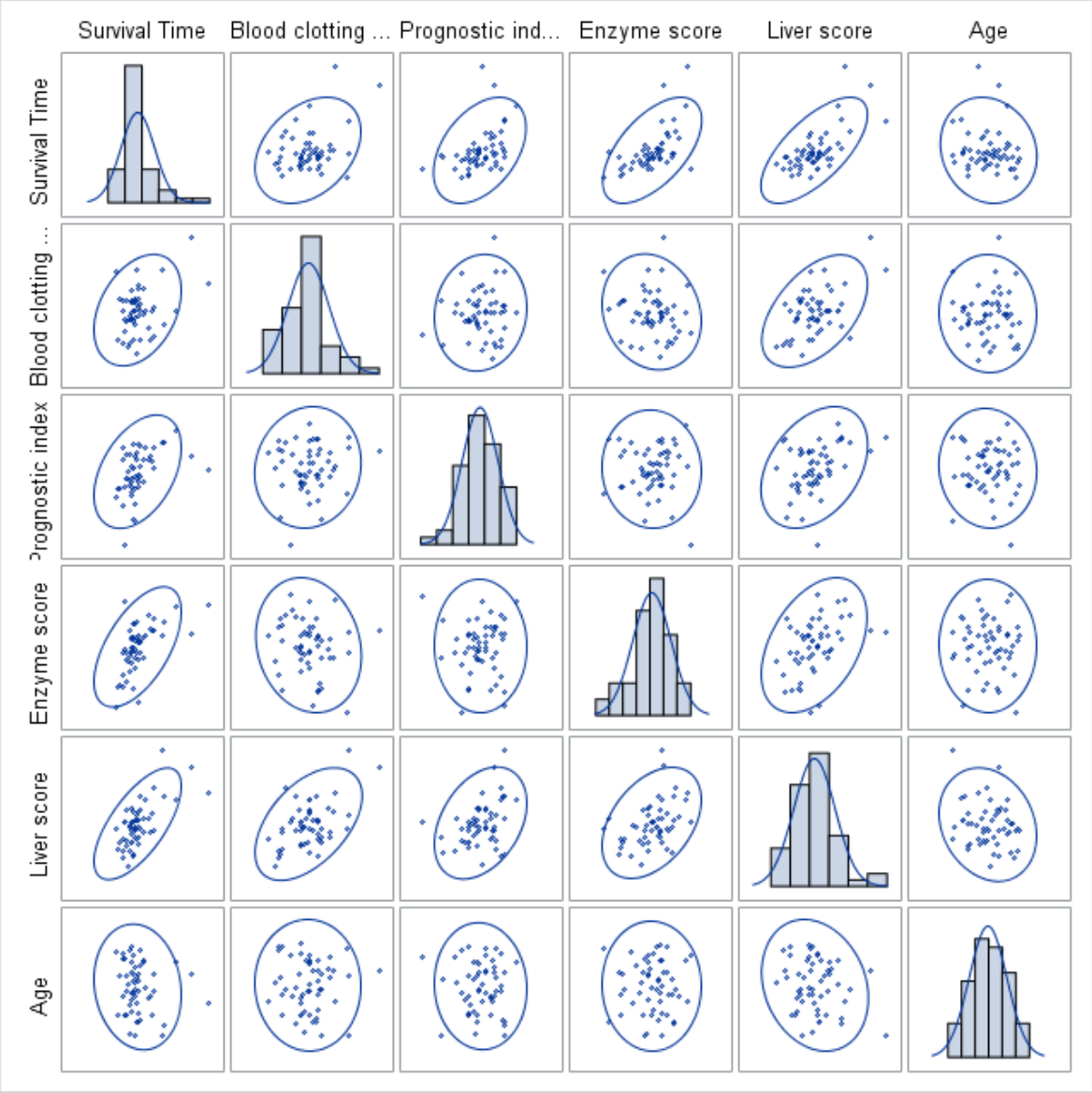


Figure 1.1 Scatter matrix for Y and X1-X5 with histogram on diagonal

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.

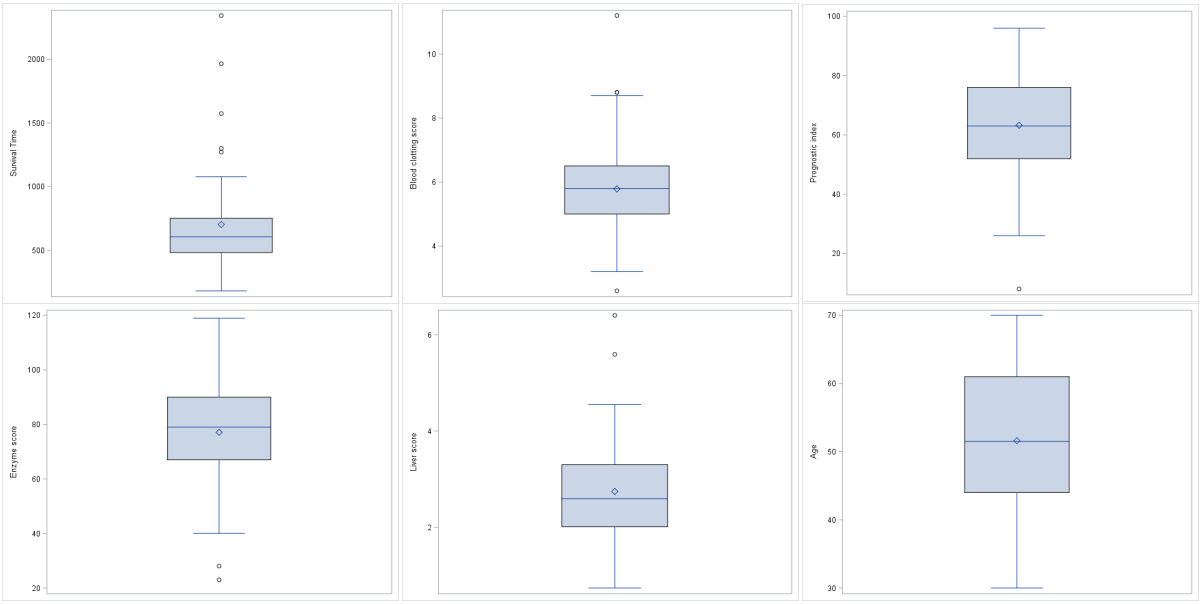


Figure 1.2 Boxplot for Y (top left), X1 (top center), X2 (top right), X3 (bottom left), X4 (bottom center), X5(bottom right)

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.

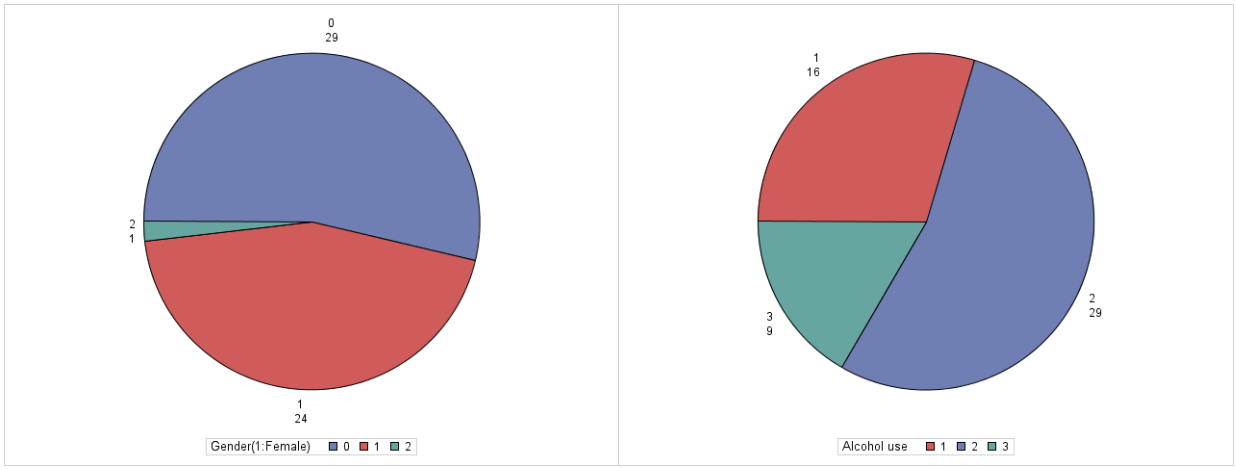


Figure 1.3 Pie Chart for Gender (left) and Alcohol use (right)

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

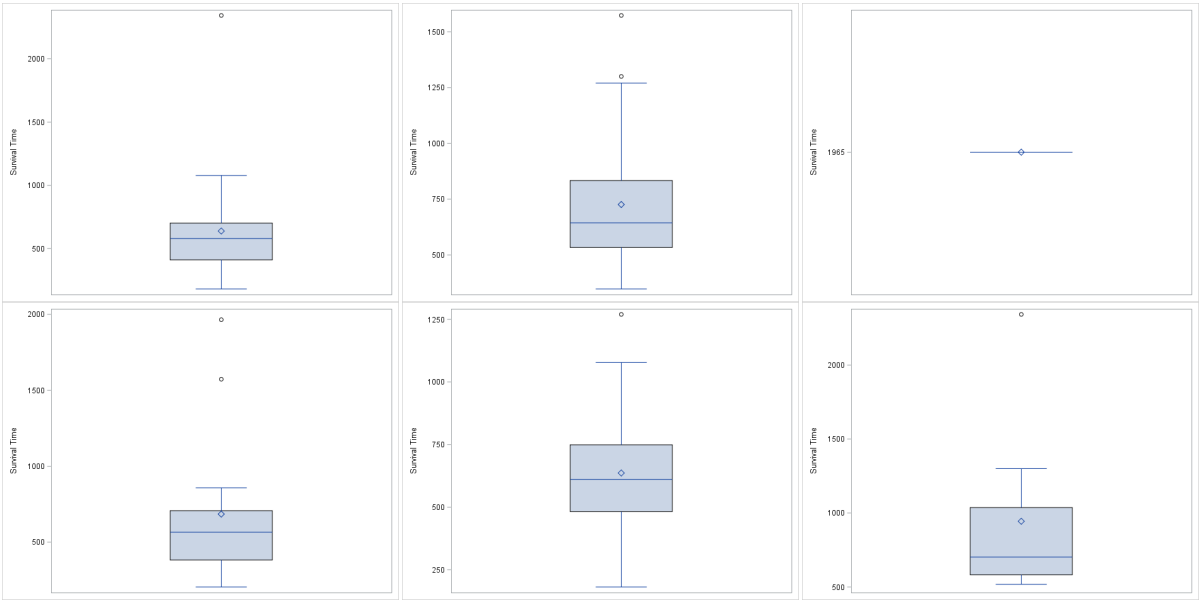


Figure 1.4 Boxplot for Y (Survival Time) across Gender = 0 (top left), Gender = 1 (top center), Gender = 2 (top right), Alcohol use = 1 (bottom left), Alcohol use = 2 (bottom center), Alcohol use = 3 (bottom right)

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

Spearman Correlation Coefficients, N = 54 Prob > |r| under H0: Rho=0						
	Y	**X1**	**X2**	**X3**	**X4**	**X5**
Y Survival Time	1.00000	0.18257 0.1864	0.50276 0.0001	0.63063 <.0001	0.60406 <.0001	-0.13188 0.3418
X1 Blood clotting score	0.18257 0.1864	1.00000	0.05590 0.6881	-0.19156 0.1652	0.32365 0.0170	-0.02356 0.8657
X2 Prognostic index	0.50276 0.0001	0.05590 0.6881	1.00000	0.11406 0.4115	0.35819 0.0078	-0.01701 0.9029
X3 Enzyme score	0.63063 <.0001	-0.19156 0.1652	0.11406 0.4115	1.00000	0.45243 0.0006	-0.04211 0.7624
X4 Liver score	0.60406 <.0001	0.32365 0.0170	0.35819 0.0078	0.45243 0.0006	1.00000	-0.15366 0.2673
X5 Age	-0.13188 0.3418	-0.02356 0.8657	-0.01701 0.9029	-0.04211 0.7624	-0.15366 0.2673	1.00000

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

3 Variables:	Y X6 X7

Simple Statistics

Variable	N	Mean	Std Dev	Median	Minimum	Maximum	Label
Y	54	702.09259	397.38581	605.50000	181.00000	2343	Survival Time
X6	54	0.48148	0.54047	0	0	2.00000	Gender(1:Female)
X7	54	1.87037	0.67422	2.00000	1.00000	3.00000	Alcohol use

Spearman Correlation Coefficients, N = 54
Prob > |r| under H0: Rho=0

	Y	X6	X7
Y Survival Time	1.00000	0.24379 0.0757	0.20537 0.1363
X6 Gender(1:Female)	0.24379 0.0757	1.00000	-0.11433 0.4104
X7 Alcohol use	0.20537 0.1363	-0.11433 0.4104	1.00000

Spearman Correlation Statistics (Fisher's z Transformation)

Variable	With Variable	N	Sample Correlation	Fisher's z	95% Confidence Limits		p Value for H0:Rho=0
Y	X6	54	0.24379	0.24880	-0.025646	0.480203	0.0756
Y	X7	54	0.20537	0.20833	-0.066022	0.448469	0.1368
X6	X7	54	-0.11433	-0.11483	-0.370741	0.158276	0.4122

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

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	5816713	1163343	21.87	<.0001
Error	48	2552807	53183		
Corrected Total	53	8369521			

Root MSE	230.61544	R-Square	0.6950
Dependent Mean	702.09259	Adj R-Sq	0.6632
Coeff Var	32.84687		

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.

Parameter Estimates							
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	-1179.36665	275.61935	-4.28	<.0001	0
X1	Blood clotting score	1	86.63045	26.90472	3.22	0.0023	1.85371
X2	Prognostic index	1	8.50111	2.13705	3.98	0.0002	1.30026
X3	Enzyme score	1	11.12416	1.95753	5.68	<.0001	1.72500
X4	Liver score	1	38.55356	49.25141	0.78	0.4376	2.76945
X5	Age	1	-2.33996	2.96912	-0.79	0.4345	1.08686

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.

Durbin-Watson D	1.932
Pr < DW	0.4004
Pr > DW	0.5996
Number of Observations	54
1st Order Autocorrelation	0.002

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.

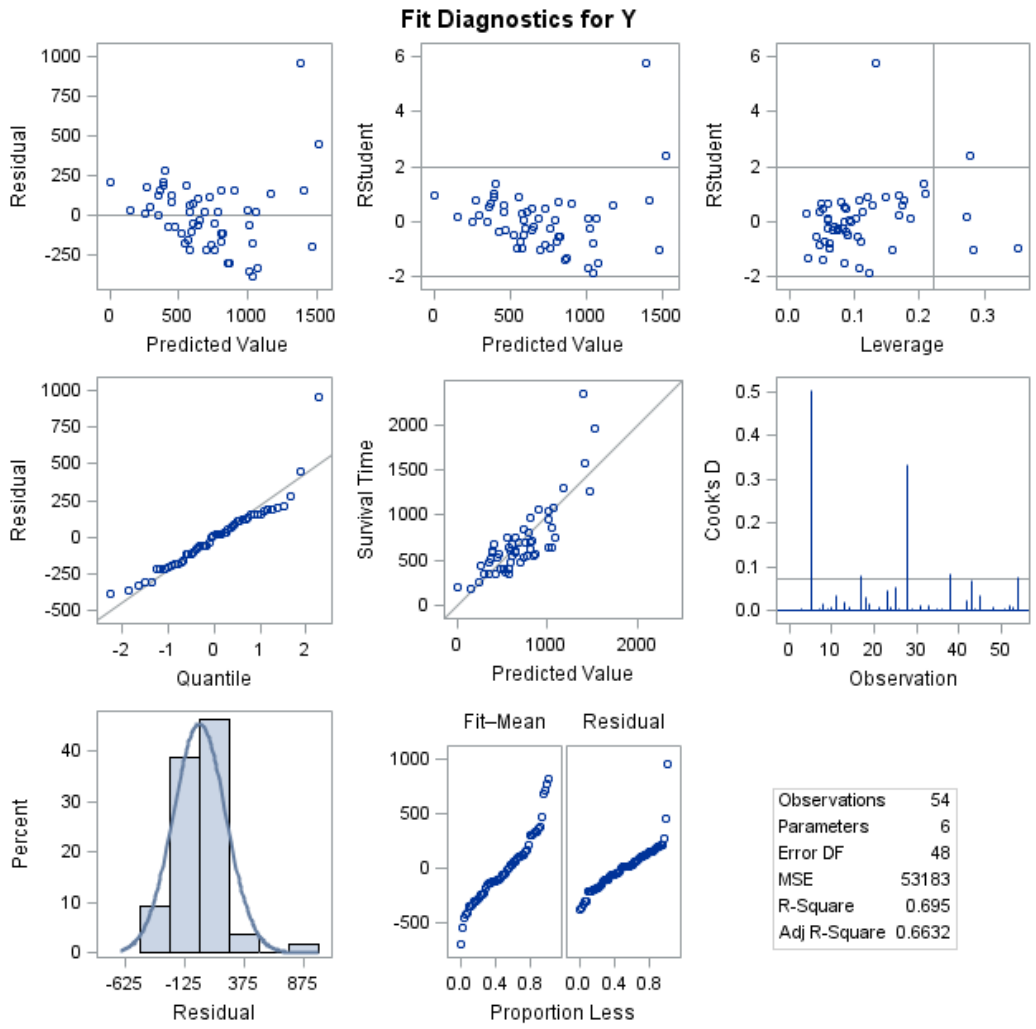


Figure 4.4 Model Diagnostics panel Survival Time on X1-X5

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.

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.824561	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.138225	Pr > D	0.0109
Cramer-von Mises	W-Sq	0.183004	Pr > W-Sq	0.0085
Anderson-Darling	A-Sq	1.419891	Pr > A-Sq	<0.0050

```
  D'AGOSTINO TEST OF NORMALITY FOR VARIABLE D, N=54
G1=2.30181   SQRTB1=2.23737   Z= 5.09984     P=0.0000
G2=10.72034  B2=12.64402      Z= 4.64802     P=0.0000
K**2=CHISQ(2 DF)=47.61248                    P=0.0000
```

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.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	9.84717	1.96943	31.97	<.0001
Error	48	2.95734	0.06161		
Corrected Total	53	12.80451			

Root MSE	0.24822	R-Square	0.7690
Dependent Mean	6.43054	Adj R-Sq	0.7450
Coeff Var	3.85996		

Parameter Estimates							
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	4.04758	0.29665	13.64	<.0001	0
X1	Blood clotting score	1	0.09087	0.02896	3.14	0.0029	1.85371
X2	Prognostic index	1	0.01298	0.00230	5.64	<.0001	1.30026
X3	Enzyme score	1	0.01613	0.00211	7.65	<.0001	1.72500
X4	Liver score	1	0.01091	0.05301	0.21	0.8378	2.76945
X5	Age	1	-0.00458	0.00320	-1.43	0.1580	1.08686

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.973216	Pr < W	0.2661
Kolmogorov-Smirnov	D	0.091974	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.05802	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.369996	Pr > A-Sq	>0.2500

```
D'AGOSTINO TEST OF NORMALITY FOR VARIABLE D, N=54
G1=0.23897   SQRTB1=0.23228   Z= 0.76514    P=0.4442
G2=-0.68789  B2=2.26508       Z=-1.34445    P=0.1788
K**2=CHISQ(2 DF)= 2.39300                  P=0.3023
```

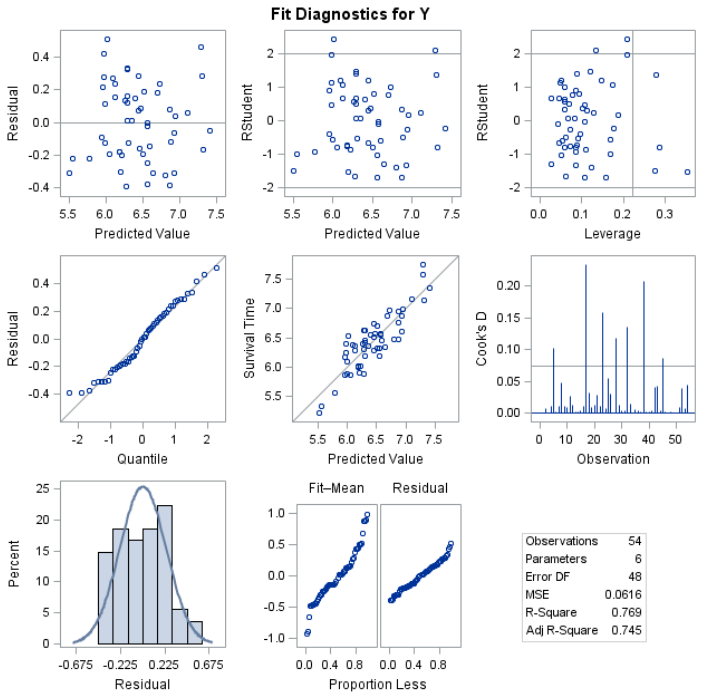
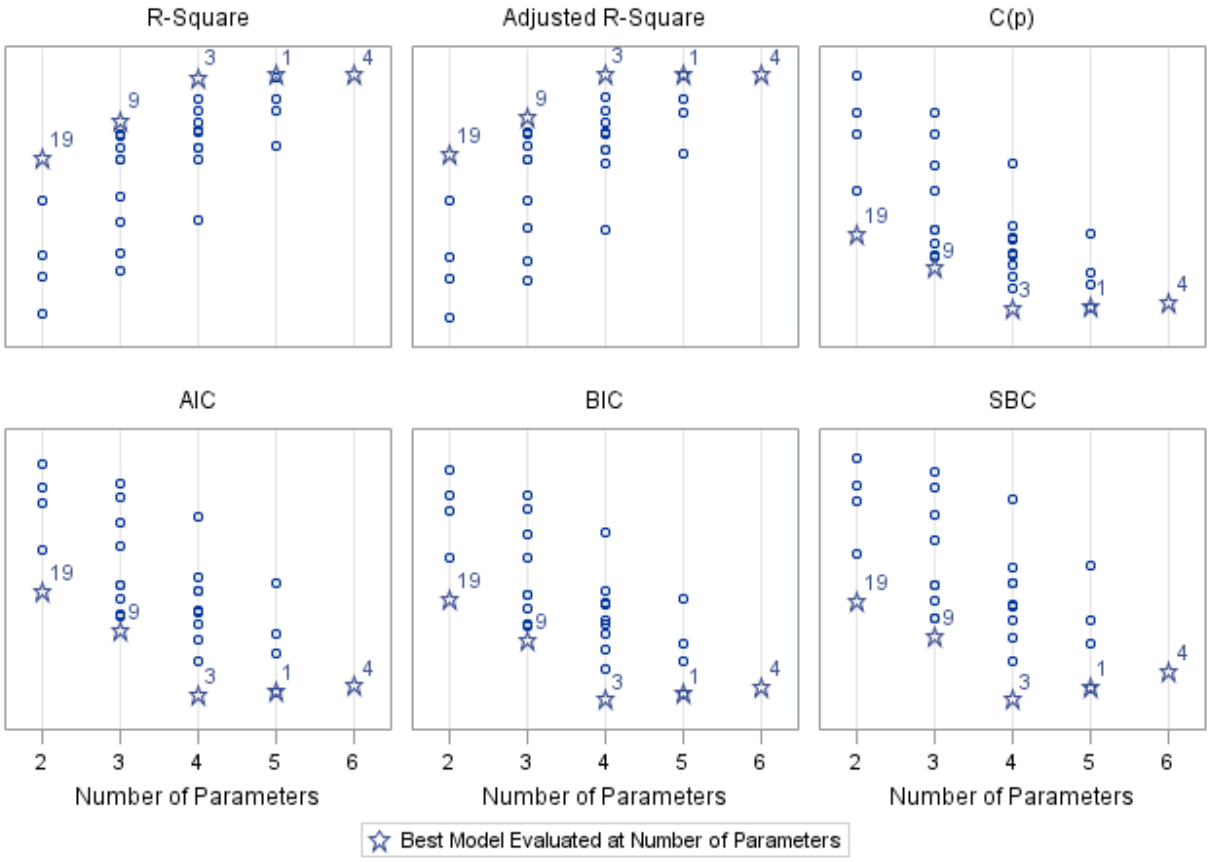


Figure 4.6 tests for log(Y)

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.

Model Index	Number in Model	Adjusted R-Square	R-Square	C(p)	AIC	BIC	SBC	Variables in Model
1	4	0.6659	0.6911	4.6128	591.9259	595.0095	601.87080	X1 X2 X3 X5
2	4	0.6658	0.6910	4.6211	591.9351	595.0170	601.88006	X1 X2 X3 X4
3	3	0.6652	0.6841	3.7091	591.1301	593.8058	599.08609	X1 X2 X3
4	5	0.6632	0.6950	6.0000	593.2409	596.7096	605.17479	X1 X2 X3 X4 X5
5	3	0.6065	0.6287	12.4246	599.8539	601.2365	607.80981	X2 X3 X4
6	4	0.5988	0.6291	14.3677	601.8013	603.0417	611.74617	X2 X3 X4 X5
7	3	0.5689	0.5933	18.0034	604.7786	605.4794	612.73451	X1 X3 X4
8	4	0.5613	0.5944	19.8243	606.6272	607.0406	616.57215	X1 X3 X4 X5
9	2	0.5444	0.5616	20.9922	606.8317	607.4334	612.79860	X3 X4
10	3	0.5353	0.5616	22.9866	608.8273	608.9950	616.78324	X3 X4 X5
11	3	0.5087	0.5365	26.9350	611.8327	611.6210	619.78865	X1 X3 X5
12	2	0.5076	0.5262	26.5635	611.0252	611.2183	616.99211	X1 X3
13	2	0.5042	0.5229	27.0785	611.3969	611.5547	617.36382	X2 X3
14	3	0.5031	0.5312	27.7757	612.4516	612.1635	620.40756	X2 X3 X5
15	2	0.4686	0.4887	32.4697	615.1416	614.9515	621.10851	X2 X4
16	3	0.4590	0.4896	34.3167	617.0388	616.2028	624.99478	X1 X2 X4
17	3	0.4582	0.4889	34.4331	617.1170	616.2719	625.07295	X2 X4 X5
18	4	0.4481	0.4898	36.2937	619.0234	617.5342	628.96830	X1 X2 X4 X5
19	1	0.4440	0.4545	35.8396	616.6300	616.8711	620.60793	X4
20	2	0.4336	0.4550	37.7695	618.5859	618.0890	624.55282	X4 X5
21	2	0.4332	0.4546	37.8267	618.6219	618.1219	624.58881	X1 X4
22	3	0.4223	0.4550	39.7618	620.5810	619.3439	628.53695	X1 X4 X5


Fit Criteria for Y with Model Index
R-Square
Adjusted R-Square
C(p)
AIC
BIC
SBC
Number of Parameters
Number of Parameters
Number of Parameters
Best Model Evaluated at Number of Parameters

Elastic Net Selection Summary				
Step	Effect Entered	Effect Removed	Number Effects In	CV PRESS
0	Intercept		1	8700002.68
1	X4		2	4811818.60
2	X3		3	4317270.60
3	X2		4	3846589.07
4	X1		5	3400866.77*
5	X5		6	3524604.60
* Optimal Value of Criterion				

LASSO Selection Summary				
Step	Effect Entered	Effect Removed	Number Effects In	CV PRESS
0	Intercept		1	8528665.83
1	X4		2	5045183.51
2	X3		3	4357539.89
3	X2		4	3826384.66
4	X1		5	3446829.64
5	X5		6	3444205.84*
* Optimal Value of Criterion				

Summary of Stepwise Selection									
Step	Variable Entered	Variable Removed	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	X4		Liver score	1	0.4545	0.4545	35.8396	43.33	<.0001
2	X3		Enzyme score	2	0.1071	0.5616	20.9922	12.45	0.0009
3	X2		Prognostic index	3	0.0672	0.6287	12.4246	9.04	0.0041
4	X1		Blood clotting score	4	0.0623	0.6910	4.6211	9.88	0.0028
5		X4	Liver score	3	0.0069	0.6841	3.7091	1.10	0.3002

Summary of Backward Elimination								
Step	Variable Removed	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	X4	Liver score	4	0.0039	0.6911	4.6128	0.61	0.4376
2	X5	Age	3	0.0070	0.6841	3.7091	1.11	0.2983

Summary of Forward Selection								
Step	Variable Entered	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	X4	Liver score	1	0.4545	0.4545	35.8396	43.33	<.0001
2	X3	Enzyme score	2	0.1071	0.5616	20.9922	12.45	0.0009
3	X2	Prognostic index	3	0.0672	0.6287	12.4246	9.04	0.0041
4	X1	Blood clotting score	4	0.0623	0.6910	4.6211	9.88	0.0028
5	X5	Age	5	0.0039	0.6950	6.0000	0.62	0.4345

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.

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.97626	Pr < W	0.3571
Kolmogorov-Smirnov	D	0.083261	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.061569	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.397103	Pr > A-Sq	>0.2500

```
  D'AGOSTINO TEST OF NORMALITY FOR VARIABLE D, N=54
G1=0.33811   SQRTB1=0.32865   Z= 1.07312     P=0.2832
G2=-0.07632  B2=2.82148       Z= 0.08379     P=0.9332
K**2=CHISQ(2 DF)= 1.15860                    P=0.5603
```

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	9.69600	3.23200	51.99	<.0001
Error	50	3.10851	0.06217		
Corrected Total	53	12.80451			

Root MSE	0.24934	R-Square	0.7572
Dependent Mean	6.43054	Adj R-Sq	0.7427
Coeff Var	3.87743		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	3.76644	0.22676	16.61	<.0001
X1	Blood clotting score	1	0.09547	0.02169	4.40	<.0001
X2	Prognostic index	1	0.01334	0.00203	6.56	<.0001
X3	Enzyme score	1	0.01644	0.00163	10.09	<.0001

Model: MODEL1
Dependent Variable: Y Survival Time

Durbin-Watson D	2.129
Pr < DW	0.6906
Pr > DW	0.3094

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

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	5725819	1908606	36.10	<.0001
Error	50	2643701	52874		
Corrected Total	53	8369521			

Root MSE	229.94353	R-Square	0.6841
Dependent Mean	702.09259	Adj R-Sq	0.6652
Coeff Var	32.75117		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	702.09259	31.29135	22.44	<.0001
X1	Blood clotting score	1	161.99278	32.06810	5.05	<.0001
X2	Prognostic index	1	158.57900	31.71590	5.00	<.0001
X3	Enzyme score	1	257.76178	31.94651	8.07	<.0001

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;
```