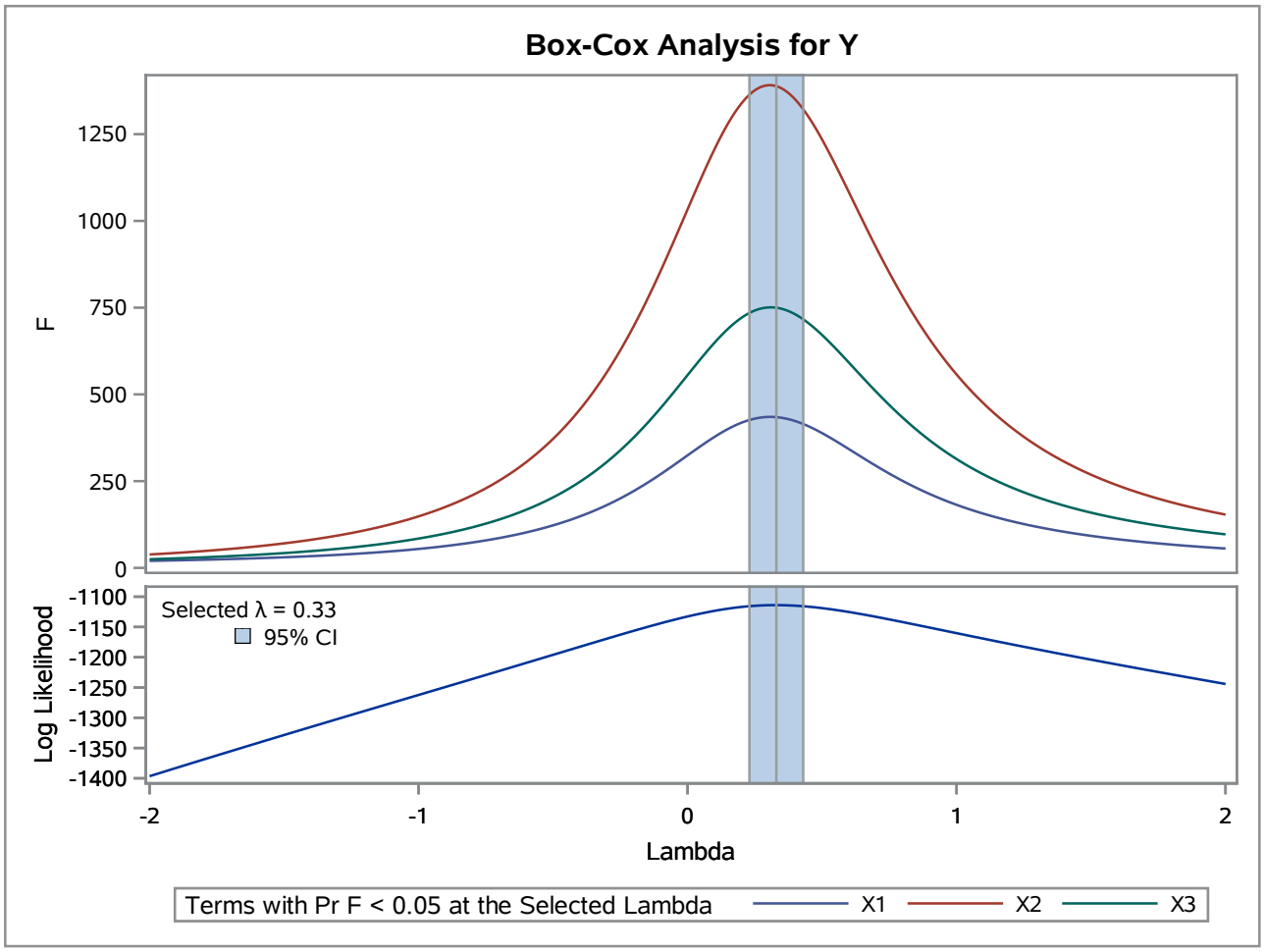
Math 5364
Data Mining 2
Homework 29
Mary Barker

```
*1. Import the file math5305Lab6Data.txt, whose columns are the
        variables Y, X_1, X_2, X_3. In Homework 27, we saw that the
        model

                Y = beta_0 + beta_1 X_1 + beta_2 X_2 + beta_3 X_3 + e

        does not satisfy the assumption e_i ~ N(0, sigma^2),i=1,...,n
        To remedy this, preform a Box-Cox transformation of Y by defining

                tildeY_i = ((Y_i)^lambda - 1) / lambda for i = 1, ... , n;

                options obs=100;

                data problem1;
                    infile '/folders/myshortcuts/sas_folder/math5305Lab6Data.txt' dlm=',
                    input Y X1 X2 X3;
                run;

        *a. Fit the model
                tildeY = beta_0 + beta_1 X_1 + beta_2 X_2 + beta_3 X_3 + e

                and let hat_tildeY and tilde_e be the predicted values and
                residuals for this transformed model;

                proc transreg data=problem1 detail(NOKNOTS NOCOEFFICIENTS);
                        model boxcox(Y/lambda = -2 to 2 by .01) = Identity(X1 X2 X3);
                        output out=myoutput;
                run;
                proc reg data=myoutput;
                        var TY TX1 TX2 TX3;
                        model TY=TX1 TX2 TX3;
                        output out = transformed_model_output
                                predicted=hat_tildeY
                                residual=tilde_e;
                run;

        *b. Plot tildeY vs hat_tildeY and tilde_e vs. hat_tildeY. Does
                curvature appear to exist in the transformed model? ;

                proc plot data=transformed_model_output;
```

```
                plot Y * hat_tildeY
                         tilde_e * hat_tildeY;
        run;

    *c. Investigate normality of the errors for the transformed model;

        proc univariate data=transformed_model_output normal;
                var tilde_e;
                qqplot tilde_e;
        run;


    *d. Investigate constancy of error variance for the transformed
            model;

        proc reg data=transformed_model_output;
            model TY=TX1 TX2 TX3/SPEC;
        run;


    *e. Do the errors for the transformed model appear to satisfy the
            assumptions of normality and constant error variance? How do
            your results compare to those of Homework 23? ;



*2. The file math5305Lab7Data.txt contains data contains data for the
        variables Y, X_1, X_2, ... , X_40. Perform a stepwise regression
        on this data set using SAS. (Hints: It may be helpful to use the
        "import data' option in the "file" menu to import this data. Also,
        make sure to specify in your glemselect procedure which variables
        are class variables. Finally, it may be convenient to use R to
        generate the model statement for this procedure.);


        options obs=2000;
        data problem2;
            infile '/folders/myshortcuts/sas_folder/math5305Lab7Data.txt' dlm=',
            input Y X1 X2 X3 X4 X5 X6 X7 X8 X9 X10 X11 X12 X13 X14 X15 X16
                           X17 X18 X19 X20 X21 X22 X23 X24 X25 X26 $ X27 X28 X2
                           X31 X32 X33 $ X34 X35 X36 X37 X38 X39 X40 $;
        run;
        proc glmselect data=problem2;
                class X26 X33 X40;
```

```
        model Y = X1-X40 / selection=stepwise ;
        output out = stepwise_results;
run;
```

The TRANSREG Procedure



Model Statement Specification Details				
Type	**DF**	**Variable**	**Description**	**Value**
Dep	1	BoxCox(Y)	Lambda Used	0.33
			Lambda	0.33
			Log Likelihood	-1113.9
			Conv. Lambda	
			Conv. Lambda LL	
			CI Limit	-1115.8
			Alpha	0.05
Ind	1	Identity(X1)	DF	1
Ind	1	Identity(X2)	DF	1
Ind	1	Identity(X3)	DF	1

The REG Procedure
Model: MODEL1
Dependent Variable: TY Y Transformation

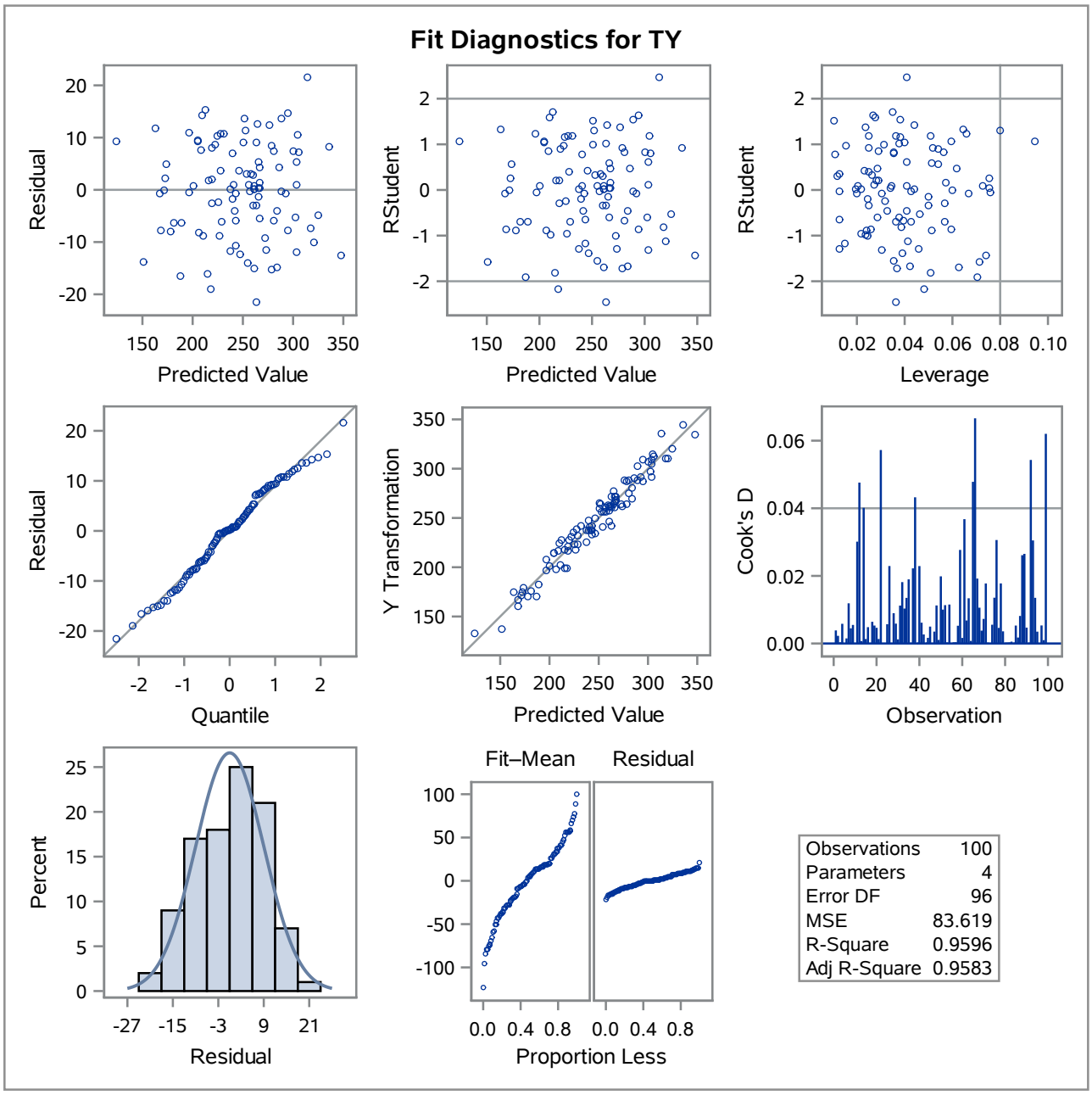
Number of Observations Read	100
Number of Observations Used	100

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	3	190475	63492	759.29	<.0001
Error	96	8027.46635	83.61944		
Corrected Total	99	198502			

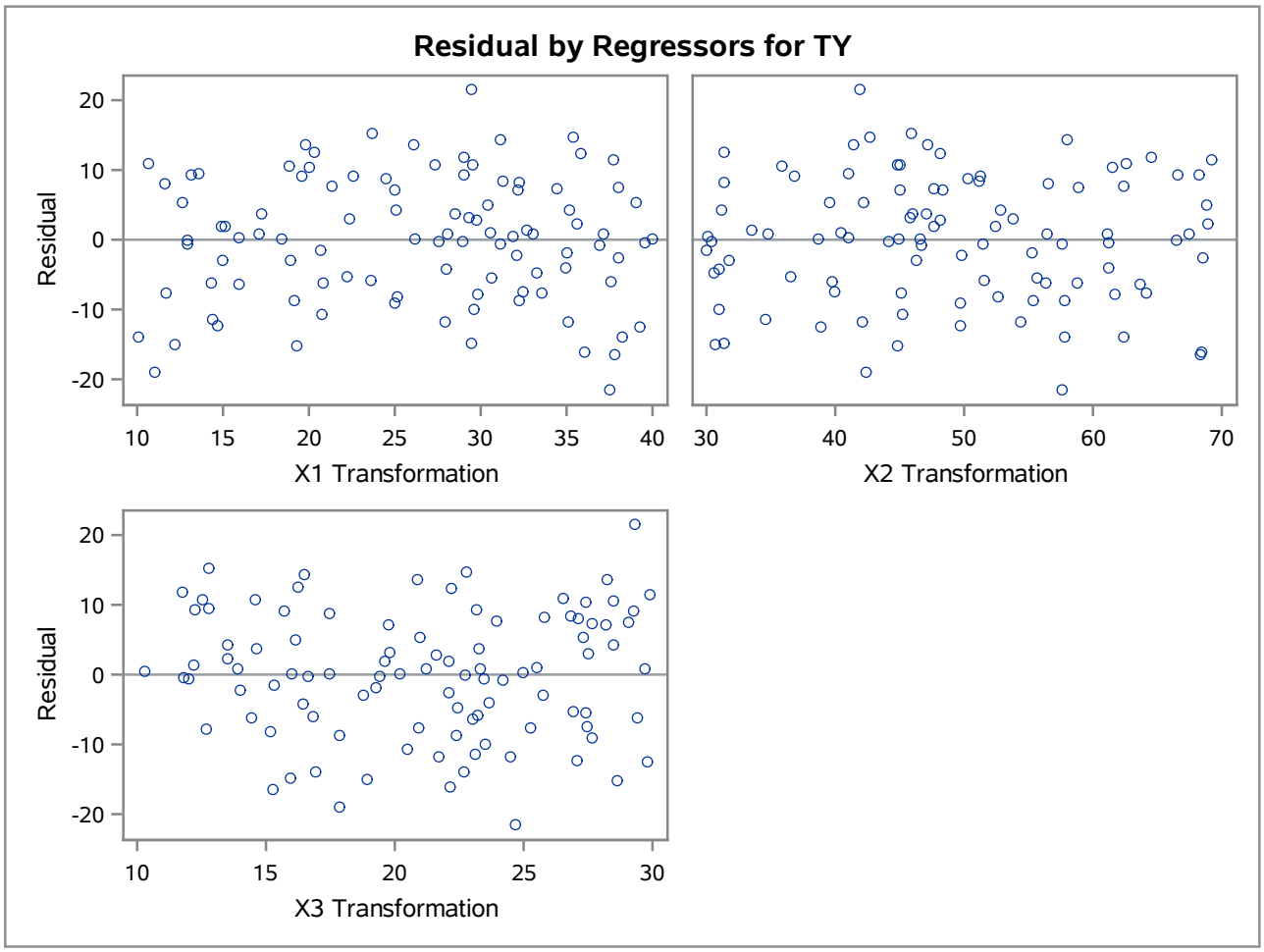
Root MSE	9.14437	**R-Square**	0.9596
Dependent Mean	247.15703	**Adj R-Sq**	0.9583
Coeff Var	3.69982		

Parameter Estimates						
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	Intercept	1	239.19831	5.84935	40.89	<.0001
TX1	X1 Transformation	1	2.23916	0.10743	20.84	<.0001
TX2	X2 Transformation	1	-3.02597	0.08121	-37.26	<.0001
TX3	X3 Transformation	1	4.62033	0.16873	27.38	<.0001

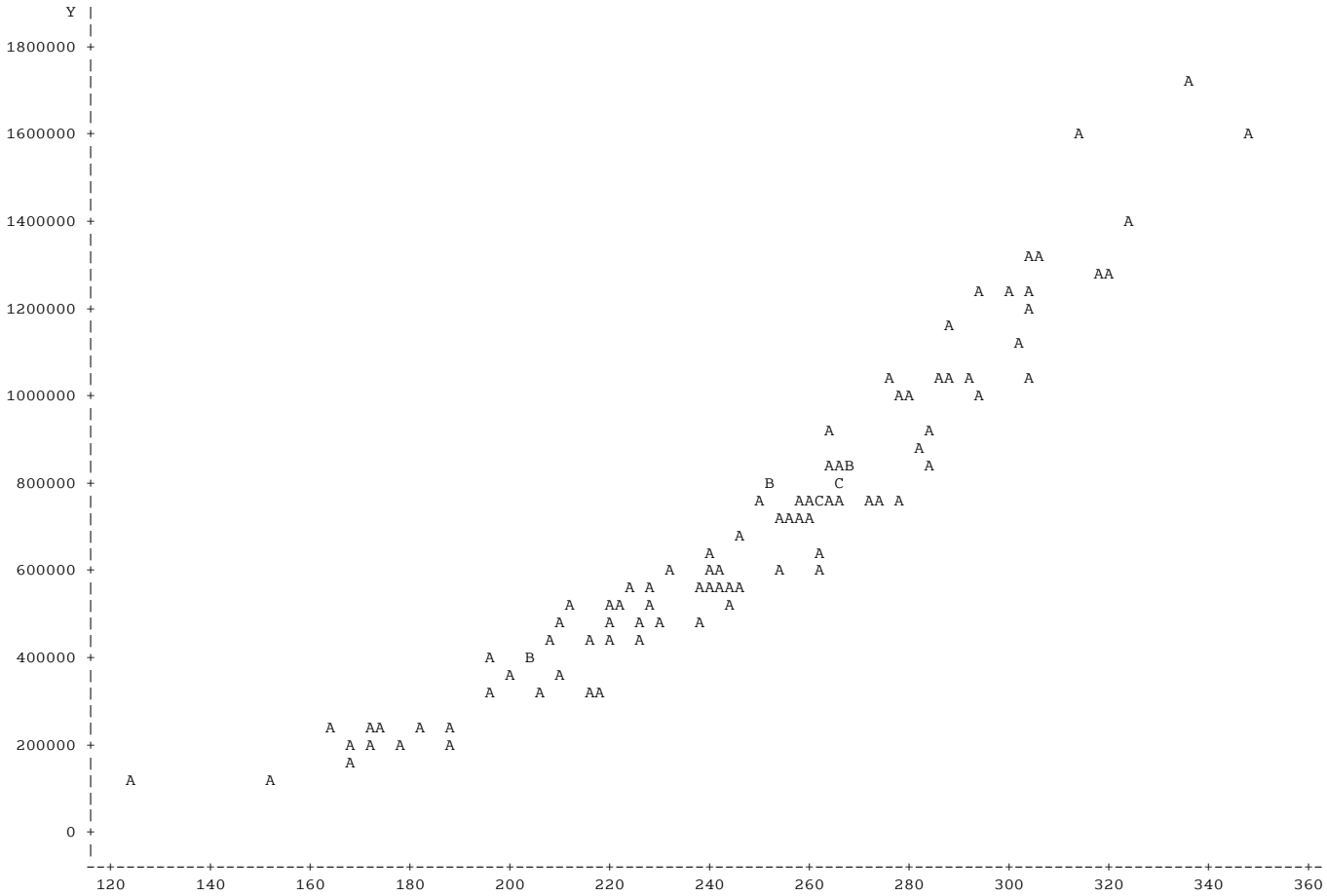
The REG Procedure
Model: MODEL1
Dependent Variable: TY Y Transformation



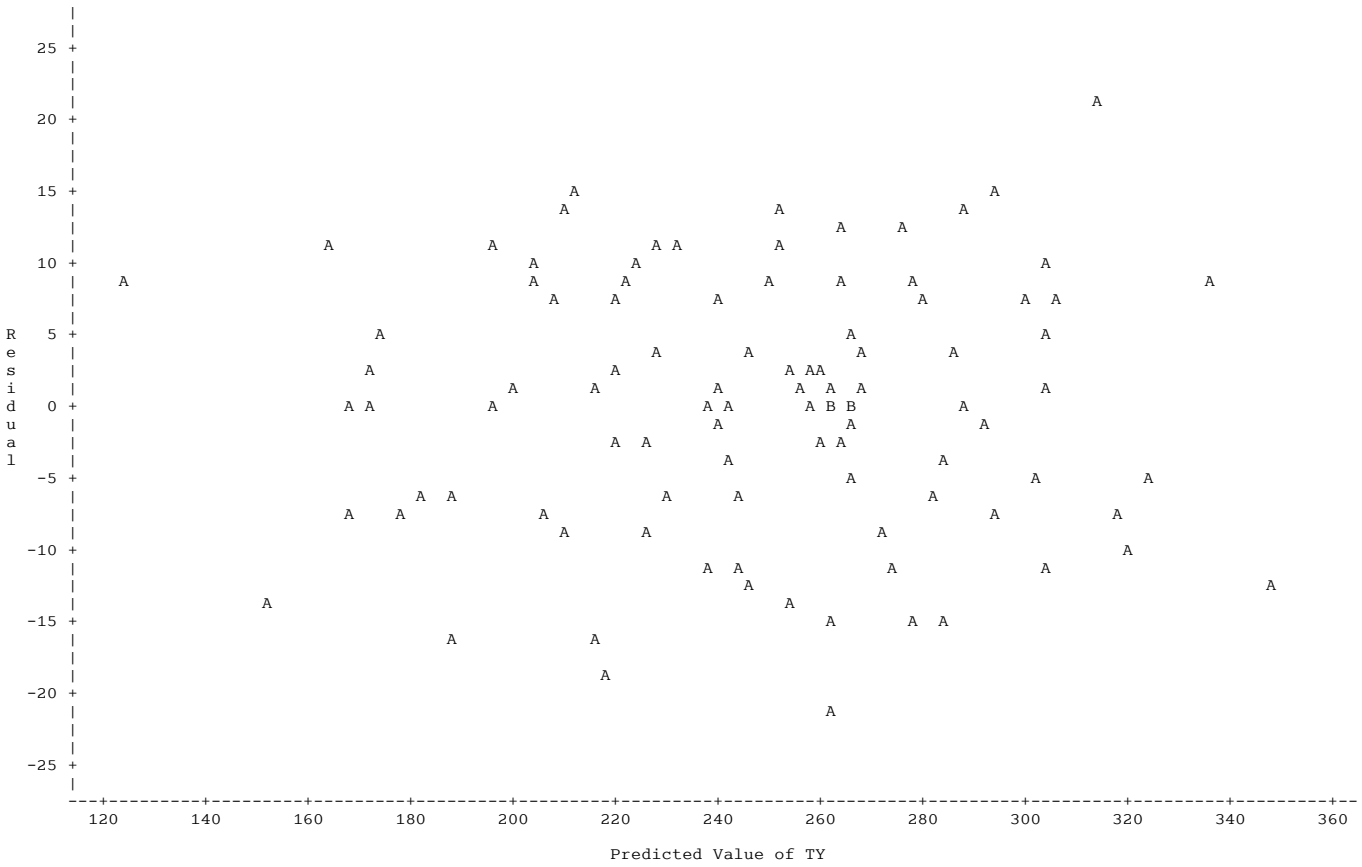
The REG Procedure
Model: MODEL1
Dependent Variable: TY Y Transformation



Plot of Y*hat_tildeY. Legend: A = 1 obs, B = 2 obs, etc.



Plot of tilde_e*hat_tildeY. Legend: A = 1 obs, B = 2 obs, etc.



The UNIVARIATE Procedure
Variable: tilde_e (Residual)

Moments			
N	100	Sum Weights	100
Mean	0	Sum Observations	0
Std Deviation	9.00474978	Variance	81.0855187
Skewness	-0.1612325	Kurtosis	-0.5836252
Uncorrected SS	8027.46635	Corrected SS	8027.46635
Coeff Variation	.	Std Error Mean	0.90047498

Basic Statistical Measures			
Location		Variability	
Mean	0.000000	Std Deviation	9.00475
Median	0.241637	Variance	81.08552
Mode	.	Range	43.10508
		Interquartile Range	13.79151

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	0	Pr > |t|	1.0000
Sign	M	4	Pr >= |M|	0.4841
Signed Rank	S	54	Pr >= |S|	0.8538

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.985728	Pr < W	0.3579
Kolmogorov-Smirnov	D	0.073927	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.068403	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.45874	Pr > A-Sq	>0.2500

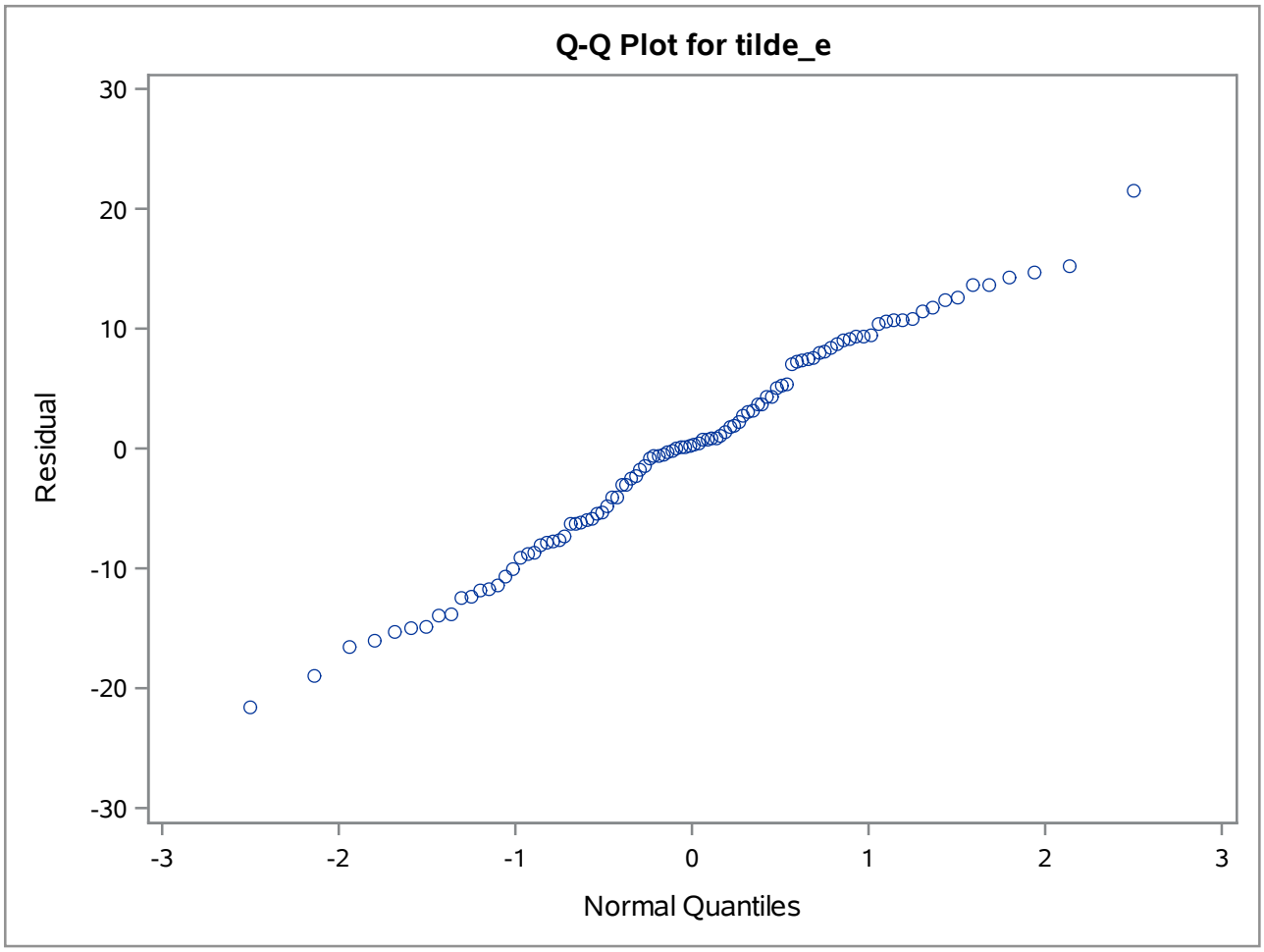
Quantiles (Definition 5)	
Level	Quantile
100% Max	21.548307
99%	18.400058
95%	13.638694
90%	11.133787
75% Q3	7.505117
50% Median	0.241637
25% Q1	-6.286393

The UNIVARIATE Procedure
Variable: tilde_e (Residual)

Quantiles (Definition 5)	
Level	**Quantile**
10%	-12.410891
5%	-15.112421
1%	-20.253009
0% Min	-21.556773

Extreme Observations			
Lowest		**Highest**	
Value	**Obs**	**Value**	**Obs**
-21.5568	92	13.6536	26
-18.9492	22	14.2779	71
-16.5513	66	14.6620	32
-16.0199	38	15.2518	88
-15.2625	59	21.5483	99

The UNIVARIATE Procedure



The REG Procedure
Model: MODEL1
Dependent Variable: TY Y Transformation

Number of Observations Read	100
Number of Observations Used	100

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	3	190475	63492	759.29	<.0001
Error	96	8027.46635	83.61944		
Corrected Total	99	198502			

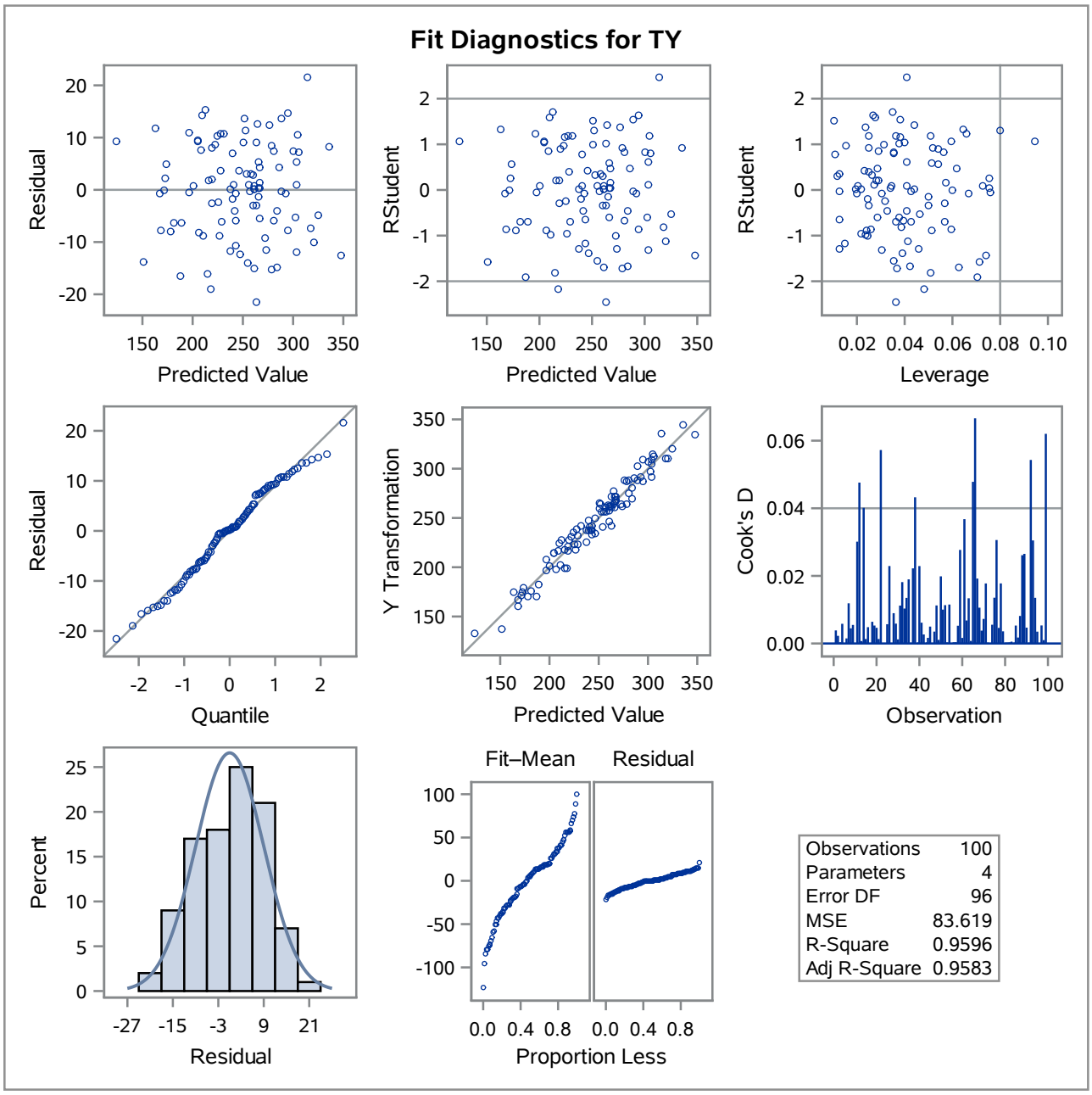
Root MSE	9.14437	**R-Square**	0.9596
Dependent Mean	247.15703	**Adj R-Sq**	0.9583
Coeff Var	3.69982		

Parameter Estimates						
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	Intercept	1	239.19831	5.84935	40.89	<.0001
TX1	X1 Transformation	1	2.23916	0.10743	20.84	<.0001
TX2	X2 Transformation	1	-3.02597	0.08121	-37.26	<.0001
TX3	X3 Transformation	1	4.62033	0.16873	27.38	<.0001

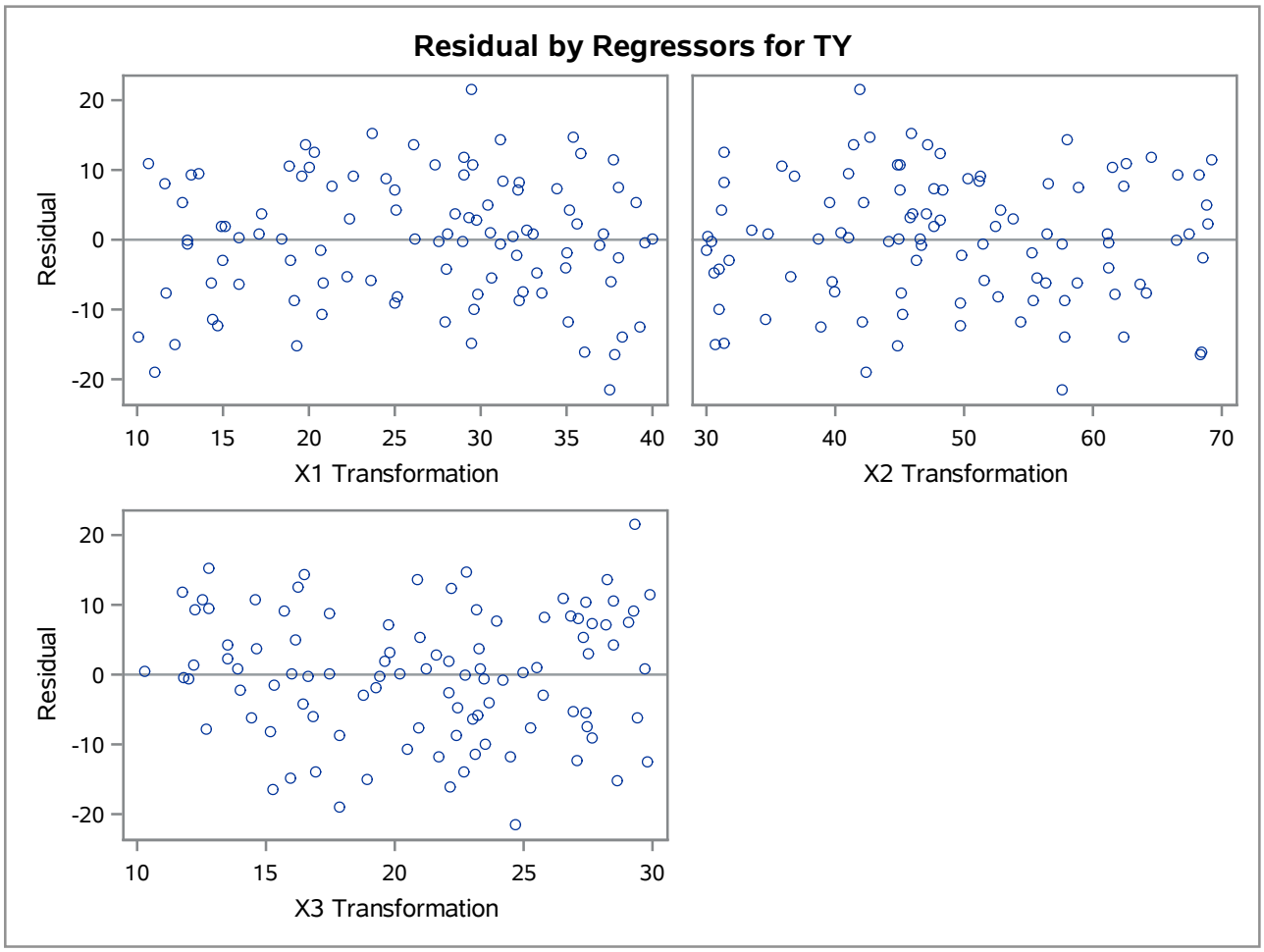
The REG Procedure
Model: MODEL1
Dependent Variable: TY Y Transformation

Test of First and Second Moment Specification		
DF	Chi-Square	Pr > ChiSq
9	8.19	0.5155

The REG Procedure
Model: MODEL1
Dependent Variable: TY Y Transformation



The REG Procedure
Model: MODEL1
Dependent Variable: TY Y Transformation



The GLMSELECT Procedure

Data Set	WORK.PROBLEM2
Dependent Variable	Y
Selection Method	Stepwise
Select Criterion	SBC
Stop Criterion	SBC
Effect Hierarchy Enforced	None

Number of Observations Read	2000
Number of Observations Used	2000

Class Level Information		
Class	**Levels**	**Values**
X26	5	Erath Monroe Schleich Sutton TomGreen
X33	5	ExtraLar ExtraSma Large Medium Small
X40	5	Blue Green Orange Red Yellow

Dimensions	
Number of Effects	41
Number of Parameters	53

The GLMSELECT Procedure

Stepwise Selection Summary					
Step	Effect Entered	Effect Removed	Number Effects In	Number Parms In	SBC
0	Intercept		1	1	15908.9724
1	X39		2	2	11980.8580
2	X16		3	3	11228.6993
3	X33		4	7	10128.0700*
* Optimal Value of Criterion					

Selection stopped at a local minimum of the SBC criterion.

Stop Details				
Candidate For	Effect	Candidate SBC		Compare SBC
Entry	X18	10131.7061	>	10128.0700
Removal	X33	11228.6993	>	10128.0700

The GLMSELECT Procedure
Selected Model

The selected model is the model at the last step (Step 3).

Effects:	Intercept X16 X33 X39

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	6	5366893	894482	5785.22
Error	1993	308148	154.61498	
Corrected Total	1999	5675040		

Root MSE	12.43443
Dependent Mean	95.68722
R-Square	0.9457
Adj R-Sq	0.9455
AIC	12091
AICC	12091
SBC	10128

Parameter Estimates					
Parameter		DF	Estimate	Standard Error	t Value
Intercept		1	-22.542150	0.921147	-24.47
X16		1	4.873465	0.121613	40.07
X33	ExtraLar	1	22.202187	0.875088	25.37
X33	ExtraSma	1	0.526099	0.866893	0.61
X33	Large	1	21.875775	0.887294	24.65
X33	Medium	1	22.667239	0.873461	25.95
X33	Small	0	0	.	.
X39		1	17.002792	0.095274	178.46