```
/* PROBLEM 1 */
Proc import out = dishwashers
datafile = "/home/u60672671/Data Files/Dishwasher.xlsx"
DBMS = xlsx;

Proc corr data = dishwashers plots = matrix;
var score price electricity gas time;
run;

/* 1) Which variables have a positive correlation with overall satisfaction (score)?
    Is the relationship of this/these variable(s) linear  with respect to score?
    If so, test whether the correlation is significant?
    State the null and alternative hypothesis,
    state the p-value and state your conclusion at α =0.05.
    */

        /* Only price has a positive correlation with overall satisfaction, a correlation coefficient
        of 0.66076.

        The relationship between price and score is linear, since the scatterplot resembles
        a linear curve.

        Null Hypothesis: p = 0
        Alternative Hypothesis:p =/= 0

        P-value = 0.0021

        At a significance level of 0.05, we would conclude that the null hypothesis should
        be rejected. Therefore, we can say that the correlation coefficient is statistically significant.


    2) Which variables have a negative correlation with overall satisfaction?
        Is the relationship of this/these variable(s) linear with respect to score? ''


        Electricity, Gas and Time have a negative correlation with overall satisfaction.
        No, none of these have a linear relationship with respect to the score.

    3) Electricity hsa a negative correlation value.
        What does this mean in terms of the score?

        This means that electricity and score have an inverse relationship.
        As electricity increases, score decreases, and so on.

        */


ods graphics on;
Proc reg data = dishwashers;
Model score = price;
run;


/*  c) Preform a liner regression of score(y) on price (x)
    (2) What is the predicted equation?

        y = 34.10332 + 0.08141x

    (3) Interpret the slope of the line?

        For every increase in the price, the score increases by 0.08141.

    (4) What is the adjusted R^2 value?
        Does the model seem to be a good fit?
        Properly interpret the  adjusted R^2.

        The adjusted r^2 value is 0.4035.
        The model does not seem to be a good fit since the points in the fit
        plot are not evenly distributed about the line, and are also not very
        close to the line.


        About 40.3% of the sample variation in the price can be explained by the linear relationship
        of the price as a function of the score.

    (5) Evaluate all three residual assumptions.
        Explain which graph you use to evaluate each assumption.
        You may number them 1, 2, 3,..., 9.
```

```
        Linearity vs. Unbiased:
        In graph 1, we can see that there is no mathematical function
        pattern. The residuals here are more or less randomly distributed.

        Normal distribution of residuals:
        In graphs 4 and 7, we can see that the data points do not follow a close linear pattern
        and do are not symmetrical/normally distributed, respectively. Therefore, the data
        does not follow this assumptuon.

        Homoscedascticity:
        In graphs 1 and 2, we can see that there is no vertical pattern in CL around
        the residuals.

    (6) Predict the score of a dishwasher with a price of $535. Include an estimate of the
        95% confidence interval of the prediction

        yhat = 34.10332 + 0.08141(535)
        yhat  = 77.65767

        RMSE = 7.92424
        77.65767 +/- 2(7.92424)

        95% CL: (61.80919, 93.50615)



    D) Preform a quadratic regression of score (y) on electricity (x)
    */


data dishwashers2;
set dishwashers;
electsq = electricity**2;
run;

ods graphics on;
Proc reg data = dishwashers2 plots = predictions(x =electricity);
Model score = electricity electsq;
run;


/*  (3) What is the predicted equation?
        The predicted equation is

        y = -1048.17517 + 35.87050x - 0.28590x^2
    (4) What is the adjusted r^2 value? Does the model seem to be a good fit?

        The adjusted r^2 value is 0.5350. The model does not seem to be a good fit
        since the prediction points are not closley hugging the fitted line.

    (5) Evaluate all three residial assumptions. Explain which graph you're using for each one

        Linearity:
        By looking at graph 1, we can see that there is no mathematical
        function pattern; the residuals are more or less randomly distributed.
        Therefore, the model meets this assumption.

        Normal Distribution of Residuals:
        In graphs 4 and 7, we can see that the residuals follow a linear slope
        pretty closely, and they are also pretty normally distributed or symmetrical
        in graph 7.

        Homoscedasticity:
        The residuals in graph 1 and 2 also do not have a vertical pattern, therefore
        this model does meet the assumption of constant variance.

        */

/*PROBLEM 2*/
Proc import out = bonecracking
datafile = "/home/u60672671/Data Files/Bone Cracking Hypercarnivores Data.xlsx"
DBMS = xlsx;

Proc reg data = bonecracking;
Model SEJ = MA;
run;

        /* a) Performa a linear regression of SEJ vs, MA.
              Write out the model (SEJ  = m*MA + b).
              Include the ANOVA table, Rsq table,
              and parameter estimates table from your SAS output.
```

```
        SEJ = (31.81161)MA - 2.76054

        b) Analyze the r-square value and the output from the 9 panel data.
           Why is a linear regression model not the best model?
           Include the 9 panel residual analysis plots.

        ADJ R^2: 0.7486

        A linear regression model is not the best model. Firstly, the observations in the
        fit plot do not closley hug the fit line. There does also
        seem to be something of a mathematical function pattern in panel 1; the
        points in this graph resemble a quadratic function. From graph 7, we can see
        that the data is also not normally distributed, being skewed left and very
        asymmetrical. In graph 4, we can see that the residuals do not follow
        a linear slope very well. Also, the residuals in the first two graphs are not
        uniformly distributed, and clump together in vertical patterns, implying
        that the model does not meet the requirement
        of constatny variance.


        c) Preform a quadratic regeression. State the model and r squared vale.
           Include the command to create the scatter plot and the model overlay.
           Briefly analyze the residual assumptions

         */
Data bonecracking2;
set bonecracking;
MA2 = MA ** 2;
run;

Proc reg data = bonecracking2 plot = predictions (x = MA);
Model SEJ = MA MA2;
run;
        /*

        SEQ = 236.05951(MA^2) -73.35032 (MA) + 8.60935

        R-Squared = 0.8006

        This model is a decent fit, but certainly not perfect. In the fit scatterplot, most of the prediction point
        hug the line. Furthermore, in panel 1, the residuals are pretty randomly strewn about
        which implies that the model is relatively unbiased. In panel 7, we can see that
        the model does not meet the criteria of being normally disributed, since this graph is asyemetrical.
        Furthermore, the data points at the end of the linear line in graph 4 curve up, further
        implying that this model may not meet the assumption of normality. The residuals
        in graphs 1 and 2 also clump together to the left, implying there may not be
        constant variance.

        */


Data bonecracking3;
Set bonecracking;
lnSEJ = log(SEJ);
run;

Proc reg data = bonecracking3;
Model lnSEJ = MA;
output out = bones predicted = predictedlnSEJ;
run;

data untransformbonecracking3;
set bones;
predictedSEJ = exp(predicredlnSEJ);
run;

proc sort data = untransformbonecracking3;
by MA;

symbol1 value = diamondfilled color = goldenrod;
symbol2 color = blueviolet i = join;
proc gplot data = untransformbonecracking3;
plot SEJ * MA = 1 predictedSEJ*MA = 2/overlay;
run;


        /* D) Perform an exponential regression using a log transformation. State the model.
        Analyze the three residual assumptions. Include the scatterplot along with
```

```
the untransformed model.

The Model: ln(y) = -0.16884  + 7.14161(x)
Adjusted R squared: 0.8004

Unbiased:
There does not seem to be a mathematical function in the residuals in graph 1,
therefore it would seem that this model is unbiased.

Normal Distributed:
The residuals in graph 4 hug the slope of a linear line quite closely, while the
data is also pretty normally distributed or symmetrical in graph 7.

Constant Variance:
The residuals in graphs 1 and 2 do hug left, but are overall pretty uniformly spread out.



E) Of the quadratic model, the exponential model, and the power model, which model has the best
fit and why?

I would say the exponential model has the best fit because it meets the three residual
assumptions most closely, while also having a nice R squared value. The data seems to be the most
normally distributed with this model, since the residuals are quite symmetrical in graph 7.
The points in the fit plot also hug the line pretty closely, especially in comparison to the two other
models. Therefore, I feel that the exponential model is the best fit.
*/
```

The CORR Procedure

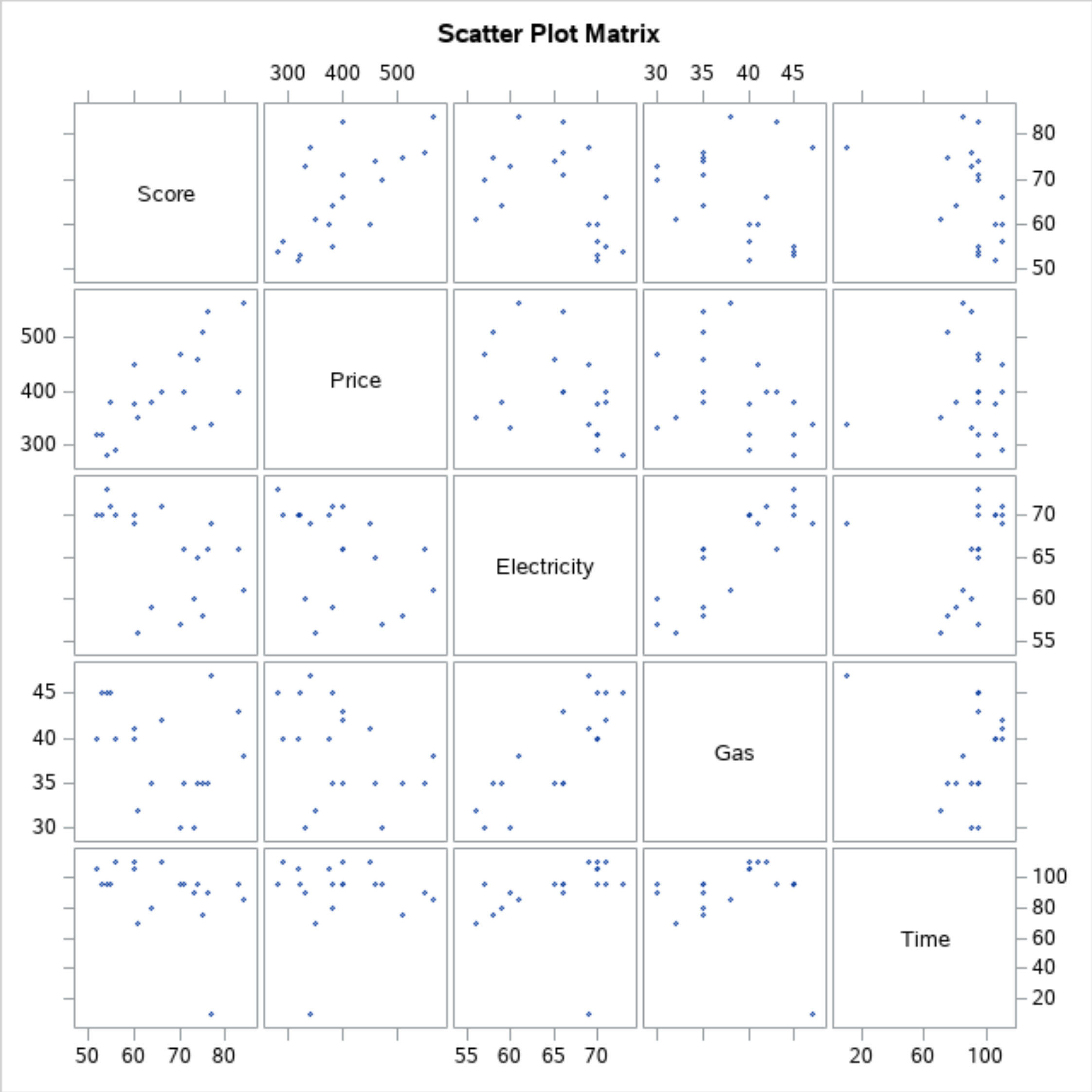
5 Variables:	Score Price Electricity Gas Time

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
Score	19	66.52632	10.25978	1264	52.00000	84.00000	Score
Price	19	398.26316	83.27187	7567	280.00000	565.00000	Price
Electricity	19	65.63158	5.44886	1247	56.00000	73.00000	Electricity
Gas	19	38.57895	5.23148	733.00000	30.00000	47.00000	Gas
Time	19	89.73684	22.32796	1705	10.00000	110.00000	Time

Pearson Correlation Coefficients, N = 19
Prob > |r| under H0: Rho=0

	Score	Price	Electricity	Gas	Time
Score **Score**	1.00000	0.66076 0.0021	-0.47335 0.0407	-0.31030 0.1960	-0.39102 0.0978
Price **Price**	0.66076 0.0021	1.00000	-0.42966 0.0664	-0.37734 0.1112	0.01050 0.9660
Electricity **Electricity**	-0.47335 0.0407	-0.42966 0.0664	1.00000	0.82645 <.0001	0.25716 0.2879
Gas **Gas**	-0.31030 0.1960	-0.37734 0.1112	0.82645 <.0001	1.00000	-0.10564 0.6669
Time **Time**	-0.39102 0.0978	0.01050 0.9660	0.25716 0.2879	-0.10564 0.6669	1.00000



The REG Procedure
Model: MODEL1
Dependent Variable: Score Score

Number of Observations Read	19
Number of Observations Used	19

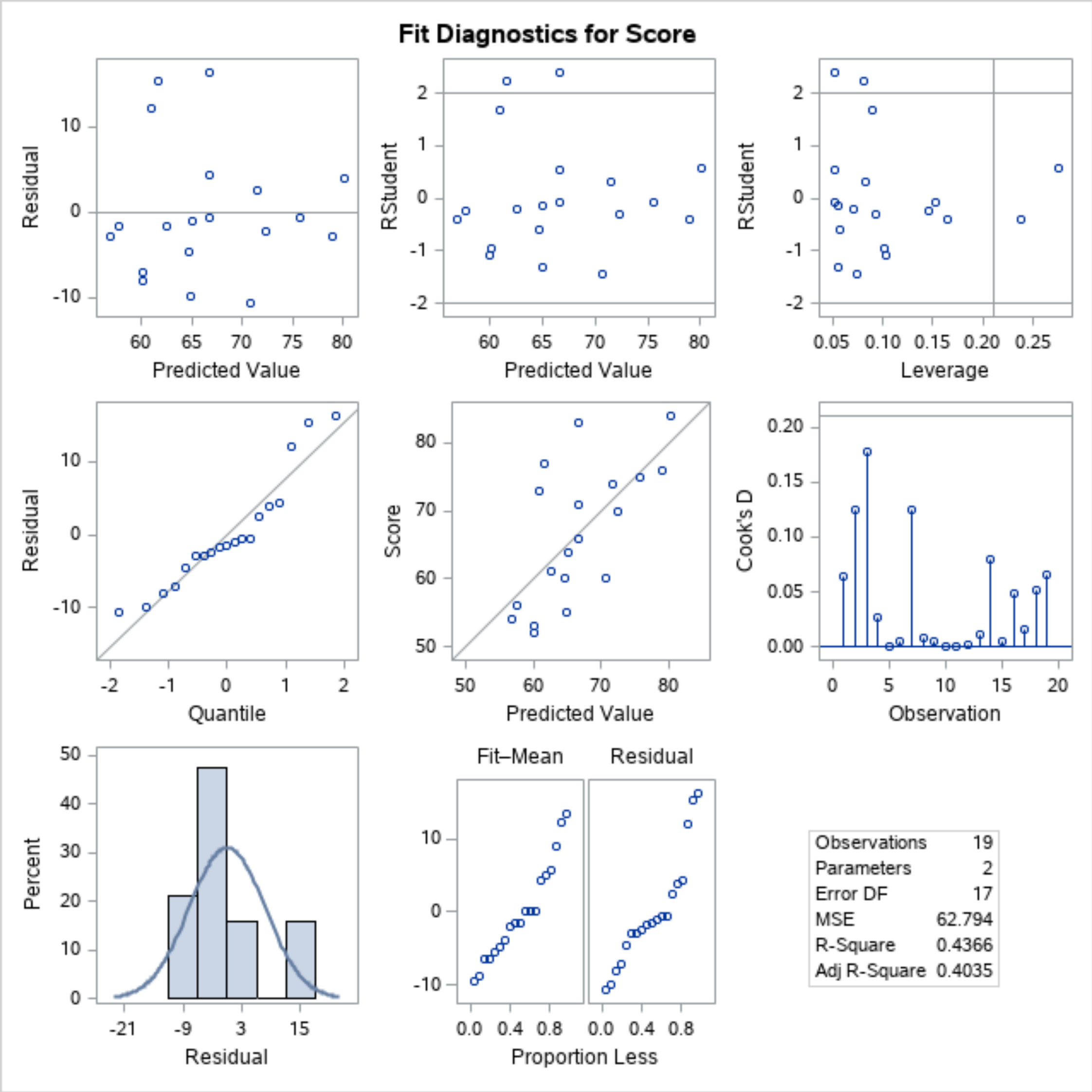
Analysis of Variance

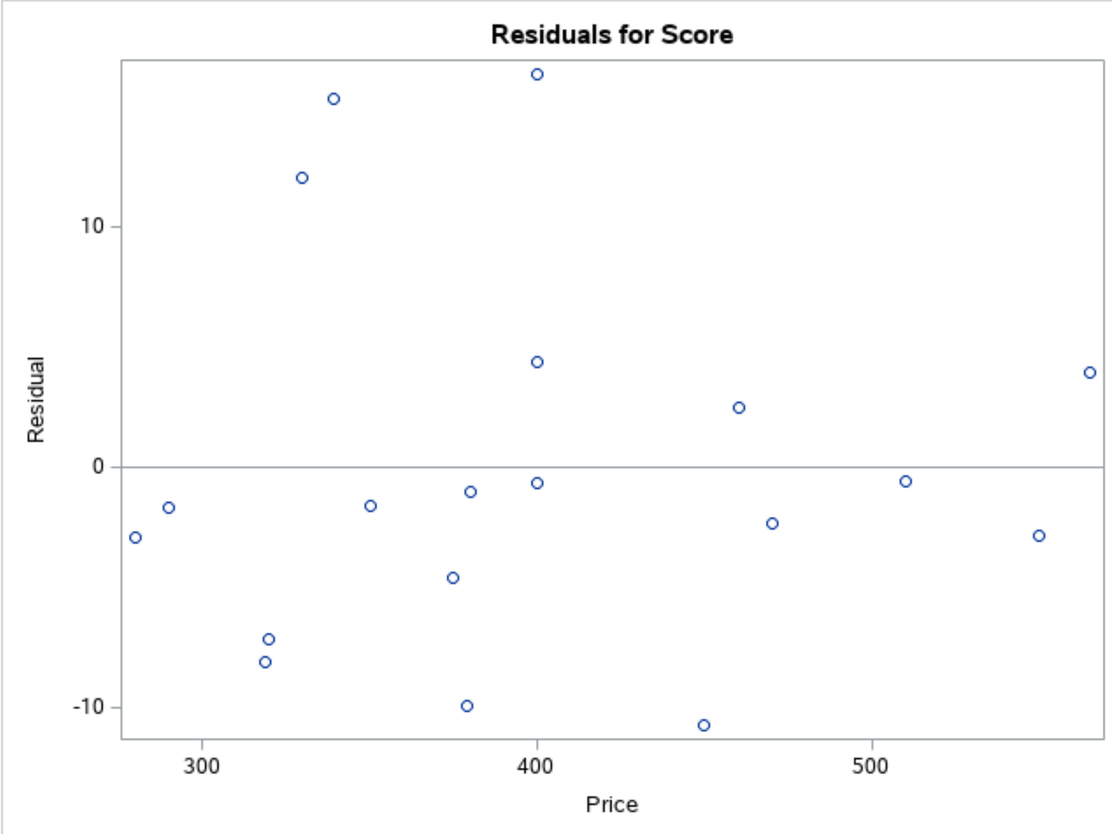
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	827.24706	827.24706	13.17	0.0021
Error	17	1067.48978	62.79352		
Corrected Total	18	1894.73684			

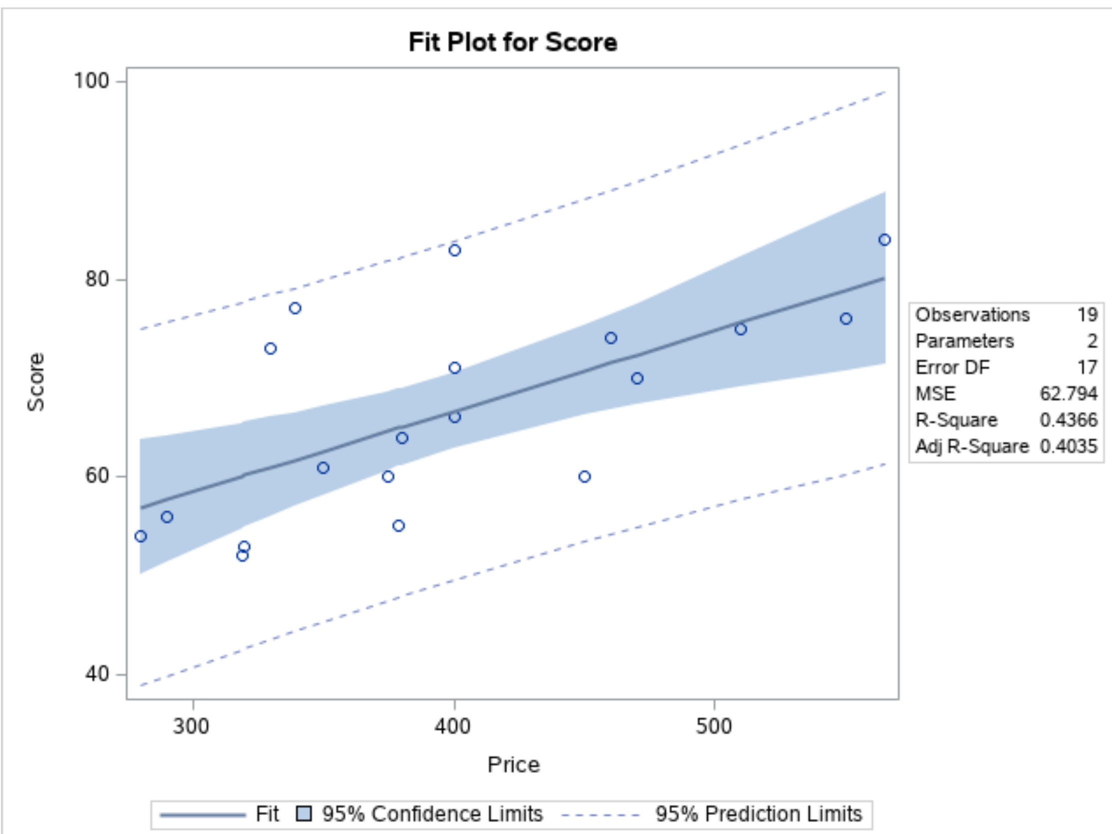
Root MSE	7.92424	**R-Square**	0.4366
Dependent Mean	66.52632	**Adj R-Sq**	0.4035
Coeff Var	11.91143		

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	34.10332	9.11602	3.74	0.0016
Price	Price	1	0.08141	0.02243	3.63	0.0021

The REG Procedure
Model: MODEL1
Dependent Variable: Score Score







The REG Procedure
Model: MODEL1
Dependent Variable: Score Score

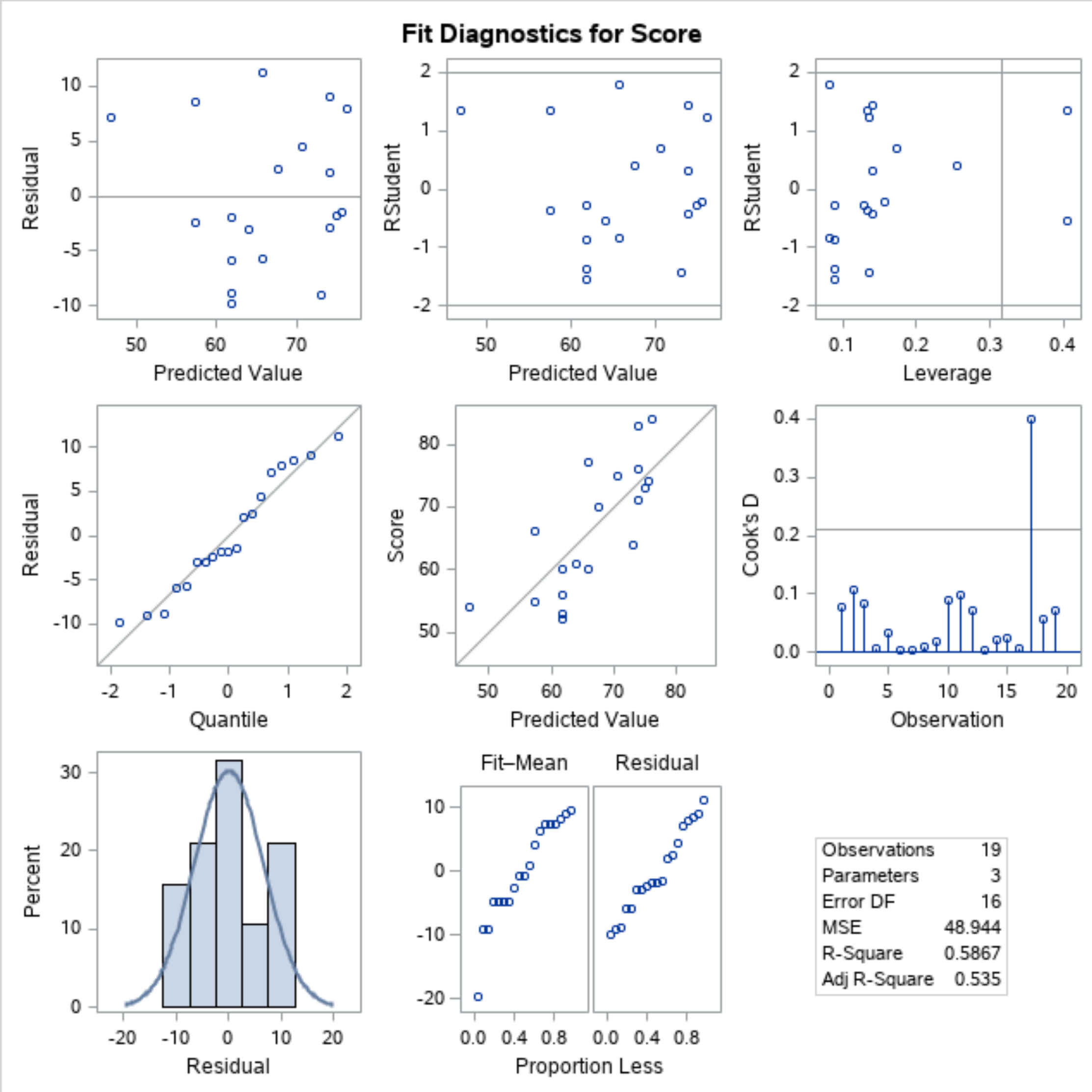
Number of Observations Read	19
Number of Observations Used	19

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	2	1111.63380	555.81690	11.36	0.0009
Error	16	783.10304	48.94394		
Corrected Total	18	1894.73684			

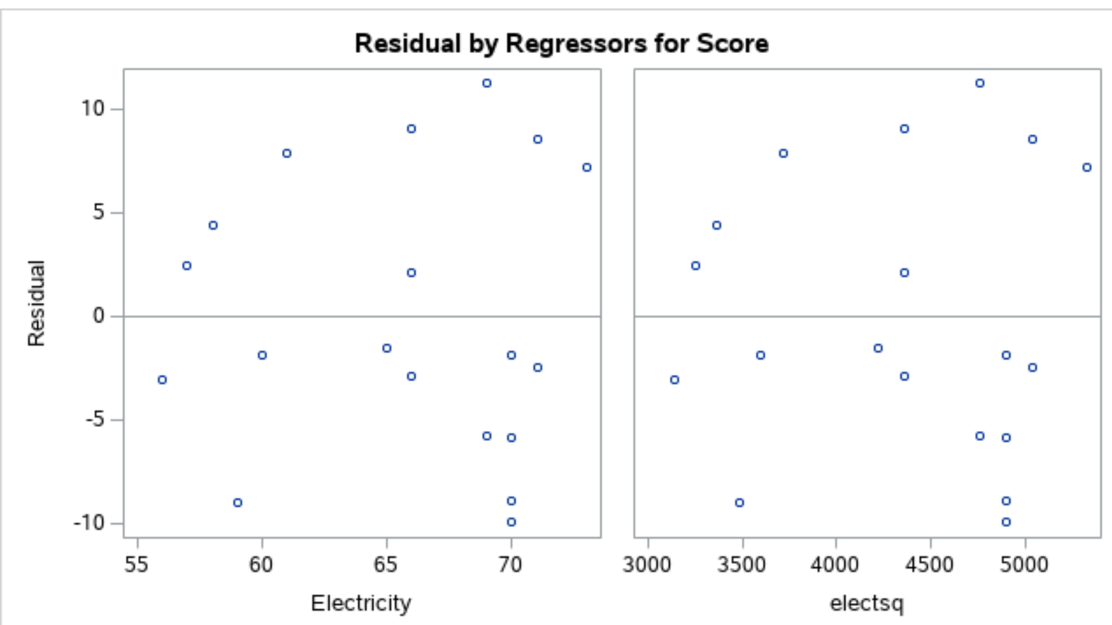
Root MSE	6.99599	**R-Square**	0.5867
Dependent Mean	66.52632	**Adj R-Sq**	0.5350
Coeff Var	10.51613		

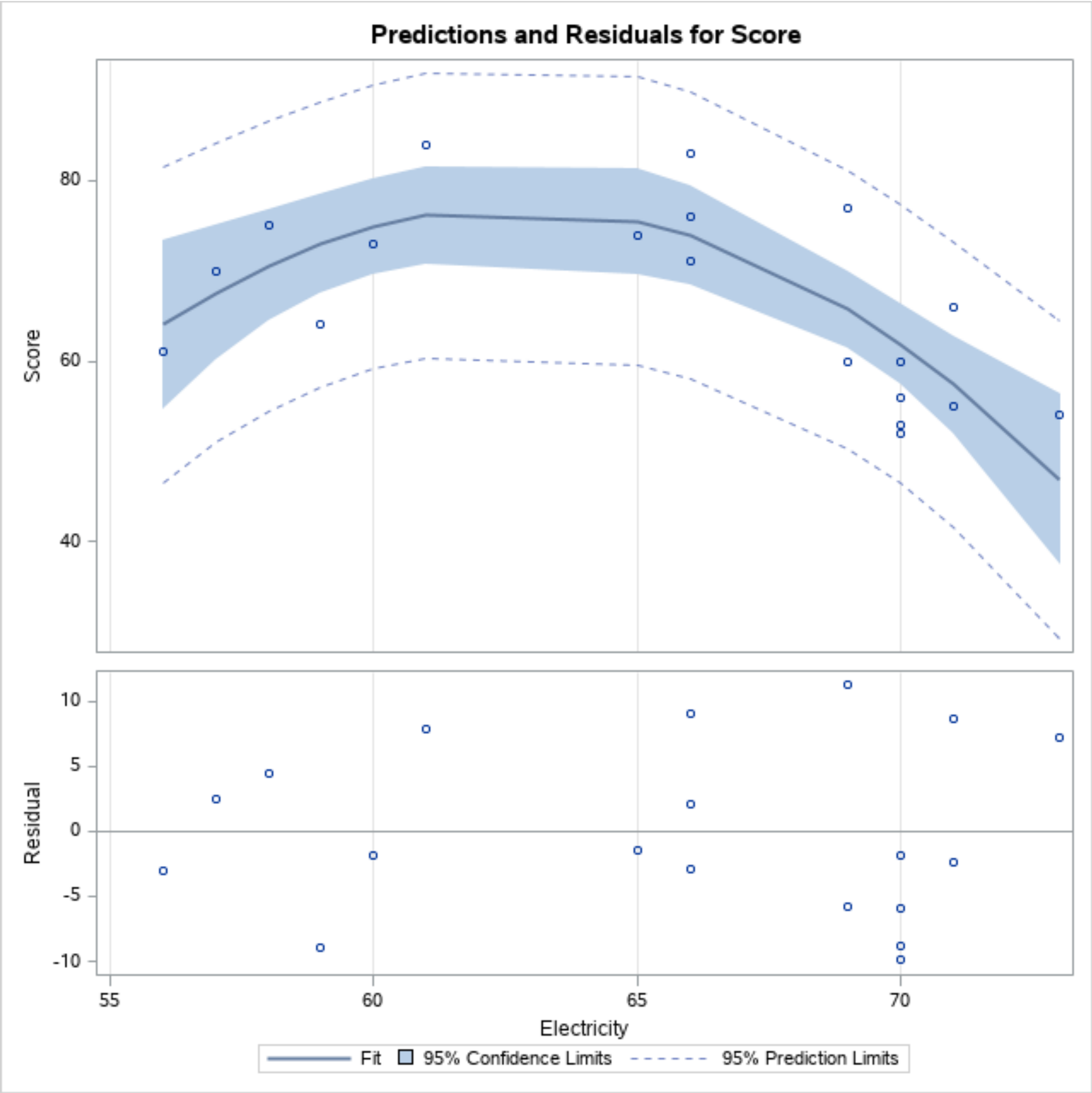
Parameter Estimates						
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	Intercept	1	-1048.17517	313.75190	-3.34	0.0041
Electricity	Electricity	1	35.87050	9.81614	3.65	0.0021
electsq		1	-0.28590	0.07630	-3.75	0.0018

The REG Procedure
Model: MODEL1
Dependent Variable: Score Score



Observations	19
Parameters	3
Error DF	16
MSE	48.944
R-Square	0.5867
Adj R-Square	0.535





The REG Procedure
Model: MODEL1
Dependent Variable: SEJ SEJ

Number of Observations Read	36
Number of Observations Used	36

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	48.94323	48.94323	105.20	<.0001
Error	34	15.81857	0.46525		
Corrected Total	35	64.76180			

Root MSE	0.68209	R-Square	0.7557
Dependent Mean	4.65333	Adj R-Sq	0.7486
Coeff Var	14.65818		

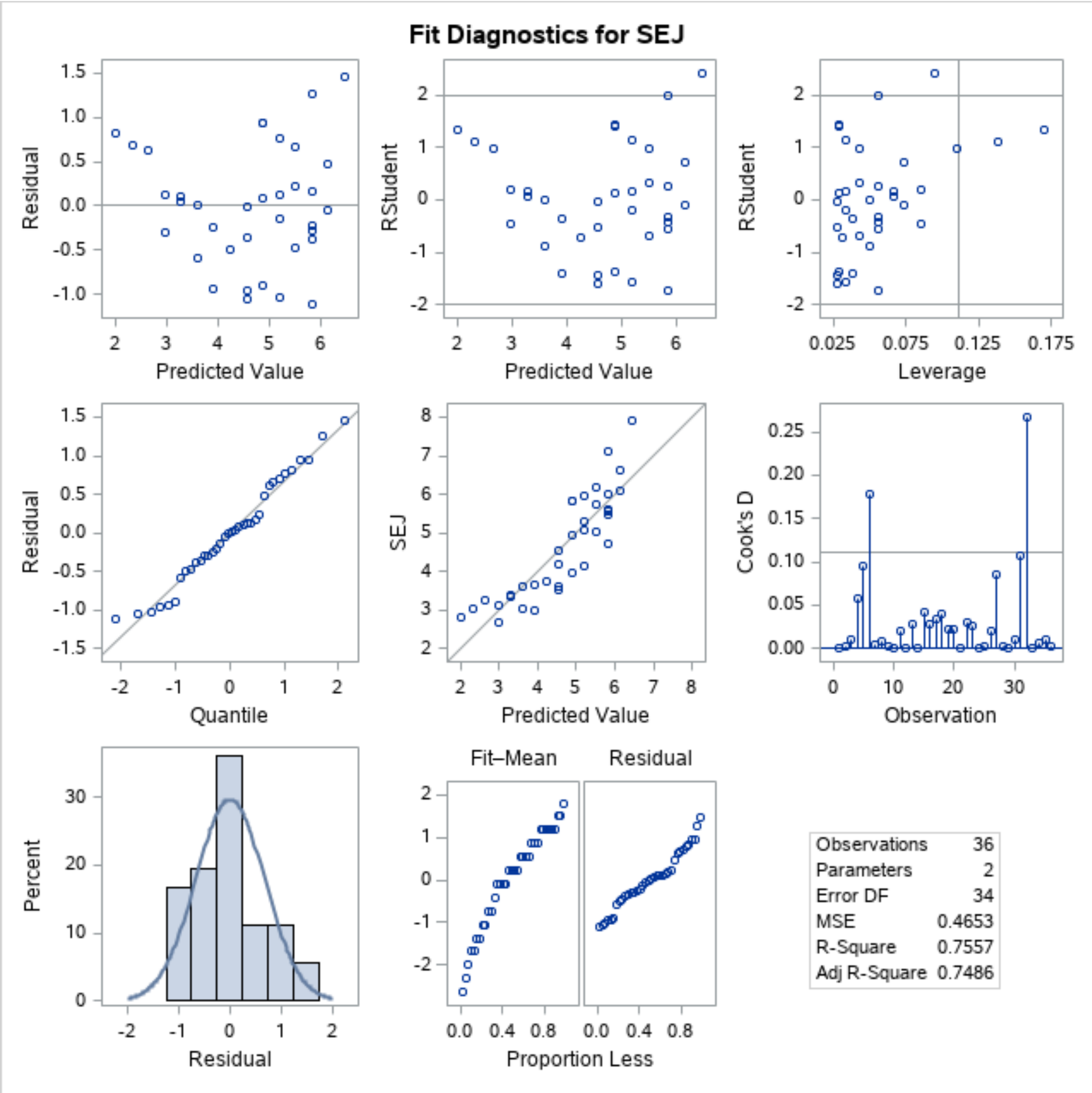
Parameter Estimates

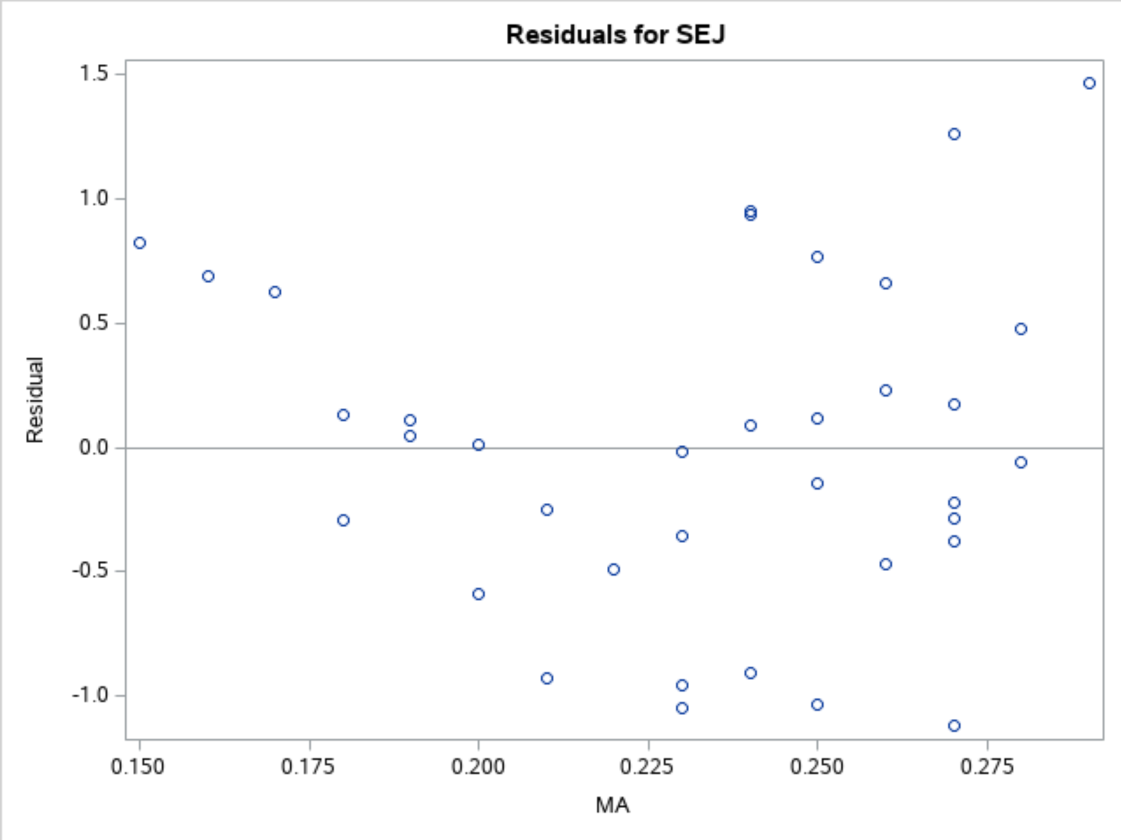
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-2.76054	0.73173	-3.77	0.0006

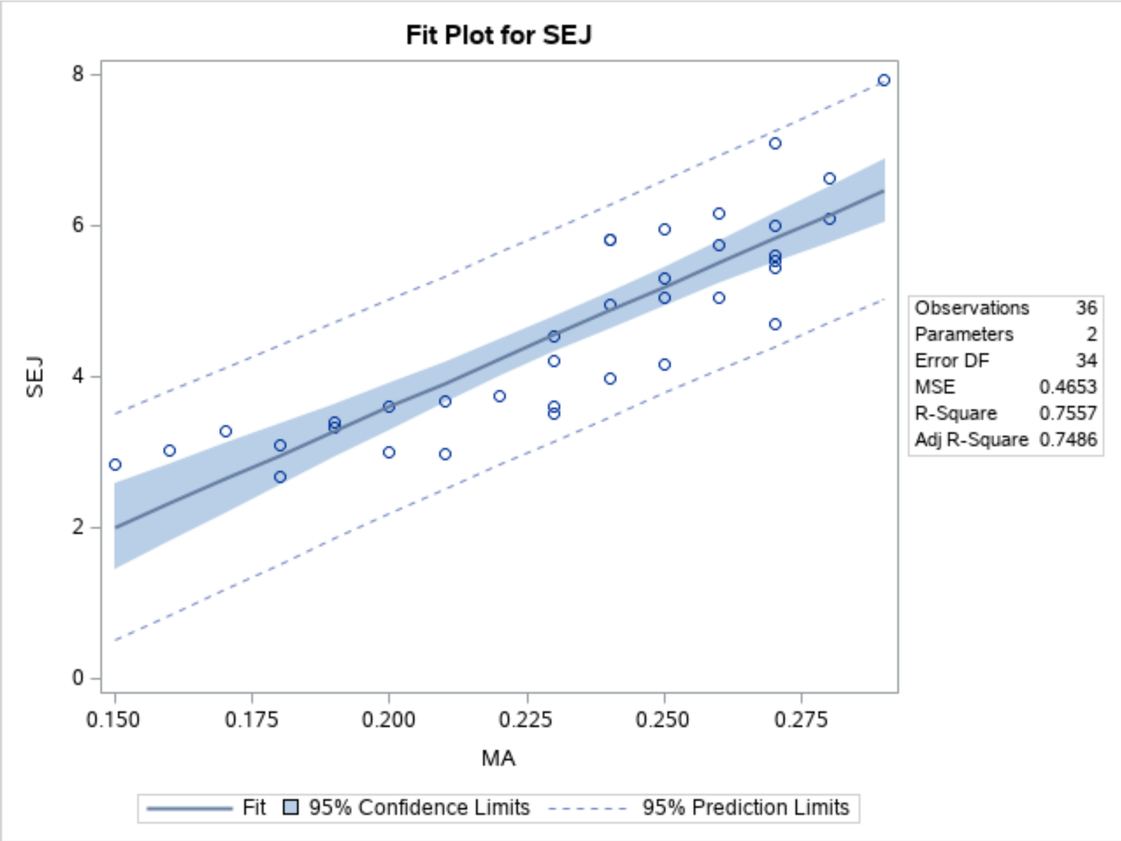
Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
MA	MA	1	31.81161	3.10158	10.26	<.0001

The REG Procedure
Model: MODEL1
Dependent Variable: SEJ SEJ







The REG Procedure
Model: MODEL1
Dependent Variable: SEJ SEJ

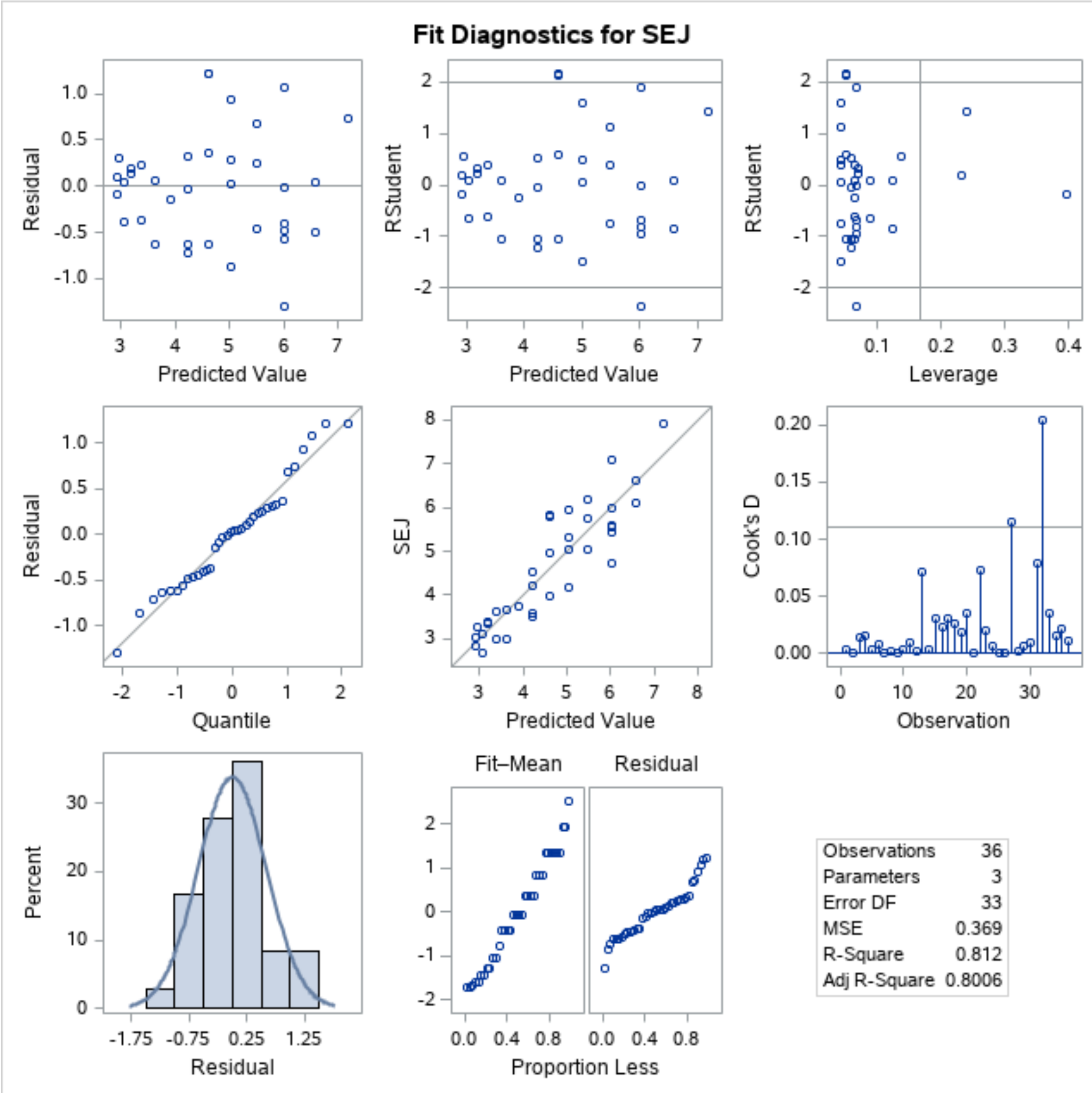
Number of Observations Read	36
Number of Observations Used	36

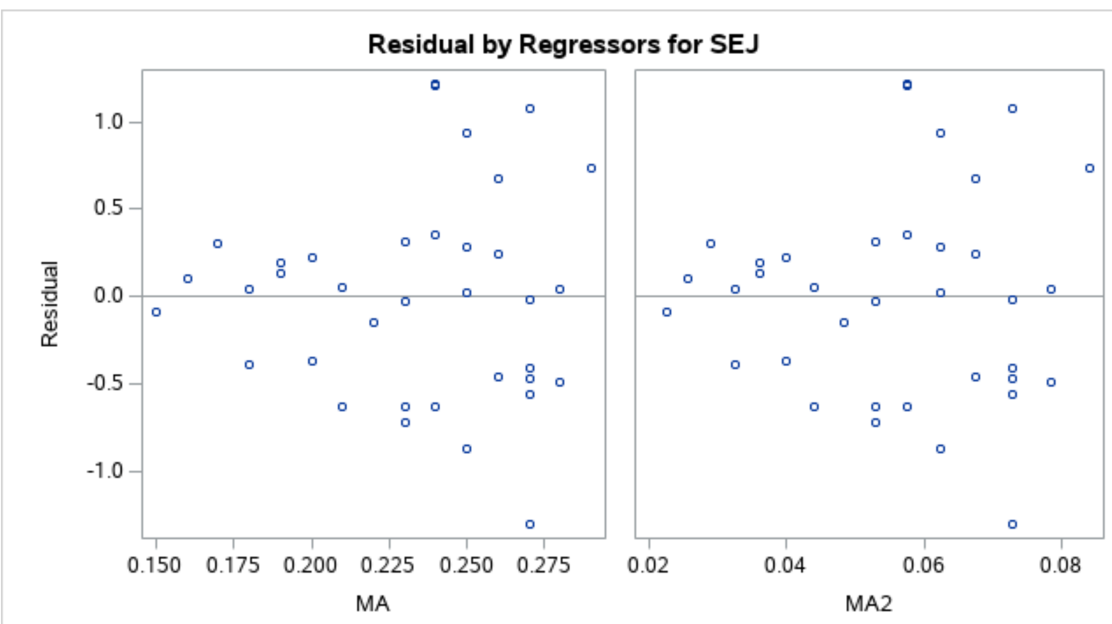
Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	2	52.58568	26.29284	71.26	<.0001
Error	33	12.17612	0.36897		
Corrected Total	35	64.76180			

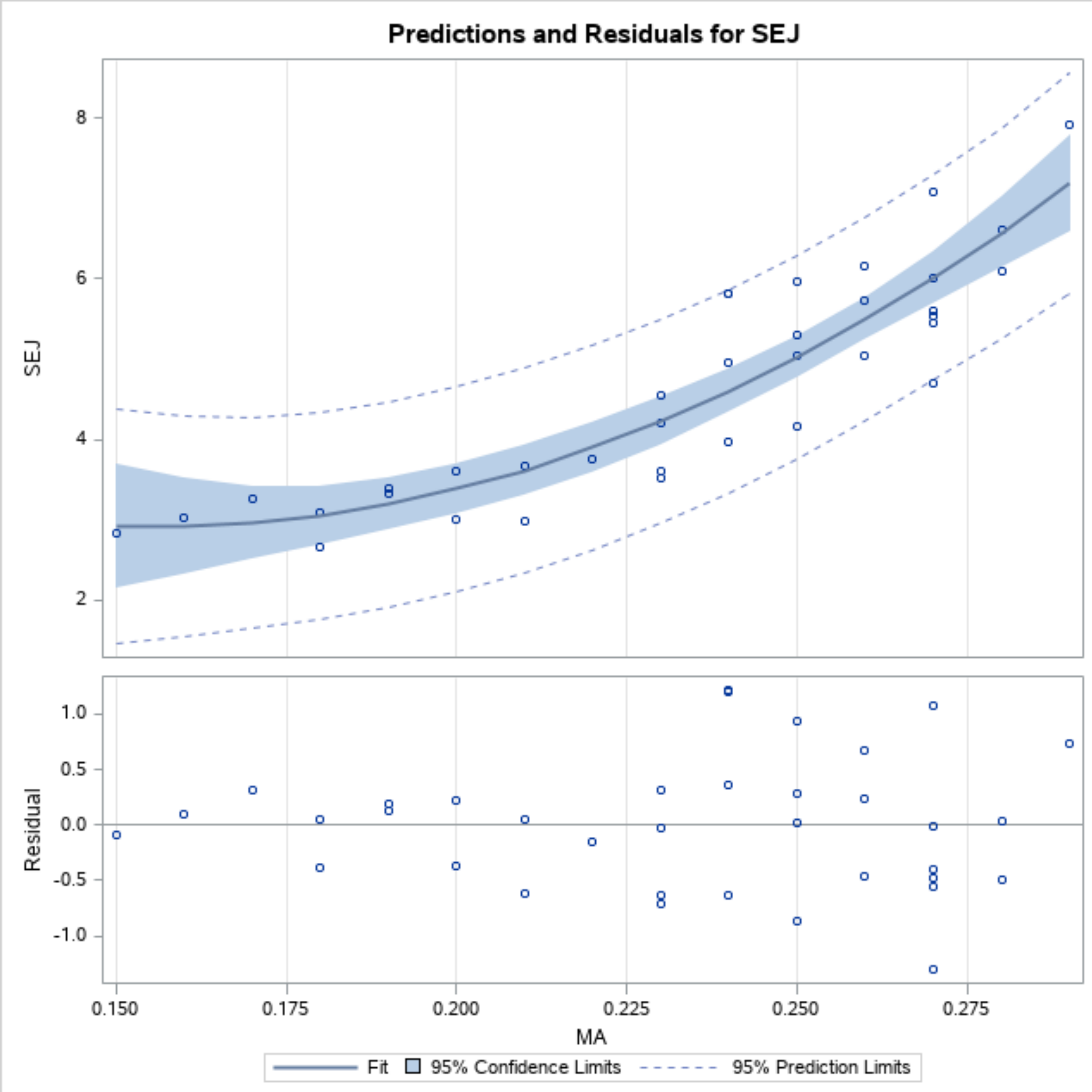
Root MSE	0.60743	**R-Square**	0.8120
Dependent Mean	4.65333	**Adj R-Sq**	0.8006
Coeff Var	13.05369		

Parameter Estimates						
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	Intercept	1	8.60935	3.67694	2.34	0.0254
MA	MA	1	-73.35032	33.58403	-2.18	0.0362
MA2		1	236.05951	75.13149	3.14	0.0035

The REG Procedure
Model: MODEL1
Dependent Variable: SEJ SEJ


Fit Diagnostics for SEJ
Residual
Predicted Value
RStudent
Leverage
Quantile
SEJ
Cook's D
Observation
Percent
Fit–Mean
Residual
Proportion Less
Observations 36
Parameters 3
Error DF 33
MSE 0.369
R-Square 0.812
Adj R-Square 0.8006


Residual by Regressors for SEJ
Residual
MA
MA2



The REG Procedure
Model: MODEL1
Dependent Variable: lnSEJ

Number of Observations Read	36
Number of Observations Used	36

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2.46668	2.46668	141.34	<.0001
Error	34	0.59339	0.01745		
Corrected Total	35	3.06007			

Root MSE	0.13211	**R-Square**	0.8061
Dependent Mean	1.49555	**Adj R-Sq**	0.8004
Coeff Var	8.83340		

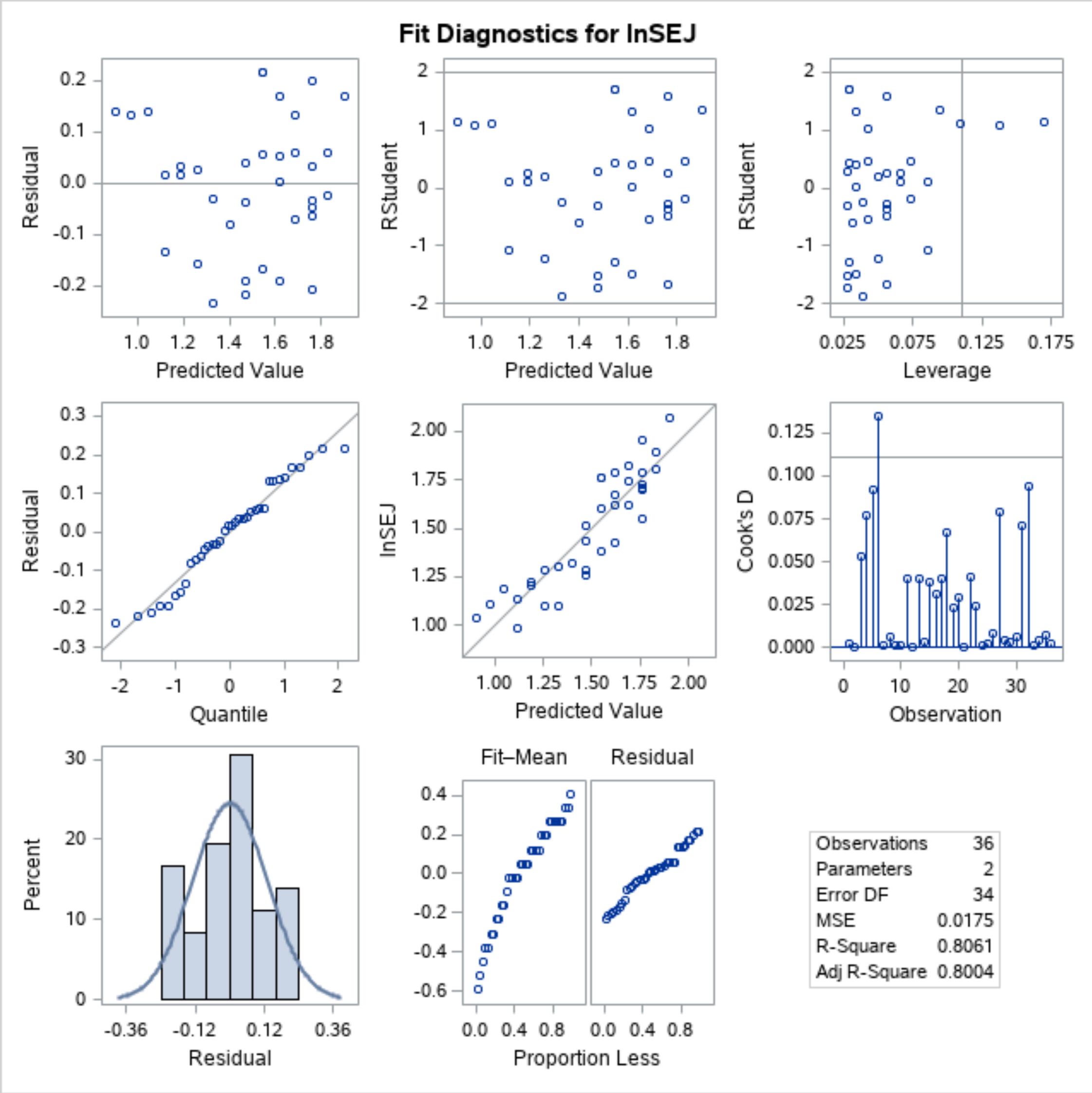
Parameter Estimates

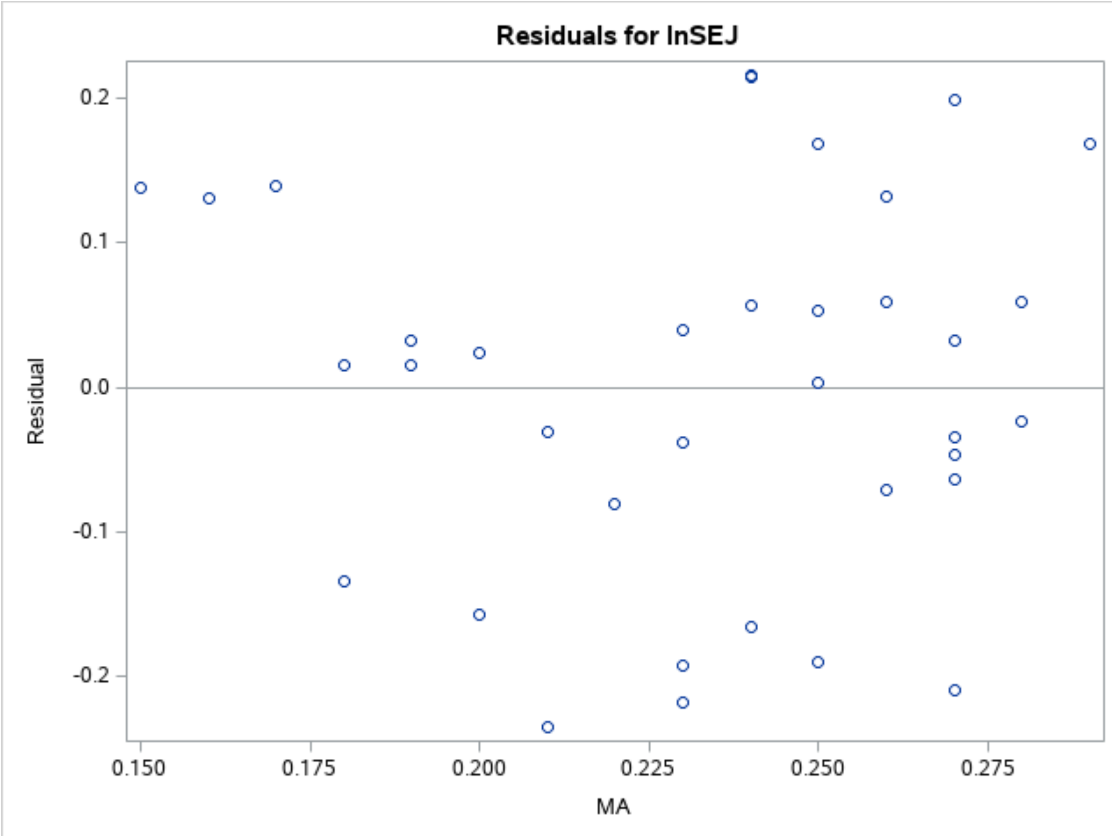
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-0.16884	0.14172	-1.19	0.2418

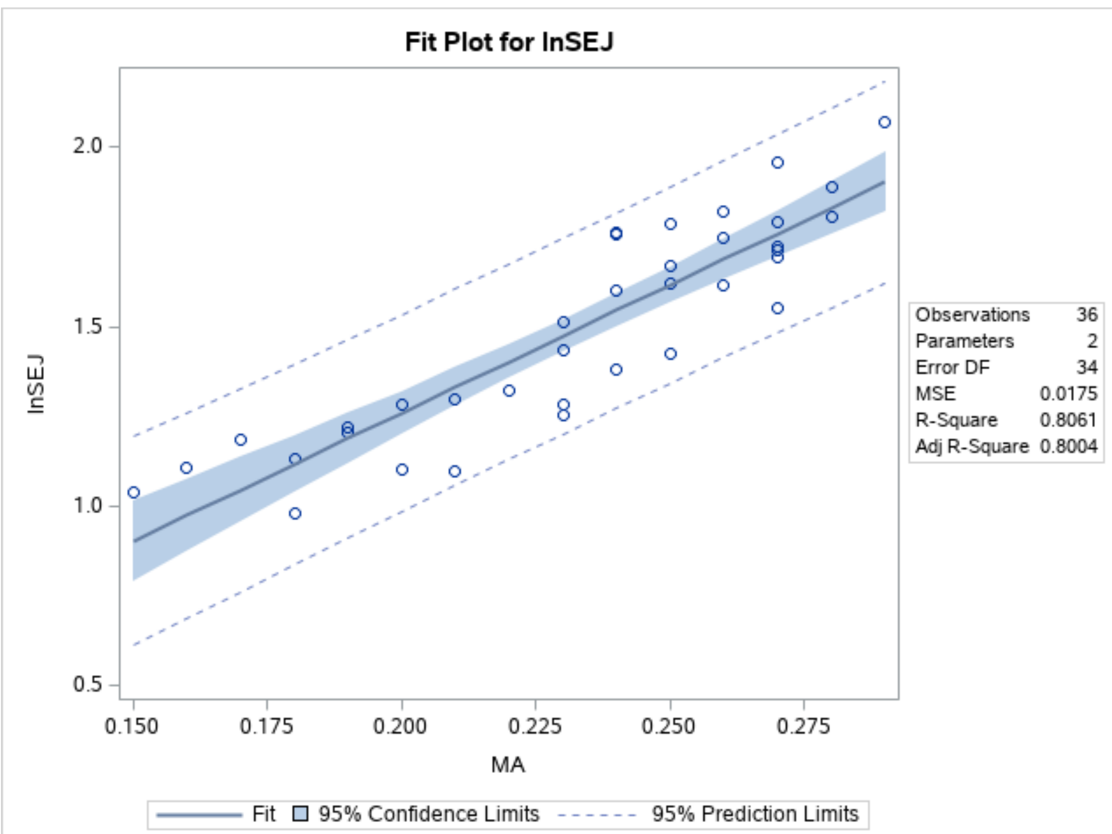
Parameter Estimates

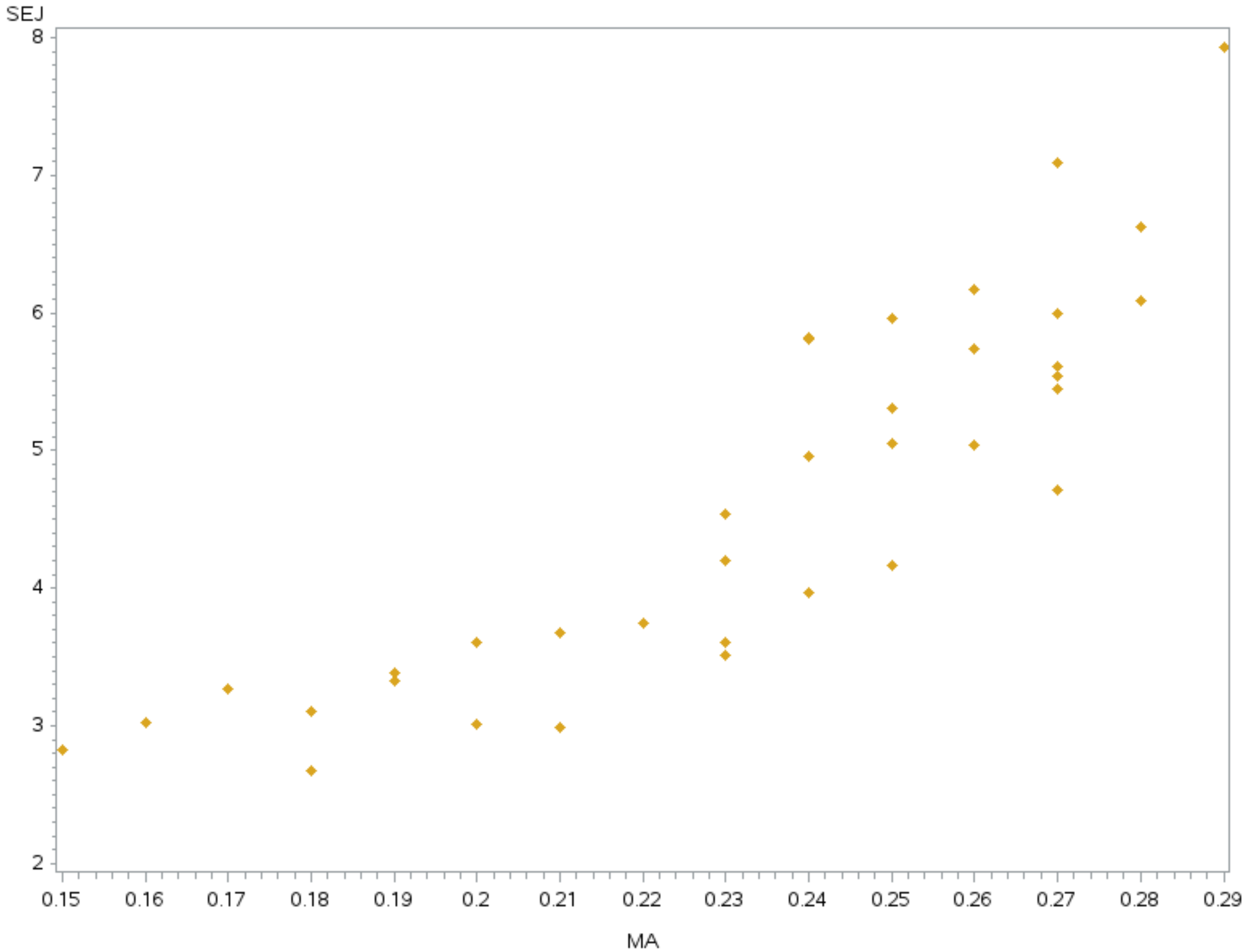
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
MA	MA	1	7.14161	0.60072	11.89	<.0001

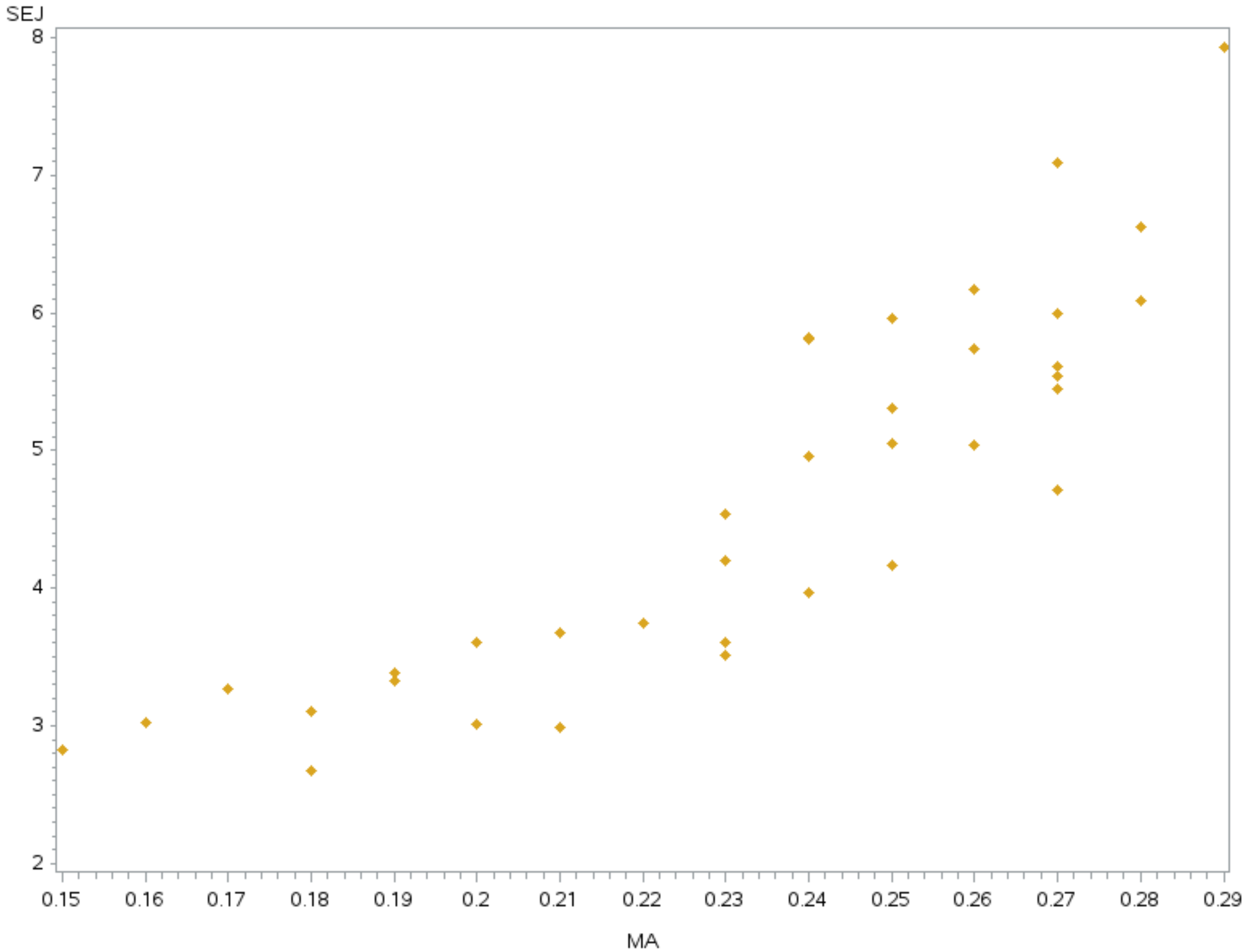
The REG Procedure
Model: MODEL1
Dependent Variable: lnSEJ




Residuals for lnSEJ
Residual
MA
0.2
0.1
0.0
-0.1
-0.2
0.150
0.175
0.200
0.225
0.250
0.275


Fit Plot for lnSEJ
lnSEJ
MA
2.0
1.5
1.0
0.5
0.150
0.175
0.200
0.225
0.250
0.275
Observations 36
Parameters 2
Error DF 34
MSE 0.0175
R-Square 0.8061
Adj R-Square 0.8004
Fit
95% Confidence Limits
95% Prediction Limits


SEJ
8
7
6
5
4
3
2
0.15
0.16
0.17
0.18
0.19
0.2
0.21
0.22
0.23
0.24
0.25
0.26
0.27
0.28
0.29
MA


SEJ
8
7
6
5
4
3
2
0.15
0.16
0.17
0.18
0.19
0.2
0.21
0.22
0.23
0.24
0.25
0.26
0.27
0.28
0.29
MA