Chapter 8 – A Closer Look at Assumptions for Simple Linear Regression

Case 8.1.1. Island Area and Number of Species – An Observational Study *R&S p. 208-209*

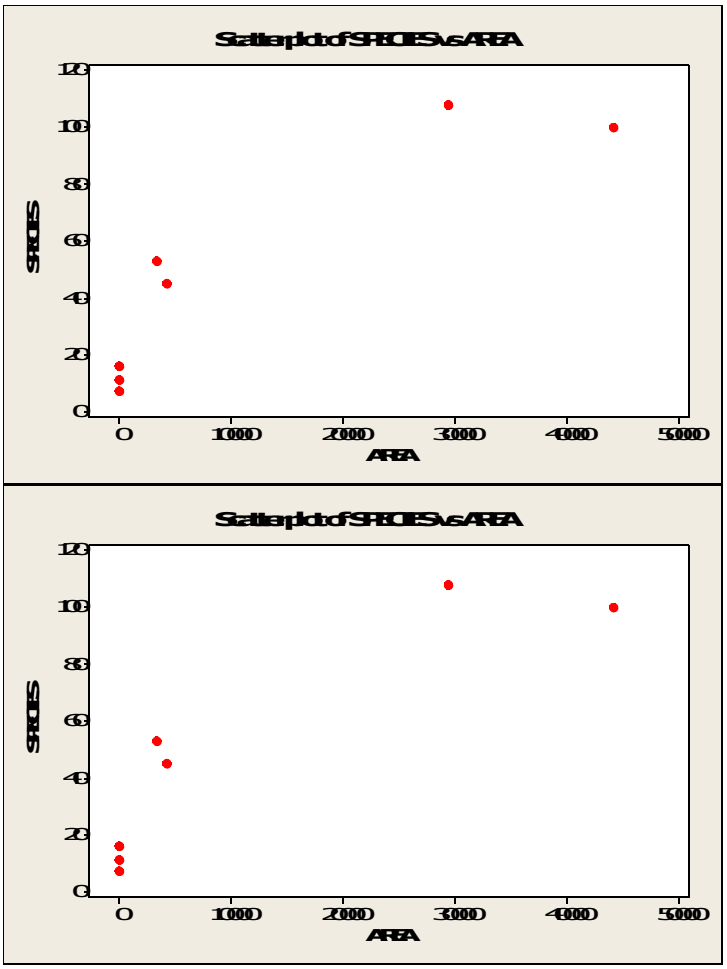
Step 1: Copy the data into a Minitab Worksheet: use these steps:

File → Open Worksheet →Browse your local directory and upload the csv file case0801.csv. The data will appear as two columns in Minitab with titles AREA and SPECIES. See the first three columns of ***R&S Display 8.1***.

Data Display

Row	AREA	SPECIES
1	44218	100
2	29371	108
3	4244	45
4	3435	53
5	32	16
6	5	11
7	1	7

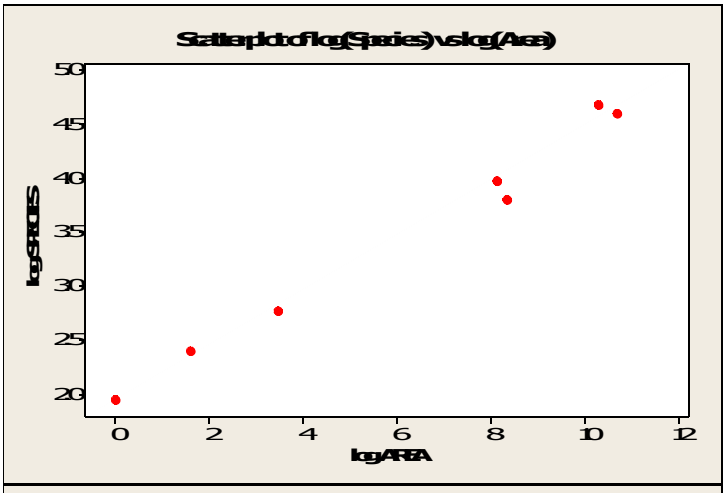
Step 2: To construct a Scatterplot of SPECIES versus AREA, go to Graph and select Simple graph option. Select SPECIES into Y variables and select AREA into X variables and click OK. You can add a title by going into Labels option. This produces this graph.

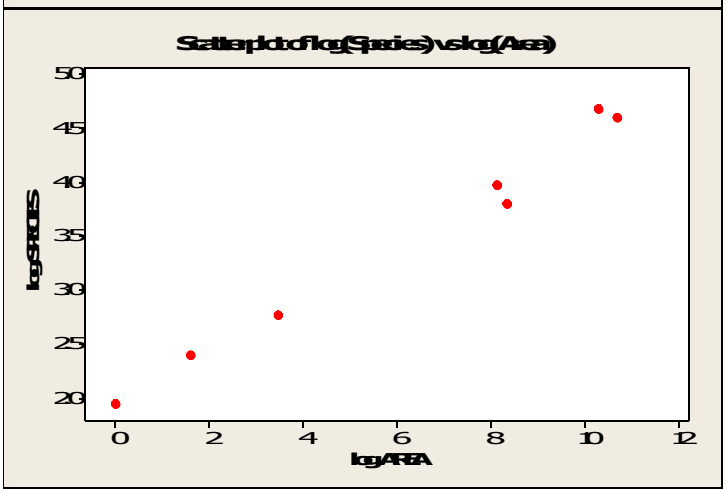


Step 3: Natural log Transform of both variables: Go to Calc→ Calculator→ Store Result in Variable C3; Under Expression, select Natural log (log base e); it will bring up LN( ) in the Expression box. Insert C1 within parenthesis, so we have LN(C1) and click ok. This will create a new column C3; you can name it ln(Area). Repeat to get ln(Species) into C4. Display this data and do a scatterplot.

Data Display

Row	AREA	SPECIES	ln(Area)	ln(Species)
1	44218.00	100.00	10.6969	4.60517
2	29371.00	108.00	10.2878	4.68213
3	4244.00	45.00	8.3533	3.80666
4	3435.00	53.00	8.1418	3.97029
5	32.00	16.00	3.4657	2.77259
6	5.00	11.00	1.6094	2.39790
7	1.00	7.00	0.0000	1.94591



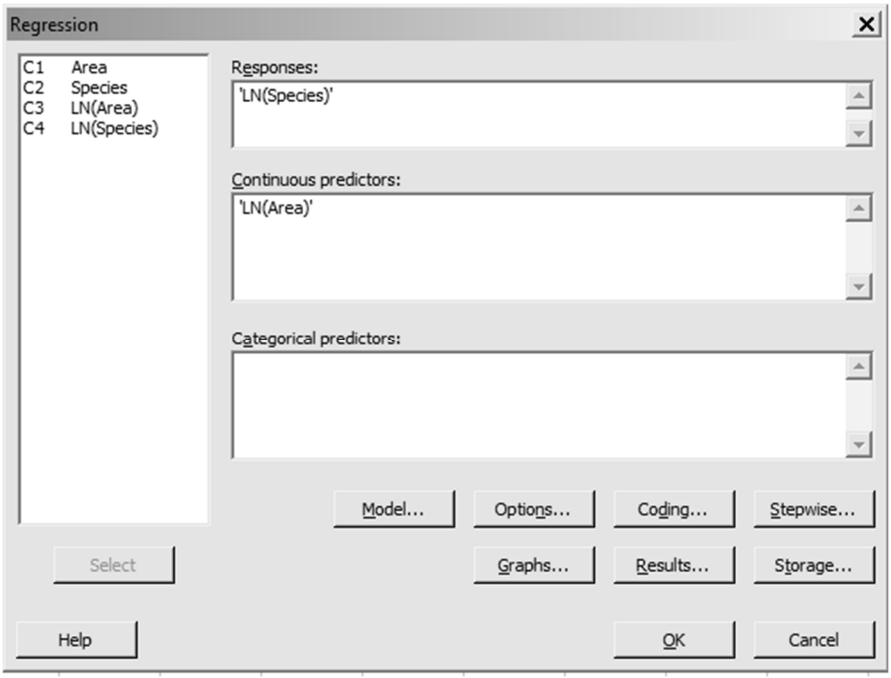


Step 4: Correlation: Go to Stat →Basic Statistics → Correlation; select variables log(SPECIES) and log(AREA) and click ok to produce this output.

Correlations: log AREA, log SPECIES

```
Pearson correlation of log AREA and log SPECIES = 0.994
P-Value = 0.000
```

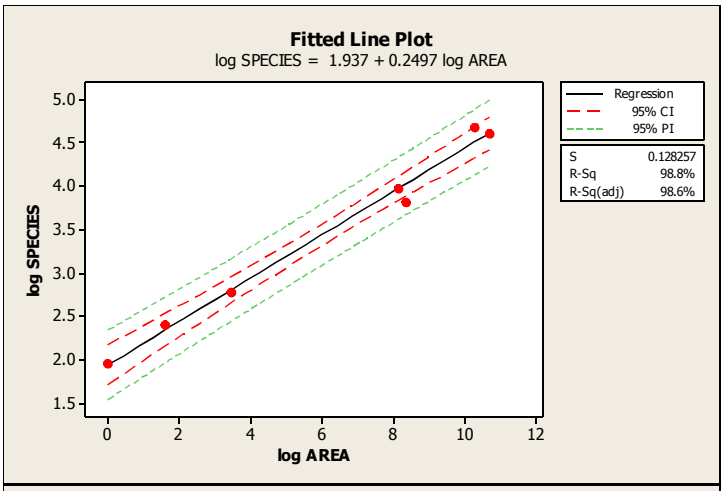
Step 4: To produce a fitted simple linear regression, go to Stat → Regression → Regression (first option for a linear regression); select log(SPECIES) into Response (Y) and log(AREA) into Predictors (X), and click Ok to produce these results.

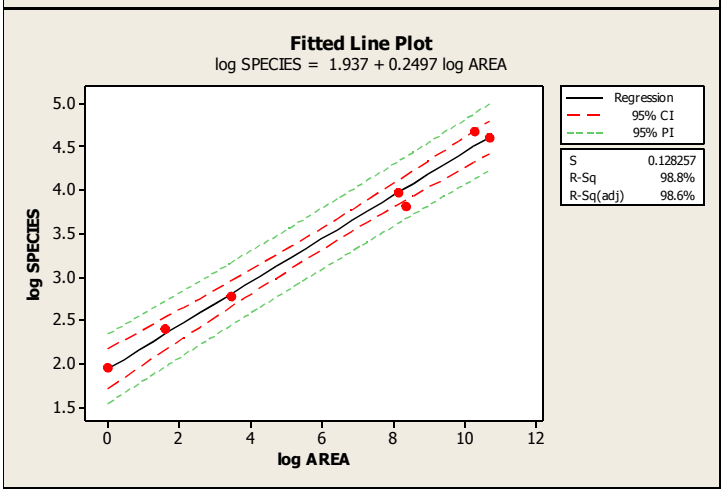


Regression Analysis: LN(Species) versus LN(Area)

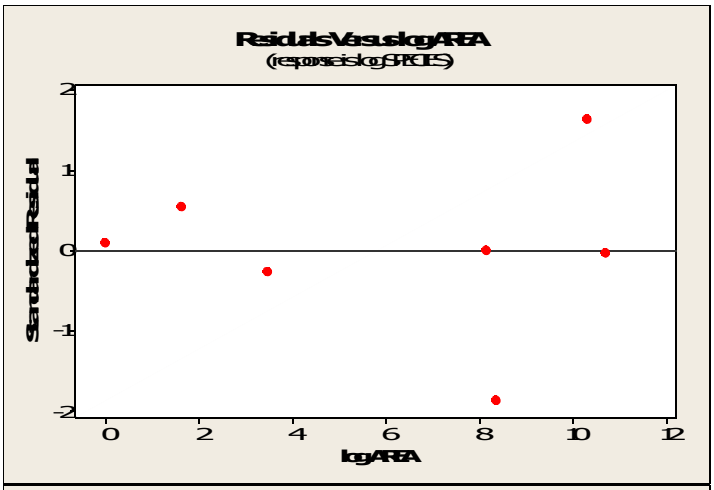
```
Analysis of Variance

Source      DF   Adj SS   Adj MS  F-Value  P-Value
Regression   1  6.99619  6.99619   425.30    0.000
  LN(Area)   1  6.99619  6.99619   425.30    0.000
Error        5  0.08225  0.01645
Total        6  7.07844


Model Summary

       S    R-sq  R-sq(adj)  R-sq(pred)
0.128257  98.84%     98.61%      97.97%


Coefficients

Term        Coef  SE Coef  T-Value  P-Value   VIF
Constant  1.9365   0.0881    21.97    0.000
LN(Area)  0.2497   0.0121    20.62    0.000  1.00


Regression Equation
```

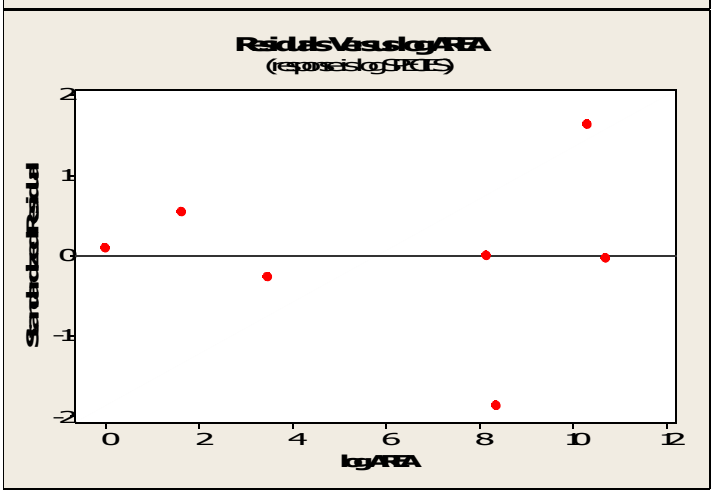
LN(Species) = 1.9365 + 0.2497 LN(Area)

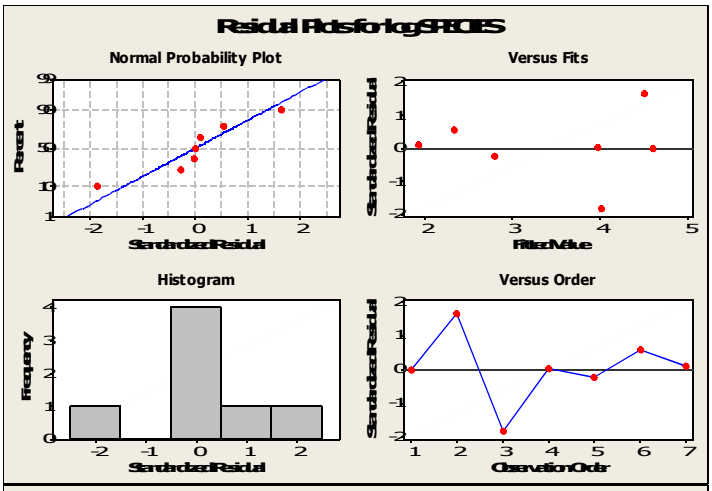

Fitted Line Plot
log SPECIES = 1.937 + 0.2497 log AREA
Regression
95% CI
95% PI
S 0.128257
R-Sq 98.8%
R-Sq(adj) 98.6%
log SPECIES
log AREA
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
0
2
4
6
8
10
12

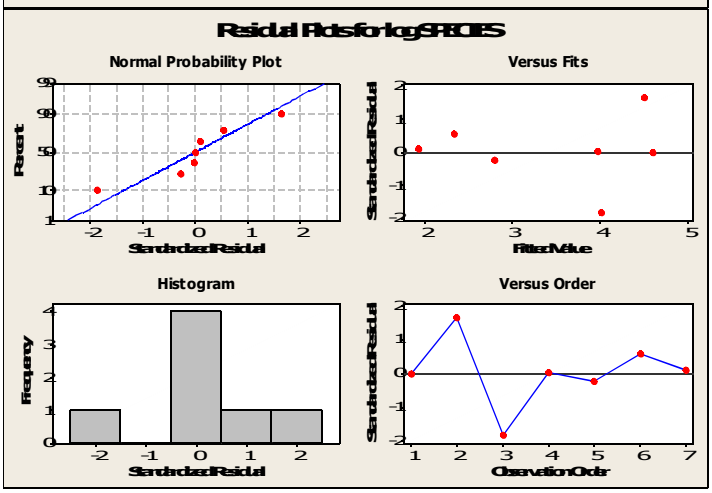

Fitted Line Plot
log SPECIES = 1.937 + 0.2497 log AREA
Regression
95% CI
95% PI
S 0.128257
R-Sq 98.8%
R-Sq(adj) 98.6%
log SPECIES
log AREA
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
0
2
4
6
8
10
12

Standardized Residual Plots





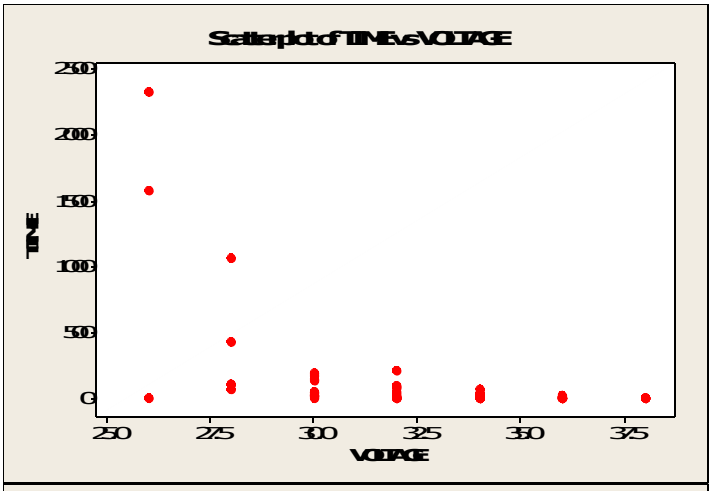

Normal Probability Plot
Versus Fits
Histogram
Versus Order

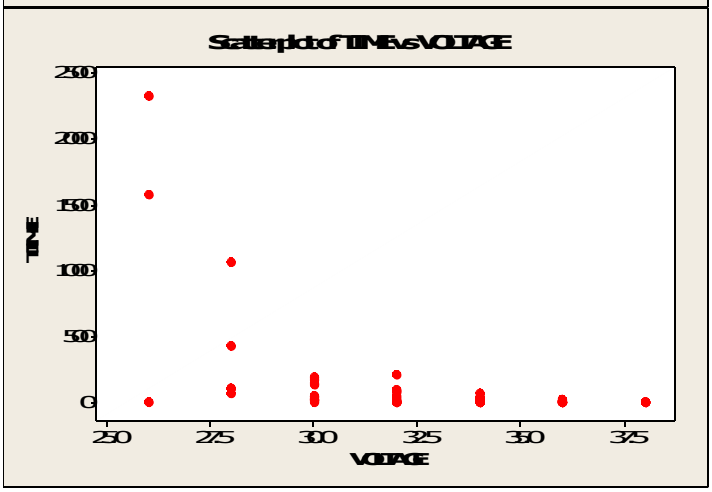

Normal Probability Plot
Versus Fits
Histogram
Versus Order

Case 8.1.2. Breakdown Times for Insulating Fluid under Different Voltages – A Controlled Experiment *R&S p. 209-211*

Step 1: Copy the data into a Minitab Worksheet: use these steps:

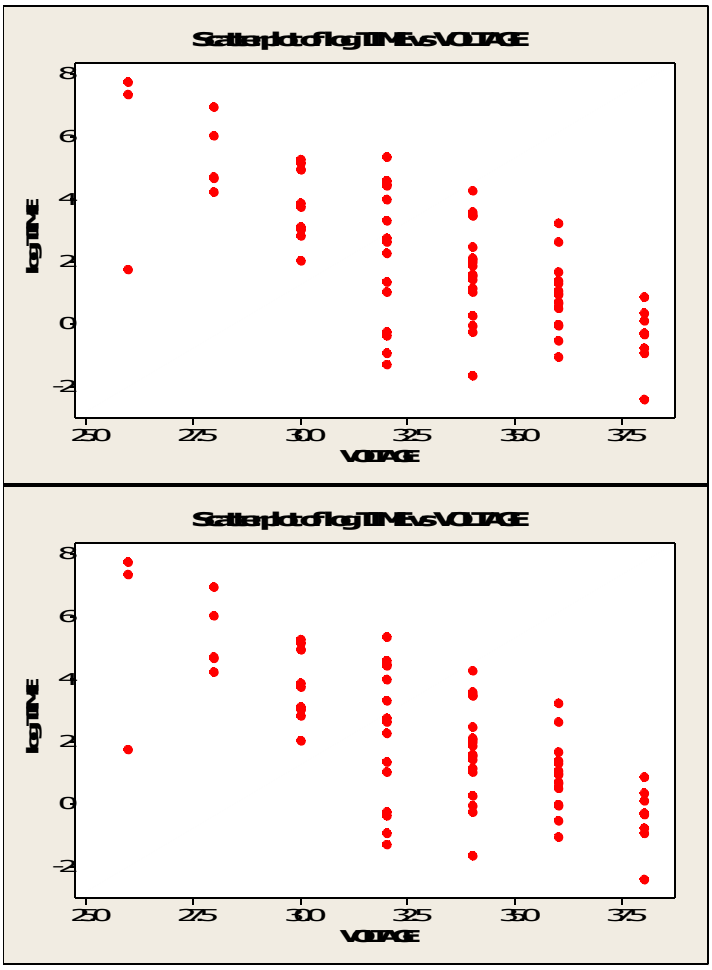
File → Open Worksheet →Browse your local directory and upload the csv file case0802.csv. The data will appear as two columns in Minitab with titles TIME, VOLTAGE and GROUP. See the data displayed by GROUP in ***R&S Display 8.3***.





Step 2: To construct a Scatterplot of log(TIME) versus VOLTAGE, first go to Calc →Calculator and save LN(TIME) into C4. Then, go to Graph and select Simple graph option. Select log TIME into Y

variables and select VOLTAGE into X variables and click OK. This produces this graph, also see ***R&S Display 8.2***.



Step 3: To produce a fitted simple linear regression, go to Stat → Regression → Regression (first option for a linear regression); select log TIME into Response (Y) and VOLTAGE into Predictors (X), and click Ok to produce these results.

Regression Analysis: LN(TIME) versus Voltage

```
Analysis of Variance

Source          DF   Adj SS   Adj MS  F-Value  P-Value
Regression       1  190.151  190.151    78.14    0.000
  Voltage        1  190.151  190.151    78.14    0.000
Error           74  180.075    2.433
  Lack-of-Fit    5    6.326    1.265     0.50    0.773
```

Pure Error	69	173.749	2.518
Total	75	370.226	

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
1.55995	51.36%	50.70%	48.50%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	18.96	1.91	9.92	0.000	
Voltage	-0.5074	0.0574	-8.84	0.000	1.00

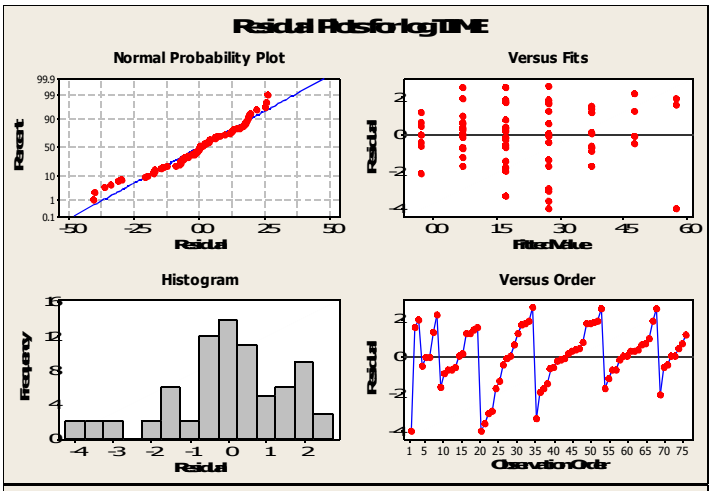
Regression Equation

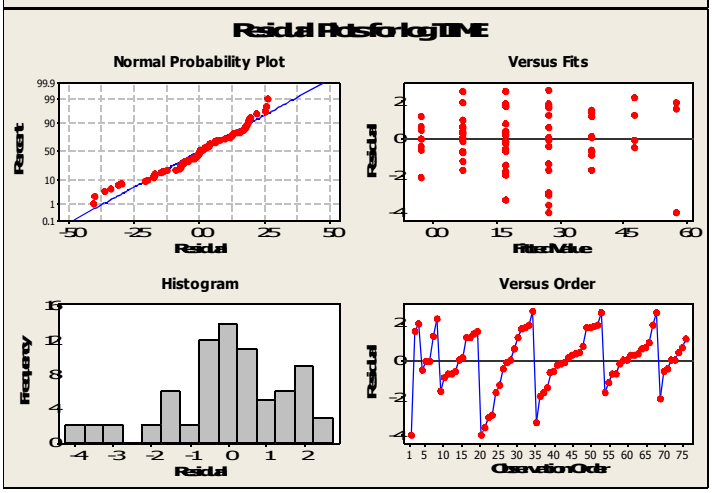
LN(TIME) = 18.96 - 0.5074 Voltage

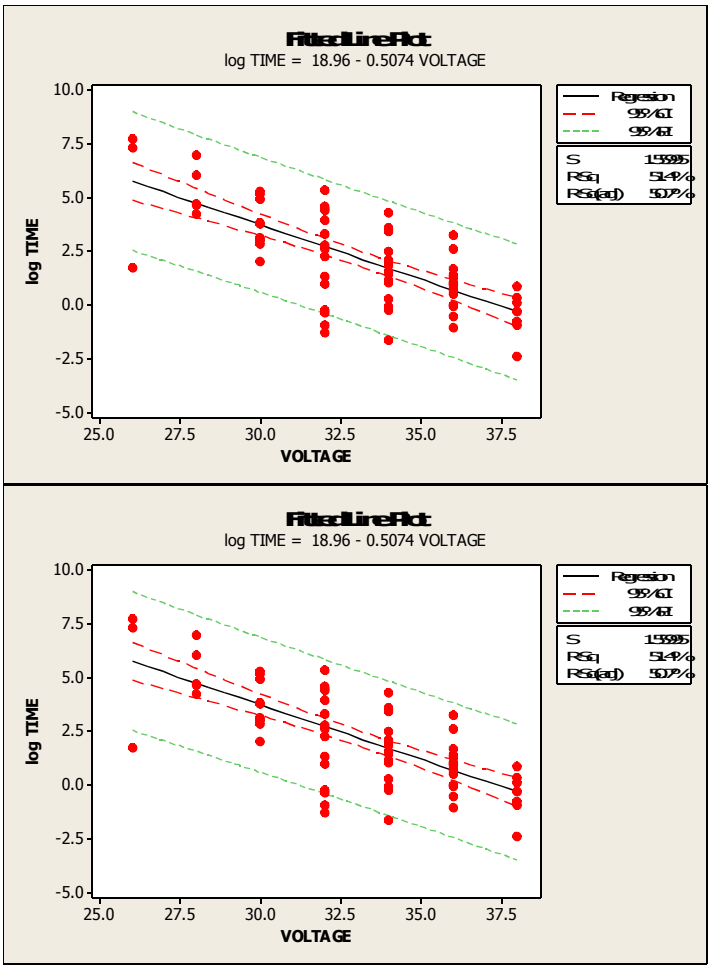
Fits and Diagnostics for Unusual Observations

Obs	LN(TIME)	Fit	Resid	Std Resid		
1	1.756	5.764	-4.008	-2.68	R	X
2	7.365	5.764	1.601	1.07		X
3	7.751	5.764	1.987	1.33		X
20	-1.309	2.720	-4.029	-2.60	R	
21	-0.916	2.720	-3.636	-2.35	R	
35	-1.661	1.705	-3.366	-2.17	R	

R Large residual
X Unusual X


Residual Plots for log TIME
Normal Probability Plot
Versus Fits
Histogram
Versus Order
Percent
Residual
Fitted Value
Frequency
Observation Order


Residual Plots for log TIME
Normal Probability Plot
Versus Fits
Histogram
Versus Order
Percent
Residual
Fitted Value
Frequency
Observation Order


Fitted Line Plot
log TIME = 18.96 - 0.5074 VOLTAGE
Regression
95% CI
95% PI
S 1.59395
R-Sq 51.4%
R-Sq(adj) 50.7%
log TIME
VOLTAGE
10.0
7.5
5.0
2.5
0.0
-2.5
-5.0
25.0
27.5
30.0
32.5
35.0
37.5
Fitted Line Plot
log TIME = 18.96 - 0.5074 VOLTAGE
Regression
95% CI
95% PI
S 1.59395
R-Sq 51.4%
R-Sq(adj) 50.7%
log TIME
VOLTAGE