Chapter 8 (More on Assumptions for the Simple Linear Regression)

Your textbook considers the following assumptions:

Linearity – This is not something I usually consider an explicit assumption, but obviously an investigator believes that the model he is using is correct. A scatter plot can help to determine if this assumption is valid.

Constant variance – The residuals represent variance, and they get pooled into a single estimate of variance. As with ANOVA, we assume that they all represent the same variance or pooling would not produce a good estimate of variance.

Normality – The regression should be normally distributed at each value of Xi, and the pooled residuals should be normally distributed. This assumption is needed for testing and placing confidence intervals.

Independence – Each observation should be unrelated to the other observations.

We will examine a dataset and determine, to the extent possible, if the assumptions are met. The dataset is CASE0802 in the text. The data gives the time in minutes to breakdown of 76 samples of an insulating fluid subjected to different constant voltages.

First, plot the data and try a regression and examine the residual plot.

```
1          ********************************************************;
2          *** Time in minutes to breakdown of 76 samples of an  ***;
3          *** insulating fluid subjected to different constant  ***;
4          *** voltages.                                         ***;
5          ********************************************************;
6
7          dm'log;clear;output;clear';
8          options nodate nocenter nonumber ps=512 ls=99 nolabel;
9          ODS HTML style=minimal rs=none
9        ! body='C:\Geaghan\Current\EXST3201\Fall2005\SAS\Fluids01.html' ;
NOTE: Writing HTML Body file:
    C:\Geaghan\Current\EXST3201\Fall2005\SAS\Fluids01.html
10
11         Title1 'Chapter 8 : Time to breakdown of 76 samples of an insulating fluid';
12         filename input1 'C:\Geaghan\Current\EXST3201\Datasets\ASCII\case0802.csv';
13
14         data Fluids; infile input1 missover DSD dlm="," firstobs=2;
15            input TIME VOLTAGE GROUP $;
16              label Time = 'Breakdown time (minutes)'
17                    Voltage = 'Voltage(kV)';
18            LogTime = log(Time);
19         datalines;
NOTE: The infile INPUT1 is:
      File Name=C:\Geaghan\Current\EXST3201\Datasets\ASCII\case0802.csv,
      RECFM=V,LRECL=256
NOTE: 76 records were read from the infile INPUT1.
      The minimum record length was 16.
      The maximum record length was 31.
NOTE: The data set WORK.FLUIDS has 76 observations and 4 variables.
NOTE: DATA statement used (Total process time):
      real time           0.03 seconds
      cpu time            0.03 seconds
20         run;
```

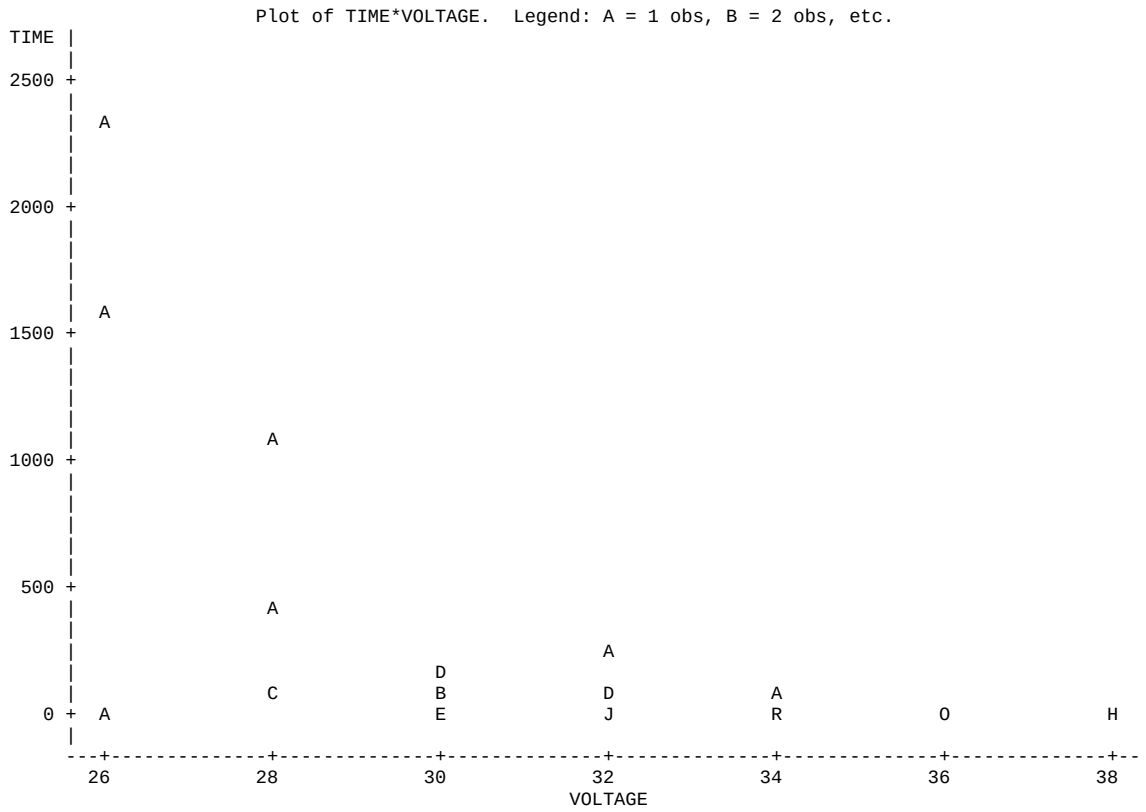
```
21
22          PROC PRINT DATA=Fluids; TITLE2 'Raw data Listing'; RUN;
NOTE: There were 76 observations read from the data set WORK.FLUIDS.
NOTE: The PROCEDURE PRINT printed page 1.
NOTE: PROCEDURE PRINT used (Total process time):
      real time           0.03 seconds
      cpu time            0.03 seconds
```

Chapter 8 : Time to breakdown of 76 samples of an insulating fluid
Raw data Listing

Obs	TIME	VOLTAGE	GROUP	LogTime	Obs	TIME	VOLTAGE	GROUP	LogTime
1	5.79	26	GROUP 1	1.75613	38	1.31	34	GROUP 5	0.27003
2	1579.52	26	GROUP 1	7.36488	39	2.78	34	GROUP 5	1.02245
3	2323.70	26	GROUP 1	7.75092	40	3.16	34	GROUP 5	1.15057
4	68.85	28	GROUP 2	4.23193	41	4.15	34	GROUP 5	1.42311
5	108.29	28	GROUP 2	4.68481	42	4.67	34	GROUP 5	1.54116
6	110.29	28	GROUP 2	4.70311	43	4.85	34	GROUP 5	1.57898
7	426.07	28	GROUP 2	6.05460	44	6.50	34	GROUP 5	1.87180
8	1067.60	28	GROUP 2	6.97317	45	7.35	34	GROUP 5	1.99470
9	7.74	30	GROUP 3	2.04640	46	8.01	34	GROUP 5	2.08069
10	17.05	30	GROUP 3	2.83615	47	8.27	34	GROUP 5	2.11263
11	20.46	30	GROUP 3	3.01847	48	12.06	34	GROUP 5	2.48989
12	21.02	30	GROUP 3	3.04547	49	31.75	34	GROUP 5	3.45789
13	22.66	30	GROUP 3	3.12060	50	32.52	34	GROUP 5	3.48186
14	43.40	30	GROUP 3	3.77046	51	33.91	34	GROUP 5	3.52371
15	47.30	30	GROUP 3	3.85651	52	36.71	34	GROUP 5	3.60305
16	139.07	30	GROUP 3	4.93498	53	72.89	34	GROUP 5	4.28895
17	144.12	30	GROUP 3	4.97065	54	0.35	36	GROUP 6	-1.04982
18	175.88	30	GROUP 3	5.16980	55	0.59	36	GROUP 6	-0.52763
19	194.90	30	GROUP 3	5.27249	56	0.96	36	GROUP 6	-0.04082
20	0.27	32	GROUP 4	-1.30933	57	0.99	36	GROUP 6	-0.01005
21	0.40	32	GROUP 4	-0.91629	58	1.69	36	GROUP 6	0.52473
22	0.69	32	GROUP 4	-0.37106	59	1.97	36	GROUP 6	0.67803
23	0.79	32	GROUP 4	-0.23572	60	2.07	36	GROUP 6	0.72755
24	2.75	32	GROUP 4	1.01160	61	2.58	36	GROUP 6	0.94779
25	3.91	32	GROUP 4	1.36354	62	2.71	36	GROUP 6	0.99695
26	9.88	32	GROUP 4	2.29051	63	2.90	36	GROUP 6	1.06471
27	13.95	32	GROUP 4	2.63548	64	3.67	36	GROUP 6	1.30019
28	15.93	32	GROUP 4	2.76820	65	3.99	36	GROUP 6	1.38379
29	27.80	32	GROUP 4	3.32504	66	5.35	36	GROUP 6	1.67710
30	53.24	32	GROUP 4	3.97481	67	13.77	36	GROUP 6	2.62249
31	82.85	32	GROUP 4	4.41703	68	25.50	36	GROUP 6	3.23868
32	89.29	32	GROUP 4	4.49189	69	0.09	38	GROUP 7	-2.40795
33	100.59	32	GROUP 4	4.61105	70	0.39	38	GROUP 7	-0.94161
34	215.10	32	GROUP 4	5.37110	71	0.47	38	GROUP 7	-0.75502
35	0.19	34	GROUP 5	-1.66073	72	0.73	38	GROUP 7	-0.31471
36	0.78	34	GROUP 5	-0.24846	73	0.74	38	GROUP 7	-0.30111
37	0.96	34	GROUP 5	-0.04082	74	1.13	38	GROUP 7	0.12222
					75	1.40	38	GROUP 7	0.33647
					76	2.38	38	GROUP 7	0.86710

```
23
24          options ps=52 ls=111;
25          proc plot data=Fluids;
26             plot Time * VOLTAGE; TITLE2 'Plot of the raw data'; run;
27
28          Title2 'Initial fit of the raw data';
29          options ps=512 ls=111;
NOTE: There were 76 observations read from the data set WORK.FLUIDS.
NOTE: The PROCEDURE PLOT printed page 2.
NOTE: PROCEDURE PLOT used (Total process time):
      real time           0.00 seconds
      cpu time            0.00 seconds
```

Chapter 8 : Time to breakdown of 76 samples of an insulating fluid
Plot of the raw data




```
30          PROC REG DATA=Fluids lineprinter;  ID group;
31             TITLE3 'Simple Regression with REG';
32             MODEL Time = VOLTAGE;
33             output out=next r=resid p=YHat;
34          RUN;
34         !      OPTIONS PS=45;
35          TITLE3 'Plot of residuals';
NOTE: The data set WORK.NEXT has 76 observations and 6 variables.
NOTE: The PROCEDURE REG printed page 3.
NOTE: PROCEDURE REG used (Total process time):
      real time           0.03 seconds
      cpu time            0.03 seconds
```

Chapter 8 : Time to breakdown of 76 samples of an insulating fluid
Initial fit of the raw data
Simple Regression with REG

The REG Procedure
Model: MODEL1
Dependent Variable: TIME

Number of Observations Read	76
Number of Observations Used	76

Analysis of Variance

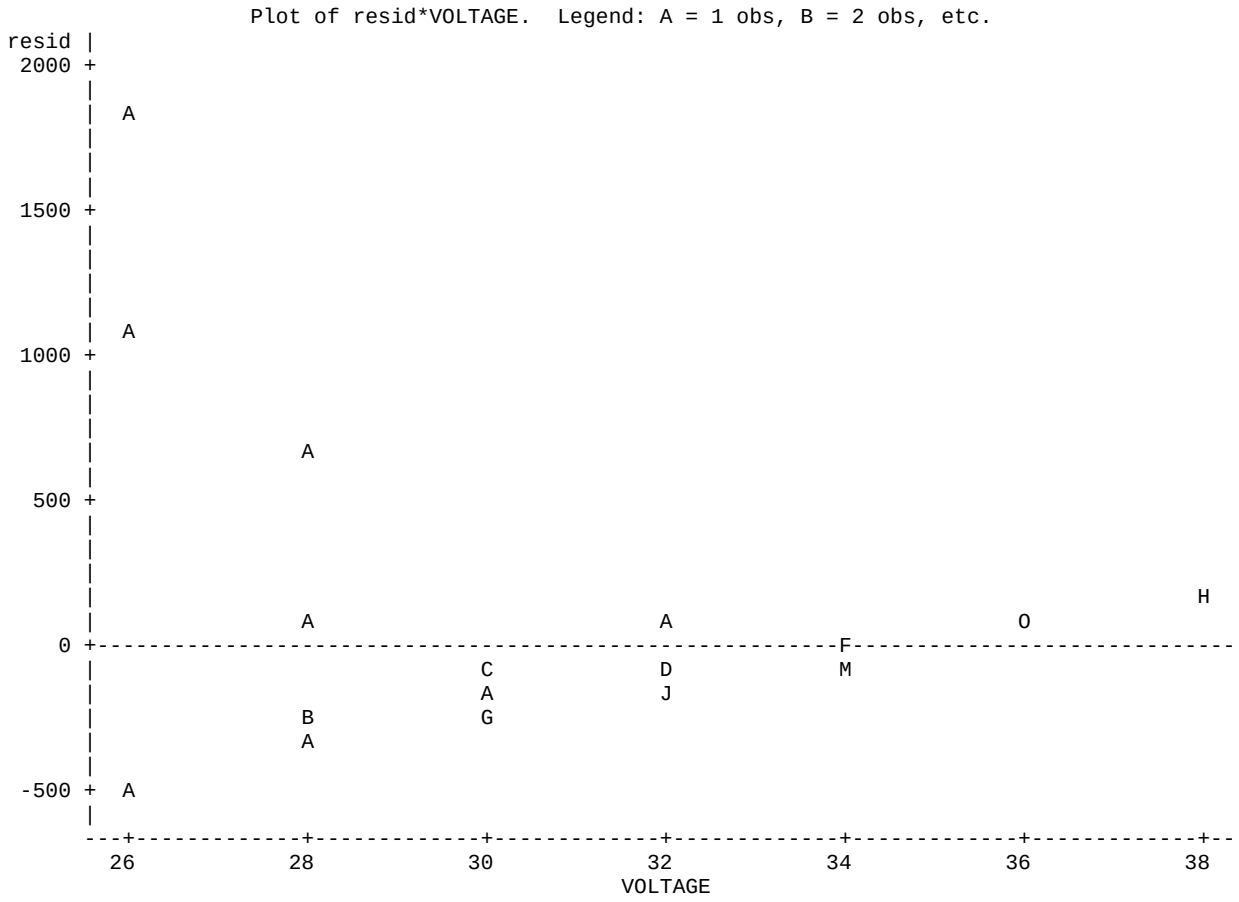
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2150408	2150408	24.27	<.0001
Error	74	6557345	88613		
Corrected Total	75	8707754			

Root MSE	297.67898	R-Square	0.2470
Dependent Mean	98.55776	Adj R-Sq	0.2368
Coeff Var	302.03504		

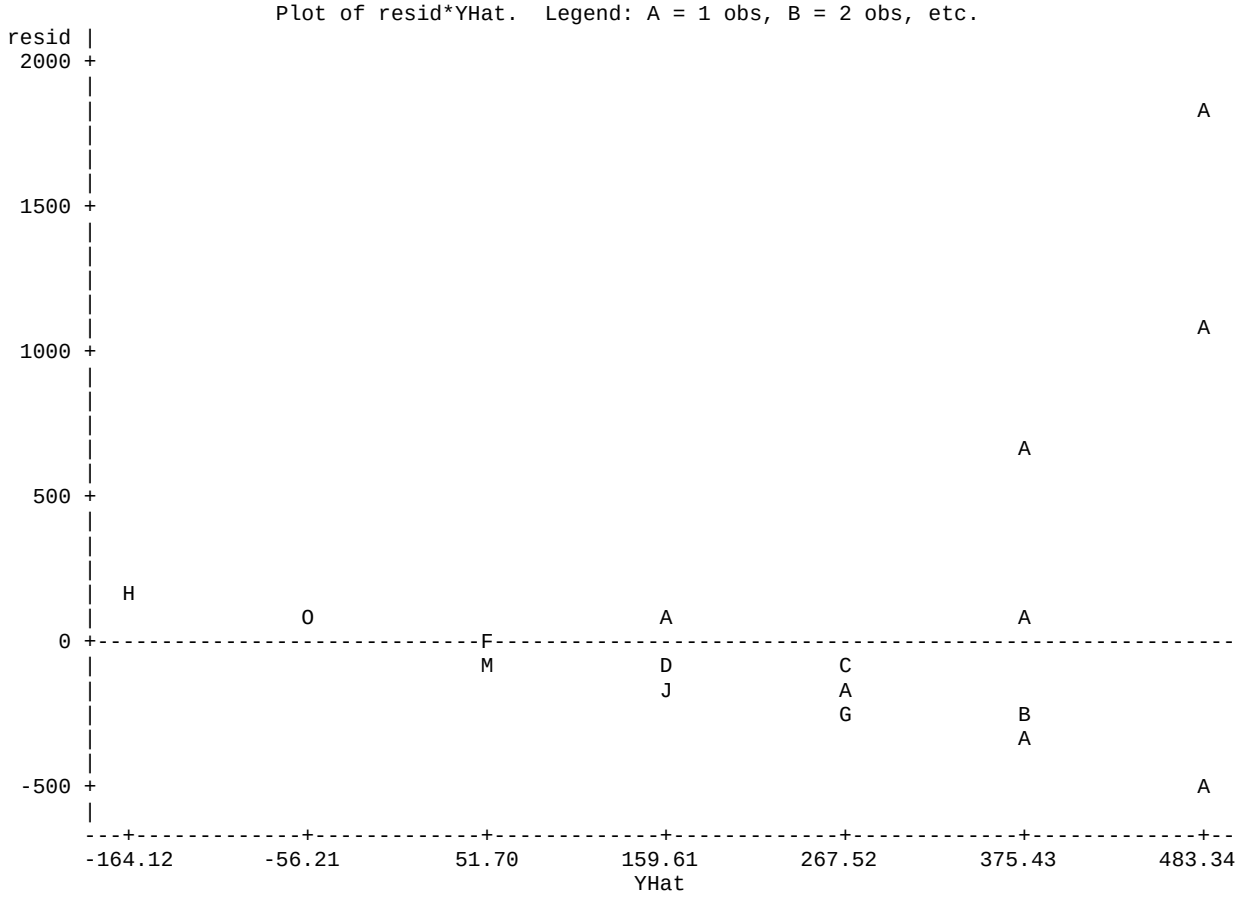
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	1886.16945	364.48121	5.17	<.0001
VOLTAGE	1	-53.95492	10.95264	-4.93	<.0001

```
36         Proc plot; PLOT resid*VOLTAGE / vref=0;
NOTE: There were 76 observations read from the data set WORK.NEXT.
NOTE: The PROCEDURE PLOT printed page 4.
NOTE: PROCEDURE PLOT used (Total process time):
      real time           0.00 seconds
      cpu time            0.00 seconds
37         Proc plot; PLOT resid*YHat / vref=0;
38         RUN;
NOTE: There were 76 observations read from the data set WORK.NEXT.
NOTE: The PROCEDURE PLOT printed page 5.
NOTE: PROCEDURE PLOT used (Total process time):
      real time           0.00 seconds
      cpu time            0.00 seconds
```



```
Chapter 8 : Time to breakdown of 76 samples of an insulating fluid
Initial fit of the raw data
Plot of residuals
```




```
39          PROC UNIVARIATE DATA=NEXT NORMAL PLOT; VAR resid;
40          RUN;
NOTE: The PROCEDURE UNIVARIATE printed pages 6-8.
NOTE: PROCEDURE UNIVARIATE used (Total process time):
      real time           0.01 seconds
      cpu time            0.01 seconds
41


Chapter 8 : Time to breakdown of 76 samples of an insulating fluid
Initial fit of the raw data
Plot of residuals

The UNIVARIATE Procedure
Variable:  resid

Moments
N                          76    Sum Weights                76
Mean                        0    Sum Observations            0
Std Deviation      295.687792    Variance            87431.2704
Skewness           4.02671557    Kurtosis            21.9169395
Uncorrected SS     6557345.28    Corrected SS        6557345.28
Coeff Variation             .    Std Error Mean      33.9177159
```

Basic Statistical Measures

Location		Variability	
Mean	0.0000	Std Deviation	295.68779
Median	-46.0272	Variance	87431
Mode	.	Range	2318
		Interquartile Range	203.20468

Tests for Location: Mu0=0

Test	-Statistic-		-----p Value------	
Student's t	t	0	Pr > |t|	1.0000
Sign	M	-9	Pr >= |M|	0.0505
Signed Rank	S	-283	Pr >= |S|	0.1440

Tests for Normality

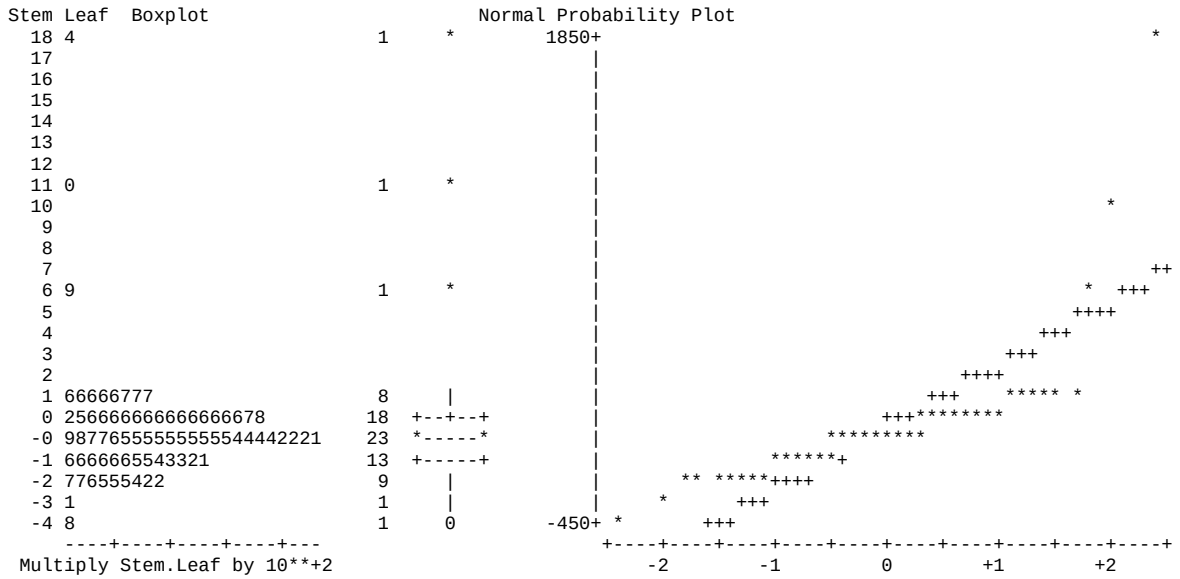
Test	--Statistic---		-----p Value------	
Shapiro-Wilk	W	0.615009	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.248565	Pr > D	<0.0100
Cramer-von Mises	W-Sq	1.204042	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	6.998223	Pr > A-Sq	<0.0050

Quantiles (Definition 5)

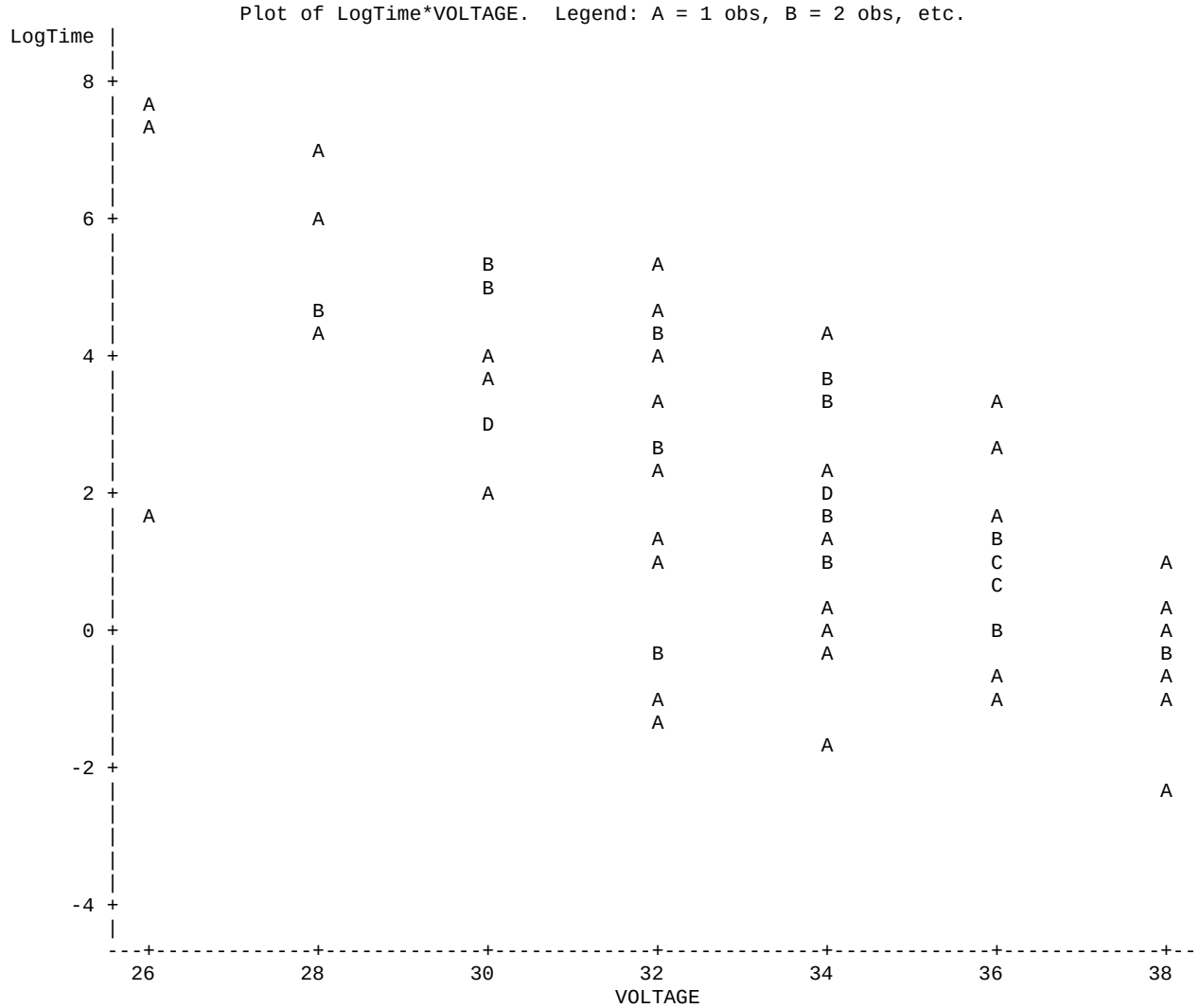
Quantile	Estimate
100% Max	1840.3584
99%	1840.3584
95%	166.4975
90%	164.8475
75% Q3	58.5327
50% Median	-46.0272
25% Q1	-144.6720
10%	-246.5019
5%	-265.1417
1%	-477.5515
0% Min	-477.5515

Extreme Observations

------Lowest-----		------Highest-----	
Value	Obs	Value	Obs
-477.552	1	165.518	75
-306.582	4	166.498	76
-267.142	5	692.168	8
-265.142	6	1096.178	2
-259.782	9	1840.358	3

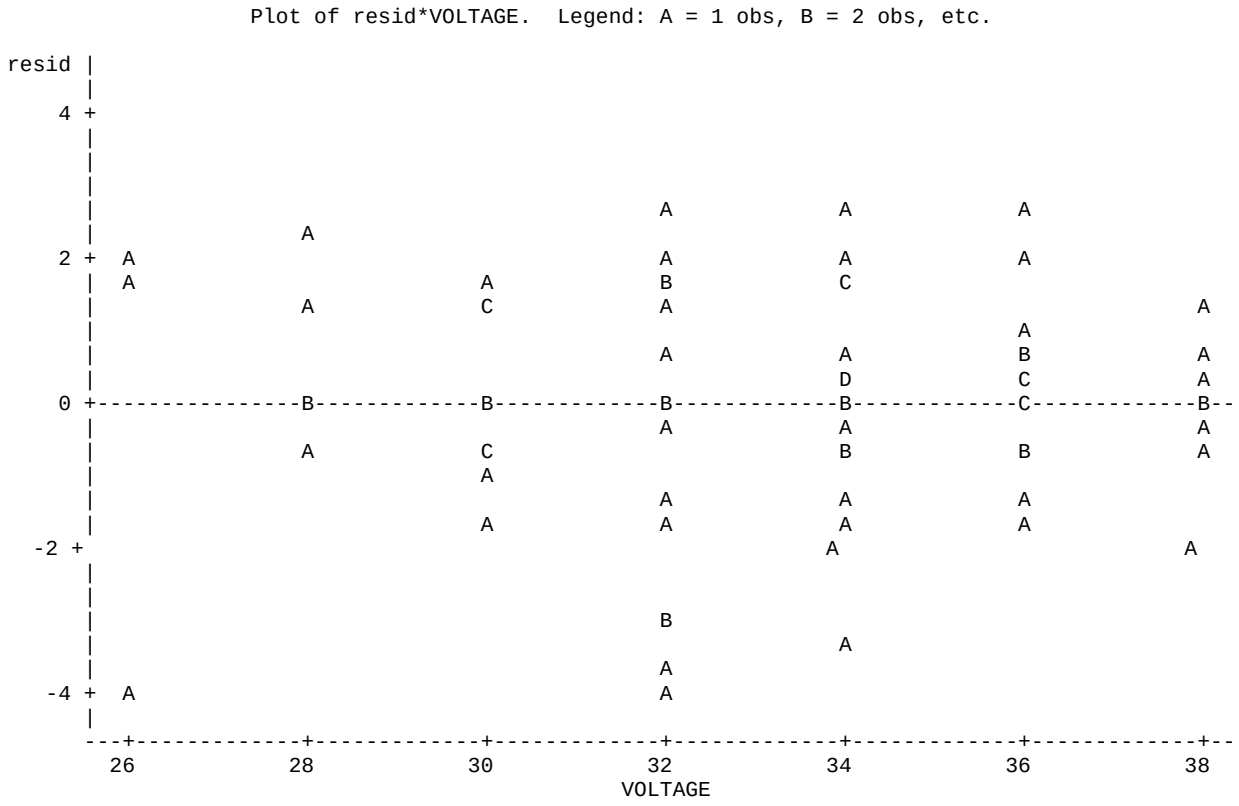


```
42         Title2 'Examination of the log transformed data';
43         options ps=52 ls=111;
44         proc plot data=Fluids;
45            plot LogTime * VOLTAGE; TITLE3 'Plot of the log transformed data';
46
47         options ps=512 ls=111;
NOTE: There were 76 observations read from the data set WORK.FLUIDS.
NOTE: The PROCEDURE PLOT printed page 9.
NOTE: PROCEDURE PLOT used (Total process time):
      real time           0.00 seconds
      cpu time            0.00 seconds


Chapter 8 : Time to breakdown of 76 samples of an insulating fluid
Examination of the log transformed data
Plot of the log transformed data
```

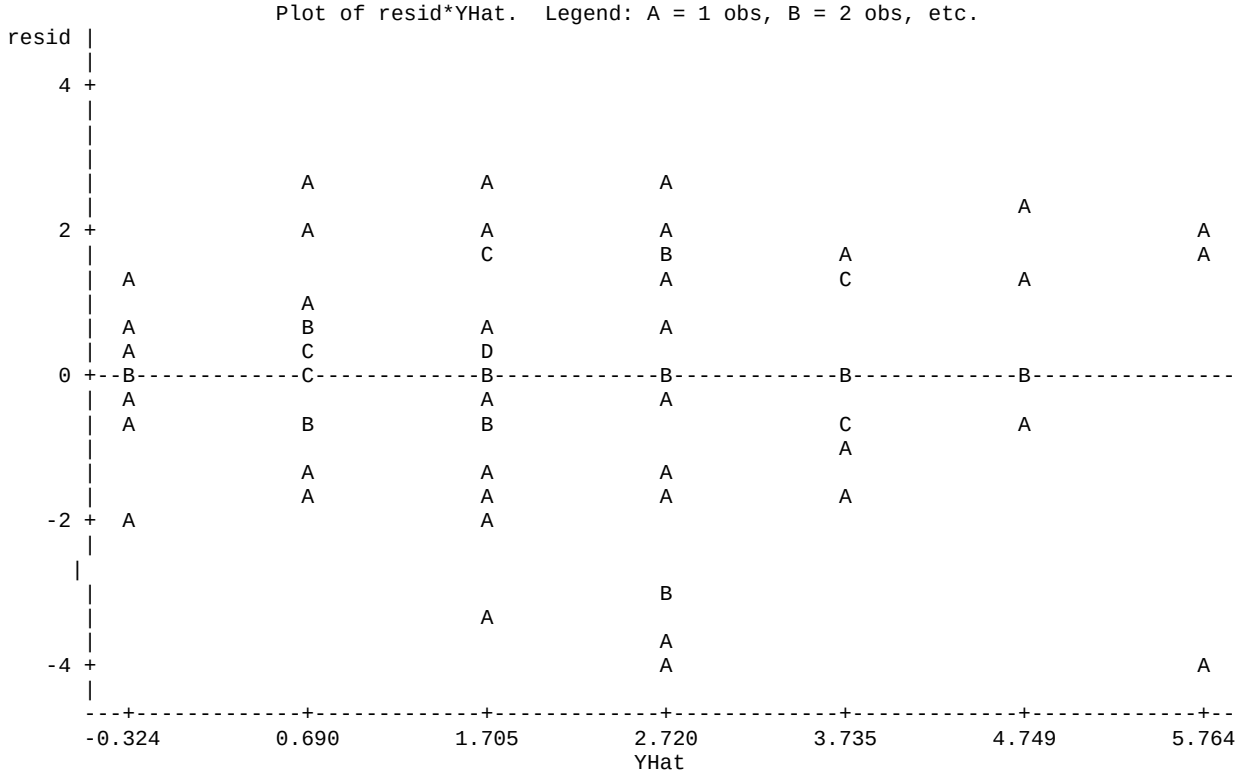



```
48         PROC REG DATA=Fluids lineprinter;  ID group;
49            TITLE3 'Transformed regression with REG';
50            MODEL LogTime = VOLTAGE;
51            output out=next r=resid p=YHat;
52         RUN;
```

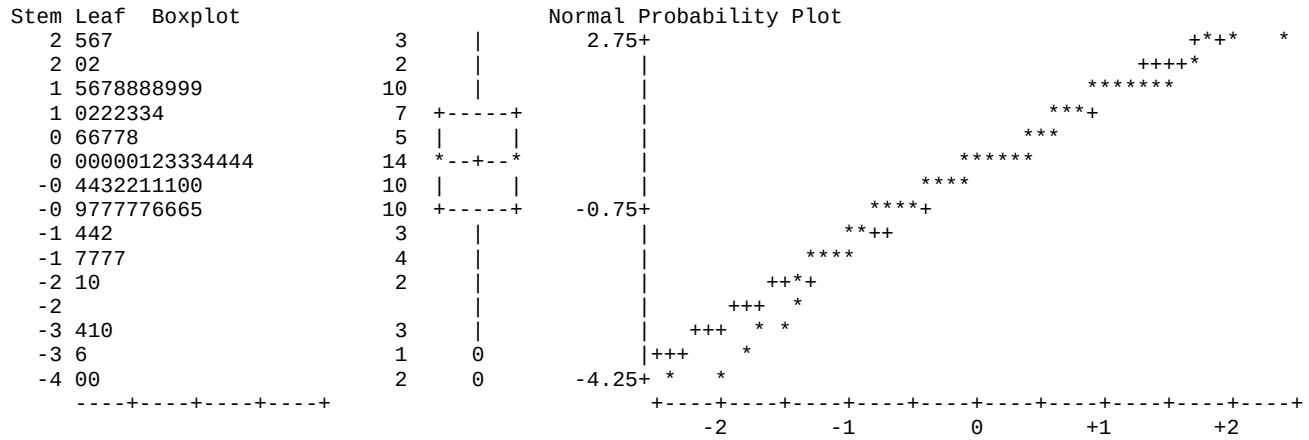
```
Chapter 8 : Time to breakdown of 76 samples of an insulating fluid
Examination of the log transformed data
Transformed regression with REG

The REG Procedure
Model: MODEL1
Dependent Variable: LogTime

Number of Observations Read          76
Number of Observations Used          76

Analysis of Variance
                                  Sum of           Mean
Source                     DF        Squares         Square    F Value    Pr > F
Model                       1      190.15149      190.15149      78.14    <.0001
Error                      74      180.07484        2.43344
Corrected Total            75      370.22633

Root MSE              1.55995    R-Square     0.5136
Dependent Mean        2.14566    Adj R-Sq     0.5070
Coeff Var            72.70267

Parameter Estimates
                     Parameter       Standard
Variable     DF       Estimate          Error    t Value    Pr > |t|
Intercept     1       18.95546        1.91002       9.92      <.0001
VOLTAGE       1       -0.50736        0.05740      -8.84      <.0001



52       !      OPTIONS PS=45;
53         TITLE3 'Plot of residuals';
NOTE: The data set WORK.NEXT has 76 observations and 6 variables.
NOTE: The PROCEDURE REG printed page 10.
NOTE: PROCEDURE REG used (Total process time):
      real time           0.04 seconds
      cpu time            0.03 seconds
54         Proc plot; PLOT resid*VOLTAGE / vref=0;
NOTE: There were 76 observations read from the data set WORK.NEXT.
NOTE: The PROCEDURE PLOT printed page 11.
NOTE: PROCEDURE PLOT used (Total process time):
      real time           0.01 seconds
      cpu time            0.01 seconds
55         Proc plot; PLOT resid*YHat / vref=0;
56         RUN;
NOTE: There were 76 observations read from the data set WORK.NEXT.
NOTE: The PROCEDURE PLOT printed page 12.
NOTE: PROCEDURE PLOT used (Total process time):
      real time           0.00 seconds
      cpu time            0.00 seconds


Chapter 8 : Time to breakdown of 76 samples of an insulating fluid
Examination of the log transformed data
Plot of residuals
```



Chaper 8 : Time to breakdown of 76 samples of an insulating fluid
Examination of the log transformed data
Plot of residuals



```
57         PROC UNIVARIATE DATA=NEXT NORMAL PLOT; VAR resid;
58         RUN;
NOTE: The PROCEDURE UNIVARIATE printed pages 13-15.
NOTE: PROCEDURE UNIVARIATE used (Total process time):
      real time           0.01 seconds
      cpu time            0.01 seconds


Chapter 8 : Time to breakdown of 76 samples of an insulating fluid
Examination of the log transformed data
Plot of residuals

The UNIVARIATE Procedure
Variable:  resid

Moments
N                          76    Sum Weights                76
Mean                        0    Sum Observations            0
Std Deviation      1.54951535    Variance           2.40099782
Skewness           -0.6488654    Kurtosis           0.30321576
Uncorrected SS     180.074836    Corrected SS       180.074836
Coeff Variation             .    Std Error Mean      0.1777416

Basic Statistical Measures
    Location                     Variability
Mean     0.000000     Std Deviation            1.54952
Median   0.036589     Variance                 2.40100
Mode      .           Range                    6.68044
                      Interquartile Range      1.91300

Tests for Location: Mu0=0
Test           -Statistic-    -----p Value------
Student's t    t         0    Pr > |t|    1.0000
Sign           M         3    Pr >= |M|   0.5666
Signed Rank    S        93    Pr >= |S|   0.6333

Tests for Normality
Test                   --Statistic---    -----p Value------
Shapiro-Wilk           W     0.956759    Pr < W      0.0112
Kolmogorov-Smirnov     D      0.10799    Pr > D      0.0267
Cramer-von Mises       W-Sq   0.11647    Pr > W-Sq   0.0699
Anderson-Darling       A-Sq  0.834933    Pr > A-Sq   0.0311

Quantiles (Definition 5)                  50% Median     0.0365886
Quantile        Estimate                  25% Q1        -0.6947034
100% Max       2.6513227                  10%           -1.9535119
99%            2.6513227                  5%            -3.3657817
95%            2.2239282                  1%            -4.0291137
90%            1.8912724                  0% Min        -4.0291137
75% Q3         1.2183016

Extreme Observations
------Lowest-----        -----Highest-----
   Value      Obs           Value      Obs
-4.02911       20         1.98695        3
-4.00784        1         2.22393        8
-3.63607       21         2.54836       68
-3.36578       35         2.58390       53
-3.09084       22         2.65132       34
```




```
60         Title2 'Testing lack of fit';
61         options ps=512 ls=111;
62         PROC GLM DATA=Fluids;  class group;
63            TITLE3 'ANOVA of the Untransformed data with GLM';
64            MODEL Time = group;
65         RUN;
67         options ps=512 ls=111;
NOTE: The PROCEDURE GLM printed pages 16-17.
NOTE: PROCEDURE GLM used (Total process time):
      real time           0.03 seconds
      cpu time            0.03 seconds


Chapter 8 : Time to breakdown of 76 samples of an insulating fluid
Testing lack of fit
ANOVA of the Untransformed data with GLM

The GLM Procedure

Class Level Information
Class         Levels    Values
GROUP              7    GROUP 1 GROUP 2 GROUP 3 GROUP 4 GROUP 5 GROUP 6 GROUP 7
Number of Observations Read          76
Number of Observations Used          76

Dependent Variable: TIME
                                        Sum of
Source                     DF          Squares     Mean Square    F Value    Pr > F
Model                       6     5082509.892      847084.982      16.12    <.0001
Error                      69     3625243.641       52539.763
Corrected Total            75     8707753.533

R-Square     Coeff Var      Root MSE     TIME Mean
0.583676      232.5697      229.2155      98.55776

Source                     DF       Type I SS     Mean Square    F Value    Pr > F
GROUP                       6     5082509.892      847084.982      16.12    <.0001

Source                     DF     Type III SS     Mean Square    F Value    Pr > F
GROUP                       6     5082509.892      847084.982      16.12    <.0001
```

Model	d.f. Error	SS Error	MS	F	Pr>F
Reduced	74	6557345.000			
Full	69	3625243.641			
Difference	5	2932101.359	586420.27180	11.161	0.000000066745
Full	69	3625243.641	52539.76291		

```
68         PROC GLM DATA=Fluids;  class group;
69            TITLE3 'Extra SS of fitting means after the slope';
70            MODEL Time = voltage group;
71         RUN;
NOTE: The PROCEDURE GLM printed pages 18-19.
NOTE: PROCEDURE GLM used (Total process time):
      real time           0.04 seconds
      cpu time            0.04 seconds



Chapter 8 : Time to breakdown of 76 samples of an insulating fluid
Testing lack of fit
Extra SS of fitting means after the slope

The GLM Procedure

Class Level Information
Class         Levels    Values
GROUP              7    GROUP 1 GROUP 2 GROUP 3 GROUP 4 GROUP 5 GROUP 6 GROUP 7

Number of Observations Read          76
Number of Observations Used          76


Dependent Variable: TIME
                                     Sum of
Source                       DF         Squares     Mean Square    F Value    Pr > F
Model                         6     5082509.892      847084.982      16.12    <.0001
Error                        69     3625243.641       52539.763
Corrected Total              75     8707753.533

R-Square     Coeff Var      Root MSE     TIME Mean
0.583676      232.5697      229.2155      98.55776

Source                       DF       Type I SS     Mean Square    F Value    Pr > F
VOLTAGE                       1     2150408.256     2150408.256      40.93    <.0001
GROUP                         5     2932101.636      586420.327      11.16    <.0001

Source                       DF     Type III SS     Mean Square    F Value    Pr > F
VOLTAGE                       0           0.000               .          .         .
GROUP                         5     2932101.636      586420.327      11.16    <.0001
```