BREAKDOWN TIMES

13. Brief Version of the Case Study

13.1 Problem Formulation

In an industrial laboratory, an experiment was conducted to obtain the times (in minutes) to breakdown of 76 samples of an insulating fluid subjected to various constant elevated test voltages. At each test voltage, a number of times to breakdown (in minutes) were observed. Elevated test voltages were employed to save time in collecting the breakdown data. In applications, the voltages are so low that the average time to a breakdown runs millions of years. The experiment was carried out at seven different voltage levels, spaced two kilovolts (kV) apart from 26 to 38 kV.

The problem was studied by W.B. Nelson and the results were published in "Graphical Analysis of Accelerated Life Test Data", *IEEE Transactions in Reliability*, R21, No.1, pages 2-11, 1972.

The data from the experiment are available in the SPSS file break.sav located in the STAT 252 directory on the FTP server.

The following is a description of the variables in the data file:

<u>**Column**</u>	<u>**Name of Variable**</u>	<u>**Description of Variable**</u>
1	VOLTAGE	Voltage Level (in kV)
3	CODE	1 when VOLTAGE = 26
2	TIME	Time to breakdown (in
	minutes)	
		2 when VOLTAGE = 28
		3 when VOLTAGE = 30
		4 when VOLTAGE = 32
		5 when VOLTAGE = 34
		6 when VOLTAGE = 36
		7 when VOLTAGE = 38

We will use SPSS to answer the following questions using the data:

1. What is the distribution of time to breakdown at a constant voltage?

2. Is it possible to develop a reliable model describing the relationship between voltage level and breakdown time?

13.2 Study Design

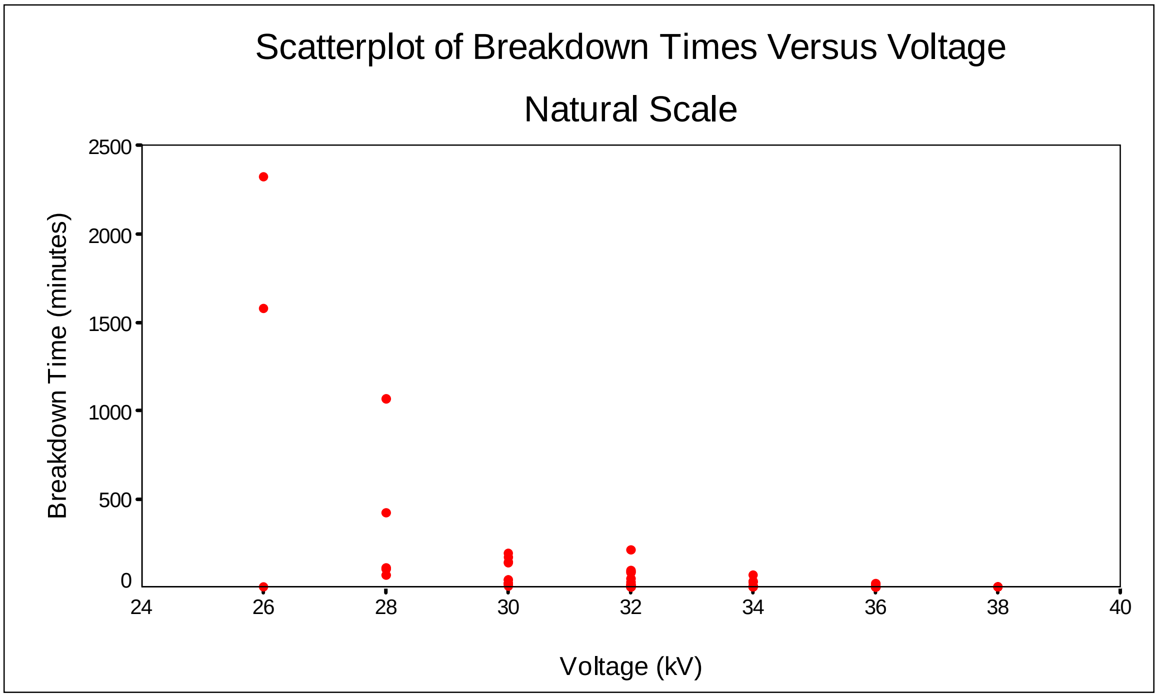
The 76 batches of the same insulating fluid constitute the experimental units in the experiment. Let us assume that the batches were randomly assigned to the seven voltage levels of 26, 28, 30, 32, 34, 36, 38 kV. The numbers of batches assigned to the different voltage levels are 3, 5, 11, 15, 19, 15, and 8, respectively. The measured responses were the times, in minutes, until breakdown.

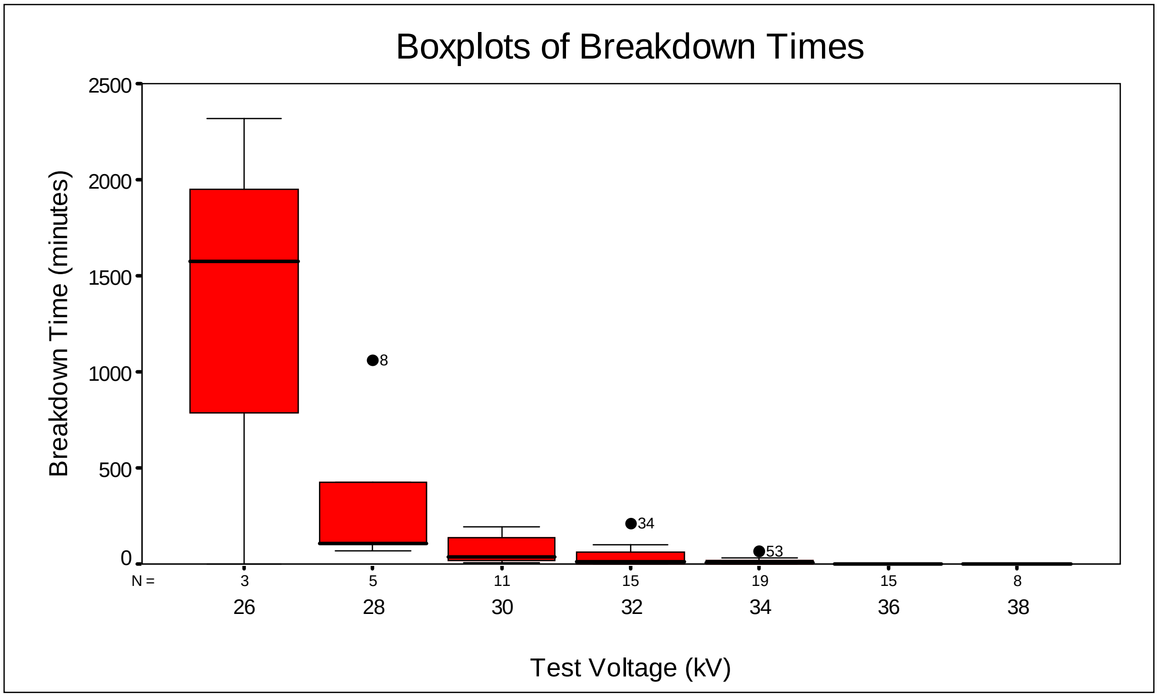
Under the assumption that the fluid batches were randomly assigned randomly to the seven voltage levels, the experiment is an example of a randomized experiment. Thus, causal inferences can be drawn from the data.

Indeed, the laboratory setting for this experiment allowed the experimenter to hold all factors constant except the voltage level. Therefore, if any significant differences among the mean breakdown times can be detected, it seems reasonable to infer that they have been caused by different voltage levels.

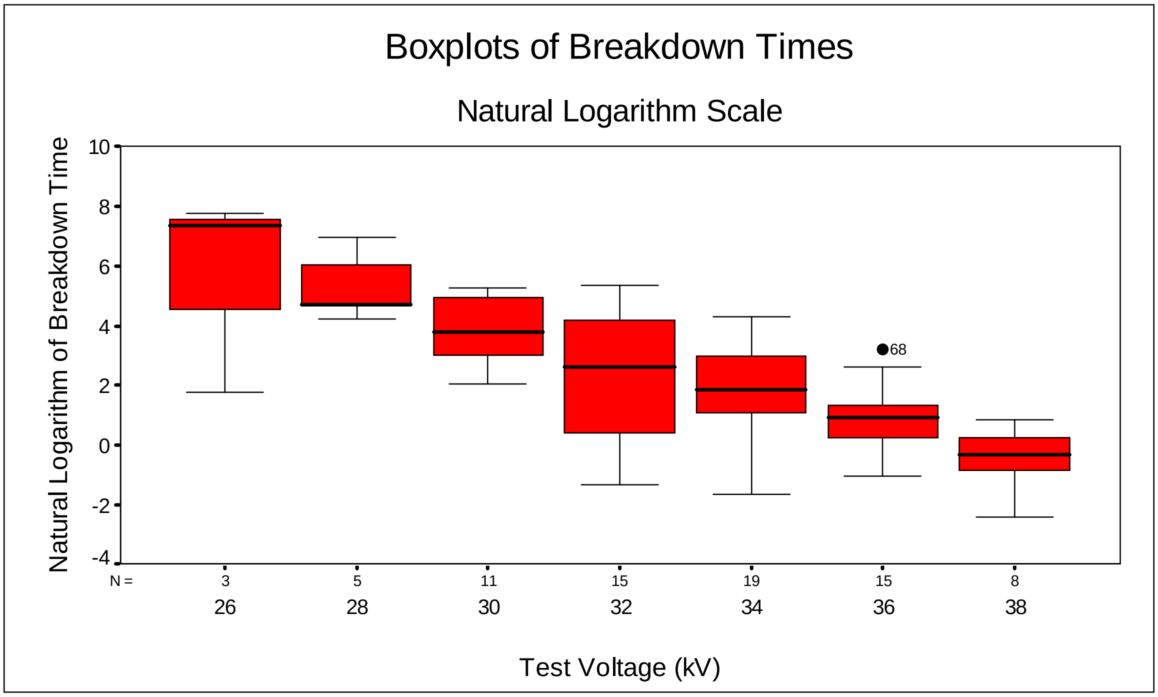
13.3 Displaying and Describing Data

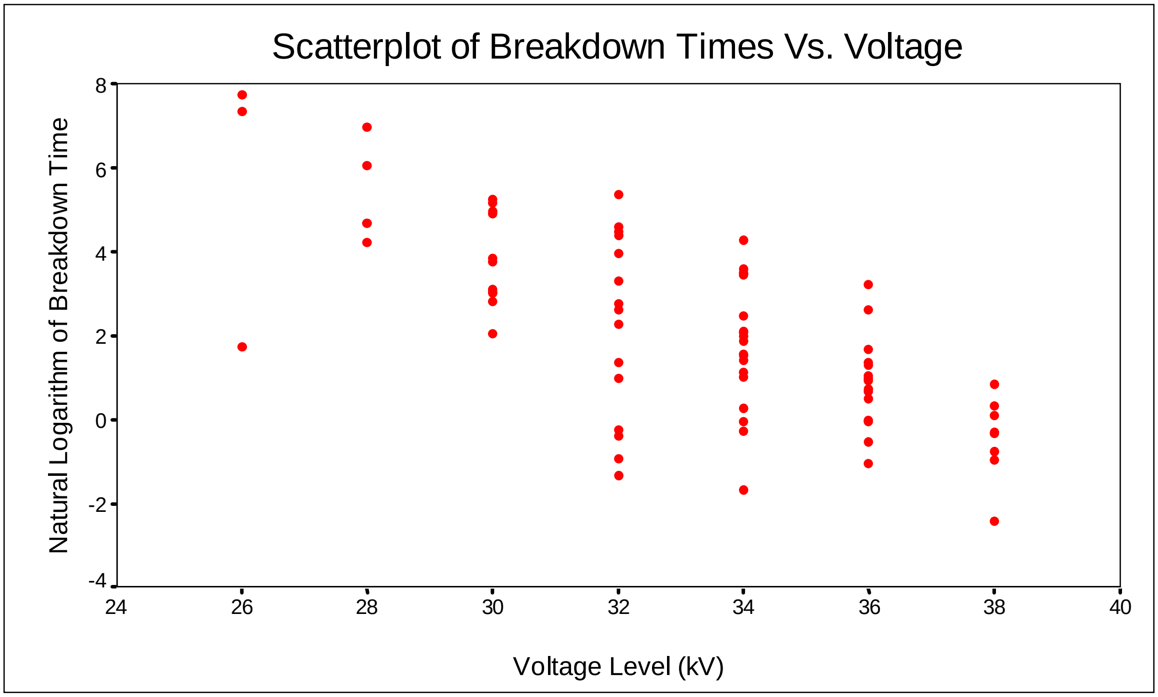
The scatterplot of the breakdown times versus voltage and side-by-side boxplots of the breakdown times for the seven experimental groups on the original scale of measurement reveal non-constant spread, non-linear pattern (exponential), skewness in the data, and outliers.





Both plots suggest transformation of the response variable. The natural logarithm transformation applied to the breakdown time removed skewness in the data, made the spreads approximately equal, and revealed a linear relationship between the log-breakdown times and voltage.





The further analysis will be conducted with log-breakdown time as the response variable and voltage as the explanatory variable.

13.4 Simple Linear Regression

As the above scatterplot displays a linear relationship between the log-breakdown time and voltage, the following simple regression model is suitable:

$$Ln(Time\,|\,Voltage) = \beta_0 + \beta_1 * Voltage + ERROR.$$

Here *Time* denotes the time until breakdown and *Voltage* is the voltage level with possible values at 26, 28, …,38. The random variable *ERROR* is assumed to follow a normal distribution with the mean of zero and an unknown standard deviation σ. The standard deviation is constant at all levels of *Voltage*. The variable *ERROR* follows a normal distribution at each voltage level.

The simple linear regression model can be stated equivalently as follows:

$$\mu\{Ln(Time\,|\,Voltage)\} = \beta_0 + \beta_1 * Voltage.$$

The above model with *Voltage* as a predictor is useful only if the slope β_1 is different from zero. The hypothesis that $\beta_1 = 0$ (the model is useful) can be tested using either t or F tests. The F-statistic is the square of the t-statistic and the corresponding p-values of the two tests are identical.

A quick glance at the scatterplot of the log-transformed breakdown times versus voltage shows that the data provide strong evidence of the utility of the linear regression model.

The SPSS simple linear regression model output for the problem has the following form:

```
                    LINEAR REGRESSION

Multiple R                  .71667
R Square                    .51361
Adjusted R Square           .50704
Standard Error             1.55995

                        Analysis of Variance

                    DF    Sum of Squares       Mean Square

Regression          1      190.15149            190.15149
Residual            74     180.07484            2.43344

F =      78.14090     Signif F =  .0000
```

According to the output, the value of the correlation coefficient between breakdown time and voltage is -0.71667. The value of R^2 (0.5136) says that 51.36% of the variation in the log-breakdown times was explained by the linear regression on voltage. The remaining variation was due to some other variables.

We analyze the ANOVA table associated with the simple regression. The sum of squares due to the regression model is reported as 190.15149, and the sum of squares due to error (residual sum of squares) is 180.07484. The residual mean square is an estimate of the variance σ^2 and is equal to 2.4334.

The value of the F statistic is equal to 78.1409 with the corresponding p-value of 0 provides very strong evidence of the utility of the model. This is what we expected by examining the scatterplot of the log-transformed responses.

Now we analyze the part of the output providing the estimates of the regression parameters.

```
--------------------- Variables in the Equation -----------------------------------

Variable          B           SE B     95% Confidence Interval B      Beta

VOLTAGE     -.507365    .057396    -.621729       -.393001        -.716665
(Constant)  18.955459   1.910019   15.149663    22.761254

Variable          T           Sig T

VOLTAGE       -8.840          .0000
(Constant)     9.924          .0000
```

According to the output, the estimated regression line of the breakdown time of insulating fluid on voltage is

$$\mu\{Ln(Time\,|\,Voltage)\} = 18.955 - 0.507 * Voltage.$$

The negative sign of the slope is logical because the relationship between the log-breakdown time and voltage is negative. Any predictions of the natural logarithm of the breakdown time based on the above estimated regression line are valid only in the range 26 to 38 kV of voltage.

The estimated regression line was obtained for the log-transformed data. We remember that if the log-transformed responses have a symmetric distribution, then taking the antilogarithm of the slope of the estimated regression line for the log-transformed data, shows a multiplicative change in the median response as the explanatory variable increases by 1 unit.

The slope of -0.507 shows the average change in the natural logarithm of the breakdown time as voltage increases by 1 kV. Thus a one kV increase in voltage is associated with a multiplicative change in median breakdown time of exp(-0.507) = 0.60. So, the median breakdown time at 28 kV is 60% of what it is at 27 kV; the median breakdown time at 29 kV is 60% of what it is at 28 kV.

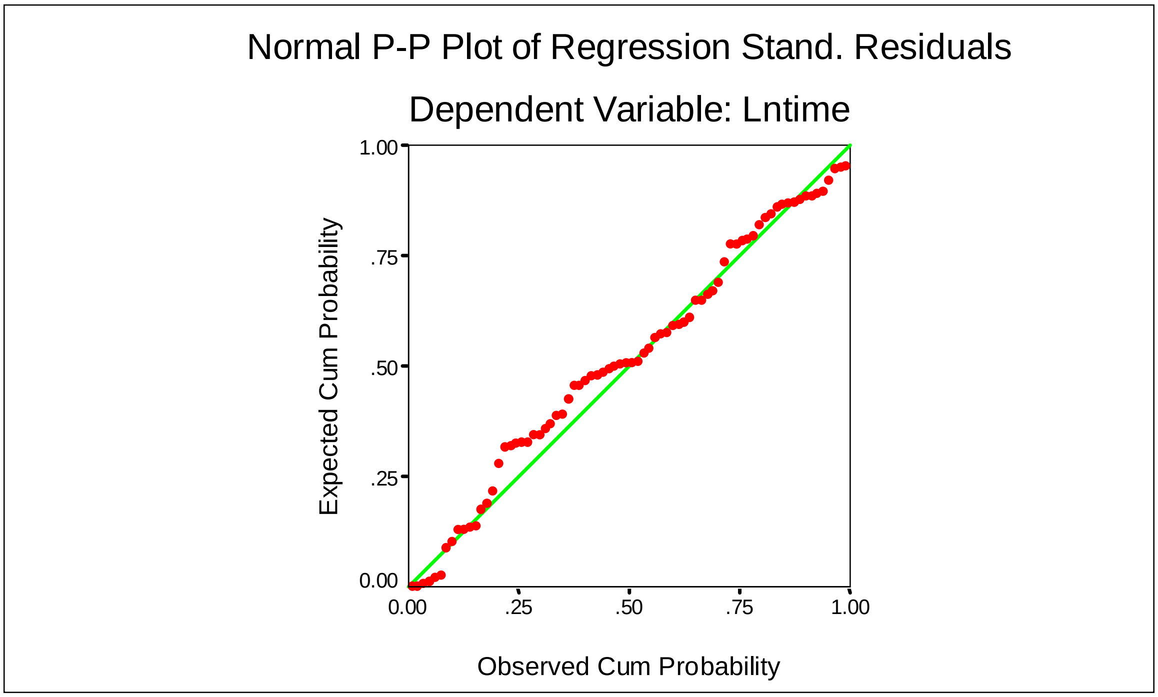
Since a 95% confidence interval for β_1 is -0.622 to -0.393, a 95% confidence interval for exp(β_1) is exp(-0.622) to exp(-0.393), or 0.54 to 0.68.

13.5 Checking the Regression Model Assumptions

The conclusions based on the regression model are valid only if the underlying assumptions are satisfied. The assumptions are the normality and constant variance.

The normality assumption can be verified by looking at the plot residuals. Indeed, if the residuals at each voltage level follow a normal distribution, then the log-breakdown times follow a normal distribution and vice versa.

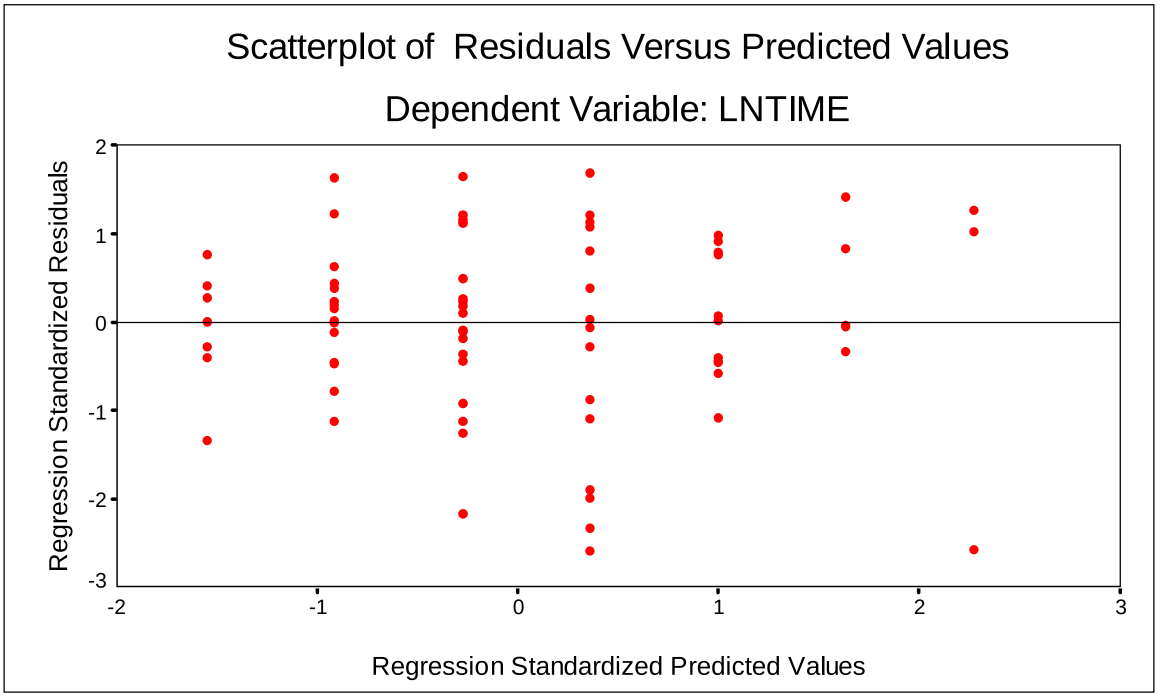
In order to assess whether the assumption is not violated with SPSS, the normal P-P plot of regression standardized residuals is obtained. If the normality assumption is not violated, points will cluster around a straight line.



As you can see, the above plot supports the normality assumption. The pattern is close enough to a straight line.

One method of checking whether the assumption of constant variance is not violated is to plot the residuals against the predicted values. We then look for a change in the spread or dispersion of the plotted points.

The scatterplot displayed below shows that the assumption of constant variance is not very likely to be violated. The spread of the plotted points is significantly different at different voltage levels.



After a logarithmic transformation of the times to breakdown, a simple linear regression model fits the insulating fluid data well. No evidence (from the normal P-P plot and residual plot) indicates that either the normality or constant variance assumption is violated. Thus mean estimation and prediction can proceed from that model, with results back-transformed to the original scale.

13.6 ANOVA Model

In this section we will apply the analysis of variance to our experiment. The separate means model (one-way ANOVA) applied to our experiment has the form

$$\mu\{Ln(Time)\,|\,Voltage\_i\} = \mu_i$$

for i=26, 28, …, 38, where μ_i denotes the mean of the group subjected to the voltage of i kV. This model is obviously more general than the simple regression model because the log-breakdown time means may or may not lie on the straight line -their values are not restricted.

The following display contains the output of the one-way analysis of variance applied to the log-transformed breakdown times.

```
Variable  LNTIME
By Variable  CODE

                         Analysis of Variance


                                  Sum of       Mean          F          F
Source                    D.F.    Squares      Squares       Ratio      Prob.

Between Groups               6    196.4774     32.7462     13.0043     .0000
Within Groups               69    173.7489      2.5181
Total                       75    370.2263



                                  Standard   Standard
Group        Count      Mean      Deviation  Error    95% Conf Inter for Mean

26 kV            3      5.6240     3.3552    1.9371    -2.7109   TO   13.9589
28 kV            5      5.3295     1.1446     .5119     3.9084   TO    6.7507
30 kV           11      3.8220     1.1112     .3350     3.0755   TO    4.5685
32 kV           15      2.2285     2.1981     .5675     1.0113   TO    3.4458
34 kV           19      1.7864     1.5252     .3499     1.0513   TO    2.5215
36 kV           15       .9022     1.1099     .2866      .2876   TO    1.5169
38 kV            8      -.4243      .9917     .3506    -1.2534   TO     .4047

Total           76      2.1457     2.2218     .2549     1.6380   TO    2.6534
```

We remember that if the log-transformed data have a symmetric distribution, then taking the antilogarithm of the mean on the log scale gives an estimate of the median on the original scale. Hence, in order to obtain the estimates of the group medians on the original scale of measurement, you have to take the antilogarithm of the above group means.

The analysis of variance F-statistic is F=13.0043, with 6 and 69 degrees of freedom, giving a reported p-value of 0. The p-value indicates strong evidence against the null hypothesis of no difference among the average log-breakdown times for the seven groups. In other words, there is strong evidence of differences among the group medians on the original scale of measurement.

13.7 Comparing the Regression and ANOVA Models

We have used two different statistical techniques to analyze the insulating fluid data: a separate-means model (one-way ANOVA) and simple linear regression model.

The separate-means (one-way ANOVA) model provided strong evidence of differences among the mean breakdown times for the seven experimental groups. As the batches of insulating fluid were randomly assigned to the different voltage levels, it was inferred that the different voltage levels must be directly responsible for the observed differences in time to breakdown.

The simple linear regression model is based on the assumption that the log-breakdown time means lie on a straight line against voltage. The separate-means model is obviously more general than the simple regression model because the log-breakdown time means may or may not lie on the straight line - their values are not restricted.

The simple linear regression model describes the relationship between the predictor variable *Voltage* and the response variable mean log-breakdown time in the form

$$\mu\{Ln(Time\,|\,Voltage)\} = 18.955 - 0.507 * Voltage.$$

Thus the simple linear regression model allows us to make predictions about the mean log-breakdown time and voltage for the voltage levels inside the experimental range 26-38 kV. Mean estimation and prediction can proceed from that model, with results back-transformed to the original scale.

Which model produces better estimates of the mean breakdown times at a given voltage level? In order to answer the question, we analyze the following table of the estimates of the group means with their corresponding standard errors.

Estimates of Group Means Using Regression and ANOVA

		Regression		**ANOVA**	
kV	**n**	**Estimate**	**St. Error**	**Estimate**	**St. Error**
26	3	5.76397	0.44673	5.6240	0.91617
28	5	4.74924	0.34463	5.3295	0.70966
30	11	3.73451	0.25362	3.8220	0.47845
32	15	2.71978	0.19036	2.2285	0.40972
34	19	1.70505	0.18575	1.7864	0.36405
36	15	0.69032	0.24315	0.9022	0.40972
38	8	-0.32441	0.33181	-0.4243	0.56104

As you can see, the standard errors from ANOVA model are uniformly larger than those obtained from the simple linear regression model. That means that the regression estimates of the mean at any particular voltage level will be more precise than the average from the batches that were tested at that level (ANOVA estimate).

Summarizing, a simple regression model should be preferred in this case because the model fits the data well. No evidence (from a residual plot and a lack-of-fit test) indicates lack of fit. The model allows for interpolation and produces better estimates of the group means.

13.8 Summary

The side-by-side boxplots and scatterplot for the data reveal non-constant spread, non-linear pattern, and presence of outliers. This pattern calls for the transformation of the response variable. After a logarithmic transformation of the times to breakdown, the data exhibit a linear pattern, approximately equal spread, and lack of outliers.

Two different statistical techniques were used to analyze the data: one-way ANOVA and simple linear regression. ANOVA provided strong evidence of differences among the mean breakdown times for the seven experimental groups. As the batches of insulating fluid were randomly assigned to the different voltage levels, it was inferred that the different voltage levels must be directly responsible for the observed differences in time to breakdown.

As the log-transformed data reveal a linear pattern, the simple linear regression model was applied. This model produced the following estimated regression line

$$\mu\{Ln(Time\,|\,Voltage)\} = 18.955 - 0.507 * Voltage.$$

Based on this equation, we can make predictions about the mean log-breakdown time and voltage for the voltage levels inside the experimental range 26-38 kV. Mean estimation and prediction can proceed from that model, with results back-transformed to the original scale. There was no evidence that the underlying model assumptions are violated.

It was demonstrated that simple linear regression model produces better estimates of the mean breakdown times at a given voltage level than ANOVA does.