10. 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. Which of the two models is more suitable for the problem? Which one provides more information about the relationship between voltage level and breakdown time? Which model produces better estimates of the mean breakdown times at a given voltage level? We will answer these and other related questions in this section.

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. In other words, there is strong evidence that voltage level had an effect on breakdown time. Moreover, the direction of the relationship is specified: voltage is the cause, breakdown time is the effect.

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 regression model establishes a linear pattern between mean log breakdown time and voltage, for all voltages in the range of 26 to 38 kV. Indeed, 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. One-way analysis of variance does not allow us to make predictions about breakdown time for the voltage levels different from seven values given in the data.

Which model produces better estimates of the mean breakdown times at a given voltage level? In order to answer the question, it is necessary to look at the standard errors of the group means in the two models.

First, let us consider the simple linear regression model. The estimates of the group means with their standard errors for the voltage levels of 26, 28,…, 38 kV were displayed at the end of **Section 6**. These estimates can be obtained by using the above estimated regression equation.

Now, consider the one-way ANOVA model. The SPSS output provided in **Section 8** gives the estimates of the means for each voltage level with their corresponding standard errors. The estimated means are simply the group averages. Unfortunately, the standard errors of the group means are based on the individual standard deviations for each group, not on the pooled estimate of standard deviation. In order to obtain the pooled standard

errors of the means, which is the one-way analysis of variance standard error, we divide the pooled estimate of standard deviation by the square root of the sample size.

According to the ANOVA output, the within-group mean square is 2.5181, so the pooled estimate of a common standard deviation is the square root of the value, which is equal to 1.586852 minutes. Now the pooled standard errors can be obtained by dividing the value by the square root of the sample sizes. The results with the values obtained from the regression model are displayed in the following table.

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. Indeed, in the simple linear regression model all 76 observations are used to estimate the mean at a particular voltage level, not just those from that particular group.

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.