poor fit to the subpopulation averages and increasing variability. It is evident that the square root transformation has not resolved the problem. A log transformation, however, works well (see Display 8.4). (*Note*: The horn-shaped pattern is always a key indicator of the need to transform, whether or not there are replicate samples at specific values of the explanatory variable.)

Transformations Indicated by Horn-Shaped Residual Plots

A horn-shaped pattern in the residual plot suggests a response transformation like the square root, the logarithm, or the reciprocal. The logarithm is the easiest to interpret. The reciprocal, $1/Y$, works better when the nonconstant spread is more severe, and the square root works better when the nonconstant spread is less severe. Judging severity of nonconstant variance in the horn-shaped residual plot is difficult and unnecessary, however. Try one of the transformations, re-fit the regression model on the transformed data, and redraw the residual plot to see if the transformation has worked. If not, then one of the others can be attempted.

8.4 INTERPRETATION AFTER LOG TRANSFORMATIONS

A data analyst must interpret regression results in a way that makes sense to the intended audience. The appropriate wording for inferential statements after logarithmic transformation depends on whether the transformation was applied to the response, to the explanatory variable, or to both.

When the Response Variable Is Logged

If $\mu\{\log(Y)|X\} = \beta_0 + \beta_1 X$, and if the distribution of the transformed responses about the regression is symmetric, then

$$\text{Median}\{Y|X\} = \exp(\beta_0)\exp(\beta_1 X).$$

Consequently,

$$\text{Median}\{Y|(X+1)\}/\text{Median}\{Y|X\} = \exp(\beta_1),$$

so an increase in X of 1 unit is associated with a multiplicative change of $\exp(\beta_1)$ in $\text{Median}\{Y|X\}$.

For example, the estimated relationship between the breakdown time (BDT) of insulating fluid and voltage is $\hat{\mu}\{\log(\text{BDT}) \mid \text{Voltage}\} = 19.0 - 0.51$ Voltage. A 1 kV increase in voltage is associated with a multiplicative change in median BDT of $\exp(-0.51)$, or 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, and so on. Since a 95% confidence interval for β_1 is -0.62 to -0.39, a 95% confidence interval for $\exp(\beta_1)$ is $\exp(-0.62)$ to $\exp(-0.39)$, or 0.54 to 0.68.

The statement Median $\{Y|(X+1)\} = 0.6 \times \text{Median}\{Y|X\}$ can also be written as Median $\{Y|(X+1)\} - \text{Median}\{Y|X\} = 0.4 \times \text{Median}\{Y|X\}$, which permits the following kind of statement: "It is estimated that the median of Y decreases by 40% for each one unit increase in X (95% confidence interval: 32% to 46%)."

When the Explanatory Variable Is Logged

The relationship $\mu\{Y|\log(X)\} = \beta_0 + \beta_1 \log(X)$ can be described in terms of multiplicative changes in X, either as a change in the mean of Y for each doubling of X or a change in the mean of Y for each ten-fold increase in X. The chosen multiple should be consistent with the range of X's in the data set.

Notice that

$$\mu\{Y|\log(2X)\} - \mu\{Y|\log(X)\} = \beta_1 \log(2),$$

so a doubling of X is associated with a $\beta_1 \log(2)$ change in the mean of Y. Similarly, a ten-fold increase in X is associated with a $\beta_1 \log(10)$ change in the mean of Y.

For the meat processing data of Section 7.1.2, $\hat{\mu}\{\text{pH}|\log(\text{Time})\} = 6.98 - 0.726 \log(\text{Time})$, so a doubling of time after slaughter is associated with a $\log(2)(-0.726) = -0.503$ unit change in pH. Since a 95% confidence interval for β_1 is from -0.805 to -0.646, a 95% CI for $\log(2)\beta_1$ is from -0.558 to -0.448. In words: It is estimated that the mean pH is reduced by 0.503 for each doubling of time after slaughter (95% confidence interval 0.448 to 0.558).

When Both the Response and Explanatory Variables Are Logged

The interpretation is a combination of the previous two. If $\mu\{\log(Y)|\log(X)\} = \beta_0 + \beta_1 \log(X)$, then $\text{Median}\{Y|X\} = \exp(\beta_0)X^{\beta_1}$. A doubling of X is associated with a multiplicative change of 2^{β_1} in the median of Y. Or, a ten-fold increase in X is associated with a 10^{β_1}-fold change in the median of Y.

For the island size and number of species data, $\hat{\mu}\{\log(\text{species})|\log(\text{area})\} = 1.94 + 0.250 \log(\text{area})$. Thus, an island area of $2A$ is estimated to have a median number of species that is $2^{0.250}$ (or 1.19) times the median number of species for an island of area A. Since a 95% confidence interval for β_1 is 0.219 to 0.281, a 95% confidence interval for the multiplicative factor in the median is $2^{0.219}$ to $2^{0.281}$, or 1.16 to 1.22.

The Need for Interpretation

An important aspect of the log transformation is that straightforward multiplicative statements emerge. Straightforward interpretations after other transformations only follow in certain instances. For example, the square root of the cross-sectional area of a tree may be re-expressed as the diameter; the reciprocal of the time to complete a race may be interpreted as the speed. In general, however, interpretation for other transformations may be awkward.

For two types of questions, interpretation is not critical. First, if the regression is used only to assess whether the distribution of the response is *associated with* the explanatory variable, it is sufficient to test the hypothesis that the slope in the regression of Y on X is zero, where Y or X are transformed variables. If the

distribution of the transformed response is associated with X, the distribution of the response is associated with X, even though that association may be difficult to describe. Secondly, if the purpose is prediction, no interpretation of the regression coefficients is needed. It is only necessary to express the prediction on the original scale, regardless of the expression used to make the prediction.

8.5 ASSESSMENT OF FIT USING THE ANALYSIS OF VARIANCE

An analysis of variance can be used to compare several models. When there are replicate response variables at several explanatory variable values, an analysis of variance F-test for comparing the simple linear regression model to the separate-means (one-way analysis of variance) model supplies a formal assessment of the goodness of fit of simple linear regression. This is called the *lack-of-fit F-test.*

8.5.1 Three Models for the Population Means

In the following three model descriptions for means, the subscript i refers to the group number (e.g., different voltage levels).

1.	Separate-means model:	$\mu\{Y\|X_i\} = \mu_i,$	for $i = 1, \ldots, I.$
2.	Simple linear regression model:	$\mu\{Y\|X_i\} = \beta_0 + \beta_1 X_i,$	for $i = 1, \ldots, I.$
3.	Equal-means model:	$\mu\{Y\|X_i\} = \mu,$	for $i = 1, \ldots, I.$

The separate-means model has no restriction on the values of any of the means. It has I different parameters—the individual group means. The simple linear regression model has two parameters, the slope and the intercept. The equal-means model has a single parameter.

These models form a *hierarchical* set. The equal-means model is a special case of the simple linear regression model, which in turn is a special case of the separate-means model. Viewed conversely, the separate-means model is a generalization of the simple linear regression model, which in turn is a generalization of the equal-means model. A generalization of one model is another model that contains the same features but uses additional parameters to describe a more complex structure.

8.5.2 The Analysis of Variance Table Associated with Simple Regression

Display 8.8 contains two different analysis of variance tables. Table (b) comes from a one-way analysis of variance (Chapter 5), showing the details of an F-test that compares the separate-means model to the equal-means model for the insulating fluid example of Section 8.1.2. Table (a) comes from a simple linear regression analysis showing the details of an F-test that compares the simple linear regression model to the equal-means model.

If β_1 is zero, the simple linear regression model reduces to the equal-means model, $\mu\{Y|X\} = \beta_0$. The hypothesis that $\beta_1 = 0$ can therefore be tested with a comparison of the sizes of the residuals from fits to these two models. An analysis of