

Part IV

**Analyzing Data with Nonlinear
Models**

In Unit One we began to see the world as data; in Unit Two we began to ask questions of data in order to find out the story it has to tell about itself, and hence about the world from which it was extracted. In Unit Three we began to make connections between sets of data, to see how the events in the situations from which the data were extracted might be related to each other. We began to analyze the relationships between sets of data by capturing those relationships in regression models, simple linear ones at first involving a dependent variable and a single explanatory variable, and then more complex linear ones with a dependent variable and several explanatory variables. This unit investigates one of the four assumptions that underlie regression modeling and at the same time seeking to develop the relationships between even more complex sets of data.

One of the main assumptions about data when you construct a regression model is that the data is sampled from a linear relationship of some sort (either two-variable or more than two-variables). If this is not true, then your resulting regression model may seem to be okay, but it will exhibit problems of one of the following types:

1. The model may be accurate for only a small slice of data. If we apply the model to data points outside this small slice, the resulting errors from the model may become larger and larger. This is related to having too small a sample of the data to notice that it really does not exhibit linearity.
2. The regression model consistently underestimates the data in certain regions and consistently overestimates it in other regions. This resulting pattern indicates that there is a better model for the data than a linear model.

In chapter 11 (page 311) we begin dealing with data that is not proportional, that is, data that violates our first regression assumption that a linear model is an appropriate fit. We will start by focusing on two-variable data and then learn how to extend this to multivariable data. Even though most real data sets are multi-dimensional, there are solid reasons for beginning our study with two-variable nonlinear data sets:

- Not all data is multidimensional - sometimes two variables are enough.
- Even in multidimensional data, we are often interested in the main effect first. That means looking at how the most significant variable relates to the dependent variable.
- In many modeling applications, the data shows one dependent variable and two independent variables with a constraint (like total cost must be less than a fixed amount). In this case, the constraint relationship between the two independent variables can be used to reduce the number of independent variables to one, making the entire data set two dimensional.
- Finally, the models we are going to discuss are easy to picture in two dimensions; in more dimensions, it is difficult to picture the models and develop an intuitive feel for what they can do. But the intuition we develop with two-variable data will help us interpret the diagnostic graphs in StatPro's regression output when we are dealing with multidimensional models.

In much the same way that straight lines have parameters that can be chosen so as to match the line closely to the data, the basic nonlinear models we will introduce have parameters that can serve the same purpose. By using these parameters to shift one of the basic models horizontally and vertically and to stretch them and flip it, we can fit this basic function to a non-proportional data set.

However, the regression routines in StatPro are only useful for producing linear models. In fact, we overcome this problem by transforming nonlinear data so that it becomes suitably linear and then applying our regression model to this straightened out data. Thus, chapter 12 (page 347) presents the key transformations that will convert many kinds of nonlinear data into linear data. This chapter also teaches us how to evaluate the quality of models built from transformed data and then how to interpret these models. The unit closes with chapter 13 (page 379) on interpreting the relationships in nonlinear models with more than one variable. We also discuss how to locate the maxima and minima of such functions.