distribution is normally distributed. That means if you take 100 samples of a non-normal distribution and calculate the mean for each of them, the distribution of the 100 mean values is normally distributed. This attribute of sampling can be applied when you take multiple samples of a population and submit the data to linear regression analysis. We will discuss this further in the following paragraphs.

Don’t look for quick ways out of this assumption. If your data set distribution is significantly different from normal, it can ruin any analysis based on parametric statistical methods. And the most painful part of this problem is you may not know when you are wrong! Testing for a normal distribution is a recommended step and is available in most statistical and data mining programs. If you don’t verify the normality of your data set, all the statistical tests may show a strong relationship in your model, but that relationship may fail miserably when you try to apply it.

We will discuss some fixes for nonindependency in the section titled “Linear Regression.”

The Assumption of Linearity

The third major assumption inherent in classical parametric procedures is that the variables have a linear effect on the response variable (the target). This means that a plot of the relationship of any variable to the response variables is a straight line. Examples of common statistical procedures that make these three assumptions are ANOVA and linear regression. Many fixes exist for handling nonlinear variables in these linear analyses. We will look at them next.

LINEAR REGRESSION

Linear regression was first proposed by Sir Francis Galton (1822–1911). Galton coined the term regression to describe the observation that the majority of very tall fathers had sons who were shorter, and most very short fathers had sons taller than them. The trend of this progression in height was toward the average (or mean) height. This phenomenon was termed regression to the mean. His analysis of this effect became known simply as regression.

The major objectives of linear regression are to

- Determine if a relationship exists between one variable and another (or a set of others);
- Describe the nature of this relationship, if it exists;
- Quantify the accuracy of this relationship;
- Evaluate the relative contributions of each variable, if multiple variables are used.

Linear regression makes all three assumptions described previously. But you can appeal to the Central Limit Theorem to correct for non-normality by taking multiple samples and working on the means rather than the original data. That is, you can generate a group of data for a given variable by drawing a group of samples for the population, finding the mean, drawing another group of samples, finding the mean, and so forth until you have