selected subsample may reflect the distribution of this population well. Similarly, results regarding diet and exercise obtained from a study performed on patients at risk for heart disease may not be generally applicable to generally healthy individuals. In this case assumptions would have to be made about how results for the at-risk population might relate to those for the healthy population.

Data used in empirical research rarely meet all (if any) of these criteria precisely. However, keeping these goals in mind can help you be precise about the types of questions you can and cannot answer reliably.

2. *Additivity and linearity.* The most important mathematical assumption of the regression model is that its deterministic component is a linear function of the separate predictors: $y = \beta_1 x_1 + \beta_2 x_2 + \cdots$.

   If additivity is violated, it might make sense to transform the data (for example, if $y = abc$, then $\log y = \log a + \log b + \log c$) or to add interactions. If linearity is violated, perhaps a predictor should be put in as $1/x$ or $\log(x)$ instead of simply linearly. Or a more complicated relationship could be expressed by including both x and x^2 as predictors.

   For example, it is common to include both `age` and `age`^{2} as regression predictors. In medical and public health examples, this allows a health measure to decline with higher ages, with the rate of decline becoming steeper as age increases. In political examples, including both `age` and `age`^{2} allows the possibility of increasing slopes with age and also U-shaped patterns if, for example, the young and old favor taxes more than the middle-aged.

   In such analyses we usually prefer to include age as a categorical predictor, as discussed in Section 4.5. Another option is to use a nonlinear function such as a spline or other generalized additive model. In any case, the goal is to add predictors so that the linear and additive model is a reasonable approximation.

3. *Independence of errors.* The simple regression model assumes that the errors from the prediction line are independent. We will return to this issue in detail when discussing multilevel models.

4. *Equal variance of errors.* If the variance of the regression errors are unequal, estimation is more efficiently performed using weighted least squares, where each point is weighted inversely proportional to its variance (see Section 18.4). In most cases, however, this issue is minor. Unequal variance does not affect the most important aspect of a regression model, which is the form of the predictor $X\beta$.

5. *Normality of errors.* The regression assumption that is generally *least* important is that the errors are normally distributed. In fact, for the purpose of estimating the regression line (as compared to predicting individual data points), the assumption of normality is barely important at all. Thus, in contrast to many regression textbooks, we do *not* recommend diagnostics of the normality of regression residuals.

   If the distribution of residuals is of interest, perhaps because of predictive goals, this should be distinguished from the distribution of the data, y. For example, consider a regression on a single discrete predictor, x, which takes on the values 0, 1, and 2, with one-third of the population in each category. Suppose the true regression line is $y = 0.2 + 0.5x$ with normally distributed errors with standard deviation 0.1. Then a graph of the data y will show three fairly sharp modes centered at 0.2, 0.7, and 1.2. Other examples of such mixture distributions arise in economics, when including both employed and unemployed people, or