10-4 Prediction Intervals and Variation

Key Concept In Section 10-3 we presented a method for using a regression equation to find a predicted value of y, but it is helpful to have some sense of the accuracy of such predictions. In this section we present a method for constructing a *prediction interval,* which is an interval estimate of a predicted value of y. (Interval estimates of parameters are *confidence intervals,* but interval estimates of variables are called *prediction intervals.*)

In Example 4(b) from the preceding section, we showed that when using the 40 pairs of shoe print lengths and heights from Data Set 2 in Appendix B, the regression equation is $\hat{y} = 80.9 + 3.22x$. Given a shoe print length of 29 cm, the best predicted height is $\hat{y} = 174.3$ cm (which is found by substituting 29 cm for x in the regression equation). For $x = 29$ cm, the best predicted height is $\hat{y} = 174.3$ cm, but we have no sense of the accuracy of that estimated height, so we need an interval estimate. Instead of using a confidence interval (which is used for an estimate of a population parameter), we use a *prediction interval*, which is an interval estimate of a predicted value of the variable y. A prediction interval serves the same role as a confidence interval, except that a prediction interval is used for an estimate of a value of a *variable,* whereas a confidence interval is used for an estimate of a value of a population parameter.

DEFINITION A confidence interval is a range of values used to estimate a population parameter, but a **prediction interval** is a range of values used to estimate a variable, such as a predicted value of y in a regression equation.

A prediction interval estimate of a predicted value $\hat{y}$ can be found using the key elements in the following box. Given the nature of the calculations, the use of technology is highly recommended.

Prediction Interval for an Individual y

Requirement

For each fixed value of x, the corresponding sample values of y are normally distributed about the regression line, and those normal distributions have the same variance.

Given a fixed and known value x_0, the prediction interval for an individual y value is

$$\hat{y} - E < y < \hat{y} + E$$

where the margin of error is

$$E = t_{\alpha/2} s_e \sqrt{1 + \frac{1}{n} + \frac{n(x_0 - \bar{x})^2}{n(\Sigma x^2) - (\Sigma x)^2}}$$

and x_0 represents the given value of x, $t_{\alpha/2}$ has $n - 2$ degrees of freedom, and s_e is the **standard error of estimate** found from Formula 10-5 or Formula 10-6. (The standard error of estimate s_e is a measure of variation of the residuals, which are the differences between the observed sample y values and the predicted values $\hat{y}$ that are found from the regression equation.)

Formula 10-5

$$s_e = \sqrt{\frac{\Sigma(y - \hat{y})^2}{n - 2}}$$

Formula 10-6

$$s_e = \sqrt{\frac{\Sigma y^2 - b_0 \Sigma y - b_1 \Sigma xy}{n - 2}}$$

(This is an equivalent form of Formula 10-5 that is good for manual calculations or writing computer programs.)