10.7: Estimation and Prediction

Learning Objectives

- To learn the distinction between estimation and prediction.
- To learn the distinction between a confidence interval and a prediction interval.
- To learn how to implement formulas for computing confidence intervals and prediction intervals.

Consider the following pairs of problems, in the context of Example 10.4.2, the automobile age and value example.

Problem 1

1. Estimate the average value of all four-year-old automobiles of this make and model.
2. Construct a 95% confidence interval for the average value of all four-year-old automobiles of this make and model.

Problem 2

1. Shylock intends to buy a four-year-old automobile of this make and model next week. Predict the value of the first such automobile that he encounters.
2. Construct a 95% confidence interval for the value of the first such automobile that he encounters.

The method of solution and answer to the first question in each pair, (1a) and (2a), are the same. When we set x equal to 4 in the least squares regression equation

$$\hat{y} = -2.05x + 32.83$$

that was computed in part (c) of Example 10.4.2, the number returned,

$$\hat{y} = -2.05(4) + 32.83 = 24.63$$

which corresponds to value $\$24,630$, is an estimate of precisely the number sought in question (1a): the mean $E(y)$ of all y values when $x = 4$. Since nothing is known about the first four-year-old automobile of this make and model that Shylock will encounter, our best guess as to its value is the mean value $E(y)$ of all such automobiles, the number 24.63 or $\$24,630$, computed in the same way.

The answers to the second part of each question differ. In question (1b) we are trying to estimate a population parameter: the mean of the all the y-values in the sub-population picked out by the value $x = 4$, that is, the average value of all four-year-old automobiles. In question (2b), however, we are not trying to capture a fixed parameter, but the value of the random variable y in one trial of an experiment: examine the first four-year-old car Shylock encounters. In the first case we seek to construct a confidence interval in the same sense that we have done before. In the second case the situation is different, and the interval constructed has a different name, prediction interval. In the second case we are trying to "predict" where a the value of a random variable will take its value.

$100(1-\alpha)\%$ Confidence Interval for the Mean Value of y at $x = x_p$

$$\hat{y}_p \pm t_{\alpha/2}\ s_\varepsilon \sqrt{\frac{1}{n} + \frac{(x_p - \overline{x})^2}{SS_{xx}}}$$

where

- x_p is a particular value of x that lies in the range of x-values in the sample data set used to construct the least squares regression line;
- $\hat{y}_p$ is the numerical value obtained when the least square regression equation is evaluated at $x = x_p$; and
- the number of degrees of freedom for $t_{\alpha/2}$ is $df = n - 2$.

The assumptions listed in Section 10.3 must hold.

The formula for the prediction interval is identical except for the presence of the number 1 underneath the square root sign. This means that the prediction interval is always wider than the confidence interval at the same confidence level and value of x. In

practice the presence of the number 1 tends to make it much wider.

$100(1-\alpha)\%$ Prediction Interval for an Individual New Value of of y at $x = x_p$

$$\hat{y}_p \pm t_{\alpha/2}\ s_\varepsilon \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}}$$

where

- x_p is a particular value of x that lies in the range of x-values in the data set used to construct the least squares regression line;
- $\hat{y}_p$ is the numerical value obtained when the least square regression equation is evaluated at $x = x_p$; and
- the number of degrees of freedom for $t_{\alpha/2}$ is $df = n - 2$.

The assumptions listed in Section 10.3 must hold.

Example 10.7.1

Using the sample data of "Example 10.4.2" in Section 10.4 , recorded in Table 10.4.3, construct a 95% confidence interval for the average value of all three-and-one-half-year-old automobiles of this make and model.

Solution

Solving this problem is merely a matter of finding the values of $\hat{y_p},\ \alpha,\ and\ t_{\alpha/2}, S_\varepsilon,\ \bar{x}\ and\ SS_{xx}$, and inserting them into the confidence interval formula given just above. Most of these quantities are already known. From Example 10.4.2, $SS_{xx} = 14\ \ and\ \ \bar{x} = 4$. From Example 10.5.2, $S\varepsilon = 1.902169814$

From the statement of the problem $x_p = 3.5$, the value of x of interest. The value of $\hat{y_p}$ is the number given by the regression equation, which by Example 10.4.2 is $\hat{y} = -2.05x + 32.83$, when $x = x_p$, that is, when $x = 3.5$. Thus here $\hat{y} = -2.05(3.5) + 32.83 = 25.655$

Lastly, confidence level 95% means that $\alpha = 1 - 0.95 = 0.05$ so $\alpha/2 = 0.025$. Since the sample size is $n = 10$, there are $n - 2 = 8$ degrees of freedom. By Figure 7.1.6, $t_{0.025} = 2.306$. Thus

$$\begin{aligned}\hat{y_p} \pm t_{\alpha/2} S_\varepsilon \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}} &= 25.655 \pm (2.306)(1.902169814)\sqrt{\frac{1}{10} + \frac{(3.5-4)^2}{14}} \\ &= 25.655 \pm 4.386403591\sqrt{0.1178571429} \\ &= 25.655 \pm 1.506\end{aligned}$$

which gives the interval $(24.149, 27.161)$

We are 95% confident that the average value of all three-and-one-half-year-old vehicles of this make and model is between $\$24,149$ and $\$27,161$.

Example 10.7.2

Using the sample data of Example 10.4.2, recorded in Table 10.4.3, construct a 95% prediction interval for the predicted value of a randomly selected three-and-one-half-year-old automobile of this make and model.

Solution

The computations for this example are identical to those of the previous example, except that now there is the extra number 1 beneath the square root sign. Since we were careful to record the intermediate results of that computation, we have immediately that the 95% prediction interval is

$$\begin{aligned}\hat{y_p} \pm t_{\alpha/2} S_\varepsilon \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}} &= 25.655 \pm 4.386403591\sqrt{1.1178571429} \\ &= 25.655 \pm 4\end{aligned}$$

which gives the interval $(21.017, 30.293)$.

We are 95% confident that the value of a randomly selected three-and-one-half-year-old vehicle of this make and model is between $\$21,017$ and $\$30,293$.

Note what an enormous difference the presence of the extra number 1 under the square root sign made. The prediction interval is about two-and-one-half times wider than the confidence interval at the same level of confidence.

KeyTakaways

- A confidence interval is used to estimate the mean value of y in the sub-population determined by the condition that x have some specific value x_p.
- The prediction interval is used to predict the value that the random variable y will take when x has some specific value x_p.