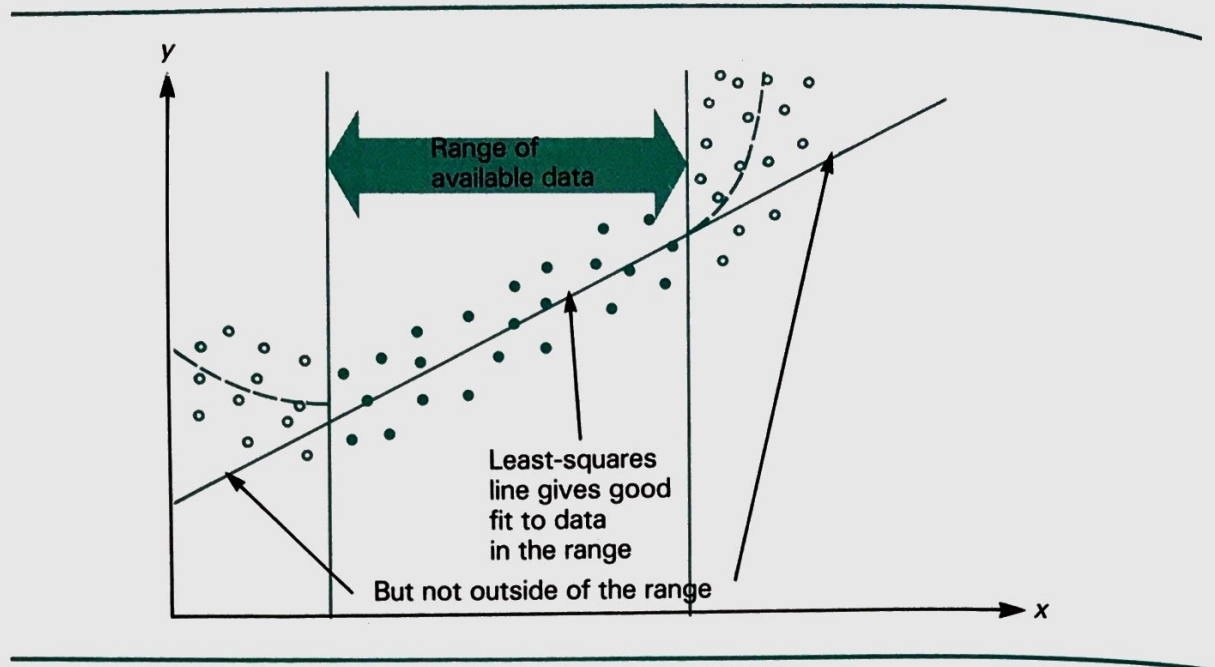


9-10 USE OF THE REGRESSION MODEL FOR PREDICTION:

As mentioned in the introduction of this chapter, there are several possible uses of a regression model. One is to understand the relationship between the two variables. As with correlation analysis, understanding a relationship between two variables in regression does not imply that one variable causes the other. Causality is a much more complicated issue and cannot be determined by a simple regression analysis.

A more common use of a regression analysis is *prediction*: providing estimates of values of the dependent variable by using the prediction equation: $\hat{Y} = b_0 + b_1 X$. It is important that prediction be done in the region of the data used in the estimation process. *You should be aware that using a regression for extrapolating outside the estimation range is risky, as the estimated relationship may not be appropriate outside this range.* This is demonstrated in Figure 9-27.

FIGURE 9-27 The Danger of Extrapolation



Point Predictions

It is very easy to produce point predictions using the estimated regression equation. All we need to do is substitute the value of X for which we want to predict Y into the prediction equation. In Example (a), suppose that American Express wants to predict charges on the card for a member who traveled 4,000 miles during a period equal to the one studied (note that $x = 4,000$ is in the range of X values used in the estimation). We use the prediction equation, Equation 9-12, but with higher accuracy for b_1 :

$$\hat{y} = 274.85 + 1.2553x = 274.85 + 1.2553(4,000) = 5,296.05 \text{ (dollars)}$$

The process of prediction in this example is demonstrated in Figure 9-28.

Prediction Intervals

Point predictions are not perfect and are subject to error. The error is due to the uncertainty in estimation as well as the natural variation of points about the regression line. A $(1 - \alpha)100\%$ prediction interval for Y is given in Equation 9-32.

A $(1 - \alpha)100\%$ prediction interval for Y :

$$\hat{y} \pm t_{\alpha/2} s \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{SS_x}} \quad (9-32)$$

As can be seen from the formula, the width of the interval depends on the distance

FIGURE 9-28 Prediction in American Express Study

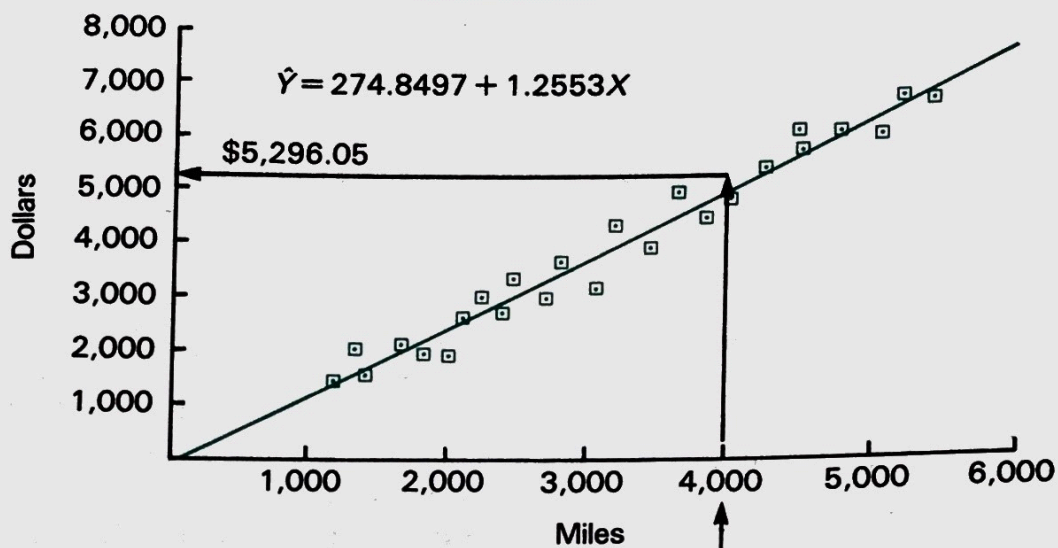
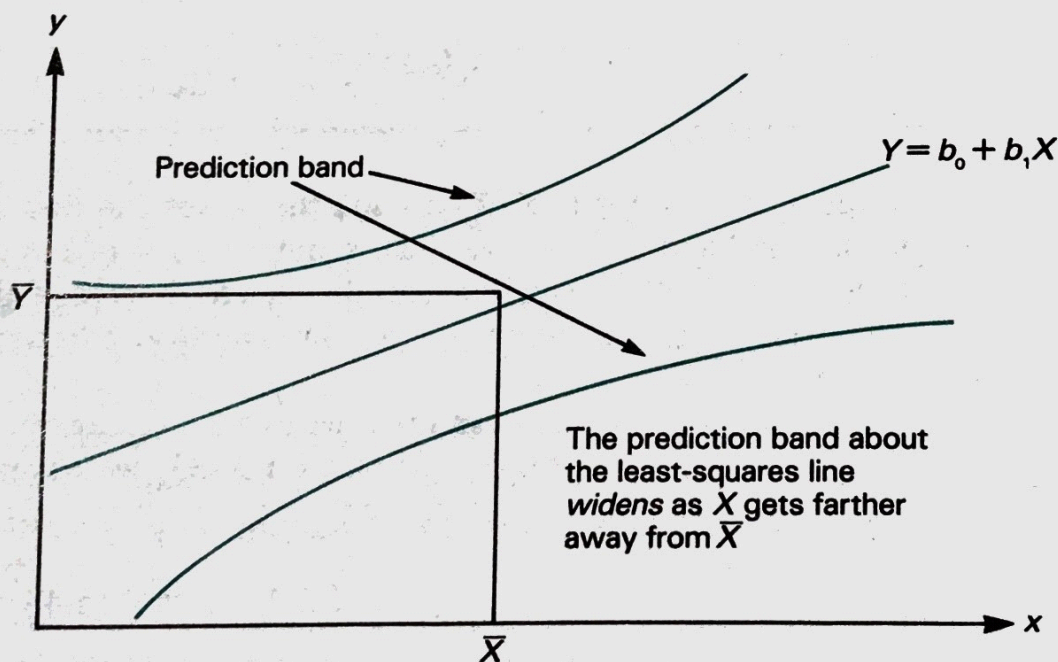


FIGURE 9-29 Prediction Band and Its Width



of our value x (for which we wish to predict Y) from the mean, $\bar{x}$. This is shown in Figure 9-29.

We will now use Equation 9-32 to compute a 95% prediction interval for the amount charged on the American Express card by a member who traveled 4,000 miles. We know that in this example, $\bar{x} = \sum x/n = 79,448/25 = 3,177.92$. We also know that $SS_x = 40,947,557.84$ and $s = 318.16$. From Appendix C, Table 3 we get the critical point for t with 23 degrees of freedom: 2.069. Applying Equation 9-32, we get:

$$5,296.05 \pm (2.069)(318.16)\sqrt{1 + 1/25 + (4,000 - 3,177.92)^2/40,947,557.84}$$

$$= 5,296.05 \pm 676.62 = [4,619.43, 5,972.67]$$

Based on the validity of the study, we are 95% confident that a cardholder who traveled 4,000 miles during a period of the given length will have charges on his or her card totaling anywhere from \$4,619.43 to \$5,972.67. What about the *average* total charge of all cardholders who traveled 4,000 miles? This is $E(Y|x = 4,000)$. The point estimate of $E(Y|x = 4,000)$ is also equal to $\hat{Y}$, but the confidence interval for this quantity is different.

A Confidence Interval for the Average Y, Given a Particular Value of X

We may compute a confidence interval for $E(Y|X)$, the expected value of Y for a given X . Here the variation is smaller because we are dealing with the average Y for a given X , rather than a particular Y . Thus, the confidence interval is narrower than a prediction interval of the same confidence level. The confidence interval for $E(Y|X)$ is given in Equation 9-33.

A $(1 - \alpha)100\%$ confidence interval for $E(Y|X)$:

$$\hat{y} \pm t_{\alpha/2}s\sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{SS_x}} \quad (9-33)$$

The confidence band for $E(Y|X)$ around the regression line looks like Figure 9-29 except that the band is narrower. The standard error of the estimator of the conditional mean, $E(Y|X)$, is smaller than the standard error of the predicted Y . Therefore, the 1 is missing from the square root quantity in Equation 9-33 as compared with Equation 9-32.

For the American Express example, let us now compute a 95% confidence interval for $E(Y|x = 4,000)$. Applying Equation 9-33, we have:

$$5,296.05 \pm (2.069)(318.16)\sqrt{1/25 + (4,000 - 3,177.92)^2/40,947,557.84} =$$

$$5,296.05 \pm 156.48 = [5,139.57, 5,452.53]$$

Being a confidence interval for a conditional mean, the interval is much narrower than the prediction interval, which has the same confidence level for covering *any given* observation at the level of X .