

In this section we consider the problem of finding a prediction interval for the actual value of  $Y$  at  $x^*$ , a given value of  $X$ . Consider, for example, a study of the relation between level of work pay ( $X$ ) and worker productivity ( $Y$ ). The actual productivity at high and medium levels of piecework pay may be of particular interest for purposes of analyzing the benefits obtained from an increase in the pay.

### Important Notes:

1.  $E(Y \mid X = x^*)$ , the expected value or average value of  $Y$  for a given value  $x^*$  of  $X$ , is what one would expect  $Y$  to be in the long run when  $X = x^*$ . Thus,  $E(Y \mid X = x^*)$  is a fixed but unknown quantity, whereas  $Y$  can take a number of values when  $X = x^*$ .
2.  $E(Y \mid X = x^*)$ , the value of the regression line at  $X = x^*$ , is entirely different from  $Y^*$ , a single realization of  $Y$  when  $X = x^*$ . In particular,  $Y^*$  need not lie on the population regression line.
3. A *confidence interval* is always reported for a parameter (e.g.,  $E(Y \mid X = x^*) = \beta_0 + \beta_1 x^*$ ), whereas a *prediction interval* is reported for the value of a random variable (e.g.,  $Y^*$ ).

### Prediction Interval for $Y^*$ at $X = x^*$

We base the prediction of  $Y$  when  $X = x^*$  (that is, of  $Y^*$ ) on

$$\hat{y}^* = \hat{\beta}_0 + \hat{\beta}_1 x^*.$$

The error in the prediction is

$$Y^* - \hat{y}^* = (\beta_0 + \beta_1 x^* + \varepsilon^*) - \hat{y}^* = \underbrace{E(Y \mid X = x^*) - \hat{y}^*}_{\text{estimation error}} + \underbrace{\varepsilon^*}_{\text{random fluctuation}}.$$

That is, the prediction error consists of the deviation between  $E(Y \mid X = x^*)$  and  $\hat{y}^*$  plus the random error  $\varepsilon^*$ , which represents the deviation of  $Y^*$  from its expected value. Thus, the variability in predicting a single  $Y^*$  will exceed that of estimating the mean response  $E(Y \mid X = x^*)$ .

### Distribution of $Y^*$

Under the assumptions of the simple linear regression model, it can be shown that

$$E(Y^* - \hat{y}^*) = E(Y - \hat{y} \mid X = x^*) = 0$$

and

$$\text{Var}(Y^* - \hat{y}^*) = \text{Var}(Y - \hat{y} \mid X = x^*) = \sigma^2 \left( 1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{XX}} \right).$$

Therefore,

$$Y^* - \hat{y}^* \sim N \left( 0, \sigma^2 \left( 1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{XX}} \right) \right).$$

### Standardization and Test Statistic

Standardizing and replacing  $\sigma$  by  $\hat{\sigma}$  gives

$$T = \frac{Y^* - \hat{y}^*}{\hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{XX}}}} \sim t_{n-2}.$$

### Prediction Interval

Thus, a  $100(1 - \alpha)\%$  prediction interval for  $Y^*$ , the actual value of  $Y$  at  $X = x^*$ , is

$$\hat{y}^* \pm t_{\alpha/2, n-2} \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{XX}}}.$$

#### Example 1

Continued with the Production data. Find a 95% prediction interval for the production run time for orders that have 200 items produced.

#### Example 2

Continued with the FreshmanGPA data. Find a 95% prediction interval for the GPA at the end of the freshman year for students who scored 30 on the ACT.