

## Section 3.5

Suppose  $\hat{y} = b_0 + b_1x$  is a *sample* regression equation

### Two Questions

- For a given value of  $x$ , say  $x_m$ , what is the mean of all the possible values of  $y$ ?
- For a given value of  $x$ , say  $x_m$ , what can we expect  $y$  to be?

**Answer** In both cases, plug  $x_m$  into the regression equation.

$$\hat{y}_m = b_0 + b_1x_m$$

**Example** Go back to the example on disposable income vs food expenditure.

- Find a point estimate of the average food expenditure of a family whose disposable income is \$33,000.
  
- Find a point prediction of the food expenditure of a family whose disposable income is \$33,000.

**Remark** Notice the terminology: point *estimate* vs point *prediction*

- Point estimate means we are estimating a *population* parameter (in this case, we use  $\hat{y}_m$  to estimate  $\mu_{y|x_m}$ ).
- Point prediction means we are approximating a single data point (in this case, a single value of  $y_m$  that goes with  $x_m$ )

**Confidence Intervals for the conditional mean of  $y$  given  $x$  ( $\mu_{y|x}$ )**

### Some Theory

- Let  $x_m$  be a fixed value of  $x$
- Then  $\hat{y}_m = b_0 + b_1x_m$  is the point estimate for the conditional mean of  $y$  given  $x_m$

- $\sigma_m^2$  is the *population variance* of all estimates of  $\hat{y}_m$ , ie, the variance of the population of all sample regression lines evaluated at  $x_m$ . It is assumed that the population of all sample regression lines are constructed from points with fixed  $x$  coordinates  $x_1, \dots, x_n$ .

$$\sigma_m^2 = \sigma_e^2 \left( \frac{1}{n} + \frac{(x_m - \bar{x})^2}{(n-1)s_x^2} \right) = \sigma_e^2 \left( \frac{1}{n} + \frac{(x_m - \bar{x})^2}{S_{xx}} \right)$$

- Estimate  $\sigma_m^2$  by the sample variance

$$s_m^2 = s_e^2 \left( \frac{1}{n} + \frac{(x_m - \bar{x})^2}{(n-1)s_x^2} \right) = s_e^2 \left( \frac{1}{n} + \frac{(x_m - \bar{x})^2}{S_{xx}} \right)$$

### Confidence Intervals for the Conditional Mean

$$\hat{y}_m - t_{\alpha/2} s_m, \hat{y}_m + t_{\alpha/2} s_m$$

where  $t$  has  $n - 2$  degrees of freedom.

**Example** Construct a 95% confidence interval estimate for the mean food expenditure for a family whose disposable income is \$33,000.

### Prediction Intervals for an individual value of $y$ given $x$

#### Some Theory

- Let  $x_p$  be a fixed value of  $x$
- Then  $\hat{y}_p = b_0 + b_1 x_p$  is the point estimate for the predicted value of  $y$  given  $x_p$
- $\sigma_p^2$  is the *population variance* of all estimates of  $\hat{y}_p$ , ie, the variance of the population of all sample regression lines evaluated at  $x_p$ . (It is assumed that the population of all sample regression lines are constructed from points with fixed  $x$  coordinates  $x_1, \dots, x_n$ .)

$$\sigma_p^2 = \sigma_e^2 \left( 1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{(n-1)s_x^2} \right) = \sigma_e^2 \left( 1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}} \right)$$

- Estimate  $\sigma_p^2$  by the sample variance

$$s_p^2 = s_e^2 \left( 1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{(n-1)s_x^2} \right) = s_e^2 \left( 1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{S_{xx}} \right)$$

## Prediction Intervals for an individual value of $y$ given $x$

$$\hat{y}_p - t_{\alpha/2} s_p, \hat{y}_p + t_{\alpha/2} s_p$$

where  $t$  has  $n - 2$  degrees of freedom.

**Example** Construct a 95% prediction interval estimate for the food expenditure for a particular family whose disposable income is \$33,000.

### Comments

- The prediction interval is longer than the confidence interval for the conditional mean.
- From the formulas we see that  $s_p^2 = s_e^2 + s_m^2$
- Even though a regression line may have a good  $r^2$  value, the regression equation still may not be adequate for making predictions.
- The  $r^2$  value is a measure of how well the line fits the points.
- You can read the section on Assessing the Quality of Prediction, but no problems will be assigned from this section.