10-4 **Variation and Prediction Intervals**

Explained and unexplained variation

In this section, we study two measures used in correlation and regression studies. (The **coefficient of determination** and the **standard error of estimate**.) We also learn how to construct a prediction interval for *y* using a regression line and a given value of *x*. To study these concepts, we need to understand and calculate the **total variation**, **explained deviation**, and the **unexplained deviation** for each ordered pair in a data set.

Assume that we have a collection of paired data containing the sample point

(*x* , *y*), that $\hat{y}$ is the predicted value of y, and that the mean of the sample *y*-values is $\bar{y}$.

The **total variation** about a regression line is the sum of the squares of the differences between the y-value of each ordered pair and the mean of y.

$$\textbf{total variation} = \sum(\boldsymbol{y} - \bar{\boldsymbol{y}})^{\boldsymbol{2}}$$

The **explained variation** is the sum of the squared of the differences between each predicted *y*-value and the mean of *y*.

$$\textbf{explained variation} = \sum(\hat{\boldsymbol{y}} - \bar{\boldsymbol{y}})^{\boldsymbol{2}}$$

The **unexplained variation** is the sum of the squared of the differences between the *y*-value of each ordered pair and each corresponding predicted *y*-value.

$$\textbf{unexplained variation} = \sum(\boldsymbol{y} - \hat{\boldsymbol{y}})^{\boldsymbol{2}}$$

The sum of the explained and unexplained variations is equal to the total variation.

Total variation = Explained variation + Unexplained variation

As its name implies, the *explained variation* can be explained by the relationship between *x* and y. The *unexplained variation* cannot be explained by the relationship between *x* and *y* and is due to chance or other variables.

Consider the advertising and sales data used throughout this section with a regression line of $\hat{y}$ = 50.729 *x* + 104.061.

Using the data point (2.0, 220) we can find the total, explained, and unexplained variation:

The Coefficient of determination

The **coefficient of determination r^2** is the ratio of the explained variation to the total variation.

$$r^2 = \frac{explaiend\ variation}{total\ variation}$$

We can compute r^2 by using the definition or by squaring the linear correlation coefficient *r*.

Ex 1)

The correlation coefficient for the following advertising expenses and company sales data is 0.913. Find the coefficient of determination. What does this tell you about the explained variation of the data about the regression line? About the unexplained variation? (r= 0.913 suggests a strong positive linear correlation)

r^2= 0.834

About 83.4% of the variation in the company sales can be explained by the variation in the advertising expenditures. About 16.6% of the variation is unexplained and is due to chance or other variables.

Advertising expenses (1000s of $), x	Company sales (1000s of $), y	xy	x^2	y^2
2.4	225	540	5.76	50,625
1.6	184	294.4	2.56	33,856
2.0	220	440	4	48,400
2.6	240	624	6.76	57,600
1.4	180	252	1.96	32,400
1.6	184	294.4	2.56	33,856
2.0	186	372	4	34,596
2.2	215	473	4.84	46,225
Sums				
15.8	1634	3289.8	32.44	337,558
$\bar{y}$ =(1634/8) =204.25	$\bar{x}$ =(15.8/8)=1.975 ,			

The Standard Error of Estimate

The **Standard Error of Estimate s_e** is the standard deviation of the observed y-values about the predicted $\hat{y}$-value for a given x-value. It is given by

$$s_e = \sqrt{\frac{\sum(y - \widehat{y})^2}{n - 2}}$$

Or as the following equivalent formula:

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

Ex 2)

The regression equation of the advertising expenses and company sales data in example 1) is

$\hat{y}$ = 50.729 *x* + 104.061

Find the standard error of estimate.

x	y	$\hat{y}$	$(y-\hat{y})^2$
2.4	225	225.81	0.6561
1.6	184	185.23	1.5129
2.0	220	205.52	209.6704
2.6	240	235.96	16.3216
1.4	180	175.08	24.2064
1.6	184	185.23	1.5129
2.0	186	205.52	381.0304
2.2	215	215.66	0.4356
Sum			635.3463

The standard error of estimate of the company sales for a specific advertising expense is about $10,290.

In chapter 7, we saw that point estimates will not give us any information about how accurate they might be. Thus, we developed confidence interval estimates to overcome this advantage. In this section we follow the same approach to construct a prediction interval.

A **prediction interval** is an interval estimate of a predicted value of *y*.

Given a linear regression equation $\hat{y} = b_0 + b_1 x$ and x_0, a specific value of x, a prediction interval for *y* is

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

Where

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

With n-2 degrees of freedom.

Ex3)

Using the results of previous example, construct a 95% prediction interval for the company sales when the advertising expenses are $2100. What can you conclude?