Objectives

- Using linear regression models

Fitting a Linear Model

Defining a Linear Relationship — Regression Analysis

Earlier, the correlation coefficient is used to measure the degree of linear relationship between two variables. However, it does not describe the exact linear association between x and y. That is the role of regression analysis. By determining a specific relationship between x and y, we may be able to use x to help predict y.

What does it mean to specify a linear relationship between two variables?

Before beginning, let's recall the equation of the line. Previously, we defined the equation of the line to be

$$y = mx + b$$

where m is the slope, and b is the y-intercept.

However, traditional statistics uses different symbols for the slope and intercept in the equation of the line. Instead of b, let b_0 be the symbol used to describe the y-intercept and b_1 be the symbol used to represent the slope of the line. Using this new set of symbols, the equation of the line becomes

$$y = b_0 + b_1 x.$$

By providing b_0 and b_1, the relationship between x and y is completely specified. The linear equation relating x to y is also referred to as a **mathematical model**. Note that the value of y is completely dependent on the value of x. Consequently, the y-variable is called the **dependent variable**. The x-variable in the model is called the **independent variable**.

The data in Figure 5.24 seem to be related. Specifying the relationship between x and y with a linear model means finding a line that best fits the data in some way. The problem is that there are many lines that could be interpreted as fitting the data.

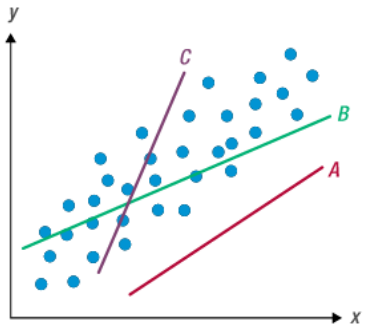
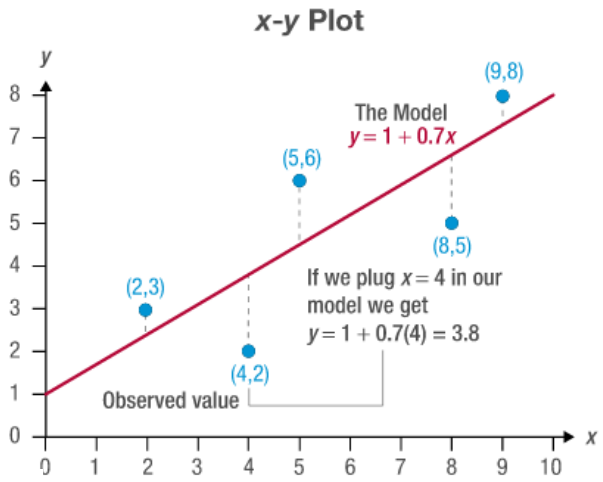


Figure 5.25

$$y = 1 + 0.7x.$$

x	y
2	3
4	2
5	6
8	5
9	8



$$\hat{y} = 1 + 0.7(2) = 2.4.$$

Figure 5.26

Table 5.4 – Observed versus Predicted Values				
Observed x	Observed y	Predicted y $\hat{y}_i = 1 + 0.7x_i$	Error $y_i - \hat{y}_i$	Squared Error $(y_i - \hat{y}_i)^2$
2	3	$2.4 = 1 + 0.7(2)$	$3 - 2.4 = 0.6$	0.36
4	2	$3.8 = 1 + 0.7(4)$	$2 - 3.8 = -1.8$	3.24
5	6	$4.5 = 1 + 0.7(5)$	$6 - 4.5 = 1.5$	2.25
8	5	$6.6 = 1 + 0.7(8)$	$5 - 6.6 = -1.6$	2.56
9	8	$7.3 = 1 + 0.7(9)$	$8 - 7.3 = 0.7$	0.49
			$\sum (y_i - \hat{y}_i) = -0.6$	$\text{SSE} = \sum (y_i - \hat{y}_i)^2 = 8.90$

error, estimated error, or residual (e_i).

$e_1 =$ Observed $y -$ Predicted $y = 3 - 2.4 = 0.6.$

$y = b_0 + b_1 x +$ error.

$\hat{y} = b_0 + b_1 x.$

error $= y - \hat{y}.$

Formula: Sum of Squared Errors (SSE)

The **sum of squared errors (SSE)** is given by

$$\text{SSE} = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - (b_0 + b_1 x_i))^2.$$

Want the smallest.

Formula: Slope and y-Intercept of the Least Squares Line

The equation for finding the slope is given by

$$b_1 = \frac{SS_{xy}}{SS_{xx}}$$

where

$$SS_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - \frac{\left(\sum x_i\right)\left(\sum y_i\right)}{n}$$

and

$$SS_{xx} = \sum (x_i - \bar{x})^2 = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}.$$

The slope can also be calculated using

$$b_1 = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2}.$$

The estimate of the intercept is given by

$$b_0 = \bar{y} - b_1\bar{x} = \frac{1}{n}\left(\sum y_i - b_1 \sum x_i\right).$$

The x_i and y_i referred to in the expressions are the observed data values of x and y, respectively.

Table 5.5 – Weekly Production		
Week	Items Produced	Cost ($)
1	22	3500
2	30	3800
3	36	4500
4	41	4200
5	27	3700
6	45	4600
7	30	3600
8	37	4550
9	32	3990
10	31	3675

$$\sum x_i = 331$$
$$\sum y_i = 40,115$$
$$\sum x_i y_i = 1,350,055$$
$$\sum x_i^2 = 11,369$$

$$b_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2} = \frac{10(1350055) - 331(40115)}{10(11369) - (331)^2} = \frac{222485}{4129} \approx 53.8835$$

$$b_0 = \bar{y} - b_1\bar{x} = \frac{1}{n}\left(\sum y_i - b_1\sum x_i\right) = \frac{1}{10}(40115 - 53.8835(331)) \approx 2227.9562.$$

Estimated Cost $= \$2227.96 + \$53.88$ (Items Produced)

↑ Dependent Variable

↑ Independent Variable

Interpreting the Regression Equation

Using the data and the estimated regression model, if the number of items produced is 0, then the predicted cost of production would be \$2227.96, the value of b_0.

$$\text{Estimated Cost} = \$2227.96 + \$53.88(0) = \$2227.96$$

The value of b_0 is always interpreted as the average value of the dependent variable, in this case, the total production cost, when the independent variable is set equal to zero. Since b_1 is the estimated slope of the line, it is interpreted as the average change in the dependent variable (total production cost) for a one unit change in the independent variable (items produced). In our example, the independent variable is expressed in units. Therefore, for every additional unit, the total production cost is expected to increase by \$53.88. If this interpretation is correct, the model's predicted cost for 27 items should be \$53.88 more than that for 26 items.

Table 5.7 – Difference in Production Cost

Predicted cost for producing 27 items	\$3682.72
Predicted cost for producing 26 items	\$3628.84
Difference	\$53.88

Definition

The **intercept coefficient**, b_0, is the average value of the dependent variable, y, when the independent variable, x, is equal to zero.

The **slope coefficient**, b_1, is the average change in the dependent variable, y, for a one-unit change in the independent variable, x.

The Importance of Errors

The usefulness of the estimated model depends on the magnitude of the prediction errors you expect the model to produce. The production model is

$$\text{Cost} = \beta_0 + \beta_1 \left(\text{Items Produced}\right) + \varepsilon_i.$$

Yet we ignored the error component when we predicted the production cost for different numbers of items. For instance, when we predicted the price of producing 26 items, we found

$$\text{Estimated Cost} = \$2227.96 + \$53.88(26) = \$3628.84.$$

Since the model is not going to be a perfect predictor, we should incorporate the possibility of error in the model. Thus,

$$\text{Estimated Cost} = \$2227.96 + \$53.88(26) + e_i = \$3628.84 + e_i$$

would have been a more precise statement. It is important to assess the magnitude of the error when the model is used for predictive purposes. If the errors are too large, then it will not be advantageous to use the model for prediction.

How Do We Summarize the Errors a Model Produces?

Computing the variation of the error data is not much different from computing the variation of any data set. Recall that the formula for sample variance is

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}.$$

In regression, we are discussing errors. For simple linear regression the degrees of freedom for the errors are $(n - 2)$, where the "2" represents the number of parameters estimated in the model (i.e., β_0 and β_1 were estimated). Making this slight adjustment for degrees of freedom produces the definition for the **variance of the error terms**, also known as the **mean squared error**.

Formula: Mean Squared Error

The variance of the error terms is also known as the **mean squared error** and is given by

$$s_e^2 = \frac{\sum (y_i - \hat{y}_i)^2}{n - 2} = \frac{\text{SSE}}{n - 2}.$$

Evaluating the Fit of a Model

The goal in constructing most linear models is to use the independent variable, x, to explain or predict the dependent variable, y. The question we want to consider is, how much of the variation in y can be explained with the model? Before determining how much variation the model explains, it will be necessary to evaluate how much variability exists in the y-variable. This quantity is called the **total sum of squares (TSS)** and represents the total variation in the dependent variable, y.

Formula: Total Sum of Squares (TSS)

The total variation in y is given by the **total sum of squares (TSS)**.

$$\text{TSS} = \sum (y_i - \bar{y})^2$$

What Is an Error?

An error $(y_i - \hat{y}_i)$ represents the model's inability to predict the variation in the dependent variable, y. If y didn't vary, for example if all y's were 6, its value would be easy to predict and the model's errors would all be zero. Adding all of the squared errors accumulates the total of all *unexplained* variation.

$$\text{SSE} = \sum (y_i - \hat{y}_i)^2 \qquad \text{TSS} = \sum (y_i - \bar{y})^2$$

The variation in y can be divided into two categories, unexplained and explained. Total variation must equal unexplained variation plus explained variation.

$$\text{TSS} = \text{Unexplained Variation} + \text{Explained Variation}$$

or

$$\text{TSS} = \text{SSE} + \text{Explained Variation}$$

Denoting explained variation as **SSR (sum of squares of regression)** produces

$$\text{TSS} = \text{SSE} + \text{SSR}$$ (TSS is the total unexplained and explained variation in y).

Solving this equation for SSR results in

$$\text{SSR} = \text{TSS} - \text{SSE}$$ (the explained variation, SSR, is equal to the total variation minus the unexplained variation).

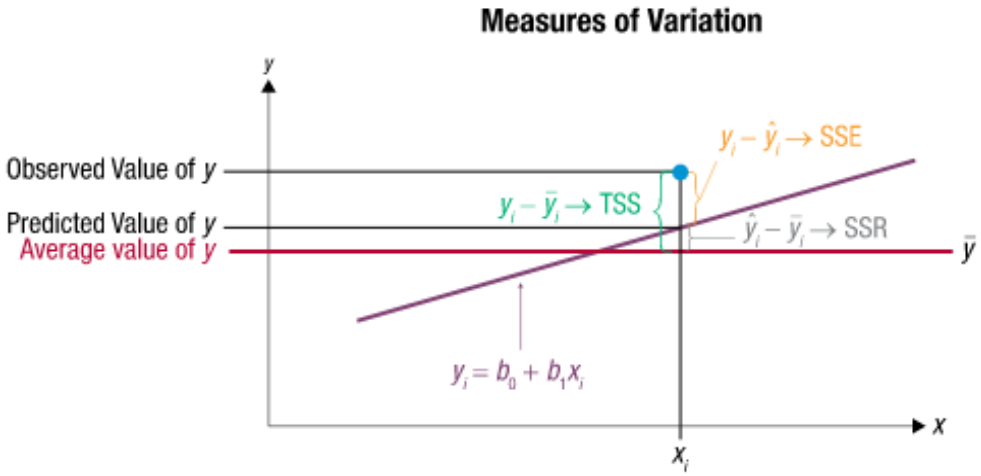

Measures of Variation
y
Observed Value of y
Predicted Value of y
Average value of y
$y_i - \hat{y}_i \rightarrow$ SSE
$y_i - \bar{y}_i \rightarrow$ TSS
$\hat{y}_i - \bar{y}_i \rightarrow$ SSR
$\bar{y}$
$y_i = b_0 + b_1 x_i$
x
x_i

Figure 5.28

ANOVA

	df	*SS*	*MS*	*F*
Regression	1	1198827.203	1198827	24.09247
Residual	8	398075.2967	49759.41	
Total	9	1596902.5		

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	2227.955922	370.1487713	6.019082	0.000317
Items Produced	53.8835069	10.97779658	4.908408	0.001181

Formula: Coefficient of Determination

The **coefficient of determination**, R^2, is given by

$$R^2 = \frac{\text{SSR}}{\text{TSS}} = 1 - \frac{\text{SSE}}{\text{TSS}}.$$

The coefficient of determination is a value between 0 and 1, inclusive. That is, $0 \leq R^2 \leq 1$.

The R^2 measurement summarizes the degree of fit on a standardized scale. The largest value R^2 can attain is 1, which will occur when the model explains all of the variation in y and consequently SSR=TSS.The smallest value of R^2 is 0, which occurs when the model does not explain any of the variation in y and consequently SSR= 0. Thus, the R^2 value is the proportion of the variation in y explained by the model.

For the production data,

$$R^2 = \frac{1{,}198{,}827.2033}{1{,}596{,}902.5} \approx 0.7507.$$

In other words, the estimated model explains about 75 percent of the variation in costs. That's pretty good. One of the interesting features of the R^2 statistic is the ability to compare the fits of two models. If one model explains 75 percent of the data and another explains 82 percent, then the second model is preferred, all other things being equal.