The Simple Regression Model

- In every regression study there is a single variable that we are trying to explain or predict, called the **dependent** variable (also called the **response** variable or the **target** variable).
- To help explain or predict the dependent variable, we use one or more **explanatory** variables (also called **independent** variables or **predictor** variables).
- If there is a single explanatory variable, the analysis is called **simple regression**.
- If there are several explanatory variables, it is called **multiple regression**

- The dependent (or response or target) variable is the single variable being explained by the regression. The explanatory (or independent or predictor) variables are used to explain the dependent variable

- A simple regression analysis includes a single explanatory variable, whereas multiple regression can include any number of explanatory variables.

SCATTERPLOTS: GRAPHING RELATIONSHIPS

- A good way to begin any regression analysis is to draw one or more scatterplots.
- A scatterplot is a graphical plot of two variables, an X and a Y.
- If there is any relationship between the two variables, it is usually apparent from the scatterplot

Example

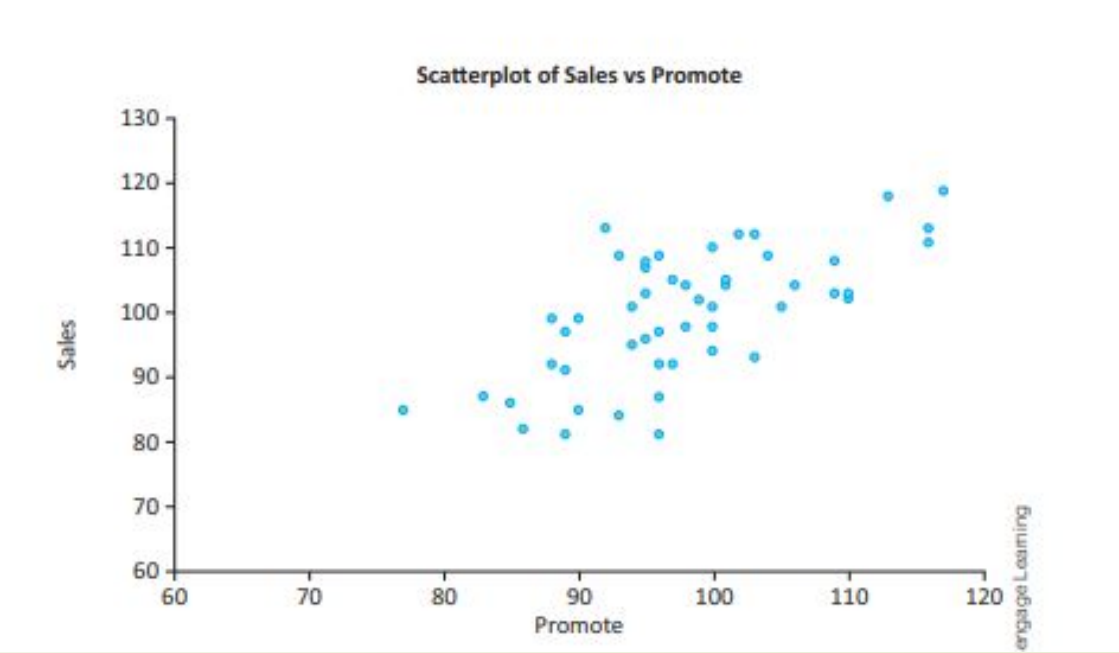
- Pharmex is a chain of drugstores that operates around the country.
- To see how effective its advertising and other promotional activities are, the company has collected data from 50 randomly selected metropolitan regions. In each region it has compared its own promotional expenditures and sales to those of the leading competitor in the region over the past year.

- There are two variables:
- ■ Promote: Pharmex's promotional expenditures as a percentage of those of the leading competitor
- ■ Sales: Pharmex's sales as a percentage of those of the leading competitor

- Note that each of these variables is an index, not a dollar amount.
- For example, if Promote equals 95 for some region, this indicates that Pharmex's promotional expenditures in that region are 95% as large as those for the leading competitor in that region.

- The company expects that there is a positive relationship between these two variables, so that regions with relatively larger expenditures have relatively larger sales.
- However, it is not clear what the nature of this relationship is.
- What type of relationship, if any, is apparent from a scatterplot?

- If it were perfect, a given value of Promote would prescribe the value of Sales exactly.
- For example, there are five regions with promotional values of 96 but all of them have different sales values.
- So the scatterplot indicates that while the variable Promote is helpful for predicting Sales, it does not lead to perfect predictions.



- This scatterplot indicates that there is indeed a positive relationship between Promote and Sales—the points tend to rise from bottom left to top right—but the relationship is not perfect.

Outliers

- Scatterplots are especially useful for identifying outliers, observations that lie outside the typical pattern of points.
- The scatterplot in Figure shows annual salaries versus years of experience for a sample of employees at a particular company.

- There is a clear linear relationship between these two variables—for all employees except the point at the top right.
- A closer look at the data reveals that this one employee is the company CEO, whose salary is well above that of all the other employees.

- An outlier is an observation that falls outside of the general pattern of the rest of the observations.

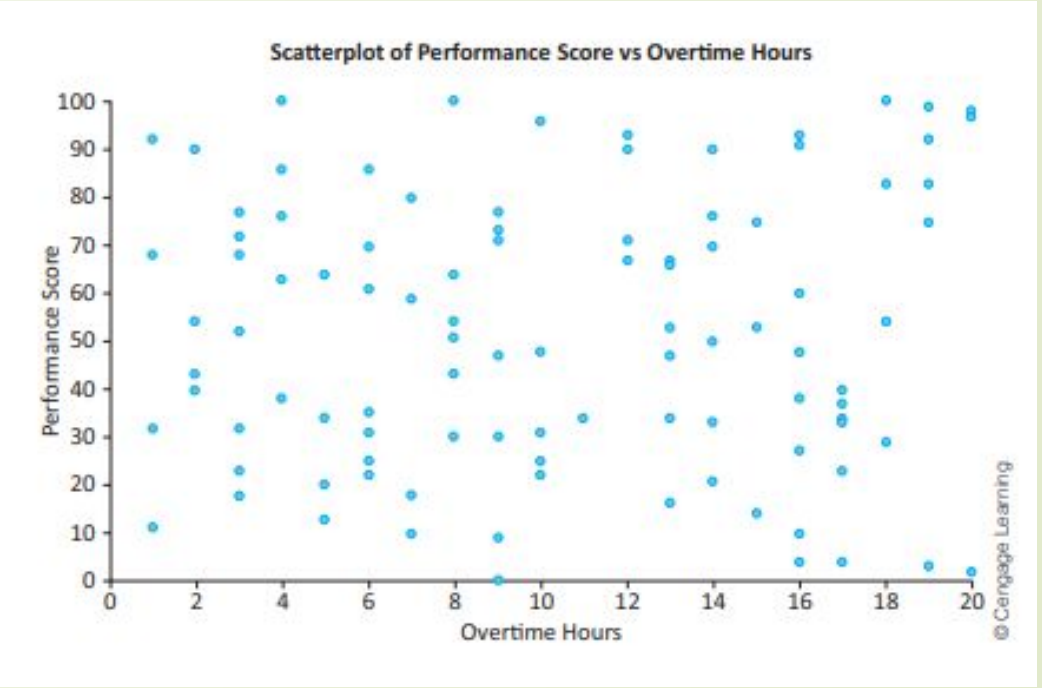
- Although scatterplots are good for detecting outliers, they do not necessarily indicate what you ought to do about any outliers you find.
- This depends entirely on the particular situation.
- If you are attempting to investigate the salary structure for typical employees at a company, then you should probably not include the company CEO.

- First, the CEO's salary is not determined in the same way as the salaries for typical employees.
- Second, if you do include the CEO in the analysis, it can greatly distort the results for the mass of typical employees.
- In other situations, however, it might not be appropriate to eliminate outliers just to make the analysis come out more nicely.

- It is difficult to generalize about the treatment of outliers, but the following points are worth noting.
- ■ If an outlier is clearly not a member of the population of interest, then it is probably best to delete it from the analysis. This is the case for the company CEO in Figure.
- ■ If it isn't clear whether outliers are members of the relevant population, you can run the regression analysis with them and again without them. If the results are practically the same in both cases, then it is probably best to report the results with the outliers included. Otherwise, you can report both sets of results with a verbal explanation of the outliers

No Relationship

- A scatterplot can provide one other useful piece of information: It can indicate that there is no relationship between a pair of variables, at least none worth pursuing.
- This is usually the case when the scatterplot appears as a shapeless swarm of points, as illustrated in Figure.



- Here the variables are an employee performance score and the number of overtime hours worked in the previous month for a sample of employees.
- There is virtually no hint of a relationship between these two variables in this plot, and if these are the only two variables in the data set, the analysis can stop right here.

CORRELATIONS: INDICATORS OF LINEAR RELATIONSHIPS

- Scatterplots provide graphical indications of relationships, whether they are linear, nonlinear, or essentially nonexistent.
- Correlations are numerical summary measures that indicate the strength of linear relationships between pairs of variables.
- A correlation between a pair of variables is a single number that summarizes the information in a scatterplot.
- A correlation can be very useful, but it has an important limitation: It measures the strength of linear relationships only.

- The usual notation for a correlation between two variables **X** and **Y** is r_{XY} .
- The formula for r_{XY} is given by

Formula for Correlation

$$r_{XY} = \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})/(n-1)}{s_X s_Y}$$

- Note that it is a sum of products in the numerator, divided by the product $s_x s_Y$ of the sample standard deviations of **X** and **Y**.

- The numerator of Equation is also a measure of association between two variables X and Y, called the ***covariance*** between X and Y.
- Like a correlation, a covariance is a single number that measures the strength of the linear relationship between two variables.
- By looking at the sign of the covariance or correlation—plus or minus—you can tell whether the two variables are positively or negatively related.
- The drawback to a covariance, however, is that its magnitude depends on the units in which the variables are measured.

- All correlations are between −1 and +1, inclusive.
- The sign of a correlation, plus or minus, determines whether the linear relationship between two variables is positive or negative.
- In this respect, a correlation is just like a covariance.
- However, the strength of the linear relationship between the variables is measured by the absolute value, or magnitude, of the correlation.
- The closer this magnitude is to 1, the stronger the linear relationship is.

- A correlation equal to 0 or near 0 indicates practically no linear relationship.
- A correlation with magnitude close to 1, on the other hand, indicates a strong linear relationship.
- At the extreme, a correlation equal to −1 or +1 occurs only when the linear relationship is perfect—that is, when all points in the scatterplot lie on a straight line.

Least Squares Estimation

Fundamental Equation for Regression

Observed Value = Fitted Value + Residual

- **The least squares line** is the line that minimizes the sum of the squared residuals. It is the line quoted in regression outputs

- A simple equation is

$$y = \beta_0 + \beta_1 x + \text{e.} \qquad (1)$$

Equation (1), which is assumed to hold in the population of interest, defines the ***simple linear regression model***.

It is also called ***the two-variable linear regression model or bivariate linear regression model because*** it relates the two variables **x** and **y**.

We now discuss the meaning of each of the quantities in (1)

- The variable **e**, called **the error term** or **disturbance** in the relationship, represents factors other than **x** that affect **y**.
- A simple regression analysis effectively treats all factors affecting **y** other than **x** as being unobserved.
- Equation (1) also addresses the issue of the functional relationship between **y** and **x**.
- If the other factors in **e** are held fixed, so that the change in **e** is zero, $\Delta e = 0$, then **x** has a linear effect on **y**:

$$\Delta y = \beta_1 \Delta x \text{ if } \Delta e = 0$$

- Thus, the change in ***y*** is simply β_1 multiplied by the change in **x**.
- This means that β_1 is the slope parameter in the relationship between ***y*** and ***x,*** holding the other factors in ***e*** fixed; it is of primary interest in applied economics.
- The intercept parameter β_0, sometimes called **the constant term.**
- The linearity of (1) implies that a one-unit change in **x** has the same effect on **y**, regardless of the initial value of **x**.

Deriving the Ordinary Least Squares Estimates

- We will address the important issue of how to estimate the parameters β_0 and β_1 in equation (1).
- To do this, we need a sample from the population.
- Let $\{(x_i, y_i): i=1, \ldots, n\}$ denote a random sample of size n from the population. Because these data come from (1), we can write

$$y_i = \beta_0 + \beta_1 x_i + \text{e.}$$

for each i.

Here, e_i is the error term for observation ***i*** because it contains all factors affecting y_i other than x_i.

As an example, x_i might be the annual income and y_i the annual savings for family ***i*** during a particular year.

If we have collected data on fifteen families, then ***n=15***.

A scatterplot of such a data set is given in Figure, along with the (necessarily fictitious) population regression function

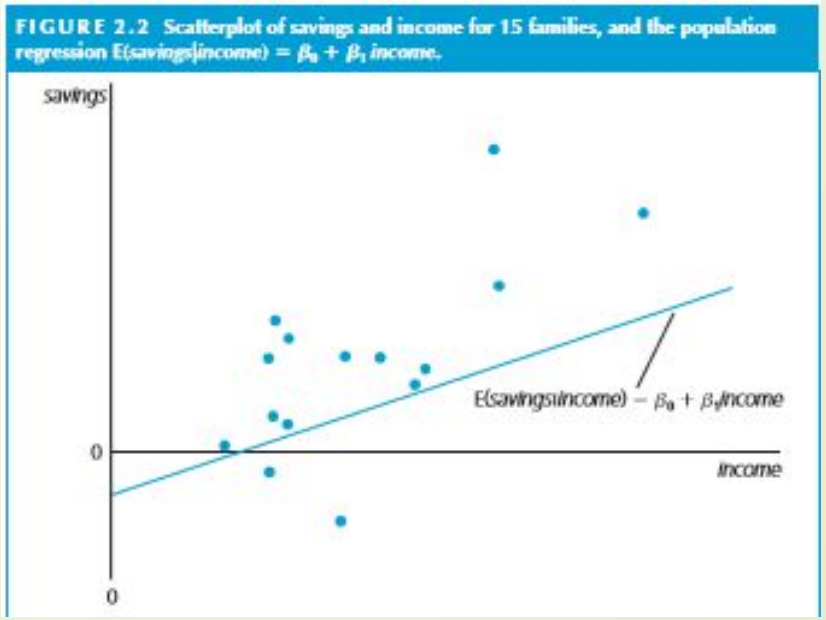


FIGURE 2.2 Scatterplot of savings and income for 15 families, and the population regression E(savings|income) = $\beta_0 + \beta_1$ income.

- We will use that **e** has zero expected value and that the covariance between **x** and **u** is zero:

E(e)=0 and Cov(x,e)= 0.

We form the ordinary least squares (OLS) regression line:

$$\hat{y} = \widehat{\beta_0} + \widehat{\beta_1}\mathrm{x}. \qquad (2)$$

where

$$\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{\mathrm{x}},$$

$$\widehat{\beta_1} = \frac{\overline{xy}}{\overline{x^2}},$$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n},\ \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n},\ \overline{xy} = \frac{\sum_{i=1}^{n} x_i y_i}{n},\ \overline{x^2} = \frac{\sum_{i=1}^{n} x_i^2}{n}.$$

- The notation $\hat{y}$, read as "**y** hat," emphasizes that the predicted values from equation (2) are estimates.
- Equation (2) is also called **the sample regression function (SRF)**.

- There are numerous residuals, it is useful to summarize them with a single numerical measure.
- This measure, called the standard error of estimate and denoted **Se** , is the standard deviation of the residuals.
- It is given by Equation

Formula for Standard Error of Estimate

$$s_e = \sqrt{\frac{\Sigma e_i^2}{n-2}}$$

- The usual empirical rules for standard deviations can be applied to the standard error of estimate.
- For example, about two-thirds of the residuals are typically within one standard error of their mean (which is zero).
- Stated another way, about two-thirds of the observed **y** values are typically within one standard error of the corresponding fitted $\hat{y}$ values.
- Similarly, about 95% of the observed **y** values are typically within two standard errors of the corresponding fitted $\hat{y}$ values

The Percentage of Variation Explained: R-Square

- R^2 is the percentage of variation of the dependent variable explained by the regression

Formula for R^2

$$R^2 = 1 - \frac{\Sigma e_i^2}{\Sigma (Y_i - \bar{Y})^2}$$

- In simple linear regression, R^2 is the square of the correlation between the dependent variable and the explanatory variable.