Linear Model

Linear model

The dependence of one variable over the other is called Linear and the equation of such dependence is called linear model

Table 14.2 (on next slide) displays data on age and price for a sample of cars of a particular make and model. We refer to the car as the Orion, but the data, obtained from the *Asian Import edition of the Auto Trader magazine, is for a real car. Ages are in years; prices* are in hundreds of dollars, rounded to the nearest hundred dollars

models

Linear Model
Twitter mining
Market basket association rule
Nearest Neighbor
Classification tree
Regression trees
Text Mining
Clustering

TABLE 14.2

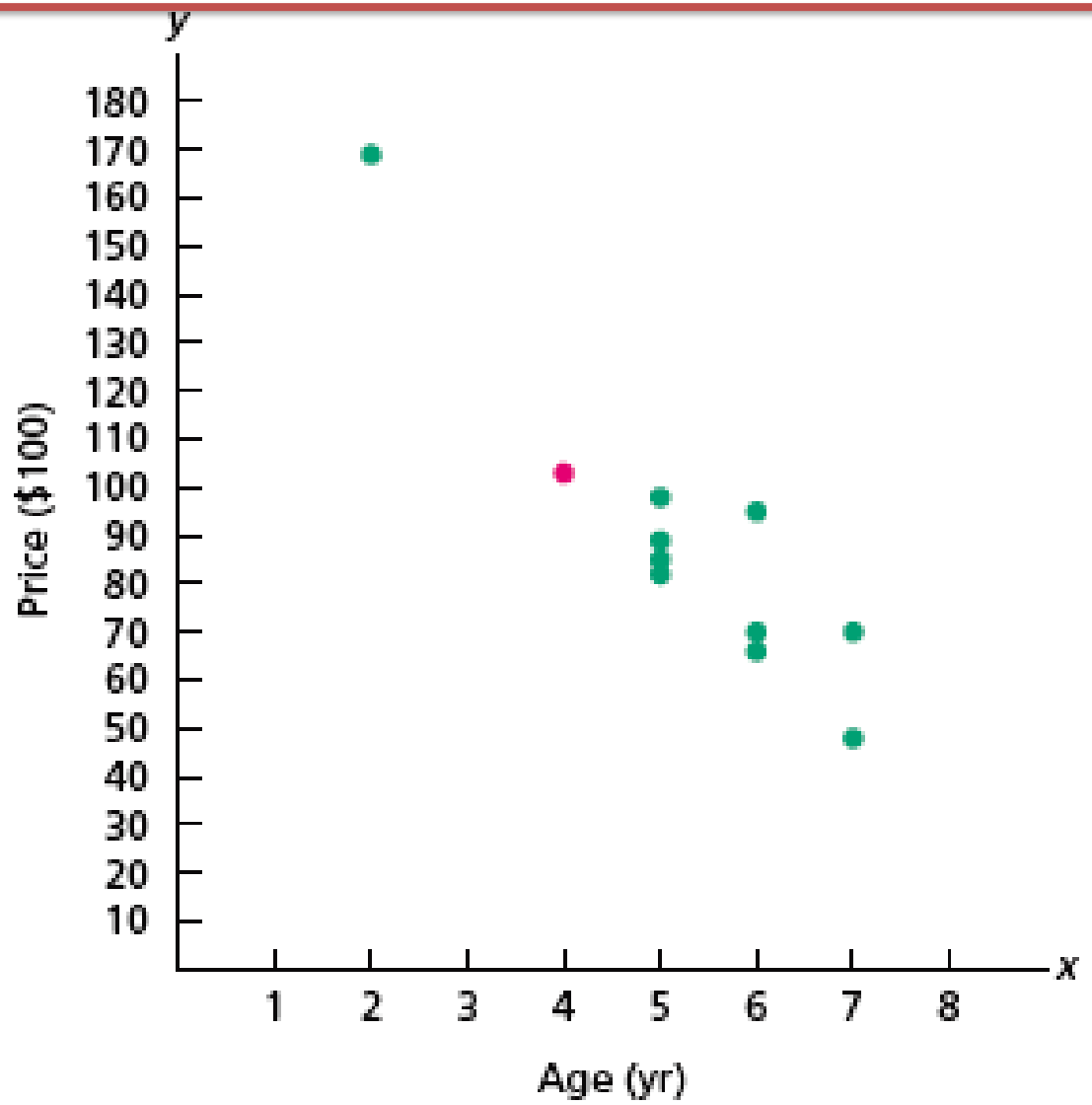
Age and price data
for a sample of 11 Orions

Car	Age (yr) x	Price ($100) y
1	5	85
2	4	103
3	6	70
4	5	82
5	5	89
6	5	98
7	6	66
8	6	95
9	2	169
10	7	70
11	7	48

Plotting the data in a *scatterplot helps us visualize any apparent relationship between* age and price. Generally speaking, a **scatterplot (or scatter diagram) is a graph** of data from two quantitative variables of a population. To construct a scatterplot, we use a horizontal axis for the observations of one variable and a vertical axis for the observations of the other. Each pair of observations is then plotted as a point. Figure 14.7 (next slide) shows a scatterplot for the age–price data in Table 14.2.

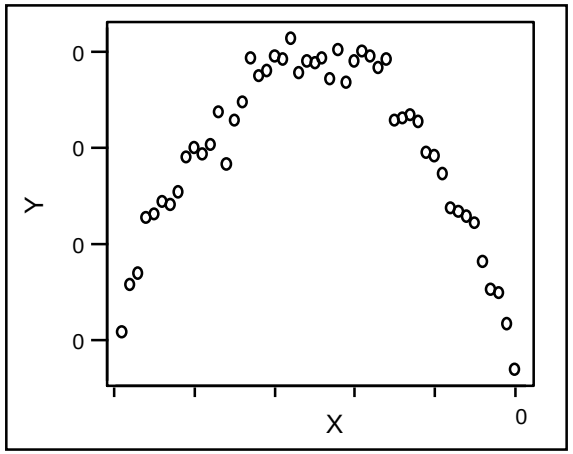
FIGURE 14.7

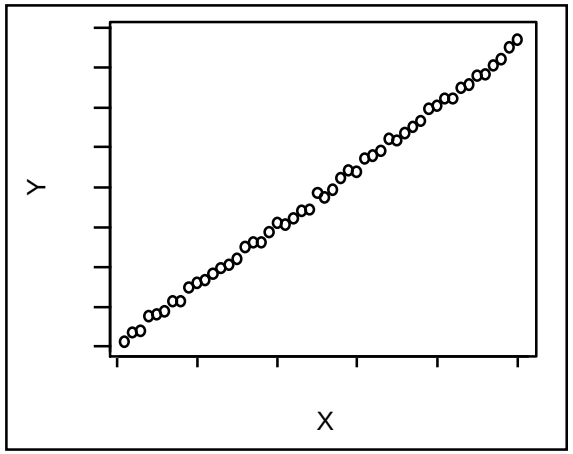
Scatterplot for the age and price data of Orions from Table 14.2

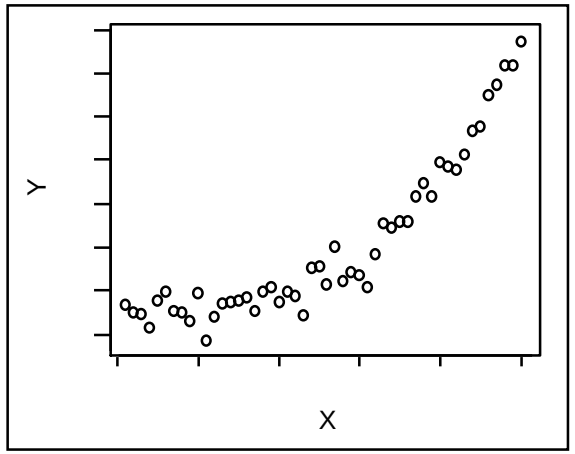


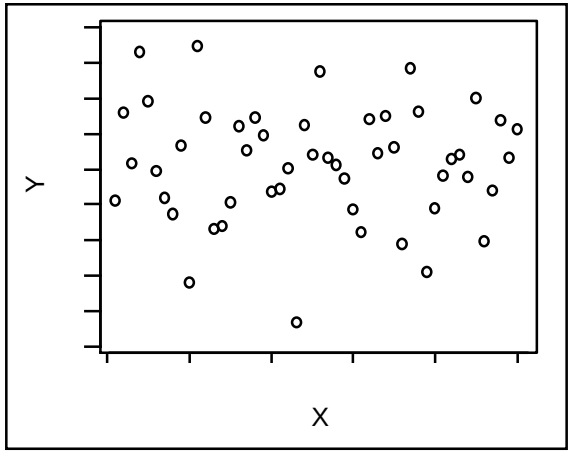
Because we could draw many different lines through the cluster of data points, we need a method to choose the “best” line. The method, called the *least-squares criterion, is based on an analysis of the errors made in using a line to fit the data points.*

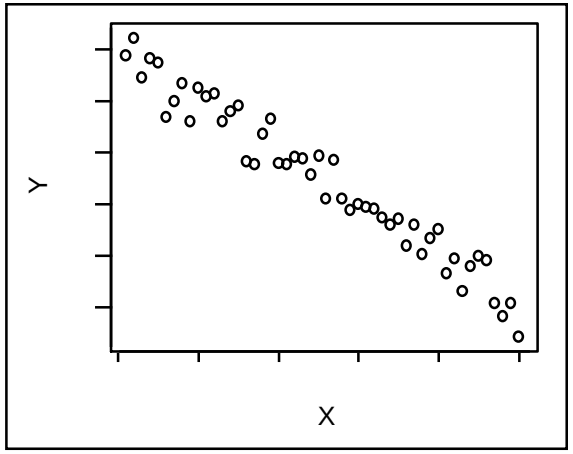
Examples of Other Scatterplots

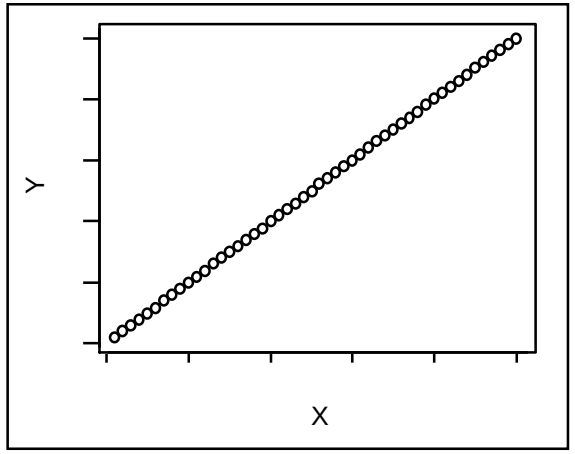












Simple Linear Model

The population simple linear model:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

where

✓ *Y* is the **dependent variable**, the variable we wish to explain or predict

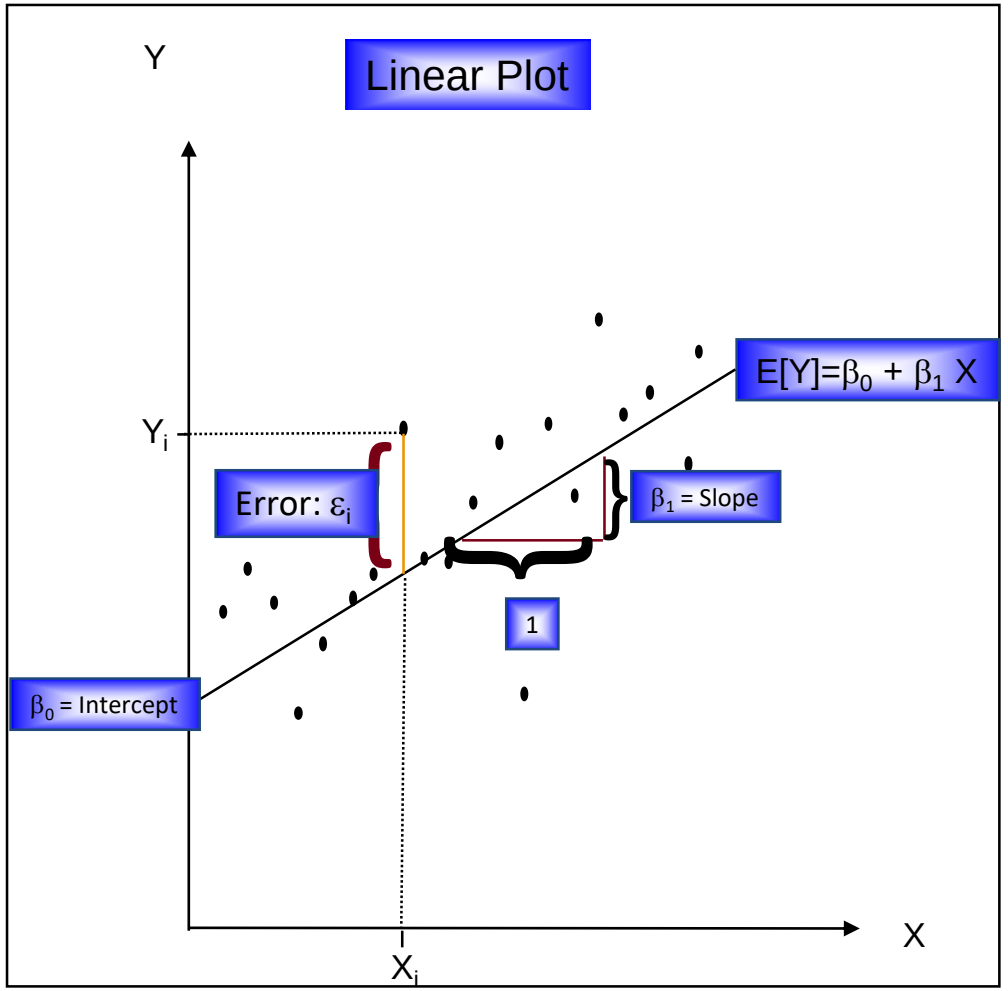
✓ *X* is the **independent variable**, also called the **predictor variable**

✓ ε is the **error term**, the only random component in the model, and thus, the only source of randomness in *Y*.

✓ β_0 is the **intercept** of the systematic component of the Linear relationship.

✓ β_1 is the **slope** of the systematic component.

Picturing the Simple Linear Model



Assumptions of Linear Model

- The relationship between X and Y is a straight-line relationship.
- The values of the independent variable X are assumed fixed (not random); the only randomness in the values of Y comes from the error term ε_i.
- The variance of Y at every value of X is the same (homogeneity of variances)
- The errors ε_i are normally distributed with mean 0 and variance σ^2. The errors are uncorrelated (not related) in successive observations. That is: $\varepsilon \sim N(0,\sigma^2)$

The Method of Least Squares

Estimation of a simple linear relationship involves finding estimated or predicted values of the intercept and slope of the linear line.

The estimated linear Linear

$$Y = b_0 + b_1X + e$$

where b_0 estimates the intercept of the population Linear line, β_0 ; b_1 estimates the slope of the population Linear line, β_1; and e stands for the observed errors - the residuals from fitting the estimated Linear line $b_0 + b_1X$ to a set of n points.

The estimated regression line:

$$\hat{Y} = b_0 + b_1 X$$

where $\hat{Y}$ (Y-hat) is the value of Y lying on the fitted regression line for a given value of X.

Estimating the intercept and slope: least squares estimation

Sum of deviation between observed and estimated values of dependent variable should be minimum

Or

Error sum of square should be minimum

$$SSE = \sum (Y_i - Y_i)^2$$

$$SSE = \sum \varepsilon^2$$

$$SSE = \sum (Y_i - \alpha - \beta X_i)^2$$

Estimating the intercept and slope: least squares estimation

Sum of deviation between observed and estimated values of dependent variable should be minimum

Or

Error sum of square should be minimum

$$\frac{dSSE}{d\beta} = \frac{d}{d\beta}\sum(Y_i - \alpha - \beta X_i)^2$$

$$\frac{dSSE}{d\alpha} = \frac{d}{d\alpha}\sum(Y_i - \alpha - \beta X_i)^2$$

Normal Equations

$$\sum_{i=1}^{n} y_i = nb_0 + b_1 \sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} x_i y_i = b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2$$

Sum of squares and cross product

$$SS_x = \sum (x - \bar{x})^2 = \sum x^2 - \frac{\left(\sum x\right)^2}{n}$$

$$SS_y = \sum (y - \bar{y})^2 = \sum y^2 - \frac{\left(\sum y\right)^2}{n}$$

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

Least – squares regressionestimators:

$$b_1 = \frac{SS_{XY}}{SS_X}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

Age and Price of Orions In the first two columns of Table 14.5, we repeat our data on age and price for a sample of 11 Orions.

---

a. Determine the Linear equation for the data.

b. Graph the Linear equation and the data points.

c. Describe the apparent relationship between age and price of Orions.

d. Interpret the slope of the Linear line in terms of prices for Orions.

e. Use the Linear equation to predict the price of a 3-year-old Orion and a 4-year-old Orion.

TABLE 14.5

Table for computing the regression equation for the Orion data

Age (yr) x	Price ($100) y	xy	x^2
5	85	425	25
4	103	412	16
6	70	420	36
5	82	410	25
5	89	445	25
5	98	490	25
6	66	396	36
6	95	570	36
2	169	338	4
7	70	490	49
7	48	336	49
58	975	4732	326

Solution

a. We first need to compute b_1 and b_0 by using Formula 14.1. We did so by constructing a table of values for x (age), y (price), xy, x^2, and their sums in Table 14.5.

The slope of the regression line therefore is

$$b_1 = \frac{S_{xy}}{S_{xx}} = \frac{\Sigma x_i y_i - (\Sigma x_i)(\Sigma y_i)/n}{\Sigma x_i^2 - (\Sigma x_i)^2/n} = \frac{4732 - (58)(975)/11}{326 - (58)^2/11} = -20.26.$$

The y-intercept is

$$b_0 = \frac{1}{n}(\Sigma y_i - b_1 \Sigma x_i) = \frac{1}{11}[975 - (-20.26) \cdot 58] = 195.47.$$

So the regression equation is $\hat{y} = 195.47 - 20.26x$.

Note: The usual warnings about rounding apply. When computing the slope, b_1, of the regression line, do not round until the computation is finished. When computing the y-intercept, b_0, do not use the rounded value of b_1; instead, keep full calculator accuracy.

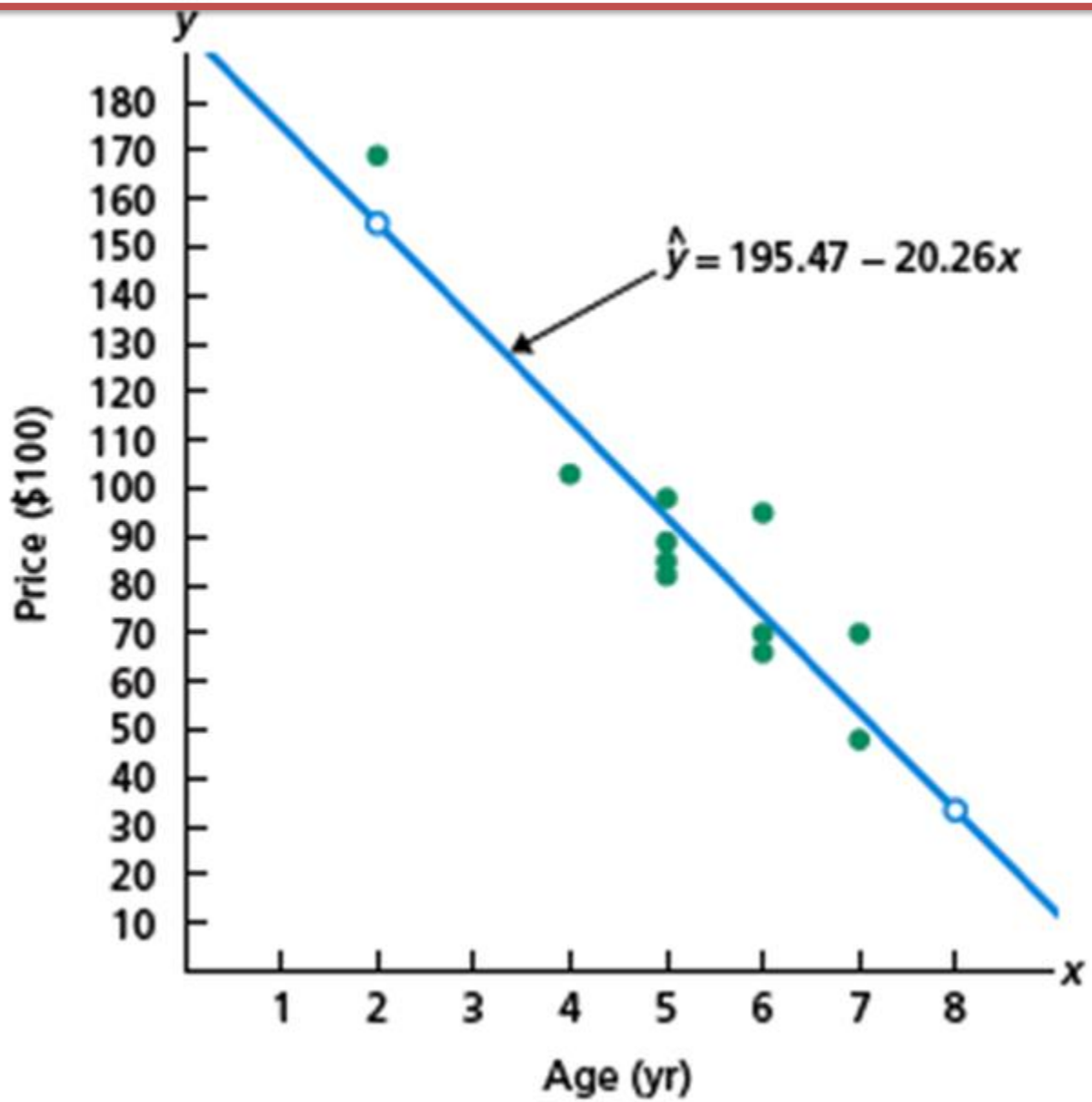
b. To graph the regression equation, we need to substitute two different x-values in the regression equation to obtain two distinct points. Let's use the x-values 2 and 8. The corresponding y-values are

$$\hat{y} = 195.47 - 20.26 \cdot 2 = 154.95 \quad \text{and} \quad \hat{y} = 195.47 - 20.26 \cdot 8 = 33.39.$$

Therefore, the regression line goes through the two points (2, 154.95) and (8, 33.39). In Fig. 14.10, we plotted these two points with open dots. Drawing a line through the two open dots yields the regression line, the graph of the regression equation. Figure 14.10 also shows the data points from the first two columns of Table 14.5.

FIGURE 14.10

Regression line and data points for Orion data



c. Because the slope of the regression line is negative, price tends to decrease as age increases, which is no particular surprise.

d. Because x represents age in years and y represents price in hundreds of dollars, the slope of -20.26 indicates that Orions depreciate an estimated \$2026 per year, at least in the 2- to 7-year-old range.

e. For a 3-year-old Orion, $x = 3$, and the regression equation yields the predicted price of

$$\hat{y} = 195.47 - 20.26 \cdot 3 = 134.69.$$

Similarly, the predicted price for a 4-year-old Orion is

$$\hat{y} = 195.47 - 20.26 \cdot 4 = 114.43.$$

Interpretation The estimated price of a 3-year-old Orion is \$13,469, and the estimated price of a 4-year-old Orion is \$11,443.

Predictor Variable and Response Variable

For a linear equation $y = b_0 + b_1x$, y is the dependent variable and x is the independent variable. However, in the context of regression analysis, we usually call y the **response variable** and x the **predictor variable** or **explanatory variable** (because it is used to predict or explain the values of the response variable). For the Orion example, then, age is the predictor variable and price is the response variable.

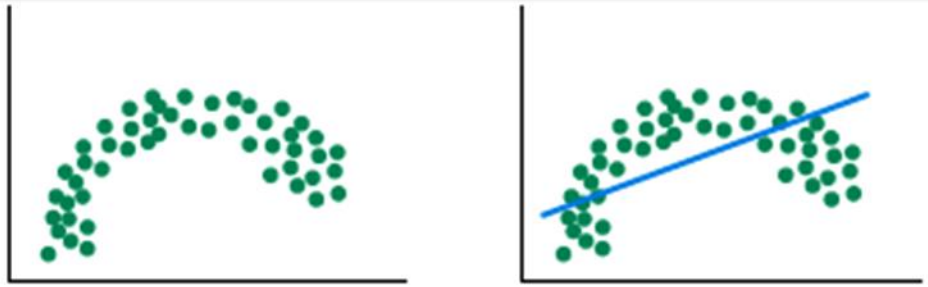
Response Variable and Predictor Variable

Response variable: The variable to be measured or observed.

Predictor variable: A variable used to predict or explain the values of the response variable.

A Warning on the Use of Linear Regression

The idea behind finding a regression line is based on the assumption that the data points are scattered about a line.† Frequently, however, the data points are scattered about a curve instead of a line, as depicted in Fig. 14.13(a).



(a) (b)

One can still compute the values of b_0 and b_1 to obtain a regression line for these data points. The result, however, will yield an inappropriate fit by a line, as shown in Fig. 14.13(b), when in fact a curve should be used. For instance, the regression line suggests that y-values in Fig. 14.13(a) will keep increasing when they have actually begun to decrease.

Criterion for Finding a Regression Line

Before finding a regression line for a set of data points, draw a scatterplot. If the data points do not appear to be scattered about a line, do not determine a regression line.

Techniques are available for fitting curves to data points that show a curved pattern, such as the data points plotted in Fig. 14.13(a). We discuss those techniques, referred to as **curvilinear regression,** in the chapter *Model Building in Regression* on the WeissStats CD accompanying this book.

Correlation coefficient

- Pearson's Correlation Coefficient is standardized covariance (unitless):

$$r = \frac{\mathrm{covariance}(x, y)}{\sqrt{\mathrm{var}\ x}\sqrt{\mathrm{var}\ y}}$$

Correlation

- Measures the relative strength of the *linear* relationship between two variables
- Unit-less
- Ranges between –1 and 1
- The closer to –1, the stronger the negative linear relationship
- The closer to 1, the stronger the positive linear relationship
- The closer to 0, the weaker any positive linear relationship

Calculating by hand…

$$\hat{r} = \frac{\mathrm{covariance}(x, y)}{\sqrt{\mathrm{var}\ x}\sqrt{\mathrm{var}\ y}} = \frac{\dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}}{\sqrt{\dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}\sqrt{\dfrac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}}$$

The End