Section 11.1 Introduction
• It is important to be able to predict values of one variable \( y \) using values of another variable \( x \).
  \[ y = \text{dependent variable} \]
  \[ x = \text{independent variable} \]

• In many cases there is evidence of a linear relationship between \( x \) and \( y \).

Example 1: There is some evidence that drinking moderate amounts of wine helps prevent heart attacks. The Minitab worksheet "May 11 lecture.mws" contains data on wine consumption (liters of alcohol from wine, per person) and yearly deaths from heart disease (deaths per 100,000 people) in 19 nations.

Make a scatterplot of this data, with \( x = \text{wine consumption} \) and \( y = \text{deaths from heart disease} \). Is there evidence of a linear relationship between the two?

• If we assume a linear relationship between \( x \) and \( y \), then
  \[ y = \beta_0 + \beta_1 x. \]  \hspace{1cm} (1)

But in reality, we can’t expect \( x \) to predict \( y \) perfectly using (1). For example there are many factors influencing rates of heart disease other than wine consumption.

A realistic prediction equation

\[ y = \beta_0 + \beta_1 x + \varepsilon \]  \hspace{1cm} (2)

where \( \varepsilon = \text{error in using (1) to predict } y \).  
\( \varepsilon \) is called a random error term.

Section 11.2 Estimating Model Parameters
• In practice, the parameters \( \beta_0, \beta_1 \) in (1) can only be estimated using a finite sample \((x_1, y_1), \ldots, (x_n, y_n)\) of data points.

• These estimates are called \( \hat{\beta}_0, \hat{\beta}_1 \), or the least-squares estimates of slope and intercept.
The predicted value of $y$ using $\hat{\beta}_0, \hat{\beta}_1$ is denoted $\hat{y}$:

$$\hat{y} =$$

Why do we use the term "least-squares"?
Formulas for $\hat{\beta}_0, \hat{\beta}_1$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$
and
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

where

$$S_{xy} = \sum_i (x_i - \bar{x})(y_i - \bar{y})$$
and
$$S_{xx} = \sum_i (x_i - \bar{x})^2.$$

Usually we will use Minitab to find $\hat{\beta}_0, \hat{\beta}_1$.

**Example 2:** In the Minitab worksheet "May 11 lecture.mws" are average degree-days and gas consumption for 16 months for a certain household. (One degree day is accumulated for each degree a day’s average temperature falls below 65°. For example, an average temperature of 20° corresponds to 45 degree-days). Gas consumption is measured in hundreds of cubic feet.

(a) What quantity should $x$ represent and what quantity should $y$ represent?

(b) Use Minitab to find $\hat{\beta}_0, \hat{\beta}_1$:
   Begin by selecting Stat—Regression—Regression.

(c) Predict the gas consumption if the degree days are 40.

(d) "Predict" the gas consumption when the degree days are 32. How does the prediction compare with the actual gas consumption when the degree days are 32?

(d) Interpret the value of $\hat{\beta}_1$ in the context of the problem.
Outliers
The parameters $\hat{\beta}_0, \hat{\beta}_1$ are greatly affected by outliers in the scatterplot.

Example 3: Does the age at which a child begins to talk predict later scores on a test of mental ability? In today’s worksheet is recorded a list of children, the age at which they began to speak, and their Gesell Adaptive Score, the result of an aptitude test taken much later.

(a) Use a fitted line plot to identify the outliers. (Choose Stat→regression→fitted line plot).

(b) Find the values of $\hat{\beta}_0, \hat{\beta}_1$
   (i) Omitting both outliers
   (ii) Omitting only child 19
   (iii) Omitting only child 18

Which outlier has the most influence on the least-squares regression line?
High leverage points
Outliers in the $x$-direction of a scatterplot are called *high leverage points*.

High influence points
• A high influence point is a high leverage point that also corresponds to an outlier in the $y$-direction.
• High influence points seriously alter the slope of the regression line.
• An outlier is NOT necessarily a point of high influence!
• Minitab identifies outliers and points of high influence.

Residuals
• Suppose we have a random sample of data points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ and a least-squares line
  \[ \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \]
  through the points.
• Let $\hat{y}_i$ be the $y$-value predicted by the $x$-value $x_i$, that is
  \[ \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \; i = 1, \ldots, n. \]
• The residuals of the data are defined by
  \[ \hat{y}_i - \bar{y}, \; i = 1, \ldots, n. \]
They are used to measure the accuracy of least-squares line.
• We define the sample variance to be
  \[ s_e^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{n-2} = \frac{SS(\text{Residual})}{n-2}. \]
• $s_e$ is called the
  ✓ sample standard deviation about the regression line, or
  ✓ the standard error of estimate, or
  ✓ the residual standard deviation.