Chapter 11: Analyzing Association Between Quantitative Variables: Regression Analysis

Section 11.1: How Can We Model How Two Variables Are Related?

Learning Objectives

1. Regression Analysis
2. The Scatterplot
3. The Regression Line Equation
4. Outliers
5. Influential Points
6. Residuals are Prediction Errors
7. Regression Model: A Line Describes How the *Mean* of y Depends on x
8. The Population Regression Equation
9. Variability about the Line
10. A Statistical Model

Learning Objective 1: Regression Analysis

- **The first step of a *regression analysis* is to identify the response and explanatory variables**
  - **We use *y* to denote the *response variable***
  - **We use *x* to denote the *explanatory variable***

Learning Objective 2: The Scatterplot

- **The first step in answering the question of association is to look at the data**
- **A *scatterplot* is a graphical display of the relationship between the response variable (y-axis) and the explanatory variable (x-axis)**

Learning Objective 2:
Example: What Do We Learn from a Scatterplot in the Strength Study?

- **An experiment was designed to measure the strength of female athletes**
- **The goal of the experiment was to find the maximum number of pounds that each individual athlete could bench press**

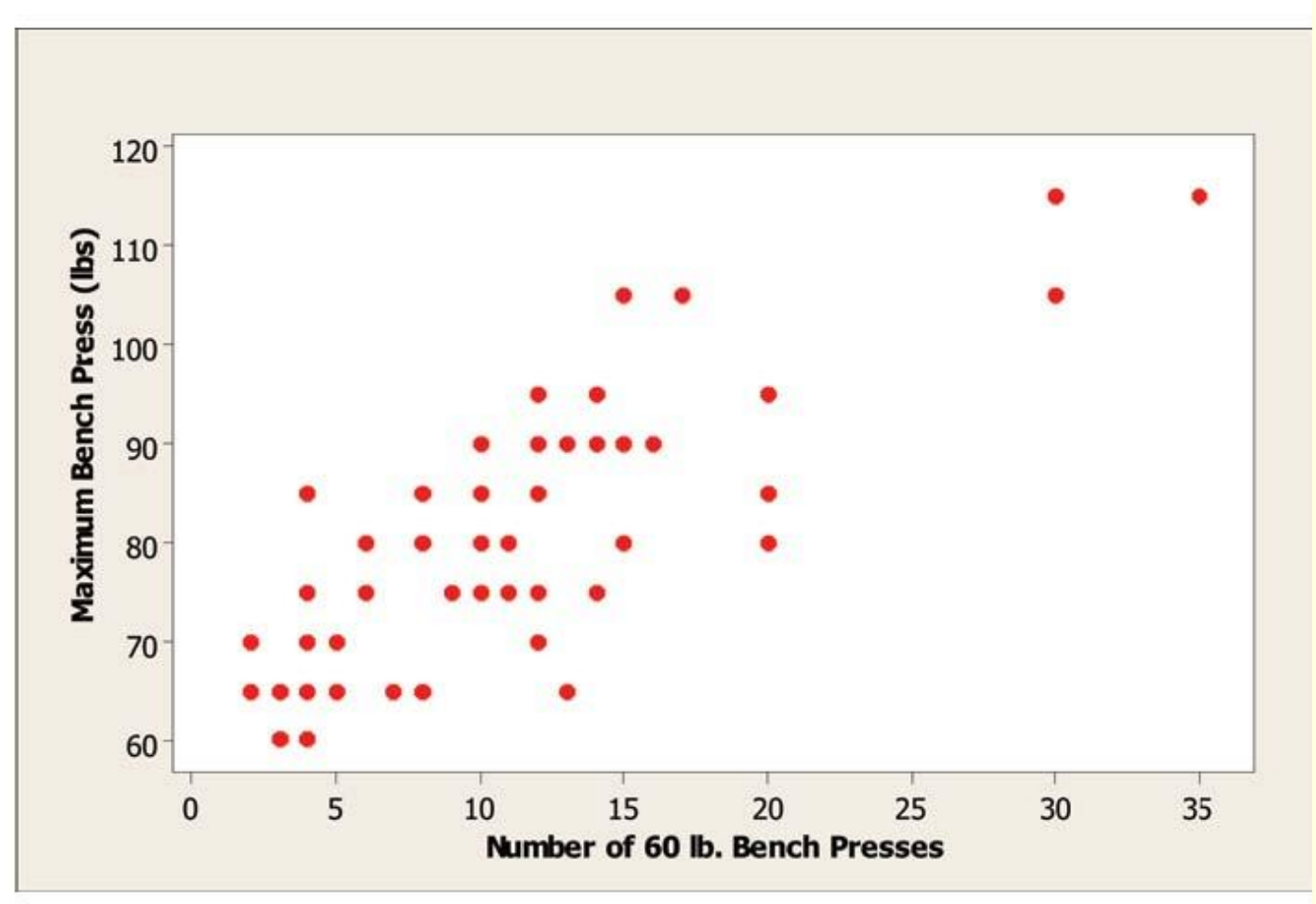
Learning Objective 2:
Example: What Do We Learn from a Scatterplot in the Strength Study?

- **57 high school female athletes participated in the study**
- **The data consisted of the following variables:**
  - **x: the number of 60-pound bench presses an athlete could do**
  - **y: maximum bench press**

Learning Objective 2:
Example: What Do We Learn from a Scatterplot in the Strength Study?

- **For the 57 girls in this study, these variable are summarized by:**
  - **x: mean = 11.0, st.deviation = 7.1**
  - **y: mean = 79.9 lbs, st.dev. = 13.3 lbs**

Learning Objective 2:
Example: What Do We Learn from a Scatterplot in the Strength Study?



Learning Objective 3: The Regression Line Equation

- **When the scatterplot shows a linear trend, a straight line can be fitted through the data points to describe that trend**
- **The regression line is:**

$$\hat{y} = a + bx$$

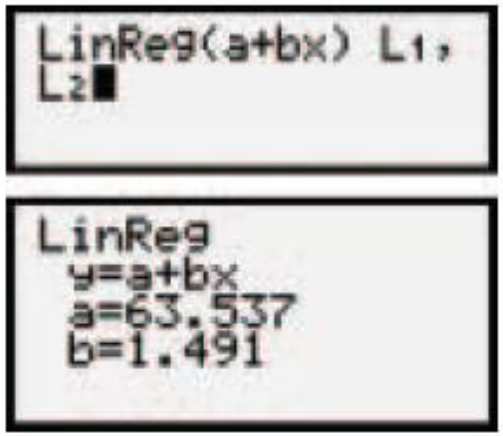
- $\hat{y}$ **is the predicted value of the response variable y**
- a **is the y-intercept and** b **is the slope**

Learning Objective 3:
Example: What Do We Learn from a Scatterplot in the Strength Study?

TABLE 12.1: MINITAB Printout for Regression Analysis of y = Maximum Bench Press (BP) and x = Number of 60-Pound Bench Presses (BP_60)

The regression equation is BP = 63.5 + 1.49 BP_60

Predictor	Coef	SE Coef	T	P
Constant	63.537	1.956	32.48	0.000
BP_60	1.4911	0.1497	9.96	0.000



Learning Objective 3:
Example: What Do We Learn from a Scatterplot in the Strength Study?

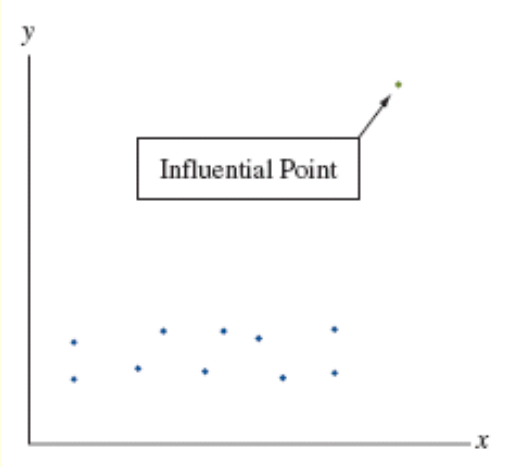
- **The MINITAB output shows the following regression equation:**
- **BP = 63.5 + 1.49 (BP_60)**
- **The y-intercept is 63.5 and the slope is 1.49**
- **The slope of 1.49 tells us that predicted maximum bench press increases by about 1.5 pounds for every additional 60-pound bench press an athlete can do**

Learning Objective 4: Outliers

- **Check for outliers by plotting the data**
- **The regression line can be pulled toward an outlier and away from the general trend of points**

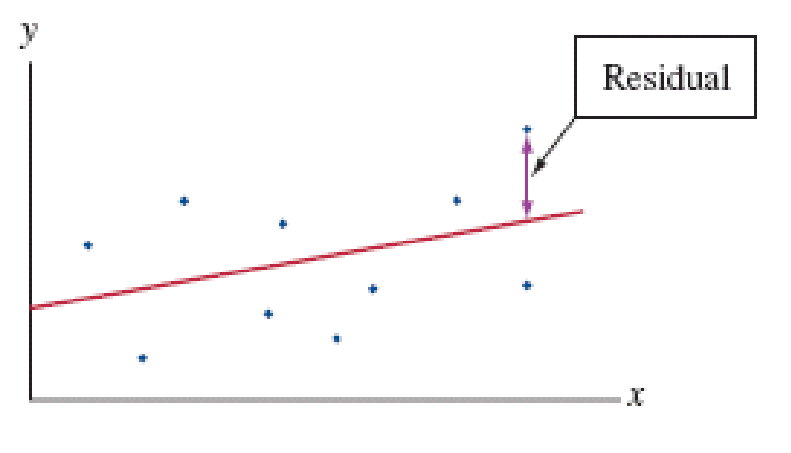
Learning Objective 5: Influential Points

- **An observation can be influential in affecting the regression line when two thing happen:**
  - **Its x value is low or high compared to the rest of the data**
  - **It does not fall in the straight-line pattern that the rest of the data have**



Learning Objective 6: Residuals are Prediction Errors

- **The regression equation is often called a prediction equation**
- **The difference $y - \hat{y}$ between an observed outcome and its predicted value is the prediction error, called a *residual***



Learning Objective 6: Residuals

- **Each observation has a residual**
- **A residual is the vertical distance between the data point and the regression line**

Learning Objective 6: Residuals

- **We can summarize how near the regression line the data points fall by**

$$sum\ of\ \ squared\ residuals =$$

$$\sum (residuals)^2 = \sum (y - \hat{y})^2$$

- **The regression line has the smallest sum of squared residuals and is called the *least squares line***

Learning Objective 7: Regression Model: A Line Describes How the *Mean* of y Depends on x

- **At a given value of x, the equation:**

$$\hat{y} = a + bx$$

  - **Predicts a single value of the response variable**
  - **But… we should not expect all subjects at that value of x to have the same value of y**
    - **Variability occurs in the y values**

Learning Objective 7: The Regression Line

- **The regression line connects the estimated *means* of y at the various x values**
- **In summary,** $\hat{y} = a + bx$

Describes the relationship between x and the *estimated means* of y at the various values of x

Learning Objective 8:
The Population Regression Equation

- **The population regression equation describes the relationship in the *population* between x and the means of y**
- **The equation is:**

$$\mu_y = \alpha + \beta x$$

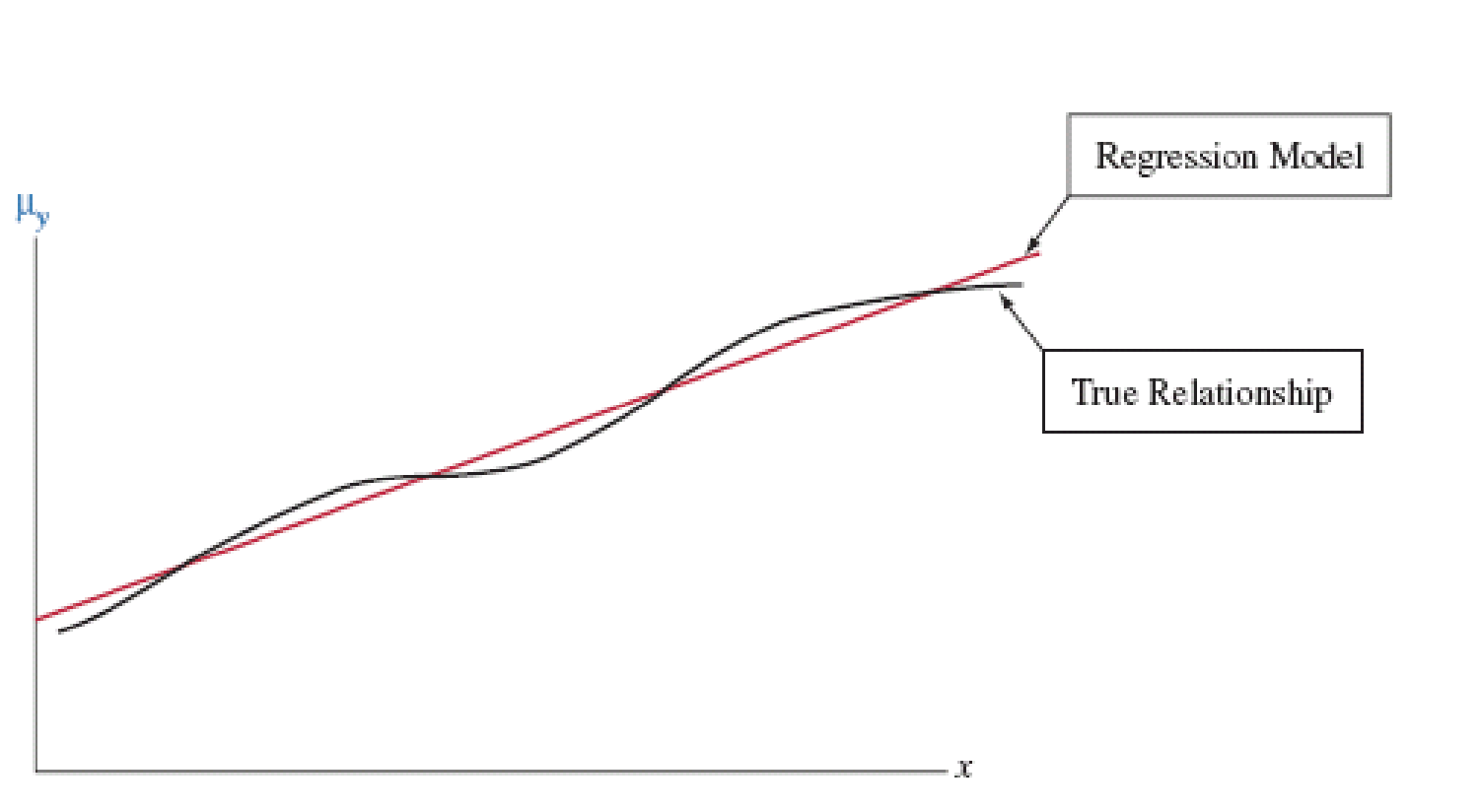
Learning Objective 8: The Population Regression Equation

- **In the population regression equation, α is a population y-intercept and β is a population slope**
  - **These are parameters**
- **In practice we estimate the population regression equation using the prediction equation for the sample data**

Learning Objective 8: The Population Regression Equation

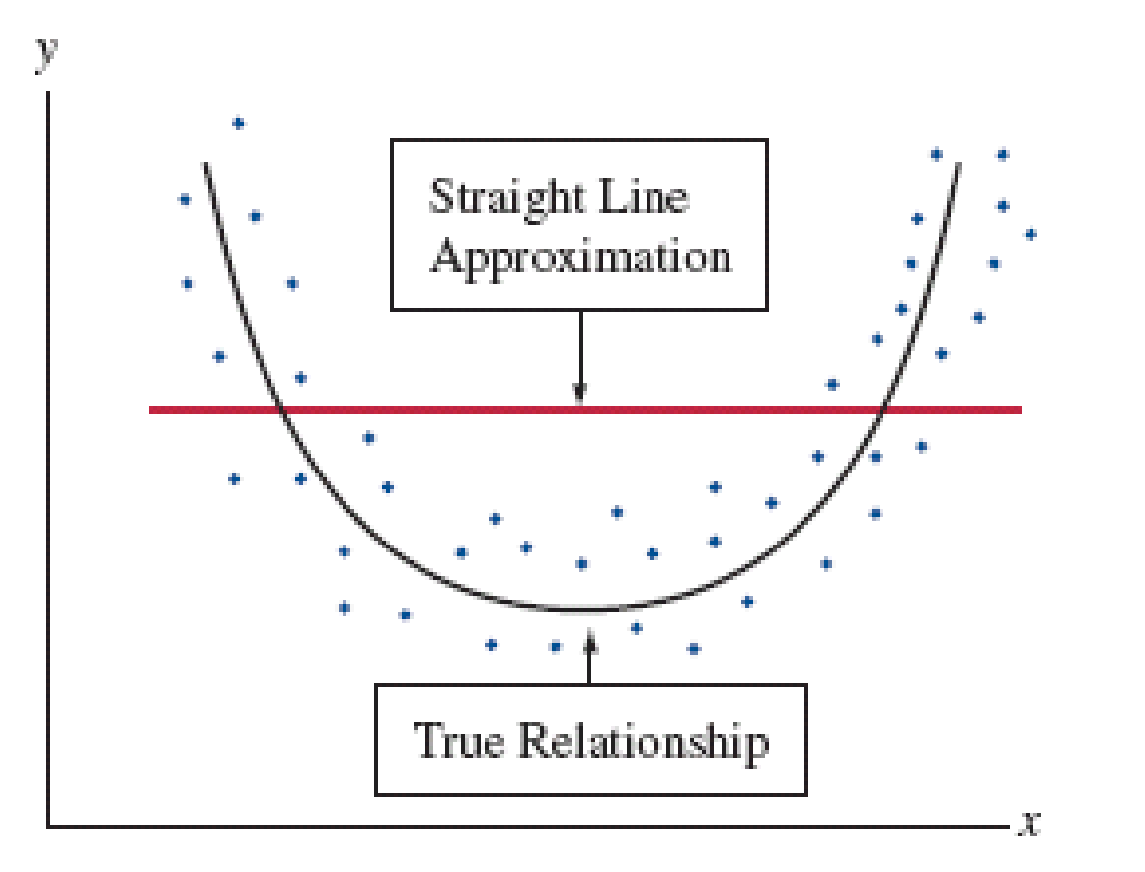
- **The population regression equation merely approximates the actual relationship between x and the population means of y**
- **It is a model**
- **A model is a simple approximation for how variables relate in the population**

Learning Objective 8:
The Regression Model



Learning Objective 8:
The Regression Model

- **If the true relationship is far from a straight line, this regression model may be a poor one**



Learning Objective 9: Variability about the Line

- **At each fixed value of x, variability occurs in the y values around their mean, μ_y**
- **The probability distribution of y values at a fixed value of x is a conditional distribution**
- **At each value of x, there is a conditional distribution of y values**
- **An additional parameter σ describes the standard deviation of each conditional distribution**

Learning Objective 10:
A Statistical Model

- **A statistical model never holds *exactly* in practice.**
- **It is merely an approximation for reality**
- **Even though it does not describe reality exactly, a model is useful if the true relationship is close to what the model predicts**

Chapter 11: Analyzing Association Between Quantitative Variables: Regression Analysis

Section 11.2: How Can We Describe Strength of Association?

Learning Objectives

1. Correlation and Slope
2. Example: What's the Correlation for Predicting Strength?
3. The Squared Correlation

Learning Objective 1: Correlation

- **The correlation, denoted by r, describes linear association**
- **The correlation ‘r’ has the same sign as the slope ‘b’**
- **The correlation ‘r’ always falls between -1 and +1**
- **The larger the absolute value of r, the stronger the linear association**

Learning Objective 1: Correlation and Slope

- **We can't use the slope to describe the strength of the association between two variables because the slope's numerical value depends on the units of measurement**

- **The correlation is a standardized version of the slope**
- **The correlation does not depend on units of measurement.**

Learning Objective 1: Correlation and Slope

- **The correlation and the slope are related in the following way:**

$$r = b\frac{s_x}{s_y}$$

Learning Objective 2:
Example: What's the Correlation for Predicting Strength?

- **For the female athlete strength study:**
  - **x: number of 60-pound bench presses**
  - **y: maximum bench press**
  - **x: mean = 11.0, st.dev.=7.1**
  - **y: mean= 79.9 lbs., st.dev. = 13.3 lbs.**
- **Regression equation:**

$$\hat{y} = 63.5 + 1.49x$$

Learning Objective 2:
Example: What's the Correlation for Predicting Strength?

$$r = b\left(\frac{s_x}{s_y}\right) = 1.49\left(\frac{7.1}{13.3}\right) = 0.80$$

- **The variables have a strong, positive association**

Learning Objective 3: The Squared Correlation

- **Another way to describe the strength of association refers to how close predictions for y tend to be to observed y values**
- **The variables are strongly associated if you can predict y much better by substituting x values into the prediction equation than by merely using the sample mean $\bar{y}$ and ignoring x**

Learning Objective 3: The Squared Correlation

- **Consider the prediction error: the difference between the observed and predicted values of y**
  - **Using the regression line to make a prediction, each error is:**

$$y - \hat{y}$$

  - **Using only the sample mean, $\bar{y}$, to make a prediction, each error is:**

$$y - \bar{y}$$

Learning Objective 3: The Squared Correlation

- **When we predict y using $\bar{y}$ (that is, ignoring x), the error summary equals:**

$$\sum (y - \bar{y})^2$$

- **This is called the *total sum of squares***

Learning Objective 3: The Squared Correlation

- **When we predict y using x with the regression equation, the error summary is:**

$$\sum (y - \hat{y})^2$$

- **This is called the *residual sum of squares***

Learning Objective 3: The Squared Correlation

- **When a strong linear association exists, the regression equation predictions tend to be much better than the predictions using $\bar{y}$**

- **We measure the *proportional reduction in error* and call it, r^2**

Learning Objective 3: The Squared Correlation

$$r^2 = \frac{\sum (y - \bar{y})^2 - \sum (y - \hat{y})^2}{\sum (y - \bar{y})^2}$$

- **We use the notation r^2 for this measure because it equals the square of the correlation r**

Learning Objective 3: The Squared Correlation Example: What Does r^2 Tell Us in the Strength Study?

- **For the female athlete strength study:**
  - **x: number of 60-pund bench presses**
  - **y: maximum bench press**
  - **The correlation value was found to be r = 0.80**
- **We can calculate r^2 from r: $(0.80)^2=0.64$**
- **For predicting maximum bench press, the regression equation has 64% less error than $\bar{y}$ has**

Learning Objective 3: The Squared Correlation

- Properties:
  - r^2 falls between 0 and 1
  - r^2=1 when $\sum(y-\hat{y})^2=0$. This happens only when all the data points fall exactly on the regression line
  - r^2=0 when $\sum(y-\hat{y})^2=\sum(y-\bar{y})^2$. This happens when the slope *b*=0, in which case each $\hat{y}=\bar{y}$
  - The closer r^2 is to 1, the stronger the linear association: the more effective the regression equation is compared to $\bar{y}$ in predicting *y*

Learning Objective 3: Correlation r and Its Square r^2

- **Both r and r^2 describe the strength of association**
- **'r' falls between -1 and +1**
  - **It represents the slope of the regression line when x and y have been standardized**
- **'r^2' falls between 0 and 1**
  - **It summarizes the reduction in sum of squared errors in predicting y using the regression line instead of using $\bar{y}$**