Chapter 13

Chapter 13
Simple Linear Regression

Chapter Topics
- Types of Regression Models
- Determining the Simple Linear Regression Equation
- Measures of Variation
- Assumptions of Regression and Correlation
- Residual Analysis
- Measuring Autocorrelation
- Inferences about the Slope

Chapter Topics (continued)
- Correlation - Measuring the Strength of the Association
- Estimation of Mean Values and Prediction of Individual Values
- Pitfalls in Regression and Ethical Issues
Purpose of Regression Analysis

- Regression Analysis is Used Primarily to Model Causality and Provide Prediction
- Predict the values of a dependent (response) variable based on values of at least one independent (explanatory) variable
- Explain the effect of the independent variables on the dependent variable

Types of Regression Models

- Positive Linear Relationship
- Negative Linear Relationship
- Relationship NOT Linear
- No Relationship

Simple Linear Regression Model

- Relationship between Variables is Described by a Linear Function
- The Change of One Variable Causes the Other Variable to Change
- A Dependency of One Variable on the Other
Simple Linear Regression Model

(continued)

Population regression line is a straight line that describes the dependence of the average value (conditional mean) of one variable on the other.

\[ Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \]

Population Y Intercept
Population Slope Coefficient
Random Error
Dependent (Response) Variable
Population Regression Line (Conditional Mean)
Independent (Explanatory) Variable

Simple Linear Regression Model

(continued)

\[ Y = \beta_0 + \beta_1 X + \varepsilon \]

Y (Observed Value of Y) = \( Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \)
\[ \varepsilon = \text{Random Error} \]
\[ \beta_1 \]
\[ \beta_0 \]
Observed Value of Y
X

Linear Regression Equation

Sample regression line provides an estimate of the population regression line as well as a predicted value of Y.

\[ Y_i = b_0 + b_1 X_i + e_i \]

Simple Regression Equation (Fitted Regression Line, Predicted Value)
\[ \hat{Y} = b_0 + b_1 X \]
Linear Regression Equation

- \( b_0 \) and \( b_1 \) are obtained by finding the values of \( b_0 \) and \( b_1 \) that minimize the sum of the squared residuals
  \[
  \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} e_i^2
  \]
- \( b_0 \) provides an estimate of \( \beta_0 \)
- \( b_1 \) provides an estimate of \( \beta_1 \)

Linear Regression Equation (continued)

\[
Y_i = b_0 + b_1 X_i + e_i
\]

Interpretation of the Slope and Intercept

- \( \beta_0 = \mu_{Y|X=0} \) is the average value of \( Y \) when the value of \( X \) is zero
- \( \beta_1 = \frac{\text{change in } \mu_{Y|X}}{\text{change in } X} \) measures the change in the average value of \( Y \) as a result of a one-unit change in \( X \)
Interpretation of the Slope and Intercept

- $b_0 = \hat{Y}(X = 0)$ is the estimated average value of $Y$ when the value of $X$ is zero.

- $b_1 = \frac{\text{change in } \hat{Y}}{\text{change in } X}$ is the estimated change in the average value of $Y$ as a result of a one-unit change in $X$.

Simple Linear Regression: Example

You wish to examine the linear dependency of the annual sales of produce stores on their sizes in square footage. Sample data for 7 stores were obtained. Find the equation of the straight line that fits the data best.

<table>
<thead>
<tr>
<th>Store</th>
<th>Square Feet</th>
<th>Annual Sales ($1000)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1,726</td>
<td>3,681</td>
</tr>
<tr>
<td>2</td>
<td>1,542</td>
<td>3,395</td>
</tr>
<tr>
<td>3</td>
<td>2,816</td>
<td>6,653</td>
</tr>
<tr>
<td>4</td>
<td>5,555</td>
<td>9,543</td>
</tr>
<tr>
<td>5</td>
<td>1,292</td>
<td>3,318</td>
</tr>
<tr>
<td>6</td>
<td>2,208</td>
<td>5,563</td>
</tr>
<tr>
<td>7</td>
<td>1,313</td>
<td>3,760</td>
</tr>
</tbody>
</table>

Scatter Diagram: Example

Excel Output
Simple Linear Regression Equation: Example

\[ \hat{Y}_i = b_0 + b_1 X_i \]

\[ = 1636.415 + 1.487 X_i \]

From Excel Printout:

<table>
<thead>
<tr>
<th>Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
</tr>
<tr>
<td>X Variable</td>
</tr>
</tbody>
</table>

Graph of the Simple Linear Regression Equation: Example

Interpretation of Results: Example

\[ \hat{Y}_i = 1636.415 + 1.487 X_i \]

The slope of 1.487 means that for each increase of one unit in \( X \), we predict the average of \( Y \) to increase by an estimated 1.487 units.

The equation estimates that for each increase of 1 square foot in the size of the store, the expected annual sales are predicted to increase by $1487.
Simple Linear Regression in PHStat

- In Excel, use PHStat | Regression | Simple Linear Regression ...
- Excel Spreadsheet of Regression Sales on Footage

Measures of Variation: The Sum of Squares

\[ SST = SSR + SSE \]

- Total Sample Variability = Explained Variability + Unexplained Variability

Measures of Variation: The Sum of Squares (continued)

- SST = Total Sum of Squares
  - Measures the variation of the \( Y \) values around their mean, \( \bar{Y} \)
- SSR = Regression Sum of Squares
  - Explained variation attributable to the relationship between \( X \) and \( Y \)
- SSE = Error Sum of Squares
  - Variation attributable to factors other than the relationship between \( X \) and \( Y \)
Measures of Variation: The Sum of Squares

\[ SST = \sum (Y_i - \bar{Y})^2 \]

\[ SSE = \sum (Y_i - \hat{Y}_i)^2 \]

\[ SSR = \sum (\hat{Y}_i - \bar{Y})^2 \]

Venn Diagrams and Explanatory Power of Regression

Variations in Sales explained by the error term or unexplained by Sizes \( (SSE) \)

Variations in Sales explained by Sizes or variations in Sizes used in explaining variation in Sales \( (SSR) \)

The ANOVA Table in Excel

<table>
<thead>
<tr>
<th>ANOVA</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>Significance F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>k</td>
<td>SSR</td>
<td>MSR = SSR/k</td>
<td>MSR/MSE</td>
<td>P-value of the F Test</td>
</tr>
<tr>
<td>Error</td>
<td>n-k-1</td>
<td>SSE</td>
<td>MSE = SSE/(n-k-1)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>n-1</td>
<td>SST</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Measures of Variation

The Sum of Squares: Example

<table>
<thead>
<tr>
<th>Degrees of freedom</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>Significance F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>30380456.12</td>
<td>30380456.12</td>
<td>81.1790902</td>
<td>0.000281201</td>
</tr>
<tr>
<td>Error</td>
<td>5</td>
<td>1871199.595</td>
<td>374239.919</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>6</td>
<td>32251655.71</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Excel Output for Produce Stores

- **SSR** (Regression Sum of Squares)
- **SSE** (Error Sum of Squares)

### The Coefficient of Determination

- $r^2 = \frac{SSR}{SST}$
- Measures the proportion of variation in $Y$ that is explained by the independent variable $X$ in the regression model

### Venn Diagrams and Explanatory Power of Regression

- $r^2 = \frac{SSR}{SSR + SSE}$
Coefficients of Determination ($r^2$) and Correlation ($r$)

$Y^2 = 1, r = +1$

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

$Y^2 = .81, r = +0.9$

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

$Y^2 = 0, r = 0$

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

Standard Error of Estimate

$S_{yx} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum (Y - \hat{Y})^2}{n-2}}$

- Measures the standard deviation (variation) of the $Y$ values around the regression equation

Measures of Variation: Produce Store Example

<table>
<thead>
<tr>
<th>Excel Output for Produce Stores</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Regression Statistics</strong></td>
</tr>
<tr>
<td>Multiple R</td>
</tr>
<tr>
<td>R Square</td>
</tr>
<tr>
<td>Adjusted R Square</td>
</tr>
<tr>
<td>Standard Error of Estimate</td>
</tr>
<tr>
<td>Observations</td>
</tr>
</tbody>
</table>

$r^2 = .94$

94% of the variation in annual sales can be explained by the variability in the size of the store as measured by square footage.
Linear Regression Assumptions

- Normality
  - Y values are normally distributed for each X
  - Probability distribution of error is normal
- Homoscedasticity (Constant Variance)
- Independence of Errors

Consequences of Violation of the Assumptions

- Violation of the Assumptions
  - Non-normality (error not normally distributed)
  - Heteroscedasticity (variance not constant)
    - Usually happens in cross-sectional data
  - Autocorrelation (errors are not independent)
    - Usually happens in time-series data
- Consequences of Any Violation of the Assumptions
  - Predictions and estimations obtained from the sample regression line will not be accurate
  - Hypothesis testing results will not be reliable
- It is Important to Verify the Assumptions

Variation of Errors Around the Regression Line

- Y values are normally distributed around the regression line.
- For each X value, the “spread” or variance around the regression line is the same.
Residual Analysis

- Purposes
  - Examine linearity
  - Evaluate violations of assumptions
- Graphical Analysis of Residuals
  - Plot residuals vs. $X$ and time

Residual Analysis for Linearity

- Not Linear
- Linear

Residual Analysis for Homoscedasticity

- Heteroscedasticity
- Homoscedasticity
Residual Analysis: Excel Output for Produce Stores Example

<table>
<thead>
<tr>
<th>Observation</th>
<th>Predicted Y</th>
<th>Residuals</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4202.34417</td>
<td>-521.344417</td>
</tr>
<tr>
<td>2</td>
<td>3928.803824</td>
<td>-533.8038245</td>
</tr>
<tr>
<td>3</td>
<td>5822.775103</td>
<td>830.2248971</td>
</tr>
<tr>
<td>4</td>
<td>9894.664688</td>
<td>-351.6646882</td>
</tr>
<tr>
<td>5</td>
<td>3557.14541</td>
<td>-239.1454103</td>
</tr>
<tr>
<td>6</td>
<td>4918.90184</td>
<td>644.0981603</td>
</tr>
<tr>
<td>7</td>
<td>3588.364717</td>
<td>171.6352829</td>
</tr>
</tbody>
</table>

Residual Analysis for Independence

- The Durbin-Watson Statistic
  - Used when data are collected over time to detect autocorrelation (residuals in one time period are related to residuals in another period)
  - Measures violation of independence assumption
    \[ D = \frac{\sum_{i=2}^{n} (e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2} \]
    Should be close to 2.
    If not, examine the model for autocorrelation.

Durbin-Watson Statistic in PHStat

- PHStat | Regression | Simple Linear Regression ...
- Check the box for Durbin-Watson Statistic
Obtaining the Critical Values of Durbin-Watson Statistic

Table 13.4  Finding Critical Values of Durbin-Watson Statistic

<table>
<thead>
<tr>
<th>$\alpha = .05$</th>
<th>$k=1$</th>
<th>$k=2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n$</td>
<td>$d_L$</td>
<td>$d_U$</td>
</tr>
<tr>
<td>15</td>
<td>1.08</td>
<td>1.36</td>
</tr>
<tr>
<td>16</td>
<td>1.10</td>
<td>1.37</td>
</tr>
</tbody>
</table>

Using the Durbin-Watson Statistic

$H_0$: No autocorrelation (error terms are independent)

$H_1$: There is autocorrelation (error terms are not independent)

Residual Analysis for Independence

Graphical Approach

- Not Independent
- Independent

- Cyclical Pattern
- No Particular Pattern

Residual is Plotted Against Time to Detect Any Autocorrelation
Inference about the Slope: t Test

- **t Test for a Population Slope**
  - Is there a linear relationship between Y and X?
- **Null and Alternative Hypotheses**
  - $H_0: \beta_1 = 0$ (no linear relationship)
  - $H_1: \beta_1 \neq 0$ (linear relationship)
- **Test Statistic**
  - $t = \frac{b_1 - \beta_1}{S_{b_1}}$ where $S_{b_1} = \frac{S_y}{\sum (X_i - \bar{X})^2}$
  - df. = $n - 2$

---

Example: Produce Store

<table>
<thead>
<tr>
<th>Data for 7 Stores:</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Store</strong></td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>7</td>
</tr>
</tbody>
</table>

**Estimated Regression Equation:**

$\hat{Y}_i = 1636.415 + 1.487X_i$

The slope of this model is 1.487.

Are square footage and annual sales linearly related?

---

Inferences about the Slope: t Test Example

- $H_0: \beta_1 = 0$
- $H_1: \beta_1 \neq 0$
- $\alpha = .05$
- df. = $7 - 2 = 5$

**Critical Value(s):**

$t = \frac{b_1 - \beta_1}{S_{b_1}}$

- Reject $H_0$ if $t < -2.5706$ or $t > 2.5706$

**Test Statistic:**

- From Excel Printout
  - $b_1 = 1.487$
  - $S_{b_1} = 0.1650$
  - $t = 9.0099$
  - $p-value = 0.00028$

**Decision:**

- Reject $H_0$ if $p-value < .025$

**Conclusion:**

- There is evidence that square footage is linearly related to annual sales.
Inferences about the Slope: Confidence Interval Example

Confidence Interval Estimate of the Slope:

\[ b_1 \pm t_{n-2}S_{b_1} \]

Excel Printout for Produce Stores

<table>
<thead>
<tr>
<th></th>
<th>Lower 95%</th>
<th>Upper 95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>475.810926</td>
<td>2797.01853</td>
</tr>
<tr>
<td>Footage</td>
<td>1.06249037</td>
<td>1.91077694</td>
</tr>
</tbody>
</table>

At 95% level of confidence, the confidence interval for the slope is (1.062, 1.911). Does not include 0.

Conclusion: There is a significant linear relationship between annual sales and the size of the store.

Inferences about the Slope: F Test

- F Test for a Population Slope
  - Is there a linear relationship between Y and X?
- Null and Alternative Hypotheses
  - \( H_0: \beta_1 = 0 \) (no linear relationship)
  - \( H_1: \beta_1 \neq 0 \) (linear relationship)
- Test Statistic
  \[ F = \frac{SSR}{SSE} \]
  \( \frac{1}{(n - 2)} \)
  - Numerator d.f.=1, denominator d.f.=n-2

Relationship between a t Test and an F Test

- Null and Alternative Hypotheses
  - \( H_0: \beta_1 = 0 \) (no linear relationship)
  - \( H_1: \beta_1 \neq 0 \) (linear relationship)
- \( (t_{n-2})^2 = F_{1,n-2} \)
- The \( p \) -value of a t Test and the \( p \) -value of an F Test are Exactly the Same
- The Rejection Region of an F Test is Always in the Upper Tail
**Inferences about the Slope: \( F \) Test Example**

- **Null Hypothesis (H\(_0\)):** \( \beta_1 = 0 \)
- **Alternative Hypothesis (H\(_1\)):** \( \beta_1 \neq 0 \)

**ANOVA**

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>( F )</th>
<th>Significance F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>30380456.12</td>
<td>30380456.12</td>
<td>81.179</td>
<td>0.000281</td>
</tr>
<tr>
<td>Residual</td>
<td>5</td>
<td>1871199.595</td>
<td>374239.919</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>6</td>
<td>32251655.71</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Test Statistic:**

- **Decision:** Reject H\(_0\).
- **Conclusion:** There is evidence that square footage is linearly related to annual sales.

**Purpose of Correlation Analysis**

- Correlation Analysis is Used to Measure Strength of Association (Linear Relationship) Between 2 Numerical Variables
  - Only strength of the relationship is concerned
  - No causal effect is implied

**Population Correlation Coefficient \( \rho \) (Rho) is Used to Measure the Strength between the Variables**
Purpose of Correlation Analysis

(continued)

- Sample Correlation Coefficient $r$ is an estimate of $\rho$ and is used to measure the strength of the linear relationship in the sample observations.

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

Sample Observations from Various $r$ Values

- $r = -1$
- $r = -0.6$
- $r = 0$
- $r = 0.6$
- $r = 1$

Features of $\rho$ and $r$

- Unit Free
- Range 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 the linear relationship
13-19

**t** Test for Correlation

- Hypotheses
  - $H_0: \rho = 0$ (no correlation)
  - $H_1: \rho \neq 0$ (correlation)

- Test Statistic
  \[
  t = \frac{r - \rho}{\sqrt{\frac{1 - r^2}{n - 2}}} 
  \]

  where

  \[
  r = \sqrt{\frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}} 
  \]

  $\rho = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}$

---

### Example: Produce Stores

Is there any evidence of linear relationship between annual sales of a store and its square footage at .05 level of significance?

**From Excel Printout**

<table>
<thead>
<tr>
<th>Regression Statistics</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple R</td>
<td>0.9705572</td>
</tr>
<tr>
<td>R Square</td>
<td>0.94198129</td>
</tr>
<tr>
<td>Adjusted R Square</td>
<td>0.9307754</td>
</tr>
<tr>
<td>Standard Error</td>
<td>611.751517</td>
</tr>
<tr>
<td>Observations</td>
<td>7</td>
</tr>
</tbody>
</table>

- $H_0: \rho = 0$ (no association)
- $H_1: \rho \neq 0$ (association)
- $\alpha = .05$
- df = 7 - 2 = 5

**Decision:**

Reject $H_0$.

**Conclusion:**

There is evidence of a linear relationship at 5% level of significance.

The value of the $t$ statistic is exactly the same as the $t$ statistic value for test on the slope coefficient.
Estimation of Mean Values

Confidence Interval Estimate for $\mu_{Y|X_i}$: The Mean of $Y$ Given a Particular $X_i$

- Size of interval varies according to distance away from mean, $\bar{X}$
- Standard error of the estimate

$\hat{Y}_i \pm t_{n-2,\alpha/2} \frac{S_{XX}}{n} \sqrt{\frac{1 + \frac{(X_i - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}{n}}$

Prediction of Individual Values

Prediction Interval for Individual Response $Y_i$ at a Particular $X_i$

- Addition of 1 increases width of interval from that for the mean of $Y$

$\hat{Y}_i \pm t_{n-2} S_{XX} \sqrt{\frac{1 + \frac{(X_i - \bar{X})^2}{n} + \frac{(X_i - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}}{n}}$

Interval Estimates for Different Values of $X$

- Prediction Interval for an Individual $Y_i$
- Confidence Interval for the Mean of $Y$

$Y_i = b_0 + b_1 X_i$
Example: Produce Stores

Data for 7 Stores:

<table>
<thead>
<tr>
<th>Store</th>
<th>Square Feet</th>
<th>Annual Sales ($000)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1,726</td>
<td>3,681</td>
</tr>
<tr>
<td>2</td>
<td>1,542</td>
<td>3,395</td>
</tr>
<tr>
<td>3</td>
<td>2,816</td>
<td>6,653</td>
</tr>
<tr>
<td>4</td>
<td>5,555</td>
<td>9,543</td>
</tr>
<tr>
<td>5</td>
<td>1,292</td>
<td>3,318</td>
</tr>
<tr>
<td>6</td>
<td>2,208</td>
<td>5,563</td>
</tr>
<tr>
<td>7</td>
<td>1,313</td>
<td>3,760</td>
</tr>
</tbody>
</table>

Consider a store with 2000 square feet.

Regression Model Obtained:

\[ \hat{Y}_i = 1636.415 + 1.487X_i \]

Example: Produce Stores

Estimation of Mean Values:

Confidence Interval Estimate for \( \mu_{\hat{Y} \mid X = X_i} \)

Find the 95% confidence interval for the average annual sales for stores of 2,000 square feet.

Predicted Sales \( \hat{Y} = 1636.415 + 1.487X_i = 4610.45 \) (in $000)

\[ \hat{X} = 2350.29 \quad S_{XX} = 611.75 \quad t_{a/2} = t_{0.025} = 2.5706 \]

\[ \hat{Y}_i \pm t_{a/2} S_{XX} \sqrt{ \frac{1}{n} + \frac{(X_i - \hat{X})^2}{\sum (X_i - \hat{X})^2} } = 4610.45 \pm 612.66 \]

\[ 3997.02 < \mu_{\hat{Y} \mid X = X_i} < 5222.34 \]

Prediction Interval for \( Y \):

Example

Prediction Interval for Individual \( Y_{X = X_i} \)

Find the 95% prediction interval for annual sales of one particular store of 2,000 square feet.

Predicted Sales \( \hat{Y} = 1636.415 + 1.487X_i = 4610.45 \) (in $000)

\[ \hat{X} = 2350.29 \quad S_{XX} = 611.75 \quad t_{a/2} = t_{0.025} = 2.5706 \]

\[ \hat{Y}_i \pm t_{a/2} S_{XX} \sqrt{ 1 + \frac{1}{n} + \frac{(X_i - \hat{X})^2}{\sum (X_i - \hat{X})^2} } = 4610.45 \pm 1687.68 \]

\[ 2922.00 < Y_{X = X_i} < 6297.37 \]
Estimation of Mean Values and Prediction of Individual Values in PHStat

- In Excel, use PHStat | Regression | Simple Linear Regression ...
  - Check the “Confidence and Prediction Interval for X=” box
  - Excel Spreadsheet of Regression Sales on Footage

Microsoft Excel Worksheet

Pitfalls of Regression Analysis

- Lacking an Awareness of the Assumptions Underlying Least-Squares Regression
- Not Knowing How to Evaluate the Assumptions
- Not Knowing What the Alternatives to Least-Squares Regression are if a Particular Assumption is Violated
- Using a Regression Model Without Knowledge of the Subject Matter

Strategy for Avoiding the Pitfalls of Regression

- Start with a scatter plot to observe possible relationship between X on Y
- Perform residual analysis to check the assumptions
- Use a histogram, stem-and-leaf display, box-and-whisker plot, or normal probability plot of the residuals to uncover possible non-normality
Strategy for Avoiding the Pitfalls of Regression

If there is violation of any assumption, use alternative methods (e.g., least absolute deviation regression or least median of squares regression) to least-squares regression or alternative least-squares models (e.g., curvilinear or multiple regression).

If there is no evidence of assumption violation, then test for the significance of the regression coefficients and construct confidence intervals and prediction intervals.

Chapter Summary

- Introduced Types of Regression Models
- Discussed Determining the Simple Linear Regression Equation
- Described Measures of Variation
- Addressed Assumptions of Regression and Correlation
- Discussed Residual Analysis
- Addressed Measuring Autocorrelation

Chapter Summary (continued)

- Described Inference about the Slope
- Discussed Correlation - Measuring the Strength of the Association
- Addressed Estimation of Mean Values and Prediction of Individual Values
- Discussed Pitfalls in Regression and Ethical Issues