Chapter 10
Correlation and Regression

10-1 Overview
10-2 Correlation
10-3 Regression

Overview
This chapter introduces important methods for making inferences about a correlation (or relationship) between two variables, and describing such a relationship with an equation that can be used for predicting the value of one variable given the value of the other variable.

We consider sample data that come in pairs.

Section 10-2
Correlation

Part 1: Basic Concepts of Correlation

Definition
A correlation exists between two variables when one of them is related to the other in some way.

Key Concept
This section introduces the linear correlation coefficient $r$, which is a numerical measure of the strength of the relationship between two variables representing quantitative data.

Because technology can be used to find the value of $r$, it is important to focus on the concepts in this section, without becoming overly involved with tedious arithmetic calculations.

Definition
The linear correlation coefficient $r$ measures the strength of the linear relationship between paired $x$- and $y$-quantitative values in a sample.

Exploring the Data
We can often see a relationship between two variables by constructing a scatterplot.

Figure 10-2 following shows scatterplots with different characteristics.
Scatterplots of Paired Data

Figure 10-2

Requirements
1. The sample of paired \((x, y)\) data is a random sample of independent quantitative data.
2. Visual examination of the scatterplot must confirm that the points approximate a straight-line pattern.
3. The outliers must be removed if they are known to be errors. The effects of any other outliers should be considered by calculating \(r\) with and without the outliers included.

Notation for the Linear Correlation Coefficient
- \(n\) represents the number of pairs of data present.
- \(\Sigma\) denotes the addition of the items indicated.
- \(\Sigma x\) denotes the sum of all \(x\)-values.
- \(\Sigma x^2\) indicates that each \(x\)-value should be squared and then those squares added.
- \((\Sigma x)^2\) indicates that the \(x\)-values should be added and the total then squared.
- \(\Sigma xy\) indicates that each \(x\)-value should be first multiplied by its corresponding \(y\)-value. After obtaining all such products, find their sum.
- \(r\) represents linear correlation coefficient for a sample.
- \(\rho\) represents linear correlation coefficient for a population.

Formula

The linear correlation coefficient \(r\) measures the strength of a linear relationship between the paired values in a sample.

\[
r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n(\Sigma x^2) - (\Sigma x)^2}n(\Sigma y^2) - (\Sigma y)^2}}
\]

Formula 10-1

Calculators can compute \(r\)

Interpreting \(r\)

Using Table A-6: If the absolute value of the computed value of \(r\) exceeds the value in Table A-6, conclude that there is a linear correlation. Otherwise, there is not sufficient evidence to support the conclusion of a linear correlation.

Using Software: If the computed \(P\)-value is less than or equal to the significance level, conclude that there is a linear correlation. Otherwise, there is not sufficient evidence to support the conclusion of a linear correlation.
Rounding the Linear Correlation Coefficient \( r \)

- Round to **three** decimal places so that it can be compared to critical values in Table A-6.
- Use calculator or computer if possible.

Example: Calculating \( r \)

Using the simple random sample of data below, find the value of \( r \).

<table>
<thead>
<tr>
<th>Data</th>
<th>( x )</th>
<th>3</th>
<th>1</th>
<th>3</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>( y )</td>
<td>5</td>
<td>8</td>
<td>6</td>
<td>4</td>
<td></td>
</tr>
</tbody>
</table>

Example: Calculating \( r \) - cont

<p>| Table 10-2 Finding Statistics Used to Calculate ( r ) |</p>
<table>
<thead>
<tr>
<th>( x )</th>
<th>( y )</th>
<th>( x \cdot y )</th>
<th>( x^2 )</th>
<th>( y^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>5</td>
<td>15</td>
<td>9</td>
<td>25</td>
</tr>
<tr>
<td>1</td>
<td>8</td>
<td>8</td>
<td>1</td>
<td>64</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>18</td>
<td>9</td>
<td>36</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>20</td>
<td>25</td>
<td>16</td>
</tr>
<tr>
<td>Total</td>
<td>12</td>
<td>23</td>
<td>61</td>
<td>44</td>
</tr>
</tbody>
</table>

\[
\begin{align*}
\sum x &= \sum y \\
\sum xy &= \sum y \\
\sum x^2 &= \sum x^2 \\
\sum y^2 &= \sum y^2 \\
\end{align*}
\]

\[
\begin{align*}
r &= \frac{n \sum xy - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2} \sqrt{n(\sum y^2) - (\sum y)^2}} \\
r &= \frac{4(61) - (12)(23)}{\sqrt{4(44) - (12)^2} \sqrt{4(141) - (23)^2}} \\
&= \frac{-32}{33.466} = -0.956
\end{align*}
\]

Example: Calculating \( r \) - cont

Given \( r = -0.956 \), if we use a 0.05 significance level we conclude that there is a linear correlation between \( x \) and \( y \) since the absolute value of \( r \) exceeds the critical value of 0.950. However, if we use a 0.01 significance level, we do not conclude that there is a linear correlation because the absolute value of \( r \) does not exceed the critical value of 0.999.

Example: Old Faithful

Given the sample data in Table 10-1, find the value of the linear correlation coefficient \( r \), then refer to Table A-6 to determine whether there is a significant linear correlation between the duration and interval of the eruption times.

Using the same procedure previously illustrated, we find that \( r = 0.926 \).

Referring to Table A-6, we locate the row for which \( n = 8 \). Using the critical value for \( \alpha = 0.05 \), we have 0.707. Because \( r = 0.926 \), its absolute value exceeds 0.707, so we conclude that there is a linear correlation between the duration and interval after the eruption times.
Properties of the Linear Correlation Coefficient $r$

1. $-1 \leq r \leq 1$
2. The value of $r$ does not change if all values of either variable are converted to a different scale.
3. The value of $r$ is not affected by the choice of $x$ and $y$. Interchange all $x$- and $y$-values and the value of $r$ will not change.
4. $r$ measures strength of a linear relationship.

Interpreting $r$: Explained Variation

The value of $r^2$ is the proportion of the variation in $y$ that is explained by the linear relationship between $x$ and $y$.

Example: Old Faithful

Using the duration/interval data in Table 10-1, we have found that the value of the linear correlation coefficient $r = 0.926$. What proportion of the variation of the interval after eruption times can be explained by the variation in the duration times?

With $r = 0.926$, we get $r^2 = 0.857$.

We conclude that 0.857 (or about 86%) of the variation in interval after eruption times can be explained by the linear relationship between the duration times and the interval after eruption times. This implies that 14% of the variation in interval after eruption times cannot be explained by the duration times.

Common Errors Involving Correlation

1. Causation: It is wrong to conclude that correlation implies causality.
2. Averages: Averages suppress individual variation and may inflate the correlation coefficient.
3. Linearity: There may be some relationship between $x$ and $y$ even when there is no linear correlation.

Part 2: Formal Hypothesis Test

Formal Hypothesis Test

- We wish to determine whether there is a significant linear correlation between two variables.
- We present two methods.
- In both methods let
  
  $H_0: \rho = 0$ (no significant linear correlation)
  $H_1: \rho \neq 0$ (significant linear correlation)

Hypothesis Test for a Linear Correlation

Figure 10-3
**Method 1: Test Statistic is t**

Test statistic:

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

Critical values:

Use Table A-3 with degrees of freedom = \( n - 2 \)

**Method 1 - cont**

P-value:

Use Table A-3 with degrees of freedom = \( n - 2 \)

Conclusion:

If the absolute value of \( t \) is > critical value from Table A-3, reject \( H_0 \) and conclude that there is a linear correlation. If the absolute value of \( t \leq \) critical value, fail to reject \( H_0 \); there is not sufficient evidence to conclude that there is a linear correlation.

**Method 2: Test Statistic is r**

Test statistic: \( r \)

Critical values: Refer to Table A-6

Conclusion:

If the absolute value of \( r \) is > critical value from Table A-6, reject \( H_0 \) and conclude that there is a linear correlation. If the absolute value of \( r \leq \) critical value, fail to reject \( H_0 \); there is not sufficient evidence to conclude that there is a linear correlation.

**Example: Old Faithful**

Using the duration/interval data in Table 10-1, test the claim that there is a linear correlation between the duration of an eruption and the time interval after that eruption. Use Method 1.

\[
T = \frac{r}{\sqrt{\frac{1 - r^2}{n-2}}}
\]

\[ t = \frac{0.926}{\sqrt{\frac{1 - 0.926^2}{8 - 2}}} = 6.008 \]

**Example: Old Faithful**

Using the duration/interval data in Table 10-1, test the claim that there is a linear correlation between the duration of an eruption and the time interval after that eruption. Use Method 2.

The test statistic is \( r = 0.926 \). The critical values of \( r = 0.707 \) are found in Table A-6 with \( n = 8 \) and \( \alpha = 0.05 \).
Method 2: Test Statistic is \( r \)

- Test statistic: \( r \)
- Critical values: Refer to Table A-6
  (8 degrees of freedom)

![Figure 10-5](image)

Example: Old Faithful

Using the duration/interval data in Table 10-1, test the claim that there is a linear correlation between the duration of an eruption and the time interval after that eruption.

Using either of the two methods, we find that the absolute value of the test statistic does exceed the critical value (Method 1: 6.008 > 2.447. Method 2: 0.926 > 0.707); that is, the test statistic falls in the critical region.

We therefore reject the null hypothesis. There is sufficient evidence to support the claim of a linear correlation between the duration times of eruptions and the time intervals after the eruptions.

Justification for \( r \) Formula

The text presents a detailed rationale for the use of Formula 10-1. The student should study it carefully.

Section 10-3
Regression

Key Concept

The key concept of this section is to describe the relationship between two variables by finding the graph and the equation of the straight line that best represents the relationship.

The straight line is called a regression line and its equation is called the regression equation.

Part 1: Basic Concepts of Regression

Regression

The regression equation expresses a relationship between \( x \) (called the independent variable, predictor variable or explanatory variable), and \( y \) (called the dependent variable or response variable).

The typical equation of a straight line \( y = mx + b \) is expressed in the form \( \hat{y} = b_0 + b_1x \), where \( b_0 \) is the \( y \)-intercept and \( b_1 \) is the slope.
Requirements

1. The sample of paired \((x, y)\) data is a random sample of quantitative data.
2. Visual examination of the scatterplot shows that the points approximate a straight-line pattern.
3. Any outliers must be removed if they are known to be errors. Consider the effects of any outliers that are not known errors.

Definitions

- **Regression Equation**
  Given a collection of paired data, the regression equation
  \[ \hat{y} = b_0 + b_1 x \]
  algebraically describes the relationship between the two variables.

- **Regression Line**
  The graph of the regression equation is called the regression line (or line of best fit, or least squares line).

Notation for Regression Equation

<table>
<thead>
<tr>
<th>Population Parameter</th>
<th>Sample Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>(y)-intercept of regression equation (\beta_0)</td>
<td>(b_0)</td>
</tr>
<tr>
<td>Slope of regression equation (\beta_1)</td>
<td>(b_1)</td>
</tr>
<tr>
<td>Equation of the regression line (y = \beta_0 + \beta_1 x)</td>
<td>(\hat{y} = b_0 + b_1 x)</td>
</tr>
</tbody>
</table>

Formulas for \(b_0\) and \(b_1\)

- **Formula 10-2**
  \[ b_1 = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2} \] (slope)

- **Formula 10-3**
  \[ b_0 = \bar{y} - b_1 \bar{x} \] (y-intercept)

The regression line fits the sample points best.

Rounding the \(y\)-intercept \(b_0\) and the Slope \(b_1\)

- Round to three significant digits.
- If you use the formulas 10-2 and 10-3, try not to round intermediate values.
Calculating the Regression Equation

In Section 10-2, we used these values to find that the linear correlation coefficient of \( r = -0.956 \). Use this sample to find the regression equation.

Calculating the Regression Equation - cont

The estimated equation of the regression line is:

\( \hat{y} = 8.75 - 1x \)

Example: Old Faithful

Given the sample data in Table 10-1, find the regression equation.

Using the same procedure as in the previous example, we find that \( b_1 = 0.234 \) and \( b_0 = 34.8 \).

Hence, the estimated regression equation is:

\( \hat{y} = 34.8 + 0.234x \)
Part 2: Beyond the Basics of Regression Predictions

In predicting a value of \( y \) based on some given value of \( x \) ... 

1. If there is not a linear correlation, the best predicted \( y \)-value is \( \bar{y} \).

2. If there is a linear correlation, the best predicted \( y \)-value is found by substituting the \( x \)-value into the regression equation.

Procedure for Predicting

Given the sample data in Table 10-1, we found that the regression equation is
\[
\hat{y} = 34.8 + 0.234x
\]
Assuming that the current eruption has a duration of \( x = 180 \) sec, find the best predicted value of \( y \), the time interval after this eruption.

We must consider whether there is a linear correlation that justifies the use of that equation. We do have a significant linear correlation (with \( r = 0.926 \)).

Example: Old Faithful

\[
\hat{y} = 34.8 + 0.234x
\]
\[
\hat{y} = 34.8 + 0.234(180) = 76.9 \text{ min}
\]

The predicted time interval is 76.9 min.

Guidelines for Using The Regression Equation

1. If there is no linear correlation, don’t use the regression equation to make predictions.
2. When using the regression equation for predictions, stay within the scope of the available sample data.
3. A regression equation based on old data is not necessarily valid now.
4. Don’t make predictions about a population that is different from the population from which the sample data were drawn.

Definitions

- **Marginal Change**
  
The marginal change is the amount that a variable changes when the other variable changes by exactly one unit.
- **Outlier**
  
An outlier is a point lying far away from the other data points.
- **Influential Point**
  
An influential point strongly affects the graph of the regression line.
Residual

The residual for a sample of paired \((x, y)\) data, is the difference \((y - \hat{y})\) between an observed sample \(y\)-value and the value of \(y\), which is the value of \(y\) that is predicted by using the regression equation.

\[
\text{residual} = \text{observed } y - \text{predicted } y = y - \hat{y}
\]

Definitions

- **Least-Squares Property**
  A straight line has the least-squares property if the sum of the squares of the residuals is the smallest sum possible.

- **Residual Plot**
  A scatterplot of the \((x, y)\) values after each of the \(y\)-coordinate values have been replaced by the residual value \(y - \hat{y}\). That is, a residual plot is a graph of the points \((x, y - \hat{y})\).

Residual Plot Analysis

If a residual plot does not reveal any pattern, the regression equation is a good representation of the association between the two variables.

If a residual plot reveals some systematic pattern, the regression equation is not a good representation of the association between the two variables.