Chapter 12 - Part I: Correlation Analysis

GOALS
When you have completed this lecture, you will be able to:

ONE
Draw and interpret a scatter diagram.

TWO
Understand and interpret the terms dependent variable, independent variable and spurious correlation.

THREE
Calculate and interpret the coefficient of correlation and the coefficient of determination.

FOUR
Conduct a test of hypothesis to determine if the population correlation is zero (thereby checking for a significant linear relationship between two variables).

Correlation Analysis

- **Correlation Analysis**: A group of statistical techniques used to measure the strength of the relationship (correlation) between two variables.
- **Scatter Diagram**: A chart that portrays the relationship between the two variables of interest.
- **Dependent Variable**: The variable that is being predicted or estimated.
- **Independent Variable**: The variable that provides the basis for estimation. It is the predictor variable.

The Coefficient of Correlation, r

- The Coefficient of Correlation (r) is a sample statistic that measures the strength of linear relationship between two variables.
  - It requires quantitative data
  - It can range from -1.00 to 1.00
  - Values of -1.00 or 1.00 indicate perfect linear correlation.
  - Values close to 0.0 indicate weak linear correlation.
  - Negative values indicate an inverse linear relationship and positive values indicate a direct linear relationship.
**Zero Correlation!**

**Formula for r & Correlation vs. Causation**

\[
    r = \frac{n(\Sigma XY) - (\Sigma X)(\Sigma Y)}{\sqrt{n(\Sigma X^2) - (\Sigma X)^2}[n(\Sigma Y^2) - (\Sigma Y)^2]}
\]

Correlation does not imply causation: Some variables are linearly related, but the relationship is not a causal relationship. In other words, a change in the independent variable doesn’t really cause a change in the dependent variable. We should be careful not to imply that a relationship between two variables is a causal relationship, unless we have conducted an experiment designed to detect such a relationship.

**Coefficient of Determination**

- The Coefficient of Determination, \( r^2 \), measures the proportion of the total variation in the dependent variable \( Y \) that is explained or accounted for by a linear relationship between \( Y \) and the independent variable \( X \).

- The coefficient of determination is the square of the coefficient of correlation, and ranges from 0 to 1.

**EXAMPLE 1**

- An ASCSU senator is concerned about the cost of textbooks. To provide insight into the problem she selects a sample of eight textbooks currently on sale in the bookstore. She is interested in whether the price of a book is related to the number of pages in the text.
### EXAMPLE 1 continued

<table>
<thead>
<tr>
<th>Book</th>
<th>Pages</th>
<th>Price ($)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>500</td>
<td>53</td>
</tr>
<tr>
<td>2</td>
<td>700</td>
<td>78</td>
</tr>
<tr>
<td>3</td>
<td>800</td>
<td>95</td>
</tr>
<tr>
<td>4</td>
<td>600</td>
<td>65</td>
</tr>
<tr>
<td>5</td>
<td>400</td>
<td>47</td>
</tr>
<tr>
<td>6</td>
<td>500</td>
<td>56</td>
</tr>
<tr>
<td>7</td>
<td>600</td>
<td>70</td>
</tr>
<tr>
<td>8</td>
<td>800</td>
<td>92</td>
</tr>
</tbody>
</table>

**Scatter diagram of # pages versus price**

Based on the value of the sample correlation coefficient, what do you think about the strength of the linear relationship between the cost of textbooks and the number of pages in the book?

- The dependent variable for the study is
  \[ Y = \quad \]
- The independent variable is \( X = \quad \)
- To measure the strength of the linear relationship between \( Y \) and \( X \), we compute the correlation coefficient.

\[
r = \frac{n(\Sigma XY) - (\Sigma X)(\Sigma Y)}{\sqrt{n(\Sigma X^2) - (\Sigma X)^2}[n(\Sigma Y^2) - (\Sigma Y)^2]}}
\]
Test the hypothesis that there is no linear relationship for the population. Use a 0.02 significance level.

- Step 1: 
  * $H_0: \rho = 0$ (no linear relationship) 
  * $H_1: \rho \neq 0$ (linear relationship)

- Step 2: $\alpha = 0.02$

- Step 3: Test statistic: 
  \[ t = \frac{r}{\sqrt{1 - r^2}} \]

- Step 4: $df = n - 2$, $\alpha = 0.02$

- Step 5: $H_0$ is rejected if $t >$ or $t <$ 

Conclusion: Since $t = \frac{-2}{\sqrt{1 - (-0.2)^2}}$, the decision is $H_0$ is rejected. 

The population correlation coefficient, $\rho$, is a point estimate for the population correlation coefficient, $\rho$. If the population correlation coefficient is zero, then there is no linear relationship between the two variables.

The sample correlation coefficient, $r$, is computed based on a sample of only 8 textbooks. If we had access to the entire population of textbooks, we could compute the correlation coefficient for the entire population.
Chapter 12 - Part II: Regression Analysis

GOALS
When you have completed this chapter, you will be able to:

ONE
Interpret the components of a regression equation and explain the concept of extrapolation.

TWO
Calculate the regression line, interpret the slope and intercept values, and use the regression equation to predict or estimate values for the dependent variable.

THREE
Calculate and interpret the standard error of estimate.

FOUR
Explain the assumptions required for regression analysis and know how to check them.

Regression Analysis

• Purpose: To determine the regression equation.
• The regression equation is used to predict the value of the dependent variable (Y) based on the independent variable (X).
• Procedure: Select a sample from the population and list the paired data for each observation; draw a scatter diagram to give a visual portrayal of the relationship; determine the regression equation.

Regression Analysis

- the regression equation: \( Y' = a + bX \), where:
  - \( Y' \) is the predicted value of \( Y \) for a given \( X \) value.
  - \( a \) is the \( Y \)-intercept, or the estimated \( Y \) value when \( X = 0 \). It only makes sense to interpret the intercept when \( X = 0 \) is in the range of the observed data.
  - \( b \) is the slope of the line, or the average change in \( Y' \) for each change of one unit in \( X \)
- the least squares principle is used to obtain \( a \) & \( b \):

\[
b = \frac{n(\sum XY) - (\sum X)(\sum Y)}{n(\sum X^2) - (\sum X)^2}
\]
\[
a = \frac{\sum Y}{n} - b \cdot \frac{\sum X}{n}
\]

EXAMPLE 1

- Develop a regression equation for the information given in EXAMPLE 1 from the previous lecture. Use the regression equation to predict the selling price of a book based on the number of pages.

\[
b = \frac{n(\sum XY) - (\sum X)(\sum Y)}{n(\sum X^2) - (\sum X)^2}
\]

\[
a = \frac{\sum Y}{n} - b \cdot \frac{\sum X}{n}
\]
**EXAMPLE 1 continued**

- So the regression equation is
  
  \[ Y' = \]

- Interpretation of the coefficients:
  
  a:
  b:

- Does the regression equation describe the relationship between X & Y perfectly for the sample data?

---

**The Standard Error of Estimate**

- The standard error of estimate measures the scatter, or dispersion, of the observed values around the line of regression

- The formulas that are used to compute the standard error:

  \[
  S_{Y\cdot X} = \sqrt{\frac{\sum(Y - Y')^2}{n - 2}}
  \]

  \[
  = \sqrt{\frac{\sum Y^2 - a(\sum Y) - b(\sum XY)}{n - 2}}
  \]

---

**EXAMPLE 1 continued**

- Determine the standard error of the estimate for the regression equation

  \[
  S_{Y\cdot X} = \sqrt{\frac{\sum Y^2 - a(\sum Y) - b(\sum XY)}{n - 2}}
  \]

---

**EXAMPLE 1 continued**

- Note: In this example, the values of the variable X, number of pages, ranged from 400 to 800.

- Caution: A sample regression equation should only be used for prediction or estimation within the range of X-values that are present in the sample. To predict or estimate outside this range of values is called extrapolation.

- Use the regression equation obtained to estimate the price of a textbook with 500 pages.

- Now estimate the price of a textbook with 50 pages.
Residuals

- A residual is the difference between the actual value for Y and the predicted value, $Y'$.  
  Residual = $Y - Y'$

- For observation #1 in Example 1, the number of pages was 500, and the price was $53. Compute the value of the residual for this observation:

- Every observation in the sample has an associated residual.

Assumptions Underlying Linear Regression

- Assumption 1: For each value of X, there is a group of Y values, and these Y values are *normally distributed*.
- Assumption 2: The *means* of these normal distributions of Y values all lie on the straight line of regression.
- Assumption 3: The *standard deviations* of these normal distributions are equal.
- Assumption 4: The Y values are statistically independent.

Checking Regression Assumptions

- To check Assumption #1: use a normal probability plot of residuals.
  - If the points fall approximately on a line, the assumption of normality seems reasonable.
- To check Assumption #2: use the scatter diagram.
- To check Assumption #3: use a residual plot.
  - plot the residuals on the vertical axis versus the corresponding value of X on the horizontal axis.
  - If the amount of vertical spread appears to change as X increases, then the assumption of equal variance is suspect.
- To check Assumption #4: consider the design of the study. The assumption holds if:
  - a simple random sample of $n$ observations (x,y) was obtained, or if for each selected value of X, or
  - a simple random sample of y-values was obtained.
EXAMPLE 1 continued

Normal Probability Plot of the Residuals
(response is PRICE)

Residuals Versus PAGES
(response is PRICE)