Econ 3790: Statistics Business and Economics

Instructor: Yogesh Uppal
Email: yuppal@ysu.edu
Chapter 14

- Covariance and Simple Correlation Coefficient
- Simple Linear Regression
Covariance

- Covariance between x and y is a measure of relationship between x and y.

\[
\text{cov}(x, y) = \frac{SS_{xy}}{n-1} = \frac{\sum (y - \bar{y})(x - \bar{x})}{n-1}
\]
Reed Auto periodically has a special week-long sale. As part of the advertising campaign Reed runs one or more television commercials during the weekend preceding the sale. Data from a sample of 5 previous sales are shown on the next slide.
Covariance

- **Example: Reed Auto Sales**

<table>
<thead>
<tr>
<th>Number of TV Ads</th>
<th>Number of Cars Sold</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>14</td>
</tr>
<tr>
<td>3</td>
<td>24</td>
</tr>
<tr>
<td>2</td>
<td>18</td>
</tr>
<tr>
<td>1</td>
<td>17</td>
</tr>
<tr>
<td>3</td>
<td>27</td>
</tr>
</tbody>
</table>
**Covariance**

<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
<th>$x - \bar{x}$</th>
<th>$y - \bar{y}$</th>
<th>$(x - \bar{x})(y - \bar{y})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>14</td>
<td>-1</td>
<td>-6</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>24</td>
<td>1</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>18</td>
<td>0</td>
<td>-2</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>17</td>
<td>-1</td>
<td>-3</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>27</td>
<td>1</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Total=10</td>
<td>Total = 100</td>
<td>(SS_{xy} = 20)</td>
<td>(SS_{xy} = 20)</td>
<td>(SS_{xy} = 20)</td>
</tr>
</tbody>
</table>

\[
\text{cov}(x, y) = \frac{SS_{xy}}{n-1} = \frac{20}{4} = 5
\]
Simple Correlation Coefficient

- Simple Population Correlation Coefficient

\[ \rho = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y} \]

\[ -1 \leq \rho \leq +1 \]

- If \( \rho < 0 \), a negative relationship between \( x \) and \( y \).
- If \( \rho > 0 \), a positive relationship between \( x \) and \( y \).
Simple Correlation Coefficient

Since population standard deviations of $x$ and $y$ are not known, we use their sample estimates to compute an estimate of $\rho$.

$$r = \frac{\text{cov}(x, y)}{s_x s_y}$$

$$-1 \leq r \leq +1$$
Simple Correlation Coefficient

Example: Reed Auto Sales

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>y</td>
<td>x - ( \bar{x} )</td>
<td>y - ( \bar{y} )</td>
<td>SS( _x )</td>
</tr>
<tr>
<td>1</td>
<td>14</td>
<td>-1</td>
<td>-6</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>24</td>
<td>1</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>18</td>
<td>0</td>
<td>-2</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>17</td>
<td>-1</td>
<td>-3</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>27</td>
<td>1</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>Total=10</td>
<td>Total=98</td>
<td></td>
<td>Total=4</td>
<td>Total= 114</td>
</tr>
</tbody>
</table>
Simple Correlation Coefficient

\[ s_x = \sqrt{\frac{(x - \bar{x})^2}{n-1}} = \sqrt{\frac{4}{4}} = 1 \]

\[ s_y = \sqrt{\frac{(y - \bar{y})^2}{n-1}} = \sqrt{\frac{114}{4}} = 5.34 \]

\[ r_{xy} = \frac{\text{cov}(x, y)}{s_x s_y} = \frac{5}{1 \times 5.34} = 0.936 \]
Chapter 14 Simple Linear Regression

- Simple Linear Regression Model
- Residual Analysis
- Coefficient of Determination
- Testing for Significance
- Using the Estimated Regression Equation for Estimation and Prediction
Simple Linear Regression Model

- The equation that describes how $y$ is related to $x$ and an error term is called the regression model.
- The simple linear regression model is:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

where:

- $\beta_0$ and $\beta_1$ are called parameters of the model,
- $\varepsilon$ is a random variable called the error term.
Simple Linear Regression Equation

- Positive Linear Relationship

![Diagram of Simple Linear Regression](image)

- Intercept $\beta_0$
- Slope $\beta_1$ is positive
- Regression line
Simple Linear Regression Equation

- Negative Linear Relationship

![Graph showing simple linear regression equation]

- **Intercept** $\beta_0$
- **Slope** $\beta_1$
- **Regression line**

Slope $\beta_1$ is negative
Simple Linear Regression Equation

- No Relationship

Regression line

Intercept $\beta_0$

Slope $\beta_1$ is 0
Interpretation of $\beta_0$ and $\beta_1$

- $\beta_0$ (intercept parameter): is the value of $y$ when $x = 0$.
- $\beta_1$ (slope parameter): is the change in $y$ given $x$ changes by 1 unit.
Estimated Simple Linear Regression Equation

The **estimated simple linear regression equation**

\[ \hat{y} = b_0 + b_1 x \]

- The graph is called the estimated regression line.
- \( b_0 \) is the \( y \) intercept of the line.
- \( b_1 \) is the slope of the line.
- \( \hat{y} \) is the estimated value of \( y \) for a given value of \( x \).
Regression Model
\[ y = \beta_0 + \beta_1 x + \varepsilon \]
Regression Equation
\[ E(y \mid x) = \beta_0 + \beta_1 x \]
Unknown Parameters \( \beta_0, \beta_1 \)

Sample Data:
\[
\begin{array}{cc}
  x & y \\
  x_1 & y_1 \\
  \cdot & \cdot \\
  x_n & y_n \\
\end{array}
\]

Estimated Regression Equation
\[ \hat{y} = b_0 + b_1 x \]

\( b_0 \) and \( b_1 \) provide point estimates of \( \beta_0 \) and \( \beta_1 \)
Least Squares Method

- Slope for the Estimated Regression Equation

\[ b_1 = \frac{SS_{xy}}{SS_y} \]

Where \( SS_{xy} = \sum (y - \bar{y})(x - \bar{x}) \)

and \( SS_x = \sum (x - \bar{x})^2 \)
Least Squares Method

- $y$-Intercept for the Estimated Regression Equation

\[ b_0 = \bar{y} - b_1 \bar{x} \]

$\bar{x} = \text{mean value for independent variable}$

$\bar{y} = \text{mean value for dependent variable}$
Estimated Regression Equation

- **Example: Reed Auto Sales**
  - Slope for the Estimated Regression Equation
    \[ b_1 = \frac{SS_{xy}}{SS_x} = \frac{20}{4} = 5 \]
  - \( y \)-Intercept for the Estimated Regression Equation
    \[ b_0 = \bar{y} - b_1 \bar{x} = 20 - 5 \times 2 = 10 \]
  - Estimated Regression Equation
    \[ \hat{y} = 10 + 5 \times x \]
Scatter Diagram and Regression Line

\[ \hat{y} = 10 + 5x \]

<table>
<thead>
<tr>
<th>Number of Ads</th>
<th>Number of Cars Sold</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>12</td>
</tr>
<tr>
<td>1.0</td>
<td>14</td>
</tr>
<tr>
<td>1.5</td>
<td>16</td>
</tr>
<tr>
<td>2.0</td>
<td>18</td>
</tr>
<tr>
<td>2.5</td>
<td>20</td>
</tr>
<tr>
<td>3.0</td>
<td>22</td>
</tr>
<tr>
<td>3.5</td>
<td>24</td>
</tr>
</tbody>
</table>
## Estimate of Residuals

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>y</td>
<td>(\hat{y})</td>
<td>(e = y - \hat{y})</td>
</tr>
<tr>
<td>1</td>
<td>14</td>
<td>15</td>
<td>-1.0</td>
</tr>
<tr>
<td>3</td>
<td>24</td>
<td>25</td>
<td>-1.0</td>
</tr>
<tr>
<td>2</td>
<td>18</td>
<td>20</td>
<td>-2.0</td>
</tr>
<tr>
<td>1</td>
<td>17</td>
<td>15</td>
<td>2.0</td>
</tr>
<tr>
<td>3</td>
<td>27</td>
<td>25</td>
<td>2.0</td>
</tr>
</tbody>
</table>
Decomposition of total sum of squares

- Relationship Among SST, SSR, SSE

\[
\text{SST} = \text{SSR} + \text{SSE}
\]

\[
\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2
\]

where:

- SST = total sum of squares
- SSR = sum of squares due to regression
- SSE = sum of squares due to error
## Decomposition of total sum of squares

\[ e = y - \hat{y} \]

<table>
<thead>
<tr>
<th></th>
<th>( SSE = \sum (y - \hat{y})^2 )</th>
<th>( \hat{y} )</th>
<th>( \hat{y} - \bar{y} )</th>
<th>( SSR = \sum (\hat{y} - \bar{y})^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>1</td>
<td>15</td>
<td>-5</td>
<td>25</td>
</tr>
<tr>
<td>-1</td>
<td>1</td>
<td>25</td>
<td>5</td>
<td>25</td>
</tr>
<tr>
<td>-2</td>
<td>4</td>
<td>20</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>15</td>
<td>-5</td>
<td>25</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>25</td>
<td>5</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td><strong>SSE=14</strong></td>
<td></td>
<td></td>
<td><strong>SSR=100</strong></td>
</tr>
</tbody>
</table>

- Check if SST = SSR + SSE
Coefficient of Determination

- The coefficient of determination is:
  \[ r^2 = \frac{SSR}{SST} \]

- \[ r^2 = \frac{100}{114} = 0.8772 \]

- The regression relationship is very strong; about 88% of the variability in the number of cars sold can be explained by the number of TV ads.

- The coefficient of determination \( (r^2) \) is also the square of the correlation coefficient \( (r) \).
Sample Correlation Coefficient

\[ r = \text{(sign of } b_1) \sqrt{\text{Coefficient of Determination}} \]

\[ r = (\text{sign of } b_1) \sqrt{r^2} \]

\[ r = + \sqrt{0.8772} = 0.936 \]
Sampling Distribution of $b_1$

- Expected value of $b_1$:
  \[ E(b_1) = b_1 \]

- Variance of $b_1$:
  \[ \sigma_1^2 = \text{Var}(b_1) = \sigma^2/SS_x \]

- Standard error of $b_1$:
  \[ \sigma_1 = \sqrt{\text{Var}(b_1)} \]
Estimate of $\sigma^2$

- The mean square error (MSE) provides the estimate of $\sigma^2$.

\[ s^2 = \text{MSE} = \frac{\text{SSE}}{(n - 2)} \]

where:

\[ \text{SSE} = \sum (y_i - \hat{y}_i)^2 \]
Interval Estimate of $\beta_1$:

- $(1-\alpha)100\%$ confidence interval for $b_1$ is:

$$b_1 \pm t_{\alpha/2} \times \sigma_1$$

- Where $t_{\alpha/2}$ is the value from $t$ distribution with $(n-2)$ degrees of freedom such that probability in the upper tail is $\alpha/2$. 
Example: Reed Auto Sales

- \( s^2 = \text{MSE} = \frac{\text{SSE}}{(n - 2)} = \frac{14}{3} = 4.67 \)

\[
\text{Var}(b_1) = \frac{s^2}{\text{SS}_x} = \frac{4.67}{4} = 1.17 \\
\sigma_1 = \sqrt{1.17} = 1.08
\]

- 95% confidence interval for \( \beta_1 \):

\[
5 \pm 3.182 \times 1.08 = 5 \pm 3.44
\]

- We can say we 95% confidence that \( \beta_1 \) will lie between 1.56 and 8.44.
Testing for Significance: \( t \) Test

- **Hypotheses**
  - \( H_0: \beta_1 = 0 \)
  - \( H_a: \beta_1 \neq 0 \)

- **Test Statistic**
  \[
  t = \frac{b_1 - 0}{SE(b_1)}
  \]

- Where \( b_1 \) is the slope estimate and \( SE(b_1) \) is the standard error of \( b_1 \).
Testing for Significance: \( t \) Test

- Rejection Rule

\[
\text{Reject } H_0 \text{ if } p\text{-value} \leq \alpha \\
\text{or } t \leq -t_{\alpha/2} \text{ or } t \geq t_{\alpha/2}
\]

where:

\( t_{\alpha/2} \) is based on a \( t \) distribution with \( n - 2 \) degrees of freedom
Testing for Significance: \( t \) Test

1. Determine the hypotheses.
   \[ H_0: \beta_1 = 0 \]
   \[ H_a: \beta_1 \neq 0 \]

2. Specify the level of significance.
   \[ \alpha = .05 \]

3. Select the test statistic.
   \[ t = \frac{b_1}{SE(b_1)} \]

4. State the rejection rule.
   Reject \( H_0 \) if \( p\)-value \leq .05
   or \( t \leq 3.182 \) or \( t \geq 3.182 \)
Testing for Significance: *t* Test

5. Compute the value of the test statistic.

\[ t = \frac{b_1}{SE(b_1)} = \frac{5}{1.08} = 4.63 \]

6. Determine whether to reject \( H_0 \).

\[ t = 4.63 > t_{\alpha/2} = 3.182. \text{ We can reject } H_0. \]
Some Cautions about the Interpretation of Significance Tests

- Rejecting $H_0: \beta_1 = 0$ and concluding that the relationship between $x$ and $y$ is significant does not enable us to conclude that there is a cause-and-effect relationship is present between $x$ and $y$.

- Just because we are able to reject $H_0: \beta_1 = 0$ and demonstrate statistical significance does not enable us to conclude that there is a linear relationship between $x$ and $y$. 