CH 12: Simple Linear Regression Model Part 1

1. The Simple Linear Regression Model (Population):

\[ y = \beta_0 + \beta_1 x + \varepsilon \]

where \( \beta_0 \) and \( \beta_1 \) are the population parameters. Moreover, \( \beta_0 \) is the y-intercept; \( \beta_1 \) is the slope; and \( \varepsilon \) is the random error in \( y \).

2. The Simple Linear Regression Model (Sample):

(A) Scatter diagram:

Given paired observations \((x_i, y_i)\), a scatter diagram uses the \( x \) and \( y \)-axes to represent the data.

(B) We use \( r \) (correlation coefficient) to measure the strength of the linear relation between the \( x \) variable and \( y \) variable:

\[ r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \]

(C) We want to find the relationship between \( x \) and \( y \) by fitting a line to the data set.

eq12.3: Estimated Simple Regression Equation: \( \hat{y} = b_0 + b_1 x \)

(D) Linear regression equation:

At each observation, the predicted value of \( y \) is given by: \( \hat{y}_i = b_0 + b_1 x_i \)

where \( b_0 \) and \( b_1 \) are regression coefficients. Moreover,

\( b_0 \) is the y-intercept: the average value of \( y \) when \( x = 0 \).

\( b_1 \) is the slope.

\( \hat{y}_i \) is the predicted value of \( y \) for observation \( i \).

\( x_i \) is the value of \( x \) for observation \( i \).

(E) We use the least squares method to compute \( b_0 \) and \( b_1 \):

(a) In this case, we minimize \( \sum(y_i - \hat{y}_i)^2 \).

(b) Using differential calculus, we can obtain the following results:

eq12.6: The Slope \( b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \)

eq12.7: The Y-intercept: \( b_0 = \bar{y} - b_1 \bar{x} \)

EX1 Given are five observations for two variables \( x \) and \( y \):

<table>
<thead>
<tr>
<th>( x )</th>
<th>( 1 )</th>
<th>( 2 )</th>
<th>( 3 )</th>
<th>( 4 )</th>
<th>( 5 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( y )</td>
<td>( 3 )</td>
<td>( 7 )</td>
<td>( 5 )</td>
<td>( 11 )</td>
<td>( 14 )</td>
</tr>
</tbody>
</table>

(a) Develop a scatter diagram and approximate the relationship between \( x \) and \( y \) by drawing a straight line through the data.

(b) Compute \( b_0 \) and \( b_1 \).

(c) Intercept the regression equation and predict the average value of \( y \) when \( x = 5 \).

3. Three important measures of variation

(1) Sum of squares Due to Error (SSE): measure of the error in using the estimated regression equation to estimate the values of the dependent variable in the sample.

\[ \text{eq12.8: Sum of squares Due to Error: } \text{SSE} = \sum(y_i - \hat{y}_i)^2 \]

(2) Total sum of squares (SST): Measure of variation of the \( y_i \) values around their mean \( \bar{y} \).

\[ \text{eq12.9: Total sum of squares: } \text{SST} = \sum(y_i - \bar{y})^2 \]

(3) Sum of squares Due to Regression (SSR): measure of variation attributable to the relationship between \( X \) and \( Y \).

\[ \text{eq12.10: Sum of Squares Due to Regression: } \text{SSR} = \sum(\hat{y}_i - \bar{y})^2 \]

\[ \text{eq12.11: Relationship Among SST,SSR,and SSE: } \text{SST} = \text{SSR} + \text{SSE} \]

4. Coefficient of Determination (Notation: \( r^2 \)): A measure of the goodness of fit of the estimated regression equation. It can be interpreted as the proportion of the variability in the dependent variable \( y \) that is explained by \( x \) in the estimated regression equation.

\[ \text{eq12.12: Coefficient of determination: } r^2 = \frac{\text{SSR}}{\text{SST}} \]

5. Standard error of estimate (Notation: \( s \)): Measures how much the data vary around the regression line. Its the square root of the mean square error (MSE).

\[ \text{eq12.16: Standard error of the estimate: } s = \sqrt{\frac{\text{SSE}}{n-2}} \]

EX 2 Given \( \text{SSR} = 66 \), \( \text{SST} = 88 \) and \( n = 22 \), (a) compute the coefficient of determination and interpret its meaning. (b) Find the standard error of estimate \( s \).
6. Population Model Assumptions:

\[ y_i = \beta_0 + \beta_1 x_i + \varepsilon \]

where \( \beta_0 \) and \( \beta_1 \) are the population parameters. Moreover, \( \beta_0 \) is the y-intercept; \( \beta_1 \) is the slope; and \( \varepsilon \) is the random error in \( y \) (assumed to be normally distributed with \( E(\varepsilon) = 0 \) and \( \text{var}(\varepsilon) = \sigma^2 \)).

Note:

7. Testing for Significance: We use \( t \)-test for the slope \( \beta_1 \) to determine the existence of a significant linear relationship between the \( x \) and \( y \) variables.

Step 1: State \( H_0 \) vs. \( H_1 \).

Step 2: Compute the test statistic and critical value.

\[ t_{\text{cal}} = \frac{b_1}{s_{b_1}} \text{ with } (n-2) \text{ degrees of freedom} \]

Step 3: Make a decision using either \( p \)-value approach or the critical value approach.

EX 3 Given \( SSR = 27.51 \), \( SST = 41.27 \), \( \sum (x_i - \bar{x})^2 = 18.38 \), \( \bar{y} = 3.0 + 0.5x_i \), and \( n = 20 \). Use the \( t \) test to test the existence of a linear relationship between \( x \) and \( y \) (\( \alpha = 0.05 \)).

Step 1: State \( H_0 \) and \( H_1 \).

Step 2 Compute the test statistic

EX 3 (cont.) Find the 95% confidence interval for the true slope \( \beta_1 \).

8. \( F \) test for significance in simple linear regression

Step 1: State \( H_0 \) and \( H_1 \).

Step 2 Compute the test statistic

\[ F = \frac{MSR}{MSE} \]

Step 3 Make a decision

9. The 100(1 - \( \alpha \))% confidence interval for the slope \( \beta_1 \):

\[ b_1 \pm t_{n-2\beta_1} \]

EX 4: To study the relationship between the size of a store \( (x, \text{in 1000 square feet}) \) and its annual sales \( (y, \text{in $1,000,000}) \). We randomly selected 14 store and obtained the following data: \( \hat{y} = 0.964 + 1.670x \), \( SST = 116.95 \), \( SSR = 105.75 \), and \( \sum (x_i - \bar{x})^2 = 37.924 \). Obtain a 95% confidence interval of the average annual sales for a store that is 4000 square feet (with \( \bar{x} = 2.921 \)).

Step 1: Find \( \hat{y} \)

Step 2: Find the standard error of the estimate \( s \):

Step 3 find the critical value \( t_{n-2}/2 \)

Step 4: Find the confidence interval

10. 100(1 - \( \alpha \))% CI for the mean value of \( y \) \( (E(y^*)) \) for a given value of \( x^* \):

\[ y^* \pm t_{n-2/2} s_p \]

with \( s_p = s \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum (x_i - \bar{x})^2}} \)

EX 4: To study the relationship between the size of a store \( (x, \text{in 1000 square feet}) \) and its annual sales \( (y, \text{in $1,000,000}) \). We randomly selected 14 store and obtained the following data: \( \hat{y} = 0.964 + 1.670x \), \( SST = 116.95 \), \( SSR = 105.75 \), and \( \sum (x_i - \bar{x})^2 = 37.924 \). Obtain a 95% confidence interval of the average annual sales for a store that is 4000 square feet (with \( \bar{x} = 2.921 \)).

Step 1 : find \( \hat{y} \)

Step 2: find the standard error of the estimate \( s \):

Step 3 find the critical value \( t_{n-2/2} \)

Step 4: Find the confidence interval

11. How to read Excel computer output for the simple regression model.

<table>
<thead>
<tr>
<th>coefs.</th>
<th>stand. error</th>
<th>t-test</th>
<th>p-value</th>
<th>lower 95%</th>
<th>upper 95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept 0.964</td>
<td>0.526</td>
<td>1.832</td>
<td>0.004</td>
<td>-0.182</td>
<td>2.111</td>
</tr>
<tr>
<td>x value 1.670</td>
<td>0.157</td>
<td>10.641</td>
<td>1.267E-7</td>
<td>1.328</td>
<td>2.012</td>
</tr>
</tbody>
</table>