CH 14: Simple Linear Regression Model Part 1

1. The Simple Linear Regression Model (Population):

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

where β_0 and β_1 are the population parameters. Moreover, β_0 is the y-intercept; β_1 is the slope; and ε 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-axies 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:

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

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

eq14.4: 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: $\widehat{y_i} = b_0 + b_1 x_i$
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.
$\widehat{y_i}$ 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 $\sum(y_i - \widehat{y_i})^2$.

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

eq14.6: The Slope $b_1 = \dfrac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sum(x_i - \overline{x})^2}$

eq14.7: The Y-intercept: $b_0 = \overline{y} - b_1 \overline{x}$

EX1 Given are five observations for two variables x and y.

(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.

eq14.8: Sum of squares Due to Error: $SSE = \sum(y_i - \hat{y}_i)^2$

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

eq14.9: Total sum of squares: $SST = \sum(y_i - \overline{y})^2$

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

eq14.10: Sum of Squares Due to Regression: $SSR = \sum(\hat{y}_i - \overline{y})^2$

eq14.11: Relationship Among SST,SSR,and SSE: $SST = SSR + SSE$

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.

eq14.12: Coefficient of determination: $r^2 = \frac{SSR}{SST}$

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).

eq14.16: Standard error of the estimate: $s = \sqrt{\frac{SSE}{n-2}}$

EX 2 Given $SSR = 66$, $SST = 88$ and $n = 22$, (a) compute the coefficient of determination and interpret its meaning. (b) Find the standard error of estimate s.

CH 14: Simple Linear Regression Model Part 2: Hypotheses Test and Confidence Intervals:

6. Population Model Assumptions:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon$$

where β_0 and β_1 are the population parameters. Moreover, β_0 is the y-intercept; β_1 is the slope; and ε is the random error in y (assumed to be normally distributed with $E(\varepsilon) = 0$ and $var(\varepsilon) = \sigma^2$).

Note:

7. Testing for Significance: We use t-test for the slope β_1 to determine the existence of a significant linear relationship between the x and y variables.

Step1: State H_0 vs. H_1.

Step2: Compute the test statistic and critical value.

$$\text{eq14.19: } t \text{ Test Statistic } t_{cal} = \frac{b_1}{S_{b_1}} \text{ with } (n-2) \text{ degrees of freedom}$$

where eq14.18: Estimated Standard Deviation of b_1: $s_{b_1} = \dfrac{s}{\sqrt{\sum(x_i - \overline{x})^2}}$

Step3: Make a decision using either p-value approach or the critical value approach.

EX 3 Given $SSR = 27.51$, $SST = 41.27$, $\sum(x_i - \overline{x})^2 = 18.38$, $\hat{y}_i = 3.0 + 0.5x_i$, 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

Step 3 Make a decision

8. F test for significance in simple linear regression

Step 1: State H_0 and H_1

Step 2 Compute the test statistic

$$\text{eq14.20: Mean Square Regression: } MSR = \frac{SSR}{\#indvar}$$

$$\text{eq14.21: } F \text{ Test Statistic: } F = \frac{MSR}{MSE}$$

Step 3 Make a decision

9. The $100(1-\alpha)\%$ confidence interval for the slope β_1:

$$b_1 \pm t_{\alpha/2} s_{b_1}$$

EX 3 (cont.) Find the 95% confidence interval for the true slope β_1

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

$$\text{eq14.24: } \hat{y^*} \pm t_{\alpha/2}\ s_{\hat{y}^*}$$

with eq12.23: $s_{\hat{y}^*} = s\sqrt{\dfrac{1}{n} + \dfrac{(x_p - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$

EX 4: To study the relationship between the size of a store (x, in 1000 square feet) and its annual sales (y, 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 $t_{\alpha/2}$

Step 4: Find the confidence interval

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