THE TEACHING OF CONFIDENCE INTERVAL AND PREDICTION INTERVAL

Pooi Ah Hin and Teh Ping Ping
University of Malaya
Malaysia

W*e attempt to use simulation to teach confidence interval for the slope parameter and prediction interval for the future observation in the simple linear regression model. Computer program in JAVA is written to illustrate the simulation in a step-by-step manner. As the observation vector is being generated, the scatter diagram, fitted line and confidence interval will be displayed. Histograms for the individual observations will be built up slowly and the proportion of confidence intervals covering the true value will be updated .The students will realize that the proportion tends to the desired value. Similarly prediction interval can be taught by using simulation .The students will realize that the average value of the proportion of the future observations falling inside the prediction interval tends to the desired value.*

1. INTRODUCTION

Suppose there are two variables x and y where x is known exactly but y is a sum of two components given by a fixed value and a random error. Assume that the fixed value is a linear function of x. Then the variable y is related to x through the following equation

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

where β_0, β_1 are unknown constants (parameters) and ε is the random error. The variable x is called the independent variable and y the dependent variable.

Let $x_1, x_2, \ldots, x_n$ be n selected values of x and $y_1, y_2, \ldots, y_n$ be respectively the corresponding values of y. Then

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \ldots, n. \qquad (1)$$

where $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are assumed to be independent and normally distributed with mean zero and variance σ^2. Equation (1) is called a simple linear regression model.

When $y_1, y_2, \ldots, y_n$ are given, the usual estimates $\hat\beta_0$, $\hat\beta_1$, and $\hat\sigma^2$ for the parameters β_0, β_1 and σ^2 are given respectively by

$$\hat\beta_1 = [\Sigma x_i y_i - n^{-1}(\Sigma x_i)(\Sigma y_i)]/[\Sigma x_i^2 - n^{-1}(\Sigma x_i)^2]$$

$$\hat\beta_0 = \frac{1}{n}\Sigma y_i - \hat\beta_1\left(\frac{1}{n}\Sigma x_i\right) \quad \text{and} \quad \hat\sigma^2 = \frac{1}{n-2}\Sigma[y_i - (\hat\beta_0 + \hat\beta_1 x_i)]^2 .$$

A $100(1 - \alpha)\%$ confidence interval for β_1 is given by

$$\{\beta_1 : \hat\beta_1 - t_{\alpha/2;n-2} S_{\hat\beta_1} < \beta_1 < \hat\beta_1 + t_{\alpha/2;n-2} S_{\hat\beta_1}\}$$

where $S_{\hat\beta_1} = \hat\sigma / \{\Sigma(x_i - \bar x)^2\}^{1/2}$, $\quad \bar x = n^{-1}\sum x_i$,

and $t_{\alpha/2;n-2}$ is the $100(1 - \alpha/2)\%$ point of the t-distribution with $(n - 2)$ degrees of freedom. An important property of this interval is that it will cover the true value of β_1 with probability $1 - \alpha$.

Let x* be a chosen value of x. When x = x*, the corresponding value of y may be denoted by y*. We refer to y* as the future observation. A $100(1 -\gamma)\%$ prediction interval for y* is given by

$$\{y^* : \hat\eta^* - t_{\gamma/2;n-2} S^* < y^* < \hat\eta^* + t_{\gamma/2;n-2} S^*\}$$

where $\hat\eta^* = \hat\beta_0 + \hat\beta_1 x^*$

and $$S^* = \hat{\sigma}\left\{1 + \frac{1}{n} + (x^* - \bar{x})^2 / \Sigma(x_j - \bar{x})^2\right\}^{1/2}.$$

An important property of this interval is that it will cover the future observation with an average probability of $1 - \gamma$.

While it is fairly easy to understand the meaning of the estimates $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\sigma}^2$, it is not easy to comprehend the properties of the confidence interval and prediction interval. One of the reasons is that in stating the properties, the word "probability" has been used. A way to explain "probability" is by means of simulation illustrated in a step-by-step manner. In the next two sections, JAVA programs shall be introduced to help the students understand the properties of the confidence interval and prediction interval.

2. JAVA PROGRAM FOR TEACHING CONFIDENCE INTERVAL

A computer program in JAVA has been written to generate the observation vector (y_1, y_2, …, y_n) repeatedly, and find the confidence interval which corresponds to each of the generated observation vectors. As the observation vectors are being generated, the histogram for the i-th observation y_i will be built up for each i = 1, 2, …, n, the confidence intervals will be drawn, and the proportion of confidence intervals which will cover the true value of the parameter will be updated.

In the program, the true values of β_0, β_1, and σ are chosen to be respectively 3.0, 2.0 and 1.0. Furthermore the value of n is chosen to be 10, $x_i = i$, i = 1, 2, …, 10, and the level α is set to 0.05.

When the program is run, a window will appear. At the bottom of the window, there is a field preceded by the string "Enter command:". The user may enter the command "GD" by typing the command name and pressing the "Enter" key.

When the command "GD" is entered, a set of values of y_1, y_2, …, y_{10} will be generated and presented in the column labeled "old" in the middle-upper portion of the window. A scatter plot of (x_i, y_i), i = 1, 2, …, 10 will appear in the left-upper portion of the window. There will be two straight lines drawn in the graph. The red line is the actual line y = 3 + 2x, while the blue line is the fitted regression line $y = \hat{\beta}_0 + \hat{\beta}_1 x$.

The confidence interval for β_1 will appear below the columns labeled as "old" and "new" respectively. This confidence interval will be plotted in the left-lower portion of the window. In the middle-lower portion of the window the number of observation vectors generated and the proportion of confidence intervals which cover the true value (i.e. 2) of the parameter β_1 will be displayed. The right portion of the window will show the histograms for y_1, y_2, …, y_{10} respectively. When a specific portion of the window is left-clicked once using the mouse, that specific portion will be enlarged and the existing contents of the window will be replaced by this enlarged diagram. With a left-click of the enlarged diagram, the original contents of the window will be restored.

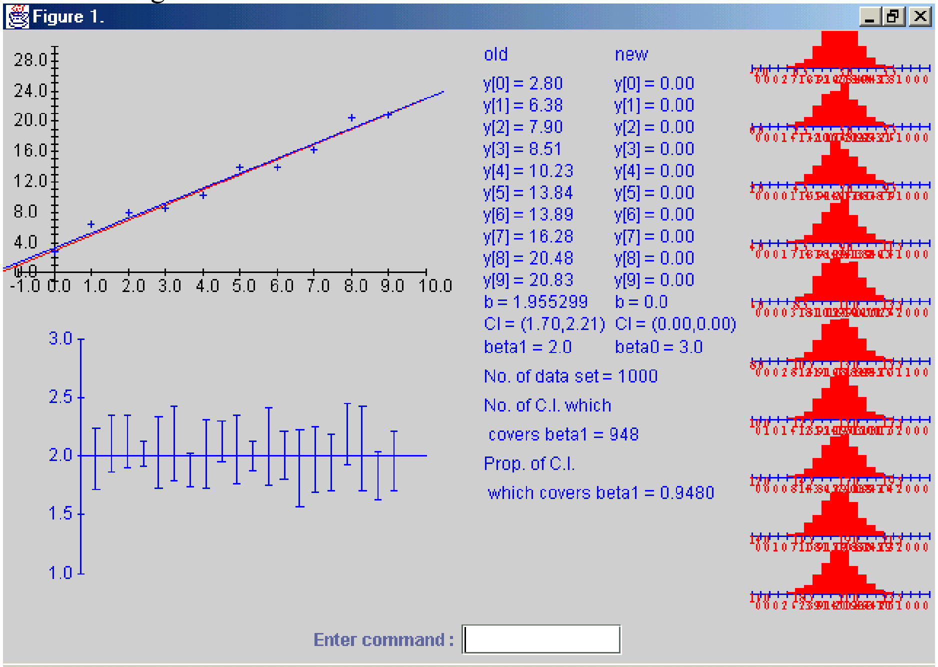
Next the following commands "GND", "DNL", "UH0", "UH1", …, "UH9", "FCI", "PCI", "UND", "UNC", "UPC", "EOD", "EOL", "SHD" may be entered one by one and in the indicated order.

The command "GND" will generate yet another observation vector which will then appear in the column labeled "new". When the command "DNL" is entered, the fitted regression line based on the new observation vector will be plotted as a white line. The command "UHi" is for drawing the histogram for y_{i+1} once more in the light of the additional value of y_{i+1}, i = 0, 1, 2, …, 9. The command "FCI" is for finding the confidence interval for the newly generated observation vector. "PCI" is the command used for plotting the newly computed confidence interval. When the command "UND" is entered, the total number of observation vectors generated will be updated. The commands "UNC" and "UPC" will respectively update the number and the proportion of confidence intervals which cover the true value of β_1. The command "EOD" is for erasing the old data from the column labeled "old". When the command "EOL" is entered, the fitted regression line (blue) based on the old observation vector will be deleted. The command “SHD” is for shifting the data from the “new” column to the “old” column.

So far two observation vectors have been generated. To continue the simulation process using more observation vectors, the commands (“GND”, “DNL”, “UH0”, …, “SHD”) may be entered repeatedly.

By going through the simulation process in a step-by-step manner, one can observe the randomness of each y_i, and the gradual change and development of the histogram of y_i towards a bell-shaped distribution centered at $\beta_0 + \beta_1 x_i$. One can also observe the randomness of the centers and widths of the confidence intervals, and the gradual change and development of the proportion of confidence intervals which cover the true value of β_1, towards the target value 0.95.

When one has a good understanding of the simulation process, one may enter the command “AUTO”. With this, the program will perform the simulation process continuously using a total of 1000 observation vectors. The window after the completion of the simulation may look like the one in Figure 1.



```
old                 new
y[0] = 2.80         y[0] = 0.00
y[1] = 6.38         y[1] = 0.00
y[2] = 7.90         y[2] = 0.00
y[3] = 8.51         y[3] = 0.00
y[4] = 10.23        y[4] = 0.00
y[5] = 13.84        y[5] = 0.00
y[6] = 13.89        y[6] = 0.00
y[7] = 16.28        y[7] = 0.00
y[8] = 20.48        y[8] = 0.00
y[9] = 20.83        y[9] = 0.00
b = 1.955299        b = 0.0
CI = (1.70,2.21)    CI = (0.00,0.00)
beta1 = 2.0         beta0 = 3.0
No. of data set = 1000
No. of C.I. which
 covers beta1 = 948
Prop. of C.I.
 which covers beta1 = 0.9480
```

Figure 1. Example of the Completed Simulation Process.

To end the execution of the program, the command “END” may be used.

3. JAVA PROGRAM FOR TEACHING PREDICTION INTERVAL

A computer program in JAVA has been written to generate the observation vector (y_1, y_2, …, y_n) repeatedly, and find the prediction interval which corresponds to each of the generated observation vectors. As the observation vectors are being generated, the histogram for the i-th observation y_i will be built up for each i = 1, 2, …, n, and the prediction intervals will be drawn. Furthermore the probability that the future observation will fall inside the prediction interval will be computed and the average value of this probability over the set of all the observation vectors generated will be updated.

In the program, the chosen values for β_0, β_1, σ, x_i, n, and α are the same as those in Section 2.

When the program is run, a window will appear. The user may first enter the command “GD”. When the value for x* has been entered, the values of the generated observations, the scatter plot, the fitted regression line and the histograms for the y_i will be displayed in the window. Also, the end points c and d and the coverage probability of the prediction interval will be displayed above the normal distribution curve for the future observation y*. Finally, the histogram for the coverage probability will be drawn.

Next the following commands “GND”, “UH0”, "UH1”, …, “UH9”, “EOD”, “SHD” may be entered one by one and in the indicated order.

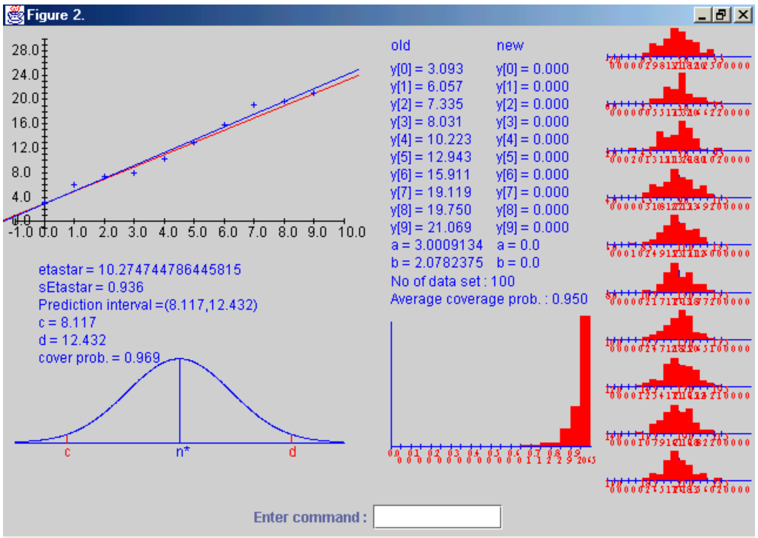
The command “GND” will generate yet another observation vector which will then appear in the column labeled “new”. Meanwhile, the fitted regression line based on the new observation vector will be plotted as a white line, the corresponding prediction interval together with its coverage probability will be displayed, and the average coverage probability together with the histogram for the coverage probability will be updated.

The commands “EOD” and “SHD” have the same roles as those in Section 2.

To continue the simulation process using more observation vectors, the commands (‘GND”, “UH0”, …, “SHD”) may be entered repeatedly.

By going through the simulation process in a step-by-step manner, one can observe the randomness of the centers, widths and coverage probabilities of the prediction intervals, and the gradual change and development of the average coverage probability towards the target value 0.95.

When one has understood the simulation process, one may enter the command “AUTO” to perform the simulation process continuously using a total of 100 observation vectors. The window after the completion of the simulation may resemble the one shown in Figure 2.



```
old                 new
y[0] = 3.093        y[0] = 0.000
y[1] = 6.057        y[1] = 0.000
y[2] = 7.335        y[2] = 0.000
y[3] = 8.031        y[3] = 0.000
y[4] = 10.223       y[4] = 0.000
y[5] = 12.943       y[5] = 0.000
y[6] = 15.911       y[6] = 0.000
y[7] = 19.119       y[7] = 0.000
y[8] = 19.750       y[8] = 0.000
y[9] = 21.069       y[9] = 0.000
a = 3.0009134       a = 0.0
b = 2.0782375       b = 0.0
No of data set : 100
Average coverage prob. : 0.950
```

```
etastar = 10.274744786445815
sEtastar = 0.936
Prediction interval =(8.117,12.432)
c = 8.117
d = 12.432
cover prob. = 0.969
```

Figure 2. . The window after the completion of the simulation