Chapter 1 Simple Linear Regression (Part 3)

1 Write an Estimated model

Statisticians/Econometricians usually write an estimated model together with some inference statistics, the following are some formats people write a model

(i)

$$\begin{array}{ccccc} \hat{Y} & = & b_0 & + & b_1 \; X \\ \text{(S.E.)} & & (s(b_0)) & & (s(b_1)) \end{array}$$

$$\hat{\sigma}^2(\text{or MSE}) = ..., \quad R^2 = ..., \\ \text{F-statistic} = ... \text{ (and others)}$$

(ii)

$$\begin{array}{ccccc} \hat{Y} & = & b_0 & + & b_1 X \\ \text{(t-values)} & & (t_0) & & (t_1) \end{array}$$

$$\hat{\sigma}^2(\text{or MSE}) = ..., \quad R^2 = ..., \\ \text{F-statistic} = ... \text{ (and others)}$$

(iii)

$$\begin{array}{ccccc} \hat{Y} & = & b_0 & + & b_1 X \\ \text{(p-values)} & & (p_0) & & (p_1) \end{array}$$

$$\hat{\sigma}^2(\text{or MSE}) = ..., \quad R^2 = ..., \\ \text{F-statistic} = ... \text{ (and others)}$$

For simple linear regression model, a plot showing the regression is also necessary.

Example 1.1 The Toluca Company manufactures refrigeration equipments as well as many replacement parts. In the past, one of the replacement parts has been produced periodically in lots of varying sizes. When a cost improvement program was undertaken, company officials wished to determine the optimum lot size for producing this part. The production of this part involves setting up the production process (which must be done no matter what is the lot size) and machining and assembly operations. One key input for the model to ascertain the optimum lot size was the relationship between lot size and labour hours required to produce the lot. To determine this relationship, data on lot size and work hours for 25 recent production runs were utilized. The production conditions were stable during the six-month period in which the 25 runs were made and were expected to continue to be the same during the next three years, the planning period for which the cost improvement program was being conducted. The data was collected and listed in **(data010301.dat)**.

Let X denote the lot size and Y work hours. Based on the problem, we consider the following regression model for the $n = 25$ observations

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, 2, ...,$$

The estimated model is (see **(R code)**)

$$\begin{array}{ccccc} \hat{Y} & = & 62.366 & + & 3.570X \\ \text{(S.E.)} & & (26.177) & & (0.347) \end{array}$$

$$\text{MSE} = 2383.392, \quad R^2 = 0.8215, \quad \text{F-statistic} = 105.9$$

2 Interval Prediction (Estimation) of $E(Y)$ (also called narrow intervals)

Recall the model is

$$Y = \underbrace{\beta_0 + \beta_1 X}_{predictable} + \underbrace{\varepsilon}_{unpredictable}$$

Note that $EY = \beta_0 + \beta_1 X$. In other words, only the mean of Y can be predicted.

Based on n observations, $(X_1, Y_1), ..., (X_n, Y_n)$ we have estimator

$$b_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}, \quad b_0 = \bar{Y} - b_1 \bar{X}.$$

The fitted model

$$\hat{Y} = b_0 + b_1 X$$

is actually a prediction of the mean $E(Y)$. We call it point prediction (estimation) of EY (for a new X) (but non-statisticians also call it "prediction of Y").

Sample distribution of $\hat{Y}$: For the simple linear regression model with independent identical normal error assumptions (i.e. model (1) in part 2 of Chapter 1), we have

$$\hat{Y} \sim N(EY, \sigma^2 [\frac{1}{n} + \frac{(X-\bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}])$$

[Proof: Since $\hat{Y}$ is linear combination of IID normal distributed random variables. We only to check its mean and variance. We know

$$E\hat{Y} = \beta_0 + \beta_1 X = EY$$

Let

$$k_i = \frac{X_j - \bar{X}}{\sum_{j=1}^n (X_j - \bar{X})^2}$$

and then $\sum_{i=1}^n k_i = 0$, and

$$b_1 = \beta_1 + \sum_{i=1}^n k_i \varepsilon_i.$$

Thus

$$\begin{aligned}
\hat{Y} &= b_0 + b_1 X = \bar{Y} - b_1 \bar{X} + b_1 X = \beta_0 + \beta_1 \bar{X} + \frac{1}{n}\sum_{i=1}^n \varepsilon_i + (X - \bar{X}) b_1 \\
&= \beta_0 + \beta_1 \bar{X} + \frac{1}{n}\sum_{i=1}^n \varepsilon_i + (X - \bar{X})(\beta_1 + \sum_{i=1}^n k_i \varepsilon_i) \\
&= \beta_0 + \beta_1 X + \sum_{i=1}^n [\frac{1}{n} + (X - \bar{X}) k_i] \varepsilon_i
\end{aligned}$$

and

$$\begin{aligned}
Var(\hat{Y}) &= Var(\sum_{i=1}^n [\frac{1}{n} + (X - \bar{X}) k_i] \varepsilon_i) \\
&= \sum_{i=1}^n [\frac{1}{n} + (X - \bar{X}) k_i]^2 \sigma^2 \\
&= \sum_{i=1}^n [\frac{1}{n^2} + 2\frac{1}{n}(X - \bar{X}) k_i + (X - \bar{X})^2 k_i^2] \sigma^2 \\
&= [\frac{1}{n} + \frac{(X - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}] \sigma^2
\end{aligned}$$

]

Since σ^2 is unknown, we need to replace σ^2 by MSE and define

$$s^2(\hat{Y}) = MSE[\frac{1}{n} + \frac{(X-\bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}].$$

Sample distribution of $(\hat{Y} - E\hat{Y})/s(\hat{Y})$: For the simple linear regression model with independent identical normal error assumptions (i.e. model (1) in part 2 of Chapter 1), we have

$$\frac{\hat{Y} - EY}{s(\hat{Y})} \sim t(n-2)$$

Confidence interval for $E(Y)$ with confidence $1-\alpha$ (also called narrow intervals), we have

$$\hat{Y} \pm t(1-\alpha/2, n-2)s(\hat{Y})$$

Prediction interval for Y **with new** X **(called wide intervals)**: Although we can not predict the value of Y, but we can find its possible range. Consider the distribution of

$$\hat{Y} - Y$$

Write

$$\hat{Y} - Y = \hat{Y} - (\beta_0 + \beta_1 X + \varepsilon) = [\hat{Y} - EY] - \varepsilon$$

It is easy to see it follows normal distribution. Its **mean** is

$$E(\hat{Y} - Y) = [E\hat{Y} - EY] - E\varepsilon = 0$$

Note that Y is a new sample

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

with ε being uncorrelated with $\varepsilon_1, ..., \varepsilon_n$ and $\mathbf{Var}(\varepsilon) = \sigma^2$. The **variance** is

$$\begin{aligned} \mathbf{Var}(\hat{Y} - Y) &= \mathbf{Var}(\hat{Y} - EY - \varepsilon) = \mathbf{Var}(\hat{Y} - EY) + \mathbf{Var}(\varepsilon) = \mathbf{Var}(\hat{Y}) + \sigma^2 \\ &= [\frac{1}{n} + \frac{(X-\bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}]\sigma^2 + \sigma^2 \end{aligned}$$

Thus

$$\hat{Y} - Y \sim N(0, [1 + \frac{1}{n} + \frac{(X - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}]\sigma^2)$$

Again, we need to replace σ by

$$s^2(pred) = MSE\Big[1 + \frac{1}{n} + \frac{(X - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\Big]$$

And

$$\frac{\hat{Y} - Y}{s(pred)} \sim t(n-2) \tag{1}$$

With confidence $1 - \alpha$, the confidence interval is

$$\hat{Y} \pm t(1 - \alpha/2, n-2) * s(pred).$$

```
R code

predict(object, newdata, interval = "none"/"confidence"/"prediction",
                                              select one
level = 0.95)
```

Pointwise confidence bands

Letting X go through a range, then the confidence interval

$$\hat{Y} \pm t(1 - \alpha/2, n-2)s(\hat{Y}) \quad \text{(narrow confidence band)}$$

or

$$\hat{Y} \pm t(1 - \alpha/2, n-2) * s(pred) \quad \text{(wide confidence band)}$$

forms a band, called confidence band.

Example 2.1 For the example 1.1, with confidence level 0.95

(a) For a new point $X = 4$, predict the expected work hours $E(Y)$ and confidence interval

(b) For a new point $X = 4$, calculate the confidence interval for Y.

(c) Draw the (narrow) pointwise confidence band

Solution: (See **(R code)**)

(a) 76.64667; [25.14723, 128.1461].

(b) [-36.72411, 190.0174].

(c) see Figure 1

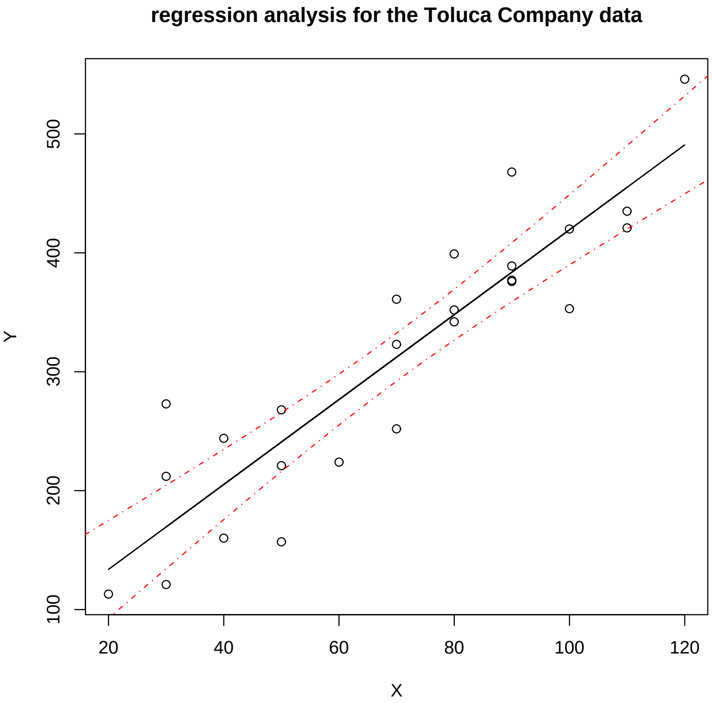

Figure 1: The 95% narrow pointwise confidence band for EY

Global confidence band [not our interest]