Alternative Versions of Regression Model

The simple linear regression model is

$$Y_i = \beta_0 X_0 + \beta_1 X_i + \varepsilon_i \qquad \text{where } X_0 \equiv 1$$

Alternative Versions of Regression Model is

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 (X_i - \bar{X}) + \beta_1 \bar{X} + \varepsilon_i \\ &= (\beta_0 + \beta_1 \bar{X}) + \beta_1 (X_i - \bar{X}) + \varepsilon_i \\ &= \beta_0^* + \beta_1 (X_i - \bar{X}) + \varepsilon_i \end{aligned}$$

Estimation of β_0 *and* β_1

The least square method for estimating the unknown parameters of the simple linear regression model can be explained as follows:

To find "good" estimators of the regression parameters β_0 and β_1, we employ the method of least squares. For the observations (X_i, Y_i) for each case, the method of least squares considers the deviation of Y_i from its expected value:

$$Y_i - (\beta_0 + \beta_1 X_i)$$

In particular, the method of least squares requires that we consider the sum of the n squared deviations. This criterion is denoted by Q:

$$Q = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 \quad , \qquad \textbf{(1.1)}$$

Differenatining (1.1) with respect to β_0 *and* β_1 **and equating to zero, we get**

$$\frac{\partial Q}{\partial \beta_0} = -2 \sum (Y_i - \beta_0 - \beta_1 X_i)$$

$$\frac{\partial Q}{\partial \beta_1} = -2 \sum X_i (Y_i - \beta_0 - \beta_1 X_i)$$

Hence

$$-2 \sum (Y_i - b_0 - b_1 X_i) = 0$$

$$-2 \sum X_i (Y_i - b_0 - b_1 X_i) = 0$$

$$\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i) = 0$$

$$\sum_{i=1}^{n} X_i(Y_i - b_0 - b_1 X_i) = 0$$

$$\sum Y_i - nb_0 - b_1 \sum X_i = 0$$

$$\sum X_i Y_i - b_0 \sum X_i - b_1 \sum X_i^2 = 0$$

Solving with respect b0 and b1, we get

$$b_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$

$$b_0 = \frac{1}{n}\left(\sum Y_i - b_1 \sum X_i\right) = \bar{Y} - b_1 \bar{X}$$

Which can be written as

$$\beta_1 = b_1 = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - \left(\sum x\right)^2}$$

$$\beta_0 = b_0 = \bar{Y} - b_1\bar{X}$$

Example 1 (Toluca Company Data)

The Toluca Company manufactures refrigeration equipment 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 labor 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 is given in the book data website:
http://www.stat.ufl.edu/~rrandles/sta4210/Rclassnotes/data/textdatasets/KutnerData/Chapter%20%201%20Data%20Sets/CH01TA01.txt

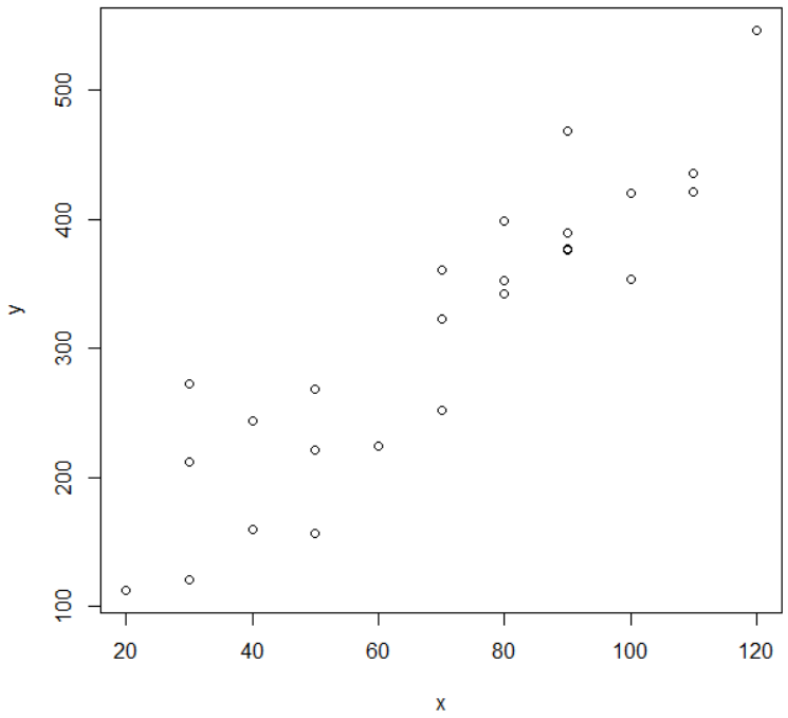
(1) Lot Size X_i	(2) Work Hours Y_i
80	399
30	121
50	221
...	...
40	244
80	342
70	323
1,750	7,807

Find the estimation of the simple linear regression model

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

Solution

The scatter plot for the data shows that the simple linear model represents a good fit for the data as follows:



Then it is easy to calculate

$$\sum (X_i - \bar{X})(Y_i - \bar{Y}) = 70{,}690$$

$$\sum (X_i - \bar{X})^2 = 19{,}800$$

$$\bar{X} = 70.0$$

$$\bar{Y} = 312.28$$

Then

$$b_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} = \frac{70{,}690}{19{,}800} = 3.5702$$

$$b_0 = \bar{Y} - b_1\bar{X} = 312.28 - 3.5702(70.0) = 62.37$$

One can use R as follows:

```
mydata = read.table("TCD.txt", header=TRUE)
# How to separate some variables the data file
x=mydata$X
y=mydata$Y
x.bar=mean(x)
y.bar=mean(y)
print(c(x.bar,y.bar))
t1=sum((x-x.bar)*(y-y.bar))
t2=sum((x-x.bar)^2)
b1=t1/t2
```

Or one can use the direct R command for regression as: lm(y~x) to get the same results.

```
Call:
lm(formula = y ~ x)
Coefficients:
(Intercept)          x
     62.37        3.57
```

As we can see from the results the estimated linear regression model is

Y = 62.4 + 3.57 X (2)

- **Interpretation of the results**

1- **When the lot size (X) increases by one units, the work hours (Y) increase by 3.57 hours.**
2- **There is 62.4 hour of the work hours (Y) do not depend on the lot size (X).**

- **The estimated simple linear regression model in equation (2) can be used to predict the work hours required for a certain lot size. For example, if the lot size is 85 units, then**

Y = 62.4 + 3.57 * 85 = 365.85 hours.

- **We can use the alternative Model as:**

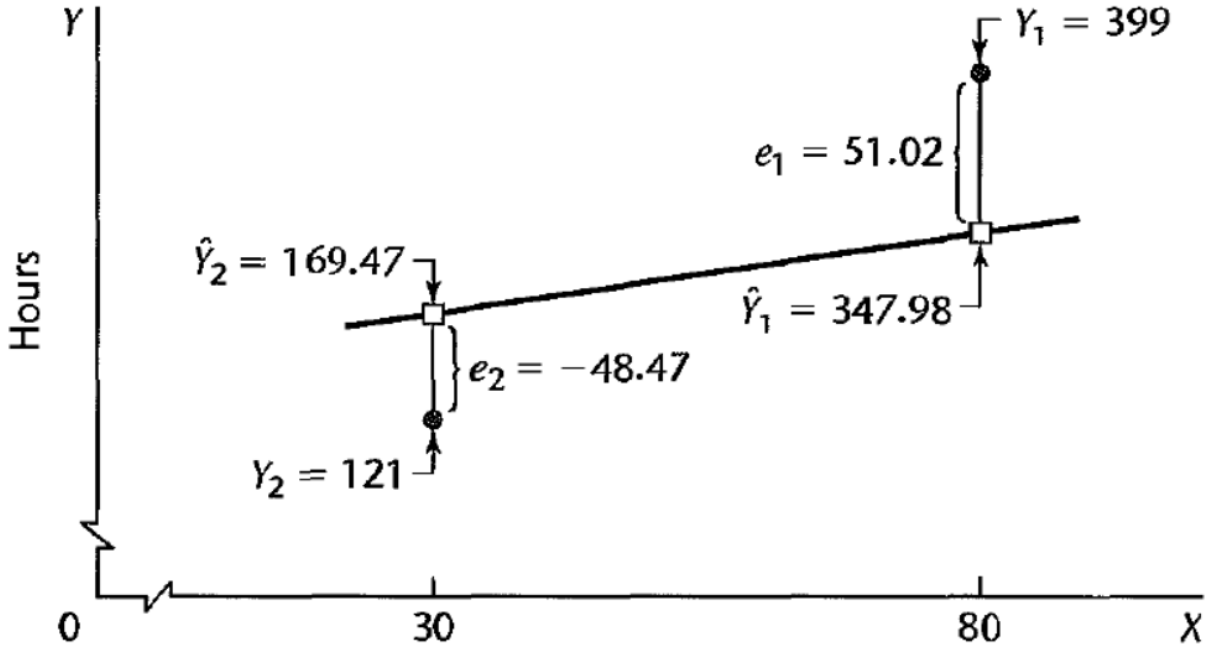
$$\hat{Y} = 312.28 + 3.5702(X - 70.0)$$

- **The residual can be calculated at each point of the intendent variable x as**

This residual is denoted by e_i and is defined in general as follows:

$$e_i = Y_i - \hat{Y}_i$$

For example, when X=30 and X=80, we calculate the residuals as:



From the Figure, we see that

$$e_i = Y_i - (b_0 + b_1 X_i) = Y_i - b_0 - b_1 X_i$$

$$e_1 = Y_1 - \hat{Y}_1 = 399 - 347.98 = 51.02$$

Similarly, we can calculate the residuals at the all point of x to get

i	*e*	*i*	*e*	*i*	*e*
1	51.02	11	-45.17	21	103.53
2	-48.47	12	-60.28	22	84.32
3	-19.88	13	5.32	23	38.83
4	-7.68	14	-20.77	24	-5.98
5	48.72	15	-20.09	25	10.72
6	-52.58	16	0.61		
7	55.21	17	42.53		
8	4.02	18	27.12		
9	-66.39	19	-6.68		
10	-83.88	20	-34.09		

These residuals can be calculated directly from R providing as:

summary(model)$res