Tutorial #1

(1.1) $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$

Question 1:

Grade point average. The director of admissions of a small college selected 120 students at random from the new freshman class in a study to determine whether a student's grade point average (GPA) at the end of the freshman year (Y) can be predicted from the ACT test score *(X)*. The results of the study follow. Assume that first-order regression model (1.1) is appropriate.

- **a. Obtain the least squares estimates of $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$, and state the estimated regression function.**
- **b. Plot the estimated regression function and the data. "Does the estimated regression function appear to fit the data well?**
- **c. Obtain a point estimate of the mean freshman GPA for students with ACT test score *X* = 30.**
- **d. What is the point estimate of the change in the mean response when the entrance test score increases by one point?**

Solution :

a. $\bar{X} = 24.725, \bar{Y} = 3.07405$

$$\sum_{i=1}^{n=120} (X_i - \bar{X})\,(Y_i - \bar{Y}) = 92.40565$$

$$\sum_{i=1}^{n=120} (X_i - \bar{X})^2 = 2379.925$$

$$\sum_{i=1}^{n=120} (Y_i - \bar{Y})^2 = 49.40545$$

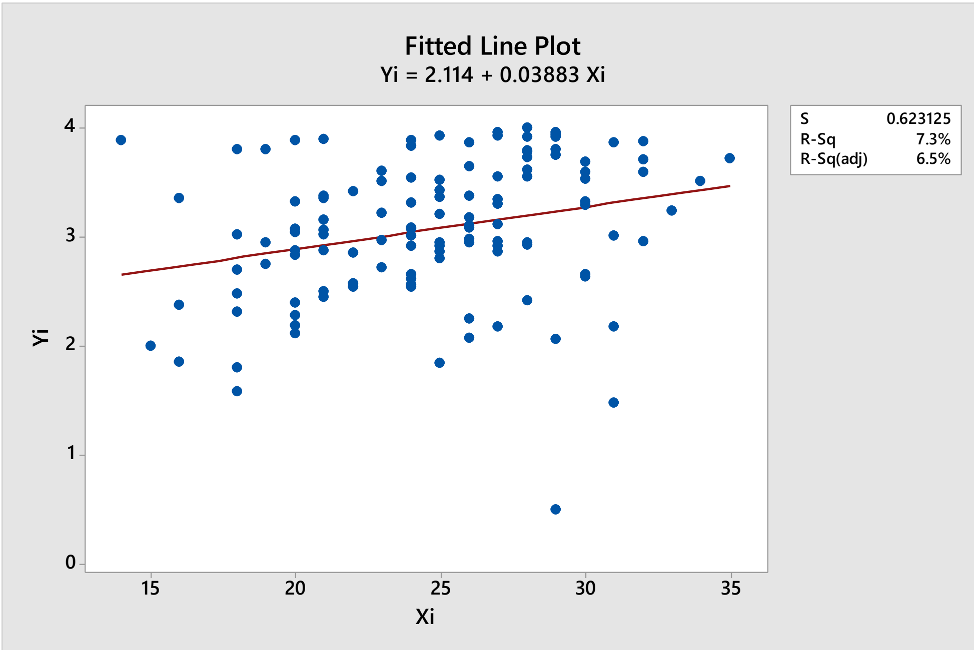
$$b_1 = \widehat{\beta_1} = \frac{\sum_{i=1}^{n=120}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n=120}(X_i - \bar{X})^2} = \frac{92.40565}{2379.925} = 0.038827$$

$$b_0 = \widehat{\beta_0} = \bar{Y} - b_1\bar{X} = 3.07405 - 0.038827 * 24.725 = 2.114049$$

$$\hat{Y} = 2.114 + 0.0388\,X$$

b. plot the estimated regression function and the data.

To plot in Minitab:* *Stat >Regression >Fitted line plot



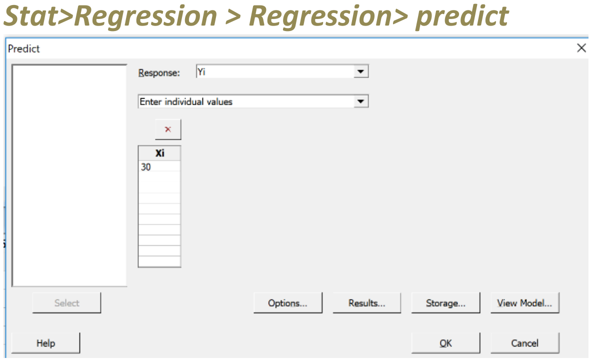
According to the plot, the estimated regression function is the line that matches the closest the given data. But, data is too spread out, the model is going to have problems giving accurate predictions because there is a lot of variance in the data. Therefore, the estimated regression function doesn't fit the data very well.

c. point estimate of $\overline{Y}$ att X=30

$$\widehat{Y_h} = 2.114 + 0.0388\ (30) = 3.278863$$

To find value of predictors for new observations in MINITAB 17

Stat>Regression > Regression> predict



d. when the entrance test score (ACT) increases by one point, the mean response (GPA) increase by 0.038827.

By use MINITAB 17 program.

Stat>Regression > Regression>Fit Regression Model
Responses: Y
continues predictors : X

Regression Analysis: Yi versus Xi

```
Analysis of Variance
Source          DF   Adj SS   Adj MS  F-Value  P-Value
Regression       1    3.588   3.5878     9.24    0.003
  Xi             1    3.588   3.5878     9.24    0.003
Error          118   45.818   0.3883
  Lack-of-Fit   19    6.486   0.3414     0.86    0.632
  Pure Error    99   39.332   0.3973
Total          119   49.405


Model Summary
       S    R-sq  R-sq(adj)  R-sq(pred)
0.623125   7.26%      6.48%       3.63%

Coefficients
Term          Coef   SE Coef  T-Value  P-Value   VIF
Constant     2.114     0.321     6.59    0.000
Xi          0.0388    0.0128     3.04    0.003  1.00

Regression Equation
Yi = 2.114 + 0.0388 Xi
```

Question 2:

Copier maintenance. The Tri-City Office Equipment Corporation sells an imported copier on a franchise basis and performs preventive maintenance and repair service on this copier. The data below have been collected from 45 recent calls on users to perform routine preventive maintenance service; for each call, X is the number of copiers serviced" عدد الناسخات التي تمت خدمتها" and (*Y*) is the total number of minutes spent by the service person" العدد الإجمالي للدقائق التي يقضيها الشخص للخدمة". Assume that first-order regression model (1.1) is appropriate.

a. Obtain the estimated regression function.
b. Plot the estimated regression function and the data. How well does the estimated regression function fit the data?
c. Interpret $b_0$in your estimated regression function. Does $b_0$provide any relevant information here? Explain.
d. Obtain a point estimate of the mean service time when X = 5 copiers are serviced.
e. Obtain the residuals e_i and the sum of the squared residuals $\sum e_i^2$. What is the relation between the sum of the squared residuals here and the quantity $Q = \sum(Y_i - b_0 - Xb_1)^2$?
f. Obtain point estimates of σ^2 and. In what units is σ expressed?

Solution :

a. $\bar{X} = 5.11111, \bar{Y} = 76.26667$

$$\sum_{i=1}^{n=45} (X_i - \bar{X})\,(Y_i - \bar{Y}) = 5118.667$$

$$\sum_{i=1}^{n=45} (X_i - \bar{X})^2 = 340.4444$$

$$\sum_{i=1}^{n=45} (Y_i - \bar{Y})^2 = 80376.8$$

Note: to calculate x_bar and y_bar in Minitab

Editor >> Enable commands

```
MTB > let k1=mean(c2)
MTB > let k2=mean(c3)
MTB > print k1 k2
```

Data Display

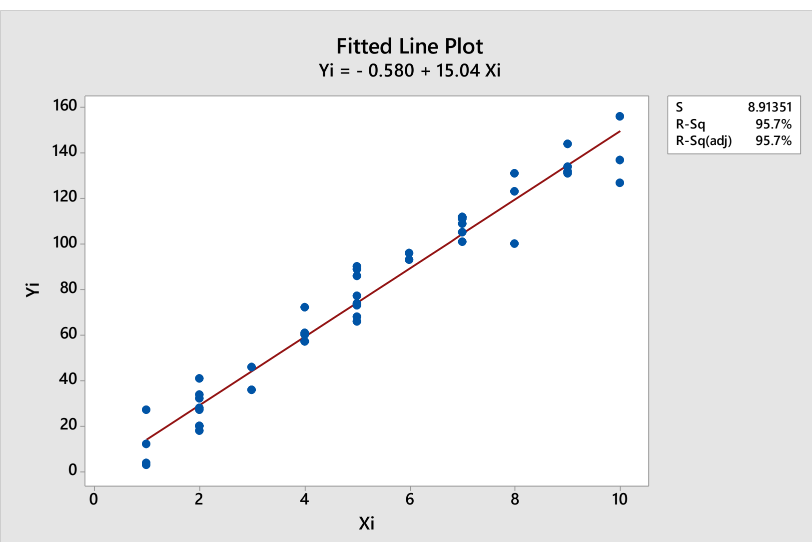
```
K1     5.11111
K2     76.2667
```

$$b_1 = \widehat{\beta_1} = \frac{\sum_{i=1}^{n=45}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n=45}(X_i - \bar{X})^2} = 15.03525$$

$$b_0 = \widehat{\beta_0} = \bar{Y} - b_1\bar{X} = -0.58016$$

$$\hat{Y} = -0.58016 + 15.03525\ X$$

b. *scatter plot and line of statistical relationship*




```
The regression equation is
Yi = - 0.580 + 15.04 Xi

S = 8.91351   R-Sq = 95.7%   R-Sq(adj) = 95.7%

Analysis of Variance
Source       DF       SS       MS       F      P
Regression    1  76960.4  76960.4  968.66  0.000
Error        43   3416.4     79.5
Total        44  80376.8
```

According to the plot, the estimated regression function matches very well the data. Almost all levels of X show the same spread and the line touches at least one point(which in most cases is close to the mean of Y) in each level of X.

c. b_0 gives the mean of the probability distribution of Y only at X=0 .Thus, in this case doesn't give any information.

d. a point estimate of the mean service time when X=5 is

$\widehat{Y_h} = -0.58016 + 15.03525\ (5) = 74.59608$ **minutes**

#To find value of predictors for new observations in MINITAB
Stat>Regression > Regression> predict
Response: Yi
Enter individual values
Xi
5

>>ok

Prediction for Yi

```
Regression Equation
Yi = -0.58 + 15.035 Xi

Variable  Setting
Xi              5

    Fit    SE Fit         95% CI                95% PI
74.5961  1.32983  (71.9142, 77.2779)  (56.4213, 92.7708)
```

e. Obtain the residuals e_i and the sum of the squared residuals $\sum e_i^2$. What is the relation between the sum of the squared residuals here and the quantity $Q = \sum(Y_i - b_0 - Xb_1)^2$? **(Given next lecture)**

$$SSE = \sum e_i^2 = \sum_{i=1}^{n=45} (y_i - \widehat{y_i})^2 = 3416.377$$

$$\sum e_i^2 = Q$$

f. Obtain point estimates of $\sigma^2 = Var(\varepsilon_i)$ and. In what units is σ expressed?
(Given next lecture)

$$\widehat{\sigma^2} = MSE = \frac{\sum e_i^2}{n-2} = \frac{3416.377}{43} = 79.45063 \text{ Minutes}^2$$

$$\hat{\sigma} = \sqrt{MSE} = \sqrt{79.45063} \text{ Minutes}$$

```
The regression equation is
Yi = - 0.580 + 15.04 Xi

S = 8.91351   R-Sq = 95.7%   R-Sq(adj) = 95.7%

Analysis of Variance
Source      DF       SS       MS       F      P
Regression   1  76960.4  76960.4  968.66  0.000
Error       43   3416.4     79.5
Total       44  80376.8
```

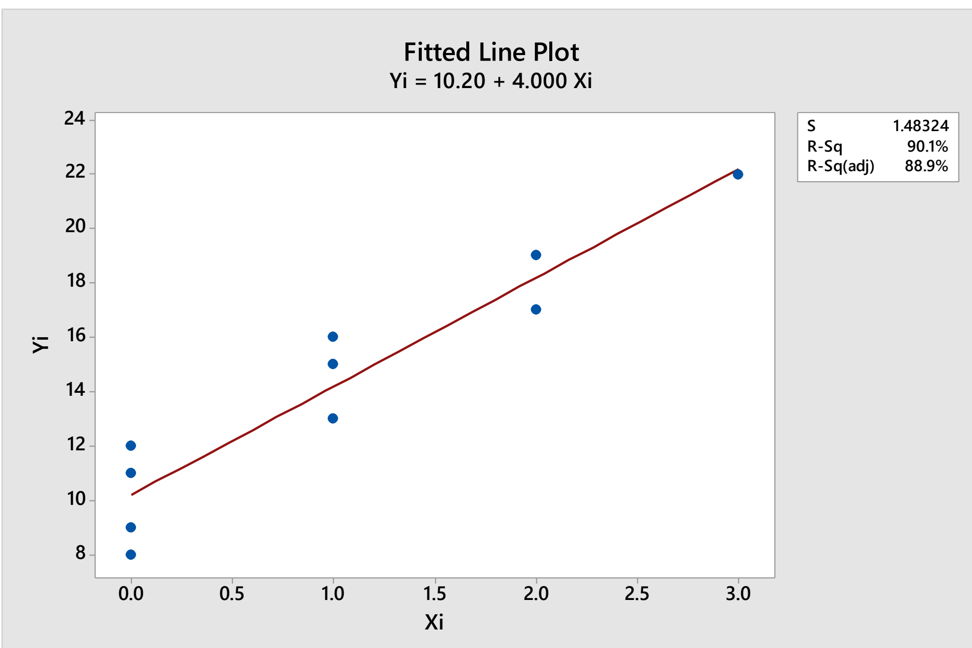
Question 3:

Airfreight breakage. A substance used in biological and medical research is shipped by airfreight to users in cartons of 1,000 ampules. The data below, involving 10 shipments, were collected on the number of times the carton was transferred from one aircraft to another over the shipment route (X) and the number of ampules found to be broken upon arrival (Y). Assume that first-order regression model (1.1) is appropriate.

a. **Plot the estimated regression function and the data. Does a linear regression function appear to give a good fit here?**
b. **Compute $\sum_{i=1}^{n} x_{ij}$, $\sum_{i=1}^{n} y_i$, $\sum_{i=1}^{n} x_i^2$, $\sum_{i=1}^{n} y_i^2$ *and* $\sum_{i=1}^{n} x_i y_i$. Use these to fit the data with the simple linear regression model $y = \beta_0 + \beta_1 x + \varepsilon$?**
c. **Obtain a point estimate of the expected number of broken ampules when X = 1 transfer is made.**
d. **Estimate the increase in the expected number of ampules broken when there are 2 transfers as compared to 1 transfer.**
e. **Verify that your fitted regression line goes through the point $(\bar{X}, \bar{Y})$.**
f. Obtain the residual for the first case. What is its relation to e_1?
g. Compute $\sum e_i^2$ and *MSE*. What is estimated by *MSE?*

Solution :

a.



We note that most of the points fall around the line of statistical relationship.

In general, the simple linear model is good to fit the data.

b. we have tow way to estimated regression function

First way

i	X_i	Y_i	$x_i y_i$	y_i^2	x_i^2
1	1	16	16	256	1
2	0	9	0	81	0
3	2	17	34	289	4
4	0	12	0	144	0
5	3	22	66	484	9
6	1	13	13	169	1
7	0	8	0	64	0
8	1	15	15	225	1
9	2	19	38	361	4
10	0	11	0	121	0
sum	10	142	182	2194	20

$$b_1 = \widehat{\beta_1} = \frac{\sum_{i=1}^{n=10}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n=10}(X_i - \bar{X})^2} = \frac{\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i / n}{\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2 / n} = \frac{\sum_{i=1}^{n} x_i y_i - n\,\bar{x}\,\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

$$b_1 = \frac{(182 - \frac{10 * 142}{10}}{(20 - (10^2/10\,)} = \frac{40}{10} = 4$$

$$b_0 = \bar{y} - b_1 \bar{x} = \frac{\sum_{i=1}^{n} y_i}{n} - b_1\left(\frac{\sum_{i=1}^{n} x_i}{n}\right) = \frac{142}{10} - 4\left(\frac{10}{10}\right) = 10.2$$

$$\boldsymbol{\hat{Y} = 10.2 + 4\,X}$$

Second way

i	X_i	Y_i	$(X_i - \bar{X})$	$(Y_i - \bar{Y})$	$(X_i - \bar{X})\,(Y_i - \bar{Y})$	$(X_i - \bar{X})^2$	$(Y_i - \bar{Y})^2$	$\hat{Y}_i$	$Y_i - \hat{Y}_i$
1	1	16	0	1.8	0	0	3.24	14.2	1.8
2	0	9	-1	-5.2	5.2	1	27.04	10.2	-1.2
3	2	17	1	2.8	2.8	1	7.84	18.2	-1.2
4	0	12	-1	-2.2	2.2	1	4.84	10.2	1.8
5	3	22	2	7.8	15.6	4	60.84	22.2	-0.2
6	1	13	0	-1.2	0	0	1.44	14.2	-1.2

7	0	8	-1	-6.2	6.2	1	38.44	10.2	-2.2
8	1	15	0	0.8	0	0	0.64	14.2	0.8
9	2	19	1	4.8	4.8	1	23.04	18.2	0.8
10	0	11	-1	-3.2	3.2	1	10.24	10.2	0.8
	mean	mean	sum	sum	sum	sum	sum		
	1	14.20	0	0	40	10	177.60		

$\bar{X} = 1, \bar{Y} = 14.2$

$\sum_{i=1}^{n=10}(X_i - \bar{X})(Y_i - \bar{Y}) = 40, \quad \sum_{i=1}^{n=10}(X_i - \bar{X})^2 = 10, \quad \sum_{i=1}^{n=10}(Y_i - \bar{Y})^2 = 177.6$

$$b_1 = \widehat{\beta_1} = \frac{\sum_{i=1}^{n=10}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n=10}(X_i - \bar{X})^2} = 4$$

$$b_0 = \widehat{\beta_0} = \bar{Y} - b_1\bar{X} = 10.2$$

$$\boldsymbol{\hat{Y} = 10.2 + 4\,X}$$

$\boldsymbol{b_0}$ it is equal to 10.2 this the intercept of **Y** axis, this value does not dependent on the number of times transferred.

$\boldsymbol{b_1}$ it is equal to 4 this is the slope of the regression, this value dependent on the number of times transferred. This means that if we increase the number of times transferred by one unit, then we should be the number of broken ampules increase approximation 4.

c. At X=1

$\widehat{Y_h} = 10.2 + 4(1) = 14.2$

d. *when the transfer increase to 2 then the increase in the expected number of ampules broken well increase by 8 to be:*

$$\widehat{Y_h} = 10.2 + 4(2) = 18.2$$

e. $\bar{X} = 1, \bar{Y} = 14.2$

$(\bar{X}, \bar{Y}) = (1, 14.2)$

If $\bar{X} = 1$, Then $\hat{Y}_{x=\bar{X}} = 10.2 + 4.0(1) = 14.20$

Therefore, we can say the regression line goes through the point $(\bar{X}, \bar{Y}) = (1, 14.2)$

Also,

$\hat{y}_i = b_0 + b_1 x_i$

If $x_i = \bar{x}$

$\hat{y}_i = b_0 + b_1\bar{x}$

$$b_0 = \bar{y} - b_1\bar{x}$$
$$\hat{y}_\imath = \bar{y} - b_1\bar{x} + b_1\bar{x} = \bar{y}$$
$$\hat{y}_\imath = \bar{y}$$

f) Obtain the residual for the first case. What is its relation to e_1?

$$\boldsymbol{e_1} = y_1 - \hat{y}_\imath = \mathbf{16 - 14.2 = 1.8}$$

g) Compute $\sum e_i^2$ and *MSE*. What is estimated by *MSE?*

$$SSE = \sum e_i^2 = \sum_{i=1}^{n=10} (y_i - \hat{y}_\imath)^2 =$$

$$\widehat{\sigma^2} = MSE = \frac{\text{SSE}}{\text{n} - 2} = \frac{\sum e_i^2}{n - 2}$$

Question 4: H.W

Plastic hardness. Refer to Problems 1.3 and 1.14. Sixteen batches of the plastic were made, and from each batch one test item was molded. Each test item was randomly assigned to one of the four predetermined time levels, and the hardness was measured after the assigned elapsed time. The results are shown below; X is the elapsed time in hours? And *Y* is hardness in Brinell units. Assume that first-order regression model (1.1) is appropriate.

a. **Obtain the estimated regression function. Plot the estimated regression function and the data. Does a linear regression function appear to give a good fit here?**
b. **Obtain a point estimate of the mean hardness when X = 40 hours.**
c. **Obtain a point estimate of the change in mean hardness when X increases by 1 hour**
d. **Obtain the residuals e_j. Do they sum to zero in accord with $\sum e_i = 0 \quad (1.17)$?**
e. **Estimate σ^2. In what units is σ expressed?**

<u>Question 5:</u>

Suppose that you are given observations y_1 and y_2 such that : $y_1 = \propto + \beta + \epsilon_1$,

$y_2 = -\propto + \beta + \epsilon_2$ The random variables ϵ_i for i = 1,2 are independent and normally distributed with mean 0 and variance σ^2 ,

a. **find the least squares estimators of the parameters $\propto$ *and* β , also verify that : they are unbiased estimators .(Hint : obtain the minimum of the sum of the ϵ_i^2 using the least squares technique.)**
b. **find variance of $\widehat{\propto}$.**

Solution :

Q = $\sum_{i=1}^{2} \epsilon_i^2 = \epsilon_1^2 + \epsilon_2^2 = (y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 = (y_1 - \propto - \beta)^2 + (y_2 + \propto - \beta)^2$

Then ,

$$\frac{\partial Q}{\partial \propto} = -2(y_1 - \propto - \beta) + 2(y_2 + \propto - \beta) \rightarrow \frac{\partial Q}{\partial \propto} = 0 \rightarrow \widehat{\propto} = \frac{y_1 - y_2}{2}$$

$$\frac{\partial Q}{\partial \beta} = -2(y_1 - \propto - \beta) - 2(y_2 + \propto - \beta) \rightarrow \frac{\partial Q}{\partial \beta} = 0 \rightarrow \widehat{\beta} = \frac{y_1 + y_2}{2}$$

To verify that , $\widehat{\propto}$ and $\widehat{\beta}$ are unbiased estimators :

$$E(\widehat{\propto}) = E\left(\frac{y_1 - y_2}{2}\right) = \frac{1}{2}[E(y_1) - E(y_2)] = \frac{1}{2}(\propto + \beta + \propto - \beta) = \propto$$

$$E(\widehat{\beta}) = E\left(\frac{y_1 + y_2}{2}\right) = \beta$$

***so*, $\widehat{\propto}$ is an unbiased estimator of $\propto$ and $\widehat{\beta}$ is an unbiased estimator of β .**

b. The variance of $\widehat{\propto}$ is calculated by taking into account that ϵ_1 *and* ϵ_2 are independent and normally distributed with a common variance of σ^2.

$$V(\widehat{\propto}) = \frac{1}{4}[var(y_1) + var(y_2)] = \frac{1}{4}[var(\varepsilon_1) + var(\varepsilon_2)] = \frac{1}{4}[2\boldsymbol{\sigma}^2] = \frac{\boldsymbol{\sigma}^2}{2}$$

Question 6:

An investigation , conducted by a mail-order company, into the relation between the sales revenues (y_i , $in\ millions\ of\ dollars$) and the price per gallon of gasoline (x_i , in cents) over a period of 10 months yields :

$\sum_{i=1}^{10} y_i = 527$, $\sum_{i=1}^{10} x_i = 6509$, $\sum_{i=1}^{10} x_i^2 = 4909311$, $\sum_{i=1}^{10} x_i\, y_i = 325243$.

Estimate the parameters β_0 and β_1 in the regression model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, where the ϵ_i are uncorrelated with a mean of zero and a common variance of σ^2 for i=1,2,…,10 .

Solution :

$\bar{y} = 52.7\ \bar{x} = 650.9$

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} x_i y_i - n\,\bar{x}\,\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = -0.0264$$

$$b_0 = \bar{y} - b_1\bar{x} = 69.9076$$

$$\hat{y} = 69.9076 - 0.0264\,x$$

Question 7:

Let X and ϵ be two independent random variables , and E(ϵ) = 0 . Let Y = $\beta_0 + \beta_1 X + \epsilon$.

Show that : $\beta_1 = \frac{cov(X,Y)}{V(X)} = Corr(X,Y)\sqrt{\frac{V(X)}{V(Y)}}$

Solution :

$$Y = \beta_0 + \beta_1 X + \epsilon \quad \rightarrow \quad E(Y) = \beta_0 + \beta_1 E(X) \quad \rightarrow \quad Y - E(Y) = \big(X - E(X)\big)\beta_1 + \epsilon$$

Hence :

$$COV(X,Y) = E[(X-E(X))(Y-E(Y))]$$

$$COV(X,Y) = E[(X-E(X))(X-E(X))\beta_1)]$$

$$COV(X,Y) = \beta_1\ E\,[X - E(X)]^2$$

$$= \beta_1\, V(X)$$

SO , $\beta_1 = \frac{COV(X,Y)}{V(X)}$

Homework :

Prove that

1. $\boldsymbol{\beta_1 = Corr(X,Y)\sqrt{\frac{V(Y)}{V(X)}}}$

2. $\boldsymbol{\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i}$

3. $\boldsymbol{\sum_{i=1}^{n} x_i e_i = 0}$

4. $\boldsymbol{\sum_{i=1}^{n} \hat{y}_i e_i = 0}$