Simple Linear Regression

<u>Example(1)</u>: (Yusra Elkamali)

Use the following variable (X) is the number of cubic feet moved as Independent variable to predict the total number of labor hours(Y) by answering the following questions:

Feet(X)	Hours(Y)	X^2	Y^2	XY
545	24	297025	576	13080
400	13.5	160000	182.25	5400
562	26.25	315844	689.0625	14752.5
540	25	291600	625	13500
220	9	48400	81	1980
344	20	118336	400	6880
569	22	323761	484	12518
340	11.25	115600	126.5625	3825
900	50	810000	2500	45000
285	12	81225	144	3420
865	38.75	748225	1501.5625	33518.75
831	40	690561	1600	33240
344	19.5	118336	380.25	6708
360	18	129600	324	6480
750	28	562500	784	21000
650	27	422500	729	17550
415	21	172225	441	8715
275	15	75625	225	4125
557	25	310249	625	13925
1028	45	1056784	2025	46260
793	29	628849	841	22997
523	21	273529	441	10983
564	22	318096	484	12408
312	16.5	97344	272.25	5148
757	37	573049	1369	28009
600	32	360000	1024	19200
796	34	633616	1156	27064
577	25	332929	625	14425
500	31	250000	961	15500
695	24	483025	576	16680
1054	40	1110916	1600	42160
486	27	236196	729	13122
442	18	195364	324	7956
1249	62.5	1560001	3906.25	78062.5
995	53.75	990025	2889.0625	53481.25
1397	79.5	1951609	6320.25	111061.5
Totals: 22520	1042.5	16842944	37960.5	790134.5

Find:

1) The regression coefficients b_0, b_1,

2) Interpret the slope of this problem.

3) Write the estimated regression equation (prediction line).

4) Predict the mean labor hours for moving 500 cubic feet

5) Sum of square of regression (SSR) and Error sum of square (SSE).

6) Coefficient of Determination (R^2) and Interpret it.

7) Standard error of the estimate:

8) Test the hypotheses that there is NO correlation between X and Y.

9) Test if there is evidence of linear relationship between X and Y.

10) Construct ANOVA table for regression to test that there is No relationship between X and Y by using F-ratio.

Solution:

1) The regression coefficients:

a) $\mathbf{b_1 = \frac{SSXY}{SSX}}$, $\mathbf{SSXY = \sum XY - \frac{(\sum X)(\sum Y)}{n} = 790134.5 - \frac{22520 * 1042.5}{36} =}$ $\mathbf{137992.8333}$

$$SSX = \sum X^2 - \frac{(\sum X)^2}{n} = 16842944 - \frac{(22520)^2}{36} = 2755432.889$$

$$b_1 = \frac{137992.8333}{2755432.889} = 0.05$$

$\bar{X}$ = 625.555556 , $\bar{Y}$ =28.95833333 , n=36

b) $\boldsymbol{b_0 = \bar{Y} - b_1 \bar{X}}$ = **28.95833333 – (0.05*625.555556) = -2.31944444**

b_0 = -2.319

2) **For every cubic foot increase in the amount moved, predicted mean labor hours are estimated to increase by 0.05 hours.**

---

3) **The prediction Line is : $\hat{Y}$= b_0+b_1X**

$\hat{Y}$= -2.319 + 0.05X

4)Predict the mean labor hours for moving 500 cubic feet

$\hat{Y}$ **= -2.319 + 0.05(500) =22.68**

---

5) **Sum of square of regression (SSR) and Error sum of square (SSE).**

SSR=

$$b_0\sum Y + b_1\sum XY - \frac{(\sum Y)^2}{n} = -2.319(1042.5) + 0.05(790134.5) - \frac{(1042.5)^2}{36} = 6900.11$$

SSE = $\sum Y^2 - b_0\sum Y - b_1\sum XY$ **= 37960.5 – (-2.319*1042.5)- 0.05(790134.5)=** 870.83

SST= SSR+SSE= 7770.94

---

6) Coefficient of Determination (R^2) and Interpret it.

$$R^2 = \frac{SSR}{SST} = \frac{6900.11}{7771.44} = 89.59\%$$

89.59% OF variation in Hours explained by variation of number of cubic feet moved.

--

7) Standard error of the estimate:

$$S_{yx} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{808.3687861}{36-2}} = 4.876018102$$

--

8) Test the hypotheses that there is NO correlation between X and Y.

$$r = \sqrt{R^2} = \sqrt{0.8959} = 0.97$$

$$r = \frac{\text{cov}(X,Y)}{S_x S_y} = \frac{\left(\sum (X-\bar{X})(Y-\bar{Y})\right)/n-1}{\sqrt{\frac{\sum (X-\bar{X})^2}{n-1}}\sqrt{\frac{\sum (Y-\bar{Y})^2}{n-1}}} = \frac{137992.8333/34}{(280.583)(14.9)} = 0.97$$

Using t Test for the Correlation Coefficient:

$H_0: \rho = 0$, $H_1: \rho \neq 0$, **t crit. = 2.0322,**

t stat. = $\frac{r-\rho}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{0.97-0}{\sqrt{\frac{1-0.97^2}{36-2}}} = 23.27$

t stat > t critical, then reject the null hypothesis test. There is high correlation between X and Y, at significant level 0.05.

---

9) Test if there is evidence of linear relationship between X and Y: $\sum(X - \overline{X})^2$ **= 2755432.889**

a. **Using t Test for the Slope:**

$H_0 : \beta_1 = 0$ **,** $H_1 : \beta_1 \neq 0$ **, t crit. = 2.0322,**

t stat. = $\frac{b_1 - \beta_1}{S_{b_1}} = \frac{0.05 - 0}{0.00303} = 16.5223$

$$S_{b_1} = \frac{S_{YX}}{\sqrt{SSX}} = \frac{S_{YX}}{\sqrt{\sum(X - \overline{X})^2}} = \frac{5.0314}{\sqrt{2755432.89}} = 0.00303$$

The decision is reject H_0 because t stat > t crit .

There is a relation between X and Y

---

10) ANOVA TABLE: used to test the significant of the Model.

ANOVA				
SS	*df*	*SS*	*MS*	*F*
Regression(SSR)	1	6900.11	6900.11	269.4
Error (SSE)	34	870.83	25.61	
Total	35	7770.94		

$H_0 : \beta_1 = 0$ **,** $H_1 : \beta_1 \neq 0$

F stat = 269.4 > F $_{1,34,0.05}$= 4.17

Reject H0 at 0.05, Model was high significant which means that there was strong linear relationship between X and Y

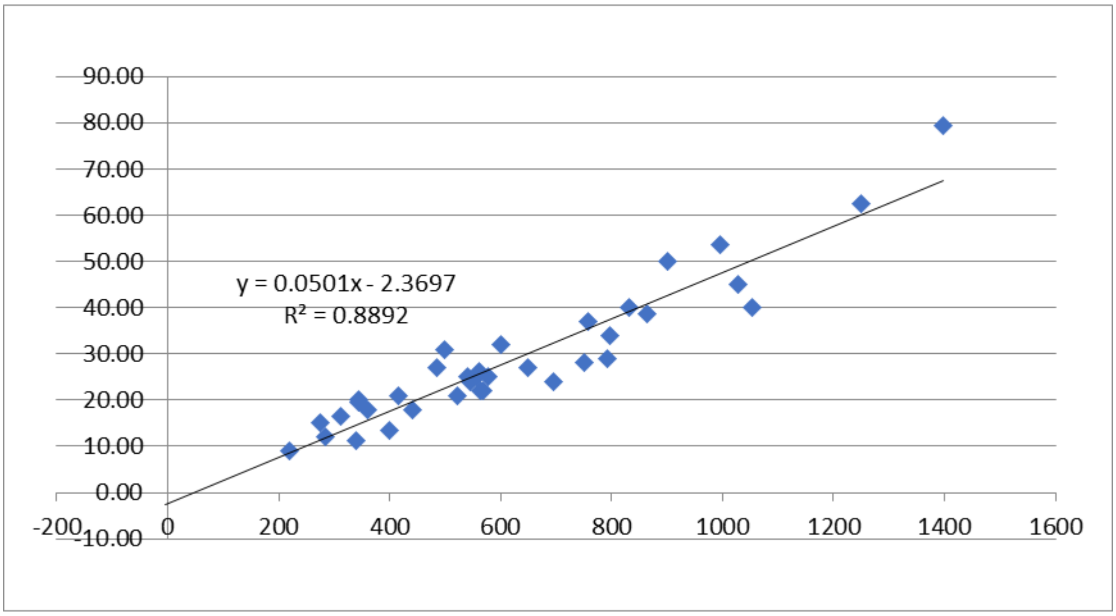
Other Solution: (Dr. Maher Badawi)

<u>Use this table for three following problems:</u>

Feet(X)	Hours(Y)	$(X-\bar{X})$	$(Y-\bar{Y})$	$(X-\bar{X})(Y-\bar{Y})$	$(X-\bar{X})^2$	$\hat{Y}$	$(\hat{Y}-\bar{Y})^2$	$(Y-\bar{Y})^2$	$(Y-\hat{Y})^2$
545	24.00	-80.556	-4.96	399.4212963	6489.197531	24.9	16.2751269	24.5851	0.85394
400	13.50	-225.56	-15.46	3486.712963	50875.30864	17.7	127.596995	238.96	17.326
562	26.25	-63.556	-2.71	172.1296296	4039.308642	25.8	10.1307227	7.33507	0.22519
540	25.00	-85.556	-3.96	338.6574074	7319.753086	24.7	18.3581883	15.6684	0.10648
220	9.00	-405.56	-19.96	8094.212963	164475.3086	8.65	412.509636	398.335	0.1239
344	20.00	-281.56	-8.96	2522.268519	79273.53086	14.9	198.820696	80.2517	26.4406
569	22.00	-56.556	-6.96	393.5324074	3198.530864	26.1	8.02202357	48.4184	17.024
340	11.25	-285.56	-17.71	5056.712963	81541.97531	14.7	204.510032	313.585	11.612
900	50.00	274.444	21.04	5774.768519	75319.75309	42.7	188.904488	442.752	53.2522
285	12.00	-340.56	-16.96	5775.25463	115978.0864	11.9	290.876963	287.585	0.00937
865	38.75	239.444	9.79	2344.560185	57333.64198	40.9	143.794713	95.8767	4.83902
831	40.00	205.444	11.04	2268.449074	42207.41975	39.2	105.857636	121.918	0.56694
344	19.50	-281.56	-9.46	2663.046296	79273.53086	14.9	198.820696	89.4601	21.5486
360	18.00	-265.56	-10.96	2910.046296	70519.75309	15.7	176.865926	120.085	5.47916
750	28.00	124.444	-0.96	-119.2592593	15486.41975	35.2	38.8404645	0.9184	51.7039
650	27.00	24.4444	-1.96	-47.87037037	597.5308642	30.2	1.49862761	3.83507	10.1284
415	21.00	-210.56	-7.96	1675.671296	44333.64198	18.4	111.190273	63.3351	6.68919
275	15.00	-350.56	-13.96	4893.171296	122889.1975	11.4	308.210263	194.835	12.9426
557	25.00	-68.556	-3.96	271.3657407	4699.864198	25.5	11.7874183	15.6684	0.27568
1028	45.00	402.444	16.04	6455.87963	161961.5309	49.1	406.204997	257.335	16.9156
793	29.00	167.444	0.04	6.976851852	28037.64198	37.3	70.3193542	0.00174	69.6223
523	21.00	-102.56	-7.96	816.1712963	10517.64198	23.8	26.3786018	63.3351	7.96551
564	22.00	-61.556	-6.96	428.3240741	3789.08642	25.9	9.50315688	48.4184	15.0204
312	16.50	-313.56	-12.46	3906.37963	98317.08642	13.3	246.582578	155.21	10.5275
757	37.00	131.444	8.04	1057.032407	17277.64198	35.5	43.3329104	64.6684	2.12837
600	32.00	-25.556	3.04	-77.73148148	653.0864198	27.7	1.63796283	9.25174	18.6753
796	34.00	170.444	5.04	859.3240741	29051.30864	37.5	72.8616645	25.4184	12.2097
577	25.00	-48.556	-3.96	192.1990741	2357.641975	26.5	5.9130458	15.6684	2.33068
500	31.00	-125.56	2.04	-256.3425926	15764.19753	22.7	39.5371405	4.1684	69.381
695	24.00	69.4444	-4.96	-344.3287037	4822.530864	32.4	12.0950705	24.5851	71.1683
1054	40.00	428.444	11.04	4730.740741	183564.642	50.4	460.38633	121.918	108.471
486	27.00	-139.56	-1.96	273.2962963	19475.75309	22	48.8458474	3.83507	25.3074
442	18.00	-183.56	-10.96	2011.462963	33692.64198	19.8	84.5022855	120.085	3.11812
1249	62.50	623.444	33.54	20911.36574	388682.9753	60.2	974.830046	1125.04	5.37961
995	53.75	369.444	24.79	9159.143519	136489.1975	47.5	342.319523	614.627	39.5614
1397	79.50	771.444	50.54	38990.08796	595126.5309	67.6	1492.59747	2554.46	141.789
				137992.8333	**2755432.889**		**6910.71887**	**7771.44**	**860.719**

12.6, page: 446

a.



b. $b_1 = \frac{SSXY}{SSX} = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sum(X-\bar{X})^2} = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sum X^2 - \frac{(\sum X)^2}{n}}$

$$b_1 = \frac{SSXY}{SSX} = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sum(X-\bar{X})^2} = \frac{137992.833}{2755432.889} = 0.05$$

$$b_0 = \bar{Y} - b_1\bar{X} = 28.9583 - 0.05(625.556) = 2.3697$$

c. For every cubic foot increase in the amount moved, predicted mean labor hours are estimated to increase by 0.05 hours.

d. $Y = b_0 + b_1 X = -2.3697 + 0.05(500) = 22.67$ hours.

e. The labor hours are affected by the amount to be moved.

12.18, page: 451

Using the result of problem 12.6,

a. $r^2 = \frac{SSR}{SST} = \frac{\sum(\hat{Y}-\bar{Y})^2}{\sum(Y-\bar{Y})^2} = \frac{6910.72}{7771.44} = 0.8892$

88.92% of the variation in labor hours can explained by the variation in cubic feet moved.

b. $S_{YX} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum(Y-\hat{Y})^2}{n-2}} = \sqrt{\frac{860.7186}{36-2}} = 5.0314$

12.44, page: 465

Using the result of problem 12.6,

b. Using t Test for the Slope:

$H_0 : \beta_1 = 0$, $H_1 : \beta_1 \neq 0$, t crit. = 2.0322,

t stat. = $\dfrac{b_1 - \beta_1}{S_{b_1}} = \dfrac{0.05 - 0}{0.00303} = 16.5223$

$$S_{b_1} = \frac{S_{YX}}{\sqrt{SSX}} = \frac{S_{YX}}{\sqrt{\sum (X - \bar{X})^2}} = \frac{5.0314}{\sqrt{2755432.89}} = 0.00303$$

t stat. = 16.5223 > t crit. = 2.0322. At the 0.05 level of significance we reject H_0 , we conclude that there is evidence of a linear relationship between the number of cubic feet moved and the labor hours.

a. Using F Test for the Slope:

$H_0 : \beta_1 = 0$, $H_1 : \beta_1 \neq 0$, F crit. = 4.13,

$$Fstat = \frac{MSR}{MSE} = \frac{SSR/1}{SSE/n-2} = \frac{\sum (\hat{Y} - \bar{Y})^2 /1}{\sum (Y - \bar{Y})^2 /n-2} = \frac{6910.72}{7771.44/34} = \frac{6910.72}{25.32} = 272.99$$

F stat. = 272.99 > F crit. = 4.13. At the 0.05 level of significance we reject H_0 , we conclude that there is evidence of a linear relationship between the number of cubic feet moved and the labor hours.

a. Using t Test for the Correlation Coefficient:

$H_0 : \rho = 0$, $H_1 : \rho \neq 0$, t crit. = 2.0322,

t stat. = $\dfrac{r - \rho}{\sqrt{\dfrac{1-r^2}{n-2}}} = \dfrac{0.97 - 0}{\sqrt{\left(\dfrac{1-(0.97)^2}{36-2}\right)}} = 23.27$

$$r = \frac{\mathrm{cov}(X,Y)}{S_x S_y} = \frac{\left(\sum (X - \bar{X})(Y - \bar{Y})\right)/n-1}{\sqrt{\dfrac{\sum (X - \bar{X})^2}{n-1}}\sqrt{\dfrac{\sum (Y - \bar{Y})^2}{n-1}}} = \frac{137992.8333/34}{(280.583)(14.9)} = 0.97$$

t stat. = 23.27 > t crit. = 2.0322. At the 0.05 level of significance we reject H_0 , we conclude that there is evidence of a linear relationship between the number of cubic feet moved and the labor hours.

Example(2): (Hesah Elawad)
Haverty's Furniture is a family business that has been selling to retail customers in the Chicago area for many years. The company advertises extensively on radio, TV, and the Internet, emphasizing low prices and easy credit terms. The owner would like to review the relationship between sales and the amount spent on advertising .Below is information on sales and advertising expense for the last four months.

Month	Advertising Expense **(X)**	**Sales Revenue (Y)**
July	2	7
August	1	3
September	3	8
October	4	10

Find:

1) The regression coefficients b_0 ,b_1 ,

2) Write the estimated regression equation (prediction line).

3) Estimate sales when $3 million is spent on advertising.

Solution(1):

Month	**(X)**	**(Y)**	**XY**	**X^2**	**Y^2**
July	2	7	14	4	49
August	1	3	3	1	9
September	3	8	24	9	64
October	4	10	40	16	100
Total	10	28	81	30	222

$$n = 4, \sum X = 10 \quad , \sum Y = 28 \quad , \sum XY = 81,$$

$$\sum X^2 = 30, \quad \sum Y^2 = 222$$

1) The regression coefficients b_0 ,b_1 ,

$$b_1 = \frac{SSXY}{SSX} = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sum X^2 - \frac{(\sum X)^2}{n}} = \frac{81 - \frac{(10)(28)}{4}}{30 - \frac{(10)^2}{4}} = \frac{81 - \frac{280}{4}}{30 - \frac{100}{4}}$$

$$= \frac{81 - 70}{30 - 25} = \frac{11}{5} = 2.2$$

The regression intercept

$$b_0 = \overline{Y} - b_1 \overline{X}$$

$$\overline{X} = \frac{\sum X_i}{n} = \frac{10}{4} = 2.5$$

$$\overline{Y} = \frac{\sum Y_i}{n} = \frac{28}{4} = 7$$

$$b_0 = \overline{Y} - b_1 \overline{X} = 7 - (2.2)(2.5) = 7 - 5.5 = 1.5$$

2) Write the estimated regression equation (prediction line).

Simple Linear Regression Equation:

$$\hat{Y} = b_0 + b_1 X = 1.5 + 2.2X$$

3) Estimate sales when $3 million is spent on advertising.

$$\hat{Y} = 1.5 + 2.2X = 1.5 + 2.2(3) = 1.5 + 6.6 = 8.1$$

Solution(2):

(X)	**(Y)**	$(X - \overline{X})$	$(Y - \overline{Y})$	$(X - \overline{X})(Y - \overline{Y})$	$(X - \overline{X})^2$	$(Y - \overline{Y})^2$
2	7	-0.5	0	0	0.25	0
1	3	-1.5	-4	6	2.25	16
3	8	0.5	1	0.5	0.25	1
4	10	1.5	3	4.5	2.25	9
10	28	0	0	11	5	26

$$n = 4, \sum X = 10 \quad , \sum Y = 28$$

$$\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}) = 11$$

$$\sum_{i=1}^{n} (X_i - \bar{X})^2 = 5$$

$$\sum_{i=1}^{n} (Y_i - \bar{Y})^2 = 26$$

$$\bar{X} = \frac{\sum X_i}{n} = \frac{10}{4} = 2 \cdot 5$$

$$\bar{Y} = \frac{\sum Y_i}{n} = \frac{28}{4} = 7$$

1) The regression coefficients b_0 ,b_1 ,

$$b_1 = \frac{SSXY}{SSX} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sum (X - \bar{X})^2} = \frac{11}{5} = 2.2$$

The regression intercept

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

$$\bar{X} = \frac{\sum X_i}{n} = \frac{10}{4} = 2.5$$

$$\bar{Y} = \frac{\sum Y_i}{n} = \frac{28}{4} = 7$$

$$b_0 = \bar{Y} - b_1 \bar{X} = 7 - (2.2)(2.5) = 7 - 5.5 = 1.5$$

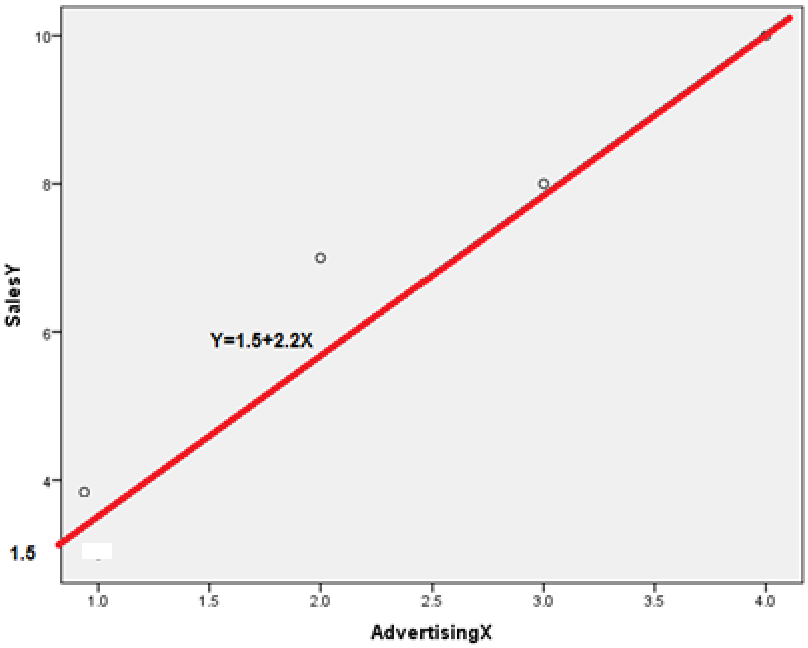
2) Write the estimated regression equation (prediction line).

Simple Linear Regression Equation:

$$\hat{Y} = b_0 + b_1 X = 1.5 + 2.2X$$

3) Estimate sales when $3 million is spent on advertising.

$$\hat{Y} = 1.5 + 2.2X = 1.5 + 2.2(3) = 1.5 + 6.6 = 8.1$$



Y=1.5+2.2X
SalesY
AdvertisingX
10
8
6
4
1.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0