Topic 3: CORRELATION AND REGRESSION

3.1 Correlation

- Scatter Plot
- Pearson product moment correlation coefficient

3.2 Simple Linear Regression (SLR)

- Regression equation using least of square method
- Interpretation of the estimated parameter
- Prediction
- Coefficient of determination

CORRELATION

- Correlation- produce a measurement that describe the strength / type of relationship between variables.
- Describe the degree / strength and the type of the relationship between the two variables.
- The relationship will range from very strong to no relation.

An Overview of Regression and Correlation

- Regression and Correlation are two concepts used to describe relationship between variables (independent and dependent variables)
- Inferential statistic – determining whether a relationship between two or more numerical or quantitative variables exists
- Independent variable – variable that can be controlled or manipulated
- Dependent variable – variable that cannot be controlled or manipulated

Example 1

Discuss relationship between monthly income and monthly expenditure / savings

- Independent variable, X : Monthly income
- Dependent variable, Y : Monthly saving / expenditure

Example 2

Discuss relationship between number of counters opened at a bank and waiting time

- Independent variable, X : Number of counters opened
- Dependent variable, Y : Waiting time

SCATTER DIAGRAM (OR SCATTERPLOT)

x-axis: **Independent Variable (X)**

y-axis: **Dependent Variable (Y)**

A plot of paired observations is called a *scatter diagram* / *Scatterplot.*

- Given a scatter plot, one must be able to draw the line of best fit.
- Purposes – enable to see the trend and predictions on the basis of the data.

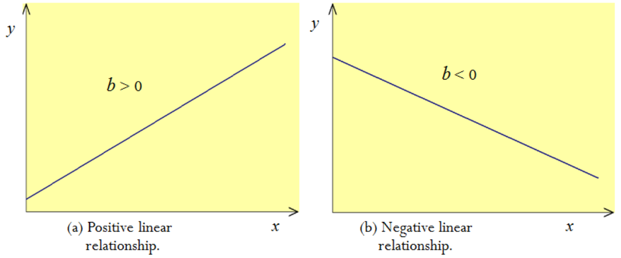
SCATTER PLOT AND ITS USES

- Initial tool to study relationship between **two quantitative** random variables.
- Indicates the **degree of linear relationship (perfect,high, moderate or low)** between two random variables.
- If the points are widely scattered, then it indicates low correlation between the variables.
- The less scattered are the points in a linear pattern, the higher is the degree of relationship.
- It also indicates whether the relationship is linear positive or linear negative.

INTERPRETING SCATTER PLOTS

- (Perfect/high/moderate/low) Positive linear relationship
- (Perfect/High/moderate/low) Negative linear relationship
- Nonlinear relationship
- No relationship

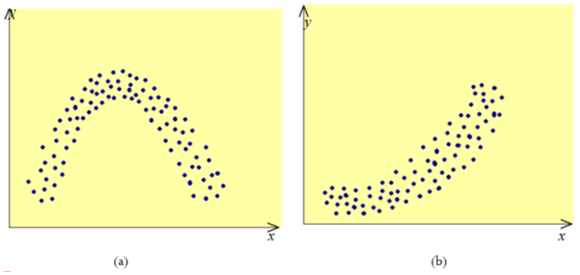
FIGURE 1: POSITIVE AND NEGATIVE LINEAR RELATIONSHIPS BETWEEN X AND Y.



(a) Positive linear relationship.

(b) Negative linear relationship.

FIGURE 2: NONLINEAR RELATIONS BETWEEN *X* AND *Y*.



(a)

(b)

FIGURE 3: LINEAR CORRELATION BETWEEN TWO VARIABLES.

Perfect positive linear relationship, $r = 1$

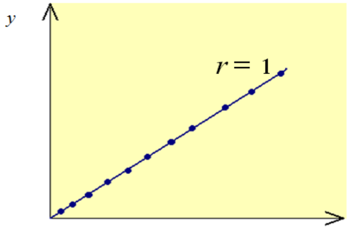


FIGURE 4: LINEAR CORRELATION BETWEEN TWO VARIABLES.

Perfect negative linear relationship, $r = -1$

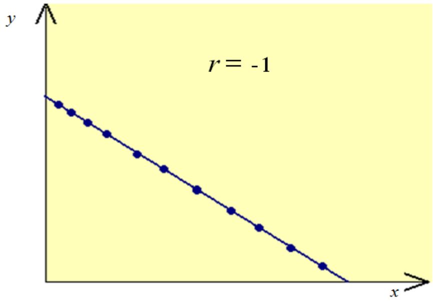


FIGURE 5: LINEAR CORRELATION BETWEEN TWO VARIABLES.

Strong positive linear relationship **(r is close to 1)**

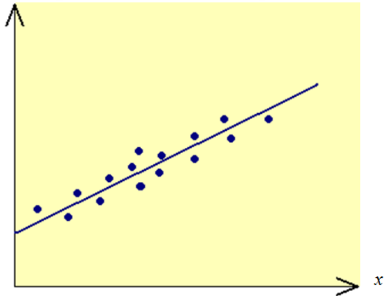


FIGURE 6: LINEAR CORRELATION BETWEEN TWO VARIABLES.

Strong negative linear relationship (r is close to -1)

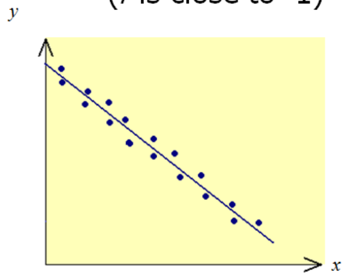


FIGURE 7: LINEAR CORRELATION BETWEEN TWO VARIABLES.

Weak positive linear relationship (r is positive but close to 0)

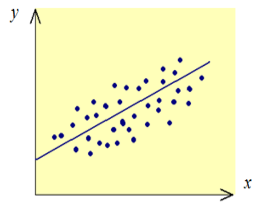


FIGURE 8: LINEAR CORRELATION BETWEEN TWO VARIABLES.

Weak negative linear relationship (r is negative and close to 0)

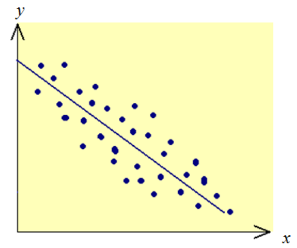
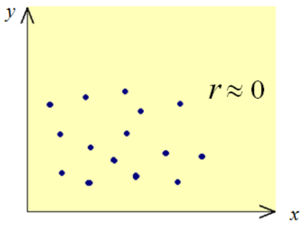


FIGURE 9: LINEAR CORRELATION BETWEEN TWO VARIABLES.

No linear relationship, $r \approx 0$



Example 3

Suppose we take a sample of seven households from a low-to-moderate-income neighborhood and collect information on their income and food expenditure for the past month. The information obtained (in hundreds of dollars) is given in Table 1.

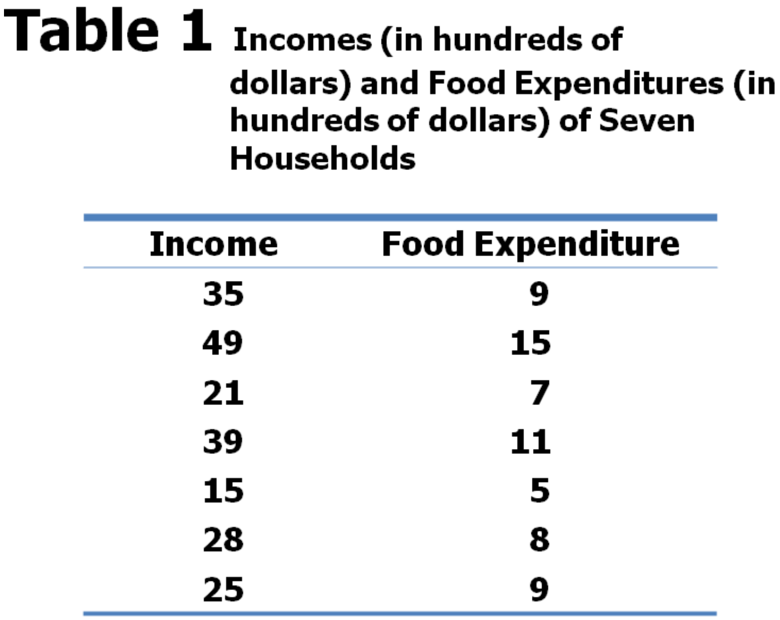


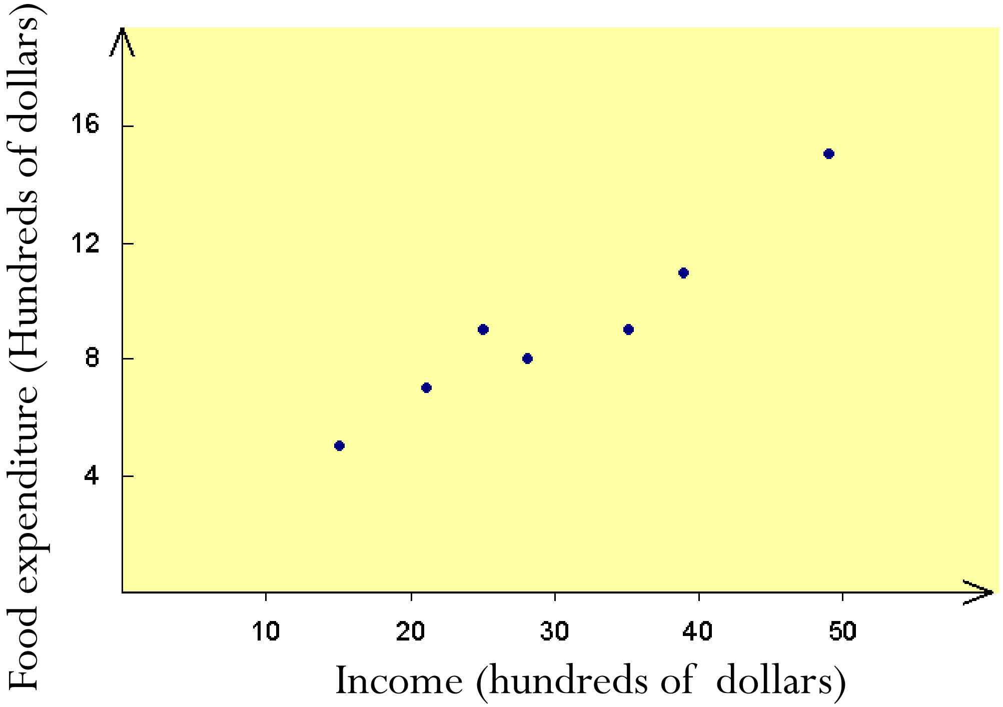
Table 1 Incomes (in hundreds of dollars) and Food Expenditures (in hundreds of dollars) of Seven Households

Income	Food Expenditure
35	9
49	15
21	7
39	11
15	5
28	8
25	9

Construct a scatter diagram of the data. Does a simple linear regression model seem appropriate in this situation? Explain.

Solution 3

The scatterplot shows the relationship between Income and Food Expenditure.



➢**By looking at the scatter diagram of Figure 3, we observe that there exists a strong positive linear relationship between food expenditure and income.**

Pearson Correlation coefficient, r

- Both variable must be quantitative and normally distributed .
- The value of the correlation coefficient always lies in the range of -1 to 1; that is **-1 ≤ r ≤ 1**.
- **r** and **b** have the **same sign**.

r	Interpretation/Explanation/Comment
r=0	No relationship
r ≤ 0.5	Low positive linear relationship
r ≥ -0.5	Low negative linear relationship
0.5 < r < 0.7	Moderate positive linear relationship
-0.7 < r < -0.5	Moderate negative linear relationship
r ≥ 0.7	High positive linear relationship
r ≤ -0.7	High negative linear relationship
r=1	Perfect positive linear relationship
r=-1	Perfect negative linear relationship

➢Calculation for r :

$$r = \frac{n\left(\sum XY\right) - \left(\sum X\right)\left(\sum Y\right)}{\sqrt{\left(n\left(\sum X^2\right) - \left(\sum X\right)^2\right)\left(n\left(\sum Y^2\right) - \left(\sum Y\right)^2\right)}}$$

OR

$$r = \frac{\sum XY - \frac{\left(\sum X\right)\left(\sum Y\right)}{n}}{\sqrt{\left(\sum X^2 - \frac{\left(\sum X\right)^2}{n}\right)\left(\sum Y^2 - \frac{\left(\sum Y\right)^2}{n}\right)}}$$

Example 4

Suppose we take a sample of seven households from a low-to-moderate-income neighborhood and collect information on their income and food expenditure for the past month. The information obtained (in hundreds of dollars) is given in Table 1.

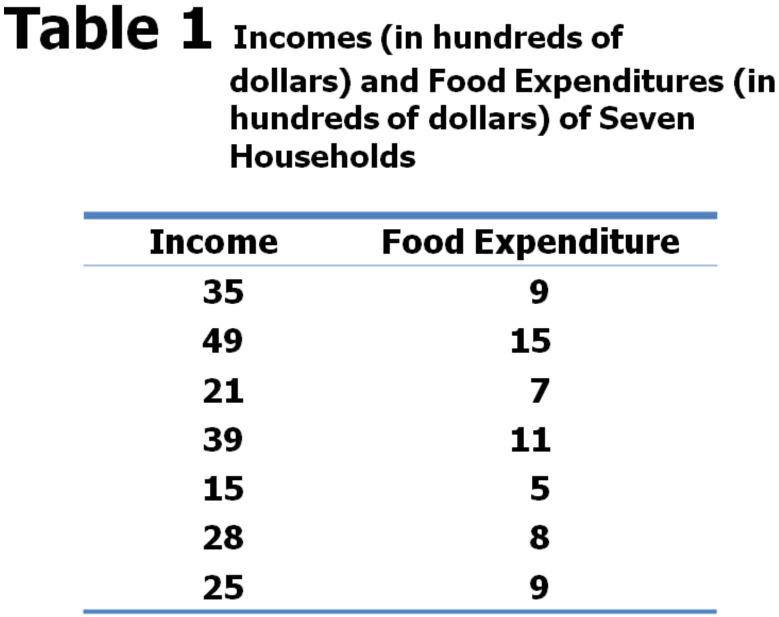


Table 1 Incomes (in hundreds of dollars) and Food Expenditures (in hundreds of dollars) of Seven Households

Income	Food Expenditure
35	9
49	15
21	7
39	11
15	5
28	8
25	9

Find the sample correlation coefficient between the income and food expenditure. Comment on the value obtained.

Solution 4

Income x	Food Expenditure y	xy	x^2	y^2
35	9	315	1225	81
49	15	735	2401	225
21	7	147	441	49
39	11	429	1521	121
15	5	75	225	25
28	8	224	784	64
25	9	225	625	81
Σx = 212	Σy = 64	Σxy = 2150	Σx^2 = 7222	Σy^2 = 646

$$r = \frac{\sum XY - \frac{\left(\sum X\right)\left(\sum Y\right)}{n}}{\sqrt{\left(\sum X^2 - \frac{\left(\sum X\right)^2}{n}\right)\left(\sum Y^2 - \frac{\left(\sum Y\right)^2}{n}\right)}}$$

$$= \frac{211.7143}{\sqrt{(801.4286)(60.8571)}} = 0.9587$$

OR

$$r = \frac{n\sum XY - \left(\sum X\right)\left(\sum Y\right)}{\sqrt{\left(n\sum X^2 - \left(\sum X\right)^2\right)\left(n\sum Y^2 - \left(\sum Y\right)^2\right)}}$$

$$= \frac{1482}{\sqrt{(5610)(426)}} = 0.9587$$

r = 0.96. The value of r=0.96 indicates that the food expenditure and the income are linearly, strong positive related.

Topic 3: CORRELATION AND REGRESSION

3.1 **Correlation**

- Scatter Plot
- Pearson product moment correlation coefficient
- Spearman's Rank Correlation Coefficient

3.2 **Simple Linear Regression (SLR)**

- Regression equation using least of square method
- Interpretation of the estimated parameter
- Prediction
- Coefficient of determination

SIMPLE LINEAR REGRESSION

Simple Regression

A ***simple regression*** model includes only two variables: **one independent** and **one dependent**. The dependent variable is the one being explained, and the independent variable is the one used to explain the variation in the dependent variable.

Linear Regression

A (simple) regression model that gives a **straight-line** relationship between two variables is called a ***linear regression*** model.

Regression- produce a prediction equation that express y (dependent) as a function of x (independent).

Regression line (Regression equation): $\hat{y} = a + bx$

➢Analyze the relationship between the two quantitative variables, X and Y

Regression line (Regression equation)

$$\hat{y} = a + bx$$

- The equation of the best-fitting line is calculated using a set of n pairs (X_i, Y_i).
- We choose our estimates a and b to estimate A and B so that the vertical distances of the points from the line, are minimized.

Interpretation of a and b

Slope, b

- change in the mean of the distribution of the response produced by a unit change in *x*.
- Change in y due to change of one unit in x.
- In general,
  - If **b positive**, if x incraeses by 1 unit y will increse by b units.
  - If **b negative**, if x incraeses by 1 unit y will decrease by b units.
  - Or for each additional unit of x, y will change by b units.

Constant term or y intercept, a

- If $x = 0$ is in the range, then a is the mean of the distribution of the response *y*.
- If $x = 0$ is not in the range, then a has no practical interpretation.

Calculation for a and b Using method of least squares

Slope, b

$$b = \frac{n\left(\sum XY\right) - \left(\sum X\right)\left(\sum Y\right)}{n\left(\sum X^2\right) - \left(\sum X\right)^2}$$

OR

$$b = \frac{\left(\sum XY\right) - \left(\frac{\left(\sum X\right)\left(\sum Y\right)}{n}\right)}{\left(\left(\sum X^2\right) - \frac{\left(\sum X\right)^2}{n}\right)}$$

Constant term or y-intercept, a

$$a = \frac{\left(\sum Y\right)}{n} - b\frac{\left(\sum X\right)}{n}$$

EXAMPLE 6

A study was carried out to determine the relationship between the age and the time (in minutes) needed to run a 12 kilometre marathon event. The following table shows the data recorded.

Age (years)	Time (minutes)
40	61
50	81
66	92
45	70
61	87
48	76
50	88
46	68

i) Find the least-squares regression equation.

ii) Interpret the values of a and b obtained in part i).

SOLUTION 6

$$n = 8 \qquad \sum Y = 623$$

$$\sum X = 406 \qquad \sum Y^2 = 49359$$

$$\sum X^2 = 21122 \qquad \sum XY = 32195$$

Solution (i)

$$a = 21.2164$$

$$b = 1.1164$$

$$\widehat{y} = a + bx$$

$$\widehat{y} = 21.2164 + 1.1164x$$

Solution (ii)

a = 21.2164

No practical interpretation for a since x=0 not in the range of X.

b= 1.1164

- If the age increases by 1 year old, the time needed to run a 12 kilometre marathon event will increase by 1.1 minutes.

Making Prediction

- We can predict the value of dependent variable if the value of independent variable is given by using the equation below.

$$\hat{y} = a + bx$$

- Substitute the value of independent variable (x) into the Regression equation.

Example 7

Suppose we take a sample of seven households from a low-to-moderate-income neighborhood and collect information on their income and food expenditure for the past month. The information obtained (in hundreds of dollars) is given in Table 1.

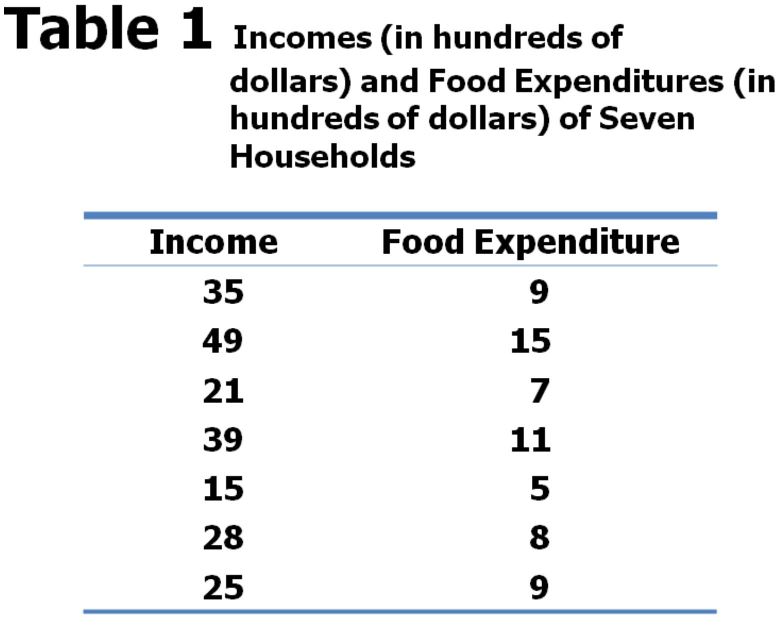


Table 1 Incomes (in hundreds of dollars) and Food Expenditures (in hundreds of dollars) of Seven Households

Income	Food Expenditure
35	9
49	15
21	7
39	11
15	5
28	8
25	9

Use the equation of the fitted line to predict what the value of food expenditure would be when the income is Forty five hundreds of dollars.

Solution 7

Income x	Food Expenditure y	xy	x^2	y^2
35	9	315	1225	81
49	15	735	2401	225
21	7	147	441	49
39	11	429	1521	121
15	5	75	225	25
28	8	224	784	64
25	9	225	625	81
Σx = 212	Σy = 64	Σxy = 2150	Σx^2 = 7222	Σy^2 = 646

$$b = \frac{\left(\sum XY\right) - \left(\frac{\left(\sum X\right)\left(\sum Y\right)}{n}\right)}{\left(\left(\sum X^2\right) - \frac{\left(\sum X\right)^2}{n}\right)} = \frac{211.7143}{801.4286} = 0.2642$$

$$a = \frac{\left(\sum Y\right)}{n} - b\frac{\left(\sum X\right)}{n}$$

$$= 9.1429 - (0.2642)(30.2857)$$

$$= 1.1414$$

The least squares regression line is

ŷ = 1.1414 + 0.2642x

Food expenditure = 1.1414 + 0.2642 income

Food expenditure = 1.1414 + 0.2642 (45)

= 13.0304 hundred of dollar.

= $1303.04

On average, all households with a monthly income of $4500 spend about $1303.04 per month on food.

Coefficient of determination,r²

- To measure strength of that linear relationship / how well the model fits.
  - ✓ **Coefficient of determination, r² = (correlation coefficient)²**

$$r^2 = \frac{\left(n\left(\sum XY\right) - \left(\sum X\right)\left(\sum Y\right)\right)^2}{\left(n\left(\sum X^2\right) - \left(\sum X\right)^2\right)\left(n\left(\sum Y^2\right) - \left(\sum Y\right)^2\right)}$$

Or

$$r^2 = \frac{\left(\sum XY - \frac{\left(\sum X\right)\left(\sum Y\right)}{n}\right)^2}{\left(\sum X^2 - \frac{\left(\sum X\right)^2}{n}\right)\left(\sum Y^2 - \frac{\left(\sum Y\right)^2}{n}\right)}$$

- **Interpretation :**

  r^2 X 100% of total variations in y is explained by x, the other (100% –(r^2 X 100%)) of variations is explained by other factors.

Example 8

Suppose we take a sample of seven households from a low-to-moderate-income neighborhood and collect information on their income and food expenditure for the past month. The information obtained (in hundreds of dollars) is given in Table 1.

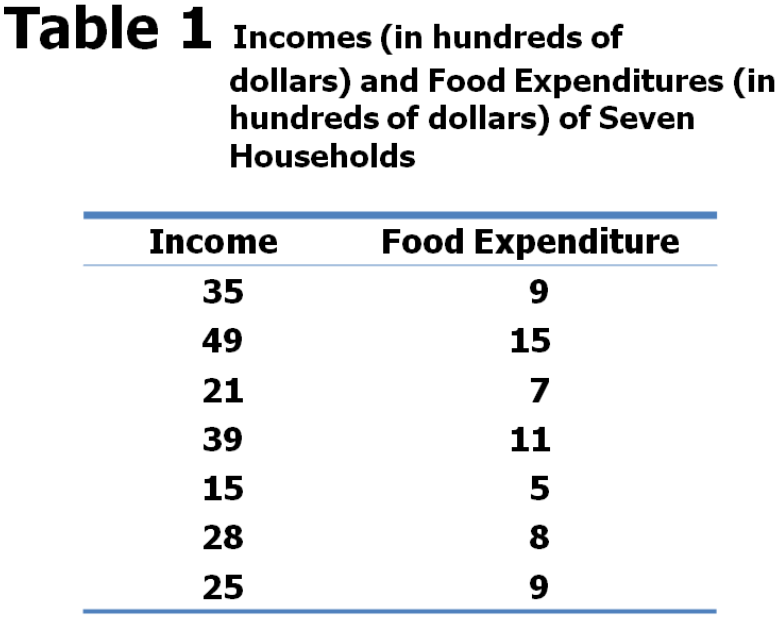
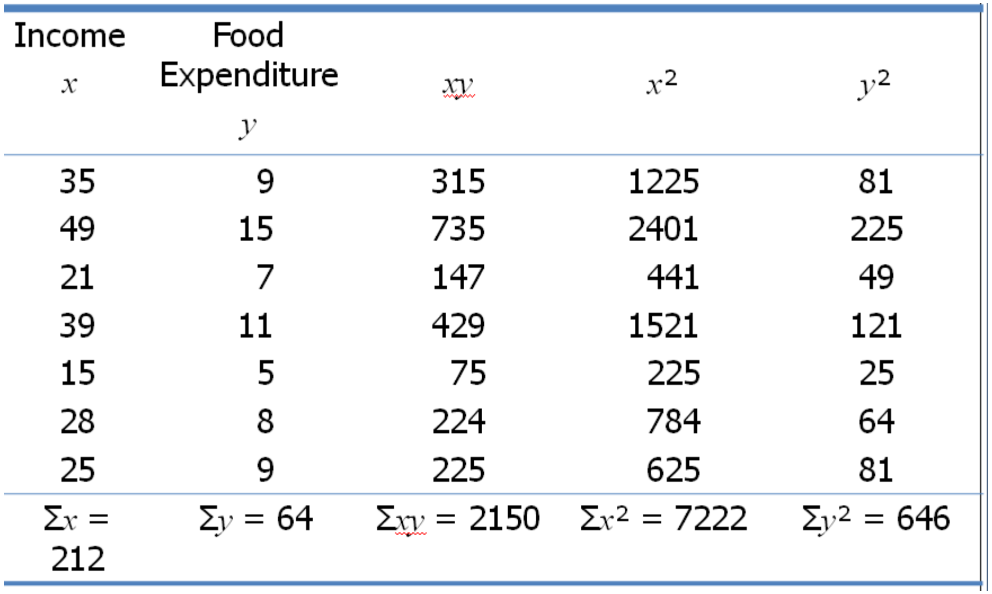


Table 1 Incomes (in hundreds of dollars) and Food Expenditures (in hundreds of dollars) of Seven Households

Income	Food Expenditure
35	9
49	15
21	7
39	11
15	5
28	8
25	9

Find the coefficient of determination and interpret the value.

Solution 8



Income x	Food Expenditure y	xy	x^2	y^2
35	9	315	1225	81
49	15	735	2401	225
21	7	147	441	49
39	11	429	1521	121
15	5	75	225	25
28	8	224	784	64
25	9	225	625	81
$\Sigma x =$ 212	Σy = 64	Σxy = 2150	Σx^2 = 7222	Σy^2 = 646

$$r^2 = \frac{\left(\sum XY - \frac{\left(\sum X\right)\left(\sum Y\right)}{n}\right)^2}{\left(\sum X^2 - \frac{\left(\sum X\right)^2}{n}\right)\left(\sum Y^2 - \frac{\left(\sum Y\right)^2}{n}\right)} = \frac{(211.7143)^2}{(801.4286)(60.8571)} = 0.9216$$

Coefficient of determination, r^2 =0.9216= 0.92

92.16% of total variations in **food expenditure** are explained by the variation in **income of the household**, the other 7.84% of variations is explained by other factors.

Exercise 1

Of the two personnel evaluation techniques available, the first requires a two-hour test interview while the second can be completed in less than an hour. The scores for each of the 15 individuals who took both tests are given in the next table.

Applicant	Test 1	Test 2
1	75	38
2	89	56
3	60	35
4	71	45
5	92	59
6	105	70
7	55	31
8	87	52
9	73	48
10	77	41
11	84	51
12	91	58
13	75	45
14	82	49
15	76	47

a) Construct a scatterplot for the data. Does the assumption of linearity appear to be reasonable?

b) Find the least-square line for the data.

c) Use the regression line to predict the score on the second test for an applicant who scored 85 on test 1.

d) Find the coefficient of correlation between these two variables.