Chapter 7
Simple linear regression and correlation

Department of Statistics and Operations Research

November 8, 2021

Plan

1. Pearson's correlation coefficient
   - Definition
   - Hypotheses testing of correlation coefficient
2. Simple linear regression
   - Least Squares and the Fitted Model
   - Properties of the regression and fitted regression lines
   - Estimation of the error variance
   - Properties of the estimates of β_0 and β_1
   - Inference
   - Coefficient of determination R^2

Plan

1. Pearson's correlation coefficient
   - Definition
   - Hypotheses testing of correlation coefficient
2. Simple linear regression
   - Least Squares and the Fitted Model
   - Properties of the regression and fitted regression lines
   - Estimation of the error variance
   - Properties of the estimates of β_0 and β_1
   - Inference
   - Coefficient of determination R^2

Definition and examples

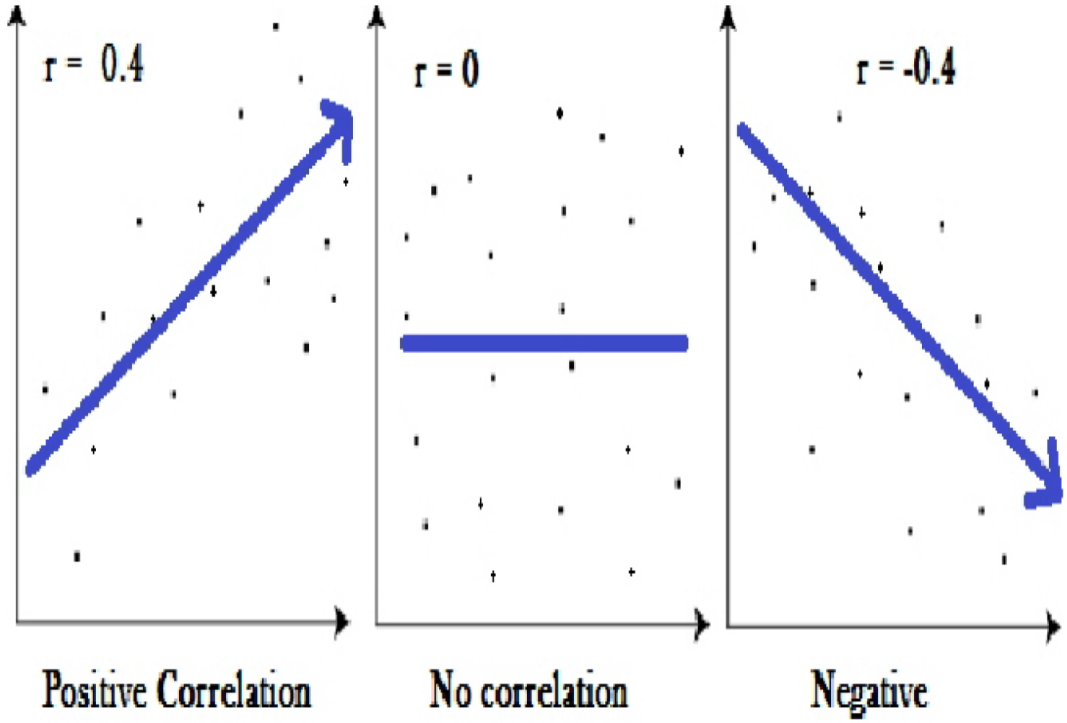
Pearson's r summarizes the relationship between two variables that have a straight line or linear relationship with each other.

(1) If the two variables have a straight line relationship in the positive direction, then r will be positive and considerably above 0.

(2) If the linear relationship is in the negative direction, so that increases in one variable, are associated with decreases in the other, then $r < 0$.

(3) If the linear relationship is constant (no correlation), then $r = 0$.

(4) The possible values of r range from -1 to $+1$, with values close to 0 signifying little relationship between the two variables.


r = 0.4
r = 0
r = -0.4
Positive Correlation
No correlation
Negative

Definition

The most common formula for computing a product-moment correlation coefficient (r) is given below

$$r = \frac{S_{XY}}{\sqrt{S_{XX}}\sqrt{S_{YY}}}$$

where

1. $S_{YY} = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n} Y_i^2 - n\overline{Y}^2$
2. $S_{XX} = \sum_{i=1}^{n}(X_i - \overline{X})^2 = \sum_{i=1}^{n} X_i^2 - n\overline{X}^2$
3. $S_{XY} = \sum_{i=1}^{n}(Y_i - \overline{Y})(X_i - \overline{X}) = \sum_{i=1}^{n} X_i Y_i - n\overline{X}\,\overline{Y}$

where $\overline{X}$ and $\overline{Y}$ are the means of X and Y respectively.

Example 1

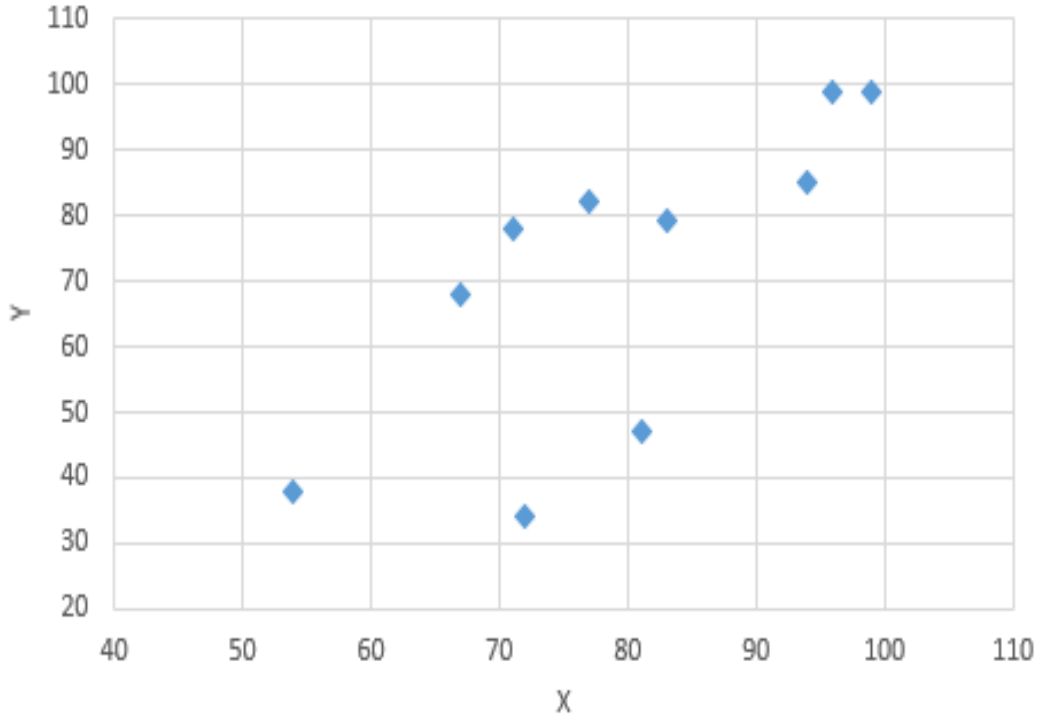
The results of a class of 10 students on midterm exam mark (X) and on the final examination mark (Y) are as follows

X	77	54	71	72	81	94	96	99	83	67
Y	82	38	78	34	47	85	99	99	79	68

1. Construct the scatter diagram.
2. Is there a linear relationship (linear association) between X and Y? Is it positive or negative?
3. Calculate the sample coefficient of correlation (r).

Solution

1) The scatter diagram



2) The scatter diagram suggests that there is a positive linear association between X and Y since there is a linear trend for which the value of Y linearly increases when the value of X increases.

3) Calculating the sample coefficient of correlation (r)

X_i	Y_i	A	B	A^2	B^2	AB
77	82	-2.4	11.1	5.76	123.21	-26.64
54	38	-25.4	-32.9	645.16	1082.41	835.66
71	78	-8.4	7.1	70.65	50.41	-59.64
72	34	-7.4	-36.9	54.76	1361.61	273.06
81	47	1.6	-23.9	2.56	571.21	-38.24
94	85	14.6	14.1	213.16	198.81	205.86
96	99	16.6	28.1	275.56	789.61	466.46
99	99	19.6	28.1	384.16	789.61	550.76
83	79	3.6	8.1	12.96	65.61	29.16
76	68	-12.4	-2.9	153.76	8.41	35.96

where $A = (X_i - \overline{X})$ and $B = (Y_i - \overline{Y})$

We have

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{794}{10} = 79.4 \text{ and } \overline{Y} = \frac{\sum_{i=1}^{n} Y_i}{n} = \frac{709}{10} = 70.9$$

$$S_{YY} = 5040.9 \text{ and } S_{XX} = 1818.4 \text{ and } S_{XY} = 2272.4$$

Then the sample coefficient of correlation is

$$r = \frac{S_{XY}}{\sqrt{S_{XX}}\sqrt{S_{YY}}} = \frac{2272.4}{\sqrt{1818.4}\sqrt{5040.9}} = 0.75056 \approx 0.75$$

Based on our rule, there is a strong positive linear relationship between X and Y. (The values of Y increase when the values of X increase).

Example 2

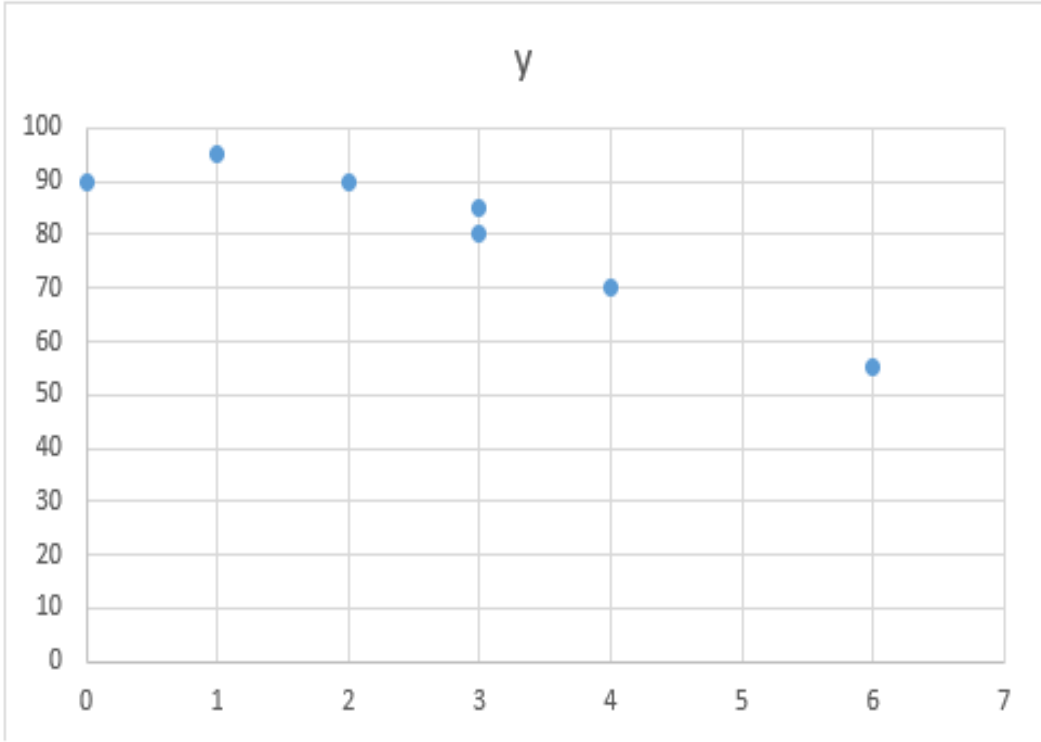
The table below shows the number of absences, x, in a Calculus course and the final exam grade, y, for 7 students.

X	1	0	2	6	4	5	3
Y	95	90	90	55	70	80	85

1. Construct the scatter diagram.
2. Is there a linear relationship (linear association) between X and Y? Is it positive or negative?
3. Calculate the sample coefficient of correlation (r).

Solution

1) The scatter diagram



2) The scatter diagram suggests that there is a negative linear association between X and Y since there is a linear trend for which the value of Y linearly decreases when the value of X increases.

3) Calculating the sample coefficient of correlation (r)
We have

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{19}{7} \text{ and } \overline{Y} = \frac{\sum_{i=1}^{n} Y_i}{n} = \frac{565}{7}$$

$$S_{YY} = \sum_{i=1}^{7} Y_i^2 - 7 \times \overline{Y}^2 = 46775 - 7 \times (\frac{565}{7})^2 = 1171.429$$

$$S_{XX} = \sum_{i=1}^{7} X_i^2 - 7 \times \overline{X}^2 = 75 - 7 \times (\frac{19}{7})^2 = 23.42857$$

$$S_{XY} = \sum_{i=1}^{7} X_i Y_i - 7 \times \overline{X}\,\overline{Y} = 1380 - 7 \times (\frac{19}{7})(\frac{565}{7}) = -153.5714$$

Then the sample coefficient of correlation is

$$r = \frac{S_{XY}}{\sqrt{S_{XX}}\sqrt{S_{YY}}} = \frac{-153.5714}{\sqrt{23.42857}\sqrt{1171.429}} = -0.9269997 \approx -0.93$$

This result shows, there is a strong negative correlation between the number of absences and the final exam grade, since r is very close to -1. Thus, as the number of absence increases, the final exam grade tends to decrease.

Hypotheses testing of correlation coefficient

The sample correlation coefficient, r, is our estimate of the unknown population correlation coefficient. The symbol for the population correlation coefficient is ρ, the Greek letter (rho).
$\rho =$ population correlation coefficient (unknown).
$r =$ sample correlation coefficient (known; calculated from sample data). The hypothesis test lets us decide whether the value of the population correlation coefficient ρ is (close to 0) or (significantly different from 0). We decide this based on the sample correlation coefficient r and the sample size n. For such test, we follow the steps below:
*Setup*1 the hypotheses

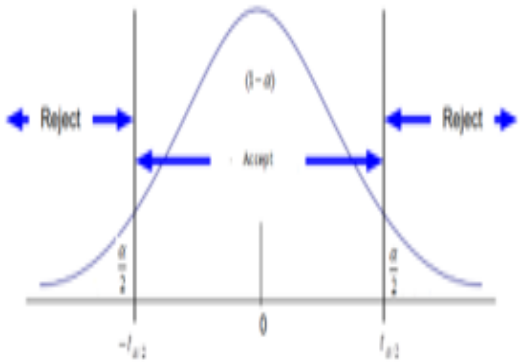
$$\begin{cases} H_0 : \rho = 0. \\ H_1 : \rho \neq 0. \end{cases}$$

Setup2 Calculate the test statistics under $H_0 : \rho = 0$ as

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Where is the simple correlation coefficient calculated from the sample and is the sample size. This statistic follows t distribution with $n - 2$ degrees of freedom.

Setup3 Specify the critical regions



Setup4 Decision When the value of the test statistic belongs to the rejection region, we reject H_0, otherwise accept H_0.

Conclusion: "There is sufficient evidence to conclude that there is a significant linear relationship between x and y because the correlation coefficient is significantly different from 0."

Example 3

Test the significance of the correlation coefficients at 5% level of significance in

a. Example 1

b. Example 2

Solution

a. **From Example 10.1**, we have $r = 0.75, \quad n = 10.$

Step 1: setup the hypotheses

$$H_0 : \rho = 0 \quad versus \quad H_1 : \rho \neq 0$$

Step 2: Calculate the test statistic as

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.75\sqrt{10-2}}{\sqrt{1-0.75^2}} = 3.207.$$

Step 3: Specify the critical regions, sine $t_{\alpha/2} = 2.306$ at 8 degrees of freedom, then the critical region as shown below

Step 4: Decision:

The calculate t=3.207 belongs to the rejection region, so we reject $H_0 : \rho = 0.$

So, we conclude that: "There is sufficient evidence to conclude that there is a significant linear relationship between midterm exam mark (X) and the final examination mark (Y) because the correlation coefficient is significantly different from 0.".

One can get the same conclusion by using p-value approach, that is

$$\text{p-value} = 2[P(T > |3.207|)] = 2[1 - P(T < 3.07)] = 2(1 - 0.9938)$$

$$= 0.013$$

which is less than 5%, so reject H0.

b. From Example 10.2:

From Example 10.2, we have $r = 0.93, \quad n = 7.$

Step 1: setup the hypotheses

$$H_0 : \rho = 0 \quad versus \quad H_1 : \rho \neq 0$$

Step 2: Calculate the test statistic as

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{-0.93\sqrt{7-2}}{\sqrt{1-0.93^2}} = -5.66.$$

Step 3: Specify the critical regions, sine $t_{\alpha/2} = 2.57$ at 5 degrees of freedom, then the critical region as shown below

Step 4: Decision:

The calculate t=-5.66 belongs to the rejection region, so we reject $H_0 : \rho = 0.$

So, we conclude that: "There is sufficient evidence to conclude that there is a significant linear relationship between number of absences (X) and the final exam grade (Y) because the correlation coefficient is significantly different from 0.".

One can get the same conclusion by using p-value approach, that is

$$\text{p-value} = 2[P(T > |-5.66|)] = 2[1 - P(T < 5.66)] = 2(1 - 0.9988)$$

$$= 0.0024$$

which is less than 5%, so reject H0.

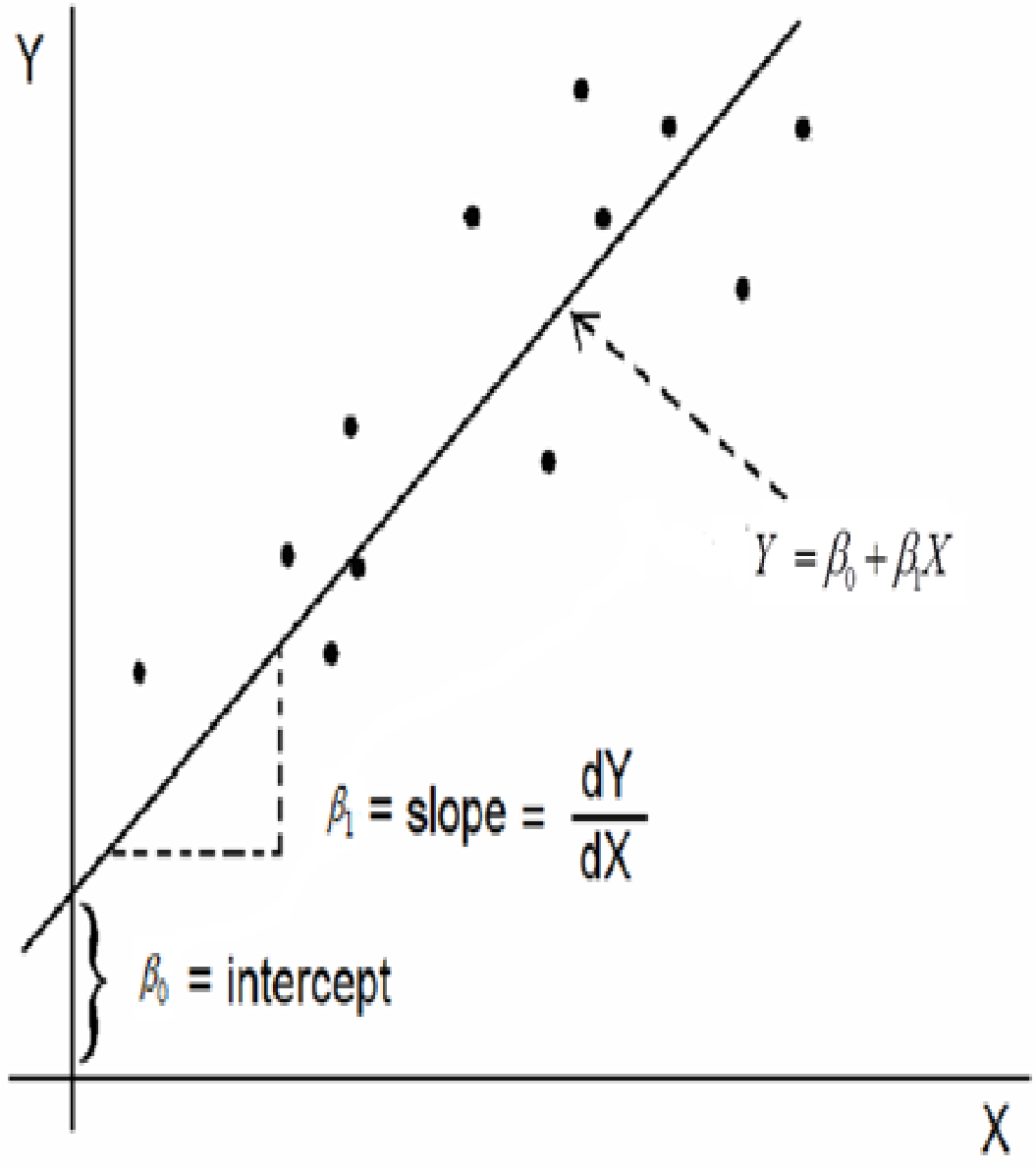
Plan

The simple linear regression model describing the linear relationship between X (independent variable/predictor variable/explanatory variable) and Y (dependent variable/response variable) is given by the following regression line.

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \qquad i = 1, \ldots, n,$$

where,

1. (X_i, Y_i) is the $i - th$ value of the X and Y,
2. e_i is the random term in the regression simple regression line and this term makes the regression analysis as a probabilistic approach,
3. (b_0, b_1) are the parameters of the simple regression line, b_0 is the constant term (intercept) and b_1 is the coefficient of the independent variable X (slope).


Y
$Y = \beta_0 + \beta_1 X$
β_1 = slope = $\frac{dY}{dX}$
β_0 = intercept
X

Least Squares and the Fitted Model

The least squares method is used to find the estimation of parameters (b_0, b_1) . The estimated line is the line that makes the sum of the squares of the vertical distances of the data points from the line as small as possible, computationally (the sum of the error equal zero), this can be seen as the expected value of the random term $E(e_i) = 0$ So, the estimated regression line can be obtained as follows:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \tag{1}$$

where,

1. Y_i is the (random) response for $i - th$ case,
2. β_0, β_1 are the parameters,
3. X_i is a known constant, the value of the predictor variable for the $i - th$ case,
4. ε_i is a random error term, such that,

$$E(\varepsilon_i) = 0, \quad Var(\varepsilon_i) = \sigma^2, \quad Cov(\varepsilon_i, \varepsilon_j) = 0, \ \ i \neq j$$

Least square estimates coefficients

Theorem

The least square estimates coefficients of the simple regression model can also be written in terms of linear form of Y_i as

$$b_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{\sum_{i=1}^{n} X_i Y_i - n\overline{X}\ \overline{Y}}{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2} = r\sqrt{\frac{S_{YY}}{S_{XX}}}$$

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

We can written b_0 and b_1 with another form:

$$b_1 = \sum_{i=1}^{n} \frac{(X_i - \overline{X})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} Y_i = \sum_{i=1}^{n} K_i Y_i$$

and

$$b_0 = \sum_{i=1}^{n} \left(\frac{1}{n} - \overline{X} K_i \right) Y_i = \sum_{i=1}^{n} L_i Y_i$$

where K_i and L_i are constants, and Y_i is a random variable with mean and variance given above:

$$K_i = \frac{(X_i - \overline{X})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \tag{2}$$

$$L_i = \frac{1}{n} - \overline{X} K_i = \frac{1}{n} - \frac{(X_i - \overline{X})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \tag{3}$$

Definition

The fitted regression line, also known as the prediction equation is:

$$\widehat{Y}_i = b_0 + b_1 X_i.$$

We shall find b_0 and b_1, the estimates of β_0 and β_1, so that the sum of the squares of the residuals is a minimum. This minimization procedure for estimating the parameters is called the method of least squares. Hence, we shall find b_0 and b_1 so as to minimize

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2$$

SSE is called the error sum of squares.

Example 4

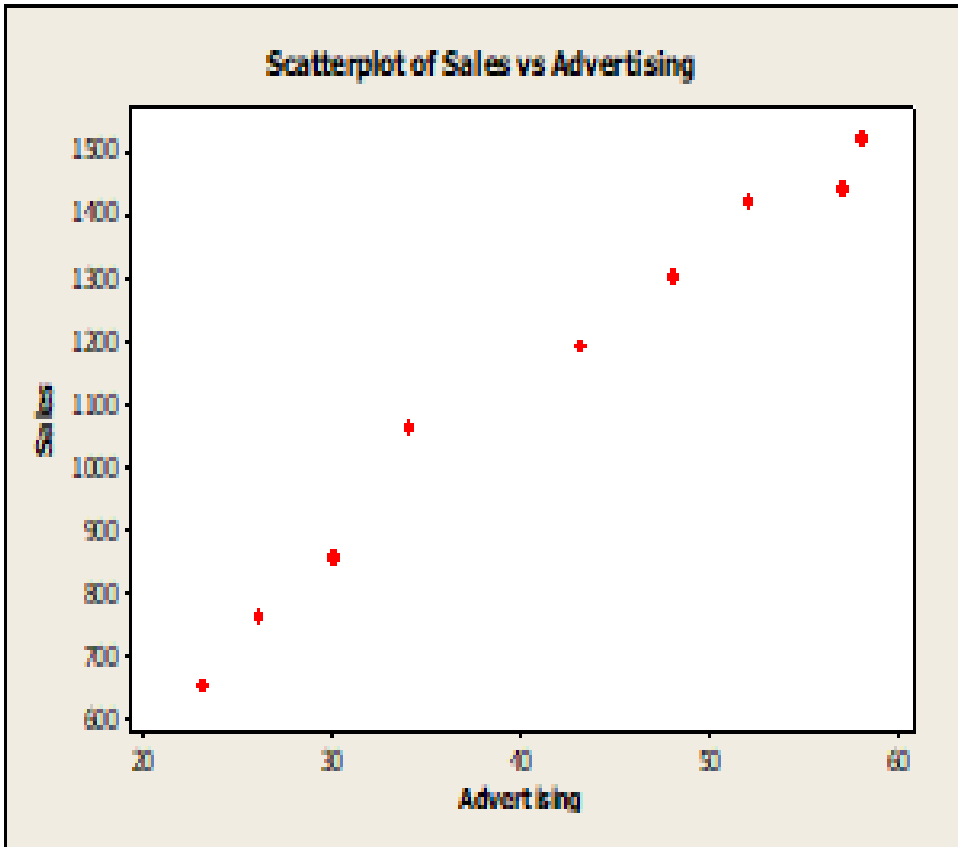
The table below shows some data from the early days of clothing company. Each row in the table shows the company sales for a year, and the amount spent on advertising in that year.

X	23	26	30	34	43	48	52	57	58
Y	651	762	856	1063	1190	1298	1421	1440	1518

1. Draw the scatter diagram of the data and write your comment about it.
2. Find the least square estimate of the simple linear regression model and interpret the result.

Solution

1. The scatter is given by



The scatter diagram shows the relation between the sales and advertising in linear and the correlation coefficient between the Advertising X and Sales Y is given by

$$r = \frac{S_{XY}}{\sqrt{S_{XX}}\sqrt{S_{YY}}} = \frac{33671.56}{\sqrt{1437.56}\sqrt{807485.6}} = 0.988,$$

where

1. $S_{YY} = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = 807485.6$
2. $S_{XX} = \sum_{i=1}^{n}(X_i - \overline{X})^2 = 1437.56$
3. $S_{XY} = \sum_{i=1}^{n}(Y_i - \overline{Y})(X_i - \overline{X}) = 33671.56$

b. From the data we have

$$\overline{X} = 41.22, \qquad \overline{Y} = 1133.22, \qquad n = 9$$

$$\sum_{i=1}^{9} X_i^2 = 16731, \qquad \sum_{i=1}^{9} Y_i^2 = 12365219, \qquad \sum_{i=1}^{9} X_i Y_i = 454097$$

$$S_{XX} = \sum_{i=1}^{9} X_i^2 - 9 \times \overline{X}^2 = 16731 - 9 \times 41.22^2 = 1437.556$$

$$S_{YY} = \sum_{i=1}^{9} Y_i^2 - 9 \times \overline{Y}^2 = 12365219 - 9 \times 1133.22^2 = 807485.6$$

$$S_{XY} = \sum_{i=1}^{9} X_i Y_i - 9 \times \overline{X}\,\overline{Y} = 454097 - 9 \times 41.22 \times 1133.22 = 33671.56$$

The least-square line

$$b_1 = \frac{S_{XY}}{S_{XX}} = \frac{33671.56}{1437.556} = 23.42279$$

$$b_0 = \overline{Y} - b_1 \times \overline{X} = 1133.22 - 23.42 \times 41.22 = 167.689$$

Finally, we have

$$\widehat{Y} = 167.689 + 23.42\,X$$

The slope b_1 can be calculated using the correlation coefficient as

$$b_1 = r\sqrt{\frac{S_{YY}}{S_{XX}}} = 0.988\sqrt{\frac{807485.6}{1437.556}} = 23.42$$

In this case, our outcome of interest is sales. If we use Advertising as the predictor variable, linear regression estimates that

$$Sales = 167.7 + 23.42\, Advertising.$$

That is, if advertising expenditure is increased by one million dollars, then sales will be expected to increase by 23.4 million dollars, and if there was no advertising we would expect sales of 167.7 million dollars.

Assumptions

Important assumptions and properties can be added to the simple linear regression line defined in (1); they are:

1. The error ε_i is normally distributed with mean 0 and variance σ^2. The last point states that the random errors are independent (uncorrelated).
2. Since the error $\varepsilon_i \sim N(0, \sigma^2)$ this also implies that:

$$E(Y_i) = \beta_0 + \beta_1 X_i, \quad Var(Y_i) = \sigma^2, Cov(Y_i, Y_j) = 0, \quad i \neq j,$$

hence the response variable Y_i is normally distributed $N(\beta_0 + \beta_1 X_i, \sigma^2)$

Properties

The fitted regression line with the corresponding errors satisfy the following properties (without proof):

1. The residuals sum equals to 0, $\sum_{i=1}^{n} e_i = 0$
2. The sum of Y equals the sum of the fitted Y,

$$\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \widehat{Y}_i$$

3. The sum of the weighted (by X) residuals is 0,

$$\sum_{i=1}^{n} X_i e_i = 0$$

4. The sum of the weighted (by Y) residuals is 0,

$$\sum_{i=1}^{n} Y_i e_i = 0$$

5. The regression line goes through the point $(\overline{X}, \overline{Y})$

Estimation of the error variance

The fitted values for the individual observations are obtained by plugging in the corresponding level of the predictor variable (X_i) into the fitted equation. The residuals are the vertical distances between the observed values (Y_i) and their fitted values $\widehat{Y}_i$, and are denoted as e_i, are given by

$$e_i = Y_i - \hat{Y}_i,\ i = 1, 2, ..., n.$$

From example 4, we have

$$e_i = Y_i - \hat{Y}_i = Y_i - 167.6829 - 23.42279X_i;\ i = 1, 2, ..., 9.$$

Example

The values of Y_i, $\widehat{Y}_i$ and e_i are given in the following table

$Advertising(X)$	$Sales(Y)$	$\widehat{Y}$	e	e^2
23	651	706.41	−55.41	3070.27
26	762	776.68	−14.68	215.36
30	856	870.37	−14.37	206.38
34	1063	964.06	98.94	9789.52
43	1190	1174.86	15.14	229.22
48	1298	1291.98	6.02	36.24
52	1421	1385.67	35.33	1248.21
57	1440	1502.78	−62.78	3941.33
58	1518	1526.2	−8.2	67.24

Then, we have

$$\sum_{i=1}^{9} e_i^2 \approx 18804$$

Theorem

An unbiased estimate of σ^2, named the mean squared error (MSE), is

$$s^2 = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2} = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{SSE}{n-2}$$

From example 4, we have

$$s^2 = \frac{\sum_{i=1}^{9} e_i^2}{9-2} = \frac{18804}{7} = 2686.286$$

To obtain an estimate of the standard deviation (which is in the units of the data), we take the square root of the error mean square

$$s = \sqrt{MSE} = \sqrt{2686.286} \approx 51.83$$

The coefficients K_i and L_i defined by (2) and (3) satisfy the following properties:

Lemma

The coefficients K_i and L_i satisfy the following properties:

$$\begin{cases} \sum_{i=1}^{n} K_i = 0 & \sum_{i=1}^{n} L_i = 1 \\ \sum_{i=1}^{n} K_i X_i = 1 \quad \textit{and} & \sum_{i=1}^{n} L_i X_i = 0 \\ \sum_{i=1}^{n} K_i^2 = \frac{1}{S_{xx}} & \sum_{i=1}^{n} L_i^2 = \frac{1}{n} + \frac{\overline{X}^2}{S_{xx}} \end{cases}$$

Lemma

1. The point estimators of β_0 and β_1 and are unbiased, i.e.

$$E(b_0) = \beta_0 \quad \textit{and} \quad E(b_1) = \beta_1$$

2. The point estimators of β_1 and β_0 and have the following variances, respectively

$$Var(\beta_1) = Var(b_1) = \frac{\sigma^2}{S_{XX}} = \frac{MSE}{S_{XX}}$$

and

$$Var(\beta_0) = Var(b_0) = MSE\left(\frac{1}{n} + \frac{\overline{X}^2}{S_{XX}}\right)$$

Example 5

Calculate the variances and standard errors of the least square estimators of coefficients of the simple linear regression in Example 4

Solution

For such data, the variances of b_1 and b_0 are given respectively by

$$Var(b_1) = \frac{\sigma^2}{S_{XX}} = \frac{MSE}{S_{XX}} = \frac{2686.276}{1437.56} \approx 1.87$$

$$Var(b_0) = MSE\left(\frac{1}{n} + \frac{\overline{X}^2}{S_{XX}}\right) = 2686.276 \times \left(\frac{1}{9} + \frac{(41.22)^2}{1437.56}\right) \approx 3473.5$$

Hence, the standard errors of b_1 and b_0 are given respectively by

$$S.E(b_1) = \sqrt{Var(b_1)} = \sqrt{1.87} \approx 1.37 \quad \textit{and}$$

$$S.E(b_0) = \sqrt{Var(b_0)} = \sqrt{3473.5} \approx 58.94$$

In this section, we discuss some statistical inferences related to the simple linear regression model, such as constructing confident intervals for the model coefficients and hypotheses testing about the coefficients using t and F tests. We assume the errors follow $N(0, \sigma^2)$. To develop the inference about the model coefficient, we need to present some the following lemmas.

Lemma (Sampling distributions)

Let b_1 and b_0 are the estimators of the slope and the intercept in the simple linear regression model, then each one of the quantities

$$T_1 = \frac{b_1 - \beta_1}{S.E(b_1)} \quad \textit{and} \quad T_0 = \frac{b_0 - \beta_0}{S.E(b_0)} \tag{4}$$

have t distribution with $(n-2)$ degrees of freedom.

Lemma (Interval estimation Concerning the Regression Coefficients)

A $100(1-\alpha)\%$ confidence interval for the parameters β_1 and β_0 in the regression line respectively given by

$$b_1 - t_{1-\frac{\alpha}{2},n-2} \times S.E(b_1) < \beta_1 < b_1 + t_{1-\frac{\alpha}{2},n-2} \times S.E(b_1)$$

and

$$b_0 - t_{1-\frac{\alpha}{2},n-2} \times S.E(b_0) < \beta_0 < b_0 + t_{1-\frac{\alpha}{2},n-2} \times S.E(b_0)$$

where $t_{1-\frac{\alpha}{2},n-2}$ is a value of the t-distribution with $n-2$ degrees of freedom and

$$S.E(b_1) = \sqrt{\frac{MSE}{S_{XX}}} \quad \textit{and} \quad S.E(b_0) = \sqrt{MSE\left(\frac{1}{n} + \frac{\overline{X}^2}{S_{XX}}\right)}$$

Example 6

Consider data in example 4, then find 95% confidence interval for both β_1 and β_0.

Solution

For such data, we have calculated

$$S.E(b_1) = 1.37 \quad \textit{and} \quad t_{1-\frac{\alpha}{2},n-2} = t_{0.975,7} = 2.365$$

Hence the 95% confidence interval of β_1 is given by

$$23.42 - 2.365 \times 1.37 < \beta_1 < 23.42 + 2.365 \times 1.37$$

We get

$$20.2 < \beta_1 < 26.7$$

This can be interpreted as: when the advertising increases by one million, the sales increase with probability 95% within $(20.2, 26.7)$ million.

Similarly, we have calculated

$$S.E(b_0) = 58.94 \quad \textit{and} \quad t_{1-\frac{\alpha}{2},n-2} = t_{0.975,7} = 2.365$$

Hence the 95% confidence interval of β_0 is given by

$$167.68 - 2.365 \times 58.94 < \beta_0 < 167.68 + 2.365 \times 58.94$$

We get

$$28.3 < \beta_0 < 307.1$$

This can be interpreted as: when we have no advertising, the sales will be with probability 95% within $(28.3, 307.1)$ million.

Hypothesis Testing of the parameters β_0 and β_1

The sampling distributions of T_1 and T_0 defined in (4) can be used to test some hypotheses concerning the coefficients of the simple linear regression model. These test are very important to check the validity of the simple linear model.

Steps for testing β_i, $i = 0, 1$

To test β_i, $i = 0, 1$, is equal a certain value, say $\beta_i^{(0)}$, we follow the steps below:

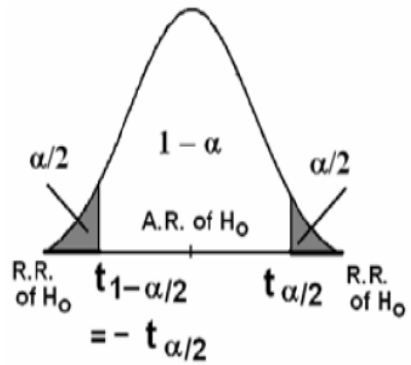
1. Setup the hypotheses

$$H_0 : \beta_i = \beta_i^{(0)} \quad vs \quad H_1 : \beta_i \neq \beta_i^{(0)}$$

2. Test statistic under H_0

$$T_i = \frac{b_i - \beta_i^{(0)}}{S.E(b_i)} \sim t_{n-1}$$

3. Critical regions



R.R: Rejection Region and A.R: Acceptance Region

4. Decision:
When the calculated T_i belongs to the shaded areas, we reject the null hypothesis H_0, otherwise accept H_0.

Remarks

1. In some applications, we may need to test

$$H_0 : \beta_i = 0 \quad vs \quad H_1 : \beta_i \neq 0$$

In these cases, you need to replace $\beta_i^{(0)}$ by zeros.

2. In some applications, we may need to test

$$H_0 : \beta_i = 0 \quad vs \quad H_1 : \beta_i > (<) 0, \;\; i = 0, 1$$

In these cases, you need to replace the critical regions to one-sided critical regions.

3. One may use the two-sided $p - value$ approach

$$p - value = 2P(T > |T_i|), \;\; i = 0, 1$$

then reject H_0 when $p - value < \alpha$, otherwise accept H_0. The one-sided $p - value$ is $p - value = P(T > |T_i|), \;\; i = 0, 1$ then reject H_0 when $p_value < \alpha$, otherwise accept H_0.

Example 7

Consider data in example 4, test the hypotheses at 5% level of significance

$$H_0 : \beta_1 = 0 \quad vs \quad H_1 : \beta_1 \neq 0$$

and

$$H_0 : \beta_0 = 0 \quad vs \quad H_1 : \beta_0 \neq 0$$

Solution

We start by testing β_1. We have the following hypothesis:

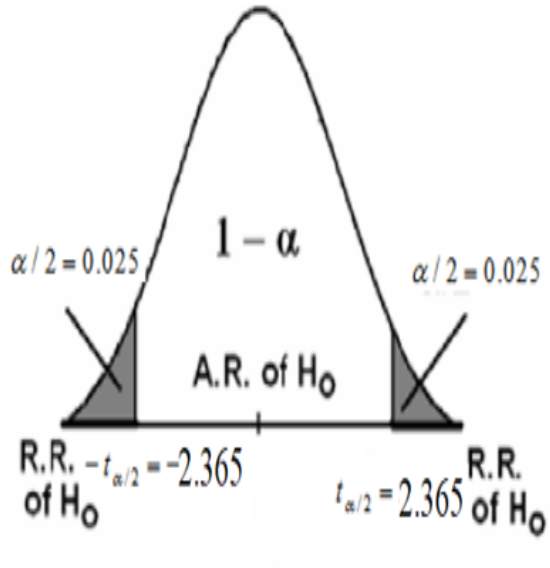
$$H_0 : \beta_1 = 0 \quad vs \quad H_1 : \beta_1 \neq 0$$

Test statistic under H_0 is given by

$$T_1 = \frac{b_1 - \beta_1^{(0)}}{S.E(b_1)} = \frac{23.42 - 0}{1.37} \approx 17.1$$

The critical regions are given by

$$t_{\frac{\alpha}{2},n-2} = t_{0.025,7} = 2.365 \quad \textit{and} \quad -t_{\frac{\alpha}{2},n-2} = -t_{0.025,7} = -2.365$$



R.R: Rejection Region and A.R: Acceptance Region

Decision: The calculated test $T_1 = 17.1$ belongs to the shaded areas, then we reject the null hypothesis H_0. As we can see from the results that, $T_1 = 17.1$. Also, the

$$p - value = 2P(T > |T_1|) = 2P(T > |17.1|) \approx 0.000 < 0.05,$$

then reject H_0.

Now, we test β_0. We have the following hypothesis:

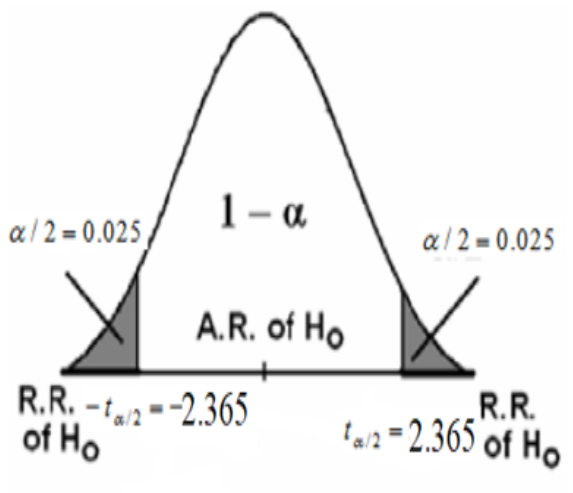
$$H_0 : \beta_0 = 0 \quad vs \quad H_1 : \beta_0 \neq 0$$

Test statistic under H_0 is given by

$$T_0 = \frac{b_0 - \beta_0^{(0)}}{S.E(b_0)} = \frac{167.68 - 0}{58.94} \approx 2.85$$

The critical regions are given by

$$t_{\frac{\alpha}{2},n-2} = t_{0.025,7} = 2.365 \quad and \quad -t_{\frac{\alpha}{2},n-2} = -t_{0.025,7} = -2.365$$



R.R: Rejection Region and A.R: Acceptance Region

Decision: The calculated test $T_1 = 2.85$ belongs to the shaded areas, then we reject the null hypothesis H_0. As we can see from the results that, $T_0 = 2.85$. Also, the

$$p - value = 2P(T > |T_0|) = 2P(T > |2.85|) \approx 0.025 < 0.05,$$

then reject H_0.

Coefficient of determination R^2

The coefficient of determination can also be obtained by squaring the Pearson correlation coefficient. This method works only for the linear regression model

$$\mu_i = \mu_0 + \mu_1 X_i, \quad i = 1, \ldots, n,$$

The method does not work in general. The coefficient of determination, R^2, represents the proportion of the total sample variation in Y (measured by the sum of squares of deviations of the sample Y values about their mean $\overline{Y}$) that is explained by (or attributed to) the linear relationship between X and Y. Some other way to calculate the coefficient of determination as

$$R^2 = \frac{SSR}{SSTOT} = 1 - \frac{SSE}{SSTOT}$$

where the total sum of squared error and the sum of squared regression error are given respectively by

$$SSTOT = \sum_{i=1}^{n} (Y_i - \overline{Y})^2 \quad \textit{and} \quad SSR = \sum_{i=1}^{n} (\widehat{Y}_i - \overline{Y})^2$$

Lemma

1. We have
$$SSTOT = SSE + SSR,$$
2. The coefficient of determination is a number between 0 and 1, inclusive. That is,
$$0 \leq R^2 \leq 1,$$
3. If $R^2 = 0$, the least squares regression line has no explanatory value,
4. If $R^2 = 1$, the regression line explains 100% of the variation in the response variable Y,
5. The simple correlation coefficient can be simply obtained as
$$r = \sqrt{R^2}$$

with sing as the sign of the estimate of the slope b_1.

Example 8

Calculate the coefficient of determination of the simple linear model in Example 4, then integrate the results. Also, calculate Pearson correlation coefficient.

Solution

From the data, we have

$$\begin{cases} SSTOT = S_{YY} = 807485.6 \\ SSE = 18804 \\ SSR = SSTOT - SSE = 807485.6 - 18804 = 788681.6 \end{cases}$$

Then the coefficient of determination equals to

$$R^2 = \frac{SSR}{SSTOT} = \frac{788681.6}{807485.6} = 0.9767$$

The result shows that 97.7% of the total variation in the sales is due to the advertising. The simple correlation coefficient is

$$r = \sqrt{0.9767} = 0.988$$