Chapter 7
Simple Linear Regression and Correlation

Department of Statistics and Operations Research



27 October 2026

Bayan Almukhlif
balmukhlef@ksu.edu.sa

Contents

1 Correlation

2 Test for Correlation ρ

3 Regression Model $\hat{y}$

4 Least Squares Method b

5 Inference on Coefficients t

6 Coefficient of Determination R^2

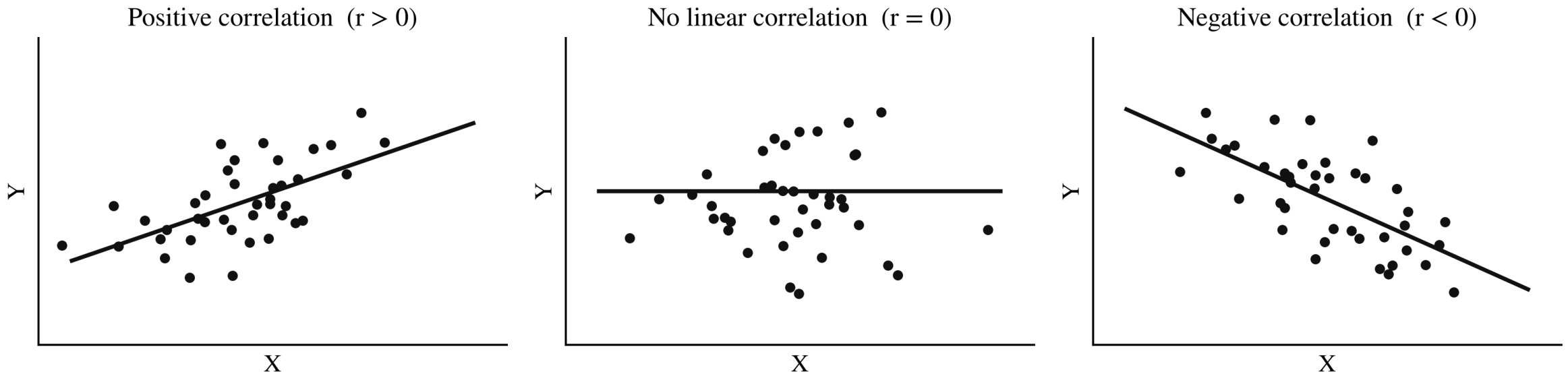
1

Correlation

Pearson's correlation coefficient · two worked examples

Pearson's Correlation Coefficient

The **correlation coefficient r** measures the strength and the direction of a **straight line** relationship between two variables X and Y. We calculate it from a sample of n pairs (x_i, y_i)



Remember

Always draw a scatter diagram first. The number r only describes a straight line pattern

Properties of r

1. **Range**
   r is always between −1 and +1

2. **Positive r**
   Y tends to increase when X increases

3. **Negative r**
   Y tends to decrease when X increases

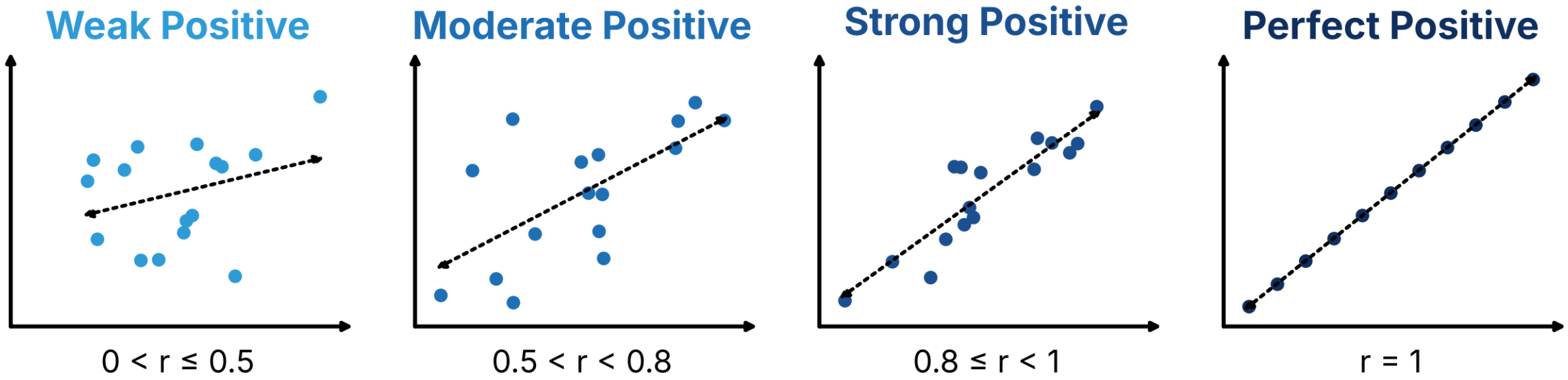
4. **r = 0**
   No linear relationship between X and Y

5. **r = 1 or r = −1**
   All points lie exactly on a straight line (up or down)

6. **Close to 0 or ±1**
   Near 0 the relationship is weak and near ±1 it is strong

Note
r = 0.6 does **not** mean a relationship twice as strong as r = 0.3. To compare strength we use r^2 (Section 6)

Scatter Plots & Correlation Examples



Definition of correlation coefficient (r)

$$r = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}$$

where $\bar{x}$ and $\bar{y}$ are the sample means and

1. $$S_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n} {x_i}^2 - n\overline{x}^2$$

2. $$S_{yy} = \sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n} {y_i}^2 - n\overline{y}^2$$

3. $$S_{xy} = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \sum_{i=1}^{n} x_i y_i - n\overline{xy}$$

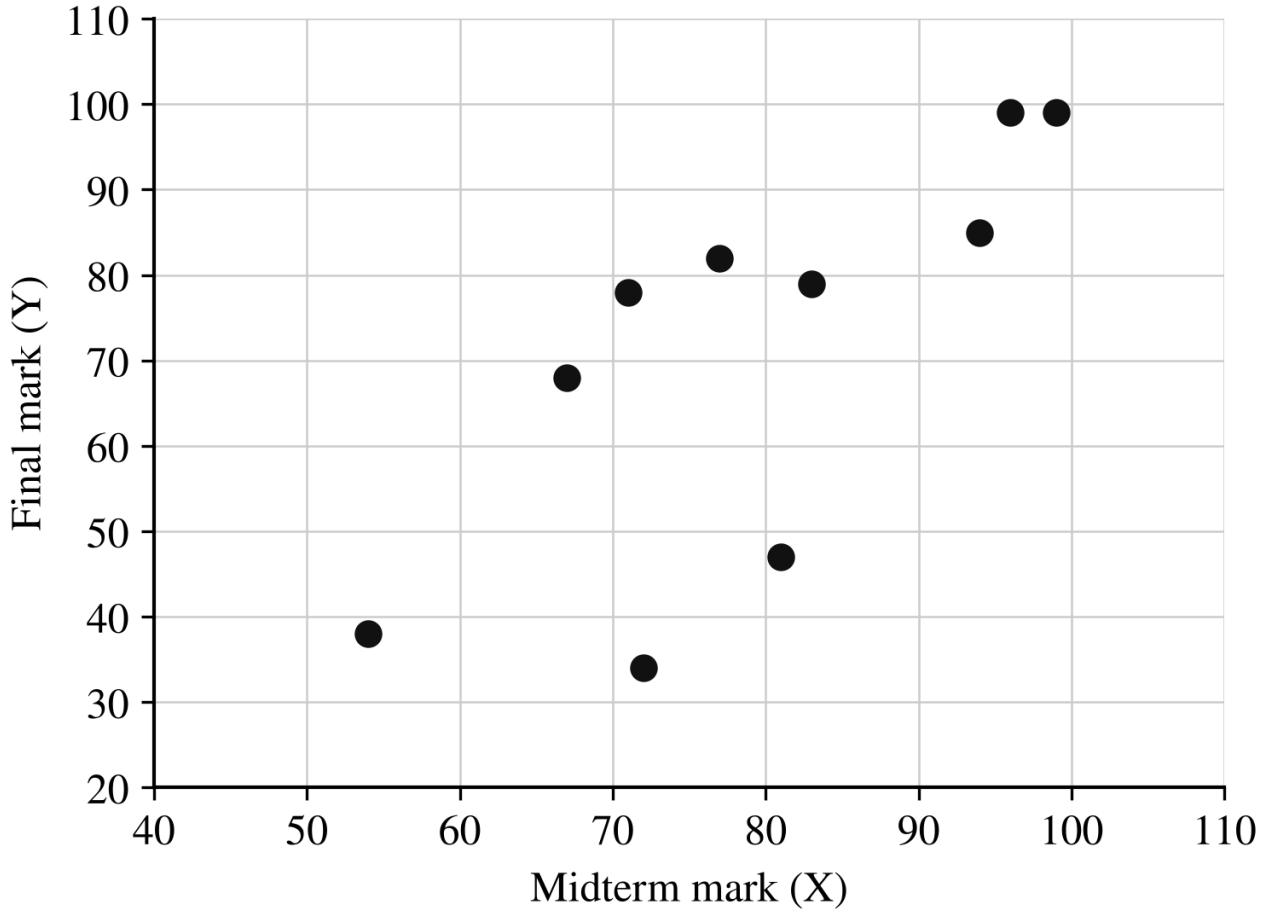
Example 1

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

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

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

Example 1: Solution



1 Scatter diagram

Each point is one student. The points go up from left to right

2 Linear relationship

Comment

The scatter diagram shows a **positive linear association** between X and Y. The value of Y increases when the value of X increases

Example 1: Solution

X	Y	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	70.56	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
67	68	-12.4	-2.9	153.76	8.41	35.96
Sum		**0.0**	**0.0**	**1818.4**	**5040.9**	**2272.4**

$A = (x_i - \bar{x})$ *and* $B = (y_i - \bar{y})$

$$\bar{x} = \frac{794}{10} = 79.4, \quad \bar{y} = \frac{709}{10} = 70.9$$

$$S_{xx} = 1818.4, \quad S_{yy} = 5040.9$$

$$S_{xy} = 2272.4$$

$$r = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} = \frac{2272.4}{\sqrt{1818.4}\sqrt{5040.9}}$$

$$r \approx 0.75$$

Conclusion

r = 0.75 shows a **strong positive** linear relationship. Y increases when X increases

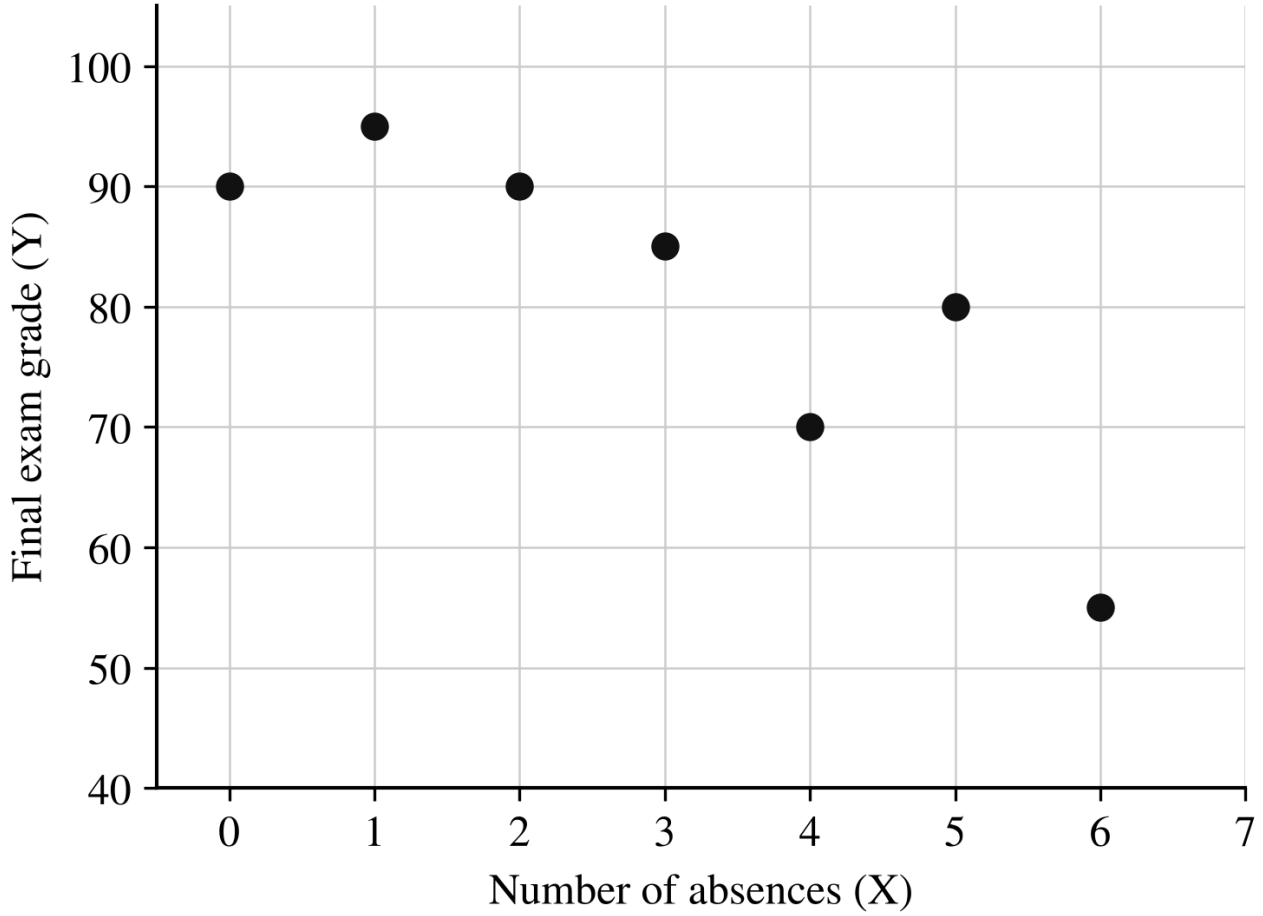
Example 2

The table shows the number of absences (X) in a Calculus course and the final exam grade (Y) for 7 students

X (absences)	1	0	2	6	4	5	3
Y (grade)	95	90	90	55	70	80	85

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

Example 2: Solution



1 Scatter diagram

Each point is one student. The points go down from left to right

2 Linear relationship

Comment

The scatter diagram shows a **negative linear association** between X and Y. The value of Y decreases when the value of X increases

Example 2: Solution

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{21}{7} = 3,$$

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{565}{7} \approx 80.71$$

$$\sum_{i=1}^{n} {x_i}^2 = 91, \quad \sum_{i=1}^{n} {y_i}^2 = 46775$$

$$\sum_{i=1}^{n} x_i y_i = 1540 \;\; n = 7$$

$$S_{xx} = \sum_{i=1}^{n} {x_i}^2 - n\bar{x}^2 = 91 - 7\times 3^2 = 28$$

$$S_{yy} = \sum_{i=1}^{n} {y_i}^2 - n\bar{y}^2 = 46775 - 7\times\left(\frac{565}{7}\right)^2 = 1171.43$$

$$S_{xy} = \sum_{i=1}^{n} x_i y_i - n\bar{x}\,\bar{y} = 1540 - 7\times 3\times\frac{565}{7} = -155$$

Sample Correlation coefficient is

$$r = \frac{-155}{\sqrt{28}\sqrt{1171.43}} \approx -0.86$$

Conclusion

r = −0.86 shows a **strong negative** correlation between the number of absences and the final exam grade. Thus, as the number of absence increases, the final exam grade tends to decrease.

2

Test for Correlation

Is the linear relationship real or only chance?

Test for ρ: Steps (1)

r: is the **sample** correlation (known). **ρ** (rho) is the **population** correlation (unknown).

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:

1 Hypotheses

$\mathbf{H}_0 : \rho = 0$ (no linear relationship)

$\mathbf{H}_1 : \rho \neq 0$ (a linear relationship)

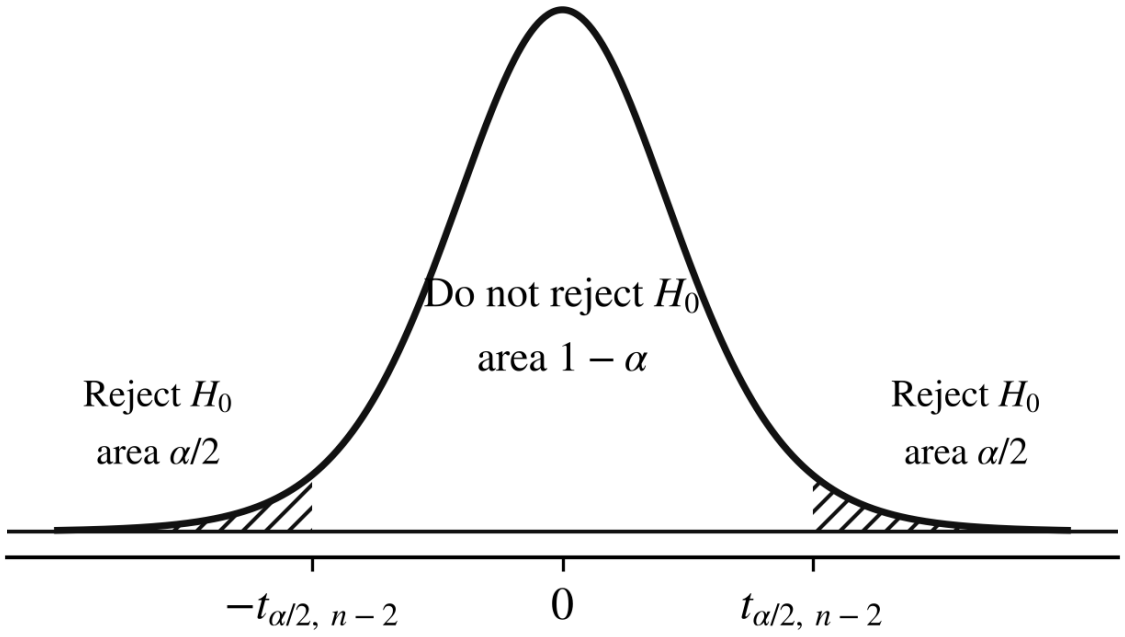
2 Test statistic

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim \mathrm{t}(\mathrm{n}-2)$$

Under H_0 the statistic t follows a t distribution with **n – 2** degrees of freedom

Test for ρ: Steps (2)

3 Specify the Critical region



Note

Testing $\rho = 0$ is the same as testing $\beta_1 = 0$ in simple linear regression

4 Decision

Reject H_0 if $\mathbf{|t| > t_{\alpha/2,\, n-2}}$, otherwise do not reject H_0. Or reject H_0 if the **p-value < α**

Conclusion if we reject H_0

There is enough evidence of a significant linear relationship between X and Y because ρ is different from 0.

Example 3 (a): Test for Example 1

Test the significance of the correlation coefficient at the 5% level of significance for the data of Example 1

1 Hypotheses

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

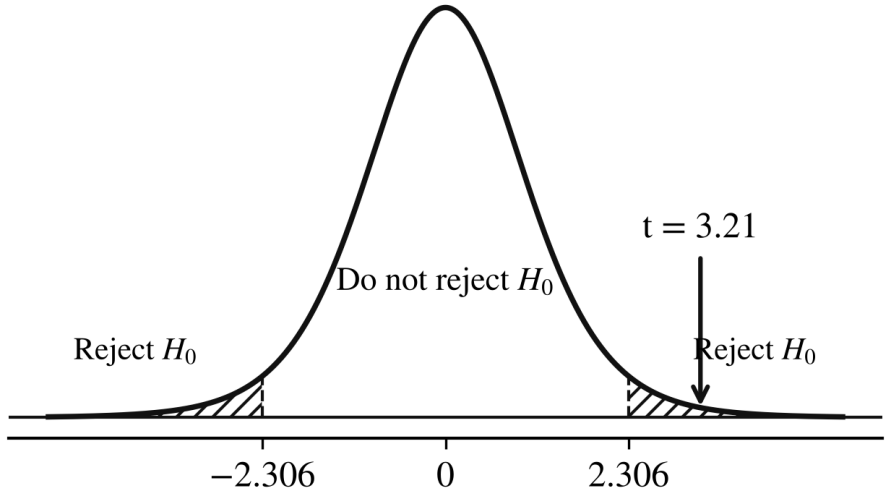
2 Test statistic

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.7506\sqrt{10-2}}{\sqrt{1-(0.7506)^2}} \approx 3.21$$

3 Critical value

Degrees of freedom = n − 2 = 8

$t_{0.025,\,8} = 2.306$



Decision

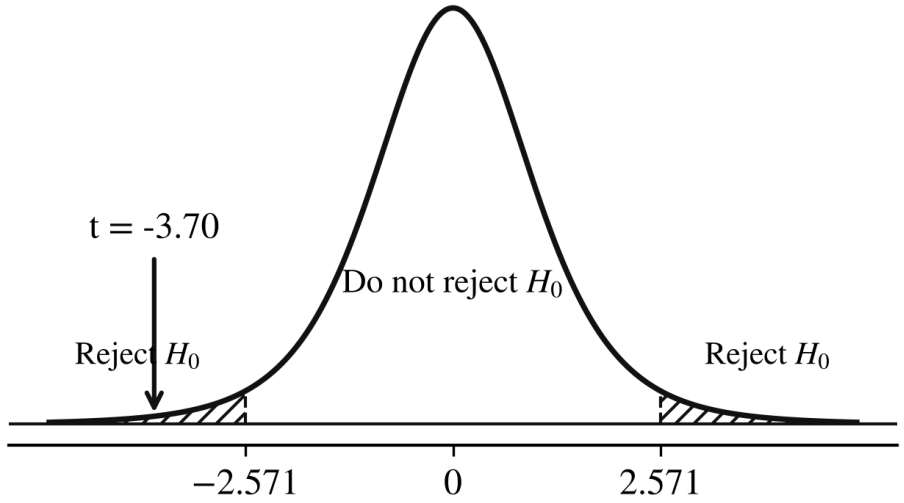
|t| = 3.21 > 2.306 so we **reject** H_0.
The p-value ≈ 0.012 is less than 0.05

Conclusion

There is a significant linear relationship between the midterm mark and the final mark

Example 3 (b): Test for Example 2

Test the significance of the correlation coefficient at the 5% level of significance for the data of Example 2



1 Hypotheses

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

2 Test statistic

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{-0.856\sqrt{7-2}}{\sqrt{1-(-0.856)^2}} \approx -3.7$$

3 Critical value

Degrees of freedom = n − 2 = 5

$t_{0.025,\,5} = 2.571$

Decision

$|t| = 3.7 > 2.571$ so we **reject** H_0. The p-value ≈ 0.014 is less than 0.05

Conclusion

There is a significant linear relationship between the number of absences and the final grade.

3

Simple Linear Regression Model

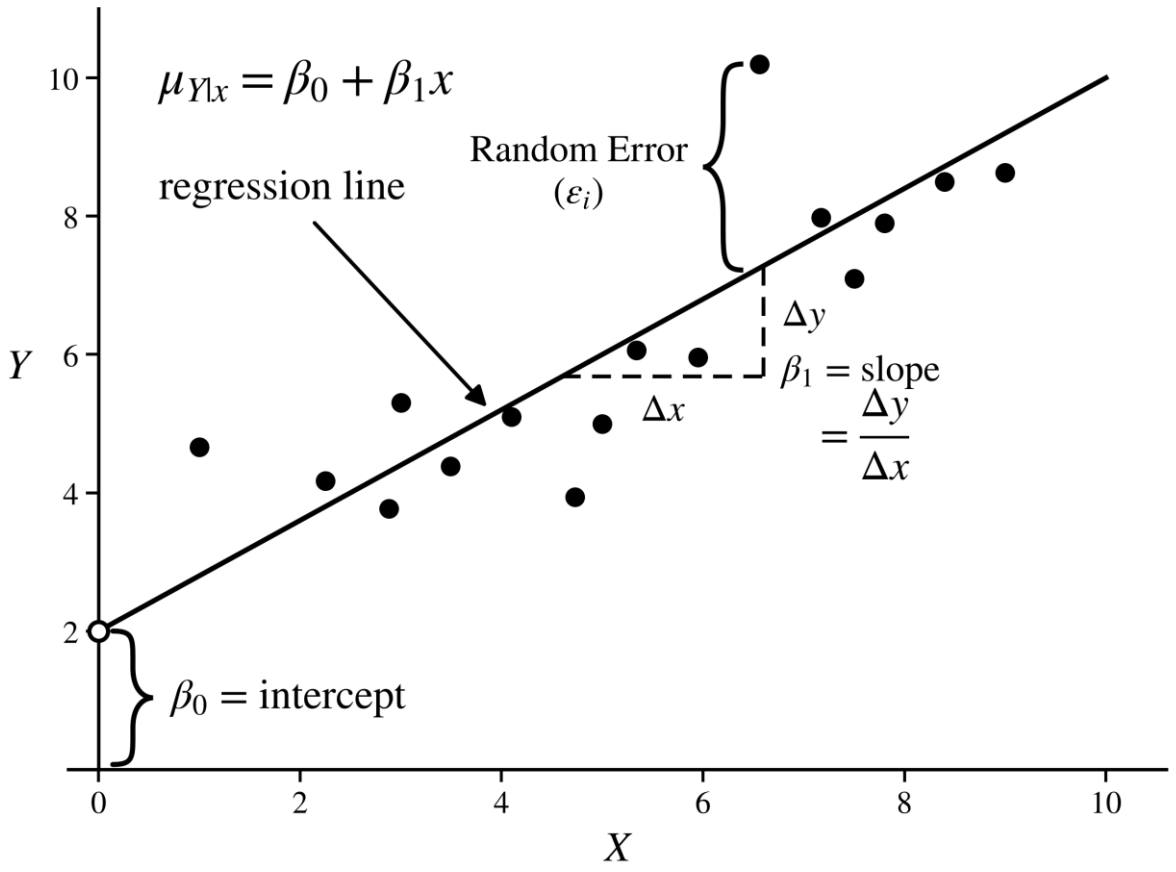
The model · its assumptions · the fitted line

The Simple Linear Regression Model

The model describes the linear relationship between X and Y is given by the following regression line

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

- $\boldsymbol{X_i}$ is the independent variable (predictor or explanatory variable)
- $\boldsymbol{Y_i}$ is the dependent variable (response variable)
- $\boldsymbol{\beta_0}$ is the intercept and $\boldsymbol{\beta_1}$ is the slope. They are the unknown parameters
- $\boldsymbol{\varepsilon_i}$ is the random error term. It makes the analysis probabilistic
- $\boldsymbol{\beta_0 + \beta_1 X_i}$ is the constant part .



Assumptions of the Model

1 The errors

The errors $\boldsymbol{\varepsilon_i}$ are independent and normally distributed $\varepsilon_i \sim N(0, \sigma^2)$. The last point states that the random errors are independent (uncorrelated).

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

2 The response

So the response Y_i is also normal $Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$

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

X_i is a known constant. Only Y_i is random

The Fitted Regression Line

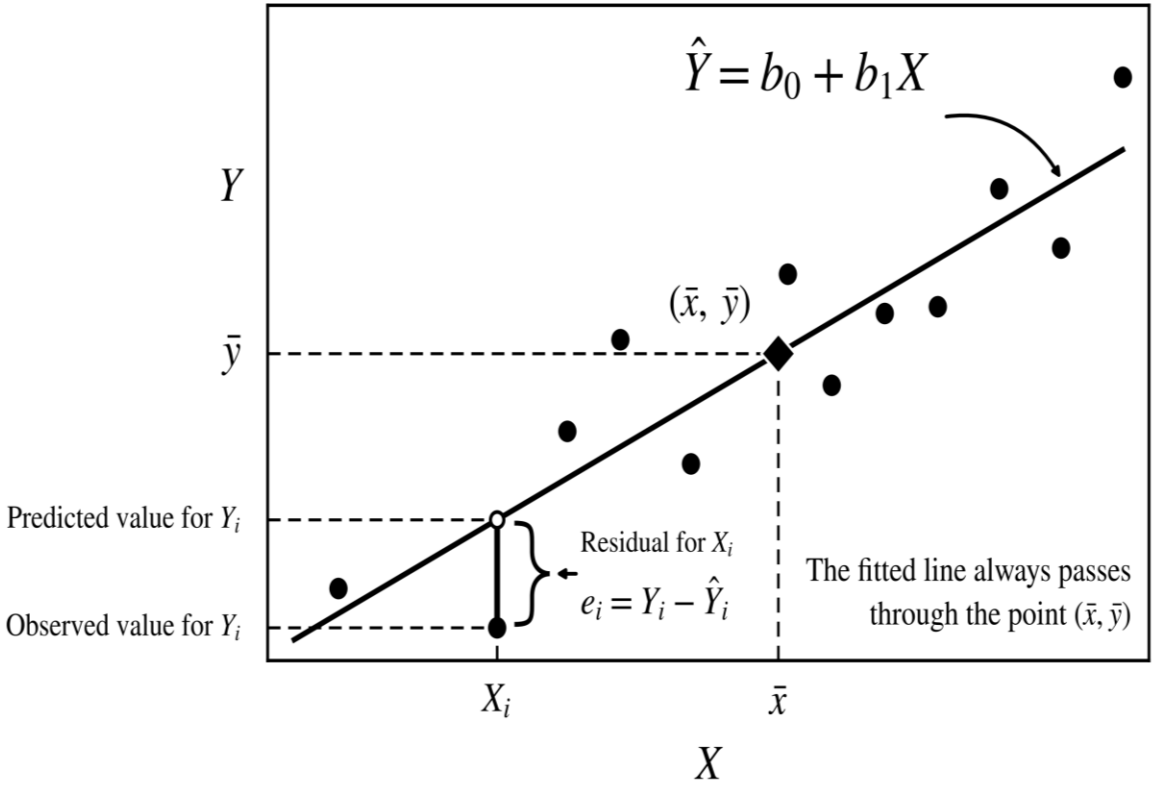
We estimate $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$ by $\mathbf{b_0}$ and $\mathbf{b_1}$ using the sample.
The **fitted line** (prediction equation) is

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

The **residual** is the vertical distance between the observed value y_i and their fitted values $\hat{y}_i$ (from a point to the line) and denoted as $\boldsymbol{e_i}$

$$e_i = y_i - \hat{y}_i, \quad i = 1, \ldots, n$$

The fitted line is an estimate of the true regression line



4

Least Squares Method

Estimates · fitted line properties · error variance

The Method of Least Squares

We find $\mathbf{b_0}$ and $\mathbf{b_1}$ so that the **sum of the squares of the residuals is as small as possible.** This is the method of least squares. We minimize

$$\text{SSE} = \sum_{i=1}^{n} {e_i}^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

Name

SSE is called the **error sum of squares**

Idea

The least squares line is the line that is closest to all points in this sense

Least Squares Estimates

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{S_{xy}}{S_{xx}}$$

$$b_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{xy}}{\sum_{i=1}^{n} {x_i}^2 - n\overline{x}^2} = r\sqrt{\frac{S_{yy}}{S_{xx}}}$$

$$b_0 = \overline{y} - b_1\overline{x}$$

Note

The fitted line always passes through the point (x̄, ȳ)

Example 4

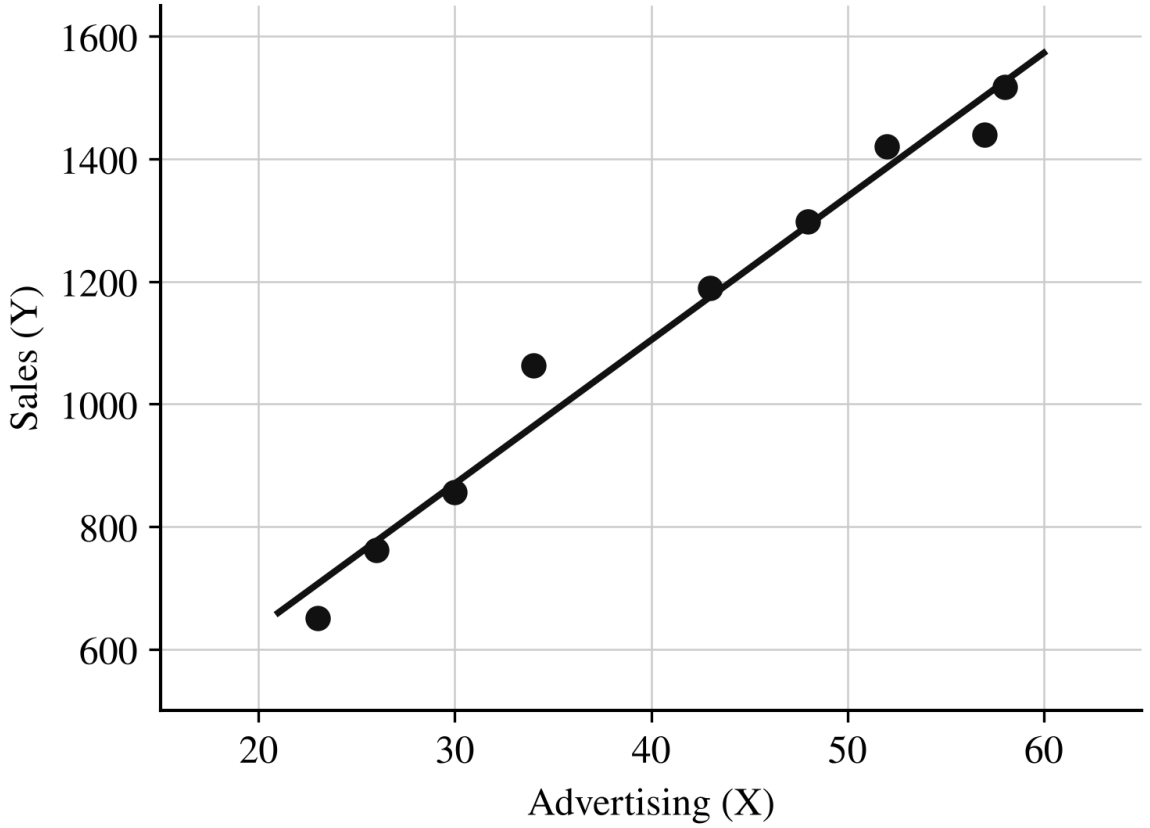
The table shows data from the early days of a clothing company. Each column shows the company sales for a year (Y) and the amount spent on advertising in that year (X). Both are in millions of dollars

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 squares estimates of the simple linear regression model and interpret the result

Example 4: Solution



The scatter diagram shows a clear **linear** relationship between sales and advertising.

The correlation coefficient between the Advertising X and Sales Y is given by

$$r = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} = \frac{33671.56}{\sqrt{1437.56}\sqrt{807485.6}} = 0.988$$

A very strong positive correlation.

From the data we have

$$\overline{x} = 41.22, \quad \overline{y} = 1133.22, \quad n = 9$$

$$\sum_{i=1}^{9} x_i^2 = 16731$$

$$\sum_{i=1}^{9} y_i^2 = 12365219$$

$$\sum_{i=1}^{9} x_i y_i = 454097$$

Example 4: Solution

Sums of squares and products

$$S_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n} {x_i}^2 - n{\overline{x}}^2$$

$$S_{xx} = \sum_{i=1}^{9} x_i^2 - 9\times\overline{x}^2 = 16731 - 9\times 41.22^2 = 1437.556$$

$$S_{yy} = \sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n} {y_i}^2 - n{\overline{y}}^2$$

$$S_{yy} = \sum_{i=1}^{9} y_i^2 - 9\times\overline{y}^2 = 12365219 - 9\times 1133.22^2 = 807485.6$$

$$S_{xy} = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y}$$

$$S_{xy} = \sum_{i=1}^{9} x_i y_i - 9\times\overline{xy} = 454097 - 9\times 41.22\times 1133.22 = 33671.56$$

The least-squares line

$$b_1 = \frac{S_{xy}}{S_{xx}} = \frac{33671.56}{1437.556} = 23.42$$

$$b_0 = \overline{y} - b_1\overline{x}$$

$$= 1133.22 - 23.42\times 41.22 = 167.68$$

$$\boldsymbol{\hat{y} = 167.68 + 23.42x}$$

Interpretation

Each one million dollar increase in advertising is expected to raise sales by 23.4 million dollars. With no advertising, expected sales are 167.7 million dollars.

Example 4: Solution (Other way to calculate b_1)

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 increases by one million dollars, sales are expected to increase by 23.4 million dollars, and if there is no advertising, we expect sales of 167.7 million dollars.

Properties of the Fitted Line

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

1. $\Sigma e_i = 0$
   The sum of the residuals is zero

2. $\Sigma y_i = \Sigma \hat{y}_i$
   The sum of the observed Y equals the sum of the fitted Y

3. $\Sigma x_i e_i = 0$
   The residuals weighted by X add to zero

4. $\Sigma \hat{y}_i e_i = 0$
   The residuals weighted by the fitted values add to zero

5. **Point $(\bar{x}, \bar{y})$**
   The regression line goes through the point of the means

Estimation of the Error Variance σ^2

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

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2} = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{\text{SSE}}{n-2} = \text{MSE}$$

The standard deviation of the errors (in the units of the data) is

$$s = \sqrt{\text{MSE}}$$

$$\text{SSE} = S_{yy} - b_1 S_{xy}$$

Why n – 2

We divide by n – 2 because two parameters (b_0 and b_1) are estimated from the data

Example 4: Error Variance

X	Y	$\hat{y}$	e	e^2
23	651	706.41	-55.41	3069.94
26	762	776.68	-14.68	215.37
30	856	870.37	-14.37	206.40
34	1063	964.06	98.94	9789.59
43	1190	1174.86	15.14	229.14
48	1298	1291.98	6.02	36.28
52	1421	1385.67	35.33	1248.36
57	1440	1502.78	-62.78	3941.55
58	1518	1526.20	-8.20	67.31
Sum			**0.00**	**18804**

$$e_i = y_i - \hat{y}_i$$

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

$$s^2 = MSE = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{18804}{9-2} = 2686.28$$

$$s = \sqrt{MSE} = \sqrt{2686.28} \approx 51.83$$

Check

$SSE = S_{yy} - b_1 S_{xy}$

$= 807485.6 - 23.4228 \times 33671.56$

$= 18804$

We get the same value

5

Inference on the Coefficients

Standard errors · confidence intervals · t tests

Other way of writing b_0 and b_1

The estimates are linear functions of the y_i

$$b_1 = \sum_{i=1}^{n} \boldsymbol{y_i} \frac{(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sum_{\mathbf{i=1}}^{\boldsymbol{n}} \boldsymbol{K_i y_i}$$

$$\boldsymbol{K_i = \frac{(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{x_i - \bar{x}}{S_{xx}}}$$

$$\sum_{i=1}^{n} K_i = 0$$

$$\sum_{i=1}^{n} K_i x_i = 1$$

$$\sum_{i=1}^{n} {K_i}^2 = \frac{1}{S_{xx}}$$

$$b_0 = \sum_{i=1}^{n} \left(\frac{1}{n} - \bar{x}K_i\right) y_i = \sum_{\mathbf{i=1}}^{\boldsymbol{n}} \boldsymbol{L_i y_i}$$

$$\boldsymbol{L_i = \frac{1}{n} - \bar{x}K_i}$$

$$\sum_{i=1}^{n} L_i = 1$$

$$\sum_{i=1}^{n} L_i x_i = 0$$

$$\sum_{i=1}^{n} {L_i}^2 = \frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}$$

K_i and L_i are constants. They satisfy these properties.

Mean and Variance of the Estimates

1 Unbiased estimates

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

2 Variances

$$\mathrm{Var}(b_1) = \frac{\sigma^2}{S_{xx}} \approx \frac{\mathrm{MSE}}{S_{xx}}$$

$$\mathrm{Var}(b_0) \approx \mathrm{MSE}\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)$$

3 Standard errors

$$\mathrm{S.E}(b_1) = \sqrt{\frac{\mathrm{MSE}}{S_{xx}}}$$

$$\mathrm{S.E}(b_0) = \sqrt{\mathrm{MSE}\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)}$$

Example 5

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

1 Slope b_1

$$MSE = \frac{\text{SSE}}{n-2} = 2686.28$$

$$\text{Var}(b_1) = \frac{\sigma^2}{S_{xx}} \approx \frac{\text{MSE}}{S_{xx}} = \frac{2686.28}{1437.56} \approx 1.87$$

$$\text{S. E}(b_1) = \sqrt{1.87} \approx 1.37$$

2 Intercept b_0

$$\text{Var}(b_0) = \text{MSE}\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)$$

$$= 2686.28 \times \left(\frac{1}{9} + \frac{41.22^2}{1437.56}\right)$$

$$\approx 3473.8$$

$$\text{S. E}(b_0) = \sqrt{3473.8} \approx 58.94$$

Inference for the Regression Model

In this section, we discuss some statistical inferences related to the simple linear regression model, such as constructing confidence intervals for the model coefficients and hypothesis 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 coefficients, we need to present the following lemmas.

Lemma (Sampling distributions)

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

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

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

Confidence Intervals for β_1 and β_0

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

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

and

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

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

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

Example 6

For the data of Example 4 find a 95% confidence interval for β_1 and for β_0.

Here S.E(b_1) = 1.37, S.E(b_0) = 58.94 and $t_{0.025,\,7}$ = 2.365

1 Interval for β_1

$$b_1 \mp t_{\alpha/2,n-2} \times \text{S. E}(b_1)$$

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

$$20.2 < \beta_1 < 26.7$$

Interpretation

We are 95% confident that each one million dollar increase in advertising raises expected sales by between 20.2 and 26.7 million dollars.

2 Interval for β_0

$$b_0 - t_{\alpha/2,n-2} \times \text{S. E}(b_0)$$

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

$$28.3 < \beta_0 < 307.1$$

Interpretation

With no advertising, we are 95% confident that the true mean sales are between 28.3 and 307.1 million dollars.

Hypothesis Tests for the Coefficients

The sampling distributions of $\boldsymbol{T_1}$ and $\boldsymbol{T_0}$ 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.

1 Hypotheses

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

2 Test statistic

$$T_i = \frac{b_i - \beta_i^{(0)}}{\text{S. E}(b_i)} \sim t_{n-2}$$

3 Critical region

Reject H_0 if $\mathbf{|T_i| > t_{\alpha/2,\, n-2}}$

Use the same two tail regions as in the test for ρ

4 Decision

If T_i is in the critical region we reject H_0

Otherwise, we do not reject H_0.

Here i = 0 or 1

Remarks on the Tests

In some applications, we may need to test

1 **Test $H_0: \beta_i = 0 \; vs \; H_1: \beta_i \neq 0$**

Replace $\beta_i^{(0)}$ by zero .

2 **One-sided tests**

For $H_0 : \beta_i < 0 \quad vs \quad H_1 : \beta_i > 0$

or $H_0 : \beta_i > 0 \quad vs \quad H_1 : \beta_i < 0$

use a one-sided critical region.

3 **p-value approach**

Two-sided p-value = $2P(T > |T_i|)$.

One-sided p-value = $P(T > |T_i|)$.

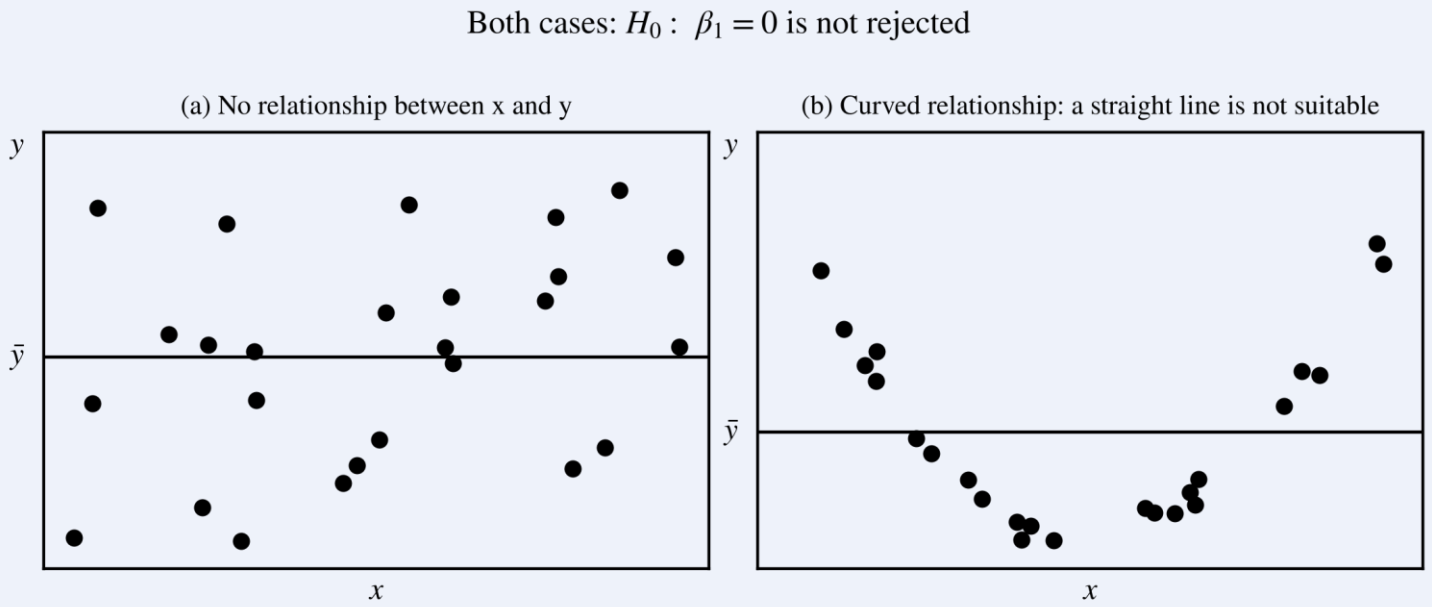
Reject H_0 if the **p-value < α**

4 **Meaning for the slope**

If we reject $H_0 : \beta_1 = 0$ there is a significant linear regression. If we do not reject there is no significant linear relationship between E(Y) and x.

Failing to Reject H_0: $\beta_1 = 0$

Failing to reject H_0: $\beta_1 = 0$ means there is no significant linear relationship between Y and x. There are two possible reasons:



(a) Changing x has little impact on changes in Y.

(b) The true relationship is nonlinear (curved), so a straight line cannot show it.

Source: Walpole, Section 11.5, Figure 11.8

Example 7: Testing the Regression Coefficients

Example 7

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

$$H_0: \beta_1 = 0 \quad \text{vs} \quad H_1: \beta_1 \neq 0$$

and

$$H_0: \beta_0 = 0 \quad \text{vs} \quad H_1: \beta_0 \neq 0$$

Solution

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

$$H_0: \beta_1 = 0 \quad \text{vs} \quad H_1: \beta_1 \neq 0$$

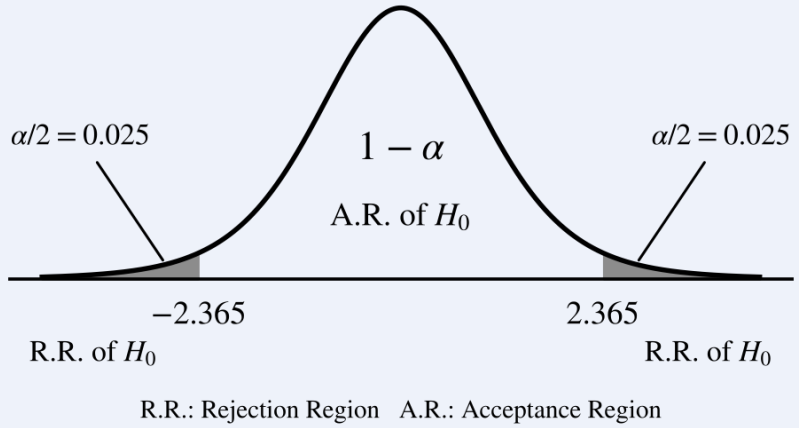
Test statistic under H_0 is given by

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

Example 7: Testing Slope (β_1)

The critical regions are given by

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



R.R.: Rejection Region 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. Also, the

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

then reject H_0. So the slope β_1 is significantly different from 0.

Example 7: Testing Intercept (β_0)

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

$$H_0: \beta_0 = 0 \quad \text{vs} \quad H_1: \beta_0 \neq 0$$

Test statistic under H_0 is given by

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

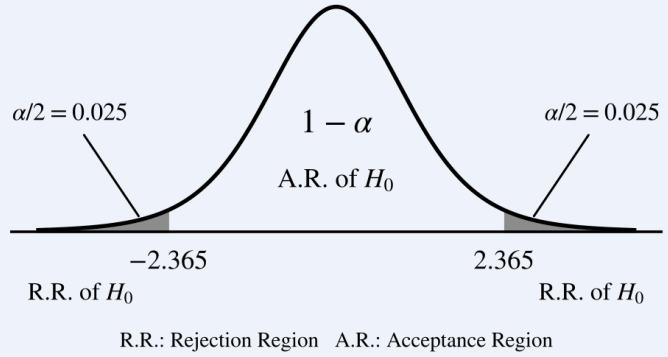
The critical regions are given by

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

Decision: The calculated test $T_0 = 2.85$ belongs to the shaded areas, then we reject the null hypothesis H_0.
Also, the $p - value = 2P(T > |T_0|)$

$$= 2P(T > |2.85|) \approx 0.025 < 0.05$$

then reject H_0. So the intercept β_0 is significant.



6

Coefficient of Determination R^2

How well does the line fit the data?

Coefficient of Determination R^2

In Simple linear regression ,the coefficient of determination can be obtained by squaring the Pearson correlation coefficient: $\mathbf{R^2 = r^2}$.

The coefficient of determination, R2, represents the **proportion of the total sample variation in Y** (measured by the sum of squares of deviations of the sample $\boldsymbol{y}$ values about their mean $\overline{\boldsymbol{y}}$) **that is explained by** (or attributed to) the linear relationship between X and Y.

Another way to calculate the coefficient of determination uses three sums of squares, as follows:

Total sum of squares

$$\text{SST} = \sum_{i=1}^{n} (y_i - \overline{y})^2$$

The total variation in the observed y values.

Regression sum of squares

$$\text{SSR} = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2$$

The variation in y that is explained by the model.

Error sum of squares

$$\text{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

The variation in y that is left unexplained by the model.

$$\text{SST} = \text{SSR} + \text{SSE}$$

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

Properties of the Coefficient of Determination R^2

1 $0 \le R^2 \le 1$

The coefficient of determination is a number between 0 and 1

2 $R^2 = 0$

The least squares regression line has no explanatory value.

3 $R^2 = 1$

The regression line explains 100% of the variation in Y. All residuals are zero

4 **Correlation**

$r = \pm\sqrt{R^2}$ with the sign of the estimate of the slope b_1.

Careful

A high R^2 does not prove that a straight line is the best model. Always look at the scatter diagram too.

Example 8

Calculate the coefficient of determination of the regression in Example 4 and interpret it the results. Also find the Pearson correlation coefficient.

Solution

$$\text{SST} = S_{yy} = 807485.6$$

$$\text{SSE} = 18803.9$$

$$\text{SSR} = \text{SST} - \text{SSE} = 788681.6$$

The coefficient of determination:

$$R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{788681.6}{807485.6} \approx 0.9767$$

The correlation coefficient

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

Interpretation

About **97.7%** of the total variation in sales is explained by advertising. The correlation is positive because the slope b_1 is positive.

Summary of the Chapter

Topic	Formula or rule
Correlation	$r = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}, \quad -1 \le r \le 1$
Test for ρ	$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}, \quad df = n-2$
Fitted line	$\hat{y} = b_0 + b_1 x, \quad b_1 = \frac{S_{xy}}{S_{xx}} = r\sqrt{\frac{S_{yy}}{S_{xx}}}, \quad b_0 = \overline{y} - b_1\overline{x}$
Error variance	$s^2 = \frac{\mathrm{SSE}}{n-2} = \mathrm{MSE}, \qquad s = \sqrt{\mathrm{MSE}}$
Standard errors	$\mathrm{S.E}(b_1) = \sqrt{\frac{\mathrm{MSE}}{S_{xx}}}, \quad \mathrm{S.E}(b_0) = \sqrt{\mathrm{MSE}\left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right)}$
Confidence interval	$b_i \pm t_{\alpha/2,n-2} \times \mathrm{S.E}(b_i)$
Test for βi	$T = \frac{b_i - \beta_i^{(0)}}{\mathrm{S.E}(b_i)} \sim t_{n-2}$
Fit of the model	$R^2 = \frac{\mathrm{SSR}}{\mathrm{SST}} = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}} = r^2$