

1A/B Least Squares Linear Regression

A least squares regression line is used to:

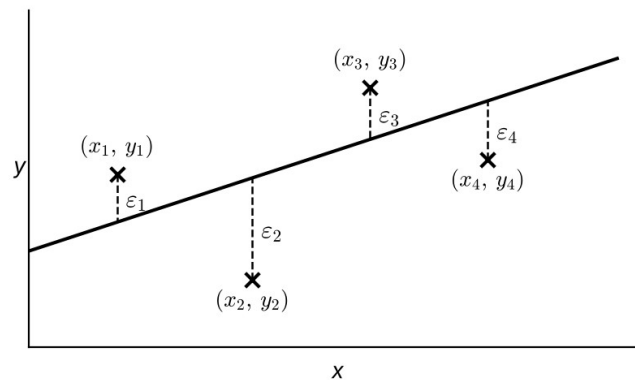
The explanatory (independent) variable is:

The response (dependent) variable is:

Interpolation:

Extrapolation:

The residual of a data point is:



The least squares regression line of y on x is the line that minimises:

The equation of the regression line of y on x is $y = a + bx$, where

$b =$

$a =$

Summary statistics:

$S_{xy} =$

$S_{xx} =$

$S_{yy} =$

1. The results from an experiment in which different masses were placed on a spring and the resulting length of the spring measured, are shown below.

Mass, x (kg)	20	40	60	80	100
Length, y (cm)	48	55.1	56.3	61.2	68

- a) Calculate S_{xx} and S_{xy} .

(You may use $\Sigma x = 300$, $\Sigma x^2 = 22\,000$, $\bar{x} = 60$, $\Sigma xy = 18\,238$, $\Sigma y^2 = 16\,879.14$, $\Sigma y = 288.6$, $\bar{y} = 57.72$)

- b) Calculate the regression line of y on x.

- c) Use your equation to predict the length of the spring when the applied mass is:

i) 58 kg

ii) 130 kg

- d) Comment on the reliability of your predictions.

2. A scientist working in agricultural research believes that there is a linear relationship between the amount of a certain food supplement given to hens and the hardness of the shells of the eggs they lay. As an experiment, controlled quantities of the supplement were added to the hens' normal diet for a period of two weeks and the hardness of the shells of the eggs laid at the end of this period was then measured on a scale from 1 to 10, with the following results:

Food supplement, f (g/day)	2	4	6	8	10	12	14
Hardness of shells, h	3.2	5.2	5.5	6.4	7.2	8.5	9.8

- a) Find the equation of the regression line of h on f.

(You may use $\Sigma f = 56$, $\Sigma h = 45.8$, $\bar{f} = 8$, $\bar{h} = 6.543$, $\Sigma f^2 = 560$, $\Sigma fh = 422.6$)

- b) Interpret what the values of a and b tell you.

3. A repair workshop finds it is having a problem with a pressure gauge it uses. It decides to have it checked by a specialist firm. The following data was obtained.

Gauge reading, x (bars)	1.0	1.4	1.8	2.2	2.6	3.0	3.4	3.8
Correct reading, y (bars)	0.96	1.33	1.75	2.14	2.58	2.97	3.38	3.75

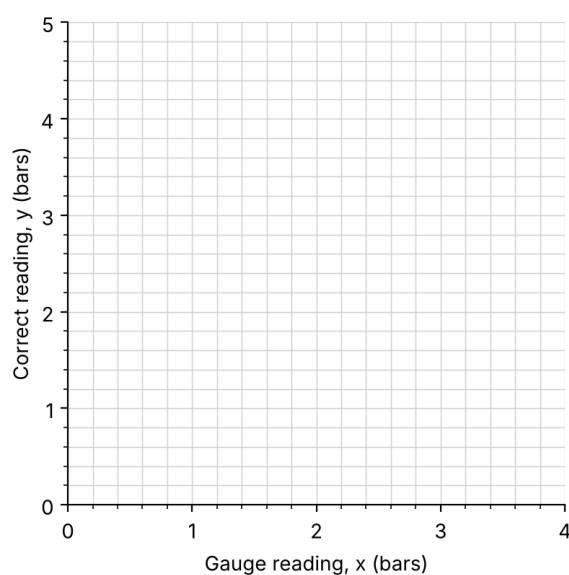
(You may use $\Sigma x = 19.2$, $\Sigma x^2 = 52.8$, $\Sigma y = 18.86$, $\Sigma y^2 = 51.30$, $\Sigma xy = 52.04$)

- a) Show that $S_{xy} = 6.776$ and find S_{xx} .

It is thought that a linear relationship of the form $y = a + bx$ could be used to describe these data.

- b) Use linear regression to find the values of a and b , giving your answers to 4 decimal places.

- c) Draw a scatter diagram to represent these data and draw the regression line on your diagram.



- d) The gauge shows a reading of 2 bar. Using the regression equation, work out what the correct reading should be.

Coding: to find the original regression line from a coded one:

4. Eight samples of carbon steel were produced with different percentages, $c\%$, of carbon in them. Each sample was heated in a furnace until it melted and the temperature, m in $^{\circ}\text{C}$, at which it melted was recorded.

The results were coded such that $x = 10c$ and $y = \frac{m - 700}{5}$. The coded results are shown in the table.

Percentage of carbon, x	1	2	3	4	5	6	7	8
Melting point, y	35	28	24	16	15	12	8	6

- a) Calculate S_{xy} and S_{xx} . (You may use $\Sigma x^2 = 204$ and $\Sigma xy = 478$.)
- b) Find the regression line of y on x .
- c) Estimate the melting point of carbon steel which contains 0.25% carbon.

1C Residuals

If a set of bivariate data has regression equation $y = a + bx$, then the residual of the data point (x_i, y_i) is:

The sum of the residuals of all the data points is:

A residual is negative when:

A residual plot shows:

If the residuals are randomly scattered about zero:

Residual patterns which suggest a linear model is not suitable:

Residuals can also be used to identify outliers:

1. The table shows the relationship between the temperature, t °C, and the sales of ice cream, s , on five days in June:

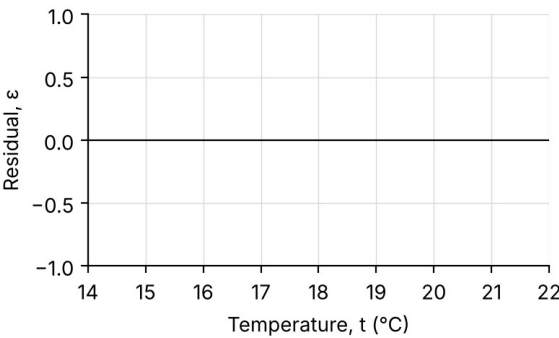
Temp, t (°C)	15	16	18	19	21
Sales, s (100s)	12.0	15.0	17.5	p	24.0

The equation of the regression line of s on t is given as $s = -17.154 + 1.9693t$.

- a) Calculate the residuals for the given regression line and hence find the value of p .

t	s	$s = -17.154 + 1.9693t$	ϵ
15	12.0		
16	15.0		
18	17.5		
19	p		
21	24.0		

b) By considering the residuals, comment on whether a linear regression model is suitable for these data.



2. The table shows the time taken, t minutes, to produce y litres of paint in a factory.

t	2.1	3.7	4.8	6.1	7.2
y	19.2	27.3	26.9	38.5	40.9

The regression line of y on t is given as $y = 9.7603 + 4.3514t$. One of the y -values was incorrectly recorded.

a) Use residuals to identify the incorrect value.

t	y	$y = 9.7603 + 4.3514t$	ϵ
2.1	19.2		
3.7	27.3		
4.8	26.9		
6.1	38.5		
7.2	40.9		

b) Comment on the validity of ignoring this outlier in your analysis.

c) Ignoring the outlier, produce a new model.

d) Use the new model to estimate the amount of paint that is produced in 4.8 minutes.

The residual sum of squares (RSS) is:

Formula for the RSS:

RSS =

The smaller the value of the RSS:

The linear regression model is the linear model which:

3. The data shows the sales, in 100s, y , of Slush at a riverside café and the number of hours of sunshine, x , on five random days during August.

x	8	10	11.5	12	12.2
y	7.1	8.2	8.9	9.2	9.5

Given that $\Sigma x = 53.7$, $\Sigma y = 42.9$, $\Sigma x^2 = 589.09$, $\Sigma y^2 = 371.75$, $\Sigma xy = 467.45$

- a) calculate the residual sum of squares (RSS).

The RSS for five random days in December is 0.0562.

- b) State, with a reason, which month is more likely to have a linear fit between the number of hours of sunshine and the sales of Slush.