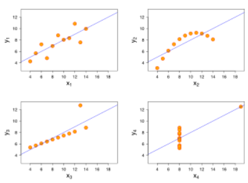
MDM4U: Mathematics of Data Management

How Can We Model Data?

Linear Regression

J. Garvin



Regression

In mathematics, *regression* is a tool that allows us to make predictions about the values one variable based on the values of another.

One of the simplest forms of regression is *linear regression*, which produces an equation describing the *line of best fit* of a data set.

Least-Squares Regression

One technique for determining the line of best fit is called *least-squares* regression.

The distance from a datum to its line of best fit is called a *residual*.

Least-squares regression ensures two criteria:

1. The sum of the residuals is zero.
2. The sum of the squares of the residuals is a minimum.

Line of Best Fit Using Least-Squares Regression

Using the method of least-squares regression, the line of best fit for the variables x and y is given by $y = ax + b$, where $a = \dfrac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2}$ and $b = \overline{y} - a\overline{x}$.

Least-Squares Regression

Example

As an experiment, eight individuals are asked to press a button when a light turns on. Their ages (in years), and their response-times (in ms), are recorded below. Calculate the line of best fit, and display the data using a scatter plot.

Age	24	28	33	41	42	59	73	76
Time	250	280	350	420	470	820	1020	1050

To find the value of a, use a table with columns for x, y, xy and x^2, where age is the independent variable (x) and response-time the dependent variable (y).

Least-Squares Regression

x	y	xy	x^2
24	250	6000	576
28	280	7840	784
33	350	11550	1089
41	420	17220	1681
42	470	19740	1764
59	820	48380	3481
73	1020	74460	5329
76	1050	79800	5776
376	4660	264990	20480

$$a = \frac{8 \times 264990 - 376 \times 4660}{8 \times 20480 - 376^2} = \frac{22985}{1404} \approx 16.371$$

Least-Squares Regression

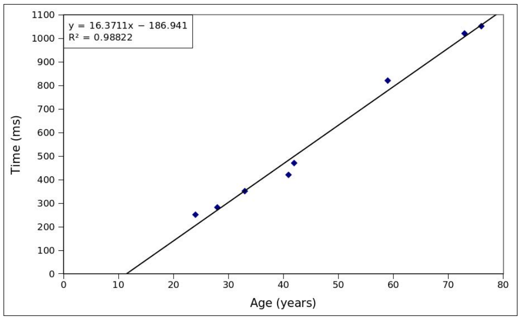
To find b, we need the values of $\overline{x}$ and $\overline{y}$.

$\overline{x} = \dfrac{376}{8} = 47$, $\overline{y} = \dfrac{4660}{8} = \dfrac{1165}{2}$.

$b = \dfrac{1165}{2} - \dfrac{22985}{1404} \times 47 = -\dfrac{262465}{1404} \approx -186.94$.

The equation of the line of best fit is $y = \dfrac{22985}{1404}x - \dfrac{262465}{1404}$, or $y \approx 16.371x - 186.94$.

Least-Squares Regression



Least-Squares Regression

Example

Calculate the line of best fit for the data below, graph it, and classify the correlation.

Time (s)	2	4	4	5	8	13
Distance (m)	10	16	15	16	21	27

x	y	xy	x^2
2	10	20	4
4	16	64	16
4	15	60	16
5	16	80	25
8	21	168	64
13	27	351	169
36	105	743	294

Least-Squares Regression

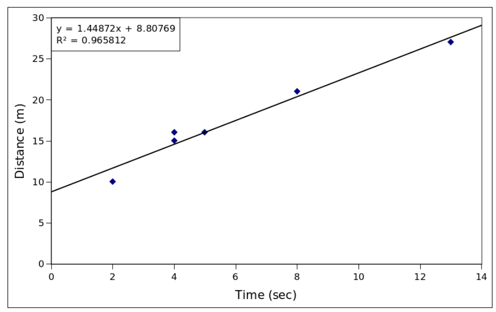
$$a = \frac{6 \times 743 - 36 \times 105}{6 \times 294 - 36^2} = \frac{113}{78} \approx 1.45$$

$$\overline{x} = \frac{36}{6} = 6,\ \overline{y} = \frac{105}{6} = \frac{35}{2}.$$

$$b = \frac{35}{2} - \frac{113}{78} \times 6 = \frac{229}{26} \approx 8.8.$$

Therefore, the line of best fit is $y = \frac{113}{78}x + \frac{229}{26}$, or $y \approx 1.45x + 8.8$.

Least-Squares Regression



Since the data are close to the line of best fit, and the trend line is increasing, the data illustrate a strong, positive correlation.

Effects of Outliers

Recall that a datum that deviates far from the other data is an *outlier*.

Since least-squares regression minimizes the squares of the residuals for *all* data, all outliers affect the calculation of the line of best fit.

Removing outliers can have a large impact on both the line of best fit, and the correlation coefficient.

Effects of Outliers

Example

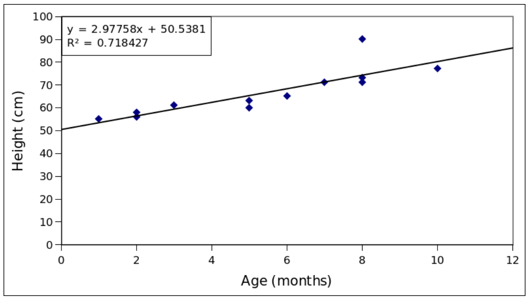
The data below summarizes the heights of twelve babies. Calculate the correlation coefficient and determine the equation of the line of best fit using least-squares regression. Use a scatter plot to visualize the data.

Age (months)	1	2	2	3	5	5
Height (cm)	55	58	56	61	63	60

Age (months)	6	7	8	8	8	10
Height (cm)	65	71	71	90	73	77

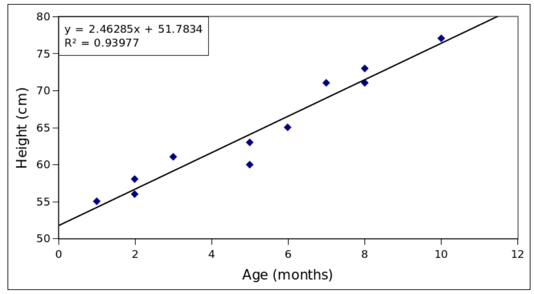
Using a spreadsheet, the equation of the line of best fit is $y \approx 2.98x + 50.54$ and the correlation coefficient is approximately 0.85.

Effects of Outliers



The data point $(8, 90)$ looks like it might be an anomaly in the data. Repeat the analysis with this point removed.

Effects of Outliers



The correlation coefficient changes to 0.97 with the outlier removed, indicating a stronger correlation.

Effects of Outliers

While the previous example illustrates how outliers can affect data analysis, it is important to assess the importance of the data itself.

An outlier cannot be removed simply because it does not fit a mathematical model. There must be some evidence to remove it.

There may be some underlying factor that is affecting the data, manifesting itself in the form of an outlier.

We will talk more about these factors later.

Questions?