1. Least Squares Lines

Background

The formulas for linear least squares fitting were independently derived by German mathematician Johann Carl Friedrich Gauss (1777-1855) and the French mathematician Adrien-Marie Legendre (1752-1833).

Theorem (Least Squares Line Fitting). Given the n data points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, the least squares line $y = a\,x + b$ that fits the points has coefficients a and b given by:

$$a = \frac{\left(\sum_{k=1}^{n} x_k\right) \sum_{k=1}^{n} y_k - n \sum_{k=1}^{n} x_k\, y_k}{\left(\sum_{k=1}^{n} x_k\right)^2 - n \sum_{k=1}^{n} x_k^2}$$

and

$$b = \frac{\left(\sum_{k=1}^{n} x_k\right) \sum_{k=1}^{n} x_k\, y_k - \sum_{k=1}^{n} x_k^2 \sum_{k=1}^{n} y_k}{\left(\sum_{k=1}^{n} x_k\right)^2 - n \sum_{k=1}^{n} x_k^2}\ .$$

Remark. The least squares line is often times called the line of regression.

Example 1. Find the standard "least squares line" $y = a\,x + b$ for the data points $(-1, 10), (0, 9), (1, 7), (2, 5), (3, 4), (4, 3), (5, 0), (6, -1)$.
Use the subroutine **Regression** to find the line. Compare with the line obtained with *Mathematica*'s **Fit** procedure.
Solution 1.

Theorem (Power Fit). Given the n data points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, the power curve $y = a\,x^m$ that fits the points has coefficients a given by:

$$a = \sum_{k=1}^{n} x_k^m\, y_k \Big/ \sum_{k=1}^{n} x_k^{2m}.$$

Remark. The case m = 1 is a line that passes through the origin.

Example 2. Find "modified least squares line" of the form $y = a\,x$ for the data points $(1, 2), (2, 3), (3, 5), (4, 6), (5, 8), (6, 9), (7, 11), (8, 12)$.
Solution 2.

Application to Astronomy

In 1601 the German astronomer Johannes Kepler (1571-1630) formulated the third law of planetary motion $T = C x^{3/2}$, where x is the distance to the sun measured in millions of kilometers, T is the orbital period measured in days, and C is a constant. The observed data pairs for the first four planets: Mercury, Venus, Earth, and Mars are

$(57.59, 87.99), (108.11, 224.70), (149.57, 365.26), (227.84, 686.98)$.

Example 3. Find the power curve $y = a x^{3/2}$ for the data points

$(57.59, 87.99), (108.11, 224.70), (149.57, 365.26), (227.84, 686.98)$.

Solution 3.

Example 1. Find the standard "least squares line" $y = a x + b$ for the data points $(-1, 10), (0, 9), (1, 7), (2, 5), (3, 4), (4, 3), (5, 0), (6, -1)$.

Solution 1.

The abscissas is

$$\{x_k\} = \{-1, 0, 1, 2, 3, 4, 5, 6\}$$

The ordinates is

$$\{y_k\} = \{10, 9, 7, 5, 4, 3, 0, -1\}$$

We can add them up..

$$\mathbf{n = Length[XY]} \quad = 8$$

$$\sum_{k=1}^{n} X_{[[k]]} \quad = 20$$

$$\sum_{k=1}^{n} Y_{[[k]]} \quad = 37$$

If you want to add up their squares then be careful using the power. Sometimes extra parenthesis helps.

$$\sum_{k=1}^{n} (X_{[[k]]})^2 \quad = 92$$

$$\sum_{k=1}^{n} X_{[[k]]} Y_{[[k]]} \quad = 25$$

Let's peek at the linear system that was solved.

$y = b + a x$

The normal equations for finding the coefficients a and b are:

$$\begin{pmatrix} 92 & 20 \\ 20 & 8 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} 25 \\ 37 \end{pmatrix}$$

The solution is

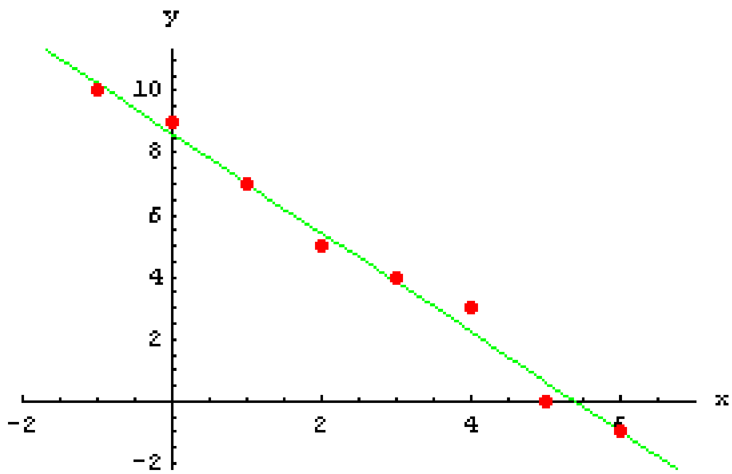
$$\begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} -\frac{45}{28} \\ \frac{121}{14} \end{pmatrix}$$

$$a = -\frac{45}{28}$$
$$b = \frac{121}{14}$$

The least squares line is

$$y = \frac{121}{14} - \frac{45\,x}{28} = 8.64286 - 1.60714\,x$$

Of course we want a graph.



Points = {{-1, 10}, {0, 9}, {1, 7}, {2, 5}, {3, 4}, {4, 3}, {5, 0}, {6, -1}}

The `least squares line` is

$$y = \frac{121}{14} - \frac{45\,x}{28} = 8.64286 - 1.60714\,x$$

The sum of the residual's squared for this example is:

$$\sum_{k=1}^{n} (Y_{[[k]]} - a\,X_{[[k]]} - b)^2 = \frac{39}{28} = 1.39286$$

Example 2. Find "modified least squares line" of the form $y = a\,x$ for the data points $(1, 2), (2, 3), (3, 5), (4, 6), (5, 8), (6, 9), (7, 11), (8, 12)$.

Solution 2.

First, enter the data points. Then form the lists of abscissa's, and ordinates.

$$\{x_k, y_k\}_{k=1}^{n} = \{\{1, 2\}, \{2, 3\}, \{3, 5\}, \{4, 6\}, \{5, 8\}, \{6, 9\}, \{7, 11\}, \{8, 12\}\}$$
$$\{x_k\}_{k=1}^{n} = \{1, 2, 3, 4, 5, 6, 7, 8\}$$
$$\{y_k\}_{k=1}^{n} = \{2, 3, 5, 6, 8, 9, 11, 12\}$$
$$n = 8$$

$$\sum_{k=1}^{n} (X_{[[k]]})^2 = 204$$

$$\sum_{k=1}^{n} X_{[[k]]}\, Y_{[[k]]} = 314$$

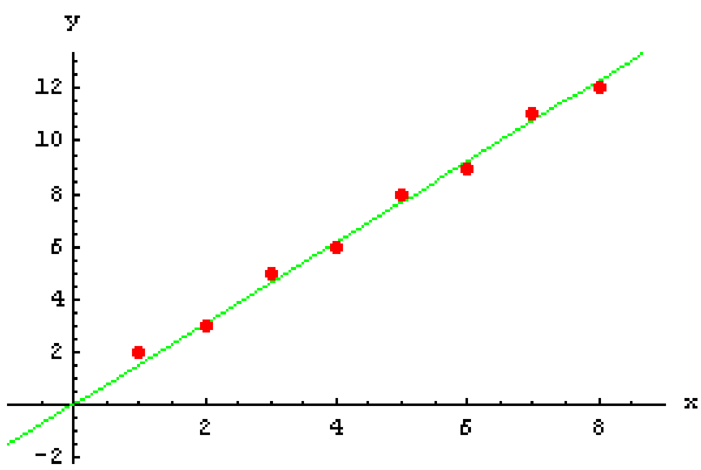
$$a = \sum_{k=1}^{n} X_{[[k]]}\, Y_{[[k]]} \Big/ \sum_{k=1}^{n} (X_{[[k]]})^2 = \frac{157}{102}$$

The "modified least squares line" is

Points = {{1, 2}, {2, 3}, {3, 5}, {4, 6}, {5, 8}, {6, 9}, {7, 11}, {8, 12}}

The `modified least squares line` is

$$y = f[x] = \frac{157\,x}{102}$$
$$y = f[x] = 1.53922\,x$$



Points = {{1, 2}, {2, 3}, {3, 5}, {4, 6}, {5, 8}, {6, 9}, {7, 11}, {8, 12}}

The `modified least squares line` is

$$y = \frac{157\,x}{102} = 1.53922\,x$$

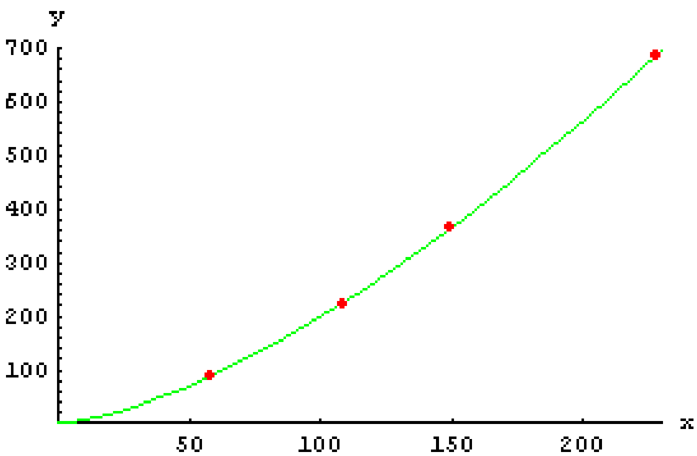
Example 3. Find the power curve $y = a x^{3/2}$ for the data points
$(57.59, 87.99), (108.11, 224.70), (149.57, 365.26), (227.84, 686.98)$.

Solution 3.

Points = {{57.59, 87.99}, {108.11, 224.7}, {149.57, 365.26}, {227.84, 686.98}}

The power curve `Fit` using the formula $y = a\ x^{3/2}$

$y = f[x] = 0.199769\, x^{3/2}$



Points = {{57.59, 87.99}, {108.11, 224.7}, {149.57, 365.26}, {227.84, 686.98}}

The power curve `Fit` using the formula $y = a\ x^{3/2}$

$y = 0.199769\, x^{3/2}$