Least Squares Linear Regression Equations

Dave Rosenman

September 16, 2017

Contents

1 Least Squares Regression Equations **2**
1.1 Best Fit Line 2
1.2 Best Fit Line Through the Origin, (0,0) 3

2 Derivations of m and c **3**
2.1 Best Fit Line 3
2.2 Best Fit Line Through the Origin, (0,0) 5

1 Least Squares Regression Equations

1.1 Best Fit Line

If you have taken n pairs of measurements $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$, the mean value of x is by definition:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

and the mean value of y is

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

The slope of the best fit line, m is given by:

$$m = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

fra

The y-intercept, c, is given by:

$$c = \bar{y} - m\bar{x}$$

The standard error in the slope, Δm, is:

$$\Delta m = \sqrt{\frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\frac{\sum_{i=1}^{n}(y_i - mx_i - c)^2}{n-2}}$$

The standard error in the y intercept, Δc is:

$$\Delta c = \sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)\frac{\sum_{i=1}^{n}(y_i - mx_i - c)^2}{n-2}}$$

The coefficient of determination, r^2 is:

$$r^2 = 1 - \frac{\sum\limits_{i=1}^{n}(y_i - mx_i - c)^2}{\sum\limits_{i=1}^{n}(y_i - \bar{y})^2}$$

1.2 Best Fit Line Through the Origin, (0,0)

If the best fit is required to pass through the origin, $(0, 0)$, then $c = 0$, and

$$m = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

and the standard error of the slope, Δm is:

$$\Delta m = \sqrt{\frac{1}{\sum_{i=1}^{n} x_i^2} \frac{\sum_{i=1}^{n}(y_i - mx_i)^2}{n-1}}$$

The coefficient of determination, r^2, is:

$$r^2 = 1 - \frac{\sum\limits_{i=1}^{n} (y_i - mx_i)^2}{\sum\limits_{i=1}^{n} y_i^2}$$

2 Derivations of m and c

2.1 Best Fit Line

For the least squares method, if we have a set of n measurements $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$, we are looking for the line $y = mx + c$ that minimizes the equation

$$S = \sum_{i=1}^{n} (y_i - mx_i - c)^2$$

(i.e. the line that minimizes the sum of the squared deviations of our y values from the values determined by $y = mx + c$).

So we are looking for the coefficients m and c that minimize the equations above. Thinking of S as $S(m, c)$, S, if S is at a minimum point:

- $\frac{\partial S}{\partial m} = 0$
- $\frac{\partial S}{\partial c} = 0$

$$\frac{\partial S}{\partial m} = -2 \sum_{i=1}^{n} x_i (y_i - mx_i - c) = 0$$

$$\sum_{i=1}^{n} x_i y_i = m \sum_{i=1}^{n} x_i^2 + c \sum_{i=1}^{n} x_i \tag{1}$$

$$\frac{\partial S}{\partial c} = -2 \sum_{i=1}^{n} (y_i - mx_i - c) = 0$$

$$\sum_{i=1}^{n} y_i = m \sum_{i=1}^{n} x_i + cn \tag{2}$$

Note: Equation (2) shows that the best fit line goes through the point $(\hat{x}, \hat{y})$, with $\hat{x}$ being the average value of your measured x coordinates and $\hat{y}$ being the average value of your measured y coordinates. From equation (2),

$$\underbrace{\frac{1}{n}\sum_{i=1}^{n} y_i}_{\bar{y}} = \underbrace{m\frac{1}{n}\sum_{i=1}^{n} x_i}_{m\bar{x}} + c$$

Back to deriving the values for m and c... To find c in terms of m, divide equation (2) by n and simplify:

$$\underbrace{\frac{1}{n}\sum_{i=1}^{n} y_i}_{\bar{y}} + c = m\,\underbrace{\frac{1}{n}\sum_{i=1}^{n} x_i}_{\bar{x}}$$

$$\begin{aligned} \bar{y} &= m\bar{x} + c \\ c &= \bar{y} - m\bar{x} \end{aligned} \tag{3}$$

Plugging in c from equation (3) into equation (1):

$$m\sum_{i=1}^{n} x_i^2 + (\bar{y} - m\bar{x})\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i$$

Factoring out m and simplifying:

$$m\left(\sum_{i=1}^{n} x_i^2 - \bar{x}\sum_{i=1}^{n} x_i\right) + \underbrace{\underbrace{\bar{y}}_{\frac{1}{n}\sum_{i=1}^{n} b y_i} \sum_{i=1}^{n} x_i}_{\sum_{i=1}^{n} y_i \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}\sum_{i=1}^{n} y_i} = \sum_{i=1}^{n} x_i y_i$$

$$m\left(\sum_{i=1}^{n} x_i^2 - \bar{x}\sum_{i=1}^{n} x_i\right) = \sum_{i=1}^{n} x_i y_i - \bar{x}\sum_{i=1}^{n} y_i$$

Solving for m:

$$m = \frac{\sum_{i=1}^{n} (x_i - \bar{x})\, y_i}{\left(\sum_{i=1}^{n} x_i^2 - \bar{x}\sum_{i=1}^{n} x_i\right)} \tag{4}$$

The denominator of equation (4) is equal to

$$\sum_{i=1}^{n} \left(x_i - \bar{x}^2\right)$$

Proof:

$$\sum_{i=1}^{n}(x_i-\bar{x})^2=\sum_{i=1}^{n}x_i^2-2\bar{x}\sum_{i=1}^{n}x_i+\sum_{i=1}^{n}\bar{x}^2 \tag{5}$$

The last term on the right in (5) is:

$$\sum_{i=1}^{n}\bar{x}^2=n\bar{x}^2=\bar{x}\cdot n\bar{x}=\bar{x}\cdot n\left(\frac{1}{n}\sum_{i=1}^{n}x_i\right)=\bar{x}\sum_{i=1}^{n}x_i \tag{6}$$

So from (5) and (6)

$$\sum_{i=1}^{n}(x_i-\bar{x})^2=\sum_{i=1}^{n}x_i^2-2\bar{x}\sum_{i=1}^{n}x_i+\bar{x}\sum_{i=1}^{n}x_i$$

$$\sum_{i=1}^{n}(x_i-\bar{x})^2=\sum_{i=1}^{n}x_i^2-\bar{x}\sum_{i=1}^{n}x_i \tag{7}$$

QED

$$m=\frac{\sum_{i=1}^{n}(x_i-\bar{x})\,y_i}{\sum_{i=1}^{n}(x_i-x)^2} \tag{8}$$

2.2 Best Fit Line Through the Origin, (0,0)

For the best fit line through the origin, $c=0$. From equation (1), $\sum_{i=1}^{n}x_iy_i=m\sum_{i=1}^{n}x_i^2+c\sum_{i=1}^{n}x_i$

$$\sum_{i=1}^{n}x_iy_i=m\sum_{i=1}^{n}x_i^2$$

$$m=\frac{\sum_{i=1}^{n}x_iy_i}{\sum_{i=1}^{n}x_i^2} \tag{9}$$