The Mothod of Least Squares

Satya Mandal, KU

March 26, 2007

The General Problem: *Given a set of data points*

$$(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots, (x_N, y_N),$$

problem is to determine a line $y = mx + c$ that fits this data best according to some criterion. Objective is to use this best fit line to make projections regarding y as a function of x. For example, x may be the price and y the revenue. Once you have some history, you can use the best fit line to make projections how revenue will change with price changes. **In this section, we determine the best fit line according the Least Square method.** We will do this from the statistical point of view.

Suppose

$$(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots, (x_N, y_N)$$

are N data points. We wish to determine the line

$$y = mx + c$$

that fits the data best according to the method of Least Square. We describe the method, as follows:

1. The vertical distance of a data point (x_i, y_i) from the line is

$$D_i = (mx_i + c) - y_i$$

2. The sum of square of these distances:

$$L = L(m,c) = D_1^2 + D_2^2 + \cdots + D_N^2 = \sum_{i=1}^{N} (mx_i + c - y_i)^2.$$

3. The line for which L is least will be called the **least square line,** also the **regration line.**

4. To determine the least square line, we need to optimize (or minimize) L with respect ot m and c. Which will be given by

$$\frac{\partial L}{\partial m} = 0 \qquad and \qquad \frac{\partial L}{\partial c} = 0.$$

   So, we have

$$\frac{\partial L}{\partial m} = \sum 2(mx_i + c - y_i)x_i = 0 \quad and \quad \frac{\partial L}{\partial c} = \sum 2(mx_i + c - y_i) = 0.$$

   Divide both equations by 2 and we get

$$\sum (mx_i + c - y_i)x_i = 0 \qquad Eqn - I$$

   and

$$\sum (mx_i + c - y_i) = 0. \qquad Eqn - II$$

5. We borrow some notations and definations from statistics:

   (a) The means are defines as follows:

$$\overline{x} = \frac{\sum x_i}{N}; \qquad \overline{y} = \frac{\sum y_i}{N}$$

   and

   (b) The variances σ_x^2, σ_y^2 of x and y values are defined as

$$\sigma_x^2 = \frac{\sum (x_i - \overline{x})^2}{N}; \qquad \sigma_y^2 = \frac{\sum (y_i - \overline{y})^2}{N}.$$

   Variance is a measure of variability. If all $x-$ values are same then $\sigma_x^2 = 0$.

(c) The covariance of x abd y is defined as

$$cov(x,y) = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{N}.$$

We use these statistical notations and continue with our optimization as follows.

6. Divide Eqn-II by N and we get

$$m\overline{x} + c - \overline{y} = 0 \qquad or \qquad c = \overline{y} - m\overline{x}.$$

7. The equation (I) is simplified to

$$m\sum x_i^2 + c\sum x_i - \sum x_i y_i = 0 \qquad\qquad Eqn - III.$$

8. Before we proceed, we establish following two identities:

(a)

$$\sum x_i^2 = \sum(x_i - \overline{x})^2 + N\overline{x}^2$$

Proof. We will use the fact that $\sum x_i = N\overline{x}$. We have

$$\sum x_i^2 = \sum[(x_i - \overline{x})^2 + 2x_i\overline{x} - \overline{x}^2] = \sum(x_i - \overline{x})^2 + 2\overline{x}\sum x_i - N\overline{x}^2$$

$$= \sum(x_i - \overline{x})^2 + 2\overline{x}(N\overline{x}) - N\overline{x}^2 = \sum(x_i - \overline{x})^2 + N\overline{x}^2.$$

(b)

$$\sum x_i y_i = \sum(x_i - \overline{x})(y_i - \overline{y}) + N\overline{xy}$$

Proof. This proof is similar to the above. Here ee will use both $\sum x_i = N\overline{x}$ and $\sum y_i = N\overline{y}$. We have

$$\sum x_i y_i = \sum[(x_i - \overline{x})(y_i - \overline{y}) + \overline{x}y_i + x_i\overline{y} - \overline{xy}]$$

$$= \sum(x_i - \overline{x})(y_i - \overline{y}) + \overline{x}\sum y_i + \overline{y}\sum x_i - N\overline{xy}$$

$$= \sum(x_i - \overline{x})(y_i - \overline{y}) + \overline{x}(N\overline{y}) + \overline{y}(N\overline{x}) - N\overline{xy} = \sum(x_i - \overline{x})(y_i - \overline{y}) + N\overline{xy}.$$

9. Now Equation (III) is rewritten as

$$m\left[\sum(x_i-\overline{x})^2+N\overline{x}^2\right]+c\sum x_i-\sum x_iy_i=0.$$

OR

$$m\left[\sum(x_i-\overline{x})^2+N\overline{x}^2\right]+cN\overline{x}-\sum x_iy_i=0.$$

OR

$$m\left[\sum(x_i-\overline{x})^2+N\overline{x}^2\right]+cN\overline{x}-\left[\sum(x_i-\overline{x})(y_i-\overline{y})+N\overline{xy}\right]=0.$$

Divide by N, we get

$$m\left[\sigma_x^2+\overline{x}^2\right]+c\overline{x}-[cov(x,y)+\overline{xy}]=0.$$

Substitute $c=\overline{y}-m\overline{x}$, we get

$$m\left[\sigma_x^2+\overline{x}^2\right]+(\overline{y}-m\overline{x})\overline{x}-[cov(x,y)+\overline{xy}]=0.$$

This reduces to

$$m\sigma_x^2-cov(x,y)=0 \qquad hence \qquad m=\frac{cov(x,y)}{\sigma_x^2}.$$

Therefore,

$$c=\overline{y}-m\overline{x}=\overline{y}-\frac{cov(x,y)}{\sigma_x^2}\overline{x}.$$

Theorem 0.1 *So, the least square line $y=mx+c$ or the regrassion line is*

$$y=\frac{cov(x,y)}{\sigma_x^2}x+\left(\overline{y}-\frac{cov(x,y)}{\sigma_x^2}\overline{x}\right)$$

Question or Problem: In least square method, we used the vertical distance of the line from the data points. It will be interesting to use perpendicular distance p_i of the points from the line. More precisely, suppose

$y = mx + c$ is a line and p_i is the perpendicular distance of the data point (x_i, y_i) from the line. Consider

$$P = \sum p_i^2.$$

Now minimize (optimize) P to determine this line.