Geometric Foundations of Data Analysis I: Week 2

Joshua Maglione

September 16, 2025

1.2 Plane of best fit

We consider two independent variables and one dependent variable now. Consider the following data points as given in Figure 1.1.

i	x_{i1}	x_{i2}	y_i
0	278	36	287
1	252	31	256
2	344	35	300
3	134	33	182
4	215	35	248
5	261	40	271
6	131	39	149
7	463	43	411
8	167	46	214
9	298	42	291
10	230	60	314
11	293	67	352
12	290	37	298
13	271	31	252
14	385	63	439
15	354	36	328

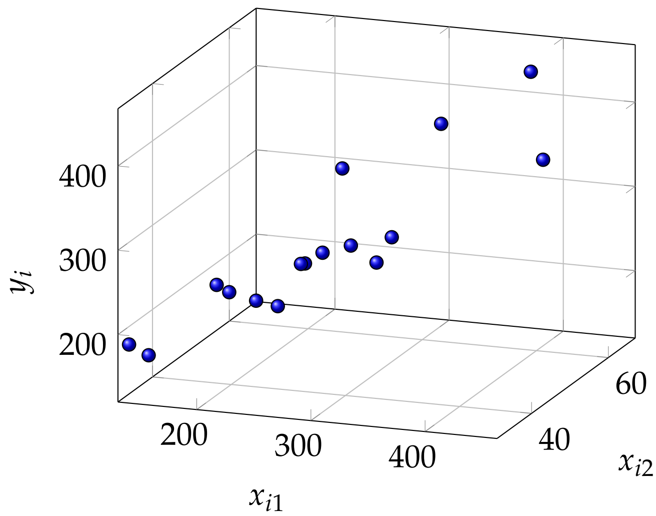

Figure 1.1: Data points in $\mathbb{R}^3$.

We can put some meaning to these data. For example, suppose a company is selling a product, and we have 16 populations of people labeled 0 through 15. The values x_{i1} are the population sizes in 100s of people; the values x_{i2} are the average yearly income in €1000 per capita; and the values y_i are the number of sales of the product. (There might be dependencies between population size and average income, but our model treats them as independent.)

It looks like though there is a plane of best fit for the data—thanks to the suggestive viewing angle. Our goal is to find a plane, given by

$$y = b_0 + b_1 x_1 + b_2 x_2.$$

We can just do what we did last time. That is, for

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} \end{pmatrix}, \qquad Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \qquad B = \begin{pmatrix} b_0 \\ b_1 \\ b_2 \end{pmatrix},$$

we need to solve for B in the equation

$$X^t X B = X^t Y.$$

Thus, if $X^t X$ is invertible, there is a unique B, which is equal to $(X^t X)^{-1} X^t Y$. For our example, we have

$$X^t X = \begin{pmatrix} 16 & 4366 & 674 \\ 4366 & 1309480 & 187024 \\ 674 & 187024 & 30330 \end{pmatrix}, \qquad X^t Y = \begin{pmatrix} 4592 \\ 1343400 \\ 200571 \end{pmatrix}.$$

Therefore, the plane of best fit is approximately

$$y = -11.3 + 0.7 x_1 + 2.6 x_2.$$

Putting all the data together we have a plane of best fit as seen in Figure 1.2.

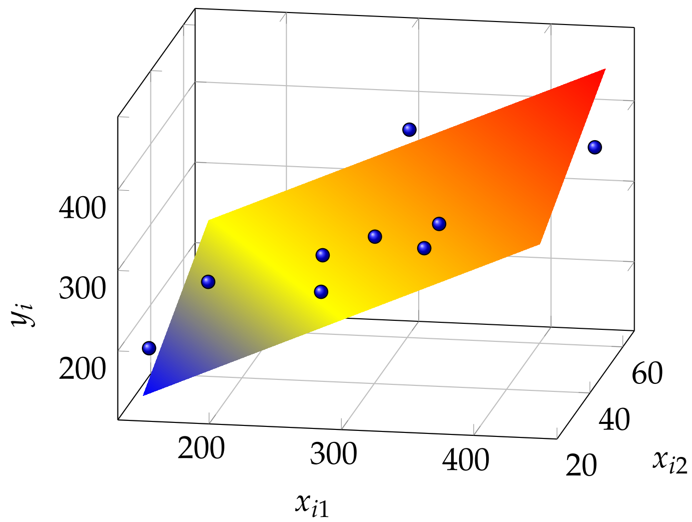

Figure 1.2: Data points together with the plane of best fit.

1.3 Hyperplane of best fit

Now we go to the general case. Suppose we have $p - 1$ independent variables and 1 dependent variable, where $p \geqslant 2$. We assume we have n data points of the form

$$(x_{i1}, x_{i2}, \dots, x_{i,p-1}, y_1) \in \mathbb{R}^p.$$

The least squares fitting for these data is a hyperplane of the form

$$y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_{p-1} x_{p-1}.$$

To solve for the values b_i, we do as we did before. We define matrices

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1,p-1} \\ 1 & x_{21} & x_{22} & \cdots & x_{2,p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{n,p-1} \end{pmatrix}, \qquad Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \qquad B = \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_{p-1} \end{pmatrix}.$$

As before, the values we want are given by the equation

$$X^{\mathrm{t}} X B = X^{\mathrm{t}} Y. \tag{1.1}$$

1.4 Why Equation (1.1) works

The heart of least squares is (Euclidean) distance. The distance between two points $x = (x_1, \ldots, x_p)$ and $y = (y_1, \ldots, y_p)$ in $\mathbb{R}^p$ is

$$d(x,y) = \|x - y\| = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_p - y_p)^2}.$$

For a vector $v = (v_1, \ldots, v_p) \in \mathbb{R}^p$, the **length** of v is

$$\|v\| = d(0, v) = \sqrt{v_1^2 + v_2^2 + \cdots + v_p^2}.$$

Recall that the dot product of two (column) vectors u and v is

$$u \cdot v = u^{\mathrm{t}} v = u_1 v_1 + u_2 v_2 + \cdots + u_p v_p.$$

Thus, the length of v is $\|v\| = \sqrt{v \cdot v}$; in other words $\|v\|^2 = v \cdot v$. In addition, if $u \cdot v = 0$, we say that u and v are **orthogonal** (or perpendicular).

The goal of least squares is to *minimize distance*; more specifically to minimize $\|Y - XB\|$. Note that the column vector Y has entries that are the *actual y_i values*, and the column vector

$$XB = \begin{pmatrix} B \cdot (1, x_{11}, x_{12}, \ldots, x_{1,p-1}) \\ B \cdot (1, x_{21}, x_{22}, \ldots, x_{2,p-1}) \\ \vdots \\ B \cdot (1, x_{n1}, x_{n2}, \ldots, x_{n,p-1}) \end{pmatrix} = \begin{pmatrix} b_0 + b_1 x_{11} + b_2 x_{12} + \cdots + b_{p-1} x_{1,p-1} \\ b_0 + b_1 x_{21} + b_2 x_{22} + \cdots + b_{p-1} x_{2,p-1} \\ \vdots \\ b_0 + b_1 x_{n1} + b_2 x_{n2} + \cdots + b_{p-1} x_{n,p-1} \end{pmatrix}$$

Therefore, $\|Y - XB\|$ is the square root of a sum of squares of the form

$$y_i - b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_{p-1} x_{i,p-1}.$$

Hence minimizing $\|Y - XB\|$ is the same as minimizing $\|Y - XB\|^2$, which is a sum of *squares*.

Proposition 1.1. *The minimal distance $\|Y - XB\|$ is achieved by solving for B in*

$$X^{\mathrm{t}} X B = X^{\mathrm{t}} Y.$$

Proof. Consider the subspace $U = \{Xu \mid u \in \mathbb{R}^p\}$ of $\mathbb{R}^n$, and observe that our desired solution XB is contained in U. Since $\|Y - XB\|$ is minimal, we must have that the vector $Y - XB$ is orthogonal to all vectors contained in U.[1] That is,

[1] To see why this is true, see Section 6.3.1 of [1], which is all about orthogonal decompositions.

$(Xu) \cdot (Y - XB) = 0$ for all $u \in \mathbb{R}^p$. In other words, we have for all $u \in \mathbb{R}^p$,

$$\begin{aligned} 0 = (Xu)^{\mathrm{t}}(Y - XB) &= u^{\mathrm{t}} X^{\mathrm{t}}(Y - XB) \\ &= u^{\mathrm{t}} \left(X^{\mathrm{t}} Y - X^{\mathrm{t}} XB \right). \end{aligned}$$

Because $u^{\mathrm{t}} \left(X^{\mathrm{t}} Y - X^{\mathrm{t}} XB \right) = 0$ for all $u \in \mathbb{R}^p$, it follows that $X^{\mathrm{t}} Y - X^{\mathrm{t}} XB = 0$. $\square$

1.5 In class exercises pt. II

1. Determine $X^{\mathrm{t}} X$ and $X^{\mathrm{t}} Y$ with

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1,p-1} \\ 1 & x_{21} & x_{22} & \cdots & x_{2,p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{n,p-1} \end{pmatrix}, \qquad Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}.$$

2. Using (1) and by taking partial derivatives of

$$S(b_0, \ldots, b_{p-1}) = \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_{p-1} x_{i,p-1}))^2, \quad (1.2)$$

   show that the hyperplane of best fit is obtained by solving $X^{\mathrm{t}} XB = X^{\mathrm{t}} Y$. (You could try this for $p = 3$ first.)

References

[1] Dan Margalit and Joseph Rabinoff, *Interactive Linear Algebra*, 2019, https://textbooks.math.gatech.edu/ila/.