Geometric Foundations of Data Analysis I: Week 3

Joshua Maglione

22 September 2025

Contents

2.8 Nonlinear fittings

Although all of our examples so far have been linear fittings, we will demonstrate that least squares fittings works in the nonlinear case. What is important is that we have a candidate equation to fit. In the linear cases, we tried to fit

$$y = b_0 + b_1x_1 + b_2x_2 + \cdots + b_{p-1}x_{p-1}.$$

Suppose we have the following data as given in Figure 2.1. Instead of trying to fit the line $y = b_0 + b_1x$, we could try to fit the parabola:

$$y = b_0 + b_1x + b_2x^2.$$

We can treat this the same way as before. Of course the quantities x and x^2 are *not* independent, but we can ignore this. Set

$$x_{i1} = x_i, \qquad\qquad x_{i2} = x_i^2.$$

Therefore, the hyperplane of best fit for the data (x_{i1}, x_{i2}, y_i) will give us the parabola of best fit. *Try this on your own!*

So one can fit any hypersurface $y = f(x_1, \dots, x_{p-1})$ to the given data. The function f in this case is called the **regression function**. This general method of analysis is known as **regression analysis**. A few questions arise:

- Which surface is "best"?
- How can we quantify "best"?
- Even in the line case $(p = 2)$, how can we quantify how well data fits our line?

x_i	y_i
2.27	2.50
5.06	-16.13
1.45	4.23
5.89	-22.46
0.48	1.37
-0.22	0.86
1.44	11.85
-1.77	-14.71
2.45	9.42
-1.54	-14.07
7.55	-55.62
1.76	4.45
5.16	-19.56
3.26	-2.79
3.23	5.20
0.85	8.09

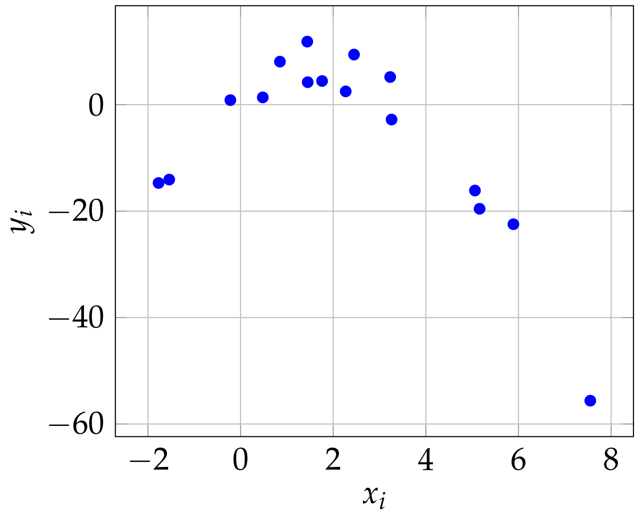

Figure 2.1: Data points demonstrating a nonlinear relationship.

2.9 Coefficient of determination (R^2 values)

We are going to make precise how well our hyperplane fits our data. Recall that hyperplanes can be replaced by hypersurfaces; see Section 2.8. First we establish some notation. Suppose we have n data points $(x_{i1}, x_{i2}, \dots, x_{i,p-1}, y_i) \in \mathbb{R}^p$. Then we define

$$\begin{aligned} &\text{(Fitted value)} && \widehat{y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_{p-1} x_{i,p-1}, \\ &\text{(Residual)} && e_i = y_i - \widehat{y}_i, \\ &\text{(Sample mean)} && \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i. \end{aligned}$$

These yield vectors in $\mathbb{R}^n$ as follows

$$\widehat{Y} = \begin{pmatrix} \widehat{y}_1 \\ \vdots \\ \widehat{y}_n \end{pmatrix} = XB, \qquad E = \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix} = Y - \widehat{Y}, \qquad \overline{Y} = \begin{pmatrix} \overline{y} \\ \vdots \\ \overline{y} \end{pmatrix} = \overline{y}\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}.$$

From our n data points, we have three points in $\mathbb{R}^n$ given by Y, $\widehat{Y}$, and $\overline{Y}$. Three points always lie on a plane, so the three point determine a triangle on such a plane. What does this triangle look like? If it is a triangle (and not a line or a single point), then the next lemma proves it must be a right triangle.

Lemma 2.2. *The vectors $E = Y - \widehat{Y}$ and $\widehat{Y} - \overline{Y}$ are orthogonal.*

Proof. Suppose $X^{\mathrm{t}}XB = X^{\mathrm{t}}Y$. We need to prove two equations. For the first,

$$0 = X^{\mathrm{t}}(Y - XB) = X^{\mathrm{t}}(Y - \widehat{Y}) = X^{\mathrm{t}}E.$$

Hence, $X^{\mathrm{t}}E = 0$. For the second,

$$\begin{aligned}
\overline{Y}^{\mathrm{t}}E &= \overline{y}\sum_{i=1}^{n}(y_i - \widehat{y}_i) \\
&= \overline{y}\sum_{i=1}^{n}(y_i - (b_0 + b_1x_{i1} + b_2x_{i2} + \cdots + b_{p-1}x_{i,p-1})) \\
&= -\frac{\overline{y}}{2}\cdot\frac{\partial S}{\partial b_0} = 0,
\end{aligned}$$

where S is defined in the Week 1 Notes, so $\overline{Y}^{\mathrm{t}}E = 0$. Thus, we have

$$(\widehat{Y} - \overline{Y})\cdot E = (XB)^{\mathrm{t}}E - \overline{Y}^{\mathrm{t}}E = 0. \qquad \square$$

Remark 2.3. One can simplify the proof for Lemma 2.2 by applying an isometry to the data, so that $\overline{y} = 0$. That is, one only needs to prove that E and $\widehat{Y}$ are orthogonal.