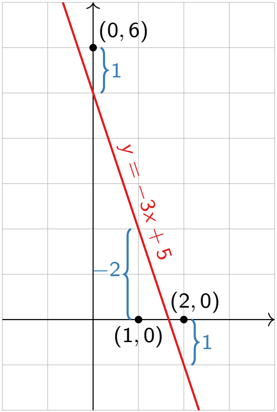
Section 6.6

Least Squares Problems

Data Modeling: Best fit line

What does it minimize?

Best fit line minimizes the **sum of the squares** of the *vertical distances from the data points* to the line.



Data modeling: best fit parabola

What least squares problem $Ax = b$ finds **the best parabola** through the points $(-1, 0.5)$, $(1, -1)$, $(2, -0.5)$, $(3, 2)$?

The general equation for a parabola is

$$ax^2 + bx + c = y.$$

So we want to solve:

$$\begin{aligned} a(-1)^2 + b(-1) + c &= 0.5 \\ a(1)^2 + b(1) + c &= -1 \\ a(2)^2 + b(2) + c &= -0.5 \\ a(3)^2 + b(3) + c &= 2 \end{aligned}$$

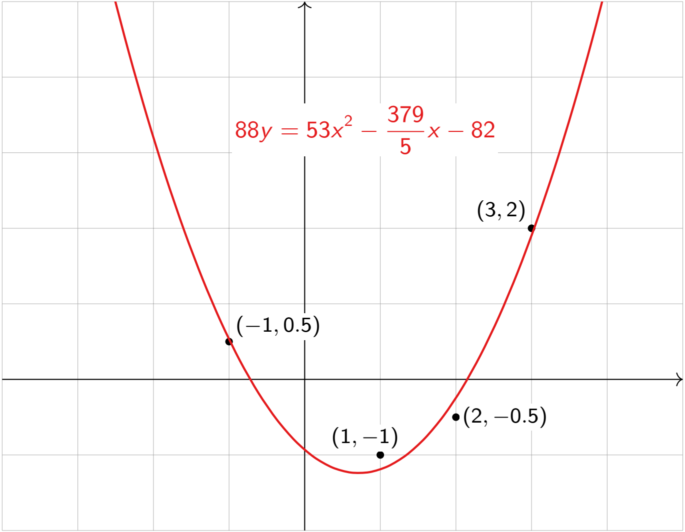
In matrix form:

$$\begin{pmatrix} 1 & -1 & 1 \\ 1 & 1 & 1 \\ 4 & 2 & 1 \\ 9 & 3 & 1 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} 0.5 \\ -1 \\ -0.5 \\ 2 \end{pmatrix}.$$

Answer: $\widehat{a} = \frac{53}{88}, \widehat{b} = \frac{379}{440}, \widehat{c} = \frac{82}{88}$ so best fit is: $53x^2 - \dfrac{379}{5}x - 82 = 88y$

Data modeling: best fit parabola

Picture



Data modeling: best fit ellipse

Find the best fit ellipse for the points $(0, 2)$, $(2, 1)$, $(1, -1)$, $(-1, -2)$, $(-3, 1)$.

The general equation for an ellipse is

$$x^2 + ay^2 + bxy + cx + dy + e = 0$$

So we want to solve:

$$\begin{aligned}
(0)^2 + \quad A(2)^2 + \quad B(0)(2) + \quad C(0) + \quad D(2) + E &= 0\\
(2)^2 + \quad A(1)^2 + \quad B(2)(1) + \quad C(2) + \quad D(1) + E &= 0\\
(1)^2 + A(-1)^2 + \quad B(1)(-1) + \quad C(1) + D(-1) + E &= 0\\
(-1)^2 + A(-2)^2 + B(-1)(-2) + C(-1) + D(-2) + E &= 0\\
(-3)^2 + \quad A(1)^2 + \quad B(-3)(1) + C(-3) + \quad D(1) + E &= 0
\end{aligned}$$

In matrix form:

$$\begin{pmatrix} 4 & 0 & 0 & 2 & 1 \\ 1 & 2 & 2 & 1 & 1 \\ 1 & -1 & 1 & -1 & 1 \\ 4 & 2 & -1 & -2 & 1 \\ 1 & -3 & -3 & 1 & 1 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \\ d \\ e \end{pmatrix} = \begin{pmatrix} 0 \\ -4 \\ -1 \\ -1 \\ -9 \end{pmatrix}.$$

Data modeling: best fit ellipse

Complete procedure

$$A = \begin{pmatrix} 4 & 0 & 0 & 2 & 1 \\ 1 & 2 & 2 & 1 & 1 \\ 1 & -1 & 1 & -1 & 1 \\ 4 & 2 & -1 & -2 & 1 \\ 1 & -3 & -3 & 1 & 1 \end{pmatrix} \qquad b = \begin{pmatrix} 0 \\ -4 \\ -1 \\ -1 \\ -9 \end{pmatrix}.$$

$$A^T A = \begin{pmatrix} 35 & 6 & -4 & 1 & 11 \\ 6 & 18 & 10 & -4 & 0 \\ -4 & 10 & 15 & 0 & -1 \\ 1 & -4 & 0 & 11 & 1 \\ 11 & 0 & -1 & 1 & 5 \end{pmatrix} \qquad A^T b = \begin{pmatrix} -18 \\ 18 \\ 19 \\ -10 \\ -15 \end{pmatrix}$$

Row reduce:

$$\left(\begin{array}{ccccc|c} 35 & 6 & -4 & 1 & 11 & -18 \\ 6 & 18 & 10 & -4 & 0 & 18 \\ -4 & 10 & 15 & 0 & -1 & 19 \\ 1 & -4 & 0 & 11 & 1 & -10 \\ 11 & 0 & -1 & 1 & 5 & -15 \end{array}\right) \rightsquigarrow \left(\begin{array}{ccccc|c} 1 & 0 & 0 & 0 & 0 & 16/7 \\ 0 & 1 & 0 & 0 & 0 & -8/7 \\ 0 & 0 & 1 & 0 & 0 & 15/7 \\ 0 & 0 & 0 & 1 & 0 & -6/7 \\ 0 & 0 & 0 & 0 & 1 & -52/7 \end{array}\right)$$

Best fit ellipse:

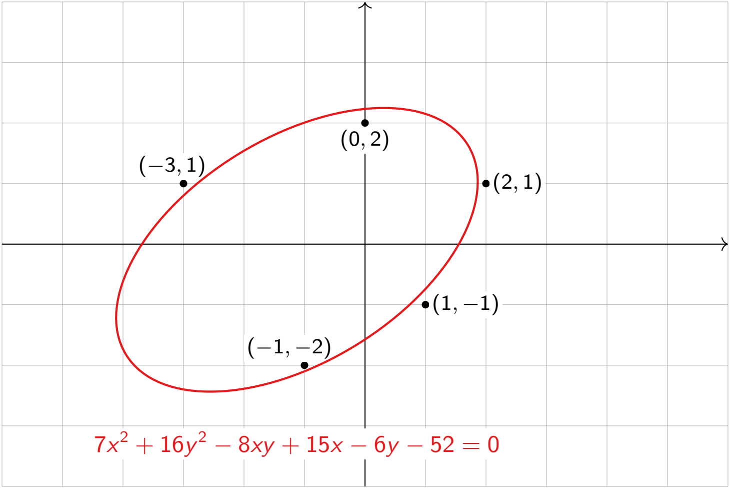
$$x^2 + \frac{16}{7}y^2 - \frac{8}{7}xy + \frac{15}{7}x - \frac{6}{7}y - \frac{52}{7} = 0$$

or

$$7x^2 + 16y^2 - 8xy + 15x - 6y - 52 = 0.$$

Data modeling: best fit ellipse

Picture



Remark: Gauss invented the method of least squares to do exactly this: he predicted the (elliptical) orbit of the asteroid Ceres as it passed behind the sun in 1801.

Extra: Best fit linear function

What least squares problem $Ax = b$ finds the best linear function $f(x, y)$ **fitting the following data**?

x	y	$f(x,y)$
1	0	0
0	1	1
−1	0	3
0	−1	4

The general equation for a linear function in two variables is

$$f(x,y) = ax + by + c.$$

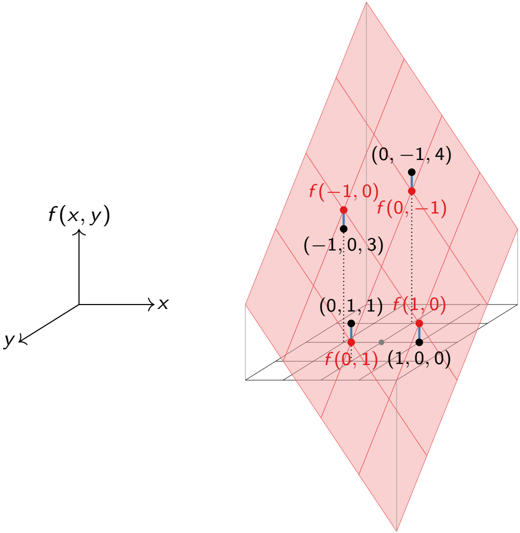
So we want to solve

$$\begin{aligned} a(1) + \quad b(0) + c &= 0 \\ a(0) + \quad b(1) + c &= 1 \\ a(-1) + \quad b(0) + c &= 3 \\ a(0) + b(-1) + c &= 4 \end{aligned}$$

In matrix form:

$$\begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ -1 & 0 & 1 \\ 0 & -1 & 1 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 3 \\ 4 \end{pmatrix}.$$

Answer: $\widehat{a} = -\frac{3}{2}, \widehat{b} = -\frac{3}{2}, \widehat{c} = 2$ so best fit is: $f(x,y) = -\dfrac{3}{2}x - \dfrac{3}{2}y + 2$



Graph of

$$f(x,y) = -\frac{3}{2}x - \frac{3}{2}y + 2$$

Multiple Regression

Generalizing the best-fit plane before:

- A **variable y depends** on
- Independent *variables u, v*

General formula:

$$y = \beta_0 f_0(u,v) + \beta_1 f_1(u,v) + \cdots + \beta_k f_k(u,v)$$

with $f_0, \ldots, f_k$ any sort of known functions and $\beta_0, \ldots, \beta_k$ unknown weights.

The best fit plane:

$$y = \beta_0 + \beta_1 u + \beta_2 v$$

A quadratic function (next week's subject):

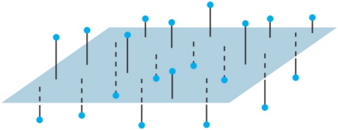
$$y = \beta_0 + \beta_1 u + \beta_2 v + \beta_3 u^2 + \beta_4 uv + \beta_5 v^2$$

Multiple regression

Expert's notation

The model to fit:

$$
\begin{aligned}
y_1 &= \beta_0 + \beta_1 u_1 + \beta_2 v_1 + \epsilon_1 \\
y_2 &= \beta_0 + \beta_1 u_2 + \beta_2 v_2 + \epsilon_2 \\
&\vdots \qquad\qquad\quad \vdots \\
y_n &= \beta_0 + \beta_1 u_n + \beta_2 v_n + \epsilon_n
\end{aligned}
$$



The equation display $y = X\beta + \varepsilon$:

Observation vector, Design matrix, Parameter vector, Residual vector

$$
\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & u_1 & v_1 \\ 1 & u_2 & v_2 \\ \vdots & \vdots & \vdots \\ 1 & u_n & v_n \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}, \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}
$$

The error

We want to *minimize the length of ε*.
In last section we don't write it as part of the equation.