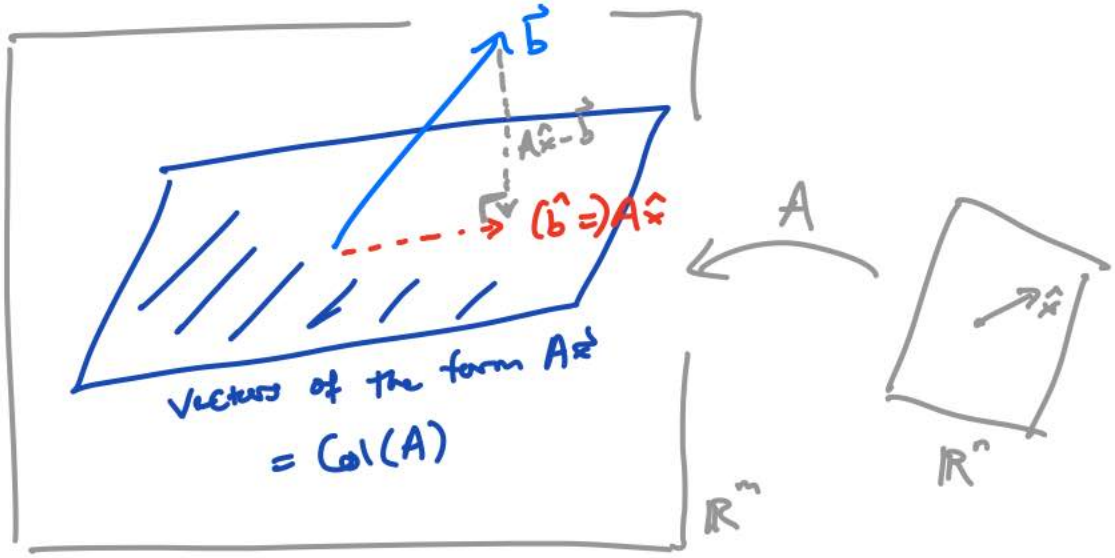
Lecture 34: Least Squares (II)

Before we continue with this material, I pause to note that the method of least squares was invented by Gauss & Legendre more than 200 years ago to account for measurement errors in astronomical tables used for ship navigation. In the more general form of linear regression, it continues to have a large number of applications in the hard and social sciences.

We begin by reviewing the normal equations. Recall that a least-squares solution to $A\vec{x} = \vec{b}$ (A = $m \times n$ matrix) is a vector $\hat{x} \in \mathbb{R}^n$ which minimizes $\|A\vec{x} - \vec{b}\|$:



Since $\hat{b} := \text{Proj}_{\text{Col}(A)} \vec{b}$ minimizes the distance to $\vec{b}$, we can take for $\hat{x}$ any (exact) solution to

$$A\hat{x} = \hat{b}.$$

Now, the space of vectors "$\perp$" to $\mathrm{Col}(A)$ is

$$\mathrm{Col}(A)^{\perp} = \mathrm{Nul}(A^T). \qquad \text{(Why?)}$$

So

$$0 = A^T(\vec{b} - \hat{b}) = A^T(\vec{b} - A\hat{x}) = A^T\vec{b} - A^TA\hat{x}$$

$$\Rightarrow \boxed{A^TA\hat{x} = A^T\vec{b}} \quad \text{(Normal Equations)}.$$

(We showed that, conversely, any solution $\hat{x}$ to this is a least-squares solution to $A\vec{x} = \vec{b}$.)

Ex 1 / Find all least-squares solutions to

$$\underbrace{\begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix}}_{A} \vec{x} = \underbrace{\begin{pmatrix} 1 \\ 3 \\ 8 \\ 2 \end{pmatrix}}_{\vec{b}}.$$

Compute

$$A^TA = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 4 & 2 & 2 \\ 2 & 2 & 0 \\ 2 & 0 & 2 \end{pmatrix}$$

and

$$A^T\vec{b} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 3 \\ 8 \\ 2 \end{pmatrix} = \begin{pmatrix} 14 \\ 4 \\ 10 \end{pmatrix}.$$

Now write the normal equations in augmented matrix form:

$$\left[\begin{array}{ccc|c} 4 & 2 & 2 & 14 \\ 2 & 2 & 0 & 4 \\ 2 & 0 & 2 & 10 \end{array}\right] \to \left[\begin{array}{ccc|c} 1 & 0 & 1 & 5 \\ 0 & 1 & -1 & -3 \\ 0 & 0 & 0 & 0 \end{array}\right] \to \begin{array}{l} x_1 + x_3 = 5 \\ x_2 - x_3 = -3 \end{array}$$

x_3 free

$$\rightsquigarrow \quad \begin{array}{l} x_1 = -x_3 + 5 \\ x_2 = x_3 - 3 \end{array} \quad \rightsquigarrow \quad \hat{x} = t\begin{pmatrix} -1 \\ 1 \\ 1 \end{pmatrix} + \begin{pmatrix} 5 \\ -3 \\ 0 \end{pmatrix}.$$

Any choice of t yields a least-squares solution.

This begs the question: when is $\hat{x}$ unique (since it clearly isn't in the example we just did)? Well,

$$\text{Solution to normal eqn. is unique} \iff \mathrm{Nul}(A^TA) = \{0\}$$

$$\iff A^TA \text{ is invertible.}$$

(A^TA square)

Claim: A^TA is invertible $\iff$ A has independent columns.

Proof: First, observe that independence of A's columns is equivalent to "$A\vec{x} = \vec{0} \Rightarrow \vec{x} = \vec{0}$", i.e. $\mathrm{Nul}(A) = \{0\}$.

($A\vec{x}$: linear comb. of A's columns w/ weights x_i)

So we must show

$$\mathrm{Nul}(A^TA) = \{\vec{0}\} \iff \mathrm{Nul}(A) = \{\vec{0}\}.$$

($\Rightarrow$): If $A\vec{x} = \vec{0}$, then $A^TA\vec{x} = \vec{0}$, whereupon (if $\mathrm{Nul}(A^TA) = \{\vec{0}\}$) $\vec{x} = \vec{0}$.

($\Longleftarrow$): If $A^TA\vec{x} = \vec{0}$, then

$$0 = \vec{x}^TA^TA\vec{x} = (A\vec{x})^T(A\vec{x}) = A\vec{x}\cdot A\vec{x} = \|A\vec{x}\|^2$$

$$\Rightarrow A\vec{x} = \vec{0}, \quad \text{and so (if } \mathrm{Nul}(A) = \vec{0}) \ \vec{x} = \vec{0}. \qquad \square$$

Corollary: If A has independent columns, $\hat{x}$ is unique and we have $\hat{x} = (A^TA)^{-1}A^T\vec{b}$.

Also note that if A has independent columns, we have a decomposition $A = QR$ = (orthogonal) $\times$ (upper triangular with positive diagonal entries).

(upper triangular with positive diagonal entries: invertible)

So

$$\hat{x} = (A^TA)^{-1}A^T\vec{b} = ((QR)^TQR)^{-1}(QR)^T\vec{b}$$

$$= (R^TQ^TQR)^{-1}R^TQ^T\vec{b} = R^{-1}(R^T)^{-1}R^TQ^T\vec{b}$$

I, since Q is orthogonal

I

$$= R^{-1}Q^T\vec{b},$$

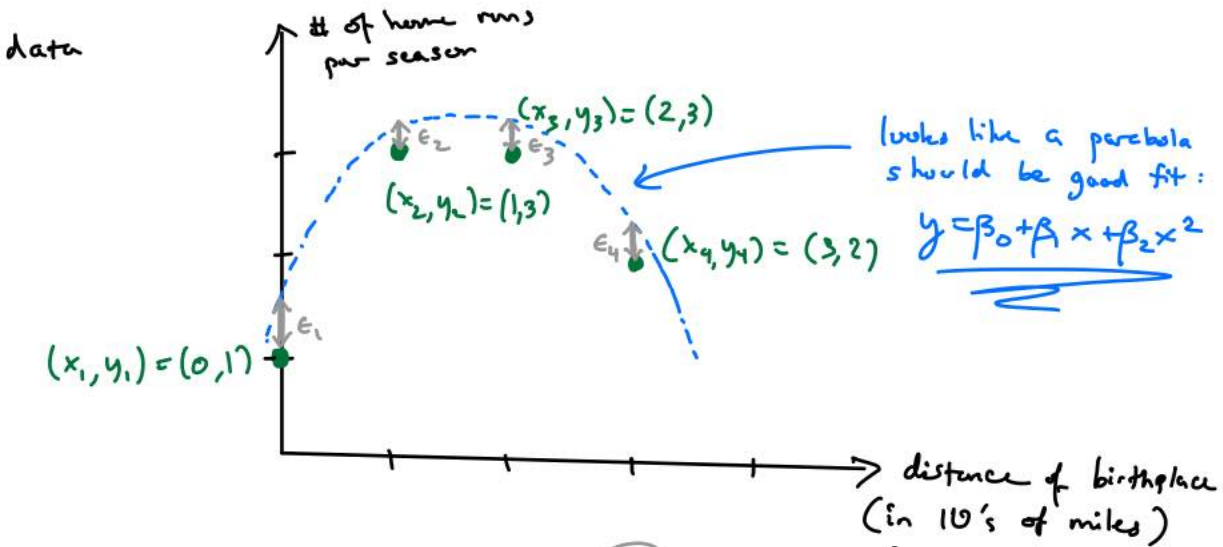
which is exactly what we had (by a different method) in the last lecture.

Linear Regression

Here "linear" refers to the use of linear methods to determine the coefficients in an equation (expressing the relationship between two variables) that gives the best fit to observed data in the

sense of least squares. The equation itself may not be linear, as we see in the next example.

Ex 2 / Having entered the sports business, Monsanto (or rather the Bayer "crop sciences division") is trying to build a better baseball player. They've tracked a few little-leaguers and got the data



looks like a parabola should be good fit:

$y = \beta_0 + \beta_1 x + \beta_2 x^2$

So write

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \epsilon_1 \\ &\vdots \qquad\quad \vdots \qquad\quad \vdots \qquad\quad \vdots \\ y_4 &= \beta_0 + \beta_1 x_4 + \beta_2 x_4^2 + \epsilon_4 \end{aligned}$$

(error)

i.e.

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ 1 & x_3 & x_3^2 \\ 1 & x_4 & x_4^2 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \end{pmatrix}$$

$$\vec{y} = X \cdot \vec{\beta} + \vec{\epsilon}$$

to minimize $\|\vec{\epsilon}\| = \|\vec{y} - X\vec{\beta}\|$, want to solve

$$\boxed{X\hat{\beta} = \hat{y}}$$

or equivalently (normal equations)

$$X^T X \hat{\beta} = X^T \vec{y},$$

which is

$$\begin{pmatrix} 4 & 6 & 14 \\ 6 & 14 & 36 \\ 14 & 36 & 98 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} 9 \\ 15 \\ 33 \end{pmatrix}.$$

Row-reducing the associated augmented matrix $\Rightarrow \hat{\beta} = \begin{pmatrix} 1.05 \\ 2.55 \\ -0.75 \end{pmatrix}$.

So use $y = 1.05 + 2.55x - 0.75x^2$

$\Rightarrow \frac{dy}{dx} = 2.55 - 1.5x \Rightarrow x = \frac{2.55}{1.5} = 1.7 \Rightarrow$

should raise baseball players 17 miles from the arch. //

While silly, this illustrates why one might <u>want</u> to fit a model to pre-existing data points — for its <u>predictive</u> power. Rather than looking at the <u>shape</u> of the data, the form of the equation might arise from physical principle:

Ex 3/ If $\begin{cases} M_A \text{ grams of radioactive substance A} \\ M_B \text{ grams of radioactive substance B} \end{cases}$

are in your backpack at time $t = 0$, then the total amount of radioactive mixture is

$$y = M_A e^{-0.02t} + M_B e^{-0.07t}.$$

(decay constants)

We know what the substances are, hence their decay constants, but not M_A & M_B (the bullies won't say). However, you measure $(t, y) = (t_1, y_1), (t_2, y_2), \ldots, (t_n, y_n)$. There is of course error in these measurements. (So we can see the point of least-squares in the context of problem Gauss studied.)

The linear model that can be used to estimate M_A & M_B is:

$$\begin{cases} y_1 = M_A e^{-0.02t_1} + M_B e^{-0.07t_1} + \epsilon_1 \\ \vdots \qquad\qquad \vdots \qquad\qquad \vdots \qquad \vdots \\ y_n = M_A e^{-0.02t_n} + M_B e^{-0.07t_n} + \epsilon_n \end{cases}$$

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} e^{-0.02t_1} & e^{-0.07t_1} \\ \vdots & \vdots \\ e^{-0.02t_n} & e^{-0.07t_n} \end{pmatrix} \begin{pmatrix} M_A \\ M_B \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

$$\vec{y} = X \cdot \vec{\beta} + \vec{\epsilon}$$

least-squares solution of this will then give approximate values of M_A & M_B. //

I'll let you read the text's example of multiple regression — it's not that different from the above.