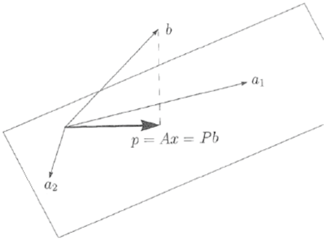
Projection onto a subspace

We can also project a vector $\boldsymbol{b}$ onto a subspace (for instance, a plane); in this case we are looking for the closest vector to $\boldsymbol{b}$ in the subspace.

It is convenient to write the subspace as the span of the vectors $\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n$, and let A denote the matrix with $\boldsymbol{a}_i$ as its *columns*. So the subspace is just col(A).

We define the projection onto the subspace as a vector $\boldsymbol{p} = A\boldsymbol{x}$, such that the distance $\|\boldsymbol{b} - A\boldsymbol{x}\|$ is minimized.



Projection onto a subspace – formula

Since $\boldsymbol{b} - A\boldsymbol{x}$ is perpendicular to every column of A:

$$A^T(\boldsymbol{b} - A\boldsymbol{x}) = \boldsymbol{0} \;\Rightarrow\; A^T A\boldsymbol{x} = A^T\boldsymbol{b}.$$

If $A^T A$ is invertible, then

$$\boldsymbol{p} = A\boldsymbol{x} = \underbrace{A(A^T A)^{-1}A^T}_{P}\,\boldsymbol{b} = P\boldsymbol{b}.$$

If A has only 1 column, then the projection matrix P simplifies to the expression on Slide 11.

Regression

Regression is a technique used to determine a relationship between independent and dependent variables in data set.

For example, suppose we have independent variables x_1 and x_2, and dependent variable y. We suspect that there is a relationship of the form $y = c_0 + c_1x_1 + c_2x_2$, and we wish to find c_i.

Suppose our data consists of the measurements (x_{i1}, x_{i2}, y_i) for $i = 1, 2, \ldots, n$. In matrix form, the proposed relationship can be written as the equation

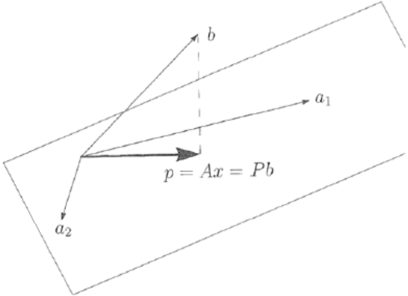
$$\underbrace{\begin{bmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} c_0 \\ c_1 \\ c_2 \end{bmatrix}}_{\boldsymbol{x}} = \underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}}_{\boldsymbol{b}}$$

Projection

Usually, there is no solution to the equation $A\boldsymbol{x} = \boldsymbol{b}$, because

- The proposed relationship may not be exact,
- There are probably measurement errors.

So the best we can do is to find an $\boldsymbol{x}$ such that the distance $\|\boldsymbol{b} - A\boldsymbol{x}\|$ is minimized. This is precisely the same as projecting $\boldsymbol{b}$ onto the column space of A!



From Cohort 12, we know that $\boldsymbol{x} = (A^T A)^{-1} A^T \boldsymbol{b}$.

Regression, example 1

For different branches of a company, let $y =$ sales revenues in \$millions, $x_1 =$ number of sales people, and $x_2 =$ sales expenditures in \$millions.

x_1	x_2	y
31	1.85	4.20
46	2.80	7.28
40	2.20	5.60
49	2.85	8.12
38	1.80	5.46
49	2.80	7.42
31	1.85	3.36
38	2.30	5.88
33	1.60	4.62
42	2.15	5.88

We suspect $y = c_0 + c_1x_1 + c_2x_2$. Estimate the values of c_i.

Regression, example 1

We form

$$A = \begin{bmatrix} 1 & 31 & 1.85 \\ 1 & 46 & 2.80 \\ 1 & 40 & 2.20 \\ 1 & 49 & 2.85 \\ 1 & 38 & 1.80 \\ 1 & 49 & 2.80 \\ 1 & 31 & 1.85 \\ 1 & 38 & 2.30 \\ 1 & 33 & 1.60 \\ 1 & 42 & 2.15 \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} 4.20 \\ 7.28 \\ 5.60 \\ 8.12 \\ 5.46 \\ 7.42 \\ 3.36 \\ 5.88 \\ 4.62 \\ 5.88 \end{bmatrix},$$

then compute

$$\boldsymbol{x} = (A^T A)^{-1} A^T \boldsymbol{b} \approx \begin{bmatrix} -2.61 \\ 0.192 \\ 0.341 \end{bmatrix}.$$

So $y \approx -2.61 + 0.192x_1 + 0.341x_2$.

General case

More precisely, this technique is called *least square regression*:

- Suspect a relationship $y = c_0 + c_1 x_1 + \cdots + c_m x_m$.
- Store the y (dependent variable) measurements into a column vector, $\boldsymbol{b}$.
- Let A be an $n \times (m+1)$ matrix. Fill in the first column of A with 1's.
- Store the x_1 measurements, in the correct order, into the second column of A, the x_2 measurements into the third column of A, etc.
- Compute $\boldsymbol{x} = (A^T A)^{-1} A^T \boldsymbol{b}$. Then the components of $\boldsymbol{x}$ are the best approximations to $c_0, c_1, \ldots, c_m$.

Regression, example 2

This is exactly what *Excel*'s `Add Trendline` function does!

Example: data for female life expectancy in the US (y) vs year (x) is shown below:

x	y
1920	54.6
1930	61.6
1940	65.2
1950	71.1
1960	73.1
1970	74.7
1980	77.5
1990	78.8
2000	79.7
2010	81.1

Regression, example 2

We can try different models:

Model 1: $y = c_0 + c_1 x$.

Just take $x_1 = x$. See MATLAB, and compare with *Excel*.

Model 2: $y = c_0 + c_1 x + c_2 x^2$.

Here we take $x_1 = x$, $x_2 = x^2$. See MATLAB.

There are many other possible models. One possibility is $y = \alpha x^\beta$, which we can transform into a linear equation by taking the log of both sides:

$$\underbrace{\log(y)}_{\text{`}y\text{'}} = \underbrace{\log(\alpha)}_{c_0} + \underbrace{\beta}_{c_1} \underbrace{\log(x)}_{x_1}.$$