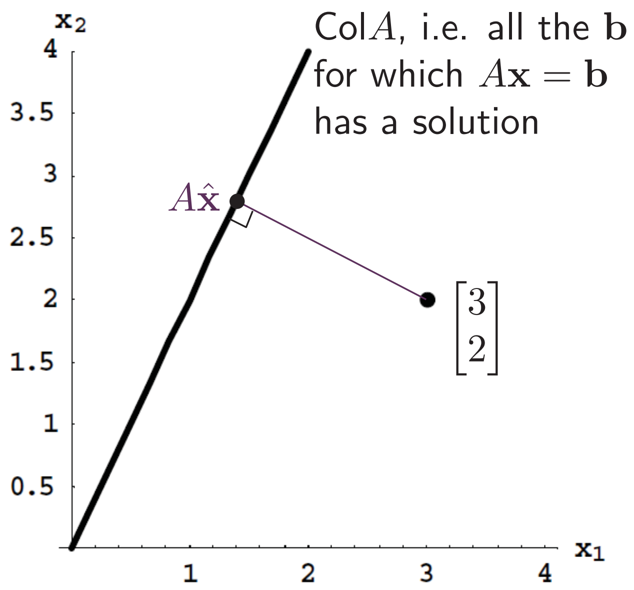
Let $A = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$. The linear system $A\mathbf{x} = \begin{bmatrix} 3 \\ 2 \end{bmatrix}$ does not have a solution, because

$\begin{bmatrix} 3 \\ 2 \end{bmatrix}$ is not in Col$A$ = Span $\left\{ \begin{bmatrix} 1 \\ 2 \end{bmatrix} \right\}$.

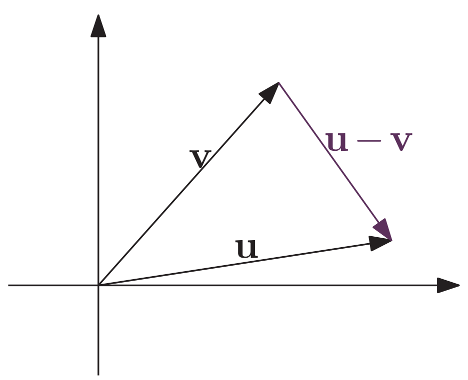


We wish to find a "closest approximate solution", i.e. a vector $\hat{\mathbf{x}}$ such that $A\hat{\mathbf{x}}$ is the unique point in Col$A$ that is "closest" to $\begin{bmatrix} 3 \\ 2 \end{bmatrix}$. This is called a least-squares solution (p17).

To do this, we have to first define what we mean by "closest", i.e. define the idea of distance.

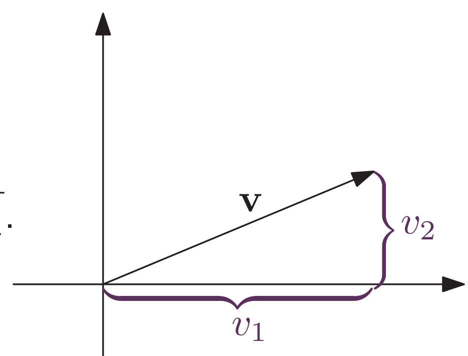
In $\mathbb{R}^2$, the distance between $\mathbf{u}$ and $\mathbf{v}$ is the length of their difference $\mathbf{u} - \mathbf{v}$.

So, to define distances in $\mathbb{R}^n$, it's enough to define the length of vectors.



In $\mathbb{R}^2$, the length of $\begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$ is $\sqrt{v_1^2 + v_2^2}$.

So we define the length of $\begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix}$ is $\sqrt{v_1^2 + \cdots + v_n^2}$.



§6.1, p368: Length, Orthogonality, Best Approximation

It is more useful to define a more general idea:

Definition: The *dot product* of two vectors $\mathbf{u} = \begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix}$ and $\mathbf{v} = \begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix}$ in $\mathbb{R}^n$ is the scalar

$$\mathbf{u} \cdot \mathbf{v} = \mathbf{u}^T \mathbf{v} = u_1 v_1 + \cdots + u_n v_n.$$

Warning: do not write $\mathbf{uv}$, which is an undefined matrix-vector product, or $\mathbf{u} \times \mathbf{v}$, which has a different meaning.

Definition: The *length* or *norm* of $\mathbf{v}$ is

$$\|\mathbf{v}\| = \sqrt{\mathbf{v} \cdot \mathbf{v}} = \sqrt{v_1^2 + \cdots + v_n^2}.$$

Definition: The *distance* between $\mathbf{u}$ and $\mathbf{v}$ is $\|\mathbf{u} - \mathbf{v}\|$.

$\mathbf{u} \cdot \mathbf{v} = \mathbf{u}^T \mathbf{v} = u_1 v_1 + \cdots + u_n v_n.$

$\|\mathbf{v}\| = \sqrt{\mathbf{v} \cdot \mathbf{v}} = \sqrt{v_1^2 + \cdots + v_n^2}.$

Distance between $\mathbf{u}$ and $\mathbf{v}$ is $\|\mathbf{u} - \mathbf{v}\|$.

Example: $\mathbf{u} = \begin{bmatrix} 3 \\ 0 \\ -1 \end{bmatrix}, \mathbf{v} = \begin{bmatrix} 8 \\ 5 \\ -6 \end{bmatrix}.$

$\mathbf{u} \cdot \mathbf{v} = 3 \cdot 8 + 0 \cdot 5 + -1 \cdot -6 = 24 + 0 + 6 = 30.$

The distance between $\mathbf{u}$ and $\mathbf{v}$ is

$$\left\| \begin{bmatrix} 3 \\ 0 \\ -1 \end{bmatrix} - \begin{bmatrix} 8 \\ 5 \\ -6 \end{bmatrix} \right\| = \left\| \begin{bmatrix} -5 \\ -5 \\ 5 \end{bmatrix} \right\| = \sqrt{(-5)^2 + (-5)^2 + 5^2} = \sqrt{75} = 5\sqrt{3}.$$

Properties of the dot product:

Let $\mathbf{u}, \mathbf{v}$ and $\mathbf{w}$ be vectors in $\mathbf{R}^n$, and let c be any scalar. Then

a. $\mathbf{u} \cdot \mathbf{v} = \mathbf{v} \cdot \mathbf{u}$ — symmetry

b. $(\mathbf{u} + \mathbf{v}) \cdot \mathbf{w} = \mathbf{u} \cdot \mathbf{w} + \mathbf{v} \cdot \mathbf{w}$

c. $(c\mathbf{u}) \cdot \mathbf{v} = c(\mathbf{u} \cdot \mathbf{v}) = \mathbf{u} \cdot (c\mathbf{v})$

(b and c: linearity in each input separately)

d. $\mathbf{u} \cdot \mathbf{u} \geq 0$, and $\mathbf{u} \cdot \mathbf{u} = 0$ if and only if $\mathbf{u} = \mathbf{0}$. — positivity; and the only vector with length 0 is $\mathbf{0}$

Combining parts b and c, one can show

$$\boxed{(c_1\mathbf{u}_1 + \cdots + c_p\mathbf{u}_p) \cdot \mathbf{w} = c_1(\mathbf{u}_1 \cdot \mathbf{w}) + \cdots + c_p(\mathbf{u}_p \cdot \mathbf{w})}$$

So b and c together says that, for fixed $\mathbf{w}$, the function $\mathbf{x} \mapsto \mathbf{x} \cdot \mathbf{w}$ is linear - this is true because $\mathbf{x} \cdot \mathbf{w} = \mathbf{w} \cdot \mathbf{x} = \mathbf{w}^T\mathbf{x}$ and matrix multiplication by $\mathbf{w}^T$ is linear.

From property c:

$$\|c\mathbf{v}\|^2 = (c\mathbf{v}) \cdot (c\mathbf{v}) = c^2\mathbf{v}.\mathbf{v} = c^2 \|\mathbf{v}\|^2,$$

so (squareroot both sides)

$$\|c\mathbf{v}\| = |c| \, \|\mathbf{v}\|.$$

For many applications, we are interested in vectors of length 1.
Definition: A *unit vector* is a vector whose length is 1.

Given $\mathbf{v}$, to create a unit vector in the same direction as $\mathbf{v}$, we divide $\mathbf{v}$ by its length $\|\mathbf{v}\|$ (i.e. take $c = \frac{1}{\|\mathbf{v}\|}$ in the equation above). This process is called normalising.

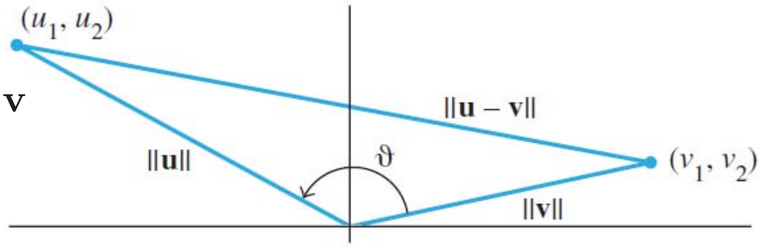
Example: Find a unit vector in the same direction as $\mathbf{v} = \begin{bmatrix} 8 \\ 5 \\ -6 \end{bmatrix}$.

Answer: $\mathbf{v} \cdot \mathbf{v} = 8^2 + 5^2 + (-6)^2 = 125$.

So a unit vector in the same direction as $\mathbf{v}$ is $\dfrac{\mathbf{v}}{\|\mathbf{v}\|} = \dfrac{1}{\sqrt{125}} \begin{bmatrix} 8 \\ 5 \\ -6 \end{bmatrix}$.

Visualising the dot product:
In $\mathbb{R}^2$, the cosine law says $\|\mathbf{u}-\mathbf{v}\|^2 = \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2 - 2\|\mathbf{u}\|\,\|\mathbf{v}\|\cos\theta$.
We can "expand" the left hand side using dot products:

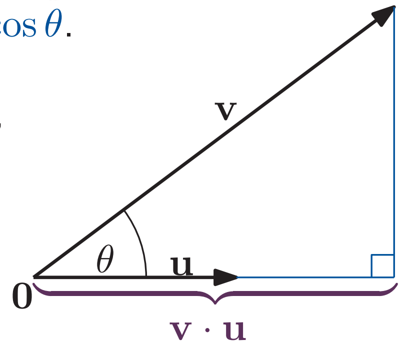
$$\begin{aligned}\|\mathbf{u}-\mathbf{v}\|^2 &= (\mathbf{u}-\mathbf{v})\cdot(\mathbf{u}-\mathbf{v})\\ &= \mathbf{u}\cdot\mathbf{u} - \mathbf{u}\cdot\mathbf{v} - \mathbf{v}\cdot\mathbf{u} + \mathbf{v}\cdot\mathbf{v}\\ &= \|\mathbf{u}\|^2 - 2\mathbf{u}\cdot\mathbf{v} + \|\mathbf{v}\|^2.\end{aligned}$$



Comparing with the cosine law, we see $\mathbf{u}\cdot\mathbf{v} = \|\mathbf{u}\|\,\|\mathbf{v}\|\cos\theta$.

In particular, if $\mathbf{u}$ is a unit vector, then $\mathbf{v}\cdot\mathbf{u} = \|\mathbf{v}\|\cos\theta$, as shown in the bottom picture.

Notice that $\mathbf{u}$ and $\mathbf{v}$ are perpendicular if and only if $\theta = \frac{\pi}{2}$, i.e. when $\cos\theta = 0$. This is equivalent to $\mathbf{u}\cdot\mathbf{v} = 0$.



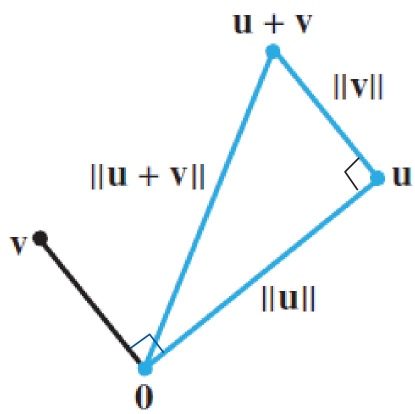
So, to generalise the idea of perpendicularity to $\mathbb{R}^n$ for $n > 2$, we make the following definition:
Definition: Two vectors $\mathbf{u}$ and $\mathbf{v}$ are *orthogonal* if $\mathbf{u}\cdot\mathbf{v} = 0$.
We also say $\mathbf{u}$ *is orthogonal to* $\mathbf{v}$.

Another way to see that orthogonality generalises perpendicularity:

Theorem 2: Pythagorean Theorem: Two vectors $\mathbf{u}$ and $\mathbf{v}$ are orthogonal if and only if $\|\mathbf{u}+\mathbf{v}\|^2 = \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2$.

Proof:

$$\begin{aligned}\|\mathbf{u}+\mathbf{v}\|^2 &= (\mathbf{u}+\mathbf{v})\cdot(\mathbf{u}+\mathbf{v})\\ &= \mathbf{u}\cdot\mathbf{u} + \mathbf{u}\cdot\mathbf{v} + \mathbf{v}\cdot\mathbf{u} + \mathbf{v}\cdot\mathbf{v}\\ &= \|\mathbf{u}\|^2 + 2\mathbf{u}\cdot\mathbf{v} + \|\mathbf{v}\|^2.\end{aligned}$$

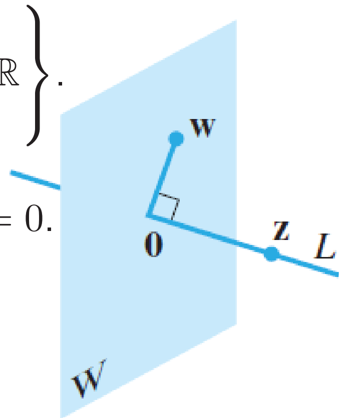


So $\|\mathbf{u}+\mathbf{v}\|^2 = \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2$ if and only if $\mathbf{u}\cdot\mathbf{v} = 0$.

Instead of $\mathbf{v}$ being orthogonal to just a single vector $\mathbf{u}$, we can consider orthogonality to a set of vectors:

Definition: Let W be a subspace of $\mathbb{R}^n$ (or more generally a subset).
A vector $\mathbf{z}$ is *orthogonal to W* if it is orthogonal to every vector in W.
The *orthogonal complement* of W, written $W^\perp$, is the set of all vectors orthogonal to W. In other words, $\mathbf{z}$ is in $W^\perp$ means $\mathbf{z} \cdot \mathbf{w} = 0$ for all $\mathbf{w}$ in W.

Example: Let W be the x_1x_3-plane in $\mathbb{R}^3$, i.e. $W = \left\{ \begin{bmatrix} a \\ 0 \\ b \end{bmatrix} \middle| a, b \in \mathbb{R} \right\}$.

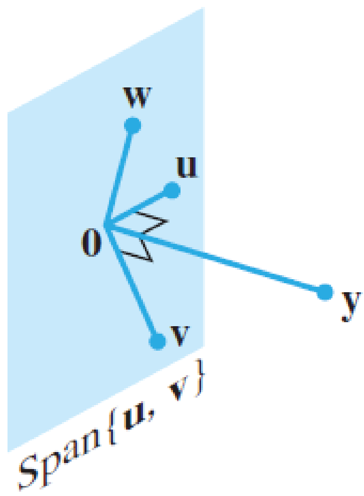


$\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$ is orthogonal to W, because $\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \cdot \begin{bmatrix} a \\ 0 \\ b \end{bmatrix} = 0 \cdot a + 1 \cdot 0 + 0 \cdot b = 0.$

We show on p13 that $W^\perp$ is Span $\left\{ \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \right\}$.

Key properties of $W^\perp$, for a subspace W of $\mathbb{R}^n$:

1. If $\mathbf{x}$ is in both W and $W^\perp$, then $\mathbf{x} = \mathbf{0}$ (ex. sheet #21 q2b).
2. If $W = \text{Span}\{\mathbf{v}_1, \dots, \mathbf{v}_p\}$, then $\mathbf{y}$ is in $W^\perp$ if and only if $\mathbf{y}$ is orthogonal to each $\mathbf{v}_i$ (same idea as ex. sheet q2a, see diagram).
3. $W^\perp$ is a subspace of $\mathbb{R}^n$ (checking the axioms directly is not hard, alternative proof p13).
4. $\dim W + \dim W^\perp = n$ (follows from alternative proof of 3, see p13).
5. If $W^\perp = U$, then $U^\perp = W$.
6. For a vector $\mathbf{y}$ in $\mathbb{R}^n$, the closest point in W to $\mathbf{y}$ is the unique point $\hat{\mathbf{y}}$ such that $\mathbf{y} - \hat{\mathbf{y}}$ is in $W^\perp$ (see p15-17).



(1 and 3 are true for any set W, even when W is not a subspace.)

Dot product and matrix multiplication:

Remember (week 2 p16, §1.4) the row-column method of matrix-vector multiplication:

Example: $\begin{bmatrix} 4 & 3 \\ 2 & 6 \\ 14 & 10 \end{bmatrix} \begin{bmatrix} -2 \\ 2 \end{bmatrix} = \begin{bmatrix} 4(-2)+3(2) \\ 2(-2)+6(2) \\ 14(-2)+10(2) \end{bmatrix} = \begin{bmatrix} -2 \\ 8 \\ -8 \end{bmatrix}.$

This last entry is $\begin{bmatrix} 14 \\ 10 \end{bmatrix} \cdot \begin{bmatrix} -2 \\ 2 \end{bmatrix}$.

In general,

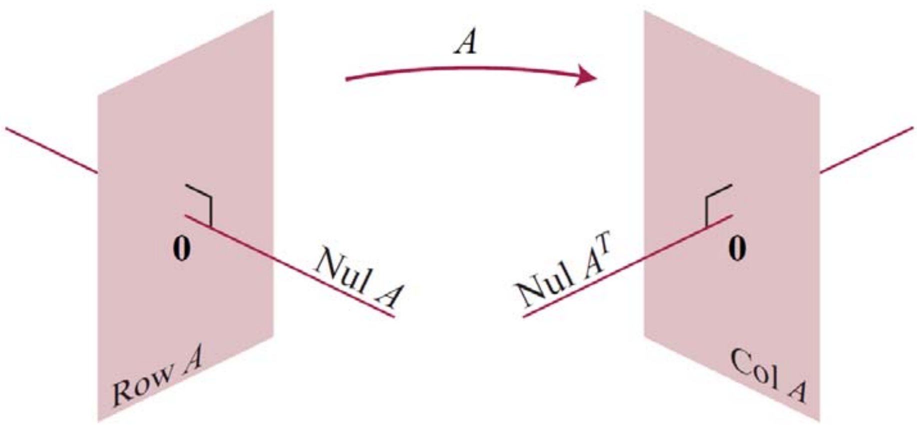
$$\begin{bmatrix} -- & \mathbf{r}_1 & -- \\ -- & \vdots & -- \\ -- & \mathbf{r}_m & -- \end{bmatrix} \mathbf{x} = \begin{bmatrix} \mathbf{r}_1 \cdot \mathbf{x} \\ \vdots \\ \mathbf{r}_m \cdot \mathbf{x} \end{bmatrix}.$$

So $\mathbf{x}$ is in the null space of A if and only if $\mathbf{r}_i \cdot \mathbf{x} = 0$ for every row $\mathbf{r}_i$ of A. By property 2 on the previous page, this precisely means $\mathbf{x}$ is in (Span $\{\mathbf{r}_1, \dots, \mathbf{r}_m\})^\perp$. So

Theorem 3: Orthogonality of Subspaces associated to Matrices:
(Row$A)^\perp$ = Nul$A$, and ...

Theorem 3: Orthogonality of Subspaces associated to Matrices: For a matrix A, (Row$A)^\perp$ = Nul$A$ and (Col$A)^\perp$ = NulA^T.

The second assertion comes from applying the first statement to A^T instead of A, remembering that RowA^T = Col$A$.



Theorem 3: Orthogonality of Subspaces associated to Matrices: For a matrix A, $(\text{Row}A)^\perp = \text{Nul}A$ and $(\text{Col}A)^\perp = \text{Nul}A^T$.

We can use this theorem to prove that $W^\perp$ is a subspace: given a subspace W of $\mathbb{R}^n$, let A be the matrix whose rows is a basis for W, so $\text{Row}A = W$. Then $W^\perp = \text{Nul}A$, and null spaces are subspaces, so $W^\perp$ is a subspace. Futhermore, the Rank Nullity Theorem says $\dim \text{Row}A + \dim \text{Nul}A = n$, so $\dim W + \dim W^\perp = n$.

The argument above also gives us a way to compute orthogonal complements:

Example: Let $W = \left\{ \begin{bmatrix} a \\ 0 \\ b \end{bmatrix} \middle| a, b \in \mathbb{R} \right\}$. A basis for W is $\left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \right\}$, so $W^\perp$ is the solutions to $\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \mathbf{x} = \mathbf{0}$, i.e. $W^\perp = \left\{ s \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \middle| s \in \mathbb{R} \right\}$.

Notice $\dim W + \dim W^\perp = 2 + 1 = 3$.

On p11, we related the matrix-vector product to the dot product:

$$\begin{bmatrix} -- & \mathbf{r}_1 & -- \\ -- & \vdots & -- \\ -- & \mathbf{r}_m & -- \end{bmatrix} \mathbf{x} = \begin{bmatrix} \mathbf{r}_1 \cdot \mathbf{x} \\ \vdots \\ \mathbf{r}_m \cdot \mathbf{x} \end{bmatrix}.$$

Because each column of a matrix-matrix product is a matrix-vector product,

$$AB = A \begin{bmatrix} | & & | \\ \mathbf{b}_1 & \dots & \mathbf{b}_p \\ | & & | \end{bmatrix} = \begin{bmatrix} | & & | \\ A\mathbf{b}_1 & \dots & A\mathbf{b}_p \\ | & & | \end{bmatrix},$$

we can also express matrix-matrix products in terms of the dot product: the (i, j)-entry of the product AB is (ith row of A) $\cdot$ (jth column of B)

$$\begin{bmatrix} -- & \mathbf{r}_1 & -- \\ -- & \vdots & -- \\ -- & \mathbf{r}_m & -- \end{bmatrix} \begin{bmatrix} | & & | \\ \mathbf{b}_1 & \dots & \mathbf{b}_p \\ | & & | \end{bmatrix} = \begin{bmatrix} \mathbf{r}_1 \cdot \mathbf{b}_1 & \dots & \mathbf{r}_1 \cdot \mathbf{b}_p \\ \vdots & & \vdots \\ \mathbf{r}_m \cdot \mathbf{b}_1 & \dots & \mathbf{r}_m \cdot \mathbf{b}_p \end{bmatrix}.$$

Closest point to a subspace:

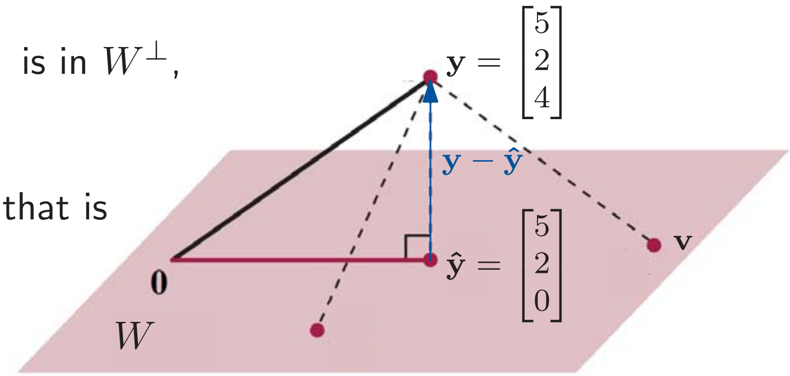
Theorem 9: Best Approximation Thoerem: Let W be a subspace of $\mathbb{R}^n$, and $\mathbf{y}$ a vector in $\mathbb{R}^n$. Then there is a unique point $\hat{\mathbf{y}}$ in W such that $\mathbf{y} - \hat{\mathbf{y}}$ is in $W^\perp$, and this $\hat{\mathbf{y}}$ is the closest point in W to $\mathbf{y}$ in the sense that $\|\mathbf{y} - \hat{\mathbf{y}}\| < \|\mathbf{y} - \mathbf{v}\|$ for all $\mathbf{v}$ in W with $\mathbf{v} \neq \hat{\mathbf{y}}$.

Example: Let $W = \text{Span}\left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \right\}$, so $W^\perp = \text{Span}\left\{ \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \right\}$. Let $\mathbf{y} = \begin{bmatrix} 5 \\ 2 \\ 4 \end{bmatrix}$.

Take $\hat{\mathbf{y}} = \begin{bmatrix} 5 \\ 2 \\ 0 \end{bmatrix}$, then $\mathbf{y} - \hat{\mathbf{y}} = \begin{bmatrix} 0 \\ 0 \\ 4 \end{bmatrix}$ is in $W^\perp$,

so $\hat{\mathbf{y}} = \begin{bmatrix} 5 \\ 2 \\ 0 \end{bmatrix}$ is unique point in W that is

closest to $\begin{bmatrix} 5 \\ 2 \\ 4 \end{bmatrix}$.

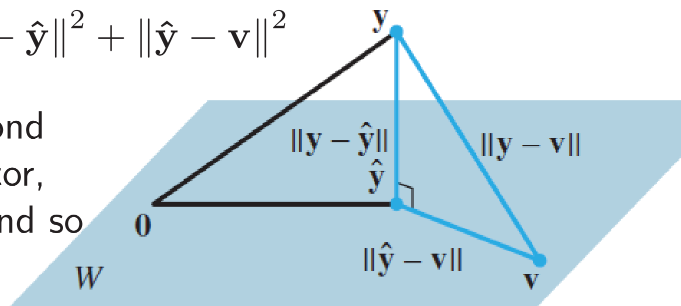


Theorem 9: Best Approximation Thoerem: Let W be a subspace of $\mathbb{R}^n$, and $\mathbf{y}$ a vector in $\mathbb{R}^n$. Then there is a unique point $\hat{\mathbf{y}}$ in W such that $\mathbf{y} - \hat{\mathbf{y}}$ is in $W^\perp$, and this $\hat{\mathbf{y}}$ is the closest point in W to $\mathbf{y}$ in the sense that $\|\mathbf{y} - \hat{\mathbf{y}}\| < \|\mathbf{y} - \mathbf{v}\|$ for all $\mathbf{v}$ in W with $\mathbf{v} \neq \hat{\mathbf{y}}$.

Partial Proof: We show here that, if $\mathbf{y} - \hat{\mathbf{y}}$ is in $W^\perp$, then $\hat{\mathbf{y}}$ is the unique closest point (i.e. it satisfies the inequality). We will not show here that there is always a $\hat{\mathbf{y}}$ such that $\mathbf{y} - \hat{\mathbf{y}}$ is in $W^\perp$. (See §6.3 on orthogonal projections, in Week 12 notes.)
We are assuming that $\mathbf{y} - \hat{\mathbf{y}}$ is in $W^\perp$. (vertical blue edge)
$\hat{\mathbf{y}} - \mathbf{v}$ is a difference of vectors in W, so it is in W. (horizontal blue edge)
So $\mathbf{y} - \hat{\mathbf{y}}$ and $\hat{\mathbf{y}} - \mathbf{v}$ are orthogonal. Apply the Pythagorean Theorem (blue triangle):

$$\|(\mathbf{y} - \hat{\mathbf{y}}) + (\hat{\mathbf{y}} - \mathbf{v})\|^2 = \|\mathbf{y} - \hat{\mathbf{y}}\|^2 + \|\hat{\mathbf{y}} - \mathbf{v}\|^2$$

The left hand side is $\|\mathbf{y} - \mathbf{v}\|^2$.
The right hand side: if $\mathbf{v} \neq \hat{\mathbf{y}}$, then the second term is the squared-length of a nonzero vector, so it is positive. So $\|\mathbf{y} - \mathbf{v}\|^2 > \|\mathbf{y} - \hat{\mathbf{y}}\|^2$ and so $\|\mathbf{y} - \mathbf{v}\| > \|\mathbf{y} - \hat{\mathbf{y}}\|$.



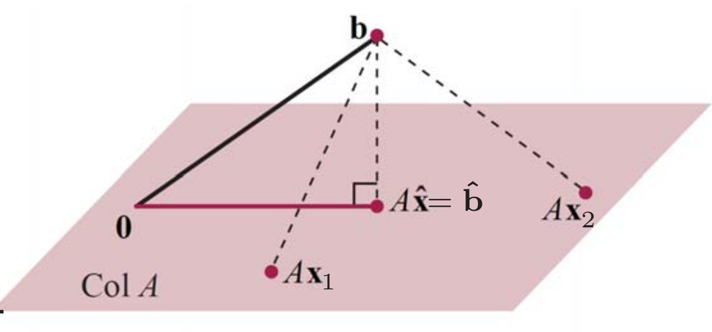
§6.5-6.6: Least Squares, Application to Regression

Remember our motivation: we have an inconsistent equation $A\mathbf{x} = \mathbf{b}$, and we want to find a "closest approximate solution" $\hat{\mathbf{x}}$ such that $A\hat{\mathbf{x}}$ is the point in Col$A$ that is closest to $\mathbf{b}$.

Definition: If A is an $m \times n$ matrix and $\mathbf{b}$ is in $\mathbb{R}^m$, then a *least-squares solution* of $A\mathbf{x} = \mathbf{b}$ is a vector $\hat{\mathbf{x}}$ in $\mathbb{R}^n$ such that $\|\mathbf{b} - A\hat{\mathbf{x}}\| \leq \|\mathbf{b} - A\mathbf{x}\|$ for all $\mathbf{x}$ in $\mathbb{R}^n$.

Equivalently: we want to find a vector $\hat{\mathbf{b}}$ in Col$A$ that is closest to $\mathbf{b}$, and then solve $A\hat{\mathbf{x}} = \hat{\mathbf{b}}$.

Because of the Best Approximation Theorem (p15-16): $\mathbf{b} - \hat{\mathbf{b}}$ is in $(\text{Col}A)^\perp$.
Because of Orthogonality of Subspaces associated to Matrices (p11-13): $(\text{Col}A)^\perp = \text{Nul}A^T$.
So we need $\hat{\mathbf{b}}$ so that $\mathbf{b} - \hat{\mathbf{b}}$ is in NulA^T.



The least-squares solutions to $A\mathbf{x} = \mathbf{b}$ are the solutions to $A\hat{\mathbf{x}} = \hat{\mathbf{b}}$ where $\hat{\mathbf{b}}$ is the unique vector such that $\mathbf{b} - \hat{\mathbf{b}}$ is in NulA^T.
Equivalently,

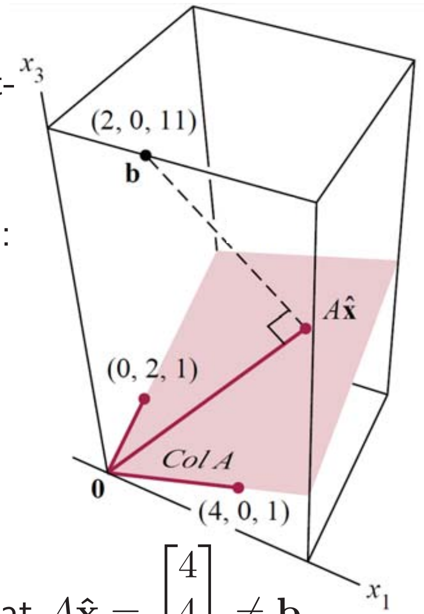
$$A^T(\mathbf{b} - \hat{\mathbf{b}}) = \mathbf{0}$$
$$A^T\mathbf{b} - A^T\hat{\mathbf{b}} = \mathbf{0}$$
$$A^T\mathbf{b} = A^T\hat{\mathbf{b}}$$
$$A^T\mathbf{b} = A^TA\hat{\mathbf{x}}$$

So we have proved:
Theorem 13: Least-Squares Theorem: The set of least-squares solutions of $A\mathbf{x} = \mathbf{b}$ is the set of solutions of the normal equations $A^TA\hat{\mathbf{x}} = A^T\mathbf{b}$.

Because of the existence part of the Best Approximation Theorem (that we will prove later), $A^TA\hat{\mathbf{x}} = A^T\mathbf{b}$ is always consistent.
Warning: The terminology is confusing: a least-squares solution $\hat{\mathbf{x}}$, satisfying $A^TA\hat{\mathbf{x}} = A^T\mathbf{b}$, is in general not a solution to $A\mathbf{x} = \mathbf{b}$. That is, usually $A\hat{\mathbf{x}} \neq \mathbf{b}$.

Theorem 13: Least-Squares Theorem: The set of least-squares solutions of $A\mathbf{x} = \mathbf{b}$ is the set of solutions of the normal equations $A^T A\hat{\mathbf{x}} = A^T\mathbf{b}$.

Example: Let $A = \begin{bmatrix} 4 & 0 \\ 0 & 2 \\ 1 & 1 \end{bmatrix}$ and $\mathbf{b} = \begin{bmatrix} 2 \\ 0 \\ 11 \end{bmatrix}$. Find a least-squares solution of the inconsistent equation $A\mathbf{x} = \mathbf{b}$.



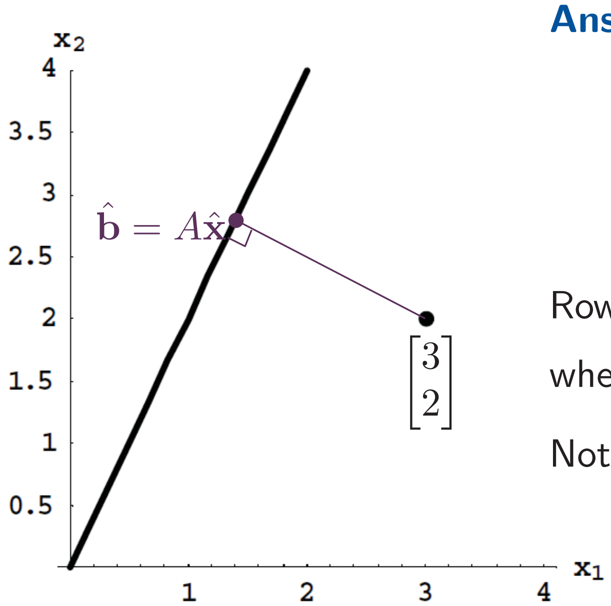
Answer: We solve the normal equations $A^T A\hat{\mathbf{x}} = A^T\mathbf{b}$:

$$\begin{bmatrix} 4 & 0 & 1 \\ 0 & 2 & 1 \end{bmatrix} \begin{bmatrix} 4 & 0 \\ 0 & 2 \\ 1 & 1 \end{bmatrix} \hat{\mathbf{x}} = \begin{bmatrix} 4 & 0 & 1 \\ 0 & 2 & 1 \end{bmatrix} \begin{bmatrix} 2 \\ 0 \\ 11 \end{bmatrix}$$

$$\begin{bmatrix} 17 & 1 \\ 1 & 5 \end{bmatrix} \hat{\mathbf{x}} = \begin{bmatrix} 19 \\ 11 \end{bmatrix}$$

By row-reducing $\left[\begin{array}{cc|c} 17 & 1 & 19 \\ 1 & 5 & 11 \end{array}\right]$, we find $\hat{\mathbf{x}} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$. Note that $A\hat{\mathbf{x}} = \begin{bmatrix} 4 \\ 4 \\ 3 \end{bmatrix} \neq \mathbf{b}$.

Example: (from p1) Let $A = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$ and $\mathbf{b} = \begin{bmatrix} 3 \\ 2 \end{bmatrix}$. Find the set of least-squares solutions of the inconsistent equation $A\mathbf{x} = \mathbf{b}$.



Answer: We solve $A^T A\hat{\mathbf{x}} = A^T\mathbf{b}$:

$$\begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix} \hat{\mathbf{x}} = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix} \begin{bmatrix} 3 \\ 2 \end{bmatrix}$$

$$\begin{bmatrix} 5 & 10 \\ 10 & 20 \end{bmatrix} \hat{\mathbf{x}} = \begin{bmatrix} 7 \\ 14 \end{bmatrix}$$

Row-reducing $\left[\begin{array}{cc|c} 5 & 10 & 7 \\ 10 & 20 & 14 \end{array}\right]$ gives $\hat{\mathbf{x}} = \begin{bmatrix} 7/5 \\ 0 \end{bmatrix} + s\begin{bmatrix} -2 \\ 1 \end{bmatrix}$ where s can take any value.

Note that $A\hat{\mathbf{x}} = A\left(\begin{bmatrix} 7/5 \\ 0 \end{bmatrix} + s\begin{bmatrix} -2 \\ 1 \end{bmatrix}\right) = \begin{bmatrix} 7/5 \\ 14/5 \end{bmatrix}$, independent of s: $A\hat{\mathbf{x}}$ is the closest point in Col$A$ to $\mathbf{b}$, which by the Best Approximation Theorem is unique.

Observations from the previous examples:

- $A^T A$ is a square matrix and is symmetric. (Exercise: prove it!)
- The normal equations sometimes have a unique solution and sometimes have infinitely many solutions, but $A\hat{\mathbf{x}}$ is unique.

When is the least-squares solution unique?

Theorem 14: Uniqueness of Least-Squares Solutions: The equation $A\mathbf{x} = \mathbf{b}$ has a unique least-squares solution if and only if the columns of A are linearly independent.

Consequences:

- The number of least-squares solutions to $A\mathbf{x} = \mathbf{b}$ does not depend on $\mathbf{b}$, only on A.
- Because $A^T A$ is a square matrix, if the least-squares solution is unique, then it is $\hat{\mathbf{x}} = (A^T A)^{-1} A^T \mathbf{b}$. This formula is useful theoretically (e.g. for deriving general expressions for regression coefficients, see Homework 6 q5).

Theorem 14: Uniqueness of Least-Squares Solutions: The equation $A\mathbf{x} = \mathbf{b}$ has a unique least-squares solution if and only if the columns of A are linearly independent.

Proof 1: The least-squares solutions are the solutions to the normal equations $A^T A\hat{\mathbf{x}} = A^T \mathbf{b}$. So

- "unique least-squares solution" is equivalent to Nul$(A^T A) = \{\mathbf{0}\}$.
- "columns of A are linearly independent" is equivalent to Nul$A = \{\mathbf{0}\}$.

So the theorem will follow if we prove the stronger fact Nul$(A^T A) =$ Nul$A$; in other words, $A^T A\mathbf{x} = \mathbf{0}$ if and only if $A\mathbf{x} = \mathbf{0}$.

- If $A\mathbf{x} = \mathbf{0}$, then $A^T A\mathbf{x} = A^T(A\mathbf{x}) = A^T\mathbf{0} = \mathbf{0}$.
- If $A^T A\mathbf{x} = \mathbf{0}$, then $\|A\mathbf{x}\|^2 = (A\mathbf{x}) \cdot (A\mathbf{x}) = (A\mathbf{x})^T(A\mathbf{x}) = \mathbf{x}^T A^T A\mathbf{x}$ $= \mathbf{x}^T(A^T A\mathbf{x}) = \mathbf{x}^T\mathbf{0} = 0$. So the length of $A\mathbf{x}$ is 0, which means it must be the zero vector.

Proof 2: The least-squares solutions are the solutions to $A\hat{\mathbf{x}} = \hat{\mathbf{b}}$ where $\hat{\mathbf{b}}$ is unique (the closest point in Col$A$ to $\mathbf{b}$). The equation $A\hat{\mathbf{x}} = \hat{\mathbf{b}}$ has a unique solution precisely when the columns of A are linearly independent.

Application: least-squares line

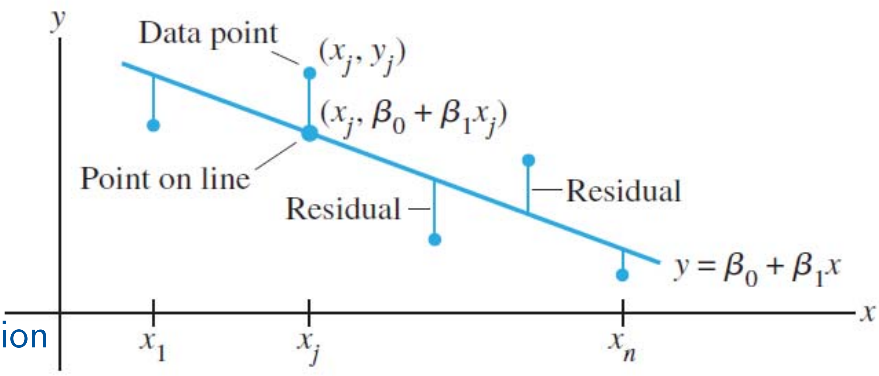
Suppose we have a model that relates two quantities x and y linearly, i.e. we expect $y = \beta_0 + \beta_1 x$, for some unknown numbers β_0, β_1.
To estimate β_0 and β_1, we do an experiment, whose results are $(x_1, y_1), \ldots, (x_n, y_n)$.
Now we wish to solve (for β_0, β_1):

$$\begin{aligned} \beta_0 + \beta_1 x_1 &= y_1 \\ \beta_0 + \beta_1 x_2 &= y_2 \\ \vdots \quad & \quad \vdots \\ \beta_0 + \beta_1 x_n &= y_n \end{aligned} \qquad \text{i.e.} \quad \underset{\substack{\| \\ \text{design matrix}}}{\underbrace{\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}}_{X}} \underset{\substack{\| \\ \text{parameter vector}}}{\underbrace{\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}}_{\boldsymbol{\beta}}} = \underset{\substack{\| \\ \text{observation vector}}}{\underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}}_{\mathbf{y}}}$$

$A\mathbf{x} = \mathbf{b}$ with different notation

We wish to solve (for β_0, β_1):

$$\underset{\substack{\| \\ \text{design matrix}}}{\underbrace{\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}}_{X}} \underset{\substack{\| \\ \text{parameter vector}}}{\underbrace{\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}}_{\boldsymbol{\beta}}} = \underset{\substack{\| \\ \text{observation vector}}}{\underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}}_{\mathbf{y}}}$$



Because experiments are rarely perfect, our data points (x_i, y_i) probably don't all lie exactly on any line, i.e. this system probably doesn't have a solution. So we ask for a least-squares solution.

A least-squares solution minimises $\|\mathbf{y} - X\boldsymbol{\beta}\|$, which is equivalent to minimising $\|\mathbf{y} - X\boldsymbol{\beta}\|^2 = (y_1 - (\beta_0 + \beta_1 x_1))^2 + \cdots + (y_n - (\beta_0 + \beta_1 x_n))^2$, the sums of the squares of the residuals. (The residuals are the vertical distances between each data point and the line, as in the diagram above).

Example: Find the equation $y = \hat{\beta}_0 + \hat{\beta}_1 x$ for the least-squares line for the following data points:

x_i	2	5	7	8
y_i	1	2	3	3

Answer: The model $X\boldsymbol{\beta} = \mathbf{y}$ is

$$\begin{bmatrix} 1 & 2 \\ 1 & 5 \\ 1 & 7 \\ 1 & 8 \end{bmatrix} \boldsymbol{\beta} = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 3 \end{bmatrix}$$

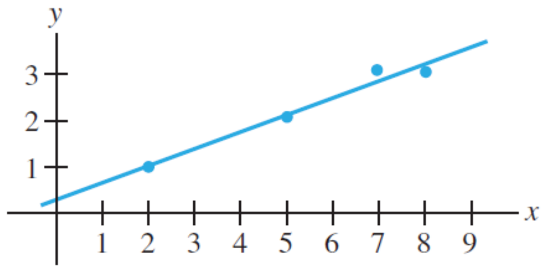
The normal equations $X^T X \hat{\boldsymbol{\beta}} = X^T \mathbf{y}$ are

$$\begin{bmatrix} 1 & 1 & 1 & 1 \\ 2 & 5 & 7 & 8 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 1 & 5 \\ 1 & 7 \\ 1 & 8 \end{bmatrix} \hat{\boldsymbol{\beta}} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 2 & 5 & 7 & 8 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 3 \\ 3 \end{bmatrix}$$

$$\begin{bmatrix} 4 & 22 \\ 22 & 142 \end{bmatrix} \hat{\boldsymbol{\beta}} = \begin{bmatrix} 9 \\ 57 \end{bmatrix}.$$

We wish to solve (for β_0, β_1):

$$\underset{X}{\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}} \underset{\boldsymbol{\beta}}{\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}} = \underset{\mathbf{y}}{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}}$$



Row-reducing gives $\hat{\boldsymbol{\beta}} = \begin{bmatrix} 2/7 \\ 5/14 \end{bmatrix}$, so the equation of the least-squares line is $y = 2/7 + 5/14x$.

Application: least-squares fitting of other curves

Suppose we model y as a more complicated function of x, i.e. $y = \beta_0 f_0(x) + \beta_1 f_1(x) + \cdots + \beta_k f_k(x)$, where $f_0, \ldots, f_k$ are known functions, and $\beta_0, \ldots, \beta_k$ are unknown parameters that we will estimate from experimental data. Such a model is still called a "linear model", because it is linear in the parameters $\beta_0, \ldots, \beta_k$.

Example: Estimate the parameters $\beta_1, \beta_2, \beta_3$ in the model $y = \beta_1 x + \beta_2 x^2 + \beta_3 x^3$, given the data

x_i	2	3	4	6	7
y_i	1.6	2.0	2.5	3.1	3.4

Answer: The model equations are $\beta_1 2 + \beta_2 2^2 + \beta_3 2^3 = 1.6$
$\beta_1 3 + \beta_2 3^2 + \beta_3 3^3 = 2.0$, and so on.

In matrix form: $\begin{bmatrix} 2 & 4 & 8 \\ 3 & 9 & 27 \\ 4 & 16 & 64 \\ 6 & 36 & 216 \\ 7 & 49 & 343 \end{bmatrix} \boldsymbol{\beta} = \begin{bmatrix} 1.6 \\ 2.0 \\ 2.5 \\ 3.1 \\ 3.4 \end{bmatrix}$. Then we solve the normal equations etc...

So in general, to estimate the parameters $\beta_0, \dots, \beta_k$ in a linear model $y = \beta_0 f_0(x) + \beta_1 f_1(x) + \dots + \beta_k f_k(x)$, we find the least-squares solution to

$$\begin{aligned}\beta_0 f_0(x_1) + \beta_1 f_1(x_1) + \dots + \beta_k f_k(x_1) &= y_1 \\ \beta_0 f_0(x_2) + \beta_1 f_1(x_2) + \dots + \beta_k f_k(x_2) &= y_2 \\ \vdots \qquad\qquad\qquad\qquad & \vdots \;\; \vdots\end{aligned}$$

more general design matrix

parameter vector with more rows

i.e. $$\begin{bmatrix} f_0(x_1) & f_1(x_1) & \dots & f_k(x_1) \\ f_0(x_2) & f_1(x_2) & \dots & f_k(x_2) \\ \vdots & \vdots & & \vdots \\ f_0(x_n) & f_1(x_n) & \dots & f_k(x_n) \end{bmatrix} \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_k \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

same observation vector

(Least-squares lines correspond to the case $f_0(x) = 1, f_1(x) = x$.)

Least-squares techniques can also be used to fit a surface to experimental data, for linear models with more than one input variable (e.g. $y = \beta_0 + \beta_1 x + \beta_2 xw$, for input variables x and w) - this is called multiple regression.