Lecture 34 (?): Least squares and linear models

Danny W. Crytser

May 21, 2014

Today's lecture

1. We'll talk about how to obtain $\text{proj}_W \mathbf{v}$ using orthonormal bases.
2. We'll introduce the least-squares approximation problem.
3. We'll look at a few applications of least-squares approximation.

Orthogonal matrices

We have discussed finding projections of vectors on subspaces.

Orthogonal matrices

We have discussed finding projections of vectors on subspaces.

Theorem

If $W \subset \mathbb{R}^n$ is a subspace and $B = \{\mathbf{b}_1, \dots, \mathbf{b}_p\}$ is an orthogonal basis for W, then for any $\mathbf{v} \in \mathbb{R}^n$:

Orthogonal matrices

We have discussed finding projections of vectors on subspaces.

Theorem

If $W \subset \mathbb{R}^n$ is a subspace and $B = \{\mathbf{b}_1, \ldots, \mathbf{b}_p\}$ is an orthogonal basis for W, then for any $\mathbf{v} \in \mathbb{R}^n$:

1. *the projection of $\mathbf{v}$ on W is given by*

$$\text{proj}_W \mathbf{v} = \frac{\mathbf{v} \cdot \mathbf{b}_1}{\mathbf{b}_1 \cdot \mathbf{b}_1}\mathbf{b}_1 + \frac{\mathbf{v} \cdot \mathbf{b}_2}{\mathbf{b}_2 \cdot \mathbf{b}_2}\mathbf{b}_2 + \ldots + \frac{\mathbf{v} \cdot \mathbf{b}_p}{\mathbf{b}_p \cdot \mathbf{b}_p}\mathbf{b}_p.$$

   The vector $\text{proj}_W \mathbf{v}$ belongs to W and is the closest vector in W to $\mathbf{v}$.

Orthogonal matrices

We have discussed finding projections of vectors on subspaces.

Theorem

If $W \subset \mathbb{R}^n$ is a subspace and $B = \{\mathbf{b}_1, \ldots, \mathbf{b}_p\}$ is an orthogonal basis for W, then for any $\mathbf{v} \in \mathbb{R}^n$:

1. *the projection of $\mathbf{v}$ on W is given by*

$$\text{proj}_W \mathbf{v} = \frac{\mathbf{v} \cdot \mathbf{b}_1}{\mathbf{b}_1 \cdot \mathbf{b}_1}\mathbf{b}_1 + \frac{\mathbf{v} \cdot \mathbf{b}_2}{\mathbf{b}_2 \cdot \mathbf{b}_2}\mathbf{b}_2 + \ldots + \frac{\mathbf{v} \cdot \mathbf{b}_p}{\mathbf{b}_p \cdot \mathbf{b}_p}\mathbf{b}_p.$$

   The vector $\text{proj}_W \mathbf{v}$ belongs to W and is the closest vector in W to $\mathbf{v}$.

2. *the distance from $\mathbf{v}$ to W is*

$$\text{dist}(\mathbf{v}, W) = ||\mathbf{v} - \text{proj}_W \mathbf{v}||.$$

We have discussed finding projections of vectors on subspaces.

Theorem

If $W \subset \mathbb{R}^n$ is a subspace and $B = \{\mathbf{b}_1, \ldots, \mathbf{b}_p\}$ is an orthogonal basis for W, then for any $\mathbf{v} \in \mathbb{R}^n$:

1. *the projection of $\mathbf{v}$ on W is given by*

$$\text{proj}_W \mathbf{v} = \frac{\mathbf{v} \cdot \mathbf{b}_1}{\mathbf{b}_1 \cdot \mathbf{b}_1}\mathbf{b}_1 + \frac{\mathbf{v} \cdot \mathbf{b}_2}{\mathbf{b}_2 \cdot \mathbf{b}_2}\mathbf{b}_2 + \ldots + \frac{\mathbf{v} \cdot \mathbf{b}_p}{\mathbf{b}_p \cdot \mathbf{b}_p}\mathbf{b}_p.$$

   The vector $\text{proj}_W \mathbf{v}$ belongs to W and is the closest vector in W to $\mathbf{v}$.

2. *the distance from $\mathbf{v}$ to W is*

$$\text{dist}(\mathbf{v}, W) = ||\mathbf{v} - \text{proj}_W \mathbf{v}||.$$

Orthogonal matrices

We have discussed finding projections of vectors on subspaces.

Theorem

If $W \subset \mathbb{R}^n$ is a subspace and $B = \{\mathbf{b}_1, \ldots, \mathbf{b}_p\}$ is an orthogonal basis for W, then for any $\mathbf{v} \in \mathbb{R}^n$:

1. *the projection of $\mathbf{v}$ on W is given by*

$$\text{proj}_W \mathbf{v} = \frac{\mathbf{v} \cdot \mathbf{b}_1}{\mathbf{b}_1 \cdot \mathbf{b}_1}\mathbf{b}_1 + \frac{\mathbf{v} \cdot \mathbf{b}_2}{\mathbf{b}_2 \cdot \mathbf{b}_2}\mathbf{b}_2 + \ldots + \frac{\mathbf{v} \cdot \mathbf{b}_p}{\mathbf{b}_p \cdot \mathbf{b}_p}\mathbf{b}_p.$$

   The vector $\text{proj}_W \mathbf{v}$ belongs to W and is the closest vector in W to $\mathbf{v}$.

2. *the distance from $\mathbf{v}$ to W is*

$$\text{dist}(\mathbf{v}, W) = ||\mathbf{v} - \text{proj}_W \mathbf{v}||.$$

When the basis is orthonormal then the formula becomes simpler–the denominators are all 1.

We have discussed finding projections of vectors on subspaces.

Theorem

If $W \subset \mathbb{R}^n$ is a subspace and $B = \{\mathbf{b}_1, \ldots, \mathbf{b}_p\}$ is an orthonormal basis for W, then for any $\mathbf{v} \in \mathbb{R}^n$:

1. *the projection of $\mathbf{v}$ on W is given by*

$$\operatorname{proj}_W \mathbf{v} = (\mathbf{v} \cdot \mathbf{b}_1)\mathbf{b}_1 + (\mathbf{v} \cdot \mathbf{b}_2)\mathbf{b}_2 + \ldots + (\mathbf{v} \cdot \mathbf{b}_p)\mathbf{b}_p.$$

   The vector $\operatorname{proj}_W \mathbf{v}$ belongs to W and is the closest vector in W to $\mathbf{v}$.

2. *the distance from $\mathbf{v}$ to W is*

$$\operatorname{dist}(\mathbf{v}, W) = ||\mathbf{v} - \operatorname{proj}_W \mathbf{v}||.$$

Orthogonal matrices

We can further simplify this computation. Let $B = \{\mathbf{b}_1, \ldots, \mathbf{b}_p\}$ be an *orthonormal* basis for W, and form the matrix $U = [\mathbf{b}_1 \ldots \mathbf{b}_p]$. This matrix has orthonormal columns, so $U^T U = I_p$. The matrix UU^T usually does not equal I_n, but it yields useful information.

Orthogonal matrices

We can further simplify this computation. Let $B = \{\mathbf{b}_1, \ldots, \mathbf{b}_p\}$ be an *orthonormal* basis for W, and form the matrix $U = [\mathbf{b}_1 \ldots \mathbf{b}_p]$. This matrix has orthonormal columns, so $U^T U = I_p$. The matrix UU^T usually does not equal I_n, but it yields useful information.

Theorem

Let $W \subset \mathbb{R}^n$ and let $B = \{\mathbf{b}_1, \ldots, \mathbf{b}_p\}$ be an orthonormal basis for W, and form the matrix $U = [\mathbf{b}_1 \ldots \mathbf{b}_p]$. Then for any $\mathbf{v} \in \mathbb{R}^n$, we have

$$\mathrm{proj}_W \mathbf{v} = (UU^T)\mathbf{v}.$$

Orthogonal matrices

We can further simplify this computation. Let $B = \{\mathbf{b}_1, \ldots, \mathbf{b}_p\}$ be an *orthonormal* basis for W, and form the matrix $U = [\mathbf{b}_1 \ldots \mathbf{b}_p]$. This matrix has orthonormal columns, so $U^T U = I_p$. The matrix UU^T usually does not equal I_n, but it yields useful information.

Theorem

Let $W \subset \mathbb{R}^n$ and let $B = \{\mathbf{b}_1, \ldots, \mathbf{b}_p\}$ be an orthonormal basis for W, and form the matrix $U = [\mathbf{b}_1 \ldots \mathbf{b}_p]$. Then for any $\mathbf{v} \in \mathbb{R}^n$, we have

$$\mathrm{proj}_W \mathbf{v} = (UU^T)\mathbf{v}.$$

That is, to project a vector $\mathbf{v}$ onto the subspace W, one need only multiply it on the left by the matrix UU^T.

Proof.

Notice that if $\mathbf{v} \in \mathbb{R}^n$ then $U^T\mathbf{v} = \begin{bmatrix} \mathbf{b}_1 \cdot \mathbf{v} \\ \mathbf{b}_2 \cdot \mathbf{v} \\ \vdots \\ \mathbf{b}_p \cdot \mathbf{v} \end{bmatrix}$, by the row-column rule for multiplying matrices.

Orthogonal matrices

Proof.

Notice that if $\mathbf{v} \in \mathbb{R}^n$ then $U^T\mathbf{v} = \begin{bmatrix} \mathbf{b}_1 \cdot \mathbf{v} \\ \mathbf{b}_2 \cdot \mathbf{v} \\ \vdots \\ \mathbf{b}_p \cdot \mathbf{v} \end{bmatrix}$, by the row-column rule for multiplying matrices. Then

$$\begin{aligned} UU^T\mathbf{v} &= U(U^T\mathbf{v}) \quad \text{(associative)} \\ &= U \begin{bmatrix} \mathbf{b}_1 \cdot \mathbf{v} \\ \mathbf{b}_2 \cdot \mathbf{v} \\ \vdots \\ \mathbf{b}_p \cdot \mathbf{v} \end{bmatrix} \\ &= (\mathbf{b}_1 \cdot \mathbf{v})\mathbf{b}_1 + \ldots + (\mathbf{b}_p \cdot \cdot \mathbf{v})\mathbf{b}_p \quad \text{(def. matrix-vector mult.)} \\ &= \text{proj}_W \mathbf{v} \quad \text{(theorem)} \end{aligned}$$

□

Orthogonal matrices

We can use this result to find distances to subspaces.

We can use this result to find distances to subspaces. Let

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 0 & -1 & 2 \\ 1 & -2 & 2 \\ 1 & 0 & -5 \end{bmatrix}$$

and let $\mathbf{b} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$.

Orthogonal matrices

We can use this result to find distances to subspaces. Let

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 0 & -1 & 2 \\ 1 & -2 & 2 \\ 1 & 0 & -5 \end{bmatrix}$$

and let $\mathbf{b} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$. Then you can check that $\mathbf{b} \notin \operatorname{col} A$; that is, the system $A\mathbf{x} = \mathbf{b}$ is inconsistent.

Orthogonal matrices

We can use this result to find distances to subspaces. Let

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 0 & -1 & 2 \\ 1 & -2 & 2 \\ 1 & 0 & -5 \end{bmatrix}$$

and let $\mathbf{b} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$. Then you can check that $\mathbf{b} \notin \operatorname{col} A$; that is, the system $A\mathbf{x} = \mathbf{b}$ is inconsistent.Let's find the closest vector in $\operatorname{col} A$ to $\mathbf{b}$.

Orthogonal matrices

The columns of A are orthogonal but they are not orthonormal—the length of each vector isn't 1.

Orthogonal matrices

The columns of A are orthogonal but they are not orthonormal—the length of each vector isn't 1. We scale each column by the reciprocal of its length, obtaining a new matrix U with orthonormal columns

$$U = \begin{bmatrix} 1/\sqrt{3} & 2/3 & 3/\sqrt{42} \\ 0 & -1/3 & 2/\sqrt{42} \\ 1/\sqrt{3} & -2/3 & 2/\sqrt{42} \\ 1/\sqrt{3} & 0 & -5/\sqrt{42} \end{bmatrix}$$

Now we can form the product UU^T:

$$U^T = \begin{bmatrix} 1/\sqrt{3} & 0 & 1/\sqrt{3} & 1/\sqrt{3} \\ 2/3 & -1/3 & -2/3 & 0 \\ 3/\sqrt{42} & 2/\sqrt{42} & 2/\sqrt{42} & -5/\sqrt{42} \end{bmatrix}$$

Orthogonal matrices

Now we can form the product UU^T:

$$U^T = \begin{bmatrix} 1/\sqrt{3} & 0 & 1/\sqrt{3} & 1/\sqrt{3} \\ 2/3 & -1/3 & -2/3 & 0 \\ 3/\sqrt{42} & 2/\sqrt{42} & 2/\sqrt{42} & -5/\sqrt{42} \end{bmatrix}$$

$$UU^T = \begin{bmatrix} 125/126 & -5/63 & 2/63 & -1/42 \\ -5/63 & 13/63 & 20/63 & -5/21 \\ 2/63 & 20/63 & 55/63 & 2/21 \\ -1/42 & -5/21 & 2/21 & 13/14 \end{bmatrix}.$$

Orthogonal matrices

Now that we know UU^T we can find $\text{proj}_W \mathbf{v}$

Orthogonal matrices

Now that we know UU^T we can find $\text{proj}_W\,\mathbf{v}$

$$\begin{aligned}\text{proj}_W\,\mathbf{v} &= UU^T\mathbf{v}\\ &= \begin{bmatrix}125/126 & -5/63 & 2/63 & -1/42\\ -5/63 & 13/63 & 20/63 & -5/21\\ 2/63 & 20/63 & 55/63 & 2/21\\ -1/42 & -5/21 & 2/21 & 13/14\end{bmatrix}\begin{bmatrix}1\\1\\1\\1\end{bmatrix}\\ &= \begin{bmatrix}58/63\\13/63\\83/63\\16/21\end{bmatrix}\end{aligned}$$

Thus the distance from $\mathbf{v}$ to W the distance from $\mathbf{v}$ to $\text{proj}_W\,\mathbf{v}$. This is

$$\text{dist}\left(\mathbf{v}, \begin{bmatrix}58/63\\13/63\\83/63\\16/21\end{bmatrix}\right) = ||(5/63, 50/63, -20/63, 5/21)|| \approx 0.891$$

Orthogonal matrices

Now that we know UU^T we can find $\text{proj}_W\,\mathbf{v}$

$$\begin{aligned}\text{proj}_W\,\mathbf{v} \\ &= \begin{bmatrix} 125/126 & -5/63 & 2/63 & -1/42 \\ -5/63 & 13/63 & 20/63 & -5/21 \\ 2/63 & 20/63 & 55/63 & 2/21 \\ -1/42 & -5/21 & 2/21 & 13/14 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \\ &= \begin{bmatrix} 58/63 \\ 13/63 \\ 83/63 \\ 16/21 \end{bmatrix}\end{aligned}$$

Thus the distance from $\mathbf{v}$ to W the distance from $\mathbf{v}$ to $\text{proj}_W\,\mathbf{v}$. This is

$$\text{dist}\left(\mathbf{v}, \begin{bmatrix} 58/63 \\ 13/63 \\ 83/63 \\ 16/21 \end{bmatrix}\right) = ||(5/63, 50/63, -20/63, 5/21)|| \approx 0.891$$

Orthogonal matrices

Now that we know UU^T we can find $\text{proj}_W \mathbf{v}$

$$\text{proj}_W \mathbf{v}$$

$$= \begin{bmatrix} 58/63 \\ 13/63 \\ 83/63 \\ 16/21 \end{bmatrix}$$

Thus the distance from $\mathbf{v}$ to W the distance from $\mathbf{v}$ to $\text{proj}_W \mathbf{v}$. This is

$$\text{dist}\left(\mathbf{v}, \begin{bmatrix} 58/63 \\ 13/63 \\ 83/63 \\ 16/21 \end{bmatrix}\right) = ||(5/63, 50/63, -20/63, 5/21)|| \approx 0.891$$

Orthogonal matrices

Now that we know UU^T we can find $\text{proj}_W \mathbf{v}$

$$\begin{aligned}\text{proj}_W \mathbf{v} &= UU^T\mathbf{v}\\ &= \begin{bmatrix} 125/126 & -5/63 & 2/63 & -1/42 \\ -5/63 & 13/63 & 20/63 & -5/21 \\ 2/63 & 20/63 & 55/63 & 2/21 \\ -1/42 & -5/21 & 2/21 & 13/14 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \\ &= \begin{bmatrix} 58/63 \\ 13/63 \\ 83/63 \\ 16/21 \end{bmatrix}\end{aligned}$$

Thus the distance from $\mathbf{v}$ to W the distance from $\mathbf{v}$ to $\text{proj}_W \mathbf{v}$. This is

$$\text{dist}\left(\mathbf{v}, \begin{bmatrix} 58/63 \\ 13/63 \\ 83/63 \\ 16/21 \end{bmatrix}\right)$$

Now that we know UU^T we can find $\text{proj}_W \mathbf{v}$

$$\text{proj}_W \mathbf{v} = UU^T\mathbf{v}$$
$$= \begin{bmatrix} 125/126 & -5/63 & 2/63 & -1/42 \\ -5/63 & 13/63 & 20/63 & -5/21 \\ 2/63 & 20/63 & 55/63 & 2/21 \\ -1/42 & -5/21 & 2/21 & 13/14 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$
$$= \begin{bmatrix} 58/63 \\ 13/63 \\ 83/63 \\ 16/21 \end{bmatrix}$$

Thus the distance from $\mathbf{v}$ to W the distance from $\mathbf{v}$ to $\text{proj}_W \mathbf{v}$. This is

$$\text{dist}\left(\mathbf{v}, \begin{bmatrix} 58/63 \\ 13/63 \\ 83/63 \\ 16/21 \end{bmatrix}\right) = ||(5/63, 50/63, -20/63, 5/21)|| \approx 0.891$$

Least-squares

The idea of finding a vector in $\operatorname{col} A$ which is close to $\mathbf{b}$ given that $\mathbf{b}$ does not belong to $\operatorname{col} A$ leads to a general question.

Least-squares

The idea of finding a vector in $\operatorname{col} A$ which is close to $\mathbf{b}$ given that $\mathbf{b}$ does not belong to $\operatorname{col} A$ leads to a general question.

Definition

If A is an $m \times n$ matrix and $\mathbf{b} \in \mathbb{R}^m$, then a **least-squares solution** to $A\mathbf{x} = \mathbf{b}$ 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 $\mathbf{x} \in \mathbb{R}^n$.

The idea of finding a vector in $\operatorname{col} A$ which is close to $\mathbf{b}$ given that $\mathbf{b}$ does not belong to $\operatorname{col} A$ leads to a general question.

Definition

If A is an $m \times n$ matrix and $\mathbf{b} \in \mathbb{R}^m$, then a **least-squares solution** to $A\mathbf{x} = \mathbf{b}$ 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 $\mathbf{x} \in \mathbb{R}^n$.

The idea is that a least-squares solution is usually *not* a solution to $A\mathbf{x} = \mathbf{b}$ but it is as close as you can get to $\mathbf{b}$ with vectors of the form $A\mathbf{x}$.

Least-squares

Proposition

If A is an $m \times n$ matrix and $\mathbf{b} \in \mathbb{R}^m$, then

$$\hat{b} = \mathrm{proj}_{\mathrm{col}\,A}\,\mathbf{b}$$

belongs to $\mathrm{col}\,A$ *and any vector $\hat{x}$ with $A\hat{x} = \hat{b}$ is a least-squares solution to $A\mathbf{x} = \mathbf{b}$.*

This proposition says that there *are* least squares solutions but it doesn't give us a fast way to compute them.

Definition

Let A be an $m \times n$ matrix and let $\mathbf{b} \in \mathbb{R}^m$. Then

$$A^T A\mathbf{x} = A^T \mathbf{b}$$

is a consistent system of equations called the **normal equations** of the system $A\mathbf{x} = \mathbf{b}$.

Least-squares

Definition

Let A be an $m \times n$ matrix and let $\mathbf{b} \in \mathbb{R}^m$. Then

$$A^T A\mathbf{x} = A^T \mathbf{b}$$

is a consistent system of equations called the **normal equations** of the system $A\mathbf{x} = \mathbf{b}$.

The normal equations are useful mainly as another way to view the least-squares problem.

Definition

Let A be an $m \times n$ matrix and let $\mathbf{b} \in \mathbb{R}^m$. Then

$$A^T A\mathbf{x} = A^T\mathbf{b}$$

is a consistent system of equations called the **normal equations** of the system $A\mathbf{x} = \mathbf{b}$.

The normal equations are useful mainly as another way to view the least-squares problem.

Theorem

The least-squares solutions to $A\mathbf{x} = \mathbf{b}$ are exactly the solutions of the normal equations $A^T A\mathbf{x} = A^T\mathbf{b}$.

Least-squares

Theorem

The least-squares solutions to $A\mathbf{x} = \mathbf{b}$ *are exactly the solutions of the normal equations* $A^T A\mathbf{x} = A^T\mathbf{b}$.

Least-squares

Theorem

The least-squares solutions to $A\mathbf{x} = \mathbf{b}$ are exactly the solutions of the normal equations $A^T A\mathbf{x} = A^T\mathbf{b}$.

Proof.

The vector $\mathbf{x}$ is a least-squares solution if and only if $\mathbf{b} - A\mathbf{x}$ is orthogonal to the column space of A. But this means that each column $\mathbf{c}_i$ is orthogonal to $\mathbf{b} - A\mathbf{x}$. This is the same as $\mathbf{c}_i \cdot A\mathbf{x} = \mathbf{c}_i \cdot \mathbf{b}$. This is equivalent to $A^T(A\mathbf{x}) = A^T(\mathbf{b})$, by the row-column rule for computing matrix products. □

What the theorem means: If you want to find the least squares solutions to $A\mathbf{x} = \mathbf{b}$, you just have to find the (actual) solutions to $A^T A\mathbf{x} = A^T\mathbf{b}$.

Least-squares

Now that all the theorems are out of the way we can solve some least-squares problems.

Now that all the theorems are out of the way we can solve some least-squares problems.

Example

Let $A = \begin{bmatrix} 1 & 1 \\ 2 & 2 \end{bmatrix}$ and $\mathbf{b} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$. The system $A\mathbf{x} = \mathbf{b}$ is inconsistent, so we solve the least-squares solution.

Now that all the theorems are out of the way we can solve some least-squares problems.

Example

Let $A = \begin{bmatrix} 1 & 1 \\ 2 & 2 \end{bmatrix}$ and $\mathbf{b} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$. The system $A\mathbf{x} = \mathbf{b}$ is inconsistent, so we solve the least-squares solution. The least-squares solutions to $A\mathbf{x} = \mathbf{b}$ are the same as the (actual) solutions to $A^T A\mathbf{x} = A^T\mathbf{b}$. The product $A^T A$ is $\begin{bmatrix} 5 & 5 \\ 5 & 5 \end{bmatrix}$ and $A^T\mathbf{b} = \begin{bmatrix} 3 \\ 3 \end{bmatrix}$. The solutions to $A^T A\mathbf{x} = A^T\mathbf{b}$ are $\left\{ \begin{bmatrix} 3/5 + x \\ -x \end{bmatrix} : x \in \mathbb{R} \right\}$. These are the least-squares solutions to $A\mathbf{x} = \mathbf{b}$: each minimizes the error $||A\mathbf{x} - \mathbf{b}||$.

Least-squares

We find the least square solution to $A\mathbf{x} = \mathbf{b}$, where

$$A = \begin{bmatrix} 1 & -3 & -3 \\ 1 & 5 & 1 \\ 1 & 7 & 2 \end{bmatrix} \text{ and } \mathbf{b} = \begin{bmatrix} -3 \\ -65 \\ -28 \end{bmatrix}.$$

We find the least square solution to $A\mathbf{x} = \mathbf{b}$, where $A = \begin{bmatrix} 1 & -3 & -3 \\ 1 & 5 & 1 \\ 1 & 7 & 2 \end{bmatrix}$ and $\mathbf{b} = \begin{bmatrix} -3 \\ -65 \\ -28 \end{bmatrix}$. Here

$$A^T A = \begin{bmatrix} 3 & 9 & 0 \\ 9 & 83 & 28 \\ 0 & 28 & 14 \end{bmatrix}$$

and $A^T\mathbf{b} = \begin{bmatrix} -3 \\ -65 \\ -28 \end{bmatrix}$.

Least-squares

We find the least square solution to $A\mathbf{x} = \mathbf{b}$, where
$A = \begin{bmatrix} 1 & -3 & -3 \\ 1 & 5 & 1 \\ 1 & 7 & 2 \end{bmatrix}$ and $\mathbf{b} = \begin{bmatrix} -3 \\ -65 \\ -28 \end{bmatrix}$. Here

$$A^T A = \begin{bmatrix} 3 & 9 & 0 \\ 9 & 83 & 28 \\ 0 & 28 & 14 \end{bmatrix}$$

and $A^T\mathbf{b} = \begin{bmatrix} -3 \\ -65 \\ -28 \end{bmatrix}$.The general solution to $A^T A\mathbf{x} = A^T\mathbf{b}$ is
$x_1 = 2 + \frac{3}{2}x_3, x_2 = -1 - \frac{1}{2}x_3$ and x_3 free.

We find the least square solution to $A\mathbf{x} = \mathbf{b}$, where $A = \begin{bmatrix} 1 & -3 & -3 \\ 1 & 5 & 1 \\ 1 & 7 & 2 \end{bmatrix}$ and $\mathbf{b} = \begin{bmatrix} -3 \\ -65 \\ -28 \end{bmatrix}$. Here

$$A^T A = \begin{bmatrix} 3 & 9 & 0 \\ 9 & 83 & 28 \\ 0 & 28 & 14 \end{bmatrix}$$

and $A^T\mathbf{b} = \begin{bmatrix} -3 \\ -65 \\ -28 \end{bmatrix}$.The general solution to $A^T A\mathbf{x} = A^T\mathbf{b}$ is $x_1 = 2 + \frac{3}{2}x_3, x_2 = -1 - \frac{1}{2}x_3$ and x_3 free. We can set $x_3 = 0$ to get *a* least-squares solution:

$$\hat{\mathbf{x}} = \begin{bmatrix} 2 \\ -1 \\ 0 \end{bmatrix}.$$

Least-squares

In the previous example, the set of least-squares solutions was infinite. There is a theorem that describes when the least-squares solution to any system $A\mathbf{x} = \mathbf{b}$ is unique:

In the previous example, the set of least-squares solutions was infinite. There is a theorem that describes when the least-squares solution to any system $A\mathbf{x} = \mathbf{b}$ is unique:

Theorem

Let A be an $m \times n$ matrix. The following statements are equivalent:

Least-squares

In the previous example, the set of least-squares solutions was infinite. There is a theorem that describes when the least-squares solution to any system $A\mathbf{x} = \mathbf{b}$ is unique:

Theorem

Let A be an $m \times n$ matrix. The following statements are equivalent:

1. *for all $\mathbf{b} \in \mathbb{R}^m$, there is a unique least-squares solution to $A\mathbf{x} = \mathbf{b}$;*

In the previous example, the set of least-squares solutions was infinite. There is a theorem that describes when the least-squares solution to any system $A\mathbf{x} = \mathbf{b}$ is unique:

Theorem

Let A be an $m \times n$ matrix. The following statements are equivalent:

1. *for all $\mathbf{b} \in \mathbb{R}^m$, there is a unique least-squares solution to $A\mathbf{x} = \mathbf{b}$;*
2. *the columns of A are linearly independent;*

Least-squares

In the previous example, the set of least-squares solutions was infinite. There is a theorem that describes when the least-squares solution to any system $A\mathbf{x} = \mathbf{b}$ is unique:

Theorem

Let A be an $m \times n$ matrix. The following statements are equivalent:

1. *for all $\mathbf{b} \in \mathbb{R}^m$, there is a unique least-squares solution to $A\mathbf{x} = \mathbf{b}$;*
2. *the columns of A are linearly independent;*
3. *the matrix $A^T A$ is invertible.*

Least-squares

In the previous example, the set of least-squares solutions was infinite. There is a theorem that describes when the least-squares solution to any system $A\mathbf{x} = \mathbf{b}$ is unique:

Theorem

Let A be an $m \times n$ matrix. The following statements are equivalent:

1. *for all $\mathbf{b} \in \mathbb{R}^m$, there is a unique least-squares solution to $A\mathbf{x} = \mathbf{b}$;*
2. *the columns of A are linearly independent;*
3. *the matrix $A^T A$ is invertible.*

If these hold then for any $\mathbf{b} \in \mathbb{R}^m$ the least-squares solution to $A\mathbf{x} = \mathbf{b}$ is given by $\hat{\mathbf{x}} = (A^T A)^{-1} A^T \mathbf{b}$.

Least-squares

In the previous example, the set of least-squares solutions was infinite. There is a theorem that describes when the least-squares solution to any system $A\mathbf{x} = \mathbf{b}$ is unique:

Theorem

Let A be an $m \times n$ matrix. The following statements are equivalent:

1. *for all $\mathbf{b} \in \mathbb{R}^m$, there is a unique least-squares solution to $A\mathbf{x} = \mathbf{b}$;*
2. *the columns of A are linearly independent;*
3. *the matrix $A^T A$ is invertible.*

If these hold then for any $\mathbf{b} \in \mathbb{R}^m$ the least-squares solution to $A\mathbf{x} = \mathbf{b}$ is given by $\hat{\mathbf{x}} = (A^T A)^{-1} A^T \mathbf{b}$.

You can think of this as a kind of "Invertible Matrix Theorem for non-square matrices."

Least-squares

If $A = \begin{bmatrix} 1 & 2 \\ 0 & 2 \\ 0 & 2 \end{bmatrix}$ then for any $\mathbf{b} \in \mathbb{R}^3$, there is a unique least-squares solution to $A\mathbf{x} = \mathbf{b}$.

If $A = \begin{bmatrix} 1 & 2 \\ 0 & 2 \\ 0 & 2 \end{bmatrix}$ then for any $\mathbf{b} \in \mathbb{R}^3$, there is a unique least-squares solution to $A\mathbf{x} = \mathbf{b}$.

Least-squares

Solving least-squares problems involving A is sped up considerably when you have a QR factorization for A.

Solving least-squares problems involving A is sped up considerably when you have a QR factorization for A.

Theorem

Let A be an $m \times n$ matrix with linearly independent columns and suppose $A = QR$ is a least-squares factorization for A.

Least-squares

Solving least-squares problems involving A is sped up considerably when you have a QR factorization for A.

Theorem

Let A be an $m \times n$ matrix with linearly independent columns and suppose $A = QR$ is a least-squares factorization for A. Then for any $\mathbf{b} \in \mathbb{R}^m$ the least-squares solution to $A\mathbf{x} = \mathbf{b}$ is unique and given by

$$\hat{\mathbf{x}} = R^{-1}(Q^T\mathbf{b}).$$

Solving least-squares problems involving A is sped up considerably when you have a QR factorization for A.

Theorem

Let A be an $m \times n$ matrix with linearly independent columns and suppose $A = QR$ is a least-squares factorization for A. Then for any $\mathbf{b} \in \mathbb{R}^m$ the least-squares solution to $A\mathbf{x} = \mathbf{b}$ is unique and given by

$$\hat{\mathbf{x}} = R^{-1}(Q^T\mathbf{b}).$$

The closest vector to $\mathbf{b}$ in colA *is* $(QQ^T)\mathbf{b}$.

Let $A = \begin{bmatrix} 1 & 2 \\ 1 & 3 \\ 1 & 2 \end{bmatrix}$ and let $\mathbf{b} = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}$.

Let $A = \begin{bmatrix} 1 & 2 \\ 1 & 3 \\ 1 & 2 \end{bmatrix}$ and let $\mathbf{b} = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}$. We have a QR factorization

$$A = \underbrace{\begin{bmatrix} 1/\sqrt{3} & -1/\sqrt{6} \\ 1/\sqrt{3} & \sqrt{2/3} \\ 1/\sqrt{3} & -1/\sqrt{6} \end{bmatrix}}_{Q} \underbrace{\begin{bmatrix} \sqrt{3} & 7/\sqrt{3} \\ 0 & \sqrt{2/3} \end{bmatrix}}_{R}.$$

Least-squares

Let $A = \begin{bmatrix} 1 & 2 \\ 1 & 3 \\ 1 & 2 \end{bmatrix}$ and let $\mathbf{b} = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}$. We have a QR factorization

$$A = \underbrace{\begin{bmatrix} 1/\sqrt{3} & -1/\sqrt{6} \\ 1/\sqrt{3} & \sqrt{2/3} \\ 1/\sqrt{3} & -1/\sqrt{6} \end{bmatrix}}_{Q} \underbrace{\begin{bmatrix} \sqrt{3} & 7/\sqrt{3} \\ 0 & \sqrt{2/3} \end{bmatrix}}_{R}.$$

Then take

$$Q^T\mathbf{b} = \begin{bmatrix} 2/(3\sqrt{3}) \\ -1/(6\sqrt{6}) + \sqrt{2/3} \end{bmatrix}.$$

Let $A = \begin{bmatrix} 1 & 2 \\ 1 & 3 \\ 1 & 2 \end{bmatrix}$ and let $\mathbf{b} = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}$. We have a QR factorization

$$A = \underbrace{\begin{bmatrix} 1/\sqrt{3} & -1/\sqrt{6} \\ 1/\sqrt{3} & \sqrt{2/3} \\ 1/\sqrt{3} & -1/\sqrt{6} \end{bmatrix}}_{Q} \underbrace{\begin{bmatrix} \sqrt{3} & 7/\sqrt{3} \\ 0 & \sqrt{2/3} \end{bmatrix}}_{R}.$$

Then take

$$Q^T\mathbf{b} = \begin{bmatrix} 2/(3\sqrt{3}) \\ -1/(6\sqrt{6}) + \sqrt{2/3} \end{bmatrix}.$$

Now the least squares solution is

$$\hat{\mathbf{x}} = R^{-1} \begin{bmatrix} 2/(3\sqrt{3}) \\ -1/(6\sqrt{6}) + \sqrt{2/3} \end{bmatrix} = \begin{bmatrix} 1/2 \\ 1/2 \end{bmatrix}$$

Modeling with least squares

In science/stats/econ you often have three things:

Modeling with least squares

In science/stats/econ you often have three things:

1. some experimental data

Modeling with least squares

In science/stats/econ you often have three things:

1. some experimental data
2. a mathematical model you have chosen to model the data

Modeling with least squares

In science/stats/econ you often have three things:

1. some experimental data
2. a mathematical model you have chosen to model the data
3. some parameters which control what the mathematical model looks like

Modeling with least squares

In science/stats/econ you often have three things:

1. some experimental data
2. a mathematical model you have chosen to model the data
3. some parameters which control what the mathematical model looks like

You want to pick the right parameters to make your model approximate the data as closely as possible.

Modeling

Example

Let's say you want to mathematically model how the height of a tree varies with its age. You collect four data points, each of which consists of an ordered pair of the form

$$(\text{age of tree in years},\ \text{height of tree in meters})\,.$$

Modeling

Example

Let's say you want to mathematically model how the height of a tree varies with its age. You collect four data points, each of which consists of an ordered pair of the form

$$(\text{age of tree in years}, \ \text{height of tree in meters})\,.$$

Let t denote age and h denote height. Let's say the *data* you collect are
$(t_1, h_1) = (1, 2), (t_2, h_2) = (2, 3), (t_3, h_3) = (4, 7), (t_4, h_4) = (5, 9).$

Modeling

Example

Let's say you want to mathematically model how the height of a tree varies with its age. You collect four data points, each of which consists of an ordered pair of the form

$$(\text{age of tree in years},\ \text{height of tree in meters})\,.$$

Let t denote age and h denote height. Let's say the *data* you collect are
$(t_1, h_1) = (1, 2), (t_2, h_2) = (2, 3), (t_3, h_3) = (4, 7), (t_4, h_4) = (5, 9)$.
Your teacher suggests that you model the data with a quadratic function

$$h = \beta_0 + \beta_1 t + \beta_2 t^2.$$

This is the *model*.

Modeling

Example

Let's say you want to mathematically model how the height of a tree varies with its age. You collect four data points, each of which consists of an ordered pair of the form

$$(\text{age of tree in years}, \text{ height of tree in meters}) .$$

Let t denote age and h denote height. Let's say the *data* you collect are

$(t_1, h_1) = (1, 2), (t_2, h_2) = (2, 3), (t_3, h_3) = (4, 7), (t_4, h_4) = (5, 9).$

Your teacher suggests that you model the data with a quadratic function

$$h = \beta_0 + \beta_1 t + \beta_2 t^2.$$

This is the *model*. Then the *parameters* are $\beta_0, \beta_1, \beta_2$. You have control over the parameters: you can set them however you like in order to most closely approximate the data.

Modeling

What does "most closely approximate the data" mean in this context? Basically it means that you are doing a least squares problem.

Modeling

What does “most closely approximate the data” mean in this context? Basically it means that you are doing a least squares problem. The height data form an *observation vector*:

$$\mathbf{y} = \begin{bmatrix} h_1 \\ h_2 \\ h_3 \\ h_4 \end{bmatrix}.$$

Modeling

What does "most closely approximate the data" mean in this context? Basically it means that you are doing a least squares problem. The height data form an *observation vector*:

$$\mathbf{y} = \begin{bmatrix} h_1 \\ h_2 \\ h_3 \\ h_4 \end{bmatrix}.$$

The model you have selected (quadratic) along with the ages of the trees determine a *design matrix*, which is denoted by X:

$$X = \begin{bmatrix} 1 & t_1 & (t_1)^2 \\ 1 & t_2 & (t_2)^2 \\ 1 & t_3 & (t_3)^2 \\ 1 & t_4 & (t_4)^2 \end{bmatrix}.$$

Modeling

What does "most closely approximate the data" mean in this context? Basically it means that you are doing a least squares problem. The height data form an *observation vector*:

$$\mathbf{y} = \begin{bmatrix} h_1 \\ h_2 \\ h_3 \\ h_4 \end{bmatrix}.$$

The model you have selected (quadratic) along with the ages of the trees determine a *design matrix*, which is denoted by X:

$X = \begin{bmatrix} 1 & t_1 & (t_1)^2 \\ 1 & t_2 & (t_2)^2 \\ 1 & t_3 & (t_3)^2 \\ 1 & t_4 & (t_4)^2 \end{bmatrix}$. Parameters form a *parameter vector* as

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}.$$

Modeling

Now we can state the basic idea behind modeling problems with least-squares: *you should pick the parameter vector* β *which makes the "prediction vector"* $X\beta$ *as close to the observed vector* $\mathbf{y}$ *as possible.*

Modeling

Now we can state the basic idea behind modeling problems with least-squares: *you should pick the parameter vector* β *which makes the "prediction vector"* $X\beta$ *as close to the observed vector* $\mathbf{y}$ *as possible.* That is, least-squares parameters $\beta_0, \beta_1, \beta_2$ are exactly the entries of the least-squares solution to $X\mathbf{x} = \mathbf{y}$, where X is the design matrix and $\mathbf{y}$ is the observation vector.

Now we can find the least-squares solution for the tree-height problem. The *observation vector* is the list of heights: $\mathbf{y} = \begin{bmatrix} 2 \\ 3 \\ 7 \\ 9 \end{bmatrix}$.

Now we can find the least-squares solution for the tree-height problem. The *observation vector* is the list of heights: $\mathbf{y} = \begin{bmatrix} 2 \\ 3 \\ 7 \\ 9 \end{bmatrix}$.

The *design matrix* is obtained by plugging in $t_1 = 1$, $t_2 = 2$, $t_3 = 4$, $t_4 = 5$ into the matrix from before:

$$X = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 4 & 16 \\ 1 & 5 & 25 \end{bmatrix}.$$

Now we can find the least-squares solution for the tree-height problem. The *observation vector* is the list of heights: $\mathbf{y} = \begin{bmatrix} 2 \\ 3 \\ 7 \\ 9 \end{bmatrix}$.

The *design matrix* is obtained by plugging in $t_1 = 1$, $t_2 = 2$, $t_3 = 4$, $t_4 = 5$ into the matrix from before:

$$X = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 4 & 16 \\ 1 & 5 & 25 \end{bmatrix}.$$

The least-squares solution to $X\mathbf{x} = \mathbf{y}$ is $\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} 0.933 \\ 0.8 \\ 0.167 \end{bmatrix}$.

Now we can find the least-squares solution for the tree-height problem. The *observation vector* is the list of heights: $\mathbf{y} = \begin{bmatrix} 2 \\ 3 \\ 7 \\ 9 \end{bmatrix}$.

The *design matrix* is obtained by plugging in $t_1 = 1$, $t_2 = 2$, $t_3 = 4$, $t_4 = 5$ into the matrix from before:

$$X = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 4 & 16 \\ 1 & 5 & 25 \end{bmatrix}.$$

The least-squares solution to $X\mathbf{x} = \mathbf{y}$ is $\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} 0.933 \\ 0.8 \\ 0.167 \end{bmatrix}$. So the least-squares model is $h(t) = 0.933 + 0.8t + 0.167t^2$.

Modeling

Let's pause to review how to construct the design matrix and the observation vector. You are assuming that there is some dependent variable y, some independent variable t (could be more than one), and that there is some relation $y = \sum_{i=0}^{q} \beta_i f_i$, where f_i are functions of the independent variable $f_i = f_i(t)$.

Modeling

Let's pause to review how to construct the design matrix and the observation vector. You are assuming that there is some dependent variable y, some independent variable t (could be more than one), and that there is some relation $y = \sum_{i=0}^{q} \beta_i f_i$, where f_i are functions of the independent variable $f_i = f_i(t)$. In the previous example we have $f_0(t) = 1, f_1(t) = t, f_2(t) = t^2$. That's the *model*. The experimental data comes to you as a list of observations $(t_1, y_1), (t_2, y_2), \ldots, (t_m, y_m)$, where t_k is some specific value of the independent variable and y_k is the value of the dependent variable you observe at when the independent variable is t_k.

In this case the observation vector is just the list of y values:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \ldots y_m \end{bmatrix}$$

In this case the observation vector is just the list of y values:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \dots y_m \end{bmatrix}$$

The design matrix X has one column for each parameter β_i, and the ith column of X is just

$$\begin{bmatrix} f_i(t_1) \\ f_i(t_2) \\ \vdots \\ f_q(t_m) \end{bmatrix}.$$

Modeling

Let's do another example. Suppose that you have experimental data $(1, 7.9), (2, 5.4), (3, -.9)$ and you wish to model this data as

$$y = A\cos x + B\sin x$$

where $A, B \in \mathbb{R}$. How do we do that?

Modeling: $y = A\cos x + B\sin x$

The data are
$(x_1, y_1) = (1, 7.9), (x_2, y_2) = (2, 5.4), (x_3, y_3) = (3, -.9)$.

Modeling: $y = A\cos x + B\sin x$

The data are
$(x_1, y_1) = (1, 7.9), (x_2, y_2) = (2, 5.4), (x_3, y_3) = (3, -.9)$. So we can write the observation vector as $\mathbf{y} = \begin{bmatrix} 7.9 \\ 5.4 \\ -.9 \end{bmatrix}$.

Modeling: $y = A\cos x + B\sin x$

The data are $(x_1, y_1) = (1, 7.9), (x_2, y_2) = (2, 5.4), (x_3, y_3) = (3, -.9)$. So we can write the observation vector as $\mathbf{y} = \begin{bmatrix} 7.9 \\ 5.4 \\ -.9 \end{bmatrix}$. The two functions are $f_1(x) = \cos(x)$ and $f_2(x) = \sin(x)$. Thus the design matrix is

$$X = \begin{bmatrix} f_1(x_1) & f_2(x_1) \\ f_1(x_2) & f_2(x_2) \\ f_1(x_3) & f_2(x_3) \end{bmatrix} = \begin{bmatrix} 0.54 & 0.84 \\ -0.42 & 0.91 \\ -0.99 & 0.14 \end{bmatrix}.$$

Modeling: $y = A\cos x + B\sin x$

The data are
$(x_1, y_1) = (1, 7.9), (x_2, y_2) = (2, 5.4), (x_3, y_3) = (3, -.9)$. So we can write the observation vector as $\mathbf{y} = \begin{bmatrix} 7.9 \\ 5.4 \\ -.9 \end{bmatrix}$. The two functions are $f_1(x) = \cos(x)$ and $f_2(x) = \sin(x)$. Thus the design matrix is

$$X = \begin{bmatrix} f_1(x_1) & f_2(x_1) \\ f_1(x_2) & f_2(x_2) \\ f_1(x_3) & f_2(x_3) \end{bmatrix} = \begin{bmatrix} 0.54 & 0.84 \\ -0.42 & 0.91 \\ -0.99 & 0.14 \end{bmatrix}.$$

So we need to find the least squares solution to $X\mathbf{x} = \begin{bmatrix} 7.9 \\ 5.4 \\ -.9 \end{bmatrix}$.

Modeling: $y = A\cos x + B\sin x$

The data are $(x_1, y_1) = (1, 7.9), (x_2, y_2) = (2, 5.4), (x_3, y_3) = (3, -.9)$. So we can write the observation vector as $\mathbf{y} = \begin{bmatrix} 7.9 \\ 5.4 \\ -.9 \end{bmatrix}$. The two functions are $f_1(x) = \cos(x)$ and $f_2(x) = \sin(x)$. Thus the design matrix is

$$X = \begin{bmatrix} f_1(x_1) & f_2(x_1) \\ f_1(x_2) & f_2(x_2) \\ f_1(x_3) & f_2(x_3) \end{bmatrix} = \begin{bmatrix} 0.54 & 0.84 \\ -0.42 & 0.91 \\ -0.99 & 0.14 \end{bmatrix}.$$

So we need to find the least squares solution to $X\mathbf{x} = \begin{bmatrix} 7.9 \\ 5.4 \\ -.9 \end{bmatrix}$. The least-squares solution is $\hat{\mathbf{x}} = \begin{bmatrix} 2.34 \\ 7.45 \end{bmatrix}$. So the best model is $y = 2.34\cos(x) + 7.45\sin(x)$.