Application of Linear Algebra on Least Squares Approximation

Kelan Lu
Doctoral Student
Univ. of North Texas
Dept. of Political Science
lillylu01@gmail.com

May 8, 2010
An Introduction of the Least Squares Problem

Least Squares problem can be illustrated into two different ways, including the geometric approach and the numerical approach.

Geometric Approach

The geometric way to understand the least squares problem is as follows. Generally, we often run into the problem that we have more than two points and try to represent our points with one straight line. However, the data points do not lie on a straight line. We can try infinitely many straight lines to fit all the data points. Under this situation, the problem of least squares is to find the line that fits the data the best.

The best fitting line is often called the least squares line or the regression line. In order to understand the best fitting line, we need to understand the concept of "residuals", which means that the directed distances between the observed data points and the corresponding points on the model line. To obtain the best fitting line, we need to minimize the sum of the squares of the residuals.

Numerical Approach

The numerical approach to understand the least squares problem is as follows. When we try to fit one line on more than two points, we tend to face a problem that the linear equation $\mathbf{A}\vec{x} = \vec{b}$ has no solution because there are more equations than the number of variables. In other words, the linear system is inconsistent and $\vec{b}$ is not in the column space of $\mathbf{A}$. In order to illustrate the inconsistency of the above linear equation, I use the following example.

Given that two variables $x$ and $y$ have linear relationship, which is $y = mx + b$. Suppose I collect the following data:
In order to find a line with equation $y = mx + b$ through the above points, we need to form the following system of linear equations.

\begin{align*}
2m + b &= 9 \\
3m + b &= 7 \\
5m + b &= 4 \\
m + b &= 1 \\
\end{align*}

The above system of linear equations can be re-written in the following matrix form:
Through row-reduction to the row reduced echelon form of the augmented matrix, I found the system is inconsistent because it is over-determined and over-determined matrix is usually inconsistent.

Therefore, we cannot find a solution $\mathbf{x}$ that will satisfy $\mathbf{A}\mathbf{x} = \mathbf{b}$.

**Define the Least Squares Problem**

What can we do to obtain a solution when there is no solution existing for the linear equation $\mathbf{A}\mathbf{x} = \mathbf{b}$? First, we can regard $\mathbf{A}\mathbf{x}$ as an approximation of $\mathbf{b}$. The smaller the distance between $\mathbf{b}$ and $\mathbf{A}\mathbf{x}$, the better approximation between them. Second, instead of solving $\mathbf{A}\mathbf{x} = \mathbf{b}$, we try to find a vector $\mathbf{\hat{x}}$ such that $\mathbf{A}\mathbf{\hat{x}}$ is good approximation of $\mathbf{b}$ and thus the value of $\|\mathbf{b} - \mathbf{A}\mathbf{\hat{x}}\|$ is as small as possible. Therefore, according to Lay (2005), the definition of a least-squares solution of $\mathbf{A}\mathbf{x} = \mathbf{b}$ is as follows: If $\mathbf{A}$ is a m x n matrix and $\mathbf{b}$ is in $\mathbb{R}^m$, a least-squares solution of $\mathbf{A}\mathbf{x} = \mathbf{b}$ is a vector $\mathbf{\hat{x}}$ in $\mathbb{R}^n$ such that $\|\mathbf{b} - \mathbf{A}\mathbf{\hat{x}}\| \leq \|\mathbf{b} - \mathbf{A}\mathbf{x}\|$ for all $\mathbf{x}$ in $\mathbb{R}^n$.

From the geometric perspective, we can deal with the least squares problem by the following logic. First, according to the definition of column space, a typical vector in $\text{Col}\mathbf{A}$ can be written as $\mathbf{A}\mathbf{x}$ for some $\mathbf{x}$ because the notation $\mathbf{A}\mathbf{x}$ stands for a linear combination of the columns of $\mathbf{A}$. Therefore, $\mathbf{A}\mathbf{x}$ will always be in the column space of $\mathbf{A}$. Since there is no solution for the equation $\mathbf{A}\mathbf{x} = \mathbf{b}$, $\mathbf{b}$ is not in the column space of $\mathbf{A}$. Therefore, we can find a $\mathbf{x}$ that makes $\mathbf{A}\mathbf{x}$ the closest point in $\text{Col}\mathbf{A}$ to $\mathbf{b}$. According
to Lay (2005), such a vector is the orthogonal projection of $\vec{b}$ onto the column space of $A$. The following section will illustrate the important concept of orthogonal projection.

**Orthogonal Projections**

**Orthogonal Vectors**

In order to understand orthogonal projections, we have to first introduce the concept of orthogonal vectors. According to Lay (2005), two vectors $\vec{u}$ and $\vec{v}$ in $\mathbb{R}^n$ are orthogonal (to each other) if their dot multiplication is equal to zero. In order to prove this statement, we use $\mathbb{R}^2$ and two lines through the origin determined by vectors $\vec{u}$ and $\vec{v}$.

Start from the origin, we draw two lines determined by vectors $\vec{u}$ and $\vec{v}$. These two lines are perpendicular to each other if and only if the distance from $\vec{u}$ to $\vec{v}$ is the same as the distance from $\vec{u}$ to $-\vec{v}$. That is to say, the two lines are perpendicular to each other if and only if the squares of the two distances above are the same.

$$
[dist(\vec{u}, \vec{v})]^2 = \|\vec{u} - \vec{v}\|^2 \\
= (\vec{u} - \vec{v}) \cdot (\vec{u} - \vec{v}) \\
= \vec{u} \cdot (\vec{u} - \vec{v}) - \vec{v} \cdot (\vec{u} - \vec{v}) \\
= \vec{u} \cdot \vec{u} - \vec{u} \cdot \vec{v} - \vec{v} \cdot \vec{u} + \vec{v} \cdot \vec{v} \\
= \|\vec{u}\|^2 + \|\vec{v}\|^2 - 2\vec{u} \cdot \vec{v}
$$

To the same token, when we calculate the distance between $\vec{u}$ to $-\vec{v}$, we can follow the same procedure shown above:
\[
[\text{dist}(\vec{u}, -\vec{v})]^2 = \|\vec{u}\|^2 + \|\vec{v}\|^2 - 2\vec{u} \cdot(-\vec{v})
\]
\[
= \|\vec{u}\|^2 + \|\vec{v}\|^2 + 2\vec{u} \cdot \vec{v}
\]

Therefore, the squared distances will be equal to each other if and only if \(\vec{u} \cdot \vec{v} = 0\).

In conclusion, the two lines determined by the two vectors \(\vec{u}\) and \(\vec{v}\) are perpendicular to each other if and only if \(\vec{u} \cdot \vec{v} = 0\).

Moreover, if a vector \(\vec{z}\) is orthogonal to every vector in a subspace \(W\) of \(R^n\), then \(\vec{z}\) is said to be orthogonal to \(W\). The set of all vectors \(\vec{z}\) that are orthogonal to \(W\) is called the orthogonal complement of \(W\) and is denoted by \(W^\perp\).

**An Orthogonal Projection in \(R^n\)**

Given a non-zero vector \(\vec{u}\) in \(R^n\), we can decompose a vector \(\vec{y}\) in \(R^n\) into the sum of two vectors, one a multiple of \(\vec{u}\) and the other orthogonal to \(\vec{u}\). Then we can write:

\[
\vec{y} = \hat{\vec{y}} + \vec{z}
\]

where \(\hat{\vec{y}} = \alpha\vec{u}\) for some scalar \(\alpha\) and \(\vec{z}\) is some vector orthogonal to \(\text{vecu}\). Then \(\vec{z}\) is equal to \(\vec{y} - \hat{\vec{y}}\), which is is orthogonal to \(\vec{u}\) if and only if

\[
\vec{0} = (\vec{y} - \alpha\vec{u})\vec{u} = \vec{y}\vec{u} - (\alpha\vec{u})\vec{u} = \vec{y}\vec{u} - \alpha(\vec{u}\vec{u})
\]

Therefore, \(\vec{z}\) is orthogonal to \(\vec{u}\) if and only if \(\alpha = \frac{\vec{y}\vec{u}}{\vec{u}\vec{u}}\) and \(\hat{\vec{y}} = \frac{\vec{y}\vec{u}}{\vec{u}\vec{u}} \cdot \vec{u}\). Thus, the vector \(\hat{\vec{y}}\) is called the orthogonal projection of \(\vec{y}\) onto \(\vec{u}\), and the vector \(\vec{z}\) is called the component of \(\vec{y}\) orthogonal to \(\vec{u}\). In this case, \(\hat{\vec{y}}\) is denoted by \(\text{proj}_L\vec{y}\) and is called the orthogonal projection of \(\vec{y}\) onto \(L\).
Let $\vec{b}$ be a nonzero vector in $R^m$, and $W$ be a subspace of $R^m$ spanned by the vectors $\vec{w}_1, \vec{w}_2, \vec{w}_3, ..., \vec{w}_n$. There is a vector $\vec{b}$ in $W$ such that (1) $\vec{b}$ is the unique vector in $W$ for which $\vec{b} - \vec{b}$ is orthogonal to $W$, and (2) $\vec{b}$ is the unique vector in $W$ closest to $\vec{b}$. This vector $\vec{b}$ is called the orthogonal projection of $\vec{b}$ onto $W$. We can express it as:

$$\hat{\vec{b}} = \text{proj}_W \vec{y}$$

The first property the vector $\vec{b}$ leads to the Orthogonal Decomposition Theorem, and the second property contributes to the Best Approximation Theorem. These two properties are critical for us to find the least-squares solutions of linear systems.

### Solution to the General Least-Squares Problem

Given that $\vec{b}$ be a nonzero vector in $R^m$, and $A$ be a subspace of $R^m$ spanned by the vectors $\vec{a}_1, \vec{a}_2, \vec{a}_3, ..., \vec{a}_n$, apply the Best Approximation Theorem to the subspace of $\text{Col}A$. Let

$$\vec{b} = \text{proj}_{\text{Col}A} \vec{y}$$

The equation $A\vec{x} = \vec{b}$ is consistent because $\vec{b}$ is in the column space of $A$. Moreover, there is a vector $\vec{x}$ in $R^n$ such that $A\vec{x} = \vec{b}$, and this vector $\vec{x}$ is the least-squares solution for the equation $A\vec{x} = \vec{b}$.

Given that $\vec{x}$ satisfies the equation $A\vec{x} = \vec{b}$, according to the Decomposition Theorem, the projection $\vec{b}$ has the property that $\vec{b} - \vec{b}$ is orthogonal to $\text{Col}A$. Therefore, $\vec{b} - A\vec{x}$ is orthogonal to each column of $A$. If $\vec{a}_j$ is any column of $A$, then $\vec{a}_j \cdot (\vec{b} - A\vec{x}) = 0$, and $\vec{a}_j^T \cdot (\vec{b} - A\vec{x}) = 0$. Since each $\vec{a}_j^T$ is a row of $A^T$,

$$A^T(\vec{b} - A\vec{x}) = 0$$
\[ A^T \vec{b} - A^T A \vec{x} = \vec{0} \]

\[ A^T A \vec{x} = A^T \vec{b} \]

Therefore, each least-squares solution of \( A \vec{x} = \vec{b} \), denoted as \( \vec{x} \), satisfies the equation

\[ A^T A \vec{x} = A^T \vec{b} \]

, which is a system of equations called the normal equations for \( A \vec{x} = \vec{b} \).

The matrix \( A^T A \) is invertible if and only if the columns of \( A \) are linearly independent. In this case, the equation \( A \vec{x} = \vec{b} \) has only one least-squares solution \( \vec{x} \), and it is given by

\[ \vec{x} = (A^T A)^{-1} A^T \vec{b} \]

Application of Least Square Lines on the Tumor Size Experiment

Suppose I am interested in the effect of radiation on the size of cancerous tumor. I expect that the size of cancerous tumor is a linear function of the number of hours of radiation. Under this expectation, I build up the following model:

\[ Y_i = \beta_0 + \beta_1 X_{1i} + u_i \]

Where \( Y_i = \) Size of the tumor, and \( X_{1i} = \) the number of hours of radiation, and \( u_i = \) the error term, which is equal to subtracting the observed size of the tumor by the modeled size of the tumor).
In order to estimate the above model, I collect the following experimental data points:

<table>
<thead>
<tr>
<th>$Y_i$</th>
<th>$X_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
</tr>
</tbody>
</table>

Use the x-coordinates of the data to build the matrix $X$ and the y-coordinates to build the vector $\mathbf{y}$:

\[
X = \begin{bmatrix}
1 & 0 \\
1 & 1 \\
1 & 2 \\
1 & 3 \\
1 & 4 \\
\end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix}
10 \\
8 \\
7 \\
5 \\
2 \\
\end{bmatrix}
\]

For the least-squares solution of $X\tilde{\beta} = \tilde{y}$, obtain the normal equation (with the new notation):

\[
X^T X \tilde{\beta} = X^T \tilde{y}
\]

That is, compute
\[
\mathbf{X}^T \mathbf{X} = \begin{bmatrix}
1 & 1 & 1 & 1 & 1 \\
0 & 1 & 2 & 3 & 4
\end{bmatrix}
\begin{bmatrix}
1 & 0 \\
1 & 1 \\
1 & 2 \\
1 & 3 \\
1 & 4
\end{bmatrix}
= \begin{bmatrix}
5 & 10 \\
10 & 30
\end{bmatrix}
\]

\[
\mathbf{X}^T \mathbf{y} = \begin{bmatrix}
1 & 1 & 1 & 1 \\
0 & 1 & 2 & 3 & 4
\end{bmatrix}
\begin{bmatrix}
10 \\
8 \\
7 \\
5 \\
2
\end{bmatrix}
= \begin{bmatrix}
32 \\
45
\end{bmatrix}
\]

The normal equations are:

\[
\begin{bmatrix}
5 & 10 \\
10 & 30
\end{bmatrix}
\begin{bmatrix}
\beta_0 \\
\beta_1
\end{bmatrix}
= \begin{bmatrix}
32 \\
45
\end{bmatrix}
\]

Hence

\[
\begin{bmatrix}
\beta_0 \\
\beta_1
\end{bmatrix}
= \begin{bmatrix}
5 & 10 \\
10 & 30
\end{bmatrix}^{-1}
\begin{bmatrix}
32 \\
45
\end{bmatrix}
= \frac{1}{50}
\begin{bmatrix}
30 & -10 \\
-10 & 5
\end{bmatrix}
\begin{bmatrix}
32 \\
45
\end{bmatrix}
= \frac{1}{50}
\begin{bmatrix}
510 \\
-95
\end{bmatrix}
= \begin{bmatrix}
10.2 \\
-1.9
\end{bmatrix}
\]

Thus the least-squares line has the equation:

\[
y = 10.2 - 1.9x
\]
In conclusion, if we do not have any radiation on the tumor, the size of the tumor will be 10.2 units. Every additional hour’s radiation on the tumor will decrease the size of the tumor by 1.9 units.
tyleapsr