Method of Least Squares

Steven J Miller
Williams College

sjml@williams.edu

http://www.williams.edu/Mathematics/sjmiller/public_html/
Introduction
Figure: xkcd: Convincing: https://xkcd.com/833/ (Extra text: And if you labeled your axes, I could tell you exactly how MUCH better.)
Data from $x_n = 5 + .2n$, $y_n = 5x_n$ plus an error randomly drawn from a normal distribution with mean zero and standard deviation 4. Best fit line of $y = 4.99x + .48$; thus $a = 4.99$ and $b = .48$. 
Spring Test (continued)

Our value of \( b \) is significantly off: \( a = 4.99 \) and \( b = .48 \).
Our value of $b$ is significantly off: $a = 4.99$ and $b = .48$.

Using absolute values for errors gives best fit value of $a$ is 5.03 and the best fit value of $b$ is less than $10^{-10}$ in absolute value.
Our value of $b$ is significantly off: $a = 4.99$ and $b = .48$.

Using absolute values for errors gives best fit value of $a$ is $5.03$ and the best fit value of $b$ is less than $10^{-10}$ in absolute value.

The difference between these values and those from the Method of Least Squares is in the best fit value of $b$ (the least important of the two parameters), and is due to the different ways of weighting the errors.
Regression

See https://web.williams.edu/Mathematics/sjmiller/public_html/probabilitylifesaver/MethodLeastSquares.pdf
Overview

Idea is to find best-fit parameters: choices that minimize error in a conjectured relationship.

Say observe $y_i$ with input $x_i$, believe $y_i = ax_i + b$. Three choices:

\[
E_1(a, b) = \sum_{n=1}^{N} (y_i - (ax_i + b))
\]

\[
E_2(a, b) = \sum_{n=1}^{N} |y_i - (ax_i + b)|
\]

\[
E_3(a, b) = \sum_{n=1}^{N} (y_i - (ax_i + b))^2.
\]
Idea is to find *best-fit* parameters: choices that minimize error in a conjectured relationship.

Say observe $y_i$ with input $x_i$, believe $y_i = ax_i + b$. Three choices:

$$E_1(a, b) = \sum_{n=1}^{N} (y_i - (ax_i + b))$$

$$E_2(a, b) = \sum_{n=1}^{N} |y_i - (ax_i + b)|$$

$$E_3(a, b) = \sum_{n=1}^{N} (y_i - (ax_i + b))^2.$$ 

Use sum of squares as calculus available.
Linear Regression

Explicit formula for values of $a, b$ minimizing error $E_3(a, b)$. From

$$\frac{\partial E_3(a, b)}{\partial a} = \frac{\partial E_3(a, b)}{\partial b} = 0 :$$

After algebra:

$$\left( \hat{a}, \hat{b} \right) = \left( \begin{array}{cc} \sum_{n=1}^{N} x_i^2 & \sum_{n=1}^{N} x_i \\ \sum_{n=1}^{N} x_i & \sum_{n=1}^{N} 1 \end{array} \right) \left( \begin{array}{c} \sum_{n=1}^{N} x_i y_i \\ \sum_{n=1}^{N} y_i \end{array} \right)$$

or

$$a = \frac{\sum_{n=1}^{N} 1 \sum_{n=1}^{N} x_n y_n - \sum_{n=1}^{N} x_n \sum_{n=1}^{N} y_n}{\sum_{n=1}^{N} 1 \sum_{n=1}^{N} x_n^2 - \sum_{n=1}^{N} x_n \sum_{n=1}^{N} x_n}$$

$$b = \frac{\sum_{n=1}^{N} x_n \sum_{n=1}^{N} x_n y_n - \sum_{n=1}^{N} x_n^2 \sum_{n=1}^{N} y_n}{\sum_{n=1}^{N} x_n \sum_{n=1}^{N} x_n - \sum_{n=1}^{N} x_n^2 \sum_{n=1}^{N} 1}.$$
Theory
Theoretical Aside: Derivation

See https://web.williams.edu/Mathematics/sjmiller/public_html/341Fa18/handouts/MethodLeastSquares.pdf

\[ E_3(a, b) = \sum_{n=1}^{N} (y_i - (ax_i + b))^2. \]

Error a function of two variables, the unknown parameters \( a \) and \( b \).

Note \( x, y \) are the data \( NOT \) the variables.

The goal is to find values of \( a \) and \( b \) that minimize the error.
Theoretical Aside: Derivation: II

One-Variable Calculus: candidates for max/min from boundary points and critical points (places where derivative vanishes).

Multivariable Calculus: Similar, need partial derivatives to vanish (partial is hold all variables fixed but one).

\[
\nabla E = \left( \frac{\partial E}{\partial a}, \frac{\partial E}{\partial b} \right) = (0, 0),
\]

or

\[
\frac{\partial E}{\partial a} = 0, \quad \frac{\partial E}{\partial b} = 0.
\]

Do not have to worry about boundary points: as \(|a|\) and \(|b|\) become large, the fit gets worse and worse.
Theoretical Aside: Derivation: III

Differentiating $E(a, b)$ yields

$$\frac{\partial E}{\partial a} = \sum_{n=1}^{N} 2 \left( y_n - (ax_n + b) \right) \cdot (-x_n)$$

$$\frac{\partial E}{\partial b} = \sum_{n=1}^{N} 2 \left( y_n - (ax_n + b) \right) \cdot (-1).$$

Setting $\frac{\partial E}{\partial a} = \frac{\partial E}{\partial b} = 0$ (and dividing by -2) yields

$$\sum_{n=1}^{N} \left( y_n - (ax_n + b) \right) \cdot x_n = 0$$

$$\sum_{n=1}^{N} \left( y_n - (ax_n + b) \right) = 0.$$

Note we can divide both sides by -2 as it is just a constant; we cannot divide by $x_i$ as that varies with $i$. 
Theoretical Aside: Derivation: IV

Rewrite as

\[
\begin{align*}
\left( \sum_{n=1}^{N} x_n^2 \right) a + \left( \sum_{n=1}^{N} x_n \right) b &= \sum_{n=1}^{N} x_n y_n \\
\left( \sum_{n=1}^{N} x_n \right) a + \left( \sum_{n=1}^{N} 1 \right) b &= \sum_{n=1}^{N} y_n.
\end{align*}
\]

Values of $a$ and $b$ which minimize the error satisfy the following matrix equation:

\[
\begin{pmatrix}
\sum_{n=1}^{N} x_n^2 & \sum_{n=1}^{N} x_n \\
\sum_{n=1}^{N} x_n & \sum_{n=1}^{N} 1
\end{pmatrix}
\begin{pmatrix} a \\ b \end{pmatrix} =
\begin{pmatrix}
\sum_{n=1}^{N} x_n y_n \\
\sum_{n=1}^{N} y_n
\end{pmatrix}.
\]

(1)
Theoretical Aside: Derivation: V

Inverse of a matrix $A$ is the matrix $B$ such that $AB = BA = I$, where $I$ is the identity matrix.

If $A = \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix}$ is a $2 \times 2$ matrix where $\det A = \alpha \delta - \beta \gamma \neq 0$, then $A$ is invertible and

$$A^{-1} = \frac{1}{\alpha \delta - \beta \gamma} \begin{pmatrix} \delta & -\gamma \\ -\beta & \alpha \end{pmatrix}.$$  \hspace{1cm} (2)

In other words, $AA^{-1} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ here.

For example, if $A = \begin{pmatrix} 1 & 2 \\ 3 & 7 \end{pmatrix}$ then $\det A = 1$ and $A^{-1} = \begin{pmatrix} 7 & -2 \\ -3 & 1 \end{pmatrix}$; we can check this by noting (through matrix multiplication) that

$$\begin{pmatrix} 1 & 2 \\ 3 & 7 \end{pmatrix} \begin{pmatrix} 7 & -2 \\ -3 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$  \hspace{1cm} (3)
Theoretical Aside: Derivation: VI

\[
\begin{pmatrix}
a \\
b
\end{pmatrix} = \left( \begin{pmatrix}
\sum_{n=1}^{N} x_n^2 & \sum_{n=1}^{N} x_n \\
\sum_{n=1}^{N} x_n & \sum_{n=1}^{N} 1
\end{pmatrix} \right)^{-1} \begin{pmatrix}
\sum_{n=1}^{N} x_n y_n \\
\sum_{n=1}^{N} y_n
\end{pmatrix}.
\]

(4)

Denote the matrix from (1) by \( M \). The determinant of \( M \) is

\[
\det M = \sum_{n=1}^{N} x_n^2 \cdot \sum_{n=1}^{N} 1 - \sum_{n=1}^{N} x_n \cdot \sum_{n=1}^{N} x_n.
\]

As

\[
\bar{x} = \frac{1}{N} \sum_{n=1}^{N} x_n,
\]

we find that

\[
\det M = N \sum_{n=1}^{N} x_n^2 - (N \bar{x})^2 = \frac{1}{N} \left( N \sum_{n=1}^{N} (x_n - \bar{x})^2 \right),
\]

where the last equality follows from algebra. If the \( x_n \) are not all equal, \( \det M \) is non-zero and \( M \) is invertible.
We rewrite (4) in a simpler form. Using the inverse of the matrix and the definition of the mean and variance, we find

\[
\begin{pmatrix}
a \\ b
\end{pmatrix} = \frac{1}{N^2 \sigma^2_x} \begin{pmatrix}
N & -N \bar{x} \\ -N \bar{x} & \sum_{n=1}^{N} x_n^2
\end{pmatrix} \begin{pmatrix}
\sum_{n=1}^{N} x_n y_n \\ \sum_{n=1}^{N} y_n
\end{pmatrix}.
\]  
(5)

Expanding gives

\[
a = \frac{N \sum_{n=1}^{N} x_n y_n - N \bar{x} \sum_{n=1}^{N} y_n}{N^2 \sigma^2_x} \]

\[
b = \frac{-N \bar{x} \sum_{n=1}^{N} x_n y_n + \sum_{n=1}^{N} x_n^2 \sum_{n=1}^{N} y_n}{N^2 \sigma^2_x} \]

\[
\bar{x} = \frac{1}{N} \sum_{n=1}^{N} x_i
\]

\[
\sigma^2_x = \frac{1}{N} \sum_{n=1}^{N} (x_i - \bar{x})^2.
\]  
(6)
As the formulas for $a$ and $b$ are so important, it is worth giving another expression for them. We also have

\[
\begin{align*}
  a &= \frac{\sum_{n=1}^{N} 1 \sum_{n=1}^{N} x_n y_n - \sum_{n=1}^{N} x_n \sum_{n=1}^{N} y_n}{\sum_{n=1}^{N} 1 \sum_{n=1}^{N} x_n^2 - \sum_{n=1}^{N} x_n \sum_{n=1}^{N} x_n} \\
  b &= \frac{\sum_{n=1}^{N} x_n \sum_{n=1}^{N} x_n y_n - \sum_{n=1}^{N} x_n^2 \sum_{n=1}^{N} y_n}{\sum_{n=1}^{N} x_n \sum_{n=1}^{N} x_n - \sum_{n=1}^{N} x_n^2 \sum_{n=1}^{N} 1}.
\end{align*}
\]
Theoretical Aside: Derivation: Remarks

Formulas for $a$ and $b$ are reasonable, as can be seen by a unit analysis. Imagine $x$ in meters and $y$ in seconds. Then if $y = ax + b$ we would need $b$ and $y$ to have the same units (seconds), and $a$ to have units seconds per meter. If we substitute we do see $a$ and $b$ have the correct units. Not a proof that we have not made a mistake, but a great reassurance. No matter what you are studying, you should always try unit calculations such as this.
There are other, equivalent formulas for $a$ and $b$, arranging the algebra in a slightly different sequence of steps. Essentially what we are doing is the following: image we are given

\begin{align*}
4 &= 3a + 2b \\
5 &= 2a + 5b.
\end{align*}

If we want to solve, we can proceed in two ways. We can use the first equation to solve for $b$ in terms of $a$ and substitute in, or we can multiply the first equation by 5 and the second equation by 2 and subtract; the $b$ terms cancel and we obtain the value of $a$. Explicitly,

\begin{align*}
20 &= 15a + 10b \\
10 &= 4a + 10b,
\end{align*}

which yields

\[ 10 = 11a, \]

or

\[ a = \frac{10}{11}. \]
Regression Extensions
Beyond the Best Fit Line

Did \( y = ax + b \).

All that matters is linear in the unknown parameters.

Could do

\[
y = a_1 f_1(x) + a_2 f_2(x) + \cdots + a_k f_k(x);
\]

do not need the functions \( f \) to be linear.
Non-linear Relations

Most relations are not linear.

Newton’s law of gravity: \( F = Gm_1 m_2 / r^2 \).

If guess force is proportional to a power of the distance: \( F = Br^a \).

Take logarithms: \( \log(F) = a \log(r) + b \) with \( b = \log B \).

Note the linear relation between \( \log(F) \) and \( \log(r) \).
City Populations

The twenty-five most populous cities (I believe this is American cities from a few years ago):

<table>
<thead>
<tr>
<th>Rank</th>
<th>City Population</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8,363,710</td>
</tr>
<tr>
<td>2</td>
<td>3,833,995</td>
</tr>
<tr>
<td>3</td>
<td>2,853,114</td>
</tr>
<tr>
<td>4</td>
<td>2,242,193</td>
</tr>
<tr>
<td>5</td>
<td>1,567,924</td>
</tr>
<tr>
<td>6</td>
<td>1,540,351</td>
</tr>
<tr>
<td>7</td>
<td>1,351,305</td>
</tr>
<tr>
<td>8</td>
<td>912,062</td>
</tr>
<tr>
<td>9</td>
<td>754,885</td>
</tr>
<tr>
<td>10</td>
<td>620,535</td>
</tr>
<tr>
<td>11</td>
<td>1,351,305</td>
</tr>
<tr>
<td>12</td>
<td>808,976</td>
</tr>
<tr>
<td>13</td>
<td>807,815</td>
</tr>
<tr>
<td>14</td>
<td>798,382</td>
</tr>
<tr>
<td>15</td>
<td>757,688</td>
</tr>
<tr>
<td>16</td>
<td>703,073</td>
</tr>
<tr>
<td>17</td>
<td>687,456</td>
</tr>
<tr>
<td>18</td>
<td>669,651</td>
</tr>
<tr>
<td>19</td>
<td>636,919</td>
</tr>
<tr>
<td>20</td>
<td>613,190</td>
</tr>
<tr>
<td>21</td>
<td>604,477</td>
</tr>
<tr>
<td>22</td>
<td>598,707</td>
</tr>
<tr>
<td>23</td>
<td>598,541</td>
</tr>
<tr>
<td>24</td>
<td>27</td>
</tr>
</tbody>
</table>
City Populations

Figure: Plot of rank versus population
City Populations

Figure: Plot of rank versus log(population)
City Populations

Figure: Plot of log(rank) versus log(population)
City Populations

Plot of 100 most populous cities

Figure: Plot of rank versus population
City Populations

Plot of 100 most populous cities: log-log plot

**Figure:** Plot of log(rank) versus log(population)
Word Counts

Figure: Plot of rank versus occurrences
Figure: Plot of log(rank) versus log(occurrences)
Examples:
Chapter 70 Aid, Kepler’s Laws, Birthday Problem
Real World Challenge: Need to assign $3,500,000 to three schools (LES, WES, MtG).

- Pre-regionalization know how much state gives each; post regionalization only know sum.

- State has formula, lots of variables, secret.

What is the goal? How do we accomplish it?
Objectives

- Fair formula that predicts well.
- Transparent, seems fair.
- Can be explained.
Solution: Method of Least Squares / Linear Regression.

Inputs: Population of Schools (LES(pop), WES(pop), MtG(pop)), Assessment of Towns (EQV(L), EQV(W)).

Formula: If $\overrightarrow{y} = X \overrightarrow{\beta}$ then

$$\overrightarrow{\beta} = (X^T X)^{-1} X^T \overrightarrow{y}.$$  

What properties do we want the solution to have?
Properties of Solution

- Want solution to exist – will it?
- Want values to be between 0 and 1 – will it?
- Want values to be stable under small changes – will it?
- Want the sum of the three percentages to add to 1 – will it?
<table>
<thead>
<tr>
<th>Introduction</th>
<th>Regression</th>
<th>Theory</th>
<th>Regression Extensions</th>
<th>Examples</th>
</tr>
</thead>
</table>

40
Theory vs Reality

Predicted, Actual and Errors for Schools:
LES: 21.7826 22.0248 -0.242194
WES: 27.8397 27.8767 -0.0369861
MtG: 50.3776 50.0984 0.279181
Sum of three predictions is 100%

Total chapter 70 funds in 2018: 3,489,437.
1% of total is 34,894.40.
.3% of total is 10,468.31.

School budgets (roughly): LES $2.7 million, WES $6.6 million, MtG $11 million.
Many non-linear relationships are linear after applying logarithms:

\[ Y = BX^a \text{ then } \log(Y) = a \log(X) + b, \quad b = \log B. \]
Many non-linear relationships are linear after applying logarithms:

\[ Y = BX^a \text{ then } \log(Y) = a \log(X) + b, \quad b = \log B. \]

Kepler’s Third Law: if \( T \) is the orbital period of a planet traveling in an elliptical orbit about the sun (and no other objects exist), then \( T^2 = \tilde{B}L^3 \), where \( L \) is the length of the semi-major axis.

Assume do not know this – can we *discover* through statistics?
Many non-linear relationships are linear after applying logarithms:

\[ Y = BX^a \text{ then } \log(Y) = a \log(X) + b, \quad b = \log B. \]

Kepler’s Third Law: if \( T \) is the orbital period of a planet traveling in an elliptical orbit about the sun (and no other objects exist), then \( T = BL^{1.5} \), where \( L \) is the length of the semi-major axis.

Assume do not know this – can we discover through statistics?
Kepler’s Third Law: Can see the 1.5 exponent!

Data: Semi-major axis: Mercury 0.387, Venus 0.723, Earth 1.000, Mars 1.524, Jupiter 5.203, Saturn 9.539, Uranus 19.182, Neptune 30.06 (the units are astronomical units, where one astronomical unit is $1.496 \cdot 10^8$ km).

Data: orbital periods (in years) are 0.2408467, 0.61519726, 1.0000174, 1.8808476, 11.862615, 29.447498, 84.016846 and 164.79132.

If $T = B L^a$, what should $B$ equal with this data? Units: bruno, millihelen, slug, smoot, .... See https://en.wikipedia.org/wiki/List_of_humorous_units_of_measurement
Kepler’s Third Law: Can see the 1.5 exponent!

Data: Semi-major axis: Mercury 0.387, Venus 0.723, Earth **1.000**, Mars 1.524, Jupiter 5.203, Saturn 9.539, Uranus 19.182, Neptune 30.06 (the units are astronomical units, where one astronomical unit is $1.496 \cdot 10^8$ km).

Data: orbital periods (in years) are 0.2408467, 0.61519726, **1.0000174**, 1.8808476, 11.862615, 29.447498, 84.016846 and 164.79132.

If $T = BL^a$, what should $B$ equal with this data? Units: bruno, millihelen, slug, smoot, .... See https://en.wikipedia.org/wiki/List_of_humorous_units_of_measurement
Kepler’s Third Law: Can see the 1.5 exponent!

If try $\log T = a \log L + b$: best fit values are...?

HOMEWORK!

Figure: Plot of $\log P$ versus $\log L$ for planets. Is it surprising $b \approx 0$ (so $B \approx 1$ or $b \approx 0$?)
Units: Goal: find good statistics to describe the world.

Figure: Harvard Bridge, about 620.1 meters.
Units: Goal: find good statistics to describe the world.

Figure: Harvard Bridge, 364.1 Smoots (± one ear).
Units: Goal: find good statistics to describe the world.

Sieze opportunities: Never know where they will lead.

Birthday Problem: Assume a year with $D$ days, how many people do we need in a room to have a 50% chance that at least two share a birthday, under the assumption that the birthdays are independent and uniformly distributed from 1 to $D$?
Birthday Problem: Assume a year with $D$ days, how many people do we need in a room to have a 50% chance that at least two share a birthday, under the assumption that the birthdays are independent and uniformly distributed from 1 to $D$?

An analysis shows the answer is approximately $D^{1/2} \sqrt{\log 4}$.

Can do simulations and try and see the correct exponent; will look not for 50% chance but the expected number of people in room for the first collision.
Try $P = BD^a$, take logs so $\log P = a \log D + b$ ($b = \log B$).

**Figure:** Plot of best fit line for $P$ as a function of $D$. We twice ran 10,000 simulations with $D$ chosen from 10,000 to 100,000. Best fit values were $a \approx 0.506167$, $b \approx -0.0110081$ (left) and $a \approx 0.48141$, $b \approx 0.230735$ (right).