Lecture 15 Linear Least-Square Regression (1)

Objectives

- Understanding data fitting problems
- Understanding residual error and sum of residual errors
- Knowing how to manually calculate the slope and intercept of a best-fit straight line with least-square regression

Introduction

- Find a straight line that passes through the following points (1,7), (2,12) and (3,15)

\[
\begin{align*}
7 & = k + b \\
12 & = 2k + b \\
15 & = 3k + b
\end{align*}
\]

Equation of a straight line: \( y = kx + b \)

Introduction

- There are more equations than unknowns. Now let’s try to solve the problem using the first two equations:

\[
\begin{align*}
7 & = k + b \\
12 & = 2k + b
\end{align*} \Rightarrow \begin{align*}
k & = 5 \\
b & = 2
\end{align*}
\]

However, it does not satisfy the last equation:

\( 3 \times 5 + 2 = 17 \neq 15 \)

Discrepancy=17-15=+2

Introduction

- Graphically it looks like this:

Introduction

Similarly, you can use equations 1 and 3 or equations 2 and 3 to solve for k and b, but the resulting lines will not pass the third point.
Introduction

So, among the three attempted solutions, which one is ‘better’?

- **discrepancy** = 2
- **discrepancy** = -1
- **discrepancy** = 2

Solution 1

Solution 2

Solution 3

Residual Error

- For a straight line $y = kx + b$ that goes through points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, the residual error for each point is defined as:

  $e_i = (kx_i + b) - y_i$

  $e_2 = (kx_2 + b) - y_2$

  $\ldots$

  $e_n = (kx_n + b) - y_n$

Measurement of “Best Fit”

1. Sum of the residual errors for all the available data: $\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (kx_i + b - y_i)$

2. Sum of the absolute values of the residual errors: $\sum_{i=1}^{n} |e_i| = \sum_{i=1}^{n} |kx_i + b - y_i|$

3. Sum of the squares of the residual errors: $S_r = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (kx_i + b - y_i)^2$

Least-Square Regression

- To determine $k$ and $b$ so that $S_r$ is minimized, we need to solve:

  $\frac{\partial S_r}{\partial k} = 2\sum_{i=1}^{n} x_i (kx_i + b - y_i) = 2k\sum_{i=1}^{n} x_i^2 + 2b\sum_{i=1}^{n} x_i - 2\sum_{i=1}^{n} x_i y_i = 0$

  $\frac{\partial S_r}{\partial b} = \sum_{i=1}^{n} (kx_i + b - y_i) = 2b\sum_{i=1}^{n} x_i + 2nb - 2\sum_{i=1}^{n} y_i = 0$

  Which is equivalent to solve the following:

  $(\sum x_i^2)k + (\sum x_i) b = \sum x_i y_i$

  $(\sum x_i) k + nb = \sum y_i$

Least-Square Regression

- Simultaneously solve for $k$ and $b$ from the two equations, we have

  $k = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2}$

  $b = \frac{\sum y_i}{n} - \frac{\sum x_i}{n} k = \bar{y} - \bar{x} k$
Example
• Find a straight line that passes through the following points (1,7), (2,12) and (3,15) using least square regression

\[
\begin{array}{ccccc}
\text{point #} & x_i & y_i & x_i^2 & x_iy_i \\
1 & 1 & 7 & 1 & 7 \\
2 & 2 & 12 & 4 & 24 \\
3 & 3 & 15 & 9 & 45 \\
\text{sum} & 6 & 34 & 14 & 76 \\
\end{array}
\]

\[
k = \frac{3 \times 76 - 6 \times 34}{3 \times 14 - 6^2} = 4 \\
b = \frac{34}{3} - \frac{6 \times 4}{3} = \frac{10}{3} = 3.333
\]

Thus, the resulting straight line is

\[
y = 4.0x + 3.333
\]

Example
• The solution \( k=4.0, b=3.333 \) that minimizes the error, as measured by the sum of the square of the error at individual points, is called the least square regression solution of the original problem:

\[
\begin{align*}
k + b &= 7 \\
2k + b &= 12 \\
3k + b &= 15
\end{align*}
\]

Exercise
The drag force experienced by a free-falling object in the air is a function of the velocity. The following data are collected during an experiment. Fit a straight line through the data points using least-square regression.

<table>
<thead>
<tr>
<th>data #</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>( v \text{ (m/s)} )</td>
<td>10</td>
<td>20</td>
<td>30</td>
<td>40</td>
<td>50</td>
<td>60</td>
<td>70</td>
<td>80</td>
</tr>
<tr>
<td>( F \text{ (N)} )</td>
<td>25</td>
<td>70</td>
<td>380</td>
<td>550</td>
<td>610</td>
<td>1220</td>
<td>830</td>
<td>1450</td>
</tr>
</tbody>
</table>