Method of Least squares

1. Recall, when fitting a straight line \( y = a + bx \) to the data \((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\), the least squares estimates \( \hat{a}, \hat{b} \) for the parameters \( a, b \) are given by

\[
\hat{b} = \frac{\sum_{i=1}^{n} x_i y_i - n \overline{x} \overline{y}}{\sum_{i=1}^{n} x_i^2 - n \overline{x}^2}, \quad \hat{a} = \overline{y} - \hat{b} \overline{x}
\]

where \( \overline{x} := \frac{1}{n} \sum_{i=1}^{n} x_i \) and \( \overline{y} := \frac{1}{n} \sum_{i=1}^{n} y_i \).

The extension of a spring is recorded when various sized masses are attached.

\[
\begin{array}{c|cccccc}
\text{Mass (g)} & 10 & 20 & 30 & 40 & 50 \\
\hline
\text{Extension (mm)} & 4.6 & 9.8 & 15.2 & 20.0 & 25.4 \\
\end{array}
\]

Fit a straight line through the data using the method of least squares.

Give an estimate of the extension for masses of (i) 25g, and (ii) 100g. Do you have any reservations about either estimate? Why?

2. (a) A sequence of \( n \) measurements \((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\) are taken in an experiment where the response variable \( y \) is thought to vary quadratically in the controlled variable \( x \). That is, for some constants \( a, b, c \), we assume that

\[
y_i = a + bx_i + cx_i^2 + \epsilon_i \quad (i = 1, \ldots, n)
\]

where \( \epsilon_i \) is the error in the \( i \)-th measurement.

Consider minimising \( S(a, b, c) := \sum_i \epsilon_i^2 \) to show that any least squares estimate \((\hat{a}, \hat{b}, \hat{c})\) for the parameters \((a, b, c)\) must satisfy

\[
\begin{align*}
n \hat{a} + \left( \sum_{i=1}^{n} x_i \right) \hat{b} + \left( \sum_{i=1}^{n} x_i^2 \right) \hat{c} &= \sum_{i=1}^{n} y_i \\
\left( \sum_{i=1}^{n} x_i \right) \hat{a} + \left( \sum_{i=1}^{n} x_i^2 \right) \hat{b} + \left( \sum_{i=1}^{n} x_i^3 \right) \hat{c} &= \sum_{i=1}^{n} x_i y_i \\
\left( \sum_{i=1}^{n} x_i^2 \right) \hat{a} + \left( \sum_{i=1}^{n} x_i^3 \right) \hat{b} + \left( \sum_{i=1}^{n} x_i^4 \right) \hat{c} &= \sum_{i=1}^{n} x_i^2 y_i
\end{align*}
\]

These are the least squares normal equations for quadratic regression.

(b) \( x := t - 123 \) (why?).

Let \( y \) be the separation and \( t \) the temperature. Define \( x := t - 123 \) (why?).

Using the method of least squares, find the best fit to the data amongst the family of quadratic curves \( y = a + bx + cx^2 \).

Estimate the separations of the parts at temperatures of (i) 125.5°C and (ii) 140°C. Is any caution required in using either estimate?
Mathematics tutorial sheet 5: Solutions

Method of Least Squares

• 1. Tabulating the data (else using a fancy calculator!)

<table>
<thead>
<tr>
<th>i</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_i$</td>
<td>10</td>
<td>20</td>
<td>30</td>
<td>40</td>
<td>50</td>
</tr>
<tr>
<td>$y_i$</td>
<td>4.6</td>
<td>9.8</td>
<td>15.2</td>
<td>20.0</td>
<td>25.4</td>
</tr>
<tr>
<td>$x_i^2$</td>
<td>100</td>
<td>400</td>
<td>900</td>
<td>1600</td>
<td>2500</td>
</tr>
</tbody>
</table>

$n = 5$ \[ \sum x_i = 150 \]

\[ \sum y_i = 75.0 \]

\[ \sum x_i y_i = 5500 \]

\[ \sum x_i^2 y_i = 27680.0 \]

and \[ \bar{x} = \sum x_i/n = 150/5 = 30, \quad \bar{y} = \sum y_i/n = 75/5 = 15. \]

• When we fit a straight line $y = a + bx$ to the data, the least squares estimates for the parameters $a, b$ are given by

\[ \hat{b} = \frac{\sum_{i=1}^{n} x_i y_i - n \bar{x} \bar{y}}{\sum_{i=1}^{n} x_i^2 - n \bar{x}^2} = \frac{2768 - 5 \times 30 \times 15}{5500 - 5 \times (30)^2} = \frac{518}{1000} = 0.518 \]

\[ \hat{a} = \bar{y} - \hat{b} \bar{x} = 15 - 0.518 \times 30 = -0.54 \]

(Note: in general, when using such formulae be careful to maintain accuracy - watch out for rounding errors!)

• Thus, the line that gives the least squares fit to the data is $y = -0.54 + (0.518)x$.

• The estimates of the extension of the spring for masses of 25 and 100 are

\[ \hat{y}_{25} = -0.54 + (0.518)25 = 12.41 \]

\[ \hat{y}_{100} = -0.54 + (0.518)100 = 51.26 \]

We should have confidence in the estimate for 25g as this is within the range of the data used to fit the curve. However, the estimate for 100g should be used with caution as it is an estimate well outside the range (10−50g) that the data covers. In this example, the spring may well have passed its elastic limit by 100g and then the estimate would be very unsatisfactory.

• 2. (a) We wish to find the quadratic that minimises the sum of the squared errors over the data. Let

\[ S(a, b, c) := \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (a + bx_i + cx_i^2 - y_i)^2. \]

Then for a minimum of $S$ at $(\hat{a}, \hat{b}, \hat{c})$ we require

\[ \frac{\partial S}{\partial a}(\hat{a}, \hat{b}, \hat{c}) = 0, \quad \frac{\partial S}{\partial b}(\hat{a}, \hat{b}, \hat{c}) = 0, \quad \frac{\partial S}{\partial c}(\hat{a}, \hat{b}, \hat{c}) = 0 \]

[and strictly speaking we should also check that the matrix of second derivatives is positive definite at the point $(\hat{a}, \hat{b}, \hat{c})$ to guarantee that we have indeed found a minimum, but you are NOT expected to check this].

• The partial derivatives are

\[ \frac{\partial S}{\partial a} = \sum_{i=1}^{n} 2(a + bx_i + cx_i^2 - y_i), \]

\[ \frac{\partial S}{\partial b} = \sum_{i=1}^{n} 2x_i(a + bx_i + cx_i^2 - y_i), \]

\[ \frac{\partial S}{\partial c} = \sum_{i=1}^{n} 2x_i^2(a + bx_i + cx_i^2 - y_i). \]
Any point \((\hat{a}, \hat{b}, \hat{c})\) where these are simultaneously equal to zero then requires

\[
\sum_{i=1}^{n} (\hat{a} + \hat{b}x_i + \hat{c}x_i^2 - y_i) = n\hat{a} + \hat{b}\sum_{i=1}^{n} x_i + \hat{c}\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} y_i = 0,
\]

\[
\sum_{i=1}^{n} x_i (\hat{a} + \hat{b}x_i + \hat{c}x_i^2 - y_i) = \hat{a}\sum_{i=1}^{n} x_i + \hat{b}\sum_{i=1}^{n} x_i^2 + \hat{c}\sum_{i=1}^{n} x_i^3 - \sum_{i=1}^{n} x_i y_i = 0,
\]

\[
\sum_{i=1}^{n} x_i^2 (\hat{a} + \hat{b}x_i + \hat{c}x_i^2 - y_i) = \hat{a}\sum_{i=1}^{n} x_i^2 + \hat{b}\sum_{i=1}^{n} x_i^3 + \hat{c}\sum_{i=1}^{n} x_i^4 - \sum_{i=1}^{n} x_i^2 y_i = 0,
\]

which yield the required least squared normal equations for quadratic regression.

- [Recall, the definition of the summation notation:
  \[
  \sum_{i=1}^{n} x_i = x_1 + x_2 + \ldots + x_n
  \]
  With this in mind, we see that
  \[
  \sum_{i=1}^{n} \{x_i + y_i\} = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i
  \]
  since we can just rearrange the order in which we add up the terms
  \[
  \{x_1 + y_1\} + \{x_2 + y_2\} + \ldots + \{x_n + y_n\} = \{x_1 + \ldots + x_n\} + \{y_1 + \ldots + y_n\}.
  \]
  And we also have the simple facts that for any constant \(\alpha\) (that is, any \(\alpha\) that doesn’t depend on the dummy variable \(i\) of the summation)
  \[
  \sum_{i=1}^{n} \alpha x_i = \alpha \sum_{i=1}^{n} x_i,
  \]
  \[
  \sum_{i=1}^{n} \alpha = n\alpha
  \]
  since
  \[
  \alpha x_1 + \alpha x_2 + \ldots + \alpha x_n = \alpha \{x_1 + \ldots + x_n\},\quad \text{etc.}
  \]
  The summation convention is very neat, but sometimes it is worth writing out exactly what it means (as just above) so you don’t forget obvious properties!]

- (b) Defining \(x = t - 123\) (so that, \(x_1 = t_1 - 123 = 120 - 123 = -3\), etc.) is chosen as \(T = 123\) is the mean of the temperatures of the data (as \(T := \sum t_i/n = \{120 + 121 + \ldots + 126\}/123\) and, as you can check, centering the temperature about the mean simplifies the calculations since we now clearly have \(\sum x_i = 0\), and also \(\sum x_i^2 = 0\) by the symmetry. The remaining summations for \(x\) values are much more manageable than would be the sums for \(t\) values (eg. 1244 is much worse to handle than 141!).

- You should find (such as by tabulating as in question 1) that
  \[
  \sum x_i^2 = 28,\quad \sum x_i^4 = 196,\quad \sum y_i = 49.0,\quad \sum x_i y_i = 55.7,\quad \sum x_i^2 y_i = 284.5
  \]
  the least squares normal equations for quadratic regression from part (a) now reduce to
  \[
  7\hat{a} + 28\hat{c} = 49.0
  \]
  \[
  28\hat{b} = 55.7
  \]
  \[
  28\hat{a} + 196\hat{c} = 284.5
  \]
  then immediately we get \(\hat{b} = 55.7/28 = 1.989285714\) and solving the remaining pair of equations gives \(\hat{a} = 2.785714286\), \(\hat{c} = 1.053571429\).
• The best estimate at $t = 125.5$ ie. $x = 2.5$, is given by

$$\hat{a} + \hat{b}(2.5) + \hat{c}(2.5)^2 = 14.34375001 = 14.3(1 \text{ d.p})$$

• The best estimate at $t = 140$ ie. $x = 17$ is

$$\hat{a} + \hat{b}(17) + \hat{c}(17)^2 = 341.0857144 = 341.1(1 \text{ d.p})$$

Much caution is required with this estimate as it is way outside the range of temperatures that the curve was fitted to.

[NOTE: when making calculations take care not to let rounding errors ruin your answers - try and use the most accurate values you have in formulas and save the rounded values for final answers only.]