Introduction: Matching a Few Parameters to a Lot of Data.

Sometimes we get a lot of data, many observations, and want to fit it to a simple model.

Low dimensional models (e.g. low degree polynomials) are easy to work with, and are quite well behaved (high degree polynomials can be quite oscillatory.)

All measurements are noisy, to some degree. Often, we want to use a large number of measurements in order to “average out” random noise.

Approximation Theory looks at two problems:

[1] Given a data set, find the best fit for a model (i.e. in a class of functions, find the one that best represents the data.)

Discrete Least Squares: Linear Approximation.

The form of Least Squares you are most likely to see: **Find the Linear Function**, \( p_1(x) = a_0 + a_1 x \), that best fits the data. The error \( E(a_0, a_1) \) we need to minimize is:

\[
E(a_0, a_1) = \sum_{i=0}^{n} [(a_0 + a_1 x_i) - y_i]^2.
\]

The first partial derivatives with respect to \( a_0 \) and \( a_1 \) better be zero at the minimum:

\[
\frac{\partial}{\partial a_0} E(a_0, a_1) = 2 \sum_{i=0}^{n} [(a_0 + a_1 x_i) - y_i] = 0
\]

\[
\frac{\partial}{\partial a_1} E(a_0, a_1) = 2 \sum_{i=0}^{n} x_i [(a_0 + a_1 x_i) - y_i] = 0.
\]

We “massage” these expressions to get the Normal Equations...

### Quadratic Model, \( p_2(x) \)

For the quadratic polynomial \( p_2(x) = a_0 + a_1 x + a_2 x^2 \), the error is given by

\[
E(a_0, a_1, a_2) = \sum_{i=0}^{n} [(a_0 + a_1 x_i + a_2 x_i^2) - y_i]^2
\]

At the minimum (best model) we must have

\[
\frac{\partial}{\partial a_0} E(a_0, a_1, a_2) = 2 \sum_{i=0}^{n} [(a_0 + a_1 x_i + a_2 x_i^2) - y_i] = 0
\]

\[
\frac{\partial}{\partial a_1} E(a_0, a_1, a_2) = 2 \sum_{i=0}^{n} x_i [(a_0 + a_1 x_i + a_2 x_i^2) - y_i] = 0
\]

\[
\frac{\partial}{\partial a_2} E(a_0, a_1, a_2) = 2 \sum_{i=0}^{n} x_i^2 [(a_0 + a_1 x_i + a_2 x_i^2) - y_i] = 0.
\]

Linear Approximation: The Normal Equations

Given a set of \( n \) data points \( (x_i, y_i) \), the objective is to find the parameters \( a_0, a_1 \) that minimize the sum of squares of residuals:

\[
E(a_0, a_1) = \sum_{i=0}^{n} [(a_0 + a_1 x_i) - y_i]^2
\]

To find the minimum, we set the first partial derivatives to zero:

\[
\frac{\partial}{\partial a_0} E(a_0, a_1) = \sum_{i=0}^{n} (a_0 + a_1 x_i - y_i) = 0
\]

\[
\frac{\partial}{\partial a_1} E(a_0, a_1) = \sum_{i=0}^{n} x_i (a_0 + a_1 x_i - y_i) = 0
\]

These equations are known as the normal equations.

For the quadratic polynomial \( p_2(x) = a_0 + a_1 x + a_2 x^2 \), the normal equations are:

\[
\begin{align*}
  a_0 (n+1) + a_1 \sum_{i=0}^{n} x_i &= \sum_{i=0}^{n} y_i \\
  a_0 \sum_{i=0}^{n} x_i + a_1 \sum_{i=0}^{n} x_i^2 &= \sum_{i=0}^{n} x_i y_i
\end{align*}
\]

Note: Even though the model is quadratic, the resulting (normal) equations are linear. — The model is linear in its parameters, \( a_0, a_1, \) and \( a_2 \).
We rewrite the Normal Equations as:

\[
\begin{bmatrix}
(n + 1) \sum_{i=0}^{n} x_i & \sum_{i=0}^{n} x_i^2 \\
\sum_{i=0}^{n} x_i & \sum_{i=0}^{n} x_i^2 & \sum_{i=0}^{n} x_i^3 \\
\sum_{i=0}^{n} x_i^2 & \sum_{i=0}^{n} x_i^3 & \sum_{i=0}^{n} x_i^4
\end{bmatrix}
\begin{bmatrix}
a_0 \\
a_1 \\
a_2
\end{bmatrix}
= 
\begin{bmatrix}
\sum_{i=0}^{n} y_i \\
\sum_{i=0}^{n} x_i y_i
\end{bmatrix}.
\]

It is not immediately obvious, but this expression can be written in the form \( A^T A \tilde{a} = A^T \tilde{y} \). Where the matrix \( A \) is very easy to write in terms of \( x_i \). [Jump Forward].

### Application: Cricket Thermometer

**Snowy Tree Cricket**

- Folk method for finding temperature (Fahrenheit)
  - **Count the number of chirps in a minute and divide by 4, then add 40**
- In 1898, A. E. Dolbear [3] noted that
  - "crickets in a field [chirp] synchronously, keeping time as if led by the wand of a conductor"
- He wrote down a formula in a scientific publication (first?)
  \[
  T = 50 + \frac{N - 40}{4}
  \]

Mathematical models for chirping of snowy tree crickets, *Oecanthulus fultoni*, are **Linear Models**

- Data from C. A. Bessey and E. A. Bessey [2] (8 crickets) from Lincoln, Nebraska during August and September, 1897 (shown on next slide)
- The least squares best fit line to the data is

\[ T = 60 + \frac{N - 92}{4.7} \]


---

**Biological Questions – Cricket Model**

How well does the line fitting the Bessey & Bessey data agree with the Dolbear model given above?

- Graph shows Linear model fits the data well
- Data predominantly below **Folk/Dolbear** model
- Possible discrepancies
  - Different cricket species
  - Regional variation
  - Folk only an approximation
- Graph shows only a few °F difference between models

When can this model be applied from a practical perspective?

- Biological thermometer has limited use
- Snowy tree crickets only chirp for a couple months of the year and mostly at night
- Temperature needs to be above 50°F
**Discrete Least Squares Application: Cricket Thermometer**

**Mathematical Questions – Cricket Model 1**

Over what range of temperatures is this model valid?
- Biologically, observations are mostly between 50°F and 85°F
- Thus, limited range of temperatures, so limited range on the Linear Model
- **Range of Linear functions** affects its **Domain**
- From the graph, **Domain** is approximately 50–200 Chirps/min

**Cricket Data Analysis**

C. A. Bessey and E. A. Bessey collected data on eight different crickets that they observed in Lincoln, Nebraska during August and September, 1897. The number of chirps/min was \( N \) and the temperature was \( T \).

Create matrices

\[
A_1 = \begin{pmatrix}
1 & N_1 \\
1 & N_2 \\
\vdots & \vdots
\end{pmatrix},
A_2 = \begin{pmatrix}
1 & N_1 & N_1^2 \\
1 & N_2 & N_2^2 \\
\vdots & \vdots & \vdots
\end{pmatrix},
A_3 = \begin{pmatrix}
1 & N_1 & N_1^2 & N_1^3 \\
1 & N_2 & N_2^2 & N_2^3 \\
\vdots & \vdots & \vdots & \vdots
\end{pmatrix},
A_4 = \begin{pmatrix}
1 & N_1 & N_1^2 & N_1^3 & N_1^4 \\
1 & N_2 & N_2^2 & N_2^3 & N_2^4 \\
\vdots & \vdots & \vdots & \vdots & \vdots
\end{pmatrix}
\]

**Cricket Linear Model**

If you compute the matrix *which you never should!*

\[
A_1^T A_1 = \begin{pmatrix} 52 & 7447 \\ 7447 & 1133259 \end{pmatrix},
\]

it has eigenvalues

\[\lambda_1 = 3.0633 \quad \text{and} \quad \lambda_2 = 1,133,308,\]

which gives the condition number

\[\text{cond}(A_1^T A_1) = 3.6996 \times 10^5.\]

Whereas

\[\text{cond}(A_1) = 608.2462.\]

In Matlab

\[A_1 \backslash T\]

gives the parameters for best linear model

\[T_1(N) = 0.2155 N + 39.7441.\]
Discrete Least Squares
Application: Cricket Thermometer

Polynomial Fits to the Data: Linear

Similarly, the matrix

\[
A_2^T A_2 = \begin{pmatrix}
52 & 7447 & 1133259 \\
7447 & 1133259 & 1.8113 \times 10^8 \\
1133259 & 1.8113 \times 10^8 & 3.0084 \times 10^{10}
\end{pmatrix},
\]

has eigenvalues

\[
\lambda_1 = 0.1957, \quad \lambda_2 = 42,706, \quad \lambda_3 = 3.00853 \times 10^{10}
\]

which gives the condition number

\[
\text{cond}(A_2^T A_2) = 1.5371 \times 10^{11}.
\]

Whereas,

\[
\text{cond}(A_2) = 3.9206 \times 10^5,
\]

and

\[
A_2 \backslash T,
\]

gives the parameters for best quadratic model

\[
T_2(N) = -0.00064076 N^2 + 0.39625 N + 27.8489.
\]

Joseph M. Mahaffy, \langle jmahaffy@mail.sdsu.edu \rangle Lecture Notes – Least Squares — (22/29)

Polynomial Fits to the Data: Quadratic

The condition numbers for the cubic and quartic rapidly get larger with

\[
\text{cond}(A_3^T A_3) = 6.3648 \times 10^{16} \quad \text{and} \quad \text{cond}(A_4^T A_4) = 1.1218 \times 10^{23}
\]

These last two condition numbers suggest that any coefficients obtained are highly suspect.

However, if done right, we are “only” subject to the condition numbers

\[
\text{cond}(A_3) = 2.522 \times 10^8, \quad \text{cond}(A_4) = 1.738 \times 10^{11}.
\]

The best cubic and quartic models are given by

\[
T_3(N) = 0.0000018977 N^3 - 0.001445 N^2 + 0.50540 N + 23.138
\]

\[
T_4(N) = -0.0000001765 N^4 + 0.00001190 N^3 - 0.003504 N^2 + 0.6876 N + 17.314
\]

Joseph M. Mahaffy, \langle jmahaffy@mail.sdsu.edu \rangle Lecture Notes – Least Squares — (24/29)
So how does one select the best model? Visually, one can see that the linear model does a very good job, and one only obtains a slight improvement with a quadratic. Is it worth the added complication for the slight improvement?

It is clear that the sum of square errors (SSE) will improve as the number of parameters increase. From statistics, it is hotly debated how much penalty one should pay for adding parameters.

**Bayesian Information Criterion**

Let \( n \) be the number of data points, \( SSE \) be the sum of square errors, and let \( k \) be the number of parameters in the model.

\[
BIC = n \ln(SSE/n) + k \ln(n).
\]

**Akaike Information Criterion**

\[
AIC = 2k + n(\ln(2\piSSE/n) + 1).
\]
The table below shows the by the Akaike information criterion that we should take a quadratic model, while using a Bayesian Information Criterion we should use a cubic model.

<table>
<thead>
<tr>
<th></th>
<th>Linear</th>
<th>Quadratic</th>
<th>Cubic</th>
<th>Quartic</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SSE</strong></td>
<td>108.8</td>
<td>79.08</td>
<td>78.74</td>
<td>78.70</td>
</tr>
<tr>
<td><strong>BIC</strong></td>
<td>46.3</td>
<td>33.65</td>
<td>33.43</td>
<td>37.35</td>
</tr>
<tr>
<td><strong>AIC</strong></td>
<td>189.97</td>
<td><strong>175.37</strong></td>
<td>177.14</td>
<td>179.12</td>
</tr>
</tbody>
</table>

Returning to the original statement, we do fairly well by using the folk formula, despite the rest of this analysis!