Deterministic Analysis of Observations - Chapter 1 of Heinz

MATH 123: Mathematical Modeling, Spring 2019
Dr. Doreen De Leon

1 Data Transformations: Linear Models - Section 1.2 of Heinz

One of the simplest modeling approaches involves a linear model. Of course, most real-world data that we might be interested in will not follow a linear function. However, we may sometimes transform the data to allow the use of linear models. This chapter considers examples of this approach.

We will evaluate the performance of the models in this chapter by determining the maximum relative error related to all of the data points. Given a set of data points, \((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\), the relative error related to each data point, denoted \(e_i\), is determined by

\[
e_i = \frac{y_i^{(mod)} - y_i}{y_i}.
\]

NOTE: This does not agree with the definition in the textbook. One reason for our use of this definition, as opposed to the one in the textbook, is that if the model is flawed, then it does not represent the “true” value for that data point.

Given any two data points \((x_1, y_1)\) and \((x_2, y_2)\), we may write the equation of the line passing through them in the form

\[
y = y_1 \frac{x - x_2}{x_1 - x_2} + y_2 \frac{x - x_1}{x_2 - x_1},
\]

known as the Lagrange form of the linear polynomial. It is easy to show that this is equivalent to the standard two-point formula

\[
y - y_1 = m(x - x_1),
\]

where

\[
m = \frac{y_2 - y_1}{x_2 - x_1}.
\]

Example 1.1. Consider the data of the height and weight of 10 students shown in Table 1.1.
Table 1.1: Height (in inches) and weight (in pounds) of 10 students.

<table>
<thead>
<tr>
<th>height</th>
<th>weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>63</td>
<td>127</td>
</tr>
<tr>
<td>64</td>
<td>121</td>
</tr>
<tr>
<td>66</td>
<td>142</td>
</tr>
<tr>
<td>69</td>
<td>157</td>
</tr>
<tr>
<td>69</td>
<td>162</td>
</tr>
<tr>
<td>71</td>
<td>156</td>
</tr>
<tr>
<td>71</td>
<td>169</td>
</tr>
<tr>
<td>72</td>
<td>165</td>
</tr>
<tr>
<td>73</td>
<td>181</td>
</tr>
<tr>
<td>75</td>
<td>208</td>
</tr>
</tbody>
</table>

The data appears fairly linear, when plotted. Based on this plot, it appears reasonable to use, e.g., the points \((71, 169)\) and \((73, 181)\) to find a line that models the data. Doing this gives

\[
y = \frac{169}{-2}x - \frac{73}{-2} + \frac{181}{2}x - \frac{71}{2} = 6x - 257.
\]

The plot is shown below in Figure 1.1 below.

Figure 1.1: Plot of data points and “interpolating” line.

Computing the relative error for each of these points gives:

The maximum relative error, then, is 8%, which is not too bad.
Table 1.2: Relative error of model for weight as a function of height.

<table>
<thead>
<tr>
<th>height</th>
<th>63</th>
<th>64</th>
<th>66</th>
<th>69</th>
<th>71</th>
<th>71</th>
<th>72</th>
<th>73</th>
<th>75</th>
</tr>
</thead>
<tbody>
<tr>
<td>weight</td>
<td>127</td>
<td>121</td>
<td>142</td>
<td>157</td>
<td>162</td>
<td>156</td>
<td>169</td>
<td>165</td>
<td>181</td>
</tr>
<tr>
<td>rel. error</td>
<td>-0.047</td>
<td>0.050</td>
<td>-0.021</td>
<td>0</td>
<td>-0.031</td>
<td>0.08</td>
<td>0</td>
<td>0.06</td>
<td>0</td>
</tr>
</tbody>
</table>

Next, we will consider an example for which we need to modify the data.

**Example 1.2.** Consider the following data, giving the weight $W$ of a fish (specifically, bass) versus its length, $l$.

Table 1.3: Data for weight and length of bass fish.

<table>
<thead>
<tr>
<th>length $l$ (in.)</th>
<th>12.5</th>
<th>12.625</th>
<th>12.625</th>
<th>14.125</th>
<th>14.5</th>
<th>14.5</th>
<th>17.25</th>
<th>17.75</th>
</tr>
</thead>
<tbody>
<tr>
<td>weight $W$ (oz.)</td>
<td>17</td>
<td>16</td>
<td>17</td>
<td>23</td>
<td>26</td>
<td>27</td>
<td>41</td>
<td>49</td>
</tr>
</tbody>
</table>

This data is plotted in Figure 1.2 below.

![Weight vs. Length of Bass](image)

Figure 1.2: The length and weight of eight bass.

The points appear to be more or less quadratic, so let’s assume that the model has the form $W = kl^2$ and see if this is correct. To do this using our linear model, we will take the natural log of both sides, giving

$$\ln(W) = \ln(k) + 2 \ln(l).$$

We then need the natural log of our data, which is given in the following table.

From the plot of this data (in Figure 1.3), we observe that the points appear to lie (more or less) on a line.

Using the points $(2.648, 3.135)$ and $(2.848, 3.714)$ gives the line

$$y = 3.714 \cdot \frac{x - 2.648}{2.848 - 2.648} + 3.313 \cdot \frac{x - 2.848}{2.648 - 2.848} \approx 2.005x - 1.996.$$
Table 1.4: Natural log of data for bass fish.

<table>
<thead>
<tr>
<th>log of length $l$ (in.)</th>
<th>2.526</th>
<th>2.536</th>
<th>2.536</th>
<th>2.648</th>
<th>2.674</th>
<th>2.674</th>
<th>2.848</th>
<th>2.876</th>
</tr>
</thead>
<tbody>
<tr>
<td>log of weight $W$ (oz.)</td>
<td>2.833</td>
<td>2.773</td>
<td>2.833</td>
<td>3.135</td>
<td>3.258</td>
<td>3.296</td>
<td>3.714</td>
<td>3.892</td>
</tr>
</tbody>
</table>

Figure 1.3: Plot of natural log of weight vs. the length of eight bass.

Therefore, $k = e^{-1.996} = 0.136$ and we see that the slope is approximately 2, as we expected. So, the model for the weight of the bass as a function of the length is

$$W = 0.136l^2.$$ 

How well does this compare to the data? We can look at the relative error in Table 1.5.

<table>
<thead>
<tr>
<th>length $l$ (in.)</th>
<th>12.5</th>
<th>12.625</th>
<th>12.625</th>
<th>14.125</th>
<th>14.5</th>
<th>14.5</th>
<th>17.25</th>
<th>17.75</th>
</tr>
</thead>
<tbody>
<tr>
<td>weight $W$ (oz.)</td>
<td>17</td>
<td>16</td>
<td>17</td>
<td>23</td>
<td>26</td>
<td>27</td>
<td>41</td>
<td>49</td>
</tr>
<tr>
<td>rel. error</td>
<td>0.25</td>
<td>0.255</td>
<td>0.275</td>
<td>0.180</td>
<td>0.100</td>
<td>0.059</td>
<td>-0.013</td>
<td>-0.126</td>
</tr>
</tbody>
</table>

The largest relative error is 0.255, or approximately 26%, which might seem to indicate that the original model is rather poor. This is not actually the case, as we will see later when we discuss an alternate way to find the best fit for the data.

We will consider one more example, that of exponential growth.

**Example 1.3.** Consider the data on the population of a certain species of bacteria given in Table 1.6.
Table 1.6: Bacteria population after $t$ hours.

<table>
<thead>
<tr>
<th>$t$ (hr.)</th>
<th>7</th>
<th>14</th>
<th>21</th>
<th>28</th>
<th>35</th>
<th>42</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P$</td>
<td>8</td>
<td>41</td>
<td>133</td>
<td>250</td>
<td>280</td>
<td>297</td>
</tr>
</tbody>
</table>

A plot of the data indicates that the data has exponential growth (at least for most of the data points). So, assuming that the model is of the form

$$P = ae^{bt},$$

we can recast this as a linear model by taking the natural log of both sides, obtaining

$$\ln(P) = \ln(a) + bt.$$ 

If we plot the logarithm of the population $P$ versus $t$, we do, in fact, see data that follows approximately a line (again, for most of the data). Let’s use the first point, $(7, \ln(8))$, and $(21, \ln(133))$ to determine the parameters for the model,

$$b \approx 0.2008 \text{ and } \ln(a) \approx 0.6740, \text{ or } a \approx 1.962.$$ 

So, our model is

$$P = 1.962e^{0.2008t}.$$ 

How well does this compare to the data? Looking at a graph with the data points and the exponential model demonstrates that the model does fit well for the last few data points, which is not surprising since the data does not appear exponential in that region.

Figure 1.4: Plot of data for the bacteria population in Table 1.6 and an exponential approximate model.

This approach really seems to magnify errors. We will investigate how to improve on this method in the future.
2 Model Development: Polynomial Models - Section 1.3 of Heinz

Lagrangean Form of Polynomials

We will start with the two easiest cases, linear polynomials and quadratic polynomials, and then generalize. The process of finding a polynomial passing through a given set of points is called interpolation.

1. Suppose the data consists of two points \((x_1, y_1)\) and \((x_2, y_2)\). Let \(P_1(x) = a_0 + a_1 x\).

Then

\[
y_i = P_1(x_i) = a_0 + a_1 x_i, \quad i = 1, 2.
\]

In addition, \(y - y_1 = m(x - x_1)\), where

\[
m = \frac{y_2 - y_1}{x_2 - x_1}.
\]

Therefore, we obtain

\[
y - y_1 = \left(\frac{y_2 - y_1}{x_2 - x_1}\right) (x - x_1),
\]

or

\[
y = \left(\frac{y_2 - y_1}{x_2 - x_1}\right) (x - x_1) + y_1.
\]

Playing with this algebraically gives the form

\[
P_1(x) = \left(\frac{x - x_2}{x_1 - x_2}\right) y_1 + \left(\frac{x - x_1}{x_2 - x_1}\right) y_2.
\]

This is the Lagrange form of the linear polynomial passing through the data points.

2. Suppose the data consists of three points: \((x_1, y_1)\), \((x_2, y_2)\), \((x_3, y_3)\). Then we may form a quadratic polynomial through the data points of the form \(P_2(x) = a_0 + a_1 x + a_2 x^2\).

So, we have

\[
y_1 = a_0 + a_1 x_1 + a_2 x_1^2
\]

\[
y_2 = a_0 + a_1 x_2 + a_2 x_2^2
\]

\[
y_3 = a_0 + a_1 x_3 + a_2 x_3^2.
\]

Solving (not fun) gives a polynomial we may write in Lagrange form as

\[
P_2(x) = \frac{(x - x_2)(x - x_3)}{(x_1 - x_2)(x_1 - x_3)} y_1 + \frac{(x - x_1)(x - x_3)}{(x_2 - x_1)(x_2 - x_3)} y_2 + \frac{(x - x_1)(x - x_2)}{(x_3 - x_1)(x_3 - x_2)} y_3.
\]

For the general case, we wish to construct a polynomial of degree at most \(n\) that passes through the \(n + 1\) points \((x_1, y_1), (x_2, y_2), \ldots, (x_{n+1}, y_{n+1})\). The idea is to express \(P_n(x)\) in the form

\[
P_n(x) = y_1 L_{n,1}(x) + y_2 L_{n,2}(x) + \cdots + y_{n+1} L_{n,n+1}(x) = \sum_{k=1}^{n+1} y_k L_{n,k}(x), \quad (2.1)
\]
where \( L_{n,k} \) satisfy
\[
L_{n,k}(x_i) = \begin{cases} 
0, & i \neq k, \\
1, & i = k.
\end{cases}
\]

How can we construct functions \( L_{n,k} \) that satisfy (2.1) for \( k = 1, 2, \ldots, n + 1 \)?

- To satisfy \( L_{n,k}(x_i) = 0 \) if \( i \neq k \), the numerator must contain
  \[
  (x - x_1)(x - x_2) \cdots (x - x_{k-1})(x - x_{k+1}) \cdots (x - x_{n+1}).
  \]

- To satisfy \( L_{n,k}(x_k) = 1 \), the denominator must equal the numerator evaluated at \( x = x_k \), i.e.,
  \[
  (x_k - x_1)(x_k - x_2) \cdots (x_k - x_{k-1})(x_k - x_{k+1}) \cdots (x_k - x_{n+1}).
  \]

Therefore,
\[
L_{n,k}(x) = \frac{(x - x_1)(x - x_2) \cdots (x - x_{k-1})(x - x_{k+1}) \cdots (x - x_{n+1})}{(x_k - x_1)(x_k - x_2) \cdots (x_k - x_{k-1})(x_k - x_{k+1}) \cdots (x_k - x_{n+1})}.
\]

The interpolating polynomial thus defined is easily described, once the form of \( L_{n,k} \) is known. The polynomial so described is known as the \( n \)th Lagrange interpolating polynomial.

**Note:** The notation \( L_{n,k} = L_k \) is used when the order of the polynomial is clear.

The following theorem guarantees that such a polynomial will always exist and be unique.

**Theorem 2.1.** If \( x_1, x_2, \ldots, x_{n+1} \) are distinct real numbers, then for arbitrary values \( y_1, y_2, \ldots, y_{n+1} \) there is a unique polynomial \( P_n(x) \) of degree at most \( n \) such that \( P_n(x_i) = y_i \) for \( i = 1, 2, \ldots, n + 1 \).

**Proof.**

- Uniqueness: Suppose that there were two such polynomials, \( P_n(x) \) and \( Q_n(x) \). Then
  \[
  (P_n - Q_n)(x_i) = 0, \ i = 1, 2, \ldots, n + 1.
  \]

  Since the degree of \( P_n - Q_n \) can be at most \( n \), it can have at most \( n \) zeros if it is not the zero polynomial. Since \( x_i, \ i = 1, 2, \ldots, n + 1 \) are distinct, \( P_n - Q_n \) has \( n + 1 \) roots, and therefore, \( P_n - Q_n = 0 \), or \( P_n = Q_n \).

- Existence: (Inductive proof) For \( n = 0 \), choose \( P_0 \) so that \( P(x_1) = y_1 \).

  Suppose we have a polynomial \( P_{k-1} \) of degree at most \( k - 1 \) such that \( P_{k-1}(x_i) = y_i \) for \( i = 1, 2, \ldots, k \). Construct the polynomial
  \[
  P_k(x) = P_{k-1}(x) + c(x - x_1)(x - x_2) \cdots (x - x_k).
  \]
Then $P_k(x)$ is a polynomial of degree at most $k$, and clearly $P_k(x)$ interpolates the data $P_k(x_i) = P_{k-1}(x_i)$ for $i = 1, 2, \ldots, k$. If we choose $c$ so that $P_k(x_{k+1}) = y_{k+1}$, i.e., so

$$P_k(x_{k+1}) = P_{k-1}(x_{k+1}) + c(x_{k+1} - x_1)(x_{k+1} - x_2) \cdots (x_{k+1} - x_k) = y_{k+1}.$$

We can solve for $c$ since $c$ is multiplied by a nonzero real number. Therefore, such a polynomial exists for any nonnegative integer $n$.

\[
\square
\]

**Problems with Polynomial Modeling**

**Erroneous Data**

Data used for the development of models may be affected by a variety of errors, such as

- noisy observations resulting from changing conditions under which measurements are made;
- errors in the tools used to measure data;
- round-off error; and
- data entry errors.

**Effect of Erroneous Data**

Polynomials that interpolate data, particularly higher order polynomials, are particularly sensitive to errors. A small error in even one data value can dramatically change the coefficients of the polynomial, and thus, its resulting shape and predictions. These errors may be magnified such that a 0.1% error in the data can cause a ten-fold error in the polynomial.

**Reduced Polynomial Models**

One way to minimize this kind of error, then, is to apply low-order polynomials, which are constructed on the basis of a reduced number of data points. Such polynomial models are called reduced polynomial models. Another advantage of using reduced order polynomials is the reduction in artificial oscillations. In the field of numerical analysis, Runge’s phenomenon is one such issue – it is the appearance of oscillations at the edges of an interval that occurs when using high-degree polynomials to interpolate data over an equispaced grid (i.e., with equally spaced grid points).

One problem with the reduced polynomial models is that they are highly dependent on the choice of data points to include. This is true for all of the examples we have done thus far.
Other Polynomial Modeling Approaches

There are some ways to avoid some of the above problems, such as

- using piecewise polynomial interpolation, such as splines;
- using least-squares fitting to fit a polynomial of lower degree to all of the data (to be discussed in the next chapter); and
- changing the spacing of the data points used to interpolate the data.

Time constraints preclude us from discussing the above.

3 Population Modeling - Section 1.4 of Heinz

We have already looked at one example of population growth, that of a population of bacteria, seen in Example 1.3. For convenience, the data is again presented below. In Example 1.3,

<table>
<thead>
<tr>
<th>$t$ (hr.)</th>
<th>7</th>
<th>14</th>
<th>21</th>
<th>28</th>
<th>35</th>
<th>42</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P$</td>
<td>8</td>
<td>41</td>
<td>133</td>
<td>250</td>
<td>280</td>
<td>297</td>
</tr>
</tbody>
</table>

we assumed that the population satisfied exponential growth of the form

$$P(t) = ae^{bt},$$

which may be written as

$$P(t) = P_0e^{r(t-t_0)},$$  \hspace{1cm} (3.1)

where $P_0$ is the population at some time $t_0$ and $r$ is the relative growth rate of the population. Equation (3.1) satisfies *Malthus' law*, which states that the rate of change in the size of a population $P$ is proportional to $P$, modeled by the differential equation

$$\frac{dP}{dt} = rP,$$

where $P(t_0) = P_0$. We saw that when we fit a line to the transformed data, the results were not very good, especially for larger time. Looking at the plot of the data again below, we see that the data seems to take more of an S-shape.
Such a shape appears to be that of a *logistic function*. A logistic function has the form

\[
y(x) = \frac{L}{1 + Ce^{-kx}}.
\]

For population modeling, the logistic function has the form

\[
P(t) = \frac{KP_0}{P_0 + (K - P_0)e^{-r(t-t_0)}},
\]

which solves the initial value problem (differential equation with initial condition)

\[
\frac{dP}{dt} = rP \left(1 - \frac{P}{K}\right),\quad P(t_0) = P_0.
\]  

In the textbook, a different form of the logistic model is used, and the given form lends itself more easily to modeling by a linear function after a transformation of the data. However, if we digress from the topic of this section, let \(t_0 = 0\) and do some algebraic manipulation, we may rewrite (3.2) as

\[
P(t) = \frac{P_0e^{rt}}{1 + \frac{P_0}{K}(e^{rt} - 1)}.
\]

Assuming that the population is known at three times, \(t = 0, t = T,\) and \(t = 2T,\) being \(P_0, P_1,\) and \(P_2,\) respectively, it is possible (after some not so fun calculations) to show that

\[
K = P_1\frac{P_0P_1 + P_1P_2 + 2P_0P_2}{P_1^2 - P_0P_2} \quad \text{and} \quad r = \frac{1}{T}\log \left(\frac{\frac{1}{P_0} - \frac{1}{K}}{\frac{1}{P_1} - \frac{1}{K}}\right).
\]

Note that if we start with (3.2) and assume that \(P_1 = P(t_0 + T)\) and \(P_2 = P(t_0 + 2T),\) we obtain the same estimates for \(r\) and \(K.\) So, using the data for the bacteria, we have that \(P(7) = 8, P(7 + 7) = P(14) = 41,\) and \(P(7 + 2(7)) = P(21) = 133,\) we obtain the estimates

\[
r \approx 0.255 \quad \text{and} \quad K \approx 242.7.
\]
In Figure 3.1, we see that the logistic curve has a similar shape as the data points, but again the error is larger for the latter three data points. It should be noted that the error is much smaller than for the exponential model obtained in Example 1.3.

Exercise: What happens if we choose different points, e.g., set $T = 14$, instead of 7? Does this improve things? Disimprove them?

**Polynomial Models**

We again consider polynomial models of the population. How do we expect these to perform, at least, say, for the first part of the data (the exponential-appearing part)? We can actually think about this analytically by considering the Taylor series for $e^{rt}$ centered at $t_0 = 0$,

$$e^{rt} = 1 + rt + \frac{1}{2}(rt)^2 + \cdots.$$ 

(3.5)

The error in an $n$th order Taylor polynomial approximation of (3.5) is given by $R_n(t)$, where

$$R_n(t) = \frac{1}{(n+1)!} e^{rci(rt)^{n+1}},$$

where $c$ is between 0 and $t$. If $t$ is sufficiently small, then $R_n(t)$ will be sufficiently small to make the $n$th order Taylor polynomial a good approximation of $e^{rt}$. 

11
4 Determining the Best Polynomial for the Job

As discussed previously, using reduced polynomial models can avoid the oscillatory errors inherent in using higher-order polynomials to interpolate the data points. This begs the question of how to decide what order polynomial to use. There are two methods we may use to qualitatively determine whether or not we can try a lower order polynomial and, if so, what order.

Look at a Plot of the Data Points for a Pattern

Looking at a plot of the data points can give a good indication what degree polynomial seems to be followed by the data; e.g., if the data appears to follow a parabolic path, then a quadratic polynomial is the best choice.

Use Divided Difference Tables

Suppose that our data satisfies exactly the parabola

\[ y = ax^2 + bx + c. \]

Then

\[ \frac{dy}{dx} = 2ax + b, \quad \frac{d^2y}{dx^2} = 2a, \quad \text{and} \quad \frac{d^3y}{dx^3} = 0. \]

In general for a polynomial of degree \( n \), it is true that

\[ \frac{d^{n+1}y}{dx^{n+1}} = 0. \]

And, since

\[ \frac{dy}{dx} = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x}, \]

taking functional differences of successive data values will give us at least some insight into how the derivatives behave. The goal is to find successive differences that will approximate 0. We define \( \Delta^k \) to represent these differences at different levels, i.e.,

\[ \Delta^k \approx \frac{d^k y}{dx^k}. \]

Example 4.1. The distance for a car to stop depends on the velocity \( v \) of the car. Table 4.1 gives some stopping distances \( D \), measure in feet, for various velocities, measured in miles per hour. The divided differences for this data are given below.

\(^1\text{Much of this material comes from Mathematical Modeling with Maple by William P. Fox, Brooks-Cole (2012).}\)
<table>
<thead>
<tr>
<th>$v$</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>50</th>
<th>60</th>
<th>70</th>
<th>80</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D$</td>
<td>42</td>
<td>73.5</td>
<td>116</td>
<td>173</td>
<td>248</td>
<td>343</td>
<td>464</td>
</tr>
</tbody>
</table>

Table 4.1: Stopping distance $D$ (in feet) for different velocities $v$ (in mph).

\[
\begin{array}{cccccc}
20 & 42 & 3.15 \\
30 & 73.5 & 0.055 & 4.25 & 0.0005833 \\
40 & 116 & 0.0725 & 5.7 & 0.0005833 & -1.25 \times 10^{-7} \\
50 & 173 & 0.09 & -6.25 \times 10^{-7} & 4.58 \times 10^{-7} & 9.722 \times 10^{-9} \\
60 & 248 & 0.10 & 1.67 \times 10^{-5} \\
70 & 343 & 0.13 \\
80 & 464 & 12.1 \\
\end{array}
\]

Since the values for $\Delta^4$ are of different signs, this approach does not work well, since the following $\Delta^k$ values are suspect. However, since the $\Delta^3$ values are sufficiently small, we can assume that they are close enough to 0 to warrant using a quadratic polynomial, which is, in fact, what the course textbook by Heinz does.