Calculus for the Life Sciences

Lecture Notes – Least Squares Analysis

Joseph M. Mahaffy,
⟨jmahaffy@mail.sdsu.edu⟩

Department of Mathematics and Statistics
Dynamical Systems Group
Computational Sciences Research Center
San Diego State University
San Diego, CA 92182-7720

http://www-rohan.sdsu.edu/∼jmahaffy

Spring 2017

Outline

Linear Least Squares Best Fit

- Linear Models section showed cricket data appear to lie on a line
- **Linear least squares best fits** a linear model to data
- **Linear regression** is another common name for this analysis
  - The term regression comes from a pioneer in the field of applied statistics who gave the least squares line this name because his studies indicated that the stature of sons of tall parents reverts or regresses toward the mean stature of the population

Cell Division in *E. coli*

Figures for Cell Cycle for *E. coli*

- Genome is a single large loop of DNA (3,800,000 base pairs)
- Replicates in both directions, starting at *oriC*
- Bacteria (prokaryotes) cell cycle differs from eukaryotic organisms – replication cycles overlap for rapid growth

Cell Cycle in *E. coli*

Replication of DNA in *E. coli*

- *Escherichia coli* can divide every 20 minutes
- Time for the DNA to replicate is the C period
- Time for the two loops of DNA to split apart, segregate, and form two new daughter cells is the D period
- The C period is 35-50 min, and the D period is over 25 min
- Replication cycle often longer than cell division time
- Up to 8 *oriCs* in a single *E. coli*

Pulse Labeling Experiment

Finding the C Period

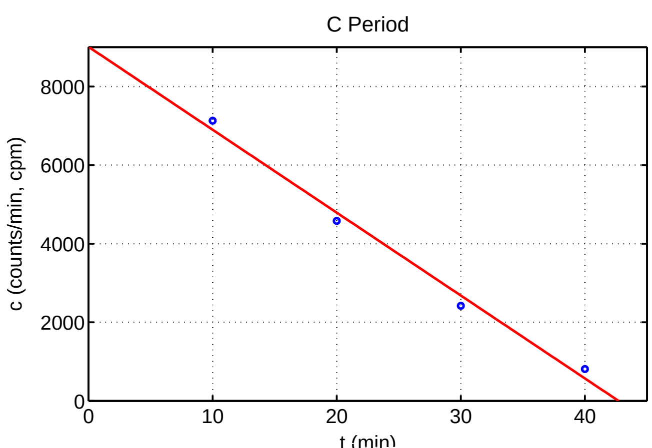
- A pulse of radioactive thymidine given to *E. coli*
- Drugs at $t = 0$ to stop new replication forks and division
- Radioactive thymidine added to existing forks
- As forks end, no new radioactive thymidine added
- Radioactive emissions, c in counts/min (cpm) measured in lab of Prof. Judith Zyskind (SDSU)

t (min)	10	20	30	40
c (cpm)	7130	4580	2420	810

Linear Model for C Period – Graph

Best fitting Linear Model for C Period

$$c = -211.2\,t + 9015$$



Fitting the Data

Linear Model

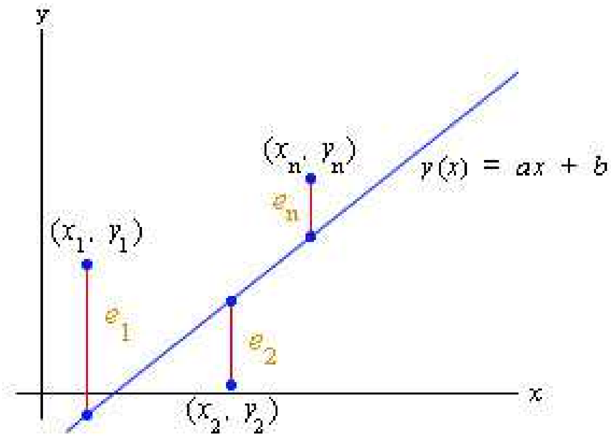
$$c = at + b$$

- Actual model uses techniques of Integral Calculus
- Linear model a reasonable approximation
- The t-intercept approximates the C period
- **Least squares best fit** minimizes sum of c-distance from data to linear model
- Minimizes distance by adjusting slope, a, and intercept, b

Data suggest that **C period** is 42.7 min

Least Squares Best Fit

The **least squares best fit** of a line to data is the best line through a set of data



Fitting the Data

- Consider a set of n data points:

$$(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$$

- Select a slope, a, and an intercept, b, that results in a line that in some sense best fits the data

$$y(x) = ax + b$$

- The least squares best fit minimizes the square of the error in the distance between the y_i values of the data points and the y value of the line
- Distance depends on selection of the slope, a, and the intercept, b

Error between Line and Data

- Error between each of the data points and the line is

$$e_i = y_i - y(x_i) = y_i - (ax_i + b), \quad i = 1, ...n$$

- Define the **Absolute Error** between each of the data points and the line as

$$|e_i| = |y_i - y(x_i)| = |y_i - (ax_i + b)|, \quad i = 1, ...n$$

- The error e_i varies as a and b vary

Sum of Square Errors

The error between each data point and the line is

$$e_i = y_i - (ax_i + b), \quad i = 1, ...n$$

Create a function depending on the slope a and intercept b of the line, which sums the square errors

$$J(a, b) = e_1^2 + e_2^2 + ... + e_n^2 = \sum_{i=1}^{n} e_i^2$$

The **Least Squares Best Fit Line** is the minimum value of the function $J(a, b)$

Minimum is determined using Calculus of two variables

Formula for Best Fitting Line

Assume data points $(x_i, y_i), i = 1, ..., n$, and line

$$y = ax + b$$

Define the mean of the x values

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

The best fitting slope satisfies

$$a = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

The best fitting intercept satisfies

$$b = \frac{1}{n}\sum_{i=1}^{n} y_i - a\bar{x} = \bar{y} - a\bar{x}$$

C Period Example (continued)

The pulse labeling experiment for *E. coli* gave data points:

$$(10, 7130), (20, 4580), (30, 2420), (40, 810)$$

The mean time is

$$\bar{t} = \frac{10 + 20 + 30 + 40}{4} = 25$$

The best slope, a, satisfies

$$a = \frac{(10-25)7130 + (20-25)4580 + (30-25)2420 + (40-25)810}{(10-25)^2 + (20-25)^2 + (30-25)^2 + (40-25)^2}$$

$$a = -211.2$$

C Period Example (continued)

Similarly, the c-intercept, b, satisfies:

$$\begin{aligned} b &= \frac{7130 + 4580 + 2420 + 810}{4} - (-211.2)25 \\ b &= 9015 \end{aligned}$$

Thus, the best fitting line is given by

$$c(t) = -211.2t + 9015$$

C Period Example - Error

With $c(t) = -211.2t + 9015$, compute the errors

For the first datum point $(t, c) = (10, 7130)$, the model predicts $c(10) = 6900$, so

$$e_1 = c_1 - c(10) = 7130 - 6903 = 227$$

$$\begin{aligned} e_2 &= c_2 - c(20) = 4580 - 4791 = -211 \\ e_3 &= c_3 - c(30) = 2420 - 2679 = -259 \\ e_4 &= c_4 - c(40) = 810 - 567 = 243 \end{aligned}$$

The sum of the square of these errors is

$$J(-211.2, 9015) = 51529 + 44521 + 67081 + 59049 = 222180$$

Juvenile Growth Model - Revisited

The linear Models section showed that Juvenile Height was approximated well with a linear model

Linear model is given by:

$$h(a) = 6.46\,a + 72.3$$

and fit the data well

Least sum of square errors is found to be

$$J(m, b) = 41.5$$

Applet for Juvenile Height Growth

Example 1 - Model Choice 1

Two researchers had only a limited set of data, the points (2,2), (5,6), and (8,3).

Researcher A felt that the model given by with y increasing with increasing x

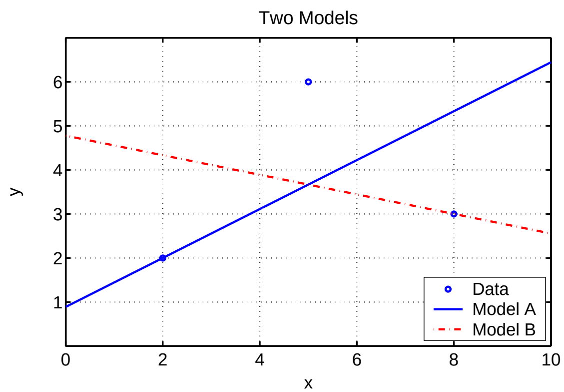
$$y = \frac{5}{9}x + \frac{8}{9}$$

Researcher B felt that the model given by with y decreasing with increasing x

$$y = -\frac{2}{9}x + \frac{43}{9}$$

Example 1 2

Graph of data and two models:



- Find the sum of square errors for each model
- Which one is better accoding to the data

Example 1 3

Solution: Recall the error for line $y = ax + b$ satisfies:

$$e_i = y_i - (ax_i + b)$$

For Model A,

$$J_A = e_1^2 + e_2^2 + e_3^2$$

$$J_A = \left(2 - \left(\tfrac{5}{9}(2) + \tfrac{8}{9}\right)\right)^2 + \left(6 - \left(\tfrac{5}{9}(5) + \tfrac{8}{9}\right)\right)^2 + \left(3 - \left(\tfrac{5}{9}(8) + \tfrac{8}{9}\right)\right)^2$$

$$J_A = 10.89$$

Example 1

For Model B,

$$J_B = e_1^2 + e_2^2 + e_3^2$$

$$J_B = \left(2 - \left(-\tfrac{2}{9}(2) + \tfrac{43}{9}\right)\right)^2 + \left(6 - \left(-\tfrac{2}{9}(5) + \tfrac{43}{9}\right)\right)^2 + \left(3 - \left(-\tfrac{2}{9}(8) + \tfrac{43}{9}\right)\right)^2$$

$$J_B = 10.89$$

Since $J_A = J_B$, the two models are equally valid

Example 1

What is the best fitting Linear model for these data?

Solution: The average x is

$$\bar{x} = \frac{2+5+8}{3} = 5$$

Best slope a satisfies:

$$a = \frac{(2-5)2+(5-5)6+(8-5)3}{(2-5)^2+(5-5)^2+(8-5)^2} = \frac{1}{6}$$

Since $\bar{y} = \frac{11}{3}$, the intercept b is

$$b = \bar{y} - a\bar{x} = \frac{11}{3} - \frac{5}{6} = \frac{17}{6}$$

The best linear model is

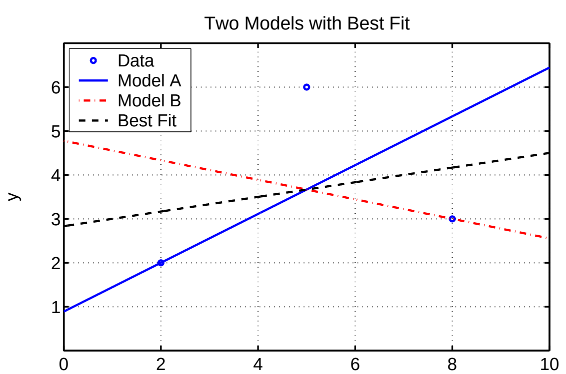
$$y = \frac{1}{6}x + \frac{17}{6}$$

Example 1

The sum of square error between the data and the best fitting model is **8.17**, which is better than other models (**10.89**)

There are clearly too few data to really produce a model

Graph of data and three models:



Actual and Absolute Error

- Error analysis is important for testing validity of a model
- Let X_e be an experimental measurement or the worst value from a model being tested
- Let X_t be a theoretical value or the best value from actual data
- The **Actual Error** is

$$\textbf{Actual Error} = X_e - X_t$$

- The **Absolute Error** is appropriate when only the magnitude of the error is needed

$$\textbf{Absolute Error} = |X_e - X_t|$$

Relative and Percent Error

- Relative and Percent error allow a better comparison of the error between data sets or within a data set with large differences in the numerical values
- Again let X_e be an experimental measurement or the worst value from a model being tested and X_t be a theoretical value or the best value from actual data
- The **Relative Error** is
$$\textbf{Relative Error} = \frac{X_e - X_t}{X_t}$$
- The **Percent Error** is the most common and divides the **Relative error** by the best expected value
$$\textbf{Percent Error} = \frac{X_e - X_t}{X_t} \times 100\%$$

Growth Model 1

Consider the growth of a fish given by the data:

t (weeks)	0	1	2	3	5	7	9
L (cm)	2.4	3.1	3.7	4.1	5.2	4.9	6.9

The formula for finding the least squares best fit linear model gives:
$$L = 0.437\,t + 2.644$$

Determine the growth rate for this model

Solution: The rate of growth is the slope of the best fitting line, so

Growth Rate = 0.437 cm/week

Growth Model 2

Find the sum of square errors

Solution: Each of the square errors is:
$$\begin{aligned}
e_1^2 &= (2.4 - 2.644)^2 = 0.0595\\
e_2^2 &= [3.1 - (0.437 + 2.644)]^2 = 0.0004\\
e_3^2 &= [3.7 - (0.874 + 2.644)]^2 = 0.0331\\
e_4^2 &= [4.1 - (1.311 + 2.644)]^2 = 0.0210\\
e_5^2 &= [5.2 - (2.185 + 2.644)]^2 = 0.1376\\
e_6^2 &= [4.9 - (3.059 + 2.644)]^2 = 0.6448\\
e_7^2 &= [6.9 - (3.933 + 2.644)]^2 = 0.1043
\end{aligned}$$

Sum of Square Errors is
$$J(0.437, 2.644) = 1.0008$$

Growth Model 3

- Some data sets have points that are erroneous due to problems with the experiment (say contamination) or simply a poorly recorded value
- Statistical tests exist to determine if point can be removed
- Hypothesis testing studied in Bio 215
- If these points are included in the model, then they can result in misleading models

Growth Model

Which point is most likely erroneous?

The point with the most error is (7, 4.9)

When this point is removed, the new least squares best fit model is

$$L = 0.492t + 2.594$$

Determine the growth rate for this model

Growth Rate = 0.492 cm/week

What is the new sum of square errors

Solution: The new sum of square errors is

$$J(a, b) = 0.0376 + 0.0002 + 0.0149 + 0.0009 + 0.0213 + 0.0149 = 0.0898$$

which is only 9% of the sum of squares error from above

Growth Model

What is the percent error between the computed growth rates?

Solution: The growth rate without the erroneous point is the best value, so

$$X_t = 0.492$$

The original growth rate is the worst value, so

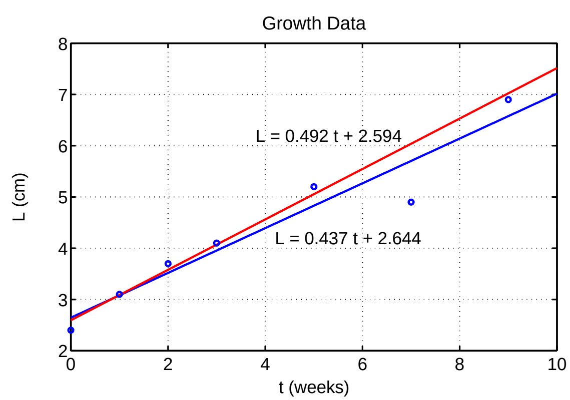
$$X_e = 0.437$$

Percent error is

$$\left(\frac{0.437 - 0.492}{0.492}\right) \times 100 = -11.2\%$$

Growth Model

Graph of data and two models:



Graph readily shows linear data and erroneous point