Case Study: Least-Squares Solutions

In this case study, large systems of equations which have no solution are considered. The problem of readjusting the North American Datum is one such large system of equations; unfortunately, it is far too large a system to deal with, so instead the general method for solving these systems will be discussed: the method of least squares.

The **method of least squares** is used when there is a system of equations $A\mathbf{x} = \mathbf{b}$ which has no solution. In such a case, a "solution" will be sought $\hat{\mathbf{x}}$ which makes the difference between $A\hat{\mathbf{x}}$ and $\mathbf{b}$ as small as possible. This "solution" is called a **least-squares solution** to the system $A\mathbf{x} = \mathbf{b}$. As is shown in Section 6.5, the least-squares solution $\hat{\mathbf{x}}$ must satisfy the **normal equations**

$$A^TA\hat{\mathbf{x}} = A^T\mathbf{b}$$

Furthermore, if the columns of A are linearly independent, then A^TA is invertible, and there is a unique least-squares solution

$$\hat{\mathbf{x}} = (A^TA)^{-1}A^T\mathbf{b}$$

As described in Section 6.6, one way in which least-squares solutions are useful is in the fitting of a curve to a set of data. First rewrite the system as $X\boldsymbol{\beta} = \mathbf{y}$, where X is called the **design matrix**, $\boldsymbol{\beta}$ is called the **parameter vector**, and $\mathbf{y}$ is called the **observation vector**. Then the least-squares solution to this system is

$$\hat{\boldsymbol{\beta}} = (X^TX)^{-1}X^T\mathbf{y}$$

A least-squares solution may be used to see a trend in data and to use that trend to make predictions.

Example 1: The following table lists the gold medalist in the men's 400 meter run for each of last 7 Olympics along with the time of the race in seconds:

Year	Name	Nation	Time
1972	Vincent Matthews	United States	44.66
1976	Alberto Juantorena	Cuba	44.26
1980	Viktor Martin	USSR	44.60
1984	Alonzo Babers	United States	44.27
1988	Steven Lewis	United States	43.87
1992	Quincy Watts	United States	43.50
1996	Michael Johnson	United States	43.49
2000	Michael Johnson	United States	43.84

One might to attempt to predict the results of this race at the 2004 Olympics by fitting an appropriate curve through the data points above. For ease in calculation, let the t-coordinate be the number of years after 1972. See Figure 1.

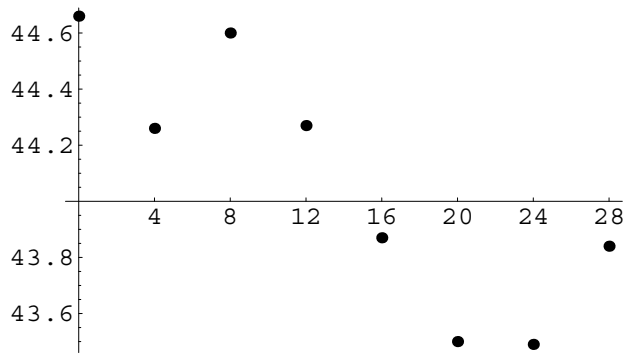

Figure 1: Olympic men's 400 meters (1972-2000)

First fit a line to these points. As shown on page 420, the design matrix X and the observation vector $\mathbf{y}$ for this system are

$$X = \begin{bmatrix} 1 & 0 \\ 1 & 4 \\ 1 & 8 \\ 1 & 12 \\ 1 & 16 \\ 1 & 20 \\ 1 & 24 \\ 1 & 28 \end{bmatrix} \quad \text{and} \quad \mathbf{y} = \begin{bmatrix} 44.66 \\ 44.26 \\ 44.60 \\ 44.27 \\ 43.87 \\ 43.50 \\ 43.49 \\ 43.84 \end{bmatrix}$$

Since the least-squares solution to this system is $\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \mathbf{y} = \begin{bmatrix} 44.615 \\ -.0395536 \end{bmatrix}$, the line $y = 44.615 - .0395536t$ is the line which best fits these points. See Figure 2. This model could be used to attempt to predict the winning time at the 2004 Olympics by letting $x = 32$. In this case $y = 44.615 - .0395536(32) = 43.3493$, so this model predicts 43.3493 seconds as the winning time in 2004.

However, it may be more appropriate to fit a curve of a different form to the data. For example the quadratic curve $y = \beta_0 + \beta_1 t + \beta_2 t^2$ may be fit. In this case the observation vector $\mathbf{y}$ is the same as above. However, the design matrix must change. As shown in Example 2 on page 422, the design matrix X has become

$$X = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 4 & 16 \\ 1 & 8 & 64 \\ 1 & 12 & 144 \\ 1 & 16 & 256 \\ 1 & 20 & 400 \\ 1 & 24 & 576 \end{bmatrix}$$

Since the least-squares solution to this system is $\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \mathbf{y} = \begin{bmatrix} 44.7087 \\ -.0629911 \\ -.000837054 \end{bmatrix}$, the curve $y = 44.7087 - .0629911t - .000837954t^2$ is the curve of that form which best fits these points. See Figure 2. This model predicts that the winning time in 2004 will be $44.7087 - .0629911(32) - .000837054(1024) = 43.5502$ seconds.

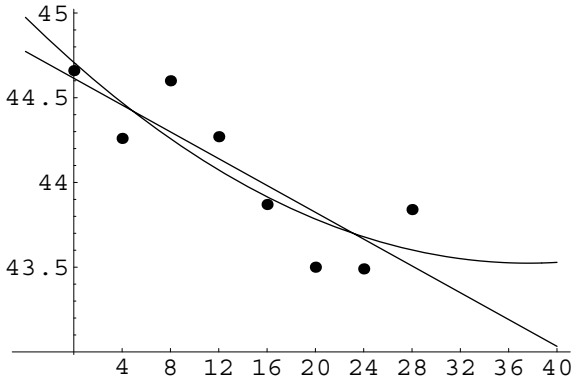

Figure 2: Olympic men's 400 meters (1972-2000) – Linear and Quadratic Curves

Example 2: The average high and low temperatures in Charlotte, North Carolina for particular days of the year are listed in the following table. For ease in calculation, the days of the year are numbered beginning with January 15, which has the lowest average high and low temperatures in the year.

Date	Day Number	High Temperature	Low Temperature
January 15	0	50	31
March 15	59	61	39
May 15	120	79	57
July 15	181	88	69
September 15	243	82	63
November 15	304	62	40

It makes sense that this data should be modelled by a periodic function, so using $\sin t$ and $\cos t$ seems reasonable. However, functions are needed whose period is 365 days, so the following curve might be tried:

$$y = \beta_0 + \beta_1 \cos\left(\frac{2\pi t}{365}\right) + \beta_2 \sin\left(\frac{2\pi t}{365}\right)$$

Fitting this curve to the data yields the design matrix

$$X = \begin{bmatrix} 1 & 1 & 0 \\ 1 & \cos\left(\frac{118\pi}{365}\right) & \sin\left(\frac{118\pi}{365}\right) \\ 1 & \cos\left(\frac{240\pi}{365}\right) & \sin\left(\frac{240\pi}{365}\right) \\ 1 & \cos\left(\frac{362\pi}{365}\right) & \sin\left(\frac{362\pi}{365}\right) \\ 1 & \cos\left(\frac{486\pi}{365}\right) & \sin\left(\frac{486\pi}{365}\right) \\ 1 & \cos\left(\frac{608\pi}{365}\right) & \sin\left(\frac{608\pi}{365}\right) \end{bmatrix}$$

For the high temperatures, use the observation vector $\mathbf{y}_H$; for low temperatures, use the observation vector $\mathbf{y}_L$:

$$\mathbf{y}_H = \begin{bmatrix} 50 \\ 61 \\ 79 \\ 88 \\ 82 \\ 62 \end{bmatrix}, \mathbf{y}_L = \begin{bmatrix} 31 \\ 39 \\ 57 \\ 69 \\ 63 \\ 40 \end{bmatrix}$$

For high temperatures, the least-squares solution to the system is $\hat{\boldsymbol{\beta}}_H = (X^T X)^{-1} X^T \mathbf{y} = \begin{bmatrix} 70.48 \\ -18.99 \\ -0.85 \end{bmatrix}$. The curve $y = 70.48 - 18.99 \cos\left(\frac{2\pi t}{365}\right) - 0.85 \sin\left(\frac{2\pi t}{365}\right)$ is the curve of that form which best fits these points. For low temperatures, the least-squares solution to the system is $\hat{\boldsymbol{\beta}}_L = (X^T X)^{-1} X^T \mathbf{y} = \begin{bmatrix} 49.99 \\ -19.51 \\ -1.696 \end{bmatrix}$. The curve $y = 49.99 - 19.51 \cos\left(\frac{2\pi t}{365}\right) - 1.696 \sin\left(\frac{2\pi t}{365}\right)$ is the curve of that form which best fits these points. See Figure 2.

These curves can be used to approximate the average high and low temperatures on a given day. For example, on April 26 ($t = 100$) the curves give the values of $72.49°$ for a high temperature and $51.24°$ for a low temperature. These values agree fairly well with the real average high and low temperatures, which are respectively $75°$ and $52°$.

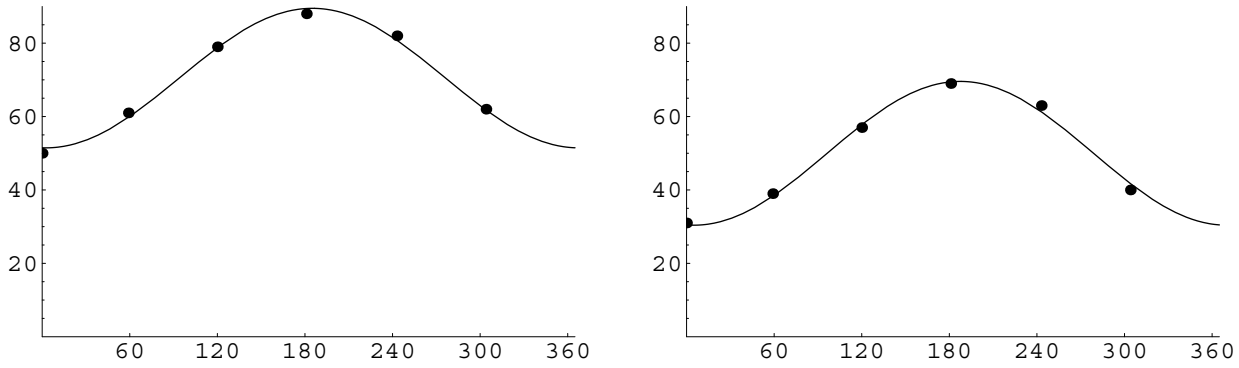

Figure 3: Average High and Low Temperatures in Charlotte NC – trigonometric curves

Questions:

1. Consider the set of points $(1, 2)$, $(2, 3)$, $(3, 5)$, $(4, 4)$, and $(5, 1)$.

   a) Find the line which best fits these points.

   b) Find the quadratic curve which best fits these points.

   c) Which curve, the line or the quadratic, do you feel gives a better approximation to the data? Why?

2. In the Olympic example above, two different curves were fit to the same data. The quadratic curve may not be a good choice, however. In Figure 2, it is seen that the values of this curve will eventually increase after about $x = 38$ (the year 2010). It does not seem reasonable to expect that the times will begin to increase steadily after 2010. Perhaps a curve which is slowly decreasing would be a better choice. Model the Olympic data in Example 2 using the following curves, and graph your curve along with the data points. Then predict the time for the 2004 Olympics and see which model you believe models the data best.

   a) $y = \beta_0 + \beta_1 \left(\frac{1}{t+10}\right)$

   b) $y = \beta_0 + \beta_1 \ln(t + 10)$

3. There was a low tide at Cape Hatteras Pier, North Carolina at 12:00 a.m. on January 1, 1999. The water levels (in feet above mean sea level) over the next twelve hours were as follows:

Time	Level
1:00 a.m.	-1.60
3:00 a.m.	0.47
5:00 a.m.	2.25
7:00 a.m.	2.47
9:00 a.m.	1.08
11:00 a.m.	-1.21

   Assuming a twelve hour space between low tides, fit an appropriate curve to this data, and use this curve to approximate how far above mean sea level the high tide was which occurred during this time interval.

4. Use Weather Channel documentaion to do your own climate study of your town. Choose dates, use model you think would be good, analyze it.

References:

1. National Ocean Service website: `www.opsd.nos.noaa.gov`

2. *The World Almanac and Book of Facts, 2001*. Mahwah: World Almanac Books, 2000.

3. Udelson, Steve. WSOC-TV Meteorologist. Private communication, 1998.