Math 2802 N1-N3 Worksheet 9

Solutions

1. Let $A=\begin{pmatrix}1&0\\0&1\\1&0\end{pmatrix}$ and $b=\begin{pmatrix}1\\2\\-3\end{pmatrix}$. Find the least-squares solution to $Ax=b$

a) using the normal equations,

b) using that the columns of A are orthogonal. Compare the two methods

Solution.

a) We have to compute

$$A^TA=\begin{pmatrix}1&0&1\\0&1&0\end{pmatrix}\begin{pmatrix}1&0\\0&1\\0&1\end{pmatrix}=\begin{pmatrix}2&0\\0&1\end{pmatrix}$$

$$A^Tb=\begin{pmatrix}1&0&1\\0&1&0\end{pmatrix}\begin{pmatrix}1\\2\\-3\end{pmatrix}=\begin{pmatrix}-2\\2\end{pmatrix}$$

Now the least-square solution is $\widehat{x}=\begin{pmatrix}-1\\2\end{pmatrix}$ because it solves

$$\begin{pmatrix}2&0\\0&1\end{pmatrix}\begin{pmatrix}\widehat{x}_1\\\widehat{x}_2\end{pmatrix}=\begin{pmatrix}-2\\2\end{pmatrix}.$$

Entries in $\widehat{x}$ are the coordinates with respect to the orthogonal vectors u_1,u_2 below.

b) Since Ax gives a linear combination of vectors $u_1=\begin{pmatrix}1\\0\\1\end{pmatrix}$ and $u_2=\begin{pmatrix}0\\1\\0\end{pmatrix}$; we are looking for the linear combination $\widehat{x}_1u_1+\widehat{x}_2u_2$ that is the closest to b. By definition the resulting combination is the projection of b onto the plane spanned by u_1,u_2, say W.

Since u_1 and u_2 are orthogonal, we have a ready-to-use formula for

$$proj_W(b)=\frac{b\cdot u_1}{u_1\cdot u_1}u_1+\frac{b\cdot u_2}{u_2\cdot u_2}u_2=-u_1+2u_2$$

We had computed all those dot products already! Note that

$$A^TA=\begin{pmatrix}u_1\cdot u_1&u_2\cdot u_1\\u_1\cdot u_2&u_2\cdot u_2\end{pmatrix}\qquad\text{and}\qquad A^Tb=\begin{pmatrix}u_1\cdot b\\u_2\cdot b\end{pmatrix}.$$

2. The table below gives the crude male death rate for lung cancer in 1950 and the per capita consumption of cigarettes in 1930 in various countries.

Country	Cigarrette Consumption (per capita)	Lung cancer deaths (per million males)
Country A	270	96
Country B	300	124
Country C	360	166
Country D	460	171

a) Obtain the straight line $y = \beta_1 x + \beta_0$ that best fits the data, having lung cancer rates as a function of cigarrette consumption.

b) According to the terminology of statistical analysis: give the design matrix, the observation vector and the unknown parameter vector.

c) In 1930, the per capita cigarette consumption in Country E was 1200. Estimate (extrapolate) the male lung cancer rate in Country E in 1950.

Solution.

a) We have to solve a least squares problem for $Ax = b$ with A,x and b as below.

Since the columns of A are not orthogonal we have no shortcut to find the entries of x. We have to use the normal equations $A^T A\widehat{x} = A^T b$.

We have $A^T b = \begin{pmatrix} 557 \\ 201540 \end{pmatrix}$ and $A^T A = \begin{pmatrix} 4 & 1390 \\ 1390 & 504100 \end{pmatrix}$. We can check that $A^T A$ is invertible; this implies that

$$\widehat{x} = (A^T A)^{-1} A^T b = \begin{pmatrix} 4 & 1390 \\ 1390 & 504100 \end{pmatrix}^{-1} \begin{pmatrix} 557 \\ 201540 \end{pmatrix} \approx \begin{pmatrix} 7.6 \\ -.38 \end{pmatrix}$$

Therefore, the best-fit line is given by $y = -.38x + 7.6$.

b) Observation and parameter vectors are simply $b = \begin{pmatrix} 96 \\ 124 \\ 166 \\ 171 \end{pmatrix}$ and $x = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$, respectively. The design matrix: $A = \begin{pmatrix} 1 & 270 \\ 1 & 300 \\ 1 & 360 \\ 1 & 460 \end{pmatrix}$

c) The line is $y = -.38x + 7.6$. For Country E we have that $x = 1200$ and thus the predicted value for the male lung cancer rate is $y = 463.6$.

3. Compute the error associated to the least-squares best fit line in Problem 2. Sketch a picture of the data and the best fit line that shows

'*the best-fit line minimizes the sum of the squares of the residuals*'

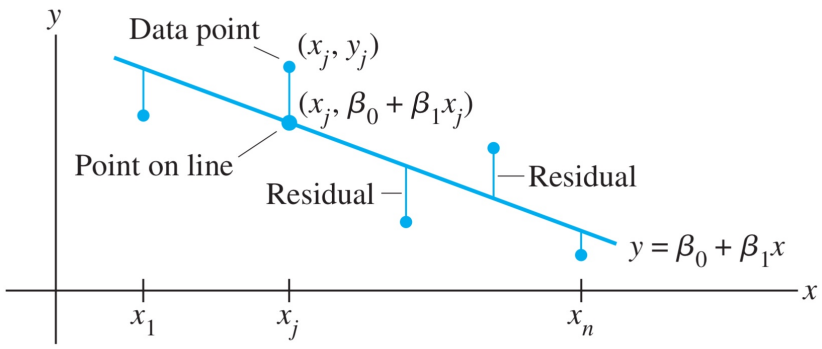
Solution.

Let $\widehat{b} = Ax$ where x is the least-solution we found. The vector $\widehat{b}$ represents the prediction of lung cancer rate (this is how we solved problem *2c)*) we have to compare it to vector b, the actual observations that have been made in countries *A-D*.

The error associated is the distance between $\widehat{b}$ and b; equivalently

$$||\widehat{b} - b|| = || \begin{pmatrix} 110.2 \\ 121.6 \\ 144.4 \\ 182.4 \end{pmatrix} - \begin{pmatrix} 96 \\ 124 \\ 166 \\ 171 \end{pmatrix} || = || \begin{pmatrix} 14.2 \\ -2.4 \\ -21.6 \\ 11.4 \end{pmatrix} || = \sqrt{803.92} \approx 28.35$$

This is a square-root of the sum of the distances between the prediction and the observation of the lung cancer rates at each country. For a representative picture:



4. The table below gives the prices per bar of chocolate according to the number of boxes of chocolate bar produced by a company. Each box contains 20 chocolate bars.

Boxes produced	1	2	3	4	5
Cost per bar	1.8	2.7	3.4	3.8	3.9

a) Obtain the parabola $y = \beta_2 x^2 + \beta_1 x$ that best fits the cost of the chocolate bar as a function of the production.

b) According to the terminology of statistical analysis: give the design matrix, the observation vector and the unknown parameter vector.

c) Estimate (extrapolate) the cost of each chocolate bar if the company would produce 10 boxes of chocolate bars.

Solution.

a) We have to solve a least squares problem for $Ax = b$ with A,x and b as below.

Since the columns of A are not orthogonal we have no shortcut to find the entries of x. We have to use the normal equations $A^T A\widehat{x} = A^T b$.

We have $A^T b = \begin{pmatrix} 52.1 \\ 201.5 \end{pmatrix}$ and $A^T A = \begin{pmatrix} 55 & 225 \\ 225 & 979 \end{pmatrix}$. We can check that $A^T A$ is invertible; this implies that

$$\widehat{x} = (A^T A)^{-1} A^T b = \begin{pmatrix} 55 & 225 \\ 225 & 979 \end{pmatrix}^{-1} \begin{pmatrix} 52.1 \\ 201.5 \end{pmatrix} \approx \begin{pmatrix} 1.76 \\ -.20 \end{pmatrix}$$

Therefore, the best-fit parabola is given by $y = 1.76x - .2x^2$.

b) Observation and parameter vectors are simply $b = \begin{pmatrix} 1.8 \\ 2.7 \\ 3.4 \\ 3.8 \\ 3.9 \end{pmatrix}$ and $x = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$, respectively. The design matrix: $A = \begin{pmatrix} 1 & 1 \\ 2 & 4 \\ 3 & 9 \\ 4 & 16 \\ 5 & 25 \end{pmatrix}$

c) The parabola is $y = 1.76x - .2x^2$. For a production of 10 boxes the predicted cost per bar is $y = 17.6 - 20 = -2.4$.

Note: This makes no sense for the applications model, the issue can be spotted from the negative value of β_2, the extrapolation would eventually give negative numbers for the cost of a product.