Biometrical Letters
Vol. 47 (2010), No. 2, 155-159

A simple algebraic derivation of normal equations

Czesław Stępniak

Institute of Mathematics, University of Rzeszów, Rejtana 16 A, 35-959 Rzeszów, Poland, e-mail: cees@univ.rzeszow.pl

Summary

Normal equations have a nice geometric interpretation but their formal derivation requires some mathematical knowledge, such as differential calculus or generalized inverse. We present a simple algebraic method based on the equivalence of certain systems of linear equations. This equivalence may be interesting in itself.

Key words: system of linear equations, equivalent systems, least squares solution, normal equations.

1. Introduction

Normal equations take their origin from Gauss (1857) and Legendre (1806). They refer to the standard Gauss-Markov model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$,where $\mathbf{y}$ is the column of observations, $\boldsymbol{\beta}$ is the column of unknown parameters, $\mathbf{X}$ is a known matrix, and $\mathbf{e}$ is an unobservable random column with zero expectation and the identity dispersion matrix. The *Least Squares* principle is to find $\boldsymbol{\beta}_0$, which minimizes the residual $\| \mathbf{X}\boldsymbol{\beta} - \mathbf{y} \|^2$ over all $\boldsymbol{\beta}$. For the history and the possible approaches to the problem we refer to Herr (1980) and deLaubenfels (2006).

The literature on the subject reveals three different approaches based respectively on:

- geometry (cf. Fisher (1915), Kruskal (1961, 1968)),
- differential calculus (cf. Scheffé (1959), Rao (1973)),
- generalized inverse matrices (cf. Rao and Mitra (1971), Bapat (2000)).

As in Rao (1973, pp. 222-223) the necessary condition for the extremum, obtained by setting the first derivatives equal to zero, needs to be

supplemented by further (algebraic or analytic) consideration. In this situation a complete formal algebraic way is still being sought. The first step in this direction was taken by Hamilton (1933).

In this work we express the well-known geometric ideas in an algebraic form. Our key tool consists in showing the equivalence of certain systems of linear equations. It throws more light on the algebraic nature of the problem and may be interesting in itself. The whole of our considerations are subject to only moderate prerequisites.

2. Preliminaries

For any matrix $\mathbf{A}$ of dimension $n \times p$ define the sets

$$\mathcal{R}(\mathbf{A}) = \{\mathbf{x} \in R^n : \mathbf{x} = \mathbf{A}\mathbf{y} \text{ for some } \mathbf{y} \in R^p\} \quad \text{(i.e. the } \textit{range} \text{ of } \mathbf{A})$$

and

$$\mathcal{N}(\mathbf{A}) = \{\mathbf{y} \in R^p : \mathbf{A}\mathbf{y} = \mathbf{0}\} \quad \text{(i.e. the } \textit{kernel} \text{ of } \mathbf{A}).$$

We note that

$$\mathbf{x}^T\mathbf{y} = \mathbf{0} \text{ for all } \mathbf{x} \in \mathcal{R}(\mathbf{A}) \text{ and } \mathbf{y} \in \mathcal{N}(\mathbf{A}^T).$$

It is clear that the range $\mathcal{R}(\mathbf{A})$ constitutes an r-dimensional linear space in R^n spanned by the columns of $\mathbf{A}$, where $r = rank(\mathbf{A})$, while $\mathcal{N}(\mathbf{A}^T)$ constitutes an $(n-r)$-dimensional space of all vectors being orthogonal to any vector in $\mathcal{R}(\mathbf{A})$ with respect to the usual inner product $(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T\mathbf{y}$. Thus any vector $\mathbf{x} \in R^n$ may be presented in the form

$$\mathbf{x} = \mathbf{x}_1 + \mathbf{x}_2, \text{ where } \mathbf{x}_1 \in \mathcal{R}(\mathbf{A}),\ \mathbf{x}_2 \in \mathcal{N}(\mathbf{A}^T) \text{ are orthogonal.}$$

Since $\mathbf{A}^T\mathbf{A}\mathbf{x} = \mathbf{0}$ if and only if $\mathbf{x}^T\mathbf{A}^T\mathbf{A}\mathbf{x} = \mathbf{0}$, and hence $\mathbf{A}\mathbf{x} = \mathbf{0}$, we get $\mathcal{N}(\mathbf{A}\mathbf{A}^T) = \mathcal{N}(\mathbf{A}^T)$.

Denote by $\mathbf{P} = \mathbf{P}_A$ the linear operator from R^n onto $\mathcal{R}(\mathbf{A})$ defined by

$$\mathbf{P}\mathbf{x} = \begin{cases} \mathbf{x}, & \text{if} \quad \mathbf{x} \in \mathcal{R}(\mathbf{A}) \\ \mathbf{0}, & \text{if} \quad \mathbf{x} \in \mathcal{N}(\mathbf{A}^T) \end{cases} \tag{1}$$

(i.e. the *orthogonal projector* onto $\mathcal{R}(\mathbf{A})$). It follows from definition (1) that $\mathbf{P}\mathbf{P} = \mathbf{P}$. The following lemma will be a key tool in our further considerations.

Lemma 1. For any matrix $\mathbf{A}$ and for any vector $\mathbf{b} \in \mathcal{R}(\mathbf{A})$ the following are equivalent:

(i) $\mathbf{Ax} = \mathbf{b}$,

(ii) $\mathbf{A}^T\mathbf{Ax} = \mathbf{A}^T\mathbf{b}$.

Proof. (i)$\Longrightarrow$(ii) is evident (without any condition on $\mathbf{b}$).
(ii)$\Longrightarrow$(i). By the assumption $\mathbf{b} \in \mathcal{R}(\mathbf{A})$, we get $\mathbf{b} = \mathbf{Ac}$ for some $\mathbf{c}$. Thus (ii) reduces to $\mathbf{A}^T\mathbf{Ax} = \mathbf{A}^T\mathbf{Ac}$ and its general solution is $\mathbf{x} = \mathbf{c} + \mathbf{x}_0$, where $\mathbf{x}_0 \in \mathcal{N}(\mathbf{A}^T\mathbf{A}) = \mathcal{N}(\mathbf{A})$. Therefore $\mathbf{x}$ is a solution of (i). □

Remark 1. The assumption $\mathbf{b} \in \mathcal{R}(\mathbf{A})$ in Lemma 1 is essential. To see this let us set

$$\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} \quad \text{and} \quad \mathbf{b} = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}.$$

Then $\mathbf{A}^T\mathbf{A} = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$ and $\mathbf{A}^T\mathbf{b} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$. Thus (ii) has a solution $\mathbf{x} = \begin{bmatrix} 0, & 0 \end{bmatrix}^T$, while (i) is inconsistent.

3. Main result

For any matrix $\mathbf{A}$ of dimension $n \times p$ and for any vector $\mathbf{b} \in R^n$ consider the linear equation

$$\mathbf{Ax} = \mathbf{b}. \tag{2}$$

The equation (2) may be consistent (if $\mathbf{b} \in \mathcal{R}(\mathbf{A})$), or inconsistent (otherwise). In the second case we are seeking such $\mathbf{x}$ that the residual vector $\mathbf{b} - \mathbf{Ax}$ is as small as possible.

Definition 1. Any vector $\widehat{\mathbf{x}} \in R^p$ is said to be the Least Squares Solution (LSS) of (2) if

$$(\mathbf{b} - \mathbf{A}\widehat{\mathbf{x}})^T(\mathbf{b} - \mathbf{A}\widehat{\mathbf{x}}) \leqslant (\mathbf{b} - \mathbf{Ax})^T(\mathbf{b} - \mathbf{Ax}) \quad \text{for any} \quad \mathbf{x} \in R^p. \tag{3}$$

The following theorem shows that this definition is not empty and reduces the LSS of an inconsistent equation (2) to the ordinary solution of a consistent one.

Theorem 1. (a) The equation (2) has at least one Least Squares Solution.

(b) Vector $\mathbf{x} \in R^P$ is a LSS of (2) if and only if

$$\mathbf{A}^T\mathbf{A}\mathbf{x} = \mathbf{A}^T\mathbf{b} \tag{4}$$

(c) The equation (4) is equivalent to

$$\mathbf{A}\mathbf{x} = \mathbf{P}\mathbf{b} \tag{5}$$

where $\mathbf{P}$ is the orthogonal projector onto $\mathcal{R}(\mathbf{A})$ defined by (1).

Remark 2. In the statistical literature the equation (4) is called the normal equations.

Proof. By properties of the projector $\mathbf{P}$ we get

$$\begin{aligned}(\mathbf{b}-\mathbf{A}\mathbf{x})^T(\mathbf{b}-\mathbf{A}\mathbf{x}) &= [\mathbf{P}\mathbf{b}+(\mathbf{I}-\mathbf{P})\mathbf{b}-\mathbf{A}\mathbf{x}]^T[\mathbf{P}\mathbf{b}+(\mathbf{I}-\mathbf{P})\mathbf{b}-\mathbf{A}\mathbf{x}]\\ &= (\mathbf{P}\mathbf{b}-\mathbf{A}\mathbf{x})^T(\mathbf{P}\mathbf{b}-\mathbf{A}\mathbf{x})+[(\mathbf{I}-\mathbf{P})\mathbf{b}]^T[(\mathbf{I}-\mathbf{P})\mathbf{b}]\\ &= (\mathbf{P}\mathbf{b}-\mathbf{A}\mathbf{x})^T(\mathbf{P}\mathbf{b}-\mathbf{A}\mathbf{x})+\mathbf{b}^T(\mathbf{I}-\mathbf{P})\mathbf{b}\\ &\geqslant \mathbf{b}^T(\mathbf{I}-\mathbf{P})\mathbf{b}\end{aligned}$$

with equality if and only if (5) holds. Moreover, by definition of $\mathbf{P}$, the equation (5) is consistent, and by Lemma 1 it is equivalent to (4). □

4. Application in linear regression

A model of linear regression with one explanatory variable may be presented in the form

$$\mathbf{y} = \mu\mathbf{1}_n + \alpha\mathbf{z} + \mathbf{e},$$

where $\mathbf{y} = (y_1, ..., y_n)^T$ is the observation vector, $\mathbf{1}_n$ means the column of n ones, $\mathbf{z} = (z_1, ..., z_n)^T$ is the vector of the observed explanatory variable, and $\mathbf{e} = (e_1, ..., e_n)^T$ is the vector of experimental errors (with the standard assumptions).

By setting in (4) $\mathbf{A} = [\mathbf{1}_n, \mathbf{z}]$ and $\mathbf{b} = \mathbf{y}$ we get

$$\mathbf{A}^T\mathbf{A} = \begin{bmatrix} n & \sum_{i=1}^n z_i \\ \sum_{i=1}^n z_i & \sum_{i=1}^n z_i^2 \end{bmatrix}$$

and

$$\mathbf{A}^T\mathbf{b} = \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n z_i y_i \end{bmatrix}.$$

Consequently the normal equations (4) with $\mathbf{x} = (\mu, \alpha)^T$ reduce to

$$\begin{bmatrix} n & \sum_{i=1}^n z_i \\ \sum_{i=1}^n z_i & \sum_{i=1}^n z_i^2 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n z_i y_i \end{bmatrix}.$$

Assuming that $r(\mathbf{A}) = 2$ we get

$$(\mathbf{A}^T\mathbf{A})^{-1} = \frac{1}{n\sum z_i^2 - (\sum z_i)^2} \begin{bmatrix} \sum_{i=1}^n z_i^2 & -\sum_{i=1}^n z_i \\ -\sum_{i=1}^n z_i & n \end{bmatrix}.$$

This leads to the Least Squares estimators

$$\widehat{\alpha} = \frac{n\sum z_i y_i - \sum z_i \sum y_i}{n\sum z_i^2 - (\sum z_i)^2}$$

and

$$\widehat{\mu} = \frac{\sum y_i}{n} - \widehat{\alpha}\frac{\sum z_i}{n}.$$

References

Bapat R.B. (2000): Linear Algebra and Linear Models, 2nd ed. Springer-Verlag, New York.

deLaubenfels R. (2006): The victory of least squares and orthogonality in statistics, Amer. Statist. 60: 315–321.

Fisher R. A. (1915): Frequency distribution of the values of the correlation coefficient in samples from infinitely large population, Biometrika 10: 507–521.

Gauss C. F. (1857): Theory of Motion of the Heavenly Bodies, [translation: C.H. Davies, Little Brown, Boston].

Hamilton C. H. (1933): Algebraic derivation of the normal equations in multiple and partial correlation, J. Amer. Statist. Assoc. 28: 204–208.

Herr D. C. (1980): On the history of the use of geometry in the general linear model, Amer. Statist. 34: 43–47.

Kruskal W. (1961): The coordinate-free approach to Gauss-Markov estimation and its application to missing and extra observations, 3th Berkeley Symp. Math. Statist. Prob. 1: 435–451.

Kruskal W. (1968): When are Gauss-Markov and least squares estimators identical? Ann. Math. Statist. 39: 70–75.

Legendre A. M. (1806): Nouvelles méthodes pour la détermination des orbites des comètes, Courcier, Paris, France.

Rao C. R. (1973): Linear Statistical Inference and its Applications, 2nd ed. Wiley, New York.

Rao C.R., Mitra S.K. (1971): Generalized Inverse of Matrices and its Applications, Wiley, New York.

Scheffé H. (1959): The Analysis of Variance, Wiley, New York.