Some equivalences in linear estimation[*]

Dmitry Danilov
Eurandom, Eindhoven University of Technology

Jan R. Magnus
CentER and Department of Econometrics & OR
Tilburg University

Russian version appeared in *Quantile*, 3, 83–90, 2007.

Key words: Linear Bayes estimation, best linear unbiased, least squares, sparse problems, large-scale optimization.

Abstract: Under normality, the Bayesian estimation problem, the best linear unbiased estimation problem, and the restricted least-squares problem are all equivalent. As a result we need not compute pseudo-inverses and other complicated functions, which will be impossible for large sparse systems. Instead, by reorganizing the inputs, we can rewrite the system as a new but equivalent system which can be solved by ordinary least-squares methods.

Corresponding author: Jan R. Magnus, CentER and Department of Econometrics & OR, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands. Phone: +31-13-466-3092, fax: +31-13-466-3066, e-mail: magnus@uvt.nl.

[*] We thank the editor of this journal for his constructive comments.

1 Introduction

Suppose we are given an $n \times 1$ vector y and an $n \times k$ matrix X with linearly independent columns. The vector y and the matrix X are assumed to be known (and non-stochastic). The problem is to determine the $k \times 1$ vector β that satisfies the equation

$$y = X\beta.$$

Let $M := I_n - X(X'X)^{-1}X'$ be the usual idempotent matrix. If $My = 0$, then the equation $y = X\beta$ has a unique solution

$$\hat{\beta} := (X'X)^{-1}X'y.$$

If $My \neq 0$, then the equation has no solution. In that case we may seek a vector $\hat{\beta}$ which, in a sense, minimizes the "error" vector $e = y - X\beta$. A convenient scalar measure of the error would be

$$e'e = (y - X\beta)'(y - X\beta),$$

and we know that $\hat{\beta}$ minimizes $e'e$ over all real k-dimensional β-vectors. The vector $\hat{\beta}$ is called the *least-squares solution* and $X\hat{\beta}$ the *least-squares approximation* to y. Thus $\hat{\beta}$ is the "best" choice for β whether the equation $y = X\beta$ is consistent or not. If $y = X\beta$ is consistent, then $\hat{\beta}$ is the solution; if $y = X\beta$ is not consistent, then $\hat{\beta}$ is the least-squares solution.

In contrast to the deterministic least-squares problem, the standard linear regression problem is framed in a random set-up where we consider

$$y = X\beta + \varepsilon, \quad \varepsilon \sim \mathrm{N}(0, \sigma^2 I_n),$$

from which we obtain the Gauss-Markov estimator and its variance matrix as

$$\hat{\beta} = (X'X)^{-1}X'y, \quad \mathrm{var}(\hat{\beta}) = \sigma^2(X'X)^{-1},$$

see e.g. Magnus, Katyshev and Peresetsky (1997). The fact that the least-squares solution and the Gauss-Markov estimator are identical is by no means obvious and was thoroughly investigated and highlighted by Rao (1971, 1973). The equivalence has led to the unfortunate usage of the term "(ordinary) least-squares *estimator*" meaning the Gauss-Markov estimator. The method of least squares, however, is a purely deterministic method which has to do with approximation, *not* with estimation.

In contrast to the classical (frequentist) approach, a Bayesian does not assume "true" β-parameters. Instead, a probability distribution of the parameters is assumed, the so-called prior distribution. The data then serve to modify the prior idea of the "truth" into a more complete idea: the posterior distribution. The formula that generates this transformation is Bayes' formula:

$$p(\beta|y) = \frac{\pi(\beta)p(y|\beta)}{p(y)},$$

where $\pi(\beta)$ denotes the prior distribution, $p(y|\beta)$ is the usual likelihood function, $p(\beta|y)$ is the posterior distribution, and $p(y)$ is a proportionality constant, which follows from the fact that $p(\beta|y)$ must integrate to one. For example, if we assume (for the moment) that Σ is known and that the relevant distributions are normal, then a Bayesian would use data

$$y|\beta \sim \mathrm{N}(X\beta, \Sigma),$$

precisely as a classical statistician. But in addition, the Bayesian would use all other information on β, however vague, in so-called priors, say

$$\beta \sim \mathrm{N}(h, H).$$

The posterior can then be shown to be $p(\beta|y) \sim \mathrm{N}(\hat{\beta}, V)$, where

$$V = (H^{-1} + X'\Sigma^{-1}X)^{-1}, \quad \hat{\beta} = V(H^{-1}h + X'\Sigma^{-1}y).$$

The mean of the posterior distribution, $\hat{\beta}$, can then be viewed as an "estimator" of β, and we see that it is a matrix-weighted average between the prior mean h and the classical GLS estimator $(X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}y$. A Bayesian prefers to talk about precision rather than variance, the latter being the inverse of the former. Then, $V^{-1} = H^{-1} + X'\Sigma^{-1}X$, in words:

$$\text{posterior precision} = \text{prior precision} + \text{data precision}.$$

Hence the precision always increases (the variance decreases) when more information is added, whether in the form of data or priors. If $H^{-1} = 0$, then there is no prior information and we obtain the classical results as a special case.

In a normal Bayesian framework, that is, a framework where both the likelihood and the priors are based on the normal distribution, the posterior is normal as well (as we have just seen), and therefore there is no mathematical difference between data and priors, although there is of course a conceptual difference. This simple observation leads to further equivalences which are explored in this note.

The reason for investigating these equivalences is both conceptual and computational. In the Bayesian set-up of Section 2, the formulae are not computable in situations where the $n \times k$ design matrix X is of "large" dimension and is "sparse". A matrix is "sparse" when it has many structural zeros. If an $n \times k$ matrix possesses s structural zeros, then the matrix can be stored as a $(nk - s) \times 3$ matrix, where the i-th row contains the row-index, column-index, and value of the i-th nonzero entry. This is often useful if storage space is more important than access speed. Recently, large sparse matrices have become important in economics, for example in studying longitudinal samples of over one million workers from more than 500,000 employing firms; see Abowd, Kramarz, and Margolis (1999) and Abowd, Creecy, and Kramarz (2002), using an algorithm due to Dongarra, Duff, Sorenson, and van der Vorst (1991), or in estimating national accounts data, especially in developing countries, where one may encounter 5000 or more variables and 20,000 observations; see Magnus, van Tongeren, and de Vos (2000), now using the Snaer software developed by Danilov and Magnus (2007).

In Section 2 we formulate our basic question, which is phrased in Bayesian terminology. In Section 3 we show the equivalence of the Bayesian problem with various optimization schemes, including restricted and unrestricted least squares, and best linear unbiased estimation. All these methods are equivalent to an unrestricted least-squares estimation problem. In Section 4 we briefly interpret our results. Finally, in Section 5, we offer some thoughts about proving matrix equalities.

2 Data and priors

We are interested in a vector $\beta = (\beta_1, \beta_2, \dots, \beta_k)'$ consisting of k latent (random) variables. Data are available on n_1 linear combinations of β. Let y_1 denote the $n_1 \times 1$ data vector. Our starting point is a measurement equation,

$$y_1|\beta \sim \mathrm{N}_{n_1}(X_1\beta, \Sigma_1). \tag{1}$$

The $n_1 \times k$ matrix X_1 often takes the form of a selection matrix, say $X_1 = (I_{n_1}, 0)$, so that $X_1\beta$ is a subvector of β, but this is not required here. Neither is it required that the matrix X_1 has full row-rank. (The condition that X_1 has full row-rank was made in Magnus, van Tongeren, and de Vos (2000, Theorem 1), but it is in fact redundant.) Measurements are unbiased in the sense that $\mathrm{E}(y_1|\beta) = X_1\beta$. The $n_1 \times n_1$ matrix Σ_1 denotes a positive definite variance matrix, typically (but not necessarily) diagonal.

In addition to the n_1 data, we have access to two further pieces of information: prior views concerning the latent variables or linear combinations thereof, and deterministic linear constraints. In particular, we have m_1 random priors:

$$R_1\beta \sim \mathrm{N}_{m_1}(h_1, H_1) \tag{2}$$

and m_2 exact restrictions (identities):

$$R_2\beta = h_2 \text{ (almost surely)}, \tag{3}$$

in total $m := m_1 + m_2$ pieces of prior information.

We assume that the $m_1 \times m_1$ matrix H_1 is positive definite (hence nonsingular) and that the $m_2 \times k$ matrix R_2 has full row-rank m_2 (so that the exact restrictions are linearly independent and thus form a consistent set of equations). We define

$$R := \begin{pmatrix} R_1 \\ R_2 \end{pmatrix}, \quad h := \begin{pmatrix} h_1 \\ h_2 \end{pmatrix}, \quad H := \begin{pmatrix} H_1 & 0 \\ 0 & 0 \end{pmatrix},$$

and assume that $\mathrm{rk}(R) = m$, which, of course, implies that both R_1 and R_2 have full row-rank. We shall see later that the rank condition on R is not a serious restriction, because of the fact that there is no mathematical difference in the normal case between priors and data, so that we can always consider all priors as data (and vice versa). Hence the condition $m \leq k$ is not restrictive either.

In order to identify all k variables from the information (data and priors) we need at least k pieces of information: $m + n_1 \geq k$. But this is not sufficient for identification, because some of the information may be on the same variables. If $m = k$, all variables are identified. If $m < k$, we introduce a semi-orthogonal $k \times (k - m)$ matrix L such that $RL = 0$ and $L'L = I_{k-m}$, and necessary and sufficient for identification is the condition

$$\mathrm{rk}(X_1 L) = k - m.$$

An alternative and equivalent condition is:

$$\mathrm{rk}\begin{pmatrix} R \\ X_1 \end{pmatrix} = k.$$

This follows because the definition of L implies that

$$\mathrm{rk}\begin{pmatrix} R \\ X_1 \end{pmatrix} = \mathrm{rk}(R) + \mathrm{rk}(X_1 L).$$

3 Equivalences

We now present six estimators of β (and their variances), all equivalent. This equivalence is based on two facts. First, a Bayesian analysis with normal data and normal priors is closely linked with a quadratic minimization problem. Second, best linear unbiased estimation is also closely linked to quadratic minimization (least squares).

3.1 Bayesian solution

Using Theorem 1 of Magnus, van Tongeren, and de Vos (2000), we see that the posterior distribution of β is given by

$$\beta | y_1 \sim \mathrm{N}_k(\hat{\beta}, V),$$

where

$$V = R^+HR^{+\prime} - R^+HR^{+\prime}X_1'\Sigma_0^{-1}X_1R^+HR^{+\prime} + CKC' \tag{4}$$

and

$$\hat{\beta} = R^+h - (R^+HR^{+\prime} + CK)X_1'\Sigma_0^{-1}(X_1R^+h - y_1). \tag{5}$$

In these expressions, the following definitions have been employed:

$$\Sigma_0 := \Sigma_1 + X_1R^+HR^{+\prime}X_1', \quad C := I_k - R^+HR^{+\prime}X_1'\Sigma_0^{-1}X_1,$$

and

$$K := \begin{cases} L(L'X_1'\Sigma_0^{-1}X_1L)^{-1}L' & \text{if } m < k, \\ 0 & \text{if } m = k. \end{cases}$$

Also, the notation A^+ denotes the Moore-Penrose inverse of a matrix A. Although this result is of theoretical interest, it is not practical for computations, especially when the dimensions are large. Hence we seek alternative, but equivalent, formulations, of these posterior moments.

3.2 Only data, no random priors

The first step towards simplification is to interpret all the random priors as data, and consider the new $n := n_1 + m_1$ "data"-vector

$$y := \begin{pmatrix} y_1 \\ h_1 \end{pmatrix}.$$

Defining

$$X := \begin{pmatrix} X_1 \\ R_1 \end{pmatrix}, \quad \Sigma := \begin{pmatrix} \Sigma_1 & 0 \\ 0 & H_1 \end{pmatrix},$$

we may write the measurement equation as

$$y|\beta \sim \mathrm{N}_n(X\beta, \Sigma)$$

together with the priors (exact restrictions)

$$R_2\beta = h_2 \text{ (almost surely).}$$

We assume that $m_2 < k$ and we define a semi-orthogonal $k \times (k - m_2)$ matrix L_2 such that $R_2L_2 = 0$ and $L_2'L_2 = I_{k-m_2}$. The identifiability condition becomes

$$\mathrm{rk}\begin{pmatrix} R_2 \\ X \end{pmatrix} = k$$

or alternatively

$$\mathrm{rk}(XL_2) = k - m_2.$$

The posterior moments of $\beta|y$ are then given by

$$\hat{\beta} = R_2^+h_2 - VX'\Sigma^{-1}(XR_2^+h_2 - y) \tag{6}$$

and

$$V = L_2\left(L_2'X'\Sigma^{-1}XL_2\right)^{-1}L_2'. \tag{7}$$

These two moments are numerically identical to the ones presented in (5) and (4).

3.3 Best linear unbiased estimation

An alternative view, also leading to the same results, is to consider the regression model

$$y \sim \mathrm{N}(X\beta, \Sigma)$$

subject to the linear constraints

$$R_2\beta = h_2,$$

where β is now a *nonrandom* parameter vector to be estimated. The best linear unbiased estimator of β is given by

$$\hat{\beta} = G^{-1}X'\Sigma^{-1}y + G^{-1}R_2'(R_2G^{-1}R_2')^{-1}(h_2 - R_2G^{-1}X'\Sigma^{-1}y) \tag{8}$$

with variance

$$V = G^{-1} - G^{-1}R_2'(R_2G^{-1}R_2')^{-1}R_2G^{-1}, \tag{9}$$

where $G := X'\Sigma^{-1}X + R_2'R_2$; see Magnus and Neudecker (1988, Theorem 13.6). Again, the expressions (8) and (9) are numerically identical to (5) and (4).

3.4 Restricted least squares

There is a close relationship between best linear unbiased estimation and least squares (Rao, 1971, 1973). As shown in Magnus and Neudecker (1988, Theorem 13.16), we can obtain $\hat{\beta}$ in this case also as the solution of

$$\begin{aligned} &\text{minimize } (y - X\beta)'\varSigma^{-1}(y - X\beta) \\ &\text{subject to } R_2\beta = h_2. \end{aligned}$$

As discussed in the Introduction, least squares is a deterministic method. All best linear unbiased estimation problems allow for an equivalent least-squares formulation, but the weighting matrix is different in each case. Of course, equivalence only holds if we use the correct weighting matrix, in this case, $\varSigma^{-1}$.

3.5 Unrestricted least squares, 1

Alternatively, we can take the restriction as part of the data (with zero variance). Then we need to find the solution of

$$\text{minimize} \begin{pmatrix} y - X\beta \\ h_2 - R_2\beta \end{pmatrix}' \begin{pmatrix} \varSigma + XX' & XR_2' \\ R_2X' & R_2R_2' \end{pmatrix}^{-1} \begin{pmatrix} y - X\beta \\ h_2 - R_2\beta \end{pmatrix},$$

as shown in Corollary 1 of Theorem 13.15 in Magnus and Neudecker (1988).

3.6 Unrestricted least squares, 2

As a final equivalence, we can explicitly solve the restrictions, thus reducing the dimension of the problem. This is achieved by writing

$$R_2 = (R_{21} : R_{22}),$$

where R_{21} is an $m_2 \times (k - m_2)$ matrix and R_{22} is a *nonsingular* $m_2 \times m_2$ matrix. Partitioning β correspondingly, we write the restriction as

$$R_{21}\beta_1 + R_{22}\beta_2 = h_2,$$

so that

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} I \\ -R_{22}^{-1}R_{21} \end{pmatrix} \beta_1 + \begin{pmatrix} 0 \\ R_{22}^{-1}h_2 \end{pmatrix} \equiv Q\beta_1 + q.$$

Let

$$X^* := \varSigma^{-1/2}X = \begin{pmatrix} \varSigma_1^{-1/2}X_1 \\ H_1^{-1/2}R_1 \end{pmatrix}, \quad y^* := \varSigma^{-1/2}y = \begin{pmatrix} \varSigma_1^{-1/2}y_1 \\ H_1^{-1/2}h_1 \end{pmatrix}.$$

Then the constrained problem

$$\begin{aligned} &\text{minimize } (y - X\beta)'\varSigma^{-1}(y - X\beta) \\ &\text{subject to } R_2\beta = h_2. \end{aligned}$$

can be written equivalently as

$$\text{minimize } \|(y^* - X^*q) - X^*Q\beta_1\|^2$$

with respect to β_1. This is a simple unrestricted least-squares problem.

4 Interpretation

The purpose of this note has been to demonstrate the equivalence between three methods under normality: least squares, best linear unbiased estimation, and Bayesian estimation. The proved equivalences are conceptually of interest, but they are also of practical importance. In today's econometrics applications, the dimensions of the matrices can become very large, for example, in financial data or national accounts data. Standard calculations such as inverses (even worse, Moore-Penrose inverses) may then become unstable or even infeasible. The use of stable and simple formulae is then essential.

Solving a restricted least-squares problem by a two-step procedure has the advantage that the dimension of the system is significantly reduced. Moreover, in many practical situations the first step can be done once and then the results of the reduction may be used for many restricted least-squares problems. In particular, the R_2 matrix is usually fixed because it represents the economic structure, while the matrices related to R_1 are priors that will vary. Hence, a method where calculations related to R_2 are done just once will be of practical importance.

The equivalences provided in this note are all exact, and no information is lost. If dimensions increase even further, then even the simplest formulation provided here may become unstable. In such cases one should seek for "robust" alternatives, that is, methods that are computationally feasible, are *not* necessarily exact, but do not deviate much from the exact solutions, also in extreme cases.

5 Some thoughts on proving matrix equalities

Suppose we wish to prove that $A = B$ for two given matrices A and B, such as (4) and (7), or (5) and (6). What methods are available to us? First, we could simply try and prove that $A = B$ by a direct method, for example by proving that $a_{ij} = b_{ij}$ for all i and j. This is usually not a good method. A second method, somewhat better, is to consider $\varDelta := A - B$ and to prove that $\varDelta = 0$. Third, and usually faster, is to consider not the matrix equation $\varDelta = 0$, but the vector equation $\varDelta x = 0$ for all vectors x. Equivalently, one could even try and prove the scalar equation $x'\varDelta'\varDelta x = 0$ for all x. This third method is essentially a geometrical method: we consider the mappings from x to Ax and Bx. If the result of the two mappings is the same for every x, then the mappings themselves must be the same too.

A fourth and somewhat different idea can be used if $\varDelta$ depends on a matrix X, so that we need to prove that $\varDelta(X) = 0$ for every X. Letting d denote the differential, it is then sufficient to prove that

$$\mathrm{d}(\varDelta(X)) = 0, \qquad \varDelta(X_0) = 0$$

for some arbitrarily and suitably chosen matrix X_0 (usually the null matrix or the identity matrix); see Exercise 13.69 in Abadir and Magnus (2005) for an example.

Interestingly, the equalities in this note are not proved by any of these methods. Instead we rely on the fact that the frameworks from which the equations arise are equivalent, and hence that the resulting expressions must be the same too. Direct proofs of the equalities are of course possible, but very laborious.

References

Abadir, K.M., and J.R. Magnus (2005), *Matrix Algebra*, Econometric Exercises Volume 1, Cambridge University Press, New York.

Abowd, J.M., R.H. Creecy, and F. Kramarz (2002), Computing person and firm effects using linked longitudinal employer-employee data, Cornell University Working Paper, Ithaca, NY.

Abowd, J.M., F. Kramarz, and D.N. Margolis (1999), High wage workers and high wage firms, *Econometrica*, 67, 251–333.

Danilov, D., and J.R. Magnus (2007), On the estimation of a large sparse Bayesian system: the Snaer project, submitted for publication.

Dongarra, J.J., I.S. Duff, D.C. Sorenson, and H.A. van der Vorst (1991), *Solving Linear Systems on Vector and Shared Memory Computers*, SIAM, Philadelphia.

Magnus, J.R., P.K. Katyshev and A.A. Peresetsky (1997). *Эконометрика. Начальный курс* (Econometrics: A First Course), Delo Publishers, Moscow. Seventh edition 2005.

Magnus, J. R. and H. Neudecker (1988). *Matrix Differential Calculus with Applications in Statistics and Econometrics*, John Wiley and Sons, Chichester/New York. Second edition (paperback) 1999, Third edition 2007. Russian translation: *Матричное дифференциальное исчисление с приложениями к статистике и эконометрике*, Fizmatlit Publishing House, Moscow, 2002.

Magnus, J.R., J.W. van Tongeren, and A.F. de Vos (2000), National account estimation using indicator ratios, *The Review of Income and Wealth*, 46, 329–350.

Rao, C.R. (1971), Unified theory of linear estimation, *Sankhyā, A*, 33, 371–477. (Corrigenda, *Sankhyā, A*, 34, 477.)

Rao, C.R. (1973), *Linear Statistical Inference and Its Applications*, 2nd edition, John Wiley, New York.