

ON AN IDENTITY DERIVED FROM UNBIASEDNESS IN LINEAR MODELS

C. E. McCulloch and S. R. Searle

Biometrics Unit and Statistics Center, Cornell University, Ithaca, NY 14853

BU-1166-MC

August, 1994

Abstract

Difficulties associated with deriving $\mathbf{AX} = \mathbf{X}$ as a necessary condition from $\mathbf{Ay} + \mathbf{c}$ being an unbiased estimator of $\mathbf{X}\boldsymbol{\beta}$ are discussed in terms of the linear model $\mathbf{y} \sim (\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$. We demonstrate two instances in which $\mathbf{AX} = \mathbf{X}$ and $\mathbf{c} = \mathbf{0}$ are not necessary conditions but then argue that these can be ignored in practice. Excluding these possibilities leads to $\mathbf{AX} = \mathbf{X}$ and $\mathbf{c} = \mathbf{0}$ being necessary.

1. Introduction

Developing results pertaining to the linear model $\mathbf{y} \sim (\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$ often involves wanting to draw conclusions about a matrix \mathbf{A} and a vector \mathbf{c} for which $\mathbf{Ay} + \mathbf{c}$ is an unbiased estimator of $\mathbf{X}\boldsymbol{\beta}$. Thus with

$$E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta} \quad \text{and} \quad \text{Var}(\mathbf{y}) = \mathbf{V} \tag{1}$$

being, respectively, the mean and dispersion matrix of \mathbf{y} , unbiasedness of $\mathbf{Ay} + \mathbf{c}$ for $\mathbf{X}\boldsymbol{\beta}$ means that

$$E(\mathbf{Ay} + \mathbf{c}) = E(\mathbf{Ay}) + \mathbf{c} = \mathbf{X}\boldsymbol{\beta} . \tag{2}$$

Assume that \mathbf{A} is nonstochastic, in which case (2) becomes

$$E(\mathbf{Ay} + \mathbf{c}) = \mathbf{AX}\boldsymbol{\beta} + \mathbf{c} = \mathbf{X}\boldsymbol{\beta} . \tag{3}$$

Sufficient conditions for (3) to be true are, obviously,

$$\mathbf{AX} = \mathbf{X} \quad \text{and} \quad \mathbf{c} = \mathbf{0} . \tag{4}$$

Developing (4) as a necessary condition often proceeds as follows. If (1) is to be true in general then

(4) must hold for all β . Setting $\beta = 0$ shows that $c = 0$ and by letting β successively be the columns of an identity matrix, of the same order as β , (3) yields (4).

There are, however, two instances under which this argument fails and $\mathbf{AX} = \mathbf{X}$ and $c = 0$ are not necessary conditions. That is, it is possible to have $E(\mathbf{Ay} + c) = \mathbf{X}\beta$ without (4). We illustrate these possibilities through examples. Throughout the paper we are interested mainly in when (2) is necessary for (4), and not the sufficiency argument when \mathbf{AX} is known equal to \mathbf{X} .

2. Situations where $\mathbf{AX} = \mathbf{X}$ and $c = 0$ are not necessary

2.1 In \mathbf{Ay} the matrix \mathbf{A} is a function of \mathbf{y} .

For $\mathbf{y}' = [y_1 \ y_2 \ y_3 \ y_4]$ being a vector of four observations, let its expected value be

$$E(\mathbf{y}) = E \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} \beta_1 \\ \beta_1 \\ \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \mathbf{X}\beta,$$

for

$$\mathbf{X} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \quad \text{and} \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}. \tag{5}$$

Example 1a: For (5) we illustrate the possibility of having \mathbf{A} depend on \mathbf{y} with $E(\mathbf{Ay} + c) = \mathbf{X}\beta$ but with $\mathbf{AX} \neq \mathbf{X}$. So long as $y_2 \neq 0$, defining (with probability 1)

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & y_1/y_2 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad \text{gives} \quad E(\mathbf{Ay}) = \mathbf{X}\beta \quad \text{but} \quad \mathbf{AX} = \begin{bmatrix} 1 & 0 \\ y_1/y_2 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \neq \mathbf{X}.$$

Even though \mathbf{A} depends on \mathbf{y} , note that \mathbf{Ay} is nevertheless a linear function of \mathbf{y} . This need not always be the case as illustrated by the following example, adapted from Harville (1981).

Example 1b: Suppose that in combination with $E(\mathbf{y})$ of (5), we also have

$$\mathbf{V} = \text{Var}(\mathbf{y}) = \begin{bmatrix} 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 1 \end{bmatrix}.$$

Then for $\mathbf{a}' = [1 - y_3 - y_4 \quad 0 \quad y_3 \quad y_3]$ we have

$$\begin{aligned} E(\mathbf{a}'\mathbf{y}) &= E(y_1 - y_1y_3 - y_1y_4 + y_3^2 + y_3y_4) \\ &= \beta_1 - (1 + \beta_1^2) - (1 + \beta_1\beta_2) + (1 + \beta_1^2) + (1 + \beta_1\beta_2) \\ &= \beta_1. \end{aligned}$$

Hence for

$$\mathbf{A} = \begin{bmatrix} 1 - y_3 - y_4 & 0 & y_3 & y_3 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$E(\mathbf{A}\mathbf{y}) = E \begin{bmatrix} \mathbf{a}'\mathbf{y} \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} \beta_1 \\ \beta_1 \\ \beta_1 \\ \beta_2 \end{bmatrix} = \mathbf{X}\boldsymbol{\beta}.$$

Notice that this cannot be written as $E(\mathbf{A}\mathbf{y})$ equaling $\mathbf{A}E(\mathbf{y}) = \mathbf{A}\mathbf{X}\boldsymbol{\beta}$ because, with \mathbf{A} being a function of \mathbf{y} , we do not have $E(\mathbf{A}\mathbf{y})$ equal to $\mathbf{A}E(\mathbf{y})$; and yet $E(\mathbf{A}\mathbf{y})$ does equal $\mathbf{X}\boldsymbol{\beta}$. And, of course,

$$\mathbf{A}\mathbf{X} = \begin{bmatrix} 1 - y_3 - y_4 & 0 & y_3 & y_3 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 - y_4 & y_3 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \neq \mathbf{X}.$$

Note in this example that \mathbf{A} is a function of \mathbf{y} , just as in Example 1a, but in contrast to Example 1a we have here $\mathbf{A}\mathbf{y}$ being a nonlinear function of \mathbf{y} . Yet in both cases $E(\mathbf{A}\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$ with $\mathbf{A}\mathbf{X} \neq \mathbf{X}$.

2.2 Estimable constraints on β

The second situation in which $E(\mathbf{A}\mathbf{y}) = \mathbf{X}\beta$ but $\mathbf{A}\mathbf{X} \neq \mathbf{X}$ can occur is when estimable constraints on β are known. We begin with an example.

Example 2: Suppose in (5) we know *a priori*, or are prepared to assume, that $\beta_1 = \beta_2$. Then for

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad \text{we have} \quad E(\mathbf{A}\mathbf{y}) = \mathbf{X}\beta \quad \text{but} \quad \mathbf{A}\mathbf{X} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ \frac{1}{2} & \frac{1}{2} \\ 0 & 1 \end{bmatrix} \neq \mathbf{X}.$$

This is so because $E(\mathbf{A}\mathbf{y})$ becomes $\mathbf{X}\beta$ through using $\beta_1 = \beta_2$ which, in the words of Christensen, “is more information than is given by the model.” Therefore, as statisticians knowing $\beta_1 = \beta_2$ *a priori*, we would usually reformulate the model as

$$E \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} \beta_1 \\ \beta_1 \\ \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \beta_1 = \mathbf{X}^* \beta^*,$$

with

$$\mathbf{X}^* = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \quad \text{and} \quad \beta^* = \beta_1.$$

For this model, $\mathbf{A}\mathbf{X}^* = \mathbf{X}^*$ and $\mathbf{c} = \mathbf{0}$ is necessary for an \mathbf{A} not dependent on \mathbf{y} to be such as to satisfy $E(\mathbf{A}\mathbf{y} + \mathbf{c}) = \mathbf{X}^* \beta^*$.

The constraints $\beta_1 = \beta_2$ in Example 2 can be generalized to

$$\mathbf{T}\mathbf{X}\beta = \mathbf{d} \tag{6}$$

for known \mathbf{T} and \mathbf{d} such that $\mathbf{T}\mathbf{X}$ is of full row rank and \mathbf{d} is in the column space of $\mathbf{T}\mathbf{X}$. That is, using the constraint $\mathbf{T}\mathbf{X}\beta = \mathbf{d}$ we can find an estimator such that $E(\mathbf{A}\mathbf{y} + \mathbf{c}) = \mathbf{X}\beta$ but $\mathbf{A}\mathbf{X} \neq \mathbf{X}$. We illustrate this situation via Example 3, and give a general proof in the Appendix. Note in passing that

\mathbf{d} being in the column space of \mathbf{TX} is needed for equations (6) to be consistent; and there is no loss of generality in requiring \mathbf{TX} to be of full row rank because if it is not then (6) has some redundant equations, since some equations in (6) will then be linear combinations of others.

Example 3: Consider the model equation $y_{ij} = \mu + \alpha_i + e_{ij}$ for the 1-way classification with 2, 2 and 3 observations in three classes. Then suppose we have constraints (based on estimable functions)

$$\begin{aligned} \alpha_1 - \alpha_2 &= 2 \\ 7\mu + 2\alpha_1 + 2\alpha_2 + 3\alpha_3 &= 0. \end{aligned} \tag{7}$$

Representing the seven observations as $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$ gives

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} \quad \text{and} \quad \mathbf{X}\boldsymbol{\beta} = \begin{bmatrix} \mu + \alpha_1 \\ \mu + \alpha_1 \\ \mu + \alpha_2 \\ \mu + \alpha_2 \\ \mu + \alpha_3 \\ \mu + \alpha_3 \\ \mu + \alpha_3 \end{bmatrix}. \tag{8}$$

Writing (7) as $\mathbf{TX}\boldsymbol{\beta} = \mathbf{d}$ we have

$$\begin{bmatrix} 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \mathbf{X}\boldsymbol{\beta} = \begin{bmatrix} 2 \\ 0 \end{bmatrix}$$

and so

$$\mathbf{T} = \begin{bmatrix} 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \quad \text{and} \quad \mathbf{d} = \begin{bmatrix} 2 \\ 0 \end{bmatrix}.$$

With \mathbf{T} having two rows, define $\mathbf{X} = [\mathbf{X}_1 \quad \mathbf{X}_2]$ for

$$\mathbf{X}_1 = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \end{bmatrix} \quad \text{and} \quad \mathbf{X}_2 = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix},$$

and this gives

$$\mathbf{TX}_1 = \begin{bmatrix} 0 & 1 \\ 7 & 2 \end{bmatrix} \quad \text{and} \quad (\mathbf{TX}_1)^{-1} = \frac{1}{7} \begin{bmatrix} -2 & 1 \\ 7 & 0 \end{bmatrix}.$$

Then define

$$\mathbf{W} = [\mathbf{I} - \mathbf{X}_1(\mathbf{TX}_1)^{-1}\mathbf{T}]\mathbf{X}_2$$

$$= \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} - \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \end{bmatrix} \frac{1}{7} \begin{bmatrix} -2 & 1 \\ 7 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}$$

$$= \frac{1}{7} \begin{bmatrix} 3 & -3 \\ 3 & -3 \\ 3 & -3 \\ 3 & -3 \\ -4 & 4 \\ -4 & 4 \\ -4 & 4 \end{bmatrix} \quad \text{and so} \quad \mathbf{W} = \frac{7}{3} \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

Now define

$$\mathbf{A} = \mathbf{W}\mathbf{W}^{\top} = \frac{1}{3} \begin{bmatrix} 3 \\ 3 \\ 3 \\ 3 \\ -4 \\ -4 \\ -4 \end{bmatrix} \mathbf{0}_{7 \times 6} \quad \text{and observe that} \quad \mathbf{A}\mathbf{X} = \frac{1}{3} \begin{bmatrix} 3 & 3 & 0 & 0 \\ 3 & 3 & 0 & 0 \\ 3 & 3 & 0 & 0 \\ 3 & 3 & 0 & 0 \\ -4 & -4 & 0 & 0 \\ -4 & -4 & 0 & 0 \\ -4 & -4 & 0 & 0 \end{bmatrix} \neq \mathbf{X}.$$

But, using

$$\begin{aligned} \mathbf{c} &= -(\mathbf{A} - \mathbf{I})\mathbf{X}_1(\mathbf{TX}_1)^{-1}\mathbf{d} = (\mathbf{X}_1 - \mathbf{AX}_1)(\mathbf{TX}_1)^{-1}\mathbf{d} \\ &= \begin{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \end{bmatrix} - \frac{1}{3} \begin{bmatrix} 3 & 3 \\ 3 & 3 \\ 3 & 3 \\ 3 & 3 \\ -4 & -4 \\ -4 & -4 \\ -4 & -4 \end{bmatrix} \end{bmatrix} \frac{1}{7} \begin{pmatrix} 0 & 1 \\ 7 & 2 \end{pmatrix} \begin{pmatrix} 2 \\ 0 \end{pmatrix} = \begin{bmatrix} 0 \\ 0 \\ -2 \\ -2 \\ 4/3 \\ 4/3 \\ 4/3 \end{bmatrix} \end{aligned}$$

it follows that

$$\mathbf{E}(\mathbf{Ay} + \mathbf{c}) = \mathbf{AX}\boldsymbol{\beta} + \mathbf{c} = \begin{bmatrix} \mu + \alpha_1 \\ \mu + \alpha_1 \\ \mu + \alpha_1 - 2 \\ \mu + \alpha_1 - 2 \\ -4(\mu + \alpha_1)/3 + 4/3 \\ -4(\mu + \alpha_1)/3 + 4/3 \\ -4(\mu + \alpha_1)/3 + 4/3 \end{bmatrix}. \quad (9)$$

At first sight this does not look much like $\mathbf{X}\boldsymbol{\beta}$ of (8): but in fact it is exactly $\mathbf{X}\boldsymbol{\beta}$. This is so because of the constraints of (7). First, $\alpha_1 - \alpha_2 = 2$ in (7) gives $\mu + \alpha_1 - 2 = \mu + \alpha_2$ in (9); and using both $\alpha_1 - \alpha_2 = 2$ and $7\mu + 2\alpha_1 + 2\alpha_2 + 3\alpha_3 = 0$ of (7) gives $-4(\mu + \alpha_1)/3 + 4/3 = \mu + \alpha_3$ in (9). Thus $\mathbf{E}(\mathbf{Ay} + \mathbf{c}) = \mathbf{X}\boldsymbol{\beta}$

3. Establishing the necessity of $\mathbf{AX} = \mathbf{X}$

From the preceding section we see there are two instances for which $\mathbf{AX} = \mathbf{X}$ and $\mathbf{c} = \mathbf{0}$ are not necessary conditions for $\mathbf{E}(\mathbf{Ay} + \mathbf{c}) = \mathbf{X}\boldsymbol{\beta}$:

- (i) \mathbf{A} being a random function of \mathbf{y} ,
- (ii) $\mathbf{TX}\boldsymbol{\beta} = \mathbf{d}$ for \mathbf{d} being a constant vector known *a priori*.

The first of these we can exclude on the grounds that allowing \mathbf{A} to depend on \mathbf{y} allows nonlinear estimators and the rationale for this investigation is to explore the theory of linear estimation in the context of the model (1). The second can be excluded by reformulating the model to incorporate any

estimable constraints. What then can we say if we exclude these two possibilities? We can now show that $\mathbf{AX} = \mathbf{X}$ and $\mathbf{c} = \mathbf{0}$ are necessary conditions. The case of nonsingular \mathbf{V} is straightforward; it is the case of singular \mathbf{V} that has caused some controversy (Christensen, 1990; Harville, 1990; Puntanen and Styan, 1990). We therefore proceed to that case.

The case of singular \mathbf{V} leads to an equation so temptingly close to (5) that it is easy to miss the distinction. For the sake of completeness we begin with the detail from Harville (1990).

When \mathbf{V} is singular, there always exists a matrix \mathbf{T} of maximal full row rank such that

$$\mathbf{TV} = \mathbf{0} .$$

Therefore

$$\text{var}(\mathbf{Ty}) = \mathbf{0}$$

and hence, with probability one, \mathbf{Ty} is some constant. Call that constant \mathbf{b} :

$$\mathbf{b} = \mathbf{Ty} .$$

Consequently $E(\mathbf{Ty})$, which is $\mathbf{TX}\boldsymbol{\beta}$, is \mathbf{b} . Thus

$$\mathbf{TX}\boldsymbol{\beta} = \mathbf{b} = \mathbf{Ty} \text{ with probability one .} \quad (10)$$

At first thought $\mathbf{TX}\boldsymbol{\beta} = \mathbf{b}$ of (10) seems to be the same kind of equation as $\mathbf{TX}\boldsymbol{\beta} = \mathbf{d}$ of (6), which would lead to $\mathbf{AX} = \mathbf{X}$ and $\mathbf{c} = \mathbf{0}$ not being necessary. However, there is a big difference and it is a difference which, when ignored, has (in the words of a referee) “led many to act as though condition (10) is a constraint on $\boldsymbol{\beta}$ in the same sense as is (6).” The difference is that in (6) the \mathbf{d} is a given constant, but in (10) the \mathbf{b} of $\mathbf{b} = \mathbf{TX}\boldsymbol{\beta}$ is not. This is because *a priori*, we do not know the value of \mathbf{b} and are unwilling to exclude any possible values for \mathbf{b} .

We can now prove that (3) leads to $\mathbf{AX} = \mathbf{X}$ and $\mathbf{c} = \mathbf{0}$. To do so, we assume \mathbf{A} does not depend on \mathbf{y} , and then discuss where our proof breaks down if \mathbf{A} does depend on \mathbf{y} .

Starting with (4), namely $\mathbf{AX}\boldsymbol{\beta} + \mathbf{c} = \mathbf{X}\boldsymbol{\beta}$, and solving $\mathbf{TX}\boldsymbol{\beta} = \mathbf{b}$ of (10) using the notation \mathbf{U} for \mathbf{TX} gives

$$\boldsymbol{\beta} = \mathbf{U}^{-1}\mathbf{b} + (\mathbf{I} - \mathbf{U}^{-1}\mathbf{U})\mathbf{z} \quad (11)$$

for arbitrary \mathbf{z} . Substituting this into (3) gives

$$(\mathbf{AX} - \mathbf{X})\mathbf{U}^{-1}\mathbf{b} + (\mathbf{AX} - \mathbf{X})(\mathbf{I} - \mathbf{U}^{-1}\mathbf{U})\mathbf{z} = -\mathbf{c} .$$

Because of the arbitrariness of \mathbf{z} , this yields

$$(\mathbf{AX} - \mathbf{X})\mathbf{U}^{-}\mathbf{b} = -\mathbf{c} \quad (12)$$

and

$$(\mathbf{AX} - \mathbf{X})(\mathbf{I} - \mathbf{U}^{-}\mathbf{U}) = \mathbf{0} \quad (13)$$

as in (2) of Harville (1990). Since we are unwilling to exclude any possible values of \mathbf{b} [contrast this with (6)], (12) must be satisfied, in particular, by $\mathbf{b} = \mathbf{0}$. Hence $\mathbf{c} = \mathbf{0}$ and then

$$(\mathbf{AX} - \mathbf{X})\mathbf{U}^{-} = \mathbf{0} . \quad (14)$$

Substituting (14) into (13) then gives

$$\mathbf{AX} = \mathbf{X} .$$

Thus, on assuming that \mathbf{A} does not depend on \mathbf{y} we have shown that $\mathbf{AX} = \mathbf{X}$ and $\mathbf{c} = \mathbf{0}$ are necessary conditions for unbiased estimation.

The preceding argument breaks down in two places when \mathbf{A} does depend on \mathbf{y} . First, in (3), taking $E(\mathbf{Ay} + \mathbf{c}) = \mathbf{AX}\boldsymbol{\beta} + \mathbf{c}$ is based upon using $E(\mathbf{Ay}) = \mathbf{AE}(\mathbf{y})$, which is not true when \mathbf{A} is a function of \mathbf{y} . Second, in going from (12) and (13) to (14) we are assuming that \mathbf{A} is not a function of \mathbf{b} ; and if \mathbf{A} were to be a function of \mathbf{y} it would be a function of \mathbf{b} , because $\mathbf{b} = \mathbf{T}\mathbf{y}$ of (10). Therefore if \mathbf{A} was a function of \mathbf{y} the argument leading from (12) and (13) to (14) and hence to $\mathbf{AX} = \mathbf{X}$ would not be valid. Hence, providing \mathbf{A} is not a function of \mathbf{y} , $\mathbf{AX} = \mathbf{X}$ and $\mathbf{c} = \mathbf{0}$ are necessary conditions for $E(\mathbf{Ay} + \mathbf{c})$ to be $\mathbf{X}\boldsymbol{\beta}$.

4. Sufficiency

It can be noted in passing that conditions (11) and (12) are sufficient for $E(\mathbf{Ay} + \mathbf{c}) = \mathbf{X}\boldsymbol{\beta}$ when \mathbf{A} does not depend on \mathbf{y} . This is so because

$$\begin{aligned} E(\mathbf{Ay} + \mathbf{c}) &= \mathbf{AX}\boldsymbol{\beta} + \mathbf{c} \\ &= \mathbf{AX}[\mathbf{U}^{-}\mathbf{b} + (\mathbf{I} - \mathbf{U}^{-}\mathbf{U})\mathbf{z}] + \mathbf{c} && \text{from (11)} \\ &= \mathbf{AX}\mathbf{U}^{-}\mathbf{b} + \mathbf{AX}(\mathbf{I} - \mathbf{U}^{-}\mathbf{U})\mathbf{z} + \mathbf{c} \\ &= \mathbf{X}\mathbf{U}^{-}\mathbf{b} - \mathbf{c} + \mathbf{X}(\mathbf{I} - \mathbf{U}^{-}\mathbf{U})\mathbf{z} + \mathbf{c} && \text{from (12) and (13)} \\ &= \mathbf{X}[\mathbf{U}^{-}\mathbf{b} + \mathbf{X}(\mathbf{I} - \mathbf{U}^{-}\mathbf{U})\mathbf{z}] \\ &= \mathbf{X}\boldsymbol{\beta} . \end{aligned}$$

5. Conclusion

The only way possible of having an \mathbf{A} such that $\mathbf{A}\mathbf{y} + \mathbf{c}$ is unbiased for $\mathbf{X}\boldsymbol{\beta}$ without also having $\mathbf{A}\mathbf{X} = \mathbf{X}$ and $\mathbf{c} = \mathbf{0}$ is either for \mathbf{A} to be a function of \mathbf{y} (which is distasteful and leads to nonlinear estimators), or for $\mathbf{TX}\boldsymbol{\beta}$ to be \mathbf{d} , some pre-data (given) constant, and this situation can be avoided by rewriting the model. If we exclude these two cases then the joint conditions $\mathbf{A}\mathbf{X} = \mathbf{X}$ and $\mathbf{c} = \mathbf{0}$ are necessary and sufficient for unbiasedness of $\mathbf{A}\mathbf{y} + \mathbf{c}$ for $\mathbf{X}\boldsymbol{\beta}$.

Acknowledgements

Thanks go to Ronald Christensen and David Harville for helpful comments and advice and to referees for ideas for some good improvements.

References

- Christensen, R. (1990). Comment on Puntanen and Styan (1989). *The American Statistician* **44**, 191-192.
- Harville, D. A. (1981). Unbiased and minimum-variance unbiased estimation of estimable functions for fixed linear models with arbitrary covariance structure. *Annals of Statistics* **9**, 633-637.
- Harville, D. A. (1990). Comment on Puntanen and Styan (1989). *The American Statistician* **44**, 192.
- Puntanen, S. and Styan, G. P. H. (1989). The equality of the ordinary least squares estimator and the best linear unbiased estimator. *The American Statistician* **43**, 153-161.
- Puntanen, S. and Styan, G. P. H. (1990). Reply to Christensen (1990) and Harville (1990). *The American Statistician* **44**, 192.
- Searle, S. R. (1982). *Matrix Algebra Useful for Statistics*. Wiley, New York.

Appendix

We now give a general proof that when

$$\mathbf{TX}\boldsymbol{\beta} = \mathbf{d} \tag{A1}$$

is known, *a priori*, we can always find an \mathbf{A} and \mathbf{c} such that $E(\mathbf{A}\mathbf{y} + \mathbf{c}) = \mathbf{X}\boldsymbol{\beta}$ but with $\mathbf{A}\mathbf{X} \neq \mathbf{X}$. For \mathbf{TX} of full row rank, (A1) can always be written (perhaps with some resequencing of rows and/or columns of \mathbf{TX} and $\boldsymbol{\beta}$) as

$$\mathbf{TX}_1\boldsymbol{\beta}_1 + \mathbf{TX}_2\boldsymbol{\beta}_2 = \mathbf{d} , \tag{A2}$$

where \mathbf{TX}_1 is nonsingular. Solving (A2) for $\boldsymbol{\beta}_1$ gives

$$\boldsymbol{\beta}_1 = (\mathbf{TX}_1)^{-1}(\mathbf{d} - \mathbf{TX}_2\boldsymbol{\beta}_2) . \tag{A3}$$

Using (A3) we now develop an \mathbf{A} such that

$$\mathbf{A}\mathbf{X}\boldsymbol{\beta} + \mathbf{c} = \mathbf{X}\boldsymbol{\beta} \tag{A4}$$

and show that $\mathbf{A}\mathbf{X} \neq \mathbf{X}$. To satisfy (A4) we want

$$\begin{aligned} \mathbf{A}\mathbf{X}\boldsymbol{\beta} + \mathbf{c} &= \mathbf{A}(\mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2) + \mathbf{c} \\ &= \mathbf{A}\left[\mathbf{X}_1(\mathbf{TX}_1)^{-1}(\mathbf{d} - \mathbf{TX}_2\boldsymbol{\beta}_2) + \mathbf{X}_2\boldsymbol{\beta}_2\right] + \mathbf{c} , \quad \text{from (A3)} \\ &= \mathbf{A}\left[\mathbf{I} - \mathbf{X}_1(\mathbf{TX}_1)^{-1}\mathbf{T}\right]\mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{A}\mathbf{X}_1(\mathbf{TX}_1)^{-1}\mathbf{d} + \mathbf{c} \end{aligned} \tag{A5}$$

to be equal to

$$\begin{aligned} \mathbf{X}\boldsymbol{\beta} &= \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 \\ &= \left[\mathbf{I} - \mathbf{X}_1(\mathbf{TX}_1)^{-1}\mathbf{T}\right]\mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{X}_1(\mathbf{TX}_1)^{-1}\mathbf{d} . \end{aligned} \tag{A6}$$

Writing $\mathbf{W} = \left[\mathbf{I} - \mathbf{X}_1(\mathbf{TX}_1)^{-1}\mathbf{T}\right]\mathbf{X}_2$, and noting that for (A5) and (A6) to be equal for all $\boldsymbol{\beta}_2$ gives

$$\mathbf{A}\mathbf{W} = \mathbf{W} . \tag{A7}$$

We then also have

$$\mathbf{c} + (\mathbf{A} - \mathbf{I})\mathbf{X}_1(\mathbf{TX}_1)^{-1}\mathbf{d} = \mathbf{0} . \tag{A8}$$

Solutions to (A7) and (A8) are

$$\mathbf{A} = \mathbf{W}\mathbf{W}^{-1} \tag{A9}$$

and

$$\mathbf{c} = -(\mathbf{A} - \mathbf{I})\mathbf{X}_1(\mathbf{TX}_1)^{-1}\mathbf{d} \quad (\text{A10})$$

as in Example 3.

Finally, \mathbf{A} of (A9) has the property that $\mathbf{AX} \neq \mathbf{X}$. Clearly \mathbf{AX} can only be equal to \mathbf{X} if $\mathbf{TAX} = \mathbf{TX}$. Since $\mathbf{TW} = \mathbf{0}$ we also have $\mathbf{TAX} = \mathbf{0}$. But $\mathbf{TX} \neq \mathbf{0}$ because we take \mathbf{TX}_1 nonsingular. Hence $\mathbf{AX} \neq \mathbf{X}$.