LINEAR ALGEBRA AND ITS APPLICATIONS

ELSEVIER Linear Algebra and its Applications 289 (1999) 225–242

Changes in the general linear model: A unified approach

S.R. Jammalamadaka *, D. Sengupta

Department of Statistics and Applied Probability, University of California, Santa Barbara, CA 93106, USA

Received 14 February 1997; accepted 21 September 1997

Submitted by G.P.H Styan

Abstract

Analysis of the general linear model with possibly rank-deficient design and dispersion matrices has sometimes generated some confusion and controversy, prompting some researchers to discuss it as quite distinct from the case of full-rank matrices. We show that linear zero functions, i.e., linear functions in observations which have zero expectations for all parameter values, provide an intuitive way of developing all the important results in connection with the general linear model, thus bridging this imaginary gap. We show that the effect of addition or deletion of a set of observations in this model can be clearly understood in statistical terms if viewed through such linear zero functions. The effect of adding or dropping a group of parameters is also explained well in this manner. Several sets of update equations were derived by a host of previous researchers in various special cases of the above set-up. The results derived here bring out the common underlying principles of these formulae and indeed help simplify most of them. These results also provide further insights into recursive residuals, design of experiments, deletion diagnostics and selection of subset models. © 1999 Published by Elsevier Science Inc. All rights reserved.

Keywords: Linear zero functions; Singular linear model; Update; Diagnostics; Design; Subset selection; Model selection; Recursive residuals

* Corresponding author.

PII: S0024-3795(97)10047-7

1. Introduction

Consider the linear model $(y, X\beta, \sigma^2 V)$ where the parameters β and σ^2 are unknown and the design matrix X is fixed. The statistical quantities of interest include the best linear unbiased estimators (BLUEs) of the estimable parametric functions, variance–covariance matrices of such estimators, the residual sum of squares and the likelihood ratio tests for testable linear hypotheses. In this article we are primarily concerned with the changes in these quantities when some observations are appended or deleted, as well as when some regressors are added or dropped.

Earlier work in this area include algebraic formulae in various special cases, given by Plackett (1950), Mitra and Bhimasankaram (1971), McGilchrist and Sandland (1979), Haslett (1985), Chib et al. (1987), and Bhimasankaram et al. (1995). Kourouklis and Paige (1981) gave a computational algorithm for recursive estimation which were later used in a number of statistical packages. Mitra and Bhimasankaram (1971) and Bhimasankaram and Jammalamadaka (1994) considered the addition or deletion of a *regressor* as well, – a problem not considered by most of the other authors. The work of McGilchrist and Sandland (1979) and Haslett (1985) make use of recursive residuals – a theoretical tool that has several other applications (see Kianifard and Swallow, 1996). We show in this article that all these results can be considerably generalized while at the same time providing much simpler and intuitive explanation of what is going on. Besides, these results also hold in the case of singular V.

The case of singular V is important for a number of reasons. It may arise because of certain exact linear constraints, noise-free measurements for a subset of the data, repetition of errors in a randomized experiment (see for instance Kempthorne, 1952, pp. 137,190 and Scheffé, 1959, pp. 299–301) or redundancy in a derived linear model (see Rowley, 1977; Bich, 1990). Also, the singularity of V may be seen as a limiting special case of a *nearly* rank-deficient dispersion matrix. Often such singular linear models have been treated in the literature by special (and relatively complex) methodology that was not needed in the full-rank case (cf. Christensen, 1987, pp. 179–200). Prominent approaches of this kind are the Inverse Partition Matrix method and the Unified Theory of Least Squares Estimation (cf. Rao, 1973, pp. 298–302). There has even been some controversy regarding many related issues like how to generalize the definition of linear unbiased estimators to the singular case (see Harville, 1981), whether the usual least squares theory would go through (see Rao, 1978) and whether a part of the model equation should be treated as a deterministic constraint. Some researchers have advocated separation of the 'statistical' part of the model from the 'non-statistical part' (see Feuerverger and Fraser, 1980). However, this approach makes it difficult to relate the singular case to the *almost* singular case, which is one of the objectives of studying the singular model in the first place.

We argue that the singular model does not need a special treatment. It is possible to derive virtually every result for linear models using linear zero functions and simple vector space arguments that hold for non-singular as well as singular V. In this article, we make use of this approach to derive the update equations in the general linear model. These results are not meant as computational formulae. Rather, we emphasize statistical interpretation, making the common underlying principles transparent and avoiding competing or contradictory intuitions. As it turns out, this clearer understanding of what is happening also leads, at the same time, to simplification of some of the messy algebra of the earlier formulae.

We now introduce a set of notations that will be used throughout the article. The mean and dispersion of a random vector $\boldsymbol{v}$ are denoted by $E(\boldsymbol{v})$ and $D(\boldsymbol{v})$, respectively, while $\mathrm{Cov}(\boldsymbol{v}, \boldsymbol{u})$ represents its covariance with another random vector $\boldsymbol{u}$. The variance of a scalar random variable z is denoted by $\mathrm{Var}(z)$. For a matrix $\boldsymbol{A}$, the notations $\rho(\boldsymbol{A})$, $\boldsymbol{A}'$, $\boldsymbol{A}^-$, $\mathscr{C}(\boldsymbol{A})$ and $\boldsymbol{P}_A$ represent its rank, transpose, a generalized inverse (g-inverse), its column space and the orthogonal projector onto its column space, respectively. The space orthogonal to $\mathscr{C}(\boldsymbol{A})$ is denoted by $\mathscr{C}(\boldsymbol{A})^{\perp}$. Unless mentioned otherwise, subscripts refer to the sample size of a model and subscripts within parantheses correspond to the number of parameters in a model.

Definition 1.1. The linear function $\boldsymbol{l}'\boldsymbol{y}$ in the linear model $(\boldsymbol{y}, \boldsymbol{X\beta}, \sigma^2 \boldsymbol{V})$ is called a linear zero function (LZF) if $E(\boldsymbol{l}'\boldsymbol{y}) = 0$ for all values of $\boldsymbol{\beta}$.

Algebraically, $\boldsymbol{l}'\boldsymbol{y}$ is a LZF if and only if $\boldsymbol{l} \in \mathscr{C}(\boldsymbol{X})^{\perp}$. Thus, every LZF is a linear function of $(\boldsymbol{I} - \boldsymbol{P}_X)\boldsymbol{y}$. The importance of the LZFs stems from the following well-known result.

Theorem 1.2 (Rao, 1968). *In the linear model $(\boldsymbol{y}, \boldsymbol{X\beta}, \sigma^2 \boldsymbol{V})$ with possibly singular $\boldsymbol{V}$, a linear function $\boldsymbol{l}'\boldsymbol{y}$ is the BLUE of its expectation if and only if it is uncorrelated with every LZF.*

This result is analogous in spirit to Basu's Theorem (Basu, 1959). In the special case of multivariate normal errors, it is clear that the LZFs are ancillary statistics. Therefore it is not surprising that the maximum likelihood estimator of an estimable linear function of $\boldsymbol{\beta}$, which coincides with its BLUE in this set-up, should be independent of (uncorrelated with) the LZFs.

Residuals and recursive residuals in the linear model are among the well-known examples of LZFs. Indeed, every LZF can be written as a linear combination of the residuals.

Theorem 1.2 can be used to derive an expression for the BLUE of $\boldsymbol{X\beta}$, by making use of the lemma stated below.

Lemma 1.3. *Let $z = (u' : v')'$ be a random vector having first and second moments such that $E(v) \in \mathscr{C}(D(v))$. Then the linear compound $u + Bv$ is uncorrelated with v if and only if $Bv = -\mathrm{Cov}(u, v)[D(v)]^- v$.*

Proof. It is easy to see that the above choice makes $u + Bv$ uncorrelated with v. In order to prove the necessity of the condition, one only needs to set the covariance equal to zero, postmultiply the resulting equation by $[D(v)]^- v$ and simplify. □

In the above, Bv does not depend on the choice of the g-inverse of $D(v)$. Choosing $u = y$ and $v = (I - P_X)y$, the BLUE of $X\beta$ is obtained as

$$\widehat{X\beta} = [I - V(I - P_X)\{(I - P_X)V(I - P_X)\}^-(I - P_X)]y. \tag{1}$$

The expression on the right-hand side simplifies to the usual ones when V is the identity matrix or when it is positive definite. Bhimasankaram and Sengupta (1996) show that the same Eq. (1) also leads to the other expressions obtained from the two methods described by Rao (1973), pp. 298–302). We define $\hat{\beta} = X^-\widehat{X\beta}$. It follows that $\hat{\beta}$ satisfies the 'normal equations' and that $p'\hat{\beta}$ is the BLUE of $p'\beta$ whenever the latter is estimable. In particular, $X\hat{\beta} = \widehat{X\beta}$. Searle (1994) considered the special choice $\hat{\beta} = X^+\widehat{X\beta}$, where X^+ is the Moore–Penrose pseudo-inverse (see Rao, 1973, p. 26).

From Eq. (1), the dispersion of $\widehat{X\beta}$ turns out to be

$$D(\widehat{X\beta}) = \sigma^2[V - V(I - P_X)\{(I - P_X)V(I - P_X)\}^-(I - P_X)V]. \tag{2}$$

In view of Lemma 1.3, (2) can be viewed as the residual variability in y, which is of course an unbiased estimator of $\widehat{X\beta}$, after removing the variability due to the LZFs. Again in the case of normal errors, it is equal to the conditional variance $D(y|(I - P_X)y)$. Whenever $A\beta$ is estimable, $D(A\hat{\beta}) = AX^-D(\widehat{X\beta})X^{-'}A'$. All these expressions simplify to the familiar ones when V is positive definite.

The LZFs provide an important interpretation of the R_0^2, the residual sum of squares, whether or not V is singular. Consider a 'basis' set of uncorrelated LZFs each with variance σ^2, where the basis set is such that it leaves out no nontrivial LZF that is uncorrelated with the LZFs in the chosen set. Then it can be shown that R_0^2 is the sum of squares of the LZFs of any such set. The error degrees of freedom associated with R_0^2 is the number of LZFs in the basis set, which happens to be $\rho(X : V) - \rho(X)$. A linear restriction of the form $A\beta = \xi$ (with $A\beta$ estimable) introduces additional LZFs in the form of $A\hat{\beta} - \xi$, where $A\hat{\beta}$ is the unrestricted BLUE of $A\beta$. Therefore the residual sum of squares under the linear restriction is given by

$$R_H^2 = R_0^2 + (A\hat{\beta} - \xi)'[D(A\hat{\beta})]^-(A\hat{\beta} - \xi). \tag{3}$$

Using the above facts, we derive in Section 2 a set of update formulae for $\widehat{X\beta}$, $D(\widehat{X\beta})$, R_0^2, R_H^2 and the associated degrees of freedom, when a set of observations is either added or deleted. The corresponding results for the addition or deletion of a set of regressors are given in Section 3. In Section 4 we demonstrate the usefulness of the results of Section 2 in designing a new observation in an existing model. We briefly mention other applications of the results of this paper, in Section 5.

2. Addition and deletion of observations

Let us denote the linear model with n observations by $(\boldsymbol{y}_n, \boldsymbol{X}_n\boldsymbol{\beta}, \sigma^2\boldsymbol{V}_n)$. Note that for $m < n$, each LZF in the sub-model $(\boldsymbol{y}_m, \boldsymbol{X}_m\boldsymbol{\beta}, \sigma^2\boldsymbol{V}_m)$ is also a LZF in the larger model $(\boldsymbol{y}_n, \boldsymbol{X}_n\boldsymbol{\beta}, \sigma^2\boldsymbol{V}_n)$. The number of uncorrelated LZFs exclusive to the larger model, which are all uncorrelated with the common LZFs, is $[\rho(\boldsymbol{X}_n : \boldsymbol{V}_n) - \rho(\boldsymbol{X}_n)] - [\rho(\boldsymbol{X}_m : \boldsymbol{V}_m) - \rho(\boldsymbol{X}_m)]$. The clue to the update relationships lies in the identification of these LZFs.

Let $\boldsymbol{y}_n$, $\boldsymbol{X}_n$ and $\boldsymbol{V}_n$ be paritioned as follows:

$$\boldsymbol{y}_n = \begin{pmatrix} \boldsymbol{y}_m \\ \boldsymbol{y}_l \end{pmatrix}, \quad \boldsymbol{X}_n = \begin{pmatrix} \boldsymbol{X}_m \\ \boldsymbol{X}_l \end{pmatrix}, \quad \boldsymbol{V}_n = \begin{pmatrix} \boldsymbol{V}_m & \boldsymbol{V}_{ml} \\ \boldsymbol{V}_{lm} & \boldsymbol{V}_l \end{pmatrix},$$

where $l = n - m$. If $\rho(\boldsymbol{X}_n) - \rho(\boldsymbol{X}_m) = l$, there are effectively l additional regressors in the larger model. (These do not affect the fit of $\boldsymbol{y}_m$, but help fit $\boldsymbol{y}_l$ exactly.) The parameters corresponding to these regressors have no relevance in the smaller model. It is easy to see that there is no new LZF exclusive to the augmented model. Consequently there need be no revision in the BLUE of any function that is estimable under the initial model. The dispersion of such a BLUE, as well as R_0^2 and R_H^2, would also remain unchanged, assuming that the linear function $\boldsymbol{A\beta}$ in the constraint is estimable under the initial model.

If the number of regressors exclusive to the larger model ($l_1 = \rho(\boldsymbol{X}_n) - \rho(\boldsymbol{X}_m)$) is strictly between 0 and l, we can identify l_1 rows of $\boldsymbol{X}_l$ which together with those of $\boldsymbol{X}_m$ would have the rank $\rho(\boldsymbol{X}_n)$. As before, these rows have no role to play in the updating of the BLUE of any function estimable under both the models. The remaining rows of $\boldsymbol{X}_l$ are in the row space of $\boldsymbol{X}_n$. Therefore we have only to consider the updating problem in the case $\mathscr{C}(\boldsymbol{X}_l') \subset \mathscr{C}(\boldsymbol{X}_m')$.

At this point we can also dispose off the pathological case $\rho(\boldsymbol{X}_n : \boldsymbol{V}_n) = \rho(\boldsymbol{X}_m : \boldsymbol{V}_m)$, which occurs when the error part in the new observations $(\boldsymbol{y}_l - \boldsymbol{X}_l\boldsymbol{\beta})$ is a linear function of its counterpart from the old observations, $(\boldsymbol{y}_m - \boldsymbol{X}_m\boldsymbol{\beta})$, for all $\boldsymbol{\beta}$ with probability 1. This case is not interesting, since there is no change in any statistic whatsoever.

Assuming that $\mathscr{C}(\boldsymbol{X}_l') \subset \mathscr{C}(\boldsymbol{X}_m')$ and $l_* = \rho(\boldsymbol{X}_n : \boldsymbol{V}_n) - \rho(\boldsymbol{X}_m : \boldsymbol{V}_m) > 0$, there is a set of new LZFs in the augmented model which is identified through the following lemma.

Lemma 2.1. *A vector of LZFs of the model* $(\boldsymbol{y}_n, \boldsymbol{X}_n\boldsymbol{\beta}, \sigma^2\boldsymbol{V}_n)$ *that is uncorrelated with all the existing LZFs of* $(\boldsymbol{y}_m, \boldsymbol{X}_m\boldsymbol{\beta}, \sigma^2\boldsymbol{V}_m)$ *is given by*

$$\boldsymbol{w}_l = \boldsymbol{y}_l - \boldsymbol{X}_l\hat{\boldsymbol{\beta}}_m - \boldsymbol{V}'_{ml}\boldsymbol{V}_m^-(\boldsymbol{y}_m - \boldsymbol{X}_m\hat{\boldsymbol{\beta}}_m). \tag{4}$$

Further, all LZFs of the larger model are linear combinations of $\boldsymbol{w}_l$ *and the LZFs of the smaller model.*

Proof. It is easy to see that $\boldsymbol{y}_l - \boldsymbol{X}_l\hat{\boldsymbol{\beta}}_m$ is indeed a LZF in the augmented model. The expression for $\boldsymbol{w}_l$ is obtained by making it uncorrelated with $(\boldsymbol{I}_m - \boldsymbol{P}_{X_m})\boldsymbol{y}_m$ as per Lemma 1.3, and simplifying it.

Let $\boldsymbol{u}'(\boldsymbol{I} - \boldsymbol{P}_{X_n})\boldsymbol{y}_n$ be a LZF in the larger model that is uncorrelated with $\boldsymbol{w}_l$ and the LZFs of the smaller model. Consequently it is uncorrelated with $(\boldsymbol{y}_l - \boldsymbol{X}_l\hat{\boldsymbol{\beta}}_m)$ and $(\boldsymbol{I} - \boldsymbol{P}_{X_m})\boldsymbol{y}_m$. Therefore

$$(\boldsymbol{I} - \boldsymbol{P}_{X_m})(\boldsymbol{V}_m : \boldsymbol{V}_{ml})(\boldsymbol{I} - \boldsymbol{P}_{X_n})\boldsymbol{u} = 0,$$
$$(\boldsymbol{V}_{lm} : \boldsymbol{V}_l)(\boldsymbol{I} - \boldsymbol{P}_{X_n})\boldsymbol{r} - \boldsymbol{X}_l\boldsymbol{X}_m^-(\boldsymbol{V}_m : \boldsymbol{V}_{ml})(\boldsymbol{I} - \boldsymbol{P}_{X_n})\boldsymbol{u} = 0.$$

The first condition is equivalent to $(\boldsymbol{V}_m : \boldsymbol{V}_{ml})(\boldsymbol{I} - \boldsymbol{P}_{X_n})\boldsymbol{u} \in \mathscr{C}(\boldsymbol{X}_m)$. It follows that

$$\begin{pmatrix} \boldsymbol{X}_m \\ \boldsymbol{X}_l \end{pmatrix}\boldsymbol{X}_m^-(\boldsymbol{V}_m : \boldsymbol{V}_{ml})(\boldsymbol{I} - \boldsymbol{P}_{X_n})\boldsymbol{u} = \begin{pmatrix} \boldsymbol{V}_m & \boldsymbol{V}_{ml} \\ \boldsymbol{V}_{lm} & \boldsymbol{V}_l \end{pmatrix}(\boldsymbol{I} - \boldsymbol{P}_{X_n})\boldsymbol{u},$$

that is, $\boldsymbol{V}(\boldsymbol{I} - \boldsymbol{P}_{X_n})\boldsymbol{u} \in \mathscr{C}(\boldsymbol{X}_n)$. This implies that $\boldsymbol{u}'(\boldsymbol{I} - \boldsymbol{P}_{X_n})\boldsymbol{y}_n$ is a trivial LZF with zero variance. □

Remark 2.2. Since all the LZFs of the larger model that are uncorrelated with those of the smaller model, are linear functions of $\boldsymbol{w}_l$, the rank of $D(\boldsymbol{w}_l)$ must be l_*.

Remark 2.3. There is no unique choice of the LZF with the properties stated in Lemma 2.1. Any linear function of $\boldsymbol{w}_l$ that has the same rank of the dispersion matrix would suffice. However, the expression in Eq. (4) is invariant under the choice of the g-inverse of $\boldsymbol{V}_m$ (see Eq. (1)).

Remark 2.4. Suppose $\boldsymbol{F}\boldsymbol{F}'$ is a rank-factorization of $D(\boldsymbol{w}_l)$, and $\boldsymbol{F}^{-\mathrm{L}}$ is a left-inverse of $\boldsymbol{F}$. Then the LZF, $\boldsymbol{F}^{-\mathrm{L}}\boldsymbol{w}_l$ can be defined as a *recursive group residual* for the observation vector $\boldsymbol{y}_l$. Several special cases of this can be found in the literature. Brown et al. (1975) and McGilchrist and Sandland (1979) considered homoscedastic $\boldsymbol{V}$ and positive definite $\boldsymbol{V}$, respectively, with $l = 1$. Haslett (1985) assumed $\boldsymbol{V}$ to be positive definite and $l \geqslant 1$. In the general case, the recursive group residual is not uniquely defined whenever $D(\boldsymbol{w}_l)$ is a singular matrix. However, $\boldsymbol{w}_l$ is uniquely defined given the order of inclusion of the observations. The components of $\boldsymbol{w}_l$ also have one-to-one correspondence with those of $\boldsymbol{y}_l$.

Remark 2.5. Let $d_l(\beta) = y_l - X_l\beta - V_{lm}V_m^-(y_m - X_m\beta)$, the part of the model error of y_l that is uncorrelated with the model error of y_m. The LZF w_l can be seen as $d_l(\hat{\beta}_m)$, the prediction of $d_l(\beta)$ based on the first m observations. The implications of this identification will be clear later in this section.

We now provide the update equations for data augmentation through the following theorem.

Theorem 2.6. *Under the set-up mentioned above, let $\mathscr{C}(X_l') \subset \mathscr{C}(X_m')$ and let $l_* = \rho(X_n : V_n) - \rho(X_m : V_m) > 0$. Suppose further that $A\beta$ is estimable with $D \times (A\hat{\beta}_m)$ not identically zero. Then*

(i) $X_m\hat{\beta}_n = X_m\hat{\beta}_m - \mathrm{Cov}(X_m\hat{\beta}_m, w_l)[D(w_l)]^- w_l$.

(ii) $D(X_m\hat{\beta}_n) = D(X_m\hat{\beta}_m) - \mathrm{Cov}(X_m\hat{\beta}_m, w_l)[D(w_l)]^- \mathrm{Cov}(X_m\hat{\beta}_m, w_l)'$.

(iii) $R_{0_n}^2 = R_{0_m}^2 + \sigma^2 w_l'[D(w_l)]^- w_l$.

(iv) $R_{H_n}^2 = R_{H_m}^2 + \sigma^2 w_{l*}'[D(w_{l*})]^- w_{l*}$, *where* $w_{l*} = w_l - \mathrm{Cov}(w_l, A\hat{\beta}_m)[D(A\hat{\beta}_m)]^- (A\hat{\beta}_m - \xi)$.

(v) *The degrees of freedom of R_0^2 and R_H^2 increase by l_* and $\rho(D(w_{l*}))$, respectively.*

Proof. Note that $X_m\hat{\beta}_m$ is an unbiased estimator of $X_m\beta$ that is already uncorrelated with the LZFs of the model of sample size m. Part (i) is proved by making it uncorrelated with the new LZFs w_l through Lemma 1.3.

Part (ii) can be derived by noting that $X_m\hat{\beta}_n$ must be uncorrelated with the increment term in part (i).

Part (iii) follows from the characterization of R_0^2 through a basis set of linear zero functions (see Section 1).

Substitution of these three update formulae into (3) leads to (iv) after some algebraic manipulation. This result can also be proved directly by showing that w_{l*} is the recursive group residual of y_l modified for the restricted model.

Part (v) is a consequence of the fact that the additional error degrees of freedom coincide with the number of nontrivial LZFs of the augmented model that are uncorrelated with the old ones as well as among themselves. □

Remark 2.7. McGilchrist and Sandland (1979) and Haslett (1985) had used the recursive residual for their update formula for R_0^2 in the case of positive definite V_n. The present derivation has a more intuitive appeal and it leads to simple expressions for the other updates as well.

Remark 2.8. Parts (i)–(iii) with $l = 1$, generalize the results of Bhimasankaram and Jammalamadaka (1994) to the case of multiple observations and possibly singular dispersion matrix. The expression in part (iv) is simpler to understand than theirs.

Remark 2.9. The variance and covariances used in the above expressions can be evaluated in a number of ways. Bhimasankaram et al. (1995) used the Inverse Partition Matrix method of Rao (1973), p. 298 for the updates in the singular dispersion case for $l = 1$. Theorem 2.6 helps interpret their algebraic formulae.

Remark 2.10. We reiterate that the results of Theorem 2.6 are useful mainly for the purposes of statistical interpretation and understanding, and should not be treated as a set of computational formulae. There is a vast literature on numerically stable methods of recursive estimation in the linear model, see for instance Gragg et al. (1979), Kourouklis and Paige (1981) and Farebrother (1988).

The vector of additional LZFs ($\boldsymbol{w}_l$) serves as the key to the updates for data augmentation. However, it is not very useful to obtain the update formulae for data deletion, since it is not readily computable from the current model $(\boldsymbol{y}_n, \boldsymbol{X}_n\boldsymbol{\beta}_n, \sigma^2\boldsymbol{V}_n)$. The following lemma provides a transformation of $\boldsymbol{w}_l$ that is useful in the present context. In the following, $\boldsymbol{d}_l(\cdot)$ is as defined in Remark 2.5.

Lemma 2.11. *The conditions of Lemma* 2.1 *are satisfied by* $\boldsymbol{r}_l = \boldsymbol{d}_l(\hat{\boldsymbol{\beta}}_n)$.

Proof. It is clear that

$$\begin{aligned}\boldsymbol{r}_l &= \boldsymbol{w}_l + \boldsymbol{d}_l(\hat{\boldsymbol{\beta}}_n) - \boldsymbol{d}_l(\hat{\boldsymbol{\beta}}_m) = \boldsymbol{w}_l + (-\boldsymbol{V}_{lm}\boldsymbol{V}_m^- : \boldsymbol{I})(\boldsymbol{X}_n\hat{\boldsymbol{\beta}}_m - \boldsymbol{X}_n\hat{\boldsymbol{\beta}}_n)\\ &= \boldsymbol{w}_l + (-\boldsymbol{V}_{lm}\boldsymbol{V}_m^- : \boldsymbol{I})\mathrm{Cov}(\boldsymbol{X}_n\hat{\boldsymbol{\beta}}_m, \boldsymbol{w}_l)[D(\boldsymbol{w}_l)]^- \boldsymbol{w}_l,\end{aligned}$$

by making use of part (i) of Theorem 2.6. Being a linear function of $\boldsymbol{w}_l$, $\boldsymbol{r}_l$ must be a vector of LZFs of the larger model that is uncorrelated with those of the smaller model. It remains to be shown that there is no other LZF of the larger model having this property. Let us suppose, for contradiction, that e is such a LZF. By virtue of Lemma 2.1, e must be of the form $\boldsymbol{g}'\boldsymbol{w}_l$ for some vector $\boldsymbol{g}$. It follows from the above decomposition of $\boldsymbol{r}_l$ that

$$\begin{aligned}\mathrm{Cov}(\boldsymbol{r}_l, e) &= \mathrm{Cov}(\boldsymbol{w}_l, e) + (-\boldsymbol{V}_{lm}\boldsymbol{V}_m^- : \boldsymbol{I})\mathrm{Cov}(\boldsymbol{X}_n\hat{\boldsymbol{\beta}}_m, \boldsymbol{w}_l'\boldsymbol{g})\\ &= \mathrm{Cov}(\boldsymbol{d}_l(\boldsymbol{\beta}), \boldsymbol{g}'\boldsymbol{d}_l(\hat{\boldsymbol{\beta}}_m))\\ &= D(\boldsymbol{d}_l(\boldsymbol{\beta}))\boldsymbol{g} - \mathrm{Cov}\Big(\boldsymbol{d}_l(\boldsymbol{\beta}), \boldsymbol{g}'(\boldsymbol{d}_l(\boldsymbol{\beta}) - \boldsymbol{d}_l(\hat{\boldsymbol{\beta}}_m))\Big)\\ &= D(\boldsymbol{d}_l(\boldsymbol{\beta}))\boldsymbol{g}.\end{aligned}$$

The last simplification is possible because $\boldsymbol{d}_l(\boldsymbol{\beta})$ is uncorrelated with $\boldsymbol{y}_m$. If e is uncorrelated with $\boldsymbol{r}_l$, $\boldsymbol{g}'\boldsymbol{d}_l(\boldsymbol{\beta})$ must be identically zero. Therefore $e = \boldsymbol{g}'\boldsymbol{d}_l(\boldsymbol{\beta})$ must be a trivial LZF which is zero with probability 1. □

The advantage of $\boldsymbol{r}_l$ over $\boldsymbol{w}_l$ is that the former is expressed in terms of the estimator in the current model. In the light of Remark 2.3, Lemma 2.11 implies that

$$\boldsymbol{X}_m\hat{\boldsymbol{\beta}}_m = \boldsymbol{X}_m\hat{\boldsymbol{\beta}}_n + \mathrm{Cov}(\boldsymbol{X}_m\hat{\boldsymbol{\beta}}_m, \boldsymbol{r}_l)[D(\boldsymbol{r}_l)]^-\boldsymbol{r}_l. \tag{5}$$

The covariance on the right hand side remains to be expressed in terms of the known quantities in the current model. Indeed, from Eq. (5) it follows that

$$\mathrm{Cov}(\boldsymbol{X}_m\hat{\boldsymbol{\beta}}_m, \boldsymbol{d}_l(\boldsymbol{\beta})) = \mathrm{Cov}(\boldsymbol{X}_m\hat{\boldsymbol{\beta}}_n, \boldsymbol{d}_l(\boldsymbol{\beta}))\mathrm{Cov}(\boldsymbol{X}_m\hat{\boldsymbol{\beta}}_m, \boldsymbol{r}_l)[D(\boldsymbol{r}_l)]^-\mathrm{Cov}(\boldsymbol{r}_l, \boldsymbol{d}_l(\boldsymbol{\beta})).$$

Since $\boldsymbol{d}_l(\boldsymbol{\beta})$ is uncorrelated with $\boldsymbol{y}_m$ while $\boldsymbol{X}_m\hat{\boldsymbol{\beta}}_m$ is a linear function of it, the left hand side is zero. On the other hand, $\mathrm{Cov}(\boldsymbol{r}_l, \boldsymbol{d}_l(\boldsymbol{\beta})) - D(\boldsymbol{r}_l)$ is the covariance of $\boldsymbol{r}_l$ with a BLUE which must be zero. Therefore the second term in the right hand side reduces to $\mathrm{Cov}(\boldsymbol{X}_m\hat{\boldsymbol{\beta}}_m, \boldsymbol{r}_l)$, which can be replaced by $-\mathrm{Cov}(\boldsymbol{X}_m\hat{\boldsymbol{\beta}}_n, \boldsymbol{d}_l(\boldsymbol{\beta}))$ in Eq. (5). This leads to the update relationships given below.

Theorem 2.12. *Let $\mathscr{C}(\boldsymbol{X}_l') \subset \mathscr{C}(\boldsymbol{X}_m')$, $l_* = \rho(\boldsymbol{X}_n : \boldsymbol{V}_n) - \rho(\boldsymbol{X}_m : \boldsymbol{V}_m) > 0$ and $\boldsymbol{A\beta}$ be estimable in either model with $D(\boldsymbol{A}\hat{\boldsymbol{\beta}}_n)$ not identically zero. Then the updated statistics for the smaller model are as follows:*

(i) $\boldsymbol{X}_m\hat{\boldsymbol{\beta}}_m = \boldsymbol{X}_m\hat{\boldsymbol{\beta}}_n - \mathrm{Cov}(\boldsymbol{X}_m\hat{\boldsymbol{\beta}}_n, \boldsymbol{d}_l(\boldsymbol{\beta}))[D(\boldsymbol{r}_l)]^-\boldsymbol{r}_l$.

(ii) $D(\boldsymbol{X}_m\hat{\boldsymbol{\beta}}_m) = D(\boldsymbol{X}_m\hat{\boldsymbol{\beta}}_n) + \mathrm{Cov}(\boldsymbol{X}_m\hat{\boldsymbol{\beta}}_n, \boldsymbol{d}_l(\boldsymbol{\beta}))[D(\boldsymbol{r}_l)]^-\mathrm{Cov}(\boldsymbol{X}_m\hat{\boldsymbol{\beta}}_n, \boldsymbol{d}_l(\boldsymbol{\beta}))'$.

(iii) $R^2_{0_m} = R^2_{0_n} - \sigma^2\boldsymbol{r}_l'[D(\boldsymbol{r}_l)]^-\boldsymbol{r}_l$.

(iv) $R^2_{H_m} = R^2_{H_n} - \sigma^2\boldsymbol{r}_{l*}'[D(\boldsymbol{r}_{l*})]^-\boldsymbol{r}_{l*}$, *where* $\boldsymbol{r}_{l*} = \boldsymbol{r}_l + \mathrm{Cov}(\boldsymbol{d}_l(\boldsymbol{\beta}), \boldsymbol{A}\hat{\boldsymbol{\beta}}_n)[D(\boldsymbol{A}\hat{\boldsymbol{\beta}}_n)]^-$ $(\boldsymbol{A}\hat{\boldsymbol{\beta}}_n - \boldsymbol{\xi})$.

(v) *The degrees of freedom of R_0^2 and R_H^2 decrease by l_* and $\rho(D(\boldsymbol{r}_{l*}))$, respectively.*

Proof. Parts (i)–(iii) and (v) follow immediately from the above discussion and Theorem 2.6. Part (iv) is proved by substituting the update formulae of parts (i) and (ii) into Eq. (3) and simplifying. □

3. Addition and deletion of regressors

With the number of observations fixed, one is often confronted with the task of comparing two models where the regressors in one is a subset of the regressors in the other. In this section we deal with the connection between the models $(\boldsymbol{y}, \boldsymbol{X}_{(k)}\boldsymbol{\beta}_{(k)}, \sigma^2\boldsymbol{V})$ and $(\boldsymbol{y}, \boldsymbol{X}_{(h)}\boldsymbol{\beta}_{(h)}, \sigma^2\boldsymbol{V})(k > h)$, where the subscript within parentheses represents the number of *regressors* in the model, and

$$\boldsymbol{X}_{(k)} = (\boldsymbol{X}_{(h)} : \boldsymbol{X}_{(j)}), \quad \boldsymbol{\beta}_{(k)} = \begin{pmatrix} \boldsymbol{\beta}_{(h)} \\ \boldsymbol{\beta}_{(j)} \end{pmatrix}.$$

For the consistency of the smaller model with the data, $(\boldsymbol{I}-\boldsymbol{P}_V)\boldsymbol{y}$ must belong to $\mathscr{C}((\boldsymbol{I}-\boldsymbol{P}_V)\boldsymbol{X}_{(h)})$. We assume that this condition holds. It follows that the data is consistent with the larger model as well.

Notice that every LZF in the larger model is a LZF in the smaller model. The number of LZFs exclusive to the smaller model is $j_* = \rho(\boldsymbol{X}_{(k)} : \boldsymbol{V}) - \rho(\boldsymbol{X}_{(h)} : \boldsymbol{V}) - \rho(\boldsymbol{X}_{(k)}) + \rho(\boldsymbol{X}_{(h)})$. It is clear that $0 \leqslant j_* \leqslant \rho(\boldsymbol{X}_{(j)})$.

Suppose $\boldsymbol{x}$ is a regressor exclusive to the larger model which is not in $\mathscr{C}(\boldsymbol{X}_{(h)} : \boldsymbol{V})$. Then $\boldsymbol{l} = (\boldsymbol{I}-\boldsymbol{P}_{X_{(k-1)}:V})\boldsymbol{x}_{(k)}$ must be a nontrivial vector. Consistency of the smaller model dictates that $\boldsymbol{l}'\boldsymbol{y} = 0$ with probability 1, while that of the larger model requires $\boldsymbol{l}'\boldsymbol{y} = (\boldsymbol{l}'\boldsymbol{x})\beta$, where β is the coefficient of $\boldsymbol{x}$ in the larger model. These two conditions can not hold simultaneously unless β is identically zero, that is, $\boldsymbol{x}$ is *useless* as a regressor. We now assume that there is no such regressor in the larger model, that is, $\rho(\boldsymbol{X}_{(k)} : \boldsymbol{V}) = \rho(\boldsymbol{X}_{(h)} : \boldsymbol{V})$. Consequently $j_* = \rho(\boldsymbol{X}_{(k)}) - \rho(\boldsymbol{X}_{(h)})$. If $j_* = 0$, the regressors exclusive to the larger model are *redundant* in the presence of the other regressors, so that the statistics under the two models are identical. The case of real interest is when $0 < j_* \leqslant \rho(\boldsymbol{X}_{(j)})$.

Consider first the problem of estimability. Notice that the only functions of $\boldsymbol{\beta}_{(h)}$ that are estimable under the larger model are linear combinations of $(\boldsymbol{I}-\boldsymbol{P}_{X_{(j)}})\boldsymbol{X}_{(h)}\boldsymbol{\beta}_{(h)}$. On the other hand, the estimable functions in the smaller model are linear combinations of $\boldsymbol{X}_{(h)}\boldsymbol{\beta}_{(h)}$. Therefore the estimable functions of $\boldsymbol{\beta}_{(h)}$ in the larger model are estimable under the smaller model, but the converse is not true in general. The rank of $(\boldsymbol{I}-\boldsymbol{P}_{X_{(j)}})\boldsymbol{X}_{(h)}$ is j_*. Therefore a necessary and sufficient condition for all the estimable functions in the *smaller* model to be estimable under the larger model is that $j_* = \rho(\boldsymbol{X}_{(j)})$. In such a case $\boldsymbol{X}_{(h)}\boldsymbol{\beta}_{(h)}$ and $\boldsymbol{X}_{(j)}\boldsymbol{\beta}_{(j)}$ are estimable under the larger model.

Even if $0 < j_* < \rho(\boldsymbol{X}_{(j)})$, there are some functions of $\boldsymbol{\beta}_{(h)}$ that are estimable under both the models. We now proceed to obtain the update of the BLUE of such a function when the last j regressors are dropped from the larger model. In order to distinguish between the least squares estimators under the two models, we use a 'tilde' for the estimators under the smaller model and the usual 'hat' for those under the larger model.

The condition $j_* = \rho(\boldsymbol{X}_{(k)}) - \rho(\boldsymbol{X}_{(h)})$ implies that there are j_* uncorrelated LZFs (subject to an ambiguity in scale) in the smaller model that are uncorrelated with all the LZFs in the larger model. A LZF with this property must have been a BLUE in the larger model. The following lemma provides an adequate set of such linear functions.

Lemma 3.1. *The linear function* $\boldsymbol{v} = (\boldsymbol{I}-\boldsymbol{P}_{X_{(h)}})\boldsymbol{X}_{(j)}\hat{\boldsymbol{\beta}}_{(j)}$, *is a vector of BLUEs in the model* $(\boldsymbol{y}, \boldsymbol{X}_{(k)}\boldsymbol{\beta}_{(k)}, \sigma^2\boldsymbol{V})$ *and a vector of LZFs in the model* $(\boldsymbol{y}, \boldsymbol{X}_{(h)}\boldsymbol{\beta}_{(h)}, \sigma^2\boldsymbol{V})$. *Further,* $\rho(D(\boldsymbol{v})) = j_*$.

Proof. The parametric function $(\boldsymbol{I}-\boldsymbol{P}_{X_{(h)}})\boldsymbol{X}_{(k)}\boldsymbol{\beta}_{(k)}$ is estimable in the larger model. The BLUE of this function is $\boldsymbol{v}$. It is easy to see that $E(\boldsymbol{v}) = 0$ under the smaller model. Since the column space of $D(\boldsymbol{X}_{(k)}\hat{\boldsymbol{\beta}}_{(k)})$ is $\mathscr{C}(\boldsymbol{X}_{(k)}) \cap \mathscr{C}(\boldsymbol{V})$ (see Bhimasankaram and Sengupta, 1996), that of $D(\boldsymbol{v})$ must be $\mathscr{C}((\boldsymbol{I}-\boldsymbol{P}_{X_{(h)}})\boldsymbol{X}_{(k)}) \cap \mathscr{C}((\boldsymbol{I}-\boldsymbol{P}_{X_{(h)}})\boldsymbol{V})$ or simply $\mathscr{C}((\boldsymbol{I}-\boldsymbol{P}_{X_{(h)}})\boldsymbol{X}_{(k)})$. Consequently $\rho(D(\boldsymbol{v})) = \rho((\boldsymbol{I}-\boldsymbol{P}_{X_{(h)}})\boldsymbol{X}_{(k)}) = j_*$. □

Once a vector of LZFs with the desired property is identified, the theorem given below follows along the lines of Theorem 2.6.

Theorem 3.2. *Let $\boldsymbol{A\beta}_{(h)}$ be estimable under the larger model. Then*
(i) $\boldsymbol{A}\tilde{\boldsymbol{\beta}}_{(h)} = \boldsymbol{A}\hat{\boldsymbol{\beta}}_{(h)} - \mathrm{Cov}(\boldsymbol{A}\hat{\boldsymbol{\beta}}_{(h)}, \boldsymbol{v})[D(\boldsymbol{v})]^{-}\boldsymbol{v}$, *where* $\boldsymbol{v} = (\boldsymbol{I}-\boldsymbol{P}_{X_{(h)}})\boldsymbol{X}_{(j)}\hat{\boldsymbol{\beta}}_{(j)}$.
(ii) $D(\boldsymbol{A}\tilde{\boldsymbol{\beta}}_{(h)}) = D(\boldsymbol{A}\hat{\boldsymbol{\beta}}_{(h)}) - \mathrm{Cov}(\boldsymbol{A}\hat{\boldsymbol{\beta}}_{(h)}, \boldsymbol{v})[D(\boldsymbol{v})]^{-}\mathrm{Cov}(\boldsymbol{A}\hat{\boldsymbol{\beta}}_{(h)}, \boldsymbol{v})'$.
(iii) $R^2_{0_{(h)}} = R^2_{0_{(k)}} + \sigma^2\boldsymbol{v}'[D(\boldsymbol{v})]^{-}\boldsymbol{v}$.
(iv) $R^2_{H_{(h)}} = R^2_{H_{(k)}} + \sigma^2\boldsymbol{v}_*'[D(\boldsymbol{v}_*)]^{-}\boldsymbol{v}_*$, *where* $\boldsymbol{v}_* = \boldsymbol{v} - \mathrm{Cov}(\boldsymbol{v}, \boldsymbol{A}\hat{\boldsymbol{\beta}}_{(h)})[D(\boldsymbol{A}\hat{\boldsymbol{\beta}}_{(h)})]^{-}(\boldsymbol{A}\hat{\boldsymbol{\beta}}_{(h)} - \boldsymbol{\xi})$.
(v) *The increase in the degrees of freedom of R^2_0 and R^2_H with the deletion of the regressors are given by j_* and $\rho(D(\boldsymbol{v}_*))$, respectively.*

Remark 3.3. The vector $\boldsymbol{v}_*$ is the BLUE of $(\boldsymbol{I}-\boldsymbol{P}_{X_{(h)}})\boldsymbol{X}_{(j)}\boldsymbol{\beta}_{(j)}$ in the larger model under the restriction $\boldsymbol{A\beta}_{(h)} = \boldsymbol{\xi}$.

Remark 3.4. Depending on the special case at hand, one may use a different form of $\boldsymbol{v}$ that would have the requisite properties. For instance, if $j_* = \rho(\boldsymbol{X}_{(j)})$, it can be chosen as $\boldsymbol{X}_{(j)}\hat{\boldsymbol{\beta}}_{(j)}$. If $j_* = j$, $\boldsymbol{v}$ can be chosen as $\hat{\boldsymbol{\beta}}_{(j)}$.

Remark 3.5. If $\boldsymbol{\beta}_{(k)}$ is entirely estimable under the original model, we have

$$\tilde{\boldsymbol{\beta}}_{(h)} = \hat{\boldsymbol{\beta}}_{(h)} - \mathrm{Cov}(\hat{\boldsymbol{\beta}}_{(h)}, \hat{\boldsymbol{\beta}}_{(j)})[D(\hat{\boldsymbol{\beta}}_{(j)})]^{-}\hat{\boldsymbol{\beta}}_{(j)},$$
$$D(\tilde{\boldsymbol{\beta}}_{(h)}) = D(\hat{\boldsymbol{\beta}}_{(h)}) - \mathrm{Cov}(\hat{\boldsymbol{\beta}}_{(h)}, \hat{\boldsymbol{\beta}}_{(j)})[D(\hat{\boldsymbol{\beta}}_{(j)})]^{-}\mathrm{Cov}(\hat{\boldsymbol{\beta}}_{(h)}, \hat{\boldsymbol{\beta}}_{(j)})'$$

These updates only involve $\hat{\boldsymbol{\beta}}_{(k)}$ and its dispersion.

In the case of adding a few regressors, the role of $\boldsymbol{v}$ has to be assumed by a vector computable in terms of the statistics of the smaller model. Such a vector is presented in the following lemma.

Lemma 3.6. *A vector of LZFs in the smaller model that is also a BLUE in the larger model is*

$$\boldsymbol{t} = \boldsymbol{X}'_{(j)}(\boldsymbol{I}-\boldsymbol{P}_{X_{(h)}})\{(\boldsymbol{I}-\boldsymbol{P}_{X_{(h)}})\boldsymbol{V}(\boldsymbol{I}-\boldsymbol{P}_{X_{(h)}})\}^{-}(\boldsymbol{I}-\boldsymbol{P}_{X_{(h)}})\boldsymbol{y}. \tag{6}$$

Further, $\rho(D(\boldsymbol{t})) = j_*$.

Proof. It is clear that $\boldsymbol{t}$ is a LZF in the smaller model. Let $\boldsymbol{l}'\boldsymbol{y}$ be a LZF in the augmented model. Then $\boldsymbol{X}'_{(j)}\boldsymbol{l} = 0$ and $\boldsymbol{X}'_{(h)}\boldsymbol{l} = 0$. Writing $\boldsymbol{l}$ as $(\boldsymbol{I}-\boldsymbol{P}_{X_{(h)}})\boldsymbol{s}$, we have by virtue of Eq. (1)

$$\begin{aligned}\mathrm{Cov}(\boldsymbol{t}, \boldsymbol{l}'\boldsymbol{y}) &= \sigma^2 \boldsymbol{X}'_{(j)}(\boldsymbol{I} - \boldsymbol{P}_{X_{(h)}})\{(\boldsymbol{I} - \boldsymbol{P}_{X_{(h)}})\boldsymbol{V}(\boldsymbol{I} - \boldsymbol{P}_{X_{(h)}})\}^- \\ &\qquad \cdot (\boldsymbol{I} - \boldsymbol{P}_{X_{(h)}})\boldsymbol{V}(\boldsymbol{I} - \boldsymbol{P}_{X_{(h)}})\boldsymbol{s} \\ &= \sigma^2 \boldsymbol{X}'_{(j)}(\boldsymbol{I} - \boldsymbol{P}_{X_{(h)}})\boldsymbol{s} = \sigma^2 \boldsymbol{X}'_{(j)}\boldsymbol{l} = 0\end{aligned}$$

In the above we have used the fact that $\mathscr{C}((\boldsymbol{I} - \boldsymbol{P}_{X_{(h)}})\boldsymbol{X}_{(j)})$ is a subset of $\mathscr{C}((\boldsymbol{I} - \boldsymbol{P}_{X_{(h)}})\boldsymbol{V})$ (identical to $\mathscr{C}((\boldsymbol{I} - \boldsymbol{P}_{X_{(h)}})\boldsymbol{V}(\boldsymbol{I} - \boldsymbol{P}_{X_{(h)}})))$, which follows from the assumption $\boldsymbol{X}_{(j)} \in \mathscr{C}(\boldsymbol{X}_{(h)} : \boldsymbol{V})$. Being uncorrelated with all LZFs in the larger model, $\boldsymbol{t}$ must be BLUE there. The rank condition follows from the fact that $\mathscr{C}(D(\boldsymbol{y} - \boldsymbol{X}_{(h)}\hat{\boldsymbol{\beta}}_{(h)})) = \mathscr{C}(\boldsymbol{V}(\boldsymbol{I} - \boldsymbol{P}_{X_{(h)}}))$ (see Bhimasankaram and Sengupta, 1996), which implies $\mathscr{C}(D(\boldsymbol{t})) = \mathscr{C}(\boldsymbol{X}'_{(j)}(\boldsymbol{I} - \boldsymbol{P}_{X_{(h)}}))$. □

Remark 3.7. From the statements of Lemmas 3.1 and 3.6 it is evident that the random vectors $\boldsymbol{v}$ and $\boldsymbol{t}$ must be related. In fact they are functions of one another:

$$\begin{aligned}\boldsymbol{v} &= \sigma^2(\boldsymbol{I} - \boldsymbol{P}_{X_{(h)}})\boldsymbol{X}_{(j)}[D(\boldsymbol{t})]^- \boldsymbol{t}, \\ \boldsymbol{t} &= \boldsymbol{X}'_{(j)}(\boldsymbol{I} - \boldsymbol{P}_{X_{(h)}})\{(\boldsymbol{I} - \boldsymbol{P}_{X_{(h)}})\boldsymbol{V}(\boldsymbol{I} - \boldsymbol{P}_{X_{(h)}})\}^- \boldsymbol{v}.\end{aligned}$$

Remark 3.8. Recall that $\mathscr{C}(\boldsymbol{X}_{(j)})$ is assumed to be a subset of $\mathscr{C}(\boldsymbol{X}_{(h)} : \boldsymbol{V})$. If $\boldsymbol{X}_{(j)} = \boldsymbol{X}_{(h)}\boldsymbol{B} + \boldsymbol{V}\boldsymbol{C}$, then $\boldsymbol{t}$ is the same as $\boldsymbol{C}'\boldsymbol{y}_{\text{res}}$, where $\boldsymbol{y}_{\text{res}}$ is the residual of $\boldsymbol{y}$ from the smaller model. (Specifically, $\boldsymbol{y}_{\text{res}} = \boldsymbol{R}\boldsymbol{y}$ where $\boldsymbol{R} = \boldsymbol{V}(\boldsymbol{I} - \boldsymbol{P}_{X_{(h)}})\{(\boldsymbol{I} - \boldsymbol{P}_{X_{(h)}})\boldsymbol{V}(\boldsymbol{I} - \boldsymbol{P}_{X_{(h)}})\}^-(\boldsymbol{I} - \boldsymbol{P}_{X_{(h)}})$, as seen from Eq. (1).) It can also be interpreted as $\boldsymbol{X}'_{(j)_{\text{res}}}\boldsymbol{V}^-\boldsymbol{y}_{\text{res}}$ where $\boldsymbol{X}_{(j)_{\text{res}}} = \boldsymbol{R}\boldsymbol{X}_{(j)}$, the 'residual' of $\boldsymbol{X}_{(j)}$ when regressed (one column at a time) on $\boldsymbol{X}_{(h)}$. Similarly, $D(\boldsymbol{t})$ is the same as $\sigma^2\boldsymbol{X}'_{(j)_{\text{res}}}\boldsymbol{V}^-\boldsymbol{X}_{(j)_{\text{res}}}$.

Remark 3.9. The expectations of $\boldsymbol{v}$ and $\boldsymbol{t}$ are linear functions of $\boldsymbol{\beta}_{(j)}$. These linear parametric functions are estimable in the linear model $(\boldsymbol{y}_{\text{res}}, \boldsymbol{X}_{(j)_{\text{res}}}, \sigma^2\boldsymbol{W})$, where $\boldsymbol{W} = \boldsymbol{R}\boldsymbol{V}$. Moreover, $\boldsymbol{v}$ and $\boldsymbol{t}$ are BLUEs of the corresponding parametric functions in this 'residual' model, which is obtained from the original (larger) model by premultiplying both the systematic and error parts by $\boldsymbol{R}$. The algebraic calculations are simplified by the fact that the regressors here reside in the column space of the dispersion matrix.

Remark 3.10. When $j = 1$ and $\boldsymbol{V}$ is positive definite, the BLUE of the new parameter in the augmented model is proportional to the 'lost' LZF. In this special case the interpretation of the BLUE as the simple regression coefficient in a 'residual' model is quite well-known. The model dispersion matrix $\sigma^2\boldsymbol{W}$ in the residual model is sometimes replaced by $\sigma^2\boldsymbol{V}$, which makes no algebraic difference.

We now have the update relations for the larger model as follows.

Theorem 3.11. *If $A\boldsymbol{\beta}_{(h)}$ is estimable under the larger model, then*
(i) $A\hat{\boldsymbol{\beta}}_{(h)} = A\tilde{\boldsymbol{\beta}}_{(h)} + \mathrm{Cov}(AX_{(k)}^{-}\boldsymbol{y}, \boldsymbol{t})[D(\boldsymbol{t})]^{-}\boldsymbol{t}$, *where $\boldsymbol{t}$ is as in Eq.* (6).
(ii) $D(A\hat{\boldsymbol{\beta}}_{(h)}) = D(A\tilde{\boldsymbol{\beta}}_{(h)}) + \mathrm{Cov}(AX_{(k)}^{-}\boldsymbol{y}, \boldsymbol{t})[D(\boldsymbol{t})]^{-}\mathrm{Cov}(AX_{(k)}^{-}\boldsymbol{y}, \boldsymbol{t})'$.
(iii) $R^2_{0_{(k)}} = R^2_{0_{(h)}} - \sigma^2\boldsymbol{t}'[D(\boldsymbol{t})]^{-}\boldsymbol{t}$.
(iv) $R^2_{H_{(k)}} = R^2_{H_{(h)}} - \sigma^2\boldsymbol{t}_*'[D(\boldsymbol{t}_*)]^{-}\boldsymbol{t}_*$, *where* $\boldsymbol{t}_* = \boldsymbol{t} - \mathrm{Cov}(\boldsymbol{t}, A\hat{\boldsymbol{\beta}}_{(h)})[D(A\hat{\boldsymbol{\beta}}_{(h)})]^{-}(A\hat{\boldsymbol{\beta}}_{(h)} - \xi)$.
(v) *The increase in the degrees of freedom of R_0^2 and R_H^2 with the deletion of the regressors are given by j_* and $\rho(D(\boldsymbol{t}_*))$, respectively.*

Proof. Since $\boldsymbol{t}$ contains j_* uncorrelated LZFs of the current model that turn into BLUEs in the larger model,

$$A\tilde{\boldsymbol{\beta}}_{(h)} = A\hat{\boldsymbol{\beta}}_{(h)} - \mathrm{Cov}(A\hat{\boldsymbol{\beta}}_{(h)}, \boldsymbol{t})[D(\boldsymbol{t})]^{-}\boldsymbol{t}.$$

Write $A\hat{\boldsymbol{\beta}}_{(h)}$ as

$$A\hat{\boldsymbol{\beta}}_{(h)} = AX_{(k)}^{-}[X_{(k)}\hat{\boldsymbol{\beta}}_{(k)}] = AX_{(k)}^{-}\boldsymbol{y} + AX_{(k)}^{-}[\boldsymbol{y} - X_{(k)}\hat{\boldsymbol{\beta}}_{(k)}].$$

The second term is a LZF in the larger model and hence is uncorrelated with $\boldsymbol{t}$. Therefore $\mathrm{Cov}(A\hat{\boldsymbol{\beta}}_{(h)}, \boldsymbol{t}) = \mathrm{Cov}(AX_{(k)}^{-}\boldsymbol{y}, \boldsymbol{t})$. Parts (i), (ii) and (iii) follow immediately. Part (iv) is proved by substituting the results of these three parts into Eq. (3). Part (v) is easy to prove. □

Remark 3.12. The vector $AX_{(k)}^{-}\boldsymbol{y}$ depends on the choice of the generalized inverse of $X_{(k)}$, but its covariance with $\boldsymbol{t}$ does not.

Remark 3.13. The vector $\boldsymbol{t}_*$ used in parts (iv) and (v) may be expressed in terms of the statistics of the original model by using parts (i) and (ii). The expression simplifies to

$$\boldsymbol{t}_* = D(\boldsymbol{t})\left[D(\boldsymbol{t}) + \mathrm{Cov}(\boldsymbol{t}, AX_{(k)}^{-}\boldsymbol{y})[D(A\tilde{\boldsymbol{\beta}}_{(h)})]^{-}\mathrm{Cov}(\boldsymbol{t}, AX_{(k)}^{-}\boldsymbol{y})'\right]^{-} \cdot \left[\boldsymbol{t} + \mathrm{Cov}(\boldsymbol{t}, AX_{(k)}^{-}\boldsymbol{y})[D(A\tilde{\boldsymbol{\beta}}_{(h)})]^{-}(A\tilde{\boldsymbol{\beta}}_{(h)} - \xi)\right].$$

Remark 3.14. There is a striking symmetry between the updates equations of Theorems 3.2 and 3.11. Although it makes intuitive sense, this symmetry was absent from the results obtained by other researchers.

4. Application: Design of a new observation

Suppose a set of m observations has already been given, and one is interested in a particular estimable function $\boldsymbol{p}'\boldsymbol{\beta}$. Consider the problem of minimizing the variance of the BLUE of $\boldsymbol{p}'\boldsymbol{\beta}$ by choosing an additional design point optimally.

In the absence of any constraint, the variance can be reduced to zero. A reasonable constraint may be to set an upper bound on the variance of the predicted value of the additional observation, calculated on the basis of the first m observations. The simple case of homoscedastic model errors admits an intuitively meaningful solution to this problem: the new row of the design matrix should be proportional to $\boldsymbol{p}$. We now derive a solution in the general case of heteroscedastic and possibly singular error dispersion matrix, by making use of the results of Section 2.

Note that in the present context $l = 1$. In order to simplify the notations, we denote $\boldsymbol{X}_m, \boldsymbol{X}_l, \boldsymbol{y}_m, \boldsymbol{y}_l, \boldsymbol{V}_m, \boldsymbol{V}_{ml}, \boldsymbol{V}_l, \boldsymbol{w}_l$ and $\boldsymbol{d}_l(\cdot)$ by $\boldsymbol{X}, \boldsymbol{x}', \boldsymbol{y}, y, \boldsymbol{V}, \boldsymbol{v}, v, w$ and $d(\cdot)$, respectively. The task is to minimize the variance of $\boldsymbol{p}'\hat{\boldsymbol{\beta}}_{m+1}$ with respect to $\boldsymbol{x}$, subject to the constraint $\mathrm{Var}(\boldsymbol{x}'\hat{\boldsymbol{\beta}}_m) \leqslant \alpha\sigma^2$ where α is a known positive number.

It is clear that the new design point $\boldsymbol{x}$ carries no information about $\boldsymbol{p}'\boldsymbol{\beta}$ if it is not in $\mathscr{C}(\boldsymbol{X}')$. Therefore $\boldsymbol{x}$ has to be of the form $\boldsymbol{X}'\boldsymbol{u}$. In such a case, choosing $\boldsymbol{x}$ is equivalent to choosing $\boldsymbol{u}$. It was shown in Section 2 that whenever $\boldsymbol{x} \in \mathscr{C}(\boldsymbol{X}')$, there must be an additional LZF with nonzero variance, unless the new observation error is perfectly correlated with the first m errors of the model. The latter case is not interesting, since the $\mathrm{Var}(\boldsymbol{x}'\hat{\boldsymbol{\beta}}_{m+1})$ happens to be the same as $\mathrm{Var}(\boldsymbol{x}'\hat{\boldsymbol{\beta}}_m)$. In the following discussion we assume that $\boldsymbol{x} = \boldsymbol{X}'\boldsymbol{u}$ and $\mathrm{Var}(w) > 0$.

In view of Part (ii) of Theorem 2.6, minimizing $\mathrm{Var}(\boldsymbol{p}'\hat{\boldsymbol{\beta}}_{m+1})$ is equivalent to maximizing $[\mathrm{Cov}(\boldsymbol{p}'\hat{\boldsymbol{\beta}}_m, w)]^2/\mathrm{Var}(w)$. Writing w as $d(\boldsymbol{\beta}) + [d(\hat{\boldsymbol{\beta}}_m) - d(\boldsymbol{\beta})]$, a sum of uncorrelated parts (see Remark 2.5), it follows that

$$
\begin{aligned}
\mathrm{Cov}(\boldsymbol{p}'\hat{\boldsymbol{\beta}}_m, w) &= \mathrm{Cov}(\boldsymbol{p}'\hat{\boldsymbol{\beta}}_m, d(\hat{\boldsymbol{\beta}}_m) - d(\boldsymbol{\beta})),\\
\mathrm{Var}(w) &= \mathrm{Var}(d(\boldsymbol{\beta})) + \mathrm{Var}(d(\hat{\boldsymbol{\beta}}_m) - d(\boldsymbol{\beta})).
\end{aligned}
$$

Let $\theta = \mathrm{Var}(d(\boldsymbol{\beta}))$ and the vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ satisfy $\boldsymbol{p} = \boldsymbol{X}'\boldsymbol{a}$ and $\boldsymbol{v} = \boldsymbol{V}\boldsymbol{b}$, respectively. Then $d(\hat{\boldsymbol{\beta}}_m) - d(\boldsymbol{\beta}) = (\boldsymbol{b} - \boldsymbol{u})'\boldsymbol{X}(\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta})$. Denoting $D(\boldsymbol{X}_m\hat{\boldsymbol{\beta}}_m)$ by $\boldsymbol{S}\boldsymbol{S}'$, we have

$$
\begin{aligned}
\mathrm{Cov}(\boldsymbol{p}'\hat{\boldsymbol{\beta}}_m, w) &= -\boldsymbol{a}'\boldsymbol{S}\boldsymbol{S}'(\boldsymbol{u} - \boldsymbol{b}),\\
\mathrm{Var}(w) &= \theta + (\boldsymbol{u} - \boldsymbol{b})'\boldsymbol{S}\boldsymbol{S}'(\boldsymbol{u} - \boldsymbol{b}),\\
\mathrm{Var}(\boldsymbol{x}'\hat{\boldsymbol{\beta}}_m) &= \boldsymbol{u}'\boldsymbol{S}\boldsymbol{S}'\boldsymbol{u}.
\end{aligned}
$$

Thus the optimization problem reduces to

$$
\max_{\boldsymbol{u}} \frac{[\boldsymbol{a}'\boldsymbol{S}'\boldsymbol{S}(\boldsymbol{u} - \boldsymbol{b})]^2}{\theta + (\boldsymbol{u} - \boldsymbol{b})'\boldsymbol{S}\boldsymbol{S}'(\boldsymbol{u} - \boldsymbol{b})}, \quad \text{s.t.} \quad \boldsymbol{u}'\boldsymbol{S}\boldsymbol{S}'\boldsymbol{u} \leqslant \alpha. \tag{7}
$$

A further simplification occurs if we let $\boldsymbol{u}_1 = \boldsymbol{P}_{S'a}\boldsymbol{S}'\boldsymbol{u}$, $\boldsymbol{b}_1 = \boldsymbol{P}_{S'a}\boldsymbol{S}'\boldsymbol{b}$, $\boldsymbol{u}_2 = \boldsymbol{S}'\boldsymbol{u} - \boldsymbol{u}_1$ and $\boldsymbol{b}_2 = \boldsymbol{S}'\boldsymbol{b} - \boldsymbol{b}_1$. The solution to Eq. (7) can be obtained from the solution to the following problem.

$$\max_{u_1,u_2} \frac{(u_1 - b_1)'(u_1 - b_1)}{\theta + (u_2 - b_2)'(u_2 - b_2)}, \tag{8}$$
$$\text{s.t.} \quad u_1 \in \mathscr{C}(S'a), \; u_2 \in \mathscr{C}((I - P_{S'a})S'), \; u_1'u_1 + u_2'u_2 \leqslant \alpha.$$

Notice that the objective function of (8) is equivalent to, but not identical with that of (7).

Sengupta (1995) arrived at a similar formulation of the problem using the Inverse Partition Matrix method.

Theorem 4.1 (Sengupta, 1995). *The solution to the optimization problem* (8) *is as follows.*

(i) *If* $b_1 = b_2 = 0$, *the maximum is attained if and only if* $u_1 = \pm(\alpha/a'SS'a)^{1/2}S'a$ *and* $u_2 = 0$.

(ii) *If* $b_1 \neq 0$ *and* $b_2 = 0$, *the maximum is attained if and only if* $u_1 = -(\alpha/b_1'SS'b_1)^{1/2}b_1$ *and* $u_2 = 0$.

(iii) *If* $b_1 = 0$ *and* $b_2 \neq 0$, *the maximum is attained if and only if* $u_2 = c_1 b_2$, *where*

$$c_1 = \frac{b_2'b_2 + \alpha + \theta}{2b_2'b_2}\left[1 - \left(1 - \frac{4\alpha b_2'b_2}{(b_2'b_2 + \alpha + \theta)^2}\right)^{1/2}\right],$$

and $u_1 = \pm[(\alpha - c_1^2 b_2'b_2)/a'SS'a]^{1/2}S'a$.

(iv) *If* b_1 *and* b_2 *are both non-zero, then the maximum is attained if and only if* $u_2 = c_2 b_2$ *where* c_2 *maximizes* $[(\alpha - c_2^2 b_2'b_2)^{1/2} + (b_1'b_1)^{1/2}]^2/[\theta + b_2'b_2(c_2 - 1)^2]$ *over the range* $0 \leqslant c_2 \leqslant (\alpha/b_2'b_2)^{1/2}$, *and* $u_1 = \pm[(\alpha - c_2^2 b_2'b_2)/b_1'b_1]^{1/2}b_2$.

Proof. The proofs of Parts (i) and (ii) are straightforward. The other two parts are proved by holding u_2 fixed, maximizing the numerator of Eq. (8) subject to the constraint $u_1'u_1 \leqslant \alpha - u_2'u_2$, and maximizing the resulting expression with respect to u_2. For details, we refer the reader to Sengupta (1995). □

Remark 4.2. The solution of Part (ii) coincides with one of the two solutions of Part (i).

Remark 4.3. Suppose r_1 is the correlation between y and $p'\hat{\beta}_m$, and r_2 is the multiple correlation of y with $X\hat{\beta}_m$. Then $b_1'b_1 = vr_1^2$ and $b_2'b_2 = v(r_2^2 - r_1^2)$. Thus the special cases of Theorem 4.1 have direct statistical interpretation.

Theorem 4.1. *leads to a choice of* $S'u$ *in each of the four special cases. The following lemma allows one to translate this into a choice of* x.

Lemma 4.4. *The condition* $S'u = S't$ *is equivalent to* $Xu = Xt + X(I - P_V)t_0$ *for some vector* t_0.

Proof. It is clear that $S'u = S't$ if and only if u is of the form $u = t + t_1$ where $S't_1 = 0$. It follows that t_1 is orthogonal to $\mathscr{C}(D(X\hat{\beta}_m))$ which is the same as $\mathscr{C}(X) \cap \mathscr{C}(V)$. Thus t_1 must be of the form $(I - P_X)t_2 + (I - P_V)t_0$. The conclusion follows. □

We are now ready for an explicit solution to the design problem.

Theorem 4.5. *The optimum choice of* x *that minimizes* $\mathrm{Var}(p'\hat{\beta}_{m+1})$ *subject to* $\mathrm{Var}(x'\hat{\beta}_m) \leqslant \alpha$ *is given as follows.*

$$x = \begin{cases} \pm(\alpha/v_p)^{1/2} p + x_0 & \text{if } r_2 = 0, \\[2ex] -(\alpha/vr_1^2)^{1/2} X'V^- v + x_0 & \text{if } r_2^2 = r_1^2 > 0, \\[2ex] [(\alpha - c_1^2 v r_2^2)/v_p]^{1/2} + c_1 X'V^- v + x_0 & \text{if } r_2 > 0 = r_1, \\[2ex] -\left[c_2 + \left\{\frac{\alpha - c_2^2 v(r_2^2 - r_1^2)}{vr_1^2}\right\}^{1/2}\right](v/v_p)^{1/2} r_1 p & \\ \quad + c_2 X'V^- v + x_0 & \text{if } r_2^2 > r_1^2 > 0, \end{cases}$$

where $v_p = \mathrm{Var}(p'\hat{\beta}_m)/\sigma^2$, x_0 *is an arbitrary vector in* $\mathscr{C}(X(I - P_V))$, r_1 *and* r_2 *are as in Remark* 4.3, *and*

$$c_1 = \left(\frac{1}{2} + \frac{\alpha + \theta}{2vr_2^2}\right)\left[1 - \left(1 - \frac{4\alpha v r_2^2}{(\alpha + \theta + vr_2^2)^2}\right)^{1/2}\right],$$

$$c_2 = \arg \max_{c \in [0, \{\alpha/v(r_2^2 - r_1^2)\}^{1/2}]} \frac{\left[\{\alpha - c^2 v(r_2^2 - r_1^2)\}^{1/2} + \{vr_1^2\}^{1/2}\right]^2}{\theta + v(r_2^2 - r_1^2)(c - 1)^2}.$$

Proof. The results follow from Theorem 4.1 and Lemma 4.4 after some algebra. □

Remark 4.6. The ambiguity in the choice of V^- can be removed by replacing $X'V^-v$ by $X'P_V V^- v$. The difference between the two terms is absorbed by the arbitrary vector x_0.

Remark 4.7. The intuitive solution of choosing x in the direction of p is optimal not only in the homoscedastic case, but whenever $r_2 = -r_1 > 0$. If $r_2 = r_1 > 0$, the opposite direction is optimal. Both of these cases correspond to the situation when the multiple correlation of y with $X\hat{\beta}_m$ is the same (in magnitude) as its correlation with $p'\hat{\beta}_m$ alone. Both the solutions are optimal when $r_2 = 0$. The assumption of uncorrelated error variances is a special case when $r_2 = 0$.

5. Other applications

Since the results obtained in Sections 2 and 3 are applicable to the case of a non-diagonal and singular dispersion matrix of the model errors, these should be useful for interpreting certain limiting cases. For instance, one can observe the effects of a transition from near-singularity to perfect singularity of V and from weak correlation to lack of correlation of the observations.

The group recursive residuals defined in Remark 2.4 have properties analogous to those of the recursive residuals that have been traditionally used in the homoscedastic set-up. An extension of the weak convergence results of Sen (1982) to the general case appears to be achievable and will be considered elsewhere.

The updates for a set of dropped observations may be used to interpret deletion diagnostics for linear regression, and to generalize them to the case of a general disperson matrix.

The updates for a set of dropped parameters help quantify the rewards and penalties of having fewer parameters, which is important for achieving parsimony. The updates for additional parameters play a similar role, and work in the reverse direction.

References

Basu, D., 1959. The family of ancillary statistics. Sankhyā Ser. A 21, 247–256.

Bhimasankaram, P., Jammalamadaka, S.R., 1994. Updates of Statistics in a general linear model: A statistical interpretation, applications. Communication in Statistics, Ser. B 23, 789–801.

Bhimasankaram, P., Sengupta, D., Ramanathan, S., 1995. Recursive inference in a general linear model. Sankhyā Ser. A 57, 227–255.

Bhimasankaram, P., Sengupta, D., 1996. The linear zero functions approach to linear models. Submitted. Also, Technical Report No. ASC/95/9, ASSC Division, Indian Statistical Institute, Calcutta.

Bich, W., 1990. Variances, covariances, restraints in mass metrology. Metrologia 27, 111–116.

Brown, R.L., Durbin, J., Evans, J.M., 1975. Techniques for testing the constancy of regression relationships over time. Journal of the Royal Statistical Society Ser. B 37, 149–163.

Chib, S., Jammalamadaka, S.R., Tiwari, R.C., 1987. Another look at some results on the recursive estimation in the general linear model. American Statistician 41, 56–58.

Christensen, R., 1987. Plane Answers to Complex Questions: The Theory of Linear Models. Springer, Berlin.

Farebrother, R.W., 1988. Linear Least Squares Computations. Marcel Dekker, New York.

Feuerverger, A., Fraser, D.A.S., 1980. Categorical information, the singular linear model. Canadian Journal of Statistics 8, 41–45.

Gragg, W.B., LeVeque, R.J., Trangenstein, J.A., 1979. Numerically stable methods for updating regressions. Journal of the American Statistical Association 74, 161–168.
Harville, D.A., 1981. Unbiased, mimimum-variance unbiased estimation of estimable functions for fixed linear models with arbitrary covariance structure. Annals of Statistics 9, 633–637.
Haslett, S., 1985. Recursive estimation of the general linear model with dependent errors, multiple additional observations. Australian Journal of Statistics 27, 183–188.
Kianifard, F., Swallow, W.H., 1996. A review of the development, application of recursive residuals in linear models. Journal of the American Statistical Association 91, 391–400.
Kempthorne, O., 1952. The Design, Analysis of Experiments. Wiley, New York, pp. 137–138, 190–191.
Kourouklis, S., Paige, C.C., 1981. A constrained least squares approach to the general Gauss-Markov linear model. Journal of the American Statistical Association 76, 620–625.
McGilchrist, C.A., Sandland, R.L., 1979. Recursive estimation of the general linear model with dependent errors. Journal of the Royal Statistical Society, Ser. B 41, 65–68.
Mitra, S.K., Bhimasankaram, P., 1971. Generalized inverses of partitioned matrices, recalculation of least squares estimators for data or model changes. Sankhyā Ser. A 33, 396–410.
Plackett, R.L., 1950. Some theorems in least squares. Biometrika 37, 149–157.
Rao, C.R., 1968. A note on a previous lemma in the theory of least squares, some further results. Sankhyā Ser. A 30, 259–266.
Rao, C.R., 1973. Linear Statistical Inference, its Applications. Wiley, New York.
Rao, C.R., 1978. Least squares theory for possibly singular models. Canadian Journal of Statistics 6, 19–23.
Rowley, J.C.R., 1977. Singularities in econometric models of wage determination based on time series data. ASA Proceedings of Business, Economic Statistics Section, pp. 616–621.
Scheffé, H., 1959. The Analysis of Variance. Wiley, New York, pp. 299–301.
Searle, S.R., 1994. Extending some results, proofs for the singular linear model. Linear Algebra and Applications 210, 139–151.
Sengupta, D., 1995. Optimal choice of a new observation in a linear model. Sankhyā Ser. A 57, 137–153.