Sample Size Requirements for Estimation in SUR Models[1]

William E. Griffiths
University of Melbourne

Christopher L. Skeels
Australian National University

Duangkamon Chotikapanich
Curtin University of Technology

February 26, 2001

[1] The authors would like to thank Trevor Breusch, Denzil Fiebig and an anonymous referee for helpful comments over the paper's development. They are, of course, absolved from any responsibility for shortcomings of the final product.

Abstract

This paper explores sample size requirements for the estimation of SUR models by (two-stage) feasible generalized least squares, maximum likelihood and Bayesian methods. It is found that the sample size requirements presented in standard treatments of SUR models are incomplete and potentially misleading. It is also demonstrated that likelihood-based methods potentially require larger sample sizes than does the two-stage estimator considered in this paper.

1 Introduction

The seemingly unrelated regressions (SUR) model [17] has become one of the most frequently used models in applied econometrics. The coefficients of individual equations in such models can be consistently estimated by ordinary least squares (OLS) but, except for certain special cases, efficient estimation requires joint estimation of the entire system. System estimators which have been used in practice include two-stage methods based on OLS residuals, maximum likelihood (ML), and, more recently, Bayesian methods; e.g. [11, 2]. Various modifications of these techniques have also been suggested in the literature; e.g. [14, 5]. It is somewhat surprising, therefore, that the sample size requirements for joint estimation of the parameters in this model do not appear to have been correctly stated in the literature. In this paper we seek to correct this situation.

The usual assumed requirement for the estimation of SUR models may be paraphrased as the sample size must be greater than the number of explanatory variables in each equation and at least as great as the number of equations in the system. Such a statement is flawed in two respects. First, the estimators considered sometimes have more stringent sample size requirements than implied by this statement. Second, the different estimators may have different sample size requirements. In particular, for a given model the maximum likelihood estimator and the Bayesian estimator with a non-informative prior may require larger sample sizes than does the two-stage estimator.

To gain an appreciation of the different sample size requirements, consider a 4-equation SUR model with 3 explanatory variables and a constant in each equation. Suppose that the explanatory variables in different equations are distinct. Although one would not contemplate using such a small number of observations, two-stage estimation can proceed if the number of observations (T) is greater than 4. For ML and Bayesian estimation $T > 16$ is required. For a 10-equation model with 3 distinct explanatory variables and a constant in each equation, two-stage estimation requires $T \geq 11$ — the conditions outlined above would suggest that 10 observations should be sufficient — whereas ML and Bayesian estimation need $T > 40$. This last example illustrates not only that the usually stated sample size requirements can be incorrect, as they are for the two-stage estimator, but also just how misleading they can be for likelihood-based estimation.

The structure of the paper is as follows. In the next section we introduce the model and notation, and illustrate how the typically stated sample size requirement can be misleading. Section 3 is comprised of two parts. The first part of Section 3 derives a necessary condition on sample size for two-stage estimation and provides some discussion of this result; the second part illustrates the result by considering its application in a variety of different situations. Section 4 derives the analogous result for ML and Bayesian estimators. It is also broken into two parts, the first of which derives and discusses the result, while the second part illustrates the result by examining the model that was the original motivation for this paper. Concluding remarks are presented in Section 5.

2 The Model and Preliminaries

Consider the SUR model written as

$$y_j = X_j\beta_j + e_j, \qquad j = 1, \ldots , M, \tag{1}$$

where y_j is a $(T \times 1)$ vector of observations on the dependent variable for the jth equation, X_j is a $(T \times K_j)$ matrix of observations on K_j explanatory variables in the jth equation. We will assume that each of the X_j have full column rank. It will also be assumed that the $(TM \times 1)$ continuously distributed random vector $e = [e_1', e_2', \ldots , e_M']'$ has mean vector zero and covariance matrix $\Sigma \otimes I_T$. We are concerned with estimation of the $(K \times 1)$ coefficient vector $\beta = [\beta_1', \ldots , \beta_M']'$, where $K = \sum_{j=1}^{m} K_j$ is the total number of unknown coefficients in the system.

The OLS estimators for the β_j are

$$b_j = (X_j'X_j)^{-1}X_jy_j, \qquad j = 1, \ldots , M.$$

The corresponding OLS residuals are given by

$$\hat{e}_j = y_j - X_jb_j = M_{X_j}y_j, \tag{2}$$

where $M_A = I - A(A'A)^{-1}A' = I - P_A$ for any matrix A of full column rank. We shall define

$$\widehat{E} = [\hat{e}_1, \hat{e}_2, \ldots , \hat{e}_M].$$

The two-stage estimator for this system of equations is given by

$$\hat{\beta} = [\widehat{X}'(\widehat{\Sigma}^{-1} \otimes I_T)\widehat{X}]^{-1}\widehat{X}'(\widehat{\Sigma}^{-1} \otimes I_T)y, \tag{3}$$

where the $(TM \times K)$ matrix $\widehat{X}$ is block diagonal with the X_j making up the blocks, $y = [y_1', y_2', \ldots, y_M']'$, and[1]

$$T\widehat{\Sigma} = \widehat{E}'\widehat{E}.$$

Clearly, $\hat{\beta}$ is not operational unless $\widehat{\Sigma}$ is non-singular and $\widehat{\Sigma}$ will be singular unless the $(T \times M)$ matrix $\widehat{E}$ has full column rank. A standard argument is that $\widehat{E}$ will have full column rank with probability one provided that (i) $T \geq M$, and (ii) $T > k_{max} = \max_{j=1,\ldots,M} K_j$. *Observe that (ii) is stronger than the $T \geq k_{max}$ requirement implicit in assuming that all X_j have full column rank*; it is required because for any $K_j \geq T$ the corresponding $\hat{e}_j$ is identically equal to zero, ensuring that $\widehat{\Sigma}$ is singular. Conditions (i) and (ii) can be summarized as[2]

$$T \geq \max(M, k_{max} + 1). \tag{4}$$

That (4) does not ensure the non-singularity of $\widehat{\Sigma}$ is easily demonstrated by considering the special case of the multivariate regression model, arising when $X_1 = \ldots = X_M = X^*$ (say) and hence $K_1 = \ldots = K_M = k^*$ (say). In this case model (1) reduces to

$$Y = X^*B + E,$$

where

$$\begin{aligned} Y &= [y_1, y_2, \ldots, y_M], \\ B &= [\beta_1, \beta_2, \ldots, \beta_M], \\ E &= [e_1, e_2, \ldots, e_M], \end{aligned} \tag{5}$$

and

$$T\widehat{\Sigma} = E'E = Y'M_{X^*}Y.$$

[1] A number of other estimators of Σ have been suggested in the literature; they differ primarily in the scaling applied to the elements of $\widehat{E}'\widehat{E}$ (see, for example, the discussion in [13, p.17]), but $\widehat{\Sigma}$ is that estimator most commonly used. Importantly, $\widehat{\Sigma}$ uses the same scaling as does the ML estimator for Σ, which is the appropriate choice for likelihood-based techniques of inference. Finally, $\widehat{\Sigma}$ was also the choice made by [12] when deriving exact finite-sample distributional results for the two-stage estimator in this model. Our results are consequently complementary to those earlier ones.

[2] Sometimes only part of the argument is presented. For example, [6, p.627] formally states only (i).

Using a full rank decomposition, we can write

$$M_{X^*} = CC',$$

where C is a $(T \times (T - k^*))$ matrix of rank $T - k^*$. Setting $W = C'Y$, we have

$$T\widehat{\Sigma} = W'W,$$

whence it follows that, in order for $\widehat{\Sigma}$ to have full rank, the $((T - k^*) \times M)$ matrix W must have full column rank, which requires that $T - k^* \geq M$ or, equivalently, that[3]

$$T \geq M + k^*. \tag{6}$$

In the special case of $M = 1$, condition (6) corresponds to condition (4), as $k^* = k_{max}$. For $M > 1$, condition (6) is more stringent in its requirement on sample size than condition (4), which begs the question as to whether even more stringent requirements on sample size exist for the two-stage estimation of SUR models. It is to this question that we turn next.

3 Two-Stage Estimation

3.1 A Necessary Condition and Discussion

Our proposition is that, subject to the assumptions given with model (1), two-stage estimation is feasible provided that $\widehat{\Sigma}$ is non-singular. Unfortunately, $\widehat{\Sigma}$ is a matrix of random variables and there is nothing in the assumptions of the model that precludes $\widehat{\Sigma}$ being singular. However, there are sample sizes that are sufficiently small that $\widehat{\Sigma}$ must be singular and in the following theorem we shall characterize these sample sizes. For all larger sample sizes $\widehat{\Sigma}$ will be non-singular with probability one and two-stage estimation feasible. That $\widehat{\Sigma}$ is only non-singular *with probability one* implies that our condition is necessary but not sufficient for the two-stage estimator to be feasible; unfortunately, no stronger result is possible.[4]

[3] It should be noted that condition (6) is not new and is typically assumed in discussion of the multivariate regression model; see, for example, [1, p. 287], or the discussion of [4] in an empirical context.

[4] A necessary and sufficient condition is only available if one imposes on the support of e restrictions which preclude the possibility $\widehat{\Sigma}$ being singular except when the sample size is sufficiently small. For example, one would have to preclude the possibility of either multicollinearity between the y_j or any $y_j = 0$, both of which would ensure a singular $\widehat{\Sigma}$. We have avoided making such assumptions here.

Theorem 1. *In the model* (1)*, a necessary condition for the estimator* (3) *to be feasible, in the sense that* $\widehat{\Sigma}$ *is non-singular with probability one, is that*

$$T \geq M + \rho - \eta, \tag{7}$$

where $\rho = \text{rank}([X_1, X_2, \ldots, X_M])$ *and* η *is the rank of the matrix* D *defined in equations* (9) *and* (10)*.*

Proof. Let $X = [X_1, X_2, \ldots, X_M]$ be the $(T \times K)$ matrix containing all the explanatory variables in all equations. We will define $\rho = \text{rank}(X) \leq T$. Next, let the orthogonal columns of the $(T \times \rho)$ matrix V comprise a basis set for the space spanned by the columns of X, so that there exists $(\rho \times k_j)$ matrices F_j such that $X_j = VF_j$ (for all $j = 1, \ldots, M$).[5] Under the assumption that X_j has full column rank, it follows that F_j must also have full column rank and so we see that $k_{max} \leq \rho \leq T$. It also follows that there exist $(\rho \times (\rho - k_j))$ matrices G_j such that $[F_j, G_j]$ is non-singular and $F_j' G_j = 0$. Given G_j we can define $(T \times (\rho - k_j))$ matrices

$$Z_j = VG_j, \quad j = 1, \ldots, M.$$

The columns of each Z_j span that part of the column space of V not spanned by the columns of the corresponding X_j. By Pythagoras' Theorem,

$$P_V = P_{X_j} + M_{X_j} Z_j (Z_j' M_{X_j} Z_j)^{-1} Z_j' M_{X_j}.$$

Observe that, because $X_j' Z_j = 0$ by construction, $M_{X_j} Z_j = Z_j$ giving $P_V = P_{X_j} + P_{Z_j}$ or, equivalently,

$$M_{X_j} = M_V + P_{Z_j}.$$

From equation (2),

$$\hat{e}_j = [M_V + P_{Z_j}] y_j,$$

so that

$$\widehat{E} = M_V Y + D, \tag{8}$$

where

$$D = [d_1, \ldots, d_M], \tag{9}$$

[5] Subsequently, if the columns of a matrix A (say) form a basis set for the space spanned by the columns of another matrix Q (say) we shall simply say that A is a basis for Q.

with

$$d_j = \begin{cases} P_{Z_j} y_j, & \text{if } \rho > k_j, \\ 0, & \text{otherwise,} \end{cases} \qquad j = 1, \dots, M. \tag{10}$$

Thus, the OLS residual for each equation can be decomposed into two orthogonal components, $M_V y_j$ and d_j. $M_V y_j$ is the OLS residual from the regression of y_j on V and d_j is the orthogonal projection of y_j on to that part of the column space of V which is not spanned by the columns of X_j. Noting that $Y' M_V D = 0$, because $M_V Z_j = 0$ $(j = 1, \dots, M)$, equation (8) implies that

$$\widehat{E}' \widehat{E} = Y' M_V Y + D' D. \tag{11}$$

It is well known that if R and S are any two matrices such that $R + S$ is defined then[6]

$$\operatorname{rank}(R + S) \leq \operatorname{rank}(R) + \operatorname{rank}(S). \tag{12}$$

Defining $\theta = \operatorname{rank}(\widehat{E}' \widehat{E})$, $\delta = \operatorname{rank}(Y' M_V Y)$ and $\eta = \operatorname{rank}(D)$, equations (11) and (12) give us

$$\theta \leq \delta + \eta.$$

Now, M_V admits a full rank decomposition of the form $M_V = HH'$, where H is a $(T \times (T - \rho))$ matrix. Consequently, $\delta = \operatorname{rank}(Y' M_V Y) = \operatorname{rank}(Y' H) \leq \min(M, T - \rho)$, with probability one, so that

$$\theta \leq \min(M, T - \rho) + \eta.$$

Clearly, $\widehat{E}' \widehat{E}$ has full rank if and only if $\theta = M$, which implies

$$M \leq \min(M, T - \rho) + \eta. \tag{13}$$

If $T - \rho \geq M$, equation (13) is clearly satisfied. Thus, the binding inequality for (13) occurs when $\min(M, T - \rho) = T - \rho$. □

As noted in equation (6), and in further discussion below, $T \geq M + \rho$ is the required condition for a class of models that includes the multivariate regression model. When some explanatory variables are omitted from some equations, the sample size requirement is less stringent, with the reduced requirement depending on $\eta = \operatorname{rank}(D)$.

[6] See, for example, [10, A.6(iv)].

Care must be exercised when applying the result in (7) because of several relationships that exist between M, T, ρ and η. We have already noted that $\rho \leq T$. Let us examine η more closely. First, because D is a $(T \times M)$ matrix, and $\eta = \text{rank}(D)$, it must be that $\eta \leq \min(M, T)$. Actually, we can write $\eta \leq \min(d, T)$, where $0 \leq d \leq M$ denotes the number of non-zero d_j; that is, d is the number of X_j which do not form a basis for X.

Second, the columns of D are a set of projections onto the space spanned by the columns of $Z = [Z_1, Z_2, \ldots, Z_M]$, a space of possibly lower dimension than ρ, say $\rho - \omega$, where $0 \leq \omega \leq \rho$. In practical terms, Z_j is a basis for that part of V spanned by the explanatory variables excluded from the j-th equation; the columns of Z span that part of V spanned by all variables excluded from at least one equation. If there are some variables common to all equations, and hence not excluded from any equations, then Z will not span the complete ρ-dimensional space spanned by V. More formally, we will write $V = [V_1, V_2]$, where the $(T \times (\rho - \omega))$ matrix V_2 is a basis for Z and the $(T \times \omega)$ matrix V_1 is a basis for a sub-space of V spanned by the columns of each X_j, $j = 1, \ldots, M$. The most obvious example for which $\omega > 0$ is when each equation contains an intercept. Another example, is the multivariate regression model, where $\omega = \rho$, so that V_2 is empty and $D = 0$. Clearly, because $T \geq \rho \geq \rho - \omega$, the binding constraint on η is not $\eta \leq \min(d, T)$ but rather $\eta \leq \min(d, \rho - \omega)$.

Note that $\eta \leq \min(d, \rho - \omega) \leq \rho - \omega \leq \rho$, which implies that $T \geq M + \rho - \eta \geq M$ is a necessary condition for $\widehat{\Sigma}$ to be non-singular. Obviously $T \geq M$ is part of (4). The shortcoming of (4) is its failure to recognize the interactions of the X_j in the estimation of (1) as a system of equations. In particular, $T \geq k^* + 1$ is an attempt to characterize the entire system on the basis of those equations which are most extreme in the sense of having the most regressors. As we shall demonstrate, such a characterization is inadequate.

Finally, it is interesting to note the relationship between the result in Theorem 1 and results in the literature for the existence of the mean of a two-stage estimator. [15] show that sufficient conditions for the existence of the mean of a two-stage estimator that uses an error covariance matrix estimated from the residuals of the corresponding unrestricted multivariate regression model are (i) the errors have finite moments of order 4, and (ii) $T > M + K^* + 1$ where K^* denotes the number of distinct regressors in the system.[7] They also provide other (alternative) sufficient conditions that are equally relevant when residuals from the restricted SUR model are used. The existence of higher order moments of a two-step estimator in a two-equation model have

[7] Clearly K^* is equivalent to ρ in the notation of this paper.

also been investigated by [8]. In every case we see that the result of Theorem 1 for the existence of the estimator is less demanding of sample size than are the results for the existence of moments of the estimator. This is not surprising. The existence of moments requires sufficiently thin tails for the distribution of the estimator. Reduced variability invariably requires increased information which manifests itself in a greater requirement on sample size.[8] This provides even stronger support for our assertion that the usually stated sample size requirements are inadequate because the existence of moments are important to many standard techniques of inference.

3.2 Applications of the Necessary Condition

In what follows we shall exploit the randomness of the d_j to obtain $\eta = \min(d, \rho - \omega)$ with probability one. Consequently, (7) reduces to

$$T \geq \begin{cases} M + \omega, & \text{for } \eta = \rho - \omega, \\ M + \rho - d, & \text{for } \eta = d. \end{cases} \tag{14}$$

Let us explore these results through a series of examples.

First, $\eta = 0$ requires either $d = 0$ or $\rho = \omega$ (or both). Both of these requirements correspond to the situation where each X_j is a basis for X, so that $\rho = k_{max} = K_j$ for $j = 1, \dots, M$.[9] Note that this does not require $X_1 = \dots = X_M$ but does include the multivariate regression model as the most likely special case. In this case, (14) reduces to

$$T \geq M + k^*, \tag{15}$$

which is obviously identical to condition (6) and so need not be explored further.

Next consider the model

$$y_1 = x_1\beta_{11} + e_1, \tag{16a}$$
$$y_2 = x_1\beta_{21} + x_2\beta_{22} + x_3\beta_{23} + e_2, \tag{16b}$$
$$y_3 = x_1\beta_{31} + x_2\beta_{32} + x_3\beta_{33} + e_3. \tag{16c}$$

Here $M = 3$, $\rho = 3$, $d = 1$ and $\omega = 1$, so that $\eta = \min(1, 3 - 1) = 1$. Such a system will require a minimum of 5 observations to estimate. If $\beta_{23} \equiv 0$, so

[8] The arguments underlying these results are summarised by [14, Chapter 4].

[9] The equality $\rho = k_{max} = K_j$ follows from our assumption of full column rank for each of the X_j. If this assumption is relaxed, condition (15) becomes $T \geq M + \rho$, which is the usual sample size requirement in rank deficient multivariate regression models; see [9, Section 6.4].

that x_3 no longer appears in equation (16b), then $\eta = d = \rho - \omega = 2$ and the sample size requirement for the system reduces to $T \geq 4$. Suppose now that, in addition to deleting x_3 from equation (16b), we include x_2 in equation (16a). In this case, $\omega = d = 2$ but $\eta = \rho - \omega = 1$ and, once again, the sample size requirement is 5. Finally, if we add x_2 and x_3 to equation (16a) and leave equation (16b) as stated, so that the system becomes a multivariate regression equation, the sample size requirement becomes $T \geq 6$. None of the changes to model (16) that have been suggested above alter the prediction of condition (4), which is that 4 observations should be sufficient to estimate the model. Hence, condition (4) typically under-predicts the actual sample size requirement for the model and it is unresponsive to certain changes in the composition of the model which do impact upon the sample size requirement of the two-stage estimator.

In the previous example we allowed d and $\rho - \omega$ and η to vary but at no time did the sample size requirement reduce to $T = M$. The next example provides a simple illustration of this situation. A common feature with the previous example will be the increasing requirement on sample size as the commonality of regressors across the system of equations increases, where again ω is the measure of commonality. Heuristically, the increase in sample size is required to compensate for the reduced information available when the system contains fewer distinct explanatory variables. Consider the two-equation models

$$\begin{aligned} y_1 &= x_1\beta_1 + e_1, \\ y_2 &= x_2\beta_2 + e_2, \end{aligned} \tag{17}$$

and

$$\begin{aligned} y_1 &= x_1\beta_1 + e_1, \\ y_2 &= x_1\beta_2 + e_2. \end{aligned} \tag{18}$$

In model (17) there are no common regressors and so $\omega = 0$, which implies that $\eta = \min(d, \rho) = \min(2, 2) = 2$. Consequently, (14) reduces to $T \geq M + \rho - \eta = 2$. That is, model (17) can be estimated using a sample of only two observations or, more importantly, one can estimate as many equations as one has observations.[10] Model (18) is a multivariate regression model and so $T \geq M + \rho = 3$.

As a final example, if $\rho = T$ then $Y'M_VY = 0$ and the estimability of the model is determined solely by η. From condition (14) we see that, in order

[10] This is a theoretical minimum sample size and should not be interpreted as a serious suggestion for empirical work!

to estimate M equations, we require $\eta = M$. But $\eta \leq \rho - \omega \leq T - \omega$ and so $M = T - \omega$ is the largest number of equations that can be estimated on the basis of T observations. This is the result observed in the comparison of models (17) and (18); it will be encountered again in Section 4, where each equation in a system contains an intercept, so that $\omega = 1$, and $M = T - 1$ is the largest number of equations that can be estimated for a given sample size. The case of $\rho = T$ would be common in large systems of equations where each equation contributes it own distinct regressors; indeed this is the context in which it arises in Section 4.

4 Maximum Likelihood and Bayesian Estimation

4.1 A Necessary Condition and Discussion

Likelihood-based estimation, be it maximum likelihood or Bayesian, requires distributional assumptions and so we will augment our earlier assumptions about e by assuming that the elements of e are jointly normally distributed. Consequently, the log-likelihood function for model (1) is

$$L = -(TM/2)\log(2\pi) - (T/2)\log|\Sigma| - (1/2)tr(S\Sigma^{-1}), \tag{19}$$

where $S = E'E$. The ML estimates for β and Σ are those values which simultaneously satisfy the first-order conditions

$$\tilde{\beta} = [\widehat{X}'(\widetilde{\Sigma}^{-1} \otimes I_T)\widehat{X}]^{-1}\widehat{X}'(\widetilde{\Sigma}^{-1} \otimes I_T)y, \tag{20}$$

and

$$T\widetilde{\Sigma} = \widetilde{E}'\widetilde{E}, \tag{21}$$

with $\text{vec}(\widetilde{E}) = y - \widehat{X}\tilde{\beta}$. This is in contrast to the two-stage estimator which obtains estimates of Σ and β sequentially. Rather than trying to simultaneously maximize L with respect to both β and Σ, it is convenient to equivalently maximize L with respect to β subject to the constraint that $\Sigma = S/T$, which will ensure that the ML estimates, $\tilde{\beta}$ and $\widetilde{\Sigma}$, satisfy equations (20) and (21). Imposing the constraint by evaluating L at $\Sigma = S/T$ gives the concentrated log-likelihood function[11]

$$L^*(\beta) = \text{constant} - \frac{T}{2}\log|S|.$$

[11] See, for example, [7, p.553].

Similarly, using a prior density function $f(\beta, \Sigma) \propto |\Sigma|^{-(M+1)/2}$, it can be shown that the marginal posterior density function for β is[12]

$$f(\beta|y) \propto |S|^{-T/2}.$$

Consequently, we see that both the ML and Bayesian estimators are obtained by minimizing the generalized variance $|S|$ with respect to β.

The approach adopted in this section will be to demonstrate that, for sufficiently small samples, there necessarily exists β's such that S is singular (has rank less than M), so that $|S| = 0$. In such cases, ML estimation cannot proceed as the likelihood function is unbounded at these points; similarly, the posterior density for β will be improper at these β's. Since $S = E'E$, S will be singular if and only if E has rank less than M. Consequently, our problem reduces to determining conditions under which there necessarily exists β such that E is rank deficient.

Theorem 2. *In the model* (1)*, augmented by a normality assumption, a necessary condition for $S = E'E$ to be non-singular, and hence for the likelihood function* (19) *to be bounded, is that*

$$T \geq M + \rho, \tag{22}$$

where $\rho = \text{rank}([X_1, X_2, \dots, X_M])$ *and E is defined in equation* (5)*.*

Proof. E will have rank less than M if there exists an $(M \times 1)$ vector $c = [c_1, c_2, \dots, c_M]'$ such that $Ec = 0$. This is equivalent to the equation

$$\Phi\alpha = 0,$$

where the $(T \times (M + K))$ matrix $\Phi = [Y, -X]$ and the $((M + K) \times 1)$ vector $\alpha = [c', c_1\beta_1', c_2\beta_2', \dots, c_M\beta_M']'$. A non-trivial solution for α, and hence a non-trivial c, requires that Φ be rank deficient. But rank $\Phi = \min(T, M + \rho)$ with probability one and so, in order for E to be rank deficient, it follows that $T < M + \rho$. A necessary condition for E to have full column rank with probability one is then the converse of the condition for E to be rank deficient. □

A number of comments are in order. First, (22) is potentially more stringent in its requirements on sample size than is (14). Theorem 1 essentially provides a spectrum of sample size requirements, $T \geq M + \rho - \eta$, where the actual requirement depends on the specific data set used for estimation of the model. The likelihood-based requirement is the most stringent of those for

[12] See, for example, [18, p.242].

the two-stage estimator, corresponding to the situation where $\eta = 0$. That η can differ from zero stems from the fact that the OLS residuals used in the construction of $\widehat{\Sigma}$ by the two-stage estimator satisfy $\hat{e}_j = M_{X_j} y_j$, whereas the likelihood-based residuals used in the construction of $\widetilde{\Sigma}$ need not satisfy the analogous $\tilde{e}_j = M_{X_j} y_j$.

Second, the proof of Theorem 2 is not readily applicable to the two-stage estimator considered in Theorem 1. In both cases we are concerned with determining the requirements for a solution to an equation of the form $Ac = 0$, with $A = \widehat{E} = [M_{X_1}y_1, M_{X_2}y_2, \ldots, M_{X_M}y_M]$ in Theorem 1 and $A = E$ in Theorem 2. For the two-stage estimator, the interactions between the various X_j are important and complicate arguments about rank. The decomposition (11) provides fundamental insight into these interactions, making it possible to use arguments of rank to obtain the necessary condition on sample size. The absence of corresponding relationships between the vectors of likelihood-based residuals means that a decomposition similar to equation (11) is not required and that the simpler proof of Theorem 2 is sufficient.

An alternative way of thinking about why the development of the previous section differs from that of this section is to recognise that the problems being addressed in the two sections are different. The difference in the two problems can be seen by comparing the criteria for estimation. For likelihood-based estimators the criterion is to choose β to minimize $|S|$, a polynomial of order $2M$ in the elements of β. For the two-stage estimator a quadratic function in the elements of β is minimized. The former problem is a higher-dimensional one for all $M > 1$ and, consequently, its solution has larger minimal information (or sample size) requirements when the estimators diverge.[13]

4.2 Applications of the Necessary Condition

In light of condition (22), it is possible to re-examine some empirical work undertaken by [3] to investigate the effect of increasing M for fixed T and K_j. This investigation was motivated by the work of [5]. The data set was such that $T = 19$ and $K_j = 3$ for all j. Each X_j contained an intercept and two other regressors that were unique to that equation. Consequently, in this model $\rho = \min(2M+1, T)$, so that $\text{rank}(\Phi) = \min(3M+1, T)$, and condition (22) predicts that likelihood-based methods should only be able to estimate systems containing up to $M = (T-1)/3 = 6$ equations.[14] Conversely, as

[13] There are special cases where the two-stage estimator and the likelihood-based estimators coincide; see [14]. Obviously, their minimal sample size requirements are the same in these cases.

[14] Strictly the prediction is $M = [(T-1)/3]$ equations, where $[x]$ denotes the integer part of x. Serendipitously $(T-1)/3$ is exactly 6 in this case.

demonstrated below, condition (14) predicts that two-stage methods should be able to estimate systems containing up to $M = T - 1 = 18$ equations. This is exactly what was found. Although the two-stage estimator for the first two equations gave relatively similar results for M all the way up to 18, the authors had difficulty with the maximum likelihood and Bayesian estimators for $M \geq 7$. With maximum likelihood estimation the software package SHAZAM ([16]) sometimes uncovered singularities and sometimes did not, but, when it did not, the estimates were quite unstable. With Bayesian estimation the Gibbs sampler broke down from singularities, or got stuck in a narrow nonsensical range of parameter values.

The statement of condition (14) implicitly assumes that the quantity of interest is sample size. It is a useful exercise to illustrate how (14) should be used to determine the maximum number of estimable equations given the sample size; we shall do so in the context of the model discussed in the previous paragraph.[15] To begin observe that (i) $d = M$, because each equation contains two distinct regressors; (ii) $\omega = 1$, because each equation contains an intercept; and (iii) $\rho = \min(2M+1, T)$, as before, so that $\rho - \omega = \min(2M, T-1)$. Unless $M > T - 1$, $d \leq \rho - \omega$, so that $\eta = d$, and condition (14) reduces to $T \geq M + \rho - d$. Clearly, condition (14) is satisfied for all $T \geq M + 1$ in this model, because $d = M$ and $T \geq \rho$ by definition. Next suppose that $M > T - 1$. Then $\eta = \rho - \omega = T - 1 < M = d$ and (14) becomes $M \leq T - \omega = T - 1$. But this is a contradiction and so, in this model, the necessary condition for $\widehat{\Sigma}$ to be non-singular with probability one is that $M \leq T - 1$.[16] It should be noted that this is one less equation than predicted by condition (4), where $T \geq M$ will be the binding constraint, as $19 = T > k_{max} + 1 = 3$ in this case.

[15] The final example of Section 3 is similar to this one except for the assumption that $\rho = T$, which is not made here.

[16] An alternative proof of this result comes from working with condition (13) directly and recognizing that the maximum number of equations that can be estimated for a given sample size will be that value at which the inequality is a strict equality. Substituting for $d = M$, $\rho = \min(2M + 1, T)$ and $\eta = \min(d, \rho - \omega)$ in (13) yields

$$M \leq \min(M, T - \min(2M + 1, T)) + \min(M, \min(2M + 1, T) - 1),$$

which will only be violated if

$$T - \min(2M + 1, T) + \min(2M + 1, T) - 1 = T - 1 < M,$$

as required.

5 Concluding Remarks

This paper has explored sample size requirements for the estimation of SUR models. We found that the sample size requirements presented in standard treatments of SUR models are, at best, incomplete and potentially misleading. We also demonstrated that likelihood-based methods potentially require much larger sample sizes than does the two-stage estimator considered in this paper.

It is worth noting that the nature of the arguments for the likelihood-based estimators is very different to that presented for the two-stage estimator. This reflects the impact of the initial least squares estimator on the behaviour of the two-stage estimator on the structure of the problem.[17] In both cases the results presented are necessary but not sufficient conditions. This is because we are discussing the non-singularity of random matrices and so there exists sets of Y (of measure zero) such that $\widehat{\Sigma}$ and $\widetilde{\Sigma}$ are singular even when the requirements presented here are satisfied. Alternatively, the results can be thought of as necessary and sufficient with probability one.

Our numerical exploration of the results derived in this paper revealed that standard packages didn't always cope well with undersized samples.[18] For example, it was not uncommon for them to locate local maxima of likelihood functions rather than correctly identify unboundedness. In the case of two-stage estimation, singularity of $\widehat{\Sigma}$ sometimes resulted in the first stage OLS estimates being reported without meaningful further comment. Consequently, we would strongly urge practitioners to check the minimal sample size requirements and, if their sample size is at all close to the minimum bound, take steps to ensure that the results provided by their computer package are valid.

References

[1] T. W. Anderson. *An Introduction to Multivariate Statistical Analysis.* John Wiley and Sons, Inc., New York, second edition, 1984.

[2] S. Chib and E. Greenberg. Markov chain Monte Carlo simulation methods in econometrics. *Econometric Theory*, 12(3):409–431, 1996.

[17] Although likelihood-based estimators are typically obtained iteratively, and may well use the same initial estimator as the two-stage estimator considered here, the impact of the initial estimator is clearly dissipated as the algorithm converges to the likelihood-based estimate.

[18] These experiments are not reported in the paper. They served merely to confirm the results derived and to ensure that the examples presented were, in fact, correct.

[3] D. Chotikapanich and W. E. Griffiths. Finite sample inference in the SUR model. Working Papers in Econometrics and Applied Statistics No. 103, University of New England, Armidale, 1999.

[4] A. Deaton. Demand analysis. In Z. Griliches and M. D. Intriligator, editors, *Handbook of Econometrics*, volume 3, chapter 30, pages 1767–1839. North Holland, New York, 1984.

[5] D. G. Fiebig and J. Kim. Estimation and inference in SUR models when the number of equations is large. *Econometric Reviews*, 19(1):105–130, 2000.

[6] W. H. Greene. *Econometric Analysis.* Prentice-Hall, Inc., Upper Saddle River, New Jersey, fourth edition, 2000.

[7] G. G. Judge, R. C. Hill, W. E. Griffiths, H. Lutkepohl, and T.-C. Lee. *Introduction to the Theory and Practice of Econometrics.* John Wiley & Sons, New York, second edition, 1988.

[8] T. Kariya and K. Maekawa. A method for approximations to the pdf's and cdf's of GLSE's and its application to the seemingly unrelated regression model. *Annals of the Institute of Statistical Mathematics*, 34:281–297, 1982.

[9] K. V. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate Analysis.* Academic Press, London, 1979.

[10] R. J. Muirhead. *Aspects of Multivariate Statistical Theory.* John Wiley and Sons, Inc., New York, 1982.

[11] D. Percy. Prediction for seemingly unrelated regressions. *Journal of Royal Statistical Society*, 44(1):243–252, 1992.

[12] P. C. B. Phillips. The exact distribution of the SUR estimator. *Econometrica*, 53(4):745–756, 1985.

[13] V. K. Srivastava and T. Dwivedi. Estimation of seemingly unrelated regression equations. *Journal of Econometrics*, 10(1):15–32, 1979.

[14] V. K. Srivastava and D. E. A. Giles. *Seemingly Unrelated Regression Equations Models: Estimation and Inference.* Marcel Dekker, New York, 1987.

[15] V. K. Srivastava and B. Raj. The existence of the mean of the estimator in seemingly unrelated regressions. *Communications in Statistics A*, 8:713–717, 1979.

[16] K. J. White. *SHAZAM User's Reference Manual Version 8.0*. McGraw-Hill, New York, 1997.

[17] A. Zellner. An efficient method of estimating seemingly unrelated regressions and tests of aggregation bias. *Journal of the American Statistical Association*, 57(297):348–368, 1962.

[18] A. Zellner. *Introduction to Bayesian Inference in Econometrics.* John Wiley and Sons, New York, 1971.