Journal of Biometrics & Biostatistics
ISSN: 2155-6180

Oliveira and Teixeira-Pinto, J Biom Biostat 2015, 6:4
DOI: 10.4172/2155-6180.1000256

Research Article Open Access

Analyzing Multiple Outcomes: Is it Really Worth the use of Multivariate Linear Regression?

Rosa Oliveira[1,2,3]* and Armando Teixeira-Pinto[1,2,4]

[1]CINTESIS
[2]Faculty of Medicine, The University of Porto
[3]School of Allied Health Technologies, Polytechnic Institute of Porto, Portugal
[4]School of Public Health,The University of Sydney, Australia

Abstract

In health related research it is common to have multiple outcomes of interest in a single study. These outcomes are often analysed separately, ignoring the correlation between them. One would expect that a multivariate approach would be a more efficient alternative to individual analyses of each outcome. Surprisingly, this is not always the case. In this article we discuss different settings of linear models and compare the multivariate and univariate approaches. We show that for linear regression models, the estimates of the regression parameters associated with covariates that are shared across the outcomes are the same for the multivariate and univariate models while for outcome-specific covariates the multivariate model performs better in terms of efficiency.

Keywords: Ordinary least squares; Correlated errors; Generalized least squares; Monte Carlo

Introduction

In biomedical research it is common that the outcome of interest is characterized by multiple variables rather than a single measure per individual. For example, in clinical trials designed to evaluate the safety and effectiveness of drug-eluting coronary stents, in particular, comparing bare-metal stents with drug-eluting stents we may be interested in multiple outcomes such as myocardial infarction, target vessel revascularization, target lesion revascularization, angiographically verified (definite) stent thrombosis, among others.

A common approach when multiple outcomes are present in a study, is to analyze each outcome independently, in a univariate framework, ignoring the most likely correlation between the outcomes and the multivariate structure of the data. At first glance, this approach may seem less efficient than applying multivariate methods, because it ignores the additional information contained on the correlation between the outcomes. Surprisingly, this is not always the case.

In one of his seminal articles, Zellner [1] introduced the concept of Seemingly Unrelated Regression (SUR) as an extension to the classical multivariate linear regression (MvLR). In the SUR model (2), each outcome of interest is allowed to be associated with it's own set of covariates. If all the outcomes are modeled using the same covariates, the SUR model reduces to the classic MvLR.

The SUR estimator proposed by Zellner [2] was shown to be, in the general case, more efficient than the Ordinary Least Squares (OLS) estimator studied the properties. Later, Srivastava [3] studied SUR model with second-order moments and also found that as the correlation increases, the relative efficiency decreases which agrees with Zellner [4] study. Breiman [5] also investigated different procedures in order for SUR to be more efficient than OLS. All authors concluded that if there is no correlation between the outcomes or if the setting is the same as in the classical multivariate linear regression where all the outcomes are modeled using the same covariates, the coefficients estimators in the multivariate setting is the OLS estimator.

The aim of this article is to show the relative efficiency of the SUR estimator when compared to the OLS estimator, in situations where some covariates are shared by all outcomes but some others are outcome specific. In section two we briefly introduce the SUR model and provide an analytical solution for the SUR estimator in the mixed setting where some covariates are shared and some are not, together with a Monte Carlo simulation for several scenarios. In section 3, two real data examples are introduced to illustrate the results and the final section discusses the implication of these results.

Seemingly Unrelated Regression

A SUR model consists of a set of linear regression equations associating different outcomes $Y_1,...,Y_K$ with covariates in which the errors of the different equations may be correlated (model (2)). In fact, the "seemingly unrelated" expression comes from the apparent lack of relationship between the equations, when each outcome has its own set of covariates, but the equations are related by the possibility of correlation between the error terms. Zellner [2] has shown that separate modeling of the equations results, in the general case, in less efficient estimates.

The SUR model can be compactly written as:

$$\begin{aligned} Y_1 &= \beta_1 X_1 + \varepsilon_1, \\ Y_2 &= \beta_2 X_2 + \varepsilon_2, \\ &\vdots \\ Y_K &= \beta_K X_K + \varepsilon_K, \quad (\varepsilon_1,...,\varepsilon_K) \sim MVN(\mathbf{0},\Sigma). \end{aligned} \tag{1}$$

or

$$\mathbf{Y} = \beta' \mathbf{X} + \varepsilon. \tag{2}$$

where, $\mathbf{Y}$, is a column vector of the stacked outcome and

***Corresponding author:** Rosa Oliveira, Faculty of Medicine, The University of Porto, CINTESIS, Rua Central de Pinheiro, 200; 4510-032 Gondomar, Porto, Portugal; Tel: 00351939355143; E-mail: rcoliveira@med.up.pt

Received September 05, 2015; **Accepted** October 26, 2015; **Published** October 30, 2015

Citation: Oliveira R, Teixeira-Pinto A (2015) Analyzing Multiple Outcomes: Is it Really Worth the use of Multivariate Linear Regression? J Biom Biostat 6: 256. doi:10.4172/2155-6180.1000256

$$X = \begin{pmatrix} X_1 & 0 & \cdots & 0 \\ 0 & X_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & X_K \end{pmatrix}$$

X_k representes the design matrix with the set of covariates associated with outcome Y_k and β is a column vector of coefficients of explanatory variables and ^{e} is a column vector of the error terms.

The usual assumption is that the errors are normally distributed with mean zero and variance-covariance matrix ^{S} given by:

$$\sum = E(\varepsilon\varepsilon^T) = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1K} \\ \sigma_{12} & \sigma_{22} & \cdots & \sigma_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1K} & \sigma_{2k} & \cdots & \sigma_{KK} \end{bmatrix} \quad (3)$$

In SUR models the several equations may be related by the fact of the errors are correlated across equations; and/or a subset of covariates are the same which allows that each of the K outcomes have a different design matrix with some of the covariates being the same, that has particular relevance.

We first describe a more familiar setting in the biostatistics literature. When the equations share the same set of covariates, i.e., $X_1 = ... = X_K$, the SUR model reduces to the classical multivariate linear regression (MvLR). In this setting the error terms associated with each outcome are, again, allowed to be correlated. By ignoring this correlation and fitting K separate regressions this would be equivalent to fitting the SUR model assuming independence of the errors. Therefore, it is obvious that if the errors are not correlated, the SUR estimator is exactly the OLS estimator obtained fitting the K regression separately.

Surprisingly, in the MvLR case (the covariates in each equation are the same) even if the errors are strongly correlated, the SUR estimator always reduces to the OLS. In other words, if the outcomes are modelled using the same covariates, the multivariate model gives the same result (both point estimates and standard errors) as fitting individual regressions for each outcome, despite the level of correlation between the errors. Although this result has been around for many years, it stills raises many eyebrows when is stated, even among experienced biostatisticians. In Appendix A we reproduce the proof of this result.

However, Zellner [1] has shown that in the case that $X_1,...,X_K$ are all different there are gains in efficiency by jointly estimating the β in the SUR model over modelling each outcome separately if the errors are correlated. The increase in efficiency is higher if the correlation between the error terms is strong. In particular in his simulations some gains are observed for $\rho > 0.3$ [1] and explanatory variables in different equations are uncorrelated.

Shared and unshared covariates

We now consider a mixed situation of the two settings described above where the outcomes have some common covariates and other are specific to each outcome. Let's suppose the multivariate linear model:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_K \end{pmatrix} \begin{pmatrix} X_1 & 0 & \cdots & 0 \\ 0 & X_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & X_K \end{pmatrix}$$

$$Y = \beta' X + \gamma' Z + \varepsilon,$$

$$\varepsilon \sim MVN(\mathbf{0}, \Sigma_1)$$

where, $X = (X_1,...,X_k)$ represent the vector of each outcome specific covariate for the k^{th}-outcome and $\mathbf{Z}_j = (Z_1,...,Z_n)$ the K shared covariates by all outcomes, and $\Sigma_1 = \Sigma \otimes \mathbf{I}_n$, with $\sum$ as previously defined.

One result appears in the particular case when $Z_1 = Z_2 = ... = Z_n$. In this case the SUR estimator of β, simplifies to the OLS, this is, if we have the same covariate for all outcomes, the estimator of β that we get from the multivariate model, is exactly the same we get from the univariate regression and, so, the correlation between the error does not play a role in the estimator.

I.e. the correlation between the error terms does not affect the estimation of β and are no efficiency gains when modelling the equations in a multivariate setting compared in applying equation by equation.

Considering what was reported formerly, we want to study the hypothetical gains in the efficiency (SUR compared with OLS) in unshared covariates and in shared covariates coefficients estimates using a multivariate model that takes into account the potential correlation between the error terms when one or more covariates are correlated.

Bearing in mind what was previously reported we expected: to have gains in the efficiency (SUR estimator compared with OLS estimator) in unshared covariates coefficients estimates, and do not have gains in the efficiency of $\hat{\beta}_{SUR}$ when compared with $\hat{\beta}_{OLS}$ in shared covariates coefficients estimates.

Theorem 1 Consider the multivariate linear model:

$$Y = \beta' X + \gamma' Z + \varepsilon,$$

$$MVN(\ ,\ )$$

where, $X = (X_1,...,X_k)$ represent the vector of each outcome specific covariate for the k^{th}-outcome and $\mathbf{Z}_j = (Z_1,...,Z_n)$ the K shared covariates by all outcomes, and $\Sigma_1 = \Sigma \otimes \mathbf{I}_n$, where $\sum$ is the variance-covariance matrix.

The best unbiased estimator is given by:

$$\hat{\beta} = [X'\left(\Sigma^{-1} \otimes \mathbf{I}_n\right)\mathbf{X}]^{-1} X'\left(\Sigma^{-1} \otimes \mathbf{I}_n\right)\mathbf{Y} + (Z'Z)^{-1} Z'Y.$$

Proof.

$$\begin{aligned}
\hat{\beta} &= \left\{\left[\mathbf{X} + (\mathbf{E}_K \otimes \mathbf{Z})\right]'\left(\Sigma^{-1} \otimes \mathbf{I}_n\right)\left[\mathbf{X} + (\mathbf{E}_K \otimes \mathbf{Z})\right]\right\}^{-1}\left[\mathbf{X} + (\mathbf{E}_K \otimes \mathbf{Z})\right]'\left(\Sigma^{-1} \otimes \mathbf{I}_n\right)\mathbf{Y} \\
&= \left\{[X'\left(\Sigma^{-1} \otimes \mathbf{I}_n\right)\mathbf{X}]^{-1} + [(\mathbf{E}'_K\Sigma^{-1}\mathbf{E}_K) \otimes (Z'Z)]^{-1}\right\}[X + \left[\mathbf{X} + (\mathbf{E}_K \otimes \mathbf{Z})\right]'\left(\Sigma^{-1} \otimes \mathbf{I}_n\right)\mathbf{Y}, \\
&\textit{by lemma (1)} \\
&= [X'\left(\Sigma^{-1} \otimes \mathbf{I}_n\right)\mathbf{X}]^{-1}\left[\mathbf{X}(\mathbf{E}_K \otimes \mathbf{Z})\right]'\left(\Sigma^{-1} \otimes \mathbf{I}_n\right)\mathbf{Y} + [(\mathbf{E}'_K\Sigma^{-1}\mathbf{E}_K) \otimes (Z'Z)]^{-1}\left[\mathbf{X} + (\mathbf{E}_K \otimes \mathbf{Z})\right]'\left(\Sigma^{-1} \otimes \mathbf{I}_n\right)\mathbf{Y} \\
&= [X'\left(\Sigma^{-1} \otimes \mathbf{I}_n\right)\mathbf{X}]^{-1} X'\left(\Sigma^{-1} \otimes \mathbf{I}_n\right)\mathbf{Y} + [X'\left(\Sigma^{-1} \otimes \mathbf{I}_n\right)\mathbf{X}]^{-1}(\mathbf{E}_K \otimes \mathbf{Z})'\left(\Sigma^{-1} \otimes \mathbf{I}_n\right)\mathbf{Y} \\
&\quad + [(\mathbf{E}'_K\Sigma^{-1}\mathbf{E}_K) \otimes (Z'Z)]^{-1} X'\left(\Sigma^{-1} \otimes \mathbf{I}_n\right)\mathbf{Y} \\
&\quad + [(\mathbf{E}'_K\Sigma^{-1}\mathbf{E}_K) \otimes (Z'Z)]^{-1}(\mathbf{E}_K \otimes \mathbf{Z})'\left(\Sigma^{-1} \otimes \mathbf{I}_n\right)\mathbf{Y}, \\
&\quad \text{for independents X and Z matrices} \\
&= [X'\left(\Sigma^{-1} \otimes \mathbf{I}_n\right)\mathbf{X}]^{-1} X'\left(\Sigma^{-1} \otimes \mathbf{I}_n\right)\mathbf{Y} \\
&\quad + \left(\mathbf{E}_K^{-1}\Sigma\mathbf{E}_K\right)^{-1} \otimes (Z'Z)^{-1}](E_K \otimes Z)'\left(\Sigma^{-1} \otimes \mathbf{I}_n\right)\mathbf{Y} \\
&= \underbrace{[X'\left(\Sigma^{-1} \otimes \mathbf{I}_n\right)\mathbf{X}]^{-1} X'\left(\Sigma^{-1} \otimes \mathbf{I}_n\right)\mathbf{Y}}_{\hat{\beta}_{SUR}} + \underbrace{(Z'Z)^{-1}Z'Y}_{\hat{\beta}_{OLS}}.
\end{aligned}$$

(complete proof in Appendix B).

Simulation study

We performed a Monte Carlo simulation study to investigate the

efficiency and bias obtained by the univariate and multivariate models, for which the true regression coefficients were known.

For the simulation, we varied the sample size and the correlations between equation errors. 10000 independent samples were generated for the different settings, combining different sample sizes (n=50, n=500, n=1000 and error correlation between the errors (0.0; 0.3; 0.6 and 0.9). Once the 10000 data sets were produced within each correlation range, each set was analysed using both a multivariate linear model, allowing correlation between the error terms, and a univariate linear regression model, ignoring the correlation between the outcomes. The empirical standard errors were obtained by computing the standard deviation of the MLEs for the regression parameters for the set of the simulation. All of the analyses were produced using R 2.11.0 GUI 1.33 Leopard. For each data set, the parameter estimate vectors from the multivariate approach and univariate approach were divided to obtain Relative of the Mean Square Error (RMSDataE).

The model used in each of these data sets was a simple three equation model.

$$Y_1 = 10 + 3X_1 + 5Z + \varepsilon_1; Y_2 = 3 + 4X_2 + Z + \varepsilon_2; Y_3 = 4 - 0.4Z + \varepsilon_3.$$

Y_i represents the outcomes, X_i the specific covariates for each outcome and Z_i the shared covariate by the three outcomes, $i = 1,2,3$. Within each outcome, the error term is independently and identically distributed (Appendix C).

X_{1i} was generated from a $N(2,3)$; X_{2i} generated from a $N(-2,4)$ and Z_i generated from a $N(1,1)$.

The results, means of the estimates for the regression parameters, of the simulations are summarised in Table 1. Overall, the regression estimates were similar, both for the situation of shared and specific covariates for the outcomes and the empirical standard errors were similar to the average of the standard errors obtained in each simulation.

In SUR setting, for the particular case of common set of covariates associated with the outcomes, SUR and OLS performed the same and there was no gain in efficiency of $\hat{\beta}_{SUR}$ when compared with $\hat{\beta}_{OLS}$, despite the correlation between the outcomes, reason why de RMSE is omitted in Table 1. However, concerning parameters estimates associated with specific covariates there is a small efficiency gain in $\hat{\beta}_{SUR}$ when compared with $\hat{\beta}_{OLS}$. This gain was only about 10% (approx.) and occurred when correlation between the outcomes was high, above 0.6 approximately (Table 1).

We obtained similar parameters estimators for both the approaches, nonetheless on the analysis of the unshared covariates coefficients estimates of the efficiency of SUR estimates compared with equation by equation estimates we obtained efficiency gain when the sample is large, that is, SUR was more efficient than OLS. Curiously, when the sample is small and the correlation low, contradictorily, there is loss in the efficiency of $\hat{\beta}_{SUR}$ when compared with $\hat{\beta}_{OLS}$. Concerning the shared covariates, SUR and OLS performed the same and, so, there was not gain in efficiency ($\hat{\beta}_{SUR}$ reduces to $\hat{\beta}_{OLS}$).

Applications

The first example 3.1 illustrates a randomised clinical trial study and shows the similar performance of the approaches described above when the each outcome has one specific covariate and all outcomes share one covariate. The second example, 3.2, illustrates a cohort study looking at cardiovascular risk factors in children. Analogous to the first example each outcome has a specific covariate and both the outcomes share, in this case, two covariates.

Restenosis comparison following coronary stenting using bare-metal stents and drug-eluting stents

Thrombotic events remain the primary cause of death after percutaneous coronary interventions; nonetheless, the growing use of stents has improved the results of percutaneous coronary revascularization [6]. However, bare metal stents cause angiographic restenosis, an undesirable side-effect. Sirolimus-eluting stent implantation, a type of Drug-eluting stents (DESs), were introduced to reduce the incidence of restenosis but there were several reports post drug-eluting stent thrombosis, after more then 360 days.

The data for this example were collected for the SIRolImUS-Eluting Stent in De Novo Native Coronary Lesions (SIRIUS) study. SIRIUS was a randomized, double-bind study conducted in United States with the purpose to assess the safety and efficacy of the sirolimus-eluting stent in the prevention of restenosis in the novo native coronary artery lesions when compared with the uncoated Bx stent.

In brief, the primary endpoint of the study was target vessel failure at 9 months after the procedure. A total of 1101 patients were randomized to either the Cypher sirolimus-eluting stent (n=533) or a bare metal stent (n=525) (43 patients were de-registered).

Procedural success was defined as successful implantation of the study device, a final vessel diameter stenosis <50%. The objective performance criterion was based on three 9 month outcomes: Reference Vessel Diameter (RVD), Post-Stent Diameter Stenosis in Lesion (LPSTDS) and Post-Stent Diameter Stenosis in Stent (SPSTDS) for a more detailed analysis concerning the prediction rule. For each of this outcomes of interest, there are specific covariates, subjects in the study were adjusted for their baselines and a common covariate, all subject in the study shares the group covariate. RVD, SPSTDS and LPSTDS are the specific covariates for RVD9, SPSTDS9 and LPSTDS9 outcomes, respectively. More detailed study design and data collection have been described by Holmes [7].

By analysing the results, the estimates produced by both OLS and SUR are present in Table 2. Most of the estimates are the same or

ρ	Coefficients														
	n=50					n=500					n=5000				
	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
0	1.007	**1.039**	1.001	1.012	**1.053**	1.001	1.000	1.000	1.004	1.000	1.000	1.002	1.000	1.000	1.002
0.3	1.009	**1.039**	1.001	1.009	**1.030**	0.998	0.991	1.000	0.998	0.989	0.997	0.983	1.000	0.994	0.973
0.6	0.996	0.996	1.000	1.000	0.981	0.993	**0.955**	1.000	0.995	**0.948**	0.991	**0.939**	1.000	0.993	**0.931**
0.9	0.988	**0.957**	1.000	0.982	**0.928**	0.987	**0.920**	1.000	0.983	**0.882**	0.980	**0.902**	1.000	0.985	**0.884**

Table 1: Ratio of the mean square error of the multivariate model (SUR) to the univariate models (OLS) averaged. over the results of 10000 simulated datasets with sample size equal to 50 for each correlation level and for data generated with a common covariate.

		Intercept			Shared			Specific		
		Std	Std:E	p value	Std	Std:E	p value	Std	Std:E	p value
RVD	**OLS**	0.489	0.046	<0.001	-0.041	0.015	0.007	0.826	0.016	<0.001
	SUR	0.512	0.046	<0.001	-0.041	0.015	0.007	0.818	0.016	<0.001
LPSTDS	**OLS**	4.553	0.331	<0.001	1.492	0.108	<0.001	0.217	0.064	<0.001
	SUR	4.6	0.214	<0.001	1.477	0.108	<0.001	0.211	0.04	<0.001
SPSTDS	**OLS**	3.109	0.531	<0.001	1.815	0.094	<0.001	0.581	0.078	<0.001
	SUR	3.904	0.335	<0.001	1.804	0.094	<0.001	0.465	0.049	<0.001

Table 2: Coefficients estimates and ratio of the mean square error of the multivariate model (SUR) to the univariate models (OLS) in SIRIUS case study results.

similar to each other. However, the standard errors for the coefficients of unshared covariates differ by a substantial percentage. For example, LPSTDS residuals are minor correlated with RVD as we saw before, but highly correlated with SPSTDS. Just like we expected there were no significant gains in efficiency in SUR estimate compared with OLS estimate in group coefficient estimate, but there was an efficiency gain in LPSTDS coefficient estimate (approximately 38%). The estimates produced by both OLS and SUR are again approximately the same. Just in SPSTDS coefficient estimate we observe differences.

More over SPSTDS residuals are minor correlated with RVD as we saw before, but highly correlated with LPSTDS. So, like in previous case, as we expected there were no significant gains in efficiency in SUR estimate compared with OLS Estimate in group coefficient estimate, but there was efficiency gains in SPSTDS coefficient estimate (approximately 38%).

In the coefficients of shared covariates the estimates gains of SUR compared with OLS aren't exactly zero because errors were assumed as independents but, in the reality, there is some dependence between them (Table 2).

Cardiovascular diseases risk factors in portuguese youngters

The data used in this example were obtained from a large cohort focusing on cardiovascular risk factors in a paediatric population [8]. We focus on one of the longitudinal evaluations of the study and for this example we will use complete and ignore the missing observations. This example does not intend to be a thoroughly analysis of the data or to address a realistic research question. Instead, we build a simple case using known factors that affect blood pressure and contrast the OLS and SUR estimators.

The dataset analysis included 770 children (363 girls and 407 boys). Blood pressure, both systolic and diastolic, demographic and anthropometric measures were available.

We modelled both, systolic (SBP) and diastolic blood pressure (DBP) using age and the z-score of body mass index (BMI). The use of age-standardised BMI is convenient because it is, by construction, independent of the other covariate age. Previous studies [9-13] have found a difference between sexes for the SBP but not for DBP. In this example, the univariate analysis has shown the same relationship, with a significant difference between boys and girls for SBP but not for DBP. Therefore, we used sex as the outcome specific covariate for SBP and we did not include it for the DBP model [14-16]. The final models are then:

$$SBP = age + sex + BMI; DBP = age + BMI$$

The estimates obtained in this last model using SUR and OLS procedures are presented in Table 3. Most of the coefficient estimates are the same or very close to each other as they are minor correlated and sex becomes significant (Table 3).

The final results presented in Table 3 quantify the associations of the two metabolites expressions that have two shared covariates (age and BMI) and one specific covariate (sex). The results show again that the standard errors of the regression coefficients for specific covariates are about 13% less for the SUR method than for the OLS method. It is worth noticing that the standard errors for the other covariates are the same for the SUR and OLS [17,18].

Conclusion

In the present paper two regression methods, multivariate and univariate approach, have been presented to explain a large amount of variance of the prediction. We corroborate that SUR estimator performs better than the equation by equation method of the OLS estimator in estimating a system of regression equations, which are related by their error terms, nonetheless, there were several points regarding these results that deserve special attention. First, it was assumed a normal population, insofar in what way disparities from normality affect the distribution of the coefficient estimators' remains to be studied. Second, if ρ and if the joint estimation technique is applied, the estimator will have a variance slightly larger then when using separate modeling of the equations in samples of moderate size. Third, it is to be noted that in this paper we investigated a problem for which independents X and Z matrices were considered, that is, $X'Z = 0$, which can always happen in real situations. For so, the efficiency of the proposed class has been studied by comparing the two estimators, both with theoretical and practical considerations. With the purpose to validate our study, numerical comparisons were made on one real data set. Improvements upon the regression estimators where

Coefficients													
		Intercept			age			BMI			sex		
		Std	Std:E	p value	Std	Std:E	p value	Std	Std:E	p value	Std	Std:E	p value
SBP	**OLS**	96.862	1.285	<0.001	1.929	0.117	<0.001	3.201	0.285	<0.001	-0.93	0.565	0.0999
	SUR	96.933	1.279	<0.001	1.93	0.117	<0.001	3.202	0.285	<0.001	-1.083	0.493	0.0285
DBP	**OLS**	49.1	1.281	<0.001	1.047	0.12	<0.001	1.305	0.29	<0.001			
	SUR	49.1	1.281	<0.001	1.047	0.12	<0.001	1.305	0.29	<0.001			

Table 3: Coefficients estimates and ratio of the mean square error of the multivariate model (SUR) to the univariate models (OLS) in Cardiovascular Diseases Risk Factors in Portuguese Youngsters study results.

achieved and we concluded that the coefficients associated with the outcome-specific covariates there were efficiency gains depending on the correlation between the outcomes. The results have shown that the standard errors of the SUR estimator is lower than the OLS estimator in outcome specific covariates when the errors are correlated between the equations, and are approximately the same in outcome shared covariates despite the correlation between the errors. In other words, there are gains in the efficiency of SUR model over separate equation by equation when contemporaneous correlation between the errors is high. The gains are obtained when correlation for the residuals was more than 0.6 (approximately). Nevertheless, the efficiency gains were minor (even when the correlation was 0.9 the efficiency gain was about 10). Moreover, even if there were efficiency gains in the estimators, we must remember that as the number of comparisons increases, the probability of making, at least, one type I error in the analysis greatly increases, corroborating the idea of performing a single model, SUR model [19].

Acknowledgments

The authors would like to thank Laura Mauri for SIRIUS data and Jorge Mota, Jose Ribeiro and Sandra Guerra (deceased) for Cardiovascular Diseases Risk Factors in Portuguese Youngters who kindly allowed data access selected for the study for helpful comments, which helped to improve the paper. This research was supported in part by a grant from Fundacão para a Ciência e a Tecnologia (FCT), Rosa Oliveira was supported by Grant PTDC/SAL-ESA/100841/2008. This article was supported by FEDER through Programa Operacional Factores de Competitividade – COMPETE and by National Funds through FCT - Fundação para a Ciência e a Tecnologia within CINTESIS, R&D Unit (reference UID/IC/4255/2013).

Appendix A

Theorem 2 *Consider the multivariate linear model:*

$$Y = \beta' X + \mathrm{e},$$

$$\varepsilon \sim MVN(\mathbf{0}, \Sigma)$$

where, $X = (X_{1k}, \ldots, X_{nk})$ represent the vector of each covariate for the k^{th}-outcome, $\sum$ is the variance-covariance matrix and $\otimes$ the Kronecker product.

If $X_{1k} = \ldots = X_{nk} = X$ The best unbiased estimator is given by:

Proof.

$$\begin{aligned}
\hat{\beta}_{GLS} &= [X'(\Sigma \otimes \mathbf{I}_n)^{-1} X]^{-1} X'(\Sigma \otimes \mathbf{I}_n)^{-1} \mathbf{Y} \\
&= [(\mathbf{I}_m \otimes \mathbf{X})'(\Sigma^{-1} \otimes \mathbf{I}_n)(\mathbf{I}_m \otimes \mathbf{X})]^{-1} (\mathbf{I}_m \otimes \mathbf{X})'(\Sigma^{-1} \otimes \mathbf{I}_n)\mathbf{Y} \\
&= (\Sigma^{-1} \otimes \mathbf{X}'X)^{-1}(\Sigma^{-1} \otimes \mathbf{X}')\mathbf{Y} \\
&= (\mathrm{I}_m \otimes (\mathbf{X}'X)^{-1}\mathbf{X}')\mathbf{Y} \\
&= \hat{\beta}_{OLS}
\end{aligned}$$

Appendix B

Theorem 1 - Complete Proof

Proof.

$$\begin{aligned}
\hat{\beta} &= \left\{\left[\mathbf{X} + (\mathbf{E}_K \otimes \mathbf{Z})\right]'(\Sigma^{-1} \otimes \mathbf{I}_n)\left[\mathbf{X} + (\mathbf{E}_K \otimes \mathbf{Z})\right]\right\}^{-1}\left[\mathbf{X} + (\mathbf{E}_K \otimes \mathbf{Z})\right]'(\Sigma^{-1} \otimes \mathbf{I}_n)\mathbf{Y} \\
&= \{X'(\Sigma^{-1} \otimes \mathbf{I}_n)\mathbf{X} + \mathbf{X}'(\Sigma^{-1} \otimes \mathbf{I}_n)(\mathbf{E}_K \otimes Z) + (\mathbf{E}_K \otimes \mathbf{Z})'(\Sigma^{-1} \otimes \mathbf{I}_n)\mathbf{X} \\
&\quad + (\mathbf{E}_K \otimes \mathbf{Z})'(\Sigma^{-1} \otimes \mathbf{I}_n)(\mathbf{E}_K \otimes Z)\}^{-1}[X + \left[\mathbf{X} + (\mathbf{E}_K \otimes \mathbf{Z})\right]'(\Sigma^{-1} \otimes \mathbf{I}_n)\mathbf{Y}, \\
&\{X'(\Sigma^{-1} \otimes \mathbf{I}_n)\mathbf{X} + \mathbf{X}'[(\Sigma^{-1}\mathbf{E}_K) \otimes Z] + [(\mathbf{E}'_K\Sigma^{-1} \otimes Z')]\mathbf{X} \\
&\quad + [(\mathbf{E}'_K\Sigma^{-1}\mathbf{E}_K) \otimes (Z'Z)]\}^{-1}\left[\mathbf{X} + (\mathbf{E}_K \otimes \mathbf{Z})\right]'(\Sigma^{-1} \otimes \mathbf{I}_n)\mathbf{Y} \\
&= \{X'(\Sigma^{-1} \otimes \mathbf{I}_n)\mathbf{X} + [(\mathbf{E}'_K\Sigma^{-1}\mathbf{E}_K) \otimes (Z'Z)]\}^{-1}\{\left[\mathbf{X} + (\mathbf{E}_K \otimes \mathbf{Z})\right]'(\Sigma^{-1} \otimes \mathbf{I}_n)\mathbf{Y} \\
&\{[X'(\Sigma^{-1} \otimes \mathbf{I}_n)\mathbf{X}]^{-1} + [(\mathbf{E}'_K\Sigma^{-1}\mathbf{E}_K) \otimes (Z'Z)]\}^{-1}\}\left[\mathbf{X} + (\mathbf{E}_K \otimes \mathbf{Z})\right]'(\Sigma^{-1} \otimes \mathbf{I}_n)\mathbf{Y} \\
&\textit{by lemma}\,(1) \\
&= [X'(\Sigma^{-1} \otimes \mathbf{I}_n)\mathbf{X}]^{-1}\left[\mathbf{X}(\mathbf{E}_K \otimes \mathbf{Z})\right]'(\Sigma^{-1} \otimes \mathbf{I}_n)\mathbf{Y} \\
&+ [(\mathbf{E}'_K\Sigma^{-1}\mathbf{E}_K) \otimes (Z'Z)]^{-1}\left[\mathbf{X} + (\mathbf{E}_K \otimes \mathbf{Z})\right]'(\Sigma^{-1} \otimes \mathbf{I}_n)\mathbf{Y} \\
&= [X'(\Sigma^{-1} \otimes \mathbf{I}_n)\mathbf{X}]^{-1}X'(\Sigma^{-1} \otimes \mathbf{I}_n)\mathbf{Y} + [X'(\Sigma^{-1} \otimes \mathbf{I}_n)\mathbf{X}]^{-1}(\mathbf{E}_K \otimes \mathbf{Z})'(\Sigma^{-1} \otimes \mathbf{I}_n)\mathbf{Y} \\
&+ [(\mathbf{E}'_K\Sigma^{-1}\mathbf{E}_K) \otimes (Z'Z)]^{-1}X'(\Sigma^{-1} \otimes \mathbf{I}_n)\mathbf{Y} \\
&+ [(\mathbf{E}'_K\Sigma^{-1}\mathbf{E}_K) \otimes (Z'Z)]^{-1}(\mathbf{E}_K \otimes \mathbf{Z})'(\Sigma^{-1} \otimes \mathbf{I}_n)\mathbf{Y}, \\
&\text{for independents X and Z matrices} \\
&= [X'(\Sigma^{-1} \otimes \mathbf{I}_n)\mathbf{X}]^{-1}X'(\Sigma^{-1} \otimes \mathbf{I}_n)\mathbf{Y} \\
&+ (\mathbf{E}'_K\Sigma^{-1}\mathbf{E}_K) \otimes (Z'Z)]^{-1}(E'_K \otimes Z')(\Sigma^{-1} \otimes \mathbf{I}_n)\mathbf{Y} \\
&= [X'(\Sigma^{-1} \otimes \mathbf{I}_n)\mathbf{X}]^{-1}X'(\Sigma^{-1} \otimes \mathbf{I}_n)\mathbf{Y} \\
&+ (\mathbf{E}_K^{-1}\Sigma\mathbf{E}_K)^{-'} \otimes (Z'Z)]^{-1}(E'_K \otimes Z')(\Sigma^{-1} \otimes \mathbf{I}_n)\mathbf{Y} \\
&= [X'(\Sigma^{-1} \otimes \mathbf{I}_n)\mathbf{X}]^{-1}X'(\Sigma^{-1} \otimes \mathbf{I}_n)\mathbf{Y} + (\mathbf{E}_K^{-1}\Sigma\mathbf{E}^{-'}{}_K E_k{}'\Sigma^{-1}) \otimes (Z'Z)^{-1}Z'\mathbf{Y} \\
&= \underbrace{[X'(\Sigma^{-1} \otimes \mathbf{I}_n)\mathbf{X}]^{-1}X'(\Sigma^{-1} \otimes \mathbf{I}_n)\mathbf{Y}}_{\hat{\beta}_{SUR}} + \underbrace{(Z'Z)^{-1}Z'Y}_{\hat{\beta}_{OLS}}.
\end{aligned}$$

Appendix C

Definition 4.1 *If* $\mathbf{X} \in \mathbb{C}^{\mathbf{m \times n}}$, *then* $\mathbf{X}^-$ *is the unique matrix in* $\mathbb{C}^{\mathbf{m \times n}}$ *such that:*

1. $\mathbf{X}\mathbf{X}^-\mathbf{X} = \mathbf{X}$
2. $\mathbf{X}^-\mathbf{X}\mathbf{X}^- = \mathbf{X}^-$
3. $(\mathbf{X}\mathbf{X}^-)' = \mathbf{X}\mathbf{X}^-$
4. $(\mathbf{X}^-\mathbf{X})' = \mathbf{X}^-\mathbf{X}$

Lemma 1 *If* $\mathbf{X}$ *and* $\mathbf{Y}$ *are two* $m \times n$ *matrices satisfying* $\mathbf{X'Y} = 0$ *and* $\mathbf{XY'} = 0$, *then*

$$(\mathbf{X} + \mathbf{Y})^{-1} = \mathbf{X}^{-1} + \mathbf{Y}^{-1}$$

Proof. We will prove that the four Moore-Penrose conditions are satisfied to demonstrate that $(\mathbf{X^{-1}} + \mathbf{Y^{-1}})$ is the generalized inverse of $(\mathbf{X} + \mathbf{Y})$:

$$\mathbf{X}^{-1}\mathbf{Y} = (\mathbf{X'X})^{-1}\mathbf{X'Y} = 0$$

$$\mathbf{Y}^{-1}\mathbf{X} = (\mathbf{Y'Y})^{-1}\mathbf{Y'X} = 0$$

$$\mathbf{YX}^{-1} = \mathbf{YX'}(\mathbf{XX'})^{-1} = \left[(\mathbf{XX'})^{-1'}\mathbf{XY'}\right]' = 0$$

$$\mathbf{XY}^{-1} = \mathbf{XY'}(\mathbf{YY'})^{-1} = \left[(\mathbf{YY'})^{-1'}\mathbf{YX'}\right]' = 0$$

So, we proved that:

1. $(\mathbf{X}^{-1} + \mathbf{Y}^{-1})(\mathbf{X} + \mathbf{Y}) = \mathbf{X}^{-1}\mathbf{X} + \mathbf{Y}^{-1}\mathbf{Y}$
2. $(\mathbf{X} + \mathbf{Y})(\mathbf{X}^{-1} + \mathbf{Y}^{-1}) = \mathbf{X}\mathbf{X}^{-1} + \mathbf{Y}\mathbf{Y}^{-1}$

Once this two matrices are both symmetric, conditions (iii) and (iv) are accomplished. To prove that conditions (i) and (ii) are satisfied, we only need to post-multiply equations 1 and 2 by $(\mathbf{X}^{-1} + \mathbf{Y}^{-1})$ and $(\mathbf{X} + \mathbf{Y})$ respectively.

References

1. Zellner A (1962) An Efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. Journal of the American Statistical Association 57: 348-368.
2. Zellner A, Huang DS (1962) Further properties of efficient estimators for seemingly unrelated regression equations. International Economic Review 3: 300-313.
3. Srivastava VK (1970) The Efficiency of estimating seemingly unrelated regression equations. Annals of the Institute of Statistical Mathematics 22: 483-493.
4. Zellner A (1971) An introduction to Bayesian inference in econometrics. John Wiley Sons Inc. New York, USA.
5. Breiman L, Friedman JH (1997) Predicting Multivariate Responses in Multiple Linear Regressions. Journal of the Royal Statistical Society, Series B: Statistical Methodology 59: 3-54.
6. Hirshfeld JW Jr, Ellis SG, Faxon DP (1998) Recommendations for the assessment and maintenance of proficiency in coronary interventional procedures: Statement of the American College of Cardiology. J Am Coll Cardiol 31: 722-743.
7. Holmes DR Jr, Leon MB, Moses JW (2003) One-year follow-up of the SIRIUS study: A randomized study with the sirolimus-eluting Bx VELOCITY in the treatment of patients with de-novo native coronary artery lesions. J Am Coll Cardiol 41: 805-813.
8. Guerra S, Ribeiro Jc, Oliveira J, Pinto AT, Twisk JWR, et al. (2003) One-Year stability of cardiovascular diseases risk factors in portuguese youngters. Pediatric Exercise Science 15: 428-439.
9. Sun SS, Grave GD, Siervogel RM, Pickoff AA, Arslanian SS, et al. (2007) Systolic blood pressure in childhood predicts hypertension and metabolic syndrome later in life. Pediatrics 119: 237-246.
10. Cox DR, Wermuth N (1992) Response models for mixed binary and quantitative variables. Biometrika 79: 441-461.
11. Johnson RA, Wichern DW (2001) Applied Multivariate Statistical Analysis. New York: Prentice Hall, USA.
12. Sclove SL (1971) Improved estimation of parameters in multivariate regression. Sankhya A 33: 61-66.
13. Teixeira-Pinto A, Normand SL (2008) Statistical methodology for classifying units on the basis of multiple-related measures. Stat Med 27: 1329-1350.
14. Teixeira-Pinto A, Normand SL (2009) Correlated bivariate continuous and binary outcomes: issues and applications. Stat Med 28: 1753-1773.
15. Srivastava VK, Giles DEA (1987) Seemingly Unrelated Regression Equations Models. New York: Marcel Dekker Inc. New York, NY, USA.
16. Srivastava VK, Maekawa K (1995) Efficiency Properties of Feasible Generalized Least Squares Estimators in SURE Models under Non-normal errors. Journal of Econometrics 66: 99-121.
17. Srivastava VK (1973) The Efficiency of an improved method of estimating seemingly unrelated regression equations. Journal of Econometrics 3: 341-350.
18. Sclove SL (1973) Improved estimation of parameters in multivariate regression. Sankhya: The Indian Journal of Statistics, Series A 33: 61-66.
19. Cutlip DE, Chhabra AG, Baim DS, Chauhan MS, Marulkar S, et al. (2004) Beyond restenosis: five-year clinical outcomes from second-generation coronary stent trials. Circulation 110: 1226-1230.