Regression Methods for Multiple Outcomes in Health Research

Oliveira, Rosa
University of Porto: Faculty of Medicine, Health Information and Decision Sciences
Alameda Prof. Hernani Monteiro
4200-319 Porto,Portugal
E-mail: rcoliveira@med.up.pt

Teixeira-Pinto, Armando
University of Porto: Faculty of Medicine, Health Information and Decision Sciences
Alameda Prof. Hernani Monteiro
4200-319 Porto,Portugal
E-mail: tpinto@med.up.pt

Abstract

We study the gains in efficiency of multivariate regression compared to multiple univariate regressions when the set of covariates are specific to each outcome. In particular, we analyze the situation where the outcomes share some of the covariates but also depend on specific covariates not shared by all outcomes. We demonstrate that for the coefficients associated with shared covariates there are efficiency gains, while for the outcome-specific covariates the efficiency gains depend on the correlation between the outcomes.

Introduction

In research problems, particularly in health research studies, it is common to collect multiple outcomes in order, for example, to examine therapeutic efficacy, treatment effectiveness or associations with various covariates of interest. For instance, in what concerns major adverse cardiovascular events after stenting, where several variables are measured in order to understand the outcome of interest, for instance myocardial infarction, vascular access site complications, stroke and contrast agent nephropathy to target vessel revascularization.
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 in the sense that such approach ignores the additional information contained on the correlation between the outcomes. Surprisingly, this is not always true. Zellner (1962) studied the general case where each outcome has its own set of covariates and designated this approach as Seemingly Unrelated Regression (SUR) models.
Considering the following setting:
$Y_1 = \boldsymbol{\beta_1} X_1 + \boldsymbol{\epsilon}_1$
$Y_2 = \boldsymbol{\beta_2} X_{22} + \boldsymbol{\epsilon}_2$
...
$Y_k = \boldsymbol{\beta_k} X_k + \boldsymbol{\epsilon}_k$.
In his work, he showed that if the errors $\boldsymbol{\epsilon}_1$, ..., $\boldsymbol{\epsilon}_k$ are correlated, the estimates that are obtained by modeling the outcomes jointly have smaller standard errors that the estimates obtained by ordinary least squares (OLS), i.e., the estimates obtained by treating the outcomes as independent and fitting each submodel separately. However, for the particular case where $X_1 = X_2 = ... = X_k$ this is no longer true. In fact, if the covariates are the same for all the outcomes, the maximum likelihood estimates (MLE) obtained in the joint model are exactly the OLS despite the correlation between the error terms or, in other words, the correlation does not carry any additional information for the estimators.
Our aim is to study the mixed situation where some covariates are shared by all outcomes but some others are unshared. We show the analytical estimates for the parameters associated with share and unshared covariates and present a Monte Carlo simulation for different settings.

The mix setting

To find appropriate constrains to the parameters we study the relationship of the gain/loss of efficiency of coefficients estimates in sets of equations of multiple multivariate regression when one or more covariates are correlated.
Consider the multivariate linear model:

$$\mathbf{Y} = \boldsymbol{\beta}^T Z + \boldsymbol{\gamma}^T \mathbf{X} + \boldsymbol{\epsilon},$$
$$\boldsymbol{\epsilon} \sim MVN(\mathbf{0}, \Sigma)$$

where, $\boldsymbol{Z}_j = (Z_{j1}, ..., Z_{jk_w^j})$ represent the vector of each outcome specific covariate for the j^{th}-outcome and $X = (X_1, ..., X_{k_z})$ the k_z shared covariates by all outcomes.
Considering what was previously reported, we want to develop a multivariate model that takes into account the potential correlation between the error terms when one or more covariates are correlated and study the hypothetical gains in the efficiency (SUR compared with OLS) in unshared covariates and shared covariates coefficients estimates.
We consider a set of individual linear multiple regression equations, each one explain-

ing a particular outcome. However this model is not appropriate if the outcomes are associated with different covariates. In this case all equations should be considered simultaneously taking into account the covariance structure of the residuals (Zellner, 1962; Srivastava, 1973).

In answering how much efficiency is gained by using GLS instead of OLS, Zellner(1962) has shown that there were gains in efficiency of SUR model over separate equation by equation. That efficiency would be attained when contemporaneous correlation between the disturbances is high and explanatory variables in different equations are uncorrelated. He found that definite gains are obtained for all sample sizes when $\boldsymbol{\rho} > 0.3$ where $\boldsymbol{\rho}$ is the contemporaneous correlation for the disturbances in the two equations, as well as the correlation increases, the relative efficiency decreases.

We obtained similar parameters estimators for both the approaches. In the unshared covariates, MSE for the parameters was smaller using multivariate approach when compared with univariate in the case correlation was high, i.e., SUR was more efficient than OLS. Concerning the shared covariates, SUR and OLS performed the same and, so, there was no gain in efficiency.

Simulation

We performed a Monte Carlo simulation study to investigate the efficiency for estimates obtained by the multivariate regression and multiple univariate regressions models. Three simulations studies were conducted, in such a way that sample size was small ($\mathbf{n} = 50$), medium ($\mathbf{n} = 500$) or large($\mathbf{n} = 1000$) and error correlation varied from 0.0 till 0.9. In each of the three simulation studies, data is generated randomly and the methods are compared.

For each level of correlation 10000 independent samples were generated and data were modeled using an univariate approach, ignoring the correlation between the outcomes and with multivariate approach, assuming the correlation between the outcomes.

This simulation study is a three-equation model with correlated errors and, to compare the different methods empirically, we simulated data in two different ways, considering different values of correlations between the errors, for which the true regression coefficients were known.

We aimed to investigate efficiency for estimates obtained by the univariate and SUR models in the particular case, where z is a shared covariate for all outcomes and X is a specific covariate:

$\hat{Y_1} = \hat{\boldsymbol{\beta}}_{\mathbf{0}} + \hat{\boldsymbol{\beta}}_{\mathbf{1}} X_1 + \hat{\boldsymbol{\beta}}_{\mathbf{2}} Z$

$\hat{Y_2} = \hat{\boldsymbol{\beta}}_{\mathbf{3}} + \hat{\boldsymbol{\beta}}_{\mathbf{4}} X_2 + \hat{\boldsymbol{\beta}}_{\mathbf{5}} Z$

$\hat{Y_3} = \hat{\boldsymbol{\beta}}_{\mathbf{6}} + \hat{\boldsymbol{\beta}}_{\mathbf{7}} Z.$

The error term is independently and identically distributed as $\boldsymbol{\epsilon} \sim MVN(\mathbf{0}, \Sigma)$.
Table 1 shows the means of the estimates for the regression parameters using simulated data with different levels of correlation ($\boldsymbol{\rho} = 0.0;\ 0.5; 0.9$) between the outcomes. The parameters estimates were similar for all the approaches, both for the situation of shared and specific covariates for 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. These values were similar to the average of the standard errors obtained in each simulation.
When the outcomes share the same covariates, coefficients for the shared parameters showed no gains in efficiency of SUR estimator compared with OLS estimator. On the other hand, concerning the parameters for the specific covariates, on the analysis of efficiency of the SUR estimates compared with the equation by equation estimates we obtained efficiency gain. Nevertheless that gains were obtained when the residuals correlation was hight (above 0.6 approximately). Moreover, that gain was only about 10%.

$\boldsymbol{\rho}$	$\hat{\boldsymbol{\beta}}_\mathbf{0}$	$\hat{\boldsymbol{\beta}}_\mathbf{1}$	$\hat{\boldsymbol{\beta}}_\mathbf{2}$	$\hat{\boldsymbol{\beta}}_\mathbf{3}$	$\hat{\boldsymbol{\beta}}_\mathbf{4}$	$\hat{\boldsymbol{\beta}}_\mathbf{5}$	$\hat{\boldsymbol{\beta}}_\mathbf{6}$	$\hat{\boldsymbol{\beta}}_\mathbf{7}$
0,0	1,007	1,039	1,001	1,012	1,053	1,000	1,000	1,000
0,5	1,000	1,016	1,002	1,000	1,014	1,001	1,000	1,000
0,9	0,988	0,957	1,000	0,982	0,928	1,000	1,000	1,000

Table 1: Mean Square error (MSE) from the simulation study: ratio of the MSE of the multivariate models to the univariate models. Results obtained from 10000 samples of size 1000 for each correlation level.

Conclusion

In the setting of the seemingly unrelated regressions and for the particular case of common set of covariates associated with the outcomes, the OLS is still the best linear unbiased estimator, despite the correlation between the outcomes. The estimates of the parameters associated with the specific covariates of each outcome show a small gain in efficiency when compared with the univariate approach, however only for high correlation (above 0.6 approximately) between the outcomes, i.e. SUR is more efficient than OLS. This suggests that if one foresees that different covariates may be associated with the outcomes, the multivariate approach offers advantages. Our result is a combination of these two methods.

References

1. Breiman L, Friedman JH: Predicting Multivariate Responses in Multiple Linear Regression. Journal of the Royal Statistical Society, Series B: Statistical Methodology 1997, 59:3-54.
2. Johnson RA, Wichern DW: Applied Multivariate Statistical Analysis. New York: Prentice Hall; 2001.
3. Sclove SL: Improved estimation of parameters in multivariate regression. Sankhya A 1971, 33:61-66.
4. Srivastava VK: The efficiency of estimating seemingly unrelated regression equations. Annals of the Institute of Statistical Mathematics 22 (1970), pp. 483-493.
5. Srivastava VK: The efficiency of an improved method of estimating seemingly unrelated regression equations. Journal of Econometrics 3 (1973), pp. 341-350.
6. Teixeira-Pinto A, Normand SLT: Statistical methodology for classifying units on the basis of multiple-related measures. Stat Med 2008, 27(9):1329-1350.
7. Teixeira-Pinto A, Normand S-L: Correlated Bivariate Continuous and Binary Outcomes: Issues and Applications. Stat Med 2009, in press.
8. Zellner A: An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. Journal of the American Statistical Association 57(298) 1962 : 348-368.2.
9. Zellner A, Huang D.S: Further properties of efficient estimators for seemingly unrelated regression equations. International Economic Review 3(3) 1962: 300- 313.