New Parameter Reduction Methods

Simple Estimators of Coefficients of Interest in OLS Regression

James Ryan

McGill University - Department of Economics

Contact Information:

Department of Economics

McGill University

855 Sherbrooke Street West, Montréal, QC H3A 2T7

Phone: +1 (917) 881 4626

Email: james.ryan2@mail.mcgill.ca

Introduction

In least-squares regression models, often an investigator wishes to focus on one or a few parameters out of the full set of regression coefficients. Omitting variables may bias the estimate of the parameter of interest, while including too many regressors leads to efficiency loss. Investigators may often select variables on an a priori basis, which may exclude important regressors, or use other modeling procedures which apply penalties to the regressors [4] in order to shrink and select regressors. A new method [2] may be employed which defines a small number of regressors of interest, computes a matrix of principal components from a matrix of remaining regressors, and estimates the parameter of interest from a constant, the variables of interest, and principal components chosen based on mean-squared error (MSE) or Akaike information criterion (AIC). Small-sample theoretical trials have demonstrated the effectiveness of this new method in estimating parameters of interest, and compares favourably with several other methods.

The focus of this study is to perform real-world analysis on empirical studies by other authors in order to determine the effectiveness of the new methods in larger sample sizes and in cases where parameters of interest must be computed. The project first conducts a literature review in order to identify studies where the new methods may be employed. From there, we examine two papers in greater detail, henceforth referred to as Barro-Lee [1] and Eminent Domain [3], and perform the new methods on these authors data sets to obtain coefficient and standard error estimates over several limiting factors. The results are promising for proving the effectiveness of the new methods, and provide a framework for further empirical analysis in the future.

Main Objectives

1. Perform a review of econometric literature in order to determine candidate studies upon which new the new method may be tested
2. Further isolate studies and compile more in-depth information about modelling and obtain data sets cited in the papers presented
3. Write and troubleshoot code which executes the new methods in MATLAB and R, as well as codes for graphical printouts
4. Obtain preliminary results for comparison with other parameter reduction methods at a later point in time

Algebraic Methods

Here, we outline some of the formulae used in the new methods for empirical analysis.

Let y_i denote the ith observation and let x_i denote the variable of interest which determines the ith observation. Additionally, denote a matrix W of other variables afecting y with ith row W_i' and error term e_i. The least-squares regression model is given as

$$y_i = c + \beta x_i + W_i'\gamma + e_i \tag{1}$$

with coefficient of interest β and parameter vector γ. A fully articulated model of the process y is not necessary; thus, since β is required while γ is not, consistent estimation can be achieved so long as the model includes regressors which span the same space as columns of W with nonzero coefficients and which are not orthogonal to x. Hence, β may be constructed from the columns of W by principal components.The regression model then becomes

$$y_i = c + \beta x_i + C_i'\nu + e_i \tag{2}$$

where C_i' is a matrix containing $\bar{k}$ principal components of W. Although W may have many columns, $\bar{k}$ may remain small while containing all the necessary information such that estimates of β are consistent across (1) and (2).

To obtain the feasible estimator, consider some finite number $d(n)$ of observables, and without loss of generality let W denote the $n \times d(n)$ matrix of variables. Then, consider an integer sequence $k(n)$ and corresponding set $Z_{k(n)}$ of $k(n)$ n-dimensional vectors obtained from the $d(n)$ observables in W. Additionally, assume that $Z_{k(n)}$ contains a $n \times 1$ vector of ones. Thus, regression of y on X and $Z_{k(n)}$ provides

$$\hat{\beta}_{Z_{k(n)}} = arg\ min_\beta (M_{Z_{k(n)}y_n} - M_{Z_{k(n)}}X_n\beta)'(M_{Z_{k(n)}y_n} - M_{Z_{k(n)}}X_n\beta) \tag{3}$$

$$= (X'M_{Z_{k(n)}}X)^{-1}(X'M_{Z_{k(n)}y}) \tag{4}$$

AIC estimates yield some $k^* \leq k$ such that $k^*(n) \subset k(n)$ to obtain the feasible estimator. Hence, the corresponding set $Z_{k^*(n)}$ of $k^*(n)$ n-dimensional vectors obtained from W, with an $n \times 1$ vector of ones, replaces $Z_{k(n)}$ in equations (3) and (4).

Data

Tables 1 and 2 present the k^* values, intercepts, and unbiased estimators of the variable of interest for the Barro-Lee and Eminent Domain data sets, respectively. The "Variable" column identifies variables of interest identified by the original authors upon which the new methods were tested.

Variable	k^*	Intercept	β
log(GDP)	2	0.03157	0.00096
Investment/GDP	2	0.02683	0.06345
Gov't/GDP	2	0.54642	-0.10192
Black Market	2	0.05134	-0.07183
Instability	6	0.04600	-0.05104

Table 1: Barro-Lee Results
k=60,n=90

Variable	k^*	Intercept	β
Case-Schiller	33	3.10503	0.18199
FHFA Index	33	0.36138	0.04942
Non-Metro	33	-36.7621	0.03042
GDP	31	-3.16444	0.00072

Table 2: Eminent Domain Results
k=97,n=3984

The Barro-Lee study focuses on the effect of the five listed variables on average growth rate of GDP per capita across 90 countries from 1965 to 1985. The variable "Black Market" refers to the black market premium in each country, and "Instability" refers to a metric of political instability across each country. The remaining variables identify the natural logarithm of GDP, investment as a proportion of GDP, and government expenditure as a proportion of GDP. Additional variables include life expectancy, population, access to education, and several others. Low k^* values indicate lesser importance of additional variables in predicting outcomes. Meanwhile, the Eminent Domain study studies the effect of the four listed variables - Case-Schiller, FHFA, and Non-Metro home price indices, as well as local GDP - on the number of Eminent Domain cases in the United States which filed in favour of the plaintiff. Additional variables include the number of cases heard by an appeals court, and descriptive characteristics of the judges (e.g. race, party affiliation, age, etc.). Higher k^* values indicate greater importance of additional variables in predicting outcomes.

Additionally, trials were conducted on the two data sets by manually ranging k from 1 to 50 and obtaining coefficient and standard error estimates for each variable of interest, as opposed to selecting a value of k^* and corresponding intercept and β values via AIC selection criterion. The results are in figures (1) through (4).

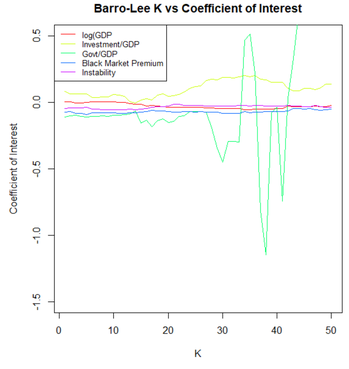

Figure 1: Barro-Lee Trials, $k \in [1, 50]$

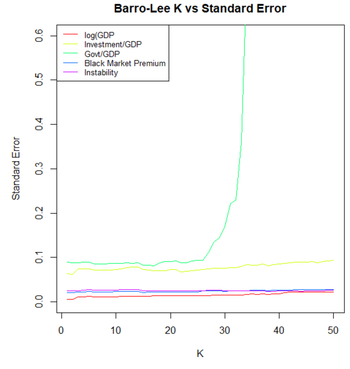

Figure 2: Barro-Lee Trials, $k \in [1, 50]$

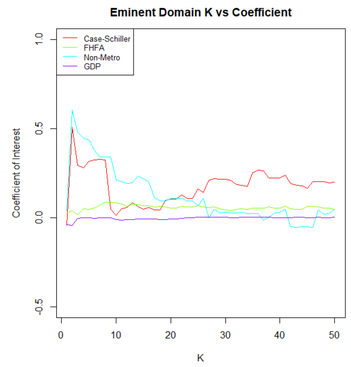

Figure 3: Eminent Domain Trials, $k \in [1, 50]$

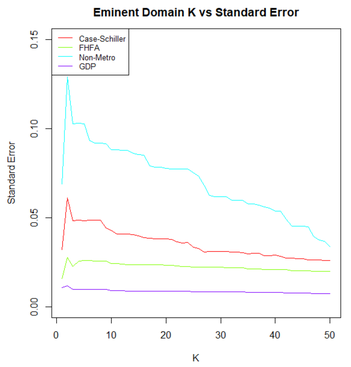

Figure 4: Eminient Domain Trials, $k \in [1, 50]$

For the Barro-Lee trials, coefficient estimates and standard errors are stable for low values of k, indicative of the optimality of k^* as computed. As k increases, the estimates become less stable, suggesting that the value of k^* for each variable yields a better predictor than obtained by including more elements of W. In the Eminent Domain case, coefficient estimates again stabilize as k tends towards k^*, and are more volatile for lower values of k. These results point to the new methods yielding unbiased estimators in most cases, and are overall promising for large sample sizes.

Conclusions

- The methods being tested generally result in effective and stable estimators for the variable of interest in both cases, indicating that parameter reduction may be useful in these cases for better identifying the effect of a particular variable
- Indicating more or less components than the optimally chosen amount result in less viable unbiased estimators, suggesting that selection of components of W based on AIC criterion results in effective OLS modeling
- Parameter reduction may be applied in this fashion to the cases listed in order to produce more accurate and simplified models, without losing critical information

Forthcoming Research

Further work must be done to compare the new methods to alternative methods of parameter reduction, relying on shrinkage and selection of variables, in order to determine their effectiveness. Moreover, additional trials on other data sets will allow for a larger base upon which to determine the effectiveness of the results. Nevertheless, the current results are promising for the effectiveness of the new methods, and provide a solid foundation for further empirical analysis.

References

[1] Robert J. Barro and Jong-Wha Lee. Sources of economic growth. *Carnegie-Rochester Conference Series on Public Policy*, 40:1–46, June 1994.

[2] John W. Galbraith and Victoria Zinde-Walsh. Simple and reliable estimators of coefficients of interest in a model with many potential confounding effects. pages 1–37, February 2018.

[3] Giorgio E Primiceri, Giannone, Domenico, and Lenza, Michele. Economic Predictions with Big Data: The Illusion of Sparsity. page 25.

[4] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1996.

Acknowledgements

Special thanks to Prof. John W. Galbraith for his guidance and assistance on this project, the McGill Arts Internship Office for their financial assistance, and Mr. Harry Samuel for his contributions to this award.