Robustness of Econometric Variable Selection Methods

Bernd Brandl

University of Vienna, Austria
bernd.brandl@univie.ac.at

1 Introduction

Variable selection in cross-country growth regression models is currently a major open research topic and has inspired theoretical and empirical literature, see [6]. There are two categories of research problems that are intimately connected. The first problem is model uncertainty and the second is data heterogeneity. Recent literature aims to overcome the first problem by applying Bayesian Model Averaging (BMA) approaches in finding important, robust and significant variables to explain economic growth. While BMA offers an appealing approach to handle model uncertainty very little research has been undertaken to consider the problem of data heterogeneity. In this paper we analyze the issue of data heterogeneity on the basis of the exclusion of countries, i.e. we will take a closer look at the robustness of approaches when countries are eliminated from the data set. We will show that results of BMA are very sensitive to small variations in data. As an alternative to BMA in the cross-country growth regression debate we suggest the use of "classical" Bayesian Model Selection (BMS). We will argue that there is much in favor of BMS and will show that BMS is less sensitive in the identification of important, robust and significant variables when small variations in data are made. Our empirical results are undertaken on the most frequently used data set in the cross-country growth debate provided by [4].

2 The Cross-Country Growth Regression Debate

The problem encountered in the literature on the empirical validity of determinants of cross-country economic growth is that the number of explanatory variables is large compared to the number of observations. In traditional literature cross-country regressions of the form

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n X_n + \epsilon \tag{1}$$

were reported. Where y represents the vector of economic growth across countries, a α is a constant, and $x_1, \ldots\ldots x_n$ are vectors of explanatory variables. In each paper

the set of explanatory variables differs according to different theoretical and technical considerations. For more details see also [4] and [2]. The aim of empirical research is to determine the importance of variables and specifications. In the context of economic growth, a regression with all variables that are argued to be important is simply not feasible for reasons of sample size and collinearity. The infeasibility inspired the development of data-driven approaches, see [1] and [3]. In recent literature the Bayesian Averaging of Classical Estimates (BACE) approach suggested by [4] appears to be most promising. Independent of using BMA or traditional regressions all literature makes the implicit assumption that the parameters are constant across countries, as stressed for example by [6]. This assumption is too strong as there are many examples where coefficients can be argued to vary extremely over countries. In the following we will observe the problem of parameter (or coefficient of variables) heterogeneity by a diagnostic analysis where we are excluding from the sample one country at a time. We will make a comparison of the robustness in the identification of important variables by applying the BACE approach and the BMS approach by optimizing the well-known Bayesian Information Criteria (BIC), proposed by [5].

3 Bayesian Model Averaging Versus Bayesian Model Selection

In the cross-country growth discussion the problem of selection and averaging arises when the relationship between y and a subset of x has to be modeled, but there is uncertainty about which subset to use. Model selection refers to using the data to select one model from the list of candidate models $M_1, \ldots, M_K$, while model averaging refers to the process of estimating some quantity under each model M_j and then averaging the estimates according to how likely each model is. This means that model M_j is used together with the data to predict $\hat{y_j}$. This implies an overall prediction $\sum_{j=1}^{k} w_j \widehat{y}_j$ where a particular weight w_j is used to express the probability that the model M_j generated the data. The BMA way is to compute the posterior probability for each model $P\left(M_j \mid Data\right)$. In the case of model selection, a model that maximizes $P\left(M_j \mid Data\right)$ is chosen. As our interest is not only in models but specifically on the importance of variables we will apply the frequently applied BACE approach to compute posterior probabilities and apply a genetic algorithm to solve the maximization problem. In contrast to other (pure) BMA approaches which demand the specification of a prior distribution for all parameters, the BACE approach requires only specification of the expected model size $\bar{k}$. The simple idea behind the BACE approach is to combine averging of estimates across models with ordinary least squares (OLS) estimation. In the BACE approach, the posterior inclusion probability (PIP) provides information about the relevance of a variable i and the sum of the posterior probabilities of all models containing variable i is calculated via

$$P\left(\beta_i \neq 0 \mid y\right) = \sum_{j=1}^{2^K} P\left(M_j \mid y\right). \tag{2}$$

Where $P\left(M_j\right)$ is the posterior model probability, i.e. the probability distribution of model M_j given the data y. It is calculated as the proportional likelihood function corrected for the degrees of freedom