Applying standard and semiparametric Bayesian IV on health economic data

Sylvia Frühwirth-Schnatter[1], Martin Halla[2], Alexandra Posekany[1], Gerald Pruckner[2], Thomas Schober[2]

Bayesian Young Statisticians Meeting (BAYSM), Milan June, 5-6, 2013
Paper no. 34

[1] WU Vienna University of Economics and Business, Institute of Statistics and Mathematics, Vienna, Austria
sylvia.fruehwirth-schnatter@wu.ac.at
alexandra.posekany@wu.ac.at
[2] Johannes Kepler Universtät, Department of Economics, Linz, Austria
martin.halla@jku.at
gerald.pruckner@jku.at
thomas.schober@jku.at

Abstract

This paper compares 3 inferential approaches applied to a very large and challenging data set from health economics. The chosen instrumental variable model aims for determining the causal effect of family size on labour and health outcomes. Contrasting frequentist and Bayesian parametric and semiparametric approaches, we find that the parametric version outperforms the rest regarding computational efficiency and estimation precision.

Keywords: instrumental variables; Bayesian econometrics; high dimensional data analysis; MCMC

1 Introduction and Model

Bayesian methods have gained interest and importance in various fields of economics, see [5]. This paper presents an example from health economics, a discipline dealing with medical data from health or health care on a micro- or macroeconomic level. A large data set containing information about 200,000-300,000 individuals forms the basis of our investigations. Our intention is to determine possible causal effects of family size on labour outcomes, like wages

and employment, and health outcomes, such as expenses on drugs and doctors or days spent in the hospital.

However, since the family size is expected to be an endogenous regressor, i. e. correlated with the noise, we employ the tool of instrumental variable (IV) analysis. In the Bayesian IV model the data is generated according to the following system of model equations, e. g. [1]:

$$\begin{aligned} x &= z'\delta + \varepsilon_1 \\ y &= \beta x + w'\gamma + \varepsilon_2 \end{aligned}$$

$$Cov(\tilde{\varepsilon_1}, \tilde{\varepsilon_2}) = C\Sigma C^T \qquad C = \begin{pmatrix} 1 & 0 \\ c & 1 \end{pmatrix} \quad \Sigma = \begin{pmatrix} \sigma_1^2 & \rho \\ \rho & \sigma_2^2 \end{pmatrix}$$

The second equation is the usual linear regression model equation with regressor variables x and covariates w, where variable x is endogenous. Therefore, the first equation is introduced to add variation to x, which is independent of ε_2, through the so-called instrument z. In our application, the endogenous variable x is the number of children in families with 2 or more children and the instrument z is an indicator whether twins are born at the second birth. As covariates, we use the information about the mother's age at first or second birth, the sex of the first born child and the age at the time of investigation for health and labour outcomes.

In the standard model, [1], a bivariate normal distribution is assumed for the error vector $(\varepsilon_1, \varepsilon_2)$. This model can be inferred straightforwardly either using standard 2-stage least squares inference or a Bayesian inference could be implemented applying standard Gibbs sampling, for which detailed descriptions are provided in [1] and [3]. Our complex data however violate the normal distribution assumption of this model, as we are faced with either binary or highly skewed data, including excess zeros. Following [2], we therefore investigate in addition whether a semi-parametric approach based on a generalised nonparametric error model is able to deal with this departure from normality. [2] suggested an infinite mixture of normals, which is realized via Dirichlet process priors, as error model. A MCMC realisation for both models is implemented in the R package 'bayesm', see [4].

However, we encountered problems with the bayesm Dirichlet process sampler, as the algorithm turned out to be unable to handle such large data sets. Thus, we developed an alternative approach of splitting the data set into roughly a dozen smaller data sets $D_1, \ldots, D_K$. We run MCMC estimation on each single data set D_k and perform a post-inference resampling step to sequentially merge the K posterior densities into a single one. Given draws from the merged posterior sample $D_1, \ldots, D_{k-1}$, we use these merged draws as a proposal in a Metropolis Hastings algorithm. Here, the likelihood ratio of the k's sample D_k defines the acceptance rate based on

$$\frac{p(\theta^{new}|D_1, \ldots, D_k)}{p(\theta^{old}|D_1, \ldots, D_k)} = \frac{p(D_k|\theta^{new})p(\theta^{new}|D_1, \ldots, D_{k-1})}{p(D_k|\theta^{old})p(\theta^{old}|D_1, \ldots, D_{k-1})}$$

where $\theta = (\delta, \beta, \ldots)$ summarises all parameters of interest. We repeat this procedure for $k = 2, \ldots, K$ to obtain final draws from the joint posterior $p(\theta|D_1, \ldots, D_K)$.

2 Results

In our analysis, we compare these two Bayesian approaches, namely a Gibbs sampler based on normal distribution assumptions, and the Dirichlet process sampler of infinite mixtures of normals against frequentist inference of the instrumental variable model, based on two-stage least squares. Table 1 summarises our findings. The Gaussian Bayesian IV model outperforms two-stage least squares by far, providing considerable improvement regarding the size of confidence intervals, the accuracy and precision of estimates.

outcome	freq. IV	Bayes IV	Bayes DP IV
Some college education	-0.031	-0.04	0.0002
	[-0.067,0.004]	[-0.046,-0.035]	[-0.0017,0.0012]
Employed	-0.023	0.009	0.0009
	[-0.057,0.011]	[0.004,0.014]	[-0.0019,0.0022]
White-collar worker	0.003	-0.029	-0.0008
	[-0.038,0.045]	[-0.036,-0.023]	[-0.0034,0.0014]
Wage	2.765	-1.45	2.72
	[-0.517,6.047]	[-1.98,-0.93]	[-6.74,11.79]
Total Expenditures on health	12.249	7.97	-60.31
	[-67.446,91.944]	[-3.14,18.82]	[-281.17,176.78]
Expenditures on doctors	-7.024	3.44	-31.87
	[-20.598,6.551]	[1.02,5.83]	[-162.65,83.46]
Expenditures on drugs	-15.910	-5.97	-0.89
	[-36.388,4.569]	[-11.01,-0.97]	[-12.89,12.27]
Days in hospital	0.191	0.09	0.05
	[-0.113,0.494]	[0.029,0.15]	[0.039,0.061]

Table 1: Summary of the results of the IV analysis. The first column contains the frequentist IV results, obtained in STATA, the second the ones for Bayesian IV with normal distribution (Gibbs sampler) and the third column results for the nonparametric Dirichlet process sampler. The intervals are the 95% confidence intervals for IV and the maximum posterior density interval covering 95% of the data in case of the Bayesian approaches.

Interestingly, despite the size of the data set or because of the size of the data set, the semi-parametric approach is not able to handle the data better than the Gibbs approach. In case of binary data, which we observe for labour data such as employment status, the Dirichlet process IV results in non-significant outsomes, almost exactly 0, while the Gibbs sampling approach is able to identify

an effect of the respective variables. Furthermore, the Dirichlet sampling method is disturbed by excess zeros which we observe for health outcomes, resulting for 2 cases in unreasonably large intervals, which we consider a clear indication that Dirichlet mixtures of normals are not suitable for these data.

To summarise, we present a challenging application of IV estimation to the analysis of large data in health economics. We compared a frequentist against two Bayesian approaches, using parametric and semi-parametric methods. Overall the simple Gibbs sampler assuming Gaussianity performs best, taking into account estimation precision as well as computation time and efficiency.

References

[1] A. Zellner, T. Ando, N. Basturk, L. Hoogerheide and H. Van Dijk. **Bayesian Analysis of Instrumental Variable Models: Acceptance-Rejection within Direct Monte Carlo**; *Econometric Reviews*; 2012; pp. 1-38.

[2] T. Conley, C. Hansen, R. McCulloch and P. Rossi. **A Semi-Parametric Bayesian Approach to the Instrumental Variable Problem** *Journal of Econometrics, Elsevier*; 2008; 144(1); pp. 276-305.

[3] T. Conley, C. Hansen and P. Rossi. **Plausibly exogenous** *The Review of Economics and Statistics*; 2012; 94(1); pp. 260-272.

[4] P. Rossi. **bayesm: Bayesian Inference for Marketing/Microeconometrics** R package version 2.2-5, 2012.

[5] P. Rossi. **Bayesian Statistics and Marketing**; *John Wiley and Sons*; 2005