Economics Letters 122 (2014) 36–39

Contents lists available at ScienceDirect

Economics Letters

journal homepage: www.elsevier.com/locate/ecolet

Bayesian endogeneity bias modeling[☆]

Gabriel Montes-Rojas [a,b,*], Antonio F. Galvao [c]

[a] CONICET–Universidad de San Andrés, Argentina

[b] Department of Economics, City University London, D306 Social Sciences Bldg, 10 Northampton Square, London EC1V 0HB, UK

[c] Department of Economics, University of Iowa, W284 Pappajohn Business Building, 21 E. Market Street, Iowa City, IA 52242, United States

HIGHLIGHTS

- Model for endogeneity bias using a prior distribution for endogenous regressors.
- Method-of-moments and Ridge penalized shrinkage regression estimator.
- Based on imposing a prior on the moment conditions for the endogenous regressors.
- Shrinkage of the endogenous regressor coefficients only.
- We show the estimator's comparison with a Bayesian estimator.

ARTICLE INFO

Article history:
Received 27 September 2013
Received in revised form
25 October 2013
Accepted 26 October 2013
Available online 4 November 2013

JEL classification:
C11
C52

Keywords:
Endogeneity
Shrinkage
Ridge regression
Method of moments

ABSTRACT

We propose to model endogeneity bias using prior distributions of moment conditions. The estimator can be obtained both as a method-of-moments estimator and in a Ridge penalized regression framework. We show the estimator's relation to a Bayesian estimator.

1. Introduction

A solution to the problem of endogeneity is to rely on exogenous information, i.e. instrumental variables (IV) (see e.g. Hausman, 1983; Angrist and Krueger, 2001) and proxy variables (see e.g. Olley and Pakes, 1996; Levinsohn and Petrin, 2003; Wooldridge, 2002, 2009) derived from additional exclusion restrictions or equations. In practice, the type of restriction and the point identification strategy chosen determine the model to be used. Alternative approaches are given in Rigobon (2003), Klein and Vella (2010), Chalak and White (2011), and Lewbel (2012). However, in many empirical applications there is disagreement and concern about the exclusion restrictions imposed. A related approach is to seek parameter set identification. The advantage of such methods is that they require weaker assumptions than those needed for consistent point estimation of the parameter of interest.

This paper suggests an alternative strategy to deal with the endogeneity problem. We propose the use of a prior distribution on the endogeneity bias when there is no availability of additional information such as instrumental or proxy variables. By modeling bias, thus, we mean to impose a prior distribution on the amount of endogeneity of the endogenous variables' coefficient estimators and then compute the distribution of the associated parameters of interest. This distribution reflects the researcher's beliefs about endogeneity. The value of our contribution does not only lie in the

[☆] The authors would like to express their appreciation to the Editor, one anonymous referee, Roger Koenker, Stephane Bonhomme, Susumu Imai, and Ehsan Soofi for helpful discussions and comments regarding this paper. All the remaining errors are our solely responsibility.

[*] Corresponding author at: Department of Economics, City University London, D306 Social Sciences Bldg, 10 Northampton Square, London EC1V 0HB, UK. Tel.: +44 0 20 7040 8919; fax: +44 0 20 70408580.

E-mail addresses: Gabriel.Montes-Rojas.1@city.ac.uk (G. Montes-Rojas), antonio-galvao@uiowa.edu (A.F. Galvao).

0165-1765/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.econlet.2013.10.034

derived estimators but in the way it proposes to think about endogeneity. In particular, it formalizes the use of prior knowledge about unobservables and their relationship with observables to quantitatively assess the degree of bias. The proposed estimators are constructed using the method-of-moments by imposing the prior distribution of a misspecified moment condition, or equivalently, in a Ridge penalized regression framework, by only penalizing the endogenous variables' coefficients. We show the connection of our estimator with a Bayesian estimator.

The proposed methodology is related to recent developments in the literature. First, it can be interpreted as in Altonji et al. (2005a,b, 2008) as a strategy to extract information from observables about the endogeneity bias. They construct an index of observables, which in combination of prior knowledge about the sign of the bias and a condition on the relationship between included (observable) and excluded (non-observable) variables can be used to identify the endogenous variable parameter. In a related work Kiviet (2011) imposes the correlation between the endogenous variable and the innovations. In this case, the identification of the true parameters cannot be achieved, but the direction and magnitude of the bias are analyzed instead. Second, the idea of using penalized regression to model endogeneity was proposed by Galvao and Montes-Rojas (2010) in the context of dynamic panel data models where the fixed-effects' shrinkage reduces the dynamic panel bias.

2. Regression shrinking of endogenous covariates

Consider the following linear regression model,

$$y_i = x_{1i}\beta_1 + x_{2i}\beta_2 + \epsilon_i, \quad i = 1, 2, \dots, n, \qquad (1)$$

where x_{1i} is a $1 \times p_1$ vector, x_{2i} is a $1 \times p_2$ vector, β_1 is a $p_1 \times 1$ vector, and β_2 is a $p_2 \times 1$ vector. x_1 contains p_1 exogenous regressors and x_2 contains p_2 endogenous regressors. Let $x = (x_1, x_2)$ and $\beta = (\beta_1, \beta_2)'$. In matrix notation $\boldsymbol{y} = \boldsymbol{x}_1\beta_1 + \boldsymbol{x}_2\beta_2 + \boldsymbol{\epsilon} = \boldsymbol{x}\beta + \boldsymbol{\epsilon}$ where $\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\epsilon}$ are the corresponding n-dimensional vectors or matrices.

We consider the following assumptions:

Assumption 1. (y_i, x_i) is an i.i.d. random sample for $i = 1, \dots, n$ with $y_i = x_{1i}\beta_1 + x_{2i}\beta_2 + \epsilon_i$.

Assumption 2. $\frac{1}{n}\boldsymbol{x}_1'\boldsymbol{x}_1 \xrightarrow{p} C_1$ (finite and non-singular), $\frac{1}{n}\boldsymbol{x}_1'\boldsymbol{x}_2 \xrightarrow{p} C_{12}$ (finite), $\frac{1}{n}\boldsymbol{x}_2'M_1\boldsymbol{x}_2 \xrightarrow{p} V_{2\cdot1}$ (finite and non-singular), and $\frac{1}{n}\boldsymbol{x}_1'M_2\boldsymbol{x}_1 \xrightarrow{p} V_{1\cdot2}$ (finite and non-singular), where $M_j = I_{p_j} - \boldsymbol{x}_j(\boldsymbol{x}_j'\boldsymbol{x}_j)^{-1}\boldsymbol{x}_j'$, $j = 1, 2$ are the orthogonal projections in least squares.

Assumption 3. For any i, $E[x_{1i}'\epsilon_i] = \mathbf{0}$ and $E[x_{2i}'\epsilon_i] = B_2$, such that $\frac{1}{n}\boldsymbol{x}_1'\boldsymbol{\epsilon} \xrightarrow{p} \mathbf{0}$ and $\frac{1}{n}\boldsymbol{x}_2'\boldsymbol{\epsilon} \xrightarrow{p} B_2$.

Let $\hat\beta^o = (\hat\beta_1^o, \hat\beta_2^o)'$ be the ordinary least squares (OLS) estimator. Simple OLS algebra and asymptotic calculations show that

$$\hat\beta_1^o \xrightarrow{p} \beta_1 - C_1^{-1}C_{12}\delta_2,$$
$$\hat\beta_2^o \xrightarrow{p} \beta_2 + \delta_2,$$

where $\delta_2 = V_{2\cdot1}^{-1}B_2$ is the OLS bias in estimating β_2. The OLS estimator is derived from the set of estimating equations

$$E[x_1'\epsilon] = \mathbf{0},$$
$$E[x_2'\epsilon] = \mathbf{0},$$

and thus the OLS bias is the result of misspecifying the second moment condition, i.e. wrongly assuming x_2 is exogenous, when in fact, $E[x_2'\epsilon] = B_2$, for a general unknown B_2. In particular, consider the following modified moment condition: $V_{2\cdot1}^{-1}E[x_2'\epsilon] = V_{2\cdot1}^{-1}B_2 = \delta_2$ and let $\delta_2 = \Lambda\beta_2$, where Λ is a $p_2 \times p_2$ diagonal matrix diag$\{\lambda_1, \lambda_2, \dots, \lambda_{p_2}\}$. In many empirical applications the signs of β_2, δ_2 and C_{12} are known, and then the sign of the elements in Λ are also known. Note that $\Lambda\beta_2$ has the role of δ_2 above and the greater a parameter λ_j is, the greater the endogeneity problem in x_j is, for $j = 1, 2, \dots, p_2$. Λ can be seen as a *proportional endogeneity tolerance parameter*.

This paper focuses on estimators that impose the prior information about δ_2 or Λ, derived through the moment conditions

$$E[x_1'\epsilon] = \mathbf{0},$$
$$V_{2\cdot1}^{-1}E[x_2'\epsilon] = \Lambda\beta_2.$$

For convenience we maintain the terminology of *exogenous* and *endogenous* variables. Nevertheless, this could be redefined for the researcher convenience.

The moment conditions can be replaced by the following estimating equations:

$$\boldsymbol{x}_1'(\boldsymbol{y} - \boldsymbol{x}_1\beta_1 - \boldsymbol{x}_2\beta_2) = \mathbf{0}, \qquad (2a)$$
$$(\boldsymbol{x}_2'M_1\boldsymbol{x}_2)^{-1}\left(\boldsymbol{x}_2'(\boldsymbol{y} - \boldsymbol{x}_1\beta_1 - \boldsymbol{x}_2\beta_2)\right) = \Lambda\beta_2. \qquad (2b)$$

The solution to this problem, if $(I_{p_2} + \Lambda)$ is non-singular, is

$$\hat\beta_1^\Lambda = \hat\beta_1^o + (\boldsymbol{x}_1'\boldsymbol{x}_1)^{-1}(\boldsymbol{x}_1'\boldsymbol{x}_2)\Lambda(I_{p_2} + \Lambda)^{-1}\hat\beta_2^o,$$
$$\hat\beta_2^\Lambda = (I_{p_2} + \Lambda)^{-1}\hat\beta_2^o.$$

Let $\hat\beta^\Lambda = (\hat\beta_1^\Lambda, \hat\beta_2^\Lambda)'$, one can notice that for this case,

$$\hat\beta_1^\Lambda \xrightarrow{p} \beta_1 - C_1^{-1}C_{12}\delta_2 + C_1^{-1}C_{12}(I_{p_2} + \Lambda)^{-1}(\Lambda\beta_2 - \delta_2),$$
$$\hat\beta_2^\Lambda \xrightarrow{p} (I_{p_2} + \Lambda)^{-1}(\beta_2 + \delta_2).$$

The proposed estimator $\hat\beta^\Lambda$ also arises as a solution to a penalized Ridge regression version of model (1), where $\{\lambda_1, \lambda_2, \dots, \lambda_{p_2}\}$ are penalties that apply to the p_2 endogenous regressors only. In this framework $\hat\beta^\Lambda$ is obtained as

$$\hat\beta^\lambda = \operatorname*{argmin}_{\beta} \sum_{i=1}^{n}(y_i - x_i\beta)^2 + (\boldsymbol{x}_2'M_1\boldsymbol{x}_2)\sum_{j=1}^{p_2}\lambda_j\beta_{2j}^2. \qquad (3)$$

Interpreting Λ as a penalty parameter allows for applications to other contexts in which our ignorance and lack of identification about some parameters determine that their estimation involves a burden for the desirable properties of the full model. Note that for the Ridge regression method, the λ_js are assumed to be non-negative, and thus they are interpreted as shrinkage parameters, that is, larger values of λ shrink the corresponding parameter estimates towards zero. However, for our purposes we could consider $\lambda >< 0$ in order to account for different types of bias.

One important case in (3) is when $\Lambda = \text{diag}\{\lambda\}$ and $\lambda > 0$ is the same positive scalar for all endogenous variables. In this case, all endogenous variables regression parameters are subject to the same penalty, that is, they shrink according to λ. This is the case of Galvao and Montes-Rojas (2010) where all fixed-effects parameters are reduced in a similar fashion. There are two important special cases of the above method. First, the model with no shrinkage, and second, the model that totally shrinks β_2. In the former case, if $\lambda = 0$, then $\hat\beta^o$ is the standard OLS estimator where both x_1 and x_2 are used. In this case, the endogeneity in x_2 produces an *endogeneity bias* in the estimators of β_1 and β_2. In the second case, if $\lambda \to \infty$, then $\hat\beta^\infty$ is the estimator where only x_1 is used and β_2 is set to 0. This will generate an *omitted variable bias* for the estimation of β_1.

3. Imposing a prior on endogeneity

In practice the estimator above is infeasible because Λ (or δ_2) is unknown and the parameters cannot be identified. If Λ is known the parameters are point identified, while if Λ is known to belong to a defined set the parameters are set identified by the range of the tuning parameter. We propose a Bayesian estimator where we impose a prior on the tuning parameter and compute the estimate of the parameters of interest based on the penalized regression. The prior of the tuning parameter thus implicitly defines a weighted set for identification and produces a posterior distribution for an estimator of β. As in a Bayesian context, the properties of the derived estimator depend on how accurate the prior is.

Our uncertainty on β_2 (and δ_2) is modeled by λ, the proportional endogeneity tuning parameter. By modeling bias, thus, we mean to impose a prior distribution on δ_2, the amount of endogeneity, through the parameter λ, and then compute the distribution of the associated parameters of interest. As an example, suppose that $\lambda \sim \chi_1^2$, so that it has mean equal to 1. Then δ_2 is of the same magnitude as β_2, that is, $E[\hat{\beta}_2^o] = \beta_2 + \delta_2 = 2\beta_2$.

The distribution of $\hat{\beta}^\lambda = (\hat{\beta}_1^\lambda, \hat{\beta}_2^\lambda)$ can then be obtained from the distribution of $\hat{\beta}^o$ and λ. Define $h(\beta|y, x)$ as the density function of this estimator. A point estimate can be based on the mean (i.e. $E_\lambda[\hat{\beta}^\lambda]$), median (i.e. $Q_{\hat{\beta}^\lambda}(0.5)$, where $Q(\cdot)$ is the quantile function) or any other statistic of interest such as mode. As a practical matter, $h(\cdot)$ can be simulated by using the asymptotic normality of $\hat{\beta}^o$ and the distribution of λ, which is assumed to be independent of (y, x).

4. Connection with a Bayesian estimator

In order to see the connection of our estimator with a Bayesian estimator, consider the loss function in a standard Ridge regression problem

$$L(\beta, \lambda) = \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda|\beta|^2.$$

Consider now a Bayesian estimator based on assuming a Gaussian likelihood for $\epsilon \equiv y - x\beta$, $p(y, x|\beta)$ where for simplicity the variance of ϵ is assumed to be known and equal to 1:

$$p(y, x|\beta, \lambda) \propto \exp\left(-1/2 \sum_{i=1}^{n} (y_i - x_i\beta)^2\right).$$

Now consider a prior on (β, λ) given by

$$p(\beta, \lambda) = \phi(\beta; 0, 1/\lambda)g(\lambda) \propto \lambda g(\lambda) \exp\left(-\frac{1}{2}\lambda|\beta|^2\right),$$

where $\phi(\cdot; m, v)$ is the Gaussian density function with mean m and variance v. Here β is assumed to have a Gaussian prior and λ a density function $g(\lambda)$. By the Bayes theorem, the posterior can be obtained by

$$p(\beta, \lambda|y, x) = p(y, x|\beta, \lambda)p(\beta, \lambda) \propto \lambda g(\lambda) \exp\left(-\frac{1}{2}L(\beta, \lambda)\right).$$

Define $y_\lambda = [y \vdots 0]$ and $x_\lambda = [x \vdots \sqrt{\lambda}]$, and note that $\hat{\beta}^\lambda$ is the OLS estimator of a regression of y_λ on x_λ. Then using (Zellner, 1971, p. 66) results

$$p(\beta, \lambda|y, x) \propto \lambda g(\lambda) \exp\left(-\frac{1}{2}((y_\lambda - x_\lambda\hat{\beta}^\lambda)'(y_\lambda - x_\lambda\hat{\beta}^\lambda) + (\beta - \hat{\beta}^\lambda)'x_\lambda' x_\lambda(\beta - \hat{\beta}^\lambda))\right).$$

Note that for fixed λ, the maximum value of $p(\beta, \lambda|y, x)$ corresponds to $\beta = \hat{\beta}^\lambda$. The posterior distribution of β, $p(\beta|y, x)$, requires integrating out λ which depends on the assumed distribution $g(\cdot)$. Note that $p(\beta|y, x)$ is different from $h(\beta|y, x)$, the density function of our estimator. In Fig. 1, we illustrate this idea. Assume for simplicity that β is unidimensional and consider a given joint posterior distribution of β and λ. For each β, $p(\beta|y, x)$ is obtained by computing the mean over all possible values of λ. For instance, for β_1, $p(\beta_1|y, x)$ corresponds to the integration over λ along the vertical line at β_1, and the same for β_2 and β_3. However, $h(\beta|y, x)$ is obtained by computing the mean over those values of λ for which $\beta = \hat{\beta}^\lambda$, i.e. only those for which for a given λ it finds the maximum density. For β_1 that only corresponds to λ_1, β_2 to λ_2 and β_3 to λ_3, as in the dashed line. Thus, the distribution of our proposed estimator uses the most likely β for each λ.

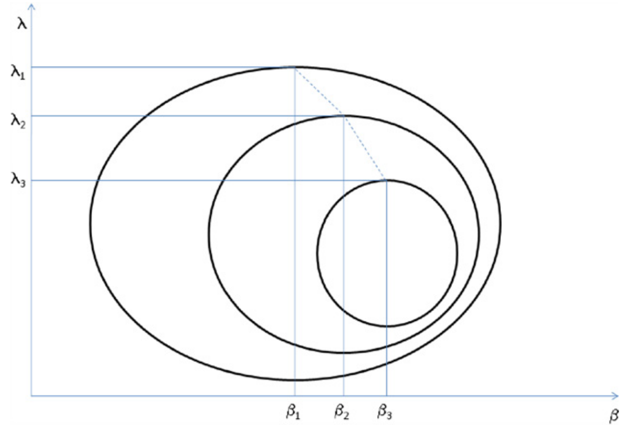

Fig. 1. Bivariate posterior distribution of (β, λ).

5. Conclusions

This paper proposed a novel way of dealing with endogeneity bias when there is no additional information such as instrumental or proxy variables. In particular, a prior is imposed on the endogeneity bias and the resulting estimators can be constructed both as a method-of-moments estimator or in a Ridge penalized regression framework.

Several extensions could be proposed for future research. First, the parameters modeling the endogeneity bias could depend on both prior information and observable variables. Second, this model could be applied to the l_1 penalization case using the asymptotic results in Knight and Fu (2000) and could be extended to model selection with several endogenous regressors. Third, the proposed model could be applied to forecasting where there is uncertainty about some parameter values and prior information would be imposed.

References

Altonji, J.G., Elder, T.E., Taber, C.R., 2005a. Selection on observed and unobserved variables: assessing the effectiveness of catholic schools. Journal of Political Economy 113, 151–184.

Altonji, J.G., Elder, T.E., Taber, C.R., 2005b. An evaluation of instrumental variable strategies for estimating the effects of catholic schooling. Journal of Human Resources 40, 791–821.

Altonji, J.G., Elder, T.E., Taber, C.R., 2008. Using selection on observed variables to assess bias from unobservables when evaluating Swan–Ganz catheterization. American Economic Review Papers and Proceedings 98, 345–350.

Angrist, J., Krueger, A., 2001. Instrumental variables and the search for identification: from supply and demand to natural experiments. Journal of Economic Perspectives 15, 69–85.

Chalak, K., White, H., 2011. An extended class of instrumental variables for the estimation of causal effects. Canadian Journal of Economics 44, 1–451.

Galvao, A.F., Montes-Rojas, G.V., 2010. Penalized quantile regression for dynamic panel data. Journal of Statistical Planning and Inference 140, 3476–3497.

Hausman, J.A., 1983. Specification and estimation of simultaneous equation models. In: Griliches, Z., Intriligator, M.D. (Eds.), Handbook of Econometrics, Vol. II. North-Holland Press, pp. 391–448.

Kiviet, J.F., 2011. Identification and inference in a simultaneous equation under alternative information sets and sampling schemes. Amsterdam School of Economics Discussion paper 2011/02.

Klein, R., Vella, F., 2010. Estimating a class of triangular simultaneous equations models without exclusion restrictions. Journal of Econometrics 154, 154–164.

Knight, K., Fu, W., 2000. Asymptotics for Lasso-type estimators. The Annals of Statistics 28, 755–770.

Levinsohn, J., Petrin, A., 2003. Estimating production functions using inputs to control for unobservables. Review of Economic Studies 70, 317–342.

Lewbel, A., 2012. Using heteroskedasticity to identify and estimate mismeasured and endogenous regressor models. Journal of Business and Economic Statistics 30, 67–80.

Olley, S., Pakes, A., 1996. The dynamics of productivity in the telecommunications equipment industry. Econometrica 64, 1263–1298.

Rigobon, R., 2003. Identification through heteroskedasticity. Review of Economics and Statistics 85, 777–792.

Wooldridge, J., 2002. Econometric Analysis of Cross Section and Panel Data. MIT Press, London.

Wooldridge, J., 2009. On estimating firm-level production functions using proxy variables to control for unobservables. Economics Letters 104, 112–114.

Zellner, A., 1971. An Introduction to Bayesian Inference in Econometrics. John Wiley & Sons, New York.