Approximately Sparse Econometric Models

David Gold
Department of Statistics, University of Washington Seattle, WA, 98195, USA

1 Introduction

An important objective of econometrics research is to conduct inference for the effect of a treatment variables on some economic response variable. In general, the value that the treatment takes for a given observation is not prescribed by the analyst. The relationship between the treatment and response may therefore be subject to *confounding*. That is, the distribution of the response given a *hypothetical intervention* in which treatment levels are prescribed by the analyst may differ from the distribution of the response given a natural distribution of treatments. Insofar as the effect of interest is specified in terms of the former distribution, the aforementioned discrepancy can induce bias in estimators derived from the latter distribution. In turn, statistical methods that address bias due to confounding are particularly relevant to econometrics studies. Two such methods are (i) the *method of instrumental variables (IV)* and (ii) inference methods for *partially linear (PL) models*.

1.1 Confounding

In this section, we briefly illustrate how confounding bias arises from the confluence of two factors: (i) the specification of a parameter of interest in terms of a hypothetical intervention in which treatment is assigned at the discretion of the analyst and (ii) the difference between such an intervention and actual circumstances. Let X, Y denote treatment and response variables, respectively, and suppose that the true data-generating mechanism of interest is

$$\mathrm{E}[Y \mid x, \boldsymbol{w}] \;=\; \beta x + f(\boldsymbol{w})\,, \tag{1.1}$$

where $\boldsymbol{W} \in \mathbb{R}^{p_w}$ are further covariates that influence Y. One possible quantity of interest is $\theta = \frac{\partial}{\partial x}\mathrm{E}[Y \mid x]$, where the reference population is considered under a hypothetical intervention in which X is set to x at the discretion of the analyst. Given that, under such a hypothetical intervention, assignment of the treatment X is independent of $\boldsymbol{W}$, the parameter of interest θ is equal to β. However, under actual circumstances, in which X may be associated with $\boldsymbol{W}$, measurements of the latter are often unavailable. Estimating θ by naïvely fitting the model $\mathrm{E}[Y \mid x] = \beta^{\dagger} x$ incurs a well-known bias that, in general, is not alleviated by asymptotics.

The method of instrumental variables addresses the foregoing issues by introducing *instrumental variables* $\boldsymbol{Z}$ associated with X through the conditional mean function $\mathrm{E}[X \mid \boldsymbol{z}]$ but otherwise unassociated with the response Y. In turn, X is allowed to be *endogenous* — that is, to have nontrivial covariance with the response error. The endogeneity of X reflects the influence of unmeasured confounders. Intuitively, the method of instrumental variables circumvents confounding bias by studying the variation in the response relative to the "part" of variation in the treatment that is due only to the effect of the instruments.

The partially linear model 1.1 is relevant when the analyst is reasonably confident that potential confounders $\boldsymbol{W}$ are measured but lacks a priori knowledge that would justify restrictions, such as linearity, on the functional relationship between the response Y and $\boldsymbol{W}$. In such cases the analyst may reasonably assume the X are not endogenous and avail herself of more general semiparametric methods.

Apart from the foregoing motivation, the present essay focuses exclusively on the statistical properties of estimators $\hat{\theta}$ derived from the IV and PL models. We refer the reader to [Pearl, 2009, Chapter 5] and Angrist and Krueger [2001] for deeper discussions of how instrumental variables models in particular come to bear on causal inference.

1.2 Inference

Inference for the treatment effect in an IV model proceeds by fitting the observed responses Y_i to estimates $\hat{\mathrm{E}}[X_i \mid \boldsymbol{z}_i]$. In linear IV models, the simplest implementation of this procedure is *Two-Stage Least Squares (2SLS)* regression of the treatment X on the instruments $\boldsymbol{Z}$ in the first stage and Y on the estimates $\hat{\mathrm{E}}[X \mid \boldsymbol{z}]$ in the second stage. Small and large sample properties of the 2SLS have been extensively researched since the 1950s, particularly to the end of alleviating the bias incurred by introducing "many instruments" — that is, letting the number $p_{\boldsymbol{z}}$ of instruments grow with the sample size n. Early work in this vein concerned the

Limited Information Maximum Likelihood (LIML) estimator [TODO: comment on LIML – I'm still working through the literature on this].

Fuller [1977] studies a modification of the LIML estimator that has finite moments and bias of order of n^{-2} [TODO: comment on issues concerning standard error estimates of Fuller estimator]. Becker [1994] proposes similarly modified estimators and obtains corrected standard errors for asymptotic approximations under a particular sequence of model parameters in which $p_{\boldsymbol{z}}$ may satsify $\lim_{n\to\infty} p_{\boldsymbol{z}}/n > 0$ — a condition that entails inconsistency of the 2SLS estimator under reasonable assumptions. More recently, Hansen et al. [2008] extend the result of Becker [1994] to the case of non-Gaussian errors. Buna and Windmeijer [2011] give an accessible survey of approximations to the finite sample bias of the 2SLS estimator in the case of a single endogenous covariate X.

References

Joshua D. Angrist and Alan B. Krueger. Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics*, 1991.

Joshua D. Angrist and Alan B. Krueger. Instrumental variables and the search for identification: From supply and demand to natural experiments. *Journal of Economic Perspectives*, 15(4):69–85, 2001.

Paul A. Becker. Alternative approximations to the distributions of the instrumental variable estimators. *Econometrica*, 1994.

Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. Inference for high-dimensional sparse econometric models. *Advances in Economics and Econometrics*, 2011.

Peter Bühlmann and Sara van de Geer. *Statistics for High-Dimensional Data.* Springer, 2011.

Maurice J.G. Buna and Frank Windmeijer. A comparison of bias approximations for the two-stage least squares (2sls) estimator. *Economic Letters*, 2011.

Wayne A. Fuller. Some properties of a modification of the limited information estimator. *Econometrica*, 1977.

Eric Gautier and Alexandre B. Tsybakov. High-dimensional instrumental variables regression and confidence sets. *ArXiv:1105.2454v4*, 2014.

Christian Hansen, Jerry Hausman, and Whitney Newey. Estimation with many instrumental variables. *Journal of Business and Economic Statistics*, 2008.

Whitney K. Newey. Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics*, 1997.

Judea Pearl. *Causality: Models, Reasoning and Inference.* Cambridge University Press, 2nd edition, 2009.

P.M. Robinson. Root-n-consistent semiparametric regression. *Econometrica*, 56(4):931–954, 1988.