cemmap
centre for microdata methods and practice

Post-selection and post-regularization inference in linear models with many controls and instruments

Victor Chernozhukov
Christian Hansen
Martin Spindler

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP02/15

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
An ESRC Research Centre

POST-SELECTION AND POST-REGULARIZATION INFERENCE IN LINEAR MODELS WITH MANY CONTROLS AND INSTRUMENTS

VICTOR CHERNOZHUKOV, CHRISTIAN HANSEN, AND MARTIN SPINDLER

Abstract. In this note, we offer an approach to estimating structural parameters in the presence of many instruments and controls based on methods for estimating sparse high-dimensional models. We use these high-dimensional methods to select both which instruments and which control variables to use. The approach we take extends Belloni et al. (2012), which covers selection of instruments for IV models with a small number of controls, and extends Belloni, Chernozhukov and Hansen (2014), which covers selection of controls in models where the variable of interest is exogenous conditional on observables, to accommodate both a large number of controls and a large number of instruments. We illustrate the approach with a simulation and an empirical example. Technical supporting material is available in a supplementary appendix.

Publication: American Economic Review 2015, Papers and Proceedings.

Online Appendix: Post-Selection and Post-Regularization Inference: An Elementary, General Approach.

1. Model and Estimation Approach

Consider the linear IV model

$$y_i = \alpha_0 d_i + x_i'\beta_0 + \varepsilon_i, \tag{1}$$

$$d_i = x_i'\gamma_0 + z_i'\delta_0 + u_i, \tag{2}$$

with $\mathrm{E}[(z_i', x_i')'\varepsilon_i] = \mathrm{E}[(z_i', x_i')'u_i] = 0$. d_i is the scalar endogenous variable and α the coefficient of interest, x_i is a p_n^x-vector of exogenous control variables, z_i is a p_n^z-vector of instruments, n is the sample size, and $p_n^x \gg n$ and $p_n^z \gg n$ are allowed. Extension to the case where d_i is a vector is straightforward and omitted for simplicity. We may

Date: January 5, 2015.

Chernozhukov: Massachussets Institute of Technology, 50 Memorial Drive, E52-361B, Cambridge, MA 02142, vchern@mit.edu. Hansen: University of Chicago Booth School of Business, 5807 S. Woodlawn Ave., Chicago, IL 60637, chansen1@chicagobooth.edu. Spindler: Munich Center for the Economics of Aging, Amalienstr. 33, 80799 Munich, Germany, spindler@mea.mpisoc.mpg.de.

have that z_i and x_i are correlated so that z_i are only valid instruments after controlling for x_i; specifically, we let $z_i = \Pi x_i + \zeta_i$, for Π a $p_n^z \times p_n^x$ matrix and ζ_i a p_n^z-vector of unobservables with $\mathrm{E}[x_i\zeta_i'] = 0$. Substituting this expression for z_i as a function of x_i into (2) and then further substituting into (1) gives a system for y_i and d_i that depends only on x_i:

$$y_i = x_i'\theta_0 + \rho_i^y, \tag{3}$$

$$d_i = x_i'\vartheta_0 + \rho_i^d, \tag{4}$$

with $\mathrm{E}[x_i\rho_i^y] = 0$ and $\mathrm{E}[x_i\rho_i^d] = 0$. This model includes the many instruments and small number of controls case by setting $p_n^x \ll n$ and can accommodate the exogenous case by setting $p_n^z = 0$ and imposing the additional condition $\mathrm{E}[d_i\varepsilon_i] = 0$.

Because the dimension of $\eta_0 = (\theta_0', \vartheta_0', \gamma_0', \delta_0')'$ may be larger than n, informative estimation and inference about α_0 is impossible without imposing restrictions on η_0. For simplicity, we provide discussion under the assumption of exact sparsity and present a generalization to approximate sparsity in the supplemental material. Specifically, we assume that

$$\|\eta_0\|_0 \leq s_n, \quad s_n^2 \log(p_n^z + p_n^x)^3/n \to 0,$$

where $\|\eta_0\|_0$ denotes the number of non-zero elements of η_0. That is, sparsity requires that, among the $p_n^x + p_n^z$ observed variables, the number of variables with non-zero coefficients is small relative to the sample size. This assumption then reduces the problem of estimating α to a problem of finding which instruments and controls to use in equations (1) and (2).

The problem that arises is that variable selection techniques are not perfect and are prone to making selection mistakes. There are two kinds of selection mistakes: A variable may be deemed relevant when in fact it has a zero coefficient and thus has no true explanatory power, or a variable may be dropped from the model despite having a non-zero coefficient. Both types of mistakes may detrimentally affect post-model-selection estimators and inference for α. When irrelevant variables are spuriously included after being deemed predictive from looking at the data, overfitting occurs and importantly the spuriously included variables are those most correlated to the noise in the sample due to data-snooping which introduces a type of "endogeneity" bias. When relevant x variables are excluded, one is left with standard omitted variables bias. When relevant z variables are excluded, one loses identification power. This last concern can be dealt with through appropriate use of weak identification robust inference as in Belloni et al. (2012).

The first type of mistake, the spurious inclusion of irrelevant variables, can be avoided through the use of modern, principled data-mining methods. For example, we use the Lasso with tuning parameters chosen as in Belloni et al. (2012), and many other options are available. These methods differ from the unprincipled data-snooping that many economists associate with the term data-mining. Specifically, modern data-mining denotes a principled search for true predictive power that guards against false discovery and overfitting, does not erroneously equate in-sample fit to out-of-sample predictive ability,

and accurately accounts for using the same data to examine many different hypotheses or models.

Of course, guarding against the first type of error comes at the cost of needing to acknowledge that the exclusion of relevant variables is likely to occur. While sensible approaches such as Lasso will accurately find strong predictors, one can show that such procedures have non-negligible probability of missing predictors with small but non-zero coefficients. Exclusion of such predictors can have substantive impacts on inference for parameters of interest such as α in our model; see, for example, Leeb and Pötscher (2008). To overcome this difficulty, one needs to base estimation and inference on procedures that are robust to this type of model selection mistake. One such approach relies on using estimating equations that are locally insensitive to this type of mistake, termed orthogonal moment functions in Belloni et al. (2013).

In the IV model with many instruments and controls, such a moment condition is given by

$$M(\alpha_0; \eta_0) = 0, \;\; M(\alpha, \eta) := \mathrm{E}\left[\psi_i(\alpha, \eta)\right] \tag{5}$$

where $\psi_i(\alpha, \eta) = (\tilde{\rho}_i^y - \tilde{\rho}_i^d \tilde{\alpha})\tilde{v}_i$ for $\eta := (\theta', \vartheta', \gamma', \delta')'$, $\tilde{\rho}_i^y := y_i - x_i'\theta$, $\tilde{\rho}_i^d := d_i - x_i'\vartheta$, and $\tilde{v}_i := x_i'\gamma + z_i'\delta - x_i'\vartheta$. When we set $\tilde{\eta} = \eta_0$, we have $\tilde{\rho}_i^y = \rho_i^y = y_i - x_i'\theta_0$, $\tilde{\rho}_i^d = \rho_i^d = d_i - x_i'\vartheta_0$, and $\tilde{v}_i = v_i := x_i'\gamma_0 + z_i'\delta_0 - x_i'\vartheta_0 = \zeta_i'\delta_0$.

We can see that small selection errors will have relatively little impact on estimation of α_0 by noting that the following orthogonality condition holds:

$$\frac{\partial}{\partial \eta} M(\alpha_0, \eta)\Big|_{\eta=\eta_0} = 0. \tag{6}$$

In other words, missing the true value η_0 by a small amount does not invalidate the moment condition. Thus, estimators $\hat{\alpha}$ of α_0 based on the empirical analog of (5),

$$\hat{M}(\hat{\alpha}, \hat{\eta}) = 0 \tag{7}$$

with $\hat{M}(\alpha, \eta) := n^{-1} \sum_{i=1}^n \left[\psi_i(\alpha, \eta)\right]$, can be shown to be "immunized" against small selection mistakes. See Belloni et al. (2013) for a general formulation of orthogonal moment funtions for use in sparse high-dimenionsal models and a number of estimation and inference results.

Note that operationally using the empirical version of (5) to estimate α_0 is equivalent to using the usual IV regression of ρ^y on ρ^d using v as instruments. Based on this argument, we suggest the following algorithm for estimating α_0 based on the "double-selection" strategy of Belloni, Chernozhukov and Hansen (2014).

Algorithm 1. *(1) Do Lasso or Post-Lasso Regression of d_i on x_i, z_i to obtain $\hat{\gamma}$ and $\hat{\delta}$. (2) Do Lasso or Post-Lasso Regression of y_i on x_i to get $\hat{\theta}$. (3) Do Lasso or Post-Lasso Regression of $\hat{d}_i = x_i'\hat{\gamma} + z_i'\hat{\delta}$ on x_i to get $\hat{\vartheta}$. (4) Let $\hat{\rho}_i^y := y_i - x_i'\hat{\theta}$, $\hat{\rho}_i^d := d_i - x_i'\hat{\vartheta}$, and $\hat{v}_i := x_i'\hat{\gamma} + z_i'\hat{\delta} - x_i'\hat{\vartheta}$. Get estimator $\hat{\alpha}$ from (7) by using standard IV regression of $\hat{\rho}_i^y$ on*

$\hat{\rho}_i^d$ with $\hat{v}_i$ as the instrument. Perform inference on α_0 using $\hat{\alpha}$ or the associated score statistic and conventional heteroscedasticity robust standard errors.

The following result summarizes the properties of $\hat{\alpha}$ obtained from Algorithm 1.

Proposition 1. *Under the stated sparsity and other regularity conditions, the estimator $\hat{\alpha}$ defined in Algorithm 1 satisfies $\sqrt{n}(\hat{\alpha}-\alpha_0) \rightsquigarrow \mathcal{N}(0, V)$ where $V = E[v_i^2]^{-2}E[\psi_i(\alpha_0, \eta_0)^2]$. The score statistic $C(\alpha_0) = n|\hat{M}(\alpha_0, \hat{\eta})|^2/(n^{-1}\sum_{i=1}^n \psi_i^2(\alpha_0, \hat{\eta}))$ satisfies $C(\alpha_0) \rightsquigarrow \chi^2(1)$. Confidence intervals based on these two results are uniformly valid for inference about α_0 over a large class of models.*

The supplementary material provides a precise statement and proof. The theoretical results do not depend on whether the Lasso estimator or the Post-Lasso estimator of Belloni and Chernozhukov (2013) is used. In the results reported in this paper, we use the Post-Lasso estimator. Note that there are other algorithms that would yield similar asymptotic properties. For example, one could follow the double-selection strategy more closely by running Lasso regression of d_i on x_i and z_i, Lasso regression of d_i on x_i, Lasso regression of y_i on x_i, and then forming a 2SLS estimator using instruments selected in the first step and controlling for the union of controls selected in the three Lasso steps.

2. Simulation Example

To illustrate the preceding discussion, we report results from a small simulation experiment. Data were generated from the model given in Section 2 with $n = 200$, $p_n^x = 300$, and $p_n^z = 150$. Other parameter values were chosen so that the infeasible, optimal instruments are "strong", perfect model selection is impossible, and the sparse model provides a good approximation. Further details are available in the supplementary material.

We provide results for four different estimators - an infeasible Oracle estimator that knows the nuisance parameters η (Oracle), two naive estimators, and the "Double-Selection" estimator. The first naive estimator follows Algorithm 1 but replaces Lasso/Post-Lasso with stepwise regression with p-value for entry of .05 and p-value for removal of .10 (Naive 1). It is well-known that this procedure fails to control model selection mistakes in which irrelevant variables are included. The second naive estimator estimates the high-dimensional nuisance functions using Post-Lasso but uses the moment condition $\mathrm{E}\left[(\rho_i^y - \rho_i^d\alpha)(x_i'\delta + z_i'\gamma)\right] = 0$ (Naive 2). This moment condition does not satisfy the orthogonality condition described above, though estimation and inference about α_0 using this condition will be valid when perfect model selection for the regression of y on x and d on x is possible.

We report the median bias (Bias), median absolute deviation (MAD), and size of 5% level tests (Size) obtained from 1000 simulation replications for each procedure. For the

Oracle, we have Bias of .006, MAD of .095, and Size of .043. For Naive 1, Bias, MAD, and Size are .160, .227, and .302 respectively; and Bias, MAD, and Size are respectively .035, .103, and .095 for Naive 2. Finally, the Double-Selection approach gives Bias of .021, MAD of .099, and Size of .054.

These results correspond to the discussion in Section I. The first naive, unprincipled procedure fails to control spurious inclusion of irrelevant variables and performs quite poorly relative to the other three approaches. The second naive procedure can be shown to be formally valid when perfect model selection is possible and performs relatively well in terms of MAD. However, the asymptotic approximation under perfect model selection provides a misleading approximation to the true sampling distribution as evidenced by the size distortion. Finally, we see that basing estimation and inference on a principled variable selection procedure and moment conditions that are immunized against small model selection mistakes produces an estimator that performs well relative to the infeasible Oracle in terms of both estimation and inference performance as measured by MAD and Size.

3. Empirical Example

We conclude with a brief empirical example where we estimate the coefficients in a simple model of demand for automobiles. We use the data and basic strategy of Berry et al. (1995). For simplicity, we consider the most basic specification

$$\begin{aligned} \log(s_{it}) - \log(s_{0t}) &= \alpha_0 p_{it} + x_{it}'\beta_0 + \varepsilon_{it} \\ p_{it} &= z_{it}'\delta_0 + x_{it}'\gamma_0 + u_{it} \end{aligned}$$

where s_{it} is the market share of product i in market t with product 0 denoting the outside option, p_{it} is price and treated as endogenous, x_{it} are observed included product characteristics, and z_{it} are instruments. One could also consider allowing random coefficients and adapting the variable selection procedures to this setting; see Gillen et al. (2014).

In their basic results, Berry et al. (1995) use five variables in x_{it}: a constant, an air conditioning dummy, horsepower divided by weight, miles per dollar, and vehicle size. They argue that characteristics of other products provide valid instruments for price and choose 10 instruments for p_{it} based on intuition and an exchangeability argument. The first five instruments are formed by deleting product i and then summing each characteristic in x across all remaining products produced by product i's firm. The other five instruments are similarly constructed by deleting all products from product i's firm and then summing each characteristic in x across all remaining products. Using these controls and instruments, the 2SLS estimate of α is -.142 with an estimated standard error of .012. One might compare this to the OLS estimate obtained treating price as exogenous given the five controls listed above which is -.089 with estimated standard of .004.

It is interesting to note that Berry et al. (1995) state, "The choice of which attributes to include in the utility function is, of course, ad hoc" (p. 872). They similarly note that one could have considered additional instruments such as higher order terms (Berry et al., 1995, p. 861). The high-dimensional methods outlined in this paper offer one strategy to help address these concerns which complements the well-founded economic intuition motivating the authors' choices. We apply our outlined methods in two scenarios. In the first, we apply the method using just the original five controls and 10 instruments. In the second, we augment the set of potential controls with a time trend, quadratics, and cubics in all continuous variables, and all first order interactions and then use sums of these characteristics as potential instruments following the original strategy. These additions give a total of 24 x-variables and 48 potential instruments. We include the intercept in all models and select over the remaining variables.

In both cases, the results suggest demand is more elastic than indicated in the baseline results. After selection using only the original variables, we estimate the price coefficient to be -.185 with an estimated standard error of .014. In this case, all five controls are selected in the log-share on controls regression, all five controls but only four instruments are selected in the price on controls and instruments regression, and four of the controls are selected for the price on controls relationship. The difference between the baseline results is thus largely driven by the difference in instrument sets. The change in the estimated coefficient is consistent with the wisdom from the many-instrument literature that inclusion of irrelevant instruments biases 2SLS toward OLS.

With the larger set of variables, our post-model-selection estimator of the price coefficient is -.221 with an estimated standard error .015. Here, we see some evidence that the original set of controls may have been overly parsimonious. In the log-share on controls regression, we have that eight control variables are selected; and we have seven controls and only four instruments selected in the price on controls and instrument regression. We also have that 13 variables are selected for the price on controls relationship. The selection of these additional variables suggests that there is important nonlinearity missed by the baseline set of variables.

Finally, we note that in terms of own-price elasticities, the results become more plausible as we move from the baseline results to the results based on variable selection with a large number of controls. Recall that facing inelastic demand is inconsistent with profit maximizing price choice within the present context, so theory would predict that demand should be elastic for all products. However, the baseline point estimates imply inelastic demand for 670 products. Using the variable selection results provides results closer to the theoretical prediction. The point estimates based on selection from only the baseline variables imply inelastic demand for 139 products, and we estimate inelastic demand for only 12 products using the results based on selection from the larger set of variables. Thus, the new methods provide the most reasonable estimates of own-price elasticities. Of course, the simple specification above suffers from the usual drawbacks of the logit demand model, but the example illustrates how the application of the methods outlined in this note may be used in estimation of structural parameters in economics and add to the plausibility of the resulting estimates.

4. Conclusion

A great deal of empirical economic research aiming to estimate causal or structural effects depends on using the right set of controls and instruments. The need for formal methods that perform this model selection and inference procedures that remain valid following model selection is likely to increase in importance as data sets become richer. We have outlined one simple approach that can be used in an instrumental variables model with many instruments and controls that extends Belloni et al. (2012) and Belloni, Chernozhukov and Hansen (2014). The approach relies on an approximate sparsity assumption and the use of high-quality variable selection procedures coupled with the use of appropriate moment functions. These ideas follow from the general framework considered in Belloni et al. (2013). For more applications of similar ideas in economics, see also Bai and Ng (2009), Belloni et al. (ArXiv, 2010*b*); Gautier and Tsybakov (2011); Belloni et al. (2010*a*); and Belloni, Chernozhukov, Hansen and Kozbur (2014) and references therein.

References

Bai, Jushan and Ng, Serena. (2009). 'Selecting Instrumental Variables in a Data Rich Environment', *Journal of Time Series Econometrics* 1(1).

Belloni, A. and Chernozhukov, V. (2013). 'Least Squares After Model Selection in High-dimensional Sparse Models', *Bernoulli* 19(2), 521–547. ArXiv, 2009.

Belloni, Alexandre, Chen, Daniel, Chernozhukov, Victor and Hansen, Christian. (2012). 'Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain', *Econometrica* 80, 2369–2429. Arxiv, 2010.

Belloni, Alexandre, Chernozhukov, Victor, Fernández-Val, Ivan and Hansen, Christian. (2013). 'Program Evaluation with High-Dimensional Data', *arXiv:1311.2645* . ArXiv, 2013.

Belloni, Alexandre, Chernozhukov, Victor and Hansen, Christian. (2010*a*). 'Inference for High-Dimensional Sparse Econometric Models', *Advances in Economics and Econometrics. 10th World Congress of Econometric Society. August 2010* III, 245–295. ArXiv, 2011.

Belloni, Alexandre, Chernozhukov, Victor and Hansen, Christian. (2014). 'Inference on Treatment Effects After Selection Amongst High-Dimensional Controls', *Review of Economic Studies* 81, 608–650. ArXiv, 2011.

Belloni, Alexandre, Chernozhukov, Victor and Hansen, Christian. (ArXiv, 2010*b*), LASSO Methods for Gaussian Instrumental Variables Models. arXiv:1012.1297.

Belloni, Alexandre, Chernozhukov, Victor, Hansen, Christian and Kozbur, Damian. (2014). 'Inference in High Dimensional Panel Models with an Application to Gun Control', *arXiv:1411.6507* . ArXiv, 2014.

Berry, Steven, Levinsohn, James and Pakes, Ariel. (1995). 'Automobile Prices in Market Equilibrium', *Econometrica* 63, 841–890.

Gautier, Eric and Tsybakov, Alexander B. (2011). 'High-Dimensional Instrumental Variables Regression and Confidence Sets', *ArXiv:1105.2454v4* .

Gillen, Benjamin J., Shum, Matthew and Moon, Hyungsik Roger. (2014). 'Demand Estimation with High-Dimensional Product Charateristics', *Advances in Econometrics* . forthcoming.

Leeb, Hannes and Pötscher, Benedikt M. (2008). 'Can one estimate the unconditional distribution of post-model-selection estimators?', *Econometric Theory* 24(2), 338–376.
URL: *http://dx.doi.org/10.1017/S0266466608080158*