EC708 Discussion 4

Yan Liu[1]

Department of Economics

Boston University

February 19, 2021

[1] Parts of the materials are from teaching slides of Shuowen Chen and Whitney Newey.

Outline

1. Hausman Test as a Pretest
2. Many Instruments Asymptotics
3. Machine Learning and Econometrics
4. More Topics on GMM

Hausman Test

Overview

Consider two estimators $\hat{\theta}_{\mathsf{I}}$ and $\hat{\theta}_{\mathsf{II}}$ and a general hypothesis H_0 against alternative H_1. Suppose

- Under H_0: both estimators are consistent while $\hat{\theta}_{\mathsf{I}}$ is more efficient;
- under H_1: only $\hat{\theta}_{\mathsf{II}}$ is consistent.

Hausman test statistic:

$$H_T = T(\hat{\theta}_{\mathsf{II}} - \hat{\theta}_{\mathsf{I}})'[\hat{V}_{\mathsf{II}} - \hat{V}_{\mathsf{I}}]^{-1}(\hat{\theta}_{\mathsf{II}} - \hat{\theta}_{\mathsf{I}}).$$

where $\hat{V}_{\mathsf{I}}, \hat{V}_{\mathsf{II}}$ are consistent estimators of the asymptotic variance of $\hat{\theta}_{\mathsf{I}}, \hat{\theta}_{\mathsf{II}}$. Under H_0, $H_T \xrightarrow{d} \chi^2_k$, where $k = \dim(\hat{\theta}_{\mathsf{I}}) = \dim(\hat{\theta}_{\mathsf{II}})$.

Hausman Test

Consider the following regression

$$\underset{T\times 1}{y} = \underset{T\times 1}{X}\beta + \underset{T\times 1}{u}.$$

We want conduct hypothesis testing

$$H_0 : \beta = \beta_0, \quad H_1 : \beta \neq \beta_0.$$

We worry that X might be endogenous. Suppose we have k valid and strong instruments:

$$\underset{T\times 1}{X} = \underset{T\times k}{Z}\Pi + \underset{T\times 1}{v}.$$

- If X is exogenous, $\hat{\beta}_{OLS}$ and $\hat{\beta}_{2SLS}$ are both consistent while $\hat{\beta}_{OLS}$ is more efficient (under conditional homoskedasticity);
- If X is endogenous, only $\hat{\beta}_{2SLS}$ is consistent.

Hausman Test

A natural thought is conduct a two-stage test:

1. Run Hausman test to test endogeneity of X.

   The null is H_0^{Hausman} : X is exogenous. Hausman test statistic is

$$H_T = T\frac{(\hat{\beta}_{2SLS} - \hat{\beta}_{OLS})^2}{\hat{V}_{2SLS} - \hat{V}_{OLS}}.$$

2. Run 2SLS based t-test if H_0^{Hausman} is rejected: $t_{2SLS} = \frac{\sqrt{T}(\hat{\beta}_{2SLS} - \beta_0)}{\sqrt{\hat{V}_{2SLS}}}$.

   Run OLS based t-test if H_0^{Hausman} is not rejected: $t_{OLS} = \frac{\sqrt{T}(\hat{\beta}_{OLS} - \beta_0)}{\sqrt{\hat{V}_{OLS}}}$.

The two-stage test statistic is

$$t_T(\beta_0) = t_{OLS}(\beta_0)1(H_T < \chi^2_{1,1-\alpha}) + t_{2SLS}(\beta_0)1(H_T > \chi^2_{1,1-\alpha}).$$

This test is problematic: severe size distortion in the second stage.

Guggenberger (2010) shows that the two-stage test is affected by three nuisance parameters:

1. $\rho = \mathrm{Corr}(u_t, v_t)$: level of endogeneity
2. $\mu/\sqrt{T}$: strength of instruments, where μ^2 is the concentration parameter. Assume $\mu/\sqrt{T} \in [\kappa, \overline{\kappa}]$ with $\kappa > 0$ (rules out weak instruments).
3. α: nominal size of the Hausman test.

The level of endogeneity is the main problem.

When and why does Hausman pretest fail?

- When X is strongly endogenous ($\sqrt{T}\rho \to \infty$):
  Hausman test always rejects H_0^{Hausman}, and t_{2SLS} is always used.
  Since instruments are strong, second-stage inference is good.
- When X is weakly endogenous ($\sqrt{T}\rho \to h < \infty$):
  Hausman test does not have enough power to detect accurately.
  Whenever it cannot reject H_0^{Hausman}, t_{OLS} is used, leading to invalid second-stage inference.

In practice we don't know how endogenous X is.

Nominal size of the Hausman test: α

- Recall we reject H_0^{Hausman} if $H_T > \chi^2_{1,1-\alpha}$.
- An increase in α means it is easier to reject.
- Hence we use t_{2SLS} more often in the second stage.

Strength of instruments: κ

- An increase in κ means the instruments are stronger.
- Hence properties of $\hat{\beta}_{2SLS}$ are further guaranteed in finite samples.

However, increasing α or κ has no effect on the conditional size of the second-stage test, conditional on not rejecting H_0^{Hausman}.

Many Instruments Asymptotics

Including more instruments reduces variances but increases bias in practice. Consider

$$
\begin{aligned}
y &= X\beta + u, \\
X &= Z\pi + v,
\end{aligned}
$$

where y and X are $T \times 1$ and Z is $T \times k$ (non-random). Suppose

- conditional homoskedasticity:

$$
\mathsf{Var}\left(\begin{pmatrix} u_t \\ v_t \end{pmatrix} \middle| Z_t \right) = \begin{pmatrix} \sigma_u^2 & \sigma_{uv} \\ \sigma_{uv} & \sigma_v^2 \end{pmatrix}.
$$

- $k/T \to \alpha \in (0,1)$. Number of instruments is non-negligible relative to the sample size.
- $\pi' Z' Z\pi / T \to Q$.

Many Instruments Asymptotics

The 2SLS estimator is

$$\hat{\beta}_{2SLS} - \beta = \left(\frac{X' P_Z X}{T} \right)^{-1} \left(\frac{X' P_Z u}{T} \right), \quad \text{where } P_Z = Z(Z'Z)^{-1} Z'.$$

Note that

$$\mathbb{E}\left[\frac{X' P_Z X}{T} \right] = \mathbb{E}\left[\frac{\pi' Z' P_Z Z \pi}{T} \right] + \mathbb{E}\left[\frac{v' P_Z v}{T} \right] = \frac{\pi' Z' Z \pi}{T} + \frac{k}{T} \sigma_v^2.$$

(The last equality uses $v' P_Z v = \text{tr}(v' P_Z v) = \text{tr}(v v' P_Z)$ and hence

$$\mathbb{E}[v' P_Z v] = \mathbb{E}[\text{tr}(v v' P_Z)] = \text{tr}\mathbb{E}[v v' P_Z] = \sigma_v^2 \text{tr}(P_Z) = \sigma_v^2 k.)$$

Similarly, we have $\mathbb{E}\left[\frac{X' P_Z u}{T} \right] = \mathbb{E}\left[\frac{v' P_Z u}{T} \right] = \frac{k}{T} \sigma_{uv}$. Therefore,

$$\hat{\beta}_{2SLS} - \beta \xrightarrow{p} (Q + \alpha \sigma_v^2)^{-1} \alpha \sigma_{uv}.$$

Extreme case: $k = T, P_Z = I_T, \hat{\beta}_{2SLS} = (X'X)^{-1} X' y = \hat{\beta}_{OLS}$.

- Thanks to a continuing decrease of cost of data collection and storage, economists now have access to big datasets.
- When p, the number of characteristics measured on a person or object, is larger than T, the sample size, the dataset is considered to be high-dimensional.
- OLS no longer feasible when $p > T$ (no unique solution).
- Usually economists specify key variables and functional forms and conduct robustness checks afterwards.
- An alternative: do semi/non-parametric econometrics. Cost: curse of dimensionality
- New alternative: select key variables using machine learning methods and conduct inference on selected specifications

Consider the following linear model

$$y = \alpha D + X\beta + u,$$

- D: treatment/policy variable of interest
- X: control variables (high-dimensional)

D is taken as exogenous after conditioning on X. We want to estimate and conduct inference on α.

- We impose approximate sparsity assumption: only s variables among X, where $s \ll T$, have non-zero coefficients, while permitting a non-zero approximation error small relative to estimation error.
- This assumption allows for imperfect model selection.

The LASSO estimator for least squares solves

$$\min_{\beta} \sum_{t=1}^{T} (y_t - \alpha d_t - x_t' \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \pi_j,$$

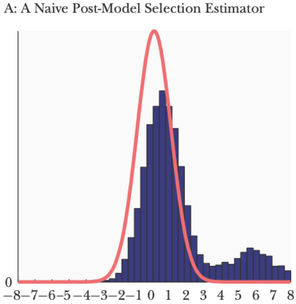
where $\lambda > 0$ is the penalty level and π_j are the penalty loadings.

- The ℓ_1-norm penalization $\lambda \sum_{j=1}^{p} |\beta_j| \pi_j$ shrinks the estimated coefficients toward zero and thus prevents overfitting.
- We exclude α from LASSO penalty.
- λ is chosen using data-driven methods (Belloni, Chen, Chernozhukov, and Hansen, 2012) or cross-validation (in prediction contexts).

Suppose LASSO selects the first two variables in X and the researcher now works with the following model

$$y = \alpha D + X_1\beta_1 + X_2\beta_2 + u.$$

A natural thought: run OLS and perform t-test on $\hat{\alpha}$. **Size distortion!**



Omitted variable bias

- LASSO targets prediction, so variables in X that are highly correlated with D will tend to be dropped since including them won't add much predictive power for the outcome given D is already in the model.
- If these variables are highly correlated with D and have nonzero coefficients in OLS, we have omitted variable bias.

Remedy: Double selection (Belloni, Chernozhukov, and Hansen, 2014)

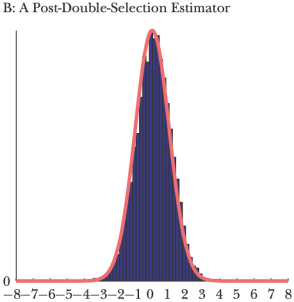
- Add an auxiliary regression $D = X\gamma + v$ and consider the system

$$y = \alpha D + X\beta + r_C + u, \tag{1}$$

$$D = X\gamma + r_D + v. \tag{2}$$

- Equation (2) aims to bring in variables that are correlated with D.

1. Run LASSO on y against D and X, excluding D from penalty. Denote the selected set of variables from X as $\hat{I}_1$.
2. Run LASSO on D against X. Denote the selected set of variables from X as $\hat{I}_2$.
3. Run OLS on y against D and $\hat{I} = \hat{I}_1 \cup \hat{I}_2$. Conduct t-test on $\hat{\alpha}$.



Consider the endogeneity model with potentially many instruments

$$y = X\beta + u,$$
$$X = Z\Pi + r_Z + v,$$

where r_Z is approximation error, $\mathbb{E}[u_t|z_t] = \mathbb{E}[u_t|z_t, r_{Zt}] = 0$, $\mathbb{E}[u_t v_t] \neq 0$.

- Variable selection in the first-stage is a pure predictive relationship: choose instruments that get the best fitted $\hat{X}$
- Therefore, we can run LASSO in the first-stage, and conventional inference from 2SLS based on selected instruments is valid.
- Belloni, Chen, Chernozhukov, and Hansen (2012) formalize the intuition and establish the asymptotic properties.

Over-identification test:

Test for moment validity in over-identified models:

$$H_0 : \mathbb{E}[m(X_t, \theta_0)] = 0, \quad H_1 : \mathbb{E}[m(X_t, \theta)] \neq 0, \ \forall \theta \in \Theta.$$

Hansen-Sargan's J-test:

$$J_T = T m_T(\hat{\theta}_T)' S_T^{-1} m_T(\hat{\theta}_T) \xrightarrow{d} \chi^2_{k-q} \text{ under } H_0.$$

- Rejecting H_0 doesn't tell you which moments are invalid.
- Rejecting H_0 doesn't necessarily mean moments are invalid. It could also be model misspecification.
- J-test tends to overreject in finite samples.

Test subsets of moment conditions:

Partition $m(X_t, \theta) = (m^1(X_t, \theta)', m^2(X_t, \theta)')'$ and G_0, S_0 conformably.

Test $H_0 : \mathbb{E}[m^1(X_t, \theta_0)] = 0$. One simple test is

$$T_1 = \min_\theta T m_T(\theta)' S_T^{-1} m_T(\theta) - \min_\theta T m_T^2(\theta)' S_T^{22^{-1}} m_T^2(\theta).$$

Under $H_0, T_1 \xrightarrow{d} \chi^2_{k_1}$, where k_1 is the dimension of $m^1(X_t, \theta)$.

An alternative version can be

$$T(\tilde{m}_T^1)' \tilde{S}_T^{-1} \tilde{m}_T^1,$$

where $\tilde{m}_T^1 = m_T^1(\hat{\theta}_T) - S_T^{12} S_T^{22^{-1}} m_T^2(\hat{\theta}_T)$ and $\tilde{S}_T$ is an estimator of the asymptotic variance of $\sqrt{T}\tilde{m}_T^1$.

Hausman test:

Let $\hat{\theta}_T$ and $\hat{\theta}_T^2$ denote optimal two-step GMM estimators using all moment conditions and using just $m^2(X_t, \theta)$, respectively.

- Under $H_0 : \mathbb{E}[m(X_t, \theta_0)] = 0$, both estimators are consistent while $\hat{\theta}_T$ is more efficient;
- under $H_1 : \mathbb{E}[m^1(X_t, \theta_0)] \neq 0, \mathbb{E}[m^2(X_t, \theta_0)] = 0$, only $\hat{\theta}_T^2$ is consistent.

If $k_1 \leq q$, a Hausman test can be formed as

$$H_T = (\hat{\theta}_T^2 - \hat{\theta}_T)'[({G_T^2}' {S_T^{22}}^{-1} G_T^2)^{-1} - (G_T' S_T^{-1} G_T)^{-1}]^{-1}(\hat{\theta}_T^2 - \hat{\theta}_T)$$

These tests are asymptotically equivalent. See Newey (1985).

Test parametric hypothesis:

$$H_0 : \underset{r\times 1}{s(\theta_0)} = 0, \quad \text{where } r < k.$$

Let $\tilde{\theta}_T = \underset{s(\theta)=0}{\arg\min}\, m_T(\theta)' S_T^{-1} m_T(\theta)$ be a restricted GMM estimator.

Let $\hat{G}_T = G_T(\hat{\theta}_T) = \partial m_T(\hat{\theta}_T)/\partial\theta'$ and $\tilde{G}_T = G_T(\tilde{\theta}_T)$. Under H_0,

$$W = Ts(\hat{\theta}_T)' \left[\frac{\partial s(\hat{\theta}_T)'}{\partial\theta} (\hat{G}_T' S_T^{-1} \hat{G}_T)^{-1} \frac{\partial s(\hat{\theta}_T)}{\partial\theta'} \right]^{-1} s(\hat{\theta}_T) \xrightarrow{d} \chi^2(q),$$

$$LR = Tm_T(\tilde{\theta}_T)' S_T^{-1} m_T(\tilde{\theta}_T) - Tm_T(\hat{\theta}_T)' S_T^{-1} m_T(\hat{\theta}_T) - W \xrightarrow{p} 0,$$

$$LM = Tm_T(\tilde{\theta}_T)' S_T^{-1} \tilde{G}_T (\tilde{G}_T' S_T^{-1} \tilde{G}_T)^{-1} \tilde{G}_T' S_T^{-1} m_T(\tilde{\theta}_T) - W \xrightarrow{p} 0.$$

Asymptotic approximation is often more accurate for LR and LM than W.

Efficiency improvement occurs because optimal weighting matrix for fewer moment conditions is not optimal for all the moment conditions.

Let $m(X_t, \theta) = (m^1(X_t, \theta)', m^2(X_t, \theta)')'$. The optimal GMM using just $m^2(X_t, \theta)$ corresponds to GMM using $m(X_t, \theta)$ with

$$W_T = \begin{pmatrix} S_T^{22^{-1}} & 0 \\ 0 & 0 \end{pmatrix} \neq S_T^{-1}.$$

Example: Linear regression model

$$E[y_t|X_t] = X_t'\beta_0.$$

OLS estimator is GMM with $m^1(Z_t, \beta) = X_t(y_t - X_t'\beta)$. We can use nonlinear functions $a(X_t)$ of X_t as additional "instrumental variables".

The optimal two-step GMM based on

$$m(Z_t, \beta) = \begin{pmatrix} m^1(Z_t, \beta) \\ m^2(Z_t, \beta) \end{pmatrix} = \begin{pmatrix} X_t \\ a(X_t) \end{pmatrix} (y_t - X_t'\beta)$$

is more efficient than OLS with heteroskedasticity.

Caveats:

- No reduction in asymptotic variance when additional moment conditions are exactly identified. That is, if we add $m^2(Z_t, \beta, \gamma)$ where γ has the same dimension as $m^2(Z_t, \beta, \gamma)$.
- Adding moment conditions can increase small sample bias (with endogeneity present).
- Adding moment conditions can increase small sample variance.
- In the example of linear regression, $(\hat{G}_T' S_T^{-1} \hat{G}_T)^{-1}$ tends to provide a poor approximation to the variance of $\hat{\beta}_T$. See Cragg (1983).

Bibliography

Belloni, A., Chen, D., Chernozhukov, V., and Hansen, C. (2012), "Sparse models and methods for optimal instruments with an application to eminent domain," *Econometrica*, 80, 2369–2429.

Belloni, A., Chernozhukov, V., and Hansen, C. (2014), "Inference on treatment effects after selection among high-dimensional controls," *The Review of Economic Studies*, 81, 608–650.

Cragg, J. G. (1983), "More efficient estimation in the presence of heteroscedasticity of unknown form," *Econometrica: Journal of the Econometric Society*, 751–763.

Guggenberger, P. (2010), "The impact of a Hausman pretest on the asymptotic size of a hypothesis test," *Econometric Theory*, 369–382.

Newey, W. K. (1985), "Generalized method of moments specification testing," *Journal of econometrics*, 29, 229–256.