TA Session 2: Pretesting and Post Model Selection Inference

Shuowen Chen[1]

EC708: PhD Econometrics I (Spring 2020)

[1] Parts of the materials are from Belloni, Chernozhukov and Hansen (2014, JEP)

Outline

- Review of Hausman Specification Test
- Hausman Test as a Pretest for Endogeneity(Guggenberger, 2010)
- Bonus: Machine Learning and Econometrics
  - Causal Inference: Double Selection Algorithm (Belloni, Chernozhukov and Hansen, 2014)
  - Prediction: Selecting among IVs in the first stage (Belloni, Chen, Chernozhukov and Hansen, 2012)

Hausman Test

Consider two estimators $\widehat{\theta}_A$ and $\widehat{\theta}_B$ and a general hypothesis H_0 against alternative H_1. If

- Under H_0: both estimators are consistent while $\widehat{\theta}_B$ is more efficient than $\widehat{\theta}_A$,
- Under H_1: only $\widehat{\theta}_A$ is consistent

Hausman test statistic:

$$H_T = T(\widehat{\theta}_B - \widehat{\theta}_A)' \left[V(\widehat{\theta}_A) - V(\widehat{\theta}_B) \right]^{-1} (\widehat{\theta}_B - \widehat{\theta}_A),$$

under H_0, $H_T \xrightarrow{d} \chi^2_K$, where $K = dim(\widehat{\theta}_B) = dim(\widehat{\theta}_A)$.

Use Hausman test as a model specification test

- Test for endogeneity
- Fixed or random effects estimators

Pretesting Question (Guggenberger, 2010, ET)

Consider the following regression

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

and we want to conduct hypothesis testing

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

but unsure about the exogeneity of X. Suppose we have K valid and strong instruments Z

$$X = Z\Pi + v$$

and conditional homoskedasticity.

- If X is exogeneous, $\widehat{\theta}_{OLS}$ and $\widehat{\theta}_{2SLS}$ are both consistent but $\widehat{\theta}_{OLS}$ is more efficient
- If X is endogenous, only $\widehat{\theta}_{2SLS}$ is consistent

A natural thought is to choose which estimator to use based on results by Hausman test on exogeneity.

Exogeneity Pretest

The null of the Hausman test:

$$H_0^{hausman} : \text{X is exogenous},$$

Hausman statistic: $H_T = \frac{T(\widehat{\theta}_{2SLS} - \widehat{\theta}_{OLS})^2}{\widehat{V}_{2SLS} - \widehat{V}_{OLS}}$

If reject $H_0^{hausman}$, use 2SLS estimator and conduct inference with 2SLS t-test. If do not reject, use OLS instead

$$t_{OLS}(\beta_0) = \frac{\sqrt{T}(\widehat{\beta}_{OLS} - \beta_0)}{\sqrt{\widehat{V}_{OLS}}}, \quad t_{2SLS}(\beta_0) = \frac{\sqrt{T}(\widehat{\beta}_{2SLS} - \beta_0)}{\sqrt{\widehat{V}_{2SLS}}}$$

In summary, the two–stage test statistic is

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

Problem: severe size distortion in the second stage. WHY?

Sources of Pretesting Problem

Three factors that affect the two–stage test

1. $\rho = corr(u_i, v_i)$: level of endogeneity
2. $\mu/\sqrt{T}$: the strength of instruments, where μ^2 is the concentration parameter. Assume $\mu/\sqrt{T} \in [\kappa, \overline{\kappa}]$: The lower bound abstracts away weak instrument complications.
3. α: nominal size of the Hausman test

The level of endogeneity is the main problem.

When and Why Does Hausman Pretest Fail?

Qualitatively speaking, X is either weakly or strongly endogenous.

- X is strongly endogeneous: Hausman test always rejects the null, and 2SLS t-stat is always used. Since instruments are strong by assumption, second–stage inference is good.
- X is weakly endogeneous: Hausman test might not be able to detect local alternatives in some directions (local power). When it cannot reject, OLS t–test is used, leading to invalid inference. Size distortion can be huge (Wong, 1997; Guggenberger, 2010)

Since in practice we don't know how endogenous X is, the two–stage test can be problematic. Increasing α and κ reduces the size distortion.

The Other Two Factors that Affect the Test

Nominal size of Hausman test: α

- Recall we reject the null if $H_T > \chi^2_{1,1-\alpha}$
- An increase in α means it's easier to reject the null
- This means we use 2SLS more often in the second stage, and hence alleviate the size distortion

Strength of Instruments: κ

- An increase in κ means the instruments are stronger
- Hence 2SLS properties are further guaranteed in finite samples.

But there is an upper limit to which we can change these parameters.

Post–Model Selection Inference

- Pretesting is a type of model selection: based on first–stage test we choose which model is work with
- Model selections are common in economics, with different selection criterion (hypothesis testing, information criterion, machine learning, etc.)
- But inference after model selection can be tricky and intractable (Davidson and McKinnon, 1993 p.97; Leeb and Pötscher, 2005)
- Subsampling and m out of n bootstrap[2] also have size problems (Andrews and Guggenberger, 2009)

[2] We will cover them later this semester.

New Paradigm: Incorporating Machine Learning

- 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 n, the sample size, the dataset is considered to be high–dimensional.
- OLS no longer feasible when $p > n$ (no unique solutions)
- Usually economists specify key variables and functional forms and conduct robustness checks afterwards
- An alternative: do semi/nonparametric econometrics[3]. Cost: curse of dimensionality
- New alternative: select key variables using machine learning methods and conduct inference on selected specifications

[3] We will probably teach these in EC709.

Causal Inference in High–Dimensional Econometrics

Consider the following linear model

$$y = \alpha D + g(X) + u,$$

- D: treatment parameter
- α: causal coefficient of interest
- X: control variables (high–dimensional)
- $g(\cdot)$: general function form

After conditioning on X, D is considered to be exogenous, and we want to estiamte and conduct inference on α.

- We impose approximate sparsity assumption: $g(X)$ can be approximated, up to some errors, by a few covariates ($< n$) whose identities are *ex-ante* unknown
- The assumption allows for imperfect model selection[4]

---

[4] In Hasuman pretest example, being unable to detect weak endogeneity and using OLS can be considered as an example.

Approximate Sparsity Assumption

$$g(X) = X\beta + r_g, \quad ||\beta||_0 \leq s, \quad \sqrt{\{\mathbb{E}(r_g^2)\}} \leq C\sqrt{s/n};$$

- The sparsity index s obeys $s^2 \log^2(\max\{p, n\})/n \to 0$.
- $||\beta||_0$: number of non-zero components of vector β
- r_g: approximation errors

The number s is defined so that the approximating error is no larger than $\sqrt{s/n}$ of the oracle estimator[5].

- The approximate sparsity is a dimension reduction assumption
- We will use LASSO to select variables among X that have non-zero coefficients

[5] Oracle means the DGP creator.

A Crash Course on LASSO

Given a collection of data $y \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n \times p}$, the LASSO estimator for least square solves the following optimization problem

$$\min_{\beta} \left\{ \frac{1}{2n} \sum_{i=1}^{n} (y_i - \alpha d_i - x_i'\beta)^2 + \lambda \sum_{j=1}^{p} |\pi_j \beta_j| \right\},$$

where $\lambda \geq 0$ is a penalty parameter and π_j is penalty loading.

- The constraint $\lambda \sum_{i=1}^{p} |\pi_i \beta_i|$ restricts the coefficients by shrinking them to zero and hence prevents overfitting.
- **We exclude D from LASSO penalty**
- The number of zero coefficients at the LASSO solution $\hat{\beta}$ depends on λ: the heavier the penalty, the more the zeros.
- λ chosen based on cross–validation or data–driven methods (Belloni, Chen, Chernozhukov and Hansen, 2012)

LASSO Coefficient for Inference?

LASSO coefficients cannot be used for inference because by construction they are biased toward zero

By Kuhn-Tucker condition, any LASSO solution $\widehat{\beta}$ satisfies $X'(y - X\widehat{\beta}) = \lambda q$, where q is subgradient of l_1 norm at $\widehat{\beta}$

$$q_i = \begin{cases} 1, & \widehat{\beta}_i > 0 \\ -1, & \widehat{\beta}_i < 0 \\ [-1, 1], & \widehat{\beta}_i = 0 \end{cases}$$

Define $B = \{i \in 1, ..., p : |q_i| = 1\}$ and assume that X_B has full rank, then $\hat{\beta}_B = (X_B' X_B)^{-1}(X_B' y - \lambda q_B)$ while $\hat{\beta}_{-B} = 0$. Therefore LASSO solution shrinks coefficients toward zero and the results depend on λ.

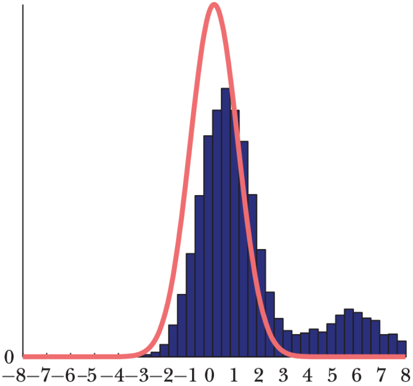
Post LASSO Model

Suppose LASSO selects the first two variables in X and researchers now work 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 $\widehat{\alpha}$. **Size Distortion**

A: A Naive Post-Model Selection Estimator



Why Post–LASSO Doesn't Give Correct Inference

Omitted variable bias

- LASSO targets prediction, so variables in X that are highly correlated with D won't be selected since adding them won't add much gain in prediction
- But if these variables actually have non–zero statistically significant coefficients in OLS, we have omitted variable bias.

Remedy: Belloni, Chernozhukov and Hansen (2014, ReStud)

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

$$y = \alpha D + X\beta + r_g + u \quad (1)$$

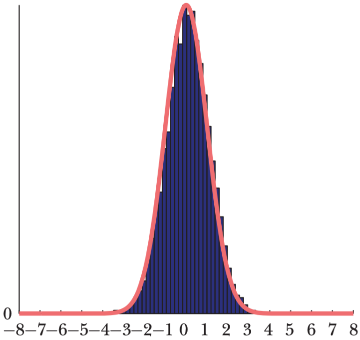
$$D = X\gamma + r_\gamma + v \quad (2)$$

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

Remedy Approach: Double Selection

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

B: A Post-Double-Selection Estimator



Machine Learning for Prediction

- Many econometric methods feature prediction steps: 2SLS, structural estimation methods like BBL
- Machine learning methods target predictions, and thus natural to exploit their advantage on prediction steps
- Consider the endogeneity model with very many candidate IVs to choose from

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

r_Z is approximation error, $\mathbb{E}[u|Z] = \mathbb{E}[v|r_Z, Z] = 0$, $\mathbb{E}[uv] \neq 0$

- The first-stage is actually a prediction problem: choose IVs that get the best fitted $\widehat{X}$, no need for inference
- Therefore we can just run LASSO in the first stage and then perform 2sls
- Belloni, Chen, Chernozhukov and Hansen (2012) formalize the intuition and establish the asymptotic properties