ECON 31200 Part B Problem Set 1

Professor James J. Heckman

Due May 14, 2020 at 5 PM CDT

1. (**10 points**) When economists build empirical models, they often "test down" to a simpler model. This practice is sometimes called "pre-testing." It is common practice in empirical economics that the testing sequences actually used are rarely reported. At the same time, proponents of RCTs insist on "pre-analysis plans" (PAP) (for example, see Christensen and Miguel (2018) and the recent NBER working paper Banerjee et al. (2020)). Explain why these proposed criteria for RCTs are a consequence of one (influential) mode of statistical inference: classical statistics. Should PAP also be applied to model specification tests (for instance, the Durbin (1954)-Wu (1973)-Hausman (1978) test)? (Hint: See posted handout, Hypothesis Testing, Part I.)

2. (**10 points**) Answer the stated questions in the handout Hypothesis Testing Part I.

3. (**10 points**) How would Bayesians, followers of the likelihood principle, and abductors build models? Is it illegitimate to use data to estimate a model, revise it and test it again on the same data? Why or why not?

4. (**5 points**) Was Feynman right in this video that social scientists (who often use pre-testing strategies) are pseudoscientists? Or did he mean their frequent failure to replicate? How could pre-testing lead to replication failure? Does publication bias erode credibility?

5. (**40 points**) For the following model,

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \equiv \begin{pmatrix} \boldsymbol{\iota} & \mathbf{X}_1 & \mathbf{X}_2 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \boldsymbol{\varepsilon}, \tag{1}$$

let N denote the number of observations, and K denote the number of covariates. Accordingly, $\mathbf{Y} \equiv (\mathrm{Y}_1, \ldots, \mathrm{Y}_N)$ is a $N \times 1$ vector, $\mathbf{X}$ is a $N \times K$ matrix containing observed covariates, $\boldsymbol{\beta}$ is a $K \times 1$ parameter vector, $\boldsymbol{\iota}$ is a $N \times 1$ vector containing 1, and $\boldsymbol{\varepsilon}$ is a $N \times 1$ vector describing unobserved variables. $\mathbf{X} \perp\!\!\!\perp \boldsymbol{\varepsilon}$ and $\mathbb{E}[\boldsymbol{\varepsilon}] = \mathbf{0}$.

Assume that $(\mathrm{X}_{1,i}, \mathrm{X}_{2,i}, \varepsilon_i)$, $i = 1, \ldots, N$ are i.i.d. random variables drawn from the following distribution:

$$\begin{pmatrix} \mathrm{X}_{1,i} \\ \mathrm{X}_{2,i} \\ \varepsilon_i \end{pmatrix} \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 & 0 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 & 0 \\ 0 & 0 & \sigma_\varepsilon^2 \end{bmatrix} \right)$$

with $\rho \in (-1, 1)$.

It is common practice in applied economics to test among competing specifications. Which of specification I or specification II is correct:

I. $\mathrm{Y}_i = \beta_0 + \mathrm{X}_{1,i}\beta_1 + \mathrm{X}_{2,i}\beta_2 + \varepsilon_i \quad i = 1, \ldots, N$.

II. $\mathrm{Y}_i = \alpha + \mathrm{X}_{1,i}\beta_1 + v_i$, where $v_i = (\varepsilon_i + X_{2,i}, \beta_2)$.

Download `PS1_Q5_Data.xlsx` from Canvas and use the provided 72 samples (simulated from models with the same $\boldsymbol{\beta}$ but with varying values of N, ρ, σ_1^2, σ_2^2, σ_ϵ^2, μ_1 and μ_2) to answer the following questions for these samples to determine which model is correct. (All models have intercepts.)

- Estimate (β_1, β_2) by OLS if $t_2 \geq 1.964$, where t_{β_2} is the t-statistic for the OLS estimate of β_2 when both X_1 and X_2 are used in the regression.
- Estimate β_1 as the value of π from the regression $\mathrm{Y} = \mathrm{X}_1\pi + v$ with $\mathbb{E}[v] = 0$ if $t_{\beta_2} < 1.964$. Denote this estimator by $\widetilde{\beta}_1$.
- The pre-test estimator is

$$\beta_1^\star \;=\; \widehat{\beta}_{1,\mathrm{OLS}}\mathbf{1}\left(t_{\beta_2} \geq 1.964\right) + \widetilde{\beta}_1\mathbf{1}\left(t_{\beta_2} < 1.964\right)$$

A. Estimate β_1 obtained from this procedure and its sampling distribution for each sample (simulated using Monte- Carlo methods for values of $\rho, \sigma_1^2, \sigma_2^2, \sigma_\varepsilon^2$, μ_1 and μ_2 which you can estimate from each sample).

B. What is the relationship of the bias for β_1 with respect to $\rho, \sigma_1^2, \sigma_2^2, \sigma_\varepsilon^2$ across samples? Characterize this bias empirically, and analytically using the posted notes on pre-testing bias.

   - Note that the DGP (or the "true model") has $\beta_1 = 1$.
   - For the analytical calculation, assume that variances are known and not estimated.
   - It may be helpful to consult Bancroft (1944).

C. How would a Bayesian approach this problem?

6. (**25 points**) Define the causal parameters for the regression model:

$$\mathrm{Y} = \beta_0 + \beta_1 \mathrm{X}_1 + \beta_2 \mathrm{X}_2 + \mathrm{U} \tag{$\star$}$$

A. Define counterfactuals.

B. Define causal parameters.

C. What is the causal effect of a unit change in X_1? The partial derivative with respect to X_1?

D. Distinguish the coefficients in ($\star$) from the coefficients in a regression of Y on $\mathrm{X}_1, \mathrm{X}_2$ and an intercept.

E. Distinguish *fixing* from *conditioning*.

F. Distinguish a thought experiment from an actual experiment.

G. Is causality defined by any particular statistical procedure?

H. Distinguish between the tasks of:

   (i) model formulation,

   (ii) identification,

   (iii) testing.

I. Answer Parts A through H for the following model:

$$\mathrm{Y} \;=\; f(\mathrm{X}, \mathrm{U}) \quad \text{where } f \;:\; (\mathrm{X}, \mathrm{U}) \;\to\; \mathrm{Y}.$$

J. When is β_1 in ($\star$) a structural parameter?

References

Bancroft, T. A. (June 1944). "On Biases in Estimation Due to the Use of Preliminary Tests of Significance". *Annals of Mathematical Statistics* 15.2, pp. 190–204. https://www.jstor.org/stable/2236199.

Banerjee, Abhijit, Esther Duflo, Amy Finkelstein, Lawrence F Katz, Benjamin A Olken, and Anja Sautmann (Apr. 2020). *In Praise of Moderation: Suggestions for the Scope and Use of Pre-Analysis Plans for RCTs in Economics*. Working Paper 26993. National Bureau of Economic Research. http://www.nber.org/papers/w26993.

Christensen, Garret and Edward Miguel (Sept. 2018). "Transparency, Reproducibility, and the Credibility of Economics Research". *Journal of Economic Literature* 56.3, pp. 920–80. https://www.aeaweb.org/articles?id=10.1257/jel.20171350.

Durbin, J. (1954). "Errors in Variables". *Revue de l'Institut International de Statistique / Review of the International Statistical Institute* 22.1, pp. 23–32. www.jstor.org/stable/1401917.

Hausman, J. A. (1978). "Specification Tests in Econometrics". *Econometrica* 46.6, pp. 1251–1271. www.jstor.org/stable/1913827.

Wu, De-Min (1973). "Alternative Tests of Independence between Stochastic Regressors and Disturbances". *Econometrica* 41.4, pp. 733–750. www.jstor.org/stable/1914093.