EC708 Discussion 8
Non-Regular Estimators

Yan Liu

Department of Economics
Boston University

March 31, 2023

Outline

1. Non-Regular Estimators
   - Hodges Super-Efficient Estimator
   - Post Model Selection Estimator
2. Selected Questions from PS3

Table of Contents

Non-Regular Estimators

Hodges Super-Efficient Estimator

Suppose $(X_1, \ldots, X_T) \overset{\text{i.i.d.}}{\sim} N(\theta_0, 1)$. Define

$$\hat{\theta}_T = \begin{cases} \bar{X}_T & \text{if } |\bar{X}_T| \geq T^{-1/4} \\ 0 & \text{if } |\bar{X}_T| < T^{-1/4} \end{cases}.$$

Why truncate? Note that $\sqrt{T}(\bar{X}_T - \theta_0) \sim N(0, 1)$.

$$\begin{aligned} \Pr(\hat{\theta}_T = 0) &= \Pr(|\bar{X}_T| < T^{-1/4}) \\ &= \Phi(\sqrt{T}(T^{-1/4} - \theta_0)) - \Phi(\sqrt{T}(-T^{-1/4} - \theta_0)). \end{aligned}$$

- If $\theta_0 \neq 0$, $\Pr(\hat{\theta}_T = 0) \to 0 \Rightarrow \hat{\theta}_T$ behaves the same as $\bar{X}_T$.
- If $\theta_0 = 0$, $\Pr(\hat{\theta}_T = 0) \to 1 \Rightarrow \hat{\theta}_T$ converges to θ_0 "arbitrarily fast".

Non-Regular Estimators

Hodges Super-Efficient Estimator

- Quadratic risk function: $\theta_0 \mapsto E_{\theta_0}(\hat{\theta}_T - \theta_0)^2$
- $\hat{\theta}_T$ “buys” its better asymptotic behavior at $\theta_0 = 0$ at the expense of erratic behavior close to zero

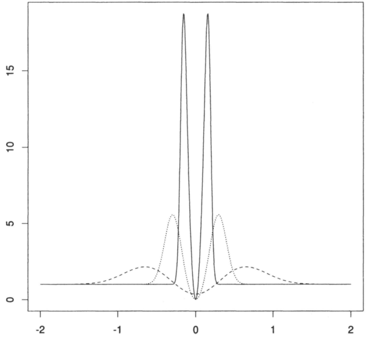


Figure 8.1. Quadratic risk functions of the Hodges estimator based on the means of samples of size 10 (dashed), 100 (dotted), and 1000 (solid) observations from the $N(\theta, 1)$-distribution.

Non-Regular Estimators

Hodges Super-Efficient Estimator

Is $\hat{\theta}_T$ regular? Consider $\theta_T = h/\sqrt{T}$ and $(X_1, \dots, X_T) \overset{\text{i.i.d.}}{\sim} N(\theta_T, 1)$. Then,

$$\sqrt{T}(\bar{X}_T - \theta_T) = \sqrt{T}\bar{X}_t - h \sim N(0, 1).$$

Hence,

$$\begin{aligned} \Pr(\hat{\theta}_T = 0) &= \Pr(|\bar{X}_T| < T^{-1/4}) \\ &= \Phi(T^{1/4} - h) - \Phi(-T^{1/4} - h) \to 1. \end{aligned}$$

This implies

$$\sqrt{T}(\hat{\theta}_T - \theta_T) \xrightarrow{d} -h \quad \Rightarrow \quad \hat{\theta}_T \text{ is not regular.}$$

Non-Regular Estimators

Post Model Selection Estimator

Consider the linear model:

$$Y_t = D_t\alpha + X_t'\beta + U_t, \quad E[U_t|D_t, X_t] = 0.$$

- D_t: treatment/policy variable of interest
- X_t: $p \times 1$ control variables, where p can be larger than T: $p \gg T$.
- Approximate sparsity assumption: D_t is exogenous after controlling for a small number $s < T$ of variables in X_t
- Natural idea: variable selection via Lasso

Lasso solves

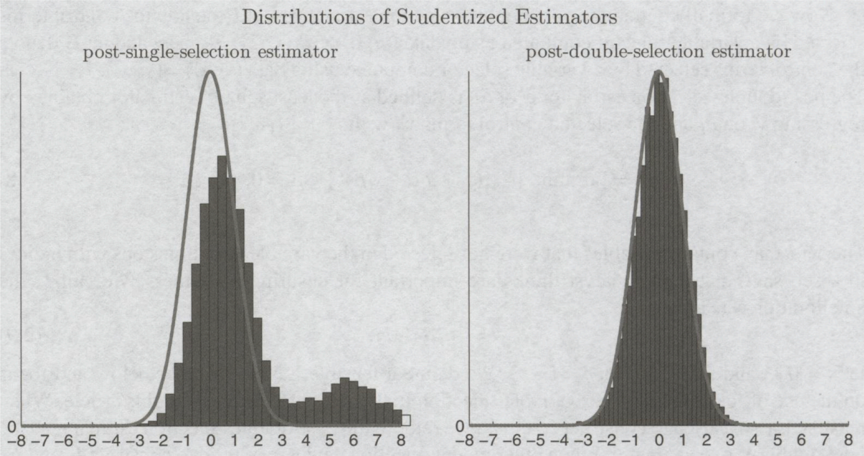
$$(\hat{\alpha}, \hat{\beta}) = \underset{\alpha \in \mathbb{R}, \beta \in \mathbb{R}^p}{\arg\min} \sum_{t=1}^{T} (Y_t - D_t \alpha - X_t' \beta)^2 + \lambda \sum_{j=1}^{p} |l_j \beta_j|,$$

where λ is penalty level and l_j's are loadings.

- Non-differentiability of the penalty function at zero induces $\hat{\beta}$ to have components set exactly to zero
- Let $\hat{I}$ denote the nonzero components of $\hat{\beta} \Rightarrow$ selected controls
- Post-single-selection estimator $\tilde{\alpha}$ of α:

$$(\tilde{\alpha}, \tilde{\beta}) = \underset{\alpha \in \mathbb{R}, \beta \in \mathbb{R}^p}{\arg\min} \left\{ \sum_{t=1}^{T} (Y_t - D_t \alpha - X_t' \beta)^2 : \beta_j = 0, \forall j \notin \hat{I} \right\}.$$

Conventional t-test based on $\tilde{\alpha}$ has size distortion (left panel)



What are the problems with the naive single selection approach?

- It ignores the relationship between D_t and X_t
- It is based on a "structural model" where the target is to learn the treatment effects given controls, while Lasso targets prediction

$\Rightarrow$ Work with a reduced form, predictive equation system:

$$Y_t = D_t\alpha + X_t'\beta + U_t = X_t'(\underbrace{\alpha\pi + \beta}_{=\theta}) + \underbrace{\alpha V_t + U_t}_{=\varepsilon_t},$$

$$D_t = X_t'\pi + V_t,$$

where $E[\varepsilon_t|X_t] = 0$ and $E[V_t|X_t] = 0$.

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

1. Select controls that predict $D_t \Rightarrow \hat{I}_1$:

$$\min_{\pi \in \mathbb{R}^p} \sum_{t=1}^{T} (D_t - X_t'\pi)^2 + \lambda \sum_{j=1}^{p} |l_j^d \pi_j|.$$

2. Select controls that predict $Y_t \Rightarrow \hat{I}_2$:

$$\min_{\theta \in \mathbb{R}^p} \sum_{t=1}^{T} (Y_t - X_t'\theta)^2 + \lambda \sum_{j=1}^{p} |l_j^y \theta_j|.$$

3. Post-double-selection estimator $\check{\alpha}$ of α:

$$(\check{\alpha}, \check{\beta}) = \underset{\alpha \in \mathbb{R}, \beta \in \mathbb{R}^p}{\arg\min} \left\{ \sum_{t=1}^{T} (Y_t - D_t\alpha - X_t'\beta)^2 : \beta_j = 0, \forall j \notin \hat{I}_1 \cup \hat{I}_2 \right\}.$$

Why is double selection important? Consider the model with one control:

$$
\begin{aligned}
Y_t &= D_t\alpha + X_t\beta + U_t, \\
D_t &= X_t\pi + V_t,
\end{aligned}
$$

where

$$
\begin{bmatrix} U_t \\ V_t \end{bmatrix} \Bigg| X_t \sim N\left(0, \begin{bmatrix} \sigma_u^2 & 0 \\ 0 & \sigma_v^2 \end{bmatrix}\right), \quad X_t \sim N(0,1).
$$

Then, X_t and D_t are jointly normal with $\sigma_d^2 = \pi^2 + \sigma_v^2$ and correlation $\rho = \pi/\sigma_d$.

Single selection drops X_t wp$\to 1$ if

$$\beta \leq \frac{\sqrt{\log T}}{\sqrt{T}} \frac{\sigma_u}{\sqrt{1-\rho^2}}.$$

- In low-dimensional settings, implemented with a conservative t-test: drop X_t if $|t| = \hat{\beta}/\,\mathrm{se}(\hat{\beta}) \leq \Phi^{-1}(1-1/(2T)) = \sqrt{2\log T}(1+o(1))$
- In high-dimensional settings, implemented with Lasso

Non-Regular Estimators

Post Model Selection Estimator

Consider a sequence $\beta_T = \frac{\sqrt{\log T}}{\sqrt{T}} \frac{\sigma_u}{\sqrt{1-\rho^2}}$.

- t-test cannot distinguish β_T from 0 and drops X_t wp$\to 1$.
- In this case, post-single-selection estimator $\tilde{\alpha}$ performs poorly:

$$\begin{aligned}
\sqrt{T}(\tilde{\alpha} - \alpha) &= \frac{\frac{1}{\sqrt{T}} \sum_{t=1}^{T} D_t(X_t\beta_T + U_t)}{\frac{1}{T}\sum_{t=1}^{T} D_t^2} \\
&= \underbrace{\sqrt{T}\beta_T \cdot \frac{\frac{1}{T}\sum_{t=1}^{T} D_t X_t}{\frac{1}{T}\sum_{t=1}^{T} D_t^2}}_{\geq \frac{1}{2}\sqrt{\log T}\frac{\sigma_u}{\sqrt{1-\rho^2}}\frac{|\rho|}{\sigma_d} \text{ wp}\to 1} + \underbrace{\frac{\frac{1}{\sqrt{T}}\sum_{t=1}^{T} D_t U_t}{\frac{1}{T}\sum_{t=1}^{T} D_t^2}}_{\xrightarrow{d} N(0, \sigma_u^2/\sigma_d^2)} \propto \sqrt{\log T} \to \infty.
\end{aligned}$$

Non-Regular Estimators

Post Model Selection Estimator

Double selection drops X_t with positive probability only if

$$\text{both } |\beta| < \frac{\sqrt{\log T}}{\sqrt{T}} \frac{\sigma_u}{\sqrt{1-\rho^2}} \text{ and } |\pi| < \frac{\sqrt{\log T}}{\sqrt{T}} \sigma_v.$$

When X_t is dropped, wp$\rightarrow 1$,

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} D_t X_t \beta \leq 2\sqrt{T} |\pi \beta| \propto \frac{\log T}{\sqrt{T}} \rightarrow 0.$$

Hence, post-double-selection estimator $\check{\alpha}$ is asymptotically normal:

$$\sqrt{T}(\check{\alpha} - \alpha) = \frac{\frac{1}{\sqrt{T}} \sum_{t=1}^{T} D_t (X_t \beta + U_t)}{\frac{1}{T} \sum_{t=1}^{T} D_t^2} \xrightarrow{d} N(0, \sigma_u^2 / \sigma_d^2).$$

- When p is small relative to T, post-double selection estimator is first-order equivalent to the full regression.
- When $p \propto T$ or $p \gg T$, this equivalence disappears, but the post-double selection method continues to be regular.

Table of Contents

PS3 Q2

2:

Consider a panel data model with $1 \leq i \leq n$ individuals, each observed for $1 \leq t \leq T$ periods. Let $Y_i = (Y_{i1}, \ldots, Y_{iT})'$, with $Y_{it} \in \mathbb{R}$, and $X_i = (X_{i1}, \ldots, X_{iT})'$ with $X_{it} \in \mathbb{R}^{d_X}$. Further, suppose

$$Y_{it} = X_{it}'\beta + \alpha_i + U_{it} \ , \tag{4}$$

where $U_i = (U_{i1}, \ldots, U_{iT})' \in \mathbb{R}^T$ is distributed according to $U_i \sim N(0, \Sigma)$ for some unknown covariance matrix Σ. Throughout this problem, assume $\{Y_i, X_i, U_i\}_{i=1}^N$ are i.i.d. and U_i is independent of X_i. We will treat β, Σ and $\alpha = (\alpha_1, \ldots, \alpha_N)' \in \mathbb{R}^N$ as the unknown parameters to be estimated.

PS3 Q2

(a) Write down the log-likelihood for this problem.

(b) For simplicity, assume $\text{Var}(U_{it}) = \sigma^2$ for all $1 \leq t \leq T$, and write down an explicit form for:

$$\sum_{i=1}^{n} \sum_{t=1}^{T} \log\{f(Y_{it}|X_i)\} \ , \tag{5}$$

where $f(Y_{it}|X_i)$ denotes the density of Y_{it} given X_i. This partial log-likelihood differs from your answer to (a) because it does not fully use U_i's joint distribution, but this likelihood can be used to estimate some of the model parameters. Explain why the partial likelihood in (5) cannot: (i) fully identify β if elements of X_{it} are time-invariant for all t, (ii) fully identify Σ.

(c) In what follows, assume X_{it} does not contain time invariant elements. Derive closed-form expressions for the maximum (partial) likelihood estimators of $(\beta, \alpha, \sigma^2)$ obtained by using part (b). Going forward, we will denote these estimators by $(\hat{\beta}, \hat{\alpha}, \hat{\sigma}^2)$.

(d) Establish the asymptotic normality of $\sqrt{N}(\hat{\beta} - \beta)$ using an asymptotic framework in which $N \to \infty$ but T remains fixed. State what assumptions you need to impose.

(e) Is $\hat{\alpha}_i$ an unbiased estimator of α_i? Is it consistent as $N \to \infty$ but T is fixed?

PS3 Q2(a)

Let $\mathbf{1}_T$ be a $T \times 1$ vector with 1 in every component. Observing that

$$Y_i | X_i \sim N(X_i\beta + \alpha_i \mathbf{1}_T, \Sigma),$$

the likelihood of Y_i given X_i is

$$L_i(\beta, \alpha, \Sigma) = f_i(Y_i | X_i; \beta, \alpha, \Sigma)$$
$$= (2\pi)^{-T/2} \det(\Sigma)^{-1/2} \exp\Big\{ -\frac{1}{2}(Y_i - X_i\beta - \alpha_i \mathbf{1}_T)'\Sigma^{-1}(Y_i - X_i\beta - \alpha_i \mathbf{1}_T)\Big\}.$$

Since $\{Y_i, X_i, U_i\}_{i=1}^N$ are i.i.d., the log-likelihood function is

$$\bar{\ell}_N(\beta, \alpha, \Sigma) = \sum_{i=1}^N \ln L_i(\beta, \alpha, \Sigma)$$
$$= -\frac{NT}{2}\ln(2\pi) - \frac{N}{2}\ln(\det(\Sigma)) - \frac{1}{2}\sum_{i=1}^N (Y_i - X_i\beta - \alpha_i \mathbf{1}_T)'\Sigma^{-1}(Y_i - X_i\beta - \alpha_i \mathbf{1}_T).$$

PS3 Q2(b)

The partial log-likelihood is

$$\tilde{\ell}_N(\beta, \alpha, \sigma^2) = \sum_{i=1}^{N} \sum_{t=1}^{T} \ln f(Y_{it}|X_{it})$$

$$= -\frac{NT}{2} \ln(2\pi) - \frac{NT}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{N} \sum_{t=1}^{T} (Y_{it} - X_{it}'\beta - \alpha_i)^2.$$

1. For simplicity, let $X_{it} \equiv X_i$. Take any $\beta^{(1)} \neq \beta^{(2)}$. Let $\alpha^{(1)}, \alpha^{(2)} \in \mathbb{R}^N$ satisfy $\alpha_i^{(2)} - \alpha_i^{(1)} = X_i'(\beta^{(1)} - \beta^{(2)})$ for each i. Then, $\tilde{\ell}_N(\beta^{(1)}, \alpha^{(1)}, \sigma^2) = \tilde{\ell}_N(\beta^{(2)}, \alpha^{(2)}, \sigma^2)$. Hence, β is not identified.
2. Since $\tilde{\ell}_N(\beta, \alpha, \sigma^2)$ does not depend on the off-diagonal elements of Σ, these elements are not identified.

PS3 Q2(c)

The FOCs are

$$\frac{\partial \tilde{\ell}_N(\hat{\beta}, \hat{\alpha}, \hat{\sigma^2})}{\partial \beta} = \frac{1}{\hat{\sigma}^2} \sum_{i=1}^{N} \sum_{t=1}^{T} X_{it}(Y_{it} - X'_{it}\hat{\beta} - \hat{\alpha}_i) = 0, \quad (1)$$

$$\frac{\partial \tilde{\ell}_N(\hat{\beta}, \hat{\alpha}, \hat{\sigma^2})}{\partial \alpha_i} = \frac{1}{\hat{\sigma}^2} \sum_{t=1}^{T} (Y_{it} - X'_{it}\hat{\beta} - \hat{\alpha}_i) = 0, \quad i = 1, \ldots, N, \quad (2)$$

$$\frac{\partial \tilde{\ell}_N(\hat{\beta}, \hat{\alpha}, \hat{\sigma^2})}{\partial \sigma^2} = -\frac{NT}{2\hat{\sigma}^2} + \frac{1}{2\hat{\sigma}^4} \sum_{i=1}^{N} \sum_{t=1}^{T} (Y_{it} - X'_{it}\hat{\beta} - \hat{\alpha}_i)^2 = 0. \quad (3)$$

PS3 Q2(c)

Let $\bar{Y}_i = \frac{1}{T}\sum_{t=1}^{T} Y_{it}$ and $\bar{X}_i = \frac{1}{T}\sum_{t=1}^{T} X_{it}$. By (2),

$$\hat{\alpha}_i = \bar{Y}_i - \bar{X}_i'\hat{\beta}, \quad i = 1, \ldots, N. \tag{4}$$

Plugging (4) into (1),

$$\hat{\beta} = \Big[\sum_{i=1}^{N}\sum_{t=1}^{T} X_{it}(X_{it} - \bar{X}_i)'\Big]^{-1} \sum_{i=1}^{N}\sum_{t=1}^{T} X_{it}(Y_{it} - \bar{Y}_i).$$

By (3),

$$\hat{\sigma}^2 = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(Y_{it} - X_{it}'\hat{\beta} - \hat{\alpha}_i)^2.$$

PS3 Q2(d)

We need to impose the following assumptions:

- $\text{rank}(E[X_i' Q_T X_i]) = d_X$;
- $E[\|X_{it}\|^2] < \infty$.

We can write

$$\sqrt{N}(\hat{\beta} - \beta) = \underbrace{\Big(\frac{1}{N} \sum_{i=1}^{N} X_i' Q_T X_i \Big)}_{\xrightarrow{p} E[X_i' Q_T X_i]}{}^{-1} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} X_i' Q_T U_i.$$

We can calculate $\text{Var}(X_i' Q_T U_i) = E[X_i' Q_T \Sigma Q_T X_i]$. By the CLT,

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} X_i' Q_T U_i \xrightarrow{d} N(0, E[X_i' Q_T \Sigma Q_T X_i]).$$

Put together, by Slutsky's theorem,

$$\sqrt{N}(\hat{\beta} - \beta) \xrightarrow{d} N(0, E[X_i' Q_T X_i]^{-1} E[X_i' Q_T \Sigma Q_T X_i] E[X_i' Q_T X_i]^{-1}).$$

PS3 Q2(e)

Let $\bar{U}_i = \frac{1}{T}\sum_{t=1}^{T} U_{it}$. We have

$$\hat{\alpha}_i - \alpha_i = \bar{Y}_i - \bar{X}_i'\hat{\beta} - \alpha_i = -\bar{X}_i'(\hat{\beta} - \beta) + \bar{U}_i.$$

Note that

$$E[\hat{\beta} - \beta | \mathbf{X}] = \Big(\sum_{i=1}^{N} X_i' Q_T X_i\Big)^{-1} \sum_{i=1}^{N} X_i' Q_T E[U_i] = 0.$$

By the law of iterated expectations,

$$E[\hat{\alpha}_i - \alpha_i] = E[\bar{X}_i'(\hat{\beta} - \beta) + \bar{U}_i] = -E[\bar{X}_i' E[\hat{\beta} - \beta | \mathbf{X}]] + E[\bar{U}_i] = 0.$$

Hence, $\hat{\alpha}_i$ is an unbiased estimator of α_i. Since $\hat{\beta} - \beta \xrightarrow{p} 0$ by part (d), as $N \to \infty$ but T is fixed, $\hat{\alpha}_i - \alpha_i \xrightarrow{p} \bar{U}_i \neq 0$. Hence, $\hat{\alpha}_i$ is not consistent for α.

PS3 Q4

4: A log-likelihood function $\bar{\ell}_T(\theta_1, \theta_2)$ is a function of two sets of parameters θ_1 and θ_2. Define $\theta_2^*(\theta_1)$ by the identity

$$\frac{\partial \bar{\ell}_T(\theta_1, \theta_2)}{\partial \theta_2}\Big|_{\theta_2 = \theta_2^*(\theta_1)} = 0$$

and define $\bar{\ell}_T^*(\theta_1) = \bar{\ell}_T(\theta_1, \theta_2^*(\theta_1))$.

(a) Show that maximizing $\bar{\ell}_T$ with respect to θ_1 and θ_2 is the same as maximizing $\bar{\ell}_T^*$ with respect to θ_1.

(b) What is $\frac{\partial \bar{\ell}_T^*(\theta_1)}{\partial\ \theta_1}$? Show how it is related to a partial derivative of $\bar{\ell}_T$. (Hint: Use the envelope theorem.)

(c) Show that $H^* = H_{11} - H_{12} H_{22}^{-1} H_{21}$ where H^* and H are the Hessians of $\bar{\ell}_T^*$ and $\bar{\ell}_T$ and H is partitioned in an obvious way.

(d) Discuss the advantages of applying the Newton-Raphson method before or after concentration.

PS3 Q4(a)

Let

$$(\hat{\theta}_1, \hat{\theta}_2) = \arg\max_{\theta_1,\theta_2} \bar{\ell}_T(\theta_1, \theta_2), \quad \tilde{\theta}_1 = \arg\max_{\theta_1} \bar{\ell}_T^*(\theta_1), \quad \tilde{\theta}_2 = \theta_2^*(\tilde{\theta}_1).$$

The FOCs for $(\hat{\theta}_1, \hat{\theta}_2)$ are

$$\frac{\partial \bar{\ell}_T(\theta_1, \theta_2)}{\partial \theta_1}\Big|_{(\theta_1,\theta_2)=(\hat{\theta}_1,\hat{\theta}_2)} = 0, \quad \frac{\partial \bar{\ell}_T(\theta_1, \theta_2)}{\partial \theta_2}\Big|_{(\theta_1,\theta_2)=(\hat{\theta}_1,\hat{\theta}_2)} = 0.$$

PS3 Q4(a)

On the other hand, by the envelope theorem, for any θ_1,

$$\frac{\partial \bar{\ell}_T^*(\theta_1)}{\partial \theta_1} = \frac{\partial \bar{\ell}_T(\theta_1, \theta_2)}{\partial \theta_1}\Big|_{\theta_2=\theta_2^*(\theta_1)} + \underbrace{\frac{\partial \bar{\ell}_T(\theta_1, \theta_2)}{\partial \theta_2}\Big|_{\theta_2=\theta_2^*(\theta_1)}}_{=0} \frac{\partial \theta_2^*(\theta_1)}{\partial \theta_1}. \tag{5}$$

Hence, the FOC for $\tilde{\theta}_1$ is

$$0 = \frac{\partial \bar{\ell}_T^*(\tilde{\theta}_1)}{\partial \theta_1} = \frac{\partial \bar{\ell}_T(\theta_1, \theta_2)}{\partial \theta_1}\Big|_{\theta_1=\tilde{\theta}_1, \theta_2=\theta_2^*(\tilde{\theta}_1)} = \frac{\partial \bar{\ell}_T(\theta_1, \theta_2)}{\partial \theta_1}\Big|_{(\theta_1,\theta_2)=(\tilde{\theta}_1,\tilde{\theta}_2)}.$$

Also, by the definition of $\theta_2^*(\theta_1)$, the FOC for $\tilde{\theta}_2$ is

$$0 = \frac{\partial \bar{\ell}_T(\theta_1, \theta_2)}{\partial \theta_2}\Big|_{\theta_1=\tilde{\theta}_1, \theta_2=\theta_2^*(\tilde{\theta}_1)} = \frac{\partial \bar{\ell}_T(\theta_1, \theta_2)}{\partial \theta_2}\Big|_{(\theta_1,\theta_2)=(\tilde{\theta}_1,\tilde{\theta}_2)}.$$

Therefore, the FOCs for $(\hat{\theta}_1, \hat{\theta}_2)$ and $(\tilde{\theta}_1, \tilde{\theta}_2)$ coincide.

PS3 Q4(c)

We know that $\theta_2^*(\theta_1)$ satisfies that for all θ_1,

$$\frac{\partial \bar{\ell}_T(\theta_1, \theta_2^*(\theta_1))}{\partial \theta_2} = 0.$$

Differentiating with respect to θ_1 yields

$$\begin{aligned}
0 &= \frac{\partial^2 \bar{\ell}_T(\theta_1, \theta_2^*(\theta_1))}{\partial \theta_2 \partial \theta_1'} \\
&= \underbrace{\frac{\partial^2 \bar{\ell}_T(\theta_1, \theta_2)}{\partial \theta_2 \partial \theta_1'}\Big|_{\theta_2 = \theta_2^*(\theta_1)}}_{=H_{21}} + \underbrace{\frac{\partial^2 \bar{\ell}_T(\theta_1, \theta_2)}{\partial \theta_2 \partial \theta_2'}\Big|_{\theta_2 = \theta_2^*(\theta_1)}}_{=H_{22}} \frac{\partial \theta_2^*(\theta_1)}{\partial \theta_1'}.
\end{aligned}$$

Hence,

$$\frac{\partial \theta_2^*(\theta_1)}{\partial \theta_1'} = -H_{22}^{-1} H_{21}. \tag{6}$$

PS3 Q4(c)

On the other hand, differentiating (5) with respect to θ_1 yields

$$H^* = \frac{\partial^2 \bar{\ell}_T^*(\theta_1)}{\partial \theta_1 \partial \theta_1'} = \underbrace{\frac{\partial^2 \bar{\ell}_T(\theta_1, \theta_2)}{\partial \theta_1 \partial \theta_1'}\bigg|_{\theta_2 = \theta_2^*(\theta_1)}}_{=H_{11}} + \underbrace{\frac{\partial^2 \bar{\ell}_T(\theta_1, \theta_2)}{\partial \theta_1 \partial \theta_2'}\bigg|_{\theta_2 = \theta_2^*(\theta_1)}}_{=H_{12}} \frac{\partial \theta_2^*(\theta_1)}{\partial \theta_1'}. \tag{7}$$

Plugging (6) into (7) yields

$$H^* = H_{11} - H_{12} H_{22}^{-1} H_{21}.$$

PS3 Q4(d)

- Newton-Raphson before concentration uses

$$\hat{\theta}_{k+1} = \hat{\theta}_k - H(\hat{\theta}_k)^{-1} s(\hat{\theta}_k).$$

$\Rightarrow$ solve a $(K_1 + K_2)$-dimensional linear system $H(\hat{\theta}_k)\Delta = -s(\hat{\theta}_k)$.

- Newton-Raphson after concentration uses

$$\tilde{\theta}_{1,k+1} = \tilde{\theta}_{1,k} - H^*(\tilde{\theta}_{1,k})^{-1} s^*(\tilde{\theta}_{1,k}).$$

$\Rightarrow$ solve a K_1-dimensional linear system $H^*(\tilde{\theta}_{1,k})\Delta = -s^*(\tilde{\theta}_{1,k})$.