Econometrics Preliminary Exam

Agricultural and Resource Economics, UC Davis

August, 2021

There are **THREE** questions. Choose and answer two of the three questions. Within each question, each part will receive equal weight in grading. You have 15 minutes to read the exam and then three hours to complete the exam.

I. Probability and Statistics

(a) Suppose a population was 95% vaccinated against a virus. An i.i.d. sample of n individuals was drawn from the population and each individual was exposed to the virus (with the same level of exposure). Suppose 462 of them got infected, among whom 342 were vaccinated.

i. What is the risk (or probability) of infection among the vaccinated? What is the risk of infection among the unvaccinated? Calculate the 95% confidence interval of the risk of infection among the vaccinated. Use critical value 1.96.

ii. Vaccine efficacy is defined as

$$\frac{\text{risk of infection among the unvaccinated - risk of infection among the vaccinated}}{\text{risk of infection among the unvaccinated}}.$$

Calculate the vaccine efficacy in this case.

(b) Consider scalar continuous random variables X and Y. Their joint distribution is

$$f(x,y) = \begin{cases} 1, & \text{if } 0 \leqslant x \leqslant a \text{ and } 0 \leqslant y \leqslant b \\ 0, & \text{otherwise} \end{cases},$$

where a and b are constants.

i. What relationship should a and b satisfy?

ii. Are X and Y independent? Why or why not?

iii. Suppose $a = b = 1$. Calculate the probability that $X - Y > 0.5$.

(c) Now we formalize the set-up in the first question restricting the analysis to the vaccinated group. Suppose there is an i.i.d. sample of n vaccinated individuals exposed to the virus to the same extent. Let $\{X_i\}_{i=1}^n$ be the random variable indicating whether individual i got infected; $X_i = 1$ if infected and $X_i = 0$ if not. Suppose the true risk of infection of the group is θ.

i. Derive the MLE estimator of θ. Call it $\hat{\theta}$. Then derive the asymptotic distribution of $\hat{\theta}$.

ii. Derive the moment generating function of the random variable X_i. Then, calculate the first two moments of X_i using the moment generating function.

iii. Derive the test statistic of the likelihood ratio test for the null $H_0 : \theta \leqslant \theta_0$ against the alternative $H_1 : \theta > \theta_0$. Show that based on the likelihood ratio test, one would reject the null if $\bar{X} \equiv \frac{1}{n}\sum_{i=1}^{n} X_i$ is too big.

II. **Linear Regression**

Consider the model $y = X\beta + e$, $E[e|X] = 0$, $E[ee'] = c^2\Omega$, where y is $n \times 1$ and X is $n \times k$. The scalar c and the matrix Ω are known. You have an *iid* random sample of size n.

(a) Is $\hat{\beta} = (X'X)^{-1}X'y$ unbiased for β? If so, prove it. If not, state additional conditions you require for unbiasedness and prove unbiasedness under those conditions.

(b) Derive the variance of $\hat{\beta}$ conditional on X. State any assumptions you need.

(c) Derive the variance of the GLS estimator $\tilde{\beta} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y$ conditional on X. State any assumptions you need.

(d) Derive the variance of $\hat{\beta} - \tilde{\beta}$ conditional on X. State any assumptions you need.

(e) Consider the estimator $\bar{\beta} = \theta\hat{\beta} + (1 - \theta)\tilde{\beta}$, where θ is a scalar constant. What value of θ minimizes the variance of $\bar{\beta}$ conditional on X? State any assumptions you need.

(f) Is $\hat{\beta}$ consistent for β? If so, prove it. If not, state additional conditions you require for consistency and prove consistency under those conditions.

(g) Find the asymptotic distribution of $\sqrt{n}(\hat{\beta} - \beta)$ as $n \to \infty$. You may carry over any assumptions you made in (e) to show consistency. State any additional assumptions you require.

(h) Suppose c and Ω are unknown. Propose a consistent estimator for Ω and prove that it is consistent. State any assumptions you need.

III. Nonlinear Estimation and Panel Data Methods

(a) **Maximum Likelihood Estimation of Binary Outcome Models**. Suppose that for $i = 1, \ldots, n$, $y_i^* = x_i'\beta_0 + u_i$, where $dim(x_i) = dim(\beta_0) = k$. The researcher can *only* observe a binary version of y_i^*, specifically $y_i = 1\{y_i^* \geqslant 0\}$, and x_i, where $1\{A\}$ equals 1 when the event A holds and zero otherwise.
Suppose that $u_i|x_i \overset{i.i.d.}{\sim} N(0,1)$.
Note: In your answer, you can use the following notation for the standard normal cdf and pdf, $\Phi(z)$ and $\phi(z) \equiv \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$, respectively. You can also maintain the cross-sectional i.i.d. assumption.

(i) Propose a maximum likelihood estimator of β_0, $\hat{\beta}$. Provide sufficient conditions for its consistency.

(ii) Derive an expression for $\sqrt{n}(\hat{\beta} - \beta_0)$ and provide sufficient conditions for its asymptotic normality as $n \to \infty$. Explain how the conditions imply the result. Make sure to state the asymptotic distribution.

(iii) Propose two different estimators of the asymptotic variance you provide in (ii). Briefly discuss the conditions required for their consistency. (A discussion of high-level conditions is sufficient, no need to provide primitive conditions.)

(iv) Propose the Wald and likelihood ratio statistics to test $H_0 : \beta_0 = c$, where c is a non-random vector. For each test statistic, state its asymptotic distribution under the null hypothesis and make sure to define all quantities that your statistic consists of.

(b) **Slope Heterogeneity in Cluster Data**. When estimating regression models from cluster data, $y_{ic} = x_{ic}'\beta_c + a_c + u_{ic}$, where for each cluster $c = 1, \ldots, G$, we observe individuals $i = 1, \ldots, n$. That is, for simplicity, we assume that there is an identical number of observations per cluster. Let $dim(x_{ic}) = dim(\beta_c) = k$ for all $c = 1, \ldots, G$.
All asymptotics in this question pertain to $n \to \infty$, while holding G fixed. You can assume the i.i.d. assumption across i within each cluster c. You can also assume that the clusters are independent, but you cannot assume that the clusters are identically distributed.

(i) Derive the probability limit for the following two estimators:

$$\hat{\beta}_{FE} = \left(\sum_{c=1}^{G} \sum_{i=1}^{n} (x_{ic} - \bar{x}_c)(x_{ic} - \bar{x}_c)' \right)^{-1} \sum_{c=1}^{G} \sum_{i=1}^{n} (x_{ic} - \bar{x}_c)(y_{ic} - \bar{y}_c) \quad (1)$$

$$\hat{\beta}_{SA} = \frac{1}{G} \left[\sum_{c=1}^{G} \left(\sum_{i=1}^{n} (x_{ic} - \bar{x}_c)(x_{ic} - \bar{x}_c)' \right)^{-1} \sum_{i=1}^{n} (x_{ic} - \bar{x}_c)(y_{ic} - \bar{y}_c) \right] \quad (2)$$

where for a random variable w_{ic}, $\bar{w}_c = \sum_{i=1}^{n} w_{ic}/n$. You can assume that the regressors are exogenous, $E[u_{ic}|x_{ic}, a_c] = 0$.

Provide all sufficient conditions to derive the probability limits invoking laws of large numbers and/or central limit theorems wherever appropriate.

Briefly describe the conditions under which both estimators would be consistent for $\beta_0 = \sum_{c=1}^{G} \beta_c/G$, the simple average of the heterogeneous slopes.

(ii) Propose a Hausman-type test of the equality of $\beta_{FE} = \beta_{SA}$, where $\beta_{FE} \equiv plim_{n\to\infty}\hat{\beta}_{FE}$ and $\beta_{SA} \equiv plim_{n\to\infty}\hat{\beta}_{SA}$. Instead of using the classical form of the statistic which requires additional restrictions to ensure the asymptotic efficiency of one of the estimators in question, you will derive the test statistic and its null distribution without imposing those additional restrictions.

To do so, follow these steps:

Step 1: derive the asymptotic joint distribution of the sampling error of $\hat{\beta}_{FE}$ and $\hat{\beta}_{SA}$ under H_0 providing sufficient conditions that ensure the validity of the asymptotic distribution;

Step 2: apply the delta method to derive the asymptotic distribution of $\sqrt{n}(\hat{\beta}_{FE} - \hat{\beta}_{SA})$ under H_0,

Step 3: write down the Hausman test statistic and state its asymptotic distribution under H_0.

ARE/ECN 240B Reference Sheet

Notation. θ_0, Θ, y_i, x_i, $s(y_i, x_i; \theta)$ and $H(y_i, x_i; \theta)$ pertain to the objects defined in the 240B lecture notes.

Assumption ULLN 1 $\sup_{\theta \in \Theta} |\sum_{i=1}^n f(y_i, x_i; \theta)/n - E[f(y_i, x_i; \theta)]| \xrightarrow{p} 0$, if the following conditions hold,

(i) *(i.i.d.)* $\{y_i, x_i\}_{i=1}^n$ is an i.i.d. sequence of random variables;

(ii) *(Compactness)* Θ is compact;

(iii) *(Continuity)* $f(y_i, x_i; \theta)$ is continuous in θ for all $(y_i, x_i')'$;

(iv) *(Measurability)* $f(y_i, x_i; \theta)$ is measurable in $(y_i, x_i')'$ for all $\theta \in \Theta$;

(v) *(Dominance)* There exists a dominating function $d(y_i, x_i)$ such that $|f(y_i, x_i; \theta)| \leqslant d(y_i, x_i)$ for all $\theta \in \Theta$ and $E[d(y_i, x_i)] < \infty$.

Assumption ULLN 2 $\sup_{\theta \in \Theta} |\sum_{i=1}^n f(y_i, x_i; \theta)/n - E[f(y_i, x_i; \theta)]| \xrightarrow{p} 0$, if the following conditions hold,

(i) (*Law of Large Numbers*) $\{y_i, x_i\}$ is i.i.d., and $E[f(y_i, x_i; \theta)] < \infty$ for all $\theta \in \Theta$, which implies $\sum_{i=1}^n f(y_i, x_i; \theta)/n \xrightarrow{p} E[f(y_i, x_i; \theta)]$.

(ii) (*Compactness of* Θ) Θ is in a compact subset of $\mathbb{R}^k$.

(iii) *(Measurability in* $(y_i, x_i')'$*)* $f(y_i, x_i; \theta)$ is measurable in $(y_i, x_i')'$ for all $\theta \in \Theta$.

(iv) (*Lipschitz Continuity*) For all $\theta, \theta' \in \Theta$, there exists $g(y_i, x_i)$, such that $|f(y_i, x_i; \theta) - f(y_i, x_i; \theta')| \leqslant g(y_i, x_i)\|\theta - \theta'\|$, for some norm $\|.\|$, and $E[g(y_i, x_i)] < \infty$.

Formula for the score statistic

$$S \equiv \left(\frac{1}{\sqrt{n}}\sum_{i=1}^n s_i(\hat{\theta}_R)\right)' A_{nR}^{-1} C_{nR}' \left\{\widehat{Avar}\left(C_{nR}A_{nR}^{-1}\sum_{i=1}^n s_i(\hat{\theta}_R)/\sqrt{n}\right)\right\}^{-1} C_{nR}A_{nR}^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n s_i(\hat{\theta}_R)\right)$$