Given Name:____________________ **Family Name:**____________________

Student Number:____________________ **Signature:**____________________

John von Neumann University
MNB Institute

Empirical Panel Data (19^{th} June, 2023)
Instructor: Chaoyi Chen

Final
VERSION CODE: A
Spring, 2023

Duration: 120 minutes
Aids Allowed: One two-sided handwritten 4 by 6 inch index card formula sheet. Only simple calculators without programming and scientific functions allowed as determined by the exam proctors. No other electronic devises permitted.

Answer all questions.

- Multiple choice questions should be chosen by marking ✓. All other questions must include all work on the exam. No credit without work.
- Sign your exam, initial each page.
- With the sole exception of simple, non-programable, non-scientific calculators, use of electronic devices is not permitted. Determination of which calculators are permitted is at the sole discretion of the exam proctors. Please turn off your cell-phone and all other electronic devices, placing these together with any personal items at the front of the room.
- You cannot leave the room during the exam without explicit permission from the proctor.
- The total marks are provided at the end of your exam.
- Hand in your entire exam together at the end of the exam.

Do not hesitate to raise your hand if you have a question.

Good Luck!

Multiple Choice

Please record all answers by marking ✓.

1. [3] Who is your instructor of this course?
   (a) Chaoyi Chen
   (b) Jerry Hausman
   (c) Joe Biden
   (d) Jackie Chan

2. [3] Which of following models is **not** a linear regression model?
   (a) $Y_i = \beta_0 + \beta_1(\frac{1}{X_i}) + \mu_i$
   (b) $Y_i = \beta_0 + \beta_1 ln(X_i) + \mu_i$
   (c) $ln(Y_i) = \beta_0 + \beta_1 ln(X_i) + \mu_i$
   (d) $ln(Y_i) = ln(\beta_0) + \beta_1 X_i + \mu_i$

3. [3] Which of following is true for panel data?
   (a) Order of data does not matter
   (b) No time dimension
   (c) Repeat observations for a specific agent over time
   (d) Combine cross-sectional and time series issues

4. [3] Consider the following panel data model:

$$y_{it} = \alpha_i + \beta' x_{it} + \varepsilon_{it},$$

   where α_i is a scalar with $E(\alpha_i) = 0$ for all $i = 1, \ldots, n$, $\beta = (\beta_1, \beta_2, \ldots, \beta_k)^\top$ is a $k \times 1$ vector of parameter, $x_{it} = (x_{it,1}, \ldots, x_{it,k})^\top$ is a $k \times 1$ vector of exogenous variables, and ε_{it} is an error term and assumed to be $i.i.d.$ with, for all $i = 1, \ldots, n$, and all $t = 1, \ldots, T$, $E(\varepsilon_{it}) = 0$, $E(\varepsilon_{it}^2) = \sigma^2$. Now, assuming $E(\alpha_i | x_{i1}, \ldots, x_{iT}) \neq 0$, the OLS estimator of β is
   (a) unbiased and consistent
   (b) unbiased but in consistent
   (c) biased but consistent
   (d) biased and inconsistent

5. [3] Continue with the same information as in Question 4. The fixed effect (or LSDV) estimators of β and α_i are
   (a) β: unbiased and consistent when either n, or T, or both tend to infinity
   α_i: unbiased and consistent when either n, or T, or both tend to infinity

(b) β: unbiased and consistent when either n, or T, or both tend to infinity
α_i: biased and inconsistent

(c) β: unbiased and consistent only when n and T both tend to infinity
α_i: unbiased and consistent when either n, or T, or both tend to infinity

(d) β: unbiased and consistent when either n, or T, or both tend to infinity
α_i: unbiased and consistent only when T or both tend to infinity

6. [3] Continue with the same information as in Question 4 but now assume $E(\alpha_i|x_{i1}, \ldots, x_{iT}) = 0$. Which of the following statement is **not** true?

(a) The OLS estimator is biased and inconsistent

(b) The LSDV estimator is unbiased and consistent

(c) The GLS estimator is unbiased and consistent

(d) The GLS estimator is the best linear unbiased estimator (BLUE)

7. [3] Continue with the same information as in Question 4 but assume we are interested in detecting whether the panel data model is a random effect model or a fixed effect model. Under the null of $E(\alpha_i|x_{i1}, \ldots, x_{iT}) = 0$, which of the following is the **correct** test statistic we use to conduct a Hausman specification test?

(a) $H = \left(\widehat{\beta}^{GLS} - \widehat{\beta}^{LSDV}\right)^\top \left(\mathrm{Var}(\widehat{\beta}^{LSDV}) - \mathrm{Var}(\widehat{\beta}^{GLS})\right)^{-1} \left(\widehat{\beta}^{GLS} - \widehat{\beta}^{LSDV}\right)$

(b) $H = \left(\widehat{\beta}^{GLS} - \widehat{\beta}^{LSDV}\right)^\top \left(\mathrm{Var}(\widehat{\beta}^{GLS}) - \mathrm{Var}(\widehat{\beta}^{LSDV})\right)^{-1} \left(\widehat{\beta}^{GLS} - \widehat{\beta}^{LSDV}\right)$

(c) $H = \left(\widehat{\beta}^{OLS} - \widehat{\beta}^{LSDV}\right)^\top \left(\mathrm{Var}(\widehat{\beta}^{LSDV}) - \mathrm{Var}(\widehat{\beta}^{OLS})\right)^{-1} \left(\widehat{\beta}^{OLS} - \widehat{\beta}^{LSDV}\right)$

(d) $H = \left(\widehat{\beta}^{OLS} - \widehat{\beta}^{LSDV}\right)^\top \left(\mathrm{Var}(\widehat{\beta}^{OLS}) - \mathrm{Var}(\widehat{\beta}^{LSDV})\right)^{-1} \left(\widehat{\beta}^{OLS} - \widehat{\beta}^{LSDV}\right)$

8. [3] For the dynamic panel fixed effect model,

$$y_{it} = \gamma y_{i,t-1} + \alpha_i + \varepsilon_{it},$$

where we assume $|\gamma| < 1$ and the initial condition $y_{i0} = 0$ for all i. The dynamic bias of the LSDV estimator is from

(a) Perfect multicollinearity

(b) Serially correlated error

(c) Spurious regression

(d) Endognenous problem

9. [3] Continue with the same information as in Question 8, we now use the Anderson and Hsiao instrumental variable (IV) approach to estimate the model. The two conditions for a valid IV, Z_{it}, are

(a) $corr(Z_{it}, \Delta y_{it-1}) = 0$ and $corr(Z_{it}, \Delta\varepsilon_{it}) \neq 0$

(b) $corr(Z_{it}, \Delta y_{it-1}) = 0$ and $corr(Z_{it}, \Delta\varepsilon_{it}) = 0$

(c) $corr(Z_{it}, \Delta y_{it-1}) \neq 0$ and $corr(Z_{it}, \Delta\varepsilon_{it}) = 0$

(d) $corr(Z_{it}, \Delta y_{it-1}) \neq 0$ and $corr(Z_{it}, \Delta\varepsilon_{it}) \neq 0$

10. [3] Continue with the same information as in Question 8 and now assume $\gamma \longrightarrow 1$. What kind of problem you may have if you propose to use the Arellano and Bond (AB) estimator?

    (a) The AB estimator is inconsistent because the presence of the spurious regression

    (b) The AB estimator is inconsistent because $\Delta\varepsilon_{it}$ is unorthogonal to y_{it-j} for $j \geq 1$

    (c) The AB estimator is still consistent but the relevance condition becomes weak

    (d) The AB estimator is still consistent but the exogeneity condition becomes weak

Long Answer

Please include all your work on the exam. No credit will be given for answers without work, even when the final answer is correct.

1. Long Answer I.

   Consider the model with a single regressor x_{it}:

$$y_{it} = \beta_1 x_{it} + \alpha_i + u_{it}$$

   where α_i represents an unobserved effect fixed over time and u_{it} is a homoskedastic error term which is independent over time (t) and individuals (i). There are N randomly sampled individuals, each observed for T = 4 time periods. Assume that $E(u_{it}|X)$ for all i and $E(u_{it}u_{is}|X) = 0$ for any t and s such that $t \neq s$, where X represent the stacked $NT \times 1$ data matrix.

   (a) [10] List the OLS objective function and derive the OLS estimator of β_0. Show OLS estimator is unbiased when $E(\alpha_i|x_{it}) = 0$ but biased when $E(\alpha_i|x_{it}) \neq 0$.

(b) [10] Derive the within-group estimator of β_0. Show the within-group estimator is unbiased when $E(\alpha_i|x_{it}) = 0$ and when $E(\alpha_i|x_{it}) \neq 0$

(c) [5] Derive the random effects estimator of β_0. Show the random effects estimator is unbiased when $E(\alpha_i|x_{it}) = 0$ but biased when $E(\alpha_i|x_{it}) \neq 0$.

(d) [5] Explain how you could test the assumption that $E(\alpha_i|x_{it}) = 0$

2. Long Answer II.

Let y_{it} denote the unemployment rate for city i at time t. You are interested in studying the effects of a state funded job training program on city unemployment rates. Let z_i denote a vector of*observed* time-constant city-specific variables that may influence the unemployment rate (these could include things like geographic location). Let x_{it} be a vector of time-varying factors that can affect the unemployment rate. The variable $prog_{it}$ is the dummy indicator for program participation: $prog_{it} = 1$ if city i participated at time t. Any sequence of program participation is possible, so that a city may participate in one year but not the next.

(a) [10] Consider the following pooled panel model

$$y_{it} = \beta_0 + z_i\gamma + x_{it}\beta_1 + \rho y_{it-1} + \delta prog_{it} + \varepsilon_{it}.$$

Discuss the merits of including $y_{i,t-1}$ and state an assumption that allows you to consistently estimate the parameters by pooled OLS.

(b) [5] Evaluate the following statement: "The model above is of limited value because the pooled OLS estimators are inconsistent if the $\{\varepsilon_{it}\}$ are serially correlated."

(c) [5] Suppose that it is more realistic to assume that program participation depends on time-constant, unobserved city heterogeneity, but not directly on past unemployment. Write down a model that allows you to estimate the effectiveness of the program in this case. Explain how to estimate the parameters, describing any minimal assumptions you need.

(d) [5] Suppose $prog_{it}$ can depend on unobserved city heterogeneity as well as the past unemployment history. Explain how to consistently estimate the effect of the program, again stating minimal assumptions.

3. Long Answer III.
Consider the following panel data model:

$$y_{it} = \alpha_i + \beta^\top x_{it} + \varepsilon_{it},$$

where α_i is a scalar with $E(\alpha_i) = 0$ for all $i = 1, \ldots, n$, $\beta = (\beta_1, \beta_2, \ldots, \beta_k)^\top$ is a $k \times 1$ vector of parameter, $x_{it} = (x_{it,1}, \ldots, x_{it,k})^\top$ is a $k \times 1$ vector of exogenous variables, and ε_{it} is an error term, $E(\varepsilon_{it}) = 0,\ E(\varepsilon_{it}^2) = \sigma^2$. We assume large n and large T setup.

(a) [10] Derive the first-differenced model. Assuming $E(\Delta y_{it} | \Delta x_{it}) = \beta^\top \Delta x_{it}$, explain how this assumption can be used to construct moment conditions, $E[g(\Delta y_{it}, \Delta x_{it}, \beta)] = 0$, for estimating β. Also write down the corresponding sample moment conditions and drive the method of moment estimator.

(b) [5] Now we assume $E(\Delta x_{it} \Delta \varepsilon_{it}) \neq 0$. Suppose we have a number of valid instruments z_{it} and the dimension of z_{it}, m, is larger than the number of parameters k. Explain how this assumption can be used to construct a new moment condition, $f(\beta) = E[f(\Delta y_{it}, \Delta x_{it}, z_{it}, \beta)] = 0$. Write down the corresponding sample moment conditions $f_{nT}(\beta)$. Explain the intuition for the GMM estimator by referring to the quadratic form:

$$Q = f_{nT}(\beta)^\top W_m f_{nT}(\beta).$$

Explain the role of the weight matrix W_m, and how should it be optimally chosen?

End of examination
Total pages: 9
Total marks: 100