Introductory Econometrics I

Final Review

Yingjie Feng

School of Economics and Management

Tsinghua University

June 1, 2024

Final Exam

- June 20 (Thursday), 14:30-16:30
- Location
  - 6A016: student ID 2017080182 – 2021013291
  - 6A017: student ID 2021013292 – 2022011397
  - 6A018: student ID 2022011398 – 2022011456
  - 6A201: student ID 2022011457 – 2022011510
  - 6A203: student ID 2022011511 – 2022012732
  - 6A205: student ID 2022012733 – 2024403080
- You can bring an A4 two-sided cheatsheet. No calculator is needed
- Regular office hour: Thursdays 4-5pm
- Additional office hour before the exam: June 19 (Wednesday), 5-6pm

Final Exam

- Usually 3 questions. Each contains several parts.
  - The 1st question may require some mathematical derivations
  - Other questions are usually based on specific empirical examples
- Focus on contents covered after midterm
  - Chapters 6, 7, 8, 9, 12, 15 and discussion about causal inference
  - All topics not covered in class will NOT be tested
  - All contents marked with a star will NOT be tested
  - Some basic knowledge from the 1st half is necessary
  - Stata commands are NOT tested, but interpreting Stata outputs might be needed
  - Take a look at the past finals, but remember the coverage of the course slightly differs in each semester

Review: Before the Midterm

- Basics of regression analysis
  - From Simple Regression to Multiple Regression
  - **Identification**: what do we want to get?
    - Zero conditional mean assumption
  - **Estimation**: how do we get it?
    - OLS regression/method of moment estimation
    - Algebraic properties of OLS
    - Statistical properties: unbiasedness, consistency
  - **Inference**: how do we characterize an estimator's uncertainty?
    - Standard errors
    - Asymptotic normality
    - Confidence intervals, t test, F test, etc.

Review: After the Mideterm

- OLS further issues (Ch6):
  - Impact of units of measurements (special case: beta coefficient)
  - Functional form (logs, quadratics, interaction) that characterize more flexible relationships between y and x
  - **Important**: Interpret results precisely (sign, magnitude, effect for different groups, etc.); characterize heterogeneous partial effects
  - What factors do you want to control for in a regression? (ceteris paribus interpretation)
  - Adjusted R-Sqaured will NOT be tested

Review: After the Mideterm

- Dummy variable approach (Ch7, important!)
  - One single dummy, dummies for multiple categories
  - Interaction with other dummies or with continuous variables
  - Dummy variable trap (perfect multicolinearity)
  - **Important**: Interpret results precisely; use dummies appropriately to check heterogeneity across groups!
  - Linear probability model
    - ⋆ Interpretation, heteroskedasiticity (and other shortcomings)

Review: After the Mideterm

- Causal inference and program evaluation (Ch2.7, 3.7e, 7.6)
  - Goal: define causality in a formal framework; understand when regression has a causal meaning and when it does not
  - Understand the potential outcomes $y(1)$ and $y(0)$, observed outcome y, treatment d in context
  - Randomized experiments: key assumption $(y_i(1), y_i(0)) \perp\!\!\!\perp d_i$; comparison of means of two groups immediately identifies ATE
  - Explain if you think the randomized experiment assumption holds or fails in context; provide solutions if it fails (e.g., control for some observed features?)
  - The technical justification of multiple regression $y = \beta_0 + \beta_1 d + \beta_2 x + e$ in the potential outcomes framework will NOT be tested
  - Question 1 (the technical one) will not be about potential outcomes

Review: After the Mideterm

- Heteroskedasticity, intra-cluster correlation, serial correlation (Ch8&12)
  - Technically, we view them as inference issues!
  - They do not directly affect unbiasedness/consistency of OLS
  - They affect how we compute standard errors, construct CIs, or do testing
  - Two general strategies:
    1. Still use OLS estimators, but use various kinds of robust standard errors
    2. Use generalized least squares and hopefully improve efficiency (we only discussed WLS for heteroskedasticity)
  - What kind of robust standard errors are appropriate for a certain research question? At what level do you want to cluster your standard errors?
  - Testing for existence of these issues is NOT required
  - For time series, you only need to know what HAC standard error is and when you want to use it

Review: After the Mideterm

- Endogeneity and solutions (Ch9&15, important!)
  - Different sources of endogeneity
    - ⋆ Omitted variables
    - ⋆ Measurement errors (do not just memorize conclusions; understand the underlying assumptions)
    - ⋆ Sample selection and simultaneity are NOT required
  - Endogeneity makes OLS biased/inconsistent
  - Don't forget one solution is multiple regression
  - Another solution: instrumental variables
    - ⋆ Two stage least squares
    - ⋆ When is an IV valid/invalid? (theoretically and empirically)
    - ⋆ Interpretation of IV estimation, comparison with OLS
    - ⋆ Test for relevance/weak IV, test for endogeneity
    - ⋆ Test for overidentifying restrictions (you need to know the basic idea, but I will not ask you to write down testing procedures)

Takeaways From This Course

- Think as a "researcher"
  - What do you want?
  - What method do you want to use?
  - Differentiate correlation and causation
- Good research
  - Be very clear about what you want (e.g., causality or statistical association?)
  - Choose an **appropriate** (maybe not the newest) approach
  - Base your analysis on theory, reasoning, background knowledge (good "story")
  - Do not "search" for results!!