PhD Econometrics Examination
January 2025

Legibility requirement: Exam-takers are responsible to write the answers clearly, and marked with appropriate order corresponding to the questions. For answers that are our of order and extremely difficult to read, graders may deduct half of the points off.

Part A
Answer any TWO of the following three questions

Q1. Given following exponential distribution:

$$f(y) = \lambda e^{-\lambda y} \quad ---------- \quad (1) \qquad where \ y > 0$$

The subscript *t* is suppressed for simplicity, t = 1, 2, 3, ….n (observation)

a. Derive the log-likelihood function.
b. Derive the MLE estimator of λ.
c. Derive the method of moment estimator of λ.
d. Derive the expected value of y, E(y).
e. For $\lambda = .06$, calculate the cumulative probability, F(8).
f. Is the pdf given above in (1) a valid pdf?

Q2 Given $y_t = \mu + u_t$, derive the following,

a. OLS, MLE, and MM estimates of μ.
b. Derive the Variance of $\widehat{\mu_{OLS}}$.
c. Derive the Variance of $\widehat{\mu_{MLE}}$ using the information matrix method.
d. Show that $\widehat{\mu_{OLS}}$ is unbiased and efficient.
e. Can you show that $\widehat{\mu_{OLS}}$ is consistent too?

Q3, Given $f(y)=3y^2$, $0 \le y \le 1$

a. Is this a valid PDF?
b. Find the mean value of y.
c. Find the median value of y.

Part B
Answer any TWO of the following three questions

Q4. Suppose you want to estimate the effect of childbearing (motherhood status) on labor force earnings for a sample of women in the U.S. using the following model

(B.1) $$Y_i = \alpha + \beta D_i + \varepsilon_i ,$$

where Y_i is the labor market earnings of woman *i,* and D_i is a dummy variable equal to one if woman *i* has had at least one child. In this way you are hoping to estimate the average treatment effect of being a mother on female labor market earnings.

a. Define and explain the five assumptions required to interpret an Ordinary Least Squares (OLS) estimate of a slope parameter as "BLUE."

b. Which of the Ordinary Least Squares assumption is likely to fail when estimating this model? Explain why? What does the mean for your estimate of the average effect of motherhood on earnings?

c. What is the "Fundamental Problem of Causal Inference"? Please define it and explain what it means for empirical analysis.

d. Define Y_{1i} as the earnings of woman *i* if she is a mother and Y_{0i} as the earnings of that same woman *i* if she is not a mother. If motherhood status was randomly assigned across women in the population then the treatment effect of motherhood on earnings would be equal to the following

   i. (B.2) $$E_i[Y_{1i} - Y_{0i}] = E[Y_i \mid D_i = 1] - E[Y_i \mid D_i = 0].$$

   ii. However, motherhood status is not randomly assigned. Using the Potential Outcomes framework, decompose the expectation in (B.2) into the *"average treatment effect on the treated"* and *"selection bias".* Also, say in words what is captured by the average treatment effect and the selection bias terms that you derive.

e. You decide to instrument for motherhood using the instrumental variable Z. Which two assumptions are necessary for this instrument to be valid? Please define them both mathematically and in words.

f. Describe how you would estimate β using two stage least squares (2SLS) and using Z as your instrumental variable. Be sure to specify the equations you would use.

g. Describe how you would estimate β using the control function approach and Z as your instrumental variable. Be sure to specify the equations you would use.

h. What is the *forbidden regression*?

i. Would the following variables be plausible instruments for motherhood. Explain why or why not and be sure to address each of the requirements for a valid instrument in your explanation.

    i. The quality of each woman's health insurance coverage
    ii. An indicator of whether or not the woman has experienced infertility
    iii. Regional differences in abortion laws (e.g. the oldest gestational age a woman can legally obtain an abortion in her region).
    iv. Availability of contraceptive services in a local area.
    v. Number of siblings the woman has
    vi. The woman's marital status

Q5. A sample of data consists of n observations on two variables, y and x. The true model is

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i \qquad \text{(B.3)}$$

where β_1 and β_2 are parameters to be estimated and ε is a disturbance term that satisfies the usual regression model assumptions.

Suppose you estimate (2) via OLS resulting in the following fitted relationship

$$y_i = b_1 + b_2 x_i + e_i \qquad \text{(B.4)}$$

a. Show that the least squares normal equations imply $\sum_i e_i = 0$ and $\sum_i x_i e_i = 0$

b. Show that the solution for the constant term is $b_1 = \bar{y} - b_2 \bar{x}$, where $\bar{y}$ and $\bar{x}$ are sample means of y and x.

c. Show that the solution for b_2 is $b_2 = \dfrac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$.

Now, let's say that a researcher using data for a sample of 3240 female employees 25 years of age and over to investigate the relationship between employees' hourly wage rates Y_i (measured in *dollars per hour*) and their age X_i (measured in *years*). The population regression equation takes the form of equation (1) $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$. Preliminary analysis of the sample data produces the following sample information:

$$N = 3240$$

$$\sum_{i=1}^{N} (y - \bar{y})^2 = 78434.97 \qquad \sum_{i=1}^{N} (x - \bar{x})^2 = 25526.17$$
$$\sum_{i=1}^{N} (x - \bar{x})(y - \bar{y}) = 3666.426 \qquad \sum_{i=1}^{N} y_i = 34379.16$$
$$\sum_{i=1}^{N} x_i = 96143.00 \qquad \sum_{i=1}^{N} y_i^2 = 443227.1$$
$$\sum_{i=1}^{N} x_i^2 = 287851.00 \qquad \sum_{i=1}^{N} x_i y_i = 1023825.00$$
$$\sum_{i=1}^{N} e_i^2 = 77908.35$$

d. Use the above information to compute OLS estimates of the intercept coefficient β_0 and the slope coefficient β_1.

e. Interpret the slope coefficient estimate you calculated in part a—i.e., explain in words what the numeric value you calculated for $\hat{\beta}_1$ means.

f. Calculate an estimate of s^2, the estimated error variance.

g. Calculate the estimate of $\text{var}(\hat{\beta}_1)$.

h. Compute the value of R^2, the coefficient of determination for the estimated OLS sample regression. Briefly explain what the value that you have calculated for R^2 means.

i. Calculate the sample value of the t-statistic for testing the null hypothesis $H_o : \beta_1 = 0$ against the alternative hypothesis $H_1 : \beta_1 \neq 0$. *(Note: You are not required to obtain or state the inference of this test. Just calculate the test-statistic itself).*

Q6. Consider the following model $Y = X_1\beta_1 + X_2\beta_2 + \varepsilon$, where X_1 is a matrix of k_1 variables and X_2 is a matrix of k_2 variables such that

$$X_1 = \begin{bmatrix} x_{11}^1 & x_{11}^2 & \dots & x_{11}^{k_1} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1n}^1 & x_{1n}^2 & \cdots & x_{1n}^{k_1} \end{bmatrix}, \; X_2 = \begin{bmatrix} x_{21}^1 & x_{21}^2 & \dots & x_{21}^{k_1} \\ \vdots & \vdots & \ddots & \vdots \\ x_{2n}^1 & x_{2n}^2 & \cdots & x_{2n}^{k_1} \end{bmatrix}.$$

Denote b_1 and b_2 as the Ordinary Least Squares (OLS) estimates for β_1 and β_2, respectively.

a. Derive the expression for the OLS estimator b_1 as a function of Y, X_1, X_2, and b_2 using the partitioned regression model.

b. Suppose you only observe X_1 but not X_2. Thus, you regress the OLS model, $Y = X_1\beta_1 + \varepsilon$.
   i. Derive the expression for the OLS estimate b_1 that you would estimate under these conditions (i.e., what is the usual OLS estimator for b_1 when you regress Y on X_1).

   ii. If $Y = X_1\beta_1 + X_2\beta_2 + \varepsilon$ is the true model, give an expression for the bias of b_1 in this circumstance (that you estimated above) as a function of X_1, X_2, and b_2.

c. Now suppose you observe both X_1 and X_2. Derive the OLS estimator for b_2 as a function of Y, X_1, X_2, and M_1 using the partitioned regression model, where M_1 is the residual maker and $M_1 = I - X_1(X_1'X_1)^{-1}X_1'$.

d. Define the Frisch-Waugh-Lovell Theorem and describe its intuition.

e. Under what conditions is the bias you solved for in b.ii. equal to zero. What does this mean in the context of the Frisch-Waugh-Lovell Theorem (i.e., what happens when you regress X_2 on X_1)

Part C
Answer any TWO of the following three questions

Q7.

[to be completed in Stata]

For this question, you will be using your knowledge on machine learning (ML) in Stata to build a predictive model of selected US female workers' hourly wage based on the 1988 National Longitudinal Survey of Young Women (NLSW).

The dataset contains n=2246 observations of a group of women in their 30s and early 40s. Contained in the dataset is demographic and socioeconomic information on the group of women, including their hourly wage (*wage*) in 1988.

__NOTE:__ You may not use any cellphones, notes, .do files from previous classes, textbooks, online resources, AI, online forums or FAQs, or any other resources whatsoever outside of Stata and the Stata help files for assistance on this question. You can only use Stata and your internal knowledge on ML in Stata to answer the questions that follow. You must also work alone. You should never pull up any web browser on your computer for any reason. Any instances of cheating or dishonesty will result in zero points for this question.

To begin, open up Microsoft Word on your computer and create a new blank document. You will be using Word to type all of your answers to the questions that follow. Be sure and save your work as you go! At the end of the exam, you will email your Word file to Olivia so that the committee can grade it.

Next, open up Stata. Create a new .do file. You will be putting all your code in this .do file for this question. Save it often so that nothing gets lost. You will also be submitting your .do file to Olivia at the end of the exam so that the committee can grade it.

To access the data, type "*sysuse nlsw88.dta*". This will automatically pull the dataset into Stata that you will be working with for this question.

Let's begin.

a. Construct a summary statistics table of the dataset. It must include mean, std. dev., min., and max. In words, provide a written description of 3-4 different variables in the dataset that have interesting summary statistics (in your opinion). Explain why the variables you select appear interesting to you in terms of their summary statistics. **In your Word doc, discuss your answers in words and include a screenshot of your summary statistics table. Additionally, show all steps in your .do file**.

b. Create tabulation tables for the variables *industry* and *occupation*, separately. The tables need to include the frequency and percent of the sample in each industry or occupation. What are the top two most common industries and top two most common occupations among women in the sample? **In your Word doc, discuss your answers in words and include screenshots of both tabulation tables. Additionally, show all steps in your .do file**.

c. Using OLS, run a regression using *wage* as the dependent variable, a set of indicator variables for both *industry* and *occupation*, and a minimum of four other relevant right hand side covariates (of your choice). Do not include any interaction or non-linear terms. Play around with different covariates to obtain the highest R-squared that you can obtain. What covariates and indicator variables are statistically significant at the 5% level or better? Discuss and interpret the sign and magnitude-of-effect on each significant covariate. **In your Word doc, discuss your answers in words and include a screenshot of your OLS regression table. Additionally, show all steps in your .do file**.

d. Create a new 0/1 indicator variable for "black" from the variable *race* (and name this new variable *black*). Additionally, create new 0/1 indicator variables for each and every industry and occupation, separately. You should have 12 new industry indicators and 13 new occupation indicators, if you have done this part correctly.

e. Then, use adaptive LASSO (linear version) to build a predictive model of *wage*. Use all possible interactions of every single variable in the dataset, using *black* and the new industry and occupation indicator variables you created above in place of *race*, *industry*, and *occupation*. Use 80% of the sample as the training sample and 20% as the test sample and use K=5 folds. Set the rseed at 10101 for both the split sample and for the LASSO model run. Run the LASSO on the training sample only. After running the adaptive LASSO in Stata, address each of the following questions:
    i. How many covariates are selected for inclusion by adaptive LASSO?
    ii. Jointly predict *wage* for the training and test samples. Name this prediction variable *y_Adapt_LASSO*.
    iii. Calculate the training and test MSE values from *y_Adapt_LASSO* in Stata.
    iv. Re-run the OLS regression from part c) with all the same covariates you included in part c), but only on the training sample. Then, jointly predict *wage* from the training and test OLS regressions and name this prediction *y_OLS*.
    v. Calculate the training and test MSE values from *y_OLS* in Stata.
    vi. Create a table that contains the previously calculated training and test MSE values from *y_Adapt_LASSO* and *y_OLS*. Your table should contain a total of four MSE values.

vii. Based on the MSE table created in the previous step, which model (OLS vs. Adaptive LASSO) would you select as your preferred predictive model of *wage*? Justify your answer in words.

viii. Include answers to parts e1) and e7) in your Word doc. Also, include screenshots of: (i) the table showing which covariates are included in the adaptive LASSO from e1); (ii) the OLS regression table from e4), and; (iii) the MSE table created in part e6).

f. Lastly, in words, describe how you would build a model to obtain the standard errors and p-values on the covariates and interactions selected by the adaptive LASSO model that you estimated in part e). Be as specific as possible. **Include your answer in your Word doc**.

g. Submit <u>both</u> your final Word doc and your final .do file to Olivia (likely via a USB drive, but Olivia will tell you the specifics). The files you submit to Olivia will constitute your final answers to this question of the core exam.

Q8. True/False/Uncertain Questions.

For each of the five statements below, state if it's True, False, or Uncertain as it is written and provide a detailed written and/or mathematical justification for the determination you make. Answers with no or insufficient justification will receive few (if any) points.

a. Economists should trust that correlations found in observational data likely represent causal relationships.

b. Randomized controlled trial (RCT) experiments are considered the "gold standard" for drawing causal inferences because treatment assignment is known and pre-determined and because the treated and control groups should be generally similar on observable and unobservable characteristics.

c. In the synthetic control method (SCM), the synthetic control is a credible counterfactual group only if the pre-treatment differences between the treatment group outcomes and the synthetic control outcomes are exactly zero.

d. The regression discontinuity design (RDD) provides credible causal estimates as long as the continuity assumption holds, meaning that the control group should be expected to experience a discontinuous "jump" in the outcome of interest at the treatment threshold.

e. In machine learning (ML), it is common to minimize the mean squared error (MSE) when determining the predictive model to select, where MSE is defined as $\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$.

Q9. In 2007, New York City (NYC) embarked on a major urban afforestation initiative to plant one million new trees in the city within a decade. Dubbed "MillionTreesNYC", the program was highly successful and the millionth tree was planted well ahead of schedule.

Urban trees provide many ecosystem service benefits, including by capturing carbon and air pollutants, water filtration, shade, reducing the "urban heat island effect", and by promoting outdoor activity and exercise. Trees can also increase property values.

Economists have used the sharp increase in the urban forest canopy in NYC as a natural experiment to study the benefits of tree cover. Typically, NYC is considered the "treatment" group and economists have studied the impact of MillionTreesNYC on various outcomes of interest (e.g., air pollution, recreation, human health, etc.).

In the questions that follow, you will be asked to think more carefully about how you could create a research design using modern causal inference to study the economic benefits of urban tree cover.

a. In words, describe in detail a credible difference-in-differences (DID) research design for identifying the causal effect of MillionTreesNYC on infant health outcomes in NYC. Carefully describe the treated and control groups and what data you would need to obtain in order to estimate the DID you construct.

b. Using math, construct a linear DID model that credibly represents the DID research design you created in part a). Include any relevant control variables and fixed effects that you think are needed in the equation. Describe, in detail, each and every variable included in your equation.

c. In words, discuss how you would "test" the parallel trends assumption necessary for a credible DID model of the effect of MillionTreesNYC on infant health outcomes. Be as specific as possible.

d. In words, discuss whether or not residential sorting behavior (i.e., NYC residents sorting where they live in response to the MillionTreesNYC program) is a concern here that would bias the DID estimate you obtain from part b). If you believe that sorting behavior <u>is a concern</u>, discuss how this might bias your results from part b) and what steps should be taken to minimize the amount of bias it might create. If you believe that sorting behavior <u>is not a concern</u>, discuss why we can credibly ignore this issue, appealing to your research design.

e. Hypothesize (as realistically as possible) <u>2-3 other confounding factors</u> that might co-vary with both the rollout of MillionTreesNYC and infant health outcomes, that could

falsely give the appearance of an effect of MillionTreesNYC on infant health when in fact no causal effect actually exists. How would you modify your research design in parts a) and b) to eliminate or control for these 2-3 confounding factors? Be as specific as possible.

f. In words, discuss 2-3 econometric placebo or falsification tests that you could perform to improve the causal interpretation of your DID results in part b). Describe each test in detail using your knowledge of econometrics and causal inference.