

PhD/MA Econometrics Examination

August 2025

Total Time: 8 hours

MA students are required to answer from A and B.

PhD students are required to answer from A, B, and C.

The answers should be presented in terms of equations, statistical details, and with necessary proofs and statistical deduction. Verbal and brief descriptive discussions will not be sufficient.

PART A

(Answer any TWO from Part A)

Q1. Use the table below. "HWSEI is a constructed variable that assigns a Hauser and Warren Socioeconomic Index (SEI) score to each occupation using the modified version of the 1990 occupational classification scheme available in the OCC1990 variable. The HWSEI variable is a measure of occupational status based upon the earnings and educational attainment associated with each category in the 1990 occupational scheme."

Source	SS	Df	MS	Number of obs	=	12104
				F(???, ???)	=	498.53
Model	1662649.59	13	127896.1	Prob > F	=	0
Residual	3101628.43	12090	256.5449	R-squared	=	???
				Adj R-squared	=	0.3483
Total	4764278.02	12103	393.6444	Root MSE	=	16.017

HWSEI	Coef.	SE.	t	P>t	[95% Conf.	Interval]
Age	1.4378	???	20.55	???	???	???
Age^2	-0.0174	0.0009	-20.44	0.0000	-0.0190	-0.0157
Female	-2.2691	???	???	???	-2.8422	-1.6959
Nursery school to grade 4	-1.4976	2.4374	-0.61	0.5390	-6.2754	3.2801
Grade 5, 6, 7, or 8	-2.3946	1.7404	-1.38	0.1690	-5.8061	1.0169
Grade 9	-1.9035	1.7752	-1.07	0.2840	-5.3831	1.5761
Grade 10	-1.0185	1.7148	-0.59	0.5530	-4.3798	2.3429
Grade 11	0.8919	1.6985	0.53	0.6000	-2.4373	4.2211
Grade 12	8.0669	1.5618	5.17	0.0000	5.0056	11.1283
1 year of college	12.1170	1.5784	7.68	0.0000	9.0230	15.2109
2 years of college	17.0264	1.6285	10.46	0.0000	13.8342	20.2186
4 years of college	26.0401	1.5930	16.35	0.0000	22.9176	29.1626
5+ years of college	35.9438	1.6119	22.3	0.0000	32.7843	39.1033
Constant	-8.8311	2.0393	-4.33	0.0000	-12.8284	-4.8337

Omitted groups: "male" and "no school or N/A"

- a) Interpret the coefficient (and other relevant information) on “Grade 11.” Grade 11 is a dummy variable for people who have completed through 11th grade, i.e. one year short of finishing high school.
- b) Now, looking at the coefficient on the variable Grade 12, is there an important lesson that we can learn?
- c) Calculate R^2 . Explain your answer. What does the number mean?
- d) Find the degrees of freedom for the F.
- e) What is the F value? What does it tell you? Can you calculate it from the information given above? If you can calculate F, interpret the value.
- f) Calculate the missing values in the row labeled “Age.”
- g) Calculate the missing values in the row labeled “Female.”
- h) What is the “Dummy Variable Trap?”

*** You should have filled in 10 blanks (???) in the table. ****

Q2. Fundamentals of OLS

- a. Write out the OLS equation in matrix form. Also, write out the matrices and state their dimensions.
- b. State the OLS assumptions in mathematical statements and in sentences (words).
- c. Show that the OLS estimator is BLUE and define BLUE. Show all parts: B, L, U, and E.
- d. What are the properties (hint: there are six) of the OLS estimator? State them in mathematics and words. Also, state any requirements which are necessary for these properties to hold.
- e. Given the properties in part d, what can you infer about the disturbances from the residuals?
- f. Write out a simple OLS model. Define your variables and describe how your model might meet or not meet all the assumptions you stated above.

Q3. The Variance of Least Squares

We know $\text{var}(b) = \sigma^2(X'X)^{-1}$, but σ^2 is an unknown parameter. Therefore in order to find $\widehat{\text{var}}(b)$, we need to find a good estimator for σ^2 .

1. Derive that estimator.
2. Write down the standard error of b .
 - a. What is the standard of error of b used for? Or *why* did we derive it?

Part B: Answer any two of the following three questions

[Short verbal descriptive answer without mathematical proofs, steps, and necessary derivation will not earn you full credit.]

Q4: OLS Statistics, Hypothesis Testing, and Interpretation

You are examining the results from an experiment intending to improve the productivity of sales for people in a firm. There are two programs being tested, sales training (*training*) and a higher sales commission rate (*commission*).

For sales training, workers are randomly assigned into different number of hours each week (*training* is a variable for number of hours). For the new higher commission rate, workers are randomly placed into two groups, those who receive a higher commission and those who continue to receive the old commission rate (*commission* = 1 for people receiving the higher rate and zero otherwise). The variable *sales* reflect annual sales revenue of this individual and *experience* reflects years worked at the firm.

There are 204 sales peoples participating in this experiment. You find a sum of squared residuals of 400.

Table 1: OLS Regression Results

	Coefficients	
	Model 1	Model 3
Intercept	44.929149 (0.554151)	44.929149 (0.554151)
Training	0.005000 (0.00100)	0.010000 (0.001000)
Commission	0.020000 (0.00300)	0.010000 (0.003000)
Experience	-0.001000 (0.002000)	0.008000 (0.002000)

Experience X Commission		0.002000
		(0.000020)
Experience X Training		-0.002000
		(0.000020)
<hr/>		
Degrees of Freedom	200	200
R-squared	0.1800	0.1900
Adjusted R-Squared	0.1795	0.1805
<hr/>		
Standard Errors in Parentheses		

Model 1:

$$\log(\text{sales}) = \beta_0 + \beta_1 \text{training} + \beta_2 \text{commission} + \beta_3 \text{experience} + \varepsilon$$

You run the OLS regression and get the following results reported in the first column of **Table 1**.

- What does the R-squared tell you about the regression?
- Calculate the 95% confidence interval on $\hat{\beta}_1$. The critical value is 1.96
- Interpret the coefficient estimate on training (*take for granted that the unbiasedness assumptions hold*).
- Interpret the coefficient estimate $\hat{\beta}_2$ from the model estimated (*take for granted that the unbiasedness assumptions hold*).
- Suppose you wanted to test the hypothesis that neither intervention had any effect on sales, clearly state the null hypothesis you would test.
- Now you estimate the following model (i.e., the same as Model 1 except that the two intervention variables are excluded from the regression)

Model 2:

$$\log(\text{sales}) = \beta_0 + \beta_1 \text{experience} + \varepsilon$$

and find a sum of squared residuals of 420. Test the hypothesis that neither intervention affects sales (the critical value at the 1% significance level for this problem is 4.506). Clearly walk through the steps to test this hypothesis, and state the conclusion of the test (*Hint: write the F-stat formula in terms of sum of squared error rather than R-squared*).

- g. Suppose the F-statistic you calculated in part (f) has a p-value of 0.015. What does this p-value tell you about the null hypothesis?

Now, a co-worker suggests that the following estimated model better explains sales.

Model 3

$$\log(\text{sales}) = \beta_0 + \beta_1 \text{training} + \beta_2 \text{commission} + \beta_3 \text{experience} + \beta_4 (\text{experience} \times \text{commission}) + \beta_5 (\text{experience} \times \text{training}) + \varepsilon$$

You run the OLS regression of Model 3 and get the results reported in Column 2 of **Table 1**.

- h. Derive and interpret the marginal effect of an additional hour of training. (Take for granted the OLS assumptions hold, and that our coefficients' estimates are significant).
- i. Explain how experience and job training interact in this sales department.
- j. Derive and interpret the marginal effect of how receiving the higher commission rate affects sales.
- k. Do more experienced or less experienced sales reps respond better to the new commission incentive? (Which coefficient and what about it answers this question).
- l. Suppose a coworker points out that training likely helps sales a lot at low levels, but the returns to training decrease (and perhaps even turn negative) as reps receive more and more of it. What variable would you add to Model 1 to estimate this effect and test this hypothesis?

Q5. In econometric analysis, we are often concerned with estimating the causal effect of some treatment variable, D_i , on an outcome variable of interest, Y_i . However, obtaining a causal estimate can be challenging.

- a. What is the “Fundamental Problem of Causal Inference”. Please define it and explain what it means for empirical analysis.
- b. Define Y_{1i} as the outcome of individual i if she/he is treated and Y_{0i} and the outcome of that same individual if she/he were not treated. If treatment were randomly assigned across individuals in the sample, then the treatment effect of D_i on Y_i is as follows:

$$E_i[Y_{1i} - Y_{0i}] = E[Y_i|D_i = 1] - E[Y_i|D_i = 0], \quad (\text{B.3})$$

where D_i is equal to one if individual i was treated and equal to zero if she/he was not treated (i.e., was in the control group).

Suppose that treatment *was not* randomly assigned. Using the Potential Outcomes Framework, decompose the expectation in equation (B.3) into the “average treatment effect” and “selection bias”. Say in words what is captured by the average treatment effect term and the selection bias terms that you derive.

- c. Suppose you are looking to estimate the returns to education using the model

$$W_i = \alpha + \beta E_i + \gamma X_i + \varepsilon_i \quad (\text{B.4})$$

Where W_i and E_i indicate individual i 's monthly earnings and educational attainment, respectively. X_i is a vector of control variables for individual i .

- d. Which OLS assumption is likely to fail when estimating this model? Why? What does this mean for your estimate on the returns to education in earnings?
- e. You decide to for instrument educational attainment using the instrumental variable, Z . Which two assumptions are necessary for an instrument to be valid? Define these assumptions in words and math.
- f. Derive the 2SLS-IV estimator for β .
- g. Would the following variables be plausible instruments for educational attainment? Please explain why or why not. Please be specific in relating the instrument to the requirements for a valid instrument.
 - i. The educational attainment of individual i 's parents.

- ii. An index variable indicating school quality in an individual's community of residence.
- iii. Quarter of individual i 's birth (i.e. born between January-March, April-June, July-September, or October-December) given that there is usually a cut-off for school entry in which children turning 5 in the last quarter must begin school the following year.

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

$$X_1 = \begin{bmatrix} x_{11}^1 & x_{11}^2 & \dots & x_{11}^{k_1} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1n}^1 & x_{1n}^2 & \dots & x_{1n}^{k_1} \end{bmatrix}, X_2 = \begin{bmatrix} x_{21}^1 & x_{21}^2 & \dots & x_{21}^{k_2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{2n}^1 & x_{2n}^2 & \dots & x_{2n}^{k_2} \end{bmatrix}.$$

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

[Short verbal descriptive answer without mathematical proofs, steps, and necessary derivation will not earn you full credit.]

Question #7

[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:
- e1) How many covariates are selected for inclusion by adaptive LASSO?
 - e2) Jointly predict *wage* for the training and test samples. Name this prediction variable *y_Adapt_LASSO*.
 - e3) Calculate the training and test MSE values from *y_Adapt_LASSO* in Stata.
 - e4) 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*.
 - e5) Calculate the training and test MSE values from *y_OLS* in Stata.
 - e6) 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.
 - e7) 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.
 - e8) 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 both 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.

Question #8

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) In machine learning (ML), the basic LASSO model is used to determine optimal covariate selection for regression model building purposes.
- b) In difference-in-differences (DID), the parallel trends assumption is an important, but not required assumption for credible estimation of some treatment effect of interest.
- c) In the synthetic control method (SCM), the synthetic control is a credible counterfactual group if the pre-treatment differences between the treatment group outcomes and the synthetic control outcomes are small.
- d) Correlations between two variables in observational data likely represent causal relationships in most cases.
- e) For instrumental variables (IV) to work, the instrument you select need only satisfy the exclusion restriction.

Question #9

In 2002, the invasive emerald ash borer (EAB) was first detected in the US. EAB is a destructive beetle that lays its eggs in North American ash trees, damaging and ultimately killing them. It is considered the most dangerous invasive species to ash tree populations in the US and has led to the loss of tens of millions of ash trees across the Midwest and Eastern US.

Ash trees were once common in many communities through the eastern half of the US and 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 exploited the introduction of EAB in a US county as a natural experiment to study the costs of deforestation in urban areas. Typically, EAB infested counties are considered the “treatment” group. Economists have studied the impact of EAB introduction on various outcomes of interest (e.g., air pollution, temperature, human health, crime rates, 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 impacts of urban deforestation caused by the EAB.

- a) In words, describe in detail a credible difference-in-differences (DID) research design for identifying the causal effect of EAB introduction on human morbidity in treated counties. 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 EAB introduction on human morbidity. Be as specific as possible.
- d) In words, discuss whether or not residential sorting behavior (i.e., people moving based on EAB spread) is a concern here that would bias the DID estimate you obtain from part b). If you believe that sorting behavior is a concern, 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 is not a concern, discuss why we can credibly ignore this issue, appealing to your research design.
- e) Hypothesize (as realistically as possible) 2-3 other confounding factors that might co-vary with both the introduction of EAB in a county and human morbidity, that could falsely give the appearance of an effect of EAB introduction on morbidity 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.