Applied Economics Field Exam

June 2019

Instructions

This is a closed book examination. No written materials are allowed. You can use a calculator.

INSTRUCTIONS FOR STUDENTS TAKING BOTH THE LABOR AND POPULATION FIELDS. You have 4 hours to complete the exam. The exam is composed of four questions. Each question is worth 100 points. You must obtain at least 75 points in at least three of the four questions to pass the labor and population exam.

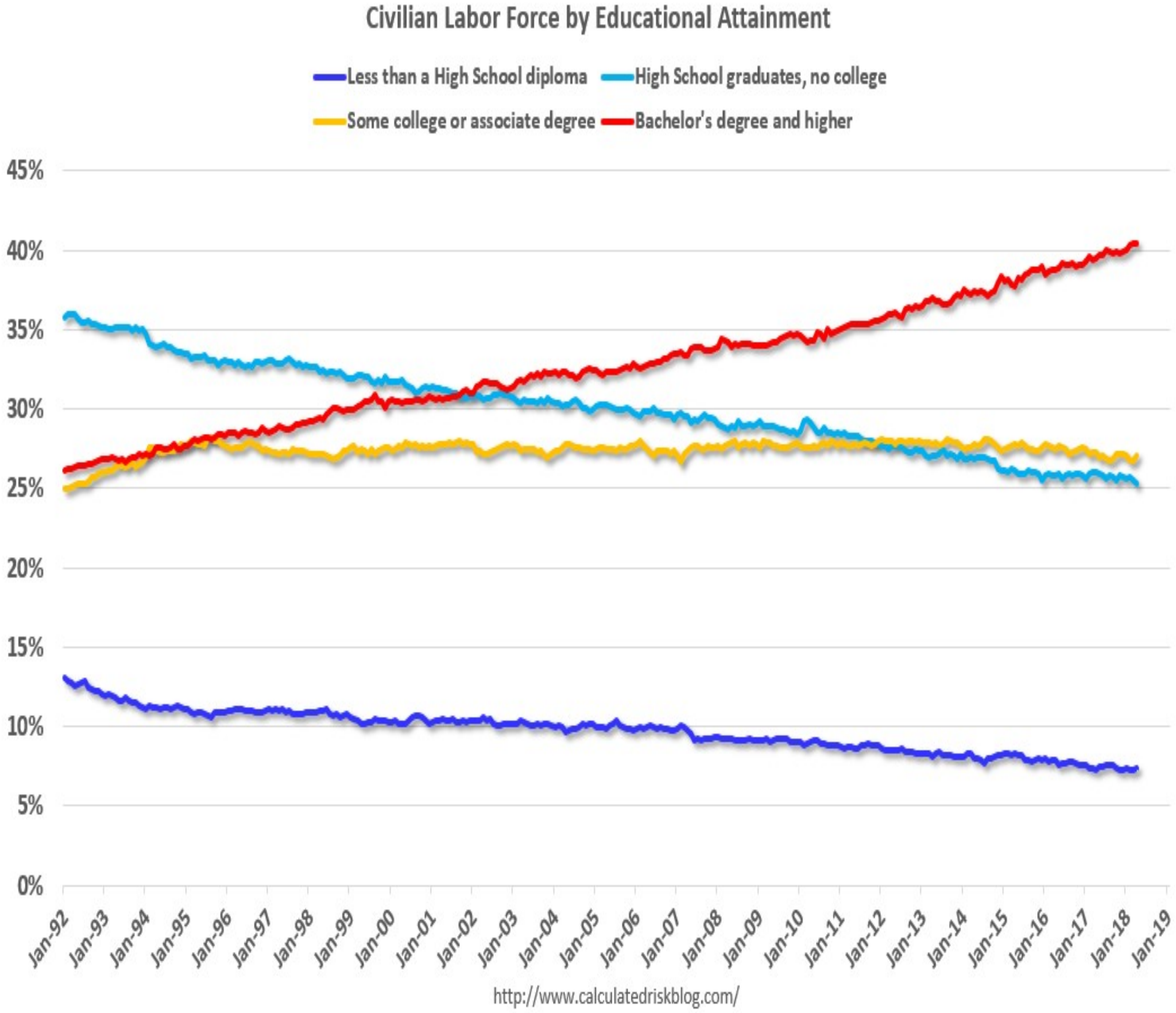
INSTRUCTIONS FOR STUDENTS TAKING ONLY ONE APPLIED FIELD. You have 3 hours to complete the exam. The exam is composed of four questions. Each question is worth 100 points. You must obtain at least 75 points in at least two of the four questions to pass the exam in your field.

Please answer each question in separate booklets.

First Question. 100 Points

Conside the figure on the next page. It documents the evolution of educational attainment in the U.S. labor force in the past 25 years. The question asks you to think about its determinants.

1. (15 points) Describe the main factors that determine the educational decisions of individuals and, hence, have contributed to an increase in the fraction of workers with a college degree and a decline in the fraction with a high school diploma or less.
2. (15 points) Rank the factors in order of importance for this period and explain why you believe this is the correct ranking.
3. (15 points) Consider only the first factor in your ranking. Write down a model of educational decisions that accounts for this factor.
4. (15 points) Using your model, show the effect of your factor on educational attainment.
5. (10 points) Using your model, derive a testable implication that enables you to determine whether your model is consistent with the data.
6. (10 points) Discuss the data you need to test the implication you just derived.
7. (5 points) Now consider the second factor in your ranking. Extend the model you developed in part 3 to account for this additional factor.
8. (5 points) Is the testable implication you derived in part 5 still valid? Explain.
9. (5 points) Derive an additional testable implication that enables you to determine whether the second factor is important to explain the observed patterns.
10. (5 points) What data do you need to test the additional implication?


Civilian Labor Force by Educational Attainment
Less than a High School diploma
High School graduates, no college
Some college or associate degree
Bachelor's degree and higher
45%
40%
35%
30%
25%
20%
15%
10%
5%
0%
Jan-92
Jan-93
Jan-94
Jan-95
Jan-96
Jan-97
Jan-98
Jan-99
Jan-00
Jan-01
Jan-02
Jan-03
Jan-04
Jan-05
Jan-06
Jan-07
Jan-08
Jan-09
Jan-10
Jan-11
Jan-12
Jan-13
Jan-14
Jan-15
Jan-16
Jan-17
Jan-18
Jan-19
http://www.calculatedriskblog.com/

Second Question. 100 Points

1. (25 pts) Answer the following questions in a few paragraphs. Be sure to cite relevant papers where appropriate.

   (a) Economists are often interested in estimating the value of amenities that are not explicitly traded in a marketplace, such as the value of neighborhood safety or of public education. Briefly describe how economists have attempted to do this. (5 pts)

   (b) Set up and describe Rosen's 1974 hedonic model (you can focus just on the consumer side.) What do we learn from the equilibrium conditions of the model? What is Rosen's "two-step" method and what are some of its identification challenges? (5 pts)

   (c) Describe Roback's 1982 model of spatial equilibrium. What is the idea behind her spatial equilibrium condition? When constructing quality of life rankings, Los Angeles often comes out near the top. Use Roback's model to explain why you think that is. (5 pts)

   (d) What is missing from Roback's model of spatial equilibrium? Discuss some recent structural models of spatial equilibrium and relate them to Roback's model. (5 pts)

   (e) Write down a simple user-cost model of house prices. What does the equilibrium relationship between house prices and rents depend on in your model? What is missing from your model? (5 pts)

2. (25 pts) Suppose you have a cross-section of data on housing transactions. The data allows you to see the street address of the house, the sale price, and the characteristics of the home, including neighborhood characteristics. You also have data on public schools, including each school's test score performance and attendance boundaries. You are interested in using this data to estimate the willingness to pay for school quality.

   (a) Suppose you regressed sale price on house characteristics and neighborhood characteristics, including the test scores for the school that the house is assigned to.

What would you expect to find? Describe some problems with this regression. (5 pts)

(b) Come up with an alternative strategy for estimating the average willingness-to-pay for school quality. Make sure to write down the estimating equation. If you are deriving your research strategy from another paper, cite that paper. How do you expect the new strategy to affect your estimates from part (a)? (10 pts)

(c) Under what assumptions will your answer from part (b) correctly identify an average willingness-to-pay? How might you test these assumptions with the data that you already have? Feel free to draw hypothetical pictures/tables you would use to test these assumptions. (10 pts)

3. (25 pts) Consider a housing market with consumers $i = 1, \ldots, N$ and houses $j = 1, \ldots, N$. Each consumer chooses one house and each house gets chosen by one consumer. The indirect utility that consumer i gets from choosing house j is:

$$V_{ij} = \alpha_i x_j + \beta_i p_j + \xi_j + \epsilon_{ij}$$

where x_j is an observed (to the econometrician) house characteristic, p_j is the observed price of the home, ξ_j is the unobserved house quality, and ϵ_{ij} is a preference shock that is iid type 1 extreme value. α_i and β_i are preference parameters given by:

$$\alpha_i = \alpha_0 + \sum_{k=1}^{K} \alpha_k z_{i,k}$$

$$\beta_i = \beta_0 + \sum_{k=1}^{K} \beta_k z_{i,k}$$

where $z_{i,k}$ are the observed demographic characteristics of consumer i. As the econometrician, you observe x_j, p_j, z_{ik}, and d_{ij}–an indicator for whether consumer i chooses house j.

(a) What is the probability (over ϵ_{ij}) that person i chooses house j? (10 pts)

(b) Let us write $\delta_j = \alpha_0 x_j - \beta_0 p_j + \xi_j$. Describe in words a computationally tractable strategy for estimating δ_j for each house. (5 pts)

(c) Suppose you correctly estimate δ_j for each house. What would happen if you regressed δ_j on x_j and p_j? Would you get unbiased estimates of α_0 and β_0? Why or why not? If not, describe an identification strategy for recovering α_0 and β_0. (10 pts)

4. (25 pts) Consider a city with $j = 1, \ldots, L$ locations, and workers can choose both where to live j and where to work k. The unit price of housing in residential location j is p_j and the wage rate in work location k is w_k. Workers are homogeneous and supply one unit of labor inelastically. A worker i who lives in location j and commutes to work at location k chooses numeraire consumption C and housing consumption H to maximize:

$$U_{ijk}(C, H) = \left(\frac{C}{\theta}\right)^{\theta} \left(\frac{H}{1-\theta}\right)^{1-\theta} \exp\left(-\kappa\tau_{jk} + \sigma\epsilon_{ijk}\right)$$

subject to budget constraint:

$$C + p_j H = w_k$$

τ_{jk} is the commute time between locations j and k and ϵ_{ijk} is a preference shock that is iid across i, j, and k. In each location k, there is a competitive firm that produces the tradeable numeraire good using labor L and capital K. The production technology in location k is:

$$Y = L^{\alpha} K^{1-\alpha} \exp\left(x_k + \xi_k\right)$$

Here, x_k is an observed characteristic of location k, and ξ_k is a productivity shock. Let r be the rental rate of capital which is the same for all firms.

(a) Derive an expression for the log indirect utility that worker i gets from living in location j and working in location k. (10 pts)

(b) Derive an expression for the equilibrium log wage rate in location k. [Hint: In equilibrium, the unit cost of production in each location has to equal 1.] (10 pts)

(c) Discuss how you would go about estimating the paramters of the model, θ, κ, σ, and α. What data would you try to collect? What assumptions on the error terms would you have to make? (5 pts)

Third Question. 100 Points

Let the standard IV model be defined by four variables defined below :

	Variable Description	**Model Equations**
1	Instrumental Variable:	$Z = f_Z(\epsilon_Z)$
2	Unobserved Pre-treatment Counfounder:	$\mathbf{V} = f_V(\epsilon_V)$
3	Treatment Choice:	$T = f_T(Z, \mathbf{V})$
4	Observed Outcome:	$Y = f_Y(T, \mathbf{V}, \epsilon_Y)$
5	Errors are mutually statistically indep.:	$\epsilon_Z \perp\!\!\!\perp \epsilon_V \perp\!\!\!\perp \epsilon_Y$

1. (5 points) Define the counterfactual outcome $Y(t)$ and the counterfactual choice $T(z)$ using the notation above.

2. (5 points) Show that the independence of error terms imply that the instrumental variable Z is statistically independent of the counterfactual choice $T(z)$ and of the counterfactual outcome $Y(t)$.

3. (10 points) Suppose that the instrument is Z is categorical such that $\text{supp}(Z) = \{z_1, ..., z_K\}$. Let the response variable $\mathbf{S}$ be a K-dimensional random vector defined by:

$$\mathbf{S} = [T(z_1), ...T(z_K)].$$

   Let the support of the response variable $\mathbf{S}$ be $\text{supp}(S) = \{\mathbf{s}_1, ..., \mathbf{s}_N\}$. Each vector $\mathbf{s}_n; n = 1, ..., N$ is called a response-type. Use the fact that $Z \perp\!\!\!\perp T(z)$ to show that the following propensity score equation holds:

$$P(T = t|Z = z) = \sum_{n=1}^{N} \mathbf{1}[T = t|Z = z, \mathbf{S} = \mathbf{s}_n] P(\mathbf{S} = \mathbf{s}_n) \qquad (1)$$

4. (5 points) Consider a binary treatment where $\text{supp}(T) \in \{0, 1\}$ and let $K = 3$ such that $\text{supp}(Z) = \{z_1, z_2, z_3\}$. Let $\mathbf{R}$ denote the *response matrix* that stacks all response-types $\mathbf{s}_n; n = 1, .., N$. Without further assumptions, draft the response matrix that comprises all possible response-types.

5. (5 points) Let the monotonicity assumption be defined as:

$$T_\omega(z_{j+1}) \geq T_\omega(z_j); j = 1, 2 \text{ for all agents } \omega \text{ in the sample space } \Omega. \tag{2}$$

Compute the response matrix that arises by assuming monotonicity (2). What is the main property of this response matrix?

6. (20 points) Vytlacil (2002) has shown that the monotonicity assumption is equivalent to a separability condition on the choice equation. In short we have that:

$$T_\omega(z) \geq T_\omega(z') \, \forall \, \omega \in \Omega \Leftrightarrow T = \mathbf{1}[P(Z) > g(\mathbf{V})],$$

where $P(Z)$ denotes the propensity score where $P(z) > P(z')$. Use the propensity score equation (1) and the response matrix $\mathbf{R}$ to sketch a proof of this equivalence.

7. (10 points) Suppose you are interested in evaluating the counterfactual outcome $E(Y(1)|\mathbf{s} = [0, 1, 1]')$. How can you estimate this counterfactual outcome using a standard Two-stage Least Square regression?

8. (5 points) Recently, your colleagues examined the evaluation of a public policy regarding homeless. Their observed data is given by:

   - Z denotes the random assignment of case worker for a homeless person.
   - A denotes the assessment score made by the case worker for the homeless.
   - T denotes the treatment assignment that is in part based on the score A. Let $T = 0$ when no services are offered and $T = 1$ when a housing assistance is offered.
   - Y is an outcome of interest.

   Let $\mathbf{V}$ be the unobserved characteristics of the homeless person and $\mathbf{U}$ be the unobserved characteristics of the case worker. It is sensible to assume that $\mathbf{V}$ causes A, T, Y and $\mathbf{U}$ causes A and T. Express the model as a Directed Acyclic Graph (DAG).

9. (10 points) Suppose you want to estimate the causal effect of T on Y using the matching assumption $Y(t) \perp\!\!\!\perp T|A$. Which causal links must be eliminated for this matching assumption to hold?

10. (5 points) Suppose $Y(t) \perp\!\!\!\perp T|A$ holds and let the propensity score be $P(a) = P(T = 1|A = a)$. Which additional assumption is necessary for the identification of the treatment effect $E(Y(1) - Y(0))$.

11. (5 points) Describe a method that uses the propensity score $P(a) = P(T = 1|A = a)$ to estimate $E(Y(1) - Y(0))$.

12. (5 points) Suppose $Y(t) \perp\!\!\!\perp T|A$ does not hold. Let Z be an index that identifies each case worker and let the propensity score associated with case worker z be $P(z) = P(T = 1|Z = z)$. Suppose the number of case workers is large enough for us to assume that $P(Z)$ is continuous and has full support in $[0, 1]$. Which additional assumption is primary for the identification of the treatment effect $E(Y(1) - Y(0))$.

13. (10 points) Describe a method that uses the propensity score $P(Z)$ of the previous item to estimate the treatment effect $E(Y(1) - Y(0))$.

Labor Field Exam 2019

Graduate Labor Economics – 262P, Prof. Till von Wachter

This is an in class exam for one hour. There is one question with 8 sub-questions. Each sub-question carries the same weight. Please write legibly. Please make an effort to write down a formula where it helps to clarify what you are talking about.

I. Instrumental Variable Estimates of the Returns to Schooling

Consider the following cross-sectional model for individual earnings:

(1) $\log y_i = \alpha_i + \beta_i S_i + \varepsilon_i$

where y_i and S_i are log earnings and years of schooling of individual i, respectively. α_i is an individual constant that may be correlated with schooling, and β_i is the return to schooling, which is allowed to vary across individuals. Let $B = E(\beta_i)$ be the average return to education in the population.

a. Under what conditions will the least squares estimate of the coefficient on schooling in model (1), β_{OLS}, be a consistent estimate of the population average return to education B?

b. What are the sources of omitted variable bias (a.k.a., 'ability' bias) and selectivity bias (a.k.a., self-selection bias) in the OLS estimator? Write down the relevant covariance terms and briefly interpret.

c. For 2 individuals, i and j, with different abilities (a_i, a_j), marginal returns to education (b_i, b_j), and marginal costs of education (r_i, r_j), graphically depict the education selection process and briefly discuss how it relates to the least squares coefficient, β_{OLS} (hint: put (log y) on the y-axis and S on the x-axis and exploit the explicit functions for returns and costs we had assumed in class).

d. Suppose there exists a variable, z_i, that differentially affects the costs of schooling across individuals for exogenous reasons (and that does not have an independent effect on the earnings):

   (2) $S_i = \theta z_i + v_i$

   Under what conditions will two-stage least squares (2SLS) estimation of equation (1) yield a consistent estimate of the average return to education B? Under these conditions, explain why 2SLS eliminates both the ability bias (this is standard) and the self-selection bias.

e. Suppose the assumptions do not hold; what parameter does 2SLS estimate and under what additional condition? Give the parameter an economic interpretation.

For the remainder of the sub-questions, consider now the case in which $\alpha_i=\alpha$ and $\beta_i=\beta$ for all i.

f. Show that IV can have a worse omitted variable bias problem than OLS when the instrument has a weak relationship to schooling.

g. Suppose S_i is measured with error that is "classically" distributed. How will this measurement error bias the least squares estimate of the returns to education, and how is this bias related to the noise-to-total variance ratio corresponding to S_i?

h. Explain how including additional control variables (X_i) into the regression may exacerbate bias from classical measurement error.