Written Preliminary PhD Examination

Department of Applied Economics
June 2017

Policy Analysis

**Instructions**
- Identify yourself by your code letter, not your name, on each page.
- Start each question’s answer at the top of a new page – do not write on the back of the answer sheets.
- Answer any FOUR of the SIX questions.
- You have four hours to complete the exam.
**Question 1:**

1. What is implicit discrimination and how does it differ from the two main economic models of discrimination (i.e. Becker’s taste for discrimination and statistical discrimination)? How does implicit discrimination differ from “spontaneous discrimination?” How would one distinguish empirically implicit discrimination from spontaneous discrimination? (20%)

2. List four major weaknesses or concerns about conventional regression models of discrimination (i.e. dummy variable method and Blinder-Oaxaca decomposition). What is Jeremiah Cotton’s main criticism with Blinder-Oaxaca decomposition and what does he recommend using to correct this issue? (20%)

3. In order to bid on public construction projects, private firms benefit from being “prequalified.” You are given the following logistic regression model relating the probability of prequalification in construction as function of firm characteristics:

\[
\text{Prob}(Y = 1) = \frac{1}{1 + \exp(- (\alpha + \beta_1 \text{DBE} + \beta_2 \text{Instate} + \beta_3 \text{Building} + \beta_4 \text{Heavy}) + \varepsilon)}
\]

Where \(Y\) indicates whether a firm was pre-qualified, \(DBE\) indicates whether the firm is considered to be a disadvantaged business enterprise due to being owned by women or members of racial minority groups, \(Instate\) indicates whether the firm is headquartered inside the state or outside the state, \(Building\) represents whether the firm is in the sub-industry of building construction, and \(Heavy\) indicates whether the firm is in the sub-industry of heavy construction.

Table 1 presents the results of estimating the coefficients on the sample of firms that applied for prequalification in the previous year by Minnesota Department of Transportation (MnDOT). Summarize the results of Table 1. (10%)

The organization of women and minority owned businesses has sued MnDOT claiming that their members are illegally discriminated against and that the plaintiffs are less likely to be prequalified for construction work than are non-minority, non-women owned business enterprises.

Indicate wherein you agree or disagree with the plaintiffs’ claims, based on the findings in Table 1. (10%)

The defendant, MnDOT, has responded by indicating that it does not discriminate and that the reasons that women and minority owned firms are less likely to be prequalified for large construction jobs is that they lack sufficient capacity to undertake these jobs.

The plaintiffs have hired you as their expert witness. Provide a detailed methodology and data analysis plan for investigating the plaintiff’s claims. (40%)
Table 1: Logistic Regression Analysis of Pre-Qualification in Construction

<table>
<thead>
<tr>
<th></th>
<th>Coefficient Estimate</th>
<th>Chi-square</th>
<th>p-value</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-0.6736*</td>
<td>6.669</td>
<td>0.0098</td>
<td>1.0000</td>
</tr>
<tr>
<td>DBE Status</td>
<td>0.6493*</td>
<td>6.280</td>
<td>0.0122</td>
<td>0.1077</td>
</tr>
<tr>
<td>Instate</td>
<td>0.9889*</td>
<td>16.546</td>
<td>0.0001</td>
<td>0.8098</td>
</tr>
<tr>
<td>Building Construction</td>
<td>-0.3717</td>
<td>2.843</td>
<td>0.0918</td>
<td>0.3706</td>
</tr>
<tr>
<td>Heavy Construction</td>
<td>-1.5555*</td>
<td>46.119</td>
<td>0.0001</td>
<td>0.4448</td>
</tr>
<tr>
<td>Mean of Dependent</td>
<td>0.3678</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Max-rescaled R-Square</td>
<td>0.1661</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p-value</td>
<td>0.0001</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of Observations</td>
<td>715</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Coefficient estimates are significant at 95% significance level.

Source: DOT Pre-Qualification of Construction File, 1999
Question 2:

The information below comes from the results of a study published in *Statistics in Medicine* (“Using the Peters-Belson Method to Measure Health Care Disparities from Complex Survey Data”). The authors estimate a model of cancer screen as a function of age, age$^2$, income, education (less than high school graduate, high school graduate, at least some college), region of residence (Northeast, West, Midwest, and South), metropolitan statistical area (in MSA or not), health insurance coverage (yes or no), and usual source of coverage (yes or no).

The Peters-Belson method assumes that the probability of a positive outcome for a person $i$ in group $R_1$ should follow the same logistic model as a person from the reference group $R_0$, so each person $i$ in group $R_1$ has a predicted outcome

$$\hat{p}_{R_0} = \frac{\exp(x_i'\hat{\beta}_{R_0})}{1 + \exp(x_i'\hat{\beta}_{R_0})}$$

Thus, using the logistic model for group $R_0$, the estimated expected proportion of individuals with $y=1$ that are in group $R_1$ is

$$\hat{p}_{R_0,R_1} = \sum_{j=1}^{n} w_j \delta_{R_0,j} \hat{p}_{R_0,j} / \sum_{j=1}^{n} w_j \delta_{R,j}$$

The estimated difference between the crude proportions of individuals in $R_0$ and $R_1$ with $y = 1$, $p_{R_0} - p_{R_1}$, which is referred to as the overall disparity, can be partitioned as

$$p_{R_0} - p_{R_1} = (p_{R_0} - \hat{p}_{R_0,R_1}) + (\hat{p}_{R_0,R_1} - p_{R_1})$$

Use this information along with the table below to answer the following questions.

1) Interpret and explain the differences in the percent explained across cancer screenings. What may contribute to the unexplained gap? (20%)

2) Does this method of computing the explained portion produce the same results as the Blinder-Oaxaca decomposition using STATA for logistic models? (20%)

3) What can be said about the statistical significance of the explained portions displayed in column six? (20%)

4) What modifications, if any, would be required to apply this methodology to understanding gender differences in labor force participation? (20%)
5) You have just discovered from the original data collectors that the Hispanic category can be of any race. Assuming that most Hispanics self-identify as “white” or “other”, how does the interpretation of the results change? (20%)
Question 3:

Answer each part of the question. Your answers should demonstrate an advanced graduate student level of economics knowledge.

1) How would capital gains be taxed under an individual income tax system based on a Haig-Simons (economic) measure of income?

2) From Slemrod-Bakija *Taxing Ourselves*: “[The amount of capital gains income that is taxed under the U.S. income tax] is the net result of massive errors (relative to economic income) in both directions, as much income in the form of capital gains ends up facing little or no tax at all, while a significant amount ends up being taxed more than once.” Discuss the quote above in reference to the U.S. income tax. Explain (a) how some capital gains income is taxed less than what would be prescribed by a Haig-Simons-based income tax and (b) how some capital gains income is over-taxed relative to Haig-Simons.

3) Under its state individual income tax, Minnesota includes net capital gains income in taxable income and subjects it to the same tax rates as apply to other income. Studies have shown that capital gains are the largest contributor of volatility to the state’s revenue system. Consequently, state policy makers sometimes propose exempting capital gains from taxable income in order to make the revenue system less volatile. Using what you know about tax policy, how would you assess such a proposal? What would be the likely impact of excluding capital gains income from the state income tax?

4) A member of the U.S. Congress proposes to exclude all interest income from the individual income tax, saying that this will increase the national saving rate in the United States.

   a. Do you expect the proposal to increase the U.S. saving rate? Explain, supporting your answer with (1) economic theory in the form of a model (graphs or equations) of individual saving behavior under an income tax and (2) any empirical evidence you care to cite.

   b. If you were a member of Congress, how would you vote on this proposal? Why? How would your knowledge of tax policy influence your decision?
Question 4:

Prior to the early nineteenth century, there were no legal restrictions on abortion in the United States. Around the 1820s, individual states started passing laws that penalized individuals who performed abortions, and all states had anti-abortion laws in place by the early 20th century. Demographers and historians have been interested in the potential effect of the introduction of these anti-abortion laws on fertility rates.

Through your affiliation with the Minnesota Population Center, you have recently received access to the 1850 complete count U.S. census. Using the information in this census, you have constructed a panel dataset of 15-year fertility histories (i.e., from 1835-1849) for the approximately 4 million married women aged 20-49 in the census, using the age of her children in 1850, and combined this with state legislative records on the year of passage of anti-abortion legislation. The resulting dataset contains approximately 25 million observations and the following variables:

- \( ID \) = woman-specific ID
- \( state \) = state of residence in 1850 (note: there were 38 states in the U.S. in 1850)
- \( year \) = ranges from 1835-1849
- \( gavebirth \) = indicator for a woman giving birth in a specific year \( t \)
- \( yearpassed \) = year the state the woman was living in in 1850 passed its anti-abortion law (13 states passed anti-abortion laws between 1835 and 1849)
- \( age \) = age of the woman in years

You also have a vector of woman-specific control variables \( X \) that contain things like literacy, husband’s occupation, urban/rural residence, household wealth, etc. You only have information on these characteristics in the year of the census (1850).

1. You decide to use a difference-in-difference strategy to estimate the effect of anti-abortion laws on fertility. Write down the specification you would use to do so. Be specific as possible. If you need to construct new variables using those listed above, describe in detail what they are and how they are constructed.

2. What is the assumption this model relies on in order to identify a causal effect of these laws on married women’s fertility? Be specific.

3. You run the DID model using OLS and your key coefficient is equal to 0.0017 (se=0.0004). Interpret this effect. Do anti-abortion laws increase fertility? Is the effect size “plausible”, in your view? (Note: the mean probability of giving birth in a single year for women in your sample is 0.219.)

4. A colleague suggests you cluster your standard errors. Do you think this is a good idea, and why or why not? If you were to cluster your standard errors, what level of clustering would be most appropriate, and why?
5. You estimate your model from (1) with clustered standard errors (at the appropriate level) and your key coefficient is 0.0017 (se=0.0050). Why didn’t the point estimate change? Why did the clustering affect the standard errors so much? Does this result change your conclusion about the effect of anti-abortion laws on fertility?

6. Another colleague suggests the addition of state-specific linear trends to your model. Write down the specification including these trends. (Hint: use a variable \( t = 1, 2, 3, \ldots, 15 \), and remember these are state-specific trends!) How does this change the key identifying assumption of your model, if at all? Which model (the model with or without these trends) do you prefer, and why?

7. You present your results at a conference, and a member of the audience suggests you also estimate the effect of anti-abortion law passage on birth spacing (i.e., the amount of time between two births to a single mother). Why would anti-abortion laws potentially affect this outcome? (Hint: Historians argue (despite the lack of evidence) that abortion was a widely practiced method of birth control for married women in the mid-19th century.)

8. You decide to start another paper investigating the effect of anti-abortion laws on birth spacing using the same dataset. Describe your empirical strategy in as much detail as possible. Start by defining the dependent variable, and get as far as you can.
Question 5:

Analysts sometimes use nonexperimental data to estimate the effect of government policies or programs. Since the mid 1960’s, the U.S. federal government has funded a preschool program called Head Start that is offered across the country for children from low-income families. Although Head Start is very popular, not every family enrolls their child and some families enroll one child and not the others.

In 1995, Currie and Thomas published a paper in the American Economic Review that investigated the effect of Head Start for participants in the NLSY, a large national survey of adults and children. The main results for white students are reported in their Table 4. The estimates of the association of Head Start participation and the student’s PPVT (a vocabulary test score obtained in early elementary school) were reported three different ways as shown by the columns in the table. The PPVT is measured as a percentile score (0-100) that shows how well each student did relative to the rest of the population.

<table>
<thead>
<tr>
<th>Variable</th>
<th>OLS-unadjusted</th>
<th>OLS-adjusted</th>
<th>Mother fixed effects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Head Start</td>
<td>-5.621</td>
<td>-0.383</td>
<td>5.875</td>
</tr>
<tr>
<td></td>
<td>(1.570)</td>
<td>(1.453)</td>
<td>(1.520)</td>
</tr>
<tr>
<td>Other preschool</td>
<td>9.077</td>
<td>1.679</td>
<td>1.173</td>
</tr>
<tr>
<td></td>
<td>(1.275)</td>
<td>(1.171)</td>
<td>(1.296)</td>
</tr>
<tr>
<td>Constant</td>
<td>31.512</td>
<td>-106.706</td>
<td>--</td>
</tr>
<tr>
<td></td>
<td>(0.783)</td>
<td>(16.306)</td>
<td></td>
</tr>
</tbody>
</table>

Note: N=2,319. Standard errors are reported in parentheses below the coefficients. OLS-adjusted regressions include controls for child age, gender, first-born status, log of household permanent income, mother’s education, mother’s test score, mother’s height, number of siblings, and grandmother’s education.

1. Interpret the first column of results, which are a comparison of mean outcomes across the Head Start and other preschool groups as compared to the omitted category of “No preschool participation.” How do the PPVT vocabulary test scores of children who attend Head Start compare to those who attend other preschools and no preschools?

2. Explain whether a comparison of mean outcomes is informative for understanding the causal effects of Head Start on the PPVT score.

3. The second column of results shows the association of Head Start participation and other preschool participation while including observable child and family characteristics. Interpret the findings. Write a linear equation for modeling the effect of Head Start
participation on the PPVT score as was used for the OLS-adjusted results and offer a critique on its usefulness for obtaining a causal estimate.

4. The researchers concluded their study by estimating a mother fixed effect model which was possible due to the existence of siblings in the data set. Interpret the findings in the final column with respect to the causal estimate of Head Start. By referring to an equation, be specific about the usefulness of this method. What key assumption(s) and what key variation are used to identify the causal estimate using this framework?

5. Based on the results in the table and your discussion above, what do you conclude about the causal effect of Head Start on children’s PPVT scores?
Question 6:

In the 2014 *AER* paper by Piketty, Saez, and Stantcheva (PSS) taxable earnings are $z$ and the tax schedule is $T(z)$. The constant top marginal tax rate is $\tau$, applicable above an exogenously given threshold income level $\bar{z}$. The government sets $\tau$ to maximize tax revenue from top taxpayers.

(a) In the standard model with only supply-side responses, let $i$’s utility be $u_i(c, z) = c - h_i(z)$, where $z$ is pre-tax income, $c = z - T(z)$ is after-tax income, and $h_i(z)$ is “labor-supply cost of earning $z$,” with $h' > 0$ and $h'' < 0$. Derive a first-order condition determining the “optimal effort choice” $h_i$ and show that individual earnings $z_i(1 - \tau)$ depend only upon the net-of-tax rate $1 - \tau$.

(b) Let $z(1 - \tau)$, without an $i$ subscript, denote the average income reported by taxpayers with income above $\bar{z}$. Following PSS, we can define the aggregate elasticity of income in the top bracket with respect to the net-of-tax as

$$e_1 = \frac{1 - \tau}{\bar{z}} \frac{dz}{d(1 - \tau)}.$$  

PSS call this the labor supply effect. Let $a = z/(z - \bar{z})$ denote the Pareto parameter of the top tail of the income distribution. Show that the optimal tax rate, which maximizes revenue from taxpayers with income above $\bar{z}$, is given by

$$\tau^* = \frac{1}{1 + a \cdot e_1}.$$  

Explain and interpret this result from a policy perspective.

(c) How would you approach the question of selecting $\bar{z}$, the threshold for the top income category?