AAEC/ECON 5126 final exam

Spring 2020 / Instructor: Klaus Moeltner

May 12, 2020

This exam is open-book, open-notes, but please work strictly on your own. Please submit the exam electronically via Canvas (or e-mail) as a single pdf. You can collect a maximum of 50 points. Each question is scored as indicated below. Vectors are given in lower-case boldface. Matrices are written in upper-case boldface.

Question I (10 points):

Consider the linear regression model, expressed for a single observation i:

$$y_i = \beta_1 + \beta_2 x_{2i}^* + \epsilon_i, \tag{1}$$

where ϵ_i has the usual CLRM properties.

Assume, however, that x_{2i}^* is measured with *proportional error* for the entire sample, with the relationship between the observed x_{2i} and the true x_{2i}^* given as:

$$x_{2i} = x_{2i}^* \left(1 + \alpha\right), \quad \text{with} \quad 0 < \alpha < 1 \tag{2}$$

Part (a), 4 points
Express the model in (1) in terms of x_{2i} for a single observation *and for the full sample.* For the full model show that the measurement error can be interpreted as introducing omitted variable bias in a regression that uses $\mathbf{x}_2$ instead of $\mathbf{x}_2^*$.

Part (b), 4 points
Using partitioned regression, show that the estimated coefficient on $\mathbf{x}_2$ (call it b_2) is biased compared to the true β_2.

Part (c), 2 points
How could this problem be fixed if α were known?

Question 2 (20 points)

Consider the following true population model for a given individual i:

$$\begin{aligned} y_i &= \beta_0 + \beta_1 T_i + \beta_2 x_i + \epsilon_i \quad \text{with} \\ \epsilon_i &\sim n\left(0, \sigma^2\right), \end{aligned} \tag{3}$$

where y_i is some continuous outcome of interest, T_i is a binary (0/1) treatment indicator, x_i is a continuous explanatory variable, and ϵ_i is a standard error term with the usual CLRM properties.

Part (a), 4 points

a What is the true treatment effect?

b Show that it can be expressed as a difference between two expectations (conditional on x_i).

Part (b), 6 points

Assume you collect a random sample of individuals from this population. In your sample, you have n_1 treated and n_0 un-treated ("control") observations. For ease of notation, let outcome and explanatory variable for a treated observation be denoted as y_{Ti} and x_{Ti}, respectively. Analogously, let y_{Ci} and x_{Ci} be outcome and explanatory variable for a given control observation.

Assume you use some matching procedure to pair each treated observation with a single control observation. You then consider the following estimator for the population treatment effect (="average treatment effect for the treated"):

$$ATT_G|\mathbf{x} = \frac{1}{n_1}\sum_{i=1}^{n_1}\left(y_{Ti} - y_{Ci}\right),$$

where subscript "G" stands for "generic," $\mathbf{x}$ collects all relevant x_i's, and the summation is over all treated observations.

a Assume that the average difference between x_{Ti} and x_{Ci} across all matched pairs equals $\delta \neq 0$. Show that, under this assumption, this generic ATT (given $\mathbf{x}$) is biased.

b Under what conditions would this bias go to zero? Provide some verbal intuition.

Part (c), 10 points

Now consider applying the linear regression model given in (3) to the *matched control observations*, that is:

$$\begin{aligned}
&y_{Ci} = \beta_0 + \beta_2 x_{Ci} + \epsilon_i \quad \text{with} \\
&\epsilon_i \sim n\left(0, \sigma^2\right),
\end{aligned} \tag{4}$$

a Assume this model produces unbiased estimates for β_0 and β_2 (after all, you used the correct functional specification, and the correct error assumptions...). Call the coefficient estimates $\hat{\beta}_0$ and $\hat{\beta}_2$, respectively. Consider the linear predictions flowing from this model plugging in either some x_{Ci} or some x_{Ti}. Call these predictions $\hat{y}_{Ci}$ and $\hat{y}_{Ti}$, respectively. Show that they are also unbiased for the corresponding $E\left(y_i|x_{Ti}\right)$ and $E\left(y_i|x_{Ci}\right)$, respectively.

b Now consider the *regression-adjusted* treatment effect estimator ATT_R, given as:

$$ATT_R|\mathbf{x} = \frac{1}{n_1}\sum_{i=1}^{n_1}\left(\left(y_{Ti} - y_{Ci}\right) - \left(\hat{y}_{Ti} - \hat{y}_{Ci}\right)\right),$$

Show that this estimator is unbiased, regardless of the (average) difference between x_{Ti} and x_{Ci}.

c In terms of unbiasedness how does this regression-adjusted matching estimator for the true treatment effect compare to directly estimating β_1 using the regression model in (3) and the entire sample of treated and controls? *(A verbal response is sufficient)*.

Question 3 (20 points)

You are involved in a research project on beach visitation in Florida (FL). Beach visitors can be divided into three groups: (1) Locals (FL residents who live within 60 miles of a given beach), (2) FL tourists (FL residents who live further away than 60 miles from a given beach), and (3) out-of-state tourists. You are interested in the true proportions of these groups for all visitors to *Siesta Key*, a large, popular beach, on a specific day. Let these true proportions be labeled as π_1, π_2, and π_3, for locals, FL tourists, and out-of-state tourists respectively. Naturally, $\sum_{j=1}^{3} \pi_j = 1, \quad j = 1 \ldots 3$. Also, let $\boldsymbol{\pi} = \begin{bmatrix} \pi_1 & \pi_2 & \pi_3 \end{bmatrix}'$.

On your day of interest, you randomly sample n visitors to Siesta Key. For each, you write down an indicator vector $\mathbf{z}_i$ that shows to which group the person belongs. For example, if the person is a local, then $\mathbf{z}_i = \begin{bmatrix} z_{1i} & z_{2i} & z_{3i} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}$. Similarly, if the person is a FL tourist, the second element in $\mathbf{z}_i$ will be "1" and the other two will be "0," and if the person is an out-of-state tourist, $z_{3i} = 1$, and $z_{1i} = z_{2i} = 0$. Let the total number of individuals sampled for each group be n_1, n_2, and n_3, respectively.

You stipulate that each individual indicator vector $\mathbf{z}_i$ follows a multinomial likelihood, given as:

$$p\left(\mathbf{z}_i|\boldsymbol{\pi}\right) = \left(\frac{1}{\prod_{j=1}^{3} z_{ji}!}\right) \prod_{j=1}^{3} \pi_j^{z_{ji}} = \prod_{j=1}^{3} \pi_j^{z_{ji}} \tag{5}$$

Part (a), 2 points

Write down the likelihood for the entire sample of n observations. You can label the entire set of n indicator vectors as $\mathbf{z}$. Simplify as much as possible.

Part (b), 5 points

As a prior for $\boldsymbol{\pi}$ you choose a Dirichlet distribution. The density and moments for the Dirichlet are given as:

$$\begin{aligned} &p\left(\boldsymbol{\pi}\right) = \left(\frac{\Gamma\left(\tilde{\alpha}\right)}{\prod_{j=1}^{3} \Gamma\left(\alpha_j\right)}\right) \prod_{j=1}^{3} \pi_j^{\alpha_j - 1}, \quad \alpha_j > 0, \, \forall j, \\ &E\left(\pi_j\right) = \frac{\alpha_j}{\tilde{\alpha}}, \quad V\left(\pi_j\right) = \frac{\alpha_j\left(\tilde{\alpha} - \alpha_j\right)}{\tilde{\alpha}^2\left(\tilde{\alpha} + 1\right)}, \quad \text{where} \\ &\tilde{\alpha} = \sum_{j=1}^{3} \alpha_j \end{aligned} \tag{6}$$

a Derive the *kernel* of the posterior distribution $p\left(\boldsymbol{\pi}|\mathbf{y}\right)$, and determine the statistical distribution for the full posterior.

b Show the posterior parameters for this distribution (label them α_j^*, $j = 1 \dots 3$).

Part (c), 5 points

Assume you have visitor information from *other nearby beaches*, with average proportions for the three visitor groups of 0.5, 0.1, and 0.4. Interpreting these averages as prior expectations, and letting $\alpha_1 = 10$, derive the prior parameters α_2 and α_3, as well as the prior variances for the three shares. Round the variances to four decimals.

Part (d), 8 points

Assume your Siesta Key sample of 200 visitors produces $n_1 = 80$, $n_2 = 10$, and $n_3 = 110$.

a Using all the information from above, compute the posterior expectations and variances for the population shares. Round all expectations to three decimals, and all variances to four decimals.

b How can you tell that the collected data has brought information to the prior?

c The town of Siesta Key is willing to sponsor an advertising campaign targeted to *in-state tourists* (group 2), if the posterior share of this group falls below 10%. What will be the town's decision?