AAEC/ECON 5126 final exam

Spring 2018 / Instructor: Klaus Moeltner

May 8, 2018

This exam is open-book, open-notes, but please work strictly on your own. Please make sure your name is on every sheet you're handing in. You have 120 minutes to complete this exam. 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 (12 points):

Consider a city located downstream of a river dam. To facilitate spawning runs of salmon, the government decides to remove the dam. This will place a large fraction of residential properties in the city in a "Special Flood Hazard Area" (SFHA) with a higher risk of flooding during storm events. You are interested in estimating the loss in property values from being located in a SFHA. You collect data on the sale price of each home i, y_i, as well as many structural and neighborhood characteristics $\mathbf{x}_i$. Your plan is to run the following CLRM:

$$y_i = \alpha + \mathbf{x}_i'\boldsymbol{\beta} + \gamma d_i + \epsilon_i, \tag{1}$$

where d_i is a binary indicator that takes the value of 1 if a residence is located in a SFHA, and a value of zero otherwise.

Part (a), 3 points

(a) Write the model at the sample level, using notation $\mathbf{y}$, $\mathbf{X}$, $\mathbf{d}$, and $\boldsymbol{\epsilon}$.

(b) Assume the functional form of the regression equation is correct. Under which conditions, involving $\mathbf{X}$, $\boldsymbol{\epsilon}$, and $\mathbf{d}$ will the OLS solution $\hat{\gamma}$ be an unbiased estimate of the true SFHA effect?

(c) Now assume you have an omitted variable problem, that is a correlation of $\boldsymbol{\epsilon}$ with one or more elements of $\mathbf{X}$. Under which condition will remain $\hat{\gamma}$ be an unbiased estimate of the true SFHA effect? How would *pre-matching the sample* before running this regression help in this case? Explain in detail.

Solution:

(a)
$$\mathbf{y} = \alpha\mathbf{i} + \mathbf{X}\boldsymbol{\beta} + \gamma\mathbf{d} + \boldsymbol{\epsilon},$$

(b) BOTH $\mathbf{X}$ and $\mathbf{d}$ are uncorrelated with $\boldsymbol{\epsilon}$ OR only $\mathbf{X}$ is correlated with $\boldsymbol{\epsilon}$ and $\mathbf{X}$ and $\mathbf{d}$ are uncorrelated with each other.

(c) **X** and **d** are uncorrelated with each other. Pre-matching can help in that it picks a sample of control homes that has a very similar distribution for all variables in **X** compared to the treated. This makes the treatment status (virtually) independent of the other control variables by design. This, in turn, guards against O.V. issues such as those described in the question.

Part (b) 9 points
Now assume that $\mathbf{x}_i$ includes all relevant control variables, such that there are no O.V. problems. ***In words***, and using as much detail as necessary, describe how you would estimate the true SFHA effect using the following alternative methods. For each case, be specific in how you would estimate the Average Treatment Effect on the Treated (ATT), and its standard error.

(a) A separate regression approach - one regression using only the treated observations, and a second regression using only controls. When would that be useful or warranted?

(b) A regression model such as (1), but using the propensity score instead of $\mathbf{x}_i$. Under which conditions involving the PS would this produce a consistent estimate of the SFHA effect?

(c) Using a matching estimator with regression correction. You can assume 1 nearest neighbor. What is the main advantage of this hybrid model over the other approaches?

Solution

(a) Run a separate regression for the treated and control homes. From each regression, generate predicted outcomes for *all* observations. Estimate the home-specific treatment effect as the difference of predicted prices for each pair of predictions. Then average these differences over all *treated* homes to obtain an estimate of the ATT. Compute standard errors using the bootstrap method. This approach is useful if you want to allow for *separate sets of coefficients* in the control and treated regression.

(b) Estimate the PS via MLE (e.g. logit model). Then insert the predicted PS into the regression models for controls and treated instead of the explanatory variables (X). This makes sense if the *estimated propensity score* itself is consistent, that is the PS equation is correctly specified and not plagued by O.V. problems. Standard errors are obtained via bootstrap methods, where the bootstrap includes both re-running the PS equation and the actual regression.

(c) For each treated, find the closest control home based on a Euclidean distance metric. (Optional response: For optimally *balanced* matching, this embeds an MLE routine to find the optimal weights for each variable in the distance metric). Then, *using matched controls only*, run a regression of prices on all observables **X** and generate predictions for both treated and matched controls. Compute the counterfactual for each treated as (price of matched control + predicted price for the treated - predicted price for the control). Compute the difference of (price of treated - counterfactual) for each treated. Average this difference over all treated. Compute standard errors using Abadie and Imbens' (2011) analytical method. (The bootstrap is inconsistent for this case).

The main advantage of this approach is its *double robustness* - the ATT will be unbiased if either the controls are perfect matches, or the auxiliary regression generates unbiased predictions of price.

Question II (18 points):

Consider the Poisson model for a random variate y with parameter λ, given as

$$\begin{aligned} p(y|\lambda) &= \frac{\lambda^y exp(-\lambda)}{y!}, \quad \text{with} \\ E(y|\lambda) &= V(y|\lambda) = \lambda, \quad \lambda > 0, y \in \{0, 1, 2, 3, ...\} \end{aligned} \tag{2}$$

Part (a), 4 points
Now consider a sample of n observations from this distribution, with each observation generically labeled $y_i, i = 1 \ldots n$. Write down the joint distribution for the sample data (in *un*-logged form). Call it $p(\mathbf{y}|\lambda)$.

Solution:

$$p(\mathbf{y}|\lambda) = \prod_{i=1}^{n} \frac{\lambda^{y_i} exp(-\lambda)}{y_i!} = \left(\prod_{i=1}^{n} \frac{1}{y_i!} \right) \lambda^{\left(\sum_{i=1}^{n} y_i\right)} exp(-n\lambda)$$

Part (b), 8 points
Suppose you stipulate a *gamma* prior density for λ with shape parameter a and inverse scale ("rate") parameter b, given as

$$\begin{aligned} p(\lambda) &= g(a, b) = \frac{b^a}{\mathbf{\Gamma}(a)} \lambda^{(a-1)} exp(-b\lambda), \quad \text{with} \\ E(\lambda) &= \frac{a}{b}, \; V(\lambda) = \frac{a}{b^2}, \quad \lambda, a, b > 0, \end{aligned} \tag{3}$$

Show that the posterior distribution of λ, given your collected data from the Poisson, is also a gamma. Show the form of the posterior shape and rate parameters (you can call them a^* and b^*).

Solution:

$$\begin{gathered} p(\lambda|\mathbf{y}) \propto \lambda^{(a-1)} exp(-b\lambda) \lambda^{\left(\sum_{i=1}^{n} y_i\right)} exp(-n\lambda) = \\ \lambda^{\left(a + \sum_{i=1}^{n} y_i - 1\right)} exp(-(b+n)\lambda) \end{gathered}$$

This describes the kernel of another gamma density with posterior shape $a^* = a + \sum_{i=1}^{n} y_i$ and posterior rate $b^* = b + n$. Therefore, we can deduce that $\lambda|\mathbf{y} \sim g(a + \sum_{i=1}^{n} y_i, b + n)$.

Part (c), 4 points
Show that the posterior expectation can be written as a weighted average of the prior expectation and the sample mean. What happens to this posterior expectation as $n \to \infty$?

Solution:

$$E(\lambda|\mathbf{y}) = \frac{a + \sum_{i=1}^{n} y_i}{b+n} = \left(\frac{b}{b+n}\right)\frac{a}{b} + \left(\frac{n}{b+n}\right)\frac{\sum_{i=1}^{n} y_i}{n}$$

The limit of the first weight is 0, and that of the second weight is 1, so as the sample size increases the posterior expectation will converge to the sample mean.

Part (d), 2 points

Suppose you are opening a small restaurant In Blacksburg. Before you start your business, you expect 20 guests / day with a variance of 10, which can be modeled as a gamma prior with shape 40 and rate 2. After 30 days of running your business, you count a total of 824 guests. You plot the daily counts, and they look exactly like a Poisson distribution.

How many guest *per day* would you expect for the following month?

Solution:
Based on the results from the previous section, the posterior expectation for daily visits can be computed as $\frac{40+824}{2+30} = \frac{864}{32} = 27$

Question 3 (20 points)

Consider a CLRM of the following form, at the observation level (dropping individual-level subscripts for simplicity):

$$\begin{aligned} &y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 x_2^2 + \epsilon, \quad \text{where} \\ &\epsilon \sim i.i.d.\left(0, \sigma^2\right) \end{aligned} \tag{4}$$

Part (a), 2 points
You are primarily interested in the marginal effect of x_1 and x_2 on the outcome variable, i.e. $\left(\frac{\partial y}{\partial x_1}\right)$, and $\left(\frac{\partial y}{\partial x_2}\right)$. Show the explicit form of these marginal effects, for a given x_1 and x_2.

Solution:

$$\begin{aligned} \frac{\partial y}{\partial x_1} &= \beta_1 + 2\beta_3 x_1 x_2^2 \\ \frac{\partial y}{\partial x_2} &= \beta_2 + 2\beta_3 x_2 x_1^2 \end{aligned}$$

Part (b), 6 points

Let $E(x_1) = \mu_1$ and $E(x_2) = \mu_2$ be the population means of the two explanatory variables. Similarly, let σ_1^2 and σ_2^2 be the two variances. Assume all of these moments are known to the analyst. Also, it is known that x_1 and x_2 are *independently* distributed.

Derive the expectation, over x_1 and x_2, of these marginal effects you obtained in the preceding part in terms of these moments. Let the solutions be labeled as $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$, respectively.

Solution:
Due to independence, we have $E\left(x_i x_j^2\right) = E(x_i)\, E\left(x_j^2\right) = \mu_i\left(\sigma_j^2 + \mu_j^2\right), i, j = 1, 2$, thus:

$$\begin{aligned} \boldsymbol{\gamma}_1 &= E\left(\frac{\partial y}{\partial x_1}\right) = \beta_1 + 2\beta_3 \mu_1\left(\sigma_2^2 + \mu_2^2\right) \\ \boldsymbol{\gamma}_2 &= E\left(\frac{\partial y}{\partial x_1}\right) = \beta_2 + 2\beta_3 \mu_2\left(\sigma_1^2 + \mu_1^2\right) \end{aligned}$$

Part (c), 4 points
Setting $x_2 = \mu_2$, at what value of x_1 is y maximized?
How do you know it's a maximum? (Assume $\beta_3 < 0$)

Solution:

$$\left(\frac{\partial y}{\partial x_1}|x_2 = \mu_2\right) = 0 \rightarrow$$
$$x_1^* = -\frac{\beta_1}{2\beta_3\mu_2^2}$$
$$\left(\frac{\partial y^2}{\partial^2 x_1}|x_2 = \mu_2\right) = 2\beta_3\mu_2^2 < 0$$

Part (d), 4 points
Using the results form part (b), solve for β_1 and β_2, then insert the resulting expressions into equation (4) in lieu of β_1 and β_2.

After some manipulation, this should produce the following "reduced-form" model:

$$y = \beta_0 + \boldsymbol{\gamma}_1 x_1 + \boldsymbol{\gamma}_2 x_2 + \beta_3 \left(f\left(x_1, x_2, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2\right)\right) + \epsilon, \tag{5}$$

where you need to fill in the explicit form of $f(.)$.

Solution:

$$y = \beta_0 + \boldsymbol{\gamma}_1 x_1 + \boldsymbol{\gamma}_2 x_2 + \beta_3 \left(x_1^2 x_2^2 - 2x_1\mu_1\left(\mu_2^2 + \sigma_2^2\right) - 2x_2\mu_2\left(\mu_1^2 + \sigma_1^2\right)\right) + \epsilon,$$

Part (e), 4 points
Now suppose that $\mu_1 = \mu_2 = 0$, and you estimate the model in (4). What is the interpretation of β_1 and β_2?

Solution:
In that case, we have $\beta_1 = \gamma_1$ and $\beta_2 = \gamma_2$, that is both coefficients can immediately be interpreted as the population expectation, over x_1 and x_2, of the marginal effects of these variables on y.