AAEC/ECON 5126 final exam

Spring 2017 / Instructor: Klaus Moeltner

May 9, 2017

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 (18 points): Endogeneity

Consider a regression of "Body Mass Index (BMI)" for teenagers on a set of explanatory variables:

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + s_i\boldsymbol{\gamma} + \epsilon_i, \quad \epsilon_i \sim n\left(0, \sigma_\epsilon^2\right) \tag{1}$$

where y_i is BMI (a well-behaved, continuous variable), $\mathbf{x}_i$ includes a set of exogenous regressors, and s_i measures the average daily amount of exercise for individual i, expressed in hours (allowing for fractional hours). The error term has the usual CLRM properties.

Unbeknownst to the researcher, the exercise variable s_i follows a second regression model:

$$s_i = \delta y_i + \eta_i \tag{2}$$

where η_i is a "well-behaved" CLRM error term. Assume that $\mathbf{x}_i$, ϵ_i, and η_i are uncorrelated with one another.

Part (a) *2 points*
Express s_i as a function of all other components in (1) and (2), except for y_i.

Solution:

$$s_i = \delta\left(\mathbf{x}_i'\boldsymbol{\beta} + s_i\boldsymbol{\gamma} + \epsilon_i\right) + \eta_i = \frac{\delta}{1-\delta\boldsymbol{\gamma}}\left(\mathbf{x}_i'\boldsymbol{\beta} + \epsilon_i\right) + \frac{1}{1-\delta\boldsymbol{\gamma}}\eta_i$$

Part (b)*4 points*
Derive the true covariance between s_i and ϵ_i.

Solution:

$$cov\left(s_i, \epsilon_i\right) = cov\left(\frac{1}{1-\delta\boldsymbol{\gamma}}\delta\left(\mathbf{x}_i'\boldsymbol{\beta} + \eta_i\right) + \frac{\delta}{1-\delta\boldsymbol{\gamma}}\epsilon_i, \epsilon_i\right) =$$
$$cov\left(\frac{\delta}{1-\delta\boldsymbol{\gamma}}\epsilon_i, \epsilon_i\right) = \frac{\delta}{1-\delta\boldsymbol{\gamma}}\sigma_\epsilon^2$$

Part (c) *2 points*
Let the full regression model in (1) for the entire sample of n observations be written as

$$\begin{aligned} &\mathbf{y} = \mathbf{M}\boldsymbol{\theta} + \boldsymbol{\epsilon}, \quad \text{where} \\ &\mathbf{M} = \begin{bmatrix} \mathbf{X} & \mathbf{s} \end{bmatrix}, \quad \boldsymbol{\theta} = \begin{bmatrix} \boldsymbol{\beta}' & \boldsymbol{\gamma}' \end{bmatrix}' \end{aligned} \tag{3}$$

Derive $cov(\mathbf{s}, \boldsymbol{\epsilon}) = E(\mathbf{s}'\boldsymbol{\epsilon})$, using your result from part (b).
(Hint: $cov(\epsilon_i, s_i) = E(\epsilon_i s_i)$.)

Solution:

$$E(\mathbf{s}'\boldsymbol{\epsilon}) = E\left(\sum_{i=1}^{n}(s_i\epsilon_i)\right) = nE(s_i\epsilon_i) = n\frac{\delta}{1-\delta\boldsymbol{\gamma}}\sigma_\epsilon^2$$

Part (d) *6 points*
Using the *partitioned form* of the OLS estimator for $\boldsymbol{\theta}$ (call it $\hat{\boldsymbol{\theta}}$) in terms of $\mathbf{X}$ and $\mathbf{s}$, show that it is biased. (Hint: You do NOT need to solve the partitioned inverse matrix involved in this operation. Also note: I'm NOT asking you to do a *partitioned regression*.)

Solution:

$$E\left(\hat{\boldsymbol{\theta}}\right) = \boldsymbol{\theta} + \begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{s} \\ \mathbf{s}'\mathbf{X} & \mathbf{s}'\mathbf{s} \end{bmatrix}^{-1} E\begin{bmatrix} \mathbf{X}'\boldsymbol{\epsilon} \\ \mathbf{s}'\boldsymbol{\epsilon} \end{bmatrix} = \boldsymbol{\theta} + \begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{s} \\ \mathbf{s}'\mathbf{X} & \mathbf{s}'\mathbf{s} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{0} \\ n\frac{\delta}{1-\delta\boldsymbol{\gamma}}\sigma_\epsilon^2 \end{bmatrix} \neq \mathbf{0}$$

Part (e) *4 points*
Describe, in words, how one could use an instrumental variable approach and Two-Stage-Least-Squares (TSLS) estimation to overcome this endogeneity problem. Can you think of one or two good instruments?

Solution:

1. Find one or more suitable instruments - variables that are (likely...) not directly related to BMI, but highly correlated with exercise. Examples: Exercise levels of parents/siblings/friends, sports facilities available at school and near-home, time spent on electronic devices / video games / watching TV, etc, and so on.
2. Regress $\mathbf{M}$ against the original $\mathbf{X}$ and instruments, and predict $\hat{\mathbf{M}}$.
3. Use $\hat{\mathbf{M}}$ in the original regression (3) in lieu of $\mathbf{M}$.

Question II (14 points): Measurement error

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

$$y_i = \beta_1 + \beta_2 x_{2i}^* + \epsilon_i, \tag{4}$$

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:

$$x_{2i} = x_{2i}^* (1 + \alpha), \quad \text{with} \quad 0 < \alpha < 1 \tag{5}$$

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

Solution:

$$\begin{aligned}
&x_{2i} = x_{2i}^* (1 + \alpha) \rightarrow x_{2i}^* = \left(\frac{1}{1+\alpha}\right) x_{2i} \\
&y_i = \beta_1 + \tilde{\beta}_2 x_{2i} + \epsilon_i \quad \text{with } \tilde{\beta}_2 = \beta_2 \left(\frac{1}{1+\alpha}\right) \\
&\mathbf{y} = \mathbf{i}\beta_1 + \tilde{\beta}_2 \mathbf{x}_2 + \boldsymbol{\epsilon}, \quad \text{or} \\
&\mathbf{y} = \mathbf{i}\beta_1 + \beta_2 \mathbf{x}_2 + \left(\boldsymbol{\epsilon} - \frac{\alpha\beta_2}{1+\alpha}\mathbf{x}_2\right)
\end{aligned}$$

This makes it clear that $\mathbf{x}_2$ is correlated with the true error term.

Part (b)*6 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.

Solution:

$$\begin{aligned}
&b_2 = \left(\mathbf{x}_2' M_0 \mathbf{x}_2\right)^{-1} \left(\mathbf{x}_2' M_0 \mathbf{y}\right) \\
&E\left(b_2 | \mathbf{x}_2\right) = \\
&E[\left(\mathbf{x}_2' M_0 \mathbf{x}_2\right)^{-1} \left(\mathbf{x}_2' M_0 \mathbf{i}\beta_1\right)] + \\
&E[\left(\mathbf{x}_2' M_0 \mathbf{x}_2\right)^{-1} \left(\mathbf{x}_2' M_0 \mathbf{x}_2\right) \beta_2] + \\
&E[\left(\mathbf{x}_2' M_0 \mathbf{x}_2\right)^{-1} \left(\mathbf{x}_2' M_0 \left(\boldsymbol{\epsilon} - \frac{\alpha\beta_2}{1+\alpha}\mathbf{x}_2\right)\right)]
\end{aligned}$$

The first term goes to zero since $M_0\mathbf{i} = \mathbf{0}$, the second term yields β_2, and the third term gives:

$$E[\left(\mathbf{x}_2' M_0 \mathbf{x}_2\right)^{-1} \left(\mathbf{x}_2' M_0 \boldsymbol{\epsilon}\right)] - E[\left(\mathbf{x}_2' M_0 \mathbf{x}_2\right)^{-1} \left(\mathbf{x}_2' M_0 \frac{\alpha\beta_2}{1+\alpha} \mathbf{x}_2\right)] = -\frac{\alpha\beta_2}{1+\alpha}$$

In total, we thus obtain:

$$E\left(b_2 | \mathbf{x}_2\right) = \beta_2 - \frac{\alpha}{1+\alpha}\beta_2 = \beta_2 \left(\frac{1}{1+\alpha}\right) = \tilde{\beta}_2$$

Part (c) *4 points*
How could this problem be fixed if α were known?

Solution:
If α is known, simply use $\mathbf{x}_2^* = \frac{1}{1+\alpha}\mathbf{x}_2$ in the regression model instead of $\mathbf{x}_2$, that is, fix the measurement error before using the variable.

Question III (18 points): Bayesian problem

You are interested in the fraction θ of all eligible VT students that participated in a recent election, with $0 \leq \theta \leq 1$. Your prior on θ is a Beta distribution, given as:

$$\begin{aligned} & p(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}, \quad \text{with} \\ & \alpha, \beta > 0, \\ & E(\theta) = \frac{\alpha}{\alpha+\beta} \\ & V(\theta) = E(\theta)\frac{\beta}{(\alpha+\beta)(\alpha+\beta+1)} \end{aligned} \tag{6}$$

You poll 100 randomly selected eligible students, and 43 of them report that they did go and vote in the election.

You decide that the sample distribution ("likelihood function") of VT voters y, out of a sample of size n of eligible students, is well characterized by a binomial density, given as:

$$p(y|\theta) = \binom{n}{y}\theta^{y}(1-\theta)^{n-y} \tag{7}$$

Part (a) *6 points*
Based on past years of elections, the average participation rate was 0.2, with a variance of 0.1. Find prior parameters α and β to fit these moments.

Solution:

$$E(\theta) = \frac{\alpha}{\alpha+\beta} = 0.2 \rightarrow \alpha = \frac{0.2}{0.8}\beta = 0.25\beta$$
$$V(\theta) = 0.2\frac{\beta}{(1.25\beta)(1.25\beta+1)} \rightarrow \beta = 0.48 \rightarrow \alpha = 0.12$$

Part (b) *6 points*
Characterize the posterior kernel for θ and show that this is the kernel of another Beta density with parameters α_1, β_1. Show the explicit forms of these posterior parameters and derive their numerical values.

Solution:

$$p(\theta|y,n) \propto \theta^{\alpha_0-1}(1-\theta)^{\beta_0-1} * \theta^{y}(1-\theta)^{n-y} =$$
$$\theta^{(\alpha_0+y-1)}(1-\theta)^{(\beta_0+n-y-1)}$$

Thus, $\theta|y,n \sim \text{Beta}(\alpha_1, \beta_1)$, with $\alpha_1 = \alpha_0 + y$, and $\beta_1 = \beta_0 + n - y$. In this case:

$$\alpha_1 = \alpha + y = 0.12 + 43 = 43.12$$
$$\beta_1 = \beta + n - y = 0.48 + 57 = 57.48$$

Part (c) *6 points*
Derive the posterior expectation and variance of θ. How do these moments compare to your priors?

Solution:

$$E(\theta|y) = \frac{43.12}{43.12 + 57.48} \approx 0.43 > E(\theta)$$
$$V(\theta|y) = 0.43\frac{57.48}{100.6 * 101.6} \approx 0.0024 << V(\theta) \quad \text{as expected.}$$