Econ 6200: Econometrics II
Prelim, April 11^{th}, 2023

This exam consists of ten questions, not of equal length or difficulty, grouped into three exercises. The questions are only partly cumulative. Each question is worth 10 points. Remember to always explain your answer.

Good luck!

1. Suppose you observe i.i.d. data from the following true process:

$$\begin{aligned} Y &= X - X^2 - 1 + \varepsilon \\ \varepsilon &\sim \mathcal{N}(0,1) \\ X &\sim \mathcal{N}(0,1), \end{aligned}$$

where X and ε are furthermore uncorrelated.

A researcher claims that in this specific example, just regressing Y on X, i.e. estimating the model

$$Y = \beta_0 + \beta_1 X + \nu$$

will tend to recover the coefficient on X in the true model, i.e. 1. Is this true:

1.1 ...in the sense of consistency, i.e. $\hat{\beta}_1 \xrightarrow{p} 1$?

1.2 ...in the sense of unbiasedness, i.e. $\mathbb{E}\hat{\beta}_1 = 1$?

If you cannot answer the above questions, proceed on the assumption that at least one answer above is "yes."

For the remainder of this exercise, we change the example by assuming that $X \sim \mathcal{N}(1,1)$. Everything else stays the same.

1.3 Without doing any calculations, do you expect that your answers to the previous questions would change?

1.4 Characterize the probability limit and asymptotic distribution (up to a variance term that you need not compute) of $\hat{\beta}_1$.

(You may assume that for $X \sim \mathcal{N}(1,1)$ one has $\mathbb{E}X^3 = 4$.)

1.5 While I called 1 the "true" coefficient on X, $\text{plim}\,\hat{\beta}_1$ is also a quantity of potential interest. Explain the distinction I make here. Can you formulate a question or decision problem for which $(\text{plim}\,\hat{\beta}_0, \text{plim}\,\hat{\beta}_1)$ is the right answer?

2 This question is about linear GMM. Suppose that there are $L = 8$ moments and $K = 5$ parameters. Throughout this question, suppose that data are i.i.d., that moments exist as needed, and furthermore that:

$$\boldsymbol{Y} = \boldsymbol{X}\theta + \varepsilon$$
$$\mathbb{E}\left(\boldsymbol{Z'}(\boldsymbol{Y} - \boldsymbol{X}\theta)\right) = \mathbf{0}$$
$$\boldsymbol{Q} \equiv \mathbb{E}(\boldsymbol{Z}'\boldsymbol{X}) \text{ is of full column rank.}$$
$$\Omega \equiv \mathbb{E}\left(\boldsymbol{Z}'\boldsymbol{Z}\varepsilon^2\right) \text{ is nonsingular.}$$

You may assume the following theorem: Let $\hat{\theta}(\hat{\boldsymbol{W}})$ be the GMM estimator and let $\hat{\boldsymbol{W}} \xrightarrow{p} \boldsymbol{W}$, where $\boldsymbol{W}$ is symmetric with full rank. Then (under above assumptions) $\hat{\theta}(\hat{\boldsymbol{W}}) \xrightarrow{p} \theta$ and $\sqrt{n}\big(\hat{\theta}(\hat{\boldsymbol{W}}) - \theta\big) \xrightarrow{d} \mathcal{N}\big(0, \text{avar}(\hat{\theta}(\hat{\boldsymbol{W}}))\big)$.

A researcher wants to estimate θ by GMM with weighting matrix

$$\tilde{\mathbf{W}} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} \mathbf{I}_5 & \underset{5\times 3}{\mathbf{0}} \\ \underset{3\times 5}{\mathbf{0}} & \underset{3\times 3}{\mathbf{0}} \end{bmatrix}.$$

2.1 Write down the optimization problem that defines $\hat{\theta}(\tilde{\mathbf{W}})$ in data matrix notation. Show in closed form that $\hat{\theta}(\tilde{\boldsymbol{W}}) = \left(\boldsymbol{X}'\boldsymbol{Z}\tilde{\boldsymbol{W}}\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Z}\tilde{\boldsymbol{W}}\boldsymbol{Z}'\boldsymbol{Y}$.

2.2 The above theorem does not immediately apply to $\hat{\theta}(\tilde{\boldsymbol{W}})$! Why not?

2.3 However, $\hat{\theta}(\tilde{\boldsymbol{W}})$ is consistent and asymptotically normal anyway. Why? (Hint: Is the closed form from 2.1 the most succinct expression of $\hat{\theta}(\tilde{\boldsymbol{W}})$?)

2.4 Is it possible that $\hat{\theta}(\tilde{\mathbf{W}})$ is efficient in the usual sense of attaining the lower bound on asymptotic variance for *all* data generating processes that fulfil our assumptions? Is it possible that $\hat{\theta}(\tilde{\mathbf{W}})$ is efficient for *some* such data generating processes? (Just give general answers, do not try to construct examples.)

2.5 Is the estimator

$$\tilde{\theta} = \hat{\theta}\big(\hat{\Omega}^{-1}\big)$$
$$\hat{\Omega} = \frac{1}{n}\sum_{i=1}^{n}\big(Z_i Z_i'(Y_i - X_i'\hat{\theta}(\tilde{\boldsymbol{W}})^2\big) \qquad \left[= \frac{1}{n}\boldsymbol{Z}'\hat{\varepsilon}\hat{\varepsilon}'\boldsymbol{Z}\right]$$

efficient? Explain why or why not.

Answer Key

1 This question was mainly meant to be about interpreting OLS as projections versus estimation of a structural linear model.

1.1 Under our distributional assumptions, we have

$$\operatorname{cov}(X,\nu)=\operatorname{cov}(X,-X^2-1)=-\operatorname{cov}(X,X^2)=0,$$

where we used that $\operatorname{cov}(X,X^2)=\mathbb{E}(X^3)-\mathbb{E}X\cdot\mathbb{E}(X^2)=0$. Note for future reference that this crucially uses symmetry of the distribution of X around 0 (hence, odd functions have expected value of 0).

1.2 This question was unintentionally hard. Thinking of $\nu=-X^2+\varepsilon$ as error term, it is *not* true that $\mathbb{E}(\nu|X)=\mathbb{E}(\nu)$, which would be needed for the standard proof of conditional unbiasedness of the slope parameter (the value of $\mathbb{E}\nu$ gets absorbed into the intercept). And conditional (on X) unbiasedness does not obtain. This is what I had really meant to ask about.

However, in an interesting twist, Brenda first discovered that the expected value of $\hat{\beta}_1$ in this specific example is indeed 1. Here is my proof, which goes well beyond what could be expected under exam conditions. Recall that a Rademacher random variable equals ± 1 with equal probability. Since X is i.i.d. $\mathcal{N}(0,1)$, $(X_1,\ldots,X_n)$ is distributionally equivalent to $R\cdot(Z_i,\ldots,Z_n)$, where Z is i.i.d. $\mathcal{N}(0,1)$ and R is *one* (constant across i) realization of a Rademacher r.v. With that in mind, write:

$$\begin{aligned}
\mathbb{E}\hat{\beta}_1 &= \mathbb{E}\left(\frac{\operatorname{cov}_n(X,-1+X-X^2+\varepsilon)}{\operatorname{var}_n(X)}\right)\\
&= 1+\mathbb{E}\left(\frac{\operatorname{cov}_n(X,-X^2+\varepsilon)}{\operatorname{var}_n(X)}\right)\\
&= 1+\frac{1}{2}\mathbb{E}\left(\frac{\operatorname{cov}_n(X,-X^2+\varepsilon)}{\operatorname{var}_n(X)}\mid R=1\right)+\frac{1}{2}\mathbb{E}\left(\frac{\operatorname{cov}_n(X,-X^2+\varepsilon)}{\operatorname{var}_n(X)}\mid R=-1\right)\\
&= 1+\frac{1}{2}\mathbb{E}\left(\frac{\operatorname{cov}_n(Z,-Z^2+\varepsilon)}{\operatorname{var}_n(Z)}\right)+\frac{1}{2}\mathbb{E}\left(\frac{\operatorname{cov}_n(-Z,-(-Z)^2+\varepsilon)}{\operatorname{var}_n(-Z)}\right)\\
&= 1+\frac{1}{2}\mathbb{E}\left(\frac{\operatorname{cov}_n(Z,-Z^2+\varepsilon)}{\operatorname{var}_n(Z)}\right)+\frac{1}{2}\mathbb{E}\left(\frac{\operatorname{cov}_n(-Z,-Z^2+\varepsilon)}{\operatorname{var}_n(Z)}\right)\\
&= 1+\frac{1}{2}\mathbb{E}\left(\frac{\operatorname{cov}_n(Z,-Z^2+\varepsilon)}{\operatorname{var}_n(Z)}\right)+\frac{1}{2}\mathbb{E}\left(\frac{-\operatorname{cov}_n(Z,-Z^2+\varepsilon)}{\operatorname{var}_n(Z)}\right)\\
&= 1,
\end{aligned}$$

using that the covariance is linear.

1.3 Yes. The calculation heavily uses symmetry of the distribution of X (and strictly speaking also continuity, since the algebra requires that $\operatorname{var}_n(X)$ is almost never zero).

1.4 First, we have

$$\begin{aligned}
&\operatorname{plim}\hat\beta_1 = \operatorname{plim}\frac{\operatorname{cov}_n(X,Y)}{\operatorname{var}_n(X)} = \operatorname{plim}\big((X-\mathbb{E}X)(-1+X-X^2+\varepsilon)\big)\\
= \;& \operatorname{plim}(-X+\mathbb{E}X+X^2-X\mathbb{E}X-X^3+X^2\mathbb{E}X+X\varepsilon-\varepsilon\mathbb{E}X)\\
= \;& -\mathbb{E}X+\mathbb{E}X+\mathbb{E}X^2-(\mathbb{E}X)^2-\mathbb{E}X^3+\mathbb{E}X^2\mathbb{E}X+\mathbb{E}(X\varepsilon)-\mathbb{E}\varepsilon)\mathbb{E}X\\
= \;& -1+1+2-1-4+2+0-0=-1.
\end{aligned}$$

The asymptotic distribution will be normal centered at (-1) and with an asymototic fariance that you were not asked to compute (it would require higher moments of X, which are trivial but tedious to evaluate).

1.5 The coefficient value of 1 is "true" if we think of $Y=-1+X-X^2+\varepsilon$ as being structural, i.e. having a causal interpretation, or at a minimum if we assume that $\mathbb{E}(Y \mid X)=1+X-X^2$. However, $\hat\beta_1$ very generally estimates the projection coefficient of Y and X and thereby the Best Linear Predictor. One question this estimand (almost) always answers is: "If we had to linearly predict Y from X under square loss, how much should our prediction vary with$X$?"

2.1 See slides on GMM.

2.2 Because $\tilde{\boldsymbol{W}}$ does not have full rank.

2.3 The weighting matrix effectively discards the last 3 moment conditions, leaving us with a just identified system (i.e., IV). Sure enough, the estimator can be simplified to $\hat\theta=(\boldsymbol{Z}'\boldsymbol{X})^{-1}\boldsymbol{Z}'\boldsymbol{Y}$.

2.4 No, the estimator is typically not efficient. However, it is efficient in the (generally unknowable) case where $\tilde{\boldsymbol{W}}$ identifies the same minimum as Ω^{-1}. Note that this happens if we know the likelihood and the first 5 conditions are the score equations.

2.5 Yes, this is a successful 2SGMM estimator. Recall that the first stage estimator only needs to be consistent!