Section 11: GMM

ARE 210

November 14, 2017

1)

a) Suppose we draw a random sample of individuals from a large population to estimate the average income. We have two unbiased observations of individual i's income μ_i: $X_i^{\mathrm{T}} = (x_{1i}, x_{2i})$. Construct the efficient GMM estimator of the population average income μ. Interpret the estimator and the effect of the optimal weighting matrix.

Write our stacked moment conditions $m(X_i;\mu) = X_i - \mu$. The optimal weighting matrix is $V[m(X_i;\mu)] = V[X_i]^{-1}$. Let $V \equiv V[X_i]$, then $V[X_i]^{-1} = \frac{1}{V_{11}V_{22} - V_{12}^2}\begin{pmatrix} V_{22} & -V_{12} \\ -V_{12} & V_{11} \end{pmatrix}$. Our GMM estimator solves

$$\hat{\mu}_{GMM} = \arg\max_{\hat{\mu}} -\frac{1}{2}\left(\frac{1}{n}\sum_{i=1}^{n} m(X_i;\mu)\right)^{\mathrm{T}} V[X_i]^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} m(X_i;\mu)\right)$$

Let $M(X_i;\mu) \equiv D_\mu m(X_i;\mu) = (-1,-1)^{\mathrm{T}}$. Then the first order condition for the maximization problem gives $\left(\frac{1}{n}\sum_{i=1}^{n} M(X_i;\hat{\mu}_{GMM})\right) V[X_i]^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} m(X_i;\hat{\mu}_{GMM})\right) = 0$. Simplifying yields

$$\hat{\mu}_{GMM} = \frac{(V_{22} - V_{12})\overline{X_1} + (V_{11} - V_{12})\overline{X_2}}{V_{11} + V_{22} - 2V_{12}}$$

where $\overline{X_j} = \frac{1}{n}\sum_{i=1}^{n} x_{ji}$.
To interpret this, note that $V_{22} - V_{12}$ is the variance of the residual from a regression of X_{2i} on X_{1i}. In other words, this estimator is an inverse residual variance weighted average of the two means.

b) (Random effects) Now assume for simplicity $V[X_i] = \mathbf{I}$. Also assume $\mu_i \sim N(\mu, \sigma^2)$, and assume for simplicity μ and σ^2 are known. What is $\hat{\mu}_{i,RE} \equiv \mathbf{E}[\mu_i|X_i]$? Compare this estimator to $\hat{\mu} = \overline{X}_i \equiv \frac{x_{1i}+x_{2i}}{2}$.

$$\begin{aligned}\mathbf{E}[\mu_i|X_i] &= \int \mu_i p(\mu_i|X_i)d\mu_i \\ &= \int \mu_i \left[p(X_i|\mu_i)p(\mu_i)/p(X_i)\right] d\mu_i\end{aligned}$$

To simplify this, note that with some algebra, we can rewrite $p(X_i|\mu_i)p(\mu_i)/p(X_i) = p(\overline{X}_i|\mu_i)p(\mu_i)/p(\overline{X}_i)$, where $\overline{X}_i = \frac{x_{1i}+x_{2i}}{2}$. We can then calculate each of these terms. $p(\overline{X}_i|\mu_i) = \sqrt{2}\phi(\sqrt{2}(\overline{X}_i - \mu_i))$, $p(\mu_i) = \frac{1}{\sigma}\phi((\mu_i - \mu)/\sigma)$, and $p(\overline{X}_i) = \frac{1}{\sqrt{\sigma^2+1/2}}\phi\left(\frac{\overline{X}_i-\mu}{\sqrt{\sigma^2+1/2}}\right)$. Substituting these and further algebra yields

$$\mathbf{E}[\mu_i|X_i] = \frac{1/\sigma^2}{2+1/\sigma^2}\mu + \frac{2}{2+1/\sigma^2}\overline{X}_i$$

which is a weighted average of the population mean, μ, and the individual's mean observed income, $\overline{X}_i$. To interpret this, consider what happens when σ^2 gets large. When σ^2 is large, it means the population mean income carries relatively little information about an individual's income compared to noisy observations about that individual's income. In this case, all the weight gets put on $\overline{X}_i$.

2) (Overidentified linear IV) Consider the linear instrumental variable model from class. $W_i \equiv \{y_i, x_i, z_i\}$ are independently and identically distributed, where y_i and x_i are scalars and z_i is a 2x1 vector, and we impose the moment restriction $\mathbf{E}[z_i(y_i - x_i\beta)] = 0$, and assume that y_i, x_i, and z_i are all mean 0.

a) Show that the GMM estimator using the weighting matrix $W = \hat{V}[z_i]^{-1}$ is just $\hat{\beta}_{2SLS} \equiv \frac{\text{Cov}[\hat{x}_i, y_i]}{V[\hat{x}_i]}$, where $\hat{x}_i = z_i^T V[z_i]^{-1}\text{Cov}[z_i, x_i]$, predicted x_i from a regression of x_i on z_i. Justify this choice of weights.
Hint: See your answer to 1) for the justification.
Note: This is called "Mahalanobis distance", which may seem weird but makes sense when we think of GMM as a minimum distance estimator, with the choice of weights allowing us to calculate the distance between the moments and 0.

First, let $m(W_i, \beta) = z_i(y_i - x_i\beta)$, and let $M(W_i, \beta) \equiv D_\beta m(X_i, \beta) = z_i x_i$. As seen previously, we can write $\hat{\beta}_{GMM} = \arg\max_b m(b)^T W m(b)$, where $m(b) = \frac{1}{n}\sum_{i=1}^n m(W_i, b)$. The first order condition of this maximization problem yields

$M(\hat{\beta}_{GMM})Wm(\hat{\beta}_{GMM}) = 0$, where $M(b) = \frac{1}{n}\sum_{i=1}^{n} M(W_i, b)$. Substituting yields

$$\text{Cov}[z_i, x_i]^T W \left(\text{Cov}[z_i, y_i] - \text{Cov}[z_i, x_i]\hat{\beta}_{GMM}\right) = 0$$
$$\hat{\beta}_{GMM} = \left(\text{Cov}[z_i, x_i]^T W \text{Cov}[z_i, x_i]\right)^{-1} \text{Cov}[z_i, x_i]^T W \text{Cov}[z_i, y_i]$$

Now, suppose $W = V[z_i]^{-1}$. Now, note that $W = V[z_i]^{-1}\mathbf{E}[z_i z_i^T]V[z_i]^{-1}$. Substituting this into our formula for $\hat{\beta}_{GMM}$ yields $\hat{\beta}_{GMM} = \hat{\beta}_{2SLS}$. Note that this inverse variance weighting of the moments is equivalent to what we saw in 1): we're weighting the moments proportionally to their residual information. Since the moments correspond to z_i, we can also think of this as normalizing z_i such that they're independent with unit variance.

b) Show how to construct the efficient GMM estimator, using $\hat{\beta}_{2SLS}$ as the first step estimator. Interpret the weighting matrix used here. Compare it to the weighting matrix in part a).

First, we would estimate $\hat{\beta}_{2SLS}$. Second, we would estimate $W_e = \mathbf{E}[m(W_i, \beta)m(W_i, \beta)^{\text{T}}] = \mathbf{E}[z_i z_i^T \epsilon_i^2]$ using $\hat{W}_e = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i\hat{\beta}_{2SLS})^2 z_i z_i^{\text{T}}$. Finally, we would estimate

$$\hat{\beta}_{GMM,e} = [\text{Cov}[z_i, x_i]^{\text{T}}\hat{W}_e\text{Cov}[z_i, x_i]]^{-1}\text{Cov}[z_i, x_i]^{\text{T}}\hat{W}_e\text{Cov}[z_i, y_i]$$

To interpret the optimal weighting matrix, note that

$$\mathbf{E}[z_i z_i^T \epsilon_i^2] = \mathbf{E}[\epsilon_i^2]\begin{pmatrix} \mathbf{E}[z_{1i}^2\epsilon_i^2]/\mathbf{E}[\epsilon_i^2] & \mathbf{E}[z_{1i}z_{2i}\epsilon_i^2]/\mathbf{E}[\epsilon_i^2] \\ \mathbf{E}[z_{1i}z_{2i}\epsilon_i^2]/\mathbf{E}[\epsilon_i^2] & \mathbf{E}[z_{2i}^2\epsilon_i^2]/\mathbf{E}[\epsilon_i^2] \end{pmatrix}$$

which you can contrast with

$$\mathbf{E}[z_i z_i^T] = \begin{pmatrix} \mathbf{E}[z_{1i}^2] & \mathbf{E}[z_{1i}z_{2i}] \\ \mathbf{E}[z_{1i}z_{2i}] & \mathbf{E}[z_{2i}^2] \end{pmatrix}$$

The two are different in that the optimal weighting matrix is just a weighted variance matrix of the z_i, where the weights used are ϵ_i^2. In other words, we'd put relatively more weight on a z_{ji} that varies relatively less when ϵ_i^2 are large.