1 GMM: General Formulas and Applications

1.1 Using GMM for Regressions

- Now, we're going to talk about how mapping things into GMM allows us to develop an asymptotic distribution that corrects for statistical problems like
  - Serial correlation
  - Conditional heteroskedasticity
- This is very useful for when we use OLS
  - We can basically 'map' OLS into the GMM framework, and this is what allows us to (easily) come up with corrections for OLS standard errors
  - Remember that OLS is consistent even if the errors do not obey OLS assumptions, but the standard errors need to be corrected. (What does consistent mean?)

1.2 OLS Review

- Recall: OLS finds parameters β to
$$\min_{\beta} E_T\left[\left(y_t - \beta' x_t\right)^2\right]$$
- How do we do this? We find β from the FOC
$$E_T\left[x_t\left(y_t - x_t'\beta\right)\right] = 0$$
- What does this say?
  - This FOC just states that the residuals, $\varepsilon_t = y_t - \beta' x_t$ are orthogonal to the x_t variables, $E_T\left[x_t \varepsilon_t\right] = 0$
- But see that this expression is just the 'moment condition' that is used for GMM
$$g_T(\beta) = E_T\left[x_t\left(y_t - x_t'\beta\right)\right] = 0$$

1.3 OLS through GMM

- With OLS, what 'case' do we have vis-a-vis GMM?
  - We're in case 1, since the dimension of β that we are trying to estimate is the same as the number of right-hand-side variables we have.
  - The number of moments equals the number of parameters.
  - We can set the sample moment conditions exactly equal to zero
  - We don't need a weighting matrix (i.e., $a = I$).
- The OLS formula is just
$$\widehat{\beta} = \left[E_T\left(x_t x_t'\right)\right]^{-1} E_T\left(x_t y_t\right)$$

- For OLS, the moment condition is given by

$$f(x_t, \beta) = x_t \left(y_t - x_t^{'} \beta \right) = x_t \varepsilon_t$$

with sample analogue

$$g_T(\beta) = E_T \left[x_t \left(y_t - x_t^{'} \beta \right) \right] = E_T \left[x_t \varepsilon_t \right].$$

- Now, recall that the asymptotic distribution for $\widehat{\beta}$ is given by

$$\widehat{\beta}_{GMM} \overset{a}{\sim} N \left[\beta, \frac{1}{T} \left(d \widehat{S}^{-1} d' \right)^{-1} \right]$$

  - where d is just the derivative of the moment condition w.r.t. β, $d = \frac{\partial g_T(x_t, \beta)}{\partial \beta}$, and $\widehat{S} = \Sigma_{j=-\infty}^{\infty} E \left[f(x_t, \beta) f(x_{t-j}, \beta)' \right]$

- So from

$$\begin{aligned} d &= -E \left(x_t x_t^{'} \right) \\ f(x_t, \beta) &= x_t \varepsilon_t, \end{aligned}$$

- GMM gives us a formula for OLS standard errors that can be written as

$$var \left(\widehat{\beta} \right) = \frac{1}{T} E \left(x_t x_t^{'} \right)^{-1} \left[\Sigma_{j=-\infty}^{\infty} E \left(\varepsilon_t x_t x_{t-j}^{'} \varepsilon_{t-j} \right) \right] E \left(x_t x_t^{'} \right)^{-1}$$

- Now, we have several cases where this formula can be applied...

1.4 OLS: Serially Uncorrelated, Homoskedastic Errors

- (Usual) OLS Assumptions:

  - No serial correlation:

$$E \left[\varepsilon_t \mid x_t, x_{t-1}, ..., \varepsilon_{t-1}, \varepsilon_{t-2}, ... \right] = 0$$

  - Homoskedasticity (same finite variance):

$$E \left[\varepsilon_t^2 \mid x_t, x_{t-1}, ..., \varepsilon_{t-1}, \varepsilon_{t-2}, ... \right] = \sigma_\varepsilon^2$$

- How do these assumptions allow us to simplify the expression

$$var \left(\widehat{\beta} \right) = \frac{1}{T} E \left(x_t x_t^{'} \right)^{-1} \left[\Sigma_{j=-\infty}^{\infty} E \left(\varepsilon_t x_t x_{t-j}^{'} \varepsilon_{t-j} \right) \right] E \left(x_t x_t^{'} \right)^{-1} ?$$

- No serial correlation: $E \left[\varepsilon_t \mid x_t, x_{t-1}, ..., \varepsilon_{t-1}, \varepsilon_{t-2}, ... \right] = 0$

  - Means that for

$$\left[\Sigma_{j=-\infty}^{\infty} E \left(\varepsilon_t x_t x_{t-j}^{'} \varepsilon_{t-j} \right) \right],$$

all terms with $j \neq 0$ fall out, so

$$\left[\Sigma_{j=-\infty}^{\infty} E \left(\varepsilon_t x_t x_{t-j}^{'} \varepsilon_{t-j} \right) \right] = E \left(\varepsilon_t^2 x_t x_t' \right)$$

- Homoskedasticity: $E\left[\varepsilon_t^2 \mid x_t, x_{t-1}, ..., \varepsilon_{t-1}, \varepsilon_{t-2}, ...\right] = \sigma_\varepsilon^2$
  - Means
$$E\left(\varepsilon_t^2 x_t x_t'\right) = \sigma_\varepsilon^2 E\left(x_t x_t'\right)$$
- So, together we get
$$\begin{aligned} var\left(\widehat{\beta}\right) &= \frac{1}{T} E\left(x_t x_t^{'}\right)^{-1} \left[\Sigma_{j=-\infty}^{\infty} E\left(\varepsilon_t x_t x_{t-j}^{'} \varepsilon_{t-j}\right)\right] E\left(x_t x_t^{'}\right)^{-1} \\ &= \frac{1}{T} \sigma_\varepsilon^2 E\left(x_t x_t'\right)^{-1} \end{aligned}$$

1.5 OLS: Heteroskedastic Errors

- What if we think we have heteroskedasticity?
  - It means we can't pull the ε_t^2 out of the expectation because it's dependent on time
$$E\left(\varepsilon_t^2 x_t x_t'\right) \neq \sigma_\varepsilon^2 E\left(x_t x_t'\right)$$
- In this case if we use
$$\begin{aligned} var\left(\widehat{\beta}\right) &= \frac{1}{T} E\left(x_t x_t^{'}\right)^{-1} \left[\Sigma_{j=-\infty}^{\infty} E\left(\varepsilon_t x_t x_{t-j}^{'} \varepsilon_{t-j}\right)\right] E\left(x_t x_t^{'}\right)^{-1} \\ &= \frac{1}{T} E\left(x_t x_t^{'}\right)^{-1} E\left(\varepsilon_t^2 x_t x_t'\right) E\left(x_t x_t^{'}\right)^{-1}, \end{aligned}$$
  we will get estimates for standard errors that are heteroskedasticity consistent.
  - These are commonly called "White standard errors"

1.6 Model Comparisons

1.7 Recall, the J-Test of Model Fit

- Recall that we have a J_T test as a *test of overidentifying restrictions*
$$TJ_T = T\left[g_T\left(\widehat{b}_{GMM}\right)' S^{-1} g_T\left(\widehat{b}_{GMM}\right)\right] \sim \chi^2\left(\#\text{ m - }\#\text{ p}\right)$$
- J_T Test: If the model is true, how often should we see a weighted sum of squared pricing errors as big as what we got?
  - If the answer is "not too often", then the model is rejected.
- But what if we want to compare one model to another? ...
- Or test how well a model does on a chosen set of pricing errors/moments/portfolios? ...

1.8 Testing Moments

- What if we want to test how well a model performs on a particular set of moments or a particular set of pricing errors?
  - E.g., we know that the unconditional CAPM performs poorly in pricing small firms
  - If you make a portfolio of the smallest decile of firms in the NYSE, the model $m = a + bR^w$ doesn't price them well - this is the "small firm effect"
  - Can a new model do better?
- How do we test this?
  - We know that g_T measures pricing errors, so we can use the asymptotic distribution for the moment conditions to find the standard errors we need.
    - Hansen's Lemma gives the sampling distribution of the moments, it is equation (11.5) in your book.
  - We can then construct a t-test to evaluate how well our new model performs on pricing small firms
  - (To test a groups of g_T, we would just use a χ^2 test.)
- But there is an easier way...
- We can use a χ^2 difference test...
- We can estimate our full model with all the moment conditions and find S.
- Next, we can set to zero the moment(s) we want to test, and re-estimate the model using the same weighting matrix S

$$\min_{b_r} g_{rT}(b_r)' S^{-1} g_{rT}(b_r)$$

  - Note $g_{rT}(b_r)$ is the vector of moment conditions with the moments we want to test set to zero (it has less moment conditions to estimate now)
- Here, we are estimating the same number of parameters, but with fewer moments
  - So $g_{rT}(b_r)' S^{-1} g_{rT}(b_r)$ will be lower than $g_T(b)' S^{-1} g_T(b)$
  - But it shouldn't be that much lower...
- But the idea is that if the moments are really zero (i.e., small pricing errors on those moments), it shouldn't be that much lower

$$T g_T(b)' S^{-1} g_T(b) - T g_{rT}(b_r)' S^{-1} g_{rT}(b_r) \sim \chi^2(r)$$

where

$$r = \#\text{ of eliminated moments}$$

1.9 Model Comparisons

- If one model can be expressed as a 'restricted' version of another, or 'unrestricted', version, we can also perform a chi-squared difference test

$$TJ_T\left(\text{restricted}\right) - TJ_T\left(\text{unrestricted}\right) \sim \chi^2\left(\#\text{ of restrictions}\right)$$

- If we use the S matrix of the unrestricted model to estimate the restricted J_T, it should rise...
  - But if the restricted model is true, it shouldn't rise by much.

1.10 General GMM Formulas

- Note that - to be even more general - our GMM estimates pick $\widehat{b}$ so that

$$a_T g_T\left(\widehat{b}\right) = 0$$

where

$$g_T\left(\widehat{b}\right) = \frac{1}{T}\sum_{t=1}^{T} f\left(x_t, b\right)$$

- The matrix a_T is just a matrix that defines which linear combinations of $g_T\left(b\right)$ will be set to zero.
  - If # parameters = # moment conditions, we can set each condition exactly to zero.
  - If # parameters < # moment conditions, we will only set some moments, or some linear combinations of moments, to zero.

1.10.1 GMM Formulas - Special Case

- The simplest form of the GMM estimates we have been working with are actually just a special case.
- We have been estimating b by

$$\min g_T\left(b\right)' W g_T\left(b\right)$$

- The FOCs of this problem are

$$\frac{\partial g_T'}{\partial b} W g_T\left(b\right) = 0$$

- So - in a more general context - for the problem $a_T g_T\left(\widehat{b}\right) = 0$, we have that

$$a_T = \frac{\partial g_T'}{\partial b} W$$

1.10.2 GMM Formulas - Efficient Estimates

- Another example: If instead we estimate b by

$$\min g_T\left(b\right)' S^{-1} g_T\left(b\right)$$

- The FOCs of this problem are

$$\frac{\partial g_T'}{\partial b} S^{-1} g_T\left(b\right) = 0$$

- So, we have that
$$a_T = \frac{\partial g_T'}{\partial b} S^{-1}$$
- Note that if we use
$$a_T g_T\left(\widehat{b}\right) = 0,$$

Our asymptotic distribution changes, so that now

$$var\left(\widehat{b}\right) = \frac{1}{T}\left(ad\right)^{-1} aSa'\left(ad\right)^{-1}$$

- If
$$\begin{aligned} a_T &= \frac{\partial g_T'}{\partial b} S^{-1} \\ &= d'S^{-1} \end{aligned}$$

We get the reduced expression that

$$\sqrt{T}\left(\widehat{b} - b\right) \to N\left[0, \left(d'S^{-1}d\right)^{-1}\right]$$

1.11 Pre-specifications

- Why might we not want to use Hansen's "efficient" estimates?
  - Perhaps if we think that particular assets suffer from measurement error or are small and illiquid. In this case, we would want to de-emphasize them.
  - Changing the value of W_{ii} in the problem
$$\min g_T\left(b\right)' W g_T\left(b\right)$$
would do this.
  - Higher values for W_{ii} force GMM to pay more attention to getting those moments estimates correctly, and vice-versa.
- We can also specify which linear combinations of moment conditions will be set to zero by changing the a_T matrix.
- Say we have two moment conditions $g_T = \left[\begin{array}{cc} g_T^1 & g_T^2 \end{array}\right]$ and use $W = I$
  - If we want GMM to pay equal attention to the two, we can set
$$a = \left[\begin{array}{cc} 1 & 1 \end{array}\right]$$
  - On the other hand, if we want GMM to pay more attention to the second one, we can set
$$a = \left[\begin{array}{cc} 1 & 10 \end{array}\right]$$
- (Note that S still shows up in the standard errors and test statistics!)

1.12 Motivations

1.12.1 Motivations - I.e., Why Do We Care?

- Robustness.
- First-stage GMM or an otherwise fixed weighting matrix GMM estimates are consistent, even if they give up something in asymptotic efficiency.
  - These estimates can also be more robust to statistical and economic issues
  - (Just like OLS estimates)
- "Efficient" estimates may not be robust if there are misspecification issues.
  - (Just like GLS estimates. Though GLS can improve efficiency, the estimates can be worse if the error covariance matrix is modelled incorrectly because GLS can focus on the misspecified areas of the model.)
- What to do?

1.12.2 Robustness

- What to do for robustness checks of our efficient estimates?
- First, calculate first-stage estimates, standard errors, and model fit tests.
- If our model is specified correctly, parameter estimates from our "efficient" procedure should not change by much and their standard errors should be tighter.
- If the results do differ, you should figure out why.
  - Are the differences due to efficiency gain?
  - Is there a model misspecification?

1.12.3 Efficient GMM

- Efficient GMM wants to focus on well-measured moments.
  - I.e., portfolios with small values of $var\left(m_{t+1}R^e_{t+1}\right)$ roughly correlate to those with small values of return variance.
- Which asset is this? The sample minimum variance portfolio.
  - And GMM's evaluation of model fit will be largely based on its ability to price this portfolio
- Since the sample mean-variance frontier is wider than the true (ex ante) mean-variance frontier, we often want to force GMM to not pay as much attention to correctly pricing what we would consider to be a 'spurious' minimum variance portfolio.

1.12.4 Economic Interest

- The optimal GMM weighting matrix focuses on linear combinations of moments with small sampling errors in both estimation and evaluation.
  - Do we care how well out model might price 90 small firms − 64 large firms − 29 medium firms? Probably not - this is an uninteresting economic moment.
- Instead we want to measure how our model performs on portfolios formed on economically interesting characteristics like size, beta, book/market, industry, etc.
  - We can force GMM to pay attention to such economically interesting moments instead.

1.12.5 What is GMM doing???

- Suppose you're just using Hansen's efficient estimator

$$\min g_T(b)' S^{-1} g_T(b)$$

- How do you know which moments (and combinations of moments) GMM is focusing on?
- Well, we can factor S^{-1} into a "square root" through a Choleski decomposition, so that where $S^{-1} = C'C$ to get

$$\min \left[g_T(b)' C'\right] \left[C g_T(b)\right]$$

- With $Cg_T(b)$, we can actually see what GMM is trying to minimize.

1.13 Estimating the Spectral Density Matrix

- Recall that the optimal weighting matrix depends on *population* moments

$$\begin{aligned} S &= \sum_{j=-\infty}^{\infty} E(u_t u_{t-j}) \\ u_t &= m_t(b) x_t - p_t \end{aligned}$$

- We estimate population moments by sample counterparts, but we obviously do not have a direct sample counterpart to the spectral density matrix.
- If $T = 100$, how many data points do you have to estimate $E\left(u_t u_{t+99}'\right)$?
  - One, $u_1 u_{100}'$
- For

$$S = \sum_{j=-\infty}^{\infty} E(u_t u_{t-j}),$$

an estimator that uses all possible autocorrelations is therefore inconsistent.

- One solution is to construct consistent estimates that downweight higher-order autocorrelations.
  - We've already seen one way to include only k autocorrelations, and to downweight the higher-order ones. (Newey-West)

1.13.1 Other Variations on Estimating S

- Remove means from u_t. Even though $E(u_t) = 0$, Hansen and Singleton (1982) advocate for it.
- Use a good first-stage W, or transform the data.
  - E.g., if two returns are highly correlated we can use R^1 and $R^2 - R^1$ instead of just R^1 and R^2
  - A Choleski decomposition $W = C'C$ can help us figure out the right W for $\min(g_T(b)' C')(Cg_T(b)$
- We can use an iterated procedure (rather than the two-step procedure) to construct S - we've already talked about this one!
- Parametric instead of non-parametric specifications of S are also valid and sometimes easier to estimate
  - E.g., modelling u as an $AR(1)$ allows us to write $S = \sum_{j=-\infty}^{\infty} E(u_t u_{t-j}) = \sigma_u^2 \sum_{j=-\infty}^{\infty} \rho^{|j|} = \sigma_u^2 \frac{1+\rho}{1-\rho}$
- Cochrane details these and some others.