17 GMM Notes

Note: These notes are slightly amplified versions of the blackboards for the GMM lectures in my Coursera Asset Pricing class. They are not meant as substitutes for the readings, and are incomplete and may have typos.

17.1 Standard error of the mean

Example. A *sample* might be

$$u_t = \begin{bmatrix} 5 \\ 4 \\ 8 \\ 8 \\ 2 \end{bmatrix}.$$

The *sample mean* is

$$E_T(u) = \frac{1}{T}\sum_{t=1}^{T} u_t = \bar{u} = 5.4$$

The sample mean is itself is a random variable. What if we sample again,

Sample #:	1	2	3	4
	5	3	5	8
	4	7	10	3
	8	7	4	6
	8	2	6	7
	2	2	3	9
$E_T(u)$	5.4	4.2	5.6	6.6

and so on. If we rerun the experiment, the *sample mean* varies.

The sample mean is a random variable. We only see one sample, but what if we could run the world over and over again? The art of statistics: from only first sample, guess what variability of sample mean is across all the samples – the distribution of the bottom row.

How do statisticians achieve such magic?

$$\bar{u} = E_t(u) = \frac{1}{T}\sum_{t=1}^{T} u_t$$

$$var(\bar{u}) = var\left(\frac{1}{T}\sum_{t=1}^{T} u_t\right) = \frac{1}{T^2} var\left(\sum_{t=1}^{T} u_t\right)$$

If 1) u_t is stationary so $var(u_1) = var(u_2)$..and 2) if $cov(u_t, u_{t+j}) = 0$, then we can find *the mean and variance of the sample mean*

$$var(\bar{u}) = \frac{1}{T^2}\left(\sum_{t=1}^{T} var\left(u_t\right)\right) = \frac{var(u_t)}{T}$$
$$\sigma(\bar{u}) = \frac{\sigma(u)}{\sqrt{T}}.$$

When you think about it, this is just amazing. With a few simple assumptions, and looking look only at our world, the first column, in which the sample mean is one number, we can learn how that result would come out differently over all possible rerunnings of history or parallel universes.

Now, what if the variables u are *not* uncorrelated over time? Again "stationary" is enough. (Don't confuse stationary with uncorrelated with i.i.d. They are different!) Then, we just include all the cross terms.

$$var(\bar{u}) = \frac{1}{T^2} var\left(\sum_{t=1}^{T} u_t\right) = \frac{1}{T}\sum_{j=-T}^{T}\frac{T-|j|}{T} cov(u_t, u_{t-j}) \tag{1}$$
$$= \frac{1}{T}\left[var(u_t) + 2\sum_{j=1}^{T}\frac{T-|j|}{T} cov(u_t, u_{t-j})\right]$$

As T gets big,

$$var(\bar{u}) \to \frac{1}{T}\sum_{j=-\infty}^{\infty} cov(u_t, u_{t-j}) = \frac{1}{T}\left[var(u_t) + 2\sum_{j=1}^{\infty} cov(u_t, u_{t-j})\right] = \frac{S}{T} \tag{2}$$

The last equality is a definition of S.

Intuition: If the covariances are big, then we don't really have as many data points as we think we do, so the sample mean is not as precise.

Not only have we found the mean and variance, but sums of random variables approach a normal distribution, even if the original random variables are not normal. Thus, we have as $T \to \infty$, $\bar{u} \to N[E(u), var(\bar{u})]$.

Use: we can use this formula to quantify the precision of a measurement, the standard error. We can use it to test – is the mean truly zero, or some other number, but we saw a larger number just by chance?

You just learned how to "compute standard error of mean for autocorrelated time series."

An example / application /useful formula: Suppose we model the autocorrelation of the u as an AR(1), $u_t = \rho u_{t-1} + \varepsilon_t$. Then

$$\sigma^2\left(\frac{1}{T}\sum_{t=1}^{T} u_t\right) \to \frac{1}{T}\sigma_u^2 \sum_{j=-\infty}^{\infty} \rho^{\|j\|} = \frac{1}{T}\sigma_u^2\left(\frac{1+\rho}{1-\rho}\right). \tag{3}$$

This is a really nice generalization of σ^2/T.

(Note: (3) is called a "parametric correction" and (2) a "nonparametric correction" for serial correlation. As you can guess, the parametric one works better in small samples if the process is close to AR(1), because you don't estimate so many extra parameters. We will come back to this issue in "Estimating the S matrix." The finite-sample version, using (1) is $var(\bar{u}) = \frac{1}{T}\left[\frac{(1+\rho)}{(1-\rho)} - \frac{2}{T}\frac{\rho\left(1-\rho^T\right)}{(1-\rho)^2}\right]$.)

The Amazing Fact behind GMM: Just about anything you could want to do in econometrics boils down to generalized version of same idea.

17.2 GMM Notation / Formula Summary

1. Model moments, true value b_0

$$g(b_0) = E\left[f(x_t, b_0)\right] = E\left[u_t\right] = E\left[(m_{t+1}(b_0)x_{t+1} - p_t\right] = 0$$

2. Sample moments:

$$g_T(b) = E_T\left[f(x_t, b)\right] = E_T\left[(m_{t+1}(b)x_{t+1} - p_t\right];\ E_T\left(\cdot\right) \equiv \frac{1}{T}\sum_{t=1}^{T}\left(\cdot\right)$$

3. GMM estimator $\hat{b}$ or b_T

$$a_T g_T(\hat{b}) = 0$$

4. Standard errors

$$\sqrt{T}(\hat{b} - b_0) \to N(0, \left(ad\right)^{-1}aSa'(ad)^{-1\prime})$$

where

$$\begin{aligned} d &= \frac{\partial g(b)}{\partial b'} \\ a &= plim\ a_T \\ S &= \sum_{j=-\infty}^{\infty} E\left[f(x_t, b_0)f(x_{t-j}, b_0)'\right] = \sum_{j=-\infty}^{\infty} E\left[u_t u_{t-j}'\right] \end{aligned}$$

5. Variance of moments:

$$\sqrt{T}g_T(\hat{b}) \to N(0, \left[I - d(ad)^{-1}a\right] S \left[I - d(ad)^{-1}a\right]').$$

$$\begin{aligned} \text{var}\left[g_T(b_0)\right] &= \frac{1}{T}S; \\ \text{var}\left[g_T(\hat{b})\right] &= \frac{1}{T}\left[I - d(ad)^{-1}a\right] S \left[I - d(ad)^{-1}a\right]' \end{aligned}$$

This formula can be used to test individual moments, or for χ^2 tests for joint significance. In particular, g_T

$$g_T' var(g_T)^{+} g_T \sim \chi^2_{N-K}$$

6. Efficient GMM. Like GLS. Use any efficient first stage to estimate S, then use

$$a = d'S^{-1}$$

with this choice,

$$\begin{aligned} var(\hat{b}) &= \frac{1}{T}(d'S^{-1}d)^{-1} \\ cov(g_T) &= \frac{1}{T}\left[S - d(d'S^{-1}d)^{-1}d'\right] \\ g_T' cov(g_T)^{+} g_T &= T g_T' S^{-1} g_T = T J_T \sim \chi^2_{\#mom-\#par} \end{aligned}$$

7. Minimization approach

$$\hat{b}_1 = \min_b g_T(b)' W g_T(b)$$

This is a choice of a_T :

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

Second stage/efficient W choice

$$\hat{b}_2 = \min_b g_T(b)' S^{-1} g_T(b)$$

First order condition: this is the efficient GMM estimator

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

Then, we can write the second stage test in this case as

$$TJ_T = T \min_{\{b\}} \left[g_T(b)' S^{-1} g_T(b) \right] \sim \chi^2_{\#mom-\#par}$$

17.3 GMM defined; standard errors and tests.

The key to GMM is simply mapping a problem into the GMM notation.

17.3.1 Population and sample moments

We write the model in a form in which it predicts that the true mean of some function of the data is zero. Typically, the model includes some free parameters. Then we choose the free parameters to make sample means close to zero, and we evaluate the model by how close the sample mean got to zero. We do standard errors and tests by exploiting the standard error of the mean idea.

Vector of moments

In asset pricing, a typical example is

$$\begin{aligned} 0 &= E(m_{t+1}(b)R^e_{t+1}) \\ 0 &= E(m_{t+1}(b)R_{t+1} - 1) \end{aligned}$$

or more generally,

$$\begin{aligned} p_t &= E_t(m_{t+1}(b)x_{t+1}) \Rightarrow \\ 0 &= E(m_{t+1}(b)x_{t+1} - p_t) \end{aligned}$$

Here we have expressed the asset pricing model by its prediction that the mean of some function of the data should be zero. There are some unknown parameters b which we will estimate.

Examples:

$$\begin{aligned} &\text{CAPM: } 0 = E\left[(a - bR^m_{t+1})R^e_{t+1}\right] \quad \{b = a, b\} \\ &\text{CCAPM: } \ 0 = E\left[\beta\left(\frac{c_{t+1}}{c_t}\right)^{-\gamma} R^e_{t+1}\right] \quad \{b = \beta, \gamma\} \end{aligned}$$

Here R^e_{t+1} is a *vector*, e.g. the 25 FF portfolios.

Those returns might even be multiplied by some instruments / managed portfolios, to make a bigger vector. With one instrument, z_t.

$$\text{Conditional} : 0 = E\left[\beta\left(\frac{c_{t+1}}{c_t}\right)^{-\gamma} \begin{bmatrix} R^e_{t+1} \\ z_t R^e_{t+1} \end{bmatrix}\right] \{b = \beta, \gamma\}$$

$$\text{Conditional} : 0 = E\left[\beta\left(\frac{c_{t+1}}{c_t}\right)^{-\gamma} \left(R^e_{t+1} \otimes \begin{bmatrix} 1 \\ z_t \end{bmatrix}\right)\right] \{b = \beta, \gamma\}$$

The second equality introduces the Kronecker product notation you often see in GMM and econometrics. If R^e is a 25×1 vector of returns, now we have a 50×1 vector of moments.

Alphas

Note:

$$E(mR^{ei}) = E(m)E(R^{ei}) + cov(m, R^{ei}) = \frac{1}{R^f}\left[E(R^{ei}) - \beta_i\lambda\right] = \frac{1}{R^f}\alpha_i$$

Thus, GMM *moments* are *alphas*. The model prediction is the same thing as $\alpha = 0$. We are trying to pick parameters by making alphas as small as possible, and then we will evaluate the model by just how small the alphas are. (Mean vs. beta graph)

Sample moments

These are *population moments*. To do empirical work, just look at *sample* counterpart,

$$E_T\left[m_{t+1}(b)R^e_{t+1}\right]$$

$$E_T \equiv \frac{1}{T}\sum_{t=1}^{T}$$

Estimation and testing strategy

So our strategy will be:

1) *Estimate* Pick $\hat{b}$ (b_T) to make the model look as good as possible – make the moments as close to zero as possible, make the alphas as small as possible, make the cross-sectional line as good as possible, give the model its best chance.

2) *Measure statistical uncertainty* We will calculate $\sigma(\hat{b})$ standard errors.

3) *Evaluate and test* the model by how close it can get to $\alpha = 0$. Are sample $\alpha \neq 0$ just due to luck?.(accounting for parameters, e.g. 25 parameters to fit 25 moments doesn't count!) .

Notation:

Now map all this into the GMM notation. The true, or population moments are

$$g(b) = E\left[f(x_t, b)\right] = E\left[u_t\right] = E\left[m_{t+1}(b)R^e_{t+1}\right] = 0.$$

The = just are many versions of the same thing.

The sample moments just replace E_T the sample mean in place of population mean

$$g_T(b) = E_T\left[f(x_t, b)\right] = ...$$

17.3.2 Estimation

There are typically more moments (25 FF returns) than parameters (2 for the CCAPM, β, γ or CAPM $a - bR^{em}$). What do we do? We *set a linear combination of moments to zero*

$$a_T g_T(\hat{b}) = 0.$$

Example: with 2 parameters b, e.g. $m = a - bR^{em}$, you can choose the parameters to fit two returns (R^m, R^f) exactly. Then you can test the model by seeing if the other moments are small. (This is what Fama and French's time-series regressions do. They give zero alpha to the factors and the risk free rate, thereby estimating factor risk premiums equal to factor means, and then evaluate the model by the remaining alphas.)

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} E_T\left(m(a,b)R^{em}\right) \\ E_T\left(m(a,b)R^{f} - 1\right) \\ E_T\left(m(a,b)R^{ehml}\right) \\ E_T(m(a,b)R^{esmb}) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

This defines GMM estimate $\hat{b}$. It answers the question, *given data, how will I estimate the free parameters?* (I'm running out of letters. The GMM notation uses b to denote the vector of parameters, $b = [a\ b]'$ in this case. That's two uses of the same letter to mean different things.)

Sometimes can solve for the estimates analytically. For example, for the CAPM, $m = a - bR^{em}$.

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} E_T\left(\left(\hat{a} - \hat{b}R^{em}\right)R^{em}\right) \\ E_T\left(\left(\hat{a} - \hat{b}R^{em}\right)R^{f} - 1\right) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} \hat{a}E_T(R^{em}) - \hat{b}E_T\left(R^{em2}\right) \\ \hat{a}E_T R^f - \hat{b}E_T\left(R^{em}R^f\right) \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

$$\begin{bmatrix} E_T(R^{em}) & E_T\left(R^{em2}\right) \\ E_T(R^f) & E_T\left(R^{em}R^f\right) \end{bmatrix} \begin{bmatrix} \hat{a} \\ -\hat{b} \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

$$\begin{bmatrix} \hat{a} \\ -\hat{b} \end{bmatrix} = \begin{bmatrix} E_T(R^{em}) & E_T\left(R^{em2}\right) \\ E_T(R^f) & E_T\left(R^{em}R^f\right) \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

Sometimes, you must solve a nonlinear system of equations For example to set $ag_T(\hat{b}) = 0$ for the consumption CAPM,

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} E_T\left(\hat{\beta}\left(\frac{C_{t+1}}{C_t}\right)^{-\hat{\gamma}} R^{em}_{t+1}\right) \\ E_T\left(\hat{\beta}\left(\frac{C_{t+1}}{C_t}\right)^{-\hat{\gamma}} R^{f} - 1\right) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} E_T\left(\left(\frac{C_{t+1}}{C_t}\right)^{-\hat{\gamma}} R^{em}_{t+1}\right) \\ E_T\left(\hat{\beta}\left(\frac{C_{t+1}}{C_t}\right)^{-\hat{\gamma}} R^{f}\right) \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

Now you have to use a computer to find β, γ. (In this case it's not hard. The first moment identifies γ, and then the second identifies β, so it's just finding a zero of a single nonlinear function.)

Think about identification. There must be at least as many equations or moments as parameters. But you must also be able to solve the moment equations for the parameters! If there is no $\hat{\gamma}$ that

solves the first equation in a given sample, you're stuck, or rather you must take a different approach. This problem can be subtle!

This all seems a bit arbitrary because I haven't told you *which a matrix to use*?

We'll come back to this question a bit, but not as much as you may hope. You're used to black boxes that tell you exactly what to do. GMM isn't like that. It's a much more flexible tool. There is a "statistically optimal" a matrix, but we often use less than optimal matrices that make a lot more economic sense and are more robust. For example, a statistically optimal technique might not let you just use the market and risk free rate to estimate a and b, or β and γ, and then "test" the model on the remaining moments. It might want you to use all the moments to estimate the parameters "efficiently." GMM lets you do the simple version.

17.3.3 Distribution

After specifying how will we produce our estimates $\hat{b}$, the next econometric question is, *what are the standard errors*? If we produce numbers ($\hat{b}$) by this recipe, what are the statistical properties of $\hat{b}$ across many samples? How do we *test* the model? Are the remaining alphas statistically close to zero? How likely is it that the remaining alphas are produced just by chance from a world where the true alphas are zero?

Results

*Econometric facts:*If model is true, if data are stationary (and some other technical assumptions – read Hansen), then $\hat{b}$ is a consistent, asymptotically normal, estimate of b.

1) Standard errors.

$$\sqrt{T}(\hat{b}-b) \to N(0, (ad)^{-1}aSa'(ad)^{-1\prime})$$

$$\begin{aligned} d &= \frac{\partial g(b)}{\partial b'} \\ a &= plim\ a_T \\ S &= \sum_{j=-\infty}^{\infty} E\left[f(x_t,b)f(x_{t-j},b)'\right] = \sum_{j=-\infty}^{\infty} E\left[u_t u_{t-j}'\right] \end{aligned}$$

2) The distribution of sample moments $g_T(\hat{b})$, *after* some linear combinations $ag_T(\hat{b}) = 0$ are set to zero is

$$\sqrt{T}g_T(\hat{b}) \to N(0, \left[I - d(ad)^{-1}a\right] S \left[I - d(ad)^{-1}a\right]').$$

Since the g_T are alphas or pricing errors, this formula gives the distribution of the alphas

3) χ^2 tests for joint significance. In particular, g_T

$$g_T'(\hat{b})var(g_T(\hat{b}))^{+}g_T(\hat{b}) \sim \chi^2$$

+ is pseudo inversion. Since we set $ag_T(\hat{b}) = 0$ in every sample, some linear combinations of $g_T(\hat{b})$ have no variance, and the variance covariance matrix is singular. But, since those same linear combinations of $g_T(\hat{b})$ are also zero, this causes no problem. The degrees of freedom of the χ^2 equals the rank of the variance-covariance matrix, equals the number of moments less the rank of a, the number of g_T that are not set to zero.

(One form of pseudo inversion uses the eigenvalue decomposition. Write the eigenvalue decomposition of the symmetric Σ as $\Sigma = Q\Lambda Q'$ with $Q'Q = I$ and Λ diagonal. Then $\Sigma^+ = Q\Lambda^+ Q'$ where

$$\Lambda = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & 0 & \\ & & & 0 \end{bmatrix} ; \Lambda^+ = \begin{bmatrix} \lambda_1^{-1} & & & \\ & \lambda_2^{-1} & & \\ & & 0 & \\ & & & 0 \end{bmatrix}$$

There are numerically better versions, but this is conceptually clear.)

17.3.4 Intuition for GMM formulas

1) Why S?

$$var(g_T) = var(\frac{1}{T}\sum_{t=1}^{T} u_t) =\rightarrow \frac{1}{T}\sum_{j=-\infty}^{\infty} E(u_t u'_{t-j}) = \frac{1}{T}S$$

1/T S gives the standard error of sample moments, (ignoring estimation). Precisely, S is the sampling variance of $g_T(b)$, not $g_T(\hat{b})$. As we saw in studying the standard error of the mean, the sums of correlations just allow for autocorrelation of u_t.

In asset pricing, $E_t(m_{t+1}R^e_{t+1}) = E_t(u_{t+1}) = 0$ so $E(u_t u_{t+j}) = 0$ under the null. However, we often measure under the alternative, i.e. we allow autocorrelation in standard errors even though the null says it shouldn't be there. Sometimes, such as in running 5 year return forecasting regressions in monthly data, there is serial correlation that we need to correct for.

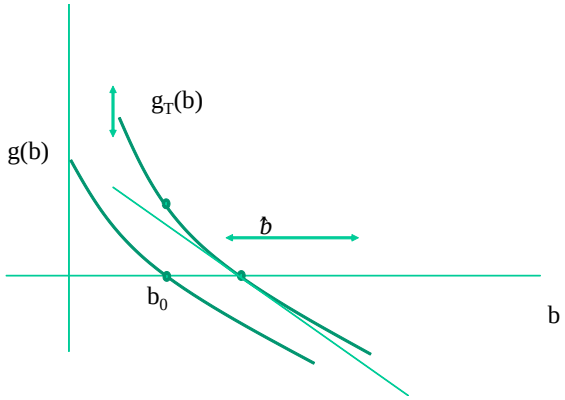
2) *Standard error formula.* This is just the delta method. For fixed b, $g_T(b)$ will vary in samples due to luck. How much does variation in g_T at each b affect variation in the estimated $\hat{b}$?

Logic: b_0 is the true value. $g(b_0)$ is zero by assumption that the model is true. $g_T(b_0)$ is not zero, because we might get a lucky draw of returns. So, look at the distribution of $\hat{b}$. Remember, b_0 is a number.

$$\begin{aligned} ag_T(\hat{b}) &= 0 \\ ag_T(b_0) + a\frac{\partial g_T}{\partial b'}(\hat{b} - b_0) &= 0 \\ (\hat{b} - b_0) &= -(ad)^{-1} ag_T(b_0) \end{aligned}$$

Thus, we have measured how variation in g_T at fixed b_0 translates into variation in $\hat{b}$ that sets sample ag_T to zero

$$var(\hat{b} - b_0) = \frac{1}{T}(ad)^{-1} aSa' (ad)^{-1\prime}$$



The picture shows you how vertical variation in $g_T(b)$ at every fixed b, due to sampling variation, translates into horizontal variation in the estimated $\hat{b}$ away from the true b_0.

In the picture, $a = 1$ (exact identification) so $\hat{b} = d^{-1}g_T(b)$ so $var(\hat{b}) = d^{-1}var(g_T(b))d^{-1\prime} = d^{-1}(S/T)d^{-1\prime}$. The actual formula adds the effect of $a \neq I$, but I'd need more dimensions to graph it.

The standard error formula uses a_T, which can vary from sample to sample, so long as it converges to some value a =plim a_T. We will often want to use data-dependent a_T matrices. ($a = d'S^{-1}$ is a prime example, as we estimate both d and S in sample.)

3) S is $cov(g_T(b))$ for *fixed* b, not $cov(g_T(\hat{b}))$ accounting for the fact that $\hat{b}$ changes in every sample too. But we pick $\hat{b}$ in each sample to set a linear combination of g_T to zero. $cov(g_T(\hat{b}))$ needs to account for that.

Example: In the exactly identified case $a = I$, the number of parameters = the number of moments, so $g_T(\hat{b}) = 0$ in every sample. Now $var(g_T(\hat{b})) = 0$, not S. That's why the formula has $S-$(lots of stuff).

"Derivation:" I'll use the delta method again.

$$ag_T(\hat{b}) = 0$$

$$g_T(\hat{b}) \approx g_T(b_0) + \frac{\partial g_T}{\partial b'}\left(\hat{b} - b_0\right) = g_T(b_0) + d\left(\hat{b} - b_0\right)$$

From above, the estimate is

$$(\hat{b} - b_0) = -(ad)^{-1} ag_T(b_0)$$

so

$$g_T(\hat{b}) \approx g_T(b_0) - d\left(ad\right)^{-1} a g_T(b_0)$$
$$g_T(\hat{b}) \approx \left(I - d\left(ad\right)^{-1} a\right) g_T(b_0)$$
$$var(g_T(\hat{b})) \approx \frac{1}{T}\left(I - d\left(ad\right)^{-1} a\right) S \left(I - d\left(ad\right)^{-1} a\right)'$$

If $a = I$, d = full rank, $cov[g_T(\hat{b})] = \frac{1}{T}\left[I - d(d)^{-1}\right] S \left[I - d(d)^{-1}\right]' = 0$. So this reduces to our intuition for the exactly identified case.

Actually, you get the same result if a and d are merely full rank (square). Again, that is the "exactly identified case" in which we're using all linear combinations of moments to identify the parameters.

$$\left[I - d(ad)^{-1}a\right] S \left[I - d(ad)^{-1}a\right] = \left[I - dd^{-1}a^{-1}a\right] S \left[I - dd^{-1}a^{-1}a\right] = 0$$

4) The χ^2 test is basically

$$\alpha' cov(\alpha)^{+} \alpha \sim \chi^2$$

This is *the* test of any asset pricing model – are the alphas jointly zero, accounting for the linear combinations of alphas that you set to zero in order to estimate free parameters? This is very much in the spirit of the "Gibbons Ross Shanken Test" if you know what that is.

17.4 Efficient GMM

So far it has been up to you to choose the a_T matrix. You have N moments $g_T(b)$ and K parameters in b. Which K linear combinations of the $g_T(b)$ do you want to set to zero?

Answer 1: Judgement! "Choose the economically important moments" (above).

Answer 2: The "statistically efficient" answer: use

$$a = d' S^{-1}.$$

If you plug $a = d'S^{-1}$ into the standard error formulas they simplify to[2]

$$var(\hat{b}) = \frac{1}{T}(d'S^{-1}d)^{-1}$$
$$cov(g_T) = \frac{1}{T}\left[S - d(d'S^{-1}d)^{-1}d'\right]$$

As before, the second term in the second equation corrects for loss of degrees of freedom due to parameter estimation.

Be careful. These formulas are often quoted as "the" GMM standard error formulas. But they only work for efficient GMM $a = d'S^{-1}$. For other a matrices, use the previous formulas.

We can use the $cov(g_T)$ formula to test whether all the moments are zero, i.e. whether the observed nonzero g_T are just due to sampling error,

$$g_T' cov(g_T)^+ g_T = T g_T' S^{-1} g_T = T J_T \sim \chi^2_{\#mom-\#par}.$$

This is the "JT test of overidentifying restrictions" The cool fact in this formula is that you can leave out the $d(d'S^{-1}d)^{-1}d'$ and get the same answer

Why are these "efficient?"

Fact 2: If $\hat{b}$ is the efficient GMM estimate, then $var(\hat{b})$ is "smallest" standard error among all linear combinations of given moments. (If you use any other a, the covariance matrix of $\hat{b}$ is this plus a positive semidefinite matrix.)

Thus: the statistical answer to picking the a matrix:

1) Choose any (reasonable!) a, to get a first stage estimate $\hat{b}_1$

2) Form an estimate of S

3) Use $a = d'S^{-1}$ in second stage.

This is just like "Use OLS first-stage, find the error covariance matrix, use GLS in a second stage."

And it's just as dangerous! A great advantage of GMM (vs., say, maximum likelihood) is that you can pick a that are "robust" not "efficient," just as you can run OLS (and correct the standard errors).

[2] The algebra is straightforward:

$$\sqrt{T}g_T(\hat{b}) \to N(0, \left[I - d(ad)^{-1}a\right] S \left[I - d(ad)^{-1}a\right]').$$
$$\left[I - d(d'S^{-1}d)^{-1}d'S^{-1}\right] S \left[I - d(d'S^{-1}d)^{-1}d'S^{-1}\right]' =$$
$$\left[I - d(d'S^{-1}d)^{-1}d'S^{-1}\right] S \left[I - S^{-1}d'(d'S^{-1}d)^{-1}d'\right] =$$
$$S - d(d'S^{-1}d)^{-1}d' - d(d'S^{-1}d)^{-1}d' + d(d'S^{-1}d)^{-1}d'S^{-1}d'(d'S^{-1}d)^{-1}d' =$$
$$S - d(d'S^{-1}d)^{-1}d'.$$

"Efficiency" means that *if* the model is exactly and precisely 100% true, *if* the data are perfectly measured and perfectly correspond to the concepts of the model, then the efficient GMM estimate has the lowest asymptotic standard error among all GMM estimates that use the same choice of moments $g_T(b)$. If not... no promises.

"Efficiency" *only means* with respect to the initial choice of moments! If you add other moments – other model predictions, or other returns – this always can help.

17.5 The minimization approach

A second natural way to make g_T "as small as possible" is

$$\hat{b}_1 = \min_b g_T(b)' W g_T(b)$$

for some "weighting matrix." The first order conditions:

$$\begin{aligned} \left\{ \frac{\partial g_T(b)'}{\partial b} W \right\} g_T(b) &= 0 \\ a_T g_T(b) &= 0 \end{aligned}$$

So this is just a way to pick an a. Also note why we used a_T before: this a will depend on the sample, but converges asymptotically.

In finance, this procedure amounts to minimizing the weighted sum of squares of the alphas. Then the minimization objective is the same as the evaluation / testing objective, is the weighted sum of squared alphas "too big?"

But as we asked "which a matrix should you use?" Now we ask "Which weighting matrix W should you choose?"

Again, the first answer is, use whatever you want to use, to produce reliable estimates that are robust to small unmodeled specification and measurement errors. GMM is a tool not a black box!

Again, there is an answer from statistical efficiency: Use S as weighting matrix in a *second stage* minimization:

$$\hat{b}_2 = \min_b g_T(b)' S^{-1} g_T(b)$$

(you need a first stage, either with fixed W or fixed a to estimate S). The first order conditions:

$$\begin{aligned} \left\{ \frac{\partial g_T(b)'}{\partial b} S^{-1} \right\} g_T(b) &= 0 \\ d' S^{-1} g_T(b) &= 0 \\ a_T g_T(b) &= 0 \end{aligned}$$

This minimization then produces the efficient GMM estimate.

Then, we can write the second stage test as

$$TJ_T = T \min_{\{b\}} \left[g_T(b)' S^{-1} g_T(b) \right] \sim \chi^2_{\#mom-\#par}$$

(This takes a little bit of clever algebra. See Hansen 1982.)

Advantage of minimization:

$$ag_T(\hat{b}) = 0$$

is a *nonlinear* set of equations. Solving that can be a bear. *Minimization* is much easier numerically.

2) Why is S the efficient weighting matrix?

A weighting matrix $\min g_T' W g_T$ uses W to force GMM to pay more attention to some rather than others.

What W should you use? Again, the right answer is, first of all, economically interesting choices, or statistically robust choices – choices that will work well with model misspecification and data mismeasurement. !

The statistical answer is, pay attention to well-measured moments. Remember,

$$var(g_T) \to \frac{1}{T} \sum_{j=-\infty}^{\infty} E(u_t u_{t-j}') = \frac{1}{T} S$$

Thus, weighting by S^{-1} means paying attention to moments that are well measured.

17.6 GMM applied to OLS

GMM allows you to correct OLS standard errors for many problems. Historically, these corrections were derived by GMM, and solved long-standing hard questions. (The White, Hansen-Hodrick, and Newey-West standard errors were famous papers, not problem sets.) This discussion is also a preview to how we will use GMM to update the distribution theory of classic regression-based methods in asset pricing.

A regression is

$$y_t = x_t' \beta + \varepsilon_t$$

$$x_t = \begin{bmatrix} 1 \\ x_t^1 \\ x_t^2 \end{bmatrix}$$

Let's map this to GMM. Moments: OLS chooses the estimate to make the right hand variables uncorrelated with the error term. Thus, we can define the OLS estimator as an exactly identified $a = I$ GMM estimate using the orthogonality condition,

$$g_T(\hat{\beta}) = E_T \left[x_t \left(y_t - x_t' \hat{\beta} \right) \right] = 0$$

$$\hat{\beta} = \left[E_T(x_t x_t') \right]^{-1} E_T(x_t y_t)$$

This is a case of linear GMM, so no search or minimization is needed to find the estimate.

This seems a little different from the usual formula. Usually we write

$$X = \begin{pmatrix} x_1' \\ x_2' \\ x_t' \end{pmatrix}$$

and then $\frac{1}{T}X'X = E_T(x_t x_t')$. Thus our $\hat{\beta}$ formula is the same as $\hat{\beta} = (X'X)^{-1} X'Y$.

Now, apply the GMM distribution theory,

$$\begin{aligned} d &= \frac{\partial g_T}{\partial \beta'} = -E(x_t x_t') \\ f(x_t, \beta) &= u_t(\beta) = x_t (y_t - x_t \beta) = x_t \varepsilon_t \\ S &= \sum_{j=-\infty}^{\infty} E\left(u_t u_{t-j}'\right) = \sum_{j=-\infty}^{\infty} E\left(\varepsilon_t x_t x_{t-j}' \varepsilon_{t-j}\right) \end{aligned}$$

$$\begin{aligned} var(\hat{\beta}) &= \frac{1}{T} d^{-1} S d^{-1\prime} \\ &\frac{1}{T} E(x_t x_t')^{-1} \sum_{j=-\infty}^{\infty} E\left(\varepsilon_t x_t x_{t-j}' \varepsilon_{t-j}\right) E(x_t x_t')^{-1} \end{aligned}$$

We're done!

Some special cases to make this familiar:

a) *Old formulas.* If ε_t serially uncorrelated, conditionally homoskedastic then the formula simplifies:

$$\begin{aligned} E(\varepsilon_t | x_t x_{t-1} .. \varepsilon_{t-1} \varepsilon_{t-1} ..) &= 0 \rightarrow \text{ no lags } j \\ E(\varepsilon_t^2 | x_t x_{t-1} .. \varepsilon_{t-1} \varepsilon_{t-1} ..) &= \sigma_\varepsilon^2 \end{aligned}$$

$$E(ab) = E(E(a|b)b)$$

means

$$E\left(\varepsilon_t x_t x_t' \varepsilon_t\right) = E\left[E\left(\varepsilon_t^2 | x_t x_t'\right)\right] = E\left(\varepsilon_t^2\right) E(x_t x_t')$$

thus,

$$\sum_{j=-\infty}^{\infty} E\left(\varepsilon_t^2 x_t x_t'\right) = \sigma_\varepsilon^2 E(x_t x_t')$$

so

$$var(\hat{\beta}) = \frac{1}{T} \sigma^2(\varepsilon) E(x_t x_t')^{-1} = \sigma^2(\varepsilon)(X'X)^{-1}.$$

b) *White errors.* Suppose ε_t are iid but conditionally heteroskedastic. If x_t is large, the variance of ε is also likely to be large. This happens in practice. Then it's easier to draw lines of different slope through the data, meaning regular standard errors are too small.

Now, we still have no lags to deal with, but we can't take out the ε. Still, it's easy enough to calculate:

$$var(\hat{\beta}) = \frac{1}{T} E(x_t x_t')^{-1} E\left(\varepsilon_t x_t x_t' \varepsilon_t\right) E(x_t x_t')^{-1}$$

Isn't this lovely? A simple formula to deal with conditional heteroskedasticity. You don't have to model the heteroskedasticity either. If ε^2 is large when x is large, the middle term will pick that up in the data for you. It will be bigger, showing the effect of conditional heteroskedasticity in raising the size of standard errors.

c) *Hansen-Hodrick / Newey-West errors*: What if errors are serially correlated? This happens all the time, as in our regression of long run returns on DP. Now, use the full formula.

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

You can't use infinite lags in practice, and high lags are poorly estimated. Thus, this is usually estimated with

$$var(\hat{\beta}) = \frac{1}{T} E(x_t x_t')^{-1} \sum_{j=-k}^{k} w_j E\left(\varepsilon_t x_t x_{t-j}' \varepsilon_{t-j}\right) E(x_t x_t')^{-1}$$

for example

$$w_j = \frac{\|k - j\|}{k}$$

and a promise to let k increase with sample size. Alternatively, you can use a parametric model of autocorrelation like the AR(1). Then you get simple formulas.

Note: What we are doing here is to find correct standard errors for OLS in the presence of error terms that violate the usual OLS assumptions. This is distinct from GLS. The OLS estimate may be "inefficient," but we're using it anyway. Thus, this is also a good example of how GMM lets you find a distribution theory for robust but sensible estimates rather than require you to use efficient but potentially fragile estimates.

17.7 Using GMM

(Note: The lectures condensed considerably from this treatment.)

Some comments on using GMM. Remember we defined the GMM estimate as

$$a_T g_T(\hat{b}) = 0$$

You can use *any* a, of enough rank. This is the genius – and danger – of GMM. *You can pick* moments you want to fit. The ability to use an arbitrary a and not just the fully efficient estimate is very useful!

You can pick "robust" "economically important" etc. moments. This ability coincides with a major change in empirical philosophy in economics over the last few decades. We do not regard models as literal "truth" but rather as quantitative parables. Thus, we want our models to fit in "sensible" directions, and to ignore "silly" predictions.

A prime example comes from macroeconomics, in which many great models have one shock such as a technology shock. This fact means that some transformation of all the variables are perfectly correlated. The real world has more than one shock and no perfect correlations. Thus the models can be "rejected" with infinite certainty. But they're still nice parables. So your choice is, either dress them up with a lot of "measurement errors" to pretend they are the literal truth, or to adopt econometric techniques that allow you to specify the "important" moments and ignore the perfect correlation.

But one should not go so far as to ignore sampling error! At least we can measure the sampling uncertainty associated with our "calibration" exercises. GMM is an ideal tool for this purpose.

A classic asset pricing example: consider the CAPM or consumption model, as discussed above (I repeat as a reminder)

$$m(a,b) = a - bR^{em} \text{ or } m = \beta\left(\frac{C_{t+1}}{C_t}\right)^{-\gamma}$$

Why don't we use the market and risk free rate to "estimate" the parameters, a, b or β, γ, and then use additional assets such as hml and smb to "test", or to "see how well the model performs on explaining the value and small firm premiums." The following a choice implements this idea.

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} E_T\left(m(a,b)R^{em}\right) \\ E_T\left(m(a,b)R^{f} - 1\right) \\ E_T\left(m(a,b)R^{ehml}\right) \\ E_T(m(a,b)R^{esmb}) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

It is a useful trick in economics. Real business cycle and DSGE models often "calibrate" preference and technology parameters β, γ, etc. by long term averages averages such as capital / labor ratios, (equity premium!) etc. Then they "evaluate" the model by its ability to capture variances and correlations – carefully avoiding the stochastic singularity implied by the one-shock structure. Good – but maybe we don't really know all that much about the parameters from the "calibration"? How sure are we about those parameters? If the model's predicted correlation between investment and output is different from that in the data, how likely is that fact just due to sampling error in the model, including the sampling error in the parameter calibration? We can implement the calibrate-verify procedure and answer all these sampling questions with

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} E_T\left(\text{model k/l - data k/l}\right) \\ E_T(\text{model } \Delta y\text{- data } \Delta y) \\ \sigma(\Delta c), \sigma(\Delta i) \\ corr(\Delta c, \Delta i, \text{ etc}) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

You don't need to test models! The whole equity premium, risk free rate puzzle comes down to

studying moment conditions

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} E_T\left(\beta\left(\frac{C_{t+1}}{C_t}\right)^{-\gamma} R^{em}\right) \\ E_T\left(\beta\left(\frac{C_{t+1}}{C_t}\right)^{-\gamma} R^{f} - 1\right) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

As this is exactly identified, there is nothing to test. But we can examine the β, γ estimates (γ being too high is the equity premium, $\beta > 1$ the risk free rate puzzle), and we can get standard errors of β, γ, to see how strongly the data require puzzling values.

GMM is a very flexible framework for all sorts of econometric problems. The "generated regressor" is the classic case.

$$\begin{aligned} y_t &= bx_t + \varepsilon_t \\ z_t &= \beta(E(y_t|x_t)) + \delta_t = \beta(\hat{b}x_t) + \delta_t \end{aligned}$$

How do we find standard errors of $\hat{\beta}$ that incorporate the sampling uncertainty of $\hat{a}$ and $\hat{b}$? Answer: write down the moments

$$\begin{aligned} E[(y_t - bx_t)x_t] &= 0 \\ E[(z_t - \beta(bx_t))z_t] &= 0. \end{aligned}$$

The weighting matrix $a = I$ will give the OLS estimates of both equations. The S matrix captures correlations between the two equations, and the off diagonal elements of the d matrix capture how b estimates enter the second equations. All together, then, the standard errors of $\hat{\beta}$ will then include the uncertainty about b.

A classic case comes up in finance: we run time series regressions of returns on factors to estimate β,

$$R_t^{ei} = a + \beta_i f_t + \varepsilon_{it} \quad t = 1, 2, ..T \ \forall i$$

and then we run cross sectional regressions of average returns on β to estimate factor risk preimums and pricing errors.

$$E(R^{ei}) = \gamma + \hat{\beta}_i \lambda + \alpha_i \ i = 1, 2, ..N$$

We can't use OLS formulas for $\sigma(\hat{\lambda})$ because 1) the errors α_i are correlated across assets i and 2) the $\hat{\beta}$ are not fixed, but generated regressors.

In finance, our moments are $E\left[(m(\text{scalar})\ R^e(\text{vector})\right]$ so $aE(mR^e) = E\left[m\ (aR^e)\right]$. *The a matrix is a choice of portfolios of the underlying assets.* By choosing an a matrix, you determine which portfolios of given assets will have zero in-sample alphas to estimate parameters. There are always some! If you say I'll use $a = [1111..]$ to treat all assets equally, you are not: you are estimating parameters to perfectly price the equally weighted portfolio in sample. There is always some perfectly priced portfolio. So look at a and choose it sensibly!

Similarly, a minimization,

$$\min g_T(\hat{b})' W g_T(\hat{b})$$

With first order condition,

$$d'W g_T(\hat{b}) = 0$$

also amounts to a choice of how to weight portfolios. Write $W = Q'Q$, then

$$\begin{aligned} &\min g_T(\hat{b})'Q'Qg_T(\hat{b}) \\ &\min E_T\left[m(\hat{b})\left[QR_t^e\right]'\right] E_T\left[m(\hat{b})\left(QR_t^{e\prime}\right)\right] \end{aligned}$$

so the choice of weighting matrix tells us which portfolios you want GMM to price as well as possible.

17.7.1 a and W choices

So what's a "good" choice of a matrix or W matrix? It will certainly help in this decision to recall the properties of our choices.

Fist stage or efficient GMM?

First stage uses an arbitrary (hopefully thoughtful) a or W matrix. Its properties are 1. the estimate $\hat{b}$ is *consistent.* 2. GMM gives standard error $\sigma(\hat{b})$ formulas, $\sigma(\hat{g}_T)$ formulas, χ^2 tests for $g = 0$. *All of these are consistent. All give correct standard errors and test statistics.* But the estimate is not *asymptotically efficient.* (if the model is 100% correct and given choice of moments)

Stopping at first stage estimates/tests is not "wrong" in that it does not introduce a bias. It is not "wrong" in the way that using $\sigma/\sqrt{T}$ for autocorrelated data is "wrong." The only criticism one could make is that it is not "efficient." You could have wrung slightly more precise estimates out of the data. Now efficiency does matter. Throwing away half the data is inefficient too. But at least check that the gain in efficiency (lower asymptotic standard errors) is not balance by a loss of robustness or common sense. You could have included more data too, which usually has a much greater increase in efficiency.

Second stage or efficient GMM is *Asymptotically efficient (given the choice of moments)* That's it. It has a lower true (not necessarily estimated) $\sigma(\hat{b})$..

The disadvantage: is it better in finite samples? is it robust to model misspecification? Does it focus in on small bid ask spreads, measurement errors, and other unmodeled ways in which the model is not a 100% accurate description of reality? (Remember, that is assumption 1! When you tell theory "Assume the model is true" you mean literally, 100% absolutely exactly true, and it can price a 1000% long-short position!)

As you can see, I'm sympathetic to robust but inefficient first stage estimates with well chosen a and W matrices, but correct standard errors. That's not always and everywhere. Asset pricing and macro cases that I deal with usually have enough data to statistically measure parameters quite well, or at least efficient estimates don't improve (lower standard errors) a lot. These cases also feature a lot of "quantitative parable" model simplification in which efficient or GLS transformations tend to seize on small specification and measurement errors and deliver garbage. However, in the

cases I'm familiar with there is a lot of serial or cross correlation of moments. Standard errors which ignore such correlation are very wrong. And the presence of such correlation is a reason that GLS or efficient estimates are often radically different from OLS or first stage estimates. A circumstance with a very nearly "true" model, very well measured data, but not much data and not much correlation of errors could easily reverse this prejudice and yield a preference for efficient procedures.

OLS vs. GLS analogy.

OLS: solves min $(Y - X\beta)'(Y - X\beta) \rightarrow \beta = (X'X)^{-1} X'Y$

What if $Y = X\beta + \varepsilon$, and ε autocorrelated/heteroskedastic?

Recall that the OLS $\hat{\beta}$ is consistent, but not efficient. The OLS standard error formula $\sigma = (X'X)^{-1} s^2$ is biased. It is often very badly biased (especially cross sectionally, and waving the "cluster" wand does not fix the problem)

The statistician's answer: Run a 2 stage procedure 1) Run OLS to find β_1. Form estimate of $E(\varepsilon\varepsilon') = \Omega$, 2) Run GLS, $\min(Y - X\beta)'\Omega^{-1}(Y - X\beta) \rightarrow \hat{\beta} = \left(X'\Omega^{-1}X\right)^{-1} X'\Omega^{-1}Y$. This GLS estimate is *consistent, efficient.* Its standard error formula $\sigma^2(\beta) = (X'\Omega^{-1}X)^{-1}s^2$ is also consistent.

The Economist's answer is, increasingly: *It is often much better to run ols, but correct standard errors.* Nearly universal in asset pricing, corporate finance. "FMB" standard errors are just this.

GLS works poorly in small samples, and often throws out the baby with bathwater, e.g. first differencing data Why? GLS assumes the model is literally 100% true, without measurement errors. See "A Brief Parable of Overdifferencing"

An alternative dominates practice: Run OLS, which is "robust" *but fix the standard errors.* In the standard regression context,

$$\begin{aligned}
\sigma^2(\hat{\beta}_{OLS}) &= var\left[\left(X'X\right)^{-1} X'Y\right] \\
&= var\left[\left(X'X\right)^{-1} X'\varepsilon\right] \\
&= \left(X'X\right)^{-1} X'\Omega X \left(X'X\right)^{-1}
\end{aligned}$$

The first stage formulas above are GMM analogues to this formula. The GMM does OLS formulas are similar. As above, this is appropriate when you have lots of data, so squeezing more information out of the data is not critical, there is lots of correlation to correct for, so OLS and GLS differ a lot, and the model is approximate or data are measured with error, so (for example) though you trust $y_{it} = x_{it}\beta + \varepsilon_{it}$, you don't trust $(y_{it} - y_{it-1}) = (x_{it} - x_{it-1})\beta + (\varepsilon_{it} - \varepsilon_{it-1})$ or even less the difference-in-difference $(y_{it} - y_{it-1}) - (y_{jt} - y_{jt-1}) = [(x_{it} - x_{it-1}) - (x_{jt} - x_{jt-1})]\beta + (\varepsilon_{it} - \varepsilon_{it-1}) - (\varepsilon_{jt} - \varepsilon_{jt-1})$, which throw out gobs of variation in the data.

17.8 Assumptions

I have not talked much at all about the assumptions behind GMM or the beautiful asymptotic distribution theory by which all these results are proved. For applied work, a few assumptions are important to keep in mind:

1. The model is true. This seems innocuous, but it is important, and can help you to diagnose problems. Not only must there be a b such that $ag(b) = 0$, there must be a $\hat{b}$ such that $a_T(\hat{b})g_T(\hat{b}) = 0$ in each sample. Loosely, b must be identified.

2. The data are stationary. This is all based on the idea that sample means converge to population means. If $x_t = x_{t-1} + \varepsilon_t$, then the sample mean of x does not converge to a population mean. GMM uses different parts of the sample to guess what would happen in different histories, and the data must be stationary to make that implication.

That's only two that most frequently go wrong.