Generalized Method of Moments

These notes provide an introduction to Generalized Method of Moments (GMM) estimation. GMM is a very popular estimation method. It was invented (at least to some extent) by economists in the early 1980s. I think that the reason for this is because economists often like only partially specifying the model (e.g., most of our arguments this semester have relied on moment conditions and asymptotic arguments rather than making distributional assumptions). In fact, many of the approaches that we have talked about this semester are actually special cases of GMM (see discussion below). GMM is discussed in Chapter 13 of the textbook.

GMM is basically an approach to estimate some population parameter, θ, when it is known that some moments that depend on the parameter are equal to 0. For example, $\mathbb{E}[g(X, Y; \theta)] = 0$ where g returns an $l \times 1$ vector and θ is a $k \times 1$ vector. The case we will emphasize in these notes is when $l > k$, so that there are more moment conditions than there are parameters to estimate. In principle, this should be useful; more moment conditions is generally better than fewer moment conditions. But it does create some additional issues that we will at least need to think through. It will also give us some opportunities.

I am mainly going to teach GMM with respect to a couple of examples: over-identified instrumental variables and difference-in-differences with multiple periods and variation in treatment timing. That said, GMM is a general framework for estimating parameters when there are more moment conditions than there are parameters to estimate.

Moment Conditions

H. 12.1-12.6

To start with, let me provide a brief introduction to instrumental variables (you will see much more about this next semester). For both linear projection and linear CEF models, we have typically written $Y = X'\beta + e$ and have that

$$\beta = \underset{b}{\operatorname{argmin}}\, \mathbb{E}\Big[(Y - X'b)^2\Big]$$

which implied that $\beta = \mathbb{E}[XX']^{-1}\mathbb{E}[XY]$. Plugging in the model for Y into this expression, it also follows that $\mathbb{E}[Xe] = 0$.

It is also possible to "invert" this type of argument. A traditional way to motivate linear regression is to suppose that a researcher is interested in a "structural" model

$$Y = X'\beta + e$$

where we allow for the possibility that $\mathbb{E}[Xe] \neq 0$. In this case, you would say that X (or at least some component(s) of X) is **endogenous**.

You might ask: How can this happen? Isn't $\mathbb{E}[Xe] = 0$ a property of linear projection? If you are

asking these questions, you have a good point; indeed, pretty much without loss of generality, you can consider the linear projection of Y on X, and it will be the case that X will be uncorrelated with the projection error.

However, please recall our example of omitted variable bias earlier in the semester. In that case, we were interested in β from the following projection model (or CEF model):

$$Y = X'\beta + W'\gamma + u$$

For example, here you might be willing to interpret β_1 as the causal effect of X_1 on Y if you could control for $X_2, \dots, X_k$ and W; or, alternatively, you might think of β as being "deep, structural" parameters that you would like to estimate and then take to some other setting. In this case, and if W were not observed, you would have $\mathbb{E}[Xu] = 0$, but this might not be so useful for recovering β. And, in particular, if you put everything that you don't observe in the error term (define $e = W'\gamma + u$), you would have that,

$$\mathbb{E}[Xe] = \mathbb{E}[XW']\gamma + \mathbb{E}[Xu] = \mathbb{E}[XW']\gamma$$

where the last term is not, in general, equal to 0, unless X and W are uncorrelated (or $\gamma = 0$ so that there is no effect of W on the outcome); see our previous discussion on omitted variable bias for more details.

Given this discussion, let's continue to think of $Y = X'\beta + e$ as a structural model. Let's first think through the case where $\mathbb{E}[Xe] = 0$. This is going to be a simple case, and in this case, you would say that X is **exogenous**. $\mathbb{E}[Xe] = 0$ is an example of a **moment equation**. To say more clearly what we mean by this, notice that you can equivalently write $\mathbb{E}[X(Y - X'\beta)] = 0$ by plugging in for e. This is $k \times 1$ vector of moment conditions. Since we are interested in the $k \times 1$ vector of parameters β, it seems reasonable to hope that we can solve the system of equations to recover β. And, in fact, under some conditions, we can,

$$0 = \mathbb{E}[XY] - \mathbb{E}[XX']\beta \implies \beta = \mathbb{E}[XX']^{-1}\mathbb{E}[XY]$$

where the main requirement is a familiar one: that $\mathbb{E}[XX']$ is invertible. This is a familiar expression for β and suggests estimating β from the regression of Y on X (as we have done many times before).

Now, let's move back to the more challenging case where $\mathbb{E}[Xe] \neq 0$. A leading case here would be that X includes one endogenous variable, say X_1, while the other variables are exogenous. To give an example, suppose that X_1 is years of education, X_2 is age, we include an intercept as X_3, and we are worried about the omitted variable W that is a person's "ability". In this case, it might be reasonable to think that $\mathbb{E}[X_1 e] \neq 0$ (because years of education is correlated with ability), but $\mathbb{E}[X_2 e] = \mathbb{E}[X_3 e] = 0$. You can immediately see that we don't have enough information from these moment conditions to recover the parameters. Here there are three parameters, β_1, β_2, and β_3, but we only have two moment equations. This is like trying to solve for three unknowns in a system of

two equations. This is sometimes referred to as being **under-identified** or that it fails the **order condition** (where the order condition is that, at a minimum, there need to be as many moment equations as there are parameters to recover). More generally, following the textbook's notation, let's partition X into X_1 and X_2 into the regressors which are endogenous (X_1) and the regressors which are exogenous (X_2), and where these are $k_1 \times 1$ and $k_2 \times 1$ vectors.

One possibility is to find another $k \times 1$ random vector that satisfies $\mathbb{E}[Ze] = 0$, even if it doesn't hold that $\mathbb{E}[Xe] = 0$. As above, it is helpful to partition Z into Z_1 and Z_2 where $Z_2 = X_2$ includes the exogenous variables in X and Z_1 are variables that are not in X. Z_1 is referred to as an **instrumental variable**. This is an extremely important concept and one that will be covered extensively next semester. In general, it's challenging to find convincing instruments, but, for now, let's just suppose that we have an instrument. If Z is $k \times 1$ and $\mathbb{E}[Ze] = 0$, this suggests that we could at least hope to recover β. In this case, we have that

$$0 = \mathbb{E}[Ze] = \mathbb{E}[Z(Y - X'\beta)] = \mathbb{E}[ZY] - \mathbb{E}[ZX']\beta \implies \beta = \mathbb{E}[ZX']^{-1}\mathbb{E}[ZY]$$

This implies that, if we can invert $\mathbb{E}[ZX']$, then we can recover β. Note that $\mathbb{E}[ZX']$ will be invertible if $\text{rank}(\mathbb{E}[ZX']) = k$ (i.e., that $\mathbb{E}[ZX']$ has full rank). Therefore, this condition is often referred to as the **rank condition** or **relevance condition**. The intuition for this condition is that the instrument needs to be correlated with the endogenous regressor. This is the condition that prevents you from using random numbers as instruments. By construction, random numbers with satisfy the exogeneity condition, $\mathbb{E}[Ze] = 0$, but they fail the rank condition.

> **Side-Comment:** I am not going to cover it here, but it would be good practice to think through how to estimate β using the above expression and to be able to show that it is consistent and asymptotically normal.

Overidentification

H. 13.2-13.4, 13.5-13.10

Now, let's suppose that Z is $l \times 1$ where $l > k$. In other words, we have more instruments than endogenous regressors. This seems like good news, but notice that the previous strategy does not work any more. In particular, $\mathbb{E}[ZX']$ is now an $l \times k$ matrix. Since $l > k$, $\mathbb{E}[ZX']$ is not square and, therefore, not invertible.

What should we do here? One idea is to just throw away some of the "extra" Z's. In practice, this could work, but it is an unsatisfactory solution as we would potentially throw away useful information if we proceeded this way. We still have that

$$\mathbb{E}[ZX']\beta = \mathbb{E}[ZY]$$

Even in this is true, however, in the over-identified case, in general, is not possible to find a $\hat{\beta}$ such that

$$0 = \frac{1}{n}\sum_{i=1}^{n} Z_i Y_i - \frac{1}{n}\sum_{i=1}^{n} Z_i X_i' \hat{\beta}$$

holds exactly. It is a bit little tricky to understand this because, if $\mathbb{E}[Ze] = 0$, then it would be the case that $0 = \mathbb{E}[ZY] - \mathbb{E}[ZX']\beta$, but, even if that's true, because we have to estimate the population expectations, in the sample, we generally won't be able to find an exact solution. It is also worth pointing out that this is an example of a moment condition, as we discussed above.

Instead, an alternative idea is to set $\hat{\beta}$ to be the value of b that makes $\frac{1}{n}\sum_{i=1}^{n} Z_i Y_i - \frac{1}{n}\sum_{i=1}^{n} Z_i X_i' b$ as close to 0 as possible. This is the idea of GMM. Re-writing this in terms of "data matrices", we are going to try to choose a value b to make $\mathbf{Z}'\mathbf{Y} - \mathbf{Z}'\mathbf{X}b$ as close to 0 (and, therefore, as close to satisfying the moment condition) as we can.

A natural way to do this is to minimizing the (weighted) Euclidean distance between the $l \times 1$ vector above and 0; that is,

$$\hat{\beta} = \underset{b}{\text{argmin}} \left(\mathbf{Z}'\mathbf{Y} - \mathbf{Z}'\mathbf{X}b\right)' \widehat{\mathbf{W}} \left(\mathbf{Z}'\mathbf{Y} - \mathbf{Z}'\mathbf{X}b\right)$$

where $\widehat{\mathbf{W}}$ is an $l \times l$ "weighting matrix" that allows us to put more weight on some moments than others (we will return to this issue below).

The above discussion is for the case where we want to estimate the parameters of linear model where (i) there is endogeneity and (ii) there is over-identification in the sense of that $l > k$. But these arguments apply more generally. Let's briefly return to our more generic setup discussed above where $0 = \mathbb{E}[g(Y, X; \theta)]$. In this case, the GMM estimator of θ is

$$\hat{\theta} = \underset{\vartheta}{\text{argmin}} \left(\frac{1}{n}\sum_{i=1}^{n} g(Y_i, X_i; \vartheta)\right)' \widehat{\mathbf{W}} \left(\frac{1}{n}\sum_{i=1}^{n} g(Y_i, X_i; \vartheta)\right)$$

which minimizes the (weighted) Euclidean distance between the sample analogue of the moment condition and 0.

Let's return to previous case (which is arguably the leading case…and is the one emphasized in the textbook) of an over-identified linear model. One additional line of algebra shows that

$$\hat{\beta}_{GMM} = \underset{b}{\text{argmin}}\ \mathbf{Y}'\mathbf{Z}\widehat{\mathbf{W}}\mathbf{Z}'Y - 2b'\mathbf{X}'\mathbf{Z}\widehat{\mathbf{W}}\mathbf{Z}'\mathbf{Y} + b'\mathbf{X}'\mathbf{Z}\widehat{\mathbf{W}}\mathbf{Z}'\mathbf{X}b$$

where I've put a "GMM" subscript on $\hat{\beta}_{GMM}$ to indicate that this is the GMM estimator of β

(there are other possible estimators here). The first order condition is given by

$$\begin{aligned} 0 &= -\mathbf{X}'\mathbf{Z}\widehat{\mathbf{W}}\mathbf{Z}'\mathbf{Y} + \mathbf{X}'\mathbf{Z}\widehat{\mathbf{W}}\mathbf{Z}'\mathbf{X}\hat{\beta}_{GMM} \\ \implies \hat{\beta}_{GMM} &= (\mathbf{X}'\mathbf{Z}\widehat{\mathbf{W}}\mathbf{Z}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\widehat{\mathbf{W}}\mathbf{Z}'\mathbf{Y} \end{aligned}$$

This expression is rather long, but supposing that $\widehat{\mathbf{W}} \xrightarrow{p} \mathbf{W}$, you can use very similar arguments to the ones that we have used before to show that $\hat{\beta}_{GMM} \xrightarrow{p} \beta$ and that

$$\sqrt{n}(\hat{\beta}_{GMM} - \beta) \xrightarrow{d} \mathcal{N}(0, \mathbf{V})$$

under the sort of conditions that you would expect (see Assumption 12.2 in the textbook) and where

$$\mathbf{V} := (\mathbb{E}[XZ']\mathbf{W}\mathbb{E}[ZX'])^{-1}\mathbb{E}[XZ']\mathbf{W}\ \mathbf{W}\mathbb{E}[ZX'](\mathbb{E}[XZ']\mathbf{W}\mathbb{E}[ZX'])^{-1}$$

There are a few things left to discuss. First, this discussion begs the question of how we should actually choose the weighting matrix $\mathbf{W}$ in practice. Let's talk intuition first. Generally, we would probably like to put more weight on the "precise" moments and less weight on the "imprecise" moments. We also would probably like to take into account the correlation between different moments, and, in general, put less weight on highly correlated moments (as they contain similar information to each other). This information on precision of moments and correlation of moments is contained in the variance of the moment conditions: $\mathbf{\Omega} := \text{var}(Ze) = \mathbb{E}[ZZ'e^2]$. And, in particular, the arguments above suggest weighting by the inverse of the variance matrix; that is, $\mathbf{\Omega}^{-1} = \mathbb{E}[ZZ'e^2]^{-1}$.

This intuition turns out to be correct. I will not provide full details here, but it turns out the setting $\mathbf{W} = \mathbf{\Omega}^{-1}$ results in what is called **efficient GMM**. You can check the arguments in Section 13.8 (in particular, Theorem 13.5 and Exercise 13.4). These arguments are not that complicated, and you will be able to understand them. I am just omitting them because we are running out of time for the semester. In addition, under this choice for the weighting matrix, the expression for $\hat{\beta}_{GMM}$ changes to

$$\hat{\beta}^o_{GMM} = (\mathbf{X}'\mathbf{Z}\hat{\mathbf{\Omega}}^{-1}\mathbf{Z}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\hat{\mathbf{\Omega}}^{-1}\mathbf{Z}'\mathbf{Y}$$

where I have included an "o" subscript to indicate that this is the efficient (or "optimal") GMM, and the expression for the asymptotic variance simplifies to

$$\mathbf{V}_0 = (\mathbb{E}[XZ']\mathbf{\Omega}^{-1}\mathbb{E}[ZX'])^{-1}$$

Notice that, the expression for $\hat{\beta}^o_{GMM}$ requires an estimate of $\mathbf{\Omega}$. This is typically done by starting out with a consistent, but possible inefficient estimator of β. For example, the most common choice would be to set $\widehat{\mathbf{W}} = (\mathbf{Z}'\mathbf{Z})^{-1}$. Making this choice actually results in the two-stage least squares

estimator (TSLS) $\hat{\beta}_{2sls}$ (which you may be familiar with from other classes, and will be discussed in more detail next semester). Setting $\widehat{\mathbf{W}} = \mathbf{I}_l$ also works. Given some preliminary estimate of β, that we'll call $\hat{\beta}$, you can compute $\hat{e}_i = Y_i - X_i'\hat{\beta}$, and then estimate $\hat{\mathbf{\Omega}} = \frac{1}{n}\sum_{i=1}^{n} Z_i Z_i' \hat{e}_i^2$ and then plug this to estimate $\hat{\beta}^o_{GMM}$.

Over-identification Tests

H. 12.31, 13.21

In most applications with over-identification, it is common to report an over-identification test. I'll continue to think about the case of a linear model with endogeneity that we have been discussing above, but similar arguments to the ones below apply to the more general moment conditions that we have also mentioned above. In Section 12.31, the textbook provides an example where $k = 1$ and $l = 2$ that gives a rough intuition of over-identification tests. In this setting, there are two moment conditions: $0 = \mathbb{E}[Z_1 e] = \mathbb{E}[Z_1(Y - X_1\beta)]$ and $0 = \mathbb{E}[Z_2 e] = \mathbb{E}[Z_2(Y - X_1\beta)]$. It is possible to recover β if you use either moment condition. But, suppose that you use the first moment condition to recover β, then, as long as the second moment condition is actually true, you should be able to plug in the recovered β and the moment condition still hold. If not, then that would indicate that at least one of the moment conditions does not hold. We will formalize this argument below and account for issues relating to sampling variance that make the arguments above not hold exactly (though they should still be close to holding) once we move from the population to the sample.

Define $J(b) = n\Big(\frac{1}{n}\mathbf{Z}'\mathbf{Y} - \frac{1}{n}\mathbf{Z}'\mathbf{X}b\Big)' \hat{\mathbf{\Omega}}^{-1}\Big(\frac{1}{n}\mathbf{Z}'\mathbf{Y} - \frac{1}{n}\mathbf{Z}'\mathbf{X}b\Big)$ which is the GMM criteria function that we minimized above to estimate β (here, we use $\hat{\mathbf{\Omega}}$ as the we weighting matrix and we also multiply by n; both of these matter for the arguments in this section). We will test $\mathbb{H}_0 : \mathbb{E}[Ze] = 0$ vs $\mathbb{H}_1 : \mathbb{E}[Ze] \neq 0$, and we will consider the test statistic $J(\hat{\beta}^o_{GMM})$. Let's consider it's behavior under $\mathbb{H}_0$. First, notice that,

$$\begin{aligned} J(\hat{\beta}^o_{GMM}) &= n\left(\frac{1}{n}\mathbf{Z}'\mathbf{Y} - \frac{1}{n}\mathbf{Z}'\mathbf{X}\hat{\beta}^o_{GMM}\right)' \hat{\mathbf{\Omega}}^{-1/2} \underbrace{\hat{\mathbf{\Omega}}^{1/2}\hat{\mathbf{\Omega}}^{-1}\hat{\mathbf{\Omega}}^{1/2}}_{=\mathbf{I}_l} \hat{\mathbf{\Omega}}^{-1/2}\left(\frac{1}{n}\mathbf{Z}'\mathbf{Y} - \frac{1}{n}\mathbf{Z}'\mathbf{X}\hat{\beta}^o_{GMM}\right) \\ &= n\left[\hat{\mathbf{\Omega}}^{-1/2}\left(\frac{1}{n}\mathbf{Z}'\mathbf{Y} - \frac{1}{n}\mathbf{Z}'\mathbf{X}\hat{\beta}^o_{GMM}\right)\right]' \hat{\mathbf{\Omega}}^{-1/2}\left(\frac{1}{n}\mathbf{Z}'\mathbf{Y} - \frac{1}{n}\mathbf{Z}'\mathbf{X}\hat{\beta}^o_{GMM}\right) \end{aligned}$$

Next, notice that,

$$\begin{aligned}
\hat{\mathbf{\Omega}}^{-1/2}\Big(\frac{1}{n}\mathbf{Z}'\mathbf{Y}-\frac{1}{n}\mathbf{Z}'\mathbf{X}\hat{\beta}\Big) &= \hat{\mathbf{\Omega}}^{-1/2}\Big(\frac{1}{n}\mathbf{Z}'\mathbf{e}-\frac{1}{n}\mathbf{Z}'\mathbf{X}(\hat{\beta}^o_{GMM}-\beta)\Big)\\
&= \hat{\mathbf{\Omega}}^{-1/2}\Big(\frac{1}{n}\mathbf{Z}'\mathbf{e}-\frac{1}{n}\mathbf{Z}'\mathbf{X}\Big(\frac{1}{n}\mathbf{X}'\mathbf{Z}\hat{\mathbf{\Omega}}^{-1}\frac{1}{n}\mathbf{Z}'\mathbf{X}\Big)^{-1}\frac{1}{n}\mathbf{X}'\mathbf{Z}\hat{\mathbf{\Omega}}^{-1}\frac{1}{n}\mathbf{Z}'\mathbf{e}\Big)\\
&= \Big(\mathbf{I}-\hat{\mathbf{\Omega}}^{-1/2}\frac{1}{n}\mathbf{Z}'\mathbf{X}\Big(\frac{1}{n}\mathbf{X}'\mathbf{Z}\hat{\mathbf{\Omega}}^{-1}\frac{1}{n}\mathbf{Z}'\mathbf{X}\Big)^{-1}\frac{1}{n}\mathbf{X}'\mathbf{Z}\hat{\mathbf{\Omega}}^{-1}\hat{\mathbf{\Omega}}^{1/2}\Big)\hat{\mathbf{\Omega}}^{-1/2}\frac{1}{n}\mathbf{Z}'\mathbf{e}\\
&= \Big(\mathbf{I}-\hat{\mathbf{\Omega}}^{-1/2}\frac{1}{n}\mathbf{Z}'\mathbf{X}\Big(\frac{1}{n}\mathbf{X}'\mathbf{Z}\hat{\mathbf{\Omega}}^{-1}\frac{1}{n}\mathbf{Z}'\mathbf{X}\Big)^{-1}\frac{1}{n}\mathbf{X}'\mathbf{Z}\hat{\mathbf{\Omega}}^{-1/2}\Big)\hat{\mathbf{\Omega}}^{-1/2}\frac{1}{n}\mathbf{Z}'\mathbf{e}
\end{aligned}$$

where the first equality plugs in $\mathbf{Y}=\mathbf{X}\beta+\mathbf{e}$, the second equality plugs in for $\hat{\beta}^o_{GMM}-\beta$ (which is straightforward to derive given the earlier expression for $\hat{\beta}^o_{GMM}$), the third equality is a little tricky as it treats the first term and second term asymmetrically (the first one just factors everything to the right while the second term keeps $\hat{\mathbf{\Omega}}^{-1/2}$ on the left while multiplying by $\mathbf{I}=\hat{\mathbf{\Omega}}^{1/2}\hat{\mathbf{\Omega}}^{-1/2}$ on the right). The last equality follows immediately.

If we define $\mathbf{R}:=\mathbf{\Omega}^{-1/2}\mathbb{E}[ZX']$, and noticing that $\hat{\mathbf{\Omega}}^{-1/2}\frac{1}{n}\mathbf{Z}'\mathbf{X}\xrightarrow{p}\mathbf{R}$, and that $\hat{\mathbf{\Omega}}^{-1/2}\frac{1}{\sqrt{n}}\mathbf{Z}'\mathbf{e}\xrightarrow{d}u\sim\mathcal{N}(0,\mathbf{I}_l)$ (to see this, note that $\frac{1}{\sqrt{n}}\mathbf{Z}'\mathbf{e}\xrightarrow{d}\mathcal{N}(0,\mathbf{\Omega})$), then it follows that

$$\begin{aligned}
\sqrt{n}\hat{\mathbf{\Omega}}^{-1/2}\Big(\frac{1}{n}\mathbf{Z}'\mathbf{Y}-\frac{1}{n}\mathbf{Z}'\mathbf{X}\hat{\beta}\Big) &= \Big(\mathbf{I}-\mathbf{R}(\mathbf{R}'\mathbf{R})^{-1}\mathbf{R}'\Big)\hat{\mathbf{\Omega}}^{-1/2}\frac{1}{n}\mathbf{Z}'\mathbf{e}+o_p(1)\\
&\xrightarrow{d}\Big(\mathbf{I}-\mathbf{R}(\mathbf{R}'\mathbf{R})^{-1}\mathbf{R}'\Big)u
\end{aligned}$$

This implies that, under $\mathbb{H}_0$,

$$\begin{aligned}
J(\hat{\beta}^o_{GMM}) &\xrightarrow{d} u'\Big(\mathbf{I}-\mathbf{R}(\mathbf{R}'\mathbf{R})^{-1}\mathbf{R}'\Big)'\Big(\mathbf{I}-\mathbf{R}(\mathbf{R}'\mathbf{R})^{-1}\mathbf{R}'\Big)u\\
&= u'\Big(\mathbf{I}-\mathbf{R}(\mathbf{R}'\mathbf{R})^{-1}\mathbf{R}'\Big)u\sim\chi^2_{l-k}
\end{aligned}$$

where the equality holds because $\Big(\mathbf{I}-\mathbf{R}(\mathbf{R}'\mathbf{R})^{-1}\mathbf{R}'\Big)$ is idempotent and symmetric. That this converges to χ^2_{l-k} holds because $u\sim\mathcal{N}(0,\mathbf{I}_l)$ and because $\operatorname{tr}\Big(\mathbf{I}-\mathbf{R}(\mathbf{R}'\mathbf{R})^{-1}\mathbf{R}'\Big)=l-k$. This establishes the behavior of $J(\hat{\beta}^o_{GMM})$ under $\mathbb{H}_0$.

On the other hand, under $\mathbb{H}_1$,

$$\frac{1}{n}\mathbf{Z}'\mathbf{Y}-\frac{1}{n}\mathbf{Z}'\mathbf{X}\hat{\beta}^o_{GMM}\nrightarrow_p 0$$

and, therefore, $J(\hat{\beta}^o_{GMM})\to\infty$ under $\mathbb{H}_1$.

Panel Data Examples

We have motivated GMM using the traditional IV approach, but having more moment conditions than parameters to estimate can occur in other settings as well. In this section, we will consider es-

timating group-time average treatment effects in a difference-in-differences framework using GMM. The discussion in this section loosely follows Marcus and Sant'Anna (JAERE, 2021).

I am going to consider the smallest scale, non-trivial version of this setting. In particular, let's suppose that there are two time periods, $t = 1$ and $t = 2$. Let's also suppose that there are three groups, $g = 2$, $g = 3$, and the never-treated group ($U = 1$) — this is somewhat awkward as we are saying that we somehow know that some units will become treated in period 3 while others will remain untreated even though we don't actually observe outcomes in period 3. That said, let's just go with it, and you can think of this as a device to keep the problem in this section simple.

Under parallel trends, there are two non-redundant moment conditions here:

$$\begin{aligned}\mathbb{E}[\Delta Y_2|G=2] - \mathbb{E}[\Delta Y_2|U=1] - ATT(2,2) &= 0 \\ \mathbb{E}[\Delta Y_2|G=3] - \mathbb{E}[\Delta Y_2|U=1] &= 0\end{aligned}$$

The first one says that $ATT(2,2)$ is equal to mean path of outcomes for group 2 relative to the mean path of outcomes for the never-treated group. The second one says that the mean path of outcomes for group 3 is the same as the mean path of outcomes for the untreated group. In this setting, the moment condition $\mathbb{E}[\Delta Y_2|G=2] - \mathbb{E}[\Delta Y_2|G=3] - ATT(2,2) = 0$ is redundant because it is linear combination of the two above moment conditions. For estimation, it will be helpful to re-write these conditional expectations as unconditional expectations; that is,

$$\begin{aligned}\mathbb{E}\left[\left(\frac{\mathbb{1}\{G=2\}}{p_2} - \frac{U}{p_U}\right)\Delta Y_2\right] - ATT(2,2) &= 0 \\ \mathbb{E}\left[\left(\frac{\mathbb{1}\{G=3\}}{p_3} - \frac{U}{p_U}\right)\Delta Y_2\right] &= 0\end{aligned}$$

(for simplicity, we are going to treat the p_g terms as being known). This fits into the GMM framework as there are two moment conditions, but there is only one parameter to estimate, $ATT(2,2)$.

Let's define

$$Y_i = \begin{pmatrix}\left(\frac{\mathbb{1}\{G_i=2\}}{p_2} - \frac{U_i}{p_U}\right)\Delta Y_{i2} \\ \left(\frac{\mathbb{1}\{G_i=3\}}{p_3} - \frac{U_i}{p_U}\right)\Delta Y_{i2}\end{pmatrix} \quad \text{and} \quad \mathbf{Y} = \begin{pmatrix}Y_1' \\ \vdots \\ Y_n'\end{pmatrix} \quad \text{and} \quad \mathbf{1} = \begin{pmatrix}1 \\ \vdots \\ 1\end{pmatrix} \quad \text{and} \quad \mathbf{R} = \begin{pmatrix}1 \\ 0\end{pmatrix}$$

so that $\mathbf{Y}$ is a $n \times 2$ matrix, $\mathbf{1}$ is an $n \times 1$ vector of ones, and $\mathbf{R}$ is a 2×1 vector (in bigger settings, you would need to make this a more complicated matrix). Given this notation, our moment conditions amount to

$$\mathbb{E}[Y - \mathbf{R}\alpha] = 0$$

and notice that we can write down the sample analogue of the moment conditions as

$$\frac{1}{n}\mathbf{Y}'\mathbf{1} - \mathbf{R}\alpha \approx 0$$

where we define $\alpha = ATT(2,2)$ (to keep the notation more concise below, and this is a scalar here), and the $\approx$ is because it is unlikely that this equation is exactly equal to 0.

Given some weighting matrix $\widehat{\mathbf{W}}$, we can estimate $ATT(2,2)$ by

$$\begin{aligned}\hat{\alpha} &= \underset{a}{\text{argmin}} \left(\frac{1}{n}\mathbf{Y}'\mathbf{1} - \mathbf{R}a\right)' \widehat{\mathbf{W}} \left(\frac{1}{n}\mathbf{Y}'\mathbf{1} - \mathbf{R}a\right) \\ &= \underset{a}{\text{argmin}} \frac{1}{n^2}\mathbf{1}'\mathbf{Y}\widehat{\mathbf{W}}\mathbf{Y}'\mathbf{1} - \frac{2}{n}a'\mathbf{R}'\widehat{\mathbf{W}}\mathbf{Y}'\mathbf{1} + a'\mathbf{R}'\widehat{\mathbf{W}}\mathbf{R}a\end{aligned}$$

Taking the first order condition (and slightly re-arranging terms), we have that

$$\begin{aligned}\mathbf{R}'\widehat{\mathbf{W}}\frac{1}{n}\mathbf{Y}'\mathbf{1} &= \mathbf{R}'\widehat{\mathbf{W}}\mathbf{R}\hat{\alpha} \\ \implies \hat{\alpha} &= \left(\mathbf{R}'\widehat{\mathbf{W}}\mathbf{R}\right)^{-1}\mathbf{R}'\widehat{\mathbf{W}}\frac{1}{n}\mathbf{Y}'\mathbf{1}\end{aligned}$$

If you take $\widehat{\mathbf{W}} = \mathbf{I}_2$ (which is a natural choice...at least to start with), notice that $\mathbf{R}'\widehat{\mathbf{W}}\mathbf{R} = \mathbf{R}'\mathbf{R} = 1$. And,

$$\mathbf{R}'\widehat{\mathbf{W}}\frac{1}{n}\mathbf{Y}'\mathbf{1} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\mathbb{1}\{G_i = 2\}}{p_2} - \frac{U_i}{p_U}\right)\Delta Y_{i2}$$

i.e., due to the nature of $\mathbf{R}$ and $\widehat{\mathbf{W}}$ being the identity matrix, we estimate $ATT(2,2)$ as the average path of outcomes for group 2 relative to the untreated group. It's interesting that we get this simplification (and perhaps surprising) — group 3 played no role here even though, under the parallel trends assumption, it seems as though we could have used group 3 as the comparison group. It turns out that the reason we are getting this result comes down to our choice of weighting matrix.

Let's now move to computing the efficient GMM estimator and see what happens. To do this, we need to first compute an estimate of the variance of the moment conditions. Given the above moment conditions, the population version of the variance is $\mathbf{\Omega} = \mathbb{E}[(Y - \mathbf{R}\alpha)(Y - \mathbf{R}\alpha)']$, and, given our preliminary estimate $\hat{\alpha}$, we can estimate $\mathbf{\Omega}$ by

$$\hat{\mathbf{\Omega}} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mathbf{R}\hat{\alpha})(Y_i - \mathbf{R}\hat{\alpha})' = \frac{1}{n}\hat{\mathbf{e}}'\hat{\mathbf{e}}$$

where $\hat{\mathbf{e}} := \mathbf{Y} - \mathbf{1} \otimes \hat{\alpha}'\mathbf{R}'$ (which just subtracts $\hat{\alpha}'\mathbf{R}'$ from each row of $\mathbf{Y}$; "$\otimes$" is the Kronecker product which we have used a handful of times before. To be clear, this is just the data matrix

version of calculating $Y_i - \mathbf{R}\hat{\alpha}$ for all n units.). Then, we can compute

$$\hat{\alpha}^o_{gmm} = \left(\mathbf{R}'\hat{\mathbf{\Omega}}^{-1}\mathbf{R}\right)^{-1}\mathbf{R}'\hat{\mathbf{\Omega}}^{-1}\frac{1}{n}\mathbf{Y}'\mathbf{1}$$

This is the efficient GMM estimate of $ATT(2,2)$. In this case, $ATT(2,2)$ combines information from both $G = 3$ and $U = 1$ for the path of outcomes absent the treatment.

Finally, in order to be able to conduct inference, let's derive the asymptotic distribution of $\sqrt{n}(\hat{\alpha} - \alpha)$. I'll focus on the case with a general weighting matrix $\widehat{\mathbf{W}} \xrightarrow{p} \mathbf{W}$, but the arguments immediately specialize to the efficient GMM case discussed above. Given the expression for $\hat{\alpha}$ above, we can re-write it as

$$\begin{aligned}
\hat{\alpha} &= \left(\mathbf{R}'\widehat{\mathbf{W}}\mathbf{R}\right)^{-1}\mathbf{R}'\widehat{\mathbf{W}}\frac{1}{n}\sum_{i=1}^{n} Y_i \\
&= \left(\mathbf{R}'\widehat{\mathbf{W}}\mathbf{R}\right)^{-1}\mathbf{R}'\widehat{\mathbf{W}}\frac{1}{n}\sum_{i=1}^{n}(\mathbf{R}\alpha + e_i) \\
&= \alpha + \left(\mathbf{R}'\widehat{\mathbf{W}}\mathbf{R}\right)^{-1}\mathbf{R}'\widehat{\mathbf{W}}\frac{1}{n}\sum_{i=1}^{n} e_i
\end{aligned}$$

where (as is implicit in the discussion above), we define $e_i := Y_i - \mathbf{R}\alpha$. This implies that

$$\begin{aligned}
\sqrt{n}(\hat{\alpha} - \alpha) &= \left(\mathbf{R}'\widehat{\mathbf{W}}\mathbf{R}\right)^{-1}\mathbf{R}'\widehat{\mathbf{W}}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} e_i \\
&= \left(\mathbf{R}'\mathbf{W}\mathbf{R}\right)^{-1}\mathbf{R}'\mathbf{W}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} e_i + o_p(1) \\
&\xrightarrow{d} \mathcal{N}(0, \mathbf{V})
\end{aligned}$$

where we define (as above) $\mathbf{\Omega} = \mathbb{E}[ee'] = \mathbb{E}[(Y - \mathbf{R}\alpha)(Y - \mathbf{R}\alpha)']$, and

$$\mathbf{V} = \left(\mathbf{R}'\mathbf{W}\mathbf{R}\right)^{-1}\mathbf{R}'\mathbf{W}\mathbf{\Omega}\mathbf{W}\left(\mathbf{R}'\mathbf{W}\mathbf{R}\right)^{-1}$$

In the case where we choose the optimal weighting matrix, $\widehat{\mathbf{W}} = \hat{\mathbf{\Omega}}^{-1}$, the asymptotic variance matrix simplifies to

$$\mathbf{V}_0 = \left(\mathbf{R}'\mathbf{\Omega}^{-1}\mathbf{R}\right)^{-1}$$

and the natural way to estimate it is by

$$\hat{\mathbf{V}}_0 = \left(\mathbf{R}'\hat{\mathbf{\Omega}}^{-1}\mathbf{R}\right)^{-1}$$

Earlier, we discussed estimating $\mathbf{\Omega}$. In practice, I think that, for inference, you should re-compute the residuals using $\hat{\mathbf{e}} := \mathbf{Y} - \mathbf{1} \otimes \hat{\alpha}^{o'}_{gmm}\mathbf{R}'$ (relative to the previous discussion, here we use $\hat{\alpha}^o_{gmm}$

rather than the preliminary estimate $\hat{\alpha}$.). Given these re-computed residuals, we can estimate $\hat{\mathbf{\Omega}} = \frac{1}{n}\hat{\mathbf{e}}'\hat{\mathbf{e}}$. Finally, given that we have an estimate of $\mathbf{V}_0$, we can compute our whole set of test statistics, confidence intervals, etc. which we have done many times before.