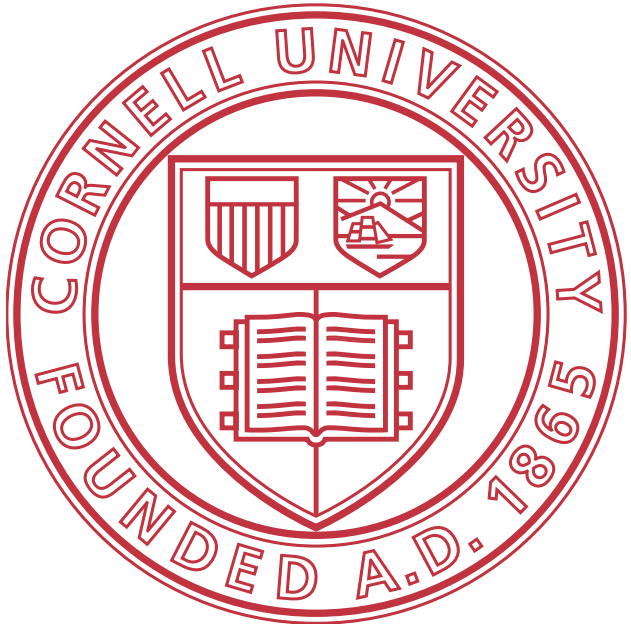
ECON 6200: Econometrics II



Generalized Method of Moments

Introduction

The Generalized Method of Moments (GMM) and extremely close relatives like Method of Simulated Moments or Indirect Inference are of great importance in applied work.

The name and clear statement of the method are due to Hansen (1982) and were cited in his Nobel prize.

There were precursors for several aspects of the theory, including the Sargan test statistic or work on nonlinear least squares and optimal instruments by Amemiya.

We will now develop a fairly general theory, though we restrict attention to linear moment conditions. Nonlinear GMM will be developed as special case of extremum estimators.

Generalized Method of Moments

What is GMM?

The Generalized Method of Moments can be thought of as generalizing and extending TSLS in several ways:

- Since we cannot set sample moment conditions exactly to zero, we must commit to a norm that we minimize. Is there a best norm?
- We allow for heteroskedasticity also in the estimation stage. (Heteroskedasticity robust standard errors for TSLS are straightforward and standardly used, but we will see that in the estimation stage, TSLS can be argued to presume homoskedasticity.)
- We consider testing instrument validity.
- It will be clear that the restriction to linear moment conditions simplifies the development but is not essential.

Compared to the Method of Moments, the modifier "Generalized" refers to the fact that we allow for, and carefully explore the implications of, overidentification, i.e. more moment conditions than parameters.

Generalized Method of Moments

Formal Statement

We know that

$$\mathbb{E}g(Y, X, Z; \theta) = 0,$$

where $\theta \in \mathbb{R}^k$ and $g(\cdot)$ is a known smooth function mapping into $\Re^\ell$, $\ell \geq k$.

The case of $\ell > k$ will be called overidentified.

We will assume $g(\cdot)$ is linear. That is not essential.

It yields OLS, IV, TSLS, and (after generalizing to multiple outcomes) SURE and simple panel data estimators as special cases.

Generalized Method of Moments

Formal Statement

We know that

$$\mathbb{E}g(Y, X, Z; \theta) = 0,$$

where $\theta \in \mathbb{R}^k$ and $g(\cdot)$ is a known smooth function mapping into $\mathfrak{R}^\ell$, $\ell \geq k$.

The case of $\ell > k$ will be called overidentified.

We will assume $g(\cdot)$ is linear. That is not essential.

It yields OLS, IV, TSLS, and (after generalizing to multiple outcomes) SURE and simple panel data estimators as special cases.

But for future reference, consider also:

- Probit: $\mathbb{E}\big(X(Y - \Phi(X'\beta))\big) = 0$.
- Best-response conditions, e.g. Euler equations:
  These are the original application of GMM (Hansen-Singleton 1983).

Generalized Method of Moments

Formal Statement

The GMM estimator is

$$\begin{aligned} \hat{\theta}(\boldsymbol{W}) &= \arg\min_{\theta} J_n(\theta) \\ J_n(\theta) &= n\overline{g}_n(\theta)'\boldsymbol{W}\overline{g}_n(\theta) \\ \overline{g}_n(\theta) &\equiv \frac{1}{n}\sum_{i=1}^{n} g(\theta;\cdot). \end{aligned}$$

Generalized Method of Moments

Formal Statement

The GMM estimator is

$$\begin{aligned} \hat{\theta}(\boldsymbol{W}) &= \arg\min_{\theta} J_n(\theta) \\ J_n(\theta) &= n\overline{g}_n(\theta)'\boldsymbol{W}\overline{g}_n(\theta) \\ \overline{g}_n(\theta) &\equiv \frac{1}{n}\sum_{i=1}^{n} g(\theta;\cdot). \end{aligned}$$

- $\boldsymbol{W}$ is a weight matrix and therefore defines the norm that we minimize. (Recall that, if we are overidentified, we cannot set $J_n(\theta) = 0$.)
- Note that we therefore really have a family of estimators. We will later think about optimal choice of $\boldsymbol{W}$.
- The scale factor in $J_n(\cdot)$ is for convenience as this criterion will converge to a nondegenerate limit.

Generalized Method of Moments

Linear Case

We next specialize to the linear case

$$g(\beta, \cdot) = Z(Y - X'\beta).$$

This covers all estimators we saw so far (we might have $Z = X$).

In data matrix notation, the estimator then minimizes

$$(\boldsymbol{Z'Y} - \boldsymbol{Z'X}\beta)'\boldsymbol{W}(\boldsymbol{Z'Y} - \boldsymbol{Z'X}\beta).$$

We can crank out the FOC:

$$\begin{aligned} -2\boldsymbol{X'ZW}(\boldsymbol{Z'Y} - \boldsymbol{Z'X}\hat{\beta}) &= 0 \\ \implies \boldsymbol{X'ZWZ'X}\hat{\beta} &= \boldsymbol{X'ZWZ'Y} \\ \implies \hat{\beta} &= (\boldsymbol{X'ZWZ'X})^{-1}\boldsymbol{X'ZWZ'Y} \\ [&= (\boldsymbol{S'_{xz}WS_{xz}})^{-1}\boldsymbol{S'_{xz}Ws_{xy}}] \end{aligned}$$

(the last line is a restatement in notation from Hayashi's textbook).

Relation to TSLS and Optimal Weights

Compare

$$\begin{aligned}\hat{\beta}_{GMM}(\boldsymbol{W}) &= (\boldsymbol{X}'\boldsymbol{Z}\boldsymbol{W}\boldsymbol{Z}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Z}\boldsymbol{W}\boldsymbol{Z}'\boldsymbol{Y} \\ \hat{\beta}_{TSLS} &= (\boldsymbol{X}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{Y}.\end{aligned}$$

The TSLS estimator is the GMM estimator with weights $\boldsymbol{W} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}$.

Relation to TSLS and Optimal Weights

Compare

$$\begin{aligned}\hat{\beta}_{GMM}(\boldsymbol{W}) &= (\boldsymbol{X}'\boldsymbol{Z}\boldsymbol{W}\boldsymbol{Z}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Z}\boldsymbol{W}\boldsymbol{Z}'\boldsymbol{Y} \\ \hat{\beta}_{TSLS} &= (\boldsymbol{X}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{Y}.\end{aligned}$$

The TSLS estimator is the GMM estimator with weights $\boldsymbol{W} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}$.

This raises questions:

- Since we have a family of weights, is there a "best" weight?
- When, if ever, is that weight $(\boldsymbol{Z}'\boldsymbol{Z})^{-1}$?

Generalized Method of Moments

Relation to TSLS and Optimal Weights

Compare

$$\begin{aligned}\hat{\beta}_{GMM}(\boldsymbol{W}) &= (\boldsymbol{X}'\boldsymbol{Z}\boldsymbol{W}\boldsymbol{Z}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Z}\boldsymbol{W}\boldsymbol{Z}'\boldsymbol{Y} \\ \hat{\beta}_{TSLS} &= (\boldsymbol{X}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{Y}.\end{aligned}$$

The TSLS estimator is the GMM estimator with weights $\boldsymbol{W} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}$.

This raises questions:

- Since we have a family of weights, is there a "best" weight?
- When, if ever, is that weight $(\boldsymbol{Z}'\boldsymbol{Z})^{-1}$?

To see the answers, it is helpful to further consider the WLS analogy.

What are the "right" weights in WLS?

With i.i.d. data, they are the inverse standard deviation.

More generally, the ideal $\boldsymbol{W}$ is the inverse variance-covariance matrix of errors ε.

Generalized Method of Moments

This raises the idea of estimating the optimal $\boldsymbol{W}$.

- Compute a preliminary estimate of β to back out residuals.
- Use these to estimate the optimal weight matrix.
- Report a final estimate $\hat{\beta}_{GMM}(\hat{\boldsymbol{W}})$.

Next steps:

- We will show that this procedure (known as "efficient" or two-step GMM) minimizes asymptotic variance.
- It will turn out that the "TSLS weights" $\boldsymbol{W} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}$ are optimal if we assume homoskedasticity.

Generalized Method of Moments

GMM Assumptions

1. We observe i.i.d. realizations (Y_i, X_i, Z_i), $i = 1, \ldots, n$.
2. $\mathbb{E}(Z(Y - X'\beta)) = 0$.
3. $\mathbb{E}(|Y^4|) < \infty$,
4. $\mathbb{E}(\|X\|^4) < \infty$,
5. $\mathbb{E}(\|Z\|^4) < \infty$,
6. $\boldsymbol{Q} \equiv \mathbb{E}(ZX')$ has full rank k,
7. $\boldsymbol{W}$ is positive definite,
8. $\Omega \equiv \mathbb{E}(ZZ'\varepsilon^2)$ is positive definite.

These are also the assumptions for TSLS, which will emerge as special case.

Asymptotic Distribution

The following algebra generalizes previous algebra for OLS:

$$\begin{aligned}
\hat{\beta}_{GMM}(\boldsymbol{W}) &= \left(\boldsymbol{X}'\boldsymbol{Z}\boldsymbol{W}\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Z}\boldsymbol{W}\boldsymbol{Z}'\boldsymbol{Y} \\
&= \beta + \left(\boldsymbol{X}'\boldsymbol{Z}\boldsymbol{W}\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Z}\boldsymbol{W}\boldsymbol{Z}'\varepsilon \\
&= \beta + \left(\tfrac{1}{n}\boldsymbol{X}'\boldsymbol{Z}\boldsymbol{W}\tfrac{1}{n}\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\tfrac{1}{n}\boldsymbol{X}'\boldsymbol{Z}\boldsymbol{W}\tfrac{1}{n}\boldsymbol{Z}'\varepsilon \\
&= \beta + \left(\mathbb{E}(XZ')\boldsymbol{W}\mathbb{E}(ZX')\right)^{-1}\mathbb{E}(XZ')\boldsymbol{W}\tfrac{1}{n}\boldsymbol{Z}'\varepsilon + o_P(1) \\
&= \beta + \left(\boldsymbol{Q}'\boldsymbol{W}\boldsymbol{Q}\right)^{-1}\boldsymbol{Q}'\boldsymbol{W}\tfrac{1}{n}\boldsymbol{Z}'\varepsilon + o_P(1)
\end{aligned}$$

Asymptotic Distribution

The following algebra generalizes previous algebra for OLS:

$$\begin{aligned}
\hat{\beta}_{GMM}(\boldsymbol{W}) &= (\boldsymbol{X}'\boldsymbol{Z}\boldsymbol{W}\boldsymbol{Z}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Z}\boldsymbol{W}\boldsymbol{Z}'\boldsymbol{Y} \\
&= \beta + (\boldsymbol{X}'\boldsymbol{Z}\boldsymbol{W}\boldsymbol{Z}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Z}\boldsymbol{W}\boldsymbol{Z}'\varepsilon \\
&= \beta + (\tfrac{1}{n}\boldsymbol{X}'\boldsymbol{Z}\boldsymbol{W}\tfrac{1}{n}\boldsymbol{Z}'\boldsymbol{X})^{-1}\tfrac{1}{n}\boldsymbol{X}'\boldsymbol{Z}\boldsymbol{W}\tfrac{1}{n}\boldsymbol{Z}'\varepsilon \\
&= \beta + (\mathbb{E}(XZ')\boldsymbol{W}\mathbb{E}(ZX'))^{-1}\mathbb{E}(XZ')\boldsymbol{W}\tfrac{1}{n}\boldsymbol{Z}'\varepsilon + o_P(1) \\
&= \beta + (\boldsymbol{Q}'\boldsymbol{W}\boldsymbol{Q})^{-1}\boldsymbol{Q}'\boldsymbol{W}\tfrac{1}{n}\boldsymbol{Z}'\varepsilon + o_P(1)
\end{aligned}$$

$$\begin{aligned}
\implies \hat{\beta}_{GMM}(\boldsymbol{W}) - \beta &\xrightarrow{p} 0, \\
\sqrt{n}(\hat{\beta}_{GMM}(\boldsymbol{W}) - \beta) &\xrightarrow{d} N\big(0, (\boldsymbol{Q}'\boldsymbol{W}\boldsymbol{Q})^{-1}\boldsymbol{Q}'\boldsymbol{W}\Omega\boldsymbol{W}\boldsymbol{Q}(\boldsymbol{Q}'\boldsymbol{W}\boldsymbol{Q})^{-1}\big).
\end{aligned}$$

Generalized Method of Moments

Asymptotic Distribution

The following algebra generalizes previous algebra for OLS:

$$\begin{aligned}
\hat{\beta}_{GMM}(\boldsymbol{W}) &= (\boldsymbol{X}'\boldsymbol{Z}\boldsymbol{W}\boldsymbol{Z}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Z}\boldsymbol{W}\boldsymbol{Z}'\boldsymbol{Y} \\
&= \beta + (\boldsymbol{X}'\boldsymbol{Z}\boldsymbol{W}\boldsymbol{Z}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Z}\boldsymbol{W}\boldsymbol{Z}'\varepsilon \\
&= \beta + (\tfrac{1}{n}\boldsymbol{X}'\boldsymbol{Z}\boldsymbol{W}\tfrac{1}{n}\boldsymbol{Z}'\boldsymbol{X})^{-1}\tfrac{1}{n}\boldsymbol{X}'\boldsymbol{Z}\boldsymbol{W}\tfrac{1}{n}\boldsymbol{Z}'\varepsilon \\
&= \beta + (\mathbb{E}(XZ')\boldsymbol{W}\mathbb{E}(ZX'))^{-1}\mathbb{E}(XZ')\boldsymbol{W}\tfrac{1}{n}\boldsymbol{Z}'\varepsilon + o_P(1) \\
&= \beta + (\boldsymbol{Q}'\boldsymbol{W}\boldsymbol{Q})^{-1}\boldsymbol{Q}'\boldsymbol{W}\tfrac{1}{n}\boldsymbol{Z}'\varepsilon + o_P(1)
\end{aligned}$$

$$\begin{aligned}
\implies \hat{\beta}_{GMM}(\boldsymbol{W}) - \beta &\xrightarrow{p} 0, \\
\sqrt{n}(\hat{\beta}_{GMM}(\boldsymbol{W}) - \beta) &\xrightarrow{d} N\big(0, (\boldsymbol{Q}'\boldsymbol{W}\boldsymbol{Q})^{-1}\boldsymbol{Q}'\boldsymbol{W}\Omega\boldsymbol{W}\boldsymbol{Q}(\boldsymbol{Q}'\boldsymbol{W}\boldsymbol{Q})^{-1}\big).
\end{aligned}$$

Similarly to before, the consistency result does not use all the assumptions, for example it only requires 2^{nd} moments. See book for details.

Generalized Method of Moments

Asymptotic Efficiency

- The matrix $\Omega = \mathbb{E}(ZZ'\varepsilon^2)$ is really the variance-covariance matrix of moment conditions.
- Intuitively, a condition with higher "own-variance" according to Ω is noisier.
- Indeed, in our WLS analogy from before, Ω parameterizes the heteroskedasticty in our fictitious rgeression with ℓ observations and k parameters.
- This suggests Ω^{-1} as efficient weighting matrix.
- Of course, we do not know Ω, but we can estimate it using residuals from a preliminary regression.
- Let's assume we can do this. Furthermore, it is easy to see in preceding algebra that results go through if $\boldsymbol{W}$ is a (possibly stochastic) sequence $\hat{\boldsymbol{W}} \xrightarrow{p} \boldsymbol{W}$.

Generalized Method of Moments

Let $\hat{\boldsymbol{W}} \xrightarrow{p} \boldsymbol{W}^* \equiv \Omega^{-1}$, then:

1. The asymptotic variance becomes

$$\boldsymbol{V} = (\boldsymbol{Q}'\boldsymbol{W}^*\boldsymbol{Q})^{-1}\boldsymbol{Q}'\boldsymbol{W}^*\Omega\boldsymbol{W}^*\boldsymbol{Q}(\boldsymbol{Q}'\boldsymbol{W}^*\boldsymbol{Q})^{-1}$$

   and simplifies to

$$\boldsymbol{V}^* = (\boldsymbol{Q}'\Omega^{-1}\boldsymbol{Q})^{-1}.$$

2. The latter is the best asymptotic variance: $\boldsymbol{V} \geq \boldsymbol{V}^*$ (i.e., $\boldsymbol{V} - \boldsymbol{V}^*$ is psd).
3. It is **only** attained by estimators that are asymptotically equivalent to $\hat{\beta}(\Omega^{-1})$.

Generalized Method of Moments

Let $\hat{\boldsymbol{W}} \xrightarrow{p} \boldsymbol{W}^* \equiv \Omega^{-1}$, then:

1. The asymptotic variance becomes

$$\boldsymbol{V} = (\boldsymbol{Q}'\boldsymbol{W}^*\boldsymbol{Q})^{-1}\boldsymbol{Q}'\boldsymbol{W}^*\Omega\boldsymbol{W}^*\boldsymbol{Q}(\boldsymbol{Q}'\boldsymbol{W}^*\boldsymbol{Q})^{-1}$$

   and simplifies to

$$\boldsymbol{V}^* = (\boldsymbol{Q}'\Omega^{-1}\boldsymbol{Q})^{-1}.$$

2. The latter is the best asymptotic variance: $\boldsymbol{V} \geq \boldsymbol{V}^*$ (i.e., $\boldsymbol{V} - \boldsymbol{V}^*$ is psd).
3. It is **only** attained by estimators that are asymptotically equivalent to $\hat{\beta}(\Omega^{-1})$.

This motivates the efficient ("two stage") GMM estimator

$$\begin{aligned} \hat{\beta}_{TSGMM} &\equiv \hat{\beta}(\hat{\boldsymbol{W}}) \\ \hat{\boldsymbol{W}} &\equiv \left(\mathbb{E}_n(ZZ'\hat{\varepsilon}^2)\right)^{-1} \\ \hat{\varepsilon} &= Y - X\hat{\beta}, \end{aligned}$$

where $\hat{\beta}$ is any consistent estimator of β, for example a GMM estimator with any reasonable weighting matrix. The industry standard is TSLS.

Generalized Method of Moments

Let $\hat{\boldsymbol{W}} \xrightarrow{p} \boldsymbol{W}^* \equiv \Omega^{-1}$, then:

1. The asymptotic variance becomes

$$\boldsymbol{V} = (\boldsymbol{Q}'\boldsymbol{W}^*\boldsymbol{Q})^{-1}\boldsymbol{Q}'\boldsymbol{W}^*\Omega\boldsymbol{W}^*\boldsymbol{Q}(\boldsymbol{Q}'\boldsymbol{W}^*\boldsymbol{Q})^{-1}$$

   and simplifies to

$$\boldsymbol{V}^* = (\boldsymbol{Q}'\Omega^{-1}\boldsymbol{Q})^{-1}.$$

2. The latter is the best asymptotic variance: $\boldsymbol{V} \geq \boldsymbol{V}^*$ (i.e., $\boldsymbol{V} - \boldsymbol{V}^*$ is psd).
3. It is **only** attained by estimators that are asymptotically equivalent to $\hat{\beta}(\Omega^{-1})$.

Note: While the previous implementation is the most popular and a frequent default, one could also use the **centered estimator**

$$\hat{\boldsymbol{W}} = \mathbb{E}_n\big((Z\hat{\varepsilon} - \mathbb{E}_n(Z\hat{\varepsilon}))(Z\hat{\varepsilon} - \mathbb{E}_n(Z\hat{\varepsilon}))'\big).$$

This literally estimates the variance, as opposed to the uncentered second moment, of $Z\varepsilon$.

Of course, the two are the same if $\mathbb{E}(Z\varepsilon) = 0$.

But the centered estimator is consistent for the variance even if the model is misspecified.

Proof of Efficiency

1. With efficient weighting, $\boldsymbol{V}$ simplifies to $\boldsymbol{V}^*$.
2. $\boldsymbol{V} \geq \boldsymbol{V}^*$ (i.e., $\boldsymbol{V} - \boldsymbol{V}^*$ is psd).
3. The inequality is strict unless the estimators are (asymptotically) equivalent.

To prove this, write

$$\boldsymbol{V} = \boldsymbol{A}'\Omega\boldsymbol{A}, \text{ where } \boldsymbol{A} = \boldsymbol{W}\boldsymbol{Q}(\boldsymbol{Q}'\boldsymbol{W}\boldsymbol{Q})^{-1}$$
$$\boldsymbol{V}^* = \boldsymbol{B}'\Omega\boldsymbol{B}, \text{ where } \boldsymbol{B} = \Omega^{-1}\boldsymbol{Q}(\boldsymbol{Q}'\Omega^{-1}\boldsymbol{Q})^{-1}$$

and observe that

$$\boldsymbol{B}'\Omega\boldsymbol{A} = (\boldsymbol{Q}'\Omega^{-1}\boldsymbol{Q})^{-1}\boldsymbol{Q}'\Omega^{-1}\Omega\boldsymbol{W}\boldsymbol{Q}(\boldsymbol{Q}'\boldsymbol{W}\boldsymbol{Q})^{-1} = \boldsymbol{V}^* = \boldsymbol{B}'\Omega\boldsymbol{B}$$
$$\implies \boldsymbol{B}'\Omega(\boldsymbol{A} - \boldsymbol{B}) = 0.$$

Thus,

$$\begin{aligned}\boldsymbol{V} = \boldsymbol{A}'\Omega\boldsymbol{A} &= (\boldsymbol{B} + (\boldsymbol{A} - \boldsymbol{B}))'\Omega(\boldsymbol{B} + (\boldsymbol{A} - \boldsymbol{B})) \\ &= \underbrace{\boldsymbol{B}'\Omega\boldsymbol{B}}_{=\boldsymbol{V}^*} + \underbrace{(\boldsymbol{A} - \boldsymbol{B})'\Omega\boldsymbol{B}}_{=0} + \underbrace{\boldsymbol{B}'\Omega(\boldsymbol{A} - \boldsymbol{B})}_{=0} + \underbrace{(\boldsymbol{A} - \boldsymbol{B})'\Omega(\boldsymbol{A} - \boldsymbol{B})}_{p.s.d.}.\end{aligned}$$

Simplification under Homoskedasticity

Assume next homoskedasticity, i.e. that

$$\mathbb{E}(\varepsilon^2|Z) = \sigma^2 \implies \Omega = \mathbb{E}(ZZ'\varepsilon^2) = \sigma^2\mathbb{E}(ZZ').$$

The estimator with ideal weighting matrix simplifies:

$$\begin{aligned}\hat{\beta}_{GMM}(\Omega^{-1}) &= \left(\boldsymbol{X}'\boldsymbol{Z}\sigma^{-2}(\mathbb{E}ZZ')^{-1}\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Z}\sigma^{-2}(\mathbb{E}ZZ')^{-1}\boldsymbol{Z}'\boldsymbol{Y} \\ &= \left(\boldsymbol{X}'\boldsymbol{Z}(\mathbb{E}ZZ')^{-1}\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Z}(\mathbb{E}ZZ')^{-1}\boldsymbol{Z}'\boldsymbol{Y}.\end{aligned}$$

But $\mathbb{E}(ZZ')$ can be estimated by $\mathbb{E}_n(ZZ') = \frac{1}{n}\boldsymbol{Z}'\boldsymbol{Z}$.

Because $\frac{1}{n}$ cancels from the expression, we can succinctly write the estimator as

$$\hat{\beta}_{GMM}(\Omega^{-1}) = \left(\boldsymbol{X}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{Y}$$

Simplification under Homoskedasticity

Assume next homoskedasticity, i.e. that

$$\mathbb{E}(\varepsilon^2|Z) = \sigma^2 \implies \Omega = \mathbb{E}(ZZ'\varepsilon^2) = \sigma^2\mathbb{E}(ZZ').$$

The estimator with ideal weighting matrix simplifies:

$$\begin{aligned}\hat{\beta}_{GMM}(\Omega^{-1}) &= \left(\boldsymbol{X}'\boldsymbol{Z}\sigma^{-2}(\mathbb{E}ZZ')^{-1}\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Z}\sigma^{-2}(\mathbb{E}ZZ')^{-1}\boldsymbol{Z}'\boldsymbol{Y} \\ &= \left(\boldsymbol{X}'\boldsymbol{Z}(\mathbb{E}ZZ')^{-1}\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Z}(\mathbb{E}ZZ')^{-1}\boldsymbol{Z}'\boldsymbol{Y}.\end{aligned}$$

But $\mathbb{E}(ZZ')$ can be estimated by $\mathbb{E}_n(ZZ') = \frac{1}{n}\boldsymbol{Z}'\boldsymbol{Z}$.

Because $\frac{1}{n}$ cancels from the expression, we can succinctly write the estimator as

$$\hat{\beta}_{GMM}(\Omega^{-1}) = \left(\boldsymbol{X}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{Y} = \hat{\beta}_{TSLS}.$$

The TSLS estimator is the efficient one under homoskedasticity!

Should we do Efficient GMM?

The case for efficient GMM is more compelling than for FGLS.

In addition, some results (notably specification testing) require efficient GMM.

Therefore, efficient GMM is common in practice.

However:

- In finite sample, estimating Ω introduces noise. Monte Carlo simulations suggest that, with small samples and moderate heteroskedasticity, TSLS may perform better.
- On the other hand, estimation of Ω could in principle be iterated, which (under conditions that at least asymptotically guarantee convergence) removes path dependency. Implementations for this exist.
  However, asymptotic analysis does not suggest a gain and one might be concerned about error propagation.

Continuous Updating GMM

For the record, with modern computing power one could also directly compute

$$\hat{\beta}_{cGMM} \equiv \arg\min_{\beta} \overline{g}_n(\beta)' \hat{\boldsymbol{W}}(\beta) \overline{g}(\beta),$$

i.e. optimize over criterion function and weighting matrix at the same time.

This is called Continuous Updating GMM.

It is not the numerically same estimator but has the same asymptotic distribution.

Should We Use All Instruments We Can Think of?

Also, now that we can use more instruments than regressors, we might be tempted to use all instruments we can think of, including "technical instruments" (polynomials of instruments etc.).

It is intutively clear that this is not right. But what are formal arguments against?

Should We Use All Instruments We Can Think of?

Also, now that we can use more instruments than regressors, we might be tempted to use all instruments we can think of, including "technical instruments" (polynomials of instruments etc.).

It is intutively clear that this is not right. But what are formal arguments against?

- Every instrument must be justified, so the "cost of assumptions" increases. Remember: Your conclusions are at most as credible as your assumptions!
- Estimating a larger weight matrix introduces more finite sample noise.
- Even without weighting, it can be shown that TSLS is inconsistent if there are many instruments in the sense that $\ell_n / n \to \alpha > 0$. For this, the instruments do not even need to be weak!
- Indeed, in practice, the previous point means our asymptotic analysis should only be invoked if $n \gg \ell$, limiting (although maybe not very tightly) the number of instruments that we can use.

Overidentification Test

What is qualitatively new in overidentified ($\ell > k$) models is that the model itself can be tested.

In other words, contrary to IV, we can test (joint) validity of moments. Why?

Overidentification Test

What is qualitatively new in overidentified ($\ell > k$) models is that the model itself can be tested.

In other words, contrary to IV, we can test (joint) validity of moments. Why?

Theorem

Under above assumptions,

$$J_n \equiv J(\hat{\beta}_{TSGMM}) \xrightarrow{d} \chi^2_{\ell-k}.$$

Overidentification Test

What is qualitatively new in overidentified ($\ell > k$) models is that the model itself can be tested.

In other words, contrary to IV, we can test (joint) validity of moments. Why?

Theorem

Under above assumptions,

$$J_n \equiv J(\hat{\beta}_{TSGMM}) \xrightarrow{d} \chi^2_{\ell-k}.$$

Intuition

- We try to set an ℓ-vector to zero but have only k free parameters to do so.
- This means we have a residual with $\ell - k$ degrees of freedom.
- However, if the model is well-specified, this residual is of order $O(n^{-1/2})$.
- If (and only if!) we use the efficient weighting matrix, it is furthermore asymptotically MVSN in a certain $(\ell - k)$-subspace.
- Then its square is $\chi^2_{\ell-k}$.

Overidentification Test

Proof of Theorem

Note: Since some expression got long, in this proof I use $\approx$ to denote that an $o_P(1)$-term was dropped.

Previous results imply that $\frac{1}{n}\boldsymbol{Z}'\hat{\varepsilon} = O_P(n^{-1/2})$, thus write

$$\begin{aligned} J_n &= n\left(\tfrac{1}{n}\boldsymbol{Z}'\hat{\varepsilon}\right)'\hat{\Omega}^{-1}\left(\tfrac{1}{n}\boldsymbol{Z}'\hat{\varepsilon}\right) \\ &\approx n\left(\tfrac{1}{n}\boldsymbol{Z}'\hat{\varepsilon}\right)'\Omega^{-1}\left(\tfrac{1}{n}\boldsymbol{Z}'\hat{\varepsilon}\right) \\ &= n\left(\boldsymbol{C}'\tfrac{1}{n}\boldsymbol{Z}'\hat{\varepsilon}\right)'(\boldsymbol{C}'\Omega\boldsymbol{C})^{-1}\left(\boldsymbol{C}'\tfrac{1}{n}\boldsymbol{Z}'\hat{\varepsilon}\right) \\ &= n\left(\boldsymbol{C}'\tfrac{1}{n}\boldsymbol{Z}'\hat{\varepsilon}\right)'\left(\boldsymbol{C}'\tfrac{1}{n}\boldsymbol{Z}'\hat{\varepsilon}\right), \end{aligned}$$

where $\Omega^{-1} = \boldsymbol{C}\boldsymbol{C}' \Leftrightarrow \Omega = (\boldsymbol{C}')^{-1}\boldsymbol{C}^{-1}$, e.g. $\boldsymbol{C}$ is the Cholesky root of Ω^{-1}.

Overidentification Test

Proof of Theorem (ctd.)

Recall $J_n \approx n\left(\boldsymbol{C}'\frac{1}{n}\boldsymbol{Z}'\hat{\varepsilon}\right)'\left(\boldsymbol{C}'\frac{1}{n}\boldsymbol{Z}'\hat{\varepsilon}\right)$. Next,

$$
\begin{aligned}
& \boldsymbol{C}'\tfrac{1}{n}\boldsymbol{Z}'\hat{\varepsilon} \\
= \; & \boldsymbol{C}'\tfrac{1}{n}\boldsymbol{Z}'(\varepsilon - \boldsymbol{X}(\hat{\beta}-\beta)) \\
= \; & \boldsymbol{C}'\tfrac{1}{n}\boldsymbol{Z}'\left(\varepsilon - \boldsymbol{X}\left(\left(\tfrac{1}{n}\boldsymbol{X}'\boldsymbol{Z}\right)\hat{\Omega}^{-1}\left(\tfrac{1}{n}\boldsymbol{Z}'\boldsymbol{X}\right)\right)^{-1}\left(\tfrac{1}{n}\boldsymbol{X}'\boldsymbol{Z}\right)\hat{\Omega}^{-1}\left(\tfrac{1}{n}\boldsymbol{Z}'\varepsilon\right)\right)
\end{aligned}
$$

Overidentification Test

Proof of Theorem (ctd.)

Recall $J_n \approx n\left(\boldsymbol{C}'\frac{1}{n}\boldsymbol{Z}'\hat{\varepsilon}\right)'\left(\boldsymbol{C}'\frac{1}{n}\boldsymbol{Z}'\hat{\varepsilon}\right)$. Next,

$$
\begin{aligned}
& \boldsymbol{C}'\tfrac{1}{n}\boldsymbol{Z}'\hat{\varepsilon} \\
= \; & \boldsymbol{C}'\tfrac{1}{n}\boldsymbol{Z}'(\varepsilon - \boldsymbol{X}(\hat{\beta}-\beta)) \\
= \; & \boldsymbol{C}'\tfrac{1}{n}\boldsymbol{Z}'\left(\varepsilon - \boldsymbol{X}\left(\left(\tfrac{1}{n}\boldsymbol{X}'\boldsymbol{Z}\right)\hat{\Omega}^{-1}\left(\tfrac{1}{n}\boldsymbol{Z}'\boldsymbol{X}\right)\right)^{-1}\left(\tfrac{1}{n}\boldsymbol{X}'\boldsymbol{Z}\right)\hat{\Omega}^{-1}\left(\tfrac{1}{n}\boldsymbol{Z}'\varepsilon\right)\right) \\
= \; & \left(\boldsymbol{I}_\ell - \boldsymbol{C}'\left(\tfrac{1}{n}\boldsymbol{Z}'\boldsymbol{X}\right)\left(\left(\tfrac{1}{n}\boldsymbol{X}'\boldsymbol{Z}\right)\hat{\Omega}^{-1}\left(\tfrac{1}{n}\boldsymbol{Z}'\boldsymbol{X}\right)\right)^{-1}\left(\tfrac{1}{n}\boldsymbol{X}'\boldsymbol{Z}\right)\hat{\Omega}^{-1}(\boldsymbol{C}')^{-1}\right)\boldsymbol{C}'\tfrac{1}{n}\boldsymbol{Z}'\varepsilon
\end{aligned}
$$

Overidentification Test

Proof of Theorem (ctd.)

Recall $J_n \approx n\left(\boldsymbol{C}'\frac{1}{n}\boldsymbol{Z}'\hat{\varepsilon}\right)'\left(\boldsymbol{C}'\frac{1}{n}\boldsymbol{Z}'\hat{\varepsilon}\right)$. Next,

$$
\begin{aligned}
& \boldsymbol{C}'\tfrac{1}{n}\boldsymbol{Z}'\hat{\varepsilon} \\
= \;& \boldsymbol{C}'\tfrac{1}{n}\boldsymbol{Z}'(\varepsilon - \boldsymbol{X}(\hat{\beta}-\beta)) \\
= \;& \boldsymbol{C}'\tfrac{1}{n}\boldsymbol{Z}'\left(\varepsilon - \boldsymbol{X}\left(\left(\tfrac{1}{n}\boldsymbol{X}'\boldsymbol{Z}\right)\hat{\Omega}^{-1}\left(\tfrac{1}{n}\boldsymbol{Z}'\boldsymbol{X}\right)\right)^{-1}\left(\tfrac{1}{n}\boldsymbol{X}'\boldsymbol{Z}\right)\hat{\Omega}^{-1}\left(\tfrac{1}{n}\boldsymbol{Z}'\varepsilon\right)\right) \\
= \;& \left(\boldsymbol{I}_\ell - \boldsymbol{C}'\left(\tfrac{1}{n}\boldsymbol{Z}'\boldsymbol{X}\right)\left(\left(\tfrac{1}{n}\boldsymbol{X}'\boldsymbol{Z}\right)\hat{\Omega}^{-1}\left(\tfrac{1}{n}\boldsymbol{Z}'\boldsymbol{X}\right)\right)^{-1}\left(\tfrac{1}{n}\boldsymbol{X}'\boldsymbol{Z}\right)\hat{\Omega}^{-1}(\boldsymbol{C}')^{-1}\right)\boldsymbol{C}'\tfrac{1}{n}\boldsymbol{Z}'\varepsilon \\
\approx \;& \left(\boldsymbol{I}_\ell - \underbrace{\boldsymbol{C}'\left(\tfrac{1}{n}\boldsymbol{Z}'\boldsymbol{X}\right)}_{\equiv\hat{\boldsymbol{R}}}\left(\left(\tfrac{1}{n}\boldsymbol{X}'\boldsymbol{Z}\right)\boldsymbol{C}\boldsymbol{C}'\left(\tfrac{1}{n}\boldsymbol{Z}'\boldsymbol{X}\right)\right)^{-1}\left(\tfrac{1}{n}\boldsymbol{X}'\boldsymbol{Z}\right)\boldsymbol{C}\boldsymbol{C}'(\boldsymbol{C}')^{-1}\right)\boldsymbol{C}'\tfrac{1}{n}\boldsymbol{Z}'\varepsilon
\end{aligned}
$$

Overidentification Test

Proof of Theorem (ctd.)

Recall $J_n \approx n\left(\boldsymbol{C}'\frac{1}{n}\boldsymbol{Z}'\hat{\varepsilon}\right)'\left(\boldsymbol{C}'\frac{1}{n}\boldsymbol{Z}'\hat{\varepsilon}\right)$. Next,

$$
\begin{aligned}
& \boldsymbol{C}'\tfrac{1}{n}\boldsymbol{Z}'\hat{\varepsilon} \\
= \;& \boldsymbol{C}'\tfrac{1}{n}\boldsymbol{Z}'(\varepsilon - \boldsymbol{X}(\hat{\beta}-\beta)) \\
= \;& \boldsymbol{C}'\tfrac{1}{n}\boldsymbol{Z}'\left(\varepsilon - \boldsymbol{X}\left(\left(\tfrac{1}{n}\boldsymbol{X}'\boldsymbol{Z}\right)\hat{\Omega}^{-1}\left(\tfrac{1}{n}\boldsymbol{Z}'\boldsymbol{X}\right)\right)^{-1}\left(\tfrac{1}{n}\boldsymbol{X}'\boldsymbol{Z}\right)\hat{\Omega}^{-1}\left(\tfrac{1}{n}\boldsymbol{Z}'\varepsilon\right)\right) \\
= \;& \left(\boldsymbol{I}_\ell - \boldsymbol{C}'\left(\tfrac{1}{n}\boldsymbol{Z}'\boldsymbol{X}\right)\left(\left(\tfrac{1}{n}\boldsymbol{X}'\boldsymbol{Z}\right)\hat{\Omega}^{-1}\left(\tfrac{1}{n}\boldsymbol{Z}'\boldsymbol{X}\right)\right)^{-1}\left(\tfrac{1}{n}\boldsymbol{X}'\boldsymbol{Z}\right)\hat{\Omega}^{-1}(\boldsymbol{C}')^{-1}\right)\boldsymbol{C}'\tfrac{1}{n}\boldsymbol{Z}'\varepsilon \\
\approx \;& \left(\boldsymbol{I}_\ell - \underbrace{\boldsymbol{C}'\left(\tfrac{1}{n}\boldsymbol{Z}'\boldsymbol{X}\right)}_{\equiv\hat{\boldsymbol{R}}}\left(\left(\tfrac{1}{n}\boldsymbol{X}'\boldsymbol{Z}\right)\boldsymbol{C}\boldsymbol{C}'\left(\tfrac{1}{n}\boldsymbol{Z}'\boldsymbol{X}\right)\right)^{-1}\left(\tfrac{1}{n}\boldsymbol{X}'\boldsymbol{Z}\right)\boldsymbol{C}\boldsymbol{C}'(\boldsymbol{C}')^{-1}\right)\boldsymbol{C}'\tfrac{1}{n}\boldsymbol{Z}'\varepsilon \\
= \;& \left(\boldsymbol{I}_\ell - \hat{\boldsymbol{R}}(\hat{\boldsymbol{R}}'\hat{\boldsymbol{R}})^{-1}\hat{\boldsymbol{R}}'\right)\boldsymbol{C}'\tfrac{1}{n}\boldsymbol{Z}'\varepsilon \\
\approx \;& \left(\boldsymbol{I}_\ell - \boldsymbol{R}(\boldsymbol{R}'\boldsymbol{R})^{-1}\boldsymbol{R}'\right)\boldsymbol{C}'\tfrac{1}{n}\boldsymbol{Z}'\varepsilon, \text{ where } \boldsymbol{R} \equiv \boldsymbol{C}'\mathbb{E}(ZX').
\end{aligned}
$$

Overidentification Test

Proof of Theorem (ctd.)

We so far have

$$\begin{aligned} J_n &\approx n\left(\boldsymbol{C}'\tfrac{1}{n}\boldsymbol{Z}'\hat{\varepsilon}\right)'\left(\boldsymbol{C}'\tfrac{1}{n}\boldsymbol{Z}'\hat{\varepsilon}\right) \\ \boldsymbol{C}'\left(\tfrac{1}{n}\boldsymbol{Z}'\hat{\varepsilon}\right) &\approx \left(\boldsymbol{I}_\ell - \boldsymbol{R}(\boldsymbol{R}'\boldsymbol{R})^{-1}\boldsymbol{R}'\right)\boldsymbol{C}'\left(\tfrac{1}{n}\boldsymbol{Z}'\varepsilon\right) \end{aligned}$$

and also observe

$$\sqrt{n}\boldsymbol{C}'\left(\tfrac{1}{n}\boldsymbol{Z}'\varepsilon\right) \xrightarrow{d} N(0, \boldsymbol{C}'\Omega\boldsymbol{C}) = N(0, \boldsymbol{C}'(\boldsymbol{C}')^{-1}\boldsymbol{C}^{-1}\boldsymbol{C}) = N(0, \boldsymbol{I}_\ell).$$

Define the r.v. $u \sim N(0, \boldsymbol{I}_\ell)$, then it follows that

$$\begin{aligned} J_n &\xrightarrow{d} \left(\left(\boldsymbol{I}_\ell - \boldsymbol{R}(\boldsymbol{R}'\boldsymbol{R})^{-1}\boldsymbol{R}'\right)u\right)'\left(\boldsymbol{I}_\ell - \boldsymbol{R}(\boldsymbol{R}'\boldsymbol{R})^{-1}\boldsymbol{R}'\right)u \\ &\sim \chi^2_{\ell-k}, \end{aligned}$$

because $\boldsymbol{I}_\ell - \boldsymbol{R}(\boldsymbol{R}'\boldsymbol{R})^{-1}\boldsymbol{R}'$ projects u onto a lower dimensional subspace.

We verify on next slide that the subspace is of dimension $\ell - k$.

Proof of Theorem (ctd.)

To see dimension of target space of $\boldsymbol{I}_\ell - \boldsymbol{R}(\boldsymbol{R}'\boldsymbol{R})^{-1}\boldsymbol{R}'$, note that this is the annihilator associated with $\boldsymbol{R} = \boldsymbol{C}'\mathbb{E}(ZX')$.

This matrix has rank k and hence its null space is of dimension $\ell - k$.

For an algebraic proof, the rank of a projection matrix (or any idempotent matrix) equals its trace, and we can use properties of the trace operator to write

$$\begin{aligned} & \operatorname{tr}\left(\boldsymbol{I}_\ell - \boldsymbol{R}(\boldsymbol{R}'\boldsymbol{R})^{-1}\boldsymbol{R}'\right) \\ = \; & \operatorname{tr}\boldsymbol{I}_\ell - \operatorname{tr}\left(\boldsymbol{R}(\boldsymbol{R}'\boldsymbol{R})^{-1}\boldsymbol{R}'\right) \\ = \; & \operatorname{tr}\boldsymbol{I}_\ell - \operatorname{tr}\left((\boldsymbol{R}'\boldsymbol{R})^{-1}\boldsymbol{R}'\boldsymbol{R}\right) \\ = \; & \operatorname{tr}\boldsymbol{I}_\ell - \operatorname{tr}\boldsymbol{I}_k \\ = \; & \ell - k. \end{aligned}$$

Visualization

Consider

$$\Omega = \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix} \Longrightarrow \Omega^{-1} = \begin{bmatrix} 1/4 & 0 \\ 0 & 1 \end{bmatrix}, \boldsymbol{C} = \begin{bmatrix} 1/2 & 0 \\ 0 & 1 \end{bmatrix}$$

$$\mathbb{E}ZX' = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \Longrightarrow \boldsymbol{R} = \begin{bmatrix} 1/2 \\ 1 \end{bmatrix}.$$

- This is a scenario with $k = 1$ endogenous regressor and $\ell = 2$ instruments.
- The first moment condition has 4 times the variance of the second one.
  In the WLS analogy, it should receive half the weight of the second condition.
  Compare $\boldsymbol{C}$.
  The column space of $\boldsymbol{R}$ is the line spanned by $[1/2, 1]'$.
- The corresponding null space is the line spanned by $[-1, 1/2]'$.

Visualization

Inference

Wald Test

Let $\theta = r(\beta)$ and $\hat{\theta} = r(\hat{\beta})$ for continuously differentiable function $r : \mathbb{R}^k \mapsto \mathbb{R}^q$. Suppose also that $\boldsymbol{R}(\beta) = \partial r(\beta)/\partial \beta'$ has full rank of q at the true value of β. Then

$$W \equiv n(\hat{\theta} - \theta)\left(\boldsymbol{R}(\hat{\beta})'\hat{\boldsymbol{V}}_{\beta}\boldsymbol{R}(\hat{\beta})\right)^{-1}(\hat{\theta} - \theta) \xrightarrow{d} \chi^2_q$$

We will develop other tests later in a more general setting.

Multiple Equation GMM

We next extend the analysis to multiple equations:

$$\begin{aligned} Y_m &= X_m\beta_m + \varepsilon_m,\ m = 1, \ldots, M \\ \mathbb{E} Z_m\varepsilon_m &= 0,\ m = 1, \ldots, M. \end{aligned}$$

- The X_m may be the same (SURE) or overlap (panel data with some time invariant regressors).
- Coefficients are not restricted across equations in this general setting, but they are in many applications.

Multiple Equation GMM

Examples

Seemingly Unrelated Regressions (this example: Griliches 1976)

$$\begin{aligned} LW69 &= \alpha_1 + \beta_1\ schooling69 + \gamma_1\ IQ + \delta_1\ experience69 + \varepsilon_1 \\ KWW &= \alpha_2 + \beta_2\ schooling69 + \gamma_2\ IQ + \varepsilon_2 \end{aligned}$$

Multiple Equation GMM

Examples

Seemingly Unrelated Regressions (this example: Griliches 1976)

$$\begin{aligned} LW69 &= \alpha_1 + \beta_1\ schooling69 + \gamma_1\ IQ + \delta_1\ experience69 + \varepsilon_1 \\ KWW &= \alpha_2 + \beta_2\ schooling69 + \gamma_2\ IQ + \varepsilon_2 \end{aligned}$$

Panel Data (this example: fictitious)

$$\begin{aligned} LW69 &= \alpha_1 + \beta_1\ schooling69 + \gamma_1\ IQ + \delta_1\ experience69 + \varepsilon_1 \\ LW80 &= \alpha_2 + \beta_2\ schooling80 + \gamma_2\ IQ + \delta_2\ experience80 + \varepsilon_2 \end{aligned}$$

Multiple Equation GMM

Define

$$\bar{\boldsymbol{Y}} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_M \end{bmatrix}, \bar{\boldsymbol{X}} = \begin{bmatrix} X_1 & 0 & \cdots & 0 \\ 0 & X_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & X_M \end{bmatrix},$$

$$\bar{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_M \end{bmatrix}, \bar{\boldsymbol{Z}} = \begin{bmatrix} Z_1 & 0 & \cdots & 0 \\ 0 & Z_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & Z_M \end{bmatrix},$$

then we have moment condition

$$\mathbb{E}(\bar{\boldsymbol{Z}}(\bar{\boldsymbol{Y}} - \bar{\boldsymbol{X}}'\bar{\beta})) = 0$$

and estimator

$$\hat{\beta}(\boldsymbol{W}) = \left(\mathbb{E}_n(\bar{\boldsymbol{X}}\bar{\boldsymbol{Z}}')\boldsymbol{W}\mathbb{E}_n(\bar{\boldsymbol{Z}}\bar{\boldsymbol{X}}')\right)^{-1}\mathbb{E}_n(\bar{\boldsymbol{X}}\bar{\boldsymbol{Z}}')\boldsymbol{W}\mathbb{E}_n(\bar{\boldsymbol{Z}}\bar{\boldsymbol{Y}}).$$

Multiple Equation GMM

The estimator

$$\hat{\beta}(\boldsymbol{W}) = \left(\mathbb{E}_n(\bar{\boldsymbol{X}}\bar{\boldsymbol{Z}}')\boldsymbol{W}\mathbb{E}_n(\bar{\boldsymbol{Z}}\bar{\boldsymbol{X}}')\right)^{-1}\mathbb{E}_n(\bar{\boldsymbol{X}}\bar{\boldsymbol{Z}}')\boldsymbol{W}\mathbb{E}_n(\bar{\boldsymbol{Z}}\bar{\boldsymbol{Y}}).$$

is as before, except that data matrix notation becomes unwieldy.

Assuming $(\bar{\boldsymbol{Y}}, \bar{\boldsymbol{X}}, \bar{\boldsymbol{Z}})$ fulfil assumptions as before, we have that

$$\begin{aligned}
\sqrt{n}(\hat{\beta}(\boldsymbol{W}) - \beta) &\xrightarrow{d} N(0, \boldsymbol{V}_\beta) \\
\boldsymbol{V}_\beta &= (\bar{\boldsymbol{Q}}'\boldsymbol{W}\bar{\boldsymbol{Q}})^{-1}\bar{\boldsymbol{Q}}'\boldsymbol{W}\bar{\Omega}\boldsymbol{W}\bar{\boldsymbol{Q}}(\bar{\boldsymbol{Q}}'\boldsymbol{W}\bar{\boldsymbol{Q}})^{-1} \\
\bar{\boldsymbol{Q}} &= \mathbb{E}(\bar{\boldsymbol{Z}}\bar{\boldsymbol{X}}') \\
\bar{\Omega} &= \mathbb{E}(\bar{\boldsymbol{Z}}\varepsilon\varepsilon'\bar{\boldsymbol{Z}}')
\end{aligned}$$

and results on efficient GMM are also as before.

Multiple Equation GMM

The estimator

$$\hat{\beta}(\boldsymbol{W}) = \left(\mathbb{E}_n(\bar{\boldsymbol{X}}\bar{\boldsymbol{Z}}')\boldsymbol{W}\mathbb{E}_n(\bar{\boldsymbol{Z}}\bar{\boldsymbol{X}}')\right)^{-1}\mathbb{E}_n(\bar{\boldsymbol{X}}\bar{\boldsymbol{Z}}')\boldsymbol{W}\mathbb{E}_n(\bar{\boldsymbol{Z}}\bar{\boldsymbol{Y}}).$$

is as before, except that data matrix notation becomes unwieldy.

Assuming $(\bar{\boldsymbol{Y}}, \bar{\boldsymbol{X}}, \bar{\boldsymbol{Z}})$ fulfil assumptions as before, we have that

$$\begin{aligned}
\sqrt{n}(\hat{\beta}(\boldsymbol{W}) - \beta) &\xrightarrow{d} N(0, \boldsymbol{V}_\beta) \\
\boldsymbol{V}_\beta &= (\bar{\boldsymbol{Q}}'\boldsymbol{W}\bar{\boldsymbol{Q}})^{-1}\bar{\boldsymbol{Q}}'\boldsymbol{W}\bar{\Omega}\boldsymbol{W}\bar{\boldsymbol{Q}}(\bar{\boldsymbol{Q}}'\boldsymbol{W}\bar{\boldsymbol{Q}})^{-1} \\
\bar{\boldsymbol{Q}} &= \mathbb{E}(\bar{\boldsymbol{Z}}\bar{\boldsymbol{X}}') \\
\bar{\Omega} &= \mathbb{E}(\bar{\boldsymbol{Z}}\varepsilon\varepsilon'\bar{\boldsymbol{Z}}')
\end{aligned}$$

and results on efficient GMM are also as before.

- This is very powerful!
  We just derived a collection of historically distinct estimators.
- Small caveat:
  Must think carefully about what assumptions on the new objects mean.

Multiple Equation GMM

Which assumptions changed?

- We must assume that $(Y_1, \ldots, Y_M, X_1, \ldots, X_M, Z_1, \ldots, Z_M)$ is i.i.d. This is stronger than assuming equation-by-equation i.i.d.
- We also assume that $\mathbb{E}(\bar{\boldsymbol{Z}}\bar{\boldsymbol{X}}') = \mathbb{E}\begin{bmatrix} Z_1X_1' & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & Z_MX_M' \end{bmatrix}$ has full rank.

Multiple Equation GMM

Which assumptions changed?

- We must assume that $(Y_1, \ldots, Y_M, X_1, \ldots, X_M, Z_1, \ldots, Z_M)$ is i.i.d.
  This is stronger than assuming equation-by-equation i.i.d.
- We also assume that $\mathbb{E}(\bar{\boldsymbol{Z}}\bar{\boldsymbol{X}}') = \mathbb{E}\begin{bmatrix} Z_1X_1' & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & Z_MX_M' \end{bmatrix}$ has full rank.
  This is **equivalent** to assuming that $\mathbb{E}(Z_mX_m')$ has full rank for each m.
  Thus, every equation is individually identified.
  (We will later weaken this assumption.)

All other assumptions did not meaningfully change.

Multiple Equation GMM

Which assumptions changed?

- We must assume that $(Y_1, \ldots, Y_M, X_1, \ldots, X_M, Z_1, \ldots, Z_M)$ is i.i.d.
  This is stronger than assuming equation-by-equation i.i.d.
- We also assume that $\mathbb{E}(\bar{\boldsymbol{Z}}\bar{\boldsymbol{X}}') = \mathbb{E}\begin{bmatrix} Z_1X_1' & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & Z_MX_M' \end{bmatrix}$ has full rank.
  This is **equivalent** to assuming that $\mathbb{E}(Z_mX_m')$ has full rank for each m.
  Thus, every equation is individually identified.
  (We will later weaken this assumption.)

All other assumptions did not meaningfully change.

When is this the same as estimating equations separately?

Because each equation is identified, we could estimate them separately.
This will lead to the same result if:

- everything is just identified or
- $\boldsymbol{W}$ is block diagonal, where its blocks correspond to equations.

In all other cases, we leverage overidentifying information across equations.

Multiple Equation GMM

It may be instructive to write out the estimator for $M = 2$:

$$
\begin{aligned}
&\hat{\beta}(\hat{\boldsymbol{W}}) \\
&= \left[\begin{bmatrix} \mathbb{E}_n(X_1 Z_1') & \mathbf{0} \\ \mathbf{0} & \mathbb{E}_n(X_2 Z_2') \end{bmatrix} \begin{bmatrix} \boldsymbol{W}_{11} & \boldsymbol{W}_{12} \\ \boldsymbol{W}_{21} & \boldsymbol{W}_{22} \end{bmatrix} \begin{bmatrix} \mathbb{E}_n(Z_1 X_1') & \mathbf{0} \\ \mathbf{0} & \mathbb{E}_n(Z_2 X_2') \end{bmatrix} \right]^{-1} \\
&\quad \cdot \begin{bmatrix} \mathbb{E}_n(X_1 Z_1') & \mathbf{0} \\ \mathbf{0} & \mathbb{E}_n(X_2 Z_2') \end{bmatrix} \begin{bmatrix} \boldsymbol{W}_{11} & \boldsymbol{W}_{12} \\ \boldsymbol{W}_{21} & \boldsymbol{W}_{22} \end{bmatrix} \begin{bmatrix} \mathbb{E}_n(Z_1 Y_1) \\ \mathbb{E}_n(Z_2 Y_2) \end{bmatrix} \\
&= \begin{bmatrix} \mathbb{E}_n(X_1 Z_1') \boldsymbol{W}_{11} \mathbb{E}_n(Z_1 X_1') & \mathbb{E}_n(X_1 Z_1') \boldsymbol{W}_{12} \mathbb{E}_n(Z_2 X_2') \\ \mathbb{E}_n(X_2 Z_2') \boldsymbol{W}_{21} \mathbb{E}_n(Z_1 X_1') & \mathbb{E}_n(X_2 Z_2') \boldsymbol{W}_{22} \mathbb{E}_n(Z_2 X_2') \end{bmatrix}^{-1} \\
&\quad \cdot \begin{bmatrix} \mathbb{E}_n(X_1 Z_1') \boldsymbol{W}_{11} \mathbb{E}_n(Z_1 Y_1) + \mathbb{E}_n(X_1 Z_1') \boldsymbol{W}_{12} \mathbb{E}_n(Z_2 Y_2) \\ \mathbb{E}_n(X_2 Z_2') \boldsymbol{W}_{21} \mathbb{E}_n(Z_1 Y_1) + \mathbb{E}_n(X_2 Z_2') \boldsymbol{W}_{22} \mathbb{E}_n(Z_2 Y_2) \end{bmatrix}.
\end{aligned}
$$

Now, what happens if $\boldsymbol{W}_{12} = \boldsymbol{W}_{21} = \mathbf{0}$?

Multiple Equation GMM

If $\boldsymbol{W}_{12} = \boldsymbol{W}_{21} = \mathbf{0}$, we have simplification

$$
\begin{aligned}
\hat{\beta}(\boldsymbol{W}) &= \begin{bmatrix} \mathbb{E}_n(X_1 Z_1')\boldsymbol{W}_{11}\mathbb{E}_n(Z_1 X_1') & \mathbb{E}_n(X_1 Z_1')\boldsymbol{W}_{12}\mathbb{E}_n(Z_2 X_2') \\ \mathbb{E}_n(X_2 Z_2')\boldsymbol{W}_{21}\mathbb{E}_n(Z_1 X_1') & \mathbb{E}_n(X_2 Z_2')\boldsymbol{W}_{22}\mathbb{E}_n(Z_2 X_2') \end{bmatrix}^{-1} \\
&\quad \cdot \begin{bmatrix} \mathbb{E}_n(X_1 Z_1')\boldsymbol{W}_{11}\mathbb{E}_n(Z_1 Y_1) + \mathbb{E}_n(X_1 Z_1')\boldsymbol{W}_{12}\mathbb{E}_n(Z_2 Y_2) \\ \mathbb{E}_n(X_2 Z_2')\boldsymbol{W}_{21}\mathbb{E}_n(Z_1 Y_1) + \mathbb{E}_n(X_2 Z_2')\boldsymbol{W}_{22}\mathbb{E}_n(Z_2 Y_2) \end{bmatrix} \\
&= \begin{bmatrix} \mathbb{E}_n(X_1 Z_1')\boldsymbol{W}_{11}\mathbb{E}_n(Z_1 X_1') & \mathbf{0} \\ \mathbf{0} & \mathbb{E}_n(X_2 Z_2')\boldsymbol{W}_{22}\mathbb{E}_n(Z_2 X_2') \end{bmatrix}^{-1} \\
&\quad \cdot \begin{bmatrix} \mathbb{E}_n(X_1 Z_1')\boldsymbol{W}_{11}\mathbb{E}_n(X_1 Y_1) \\ \mathbb{E}_n(X_2 Z_2')\boldsymbol{W}_{22}\mathbb{E}_n(X_2 Y_2) \end{bmatrix} \\
&= \begin{bmatrix} \left(\mathbb{E}_n(X_1 Z_1')\boldsymbol{W}_{11}\mathbb{E}_n(Z_1 X_1')\right)^{-1}\mathbb{E}_n(X_1 Z_1')\boldsymbol{W}_{11}\mathbb{E}_n(Z_1 Y_1) \\ \left(\mathbb{E}_n(X_2 Z_2')\boldsymbol{W}_{22}\mathbb{E}_n(Z_2 X_2')\right)^{-1}\mathbb{E}_n(X_2 Z_2')\boldsymbol{W}_{22}\mathbb{E}_n(Z_2 Y_2) \end{bmatrix} \\
&= \begin{bmatrix} \hat{\beta}_1(\boldsymbol{W}_{11}) \\ \hat{\beta}_2(\boldsymbol{W}_{22}) \end{bmatrix}.
\end{aligned}
$$

The multiple-equation GMM estimator just stacks single-equation estimators.

This reflects the absence of any nonzero cross-equation weights.

Multiple Equation GMM

When should we estimate equations jointly?

Under a naive interpretation of the result, we should estimate all the world's equations jointly:
If they are in fact unrelated, the efficient weighting matrix will pick that up.

It is intuitively clear that this recommendation is not right. But why?

- One faces an escalating number of nuisance parameters (entries of $\boldsymbol{W}$).
- Most importantly, model misspecification is "contagious":
  The estimator's probability limit equals

$$\operatorname{plim} \hat{\beta}(\boldsymbol{W}) = \beta + \left(\mathbb{E}(\bar{\boldsymbol{X}}\bar{\boldsymbol{Z}}')\boldsymbol{W}\mathbb{E}(\bar{\boldsymbol{Z}}\bar{\boldsymbol{X}}')\right)^{-1}\mathbb{E}(\bar{\boldsymbol{X}}\bar{\boldsymbol{Z}}')\boldsymbol{W}\mathbb{E}(\bar{\boldsymbol{Z}}\varepsilon).$$

  If any one entry of $\mathbb{E}(\bar{\boldsymbol{Z}}\varepsilon)$ is nonzero, then (in general) every entry of the r.h. matrix product is...
- ...except if $\boldsymbol{W}$ is block diagonal corresponding to equations.
  (You can verify this claim along the lines of the preceding slides.)
- Therefore, joint estimation requires all equations to be well-specified.
- But if they are, then joint estimation is efficient.

Multiple Equation GMM

Example: Multiple Regression...

- If $X_1 = \ldots = X_M = Z_1 = \ldots = Z_M$, then this is just multiple regression.
- That is, we regress different $Y_1, \ldots, Y_M$ on the same exogenous regressors.
- Can verify that the estimator just stacks OLS in this case.

Multiple Equation GMM

Example: Multiple Regression...

- If $X_1 = \ldots = X_M = Z_1 = \ldots = Z_M$, then this is just multiple regression.
- That is, we regress different $Y_1, \ldots, Y_M$ on the same exogenous regressors.
- Can verify that the estimator just stacks OLS in this case.

...vs Seemingly Unrelated Regression

- Suppose now some regressors are dropped from some equations, i.e. we think their coefficients are zero.
- But we still consider them exogenous in all equations.
- Then we can use them as overidentifying instruments.
- Formally, let $Z_1 = \ldots = Z_M = \bigcup_{m=1}^{M} X_m$ and the X_m not all the same.
- This is the Seemingly Unrelated Regression (SUR) estimator. (Historically, it would assume homoskedasticity, but conceptually that is a distinct issue.)

Multiple Equation GMM

A Taxonomy of Special Cases

For your interest, as we add assumptions to the basic model, we recover numerous estimators that were historically developed independently.

In several cases, the assumptions also allow to simplify expressions, notably by more extensive use of Kronecker products. See Hayashi's textbook.

- If we assume homoskedasticity, we get Full Information Instrumental Variables Efficient (FIVE) estimation (Brundy and Jorgenson, 1971).
- If we additionally assume $Z_1 = \ldots = Z_M$, we get Three-Stage Least Squares (3SLS, Zellner/Theil 1962).
  In modern terminology, this is a two-stage estimator. It can be expressed as TSLS with "pre-pre-estimation" of cross-equation correlation of errors, hence the original name.
- SUR is the final specialization as we set $Z_m = \bigcup_{m=1}^{M} X_m$.

Multiple Equation GMM

A Taxonomy of Special Cases

For your interest, as we add assumptions to the basic model, we recover numerous estimators that were historically developed independently.

In several cases, the assumptions also allow to simplify expressions, notably by more extensive use of Kronecker products. See Hayashi's textbook.

- If we assume homoskedasticity, we get Full Information Instrumental Variables Efficient (FIVE) estimation (Brundy and Jorgenson, 1971).
- If we additionally assume $Z_1 = \ldots = Z_M$, we get Three-Stage Least Squares (3SLS, Zellner/Theil 1962).
  In modern terminology, this is a two-stage estimator. It can be expressed as TSLS with "pre-pre-estimation" of cross-equation correlation of errors, hence the original name.
- SUR is the final specialization as we set $Z_m = \bigcup_{m=1}^{M} X_m$.

Multiple Equation GMM

Common Coefficients

Next, consider

$$\begin{aligned} Y_m &= X_m\beta + \varepsilon_m,\ m = 1, \ldots, M \\ \mathbb{E} Z_m \varepsilon_m &= 0,\ m = 1, \ldots, M. \end{aligned}$$

- The main intuition here is that the substantively same (but not constant) covariate is observed across equations.
- We do not need to literally revisit every covariate in every equation: Some components of X_m could be zero a.s.

Multiple Equation GMM

Define

$$\bar{\boldsymbol{Y}} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_M \end{bmatrix}, \bar{\boldsymbol{X}} = \begin{bmatrix} X_1 & \cdots & X_M \end{bmatrix}, \bar{\boldsymbol{Z}} = \begin{bmatrix} Z_1 & 0 & \cdots & 0 \\ 0 & Z_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & Z_M \end{bmatrix},$$

then we have moment condition

$$\mathbb{E}(\bar{\boldsymbol{Z}}(\bar{\boldsymbol{Y}} - \bar{\boldsymbol{X}}'\beta)) = 0$$

and estimator

$$\hat{\beta}(\boldsymbol{W}) = \left(\mathbb{E}_n(\bar{\boldsymbol{X}}\bar{\boldsymbol{Z}}')\boldsymbol{W}\mathbb{E}_n(\bar{\boldsymbol{Z}}\bar{\boldsymbol{X}}')\right)^{-1}\mathbb{E}_n(\bar{\boldsymbol{X}}\bar{\boldsymbol{Z}}')\boldsymbol{W}\mathbb{E}_n(\bar{\boldsymbol{Z}}\bar{\boldsymbol{Y}}).$$

This looks like before, but the different definition of $\bar{\boldsymbol{X}}$ changes identification:

$$\left[\mathbb{E}(\bar{\boldsymbol{Z}}\bar{\boldsymbol{X}}') \text{ has full rank }\right] \overset{???}{\iff} \left[\mathbb{E}(Z_m X_m') \text{ has full rank }, m = 1, \ldots, M\right].$$

Multiple Equation GMM

Define

$$\bar{\boldsymbol{Y}} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_M \end{bmatrix}, \bar{\boldsymbol{X}} = \begin{bmatrix} X_1 & \cdots & X_M \end{bmatrix}, \bar{\boldsymbol{Z}} = \begin{bmatrix} Z_1 & 0 & \cdots & 0 \\ 0 & Z_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & Z_M \end{bmatrix},$$

then we have moment condition

$$\mathbb{E}(\bar{\boldsymbol{Z}}(\bar{\boldsymbol{Y}} - \bar{\boldsymbol{X}}'\beta)) = 0$$

and estimator

$$\hat{\beta}(\boldsymbol{W}) = \left(\mathbb{E}_n(\bar{\boldsymbol{X}}\bar{\boldsymbol{Z}}')\boldsymbol{W}\mathbb{E}_n(\bar{\boldsymbol{Z}}\bar{\boldsymbol{X}}')\right)^{-1}\mathbb{E}_n(\bar{\boldsymbol{X}}\bar{\boldsymbol{Z}}')\boldsymbol{W}\mathbb{E}_n(\bar{\boldsymbol{Z}}\bar{\boldsymbol{Y}}).$$

This looks like before, but the different definition of $\bar{\boldsymbol{X}}$ changes identification:

$$\left[\mathbb{E}(\bar{\boldsymbol{Z}}\bar{\boldsymbol{X}}') \text{ has full rank }\right] \Longleftarrow \left[\mathbb{E}(Z_m X_m') \text{ has full rank }, m = 1, \ldots, M\right].$$

Multiple Equation GMM

Again,

$$\left[\mathbb{E}(\bar{\boldsymbol{Z}}\bar{\boldsymbol{X}}') \text{ has full rank } \right] \Longleftarrow \left[\mathbb{E}(Z_m X_m') \text{ has full rank }, m = 1, \ldots, M\right].$$

To see this, write out

$$\bar{\boldsymbol{Z}}\bar{\boldsymbol{X}}' = \begin{bmatrix} Z_1 X_1' \\ \vdots \\ Z_M X_M' \end{bmatrix}.$$

The columns of the r.h. matrix (or its expectation) can be linearly independent even if the columns of no individual matrix $Z_m X_m'$ are, but not conversely.

As a result, the "rank condition" became much less demanding.

Of course, this reflects that we just added a strong identifying assumption.

Multiple Equation GMM

Common coefficients allows for identification of many parameters that would not otherwise be identifiable.

An important application is panel data:

If we impose the assumptions that characterized SUR a few slides ago, we arrive at the **Random Effects** estimator.

If we furthermore assume that ε_m is uncorrelated across m, the estimator simplifies to **Pooled OLS**.

Thus, the difference between those estimators is about efficiency, not about identification.

We will revisit the Pooled OLS and Random Effects estimators soon.

WLS/FGLS versus Efficient GMM

Efficient GMM vaguely resembles Feasible Generalized Least Squares.

Indeed, as a grad student I was quite confused about why the one is commonly used and the other one not.

Building on Hansen, the next slides elaborate the similarities but also the crucial, difference.

WLS/FGLS versus Efficient GMM

As a reminder, weighted least squares minimizes

$$(\boldsymbol{Y} - \boldsymbol{X}\beta)' \underset{n\times n}{\boldsymbol{W}} (\boldsymbol{Y} - \boldsymbol{X}\beta),$$

where the weighting matrix $\boldsymbol{W}$ gives differential weights to different observations.

If we know that $\mathbb{E}(\varepsilon\varepsilon' \mid \boldsymbol{X}) = \sigma^2 \cdot \Omega$ for known Ω, then setting $\boldsymbol{W} = \Omega^{-1}$ is variance minimizing and the resulting estimator is in fact BLUE by Gauss-Markov.

To see the analogy to WLS as you may have seen it in undergraduate courses, note that we could equivalently minimize

$$\big(\boldsymbol{C}(\boldsymbol{Y} - \boldsymbol{X}\beta)\big)'\big(\boldsymbol{C}(\boldsymbol{Y} - \boldsymbol{X}\beta)\big),$$

where $\boldsymbol{C}$ is the Cholesky root of $\boldsymbol{W}$. In particular, if observations are uncorrelated,

$$\Omega = \begin{bmatrix} \sigma_1^2 & \dots & 0 \\ 0 & \ddots & \vdots \\ 0 & \dots & \sigma_n^2 \end{bmatrix}, \Omega^{-1} = \begin{bmatrix} 1/\sigma_1^2 & \dots & 0 \\ 0 & \ddots & \vdots \\ 0 & \dots & 1/\sigma_n^2 \end{bmatrix}, \boldsymbol{C} = \begin{bmatrix} 1/\sigma_1 & \dots & 0 \\ 0 & \ddots & \vdots \\ 0 & \dots & 1/\sigma_n \end{bmatrix}$$

and the minimand can be expressed as $\sum_{i=1}^n \big((y_i - x_i'\beta)/\sigma_i\big)^2$.

WLS/FGLS versus Efficient GMM

What to do if we do not know Ω?

We could attempt to estimate it, but unrestricted estimation of an $n \times n$ matrix from n data points does not sound good.

Two possible ways out:

- If we have a specific model (up to parameter values) for how $\mathbb{E}(\varepsilon \mid X = x)$ changes with x, we could estimate that.
- Else we could run a generic "flexible" regression of $\hat{u}_i^2$ on X; such regressions are run in order to conduct Breusch-Pagan and similar tests for heteroskedasticity.

However, the first case is rare, and in the second case we incur a lot of effort – and concerns about error propagation – for a reweighting that only really matters if weights are very different from even.

This explains why the approach has a name ("Feasible Generalized Least Squares" or FGLS) but is not common in practice.

WLS/FGLS versus Efficient GMM

What does this have to do with GMM?

The GMM estimator can be written as weighted OLS estimator in a fictitious regression:

Set $\mu = \boldsymbol{Z}'\boldsymbol{Y}$ and $\boldsymbol{G} = \boldsymbol{Z}'\boldsymbol{X}$, then we want to minimize

$$(\mu - \boldsymbol{G}\beta)' \underset{\ell \times \ell}{\boldsymbol{W}} (\mu - \boldsymbol{G}\beta),$$

and this looks like weighted least squares (WLS) with k regressors, ℓ observations, and weight matrix $\boldsymbol{W}$. Indeed, the closed-form WLS estimator

$$\hat{\beta}_{WLS} = (\boldsymbol{G}'\boldsymbol{W}\boldsymbol{G})^{-1}\boldsymbol{G}'\boldsymbol{W}\mu$$

exacrly recovers the GMM estimator upon undoing the transformations.

So is efficient GMM kind of like FGLS? And if so, why is it much more common?

WLS/FGLS versus Efficient GMM

So is efficient GMM kind of like FGLS? And if so, why is it much more common?

The analogy is useful but has limitations.

- Every moment condition in the original problem is an "observation" in the fictitious regression.
- In our WLS analogy, the "error" is $\eta \equiv \mu - \boldsymbol{G}\beta$.
- Its variance-covariance matrix is $\ell \times \ell$!
  Compare also sizing of $\boldsymbol{W}$ in objective functions

$$(\boldsymbol{Y} - \boldsymbol{X}\beta)' \underset{n \times n}{\boldsymbol{W}} (\boldsymbol{Y} - \boldsymbol{X}\beta)$$

versus

$$(\mu - \boldsymbol{G}\beta)' \underset{\ell \times \ell}{\boldsymbol{W}} (\mu - \boldsymbol{G}\beta).$$