Introduction to Econometrics II: Recitation 2

Seung-hun Lee*

January 31st, 2022

1 Instrumental Variables Method

1.1 Two stage least squares

Here we relax the assumption that we have as many instrument as the exogenous variables. Suppose the structural equation and the first-stage regression is as follows.

$$y = X\beta + e = X_1\beta_1 + X_2\beta_2 + e \quad \text{(Structural)}$$
$$X = Z\Gamma + v \quad \text{(First Stage)}$$

where $Z \in \mathbb{R}^{n\times l}, \Gamma \in \mathbb{R}^{l\times k}$ and within the (First stage) equation, we have a following reduced form for X_2

$$X_2 = Z\pi_2 + v_2$$

We keep the similar assumption we had before - $E[z_i e_i] = 0, E[x_{1i}' e_i] = 0, E[z_i v_{2i}] = 0$ and $E[x_{2i} e_i] \neq 0$. From the above equation, we can get

$$E[x_{2i} e_i] = E[z_i e_i]\pi_2 + E[v_{2i} e_i]$$

The implication of the above equation is that the endogeneity of the X_2 regressors come from the $E[v_{2i} e_i]$ part, assuming that Z is a valid instrumental variable. Moreover, the reduced form for X_2 implies that X_2 is composed of the parts spanned by the space of Z and the other part that is orthogonal to Z. Thus, the goal of the 2SLS estimators is to 'tease out' part of the

*Contact me at sl4436@columbia.edu if you spot any errors or have suggestions on improving this note.

X_2 variables that can be explained by Z and use a generated regressor from this process to derive the estimator of interest in the structural equation.

To produce the IV regression, we do as follows

1. Regress the first stage and obtain the first stage estimator $\widehat{\Gamma} = (Z'Z)^{-1}Z'X$. Then, get the predicted value of X, denoted as $\widehat{X} = Z(Z'Z)^{-1}Z'X = P_Z X$. By doing so, we provide a way to map the space of l-dimensional vectors to k-dimensional vectors.

   Before going further, here are the properties of the projection matrix P_Z and the residual matrix M_Z. You will use this quite a lot so here is a primer:

   Property 1.1 (Properties of Projection Matrix $P_Z = Z(Z'Z)^{-1}Z'$)**.** *Note that*

   - *Symmetric:* $P_Z' = (Z(Z'Z)^{-1}Z')' = Z(Z'Z)^{-1}Z' = P_Z$
   - *Idempotent:* $P_Z^2 = Z(Z'Z)^{-1}Z'Z(Z'Z)^{-1}Z' = Z(Z'Z)^{-1}Z' = P_Z$

   Property 1.2 (Properties of Residual Matrix $M_Z = I - Z(Z'Z)^{-1}Z'$)**.** *Note that*

   - *Symmetric:* $M_Z' = I' - (Z(Z'Z)^{-1}Z')' = I - Z(Z'Z)^{-1}Z' = I - P_Z = M_Z$
   - *Idempotent:* $M_Z^2 = (I - P_Z)(I - P_Z) = I - P_Z - P_Z + P_Z'P_Z = I - P_Z = M_Z$

2. In the structural equation, replace X with $\widehat{X}$ and obtain

$$\begin{aligned}\hat{\beta}_{2SLS} &= (\widehat{X}'\widehat{X})^{-1}\widehat{X}'y = (X'P_Z'P_ZX)^{-1}(X'P_Z'y) \\ &= (X'P_ZX)^{-1}X'P_Zy = (\widehat{X}'X)^{-1}\widehat{X}'y\end{aligned}$$

   which is effectively replacing Z in the previous approach with $\widehat{X}$.

To show the asymptotic properties of the 2SLS estimators, we need the following set of assumptions

Assumption 1.1 (2SLS Assumptions)**.** *This is a list of required assumptions.*

T1 (y_i, x_i, z_i) *are IID*

T2 *Finite second moments:* $E||y_i^2|| < \infty, E||x_i^2|| < \infty, E||z_i^2|| < \infty$

T3 $E(z_i z_i') > 0$

T4 $rank[E(z_i x_i')] = k$

T5 $E(z_i e_i) = 0$

T6 *Finite fourth moments:* $E||y_i^4|| < \infty, E||x_i^4|| < \infty, E||z_i^4|| < \infty$

T7 $E(z_i z_i' e_i^2) = \Omega > 0$

Then, we can show the consistency and asymptotic normality of 2SLS estimators.

Theorem 1.1 (Consistency of 2SLS)**.** *Under assumptions **T1-T5**,* $\hat{\beta}_{2SLS} \xrightarrow{p} \beta$

Proof. We can rewrite 2SLS estimators as $\hat{\beta}_{2SLS} = \beta + (X'P_Z X)^{-1} X' P_Z e$, which becomes

$$\beta + \left[\frac{X'Z}{n}\left(\frac{Z'Z}{n}\right)^{-1}\frac{Z'X}{n}\right]^{-1}\left[\frac{X'Z}{n}\left(\frac{Z'Z}{n}\right)^{-1}\frac{Z'e}{n}\right]$$

Define $Q_{ZX} = E(z_i x_i'), Q_{XX} = E(x_i x_i')$. Then by weak law of large numbers,

$$\frac{Z'X}{n} \xrightarrow{p} Q_{ZX}, \frac{X'X}{n} \xrightarrow{p} Q_{XX}, \frac{Z'e}{n} \xrightarrow{p} 0$$

By applying Slutsky's theorem and continuous mapping theorem, all the terms right of β converge in probability to 0. Thus, $\hat{\beta}_{2SLS} \xrightarrow{p} \beta$ □

Theorem 1.2 (Limiting Distribution of 2SLS)**.** *Under assumptions **T1-T7**, the limiting distribution of the 2SLS estimator is characterized by* $\sqrt{n}(\hat{\beta}_{2SLS} - \beta) \xrightarrow{d} N(0, V_\beta)$

Proof. Note that

$$\sqrt{n}(\hat{\beta}_{2SLS} - \beta) = \left[\frac{X'Z}{n}\left(\frac{Z'Z}{n}\right)^{-1}\frac{Z'X}{n}\right]^{-1}\left[\frac{X'Z}{n}\left(\frac{Z'Z}{n}\right)^{-1}\frac{Z'e}{\sqrt{n}}\right]$$

By CLT, $\frac{Z'e}{\sqrt{n}} = \frac{1}{\sqrt{n}}\sum_{i=1}^n z_i e_i \xrightarrow{d} N(0, \Omega)$. Then apply Slutsky theorem and continuous mapping theorem to obtain

$$\sqrt{n}(\hat{\beta}_{2SLS} - \beta) \xrightarrow{d} N(0, \underbrace{A_n \Omega A_n'}_{=V_\beta})$$

where $A_n = \left[\frac{X'Z}{n}\left(\frac{Z'Z}{n}\right)^{-1}\frac{Z'X}{n}\right]^{-1}\left[\frac{X'Z}{n}\left(\frac{Z'Z}{n}\right)^{-1}\right]$ □

Before going further, here are some remarks about 2SLS estimators:

- **Correct Standard Errors** When estimating $\Omega = E(z_i z_i' e_i^2)$, we are not aware of the true error. So we need to estimate this as well. Note that we need to use a proper residual. Namely, we must use

$$\hat{e}_i = y_i - x_i'\hat{\beta}_{2SLS}$$

This is correct, as $\hat{\beta}_{2SLS} \xrightarrow{p} \beta$. However, we frequently make a mistake of using

$$\tilde{e}_i = y_i - \hat{x}_i'\hat{\beta}_{2SLS}$$

Since $\tilde{x}_i$ is not exactly x_i, this converges in probability to something else.

> **Comment 1.1** (on STATA)**.** *To see this, compare the standard errors from doing* `ivregress 2sls y (x=z)` *and 'hard-coding' 2SLS by using* `reg x z`*, then* `predict xhat`*, and then coding* `reg y xhat`*.*

- **Nonlinear Extension**: If we are instead interested in the properties of $g(\hat{\beta}_{2SLS})$, where $g(\cdot)$ is not necessarily linear, we can apply delta method here.

- **Too Many IV?**: In practice, when we have too many IV's (large dimension for z_i), it may be possible for the instruments to 'perfectly' predict x_i. In other words, $\hat{x}_i$ becomes nearly identical to x_i. This can become a problem because the endogeneity bias that plagued original structural equation may not be mitigated.
One way of getting around this is to use a regression method that involves penalty mechanism - like LASSO, for instance. We will also formally treat this issue when we learn over-identification tests in the near future.

- **Note on generated regressors**: The 2SLS estimators (and control functions, to be covered later) is an example of a generated regressor. Specifically, we have

$$y = W\beta + e,$$

where W is an unobserved (latent) variable. We do know that they are related to an observable variable Z in the sense that

$$W = Z\gamma + u$$

where γ is unknown (which is why w_i is latent). If γ is consistently estimable by $\hat{\gamma}$, we can instead use

$$\widehat{W} = Z\hat{\gamma}$$

So the regression can be rewritten as

$$y = W\beta + e \iff y = \widehat{W}\beta + (e + (W - \widehat{W})\beta)$$

Hence the OLS regression can be written as

$$\hat{\beta} = (\widehat{W}'\widehat{W})^{-1}\widehat{W}'y \iff \hat{\beta} - \beta = (\widehat{W}'\widehat{W})^{-1}\widehat{W}'(e + (W - \widehat{W})\beta)$$

In large samples, $\widehat{W} \xrightarrow{p} W$ since $\hat{\gamma}$ is a consistent estimator. So $\hat{\beta} - \beta \xrightarrow{p} 0$. so consistency is guaranteed even if we use generated regressors.

Inference, however is a different question. In case where $\beta = 0$, the asymptotic distribution is fairly straightforward. We can write

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{D} N(0, V_\beta)$$

where $V_\beta = (\gamma' E[Z'Z]\gamma)^{-1} (\gamma' E[Z'ee'Z]\gamma) (\gamma' E[Z'Z]\gamma)^{-1}$, the true distribution. This is because $\beta = 0$ implies that sampling uncertainty from using a generated regressor has no impact in later-stage regressions. On the other hand, there is a problem if $\beta \neq 0$. Note that

$$\begin{aligned}
\hat{\beta} - \beta &= (\widehat{W}'\widehat{W})^{-1}\widehat{W}'(e + (W - \widehat{W})\beta) \\
\hat{\gamma} - \gamma &= (Z'Z)^{-1}Z'u \\
\widehat{W} - W &= Z(Z'Z)^{-1}Z'u
\end{aligned}$$

So the asymptotic variance can be written in this manner

$$\begin{aligned}
\hat{\beta} - \beta &= (\hat{\gamma}'Z'Z\hat{\gamma})^{-1}\hat{\gamma}'Z'[e - Z(Z'Z)^{-1}Z'u\beta] \\
&= (\hat{\gamma}'Z'Z\hat{\gamma})^{-1}\hat{\gamma}'Z'[\underbrace{e - u\beta}_{\epsilon}]
\end{aligned}$$

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{D} N(0, V_\beta^*)$$

where $V_\beta^* = (\gamma' E[Z'Z]\gamma)^{-1} (\gamma' E[Z'\epsilon\epsilon'Z]\gamma) (\gamma' E[Z'Z]\gamma)^{-1}$. Here, the sampling uncertainty is playing a role through ϵ and affects later-stage regressions.

1.2 Control function approach

This is another way to derive a 2SLS estimator. As before, I will assume $E(x_{1i}e_i) = 0, E(x_{2i}e_i) \neq 0$ and write the structural and reduced form regression as

$$y_i = x_{1i}'\beta_1 + x_{2i}'\beta_2 + e_i \qquad \text{(Structural)}$$
$$x_{2i} = \Gamma_{12}'x_{1i} + \Gamma_{22}'z_{2i} + u_{2i} \qquad \text{(Reduced Form)}$$

Also, we have a $z_i \in (x_{1i}\ z_{2i})' \in \mathbb{R}^l$ that satisfies $E(z_ie_i) = 0$. The key driving idea for the control function method is that we can write $E(x_{2i}e_i)$ differently. Specifically,

$$\begin{aligned}
E[x_{2i}e_i] &= E[(\Gamma_{12}'x_{1i} + \Gamma_{22}'\beta_2 + u_{2i})e_i] \\
&= \Gamma_{12}'E(x_{1i}e_i) + \Gamma_{22}'E(z_{2i}e_i) + E(u_{2i}e_i) \\
&= E(u_{2i}e_i) \ (\because \text{The first by exogeneity, the second by IV conditions}) \\
&\therefore E[x_{2i}e_i] \neq 0 \iff E[u_{2i}e_i] \neq 0
\end{aligned}$$

From here, we consider a linear projection of e_i onto u_{2i}, which we write as

$$e_i = u_{2i}'a + \epsilon_i \qquad \text{(LP)}$$

where $E(u_{2i}\epsilon_i) = 0$ and the population analogue of $a = E(u_{2i}u_{2i}')^{-1}E(u_{2i}e_i)$. What we now do is to substitute the e_i term in the (Structural) equation with the (LP) equation to get

$$y_i = x_{1i}'\beta_1 + x_{2i}'\beta_2 + u_{2i}'a + \epsilon_i \qquad \text{(CFA)}$$

where the following are satisfied

$$\begin{aligned}
E[x_{1i}\epsilon_i] &= E[x_{1i}(e_i - u_{2i}'a)] = 0 - 0 = 0 \\
E[u_{2i}\epsilon_i] &= E[u_{2i}(e_i - u_{2i}'a)] \\
&= E[u_{2i}e_i] - E[u_{2i}u_{2i}']E[u_{2i}u_{2i}']^{-1}E[u_{2i}e_i] = 0 \\
E[x_{2i}\epsilon_i] &= E[x_{2i}(e_i - u_{2i}'a)] = E[x_{2i}e_i] - E[x_{2i}u_{2i}']a \\
&= E[u_{2i}e_i] - \Gamma_{12}'E[x_{1i}u_{2i}']a - \Gamma_{22}'E[z_{2i}u_{2i}']a - E[u_{2i}u_{2i}']a \\
&= E[u_{2i}e_i] - 0 - 0 - E[u_{2i}e_i] = 0 \\
&(\because \text{LS assumption from Reduced Form, Recitation 1})
\end{aligned}$$

So the key takeaway is that now x_{2i} is exogenous with ϵ_i. In words, x_{2i} is correlated with e_i through u_{2i} and ϵ_i is the error after e_i has been projected onto u_{2i}.

If we know the true u_{2i}, the (CFA) equation can be regressed with an OLS straight away. This is not usually the case, so we need to use the residual $\hat{u}_{2i}$ which we can get from the (Reduced Form) equation. We work in these steps

1. **Obtain** $\hat{u}_{2i}$: This is done by regressing (Reduced Form) equation.
2. **Work with (CFA)**: However, instead of u_{2i}, use $\hat{u}_{2i}$. Then we can run an OLS on the REWRITTEN (CFA) equation.

Once this is done, we can show that the estimates from control function approach is numerically identical to 2SLS.[1]. One thing to note is that with this setup, we can conduct a test of endogeneity. Formally we want to test

$$H_0 : E(x_{2i}e_i) = 0,\ H_1 : E(x_{2i}e_i) \neq 0$$

If it is the case that $E(x_{2i}e_i) = 0$, then $E(u_{2i}e_i) = 0$. Then, by how we constructed a, this implies that $a = 0$. We can then show that the Wald statistics, under H_0, is distributed as

$$\hat{a}'(var(\hat{a}))^{-1}\hat{a} \xrightarrow{d} \chi^2_{k_2}$$

To test the null against the alternative hypothesis, define $C_{1-\alpha}$ as $\Pr(\chi^2_{k_2} \leq C_{1-\alpha}) = 1 - \alpha$. Then, we can reject the H_0 hypothesis if $\hat{a}'(var(\hat{a}))^{-1}\hat{a} > C_{1-\alpha}$.

Example 1.1 (Wooldridge, 2015 *Journal of Human Resources*)**.** *So when and why do we use control function approach in applied economics? Accroding to Wooldridge (2015), control function approach complements standard instrumental variable approach by directly allowing for the study of the nature of self-selection into treatment. Also, it is flexible in that control function approaches can be applied more easily in nonlinear setups, it does not depend on assumptions on maximum likelihood estimations being correct, and is computationally easier. This can also be used to identify average partial effects at different quartiles (See Table 4 in that paper).*

[1]For this, you need to use Frisch-Waugh-Lovell theorem to invoke this. There is good exercise on this topic in last year's problem set and one of the question this year demonstrates this empirically.

1.3 Nonlinear IV

We assume a model that is nonlinear in endogenous variables but still linear in parameters. This would be something like

$$y_1 = \gamma_{12} y_2 + \gamma_{13} y_2^2 + \delta_{11} z_1 + u_1$$
$$y_1 = \gamma_{12} y_2 + \delta_{22} z_2 + u_2$$

where $E[u_1|\mathbf{z}] = 0$ and $E[u_2|\mathbf{z}] = 0$. Thinking of this as a supply and demand equation would make the discussion a bit easier. Nonlinearity comes from the newly endogenous y_2^2 variable, since the inclusion of this variable implies that there is no way to define two endogenous variables such that the system is a two-equation system linear in two endogenous variables.

So what can we do here? One possible approach is to ignore the fact that endogenous y_2 appears differently in different equations by defining $y_3 = y_2^2$. Then what is our candidate for the IV estimator? Assuming that structural system is linear ($\gamma_{13} = 0$), we can write

$$y_2 = \pi_{21} z_1 + \pi_{22} z_2 + v_2$$

Taking squares and conditional expectation, we get

$$E[y_2^2|\mathbf{z}] = \pi_{21}^2 z_1^2 + \pi_{22}^2 z_2^2 + 2\pi_{21}\pi_{22} z_1 z_2 + E[v_2^2|\mathbf{z}]$$

If $E[v_2^2|\mathbf{z}]$ is a constant (homoskedastic), then we can find that y_2 is relevant to $z_1^2, z_2^2, z_1 z_2$ variables. This, however, came from a restrictive condition that $\gamma_{13} = 0$. In practice, we would usually consider $z_1, z_2, z_1^2, z_2^2, z_1 z_2$ all included with the linear projection

$$y_3 = \pi_{31} z_1 + \pi_{32} z_2 + \pi_{33}^2 z_1^2 + \pi_{34}^2 z_2^2 + 2\pi_{35} z_1 z_2 + v_3$$

A typical estimation method is as follows: Assume that the structural equation has

$$y_i = x_{1i}\beta_1 + x_{2i}\beta_2 + x_{2i}^2\beta_3 + e_i$$

while there are many candidates for an IV, one mistake that is tempting is to include $x_{1i}, \hat{x}_{2i}, (\hat{x}_{2i})^2$ in the second stage regression. This ignores the fact that the linear projection of the squared is not equal to the square of the linear projection, leading to an inconsistent estimation. Consistent estimation must project the original and squared values separately

1.4 Correlated Random Coefficients Models

Suppose we are interested in the effect of X_i that can differ across each individual i. We can write this model as

$$y_i = x_{2i}\beta_i + e_i, \ (E[x_{2i}e_i] \neq 0)$$

where $\beta_i = \beta + w_i$ with $E[\beta_i] = \beta$. The model of interest can be rewritten as

$$y_i = x_{2i}\beta + e_i + x_{2i}w_i,$$

Even if we assume that $E[e_i|z_i] = E[w_i|z_i] = 0$ IV estimator is not consistent if $E[x_{2i}w_i|z_i]$ depends on z_i. For a consistent estimation, what we need is that the conditional covariance, $cov(x_{2i}, w_i|z_i) = cov(x_{2i}, w_i)$. Unconditional covariance can take any value as it only affects the consistency of the intercept term.

This model can be approached with control function approach. The logic below is from Garen (2012). The model for x_{2i} is

$$x_{2i} = z_i\pi_2 + v_{2i}$$

where we assume $E[v_{2i}|z_i] = 0, E[e_i|v_{2i}] = \rho_e v_{2i}, E[w_i|v_{2i}] = \rho_w v_{2i}$ to get to normality and independence assumptions. We return to the structural equation and take conditional expectation on x_{2i}, v_{2i} to get

$$E[y_i|x_{2i}, v_{2i}] = x_{2i}\beta + \rho_e v_{2i} + \rho_w v_{2i}x_{2i}$$

and the equation above is estimable with regressing y_i on $x_{2i}, \hat{v}_{2i}$, and $\hat{v}_{2i}x_i$. Since this is a generated regressor, standard error needs correction. But like in general case of control function, test of exogeneity is possible using $H_0 : \rho_e = 0, \rho_w = 0$.

2 Linear Generalized Method of Moments

GMM methods utilize the method of moments estimators to identify the values of the parameters of interest. It can be generalized in the sense that the number of moment conditions can be greater than the number of unknown parameters. In fact, we can show that OLS, IV, MLE estimators are part of GMM.

2.1 Framework

The moment equation takes the following form: Let w_i be IID across $i = 1, .., n$, $g_i(w_i, \beta)$ be a $l \times 1$ function of the ith observation, and $\beta \in \mathbb{R}^{k \times 1}$ be the parameter of interest. Note that $l \geq k$. Then, the **moment equation model** is characterized by

$$E[g(w_i, \beta)] = 0$$

We say β is identified if there is a unique mapping from the data distribution to β, or that there is a unique β satisfying $E[g(w_i, \beta)] = 0$. When $l = k$, then we are in a just-identified case. If $l > k$, then we are in the over-identified case in the sense that there is excess information. We will not bother with the case of $l < k$, as they are under-identified and the moment equation model is unsolvable.

- **Just-identified**: In this case, we can work with the sample analogue of $g(w_i.\beta)$ straight away. Define $\bar{g}_n(\beta)$ as

  $$\bar{g}_n(\beta) = \frac{1}{n} \sum_{i=1}^{n} g_i(\beta)$$

  The **method of moments estimators** $\hat{\beta}_{MM}$ is defined as the parameter value which sets $\bar{g}_n(\beta) = 0$. In other expression:

  $$\bar{g}_n(\hat{\beta}_{MM}) = \frac{1}{n} \sum_{i=1}^{n} g_i(\hat{\beta}_{MM}) = 0$$

- **General Case**: If we can generalize to include the case where $l > k$, we are likely to run into a situation where $\hat{\beta}_{MM}$ cannot be found. This is because there may be no choice of β that sets the moment equations to 0, implying the nonexistence of the method of moments estimator. In such case, we work with a different approach.

  Define $J(\beta)$ as

  $$J(\beta) = n\bar{g}_n(\beta)' W \bar{g}_n(\beta)$$

  where $W \in \mathbb{R}^{l \times l}$ is a positive definite weight matrix that is given. n does not really affect our estimation, but it makes the analysis of the asymptotic features much easier. The **generalized method of moments estimator** is defined as the minimizer of the GMM criterion above, or

  $$\hat{\beta}_{GMM} = \arg\min_{\beta} J_n(\beta)$$

Note that when $l = k$, then method of moments estimator solve $\bar{g}_n(\hat{\beta}_{MM}) = 0$. Given that $J(\beta)$ is a positive definite matrix, the method of moments estimator in this case also minimizes $J(\beta)$. Thus, method of moments estimator is a special case of GMM estimator.

If $l > k$, then the problem we should solve is

$$\min_{\beta} J_n(\beta) = n\bar{g}(\beta)' W \bar{g}(\beta)$$
$$\implies \frac{\partial J_n(\beta)}{\partial \beta} = 2n \frac{\partial \bar{g}(\beta)}{\partial \beta}' W \bar{g}(\beta) = 0$$

To obtain this, I have used the following properties of matrix differentiation.

Property 2.1 (Matrix Differentiation Tricks)**.** *Let Let $x \in \mathbb{R}^{k\times 1}$ and $A \in \mathbb{R}^{k\times l}, y \in \mathbb{R}^{l\times 1}$. Then we have*

$$\frac{\partial x' A y}{\partial x} = Ay, \frac{\partial x' A y}{\partial y} = A' x$$

Let $x \in \mathbb{R}^{k\times 1}$ and $A \in \mathbb{R}^{k\times k}$. Then the following holds

$$\frac{\partial x' A x}{\partial x} = (A + A')x$$

Example 2.1 (Various Examples of GMM)**.** *These are some types of a GMM estimator.*

- *OLS: In a data generating process $y_i = x_i'\beta + e_i,\ x_i \in \mathbb{R}^k$ and the moment condition $E(x_i e_i) = 0$, we have k parameters β_k and k equations for each of the k variables. We can rewrite the moment condition as*

$$E(x_i(y_i - x_i'\beta)) = 0$$

And the method of moments estimators imply that we should solve

$$\frac{1}{n}\sum_{i=1}^{n} x_i(y_i - x_i'\beta) = 0 \iff \hat{\beta}_{OLS} = \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} x_i y_i$$

- *MLE: If w_i is IID across $i = 1, ..., n$, then we can write the joint likelihood function as*

$$\prod_{i=1}^{n} f(w_i|\beta)$$

and thus, the log-likelihood function

$$\sum_{i=1}^{n} \log f(w_i|\beta)$$

When we take partial differentiation w.r.t β,

$$\sum_{i=1}^{n} \frac{\partial \log f(w_i|\beta)}{\partial \beta} = 0 \implies \frac{1}{n} \sum_{i=1}^{n} \frac{\partial \log f(w_i|\beta)}{\partial \beta} = 0$$

which is equivalent to $E\left(\frac{\partial \log f(w_i|\beta)}{\partial \beta}\right) = 0$. Practically, $E\left(\frac{\partial \log f(w_i|\beta)}{\partial \beta}\right) = 0$ becomes the moment condition applicable to MLE.

- *IV: Suppose we have a data generating process $y_i = x_i'\beta + e_i$ with x_i being k dimensions. Suppose we have an $l > k$ dimensional IV with $E(z_i e_i) = 0$ $\bar{g}(\beta)$ in our context would be*

$$\frac{1}{n} \sum_{i=1}^{n} (z_i y_i - z_i x_i'\beta) = \frac{Z'y}{n} - \frac{Z'X\beta}{n}$$

Then, we can write our $J_n(\beta)$ as

$$n \left(\frac{Z'y}{n} - \frac{Z'X\beta}{n}\right)' W \left(\frac{Z'y}{n} - \frac{Z'X\beta}{n}\right)$$

By solving the minimization problem we can obtain

$$\frac{\partial J_n(\beta)}{\partial \beta} = -\frac{2}{n}(X'ZWZ'y) + \frac{2}{n}(X'ZWZ'X)\beta = 0$$
$$\implies \hat{\beta} = (X'ZWZ'X)^{-1}(X'ZWZ'y)$$

In fact, we can show that by setting $W = \left(\frac{Z'Z}{n}\right)^{-1}$, the above estimator is in fact, a 2SLS estimator.

2.2 Limiting Distribution of GMM

From above, we know that $\hat{\beta}_{GMM} = (X'ZWZ'X)^{-1}(X'ZWZ'y)$ for the case of an overidentified IV model. We can rewrite this by replacing y with $X\beta + e$, As a result, the limiting distribution of $\hat{\beta}_{GMM}$ is characterized by

$$\sqrt{n}(\hat{\beta}_{GMM} - \beta) = \left(\frac{X'Z}{n} W \frac{Z'X}{n}\right)^{-1} \left(\frac{X'Z}{n} W \frac{Z'e}{\sqrt{n}}\right)$$

For convenience, I assume the following

Assumption 2.1. *Assume that*

1. $E(z_i x_i') = Q$, *and that* $\frac{Z'X}{n} \xrightarrow{p} Q$
2. $\frac{Z'e}{\sqrt{n}} \xrightarrow{d} N(0, \Omega)$, *where* $\Omega = E(z_i z_i' e_i^2)$
3. *(If we are willing to assume W depends on n, thus* W_n*):* $W_n \xrightarrow{p} W$, *where W is a positive definite weight matrix*

Then we can say the following about the distribution of the GMM estimator

Theorem 2.1 (Limiting Distribution of the GMM estimator)**.** *If the above assumptions are satisfied, the limiting distribution of the GMM estimator can be characterized by*

$$\sqrt{n}(\hat{\beta}_{GMM} - \beta) \xrightarrow{d} N(0, (Q'WQ)^{-1}(Q'W\Omega W'Q)(Q'WQ)^{-1})$$

Where $(Q'WQ)^{-1}$ *can be replaced with A, if we stick to the notation in class.*

So far, we haven't said too much about W. We just assumed that they are fixed. Even if we suppose that W depends on n somehow, the above theorem still holds, provided that W_n converges in probability to W. Then, the natural question should be "What is the optimal selection of W?". This naturally moves us to...

2.3 Efficient GMM

To select an optimal W matrix, it must be that the resulting variance should be the smallest. If we let $W = \Omega^{-1}$ and work with $(Q'WQ)^{-1}(Q'W\Omega W'Q)(Q'WQ)^{-1} - (Q'\Omega Q)^{-1}$, we

can see that it is positive semidefinite. If we put $W = \Omega^{-1}$ and recalculate the variance, we get that the efficient GMM has a limiting distribution characterized by

$$\sqrt{n}(\hat{\beta}_{GMM} - \beta) \xrightarrow{d} N(0, (Q'WQ)^{-1})$$

Note that this weighting matrix, which can be rewritten as

$$W = \Omega^{-1} = E(z_i z_i e_i^2)^{-1}$$

is not exactly same as the weighting matrix we used for deriving the 2SLS estimator from GMM, which is $\left(\frac{Z'Z}{n}\right)^{-1}$. In the $W = E(z_i z_i e_i^2)^{-1}$ setup, we allowed for heteroskedasticity. If we impose conditional homoskedasticity in the sense that $E(e_i^2|z_i) = \sigma^2$, we can rewrite $E(z_i z_i' e_i^2)$ as

$$\begin{aligned} E(z_i z_i' e_i^2) &= E(E(z_i z_i' e_i^2|z_i)) = E(z_i z_i' E(e_i^2|z_i)) \\ &= E(z_i z_i' \sigma^2) = E(z_i z_i')\sigma^2 \end{aligned}$$

Thus, $E(z_i z_i' e_i^2)^{-1} = (Z'Z)^{-1}(\sigma^2)^{-1}$. However, scaling does not affect how the estimator is defined (by calculating $\hat{\beta}_{GMM}$ the scalar scales disappear). So $E(z_i z_i' e_i^2)^{-1}$ is effectively $(Z'Z)^{-1}$ and we get the conclusion that in a conditional homoskedastic setting, 2SLS estimator is the efficient GMM.

2.4 Two-step Optimal GMM

Since we have no idea what the true value of e_i is, the true value of Ω is naturally unknown. Therefore, we require a consistent estimator, denoted as $\widehat{W}$, for $W = \Omega^{-1}$. To do so, we need a preliminary consistent estimator for the true β, which I will denote with $\tilde{\beta}$. You can use any W matrix for this step (I or $(Z'Z)^{-1}$). Then, define

$$\widehat{\Omega} = \frac{1}{n}\sum_{i=1}^{n} g(w_i.\tilde{\beta})g(w_i.\tilde{\beta})'$$

since $\Omega = E[g(w_i, \beta)g(w_i, \beta)']$. Then we can identify what $\widehat{\Omega}^{-1}$ is. This will allow us to compute the efficient GMM estimator $\hat{\beta}_{GMM}$

Definition 2.1 (Optimal Two-step GMM Estimator)**.** We can compute Optimal Two-step GMM in these steps

1. Compute a preliminary, but consistent estimator for the true β. Denote this as $\tilde{\beta}$.
2. Using $\Omega = E[g(w_i, \beta)g(w_i, \beta)']$, create a sample analogue of this, defined as $\widehat{\Omega} = \frac{1}{n}\sum_{i=1}^{n} g(w_i.\tilde{\beta})g(w_i.\tilde{\beta})'$. We can find our $\widehat{\Omega}^{-1}$ here.
3. Using this $\widehat{\Omega}^{-1}$, construct an efficient GMM estimator $\hat{\beta}_{GMM}$

There is another way to come up with a $\widehat{\Omega}^{-1}$ in this context. Define $\bar{g}(\beta) = \frac{1}{n}\sum_{i=1}^{n} g(w_i, \tilde{\beta})$. Then, an alternative definition of $\widehat{\Omega}$ can be written as

$$\widehat{\Omega}^{+} = \frac{1}{n}\sum_{i=1}^{n}(g(w_i, \tilde{\beta}) - \bar{g}(\beta))(g(w_i, \tilde{\beta}) - \bar{g}(\beta))'$$

So why use this? For one thing, both $\widehat{\Omega}$ and $\widehat{\Omega}^{+}$ converge in probability to $E[g(w_i, \beta)g(w_i, \beta)']$. This is because

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}(g(w_i, \tilde{\beta}) - \bar{g}(\beta))(g(w_i, \tilde{\beta}) - \bar{g}(\beta))' \\
&= \frac{1}{n}\sum_{i=1}^{n} g(w_i, \tilde{\beta})g(w_i, \tilde{\beta})' - \frac{1}{n}\sum_{i=1}^{n} g(w_i, \tilde{\beta})\bar{g}(\beta)' - \frac{1}{n}\sum_{i=1}^{n}\bar{g}(\beta)g(w_i, \tilde{\beta})' + \frac{1}{n}\sum_{i=1}^{n}\bar{g}(\beta)\bar{g}(\beta)' \\
&= \frac{1}{n}\sum_{i=1}^{n} g(w_i, \tilde{\beta})g(w_i, \tilde{\beta})' - \bar{g}(\beta)\bar{g}(\beta)' \quad \left(\because \bar{g}(\beta) = \frac{1}{n}\sum_{i=1}^{n} g(w_i, \tilde{\beta})\right)
\end{aligned}$$

Under the assumption that $E[g(w_i, \beta)] = 0$, (or $E(z_i e_i) = 0$) is a correct specification, we have $\bar{g}(\beta) \xrightarrow{p} E(g(w_i, \beta)) = 0$. So $\widehat{\Omega}^{+}$ and $\widehat{\Omega}$ both converge to $E[g(w_i, \beta)g(w_i, \beta)']$.

The difference? If it is the case that $E[g(w_i, \beta)] \neq 0$ we view $\widehat{\Omega}^{+}$ as a robust estimator. Note that

- $\widehat{\Omega} = \frac{1}{n}\sum_{i=1}^{n} g(w_i.\tilde{\beta})g(w_i.\tilde{\beta})' \xrightarrow{p} E[g(w_i, \tilde{\beta})g(w_i, \tilde{\beta})']$
- $\widehat{\Omega}^{+} = \frac{1}{n}\sum_{i=1}^{n} g(w_i, \tilde{\beta})g(w_i, \tilde{\beta})' - \bar{g}(\beta)\bar{g}(\beta)' \xrightarrow{p} E[g(w_i, \tilde{\beta})g(w_i, \tilde{\beta})'] - E[g(w_i, \tilde{\beta})]E[g(w_i, \tilde{\beta})]' = var[g(w_i, \tilde{\beta})]$

In other words, $\widehat{\Omega}$ is inconsistent in case where $E[g(w_i, \tilde{\beta})] = 0$ is not guaranteed. Therefore, for tests, such as overidentification tests, it is much more desirable to use $\widehat{\Omega}^{+}$.

Since we know how to find the optimal $\widehat{W}$, we can estimate the asymptotic variance of the GMM estimators. This can be done by replacing matrices in the original variance with

their sample counterparts. In general, we can estimate by

$$\widehat{V}_{GMM} = \left(\widehat{Q}'\widehat{W}\widehat{Q}\right)^{-1}\left(\widehat{Q}'\widehat{W}\widehat{\Omega}\widehat{W}\widehat{Q}\right)\left(\widehat{Q}'\widehat{W}\widehat{Q}\right)^{-1}$$

where $\widehat{Q} = \frac{1}{n}\sum_{i=1}^{n} z_i x_i' = \frac{Z'X}{n}$, $\widehat{W}$ is expressed by either $\widehat{\Omega}$ or $\widehat{\Omega}^{+}$. The residuals used in this estimation is defined as $\hat{e}_i = y_i - x_i'\hat{\beta}_{GMM}$

One alternative to the two-step GMM estimator is contructed by letting weight matrix be an explicit function β. The criterion function is now

$$J(\beta) = n\bar{g}_n(\beta)'\left(\frac{1}{n}\sum_{i=1}^{n} g(w_i,\beta)g(w_i,\beta)'\right)^{-1}\bar{g}_n(\beta)$$

The $\hat{\beta}$ that minimizes this function is the continuously-updated GMM estimator. However, this setup is nonlinear, implying that acquiring this estimator requires numerical methods.