Conditional Expectation Function (CEF)

- We begin by thinking about population relationships.

- CEF Decomposition Theorem: Given some outcome Y_i and some covariates X_i there is always a decomposition

$$Y_i = E(Y_i/X_i) + \epsilon_i \tag{1}$$

where

$$E(\epsilon_i/X_i) = 0 \tag{2}$$

- Proof:

$$\begin{aligned} E(\epsilon_i/X_i) &= E[(Y_i - E(Y_i/X_i))/X_i] && (3)\\ &= E(Y_i/X_i) - E[E(Y_i/X_i)/X_i] && (4)\\ &= 0 && (5) \end{aligned}$$

- The last step uses the *Law of Iterated Expectations*:

$$E(Y) = E[E(Y/X)] \tag{6}$$

where the outer expectation is over X. For example, the average outcome is the weighted average of the average outcome for men and the average outcome for women where the weights are the proportion of each sex in the population.

- The CEF Decomposition Theorem implies that ϵ_i is uncorrelated with any function of X_i.

- Proof: Let $h(X_i)$ be any function of X_i.

$$E[h(X_i)\epsilon_i] = E\{E[(h(X_i)\epsilon_i)/X_i]\} = E\{h(X_i)E(\epsilon_i/X_i)\} = 0 \tag{7}$$

- We refer to the $E(Y_i/X_i)$ as the CEF.

- **Best Predictor**: $E(Y_i/X_i)$ is the Best (Minimum mean squared error – MMSE) predictor of Y_i in that it minimises the function

$$E((Y_i - h(X_i))^2 \tag{8}$$

where $h(X_i)$ is any function of X_i.

- A regression model is a particular choice of function for $E(Y_i/X_i)$.

- Linear regression:

$$Y_i = \beta_1 + \beta_2 X_{2i} + ... + \beta_K X_{Ki} + \varepsilon \tag{9}$$

- Of course, the linear model may not be correct.

Linear Predictors

- A linear predictor (with only one regressor) takes the form

$$E^*(Y_i/X_i) = \beta_1 + \beta_2 X_{2i} \tag{10}$$

- Suppose we want the Best Linear Predictor (BLP) for Y_i to minimise

$$E((Y_i - E^*(Y_i/X_i))^2 \tag{11}$$

- The solution is

$$\begin{aligned} \beta_1^* &= \mu_Y - \beta_2^* \mu_X & (12) \\ \beta_2^* &= cov(X_i, Y_i)/Var(X_i) & (13) \end{aligned}$$

- In the multivariate case with

$$E^*(Y_i/X_i) = X_i'\beta = \beta_1 + \beta_2 X_{2i} + ... + \beta_K X_{Ki} \tag{14}$$

, we have

$$\beta^* = E(X_i X_i')^{-1} E(X_i Y_i) \tag{15}$$

- We can see this by taking the first order condition for (11)

$$X_i(Y_i - X_i'\beta) = 0 \tag{16}$$

- The error term $\varepsilon_i = Y_i - E^*(Y_i/X_i)$ satisfies $E(\varepsilon_i X_i) = 0$.

- $E^*(Y/X)$ is the best linear approximation to $E(Y/X)$.

- If $E(Y_i/X_i)$ is linear in X_i, then $E(Y_i/X_i) = E^*(Y_i/X_i)$.

- Will only be the case that $E(\varepsilon_i/X_i) = 0$ if the CEF is linear. However $E(\varepsilon_i X_i) = 0$ in all cases.

Example: Bivariate Normal Distribution

- Assume Z_1 and Z_2 are two standard normal variables and

$$X_1 = \mu_1 + \sigma_1 Z_1 \quad (17)$$
$$X_2 = \mu_2 + \sigma_2(\rho Z_1 + \sqrt{(1-\rho^2)} Z_2) \quad (18)$$

- Then (X_1, X_2) are bivariate normal with $X_j \tilde{} N(\mu_j, \sigma_j^2)$.

- The covariance between (X_1, X_2) is $\rho\sigma_1\sigma_2$.

- Then using (12) and (13), the BLP

$$E^*(X_2/X_1) = \mu_2 - \beta\mu_1 + \beta X_1 \quad (19)$$
$$= \mu_2 + \beta(X_1 - \mu_1) \quad (20)$$

  where

$$\beta = \rho\sigma_2/\sigma_1 \quad (21)$$

- Note that using properties of the Normal distribution,

$$E(X_2/X_1) = \mu_2 + \beta(X_1 - \mu_1) \quad (22)$$

  so the CEF is linear in this case.

- This is not true with other distributions.

CEF and Regression Summary

- If the CEF is linear, then the population regression function is exactly it.
- If the CEF is non-linear, the regression function provides the best linear approximator to it.
- The CEF is linear only in special cases such as:

1. Joint Normality of Y and X.
2. Saturated regression models – models with a separate parameter for every possible combination of values that the regressors can take. This occurs when there is a full set of dummy variables and interactions between the dummies.

For example, suppose we have a dummy for female (x_1) and a dummy for white (x_2). The CEF is

$$E(Y/x_1, x_2) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \delta x_1 x_2 \tag{23}$$

We can see that there is a parameter for each possible set of values

$$\begin{aligned} E(Y/x_1 &= 0, x_2 = 0) = \alpha && (24)\\ E(Y/x_1 &= 0, x_2 = 1) = \alpha + \beta_2 && (25)\\ E(Y/x_1 &= 1, x_2 = 0) = \alpha + \beta_1 && (26)\\ E(Y/x_1 &= 1, x_2 = 1) = \alpha + \beta_1 + \beta_2 + \delta && (27) \end{aligned}$$

Linear Regression

- The regression model is

$$Y_i = X_i'\beta + \varepsilon_i \tag{28}$$

- We assume that $E(\varepsilon_i / X_i) = 0$ if the CEF is linear.

- We assume only that $E(X_i \varepsilon_i) = 0$ if we believe the CEF is nonlinear.

- The population parameters are

$$\beta = E(X_i X_i')^{-1} E(X_i Y_i) \tag{29}$$

- Here, Y_i is a scalar and X_i is $K * 1$ where K is the number of X variables.

- In matrix notation,

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, X = \begin{pmatrix} X_1' \\ X_2' \\ \vdots \\ X_n' \end{pmatrix}, \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \tag{30}$$

- So, we can write the model as

$$Y = X\beta + \varepsilon \tag{31}$$

where Y is $n * 1$, X is $n * K$, β is $K * 1$, and ε is $n * 1$.

Best Linear Predictor

- We observe n observations on $\{Y_i, X_i\}$ for $i = 1, ..., n$.

- We derive the OLS estimator as the BLP as it solves the sample analog of min $E((Y_i - X_i'\beta)^2)$.

- In fact it minimises $1/n \sum_i (Y_i - X_i'b)^2$.

- In matrix notation, this equals $(Y - Xb)'(Y - Xb) = Y'Y - 2b'X'Y + b'X'Xb$

- FOC: $X'Y - X'Xb = X'(Y - Xb) = 0$.

- So long as $X'X$ is invertible (X is of full rank), $b = (X'X)^{-1}X'Y$

Regression Basics

- The fitted value of a regression

$$\widehat{Y} = Xb = X(X'X)^{-1}X'Y = PY \tag{32}$$

where $P = X(X'X)^{-1}X'$.

- The residuals

$$\begin{aligned} e &= Y - Xb && (33) \\ &= Y - X(X'X)^{-1}X'Y && (34) \\ &= (I - P)Y = MY && (35) \end{aligned}$$

- M and P are symmetric idempotent matrices.

- Any matrix A is idempotent if it is square and $AA = A$.

- P is called the projection matrix that projects Y onto the columns of X to produce the set of fitted values

$$\widehat{Y} = Xb = PY \quad (36)$$

- What happens when you project X onto X?

- M is the residual-generating matrix as $e = MY$. We can easily show that

$$MY = M\varepsilon \tag{37}$$

so although the true errors are unobserved, we can obtain a certain linear combination.

Frisch-Waugh-Lovell Theorem

- Consider the partitioned regression

$$Y = X_1\beta_1 + X_2\beta_2 + \varepsilon \tag{38}$$

- Here X_1 is $n * K_1$ and X_2 is $n * K_2$.

- The FOC are

$$\begin{pmatrix} X_1' \\ X_2' \end{pmatrix} (Y - X_1 b_1 - X_2 b_2) = 0 \tag{39}$$

- Therefore

$$\begin{aligned} X_1'(Y - X_1 b_1 - X_2 b_2) &= 0 \\ X_2'(Y - X_1 b_1 - X_2 b_2) &= 0 \end{aligned} \tag{40}$$

and

$$X_1' X_1 b_1 + X_1' X_2 b_2 = X_1' Y \tag{41}$$
$$X_2' X_1 b_1 + X_2' X_2 b_2 = X_2' Y \tag{42}$$

- Solving these simultaneous equations (you should be able to do this) we get

$$b_1 = (X_1' M_2 X_1)^{-1} X_1' M_2 Y \tag{43}$$
$$b_2 = (X_2' M_1 X_2)^{-1} X_2' M_1 Y \tag{44}$$

where $M_1 = I - X_1(X_1'X_1)^{-1}X_1'$ and $M_2 = I - X_2(X_2'X_2)^{-1}X_2'$.

- Implies can estimate b_1 by regressing M_2Y on M_2X_1.

- That is, can first regress Y and X_1 on X_2 and then regress residuals on residuals.

- Note that one can also do this by regressing only X_1 on X_2 and then regressing Y on the residuals.

Application to the Intercept

- Consider the case where X_1 is the intercept (a column vector of ones) $X_1 = l$.

- Now

$$M_1 = I - l(l'l)^{-1}l' = I - \frac{1}{n}ll' \tag{45}$$

- Note that $\frac{1}{n}ll'$ is an $n * n$ matrix with each element equal to $1/n$.

- Therefore, in this case,

$$M_1 X_2 = X_2 - \overline{X}_2 \tag{46}$$

- Implies that regressing Y and X_2 on a constant and then regressing residuals on residuals is the same as taking deviations from means.

- Also implies that if you remove means from all variables, you do not need to include a constant term.

Example: Non-Linear and Linear CEF -- Birth Order and IQ in 5-Child Families (N=10214)

1. X includes a constant and a linear Birth Order variable.

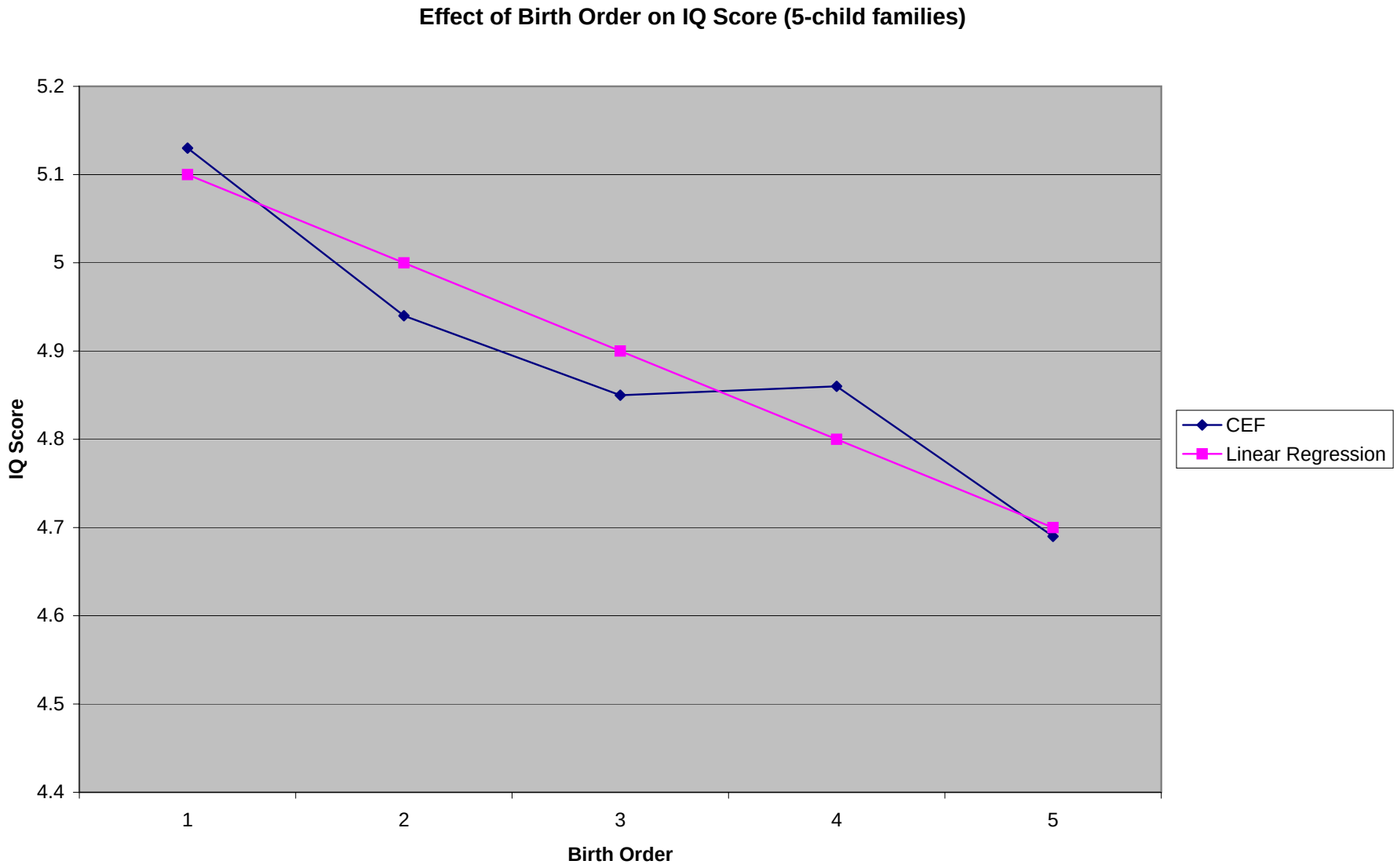
$\beta_1 = 5.2,\ \beta_2 = -0.10$

X_1	X_2	E(Y/X)	$X\hat{\beta}$	E(ε/x)
1	1	5.13	5.1	0.03
1	2	4.94	5	-0.06
1	3	4.85	4.9	-0.05
1	4	4.86	4.8	0.06
1	5	4.69	4.7	-0.01

2. X includes a constant and dummy variables for Birth Order = 2, 3, 4, 5.

$\beta_1 = 5.13,\ \beta_2 = -0.19,\ \beta_3 = -0.28,\ \beta_4 = -0.27,\ \beta_5 = -0.44$

X_1	X_2	X_3	X_4	X_5	E(Y/X)	$X\hat{\beta}$	E(ε/x)
1	0	0	0	0	5.13	5.13	0
1	1	0	0	0	4.94	4.94	0
1	0	1	0	0	4.85	4.85	0
1	0	0	1	0	4.86	4.86	0
1	0	0	0	1	4.69	4.69	0


Effect of Birth Order on IQ Score (5-child families)
IQ Score
5.2
5.1
5
4.9
4.8
4.7
4.6
4.5
4.4
1
2
3
4
5
Birth Order
CEF
Linear Regression

Instrumental Variables

- The regression model is $Y_i = X_i'\beta + \varepsilon_i$
- OLS is consistent if $E(X_i\varepsilon_i) = 0$.
- If this assumption does not hold, OLS is *endogenous*.

Example 1: Simultaneous Equations

- The simplest supply and demand system:

$$q_i^s = \beta_s p_i + \varepsilon_s \quad (1)$$
$$q_i^d = \beta_d p_i + \varepsilon_d \quad (2)$$
$$q_i^s = q_i^d \quad (3)$$

- In equilibrium,

$$p_i = \frac{\varepsilon_d - \varepsilon_s}{\beta_s - \beta_d} \tag{4}$$

- Obviously, p_i is correlated with the error term in both equations.

Example 2: Omitted Variables

- Suppose $Y = X_1\beta_1 + X_2\beta_2 + \varepsilon$ but we exclude X_2 from the regression.

- Here X_1 is $n * K_1$ and X_2 is $n * (K - K_1)$.

- The new error term $v = X_2\beta_2 + \varepsilon$ is correlated with X_1 unless $X_1$and X_2 are orthogonal or $\beta_2 = 0$.

Example 3: Measurement Error

- To see the effect of measurement error, consider the standard regression equation where there are no other control variables

$$y_i = \alpha + \beta x_i + \epsilon_i \tag{5}$$

- However, we observe

$$\widetilde{x}_i = x_i + u_i \tag{6}$$

where u_i is mean zero and independent of all other variables. Substituting we get

$$y_i = \alpha + \beta(\widetilde{x}_i - u_i) + \varepsilon_i = \alpha + \beta\widetilde{x}_i + v_i \tag{7}$$

- The new error term $v_i = \varepsilon_i - \beta u_i$ is correlated with $\widetilde{x}_i$.

Overview of Instrumental Variables

- The basics of IV can be understood with one x and one z.

- Consider the standard regression equation where there are no other control variables

$$y_i = \alpha + \beta x_i + \epsilon_i \tag{8}$$

- Let's define the sample covariance and variance matrices:

$$cov(x_i, y_i) = \frac{1}{n-1} \sum_i (x_i - \overline{x_i})(y_i - \overline{y_i}) \tag{9}$$

$$var(x_i) = \frac{1}{n-1} \sum_i (x_i - \overline{x_i})^2 \tag{10}$$

- The OLS estimator of β is

$$\widehat{\beta}_{OLS} = \frac{cov(x_i, y_i)}{var(x_i)} = \frac{cov(x_i, \alpha + \beta x_i + \epsilon_i)}{var(x_i)} = \beta + \frac{cov(x_i, \epsilon_i)}{var(x_i)} \quad (11)$$

  If x_i and ϵ_i are uncorrelated ($E(x_i\epsilon_i) = 0$), the probability limit of the second term is zero and OLS is a consistent estimator (as the sample size increases, the probability that the OLS estimate is not arbitrarily close to β goes to zero).

- However if x_i and ϵ_i are correlated ($E(x_i\epsilon_i) \neq 0$), the probability limit of $cov(x_i, \epsilon_i)$ does not equal zero and OLS is inconsistent.

- An *instrumental variable*, z_i, is one which is correlated with x_i but not with ϵ_i. The instrumental variables (IV) estimator of β is

$$\widehat{\beta}_{IV} = \frac{cov(z_i, y_i)}{cov(z_i, x_i)} = \frac{cov(z_i, \alpha + \beta x_i + \epsilon_i)}{cov(z_i, x_i)} = \beta + \frac{cov(z_i, \epsilon_i)}{cov(z_i, x_i)} \quad (12)$$

- Given the assumption that z_i and ϵ_i are uncorrelated ($E(z_i\epsilon_i) = 0$), the probability limit of the second term is zero and the IV estimator is a consistent estimator.

- When there is only one instrument, the IV estimator can be calculated using the following procedure: (1) Regress x_i on z_i

$$x_i = \mu + \pi z_i + v_i \tag{13}$$

and form $\widehat{x}_i = \widehat{\mu} + \widehat{\pi} z_i$

Then (2) estimate $\widehat{\beta}$ by running the following regression by OLS

$$y_i = a + \beta \widehat{x}_i + e_i \tag{14}$$

This process is called *Two Stage Least Squares (2SLS)*.

- It is quite easy to show this equivalence:

$$\widehat{\beta}_{2SLS} = \frac{cov(\widehat{x}_i, y_i)}{var(\widehat{x}_i)} = \frac{cov(\widehat{\mu} + \widehat{\pi} z_i, y_i)}{var(\widehat{\mu} + \widehat{\pi} z_i)} \tag{15}$$

$$= \frac{\widehat{\pi} cov(z_i, y_i)}{\widehat{\pi}^2 var(z_i)} = \frac{cov(z_i, y_i)}{\widehat{\pi} var(z_i)} \tag{16}$$

  Given

$$\widehat{\pi} = \frac{cov(z_i, x_i)}{var(z_i)} \tag{17}$$

  This implies that

$$\widehat{\beta}_{2SLS} = \frac{cov(z_i, y_i)}{cov(z_i, x_i)} = \widehat{\beta}_{IV} \tag{18}$$

- The *First Stage* refers to the regression $x_i = \mu + \pi z_i + v_i$.

- The *Reduced Form* refers to the regression $y_i = \theta + \delta z_i + u_i$.

- The *Indirect Least Squares (ILS)* estimator of β is $\widehat{\delta}/\widehat{\pi}$. This also equals $\widehat{\beta}_{IV}$.

- To see this, note that

$$\widehat{\beta}_{ILS} = \frac{\widehat{\delta}}{\widehat{\pi}} = \frac{cov(z_i, y_i)}{var(z_i)} \frac{var(z_i)}{cov(z_i, x_i)} = \frac{cov(z_i, y_i)}{cov(z_i, x_i)} \tag{19}$$

- OLS is often inconsistent because there are omitted variables. IV allows us to consistently estimate the coefficient of interest without actually having data on the omitted variables or even knowing what they are.

- Instrumental variables use only part of the variability in x – specifically, a part that is uncorrelated with the omitted variables – to estimate the relationship between x and y.

- A good instrument, z, is correlated with x for a clear reason, but uncorrelated with y for reasons beyond its effect on x.

Examples of Instruments

- Distance from home to nearest fast food as instrument for obesity.
- Twin births and sibling sex composition as instruments for family size
- Compulsory schooling laws as instruments for education.
- Tax rates on cigarettes as instruments for smoking.
- Weather shocks as instruments for income in developing countries
- Month of birth as instrument for school starting age.

The General Model

- With multiple instruments (overidentification), we could construct several IV estimators.

- 2SLS combines instruments to get a single more precise estimate.

- In this case, the instruments must all satisfy assumptions $E(z_i\epsilon_i) = 0$.

- We can write the models as

$$\begin{aligned} Y &= X\beta + \varepsilon \\ X &= Z\pi + v. \end{aligned}$$

  X is a matrix of exogenous and endogenous variables $(n \times K)$.

Z is a matrix of exogenous variables and instruments $(n \times Q), Q \geq K$.

The 2SLS estimator is

$$\widehat{\beta}_{2SLS} = (X'P_Z X)^{-1} X'P_Z Y \tag{20}$$

where

$$P_Z = Z(Z'Z)^{-1}Z' \tag{21}$$

- It can be shown that the 2SLS estimator is the most efficient IV estimator.

- The **Order Condition** for identification is that there must be at least as many instruments as endogenous variables: $Q \geq K$. This is a necessary but not sufficient condition.

- The **Rank Condition** for identification is that $rank(Z'X) = K$. This is a sufficient condition and it ensures that there is a first stage relationship.

- Example: Suppose we have 2 endogenous variables

$$\begin{aligned} x_1 &= a_1 + a_2 z_1 + a_3 z_2 + u_1 \\ x_2 &= b_1 + b_2 z_1 + b_3 z_2 + u_2 \end{aligned}$$

The order condition is satisfied. However, if $a_2 = 0$ and $b_2 = 0$, the rank condition fails and the model is unidentified. If $a_2 = 0$ and $b_3 = 0$ and the other parameters are non-zero, the rank condition passes and the model is identified.

Variance of 2SLS Estimator

- Recall the 2SLS estimator

$$\begin{aligned} \widehat{\beta}_{2SLS} &= (X' P_Z X)^{-1} X' P_Z y \\ &= \left(\widehat{X}'\widehat{X}\right) \widehat{X}' Y \end{aligned}$$

where $\widehat{X} = P_Z X$ is the predicted value of X from the first stage regression.

- Given this is just the parameter from an OLS regression of Y on $\widehat{X}$, the estimated covariance matrix under homoskedasticity takes the same form as OLS:

$$\widehat{Var}(\widehat{\beta}_{2SLS}) = \widehat{\sigma}^2 \left(\widehat{X}'\widehat{X}\right)^{-1} = \widehat{\sigma}^2 (X' P_Z X)^{-1}$$

where

$$\widehat{\sigma}^2 = \frac{1}{n-K}(Y - X\widehat{\beta}_{2SLS})'(Y - X\widehat{\beta}_{2SLS})$$

- Note that $\widehat{\sigma}^2$ uses X rather than $\widehat{X}$. Shows that standard errors from doing 2SLS manually are incorrect.

- We can simplify this further in the case of the bivariate model from equations (8) and (13).

- In this case, the element in the second row and second column of $X'P_ZX = (X'Z)(Z'Z)^{-1}(Z'X)$ simplifies to (algebra is a bit messy)

$$\frac{n^2 cov(z_i, x_i)^2}{nvar(z_i)}$$

  implying that the relevant element of

$$(X'P_ZX)^{-1} = \frac{1}{n\rho_{xz}^2\sigma_x^2} \tag{22}$$

  where the correlation between x and z equals

$$\rho_{xz} = \frac{cov(z_i, x_i)}{\sigma_x\sigma_z}$$

- Equation (22) tells us that the 2SLS variance

1. Decreases at a rate of $1/n$.

2. Decreases as the variance of the explanatory variable increases.

3. Decreases with the correlation between x and z. If this correlation approaches zero, the 2SLS variance goes to infinity.

4. Is higher than the OLS variance as, for OLS, $\rho_{xz} = 1$ as OLS uses x as an instrument for itself.

Hausman Tests

- Also referred to as Wu-Hausman, or Durbin-Wu-Hausman tests.

- Have wide applicability to cases where there are two estimators and

1. Estimator 1 is consistent and efficient under the null but inconsistent under the alternative.

2. Estimator 2 is consistent in either case but is inefficient under the null.

- We will only consider 2SLS and OLS cases.

- The null hypothesis is that $E(X'\varepsilon) = 0$.

- Suppose we have our model $Y = X\beta + \varepsilon$.

- If $E(X'\varepsilon) = 0$ the OLS estimator provides consistent estimates.

- If $E(X'\varepsilon) \neq 0$ and we have valid instruments, 2SLS is consistent but OLS is not.

- If $E(X'\varepsilon) = 0$ 2SLS remains consistent but is less efficient than OLS.

- Hausman suggests the following test statistic for whether OLS is consistent:

  $$h = \left(\widehat{\beta}_{OLS} - \widehat{\beta}_{2SLS}\right)' \left[V\left(\widehat{\beta}_{2SLS}\right) - V\left(\widehat{\beta}_{OLS}\right)\right]^{-1} \left(\widehat{\beta}_{OLS} - \widehat{\beta}_{2SLS}\right)$$

  which has an asymptotic chi square distribution.

- Note that a nice feature is that one does not need to estimate the covariance of the two estimators.

Hausman Test as Vector of Contrasts (1)

- Compare the OLS estimator $\widehat{\beta}_{OLS} = (X'X)^{-1}X'Y$ to the 2SLS estimator $\widehat{\beta}_{2SLS} = (X'P_ZX)^{-1}X'P_zY$ where P_z is symmetric $n * n$ matrix with rank of at least K.

- Under the null hypothesis $E(X'\varepsilon) = 0$,both are consistent.

$$\begin{aligned}
\widehat{\beta}_{2SLS} - \widehat{\beta}_{OLS} &= (X'P_zX)^{-1}X'P_zY - (X'X)^{-1}X'Y \\
&= (X'P_zX)^{-1}\left[X'P_zY - (X'P_zX)(X'X)^{-1}X'Y\right] \\
&= (X'P_zX)^{-1}X'P_z\left[I - X(X'X)^{-1}X'\right]Y \\
&= (X'P_zX)^{-1}X'P_zM_XY \qquad (23)
\end{aligned}$$

- The probability limit of this difference will be zero when

$$p\lim\frac{1}{n}X'P_zM_XY = 0 \qquad (24)$$

- We can partition the X matrix as $X = [X_1 X_2]$ where X_1 is an $n * G$ matrix of potentially endogenous variables and X_2 is an $n * (K - G)$ matrix of exogenous variables.

- We have instruments Z where $Z = [Z^* X_2]$ an $n * Q$ matrix $(Q \geq K)$.

- Letting hats denote the first stage predicted values, clearly $\widehat{X_2} = X_2$ and $X_2 M_x$ is zero for the rows of M_x corresponding to X_2.

- Therefore checking that $p\lim \frac{1}{n} X' P_z M_X Y = 0$ reduces to checking whether $p\lim \frac{1}{n} X_1' P_z M_X Y = p\lim \frac{1}{n} \widehat{X_1}' M_X Y = 0$.

- We can implement this test using an F-test on δ in the regression:

$$Y = X\beta + \widehat{X_1}\delta + error \tag{25}$$

- Denoting $\delta = 0$ as the restricted model, the F-statistic is

$$H = \frac{RSS_r - RSS_u}{RSS_u/(n-K-G)} \tag{26}$$

- Note from (23), that we can also do the test by regressing $M_X Y$ on $\widehat{X}$ and testing whether the parameters are zero.

Hausman Test as Vector of Contrasts (2)

- Compare the OLS estimator $\widehat{\beta}_{OLS} = (X'X)^{-1}X'Y$ to a different OLS estimator where Z^* is added as a control:

$$Y = X\beta + Z^*\gamma + v \tag{27}$$

- Because of the exclusion restriction, Z^* should have no explanatory power when X is exogenous.

- Using the Frisch-Waugh-Lovell theorem, the resulting estimate of β is $\widehat{\beta}_p$:

$$\widehat{\beta}_p = (X'M_{Z*}X)^{-1}X'M_{z*}Y \tag{28}$$

- Subtracting the OLS estimator,

$$\begin{aligned}
\widehat{\beta}_p - \widehat{\beta}_{OLS} &= (X'M_{z*}X)^{-1}X'M_{z*}Y - (X'X)^{-1}X'Y \\
&= (X'M_{z*}X)^{-1}\left[X'M_{z*}Y - (X'M_{z*}X)(X'X)^{-1}X'Y\right] \\
&= (X'M_{z*}X)^{-1}X'M_{z*}\left[I - X(X'X)^{-1}X'\right]Y \\
&= (X'M_{z*}X)^{-1}X'M_{z*}M_X Y \qquad (29)
\end{aligned}$$

- By analogy with equation (25), we can implement an F-test for whether $p\lim \frac{1}{n}X'M_{z*}M_X Y = 0$ by testing whether $\delta = 0$ in the regression:

$$Y = X\beta + M_{z*}X\delta + error \tag{30}$$

- One can show (not easily) that the resulting F-statistic is identical to that above.

- Alternatively, we can regress $M_X Y$ on $M_{z*}X$ and test whether the parameters are zero.

Two-Sample 2SLS

- Suppose have 2 samples from the same population.
- X is a matrix of exogenous and endogenous variables $(N \times K)$. Z is a matrix of exogenous variables and instruments $(N \times Q), Q \geq K$.
- Sample 1 contains Y and Z but not the endogenous elements of X.
- Sample 2 contains X and Z but not Y.
- Using subscript to denote sample, can implement TS2SLS by

1. Do the first stage regression(s) using Sample 2 observations

$$X_2 = Z_2\pi + v_2 \tag{31}$$

and estimate

$$\widehat{\pi} = (Z_2'Z_2)^{-1}Z_2'X_2 \tag{32}$$

2. Form the predicted value of X_1 as

$$\widehat{X_1} = Z_1\widehat{\pi} \tag{33}$$

3. Regress Y on $\widehat{X_1}$ using Sample 1 observations.

$$\widehat{\beta}_{TS2SLS} = (\widehat{X_1}'\widehat{X_1})^{-1}\widehat{X_1}'Y \tag{34}$$

where

$$\widehat{X_1} = Z_1(Z_2'Z_2)^{-1}Z_2'X_2 \tag{35}$$

- Note that with one endogenous variable and one instrument, we can take an ILS approach.

1. Do step 1. as with TS2SLS to get $\widehat{\pi}$.

2. Estimate the Reduced Form using Sample 1

$$Y = Z_1\delta + u_1 \tag{36}$$

3. Take the ratio of the Reduced Form and First Stage Estimates:

$$\widehat{\beta}_{ILS} = \frac{\widehat{\delta}}{\widehat{\pi}} \tag{37}$$

- To see this note that

$$\begin{aligned}
\widehat{\beta}_{TS2SLS} &= (\widehat{X}_1'\widehat{X}_1)^{-1}\widehat{X}_1'Y \\
&= [X_2'Z_2(Z_2'Z_2)^{-1}Z_1'Z_1(Z_2'Z_2)^{-1}Z_2'X_2]^{-1}X_2'Z_2(Z_2'Z_2)^{-1}Z_1'Y \\
&= (Z_2'X_2)^{-1}(Z_2'Z_2)(Z_1'Z_1)^{-1}(Z_2'Z_2)(X_2'Z_2)^{-1}X_2'Z_2(Z_2'Z_2)^{-1}Z_1' \\
&= (Z_2'X_2)^{-1}(Z_2'Z_2)(Z_1'Z_1)^{-1}Z_1'Y \\
&= [(Z_2'Z_2)^{-1}Z_2'X_2]^{-1}(Z_1'Z_1)^{-1}Z_1'Y \\
&= \widehat{\pi}^{-1}\widehat{\delta} \qquad (3
\end{aligned}$$

- This derivation is valid so long as $Q = K$ so $X_2'Z_2$ and $Z_2'X_2$ are square matrices that we can invert.

- It also uses the matrix inversion rule: $(ABC)^{-1} = C^{-1}B^{-1}A^{-1}$

Example: Devereux and Hart 2010

- Return to education using UK change in compulsory school law.
- If born 1933 or after have minimum leaving age of 15 rather than 14.
- New Earnings Survey has earnings and cohort (Y and Z).
- General Household Survey has education and cohort (X and Z).
- In General Household Survey estimate

$$Education = \alpha_0 + \alpha_1 1(YOB \geq 1933) + W\alpha_2 + e_1 \qquad (39)$$

- In New Earnings Survey estimate

$$Log(earnings) = \gamma_0 + \gamma_1 1(YOB \geq 1933) + W\gamma_2 + e_2 \qquad (40)$$

- Then the TS2SLS estimator of the return to education is $\widehat{\beta} = \frac{\widehat{\gamma_1}}{\widehat{\alpha_1}}$.

- To calculate the standard error we use the delta method.

The Delta Method

- This is a method for estimating variances of functions of random variables using taylor-series expansions.

$$f(x,y) = f(x_0,y_0) + \frac{\partial f(x,y)}{\partial x} |_{x_0,y_0} (x - x_0) + \frac{\partial f(x,y)}{\partial y} |_{x_0,y_0} (y - y_0) + ...$$

- For the case where $f(x,y) = y/x$, $\frac{\partial f(x,y)}{\partial x} = \frac{-y}{x^2}$ and $\frac{\partial f(x,y)}{\partial y} = \frac{1}{x}$.

- Therefore, evaluating at the means of x and y,

$$\frac{y}{x} \simeq \frac{\mu_y}{\mu_x} - \frac{\mu_y}{\mu_x^2}(x - \mu_x) + \frac{1}{\mu_x}(y - \mu_y) \tag{41}$$

- Then,

$$var\left(\frac{y}{x}\right) \simeq \frac{\mu_y^2}{\mu_x^4} var(x) + \frac{1}{\mu_x^2} var(y) - 2\frac{\mu_y}{\mu_x^3} cov(x,y) \tag{42}$$

- In our case

$$var(\widehat{\beta}) \simeq \frac{\widehat{\gamma}_1^2}{\widehat{\alpha}_1^4} var(\widehat{\alpha}_1) + \frac{1}{\widehat{\alpha}_1^2} var(\widehat{\gamma}_1) \tag{43}$$

- Note that the covariance term disappears because the parameters are estimated from 2 independent samples.

The Method of Moments (MOM)

- A *population moment* is just the expectation of some continuous function of a random variable:

$$\gamma = E[g(x_i)] \tag{1}$$

- For example, one moment is the mean: $\mu = E(x_i)$.

- The variance is a function of two moments:

$$\begin{aligned} \sigma^2 &= E[x_i - E(x_i)]^2 && (2) \\ &= E(x_i^2) - [E(x_i)]^2 && (3) \end{aligned}$$

- We also refer to functions of moments as moments.

- A *sample moment* is the analog of a population moment from a particular random sample

$$\widehat{\gamma} = \frac{1}{n}\sum_i g(x_i) \tag{4}$$

- So, the sample mean is $\widehat{\mu} = \frac{1}{n}\sum_i x_i$.

- The idea of MOM is to estimate a population moment using the corresponding sample moment.

- For example, the MOM estimator of the variance using (3) is

$$\widehat{\sigma}^2 = \left(\frac{1}{n}\sum_i x_i^2\right) - \left[\frac{1}{n}\sum_i x_i\right]^2 \tag{5}$$

$$= \frac{1}{n}\sum_i (x_i - \overline{x_i})^2 \tag{6}$$

- This is very similar to our usual estimator of the variance

$$\frac{1}{n-1}\sum_i (x_i - \overline{x_i})^2 \tag{7}$$

- The MOM estimator is biased but is consistent.

- Alternatively, we could calculate the MOM estimator directly using (2)

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_i (x_i - \overline{x_i})^2 \tag{8}$$

OLS as Methods of Moments Estimator

- Our population parameters for linear regression were

$$\beta = E(X_i X_i')^{-1} E(X_i Y_i) \tag{9}$$

- Can derive method of moments estimator by replacing population moments $E(X_i X_i')$ and $E(X_i Y_i)$ by sample moments:

$$b = \left[\frac{1}{n}\sum_i X_i X_i'\right]^{-1} \frac{1}{n}\sum_i X_i Y_i = (X'X)^{-1}X'Y \tag{10}$$

- Or alternatively, we can use the population moment condition

$$E(X_i \varepsilon_i) = E(X_i(Y_i - X_i'\beta)) \tag{11}$$

- The MOM approach is to choose an estimator b so that it sets the sample analog of (11) to zero:

$$\frac{1}{n}\sum_i X_i(Y_i - X_i'b) = 0 \tag{12}$$

This implies that

$$\frac{1}{n}\sum_i X_i Y_i = \frac{1}{n}\sum_i X_i X_i' b \tag{13}$$

So

$$b = \left[\frac{1}{n}\sum_i X_i X_i'\right]^{-1} \frac{1}{n}\sum_i X_i Y_i = (X'X)^{-1}X'Y \tag{14}$$

- Note that this is the OLS estimator.

Generalized Method of Moments (GMM)

- We saw earlier that the OLS estimator solves the moment condition

$$E(X_i(Y_i - X_i'\beta)) = 0 \tag{15}$$

- This moment condition was motivated by the condition $E(X_i\varepsilon_i) = 0$.

- This type of approach can be extended.

- For example, we may know that $E(Z_i\varepsilon_i) = 0$ where Z_i may include some of the elements of X_i.

- The idea of GMM is to substitute out the error term with a function of data and parameters.

- Then find the parameter values that make the conditions hold in the sample.

- Let $\varepsilon_i(\beta) = (Y_i - X_i'\beta)$. We find the parameter such that

$$\frac{1}{n}\sum_i g_i(\beta) = \frac{1}{n}\sum_i Z_i\varepsilon_i(\beta) = \frac{1}{n}Z'(Y - X\beta) \tag{16}$$

  is as close as possible to zero.

- A first guess might be the MOM estimator

$$\widehat{\beta} = (Z'X)^{-1}Z'Y \tag{17}$$

  but this only works if $Z'X$ is invertible and this is only the case if it is a square matrix.

- MOM only works when the number of moment conditions equals the number of parameters to be estimated.

- Instead GMM solves the following problem:

$$\min \left(\frac{1}{n}Z'(Y - X\beta)\right)' W \left(\frac{1}{n}Z'(Y - X\beta)\right) \tag{18}$$

- Here W is called the weight matrix and is some positive definite (PD) square matrix.

- Taking the first order conditions, we get

$$\widehat{\beta} = (X'ZWZ'X)^{-1}X'ZWZ'Y \tag{19}$$

- To see this, note that $(Z'(Y - X\beta))' W (Z'(Y - X\beta))$

$$\begin{aligned} & = (Z'Y - Z'X\beta)'W(Z'Y - Z'X\beta) \\ = \quad & Y'ZWZ'Y - Y'ZWZ'X\beta - \beta'X'ZWZ'Y + \beta'X'ZWZ'X\beta \\ = \quad & Y'ZWZ'Y - 2\beta'X'ZWZ'Y + \beta'X'ZWZ'X\beta \end{aligned}$$

This uses the fact that the transpose of a scalar is itself. Then, taking first order conditions

$$-2X'ZWZ'Y + 2X'ZWZ'X\beta = 0 \tag{20}$$

- $X'ZWZ'X$ will be invertible so long as the number of moment conditions, Q (elements of Z) is as least as big as the number of parameters, K (elements of X).

- For example, not invertible if

$$\begin{aligned} Y_i &= X_{1i}\beta_1 + X_{2i}\beta_2 + \varepsilon_i \quad (21) \\ Z_i &= X_{1i}\alpha_1 \quad (22) \end{aligned}$$

- When $Q > K$, GMM estimates will not cause all moment conditions to equal zero but will get them as close to zero as possible.

- When $Q = K$, as we would expect from (17),

$$\widehat{\beta} = (Z'X)^{-1}Z'Y \tag{23}$$

To see this note that when $Q = K$, $Z'X$ is a square matrix so

$$(X'ZWZ'X)^{-1} = (Z'X)^{-1}W^{-1}(X'Z)^{-1} \tag{24}$$

(remember $(ABC)^{-1} = C^{-1}B^{-1}A^{-1}$). Also note that in this case, W plays no role.

- This is exactly the IV estimator we saw earlier.

- If $X = Z$, the GMM estimator is exactly the OLS estimator.

Consistency of GMM Estimator

- The GMM estimator

$$\begin{aligned}
\widehat{\beta} &= (X'ZWZ'X)^{-1}X'ZWZ'Y && (25)\\
&= \beta + (X'ZWZ'X)^{-1}X'ZWZ'\varepsilon && (26)\\
&= \beta + \left(\frac{X'Z}{n}W\frac{Z'X}{n}\right)^{-1}\left(\frac{X'Z}{n}W\frac{Z'\varepsilon}{n}\right) && (27)
\end{aligned}$$

- Using the Law of Large Numbers (LLN),

$$\begin{aligned}
\frac{X'Z}{n} &= 1/n\sum_i X_i Z_i' \rightarrow \mathbf{\Sigma}_{XZ} && (28)\\
\frac{Z'X}{n} &= 1/n\sum_i Z_i X_i' \rightarrow \mathbf{\Sigma}_{ZX} && (29)\\
\frac{Z'\varepsilon}{n} &= 1/n\sum_i Z_i \varepsilon_i \rightarrow E(Z_i\varepsilon_i) && (30)
\end{aligned}$$

- Denote

$$H = (\mathbf{\Sigma}_{XZ} W \mathbf{\Sigma}_{ZX})^{-1} \mathbf{\Sigma}_{ZX} W \tag{31}$$

- Then

$$\widehat{\beta} - \beta \to HE(Z_i \varepsilon_i) = 0 \tag{32}$$

  showing consistency of GMM for any PD weighting matrix, W.

Choice of Weight Matrix

- Under some regulatory conditions, the GMM $\widehat{\beta}$ is also asymptotically normally distributed for any PD W.

- If the model is overidentified ($Q > K$), the choice of weight matrix affects the asymptotic variance and also the coefficient estimates in finite samples.

- The "best" choice for W is the inverse of the covariance of the moments i.e. the inverse of the covariance matrix of

$$Z'(Y - X\beta) = \sum_i Z_i \varepsilon_i \tag{33}$$

- However, this is unknown and needs to be estimated in the data. We can use a 3-step procedure

1. Choose a weight matrix and do GMM. Any PD weighting matrix will give consistent estimates. A good initial choice is

$$W = (Z'Z/n)^{-1} \tag{34}$$

This gives the estimator

$$\begin{aligned} \widehat{\beta} &= (X'Z(Z'Z)^{-1}Z'X)^{-1}X'Z(Z'Z)^{-1}Z'Y & (35) \\ &= (X'P_zX)^{-1}X'P_zY & (36) \end{aligned}$$

This is exactly the 2SLS estimator we saw earlier.

2. Take the residuals and use them to estimate the covariance of the moments

$$\widehat{Var}(\sum_i Z_i\varepsilon_i) = \frac{1}{n}\sum_i e_i^2 Z_i Z_i' \tag{37}$$

where

$$e_i = Y_i - X_i'\widehat{\beta} \tag{38}$$

3. Do GMM with $\widehat{W}$ as weight matrix where $\widehat{W} = (\frac{1}{n}\sum_i e_i^2 Z_i Z_i')^{-1}$. Low variance moments are given higher weight in estimation than high variance moments.

- Note that if the errors are homoskedastic, this is just the 2SLS estimator.

Variance of GMM Estimator

- In the general case, the GMM estimator is

$$\min g(\beta)' W g(\beta) \tag{39}$$

where the moment conditions are $g(\beta) = 0$.

- The variance of the GMM estimator is

$$\frac{1}{n}(G'WG)^{-1}G'W\Psi WG(G'WG)^{-1} \tag{40}$$

where $G = \frac{\partial g(\beta)}{\partial \beta}$ and Ψ is the variance-covariance matrix of the moments.

- In the OLS case,

$$
\begin{aligned}
g(\beta) &= E(X_i\varepsilon_i) = 0 && (41)\\
\widehat{g}(\beta) &= 1/n\sum_i X_i(Y_i - X_i'\beta) = 0 && (42)\\
W &= I && (43)\\
G &= \frac{\partial \widehat{g}(\beta)}{\partial \beta} = \frac{X'X}{n} && (44)\\
\Psi &= E[(X_i\varepsilon_i)(X_i\varepsilon_i)'] = E[\varepsilon_i^2 X'X] && (45)\\
\widehat{\Psi} &= \widehat{\sigma}_\varepsilon^2 \frac{X'X}{n} && (46)
\end{aligned}
$$

where the last step assumes homoskedasticity. Putting these together we get

$$\widehat{V}(\widehat{\beta}) = \widehat{\sigma}_\varepsilon^2 (X'X)^{-1} \quad (47)$$

- In the 2SLS case,

$$
\begin{aligned}
g(\beta) &= E(Z_i\varepsilon_i) = 0 && (48)\\
\widehat{g}(\beta) &= 1/n\sum_i Z_i(Y_i - X_i'\beta) = 0 && (49)\\
W &= \left(\frac{Z'Z}{n}\right)^{-1} && (50)\\
G &= \frac{\partial \widehat{g}(\beta)}{\partial \beta} = \frac{Z'X}{n} && (51)\\
\Psi &= E[(Z_i\varepsilon_i)(Z_i\varepsilon_i)'] = E[\varepsilon_i^2 Z'Z] && (52)\\
\widehat{\Psi} &= \widehat{\sigma}_\varepsilon^2 \frac{Z'Z}{n} && (53)
\end{aligned}
$$

  where the third and last steps assume homoskedasticity. Putting these together we get

$$\widehat{V}(\widehat{\beta}) = \widehat{\sigma}_\varepsilon^2 (X'P_zX)^{-1} \quad (54)$$

- This formula ignores the fact that the weight matrix is estimated and so

may understate the true variance.

Why Use GMM in Linear Models?

- When the model is just identified, GMM coincides with IV or OLS. So no reason to use GMM.

- In overidentified models with homoskedastic errors,

$$\widehat{W} = (\frac{1}{n}\sum_i e_i^2 Z_i Z_i')^{-1} = \widehat{\sigma}_e^2 \frac{1}{n}\sum_i (Z_i Z_i')^{-1}$$

  and the GMM estimator coincides with 2SLS. So no reason to use GMM.

- In overidentified models with heteroskedasticity, GMM is more efficient than 2SLS.

- Also, in time series models with serial correlation, GMM is more efficient than 2SLS.

- When estimating a system of equations, GMM is particularly useful. You will see this in Kevin Denny's section of the course.

Relationship of GMM to Maximum Likelihood

Maximum Likelihood Interpretation of OLS

- The regression model is

$$Y_i = X_i'\beta + \varepsilon_i \tag{55}$$

- Assume that

$$\varepsilon_i / X_i \tilde{} i.i.d. N(0, \sigma^2) \tag{56}$$
$$X_i \tilde{} i.i.d. g(x) \tag{57}$$

- The likelihood function is the joint density of the observed data evaluated at the observed data values.

- The joint density of $\{Y_i, X_i\}$ is

$$f(y, x) = f(y/x)g(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}(y - x'\beta)^2\right\} g(x) \tag{58}$$

- The likelihood function is

$$L = \prod \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}(Y_i - X_i'\beta)^2\right\} g(X_i) \tag{59}$$
$$= \frac{1}{(\sigma\sqrt{2\pi})^n} \exp\left\{-\frac{1}{2\sigma^2}\sum_i (Y_i - X_i'\beta)^2\right\} \prod g(X_i) \tag{60}$$

- Taking Logs,

$$LogL = -\frac{n}{2}Log(\sigma^2) - \frac{n}{2}Log(2\pi) - \frac{1}{2\sigma^2}\sum_i (Y_i - X_i'\beta)^2 + \sum_i Logg(X_i)$$

- Ignoring the last term which is not a function of β or σ,

$$\begin{aligned} LogL &= -\frac{n}{2}Log(\sigma^2) - \frac{n}{2}Log(2\pi) - \frac{1}{2\sigma^2}(Y - X\beta)'(Y - X\beta) \\ &= -\frac{n}{2}Log(\sigma^2) - \frac{n}{2}Log(2\pi) - \frac{1}{2\sigma^2}\left\{Y'Y - 2\beta'X'Y + \beta'X'X\beta\right\} \end{aligned}$$

- Taking first order conditions of this scalar with respect to β and σ^2

$$\frac{1}{\widehat{\sigma}^2}\left\{X'Y + X'X\widehat{\beta}\right\} = 0 \quad (61)$$

$$-\frac{n}{2\widehat{\sigma}^2} + \frac{1}{2\widehat{\sigma}^4}(Y - X\widehat{\beta})'(Y - X\widehat{\beta}) = 0 \quad (62)$$

- These imply the MLE

$$\widehat{\beta} = (X'X)^{-1}X'Y \tag{63}$$

$$\widehat{\sigma}^2 = \frac{(Y - X\widehat{\beta})'(Y - X\widehat{\beta})}{n} \tag{64}$$

- Note that when taking the first FOC, we are just minimising the sum of squared errors.

GMM Interpretation of Maximum Likelihood

- In the general case, the GMM estimator is

$$\min g(\beta)'W g(\beta) \tag{65}$$

where the moment conditions are $g(\beta) = 0$.

- The FOC are

$$2W\frac{\partial g}{\partial \beta}g(\beta) = 0 \tag{66}$$

- Consider the following moment:

$$g(\beta) = \frac{\partial LogL}{\partial \beta} \tag{67}$$

so

$$\frac{\partial g}{\partial \beta} = \frac{\partial^2 LogL}{\partial \beta \partial \beta'} \tag{68}$$

- The optimal weight matrix is the inverse of the variance-covariance matrix of the moments. In this case,

$$V[g(\beta)] = V\left[\frac{\partial LogL}{\partial \beta}\right] = -E\left[\frac{\partial^2 LogL}{\partial \beta \partial \beta'}\right] \tag{69}$$

and, so, the best estimate of the optimal weighting matrix is

$$\left(\frac{\partial^2 LogL}{\partial\beta\partial\beta'}\right)^{-1} \tag{70}$$

- Substituting these into the FOC, we find that the GMM estimator is defined by $\frac{\partial LogL}{\partial\beta} = 0$, the same as ML.

- So the ML estimator can be seen as a GMM estimator with a particular set of moment equations.

Limited Information Maximum Likelihood (LIML)

- ML version of 2SLS.

- Assumes joint normality of the error terms.

- LIML estimate exactly the same as 2SLS if model is just identified.

- LIML and 2SLS are asymptotically equivalent.

- LIML has better small sample properties than 2SLS in over-identified models.