Introduction

- Grades:
  - Midterm 30%
  - Assignments 20% (about 3 or 4, may use Matlab)
  - Final 50%

- Econometrics is interested in drawing *inferences* about parameters. We have:
  - A sample of observations $x = \{x_1, x_2, \dots, x_n\}$ with joint distribution $f_n(x)$
  - A model (based on econ theory)
    - Parametric: $p_n(x|\theta)$, the likelihood function of x given a parameter θ $\rightarrow$ this is the focus of this course
      - Classic
      - Bayesian
    - Non parametric: not a fixed θ, but a potentially infinite dimension of parameters
    - Semi-parametric: in-between parametric and non-parametric
  - Goal: to draw inference on θ given x.

- Parametric Methods
  1. Classic inference
     - θ has a true value θ_0, which is unknown
       - Assuming that θ_0 exists means that $p_n(x|\theta_0) = f_n(x)$
     - Need to find θ_0
       - Find estimated $\hat{\theta}$
       - Find distribution of $\hat{\theta}$, $p(\hat{\theta}|\theta_0)$
         - Exact finite sample distribution $\rightarrow$ but this is rare
         - Approximations
           - Asymptotic, $n \rightarrow \infty$
           - Bootstrap (draw pseudo-random samples)
  2. Bayesian inference: assumes that θ is a random variable from a probability distribution
     - θ is a r.v.
       - No true value θ_0
       - θ has an *a priori* density $\pi(\theta)$
     - Goal is to find an *a posteriori* density $p_n(\theta|x)$

$$\underbrace{\pi(\theta)}_{a\ priori}\ \underbrace{p_n(x|\theta)}_{\text{likelihood}} = \underbrace{p(x,\theta)}_{\text{joint}}$$

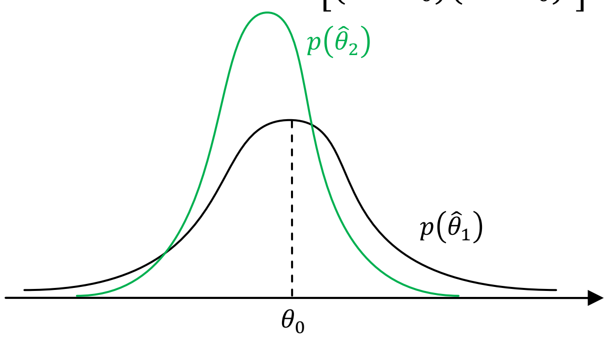
with

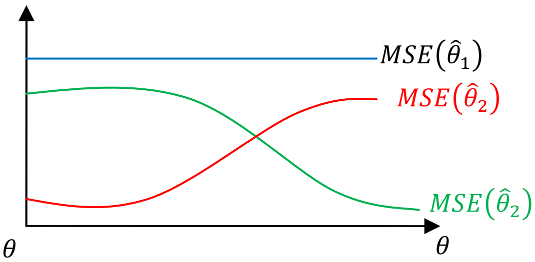
$$p_n(\theta|x) = \frac{p(x,\theta)}{\int p(x,\theta)d\theta}$$

- Properties of estimators (Note: an ***estimator*** is a ***statistic***.)
  - Unbiasness: $E\hat{\theta} = \theta_0$ (1st moment)
  - Efficiency: $\hat{\theta}_1$ is more efficient than $\hat{\theta}_2$ if $Var(\hat{\theta}_1) \leq Var(\hat{\theta}_2)$
    - Classify the estimator according to a loss function
      - Most common: MSE

$$E\left[\left(\hat{\theta}-\theta_0\right)^2\right] = Var\left(\hat{\theta}\right) + Bias^2$$

$$E\left[\left(\hat{\theta}-\theta_0\right)\left(\hat{\theta}-\theta_0\right)'\right] = Var\left(\hat{\theta}\right) + (Bias)(Bias)'$$





- MSE does not provide complete classification
- Clearly, $\hat{\theta}_1$ is ***not admissible*** because it is dominated by $\hat{\theta}_2$ and $\hat{\theta}_3$ across all possible values of θ. However, we cannot rank $\hat{\theta}_2$ and $\hat{\theta}_3$.

Linear Regression Model (Chpt 1)

- Basic assumptions
  - Dependent/LHS variables (Regressands): y
  - Explanatory/RHS variables (Regressors): $\mathbf{x}^k,\ k = 1,2,\dots,K$
  - Goal: explain y as a function of $\mathbf{x}^k,\ k = 1,2,\dots,K$

- Data:
  - Indices:
    - Cross-section: $i = 1,2,\dots,n$
    - Time series: $t = 1,2,\dots,T$
  - Dependent variable: y_i

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{bmatrix}_{(n\times 1)}$$

  - Independent variable: x_{ik}

$$\underbrace{\mathbf{x}^k = \begin{bmatrix} x_{1k} \\ \vdots \\ x_{ik} \\ \vdots \\ x_{nk} \end{bmatrix}_{(n\times 1)}}_{\text{same variable, different observations}} \qquad \underbrace{\mathbf{x}_i = \begin{bmatrix} x_{i1} \\ \vdots \\ x_{ik} \\ \vdots \\ x_{iK} \end{bmatrix}_{(K\times 1)}}_{\text{same observation, different variables}}$$

$$\longrightarrow \underset{(n\times K)}{\mathbf{X}} = [\mathbf{x}^1 \quad \mathbf{x}^2 \quad \cdots \quad \mathbf{x}^K] = \begin{bmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_n' \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1K} \\ x_{21} & x_{22} & \cdots & x_{2K} \\ \vdots & & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nK} \end{bmatrix}_{(n\times K)}$$

$$\begin{array}{cc} \begin{bmatrix} x_{11} & \cdots & x_{1k} & \cdots & x_{1K} \\ \vdots & \ddots & \vdots & & \vdots \\ x_{i1} & \cdots & x_{ik} & \cdots & x_{iK} \\ \vdots & & \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nk} & \cdots & x_{nK} \end{bmatrix} & \begin{array}{c} \\ \\ \mathbf{x}_i' \\ \\ \\ \end{array} \\ \begin{array}{ccccc} \phantom{x_{11}} & & \mathbf{x}^k & & \phantom{x_{1K}} \end{array} & \end{array}$$

- We want to explain y as a linear function of x^k

$$\mathbf{y} \approx \sum_{k=1}^{K} b_k \mathbf{x}^k \iff y_i \approx \sum_{k=1}^{K} b_k x_{ik}$$

i.e. equality between vectors in $\mathbb{R}^n$ component by component.

$$y_i = \sum_{k=1}^{K} b_k x_{ik} + \underbrace{u_i}_{\text{error term}}$$

  - The error term (or disturbance) is connected to the model
    - Note: do not confuse the ***error term*** with the ***residual*** (the realized error)

- The *residual* is connected to the estimation method
  - E.g. OLS estimation → $\hat{b}_k$ with OLS residual
$$y_i - \sum_{k=1}^{K} \hat{b}_k x_{ik} \equiv u_i$$
    GLS estimation → b_k^* with GLS residual
$$y_i - \sum_{k=1}^{K} b_k^* x_{ik} \equiv u_i^*$$
- We need some assumptions about the probability distribution of the r.v.

- Model is implied by restrictive assumptions
  - **Assumption 1.1.** Linearity
$$\mathbf{y} = \mathbf{X}'\mathbf{b} + \mathbf{u}$$
    - Example in the book about wage equation
$$WAGE_i \approx e^{b_1} e^{b_2 S_i} e^{b_3 TEN_i} e^{b_4 EXP_i}$$
$$\log(WAGE_i) = b_1 + b_2 S_i + b_3 TEN_i + b_4 EXP_i + u_i$$
  - **Assumption 1.2.** Exogeneity
$$E(\mathbf{u}|\mathbf{X}) = 0$$
    - Reminder: Law of Iterated Expectations.
$$E\left[\underbrace{E(Z|W)}_{\text{r.v.that depends on } W}\right] = EZ$$
      Here,
$$E(\mathbf{u}|\mathbf{X}) = 0 \underset{\not\Leftarrow}{\Rightarrow} E\mathbf{u} = 0$$
      This is the exogeneity assumption. Note that the implication doesn't go the other way.
    - What happens if $E\mathbf{u} = 0$ but $E(\mathbf{u}|\mathbf{X}) \neq 0$?
      - ***Endogeneity*** or ***simultaneity*** problem. This happens when
        - Some factors are observed by economic agents but not by the econometrician.
        - These observations are taken into account by the agent to determine x_i

Exogeneity Assumption of the Linear Regression Model

❖ The linear regression model

$$y \approx \sum_{k=1}^{K} b_k X^k$$

2 assumptions:

$$(1) \quad y_i = \sum_{k=0}^{K} b_k x_{ik} + u_i\,, \qquad i = 1, \dots, N \quad \text{linearity}$$

$$(2) \quad E(u_i|X) = 0 \;\Rightarrow\; E(u_i) = 0 \qquad \text{(strict) exogeneity}$$

➢ Implications of exogeneity assumption.

1. $E(u_i) = 0$. This is by the ***Law of Iterated Expectations***. Consider

$$\begin{aligned} E\big(E(u_i|X)\big) &= \sum_x \big(E(u_i|X = x)\big)P(X = x) \\ &= \sum_x \left(\sum_u u \cdot P(u_i = u|X = x)\right) P(X = x) \\ &= \sum_x \sum_u u \cdot P(u_i = u|X = x)\, P(X = x) \\ &= \sum_x \sum_u u \cdot P(u_i = u, X = x) \\ &= \sum_u u \cdot \sum_x P(u_i = u, X = x) \\ &= \sum_u u \cdot P(u_i = u) \\ &= E(u_i) \end{aligned}$$

$\because\; E(u_i|X) = 0$ by the exogeneity assumption

$\therefore\; E\big(E(u_i|X)\big) = E(0) = 0 = E(u_i)$

2. In any observation, each regressor is <u>orthogonal</u> to each one of the error terms, i.e.

$$E\big(x_{jk}u_i\big) = 0, \qquad \forall i, j = 1, \dots, n, \qquad \forall k = 1, \dots, K$$

$$E\big(\mathbf{x}_j u_i\big) = \begin{bmatrix} E\big(x_{j1}u_i\big) \\ E\big(x_{j2}u_i\big) \\ \vdots \\ E\big(x_{jK}u_i\big) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}_{(K\times 1)}, \qquad \forall i, j = 1, \dots, n$$

- *Proof*. Since x_{jk} is an element of X, the exogeneity assumption implies that

$$E\big(u_i|x_{jk}\big) = E\big[E(u_i|X)|x_{jk}\big] = 0$$

Then, it follows that

$$E\big(x_{jk}u_i\big) = E\big[E\big(x_{jk}u_i|x_{jk}\big)\big] = E\big[x_{jk}E\big(u_i|x_{jk}\big)\big] = 0.$$

3. Regressors are <u>uncorrelated</u> with the error terms:

$$Cov\big(x_{jk}, u_i\big) = \underbrace{E\big(x_{jk}u_i\big)}_{=0} - E\big(x_{jk}\big)\underbrace{E(u_i)}_{=0} = 0 \;\Rightarrow\; \rho_{x_{jk},u_i} = \frac{Cov\big(x_{jk}, u_i\big)}{\sigma_{x_{jk}}\sigma_{u_i}} = 0, \qquad \forall i, j, k$$

❖ Example. Production.

$$\ln(Q_i) = b_1 + \underbrace{b_2}_{\substack{\text{elasticity}\\ \text{coefficient}}} \ln(L_i) + u_i$$

- $E(u_i|L_i) = 0$ (OK to assume $E(u_i|L_i) = 0$)
- Suppose the firm has information that is not observed by the econometrician

$$u_i = \underbrace{v_i}_{\substack{\text{observed by the firm}\\ \text{but not by the econometrician}}} + \underbrace{w_i}_{\substack{\text{unobserved}\\ \text{by both parties}}}$$

Then,

$$Q_i = L_i^{b_2} \underbrace{e^{b_1} e^{v_i}}_{=A_i} e^{w_i}$$

- Problem of the firm $\rightarrow$ maximize expected profit

$$\max_{L_i}\left[pA_i L_i^{b_2} E(e^{w_i}) - WL_i\right]$$

FOC w.r.t. to L_i

$$pA_i b_2 L_i^{b_2-1} E(e^{w_i}) = W$$

$$\Leftrightarrow \quad L_i = \left(\frac{W}{p}\right)^{\frac{1}{b_2-1}} \left(b_2 A_i E(e^{w_i})\right)^{\frac{1}{-b_2+1}}$$

$$\ln L_i = \underbrace{\frac{1}{b_2-1}\ln\left(\frac{W}{p}\right) + \frac{1}{1-b_2}\ln\left(b_2 E(e^{w_i})\right)}_{=a} + \frac{1}{1-b_2}\ln A_i$$

$$\Leftrightarrow \quad = a + \frac{1}{1-b_2}\underbrace{\ln A_i}_{=b_1+v_i}$$

$$= a + \frac{1}{1-b_2}(b_1 + v_i)$$

$$\Leftrightarrow \quad v_i = (1-b_2)\ln L_i - (1-b_2)a - b_1$$

- Exogeneity assumption:

$$E(\ln(Q_i)\,|L_i) = b_1 + b_2 \ln L_i + \underbrace{E(v_i + w_i|L_i)}_{=0 \text{ by exogeneity}}$$

$$= b_1 + b_2 \ln L_i + E(w_i|L_i) + E(v_i|L_i)$$

- Note that the exogeneity assumption can be stated in two ways
  1. $E(u_i|X) = 0$
  2. $E(y|X) = Xb + \mathbf{0}$

Because w_i is unobserved, it is likely OK to assume $E(w_i|L_i) = 0$. But,

$$E(v_i|L_i) = E[(1-b_2)\ln L_i - a(1-b_2) - b_1|L_i]$$
$$= \underbrace{-a(1-b_2) - b_1}_{=\alpha} + (1-b_2)\ln L_i$$

Then,

$$E(\ln Q_i\,|L_i) = b_1 + b_2 \ln L_i + \alpha + (1-b_2)\ln L_i$$
$$= b_1 + \alpha + 1 \cdot \ln L_i$$

- Therefore, exogeneity implies

$$E(u_i|X_i) = 0 \quad \Leftrightarrow \quad E(y_i|X_i) = X_i'b$$

when $y_i = X_i'b + u_i$

- Here

$$\ln Q_i = b_1 + b_2 \ln L_i + u_i \quad \Rightarrow \quad E(\ln Q_i \,|L_i) = b_1 + b_2 \ln L_i$$

- More generally, when there is a link between the error term, u_i, and the explanatory variables X_i, there will be a ***simultaneity issue*** or ***endogeneity issue***.
  - This issue leads to ***simultaneity bias***
  - In our example, $(1 - b_2)$

- Notations again

$$y_i = x_i' b + u_i, \qquad X = [X^1 \quad X^2 \quad \cdots \quad X^K] = \begin{bmatrix} x_1' \\ x_2' \\ \vdots \\ x_N' \end{bmatrix}$$

- X^k $(N, 1)$-vector, with <u>superscript</u> denoting the k^{th} RHS variable
- x_i $(K, 1)$-vector, with <u>subscript</u> denoting the i^{th} observation

Property of the Matrix X

- **Assumption 1.3.** No multicollinearity. $Rank(X) = K$, i.e. full-rank (or full column rank). It means that the K columns of X (i.e. X^k, $k = 1, \dots, K$) are linearly independent.
  - Note. Max. linearly independent rows = max. linearly indep. Columns = max. size of non-zero (or non-degenerate) minors (or submatrices)

  - Interpretation 1. If there is k_0 such that

$$X^{k_0} = \sum_{k \neq k_0} \alpha_k X^k$$

    X^{k_0} does not help me explain y.
    - In theory, this assumption has cost 0 (i.e. can make this wlog).
    - In practice, we can have almost-perfect multicollinearity. This will create problems when we try to invert the matrix $(X'X)$.

  - Interpretation 2. Suppose $Xb = 0$. Then the no multicollinearity implies necessarily that $b = 0$, i.e. $b_k = \mathbf{0}$. So when we make the assumption of no multicollinearity, we are also assuming that the model is meaningful.

  - Interpretation 3. A theorem in matrix algebra states that if the matrix X is of full column rank, then $X'X$ is non-singular.
    - *Proof.* Let α be a $K \times 1$ vector.

$$\alpha'(X'X)\alpha = (X\alpha)'(X\alpha) = \Bigg\| \underbrace{X\alpha}_{(N\times 1)} \Bigg\|^2 \geq 0$$

      This means that $(X'X)$ is a positive matrix. Thus,

$$\alpha' X' X \alpha = 0 \iff X\alpha = 0$$

$$\because \ X \text{ is full rank}$$

$$\therefore \ \forall k = 1, \dots, K : X^k \text{ is linearly independent}$$

$$\iff \left(\sum_{k=1}^{K} \alpha_k X^k = 0 \ \Rightarrow \ \alpha_k = 0, \qquad \forall k = 1, \dots, K \right)$$

$$\iff (X\alpha = 0 \ \Rightarrow \ \alpha = 0)$$

      - This is a useful way to show that a matrix is non-singular. That is, it is equivalent to showing that it is positive definite (as long as I know it is positive).
        - Let M be an $n \times n$ matrix. Then,
          M is ***positive*** (or ***positive semi-definite***) if for all $(n \times 1)$-vector α

$$\alpha' M \alpha \geq 0$$

          M is ***positive definite*** if for all $(n \times 1)$-vector α

$$\alpha' M \alpha \geq 0 \ \text{ and } \ \alpha' M \alpha = 0 \iff \alpha = 0$$

  - Interpretation 4. $X'X = \sum_{i=1}^{N} x_i x_i'$. Suppose x_i are iid. Then by the WLLN

$$\underbrace{\frac{1}{N} \sum_{i=1}^{N} x_i x_i'}_{\text{sample mean}} \xrightarrow{p} \underbrace{E(x_i x_i')}_{\text{true mean}}$$

- **Slutsky's Theorem.**

$$\underbrace{\det\left(\frac{1}{N}\sum_{i=1}^{N} x_i x_i'\right)}_{d_N} \xrightarrow{p} \underbrace{\det[E(x_i x_i')]}_{d}$$

Recall convergence in probability

$$P(|d_N - d| > \epsilon) \xrightarrow{N\to\infty} 0, \qquad \forall \epsilon > 0$$

If we know that $d > 0$,

$$P\left(|d_N - d| > \frac{d}{2}\right) \xrightarrow{N} 0 \;\Leftrightarrow\; P\left(|d_N - d| \le \frac{d}{2}\right) \xrightarrow{N} 1$$

$$\Rightarrow\; P\left(d_N \ge \frac{d}{2}\right) \xrightarrow{N} 1$$

So $Rank(X) = K \;\Leftrightarrow\; \det(X'X) > 0$. Thus, for N large enough, $\frac{1}{N}\sum_{i=1}^{N} x_i x_i'$ should be invertible (as long as $\det[E(x_i x_i')] > 0$).

- Question: Why should we maintain this assumption?
  - $E(x_i x_i')$ is $\oplus$ (positive semi-definite):

$$\alpha' E(x_i x_i')\alpha = E[(\alpha' x_i)(x_i \alpha)] = E[(x_i'\alpha)^2] \ge 0$$

  - Positive definite?

$$\alpha' E(x_i x_i')\alpha = 0 \;\Leftrightarrow\; E[(x_i'\alpha)^2] = 0 \;\Leftrightarrow\; x_i'\alpha = 0 \;\; a.s.$$

  $\alpha \neq 0$ only if there are multicollinearities.

Geometric & Statistical Interpretations of Least Squares

❖ Naïve approach:

$$y_i \approx \sum_{k=1}^{K} b_k x_{ik}$$

Choose $\hat{b}$ such that

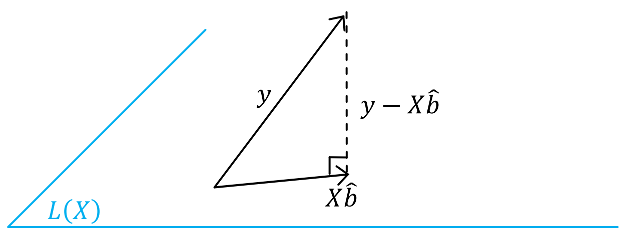
$$\sum_{i=1}^{n}\left(y_i - \sum_{k=1}^{K}\hat{b}_k x_{ik}\right)^2 \le \sum_{i=1}^{n}\left(y_i - \sum_{k=1}^{K} b_k x_{ik}\right)^2, \qquad \forall b_k$$

➢ Note: by definition of OLS,

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sum_{k=1}^{K}\hat{b}_k x_{ik}\right)^2 \le \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sum_{k=1}^{K} b_k^0 x_{ik}\right)^2 \iff \frac{1}{n}\sum_{(i=1)}^{n}\hat{u}_i^2 \le \frac{1}{n}\sum_{i=1}^{n} u_i^2$$

So the problem is

$$\min_{b\in\mathbb{R}^K}\sum_{i=1}^{n}\left(y_i - \sum_{k=1}^{K} b_k x_{ik}\right)^2 \iff \min_{b\in\mathbb{R}^K}\left[(y - Xb)'\underbrace{(y - Xb)}_{\in\mathbb{R}^n}\right].$$



$L(X) = \{Xb : b \in \mathbb{R}^K\} = \{\sum_{k=1}^{K} b_k X^k : b_k \in \mathbb{R},\ \forall k\}$ and $Xb \in \mathbb{R}^n$
$L(X)$ is a subspace of $\mathbb{R}^n$

➢ $\hat{b}$ is characterized by K orthogonal relationships

$$\underbrace{\left(y - X\hat{b}\right)}_{\in\mathbb{R}^n} \perp \underbrace{X^k}_{\in\mathbb{R}^n}, \qquad \forall k \iff X^{k\prime}\cdot\left(y - X\hat{b}\right) = 0$$

$$\iff \underbrace{X'\left(y - X\hat{b}\right) = 0}_{\text{system of } K \text{ linear equations}}$$

$$\iff X'y = X'X\hat{b}$$

$$\iff \boxed{\hat{b} = (X'X)^{-1}X'y}$$

❖ Notations: matrix of orthogonal projection

$\boldsymbol{P_X}$: ***matrix of orthogonal projection on*** $\boldsymbol{L(X)}$

$$P_X y = X\hat{b} = (X'X)^{-1}X'y$$

➢ Should be true for all y

➢ Can identify P_X as

$$P_X = X(X'X)^{-1}X'$$

➢ Matrix of orthogonal projection on the orthogonal of $L(X)$, i.e. $L(X)^{\perp}$:

$$M_X = I_n - P_X \ \Rightarrow\ M_X y = (I_n - P_X)y$$

$$
\begin{aligned}
&= y - P_X y \\
&= y - X\hat{b} \\
&= \hat{u} \;\longrightarrow\; \text{OLS residual}
\end{aligned}
$$

- Note that

$$
\begin{aligned}
y = Xb^0 + u \;\;&\Leftrightarrow\;\; u = y - Xb^0 \\
&\Rightarrow\;\; M_X u = \hat{u} - \underbrace{M_X X b^0}_{=0\ \because Xb^0 \in L(X)}
\end{aligned}
$$

Mathematically,

$$
\begin{aligned}
M_X X b^0 &= I_n X b^0 - X \underbrace{(X'X)^{-1}X'X}_{=I_n} b^0 \\
&= Xb^0 - Xb^0 \\
&= 0
\end{aligned}
$$

- Properties of Projection Matrices
  - *Idempotence* and *symmetry* are necessary and sufficient conditions for projection matrices
    - Symmetry: $A = A'$
    - Idempotence: $A^{\ell} = A$ for any $\ell \in \mathbb{N}$

- Application: the Frish-Waugh Theorem (cf. P.72, ex.4)
  - Motivation:

$$
y = Xb + Zc + u = [X \,\vdots\, Z] \binom{b}{c} + u
$$

where X is $(n \times K_1)$ and Z is $(n \times K_2)$.

Geometric Interpretation of Linear Regression (cont'd)

- Consider the true model:

$$y = Xb + Zc + u = [X \vdots Z]\begin{pmatrix} b \\ c \end{pmatrix} + u$$

  - Question: How bad/wrong is it to regress Y on X only?
  - Assumptions for the true model:
    (i) $E(u|X, Z) = 0$
    (ii) $Rank([X \vdots Z]) = K_1 + K_2$
  - Assumptions for the "reduced" model without Z: $y = X\beta + v$
    (1) $E(v|X) = 0$
    (2) $Rank(X) = K_1$

- Compare (i) and (ii) to (1) and (2).
  - Easy to see that (ii) ⇒ (2)
  - What about (i) v.s. (1)?

$$\begin{aligned} E(y|X) &= E(Xb + Zc + u|X) \\ &= Xb + E(Z|X)c + \underbrace{E(u|X)}_{=0} \\ &= Xb + E(Z|X)c \end{aligned}$$

    - In order to get (1), I need $E(Z|X)$ to be linear in X, i.e.

$$E(Z|X) = X\Gamma$$

      Thus, the true model doesn't imply the "reduced" model

- Even if the "reduced" model is not necessarily implied by the true model, I can still perform OLS:

$$OLS_{Reduced} \qquad \hat{\beta} = (X'X)^{-1}X'y$$

$$OLS_{true} \qquad \begin{pmatrix} \hat{b} \\ \hat{c} \end{pmatrix} = \left[\begin{pmatrix} X' \\ Z' \end{pmatrix}(X \quad Z)\right]^{-1}\begin{pmatrix} X' \\ Z' \end{pmatrix} y$$

  - $\hat{b} =?$ $\hat{b}$ $v.s.$ $\hat{\beta}$?

- **Frish-Waugh Theorem.**

$$\hat{b} = (X'M_2X)^{-1}X'M_2y$$

  - *Proof.*

$$\begin{pmatrix} \hat{b} \\ \hat{c} \end{pmatrix} = \arg\min_{\binom{b}{c}} \|y - (Xb + Zc)\|^2$$

    - Step 1. Concentrate w.r.t. c. For given b, minimize w.r.t. c only. Get $c(b)$

$$\min_c \|(y - Xb) - Zc\|^2$$

      Get $c(b)$ such that

$$Zc(b) = P_Z(y - Xb)$$

    - Step 2. Minimize the concentrated objective function w.r.t. b.

$$\min_b \|y - Xb - P_Z(y - Xb)\|^2 \iff \min_b \left\| \underbrace{M_Z y}_{\tilde{y}} - \underbrace{M_Z X}_{\tilde{X}} b \right\|^2$$

$$\Rightarrow \hat{b} = \left(\tilde{X}'\tilde{X}\right)^{-1}\tilde{X}'y = (X'M_ZX)^{-1}X'M_Zy$$

- Remark. In Practice,
$$y/Z \ \Rightarrow \ P_Z y \ \Rightarrow \ M_Z y \ residual$$
$$X^k/Z \ \Rightarrow \ P_Z X^k \ \Rightarrow \ M_Z X^k \ residual$$
Then,
$$M_Z y \Big/ \underbrace{M_Z X^k}_{M_Z X}, \qquad \forall k$$
- Remark. Corollary of FW Thm. If $M_Z X = X$, then
$$\hat{b} = (X'X)^{-1}X'y, \qquad (= \hat{\beta})$$
So $M_Z X = X$. This means that
$$\begin{aligned} X \perp Z \ &\Leftrightarrow \ X^k \perp Z^\ell, \qquad \forall k, \ell \\ &\Leftrightarrow \ \sum_{i=1}^{n} x_{ik} z_{i\ell} = 0, \qquad \forall k, \ell \\ &\Leftrightarrow \ \frac{1}{n}\sum_{i=1}^{n} x_{ik} z_{i\ell} = 0, \qquad \forall k, \ell \end{aligned}$$
But this does not mean that X^k and Z^ℓ are not correlated. Not quite Cov between X^k and Z^ℓ, only orthogonality condition.

- Next we consider the statistical interpretation of OLS
  - Finite sample properties
  - Interpretation

Finite Sample Properties of OLS

- Recall the standard assumptions
  (1) $y = Xb + u$
  (2) $E(u|X) = 0,\ a.s.$
  (3) $|X'X| > 0,\ a.s.$

  2 additional assumptions:
  (4) $Var(u|X) = \sigma^2 I_n,\ a.s.\ \rightarrow$ ***homoscedasticity***
  (5) $u|X \sim \mathcal{N}(0, \sigma^2 I_n)$

- Properties:
  - Under assumption (1) – (3)
$$E(\hat{b}) = b$$
    - *Proof.* Use the law of iterated expectations:
$$\begin{aligned} E(\hat{b}) &= E\left[E(\hat{b}|X)\right] \\ &= E\left[E[(X'X)^{-1}X'y|X]\right] \\ &= E[(X'X)^{-1}X'E(y|X)] \\ &= E[(X'X)^{-1}X'Xb] \\ &= b \end{aligned}$$
  - Under assumption (1) – (4)
$$Var(\hat{b}) = E[(X'X)^{-1}\sigma^2]$$
    - *Proof.*
$$\begin{aligned} Var(\hat{b}) &= E\left[(\hat{b} - E\hat{b})(\hat{b} - E\hat{b})'\right] \\ &= E\left[(\hat{b} - b)(\hat{b} - b)'\right] \\ &= E[(X'X)^{-1}X'uu'X(X'X)^{-1}] \\ &= E\left[(X'X)^{-1}X' \underbrace{E[uu'|X]}_{=\sigma^2 I_n} X(X'X)^{-1}\right] \\ &= \sigma^2 E[(X'X)^{-1}] \end{aligned}$$
      where
$$\begin{aligned} \hat{b} - b &= (X'X)^{-1}X'y - b \\ &= (X'X)^{-1}X'(Xb + u) - b \\ &= (X'X)^{-1}X'u \end{aligned}$$

- **Gauss-Markov Theorem (BLUE).**
  Under assumptions (1) – (4), the OLS $\hat{b}$ is ***BLUE*** (***best linear unbiased estimator***). That is, for any linear (w.r.t. y) estimator,
$$\tilde{b} = Cy, \qquad s.t.\ E\tilde{b} = b$$
  we have
$$Var(\hat{b}) \ll Var(\tilde{b}) \ \Leftrightarrow\ \left(Var(\tilde{b}) - Var(\hat{b})\right) \text{ is psd.}$$

- **Cramer-Rao Theorem.**
  Under assumptions (1) – (5), $\hat{b}$ is ***BUE*** (***best unbiased estimator***).

- Property: Under assumption (1) – (5)
$$\hat{b} - b \big| X \sim \mathcal{N}(0, \sigma^2 (X'X)^{-1})$$
  - *Proof.* Since $u \sim \mathcal{N}(0, \sigma^2 I)$,
$$\hat{b} - b = (X'X)^{-1} X'u$$
Note this is not the same as the central limit theorem, because we have a finite sample.

- Remark. If we want the unconditional distribution of $\left(\hat{b} - b\right)$, there are two options:
  - Assume fixed regressors
  - Asymptotic theory (i.e. as $n \to \infty$)
- Remark. Standardize $\left(\hat{b} - b \big| X\right)$:
$$\frac{(X'X)^{1/2}}{\sigma} \left(\hat{b} - b \big| X\right) \sim \mathcal{N}(0, I)$$
  - This does not depend on X.
  - This is the unconditional distribution, i.e.
$$\frac{(X'X)^{1/2}}{\sigma} \left(\hat{b} - b\right) \sim \mathcal{N}(0,1)$$
  - In general, σ is unknown! $\rightarrow$ we need an estimator of σ to use the previous result.
$$\hat{\sigma}^2 = \frac{1}{n-K} \sum_{i=1}^{n} \hat{u}^2 = \frac{1}{n-K} u' M_X u$$
    - *Proof* ($\hat{\sigma}^2$ is unbiased).
$$E(\hat{\sigma}^2) = \frac{1}{n-K} E \underbrace{(u' M_X u)}_{\text{univariate}} = \frac{1}{n-K} E[\operatorname{tr}(u' M_X u)] = \frac{1}{n-K} E(\operatorname{tr}(uu' M_X))$$
$$= \frac{1}{n-K} E[E(\operatorname{tr}(uu' M_X))|X] = \frac{1}{n-K} E\left[\operatorname{tr}\left(\underbrace{E(uu'|X)}_{=\sigma^2 I} M_X\right)\right]$$
$$= \frac{\sigma^2}{n-K} E[\operatorname{tr}(M_X)]$$

$$M_X = I_n - X(X'X)^{-1}X'$$

$$\begin{aligned} \operatorname{tr}(I_n - X(X'X)^{-1}X') &= \operatorname{tr}(I_n) - \operatorname{tr}(X(X'X)^{-1}X') \\ &= n - \operatorname{tr}\left[\underbrace{(X'X)(X'X)^{-1}}_{I_K}\right] \\ &= n - K \end{aligned}$$

where the trace has the following property:
$$\operatorname{tr}(ABC) = \operatorname{tr}(BAC) = \operatorname{tr}(CAB)$$

Statistical Interpretations of OLS

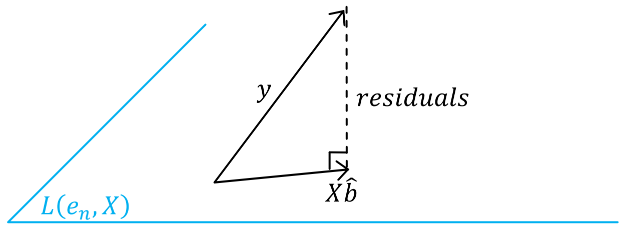
❖ Recall the model

$$y_i = \underbrace{a}_{\text{constant term}} + \sum_{k=1}^{K} b_k x_{ik} + u_i$$

Matrix of explanatory variables:

$$\left[\begin{pmatrix}1\\1\\\vdots\\1\end{pmatrix} X^1 X^2 \cdots X^K\right] = \left[\underbrace{e_n}_{\text{vector of 1's}} \;\vdots\; X\right]$$

➢ Adding a column of 1's to the regressors makes the linear regression a *affine* regression.



➢ Orthogonal conditions between the residuals and the explanatory variables.

$$\begin{cases} e_n'(y - \hat{a}e_n - X\hat{b}) = 0 \\ X^{k\prime}(y - \hat{a}e_n - X\hat{b}) = 0 \end{cases}, \qquad \forall k = 1, \dots, K$$

We have $K + 1$ linear equations to find $K + 1$ parameters $\hat{b}$ and $\hat{a}$. Note that

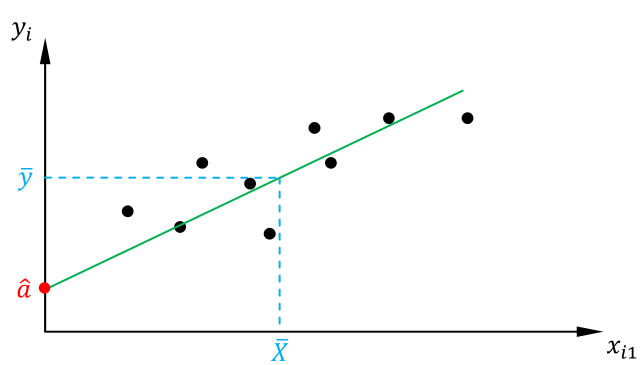
$$e_n'(y - \hat{a}e_n - X\hat{b}) = 0 \;\Leftrightarrow\; \sum_{i=1}^{n} y_i - n\hat{a} - \sum_{k=1}^{K} \hat{b}_k \left(\sum_{i=1}^{n} x_{ik}\right) = 0$$

$$\text{divide by } n \;\Leftrightarrow\; \bar{y} - \hat{a} - \sum_{k=1}^{K} \hat{b}_k \bar{X}^k = 0$$

$$\Leftrightarrow\; \boxed{\hat{a} = \bar{y} - \sum_{k=1}^{K} \hat{b}_k \bar{X}^K}$$

$$\Leftrightarrow\; \boxed{\bar{\hat{u}} = 0}$$

- $\bar{\hat{u}}$ is the empirical average of the OLS residuals. So by including a constant term, we are imposing that on average ??? is correct.



- With a constant term, the average residual is zero, and the average point belongs to the regression line
- If we do not introduce the constant term,
$$\hat{y} = X\hat{b}$$
Then the regression line goes through $(0,0)$. However, this does not minimize the sum of squares.
    - If, in addition, you want the average point to belong to the regression line, then the regression line is completely determined.

➢ Plug in the expression for $\hat{a}$

$$y - \hat{a}e_n - X\hat{b} = \underbrace{(y - \bar{y}e_n)}_{\substack{\tilde{y} \\ \text{mean deviation}}} - \sum_{k=1}^{K} \hat{b}_k \underbrace{(X^k - \bar{X}^k e_n)}_{\substack{\tilde{X}^k \\ \text{mean deviation}}}$$

Rewrite the orthogonal condition

$$X'(\tilde{y} - \tilde{X}\hat{b}) = 0 \iff \frac{1}{n}X'(\tilde{y} - \tilde{X}\hat{b}) = 0$$
$$\iff \frac{1}{n}X^{k\prime}(\tilde{y} - \tilde{X}\hat{b}) = 0, \qquad \forall k = 1, \dots, K$$

where

$$\frac{1}{n}X^{k\prime}\tilde{y} = \frac{1}{n}\sum_{i=1}^{n} x_{ik}(y_i - \bar{y})$$

is the sample (empirical) covariance between x_{ik} and y_i. Notation: $Cov_{emp}(x_{ik}, y_i)$.

$$\hat{b} = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'y = \underbrace{\left(\frac{1}{n}\tilde{X}'\tilde{X}\right)^{-1}}_{Var_{emp}(x_i)} \underbrace{\left(\frac{1}{n}\tilde{X}'\tilde{y}\right)}_{Cov_{emp}(x_i, y_i)}$$

- The OLS estimator is the ratio of the $Cov_{emp}(x_i, y_i)$ and $Var_{emp}(x_i)$.
- We can make the interpretation of Var-Cov only because we include the constant term in the regression. Without the constant term, we only have the orthogonality.

Geometric Interpretation of OLS in the Space of Random Variables

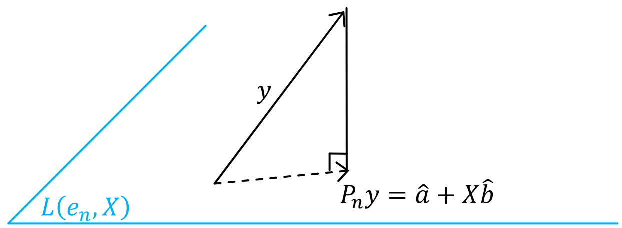
❖ Recall

$$(\hat{a},\hat{b}) = \underbrace{\arg\min_{a,b}\left[\frac{1}{n}\sum_{i=1}^{n}(y_i - a - x_i'b)^2\right]}_{\text{space of } \mathbb{R}^n}$$

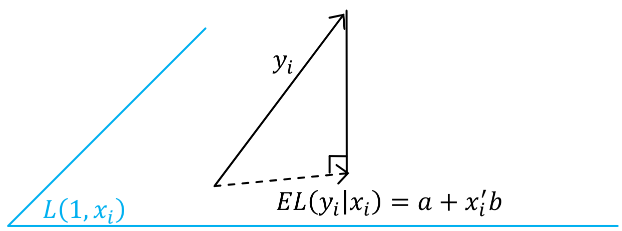
$$\downarrow \qquad\qquad\qquad \downarrow$$

$$(a^0,b^0) \overset{?}{=} \arg\min_{a,b}\underbrace{E(y_i - a - x_i'b)}_{\text{space of r.v.}}$$

➢ In $\mathbb{R}^n$:



➢ In space of r.v. that are square integrable,



- The norm (or distance) is

$$\|u\| = \sqrt{Eu^2},$$
$$\langle u,v\rangle = E(uv), \qquad \text{(the inner product)}$$

The inner product is the extension of the norm, so $\langle u,u\rangle = \|u\|$.

❖ Computation of $EL(y_i|x_i)$:

$$\begin{cases}[y_i - (a + x_i'b)] \perp 1 \\ [y_i - (a + x_i'b)] \perp x_{ik} \quad \forall k = 1,\dots,K\end{cases} \Leftrightarrow \begin{cases}E(y_i - a - x_i'b) = 0 & (*) \\ E[(y_i - a - x_i'b)x_{ik}] = 0 & \forall k\end{cases}$$

From $(*)$ we get

$$a = E(y_i) - E(x_i')b$$

Plug it into the second expectation:

$$E[(y_i - \{Ey_i - a - Ex_i'b\})x_i] = 0 \;\Rightarrow\; E[(y_i - Ey_i)x_i] - E\left[\underbrace{\overbrace{(x_i - Ex_i)'}^{1\times K}\overbrace{b}^{K\times 1}}_{1\times 1}x_i\right] = 0$$

$$\Rightarrow \underbrace{E[x_i(x_i - Ex_i)']}_{Var(x_i)}b = \underbrace{E[(y_i - Ey_i)x_i]}_{Cov(x_i,y_i)}$$

If $Var(x_i)$ is nonsingular,

$$b = \left(Var(x_i)\right)^{-1}Cov(x_i,y_i)$$

➢ This is NOT an estimate!!! It is a population value (as opposed to a sample value)

- Conclusion:

$$EL(y_i|x_i) = a^0 + x_i'b^0$$

with

$$\begin{cases} a^0 = Ey_i - Ex_i'b^0 \\ b^0 = \left(Var(x_i)\right)^{-1} Cov(x_i, y_i) \end{cases}$$

same formula as the one we got for the estimates $\hat{a}, \hat{b}$ in the space of $\mathbb{R}^n$. But now we have population moments.

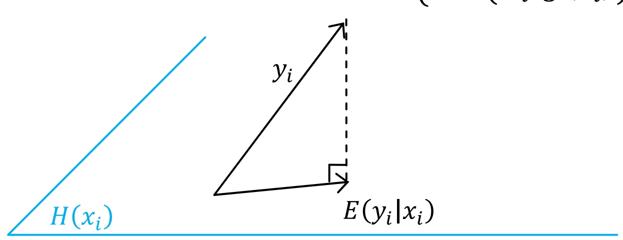
- Remark. Why do we have $Var(x_i)$ nonsingular?
  - $Var(x_i)$ is PSD $\alpha \in \mathbb{R}^k$, $\alpha' Var(x_i)\alpha = Var(\alpha' x_i) \geq 0$
  - $Var(\alpha' x_i) = 0 \;\Leftrightarrow\; \underbrace{\alpha' x_i}_{r.v.}$ constant
  - Therefore, $Var(x_i)$ is nonsingular if and only if no linear combination of x_i is constant (except if $\alpha = 0$)
  - $\frac{1}{n}\tilde{X}'\tilde{X} = Var_{emp}(\tilde{X}) \xrightarrow{p} Var(x_i)$ [under iid assumption]

- Remark. $y_i = a^0 + x_i'b^0 + u_i$
  - $Eu_i = 0$
  - $E(u_i x_i) = 0$ or $Cov(x_i, u_i) = 0$
  - a^0, b^0 are defined by the above
  - These are not assumptions
  - $EL(y_i|x_i)$ is the best linear predictor of y_i as *an affine combination of* x_i; that is, it is the solution of the minimization of

$$E\left[\left(y_i - EL(y_i|x_i)\right)^2\right] = E[(y_i - a - x_i'b)^2]$$

- Remark. $E(y_i|x_i)$ is the best predictor of y_i as *a function of* x_i; that is, it is the solution of the minimization of

$$E\left[\left(y_i - f(x_i)\right)^2\right], \qquad y_i = f(x_i) + w_i, \qquad \begin{cases} Ew_i = 0 \\ Cov\left(w_i, g(x_i)\right) = 0, \qquad \forall g \end{cases}$$



Claim: the solution of the minimization problem is

$$\left(y_i - E(y_i|x_i)\right) \perp g(x_i), \qquad \forall g$$

Proof. Consider

$$E\left(y_i g(x_i) - E(y_i|x_i)g(x_i)\right) \stackrel{?}{=} 0$$

$$\Rightarrow \; E\left(y_i g(x_i)\right) - E[E(y_i g(x_i)|x_i)] = E\left(y_i g(x_i)\right) - E\left(y_i g(x_i)\right) = 0$$

- **Corollary 1.** $EL(y_i|x_i) = E(y_i|x_i)$ if and only if $E(y_i|x_i)$ is affine with respect to x_i.

- **Corollary 2.**

$$EL\left\{E(y_i|x_i)\middle|x_i\right\} = EL(y_i|x_i)$$

Proof. Need to show that

$$\left[\underbrace{E(y_i|x_i)}_{\text{what I project}} - \underbrace{EL(y_i|x_i)}_{\substack{\text{candidate for}\\ \text{projection}}}\right] \perp (1, x_i)$$

Introduce y_i:

$$\underbrace{[E(y_i|x_i) - y_i]}_{\perp(1,x_i)} - \underbrace{[y_i - EL(y_i|x_i)]}_{\perp(1,x_i)}$$

- Final comments.
  - Exogeneity assumption in $y_i = a^0 + x_i' b^0 + u_i$:
    - Strict exogeneity: $E(u_i|x_i) = 0$
    - Weaker exogeneity: $E(u_i) = 0$, $Cov(f(x_i), u_i) = 0$ for all f
  - Is it true that $E(y_i|x_i)$ is linear?
    - True with: $E(u_i|x_i) = 0 \;\; \Rightarrow \;\; x_i$ and u_i are stochastically independent
  - What do we do when it is not true (i.e. $E(y_i|x_i)$ is not linear)?
    - Add some terms to account for nonlinear effects, e.g. x_{ik}^2 or $x_{ik}x_{i\ell}$, or more complicated functional form

Large Sample Theory (Chapter 2.3)

- Maintained assumption: (y_i, x_i) are jointly identically distributed (i.d.)
  - Consequence: (y_i, x_i, u_i) are jointly identically distributed
  - Remark. $Var(u_i) = \sigma^2$, which is a constant (i.e. independent of i). But there can be heterogeneity at the conditional level, i.e.

$$Var(u_i|x_i) = \sigma^2(x_i).$$

    - With the assumption $E(u_i|x_i) = 0$, we can write:

$$Var(u_i) = E(u_i^2) = E\left(E(u_i^2|x_i)\right) = E\big(Var(u_i|x_i)\big) \;\Rightarrow\; \sigma^2 = E\big(\sigma^2(x_i)\big)$$

- Law of Large Numbers (LLN). Consider z_i, $i \in \mathbb{N}$, i.d. and integrable (i.e. $E|z_i| < \infty$).
  - LLN:

$$\frac{1}{n}\sum_{i=1}^{n} z_i = \bar{z}_n \longrightarrow Ez_i = Ez_1$$

  - **Theorem 1 (SLLN).** Let $(z_i)_{i\in\mathbb{N}}$ iid and integrable.

$$\bar{z}_n \overset{a.s.}{\longrightarrow} Ez_i$$

    Recall the definition of almost sure convergence:

$$P\left(\lim_{n\to\infty}(\bar{z}_n - Ez_i) = 0\right) = 1.$$

  - L^2-LLN:

$$\bar{z}_n \overset{L^2}{\longrightarrow} Ez_i$$

    Recall the definition of L^2 convergence:

$$w_n \overset{L^2}{\longrightarrow} w_0 \;\Leftrightarrow\; E(w_n - w_0)^2 \underset{n\to\infty}{\longrightarrow} 0$$

    This requires w_n to be L^2 integrable.
    - Note:

$$E(w_n - w_0)^2 = \underbrace{Var(w_n - w_0)}_{\geq 0} + \underbrace{[E(w_n - w_0)]^2}_{\geq 0}$$

      So

$$E(w_n - w_0)^2 \underset{n\to\infty}{\longrightarrow} 0 \;\Leftrightarrow\; \begin{cases} Var(w_n - w_0) \underset{n\to\infty}{\longrightarrow} 0 \\ E(w_n - w_0) \underset{n\to\infty}{\longrightarrow} 0 \end{cases}$$

      In our case: Ez_i is a constant.

$$\bar{z}_n \overset{L^2}{\longrightarrow} Ez_i \;\Leftrightarrow\; Var(\bar{z}_n) \longrightarrow 0$$

$$\begin{aligned} Var(\bar{z}_n) &= Var\left(\frac{1}{n}\sum_{i=1}^{n} z_i\right) \\ &= \frac{1}{n^2}\sum_{i=1}^{n} Var(z_i) + \frac{1}{n^2}\sum_{\substack{i,j=1 \\ i\neq j}}^{n} Cov\big(z_i, z_j\big) \end{aligned}$$

$$= \underbrace{\frac{\sigma^2}{n}}_{\xrightarrow[n\to\infty]{}0} + \frac{2}{n^2} \underbrace{\sum_{1\le i<j\le n} Cov(z_i, z_j)}_{n(n-1) \text{ terms}}$$

- **Theorem 2 (L^2-LLN).** If $(z_i)_{i\in\mathbb{N}}$ such that $Ez_i^2 < \infty$, and Ez_i and $Var(z_i)$ independent on i, then

$$\bar{z}_n \xrightarrow{L^2} Ez_i \iff \frac{1}{n^2} \sum_{1\le i<j\le n} Cov(z_i, z_j) \xrightarrow[n\to\infty]{} 0$$

- **Theorem 3.** Both the SLLN and L^2-LLN imply the WLLN (convergence in probability).
  - Note that there is no clear logical relation between almost sure convergence and L^2 convergence.

- Consistency of OLS estimators (i.e. estimator of b and σ^2)

$$y_i = x_i'b + u_i, \qquad with \begin{cases} (y_i, x_i) \ i.d. \\ E(u_i|x_i) = 0 \\ \sigma^2 = Var(u_i) \end{cases}$$

  - We know 1 estimator, i.e. the OLS, by solving

$$\underbrace{X'}_{K\times n}(y - Xb) = 0$$

    Want to compare the OLS estimator with the IV-estimation

$$W'(y - Xb) = 0$$

    where $W = [w^1 \ w^2 \ \cdots \ w^H]$
    - OLS is a special case for IV-estimation.

  - Motivation:

$$E(y_i - x_i'b^0|x_i) = 0 \iff E[f(x_i)(y_i - n_i'b^0)] = 0, \qquad \forall f$$

    where b^0 is the true unknown value.

$$f_h(\cdot), \qquad h = 1, \dots, H, \qquad s.t. \ w_{ih} = f_h(x_i)$$

The IV Estimator

- Consider a matrix

$$W_{(n\times H)} = [w^1 \quad w^2 \quad \cdots \quad w^H]$$

If I assume:

$$E(u_i|x_i) = 0 \iff E(y_i - x_i'b|x_i) = 0$$
$$\iff E[f(x_i)(y_i - x_i'b)] = 0, \qquad \forall f$$

If I define $w_{ih} = f_h(x_i)$.

- Can we find $\hat{b}_w$ such that

$$\frac{1}{n}\sum_{i=1}^{n}[w_{ih}(y_i - x_i'b)] = 0, \qquad \forall h \iff W'(y - X\hat{b}_w) = 0$$
$$\iff \underbrace{W'X}_{H\times K}\hat{b}_w = W'y$$

Need $H \geq K$, because otherwise there will be more parameters than equations. If the number of instruments is the same as the regressors, then the matrix $W'X$ is invertible. However, if there are more instruments than regressors, then we need to use "pseudo-inverse".

  - Consider a ***left pseudo-inverse*** of $W'X$, call it Π_n:

$$\underbrace{\Pi_n}_{K\times H}(W'X) = I_K$$

- *Definition.* Under the maintained assumptions

$$Rank(W) = H$$
$$Rank(W'X) = K$$

Define:

$$\hat{b}_w = \Pi_n W'y$$

where Π_n is as defined above.

  - Example of matrix Π_n. Let Ω be a positive definite matrix of size H

$$\Pi_n = \underbrace{\left(\underbrace{X'W}_{K\times H}\underbrace{\Omega}_{H\times H}\underbrace{W'X}_{H\times K}\right)}_{K\times K}^{-1}\underbrace{X'W}_{K\times H}\underbrace{\Omega}_{H\times H}$$

The inverse exists because whenever a full-rank matrix is multiplied by another full-rank matrix, the product is still full rank.

Note that this is a *class* of pseudo-inverse, because a different Ω will produce a different Π_n.

Special case where $H = K$:

$$\Pi_n = (W'X)^{-1}\Omega^{-1}(X'W)^{-1}(X'W)\Omega = (W'X)^{-1}$$
$$\Rightarrow \hat{b}_w = (W'X)^{-1}W'y$$

If we choose $W = X$, then we'll get the OLS estimate.

In general, $H > K$, so there is no unique $\hat{b}_w \rightarrow$ we can talk about optimal choice of Ω.

❖ Is $\hat{b}_w$ consistent? Recall the model $y = Xb^0 + u$.

$$\begin{aligned} \hat{b}_w = \Pi_n W'y \;\; \Rightarrow \;\; \hat{b}_w &= \Pi_n W'[Xb^0 + u] \\ &= \Pi_n W'Xb^0 + \Pi_n W'u \\ &= b^0 + \underbrace{\Pi_n W'u}_{\xrightarrow{?} 0} \end{aligned}$$

- **Assumption 1.** $n\Pi_n \xrightarrow{p} \Pi$, where Π is a fixed full-rank matrix.
  - Why is this assumption reasonable?
    - Case where $H = K$.

      $$\Pi_n = (W'X)^{-1} \;\; \Rightarrow \;\; n\Pi_n \left[\frac{W'X}{n}\right]^{-1}$$

      Remember that W' is $(H \times n)$ and X is $(n \times K)$. By LLN:

      $$\frac{W'X}{n} = \frac{1}{n}\sum_{i=1}^{n} w_i x_i' \xrightarrow{p} E[w_i x_i']$$

      Here, Assumption 1 is equivalent to the LLN for $(W'X)$ (or $w_i x_i'$).

    - Case where $H > K$.

      $$n\Pi_n = \left(\frac{X'W}{n}\Omega\frac{W'X}{n}\right)^{-1}\frac{X'W}{n}\Omega$$

      where

      $$\frac{X'W}{n} \xrightarrow{p} E(x_i w_i'), \qquad \frac{W'X}{n} \xrightarrow{p} E(w_i x_i').$$

      So, same here, we need LLN for $(w_i x_i')$

  - In sum, the assumption is used to

    $$\begin{aligned} \hat{b}_w &= b^0 + \Pi_n W'u \\ &= b^0 + (n\Pi_n)\left(\frac{W'u}{n}\right) \end{aligned}$$

- **Assumption 2.** The WLLN for $(w_i u_i)$

  $$\frac{1}{n}W'u \xrightarrow{p} E(w_i u_i)$$

- **Assumption 3.** $E(w_i u_i) = 0$. That is, w_i's are valid instruments. Non-correlation between w_i and u_i

Therefore, $\hat{b}_w \xrightarrow{p} b^0$.

❖ **Theorem.** Any IV estimator $\hat{b}_w = \Pi_n W'y$ such that
  - $\Pi_n W'X = I_K$
  - $n\Pi_n \xrightarrow{p} \Pi$
  - W are valid instruments

- WLLN for $(w_i u_i)_i$

are satisfied is weakly consistent, i.e. $\hat{b}_w \xrightarrow{p} b^0$.

- Consistent estimator of σ^2.

$$\sigma^2 = E(u_i^2) = \text{plim}\left[\frac{1}{n}\sum_{i=1}^{n} u_i^2\right]$$

  - The problem here is that u_i's are not observed.
  - $\hat{u}_i$ are residuals and observed.

- **Theorem.** Given WLLN for (u_i^2), $(x_i x_i')$, and $(x_i u_i)$. If $\hat{b}$ is a consistent estimator of b^0, then

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \hat{u}_i^2 \xrightarrow{p} \sigma^2$$

where $\hat{u}_i = y_i - x_i'\hat{b}$.

  - *Proof.*

$$\begin{aligned}
\hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n} \hat{u}_i^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} \left(y_i - x_i'\hat{b}\right)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} \left[x_i' b^0 + u_i - x_i'\hat{b}\right]^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} u_i^2 + \frac{1}{n}\sum_{i=1}^{n} x_i'\left(b^0 - \hat{b}\right)^2 + \frac{2}{n}\sum_{i=1}^{n}\left[u_i x_i'\left(b^0 - \hat{b}\right)\right] \\
&\xrightarrow{p} \sigma^2
\end{aligned}$$

  - Remark. The above estimator $\hat{\sigma}^2$ is consistent, but usually biased in small/finite samples.
  - For OLS,

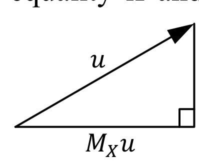
$$\frac{1}{n}\sum_{i=1}^{n} \hat{u}_i^2 \leq \frac{1}{n}\sum_{i=1}^{n} u_i^2 \quad \Rightarrow \quad E\hat{\sigma}^2 < \sigma^2$$

    - *Proof.*

$$\sum_{i=1}^{n} \hat{u}_i^2 = \|\hat{u}\|^2 = \|M_X y\|^2 = \|M_X u\|^2 \leq \|u\|^2$$

The last inequality comes from the orthogonal projection. We have equality if and only if $M_X u = u$.

$$E\|M_X u\|^2 < E\|u\|^2 \quad \Leftrightarrow \quad nE\hat{\sigma}^2 < n\sigma^2.$$



- What is the difference between the two?

$$E[\|M_X u\|^2 | X] = E[u' M_X' M_X u | X]$$

$$
\begin{aligned}
&= E[u'M_X u|X] \\
&= E[\mathrm{tr}(u'M_X u)\,|X] \\
&= \mathrm{tr}(E[M_X u'u|X]) \\
&= \mathrm{tr}\left(M_X \underbrace{E(u'u|X)}_{\sigma^2}\right) \\
&= (n-K)\sigma^2
\end{aligned}
$$

For OLS estimator under the assumption of spherical variance (i.e. $E(u'u|X) = \sigma^2$),

$$E(\|\hat{u}\|^2|X) = (n-K)\sigma^2 \;\; \Rightarrow \;\; \frac{1}{n-K}\|\hat{u}\|^2 = \frac{1}{n-K}\sum_{i=1}^{n}\hat{u}_i^2$$

So the unbiased estimator $\frac{1}{n}\sum_{i=1}^{n}\hat{u}_i^2$ underestimates σ^2.

When $n \gg (n-K)$ and n become closer and closer to each other.

Spherical and Non-Spherical Variance of the Error Term

- Recall that $Var(u) = \sigma^2 I$. Today we want to consider the "non-spherical" case.
  - Can we standardize the multivariate u?

$$u_{n,1} \sim \mathcal{N}(0, \Omega)$$
$$\Omega = \Omega^{1/2}\Omega^{1/2'}$$

- Any symmetric matrix Ω can be decomposed as:

$$\Omega = P\Lambda P'$$

where $PP' = I$ (i.e. P is orthogonal matrix) and Λ is diagonal.

  - Note that Ω is symmetric but also positive definite. So all its eigenvalues are strictly positive; that is, $\Lambda^{1/2}$ is well defined.

$$\begin{aligned} P\Lambda P' &= P\Lambda^{1/2}\Lambda^{1/2}P' \\ &= \left(P\Lambda^{1/2}\right)\left(\Lambda^{1/2}P'\right) \\ &= \Omega^{1/2}\Omega^{1/2'} \end{aligned}$$

Notation: $\Omega^{-1/2} = \left(\Omega^{1/2}\right)^{-1}$. Then,

$$\begin{aligned} Var\left(\Omega^{-1/2}u\right) &= \Omega^{-1/2}Var(u)\Omega^{-1/2'} \\ &= \Omega^{-1/2}\Omega\Omega^{-1/2'} \\ &= \Omega^{-1/2}\Omega^{1/2}\Omega^{1/2'}\Omega^{-1/2'} \\ &= I \end{aligned}$$

$$\Rightarrow \quad \Omega^{-1/2}u \sim \mathcal{N}(0, I)$$
$$\Rightarrow \quad \left(\Omega^{-1/2}u\right)'\left(\Omega^{-1/2}u\right) = u'\Omega^{-1}u \sim \chi^2(n)$$

  - Remark. The shape of a confidence region (because we're considering a vector)

$$P(u'\Omega^{-1}u \le q_{1-\alpha}) = \alpha$$

where q is the appropriate quantile of $\chi^2(n)$ distribution.

    - In the spherical case: $\Omega = \sigma^2 I$,

$$P\left(\frac{u'u}{\sigma^2} \le q_{1-\alpha}\right) = \alpha$$

$\rightarrow$ shape = sphere centered around 0 with ray $\sigma\sqrt{q_{1-\alpha}}$

    - More generally,

$$\Omega = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & 0 & \cdots & \sigma_n^2 \end{pmatrix} \Rightarrow P\left(\sum_{i=1}^{n}\frac{u_i^2}{\sigma_i^2} \le q_{1-\alpha}\right) = \alpha$$

$\rightarrow$ shape = ellipse centered around 0.

Asymptotic Probability Distribution

- Reminder:
  - Suppose $z_{i_{H,1}}$ identically distributed. Then, WLLN says
$$\bar{z}_n = \frac{1}{n}\sum_{i=1}^{n} z_i \xrightarrow{p} Ez_i = \mu$$
  - If $Var(z_i) < \infty$ and z_i iid, then
$$Var(\bar{z}_n) = \frac{1}{n}Var(z_i)$$
Suppose we rescale by $\alpha \neq 1/2$,
$$n^\alpha(\bar{z}_n - \mu) \to \begin{cases} 0 & \text{if } \alpha < \frac{1}{2} \\ \infty & \text{if } \alpha > \frac{1}{2} \end{cases}$$

- ***Central Limit Theorem*** (Lindeberg-Levy).
If $z_{i_{H,1}}$ is iid with $Ez_i = \mu$ and $Var(z_i) = \Sigma_{H,H}$, then
$$\sqrt{n}(\bar{z}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \Sigma).$$

  - Remark. Convergence in distribution. $V_n \xrightarrow{d} V$ if and only if we have something like
$$P(V_n \in A) \underset{n\to\infty}{\longrightarrow} P(V \in A)$$
If $\dim(V_n) = 1$, then
$$V_n \xrightarrow{d} V \iff \forall x \text{ where the function is well-defined } \begin{cases} x \to P(V \leq x) \\ F_{V_n}(x) \to F_V(x) \end{cases}$$
or $P(V_n \leq x) \to P(V \leq x)$
Recall that
$$V_n \xrightarrow{p} V \underset{\not\Leftarrow}{\Rightarrow} V_n \xrightarrow{d} V.$$
When V is <u>deterministic</u>,
$$V_n \xrightarrow{p} V \iff V_n \xrightarrow{d} V$$
because in this case the joint distributions of V_n and V are known.
$$\begin{cases} V_n \xrightarrow{d} V & \text{(not deterministic)} \\ Z_n \xrightarrow{d} a & \text{(deterministic)} \end{cases} \Rightarrow \begin{pmatrix} V_n \\ Z_n \end{pmatrix} \xrightarrow{d} \begin{pmatrix} V \\ a \end{pmatrix}$$

<u>But (!!!)</u>
$$\begin{cases} V_n \xrightarrow{d} V & (r.v.) \\ Z_n \xrightarrow{d} a & (r.v.) \end{cases} \begin{matrix} \not\Rightarrow \\ \Leftarrow \end{matrix} \begin{pmatrix} V_n \\ Z_n \end{pmatrix} \xrightarrow{d} \begin{pmatrix} V \\ Z \end{pmatrix}$$

- **Corollary.**
$$\left.\begin{matrix} z_n \xrightarrow{d} \mathcal{N}(0, \Sigma) \\ A_n \xrightarrow{p} A \end{matrix}\right\} \Rightarrow A_n z_n \xrightarrow{d} \mathcal{N}(0, A\Sigma A')$$

- Asymptotic distribution of IV estimation.

$$\hat{b}_W = \Pi_n W'y, \qquad \Pi_n \ s.t.\ \Pi_n W'X = I$$
$$\Rightarrow \ \hat{b}_W = b^0 + \Pi_n W'u, \qquad \because y = Xb^0 + u$$

And

$$\sqrt{n}\Pi_n W'u = \underbrace{n\Pi_n}_{\to\Pi} \underbrace{\frac{W'u}{\sqrt{n}}}_{=\frac{1}{\sqrt{n}}\sum_{i=1}^n w_i u_i}$$

  - Assume: (w_i, y_i, x_i) jointly iid, and w_i are valid instruments
$$Var(w_i u_i) = E[(w_i u_i)(w_i u_i)']$$
$$\Sigma_W = E(u_i^2 w_i w_i')$$
  - **Theorem.**
$$\sqrt{n}\left(\hat{b}_W - b^0\right) \xrightarrow{d} \mathcal{N}(0, \Pi\Sigma_W\Pi')$$
    - Interpretation. For n large enough, the probability distribution of $\hat{b}_W$ can be approximated by $\mathcal{N}\left(b^0, \frac{1}{n}\Pi\Sigma_W\Pi'\right)$.
    - Can assess this approximation by Monte Carlo.

- **Lemma** (in Davidson's Book**).**

$$\min_{\Pi}(\Pi\Sigma\Pi'), \qquad s.t.\ \Pi L = I_K \text{ for some given matrix } L$$

where $\Sigma_{H,H}$ is positive definite, $L_{H,K}$ has rank K. The solution to the above problem is
$$\Pi^* = (L'\Sigma^{-1}L)^{-1}L'\Sigma^{-1}.$$

  - Remark. Π^* is the solution of the minimization if and only if
$$\Pi^* L = I$$
$$\Pi^*\Sigma\Pi^* \ll \Pi\Sigma\Pi, \qquad \forall\Pi : \Pi L = I$$
    - What does "$\ll$" mean? Consider a vector $v = \alpha + \beta$, where α and β are uncorrelated.
$$Var(v) = Var(\alpha) + Var(\beta)$$
We say
$$\begin{aligned} Var(v) \gg Var(\alpha) \ &\Leftrightarrow \ \left(Var(v) - Var(\alpha)\right) \text{ is psd} \\ &\Leftrightarrow \ \forall x : x'\left(Var(v) - Var(\alpha)\right)x \geq 0 \end{aligned}$$
    - In our case,
$$Var\left[\sqrt{n}\left(\hat{b}_W - b^0\right)\right] = \Omega.$$
Suppose there is another estimator b^* such that
$$Var\left[\sqrt{n}(b^* - b^0)\right] = \Omega^*.$$
Then,
$$\begin{aligned} \Omega^* \ll \Omega \ &\Leftrightarrow \ \forall a : a'(\Omega - \Omega^*)a \geq 0 \\ &\Leftrightarrow \ \forall a : a'\Omega a = Var\left[\sqrt{n}a\left(\hat{b}_W - b^0\right)\right] \geq Var\left[\sqrt{n}a(b^* - b^0)\right] = a'\Omega^* a \end{aligned}$$
i.e. b^* is better than $\hat{b}_W$ in terms of variance.

  - *Proof* of the lemma. Suppose $\Pi = \Pi^* + D$. We have
$$\Pi L = \Pi^* L = I \ \Rightarrow \ DL = 0.$$

Then,

$$\begin{aligned}\Pi\Sigma\Pi' &= (\Pi^* + D)\Sigma(\Pi^* + D)' \\ &= \Pi^*\Sigma\Pi^{*\prime} + \Pi^*\Sigma D' + D\Sigma\Pi^{*\prime} + D\Sigma D'\end{aligned}$$

We want to show that

$$\Pi\Sigma\Pi' - \Pi^*\Sigma\Pi^{*\prime} \gg 0 \iff \Pi^*\Sigma D' + D\Sigma\Pi^{*\prime} + D\Sigma D' \gg 0$$

Note that

$$D\Sigma D' \gg 0 \Leftarrow \alpha'(D\Sigma D')\alpha = (D'\alpha)'\Sigma(D'\alpha) \geq 0.$$

It is enough to show that

$$\Pi^*\Sigma D' + D\Sigma\Pi^{*\prime} \gg 0$$

However, this is not easy, so we would instead show that

$$\Pi^*\Sigma D' + D\Sigma\Pi^{*\prime} = 0$$

Asymptotic Variance of IV Estimator

- Recall from last time
  - **Theorem.**
  $$\sqrt{n}\left(\hat{b}_W - b^0\right) \xrightarrow{d} \mathcal{N}(0, \Pi\Sigma_W\Pi')$$
  - **Lemma.**
  $$\min_{\Pi}(\Pi\Sigma\Pi'), \qquad s.t.\ \Pi L = I_K$$
  with Σ positive definite and $rank(L) = K$.

  - *Proof of Lemma.*
  $$\Pi = \Pi^* + D \ \ \Rightarrow \ \ DL = 0$$
  $$\begin{aligned}\Pi\Sigma\Pi' &= (\Pi^* + D)\Sigma(\Pi^* + D)' \\ &= \Pi^*\Sigma\Pi^{*\prime} + D\Sigma\Pi^{*\prime} + \Pi^*\Sigma D' + D\Sigma D'\end{aligned}$$
  To conclude that $\Pi\Sigma\Pi' - \Pi^*\Sigma\Pi^{*\prime} \gg 0$ (because $D\Sigma D' \gg 0$ and $\Pi^*\Sigma\Pi^{*\prime} \gg 0$), it is enough to show that
  $$D\Sigma\Pi^{*\prime} + \Pi^*\Sigma D' = 0$$
  But since $(D\Sigma\Pi^{*\prime})' = \Pi^*\Sigma D$, it is enough to show that
  $$D\Sigma\Pi^{*\prime} = 0 \quad \text{or} \quad \Pi^*\Sigma D' = 0$$
  Idea: define Π^* such that
  $$\Pi^*\Sigma D' = 0.$$
  Note that $DL = L'D' = 0$. Then, for any matrix A, define
  $$\Pi^* := AL'\Sigma^{-1}$$
  such that $\Pi^*\Sigma D' = 0$. We also need to make sure Π^* as defined is a valid candidate, i.e.
  $$\Pi^* L = I_K \ \ \Leftrightarrow \ \ AL'\Sigma L = I_K$$
  where the matrix $L'\Sigma L$ has full rank and thus invertible. Therefore, let
  $$A = (L'\Sigma L)^{-1}.$$
  Therefore, the solution to the minimization problem is
  $$\Pi^* = (L'\Sigma L)^{-1}L'\Sigma^{-1}.$$
  This completes the proof.

- The “best” IV estimator (i.e. the one with the smallest asymptotic variance).
  $$\Pi^* = \{E(x_i w_i')[E(u_i^2 w_i w_i')]^{-1}E(x_i w_i')'\}^{-1}E(x_i w_i')[E(u_i^2 w_i w_i')]^{-1}$$
  Thus,
  $$Var\left(\sqrt{n}\hat{b}_W^*\right) = [E(x_i w_i')\Sigma_W E(w_i x_i')]^{-1}$$
  with $\Sigma_W = E(u_i^2 w_i w_i')$.

  So the optimal IV estimator (for given W):
  $$\hat{b}_W^* = \Pi_n^* W'y$$
  where $\Pi_n^* = \left[(X'W)\hat{\Sigma}^{-1}(W'X)\right]^{-1}X'W\hat{\Sigma}^{-1}$.

  Recall that the difference between Π_n^* and Π^* is that $n\Pi_n^* \xrightarrow{d} \Pi^*$. So to apply LLN,
  $$n\Pi_n^* = \left[\frac{(X'W)}{n}\hat{\Sigma}^{-1}\frac{(W'X)}{n}\right]^{-1}\frac{X'W}{n}\hat{\Sigma}^{-1}.$$

Feasible Estimate of the IV Estimator

- Recall optimal IV for given W:
$$\Pi_n = \left[X'W\hat{\Sigma}^{-1}W'X\right]^{-1}X'W\hat{\Sigma}^{-1} \quad \text{and} \quad \hat{b}_W = \Pi_n W'y$$

- Recall $\hat{\Sigma}$ is a consistent estimator of $\Sigma_W = E(u_i^2 w_i w_i')$
  - Note: it is enough to provide estimate of $a\Sigma_W$, e.g. $a\hat{\Sigma}_W$

(1) Conditional homoscedasticity
- $w_i = f(x_i)$
- Endogeneity: $w_i \neq f(x_i)$ because $\neg(x_i \perp u_i)$
$$\Sigma_W = E[u_i^2 w_i w_i'] = E\left\{\underbrace{E(u_i^2|w_i)}_{\sigma^2} w_i w_i'\right\} = \sigma^2 E[w_i w_i']$$

Take $\hat{\Sigma} = \frac{\sigma^2}{n}W'W$. Then, in the conditional homoscedasticity case
$$\Pi_n = (X'P_W X)^{-1}X'W(W'W)^{-1} \;\Rightarrow\; \hat{b}_W = (X'P_W X)^{-1}X'P_W y$$
where $P_W = W(W'W)^{-1}W'$.

- Remark. This formula is very similar to the Frisch-Waugh, and leads to a 2-step procedure:
  - (i) "Projection": $P_W X$, i.e. OLS X^k onto W
  - (ii) "OLS": y onto $P_W X$

- **Theorem.** 2S-OLS is the optimal IV for given W under conditional homoscedasticity.

- Special case where $w_i = f(x_i)$, i.e. X are exogenous.
$$y = Xb + u$$
$$Var(y|X) = Var(u|X) = \sigma^2 I, \qquad \text{[spherical]}$$

Then, we have
$$Var(\hat{b}_W|X) = (X'P_W X)^{-1}X'P_W Var(y|X) P_W X(X'P_W X)^{-1}$$
$$= \sigma^2(X'P_W X)^{-1}$$

- If we want to compare OLS and IV variances:
$$Var(\hat{b}_W|X) = \sigma^2(X'P_W X)^{-1}$$
$$Var(\hat{b}_{OLS}|X) = \sigma^2(X'X)^{-1}$$

We can conclude that $Var(\hat{b}_W|X)$ is bigger, because whenever we do an orthogonal projection, the length of a vector becomes smaller (norm-wise). So
$$(X'P_W)(P_W X) \ll (X'X) \;\Leftrightarrow\; X'\underbrace{(M_W)}_{\gg 0}X \gg 0$$
$$\Rightarrow\; (X'P_W X)^{-1} \gg (X'X)^{-1}.$$

Therefore, OLS is always as good as IV, and sometimes better.

- Why do we do IV then?
  - When X are endogenous. [Recall: $w_i \neq f(x_i)$ in this case]
  - When X are exogenous and conditional homoscedastic:
$$E(u_i^2|x_i) = \sigma(x_i)$$

So that we capture some information that is left in the σ^2.

❖ **Theorem.** If $\hat{u}_i = y_i - x_i'\hat{b}$, where $\hat{b}$ is a consistent estimator, then

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} \hat{u}_i^2 w_i w_i' \longrightarrow \Sigma_W$$

under the appropriate LLN.

➢ *Proof.*

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} \hat{u}_i^2 w_i w_i' &= \frac{1}{n}\sum_{i=1}^{n}\left[u_i + x_i'\left(b-\hat{b}\right)\right]^2 w_i w_i' \\ &= \frac{1}{n}\sum_{i=1}^{n} u_i^2 w_i w_i' + \frac{1}{n}\sum_{i=1}^{n}\left[x_i'\left(b-\hat{b}\right)\right]^2 w_i w_i' + \frac{2}{n}\sum_{i=1}^{n} u_i \underbrace{x_i'\left(b-\hat{b}\right)}_{\in\mathbb{R}} w_i w_i'\end{aligned}$$

Need LLN for $x_i x_i' w_i w_i'$ and $x_i u_i w_i w_i'$, and $u_i^2 w_i w_i'$.
This is more restrictive because we need moments of order 4.

Asymptotic Variance (cont'd)

- $\Sigma = E[u_i^2 w_i w_i']$ with estimator
$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(\hat{u}_i^2 w_i w_i')$$
  - Remark. Eicker-White estimator
$$\hat{\Sigma} = \frac{1}{n}W'\hat{\Delta}W$$
where
$$\hat{\Delta} = \begin{pmatrix} \hat{u}_1^2 & & & 0 \\ & \hat{u}_2^2 & & \\ & & \ddots & \\ 0 & & & \hat{u}_n^2 \end{pmatrix}$$
Coefficient (k,ℓ) of $\hat{\Sigma}$ is
$$\frac{1}{n}\sum_{i=1}^{n} w_{ik}w_{i\ell}\hat{u}_i^2$$
Note that $\hat{\Delta}$ is not an estimator of Σ.
  - The feasible estimator is then,
$$\hat{b}_W = \left(X'W\hat{\Sigma}^{-1}W'X\right)^{-1}X'W\hat{\Sigma}^{-1}W'y$$
with $\hat{\Sigma} = W'\hat{\Delta}W$.
    - It has the same asymptotic distribution as the infeasible estimator
$$b_W^* = \left[X'W\underbrace{(W'\Omega W)}_{\Sigma}^{-1}W'X\right]^{-1}X'W(W'\Omega W)^{-1}W'y$$
where $\Omega = Var(y|X)$. The variance of b_W^* is
$$\begin{aligned} Var(b_W^*|X) &= Var(Ay|X) \\ &= A\,Var(y|X)A' \\ &= A\Omega A' \\ &= [X'W(W'\Omega W)^{-1}W'X]^{-1} \end{aligned}$$
where $A = [X'W(W'\Omega W)^{-1}W'X]^{-1}X'W(W'\Omega W)^{-1}W'$.

- The optimal instruments (i.e. the best matrix W).
  - **Theorem.** $[X'W(W'\Omega W)^{-1}W'X]^{-1}$ is minimum for $W = \Omega^{-1}X$
    - $\Omega^{-1}X$ is the optimal instrument
    - $b_W^* = (X'\Omega^{-1}X)^{-1}X'\Omega y$. This is the ***generalized least squares*** (***GLS***) estimator
    - $Var(b_W^*|X) = (X'\Omega^{-1}\mathrm{X})^{-1}$

    - Comments:
      - GLS is characterized as the optimal IV estimator when $Var(u|X) = \Omega$ and $W = f(X)$
      - GLS is infeasible (because Ω is unknown)
      - The theorem is also true when Ω is not diagonal.

    - *Proof.* We want to show that, for any W,

$$X'W(W'\Omega W)^{-1}W'X \ll X'\Omega^{-1}X \quad (*)$$

Note that the optimal W is given by

$$W^* = \Omega^{-1}X = \Omega^{-1/2'}\underbrace{\Omega^{-1/2}X}_{\tilde{X}}$$

Write any other W as

$$W = \Omega^{-1/2'}Z$$

Then,

$$\begin{aligned} (*) \quad &\Leftrightarrow \quad X'\Omega^{-1/2'}Z(Z'Z)^{-1}Z'\Omega^{-1/2}X \ll X'\Omega X \\ &\Leftrightarrow \quad \tilde{X}'P_Z\tilde{X} \ll \tilde{X}'\tilde{X} \end{aligned}$$

- Interpretation of GLS:
  - IV estimator with $W = \Omega^{-1}X$,

$$W'(y - Xb_W^*) = 0 \quad \Leftrightarrow \quad \underbrace{X'\Omega^{-1}}_{\substack{\text{oblique} \\ \text{projection}}} (y - Xb_W^*) = 0$$

$$y = Xb + u$$

with $Var(u|X) = \Omega = \Omega^{1/2}\Omega^{1/2'}$

$$\Rightarrow \quad \Omega^{-1/2}y = \Omega^{-1/2}Xb + \underbrace{\Omega^{-1/2}u}_{v}$$

with $Var(v|X) = I$.

Now that we have spherical errors, we know that OLS is optimal.

$$\left(X'\Omega^{-1/2'}\Omega^{-1/2}X\right)^{-1}\left(X'\Omega^{-1/2'}\Omega^{-1/2}y\right) = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y$$

  - This is the GLS formula!
  - GLS is simply OLS on standardized errors.

- Feasible GLS
  - Ω can be
    - diagonal, if iid

$$\Omega = \begin{pmatrix} \sigma^2(x_1) & & & 0 \\ & \sigma^2(x_2) & & \\ & & \ddots & \\ 0 & & & \sigma^2(x_n) \end{pmatrix}$$

and

$$\Omega^{-1/2}y = \Omega^{-1/2}Xb + \Omega^{-1/2}u \quad \Leftrightarrow \quad \begin{pmatrix} \vdots \\ \dfrac{y_i}{\sigma^2(x_i)} \\ \vdots \end{pmatrix} = \begin{pmatrix} & \vdots & \\ \cdots & \dfrac{x_{ik}}{\sigma(x_i)} & \cdots \\ & \vdots & \end{pmatrix} b + \begin{pmatrix} \vdots \\ \dfrac{u_i}{\sigma(x_i)} \\ \vdots \end{pmatrix}$$

Then, GLS is

$$\min_b \underbrace{\left[\sum_{i=1}^{n}\left(\frac{y_i}{\sigma(x_i)} - \frac{x_i'b}{\sigma(x_i)}\right)^2\right]}_{\substack{\text{sum of weighted squares} \\ \text{of residuals}}}$$

This is the WLS (weighted least squares).

  - serial correlation, GLS is usually useless.

$$\Omega^{-1/2}y = \Omega^{-1/2}\begin{pmatrix} \vdots \\ y_t \\ \vdots \end{pmatrix} = \Omega^{-1/2}Xb + \Omega^{-1/2}u$$

    If Ω is not diagonal, then $\tilde{y} \equiv \Omega^{-1/2}y$ contains mixture of different observations of y_t. This does not make sense.

- Feasible GLS or optimal WLS:

$$\hat{b} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y$$

  Here the Eicker-White does not help because we need an estimator of Ω.
  We need some assumptions about the covariance structure of u, e.g.

$$\sigma^2(x_i) = E[u_i^2 | x_i] = z_i'\alpha$$

  where $z_i = f(x_i)$.

  - Summary:
    (1) OLS of y_i on x_i $\rightarrow$ $\hat{b}$ and $\hat{u}_i$
    (2) OLS of $\hat{u}_i^2$ on z_i $\rightarrow$ $\hat{\alpha}$ and $\sigma^2(x_i)$
    (3) $\min_b \sum_{i=1}^{n} \left[\frac{y_i - x_i'b}{\hat{\sigma}(x_i)}\right]^2$ $\rightarrow$ $\hat{\hat{b}}$ which is asymptotically equivalent to b_{GLS}.

Asymptotic Tests

- Wald tests of hypothesis about b:

$$\sqrt{n}(\hat{b}_W - b^0) \xrightarrow{d} \mathcal{N}\left(0, A\,Var(\hat{b}_W)\right)$$

$$H_0 : \underbrace{g(b^0)}_{(p,1)} = 0, \qquad p \le K$$

Here $g(\cdot)$ is (either linear or non-linear) restrictions on the K elements in $\hat{b}_W$.

- Example 1. Production function

$$\ln Q_i = b_1 + b_2 \ln K_i + b_3 \ln L_i + u_i$$

Testing constant returns to scale: $H_0 : b_2 + b_3 - 1 = 0$.

- General linear hypothesis:

$$H_0 : Rb - r = 0$$

where R is (p, K), r is $(p, 1)$, and $rank(R) = p$. So we are testing p linear restrictions on K parameters. The full column rank assumption means that we're not testing the same restriction twice.

  - In Example 1, we have

$$R = (0 \quad 1 \quad 1), \qquad r = 1, \qquad b = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix}$$

    So we're testing one ($p = 1$) restriction on three parameters.

- Example 2. Non-linear restrictions.

$$y_t = a + \theta_0 x_t + \theta_1 x_{t-1} + \theta_2 x_{t-2} + \cdots + u_t$$

  - Introduce the lag-operator L:

$$y_t = a + \sum_{h=0}^{?} (\theta_h L^h) x_t + u_t$$

    where $L^h x_t = x_{t-h}$. Treating the sum as a polynomial, we have

$$(\theta_0 + \theta_1 L + \cdots) = \frac{c_0 + c_1 L}{1 + a_1 L} \;\Rightarrow\; (\theta_0 + \theta_1 L + \cdots)(1 + a_1 L) = c_0 + c_1 L$$

    This means we have a finite number (i.e. 3 in this case) of the restrictions. This allows us to accommodate an infinite number of parameters:

$$\begin{aligned} \theta_0 &= c_0, && \text{[order 0]} \\ \theta_0 a_1 + \theta_1 &= c_1, && \text{[order 1]} \\ \theta_1 a_1 + \theta_2 &= 0, && \text{[order 2]} \\ \theta_2 a_1 + \theta_3 &= 0, && \text{[order 3]} \\ &\vdots \end{aligned}$$

  - 3 free parameters $\theta_0, \theta_1, \theta_2$
  - All the other ones are functions of them:

$$\theta_3 = -\theta_2 a_1 = -\frac{\theta_2^2}{\theta_1}$$

    which is nonlinear.
  - Formulate the hypothesis as

$$H_0 : \theta_3 = \frac{\theta_2^2}{\theta_1} \iff H_0 : \underbrace{\theta_1\theta_3 - \theta_2^2 = 0}_{g(\theta)=0}$$

Use Taylor expansion:

$$g(\theta) \approx g(\theta^0) + \frac{\partial g(\theta^0)}{\partial \theta'}(\theta - \theta^0)$$

where θ^0 is a vector of the true values.

- General case:

$$\begin{aligned} g : \mathbb{R}^K &\to \mathbb{R}^p \\ b &\mapsto g(b) \end{aligned}$$

Then,

$$\underbrace{\frac{\partial g_{(p,1)}}{\partial b'_{(1,K)}}}_{(p,K)} = \begin{bmatrix} & \vdots & \\ \cdots & \dfrac{\partial g_i}{\partial b_k} & \cdots \\ & \vdots & \end{bmatrix}$$

Recall that

$$\left(\frac{\partial g}{\partial b'}\right)' = \frac{\partial g'}{\partial b}$$

We require that

$$rank\left(\frac{\partial g(b^0)}{\partial b'}\right) = p$$

Asymptotic Testing

- We have some estimator $\hat{b}_W$ of the true parameter b^0, and

$$\sqrt{n}\left(\hat{b}_W - b^0\right) \xrightarrow{d} \mathcal{N}\left(0, AVar\left(\hat{b}_W\right)\right)$$

  - $H_0 : g(b^0) = 0$, where $g(\cdot)$ is a vector of size p
    - If g is linear, $g(b^0) = Rb^0$ with $rank(R) = p$
    - If g is nonlinear, we require $rank\left(\frac{\partial g(b^0)}{\partial b'}\right) = p$

- What is the pdf of $g\left(\hat{b}_W\right)$?
  - ***Delta method***.

    Suppose that $\sqrt{n}\left(\hat{b}_W - b^0\right) \xrightarrow{d} \mathcal{N}\left(0, AVar\left(\hat{b}_W\right)\right)$, and that $g \in \mathcal{C}^1$. Then,

$$\sqrt{n}\left(g\left(\hat{b}_W\right) - g(b^0)\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{\partial g(b^0)}{\partial b'} AVar\left(\hat{b}_W\right) \frac{\partial g'(b^0)}{\partial b}\right)$$

    - *Proof.* For simplicity, assume that we're in dimension one, i.e. $p = 1$. We use instead of the Taylor expansion, the mean value theorem.

$$g\left(\hat{b}_W\right) = g(b^0) + \frac{\partial g\left(\tilde{b}\right)}{\partial b'}\left(\hat{b}_W - b^0\right)$$

      where $\tilde{b}$ is between b^0 and $\hat{b}_W$. Rescaling by $\sqrt{n}$, we get

$$\sqrt{n}\left(g\left(\hat{b}_W\right) - g(b^0)\right) = \frac{\partial g\left(\tilde{b}\right)}{\partial b'} \underbrace{\sqrt{n}\left(\hat{b}_W - b^0\right)}_{\xrightarrow{d}\mathcal{N}}$$

      From $\tilde{b}$ is between $\hat{b}_W$ and b^0 and $\hat{b}_W \xrightarrow{p} b^0 \;\Rightarrow\; \tilde{b} \xrightarrow{p} b^0$. By assumption, $\partial g/\partial b'$ is continuous, we have

$$\frac{\partial g\left(\tilde{b}\right)}{\partial b'} \xrightarrow{p} \frac{\partial g(b^0)}{\partial b'}.$$

      Recall that

$$\left.\begin{matrix} X_n \xrightarrow{d} X \\ Y_n \xrightarrow{p} a \end{matrix}\right\} \;\Rightarrow\; \begin{pmatrix} X_n \\ Y_n \end{pmatrix} \xrightarrow{d} \begin{pmatrix} X \\ a \end{pmatrix}.$$

      This completes the proof.

- Under $H_0 : g(b^0) = 0$,

$$\sqrt{n}\, g\left(\hat{b}_W\right) \xrightarrow{d} \mathcal{N}\left(0, AVar\left(g\left(\hat{b}_W\right)\right)\right)$$

$$\Rightarrow\; w_n = \left[\sqrt{n}\, g'\left(\hat{b}_W\right)\right]\left[AVar\left(g\left(\hat{b}_W\right)\right)\right]^{-1}\left[\sqrt{n}\, g\left(\hat{b}_W\right)\right] \xrightarrow{d} \chi^2(p)$$

  - Critical region (i.e. region where H_0 is REJECTED) of the test:

$$C_n = \{w_n > \chi^2_{1-\alpha}(p)\}$$

    the quantile of $\chi^2(p)$ distribution with level $(1-\alpha)$.

- Two properties of asymptotic tests:

- Property 1 (under H_0). If H_0 is true, $\Pr(C_n) \underset{n\to\infty}{\longrightarrow} \alpha$, [test result is true at the asymptotic level, cf. the Monte Carlo exercise of hw2].
- Property 2 (under H_a). If H_0 is not true, $\Pr(C_n) \underset{n\to\infty}{\longrightarrow} 1$, [consistent test].
  - *Proof.*

$$w_n = n \underbrace{\underbrace{g'(\hat{b}_W)}_{\xrightarrow{p} g'(b^0)} \underbrace{\left[AVar\left(g(\hat{b}_W)\right)\right]^{-1}}_{\xrightarrow{p} \text{a p.d. matrix}} g(\hat{b}_W)}_{\to \text{number}>0} \underset{n\to\infty}{\longrightarrow} +\infty$$

From $w_n \underset{n\to\infty}{\longrightarrow} \infty$ we conclude that $\Pr(C_n) = 1$.

Asymptotic Tests (cont'd)

- $H_0 : g(b) = 0$, g is $(p \times 1)$, b is $(K \times 1)$, $p \leq K$
  - $\hat{b}_W$ is the unconstraint estimate – $\hat{b}_W$ does not use the information contained in H_0
  - $g(\hat{b}_W)$ close to 0?
$$w_n = n\, g'(\hat{b}_W) \left[AVar\left(g(\hat{b}_W)\right)\right]^{-1} g(\hat{b}_W)$$
    Under H_0, $w_n \xrightarrow{d} \chi^2(p)$
  - C_n is the critical region,
$$C_n = \{w_n > \chi^2_{1-\alpha}(p)\}$$
    - $P(C_n) \xrightarrow{n} \alpha$ when H_0 is true [correct asymptotic size]
    - $P(C_n) \xrightarrow{n} 1$ when H_0 is not true [consistency]

- So far, we've been focusing on
$$H_0 : g(b) = 0 \quad \text{vs} \quad H_a : g(b) \neq 0$$
  What if we want to test something more challenging?
  - Idea: Consider H_a that depends on n and gets closer to H_0 as n increases
$$H_a : g(b) = \frac{\delta}{\sqrt{n}}, \qquad \delta \in \mathbb{R}^p \setminus \{\mathbf{0}\}$$
    This is a sequence of ***local alternatives***.
  - Note that $g(b)$ is not fixed, but $\sqrt{n}g(b)$ is fixed (assuming $g(b)$ converges at rate $\sqrt{n}$)
$$\sqrt{n}\left(g(\hat{b}_W) - g(b)\right) \xrightarrow{d} \mathcal{N}\left(0, AVar\left(g(\hat{b}_W)\right)\right)$$
$$\Updownarrow$$
$$\sqrt{n}g(\hat{b}_W) - \sqrt{n}g(b) \xrightarrow{d} \mathcal{N}\left(0, AVar\left(g(\hat{b}_W)\right)\right)$$
    Under the sequence of local alternatives: $\sqrt{n}g(b) = \delta$,
$$\sqrt{n}\left(g(\hat{b}_W)\right) \xrightarrow{d} \mathcal{N}\left(\delta, AVar\left(g(\hat{b}_W)\right)\right)$$
$$\Rightarrow \sqrt{n}\left[AVar\left(g(\hat{b}_W)\right)\right]^{-1/2} g(\hat{b}_W) \xrightarrow{d} \mathcal{N}\left(AVar\left(g(\hat{b}_W)\right)^{-1/2} \delta, I\right)$$
$$\Rightarrow \underbrace{ng'(\hat{b}_W) AVar\left(g(\hat{b}_W)\right)^{-1} g(\hat{b}_W)}_{w_n} \xrightarrow{d} \chi^2\left(p, \delta' AVar\left(g(\hat{b}_W)\right)^{-1} \delta\right)$$
    - Recall that if
$$z_{(p\times 1)} \sim \mathcal{N}(\mu, I) \;\Rightarrow\; z'z \sim \chi^2(p, \mu'\mu)$$
      Non-central χ^2 with p degrees of freedom and non-centrality parameter $\mu'\mu$.

- Property 3. Under the sequence of local alternatives,
$$g(b) = \frac{\delta}{\sqrt{n}}, \qquad \delta \in \mathbb{R}^p \setminus \{\mathbf{0}\}$$
  we have
$$w_n \xrightarrow{d} \chi^2\left(p, \delta' AVar\left(g(\hat{b}_W)\right)^{-1} \delta\right).$$

- Asymptotic power of the test under the sequence of local alternatives

$$P(C_n) \xrightarrow{n} P\left(\chi^2\left(p, \delta' AVar\left(g(\hat{b}_W)\right)^{-1} \delta\right)\right) > \chi^2_{1-\alpha}(p) > \alpha$$

  - The larger the non-centrality parameter, $\delta' AVar\left(g(\hat{b}_W)\right)^{-1} \delta$, the more powerful the test is. The parameter is large in two cases:
    - δ is large – but this is not very useful, as we want δ to be close to zero
    - $AVar\left(g(\hat{b}_W)\right)$ is small:

$$AVar\left(g(\hat{b}_W)\right) = \frac{\partial g(b)}{\partial b'} \underbrace{AVar(\hat{b}_W)}_{\text{make small}} \frac{\partial g'(b)}{\partial b}$$

We want to pick the efficient estimator which is associated with the "smallest" asymptotic variance.

- Wald confidence sets:

$$\sqrt{n}\left[g(\hat{b}_W) - g(b)\right] \xrightarrow{d} \mathcal{N}\left(0, AVar\left(g(\hat{b}_W)\right)\right)$$

$$\Rightarrow \quad n\left[g(\hat{b}_W) - g(b)\right]' \widehat{AVar}\left(g(\hat{b}_W)\right)^{-1} \left[g(\hat{b}_W) - g(b)\right] \xrightarrow{d} \chi^2(p)$$

- Confidence set about $g(b)$ with level $(1-\alpha)$ asymptotically:

$$I_n = \left\{h \in \mathbb{R}^p : n\left(g(\hat{b}_W) - h\right)' \widehat{AVar}\left(g(\hat{b}_W)\right)^{-1} \left(g(\hat{b}_W) - h\right) < \chi^2_{1-\alpha}(p)\right\}$$

$$P(g(b) \in I_n) \xrightarrow[n\to\infty]{} 1 - \alpha$$

- Finding feasible asymptotic variance

$$AVar\left(g(\hat{b}_W)\right) = \frac{\partial g(b)}{\partial b'} AVar(\hat{b}_W) \frac{\partial g'(b)}{\partial b}$$

$$\Rightarrow \quad \widehat{AVar}\left(g(\hat{b}_W)\right) = \frac{\partial g(\hat{b}_W)}{\partial b'} \widehat{AVar}(\hat{b}_W) \frac{\partial g'(\hat{b}_W)}{\partial b}$$

where

$$AVar(\hat{b}_W) = [E(x_i W_i') \Sigma^{-1} E(W_i x_i')]^{-1}$$

with

$$\Sigma = E(u_i^2 w_i w_i')$$

Therefore,

$$\Rightarrow \quad \widehat{AVar}(\hat{b}_W) = \left[\left(\frac{1}{n} X'W\right) \hat{\Sigma}^{-1} \left(\frac{1}{n} W'X\right)\right]^{-1}$$

f.i When $X = W$ [OLS]

$$\Sigma = E(u_i^2 x_i x_i') \quad \Rightarrow \quad \hat{\Sigma} = \frac{1}{n} X' \hat{\Delta} X$$

which is the HCC estimator with

$$\hat{\Delta} = \begin{pmatrix} \hat{u}_1^2 & & & 0 \\ & \hat{u}_2^2 & & \\ & & \ddots & \\ 0 & & & \hat{u}_n^2 \end{pmatrix}$$

or in the homoscedastic case

$$\hat{\Sigma} = \frac{\hat{\sigma}^2}{n} X'X$$

- Restricted least squares under a linear hypothesis

$$H_0 : R_{(p\times K)} b = r_{(p\times 1)}, \qquad rank(R) = p$$

Under homoscedasticity,

$$w_n = \frac{\left(R\hat{b} - r\right)[R(X'X)^{-1}R']^{-1}\left(R\hat{b} - r\right)}{\hat{\sigma}^2}$$

Here $\hat{b}$ is the OLS estimator, and $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \hat{u}_i^2$.

  - Consider the constraint optimization problem:

$$\min_b \left[\sum_{i=1}^{n} (y_i - x_i'b)^2\right] = (y - Xb)'(y - Xb)\,, \qquad s.t.\;\; Rb = r$$

Let $\lambda_{(p\times 1)}$ be a vector of Lagrange multipliers.

$$\mathcal{L} = (y - Xb)'(y - Xb) + \lambda'(Rb - r)$$

$$\underbrace{\frac{\partial \mathcal{L}}{\partial b}}_{(K\times 1)} = -2X'\left(y - X\hat{b}_c\right) + R'\lambda = 0$$

$$\Rightarrow\;\; 2X'\left(y - X\hat{b}_c\right) = R'\lambda \;\;\Leftrightarrow\;\; \hat{b}_c = \hat{b} - \frac{1}{2}(X'X)^{-1}R'\lambda$$

So $\hat{b}_c \neq \hat{b}$ as long as $\lambda \neq 0$, i.e. anytime the constraints are binding.

$$\underbrace{R\hat{b}_c}_{r} = R\hat{b} - \frac{1}{2}\underbrace{R(X'X)^{-1}R'}_{\text{invertible}}\lambda \;\;\Rightarrow\;\; \lambda = 2[R(X'X)^{-1}R']^{-1}\left(R\hat{b} - r\right)$$

Finally,

$$\hat{b}_c = \hat{b} - (X'X)^{-1}R[R(X'X)^{-1}R']^{-1}\left(R\hat{b} - r\right)$$

Difference between the adjusted values:

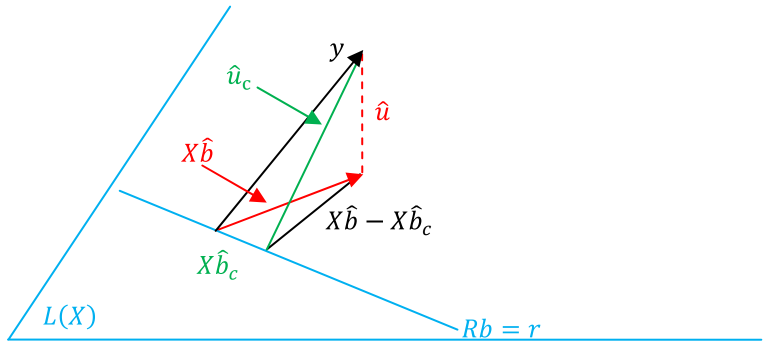
$$X\hat{b} - X\hat{b}_c = X(X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}\left(R\hat{b} - r\right)$$

$$\begin{aligned} \Rightarrow\;\; \left\|X\hat{b} - X\hat{b}_c\right\|^2 &= \left(X\hat{b} - X\hat{b}_c\right)'\left(X\hat{b} - X\hat{b}_c\right) \\ &= \left(R\hat{b} - r\right)'[R(X'X)^{-1}R']^{-1}\left(R\hat{b} - r\right) \end{aligned}$$

This is the numerator of the Wald test statistic w_n, i.e.

$$w_n = \frac{\left\|X\hat{b} - X\hat{b}_c\right\|^2}{\hat{\sigma}^2}$$

  - Geometric interpretation



- Pythagorean theorem:

$$\begin{aligned}\|\hat{u}_c\|^2 &= \|\hat{u}\|^2 + \left\|X\hat{b}-X\hat{b}_c\right\|^2 \\ &= \|\hat{u}\|^2 + \|\hat{u}-\hat{u}_c\|^2 \\ \Rightarrow \quad w_n &= \frac{\left\|X\hat{b}-X\hat{b}_c\right\|^2}{\hat{\sigma}^2} \\ &= \frac{\|\hat{u}_c\|^2 - \|\hat{u}\|^2}{\underbrace{\hat{\sigma}^2}_{=\|\hat{u}\|^2/n}} \\ &= n\cdot\frac{SSR_0 - SSR}{SSR}\end{aligned}$$

- The Gaussian case (small sample).
  Assume $u \sim \mathcal{N}(0,\sigma^2 I_n)$, or $u/\sigma \sim \mathcal{N}(0,I_n)$. Then,

$$y\big|X \sim \mathcal{N}(Xb,\sigma^2 I_n) \;\;\Rightarrow\;\; \hat{b} = (X'X)^{-1}X'y\big|X \sim \mathcal{N}$$

  and $\hat{b}_c|X \sim \mathcal{N}$ (as a linear transformation of $\hat{b}$)

$$\hat{u} = M_X y = M_X u$$

$$\left\|\frac{\hat{u}}{\sigma}\right\|^2 \sim \chi^2(n-K)$$

$$\left\|\frac{\hat{u}_c}{\sigma}\right\|^2 \sim \chi^2\big(n-(K-p)\big)$$

  From the Pythagorean theorem:

$$\underset{\chi^2(n-(K-p))}{\underset{\downarrow}{\frac{\|\hat{u}_c\|^2}{\sigma^2}}} = \underset{\chi^2(n-K)}{\underset{\downarrow}{\frac{\|\hat{u}\|^2}{\sigma^2}}} + \underset{\chi^2(p)}{\underset{\downarrow}{\frac{\|\hat{u}_c-\hat{u}\|^2}{\sigma^2}}}$$

$$\Rightarrow \frac{1}{n}w_n\cdot\frac{n-K}{p} = \frac{[SSR_0-SSR]/p}{SSR/(n-K)} \sim \mathcal{F}(p,n-K)$$

  under the null. Note that

$$\frac{[SSR_0-SSR]}{p} \sim \frac{\chi^2(p)}{p}$$

$$\frac{SSR}{n-K} \sim \frac{\chi^2(n-K)}{n-K}$$

  If H_0 is not true, $\|\hat{u}_c\|^2$ does not mean 0 anymore, because we imposed the incorrect restriction $Rb=r$.

- Fisher test of H_0:

$$C_n^F = \left\{\frac{(SSR_0 - SSR)/p}{SSR/(n-K)} > F_{1-\alpha}(p, n-K)\right\}$$

  and $P(C_n^F) = \alpha$ under H_0
  - ♦ This is exact!! Because I have the exact finite sample distribution of test statistic.

Asymptotic Test (cont'd)

- Recall from last time the Fisher test of $H_0 : Rb_{(K\times 1)} = r_{(p\times 1)}$

$$C_n^F = \left\{\frac{(SSR_0 - SSR)/p}{SSR/(n-K)} > F_{1-\alpha}(p, n-K)\right\}$$

$P(C_n^F) = \alpha$ under H_0.

  - Asymptotically,

$$C_n^F = \left\{w_n > \frac{np}{n-K} \cdot F_{1-\alpha}(p, n-K)\right\}$$

    - We're interested in knowing whether

$$\frac{np}{n-K} \cdot F_{1-\alpha}(p, n-K) \xrightarrow{?} \chi^2_{1-\alpha}(p)$$

If this is the case, then Wald testing and Fisher testing are equivalent asymptotically.

  - Proof of the above convergence.

$$\frac{np}{n-K} F(p, n-K) = \frac{np}{n-K} \cdot \frac{\chi^2(p)/p}{\chi^2(n-K)/(n-K)} = \frac{\chi^2(p)}{n^{-1}\chi^2(n-K)}$$

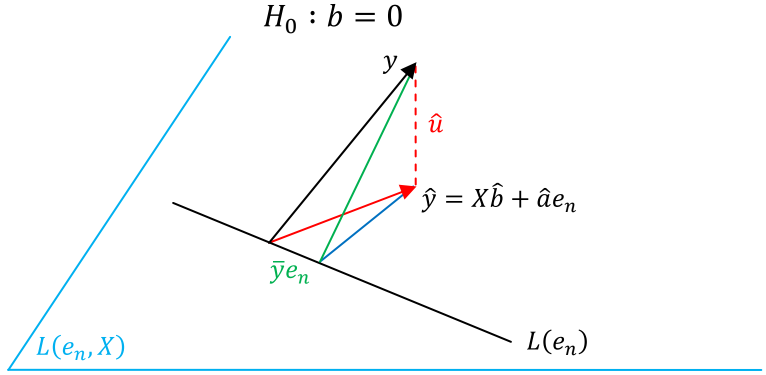
Note that the denominator converges to 1:

$$\frac{1}{n}\chi^2(n-K) = \underbrace{\frac{n-K}{n}}_{\xrightarrow{n\to\infty} 1} \cdot \underbrace{\frac{1}{n-K}\sum_{t=0}^{n-K} z_i^2}_{\xrightarrow{p} E(z_i^2) = Var(z_i^2) = 1}, \qquad z_i \sim \mathcal{N}(0,1)$$

$$\xrightarrow{p} 1$$

- Connection with R^2
  - R^2 only makes sense when there is a constant in the regression model:

$$y = ae_n + Xb + u$$
$$H_0 : b = 0$$



Using the Pythagorean theorem:

$$\underbrace{\frac{1}{n}\|\hat{u}_0\|^2}_{\substack{\frac{1}{n}\sum_{i=0}^{n}(y_i - \bar{y})^2 \\ \text{total variance}}} = \underbrace{\frac{1}{n}\|\hat{u}\|^2}_{\substack{\frac{1}{n}\sum_{i=1}^{n}\hat{u}^2 \\ \text{with } \bar{\hat{u}}=0 \\ \text{residual variance}}} + \underbrace{\frac{1}{n}\|\hat{y} - \bar{y}e_n\|^2}_{\substack{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2 \\ \text{explained variance}}}$$

$$\Rightarrow \ Total\ variance = Residual\ variance + Explained\ variance$$

By definition,

$$R^2 = \frac{Explained\ variance}{Total\ variance} = \frac{\|\hat{y} - \hat{y}e_n\|^2}{\|\hat{u}_0\|^2}$$

Recall,

$$w_n = n\frac{n^{-1}SSR_0 - n^{-1}SSR}{n^{-1}SSR} = \frac{nR^2}{1 - R^2}$$

Critical region (asymptotically):

$$C_n = \left\{\frac{nR^2}{1 - R^2} > \chi^2_{1-\alpha}(K)\right\} \quad \text{or} \quad C_n^* = \{nR^2 > \chi^2_{1-\alpha}(K)\}$$

Sufficient to show that

$$\frac{1}{1 - R^2} = \frac{\|\hat{u}_0\|^2}{\|\hat{u}\|^2} \xrightarrow{p} 1$$

This is true under H_0.

Testing Conditional Homoscedasticity

- OLS: $\hat{b} = (X'X)^{-1}X'y$ with

$$Var(\hat{b}|X) = (X'X)^{-1}X'Var(y|X)X(X'X)^{-1}$$
$$\frac{1}{n}AVar(\hat{b}) = [E(x_i x_i')]^{-1}E(x_i x_i' u_i^2)[E(x_i x_i')]^{-1}$$

can be consistently estimated by HCC

$$HCC = \left[\sum_{i=1}^{n} x_i x_i'\right]^{-1}\left[\sum_{i=1}^{n} x_i x_i' \hat{u}_i^2\right]\left[\sum_{i=1}^{n} x_i x_i'\right]^{-1}$$

  - Test of $H_0 : Var(u_i|x_i) = \sigma^2$ for any i, i.e. H_0 : conditional homoscedasticity
  - Idea is to compare

$$\frac{1}{n}\sum_{i=1}^{n} x_i x_i' \hat{u}_i^2 \quad \text{and} \quad \hat{\sigma}^2 \frac{1}{n}\sum_{i=1}^{n} x_i x_i'$$

- White (1980)

$$\frac{1}{n}\sum_{i=1}^{n} x_i x_i'(\hat{u}_i^2 - \hat{\sigma}^2) \xrightarrow{p} 0 \text{ ?}$$

Here $x_i x_i'$ is a (K, K) matrix with $K(K+1)/2$ different terms

  - Define ψ_i that contains all the different terms of $x_i x_i'$.

$$c_n \equiv \frac{1}{n}\sum_{i=1}^{n} \psi_i(\hat{u}_i^2 - \hat{\sigma}^2) \xrightarrow{p} 0 \text{ ?}$$

Testing for Conditional Homoscedasticity

- Recall from last time:

$$c_n = \frac{1}{n}\sum_{i=1}^{n} \psi_i(\hat{u}_i^2 - \hat{\sigma}^2) \xrightarrow{p} 0$$

  where ψ_i is a vector of different terms in $x_i x_i'$.

$$\sqrt{n} c_n \xrightarrow{d} \mathcal{N}(0, B)$$

  Define

$$\xi_n = n c_n' \hat{B}^{-1} c_n \xrightarrow{d} \chi^2(m)$$

  where m is the number of non-constant terms in ψ_i.

- In practice, we perform an auxiliary regression:

$$\hat{u}_i^2 = \alpha + \psi_i'\gamma + \epsilon_i$$

  - The assumption we're interested in is

$$H_0 : \gamma = 0$$

    That is, if there is homoscedasticity, then the regressors should not be able to explain much of the residuals.
    - Constrained estimator: $\hat{\alpha}_c = \hat{\sigma}^2$
    - Unconstrained OLS estimators: $\hat{\alpha} + \psi_i'\hat{\gamma}$ and the associated test statistic $\xi_n = nR^2$
    - Reject homoscedasticity if and only if $nR^2 > \chi^2_{1-\alpha}(m)$

Dynamic Regression Model

- General framework
  - Need for ergodic stationarity
  - Dynamic regression model:
$$y_t = x_t'b + u_t, \qquad Eu_t = 0$$
    - x_t is still called the explanatory variable, but there are two kinds
      - Lagged values of y_t: $y_{t-1}, y_{t-2}, \dots, y_{t-p}$
      - Other variables: $\eta_t, \eta_{t-1}, \dots, \eta_{t-q}$, where η_t is $K \times 1$

      So the regress model is
$$y_t = \underbrace{\sum_{j=1}^{p} \alpha_j y_{t-j} + \sum_{i=0}^{q} \eta_{t-i}'\gamma_i}_{x_t'b} + u_t$$
      So here,
$$b^0 = [Var(x_t)]^{-1} Cov(x_t, y_t)$$
      where $Cov(x_t, y_t)$ contain things like
$$Cov(\eta_{t-h}, \eta_{t-\ell}), Cov(y_{t-i}, y_{t-j}), \dots$$

- *Definition.* A stochastic process (z_t) is (***strictly***) ***stationary*** if for all r and all t, the joint probability distribution of $(z_t, z_{t+h_1}, z_{t+h_2}, \dots, z_{t+h_r})$ depends on $h_1, h_2, \dots, h_r$ but not on t.
- *Definition.* A process is ***weakly stationary*** (or ***covariance stationary***) when $E(z_t)$ and $Cov(z_t, z_{t+h})$ do not depend on h.
  - Note.
    - (z_t) is iid $\Rightarrow$ (z_t) is a stationary process
    - (z_t) is stationary $\Rightarrow$ (z_t) is identically distributed with some serial dependence

  - Stationarity is not sufficient to get LLN
    - Example. (x_t) iid, and z is independent of x_t
$$y_t = x_t + z$$
      Here (y_t) is stationary:
$$Cov(y_t, y_{t+h}) = Cov(x_t + z, x_{t+h} + z) = Var(z)$$
      This implies that $(x_t + z, x_{t+h_1} + z, x_{t+h_2} + z, \dots, x_{t+h_r} + z)$ has the same probability distribution as any $(x_{\bar{t}} + z, x_{\bar{t}+h_1} + z, x_{\bar{t}+h_2} + z, \dots, x_{\bar{t}+h_r} + z)$.
$$\frac{1}{T}\sum_{t=1}^{T} y_t \overset{?}{\longrightarrow} E(y_t) = E(x_t) + E(z)$$
$$= \frac{1}{T}\sum_{t=1}^{T}(x_t + z) = \frac{1}{T}\sum_{t=1}^{T} x_t + z \overset{p}{\longrightarrow} E(x_t) + z \neq E(y_t)$$
      Unless $z = E(z)$, i.e. z is a constant. But this is not true in general.
    - To avoid this situation, we would need to assume ergodicity.

  - ***Ergodicity*** (informal definition): A random event involving every member of the sequence has either probability 0 or 1.

- Example (cont'd with the previous). If $P(z < a)$ is either 0 or 1, then z is a constant (i.e. z is deterministic). So the counter-example does not work any more.

❖ **Ergodic Theorem.** If (z_t) is stationary, ergodic, and integrable, then

$$\frac{1}{T}\sum_{t=0}^{T} z_t \xrightarrow{p} E(z_t).$$

➢ Hayashi (page 101): "A stationary process is ergodic if it is asymptotically independent"

$$\lim_{n\to\infty} Cov\left(f\left(z_{t+h_1}, z_{t+h_2}, \dots, z_{t+h_r}\right), g\left(z_{t+n+h_1}, z_{t+n+h_2}, \dots, z_{t+n+h_r}\right)\right) = 0, \qquad \forall f, g$$

❖ **Theorem.** Let (z_t) be stationary, ergodic with finite variance process.

$$\frac{1}{T}\sum_{t=1}^{T} Cov(z_1, z_t) \xrightarrow{T\to\infty} 0$$

➢ The converse (non-correlation $\Rightarrow$ asymptotic independence) is true only for the Gaussian processes.

➢ Note 1. If we know that

$$Cov(z_1, z_t) \xrightarrow{t\to\infty} 0 \;\Rightarrow\; \frac{1}{T}\sum_{t=1}^{T} Cov(z_1, z_t) \xrightarrow{T\to\infty} 0$$

But the converse is not true, since

$$\frac{1}{T}\sum_{t=1}^{T} Cov(z_1, z_t) = Cov\left(z_1, \frac{1}{T}\sum_{t=1}^{T} z_t\right)$$

➢ Note 2. With $y_t = x_t + z$, $Cov(y_1, y_t) = Var(z)$

$$\frac{1}{T}\sum_{t=1}^{T} Cov(y_1, y_t) \xrightarrow{?} 0$$

❖ Need for Martingale Difference Sequence (MDS)

➢ Example. OLS: $y_t = x_t' b + u_t$ with

$$\hat{b} = (X'X)^{-1}X'y = b^0 + (X'X)^{-1}X'u$$

$$\Rightarrow \sqrt{T}\left(\hat{b} - b^0\right) = \underbrace{\left[\frac{X'X}{T}\right]^{-1}}_{\to E\left(x_i x_i'\right)} \cdot \underbrace{\frac{X'u}{\sqrt{T}}}_{\substack{\frac{1}{T}\sum_{t=1}^{T} x_t u_t \\ CLT}}$$

Think about $(\mathcal{F}_t)$ filtration, i.e. an increasing sequence of σ-fields

$$\mathcal{F}_t \subset \mathcal{F}_{t+1}$$

- Interpretation: $\mathcal{F}_t$ contains everything we know at time t, i.e. z_τ where $\tau \le t$. In the dynamic regression model, $y_t = x_t' b + u_t$,

$$\mathcal{F}_{t-1} = \sigma\,\underbrace{\left(y_\tau, \tau < t; x_s, s \le t\right)}_{\text{predetermined variables}}$$

This is the smallest σ-field containing all the predetermined variables.

$$E[u_t|\mathcal{F}_{t-1}] = 0.$$

- *Definition.* (z_t) is ***$\mathcal{F}_t$-adapted*** if $z_t \in \mathcal{F}_t$.
  - We say that $(z_t, \mathcal{F}_t)$ is an adapted sequence.

- *Definition.* M_t is a ***martingale*** with respect to $\mathcal{F}_t$ if M_t is $\mathcal{F}_t$-adapted, integrable, and $E(M_t|\mathcal{F}_{t-1}) = M_{t-1}$.

- *Definition.* ϵ_t is a ***martingale difference sequence*** (***MDS***) if ϵ_t is $\mathcal{F}_t$-adapted, integrable, and $E(\epsilon_t|\mathcal{F}_{t-1}) = 0$.

Time Series (cont'd)

- A word on m.d.s.:
  - M_t is martingale with respect to $\mathcal{F}_t \;\rightarrow\; E(M_t|\mathcal{F}_{t-1}) = M_{t-1}$
    - $\epsilon_t = M_t - M_{t-1}$
    - $E(\epsilon_t|\mathcal{F}_{t-1}) = E(M_t|\mathcal{F}_{t-1}) - E(M_{t-1}|\mathcal{F}_{t-1}) = M_{t-1} - M_{t-1} = 0$

- **Theorem.** If $(\epsilon_t, \mathcal{F}_t)$ is m.d.s. and $(x_t, \mathcal{F}_t)$ is adapted, then
  (i) $Cov(\epsilon_t, x_{t-1}) = 0$
  (ii) $(\epsilon_t x_{t-1})$ is a m.d.s. with respect to $\mathcal{F}_t$

  - *Proof.* Statement (i):
$$\begin{aligned} Cov(\epsilon_t, x_{t-1}) &= E(\epsilon_t x_{t-1}) - \underbrace{E\epsilon_t}_{=0} E x_{t-1} \\ &= E[E(\epsilon_t x_{t-1}|\mathcal{F}_{t-1})] \\ &= E[x_{t-1} E(\epsilon_t|\mathcal{F}_{t-1})] \\ &= 0 \end{aligned}$$
    Statement (ii) can be proved similarly.

- In cross-section, we assume that ϵ_t is serially independent
$$\begin{aligned} &\Rightarrow \;\; \epsilon_t, \epsilon_{t-1} \text{ are independent} \\ &\Leftrightarrow \;\; Cov\big(f(\epsilon_t), g(\epsilon_{t-1})\big) = 0, \qquad \forall f, g \end{aligned}$$
  - ϵ_t is mds (with respect to "natural filtration" $\mathcal{F}_t = \{\epsilon_\tau : t \geq \tau\}$)
$$\Rightarrow \;\; E(\epsilon_t|\mathcal{F}_{t-1}) = 0 \text{ and } Cov\big(\epsilon_t, g(\epsilon_\tau)\big) = 0, \qquad \forall g, \;\; \forall \tau < t$$
  - ϵ_t is serially uncorrelated if and only if
$$Cov(\epsilon_t, \epsilon_\tau) = 0, \qquad \forall \tau < t$$
  - Serial independence is stronger than mds (serial uncorrelation with any function of the past), which is in turn stronger than serial uncorrelation (with the past)
    - mds gives use CLT with serial dependence
    - serial uncorrelation gives WLLN

- **Theorem (WLLN).** If (ϵ_t) is a stationary mds, then
$$\frac{1}{T}\sum_{t=1}^{T} \epsilon_t \xrightarrow{p} 0$$
  - Note:
    - (ϵ_t) is mds $\nRightarrow$ $f(\epsilon_t)$ is mds
    - (ϵ_t) is stationary and ergodic $\Rightarrow$ $\forall f$, $f(\epsilon_t)$ is stationary and ergodic

- **Theorem (CLT).** If (ϵ_t) is squared integrable, stationary mds such that $\frac{1}{T}\sum_{t=1}^{T} \epsilon_t^2 \xrightarrow{p} \sigma^2$, then
$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T} \epsilon_t \xrightarrow{d} \mathcal{N}(0, \sigma^2)$$

General Method of Moments (GMM)

- GMM Orthogonality Condition (cf. Hansen (1982) *Econometrica*)
  - General idea: Estimation is based on
    - Observation of a sequence (z_t) which is stationary and ergodic
    - Structural knowledge about $f(z_t, \theta)$ (where f is known but θ is unknown) such that the true unknown value of θ, say θ^0, is characterized by $f(z_t, \theta^0)$ a mds with respect to $(\mathcal{F}_t)$

  - Two different cases

    (1) z_t is iid, i.e. $E[f(z_t, \theta^0)] = 0 \;\rightarrow\;$ unconditional moment restriction (UMR)

    (2) $f(z_t, \theta) = w_t \cdot u_t(\theta)$, where
    - u_t is the error term with $E[u_t(\theta^0|\mathcal{F}_{t-1})] = 0$. Here $\mathcal{F}_{t-1}$ is the predetermined information. $\rightarrow$ conditional moment restriction (CMR)
    - $w_t \in \mathcal{F}_{t-1}$. Note that
    $$\left.\begin{array}{r} E[u_t(\theta^0|\mathcal{F}_{t-1})] = 0 \\ w_t \in \mathcal{F}_{t-1} \end{array}\right\} \rightarrow \text{ UMR}$$
    But this is not the only way to come up with a UMR, e.g. use $g(w_t)$ where g is any function will also work.
    - z_t : all the variables entering into $u_t(\theta)$ and w_t
    $$\Rightarrow \;\; E(f(z_t, \theta^0)|\mathcal{F}_{t-1}) = 0$$

GMM (cont'd)

- GMM orthogonality conditions
  - Unconditional Moment Restriction: $E[f(z_t,\theta^0)]=0$, where f is known and z_t is iid
  - Conditional Moment Restriction: $E[u_t(\theta^0)|\mathcal{F}_{t-1}]=0$
    - Pick $w_{t-1}\in\mathcal{F}_{t-1}$. Then we have the UMR:
$$E\left[\underbrace{w_{t-1}u_t(\theta^0)}_{f(z_t,\theta^0)}\right]=0$$
    - In general, any function $g(w_t)$ will work. So from one CMR we can potentially derive an infinite number of UMR's.

- Example 1. Dynamic Regression Model.
$$y_t=x_t'b+u_t,\qquad u_t(\theta)=y_t-x_t'b$$
$$w_{t-1}\to\begin{cases}x_t & \text{where there is no simultaneity issues}\\ \eta_t & \text{where }\begin{cases}\text{simultaneity issues}\\ \text{conditional heteroscedasticity}\end{cases}\end{cases}$$
Predetermined variables (i.e. variables that belong to $\mathcal{F}_{t-1}$)
  - $y_\tau,\ \tau<t$
  - $\eta_\tau,\ \tau\le t$ if exogenous

- Example 2. Euler equations
$$\max_{C_{t+h},\theta}\sum_{h=1}^{\infty}\beta^h E(U(C_{t+h},\theta)|\mathcal{F}_t)$$
  - Constraints:
$$W_{t+h}=(W_{t+h-1}-C_{t+h-1})R_{t+h}$$
where R_{t+h} is the returns received between period $(t+h-1)$ and $(t+h)$.
  - Differentiate wrt C_{t+h} to get FOC:
$$-U'(C_{t+h-1})+\beta E[U'(C_{t+h})R_{t+h}|\mathcal{F}_{t+h-1}]=0$$
$$\Leftrightarrow\ E\left[\beta\frac{U'(C_{t+h})}{U'(C_{t+h-1})}\cdot R_{t+h}-1\middle|\mathcal{F}_{t+h-1}\right]=0$$

- More examples in Hayashi, Chapter 3.1, 3.2, on simultaneity issues and relevant instruments

- Identification.
$$E[f(z_t,\theta^0)]=0$$
  - *Case 1.* $f(z_t,\theta)=w_t(y_t-x_t'\theta)$.
$$E(w_ty_t)=E\left(\underbrace{w_t}_{H\times1}\ \underbrace{x_t'}_{1\times p}\right)\underbrace{\theta}_{p\times1}$$
Here, $H\ge p$. We need w_tx_t' to have full column rank, i.e. rank p. Hence, we ensure that
$$E\big(f(z_t,\theta^0)\big)=0\ \Leftrightarrow\ \theta=\theta^0$$
  - *Case 2.* Non-linear regression model
$$y_t=h(x_t,\theta)+u_t$$

$$f(z_t,\theta) = w_t\big(y_t - h(x_t,\theta)\big)$$

Locally, we can re-interpret the non-linear regression model as a linear one.

- Rank condition:

$$rank\left(E\left[w_t \frac{\partial h(x_t,\theta^0)}{\partial \theta'}\right]\right) = p$$

- *Case 3*. General Case: $E\big(f(z_t,\theta^0)\big) = 0$
  - Identification assumptions:

    (1) Rank condition:

    $$rank\left(E\left[\frac{\partial f(z_t,\theta^0)}{\partial \theta'}\right]\right) = p$$

    where f is a functional vector of size $H \geq p$.

    (2) $E\big(f(z_t,\theta^0)\big) = 0 \;\Leftrightarrow\; \theta = \theta^0$.

- Note. Order condition (necessary but not sufficient condition for identification)
  - p is the number of parameters and H is the number of moment conditions
    - $p = H \;\rightarrow$ ***just-identified case***
    - $p < H \;\rightarrow$ ***over-identified case***
    - $p > H \;\rightarrow$ ***under-identified case***
  - From the identification point of view, more condition is better to hope that the rank condition is satisfied.

❖ Assumption. $Var\big(f(z_t,\theta^0)\big)$ is non-singular.

- Example. $f(z_t,\theta^0) = w_t u_t(\theta^0)$

$$\begin{aligned} Var\big(f(z_t,\theta^0)\big) &= E[f(z_t,\theta^0)f'(z_t,\theta^0)] \\ &= E[w_t w_t' u_t^2(\theta^0)] \\ &= E\left[w_t w_t' \underbrace{E(u_t^2(\theta^0)|\mathcal{F}_{t-1})}_{\sigma_{t-1}^2(\theta^0)}\right] \\ &= E[w_t w_t' \sigma_{t-1}^2(\theta^0)] \end{aligned}$$

To check that it is non-singular, we compute

$$\alpha' Var\big(f(z_t,\theta^0)\big)\alpha = E[(\alpha' w_t)^2 \sigma_{t-1}^2(\theta^0)] = 0$$
$$\Rightarrow \; (\alpha' w_t)^2 \sigma_{t-1}^2(\theta^0) = 0, \qquad a.s.$$

- Assumption. $P(\sigma_{t-1}^2(\theta^0) = 0) = 0$ and $E[w_t w_t']$ non-singular (or no redundant IV in w_t)

Consistent GMM Estimation

- *Definition.* $E[f(z_t,\theta)] = 0$ where f is $H \times 1$ and θ is $p \times 1$
  - *Case 1.* Just-identified ($H = p$).

$$\hat{\theta} \text{ is the solution of } \left\{\frac{1}{T}\sum_{t=1}^{T} f(z_t,\theta) = 0\right\}$$

    - p equations for p unknowns
    - Can "hope" to find such $\hat{\theta}$

  - *Case 2.* Over-identified ($H > p$). We solve an approximation problem.

$$\min_{\theta}\left\{\bar{f}_T(\theta)' W_T \bar{f}_T(\theta)\right\}$$

  where
    - $\bar{f}_T(\theta) = \frac{1}{T}\sum_{t=1}^{T} f(z_t,\theta)$
    - W_T is a $H \times H$ positive definite matrix which is called weighting matrix.
      - We get $\hat{\theta}_T(W_T)$ for each matrix W_T, i.e. a different GMM estimate. For notational simplicity, we drop the argument and write only $\hat{\theta}_T$
    - In the over-identified case, there is more "freedom" in the choice of W_T. In contrast, in the just-identified case, whatever W_T you pick, the solution of $\hat{\theta}$ is going to be the same.

  FOC

$$\underbrace{\frac{\partial \bar{f}_T'(\hat{\theta}_T)}{\partial \theta}}_{p\times H} \underbrace{W_T}_{H\times H} \underbrace{\bar{f}_T(\hat{\theta}_T)}_{H\times 1} = 0$$

  Redefine

$$\frac{\partial \bar{f}_T'(\hat{\theta}_T)}{\partial \theta} W_T \coloneqq A_T, \qquad \text{[selection matrix]}$$

  Then,

$$A_T \bar{f}_T(\hat{\theta}_T) = 0$$

    - We started with H moment conditions and we selected p linear combinations of them.
    - The p rows of A_T are in the space spanned by the p vectors $\partial \bar{f}_T(\hat{\theta}_T)/\partial \theta_j$, $j = 1, \dots, p$.

- Consistency. $Q_T(\theta) \coloneqq \bar{f}_T'(\theta) W_T \bar{f}_T(\theta)$

$$\hat{\theta}_T = \arg\min\left\{\bar{f}_T'(\theta) W_T \bar{f}_T(\theta)\right\}$$

  - Intuition:
    - z_t is ergodic stationary
    - From ergodic theorem:

$$\bar{f}_T(\theta) = \frac{1}{T}\sum_{t=1}^{T} f(z_t,\theta) \xrightarrow{p} E[f(z_t,\theta)]$$

    - Assumption on W_T: $W_T \xrightarrow{p} W$, where W is positive definite.

  - Criterion function

$$\underbrace{Q_T(\theta)}_{\text{sample criterion}} \xrightarrow{p} \underbrace{Q_\infty(\theta)}_{\text{asymptotic criterion}} = E[f(z_t,\theta)]'WE[f(z_t,\theta)]$$

Question:

$$\begin{array}{ccc} Q_T(\theta) & \xrightarrow{p} & Q_\infty(\theta) \\ \downarrow \min & & \downarrow \min \\ \hat{\theta}_T & ?\xrightarrow{p}? & \theta^0 \end{array}$$

- **Theorem.** Suppose
  - $\theta \in \Theta$ where Θ is a compact subset of $\mathbb{R}^p$
  - $Q_T(\cdot)$ is continuous with respect to θ
  - $Q_T(\theta) \xrightarrow{p} Q_\infty(\theta)$ uniformly with respect to θ
  - θ^0 is unique solution of $\min_{\theta\in\Theta}\{Q_\infty(\theta)]$

  then, we have

  $$\hat{\theta}_T = \arg\min_{\theta\in\Theta} Q_T(\theta) \xrightarrow{p} \theta^0$$

  - This is the general result for consistency of extremum estimators

- Special case of GMM.
  - Assumption: stationarity, ergodicity, W positive definite as previously stated
  - $Q_T(\theta) = \bar{f}_T'(\theta)W_T\bar{f}_T(\theta)$ satisfies
    - Continuity: $f(z_t,\theta)$ with respect to θ
    - Uniform convergence: LLN applied to $f(z_t,\theta)$, uniform WLLN for $\bar{f}_T(\theta)$

Consistency of Extremum Estimators (cont'd)

- An ***extremum estimator*** is
$$\hat{\theta}_T = \arg\min_{\theta}[Q_T(\theta)]$$
  - A special case is the minimum distance estimator
$$Q_T(\theta) = \bar{f}'_T(\theta) W_T \bar{f}_T(\theta)$$
where $E[f(z_t, \theta)] = 0$, and
$$\bar{f}_T(\theta) = \frac{1}{T}\sum_{t=1}^{T} f(z_t, \theta)$$
The GMM is
$$Q_\infty(\theta) = E[f(z_t, \theta)]' W E[f(z_t, \theta)]$$
  - Another case is the M-estimator
$$Q_T = \frac{1}{T}\sum_{t=1}^{T} m(z_t, \theta), \qquad Q_\infty(\theta) = E[m(z_t, \theta)]$$
Examples are: OLS, NLS, WLS, WNLS, MLE
$$y_t = h(x_t, \theta) + u_t$$
$$Q_T(\theta) = \frac{1}{T}\sum_{t=1}^{T}\big(y_t - h(x_t, \theta)\big)^2$$

- Consistency of GMM-estimator (as a special case of extremum estimator)
  - Regularity: stationarity, ergodicity, positive definiteness of W
  - $\theta \in \Theta \subset \mathbb{R}^p$, where Θ is compact
  - Continuity of $f(z_t, \cdot)$
  - Uniform convergence for $\bar{f}_T(\theta)$. In this case, we require the uniform LLN for $\bar{f}_T(\theta)$
    - Sufficient condition for uniform convergence of $\bar{f}_T(\cdot)$:
$$E\left[\sup_{\theta\in\Theta}\|f(z_t, \theta)\|\right] < \infty$$

- Asymptotic normality.
  - We need to apply the mean value theorem to the FOC
$$\frac{\partial Q_T(\hat{\theta}_T)}{\partial \theta} = 0$$
Do a mean-value expansion.
    - Recall the MVT: Let $h : \mathbb{R}^p \to \mathbb{R}^q$ continuously differentiable. Then there exists $\bar{\theta} \in [\hat{\theta}_T - \theta^0]$ such that
$$h(\hat{\theta}_T) = h(\theta^0) + \frac{\partial h(\bar{\theta})}{\partial \theta'}(\hat{\theta}_T - \theta^0).$$
Note that $\bar{\theta}$ could be different for each component $h(\cdot)$, which is a vector.

  - The GMM case:
$$Q_T(\theta) = \bar{f}'_T(\theta) W_T \bar{f}_T(\theta)$$

The FOC is

$$\frac{\partial Q_T(\hat{\theta}_T)}{\partial \theta} = 2\frac{\partial \bar{f}_T'(\hat{\theta}_T)}{\partial \theta} W_T \bar{f}_T(\hat{\theta}_T) = 0$$

Focus on $\bar{f}_T(\hat{\theta}_T)$:

$$\bar{f}_T(\theta^0) + \frac{\partial \bar{f}_T(\bar{\theta})}{\partial \theta'}(\hat{\theta}_T - \theta^0)$$

The FOC becomes

$$\underbrace{\left[\frac{\partial \bar{f}_T'(\hat{\theta}_T)}{\partial \theta} W_T \frac{\partial \bar{f}_T(\bar{\theta})}{\partial \theta'}\right]}_{\xrightarrow{p} \Gamma' W \Gamma} \sqrt{T}(\hat{\theta}_T - \theta^0) = -\underbrace{\frac{\partial \bar{f}_T'(\hat{\theta}_T)}{\partial \theta}}_{\xrightarrow{p} \Gamma'} W_T \underbrace{\sqrt{T}\bar{f}_T(\theta^0)}_{\xrightarrow{d} \mathcal{N}(0,\Omega)}$$

Convergence:

$$\frac{\partial \bar{f}_T'(\hat{\theta}_T)}{\partial \theta} \xrightarrow{p} \underbrace{E\left(\frac{\partial f'(z_t, \theta^0)}{\partial \theta}\right)}_{\Gamma'}, \qquad W_T \xrightarrow{p} W, \qquad \frac{\partial \bar{f}_T(\bar{\theta})}{\partial \theta'} \xrightarrow{p} \Gamma$$

We know that $\Gamma' W \Gamma$ is invertible, because Γ is full column rank and W is positive definite. Then,

$$\sqrt{T}(\hat{\theta}_T - \theta^0) = -(\Gamma' W \Gamma)^{-1}\Gamma' W \sqrt{T}\bar{f}_T(\theta^0) + o_p(1)$$

By CLT, we have

$$\sqrt{T}(\hat{\theta}_T - \theta^0) \xrightarrow{d} \mathcal{N}\left(0\,,\ (\Gamma' W \Gamma)^{-1}\Gamma' W \Omega W \Gamma (\Gamma' W \Gamma)^{-1}\right)$$

Consistent GMM Estimator

❖ **Theorem (Asymptotic distribution of GMM estimator).** Under

- Consistency of GMM estimator
- $f(z_t, \theta)$ is continuously differentiable with respect to θ
- Rank assumption with respect to Γ:

$$rank\underbrace{\left(E\left[\frac{\partial f(z_t,\theta^0)}{\partial \theta'}\right]\right)}_{\Gamma} = p$$

- $f(z_t, \theta)$ MDS
- CLT for MDS

We have

$$\sqrt{T}\left(\hat{\theta}_T - \theta^0\right) \xrightarrow{d} \mathcal{N}(0, V)$$

where

$$V = (\Gamma' W \Gamma)^{-1} \Gamma' W \Omega W' \Gamma (\Gamma' W \Gamma)^{-1}$$

➢ Mean-value expansion of the FOC of

$$Q_T(\theta) = \bar{f}_T'(\theta) W \bar{f}_T(\theta)$$

FOC:

$$2\frac{\partial \bar{f}_T'(\hat{\theta}_T)}{\partial \theta} W \bar{f}_T(\hat{\theta}_T) = 0$$

Note that for $\tilde{\theta}_T$ in between θ^0 and $\hat{\theta}_T$ (may be different from different elements of $\bar{f}_T$),

$$\bar{f}_T(\hat{\theta}_T) = \bar{f}_T(\theta^0) + \frac{\partial \bar{f}_T(\tilde{\theta}_T)}{\partial \theta'}(\hat{\theta}_T - \theta^0)$$

$$\Leftrightarrow \frac{\partial \bar{f}_T(\hat{\theta}_T)}{\partial \theta} W_T \frac{\partial \bar{f}_T(\tilde{\theta}_T)}{\partial \theta'} \sqrt{T}\left(\hat{\theta}_T - \theta^0\right) = -\frac{\partial \bar{f}_T'(\hat{\theta}_T)}{\partial \theta} W_T \underbrace{\sqrt{T}\bar{f}_T(\theta^0)}_{\xrightarrow{d} \mathcal{N}(0,\Omega)}$$

By assumption,

$$\begin{aligned}
\hat{\theta}_T \xrightarrow{p} \theta^0 \;\; &\Rightarrow \;\; \tilde{\theta}_T \xrightarrow{p} \theta^0 \\
&\Rightarrow \;\; \frac{\partial \bar{f}_T(\hat{\theta}_T)}{\partial \theta'} \xrightarrow{p} \Gamma \\
&\Rightarrow \;\; \frac{\partial \bar{f}_T(\tilde{\theta}_T)}{\partial \theta'} \xrightarrow{p} \Gamma
\end{aligned}$$

Then,

$$\frac{\partial \bar{f}_T'(\hat{\theta}_T)}{\partial \theta} W_T \frac{\partial \bar{f}_T(\tilde{\theta}_T)}{\partial \theta'} \xrightarrow{p} \Gamma' W \Gamma$$

$$\frac{\partial \bar{f}_T'(\hat{\theta}_T)}{\partial \theta} W_T \xrightarrow{p} \Gamma' W$$

$$\frac{\partial \bar{f}_T'(\hat{\theta}_T)}{\partial \theta} W_T \frac{\partial \bar{f}_T(\tilde{\theta}_T)}{\partial \theta'} \sqrt{T}\left(\hat{\theta}_T - \theta^0\right) \xrightarrow{d} \mathcal{N}\left(0\,,\, \Gamma' W \Omega W' \Gamma\right)$$

$$\Rightarrow \;\; \sqrt{T}\left(\hat{\theta}_T - \theta^0\right) \xrightarrow{d} \mathcal{N}(0, V)$$

where

$$V = (\Gamma' W \Gamma)^{-1} \Gamma' W \Omega W' \Gamma (\Gamma' W \Gamma)^{-1}.$$

- Efficient GMM estimation
  - How to pick the efficient weighting matrix W $\rightarrow$ want to minimize the asymptotic variance of the GMM estimator $\hat{\theta}_T$
    - The only flexibility we have is in choosing W. Since W is symmetric, the idea is to pick $W = \Omega^{-1}$ so that two of the "blocks" in V cancels out with each other:

$$W = \Omega^{-1} \;\Rightarrow\; V = \underbrace{(\Gamma'\Omega^{-1}\Gamma)^{-1}\,\underbrace{\Gamma'\,\underbrace{\Omega^{-1}\Omega}_{I}\,\Omega^{-1}\Gamma}_{\Gamma'\Omega^{-1}\Gamma}}_{I}\,(\Gamma'\Omega^{-1}\Gamma)^{-1}$$

$$\Rightarrow\; V_{opt} = (\Gamma'\Omega^{-1}\Gamma)^{-1}$$

So we have the efficient GMM estimator with

$$AVar(\hat{\theta}_T^*) = (\Gamma'\Omega^{-1}\Gamma)^{-1}$$

  - Example. $f(z_t, \theta) = w_t \underbrace{(y_t - x_t'\theta)}_{u_t(\theta)}$

$$\Omega = E[w_t w_t' u_t^2(\theta^0)]$$
$$= E[w_t w_t' \sigma_t^2(\theta^0)], \qquad \sigma_t^2(\theta^0) = Var\big(u_t(\theta^0)\big)$$

Conditional homoscedasticity (given w_t)
    - $\sigma_t^2(\theta^0) = \sigma^2$
    - Ω proportional to $E(w_t w_t')$
    - Weighting matrix

$$W_T = \left(\frac{1}{T}\sum_t w_t w_t'\right)^{-1} = (W'W)^{-1}$$

The minimization problem is

$$\min_\theta \left[\frac{1}{T}\sum_{t=1}^{T} w_t(y_t - x_t'\theta)\right]' (W'W)^{-1} \left[\frac{1}{T}\sum_{t=1}^{T} w_t(y_t - x_t'\theta)\right]$$
$$\Leftrightarrow\; \min_\theta [W'(y - X\theta)]'(W'W)^{-1}[W'(y - X\theta)]$$
$$\Leftrightarrow\; \min_\theta \left[(y - X\theta)'\underbrace{W(W'W)^{-1}W'}_{=P_W = P_W'P_W}(y - X\theta)\right]$$
$$\Leftrightarrow\; \min_\theta \big[[P_W(y - X\theta)]'[P_W(y - X\theta)]\big]$$
$$\Leftrightarrow\; \min_\theta \|P_W(y - X\theta)\|^2$$

This means that the efficient GMM estimator corresponds to OLS of $P_W y$ on $P_W X$, or the 2SLS estimator of y on $P_W X$.

- The general case
  - Efficient weighting matrix $W_T \xrightarrow{p} \Omega^{-1}$, where $\Omega = Var\big(f(z_t, \theta^0)\big)$.
  - How do we estimate it?
    - 2-step GMM
    - Iterated GMM
    - Continuously updated GMM

- 2-step GMM
  - Step 1: get a consistent GMM estimator with an arbitrary weighting matrix W_T:
  $$\min_\theta\left[\bar{f}_T'(\theta)W_T\bar{f}_T(\theta)\right] \ \longrightarrow\ \tilde{\theta}_T \text{ consistent}$$
  Usually, we pick $W_T = I$.
  - Step 2: use $\tilde{\theta}_T$ to get a consistent estimator of Ω:
  $$\Omega_T(\tilde{\theta}_T) = \frac{1}{T}\sum_{t=1}^{T} f(z_t, \tilde{\theta}_T)f'(z_t, \tilde{\theta}_T)$$
  Use $\Omega_T(\tilde{\theta}_T)$ as the weighting matrix and
  $$\min_\theta\left[\bar{f}_T'(\theta)\left[\Omega_T(\tilde{\theta}_T)\right]^{-1}\bar{f}_T(\theta)\right] \ \longrightarrow\ \hat{\theta}_T \text{ efficient estimator}$$
  Note that $\Omega_T(\tilde{\theta}_T)$ does not depend on θ.

- Motivation for other practical GMM estimation methods:
  - In practice, $\hat{\theta}_T$ or 2S-GMM does not have good finite sample properties. So here are some improvements:
    - Demean $f(z_t, \tilde{\theta}_T)$, as in practice it's not always equal to zero
    $$\Omega_T^*(\tilde{\theta}_T) = \frac{1}{T}\sum_{t=1}^{T}\left[f(z_t, \tilde{\theta}_T) - \bar{f}_T(\tilde{\theta}_T)\right]\left[f(z_t, \tilde{\theta}_T) - \bar{f}_T(\tilde{\theta}_T)\right]'$$
    - Iterated GMM: idea is to keep running GMM until you find $\hat{\theta}_{T,k}$ close enough to $\hat{\theta}_{T,k+1}$.
      - Step k: $\Omega_T(\hat{\theta}_{T,k})$ or $\Omega_T^*(\hat{\theta}_{T,k})$
      $$\min_\theta\left[\bar{f}_T'(\theta)\Omega_T^{-1}(\hat{\theta}_{T,k})\bar{f}_T(\theta)\right] \ \longrightarrow\ \hat{\theta}_{T,k+1}$$
      Continue this process until $\hat{\theta}_{T,k+1}$ is close enough to $\hat{\theta}_{T,k}$ (e.g. $\bar{f}_T(\hat{\theta}_{T,k})$ is close to zero)
    - CU-GMM: integrate all the steps into one single minimization problem
    $$\min_\theta\left[\bar{f}_T'(\theta)\Omega_T^{-1}(\theta)\bar{f}_T(\theta)\right]$$
    Note that we're now not minimizing a quadratic form, so this estimator is of a different class of estimators.
      - Finite sample properties of this estimator are very good.
      - Consistent and asymptotically efficient.
      - However, in practice, there are some "local" optima where your optimization might get stuck $\rightarrow$ need a lot of robustness checks.

- Weighted least squares
$$y_t = h(x_t, \theta) + u_t, \qquad E(u_t|x_t) = 0$$
  - Weighted non-linear least squares:
  $$\min\left[\sum_{t=1}^{T}\alpha_t\left(y_t - h(x_t, \theta)\right)^2\right]$$

FOC:

$$\sum_{t=1}^{T} \underbrace{\alpha_t \frac{\partial h(x_t, \hat{\theta})}{\partial \theta}}_{w_t \ (p\times 1)} \left(y_t - h(x_t, \hat{\theta})\right) = 0$$

$$\Rightarrow \sum_{t=1}^{T} w_t \left(y_t - h(x_t, \hat{\theta})\right)^2$$

Just-identification!

- We can always reinterpret WNLS as GMM with instruments

$$w_t = \alpha_t \frac{\partial h(x_t, \hat{\theta})}{\partial \theta}$$

- More generally, any M-estimator can be reinterpreted as GMM when looking at the FOC
- In practice, in the linear case, the optimal weights are inverse of the variance (which is unknown). Since it is unknown, it has to be estimated in the first step.
  - Efficient WLS:

$$\min \sum \left[\frac{\left(y_t - h(x_t, \theta)\right)^2}{\widehat{Var}(u_t|x_t)} \right]$$

where $\widehat{Var}(u_t|x_t)$ is the result of a first step.

GMM v.s. Maximum Likelihood (ML)

- Framework for ML
  - We have n iid observations: $y_1, \dots, y_n$
  - Parametric model

$$Y_i \sim \underbrace{\ell(y_i, \theta)}_{\text{a pdf}}, \qquad \theta \in \Theta \subset \mathbb{R}^p$$

    For example, $\theta = (\mu \quad \sigma^2)'$
  - (Joint) Density function for $(y_1, \dots, y_n)$

$$\ell_n(y_1, \dots, y_n; \theta) = \prod_{i=1}^{n} \ell(y_i, \theta)$$

  - Likelihood: $\theta \rightarrow \ell_n(y_1, \dots, y_n; \theta)$
  - MLE $\hat{\theta}$:

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \prod_{i=1}^{n} \ell(y_i, \hat{\theta}) = \arg\max_{\theta \in \Theta} \sum_{i=1}^{n} \ln\left(\ell(y_i, \hat{\theta})\right)$$

    FOC:

$$\sum_{i=1}^{n} \frac{\partial \ln\left(\ell(y_i, \theta)\right)}{\partial \theta} = 0 \quad \Leftrightarrow \quad \frac{1}{n} \sum_{i=1}^{n} \frac{\partial \ln\left(\ell(y_i, \theta)\right)}{\partial \theta} = 0$$

    Note that the latter is the mean of some moment conditions

$$E\left[\frac{\partial \ln\left(\ell(y_i, \theta^0)\right)}{\partial \theta}\right] = 0$$

  - What about other orthogonality conditions or other moment conditions?
    - Consider $f(y_i, \theta)$ such that $E\left(f(y_i, \theta^0)\right) = 0$. Perform an affine regression of $\frac{\partial \ln \ell}{\partial \theta}$ on f (since both terms have mean zero, there's no need for a constant term):

$$\frac{\partial \ln\left(\ell(Y, \theta^0)\right)}{\partial \theta} = \beta f(Y, \theta^0) + u$$

      with

$$\beta = Cov\left[\frac{\partial \ln\left(\ell(Y, \theta^0)\right)}{\partial \theta}, f(Y, \theta^0)\right]\left[Var\left(f(Y, \theta^0)\right)\right]^{-1}$$

      By definition,

$$E[f(Y, \theta^0)] = 0 \quad \Leftrightarrow \quad \int f(Y, \theta^0)\ell(Y, \theta^0)dY = 0$$

$$\Rightarrow \underbrace{\int \frac{\partial f(Y, \theta^0)}{\partial \theta'} \ell(Y, \theta^0)dY}_{E\left[\frac{\partial f(Y, \theta^0)}{\partial \theta'}\right] = \Gamma} + \underbrace{\int f(Y, \theta^0) \underbrace{\frac{1}{\ell(Y, \theta^0)} \frac{\partial \ell(Y, \theta^0)}{\partial \theta}}_{\frac{\partial \ln\left(\ell(Y, \theta^0)\right)}{\partial \theta}} \ell(Y, \theta^0)dY}_{Cov\left[f(Y, \theta^0), \frac{\partial \ln(\ell)}{\partial \theta}\right]}$$

$$\Rightarrow \frac{\partial \ln\left(\ell(Y, \theta^0)\right)}{\partial \theta} = -\Gamma'\left[Var\left(f(Y, \theta^0)\right)\right] f(Y, \theta^0) + u$$

      The explained variance is equal to

$$\Gamma'\left[Var\left(f(Y,\theta^0)\right)\right]^{-1}Var\left(f(Y,\theta^0)\right)\left[Var\left(f(Y,\theta^0)\right)\right]^{-1}\Gamma = \Gamma'\left[Var\left(f(Y,\theta^0)\right)\right]^{-1}\Gamma$$

This is the inverse of the variance of efficient GMM with moment condition f.

$$\Gamma = E\left(\frac{\partial f}{\partial \theta}\right) = -Cov\left(f, \frac{\partial \ln \ell}{\partial \theta}\right) = -\mathcal{J}^{-1} = \text{Cramer-Rao lower bound}$$

"Best GMM" is the one where we choose f to maximize the explained variance, i.e. choose $f(Y,\theta) = \frac{\partial \ln \ell(Y,\theta)}{\partial \theta}$. Therefore, GMM is MLE. More specifically, in this case,

$$Var(f) = Var\left(\frac{\partial \ln \ell}{\partial \theta}\right) = \mathcal{J}$$

where $\mathcal{J}$ is the Fisher information matrix.

- Conclusion:
  Provided the parametric form of the density of the data $(y_1, \dots, y_n)$ is known (here we focused on iid).
  - Result: GMM with optimal orthogonality conditions $\frac{\partial \ln \ell}{\partial \theta}$ is numerically equivalent to ML.