AET 1 notes

Nicholas Chiang

June 13, 2024

Lecture notes

1. **Lecture 1: Regression and Projection**

   (a) Basic setup:
   Dependent variable y, k regressors $(x_1, ..., x_k)$
   Note: use lowercase letters for scalars, bold lowercase for vectors and uppercase letters for matrices.

   The random variables $(y, \boldsymbol{x})$ have a joint distribution F which is the population. The population is infinitely large.
   The distribution is unknown. We want to learn about features of F from the sample.

   To study how the distribution of y varies with the variables $\boldsymbol{x}$ in the population, we look at $f(y|\boldsymbol{x})$, the conditional density of y given $\boldsymbol{x}$.

   (b) The Conditional Expectation Function (CEF) is the conditional mean of y given $\boldsymbol{x}$:

   $$m(\boldsymbol{x}) = \mathbb{E}(y|\boldsymbol{x}) = \int_{-\infty}^{\infty} yf(y|\boldsymbol{x})dy$$

   The CEF is a random variable as it is a function of the random variable $\boldsymbol{x}$.
   Note that the expectation of $\boldsymbol{x}$ is

   $$E(\boldsymbol{x}) = \int_{-\infty}^{\infty} xf(\boldsymbol{x})d\boldsymbol{x}$$

   (c) Law of iterated expectations:
   Simple LIE: If $\mathbb{E}|y| < \infty$ then for any random vector $\boldsymbol{x}$,

   $$\mathbb{E}(\mathbb{E}(y|\boldsymbol{x})) = \mathbb{E}(y)$$

   When $\boldsymbol{x}$ is continuous,

   $$\mathbb{E}(\mathbb{E}(y|\boldsymbol{x})) = \int_{\mathbb{R}^k} \mathbb{E}(y|\boldsymbol{x})f_x(\boldsymbol{x})d\boldsymbol{x}$$

   Proof: Since $\mathbb{E}(y|\boldsymbol{x})$ is a function of the random vector $\boldsymbol{x}$ only, to calculate its expectation we integrate with respect to the density $f_x(\boldsymbol{x})$ of $\boldsymbol{x}$, that is

   $$\mathbb{E}(\mathbb{E}(y|\boldsymbol{x})) = \int_{\mathbb{R}^k} \mathbb{E}(y|\boldsymbol{x})f_x(\boldsymbol{x})d\boldsymbol{x}$$

   Then, noting that $f_{y|\boldsymbol{x}x}(y|\boldsymbol{x})f_x(\boldsymbol{x}) = f(y, \boldsymbol{x})$, the above expectation equals

   $$\int_{\mathbb{R}^k} (\int_{\mathbb{R}} yf_{y|\boldsymbol{x}}(y|\boldsymbol{x})dy)f_x(\boldsymbol{x})d\boldsymbol{x} = \int_{\mathbb{R}^k}\int_{\mathbb{R}} y(f(y, \boldsymbol{x})dyd\boldsymbol{x} = \mathbb{E}(y)$$

   which is the unconditional mean of y.

   General law of iterated expectations:
   If $\mathbb{E}|y| < \infty$ then for any random vectors $\boldsymbol{x_1}$ and $\boldsymbol{x_2}$,

   $$\mathbb{E}(\mathbb{E}(y|\boldsymbol{x_1}, \boldsymbol{x_2})|\boldsymbol{x_1}) = \mathbb{E}(y|\boldsymbol{x_1})$$

   “the smaller information set wins”.

(d) Conditioning theorem

When we condition on a random vector $\boldsymbol{x}$, we can effectively treat it as if it is constant.
E.g., $\mathbb{E}(f(\boldsymbol{x})|\boldsymbol{x}) = f(\boldsymbol{x})$ for any function $f(.)$.
The theorem is: If $\mathbb{E}|y| < \infty$ then

$$\mathbb{E}(g(\boldsymbol{x})y|\boldsymbol{x}) \ g(\boldsymbol{x})\mathbb{E}(y|\boldsymbol{x})$$

If in addition $\mathbb{E}|g(\boldsymbol{x})y| < \infty$ then

$$\mathbb{E}(g(\boldsymbol{x})y) = \mathbb{E}(g(\boldsymbol{x})\mathbb{E}(y|\boldsymbol{x}))$$

Proof:

$$\mathbb{E}(g(\boldsymbol{x})y|\boldsymbol{x}) = \int_{\mathbb{R}} g(\boldsymbol{x})yf_{y|\boldsymbol{x}}(y|\boldsymbol{x})dy = g(\boldsymbol{x})\int_{\mathbb{R}} f_{y|\boldsymbol{x}}(y|\boldsymbol{x})dy = g(\boldsymbol{x})\mathbb{E}(y|x)$$

(e) Properties of the CEF error:

The **CEF error** e is the difference between y and the CEF evaluated at the random vector $\boldsymbol{x}$:

$$e = y - m(\boldsymbol{x})$$

e is derived from the joint distribution $(y, \boldsymbol{x})$ and is a function of both y and $\boldsymbol{x}$.

If $\mathbb{E}|y| < \infty$, then

i. $\mathbb{E}(e|\boldsymbol{x}) = 0$ (i.e., CEF error has zero unconditional mean). Can show this by substituting the definition of e and taking expectations. Note that e and x can still be jointly dependent even if this holds.
ii. $\mathbb{E}(e) = 0$ (i.e., CEF error has zero conditional mean). Can show this using LIE.
iii. If $\mathbb{E}|y|^r < \infty$ for $r \geq 1$ then $\mathbb{E}|e|^r < \infty$. This is a regularity condition. See proof below.
iv. For any function $h(\boldsymbol{x})$ such that $\mathbb{E}|h(\boldsymbol{x})e| < \infty$ then $\mathbb{E}(h(\boldsymbol{x})e) = 0$. Can show this using LIE.

Proof of property (iii) above:
Applying Minkowski's inequality to $e = y - m(\boldsymbol{x})$,

$$(\mathbb{E}|e|^r)^{1/r} = (\mathbb{E}|y - m(x)|^r)^{1/r} \leq (\mathbb{E}|y|^r)^{1/r} + (\mathbb{E}|m(\boldsymbol{x})|^r)^{1/r} < \infty$$

since $\mathbb{E}|y|^r < \infty$ by assumption and $\mathbb{E}|m(\boldsymbol{x})|^r < \infty$ by the conditional expectation inequality.

For the variance of the CEF error:

$$\sigma^2 = var(e) = \mathbb{E}((e - \mathbb{E}e)^2) = \mathbb{E}(e^2)$$

Note that property (iii) above implies that:

i. If $\mathbb{E}(y^2) < \infty$ then $\sigma^2 < \infty$
ii. If $\mathbb{E}(y^2) < \infty$ then

$$var(y) \geq var(y - \mathbb{E}(y|\boldsymbol{x_1})) \geq (var(y - \mathbb{E}(y|\boldsymbol{x_1}, \boldsymbol{x_2}))$$

The CEF is the best predictor of y in the sense of achieving the lowest mean squared prediction error. This holds regardless of the joint distribution of $(y, \boldsymbol{x})$.

The mean squared prediction error is defined as

$$\mathbb{E}((y - g(\boldsymbol{x}))^2)$$

where we define the best prediction predictor as the function $g(\boldsymbol{x})$ that minimises the metric above. Proof that the CEF $m(\boldsymbol{x})$ is the best predictor of y:

$$\begin{aligned}
\mathbb{E}((y - g(\boldsymbol{x}))^2) &= \mathbb{E}((e + m(\boldsymbol{x}) - g(\boldsymbol{x}))^2) \\
&= \mathbb{E}(e^2) + 2\mathbb{E}(e(m(\boldsymbol{x}) - g(\boldsymbol{x}))) + \mathbb{E}((m(\boldsymbol{x}) - g(\boldsymbol{x}))^2) \text{ (can transpose since all scalars)} \\
&= \mathbb{E}(e^2) + \mathbb{E}((m(\boldsymbol{x}) - g(\boldsymbol{x}))^2) \text{ the last term is a square } \geq 0 \\
&\geq \mathbb{E}(e^2) = \mathbb{E}((y - m\boldsymbol{x}))^2)
\end{aligned}$$

Hence if $\mathbb{E}(y^2) < \infty$, then for any predictor $g(\boldsymbol{x})$,

$$\mathbb{E}((y - g(\boldsymbol{x}))^2) \geq \mathbb{E}((y - m(\boldsymbol{x}))^2)$$

where $m(\boldsymbol{x}) = \mathbb{E}(y|\boldsymbol{x})$.
(i.e., any other predictor has a greater mean square prediction error).

(f) Regression derivative:
It is typical to consider marginal changes in a single regressor, say x_1, holding the remaining regressors fixed. We define the marginal effect of a change in x_1, holding the variables $x_2, ..., x_k$ fixed, as:

$$\nabla_1 m(\boldsymbol{x}) = \begin{cases} \frac{\partial m(x_1,...,x_k)}{\partial x_1} & \text{if } x_1 \text{ is continuous} \\ m(1, x_2, ..., x_k) - m(0, x_2, ..., x_k) & \text{if } x_1 \text{ is binary} \end{cases}$$

Collecting the k effects into one $k \times 1$ vector, we define the **regression derivative** with respect to **x**:

$$\nabla m(\boldsymbol{x}) = \begin{bmatrix} \nabla_1 m(\boldsymbol{x}) \\ \nabla_2 m(\boldsymbol{x}) \\ ... \\ \nabla_k m(\boldsymbol{x}) \end{bmatrix}$$

(g) Linear regression model: A linear regression model restricts the CEF to be linear in $\boldsymbol{x}$, i.e.,

$$\mathbb{E}(y|\boldsymbol{x}) = m(\boldsymbol{x}) = x_1\beta_1 + ... + x_k\beta_k + \beta_{k+1}$$

This is also called the linear CEF model. The CEF can be written as

$$m(\boldsymbol{x}) = \boldsymbol{x'\beta}$$

where $\boldsymbol{x} = (x_1, x_2..., x_{k-1}, 1)'$ with dimensions $(k \times 1)$ and
$\boldsymbol{\beta} = (\beta_1...\beta_k)'$ with dimensions $(k \times 1)$

The regression derivative (i.e., the gradient vector of partial derivatives) is $\nabla m(\boldsymbol{x}) = \boldsymbol{\beta}$

To summarise, for the linear CEF model:

$$y = \boldsymbol{x}'\boldsymbol{\beta} + e$$
$$\mathbb{E}(e|\boldsymbol{x}) = 0$$

and for the homoskedastic CEF model, we add

$$\mathbb{E}(e^2|\boldsymbol{x}) = \sigma^2$$

The linear CEF model can have nonlinear effects if we include polynomial and interaction terms of $\boldsymbol{x}$, etc.

A linear predictor for y is a function of the form $\boldsymbol{x'\beta}$ for some $\boldsymbol{\beta} \in \mathbb{R}^k$.
We still allow the CEF to be unrestricted (we do not need to assume that the CEF is linear).
The mean squared prediction error is

$$S(\boldsymbol{\beta}) = \mathbb{E}((y - \boldsymbol{x'\beta})^2)$$

Assumption 2.1: To find the best linear predictor, we require the following regularity conditions:

i. $\mathbb{E}(y^2) < \infty$
ii. $\mathbb{E}||x||^2 < \infty$
iii. $\boldsymbol{Q}_{xx} = \mathbb{E}(\boldsymbol{xx'})$ (the "design matrix") is positive definite. This means that $\boldsymbol{Q}_{xx}$ is invertible and hence there is a unique solution to $\boldsymbol{Q}_{xx}\boldsymbol{\beta} = \boldsymbol{Q}_{xy}$.

The best linear predictor of y is found by selecting the vector $\boldsymbol{\beta}$ to minimise $S(\boldsymbol{\beta})$.

Definition: The **best linear predictor** of y given $\boldsymbol{x}$ is

$$P(y|\boldsymbol{x}) = \boldsymbol{x}'\boldsymbol{\beta}$$

where $\boldsymbol{\beta}$ minimises the mean squared prediction error

$$S(\boldsymbol{\beta}) = \mathbb{E}((y - \boldsymbol{x}'\boldsymbol{\beta})^2)$$

The minimiser

$$\boldsymbol{\beta} = argmin_{\boldsymbol{b}\in\mathbb{R}^k} S(\boldsymbol{b})$$

is called the **Linear Projection Coefficient**.

As this is a quadratic minimisation problem, we can explicitly solve for the minimiser. The mean squared prediction error is

$$S(\boldsymbol{\beta}) = \mathbb{E}(y^2) - 2\boldsymbol{\beta}\mathbb{E}(\boldsymbol{x}y) + \boldsymbol{\beta}'E(\boldsymbol{x}\boldsymbol{x}')\boldsymbol{\beta}$$

The FOC with respect to $\boldsymbol{\beta}$ is

$$0 = -2\mathbb{E}(\boldsymbol{x}y) + 2\mathbb{E}(\boldsymbol{x}\boldsymbol{x}')\boldsymbol{\beta}$$

And hence

$$\begin{aligned}\mathbb{E}(\boldsymbol{x}y) &= 2\mathbb{E}(\boldsymbol{x}\boldsymbol{x}')\boldsymbol{\beta}\\ \boldsymbol{Q}_{xy} &= \boldsymbol{Q}_{xx}\boldsymbol{\beta}\\ \boldsymbol{\beta} &= \boldsymbol{Q}_{xx}^{-1}\boldsymbol{Q}_{xy}\\ \boldsymbol{\beta} &= (\mathbb{E}(\boldsymbol{x}\boldsymbol{x}'))^{-1}\mathbb{E}(\boldsymbol{x}y)\end{aligned}$$

Note: Because $\boldsymbol{Q}_{xx}$ is a $k \times k$ matrix and $\boldsymbol{Q}_{xy}$ is a $k \times 1$ vector, alternative expressions like $\frac{\mathbb{E}(xy)}{\mathbb{E}(xx')}$ or $\mathbb{E}(\boldsymbol{x}y)(\mathbb{E}(\boldsymbol{x}\boldsymbol{x}'))^{-1}$ are incoherent and incorrect.

The expression for the best linear predictor (or the **linear projection** of y on $\boldsymbol{x}$) is:

$$P(y|\boldsymbol{x}) = \boldsymbol{x}'(\mathbb{E}(\boldsymbol{x}\boldsymbol{x}'))^{-1}\mathbb{E}(\boldsymbol{x}y)$$

The **projection error** is

$$e = y - \boldsymbol{x}'\beta$$

This equals the error from the regression equation iff the conditional mean is linear in $\boldsymbol{x}$; otherwise they are distinct.
Notes:

i. If the true CEF is nonlinear, the best linear predictor is not the best predictor.
ii. If the true CEF is nonlinear, the projection error is not the CEF error.

Properties of the projection error (these are distinct from the properties of the CEF error): $\mathbb{E}(\boldsymbol{x}e) = 0$. Note that this is a set of k equations, one for each regressor. It is thus equivalent to $\mathbb{E}(x_j e) = 0$, for $j = 1, ..., k$.
Proof:

$$\begin{aligned}\mathbb{E}(\boldsymbol{x}e) &= \mathbb{E}(\boldsymbol{x}(y - \boldsymbol{x}'\boldsymbol{\beta}))\\ &= \mathbb{E}(\boldsymbol{x}y) - \mathbb{E}(\boldsymbol{x}\boldsymbol{x}')(\mathbb{E}(\boldsymbol{x}\boldsymbol{x}'))^{-1}\mathbb{E}(\boldsymbol{x}y) = 0\end{aligned}$$

When the regression vector $\boldsymbol{x}$ contains a constant (e.g., $x_k = 1$), the projection error has mean zero, i.e., $\mathbb{E}(e) = 0$.
Also, since $cov(x_j, e) = \mathbb{E}(x_j e) = \mathbb{E}(x_j)\mathbb{E}(e)$, the above taken together imply that the variables x_j and e are uncorrelated.

(h) **Properties of the linear projection model**
Under Assumption 2.1:

i. The moments $\mathbb{E}(\boldsymbol{x}\boldsymbol{x}')$ and $\mathbb{E}(\boldsymbol{x}y)$ exist with finite elements.
Proof: by the expectation inequality,

$$||\mathbb{E}(\boldsymbol{x}\boldsymbol{x}')|| \leq \mathbb{E}||\boldsymbol{x}\boldsymbol{x}'|| = \mathbb{E}(||\boldsymbol{x}||^2) < \infty$$

Similarly, using the expectation inequality and the CS inequality and Assumption 2.1,

$$||\mathbb{E}(\boldsymbol{x}y)|| \leq \mathbb{E}||\boldsymbol{x}y|| \leq (\mathbb{E}(||\boldsymbol{x}||^2))^{\frac{1}{2}}(\mathbb{E}(y^2))^{\frac{1}{2}} < \infty$$

and hence the moments $\mathbb{E}(\boldsymbol{x}y)$ and $\mathbb{E}(\boldsymbol{x}\boldsymbol{x}')$ are finite and well defined.

ii. The Linear Projection Coefficient exists, is unique, and equals

$$\beta = (\mathbb{E}(\boldsymbol{x}\boldsymbol{x}'))^{-1}\mathbb{E}(\boldsymbol{x}y)$$

Proof: The coefficient $\boldsymbol{\beta} = (\mathbb{E}(\boldsymbol{x}\boldsymbol{x}'))^{-1}\mathbb{E}(\boldsymbol{x}y)$ is well defined since $(\mathbb{E}(\boldsymbol{x}\boldsymbol{x}'))^{-1}$ exists under Assumption 2.1.

iii. The best linear predictor of y given $\boldsymbol{x}$ is

$$P(y|\boldsymbol{x}) = \boldsymbol{x}'(\mathbb{E}(\boldsymbol{x}\boldsymbol{x}'))^{-1}\mathbb{E}(\boldsymbol{x}y)$$

iv. The projection error $e = y - \boldsymbol{x}'\boldsymbol{\beta}$ exists and satisfies

$$\mathbb{E}(e^2) < \infty, \qquad \mathbb{E}(\boldsymbol{x}e) = 0$$

Proof:

$$\begin{aligned}\mathbb{E}(e^2) &= \mathbb{E}((y - \boldsymbol{x}'\boldsymbol{\beta})^2)\\ &= \mathbb{E}(y^2) - 2\mathbb{E}(y\boldsymbol{x}')\boldsymbol{\beta} + \boldsymbol{\beta}\mathbb{E}(\boldsymbol{x}\boldsymbol{x}')\boldsymbol{\beta}\\ &= \mathbb{E}(y^2) - 2\mathbb{E}(y\boldsymbol{x}')(\mathbb{E}(\boldsymbol{x}\boldsymbol{x}'))^{-1}\mathbb{E}(\boldsymbol{x}y)\\ &\leq \mathbb{E}(y^2) < \infty\end{aligned}$$

v. If $\boldsymbol{x}$ contains a constant, then

$$\mathbb{E}(e) = 0$$

vi. If $\mathbb{E}|y|^r < \infty$ and $\mathbb{E}||x||^r < \infty$ for $r \geq 2$ then $\mathbb{E}|e|^r < \infty$.
Proof: Apply Minkowski's inequality to $e = y - \boldsymbol{x}'\boldsymbol{\beta}$.

In summary, for the Linear Projection Model:

$$\begin{aligned}&y = \boldsymbol{x}'\boldsymbol{\beta} + e\\ &\mathbb{E}(\boldsymbol{x}e) = 0\\ &\boldsymbol{\beta} = (\mathbb{E}(\boldsymbol{x}\boldsymbol{x}'))^{-1}\mathbb{E}(\boldsymbol{x}y)\end{aligned}$$

A related concept is Best Linear Approximation, which constructs a linear approximation of $\boldsymbol{x}'\boldsymbol{\beta}$ to CEF $m(\boldsymbol{x})$, and the mean squared approximation error is

$$d(\boldsymbol{\beta}) = \mathbb{E}((m(\boldsymbol{x}) - \boldsymbol{x}'\boldsymbol{\beta})^2)$$

where $m(\boldsymbol{x}) = E[y|\boldsymbol{x}]$.
The best linear approximation (which minimises $\mathbb{E}[(m(\boldsymbol{x}) - \boldsymbol{x}'\boldsymbol{\beta})^2]$) has the same solution as the best linear predictor (which minimises $\mathbb{E}[(y - \boldsymbol{x}'\boldsymbol{\beta})^2)]$.

(i) Omitted variable bias

Assume we have the linear CEF model, i.e., assume the true CEF is linear.
Suppose the CEF is $\boldsymbol{x}_1'\boldsymbol{\beta}_1 + \boldsymbol{x}'2\boldsymbol{\beta}_2$ (these can be two separate paritions of vectors of regressors).

Let the regressors be partitioned as

$$\boldsymbol{x} = \begin{pmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \end{pmatrix}$$

We can write the projection of y on $\boldsymbol{x}$ as

$$\begin{aligned} y &= \boldsymbol{x}'\boldsymbol{\beta} + e \\ &= \boldsymbol{x}_1'\boldsymbol{\beta}_1 + \boldsymbol{x}_2'\boldsymbol{\beta}2 + e \\ \mathbb{E}(\boldsymbol{x}e) &= 0 \end{aligned}$$

Suppose we project y on $\boldsymbol{x}_1$ only (e.g., if $\boldsymbol{x}_2$ is not observed). We have

$$\begin{aligned} y &= \boldsymbol{x}_1'\boldsymbol{\gamma}_1 + u \\ \mathbb{E}(\boldsymbol{x}_1 u) &= 0 \end{aligned}$$

The long and short regressions have different coefficients on $\boldsymbol{x}_1$, and the coefficients are the same only under one of the following conditions:

i. The projection of $\boldsymbol{x}_2$ on $\boldsymbol{x}_1$ yields a set of zero coefficients (i.e., they are uncorrelated), or
ii. The coefficient on $\boldsymbol{x}_2$ i.e., $(\boldsymbol{\beta}_2)$ is 0.

To avoid omitted variable bias, we should include all potentially relevant variables in estimated variables.

(j) Causal effects

A variable x_1 can be said to have a causal effect on y if the latter changes when all other inputs are held constant.
Consider the formulation where the full model for y is

$$y = h(x_1, \boldsymbol{x_2}, \boldsymbol{u})$$

We define the causal effect of x_1 as the change in y due to a change in x_1 holding $\boldsymbol{x_2}$ and $\boldsymbol{u}$ constant. The causal effect of x_1 on y is

$$C(x_1, \boldsymbol{x_2}, \boldsymbol{u}) = \nabla_1 h(x_1, \boldsymbol{x_2}, \boldsymbol{u})$$

This is also called a structural model/ structural effect.

Since the causal effect varies across individuals and is not observable and cannot be measured on the individual level, we focus on aggregate causal effects, in particular the average causal effect.

In the model above, the average causal effect of x_1 on y conditional on x_2 is:

$$\begin{aligned} ACE(x_1, \boldsymbol{x}_2) &= \mathbb{E}(C(x_1, \boldsymbol{x_2}, \boldsymbol{u})|x_1, \boldsymbol{x}_2) \\ &= \int_{\mathbb{R}^l} \nabla_1 h(x_1, \boldsymbol{x_2}, \boldsymbol{u}) f(\boldsymbol{u}|x_1, \boldsymbol{x}_2) d\boldsymbol{u} \end{aligned}$$

The ACE is the population average of the causal effect. What we are doing here is integrating out $\boldsymbol{u}$, which is the set of unobserved characteristics. $f(\boldsymbol{u}|x_1, \boldsymbol{x}_2)$ is the conditional density of $\boldsymbol{u}$ given $x_1, \boldsymbol{x}_2$.

Relationship between the ACE and the regression derivative

In the model above, the CEF is

$$\begin{aligned} m(x_1, \boldsymbol{x}_2) &= \mathbb{E}(h(x_1, \boldsymbol{x_2}, \boldsymbol{u})|x_1, \boldsymbol{x}_2) \\ &= \int_{R^l} h(x_1, \boldsymbol{x_2}, \boldsymbol{u}) f(\boldsymbol{u}|x_1, \boldsymbol{x}_2) d\boldsymbol{u} \end{aligned}$$

Applying the marginal effect operator, the regression derivative is

$$\begin{aligned}\nabla_1 m(x_1, \boldsymbol{x}_2) &= \int_{R^l} \nabla_1 h(x_1, \boldsymbol{x_2, u}) f(\boldsymbol{u}|x_1, \boldsymbol{x}_2) d\boldsymbol{u} \\ &+ \int_{R^l} h(x_1, \boldsymbol{x_2, u}) \nabla_1 f(\boldsymbol{u}|x_1, \boldsymbol{x}_2) d\boldsymbol{u} \\ &= ACE(x_1, \boldsymbol{x}_2) + \int_{R^l} h(x_1, \boldsymbol{x_2, u}) \nabla_1 f(\boldsymbol{u}|x_1, \boldsymbol{x}_2) d\boldsymbol{u}\end{aligned}$$

Hence, in general the regression derivative does not equal the ACE; the difference is the second term. When both are equal, the regression analysis can be interpreted causally (in the ACE sense).

When the second term is zero, the Conditional Independence Assumption (CIA) is satisfied, i.e., conditional on $\boldsymbol{x}_2$, the random variables x_1 and $\boldsymbol{u}$ are statistically independent.
When the CIA holds, $f(\boldsymbol{u}|x_1, \boldsymbol{x}_2) = f(\boldsymbol{u}|\boldsymbol{x}_2)$ does not depend on x_1, and thus $\nabla_1 f(\boldsymbol{u}|x_1, \boldsymbol{x}_2) = 0$, and

$$\nabla_1 m(x_1, \boldsymbol{x_2}) = ACE(x_1, \boldsymbol{x_2}) \qquad \text{(Theorem 2.11)}$$

i.e., the regression derivative equals the average causal effect for x_1 on y conditional on $\boldsymbol{x_2}$.

(k) Matrix algebra and inequalities

For $n \times n$ matrices A and B, and constant c:

i. $trace(AB) = trace(BA)$
ii. $trace(cA) = c\ trace(A)$
iii. $trace(A'A) = (vec(A))^2$
iv. $||vec(A)|| = trace(A'A)^2$

where the trace is the sum of diagonal elements, and $vec(A)$ is the matrix formed by stacking all the columns of A.

The Euclidean norm of a $m \times 1$ vector $\mathbf{a}$ is

$$\begin{aligned}||\mathbf{a}|| &= (\mathbf{a}'\mathbf{a}^{\frac{1}{2}}) \\ &= (\sum_{i=1}^{m} a_i^2)^{\frac{1}{2}}\end{aligned}$$

The Euclidean norm of a $m \times n$ matrix A is

$$\begin{aligned}||A|| &= ||vec(A)|| \\ &= trace(A'A\text{"})^{\frac{1}{2}} \\ &= (\sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}^2)^{\frac{1}{2}}\end{aligned}$$

For any $m \times 1$ vectors $\mathbf{a}$ and $\mathbf{b}$,

$$\begin{aligned}||\mathbf{a}'\mathbf{b}|| &= ||\mathbf{a}||\ ||\mathbf{b}|| \\ ||\mathbf{a}\mathbf{a}'|| &= ||\mathbf{a}||^2\end{aligned}$$

Schwarz Inequality: For any $m \times 1$ vectors $\mathbf{a}$ and $\mathbf{b}$,

$$|\mathbf{a}'\mathbf{b}| \leq ||\mathbf{a}||\ ||\mathbf{b}||$$

Schwarz Matrix Inequality: For any $m \times n$ matrices A and B,

$$||A'B|| \leq ||A||\ ||B||$$

Triangle inequality: For any $m \times n$ matrices A and B,

$$||A + B|| \leq ||A|| + ||B||$$

Cauchy-Schwarz Inequality: For any random $m \times n$ matrices X and Y,

$$\mathbb{E}||X'Y|| \leq (\mathbb{E}||X||^2)^{\frac{1}{2}}(\mathbb{E}||Y||^2)^{\frac{1}{2}}$$

Holder's Inequality (this is a generalisation of CS Inequality): If $p > 1$ and $q > 1$ and $\frac{1}{p} + \frac{1}{q} = 1$, then for any random $m \times n$ matrices X and Y,

$$\mathbb{E}||X'Y|| \leq (E||X||^p)^{\frac{1}{p}}(\mathbb{E}||Y||^q)^{\frac{1}{q}}$$

Minkowski's Inequality: For any random $m \times x$ matrices X and Y,

$$(\mathbb{E}||X + Y||^p)^{\frac{1}{p}} \leq (\mathbb{E}||X||^p)^{\frac{1}{p}} + (\mathbb{E}||Y||^p)^{\frac{1}{p}}$$

Jensen's inequality: If $g(.) : \mathbb{R} \to \mathbb{R}$ is convex, then for any random variable x for which $\mathbb{E}|x| \leq \infty$ and $\mathbb{E}|g(x)| < \infty$,

$$g(\mathbb{E}(x)) \leq \mathbb{E}(g(x))$$

Expectation Inequality: For any random variable x for which $\mathbb{E}|x| \leq \infty$,

$$|\mathbb{E}(x)| \leq \mathbb{E}|x|$$

Conditional Expectation Inequality: For any $r \geq 1$ such that $\mathbb{E}|y|^r \leq \infty$, then

$$\mathbb{E}(|\mathbb{E}(y|x)|^r) \leq \mathbb{E}|y|^r < \infty$$

(l) Additional notes from Hansen Chapters 1-2

Conditional variance:
If $\mathbb{E}[\boldsymbol{y}^2] \leq \infty$, the conditional variance of y given $\boldsymbol{x}$ is $var[y|\boldsymbol{x}] = \mathbb{E}[(y - \mathbb{E}[y|\boldsymbol{x}])^2|\boldsymbol{x}]$
If $\mathbb{E}[e^2] \leq \infty$, the conditional variance of e given $\boldsymbol{x}$ is $\sigma^2(\boldsymbol{x}) = var[e|\boldsymbol{x}] = \mathbb{E}[e^2|\boldsymbol{x}]$
If $\mathbb{E}[y^2] \leq \infty$, then $var[y] = \mathbb{E}[var[y|\boldsymbol{x}]] + var[\mathbb{E}[y|\boldsymbol{x}]]$ (i.e., sum of within-group variance and across-group variance).

$\boldsymbol{\beta} = \mathbb{E}[\boldsymbol{xx}']^{-1}\mathbb{E}[\boldsymbol{xy}]$ exists and is unique as long as $\mathbb{E}[\boldsymbol{xx}']$ is invertible.

$\boldsymbol{Q_{xx}} = \mathbb{E}[\boldsymbol{xx}']$ is the "design matrix". For any nonzero $\boldsymbol{\alpha} \in \mathbb{R}^k, \boldsymbol{\alpha}'\boldsymbol{Q_{xx}} = \mathbb{E}[\boldsymbol{\alpha}'\boldsymbol{xx}'\boldsymbol{\alpha}] = \mathbb{E}[(\boldsymbol{\alpha}'\boldsymbol{x})^2] \geq 0$, i.e., $\boldsymbol{Q_{xx}}$ is positive semidefinite by construction and hence invertible.

To prove that for $f(x) = a - 2b'\boldsymbol{x} - \boldsymbol{x}'C\boldsymbol{x}$, the unique minimiser is $\boldsymbol{x} = C^{-1}b$:
The FOC is

$$\begin{aligned}\frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}} &= -2b + 2C\boldsymbol{x} = 0\\ Cx &= b, \boldsymbol{x} = C^{-1}b\end{aligned}$$

The SOC is positive and hence this is a minimum point:

$$2C > 0$$

The linear predictor error variance σ^2 is:

$$\begin{aligned}\sigma^2 &= \mathbb{E}[e^2]\\ &= \mathbb{E}[(y - \boldsymbol{x}'\boldsymbol{\beta})^2]\\ &= \mathbb{E}[y^2] - 2E[y\boldsymbol{x}'\boldsymbol{\beta}] + \boldsymbol{\beta}'\mathbb{E}[\boldsymbol{xx}']\boldsymbol{\beta}\\ &= Q_{yy} - 2\boldsymbol{Q_{yx}}\boldsymbol{Q_{xx}^{-1}}\boldsymbol{Q_{xy}} + \boldsymbol{Q_{yx}}\boldsymbol{Q_{xx}^{-1}}\boldsymbol{Q_{xx}}\boldsymbol{Q_{xx}^{-1}}\boldsymbol{Q_{xy}}\\ &= Q_{yy} - \boldsymbol{Q_{yx}}\boldsymbol{Q_{xx}^{-1}}\boldsymbol{Q_{xy}} \equiv Q_{yy \cdot x}\end{aligned}$$

We can separate the constant from the other regressors and write

$$\begin{aligned} &y + \boldsymbol{x}'\boldsymbol{\beta} + \alpha + e \\ &\mathbb{E}[y] = \mathbb{E}[\boldsymbol{x}'\boldsymbol{\beta}] + \mathbb{E}[\alpha] + \mathbb{E}[e] \end{aligned}$$

Since $\mu_y = \boldsymbol{\mu_x'}\boldsymbol{\beta} + \alpha$, we can subtract this from the equation above to get:

$$\begin{aligned} y - \mu y &= (\boldsymbol{x} - \boldsymbol{\mu_x})'\boldsymbol{\beta} + e \\ \beta &= (\mathbb{E}[(\boldsymbol{x} - \boldsymbol{\mu_x})(\boldsymbol{x} - \boldsymbol{\mu_x})'])^{-1}\mathbb{E}[(\boldsymbol{x} - \boldsymbol{\mu_x})(Y - \mu_y)] \\ &= var[\boldsymbol{x}]^{-1}cov(\boldsymbol{x}, y) \end{aligned}$$

2. **Lecture 2: Algebra of Least Squares**

(a) Recap: Rules for differentiation
For vectors $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{x}, \boldsymbol{\beta}$ and symmetric matrix A:

$$\begin{aligned} \frac{\partial \boldsymbol{a'b}}{\partial \boldsymbol{b}} &= \boldsymbol{a} \\ \frac{\partial \boldsymbol{\beta}' A \boldsymbol{\beta}}{\partial \boldsymbol{\beta}} &= 2A\boldsymbol{\beta} \\ \frac{\partial A\boldsymbol{x}}{\partial \boldsymbol{x}} &= A \end{aligned}$$

(b) Samples
We are interested in estimating the parameters of the linear projection model, in particular the projection coefficient, using a data sample:

$$\boldsymbol{\beta} = (\mathbb{E}(\boldsymbol{xx}'))^{-1}\mathbb{E}(\boldsymbol{x}y)$$

Notationally, we want to distinguish observations from the underlying random variables, so we **denote observations using subscript** i which runs from 1 to n, where n is the sample size. The i^{th} observation is $(y_i, \boldsymbol{x_i}$.
We assume that the observations $\{(y_i, \boldsymbol{x_i}), i = 1, ..., n\}$ are identically distributed, and are draws from a common distribution/ data generating process F.

Under this assumption, the linear projection model applies to the random observations $(y_i, \boldsymbol{x_i})$ (i.e., we are projecting the observations of y on the observations of $\boldsymbol{x}$). We can then write the model as

$$y_i = \boldsymbol{x_i'}\boldsymbol{\beta} + e_i$$

where

$$\begin{aligned} \boldsymbol{\beta} &= argmin_{\boldsymbol{b} \in \mathbb{R}^k} S(\boldsymbol{b}) \\ S(\boldsymbol{\beta}) &= \mathbb{E}((y_i - \boldsymbol{x_i'}\boldsymbol{\beta})^2) \\ \boldsymbol{\beta} &= (\mathbb{E}(\boldsymbol{x_i}\boldsymbol{x_i'}))^{-1}\mathbb{E}(\boldsymbol{x_i}y_i) \end{aligned}$$

(c) Moment Estimators
$\boldsymbol{\beta}$ is written as a function of some population expectations.
A **moment estimator** is constructed as the same function, but of the **sample moments**.

e.g., to estimate the sample mean $\mu = \mathbb{E}(y_i) = \int_{-\infty}^{\infty} ydF(y)$, a natural estimator is the sample mean $\hat{\mu} = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

More generally, suppose we are interested in estimating $\boldsymbol{\beta}$, which is a possibly nonlinear function of a set of moments

$$\begin{aligned} \boldsymbol{\beta} &= \boldsymbol{g}(\boldsymbol{\mu}) \\ \boldsymbol{\mu} &= \mathbb{E}(\boldsymbol{h}(\boldsymbol{y_i})) \end{aligned}$$

Then a natural estimator (or "plug-in" estimator/ moment-based estimator) is

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{h}(\boldsymbol{y_i})$$

e.g., if we want to estimate the population variance

$$\sigma^2 = var(y_i) = \mathbb{E}(y_i^2) - (\mathbb{E}(y_i))^2$$

we can use the sample counterparts to form the estimator

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - (\frac{1}{n}\sum_{i=1}^{n} y_i)^2$$

(d) Least Squares Estimator

The moment estimator of the expected squared error $S(\boldsymbol{\beta})$ is

$$\hat{S}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2 = \frac{1}{n}SSE(\boldsymbol{\beta})$$

where the SSE function is

$$SSE(\boldsymbol{\beta}) = \sum_{i=1}^{n}(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2$$

Note: the population moment is $S(\boldsymbol{\beta}) = \mathbb{E}[(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2]$ while the sample moment is $\hat{S}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2$.

The **least-squares estimator $\hat{\boldsymbol{\beta}}$** is

$$\hat{\boldsymbol{\beta}} = argmin_{\boldsymbol{\beta}\in\mathbb{R}^k}\hat{S}(\boldsymbol{\beta})$$

where

$$\hat{S}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2$$

This is also called the OLS estimator.

Where there is only one regressor x_i and no intercept (i.e., both x and β are scalars, then

$$\begin{aligned} SSE(\beta &= \sum_{i=1}^{n}(y_i - x_i\beta)^2 \\ &= (\sum_{i=1}^{n} y_i^2) - 2\beta(\sum_{i=1}^{n} x_i y_i) + \beta^2(\sum_{i=1}^{n} x_i^2) \end{aligned}$$

where the minimiser of this quadratic function is

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

In an intercept-only model, we set $x_i = 1$, and the minimiser is

$$\hat{\beta} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} 1} = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}$$

When there are $k > 1$ regressors, we have

$$SSE(\boldsymbol{\beta}) = \sum_{i=1}^{n} y^2 - 2\boldsymbol{\beta}'\sum_{i=1}^{n}\boldsymbol{x}_i y_i + \boldsymbol{\beta}'\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'\boldsymbol{\beta}$$

The FOC consists of the k equations

$$0 = \frac{\partial SSE(\hat{\boldsymbol{\beta}}}{\partial \boldsymbol{\beta}} = -2\sum_{i=1}^{n} \boldsymbol{x}_i y_i + 2\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' \hat{\boldsymbol{\beta}}$$

And hence the linear projection minimiser is

$$\hat{\beta} = (\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i')^{-1} (\sum_{i=1}^{n} \boldsymbol{x}_i y_i)$$

We require $\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' > 0$ (i.e., this matrix is positive definite) for the minimiser to exist and be unique.
We can also check that this is a minimiser by looking at the SOC:

$$\frac{\partial SSE(\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} = 2\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' \qquad \text{(which is also PSD, i.e., a minimiser)}$$

Alternatively, we can construct the estimator using sample moments.
Since $\boldsymbol{\beta} = (\mathbb{E}(\boldsymbol{x}_i \boldsymbol{x}_i'))^{-1} \mathbb{E}(\boldsymbol{x}_i y_i)$, and the population moments are

$$\begin{aligned} \boldsymbol{Q_{xx}} &= \mathbb{E}(\boldsymbol{x}_i \boldsymbol{x}_i') \\ \boldsymbol{Q_{xy}} &= \mathbb{E}(\boldsymbol{x}_i y_i) \end{aligned}$$

Then we can construct the moment estimator of $\boldsymbol{\beta}$ using sample moments:

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \hat{\boldsymbol{Q}}_{\boldsymbol{xx}}^{-1} \hat{\boldsymbol{Q}}_{\boldsymbol{xy}} \\ &= (\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i')^{-1} (\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i y_i) \\ &= (\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i')^{-1} (\sum_{i=1}^{n} \boldsymbol{x}_i y_i) \end{aligned}$$

The fitted value is

$$\hat{y}_i = \boldsymbol{x}_i' \hat{\boldsymbol{\beta}}$$

and the residual

$$\hat{e}_i = y_i - \hat{y}_i = y_i - \boldsymbol{x}_i' \hat{\boldsymbol{\beta}}$$

These are "fitted" rather than "predicted" because the fitted value is a function of the entire sample, including y_i, and cannot be interpreted as a valid prediction of y_i (since we are not doing out-of-sample prediction).
The residual $\hat{e}$ is a by-product of estimation, but the error e_i is unobservable.
We will show later that

$$\sum_{i=1}^{n} \boldsymbol{x}_i \hat{e}_i = 0$$

And when $\boldsymbol{x}_i$ contains a constant,

$$\frac{1}{n}\sum_{i=1}^{n} \hat{e}_i = 0$$

(e) Demeaned regressors
We can rewrite the linear projection model by separating out the intercept term:

$$y_i = \boldsymbol{x}_i' \boldsymbol{\beta} + \alpha + e_i$$

The FOCs are

$$\begin{aligned} &\sum_{i=1}^{n} (y_i - \boldsymbol{x}_i' \hat{\boldsymbol{\beta}} - \hat{\alpha}) = 0 \\ &\sum_{i=1}^{n} \boldsymbol{x}_i (y_i - \boldsymbol{x}_i' \hat{\boldsymbol{\beta}} - \hat{\alpha}) = \mathbf{0} \end{aligned}$$

The first equation implies

$$\hat{\alpha} = \bar{y} - \bar{\boldsymbol{x}}'\hat{\boldsymbol{\beta}}$$

Sub into the second equation and solve ofr $\hat{\boldsymbol{\beta}}$ to find

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\sum_{i=1}^{n} \boldsymbol{x}_i(\boldsymbol{x}_i - \bar{\boldsymbol{x}})')^{-1}(\sum_{i=1}^{n} \boldsymbol{x}_i(y_i - \bar{y})) \\ &= (\sum_{i=1}^{n} (\boldsymbol{x}_i - \bar{x})(\boldsymbol{x}_i - \bar{x})')^{-1}(\sum_{i=1}^{n} (\boldsymbol{x}_i - \bar{\boldsymbol{x}})(y_i - \bar{y}))\end{aligned}$$

This is the demeaned formula for the LS estimator.

(f) Model in matrix notation
As there are n observations, we can stack the n equations together as

$$\begin{aligned}y_1 &= \boldsymbol{x}_1'\boldsymbol{\beta} + e_1 \\ &\dots \\ y_n &= \boldsymbol{x}_n'\boldsymbol{\beta} + e_n\end{aligned}$$

And we can define the stacked vectors as

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{e}$$

This also allows the sample sums to be written in matrix notation, so we have the LS estimator:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{y})$$

The residual vector is

$$\hat{\boldsymbol{e}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$$

The SSE is

$$SSE(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta})$$

We can show that $\boldsymbol{X}'\hat{\boldsymbol{e}} = \boldsymbol{0}$:

$$\begin{aligned}\boldsymbol{X}'\hat{\boldsymbol{e}} &= \boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \\ &= (\boldsymbol{X}'\boldsymbol{y}) - \boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \\ &= \boldsymbol{X}'\boldsymbol{y} - \boldsymbol{X}'\boldsymbol{y} = \boldsymbol{0}\end{aligned}$$

Another exercise: Let $\hat{\boldsymbol{e}}$ be the OLS residual from a regression of y on $\boldsymbol{X} = [\boldsymbol{X}_1\ \boldsymbol{X}_2]$. Then,

$$\boldsymbol{X}_2'\hat{\boldsymbol{e}} = \boldsymbol{\Gamma}'\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$$

where $\boldsymbol{X}_2 = \boldsymbol{X\Gamma}$ and Γ is a $k \times k_2$ matrix $\begin{bmatrix}0 \\ I\end{bmatrix}$.
Then,

$$\begin{aligned}\boldsymbol{X}_2'\hat{\boldsymbol{e}} &= \boldsymbol{\Gamma}'\boldsymbol{X}'\boldsymbol{y} - \boldsymbol{\Gamma}'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \\ &= \boldsymbol{0}\end{aligned}$$

(g) Projection matrix
Define the projection matrix as

$$\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$$

which is symmetric and idempotent.

We can also show that $\boldsymbol{PX} = \boldsymbol{X}$.
More generally, for any matrix that can be written as $\boldsymbol{Z} = \boldsymbol{X\Gamma}$ (where $\boldsymbol{\Gamma}$ is a matrix of constants), we have

$$\boldsymbol{PZ} = \boldsymbol{Z}$$

(e.g., this holds if we partition $\boldsymbol{X}$ into two matrices $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$, then $\boldsymbol{PX}_1 = \boldsymbol{X}_1$).

The matrix $\boldsymbol{P}$ creates the fitted values in a regression:

$$\boldsymbol{Py} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{y}}$$

A special example occurs in the intercept-only model, when $\boldsymbol{X} = \mathbf{1}_n$ is a vector of ones. Then,

$$\begin{aligned}\boldsymbol{P} &= \mathbf{1}_n(\mathbf{1}_n'\mathbf{1}_n)^{-1}\mathbf{1}_n' \\ &= \frac{1}{n}\mathbf{1}_n\mathbf{1}_n'\end{aligned}$$

And

$$\begin{aligned}\boldsymbol{Py} &= \frac{1}{n}\mathbf{1}_n\mathbf{1}_n'\boldsymbol{y} \\ &= \mathbf{1}_n\bar{y}\end{aligned}$$

Properties of the $\boldsymbol{P}$ matrix, for any $n \times k$ $\boldsymbol{X}$, with $n \geq k$:

i. $\boldsymbol{P}$ is symmetric
ii. $\boldsymbol{P}$ is idempotent
iii. $tr(\boldsymbol{P}) = k$.
Proof:

$$\begin{aligned}tr(\boldsymbol{P}) &= tr(\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}') \\ &= tr((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}) \qquad \text{since } tr(AB) = tr(BA) \\ &= tr(\boldsymbol{I}_k) = k\end{aligned}$$

iv. The eigenvalues of $\boldsymbol{P}$ are 1 and 0. There are k eigenvalues equalling 1, and $n-k$ eigenvalues equalling 0.
Proof: The eigenvalues of an idempotent matrix are all 1 and 0, and the trace of a matrix is the sum of its eigenvalues, and the trace of $\boldsymbol{P} = k$ as shown earlier.
v. $rank(\boldsymbol{P}) = k$
Proof: $\boldsymbol{P}$ is positive semi-definite (PSD) since all its eigenvalues are nonnegative. The rank of a positive semidefinite matrix equals the number of strictly positive eigenvalues.

(h) Orthogonal Projection
Define the orthogonal projection matrix (annihilator matrix) as

$$\begin{aligned}\boldsymbol{M} &= \boldsymbol{I}_n - \boldsymbol{P} \\ &= \boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\end{aligned}$$

Note that $\boldsymbol{M}$ and $\boldsymbol{X}$ are orthogonal:

$$\boldsymbol{MX} = (\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{X} = \mathbf{0}$$

More generally, for any matrix $\boldsymbol{Z}$ in the range space of $\boldsymbol{X}$, we have

$$\boldsymbol{MZ} = \boldsymbol{Z} - \boldsymbol{PZ} = \mathbf{0}$$

For example, $\boldsymbol{MX}_1 = \mathbf{0}$; $\boldsymbol{MP} = \mathbf{0}$.

Properties of the $\boldsymbol{M}$ matrix:

i. $\boldsymbol{M}$ is symmetric
ii. $\boldsymbol{M}$ is idempotent
iii. $tr(\boldsymbol{M}) = n - k$

iv. The eigenvalues of $\boldsymbol{M}$ are 1 and 0. There are $n-k$ eigenvalues equalling 1, and k eigenvalues equalling 0.

v. $rank(\boldsymbol{M}) = n-k$.

The $\boldsymbol{M}$ matrix creates LS residuals:

$$\boldsymbol{My} = \boldsymbol{y} - \boldsymbol{Py} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{e}}$$

The residuals can also be expressed as

$$\hat{\boldsymbol{e}} = \boldsymbol{My} = \boldsymbol{M}(\boldsymbol{X\beta} + \boldsymbol{e}) = \boldsymbol{Me}$$

which is free from dependence on $\boldsymbol{\beta}$, i.e., the residuals are just a linear function of the error terms.

Under the special case of $\boldsymbol{X} - \mathbf{1}_n$,

$$\begin{aligned}\boldsymbol{M} &= \boldsymbol{I}_n - \boldsymbol{P} \\ &= \boldsymbol{I}_n - \mathbf{1}_n(\mathbf{1}_n'\mathbf{1}_n)^{-1}\mathbf{1}_n'\end{aligned}$$

which is a $n \times n$ matrix with 0 as diagonal elements and -1 as off-diagonal elements.

The demeaned value of $\bar{y}$ can be expressed as

$$\boldsymbol{My} = \boldsymbol{y} - \mathbf{1}_n\bar{y}$$

(i) Estimation of Error Variance
Consider the estimation of error variance $\sigma^2 = \mathbb{E}(e_i^2)$.
A natural estimator is the moment estimator

$$\tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} e_i^2$$

but this is infeasible as e_i is not observed. The feasible estimator is

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \hat{e}_i^2$$

In matrix notation, the infeasible estimator is

$$\tilde{\sigma}^2 = \frac{1}{n}\boldsymbol{e}'\boldsymbol{e}$$

while the feasible estimator is

$$\hat{\sigma}^2 = \frac{1}{n}\hat{\boldsymbol{e}}'\hat{\boldsymbol{e}}$$

In terms of the relationship between the two estimators, we have

$$\hat{\boldsymbol{e}} = \boldsymbol{My} = \boldsymbol{Me}$$

and hence

$$\begin{aligned}\hat{\sigma^2} &= \frac{1}{n}\hat{\boldsymbol{e}}'\hat{\boldsymbol{e}} \\ &= \frac{1}{n}\boldsymbol{y}'\boldsymbol{MMy} \\ &= \frac{1}{n}\boldsymbol{e}'\boldsymbol{Me}\end{aligned}$$

We can also show that the difference between the estimators is:

$$\begin{aligned}\tilde{\sigma}^2 - \hat{\sigma^2} &= \frac{1}{n}\boldsymbol{e}'\boldsymbol{e} - \frac{1}{n}\boldsymbol{e}'\boldsymbol{Me} \\ &= \frac{1}{n}\boldsymbol{e}'\boldsymbol{Pe} && \geq 0\end{aligned}$$

since $\boldsymbol{P}$ is positive semi-definite. Hence, the feasible estimator is numerically smaller than the infeasible estimator.

We can also break down $\boldsymbol{y}$ into

$$\boldsymbol{y} = \boldsymbol{P}\boldsymbol{y} + \boldsymbol{M}\boldsymbol{y} = \hat{\boldsymbol{y}} + \hat{\boldsymbol{e}}$$

This decomposition is orthogonal, i.e.,

$$\hat{\boldsymbol{y}}'\hat{\boldsymbol{e}} = (\boldsymbol{P}\boldsymbol{y})'(\boldsymbol{M}\boldsymbol{y}) = \boldsymbol{y}'\boldsymbol{P}\boldsymbol{M}\boldsymbol{y} = 0$$

since $\boldsymbol{P}$ and $\boldsymbol{M}$ are orthogonal so $\boldsymbol{P}\boldsymbol{M} = \boldsymbol{0}$
It follows that

$$\boldsymbol{y}'\boldsymbol{y} = \hat{\boldsymbol{y}}'\hat{\boldsymbol{y}} + 2\hat{\boldsymbol{y}}'\hat{\boldsymbol{e}} + \hat{\boldsymbol{e}}'\hat{\boldsymbol{e}} = \hat{\boldsymbol{y}}'\hat{\boldsymbol{y}} + \hat{\boldsymbol{e}}'\hat{\boldsymbol{e}}$$

Or, in summation notation

$$\sum_{i=1}^{n}(y_i - \bar{y}) = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}\hat{e}_i^2$$

or, TSS = ESS + RSS.

The coefficient of determination, or **R-squared** is

$$R^2 = \frac{ESS}{TSS}$$

which is often described as the fraction of the sample variance of $\boldsymbol{y}$ which is explained by the LS fit. It is a crude measure of regression fit and increases when regressors are added to a regression.

(j) Regression Components
Sometimes we would like to focus on some of the regression coefficients only.
If so, we can partition $\boldsymbol{X}$ and $\boldsymbol{\beta}$ as follows:

$$\boldsymbol{X} = [\boldsymbol{X}_1 \qquad \boldsymbol{X}_2] \qquad \text{(note that the partitions can be of any size within } \boldsymbol{X}\text{)}$$
$$\boldsymbol{\beta} = \begin{pmatrix}\boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2\end{pmatrix}$$

The regression model can be rewritten as

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2 + \boldsymbol{\beta}_2 + \boldsymbol{e}$$

The LS estimator by definition is found by the joint minimisation

$$(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2) = arg \min_{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2} SSE(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$$

where

$$SSE(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = (\boldsymbol{y} - \boldsymbol{X}_1\boldsymbol{\beta}_1 - \boldsymbol{X}_2\boldsymbol{\beta}_2)'(\boldsymbol{y} - \boldsymbol{X}_1\boldsymbol{\beta}_1 - \boldsymbol{X}_2\boldsymbol{\beta}_2)$$

We can obtain an equivalent expression for $\hat{\boldsymbol{\beta}}_1$ by concentration (i.e., nested minimisation).
Equivalently, the solution can also be written as

$$\hat{\boldsymbol{\beta}}_1 = arg \min_{\boldsymbol{\beta}_1}(\min_{\boldsymbol{\beta}_2} SSE(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2))$$

Where the inner problem is the LS regression of $\boldsymbol{y} - \boldsymbol{X}_1\boldsymbol{\beta}_1$ on $\boldsymbol{X}_2$, and the inner solution is

$$arg \min_{\boldsymbol{\beta}_2} SSE(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = (\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}(\boldsymbol{X}_2'(\boldsymbol{y} - \boldsymbol{X}_1\boldsymbol{\beta}_1))$$

The residuals from the inner regression are

$$\begin{aligned}\boldsymbol{y} - \boldsymbol{X}_1\boldsymbol{\beta}_1 - \boldsymbol{X}_2(\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}(\boldsymbol{X}_2'(\boldsymbol{y} - \boldsymbol{X}_1\boldsymbol{\beta}_1)) &= (\boldsymbol{M}_2\boldsymbol{y} - \boldsymbol{M}_2\boldsymbol{X}_1\boldsymbol{\beta}_1) \\ &= \boldsymbol{M}_2(\boldsymbol{y} - \boldsymbol{X}_1\boldsymbol{\beta}_1)\end{aligned}$$

where

$$\boldsymbol{M}_2 = \boldsymbol{I}_n - \boldsymbol{X}_2(\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'$$

The minimised value of the inner problem is

$$\min_{\boldsymbol{\beta}_2} SSE(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = (\boldsymbol{y} - \boldsymbol{X}_1\boldsymbol{\beta}_1)'\boldsymbol{M}_2\boldsymbol{M}_2(\boldsymbol{y} - \boldsymbol{X}_1\boldsymbol{\beta}_1)$$

We can now invoke the OLS formula and write $\hat{\boldsymbol{\beta}}_1$ as

$$\begin{aligned}\hat{\boldsymbol{\beta}}_1 &= [(\boldsymbol{M}_2\boldsymbol{X}_1)'(\boldsymbol{M}_2\boldsymbol{X}_1)]^{-1}[(\boldsymbol{M}_2\boldsymbol{X}_1)'(\boldsymbol{M}_2\boldsymbol{y}] \\ &= [\boldsymbol{X}_1\boldsymbol{M}_2\boldsymbol{X}_1]^{-1}[\boldsymbol{X}_1'\boldsymbol{M}_2\boldsymbol{y}]\end{aligned}$$

We can use a similar approach to obtain $\hat{\boldsymbol{\beta}}_2$.

The key result is that the least-squares estimator has the algebraic solution:

$$\begin{aligned}\hat{\boldsymbol{\beta}}_1 &= [\boldsymbol{X}_1\boldsymbol{M}_2\boldsymbol{X}_1]^{-1}[\boldsymbol{X}_1'\boldsymbol{M}_2\boldsymbol{y}] \\ \hat{\boldsymbol{\beta}}_2 &= [\boldsymbol{X}_2\boldsymbol{M}_1\boldsymbol{X}_2]^{-1}[\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{y}]\end{aligned}$$

Intuitively, $\boldsymbol{M}_2\boldsymbol{y}$ is the remaining information in $\boldsymbol{y}$ when we "project out" the information about $\boldsymbol{X}_2$, and hence $\boldsymbol{M}_2\boldsymbol{y}$ is orthogonal to $\boldsymbol{X}_2$.
A similar argument holds for $\boldsymbol{M}_1\boldsymbol{y}$.
Hence, a simple regression of $\boldsymbol{M}_2\boldsymbol{y}$ on $\boldsymbol{M}_2\boldsymbol{X}_1$ gives us $\hat{\boldsymbol{\beta}}_1$.
The LS estimators $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ can be found by a two-step regression procedure.
From the discussion above, we have:

$$\begin{aligned}\hat{\boldsymbol{\beta}}_2 &= [\boldsymbol{X}_2\boldsymbol{M}_1\boldsymbol{X}_2]^{-1}[\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{y}] \\ &= (\tilde{\boldsymbol{X}}_2'\tilde{\boldsymbol{X}}_2)^{-1}(\tilde{\boldsymbol{X}}_2'\tilde{\boldsymbol{e}}_1)\end{aligned}$$

where we define

$$\tilde{\boldsymbol{X}}_2' = \boldsymbol{M}_1\boldsymbol{X}_2, \qquad \tilde{\boldsymbol{e}}_1 = \boldsymbol{M}_1\boldsymbol{y}$$

Hence the coefficient estimate $\tilde{\boldsymbol{\beta}}_2$ is algebraically equal to the LS regression of $\tilde{\boldsymbol{e}}_1$ on $\tilde{\boldsymbol{X}}_2$.

This gives us the **Frisch-Waugh-Lovell** theorem:
The OLS estimator of $\boldsymbol{\beta}_2$ and the OLS residuals $\hat{\boldsymbol{e}}$ may be equivalently computed by their the OLS regression or via the following algorithm:

i. Regress $\boldsymbol{y}$ on $\boldsymbol{X}_1$, obtain residuals $\tilde{\boldsymbol{e}}_1 = \boldsymbol{M}_1\boldsymbol{y}$
ii. Regress $\boldsymbol{X}_2$ on $\boldsymbol{X}_1$, obtain residuals $\tilde{\boldsymbol{X}}_2 = \boldsymbol{M}_1\boldsymbol{X}_2$
iii. Regress $\tilde{\boldsymbol{e}}_1$ on $\tilde{\boldsymbol{X}}_2$, obtain OLS estimates $\hat{\boldsymbol{\beta}}_2$ and residuals $\hat{\boldsymbol{e}}$.

Application of FWL theorem:
If we partition $\boldsymbol{X} = [\boldsymbol{X}_1 \ \boldsymbol{X}_2]$, where $\boldsymbol{X}_1 = \boldsymbol{1}_n$ is a vector of ones and $\boldsymbol{X}_2$ is a matrix of observed regressors, then to obtain an algebraic expression for $\hat{\boldsymbol{\beta}}_2$ we will define

$$\boldsymbol{M}_1 = \boldsymbol{I}_n - \boldsymbol{1}_n(\boldsymbol{1}_n'\boldsymbol{1}_n)^{-1}\boldsymbol{1}_n'$$

And hence

$$\begin{aligned}\tilde{\boldsymbol{X}}_2 &= \boldsymbol{M}_1\boldsymbol{X}_2 = \boldsymbol{X}_2 - \bar{\boldsymbol{X}}_2 \\ \boldsymbol{M}_1\boldsymbol{y} &= \boldsymbol{y} - \bar{\boldsymbol{y}}\end{aligned}$$

And hence the slope coefficient $\hat{\boldsymbol{\beta}}_2$ is the same as the OLS estimate from a regression of $y_i - \bar{y}$ on $\boldsymbol{x}_{2i} - \bar{\boldsymbol{x}}_2$, which is the same as what we derived earlier for the demeaned regression.

3. Lecture 3: Least Squares Regression

Note: from this section on, the vectors are no longer bolded for simplicity of notetaking, but the same notation is implied as in the previous sections.

(a) Random Sampling

It is common to describe the iid observations as a **random sample**. Violating this assumption will complicate inferences and require specialised treatment (e.g., clustered dependence, which will be discussed later).

e.g., for the intercept-only model:

$$\begin{aligned} y_i &= \mu + e_i \\ \mathbb{E}(e_i) &= 0 \end{aligned}$$

The projection coefficient is $\mu = \mathbb{E}(y_i)$ and the LS estimator is $\hat{\mu} = \bar{y}$.
The mean of the estimator is

$$\mathbb{E}(\bar{y}) = \mathbb{E}(\frac{1}{n}\sum_{i=1}^{n} y_i) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(y_i) = \mu$$

where the last equality follows from y_i being identically distributed.
The estimator is unbiased (recall that an estimator $\hat{\theta}$ for θ is unbiased if $\mathbb{E}(\hat{\theta}) = \theta$).

The variance of the estimator is

$$\begin{aligned} var(\bar{y}) &= \mathbb{E}(\bar{y} - \mu)^2 \\ &= \mathbb{E}((\frac{1}{n}\sum_{i=1}^{n} e_i)(\frac{1}{n}\sum_{j=1}^{n} e_j)) \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}(e_i e_j) \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 \qquad \text{(by iid assumption)} \\ &= \frac{1}{n}\sigma^2 \end{aligned}$$

(b) Linear Regression Model

Assumption 4.2: Linear Regression Model
The observations (y_i, x_i) satisfy the linear regression equation

$$\begin{aligned} y_i &= x_i'\beta + e_i \qquad \text{(note that } i \text{ denotes observations)} \\ \mathbb{E}(e_i|x_i) &= 0 \end{aligned}$$

The variables have finite second moments

$$\begin{aligned} \mathbb{E}(y_i^2) &< \infty \\ \mathbb{E}||x_i||^2 &< \infty \end{aligned}$$

and an invertible design matrix (which is positive definite):

$$Q_{xx} = \mathbb{E}(x_i x_i') > 0$$

The main parameter of interest is $\beta = \mathbb{E}(x_i x_i')^{-1}\mathbb{E}(x_i y_i)$.

We can write the conditional variance of e as a function of x_i:

$$\mathbb{E}(e_i^2|x_i) = \sigma^2(x_i) = \sigma_i^2$$

This is known as a heteroskedastic regression.
To prove certain results, we impose a strong assumption that the conditional variance is a constant.
Assumption 4.3: Homoskedastic Linear Regression Model
In addition to Assumption 4.2,

$$\mathbb{E}(e_i^2|x_i) = \sigma^2(x_i) = \sigma^2$$

is independent of x_i.

(c) Mean of LS estimator
We first calculate the mean of $\hat{\beta}$ conditional on X, which consists of the entire sample of regressor values.

Importantly, the independent distribution assumption implies that

$$\mathbb{E}(y_i|X) = \mathbb{E}(y_i|x_i) = x_i'\beta$$

That is, the conditional expectation of y_i given $\{x_i, ..., x_n\}$ only depends on x_i.
We then have

$$\mathbb{E}(y|X) = \begin{bmatrix} ... \\ \mathbb{E}(y_i|X) \\ ... \end{bmatrix} = \begin{bmatrix} ... \\ x_i'\beta \\ ... \end{bmatrix} = X\beta$$

We can then show that $\hat{\beta} = (X'X)^-1(X'y)$ is unbiased for β conditional on X:

$$\begin{aligned} \mathbb{E}(\hat{\beta}|X) &= \mathbb{E}((X'X)^{-1}X'y|X) \\ &= (X'X)^{-1}X'\mathbb{E}(y|X) \\ &= (X'X)^{-1}(X'X)\beta = \beta \end{aligned}$$

that is, for any realisation of the regressor matrix X, the conditional distribution of $\hat{\beta}$ is centered at β.
If the regularity condition $\mathbb{E}||\hat{\beta}|| < \infty$ is satisfied, we can apply the LIE to obtain the unconditional variance of $\hat{\beta}$:

$$\mathbb{E}(\hat{\beta}) = \mathbb{E}(\mathbb{E}(\hat{\beta}|X)) = \beta$$

(there is a possibility that the design matrix is not invertible and hence there is an undefined $\hat{\beta}$, but for simplicity we will assert that LIE can be applied and $\mathbb{E}(\hat{\beta}) = \beta$.

(d) Variance of LS estimator
The variance of $\hat{\beta}$ is a $k \times k$ covariance matrix, where is diagonal elements represent the variance of each element in $\hat{\beta}$.
We can use the general formula

$$var(Z|X) = \mathbb{E}((Z - \mathbb{E}(Z|X))(Z - \mathbb{E}(Z|X))'|X)$$

so that

$$\begin{aligned} var(\hat{\beta}|X) &= \mathbb{E}((\hat{\beta} - \beta)(\hat{\beta} - \beta|X) \\ &= (X'X)^{-1}X'DX(X'X)^{-1} \end{aligned}$$

where $D = \mathbb{E}(ee'|X)$
D is a diagonal matrix:

$$D = diag(\sigma_1^2, ...\sigma_n^2) = \begin{pmatrix} \sigma_1^2 & 0 & ... & 0 \\ 0 & \sigma_2^2 & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & \sigma_n^2 \end{pmatrix}$$

In the special case of the homoskecastic regression model, since $\mathbb{E}(e_i^2|x) = \sigma_i^2 = \sigma^2$, we have $D = I_n\sigma^2$.

The variance of the least-squares estimator is thus

$$\begin{aligned} V_{\hat{\beta}} &= var(\hat{\beta}|X) \\ &= (X'X)^{-1}(X'DX)(X'X)^{-1} \end{aligned}$$

In the homoskedastic linear regression model with iid sampling,

$$V_{\hat{\beta}} = \sigma^2(X'X)^{-1}$$

Note: If we generalise the covariance matrix even further, then for $var[e|x] = \Sigma\sigma^2$ (where Σ is any variance-covariance matrix), then we have

$$var[\hat{\beta}|X] = \sigma^2(X'X)^{-1}(X'\Sigma X)(X'X)^{-1}$$

(e) Proof that LS estimator has the lowest variance

Restricting ourselves to the homoskedastic linear regression model, the Gauss-Markov theorem states that among linear unbiased estimators of β, the LS estimator has the lowest variance.
In other words,

$$var[\tilde{\beta}|X] \geq \sigma^2(X'X)^{-1}$$

for the homoskedastic regression model.

We can demonstrate this as follows:
We write a linear estimator (i.e., linear in y) as

$$\tilde{\beta} = A'y$$

where A is a $n \times k$ function of x. We also require that $A'X = I_k$ for $\tilde{\beta}$ to be unbiased.
The mean of $\tilde{\beta}$ is

$$\mathbb{E}(\tilde{\beta}|X) = A'\mathbb{E}(y|X) = A'X\beta$$

The variance of $\tilde{\beta}$ is

$$\begin{aligned}
var(\tilde{\beta}|X) &= \mathbb{E}((\tilde{\beta} - A'X\beta)(\tilde{\beta} - A'X\beta)'|X) \\
&= \mathbb{E}((A'y - A'X\beta)(A'y - A'X\beta)'|X) \\
&= \mathbb{E}(A'e)(A'e)'|x) \\
&= A'E(ee'|X)A \\
&= A'DA
\end{aligned}$$

Therefore, under homoskedasticity, $var(\tilde{\beta}|X) = var(A'y|X) = A'DA = A'A\sigma^2$.

To prove the theorem, we express A as the proposed optimal choice $X(X'X)^{-1}$ plus some deviation C (an $n \times k$ matrix). Then,

$$\begin{aligned}
A'A - (X'X)^{-1} &= (C + X(X'X)^{-1})'(C + X(X'X)^{-1}) - (X'X)^{-1} \\
&= C'C + C'X(X'X)^{-1} + (X'X)^{-1}X'C + (X'X)^{-1}X'X(X'X)^{-1} - (X'X)^{-1} \\
&= C'C
\end{aligned}$$

The matrix $C'C$ is positive semidefinite since it is a quadratic form, so we have shown the result as required.

(f) Residuals
The conditional mean of the residuals is

$$\mathbb{E}(\hat{e}|X) = \mathbb{E}(Me|X) = M\mathbb{E}(e|X) = 0$$

The variance of the residuals is

$$var(\hat{e}|X) = var(Me|X) = Mvar(e|X)M = MDM$$

Note: The last equality is because for any random vector t and non-random matrix A, $var(Ay) = Avar(y)A'$.

Under homoskedasticity,

$$var(\hat{e}|x) = M\sigma^2$$

(g) Estimation of error variance
Consider the error variance parameter $\sigma^2 - \mathbb{E}(e_i^2) = \mathbb{E}(\mathbb{E}(e_i^2|x_i))$ A feasible moment estimator is

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\hat{e}_i^2$$

The mean of this estimator is

$$\begin{aligned}\mathbb{E}(\hat{\sigma}^2|X) &= \frac{1}{n}tr(\mathbb{E}(Mee'|x) \qquad \text{(since the trace of a scalar is the scalar itself)}\\ &= \frac{1}{n}tr(M\mathbb{E}(ee'|x))\\ &= \frac{1}{n}tr(M\sigma^2)\\ &= \sigma^2(\frac{n-k}{n})\end{aligned}$$

and hence $\hat{\sigma}^2$ is biased toward zero.
A classic method to obtain an unbiased estimator is to rescale the estimator, i.e., we define the "biased-corrected estimator"

$$s^2 = \frac{1}{n-k}\sum_{i=1}^{n}\hat{e}_i^2$$

So that we have

$$\begin{aligned}\mathbb{E}(s^2|X) &= \sigma^2\\ \mathbb{E}(s^2) &= \sigma^2\end{aligned}$$

(h) Covariance matrix estimation under homoskedasticity
Recall that the conditional variance of $\hat{\beta}$ under homoskedasticity is

$$V_{\hat{\beta}}^0 = (X'X)^{-1}\sigma^2$$

where

$$\sigma^2 = \mathbb{E}(e_i^2)$$

(we use superscript 0 to emphasise that this is the variance under homoskedasticity).
An estimator of this variance is

$$V_{\hat{\beta}}^0 = (X'X)^{-1}s^2$$

which is conditionally unbiased for V_{β}^0.

(i) Covariance matrix under heteroskedasticity
Under heteroskedasticity, the conditional variance of $\hat{\beta}$ is

$$V_{\hat{\beta}} = (X'X)^{-1}(X'DX)(X'X)^{-1}$$

where $D = \mathbb{E}(ee'|X)$.

We have several possible estimators:

i. HC0 estimator

$$\hat{V}_{\hat{\beta}}^{HC0} = (X'X)^{-1}(\sum_{i=1}^{n}x_i x_i'\hat{e}_i^2)(X'X)^{-1}$$

This estimator is biased toward zero because $\mathbb{E}(\hat{e}_i^2|x_i) = \frac{n-k}{n}\sigma_i^2$.

ii. HC1 estimator

$$\hat{V}_{\hat{\beta}}^{HC1} = (\frac{n}{n-k})(X'X)^{-1}(\sum_{i=1}^{n}x_i x_i'\hat{e}_i^2)(X'X)^{-1}$$

This estimator corrects for the bias in the HC0 estimator.
This is an ad hoc estimator as we can't really show that it is unbiased (for the homoskedastic case, we could use the conditioning theorem to take out $I\sigma^2$, but this is not possible in the heteroskedastic case).
Most empirical work uses the HC1 estimator when researchers suspect that heteroskedasticity is present.

iii. HC2 estimator

$$\hat{V}_{\hat{\beta}}^{HC2} = (X'X)^{-1}(\sum_{i=1}^{n} x_i x_i' \bar{e}_i^2)(X'X)^{-1}$$
$$= (X'X)^{-1}())\sum_{i=1}^{n}(1-h_{ii})^{-1} x_i x_i' \hat{e}_i^2)(X'X)^{-1}$$

where $h_i i$ are the "leverage values": the diagonal elements of $P = X(X'X)^{-1}X'$.
This estimator does the scaling within the sum rather than outside the sum.

iv. HC3 estimator

$$\hat{V}_{\hat{\beta}}^{HC3} = (X'X)^{-1}(\sum_{i=1}^{n} x_i x_i' \tilde{e}_i^2)(X'X)^{-1}$$
$$= (X'X)^{-1}(\sum_{i=1}^{n}(1-h_{ii})^{-2} x_i x_i' \hat{e}_i^2)(X'X)^{-1}$$

Similarly, this estimator does the scaling within the sum rather than outside the sum.

We can show that, in a positive definite sense,

$$\hat{V}_{\hat{\beta}}^{HC0} < \hat{V}_{\hat{\beta}}^{HC2} < \hat{V}_{\hat{\beta}}^{HC3}$$

And, under homoskedasticity, the HC2 estimator is unbiased.
And the HC3 estimator is conservative in the sense that it is weakly larger (in expectation) than the correct variance of $\hat{\beta}$.
In practice, HC1 is the norm in empirical work.

(j) **Standard errors**
A standard error $s(\hat{\beta})$ for a real-valued estimator $\hat{\beta}$ is an estimator of the standard deviation of the distribution of $\hat{\beta}$.
We take the diagonal elements of the covariance estimator, and obtain:

$$s(\hat{\beta}_j) = \sqrt{\hat{V}_{\hat{\beta}_j}} = \sqrt{[\hat{V}_{\hat{\beta}}]_{jj}}$$

Under homoskedasticity, we have

$$s(\hat{\beta}_j) = \sqrt{[(X'X)^{-1}]_{jj}}$$

(k) **Clustered sampling**

In clustered contexts, it is convenient to double-index the observations as (y_{ig}, x_{ig}) where $g = 1, ..., G$ indexes the cluster and $i = 1, ..., n_g$ indexes the individual within the g^{th} cluster.
The total number of observations per cluster (n_g) may differ across clusters.
There are G clusters and $n = \sum_{g=1}^{G} n_g$ observations in total.

An alternative to the double index notation is cluster-level notation.
Let $y_g = (y_{1_g}, ..., y_{n_g g})$ and $X_g = (x_{1_g}, ..., x_{n_g g})'$ denote the $n_g \times 1$ vector of dependent variables and $n_g \times k$ matrix of regressors for the g^{th} cluster.
Then, a linear regression model can be written down for the individual observations as

$$y_{ig} = x_{ig}'\beta + e_{ig}$$

and using cluster notation as

$$y_g = X_g\beta + e_g$$

where $e_g = (e_{1_g}, ..., e_{n_g g})$ is a $n_g \times 1$ error vector. We can also stack the observations into full sample matrices and write the model as

$$y = X\beta + e$$

The OLS estimator can be written as

$$\begin{aligned}\hat{\beta} &= (\sum_{g=1}^{G}\sum_{i=1}^{n_g} x_{ig}x'_{ig})^{-1}(\sum_{g=1}^{G}\sum_{i=1}^{n_g} x_{ig}y_{ig})\\ &= (\sum_{g=1}^{G} X'_gX_g)^{-1}(\sum_{g=1}^{G} X'_gy_g)\\ &= (X'X)^{-1}(X'y)\end{aligned}$$

The OLS residuals are $\hat{e}_{ig} = y_{ig} - x'_{ig}\hat{\beta}$ in individual-level notation and $\hat{e}_g = y_g - X_g\hat{\beta}$ in cluster level notation.
The standard clustering assumption is that the clusters are known to the researcher, and that the observations are independent across clusters, i.e., the clusters (y_g, X_g) are mutually independent across clusters g.

The model is a linear regression under the assumption

$$\mathbb{E}(e_g|X_g) = 0$$

or expressed as

$$\mathbb{E}(e_{ig}|X_g) = 0$$

for $i = 1, ..., n_g$.
Note that we can't shrink the conditioning set any further (e.g., to $\mathbb{E}(e_{ig}|X_{jg})$) because observations within a cluster are mutually dependent.
This assumption allows for observations within a cluster to be mutually dependent, but we still impose a zero-conditional-mean assumption of the error conditional on all regressor values of all observations within a cluster.

Under the assumption above, we can show that the OLS estimator is conditionally unbiased under the clustered linear regression model.
Specifically, we have:

$$\hat{\beta} - \beta = (\sum_{g=1}^{G} X'_gX_g)^{-1}(\sum_{g=1}^{G} X'_ge_g)$$

Taking conditional expectations, we have:

$$\begin{aligned}\mathbb{E}(\hat{\beta} - \beta|X) &= (\sum_{g=1}^{G} X'_gX_g)^{-1}(\sum_{g=1}^{G} X'_g\mathbb{E}(e_g|X))\\ &= (\sum_{g=1}^{G} X'_gX_g)^{-1}(\sum_{g=1}^{G} X'_g\mathbb{E}(e_g|X_g)) \qquad \text{(because of independence across clusters)}\\ &= 0\end{aligned}$$

To find the covariance matrix of $\hat{\beta}$, we let

$$\Sigma_g = \mathbb{E}(e_ge'_g|X_g)$$

which is a $n_g \times n_g$ covariance matrix of errors within the g^{th} cluster.
Since the observations are independent across clusters,

$$\begin{aligned}var((\sum_{g=1}^{G} X'_ge_g)|X) &= \sum_{g=1}^{G} var(X'_ge_g|X_g)\\ &= \sum_{g=1}^{G} X'_g\mathbb{E}(e_ge'_g|X_g)X_g\\ &= \sum_{g=1}^{G} X'_g\Sigma_gX_g\\ &\overset{\text{def}}{=} \Omega_n \qquad \text{(this is a } k \times k \text{ matrix)}\end{aligned}$$

It follows that $V_{\hat{\beta}} = var(\hat{\beta}|X) = (X'X)^{-1}\Omega_n(X'X)^{-1}$, which is a $n \times n$ matrix.

In a special case where

i. All clusters have the same number of observations $n_g = N$
ii. There is homoskedasticity, i.e., $\mathbb{E}(e_{ig}^2|x_g) = \sigma^2$
iii. There is the same covariance between different observations within a cluster, i.e., $\mathbb{E}(e_{ig}e_{lg}|x_g) = \sigma^2\rho$ for $i \neq l$
iv. Regressors x_{ig} do not vary within a cluster

Then the exact variance of the OLS estimator is

$$V_{\hat{\beta}} = (X'X)^{-1}\sigma^2(1 + \rho(N-1))$$

Note that if $\rho > 0$, this shows that the actual variance is appropriately a multiple ρN of the conventional formula with homoskedastic errors and no clustering.

To estimate $V_{\hat{\beta}}$ allowing for general correlation within clusters, we extend the robust White formula. Arellano (1987) uses this feasible estimator for Σ_g:

$$\begin{aligned}\hat{\Omega}_n &= \sum_{g=1}^{G} X_g'\hat{e}_g\hat{e}_g'X_g \\ &= \sum_{g=1}^{G}\sum_{i=1}^{n_g}\sum_{l=1}^{n_g} x_{ig}x_{lg}'\hat{e}_{ig}\hat{e}_{lg} \\ &= \sum_{g=1}^{G}(\sum_{i=1}^{n_g} x_{ig}\hat{e}_{ig})(\sum_{i=1}^{n_g} x_{ig}\hat{e}_{lg})'\end{aligned}$$

A natural cluster covariance matrix is thus

$$\hat{V}_{\hat{\beta}} = a_n(X'X)^{-1}\hat{\Omega}_n(X'X)^{-1}$$

where a_n is a possible finite-sample adjustment. The Stata cluster command uses

$$a_n = (\frac{n-1}{n-k})(\frac{G}{G-1})$$

This is an ad hoc generalisation which nests the adjustment in the HC1 estimator.
when $G = n$ (no cluster dependence), we have the HC1 estimator $a_n = \frac{n}{n-k}$.

General notes on clustering:

i. The effective sample size should be viewed as the number of clusters G as they are analogous to observations in an iid sample.
ii. When cluster sizes are highly heterogenous, the regression should be viewed as roughly equivalent to the heteroskedasticity-robust case with an extremely high degree of heteroskedasticity.
iii. Clustering may reduce bias (in the variance estimators), but over-clustering may also add noise and contribute to increased standard errors.

4. **Lecture 4: Normal Regression Model**

(a) Summary of commonly used distributions:

Normal distribution:
A random variable X has the **standard normal distribution**, or Gaussian distribution, written $X \sim N(0,1)$, if it has the density

$$\phi(x) = \frac{1}{\sqrt{2\pi}}exp(-\frac{x^2}{2}), \qquad -\infty < x < \infty$$

If a variable $X \sim N(0,1)$, then:

i. All integer moments of X are finite
ii. All odd moments of $X = 0$
iii. For any positive integer m, $\mathbb{E}[X^{2m}] = (2m-1)!! = (2m-1) \times (2m-3)... \times 1$
iv. For any $r > 0$,

$$\mathbb{E}|X|^r = \frac{2^{r/2}}{\sqrt{\pi}}\Gamma(\frac{r+1}{2})$$

where $\Gamma(t) = \int_0^\infty u^{t-1}e^{-u}$ is the gamma function.

A random variable X has a **univariate normal distribution**, written $X \sim N(\mu, sigma^2)$, for $\mu \in \mathbb{R}$ and $\sigma^2 > 0$, if it has the density

$$f(X) = \frac{1}{\sqrt{2\pi\sigma^2}} exp(-\frac{(x-\mu)^2}{2\sigma^2}, \qquad -\infty < x < \infty$$

The mean and variance of X are μ and σ^2.

The k-vector $\boldsymbol{X}$ has a **multivariate standard normal distribution**, written $\boldsymbol{X} \sim (\boldsymbol{0}, \boldsymbol{I}_k)$, if it has the joint density

$$f(x) = \frac{1}{(2\pi)^{k/2}} exp(-\frac{\boldsymbol{x}'\boldsymbol{x}}{2}), \qquad \boldsymbol{x} \in \mathbb{R}^k$$

The mean and covariance matrix of $\boldsymbol{X}$ are $\boldsymbol{0}$ and $\boldsymbol{I}$ respectively. The elements of $\boldsymbol{X}$ are mutually independent standard normal random variables.

The k-vector $\boldsymbol{X}$ has a **multivariate normal distribution**, written $\boldsymbol{X} \sim (\boldsymbol{\mu}, \boldsymbol{\Sigma})$, if it has the joint density

$$f(x) = \frac{1}{(2\pi)^{k/2} det(\boldsymbol{\Sigma})^{1/2}} exp(-\frac{(\boldsymbol{x}' - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{X} - \boldsymbol{\mu})}{2}), \qquad \boldsymbol{x} \in \mathbb{R}^k$$

The mean and covariance matrix of $\boldsymbol{X}$ are $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ respectively. The elements of $\boldsymbol{X}$ are mutually independent normal random variables.

Affine functions of normal random vectors are also multivariate normal.
Theorem 5.4: If $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\boldsymbol{Y} = \boldsymbol{a} + \boldsymbol{BX}$, then $\boldsymbol{Y} \sim N(\boldsymbol{a} + \boldsymbol{B\mu}, \boldsymbol{B\Sigma B'})$.
If bX is multivariate normal, then each component of $\boldsymbol{X}$ is univariate normal.
Theorem 5.5: If $\boldsymbol{X} = (\boldsymbol{X}_1', \boldsymbol{X}_2')'$ is multivariate normal, $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are uncorrelated iff they are independent.

Chi-square distribution
Let $\boldsymbol{X} \sim N(\boldsymbol{0}, \boldsymbol{I}_r)$ be multivariate standard normal and define $\boldsymbol{Q} = \boldsymbol{X}'\boldsymbol{X}$. Then the distribution of $\boldsymbol{Q}$ is **chi-square** with r degrees of freedom, written as $\boldsymbol{Q} \sim \chi^2_r$.
The mean and variance of $\boldsymbol{Q} \sim \chi^2_r$ are r and $2r$ respectively.
The density of χ^2_r is

$$f(x) = \frac{1}{2^{r/2}\Gamma(\frac{r}{2})} x^{r/2-1} e^{-x/2}, \qquad x > 0$$

t-distribution
Let $Z \sim N(0,1)$ and $Q \sim \chi^2_r$ be independent, and define $T = Z/\sqrt{Q/r}$.
The distribution of T is the **student t** with r degrees of freedom, and is written $T \sim t_r$.
The t-distribution has thicker tails than the standard normal distribution due to Q.
As $r \to \infty$, the t-distribution converges to the standard normal distribution.
If $Z \sim N(0,1)$ and $Q \sim \chi^2_k$ are independent, then $Z/\sqrt{Q/k} \sim t_k$
The density of T is

$$f(x) = \frac{\Gamma(\frac{r+1}{2})}{\sqrt{r\pi}\Gamma(\frac{r}{n})}(1 + \frac{x^2}{r})^{-(\frac{r+1}{2})}, \qquad \infty < x < \infty$$

F-distribution Let $Q_m \sim \chi^2_m$ and $Q_r \sim \chi^2_r$ be independent. The distribution of $F = (Q_m/m)/(Q_r/r)$ is the **F distribution**, and $F \sim F_{m,r}$.
The density of F is

$$f(x) = \frac{(\frac{m}{r})^{m/2} x^{m/2-1} \Gamma(\frac{m+r}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{r}{2})(1+\frac{m}{r}x)^{(m+r)/2}}, \qquad x > 0$$

If $m = 1$, then the F distribution equals the squared student t distribution.
As $r \to \infty$, the F distribution simplifies to $F \to Q_m/m$, a normalised χ^2/m.

(b) Asymptotic theory

A sequence a_n has the **limit** a, if for all $\delta > 0$ there is some $n_\delta < \infty$ such that for all $n \geq n_\delta$, $|a_n - a| \leq \delta$.

A random variable $z_n \in \mathbb{R}$ **converges in probability** to z as $n \to \infty$, denoted $z_n \xrightarrow{\mathrm{p}} z$, or $plim_{n\to\infty} z_n = z$, if for all $\delta > 0$,

$$\lim_{n\to\infty} \mathbb{P}(|z_n - z| \leq \delta) = 1$$

Weak law of large numbers:
If y_i are independent and identically distributed and $\mathbb{E}|y| < \infty$, then as $n \to \infty$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \xrightarrow{p} \mathbb{E}(y)$$

An estimator $\hat{\theta}$ of a parameter θ is **consistent** if $\hat{\theta} \xrightarrow{p} \theta$.

If y_i are independent and identically distributed and $\mathbb{E}|y| < \infty$, then $\hat{\mu} = \bar{y}$ is consistent for the population mean μ.

Let z_n be a random vector with distribution $F_n(u) = \mathbb{P}(z_n \leq u)$. z_n **converges in distribution** to z as $n \to \infty$, denoted $z_n \xrightarrow{d} z$, if for all u at which $F(u) = \mathbb{P}(z \leq u)$ is continuous, $F_n(u) \to F(u)$ as $n \to \infty$.
Under these conditions, it is also said that F_n **converges weakly** to F. It is common to refer to z and its distribution $F(u)$ as the **asymptotic distribution** of z_n.

Theorem 6.11 Lindeberg-Levy Central Limit Theorem: If y_i are iid and $\mathbb{E}(y_i^2) < \infty$, then as $n \to \infty$

$$\sqrt{n}(\bar{y} - \mu) \xrightarrow{d} N(0, \sigma^2)$$

where $\mu = \mathbb{E}(y)$ and $\sigma^2 = \mathbb{E}(y_i - \mu)^2$.

Moments of transformations
If we want to estimate a parameter μ which is the expected value of a transformation of a random vector y, i.e.:

$$\mu = \mathbb{E}(h(y))$$

We can define the random variable $z = h(y)$, then $\mu = \mathbb{E}(z)$ is just a simple moment of z. This suggests the moment estimator

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} z_i = \frac{1}{n}\sum_{i=1}^{n} h(y_i)$$

Since $\hat{\mu}$ is a sample average, and transformations of iid variables are also iid, the asymptotic results of the previous sections apply.
If y_i are iid, $\mu = \mathbb{E}(h(y))$ and $\mathbb{E}||h(y)|| < \infty$, then for $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} h(y_i)$, as $n \to \infty$, $\hat{\mu} \xrightarrow{p} \mu$.

Continuous Mapping Theorem: If $z_n \xrightarrow{p} c$ as $n \to \infty$ and $g(.)$ is continuous at c, then $g(z_n) \xrightarrow{p} g(c)$ as $n \to \infty$.
If y_i are iid, $\theta = g(\mathbb{E}(h(y))), \mathbb{E}||h(y)|| < \infty$ and $g(u)$ is continuous at $u = \mu$, for $\hat{\theta} = g(\hat{\mu})$ with $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} h(y_i)$, then $\hat{\theta} \xrightarrow{p} \theta$ as $n \to \infty$.
If $z_n \xrightarrow{d} z$ as $n \to \infty$ and $g : \mathbb{R}^m \to \mathbb{R}^k$ has the set of discontinuity points D_g such that $\mathbb{P}(z \in D_g) = 0$, then $g(z_n) \xrightarrow{d} g(z)$ as $n \to \infty$.

A special case of the CMT is **Slutsky's theorem**:
If $z_n \xrightarrow{d} z$ and $c_n \xrightarrow{p} c$ as $n \to \infty$, then

i. $z_n + c_n \xrightarrow{d} z + c$
ii. $z_n c_n \xrightarrow{d} zc$
iii. $\frac{z_n}{c_n} \xrightarrow{d} \frac{z}{c}$ if $c \neq 0$

Delta Method:
Theorem 6.23: If $\sqrt{n}(\hat{\mu} - \mu) \xrightarrow{d} \xi$, where $g(u)$ is continuously differentiable in a neighbourhood of μ, then as $n \to \infty$

$$\sqrt{n}(g(\hat{\mu}) - g(\mu)) \xrightarrow{d} G'\xi$$

where $G(u) = \frac{\partial}{\partial u} g(u)'$ and $G = G(\mu)$. In particular, if $\xi \sim N(0, V)$ then as $n \to \infty$

$$\sqrt{n}(g(\hat{\mu}) - g(\mu)) \xrightarrow{d} N(0, G'VG)$$

Theorem 6.24: If y_i are iid, $\mu = \mathbb{E}(h(y)), \theta = g(\mu), \mathbb{E}||h(y)||^2 < \infty$ and $G(u) = \frac{\partial}{\partial u} g(u)'$ is continuous in a neighbourhood of μ, for $\hat{\theta} = g(\hat{\mu})$ with $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} h(y_i)$, then as $n \to \infty$

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, V_\theta)$$

where $V_\theta = G'VG$, $V = \mathbb{E}((h(y) - \mu)(h(y) - \mu)')$ and $G = G(\mu)$.

Covariance matrix estimation
To use Theorem 6.24, we need an estimator of the asymptotic variance matrix $V_\theta = G'VG$. The natural plug-in estimator is

$$\begin{aligned}\hat{V}_\theta &= \hat{G}'\hat{V}\hat{G} \\ \hat{G} &= G(\hat{\mu} \\ \hat{V} &= \frac{1}{n}\sum_{i=1}^{n}(h(y_i) - \hat{\mu})(h(y_i) - \hat{\mu})'\end{aligned}$$

Under the assumptions of Theorem 6.24, the WLLN implies $\hat{\mu} \xrightarrow{p} \mu$, $\hat{V} \xrightarrow{p} V$. The CMT implies $\hat{G} \xrightarrow{p} G$ and $\hat{V}_\theta = \hat{G}'\hat{V}\hat{G} \xrightarrow{p} G'VG = V_\theta$. We have established that $\hat{V}_\theta$ is consistent for V_θ.
Theorem 6.25: Under the assumptions of Theorem 6.24, $\hat{V}_\theta \xrightarrow{p} V_\theta$ as $n \to \infty$.

(c) Normal regression model
The normal regression model is the linear regression model (with linear CEF) with an independent normal error:

$$\begin{aligned} y &= x'\beta + e \\ e &\sim N(0, \sigma^2)\end{aligned}$$

This means that the conditional distribution of y given x is normal, because it is a linear function of e.
Also, because x and e are independent, the errors are homoskedastic, i.e.:

$$\mathbb{E}(e^2|x) = \mathbb{E}(e^2) = \sigma^2$$

Normal regression is a parametric model. The **likelihood** is the joint probability density of the data, evaluated at the observed sample, and viewed as a function of the parameters.

The **maximum likelihood estimator** is the value which maximises the likelihood function.

The normal regression model is equivalent to the statement that the conditional density of y on x takes the form

$$f(y|x) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} exp(-\frac{1}{2\sigma^2}(y - x'\beta)^2)$$

Under the assumption that the observations are mutually independent, this implies that the conditional density of $(y_1, ...y_n)$ given $(x_1, ...x_n)$ is

$$\begin{aligned} f(y_1, ..., y_n|x_1, ..., x_n) &= \Pi_{i=1}^n f(y_i|x_i) \\ &= \Pi_{i=1}^n \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} exp(-\frac{1}{2\sigma^2}(y_i - x_i'\beta)^2) \\ &= \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} exp(-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - x_i'\beta)^2) \\ &\overset{\text{def}}{=} L(\beta, \sigma^2) \end{aligned}$$

and is called the **likelihood function**.
For convenience, we work with the **log-likelihood function**:

$$\begin{aligned} log\ f(y_1, ..., y_n|x_1, ..., x_n) &= \frac{n}{2}log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - x_i'\beta)^2 \\ &\overset{\text{def}}{=} logL(\beta, \sigma^2) \end{aligned}$$

The **maximum likelihood estimator** $(\hat{\beta}_{mle}, \hat{\sigma}^2_{mle})$ is the value which maximises the log-likelihood. In other words,

$$(\hat{\beta}_{mle}, \hat{\sigma}^2_{mle}) = argmax_{\beta \in \mathbb{R}^k, \sigma^2 > 0} logL(\beta, sigma^2)$$

In the normal regression model, we can explicitly solve for $(\hat{\beta}_{mle}, \hat{\sigma}^2_{mle})$.
The maximisers $(\hat{\beta}_{mle}, \hat{\sigma}^2_{mle})$ jointly solve the FOC

$$\begin{aligned} 0 = \frac{\partial}{\partial\beta} logL(\beta_{mle}, \sigma^2)|_{\beta=\hat{\beta}_{mle}, \sigma^2=\hat{\sigma}^2_{mle}} &= \frac{1}{\hat{\sigma}^2_{mle}}\sum_{i=1}^n x_i(y_i - x_i'\hat{\beta}_{mle}) \\ 0 = \frac{\partial}{\partial\sigma^2} logL(\beta_{mle}, \sigma^2)|_{\beta=\hat{\beta}_{mle}, \sigma^2=\hat{\sigma}^2_{mle}} &= -\frac{n}{2\hat{\sigma}^2_{mle}} + \frac{1}{\hat{\sigma}^4_{mle}}\sum_{i=1}^n x_i(y_i - x_i'\hat{\beta}_{mle})^2 \end{aligned}$$

Solving the FOCs, we have

$$\begin{aligned} \hat{\beta}_{mle} &= (\sum_{i=1}^n x_i x_i')^{-1}(\sum_{i=1}^n x_i y_i) = \hat{\beta}_{ols} \\ \hat{\sigma}^2_{mle} &= \frac{1}{n}\sum_{i=1}^n (y_i - x_i'\hat{\beta}_{mle})^2 = \frac{1}{n}\sum_{i=1}^n (y_i - x_i'\hat{\beta}_{ols})^2 = \frac{1}{n}\sum_{i=1}^n \hat{e}^2 = \hat{\sigma}^2_{ols} \end{aligned}$$

The MLE for β and σ^2 are algebraically identical to the OLS estimator (for σ^2, it is identical to the moment estimator that is not bias-corrected).
Note that $\hat{\beta}$ is only the MLE when the error e has a normal distribution, and not otherwise. So in this case, $\hat{\beta}$ is both the minimiser of OLS errors and the maximiser of the likelihood function.

Plugging the estimators into the maximisation problem, we obtain the maximised log-likelihood

$$logL(\hat{\beta}_{mle}, \hat{\sigma}^2_{mle}) = -\frac{n}{2}log(2\pi\hat{\sigma}^2_{mle}) - \frac{n}{2}$$

The log-likelihood is typically reported as a measure of fit.

(d) Distribution of coefficient vector
The normality assumption $e_i|x_i \sim N(0, \sigma^2)$ combined with independence of the observations imply

$$e|X \sim N(0, I_n\sigma^2)$$

Recall that the OLS estimator satisfies

$$\hat{\beta} - \beta = (X'X)^{-1}X'e$$

which is a linear function of e.

Therefore, conditional on X, we can apply the properties of the normal distribution to obtain

$$\begin{aligned} \hat{\beta} - \beta|_X &\sim (X'X)^{-1}X'\ N(0, I_n\sigma^2) \\ &\sim N(0, \sigma^2(X'X)^{-1}X'X(X'X)^{-1}) \\ &= N(0, \sigma^2(X'X)^{-1}) \end{aligned}$$

Alternatively, we can say that the OLS estimator has an exact normal distribution:

$$\hat{\beta}|_X \sim N(\beta, \sigma^2(X'X)^{-1})$$

It follows that each element of $\hat{\beta}$ is univariate normal:

$$\hat{\beta}_j|_X \sim N(\beta_j, \sigma^2[X'X)^{-1}]_{jj}) \tag{Eq 5.10}$$

(e) Distribution of residual vector
Recall that $\hat{e} = Me$. This shows that $\hat{e}$ is a linear function of e.
Therefore, conditional on X, we can apply the properties of the normal distribution to obtain

$$\hat{e} = Me|_x \sim N(0, \sigma^2 MM) = N(0, \sigma^2 M)$$

Furthermore, it is useful to understand the joint distribution of $\hat{\beta}$ and $\hat{e}$. We can stack them (as a $k + n \times 1$ vector):

$$\begin{pmatrix} \hat{\beta} - \beta \\ \hat{e} \end{pmatrix} = \begin{pmatrix} (X'X)^{-1}X'e \\ Me \end{pmatrix} = \begin{pmatrix} (X'X)^{-1}X' \\ M \end{pmatrix} e$$

This matrix has a multivariate normal distribution with coveriance matrix

$$\begin{aligned} &\begin{bmatrix} (X'X)^{-1}X' \\ M \end{bmatrix} I_n\sigma^2 \begin{bmatrix} X(X'X)^{-1} & M \end{bmatrix} \\ &= \begin{bmatrix} \sigma^2(X'X)^{-1}X'X(X'X)^{-1} & \sigma^2(X'X)^{-1}X'M \\ \sigma^2 MX(X'X)^{-1} & M\sigma^2 M \end{bmatrix} \\ &= \begin{bmatrix} \sigma^2(X'X)^{-1} & 0 \\ 0 & \sigma^2 M \end{bmatrix} \end{aligned}$$

because $X'M = 0$.
Since the off-diagonal block is zero, $\hat{\beta}$ and $\hat{e}$ are uncorrelated.
We also know that they are normally distributed, and hence it follows from the property of the multivariate normal distribution that $\hat{\beta}$ and $\hat{e}$ are statistically independent.
Specficially, in the linear regression model, $\hat{e}|_X \sim N(0, \sigma^2 M)$ and is independent of $\hat{\beta}$.

(f) Distribution of variance estimator
Consider the variance estimator

$$s^2 = \frac{1}{n-k}\sum_{i=1}^{n} \hat{e}_i^2$$

We know that $(n-k)s^2 = \hat{e}'\hat{e} = e'Me$.
To derive its distribution we need to do **spectral decomposition** of M.

Spectral Decomposition: If A is a $k \times k$ real symmetric matrix, then $A = H\Lambda H'$ where H contains the eigenvectors and Λ is a diagonal matrix with the eignevalues on the diagonal. The eigenvalues aere all real and the eigenvector matrix satisfies $H'H = I_k$.
Since M has $n-k$ eigenvalues with value 1 and k eigenvalues with value 0, we can write

$$\Lambda = \begin{bmatrix} I_{n-k}^{0} \\ 0_{k}^{0} \end{bmatrix}$$

We can then derive the following:

$$\begin{aligned}
(n-k)s^2 &= e'Me \\
&= e'H \begin{bmatrix} I_{n-k} & 0 \\ 0 & 0 \end{bmatrix} H'e \\
&= u' \begin{bmatrix} I_{n-k} & 0 \\ 0 & 0 \end{bmatrix} u \\
&= u_1'u_1
\end{aligned}$$

where

i. $u = H'e \sim N(0, I_n\sigma^2)$

ii. We partition $u = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$ where $u_1 \sim N(0, I_{n-k}\sigma^2)$, then $u'u = u_1^2 + u_2^2$.

Note that:

$$\begin{aligned}
\mathbb{E}[H'e] &= \mathbb{E}[\mathbb{E}[H'e|X]] \\
&= H'\mathbb{E}[\mathbb{E}[e|x]] \\
&= 0
\end{aligned}$$

and

$$\begin{aligned}
var[H'e] &= \mathbb{E}[(H'e - \mathbb{E}[H'e])^2] \\
&= \mathbb{E}[eHH'e] \\
&= I_n\sigma^2
\end{aligned}$$

which is normally distributed since it is a affine transformation of e which is normally distributed.

Now, we have

$$\frac{(n-k)}{\sigma^2}s^2 = (\frac{1}{\sigma}u_1')(\frac{1}{\sigma}u_1)$$

This is the sum of $n-k$ independent squared standard normal random variables with variance 1 mean 0.

Since the sum of squared mutually independent standard normal distributions is chi-square distributed, we have that in the linear regression model,

$$\frac{(n-k)s^2}{\sigma^2} \sim \chi^2_{n-k}$$

and is independent of $\hat{\beta}$.
Recall that we showed that

$$s^2 = \frac{1}{n-k}\sum_{i=1}^{n}\hat{e}_i^2$$

which is the sum of n squared residuals, can be re-expressed as a sum of $n-k$ independent squared normal variables.
We can also verify this by computing

$$\mathbb{E}(\frac{n-k}{\sigma^2}s^2) = \frac{n-k}{\sigma^2}\mathbb{E}(s^2) = \frac{n-k}{\sigma^2}\sigma^2 = n-k$$

which is the mean of the χ^2_{n-k} distribution.

(g) t-statistic

From (Eq 5.10) above, we have

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\sigma^2[(X'X)^{-1}]_{jj}}} \sim N(0,1)$$

This is sometimes called a standardised statistic, as the distribution is the standard normal.
Now, replace the unknown variance σ^2 with its estimator s^2. The **t-statistic** or t-ratio is:

$$T = \frac{\hat{\beta}_j - \beta_j}{\sqrt{s^2[(X'X)^{-1}]_{jj}}} = \frac{\hat{\beta} - \beta_j}{s(\hat{\beta}_j)}$$

where $s(\hat{\beta}_j)$ is the classical homoskedastic standard error for $\hat{\beta}_j$.

By algebraic rescaling we can write the t-statistic as:

$$\begin{aligned} T &= \frac{\hat{\beta}_j - \beta_j}{\sqrt{\sigma^2[(X'X)^{-1}]_{jj}}} / \sqrt{\frac{(n-k)s^2}{\sigma^2}/(n-k} \\ &\sim \frac{N(0,1)}{\sqrt{\chi^2_{n-k}/(n-k)}} \\ &\sim t_{n-k} \end{aligned}$$

Since we know that s^2 is independent of $\hat{\beta}$, and a standard normal distribution divided by the square root of a chi-square distribution divided by its dof is t-distributed as long as both distributions are independent, then the t-statistic follows a t-distribution with $n-k$ degrees of freedom.
The t-ratio is **pivotal**, meaning that it does not depend on unknowns (since $n-k$ is known).

(h) Likelihood ratio test

We partition the regression model as

$$y_i = x'_{1i}\beta_1 + x'_{2i}\beta_2 + e_i$$

where $dim(x_{1i}) = k_1, dim(x_{2i}) = q, k = k_1 + q$.

Suppose the null hypothesis is

$$H_0 : \beta_2 = 0$$

If the null is true, we can write the regression model as

$$y_i = x'_{1i}\beta_1 + e_i$$

which we can also call the null (or constrained) model.
The alternative is that at least one element of β_2 (recall that β_2 is a vector) is non-zero, and is written

$$H_1 : \beta_2 \neq 0$$

The likelihood ratio is the ratio of the maximised likelihood function under H_1 and H_0 respectively.
The unconstrained maximised log-likelihood is

$$logL(\hat{\beta}, \hat{\sigma}^2) = -\frac{n}{2}log(2\pi\hat{\sigma}^2) - \frac{n}{2}$$

In the constrained model, the MLE is

$$\tilde{\beta}_1 = (X'_1X_1)^{-1}X'_1y$$

with residual

$$\tilde{e}_i = y_i - x'_{1i}\tilde{\beta}_1$$

and error variance estimate

$$\tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\tilde{e}_i^2$$

where the tildes refer to the constrained estimates.
The constrained maximised log-likelihood is

$$logL(\tilde{\beta},\tilde{\sigma}^2) = -\frac{n}{2}log(2\pi\tilde{\sigma}^2) - \frac{n}{2}$$

The likelihood ratio is the difference between the two log likelihoods:

$$\begin{aligned} LR &= 2((-\frac{n}{2}log(2\pi\hat{\sigma}^2) - \frac{n}{2}) - (-\frac{n}{2}log(2\pi\tilde{\sigma}^2) - \frac{n}{2})) \\ &= nlog(\frac{\tilde{\sigma}^2}{\hat{\sigma}^2}) \end{aligned}$$

The LR test rejects for large values of LR, or equivalently for large values of

$$F = \frac{(\tilde{\sigma}^2 - \hat{\sigma}^2)/q}{\hat{\sigma}^2/(n-k)}$$

This is the F statistic for the test of hypothesis H_0 against H_1.
Under H_0, the F-statistic has an exact distribution

$$F = \frac{e'(M_1 - M)e/q}{e'Me/(n-k)} \sim \frac{\chi_q^2/q}{\chi_{n-k}^2/(n-k)} \sim F_{q,n-k}$$

(i) Introduction to large sample asymptotics

We want to find an approximation to sampling distributions without requiring the assumption of normality.

For example, let y_i and x_i be drawn from the joint density

$$f(y|x)f(x) = f(x,y) = \frac{1}{2\pi xy}exp(-\frac{1}{2}(logy - logx)^2)exp(-\frac{1}{2}(logx)^2)$$

The lognormal distribution has density function

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}}exp(-\frac{(logx-\mu)^2}{2\sigma^2})$$

so this is essentially a model $\log y = logx + loge$, or equivalently $y = xe$.
Notes:

i. Since $logx$ is normally distributed $\sim N(\mu,\sigma^2)$, x is lognormally distributed $\mu,\sigma^2)$. The same is true for e.
ii. Since the model is $y = xe$, errors are multiplicative.
iii. In this model, e is independent of x.

Let $\hat{\beta}$ be the slope coefficient estimate from a LS regression of y on x and a constant.

i. If we transform the model to a normal regression model, and then regress $logy$ on $logx$, we will get an exact sampling distribution that is normally distributed with mean at 1, and the error term will be $loge$.

ii. But if we just regress y on x, the estimated $\hat{\beta}$ may not correspond well to the true projection coefficient.
e is no longer mean zero since the model is $y = xe$ and

$$\mathbb{E}[y|x] = \mathbb{E}[xe|x] = X\mathbb{E}[e|x]$$

Additionally, $\mathbb{E}[\hat{\beta} - \beta|x]$ is no longer 0, and $\hat{\beta}$ is biased and lognormal.

iii. The sampling distribution of $\hat{\beta}$ is a function of the joint distribution of (y, x) and the sample size n.

Hence, to approximate the sampling distribution, we will need to use asymptotic theory, which approximates by taking the limit of the finite sample distribution as the sample size tends to infinity.

(j) Introduction to asymptotic theory for least squares

The asymptotic theory of LS estimation applies equally to the projection model (potentially nonlinear CEF) and the linear CEF model.

We focus on the results for the broader projection model.
Recall that

$$y_i = x_i'\beta + e_i$$

for $i = 1, ..., n$, where the linear projection coefficient is

$$\beta = (\mathbb{E}(x_i x_i'))^{-1}\mathbb{E}(x_i y_i)$$

Assumption 7.1

i. The observations $(y_i, x_i), i = 1, ..., n$ are iid.
ii. $\mathbb{E}(y^2) < \infty$.
iii. $\mathbb{E}(||x||^2) < \infty$.
iv. $Q_{xx} = \mathbb{E}(xx')$ is positive definite (i.e., invertible).

The distributional results will require a strengthening of these assumptions to finite fourth moments (see Lecture 5).

We use the WLLN and CMT to show that the LS estimator $\hat{\beta}$ is consistent for the projection coefficient β.
The derivation is based on three components:

i. The OLS estimator can be written as a continuous function of a set of sample moments.
ii. The WLLN shows that sample moments converge in probability to population moments.
iii. CMT states that continuous functions preserve convergence in probability.

Step 1: Observe that the OLS estimator is a function of the sample moments:

$$\hat{\beta} = (\frac{1}{n}\sum_{i=1}^{n} x_i x_i')^{-1}(\frac{1}{n}\sum_{i=1}^{n} x_i y_i) = \hat{Q}_{xx}^{-1}\hat{Q}_{xy}$$

where $\hat{Q}_{xx} = \frac{1}{n}\sum_{i=1}^{n} x_i x_i'$ and $\hat{Q}_{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i$

Step 2: By WLLN, as $n \to \infty$, the sample momdents converge in probability to the population moments:

$$\hat{Q}_{xx} \xrightarrow{p} \mathbb{E}(x_i x_i') = Q_{xx}$$
$$\hat{Q}_{xy} \xrightarrow{p} \mathbb{E}(x_i y_i) = Q_{xy}$$

Step 3: We can combine these equations using CMT, so as $n \to \infty$,

$$\hat{\beta} = \hat{Q}_{xx}^{-1}\hat{Q}_{xy} \xrightarrow{p} Q_{xx}^{-1}Q_{xy} = \beta$$

We can also write $\hat{\beta} = g(\hat{Q}_{xx}^{-1}\hat{Q}_{xy})$, where $g(A, b) = A^{-1}b$ is a continuous function of A and b at all values of the arguments such that A^{-1} exists. By Assumption 7.1, Q_{xx} is invertible and hence we can use the CMT since $g(A, b)$ is continuous at $A = Q_{xx}$.

We can also show consistency in an alternative way by focusing on $\hat{\beta} - \beta$:

$$\hat{\beta} - \beta = \hat{Q}_x^{-1}x\hat{Q}_{xe}$$

where

$$\hat{Q}_{xe} = \frac{1}{n}\sum_{i=1}^{n} x_i e_i$$

By WLLN,

$$\hat{Q}_{xe} \xrightarrow{p} \mathbb{E}(x_i e_i) = 0$$

and hence

$$\begin{aligned}\hat{\beta} - \beta &= \hat{Q}_{xx}^{-1}\hat{Q}_{xe} \\ &\xrightarrow{p} Q_{xx}^{-1}0 \\ &= 0\end{aligned}$$

In stochastic order notation, we can equivalently write this as

$$\hat{\beta} = \beta + o_p(1)$$

i.e., the deviation diminishes as n increases.

5. Lecture 5: Asymptotic theory for least squares

(a) Asymptotic normality

Before we can do hypothesis testing, we must first derive the asymptotic distribution of $\hat{\beta}$. To do this, we need to:

i. Write the estimator as a function of sample moments
ii. One of the moments must be written as a sum of zero-mean random vectors and normalised so that the CLT can be applied.

Specifically:

$$\hat{\beta} - \beta = \hat{Q}_{xx}^{-1}\hat{Q}_{xe} \qquad\qquad \sqrt{n}(\hat{\beta} - \beta) = (\frac{1}{n}\sum_{i=1}^{n} x_i x_i')^{-1})(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_i e_i)$$

where the latter is mean zero.

Recall that the CLT is:
Lindeberg-Levy Central Limit Theorem: If y_i are iid and $\mathbb{E}(y_i^2) < \infty$, then as $n \to \infty$

$$\sqrt{n}(\bar{y} - \mu) \xrightarrow{d} N(0, \sigma^2)$$

where $\mu = \mathbb{E}(y)$ and $\sigma^2 = \mathbb{E}(y_i - \mu)^2$.

In multivariate form, we have:
Multivariate Lindeberg-Levy Central Limit Theorem: If $\boldsymbol{y}_i \in \mathbb{R}^k$ are iid and $\mathbb{E}||(\boldsymbol{y}_i)||^2 < \infty$, then as $n \to \infty$

$$\sqrt{n}(\bar{\boldsymbol{y}} - \boldsymbol{\mu}) \xrightarrow{d} N(\boldsymbol{0}, \boldsymbol{V})$$

where $\boldsymbol{\mu} = \mathbb{E}(\boldsymbol{y})$ and $\boldsymbol{V} = \mathbb{E}((\boldsymbol{y} - \boldsymbol{\mu})(\boldsymbol{y} - \boldsymbol{\mu})')$.

We apply the multivariate CLT on the random vector $\boldsymbol{x}_i e_i$, but before we can do this, we need to show that $\mathbb{E}||\boldsymbol{x}_i e_i||^2 < \infty$. To do this, we strengthen Assumption 7.1:

Note: for the subsequent lines, the vectors are no longer bolded for simplicity of notetaking, but the same notation is implied as in the previous sections.

Assumption 7.2

i. The observations $(y_i, x_i), i = 1, ..., n$ are iid.
ii. $\mathbb{E}(y^4) < \infty$.
iii. $\mathbb{E}(||x||^4) < \infty$.
iv. $Q_{xx} = \mathbb{E}(xx')$ is positive definite (i.e., invertible).

With these assumptions we can now show $\mathbb{E}||x_i e_i||^2$. Specifically:

$$\begin{aligned}\mathbb{E}||x_i e_i||^2 &= \mathbb{E}(||x_i||^2 e_i^2)\\ &\leq (\mathbb{E}||x_i||^4)^{\frac{1}{2}}(\mathbb{E}(e_i^4))^{\frac{1}{2}}\\ &< \infty\end{aligned}$$

Notes:

i. The first equality is due to $||ax|| = |a|\ ||x||$ as a is a scalar.
ii. The second inequality is because by CS inequality,

$$(\mathbb{E}[x_k^2 x_l^2])^{\frac{1}{2}}(\mathbb{E}[e^4])^{\frac{1}{2}} \leq (\mathbb{E}[x_j^4])^{\frac{1}{4}}(\mathbb{E}[x_l^4])^{\frac{1}{4}}(\mathbb{E}[e^4])^{\frac{1}{2}}$$

iii. The final inequality is due to Assumption 7.2.iii and because if $\mathbb{E}|y|^r < \infty$ for $r \geq 1$ then $\mathbb{E}|e|^r < \infty$.

Therefore, we can apply the multivariate CLT to obtain

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_i e_i \xrightarrow{d} N(0, \Omega)$$

where $\Omega := \mathbb{E}((x_i x_i' e_i^2))$, which is $(k \times k)$. Note that since we assume iid, e_i is diagonal.

$\Omega < \infty$ (all elements are finite) because

$$||\Omega|| = ||\mathbb{E}(x_i x_i' e_i^2)|| \leq \mathbb{E}(||x_i x_i' e_i^2||) = \mathbb{E}(||(x_i e_i)(x_i e_i)'||) = \mathbb{E}(||x_i e_i||^2) \leq \infty$$

Notes:

i. The first inequality is due to the expectation inequality and Jensen's inequality since the absolute function is convex.
ii. The second inequality is due to the identity $||xx'|| = ||x||^2$ for a $m \times 1$ vector.

Hence, under Assumption 7.2,

$$\Omega < \infty$$

and

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_i e_i \xrightarrow{d} N(0, \Omega)$$

as $n \to \infty$.

Then, we have

$$\begin{aligned}\sqrt{n}(\hat{\beta} - \beta) &\xrightarrow{d} Q_{xx}^{-1} N(0, \Omega)\\ &= N(0, Q_{xx}^{-1}\Omega Q_{xx}^{-1})\end{aligned}$$

and the asymptotic distribution of the LS estimator is:

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

where

$$V_\beta = Q_{xx}^{-1}\Omega Q_{xx}^{-1}$$

$Q_{xx} = \mathbb{E}(x_i x_i')$ and $\Omega = \mathbb{E}(x_i x_i' e_i^2)$ (note that this is an unconditional distribution).

In stochastic order notation, this implies that

$$\hat{\beta} = \beta + O_p(n^{-1/2})$$

which is stronger than the consistency result we saw in Lecture 4:

$$\hat{\beta} = \beta + o_p(1).$$

The matrix $V_\beta = Q_{xx}^{-1} \Omega Q_{xx}^{-1}$ is the variance of the asymptotic distribution $\sqrt{n}(\hat{\beta} - \beta)$.

i. Consequently, V_β is often referred to as the asymptotic covariance matrix of $\hat{\beta}$.
ii. This expression is often called a sandwich form.

We can compare the asymptotic variance

$$V_\beta = Q_{xx}^{-1} \Omega Q_{xx}^{-1}$$

and the finite-sample conditional variance of $\hat{\beta}$ is

$$V_{\hat{\beta}} = var(\hat{\beta}|X) = (X'X)^{-1}(X'DX)(X'X)^{-1}$$

where $D = \mathbb{E}(ee'|X)$

Since $V_{\hat{\beta}}$ shrinks to zero as $n \to \infty$, we can rescale $V_{\hat{\beta}}$:

$$nV_{\hat{\beta}} = (\frac{1}{n}X'X)^{-1}(\frac{1}{n}X'DX)(\frac{1}{n}X'X)^{-1}$$

As $n \to \infty$, $nV_{\hat{\beta}} \xrightarrow{p} V_\beta$.
We can see this because:

i. $V_{\hat{\beta}} = Q_{xx}^{-1} \mathbb{E}(x_i x_i' e_i^2) Q_{xx}^{-1}$
ii. $(\frac{1}{n}X'X)^{-1} \xrightarrow{p} Q_{xx}^{-1}$
iii. $(\frac{1}{n}X'DX) \xrightarrow{p} \frac{1}{n}\mathbb{E}(x_i x_i' e_i^2)$

Under homoskedasticity, we have

$$cov(x_i x_i' e_i^2) = 0$$

so the asymptotic variance simplies to

$$\begin{aligned} \Omega &= \mathbb{E}(x_i x_i')\mathbb{E}(e_i^2) = Q_{xx}\sigma^2 \\ V_\beta &= Q_x^{-1} x \Omega Q_{xx}^{-1} = Q_{xx}^{-1}\sigma^2 \equiv V_\beta^0 \end{aligned}$$

and we can show that $nV_{\hat{\beta}}^0 \xrightarrow{p} V_\beta^0$ where $V_\beta^0 = (X'X)^{-1}\sigma^2$.

(b) Consistency of error variance estimators

We can show that the estimators $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n \hat{e}_i^2$ and $s^2 - \frac{1}{n-k}\sum_{i=1}^n \hat{e}_i^2$ are consistent for $\sigma^2 = \mathbb{E}(e_i^2)$.

We do this by writing

$$\begin{aligned} \hat{e}_i &= y_i - x_i'\hat{\beta} \\ &= e_i + x_i'\beta - x_i'\hat{\beta} \\ &= e_i - x_i'(\hat{\beta} - \beta) \end{aligned}$$

And hence the squared residual is

$$\hat{e}_i^2 = e_i^2 - 2e^i x_i'(\hat{\beta} - \beta) + (\hat{\beta} - \beta)' x_i x_i' (\hat{\beta} - \beta)$$

We then obtain

$$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n} e_i^2 - 2(\frac{1}{n}\sum_{i=1}^{n} e_i x_i')(\hat{\beta}-\beta) \\ &+ (\hat{\beta}-\beta)'(\frac{1}{n}\sum_{i=1}^{n} x_i x_i')(\hat{\beta}-\beta)\end{aligned}$$

By the WLLN,

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n} e_i^2 \xrightarrow{p} \sigma^2 \\ &\frac{1}{n}\sum_{i=1}^{n} e_i x_i' \xrightarrow{p} \mathbb{E}(e_i x_i') = 0 \\ &\frac{1}{n}\sum_{i=1}^{n} x_i x_i' \xrightarrow{p} \mathbb{E}(x_i x_i') = Q_{xx}\end{aligned}$$

and we know $\hat{\beta} \xrightarrow{p} \beta$.
Theorem 7.4: Under Assumption 7.1, $\hat{\sigma}^2 \xrightarrow{p} \sigma^2$ and $s^2 \xrightarrow{p} \sigma^2$ as $n \to \infty$.

(c) Covariance matrix estimation

Under homoskedasticity:

i. $V_\beta^0 = Q_{xx}^{-1}\sigma^2$, and a plug-in estimator is $\hat{V}_\beta^0 = \hat{Q}_{xx}^{-1}s^2$
where $\hat{V}_\beta^0 = \hat{Q}_{xx}^{-1}s^2 = n(X'X)^{-1}s^2 = n\hat{V}_{\hat{\beta}}^0$

ii. By WLLN and CMT we have

$$\hat{V}_\beta^0 = \hat{Q}_{xx}^{-1}s^2 \xrightarrow{p} Q_{xx}^{-1}\sigma^2 = V_\beta^0$$

where $\hat{V}_\beta^0 = n\hat{V}_{\hat{\beta}}^0$.

Theorem 7.5: Under Assumption 7.1, $\hat{V}_\beta^0 \xrightarrow{p} V_\beta^0$ as $n \to \infty$.

Under heteroskedasticity:

i. Recall that $V_\beta = Q_{xx}^{-1}\Omega Q_{xx}^{-1}$ where $\Omega = \mathbb{E}(x_i x_i' e_i^2)$

ii. A plug-in estimator is $\hat{V}_\beta = \hat{Q}_{xx}^{-1}\hat{\Omega}\hat{Q}_{xx}^{-1}$ where

$$\hat{\Omega} = \frac{1}{n}\sum_{i=1}^{n} x_i x_i' \hat{e}_i^2$$

where $\hat{V}_\beta = \hat{Q}_{xx}^{-1}\hat{\Omega}\hat{Q}_{xx}^{-1} = n(X'X)^{-1}\hat{\Omega}(X'X)^{-1} = n\hat{V}_{\hat{\beta}}^{HC0}$.

Theorem 7.6: Under Assumption 7.1, $\hat{\Omega} \xrightarrow{p} \Omega$ and $\hat{V}_\beta^{HC0} \xrightarrow{p} V_\beta$ as $n \to \infty$.

(d) Functions of parameters

Suppse the researcher is interested in a specific transformation of the coefficient vector β.
Express $\theta = r(\beta)$ for some function $r : \mathbb{R}^k \to \mathbb{R}^q$.

The estimate of θ is $\hat{\theta} = r(\hat{\beta})$.
By CMT we have
Theorem 7.8: Under Assumption 7.1, if $r(\beta)$ is continuous at the true value of β, then as $n \to \infty$, $\hat{\theta} \xrightarrow{p} \theta$.

Furthermore, if the transformation is sufficiently smooth (Assumption 7.3), then by the Delta method we can show that $\hat{\theta}$ is asymptotically normal.

Assumption 7.3: $r(\beta) : \mathbb{R}^k \to \mathbb{R}^q$ is continuously differentiable at the true value of β and $R = \frac{\partial}{\partial\beta} r(\beta)'$ has rank q.
Note: R is a $k \times q$ matrix.

Theorem 7.9: Asymptotic Distribution of Functions of Parameters
Under Assumptions 7.2 and 7.3, as $n \to \infty$,

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, V_\theta) \tag{Eq 7.25}$$

where

$$V_\theta = R' V_\beta R.$$

A special case is a linear transformation (i.e., $r(\beta)$ is a linear function of β; $r(\beta) = R\beta$).
In this case, we don't need to rely on the Delta method, and can just use the linear transformation of normally distributed random variables to obtain (Eq 7.25).

An even more special case is when R is a "selector matrix", e.g., $R = \begin{pmatrix} I \\ 0 \end{pmatrix}$
Then, we can partition $\beta = (\beta_1', \beta_2')'$ so that $R'\beta = \beta_1$ for $\beta = (\beta_1', \beta_2')$. Then

$$V^\theta = \begin{pmatrix} I & 0 \end{pmatrix} V_\beta \begin{pmatrix} I \\ 0 \end{pmatrix} = V_{11}$$

which is the upper-left sub-matrix of V_{11}. In this case,

$$\sqrt{n}(\hat{\beta}_1 - \beta_1) \xrightarrow{d} N(0, V_{11})$$

That is, subsets of $\hat{\beta}$ are approximately normal with variances given by the conformable subcomponents of V.

To estimate the asymptotic variance $V_\theta = R' V_\beta R$, we use the plug-in estimator

$$\hat{V}_\theta = \hat{R}' \hat{V}_\beta \hat{R}$$

where

$$\hat{R} = \frac{\partial}{\partial\beta} r(\hat{\beta})'$$

which is a derivative evaluated at $\hat{\beta}$.

$$\hat{R} = \frac{\partial}{\partial\beta} r(\hat{\beta})' \xrightarrow{p} \frac{\partial}{\partial\beta} r(\beta)' = R$$

since $\hat{\beta} \xrightarrow{p} \beta$ and the function $\frac{\partial}{\partial\beta} r(\beta)'$ is continuous in β.
We have the following result:
Theorem 7.10: Under Assumptions 7.2 and 7.3, as $n \to \infty$,

$$\hat{V}_\theta \xrightarrow{p} V_\theta$$

(e) Asymptotic standard errors

The standard error for the j^{th} element of β is

$$s(\hat{\beta}_j) = \sqrt{\hat{V}_{\hat{\beta}_j}} = \sqrt{[\hat{V}_{\hat{\beta}}]_{jj}}$$

When the justification is based on asymptotic theory, $s(\hat{\beta}_j)$ is called an **asymptotic standard error**.

Standard errors for $\hat{\theta}$ are constructed similarly. When θ is real-valued, we have

$$s(\hat{\theta}) = \sqrt{\hat{R}'\hat{V}_{\hat{\beta}}\hat{R}} = \sqrt{n^{-1}\hat{R}'\hat{V}_{\beta}\hat{R}}$$

where $\hat{R} = \frac{\partial}{\partial \beta} r(\hat{\beta})'$.

In Stata, the *nlcom* command can be used after estimation to report the estimate, asymptotic standard error and 95% confidence intervals.

(f) t-statistics

Let the parameter of interest be a real-valued $\theta = r(\beta)$. We consider the statistic

$$T(\theta) = \frac{\hat{\theta} - \theta}{s(\hat{\theta})}$$

Recall that

$$s(\hat{\theta}) = \sqrt{\hat{R}'\hat{V}_{\hat{\beta}}\hat{R}} = \sqrt{n^{-1}\hat{R}'\hat{V}_{\beta}\hat{R}}$$

where $\hat{V}_{\beta} \xrightarrow{p} V_{\beta}$, the asymptotic covariance matrix of $\hat{\beta}$.

We rewrite T as

$$T(\theta) = \frac{\sqrt{n}(\hat{\theta} - \theta)}{\sqrt{\hat{R}\hat{V}_{\beta}\hat{R}}} = \frac{\sqrt{n}(\hat{\theta} - \theta)}{\sqrt{\hat{V}_{\theta}}}$$

since $\hat{V}_{\theta} = \hat{R}'\hat{V}_{\beta}\hat{R}$.
The asymptotic distribution of T is standard normal.

By Theorems 7.9 and 7.10, $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, V_{\theta})$ and $\hat{V}_{\theta} \xrightarrow{p} V_{\theta}$. Thus

$$\begin{aligned} T(\theta) &= \frac{\hat{\theta} - \theta}{s(\hat{\theta})} \\ &= \frac{\sqrt{n}(\hat{\theta} - \theta)}{\sqrt{\hat{V}_{\theta}}} \\ &\xrightarrow{d} \frac{N(0, V_{\theta})}{\sqrt{V_{\theta}}} \\ &= Z \sim N(0,1) \end{aligned}$$

Note that $T(\theta)$ is asymptotically pivotal as its asymptotic distribution does not depend on parameters.
This calculation requires that $V_{\theta} > 0$, otherwise we cannot employ the CMT. Formally we add the following assumption:
Assumption 7.4: $V_{\theta} = R'V_{\beta}R > 0$.

Alternatively, if we assume a more primitive condition $\Omega = \mathbb{E}(x_i x_i' e_i^2) > 0$, then since $Q_{xx} > 0$ it follows that $V_{\beta} > 0$. Then, since R has full rank under Assumption 7.3, it follows that Assumption 7.4 will hold.

It is also useful to consider the distribution of the **absolute t-ratio** $|T(\theta)|$. Since $T(\theta) \xrightarrow{d} Z$, the CMT yields $|T(\theta)| \xrightarrow{d} |Z|$. The distribution function of $|Z|$ is

$$\begin{aligned} P(|Z| \leq u) &= P(-u \leq Z \leq u) \\ &= P(Z \leq u) - P(Z < -u) \\ &= \Phi(u) - \Phi(-u) \\ &= 2\Phi(u) - 1 \end{aligned} \qquad \text{(Eq 7.34)}$$

where Φ is the standard normal CDF.

Hence we have:
Theorem 7.11: Under Assumptions 7.2, 7.3, 7.4, $T(\theta) \xrightarrow{d} Z \sim (0,1)$ and $|t_n(\theta)| \xrightarrow{d} |Z|$.

(g) Confidence intervals

We have considered $\hat{\theta}$ which is a point estimator for θ.
A broader concept is a **set estimator** $\hat{C}$ which is a collection of values in $\mathbb{R}^q$.

When θ is real-valued, we often focus on $\hat{C} = [\hat{L}, \hat{U}]$ which is an **interval estimator** for θ.

The **coverage probability** of $\hat{C} = [\hat{L}, \hat{U}]$ is $P(\theta \in \hat{C}]$

i. $\hat{C}$ is a function of the data and hence is random.
ii. We consider parameter θ to be fixed.

An interval estimator is called a **confidence interval** when the goal is to set the coverage probability to equal a pre-specified target such as 90% or 95%.

i. $\hat{C}$ is called a $1-\alpha$ confidence interval if $\inf_\theta P_\theta(\theta \in \hat{C}) = 1 - \alpha$.

CIs are usually obtained by "inverting" the information from a test statistic.
For example, let

$$\hat{C} = \{\theta : |T(\theta)| \leq C\} = \{\theta : -c \leq \frac{\hat{\theta} - \theta}{s(\hat{\theta})} \leq c\}$$

where c is the $1-\alpha$ quantile of the distribution of $|Z|$.
To find c, we can just apply (Eq 7.34):

$$\begin{aligned} 2\Phi(c) - 1 &= 1 - \alpha \\ 2\Phi(c) &= 2 - \alpha \\ c &= \Phi^{-1}(1 - \frac{\alpha}{2}) \end{aligned}$$

We know that the asymptotic coverage probability of this CI is

$$P(\theta \in \hat{C}) = P(|T(\theta)| \leq c) \to P(|Z| \leq c) = 1 - \alpha$$

Because the t-ratio is asymptotically pivotal, the asymptotic coverage probability is independent of the parameter θ. This allows us to "invert" the information to construct the CI explicitly.

Specifically, we rearrange each of the inequalities to obtain:

$$\{\theta : \frac{\hat{\theta} - \theta}{s(\hat{\theta})} \leq c \qquad \text{and} \qquad \frac{\hat{\theta} - \theta}{s(\hat{\theta}} \geq -c\}$$

Then, put the random variables on one side of the inequality and the parameter on the other side to obtain

$$\{\theta : \hat{\theta} - c \cdot s(\hat{\theta}) \leq \theta \qquad \text{and} \qquad \hat{\theta} + c \cdot s(\hat{\theta}) \geq \theta\}$$

Combining the inequalities, we have

$$\hat{C} = [\hat{\theta} - c \cdot s(\hat{\theta}), \hat{\theta} + c \cdot s(\hat{\theta})] \tag{Eq 7.35}$$

which still has 95% coverage probability.

Theorem 7.12: Under Assumptions 7.2, 7.3 and 7.4, for $\hat{C}$ defined in (Eq 7.35), with $c = \Phi^{-1}(1 - \alpha/2)$, $P(\theta \in C) \to 1 - \alpha$.
For $c = 1.96, P(\theta \in \hat{C}) \to 0.95$.

(h) Wald statistic

Suppose we are interested in a parameter vector θ which is a transformation of β. We are interested in the **Wald statistic**, which is a quadratic form built on $\hat{\theta}$ and θ.

Let $\theta = r(\beta) : \mathbb{R}^k \to R^q$ be any parameter vector of interest, $\hat{\theta}$ be its estimator and $\hat{V}_{\hat{\theta}}$ its covariance matrix estimator. Consider the quadratic form

$$W(\theta) = (\hat{\theta} - \theta)'\hat{V}_{\hat{\theta}}^{-1}(\hat{\theta} - \theta) = n(\hat{\theta} - \theta)'\hat{V}_{\theta}^{-1}(\hat{\theta} - \theta)$$

where $\hat{V}_\theta = n\hat{V}_{\hat{\theta}}$. When $q = 1$, then $W(\theta) = T(\theta)^2$ is the square of the t-ratio. When $q > 1$, $W(\theta)$ is typically called the **Wald Statistic**.

The asymptotic distribution of $W(\theta)$ can be derived given Theorem 7.9 and 7.10.

$$W(\theta) = \sqrt{(}\hat{\theta} - \theta)'\hat{V}_{\theta}^{-1}\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} Z'V_{\theta}^{-1}Z$$

We then need to use Assumption 7.4 that $V_\theta > 0$ (positive definite). We can then perform a Cholesky Decomposition to decompose the “square roots” of V_θ:

Cholesky Decomposition: If A is $k \times k$ and positive definite then $A = LL'$ where L is lower triangular and full rank, and non-singular.

Because $V_\theta > 0$, we can write $V_\theta = CC'$ and $V_\theta^{-1} = C'^{-1}C^{-1}$.
Then, $Z'V_\theta^{-1}Z = Z'C'^{-1}C^{-1}Z = (C^{-1}Z)'(C^{-1}Z)$.
$C^{-1}Z \xrightarrow{d} N(0, C^{-1}V_\theta C^{-1}]) = N(0, C^{-1}CC'C^{-1'}) = N(0, I_q)$
Therefore, since $Z'V_\theta^{-1}Z$ is the sum of squared mutually independent standard normal distributions (as can be seen from the quadratic form), it follows a chi-squared distribution with q degrees of freedom.
Theorem 7.13: Under Assumptions 7.2, 7.3 and 7.4, as $n \to \infty$,

$$W(\theta) \xrightarrow{d} \chi^2_q$$

Under homoskedasticity, the Wald statistic is

$$W^0(\theta) = (\hat{\theta} - \theta)'(\hat{V}^0_{\hat{\theta}})^{-1}(\hat{\theta} - \theta) = n(\hat{\theta} - \theta)'(\hat{V}^0_\theta)^{-1}(\hat{\theta} - \theta)$$

which has the same asymptotic distribution as above.

(i) Confidence regions

A confidence region $\hat{C}$ is a set estimator for $\theta \in R^q$ when $q > 1$.
We can “invert” the information from the Wald statistic to construct a confidence region, which is an ellipse:

$$\hat{C} = \{\theta : W(\theta) \leq c_{1-\alpha}\}$$

with $c_{1-\alpha}$ the $1 - \alpha$ quantile of the χ^2_q distribution (thus $F_q(c_{1-\alpha}) = 1 - \alpha$).

Theorem 7.13 implies

$$P(\theta \in \hat{C}) \to P(\chi^2_q \leq c_{1-\alpha}) = 1 - \alpha$$

6. **Lecture 6: Restricted estimation and hypothesis testing**

(a) Restricted estimation

We often impose a constraint on β in the linear projection model:

$$y_i = x_i'\beta + e_i$$
$$\mathbb{E}(x_i e_i) = 0$$

In general, a set of q linear constraints (or restrictions) on β takes the form

$$R'\beta = c$$

where R is $k \times q$, rank $(R) = q < k$, and c is $q \times 1$.

Assumption 8.1: $R'\beta = c$ where R is $k \times q$ with rank$(R) = q$.

i. This is analogous to what we examined before (functions of parameters).
ii. As in before, we assume that there are fewer constraints than the number of parameters ($q < k$) and R has full rank.
iii. Full rank means that the constraints are linearly independent (there are no redundant or contradictory constraints).

Examples of constraints:

i. $\beta_1 = \beta_2 = 0$
ii. $\beta_1 = 1, \beta_2 = -1$
iii. $\beta_1 = -\beta_2$
iv. $\beta_1 + \beta_2 = 0$ (equivalent to the preceding constraint).

In each case, we can define R accordingly.

We can define the restricted parameter space as

$$B_R = \{\beta : R'\beta = c\}$$

That is, the set of values β that satisfy the constraints.

A **constrained (or restricted)** estimator is an estimator that satisfies Assumption 8.1.
A typical reason to impose a constraint is that we believe (or have information) that the constraint is true.

i. By imposing it we hope to improve estimation efficiency.
ii. The goal is to obtain consistent estimates with reduced variance relative to the unconstrained estimator.

(b) Constrained least squares

The **constrained LS (CLS) estimator** is

$$\tilde{\beta}_{cls} = \arg \min_{R'\beta = c} SSE(\beta)$$

where

$$SSE(\beta) = \sum_{i=1}^{n} (y_i - x_i'\beta)^2 = y'y - 2y'X\beta + \beta'X'X\beta$$

by notation, we use $\tilde{\beta}$ to indicate that we are dealing with a constrained estimator.
As this is a constrained maximisation problem, the Lagrangian is

$$\mathcal{L}(\beta, \lambda) = \frac{1}{2} SSE(\beta) + \lambda'(R'\beta - c)$$

over (β, λ), where λ is a $q \times 1$ vector of Lagrange multipliers.
The FOCs are

$$\frac{\partial}{\partial \beta} \mathcal{L}(\tilde{\beta}_{cls}, \tilde{\lambda}_{cls}) = -X'y + X'X\tilde{\beta}_{cls} + R\tilde{\lambda}_{cls} = 0 \tag{Eq 8.6}$$

and

$$\frac{\partial}{\partial \lambda} \mathcal{L}(\tilde{\beta}_{cls}, \tilde{\lambda}_{cls}) = R'\tilde{\beta}_{cls} - c = 0 \tag{Eq 8.7}$$

Premultiplying (Eq 8.6) by $R'(X'X)^{-1}$, we obtain

$$-R'\hat{\beta}_{ols} + R'\tilde{\beta}_{cls} + R'(X'X)^{-1}R\tilde{\lambda}_{cls} = 0$$

where $\hat{\beta} = (X'X)X'y$ is the unconstrained LS estimator.

i. Note that in unconstrained estimation, $\lambda = 0$ and we premultiply (Eq 8.6) by $(X'X)^{-1}$ to obtain the solution of the OLS estimator (we can think of $R = I$ and $C = \beta$).

ii. Here, we apply a similar trick by premultiplying by $R'(X'X)^{-1}$ (the $(X'X)^{-1}$ matrix "rotated" by R).

Substituing in (Eq 8.7), we have

$$\tilde{\beta}_{cls} = \hat{\beta}_{ols} - (X'X)^{-1}R[R'(X'X)^{-1}R]^{-1}(R'\hat{\beta}_{ols} - c)$$

i. The CLS estimator equals the OLS estimator minus some deviation.

ii. The deviation is a linear function of $R'\hat{\beta}_{ols} - c$ (how much the unconstrained estimator deviates from the constraint).

iii. In Stata, CLS is implemented using the *cnsreg* command.

(c) Finite sample properties of the CLS estimator

Consider the linear CEF model

$$\begin{aligned} y_i &= x_i'\beta + e_i \\ \mathbb{E}(e_i|x_i) &= 0 \end{aligned}$$

Suppose Assumption 8.1 holds, i.e., $R'\beta = c$.
We can then show the following:
Theorem 8.1: Define $P = X(X'X)^{-1}X'$ and

$$A = (X'X)^{-1}R(R'(X'X)^{-1}R)^{-1}R'(X'X)^{-1}$$

(note that A is symmetric).
Then, we can show that

i. $R'\hat{\beta} - c = R'(X'X)^{-1}X'e$
Proof:

$$\begin{aligned} R'\hat{\beta} - c &= R'(\hat{\beta} - \beta) \\ &= R'((X'X)^{-1}X'y) - (X'X)^{-1}X'X\beta) \\ &= R'(X'X)^{-1}X'e \end{aligned}$$

ii. $\tilde{\beta}_{cls} - \beta = ((X'X)^{-1}X' - AX')e$
Proof:

$$\begin{aligned} \tilde{\beta}_{cls} - \beta &= \hat{\beta} - (X'X)^{-1}R[R'(X'X)^{-1}R]^{-1}(R'\hat{\beta} - c) - \beta \\ &= (X'X)^{-1}X'e - (X'X)^{-1}R[R'(X'X)^{-1}R]^{-1}(R'\hat{\beta} - c) \\ &= (X'X)^{-1}X'e - (X'X)^{-1}R[R'(X'X)^{-1}R]^{-1}R'(X'X)^{-1}X'e \\ &= ((X'X)^{-1}X' - AX')e \end{aligned}$$

Note: This is just the OLS deviation $\hat{\beta} - \beta$ adjusted by a linear function of e, $(AX'e)$.

The CLS estimator is unbiased for β:
Theorem 8.2: In the linear regression model,

$$\mathbb{E}(\tilde{\beta}_{cls}|X) = \beta$$

Proof:

$$\tilde{\beta}_{cls} - \beta = ((X'X)^{-1}X' - AX')e$$

then, just take expectations and apply conditioning theorem and LIE.

If we assume homoskedasticity, we can show that:
Theorem 8.3: In the homoskedastic linear regression model with $\mathbb{E}(e_i^2|x_i) = \sigma^2$, under Theorem 8.1:

$$\begin{aligned} V_{\tilde{\beta}}^0 &= var(\tilde{\beta}_{cls}|X) \\ &= ((X'X)^{-1} - (X'X)^{-1}R(R'(X'X)^{-1}R)^{-1})^{-1}R'(X'X)^{-1})\sigma^2 \\ &= ((X'X)^{-1} - A)\sigma^2 \end{aligned}$$

Examining the $k \times k$ matrix A:

i. It can be expressed as $A = G'BG$ where $G = R'(X'X)^{-1}$ and $B = R'(X'X)^{-1}R > 0$.
ii. A is positive semi-definite.
iii. G has full column rank if $q = k$.
iv. $(X'X)^{-1} - A \geq 0$ due to the definition of variance.
v. We can express $(X'X)^{-1} - A = C'C$, where $C' \equiv (X'X)^{-1}X' - AX'$.

Therefore we have shown that CLS is unbiased, and CLS has lower variance (in a positive definite sense) than OLS and thus is more efficient. Note that all the above results assume that Assumption 8.1 is true.

(d) Hausman equality

Under Assumption 8.1 and homoskedasticity, CLS is the efficient estimator, and OLS is inefficient. We can show that the variance of the difference between both estimators is equal to the difference between the variances.
Specifically,

$$\begin{aligned}(\hat{\beta}_{ols} - \tilde{\beta}_{cls}, \tilde{\beta}_{cls}) &= \mathbb{E}((\hat{\beta}_{ols} - \tilde{\beta}_{cls})(\tilde{\beta}_{cls} - \beta)') \\ &= \mathbb{E}((AX')(X(X'X)^{-1} - XA))\sigma^2 = 0\end{aligned}$$

Thus, the deviation of OLS from CLS is orthogonal to the CLS itself.
In other words, $\hat{\beta}_{ols} - \tilde{\beta}_{cls}$ and $\tilde{\beta}_{cls}$ are conditionally uncorrelated and hence independent.

One corollary is

$$cov(\hat{\beta}_{ols}, \tilde{\beta}_{cls}) = var(\tilde{\beta}_{cls})$$

This is because $cov(\hat{\beta}_{ols}, \tilde{\beta}_{cls}) = cov(\hat{\beta}_{ols} - \tilde{\beta}_{cls}, \tilde{\beta}_{cls}) + cov(\tilde{\beta}_{cls}, \tilde{\beta}_{cls}) = 0 + var(\tilde{\beta}_{cls})$

We then have

$$\begin{aligned}var(\hat{\beta}_{ols} - \tilde{\beta}_{cls}) &= var(\hat{\beta}_{ols}) \\ &= var(\hat{\beta}_{ols}) - 2cov(\hat{\beta}, \tilde{\beta}_{cls}) + var(\tilde{\beta}_{cls}) \\ &= var(\hat{\beta}_{ols}) - 2var(\tilde{\beta}_{cls}) + var(\tilde{\beta}_{cls}) \\ &= var(\hat{\beta}_{ols}) - var(\tilde{\beta}_{cls})\end{aligned}$$

This is known as **Hausman Equality** where the variance of the difference between estimators is equal to the difference between the variances.
It occurs (generically) when we are comparing an efficient and inefficient estimator.
This will later allow us to form a test statistic (such as the t-statistic) easily.

(e) Minimum distance

In CLS, we find a numerical estimate which satisfies the constraint such that it is “as close as possible” to the unconstrained estimate (in the SSE sense).
More generally, a **minimal distance (md)** estimate tries to find a parameter value satisfiying the constraint which is as close as possible to the unconstrained estimator.

Let $\hat{\beta}$ be the unconstrained estimator. For some $(k \times k)$ positive definite **weight matrix** $\hat{W} > 0$, define the **quadratic criterion function**

$$J(\beta) = n(\hat{\beta} - \beta)'\hat{W}(\hat{\beta} - \beta)$$

which is a (squared) Euclidean distance between $\hat{\beta}$ and β.

i. $J(\beta)$ is small if β is close to $\hat{\beta}$
ii. $J(\beta)$ is minimised at zero only if $\beta = \hat{\beta}$

A minimum distance estimator $\tilde{\beta}_{md}$ for β is

$$\tilde{\beta}_{md} = arg \min_{R'\beta=c} J(\beta)$$

which is a constrained estimator.

We can show that the CLS estimator is a special case when $\hat{W} = \hat{Q}_{xx}$, that is, the CLS estimator minimises the criterion

$$J^0(\beta) = n(\hat{\beta}-\beta)'\hat{Q}_{xx}(\hat{\beta}-\beta)$$

To see this, rewrite the least squares criterion as:

$$\begin{aligned} SSE(\beta) &= \sum_{i=1}^{n}(y_i - x_i'\beta)^2 \\ &= \sum_{i=1}^{n}(x_i'\hat{\beta} + \hat{e}_i - x_i'\beta)^2 \\ &= \sum_{i=1}^{n}\hat{e}_i^2 + (\hat{\beta}-\beta)'(\sum_{i=1}^{n} x_i x_i')(\hat{\beta}-\beta) \\ &= n\hat{\sigma}^2 + J^0(\beta) \end{aligned}$$

that is, we break down the SSE into:

i. The minimised value under unconstrained minimisation, and
ii. The penalty due to the distance between the unconstrained solution and other candidate values.

Only the later term is a function of β.

Therefore, the CLS estimator is a special kind of minimum distance estimator.
We can solve for the minimum distance estimator $\tilde{\beta}_{md}$ by the method of Lagrange multipliers.
The Lagrangian is:

$$\mathcal{L}(\beta,\lambda) = \frac{1}{2}K(\beta,\hat{W}) + \lambda'(R'\beta - c)$$

and the solution is

$$\begin{aligned} \tilde{\lambda}_{md} &= n(R'\hat{W}^{-1}R)^{-1}(R'\hat{\beta} - c) \\ \tilde{\beta}_{md} &= \hat{\beta} - \hat{W}^{-1}R(R'\hat{W}^{-1}R)^{-1}(R'\hat{\beta} - c) \end{aligned}$$

setting $\hat{W} = (X'X)^{-1}$, the solution becomes the CLS solution.

(f) Effective minimum distance estimator

The asymptotic covariance matrix of the minimum distance estimator depends on the weight matrix W, and is written as $V_\beta(W)$.
The asymptotically optimal weight matrix is the one that minimises $V_\beta(W)$.
It turns out that the optimal weight matrix is $W = V_\beta^{-1}$, i.e., the weight matrix is the inverse of the variance of β.
In practice, we can use the feasible weight matrix $\hat{W} = \hat{V}_\beta^{-1}$.

The **efficient minimum distance estimator** is:

$$\tilde{\beta}_{emd} = \hat{\beta} - \hat{V}_\beta R(R'\hat{V}_{beta}R)^{-1}(R'\hat{\beta} - c) \qquad \text{(Eq 8.25)}$$

Theorem 8.9: Efficient Minimum Distance Estimator
Under Assumptions 7.2 and 8.1,

$$\sqrt{n}(\tilde{\beta}_{emd} - \beta) \xrightarrow{d} N(0, V_{\beta,emd})$$

as $n \to \infty$, where

$$V_{\beta,emd} = V_\beta - V_\beta R(R'V_\beta R)^{-1}R'V_\beta$$

Since

$$V_{\beta,emd} \leq V_\beta$$

the estimator in (Eq 8.25) has a lower asymptotic variance than the unrestricted estimator. Furthermore, for any W,

$$V_{\beta,emd} \leq V_\beta(W)$$

so (Eq 8.25) is asymptotically efficient in the class of minimum distance estimators.

We can verify that under homoskedasticity, the CLS estimator is the efficient minimum distance estimator.

(g) Hypothesis testing

Hypothesis tests attempt to asses whether there is evidence to contradict a proposed parametric restriction. Let

$$\theta = r(\beta)$$

be a $q \times 1$ parameter of interest, where $r : \mathbb{R}^k \to \Theta \in \mathbb{R}^q$ is some transformation.
A point hypothesis concerning θ is a proposed restriction such as $\theta = \theta_0$, where θ_0 is a hypothesised (known) value.

More generally, letting $\beta \in B \in \mathbb{R}^k$ be the parameter space (β is a vector of restrictions), a hypothesis is a restriction $\beta \in B_0$ where B_0 is a proper subset of B. This specialises to 9.1 below by setting $B_0 = \{\beta \in B : r(\beta) = \theta_0\}$.

The hypothesis to be tested is called the null hypothesis:
Definition 9.1: The **null hypothesis**, written H_0, is the restriction $\theta = \theta_0$ or $\beta \in B_0$.
Definition 9.2: The **alternative hypothesis**, written H_1, is the set $\{\theta \in \Theta : \theta \neq \theta_0\}$, or $\{\beta \in B : \beta \neq B_0\}$.

The decision (either "accept H_0" or "reject H_0") is a mapping from the sample space to the decision set.
This splits the sample space into two regions (S_0, S_1).

i. If the observed sample falls into S_0 (**acceptance region**), we accept H_0.
ii. If the sample falls into S_1 (**rejection region**), we reject H_0.

Express this mapping as a real-valued function called a **test statistic**

$$T = T((y_1, x_1)...(y_n, x_n))$$

relative to a **critical value** c.

The decision rule is

i. Accept H_0 if $T \leq c$
ii. Reject H_0 if $T > c$

A test statistic should be designed so that small values are likely when H_0 is true and large values are likely when H_1 is true. An example is

$$T = |T(\theta_0)|$$

where

$$T(\theta) = \frac{\hat{\theta} - \theta}{s(\hat{\theta})}$$

A false rejection of H_0 (rejecting H_0 when H_0 is true) is a **Type 1 error**.
The finite-sample **size** of the test is

$$P(\text{Reject } H_0|H_0 \text{ true}) = P(T > c|H_0 \text{ true}) \tag{Eq 9.4}$$

which is more formally defined as the supremum of (Eq 9.4) across all data distributions that satisfy H_0.
We want to limit the incidence of Type 1 error by bounding the size of the test.
Suppose the test statistic has an asymptotic distribution under H_0, i.e., when H_0 is true, $T \xrightarrow{d} \xi$, and let $G(u) = P(\xi \leq u)$ denote the distribution of ξ. We call ξ (or G) the **asymptotic null distribution**.

i. If ξ (or G) does not depend on unknown parameters, then T is **asymptotically pivotal**.
ii. We define the **asymptotic size** of the test as

$$\begin{aligned}\lim_{n\to\infty} P(T > c|H_0 \text{ true}) &= P(\xi > c)\\ &= 1 - G(c)\end{aligned}$$

iii. We can pre-select a **significance level** $\alpha \in (0, 1)$ and then select c (the critical value) such that the size of the test is no larger than α. Set c equal to the $1 - \alpha$ quantile of the distribution G. Note: We generally use two-sided critical values unless the parameter space is known to satisfy a one-sided restriction.

For example, for the t-test:

$$\begin{aligned}H_0 &: \theta = \theta_0\\ H_1 &: \theta \neq \theta_0\\ T(\theta_0) &= |\frac{\hat{\theta} - \theta_0}{s(\hat{\theta})}|\end{aligned}$$

Theorem 9.1: Under Assumptions 7.2, 7.3 and $H_0 : \theta = \theta_0$,

$$T(\theta_0) \xrightarrow{d} Z$$

For c satisfying $\alpha = 2(1 - \Phi(c))$,

$$P(|T(\theta_0)| > c|H_0) \to \alpha$$

and the test "Reject H_0 if $|T(\theta_0)| > c$" has asymptotic size α, which is the preselected significance level.

A false acceptance of H_0 (accepting H_0 when H_1 is true) is called a **Type 2 error**. The **power** of the test is the rejection probability under H_1:

$$\pi(\theta) = P(\text{Reject } H_0|H_1 \text{ true}) = P(T > c|H_1 \text{ true})$$

$\pi(\theta)$ is the **power function** and it depends on the true value of the parameter θ.
The goal of test construction is to have **high power subject to the constraint that the size of the test is lower than the pre-specified significance level** (although to be able to make this choice requires us to have multiple test statistics to choose from).
For a well-behaved test, the power is increasing both as θ moves away from θ_0 and as the sample size n increases.

Given a test statistic T, reducing the critical value c reduces the likelihood of a Type 1 error, but increases the likelihood of a Type 2 error. Thus the choice of c involves a tradeoff between size and power. The size can't be arbitrarily small because then the power will be small too.
Other notes:

i. When we reject the null, we say that the statistic is **statistically significant** at significance level α. Otherwise, it is **statistically insignificant**.
ii. When we reject the null, it doesn't necessarily imply that the null is false (remember the Type 1 error). When we accept the null, it doesn't necessarily imply that the null is true (remember the Type 2 error).

iii. The test does not tell the probability of H_0 or H_1 being true. In the frequentist approach, H_0 is either true or not true.

iv. Statistical vs. economic significance: Results can be economically insignificant even if strongly statistically significant. One solution is to focus whenever possible on confidence intervals and the economic meaning of the coefficients.

p-values measure the strength of the evidence against the null hypothesis.
The asymptotic p-value is

$$p = 1 - G(T)$$

and is just an alternative representation of the evidence contained in the test statistic.

P-values can be interpreted as: If H_0 is true, what is the probability that you obtain an event that is at least as extreme as the one from the current sample?
The lower the p-value, the stronger the evidence against the null.

Furthermore, the asymptotic p-value has a convenient asymptotic null distribution. Since $T \xrightarrow{d} \xi$ under H_0, then $p = 1 - G(T) \xrightarrow{d} 1 - G(\xi)$, which has the distribution

$$\begin{aligned}
P(1 - G(\xi) \leq u) &= P(1 - u \leq G(\xi)) \\
&= 1 - P(\xi \leq G^{-1}(1-u)) \\
&= 1 - G(G^{-1}(1-u)) \\
&= 1 - (1-u) \\
&= u
\end{aligned}$$

Therefore, if H_0 is true, the p-value has an asymptotic uniform distribution on $[0, 1]$, i.e., $p \xrightarrow{d} U[0, 1]$.
We can then set up a decision rule based on the p-value. Based on the preselected significance level α:

i. Reject null if $p < \alpha$

ii. Accept null if $p \geq \alpha$

This decision rule has a size of α.
In general, when a coefficient θ is of interest, it is constructive to focus on the point estimate, s.e.'s and confidence intervals.

(h) Wald tests

Wald tests are a generalisation of the t-test. They can be used when we have $q > 1$ restrictions, where the discrepancy $\hat{\theta} - \theta_0$ is a vector and has more than one measure of length.

The **Wald statistic** is just a quadratic form evaluated at H_0. It measures the weighted Euclidean measure of the length of the vector $\hat{\theta} - \theta_0$ (deviation of $\hat{\theta}$ from hypothesised value θ_0).

$$W = W(\theta_0) = (\hat{\theta} - \theta_0)' \hat{V}_{\hat{\theta}}^{-1} (\hat{\theta} - \theta_0)$$

where $\hat{V}_{\hat{\theta}} = \hat{R}' \hat{V}_{\hat{\beta}} \hat{R}$ is an estimator of $V_{\hat{\theta}}$ and $\hat{R} = \frac{\partial}{\partial \beta} r(\hat{\beta})'$. Notice that we can write W alternatively as

$$W = n(\hat{\theta} - \theta_0)' \hat{V}_{\theta}^{-1} (\hat{\theta} - \theta_0)$$

where the feasible asymptotic variance $\hat{V}_{\theta} = n\hat{V}_{\hat{\theta}}$.

We can write W as a function of $\hat{\beta}$ generally as

$$W = (r(\hat{\beta}) - \theta_0)' (\hat{R} \hat{V}_{\hat{\beta}} \hat{R})^{-1} (r(\hat{\beta}) - \theta_0)$$

When the restrictions are linear, this becomes

$$W = (R'\hat{\beta} - \theta_0)' (\hat{R} \hat{V}_{\hat{\beta}} \hat{R})^{-1} (R'\hat{\beta} - \theta_0)$$

Theorem 9.2: Under Assumptions 7.2, 7.3, 7.4, and $H_0 : \theta = \theta_0$, then

$$W \xrightarrow{d} \chi_q^2 \qquad \text{(see Theorem 7.13 in Lecture 5)}$$

and for c satisfying $\alpha = 1 - G_q(c)$,

$$P(W > c|H_0) \to a$$

so the test "Reject H_0 if $W > c$" has asymptotic size α.
Notes:

i. The asymptotic p-value of W is $p = 1 - G_q(w)$
ii. The F version of the Wald statistic is $F = \frac{W}{q}$.

(i) Criterion-based tests

These are an alternative class of tests analogous to the likelihood ratio test, based on the discrepancy between the criterion function J minimised with and without the restriction.

The unconstrained estimator is

$$\hat{\beta} = arg \min_{\beta \in B} J(\beta)$$

The constrained estimator (imposing the restriction under H_0) is

$$\tilde{\beta} - arg \min_{\beta \in B_0} J(\beta)$$

And hence the **criterion-based statistic** (or distance statistic, or minimum-distance statistic) is

$$\begin{aligned} J &= \min_{\beta \in B_0} J(\beta) - \min_{\beta \in B} J(\beta) \\ &= J(\tilde{\beta}) - J(\hat{\beta}) \end{aligned}$$

Note that $J \geq 0$.
If J is very large, the null hypothesis is unlikely to be true.

The statistic J measures the cost (on the criterion) of imposing the null restriction.
To construct this statistic, we typically need to estimate both $\hat{\beta}$ and $\tilde{\beta}$ (unlike the Wald statistic where only the unconstrained model is estimated).

(j) Minimum distance tests

The **minimum distance test** uses a quadratic form criterion

$$J(\beta) = n(\hat{\beta} - \beta)'\hat{W}(\hat{\beta} - \beta)$$

where $\hat{\beta}$ is the unrestricted estimator.

As in the case of restricted estimation, $J(\hat{\beta}) = 0$. Therefore the **minimum distance statistic** is

$$J = J(\tilde{\beta}_{md}) = n(\hat{\beta} - \tilde{\beta}_{emd})'\hat{W}(\hat{\beta} - \tilde{\beta}_{emd})$$

It measures the weighted Euclidean measure of the length of the vector $\hat{\beta} - \tilde{\beta}_{md}$ (the discrepancy between the unconstrained estimator $\hat{\beta}$ and the constrained estimator $\tilde{\beta}_{md}$).

The **efficient minimum distance statistic** is

$$J^* = n(\hat{\beta} - \tilde{\beta}_{emd})'\hat{W}(\hat{\beta} - \tilde{\beta}_{emd})$$

We can show that under linear restrictions, $J^* = W$ and the minimum distance and Wald tests are equivalent.

For nonlinear restrictions, both statistics are different.

Theorem 9.4: Under Assumptions 7.2, 7.3, 7.4, and $H_0 : \theta = \theta_0$,

$$J^* \xrightarrow{d} \chi^2_q$$

Under homoskedasticity, the criterion for the emd statistic is

$$J^0(\beta) = n(\hat{\beta} - \beta)'\hat{Q}_{xx}(\hat{\beta} - \beta)/s^2$$

and the constrained estimator $\tilde{\beta}_{emd}$ is just the CLS estimator.

(k) F tests

The F statistic is a criterion-based statistic as it involves estimating both $\tilde{\beta}$ and $\hat{\beta}$. The statistic is

$$F = \frac{(\tilde{\sigma}^2 - \hat{\sigma}^2)/q}{\hat{\sigma}^2/(n-k)}$$

where $\tilde{\beta}_{cls}$ stands for the CLS model and $\hat{\beta}$ stands for the unconstrained model.

Alternatively, we can write

$$F = \frac{SSE(\tilde{\beta}_{cls} - SSE(\hat{\beta})}{qs^2}$$

where

$$SSE(\beta) = \sum_{i=1}^{n}(y_i - x_i'\beta)^2$$

We can then rewrite

$$F = J^0/q$$

Therefore, the F statistic is identical to the homoskedastic minimum distance statistic divided by the number of restrictions, q.

(l) Problems with tests of nonlinear hypotheses

When the restrictions are linear, **all of the above tests are equivalent**.

When the restrictions are nonlinear, the Wald tests may perform poorly in **finite samples**. Intuitively, the Wald test relies solely on the unconstrained model.

Example 1: Take the model

$$y_i = \beta + e_i$$
$$e_i \sim N(0, \sigma^2)$$

and consider the hypothesis

$$H_0 : \beta = 1$$

The Wald test for H_0 is

$$W = n\frac{(\hat{\beta} - 1)^2}{\hat{\sigma}^2}$$

where $\hat{\beta}$ and $\hat{\sigma}^2$ are the sample mean and variance of y_i.

Note that H_0 is equivalent to the hypothesis that

$$H_0(s) : \beta^s = 1$$

for any positive integer s (although this makes the restriction unnecessarily nonlinear).
Let $r(\beta) = \beta^s$, and noting $R = s\beta^{s-1}$, the Wald test for $H_0(s)$ is

$$W(s) = n\frac{(\hat{\beta}^s - 1)^2}{\hat{\sigma}^2 s^2 \hat{\beta}^{2s-2}}$$

(in other words, this overcomplicates the earlier restriction; we could have simplified by leaving the restriction as $s = 1$, then the Wald statistic would be what we had earlier for H_0).
We have the same evidence, the same hypothesis, but the statistic $W(s)$ varies with s numerically.

Note that the asymptotic result says that $W(s) \xrightarrow{d} \chi^2_1$ under H_0 for any s. That is, the value of s does not matter asymptotically. The asymptotic 5% critical value is 3.84.

We can use **Monte Carlo simulation** to study the exact distributions of statistical procedures in finite samples:

i. Choose the DGP and true parameter values β, σ^2. Choose the sample size n.
ii. Use the model to create B random samples (each of sample size n) via simulation (these are also called B replications).
iii. In each replication, use the random sample to estimate $\hat{\beta}, \hat{\sigma}^2$. Then, compute $W(s; \hat{\beta}, \hat{\sigma}^2)$ and reject or accept $H_0(s)$ based on the significance level α.
iv. The rejection frequency of a test is then computed as the total number of rejections divided by B.

More generally, suppose we are interested in an estimator $\hat{\theta}$. Monte-Carlo simulations can be used to calculate the bias, mean-squared error, and variance of the distribution of $\hat{\theta} - \theta$.

$$\begin{aligned}
\widehat{Bias(\hat{\theta})} &= \frac{1}{B}\sum_{b=1}^{B} T_b = \frac{1}{B}\sum_{b=1}^{B} \hat{\theta}_b - \theta \\
\widehat{MSE(\hat{\theta})} &= \frac{1}{B}\sum_{b=1}^{B} (T_b)^2 = \frac{1}{B}\sum_{b=1}^{B} (\hat{\theta}_b - \theta)^2 \\
\widehat{Var(\hat{\theta})} &= \widehat{MSE(\hat{\theta})} - (\widehat{Bias(\hat{\theta})})^2
\end{aligned}$$

Suppose we are interested in the Type 1 error associated with an asymptotic 5% two-sided t-test. We would then set $T = |\hat{\theta} - \theta|/s(\hat{\theta})$ and calculate

$$\hat{P} = \frac{1}{B}\sum_{b=1}^{B} 1(T_b \geq 1.96)$$

Returning to the Wald test, the results (not replicated here) show that when $s = 1$, the test performs well, but when $s > 1$, the test over-rejects in small samples.
Therefore, the Wald test is sensitive to the choice of s in finite samples.

Example 2: Take the model

$$\begin{aligned}
y_i &= \beta_0 + x_{1i}\beta_1 + x_{2i}\beta_2 + e_1 && \text{(Eq 9.14)} \\
\mathbb{E}(x_i e_i) & \qquad\qquad\qquad 0
\end{aligned}$$

and the hypothesis

$$H_0 : \frac{\beta_1}{\beta_2} = \theta_0$$

Let $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ be the least-squares estimator of (Eq 9.14), let $\hat{V}_{\hat{\beta}}$ be an estimator of the covariance matrix for $\hat{\beta}$ and set $\hat{\theta} = \hat{\beta}_1/\hat{\beta}_2$. Define

$$\hat{R}_1 = \begin{pmatrix} 0 & \frac{1}{\hat{\beta}_2} & -\frac{\hat{\beta}_1}{\hat{\beta}_2^2} \end{pmatrix}'$$

so that the standard error for $\hat{\theta}$ is $s(\hat{\theta}) = (\hat{R}_1' \hat{V}_{\hat{\beta}} \hat{R}_1)^{1/2}$. In this case a t-statistic for H_0 is

$$T_1 = \frac{(\frac{\hat{\beta}_1}{\hat{\beta}_2} - \theta_0)}{s(\hat{\theta})}$$

An alternative statistic can be constructed through reformulating the null hypothesis as

$$H_0 = \beta_1 - \theta_0 \beta_2 = 0$$

A t-statistic based on this formulation of the hypothesis is

$$T_2 = \frac{\hat{\beta}_1 - \theta_0 \hat{\beta}_2}{(R_2' \hat{V}_{\hat{\beta}} R_2)^{1/2}}$$

where

$$R_2 = \begin{pmatrix} 0 & 1 & -\theta_0 \end{pmatrix}'$$

In fact, we typically work with T_2. An example is $\beta_1 - 2\beta_2 = 0$, which is just a linear restriction.

We can compare the finite sample performance of the Wald test based on T_1 and T_2.
In the Monte Carlo simulation:

i. Let x_{1i} and x_{2i} be mutually independent $N(0,1)$ variables.
ii. e_i is an independent $N(0, \sigma^2$ draw with $\sigma = 3$
iii. Set $\beta_0 = 0, \beta_1 = 1$
iv. The free parameters are β_2 (which gives θ_0) and sample size n.

The results (not replicated here) show that the Wald test based on T_1 performs poorly in finite samples.

The conclusion is that whenever possible, **the Wald test should not be used to test nonlinear hypotheses**.
A simple solution is to use a criterion-based test or minimum distance test, which are invariant to the algebraic formulation of the null hypothesis.

(m) Confidence intervals by test inversion

Confidence intervals "repackage" the information contained in a test statistic.
Given a test statistic $T(\theta)$ and critical value c, the acceptance region "accept if $T(\theta) \leq c$" is identical to the confidence interval $\hat{C} = [\theta : T(\theta) \leq c]$.

As discussed in Lecture 5 general method for finding CIs is known as **test statistic inversion**.

Now suppose a parameter of interest $\theta = r(\beta)$ is a nonlinear function of β. In this case,

i. $\hat{\beta} = r(\hat{\beta})$ is the plug-in estimator
ii. $s(\hat{\theta}) = \sqrt{\hat{R}' \hat{V}_{\hat{\beta}} \hat{R}}$ is the standard error based on the Delta method.

We can still invert a t-statistic $|\frac{\hat{\theta} - \theta}{SE(\hat{\theta})}|$ to construct a CI, but the choice of parameterisation matters greatly under nonlinear hypotheses.
A solution is to **rewrite the hypothesis as a linear restriction**, for example if the hypothesis is $\theta = \frac{\beta_1}{\beta_2}$, we can rewrite the hypothesis as $\beta_1 - \theta \beta_2 = 0$.
The t-statistic is then

$$T(\theta) = \frac{\hat{\beta}_1 - \hat{\beta}_2 \theta}{(R' \hat{V}_{\hat{\beta}} R)^{1/2}}$$

where $R = \begin{pmatrix} -1 \\ -\theta \end{pmatrix}$ which depends on θ. Therefore, θ appears in multiple elements of $T(\theta)$. To find

$$\hat{C} = \{\theta : |T(\theta)| \leq 1.96\}$$

we need to use a grid search method over θ, i.e., loop over each grid point of θ and depermine where $|T(\theta)| \leq 1.96$ at each gridpoint of θ.

(n) Multiple tests and Bonferroni corrections

The problem of multiple testing occurs when we find that one statistic appears to be "significant" after examining a large number of statistics.

Suppose we examine a set of k coefficients, SEs and t-ratios, and consider the "significance" of each statistic. Individually, each t-test has asymptotic size α.

Under the joint null hypothesis that a set of k hypotheses are all true, the probability that we observe at least one p-value smaller than α is bounded by αk. Therefore, the decision rule "reject the joint null hypothesis if one of the p-values $< \alpha$" severely over-rejects relative to the pre-selected significance level α.

The appropriate approach is to use an F-test, for example, which takes into account the correlation between the k test statistics.

7. Lecture 7: Instrumental Variables

(a) Overview

There is **endogeneity** in the linear model

$$y_i = x_i'\beta + e_i \qquad \text{the "structural equation"}$$

if β is the parameter of interest and

$$\mathbb{E}(x_i e_i) \neq 0 \tag{Eq 12.2}$$

We will call the above a "structural equation" and β a "structural parameter", to distinguish it from the regrression and projection models.
When (Eq 12.2) holds, it is typical to say that x_i is **endogenous** for β.

The parameter β is not necessarily a linear projection coefficient.
In fact, endogeneity cannot happen if the parameter is defined by the linear projection coefficient.
To see this, define

$$\beta^* = \mathbb{E}(x_i x_i')^{-1}\mathbb{E}(x_i y_i)$$

and the linear projection equation is

$$\begin{aligned} y_i &= x_i'\beta^* + e_i^* \\ \mathbb{E}(x_i e_i^*) &= 0 \end{aligned}$$

which implies no endogeneity for β^*.

Also, under endogeneity, β^* (from the linear projection coefficient) $\neq \beta$ (from the structural equation).

$$\begin{aligned} \beta^* &= (\mathbb{E}(x_i x_i'))^{-1}\mathbb{E}(x_i y_i) \\ &= (\mathbb{E}(x_i x_i'))^{-1}\mathbb{E}(x_i(x_i'\beta + e_i) \\ &= \beta + (\mathbb{E}(x_i x_i'))^{-1}\mathbb{E}(x_i e_i) \\ &\neq \beta \qquad \text{because } \mathbb{E}(x_i e_i) \neq 0 \end{aligned}$$

Hence endogeneity implies that the LS estimator is inconsistent for the structural parameter.

$$\hat{\beta} \xrightarrow{p} (\mathbb{E}(x_i x_i'))^{-1}\mathbb{E}(x_i y_i) = \beta^* \neq \beta$$

The inconsistency of the LS estimator is referred to as **endogeneity bias** (note: the actual issue is inconsistency, not bias).

(b) Examples

i. Example 1: Measurement error in the regressor

Suppose $(y_i, \boldsymbol{z}_i)$ are joint random variables, $\mathbb{E}(y_i|\boldsymbol{z}_i) = \boldsymbol{z}_i'\boldsymbol{\beta}$ is linear, $\boldsymbol{\beta}$ is the structural parameter, and $\boldsymbol{z}_i$ is not observed.

Instead, we observe $\boldsymbol{x}_i = \boldsymbol{z}_i + \boldsymbol{u}_i$, where $\boldsymbol{u}_i$ is a $k \times 1$ measurement error, independent of $e_i := y_i - E(y_i|\boldsymbol{z}_i)$ and $\boldsymbol{z}_i$.
Note: Since $\boldsymbol{z}$ is a vector of k regressors, $\boldsymbol{x}$ is a vector of k measurement errors. For subsequent discussion we drop the bold notation for easier notetaking.

This is an example of a **latent variable model**, where “latent” refers to a structural variable which is unobserved.

The model $x_i = z_i = u_i$ with z_i and u_i independent and $\mathbb{E}(u_i) = 0$ is known as **classical measurement error**. This means that x_i is a noisy but unbiased measure of z_i.

By substitution we can express y_i as a function of the observed variable x_i:

$$\begin{aligned} y_i &= z_i'\beta + e_i \\ &= (x_i - u_i)'\beta + e_i \\ &= x_i'\beta + v_i \end{aligned}$$

where $v_i = e_i - u_i'\beta$

We can then express

$$y_i = x_i'\beta + v_i$$

but the error v_i is not a projection error, because

$$\mathbb{E}(x_i v_i) = \mathbb{E}[(z_i + u_i)(e_i - u_i'\beta)] = \mathbb{E}(u_i u_i')\beta \neq 0$$

as long as $\beta \neq 0$ and $E(u_i u_i') \neq 0$. The LS estimator is then inconsistent for β.

In a simplified case, we can calculate the form of the projection coefficient. Suppose that $k = 1$ and there is only one regressor. Substituting in the above result on $\mathbb{E}(x_i v_i)$, we have

$$\beta^* = \beta + \frac{\mathbb{E}(x_i v_i)}{x_i^2} = \beta(1 - \frac{\mathbb{E}(u_i^2}{\mathbb{E}(u_i^2)})$$

Since $\frac{\mathbb{E}(u_i^2)}{\mathbb{E}(x_i^2)} < 1$, the projection coefficient shrinks the structural parameter β toward zero. This is called **measurement error bias** or **attenuation bias**.

ii. Example 2: Supply and Demand

The variables q_i and p_i are determined jointly by the demand equation

$$q_i = -\beta_1 p_i + e_{1i}$$

and the supply equation

$$q_i = \beta_2 p_i + e_{2i}$$

The structural parameters are β_1, β_2.
Assume that the demand and supply shocks/ errors

$$e_i = \begin{pmatrix} e_{1i} \\ e_{2i} \end{pmatrix}$$

is iid, $\mathbb{E}(e_i) = 0$ and $\mathbb{E}(e_i e_i') = I_2$ (we just use a 2x2 example for this example).

If we regress q_i on p_i, the projection coefficient equals neither structural parameter.
Specifically, we can solve for q_i and p_i in terms of the errors. In matrix notation,

$$\begin{bmatrix} 1 & \beta_1 \\ 1 & -\beta_2 \end{bmatrix} \begin{pmatrix} q_i \\ p_i \end{pmatrix} = \begin{pmatrix} e_{1i} \\ e_{2i} \end{pmatrix}$$

Therefore,

$$\begin{aligned} \begin{pmatrix} q_i \\ p_i \end{pmatrix} &= \begin{bmatrix} 1 & \beta_1 \\ 1 & -\beta_2 \end{bmatrix}^{-1} \begin{pmatrix} e_{1i} \\ e_{2i} \end{pmatrix} \\ &= \begin{bmatrix} \beta_2 & \beta_1 \\ 1 & -1 \end{bmatrix} \begin{pmatrix} e_{1i} \\ e_{2i} \end{pmatrix} \left(\tfrac{1}{\beta_1+\beta_2}\right) \\ &= \begin{pmatrix} (\beta_2 e_{1i} + \beta_1 e_{2i})/(\beta_1+\beta_2) \\ (e_{1i} - e_{2i})/(\beta_1+\beta_2) \end{pmatrix} \end{aligned}$$

The projection of q_i on p_i yields

$$\begin{aligned} q_i &= \beta^* p_i + e_i^* \\ \mathbb{E}(p_i e_i^*) &= 0 \end{aligned}$$

where we can verify that

$$\beta^* = \frac{\mathbb{E}(p_i q_i)}{\mathbb{E}(p_i^2)} = \frac{\beta_2 - \beta_1}{2}$$

Thus the projection coefficient β^* equals neither structural parameter.
The fact that the projection coefficient is neither the supply nor demand slope is called **simultaneous equations bias**. This occurs generally when some variables are jointly determined, as in a market equilibrium.

Generally, when both the dependent variable and a regressor simultaneously determined, then the variables should be treated as endogenous.

iii. Example 3: Choice variables as regressors
Take the classic wage equation

$$log(wage) = \beta education + \epsilon$$

with β interpreted as the average causal effect of education on wage.

If wages are affected by unobserved ability, and individuals with high ability self-select into higher education, then e contains unobserved ability, so *education* and e will be positively correlated and hence *education* is endogenous.

The positive correlation means that the projection coefficient β^* will be upward biased relative to the structural parameter β.

This type of endogeneity occurs generally when y and x are both choices made by an economic agent, even if they are made at different points in time.

(c) Instruments

We defined endogeneity as the context where the regressor is correlated with the equation error.
In most applications, we only treat a subset of the regressors as endogenous. Some of the regressors will be treated as **exogenous** (they are uncorrelated with the equation error).

Specifically, make the partition

$$x_i = \begin{pmatrix} x_{1i} \\ x_{2i} \end{pmatrix} \begin{matrix} k_1 \\ k_2 \end{matrix}$$

and similarly

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \begin{matrix} k_1 \\ k_2 \end{matrix}$$

so that the structural equation is

$$\begin{aligned} y_i &= x_i'\beta + e_i \qquad \text{(Eq 12.4)} \\ &= x_{1i}'\beta_1 + x_{2i}'\beta_2 + e_i \end{aligned}$$

The regressors are assumed to satisfy

$$\begin{aligned} \mathbb{E}(x_{1i}e_i) &= 0 \\ \mathbb{E}(x_{2i}e_i) &\neq 0 \end{aligned}$$

Thus the number of regressors $k = k_1 + k_2$.

i. x_{1i} are exogenous variables
ii. x_{2i} are endogenous variables

for the structural parameter β.

In matrix notation, we can write the structural equation as

$$\begin{aligned} y &= X\beta + e \\ &= X_1\beta_1 + X_2\beta_2 + e \end{aligned}$$

In most applications, k_2 is small (1 or 2). To consistently estimate β we need additional information and these are called **instruments**.

Definition 12.1: The $l \times 1$ random vector z_i is an **instrumental variable** for (Eq 12.4) if

$$\begin{aligned} \mathbb{E}(z_ie_i) &= 0 \qquad \text{(Eq 12.5)} \\ \mathbb{E}(z_iz_i') &> 0 \qquad \text{(Eq 12.6)} \\ rank(\mathbb{E}(z_ix_i')) &= k \qquad \text{(Eq 12.7)} \end{aligned}$$

i. (Eq 12.5): The instruments are exogenous in the sense that they are uncorrelated with the regression error and determined outside the model for y_i.
ii. (Eq 12.6): This is a normalisation which excludes linearly redundant instruments that can be expressed as a linear combination of some other instruments
iii. (Eq 12.7): This is the **relevance condition**, which is necessary for the identification of the model and a unique solution. A necessary condition for this is that the number of instruments is greater than the number of endogenous regressors ($l > k$).

Note that the regressors x_{1i} satisfy condition 12.5 and thus should be included as instrumental variables. They are thus a subset of the variables z_i.
Notationally, we make the partition

$$z_i = \begin{pmatrix} z_{1i} \\ z_{2i} \end{pmatrix} = \begin{pmatrix} x_{1i} \\ z_{2i} \end{pmatrix} \begin{matrix} k_1 \\ l_2 \end{matrix}$$

Thus the number of instruments $l = k_1 + l_2$.

i. $x_{1i} = z_{1i}$ are called **included exogenous variables** (or simply exogenous variables).
ii. z_{2i} are called excluded exogenous variables (or simply instruments).
iii. z_{2i} can still be added to the structural equation, but they will have true zero coefficients.

The model is

i. Just-identified if $l = k$ (and $l_2 = k_2$)
ii. Over-identified if $l > k$ (and $l_2 > k_2$)

What variables can be used as instrumental variables?

i. From the definition $\mathbb{E}(z_i e_i) = 0$ we see that the instrument must be correlated with the equation error, meaning that it is excluded from the structural equation (sometimes called **exclusion restriction**).
ii. From the rank condition (12.7), the instrument must be correlated with the endogenous variables x_{2i} after controlling for x_{1i} (called relevance).

These two requirements are typically interpreted as requiring:

i. The instruments are determined outside the system for (y_i, x_{2i}).
ii. The instruments causally determine x_{2i}.
iii. The instruments do not causally determine y_i except through x_{2i}.

Revisiting the examples above:

i. **Measurement error**: A common choice for an instrument z_{2i} is an alternative measurement for z_i. For this z_{2i} to satisfy the property of IV, the measurement error in z_{2i} must be independent of that in x_i.
ii. **Supply and demand**: An instrument for price in a demand equation is a variable z_{2i} that influences supply but not demand (e.g., weather).
iii. **Choice variable as regressor**: This will be discussed later.

(d) Reduced form estimation

The reduced form is the relationship between the regressors x_i which are $(k \times 1)$ and the instruments z_i which are $(l \times 1)$. A linear reduced form model for x_i is

$$x_i = \Gamma' z_i + u_i \tag{Eq 12.9}$$

This is a multivariate regression with k dependent variables. We can also view it as k equations stacked over one another.

The $l \times k$ coefficient matrix Γ can be defined by linear projection:

$$\Gamma = \mathbb{E}(z_i z_i')^{-1}\mathbb{E}(z_i x_i')$$

so that

$$\mathbb{E}(z_i u_i') = 0$$

by definition of linear projection.
In matrix notation, we can write (Eq 12.9) as

$$X = Z\gamma + U$$

where X and U are $n \times k$, Z is $n \times l$, Γ is $l \times k$.
Note that Γ is well-defined and unique under (Eq 12.6).

We can also construct a reduced form equation for y_i, which expresses y_i as a function of exogenous variables only.
Substituting (Eq 12.9) into (Eq 12.4), we have

$$\begin{aligned} y_i &= (\Gamma' z_i + u_i)'\beta + e_i && \text{(Eq 12.14)} \\ &= z_i'\lambda + v_i \end{aligned}$$

where

$$\lambda = \Gamma\beta$$

and

$$v_i = u_i'\beta + e_i$$

Observe that

$$\mathbb{E}(z_i v_i) = \mathbb{E}(z_i u_i')\beta + \mathbb{E}(z_i e_i) = 0$$

and hence (Eq 12.14) is a projection equation.
This means that we can write the reduced form coefficient as

$$\lambda = \mathbb{E}(z_i z_i')^{-1}\mathbb{E}(z_i y_i)$$

which is well-defined and unique under (Eq 12.6).

Overall, the reduced form equations are

$$\begin{aligned} y_i &= z_i'\lambda + v_i \\ x_i &= \Gamma' z_i + u_i \end{aligned}$$

where $\lambda = \Gamma\beta$.

The reduced form equations can be estimated by least squares.
First step: the OLS estimator for Γ is

$$\hat{\Gamma} = (\sum_{i=1}^{n} z_i z_i')^{-1}(\sum_{i=1}^{n} z_i x_i')$$

(we can do k regressions and stack the estimates horizontally).

We then have

$$x_i = \hat{\Gamma}' z_i + \hat{u}_i$$

In matrix notation,

$$\begin{aligned} \hat{\Gamma} &= (Z'Z)^{-1}(Z'X) \\ X &= Z\hat{\Gamma} + \hat{U} \end{aligned}$$

Second step: the OLS estimate for λ is

$$\begin{aligned} \hat{\lambda} &= (\sum_{i=1}^{n} z_i z_i')^{-1}(\sum_{i=1}^{n} z_i y_i) \\ y_i &= z_i'\hat{\lambda} + \hat{v}_i \\ &= z_{1i}'\hat{\lambda}_1 + z_{2i}'\hat{\lambda}_2 + \hat{v}_i \end{aligned}$$

In matrix notation,

$$\begin{aligned} \hat{\lambda} &= (Z'Z)^{-1}(Z'y) \\ y &= Z\hat{\lambda} + \hat{v} \\ &= Z_1\hat{\lambda}_1 + Z_2\hat{\lambda}_2 + \hat{v} \end{aligned}$$

(e) Identification

A parameter is identified if it is a unique function of the probability distribution of the observables.
One way to show this is to write it as an explicit function of population moments.
For example, for the reduced form coefficients Γ and λ:

$$\begin{aligned} \Gamma &= \mathbb{E}(z_i z_i')^{-1}\mathbb{E}(z_i x_i') \\ \lambda &= \mathbb{E}(z_i z_i')^{-1}\mathbb{E}(z_i y_i) \end{aligned}$$

given that Definition 12.1 holds.

We are interested in the structural parameter β.
It relates to the reduced-form parameters by the relation $\lambda = \Gamma\beta$ and is identified if it is uniquely determined by this relation (note that Γ is not square unless $l = k$, so we cannot invert it).

This is a set of l equations with k unknowns, with $l \geq k$.
There is a unique solution for β iff $rank(\Gamma) = k$.

The identification equation $\lambda = \Gamma\beta$ is the same as

$$\mathbb{E}(z_i y_i) = \mathbb{E}(z_i x_i')\beta$$

Which has a unique solution iff

$$rank(\mathbb{E}(z_i x_i')) = k$$

which justifies (Eq 12.7) in Definition 12.1. This is called the **relevance condition**.

When $l = k$, Γ is invertible if Γ has full rank, so $\beta = \Gamma^{-1}\lambda$.
When $l > k$, the solution $\beta = (\Gamma'\Gamma)^{-1}\Gamma'\lambda$ is equivalent to applying least-squares to the system of equations $\lambda = \Gamma\beta$. This is l equations with k unknowns and no error.

(f) Instrumental variables estimator

We first consider the case where the model is just identified ($l = k$).

The assumption that z_i is an instrumental variable implies that $\mathbb{E}(z_i e_i) = 0$.
Making the substitution for e_i,

$$\begin{aligned}\mathbb{E}(z_i(y_i - x_i'\beta)) &= 0\\ \mathbb{E}(z_i y_i) - \mathbb{E}(z_i x_i')\beta &= 0\end{aligned}$$

Because $\mathbb{E}(z_i x_i')$ is invertible ($l = k$ and full rank), the solution for β is

$$\beta = (\mathbb{E}(z_i x_i'))^{-1}\mathbb{E}(z_i y_i)$$

The **instrumental variables (IV)** estimator replaces the population moments by sample versions:

$$\begin{aligned}\hat{\beta}_{iv} &= (\frac{1}{n}\sum_{i=1}^{n} z_i x_i')^{-1}(\frac{1}{n}\sum_{i=1}^{n} z_i y_i)\\ &= (\sum_{i=1}^{n} z_i x_i')^{-1}(\sum_{i=1}^{n} z_i y_i)\\ &= (Z'X)^{-1}(Z'y)\end{aligned}$$

In Stata, this can be done using $ivregress\ 2sls$.

More generally, it is common to refer to any estimator of the form

$$\hat{\beta}_{iv} = (W'X)^{-1}(W'y)$$

given a $n \times k$ matrix W as an IV estimator for β using the instrument W.

Alternatively, we can start with the solution $\beta = \Gamma^{-1}\lambda$ and replace the components with LS estimates. This is the **indirect least squares (ILS)** estimator.

$$\begin{aligned}\hat{\beta}_{ils} &= \hat{\Gamma}^{-1}\hat{\lambda}\\ &= ((Z'Z)^{-1})(Z'X))^{-1}((Z'Z)^{-1})Z'y))\\ &= (Z'X)^{-1})(Z'Z)(Z'Z)^{-1})(Z'y)\\ &= (Z'X)^{-1}(Z'y)\end{aligned}$$

the IV and ILS estimates are identical.

Given the IV estimator we define the residual vector

$$\hat{e} = y - X\hat{\beta}_{iv}$$

which satisfies

$$Z'\hat{e} = Z'y - Z'X(Z'X)^{-1}(Z'y) = 0$$

Since Z includes an intercept, this means that the residuals sum to zero, and are uncorrelated with the included and excluded instruments.

The IV estimator also has a demeaned representation like OLS:

$$\begin{aligned}\hat{\beta}_{iv} &= (\sum_{i=1}^{n} z_i(x_i - \bar{x})')^{-1}(\sum_{i=1}^{n} z_i(y_i - \bar{y})) \\ &= (\sum_{i=1}^{n}(z_i - \bar{z})(x_i - \bar{x})')^{-1}(\sum_{i=1}^{n}(z_i - \bar{z})(y_i - \bar{y}))\end{aligned}$$

Similar to OLS, the estimate for the slope coefficients is the IV with demeaned data and no intercept.

(g) Wald estimator

In many cases, the instrument is a binary variable.

Suppose the model has just one endogenous regressor and no other regressors beyond the intercept:

$$\begin{aligned}y_i &= x_i\beta + \alpha + e_i \\ \mathbb{E}(e_i|z_i) &= 0\end{aligned}$$

with z_i binary.

Taking expectations of the structural equation given $z_i = 1$ and $z_i = 0$ respectively, we have

$$\mathbb{E}(y_i|z_i = 1) = \mathbb{E}(x_i|z_i = 1)\beta + \alpha\mathbb{E}(y_i|z_i = 0) = \mathbb{E}(x_i|z_i = 0)\beta + \alpha$$

Subtracting and dividing, we have

$$\beta = \frac{\mathbb{E}(y_i|z_i = 1) - \mathbb{E}(y_i|z_i = 0)}{\mathbb{E}(x_i|z_i = 1) - \mathbb{E}(x_i|z_i = 0)}$$

The natural moment estimator for β replaces the expectations by the averages within the "grouped data" where $z_i = 1$ and $z_i = 0$ respectively. That is, we define the grouped means

$$\begin{aligned}\bar{y}_1 &= \frac{\sum_{i=1}^{n} z_i y_i}{\sum_{i=1}^{n} z_i} & \bar{y}_0 &= \frac{\sum_{i=1}^{n}(1 - z_i)y_i}{\sum_{i=1}^{n}(1 - z_i)} \\ \bar{x}_1 &= \frac{\sum_{i=1}^{n} z_i x_i}{\sum_{i=1}^{n} z_i} & \bar{x}_0 &= \frac{\sum_{i=1}^{n}(1 - z_i)x_i}{\sum_{i=1}^{n}(1 - z_i)}\end{aligned}$$

and the moment estimator

$$\hat{\beta} = \frac{\bar{y}_1 - \bar{y}_0}{\bar{x}_1 - \bar{x}_0}$$

This is known as the **Wald estimator**.

i. β is the expected change in y_i due to changing z_i, divided by the expected change in x_i due to changing z_i.
ii. Therefore we have only used the exogenous variation driven by z_i to estimate the structural relationship between x_i and y_i, and we have discarded the rest of the information.

(h) Two-stage least squares

2SLS allows the general case of $l \geq k$. The reduced-form equations give

$$\begin{aligned}y_i &= z_i'\Gamma\beta + v_i \\ \mathbb{E}(z_i v_i) &= 0\end{aligned}$$

Defining $w_i = \Gamma' z_i$, which is a $k \times 1$ vector (since Γ is $k \times l$ and z_i is $l \times 1$), we can write this as

$$\begin{aligned} y_i &= w_i'\beta + v_i \\ \mathbb{E}(w_i v_i) &= 0 \end{aligned}$$

The key insight is that we have "condensed" the $l \times 1$ instrument vector z_i into a $k \times 1$ instrument vector w_i.

Suppose Γ were known. Then, we can just perform least-squares regression of y_i on w_i:

$$\begin{aligned} \hat{\beta} &= (W'W)^{-1}(W'y) \\ &= (\Gamma' Z' Z \Gamma)^{-1}](\Gamma' Z' y) \end{aligned}$$

while this is infeasible, we can estimate Γ from the reduced-form regression of X on Z. This gives

$$\hat{\Gamma} = (Z'Z)^{-1}(Z'X)$$

We then obtain

$$\begin{aligned} \hat{\beta}_{2sls} &= (\hat{\Gamma} Z' Z \hat{\Gamma})^{-1}(\hat{\Gamma}' Z' y) \\ &= (X'Z(Z'Z)^{-1}Z'Z(Z'Z)^{-1}Z'X)^{-1}X'Z(Z'Z)^{-1}Z'y \\ &= (X'Z(Z'Z)^{-1}Z'X)^{-1}X'Z(Z'Z)^{-1}Z'y \end{aligned}$$

This is the **two-stage least squares (2SLS) estimator**.
It is identical to the following two-step procedure:

i. Regress X on Z and obtain $\hat{\Gamma}$ as well as fitted values $Z\hat{\Gamma}$ (a $n \times k$ matrix).
ii. Regress y on $Z\hat{\Gamma}$ to obtain $\hat{\beta}$.

If the model is just-identified (i.e., $l = k$), then the 2SLS estimator simplifies to the IV estimator. In particular, since $X'Z$ and $Z'X$ will be square and invertible, we can factor

$$\begin{aligned} (X'Z(Z'Z)^{-1}X'Z)^{-1} &= (Z'X)^{-1}((Z'Z)^{-1})^{-1}(X'Z)^{-1} \\ &= (Z'X)^{-1}(Z'Z)(X'Z)^{-1} \end{aligned}$$

Then,

$$\begin{aligned} \hat{\beta}_{2sls} &= (X'Z(Z'Z)^{-1}Z'X)^{-1}X'Z(Z'Z)^{-1}Z'y \\ &= (Z'X)^{-1}(Z'Z)(X'Z)^{-1}X'Z(Z'Z)^{-1}Z'y \\ &= (Z'X)^{-1}(Z'Z)(Z'Z)^{-1}Z'y \\ &= (Z'X)^{-1}Z'y \\ &= \hat{\beta}_{iv} \end{aligned}$$

There are several alternative representations of the 2SLS estimator.

i. First, defining the projection matrix

$$P_z = Z(Z'Z)^{-1}Z'$$

We can write the estimator more compactly as

$$\hat{\beta}_{2sls} = (X'P_Z X)^{-1}X'P_Z y$$

This is useful for representation and derivations, but not useful for computation as the $n \times n$ matrix P_Z is too large to compute when n is large.

ii. Second, define the fitted values for X from the reduced form

$$\hat{X} = P_Z X = Z\hat{\Gamma}$$

Then the estimator can be written as

$$\hat{\beta}_{2sls} = (\hat{X}'X)^{-1}\hat{X}'y$$

This is an IV estimator as defined in the previous section, using $\hat{X}$ as the instrument.

iii. Third, since P_Z is idempotent, we can write the estimator as

$$\begin{aligned}\hat{\beta}_{2sls} &= (X'P_ZP_ZX)^{-1}X'P_zy \\ &= (\hat{X}'\hat{X})^{-1}\hat{X}'y\end{aligned}$$

This is the LS estimator obtained by regressing y on fitted values $\hat{X}$.
The estimator can be computed by:

A. Step 1: Regress X on Z, vis., $\hat{\Gamma} = (Z'Z)^{-1}(Z'X)$ and $\hat{X} = Z\hat{\Gamma} = P_ZX$.

B. Step 2: Regress y on $\hat{X}$, vis., $\hat{\beta}_{2sls} = (\hat{X}'\hat{X})^{-1}\hat{X}'y$

For the prejection $\hat{X}$, $X = [X_1, X_2]$ and $Z = [X_1, Z_2]$. Notice that $\hat{X}_1 = P_ZX_1 = X_1$ since X_1 lies in the span of Z. Then

$$\hat{X} = [\hat{X}_1\hat{X}_2] = [X_1, \hat{X}_2]$$

Thus in the second stage, we regress y on X_1 and $\hat{X}_2$, so only the endogenous variables X_2 are replaced by their fitted values:

$$\hat{X}_2 = X_1\hat{\Gamma}_{12} + Z_2\hat{\Gamma}_{22}$$

The least squares estimator can be written as

$$y = X_1\hat{\beta}_1 + \hat{X}_2\hat{\beta}_2 + \hat{\epsilon}$$

iv. Fourth, let $P_1 = X_1(X_1'X_1)^{-1}X_1'$. Applying the FWL theorem we obtain

$$\begin{aligned}\hat{\beta}_2 &= (\hat{X}_2'(I_n - P_1)\hat{X}_2)^{-1}(\hat{X}_2'(I_n - P_1)y) \\ &= (X_2'P_Z(I_n - P_1)P_zX_2)^{-1}(X_2'P_Z(I_n - P_1)y) \\ &= (X_2'(P_Z = P_1)X_2)^{-1}(X_2'(P_Z - P_1)y)\end{aligned}$$

since $P_ZP_1 = P_1$.

v. Fifth, the projection matrix P_Z can be replaced by the projection onto the pair $[X_1, \tilde{Z}_2]$ where $\tilde{Z}_2 = (I_n - P_1)Z_2$ is Z_2 projected orthogonal to X_1.
Since X_1 and $\tilde{Z}_2$ are orthogonal, $P_Z = P_1 + P_2$ where $P_2 = \tilde{Z}_2(\tilde{Z}_2'\tilde{Z}_2)^{-1}\tilde{Z}_2'$.
Thus $P_Z - P_1 = P_2$ and

$$\begin{aligned}\hat{\beta}_2 &= (X_2'P_2X_2)^{-1}(X_2'P_2y) \\ &= (X_2'\tilde{Z}_2(\tilde{Z}_2'\tilde{Z}_2)^{-1}\tilde{Z}_2'X_2)^{-1}(X_2'\tilde{Z}_2(\tilde{Z}_2'\tilde{Z}_2)^{-1}\tilde{Z}_2'y)\end{aligned}$$

Note: We can show $P_Z = P_1 + P_2$ using the FWL theorem).

Given the 2SLS estimator we define the residual vector

$$\hat{e} = y - X\hat{\beta}_{2sls}$$

When the model is overidentified, the instruments and residuals are not orthogonal. That is,

$$Z'\hat{e} \neq 0$$

It does, however, satisfy

$$\begin{aligned}\hat{X}'\hat{e} &= \hat{\Gamma}'Z'\hat{e} \\ &= X'Z(Z'Z)^{-1}Z'\hat{e} \\ &= X'Z(Z'Z)^{-1}Z'y - X'Z(Z'Z)^{-1}Z'X\hat{\beta}_{2sls} \\ &= 0\end{aligned}$$

(i) Consistency and asymptotic distribution of 2SLS

Assumption 12.1

i. The observations $(y_i, x_i, z_i), i = 1, ..., n$ are iid

ii. $\mathbb{E}(y^2) < \infty$
iii. $\mathbb{E}||x||^2 < \infty$
iv. $\mathbb{E}||z||^2 < \infty$
v. $\mathbb{E}(zz')$ is positive definite.
vi. $\mathbb{E}(zx')$ has full rank k.
vii. $\mathbb{E}(ze) = 0$.

Assumptions 12.1.5-7 are identical to Definition 12.1.

Theorem 12.1: Under Assumption 12.1, $\hat{\beta}_{2sls} \xrightarrow{p} \beta$ as $n \to \infty$.
Proof: Similar to the proof for the LS estimator, take the structural equation $y = X\beta + e$ in matrix format and substitute it into the expression for the estimator. We obtain

$$\begin{aligned}\hat{\beta}_{2sls} &= (X'Z(Z'Z)^{-1}Z'X)^{-1}X'Z(Z'Z)^{-1}Z'(X\beta + e)\\ &= \beta + (X'Z(Z'Z)^{-1}Z'X)^{-1}X'Z(Z'Z)^{-1}Z'e \qquad \text{(Eq 12.41)}\end{aligned}$$

This separates out the stochastic component. Rewriting and applying the WLLN and CMT

$$\begin{aligned}\hat{\beta}_{2sls} - \beta &= ((\frac{1}{n}X'Z)(\frac{1}{n}Z'Z)^{-1}(\frac{1}{n}Z'X))^{-1}(\frac{1}{n}X'Z)(\frac{1}{n}Z'Z)^{-1}(\frac{1}{n}Z'e)\\ &\xrightarrow{p} (Q_{xz}Q_{zz}^{-1}Q_{zx})^{-1}Q_{xz}Q_{zz}^{-1}\mathbb{E}(z_ie_i) = 0 \qquad \text{(by WLLN)}\end{aligned}$$

where

$$\begin{aligned}Q_{xz} &= \mathbb{E}(x_iz_i')\\ Q_{zz} &= \mathbb{E}(z_iz_i') > 0\\ Q_{zx} &= \mathbb{E}(z_ix_i')\end{aligned}$$

The WLLN holds under the iid assumption (Assumption 12.1.1) and the finite second moment assumption (Assumptions 12.1.2-12.1.4).
The continuous mapping theorem applies if the matrices Q_{zz} and $Q_{xz}Q_{zz}^{-1}Q_{zx}$ are invertible, which hold under the identification assumptions (Assumptions 12.1.5-6).
The final inequality uses Assumption 12.1.7.

Assumption 12.2 In addition to Assumption 12.1,

i. $\mathbb{E}(y^4) < \infty$
ii. $\mathbb{E}||z||^4 < \infty$
iii. $\Omega = \mathbb{E}(zz'e^2)$ is positive definite.

Theorem 12.2: Under Assumption 12.2, as $n \to \infty$,

$$\sqrt{n}(\hat{\beta}_{2sls} - \beta) \xrightarrow{d} N(0, V_\beta)$$

where

$$V_\beta = (Q_{xz}Q_{zz}^{-1}Q_{zx})^{-1}(Q_{xz}Q_{zz}^{-1}\Omega Q_{zz}^{-1}Q_{zx})(Q_{xz}Q_{zz}^{-1}Q_{zx})^{-1}$$

Under the homoskedasticity condition we have the simplifications $\Omega = Q_{zz}\sigma^2$ and $V_\beta = V_\beta^0 \overset{def}{=} (Q_{xz}Q_{zz}^{-1}Q_{zx})^{-1}\sigma^2$.

Proof of Theorem 12.2:
The derivation of the asymptotic distribution builds on the proof of consistency. Using (Eq 12.41),

$$\sqrt{n}(\hat{\beta}_{2sls} - \beta) = ((\frac{1}{n}X'Z)(\frac{1}{n}Z'Z)^{-1}(\frac{1}{n}Z'X))^{-1}(\frac{1}{n}X'Z)(\frac{1}{n}Z'Z)^{-1}(\frac{1}{\sqrt{n}}Z'e)$$

We apply the WLLN and CMT for the moment matrices involving Z and Z, the same as in the proof of consistency. In addition, by the CLT for iid observations,

$$\frac{1}{\sqrt{n}}Z'e = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} z_ie_i \xrightarrow{d} N(0, \Omega)$$

because the vector z_ie_i is iid and mean zero under Assumptions 12.1.1 and 12.1.7, and has a finite second moment (we will prove this below).

We obtain

$$\begin{aligned}\sqrt{n}(\hat{\beta}_{2sls} - \beta) &= ((\frac{1}{n}X'Z)(\frac{1}{n}Z'Z)^{-1}(\frac{1}{n}Z'X))^{-1}(\frac{1}{n}X'Z)(\frac{1}{n}Z'Z)^{-1}(\frac{1}{\sqrt{n}}Z'e) \\ &\xrightarrow{d} (Q_{xz}Q_{zz}^{-1}Q_{zx})^{-1}Q_{xz}Q_{zz}^{-1}N(0,\Omega) = N(0,V_\beta)\end{aligned}$$

as stated.

Proof that z_ie_i has a finite second moment:
By Minkowski's inequality,

$$\begin{aligned}(\mathbb{E}(e^4))^{1/4} &= (\mathbb{E}((y - x'\beta)^4))^{1/4} \\ &\leq (\mathbb{E}(y^4))^{1/4} + ||\beta||(\mathbb{E}||x||^4)^{1/4} < \infty\end{aligned}$$

under Assumptions 12.2.1 and 12.2.2. Then, by the CS inequality,

$$\mathbb{E}||ze||^2 \leq (\mathbb{E}||z||^4)^{1/2}(\mathbb{E}(e^4))^{1/2} < \infty$$

using Assumptions 12.2.3.

(j) Determinants of 2SLS variance

The homoskedastic formula for the asymptotic variance of the 2SLS estimator is:

$$\begin{aligned}V_\beta^0 &= (Q_{xz}Q_{zz}^{-1}Q_{zx})^{-1}\sigma^2 \\ &= (\mathbb{E}(x_iz_i')(\mathbb{E}(z_iz_i'))^{-1}\mathbb{E}(z_ix_i'))^{-1}\mathbb{E}(e_i^2)\end{aligned}$$

The variance decreases as:

i. The variance of e_i (the error in the y equation) decreases
ii. The variance of x_i increases
iii. The correlation between x_i and z_i increases

The variance is not affected by the variance/ covariance structure of z_i, and is invariant to rotations of z_i (including rescaling z_i). Intuitively, this is because z_i is in both the numerator and denominator of the formula for the asymptotic variance.
The asymptotic variance decreases as the number of instruments increases. However, the finite sample bias of the 2SLS estimator tends to increase as the number of instruments increases. There is thus a tradeoff between bias and variance.

(k) Covariance matrix estimation

The technique is similar to LS estimation.

$$\hat{V}_\beta = (\hat{Q}_{xz}\hat{Q}_{zz}^{-1}\hat{Q}_{zx})^{-1}(\hat{Q}_{xz}\hat{Q}_{zz}^{-1}\hat{\Omega}\hat{Q}_{zz}^{-1}\hat{Q}_{zx})(\hat{Q}_{xz}\hat{Q}_{zz}^{-1}\hat{Q}_{zx})^{-1}$$

where

$$\begin{aligned}\hat{Q}_{zz} &= \frac{1}{n}\sum_{i=1}^n z_iz_i' = \frac{1}{n}Z'Z \\ \hat{Q}_{xz} &= \frac{1}{n}\sum_{i=1}^n x_iz_i' = \frac{1}{n}X'Z \\ \hat{\Omega} &= \frac{1}{n}\sum_{i=1}^n z_iz_i'\hat{e}_i^2 \\ \hat{e}_i &= y_i - x_i'\hat{\beta}_{2sls}\end{aligned}$$

The homoskedastic variance matrix is estimated by

$$\hat{V}_\beta^0 = (\hat{Q}_{xz}\hat{Q}_{zz}^{-1}\hat{Q}_{zx})^{-1}\hat{\sigma}^2$$
$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n \hat{e}_i^2$$

Theorem 12.3: Under Assumption 12.2, as $n \to \infty$,

$$\hat{V}_\beta^0 \xrightarrow{p} V_\beta^0$$
$$\hat{V}_\beta \xrightarrow{p} V_\beta$$

In Stata, the *ivregress* command calculates the covariance matrix estimator (homoskedastic or heteroskedastic).

It is important that the covariance matrix be constructed using the correct residual formula:

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

This is different than what would be obtained if the "two-stage" computation method is used.
Specifically, for the two stage method:

i. First, we estimate the reduced form

$$x_i = \hat{\Gamma}' z_i + \hat{u}_i$$

to obtain the predicted values

$$\hat{x}_i = \hat{\Gamma}' z_i$$

ii. Second, we regress y_i on $\hat{x}_i$, giving

$$y_i = \hat{x}_i'\hat{\beta}_{2sls} + \hat{v}_i \qquad \text{(Eq 12.43)}$$

where $\hat{v}_i$ are LS residuals.

For the covariance matrix constructed using $\hat{v}_i$, the homoskedastic formula is

$$\hat{V}_\beta = (\frac{1}{n}\hat{X}'\hat{X})^{-1}\hat{\sigma}_v^2 = (\hat{Q}_{xz}\hat{Q}_{zz}^{-1}\hat{Q}_{zx})^{-1}\hat{\sigma}_v^2$$
$$\hat{\sigma}_v^2 = \frac{1}{n}\sum_{i=1}^n \hat{v}_i^2$$

which is proportional to $\hat{\sigma}_v^2$ rather than $\hat{\sigma}^2$.
This is important because $\hat{v}_i$ differs from $\hat{e}_i$:

$$\begin{aligned}\hat{v}_i &= y_i - x_i'\hat{\beta}_{2sls} + (x_i - \hat{x}_i)'\hat{\beta}_{2sls}\\ &= \hat{e}_i + \hat{u}_i'\hat{\beta}_{2sls}\\ &\neq \hat{e}_i\end{aligned}$$

This means that the covariance matrix estimator and standard errors based on (Eq 12.43) will be incorrect.

(l) Finite sample theory
This is not our focus for IV and 2SLS; even though the errors are normal, IV-type estimators are non-linear functions of these errors and thus the estimators are non-normally distributed in finite samples.

We will rely on asymptotic theory when we consider IV and 2SLS estimators.

(m) Control function regression

Control function regression (i.e., OLS with an auxillary regressor) is an alternative way of computing the 2SLS estimator by least-squares.
It allows us to use the entire variation in X, is useful in more complicated nonlinear contexts, and also in the linear model to construct tests for heterogeneity.

The structural and reduced-form equations for the standard IV model are:

$$\begin{aligned} y_i &= x_{1i}'\beta_1 + x_{2i}'\beta_2 + e_i \\ x_{2i} &= \Gamma_{12}' z_{1i} + \Gamma_{22}' z_{2i} + u_{2i} \end{aligned}$$

where x_{2i} is an endogenous regressor, $x_{1i} = z_{1i}$ are the included exogenous regressors, and z_{2i} are the excluded instruments.
Since the IV assumption specifies that $\mathbb{E}(z_i e_i) = 0$, x_{2i} is endogenous iff u_{2i} and e_i are correlated. This is because

$$\begin{aligned} \mathbb{E}[x_2 e] &= \Gamma_{12}' \mathbb{E}[z_1 e] + \Gamma_{22}' \mathbb{E}[z_2 e] + \mathbb{E}[u_2 e] \\ &= \mathbb{E}[u_2 e] \end{aligned}$$

The method of control function breaks down e_i into a component that is correlated with x_{2i} and another component that is uncorrelated with x_{2i}. Then, we can estimate the structural equation using OLS by explicitly controlling for the component that is correlated with x_{2i}, which is called a **control function**.

To obtain the control function, consider the linear projection of e_i on u_{2i}:

$$\begin{aligned} &e_i = u_{2i}'\alpha + \epsilon_i \\ &\alpha = (\mathbb{E}(u_{2i}u_{2i}'))^{-1}\mathbb{E}(u_{2i}e_i) \\ &\mathbb{E}(u_{2i}\epsilon_i) = 0 \end{aligned}$$

Substituting this into the structural form equation we find

$$\begin{aligned} y_i &= x_{1i}'\beta_1 + x_{2i}'\beta_2 + u_{2i}'\alpha + \epsilon_i \\ \mathbb{E}(x_{1i}\epsilon_i) &= \mathbb{E}(x_{1i}(e - u_{2i}'\alpha)) \\ &= \mathbb{E}[x_{1i}e - x_{1i}u_{2i}'\alpha] = 0 \\ \mathbb{E}(x_{2i}\epsilon_i) &= \mathbb{E}((\Gamma_{12}'z_i + \Gamma_{22}'z_2 + u_2)\epsilon) \\ &= \mathbb{E}[u_2\epsilon] = 0 \\ \mathbb{E}(u_{2i}\epsilon_i) &= 0 \end{aligned}$$

Notice that x_{2i} is uncorrelated with ϵ_i. This is because x_{2i} is correlated with e_i only through u_{2i}, and ϵ_i is the error after e_i has been projected orthogonal to u_{2i}.

We can estimate u_{2i} by the reduced-form residual:

$$\hat{u}_{2i} = x_{2i} - \hat{\Gamma}_{12}' z_{1i} - \hat{\Gamma}_{22}' z_{2i}$$

Then we can simply regress y_i on $x_{1i}, x_{2i}, \hat{u}_{2i}$. We write this as

$$y_i = x_i'\hat{\beta} + \hat{u}_{2i}'\hat{\alpha} + \hat{\epsilon}_i$$

or in matrix notation as

$$y = X\hat{\beta} + \hat{U}_2\hat{\alpha} + \hat{\epsilon}$$

This turns out to be algebraically identical to the 2SLS estimator (this only applies for linear models). In other words we can regress Y on the raw X with the addition of a control function, and still get the $\hat{\beta}_{2sls}$ coefficients.

(n) Endogeneity tests

The endogeneity tests test whether x_{2i} is an endogenous regressor (i.e., if IV is even necessary). The null and alternative hypotheses are

$$\begin{aligned} H_0 &: \mathbb{E}(x_{2i}e_i) = 0 \\ H_1 &: \mathbb{E}(x_{2i}e_i) \neq 0 \end{aligned}$$

Throughout we maintain the assumption that $\mathbb{E}(z_i e_i) = 0$ (i.e., instruments are valid), otherwise we have no benchmarks and can't test for the endogeneity of x_{2i}.

Recall the control function regression

$$\begin{aligned} y_i &= x'_{1i}\beta_1 + x'_{2i}\beta_2 + u'_{2i}\alpha + \epsilon_i \\ \alpha &= (\mathbb{E}(u_{2i}u'_{2i}))^{-1}\mathbb{E}(u_{2i}e_i) \end{aligned}$$

and $E(x_{2i}e_i) = 0$ iff $\mathbb{E}(u_{2i}e_i) = 0$

So the hypothesis can be restated as

$$\begin{aligned} H_0 &: \alpha = 0 \\ H_1 &: \alpha \neq 0 \end{aligned}$$

We can construct a Wald statistic W for $\alpha = 0$ in the control function regression. Under H_0, W is asymptotically chi-square with k_2 degrees of freedom (the number of endogenous regressors).

Theorem 12.14: Under $H_0, W \xrightarrow{d} \chi^2_{k_2}$. Let $c_{1-\alpha}$ solve $P(\chi^2_{k_2} \leq c_{1-\alpha} = 1 - \alpha)$. The test "Reject H_0 if $W > c_{1-\alpha}$" has asymptotic size α.

This can be computed in Stata using the *estat endogenous* command after *ivregress* when the latter uses a robust covariance option.

There is an alternative way to derive a test for endogeneity. Under H_0, both OLS and 2SLS are consistent estimators, but under H_1, they converge to different values. Thus we can use the difference between both estimators as a test statistic.

This is based on the "Hausman" equality-like result earlier (i.e., variance of difference = difference of variances).

We form a (homoskedastic) Wald statistic using the difference in the variances of the OLS and 2SLS estimators to construct the covariance matrix. Under H_0, it follows an exact F distribution (under strong assumptions). When there is only one parameter to be tested, this reduces to a t-statistic.

The general class of tests are called **Durbin-Wu-Hausman tests**.

(o) Overidentification tests

When $l > k$, the model is overidentified, meaning that there are more moments than free parameters. This is a restriction and is testable (i.e., we can test whether some of the instruments are invalid). Such tests are called **overidentification tests**.

The IV model specifies the $l \times 1$ moment conditions:

$$\mathbb{E}(z_i e_i) = 0$$

Equivalently,

$$\mathbb{E}(z_i y_i) - \mathbb{E}(z_i x'_i)\beta = 0$$

There are l equations but k unknowns, where $l > k$.

As an illustration, suppose there is a single endogenous regressor x_{2i} and two instruments z_{1i} and z_{2i} (there is no x_{1i}). Then, the model specifies that

$$\mathbb{E}(z_{1i}y_i) = \mathbb{E}(z_{1i}x_{2i})\beta$$

and

$$\mathbb{E}(z_{2i}y_i) = \mathbb{E}(z_{2i}x_{2i})\beta$$

Thus the scalar β has to solve both equations.

There may not be a single β that satisfies both equations, but we could solve β using only either one of the equations using an IV estimator.

If the overidentification hypothesis is correct, i.e., instruments are valid in the sense that

$$\mathbb{E}(z_i e_i) = 0$$

then we should expect that, in large samples, both IV estimators are consistent for β.
If the hypothesis is false, then both IV estimators will converge to different probability limits.

For a general overidentification test, the hypotheses are

$$H_0 : \mathbb{E}(z_i e_i) = 0$$
$$H_1 : \mathbb{E}(z_i e_i) \neq 0$$

For now we also add the conditional homoskedasticity assumption

$$\mathbb{E}(e_i^2|z_i) = \sigma^2$$

We will relax this assumption when we discuss the GMM approach.

Consider a linear projection of the error e_i on the instruments z_i

$$e_i = z_i'\alpha + \epsilon_i$$

with $\alpha = (\mathbb{E}(z_i z_i'))^{-1}\mathbb{E}(z_i e_i)$
We can now write H_0 as $\alpha = 0$. As e is not observed, replace it with 2SLS residuals $\hat{e}$, and estimate α by LS regression:

$$\hat{\alpha} = (Z'Z)^{-1}Z'\hat{e}$$

Sargan (1958) proposed testing H_0 via a **score test**, which takes the form

$$S = \hat{\alpha}'(v\hat{a}r(\hat{\alpha}))^{-\hat{\alpha}} = \frac{\hat{e}'Z(Z'Z)^{-1}Z'\hat{e}}{\hat{\sigma}^2}$$

where $\hat{\sigma}^2 = \frac{1}{n}\hat{e}'\hat{e}$.
This is an asymptotic test of the overidentifying restrictions under the assumption of conditional homoskedasticity.
Note: The "naive Wald" intuition for this test is that

$$\begin{aligned} W &= \hat{\alpha}'\hat{V}_{\hat{\alpha}}^{-1}\hat{\alpha} \\ &= \hat{e}'Z(Z'Z)^{-1}[(Z'Z)^{-1}\hat{\sigma}^2]^{-1}(Z'Z)^{-1}Z'\hat{e} \\ &= \frac{\hat{e}P_Z\hat{e}}{\hat{\sigma}^2} \end{aligned}$$

Note: A score test is based on the intuition that tests whether a derivative (a "score"; e.g., a FOC in a constrained likelihood function, or something similar) is equal to zero. When there are multiple constraints, the score is often expressed in quadratic form. If the constraints are zero, then the "score" should be close to zero.

Note: Under homoskedasticity, the score test and Wald test are equivalent.

The Sargan test rejects H_0 when $S > c$ for some critical value c. The result is summarised below:
Theorem 12.16: Under Assumption 12.2 and $\mathbb{E}(e_i^2|z_i) = \sigma^2$, then as $n \to \infty$

$$S \xrightarrow{d} \chi^2_{l-k}$$

where S is the Sargan test statistic.
For c satisfying $\alpha = 1 - G_{l-k}(c)$,

$$P(S > c|H_0) \to \alpha$$

so the test "Reject H_0 if $S > c$" has asymptotic size α.

The Sargan test can be implemented in Stata using the command *estat overid* after running *ivregress* $2sls$.

The Sargan statistic can be generalised to a GMM overidentification statistic, which allows heteroskedasticity (see Lecture 8).

8. **Lecture 8: Generalised Method of Moments**

   (a) Generalised Method of Moments: Introduction

   GMM generalises the classical **method of moments (MOM)** estimator by allowing for models that have more moment equations than unknown parameters and which are thus overidentified.

   GMM includes both linear and nonlinear models.

   All the models introduced so far can be written as **moment equation models**, where the population parameters solve a system of moment equations.

   Let $g_i(\beta)$ be a known $l \times 1$ function of the i^{th} observation and a $k \times 1$ parameter β. A moment equation model is summarised by the moment equations

   $$\mathbb{E}(g_i(\beta)) = 0 \qquad \text{(Eq 13.1)}$$

   and a parameter space $\beta \in B$.

   For example, in the IV model, $g_i(\beta) = z_i(y_i - x_i'\beta)$.

   In general, β is identified if there is a unique mapping from the data distribution to β.

   i. In (Eq 13.1), the data distribution is summarised by population moments. This means that there is a unique β satisfying (Eq 13.1).
   ii. It is necessary for $l \geq k$ for there to be a unique solution. In general, we assume $l \geq k$.
   iii. If $l = k$ the model is just identified.
   iv. If $l > k$ the model is over-identified.
   v. If $l < k$ the model is under-identified.

   (b) Method of moments estimators

   In this section we consider $l = k$ ("classical method of moments").

   We define the sample analog of (Eq 13.1):

   $$\bar{g}_n(\beta) = \frac{1}{n}\sum_{i=1}^{n} g_i(\beta) \qquad \text{(Eq 13.2)}$$

The **method of moments estimator (MME)** $\hat{\beta}_{mm}$ for β is defined as the parameter value which sets $\bar{g}_n(\beta) = 0$. Thus

$$\bar{g}_n(\hat{\beta}_{mm}) = \frac{1}{n}\sum_{i=1}^{n} g_i(\hat{\beta}_{mm}) = 0 \qquad \text{(Eq 13.3)}$$

(Eq 13.3) are known as the **estimating equations**.

In some contexts, there is a numerical solution. In other cases, the solution must be found numerically.

Examples:

i. Mean: Set $g_i(\mu) = y_i - \mu$. The MME is $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

ii. Mean and variance (2 parameters and 2 equations): Set

$$g_i(\mu, \sigma^2) = \begin{pmatrix} y_i - \mu \\ (y_i - \mu)^2 - \sigma^2 \end{pmatrix}$$

The MME are $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\mu})^2$.

iii. OLS: Set $g_i(\beta) = x_i(y_i - x_i'\beta)$. The MME is $\hat{\beta} = (X'X)^{-1}(X'y)$.

iv. OLS and variance: Set

$$g_i(\beta, \sigma^2) = \begin{pmatrix} x_i(y_i - x_i'\beta) \\ (y_i - x_i'\beta)^2 - \sigma^2 \end{pmatrix}$$

The MME is $\hat{\beta} = (X'X)^{-1}(X'y)$ and $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i'\hat{\beta})^2$.

v. IV: Set $g_i(\beta) = z_i(y_i - x_i'\beta)$. The MME is $\hat{\beta} = (\sum_{i=1}^{n} z_i x_i')^{-1}(\sum_{i=1}^{n} z_i y_i)$.

(c) Overidentified moment equations

In the IV model,

$$g_i(\beta) = z_i(y_i - x_i'\beta)$$

Thus (Eq 13.2) is

$$\bar{g}_n(\beta) = \frac{1}{n}\sum_{i=1}^{n} g_i(\beta) = \frac{1}{n}\sum_{i=1}^{n} z_i(y_i - x_i'\beta) = \frac{1}{n}(Z'y - Z'X\beta) \qquad \text{(Eq 13.4)}$$

If $l > k$, $Z'X$ is no longer a square matrix and hence is not invertible.
When the model is overidentified, there are more equations than parameters and hence there is no choice of β that sets (Eq 13.4) to zero. Thus we cannot apply the method of moments estimator.

Instead we can try to find an estimator that makes (Eq 13.4) as close to zero as possible.

One way of thinking about this is to define the vector $\mu = Z'y$ (which is observed), the matrix $G = Z'X$ (which is observed), and the "error" $\eta = \mu - G\beta$. We can then rewrite (Eq 13.4) as

$$\mu = G\beta + \eta$$

which looks like a regression with the $l \times 1$ variable μ, $l \times k$ regressor matrix G, and $l \times 1$ error vector η.
The goal is to make η as small as possible, i.e., minimise $\eta'\eta$.

A simple method to do this is to use least-squares regression of μ on G. The LS solution is

$$\hat{\beta} = (G'G)^{-1}(G'\mu)$$

which is analogous to the discussion of identification when $l > k$ in the IV framework.

More generally, when errors are non-homogenous it can be more efficient to estimate by generalised least squares. Thus for some weight matrix W, we consider the estimator

$$\begin{aligned}\hat{\beta} &= (G'WG)^{-1}(G'W\mu)\\ &= (X'ZWZ'X)^{-1}(X'ZWZ'y)\end{aligned}$$

which minimises the **weighted** sum of squares

$$\eta'W\eta$$

This solution is known as the generalised method of moments (GMM).

The estimator is typically defined as follows:
Given a set of moment equations

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

and an $l \times l$ weight matrix $W > 0$, the **GMM criterion function** is defined as

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

i. *Note: the factor n is not important for the definition of the estimator, but is convenient for distribution theory.*
ii. The criterion $J(\beta)$ is the weighted sum of squared moment equation errors.
iii. When $W = I_l$ (no weighting), then $J(\beta)$ is the simple sum of these squared errors:

$$J(\beta) = n \cdot \bar{g}_n(\beta)'\bar{g}_n(\beta) = n \cdot ||\bar{g}_n(\beta||^2$$

iv. $J(\beta) \geq 0$ because $W > 0$. When $l > k$, $J(\beta) > 0$.

The GMM estimator is defined as the minimiser of the GMM criterion $J(\beta)$.
Definition 13.1: The Generalised Method of Moments estimator is

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

Special case: When $l = k$, the method of moments estimator $\hat{\beta}_{gmm}$ yields $J_n(\beta_{mm}) = 0$. Therefore in this case, $\hat{\beta}_m m = \hat{\beta}_{gmm}$. Also, the weight matrix does not matter in this case.

In the following discussion, we focus on linear moment equations, particularly the overidentified IV model

$$g_i(\beta) = z_i(y_i - x_i'\beta) \qquad \text{(Eq 13.5)}$$

where z_i is $l \times 1$ and x_i is $k \times 1$.

The GMM method can be readily applied to nonlinear moment equations.

(d) GMM estimator

Given (Eq 13.5) and the sample analog (Eq 13.4), the GMM criterion can be written as

$$J(\beta) = n(Z'y - Z'X\beta)'W(Z'y - Z'X\beta).$$

The GMM estimator minimises $J(\beta)$. The FOCs are

$$\begin{aligned}0 &= \frac{\partial}{\partial\beta}J(\hat{\beta})\\ &= 2\frac{\partial}{\partial\beta}\bar{g}_n(\hat{\beta})'W\bar{g}_n(\hat{\beta})\\ &= -2(\frac{1}{n}X'Z)W(\frac{1}{n}Z'(y - X\hat{\beta}))\end{aligned}$$

The solution is given by:
Theorem 13.1: For the overidentified IV model

$$\hat{\beta}_{gmm} = (X'ZWZ'X)^{-1}(X'ZWZ'y) \qquad \text{(Eq 13.6)}$$

i. $(X'ZWZ'X)$ is invertible since $W > 0$ and $Z'X$ is full rank.
ii. The solution depends on W only up to scale (intuitively, W is in both the numerator and denominator). The multiplicative constant does not matter.
iii. When the weight matrix W is fixed by the user, $\hat{\beta}_{gmm}$ is called a **one-step GMM estimator**.

The GMM estimator (Eq 13.6) resembles the 2SLS estimator. In fact, they are equal when $W = (Z'Z)^{-1}$.
Theorem 13.2: If $W = (Z'Z)^{-1}$ then $\hat{\beta}_{gmm} = \hat{\beta}_{2sls}$.
Furthermore, if $k = l$ then $\hat{\beta}_{gmm} = \hat{\beta}_{iv}$.

(e) Distribution of GMM estimator

Let

$$Q = \mathbb{E}(z_i x_i')$$

and

$$\Omega = \mathbb{E}(z_i z_i' e_i^2) = \mathbb{E}(g_i g_i')$$

where $g_i = z_i e_i$. Then

$$(\frac{1}{n}X'Z)W(\frac{1}{n}Z'X) \xrightarrow{p} Q'WQ$$

and

$$(\frac{1}{n}X'Z)W(\frac{1}{\sqrt{n}Z'e}) \xrightarrow{d} Q'W \cdot N(0, \Omega)$$

The GMM estimator is asymptotically normal with a "sandwich form" asymptotic variance.
Theorem 13.3: Asymptotic distribution of GMM estimator.
Under Assumption 12.2, as $n \to \infty$,

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

where

$$V_\beta = (Q'WQ)^{-1}(Q'W\Omega WQ)(Q'WQ)^{-1} \qquad \text{(Eq 13.7)}$$

(f) Efficient GMM

The asymptotic variance V_β (Eq 13.7) depends on the weight matrix W.

The asymptotically optimal weight matrix W_0 is one which minimises V_β (in the positive definite sense).
This turns out to be $W_0 = \Omega^{-1}$ (because this lets us get rid of the Ω in the variance of β.

When we use $W = W_0 = \Omega^{-1}$, we call the estimator the **efficient GMM estimator**.

$$\hat{\beta}_{gmm} = (X'Z\Omega^{-1}Z'X)^{-1}(X'Z\Omega^{-1}Z'y)$$

The asymptotic variance becomes

$$V_\beta = (Q'\Omega^{-1}Q)^{-1}(Q'\Omega^{-1}\Omega\Omega^{-1}Q)(Q'\Omega^{-1}Q)^{-1} = (Q'\Omega^{-1}Q)^{-1}$$

Theorem 13.4: Asymptotic Distribution of GMM with Efficient Weight Matrix.
Under Assumption 12.2 and $W = \Omega^{-1}$, as $n \to \infty$

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

where

$$V_\beta = (Q'\Omega^{-1}Q)^{-1}$$

Theorem 13.5: Efficient GMM.
Under Assumption 12.2, for any $W > 0$,

$$(QW'Q)^{-1})(Q'W\Omega WQ)(Q'WQ)^{-1} - (Q'\Omega^{-1}Q)^{-1} > 0$$

Thus if $\hat{\beta}_{gmm}$ is the efficient GMM estimator and $\tilde{\beta}_{gmm}$ is another GMM estimator, then

$$avar(\hat{\beta}_{gmm}) \leq avar(\tilde{\beta}_{gmm})$$

In fact, as part of a more general result about efficiency, in the linear model no estimator has better asymptotic efficiency than the efficient linear GMM estimator.

As a special case, for the linear model we introduced the 2SLS estimator, which has weight matrix $\hat{W} = (Z'Z)^{-1}$ or equivalently $\hat{W} = (\frac{1}{n}Z'Z)^{-1}$ since scaling does not matter.
Since $\hat{W} \xrightarrow{p} (\mathbb{E}(z_i z_i'))^{-1}$ this is asymptotically equivalent to using the weight matrix $W = (\mathbb{E}(z_i z_i'))^{-1}$.

In contrast, the efficient weight matrix is $W_0 = \Omega^{-1} = (\mathbb{E}(z_i z_i' e_i^2))^{-1}$.

i. Under the assumption of homoskedasticity where $\mathbb{E}(e_i^2|z_i) = \sigma^2$, the efficient weight matrix becomes $W_0 = \Omega^{-1} = (\mathbb{E}(z_i z_i'))^{-1}\sigma^{-2}$, or equivalently $W = (\mathbb{E}(z_i z_i'))^{-1}$ since σ^2 is a constant and scaling does not matter.

Therefore we have:
Theorem 13.6: Under Assumption 12.2 and $\mathbb{E}(e_i^2|z_i) = \sigma^2$ then $\hat{\beta}_{2sls}$ is efficient GMM (i.e., under homoskedasticity, $\hat{\beta}_{2sls}$ is the efficient GMM estimator).

This also means that under heteroskedasticity, 2SLS is less efficient than the efficient GMM estimator.

(g) Estimation of the efficient weight matrix

To estimate the efficient weight matrix for the efficient GMM estimator, the convention is to form an estimate $\hat{\Omega}$ of Ω and then set $\hat{W} = \hat{\Omega}^{-1}$.

The **two-step GMM estimator proceeds** by using a one-step consistent estimate of β to construct the weight matrix estimator $\hat{W}$.

In the linear model, the natural one-step estimator for β is the 2SLS estimator $\hat{\beta}_{2sls}$.

We set $\tilde{e}_i = y_i - x_i'\beta_{2sls}$, $\tilde{g}_i = g_i(\tilde{\beta}) = z_i\tilde{e}_i$ and $\bar{g}_n = n^{-1}\sum_{i=1}^n \tilde{g}_i$.

Two moment estimators of the optimal weight matrix Ω are

$$\hat{\Omega} = \frac{1}{n}\sum_{i=1}^n \tilde{g}_i\tilde{g}_i' \qquad \text{(Eq 13.8)}$$

and

$$\hat{\Omega}^* = \frac{1}{n}\sum_{i=1}^n (\tilde{g}_i - \bar{g}_n)(\tilde{g}_i - \bar{g}_n)' \qquad \text{(Eq 13.9)}$$

(Eq 13.8) is an uncentered covariance matrix estimator, while (Eq 13.9) is the centered version. Both are consistent for Ω. When the model is just identified, $\bar{g}_n = 0$ and both are identical.

Given the choice of covariance matrix estimator we set $\hat{W} = \hat{\Omega}^{-1}$ or $\hat{W} = \hat{\Omega}^{*-1}$. Given this weight matrix, we then construct the two-step GMM estimator as (Eq 13.6) using the weight matrix $\hat{W}$.

The two-step GMM estimator is asymptotically efficient:
Theorem 13.7: Under Assumption 12.2 and $\Omega > 0$, if $\hat{W} = \hat{\Omega}^{-1}$ or $\hat{W} = \hat{\Omega}^{*-1}$ where the latter are defined in (Eq 13.8) and (Eq 13.9), then as $n \to \infty$

$$V_\beta = (Q'\Omega^{-1}Q)^{-1}$$

In Stata, the two-step GMM estimator can be obtained using the *ivregress gmm* command.

(h) Wald test

The Wald test can be constructed in the same way as done previously.

For a given function $r(\beta) : \mathbb{R}^k \to \Theta \subset \mathbb{R}^q$ we define the parameter $\theta = r(\beta)$. The GMM estimator of θ is $\hat{\theta}_{gmm} = r(\hat{\beta}_{gmm})$. By the delta method, it is asymptotically normal with covariance matrix

$$V_\theta = R'V_\beta R$$
$$R = \frac{\partial}{\partial\beta} r(\beta)'$$

An estimator of the asymptotic covariance matrix is

$$\hat{V}_\theta = \hat{R}'\hat{V}_\beta \hat{R}$$
$$\hat{R} = \frac{\partial}{\partial\beta} r(\hat{\beta}_{gmm})'$$

When θ is scalar then an asymptotic standard error for $\hat{\theta}_{gmm}$ is formed as $\sqrt{n^{-1}\hat{V}_\theta}$.

A standard test of the hypothesis

$$H_0 : \theta = \theta_0$$

against

$$H_1 : \theta \neq \theta_0$$

is based on the Wald statistic

$$W = n(\hat{\theta} - \theta_0)'\hat{V}_\theta^{-1}(\hat{\theta} - \theta_0)$$

Let $G_q(u)$ denote the χ^2_q distribution function.
Theorem 13.8: Under Assumptions 12.2 and 7.3, and if H_0 holds, then as $n \to \infty$,

$$W \xrightarrow{d} \chi^2_q$$

For c satisfying $\alpha = 1 - G_q(c)$,

$$P(W > c|H_0) \to \alpha$$

so the test "Reject H_0 if $W > c$" has asymptotic size α.

In Stata, the commands *test* and *testparm* can be used after *ivregress gmm* to implement Wald tests of linear hypotheses. The commands *nlcom* and *testnl* can be used after *ivregress gmm* to implement Wald tests of nonlinear hypotheses.

(i) Distance test

When the function $r(\beta)$ is nonlinear, it is better to use a criterion-based statistic (as the Wald test may perform poorly in finite samples).

This is sometimes called the **GMM distance statistic** and sometimes called a LR-like statistic.

The idea is to compare the unrestricted and restricted estimators by contrasting the criterion functions. The unrestricted estimator takes the form

$$\hat{\beta}_{gmm} = arg \min_{\beta} J(\beta)$$

where

$$\hat{J}(\beta) = n \cdot \bar{g}_n(\beta)'\hat{\Omega}^{-1}\bar{g}_n(\beta)$$

is the unrestricted GMM criterion with an efficient weight matrix estimate $\hat{\Omega}$. The minimised value of the criterion is

$$\hat{J} = \hat{J}(\hat{\beta}_{gmm})$$

The estimator subject to $r(\beta) = \theta_0$ is

$$\hat{\beta}_{cgmm} = arg \min_{r(\beta)=\theta_0} \tilde{J}(\beta)$$

where

$$\tilde{J}(\beta) = n \cdot \bar{g}_n(\beta)'\tilde{\Omega}^{-1}\bar{g}_n(\beta)$$

which depends on an efficient weight matrix estimate, either $\hat{\Omega}$ (the same as the unrestricted estimator) or $\tilde{\Omega}$ (the iterated weight matrix from constrained estimation). The minimised value of the criterion is

$$\tilde{J} = \tilde{J}(\hat{\beta}_{cgmm})$$

The GMM distance (or LR-like) statistic is the difference in the criterion functions.

$$D = \tilde{J} - \hat{J}$$

which is always nonnegative.
The distance test shares the useful feature of LR tests in that it is a natural by-product of the computation of alternative models.

Theorem 13.12: Under Assumptions 12.2 and 7.3, and if H_0 holds, then as $n \to \infty$,

$$D \xrightarrow{d} \chi_q^2$$

For c satisfying $\alpha = 1 - G_q(c)$,

$$P(D > c|H_0) \to \alpha$$

so the test "Reject H_0 if $D > c$" has asymptotic size α.

(j) Overidentification test

For the previously discussed Sargan test which assumes homoskedasticity, we can generalise the test to cover heteroskedasticity.

The null hypothesis is

$$H_0 : \mathbb{E}(z_i e_i) = 0$$

The sample analog

$$\bar{g}_n \xrightarrow{p} \mathbb{E}(z_i e_i)$$

and thus $\bar{g}_n$ can be used to assess whether the null hypothesis is true.

Assuming we use an efficient weight matrix estimate, the criterion function at the parameter estimates is

$$J = J(\hat{\beta}_{gmm}) = n\bar{g}_n'\hat{\Omega}^{-1}\bar{g}_n$$

This is a quadratic form in $\bar{g}_n$ and is thus a natural test statistic.

Theorem 13.14: Under Assumption 12.2, then as $n \to \infty$,

$$J = J(\hat{\beta}_{gmm}) \xrightarrow{d} \chi^2_{l-k}$$

For c satisfying $\alpha = 1 - G_{l-k}(c)$,

$$P(J > c)|H_0) \to \alpha$$

so the test "Reject H_0 if $J > c$" has asymptotic size α.

The degrees of freedom are the number of overidentifying restrictions.

It is advisable to report the statistic J whenever GMM is the estimation method.

In Stata, the command *estat overid* after *ivregress gmm* can be used to implement the test.

(k) Endogeneity test

Endogeneity tests are simple to implement in the GMM framework. The model is

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

where the maintained assumption is that the regressors x_{1i} and included instruments z_{2i} are exogenous so that $\mathbb{E}(x_{1i}e_i) = 0$ and $\mathbb{E}(z_{2i}e_i) = 0$. The question is whether or not x_{2i} is endogenous.

The null hypothesis is

$$H_0 : \mathbb{E}(x_{2i}e_i) = 0$$

and the alternative is

$$H_1 : \mathbb{E}(x_{2i}e_i) \neq 0$$

The GMM test is constructed as follows. First, estimate the model by efficient GMM using (x_{1i}, z_{2i}) as instruments for (x_{1i}, x_{2i}). Let $\hat{J}$ denote the resulting GMM criterion. The test statistic is the difference in the criterion functions:

$$C = \hat{J} - \tilde{J}$$

Theorem 13.16: Under Assumption 12.2 and if $\mathbb{E}(z_{2i}x_{2i}')$ has full rank k_2, then as $n \to \infty$,

$$C \xrightarrow{d} \chi^2_{k_2}$$

For c satisfying $\alpha = 1 - G_{k_2}(c)$,

$$P(C > c|H_0) \to \alpha$$

so the test "Reject H_0 if $C > c$ has asymptotic size α.

In Stata, the command *estat endogenous* after *ivregress gmm* can be used to implement the test.

9. **Lecture 9: Panel data**

(a) Introduction and notation

In **panel data** or **longitudinal data** there are multiple observations for each individual.
More broadly, panel data methods can be applied to any context with cluster-type dependence.

A typically maintained assumption for micro panels is that individuals are mutually dependent while the observations for a given individual are correlated across time periods.
This means that the observations follow a clustered dependence structure.
Current econometric practice is to use cluster-robust covariance matrix estimators when possible.

For asymptotics, we usually consider $N \to \infty$, T fixed for micro panels.

We index observations by both individual $i = 1, ..., N$ and the time period $t = 1, ..., T$.

In balanced panels, there are an equal number T of observations for each panel. The total number of observations is $n = NT$.
In unbalanced panels, we observe a total number T_i of observations for each individual. The total number of observations is $n = \sum_{i=1}^{N} T_i$.

We focus on panel data regression models whose observations are pairs (y_{it}, x_{it}) where y_{it} is the dependent variable and x_{it} is a k-vector of regressors.

We can cluster the observations at the level of the individual:

i. $\boldsymbol{y}_i : T_i \times 1$ stacked observations on y_{it} for $t \in S_i$
ii. $\boldsymbol{X}_i : T_i \times k$ matrix of stacked x'_{it} for $t \in S_i$

We can also use matrix notation for the full sample: Let $\boldsymbol{y} = (\boldsymbol{y}'_1, ..., \boldsymbol{y}'_N)'$ denote the $n \times 1$ vector of stacked $\boldsymbol{y}_i$, and set $\boldsymbol{X} = (\boldsymbol{X}'_1, ..., \boldsymbol{X}'_N)'$ similarly.

Note: from this section on, the vectors are no longer bolded for simplicity of notetaking, but the same notation is implied as in the previous sections.

(b) Pooled regression

The simplest model in panel regression is pooled regression:

$$\begin{aligned} y_{it} &= x'_{it}\beta + e_{it} \\ \mathbb{E}(x_{it}e_{it}) &= 0 \end{aligned} \tag{Eq 17.1}$$

where β is a $k \times 1$ coefficient vector and e_{it} is an error.
Note that by the projection assumption (Eq 17.1), β is the linear projection coefficient and e_{it} is the projection error.

The model can be written at the level of the individual as

$$\begin{aligned} y_i &= X_i\beta + e_i \\ \mathbb{E}(X'_i e_i) &= 0 \end{aligned}$$

Note: $\mathbb{E}(X'_i e_i) = \sum_{t=1}^{T} \mathbb{E}(x_{it}e_{it}) = 0$, which is implied by (Eq 17.1).
Note: e_i is $T_i \times 1$.

The equation for the full sample is

$$y = X\beta + e$$

where e is $n \times 1$ (recall that $n = NT$).

The standard estimator for β in the pooled regression model is least squares. This is essentially just OLS run on all observations in the data.

$$\begin{aligned}\hat{\beta}_{pool} &= (\sum_{i=1}^{N}\sum_{t\in S_i} x_{it}x_{it}')^{-1}(\sum_{i=1}^{N}\sum_{t\in S_i} x_{it}y_{it}) \\ &= (\sum_{i=1}^{N} X_i'X_i)^{-1}(\sum_{i=1}^{N} X_i'y_i) \\ &= (X'X)^{-1}(X'y)\end{aligned}$$

The vector of least-squares residuals for the i^{th} individual is $\hat{e}_i = y_i - X_i\hat{\beta}_{pool}$. While it is the conventional LS estimator, in the context of panel data it is called the **pooled regression estimator**.

By linearity and the cluster-level notation, we can write the estimator as

$$\begin{aligned}\hat{\beta}_{pool} &= (\sum_{i=1}^{N} X_i'X_i)^{-1}(\sum_{i=1}^{N} X_i'(X_i\beta + e_i)) \\ &= \beta + (\sum_{i=1}^{N} X_i'X_i)^{-1}(\sum_{i=1}^{N} X_i'e_i)\end{aligned}$$

Then,

$$\mathbb{E}(\hat{\beta}_{pool}|X) = \beta + (\sum_{i=1}^{N} X_i'X_i)^{-1}(\sum_{i=1}^{N} X_i'\mathbb{E}(e_i|X_i)) = \beta$$

if the following assumption holds for all i and t:

$$\mathbb{E}(e_{it}|X_i) = 0 \qquad \text{(Eq 17.2)}$$

This assumption is called **strict mean independence** of e_{it}:

i. This occurs when the errors e_{it} are mean independent of all regressors x_{ij} for all time periods $j = 1, ..., T$.
   x_{it} is exogenous in the sense discussed in the IV framework.

ii. This assumption is stronger than:
   A. Pairwise mean independence $\mathbb{E}(e_{it}|x_{it}) = 0$
   B. Projection assumption (Eq 17.1) $\mathbb{E}(e_{it}x_{it}) = 0$

iii. This assumption requires that neither lagged nor future values of x_{it} hep to forecast e_{it}.
   A. For example, it excludes lagged dependent variables (such as y_{it-1}) from x_{it} (no dynamics).
   B. Otherwise e_{it} can be correlated with x_{it+1}.
   C. This point is important in dynamic models. For example, suppose e_{it} is interpreted as a preference or income shock. It may affect future characteristics of the individual. Then, future characteristics are not really exogenous to current shocks.

This stronger assumption is the cost of allowing within-individual clustered dependence. If we assume that within-individual observations are mutually independent, then $\mathbb{E}(e_{it}|X_i) = \mathbb{E}(e_{it}|x_{it}) = 0$ by pairwise mean independence.

The pooled estimator is unbiased for β if strict mean independence holds.

i. If e_{it} is serially uncorrelated and homoskedastic, the covariance estimator takes a classical form (i.e., the OLS form; $\sigma^2(X'X)^{-1}$).

ii. If e_{it} is heteroskedastic but serially uncorrelated, we can use a heteroskedasticity-robust covariance matrix estimator.

In general, however, we expect the errors e_{it} to be correlated across time for a given individual. The conventional solution is to use a cluster-robust covariance matrix estimator, which allows arbitrary within-cluster dependence.

Cluster-robust covariance matrix estimators for pooled regression take the form

$$\hat{V}_{pool} = (X'X)^{-1}(\sum_{i=1}^{N} X_i'\hat{e}_i\hat{e}_i'X_i)(X'X)^{-1}$$

Similar to the clustering case, this can be multiplied by a DOF adjustment. The adjustment used by the Stata *regress cluster(id)* command (where *id* indicates the individual variable) is

$$\hat{V}_{pool} = \frac{(n-1)}{(n-k)}(\frac{N}{N-1})(X'X)^{-1}(\sum_{i=1}^{N} X_i'\hat{e}_i\hat{e}_i'X_i)(X'X)^{-1}$$

Note that these are the same as the formulas discussed in Lecture 3.

(c) One-way error component model

Instead of allowing arbitrary clustered dependence, we can explicitly model the correlation structure of the regression error e_{it}.

The most common choice is an **error-components structure**. The simplest form is

$$e_it = u_i + \epsilon_{it} \qquad \text{(Eq 17.4)}$$

where u_i is an individual-specific effect and ϵ_{it} are iid errors.

This is known as a **one-way error component model**.

In vector notation,

$$e_i = 1_i u_i + \epsilon_i$$

where 1_i is a $T_i \times 1$ vector of 1's.

The one-way error component regression model is

$$y_{it} = x_{it}'\beta + u_i + \epsilon_{it}$$

or

$$y_i = X_i\beta + 1_i u_i + \epsilon_i$$

(d) Random effects

The **random effects (RE) model** assumes that u_i and ϵ_{it} in (Eq 17.4) are conditionally mean zero, uncorrelated, and homoskedastic.

Assumption 17.1: (Random Effects). Model (17.4) holds with

$$\begin{aligned}
\mathbb{E}(\epsilon_{it}|X_i) &= 0 && \text{(mean zero errors)} && \text{(Eq 17.5)}\\
\mathbb{E}(\epsilon_{it}^2|X_i) &= \sigma_\epsilon^2 && \text{(homoskedastic errors)} && \text{(Eq 17.6)}\\
\mathbb{E}(\epsilon_{ij}\epsilon_{it}|X_i) &= 0 && \text{(serially uncorrelated errors)} && \text{(Eq 17.7)}\\
\mathbb{E}(u_i|X_i) &= 0 && \text{(mean zero individual effect)} && \text{(Eq 17.8)}\\
\mathbb{E}(u_i^2|X_i) &= \sigma_u^2 && \text{(homoskedastic individual effect)} && \text{(Eq 17.9)}\\
\mathbb{E}(u_i\epsilon_{it}|X_i) &= 0 && \text{(individual-specific effect and iid errors mutually uncorrelated)} && \text{(Eq 17.10)}
\end{aligned}$$

where (Eq 17.7) holds for all $j \neq t$.
Note that

$$\begin{aligned}
\mathbb{E}[\epsilon_i\epsilon_i'|X_i] &= \mathbb{E}[(1_iu_i + \epsilon_i)(1_iu_i + \epsilon_i)'|X_i]\\
&= 1_i1_i'\sigma_u^2 + I_i\sigma_\epsilon^2 \qquad \text{(since } \mathbb{E}(u_i\epsilon_{it}|X_i) = 0)
\end{aligned}$$

The random effects specification in Assumption 17.1 implies that the vector of errors e_i for individual i has the covariance structure

$$\begin{aligned}
\mathbb{E}(e_i|X_i) &= 0 \\
\mathbb{E}(e_i e_i'|X_i) &= 1_i 1_i' \sigma_u^2 + I_i \sigma_\epsilon^2 \\
&= \begin{pmatrix} \sigma_u^2 + \sigma_\epsilon^2 & \sigma_u^2 & ... & \sigma_u^2 \\ \sigma_u^2 & \sigma_u^2 + \sigma_\epsilon^2 & ... & \sigma_u^2 \\ ... & ... & ... & ... \\ \sigma_u^2 & \sigma_u^2 & ... & \sigma_u^2 + \sigma_\epsilon^2 \end{pmatrix} \\
&= \sigma_\epsilon^2 \Omega_i
\end{aligned}$$

where I_i is an identity matrix of dimension T_i. Note that

$$\Omega_i = I_i + 1_i 1_i' \sigma_u^2 / \sigma_\epsilon^2$$

Also, if $\sigma_u^2 = 0$, this is just OLS without any cluster dependence.

The RE model is equivalent to an **equi-correlation** model:

$$\begin{aligned}
\mathbb{E}(e_{it}|X_i) &= 0 \\
\mathbb{E}(e_{it}^2|X_i) &= \sigma^2 \\
\mathbb{E}(e_{ij} e_{it}|X_i) &= \mathbb{E}((u_i + \epsilon_{ij})(u_i + \epsilon_{it})|X_i) \\
&= \sigma_u^2 \\
&= \rho\sigma^2 \qquad \text{for } j \neq t
\end{aligned}$$

This arises by writing

$$\begin{aligned}
&\sigma_u^2 = \rho\sigma^2 \\
&\sigma_\epsilon^2 = (1-\rho)\sigma^2 \qquad \text{where } \sigma^2 := \mathbb{E}(e_{it}^2|X_i)
\end{aligned}$$

and hence

$$\rho = corr(e_{ij}, eit) = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_\epsilon^2} < 1$$

Given the error structure, OLS is not efficient and the natural estimator for β is GLS.
Suppose σ_u^2 and σ_ϵ^2 are known. The GLS estimator is:

$$\hat{\beta}_{gls} = (\sum_{i=1}^{N} X_i' \Omega_i^{-1} X_i)^{-1} (\sum_{i=1}^{N} X_i' \Omega_i^{-1} y_i)$$

Under Assumption 17.1, $\hat{\beta}_{gls}$ is unbiased for β:

$$\begin{aligned}
\hat{\beta}_{gls} - \beta &= (\sum_{i=1}^{N} X_i' \Omega_i^{-1} X_i)^{-1} \\
\mathbb{E}(\hat{\beta}_{gls} - \beta|X) &= (\sum_{i=1}^{N} X_i' \Omega_i^{-1} X_i)^{-1} (\sum_{i=1}^{N} X_i' \Omega_i^{-1} \mathbb{E}(e_i|X_i) = 0
\end{aligned}$$

The variance of $\hat{\beta}_{gls}$ is

$$V_{gls} = (\sum_{i=1}^{n} X_i' \Omega_i^{-1} X_i)^{-1} \sigma^2 \epsilon \qquad \text{(Eq 17.11)}$$

Comparing $\hat{\beta}_{gls}$ with $\hat{\beta}_{pool}$, under Assumption 17.1 the latter is also unbiased for β and has variance

$$V_{pool} = (\sum_{i=1}^{n} X_i' X_i)^{-1} (\sum_{i=1}^{n} X_i' \Omega_i X_i) (\sum_{i=1}^{n} X_i' X_i)^{-1} \qquad \text{(Eq 17.12)}$$

Under Assumption 17.1,

$$V_{gls} \leq V_{pool}$$

and hence the random effects estimator is more efficient than the pooled estimator. The two variance matrices are identical when there is no individual-specific effect ($\sigma_u^2 = 0$). In this case,

$$V_{gls} = V_{pool} = (\sum_{i=1}^{n} X_i' X_i)^{-1} \sigma_\epsilon^2 \qquad \text{(when } \Omega = I\text{)}$$

Under the assumption that the random effects model is a useful approximation but not literally true, we may consider a cluster-robust covariance matrix estimator such as

$$\hat{V}_{gls} = (\sum_{i=1}^{N} X_i' \Omega_i^{-1} X_i)^{-1} (\sum_{i=1}^{N} X_i' \Omega_i^{-1} \hat{e}_i \hat{e}_i' \Omega_i^{-1} X_i)(\sum_{i=1}^{n} X_i' \Omega_i^{-1} X_i)^{-1} \qquad \text{(Eq 17.14)}$$

where $\hat{e}_i = y_i - X_i \hat{\beta}_{gls}$. This may be rescaled by a DOF adjustment if desired.

The RE estimator can be obtained using the Stata command *xtreg*. The default covariance matrix estimator is (Eq 17.11). To use (17.14) we can use the command *xtreg vce*(*robust*).

(e) Fixed effects

Again, consider the one-way error component regression model.

In many applications it is useful to interpret the individual-specific effect u_i as a time-invariant unobserved missing variable.
It is then natural to expect u_i to be correlated with the regressors x_{it}.

If the stochastic structure of u_i is treated as unknown and possibly correlated with x_{it} then u_i is called a **fixed effect**.

There will be issues of omitted variable bias and endogeneity. Not controlling for u_i will result in biased and inconsistent estimates.

With an unstructured individual effect u_i, a sufficient condition for the identification of β is:
Definition 17.1: The regressor x_{it} is strictly exogenous for the error ϵ_{it} if

$$\mathbb{E}(x_{is}\epsilon_{it}) = 0 \qquad \text{(Eq 17.17)}$$

for all $s = 1, ..., T$.

This is a strong (projection) condition, meaning that all past, current and future ϵ_{it} are orthogonal to the regressors.
Again this rules out dynamics, and is not appropriate for dynamic models.
We can consider estimation moethods under the weaker assumption of "predetermined regressors": $\mathbb{E}(x_{it-s}\epsilon_{it}) = 0$ for all $s \geq 0$ (not covered here).

(Eq 17.17) is a projection analog of strict mean independence of ϵ_{it}:

$$\mathbb{E}(\epsilon_{it}|X_i) = 0 \qquad \text{(Eq 17.18)}$$

Note that (Eq 17.18) implies (Eq 17.17).

Note that we impose an assumption on ϵ_{it}, which is not the same as e_{it}, which is defined as $e_{it} := u_i + \epsilon_{it}$.
No assumption is made on u_i.
Therefore, (Eq 17.17) is much weaker than (Eq 17.2) or Assumption 17.1.

(f) Within transformation

When the relationship between u_i and x_{it} is fully unstructured, the only way to consistently estimate β is by an estimator which is invariant to u_i.

One transformation to eliminate u_i is the **within transformation**.
Define the mean of a variable for a given individual as

$$\bar{y}_i = \frac{1}{T_i}\sum_{t\in S_i} y_{it}$$

The within transformation is (note the dot notation):

$$\dot{y}_{it} = y_{it} - \bar{y}_i$$

We refer to $\dot{y}_{it}$ as demeaned values.

We can write the individual-specific mean as

$$\bar{y}_i = (1_i'1_i)^{-1}1_i'y_i$$

(analogous to projecting y on the intercept in Lectures 2 and 3).

Stacking the observations we can write

$$\begin{aligned}\dot{y}_i &= y_i - 1_i\bar{y}_i \\ &= y_i - 1_i(1_i'1_i)^{-1}1_i'y_i \\ &= M_i y_i\end{aligned}$$

where

$$M_i = I_i - 1_i(1_i'1_i)^{-1}1_i'$$

is the individual-specific demeaning operator.

Similarly we can define

$$\begin{aligned}\bar{x}_i &= \frac{1}{T_i}\sum_{t\in S_i} x_{it} \\ \dot{x}_{it} &= x_{it} - \bar{x}_i \\ \dot{X}_i &= M_i X_i\end{aligned}$$

Applying these operations to the one-way error component regression model

$$y_{it} = x_{it}'\beta + u_i + \epsilon_{it} \qquad \text{(Eq 17.15)}$$

we have

$$\bar{y}_i = \bar{x}_i'\beta + u_i + \bar{\epsilon}_i$$

where $\bar{\epsilon}_i = \frac{1}{T_i}\sum_{t\in S_i}\epsilon_{it}$. Subtracting from (Eq 17.15), we have

$$\dot{y}_{it} = \dot{x}_{it}'\beta + \dot{\epsilon}_{it} \qquad \text{(Eq 17.21)}$$

where $\dot{\epsilon}_{it} = \epsilon_{it} - \bar{\epsilon}_{it}$ and the individual effect u_i has been eliminated.

We can write this in vector notation as well. From the model

$$y_i = X_i\beta + 1_i u_i + \epsilon_i \qquad \text{(Eq 17.16)}$$

we can apply the demeaning operator M_i to both sides and obtain

$$\dot{y}_i = \dot{X}_i\beta + \dot{\epsilon}_i \qquad \text{(Eq 17.22)}$$

The individual-specific effect u_i has been eliminated since $M_i 1_i = 0$. (Eq 17.22) is a vector version of (Eq 17.21).

Because u_i is eliminated, estimators constructed from (Eq 17.21) and (Eq 17.22) will be invariant to the values of u_i.
One consequence is that all time-invariant regressors ($x_{it} = x_i$) are also eliminated.
However, we can introduce some structure, e.g., assume that some time-invariant regressors are orthogonal to u_i, and $\mathbb{E}(z_i u_i) = 0$. This allows the effects of time-invariant regressors to be estimated.

Another consequence is that the within transformation can greatly reduce the variance of the regressors. This may reduce the precision of the estimators.

(g) Fixed effects estimator

The **fixed-effects (FE) estimator** or the **within estimator** applies least-squares to the demeaned equation (Eq 17.21 or 17.22).

$$\begin{aligned}\hat{\beta}_{fe} &= (\sum_{i=1}^{N}\sum_{t\in S_i} \dot{x}_{it}\dot{x}_{it}')^{-1}(\sum_{i=1}^{N}\sum_{t\in S_i} \dot{x}_{it}\dot{y}_{it})\\ &= (\sum_{i=1}^{N} \dot{X}_i'\dot{X}_i)^{-1}(\sum_{i=1}^{N} \dot{X}_i'\dot{y}_i)\\ &= (\sum_{i=1}^{N} X_i'M_iX_i)^{-1}(\sum_{i=1}^{N} X_i'M_iy_i)\end{aligned}$$

The above implicitly assumes that the matrix

$$\sum_{i=1}^{N} \dot{X}_i'\dot{X}_i$$

is full rank, where $\dot{X}_i = M_iX_i$. This rules out time-invariant regressors.

The FE residuals are

$$\begin{aligned}\hat{e}_{it} &= \dot{y}_{it} - \dot{x}_{it}'\hat{\beta}_{fe}\\ \hat{e}_i &= \dot{y}_i - \dot{X}_i\hat{\beta}_{fe} \qquad \text{(Eq 17.23)}\end{aligned}$$

Under the strict mean independence assumption of ϵ_{it} (Eq 17.18), by linearity and the fact that $M_i 1_i = 0$, we can write

$$\hat{\beta}_{fe} - \beta = (\sum_{i=1}^{N} X_i'M_iX_i)^{-1}(\sum_{i=1}^{N} X_i'M_i\epsilon_i)$$

which gives

$$\mathbb{E}(\hat{\beta}_{fe} - \beta|X) = (\sum_{i=1}^{N} X_i'M_iX_i)^{-1}(\sum_{i=1}^{N} X_i'M_i\mathbb{E}(\epsilon_i|X_i)) = 0$$

Thus the FE estimator is unbiased for β.

Let

$$\Sigma_i = \mathbb{E}(\epsilon_i\epsilon_i'|X)$$

denote the $T_i \times T_i$ conditional covariance matrix of the idiosyncratic errors. The variance of $\hat{\beta}_{fe}$ is

$$V_{fe} = var(\hat{\beta}_{fe}|X) = (\sum_{i=1}^{N} \dot{X}_i'\dot{X}_i)^{-1}(\sum_{i=1}^{N} \dot{X}_i'\Sigma_i\dot{X}_i)(\sum_{i=1}^{N} \dot{X}_i'\dot{X}_i)^{-1} \qquad \text{(Eq 17.24)}$$

This expression simplifies when the idiosyncratic errors are homoskedastic and serially uncorrelated:

$$\mathbb{E}(\epsilon_{it}^2|X_i) = \sigma_\epsilon^2 \quad \text{(Eq 17.25)}$$
$$\mathbb{E}(\epsilon_{ij}\epsilon_{it}|X_i) = 0 \quad \text{(Eq 17.26)}$$

for all $j \neq t$. In this case, $\Sigma_i = I_i\sigma_\epsilon^2$ and (Eq 17.24) simplies to

$$V_{fe}^0 = \sigma_\epsilon^2(\sum_{i=1}^N \dot{X}_i'\dot{X}_i)^{-1} \quad \text{(Eq 17.27)}$$

Suppose there are no individual-specific effects ($u_i = 0$) so that both the FE and pooled estimators are unbiased for β. Under (Eq 17.25) and (Eq 17.26), we have

$$V_{fe}^0 = \sigma_\epsilon^2(\sum_{i=1}^N \dot{X}_i'\dot{X}_i)^{-1} \geq \sigma_\epsilon^2(\sum_{i=1}^N X_i'X_i)^{-1} = V_{pool}$$

i.e., the FE estimator is less efficient than the pooled estimator.
Intuitively, this is because the demeaned regressors have less variation than the original regressors. We can also show this by

$$\begin{aligned}\sum X_i'X_i - \sum \dot{X}_i'\dot{X}_i &= \sum(X_i - M_iX_i)'(X_i - M_iX_i)\\ &= \sum(P_iX_i)'(P_iX_i)\\ &= \sum X_i'P_iX_i \geq 0\end{aligned}$$

In Stata, the FE estimator is implemented by *xtreg fe*. The cluster-robust covariance matrix estimator can be obtained by the option *vce(robust)*.

(h) Differenced estimator

The **first-differencing transformation** also eliminates the individual-specific effect:

$$\Delta y_{it} = y_{it} - y_{it-1}$$

This can be applied to all but the first observation.
One advantage of FD is that we don't need to impose strict exogeneity of ϵ_{it} (as was the case for (Eq 17.17)) and hence can include some dynamics.

At the level of the individual this can be written as

$$\Delta y_i = D_i y_i$$

where D_i is the $(T_i \times 1) \times T_i$ matrix differencing operator

$$D_i = \begin{bmatrix} -1 & 1 & 0 & ... & 0 & 0\\ 0 & -1 & 1 & ... & 0 & 0\\ ... & & & & & \\ 0 & 0 & 0 & ... & -1 & 1 \end{bmatrix}$$

Applying the transformation to the one-way error components regression model, we obtain

$$\Delta y_{it} = \Delta x_{it}'\beta + \Delta\epsilon_{it}$$
$$\Delta y_i = \Delta X_i\beta + \Delta\epsilon_i \quad \text{(Eq 17.29)}$$

The **FD estimator** is the LS estimator applied to the differenced equation:

$$\begin{aligned}\hat{\beta}_\Delta &= (\sum_{i=1}^N\sum_{t\geq 2}\Delta x_{it}\Delta x_{it}')^{-1}(\sum_{i=1}^N\sum_{t\geq 2}\Delta x_{it}\Delta y_{it})\\ &= (\sum_{i=1}^N \Delta X_i'\Delta X_i)^{-1}(\sum_{i=1}^N \Delta X_i'\Delta y_i)\\ &= (\sum_{i=1}^N X_i'D_i'D_iX_i)^{-1}(\sum_{i=1}^N X_i'D_i'D_iy_i) \quad \text{(Eq 17.30)}\end{aligned}$$

Note that D is not symmetric and not idempotent.

When $t = 2$, the FD and FE estimators are equal, but they differ for $T > 2$. Under the assumption that ϵ_{it} is iid, we can show that the FE estimator is more efficient than the FD estimator.

(i) Dummy variables regression

An alternative way to estimate the FE model is by LS of y_i on x_{it} and a full set of dummy variables, one for each individual in the sample.
This is algebraically equivalent to the FE/ within estimator.

Consider the error-component model with regressors, which can be written as

$$y_{it} = x_{it}'\beta + d_i'u + \epsilon_{it}$$

where d_i is a $(1 \times N)$ vector where all entries are zeros and the i^{th} element is 1 (e.g., $[0\ 0\ 0\ 1\ 0\ 0]'$), and $u_i = d_i'u$.

In matrix notation,

$$y = X\beta + Du + \epsilon$$

Where D is the $n \times N$ matrix obtained by stacking all the vectors d_i.

We estimate (β, u) by LS, and write the estimates as

$$y = X\hat{\beta} + D\hat{u} + \hat{\epsilon}$$

We call this the **dummy variable estimator** of the fixed effects model.

By the FWL theorem, the dummy variable estimator $\hat{\beta}$ and residuals $\hat{\epsilon}_{it}$ can be obtained by:

i. Regress y on D, obtain residuals
ii. Regress X on D, obtain residuals
iii. Regress residuals from (i) on residuals from (ii)

Since (i) is a within transformation for y, and (ii) is a within transformation for X, this is equivalent to the FE estimator.

Theorem 17.1: The fixed effects estimator of β is algebraically identical to the dummy variable estimator of β. The two estimators also yield the same residuals.

When the number of individuals N is large, the within transformation is easier for computation than the dummy variable approach.

(j) Fixed effects covariance matrix estimation

When ϵ_{it} is serially uncorrelated and homoskedastic, the covariance matrix estimator is

$$\hat{V}_{fe}^{0} = \hat{\sigma}_{\epsilon}^2 (\dot{X}'\dot{X})^{-1} \qquad \text{(Eq 17.36)}$$

with

$$\hat{\sigma}_{\epsilon}^2 = \frac{1}{n - N - k} \sum_{i=1}^{n} \sum_{t \in S_i} \hat{\epsilon}_{it}^2 = \frac{1}{n - N - k} \sum_{i=1}^{N} \hat{\epsilon}_i' \hat{\epsilon}_i$$

The $N + k$ degrees of freedom adjustment is motivated by the dummy variable representation. Indeed, we can verify that $\hat{\sigma}_{\epsilon}^2$ is unbiased for σ_{ϵ}^2 under (Eq 17.18, 17.25, 17.26).

In the general case where ϵ_{it} is serially correlated and heteroskedastic, we can use the cluster-robust covariance matrix estimator:

$$\hat{V}_{fe}^{cluster} = (\dot{X}'\dot{X})^{-1} (\sum_{i=1}^{N} \dot{X}_i' \hat{\epsilon}_i \hat{\epsilon}_i' \dot{X}_i) (\dot{X}'\dot{X})^{-1} \qquad \text{(Eq 17.38)}$$

Potential DOF adjustments include:

$$\hat{V}_{fe}^{cluster} = (\frac{N}{N-1})(\dot{X}'\dot{X})^{-1}(\sum_{i=1}^{N}\dot{X}_i'\hat{\epsilon}_i\hat{\epsilon}_i'\dot{X}_i)(\dot{X}'\dot{X})^{-1} \qquad \text{(Eq 17.39)}$$

$$\hat{V}_{fe}^{cluster} = (\frac{n-1}{n-N-k})(\frac{N}{N-1})(\dot{X}'\dot{X})^{-1}(\sum_{i=1}^{N}\dot{X}_i'\hat{\epsilon}_i\hat{\epsilon}_i'\dot{X}_i)(\dot{X}'\dot{X})^{-1} \qquad \text{(Eq 17.40)}$$

Note that in typical micropanel applications, N is very large and k is modest, thus the adjustment in (Eq 17.39) is minor, while that in (Eq 17.40) can be substantial when T is small.

(k) Other issues about the FE estimator

Intercept
Some packages may report an intercept estimator from FE estimation. For example, in $xtreg,\ fe$ it is defined as

$$\hat{\alpha} = \bar{y} - \bar{x}'\hat{\beta}_{fe}$$

where $\bar{y}$ and $\bar{x}$ are averages from the full sample.

There is no clear interpretation or use for this estimate. It may be better to focus on the slope coefficients.

Estimation of fixed effects
The fixed effects themselves may be interesting. For example, we may want to measure the distribution of u_i to understand its heterogeneity, or do prediction.

The LS estimates of the fixed effects (based on the dummy variables regressions) can be obtained from the individual-specific means:

$$\hat{u}_i = \frac{1}{T_i}\sum_{i=1}^{N}(y_{it} - x_i'\hat{\beta}_{fe}) = \bar{y} - \bar{x}_i'\hat{\beta}_{fe} \qquad \text{(Eq 17.51)}$$

and does not require a regression with $N + k$ regressors.

If an intercept has been estimated, it should be subtracted from (Eq 17.51). In this case the estimated fixed effects are

$$\hat{u}_i = \bar{y} - \bar{x}_i'\hat{\beta}_{fe} - \hat{\alpha}_{fe}$$

Note that when the number of time series observations T_i is small, $\hat{u}_i$ will be an imprecise estimator of u_i and hence calculations based on $\hat{u}_i$ should be interpreted cautiously.

(l) Asymptotic distribution of the FE estimator

Assumption 17.2:

i. $y_{it} = x_{it}'\beta + u_i + \epsilon_{it}$ for $i = 1, ..., N$ and $t = 1, ..., T$ with $T \geq 2$
ii. The variables (ϵ_i, X_i), $i = 1, ..., N$ are iid
iii. $\mathbb{E}(x_{is}\epsilon_{it}) = 0$ for all $s = 1, ..., T$ (i.e., strict exogeneity)
iv. $Q_T = \mathbb{E}(\dot{X}_i'\dot{X}_i) > 0$ (i.e., the design demeaned matrix is positive definite; this is required for identification)
v. $\mathbb{E}(\epsilon_{it}^4) < \infty$
vi. $\mathbb{E}||x_{it}||^4 < \infty$

Theorem 17.2: Under Assumption 17.2, as $N \to \infty$,

$$\sqrt{N}(\hat{\beta}_{fe} - \beta) \xrightarrow{d} N(0, V_\beta)$$

where

$$V_\beta = Q_T^{-1}\Omega_T Q_T^{-1}$$
$$\Omega_T = \mathbb{E}(\dot{X}_i'\epsilon_i\epsilon_i'\dot{X}_i)$$

This asymptotic distribution is derived as the number of individuals N diverges to infinity while the number of time periods T is held fixed.
Intuitively:

i. FE regression is estimating $N+k$ coefficients (in dummy variables regression)
ii. The theory specfies that $N \to \infty$
iii. The number of estimated parameters diverges to infinity at the same rate as sample size. The way we do this is that we purge the N coefficients using a transformation, so we only deal with k coefficients.

Formally, the assumptions imply that the variables $(\dot{X}_i, \epsilon_i)$ are iid across i and have finite fourth moments. Thus by the WLLN

$$\frac{1}{N}\sum_{i=1}^{N}\dot{X}_i'\dot{X}_i \xrightarrow{p} \mathbb{E}(\dot{X}_i'\dot{X}_i) = Q_T$$

The random vectors $\dot{X}_i'\epsilon_i$ are iid. Assumption 17.2.3 implies

$$\mathbb{E}(\dot{X}_i'\epsilon_i) = \sum_{t=1}^{T}\mathbb{E}(\dot{x}_{it}\epsilon_{it}) = \sum_{t=1}^{T}\mathbb{E}(x_{it}\epsilon_{it}) - \sum_{t=1}^{T}\sum_{j=1}^{T}\mathbb{E}(x_{ij}\epsilon_{it}) = 0 - 0 = 0$$

because of strict exogeneity, so the random vectors $\dot{X}_i'\epsilon_i$ are mean zero.

Assumptions (17.2.5-6) imply that $\dot{X}_i'\epsilon_i$ has a finite covariance matrix, which is Ω_T. The assumptions for the CLT (Theorem 6.11) hold, thus

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\dot{X}_i'\epsilon_i \xrightarrow{d} N(0, \Omega_T)$$

Together we find

$$\sqrt{N}(\hat{\beta}_{fe} - \beta) = (\frac{1}{N}\sum_{i=1}^{N}\dot{X}_i'\dot{X}_i)^{-1}(\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\dot{X}_i'\epsilon_i) \xrightarrow{d} Q_T^{-1}N(0,\Omega_T) = N(0, V_\beta)$$

(m) Hausman test for random vs fixed effects

The RE model is a special case of the FE model. We can test the null hypothesis of random effects versus the alternative of fixed effects.

Under the null, the RE estimator is more efficient than the FE estimator. The Hausman statistic is

$$H = (\hat{\beta}_{fe} - \hat{\beta}_{re})'v\hat{a}r(\hat{\beta}_{fe} - \hat{\beta}_{re})^{-1}(\hat{\beta}_{fe} - \hat{\beta}_{re})$$
$$= (\hat{\beta}_{fe} - \hat{\beta}_{re})'(\hat{V}_{fe} - \hat{V}_{re})^{-1}(\hat{\beta}_{fe} - \hat{\beta}_{re})$$

where both $\hat{V}_{fe}$ and $\hat{V}_{re}$ take the classical (non-robust) form.

An asymptotic $100\alpha\%$ test rejects if H exceeds the $1-\alpha^{th}$ quantile of the χ^2_k distribution, where $k = dim(\beta)$. If the test rejects, this is evidence that the individual effect u_i is correlated with the regressors, so the random effects model is not appropriate. If the test fails to reject, the evidence says that the random effects hypothesis cannot be rejected.

The test can be implemented on a subset of the coefficients β.

i. E.g., the regressors contain some time-invariant elements so that the RE estimator contains more coefficients than the FE estimator.

ii. In this case, the test should be implemented only on the coefficients on the time-varying regressors.

However, it is not advised to use the outcome from the Hausman test to select whether we use the FE or RE estimator. This approach is known as a **pretest estimator** and is biased because the result of the test is random and correlated with the estimators.

Current econometric practice is to prefer robustness over efficiency, and random effects are only used in contexts where FE estimation is unknown or challenging (e.g., nonlinear models).

(n) Time trends

In general we expect that economic agents will experience common shocks during the same time period, and it is often desirable to include time effects in a panel regression model.

The simplest specification is a linear time trend:

$$y_i t = x_i' t \beta + \gamma t + u_i + \epsilon_{it}$$

More flexible specifications (e.g., quadratic) can be used.

For estimation, it is appropriate to include the time trend t as an element of the regressor vector x_{it} and then apply fixed effects.

The time trends may be individual-specific. A linear time trend specification only extracts a common time trend. We can include an interaction effect such as

$$y_{it} = x_{it}'\beta + \gamma_i t + u_i + \epsilon_{it}$$

In a FE specification, (γ_i, u_i) are treated as possibly correlated with the regressors.

We can perform fixed-effect estimation using the FWL approach.

i. First, for each individual, regress x_{it} (y_{it}) on t to obtain individual-level detrended observations (residuals) $\dot{x}_{it}$ $(\dot{y}_{it})$:

$$x_{it} = \hat{\alpha}_i + \hat{\gamma}_i t + \dot{x}_{it}$$

ii. Then, perform LS regression:

$$\hat{\beta}_{FE} = (\dot{X}'\dot{X})^{-1})(\dot{X}'\dot{y})$$

where the elements of $\dot{X}$ and $\dot{y}$ are the individual-level detrended observations.

The functional forms imposed by linear time trends are very limiting. We can consider the most flexible specification where the trend is allowed to take any arbitrary shape, but will require that it is common rather than individual-specific (otherwise it cannot be identified).

The model is a **two-way error component model**:

$$y_{it} = x_{it}'\beta + v_t + u_i + \epsilon_{it} \tag{Eq 17.63}$$

where v_t is an unobserved time-specific effect.

The **two-way within transformation** is used.
Define the time-specific mean at time t as

$$\tilde{y}_t = \frac{1}{N_t}\sum_{i \in S_t} y_{it}$$

where N_t is the number of individuals at time t.

For the case of balanced panels we compute the two-way within transformation as

$$\ddot{y}_{it} = y_{it} - \bar{y}_i - \tilde{y}_t + \bar{y} \tag{Eq 17.65}$$

where $\bar{y} = n^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} y_{it}$ is the full-sample mean.

Suppose there are no regressors:

$$y_{it} + v_t + u_i + \epsilon_{it}$$

then

$$\begin{aligned} \bar{y}_i &= \bar{v} + u_i + \bar{\epsilon}_i \\ \tilde{y}_i &= v_t + \bar{u} + \tilde{\epsilon}_t \\ \bar{y} &= \bar{v} + \bar{u} + \bar{\epsilon} \end{aligned}$$

and hence

$$\begin{aligned} \ddot{y}_{it} &= v_t + u_i + \epsilon_{it} - (\bar{v} + u_i + \bar{\epsilon}_i) - (v_t + \bar{u} + \tilde{\epsilon}_t) + \bar{v} + \bar{u} + \bar{\epsilon} \\ &= \epsilon_{it} - \bar{\epsilon}_i - \tilde{\epsilon}_t + \bar{\epsilon} \\ &= \ddot{\epsilon}_{it} \end{aligned}$$

so the individual and time effects are eliminated.

The two-way within transformation applied to (Eq 17.63) yields

$$\ddot{y}_{it} = \ddot{x}'_{it}\beta + \ddot{\epsilon}_{it} \tag{Eq 17.66}$$

which is invariant to v_i and u_i. The **two-way within estimator** of β is LS applied to (Eq 17.66). The coefficients are only identified for regressors which have variation both across individuals and across time.

The estimator is equivalent to the two-way dummy variables regression with individual and time effects dummies:

$$y_{it} = x'_{it}\beta + \tau'_t v + d'_i u + \epsilon_{it}$$

Another way to estimate the model is to write it as a one-way fixed effects model with time effects as dummies:

$$y_{it} = x'_{it}\beta + \tau'_t v + u_i + \epsilon_{it}$$

This can be estimated by standard one-way FE estimation methods such as *xtreg*. Given that T is small, there are not many coefficients estimated.
To prevent multicollinearity we need to omit one time dummy variable (such as using $t = 1$ as the baseline time period).

(o) Difference-in-differences: Introduction

We mainly discuss very basic, conventional DID estimation in the context of panel/ repeated cross-section data.
Conventionally, estimation is typically a two-way panel data regression with a policy indicator as a regressor. Clustered variance estimation is generally recommended for inference.

A **difference-in-difference estimator** compares the change in the treatment sample with a change in a control sample.
This requires the **parallel trend assumption (trends are parallel between the control and treatment groups)**. Otherwise, the estimate is not interpretable as a policy effect.

The standard errors are calculated by clustering (e.g., by restaurant in the case of Card and Kreuger 1994).

The regression is algebraically identical to the two-way FE regression

$$y_{it} = \theta D_{it} + u_i + v_t + \epsilon_{it}$$

where u_i is a restaurant FE and v_t is a time FE.

We can estimate this by a one-way FE regression with time dummies, and can also augment the regression with additional controls x_{it}.

(p) Identification

Consider the diff-in-diff equation

$$y_{it} = \theta D_{it} + x'_{it}\beta + u_i + v_t + \epsilon_{it} \qquad \text{(Eq 18.5)}$$

for $i = 1, ..., N$ and $t = 1, ..., T$

We are interested in conditions under which the coefficient θ is the causal impact of the treatment D_{it} on the outcome y_{it}.

We can write $y = h(D, x, e)$. Model (18.5) specifies that $h(D, x, e)$ is separable and linear in its arguments, and that the unobservables consist of individual-specific, time-specific and idiosyncratic effects. These are strong assumptions.

Recall the two-way within transformation (17.65) and set $\ddot{z}_{it} = (\ddot{D}_{it}, \ddot{x}'_{it})'$

Theorem 18.1 Suppose the following conditions hold:

i. $y_{it} = \theta D_{it} + x'_{it}\beta + u_i + v_t + \epsilon_{it}$, i.e., the observables, individual effects, time effects and idiosyncratic effects are additively separable)
ii. $\mathbb{E}(\ddot{z}_{it}\ddot{z}'_{it}) > 0$
iii. $\mathbb{E}(x_{it}\epsilon_{is} = 0)$ for all t and s (exogeneity assumption)
iv. Conditional on $x_{i1}, x_{i2}, ..., x_{iT}$ the random variables D_{it} and e_{is} are statistically independent for all t and s (i.e., see Week 1 notes; the treatment is conditionally independent of the error).

Then the coefficient θ in (18.5) equals the average causal effect for D on y conditional on x.

To show Theorem 18.1, we apply the two-way within transformation (Eq 17.65) to (Eq 18.5). We obtain

$$\ddot{y}_{it} = \theta \ddot{D}_{it} + \ddot{x}_{it}\beta + \ddot{\epsilon}_{it}$$

Under Condition (ii) the projection coefficients (θ, β) are uniquely defined, and under Condition (iii) and Condition (iv) they equal the linear regression coefficients. Thus θ is the regression derivative with respect to D.

Condition (iv) implies that conditional on $\ddot{x}_{it}$ the random variables $\ddot{D}_{it}$ and $\ddot{\epsilon}_{it}$ are statistically independent. Theorem 2.11 shows that this implies that the regression derivative θ equals the average causal effect as stated.

Assumption (iv) requires that the treatment be independent of ϵ_{it}, where the latter is a collection of factors not controlled for in the regression. This is a strong assumption.

(q) Inference

Many diff-in-diff applications use highly aggregate data because they are investigating the impact of policy changes that occur at an aggregate level.

We typically use clustering methods to calculate standard errors, with clustering applied at a high level of aggregation.

Recall the Moulton formula in an equi-correlation model with clustered dependence:

$$V_{\hat{\beta}} = (X'X)^{-1}\sigma^2(1 + \rho(N - 1))$$

where N is the number of individuals in each group, and ρ is the correlation across individuasl within the group.
This inflates the "usual" variance by the factor $(1 + \rho(N - 1))$.

Another challenge is when treatment ($D_{it} = 1$) applies to only a small number of units i. An extreme case is when there is only one treated unit. In this case, the robust covariance matrix estimator is singular and we have problematic inference. See Conley and Taber (2011) for example for other approaches to the problem.

10. Lecture 10: Nonlinear methods and M-estimators

(a) Introduction

An **m-estimator** is defined as a minimiser of a **sample average**

$$\hat{\theta} = arg \min_{\theta \in \Theta} S_n(\theta)$$
$$S_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \rho(Y_i, X_i, \theta)$$

i. $S_n(\theta) \equiv \frac{1}{n} \sum_{i=1}^{n} \rho(Y_i, X_i, \theta) \equiv \rho(Y, X, \theta)$ is some function of data (Y, X) and parameter $\theta \in \Theta$. For notational simplicity we write each component as $\rho_i(\theta) := \rho(Y_i, X_i, \theta)$

ii. $S_n(\theta)$ is called the **criterion function** or **objective function**.

iii. "m-estimator" is a broad class of estimators that include maximum likelihood estimators as a special case ("m" stands for "maximum likelihood type").

Examples:

i. Ordinary Least Squares: $\rho_i(\theta) = (Y_i - X_i'\theta)^2$, which is the SSE.

ii. Nonlinear Least Squares: $\rho_i(\theta) = (Y_i - m(X_i, \theta))^2$, where $m(.)$ is a nonlinear function.

iii. Least Absolute Deviations: $\rho_i(\theta) = |Y_i - X_i'\theta|$

iv. Quantile Regression: $(Y_i - X_i'\theta)(\tau - \infty\{(Y_i - X_i'\theta) < 0\})$

v. Maximum likelihood: $\rho_i(\theta) = -log\ f(Y_i|X_i, \theta)$, where f is the density of Y given X.

Recall the definition of the OLS estimator:
Definition 3.1: The least-squares estimator $\hat{\beta}$ is

$$\hat{\beta} = arg \min_{\beta \in \mathbb{R}^k} \hat{S}(\beta)$$

where

$$S(\hat{\beta}) = \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i'\beta)^2$$

The maximum likelihood estimator includes many standard estimators of limited-dependent-variable models such as the **probit model** for a binary dependent variable

$$P[Y = 1|X] = \Phi(X'\theta)$$
$$\rho_i(\theta) = -Y_i log(\Phi(X_i'\theta)) - (1 - Y_i) log(1 - \Phi(X_i'\theta))$$

where $\Phi(u)$ is the normal CDF and $\rho_i(\theta)$ is the negative log-density function.

Some estimators are **not m-estimators**, e.g., GMM estimators. For GMM, recall that the criterion function is a quadratic form of moment equations (which are sample averages).

The GMM estimator is typically defined as follows: Given as set of moment equations

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

and an $l \times l$ weight matrix $W > 0$, the **GMM criterion function** is defined as

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

so we see that as opposed to m-estimators, in GMM we are not optimising $\bar{g}$ but instead a quadratic form composed of $\bar{g}$.
However, both GMM and m-estimators belong to **extremum estimators**.

(b) Identification and estimation

A parameter vector θ is **identified** if it is uniquely determined by the probability distribution of the observations (i.e., we can express θ as a unique function of the observables).
This is a property of the probability distribution, not of the estimator.

However, when we discuss a specific estimator, it is common to describe identification in terms of the criterion function:
Assume $\mathbb{E}[\rho(Y, X, \theta)] < \infty$. Define

$$\begin{aligned} S(\theta) &= \mathbb{E}[S_n(\theta)] = \mathbb{E}[\rho(Y, X, \theta)] \\ \theta_0 &= arg \min_{\theta \in \Theta} S(\theta) \end{aligned}$$

i. $S(\theta)$ is the population criterion function.
ii. θ_0 is the population minimiser of $S(\theta)$.
e.g., if we miimise $\mathbb{E}[(y_i - X - \beta)^2]$, θ_0 maps to β, where $\beta = (\mathbb{E}(xx'))^{-1}E(xy)$.
$S_n(\theta)$ would be the simple average version of the SSE.

We say that θ is **identified/ point identified** by $S(\theta)$ if θ_0 is unique.

In linear models we may derive conditions under which a parameter is identified (e.g., in the case of the linear projection model).
However, this is generally not possible for nonlinear models, and identification needs to be examined (or assumed) on a model-by-model basis.

Note that even when a parameter is identified, we still need to numerically solve for the m-estimator $\hat{\theta}$, which minimises $S_n(\theta)$.

(c) Consistency

There are difficulties with estimating parameters consistently for nonlinear models:

i. There is no explicit algebraic expression for nonlinear estimators, and the nonlinear estimator needs to be solved numerically from the sample. There may be issues of local extrema, etc. that prevent numerical convergence.
ii. Due to the lack of algebraic expressions, we cannot apply WLLN to the formula (which is unavailable) of the estimator to prove consistency.

What is available to us is that an m-estimator minimises the criterion function $S_n(\theta)$, which is itself a sample average. For any given θ, the WLLN shows that

$$S_n(\theta) \xrightarrow{p} S(\theta)$$

where

$$\begin{aligned} S_n(\theta) &= \frac{1}{n} \sum_{i=1}^{n} \rho(Y_i, X_i, \theta) \\ S(\theta) &= \mathbb{E}[\rho(Y, X, \theta)] \end{aligned}$$

We need to show that the minimiser of $S_n(\theta)$ (i.e., $\hat{\theta}$) converges in probability to the minimiser of $S(\theta)$, i.e., θ_0. The WLLN by itself is not sufficient to show this result.

We know that $S_n(\theta)$ has **pointwise convergence** to $S(\theta)$, i.e., for each given point θ, $S_n(\theta)$ converges in probability to $S(\theta)$. However, the speed of convergence is different at different points of θ. We may have extreme scenarios where $\hat{\theta}$ diverges from θ_0 as $n \to \infty$.
What we are trying to prove is **uniform convergence**, i.e., $S_n(\theta)$ is restrained within the neighbourhood of $S(\theta) - \epsilon, S(\theta) + \epsilon$ for all values of θ.

Definition 22.1: $S_n(\theta)$ **converges in probability** to $S(\theta)$ **uniformly** over $\theta \in \Theta$ if

$$\sup_{\theta \in \Theta} |S_n(\theta) - S(\theta)| \xrightarrow{p} 0$$

as $n \to \infty$.

With uniform convergence of $S_n(\theta)$,, we can show the consistency of the m-estimator $\hat{\theta}$ for θ_0.
Theorem 22.1: $\hat{\theta} \xrightarrow{p} \theta_0$ as $n \to \infty$ if

i. $S_n(\theta)$ converges in probability to $S(\theta)$ uniformly over $\theta \in \Theta$, **and**
ii. θ_0 uniquely minimises $S(\theta)$ in the sense that for all $\epsilon > 0$,

$$\inf_{\theta : ||\theta - \theta_0|| \geq \epsilon} S(\theta) > S(\theta_0)$$

Proof:

i. Step 1: We show that $S(\hat{\theta}) \xrightarrow{p} S(\theta_0)$.

We have $S(\hat{\theta}) - S(\theta_0) \geq 0$ because θ_0 uniquely minimises $S(\theta)$.
We decompose the difference into the sum of three pairs:

$$\begin{aligned} S(\hat{\theta}) - S(\theta_0) &= S(\hat{\theta}) - S_n(\hat{\theta}) + S_n(\hat{\theta}) - S_n(\theta_0) + S_n(\theta_0) - S(\theta_0) \\ &\leq S(\hat{\theta}) - S_n(\hat{\theta}) + S_n(\theta_0) - S(\theta_0) \\ &\leq \sup_{\theta \in \Theta} |S_n(\theta) - S(\theta)| + \sup_{\theta \in \Theta} |S_n(\theta) - S(\theta)| \\ &\leq 2 \sup_{\theta \in \Theta} |S_n(\theta) - S(\theta)| \end{aligned}$$

where the first inequality is due to $\hat{\theta}$ being the minimiser of $S_n(\theta)$ and hence $S_n(\theta) \leq S_n(\theta_0)$.
Because $\sup_{\theta \in \Theta} |S_n(\theta) - S(\theta)| \xrightarrow{p} 0$ by uniform convergence, and

$$0 \leq S(\hat{\theta}) - S(\theta_0) \leq 2 \sup_{\theta \in \Theta} |S_n(\theta) - S(\theta)|$$

we have $S(\hat{\theta}) - S(\theta_0) \xrightarrow{p} 0$.

ii. Step 2: We need to show that $\lim_{n \to \infty} P(|\hat{\theta}_n - \theta_0| < \epsilon) = 1$ for all $\epsilon > 0$.

For any $\epsilon > 0$ there exists $\delta > 0$ such that $|\hat{\theta}_n - \theta_0| > \epsilon$ iff $|S(\hat{\theta}_n) - S(\theta_0)| > \delta$.
We then obtain $0 \leq P(|\hat{\theta}_n - \theta_0| > \epsilon) = P(|S(\hat{\theta}_n) - S(\theta_0)| > \delta)$.
By $S(\hat{\theta}) \xrightarrow{p} S(\theta_0)$, we have $\lim_{n \to \infty} P(|S(\hat{\theta}_n) - S(\theta_0)| > \delta) = 0$.
Therefore $\lim_{n \to \infty} P(|\hat{\theta}_n - \theta_0| > \epsilon) = 0$, or $\hat{\theta} \xrightarrow{p} \theta_0$, as claimed.

(d) Uniform LLN

Uniform convergence is a **high-level assumption** in that we cannot verify it from the more basic features of the model/data. The low-level sufficient conditions are:

Theorem 22.2: Uniform Law of Large Numbers (ULLN). Assume

i. (Y_i, X_i) are iid.
ii. $\rho(Y, X, \theta)$ is continuous in $\theta \in \Theta$ with probability one (i.e., almost surely; this is stronger than convergence in probability or weak convergence).
iii. $|\rho(Y, X, \theta)| \leq G(Y, X)$ where $\mathbb{E}[G(Y, X)] \leq \infty$
iv. Θ is compact (i.e., the parameter space is closed and bounded.

Then, $\sup_{\theta\in\Theta}|S_n(\theta)-S(\theta)| \xrightarrow{p} 0$ (i.e., uniform convergence).

Regularity condition (iii) bounds $\rho(Y,X,\theta)$ by a pure function of the data. It implies that $|\mathbb{E}(\rho(Y,X,\theta))| < \infty$.
$(G(Y,X))$ is called an envelope (out of scope for this course). See Hansen's Probability and Statistics for Economists, Theorem 18.2.

Definition 6.4: A random variable $z_n \in \mathbb{R}$ **converges in probability** to z as $n \to \infty$, denoted $z_n \xrightarrow{p} z$, or $plim_{n\to\infty} z_n = z$, if for all $\delta > 0$

$$\lim_{n\to\infty} P(|z_n - z| \leq \delta) = 1$$

Definition 6.6: A random variable $z_n \in \mathbb{R}$ **converges almost surely** to z as $z \to \infty$, denoted $z_n \xrightarrow{\text{a.s.}} z$ if for every $\delta > 0$

$$P(\lim_{n\to\infty} |z_n - z| \leq \delta) = 1$$

Or equivalently

$$P[\lim_{n\to\infty} z_n = z] = 1$$

Theorem 22.3: $\hat{\theta} \xrightarrow{p} \theta_0$ as $n \to \infty$ if

i. (Y_i, X_i) are iid
ii. $\rho(Y,X,\theta)$ is continuous in $\theta \in \Theta$ with probability one
iii. $|\rho(Y,X,\theta)| \leq G(Y,X)$ where $\mathbb{E}[G(Y,X)] < \infty$
iv. Θ is compact
v. θ_0 uniquely minimises $S(\theta)$.

(e) Asymptotic distribution

We define the first-order dervatives (the **score**) as

$$\psi(Y,X,\theta) = \frac{\partial}{\partial\theta}\rho(Y,X,\theta)$$
$$\bar{\psi}_n(\theta) = \frac{\partial}{\partial\theta}S_n(\theta)$$
$$\psi(\theta) = \frac{\partial}{\partial\theta}S(\theta)$$

We also define

$$\psi_i(\theta) = \psi(Y_i, X_i, \theta)$$
$$\psi_i = \psi_i(\theta_0)$$

We show the result as follows:
First, by definition $\hat{\theta}$ minimises $S_n(\theta)$, which implies that the FOC is satisfied:

$$0 = \bar{\psi}_n(\hat{\theta})$$

We then expand the RHS as a first order Taylor expansion about θ_0. This is valid when $\hat{\theta}$ is in the neighbourhood of θ_0, which holds for n sufficiently large.
(Note: since the slope is 0 at $\hat{\theta}$, $\bar{\psi}(\theta) = \frac{\partial}{\partial\theta}S_n(\theta)$).

$$0 = \bar{\psi}_n(\hat{\theta}) \approx \bar{\psi}_n(\theta_0) + \frac{\partial^2}{\partial\theta\partial\theta'}S_n(\theta_0)(\hat{\theta} - \theta_0)$$

Rewriting, we obtain

$$\sqrt{n}(\hat{\theta} - \theta_0) \approx -(\frac{\partial^2}{\partial\theta\partial\theta'}S_n(\theta_0))^{-1}(\sqrt{n}\bar{\psi}_n(\theta_0)$$

Consider the two components.

i. First, by the WLLN

$$\frac{\partial^2}{\partial\theta\partial\theta'}S_n(\theta_0) = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\theta\partial\theta'}\rho(Y_i, X_i, \theta_0) \xrightarrow{p} \mathbb{E}[\frac{\partial^2}{\partial\theta\partial\theta'}\rho_i(Y, X, \theta_0)] \overset{\text{def}}{=} \boldsymbol{Q}$$

where $\boldsymbol{Q}$ is the design matrix.
Note: Recall that in a linear projection model, we have

$$\sqrt{n}(\hat{\beta} - \beta) = (X'X)^{-1}(\sqrt{n}X'e) = (\frac{1}{n}\sum x_i x_i')^{-1}(\sqrt{n}\frac{1}{n}\sum x_i e_i)$$

Note: $\frac{\partial^2}{\partial\theta\partial\theta'}S_n(\theta_0)$ captures the concavity of $S_n(\theta_0)$ at θ_0.
Note: $\psi_n(\theta_0)$ is not zero. It captures the deviation from the FOC that is due to random sampling. For example, suppose we evaluate $H_0 : \theta = \theta_0$. If θ_0 is indeed the true parameter value, we will have an asymptotic distribution governed by this Taylor expansion under H_0. Otherwise we get a score that is very different from zero and reject H_0.

ii. Second,

$$\sqrt{n}\bar{\psi}_n(\theta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_o$$

since θ_0 minimises $S(\theta) = \mathbb{E}(\rho_i(\theta))$, it satisfies the FOC

$$0 = \psi(\theta_0) = \mathbb{E}[\psi(Y, X, \theta_0)]$$

Thus the summands in $\sum_{i=1}^{n}\psi_i$ are mean zero. Applying the CLT, this sum converges in distribution to $N(0, \Omega)$ where $\Omega = \mathbb{E}[\psi_i, \psi_i']$. We deduce that

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \boldsymbol{Q}^{-1}N(0, \Omega) = N(0, \boldsymbol{Q}^{-1}\Omega\boldsymbol{Q}^{-1})$$

The asymptotic variance is in sandwich form. In some cases it simplifies, e.g., in MLE, we have $\boldsymbol{Q} = \Omega$ so $\boldsymbol{Q}^{-1}\Omega\boldsymbol{Q}^{-1} = \Omega^{-1}$.