DegreeLive

Bachelor of Commerce (B2001) | Level 3 • Part B Major | Monash University

Principles of econometrics

Course Code: ETC3400 • Module 1 Notes

Module 1: Matrix Algebra for Econometrics & Finite-Sample OLS Theory

1.1 Matrix and Vector Foundations for Econometric Analysis

Modern econometric theory formulates linear and nonlinear statistical models in compact matrix notation because scalar summation notation becomes intractable once multiple regressors, parameter restrictions, and covariance structures are introduced. Throughout ETC3400, column vectors are denoted by lowercase letters (y, x_i, β, u), matrices by uppercase letters (X, P_X, M_X, Σ), and the transpose of a matrix A by A'.

Let A be an $m \times k$ real matrix with (i,j) element a_{ij}, and let B be a $k \times p$ real matrix. Matrix multiplication $C = AB$ yields an $m \times p$ matrix with element $c_{ij} = \sum_{s=1}^{k} a_{is} b_{sj}$, and transposition reverses the order of multiplication so that $(AB)' = B'A'$. When A is square ($m = k$) and non-singular ($\det(A) \neq 0$), its unique inverse A^{-1} satisfies $A^{-1}A = AA^{-1} = I_k$, where I_k is the $k \times k$ identity matrix, and $(AB)^{-1} = B^{-1}A^{-1}$ whenever both A and B are invertible.

Trace, Rank, Symmetry, and Idempotence

Four structural matrix concepts govern the finite-sample distribution theory of Ordinary Least Squares (OLS):

- **Trace:** For a square $n \times n$ matrix A, the trace is the sum of its principal diagonal elements, $\text{tr}(A) = \sum_{i=1}^{n} a_{ii}$. The trace operator is linear ($\text{tr}(aA + bB) = a\,\text{tr}(A) + b\,\text{tr}(B)$) and satisfies the cyclic permutation property $\text{tr}(ABC) = \text{tr}(CAB) = \text{tr}(BCA)$ whenever the products are conformable and square.
- **Rank:** The column rank of an $n \times k$ matrix X is the maximum number of linearly independent columns of X, which always equals its row rank. We say X has full column rank if $\text{rank}(X) = k \leq n$. A fundamental theorem states that $\text{rank}(X'X) = \text{rank}(X)$, so $X'X$ is invertible if and only if X has full column rank k.
- **Symmetric and Idempotent Matrices:** A square matrix A is symmetric if $A' = A$ and idempotent if $A^2 = AA = A$. For any symmetric idempotent matrix A, every eigenvalue equals either 0 or 1, which implies the crucial identity $\text{rank}(A) = \text{tr}(A)$.
- **Positive Definiteness:** A symmetric $k \times k$ matrix A is positive definite ($A > 0$) if the quadratic form $c'Ac > 0$ for all non-zero vectors $c \in \mathbb{R}^k$, and positive semi-definite ($A \geq 0$) if $c'Ac \geq 0$ for all $c \in \mathbb{R}^k$. If X is $n \times k$ with full column rank k, then for any $c \neq 0$, $Xc \neq 0$ and therefore $c'(X'X)c = (Xc)'(Xc) = \|Xc\|^2 > 0$, proving that $X'X$ is strictly positive definite.

Matrix Operation / Property	Algebraic Identity	Primary Econometric Application
Trace Cyclic Property	$\text{tr}(ABC) = \text{tr}(BCA)$	Deriving $E(\hat{u}'\hat{u} \mid X) = \sigma^2(n-k)$
Quadratic Form Expectation	$E(u'Au) = \text{tr}(A\Sigma) + \mu'A\mu$	Unbiasedness of s^2 and variance of quadratic forms
Symmetric Idempotence	$A = A',\ A^2 = A \implies \text{rank}(A) = \text{tr}(A)$	Degrees of freedom of P_X, M_X, and chi-squared tests
Matrix Differentiation (Linear)	$\partial(a'x)/\partial x = a$	First-order condition for normal equations

Degree.live

Matrix Operation / Property	Algebraic Identity	Primary Econometric Application
Matrix Differentiation (Quadratic)	$\partial(x'Ax)/\partial x = (A + A')x = 2Ax$ (symmetric A)	Minimising the sum of squared residuals $S(\beta)$
Partitioned Matrix Inverse	Schur complement $(X_2'M_1X_2)^{-1}$	Proof of the Frisch-Waugh-Lovell Theorem

Partitioned Matrices and the Schur Complement

Suppose a symmetric non-singular matrix A is partitioned into conformable blocks:

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$$

Where:

- A_{11} is a $k_1 \times k_1$ invertible submatrix and A_{22} is a $k_2 \times k_2$ invertible submatrix.
- $A_{12} = A_{21}'$ is a $k_1 \times k_2$ off-diagonal block capturing cross-products between two regressor groups.

Provided A_{11} and the Schur complement $D = A_{22} - A_{21}A_{11}^{-1}A_{12}$ are non-singular, the inverse of A is given explicitly by:

$$A^{-1} = \begin{pmatrix} A_{11}^{-1} + A_{11}^{-1}A_{12}D^{-1}A_{21}A_{11}^{-1} & -A_{11}^{-1}A_{12}D^{-1} \\ -D^{-1}A_{21}A_{11}^{-1} & D^{-1} \end{pmatrix}$$

Where:

- $D = A_{22} - A_{21}A_{11}^{-1}A_{12}$ is the Schur complement of A_{11} in A.
- The bottom-right block D^{-1} isolates the covariance structure of the second parameter subvector after partialling out the first block.

Vector Differentiation Rules

For a scalar-valued function $f(\beta)$ of a $k \times 1$ parameter vector $\beta = (\beta_1, \dots, \beta_k)'$, the gradient vector $\partial f/\partial \beta$ is the $k \times 1$ column vector with j-th entry $\partial f/\partial \beta_j$. For a $k \times 1$ constant vector a and a $k \times k$ constant matrix A, expanding $a'\beta = \sum_{j=1}^{k} a_j\beta_j$ and $\beta' A\beta = \sum_{i=1}^{k}\sum_{j=1}^{k} a_{ij}\beta_i\beta_j$ yields:

$$\frac{\partial(a'\beta)}{\partial\beta} = a, \qquad \frac{\partial(\beta' A\beta)}{\partial\beta} = (A + A')\beta, \qquad \frac{\partial^2(\beta' A\beta)}{\partial\beta\,\partial\beta'} = A + A'$$

Where:

- a is a fixed $k \times 1$ vector independent of β.
- A is a fixed $k \times k$ matrix; when A is symmetric ($A' = A$), the gradient simplifies to $2A\beta$ and the Hessian matrix equals $2A$.

DegreeLive

1.2 The Classical Linear Regression Model in Matrix Notation

Consider a sample of n observations $\{(y_i, x_{i1}, \dots, x_{ik})\}_{i=1}^{n}$ generated by a linear population model, such as modelling household expenditure across Australian states using ABS Household Expenditure Survey microdata or firm output using Australian corporate panel records. For each observation $i = 1, \dots, n$:

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + u_i = x_i' \beta + u_i$$

Where:

- y_i is the scalar dependent variable for observation i.
- $x_i = (x_{i1}, x_{i2}, \dots, x_{ik})'$ is the $k \times 1$ column vector of regressors for observation i, typically with $x_{i1} = 1$ representing the intercept.
- $\beta = (\beta_1, \dots, \beta_k)'$ is the $k \times 1$ vector of unknown population parameters.
- u_i is the unobserved scalar disturbance (error) term for observation i.

Stacking all n observations vertically produces the matrix representation of the linear regression model:

$$y = X\beta + u$$

Where:

- $y = (y_1, \dots, y_n)'$ is the $n \times 1$ vector of outcomes.
- X is the $n \times k$ design (regressor) matrix whose i-th row is x_i' and whose j-th column contains the n observations on regressor j.
- $u = (u_1, \dots, u_n)'$ is the $n \times 1$ vector of unobserved disturbances.

Finite-Sample Classical Assumptions (CLRM)

Exact finite-sample properties of OLS rely on the following four Gauss-Markov assumptions, supplemented by a fifth normality assumption when conducting exact small-sample inference:

1. **A1 (Linearity in Parameters):** The data generating process satisfies $y = X\beta + u$, where the functional relationship between y, X, and u is linear in the unknown parameter vector β.
2. **A2 (Full Column Rank / No Perfect Multicollinearity):** With probability one, $\text{rank}(X) = k \leq n$. No column of X is an exact linear combination of the remaining columns, guaranteeing that $X'X$ is positive definite and invertible.
3. **A3 (Strict Exogeneity):** The conditional expectation of the entire error vector given the entire regressor matrix is zero: $E(u \mid X) = 0$. In particular, for every pair of observations i and j, $E(u_i \mid x_1, \dots, x_n) = 0$. By the Law of Iterated Expectations, this also implies $E(u) = 0$ and $E(X'u) = 0$.
4. **A4 (Spherical Disturbances — Homoskedasticity and No Autocorrelation):** The conditional covariance matrix of u given X is proportional to the identity matrix: $\text{Var}(u \mid X) = E(uu' \mid X) = \sigma^2 I_n$, with $0 < \sigma^2 < \infty$. This requires both constant conditional variance $\text{Var}(u_i \mid X) = \sigma^2$ for all i (homoskedasticity) and zero conditional covariance $\text{Cov}(u_i, u_j \mid X) = 0$ for all $i \neq j$ (no serial or spatial correlation).

DegreeLive

5. **A5 (Conditional Normality — for Exact Inference):** $u \mid X \sim N(0, \sigma^2 I_n)$. Combined with A1–A4, A5 yields exact t and F sampling distributions for any sample size $n > k$.

1.3 Derivation of the OLS Estimator and Algebraic Properties

Given an arbitrary candidate parameter vector $b \in \mathrm{R}^k$, define the residual vector $e(b) = y - Xb$. The Ordinary Least Squares estimator $\hat{\beta}$ minimises the Sum of Squared Residuals (SSR) objective function $S(b)$:

$$S(b) = e(b)'e(b) = (y - Xb)'(y - Xb) = y'y - 2b'X'y + b'(X'X)b$$

Where:

- $S(b)$ is the scalar sum of squared residuals evaluated at b.
- $b'X'y = y'Xb$ because a 1×1 scalar equals its own transpose.

Applying the matrix differentiation rules with respect to b yields the first-order necessary condition (the OLS normal equations):

$$\frac{\partial S(b)}{\partial b}\Big|_{b=\hat{\beta}} = -2X'y + 2(X'X)\hat{\beta} = 0 \iff (X'X)\hat{\beta} = X'y$$

Where:

- $(X'X)\hat{\beta} = X'y$ is a system of k linear equations in k unknowns known as the normal equations.
- The Hessian matrix of $S(b)$ is $\partial^2 S(b)/(\partial b\, \partial b') = 2(X'X)$, which is strictly positive definite under Assumption A2, ensuring that $\hat{\beta}$ is the unique global minimiser of $S(b)$.

Pre-multiplying both sides of the normal equations by $(X'X)^{-1}$ gives the closed-form OLS estimator:

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

Where:

- $\hat{\beta}$ is the $k \times 1$ OLS estimator vector.
- $(X'X)^{-1}X'$ is the $k \times n$ Moore-Penrose left pseudoinverse of X when X has full column rank.

Substituting $y = X\beta + u$ into the formula for $\hat{\beta}$ separates the true parameter vector from the sampling error:

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

Where:

- $(X'X)^{-1}X'u$ is the $k \times 1$ sampling error vector $\hat{\beta} - \beta$, expressed as a linear transformation of the disturbance vector u given X.

DegreeLive

Algebraic Properties of OLS Fitted Values and Residuals

Let $\hat{y} = X\hat{\beta}$ denote the $n \times 1$ vector of fitted values and $\hat{u} = y - \hat{y} = y - X\hat{\beta}$ denote the $n \times 1$ vector of OLS residuals. Notice immediately from the normal equations that:

$$X'\hat{u} = X'(y - X\hat{\beta}) = X'y - (X'X)\hat{\beta} = 0$$

Where:

- $X'\hat{u} = 0$ states that the OLS residual vector $\hat{u}$ is orthogonal to every column of the regressor matrix X.
- Consequently, the fitted vector $\hat{y} = X\hat{\beta}$ is also orthogonal to the residuals: $\hat{y}'\hat{u} = \hat{\beta}'X'\hat{u} = 0$.
- Whenever X includes a column of ones ($\iota_n = (1, \dots, 1)'$), the first element of $X'\hat{u} = 0$ implies $\sum_{i=1}^{n} \hat{u}_i = 0$ and hence the sample mean of the fitted values equals the sample mean of y: $\bar{\hat{y}} = \bar{y}$.

1.4 Geometry of Least Squares: Projection and Annihilator Matrices

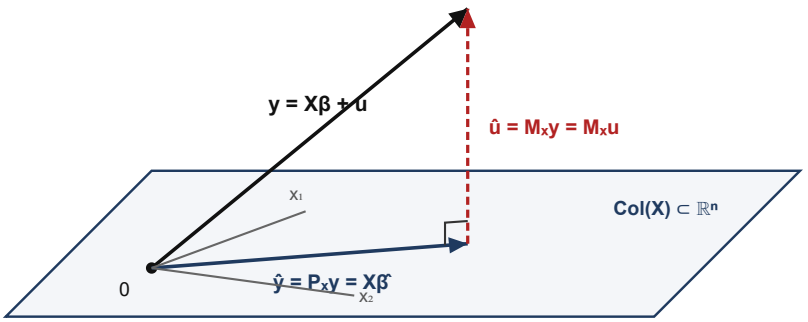
Least squares estimation has a transparent geometric interpretation in n-dimensional Euclidean space R^n. The columns of X span a k-dimensional linear subspace $\mathrm{C}(X) = \{Xb : b \in \mathrm{R}^k\} \subset \mathrm{R}^n$, called the column space of X. Because the observed outcome vector $y \in \mathrm{R}^n$ generally contains noise outside $\mathrm{C}(X)$, finding $\hat{\beta}$ to minimise $\|y - Xb\|^2$ is equivalent to finding the vector $\hat{y} \in \mathrm{C}(X)$ that is closest in Euclidean distance to y. By the Pythagorean theorem, this closest point is the orthogonal projection of y onto $\mathrm{C}(X)$.

Define the $n \times n$ **projection matrix** (also called the hat matrix) P_X and the $n \times n$ **annihilator matrix** (residual-maker matrix) M_X as:

$$P_X = X(X'X)^{-1}X', \qquad M_X = I_n - P_X = I_n - X(X'X)^{-1}X'$$

Where:

- P_X maps any vector $v \in \mathrm{R}^n$ onto its orthogonal projection $P_X v \in \mathrm{C}(X)$, so $\hat{y} = X\hat{\beta} = P_X y$.
- M_X maps any vector $v \in \mathrm{R}^n$ onto its component orthogonal to $\mathrm{C}(X)$, so $\hat{u} = y - P_X y = M_X y$.



Fundamental Properties of P_X and M_X

1. **Symmetry and Idempotence:** Both P_X and M_X are symmetric and idempotent:

$$P_X^{'} = P_X, quad P_X^2 = X(X^{'}X)^{-1}X^{'}X(X^{'}X)^{-1}X^{'} = P_X$$

$$M_X^{'} = M_X, quad M_X^2 = (I_n - P_X)(I_n - P_X) = I_n - 2P_X + P_X^2 = M_X$$

   Where $P_X^2 = P_X$ means projecting an already-projected vector leaves it unchanged.

2. **Mutual Orthogonality and Annihilation:** $P_X M_X = P_X(I_n - P_X) = P_X - P_X^2 = 0$. Furthermore, $P_X X = X$ (any vector in $\mathrm{C}(X)$ projects onto itself) and $M_X X = X - P_X X = 0$ (M_X annihilates every column of X).

3. **Relationship Between Residuals and True Errors:** Because $M_X X = 0$, substituting $y = X\beta + u$ into $\hat{u} = M_X y$ yields:

$$\hat{u} = M_X(X\beta + u) = M_X X\beta + M_X u = M_X u$$

   Where $\hat{u} = M_X u$ shows that OLS residuals depend solely on the true disturbance vector u and the design matrix X, never on the unknown parameter vector β.

4. **Traces and Degrees of Freedom:** Using the cyclic permutation property of the trace:

$$\mathrm{tr}(P_X) = \mathrm{tr}(X(X^{'}X)^{-1}X^{'}) = \mathrm{tr}((X^{'}X)^{-1}X^{'}X) = \mathrm{tr}(I_k) = k$$

$$\mathrm{tr}(M_X) = \mathrm{tr}(I_n - P_X) = \mathrm{tr}(I_n) - \mathrm{tr}(P_X) = n - k$$

   Where k is the model degrees of freedom (dimension of $\mathrm{C}(X)$) and $n - k$ is the residual degrees of freedom (dimension of the orthogonal complement $\mathrm{C}(X)^{\perp}$).

5. **Pythagorean Decomposition of Total Variation:** Because $y = P_X y + M_X y$ and $(P_X y)^{'}(M_X y) = y^{'}P_X M_X y = 0$, the squared norm decomposes orthogonally as $y^{'}y = \hat{y}^{'}\hat{y} + \hat{u}^{'}\hat{u}$. When X contains an intercept ι_n, letting $M_0 = I_n - \frac{1}{n}\iota_n\iota_n^{'}$ be the demeaning matrix gives $TSS = y^{'}M_0 y$, $SSE = \hat{y}^{'}M_0\hat{y}$, and $SSR = \hat{u}^{'}\hat{u}$, satisfying $TSS = SSE + SSR$.

Worked Example 1: Matrix OLS Estimation and Projection Verification

Suppose an econometrician analyses weekly log earnings y (in natural log of AUD) against an intercept ($x_1 = 1$) and years of post-secondary education (x_2) for $n = 4$ Australian workers, with design matrix X and outcome vector y:

DegreeLive

$$X = \begin{pmatrix} 1 & 2 \\ 1 & 4 \\ 1 & 6 \\ 1 & 8 \end{pmatrix}, \qquad y = \begin{pmatrix} 6.2 \\ 6.6 \\ 7.1 \\ 7.7 \end{pmatrix}$$

Where X is 4×2 ($n = 4, k = 2$) and y is 4×1.

Step 1: Compute $X'X$ and $X'y$.

$$X'X = \begin{pmatrix} 4 & 20 \\ 20 & 120 \end{pmatrix}, \qquad X'y = \begin{pmatrix} 6.2 + 6.6 + 7.1 + 7.7 \\ 2(6.2) + 4(6.6) + 6(7.1) + 8(7.7) \end{pmatrix} = \begin{pmatrix} 27.6 \\ 143.0 \end{pmatrix}$$

Where $\det(X'X) = 4(120) - 20^2 = 80 \neq 0$, confirming full column rank $k = 2$.

Step 2: Invert $X'X$ and solve for $\hat{\beta}$.

$$(X'X)^{-1} = \frac{1}{80}\begin{pmatrix} 120 & -20 \\ -20 & 4 \end{pmatrix} = \begin{pmatrix} 1.50 & -0.25 \\ -0.25 & 0.05 \end{pmatrix}$$

$$\hat{\beta} = (X'X)^{-1}X'y = \begin{pmatrix} 1.50(27.6) - 0.25(143.0) \\ -0.25(27.6) + 0.05(143.0) \end{pmatrix} = \begin{pmatrix} 5.65 \\ 0.25 \end{pmatrix}$$

Where $\hat{\beta}_0 = 5.65$ is the estimated intercept and $\hat{\beta}_1 = 0.25$ implies an approximate 25% increase in earnings per additional year of post-secondary education in this stylised sample.

Step 3: Compute fitted values, residuals, s^2, and orthogonality checks.

Computing $\hat{y} = X\hat{\beta} = (6.15, 6.65, 7.15, 7.65)'$ yields residuals $\hat{u} = y - \hat{y} = (0.05, -0.05, -0.05, 0.05)'$. Directly verifying the normal equation orthogonality condition gives $\sum \hat{u}_i = 0$ and $\sum x_{i2}\hat{u}_i = 2(0.05) + 4(-0.05) + 6(-0.05) + 8(0.05) = 0$. The residual sum of squares is $SSR = \hat{u}'\hat{u} = 4(0.0025) = 0.01$, yielding $s^2 = SSR/(n-k) = 0.01/2 = 0.005$ and estimated covariance matrix $\widehat{\mathrm{Var}}(\hat{\beta} \mid X) = s^2(X'X)^{-1}$ with diagonal entries 0.0075 and 0.00025. Hence $\mathrm{se}(\hat{\beta}_1) = \sqrt{0.00025} \approx 0.01581$, and the t-statistic for $H_0 : \beta_1 = 0$ is $t = 0.25/0.01581 = 15.81$.

1.5 Partitioned Regression and the Frisch-Waugh-Lovell (FWL) Theorem

In empirical econometrics, researchers frequently focus on a subset of explanatory variables β_2 while controlling for a nuisance or control block β_1 (such as time trends, quarterly seasonal indicators in ABS macroeconomic series, or firm fixed effects). Partition the $n \times k$ regressor matrix $X = [X_1 \;\; X_2]$, where X_1 is $n \times k_1$, X_2 is $n \times k_2$, and $k_1 + k_2 = k$. The regression model becomes:

DegreeLive

$$y = X_1\beta_1 + X_2\beta_2 + u$$

Where:

- β_1 is the $k_1 \times 1$ parameter vector associated with control variables X_1.
- β_2 is the $k_2 \times 1$ parameter vector of primary interest associated with X_2.

Theorem (Frisch-Waugh-Lovell): Let $M_1 = I_n - X_1(X_1^{'}X_1)^{-1}X_1^{'}$ be the annihilator matrix for X_1. Define the partialled-out dependent vector $\widetilde{y} = M_1 y$ (residuals from regressing y on X_1) and the partialled-out regressor block $\widetilde{X_2} = M_1 X_2$ (residuals from regressing each column of X_2 on X_1). Then:

1. The OLS estimate $\hat{\beta}_2$ from the full regression of y on $[X_1 \;\; X_2]$ equals the OLS estimate from regressing $\widetilde{y}$ on $\widetilde{X_2}$:

$$\hat{\beta}_2 = (X_2^{'}M_1X_2)^{-1}X_2^{'}M_1y = (\widetilde{X_2}\widetilde{X_2})^{-1}\widetilde{X_2}\widetilde{y}$$

2. The residual vectors from the full regression and the partialled-out regression are identical:

$$\hat{u} = M_X y = M_1 y - M_1 X_2 \hat{\beta}_2 = \widetilde{y} - \widetilde{X_2}\hat{\beta}_2$$

Proof of the Frisch-Waugh-Lovell Theorem

Write the full OLS normal equations $X^{'}\hat{u} = 0$ in partitioned form using $\hat{u} = y - X_1\hat{\beta}_1 - X_2\hat{\beta}_2$:

$$\begin{pmatrix} X_1^{'} \\ X_2^{'} \end{pmatrix}(y - X_1\hat{\beta}_1 - X_2\hat{\beta}_2) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

From the first block row, $X_1^{'}y = X_1^{'}X_1\hat{\beta}_1 + X_1^{'}X_2\hat{\beta}_2$. Solving for $\hat{\beta}_1$ gives:

$$\hat{\beta}_1 = (X_1^{'}X_1)^{-1}X_1^{'}(y - X_2\hat{\beta}_2)$$

Where $\hat{\beta}_1$ equals the OLS regression of $y - X_2\hat{\beta}_2$ on X_1. Substituting this expression back into the residual vector $\hat{u}$ yields:

$$\hat{u} = y - X_1(X_1^{'}X_1)^{-1}X_1^{'}(y - X_2\hat{\beta}_2) - X_2\hat{\beta}_2 = (I_n - P_1)(y - X_2\hat{\beta}_2) = M_1y - M_1X_2\hat{\beta}_2$$

Where $P_1 = X_1(X_1^{'}X_1)^{-1}X_1^{'}$. Now substitute $\hat{u} = M_1y - M_1X_2\hat{\beta}_2$ into the second block row of the normal equations, $X_2^{'}\hat{u} = 0$:

DegreeLive

$$X_2'(M_1 y - M_1 X_2 \hat{\beta}_2) = 0 \implies (X_2' M_1 X_2)\hat{\beta}_2 = X_2' M_1 y$$

Because M_1 is symmetric and idempotent ($M_1 = M_1' M_1$), we have $X_2' M_1 X_2 = (M_1 X_2)'(M_1 X_2) = \tilde{X}_2'\tilde{X}_2$ and $X_2' M_1 y = (M_1 X_2)'(M_1 y) = \tilde{X}_2'\tilde{y}$. Under Assumption A2, $\tilde{X}_2'\tilde{X}_2$ is invertible, establishing both $\hat{\beta}_2 = (\tilde{X}_2'\tilde{X}_2)^{-1}\tilde{X}_2'\tilde{y}$ and the residual equivalence $\hat{u} = \tilde{y} - \tilde{X}_2\hat{\beta}_2$. Notice carefully that although the point estimate $\hat{\beta}_2$ and residuals $\hat{u}$ are identical, the residual degrees of freedom for estimating σ^2 in the partialled-out regression must still be $n - k_1 - k_2$ rather than $n - k_2$, because k_1 degrees of freedom were consumed by M_1.

Empirical Applications of FWL

- **Demeaning (Intercept and Fixed Effects):** When $X_1 = \iota_n$, $M_1 = I_n - \frac{1}{n}\iota_n \iota_n'$ subtracts the sample mean from y and every column of X_2. In panel models, setting X_1 to a matrix of entity dummy variables yields the Within (Fixed Effects) estimator by demeaning each variable within each cross-sectional unit.
- **Detrending and Deseasonalising Macroeconomic Time Series:** When regressing Australian quarterly consumption growth on disposable income growth, including a linear time trend and three seasonal dummies in the regression yields the exact same income elasticity as first detrending and deseasonalising both series separately and then regressing the residuals on one another.
- **Omitted Variable Bias Formula:** If X_2 is omitted from a regression of y on X_1, the short-regression estimator $\tilde{\beta}_1 = (X_1'X_1)^{-1}X_1'y$ satisfies $E(\tilde{\beta}_1 \mid X) = \beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2$, where $(X_1'X_1)^{-1}X_1'X_2$ is the matrix of auxiliary regression slopes of X_2 on X_1.

Worked Example 2: Numerical Verification of the FWL Theorem

Return to the data of Worked Example 1, partitioning $X = [X_1 \;\; X_2]$ where $X_1 = (1,1,1,1)'$ and $X_2 = (2,4,6,8)'$. The annihilator matrix $M_1 = I_4 - \frac{1}{4}\iota_4\iota_4'$ subtracts sample means $\bar{x}_2 = 5$ and $\bar{y} = 6.9$:

$$\tilde{X}_2 = M_1 X_2 = \begin{pmatrix} 2-5 \\ 4-5 \\ 6-5 \\ 8-5 \end{pmatrix} = \begin{pmatrix} -3 \\ -1 \\ 1 \\ 3 \end{pmatrix}, \qquad \tilde{y} = M_1 y = \begin{pmatrix} 6.2-6.9 \\ 6.6-6.9 \\ 7.1-6.9 \\ 7.7-6.9 \end{pmatrix} = \begin{pmatrix} -0.7 \\ -0.3 \\ 0.2 \\ 0.8 \end{pmatrix}$$

Where $\tilde{X}_2'\tilde{X}_2 = (-3)^2 + (-1)^2 + 1^2 + 3^2 = 20$ and $\tilde{X}_2'\tilde{y} = (-3)(-0.7) + (-1)(-0.3) + 1(0.2) + 3(0.8) = 2.1 + 0.3 + 0.2 + 2.4 = 5.0$. Consequently:

$$\hat{\beta}_2 = \frac{\tilde{X}_2'\tilde{y}}{\tilde{X}_2'\tilde{X}_2} = \frac{5.0}{20} = 0.25$$

Computing the partialled-out residuals gives $\tilde{y} - \hat{\beta}_2\tilde{X}_2 = (-0.7, -0.3, 0.2, 0.8)' - 0.25(-3, -1, 1, 3)' = (0.05, -0.05, -0.05, 0.05)'$, matching both $\hat{\beta}_2$ and $\hat{u}$ from Worked Example 1 exactly.

1.6 Finite-Sample Properties and Proof of the Gauss-Markov Theorem

Before establishing optimality, we derive the exact conditional mean and covariance matrix of $\hat{\beta}$ and construct an unbiased estimator of σ^2.

Unbiasedness and Conditional Covariance Matrix of OLS

Using $\hat{\beta} = \beta + (X'X)^{-1}X'u$ and treating X as fixed when conditioning on X:

$$E(\hat{\beta} \mid X) = \beta + (X'X)^{-1}X'E(u \mid X) = \beta + (X'X)^{-1}X' \cdot 0 = \beta$$

Where Assumption A3 ($E(u \mid X) = 0$) is used directly; by the Law of Iterated Expectations, $E(\hat{\beta}) = E[E(\hat{\beta} \mid X)] = \beta$, so OLS is both conditionally and unconditionally unbiased.

Under Assumption A4 ($\text{Var}(u \mid X) = \sigma^2 I_n$), the conditional covariance matrix of $\hat{\beta}$ given X is:

$$\text{Var}(\hat{\beta} \mid X) = E\left[(\hat{\beta} - \beta)(\hat{\beta} - \beta)' \mid X\right] = (X'X)^{-1}X'E(uu' \mid X)X(X'X)^{-1}$$

$$= (X'X)^{-1}X'(\sigma^2 I_n)X(X'X)^{-1} = \sigma^2(X'X)^{-1}$$

Where:

- $\text{Var}(\hat{\beta} \mid X)$ is a $k \times k$ symmetric positive definite covariance matrix.
- The j-th diagonal entry $\sigma^2[(X'X)^{-1}]_{jj}$ equals $\text{Var}(\hat{\beta}_j \mid X)$.

Unbiased Estimation of Error Variance σ^2

Because σ^2 is unknown in practice, we estimate it using the OLS residual sum of squares $\hat{u}'\hat{u}$. Since $\hat{u} = M_X u$ and M_X is symmetric and idempotent ($M_X'M_X = M_X$), the residual sum of squares is a quadratic form in u: $\hat{u}'\hat{u} = u'M_X u$. Taking the conditional expectation given X and noting that a scalar equals its own trace:

$$E(\hat{u}'\hat{u} \mid X) = E\left[\text{tr}(u'M_X u) \mid X\right] = E\left[\text{tr}(M_X uu') \mid X\right] = \text{tr}\left(M_X E(uu' \mid X)\right)$$

$$= \text{tr}(M_X \sigma^2 I_n) = \sigma^2 \,\text{tr}(M_X) = \sigma^2(n-k)$$

Where $\text{tr}(M_X) = n - k$ was established in Section 1.4. Dividing $\hat{u}'\hat{u}$ by $n - k$ therefore defines the unbiased variance estimator:

$$s^2 = \hat{\sigma}^2 = \frac{\hat{u}'\hat{u}}{n-k}, \qquad E(s^2 \mid X) = \sigma^2$$

Where $s = \sqrt{s^2}$ is the standard error of the regression (SER). Notice that replacing $n - k$ by n (as in the Gaussian MLE $\tilde{\sigma}^2 = \hat{u}'\hat{u}/n$) introduces a downward finite-sample bias of $-k\sigma^2/n$.

DegreeLive

Statement and Proof of the Gauss-Markov Theorem

Theorem (Gauss-Markov): Under Assumptions A1–A4, the OLS estimator $\hat{\beta} = (X'X)^{-1}X'y$ is the **Best Linear Unbiased Estimator (BLUE)** of β. That is, for any alternative estimator $\tilde{\beta} = Cy$ that is linear in y (where C is a $k \times n$ matrix depending only on X) and conditionally unbiased ($E(\tilde{\beta} \mid X) = \beta$ for all $\beta \in \mathrm{R}^k$), the difference between their covariance matrices is positive semi-definite:

$$\mathrm{Var}(\tilde{\beta} \mid X) - \mathrm{Var}(\hat{\beta} \mid X) \geq 0$$

Where $A \geq 0$ means that for every $k \times 1$ vector $c \in \mathrm{R}^k$, $\mathrm{Var}(c'\tilde{\beta} \mid X) \geq \mathrm{Var}(c'\hat{\beta} \mid X)$.

Proof: Write the linear matrix C as the sum of the OLS matrix $(X'X)^{-1}X'$ and a $k \times n$ discrepancy matrix D:

$$C = (X'X)^{-1}X' + D$$

Where $D = C - (X'X)^{-1}X'$. Compute the conditional expectation of $\tilde{\beta} = Cy$ using $y = X\beta + u$ and Assumption A3 ($E(u \mid X) = 0$):

$$E(\tilde{\beta} \mid X) = E\left[\left((X'X)^{-1}X' + D\right)(X\beta + u) \mid X\right] = (I_k + DX)\beta + CE(u \mid X) = \beta + DX\beta$$

For $\tilde{\beta}$ to be unbiased for every possible vector $\beta \in \mathrm{R}^k$, we must have $DX\beta = 0$ for all β, which holds if and only if:

$$DX = 0$$

Next, compute the conditional covariance matrix of $\tilde{\beta}$ under Assumption A4 ($\mathrm{Var}(u \mid X) = \sigma^2 I_n$):

$$\mathrm{Var}(\tilde{\beta} \mid X) = C\,\mathrm{Var}(u \mid X)\,C' = \sigma^2 CC' = \sigma^2\left((X'X)^{-1}X' + D\right)\left(X(X'X)^{-1} + D'\right)$$

Expanding the product gives four terms:

$$\mathrm{Var}(\tilde{\beta} \mid X) = \sigma^2\left[(X'X)^{-1}X'X(X'X)^{-1} + (X'X)^{-1}X'D' + DX(X'X)^{-1} + DD'\right]$$

Applying the unbiasedness condition $DX = 0$ (and its transpose $X'D' = 0$) eliminates both cross-terms, leaving:

$$\mathrm{Var}(\tilde{\beta} \mid X) = \sigma^2(X'X)^{-1} + \sigma^2 DD' = \mathrm{Var}(\hat{\beta} \mid X) + \sigma^2 DD'$$

DegreeLive

Where $\sigma^2 DD'$ is symmetric and positive semi-definite because for any $c \in \mathrm{R}^k$, $c'(\sigma^2 DD')c = \sigma^2(D'c)'(D'c) = \sigma^2\|D'c\|^2 \geq 0$. Equality holds if and only if $D'c = 0$; in particular, $\mathrm{Var}(\tilde{\beta} \mid X) = \mathrm{Var}(\hat{\beta} \mid X)$ only when $D = 0$, i.e., $\tilde{\beta} = \hat{\beta}$ uniquely. Therefore, OLS achieves the minimum variance among all linear unbiased estimators.

1.7 Exact Normal-Theory Finite-Sample Inference

The Gauss-Markov theorem requires no distributional assumption on u beyond first and second conditional moments, but finite-sample hypothesis testing requires knowing the exact sampling distribution of $\hat{\beta}$ and s^2. Imposing Assumption A5 ($u \mid X \sim N(0, \sigma^2 I_n)$) ensures that linear combinations of u are jointly normal and quadratic forms in u follow scaled chi-squared distributions.

Exact Distributions of $\hat{\beta}$, s^2, and the t-Statistic

1. **Normality of $\hat{\beta}$:** Since $\hat{\beta} - \beta = (X'X)^{-1}X'u$ is a linear transformation of a multivariate normal vector conditional on X,

   $$\hat{\beta} \mid X \sim N\left(\beta,\ \sigma^2(X'X)^{-1}\right)$$

   In fact, under Assumptions A1–A5, $\hat{\beta}$ attains the Cramér-Rao Lower Bound and is the Minimum Variance Unbiased Estimator (MVUE) among all unbiased estimators, linear or nonlinear.

2. **Independence of $\hat{\beta}$ and $\hat{u}$:** Both $\hat{\beta} - \beta = (X'X)^{-1}X'u$ and $\hat{u} = M_X u$ are jointly normal conditional on X. Their conditional covariance matrix is:

   $$\mathrm{Cov}(\hat{\beta}, \hat{u} \mid X) = E\left[(X'X)^{-1}X'uu'M_X \mid X\right] = \sigma^2(X'X)^{-1}X'M_X = 0$$

   Because $X'M_X = 0$ and uncorrelated jointly normal random vectors are statistically independent, $\hat{\beta}$ and $\hat{u}$ (and hence $\hat{\beta}$ and s^2) are conditionally independent given X.

3. **Chi-Squared Distribution of s^2:** Standardise $z = u/\sigma \mid X \sim N(0, I_n)$. Because M_X is symmetric and idempotent with rank $n - k$, there exists an orthogonal matrix Q such that $Q'M_XQ = \mathrm{diag}(I_{n-k}, 0_k)$. Letting $w = Q'z \sim N(0, I_n)$ yields:

   $$\frac{(n-k)s^2}{\sigma^2} = \frac{\hat{u}'\hat{u}}{\sigma^2} = z'M_Xz = w'Q'M_XQw = \sum_{j=1}^{n-k} w_j^2 \sim \chi^2_{n-k}$$

   Where χ^2_{n-k} is the chi-squared distribution with $n - k$ degrees of freedom.

4. **Exact t-Distribution for a Single Linear Combination:** To test $H_0 : c'\beta = r_0$ against $H_1 : c'\beta \neq r_0$ for a known non-zero $k \times 1$ vector c (e.g., a single coefficient when c is a unit basis vector e_j), the numerator satisfies $(c'\hat{\beta} - r_0)/\sqrt{\sigma^2 c'(X'X)^{-1}c} \sim N(0,1)$ under H_0, independently of $(n-k)s^2/\sigma^2 \sim \chi^2_{n-k}$. Substituting s^2 for σ^2 gives:

DegreeLive

$$t = \frac{c'\hat{\beta} - r_0}{\cancel{s^2 c'(X'X)^{-1}c}} = \frac{c'\hat{\beta} - r_0}{\text{se}(c'\hat{\beta})} \sim t_{n-k}$$

Where t_{n-k} is Student's t-distribution with $n - k$ degrees of freedom.

Exact F-Test for Multiple Linear Restrictions

Now consider $q \leq k$ linearly independent joint restrictions written in matrix form as:

$$H_0 : R\beta = r \qquad \text{vs.} \qquad H_1 : R\beta \neq r$$

Where:

- R is a known $q \times k$ matrix with full row rank $\text{rank}(R) = q$.
- r is a known $q \times 1$ vector of hypothesised values.

Under H_0 and Assumptions A1–A5, $R\hat{\beta} - r = R(\hat{\beta} - \beta) \mid X \sim N(0, \sigma^2 R(X'X)^{-1}R')$. Standardising the quadratic form by σ^2 yields a χ^2_q variable independent of $(n-k)s^2/\sigma^2 \sim \chi^2_{n-k}$. Forming the ratio of independent chi-squared variables divided by their respective degrees of freedom produces the exact F-statistic:

$$F = \frac{(R\hat{\beta} - r)' \left[R(X'X)^{-1}R'\right]^{-1} (R\hat{\beta} - r)}{q \cdot s^2} \sim F_{q,\, n-k}$$

Where:

- q is the number of linear restrictions (numerator degrees of freedom).
- $n - k$ is the residual degrees of freedom (denominator degrees of freedom).

Restricted Least Squares (RLS) and Goodness-of-Fit Equivalence

Equivalently, let $\tilde{\beta}_{RLS}$ minimise $S(b) = (y - Xb)'(y - Xb)$ subject to the linear constraint $Rb = r$. Forming the Lagrangian $\text{L}(b, \lambda) = (y - Xb)'(y - Xb) + 2\lambda'(Rb - r)$ and solving the first-order conditions gives the Restricted Least Squares estimator:

$$\tilde{\beta}_{RLS} = \hat{\beta} - (X'X)^{-1}R' \left[R(X'X)^{-1}R'\right]^{-1} (R\hat{\beta} - r)$$

Where $\hat{\beta}$ is the unconstrained OLS estimator. Substituting $\tilde{\beta}_{RLS}$ into the restricted residual sum of squares $SSR_R = (y - X\tilde{\beta}_{RLS})'(y - X\tilde{\beta}_{RLS})$ and comparing against the unrestricted sum of squared residuals $SSR_U = \hat{u}'\hat{u}$ yields the identity:

$$SSR_R - SSR_U = (R\hat{\beta} - r)' \left[R(X'X)^{-1}R'\right]^{-1} (R\hat{\beta} - r)$$

DegreeLive

Consequently, the matrix Wald-type F-statistic is algebraically identical to the classic nested-model F-statistic:

$$F = \frac{(SSR_R - SSR_U)/q}{SSR_U/(n-k)} = \frac{(R_U^2 - R_R^2)/q}{(1-R_U^2)/(n-k)}$$

Where R_U^2 and R_R^2 are the unrestricted and restricted coefficients of determination (valid when both models contain the same dependent variable and an intercept). To penalise unnecessary regressors that leave SSR_U scarcely changed, the **Adjusted** R^2 is defined as $\bar{R}^2 = 1 - \frac{SSR/(n-k)}{TSS/(n-1)} = 1 - \frac{n-1}{n-k}(1-R^2)$. Adding a single regressor increases $\bar{R}^2$ if and only if the square of its t-statistic exceeds 1 ($|t| > 1$).

Worked Example 3: Matrix F-Test for Joint CAPM Restrictions on ASX Equities

Consider an empirical Capital Asset Pricing Model (CAPM) test augmented with firm size across $n = 63$ ASX-listed Australian companies ($k = 3$ regressors: intercept β_0, market beta β_1, and log market capitalisation β_2). Suppose OLS estimation produces $\hat{\beta} = (0.012, 1.150, -0.008)'$, $s^2 = 0.0016$, and:

$$(X'X)^{-1} = \begin{pmatrix} 0.050 & -0.020 & -0.001 \\ -0.020 & 0.250 & 0.005 \\ -0.001 & 0.005 & 0.010 \end{pmatrix}$$

We test the joint hypothesis $H_0 : \beta_0 = 0$ and $\beta_1 = 1$ ($q = 2$ restrictions) using R and r:

$$R = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}, \qquad r = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

Step 1: Compute the discrepancy vector $R\hat{\beta} - r$ and extract $R(X'X)^{-1}R'$.

$$R\hat{\beta} - r = \begin{pmatrix} 0.012 - 0 \\ 1.150 - 1 \end{pmatrix} = \begin{pmatrix} 0.012 \\ 0.150 \end{pmatrix}, \qquad R(X'X)^{-1}R' = \begin{pmatrix} 0.050 & -0.020 \\ -0.020 & 0.250 \end{pmatrix}$$

Step 2: Invert the 2×2 matrix $R(X'X)^{-1}R'$.

The determinant is $0.050(0.250) - (-0.020)^2 = 0.0125 - 0.0004 = 0.0121$, so:

$$[R(X'X)^{-1}R']^{-1} = \frac{1}{0.0121}\begin{pmatrix} 0.250 & 0.020 \\ 0.020 & 0.050 \end{pmatrix}$$

Step 3: Evaluate the quadratic form and the F-statistic.

DegreeLive

$$(R\hat{\beta} - r)' \left[R(X'X)^{-1}R'\right]^{-1} (R\hat{\beta} - r) = \frac{0.250(0.012)^2 + 2(0.020)(0.012)(0.150) + 0.050(0.150)^2}{0.0121}$$

$$= \frac{0.000036 + 0.000072 + 0.001125}{0.0121} = \frac{0.001233}{0.0121} \approx 0.101901$$

$$F = \frac{0.101901}{2 \times 0.0016} = \frac{0.101901}{0.0032} \approx 31.84$$

Comparing $F = 31.84$ against the 5% critical value $F_{0.05;\,2,\,60} \approx 3.15$ (with $n - k = 60$ denominator degrees of freedom), we strongly reject the joint null hypothesis H_0 at the 5% and 1% significance levels.

1.8 Numerical Conditioning, Multicollinearity, and Simulation Principles

In applied econometric computing (for example, in R or Python), solving the normal equations directly via explicit inversion of $X'X$ can suffer from severe floating-point rounding error when columns of X are nearly collinear. The degree of numerical sensitivity is governed by the **condition number** $\kappa(X'X) = \lambda_{\max}/\lambda_{\min}$, the ratio of the largest to smallest eigenvalues of $X'X$. Because $\kappa(X'X) = [\kappa(X)]^2$, squaring the design matrix doubles the loss of numerical precision. Production econometric software therefore computes $\hat{\beta}$ via QR decomposition ($X = QR$, where $Q'Q = I_k$ and R is upper triangular, solving $R\hat{\beta} = Q'y$) or Singular Value Decomposition (SVD), avoiding explicit formation of $X'X$.

Statistically, imperfect multicollinearity arises when regressors exhibit strong linear correlation without violating $\text{rank}(X) = k$. Letting R_j^2 denote the R^2 from regressing column x_j on all other columns of X, the partitioned inverse formula yields the exact variance expression for $\hat{\beta}_j$:

$$\text{Var}(\hat{\beta}_j \mid X) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2(1 - R_j^2)} = \frac{\sigma^2}{SST_j} \cdot \text{VIF}_j$$

Where:

- $SST_j = \sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$ is the total sample variation in regressor j.
- $\text{VIF}_j = 1/(1 - R_j^2)$ is the Variance Inflation Factor for regressor j.

Crucially, imperfect multicollinearity does not bias $\hat{\beta}$ or invalidate the Gauss-Markov theorem, t-tests, or F-tests; rather, it inflates individual standard errors while joint F-tests can still detect strong collective significance. Finally, Monte Carlo simulation provides a computational laboratory for verifying finite-sample properties: by drawing M independent artificial samples $u^{(m)} \sim N(0, \sigma^2 I_n)$ for fixed X, computing $\hat{\beta}^{(m)}$ and $s^{2(m)}$ across $m = 1, \ldots, M$, analysts can directly observe that the empirical mean of $\hat{\beta}^{(m)}$ converges to β, empirical variance matches $\sigma^2(X'X)^{-1}$, and the empirical rejection rate of exact t-tests under H_0 equals the nominal significance level α.

Degree Live