Econometrics I

Personal Study Notes and Formal Derivations

Victor Alves

2026-06

Contents

Disclaimer

- **Incomplete Document:** This file constitutes supporting material under active development. Several sections and asymptotic derivations are still being revised, expanded, and supplemented.
- **AI-Assisted Construction:** This material was structured, reviewed, and expanded with the assistance of Artificial Intelligence models for didactic organization and Markdown/LaTeX formatting rigor.
- **Margin of Error:** Due to the technical nature of the matrix and asymptotic proofs, the text may contain typographical errors, algebraic omissions, or theoretical inaccuracies not yet reviewed by the author. It should not be used as the sole definitive bibliographic source.

Chapter 1: The Classical Linear Regression Model (CLRM)

1.1 Motivation and Geometric Interpretation

The Classical Linear Regression Model (CLRM) addresses the problem of estimating the conditional mean of a dependent variable (y) given a set of independent variables ($x_1, \ldots, x_k$), aiming to describe the statistical dependence between them.

Geometric Interpretation in $\mathbb{R}^n$:

Let $\mathbf{y} \in \mathbb{R}^n$ be the vector of observations of the dependent variable and $\mathbf{X} \in \mathbb{R}^{n \times K}$ the data matrix. Geometrically, the Ordinary Least Squares (OLS) estimator seeks a vector of fitted values $\hat{\mathbf{y}}$ that represents the **orthogonal projection** of $\mathbf{y}$ onto the subspace spanned by the columns of $\mathbf{X}$ (the column space). The residual vector $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$ is the vector of smallest Euclidean norm connecting $\mathbf{y}$ to the subspace, ensuring that $\mathbf{e}$ is orthogonal to every column of $\mathbf{X}$.

1.2 CLRM Assumptions and Their Roles

Assumption 1.1: Linearity in Parameters

- **Role:** Defines the functional structure of the model, ensuring that the dependent variable is a weighted sum of the regressors and an additive disturbance.

- **Counterexample:** A production model where parameters are exponents, such as $Y = \beta_0 X_1^{\beta_1} \varepsilon$, requires logarithmic transformation to be linearized; if not linearizable, algebraic OLS methods fail in direct identification.

Assumption 1.2: No Perfect Multicollinearity (Full Rank)

- **Role:** Ensures that the explanatory variables vary in a linearly independent manner, allowing the inversion of the matrix ($\mathbf{X}'\mathbf{X}$) and the unique identification of parameters.
- **Counterexample:** Including the same variable in different units (e.g., meters and centimeters) or falling into the "dummy variable trap" (including dummies for all categories and an intercept) renders the matrix singular, preventing the estimator from being computed.

Assumption 1.3: Strict Exogeneity (Zero Conditional Mean)

- **Role:** Guarantees that the error term does not contain information systematically related to the regressors, being the necessary condition for the unbiasedness of the estimator.
- **Counterexample:** Omitting a relevant variable correlated with the included regressors causes the error to absorb this effect, generating correlation between $\mathbf{X}$ and ε and resulting in systematically biased estimates (omitted variable bias).

Assumption 1.4: Spherical Errors (Homoskedasticity and No Autocorrelation)

- **Role:** Establishes that the error variance is constant across observations and that errors are independent of each other, simplifying the covariance structure to a scalar matrix ($\sigma^2\mathbf{I}$).
- **Counterexample:** In time series data, shocks in one period may persist into the next (autocorrelation), or in cross-sectional data, variance may increase with income level (heteroskedasticity), invalidating standard errors and OLS efficiency.

1.3 Formal Statement of the CLRM

Let a measure and probability space be associated with the data generating process. The Classical Linear Regression Model (CLRM) is defined by the following structure:

1. **Population Linearity:** The response vector $\mathbf{y} \in \mathbb{R}^n$ satisfies the equation:

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon \tag{1.1}$$

   where $\mathbf{X}$ is an $n \times K$ matrix of regressors, $\beta \in \mathbb{R}^K$ is the vector of fixed parameters, and $\varepsilon \in \mathbb{R}^n$ is the vector of stochastic disturbances.

2. **Identifiability:** The matrix $\mathbf{X}$ has full column rank, such that $\text{rank}(\mathbf{X}) = K < n$ with probability 1.

3. **Exogeneity:** $E[\varepsilon|\mathbf{X}] = \mathbf{0}_{n\times 1}$.

4. **Sphericity:** $E[\varepsilon\varepsilon'|\mathbf{X}] = \sigma^2\mathbf{I}_n$, where $\mathbf{I}_n$ is the identity matrix of order n and $\sigma^2 \in (0, \infty)$.

1.4 Derivation of the OLS Estimator

Theorem (Ordinary Least Squares Estimator). Under Assumptions 1.1 to 1.4, the OLS estimator is given by:

$$\boxed{\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}} \tag{1.2}$$

Proof:

We define the objective function $S(\mathbf{b})$ as the squared Euclidean norm of the residual vector:

$$S(\mathbf{b}) = (\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b}) \tag{1.3}$$

Expanding the quadratic form:

$$S(\mathbf{b}) = \mathbf{y}'\mathbf{y} - \mathbf{b}'\mathbf{X}'\mathbf{y} - \mathbf{y}'\mathbf{X}\mathbf{b} + \mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b} \tag{1.4}$$

Since $\mathbf{b}'\mathbf{X}'\mathbf{y}$ is a scalar (dimension 1×1), it is identical to its transpose ($\mathbf{y}'\mathbf{X}\mathbf{b}$). Hence:

$$S(\mathbf{b}) = \mathbf{y}'\mathbf{y} - 2\mathbf{b}'\mathbf{X}'\mathbf{y} + \mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b} \tag{1.5}$$

To find the minimum, we differentiate $S(\mathbf{b})$ with respect to $\mathbf{b}$ and set it to zero (First Order Condition):

$$\frac{\partial S(\mathbf{b})}{\partial \mathbf{b}} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{0} \tag{1.6}$$

Isolating $\mathbf{b}$:

$$\mathbf{X}'\mathbf{X}\hat{\beta} = \mathbf{X}'\mathbf{y} \tag{1.7}$$

By Assumption 1.2 (Full Rank), $(\mathbf{X}'\mathbf{X})$ is positive definite and thus invertible:

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{1.8}$$

□

1.5 Proof of Unbiasedness of OLS

Theorem (Unbiasedness of OLS). Under Assumption 1.3 (Strict Exogeneity), the OLS estimator is unbiased:

$$\boxed{E[\hat{\beta}|\mathbf{X}] = \beta} \tag{1.9}$$

Proof:

Substituting the structural equation (1.1) into (1.2):

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\beta + \varepsilon) \tag{1.10}$$

Distributing:

$$\hat{\beta} = \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon \tag{1.11}$$

Applying the conditional expectation given $\mathbf{X}$:

$$E[\hat{\beta}|\mathbf{X}] = \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E[\varepsilon|\mathbf{X}] \tag{1.12}$$

By Assumption 1.3, $E[\varepsilon|\mathbf{X}] = \mathbf{0}$:

$$E[\hat{\beta}|\mathbf{X}] = \beta \tag{1.13}$$

□

1.6 Implications of Strict Exogeneity

1.6.1 Motivation

Exogeneity is the assumption that enables the transition from mere **statistical association** to **causal or structural inference**. In practical terms, it solves the problem of isolating the effect of a specific explanatory variable on the dependent variable, ensuring that the error term does not contain factors that vary systematically with the regressors.

Geometric Interpretation in $\mathbb{R}^n$:

Let $\mathbf{X}$ be the matrix of regressors. Exogeneity implies that, in the sample space, the vector of population errors ε is "orthogonal" (in expectation) to the subspace spanned by the columns of $\mathbf{X}$. This means that the projection of $\mathbf{y}$ onto the column space of $\mathbf{X}$ recovers exactly the structural component $\mathbf{X}\beta$, without contamination from unobserved factors that could "pull" the projection in wrong directions.

1.6.2 Assumptions and Their Roles

Assumption 1.5: Strict Exogeneity (Zero Conditional Mean)

- **Role:** Guarantees that the error is not predictable from any information contained in the regressors, establishing mean independence.
- **Counterexample:** If we study the effect of education (x) on wages (y), and omit "innate ability" (a), the latter will reside in the error ε. Since more able individuals tend to pursue more education, $\text{Cov}(x, \varepsilon) \neq 0$. The estimator will attribute to education the wage gain that actually comes from ability, generating **omitted variable bias**.

1.6.3 Implications of Strict Exogeneity

Implication I: Zero Unconditional Mean ($E[\varepsilon] = \mathbf{0}$)

By the Law of Iterated Expectations:

$$E[\varepsilon] = E[E[\varepsilon|\mathbf{X}]] = E[\mathbf{0}] = \mathbf{0} \tag{1.14}$$

Implication II: Orthogonality and Zero Covariance

For any function $h(\mathbf{X})$ of the regressors:

$$E[h(\mathbf{X})\varepsilon] = E[E[h(\mathbf{X})\varepsilon|\mathbf{X}]] = E[h(\mathbf{X})E[\varepsilon|\mathbf{X}]] = \mathbf{0} \tag{1.15}$$

In particular, $\text{Cov}(\mathbf{X}, \varepsilon) = \mathbf{0}$.

Implication III: Identification of the Parameter β

Taking the conditional expectation of model (1.1):

$$E[\mathbf{y}|\mathbf{X}] = \mathbf{X}\beta + E[\varepsilon|\mathbf{X}] = \mathbf{X}\beta \tag{1.16}$$

Premultiplying by $\mathbf{X}'$ and taking the unconditional expectation:

$$E[\mathbf{X}'\mathbf{y}] = E[\mathbf{X}'\mathbf{X}]\beta \tag{1.17}$$

Isolating β:

$$\beta = (E[\mathbf{X}'\mathbf{X}])^{-1}E[\mathbf{X}'\mathbf{y}] \tag{1.18}$$

Equation (1.18) shows that the population parameter β is uniquely determined by the observable moments of $\mathbf{y}$ and $\mathbf{X}$. Without exogeneity, the OLS estimator would converge to a value $\beta^* \neq \beta$, resulting in **inconsistency**.

1.7 Ordinary Least Squares (OLS) as Orthogonal Projection

1.7.1 Motivation

The OLS process solves the problem of **approximation in subspaces**. When we have a system of linear equations $\mathbf{y} = \mathbf{X}\beta$ that is overidentified (more equations than unknowns, $n > K$), there is generally no exact solution due to noise in the data. OLS seeks the parameter vector that produces the "best fit," minimizing the Euclidean distance between the observed vector and the predicted vector.

1.7.2 Assumptions and Their Roles

Assumption 1.6: Full Column Rank

- **Role:** Ensures that the columns of $\mathbf{X}$ are linearly independent, guaranteeing that the Gram matrix $(\mathbf{X}'\mathbf{X})$ is invertible.
- **Counterexample:** If one variable is a copy of another, the subspace S will have dimension less than K. This creates a "line" or "plane" of possible solutions for β, making the estimator indeterminate.

Assumption 1.7: $n > K$ (Positive Degrees of Freedom)

- **Role:** Necessary for the residual to not be trivially zero and to allow estimation of the noise variance.
- **Counterexample:** If $n = K$, the "approximation" becomes an exact interpolation; if $n < K$, the system is underdetermined and there are infinitely many solutions that zero the error, preventing the identification of individual effects.

1.7.3 Formal Statement and Derivation

The OLS estimator, denoted by $\hat{\beta}$, is defined as:

$$\hat{\beta} = \underset{\mathbf{b} \in \mathbb{R}^K}{\arg\min}\ S(\mathbf{b}) \tag{1.19}$$

where $S(\mathbf{b}) = \|\mathbf{y} - \mathbf{X}\mathbf{b}\|^2 = (\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b})$.

Theorem (Uniqueness of OLS). Under Assumption 1.6, the OLS estimator is unique and given by (1.2).

Proof:

The Hessian matrix is $\mathbf{H} = 2\mathbf{X}'\mathbf{X}$. By Assumption 1.6, $\mathbf{X}$ has linearly independent columns. If $\mathbf{v} \neq \mathbf{0}$, then $\mathbf{X}\mathbf{v} \neq \mathbf{0}$, which implies $\mathbf{v}'(\mathbf{X}'\mathbf{X})\mathbf{v} = \|\mathbf{X}\mathbf{v}\|^2 > 0$. Hence, $2\mathbf{X}'\mathbf{X}$ is **positive definite**, guaranteeing that the critical point is a **strict and unique minimum**. □

1.8 Gauss-Markov Theorem

1.8.1 Motivation

The Gauss-Markov Theorem addresses the problem of **estimator selection**. Given that there are infinitely many ways to combine the data to estimate a parameter β in an unbiased manner, the theorem identifies which of these combinations produces the smallest uncertainty (variance).

Geometric Interpretation in $\mathbb{R}^n$:

In the sample space, the OLS estimator $\hat{\beta}$ corresponds to the orthogonal projection of the observation vector $\mathbf{y}$ onto the subspace spanned by the columns of $\mathbf{X}$. Any other linear unbiased estimator can be seen as an "oblique" projection or a transformation that does not minimize the Euclidean distance. The theorem guarantees that the "ellipse of uncertainty" (covariance matrix) of any other linear unbiased estimator will contain the ellipse of uncertainty of OLS.

1.8.2 Formal Statement

Gauss-Markov Theorem. Under Assumptions 1.1 to 1.4 (Linearity, Full Rank, Strict Exogeneity, and Sphericity), the OLS estimator $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ has the smallest covariance matrix (in the positive definite matrix sense) within the class of all linear unbiased estimators of β.

1.8.3 Granular Proof

Definition 1 (OLS Estimator):

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{A}\mathbf{y}, \quad \mathbf{A} \equiv (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \tag{1.20}$$

Definition 2 (Alternative Linear Estimator):

Let $\tilde{\beta}$ be any linear estimator defined by $\tilde{\beta} = \mathbf{C}\mathbf{y}$, where $\mathbf{C}$ is a $K \times n$ matrix that may depend on $\mathbf{X}$, but not on $\mathbf{y}$. We define $\mathbf{D}$ such that $\mathbf{C} = \mathbf{A} + \mathbf{D}$.

Lemma 1 (Unbiasedness Condition): For $\tilde{\beta}$ to be unbiased, we must have $\mathbf{D}\mathbf{X} = \mathbf{0}$.

Proof:

$$E[\tilde{\beta}|\mathbf{X}] = E[(\mathbf{A} + \mathbf{D})\mathbf{y}|\mathbf{X}] = (\mathbf{A} + \mathbf{D})\mathbf{X}\beta + (\mathbf{A} + \mathbf{D})E[\varepsilon|\mathbf{X}] \tag{1.21}$$

$$E[\tilde{\beta}|\mathbf{X}] = \mathbf{A}\mathbf{X}\beta + \mathbf{D}\mathbf{X}\beta = \beta + \mathbf{D}\mathbf{X}\beta \tag{1.22}$$

For $E[\tilde{\beta}|\mathbf{X}] = \beta$ for any β, it is necessary that $\mathbf{D}\mathbf{X} = \mathbf{0}$. □

Step 1: Variance of OLS:

$$\text{Var}(\hat{\beta}|\mathbf{X}) = \mathbf{A}\,\text{Var}(\mathbf{y}|\mathbf{X})\mathbf{A}' = \sigma^2\mathbf{A}\mathbf{A}' = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \tag{1.23}$$

Step 2: Variance of the Alternative Estimator:

$$\text{Var}(\tilde{\beta}|\mathbf{X}) = \sigma^2\mathbf{C}\mathbf{C}' = \sigma^2(\mathbf{A} + \mathbf{D})(\mathbf{A} + \mathbf{D})' \tag{1.24}$$

$$\text{Var}(\tilde{\beta}|\mathbf{X}) = \sigma^2(\mathbf{A}\mathbf{A}' + \mathbf{A}\mathbf{D}' + \mathbf{D}\mathbf{A}' + \mathbf{D}\mathbf{D}') \tag{1.25}$$

Step 3: Simplification of Cross Terms:

$$\mathbf{A}\mathbf{D}' = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}' = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{D}\mathbf{X})' = \mathbf{0} \tag{1.26}$$

Similarly, $\mathbf{D}\mathbf{A}' = \mathbf{0}$.

Step 4: Efficiency Comparison:

$$\text{Var}(\tilde{\beta}|\mathbf{X}) = \sigma^2\mathbf{A}\mathbf{A}' + \sigma^2\mathbf{D}\mathbf{D}' = \text{Var}(\hat{\beta}|\mathbf{X}) + \sigma^2\mathbf{D}\mathbf{D}' \tag{1.27}$$

Since $\mathbf{D}\mathbf{D}'$ is an outer product matrix, it is **positive semidefinite (PSD)**. Hence:

$$\text{Var}(\tilde{\beta}|\mathbf{X}) - \text{Var}(\hat{\beta}|\mathbf{X}) = \sigma^2\mathbf{D}\mathbf{D}' \succeq \mathbf{0} \tag{1.28}$$

Conclusion: OLS is **BLUE** (Best Linear Unbiased Estimator). □

1.9 Projection and Annihilator Matrices

1.9.1 Motivation

The central problem of linear regression is to decompose a vector of observed data $\mathbf{y}$ into two mutually exclusive components: one that belongs to the subspace spanned by our explanatory data (the "explained" part) and another that is orthogonal to that subspace (the noise or "error"). The projection and annihilator matrices are the linear operators that perform this separation.

Geometric Interpretation in $\mathbb{R}^n$:

Let S be the subspace spanned by the columns of the matrix $\mathbf{X}$ (the column space of $\mathbf{X}$).

1. The **Projection Matrix (P)** acts as a "map" that orthogonally projects any vector $\mathbf{y}$ onto S, resulting in the vector of fitted values $\hat{\mathbf{y}}$.
2. The **Annihilator Matrix (M)** projects the vector $\mathbf{y}$ onto the orthogonal complement of S, resulting in the residual vector $\mathbf{e}$.

Geometrically, the vector $\mathbf{y}$ is the hypotenuse of a right triangle whose legs are $\hat{\mathbf{y}}$ and $\mathbf{e}$, ensuring the orthogonal decomposition $\mathbf{y} = \hat{\mathbf{y}} + \mathbf{e}$.

1.9.2 Formal Statement

Let $\mathbf{X} \in \mathbb{R}^{n\times K}$ with $\text{rank}(\mathbf{X}) = K$. We define:

1. **Projection Matrix (or "Hat Matrix"):**

$$\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \tag{1.29}$$

2. **Annihilator Matrix (or "Residual Maker"):**

$$\mathbf{M} = \mathbf{I}_n - \mathbf{P} \tag{1.30}$$

1.9.3 Properties of Projection Matrices

Property I: Symmetry

$$\mathbf{P}' = \mathbf{P}, \quad \mathbf{M}' = \mathbf{M} \tag{1.31}$$

Property II: Idempotence

$$\mathbf{PP} = \mathbf{P}, \quad \mathbf{MM} = \mathbf{M} \tag{1.32}$$

Property III: Orthogonality and Conservation

$$\mathbf{PX} = \mathbf{X}, \quad \mathbf{MX} = \mathbf{0}, \quad \mathbf{PM} = \mathbf{0} \tag{1.33}$$

Property IV: Trace and Rank

$$\text{tr}(\mathbf{P}) = K, \quad \text{tr}(\mathbf{M}) = n - K \tag{1.34}$$

Since for idempotent matrices $\text{rank}(\mathbf{P}) = \text{tr}(\mathbf{P})$, the rank of the projection matrix is exactly the number of regressors K.

1.9.4 Orthogonal Decomposition of y

Using the matrices $\mathbf{P}$ and $\mathbf{M}$:

$$\mathbf{y} = \mathbf{P}\mathbf{y} + \mathbf{M}\mathbf{y} = \hat{\mathbf{y}} + \mathbf{e} \tag{1.35}$$

where $\hat{\mathbf{y}} = \mathbf{P}\mathbf{y}$ are the fitted values and $\mathbf{e} = \mathbf{M}\mathbf{y}$ are the residuals.

Verification of Orthogonality:

$$\hat{\mathbf{y}}'\mathbf{e} = (\mathbf{P}\mathbf{y})'(\mathbf{M}\mathbf{y}) = \mathbf{y}'\mathbf{P}'\mathbf{M}\mathbf{y} = \mathbf{y}'\mathbf{P}\mathbf{M}\mathbf{y} = \mathbf{0} \tag{1.36}$$

1.10 Partitioned Regression and the Frisch-Waugh-Lovell Theorem

1.10.1 Motivation

The concrete problem solved by partitioned regression is the isolation of the marginal effect of a subset of variables, "cleaning" the influence of other regressors (controls). Researchers often face the dilemma of the "Short Regression" (estimated with available data) versus the "Long Regression" (the ideal population model), which generates systematically biased estimates when relevant variables are omitted.

Geometric Interpretation in $\mathbb{R}^n$:

Let S be the subspace spanned by the columns of $\mathbf{X} = [\mathbf{X}_1 \mid \mathbf{X}_2]$. The long regression estimator projects the vector $\mathbf{y}$ onto the total space S. The short regression projects $\mathbf{y}$ only onto the subspace $S_1 \subset S$ spanned by $\mathbf{X}_1$. If $\mathbf{X}_1$ and $\mathbf{X}_2$ are not orthogonal, the projection onto S_1 captures part of the variation that geometrically belongs to the direction of $\mathbf{X}_2$, "contaminating" the coefficient of $\mathbf{X}_1$. The Frisch-Waugh-Lovell (FWL) Theorem shows that the long regression coefficient can be obtained by projecting the residuals of one short regression onto the residuals of another, performing what we call *partialling out*.

1.10.2 Assumptions and Their Roles

Assumption 1.8: Relevance of the Omitted Variable ($\beta_2 \neq \mathbf{0}$)

- **Role:** Ensures that the excluded variable has a structural effect on the dependent variable in the population model.
- **Counterexample:** If we include irrelevant variables (whose true parameter is zero), the short regression will be unbiased, though less efficient.

Assumption 1.9: Correlation between Regressors ($\mathbf{X}_1'\mathbf{X}_2 \neq \mathbf{0}$)

- **Role:** Establishes that the included and omitted variables are linearly dependent in the sample, allowing the included variable to act as a partial *proxy* for the omitted one.
- **Counterexample:** If the regressors are orthogonal ($\mathbf{X}_1'\mathbf{X}_2 = \mathbf{0}$), omitting a relevant variable does not bias the coefficient of the other, only increases the residual variance.

1.10.3 Formal Statement

Consider the Partitioned Linear Regression Model:

$$\mathbf{y} = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \varepsilon \tag{1.37}$$

where $\mathbf{y} \in \mathbb{R}^n$, $\mathbf{X}_1 \in \mathbb{R}^{n \times k_1}$, $\mathbf{X}_2 \in \mathbb{R}^{n \times k_2}$, and $\varepsilon \in \mathbb{R}^n$. We assume strict exogeneity and full rank for $\mathbf{X} = [\mathbf{X}_1 \mid \mathbf{X}_2]$.

We define:

1. **Short Regression Estimator:** $\tilde{\beta}_1 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y}$.
2. **Annihilator Matrix of $\mathbf{X}_1$:** $\mathbf{M}_1 = \mathbf{I} - \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'$.

Theorem (Omitted Variable Bias). Under Assumptions 1.8 and 1.9:

$$\boxed{E[\tilde{\beta}_1|\mathbf{X}] = \beta_1 + (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\beta_2} \tag{1.38}$$

Proof:

$$E[\tilde{\beta}_1|\mathbf{X}] = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'E[\mathbf{y}|\mathbf{X}] = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'(\mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2) \tag{1.39}$$

$$E[\tilde{\beta}_1|\mathbf{X}] = \beta_1 + (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\beta_2 \tag{1.40}$$

□

1.10.4 Frisch-Waugh-Lovell (FWL) Theorem

Theorem (Frisch-Waugh-Lovell). The OLS estimator for β_2 in model (1.37) is identical to the estimator obtained by regressing the residuals of $\mathbf{y}$ (with respect to $\mathbf{X}_1$) on the residuals of $\mathbf{X}_2$ (with respect to $\mathbf{X}_1$):

$$\boxed{\hat{\beta}_2 = (\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{M}_1\mathbf{y}} \tag{1.41}$$

Proof (sketch): The partitioned normal equations are:

$$\begin{bmatrix} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{X}_2 \\ \mathbf{X}_2'\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1'\mathbf{y} \\ \mathbf{X}_2'\mathbf{y} \end{bmatrix} \tag{1.42}$$

From the first row: $\hat{\beta}_1 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'(\mathbf{y} - \mathbf{X}_2\hat{\beta}_2)$. Substituting into the second row and rearranging:

$$\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2\hat{\beta}_2 = \mathbf{X}_2'\mathbf{M}_1\mathbf{y} \tag{1.43}$$

Isolating $\hat{\beta}_2$, we obtain (1.41). □

1.11 Relative Efficiency: Short vs. Long Regression

1.11.1 Motivation

The concrete problem here is the **bias-variance trade-off** in model selection. When deciding whether to include a set of controls ($\mathbf{X}_2$) to estimate the effect of interest ($\mathbf{X}_1$), we face a dilemma: the **Long Regression** (with $\mathbf{X}_2$) eliminates omitted variable bias, but the **Short Regression** (without $\mathbf{X}_2$) may offer more "stable" estimates (lower variance), especially if there is collinearity among the regressors.

Geometric Interpretation in $\mathbb{R}^n$:

Think of the variance of an estimator as the "volume" of its confidence ellipsoid.

1. In the **Short Regression**, we project $\mathbf{y}$ directly onto $\text{Col}(\mathbf{X}_1)$. Precision depends only on the dispersion of $\mathbf{X}_1$.

2. In the **Long Regression**, the FWL Theorem tells us that we estimate the coefficient of $\mathbf{X}_1$ by projecting $\mathbf{y}$ onto the part of $\mathbf{X}_1$ that is **orthogonal** to $\mathbf{X}_2$ (i.e., $\mathbf{M}_2\mathbf{X}_1$).

If $\mathbf{X}_1$ and $\mathbf{X}_2$ are highly correlated, the residual vector $\mathbf{M}_2\mathbf{X}_1$ will have a very small norm. Geometrically, we are trying to extract information from a very narrow "shadow" of $\mathbf{X}_1$, which amplifies uncertainty and increases the variance of the estimator.

1.11.2 Formal Statement

Let the DGP be given by the Long Regression (1.37). Under homoskedasticity, the variances of the estimators for β_1 are:

1. **Long Regression:** $\text{Var}(\hat{\beta}_1|\mathbf{X}) = \sigma^2(\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1)^{-1}$.
2. **Short Regression:** $\text{Var}(\tilde{\beta}_1|\mathbf{X}) = \sigma^2(\mathbf{X}_1'\mathbf{X}_1)^{-1}$.

Theorem (Efficiency of the Short Regression). Under homoskedasticity:

$$\boxed{\text{Var}(\hat{\beta}_1|\mathbf{X}) \succeq \text{Var}(\tilde{\beta}_1|\mathbf{X})} \tag{1.44}$$

Proof:

$$\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1 = \mathbf{X}_1'\mathbf{X}_1 - \mathbf{X}_1'\mathbf{P}_2\mathbf{X}_1 \preceq \mathbf{X}_1'\mathbf{X}_1 \tag{1.45}$$

By the Matrix Inversion Lemma (if $\mathbf{A} \preceq \mathbf{B}$, then $\mathbf{A}^{-1} \succeq \mathbf{B}^{-1}$):

$$\sigma^2(\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1)^{-1} \succeq \sigma^2(\mathbf{X}_1'\mathbf{X}_1)^{-1} \tag{1.46}$$

□

1.11.3 Practical Implications

The short regression is always **more efficient** (has lower variance) than the long regression, except in the case of orthogonality ($\mathbf{X}_1'\mathbf{X}_2 = \mathbf{0}$), where the variances are equal.

Indispensable Assumption: Homoskedasticity ($\sigma^2\mathbf{I}$) is indispensable for the mathematical proof. In the presence of heteroskedasticity, the short variance matrix becomes a "sandwich" and it is theoretically possible for the long regression to be more efficient if $\mathbf{X}_2$ helps explain the variance structure of the error, although this rarely occurs with standard OLS.

1.12 Estimator of the Residual Variance (s^2) and Distribution of Residuals

1.12.1 Motivation

The concrete problem here is **measuring uncertainty**. The Ordinary Least Squares (OLS) estimator provides the "best fit," but does not reveal the magnitude of the random error (σ^2). Without an estimate of σ^2, it is impossible to compute standard errors or perform hypothesis tests. The estimator s^2 solves this by correcting the downward bias that would occur if we used the simple average of the squared residuals.

Geometric Interpretation in $\mathbb{R}^n$:

The vector of observations $\mathbf{y}$ inhabits a space of dimension n. The model projects $\mathbf{y}$ onto a subspace of dimension K (spanned by the regressors). The residual vector $\mathbf{e}$ is the projection of $\mathbf{y}$ onto the **orthogonal complement** of that subspace. Since this complement has dimension $n - K$, the residual vector has only $n - K$ independent directions in which to vary. Dividing the sum of squared residuals by $n - K$ (degrees of freedom) is equivalent to calculating the average variance per available dimension in the residual space.

1.12.2 Additional Assumptions

Assumption 1.10: Normality of Errors ($\varepsilon|\mathbf{X} \sim N(\mathbf{0}, \sigma^2\mathbf{I}_n)$)

- **Role:** Necessary to derive the exact distribution of residuals and for finite-sample inference.
- **Counterexample:** Without normality, the distribution of residuals is not exactly known, although asymptotically it approaches a normal distribution.

1.12.3 Formal Statement

Let the model be $\mathbf{y} = \mathbf{X}\beta + \varepsilon$, with $\varepsilon \sim N(\mathbf{0}, \sigma^2\mathbf{I}_n)$ and $\text{rank}(\mathbf{X}) = K$. We define:

1. **Projection Matrix (P):** $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.
2. **Annihilator Matrix (M):** $\mathbf{M} = \mathbf{I}_n - \mathbf{P}$.
3. **Residual Vector (e):** $\mathbf{e} = \mathbf{M}\mathbf{y}$.
4. **Variance Estimator (s^2):** $s^2 = \frac{\mathbf{e}'\mathbf{e}}{n-K}$.

1.12.4 Proof of Unbiasedness of s^2

Auxiliary Lemma 1 (Properties of Matrix M): $\mathbf{M}$ is symmetric ($\mathbf{M}' = \mathbf{M}$), idempotent ($\mathbf{M}\mathbf{M} = \mathbf{M}$), and orthogonal to $\mathbf{X}$ ($\mathbf{M}\mathbf{X} = \mathbf{0}$).

Auxiliary Lemma 2 (Expectation of Quadratic Forms): For a random vector $\mathbf{z}$ with mean μ and covariance matrix $\mathbf{\Sigma}$, $E[\mathbf{z}'\mathbf{A}\mathbf{z}] = \text{tr}(\mathbf{A}\mathbf{\Sigma}) + \mu'\mathbf{A}\mu$.

Theorem (Unbiasedness of s^2). Under Assumptions 1.1 to 1.4:

$$\boxed{E[s^2|\mathbf{X}] = \sigma^2} \tag{1.47}$$

Proof:

Step 1: Express the Sum of Squared Residuals (SSR) in terms of the population error:

$$\mathbf{e} = \mathbf{M}\mathbf{y} = \mathbf{M}(\mathbf{X}\beta + \varepsilon) = \mathbf{M}\mathbf{X}\beta + \mathbf{M}\varepsilon = \mathbf{M}\varepsilon \tag{1.48}$$

Hence, the SSR is:

$$\mathbf{e}'\mathbf{e} = \varepsilon'\mathbf{M}'\mathbf{M}\varepsilon = \varepsilon'\mathbf{M}\varepsilon \tag{1.49}$$

Step 2: Apply the Expectation Operator:

$$E[\mathbf{e}'\mathbf{e}|\mathbf{X}] = E[\varepsilon'\mathbf{M}\varepsilon|\mathbf{X}] \tag{1.50}$$

Applying Lemma 2 (with $\mu = \mathbf{0}$ and $\mathbf{\Sigma} = \sigma^2\mathbf{I}$):

$$E[\mathbf{e}'\mathbf{e}|\mathbf{X}] = \text{tr}(\mathbf{M} \cdot \sigma^2\mathbf{I}) + \mathbf{0}'\mathbf{M}\mathbf{0} = \sigma^2 \,\text{tr}(\mathbf{M}) \tag{1.51}$$

Step 3: Compute the Trace of M:

$$\text{tr}(\mathbf{M}) = \text{tr}(\mathbf{I}_n - \mathbf{P}) = n - \text{tr}(\mathbf{P}) = n - \text{tr}\left(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right) \tag{1.52}$$

By the cyclic property of the trace ($\text{tr}(\mathbf{A}\mathbf{B}) = \text{tr}(\mathbf{B}\mathbf{A})$):

$$\operatorname{tr}(\mathbf{M}) = n - \operatorname{tr}\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\right) = n - \operatorname{tr}(\mathbf{I}_K) = n - K \tag{1.53}$$

Step 4: Finalize the Proof:

$$E[\mathbf{e}'\mathbf{e}|\mathbf{X}] = \sigma^2(n-K) \tag{1.54}$$

Taking the expectation of s^2:

$$E[s^2|\mathbf{X}] = E\left[\frac{\mathbf{e}'\mathbf{e}}{n-K}\middle|\mathbf{X}\right] = \frac{\sigma^2(n-K)}{n-K} = \sigma^2 \tag{1.55}$$

□

1.12.5 Proof of the Distribution of Residuals

Auxiliary Lemma 3 (Transformation of Normal Vectors): If $\mathbf{z} \sim N(\mu, \mathbf{\Sigma})$, then for a constant matrix $\mathbf{A}$, the vector $\mathbf{Az} \sim N(\mathbf{A}\mu, \mathbf{A\Sigma A}')$.

Theorem (Distribution of Residuals). Under Assumption 1.10 (normality):

$$\boxed{\mathbf{e}|\mathbf{X} \sim N(\mathbf{0}, \sigma^2\mathbf{M})} \tag{1.56}$$

Proof:

Step 1: Identify the linear form:

From equation (1.48), $\mathbf{e} = \mathbf{M}\varepsilon$. Since $\varepsilon \sim N(\mathbf{0}, \sigma^2\mathbf{I}_n)$ by assumption, the residuals are a linear transformation of a normal vector.

Step 2: Compute the First Moment:

$$E[\mathbf{e}|\mathbf{X}] = \mathbf{M}E[\varepsilon|\mathbf{X}] = \mathbf{M}\cdot\mathbf{0} = \mathbf{0} \tag{1.57}$$

Step 3: Compute the Second Moment (Covariance Matrix):

$$\operatorname{Var}(\mathbf{e}|\mathbf{X}) = \mathbf{M}\operatorname{Var}(\varepsilon|\mathbf{X})\mathbf{M}' = \mathbf{M}(\sigma^2\mathbf{I}_n)\mathbf{M} = \sigma^2(\mathbf{MM}) = \sigma^2\mathbf{M} \tag{1.58}$$

Conclusion: The residual vector follows a multivariate normal distribution:

$$\mathbf{e}|\mathbf{X} \sim N(\mathbf{0}, \sigma^2\mathbf{M}) \tag{1.59}$$

□

1.12.6 Independence between $\hat{\beta}$ and e

Theorem (Independence between $\hat{\beta}$ and e). Under Assumption 1.10 (normality):

$$\boxed{\hat{\beta} \perp \mathbf{e} \mid \mathbf{X}} \tag{1.60}$$

Proof:

$\hat{\beta} = \beta + \mathbf{A}\varepsilon$, where $\mathbf{A} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. $\mathbf{e} = \mathbf{M}\varepsilon$.

We compute the covariance between the random terms:

$$\text{Cov}(\mathbf{A}\varepsilon, \mathbf{M}\varepsilon|\mathbf{X}) = \mathbf{A}\,\text{Var}(\varepsilon|\mathbf{X})\mathbf{M}' = \sigma^2\mathbf{AM} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{M} = \mathbf{0} \tag{1.61}$$

Since $\hat{\beta}$ and $\mathbf{e}$ are linear transformations of a normal vector, zero covariance implies **stochastic independence**. □

1.13 Summary of CLRM Properties

Property	Expression	Condition
OLS Estimator	$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$	Full Rank
Unbiasedness	$E[\hat{\beta}\|\mathbf{X}] = \beta$	Strict Exogeneity
Variance of OLS	$\text{Var}(\hat{\beta}\|\mathbf{X}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$	Sphericity
BLUE	$\text{Var}(\hat{\beta}) \preceq \text{Var}(\tilde{\beta})$	Gauss-Markov
Estimator of σ^2	$s^2 = \frac{\mathbf{e}'\mathbf{e}}{n-K}$	Sphericity
Unbiasedness of s^2	$E[s^2\|\mathbf{X}] = \sigma^2$	Sphericity
Distribution of Residuals	$\mathbf{e} \sim N(\mathbf{0}, \sigma^2\mathbf{M})$	Normality
Independence	$\hat{\beta} \perp \mathbf{e} \mid \mathbf{X}$	Normality

Chapter 2: Hypothesis Testing in the Classical Normal Linear Regression Model

This chapter establishes the theoretical foundations for statistical inference in the Classical Normal Linear Regression Model (CNLRM), covering from the derivation of the t-statistic for hypotheses about a single parameter to the F-statistic for multiple linear restrictions, including confidence intervals and the trinity of asymptotic tests. The material is organized in progressive sections, moving from geometric intuition to rigorous algebraic formalism.

2.1 The Classical Normal Linear Regression Model (CNLRM)

2.1.1 Motivation

The concrete problem here is the **validity of inference**. Until now, the Ordinary Least Squares (OLS) method allowed us to obtain point estimates ($\hat{\beta}$), but did not tell us how likely it would be to obtain these values if the true relationship were null. By adding the normality assumption, we transform the model from an algebraic approximation tool into a complete probabilistic model. This allows us to determine the exact distribution of the estimators and construct tests that quantify uncertainty under unknown variance.

Geometric Interpretation in $\mathbb{R}^n$:

Under the normality assumption, the vector of population errors ε is not only orthogonal to the column space of $\mathbf{X}$ in expectation, but its probability density is spherically symmetric in $\mathbb{R}^n$. When we project $\mathbf{y}$ onto the subspace spanned by $\mathbf{X}$, the distance between the observed value and the projection (the residual) follows a distribution related to χ^2, allowing us to "measure" the error in a standardized way.

2.1.2 CNLRM Assumptions

For the construction of exact tests, we add the following assumption to the CLRM:

Assumption 2.1: Normality of Errors ($\varepsilon|\mathbf{X} \sim N(\mathbf{0}, \sigma^2\mathbf{I}_n)$)

- **Role:** Ensures that any linear combination of y (such as the estimators $\hat{\beta}$) also has a normal distribution. This is the basis for the ratio between the estimation error and the standard error to follow a Student's t distribution.

- **Counterexample:** If the errors follow a heavy-tailed distribution (such as Cauchy), the OLS estimator can still be computed, but the test statistic will not follow a t distribution in small samples, leading to incorrect conclusions about the significance of variables.

2.2 Inference about a Single Parameter: The t-Statistic

2.2.1 Motivation

The concrete problem solved by the t-statistic is hypothesis testing about a single regression coefficient when the population variance σ^2 is unknown. Instead of substituting σ^2 with an ad hoc estimator, the t-statistic incorporates the additional uncertainty from estimating the variance, resulting in a distribution with heavier tails than the Normal.

Geometric Interpretation in $\mathbb{R}^2$:

For a simple regression, we can visualize the estimator $\hat{\beta}_2$ as a random variable centered at β_2. The standard error $se(\hat{\beta}_2)$ measures the dispersion of this variable. The t-statistic standardizes the distance between the estimate and the tested value, dividing it by the estimated standard error. The denominator s (root of the estimated variance) introduces additional variability, making the resulting distribution have heavier tails than the Normal.

2.2.2 Formal Statement

Consider the Simple Linear Regression Model defined for n observations:

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i, \quad i = 1, \dots, n \tag{2.1}$$

where y_i is the dependent variable, x_i is the non-stochastic regressor, β_1 is the intercept, β_2 is the slope, and ε_i is the stochastic error.

We assume the following conditions of the Classical Normal Model:

1. **Exogeneity and Normality:** $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$, where σ^2 is constant (homoskedastic) and unknown.
2. **Identifiability:** There is variation in x, such that $\sum_{i=1}^{n}(x_i - \bar{x})^2 > 0$.

Theorem (Distribution of the t-Statistic). Under the CNLRM assumptions, the statistic:

$$\boxed{T = \frac{\hat{\beta}_2 - \beta_2^0}{se(\hat{\beta}_2)} \sim t_{n-2}} \tag{2.2}$$

where $se(\hat{\beta}_2) = \frac{s}{\sqrt{\sum (x_i - \bar{x})^2}}$ and $s^2 = \frac{\sum_{i=1}^{n} \hat{e}_i^2}{n-2}$, follows a Student's t distribution with $n-2$ degrees of freedom.

2.2.3 Formal Derivation

Step 1: Derivation of the OLS Estimators ($\hat{\beta}_1, \hat{\beta}_2$)

We minimize the sum of squared residuals $S(b_1, b_2) = \sum_{i=1}^{n}(y_i - b_1 - b_2 x_i)^2$.

The First Order Conditions (FOC) are:

$$\frac{\partial S}{\partial b_1} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) = 0 \tag{2.3}$$

$$\frac{\partial S}{\partial b_2} = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i)x_i = 0 \tag{2.4}$$

From (2.3):

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x} \tag{2.5}$$

Substituting (2.5) into (2.4):

$$\hat{\beta}_2 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \tag{2.6}$$

Step 2: Variance Estimator (s^2)

We define the residual as $\hat{e}_i = y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i$. The unbiased estimator of the variance is:

$$s^2 = \frac{\sum_{i=1}^{n} \hat{e}_i^2}{n-2} \tag{2.7}$$

(Justification: We lose 2 degrees of freedom when estimating β_1 and β_2.)

Step 3: Sampling Distribution and Construction of the Test Statistic

Under the normality assumption, since $\hat{\beta}_2$ is a linear function of y (which is normal):

$$\hat{\beta}_2 \sim N\left(\beta_2, \frac{\sigma^2}{\sum(x_i - \bar{x})^2}\right) \tag{2.8}$$

To test $H_0 : \beta_2 = \beta_2^0$, we standardize the estimator:

$$Z = \frac{\hat{\beta}_2 - \beta_2^0}{\sigma / \sqrt{\sum(x_i - \bar{x})^2}} \sim N(0,1) \tag{2.9}$$

Since σ is unknown, we use s. The associated Chi-square variable is:

$$V = \frac{(n-2)s^2}{\sigma^2} \sim \chi^2_{n-2} \tag{2.10}$$

Auxiliary Lemma (Definition of the Student's t Distribution):

If $Z \sim N(0,1)$ and $V \sim \chi^2_\nu$ are independent, then $T = \frac{Z}{\sqrt{V/\nu}} \sim t_\nu$.

Applying the lemma with $\nu = n - 2$:

$$T = \frac{\frac{\hat{\beta}_2 - \beta_2^0}{\sigma / \sqrt{\sum(x_i - \bar{x})^2}}}{\sqrt{\frac{(n-2)s^2}{\sigma^2}/(n-2)}} = \frac{\hat{\beta}_2 - \beta_2^0}{s / \sqrt{\sum(x_i - \bar{x})^2}} \tag{2.11}$$

Defining the estimated Standard Error as $se(\hat{\beta}_2) = \frac{s}{\sqrt{\sum(x_i - \bar{x})^2}}$:

$$T = \frac{\hat{\beta}_2 - \beta_2^0}{se(\hat{\beta}_2)} \sim t_{n-2} \tag{2.12}$$

□

2.3 Distribution of Linear Combinations of OLS

2.3.1 Motivation

The concrete problem here is **inference about multiple simultaneous hypotheses**. Frequently, economic or scientific theory does not predict only the value of a single parameter, but a relationship between them (e.g., constant returns to scale, where the sum of coefficients must be unity).

Geometric Interpretation in $\mathbb{R}^K$:

Let $\beta \in \mathbb{R}^K$ be the vector of population parameters. A single linear restriction defines a hyperplane of dimension $K-1$ in the parameter space. A set of J linearly independent restrictions defines the intersection of J hyperplanes, resulting in an affine subspace of dimension $K-J$. Testing whether the restriction is valid is equivalent to measuring the probabilistic distance between the unrestricted estimator $\hat{\beta}$ (the point that minimizes the sum of squares in the total space) and the restricted subspace.

2.3.2 Additional Assumptions

Assumption 2.2: Full Column Rank ($\text{rank}(\mathbf{X}) = K$)

- **Role:** Guarantees the identifiability of $\hat{\beta}$ and the invertibility of $(\mathbf{X}'\mathbf{X})$.
- **Counterexample:** If the columns of $\mathbf{X}$ are linearly dependent, there is no unique point in the parameter space to compare with the restrictions.

Assumption 2.3: Normality of Errors ($\varepsilon|\mathbf{X} \sim N(\mathbf{0}, \sigma^2\mathbf{I}_n)$)

- **Role:** This is the generating assumption. Since the estimator is a linear transformation of the errors, the normality of ε is inherited by $\hat{\beta}$ in any finite sample.
- **Counterexample:** Without normality, $\hat{\beta}$ follows a normal distribution only asymptotically ($n \to \infty$). In small samples, the distribution would be unknown, invalidating exact tests.

Assumption 2.4: Independence and Rank of Restrictions ($\text{rank}(\mathbf{R}) = J < K$)

- **Role:** Ensures that the restrictions are not redundant or contradictory, guaranteeing that the variance-covariance matrix of the restrictions is invertible.
- **Counterexample:** If $\mathbf{R}$ does not have full rank, the "ellipse of uncertainty" of the restrictions will be collapsed in some dimension, preventing the calculation of the test statistic.

2.3.3 Formal Statement

Consider the model $\mathbf{y} = \mathbf{X}\beta + \varepsilon$, with $\mathbf{y} \in \mathbb{R}^n$, $\mathbf{X} \in \mathbb{R}^{n \times K}$, $\text{rank}(\mathbf{X}) = K$, and $\varepsilon \sim N(\mathbf{0}, \sigma^2\mathbf{I}_n)$.

Consider a set of J linear restrictions on β of the form:

$$\mathbf{R}\beta = \mathbf{q} \tag{2.13}$$

where $\mathbf{R}$ is a $J \times K$ matrix of constants of rank J, and $\mathbf{q}$ is a $J \times 1$ vector of constants.

The unrestricted OLS estimator is $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.

Theorem (Distribution of the Restriction Vector). Under the CNLRM assumptions, the vector of sample restrictions $\mathbf{m} = \mathbf{R}\hat{\beta} - \mathbf{q}$ follows a multivariate normal distribution:

$$\boxed{\mathbf{m}|\mathbf{X} \sim N\left(\mathbf{R}\beta - \mathbf{q},\, \sigma^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\right)} \tag{2.14}$$

2.3.4 Formal Derivation

Auxiliary Lemma 1 (Affine Transformation of Normal Vectors):

If $\mathbf{z} \sim N(\mu, \mathbf{\Sigma})$ and $\mathbf{w} = \mathbf{a} + \mathbf{B}\mathbf{z}$ (where $\mathbf{a}$ and $\mathbf{B}$ are constants), then $\mathbf{w} \sim N(\mathbf{a} + \mathbf{B}\mu, \mathbf{B}\mathbf{\Sigma}\mathbf{B}')$.

Step 1: Distribution of y conditional on X

Since $\mathbf{y} = \mathbf{X}\beta + \varepsilon$, and $\varepsilon \sim N(\mathbf{0}, \sigma^2\mathbf{I}_n)$:

$$E[\mathbf{y}|\mathbf{X}] = \mathbf{X}\beta, \qquad \text{Var}(\mathbf{y}|\mathbf{X}) = \sigma^2\mathbf{I}_n \tag{2.15}$$

Hence:

$$\mathbf{y}|\mathbf{X} \sim N(\mathbf{X}\beta, \sigma^2\mathbf{I}_n) \tag{2.16}$$

Step 2: Distribution of the Estimator $\hat{\beta}$

The OLS estimator is $\hat{\beta} = \mathbf{B}\mathbf{y}$, where $\mathbf{B} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.

Applying Lemma 1:

$$E[\hat{\beta}|\mathbf{X}] = \beta, \qquad \text{Var}(\hat{\beta}|\mathbf{X}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \tag{2.17}$$

Hence:

$$\hat{\beta}|\mathbf{X} \sim N\left(\beta,\, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\right) \tag{2.18}$$

Step 3: Distribution of the Restriction Matrix m

We define the discrepancy vector $\mathbf{m} = \mathbf{R}\hat{\beta} - \mathbf{q}$.

Since $\mathbf{m}$ is an affine function of $\hat{\beta}$, we apply Lemma 1 again:

$$E[\mathbf{m}|\mathbf{X}] = \mathbf{R}\beta - \mathbf{q}, \qquad \text{Var}(\mathbf{m}|\mathbf{X}) = \sigma^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}' \tag{2.19}$$

Therefore:

$$\mathbf{m}|\mathbf{X} \sim N\left(\mathbf{R}\beta - \mathbf{q},\, \sigma^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\right) \tag{2.20}$$

□

2.4 Cholesky Decomposition

2.4.1 Motivation

The concrete problem solved by the Cholesky decomposition is the **"standardization" of multivariate linear systems**. Just as in scalar statistics we divide a variable by its standard deviation to obtain unit variance, in matrices we seek an operator $\mathbf{L}$ such that, when applied to a system with variance-covariance matrix $\mathbf{\Sigma}$, it results in a system with identity $\mathbf{I}$.

Geometric Interpretation in $\mathbb{R}^n$:

A covariance matrix $\mathbf{\Sigma}$ defines the geometry of an ellipsoid of uncertainty. The Cholesky decomposition extracts a lower triangular matrix $\mathbf{L}$ that functions as a coordinate system. Multiplying a white noise vector (spherical) by $\mathbf{L}$ "stretches" and "rotates" this vector so that it assumes the shape of the ellipsoid defined by $\mathbf{\Sigma}$. Conversely, multiplying correlated data by $\mathbf{L}^{-1}$ (the *whitening* process) collapses the ellipsoid back to a unit sphere.

2.4.2 Formal Statement

Let $\mathbf{A} \in \mathbb{R}^{K \times K}$ be a square, symmetric, and positive definite matrix. The **Cholesky Theorem** states that there exists a unique lower triangular matrix $\mathbf{L} \in \mathbb{R}^{K \times K}$, with strictly positive diagonal elements ($L_{ii} > 0$), such that:

$$\boxed{\mathbf{A} = \mathbf{L}\mathbf{L}'} \tag{2.21}$$

where $\mathbf{L}$ is referred to as the **Cholesky square root** of $\mathbf{A}$.

2.4.3 Algorithmic Derivation

Notation Definition:

- A_{ij}: element in row i and column j of $\mathbf{A}$.
- L_{ij}: element in row i and column j of $\mathbf{L}$.
- $\mathbf{L}$ is lower triangular $\implies L_{ij} = 0$ for $j > i$.

Step 1: Expansion of the Matrix Product

By the definition of matrix multiplication:

$$A_{ij} = \sum_{k=1}^{K} L_{ik} L_{jk} \tag{2.22}$$

Since $\mathbf{L}$ is lower triangular:

$$A_{ij} = \sum_{k=1}^{j} L_{ik} L_{jk} \tag{2.23}$$

Step 2: Diagonal Elements ($i = j$)

For a generic diagonal element L_{ii}:

$$A_{ii} = \sum_{k=1}^{i-1} L_{ik}^2 + L_{ii}^2 \tag{2.24}$$

Isolating:

$$L_{ii} = \sqrt{A_{ii} - \sum_{k=1}^{i-1} L_{ik}^2} \tag{2.25}$$

Step 3: Off-Diagonal Elements ($i > j$)

From (2.23) and isolating L_{ij}:

$$L_{ij} = \frac{1}{L_{jj}} \left(A_{ij} - \sum_{k=1}^{j-1} L_{ik} L_{jk} \right) \tag{2.26}$$

Auxiliary Lemma (Uniqueness):

Suppose $\mathbf{A} = \mathbf{L}_1\mathbf{L}_1' = \mathbf{L}_2\mathbf{L}_2'$. Since $\mathbf{A}$ is positive definite, it is invertible. This implies $\mathbf{L}_1^{-1}\mathbf{L}_2 = \mathbf{L}_1'(\mathbf{L}_2')^{-1}$. The left-hand side is lower triangular and the right-hand side is upper triangular. A matrix that is simultaneously lower and upper triangular must be diagonal. Since the diagonals are positive, the only solution is the identity, hence $\mathbf{L}_1 = \mathbf{L}_2$.

2.5 Multivariate Standardization and the χ^2 Statistic

2.5.1 Motivation

The concrete problem solved here is the **comparison of distances in correlated spaces**. When we test multiple simultaneous hypotheses (e.g., $\mathbf{R}\beta = \mathbf{q}$), the estimation error in one parameter may be correlated with the error in another. If we tried to sum the squares of these errors directly, we would be ignoring the geometry of the system.

Geometric Interpretation in $\mathbb{R}^J$:

Imagine that the uncertainty of the restrictions forms an ellipsoid of probability density. The Cholesky Decomposition extracts the dependence structure of this ellipsoid.

1. **Normal to Standard:** The Cholesky factor acts as a "whitening" operator, rotating and collapsing the uncertainty ellipsoid into a perfect unit-radius sphere.
2. **Standard to Chi-square:** Once we have a unit sphere, the sum of the squares of the distances from the center to the surface follows, by definition, a χ^2 distribution.

2.5.2 Formal Statement

Let $\mathbf{m} = \mathbf{R}\hat{\beta} - \mathbf{q}$ be the discrepancy vector of the restrictions of dimension $J \times 1$. Under the null hypothesis $H_0 : \mathbf{R}\beta = \mathbf{q}$, we assume:

1. **Distribution:** $\mathbf{m} \sim N(\mathbf{0}, \mathbf{V}_m)$, where $\mathbf{V}_m = \sigma^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'$.
2. **Variance Structure:** $\mathbf{V}_m$ is a $J \times J$ symmetric and positive definite matrix.
3. **Decomposition:** There exists a lower triangular matrix $\mathbf{L}$ such that $\mathbf{V}_m = \mathbf{L}\mathbf{L}'$.

Theorem (Wald Statistic). The Wald statistic for testing $H_0 : \mathbf{R}\beta = \mathbf{q}$ is:

$$\boxed{W = \mathbf{m}'\mathbf{V}_m^{-1}\mathbf{m} \sim \chi^2_J} \tag{2.27}$$

2.5.3 Formal Derivation

Step 1: Transformation to Standard Normal (Whitening)

We define the transformed vector $\mathbf{z} = \mathbf{L}^{-1}\mathbf{m}$.

By the Affine Transformation Lemma:

$$E[\mathbf{z}|\mathbf{X}] = \mathbf{L}^{-1}E[\mathbf{m}|\mathbf{X}] = \mathbf{0} \tag{2.28}$$

$$\text{Var}(\mathbf{z}|\mathbf{X}) = \mathbf{L}^{-1}\mathbf{V}_m(\mathbf{L}^{-1})' = \mathbf{L}^{-1}(\mathbf{L}\mathbf{L}')(\mathbf{L}^{-1})' = \mathbf{I}_J \tag{2.29}$$

Hence:

$$\mathbf{z} \sim N(\mathbf{0}, \mathbf{I}_J) \tag{2.30}$$

Step 2: Transformation to Chi-square (χ^2)

We define the test statistic W as the inner product (sum of squares) of the standardized vector:

$$W = \mathbf{z}'\mathbf{z} = \mathbf{m}'(\mathbf{L}^{-1})'\mathbf{L}^{-1}\mathbf{m} = \mathbf{m}'\mathbf{V}_m^{-1}\mathbf{m} \tag{2.31}$$

Auxiliary Lemma (Distribution of Quadratic Forms):

If $\mathbf{z} \sim N(\mathbf{0}, \mathbf{I}_J)$, then $\mathbf{z}'\mathbf{z} = \sum_{j=1}^{J} z_j^2 \sim \chi^2_J$.

Applying the lemma:

$$W \sim \chi^2_J \tag{2.32}$$

□

2.6 The Snedecor F Statistic

2.6.1 Motivation

Until now, we have derived test statistics that depended on the population parameter σ^2. However, in real applications, this parameter is invariably unknown. The problem solved by the F statistic is to allow simultaneous inference about J linear restrictions using an estimate of the variance based on the model residuals (s^2).

Geometric Interpretation in $\mathbb{R}^n$:

The F statistic can be viewed as the ratio between two normalized orthogonal distances in $\mathbb{R}^n$. The numerator measures the (squared) length of the projection of $\mathbf{y}$ that is "lost" when we impose the restrictions of H_0. The denominator measures the (squared) length of the residual vector of the unrestricted model, which represents the inherent variability of the error. The independence between these two measures ensures that the ratio follows an F distribution.

2.6.2 Formal Statement

Consider the model $\mathbf{y} = \mathbf{X}\beta + \varepsilon$ with $\varepsilon \sim N(\mathbf{0}, \sigma^2\mathbf{I}_n)$ and $\text{rank}(\mathbf{X}) = K$. Consider the null hypothesis $H_0 : \mathbf{R}\beta = \mathbf{q}$, where $\mathbf{R}$ is $J \times K$ with $\text{rank}(\mathbf{R}) = J$.

We define:

1. **Variance Estimator (s^2):** $s^2 = \frac{\mathbf{e}'\mathbf{e}}{n-K}$, where $\mathbf{e} = \mathbf{M}\mathbf{y}$.
2. **Discrepancy Vector (m):** $\mathbf{m} = \mathbf{R}\hat{\beta} - \mathbf{q}$.

Theorem (F Statistic). Under the CNLRM assumptions, the statistic:

$$\boxed{F = \frac{\mathbf{m}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}\mathbf{m}/J}{s^2} \sim F_{J,n-K}} \tag{2.33}$$

follows a Snedecor F distribution with J and $n-K$ degrees of freedom.

2.6.3 Formal Derivation

Step 1: Distribution of s^2

We start from the definition of residuals: $\mathbf{e} = \mathbf{M}\mathbf{y}$, where $\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.

Substituting the model structure:

$$\mathbf{e} = \mathbf{M}(\mathbf{X}\beta + \varepsilon) = \mathbf{M}\varepsilon \tag{2.34}$$

The sum of squared residuals (SSR) is:

$$\mathbf{e}'\mathbf{e} = \varepsilon'\mathbf{M}\varepsilon \tag{2.35}$$

Auxiliary Lemma 2 (Distribution of Quadratic Forms):

If $\mathbf{z} \sim N(\mathbf{0}, \mathbf{I})$ and $\mathbf{A}$ is idempotent with $\text{rank}(\mathbf{A}) = \nu$, then $\mathbf{z}'\mathbf{A}\mathbf{z} \sim \chi^2_\nu$.

Applying the Lemma:

$$\frac{(n-K)s^2}{\sigma^2} \sim \chi^2_{n-K} \tag{2.36}$$

Step 2: Independence between Numerator and Denominator

The Wald statistic W depends on $\hat{\beta}$, while s^2 depends on $\mathbf{e}$.

$$\hat{\beta} = \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon, \qquad \mathbf{e} = \mathbf{M}\varepsilon \tag{2.37}$$

We compute the covariance:

$$\text{Cov}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon, \mathbf{M}\varepsilon|\mathbf{X}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{M} = \mathbf{0} \tag{2.38}$$

Since $\hat{\beta}$ and $\mathbf{e}$ are linear transformations of a normal vector, zero covariance implies **stochastic independence**.

Step 3: Derivation of the F Statistic

We define $W^* = \mathbf{m}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}\mathbf{m}$. Note that $W = W^*/\sigma^2 \sim \chi^2_J$.

By the definition of the F distribution (ratio of two independent χ^2 variables divided by their respective degrees of freedom):

$$F = \frac{W/J}{\frac{(n-K)s^2}{\sigma^2}/(n-K)} \tag{2.39}$$

Canceling σ^2:

$$F = \frac{\mathbf{m}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}\mathbf{m}/J}{s^2} \tag{2.40}$$

Applying the derived distributions:

$$F \sim \frac{\chi^2_J/J}{\chi^2_{n-K}/(n-K)} = F_{J,n-K} \tag{2.41}$$

□

2.7 The Feasible Wald Test

2.7.1 Motivation

The concrete problem solved here is the **operationalization of multiple hypothesis tests** when the noise variance (σ^2) is unknown. While the theoretical Wald statistic uses the true variance, the feasible test substitutes this parameter with its sample estimate (s^2), allowing the researcher to perform tests of linear restrictions using only the sample data.

Geometric Interpretation in $\mathbb{R}^J$:

Geometrically, the Wald statistic is a measure of the **weighted Euclidean distance** between the vector of sample restrictions ($\mathbf{R}\hat{\beta}$) and the vector of restrictions under the null hypothesis ($\mathbf{q}$). The use of s^2 in the denominator acts as an "adaptive ruler": if the estimated noise in the data is high, the measured distance between the coefficients and the hypothesis must be significantly larger to be considered statistically relevant. The F statistic is nothing more than this distance normalized by the number of dimensions (restrictions) being tested simultaneously.

2.7.2 Formal Statement

Consider the model $\mathbf{y} = \mathbf{X}\beta + \varepsilon$, where $\varepsilon \sim N(\mathbf{0}, \sigma^2\mathbf{I}_n)$ and $\text{rank}(\mathbf{X}) = K$. Consider the null hypothesis $H_0 : \mathbf{R}\beta = \mathbf{q}$, where $\mathbf{R}$ is a $J \times K$ matrix of constants of rank J and $\mathbf{q}$ is a $J \times 1$ vector.

We define:

1. **Discrepancy Vector:** $\mathbf{m} = \mathbf{R}\hat{\beta} - \mathbf{q}$.
2. **Residual Variance Estimator:** $s^2 = \frac{\mathbf{e}'\mathbf{e}}{n-K}$, where $\mathbf{e} = \mathbf{y} - \mathbf{X}\hat{\beta}$.
3. **Feasible Wald Statistic (W):**

$$W = \frac{\mathbf{m}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}\mathbf{m}}{s^2} \tag{2.42}$$

Theorem (Wald-F Relationship). Under H_0 and normality, the statistic $F = W/J$ follows a Snedecor F distribution with J and $n - K$ degrees of freedom:

$$\boxed{F = \frac{W}{J} \sim F_{J,n-K}} \tag{2.43}$$

2.7.3 Formal Derivation

Auxiliary Lemma 3 (Distribution of m): Under H_0 and normality, $\mathbf{m}|\mathbf{X} \sim N(\mathbf{0}, \sigma^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}')$.

Step 1: Construction of the χ^2 Quadratic Form for the Numerator

Consider the theoretical Wald statistic (W_σ) using the population variance σ^2:

$$W_\sigma = \mathbf{m}'[\text{Var}(\mathbf{m}|\mathbf{X})]^{-1}\mathbf{m} \tag{2.44}$$

Substituting the variance from Lemma 3:

$$W_\sigma = \frac{\mathbf{m}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}\mathbf{m}}{\sigma^2} \tag{2.45}$$

By the theory of quadratic forms of normal vectors, since $\mathbf{m}$ has dimension J:

$$W_\sigma \sim \chi^2_J \tag{2.46}$$

Step 2: Construction of the χ^2 Quadratic Form for the Denominator

We define the chi-square variable based on the sum of squared residuals (SSR):

$$V = \frac{(n-K)s^2}{\sigma^2} \sim \chi^2_{n-K} \tag{2.47}$$

Step 3: Stochastic Independence

As demonstrated in Section 1.6.3, $\hat{\beta}$ (and hence $\mathbf{m}$) is independent of the residual vector $\mathbf{e}$ (and hence s^2) under the normality assumption. Therefore, W_σ and V are independent.

Step 4: Definition of the F Ratio

By definition, an F variable is the ratio of two independent χ^2 variables, each divided by its respective degrees of freedom:

$$F = \frac{W_\sigma/J}{V/(n-K)} \tag{2.48}$$

Substituting (2.45) and (2.47) into (2.48):

$$F = \frac{\frac{\mathbf{m}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}\mathbf{m}}{\sigma^2 \cdot J}}{\frac{s^2}{\sigma^2}} \tag{2.49}$$

Canceling the unknown parameter σ^2:

$$F = \frac{\mathbf{m}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}\mathbf{m}}{s^2 \cdot J} \tag{2.50}$$

Recalling the definition of W in (2.42):

$$F = \frac{W}{J} \tag{2.51}$$

□

2.8 Confidence Intervals and Ellipsoidal Regions

2.8.1 Motivation

Point estimators, although unbiased, do not quantify the intrinsic uncertainty of the sampling process. The Confidence Interval solves the problem of providing a **set of plausible values** for the population parameter, based on the precision of the estimator.

Geometric and Spherical Interpretation:

Consider the parameter space $\mathbb{R}^K$.

1. **Univariate Case:** The CI is a line segment centered at the estimate.
2. **Multivariate Case (Spherical Interpretation):** Under the assumption that the parameter estimates are independent and have the same variance (spherical errors), the joint confidence region in $\mathbb{R}^2$ is a **circle** centered at $(\hat{\beta}_1, \hat{\beta}_2)$.
3. **Covariance and Slope:** When there is covariance between the estimators ($\text{Cov}(\hat{\beta}_1, \hat{\beta}_2) \neq 0$), the circle deforms into an **ellipse**. The slope of the principal axis of this ellipse is governed by the covariance structure: if the correlation is positive, the ellipse tilts to the right; if negative, to the left.

2.8.2 Formal Statement

Let $(X_1, \ldots, X_n)$ be a random sample from a population with parameter $\theta \in \Theta$. We define the Confidence Interval of level $1 - \alpha$ as the random set $\hat{C} = [\hat{L}, \hat{U}]$ such that:

$$P(\theta \in [\hat{L}, \hat{U}]) = 1 - \alpha \tag{2.52}$$

where $\hat{L}$ and $\hat{U}$ are statistics (functions of the data).

Theorem (Confidence Interval for β_j). Under the CNLRM assumptions:

$$\boxed{CI_{1-\alpha}(\beta_j) = \hat{\beta}_j \pm t_{n-K,\alpha/2} \cdot se(\hat{\beta}_j)} \tag{2.53}$$

Theorem (Confidence Interval for Linear Combination). For $\mathbf{c}'\beta$:

$$\boxed{CI_{1-\alpha}(\mathbf{c}'\beta) = \mathbf{c}'\hat{\beta} \pm t_{n-K,\alpha/2} \cdot \sqrt{s^2 \mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}}} \tag{2.54}$$

Theorem (Joint Confidence Region). The joint confidence region of level $1 - \alpha$ for β is:

$$\boxed{\hat{C}_{\text{joint}} = \left\{\beta : (\hat{\beta} - \beta)'(\mathbf{X}'\mathbf{X})(\hat{\beta} - \beta) \leq K \cdot s^2 \cdot F_{K,n-K,1-\alpha}\right\}} \tag{2.55}$$

2.8.3 Formal Derivation (Univariate Case)

Auxiliary Lemma 4 (Distribution of the Sample Mean):

If $X_i \sim N(\mu, \sigma^2)$, then $\bar{X} \sim N(\mu, \sigma^2/n)$.

Definition of the Pivot Variable:

We define the standardized variable Z:

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1) \tag{2.56}$$

Note that the distribution of Z does not depend on μ.

Step 1: Establishing the Central Probability

Given a significance level α, we select the quantile $z_{\alpha/2}$ such that:

$$P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha \tag{2.57}$$

Step 2: Substitution of the Pivot Variable

Substituting (2.56) into (2.57):

$$P\left(-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = 1 - \alpha \tag{2.58}$$

Step 3: Isolating the Parameter μ

Multiplying all terms by the standard error $\sigma/\sqrt{n}$:

$$P\left(-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \bar{X} - \mu \leq z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha \tag{2.59}$$

Subtracting $\bar{X}$ and rearranging:

$$P\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha \tag{2.60}$$

Conclusion: The interval $\hat{C} = \left[\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$ is the $1 - \alpha$ CI for μ.

When σ^2 is unknown, we substitute $z_{\alpha/2}$ with $t_{n-1,\alpha/2}$, resulting in (2.53).

2.8.4 Generalization to Ellipsoidal Regions

When dealing with a vector β of dimension K, the probabilistic distance is measured via the Mahalanobis quadratic form. Under the null hypothesis that β is the true value, the Wald statistic W is:

$$W(\beta) = (\hat{\beta} - \beta)'[\text{Var}(\hat{\beta})]^{-1}(\hat{\beta} - \beta) \sim \chi^2_K \tag{2.61}$$

The joint confidence region is the set:

$$\hat{C}_{\text{joint}} = \{\beta : W(\beta) \leq \chi^2_{K,1-\alpha}\} \tag{2.62}$$

When σ^2 is unknown, we use s^2 and the F distribution, resulting in (2.55).

2.9 Joint Significance Test and the Algebra of Sums of Squares

2.9.1 Motivation

The concrete problem solved by the joint significance test is the **collective validity of the regressors**. Frequently, individual variables may not be statistically significant in isolation (due to multicollinearity, for example), but together they explain a relevant portion of the variance of the phenomenon. The test seeks to answer: "Is the proposed model better than just using the sample mean to predict the dependent variable?"

Geometric Interpretation in $\mathbb{R}^n$:

In the sample space, the Total Sum of Squares (SST) represents the squared length of the vector of observations centered at their mean. OLS decomposes this vector into two orthogonal components: the Explained Sum of Squares (SSE), which is the projection of the vector onto the subspace spanned by the regressors, and the Residual Sum of Squares (SSR), which is the orthogonal distance of the vector to the subspace. The R^2 is the squared cosine of the angle between the observed vector and its projection, indicating the angular "proximity" between the data and the model.

2.9.2 Formal Statement

Consider the linear regression model $\mathbf{y} = \mathbf{X}\beta + \varepsilon$ with n observations and K parameters (including the intercept). We define:

1. **Total Sum of Squares (SST):** $SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$.
2. **Explained Sum of Squares (SSE):** $SSE = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$.
3. **Residual Sum of Squares (SSR):** $SSR = \sum_{i=1}^{n}\hat{e}_i^2$, where $\hat{e}_i = y_i - \hat{y}_i$.
4. **Coefficient of Determination (R^2):** $R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$.

Decomposition Theorem: If the model contains an intercept, then $SST = SSE + SSR$.

2.9.3 Derivation of the F Statistic via R^2

Step 1: Comparison between Restricted and Unrestricted Models

Let an unrestricted model (ur) with K parameters and a restricted model (r) with $K - J$ parameters, where J is the number of exclusion restrictions imposed (e.g., $\beta_j = 0$).

Proposition: $R^2_{ur} \geq R^2_r$ and $SSR_{ur} \leq SSR_r$.

Proof: The unrestricted OLS estimator minimizes the sum of squared residuals over the space $\mathbb{R}^K$. The restricted model minimizes the same function over a lower-dimensional subspace ($K - J$). Since the set of solutions of the restricted model is contained in the set of the unrestricted model, the minimum value attained by the unrestricted model will necessarily be less than or equal to that of the restricted model. □

Step 2: Construction of the F Statistic via SSR

To test the joint validity of J linear restrictions:

$$F = \frac{(SSR_r - SSR_{ur})/J}{SSR_{ur}/(n - K)} \tag{2.63}$$

Step 3: Construction of the F Statistic via R^2

From the definition of R^2: $SSR_{ur} = SST(1 - R^2_{ur})$ and $SSR_r = SST(1 - R^2_r)$.

Substituting into (2.63):

$$F = \frac{(R^2_{ur} - R^2_r)/J}{(1 - R^2_{ur})/(n-K)} \sim F_{J,n-K} \tag{2.64}$$

Step 4: Special Case of Global Significance

In the test where $H_0 : \beta_2 = \beta_3 = \cdots = \beta_K = 0$ (all slopes are zero), the restricted model is $y = \beta_1 + \varepsilon$. In this case, $R^2_r = 0$. Substituting into (2.64):

$$\boxed{F = \frac{R^2/(K-1)}{(1-R^2)/(n-K)} \sim F_{K-1,n-K}} \tag{2.65}$$

2.10 Summary of Test Statistics

Hypothesis	Statistic	Distribution	Degrees of Freedom
$H_0 : \beta_j = \beta_j^0$	$t = \frac{\hat{\beta}_j - \beta_j^0}{se(\hat{\beta}_j)}$	t_{n-K}	$n-K$
$H_0 : \mathbf{R}\beta = \mathbf{q}$ (with σ^2 known)	$W = \mathbf{m}'\mathbf{V}_m^{-1}\mathbf{m}$	χ^2_J	J
$H_0 : \mathbf{R}\beta = \mathbf{q}$ (with σ^2 unknown)	$F = \frac{W^*/J}{s^2}$	$F_{J,n-K}$	$J,\ n-K$
$H_0 : \beta_2 = \cdots = \beta_K = 0$	$F = \frac{R^2/(K-1)}{(1-R^2)/(n-K)}$	$F_{K-1,n-K}$	$K-1,\ n-K$

2.11 Role of Assumptions

1. **Normality of Errors (Assumption 2.1):** Indispensable to guarantee that test statistics follow exact distributions in finite samples. Without normality, t, F, and W have only asymptotic validity.
2. **Full Rank (Assumption 2.2):** Indispensable to guarantee the identifiability of parameters and the invertibility of $(\mathbf{X}'\mathbf{X})$, necessary for computing variances.
3. **Rank of Restrictions (Assumption 2.4):** Indispensable to guarantee that the matrix $\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'$ is invertible, allowing the calculation of the Wald and F statistics.
4. **Sphericity ($E[\varepsilon\varepsilon'] = \sigma^2\mathbf{I}$):** Simplifies the variance structure, allowing closed-form statistics. Under heteroskedasticity, formulas require robust adjustments.
5. **Homoskedasticity:** Indispensable for the relationship $F = W/J$. If the variance were heteroskedastic, the term $\text{Var}(\mathbf{m}|\mathbf{X})$ would involve the matrix $\mathbf{\Omega}$, preventing the isolation of the scalar s^2 and breaking the identity.
6. **Presence of Intercept:** Necessary for the decomposition $SST = SSE + SSR$. If the intercept is omitted, SST is not comparable, making the statistic based on R^2 mathematically invalid.

Chapter 3: Heteroskedasticity

3.1 The Heteroskedasticity Problem

3.1.1 Motivation and Geometric Interpretation

The Ordinary Least Squares (**OLS**) estimator is built upon the premise that errors possess constant variance — the so-called **homoskedasticity**. **Heteroskedasticity** arises when this variance is not uniform but varies systematically with the explanatory variables.

In empirical contexts, this situation is frequent. When modeling savings as a function of income, for example, higher-income families exhibit much greater dispersion in their saving habits than low-income families — the uncertainty about behavior increases with the level of the variable itself.

Geometric Interpretation in $\mathbb{R}^2$:

In a scatter plot (X, Y):

- Under **homoskedasticity**, the cloud of points exhibits uniform thickness around the regression line.
- Under **heteroskedasticity**, this cloud assumes a cone or fan shape: the vertical dispersion of residuals expands as we move along the X axis.

The OLS estimator fits the line through the center of this cloud but assigns equal weight to all observations — ignoring that points in regions of high dispersion are less precise informants about the location of the true line than those in regions of low dispersion.

3.1.2 Classical Assumptions and Their Roles

Assumption 3.1: Linearity in Parameters

- **Role:** Ensures that the population relationship is expressed as a linear combination of fixed parameters and an additive error, enabling the OLS estimator.
- **Counterexample:** If the model is intrinsically non-linear (e.g., $Y = e^{\beta X} \cdot u$), OLS fails to capture the curvature, generating systematically biased residuals.

Assumption 3.2: No Perfect Multicollinearity

- **Role:** Ensures that $\mathbf{X}$ has full rank, enabling the inversion of $(\mathbf{X}'\mathbf{X})$ and the unique identification of parameters.
- **Counterexample:** If $X_2 = 2X_1$, there are infinitely many lines that minimize the sum of squares, making β non-identifiable.

Assumption 3.3: Strict Exogeneity ($E[\mathbf{u}|\mathbf{X}] = \mathbf{0}$)

- **Role:** Ensures that the error does not carry systematic information correlated with the regressors — a fundamental condition for the unbiasedness of OLS.
- **Counterexample:** With endogeneity ($E[\mathbf{u}|\mathbf{X}] \neq \mathbf{0}$), the estimator absorbs into β effects belonging to the error, generating bias even in infinite samples.

Assumption 3.4: Heteroskedasticity ($E[\mathbf{u}\mathbf{u}'|\mathbf{X}] = \mathbf{\Omega}$)

- **Role:** Replaces the homoskedasticity assumption. The matrix $\mathbf{\Omega}$ is diagonal with elements $\text{diag}(\mathbf{\Omega}) = \{\sigma_1^2, \ldots, \sigma_n^2\}$ not necessarily equal.
- **Counterexample:** Erroneously assuming $E[\mathbf{u}\mathbf{u}'|\mathbf{X}] = \sigma^2\mathbf{I}$ produces incorrect standard error formulas, invalidating all significance tests (t and F).

3.1.3 Formal Statement of the Problem

Consider the linear population model in matrix notation:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{u} \tag{3.1}$$

where $\mathbf{y} \in \mathbb{R}^n$, $\mathbf{X} \in \mathbb{R}^{n \times k}$ with $\text{rank}(\mathbf{X}) = k < n$, $\beta \in \mathbb{R}^k$, and $\mathbf{u} \in \mathbb{R}^n$. Conditionally on $\mathbf{X}$, we assume:

1. $E[\mathbf{u}|\mathbf{X}] = \mathbf{0}$ (strict exogeneity);
2. $E[\mathbf{u}\mathbf{u}'|\mathbf{X}] = \mathbf{\Omega} = \text{diag}(\sigma_1^2, \ldots, \sigma_n^2)$ (heteroskedasticity).

The OLS estimator is given by:

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{3.2}$$

Objective of this section: To demonstrate that, under heteroskedasticity, $\hat{\beta}$ remains unbiased and consistent, but:

- the classical formula $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ becomes invalid for the variance;
- the estimator ceases to be the Best Linear Unbiased Estimator (**BLUE**).

3.1.4 Formal Derivation of Consequences

Step 1: Decomposition of the Estimator

Substituting (3.1) into (3.2) and using $(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X}) = \mathbf{I}_k$:

$$\hat{\beta} = \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u} \tag{3.3}$$

Step 2: Unbiasedness

Applying the conditional expectation in (3.3) and using Assumption 3.3:

$$E[\hat{\beta}|\mathbf{X}] = \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\, E[\mathbf{u}|\mathbf{X}] = \beta \tag{3.4}$$

Partial conclusion: OLS is unbiased under heteroskedasticity — the variance of the errors does not affect the mean of the estimator.

Step 3: True Variance — Sandwich Form

For an unbiased estimator, $\text{Var}[\hat{\beta}|\mathbf{X}] = E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'|\mathbf{X}]$. From (3.3):

$$\hat{\beta} - \beta = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u} \tag{3.5}$$

Computing the outer product and extracting the deterministic terms:

$$\text{Var}[\hat{\beta}|\mathbf{X}] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\, E[\mathbf{u}\mathbf{u}'|\mathbf{X}]\, \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

Substituting Assumption 3.4:

$$\boxed{\text{Var}[\hat{\beta}|\mathbf{X}] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Omega}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}} \tag{3.6}$$

This is the **sandwich form** of the true variance of the OLS estimator.

Step 4: Invalidity of Classical Inference

Classical inference assumes $E[\mathbf{u}\mathbf{u}'|\mathbf{X}] = \sigma^2\mathbf{I}$. Substituting into (3.6):

$$\text{Var}_{\text{classical}}[\hat{\beta}|\mathbf{X}] = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

If $\mathbf{\Omega} \neq \sigma^2\mathbf{I}$, this expression is **not** the correct variance estimator. Since the t and F statistics depend on the standard errors in the denominator, using the classical formula under heteroskedasticity produces biased tests, leading to incorrect conclusions about significance.

3.1.5 Synthesis of the Role of Assumptions

- **Strict exogeneity** ($E[\mathbf{u}|\mathbf{X}] = \mathbf{0}$) was indispensable for guaranteeing unbiasedness in (3.4).
- **Homoskedasticity** is the assumption that "fails" in this context: without it, OLS loses efficiency within the BLUE class (the Gauss-Markov Theorem no longer applies), making it necessary to use:
  - **robust standard errors** (EHW estimator) for valid inference; or
  - the **Generalized Least Squares (GLS) estimator** to recover efficiency.

3.2 The Eicker-Huber-White (EHW) Estimator

3.2.1 Motivation and Geometric Interpretation

Although $\hat{\beta}$ remains unbiased and consistent under heteroskedasticity, the classical variance formula $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ fails to capture the true uncertainty — typically underestimating standard errors and inflating t-statistics, leading to false positives in significance tests.

The **Eicker-Huber-White (EHW)** estimator solves this problem by allowing each observation to have its own variance, providing a robust basis for inference **without requiring knowledge of the functional form of heteroskedasticity**.

Geometric Interpretation in $\mathbb{R}^n$:

The EHW estimator can be interpreted as a non-parametric reconstruction of the **ellipsoid of uncertainty** around $\hat{\beta}$. Instead of assuming that this ellipsoid is a uniform hypersphere (homoskedasticity), the EHW estimator estimates its irregular shape using the squared OLS residuals as a local proxy for the variance of each observation.

3.2.2 Specific Assumptions of the EHW Estimator

Assumption 3.5: Contemporary Orthogonality ($E[\mathbf{x}_i u_i] = \mathbf{0}$)

- **Role:** Minimum exogeneity condition for OLS consistency. Ensures that the regressors do not carry systematic information about the error.
- **Counterexample:** With endogeneity ($E[\mathbf{x}_i u_i] \neq \mathbf{0}$), the estimator converges to a value distinct from β — no variance correction recovers the validity of inference.

Assumption 3.6: Independence between Observations ($E[u_i u_j|\mathbf{X}] = 0$ for $i \neq j$)

- **Role:** Ensures that $\mathbf{\Omega}$ is diagonal. The EHW estimator only needs to estimate the n individual variances σ_i^2.
- **Counterexample:** With autocorrelation or spatial dependence, the off-diagonal elements of $\mathbf{\Omega}$ are non-zero. The EHW estimator, by ignoring them, becomes inconsistent — requiring clustered standard errors (Arellano) or Newey-West.

Assumption 3.7: Finite Fourth-Order Moments ($E[u_i^4 \mathbf{x}_i \mathbf{x}_i'] < \infty$)

- **Role:** Technical condition for the Law of Large Numbers and the CLT to apply to the quadratic product $u_i^2 \mathbf{x}_i \mathbf{x}_i'$.
- **Counterexample:** With excessively heavy-tailed distributions (e.g., Cauchy), the variance of the variance estimator is not well-defined, preventing convergence of the robust estimator.

3.2.3 Formal Statement

Consider the linear model $\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$ with $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. Under $E[\mathbf{u}|\mathbf{X}] = \mathbf{0}$ and $E[\mathbf{u}\mathbf{u}'|\mathbf{X}] = \mathbf{\Omega} = \text{diag}(\sigma_1^2, \ldots, \sigma_n^2)$, we wish to construct an estimator $\widehat{\text{Var}}[\hat{\beta}]$ such that:

$$\text{plim}\Big(n\,\widehat{\text{Var}}[\hat{\beta}]\Big) = \text{Avar}\Big[\sqrt{n}(\hat{\beta} - \beta)\Big]$$

3.2.4 Formal Derivation of the EHW Estimator

Step 1: Decomposition of the Estimation Error

From the definition of OLS:

$$\hat{\beta} - \beta = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u} \tag{3.7}$$

Step 2: Asymptotic Scaling

Premultiplying (3.7) by $\sqrt{n}$:

$$\sqrt{n}(\hat{\beta} - \beta) = \left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1} \left(\frac{\mathbf{X}'\mathbf{u}}{\sqrt{n}}\right) \tag{3.8}$$

Step 3: Central Limit Theorem for the Score

Let $\mathbf{w}_i = \mathbf{x}_i u_i$ with $E[\mathbf{w}_i] = \mathbf{0}$ (Assumption 3.5). By the multivariate CLT (under Assumption 3.7):

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{x}_i u_i \xrightarrow{d} N(\mathbf{0}, \mathbf{\Sigma}) \tag{3.9}$$

where the variance matrix of the score is:

$$\mathbf{\Sigma} = \operatorname{plim}\left(\frac{1}{n} \sum_{i=1}^{n} u_i^2 \mathbf{x}_i \mathbf{x}_i'\right) = E[u_i^2 \mathbf{x}_i \mathbf{x}_i'] \tag{3.10}$$

Step 4: Asymptotic Variance — Sandwich Form

Let $\mathbf{Q} = \operatorname{plim}(\mathbf{X}'\mathbf{X}/n) = E[\mathbf{x}_i \mathbf{x}_i']$. By Slutsky's Lemma applied to (3.8):

$$\boxed{\operatorname{Avar}\left[\sqrt{n}(\hat{\beta} - \beta)\right] = \mathbf{Q}^{-1}\mathbf{\Sigma}\mathbf{Q}^{-1}} \tag{3.11}$$

$\mathbf{\Sigma}$ is the "filling" and $\mathbf{Q}^{-1}$ are the "bread slices" of the asymptotic sandwich structure.

Step 5: Estimation of $\mathbf{\Sigma}$ by White's Estimator

Since u_i is unobservable, White (1980) proposed replacing it with the OLS residual $\hat{u}_i = y_i - \mathbf{x}_i'\hat{\beta}$:

$$\hat{\mathbf{\Sigma}} = \frac{1}{n} \sum_{i=1}^{n} \hat{u}_i^2 \mathbf{x}_i \mathbf{x}_i' \tag{3.12}$$

The consistency of $\hat{\mathbf{\Sigma}}$ follows from:

- $\hat{\beta} \xrightarrow{p} \beta$;
- expansion $\hat{u}_i = u_i - \mathbf{x}_i'(\hat{\beta} - \beta)$;
- the second- and third-order terms involve $(\hat{\beta} - \beta) \xrightarrow{p} \mathbf{0}$, and the adjacent sums are $O_p(1)$ by the finiteness of fourth-order moments (Assumption 3.7).

Therefore:

$$\operatorname{plim}(\hat{\mathbf{\Sigma}}) = \mathbf{\Sigma} \tag{3.13}$$

Step 6: Final Robust Estimator

Substituting the sample analogs into (3.11) and undoing the scaling by n:

$$\boxed{\widehat{\operatorname{Var}}_{\text{EHW}}[\hat{\beta}] = (\mathbf{X}'\mathbf{X})^{-1} \left(\sum_{i=1}^{n} \hat{u}_i^2 \mathbf{x}_i \mathbf{x}_i'\right) (\mathbf{X}'\mathbf{X})^{-1}} \tag{3.14}$$

This is the expression for the robust variance-covariance estimator **HC0** of White (1980).

3.2.5 Role of Assumptions in the EHW Estimator

Independence between observations (Assumption 3.6) was indispensable for ensuring that $\mathbf{\Sigma}$ depended only on the contemporaneous products $u_i^2 \mathbf{x}_i \mathbf{x}_i'$. With autocorrelation or spatial dependence, the covariance terms $E[u_i u_j \mathbf{x}_i \mathbf{x}_j']$ $(i \neq j)$ would be non-zero, and the estimator (3.12) would be inconsistent for the true variance — requiring alternatives such as clustered standard errors (Arellano) or Newey-West.

3.3 Asymptotic Properties of the EHW Estimator

3.3.1 Motivation

In large samples, it is not necessary for the errors to be normally distributed to perform valid inference. The asymptotic challenge that the EHW estimator solves is the construction of correct hypothesis tests when the error variance is unknown and potentially heteroskedastic.

Geometric Interpretation:

The sampling distribution of $\hat{\beta}$ is a cloud of points that contracts as n grows (consistency). When scaled by $\sqrt{n}$, this cloud converges to a Normal distribution. The EHW estimator allows us to accurately calculate the "width" (variance) of this Normal even when the dispersion of the original data is irregular — ensuring that confidence intervals and t-tests do not distort the true precision of the estimate.

3.3.2 Asymptotic Assumptions

Assumption 3.8: i.i.d. Sampling of $(\mathbf{x}_i, y_i)$

- **Role:** Enables the direct application of the Weak Law of Large Numbers (WLLN) and the CLT in their classical forms.
- **Counterexample:** With temporal dependence (autocorrelation), the sums do not converge to the simple expectations, requiring Newey-West corrections or time series approaches.

Assumption 3.9: Finite Fourth-Order Moments ($E[\|\mathbf{x}_i\|^4] < \infty$ and $E[u_i^4] < \infty$)

- **Role:** Ensures that the variance of $u_i^2 \mathbf{x}_i \mathbf{x}_i'$ is finite, allowing the sandwich filling to converge in probability to the population value.
- **Counterexample:** With heavy-tailed distributions (e.g., Cauchy), the mean squared error does not converge, making the robust standard error unstable.

Assumption 3.10: Non-Singular Moment Matrix ($\mathbf{Q} = E[\mathbf{x}_i \mathbf{x}_i']$ positive definite)

- **Role:** Ensures that the regressors are not asymptotically collinear, guaranteeing that $\mathbf{Q}^{-1}$ exists and that β remains identified in the limit.
- **Counterexample:** If a regressor becomes constant or asymptotically collinear, the variance of the estimator explodes to infinity.

3.3.3 Formal Statement

Under i.i.d. sampling, finite fourth-order moments, and contemporaneous exogeneity ($E[\mathbf{x}_i u_i] = \mathbf{0}$), the OLS estimator $\hat{\beta}_n = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ and the robust estimator $\widehat{\mathrm{Var}}_{\mathrm{EHW}} = (\mathbf{X}'\mathbf{X})^{-1}\big(\sum_i \hat{u}_i^2 \mathbf{x}_i \mathbf{x}_i'\big)(\mathbf{X}'\mathbf{X})^{-1}$ satisfy:

1. **Consistency:** $\hat{\beta}_n \xrightarrow{p} \beta$
2. **Asymptotic Normality:** $\sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{d} N(\mathbf{0}, \mathbf{V}_\beta)$, where $\mathbf{V}_\beta = \mathbf{Q}^{-1}\mathbf{\Sigma}\mathbf{Q}^{-1}$
3. **Consistency of the Robust Estimator:** $\mathrm{plim}(n\,\widehat{\mathrm{Var}}_{\mathrm{EHW}}) = \mathbf{V}_\beta$

3.3.4 Formal Derivation

Step 1: Consistency of $\hat{\beta}_n$

We write the estimation deviation in scaled form:

$$\hat{\beta}_n - \beta = \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i u_i\right) \tag{3.15}$$

By **Khinchin's WLLN** (under Assumption 3.8):

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i' \xrightarrow{p} \mathbf{Q}, \qquad \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i u_i \xrightarrow{p} E[\mathbf{x}_i u_i] = \mathbf{0}$$

By the **Continuous Mapping Theorem** applied to (3.15):

$$\operatorname{plim}(\hat{\beta}_n - \beta) = \mathbf{Q}^{-1}\cdot\mathbf{0} = \mathbf{0} \quad \Longrightarrow \quad \hat{\beta}_n \xrightarrow{p} \beta$$

Step 2: Asymptotic Distribution

Premultiplying (3.15) by $\sqrt{n}$:

$$\sqrt{n}(\hat{\beta}_n - \beta) = \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{x}_i u_i\right) \tag{3.16}$$

By the **Multivariate CLT** (under Assumptions 3.8 and 3.9):

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{x}_i u_i \xrightarrow{d} N(\mathbf{0}, \mathbf{\Sigma}), \qquad \mathbf{\Sigma} = E[u_i^2\mathbf{x}_i\mathbf{x}_i']$$

By **Slutsky's Lemma**, combining $\hat{\mathbf{Q}}_n \xrightarrow{p} \mathbf{Q}$ and the convergence in distribution of the score:

$$\boxed{\sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{d} N(\mathbf{0},\, \mathbf{Q}^{-1}\mathbf{\Sigma}\mathbf{Q}^{-1})} \tag{3.17}$$

Step 3: Consistency of the Filling ($\hat{\mathbf{\Sigma}} \xrightarrow{p} \mathbf{\Sigma}$)

White's estimator for the filling is $\hat{\mathbf{\Sigma}} = \frac{1}{n}\sum_i \hat{u}_i^2\mathbf{x}_i\mathbf{x}_i'$. Expanding $\hat{u}_i = u_i - \mathbf{x}_i'(\hat{\beta}_n - \beta)$:

$$\hat{u}_i^2 = u_i^2 - 2u_i\mathbf{x}_i'(\hat{\beta}_n - \beta) + (\hat{\beta}_n - \beta)'\mathbf{x}_i\mathbf{x}_i'(\hat{\beta}_n - \beta)$$

Upon substituting into the sample sum:

- The first term converges by the WLLN: $\frac{1}{n}\sum_i u_i^2\mathbf{x}_i\mathbf{x}_i' \xrightarrow{p} \mathbf{\Sigma}$.
- The second- and third-order terms involve $(\hat{\beta}_n - \beta) \xrightarrow{p} \mathbf{0}$, and the adjacent sums are $O_p(1)$ by the finiteness of fourth-order moments (Assumption 3.9).

Therefore:

$$\operatorname{plim}(\hat{\mathbf{\Sigma}}) = \mathbf{\Sigma} \tag{3.18}$$

Step 4: Consistency of the Robust Variance

Rewriting $n\,\widehat{\mathrm{Var}}_{\mathrm{EHW}}$ in scaled form:

$$n\,\widehat{\mathrm{Var}}_{\mathrm{EHW}} = \left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1} \hat{\mathbf{\Sigma}} \left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1}$$

By the Continuous Mapping Theorem, combining $\hat{\mathbf{Q}}_n \xrightarrow{p} \mathbf{Q}$ and $\hat{\mathbf{\Sigma}} \xrightarrow{p} \mathbf{\Sigma}$:

$$\boxed{\mathrm{plim}(n\,\widehat{\mathrm{Var}}_{\mathrm{EHW}}) = \mathbf{Q}^{-1}\mathbf{\Sigma}\mathbf{Q}^{-1} = \mathbf{V}_\beta} \tag{3.19}$$

3.3.5 Role of Assumptions in Asymptotic Properties

The **finiteness of fourth-order moments** (Assumption 3.9) was indispensable in Step 3. Without it, the WLLN cannot be applied to the product $u_i^2\mathbf{x}_i\mathbf{x}_i'$, preventing the convergence of $\hat{\mathbf{\Sigma}}$ to the population value and invalidating all asymptotic inference.

i.i.d. sampling (Assumption 3.8) ensures that sample means converge to population expectations. With dependence, the covariance structure of the score becomes more complex, requiring robust methods for time series or panel data.

3.4 Hypothesis Testing with the EHW Estimator

3.4.1 Motivation and Geometric Interpretation

The Wald test addresses the problem of simultaneously testing multiple linear hypotheses about β (e.g., $\beta_1 = 0$ and $\beta_2 = 1$). Instead of evaluating each parameter in isolation, it measures the **statistical distance** between the unrestricted estimator $\hat{\beta}$ and the constraint space imposed by the null hypothesis.

Geometric Interpretation in $\mathbb{R}^k$:

$\hat{\beta}$ is a point in the parameter space; the null hypothesis defines a linear subspace (hyperplane) of dimension $k - q$. If H_0 is true, $\hat{\beta}$ should be "close" to this hyperplane. The distance, however, is not Euclidean — it is weighted by the uncertainty of the estimator. Under heteroskedasticity, the confidence ellipsoid around $\hat{\beta}$ has irregular axes; the EHW estimator correctly estimates the curvature of this ellipsoid, ensuring that the Wald distance reflects the true precision of the data.

3.4.2 Assumptions of the Test

Assumption 3.11: Independent Linear Restrictions ($H_0 : \mathbf{R}\beta = \mathbf{r}$)

- **Role:** Defines q simultaneous linear conditions with $\mathbf{R} \in \mathbb{R}^{q\times k}$ of full row rank ($\mathrm{rank}(\mathbf{R}) = q$).
- **Counterexample:** If $\mathrm{rank}(\mathbf{R}) < q$, the restrictions are redundant or contradictory, making $(\mathbf{R}\mathbf{V}_\beta\mathbf{R}')$ singular and the statistic undefined.

Assumption 3.12: Consistency of the EHW Estimator

- **Role:** Ensures that the weighting matrix of the Wald quadratic form converges to the true asymptotic variance $\mathbf{V}_\beta$.
- **Counterexample:** With autocorrelation or violation of the EHW assumptions, the estimator converges to an incorrect matrix, causing the statistic W not to follow the χ^2_q distribution, invalidating the test.

Assumption 3.13: Asymptotic Normality of $\sqrt{n}(\hat{\beta} - \beta)$

- **Role:** Allows the quadratic form of the Wald statistic to converge to a χ^2 distribution in large samples.
- **Counterexample:** Without finite fourth-order moments, the CLT fails and the Wald statistic may exhibit heavy-tailed distributions, inflating false rejection rates.

3.4.3 Formal Statement

Under heteroskedasticity $E[\mathbf{u}\mathbf{u}'|\mathbf{X}] = \mathbf{\Omega} = \text{diag}(\sigma_i^2)$, we wish to test:

$$H_0 : \mathbf{R}\beta - \mathbf{r} = \mathbf{0} \tag{3.20}$$

where $\mathbf{R} \in \mathbb{R}^{q \times k}$ with $\text{rank}(\mathbf{R}) = q$ and $\mathbf{r} \in \mathbb{R}^q$. The **robust Wald statistic** is:

$$W = (\mathbf{R}\hat{\beta} - \mathbf{r})' \left[\mathbf{R} \, \widehat{\text{Var}}_{\text{EHW}}[\hat{\beta}] \, \mathbf{R}' \right]^{-1} (\mathbf{R}\hat{\beta} - \mathbf{r}) \tag{3.21}$$

Objective: To prove that, under H_0, $W \xrightarrow{d} \chi^2_q$ as $n \to \infty$.

3.4.4 Formal Derivation

Auxiliary Lemma (Quadratic Forms of Gaussian Vectors): If $\mathbf{z} \sim N(\mathbf{0}, \mathbf{\Sigma})$ with $\mathbf{\Sigma}$ non-singular, then $\mathbf{z}'\mathbf{\Sigma}^{-1}\mathbf{z} \sim \chi^2_{\dim(\mathbf{z})}$.

Step 1: Asymptotic Behavior of the Discrepancy Vector

Under H_0, substituting $\mathbf{r} = \mathbf{R}\beta$:

$$\mathbf{R}\hat{\beta} - \mathbf{r} = \mathbf{R}(\hat{\beta} - \beta) \tag{3.22}$$

Premultiplying by $\sqrt{n}$:

$$\sqrt{n}(\mathbf{R}\hat{\beta} - \mathbf{r}) = \mathbf{R} \left[\sqrt{n}(\hat{\beta} - \beta) \right] \tag{3.23}$$

Step 2: Limiting Distribution of the Discrepancy Vector

By the asymptotic normality of $\sqrt{n}(\hat{\beta} - \beta)$ proved in (3.17), and by the property of linear transformations of Gaussian vectors:

$$\sqrt{n}(\mathbf{R}\hat{\beta} - \mathbf{r}) \xrightarrow{d} N(\mathbf{0}, \, \mathbf{R}\mathbf{V}_\beta\mathbf{R}') \tag{3.24}$$

Step 3: Consistency of the Weighting Matrix

Defining $\hat{\mathbf{V}}_n = n \, \widehat{\text{Var}}_{\text{EHW}}[\hat{\beta}]$, by the consistency proved in (3.19) and the Continuous Mapping Theorem:

$$\text{plim}(\mathbf{R}\hat{\mathbf{V}}_n\mathbf{R}') = \mathbf{R}\mathbf{V}_\beta\mathbf{R}' \tag{3.25}$$

Step 4: Asymptotic Formulation of the Wald Statistic

Rewriting (3.21) with the scaling factor n balanced:

$$W = \left[\sqrt{n}(\mathbf{R}\hat{\beta} - \mathbf{r}) \right]' \left[\mathbf{R}\hat{\mathbf{V}}_n\mathbf{R}' \right]^{-1} \left[\sqrt{n}(\mathbf{R}\hat{\beta} - \mathbf{r}) \right] \tag{3.26}$$

By **Slutsky's Lemma**, combining the convergence in distribution of (3.24) and the convergence in probability of (3.25):

$$W \xrightarrow{d} \mathbf{z}' \left[\mathbf{R}\mathbf{V}_\beta\mathbf{R}' \right]^{-1} \mathbf{z}, \qquad \mathbf{z} \sim N(\mathbf{0}, \, \mathbf{R}\mathbf{V}_\beta\mathbf{R}') \tag{3.27}$$

Step 5: Convergence to χ^2_q

Applying the Auxiliary Lemma directly to (3.27), since $\mathbf{R}\mathbf{V}_\beta\mathbf{R}'$ has full rank q:

$$\boxed{W \xrightarrow{d} \chi^2_q} \tag{3.28}$$

3.4.5 Role of Assumptions

The **full rank of R** ($\text{rank}(\mathbf{R}) = q$) is indispensable. Without it, the matrix $\mathbf{R}\mathbf{V}_\beta\mathbf{R}'$ would be singular, its inversion in (3.26) would be undefined, and the Wald statistic would not be mathematically well-posed.

3.5 The Inference Dilemma: χ^2 versus F Distribution

The choice between critical values from the χ^2 distribution or the F distribution reflects the balance between pure asymptotic theory and finite-sample performance.

Algebraic Relationship between W and F:

The two statistics maintain the direct identity:

$$F = \frac{W}{q}$$

where q is the number of restrictions. In the classical framework (homoskedasticity and exact normality), W/q follows strictly $F_{q,n-k}$. Under heteroskedasticity, the exact distribution of W/q in finite samples is unknown, but its asymptotic limit converges to χ^2_q/q.

Why does asymptotic theory point to χ^2?

There is no theorem guaranteeing that the Wald statistic computed with robust standard errors follows an F distribution in small samples with heteroskedasticity. As $n - k \to \infty$, the $F_{q,n-k}$ distribution multiplied by q converges exactly to χ^2_q — in large samples, the two approaches produce numerically identical conclusions.

Why does practice use the F distribution?

The F distribution has heavier tails than the χ^2. For the same significance level α, its critical values are larger, generating more conservative p-values. This offers practical protection against false positives when n is not sufficiently large to guarantee complete asymptotic convergence of White's estimator. In other words: if the choice between F and χ^2 materially alters the rejection decision, this is a sign that the asymptotic approximation of the EHW estimator is fragile in that sample — making the more cautious stance provided by the F distribution the prudent recommendation.

3.6 Generalized Least Squares (GLS) Estimator

3.6.1 Motivation and Geometric Interpretation

The Ordinary Least Squares (**OLS**) estimator is built under the assumption of spherical errors (constant variance and no serial correlation). In real data, however, observations often exhibit heterogeneous variances (heteroskedasticity) or stochastic dependence (autocorrelation). In these scenarios, OLS loses its efficiency property within the class of linear unbiased estimators, although it remains unbiased and consistent.

The Generalized Least Squares (**GLS**) estimator resolves this inefficiency. Instead of assigning the same mathematical weight to all observations, GLS incorporates the structure of the population covariance matrix of the errors to transform the original system, applying weights inversely proportional to the variability and codependence of each observation.

Geometric Interpretation in $\mathbb{R}^n$:

Geometrically, OLS orthogonally projects the data vector $\mathbf{y}$ onto the column space of $\mathbf{X}$, denoted $\mathcal{C}(\mathbf{X})$, under the standard Euclidean metric. Under non-spherical errors, the uncertainty region around $\mathbf{y}$ ceases to be a hypersphere and assumes the shape of an **ellipsoid** in $\mathbb{R}^n$.

GLS acts by applying a linear transformation through a weighting matrix that rotates and rescales the axes of the data space. This operation deforms the error ellipsoid, converting it into a unit hypersphere compatible with the classical Gauss-Markov assumptions. The standard orthogonal projection is then performed in this transformed space.

3.6.2 Assumptions of GLS

Assumption 3.14: Linearity and Full Rank

- **Role:** Defines the population structural relationship $\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$ and ensures that the transformed sample moment matrix is invertible, guaranteeing the unique identification and computability of β.
- **Counterexample:** If $\text{rank}(\mathbf{X}) < k$, there is perfect multicollinearity. The column space $\mathcal{C}(\mathbf{X})$ exhibits dimensional redundancy, generating infinitely many solutions to the minimization problem and rendering the estimator undefined.

Assumption 3.15: Strict Exogeneity ($E[\mathbf{u}|\mathbf{X}] = \mathbf{0}$)

- **Role:** Ensures that the conditional error has zero mean for any realization of $\mathbf{X}$, being the fundamental pillar for unbiasedness in finite samples.
- **Counterexample:** If $E[\mathbf{u}|\mathbf{X}] \neq \mathbf{0}$, the estimator will absorb into the vector β systematic effects belonging to the noise, generating bias in finite samples and asymptotic inconsistency.

Assumption 3.16: Generalized Covariance Matrix ($E[\mathbf{u}\mathbf{u}'|\mathbf{X}] = \sigma^2\mathbf{\Omega}$)

- **Role:** Determines the stochastic structure of the errors, where $\mathbf{\Omega}$ is an $n \times n$ symmetric positive definite (SPD) matrix.
- **Counterexample:** If $\mathbf{\Omega}$ is not positive definite, it will be singular (having at least one zero eigenvalue), implying the existence of a linear combination of errors with zero variance — a deterministic error. This precludes the existence of $\mathbf{\Omega}^{-1}$, necessary for GLS computation.

3.6.3 Formal Statement and Derivation

3.6.3.1 Formal Statement Consider the linear population model in matrix notation:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{u} \tag{3.29}$$

where $\mathbf{y} \in \mathbb{R}^n$, $\mathbf{X} \in \mathbb{R}^{n\times k}$ with $\text{rank}(\mathbf{X}) = k$, $\beta \in \mathbb{R}^k$, and $\mathbf{u} \in \mathbb{R}^n$. Conditionally on $\mathbf{X}$, we assume:

1. $E[\mathbf{u}|\mathbf{X}] = \mathbf{0}$
2. $E[\mathbf{u}\mathbf{u}'|\mathbf{X}] = \sigma^2\mathbf{\Omega}$, with $\mathbf{\Omega}$ symmetric and positive definite.

The Generalized Least Squares estimator is defined by:

$$\boxed{\hat{\beta}_{GLS} = (\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{y}} \tag{3.30}$$

Objective: To demonstrate, via linear transformation of the model, that $\hat{\beta}_{GLS}$ is unbiased and to derive its exact finite-sample variance-covariance matrix.

3.6.3.2 Formal Derivation Step 1: Factorization of the Covariance Matrix

Since $\mathbf{\Omega}$ is symmetric and positive definite, its inverse $\mathbf{\Omega}^{-1}$ shares the same properties. By the Cholesky decomposition, there exists a non-singular square matrix $\mathbf{P}$ of dimension $n \times n$ such that:

$$\mathbf{\Omega}^{-1} = \mathbf{P}'\mathbf{P} \tag{3.31}$$

From this factorization, the sphericity property follows:

$$\mathbf{P\Omega P'} = \mathbf{P}(\mathbf{P'P})^{-1}\mathbf{P'} = \mathbf{P}\mathbf{P}^{-1}(\mathbf{P'})^{-1}\mathbf{P'} = \mathbf{I}_n \tag{3.32}$$

Step 2: Linear Transformation of the Model

Premultiplying the structural equation (3.29) by $\mathbf{P}$:

$$\mathbf{Py} = \mathbf{PX}\beta + \mathbf{Pu} \tag{3.33}$$

Defining the transformed objects:

$$\mathbf{y}^* = \mathbf{Py}, \qquad \mathbf{X}^* = \mathbf{PX}, \qquad \mathbf{u}^* = \mathbf{Pu} \tag{3.34}$$

we obtain the transformed model:

$$\mathbf{y}^* = \mathbf{X}^*\beta + \mathbf{u}^* \tag{3.35}$$

Step 3: Verification of Gauss-Markov Conditions in the Transformed Model

Conditional mean:

$$E[\mathbf{u}^*|\mathbf{X}] = \mathbf{P}\,E[\mathbf{u}|\mathbf{X}] = \mathbf{0} \tag{3.36}$$

Conditional variance:

$$\mathrm{Var}[\mathbf{u}^*|\mathbf{X}] = \mathbf{P}\,E[\mathbf{uu'}|\mathbf{X}]\,\mathbf{P'} = \mathbf{P}(\sigma^2\mathbf{\Omega})\mathbf{P'} = \sigma^2(\mathbf{P\Omega P'})$$

Applying identity (3.32):

$$\mathrm{Var}[\mathbf{u}^*|\mathbf{X}] = \sigma^2\mathbf{I}_n \tag{3.37}$$

Partial conclusion: The transformed model (3.35) satisfies the Gauss-Markov conditions of homoskedasticity and no serial correlation.

Step 4: Application of OLS to the Transformed Model

Once the classical assumptions hold for (3.35), OLS applied to this model is the Best Linear Unbiased Estimator (BLUE):

$$\hat{\beta} = (\mathbf{X}^{*'}\mathbf{X}^*)^{-1}\mathbf{X}^{*'}\mathbf{y}^* \tag{3.38}$$

Step 5: Return to the Original Variables

Substituting the definitions (3.34) into (3.38):

$$\hat{\beta}_{GLS} = [(\mathbf{PX})'(\mathbf{PX})]^{-1}(\mathbf{PX})'(\mathbf{Py}) = [\mathbf{X'P'PX}]^{-1}\mathbf{X'P'Py}$$

Applying the factorization $\mathbf{P'P} = \mathbf{\Omega}^{-1}$ from (3.31):

$$\hat{\beta}_{GLS} = (\mathbf{X'\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X'\Omega}^{-1}\mathbf{y}$$

which coincides with the statement (3.30). □

Step 6: Finite-Sample Variance

Since $\hat{\beta}_{GLS}$ operates as OLS on the transformed model, its conditional variance is given directly by the classical formula applied to spherical variables:

$$\text{Var}[\hat{\beta}_{GLS}|\mathbf{X}] = \sigma^2(\mathbf{X}^{*'}\mathbf{X}^*)^{-1}$$

Substituting $\mathbf{X}^* = \mathbf{PX}$ and applying factorization (3.31):

$$\boxed{\text{Var}[\hat{\beta}_{GLS}|\mathbf{X}] = \sigma^2[\mathbf{X'P'PX}]^{-1} = \sigma^2(\mathbf{X'\Omega^{-1}X})^{-1}} \tag{3.39}$$

3.6.4 Relative Efficiency: The Generalized Gauss-Markov Theorem

3.6.4.1 Statement of the Theorem We will formally prove that $\hat{\beta}_{GLS}$ is strictly more efficient than OLS ($\hat{\beta}_{OLS}$) under non-spherical errors. This is equivalent to proving that the difference between their variance-covariance matrices is positive semidefinite (PSD):

$$\mathbf{\Delta} = \text{Var}(\hat{\beta}_{OLS}|\mathbf{X}) - \text{Var}(\hat{\beta}_{GLS}|\mathbf{X}) \succeq \mathbf{0} \tag{3.40}$$

3.6.4.2 Formal Derivation Notation and Preliminary Lemmas:

1. **Lemma 1 (Spectral Decomposition):** Since $\mathbf{\Omega}$ is SPD, there exists a unique SPD matrix $\mathbf{\Omega}^{1/2}$ such that $\mathbf{\Omega}^{1/2}\mathbf{\Omega}^{1/2} = \mathbf{\Omega}$ and $\mathbf{\Omega}^{-1/2}\mathbf{\Omega}^{-1/2} = \mathbf{\Omega}^{-1}$.
2. **Lemma 2 (Inversion of Inequalities):** For two SPD matrices A and B, $A \succeq B \iff B^{-1} \succeq A^{-1}$.
3. **Lemma 3 (Projection Matrix):** The matrix $P_Z = Z(Z'Z)^{-1}Z'$ is symmetric and idempotent ($P_Z^2 = P_Z$), and its complement $M_Z = I - P_Z$ is also symmetric and idempotent, hence PSD.

Step 1: Expression of the Variances

The variance of the OLS estimator under non-spherical errors is:

$$\text{Var}(\hat{\beta}_{OLS}|\mathbf{X}) = \sigma^2(\mathbf{X'X})^{-1}\mathbf{X'\Omega X}(\mathbf{X'X})^{-1} \tag{3.41}$$

The variance of the GLS estimator is:

$$\text{Var}(\hat{\beta}_{GLS}|\mathbf{X}) = \sigma^2(\mathbf{X'\Omega^{-1}X})^{-1} \tag{3.42}$$

Step 2: Comparison of Inverses

By **Lemma 2**, proving $\text{Var}(\hat{\beta}_{OLS}) \succeq \text{Var}(\hat{\beta}_{GLS})$ is equivalent to proving:

$$[\text{Var}(\hat{\beta}_{GLS})]^{-1} \succeq [\text{Var}(\hat{\beta}_{OLS})]^{-1} \tag{3.43}$$

Substituting expressions (3.41) and (3.42) and simplifying σ^2:

$$\mathbf{X'\Omega^{-1}X} \succeq (\mathbf{X'X})(\mathbf{X'\Omega X})^{-1}(\mathbf{X'X}) \tag{3.44}$$

Step 3: Identification of the Projection Matrix

Define the auxiliary matrices $A = \mathbf{\Omega}^{-1/2}\mathbf{X}$ and $B = \mathbf{\Omega}^{1/2}\mathbf{X}$.

We rewrite the terms of inequality (3.44):

$$\mathbf{X'\Omega^{-1}X} = \mathbf{X'\Omega^{-1/2}\Omega^{-1/2}X} = A'A \tag{3.45}$$

$$\mathbf{X'X} = \mathbf{X'\Omega^{-1/2}\Omega^{1/2}X} = A'B \tag{3.46}$$

$$\mathbf{X'\Omega X} = \mathbf{X'\Omega^{1/2}\Omega^{1/2}X} = B'B \tag{3.47}$$

Substituting (3.45), (3.46), and (3.47) into inequality (3.44), the difference between the terms becomes:

$$\mathbf{\Delta} = A'A - (A'B)(B'B)^{-1}(B'A) \tag{3.48}$$

Factoring A' on the left and A on the right:

$$\mathbf{\Delta} = A'[I - B(B'B)^{-1}B']A \tag{3.49}$$

Step 4: Proof of Positive Semidefiniteness

The term in brackets in (3.49) is exactly the annihilator matrix $M_B = I - P_B$, where $P_B = B(B'B)^{-1}B'$ is the projection matrix onto the column space of B:

$$\mathbf{\Delta} = A'M_BA \tag{3.50}$$

Verification of properties of M_B:

1. **Symmetry:** $M_B' = (I - P_B)' = I' - P_B' = I - P_B = M_B$.
2. **Idempotence:** $M_BM_B = M_B$.
3. **Positive Semidefiniteness:** For any vector $v \in \mathbb{R}^k$, define $z = Av$. Then:

$$v'\mathbf{\Delta}v = v'A'M_BAv = z'M_Bz \tag{3.51}$$

Since M_B is idempotent and symmetric:

$$z'M_Bz = z'M_B'M_Bz = \|M_Bz\|^2 \geq 0 \tag{3.52}$$

Conclusion:

Since $v'\mathbf{\Delta}v \geq 0$ for any vector v, the matrix $\mathbf{\Delta}$ is PSD. This proves that the inverse of the GLS variance dominates the inverse of the OLS variance, confirming the superior efficiency of GLS. The **SPD assumption of $\mathbf{\Omega}$** was indispensable for guaranteeing the existence of $\mathbf{\Omega}^{1/2}$ and the validity of the orthogonal projection in the transformed space.

3.6.5 Asymptotic Properties of GLS

3.6.5.1 Asymptotic Statement In finite samples, GLS requires complete knowledge of the parametric structure of $\mathbf{\Omega}$. Under usual regularity conditions, as $n \to \infty$ the GLS estimator satisfies:

1. **Consistency:** $\text{plim}(\hat{\beta}_{GLS}) = \beta$
2. **Asymptotic Normality:** $\sqrt{n}(\hat{\beta}_{GLS} - \beta) \xrightarrow{d} N\big(\mathbf{0},\, \sigma^2\mathbf{Q}_{GLS}^{-1}\big)$

3.6.5.2 Formal Derivation Step 1: Formulation of the Estimation Deviation

Substituting $\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$ into (3.30) and canceling the term $(\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X}) = \mathbf{I}_k$:

$$\hat{\beta}_{GLS} - \beta = (\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{u} \tag{3.53}$$

Step 2: Consistency

Normalizing (3.53) by n:

$$\hat{\beta}_{GLS} - \beta = \left(\frac{\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X}}{n}\right)^{-1}\left(\frac{\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{u}}{n}\right) \tag{3.54}$$

By the Weak Law of Large Numbers, under regularity conditions:

$$\operatorname{plim}\left(\frac{\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X}}{n}\right) = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} E[\mathbf{x}_i^*\mathbf{x}_i^{*\prime}] = \mathbf{Q}_{GLS} \tag{3.55}$$

where $\mathbf{Q}_{GLS}$ is finite and positive definite. For the average score, by strict exogeneity:

$$\operatorname{plim}\left(\frac{\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{u}}{n}\right) = E[\mathbf{x}_i^* u_i^*] = \mathbf{0} \tag{3.56}$$

By the Continuous Mapping Theorem applied to (3.54):

$$\operatorname{plim}(\hat{\beta}_{GLS} - \beta) = \mathbf{Q}_{GLS}^{-1}\cdot\mathbf{0} = \mathbf{0} \quad\Longrightarrow\quad \hat{\beta}_{GLS} \xrightarrow{p} \beta \tag{3.57}$$

Step 3: Asymptotic Distribution

Premultiplying (3.53) by $\sqrt{n}$:

$$\sqrt{n}(\hat{\beta}_{GLS} - \beta) = \left(\frac{\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X}}{n}\right)^{-1}\left(\frac{\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{u}}{\sqrt{n}}\right) \tag{3.58}$$

For the rescaled score vector $\mathbf{w}_n = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{x}_i^* u_i^*$, its exact conditional variance is:

$$\operatorname{Var}[\mathbf{w}_n|\mathbf{X}] = \frac{1}{n}\mathbf{X}'\boldsymbol{\Omega}^{-1}E[\mathbf{u}\mathbf{u}'|\mathbf{X}]\boldsymbol{\Omega}^{-1}\mathbf{X} = \sigma^2\left(\frac{\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X}}{n}\right)$$

which converges to $\sigma^2\mathbf{Q}_{GLS}$. Under finite fourth-order moment conditions, the Multivariate Central Limit Theorem provides:

$$\frac{\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{u}}{\sqrt{n}} \xrightarrow{d} N\big(\mathbf{0},\, \sigma^2\mathbf{Q}_{GLS}\big) \tag{3.59}$$

Step 4: Application of Slutsky's Lemma

Combining (3.55) and (3.59) into (3.58) via Slutsky's Lemma:

$$\sqrt{n}(\hat{\beta}_{GLS} - \beta) \xrightarrow{d} \mathbf{Q}_{GLS}^{-1}\cdot N\big(\mathbf{0},\, \sigma^2\mathbf{Q}_{GLS}\big)$$

By the property of linear transformations of Gaussian vectors:

$$\boxed{\sqrt{n}(\hat{\beta}_{GLS} - \beta) \xrightarrow{d} N\big(\mathbf{0},\, \sigma^2\mathbf{Q}_{GLS}^{-1}\big)} \tag{3.60}$$

3.6.6 Role of Assumptions in GLS Properties

The **symmetry and positive definiteness of $\boldsymbol{\Omega}$** (Assumption 3.16) are the formal core of the entire derivation. Without them:

- the factorization $\mathbf{P}'\mathbf{P} = \boldsymbol{\Omega}^{-1}$ is undefined or generates a singular transformation matrix;
- mapping to the spherical space becomes impossible;
- the inversion of the moment matrix in (3.39) loses validity.

The **asymptotic full rank of $\mathbf{Q}_{GLS}$** is indispensable for the validity of matrix inversion in the limit. If $\mathbf{Q}_{GLS}$ were singular, the asymptotic variance would be undefined, invalidating inference in large samples.

3.7 Feasible Generalized Least Squares (FGLS) and the Relationship with EHW

3.7.1 Motivation and Geometric Interpretation

The Generalized Least Squares (**GLS**) estimator is an idealized theoretical construction: it requires full knowledge of the population variance-covariance matrix $\mathbf{\Omega}$, a condition rarely met in empirical data. The Feasible Generalized Least Squares (**FGLS**) estimator addresses this practical limitation by substituting the unknown matrix $\mathbf{\Omega}$ with a consistent estimator $\hat{\mathbf{\Omega}} = \mathbf{\Omega}(\hat{\theta})$, constructed from the residuals of an initial parametric specification estimated by OLS.

Comparative Geometric Interpretation:

Imagine the uncertainty cloud of the original data in $\mathbb{R}^n$ as an elongated ellipsoid due to heteroskedasticity.

- **GLS/FGLS** applies a linear transformation that "spherifies" this deformed cloud back to a perfectly spherical geometry before projecting the vector $\mathbf{y}$ onto the column space of $\mathbf{X}$.
- **EHW** accepts the standard Euclidean projection of OLS onto the original elliptical cloud and merely inflates or rotates the axes of the confidence ellipsoid of $\hat{\beta}_{OLS}$ so that the reported confidence intervals are not misleadingly narrow.

Fragility of FGLS in Finite Samples:

While theoretical GLS is guaranteed to be the Best Linear Unbiased Estimator (**BLUE**) in finite samples, FGLS **loses the unbiasedness property** in finite samples due to the randomness introduced by the estimation of $\hat{\theta}$. Under small samples or if the heteroskedasticity structure is misspecified, FGLS may exhibit sample variance significantly **greater than that of ordinary OLS**, becoming inefficient.

Crucial Distinction between FGLS and EHW:

EHW **does not constitute a variant or subclass of GLS**:

- The GLS/FGLS framework acts directly on the **core of the estimator**, altering the coefficient vector $\hat{\beta}$ to extract efficiency gains through orthogonal reweighting of the data.
- EHW **preserves the OLS estimate vector intact** — accepting its inefficiency under non-spherical errors — and merely corrects the variance-covariance matrix *ex-post*, ensuring only that inference remains asymptotically valid.

3.7.2 Assumptions of FGLS

Assumption 3.17: Structural Consistency of the Variance Estimator ($\operatorname{plim} \hat{\mathbf{\Omega}} = \mathbf{\Omega}$)

- **Role:** Ensures that, as the sample size grows, the estimated matrix $\hat{\mathbf{\Omega}}$ collapses onto the true population structure, allowing FGLS to recover the equivalence and asymptotic efficiency of idealized GLS.
- **Counterexample:** If the parametric function chosen to model the variance is incorrect (e.g., modeling heteroskedasticity as linear when it is exponential), $\hat{\mathbf{\Omega}}$ will converge to an erroneous matrix, and asymptotic FGLS will be less efficient than simple OLS.

Assumption 3.18: Strict Exogeneity ($E[\mathbf{u}|\mathbf{X}] = \mathbf{0}$)

- **Role:** Indispensable condition for ensuring FGLS consistency. Unlike ordinary OLS — which requires only contemporaneous orthogonality $E[\mathbf{x}_i u_i] = \mathbf{0}$ for consistency — the linear transformation applied by FGLS mixes residuals and regressors from different observations.
- **Counterexample:** In time series models with lagged dependent variables (e.g., autoregressive models), the errors are orthogonal to contemporaneous regressors but correlated with future regressors. In this scenario, OLS with EHW correction remains consistent, while FGLS becomes severely **inconsistent**.

3.7.3 Formal Statement and Derivation

3.7.3.1 Formal Statement Consider the linear population model:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{u} \tag{3.61}$$

where $E[\mathbf{u}|\mathbf{X}] = \mathbf{0}$ and $\text{Var}[\mathbf{u}|\mathbf{X}] = \sigma^2\mathbf{\Omega}(\theta)$, with $\theta \in \mathbb{R}^p$ containing the unknown variance parameters. The FGLS estimator is defined by:

$$\boxed{\hat{\beta}_{FGLS} = (\mathbf{X}'\hat{\mathbf{\Omega}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{\Omega}}^{-1}\mathbf{y}} \tag{3.62}$$

where $\hat{\mathbf{\Omega}} = \mathbf{\Omega}(\hat{\theta})$ and $\hat{\theta}$ is a consistent estimator such that $\hat{\theta} \xrightarrow{p} \theta$.

Objectives:

1. Demonstrate the asymptotic equivalence between $\hat{\beta}_{FGLS}$ and $\hat{\beta}_{GLS}$.
2. Prove by algebraic contradiction that EHW cannot be mapped as a special case of GLS reweighting.

3.7.3.2 Formal Derivation Step 1: Decomposition and Scaling of the Estimation Error

Substituting $\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$ into (3.62):

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

Multiplying by $\sqrt{n}$:

$$\sqrt{n}(\hat{\beta}_{FGLS} - \beta) = \left(\frac{\mathbf{X}'\hat{\mathbf{\Omega}}^{-1}\mathbf{X}}{n}\right)^{-1}\left(\frac{\mathbf{X}'\hat{\mathbf{\Omega}}^{-1}\mathbf{u}}{\sqrt{n}}\right) \tag{3.63}$$

Step 2: Application of Slutsky's Lemma

Since $\hat{\theta} \xrightarrow{p} \theta$, the Continuous Mapping Theorem guarantees that $\hat{\mathbf{\Omega}}^{-1} \xrightarrow{p} \mathbf{\Omega}^{-1}$. We evaluate the limits:

1. By the Law of Large Numbers:

$$\text{plim}\left(\frac{\mathbf{X}'\hat{\mathbf{\Omega}}^{-1}\mathbf{X}}{n}\right) = \text{plim}\left(\frac{\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X}}{n}\right) = \mathbf{Q}_{GLS} \tag{3.64}$$

2. By the consistency of $\hat{\mathbf{\Omega}}$ and strict exogeneity:

$$\frac{\mathbf{X}'\hat{\mathbf{\Omega}}^{-1}\mathbf{u}}{\sqrt{n}} - \frac{\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{u}}{\sqrt{n}} \xrightarrow{p} \mathbf{0} \tag{3.65}$$

By **Slutsky's Lemma**, the asymptotic distribution of FGLS converges exactly to that of theoretical GLS:

$$\sqrt{n}(\hat{\beta}_{FGLS} - \beta) \xrightarrow{d} N\left(\mathbf{0},\, \sigma^2\mathbf{Q}_{GLS}^{-1}\right) \tag{3.66}$$

Partial Conclusion: The equivalence and efficiency of FGLS are strictly **asymptotic** properties. In finite samples, $\text{Var}[\hat{\beta}_{FGLS}|\mathbf{X}] \neq \sigma^2(\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X})^{-1}$ due to the remaining sample variance from $\hat{\theta}$.

Step 3: Algebraic Distinction of EHW

The EHW estimator (HC0) is defined by:

$$\widehat{\text{Var}}_{EHW} = (\mathbf{X}'\mathbf{X})^{-1}\left(\mathbf{X}'\hat{\mathbf{\Phi}}\mathbf{X}\right)(\mathbf{X}'\mathbf{X})^{-1}, \qquad \hat{\mathbf{\Phi}} = \text{diag}(\hat{u}_1^2, \ldots, \hat{u}_n^2) \tag{3.67}$$

Proof by Contradiction: Suppose EHW is identical to a GLS estimation with arbitrary weight matrix $\mathbf{W}$. Then the coefficient estimator should be:

$$\hat{\beta}_{EHW} = (\mathbf{X}'\mathbf{W}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{-1}\mathbf{y} \tag{3.68}$$

But by definition, the estimator that accompanies the EHW matrix is ordinary OLS:

$$\hat{\beta}_{EHW} = \hat{\beta}_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{3.69}$$

For (3.68) and (3.69) to be algebraically identical for any $\mathbf{y}$, the equality of operators requires:

$$(\mathbf{X}'\mathbf{W}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{-1} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

This identity is satisfied **if, and only if**, $\mathbf{W} = \sigma^2\mathbf{I}_n$ (pure homoskedasticity). Under heteroskedasticity, there is no $\mathbf{W} \neq \sigma^2\mathbf{I}_n$ that preserves the OLS estimator. This proves that EHW operates **exclusively as an *ex-post* inference correction**, and not as an efficient reweighting of the data.

3.8 Inference Foundation: χ^2 versus F in Each Framework

3.8.1 Unified Definition of the Wald Statistic

Under the null hypothesis of q linear restrictions $H_0 : \mathbf{R}\beta - \mathbf{r} = \mathbf{0}$, with $\text{rank}(\mathbf{R}) = q$, the Wald statistic is:

$$W = (\mathbf{R}\hat{\beta} - \mathbf{r})' \left[\mathbf{R}\widehat{\text{Var}}(\hat{\beta})\mathbf{R}'\right]^{-1} (\mathbf{R}\hat{\beta} - \mathbf{r}) \tag{3.70}$$

3.8.2 The EHW Scenario: Purely Asymptotic Inference

In the EHW scenario, we maintain $\hat{\beta} = \hat{\beta}_{OLS}$ and $\widehat{\text{Var}}(\hat{\beta}) = \widehat{\text{Var}}_{EHW}$.

3.8.2.1 Convergence to χ^2_q Under H_0:

$$\mathbf{R}\hat{\beta}_{OLS} - \mathbf{r} = \mathbf{R}(\hat{\beta}_{OLS} - \beta)$$

Multiplying by $\sqrt{n}$ and invoking the CLT:

$$\sqrt{n}\mathbf{R}(\hat{\beta}_{OLS} - \beta) \xrightarrow{d} N\left(\mathbf{0}, \mathbf{R}\mathbf{V}_{EHW}\mathbf{R}'\right)$$

Since $\text{plim}(n\widehat{\text{Var}}_{EHW}) = \mathbf{V}_{EHW}$, Slutsky's Lemma provides:

$$\boxed{W \xrightarrow{d} \chi^2_q} \tag{3.71}$$

3.8.2.2 Limitation in Finite Samples Under EHW with generic heteroskedasticity, the exact distribution of (3.70) in finite samples is **completely unknown**. There is no mathematical proof that the denominator of the sandwich form is distributed as a Chi-square independent of the numerator.

> **Practical Note:** The practice of computing $F = W/q$ and testing against $F_{q,n-k}$ in the EHW context is a **heuristic and conservative approximation** for small samples, lacking support in finite-sample theory.

3.8.3 The GLS Scenario: Exact Finite-Sample Inference

In theoretical GLS, the data are filtered by $\mathbf{P}$, restoring sphericity.

3.8.3.1 Derivation of the F Statistic

1. In the transformed model $\mathbf{y}^* = \mathbf{X}^*\beta + \mathbf{u}^*$, we assume normality: $\mathbf{u}^* \sim N(\mathbf{0}, \sigma^2\mathbf{I}_n)$.
2. Under normality:
$$\hat{\beta}_{GLS} \sim N\left(\beta,\, \sigma^2(\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X})^{-1}\right)$$
3. The restriction error is:
$$\mathbf{m} = \mathbf{R}\hat{\beta}_{GLS} - \mathbf{r} \sim N\left(\mathbf{0},\, \sigma^2\mathbf{R}(\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{R}'\right)$$
4. The normalized quadratic form:
$$\frac{W_{\sigma^2}}{q} = \frac{(\mathbf{R}\hat{\beta}_{GLS} - \mathbf{r})' \left[\mathbf{R}(\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{R}'\right]^{-1} (\mathbf{R}\hat{\beta}_{GLS} - \mathbf{r})}{q \cdot \sigma^2} \sim \frac{\chi^2_q}{q} \tag{3.72}$$
5. The estimator of σ^2 is:
$$s^2_{GLS} = \frac{\hat{\mathbf{u}}^{*'}_{GLS}\hat{\mathbf{u}}^*_{GLS}}{n-k} \quad \Longrightarrow \quad \frac{(n-k)s^2_{GLS}}{\sigma^2} \sim \chi^2_{n-k} \tag{3.73}$$
6. By independence between numerator and denominator (property of orthogonal projections under normality):
$$F = \frac{W_{\sigma^2}/q}{s^2_{GLS}/\sigma^2} = \frac{(\mathbf{R}\hat{\beta}_{GLS} - \mathbf{r})' \left[\mathbf{R}(\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{R}'\right]^{-1} (\mathbf{R}\hat{\beta}_{GLS} - \mathbf{r})}{q \cdot s^2_{GLS}} \tag{3.74}$$
7. **Conclusion:**
$$\boxed{F \sim F_{q,\,n-k}} \tag{3.75}$$

3.8.4 Comparative Summary of Inference Environments

Dimension	EHW (Robust OLS)	GLS (Theoretical)
Coefficient Estimator	$\hat{\beta}_{OLS}$ (inefficient under non-spherical errors)	$\hat{\beta}_{GLS}$ (BLUE in finite samples)
Theoretical Foundation	Pure Asymptotic Theory ($n \to \infty$)	Exact Finite-Sample Distribution
Statistic Distribution	χ^2_q (CLT collapse)	$F_{q,n-k}$ (exact ratio of two χ^2)
Assumption about Errors	Ignores and tolerates the fine structure	Requires normality and knowledge of $\mathbf{\Omega}$

Indispensable Conclusion:

- In **GLS**, the transition to the F distribution is algebraically legitimate: the linear transformation cleans the covariance structure, restoring sphericity and allowing the emergence of an independent Chi-square in the denominator (s^2_{GLS}).
- In **EHW**, we stop at χ^2 because the refusal to model the error structure prevents the factorization and isolation of independent Chi-square terms in small samples, forcing inference to depend exclusively on the Gaussian asymptotic limit.

3.9 Comparative Table: Treatment of Pure Heteroskedasticity

The scope is restricted to pure heteroskedasticity: $\text{Var}(\mathbf{u}|\mathbf{X}) = \sigma^2\mathbf{\Omega}$, with $\mathbf{\Omega}$ diagonal and entries $\omega_{ii} = h(\mathbf{x}_i) > 0$. EHW refers to OLS with White's robust adjustment.

Criterion	EHW (Robust OLS)	GLS (Theoretical)	FGLS (Feasible)
Estimator of β and Stages	$\hat{\beta}_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. **1 stage:** direct orthogonal projection.	$\hat{\beta}_{GLS} = (\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{y}$. **1 stage:** requires known $\mathbf{\Omega}$.	$\hat{\beta}_{FGLS} = (\mathbf{X}'\hat{\mathbf{\Omega}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{\Omega}}^{-1}\mathbf{y}$. **2 stages:** estimates $\hat{\mathbf{\Omega}}$ via OLS residuals; then estimates β via WLS.
Assumption about $\mathbf{\Omega}$	**Arbitrary and unknown.** Imposes no functional form on $h(\mathbf{x}_i)$. Consistency depends only on contemporaneous exogeneity.	**Known *ex ante*. $\mathbf{\Omega}$** known up to σ^2. Impractical in real contexts.	**Parametric structure.** Assumes $\omega_{ii} = h(\mathbf{x}_i, \theta)$ with θ consistently estimated. Misspecification causes inefficiency; in dynamic models, may generate inconsistency.
Finite Samples	**Unbiased but inefficient.** $E[\hat{\beta}_{OLS}\|\mathbf{X}] = \beta$. Not BLUE.	**BLUE — Aitken's Theorem.** Best linear unbiased estimator under metric weighted by $\mathbf{\Omega}^{-1}$.	**Generally biased.** Dependence between $\hat{\mathbf{\Omega}}$ and errors invalidates unbiasedness. No analogue to Aitken's Theorem.
Asymptotic Properties	**Consistent and $\sqrt{n}$-normal.** $\sqrt{n}(\hat{\beta}_{OLS} - \beta) \xrightarrow{d} N(\mathbf{0}, \mathbf{V}_{EHW})$, with $\mathbf{V}_{EHW} = \mathbf{Q}^{-1}\mathbf{B}\mathbf{Q}^{-1}$.	**Consistent and asymptotically efficient.** $\mathbf{V}_{GLS} = (E[\mathbf{x}_i\mathbf{x}_i'/h(\mathbf{x}_i)])^{-1} \preceq \mathbf{V}_{EHW}$.	**Asymptotically equivalent to GLS.** If $\text{plim}(\hat{\mathbf{\Omega}}) = \mathbf{\Omega}$, achieves the same efficiency as theoretical GLS.
Rank Requirements	$\text{rank}(\mathbf{X}) = k$; $\mathbf{Q} = \text{plim}(\mathbf{X}'\mathbf{X}/n)$ SPD.	Same as EHW, plus $\mathbf{\Omega}$ SPD for existence of $\mathbf{P}$.	GLS requirements + identifiability of θ in $h(\mathbf{x}_i, \theta)$.
Inference	**Asymptotic and doubly robust.** Based on χ^2 or Normal. Dispenses with specification of $h(\cdot)$.	**Exact under Normality.** t and F with exact distributions in the transformed space.	**Asymptotic and sensitive.** t and F valid only for large n. Sensitive to misspecification of $h(\cdot)$.
Main Limitation	**Efficiency loss.** $\mathbf{V}_{EHW} - \mathbf{V}_{GLS} \succeq \mathbf{0}$. Greater cost with pronounced heteroskedasticity.	**Practical infeasibility. $\mathbf{\Omega}$** rarely known in empirical contexts.	**Risk of misspecification.** Mild misspecification generates inefficiency; severe model misspecification can generate inconsistency.

3.9.1 Methodological Notes

1. **Sandwich Estimator (EHW):** The asymptotic variance $\mathbf{V}_{EHW} = \mathbf{Q}^{-1}\mathbf{B}\mathbf{Q}^{-1}$ is estimated by $\hat{\mathbf{V}}_{EHW} = (\mathbf{X}'\mathbf{X})^{-1}\left(\sum_i \hat{u}_i^2\mathbf{x}_i\mathbf{x}_i'\right)(\mathbf{X}'\mathbf{X})^{-1}$, without imposing structure on $\mathbf{\Omega}$.
2. **Aitken's Transformation:** GLS premultiplies the model by $\mathbf{P}$ such that $\mathbf{P}'\mathbf{P} = \mathbf{\Omega}^{-1}$, producing spherical errors $\text{Var}(\mathbf{P}\mathbf{u}|\mathbf{X}) = \sigma^2\mathbf{I}$ and restoring the Gauss-Markov conditions in the transformed space.
3. **FGLS Usage Frontier:** In static models with strict exogeneity, FGLS is asymptotically preferable to EHW if $h(\cdot)$ is reasonably specified. In dynamic models, the recommendation reverses: EHW

with OLS is methodologically safer.

Chapter 4: Endogeneity

4.1 Sources of Endogeneity

The problem of **endogeneity** arises when the error term of a regression model is correlated with the regressors — violating the condition of strict or contemporaneous exogeneity. Three classical sources are responsible for this violation: (i) omission of a relevant variable, (ii) measurement error in the regressors, and (iii) simultaneity (or reverse causality). Each of these sources corrupts the consistency of the Ordinary Least Squares (OLS) estimator, albeit through distinct mechanisms.

4.1.1 Omitted Variable

4.1.1.1 Motivation and Geometric Interpretation The omitted variable problem emerges when the researcher wishes to estimate the *ceteris paribus* effect of a set of regressors $\mathbf{X}_1$ on $\mathbf{y}$, but is unable to observe or include in the model a matrix of covariates $\mathbf{X}_2$ that simultaneously affects $\mathbf{y}$ and is correlated with $\mathbf{X}_1$.

Geometric Interpretation in $\mathbb{R}^n$:

The OLS estimator of the complete model — the *long regression* — orthogonally projects $\mathbf{y}$ onto the hyperplane defined by the combined column space $\text{Col}(\mathbf{X}_1, \mathbf{X}_2)$. By omitting $\mathbf{X}_2$, we force the projection of $\mathbf{y}$ solely onto the subspace $\text{Col}(\mathbf{X}_1)$. If $\mathbf{X}_1$ and $\mathbf{X}_2$ are not orthogonal, the portion of the variation in $\mathbf{y}$ that should be attributed to $\text{Col}(\mathbf{X}_2)$ is partially absorbed by $\text{Col}(\mathbf{X}_1)$. The coefficient associated with $\mathbf{X}_1$ improperly incorporates the effect of $\mathbf{X}_2$, producing an estimate that fails to isolate the true structural parameter.

4.1.1.2 Assumptions and Their Roles **Assumption 4.1: Linearity in Parameters and Structural Specification**

- **Role:** Defines that the population Data Generating Process (DGP) obeys a linear and additive relationship in the observed and unobserved components.
- **Counterexample:** If the underlying relationship is strictly non-linear (e.g., $\mathbf{y} = \exp(\mathbf{X}\beta)$), OLS coefficients lose the interpretation of constant marginal effects, invalidating the additive decomposition of the error.

Assumption 4.2: No Perfect Multicollinearity

- **Role:** Ensures that the Gram matrix $\mathbf{X}'\mathbf{X}$ has full column rank ($\text{rank}(\mathbf{X}) = k$), guaranteeing the existence of its inverse.
- **Counterexample:** If a column of $\mathbf{X}_1$ is an exact linear combination of the others, the determinant of the Gram matrix collapses ($|\mathbf{X}'\mathbf{X}| = 0$) and the estimator becomes algebraically indeterminate.

Assumption 4.3: Strict Exogeneity of the Long Model

- **Role:** Establishes that, in the model conditioned on all structurally relevant variables, the disturbance vector ϵ has zero conditional mean: $E[\epsilon \mid \mathbf{X}_1, \mathbf{X}_2] = \mathbf{0}$.
- **Counterexample:** Without this assumption, even the inclusion of $\mathbf{X}_2$ does not restore the consistency of the estimator, as the error would retain factors correlated with the regressors.

4.1.1.3 Formal Statement and Derivation Consider the true structural model — the *long regression*:

$$\mathbf{y} = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \epsilon \tag{4.1}$$

where $\mathbf{y} \in \mathbb{R}^n$, $\mathbf{X}_1 \in \mathbb{R}^{n \times k_1}$, $\mathbf{X}_2 \in \mathbb{R}^{n \times k_2}$, and $\epsilon \in \mathbb{R}^n$. We assume:

- (i) $E[\epsilon \mid \mathbf{X}_1, \mathbf{X}_2] = \mathbf{0}$ (strict exogeneity in the long model);
- (ii) $\text{rank}(\mathbf{X}_1'\mathbf{X}_1) = k_1$ and $\text{rank}(\mathbf{X}_2'\mathbf{X}_2) = k_2$ (full rank);
- (iii) $\beta_2 \neq \mathbf{0}$ (structural relevance of the omitted variable);
- (iv) $\mathbf{X}_1'\mathbf{X}_2 \neq \mathbf{0}$ (sample non-orthogonality between regressor blocks).

The researcher estimates the *short regression*:

$$\mathbf{y} = \mathbf{X}_1\gamma_1 + \mathbf{u} \tag{4.2}$$

where $\mathbf{u} = \mathbf{X}_2\beta_2 + \epsilon$ is the composite error. The OLS estimator is $\tilde{\gamma}_1 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y}$.

Theorem (Omitted Variable Bias). Under conditions (iii) and (iv), $E[\tilde{\gamma}_1 \mid \mathbf{X}_1, \mathbf{X}_2] \neq \beta_1$; the estimator $\tilde{\gamma}_1$ is biased for the vector of structural parameters β_1.

Proof:

Step 1. Write the short regression estimator:

$$\tilde{\gamma}_1 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y} \tag{4.3}$$

Step 2. Substitute the structural DGP (4.1) into (4.3):

$$\tilde{\gamma}_1 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'(\mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \epsilon) \tag{4.4}$$

Step 3. Distribute the matrix product:

$$\tilde{\gamma}_1 = \beta_1 + (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\beta_2 + (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\epsilon \tag{4.5}$$

Step 4. Apply the conditional expectation on $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2]$:

$$E[\tilde{\gamma}_1 \mid \mathbf{X}] = \beta_1 + (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\beta_2 + (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'E[\epsilon \mid \mathbf{X}]$$

The third block vanishes by Assumption 4.3, resulting in:

$$\boxed{E[\tilde{\gamma}_1 \mid \mathbf{X}] = \beta_1 + (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\beta_2} \tag{4.6}$$

Analysis of the Bias Term:

The term $\mathbf{B} = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\beta_2$ is the **omitted variable bias** (OVB). The expression $(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2$ is algebraically equivalent to the OLS coefficient of an auxiliary regression of $\mathbf{X}_2$ on $\mathbf{X}_1$, denoted $\tilde{\delta}_{21}$. Hence:

$$E[\tilde{\gamma}_1 \mid \mathbf{X}] = \beta_1 + \tilde{\delta}_{21}\beta_2 \tag{4.7}$$

The bias exists if, and only if, $\tilde{\delta}_{21} \neq \mathbf{0}$ (non-orthogonality) **and** $\beta_2 \neq \mathbf{0}$ (structural relevance). If the included and omitted regressors are sample-orthogonal, $\tilde{\delta}_{21} = \mathbf{0}$ and the estimator remains unbiased — although the error variance is still affected. □

4.1.1.4 Asymptotic Properties under Omission Theorem (Inconsistency and Limiting Distribution). If $E[\mathbf{x}_{1i}\mathbf{x}'_{2i}] = \mathbf{Q}_{12} \neq \mathbf{0}$ and $\beta_2 \neq \mathbf{0}$:

$$\boxed{\text{plim } \hat{\gamma}_1 = \beta_1 + \mathbf{Q}_{11}^{-1}\mathbf{Q}_{12}\beta_2 \neq \beta_1} \tag{4.8}$$

$$\boxed{\sqrt{n}\left(\hat{\gamma}_1 - \text{plim } \hat{\gamma}_1\right) \xrightarrow{d} N\left(\mathbf{0},\, \mathbf{Q}_{11}^{-1}\mathbf{\Omega}\mathbf{Q}_{11}^{-1}\right)} \tag{4.9}$$

where $\mathbf{Q}_{11} = E[\mathbf{x}_{1i}\mathbf{x}'_{1i}]$, $\mathbf{Q}_{12} = E[\mathbf{x}_{1i}\mathbf{x}'_{2i}]$, and $\mathbf{\Omega} = E[u_i^{*2}\mathbf{x}_{1i}\mathbf{x}'_{1i}]$.

Proof (sketch): Applying the Weak Law of Large Numbers and Slutsky's Theorem to the scaled form of the estimator yields (4.8). The Central Limit Theorem provides (4.9). The complete detail follows the pattern established in the heteroskedasticity chapter. ■

4.1.1.5 Inference Consequences Under endogeneity by omission, the t-statistic for $H_0 : \beta_{1,j} = \beta_{1,j}^0$ satisfies:

$$\boxed{|T| \xrightarrow{p} \infty \quad \text{as } n \to \infty} \tag{4.10}$$

The test ceases to be an instrument of scientific inference and becomes a detector of omitted correlations. The same holds for the F-test, which diverges to $+\infty$ at a rate n, producing spurious rejections of any linear restrictions on β_1.

4.1.2 Measurement Error

4.1.2.1 Motivation and Geometric Interpretation The measurement error problem arises from the discrepancy between the theoretical variable prescribed by economic theory (such as "permanent income" or "human capital") and the variable actually recorded in the data (such as "declared monthly income" or "years of schooling"). While measurement error in the dependent variable only expands the residual variance without compromising consistency, measurement error in the regressors makes the OLS estimator biased and inconsistent even in infinite samples.

Geometric Interpretation in $\mathbb{R}^2$:

Adding noise to the horizontal coordinate x_i^* artificially increases the horizontal dispersion of the observed sample. Since OLS minimizes the sum of squared vertical deviations, the estimated line is "pulled" toward the horizontal axis to accommodate this greater fictitious lateral dispersion. The result is a flatter line than the true structural relationship. This phenomenon of parameter compression toward zero is called **attenuation bias**.

4.1.2.2 Assumptions of the Measurement Error Model Assumption 4.4: Classical Measurement Error (CME) Model

- **Role:** The measurement noise is orthogonal to the true value of the latent variable and to the structural error.
- **Counterexample:** If the error were correlated with the true value (e.g., high-income individuals underreporting earnings proportionally to wealth), the bias would no longer be attenuation and could inflate or invert the sign of the parameter unpredictably.

Assumption 4.5: Mutual Independence of Noises

- **Role:** The measurement noise $\mathbf{w}$ and the structural disturbance ϵ are mutually independent: $\text{Cov}(\mathbf{w}, \epsilon) = \mathbf{0}$.

- **Counterexample:** If the factor causing measurement error in $\mathbf{x}$ also directly affected $\mathbf{y}$ (e.g., a negligent interviewer who erroneously records both education and salary), an additional spurious correlation would emerge, preventing the analytical isolation of the attenuation effect.

4.1.2.3 Formal Statement and Derivation Consider the latent structural model:

$$y_i = \beta x_i^* + \epsilon_i \tag{4.11}$$

where x_i^* is the unobserved true regressor and $\beta \in \mathbb{R}$ is the parameter of interest. The observed proxy variable is:

$$x_i = x_i^* + w_i \tag{4.12}$$

We assume:

- (i) $E[\epsilon_i] = E[w_i] = 0$;
- (ii) $\mathrm{Cov}(x_i^*, \epsilon_i) = 0$;
- (iii) $\mathrm{Cov}(x_i^*, w_i) = 0$ and $\mathrm{Cov}(w_i, \epsilon_i) = 0$;
- (iv) $\mathrm{Var}(x_i^*) = \sigma_{x^*}^2 > 0$ and $\mathrm{Var}(w_i) = \sigma_w^2 > 0$.

Theorem (Attenuation). The estimator $\hat{\beta} = \sum x_i y_i / \sum x_i^2$ satisfies:

$$\boxed{\operatorname{plim} \hat{\beta} = \beta \left(\frac{\sigma_{x^*}^2}{\sigma_{x^*}^2 + \sigma_w^2} \right) = \beta\lambda} \tag{4.13}$$

with $\lambda \in (0, 1)$, proving strict inconsistency by attenuation.

Proof:

Step 1. Substitute $x_i^* = x_i - w_i$ into (4.11):

$$y_i = \beta x_i + (\epsilon_i - \beta w_i) \tag{4.14}$$

Step 2. Define the composite error $v_i \equiv \epsilon_i - \beta w_i$, so that $y_i = \beta x_i + v_i$.

Step 3. Compute $\mathrm{Cov}(x_i, v_i)$ using bilinearity:

$$\mathrm{Cov}(x_i, v_i) = \mathrm{Cov}(x_i^* + w_i,\, \epsilon_i - \beta w_i) = -\beta\sigma_w^2 \tag{4.15}$$

Step 4. Compute $\mathrm{Var}(x_i)$:

$$\mathrm{Var}(x_i) = \mathrm{Var}(x_i^* + w_i) = \sigma_{x^*}^2 + \sigma_w^2 \tag{4.16}$$

Step 5. Apply the result $\operatorname{plim} \hat{\beta} = \beta + \mathrm{Cov}(x, v)/\mathrm{Var}(x)$:

$$\operatorname{plim} \hat{\beta} = \beta + \frac{-\beta\sigma_w^2}{\sigma_{x^*}^2 + \sigma_w^2} = \beta \left(\frac{\sigma_{x^*}^2}{\sigma_{x^*}^2 + \sigma_w^2} \right) \tag{4.17}$$

□

4.1.2.4 Inference Consequences Measurement error simultaneously corrupts the centrality of the limiting distribution and the consistency of the residual variance:

1. **Power collapse ($H_0 : \beta = 0$):** The attenuation bias compresses the estimator toward zero, reducing test power and increasing the Type II error rate.
2. **Type I error explosion ($H_0 : \beta = \beta_{\text{structural}}$):** The numerator of the t-statistic ceases to be centered at zero; as n grows, the test diverges and systematically rejects true hypotheses. Formally:

$$\boxed{|t_n| \xrightarrow{p} \infty \quad \text{as } n \to \infty \text{ under } H_0 : \beta = \beta_0} \tag{4.18}$$

4.1.3 Simultaneity

4.1.3.1 Motivation and Geometric Interpretation The simultaneous equations problem arises when explanatory variables are treated as exogenously fixed when they are actually determined by an integrated systemic equilibrium. The canonical example is the competitive market: price P affects quantity demanded Q, but Q feeds back into price through the market equilibrium mechanism.

Geometric Interpretation in $\mathbb{R}^2$:

The observed data do not belong individually to either the demand or supply curves; they constitute equilibrium points resulting from the stochastic intersection of both. An OLS line fitted to these points does not represent the price elasticity of demand nor that of supply, but an ambiguous linear combination of the two, without structural interpretation.

4.1.3.2 Assumptions of the Simultaneous System **Assumption 4.6: Structural Completeness of the System**

- **Role:** Requires as many independent equations as there are endogenous variables, ensuring that the matrix of structural coefficients is invertible and that the system has a unique reduced-form solution.
- **Counterexample:** Without this condition, the structural parameters are not identifiable.

Assumption 4.7: Exogeneity of Predetermined Variables

- **Role:** The exogenous variables $\mathbf{z}$ are orthogonal to the contemporaneous structural shocks: $E[\mathbf{z}_t \mathbf{u}_t] = \mathbf{0}$.
- **Counterexample:** Without this condition, the absence of a stable informational anchor prevents the identification of the system.

4.1.3.3 Formal Statement and Derivation Consider the two-equation structural system:

$$y_{1t} = \alpha_1 y_{2t} + \beta_1 x_t + u_{1t} \tag{4.19}$$

$$y_{2t} = \alpha_2 y_{1t} + u_{2t} \tag{4.20}$$

where y_{1t} and y_{2t} are endogenous, x_t is strictly exogenous, and $1 - \alpha_1 \alpha_2 \neq 0$.

The OLS estimator of α_1 in (4.19) is $\hat{\alpha}_1 = \sum y_{2t} y_{1t} / \sum y_{2t}^2$.

Theorem (Inconsistency by Simultaneity). The estimator $\hat{\alpha}_1$ converges in probability to $\alpha^* \neq \alpha_1$:

$$\boxed{\text{plim } \hat{\alpha}_1 = \alpha_1 + \frac{\alpha_2 \sigma_1^2 + \sigma_{12}}{(1 - \alpha_1 \alpha_2) \operatorname{Var}(y_{2t})}} \tag{4.21}$$

Proof (sketch): Solving the system to obtain the reduced form of y_{2t} and computing $\text{Cov}(y_{2t}, u_{1t})$ yields (4.21). The complete detail follows the established pattern. ■

4.1.3.4 Inference Consequences Under simultaneity, the convergence of the estimator to a displaced pseudo-parameter corrupts inference:

1. **Distortion of the nominal level:** The OLS estimator converges to $\beta^* \neq \beta$; the limiting distribution of T is shifted away from $N(0,1)$.
2. **Spurious rejection of structural models:** The standard error converges to zero at rate $O(1/\sqrt{n})$, while the distance $|\beta^* - \beta|$ remains constant. The test statistic diverges:

$$\boxed{|t_n| \xrightarrow{p} \infty \quad \text{as } n \to \infty} \tag{4.22}$$

The solution is to migrate to Instrumental Variable estimators, which restore consistency by instrumenting the endogenous variables with regressors exogenous to the structural shock.

4.2 The Instrumental Variables (IV) and Two-Stage Least Squares (2SLS) Estimators

This section presents the mathematical and geometric formalization of the **Instrumental Variables (IV)** estimator and its generalization to overidentified models via **Two-Stage Least Squares (2SLS)**. The objective is to establish the statistical properties in finite samples and the asymptotic limiting behavior from the perspective of classical estimation theory.

4.2.1 The Identification Problem and the Geometry of Instruments

4.2.1.1 Motivation and Geometric Interpretation Identification is a logical prerequisite to estimation: it determines whether the information contained in the joint distribution of observable data, combined with the theoretical restrictions imposed by the economic model, is sufficient to uniquely deduce the structural parameters of interest.

The Ordinary Least Squares (OLS) estimator relies on the assumption of strict orthogonality between the regressors and the error term. When **endogeneity** occurs — whether through omitted variables, measurement error, or simultaneity — this condition is violated.

Consider the population structural model in matrix form:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{u} \tag{4.23}$$

where $\mathbf{y} \in \mathbb{R}^n$, $\mathbf{X} \in \mathbb{R}^{n \times k}$, $\beta \in \mathbb{R}^k$, and $\mathbf{u} \in \mathbb{R}^n$. The classical exogeneity assumption is violated in the population, that is, $E[\mathbf{X}'\mathbf{u}] \neq \mathbf{0}$.

Geometric Interpretation in $\mathbb{R}^n$:

If the space spanned by the regressors $\mathbf{X}$ is correlated with the error vector $\mathbf{u}$, the orthogonal projection of $\mathbf{y}$ directly onto $\text{Col}(\mathbf{X})$ captures not only the causal effect of β, but also the variation of the stochastic disturbance $\mathbf{u}$ transmitted through $\mathbf{X}$. OLS becomes inconsistent because it attributes to the parameter vector β covariances that actually belong to the population error.

The introduction of an instrument matrix $\mathbf{Z} \in \mathbb{R}^{n \times L}$ functions as an external reference axis in the Euclidean space. The instrument $\mathbf{Z}$ provides a source of variation for $\mathbf{X}$ that is perfectly orthogonal to $\mathbf{u}$. The central idea is to use only the portion of the variability of $\mathbf{X}$ that is provably exogenous to identify and isolate the true causal effect on $\mathbf{y}$.

4.2.1.2 Identification Assumptions Let $\mathbf{X}$ be the $n \times k$ matrix of regressors and $\mathbf{Z}$ the $n \times L$ matrix of instruments. For the vector of structural parameters β to be identified, two classical conditions must be satisfied:

Assumption 4.13: Order Condition ($L \geq k$)

- **Role:** This is a necessary (but not sufficient) condition for the system of population moment equations to have a unique solution. It requires that the number of external exogenous pieces of information (instruments) be at least equal to the number of unknown structural parameters to be estimated.
- **Identification Failure:** If $L < k$, the model is *underidentified*. There will be infinitely many parameter combinations consistent with the data, making any unique point estimate impossible.

Assumption 4.14: Rank Condition (Relevance)

- **Role:** Ensures that the instruments have a real and relevant linear statistical relationship with the endogenous regressors. Mathematically, it ensures that the population cross-moment matrix $\mathbf{Q}_{ZX} = E[\mathbf{z}_i \mathbf{x}_i']$ has full column rank, that is, $\text{rank}(\mathbf{Q}_{ZX}) = k$.
- **Identification Failure:** If the instrument $\mathbf{Z}$ is linearly independent of or completely orthogonal to $\mathbf{X}$ in the population ($\mathbf{Q}_{ZX} = \mathbf{0}$), the model is *unidentified*. Algebraically, the linear system will not have a unique solution and the estimator will collapse.

Assumption 4.15: Exogeneity of Instruments ($E[\mathbf{z}_i u_i] = \mathbf{0}$)

- **Role:** Fundamental condition for the consistency of the IV estimator: the instruments are not correlated with the structural error.
- **Counterexample:** If $E[\mathbf{z}_i u_i] \neq \mathbf{0}$, the IV estimator inherits the endogeneity of the instruments and becomes inconsistent — the bias may be even more severe than that of OLS.

4.2.1.3 Analytical Identification Scenarios Consider the linear structural model defined in (4.23). The identification of the vector β from the instrument matrix $\mathbf{Z}$ falls into three distinct analytical scenarios:

Scenario	Condition	Implication
Underidentification	$L < k$	Number of moment conditions is less than the number of unknowns. The structural parameter **cannot be determined**.
Exact Identification	$L = k$	There is exactly one instrument for each regressor. The classical IV estimator is obtained by direct algebraic solution.
Overidentification	$L > k$	There are more instruments than parameters. The system has more equations than unknowns, requiring optimal weighting (2SLS or GMM).

4.2.1.4 Algebraic Derivation under Exact Identification ($L = k$) We postulate the population moment condition based on the strict exogeneity of the instrument:

$$E[\mathbf{Z}'\mathbf{u}] = \mathbf{0} \quad \Longrightarrow \quad E[\mathbf{Z}'(\mathbf{y} - \mathbf{X}\beta)] = \mathbf{0} \tag{4.24}$$

By the principle of sample analogy, we replace the population expectation with its respective sample moment under exact identification ($L = k$):

$$\frac{1}{n}\mathbf{Z}'(\mathbf{y} - \mathbf{X}\hat{\beta}_{IV}) = \mathbf{0} \tag{4.25}$$

Multiplying by n and applying the distributive property:

$$\mathbf{Z}'\mathbf{y} - \mathbf{Z}'\mathbf{X}\hat{\beta}_{IV} = \mathbf{0} \quad \Longrightarrow \quad \mathbf{Z}'\mathbf{X}\hat{\beta}_{IV} = \mathbf{Z}'\mathbf{y} \tag{4.26}$$

Under **Assumption 4.14 (Rank Condition)** with $L = k$, the square matrix $\mathbf{Z}'\mathbf{X}$ of dimension $k \times k$ has full rank ($\text{rank}(\mathbf{Z}'\mathbf{X}) = k$), hence its inverse exists uniquely. Premultiplying by $(\mathbf{Z}'\mathbf{X})^{-1}$:

$$\boxed{\hat{\beta}_{IV} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}} \tag{4.27}$$

4.2.2 Finite-Sample Properties: Expectation and Variance

4.2.2.1 Motivation The proof of asymptotic consistency guarantees the appropriate behavior of the IV estimator as $n \to \infty$, but does not ensure the absence of bias in small samples. Unlike the OLS estimator under strict exogeneity, the IV estimator is **known to be biased in finite samples**.

4.2.2.2 Assumptions for Variance Derivation Assumption 4.16: Conditional Homoskedasticity and No Autocorrelation

$$E[\mathbf{u}\mathbf{u}'|\mathbf{Z}, \mathbf{X}] = \sigma^2\mathbf{I}_n \tag{4.28}$$

where $\sigma^2 > 0$ is the scalar conditional variance.

4.2.2.3 Formal Statement and Derivation Theorem (Finite-Sample Bias). The IV estimator is biased in finite samples:

$$E[\hat{\beta}_{IV}|\mathbf{Z}, \mathbf{X}] = \beta + (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'E[\mathbf{u}|\mathbf{Z}, \mathbf{X}] \tag{4.29}$$

Proof:

Substituting the structural equation (4.23) into the definition of the estimator (4.27):

$$\hat{\beta}_{IV} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'(\mathbf{X}\beta + \mathbf{u}) = \beta + (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{u} \tag{4.30}$$

Applying the conditional expectation $E[\cdot|\mathbf{Z}, \mathbf{X}]$:

$$E[\hat{\beta}_{IV}|\mathbf{Z}, \mathbf{X}] = \beta + (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'E[\mathbf{u}|\mathbf{Z}, \mathbf{X}] \tag{4.31}$$

Since $\mathbf{X}$ contains endogenous regressors correlated with the error, $E[\mathbf{u}|\mathbf{Z}, \mathbf{X}] \neq \mathbf{0}$. Therefore, the conditional bias is:

$$\boxed{\text{Bias}(\hat{\beta}_{IV}|\mathbf{Z}, \mathbf{X}) = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'E[\mathbf{u}|\mathbf{Z}, \mathbf{X}]} \tag{4.32}$$

Note: This finite-sample bias severely approaches the OLS bias as the relevance of the instruments diminishes, characterizing the *weak instruments* problem.

Theorem (Conditional Variance of the IV Estimator). Under Assumption 4.16 (conditional homoskedasticity):

$$\boxed{\mathrm{Var}(\hat{\beta}_{IV}|\mathbf{Z},\mathbf{X}) = \sigma^2 \left(\mathbf{Z}'\mathbf{X}\right)^{-1} \left(\mathbf{Z}'\mathbf{Z}\right) \left(\mathbf{X}'\mathbf{Z}\right)^{-1}} \tag{4.33}$$

Proof:

By definition:

$$\mathrm{Var}(\hat{\beta}_{IV}|\mathbf{Z},\mathbf{X}) = E\left[\left(\hat{\beta}_{IV} - E[\hat{\beta}_{IV}|\mathbf{Z},\mathbf{X}]\right)\left(\hat{\beta}_{IV} - E[\hat{\beta}_{IV}|\mathbf{Z},\mathbf{X}]\right)'\middle|\mathbf{Z},\mathbf{X}\right] \tag{4.34}$$

Substituting (4.30) and (4.31) into the definition and centering the analysis on the conditional distribution of the error:

$$\mathrm{Var}(\hat{\beta}_{IV}|\mathbf{Z},\mathbf{X}) = \left(\mathbf{Z}'\mathbf{X}\right)^{-1} \mathbf{Z}' E\left[\mathbf{u}\mathbf{u}'|\mathbf{Z},\mathbf{X}\right] \mathbf{Z} \left(\mathbf{X}'\mathbf{Z}\right)^{-1} \tag{4.35}$$

Invoking Assumption 4.16, the core of the expectation collapses to $\sigma^2\mathbf{I}_n$:

$$\mathrm{Var}(\hat{\beta}_{IV}|\mathbf{Z},\mathbf{X}) = \sigma^2 \left(\mathbf{Z}'\mathbf{X}\right)^{-1} \mathbf{Z}'\mathbf{Z} \left(\mathbf{X}'\mathbf{Z}\right)^{-1} \tag{4.36}$$

□

4.2.3 Generalization to the Overidentified Case: The 2SLS Estimator

4.2.3.1 Motivation and Geometric Interpretation When the number of instruments exceeds the number of regressors ($L > k$), the matrix $\mathbf{Z}'\mathbf{X}$ becomes rectangular of dimension $L \times k$ and cannot be directly inverted. The Two-Stage Least Squares (**2SLS**) estimator solves this problem through a double geometric projection procedure that efficiently combines all available information from the instruments.

Geometric Interpretation in $\mathbb{R}^n$:

The sample space $\mathbb{R}^n$ contains three fundamental subspaces:

1. Col($\mathbf{Z}$): the space spanned by the instruments (exogenous);
2. Col($\mathbf{X}$): the space spanned by the original regressors (contaminated by endogeneity);
3. Col($\hat{\mathbf{X}}$): the space spanned by the projected values of $\mathbf{X}$ onto Col($\mathbf{Z}$) — the "clean part" of the regressors.

2SLS operates in two geometric stages:

- **First Stage:** Orthogonally project each column of $\mathbf{X}$ onto Col($\mathbf{Z}$), eliminating the component correlated with the structural error. The result is $\hat{\mathbf{X}} = \mathbf{P}_Z\mathbf{X}$, the best linear approximation of $\mathbf{X}$ in the exogenous subspace.
- **Second Stage:** Project $\mathbf{y}$ onto Col($\hat{\mathbf{X}}$), obtaining the estimator that efficiently combines all available instruments.

4.2.3.2 Additional Assumptions for 2SLS Assumption 4.17: Rank Condition in the Overidentified Case

- **Role:** Ensures that $\mathrm{rank}(\mathbf{Z}'\mathbf{X}) = k$, guaranteeing that the instruments are sufficiently correlated with the endogenous regressors. In the overidentified case ($L > k$), this condition implies that $\mathbf{X}'\mathbf{P}_Z\mathbf{X}$ is non-singular.
- **Identification Failure:** If $\mathrm{rank}(\mathbf{Z}'\mathbf{X}) < k$, the instruments are weak or irrelevant for some regressors, and the matrix $\mathbf{X}'\mathbf{P}_Z\mathbf{X}$ becomes singular, preventing computation of the estimator.

Assumption 4.18: Finite Fourth-Order Moments

- **Role:** Guarantees the existence of asymptotic variances and the applicability of the Central Limit Theorem.
- **Counterexample:** With heavy-tailed distributions, the sample moment matrices may not converge properly, invalidating asymptotic inference.

4.2.3.3 Formal Statement and Derivation We define the fundamental orthogonal projectors in the Euclidean space $\mathbb{R}^n$:

- Projection matrix onto Col($\mathbf{Z}$):
$$\mathbf{P}_Z = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}' \tag{4.37}$$
- Complementary annihilator matrix:
$$\mathbf{M}_Z = \mathbf{I}_n - \mathbf{P}_Z \tag{4.38}$$

First Stage: Project the matrix of endogenous regressors $\mathbf{X}$ onto the exogenous space generated by $\mathbf{Z}$:

$$\hat{\mathbf{X}} = \mathbf{P}_Z\mathbf{X} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X} \tag{4.39}$$

The orthogonal decomposition ensures that $\mathbf{X} = \hat{\mathbf{X}} + \mathbf{M}_Z\mathbf{X}$, where $\hat{\mathbf{X}}$ represents the clean exogenous portion of the regressors, and $\mathbf{M}_Z\mathbf{X}$ captures the discarded endogenous variation.

Second Stage: Perform the regression of $\mathbf{y}$ on $\hat{\mathbf{X}}$ via OLS:

$$\hat{\beta}_{2SLS} = (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\mathbf{y} \tag{4.40}$$

Substituting (4.39) into (4.40) and using the idempotence and symmetry of $\mathbf{P}_Z$ ($\mathbf{P}'_Z\mathbf{P}_Z = \mathbf{P}_Z$):

$$\hat{\beta}_{2SLS} = (\mathbf{X}'\mathbf{P}'_Z\mathbf{P}_Z\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}'_Z\mathbf{y} = (\mathbf{X}'\mathbf{P}_Z\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_Z\mathbf{y} \tag{4.41}$$

Expanding the projection matrix expression yields the traditional closed form of the 2SLS estimator:

$$\boxed{\hat{\beta}_{2SLS} = \left[\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}\right]^{-1}\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}} \tag{4.42}$$

4.2.3.4 Asymptotic Properties of 2SLS Under the instrument validity conditions (Assumptions 4.13, 4.14, and 4.15) and asymptotic regularity, the 2SLS estimator satisfies:

1. **Consistency:** $\hat{\beta}_{2SLS} \xrightarrow{p} \beta$
2. **Asymptotic Normality:**
$$\sqrt{n}(\hat{\beta}_{2SLS} - \beta) \xrightarrow{d} N\left(\mathbf{0},\, \sigma^2\left(\mathbf{Q}_{XZ}\mathbf{Q}_{ZZ}^{-1}\mathbf{Q}_{ZX}\right)^{-1}\right) \tag{4.43}$$
where $\mathbf{Q}_{XZ} = E[\mathbf{x}_i\mathbf{z}_i']$, $\mathbf{Q}_{ZZ} = E[\mathbf{z}_i\mathbf{z}_i']$, and $\mathbf{Q}_{ZX} = \mathbf{Q}'_{XZ}$.
3. **Variance Estimator (under homoskedasticity):**
$$\widehat{\mathrm{Var}}(\hat{\beta}_{2SLS}) = \hat{\sigma}^2\left(\mathbf{X}'\mathbf{P}_Z\mathbf{X}\right)^{-1}, \quad \hat{\sigma}^2 = \frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{n-k} \tag{4.44}$$
4. **Robust Variance to Heteroskedasticity:**
$$\widehat{\mathrm{Var}}_{\mathrm{robust}}(\hat{\beta}_{2SLS}) = (\mathbf{X}'\mathbf{P}_Z\mathbf{X})^{-1}\left(\sum_{i=1}^{n}\hat{u}_i^2\hat{\mathbf{x}}_i\hat{\mathbf{x}}_i'\right)(\mathbf{X}'\mathbf{P}_Z\mathbf{X})^{-1} \tag{4.45}$$

4.2.4 Relative Efficiency and the Cost of Instrumentation

4.2.4.1 Motivation Under the validity of the classical Gauss-Markov assumptions — where all original regressors are strictly exogenous — the OLS estimator is the Best Linear Unbiased Estimator (BLUE). Therefore, the substitution of $\mathbf{X}$ by its instrumented projection $\hat{\mathbf{X}}$ generates an inevitable loss of information, inflating the sample variance.

Geometric Interpretation:

In $\mathbb{R}^n$, $\text{Col}(\hat{\mathbf{X}})$ is a subspace of $\text{Col}(\mathbf{X})$ when $\mathbf{X}$ is exogenous (since $\hat{\mathbf{X}}$ is a linear combination of $\mathbf{X}$ through projection). Projecting $\mathbf{y}$ onto a smaller subspace reduces the explained variation and increases the uncertainty about the parameters. The smaller the correlation between $\mathbf{X}$ and $\mathbf{Z}$ (weak instruments), the smaller $\text{Col}(\hat{\mathbf{X}})$ and the greater the efficiency loss.

4.2.4.2 Assumptions for Efficiency Comparison **Assumption 4.19: Exogeneity of X (Benchmark)**

- **Role:** Establishes the counterfactual scenario where OLS is consistent, allowing variance comparison.
- **Counterexample:** If $\mathbf{X}$ is endogenous, OLS is inconsistent and variance comparison loses practical meaning.

Assumption 4.20: Conditional Homoskedasticity

- **Role:** Ensures that $\text{Var}(\mathbf{u}|\mathbf{X}, \mathbf{Z}) = \sigma^2\mathbf{I}_n$, simplifying variance expressions.
- **Counterexample:** Under heteroskedasticity, variance formulas require robust adjustments, and the ordering relationship may be more complex.

4.2.4.3 Formal Statement and Derivation **Theorem (Cost of Instrumentation).** Under the assumption of exogeneity of $\mathbf{X}$ and conditional homoskedasticity, the variance-covariance matrices satisfy the ordering:

$$\boxed{\text{Var}(\hat{\beta}_{2SLS}|\mathbf{X}, \mathbf{Z}) \succeq \text{Var}(\hat{\beta}_{OLS}|\mathbf{X})} \tag{4.46}$$

Proof:

Auxiliary Lemma: If $\mathbf{A}$ and $\mathbf{B}$ are symmetric and strictly positive definite matrices such that $\mathbf{A} \succeq \mathbf{B}$, then $\mathbf{B}^{-1} \succeq \mathbf{A}^{-1}$.

Step 1: Formulation of conditional variances

Under homoskedasticity ($\text{Var}(\mathbf{u}|\mathbf{X}, \mathbf{Z}) = \sigma^2\mathbf{I}_n$):

$$\text{Var}(\hat{\beta}_{OLS}|\mathbf{X}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \tag{4.47}$$

$$\text{Var}(\hat{\beta}_{2SLS}|\mathbf{X}, \mathbf{Z}) = \sigma^2(\mathbf{X}'\mathbf{P}_Z\mathbf{X})^{-1} \tag{4.48}$$

Step 2: Analysis of the difference of information matrices

We evaluate the difference matrix $\mathbf{\Delta}$ between the inverses of the variances (scaled by $1/\sigma^2$):

$$\mathbf{\Delta} = \mathbf{X}'\mathbf{X} - \mathbf{X}'\mathbf{P}_Z\mathbf{X} = \mathbf{X}'(\mathbf{I}_n - \mathbf{P}_Z)\mathbf{X} = \mathbf{X}'\mathbf{M}_Z\mathbf{X} \tag{4.49}$$

Step 3: Demonstration of the positive semidefinite character of $\mathbf{M}_Z$

The annihilator matrix $\mathbf{M}_Z$ is symmetric and idempotent ($\mathbf{M}_Z = \mathbf{M}_Z'\mathbf{M}_Z$). Taking an arbitrary non-zero vector $\mathbf{v} \in \mathbb{R}^n$:

$$\mathbf{v}'\mathbf{M}_Z\mathbf{v} = \mathbf{v}'\mathbf{M}_Z'\mathbf{M}_Z\mathbf{v} = (\mathbf{M}_Z\mathbf{v})'(\mathbf{M}_Z\mathbf{v}) = \|\mathbf{M}_Z\mathbf{v}\|^2 \geq 0 \tag{4.50}$$

Since the quadratic form is non-negative for any vector $\mathbf{v}$, we conclude that $\mathbf{M}_Z \succeq \mathbf{0}$. Consequently:

$$\mathbf{X}'\mathbf{M}_Z\mathbf{X} \succeq \mathbf{0} \quad \Longrightarrow \quad \mathbf{X}'\mathbf{X} \succeq \mathbf{X}'\mathbf{P}_Z\mathbf{X} \tag{4.51}$$

Step 4: Inversion and conclusion

Invoking the Auxiliary Lemma on (4.51):

$$(\mathbf{X}'\mathbf{P}_Z\mathbf{X})^{-1} \succeq (\mathbf{X}'\mathbf{X})^{-1} \tag{4.52}$$

Multiplying by the scalar $\sigma^2 > 0$:

$$\sigma^2(\mathbf{X}'\mathbf{P}_Z\mathbf{X})^{-1} \succeq \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \quad \Longrightarrow \quad \mathrm{Var}(\hat{\beta}_{2SLS}) \succeq \mathrm{Var}(\hat{\beta}_{OLS}) \tag{4.53}$$

□

4.2.4.4 Practical Implications The efficiency cost of 2SLS relative to OLS is greater when:

1. **The correlation between X and Z is smaller** (weak instruments): $\mathrm{Col}(\hat{\mathbf{X}})$ shrinks, reducing the information available for estimation.
2. **The number of redundant instruments is larger**: irrelevant instruments add noise without increasing the exogenous variation of $\mathbf{X}$.
3. **The sample size n is smaller**: in small samples, the loss of degrees of freedom in the first stage becomes more severe.

This *trade-off* between **consistency** (2SLS) and **efficiency** (OLS) is the core of the choice between estimators: the researcher accepts greater sample variance in exchange for eliminating endogeneity bias.

4.2.5 Asymptotic Properties of the IV Estimator

4.2.5.1 Motivation To overcome finite-sample pathologies — such as the absence of defined moments under weak identification — statistical inference relies strictly on the asymptotic limiting behavior as $n \to \infty$.

4.2.5.2 Asymptotic Assumptions **Assumption 4.21: Ergodically Random Sampling**

- **Role:** The observations of the data vector $(\mathbf{y}_i, \mathbf{x}_i, \mathbf{z}_i)$ are independent and identically distributed (i.i.d.) with finite fourth-order moments, enabling the application of the Weak Law of Large Numbers and the Central Limit Theorem.
- **Counterexample:** In time series with dependence or non-stationarity, sample means may not converge to population expectations.

Assumption 4.22: Asymptotic Exogeneity of the Instrument

$$\mathrm{plim}\left(\frac{1}{n}\mathbf{Z}'\mathbf{u}\right) = E[\mathbf{z}_i u_i] = \mathbf{0} \tag{4.54}$$

- **Role:** Fundamental condition for the consistency of the IV estimator: the population correlation between instruments and error is zero.
- **Counterexample:** If $E[\mathbf{z}_i u_i] \neq \mathbf{0}$, the IV estimator is inconsistent.

Assumption 4.23: Asymptotic Relevance of the Instrument

$$\text{plim}\left(\frac{1}{n}\mathbf{Z}'\mathbf{X}\right) = E[\mathbf{z}_i\mathbf{x}_i'] = \mathbf{\Sigma}_{ZX} \tag{4.55}$$

where $\mathbf{\Sigma}_{ZX}$ is a finite population coefficient matrix with full rank equal to k.

- **Role:** Ensures that the instruments are sufficiently correlated with the endogenous regressors to identify the parameters.
- **Counterexample:** If $\text{rank}(\mathbf{\Sigma}_{ZX}) < k$, the model is underidentified and the estimator does not converge.

4.2.5.3 Formal Statement and Derivation of Consistency **Theorem (Consistency of IV).** Under Assumptions 4.21 to 4.23, the classical instrumental variables estimator is consistent in probability:

$$\boxed{\hat{\beta}_{IV} \xrightarrow{p} \beta} \tag{4.56}$$

Proof:

Starting from the stochastic form derived in (4.30), we rescale the sample operators by dividing by the sample size n:

$$\hat{\beta}_{IV} = \beta + \left(\frac{\mathbf{Z}'\mathbf{X}}{n}\right)^{-1}\left(\frac{\mathbf{Z}'\mathbf{u}}{n}\right) \tag{4.57}$$

We apply the plim operator to both sides. By Slutsky's Theorem, since the inverse function is continuous over non-singular matrices:

$$\text{plim}(\hat{\beta}_{IV}) = \beta + \left[\text{plim}\left(\frac{\mathbf{Z}'\mathbf{X}}{n}\right)\right]^{-1} \cdot \text{plim}\left(\frac{\mathbf{Z}'\mathbf{u}}{n}\right) \tag{4.58}$$

Invoking the Weak Law of Large Numbers:

1. By asymptotic relevance (Assumption 4.23): $\text{plim}\left(\frac{\mathbf{Z}'\mathbf{X}}{n}\right) = \mathbf{\Sigma}_{ZX}$ (non-singular).
2. By asymptotic exogeneity (Assumption 4.22): $\text{plim}\left(\frac{\mathbf{Z}'\mathbf{u}}{n}\right) = \mathbf{0}$.

Substituting:

$$\text{plim}(\hat{\beta}_{IV}) = \beta + \mathbf{\Sigma}_{ZX}^{-1} \cdot \mathbf{0} = \beta \tag{4.59}$$

□

4.2.5.4 Asymptotic Normality **Theorem (Asymptotic Normality of IV).** Under Assumptions 4.21 to 4.23 and assuming population homoskedasticity, the limiting distribution of the estimation error scaled by $\sqrt{n}$ satisfies:

$$\boxed{\sqrt{n}(\hat{\beta}_{IV} - \beta) \xrightarrow{d} N(\mathbf{0}, \mathbf{V})} \tag{4.60}$$

where the asymptotic variance matrix is defined by:

$$\mathbf{V} = \sigma^2\mathbf{\Sigma}_{ZX}^{-1}\mathbf{\Sigma}_{ZZ}\mathbf{\Sigma}_{XZ}^{-1} \tag{4.61}$$

with $\mathbf{\Sigma}_{ZZ} = E[\mathbf{z}_i\mathbf{z}_i']$ and $\mathbf{\Sigma}_{XZ} = \mathbf{\Sigma}_{ZX}'$.

Proof:

Subtracting β from both sides of (4.57) and scaling by $\sqrt{n}$:

$$\sqrt{n}(\hat{\beta}_{IV} - \beta) = \left(\frac{\mathbf{Z}'\mathbf{X}}{n}\right)^{-1} \left(\frac{\mathbf{Z}'\mathbf{u}}{\sqrt{n}}\right) \tag{4.62}$$

We analyze the limiting behavior of each block:

1. By the Continuous Mapping Theorem: $\left(\frac{\mathbf{Z}'\mathbf{X}}{n}\right)^{-1} \xrightarrow{p} \mathbf{\Sigma}_{ZX}^{-1}$.
2. Define $\mathbf{w}_i = \mathbf{z}_i u_i$. By exogeneity, $E[\mathbf{w}_i] = \mathbf{0}$. Under conditional homoskedasticity:

$$\mathbf{\Omega} = E[\mathbf{w}_i\mathbf{w}_i'] = E[u_i^2\mathbf{z}_i\mathbf{z}_i'] = \sigma^2 E[\mathbf{z}_i\mathbf{z}_i'] = \sigma^2\mathbf{\Sigma}_{ZZ} \tag{4.63}$$

By the Lindeberg-Lévy Central Limit Theorem:

$$\frac{\mathbf{Z}'\mathbf{u}}{\sqrt{n}} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{z}_i u_i \xrightarrow{d} N(\mathbf{0},\, \sigma^2\mathbf{\Sigma}_{ZZ}) \tag{4.64}$$

Combining the limits via Slutsky's Lemma:

$$\sqrt{n}(\hat{\beta}_{IV} - \beta) \xrightarrow{d} \mathbf{\Sigma}_{ZX}^{-1} \cdot N(\mathbf{0},\, \sigma^2\mathbf{\Sigma}_{ZZ}) \tag{4.65}$$

By the properties of linear transformations of Gaussian vectors:

$$\sqrt{n}(\hat{\beta}_{IV} - \beta) \xrightarrow{d} N\left(\mathbf{0},\, \sigma^2\mathbf{\Sigma}_{ZX}^{-1}\mathbf{\Sigma}_{ZZ}\mathbf{\Sigma}_{XZ}^{-1}\right) \tag{4.66}$$

□

4.2.6 Hypothesis Testing and Asymptotic Inference

4.2.6.1 Motivation The formulation of asymptotic test statistics aims to formally quantify the statistical uncertainty around the point estimator. It assesses whether the distance between the sample estimate and the restriction vector arises solely from sampling fluctuations or reflects a population discrepancy.

4.2.6.2 Assumptions of the Wald Test **Assumption 4.24: Independent Linear Restrictions**

- **Role:** Defines q simultaneous linear conditions with $\mathbf{R} \in \mathbb{R}^{q\times k}$ of full row rank ($\text{rank}(\mathbf{R}) = q$).
- **Counterexample:** If $\text{rank}(\mathbf{R}) < q$, the restrictions are redundant or contradictory, making the covariance matrix singular and the statistic undefined.

4.2.6.3 Formal Statement and Derivation Consider the multivariate linear null hypothesis:

$$H_0 : \mathbf{R}\beta = \mathbf{r} \tag{4.67}$$

where $\mathbf{R} \in \mathbb{R}^{q\times k}$ with $\text{rank}(\mathbf{R}) = q$, and $\mathbf{r} \in \mathbb{R}^q$.

Theorem (Wald Statistic). Under the asymptotic regularity conditions of the IV estimator, the quadratic Wald statistic:

$$W = (\mathbf{R}\hat{\beta}_{IV} - \mathbf{r})' \left[\mathbf{R}\left(\widehat{\text{Avar}}(\hat{\beta}_{IV})\right)\mathbf{R}'\right]^{-1} (\mathbf{R}\hat{\beta}_{IV} - \mathbf{r}) \tag{4.68}$$

converges asymptotically in distribution to a Chi-square variable with q degrees of freedom under the null hypothesis:

$$\boxed{W \xrightarrow{d} \chi^2_q} \tag{4.69}$$

Proof:

Theoretical Lemma: If $\mathbf{v} \sim N(\mathbf{0}, \mathbf{\Sigma})$ with $\mathbf{\Sigma}$ non-singular of order q, then $\mathbf{v}'\mathbf{\Sigma}^{-1}\mathbf{v} \sim \chi^2_q$.

Step 1: Limiting behavior under the null hypothesis

By the asymptotic normality theorem (4.60):

$$\sqrt{n}(\hat{\beta}_{IV} - \beta) \xrightarrow{d} N(\mathbf{0}, \mathbf{V}) \tag{4.70}$$

Premultiplying by $\mathbf{R}$:

$$\sqrt{n}(\mathbf{R}\hat{\beta}_{IV} - \mathbf{R}\beta) \xrightarrow{d} N(\mathbf{0}, \mathbf{RVR}') \tag{4.71}$$

Under H_0, substitute $\mathbf{R}\beta = \mathbf{r}$:

$$\sqrt{n}(\mathbf{R}\hat{\beta}_{IV} - \mathbf{r}) \xrightarrow{d} N(\mathbf{0}, \mathbf{RVR}') \tag{4.72}$$

Step 2: Consistent estimator of the asymptotic variance

We define the consistent estimator:

$$\hat{\mathbf{V}} = \hat{\sigma}^2 \left(\frac{\mathbf{Z}'\mathbf{X}}{n}\right)^{-1} \left(\frac{\mathbf{Z}'\mathbf{Z}}{n}\right) \left(\frac{\mathbf{X}'\mathbf{Z}}{n}\right)^{-1} \tag{4.73}$$

where $\hat{\sigma}^2 = \frac{1}{n}(\mathbf{y} - \mathbf{X}\hat{\beta}_{IV})'(\mathbf{y} - \mathbf{X}\hat{\beta}_{IV}) \xrightarrow{p} \sigma^2$.

Since $\hat{\mathbf{V}} \xrightarrow{p} \mathbf{V}$, the Continuous Mapping Theorem ensures:

$$\left[\mathbf{R}\hat{\mathbf{V}}\mathbf{R}'\right]^{-1} \xrightarrow{p} [\mathbf{RVR}']^{-1} \tag{4.74}$$

Step 3: Construction of the quadratic form

Rewriting (4.68) to highlight the $\sqrt{n}$ factors:

$$W = \left[\sqrt{n}(\mathbf{R}\hat{\beta}_{IV} - \mathbf{r})\right]' \left[\mathbf{R}\hat{\mathbf{V}}\mathbf{R}'\right]^{-1} \left[\sqrt{n}(\mathbf{R}\hat{\beta}_{IV} - \mathbf{r})\right] \tag{4.75}$$

Substituting (4.72) and (4.74), Slutsky's Lemma guarantees:

$$W \xrightarrow{d} \mathbf{v}' \left[\mathbf{RVR}'\right]^{-1} \mathbf{v}, \qquad \mathbf{v} \sim N(\mathbf{0}, \mathbf{RVR}') \tag{4.76}$$

Invoking the Theoretical Lemma:

$$W \xrightarrow{d} \chi^2_q \tag{4.77}$$

□

4.2.7 Instrument Validation: The Sargan Overidentification Test

4.2.7.1 Motivation The Sargan test (or overidentifying restrictions test) is applicable exclusively to the scenario where the model has more instruments than endogenous regressors ($L > k$). When the model is exactly identified ($L = k$), the sample residuals are orthogonal to the instruments by algebraic construction, making any test of hypothesis validation impossible.

In the overidentified scenario, the 2SLS estimator selects a linear combination of the instruments. Therefore, it becomes possible to test whether the null hypothesis of full exogeneity of the instrument block is statistically sustainable. The test principle is based on projecting the residuals of the estimated model onto the full matrix of available instruments $\mathbf{Z}$. If the instruments are indeed exogenous, the explanatory power of this auxiliary regression should be statistically close to zero.

4.2.7.2 Assumptions of the Sargan Test **Assumption 4.25: Overidentification ($L > k$)**

- **Role:** Necessary condition for the applicability of the test. Requires that there be more instruments than endogenous regressors.
- **Counterexample:** If $L = k$, the test is degenerate and cannot be computed.

Assumption 4.26: Joint Exogeneity of Instruments (H_0)

- **Role:** Establishes that all instruments are orthogonal to the structural error: $E[\mathbf{z}_i u_i] = \mathbf{0}$ for all i.
- **Counterexample:** If at least one instrument is endogenous, the Sargan statistic detects the violation and rejects H_0.

4.2.7.3 Computation Algorithm and Test Statistic The Sargan test is operationalized by the following algorithmic protocol:

1. Estimate the structural model via 2SLS and extract the sample residual vector:
$$\hat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\hat{\beta}_{2SLS} \tag{4.78}$$
2. Run an auxiliary regression of the residuals on the full instrument matrix $\mathbf{Z}$ via OLS:
$$\hat{\mathbf{u}} = \mathbf{Z}\gamma + \epsilon \tag{4.79}$$
3. Compute the coefficient of determination (R_u^2) of this auxiliary projection:
$$R_u^2 = \frac{\hat{\mathbf{u}}'\mathbf{P}_Z\hat{\mathbf{u}}}{\hat{\mathbf{u}}'\hat{\mathbf{u}}} \tag{4.80}$$
4. The Sargan test statistic (S) is obtained by multiplying the sample size by the coefficient of determination:
$$S = nR_u^2 = \frac{\hat{\mathbf{u}}'\mathbf{P}_Z\hat{\mathbf{u}}}{\hat{\sigma}^2} \tag{4.81}$$
where $\hat{\sigma}^2 = \hat{\mathbf{u}}'\hat{\mathbf{u}}/n$ is the estimator of the variance of the auxiliary regression error.

4.2.7.4 Asymptotic Distribution of the Sargan Test **Theorem (Distribution of the Sargan Test).** Under the null hypothesis of total exogeneity of the instruments ($H_0 : E[\mathbf{z}_i u_i] = \mathbf{0}$), the Sargan statistic converges asymptotically in distribution to a Chi-square variable whose degrees of freedom correspond exactly to the number of overidentifying restrictions (excess instruments):

$$\boxed{S \xrightarrow{d} \chi^2_{L-k}} \tag{4.82}$$

Interpretation: Rejection of the null hypothesis ($S > \chi^2_{\text{critical}}$) indicates that at least one of the instruments fails the asymptotic exogeneity condition (being correlated with the structural error term) or that the structural model has been incorrectly specified due to omitted variables.

4.2.7.5 Limitations and Practical Considerations

1. **Power of the test:** The Sargan test has low power to detect weakly endogenous instruments in small samples.
2. **Correct model specification:** Rejection may arise from either invalid instruments or misspecification of the structural equation (e.g., omission of relevant variables).
3. **Robust version to heteroskedasticity:** The Hansen test (J-statistic) is a generalization of the Sargan test that does not require homoskedasticity, being preferable in contexts with heteroskedastic errors.
4. **Interpretation of non-rejection:** Non-rejection of the null hypothesis **does not** prove that the instruments are exogenous; it only indicates that the test did not find sufficient statistical evidence to reject the joint validity of the instruments.

Chapter 5: Generalized Method of Moments (GMM) Estimator

5.1 Construction and Properties of the Estimator

5.1.1 Motivation and Geometric Interpretation

5.1.1.1 The Overidentification Problem The **Generalized Method of Moments (GMM)**, proposed by Lars Hansen (1982), represents a fundamental milestone in modern econometrics by solving the **overidentification** problem in parametric models. In the classical method of moments (Pearson, 1894), the researcher seeks an estimator that satisfies a system of equations where the number of moment conditions (L) is exactly equal to the number of unknown parameters (K). In such **just-identified** systems, there is typically a unique solution that zeros out the sample moment conditions.

However, economic theory or statistical properties often provide more restrictions (moments) than strictly necessary to identify the parameters ($L > K$). In these cases, the system of equations becomes **overdetermined** and, due to inherent sample variability, it is impossible to find a single parameter vector that satisfies all equations simultaneously. GMM resolves this difficulty by transforming a root-finding problem into an **optimization problem**, minimizing a metric of "distance" between the sample moments and the null vector.

5.1.1.2 Geometric Interpretation To develop intuition, consider the following geometric scenario:

- In $\mathbb{R}^2$, we seek a point θ that should lie at the intersection of three lines (representing three moment conditions).
- In a finite sample, these lines will rarely intersect at a single point due to stochastic noise, forming a small triangle instead of a common vertex.
- Geometrically, the GMM estimator locates the point $\hat{\theta}$ that minimizes the weighted sum of squared distances to these lines.

The **weighting matrix W** acts as a metric that deforms the space, assigning greater weight to moment conditions that have smaller variance (are more "reliable"). In $\mathbb{R}^3$, GMM can be visualized as the search for the point in a subspace that minimizes the orthogonal distance to a series of hyperplanes defined by the moment conditions.

5.1.2 Fundamental Assumptions of GMM

For the formal construction of the estimator, we define:

- $\theta \in \Theta \subseteq \mathbb{R}^K$: vector of parameters
- $\mathbf{w}_i$: vector of observed data
- $\mathbf{g}(\mathbf{w}_i, \theta)$: moment function of dimension $L \times 1$

Assumption 5.1: Population Orthogonality (Moment Condition)

- **Role:** Constitutes the theoretical foundation of the model. It states that, at the true parameter value θ_0, the mathematical expectation of the moment function is zero:

$$E[\mathbf{g}(\mathbf{w}_i, \theta_0)] = \mathbf{0} \tag{5.1}$$

- **Counterexample:** In a linear regression model $\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$ where we omit a variable $\mathbf{Z}$ correlated with $\mathbf{X}$, the error $\mathbf{u}$ will have $E[\mathbf{u}|\mathbf{X}] \neq \mathbf{0}$. In this case, the moment condition $E[\mathbf{X}'\mathbf{u}] = \mathbf{0}$ fails and any estimator based on it will be inconsistent.

Assumption 5.2: Identification (Rank Condition)

- **Role:** Ensures that the moment condition provides sufficient information to distinguish θ_0 from any other value in the parameter space. Mathematically, it requires that the matrix of expected derivatives has full rank K:

$$\mathbf{G} = E\left[\frac{\partial \mathbf{g}(\mathbf{w}_i, \theta_0)}{\partial \theta'}\right], \quad \text{rank}(\mathbf{G}) = K \tag{5.2}$$

- **Counterexample:** If we try to estimate the return to education using an instrument that has no correlation with schooling, the rank of the covariance matrix between the instrument and the regressor will be zero. The parameter cannot be located because the objective function will be "flat" with respect to it.

Assumption 5.3: Positive Definite Weighting Matrix

- **Role:** Ensures that the quadratic objective function is strictly convex (in the linear case) and has a well-defined global minimum.
- **Counterexample:** If we use a zero or singular matrix $\mathbf{W}$, the objective function would not penalize deviations in certain directions of the moment conditions, allowing infinitely many solutions or preventing numerical convergence of the optimizer.

5.1.3 Formal Formulation of the Estimator

5.1.3.1 General Definition Let $\{\mathbf{w}_i\}_{i=1}^n$ be a sequence of independent and identically distributed (i.i.d.) random vectors belonging to a probability space $(\Omega, \mathcal{F}, P)$. Let $\theta_0 \in \Theta$ be a vector of parameters in a compact parameter space $\Theta \subset \mathbb{R}^K$. We define a measurable moment function $\mathbf{g} : \mathbb{R}^M \times \Theta \to \mathbb{R}^L$ such that $L \geq K$.

The GMM estimator $\hat{\theta}_n$ is defined as the argument that minimizes the quadratic objective function:

$$\boxed{Q_n(\theta) = \bar{\mathbf{g}}_n(\theta)' \mathbf{W}_n \bar{\mathbf{g}}_n(\theta)} \tag{5.3}$$

where:

$$\bar{\mathbf{g}}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{g}(\mathbf{w}_i, \theta) \tag{5.4}$$

and $\mathbf{W}_n$ is a stochastic weighting matrix such that $\mathbf{W}_n \xrightarrow{p} \mathbf{W}$, with $\mathbf{W}$ being symmetric and positive definite.

5.1.3.2 Derivation of the Estimator for the Linear Case Consider the linear case where:

$$\mathbf{g}(\mathbf{w}_i, \theta) = \mathbf{z}_i(y_i - \mathbf{x}_i'\theta) \tag{5.5}$$

with $\mathbf{z}_i$ an $L \times 1$ vector of instruments and $\mathbf{x}_i$ a $K \times 1$ vector of regressors.

Step 1: Definition of the Sample Moment

In matrix notation, defining $\mathbf{y}$ as the $n \times 1$ vector, $\mathbf{X}$ as the $n \times K$ matrix, and $\mathbf{Z}$ as the $n \times L$ instrument matrix:

$$\bar{\mathbf{g}}_n(\theta) = \frac{1}{n}\mathbf{Z}'(\mathbf{y} - \mathbf{X}\theta) \tag{5.6}$$

Step 2: Substitution into the Objective Function

Substituting (5.6) into (5.3):

$$Q_n(\theta) = \left[\frac{1}{n}\mathbf{Z}'(\mathbf{y} - \mathbf{X}\theta)\right]' \mathbf{W}_n \left[\frac{1}{n}\mathbf{Z}'(\mathbf{y} - \mathbf{X}\theta)\right] \tag{5.7}$$

Step 3: First Order Condition (FOC)

Differentiating $Q_n(\theta)$ with respect to θ and setting it to the null vector:

$$\frac{\partial Q_n(\theta)}{\partial \theta} = -\frac{2}{n^2}\mathbf{X}'\mathbf{Z}\mathbf{W}_n\mathbf{Z}'\mathbf{y} + \frac{2}{n^2}\mathbf{X}'\mathbf{Z}\mathbf{W}_n\mathbf{Z}'\mathbf{X}\theta = \mathbf{0} \tag{5.8}$$

Step 4: Isolating the Estimator

$$\mathbf{X}'\mathbf{Z}\mathbf{W}_n\mathbf{Z}'\mathbf{X}\hat{\theta}_n = \mathbf{X}'\mathbf{Z}\mathbf{W}_n\mathbf{Z}'\mathbf{y} \tag{5.9}$$

Assuming that $(\mathbf{X}'\mathbf{Z}\mathbf{W}_n\mathbf{Z}'\mathbf{X})$ is non-singular (Assumption 5.2):

$$\boxed{\hat{\theta}_n = (\mathbf{X}'\mathbf{Z}\mathbf{W}_n\mathbf{Z}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\mathbf{W}_n\mathbf{Z}'\mathbf{y}} \tag{5.10}$$

5.1.3.3 Role of Assumptions in the Derivation The above derivation highlights the critical role of each assumption:

1. **Identification (Assumption 5.2):** Was indispensable in Step 4. If $(\mathbf{X}'\mathbf{Z}\mathbf{W}_n\mathbf{Z}'\mathbf{X})$ is singular (rank $< K$), the inverse does not exist and the estimator cannot be uniquely determined.
2. **Orthogonality (Assumption 5.1):** Ensures that, when substituting $\mathbf{y} = \mathbf{X}\theta_0 + \mathbf{u}$ into (5.10), the term associated with the error will collapse asymptotically, validating the consistency of the estimator:
$$\operatorname{plim} \hat{\theta}_n = \theta_0 \tag{5.11}$$
3. **Positive Definite Matrix (Assumption 5.3):** Ensures that the FOC (5.8) characterizes a global minimum (not just a saddle point or maximum), guaranteeing the convexity of the objective function.

5.1.4 Moments of the GMM Estimator

5.1.4.1 Motivation While the asymptotic properties of the GMM estimator (consistency and asymptotic normality) describe its limiting behavior as $n \to \infty$, the **conditional expectation and variance** allow us to analyze the statistical properties of the estimator in **finite samples**.

The central problem is **finite-sample bias**. In Instrumental Variables (IV) and overidentified GMM models, although the estimator is consistent, it often exhibits bias in small samples if the instruments are "weak" or if there is strong correlation between the endogenous regressors and the structural error.

Geometric Interpretation in $\mathbb{R}^K$:

Geometrically, the conditional expectation locates the "center of mass" of the distribution of the estimator $\hat{\beta}$ in the parameter space $\mathbb{R}^K$, given a fixed set of instruments $\mathbf{Z}$. If the estimator is unbiased, this center

coincides with β_0. The conditional variance defines the **concentration ellipsoid** around this expectation; the volume and orientation of this ellipsoid are determined by the curvature of the objective function and the dispersion of the errors, weighted by the metric structured in $\mathbf{W}$.

5.1.4.2 Additional Assumptions Assumption 5.4: Conditional Strict Exogeneity

- **Role:** Ensures that the error vector $\mathbf{u}$ has zero conditional mean for any realization of the instruments and regressors, allowing the expectation operator to pass through the bilinear form of the estimator. Mathematically: $E[\mathbf{u}|\mathbf{Z}, \mathbf{X}] = \mathbf{0}$.
- **Counterexample:** In models with predetermined regressors (such as lags of the dependent variable y_{t-1}), $E[u_t|y_{t-1}] = 0$ may hold contemporaneously, but $E[u_{t-1}|y_{t-1}] \neq 0$. This implies that $E[\hat{\beta}|\mathbf{X}] \neq \beta_0$, breaking the unbiasedness property.

Assumption 5.5: Conditional Homoskedasticity (Sphericity)

- **Role:** Simplifies the structure of the conditional variance-covariance matrix of the estimator, assuming that $\text{Var}(\mathbf{u}|\mathbf{Z}, \mathbf{X}) = \sigma^2\mathbf{I}_n$. Without this, the conditional variance assumes the expanded "sandwich" form.
- **Counterexample:** If $\text{Var}(u_i|\mathbf{z}_i) = \sigma^2 z_i^2$, the precision of the estimator will vary with the magnitude of $\mathbf{Z}$. Ignoring this heteroskedasticity leads to incorrect standard errors and invalid inference in finite samples.

Assumption 5.6: Non-Stochasticity of W (or Conditioning)

- **Role:** For the exact derivation of the expectation in finite samples, the weighting matrix $\mathbf{W}$ must be treated as fixed or dependent only on $\mathbf{Z}$.
- **Counterexample:** If $\mathbf{W}$ is estimated iteratively in two steps ($\mathbf{W} = \hat{\mathbf{\Omega}}^{-1}$), it becomes a function of $\hat{\mathbf{u}}$, which depends on $\mathbf{y}$. This creates a complex stochastic dependence between $\mathbf{W}$ and $\mathbf{u}$, making linear factorization of the expectation impossible.

5.1.4.3 Formal Derivation Statement: Consider the linear model $\mathbf{y} = \mathbf{X}\beta_0 + \mathbf{u}$. Let $\hat{\beta}_{GMM}$ be the estimator defined in (5.10). Under the assumptions of strict exogeneity (Assumption 5.4), rank identification ($\text{rank}(\mathbf{X}'\mathbf{Z}\mathbf{W}\mathbf{Z}'\mathbf{X}) = K$), and non-stochasticity of $\mathbf{W}$ conditioned on $\mathbf{Z}$ (Assumption 5.6), the estimator takes the form:

$$\hat{\beta} = (\mathbf{X}'\mathbf{Z}\mathbf{W}\mathbf{Z}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\mathbf{W}\mathbf{Z}'\mathbf{y} \tag{5.12}$$

Theorem (Conditional Expectation). Under Assumption 5.4 (strict exogeneity):

$$\boxed{E[\hat{\beta}|\mathbf{Z}, \mathbf{X}] = \beta_0} \tag{5.13}$$

Proof:

Substituting $\mathbf{y} = \mathbf{X}\beta_0 + \mathbf{u}$ into (5.12):

$$\hat{\beta} = \beta_0 + (\mathbf{X}'\mathbf{Z}\mathbf{W}\mathbf{Z}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\mathbf{W}\mathbf{Z}'\mathbf{u} \tag{5.14}$$

Applying $E[\cdot|\mathbf{Z}, \mathbf{X}]$ and using linearity:

$$E[\hat{\beta}|\mathbf{Z}, \mathbf{X}] = \beta_0 + (\mathbf{X}'\mathbf{Z}\mathbf{W}\mathbf{Z}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\mathbf{W}\mathbf{Z}'E[\mathbf{u}|\mathbf{Z}, \mathbf{X}] \tag{5.15}$$

By Assumption 5.4, $E[\mathbf{u}|\mathbf{Z}, \mathbf{X}] = \mathbf{0}$:

$$E[\hat{\beta}|\mathbf{Z}, \mathbf{X}] = \beta_0 \tag{5.16}$$

□

Theorem (Conditional Variance). Under Assumption 5.5 (homoskedasticity), the conditional variance of the GMM estimator is:

$$\boxed{\text{Var}(\hat{\beta}|\mathbf{Z},\mathbf{X}) = \sigma^2(\mathbf{X}'\mathbf{Z}\mathbf{W}\mathbf{Z}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\mathbf{W}(\mathbf{Z}'\mathbf{Z})\mathbf{W}\mathbf{Z}'\mathbf{X}(\mathbf{X}'\mathbf{Z}\mathbf{W}\mathbf{Z}'\mathbf{X})^{-1}} \tag{5.17}$$

Proof:

By definition:

$$\text{Var}(\hat{\beta}|\mathbf{Z},\mathbf{X}) = E[(\hat{\beta}-\beta_0)(\hat{\beta}-\beta_0)'|\mathbf{Z},\mathbf{X}] \tag{5.18}$$

Defining $\mathbf{M} = (\mathbf{X}'\mathbf{Z}\mathbf{W}\mathbf{Z}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\mathbf{W}\mathbf{Z}'$:

$$\text{Var}(\hat{\beta}|\mathbf{Z},\mathbf{X}) = \mathbf{M}E[\mathbf{u}\mathbf{u}'|\mathbf{Z},\mathbf{X}]\mathbf{M}' \tag{5.19}$$

Under homoskedasticity (Assumption 5.5): $E[\mathbf{u}\mathbf{u}'|\mathbf{Z},\mathbf{X}] = \sigma^2\mathbf{I}_n$.

Substituting $\mathbf{M}$ and simplifying:

$$\text{Var}(\hat{\beta}|\mathbf{Z},\mathbf{X}) = \sigma^2(\mathbf{X}'\mathbf{Z}\mathbf{W}\mathbf{Z}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\mathbf{W}(\mathbf{Z}'\mathbf{Z})\mathbf{W}\mathbf{Z}'\mathbf{X}(\mathbf{X}'\mathbf{Z}\mathbf{W}\mathbf{Z}'\mathbf{X})^{-1} \tag{5.20}$$

□

Special Case: 2SLS Estimator

If we choose $\mathbf{W} = (\mathbf{Z}'\mathbf{Z})^{-1}$, the conditional variance simplifies to:

$$\boxed{\text{Var}(\hat{\beta}_{2SLS}|\mathbf{Z},\mathbf{X}) = \sigma^2(\mathbf{X}'\mathbf{P}_Z\mathbf{X})^{-1}} \tag{5.21}$$

where $\mathbf{P}_Z = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$ is the orthogonal projection matrix onto Col($\mathbf{Z}$).

5.1.4.4 Role of Assumptions in the Derivation

1. **Strict Exogeneity (Assumption 5.4):** Was indispensable in Step 2 of the expectation proof. If the error is conditionally correlated with the instruments or regressors, $E[\mathbf{u}|\mathbf{Z},\mathbf{X}] \neq \mathbf{0}$, generating bias in finite samples.
2. **Non-Stochasticity of W (Assumption 5.6):** Allowed $\mathbf{W}$ to be treated deterministically in the expectation operator. In the classical two-step GMM, this premise is violated, introducing additional bias in small samples.
3. **Homoskedasticity (Assumption 5.5):** Was used to simplify the conditional variance. Under heteroskedasticity, the general expression (5.17) requires the substitution of $\sigma^2\mathbf{I}_n$ by $\mathbf{\Omega}$:

$$\text{Var}(\hat{\beta}|\mathbf{Z},\mathbf{X}) = (\mathbf{X}'\mathbf{Z}\mathbf{W}\mathbf{Z}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\mathbf{W}\mathbf{Z}'\mathbf{\Omega}\mathbf{Z}\mathbf{W}\mathbf{Z}'\mathbf{X}(\mathbf{X}'\mathbf{Z}\mathbf{W}\mathbf{Z}'\mathbf{X})^{-1} \tag{5.22}$$

5.1.5 Relative Efficiency of the GMM Estimator

5.1.5.1 Motivation The question of **relative efficiency** in the Generalized Method of Moments (GMM) lies in the choice of the ideal "metric" to optimally combine the information contained in different moment conditions. When the model is overidentified ($L > K$), there are infinitely many ways to weight the sample moments to obtain a consistent estimator. However, not all choices of weights are equivalent in terms of asymptotic variance.

The central problem is to determine whether an estimator structured with a weight matrix $\mathbf{W}_1$ is statistically superior to another structured with $\mathbf{W}_2$. In multivariate statistics, this is formalized by the **extended Gauss-Markov efficiency**: an estimator is globally more efficient than another if the difference between their respective variance-covariance matrices is a **Positive Semidefinite (PSD)** matrix.

Geometric Interpretation: Projection in Hilbert Space

Geometrically, the search for the optimal GMM estimator can be visualized as an **orthogonal projection** in a Hilbert space deformed by the intrinsic variance of the moments. If we think of the moments as vectors in $\mathbb{R}^L$, the GMM estimator projects the data onto the subspace spanned by the derivatives of the moments (the Jacobian matrix $\mathbf{G}$). The choice of the weight matrix $\mathbf{W}$ defines the angle and metric of this projection. The "Optimal" estimator is the one that adopts the inverse of the moment covariance matrix ($\mathbf{\Lambda}^{-1}$) as the metric, performing an orthogonal projection in the statistical sense, minimizing the volume of the error ellipsoid. Any other choice of $\mathbf{W}$ results in an "oblique" projection, presenting necessarily larger or equal variance.

5.1.5.2 Assumptions for Efficiency **Assumption 5.7: Comparability of Variances (PSD Ordering)**

- **Role:** Allows establishing an unambiguous partial order between estimators. We define that $\text{Var}(\hat{\theta}_2) \succeq \text{Var}(\hat{\theta}_1)$ if, for any non-zero constant vector $\mathbf{c} \in \mathbb{R}^K$, the variance of the linear combination $\mathbf{c}'\hat{\theta}_2$ is greater than or equal to that of $\mathbf{c}'\hat{\theta}_1$.
- **Counterexample:** Without ordering based on the PSD property, we could have an estimator with smaller variance for the first parameter but larger for the second, preventing a global declaration of superiority between estimation methods.

Assumption 5.8: Invertibility of the Moment Covariance Matrix ($\mathbf{\Lambda}$)

- **Role:** Strictly necessary to define the optimal weight matrix $\mathbf{W}^* = \mathbf{\Lambda}^{-1}$. Ensures the absence of perfectly redundant moment conditions that would generate singularity in the matrix.
- **Counterexample:** If two moment conditions were perfectly correlated in the population, $\mathbf{\Lambda}$ would be singular and the weight assigned to that specific direction would be mathematically undefined, collapsing the determination of the efficient estimator.

5.1.5.3 Formal Statement and Derivation Let $\hat{\theta}_A$ be the consistent GMM estimator that uses a symmetric positive definite weight matrix $\mathbf{A}$, and $\hat{\theta}_{opt}$ be the estimator that adopts the optimal weight matrix $\mathbf{W}^* = \mathbf{\Lambda}^{-1}$. Their respective asymptotic variances are given by:

1. **Variance for $\mathbf{W} = \mathbf{A}$ ("Sandwich" Form):**

$$\mathbf{V}(\mathbf{A}) = (\mathbf{G}'\mathbf{A}\mathbf{G})^{-1}\mathbf{G}'\mathbf{A}\mathbf{\Lambda}\mathbf{A}\mathbf{G}(\mathbf{G}'\mathbf{A}\mathbf{G})^{-1} \tag{5.23}$$

2. **Variance for $\mathbf{W} = \mathbf{\Lambda}^{-1}$ (Optimal Form):**

$$\mathbf{V}(\mathbf{\Lambda}^{-1}) = (\mathbf{G}'\mathbf{\Lambda}^{-1}\mathbf{G})^{-1} \tag{5.24}$$

Theorem (Hansen, 1982). Under regularity conditions, the difference matrix $\mathbf{D} = \mathbf{V}(\mathbf{A}) - \mathbf{V}(\mathbf{\Lambda}^{-1})$ is Positive Semidefinite ($\mathbf{D} \succeq \mathbf{0}$) for any choice of $\mathbf{A}$.

$$\boxed{\mathbf{V}(\mathbf{A}) \succeq \mathbf{V}(\mathbf{\Lambda}^{-1})} \tag{5.25}$$

Proof:

We will prove that $\mathbf{V}(\mathbf{\Lambda}^{-1})^{-1} \succeq \mathbf{V}(\mathbf{A})^{-1}$, which is equivalent for positive definite matrices.

Step 1: Cholesky Decomposition and Change of Basis

Let $\mathbf{\Lambda} = \mathbf{\Lambda}^{1/2}\mathbf{\Lambda}^{1/2}$ be the Cholesky decomposition. We define:

$$\mathbf{H} = \mathbf{\Lambda}^{-1/2}\mathbf{G}, \qquad \mathbf{K} = \mathbf{\Lambda}^{1/2}\mathbf{A}\mathbf{G} \tag{5.26}$$

Step 2: Rewriting the Inverse Variances

The inverse of the optimal variance:

$$\mathbf{V}(\mathbf{\Lambda}^{-1})^{-1} = \mathbf{G}'\mathbf{\Lambda}^{-1}\mathbf{G} = \mathbf{H}'\mathbf{H} \tag{5.27}$$

The inverse of the "sandwich" variance:

$$\mathbf{V}(\mathbf{A})^{-1} = (\mathbf{G}'\mathbf{A}\mathbf{G})(\mathbf{G}'\mathbf{A}\mathbf{\Lambda}\mathbf{A}\mathbf{G})^{-1}(\mathbf{G}'\mathbf{A}\mathbf{G}) \tag{5.28}$$

Expanding with the auxiliary definitions:

$$\mathbf{G}'\mathbf{A}\mathbf{G} = \mathbf{H}'\mathbf{K}, \qquad \mathbf{G}'\mathbf{A}\mathbf{\Lambda}\mathbf{A}\mathbf{G} = \mathbf{K}'\mathbf{K} \tag{5.29}$$

Substituting into (5.28):

$$\mathbf{V}(\mathbf{A})^{-1} = \mathbf{H}'\mathbf{K}(\mathbf{K}'\mathbf{K})^{-1}\mathbf{K}'\mathbf{H} = \mathbf{H}'\mathbf{P}_K\mathbf{H} \tag{5.30}$$

where $\mathbf{P}_K = \mathbf{K}(\mathbf{K}'\mathbf{K})^{-1}\mathbf{K}'$ is the orthogonal projection matrix onto $\text{Col}(\mathbf{K})$.

Step 3: Analysis of the Matrix Difference

$$\mathbf{\Delta} = \mathbf{V}(\mathbf{\Lambda}^{-1})^{-1} - \mathbf{V}(\mathbf{A})^{-1} = \mathbf{H}'\mathbf{H} - \mathbf{H}'\mathbf{P}_K\mathbf{H} = \mathbf{H}'(\mathbf{I} - \mathbf{P}_K)\mathbf{H} \tag{5.31}$$

Step 4: Property of Positive Semidefiniteness

Let $\mathbf{M}_K = \mathbf{I} - \mathbf{P}_K$ be the annihilator matrix. By the property of orthogonal projections, $\mathbf{M}_K$ is symmetric and idempotent ($\mathbf{M}_K^2 = \mathbf{M}_K$), hence $\mathbf{M}_K \succeq \mathbf{0}$.

For any vector $\mathbf{z} \in \mathbb{R}^K$:

$$\mathbf{z}'\mathbf{\Delta}\mathbf{z} = \mathbf{z}'\mathbf{H}'\mathbf{M}_K\mathbf{H}\mathbf{z} = (\mathbf{H}\mathbf{z})'\mathbf{M}_K(\mathbf{H}\mathbf{z}) \geq 0 \tag{5.32}$$

Hence, $\mathbf{\Delta} \succeq \mathbf{0}$, which implies $\mathbf{V}(\mathbf{A}) \succeq \mathbf{V}(\mathbf{\Lambda}^{-1})$. □

5.1.5.4 Role of Assumptions in the Derivation The assumption of **Invertibility of $\mathbf{\Lambda}$ (Assumption 5.8)** was indispensable for performing the structural decomposition described in Step 1, enabling the construction of the complementary projection matrix $\mathbf{M}_K$. Without the validity of this premise, the geometric equivalence between statistical optimality and projection orthogonality would be broken, and the optimal estimator could not be uniquely defined.

5.1.6 Asymptotic Properties of the GMM Estimator

5.1.6.1 Motivation The **asymptotic properties** of the GMM estimator describe its limiting behavior as $n \to \infty$. While finite-sample analysis deals with the exact distribution — often unknown or analytically intractable for non-linear models — asymptotic theory provides standardized and tractable approximations for conducting valid statistical inference.

The central problem solved here is guaranteeing **consistency** and deriving a **stable limiting distribution**. Consistency ensures that, as the sample size grows, the estimator will converge in probability to θ_0. Asymptotic normality guarantees that the distribution of the estimation error scaled by $\sqrt{n}$ approaches a multivariate normal distribution, underpinning conventional hypothesis tests.

5.1.6.2 Assumptions for Asymptotic Properties **Assumption 5.9: Global Identification (Uniqueness)**

- **Role:** Ensures that θ_0 is the only point in the parameter space capable of zeroing the population moment conditions: $E[\mathbf{f}(\mathbf{w}_i, \theta)] = \mathbf{0} \iff \theta = \theta_0$.
- **Counterexample:** If there are multiple candidate values that simultaneously satisfy the population moment conditions, the objective function will exhibit multiple global minima and the estimator may oscillate indefinitely between them, breaking consistency.

Assumption 5.10: Stationarity and Ergodicity

- **Role:** Allows the application of the Law of Large Numbers, ensuring that the sample means of the moment conditions and their derivatives converge in probability to their population expectations.
- **Counterexample:** If the data follow non-stationary processes (such as random walks with unit roots), the sample means will not converge to finite constants, invalidating the premise that sample moments mimic the population structure.

Assumption 5.11: Compactness of the Parameter Space Θ

- **Role:** Indispensable topological condition to ensure **uniform convergence** of the sample objective function to the population one. Prevents the numerical optimizer from diverging to infinity.
- **Counterexample:** In an objective function that is openly decreasing toward the infinite boundaries of the parameter space, the optimization algorithm would fail to establish a stopping point, invalidating the formal proof of consistency.

Assumption 5.12: Existence of Second-Order Moments

- **Role:** Requires that $\mathbf{S} = E[\mathbf{f}(\mathbf{w}_i, \theta_0)\mathbf{f}(\mathbf{w}_i, \theta_0)']$ be finite and positive definite, enabling the application of the Central Limit Theorem.
- **Counterexample:** If the errors follow heavy-tailed distributions with infinite variance (such as the Cauchy distribution), the $\sqrt{n}$ normalization would fail and the estimator would exhibit a non-Gaussian limiting distribution.

5.1.6.3 Formal Statement and Derivation Let $\hat{\theta}_n = \text{argmin}_{\theta \in \Theta} Q_n(\theta)$, where $Q_n(\theta) = \mathbf{g}_n(\theta)' \mathbf{W}_n \mathbf{g}_n(\theta)$. We define $\mathbf{G}_n(\theta) = \partial \mathbf{g}_n(\theta)/\partial \theta'$ $(L \times K)$ and assume that its population limit $\mathbf{G}_0 = E[\partial \mathbf{f}(\mathbf{w}_i, \theta_0)/\partial \theta']$ has full rank K. Under regularity conditions, $\mathbf{W}_n \xrightarrow{p} \mathbf{W}$ (positive definite) and $\sqrt{n}\mathbf{g}_n(\theta_0) \xrightarrow{d} N(\mathbf{0}, \mathbf{S})$.

Theorem (Asymptotic Normality of GMM). The scaled estimation error of the GMM estimator satisfies:

$$\boxed{\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N\left(\mathbf{0},\, (\mathbf{G}_0'\mathbf{W}\mathbf{G}_0)^{-1}\mathbf{G}_0'\mathbf{W}\mathbf{S}\mathbf{W}\mathbf{G}_0(\mathbf{G}_0'\mathbf{W}\mathbf{G}_0)^{-1}\right)} \tag{5.33}$$

Proof:

Step 1: First Order Condition (FOC)

Assuming that $\hat{\theta}_n$ lies in the interior of the parameter space:

$$\frac{\partial Q_n(\hat{\theta}_n)}{\partial \theta} = 2\mathbf{G}_n(\hat{\theta}_n)'\mathbf{W}_n\mathbf{g}_n(\hat{\theta}_n) = \mathbf{0} \tag{5.34}$$

Step 2: Expansion via the Mean Value Theorem

We expand $\mathbf{g}_n(\hat{\theta}_n)$ around θ_0:

$$\mathbf{g}_n(\hat{\theta}_n) = \mathbf{g}_n(\theta_0) + \mathbf{G}_n(\bar{\theta})(\hat{\theta}_n - \theta_0) \tag{5.35}$$

where $\bar{\theta}$ lies on the segment between $\hat{\theta}_n$ and θ_0.

Step 3: Substitution and Scaling

Substituting (5.35) into (5.34) and multiplying by $\sqrt{n}$:

$$\mathbf{G}_n(\hat{\theta}_n)'\mathbf{W}_n\sqrt{n}\mathbf{g}_n(\theta_0) + \mathbf{G}_n(\hat{\theta}_n)'\mathbf{W}_n\mathbf{G}_n(\bar{\theta})\sqrt{n}(\hat{\theta}_n - \theta_0) = \mathbf{0} \tag{5.36}$$

Step 4: Isolating the Scaled Error

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -\left[\mathbf{G}_n(\hat{\theta}_n)'\mathbf{W}_n\mathbf{G}_n(\bar{\theta})\right]^{-1}\mathbf{G}_n(\hat{\theta}_n)'\mathbf{W}_n\sqrt{n}\mathbf{g}_n(\theta_0) \tag{5.37}$$

Step 5: Application of Stochastic Limits

- $\hat{\theta}_n \xrightarrow{p} \theta_0$ and $\bar{\theta} \xrightarrow{p} \theta_0$, hence $\mathbf{G}_n(\hat{\theta}_n) \xrightarrow{p} \mathbf{G}_0$ and $\mathbf{G}_n(\bar{\theta}) \xrightarrow{p} \mathbf{G}_0$.
- $\mathbf{W}_n \xrightarrow{p} \mathbf{W}$.
- $\sqrt{n}\mathbf{g}_n(\theta_0) \xrightarrow{d} N(\mathbf{0}, \mathbf{S})$.

By Slutsky's Theorem:

$$\left[\mathbf{G}_n(\hat{\theta}_n)'\mathbf{W}_n\mathbf{G}_n(\bar{\theta})\right]^{-1}\mathbf{G}_n(\hat{\theta}_n)'\mathbf{W}_n \xrightarrow{p} (\mathbf{G}_0'\mathbf{W}\mathbf{G}_0)^{-1}\mathbf{G}_0'\mathbf{W} \tag{5.38}$$

Step 6: Consolidation of the Limiting Variance

Combining (5.38) with the CLT via continuous mapping:

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N\left(\mathbf{0},\, (\mathbf{G}_0'\mathbf{W}\mathbf{G}_0)^{-1}\mathbf{G}_0'\mathbf{W}\mathbf{S}\mathbf{W}\mathbf{G}_0(\mathbf{G}_0'\mathbf{W}\mathbf{G}_0)^{-1}\right) \tag{5.39}$$

□

5.1.6.4 Special Case: Optimal Weighting ($\mathbf{W} = \mathbf{S}^{-1}$) Under the optimal choice of the weight matrix, the asymptotic variance simplifies dramatically:

$$\mathbf{V}_{opt} = (\mathbf{G}_0'\mathbf{S}^{-1}\mathbf{G}_0)^{-1} \tag{5.40}$$

and the limiting distribution becomes:

$$\boxed{\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N\left(\mathbf{0},\, (\mathbf{G}_0'\mathbf{S}^{-1}\mathbf{G}_0)^{-1}\right)} \tag{5.41}$$

5.1.6.5 Role of Assumptions in the Derivation

1. **Stationarity and Ergodicity (Assumption 5.10):** Was indispensable in Step 5 to ensure that the sample moment matrices collapsed to stable and finite limits.
2. **Full Rank of $\mathbf{G}_0$:** Ensured the existence of the inversion operation performed in Step 4, allowing the isolation of the scaled error.
3. **Global Identification (Assumption 5.9):** Guaranteed that the population objective function has a unique minimum at θ_0, ensuring the consistency of the estimator.

5.1.7 Hansen Test (Overidentification Test)

5.1.7.1 Motivation The fundamental specification test in the GMM universe is the **Overidentifying Restrictions Test**, routinely called the **Hansen J Test**. The problem this test solves is the statistical verification of the orthogonality assumptions of the model when $L > K$.

In just-identified systems ($L = K$), the classical estimator has sufficient degrees of freedom to strictly zero out the sample equations. However, under overidentification, GMM minimizes a weighted distance metric, making the sample moment vector as close as possible to the null vector, but rarely zero in its entirety. The J Test formally evaluates whether this observed "residual distance" is statistically negligible (purely due to sample noise) or sufficiently large to indicate that the population moment conditions are violated.

Geometric Interpretation in $\mathbb{R}^L$:

Visualizing the sample moment conditions as a vector positioned in a space of dimension L, the GMM estimator projects the origin onto the subspace parameterized by the model. The J statistic quantifies the squared Euclidean distance (corrected by the optimal covariance matrix metric) between the optimized point and the absolute origin. Excessive distances indicate that the generated subspace does not encompass the population target, leading to the rejection of the specification.

5.1.7.2 Assumptions for the J Test **Assumption 5.13: Orthogonality under the Null Hypothesis (H_0)**

- **Role:** Sets the scenario of validity of the structural model. Assumes that there exists $\theta_0 \in \Theta$ such that $E[\mathbf{g}(\mathbf{w}_i, \theta_0)] = \mathbf{0}$.
- **Counterexample:** If any instrument contains direct correlation with the structural error, the respective moment condition will fail to converge to zero. The J statistic will diverge with n, forcing the rejection of the model.

Assumption 5.14: Optimal Weighting

- **Role:** Mathematical requirement for the test statistic to collapse into a standard chi-square distribution. Requires that $\mathbf{W}_n \xrightarrow{p} \mathbf{S}^{-1}$.
- **Counterexample:** If an arbitrary weight matrix (such as $\mathbf{I}_L$) is adopted under heteroskedasticity, the moments will receive distorted weights, causing the statistic to follow a linear combination of chi-square variables with unknown weights.

Assumption 5.15: Overidentification ($L > K$)

- **Role:** Ensures the existence of positive degrees of freedom for the test, defined by $df = L - K$.
- **Counterexample:** If $L = K$, the minimized objective function will reach zero by algebraic construction ($J = 0$), nullifying the power of the test.

5.1.7.3 Formal Statement and Derivation Let $\hat{\theta}_n$ be the efficient GMM estimator with $\mathbf{W}_n \xrightarrow{p} \mathbf{S}^{-1}$, where $\mathbf{S} = E[\mathbf{g}(\mathbf{w}_i, \theta_0)\mathbf{g}(\mathbf{w}_i, \theta_0)']$. The Hansen J Test statistic is:

$$J_n = n \cdot \mathbf{g}_n(\hat{\theta}_n)' \hat{\mathbf{S}}_n^{-1} \mathbf{g}_n(\hat{\theta}_n) \tag{5.42}$$

Theorem (Hansen, 1982). Under the null hypothesis of correct specification of the moment conditions:

$$\boxed{J_n \xrightarrow{d} \chi^2_{L-K}} \tag{5.43}$$

Proof:

Step 1: Linear Expansion of the Optimized Moment Vector

$$\mathbf{g}_n(\hat{\theta}_n) \approx \mathbf{g}_n(\theta_0) + \mathbf{G}_n(\theta_0)(\hat{\theta}_n - \theta_0) \tag{5.44}$$

Step 2: Substitution of the Asymptotic Estimation Error

Under optimal weighting, the estimation error is:

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -(\mathbf{G}_0'\mathbf{S}^{-1}\mathbf{G}_0)^{-1}\mathbf{G}_0'\mathbf{S}^{-1}\sqrt{n}\mathbf{g}_n(\theta_0) + o_p(1) \tag{5.45}$$

Step 3: Consolidation in Projection Form

Multiplying (5.44) by $\sqrt{n}$ and substituting (5.45):

$$\sqrt{n}\mathbf{g}_n(\hat{\theta}_n) \approx \left[\mathbf{I}_L - \mathbf{G}_0(\mathbf{G}_0'\mathbf{S}^{-1}\mathbf{G}_0)^{-1}\mathbf{G}_0'\mathbf{S}^{-1}\right]\sqrt{n}\mathbf{g}_n(\theta_0) \tag{5.46}$$

Step 4: Standardization and Change of Basis

Premultiplying by $\mathbf{S}^{-1/2}$ via Cholesky decomposition:

$$\mathbf{S}^{-1/2}\sqrt{n}\mathbf{g}_n(\hat{\theta}_n) \approx \left[\mathbf{I}_L - \mathbf{H}(\mathbf{H}'\mathbf{H})^{-1}\mathbf{H}'\right]\mathbf{S}^{-1/2}\sqrt{n}\mathbf{g}_n(\theta_0) = \mathbf{P}_M\mathbf{Z}_n \tag{5.47}$$

where $\mathbf{H} = \mathbf{S}^{-1/2}\mathbf{G}_0$, $\mathbf{P}_M = \mathbf{I}_L - \mathbf{H}(\mathbf{H}'\mathbf{H})^{-1}\mathbf{H}'$ is the projection matrix onto the null subspace of $\mathbf{H}'$, and $\mathbf{Z}_n = \mathbf{S}^{-1/2}\sqrt{n}\mathbf{g}_n(\theta_0) \xrightarrow{d} N(\mathbf{0}, \mathbf{I}_L)$.

Step 5: Identification of the Quadratic Form

Rewriting J_n:

$$J_n = \left(\mathbf{S}^{-1/2}\sqrt{n}\mathbf{g}_n(\hat{\theta}_n)\right)'\left(\mathbf{S}^{-1/2}\sqrt{n}\mathbf{g}_n(\hat{\theta}_n)\right) \approx \mathbf{Z}_n'\mathbf{P}_M'\mathbf{P}_M\mathbf{Z}_n = \mathbf{Z}_n'\mathbf{P}_M\mathbf{Z}_n \tag{5.48}$$

since $\mathbf{P}_M$ is symmetric and idempotent ($\mathbf{P}_M^2 = \mathbf{P}_M$).

Step 6: Derivation of Degrees of Freedom

The trace of $\mathbf{P}_M$ is:

$$\operatorname{tr}(\mathbf{P}_M) = \operatorname{tr}(\mathbf{I}_L) - \operatorname{tr}\left(\mathbf{H}(\mathbf{H}'\mathbf{H})^{-1}\mathbf{H}'\right) = L - \operatorname{tr}\left((\mathbf{H}'\mathbf{H})^{-1}\mathbf{H}'\mathbf{H}\right) = L - K \tag{5.49}$$

Therefore, by the theorem of quadratic forms of spherical normal vectors:

$$J_n \xrightarrow{d} \chi^2_{L-K} \tag{5.50}$$

□

5.1.7.4 Role of Assumptions in the Derivation

1. **Optimal Weighting (Assumption 5.14):** Was indispensable in Steps 4 and 5. If the weight matrix differed from $\mathbf{S}^{-1}$, the central matrix in the final quadratic form would lose the idempotence property, preventing simplification into a clean chi-square distribution.
2. **Overidentification (Assumption 5.15):** Ensures that the dimension of the null subspace is positive ($L - K > 0$), providing degrees of freedom for the test.

5.1.8 Parametric Hypothesis Tests

5.1.8.1 Motivation Parametric hypothesis testing solves the fundamental problem of validating economic theories under statistical uncertainty. Suppose a theory stipulates a specific restriction on the parameters (for example, constant returns to scale imposing that the sum of certain coefficients equals one). The unrestricted sample estimator $\hat{\theta}_n$ will rarely fall exactly on the restriction due to sampling fluctuations. The statistical test quantifies whether the observed deviation is probabilistically tolerable or indicates rejection of the theoretical hypothesis.

Geometric Interpretation in $\mathbb{R}^K$:

Consider the topography of the GMM objective function, $Q_n(\theta)$, as a hyperbolic surface ("bowl") in the parameter space. The unrestricted estimator $\hat{\theta}_n$ lies exactly at the lower vertex (bottom of the bowl). When we impose a set of theoretical restrictions defining a subspace or hyperplane, the restricted estimator $\tilde{\theta}_n$ will correspond to the lowest point on the surface *that intersects the restriction hyperplane*. The tests evaluate, from different perspectives, the loss of optimality (the gain in height in the "bowl") resulting from the displacement from the unrestricted minimum to the restricted minimum.

5.1.8.2 Assumptions for Parametric Tests **Assumption 5.16: Asymptotic Normality of the Estimator**

- **Role:** Ensures that, in large samples, the scaled sampling distribution of the estimation error follows a Gaussian law, allowing quadratic forms to be mapped directly to stable chi-square distributions.
- **Counterexample:** Under violation of fourth-order moments, the limiting distribution of the estimator would exhibit non-standardized variability, invalidating p-values derived from the χ^2 table.

Assumption 5.17: Smoothness and Full Rank of the Restriction

- **Role:** The set of restrictions structured as $\mathbf{r}(\theta) = \mathbf{0}$ must be continuously differentiable to allow stable linear approximations via Taylor Expansion (Delta Method).
- **Counterexample:** If the restriction contains points of discontinuity or geometric corners, the derivative matrix will fail to be defined, collapsing the calculation of the test covariance matrix.

5.1.8.3 Wald Test **Theorem (Wald Test).** Under the null hypothesis $H_0 : \mathbf{r}(\theta_0) = \mathbf{0}$, the Wald statistic:

$$W_n = n \cdot \mathbf{r}(\hat{\theta}_n)' \left[\mathbf{R}(\hat{\theta}_n) \hat{\mathbf{V}}_\theta \mathbf{R}(\hat{\theta}_n)' \right]^{-1} \mathbf{r}(\hat{\theta}_n) \tag{5.51}$$

where $\mathbf{R}(\theta) = \partial \mathbf{r}(\theta) / \partial \theta'$ $(q \times K)$, satisfies:

$$\boxed{W_n \xrightarrow{d} \chi^2_q} \tag{5.52}$$

Proof (sketch): From the asymptotic normality result, $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(\mathbf{0}, \mathbf{V}_\theta)$. By the Delta Method, $\sqrt{n}\mathbf{r}(\hat{\theta}_n) \xrightarrow{d} N(\mathbf{0}, \mathbf{R}\mathbf{V}_\theta\mathbf{R}')$. The quadratic form with the inverse variance converges to χ^2_q. □

5.1.8.4 GMM Distance Test An alternative robust methodology consists of the direct evaluation of the increment suffered by the minimized objective function when moving from the unrestricted to the restricted model:

$$D_n = n\left[Q_n(\tilde{\theta}_n) - Q_n(\hat{\theta}_n)\right] \tag{5.53}$$

where $\tilde{\theta}_n$ denotes the minimizer of the objective function under the imposition of the restriction $\mathbf{r}(\theta) = \mathbf{0}$.

Theorem (Distance Test). Under the null hypothesis and using the **same optimal weighting matrix** for both models:

$$\boxed{D_n \xrightarrow{d} \chi^2_q} \tag{5.54}$$

Indispensable Assumption: For the distance statistic to converge rigorously to χ^2_q, it is **strictly mandatory** that the **same optimal sample weighting matrix** ($\hat{\mathbf{S}}_n^{-1}$) be employed in the estimation of both models (restricted and unrestricted). If different weight matrices were used, the difference between the objective functions would incorporate spurious metric variations, destroying the idempotent structure necessary for convergence to a pure χ^2.

5.2 Special Cases of GMM

The Generalized Method of Moments (GMM) constitutes a unifying framework that encompasses various classical estimators as special cases, depending on the choice of the weighting matrix $\mathbf{W}_n$ and the definition of the moment conditions. This section explores nine special cases, organized in increasing order of generality, demonstrating how GMM subsumes them as particular cases.

5.2.1 Perfectly Identified GMM (Just-Identified)

5.2.1.1 Motivation and Geometric Interpretation **Perfectly identified GMM** (or just-identified) solves the problem of estimating parameters in systems where the number of moment conditions is strictly equal to the number of parameters ($L = K$). While general GMM deals with overidentification by transforming it into a quadratic optimization problem, the perfectly identified case reduces to a classical root-finding problem for a system of simultaneous equations.

Geometric Interpretation in $\mathbb{R}^K$:

Each sample moment condition can be visualized as a surface (or hyperplane, in the linear case) of dimension $K-1$ in $\mathbb{R}^K$. In a perfectly identified model ($L = K$), the GMM estimator locates the exact point where these K surfaces intersect. Unlike the overidentified case ($L > K$), where surfaces rarely meet at a single point due to sample noise, in the $L = K$ case there exists a $\hat{\theta}$ that zeros all equations simultaneously. Geometrically, the estimator is the **common vertex** of these sample surfaces, making the choice of the distance metric (the matrix $\mathbf{W}$) irrelevant for the final solution.

5.2.1.2 Assumptions **Assumption 5.18: Exact Identification (Just-Identification)**

- **Role:** Ensures that the number of moment conditions L is strictly equal to the number of parameters K ($L = K$). This allows the system of equations to be solved without the need to weight the relative statistical importance of different moments.
- **Counterexample:** If $L > K$, the system becomes overdetermined and does not have a common root for all sample moments, requiring a weight matrix $\mathbf{W}$ to minimize a quadratic distance.

Assumption 5.19: Non-Singular Jacobian

- **Role:** Ensures that the K moment conditions are not linearly dependent. Formally, $\text{rank}(\mathbf{G}_0) = K$, where $\mathbf{G}_0 = E[\partial \mathbf{g}(\mathbf{w}_i, \theta_0)/\partial \theta']$.
- **Counterexample:** If one moment condition is an exact linear function of another, the system will have infinitely many solutions, preventing the location of a unique value for $\hat{\theta}$.

5.2.1.3 Formal Statement and Derivation Let $\{\mathbf{w}_i\}_{i=1}^n$ be a sequence of stationary and ergodic random vectors. Let $\theta_0 \in \Theta \subset \mathbb{R}^K$ be the true vector. We define $\mathbf{g} : \mathbb{R}^M \times \Theta \to \mathbb{R}^L$ with $L = K$.

The perfectly identified GMM estimator $\hat{\theta}_n$ is defined as the argument that minimizes:

$$Q_n(\theta) = \bar{\mathbf{g}}_n(\theta)' \mathbf{W}_n \bar{\mathbf{g}}_n(\theta) \tag{5.159}$$

where $\bar{\mathbf{g}}_n(\theta) = \frac{1}{n} \sum_{i=1}^n \mathbf{g}(\mathbf{w}_i, \theta)$ and $\mathbf{W}_n$ is positive definite. Under just-identification, the estimator satisfies $\bar{\mathbf{g}}_n(\hat{\theta}_n) = \mathbf{0}$.

Theorem (Independence of the Weight Matrix). Under $L = K$ and Assumption 5.19, the estimator $\hat{\theta}_n$ is independent of $\mathbf{W}_n$ and satisfies:

$$\boxed{\bar{\mathbf{g}}_n(\hat{\theta}_n) = \mathbf{0}} \tag{5.160}$$

Proof:

By the definition of a positive definite matrix, $\mathbf{x}'\mathbf{W}\mathbf{x} = 0 \iff \mathbf{x} = \mathbf{0}$. If $L = K$ and Assumption 5.19 is satisfied, there exists $\hat{\theta}_n$ such that $\bar{\mathbf{g}}_n(\hat{\theta}_n) = \mathbf{0}$ (by the Inverse Function Theorem). Substituting into (5.159), we obtain $Q_n(\hat{\theta}_n) = 0$, which is the global minimum, regardless of $\mathbf{W}_n$. □

Theorem (Asymptotic Normality). Under regularity conditions:

$$\boxed{\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N\left(\mathbf{0},\, \mathbf{G}_0^{-1}\mathbf{S}(\mathbf{G}_0')^{-1}\right)} \tag{5.161}$$

where $\mathbf{S} = E[\mathbf{g}(\mathbf{w}_i, \theta_0)\mathbf{g}(\mathbf{w}_i, \theta_0)']$.

Proof (sketch): Expanding $\bar{\mathbf{g}}_n(\hat{\theta}_n)$ around θ_0 via the Mean Value Theorem, isolating $\sqrt{n}(\hat{\theta}_n - \theta_0)$, and applying the CLT with $\sqrt{n}\bar{\mathbf{g}}_n(\theta_0) \xrightarrow{d} N(\mathbf{0}, \mathbf{S})$ yields (5.161). □

5.2.1.4 Role of Assumptions The assumption of **Exact Identification ($L = K$)** was indispensable for ensuring that the Jacobian matrix $\mathbf{G}_0$ was square, allowing direct inversion in (5.161). If $L > K$, the matrix would not be square, preventing inversion and requiring the use of the weight matrix $\mathbf{W}$ to define the estimator through the first-order conditions of the quadratic form.

5.2.2 Optimal GMM (Efficient)

5.2.2.1 Motivation and Geometric Interpretation **Optimal GMM** solves the problem of statistical inefficiency in **overidentified** systems ($L > K$). In basic GMM, any positive definite weighting matrix $\mathbf{W}_n$ produces a consistent estimator, but the precision (variance) of the estimates depends critically on how different moments are weighted.

Geometric Interpretation in $\mathbb{R}^L$:

Visualize the sample moment vector $\bar{\mathbf{g}}_n(\theta)$ as a point in a space of dimension L. The weighting matrix $\mathbf{W}$ defines the **distance metric** (the geometric deformation of the error ellipsoid). If the moments have distinct variances or are correlated, simple Euclidean distance ($\mathbf{W} = \mathbf{I}$) ignores that some information is more "noisy" than others. Optimal GMM uses $\mathbf{S}^{-1}$ to rotate and scale the space, so that the projection occurs in the direction of **minimum variance**.

5.2.2.2 Assumptions Assumption 5.20: Overidentification

- **Role:** Ensures that $L \geq K$, allowing the choice of $\mathbf{W}$ to be statistically relevant for asymptotic efficiency.
- **Counterexample:** If $L = K$, the system has a unique solution with $\bar{\mathbf{g}}_n(\hat{\theta}) = \mathbf{0}$, making $\mathbf{W}$ irrelevant.

Assumption 5.21: Consistency of the Weighting Matrix

- **Role:** Requires that $\mathbf{W}_n \xrightarrow{p} \mathbf{S}^{-1}$, where $\mathbf{S} = E[\mathbf{m}(\mathbf{w}_i, \theta_0)\mathbf{m}(\mathbf{w}_i, \theta_0)']$.
- **Counterexample:** If $\mathbf{W}_n$ converges to $\mathbf{A} \neq \mathbf{S}^{-1}$, the estimator will still be consistent, but the asymptotic variance will be larger (in the PSD sense) than that of the optimal GMM.

5.2.2.3 Formal Statement and Derivation

The optimal GMM estimator $\hat{\theta}_{opt}$ minimizes:

$$Q_n(\theta) = \bar{\mathbf{g}}_n(\theta)'\hat{\mathbf{S}}_n^{-1}\bar{\mathbf{g}}_n(\theta) \tag{5.162}$$

where $\hat{\mathbf{S}}_n \xrightarrow{p} \mathbf{S} = E[\mathbf{m}(\mathbf{w}_i, \theta_0)\mathbf{m}(\mathbf{w}_i, \theta_0)']$.

Theorem (Efficiency of Optimal GMM). Under regularity conditions:

$$\boxed{\sqrt{n}(\hat{\theta}_{opt} - \theta_0) \xrightarrow{d} N\left(\mathbf{0},\, (\mathbf{G}_0'\mathbf{S}^{-1}\mathbf{G}_0)^{-1}\right)} \tag{5.163}$$

and for any other positive definite matrix $\mathbf{W}$:

$$\boxed{\mathbf{V}(\mathbf{W}) \succeq \mathbf{V}(\mathbf{S}^{-1})} \tag{5.164}$$

Proof (sketch): The FOC of GMM is $\mathbf{G}_n(\hat{\theta}_n)'\mathbf{W}_n\bar{\mathbf{g}}_n(\hat{\theta}_n) = \mathbf{0}$. Expanding $\bar{\mathbf{g}}_n(\hat{\theta}_n)$ around θ_0 and applying Slutsky yields the variance $\mathbf{V}(\mathbf{W}) = (\mathbf{G}_0'\mathbf{W}\mathbf{G}_0)^{-1}\mathbf{G}_0'\mathbf{W}\mathbf{S}\mathbf{W}\mathbf{G}_0(\mathbf{G}_0'\mathbf{W}\mathbf{G}_0)^{-1}$. For $\mathbf{W} = \mathbf{S}^{-1}$, $\mathbf{V}(\mathbf{S}^{-1}) = (\mathbf{G}_0'\mathbf{S}^{-1}\mathbf{G}_0)^{-1}$. The proof of $\mathbf{V}(\mathbf{W}) \succeq \mathbf{V}(\mathbf{S}^{-1})$ follows by the projection matrix argument presented in Section 5.1.5. □

5.2.2.4 Role of Assumptions

The assumption of **Consistency of $\mathbf{W}_n$ to $\mathbf{S}^{-1}$** was indispensable. If $\mathbf{W}$ does not collapse exactly to the inverse of the population variance of the moments, the algebraic simplification fails, preventing the estimator from reaching the lower bound of asymptotic variance.

5.2.3 Method of Moments (MM)

5.2.3.1 Motivation and Geometric Interpretation

The **Method of Moments (MM)**, systematized by Karl Pearson (1895), solves the problem of estimating population parameters when one does not wish (or cannot) assume a strict functional form for the distribution of the data, but has grounded knowledge about the mathematical expectations of certain functions of the random variables.

Geometric Interpretation in $\mathbb{R}^K$:

Each population moment condition $E[\mathbf{g}(\mathbf{w}_i, \theta)] = \mathbf{0}$ establishes a surface (or hyperplane) of dimension $K - 1$ in the parameter space. In MM, with exactly K equations for K parameters, the estimator $\hat{\theta}_{MM}$ is located at the **unique intersection** of these K sample-estimated surfaces.

5.2.3.2 Assumptions Assumption 5.22: Exact Identification

- **Role:** Ensures that $L = K$, allowing a uniquely determined solution.
- **Counterexample:** If $L > K$, the system becomes overdetermined and there will be no θ that zeros all sample moments simultaneously.

Assumption 5.23: Rank Condition

- **Role:** Ensures that $\mathbf{G}_0 = E[\partial \mathbf{g}(\mathbf{w}_i, \theta_0)/\partial \theta']$ is non-singular.
- **Counterexample:** If we try to estimate the mean and variance using only the first moment and a linear function of it, the Jacobian matrix will have rank 1, making the parameters indistinguishable.

5.2.3.3 Formal Statement and Derivation The method of moments estimator $\hat{\theta}_n$ is defined as the exact solution of the system of K sample equations:

$$\boxed{\bar{\mathbf{g}}_n(\hat{\theta}_n) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{g}(\mathbf{w}_i, \hat{\theta}_n) = \mathbf{0}} \tag{5.165}$$

Theorem (Asymptotic Normality of MM). Under regularity conditions:

$$\boxed{\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N\left(\mathbf{0},\, \mathbf{G}_0^{-1}\mathbf{S}(\mathbf{G}_0')^{-1}\right)} \tag{5.166}$$

Proof: Expanding $\bar{\mathbf{g}}_n(\hat{\theta}_n)$ around θ_0 via the Mean Value Theorem:

$$\mathbf{0} = \bar{\mathbf{g}}_n(\theta_0) + \mathbf{G}_n(\bar{\theta})(\hat{\theta}_n - \theta_0)$$

Multiplying by $\sqrt{n}$, isolating, and applying CLT + Slutsky yields (5.166). □

5.2.3.4 Role of Assumptions The assumption of **Just-Identification ($L = K$)** was indispensable for ensuring that the Jacobian matrix $\mathbf{G}_0$ was square, allowing its direct inversion without resorting to projections based on external weighting matrices.

5.2.4 Ordinary Least Squares (OLS)

5.2.4.1 Motivation and Geometric Interpretation OLS solves the problem of estimating the linear relationship between a dependent variable and regressors under the assumption of exogeneity. In the GMM context, it is the special case of **exact identification** ($L = K$) where the moment function is defined by the orthogonality condition between the regressors and the error.

Geometric Interpretation in $\mathbb{R}^n$:

The OLS estimator performs an **orthogonal projection** of $\mathbf{y}$ onto the linear subspace spanned by the columns of $\mathbf{X}$. The residual vector $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$ is perpendicular to each column of $\mathbf{X}$, validating the sample moment condition $\mathbf{X}'\mathbf{e} = \mathbf{0}$.

5.2.4.2 Assumptions Assumption 5.24: Population Orthogonality

- **Role:** States that $E[\mathbf{x}_i \varepsilon_i] = \mathbf{0}$.
- **Counterexample:** In a supply and demand model, if price is determined simultaneously with quantity, OLS will be inconsistent.

Assumption 5.25: No Perfect Multicollinearity

- **Role:** Ensures that $\mathbf{Q}_{xx} = E[\mathbf{x}_i\mathbf{x}_i']$ is invertible.
- **Counterexample:** Including "weight in kg" and "weight in pounds" as distinct regressors generates singularity.

5.2.4.3 Formal Statement and Derivation The OLS estimator $\hat{\beta}_n$ is the GMM that uses $\mathbf{g}(\mathbf{w}_i, \beta) = \mathbf{x}_i(y_i - \mathbf{x}_i'\beta)$ and $\mathbf{W}_n = \mathbf{I}_K$.

Defining $\bar{\mathbf{g}}_n(\beta) = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i(y_i - \mathbf{x}_i'\beta)$, the FOC is $\bar{\mathbf{g}}_n(\hat{\beta}_n) = \mathbf{0}$:

$$\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i(y_i - \mathbf{x}_i'\hat{\beta}_n) = \mathbf{0} \tag{5.167}$$

Isolating:

$$\boxed{\hat{\beta}_n = \left(\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i'\right)^{-1}\left(\sum_{i=1}^n \mathbf{x}_i y_i\right)} \tag{5.168}$$

Theorem (Consistency and Normality). Under exogeneity:

$$\operatorname{plim}\hat{\beta}_n = \beta_0 \tag{5.169}$$

$$\sqrt{n}(\hat{\beta}_n - \beta_0) \xrightarrow{d} N\left(\mathbf{0},\, \mathbf{Q}_{xx}^{-1}\mathbf{\Omega}\mathbf{Q}_{xx}^{-1}\right) \tag{5.170}$$

where $\mathbf{\Omega} = E[\varepsilon_i^2\mathbf{x}_i\mathbf{x}_i']$.

5.2.4.4 Role of Assumptions The assumption of **Orthogonality ($E[\mathbf{x}\varepsilon] = \mathbf{0}$)** was indispensable. Without it, the term $\frac{1}{n}\mathbf{X}'\varepsilon$ would converge to a non-zero value, generating inconsistency.

5.2.5 Robust Variance (Eicker-Huber-White)

5.2.5.1 Motivation and Geometric Interpretation The **EHW** estimator solves the problem of **invalid inference under unknown heteroskedasticity**. OLS remains consistent under heteroskedasticity, but the classical formula $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ becomes biased.

Geometric Interpretation in $\mathbb{R}^n$:

Under homoskedasticity, the error vector $\mathbf{u}$ resides in a **probability sphere**. Under heteroskedasticity, this sphere deforms into an **ellipsoid** with non-uniform axes. EHW projects the variability of these individual deviations back to $\mathbb{R}^K$, adjusting the "width" of the confidence ellipses for each direction.

5.2.5.2 Assumptions Assumption 5.26: Independence of Observations

- **Role:** Ensures that $\mathbf{\Omega} = E[\mathbf{u}\mathbf{u}']$ is diagonal.
- **Counterexample:** In time series with autocorrelation, the standard EHW would underestimate the variance.

Assumption 5.27: Existence of Fourth-Order Moments

- **Role:** Ensures the convergence of $\frac{1}{n}\sum \mathbf{x}_i\mathbf{x}_i'\hat{e}_i^2$.
- **Counterexample:** Heavy-tailed distributions prevent stable convergence.

5.2.5.3 Formal Statement and Derivation Let $\hat{\beta}_n$ be the OLS estimator and $\hat{e}_i = y_i - \mathbf{x}_i'\hat{\beta}_n$. The EHW estimator is:

$$\boxed{\hat{\mathbf{V}}_{EHW} = n(\mathbf{X}'\mathbf{X})^{-1}\left(\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i'\hat{e}_i^2\right)(\mathbf{X}'\mathbf{X})^{-1}} \tag{5.171}$$

Theorem (Consistency of EHW). Under Assumptions 5.26 and 5.27:

$$\hat{\mathbf{V}}_{EHW} \xrightarrow{p} \mathbf{Q}^{-1}\mathbf{S}\mathbf{Q}^{-1} \tag{5.172}$$

where $\mathbf{Q} = E[\mathbf{x}_i\mathbf{x}_i']$ and $\mathbf{S} = E[u_i^2\mathbf{x}_i\mathbf{x}_i']$.

Proof (sketch): $\sqrt{n}(\hat{\beta}_n - \beta_0) \xrightarrow{d} N(\mathbf{0}, \mathbf{Q}^{-1}\mathbf{S}\mathbf{Q}^{-1})$. The consistency of $\hat{\mathbf{S}} = \frac{1}{n}\sum \mathbf{x}_i\mathbf{x}_i'\hat{e}_i^2$ for $\mathbf{S}$ follows from the expansion $\hat{e}_i = u_i - \mathbf{x}_i'(\hat{\beta}_n - \beta_0)$ and the consistency of $\hat{\beta}_n$. □

5.2.5.4 Role of Assumptions **Independence (Assumption 5.26)** was indispensable. If there were serial correlation, the "filling" of the sandwich would accumulate cross-covariance terms, invalidating the purely additive form of (5.171).

5.2.6 Generalized Least Squares (GLS)

5.2.6.1 Motivation and Geometric Interpretation **GLS** solves the problem of **efficient estimation** in the presence of a known covariance structure $\mathbf{\Omega}$ for the errors, allowing observations with higher variance to receive lower relative weight.

Geometric Interpretation in $\mathbb{R}^n$:

In GLS, the space is modified by a transformation $\mathbf{P}$ such that $\mathbf{P}'\mathbf{P} = \mathbf{\Omega}^{-1}$. This is equivalent to performing an **orthogonal projection in a non-Euclidean metric space**, where distances are weighted by the inverse of the covariance matrix.

5.2.6.2 Assumptions **Assumption 5.28: Positive Definite Covariance Structure**

- **Role:** Assumes $\text{Var}(\mathbf{u}|\mathbf{X}) = \sigma^2\mathbf{\Omega}$, with $\mathbf{\Omega}$ SPD.
- **Counterexample:** If $\mathbf{\Omega}$ is singular, correct weighting of observations becomes impossible.

5.2.6.3 Formal Statement and Derivation GLS is the GMM with $\mathbf{g}_n(\beta) = \frac{1}{n}\mathbf{X}'\mathbf{\Omega}^{-1}(\mathbf{y} - \mathbf{X}\beta)$ and $\mathbf{W}_n = \mathbf{I}_K$ (just-identified).

The FOC $\mathbf{g}_n(\hat{\beta}_{GLS}) = \mathbf{0}$ yields:

$$\boxed{\hat{\beta}_{GLS} = (\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{y}} \tag{5.173}$$

Theorem (Asymptotic Normality of GLS). Under exogeneity:

$$\sqrt{n}(\hat{\beta}_{GLS} - \beta_0) \xrightarrow{d} N\left(\mathbf{0}, \sigma^2\mathbf{Q}_{gls}^{-1}\right) \tag{5.174}$$

where $\mathbf{Q}_{gls} = E[\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X}/n]$.

5.2.6.4 Role of Assumptions The assumption of **Exogeneity** ($E[\mathbf{x}_i u_i] = \mathbf{0}$) was indispensable. If the regressors are endogenous, GLS will converge to a biased value, and the efficiency correction will only alter the weight of the bias, without removing it.

5.2.7 Instrumental Variables (IV)

5.2.7.1 Motivation and Geometric Interpretation **IV** solves the problem of **endogeneity** when $E[\mathbf{x}_i \varepsilon_i] \neq \mathbf{0}$. In GMM, it is the case of **exact identification** ($L = K$) with instruments $\mathbf{Z}$.

Geometric Interpretation in $\mathbb{R}^n$:

OLS projects $\mathbf{y}$ onto Col($\mathbf{X}$). If $\mathbf{X}$ is correlated with the error, the subspace is "tilted." IV uses $\mathbf{Z}$ as an external reference, projecting $\mathbf{y}$ onto Col($\mathbf{X}$) only in the direction that is orthogonal to $\mathbf{Z}$.

5.2.7.2 Assumptions **Assumption 5.29: Exogeneity of the Instrument**

- **Role:** $E[\mathbf{z}_i \varepsilon_i] = \mathbf{0}$.
- **Counterexample:** If the instrument is endogenous, IV will be inconsistent.

Assumption 5.30: Relevance of the Instrument

- **Role:** $\text{rank}(E[\mathbf{z}_i \mathbf{x}_i']) = K$.
- **Counterexample:** Weak instruments generate infinite variance.

5.2.7.3 Formal Statement and Derivation IV is the GMM with $\mathbf{g}_n(\beta) = \frac{1}{n}\mathbf{Z}'(\mathbf{y} - \mathbf{X}\beta)$ and $L = K$.

The FOC $\mathbf{g}_n(\hat{\beta}_{IV}) = \mathbf{0}$ yields:

$$\boxed{\hat{\beta}_{IV} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}} \tag{5.175}$$

Theorem (Consistency and Normality of IV). Under Assumptions 5.29 and 5.30:

$$\text{plim}\, \hat{\beta}_{IV} = \beta_0 \tag{5.176}$$

$$\sqrt{n}(\hat{\beta}_{IV} - \beta_0) \xrightarrow{d} N\left(\mathbf{0},\, \sigma^2 \mathbf{Q}_{zx}^{-1} \mathbf{Q}_{zz} \mathbf{Q}_{xz}^{-1}\right) \tag{5.177}$$

5.2.7.4 Role of Assumptions **Relevance (Assumption 5.30)** was indispensable. If $\mathbf{Z}'\mathbf{X}$ were singular, the inverse would not exist and the IV estimator would be undefined.

5.2.8 Two-Stage Least Squares (2SLS)

5.2.8.1 Motivation and Geometric Interpretation **2SLS** simultaneously solves **endogeneity** and **overidentification**. When $L > K$, the system is overidentified, and 2SLS emerges as the special case of GMM with $\mathbf{W}_n = (\mathbf{Z}'\mathbf{Z})^{-1}$.

Geometric Interpretation in $\mathbb{R}^n$:

1. **First Stage:** Project $\mathbf{X}$ onto Col($\mathbf{Z}$): $\hat{\mathbf{X}} = \mathbf{P}_Z \mathbf{X}$.
2. **Second Stage:** Regress $\mathbf{y}$ on $\hat{\mathbf{X}}$.

5.2.8.2 Assumptions Assumption 5.31: Exogeneity of the Instrument

- **Role:** $E[\mathbf{Z}'\mathbf{u}] = \mathbf{0}$.

Assumption 5.32: Relevance of the Instrument

- **Role:** $\text{rank}(E[\mathbf{Z}'\mathbf{X}]) = K$.

Assumption 5.33: Homoskedasticity

- **Role:** Ensures that $\mathbf{W}_n = (\mathbf{Z}'\mathbf{Z})^{-1}$ is optimal.

5.2.8.3 Formal Statement and Derivation 2SLS is the GMM with $\bar{\mathbf{g}}_n(\beta) = \frac{1}{n}\mathbf{Z}'(\mathbf{y} - \mathbf{X}\beta)$ and $\mathbf{W}_n = (\mathbf{Z}'\mathbf{Z})^{-1}$.

Minimizing $Q_n(\beta) = (\mathbf{y} - \mathbf{X}\beta)'\mathbf{P}_Z(\mathbf{y} - \mathbf{X}\beta)$:

$$\boxed{\hat{\beta}_{2SLS} = (\mathbf{X}'\mathbf{P}_Z\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_Z\mathbf{y}} \tag{5.178}$$

Theorem (Asymptotic Normality of 2SLS). Under Assumptions 5.31 to 5.33:

$$\sqrt{n}(\hat{\beta}_{2SLS} - \beta_0) \xrightarrow{d} N\left(\mathbf{0},\, \sigma^2(\mathbf{\Sigma}_{xz}\mathbf{\Sigma}_{zz}^{-1}\mathbf{\Sigma}_{zx})^{-1}\right) \tag{5.179}$$

5.2.8.4 Role of Assumptions Relevance (Assumption 5.32) was indispensable. If $\mathbf{Z}$ is not correlated with $\mathbf{X}$, the matrix $(\mathbf{X}'\mathbf{P}_Z\mathbf{X})$ will be singular, making the estimator impossible to define.

5.2.9 Three-Stage Least Squares (3SLS)

5.2.9.1 Motivation and Geometric Interpretation 3SLS solves the problem of **statistical inefficiency** in overidentified systems of simultaneous equations. While 2SLS estimates each equation in isolation, 3SLS treats the system as a single block, using the covariance matrix of the residuals to weight the global system.

Geometric Interpretation in $\mathbb{R}^{nL}$:

3SLS operates in a subspace of dimension nL where the equations are "coupled." The geometry involves an **oblique projection** adjusted by the cross-covariance matrix $\mathbf{\Sigma}$, resulting in a confidence ellipsoid that is "contracted" relative to 2SLS.

5.2.9.2 Assumptions Assumption 5.34: Systemic Exogeneity

- **Role:** $E[\mathbf{z}_t u_{jt}] = \mathbf{0}$ for all j.
- **Counterexample:** If an instrument is correlated with the shock of another equation, 3SLS will spread the bias throughout the entire system.

Assumption 5.35: Non-Singularity of the Cross-Covariance

- **Role:** $\mathbf{\Sigma} = E[\mathbf{u}_t\mathbf{u}_t']$ must be SPD.
- **Counterexample:** If the errors of two equations are perfectly correlated, $\mathbf{\Sigma}$ will be singular.

5.2.9.3 Formal Statement and Derivation Consider the stacked system $\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$, with $\mathbf{X} = \text{diag}(\mathbf{X}_1, \ldots, \mathbf{X}_L)$ and $\overline{\mathbf{Z}} = \mathbf{I}_L \otimes \mathbf{Z}$.

3SLS is the GMM with $\mathbf{g}_n(\beta) = \frac{1}{n}\overline{\mathbf{Z}}'(\mathbf{y} - \mathbf{X}\beta)$ and $\mathbf{W}_n = [\frac{1}{n}\overline{\mathbf{Z}}'(\hat{\mathbf{\Sigma}} \otimes \mathbf{I}_n)\overline{\mathbf{Z}}]^{-1} = \hat{\mathbf{\Sigma}}^{-1} \otimes (\mathbf{Z}'\mathbf{Z})^{-1}$.

The FOC yields:

$$\boxed{\hat{\beta}_{3SLS} = [\mathbf{X}'(\hat{\mathbf{\Sigma}}^{-1} \otimes \mathbf{P}_Z)\mathbf{X}]^{-1}\mathbf{X}'(\hat{\mathbf{\Sigma}}^{-1} \otimes \mathbf{P}_Z)\mathbf{y}} \tag{5.180}$$

Theorem (Asymptotic Normality of 3SLS). Under Assumptions 5.34 and 5.35:

$$\sqrt{n}(\hat{\beta}_{3SLS} - \beta_0) \xrightarrow{d} N\left(\mathbf{0},\ \{E[\mathbf{X}'(\mathbf{\Sigma}^{-1} \otimes \mathbf{P}_Z)\mathbf{X}]\}^{-1}\right) \tag{5.181}$$

5.2.9.4 Role of Assumptions The assumption of **Non-Singularity of $\mathbf{\Sigma}$ (Assumption 5.35)** was indispensable. Without the invertibility of $\mathbf{\Sigma}$, we could not define the optimal weight metric $\hat{\mathbf{\Sigma}}^{-1}$ that couples the equations. Without this coupling, the estimator would collapse to equation-by-equation 2SLS, losing the efficiency gains from the correlation between errors.

5.2.10 Generalized Nonlinear Least Squares (GNLLS)

5.2.10.1 Motivation and Geometric Interpretation **Generalized Nonlinear Least Squares (GNLLS)** solves the problem of estimating parameters in models where the functional relationship between variables is inherently **non-linear** and the stochastic errors exhibit a **non-spherical** structure (heteroskedasticity or autocorrelation). While classical OLS and GLS are limited to linear response surfaces (hyperplanes), GNLLS allows the researcher to model complex phenomena, such as logistic growth curves or exponential decay, without sacrificing statistical efficiency in the presence of correlated errors.

In the **GMM** context, GNLLS emerges as a special case where the moment conditions are derived from the gradient of the non-linear function, weighted by the inverse of the error covariance matrix.

Geometric Interpretation in $\mathbb{R}^3$:

Imagine a cloud of points in $\mathbb{R}^3$ that appears to be distributed along a curved surface (such as a saddle or a rotated parabola). The GNLLS estimator seeks the parametric surface that best fits these points. Due to the covariance structure $\mathbf{\Omega}$, the "best" proximity is not measured by simple vertical Euclidean distance, but by a metric that "stretches" or "compresses" the sample space to compensate for the unequal variability of the errors. Visually, GNLLS performs an **oblique projection** onto the non-linear manifold, where the projection angle is dictated by the correlation structure among the errors.

5.2.10.2 Assumptions **Assumption 5.36: Differentiability and Continuity of the Model**

- **Role:** Ensures that the response surface $h(\mathbf{x}_i, \theta)$ is smooth enough to allow the use of gradient-based optimization methods and the application of Taylor expansion in asymptotic theory.
- **Counterexample:** If $h(\cdot)$ were a step function (discrete), the gradient would be zero or undefined over almost the entire domain, making it impossible to locate the minimum via first-order conditions.

Assumption 5.37: Global Identification (Uniqueness)

- **Role:** Ensures that there exists only one vector θ_0 in the parameter space Θ that minimizes the population cost function.
- **Counterexample:** If the model is overparameterized (e.g., $y = \theta_1\theta_2 x + \varepsilon$), there will be a continuum of combinations of θ_1 and θ_2 that produce the same fit, making the Jacobian matrix singular.

Assumption 5.38: Non-Singularity of the Covariance Matrix $\mathbf{\Omega}$

- **Role:** Ensures that the optimal weight matrix of GNLLS, $\mathbf{W} = \mathbf{\Omega}^{-1}$, exists and is positive definite, allowing the "spherification" of the transformed errors.
- **Counterexample:** If two errors were perfectly correlated, $\mathbf{\Omega}$ would be singular, preventing the numerical computation of the objective function.

5.2.10.3 Formal Statement and Derivation Consider the non-linear model $y_i = h(\mathbf{x}_i, \theta_0) + \varepsilon_i$, where $E[\varepsilon_i|\mathbf{x}_i] = 0$. Let $\mathbf{y} \in \mathbb{R}^n$ be the dependent vector and $\mathbf{h}(\mathbf{X}, \theta) \in \mathbb{R}^n$ the vector of functions. Assume that $E[\varepsilon\varepsilon'|\mathbf{X}] = \sigma^2\mathbf{\Omega}$, where $\mathbf{\Omega}$ is a known SPD matrix.

The GNLLS estimator $\hat{\theta}_n$ minimizes:

$$Q_n(\theta) = [\mathbf{y} - \mathbf{h}(\mathbf{X}, \theta)]'\mathbf{\Omega}^{-1}[\mathbf{y} - \mathbf{h}(\mathbf{X}, \theta)] \tag{5.186}$$

Step 1: First Order Condition (FOC)

Differentiating and setting to zero:

$$\nabla_\theta Q_n(\hat{\theta}_n) = -2\mathbf{J}(\hat{\theta}_n)'\mathbf{\Omega}^{-1}[\mathbf{y} - \mathbf{h}(\mathbf{X}, \hat{\theta}_n)] = \mathbf{0} \tag{5.187}$$

where $\mathbf{J}(\theta) = \partial\mathbf{h}(\mathbf{X}, \theta)/\partial\theta'$ is the $n \times K$ Jacobian matrix.

Step 2: Taylor Expansion (Linearization)

Expanding $\mathbf{h}(\mathbf{X}, \hat{\theta}_n)$ around θ_0:

$$\mathbf{h}(\mathbf{X}, \hat{\theta}_n) \approx \mathbf{h}(\mathbf{X}, \theta_0) + \mathbf{J}(\theta_0)(\hat{\theta}_n - \theta_0) \tag{5.188}$$

Substituting into (5.187) and using $\mathbf{y} = \mathbf{h}(\mathbf{X}, \theta_0) + \varepsilon$:

$$\mathbf{J}(\hat{\theta}_n)'\mathbf{\Omega}^{-1}[\varepsilon - \mathbf{J}(\theta_0)(\hat{\theta}_n - \theta_0)] \approx \mathbf{0} \tag{5.189}$$

Step 3: Isolating the Estimation Error

$$\mathbf{J}(\hat{\theta}_n)'\mathbf{\Omega}^{-1}\mathbf{J}(\theta_0)(\hat{\theta}_n - \theta_0) \approx \mathbf{J}(\hat{\theta}_n)'\mathbf{\Omega}^{-1}\varepsilon \tag{5.190}$$

Multiplying by $\sqrt{n}$ and premultiplying by the inverse:

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \approx \left[\frac{1}{n}\mathbf{J}(\hat{\theta}_n)'\mathbf{\Omega}^{-1}\mathbf{J}(\theta_0)\right]^{-1} \frac{1}{\sqrt{n}}\mathbf{J}(\hat{\theta}_n)'\mathbf{\Omega}^{-1}\varepsilon \tag{5.191}$$

Step 4: Stochastic Limits and Asymptotic Distribution

- $\hat{\theta}_n \xrightarrow{p} \theta_0 \Rightarrow \mathbf{J}(\hat{\theta}_n) \xrightarrow{p} \mathbf{J}_0$
- $\frac{1}{n}\mathbf{J}_0'\mathbf{\Omega}^{-1}\mathbf{J}_0 \xrightarrow{p} \mathbf{H}_0$
- $\frac{1}{\sqrt{n}}\mathbf{J}_0'\mathbf{\Omega}^{-1}\varepsilon \xrightarrow{d} N(\mathbf{0}, \sigma^2\mathbf{H}_0)$

By Slutsky's Theorem:

$$\boxed{\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N\left(\mathbf{0},\, \sigma^2\mathbf{H}_0^{-1}\right)} \tag{5.192}$$

5.2.10.4 Role of Assumptions The **Non-Singularity of Ω (Assumption 5.38)** was indispensable for defining the metric of the objective function and allowing the scaling of the error. **Differentiability (Assumption 5.36)** was essential for the linear expansion; without it, the estimator could not be approximated by a normal form.

5.2.11 Maximum Likelihood Estimator (MLE)

5.2.11.1 Motivation and Geometric Interpretation The **Maximum Likelihood Method (MLE)** solves the problem of estimating population parameters when the researcher assumes a complete functional form for the conditional or joint probability distribution of the data. While pure GMM is often used in semiparametric contexts (knowing only moment conditions about unknown distributions), MLE is the parametric reference estimator that, under correct specification, attains the **Cramér-Rao lower bound**, being the asymptotically most efficient estimator possible.

In the **GMM** context, MLE can be derived as a special case of **exact identification** ($L = K$), where the population moment conditions are not chosen arbitrarily but extracted directly from the adopted probabilistic structure: the moments are the components of the **score function** (the gradient vector of the log-likelihood).

Geometric Interpretation in $\mathbb{R}^K$:

Consider the log-likelihood function as a hyperbolic "mountain surface" projected over the compact parameter space of dimension K. The MLE estimator locates the **peak (global maximum)** of this surface. Geometrically, at this peak, the tangent plane is perfectly horizontal, meaning that the **gradient vector (score)** $\mathbf{s}(\mathbf{w}, \theta)$ is zero. The **Fisher Information Matrix** $\mathcal{I}(\theta)$ acts as a measure of curvature (the negative Hessian matrix) of this surface at the top: the sharper the mountain (greater negative curvature), the more information the data provide about the precise location of the parameter.

5.2.11.2 Assumptions Let $f(\mathbf{x}, \theta)$ be the probability density and $\theta \in \Theta$ the parameter vector.

Assumption 5.39: Correct Model Specification

- **Role:** Ensures that the assumed parametric density $f(\mathbf{x}, \theta)$ coincides with the true population density generating the data at $\theta = \theta_0$. It is necessary to ensure consistency and the attainment of maximum Fisher efficiency.
- **Counterexample:** If the actual data follow a Student's t distribution (heavy tails), but we model via MLE assuming a Normal distribution, the estimator may be inconsistent. In the case of Quasi-MLE (QMLE), it may retain consistency under certain conditions but will lose efficiency, requiring robust variance-covariance matrices.

Assumption 5.40: Identifiability (Uniqueness)

- **Role:** Ensures an injective correspondence between the parameter space and the family of distributions, such that $f(\mathbf{x}, \theta_1) \neq f(\mathbf{x}, \theta_2)$ if and only if $\theta_1 \neq \theta_2$.
- **Counterexample:** In a linear model where $y_i = (\theta_1 + \theta_2)x_i + \epsilon_i$, it is impossible to isolate the individual effects. The likelihood surface will exhibit a perfectly flat "ridge" (infinite maxima), resulting in a singular information matrix.

Assumption 5.41: Fisher Regularity Conditions

- **Role:** Allows the interchange between integration over the support and differentiation in the parameter space. It is indispensable for proving that the mathematical expectation of the score is identically zero and for establishing the equality of the information matrix.
- **Counterexample:** If the support of the distribution intrinsically depends on the parameter (for example, a Uniform distribution on the interval $[0, \theta]$), the density will not be continuously differentiable at the boundaries of the support, breaking the standard asymptotic normality with convergence rates different from $\sqrt{n}$.

5.2.11.3 Formal Statement and Derivation Let $\{\mathbf{x}_i\}_{i=1}^n$ be a simple random sample (i.i.d.) with population density given by $f(\mathbf{x}_i, \theta_0)$. We define:

1. **Likelihood Function:** $L(\theta) = \prod_{i=1}^n f(\mathbf{x}_i, \theta)$.
2. **Log-Likelihood:** $\ell(\theta) = \ln L(\theta) = \sum_{i=1}^n \ln f(\mathbf{x}_i, \theta)$.
3. **Score Function:** $\mathbf{s}(\mathbf{x}_i, \theta) = \nabla_\theta \ln f(\mathbf{x}_i, \theta)$ $(K \times 1)$.
4. **Fisher Information Matrix (Unit):** $\mathcal{I}(\theta) = E[\mathbf{s}(\mathbf{x}_i, \theta)\mathbf{s}(\mathbf{x}_i, \theta)'] = -E[\nabla^2_{\theta\theta'} \ln f(\mathbf{x}_i, \theta)]$.

Theorem (Asymptotic Normality of MLE). Under Fisher's regularity conditions, the maximum likelihood estimator $\hat{\theta}_n$ satisfies:

$$\boxed{\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N\left(\mathbf{0}, \mathcal{I}(\theta_0)^{-1}\right)} \tag{5.186}$$

Proof:

Step 1: Sample Moment Condition

MLE can be defined as the estimator that solves the system of sample moment equations $\mathbf{g}_n(\hat{\theta}_n) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{s}(\mathbf{x}_i, \hat{\theta}_n) = \mathbf{0}$. By the necessary condition for the existence of a maximum at an interior point:

$$\sum_{i=1}^{n} \mathbf{s}(\mathbf{x}_i, \hat{\theta}_n) = \mathbf{0} \tag{5.187}$$

Step 2: Score Property (Zero Expectation)

We prove that $E[\mathbf{s}(\mathbf{x}_i, \theta_0)] = \mathbf{0}$.

By the regularity assumption (Assumption 5.41), we interchange integral and derivative:

$$E[\mathbf{s}(\mathbf{x}_i, \theta_0)] = \int \frac{\nabla_\theta f(\mathbf{x}, \theta_0)}{f(\mathbf{x}, \theta_0)} f(\mathbf{x}, \theta_0) d\mathbf{x} = \nabla_\theta \int f(\mathbf{x}, \theta_0) d\mathbf{x}$$

Since $\int f(\mathbf{x}, \theta_0) d\mathbf{x} = 1$ over the entire support by definition of density:

$$E[\mathbf{s}(\mathbf{x}_i, \theta_0)] = \nabla_\theta(1) = \mathbf{0} \tag{5.188}$$

Step 3: Linear Expansion via the Mean Value Theorem

We expand the system of sample score equations (5.187) around the true parameter θ_0:

$$\mathbf{0} = \sum_{i=1}^{n} \mathbf{s}(\mathbf{x}_i, \theta_0) + \left[\sum_{i=1}^{n} \nabla_{\theta'} \mathbf{s}(\mathbf{x}_i, \bar{\theta})\right] (\hat{\theta}_n - \theta_0) \tag{5.189}$$

where $\bar{\theta}$ is a stochastic mean vector lying on the segment between $\hat{\theta}_n$ and θ_0.

Step 4: Scaling and Matrix Isolation

Multiplying expression (5.189) by $\frac{1}{\sqrt{n}}$ and isolating the scaled sample error:

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{s}(\mathbf{x}_i, \theta_0) = -\left[\frac{1}{n}\sum_{i=1}^{n} \nabla_{\theta'} \mathbf{s}(\mathbf{x}_i, \bar{\theta})\right] \sqrt{n}(\hat{\theta}_n - \theta_0) \tag{5.190}$$

Premultiplying by the inverse of the average sample Hessian matrix (guaranteed asymptotically by full rank):

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -\left[\frac{1}{n}\sum_{i=1}^{n} \nabla^2_{\theta\theta'} \ln f(\mathbf{x}_i, \bar{\theta})\right]^{-1} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{s}(\mathbf{x}_i, \theta_0)\right) \tag{5.191}$$

Step 5: Application of Limit Theorems

- **CLT:** Since the individual scores $\mathbf{s}(\mathbf{x}_i, \theta_0)$ are i.i.d. with zero expectation and covariance matrix $\mathcal{I}(\theta_0)$:

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{s}(\mathbf{x}_i, \theta_0) \xrightarrow{d} N\left(\mathbf{0}, \mathcal{I}(\theta_0)\right)$$

- **WLLN:** By the consistency of $\hat{\theta}_n \xrightarrow{p} \theta_0$ and continuity of the second derivatives:

$$\frac{1}{n}\sum_{i=1}^{n} \nabla^2_{\theta\theta'} \ln f(\mathbf{x}_i, \bar{\theta}) \xrightarrow{p} E\left[\nabla^2_{\theta\theta'} \ln f(\mathbf{x}_i, \theta_0)\right] = -\mathcal{I}(\theta_0) \tag{5.192}$$

Step 6: Final Result via Slutsky

Applying Slutsky's Theorem:

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} -\left[-\mathcal{I}(\theta_0)\right]^{-1} N\left(\mathbf{0}, \mathcal{I}(\theta_0)\right) \tag{5.193}$$

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N\left(\mathbf{0}, \mathcal{I}(\theta_0)^{-1}\right) \tag{5.194}$$

□

5.2.11.4 Role of Assumptions The **Regularity assumption (Assumption 5.41)** was indispensable for geometrically sustaining Fisher's identity ($\mathbf{S} = -\mathbf{G}$). Without this interchange of operators, the limiting variance of the score would not coincide with the negative population Hessian, forcing the appearance of a sandwich-type structure and breaking the direct asymptotic efficiency.

5.2.12 Logit Model

5.2.12.1 Motivation and Geometric Interpretation The **Logit** model solves the problem of modeling the conditional probability of occurrence of a strictly binary dependent variable ($y_i \in \{0, 1\}$) from a set of exogenous regressor vectors. The traditional Linear Probability Model (LPM), estimated via OLS, exhibits severe structural flaws: it generates predictions outside the unit interval $[0, 1]$ and is intrinsically heteroskedastic. Logit overcomes these constraints by mapping the linear index of regressors to the interval $[0, 1]$ using the logistic Cumulative Distribution Function (CDF).

In the unified **GMM** view, Logit represents a case of **exact identification** ($L = K$), where the population moment condition is defined by the score of a conditional Bernoulli distribution.

Geometric Interpretation in $\mathbb{R}^2$:

The observed data lie exclusively at the discrete limits $y = 0$ or $y = 1$. Logit fits an S-shaped sigmoid curve to approximate the conditional expectation $E[y_i|\mathbf{x}_i]$. From the GMM perspective, estimation seeks to rotate and translate this sigmoidal profile so that $\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{p}_i)\mathbf{x}_i = \mathbf{0}$.

5.2.12.2 Assumptions Let $\mathbf{x}_i \in \mathbb{R}^K$ be the vector of regressors and $y_i \in \{0, 1\}$ the observed binary response.

Assumption 5.42: Stochastic Independence (Random Sampling)

- **Role:** Allows the additive factorization of the log-likelihood as the sum of individual log-densities, enabling the application of the Law of Large Numbers over sums of sample moments.
- **Counterexample:** Under active temporal or spatial dependence, the simple Fisher covariance matrix would fail by ignoring cross-interactions, requiring robust variance corrections of the HAC type.

Assumption 5.43: Rank Identifiability (No Collinearity)

- **Role:** Requires that the population moment matrix of the regressors $\mathbf{Q}_{xx} = E[\mathbf{x}_i\mathbf{x}_i']$ have full rank K. Ensures that the Hessian matrix is strictly negative definite over the entire support, guaranteeing the existence of a unique stable global maximum.

- **Counterexample:** In the presence of perfect multicollinearity, the likelihood surface will exhibit an infinite flat valley. The Jacobian moment matrix will become singular, preventing analytical inversion and identification of the vector β.

Assumption 5.44: Logistic CDF Specification

- **Role:** Postulates that the true data generating process obeys the structural relationship $P(y_i = 1|\mathbf{x}_i) = \Lambda(\mathbf{x}_i'\beta_0)$, where $\Lambda(z) = e^z/(1+e^z)$. Fixes the exact functional form of the orthogonality conditions.
- **Counterexample:** If the actual generating process follows a Probit (Normal) distribution, Logit estimators will maintain directional consistency of signs, but the calculated marginal effects will exhibit systematic biases caused by the relatively heavier tails of the logistic distribution.

5.2.12.3 Formal Statement and Derivation Let $\{y_i, \mathbf{x}_i\}_{i=1}^n$ be an i.i.d. sample where $y_i|\mathbf{x}_i \sim$ Bernoulli(p_i) with probability parameterized by $p_i = \Lambda(\mathbf{x}_i'\beta_0)$. We define:

1. **Logistic Function:** $\Lambda(z) = \frac{e^z}{1+e^z}$.
2. **Derivative Identity:** $\frac{d\Lambda(z)}{dz} = \Lambda(z)(1-\Lambda(z))$.
3. **GMM Moment Vector:** $\mathbf{g}(\mathbf{w}_i, \beta) = \mathbf{x}_i(y_i - \Lambda(\mathbf{x}_i'\beta))$ $(K \times 1)$.

Theorem (Asymptotic Normality of Logit). The GMM/MLE estimator $\hat{\beta}_n$ obtained by solving $\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i(y_i - \Lambda(\mathbf{x}_i'\hat{\beta}_n)) = \mathbf{0}$ converges in the limiting distribution to:

$$\boxed{\sqrt{n}(\hat{\beta}_n - \beta_0) \xrightarrow{d} N\left(\mathbf{0}, \mathcal{I}(\beta_0)^{-1}\right)} \tag{5.195}$$

where the Fisher Information Matrix is expressed as $\mathcal{I}(\beta_0) = E\left[\Lambda(\mathbf{x}_i'\beta_0)(1-\Lambda(\mathbf{x}_i'\beta_0))\mathbf{x}_i\mathbf{x}_i'\right]$.

Proof:

Step 1: Extraction of the Moment via Bernoulli Log-Likelihood

The probability density function for a conditional observation is given by $f(y_i|\mathbf{x}_i;\beta) = p_i^{y_i}(1-p_i)^{1-y_i}$. Its respective individual log-likelihood is:

$$\ell_i(\beta) = y_i \ln(p_i) + (1-y_i)\ln(1-p_i) \tag{5.196}$$

Substituting the sigmoidal parameterization $p_i = \Lambda(\mathbf{x}_i'\beta)$ and differentiating:

$$\nabla_\beta \ell_i(\beta) = y_i \frac{1}{p_i}\nabla_\beta p_i - (1-y_i)\frac{1}{1-p_i}\nabla_\beta p_i \tag{5.197}$$

Employing $\nabla_\beta p_i = p_i(1-p_i)\mathbf{x}_i$:

$$\nabla_\beta \ell_i(\beta) = y_i \frac{p_i(1-p_i)\mathbf{x}_i}{p_i} - (1-y_i)\frac{p_i(1-p_i)\mathbf{x}_i}{1-p_i} \tag{5.198}$$

Performing the algebraic cancellations:

$$\mathbf{s}_i(\beta) = y_i(1-p_i)\mathbf{x}_i - (1-y_i)p_i\mathbf{x}_i = \mathbf{x}_i(y_i - p_i) = \mathbf{x}_i(y_i - \Lambda(\mathbf{x}_i'\beta)) \tag{5.199}$$

This constitutes the exact-identification GMM moment condition vector: $\mathbf{g}_i(\beta) = \mathbf{x}_i(y_i - \Lambda(\mathbf{x}_i'\beta))$.

Step 2: Verification of the Population Orthogonality Condition

We evaluate the conditional expectation of the constructed moment under the true parameter β_0:

$$E[\mathbf{g}_i(\beta_0)|\mathbf{x}_i] = \mathbf{x}_i\left(E[y_i|\mathbf{x}_i] - \Lambda(\mathbf{x}_i'\beta_0)\right)$$

Under Assumption 5.44, $E[y_i|\mathbf{x}_i] = \Lambda(\mathbf{x}_i'\beta_0)$. Therefore:

$$E[\mathbf{g}_i(\beta_0)|\mathbf{x}_i] = \mathbf{x}_i(\mathbf{0}) = \mathbf{0} \tag{5.200}$$

By direct application of the Law of Iterated Expectations:

$$E[\mathbf{g}_i(\beta_0)] = E[E[\mathbf{g}_i(\beta_0)|\mathbf{x}_i]] = \mathbf{0} \tag{5.201}$$

Step 3: Linearization and Expansion via MVT

We expand the system of sample normal equations via the Mean Value Theorem:

$$\mathbf{0} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{g}_i(\beta_0) + \left[\frac{1}{n}\sum_{i=1}^{n}\nabla_{\beta'}\mathbf{g}_i(\bar{\beta})\right]\sqrt{n}(\hat{\beta}_n - \beta_0) \tag{5.202}$$

We compute the sample Jacobian of the moment vector:

$$\nabla_{\beta'}\mathbf{g}_i(\beta) = -\Lambda(\mathbf{x}_i'\beta)(1 - \Lambda(\mathbf{x}_i'\beta))\mathbf{x}_i\mathbf{x}_i' \tag{5.203}$$

Step 4: Convergence and Variance Structure

- **WLLN:** The average sample Hessian matrix converges to $\mathbf{G}_0 = -\mathcal{I}(\beta_0)$.
- **CLT:** $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{g}_i(\beta_0) \xrightarrow{d} N(\mathbf{0}, \mathbf{S})$, where $\mathbf{S} = \mathcal{I}(\beta_0)$.

Step 5: Isolation and Limiting Mapping

$$\sqrt{n}(\hat{\beta}_n - \beta_0) \xrightarrow{d} \mathcal{I}(\beta_0)^{-1}N\left(\mathbf{0}, \mathcal{I}(\beta_0)\right) \tag{5.204}$$

$$\sqrt{n}(\hat{\beta}_n - \beta_0) \xrightarrow{d} N\left(\mathbf{0}, \mathcal{I}(\beta_0)^{-1}\right) \tag{5.205}$$

□

5.2.12.4 Role of Assumptions The premise of **Correct CDF Specification (Assumption 5.44)** was indispensable in Step 4. If the distribution were incorrect, the conditional equality $\text{Var}(y_i|\mathbf{x}_i) = p_i(1-p_i)$ would be violated, generating a mismatch between the moment covariance matrix $\mathbf{S}$ and the population Jacobian $\mathbf{G}_0$. The model would require the robust sandwich formulation (QMLE), under penalty of underestimating or overestimating standard errors.

5.2.13 Probit Model

5.2.13.1 Motivation and Geometric Interpretation The **Probit** model addresses the same problem of modeling binary discrete choices ($y_i \in \{0, 1\}$). Probit differs from the Logit model by mapping the linear index of exogenous regressors $\mathbf{x}_i'\beta$ to the probabilistic interval $[0, 1]$ through the Cumulative Distribution Function (CDF) of the Standard Normal distribution, denoted by $\Phi(\cdot)$.

Geometric Interpretation in $\mathbb{R}^K$:

In the vector space of explanatory variables, the Probit model establishes a **separation hyperplane** geometrically defined by the equation $\mathbf{x}_i'\beta = 0$. The conditional probability of success is represented by a **sigmoidal surface** that asymptotically extends between the ceiling and floor probabilistic boundaries. Estimation via GMM seeks to rotate and displace this separation hyperplane in the K-dimensional space of regressors so that the weighted score vector is zero in the sample.

5.2.13.2 Assumptions Let $\mathbf{x}_i \in \mathbb{R}^K$ be the vector of regressors and $y_i \in \{0, 1\}$ the discrete response.

Assumption 5.45: Identifiability (Full Rank of the Regressor Matrix)

- **Role:** Requires that the population moment matrix $E[\mathbf{x}_i \mathbf{x}_i']$ be strictly positive definite, ensuring linearly independent statistical information to discriminate the effects of each parametric component.
- **Counterexample:** If there is perfect collinearity among the components of the regressor vector, the Jacobian moment matrix will become singular, generating a linear trough of infinite solutions and making inversion for asymptotic variance calculation impossible.

Assumption 5.46: Gaussian Specification of the Latent Error

- **Role:** Assumes that the model is generated by a latent variable structure $y_i^* = \mathbf{x}_i'\beta_0 + \epsilon_i$, where the unobserved errors strictly follow the distribution $\epsilon_i \sim N(0, 1)$. Ensures that the moment conditions extracted from the score are centered at zero in the population.
- **Counterexample:** If the actual underlying disturbances follow asymmetric or heavy-tailed distributions (such as the Cauchy distribution), the use of the normal CDF $\Phi(\cdot)$ will cause a break in the probabilistic bridge, generating inconsistent and severely biased estimators.

Assumption 5.47: Strict Exogeneity of Regressors

- **Role:** Requires that the distribution of the latent error be independent of the regressors, implying the conditional condition $E[\epsilon_i|\mathbf{x}_i] = 0$.
- **Counterexample:** If one of the regressors is endogenous (such as omitting ability variables in educational choice models), the population orthogonality condition will fail because the latent error term will carry systematic information about $\mathbf{x}_i$, making the GMM/MLE estimator inconsistent.

5.2.13.3 Formal Statement and Derivation Let $\{y_i, \mathbf{x}_i\}_{i=1}^n$ be a simple random sample (i.i.d.) with $y_i \in \{0, 1\}$ and $\mathbf{x}_i \in \mathbb{R}^K$. We define $\Phi(z) = \int_{-\infty}^{z} \phi(t)dt$, where $\phi(t)$ represents the probability density function (PDF) of the standard normal.

Theorem (Asymptotic Normality of Probit). Under regularity conditions and correct Gaussian specification, the estimator $\hat{\beta}_n$ satisfies the convergence:

$$\boxed{\sqrt{n}(\hat{\beta}_n - \beta_0) \xrightarrow{d} N\left(\mathbf{0}, \mathcal{I}(\beta_0)^{-1}\right)} \tag{5.206}$$

where:

$$\mathcal{I}(\beta_0) = E\left[\frac{\phi(\mathbf{x}_i'\beta_0)^2}{\Phi(\mathbf{x}_i'\beta_0)(1 - \Phi(\mathbf{x}_i'\beta_0))}\mathbf{x}_i\mathbf{x}_i'\right] \tag{5.207}$$

Proof:

Step 1: Formulation of the Structural Score

The conditional density of Probit for an observation is expressed as:

$$f(y_i|\mathbf{x}_i, \beta) = [\Phi(\mathbf{x}_i'\beta)]^{y_i}[1 - \Phi(\mathbf{x}_i'\beta)]^{1-y_i} \tag{5.208}$$

Its individual log-likelihood is:

$$\ln L_i(\beta) = y_i \ln \Phi(\mathbf{x}_i'\beta) + (1 - y_i) \ln[1 - \Phi(\mathbf{x}_i'\beta)] \tag{5.209}$$

The score vector $\mathbf{s}_i(\beta)$ is:

$$\mathbf{s}_i(\beta) = \left[\frac{y_i\phi(\mathbf{x}_i'\beta)}{\Phi(\mathbf{x}_i'\beta)} - \frac{(1 - y_i)\phi(\mathbf{x}_i'\beta)}{1 - \Phi(\mathbf{x}_i'\beta)}\right]\mathbf{x}_i \tag{5.210}$$

Reducing to the common denominator:

$$\mathbf{s}_i(\beta) = \frac{[y_i - \Phi(\mathbf{x}_i'\beta)]\,\phi(\mathbf{x}_i'\beta)}{\Phi(\mathbf{x}_i'\beta)[1 - \Phi(\mathbf{x}_i'\beta)]}\mathbf{x}_i \tag{5.211}$$

Step 2: Numerical Equivalence via Just-Identified GMM

Since the number of moments generated by the score exactly matches the dimension of the parameter vector ($L = K$), the GMM estimator simply zeros the sample moment vector:

$$\mathbf{g}_n(\hat{\beta}_n) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{s}_i(\hat{\beta}_n) = \mathbf{0} \tag{5.212}$$

Since equation (5.212) represents exactly the First Order Conditions (FOC) of log-likelihood optimization, it follows that $\hat{\beta}_{GMM} = \hat{\beta}_{MLE}$ for any choice of positive definite weighting matrix $\mathbf{W}_n$.

Step 3: Linearization via the Mean Value Theorem

We perform the first-order linear expansion of $\mathbf{g}_n(\hat{\beta}_n)$ around the true point β_0:

$$\mathbf{0} = \mathbf{g}_n(\hat{\beta}_n) = \mathbf{g}_n(\beta_0) + \mathbf{G}_n(\bar{\beta})(\hat{\beta}_n - \beta_0) \tag{5.213}$$

where $\bar{\beta}$ lies between the estimator and the population target. Multiplying by $\sqrt{n}$ and isolating the sample error:

$$\sqrt{n}(\hat{\beta}_n - \beta_0) = -\left[\mathbf{G}_n(\bar{\beta})\right]^{-1}\sqrt{n}\mathbf{g}_n(\beta_0) \tag{5.214}$$

Step 4: Mapping of Stochastic Limits

We define the population limit Jacobian as $\mathbf{G} = E[\partial\mathbf{s}_i(\beta_0)/\partial\beta']$ and the limit variance of the moments as $\mathbf{S} = E[\mathbf{s}_i(\beta_0)\mathbf{s}_i(\beta_0)']$.

- **WLLN:** $\mathbf{G}_n(\bar{\beta}) \xrightarrow{p} \mathbf{G}$.
- **CLT:** $\sqrt{n}\mathbf{g}_n(\beta_0) \xrightarrow{d} N(\mathbf{0}, \mathbf{S})$.

Applying Slutsky's Theorem:

$$\sqrt{n}(\hat{\beta}_n - \beta_0) \xrightarrow{d} N\left(\mathbf{0},\, \mathbf{G}^{-1}\mathbf{S}(\mathbf{G}')^{-1}\right) \tag{5.215}$$

Step 5: Fisher Information Matrix Identity and Simplification

Since the moment conditions were derived from the analytical score of a perfectly specified density (Assumption 5.46), the Fisher Information Matrix Identity holds, establishing that $\mathbf{S} = -\mathbf{G} = \mathcal{I}(\beta_0)$. Substituting:

$$\text{Asy.Var}(\hat{\beta}_n) = \mathcal{I}(\beta_0)^{-1}\mathcal{I}(\beta_0)\mathcal{I}(\beta_0)^{-1} = \mathcal{I}(\beta_0)^{-1} \tag{5.216}$$

□

5.2.13.4 Role of Assumptions The assumption of **Rank Identifiability (Assumption 5.45)** is strictly indispensable for the validity of the estimator. If the regressors exhibited linear dependence, the population Jacobian matrix $\mathbf{G}$ would be singular. This would prevent the matrix inversion operation performed in Step 3, collapsing the derivation of the Gaussian limiting distribution.

5.3 Advanced Reference Guide: Unification of Estimators under GMM and M-Estimation

This reference guide analyzes the main econometric estimators through the lenses of the Generalized Method of Moments (GMM) and M-Estimation, unifying geometric, algebraic, and asymptotic intuition under a single framework.

5.3.1 Notation and Preliminary Definitions

To ensure rigorous consistency and avoid classical literature conflicts, the following strict symbol convention is adopted:

Symbol	Definition
$\mathbf{y}$	Vector of observations of the dependent variable $(n \times 1)$
$\mathbf{X}$	Matrix of regressors $(n \times k)$
$\mathbf{Z}$	Matrix of instruments $(n \times l)$
$\mathbf{u} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$	Vector of structural disturbances/errors
$\mathbf{g}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{g}_i(\boldsymbol{\theta})$	Sample moment function $(l \times 1)$. In the linear case: $\mathbf{g}_n(\boldsymbol{\beta}) = \frac{1}{n}\mathbf{Z}'\mathbf{u} = \frac{1}{n}\mathbf{Z}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$
$\mathbf{S} = E[\mathbf{g}_i(\boldsymbol{\theta})\mathbf{g}_i(\boldsymbol{\theta})']$	Population covariance matrix of the **moment conditions** $(l \times l)$. Consistent estimator: $\hat{\mathbf{S}}$
$\boldsymbol{\Omega} = E[\mathbf{u}\mathbf{u}' \mid \mathbf{X}]$	Population covariance matrix of the **model errors/disturbances** $(n \times n)$
$\mathbf{W}$	Positive definite stochastic weight matrix $(l \times l)$ used to weight moments in overidentified GMM
$\mathbf{P}_Z = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$	Orthogonal projection matrix onto Col($\mathbf{Z}$) (instrument *hat matrix*)

5.3.2 Comparative Table of Estimators

The following table organizes the main econometric estimators according to their formulation as a special case of GMM or M-Estimation, detailing moment conditions, identification structure, weight matrix, closed-form expression (when available), and asymptotic variance.

Table 7: Comparative Table of Estimators under GMM and M-Estimation

Estimator	Moment Condition	Identification	Matrices	Weight Matrix	Formula	Derivation	Variance
MM	$E[\mathbf{h}(\mathbf{y}_i, \boldsymbol{\theta})] = 0$	Exactly ($l = k$)	$\mathbf{X}$: Model; $\mathbf{Z}$: N/A	$\mathbf{W} = \mathbf{I}_k$	Implicit: $\mathbf{g}_n(\hat{\boldsymbol{\theta}}) = 0$	F.O.C. yields $\mathbf{g}_n = 0$; $\mathbf{W}$ irrelevant	$\frac{1}{n}\mathbf{G}_0^{-1}\mathbf{S}(\mathbf{G}_0')^{-1}$
OLS	$E[\mathbf{x}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta})] = 0$	Exactly ($l = k$)	$\mathbf{Z} = \mathbf{X}$	$\mathbf{W} = \mathbf{I}_k$	$\hat{\beta}_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$	$\frac{1}{n}\mathbf{X}'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = 0$	$\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$
IV	$E[\mathbf{z}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta})] = 0$	Exactly ($l = k$)	$\mathbf{Z}$: Instruments; $\mathbf{X}$: Regressors	$\mathbf{W} = \mathbf{I}_k$	$\hat{\beta}_{IV} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}$	$\frac{1}{n}\mathbf{Z}'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = 0$	$\sigma^2(\mathbf{Z}'\mathbf{X})^{-1}(\mathbf{Z}'\mathbf{Z})(\mathbf{X}'\mathbf{Z})^{-1}$
2SLS	$E[\mathbf{z}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta})] = 0$	Overid. ($l > k$)	$\mathbf{Z}$: All instr.; $\mathbf{X}$: Regressors	$\mathbf{W} = (\frac{1}{n}\mathbf{Z}'\mathbf{Z})^{-1}$	$\hat{\beta}_{2SLS} = (\mathbf{X}'\mathbf{P}_Z\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_Z\mathbf{y}$	Minimizes $(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{P}_Z(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$	$\sigma^2(\mathbf{X}'\mathbf{P}_Z\mathbf{X})^{-1}$
GMM (2-Step)	$E[\mathbf{z}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta})] = 0$	Overid. ($l > k$)	$\mathbf{Z}$: Instruments; $\mathbf{X}$: Regressors	$\mathbf{W} = \hat{\mathbf{S}}^{-1}$	$\hat{\beta}_{GMM} = (\mathbf{X}'\mathbf{Z}\hat{\mathbf{S}}^{-1}\mathbf{Z}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\hat{\mathbf{S}}^{-1}\mathbf{Z}'\mathbf{y}$	Minimizes $(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{Z}\mathbf{W}\mathbf{Z}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$	$(\mathbf{X}'\mathbf{Z}\mathbf{S}^{-1}\mathbf{Z}'\mathbf{X})^{-1}$
EHW	$E[\mathbf{x}_i u_i] = 0$	Exactly ($l = k$)	$\mathbf{Z} = \mathbf{X}$	$\mathbf{W} = \mathbf{I}_k$	$\hat{\beta}_{EHW} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$	Same as OLS	$(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\hat{\boldsymbol{\Omega}}_{diag}\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}$
GLS	$E[\mathbf{x}_i^*(y_i^* - \mathbf{x}_i^{*\prime}\boldsymbol{\beta})] = 0$	Exactly (transformed)	$\mathbf{X}^* = \boldsymbol{\Omega}^{-1/2}\mathbf{X}$; $\mathbf{y}^* = \boldsymbol{\Omega}^{-1/2}\mathbf{y}$	$\mathbf{W} = \mathbf{I}_k$	$\hat{\beta}_{GLS} = (\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{y}$	Minimizes $(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Omega}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$	$(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}$
FGLS	$E[\mathbf{x}_i u_i/\sigma_i^2] = 0$	Exactly Identified	Same as GLS with $\hat{\boldsymbol{\Omega}}$	$\mathbf{W} = \mathbf{I}_k$	$\hat{\beta}_{FGLS} = (\mathbf{X}'\hat{\boldsymbol{\Omega}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\boldsymbol{\Omega}}^{-1}\mathbf{y}$	Substitute $\boldsymbol{\Omega}$ with consistent estimator	$(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}$ (asymptotically)
3SLS	$E[\mathbf{Z} \otimes \varepsilon] = 0$	Overid. (System)	$\mathbf{Z}$: Block instr.; $\mathbf{X}$: Block-diagonal	$\mathbf{W} = (\hat{\boldsymbol{\Sigma}}^{-1} \otimes (\mathbf{Z}'\mathbf{Z})^{-1})$	$\hat{\beta}_{3SLS} = [\mathbf{X}'(\hat{\boldsymbol{\Sigma}}^{-1} \otimes \mathbf{P}_Z)\mathbf{X}]^{-1}\mathbf{X}'(\hat{\boldsymbol{\Sigma}}^{-1} \otimes \mathbf{P}_Z)\mathbf{y}$	GMM on simultaneous system	$[\mathbf{X}'(\boldsymbol{\Sigma}^{-1} \otimes \mathbf{P}_Z)\mathbf{X}]^{-1}$
GNLLS	$E[\nabla_{\boldsymbol{\beta}} h(\mathbf{x}_i, \boldsymbol{\beta}) \cdot \sigma_i^{-2} u_i] = 0$	Exactly Identified	$\mathbf{X}$: Non-linear; $\mathbf{Z} = \nabla_{\boldsymbol{\beta}} h$	$\mathbf{W} = \boldsymbol{\Omega}^{-1}$	Implicit: $\sum_i \nabla h(\mathbf{x}_i, \hat{\boldsymbol{\beta}})\sigma_i^{-2}[y_i - h(\mathbf{x}_i, \hat{\boldsymbol{\beta}})] = 0$	Minimizes $(\mathbf{y} - \mathbf{h})'\boldsymbol{\Omega}^{-1}(\mathbf{y} - \mathbf{h})$	$(\mathbf{G}_0'\boldsymbol{\Omega}^{-1}\mathbf{G}_0)^{-1}$
MLE	$E[\mathbf{s}(w_i, \boldsymbol{\theta})] = 0$	N/A (Score-Exact)	N/A (Based on density $f(y \mid \mathbf{x})$)	N/A	$\sum_i \mathbf{s}_i(\hat{\boldsymbol{\theta}}) = \sum_i \frac{\partial \ln f(y_i \mid \mathbf{x}_i, \hat{\boldsymbol{\theta}})}{\partial \hat{\boldsymbol{\theta}}} = 0$	Maximizes $\sum \ln f_i(\boldsymbol{\theta})$	$\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}_0)^{-1}$ (or $\mathbf{H}^{-1}\mathbf{J}\mathbf{H}^{-1}$ for QMLE)
Logit	$E[\mathbf{x}_i(y_i - \Lambda(\mathbf{x}_i'\boldsymbol{\beta}))] = 0$	N/A (M-Estimator)	$\mathbf{X}$: Logistic regressors	N/A	$\sum_i \mathbf{x}_i(y_i - \frac{\exp(\mathbf{x}_i'\hat{\boldsymbol{\beta}})}{1+\exp(\mathbf{x}_i'\hat{\boldsymbol{\beta}})}) = 0$	Bernoulli MLE with logistic CDF	$(\mathbf{X}'\mathbf{V}\mathbf{X})^{-1}$, where $V_{ii} = \Lambda_i(1 - \Lambda_i)$
Probit	Probit score = 0	N/A (M-Estimator)	$\mathbf{X}$: Normal CDF regressors	N/A	$\sum_i \mathbf{x}_i \frac{\phi(\mathbf{x}_i'\hat{\boldsymbol{\beta}})[y_i - \Phi(\mathbf{x}_i'\hat{\boldsymbol{\beta}})]}{\Phi(\mathbf{x}_i'\hat{\boldsymbol{\beta}})[1 - \Phi(\mathbf{x}_i'\hat{\boldsymbol{\beta}})]} = 0$	Bernoulli MLE with normal CDF	$(\mathbf{X}'\mathbf{V}\mathbf{X})^{-1}$, where $V_{ii} = \frac{\phi_i^2}{\Phi_i(1-\Phi_i)}$

5.3.3 Methodological Observations

1. **GMM as a Unifying Framework:** The Generalized Method of Moments (GMM) subsumes all listed estimators as special cases, depending on the choice of the weighting matrix $\mathbf{W}$ and the definition of the moment conditions. The table above explicitly shows this hierarchy, from traditional MM (exact identification) to optimal GMM (overidentification with efficient weighting).
2. **M-Estimation and the Score:** Estimators based on objective functions (MLE, Logit, Probit) can be reinterpreted as cases of M-Estimation where the moment condition is the gradient of the objective function (score). Under correct specification, the asymptotic variance simplifies to the inverse of the Fisher information matrix; under misspecification, the robust sandwich form (QMLE) is necessary.
3. **Identification and Weight Matrix:** The crucial distinction between exact identification ($l = k$) and overidentification ($l > k$) determines the relevance of the weight matrix $\mathbf{W}$. In exactly identified systems, $\mathbf{W}$ is irrelevant for point estimation; in overidentified systems, the choice of $\mathbf{W}$ affects both efficiency and (in finite samples) point estimation.
4. **Special Case of 2SLS:** 2SLS emerges as the one-step GMM with $\mathbf{W} = (\mathbf{Z}'\mathbf{Z})^{-1}$, which is the optimal weight matrix under homoskedasticity. Under heteroskedasticity, the two-step GMM with $\mathbf{W} = \hat{\mathbf{S}}^{-1}$ is asymptotically more efficient.
5. **Robustness and QMLE:** When parametric specification is doubtful (misspecification of density in MLE, or heteroskedasticity in GMM), the asymptotic variance assumes the sandwich form $\mathbf{H}^{-1}\mathbf{J}\mathbf{H}^{-1}$, ensuring valid inference even under more flexible conditions.
6. **GNLLS and Non-linearity:** In non-linear models, the Jacobian matrix $\mathbf{G}_0 = E[\nabla_{\boldsymbol{\beta}} h(\mathbf{x}_i, \boldsymbol{\beta}_0)]$ acts as the analog of instruments, and the asymptotic variance follows the structure $(\mathbf{G}_0'\boldsymbol{\Omega}^{-1}\mathbf{G}_0)^{-1}$, generalizing GLS to curved surfaces.