Topic 2: Ordinary Least Squares (OLS)

Magíster en Economía

Teoría Econométrica I (Econometric Theory I)

Prof. Luis Chancí

www.luischanci.com

Outline

1. **Preliminaries: Conditional Modeling**
   The CEF and regression fundamentals
2. **The Linear Regression Model (LRM)**
   Objective function and normal equations
3. **Geometry and FWL Theorem**
   Projection matrices and “partialling out”
4. **Finite-Sample Properties**
   Unbiasedness, Variance, and Gauss-Markov
5. **Constrained Least Squares (CLS)**
   Imposing linear restrictions
6. **Classical Inference & Prediction**
   Exact distributions, t/ tests, and forecasting

References

These slides closely follow the lecture notes I prepared for OLS, which are primarily based on:

- **Hansen, B. (2022).** Econometrics. Princeton University Press.
- **Davidson, R., & MacKinnon, J. G. (2004).** Econometric Theory and Methods. Oxford University Press.

> *Our goal is to look "under the hood" of OLS. We will separate the algebraic requirements needed to compute estimates from the statistical assumptions required for inference.*

1. Regression as Conditional Modeling

Conditional Expectation Function (CEF)

Let the data be denoted by , where

The joint density can always be factored into a conditional and a marginal component:

When we talk about **regression**, our main object of interest is the conditional component,

That is, regression seeks to characterize how the distribution of the outcome systematically varies with the covariates .

Conditional Expectation Function (CEF)

Rather than modeling the full conditional density, we can focus on one of its moments.

- The natural starting point is the conditional mean. In particular, the **Conditional Expectation Function (CEF)** is defined as
- This object plays a central role in econometrics because it summarizes how the average value of the outcome changes with the covariates.

Much of econometrics can be viewed as (i) estimating the CEF, or some feature of it, and (ii) asking what additional structure is needed to give that object a causal interpretation.

The Linear CEF

Now define the regression error as the deviation from the conditional mean:

> **Important: This decomposition is a *definition*, not an assumption. By construction, .**

A leading special case arises when the CEF is **linear**: .

This gives the **Linear Regression Model**:

Orthogonality and Its Implications

Because , the **Law of Iterated Expectations** implies

More generally, we obtain unconditional orthogonality: and in particular, .

This orthogonality condition is the foundation of **moment-based estimation**.

> If orthogonality fails, for example because of endogeneity arising from omitted variables, simultaneity, or measurement error, OLS will typically converge to the wrong object. Other estimators, such as IV and GMM, are designed for this case.

Statistical vs. Structural Models

- **Statistical Models**. Up to this point, everything has been a **statistical statement**. The decomposition describes conditional averages in the observable data, but it does not by itself have causal content.
- **Structural Models** are motivated by economic theory and are intended to represent behavioral, technological, or institutional relationships. This is where **identification** becomes central.
  - If a parameter is **identified**, then in principle it can be recovered from the observable distribution.
  - If it is **not identified**, then more data alone will not solve the problem; additional assumptions or structure are needed.
  - The empirical question is not only how to estimate , but also whether the object being estimated has the interpretation we want to give it.

2. The Linear Regression Model

Linearity

A natural and highly tractable way to model the Conditional Expectation Function is to approximate it with a linear function. This leads to the **Linear Regression Model**.

> **Assumption 1 — Linearity**

In matrix form, stacking all observations:

where , , , .

This linear representation is the starting point for the algebra of OLS and for the finite-sample and asymptotic results that follow later.

Interpretation of

The coefficient is usually interpreted as a **ceteris paribus effect**.

- The interpretation depends on the functional form of the model:

Model	Dependent variable	Regressor	Interpretation
Level–level			
Log–level			
Log–log			

- Also on the nature of the regressor. If is a dummy variable, measures the difference in the conditional mean between the group for which the condition holds.
  For example, if is used to indicate race, gender, or treatment status, then captures the corresponding conditional differential relative to the baseline category.

The OLS Objective Function

For observation , the model error is . After estimation, the sample analog, is called the **residual**.

To estimate , Ordinary Least Squares (OLS) chooses the parameter vector that minimizes the sum of squared residuals:

The logic is straightforward: squaring the residuals penalizes large errors and treats positive and negative deviations symmetrically.

Expanding the objective function gives .

The FOC

Taking derivatives with respect to and setting them equal to zero yields the first-order condition:

This first-order condition is the key step that leads to the normal equations.

Rearranging the first-order condition gives the fundamental **normal equations**:

But, to solve this system for a unique , we need an invertibility condition.

The OLS Estimator

> **Assumption 2 — Full Rank**
>
> The regressor matrix satisfies . Equivalently, there is no exact multicollinearity among the regressors.

Under Assumption 2, the matrix is symmetric positive definite and therefore invertible. Hence,

Two immediate implications are worth keeping in mind. (i) The second-order condition is , so the OLS solution is the unique global minimum. (ii) The normal equations also imply , which means that the residual vector is orthogonal to every column of the regressor matrix.

3. Geometry and FWL Theorem

Projection Matrices and

Once the OLS estimator has been obtained, the fitted values and residuals admit a useful geometric interpretation. The fitted values () and residuals () can be written compactly using two matrices:

- **Projection Matrix ():** Maps onto the column space of , .
- **Annihilator or residual-maker Matrix ():** Maps onto the orthogonal complement of .

So, OLS splits the observed outcome vector into two parts: .

Projection Matrices and

The matrices and summarize the key algebraic structure of OLS.

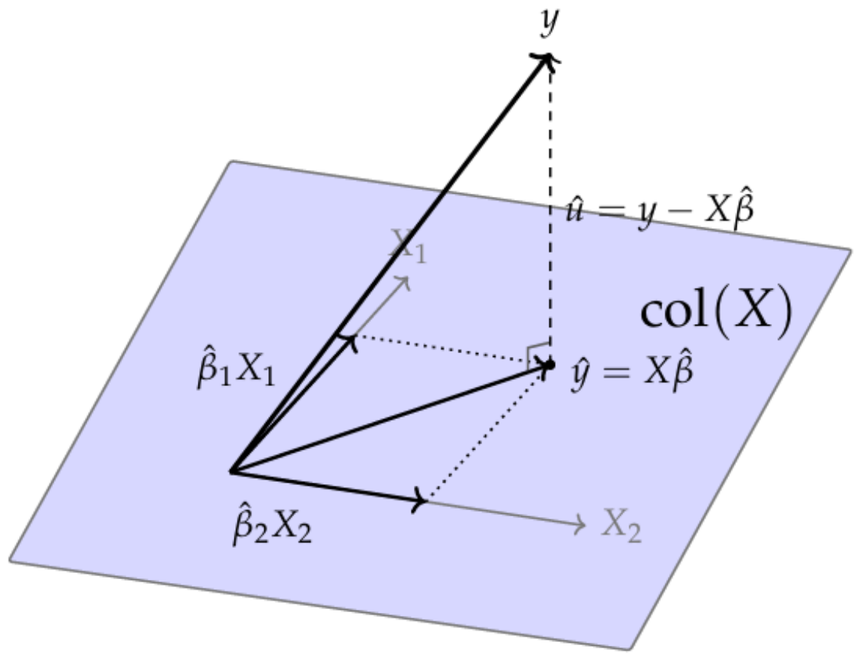
Main properties:

These results tell us that both matrices are **symmetric** and **idempotent**, and that they project onto orthogonal subspaces.

Geometric Interpretation

> **The geometric meaning of OLS: it chooses the point in that is closest to the observed vector .**

In the figure, the point is the orthogonal projection of onto the regressor space, and the residual vector is the gap between the observed outcome vector and its projection.



OLS **minimizes** — the squared distance from to .

Frisch–Waugh–Lovell (FWL) Theorem

The projection machinery developed above leads directly to one of the most useful algebraic results in linear regression: the **Frisch–Waugh–Lovell (FWL) theorem**.

FWL formalizes the concept of **partialling out**.

- Suppose we partition the regressor matrix as .
- Let contain the nuisance/control regressors, and contain the regressors of **primary interest**.
- Our objective is to estimate .

To do this, we first define the residual-maker matrix for :

Frisch–Waugh–Lovell (FWL) Theorem

> **Theorem — Frisch–Waugh–Lovell (FWL)**
>
> Partition . The OLS estimate of from the **full regression** of on equals:
>
> Moreover, the **residuals** from the full and auxiliary regressions are **identical**.

where, and are residualized variables, and .

> This theorem says that the coefficient on can be obtained either from the full regression or from a regression in which the linear influence of has first been removed from both and .

Decomposition of Variation and

The projection results developed above also give rise to one of the most familiar summaries of regression fit.

Since , it follows that

This is the basic decomposition of variation in linear regression.

> The key point is that OLS decomposes the observed **variation** in into a part captured by the linear model and a part left in the residuals.

TSS is the total variation in the dependent variable, with degrees of freedom; **ESS** is the variation accounted for by the regressors, with degrees of freedom when an intercept is included; and **SSR** is the unexplained variation, or residual variation, with degrees of freedom

and Adjusted

This decomposition motivates the **coefficient of determination**,

Thus, measures the fraction of the sample variation in that is accounted for by the linear regression model.

To account for the loss of degrees of freedom when more regressors are added:

Adding an uninformative regressor may reduce . For this reason, adjusted is often more informative than when comparing models with different numbers of regressors.

4. Finite-Sample Properties of OLS

Unbiasedness

Having established the algebraic mechanics of OLS, we now evaluate its statistical properties.

Is a reliable estimator of the true population parameter ? We begin by analyzing its finite-sample bias.

Recall the OLS estimator: . Substituting the true data-generating process, , into this expression yields:

Taking expectations conditional on the regressor matrix , and treating as fixed given itself, we obtain:

Unbiasedness

For OLS to be completely unbiased, the second term in the previous equation, , must vanish.

> **Assumption 3 — Zero Conditional Mean**
>
> That is, conditional on the full regressor matrix, the unobserved error term has a mean of exactly zero.

Note: This is arguably the most critical assumption in microeconometrics/applied econometrics (your next course). This assumption states that the regressors carry no systematic information about the unobserved shocks. So, if this assumption fails (whether due to omitted variables, measurement error, or simultaneous equations), we enter the domain of **endogeneity**. In this case, the second term in our derivation no longer evaluates to zero, making OLS fundamentally biased. Resolving this requires alternative identification strategies, such as Instrumental Variables.

Unbiasedness and the Variance of OLS

Under Assumptions 1, 2, and 3, OLS is conditionally unbiased:

This is a reassuring repeated-sampling property: across many hypothetical samples, the OLS estimator is perfectly centered at the true population parameter.

Unbiasedness alone tells us nothing about **precision**. An estimator can be unbiased but highly dispersed, making any single estimate unreliable. Therefore, the next crucial object of interest is the **variance-covariance matrix** of OLS.

The Variance of OLS

By definition, the conditional variance of the estimator is the expected outer product of its deviation from the mean, .

Recall from our previous derivation that the estimation error is . Also, we are conditioning on , so the matrices involving are treated as non-stochastic constants. Therefore,

- This is known as **the "sandwich" form**: The "Bread", , is driven entirely by the observed regressors, and the the "Meat", , is the covariance matrix of the unobserved errors.
- To obtain the classical, mathematically tractable textbook expression (and to prove the Gauss-Markov theorem), we must introduce an additional, highly restrictive assumption about the "meat".

Assumption 4: Spherical Disturbances

> **Assumption 4 — Spherical Disturbances**
>
> Equivalently, conditional on , the disturbances are **homoskedastic** and **mutually uncorrelated**.

Under this assumption, the general sandwich formula collapses to

- This is the familiar variance formula from the classical linear regression model.
- The usefulness of Assumption 4 is that it gives a simple closed-form expression for the variance of OLS. The drawback is that it is often too restrictive in empirical work.

Variance of OLS and Practical Extensions

Under Assumptions 1–4, we obtain the exact finite-sample variance:

In empirical practice, however, the spherical-disturbance assumption (A4) frequently fails. When it does, we rely on robust "sandwich" estimators:

- **White's heteroskedasticity-robust estimator**:
- **Cluster-robust estimator** (for within-group dependence):

> **Notes: Unlike the classical formula, these robust estimators are not exact in finite samples. They are justified entirely by large-sample asymptotic theory (which we will cover later).**

Estimating

To make the classical variance formula feasible, we need an estimator for the unknown disturbance variance .

The standard choice is based on the residual sum of squares:

- We divide by rather than . Because OLS chooses to minimize the sum of squared residuals, is mechanically smaller than the true unobserved . This adjustment perfectly corrects for that downward bias.
- Under Assumptions 1–4, , so is an unbiased estimator of the disturbance variance (homework).
- Therefore, the feasible variance estimator of OLS is .
- In practice, the reported standard errors are simply the square roots of the diagonal elements of this estimated matrix:

Gauss–Markov Theorem

Theorem — Gauss–Markov (BLUE)

Under Assumptions 1–4, the OLS estimator is **BLUE**: the **Best Linear Unbiased Estimator**.

That is, for any other linear unbiased estimator (where is a matrix function of), the difference in their variance-covariance matrices is positive semi-definite:

This guarantees that OLS is not just unbiased—it is the **most efficient** estimator within its class, providing the tightest possible sampling distribution.

Important: The theorem does not state that OLS is optimal among *all* possible estimators. It claims optimality only within the restricted class of estimators that are both linear in and unbiased. If a researcher is willing to accept a small amount of bias (e.g., Ridge regression) or use nonlinear procedures, they may achieve a strictly lower Mean Squared Error (MSE).

Proof Sketch: Gauss–Markov Theorem

(Please refer to the Lecture Notes for the complete proof).

1. Consider any arbitrary linear estimator, .
2. We can always express its weight matrix as the OLS weights plus some deviation matrix :
3. For to be **unbiased**, we must have . This requires . Substituting our expression for yields:
4. Now, compute the conditional variance. Because , the cross-terms vanish perfectly ():
5. Any matrix of the form is positive semi-definite (). Therefore, it mathematically follows that:
6. **Conclusion:** OLS strictly has the smallest variance among all linear unbiased estimators.

5. Constrained Least Squares (CLS)

Imposing Linear Restrictions

Up to this point, we have derived the OLS estimator unconditionally, allowing the data to freely dictate the parameter estimates.

In many empirical applications, however, economic theory suggests that the parameters must satisfy exact linear relationships. Common examples include:

- **Constant returns to scale** in a Cobb-Douglas production function (e.g.,).
- **Adding-up restrictions** in demand systems (budget shares summing to one).
- Exclusion restrictions (forcing certain coefficients to be exactly zero).

Suppose that our parameter vector is subject to linear restrictions:

CLS: Setup and Lagrangian

We now want to minimize the Sum of Squared Residuals (SSR) strictly over the subset of parameter values that satisfy these theoretical restrictions.

To solve this constrained optimization problem, we set up a Lagrangian: where is the vector of shadow prices (Lagrange multipliers) associated with the constraints.

Solving the first-order conditions with respect to both and yields the closed-form CLS estimator:

Note: The formula is often written with a plus sign by reversing the final term to .

Properties of CLS

- The algebraic intuition is: it starts with the unrestricted and adjusts it by the exact minimum distance required to force the restrictions to hold.
- If the unrestricted estimator happens to already satisfy the rule (), the adjustment term evaluates to zero and .
- The statistical properties depend entirely on whether the imposed theory is true:
  - **If the restrictions are CORRECT:** (i) CLS remains perfectly **unbiased**; (ii) CLS is **more efficient** than unrestricted OLS (its variance is smaller in the positive semi-definite matrix sense).
  - **If the restrictions are FALSE:** (i) CLS becomes fundamentally **biased**; (ii) We face a harsh bias-variance tradeoff: we gain mechanical efficiency (smaller variance) but force the estimator to converge to the wrong parameter values, causing misspecification.
- Because imposing false restrictions damages the consistency of our estimator, we must formally test whether the data supports the theory. This leads us directly to the -test.

6. Inference in OLS

Assumption 5: Normality

Up to this point, the Gauss–Markov theorem guaranteed that OLS is BLUE relying only on assumptions about the conditional mean and variance.

However, obtaining a point estimate is not enough. We are also interested in inference. To conduct hypothesis tests and compute exact tail probabilities in finite samples, we must fully specify the data-generating process.

To conduct exact finite-sample inference, we add a distributional assumption to the classical linear model:

> **Assumption 5 — Normality**

Distributional Results

The normal linear regression model is theoretically elegant because the linear algebra of OLS maps directly into standard probability distributions.

Under Assumptions 1–5:

(i)

(ii)

(iii) and are **independent** conditional on .

Distributional Results

Why are these results true?

- **(i)** Under normality, the estimator is strictly a linear transformation of a normal vector, which is always normally distributed.
- **(ii)** The residual sum of squares can be written as a quadratic form in the idempotent matrix : . Because has rank , this maps to the sum of squared independent standard normal variables.
- **(iii)** The estimator depends on through , while the residuals depend on through . Since , these components are orthogonal; under joint normality, zero covariance (orthogonality) guarantees strict independence.

Homework: Prove (ii) and (iii) rigorously using the spectral decomposition of the idempotent matrix .

The -Test

Suppose we want to test a null hypothesis on a **single coefficient**:

From the corollary above, we know the exact distribution of that specific coefficient:

- If the population variance were known, we could construct a standard normal -statistic. Because it is unknown, we replace it with our unbiased estimator .
- That replacement matters because is itself a random variable. The feasible statistic becomes the ratio of a standard normal variable to the square root of an independent variable (divided by its degrees of freedom). This exact construction defines a Student- distribution.

The -Test

Therefore, the feasible test statistic is:

The decision rule is to reject at significance level if .

A particularly common application is testing , which determines whether regressor has a statistically discernible effect distinct from zero.

The -Test and Wald Test

Now consider testing multiple restrictions simultaneously with a **joint null hypothesis**: where is a matrix of rank (matching our setup from Constrained Least Squares).

The -test asks whether imposing these linear restrictions makes the model fit the data substantially worse. The classical -statistic compares the restricted and unrestricted Sum of Squared Residuals (SSR):

The -Test and Wald Test

An equivalent formulation uses only the unrestricted estimates. This is the **Wald form**:

This statistic is numerically identical to the SSR-based -statistic. ::: {.vspace} :::

The interpretation is intuitive: it measures how far the unrestricted estimates are from the hypothesized values , weighted by the precision (variance-covariance) of those estimates.

Asymptotic Connection

Under Assumption 5 (Normality), we know that exactly:

It is theoretically useful to define the corresponding Wald statistic (without dividing by):

In large samples, even if the normality assumption fails, the Central Limit Theorem ensures that:

This is why the Wald principle extends so naturally to asymptotic inference later in the course.

The -Value

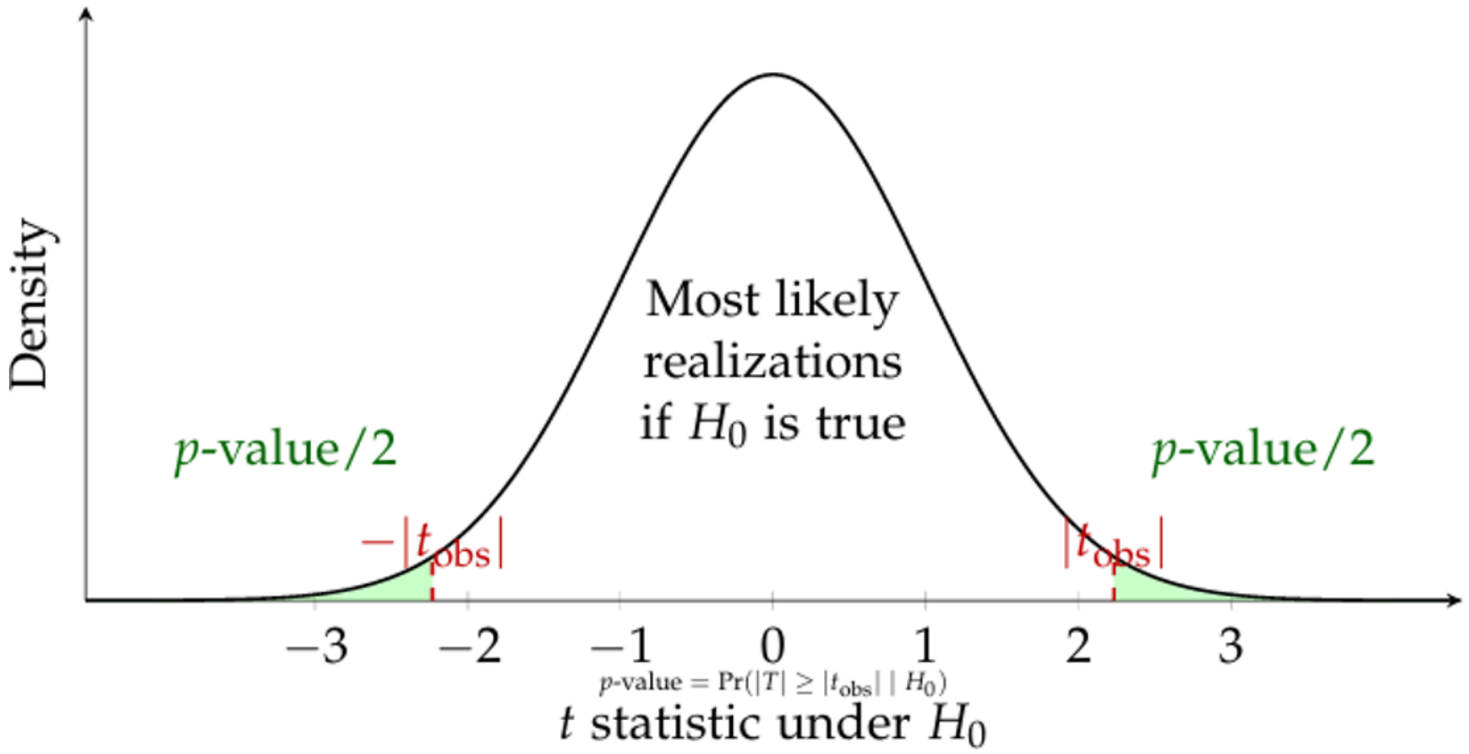
The previous slides presented hypothesis testing in its critical-value form.

The -value expresses the exact same logic from a different angle.

Observed significance level: The -value is the probability of obtaining a test statistic at least as extreme as the one actually observed in the sample:

For the two-sided -test, this becomes:

The -Value



Equivalently, the decision rule simplifies to:

7. OLS and Gaussian Quasi-MLE

Connection to Maximum Likelihood

- Under Assumption 5 (Normality), the conditional density of given is strictly normal.
- We will formally study **Maximum Likelihood Estimation (MLE)** later in the course. But, by now, we can take advantage of our current assumption.
- Let's introduce a fundamental mathematical object: the **conditional log-likelihood function** for the entire sample of independent observations.
- Because the observations are independent, the joint log-likelihood is simply the sum of the individual log-densities. In matrix notation, this is:

The Conditional Log-Likelihood

- Notice that the final term contains the exact OLS objective function, .
- We can elegantly rewrite the log-likelihood as:

> Because the variance is strictly positive, maximizing this log-likelihood with respect to is mathematically identical to minimizing the sum of squared residuals . Therefore, **under normality, OLS and MLE are exactly the same estimator for .**

Gaussian Quasi-MLE (QMLE)

If we incorrectly assume normality (Assumption 5 fails) and maximize the Gaussian likelihood function anyway, this is known as **Gaussian Quasi-Maximum Likelihood Estimation (QMLE)**.

Because the first-order conditions for in the Gaussian likelihood depend only on the linear residuals, maximizing this “wrong” likelihood still perfectly yields the OLS estimator for .

This is a profound theoretical result:

- The consistency of the OLS estimator for the conditional mean does not depend on the normality assumption.
- This establishes the foundation for Quasi-Likelihood methods later in the course.
  For example, in count data, maximizing a Poisson likelihood (even if the true data is not Poisson-distributed) will still yield consistent parameter estimates as long as the conditional mean is correctly specified.

8. Prediction and Forecast Evaluation

Plug-in Predictor and Uncertainty

Given a newly observed covariate vector , the natural out-of-sample predictor simply replaces the unknown parameters with our OLS estimates:

When forecasting, we must clearly distinguish between two different prediction targets, each with different sources of uncertainty:

1. **Targeting the Conditional Mean:**
   - *Source of error:* Purely estimation uncertainty (because).
2. **Targeting the Realized Outcome:**
   - *Source of error:* Estimation uncertainty PLUS the irreducible future shock ().

Mean Squared Forecast Error (MSFE)

To quantify the precision of our prediction for the realized outcome (under the assumption of homoskedasticity), we compute the MSFE:

This decomposes the forecast variance into two parts: (i) , The irreducible fundamental uncertainty from the future shock (no matter how much data we have, this remains); (ii) , which is the estimation uncertainty.

Notice that the estimation uncertainty depends on a quadratic form of . This means that predicting outcomes for covariate profiles that are very far from the historical sample average will mechanically result in much larger forecast standard errors.

Forecast Evaluation Metrics

Once out-of-sample predictions are generated, how do we evaluate model performance over hold-out periods?

Common **accuracy measures**:

Measure	Formula	Notes
RMSE		Scale-dependent; quadratically penalizes large errors.
MAE		Scale-dependent; penalizes linearly, more robust to outliers.
Theil's	Normalized RMSE ratio	Dimensionless; often compares the model against a naive "random walk" baseline.

Summary

Key Assumptions: A Cumulative Foundation

The properties of OLS build sequentially. Each additional assumption buys us a stronger statistical guarantee.

Assumption	Content	Enables
A1: Linearity		OLS can be formulated algebraically
A2: Full Rank		Unique closed-form OLS solution
A3: Zero Cond. Mean		Finite-sample unbiasedness
A4: Spherical Errors		Gauss–Markov (BLUE);
A5: Normality		Exact and distributions

Key Results: Mechanics and Geometry

The OLS estimator:

Geometric tools:

- Projection Matrix:
- Annihilator Matrix:
- Orthogonal Decomposition: , ,

Frisch-Waugh-Lovell (FWL) Theorem: Partialling out control variables from both and (using the annihilator) is mathematically equivalent to the full regression. The estimated coefficient on remains exactly the same.

Key Results: Statistical Properties

Finite-sample properties (Relying strictly on Assumptions 1–4):

- **Unbiasedness**:
- **Variance**:
- **BLUE**: OLS is the most efficient estimator in its class (Gauss–Markov Theorem).

Exact Inference (Adding Assumption 5 - Normality):

- The estimator is normally distributed:
- This permits exact -tests (for single coefficients) and -tests (for joint linear restrictions) in finite samples.