Ordinary Least Squares (OLS)

Review, Inference, Constraints, Prediction, and Related Models

by

Luis Chancí[1]

1 Introduction

These notes begin our study of the econometric machinery developed throughout the course. As emphasized in the syllabus, the goal is not only to use estimators but to look "under the hood" and understand how they are constructed, what they estimate, and which assumptions justify their statistical properties. Ordinary Least Squares (OLS) is the natural starting point for this task, serving as the workhorse benchmark for the field. Many of the core concepts that appear later in Instrumental Variables (IV), Generalized Method of Moments (GMM), and Maximum Likelihood Estimation (MLE) are already present here in their most intuitive form: optimization, orthogonality, identification, and sampling uncertainty.

This set of notes is written with a deliberately gradual strategy that follows the spirit of Wooldridge: assumptions are introduced only when they become necessary for a particular result, rather than being imposed all at once at the outset. This approach makes it easier to distinguish between the algebraic requirements needed to define the estimator, the statistical conditions required to establish properties such as unbiasedness or efficiency, and the distributional assumptions needed for exact finite-sample inference. From a pedagogical perspective, OLS is especially useful because it allows us to separate clearly mechanical derivation from statistical justification.

[1]Contact me at: luis.chanci@usach.cl, luischanci@santotomas.cl, or lchanci1@binghamton.edu. These notes mirror my natural thought process for teaching this material, blending the accessible style of Wooldridge or Angrist with the formal rigor of Hansen or MacKinnon. I greatly appreciate the professors who influenced my perspective, especially S. Kumbhakar (State University of New York at Binghamton) and R. Chumacero (Universidad de Chile). In the spirit of transparency, I used AI help (like Gemini) to check spelling and grammar. This version: 2026.

The exposition begins by framing regression as conditional modeling and by clarifying the link between the Conditional Expectation Function (CEF) and the linear regression model. This opening section draws on the style of introductory discussions found in texts such as Angrist and Pischke and Cameron and Trivedi, which are useful for establishing some of the basic language and concepts that will recur throughout the set of notes when discussing estimators and econometric methods. After deriving the OLS estimator from its optimization problem and studying its representation through the normal equations, we examine the geometry of least squares using projection matrices. This provides a transparent way to understand fitted values, residuals, and the Frisch–Waugh–Lovell theorem before turning to the estimator's finite-sample properties. In particular, we study the conditions required for unbiasedness, derive the variance-covariance matrix of the OLS estimator, and prove the Gauss–Markov result.

Once the basic structure is established, we consider useful variations within the same framework, including Constrained Least Squares (CLS), and then proceed to finite-sample inference under normality to cover standard t and F tests. We also include a brief preview of robust inference, while leaving the full treatment of asymptotic methods for a later set of notes. Before moving to more general estimators later in the semester, it is essential to understand why OLS takes the form it does, why it works when it works, and exactly where the machinery breaks when its assumptions are weakened.

The structure of these notes is as follows. Section 2 introduces regression as a problem of conditional modeling and motivates the population objects that underlie the analysis. Section 3 presents the Linear Regression Model (LRM), derives the OLS estimator from its objective function, and develops its algebraic and geometric properties through the normal equations, projection matrices, the Frisch–Waugh–Lovell theorem, and the decomposition of variation underlying R^2. It then studies the finite-sample properties of the estimator. Section 4 extends the framework to constrained least squares. Section 5 develops classical finite-sample inference, including t tests and F/Wald tests, while also briefly previewing robust inference under weaker assumptions. Section 6 introduces the connection between OLS and Gaussian quasi-maximum likelihood. Section 7 concludes with prediction and forecast evaluation. Key references for these notes include Hansen (2022) and Davidson & MacKinnon (2004).

2 Preliminaries: Regression as Conditional Modeling

Having motivated OLS as the first piece of the econometric machinery studied in this course, we now turn to the population object that gives regression its meaning. Before deriving estimators or imposing distributional assumptions, it is essential to clarify what regression seeks to learn from the data. The starting point is *conditional modeling*: we observe outcomes and covariates jointly and seek to characterize how the distribution of the outcome varies with the covariates. This setup provides the conceptual bridge between the observed data, the parameters of interest, and the assumptions that will later justify estimation and inference.

A central theme in modern econometrics is *identification*. Informally, identification asks whether the data-generating process, together with our maintained assumptions, contains enough information to determine the population parameter of interest. If a parameter is identified, then, at least in principle, we can learn about it from a sample. If it is not identified, no amount of additional data will resolve the ambiguity unless we strengthen the model or impose additional assumptions.

A natural starting point is the *Conditional Expectation Function* (CEF), $\mathbb{E}[y_i \mid x_i]$, which lies at the core of regression analysis. From a statistical perspective, the CEF summarizes how the average value of the outcome varies with the covariates. Much of econometrics can be viewed as: (i) estimating the CEF, or some feature of it; and then (ii) asking what additional structure is needed to interpret that object in causal or structural terms.

To put these concepts into a formal framework, let the observational data be a sample $\{w_i\}_{i=1}^n$, where $w_i = (y_i, x_i)$, $y_i \in \mathbb{R}$ is the outcome of interest, and $x_i \in \mathbb{R}^k$ is a vector of observed covariates. Suppose their joint distribution admits a density or probability mass function indexed by parameters θ, written as $f(y_i, x_i; \theta)$. By the rules of probability, the joint distribution can always be factored into a conditional and a marginal component:

$$f(y_i, x_i; \theta) = f(y_i \mid x_i; \theta_1)\, f(x_i; \theta_2).$$

Regression analysis focuses primarily on the conditional component, $f(y_i \mid x_i; \theta_1)$, because this is the part of the joint distribution that describes how the outcome varies with the covariates. In many applications, the marginal distribution of x_i can be treated as secondary for inference on θ_1.[2]

Rather than modeling the full conditional distribution from the start, econometric

[2] In more advanced treatments, this separation is related to weak exogeneity for inference on the parameters governing the conditional distribution.

analysis often focuses first on its conditional mean:

$$m(x_i) \equiv \mathbb{E}[y_i \mid x_i].$$

Define the regression error as the deviation from that conditional mean:

$$u_i \equiv y_i - m(x_i).$$

Then we can always write

$$y_i = m(x_i) + u_i.$$

Here it is important to highlight that this decomposition is a definition, not an assumption. Moreover, because u_i is defined as the difference between y_i and its conditional mean, it follows immediately that

$$\mathbb{E}[u_i \mid x_i] = 0.$$

By the Law of Iterated Expectations (LIE), this also implies $\mathbb{E}[u_i] = 0$. More generally, for any measurable function $h(\cdot)$ such that the expectation exists,[3]

$$\mathbb{E}[h(x_i)u_i] = 0.$$

This orthogonality property plays a central role in econometrics. It is the foundation for *moment-based* estimation methods, which exploit population restrictions of this form and replace them with their sample analogs. Later in the course, we will return to this idea when studying the Method of Moments (MM) and the Generalized Method of Moments (GMM).

At this stage, it is useful to distinguish between *statistical* and *structural* statements. The decomposition

$$y_i = m(x_i) + u_i$$

is purely statistical: it describes conditional averages in the observable data, but by itself it does not deliver a causal interpretation. A structural model, in contrast, is motivated by economic theory and is intended to represent behavioral, technological, or institutional relationships. Identification becomes economically meaningful when we ask whether the parameters of such a model can be recovered from the observable distribution, and under which assumptions.

A leading special case, and the one that motivates these notes, is the assumption

[3] These statements require mild regularity conditions, such as integrability, to ensure that the relevant expectations are well defined.

that the CEF is linear:

$$m(x_i) = x_i'\beta.$$

This is a substantive modeling restriction, not a purely algebraic rewrite. If it holds, then the model can be written as

$$y_i = x_i'\beta + u_i,$$

where, by construction,

$$\mathbb{E}[u_i \mid x_i] = 0.$$

This is the linear regression model. In this case, β summarizes how the conditional mean of y_i varies with the covariates.

The zero conditional mean restriction also implies the unconditional moment condition

$$\mathbb{E}[x_i u_i] = 0.$$

This implication reveals an important hierarchy:

$$\mathbb{E}[u_i \mid x_i] = 0 \quad \Longrightarrow \quad \mathbb{E}[x_i u_i] = 0,$$

but the converse is generally false. Thus, simple orthogonality between regressors and errors is weaker than a zero conditional mean restriction. This distinction is fundamental in econometrics because OLS is closely tied to orthogonality conditions, while more general estimators such as IV and GMM are built around broader moment restrictions.

In particular, when the linear model is correctly specified and the relevant orthogonality conditions hold, OLS is a natural estimator because it enforces in the sample the same type of orthogonality that characterizes the population problem. When these conditions fail (for example, because of omitted variables, simultaneity, or measurement error), OLS generally does not identify the parameter of interest and is typically inconsistent. In such cases, we must look for alternative sources of orthogonality. Instrumental Variables (IV) and the Generalized Method of Moments (GMM) do exactly this by using restrictions of the form

$$\mathbb{E}[z_i u_i] = 0,$$

where the instruments z_i need not coincide with x_i. We will return to this point later in the course.

In this sense, OLS is not only the benchmark estimator for linear regression; it is

also the simplest member of a much broader class of moment-based estimators. For that reason, we begin by studying it carefully, using the linear model as a foundation for the more general econometric methods developed later in the course.

The next section takes this linear CEF as a statistical approximation and shows how OLS is used to estimate it from the data before any structural interpretation is imposed.

3 The Linear Regression Model (LRM)

The previous section identified the Conditional Expectation Function as the central population object in regression analysis. We now introduce OLS as the tool used to estimate a linear approximation to that object from the data. A natural and highly tractable way to model the Conditional Expectation Function (CEF) is to approximate it with a linear function. In this approach, the conditional mean of y_i is written as linear in an unknown parameter vector $\beta \in \mathbb{R}^k$ that must be estimated from the data. This leads to the familiar Linear Regression Model (LRM):

$$y_i = x_i'\beta + u_i.$$

Textbooks often introduce this representation as the first core assumption of the model.

Assumption 1 (Linearity)**.** $y_i = x_i'\beta + u_i$ for $i = 1, \ldots, N$.

As noted earlier, one may also view linearity not as a literal assumption about the true conditional mean, but as a working approximation to an unknown function.

A brief note on the interpretation of the coefficients in the linear model is useful at this stage. In empirical work, the coefficients in β are often interpreted as *ceteris paribus* effects: they describe how the outcome changes with one regressor, holding the remaining covariates fixed. The exact interpretation depends on the functional form. For instance, a log–log specification yields elasticities, while a log–level specification yields semi-elasticities.

Once the model is specified, the next question is how to estimate β from a sample of N observations. The basic idea behind **Ordinary Least Squares (OLS)** is straightforward: we seek the parameter vector that makes the fitted values $x_i'\beta$ as close as possible to the observed outcomes y_i. The discrepancy

$$u_i = y_i - x_i'\beta$$

is the model error for observation i. Let $\hat{\beta}$ denote the parameter vector obtained from the estimation. Then, after estimation, the sample analog

$$\hat{u}_i = y_i - x_i'\hat{\beta}$$

is called the residual.

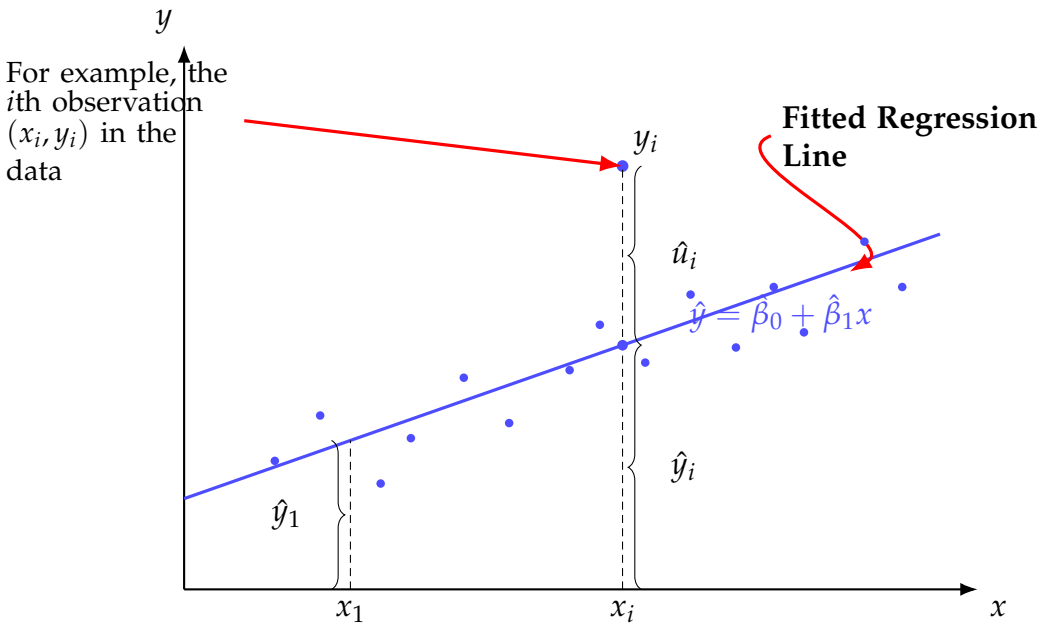

Figure 1: Illustration of fitted values and residuals in a simple linear regression.

Since we have a sample of N observations, we need a single criterion to summarize the overall fit of the model. One possible idea would be to sum the errors, $\sum_i u_i$, but this is not useful as a fitting criterion because large positive and negative discrepancies can cancel each other out. For example, the vectors $u = (1, -1, 2, -2)'$ and $u = (10, -10, 30, -30)'$ both sum to zero, even though the second clearly reflects a much poorer fit.

A natural solution is to square the errors before adding them. Squaring ensures that positive and negative discrepancies are treated symmetrically and gives greater weight to large errors. This leads to the least squares optimization problem:

$$\hat{\beta} = \arg\min_{\beta} u'u,$$

where $u = (u_1, \ldots, u_N)'$ is the vector of model errors. Furthermore, after stacking

the N observations, we can write

$$u = y - X\beta,$$

where $y \in \mathbb{R}^{N\times 1}$ is the vector of dependent variables and $X \in \mathbb{R}^{N\times k}$ is the matrix of regressors.

3.1 Objective Function and Normal Equations

We express the sum of squared errors using matrix notation by stacking the N observations. Thus, the OLS estimator chooses $\hat{\beta}$ to minimize the objective function:

$$S(\beta) = (y - X\beta)'(y - X\beta).$$

To find the vector $\hat{\beta}$ that minimizes the sum of squared errors, we expand the objective function:

$$S(\beta) = (y - X\beta)'(y - X\beta) = y'y - 2\beta'X'y + \beta'X'X\beta.$$

To find the minimizer, we differentiate $S(\beta)$ with respect to β and set the derivative equal to zero.[4] This yields the first-order condition (first-order necessary condition, FONC):

$$\left.\frac{\partial S(\beta)}{\partial \beta}\right|_{\beta=\hat{\beta}} = -2X'y + 2X'X\hat{\beta} = 0.$$

Rearranging, this leads to the fundamental **normal equations**:

$$X'X\hat{\beta} = X'y.$$

To solve this system for a unique $\hat{\beta}$, we require a specific algebraic condition. While often listed as a statistical assumption, it is fundamentally a mathematical requirement for matrix invertibility:

Assumption 2 (Full Rank)**.** The regressor matrix $X \in \mathbb{R}^{N\times k}$ satisfies $\text{rank}(X) = k$. Equivalently, the columns of X are linearly independent, so there is no exact multicollinearity.

Under Assumption 2, the matrix $X'X$ is symmetric positive definite and therefore invertible. Pre-multiplying the normal equations by $(X'X)^{-1}$ gives the unique closed-form solution:

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

[4] See the mathematical appendix for a brief review of the matrix calculus used here.

A direct result of these normal equations is the sample orthogonality condition

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

This means that the OLS residual vector is orthogonal to every column of the regressor matrix. Later on, this property will reappear in the geometry of OLS, in the Frisch–Waugh–Lovell theorem, and in the logic of moment-based estimation.

The second-order condition (SOC) confirms that this solution is indeed the unique global minimizer, since the Hessian matrix of the objective function is

$$\frac{\partial^2 S(\beta)}{\partial \beta \, \partial \beta'} = 2X'X,$$

which is positive definite under Assumption 2. Therefore, the SOC for a minimum is satisfied.

Because this estimator is obtained by minimizing the sum of squared residuals, we denote it by $\hat{\beta}_{\text{OLS}}$. Later in the course, we will encounter other estimators derived from different objective functions or estimating equations, such as Maximum Likelihood and GMM. In those cases, subscripts will help us distinguish the estimation method being used. More broadly, our strategy throughout the course will follow two steps. First, we construct estimators mechanically by solving an optimization problem, or an equivalent system of estimating equations, as we did here. Second, we study the statistical behavior of those estimators under explicit assumptions, focusing on their finite-sample and asymptotic properties. This provides a systematic way to compare methods and to understand when a particular estimator is well suited to the problem at hand.

Remark 1 (Computational remark). Although the formula $\hat{\beta} = (X'X)^{-1}X'y$ is central for theory and exposition, numerical software typically does not compute OLS by forming the inverse of $X'X$ explicitly. In practice, routines in R, Julia, or Stata usually rely on QR decomposition, and in some contexts Cholesky decomposition, because these approaches are numerically more stable.

Remark 2 (Estimator vs. Estimate). It is useful to distinguish between two related concepts. An *estimator* is a rule or formula (that is, a random object such as $(X'X)^{-1}X'y$) that maps data into a parameter value. We will see how alternative objective functions (such as those in GMM or MLE) lead to different formulas. An *estimate* is the numerical value produced when that rule is applied to a particular realized sample.

3.2 Projection Matrices and the Geometry of OLS

Once $\hat{\beta}$ has been obtained, we can compute the fitted values for the sample:

$$\hat{y} = X\hat{\beta}.$$

Geometrically, $\hat{y}$ is the orthogonal projection of the observed vector y onto the column space of X, denoted $\text{col}(X)$. The residual vector

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

is therefore the component of y that lies orthogonal to $\text{col}(X)$. In other words, OLS decomposes the observed outcome vector into a part accounted for by the regressors and a part left unexplained by the linear projection.

To formalize this geometry, it is convenient to define two matrices. The first is the *projection matrix*

$$P \equiv X(X'X)^{-1}X',$$

which maps any vector in $\mathbb{R}^N$ onto $\text{col}(X)$. The second is the *residual-maker matrix*, or *annihilator*,

$$M \equiv I_N - P,$$

where I_N is the $N \times N$ identity matrix. The matrix M maps vectors onto the orthogonal complement of $\text{col}(X)$. These two matrices appear repeatedly in the algebra of OLS and are worth studying carefully.

Lemma 3.1 (Properties of P and M)**.** *Under Assumption 2, let*

$$P = X(X'X)^{-1}X' \qquad \textit{and} \qquad M = I_N - P.$$

Then:

(i) ***Symmetry:*** $P' = P$ *and* $M' = M$.
(ii) ***Idempotency:*** $P^2 = P$ *and* $M^2 = M$.
(iii) ***Projection:*** $PX = X$ *and* $MX = 0$.
(iv) ***Orthogonality:*** $PM = MP = 0$.
(v) ***Trace:*** $\text{tr}(P) = k$ *and* $\text{tr}(M) = N - k$.

Proof. (i) Symmetry. Since $X'X$ is symmetric, its inverse $(X'X)^{-1}$ is also symmetric. Therefore,

$$P' = \left[X(X'X)^{-1}X'\right]' = X\left[(X'X)^{-1}\right]'X' = X(X'X)^{-1}X' = P.$$

Since $M = I_N - P$, it follows immediately that

$$M' = I_N' - P' = I_N - P = M.$$

(ii) Idempotency. Using the definition of P,

$$P^2 = X(X'X)^{-1}X'X(X'X)^{-1}X' = X(X'X)^{-1}(X'X)(X'X)^{-1}X' = X(X'X)^{-1}X' = P.$$

Similarly,

$$M^2 = (I_N - P)^2 = I_N - 2P + P^2 = I_N - P = M.$$

(iii) Projection. Pre-multiplying X by P gives

$$PX = X(X'X)^{-1}(X'X) = X.$$

Hence,

$$MX = (I_N - P)X = X - PX = X - X = 0.$$

(iv) Orthogonality. We have

$$PM = P(I_N - P) = P - P^2 = P - P = 0.$$

Likewise,

$$MP = (I_N - P)P = P - P^2 = 0.$$

(v) Trace. First,

$$\text{tr}(P) = \text{tr}\big(X(X'X)^{-1}X'\big).$$

Using the cyclic property of the trace,

$$\text{tr}\big(X(X'X)^{-1}X'\big) = \text{tr}\big(X'X(X'X)^{-1}\big) = \text{tr}(I_k) = k.$$

Therefore,

$$\text{tr}(M) = \text{tr}(I_N - P) = \text{tr}(I_N) - \text{tr}(P) = N - k.$$

□

[Homework: Verify the cyclic property used in part (v): prove that $\text{tr}(AB) = \text{tr}(BA)$ for any conformable matrices by summing over scalar indices, and then use this result to show explicitly that $\text{tr}(X(X'X)^{-1}X') = k$.**]**

Using these operators, the fitted values and residuals can be written compactly as

$$\hat{y} = Py \qquad \text{and} \qquad \hat{u} = My.$$

These expressions summarize the geometry of OLS. The matrix P extracts the component of y that lies in $\text{col}(X)$, while M extracts the component orthogonal to that space. Since $PM = 0$, the two components are orthogonal:

$$\hat{y}'\hat{u} = (Py)'(My) = y'P'My = y'PMy = 0.$$

Thus, OLS delivers the orthogonal decomposition

$$y = \hat{y} + \hat{u},$$

where $\hat{y}$ is the explained component and $\hat{u}$ is the unexplained component relative to the linear span of the regressors.

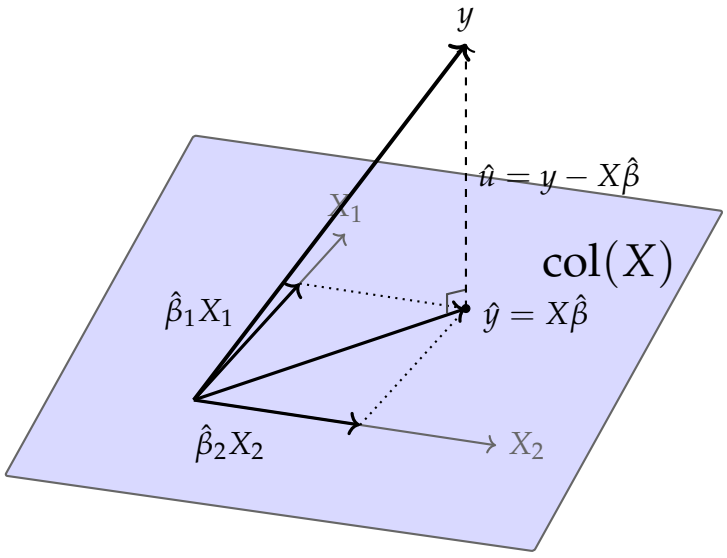

Figure 2: Geometric interpretation of OLS. The observed vector y is orthogonally projected onto $\text{col}(X)$ to produce the fitted values $\hat{y}$. The residual vector $\hat{u}$ lies in the orthogonal complement of $\text{col}(X)$.

3.3 Frisch–Waugh–Lovell Theorem

Having introduced projection matrices and the geometric interpretation of OLS, we are now in a position to derive a result that is both algebraically elegant and empirically useful: the Frisch–Waugh–Lovell (FWL) theorem. The theorem formalizes the idea of *partialling out*, that is, isolating the variation in a regressor after removing the component linearly associated with other covariates.

Suppose the regressor matrix is partitioned as

$$X = [X_1 \;\; X_2],$$

where X_1 contains a first group of regressors and X_2 contains the regressors whose coefficients are of primary interest. The theorem shows that the OLS estimate of the coefficient vector on X_2 can be obtained in two equivalent ways: either from the full regression of y on $[X_1 \;\; X_2]$, or by first removing from both y and X_2 the linear component explained by X_1, and then regressing the resulting residualized outcome on the residualized regressors.

Theorem 3.2 (Frisch–Waugh–Lovell)**.** *Partition the regressor matrix as* $X = [X_1 \;\; X_2]$ *and the parameter vector as*

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}.$$

Let

$$M_1 = I_N - X_1(X_1'X_1)^{-1}X_1'$$

be the residual-maker matrix associated with X_1*, and define*

$$\tilde{X}_2 = M_1 X_2, \qquad \tilde{y} = M_1 y.$$

Then the OLS estimate of β_2 *from the full regression of* y *on* $[X_1 \;\; X_2]$ *is*

$$\hat{\beta}_2 = (\tilde{X}_2'\tilde{X}_2)^{-1}\tilde{X}_2'\tilde{y}.$$

Moreover, the residuals from the full regression and from the auxiliary regression of $\tilde{y}$ *on* $\tilde{X}_2$ *are identical.*

Proof sketch. The normal equations for the full regression are

$$\begin{pmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{pmatrix} \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} X_1'y \\ X_2'y \end{pmatrix}.$$

From the first block equation,

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

Substitute this expression into the second block equation:

$$X_2'X_1\hat{\beta}_1 + X_2'X_2\hat{\beta}_2 = X_2'y.$$

After substitution and rearrangement,

$$X_2'M_1X_2\,\hat{\beta}_2 = X_2'M_1y.$$

Since M_1 is symmetric and idempotent, this can be written as

$$(M_1X_2)'(M_1X_2)\,\hat{\beta}_2 = (M_1X_2)'(M_1y),$$

which implies

$$\hat{\beta}_2 = (\tilde{X}_2'\tilde{X}_2)^{-1}\tilde{X}_2'\tilde{y}.$$

Finally, the residual from the full regression is

$$\hat{u} = y - X_1\hat{\beta}_1 - X_2\hat{\beta}_2.$$

Using the first normal equation, one can show that this is equal to

$$\hat{u} = M_1y - M_1X_2\hat{\beta}_2 = \tilde{y} - \tilde{X}_2\hat{\beta}_2,$$

which is exactly the residual vector from the auxiliary regression. □

[Homework: Fill in the algebra omitted in the proof. In particular, verify carefully that substituting the first block equation into the second yields

$$X_2'M_1X_2\,\hat{\beta}_2 = X_2'M_1y,$$

and then show that the residuals from the full and auxiliary regressions are identical.**]**

A useful special case occurs when $X_1 = \iota_N$, the $N \times 1$ vector of ones. In that case,

$$M_1 = I_N - \frac{1}{N}\iota_N \iota_N'$$

is the demeaning matrix. The FWL theorem then shows that including an intercept in the regression is equivalent to working with variables expressed as deviations from their sample means.

3.4 Decomposition of Variation and R^2

The projection results developed above also give rise to one of the most familiar summaries of regression fit. When the model includes an intercept, the fitted values and residuals satisfy an orthogonal decomposition around the sample mean. Let

$$M_\iota = I_N - \frac{1}{N}\iota_N \iota_N'$$

denote the demeaning matrix, where ι_N is the $N \times 1$ vector of ones. This matrix should be distinguished from the residual-maker matrix

$$M = I_N - P,$$

which removes the linear component associated with all regressors, not just the intercept. Since

$$y = \hat{y} + \hat{u},$$

we can write the centered decomposition as

$$M_\iota y = M_\iota \hat{y} + M_\iota \hat{u}.$$

Because the residuals are orthogonal to the fitted values, and because with an intercept we have $M_\iota \hat{u} = \hat{u}$, it follows that

$$y' M_\iota y = \hat{y}' M_\iota \hat{y} + \hat{u}'\hat{u}.$$

Therefore, the total variation in the dependent variable can be decomposed as

$$\text{TSS} = \text{ESS} + \text{SSR},$$

where $\text{TSS} = y'M_\iota y$ is the total sum of squares, $\text{ESS} = \hat{y}'M_\iota \hat{y}$ is the explained sum of squares, and $\text{SSR} = \hat{u}'\hat{u}$ is the sum of squared residuals. This decomposition motivates the *coefficient of determination*,

$$R^2 = 1 - \frac{\text{SSR}}{\text{TSS}} = \frac{\text{ESS}}{\text{TSS}},$$

which measures the fraction of the sample variation in y that is accounted for by the linear regression model. Although R^2 is often useful as a descriptive measure of fit, it should not be interpreted as evidence that the specification is correct. Moreover, R^2 weakly increases whenever an additional regressor is included, regardless of whether that regressor is actually informative.

Remark 3 (Adjusted R^2). Because R^2 tends to increase mechanically as regressors are added, empirical work often reports the adjusted coefficient of determination:

$$\bar{R}^2 = 1 - (1 - R^2)\frac{N-1}{N-k}.$$

The adjustment penalizes the inclusion of additional regressors through the loss of degrees of freedom. As a result, adding a variable that contributes little to the fit of the model may reduce $\bar{R}^2$. For this reason, adjusted R^2 is often more informative than R^2 when comparing specifications with different numbers of regressors.

3.5 Finite-Sample Properties of the OLS Estimator

After presenting the OLS estimator and showing how it is derived, the next question is whether it is a good estimator. There are two broad ways to evaluate an estimator. One is through its *finite-sample properties*, such as bias and variance. The other is through its *large-sample properties*, such as consistency and asymptotic normality. In this subsection, we follow the classical route and focus on the finite-sample behavior of OLS, leaving its asymptotic, or large-sample, properties for a separate set of notes.

We begin with unbiasedness. Recall that

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

Substituting the linear model $y = X\beta + u$ into this expression gives

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

Taking conditional expectations given X, we obtain

$$\mathbb{E}[\hat{\beta} \mid X] = \beta + (X'X)^{-1}X'\mathbb{E}[u \mid X].$$

This expression makes clear that OLS will be conditionally unbiased if the second term is equal to zero. This motivates the next assumption.

Assumption 3 (Zero Conditional Mean)**.** $\mathbb{E}[u \mid X] = 0$.

Assumption 3 is one of the central conditions in the classical linear model. Informally, it requires that the regressors carry no systematic information about the error term once we condition on the full regressor matrix X. When this condition fails, OLS is generally biased, and we enter the domain of endogeneity, where alternative methods such as IV become necessary.

Proposition 3.3 (Unbiasedness)**.** *Under Assumptions 1, 2, and 3,*

$$\mathbb{E}[\hat{\beta} \mid X] = \beta.$$

Proof.

$$\mathbb{E}[\hat{\beta} \mid X] = \beta + (X'X)^{-1}X'\mathbb{E}[u \mid X] = \beta.$$

□

A brief remark is useful here. In the earlier discussion of conditional modeling, we defined the regression error relative to the conditional mean, in which case $\mathbb{E}[u_i \mid x_i] = 0$ holds by construction. Here, however, we are working with the sample regression problem and treating $\mathbb{E}[u \mid X] = 0$ as an explicit assumption for the finite-sample analysis of OLS. This is a stronger statement because it conditions on the full regressor matrix X, not only on the covariates of a single observation. This distinction is worth keeping in mind: in practice, whether such orthogonality is plausible is an empirical and substantive question, not merely an algebraic one.

Unbiasedness is an important property, but it is not sufficient to rank estimators. We also care about *efficiency*, which leads us to study the variance of $\hat{\beta}$. From

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

it follows that

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

At this point we need additional structure on the conditional covariance matrix of the disturbances.

Assumption 4 (Spherical Disturbances)**.**

$$\mathbb{E}[uu' \mid X] = \sigma^2 I_N \qquad \text{for some } \sigma^2 > 0.$$

Equivalently, conditional on X, the disturbances are homoskedastic and mutually uncorrelated.

Under Assumption 4, the conditional variance of OLS takes a particularly simple form.

Theorem 3.4 (Conditional Variance of OLS)**.** *Under Assumptions 1, 2, 3, and 4,*

$$\operatorname{Var}(\hat{\beta} \mid X) = \sigma^2 (X'X)^{-1}.$$

Proof. Substituting $\mathbb{E}[uu' \mid X] = \sigma^2 I_N$ into the general variance expression yields

$$\operatorname{Var}(\hat{\beta} \mid X) = (X'X)^{-1} X' (\sigma^2 I_N) X (X'X)^{-1} = \sigma^2 (X'X)^{-1} X'X (X'X)^{-1} = \sigma^2 (X'X)^{-1}.$$

□

The formula above still depends on the unknown disturbance variance σ^2, so we need an estimator for it.

Proposition 3.5 (Unbiased Estimator of σ^2)**.** *Under Assumptions 1, 2, 3, and 4,*

$$\tilde{\sigma}^2 = \frac{\hat{u}'\hat{u}}{N - k}$$

satisfies

$$\mathbb{E}[\tilde{\sigma}^2 \mid X] = \sigma^2.$$

Proof. Since $\hat{u} = My$ and $MX = 0$ (Lemma 3.1(iii)),

$$\hat{u} = M(X\beta + u) = Mu.$$

Hence,

$$\hat{u}'\hat{u} = u'Mu,$$

because M is symmetric and idempotent. Taking expectations conditional on X,

$$\mathbb{E}[\hat{u}'\hat{u} \mid X] = \mathbb{E}[u'Mu \mid X] = \mathbb{E}[\operatorname{tr}(Muu') \mid X] = \operatorname{tr}\big(M\mathbb{E}[uu' \mid X]\big) = \sigma^2 \operatorname{tr}(M).$$

Using Lemma 3.1(v), $\operatorname{tr}(M) = N - k$, so

$$\mathbb{E}[\hat{u}'\hat{u} \mid X] = \sigma^2(N - k).$$

Dividing both sides by $N - k$ gives the result. □

Replacing σ^2 by $\tilde{\sigma}^2$ yields the usual estimator of the conditional variance-covariance matrix:

$$\widehat{\operatorname{Var}}(\hat{\beta} \mid X) = \tilde{\sigma}^2(X'X)^{-1}.$$

Preview: Robust Inference. The classical formula above relies on homoskedasticity and the absence of correlation across disturbances. Most statistical software packages use this expression by default. In empirical work, however, these conditions may fail. In that case, researchers often use alternative variance estimators that remain valid under weaker assumptions. For example, White's heteroskedasticity-robust estimator is

$$\widehat{\operatorname{Var}}_{HC}(\hat{\beta} \mid X) = (X'X)^{-1}\left(\sum_{i=1}^{N} \hat{u}_i^2 x_i x_i'\right)(X'X)^{-1}.$$

A related extension for dependent observations is the HAC estimator, where HAC stands for *heteroskedasticity and autocorrelation consistent*. Likewise, when dependence is expected within groups or clusters, a common estimator is the cluster-robust variance estimator:

$$\widehat{\operatorname{Var}}_{CR}(\hat{\beta} \mid X) = (X'X)^{-1}\left(\sum_{g=1}^{G} X_g'\hat{u}_g\hat{u}_g'X_g\right)(X'X)^{-1}.$$

Here $g = 1, \ldots, G$ indexes clusters, X_g is the regressor matrix for cluster g, and $\hat{u}_g$ is the corresponding residual vector. In practice, finite-sample correction factors are often added, but the expression above captures the basic idea. A key applied implication is that ignoring clustering when it is present often leads to standard errors that are too small and, therefore, to over-rejection of null hypotheses.

These robust estimators are extremely important in applied work, but they belong more naturally to large-sample inference. For now, the key point is that the classical OLS variance formula is exact under spherical disturbances, while robust alternatives are designed to remain valid when those assumptions are relaxed. We return to this topic in detail in a later set of notes on asymptotic inference.

We can now state the central efficiency result for OLS.

Theorem 3.6 (Gauss–Markov)**.** *Under Assumptions 1, 2, 3, and 4, the OLS estimator is* BLUE*: it is the Best Linear Unbiased Estimator. That is, for any other linear unbiased estimator* $\tilde{\beta} = Cy$*,*

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

Proof. Let

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

for some matrix D. Since $\tilde{\beta} = Cy$ is unbiased, we must have

$$\mathbb{E}[\tilde{\beta} \mid X] = CX\beta = \beta \qquad \text{for all } \beta,$$

which implies

$$CX = I_k.$$

Because $(X'X)^{-1}X'X = I_k$, it follows that

$$DX = 0.$$

Next, using Assumption 4,

$$\mathrm{Var}(\tilde{\beta} \mid X) = C\,\mathbb{E}[uu' \mid X]\,C' = \sigma^2 CC'.$$

Expanding CC' gives

$$\begin{aligned} CC' &= \left[(X'X)^{-1}X' + D\right]\left[X(X'X)^{-1} + D'\right] \\ &= (X'X)^{-1} + (X'X)^{-1}X'D' + DX(X'X)^{-1} + DD'. \end{aligned}$$

Since $DX = 0$, both cross terms vanish, so

$$CC' = (X'X)^{-1} + DD'.$$

Therefore,

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

because DD' is positive semidefinite. □

The Gauss–Markov theorem does not say that OLS is optimal among all possible estimators. Rather, it says that within the class of *linear unbiased* estimators, OLS has the smallest variance. This is why the theorem is important: under the classical

assumptions, OLS is not only easy to compute and interpret, but also efficient within a well-defined class of competitors. If we broaden the comparison class to allow biased estimators, such as Ridge regression, or to allow nonlinear procedures, OLS need not be optimal. The theorem is therefore precise and powerful, but its scope should be interpreted carefully.

4 Constrained Least Squares (CLS)

We can now extend the least squares framework by allowing for explicit restrictions on the parameter vector. In many applications, economic theory implies exact linear restrictions on β. For example, a production function may satisfy constant returns to scale, or a demand system may impose adding-up restrictions. The framework developed here will also be useful later when we study inference for such restrictions. Thus, to formalize this, suppose that β is subject to q linear restrictions of the form

$$Q'\beta = c,$$

where $Q \in \mathbb{R}^{k\times q}$ has full column rank and $c \in \mathbb{R}^q$ is a known vector.

The *Constrained Least Squares* (CLS) estimator minimizes the sum of squared residuals subject to these restrictions. Its Lagrangian is

$$\mathcal{L}(\beta, \lambda) = (y - X\beta)'(y - X\beta) + 2\lambda'(Q'\beta - c),$$

where $\lambda \in \mathbb{R}^q$ is the vector of Lagrange multipliers.

Solving the first-order conditions yields the CLS estimator:

$$\hat{\beta}_{CLS} = \hat{\beta}_{OLS} + (X'X)^{-1}Q\left[Q'(X'X)^{-1}Q\right]^{-1}(c - Q'\hat{\beta}_{OLS}).$$

This expression is intuitive. The constrained estimator starts from the unrestricted OLS estimator and then adjusts it just enough to satisfy the restrictions $Q'\beta = c$. If the unrestricted estimator already satisfies the restrictions, then $\hat{\beta}_{CLS} = \hat{\beta}_{OLS}$.

[Homework: Verify the CLS formula by solving the first-order conditions explicitly: differentiate $\mathcal{L}$ with respect to β and λ, set both equal to zero, and solve by substitution to obtain $\hat{\beta}_{CLS}$ and the Lagrange multiplier

$$\hat{\lambda} = \left[Q'(X'X)^{-1}Q\right]^{-1}(Q'\hat{\beta}_{OLS} - c).$$

]

Under the classical assumptions, if the restrictions are correct, CLS remains unbiased and is more efficient than unrestricted OLS in the positive semidefinite sense. Intuitively, imposing valid information reduces the admissible parameter space and can therefore improve precision. However, if the restrictions are false, the constrained estimator is generally biased.

This observation motivates a natural question: are the restrictions supported by the data? In the following inference section, we will study how to test linear restrictions using the standard F-test.

5 Inference in OLS

Up to this point, we have focused on the construction of the OLS estimator and on its finite-sample properties, such as unbiasedness and efficiency under the classical assumptions. In empirical work, however, estimation alone is not enough. We also want to assess the uncertainty surrounding the estimated coefficients and determine whether the data provide evidence in favor of or against economically meaningful hypotheses. This is the purpose of statistical inference.

Inference in the linear regression model is based on the sampling distribution of the estimator and of related quadratic forms. To obtain exact finite-sample results, we need assumptions that go beyond those required for unbiasedness or for the Gauss–Markov theorem. In particular, the classical t and F tests rely on a normality assumption for the disturbance term. This assumption is introduced here only to derive exact finite-sample distributions; it is not needed to establish unbiasedness or the BLUE property of OLS.

Assumption 5 (Normality)**.**

$$u \mid X \sim \mathcal{N}(0, \sigma^2 I_N).$$

Remark 4 (Roadmap of Results by Assumption Set). It is useful to keep track of which conclusions require which assumptions:

- **Assumptions 1–3** imply unbiasedness of $\hat{\beta}$ (Proposition 3.3).
- **Assumptions 1–4** imply

$$\text{Var}(\hat{\beta} \mid X) = \sigma^2 (X'X)^{-1},$$

and yield the Gauss–Markov result that OLS is BLUE (Theorem 3.6).

- **Assumptions 1–5** imply exact finite-sample t, F, and χ^2 distributions for the standard test statistics.

The next result summarizes the key distributional facts that underlie classical inference in the normal linear regression model.

Corollary 5.1 (Distributional Results Under the Normal Linear Regression Model)**.** *Under Assumptions 1–5:*

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

(ii)
$$\frac{\hat{u}'\hat{u}}{\sigma^2} \sim \chi^2(N-k).$$

(iii) $\hat{\beta}$ *and* $\hat{u}'\hat{u}$ *are independent conditional on* X*.*

Proof sketch. Part (i) follows from the representation

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

Conditional on X, the matrix $(X'X)^{-1}X'$ is fixed. Since Assumption 5 implies

$$u \mid X \sim \mathcal{N}(0, \sigma^2 I_N),$$

it follows that $\hat{\beta} \mid X$ is normally distributed as a linear transformation of a normal vector. Its conditional mean is β by Proposition 3.3, and its conditional variance is $\sigma^2(X'X)^{-1}$ by Theorem 3.4. Parts (ii) and (iii) follow from standard results on quadratic forms in normal vectors and are left as a homework exercise below. □

[Homework: Prove parts (ii) and (iii) using the spectral decomposition of the idempotent matrix M, which has $N-k$ eigenvalues equal to 1 and k eigenvalues equal to 0, or by applying Cochran's theorem directly to the quadratic form

$$\frac{u'Mu}{\sigma^2}.$$

]

5.1 t-Test

We now use the distributional results above to construct the first classical test of hypothesis in the linear model. Suppose we want to test

$$H_0 : \beta_j = \beta_{j,0}$$

for a single coefficient β_j.

Under Assumption 5, Corollary 5.1(i) implies that

$$\hat{\beta}_j \mid X \sim \mathcal{N}\left(\beta_j, \sigma^2[(X'X)^{-1}]_{jj}\right).$$

Therefore, if σ^2 were known, we could standardize and obtain

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

In practice, however, σ^2 is unknown and must be replaced by its estimator $\tilde{\sigma}^2 = \hat{u}'\hat{u}/(N-k)$. Using Corollary 5.1(ii)–(iii), the numerator is normal and independent of the estimator of the disturbance variance. As a result, the feasible test statistic

$$t_j = \frac{\hat{\beta}_j - \beta_{j,0}}{\widehat{\text{se}}(\hat{\beta}_j)} \sim t_{N-k},$$

where

$$\widehat{\text{se}}(\hat{\beta}_j) = \sqrt{\tilde{\sigma}^2[(X'X)^{-1}]_{jj}}.$$

This is the classical *t-test*. It allows us to test hypotheses about a single coefficient by comparing the observed value of t_j with the appropriate critical values from the Student's t distribution.

5.2 F-Test and Wald Test

The t-test is designed for a single restriction. In many applications, however, we want to test several linear restrictions jointly. Consider the null hypothesis

$$H_0 : R\beta = r,$$

where R is a $q \times k$ matrix of rank q, and $r \in \mathbb{R}^q$ is a known vector.

A classical way to test this hypothesis is to compare the fit of the restricted and

unrestricted models. Let SSR_R and SSR_U denote the sums of squared residuals from the restricted and unrestricted regressions, respectively. Under the null hypothesis, the statistic

$$F = \frac{(\text{SSR}_R - \text{SSR}_U)/q}{\text{SSR}_U/(N-k)} \sim F_{q,N-k}.$$

An equivalent representation is obtained from the unrestricted estimator alone. This is the *Wald form*:

$$F_W = \frac{(R\hat{\beta} - r)' \left[R(X'X)^{-1}R'\right]^{-1} (R\hat{\beta} - r)}{\tilde{\sigma}^2 \, q}.$$

Under the classical linear model, this statistic is numerically equal to the SSR-based F statistic above. Under Assumption 5, it also has the exact finite-sample distribution

$$F_W \sim F_{q,N-k}.$$

It is often useful to define the corresponding Wald statistic as

$$W = \frac{(R\hat{\beta} - r)' \left[R(X'X)^{-1}R'\right]^{-1} (R\hat{\beta} - r)}{\tilde{\sigma}^2},$$

so that

$$F_W = \frac{W}{q}.$$

Thus, the F-test and the Wald test are two equivalent ways of evaluating the same joint hypothesis in the classical linear model: one compares restricted and unrestricted fit directly, while the other uses only the unrestricted estimates and their variance-covariance matrix.

In large samples, the exact finite-sample normality assumption is no longer needed. Under standard regularity conditions,

$$W \xrightarrow{d} \chi^2(q),$$

so the Wald approach extends naturally to asymptotic inference.

Remark. The t and F tests developed in this section rely on the classical assumptions, including spherical disturbances and, for exact finite-sample results, normality. In applied work, these conditions often fail, especially because of heteroskedasticity or within-group dependence. In such cases, researchers typically rely on heteroskedasticity-

robust or cluster-robust standard errors and conduct inference using large-sample approximations. We return to these issues in a later set of notes on asymptotic inference.

5.3 Using the p-value for the test

The previous subsections presented hypothesis testing in its *critical-value* form: choose a significance level α, compute a test statistic, and reject H_0 if the statistic falls in the corresponding rejection region. The *p-value* expresses the same logic from a slightly different angle. Instead of fixing α first and then asking whether the statistic is sufficiently extreme, we start from the observed statistic and ask how unusual it would be under the null hypothesis.

Definition (observed significance level). Fix a null hypothesis H_0 and a test statistic T whose distribution under H_0 is known, either exactly or approximately. After observing the realized statistic t_{obs}, the p-value is defined as the probability—computed *under* H_0—of obtaining a value of the statistic at least as extreme as the one observed:

$$p\text{-value} = \Pr_{H_0}(\text{a realization at least as extreme as } t_{\text{obs}}) .$$

Equivalently, the p-value is the smallest significance level α at which the null hypothesis would be rejected.

Example: two-sided t-test. Consider the test of

$$H_0 : \beta_j = \beta_{j,0} \qquad \text{against} \qquad H_1 : \beta_j \neq \beta_{j,0}.$$

From Subsection 5.1, under the classical assumptions with normality, the statistic

$$t_j = \frac{\hat{\beta}_j - \beta_{j,0}}{\widehat{\text{se}}(\hat{\beta}_j)}$$

has a Student-t distribution with $N - k$ degrees of freedom under H_0. For a two-sided alternative, "more extreme" means farther from zero in absolute value. Therefore,

$$p\text{-value} = 2 \Pr(t_{N-k} \geq |t_{\text{obs}}|) = 2\Big[1 - F_{t_{N-k}}(|t_{\text{obs}}|)\Big],$$

where $F_{t_{N-k}}(\cdot)$ denotes the cumulative distribution function of the Student-t distribution. The decision rule is equivalent to the critical-value rule:

$$\text{reject } H_0 \text{ at level } \alpha \qquad \Longleftrightarrow \qquad p\text{-value} < \alpha.$$

A quick note on the interpretation. The p-value is not the probability that H_0 "is true". Rather, it is a probability statement about the *sampling behavior of the statistic under the null hypothesis*. A large p-value ("rejecting the null") means that the observed statistic lies in a region that is not unusual under H_0, so the data do not provide strong evidence against it.

Graphically, in a two-sided test, the p-value is the total tail area to the left of $-|t_{\text{obs}}|$ and to the right of $|t_{\text{obs}}|$ under the null distribution. Figure 3 illustrates this idea.

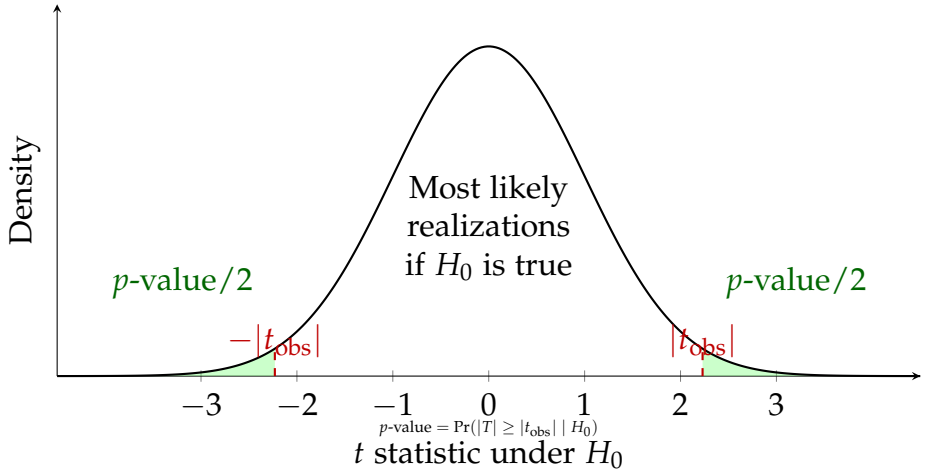

Figure 3: Two-sided p-value for a t-test. The shaded regions represent the probability, under H_0, of obtaining a statistic at least as extreme as $\pm|t_{\text{obs}}|$. For visual simplicity, the curve is drawn as a smooth symmetric proxy for the Student-t density. The exact numerical p-value is still computed from the t_{N-k} distribution given in the text.

Remark 5 (A note on p-hacking). Because empirical practice often emphasizes thresholds such as $p < 0.05$, researchers may be tempted to search across many specifications (controls, functional forms, subsamples, or outlier rules) until a "significant" result appears. This practice, commonly called *p-hacking*, undermines the nominal interpretation of the reported p-value because it ignores the multiple comparisons implicit in the specification search.

[Homework: Consider the individual null hypothesis

$$H_0 : \beta_j = 0.$$

Show that testing this null with the usual t-statistic is equivalent to testing it with the

F-statistic for a single linear restriction. In particular:

(i) prove that

$$t_j^2 = F;$$

(ii) explain why both tests must produce the same p-value.

Hint: Write the F-statistic in Wald form with $q = 1$, and then compare it directly to the squared t-statistic. Finally, recall that if $T \sim t_\nu$, then $T^2 \sim F_{1,\nu}$.]

6 A First Look at Gaussian Quasi-Maximum Likelihood

Before leaving OLS, it is useful to briefly note its connection with Maximum Likelihood Estimation (MLE), a topic that we will study in detail later in the course. Under Assumption 5, the conditional log-likelihood for the sample can be written as

$$\ell_N(\beta, \sigma^2; y \mid X) = -\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta).$$

Equivalently, using the least squares objective,

$$\ell_N(\beta, \sigma^2; y \mid X) = -\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}S(\beta).$$

This expression shows that, for a given value of σ^2, the log-likelihood depends on β only through the sum of squared residuals. Therefore, maximizing the Gaussian likelihood with respect to β is equivalent to minimizing $S(\beta)$. As a result, the MLE for β under normality coincides exactly with the OLS estimator.

This connection is useful for two reasons. First, it provides an alternative derivation of OLS under stronger distributional assumptions. Second, it anticipates a broader idea that will become important later in the course: even when the disturbance term is not actually normal, maximizing a Gaussian likelihood still leads to the OLS estimator for β. In that case, the estimator is often called the *Gaussian quasi-maximum likelihood estimator* (Gaussian QMLE).

7 Prediction and Forecast Evaluation

So far, the emphasis of these notes has been on estimation and inference for the coefficients of the linear regression model. In many empirical applications, however, the ultimate objective is not only to interpret coefficients or test hypotheses, but also to predict outcomes. This provides a natural final step for the set of notes: once a model has been estimated, how can it be used to generate predictions, and how should those predictions be evaluated?

Although the notation in this section is suggestive of a future observation and therefore resembles a forecasting setup, the same logic applies more generally to out-of-sample prediction. Suppose we observe a new regressor vector x_{N+j}. A natural predictor of the conditional mean of the outcome is

$$\hat{y}_{N+j} = x'_{N+j}\hat{\beta}.$$

This is often called the *plug-in predictor*, since it replaces the unknown parameter vector β with its estimator $\hat{\beta}$.

At this point, it is useful to distinguish between two related objects. The first is the conditional mean,

$$\mathbb{E}[y_{N+j} \mid x_{N+j}] = x'_{N+j}\beta,$$

which captures the systematic component of the outcome under the linear model. The second is the realized future outcome,

$$y_{N+j} = x'_{N+j}\beta + u_{N+j},$$

which also contains the future disturbance term. This distinction matters because the uncertainty involved in predicting the conditional mean is smaller than the uncertainty involved in predicting the actual realization.

When the goal is to predict the conditional mean, uncertainty arises only from estimating β. By contrast, when the goal is to predict the realized value of y_{N+j}, there is an additional source of uncertainty coming from the new disturbance u_{N+j}. Under homoskedasticity, the mean squared forecast error is therefore

$$\mathbb{E}\left[(\hat{y}_{N+j} - y_{N+j})^2 \mid x_{N+j}\right] = \sigma^2\left[1 + x'_{N+j}(X'X)^{-1}x_{N+j}\right].$$

The first term reflects the irreducible uncertainty from the future shock, while the second captures the estimation uncertainty associated with $\hat{\beta}$.

Once predictions have been generated, the next question is how to evaluate their

quality. Common measures of predictive accuracy include the root mean squared error (RMSE), the mean absolute error (MAE), and Theil's U. These statistics are useful for summarizing out-of-sample performance and comparing competing models on a holdout sample.

At the same time, a purely numerical comparison of forecasting errors does not tell us whether one model is significantly more accurate than another. For that purpose, one can use the Diebold–Mariano (DM) test, which evaluates whether the mean loss differential between two forecasting methods is equal to zero. If we define

$$d_{N+j} = g(\hat{u}_{1,N+j}) - g(\hat{u}_{2,N+j}),$$

where $g(\cdot)$ is a chosen loss function, then the DM test assesses whether the average of d_{N+j} differs significantly from zero. In multi-step forecasting settings, its standard errors must account for the serial correlation naturally induced in the loss differential.

This topic connects the linear regression model to one of its most common practical uses: generating predictions and comparing forecasting performance across alternative specifications.

Appendix A Matrix Algebra Reference

This appendix collects a few matrix results that were used repeatedly in the set of notes. The goal is not to provide a full review of matrix algebra, but to summarize the identities that entered directly in the derivation of OLS, the geometry of projections, the proof of the unbiased estimator of σ^2, and the finite-sample distributional results under normality.

First, in deriving the OLS estimator from the least squares objective, we used basic matrix derivatives. For vectors $a, b \in \mathbb{R}^k$ and a symmetric matrix $B \in \mathbb{R}^{k \times k}$,

$$\frac{\partial (a'b)}{\partial a} = b, \qquad \frac{\partial (a'Ba)}{\partial a} = 2Ba.$$

These identities justify the first-order condition of the quadratic objective function $S(\beta) = (y - X\beta)'(y - X\beta)$.

Second, in several proofs we used the cyclic property of the trace operator. For conformable matrices,

$$\operatorname{tr}(ABC) = \operatorname{tr}(BCA) = \operatorname{tr}(CAB).$$

In particular,

$$\operatorname{tr}(AB) = \operatorname{tr}(BA)$$

whenever both products are defined and square. This property was used, for example, in the proof that

$$\mathbb{E}[\hat{u}'\hat{u} \mid X] = \sigma^2(N - k).$$

Third, we repeatedly relied on the notion of positive definiteness. A symmetric matrix A is *positive definite*, written $A \succ 0$, if

$$x'Ax > 0 \qquad \text{for all } x \neq 0,$$

and *positive semidefinite*, written $A \succeq 0$, if

$$x'Ax \geq 0 \qquad \text{for all } x.$$

Under the full-rank assumption, $X'X$ is positive definite, which guarantees the uniqueness of the OLS solution. More generally, for any matrix D, the matrix DD' is positive semidefinite, a fact used in the proof of the Gauss–Markov theorem.

Finally, in the inference section we invoked a standard result for symmetric idempotent matrices. If M is symmetric and idempotent with rank r, then it admits a

spectral decomposition

$$M = Q\Lambda Q',$$

where Q is orthogonal and $\Lambda = \text{diag}(1, \ldots, 1, 0, \ldots, 0)$ has r ones and the remaining diagonal elements equal to zero. When

$$u \mid X \sim \mathcal{N}(0, \sigma^2 I_N),$$

this implies

$$\frac{u'Mu}{\sigma^2} \sim \chi^2(r).$$

This result underlies the distribution of $\hat{u}'\hat{u}/\sigma^2$ in the normal linear regression model.

Appendix B Notation Summary

Throughout the notes, the linear regression model is written as

$$y = X\beta + u,$$

where $y \in \mathbb{R}^{N\times 1}$ is the vector of dependent variables, $X \in \mathbb{R}^{N\times k}$ is the regressor matrix, $\beta \in \mathbb{R}^{k\times 1}$ is the parameter vector, and $u \in \mathbb{R}^{N\times 1}$ is the disturbance vector. The i-th row of X is denoted x_i'.

The OLS estimator is

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

The fitted values are

$$\hat{y} = X\hat{\beta} = Py,$$

and the residuals are

$$\hat{u} = y - X\hat{\beta} = My,$$

where

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

is the projection matrix onto $\text{col}(X)$ and

$$M = I_N - P$$

is the residual-maker matrix. In the discussion of centering and the decomposition of

variation, we also used the demeaning matrix

$$M_\iota = I_N - \frac{1}{N}\iota_N \iota_N',$$

where ι_N is the $N \times 1$ vector of ones. This matrix should not be confused with M, which residualizes with respect to the full regressor matrix. In the Frisch–Waugh–Lovell theorem, we additionally used

$$M_1 = I_N - X_1(X_1'X_1)^{-1}X_1',$$

the residual-maker matrix associated with the subset of regressors X_1.

Under spherical disturbances, the unbiased estimator of the disturbance variance is

$$\tilde{\sigma}^2 = \frac{\hat{u}'\hat{u}}{N-k}.$$

This quantity is used to construct the estimated variance-covariance matrix of OLS and the classical t and F statistics.

A final notation point is worth keeping in mind. In these notes, u denotes the model disturbance, while $\hat{u}$ denotes the residual obtained after estimation. This distinction is important because many derivations begin with the population model $y = X\beta + u$, whereas empirical implementation uses the residual vector $y - X\hat{\beta}$.

Appendix C Distributional Results Used in Classical OLS Inference

This appendix collects a few distributional results used in the derivation of the classical finite-sample inference results for OLS. The goal is not to provide a general review of probability theory, but rather to summarize the specific facts needed for Corollary 5.1 and for the construction of the classical t and F tests.

We begin with a standard fact about linear transformations of normal random vectors.

Proposition C.1 (Linear transformation of a normal vector)**.** *Let $u \mid X \sim \mathcal{N}(0, \sigma^2 I_N)$, and let A be a fixed matrix conditional on X. Then*

$$Au \mid X \sim \mathcal{N}\left(0, \sigma^2 AA'\right).$$

This result is used in the inference section to show that

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

is conditionally normal under Assumption 5, since $(X'X)^{-1}X'$ is fixed given X.

A second key result concerns quadratic forms in normal vectors.

Proposition C.2 (Quadratic form for a symmetric idempotent matrix)**.** *Let $u \mid X \sim \mathcal{N}(0, \sigma^2 I_N)$, and let M be a symmetric idempotent matrix with rank r. Then*

$$\frac{u'Mu}{\sigma^2} \sim \chi^2(r).$$

In the OLS model, the residual-maker matrix

$$M = I_N - X(X'X)^{-1}X'$$

is symmetric and idempotent, with rank $N - k$. Therefore,

$$\frac{\hat{u}'\hat{u}}{\sigma^2} = \frac{u'Mu}{\sigma^2} \sim \chi^2(N - k),$$

which is the result used in Corollary 5.1(ii).

The next fact explains why the numerator and denominator of the classical t statistic are independent.

Proposition C.3 (Orthogonal normal components are independent)**.** *Let $u \mid X \sim \mathcal{N}(0, \sigma^2 I_N)$. If A and B are fixed matrices conditional on X such that*

$$AB' = 0,$$

then the random vectors Au and Bu are independent conditional on X.

In the OLS model, this applies to

$$(X'X)^{-1}X'u \qquad \text{and} \qquad Mu,$$

because

$$X'M = 0.$$

Hence, under normality, $\hat{\beta}$ and $\hat{u}'\hat{u}$ are independent conditional on X.

These results lead directly to the classical definitions of the Student-t and F distributions.

Definition C.4 (Student-t distribution)**.** Let

$$Z \sim \mathcal{N}(0,1), \qquad V \sim \chi^2(\nu),$$

and suppose that Z and V are independent. Then

$$T = \frac{Z}{\sqrt{V/\nu}}$$

has a Student-t distribution with ν degrees of freedom, written

$$T \sim t_\nu.$$

This is exactly the structure of the classical t statistic in the linear regression model:

$$t_j = \frac{\hat{\beta}_j - \beta_{j,0}}{\sqrt{\tilde{\sigma}^2[(X'X)^{-1}]_{jj}}} \sim t_{N-k}.$$

Definition C.5 (F distribution)**.** Let

$$V_1 \sim \chi^2(q), \qquad V_2 \sim \chi^2(\nu),$$

and suppose that V_1 and V_2 are independent. Then

$$F = \frac{V_1/q}{V_2/\nu}$$

has an F distribution with q and ν degrees of freedom, written

$$F \sim F_{q,\nu}.$$

This is the distributional structure underlying the classical joint test of linear restrictions in the regression model.

A useful special case links the t and F distributions directly.

Proposition C.6 (Relation between t and F)**.** *If*

$$T \sim t_\nu,$$

then

$$T^2 \sim F_{1,\nu}.$$

This identity helps explain why, in the special case of a single restriction, the square of the usual t statistic is equal to the corresponding F statistic.

Overall, these results supply the probability tools behind the exact finite-sample inference developed in Section 5. In particular, they explain why OLS coefficient estimates are normally distributed under Assumption 5, why the residual sum of squares generates a chi-square variable, and why the resulting test statistics follow the classical t and F distributions.7