OLS in Matrix Form

Noted by Md. Inzamamul Lohani

Software Engineering, IIT, DU

November 22, 2026

Contents

1 The True Model **3**

1.1 Dimensions 3

2 Estimates, Fitted Values, and Residuals **3**

2.1 Disturbances versus residuals 3

3 The OLS Criterion: Minimize the Sum of Squared Residuals **4**

3.1 Why is $\mathbf{y}'\mathbf{X}\hat{\boldsymbol{\beta}}$ a scalar? 4

4 Differentiating RSS with Respect to the Coefficient Vector **5**

4.1 Useful matrix differentiation rules 5

4.2 Deriving the quadratic-form rule 5

4.3 Differentiate the three RSS terms 5

5 Normal Equations and the OLS Estimator **6**

5.1 When might the inverse fail to exist? 6

6 Checking That the Stationary Point Is a Minimum **6**

7 A Fundamental OLS Property: $\mathbf{X}'\mathbf{e} = 0$ **7**

7.1 Column-by-column interpretation 7

8 Consequences When the Model Includes a Constant **8**

9 Key Ideas at a Glance **8**

10 Practice Questions **8**

Scope of These Notes **9**

1 The True Model

The population regression model in matrix notation is

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

Here:

- $\mathbf{X}$ is an $n \times k$ matrix containing observations on k independent variables for n observations. If the model includes a constant (intercept), one column of $\mathbf{X}$ consists entirely of ones.
- $\mathbf{y}$ is an $n \times 1$ vector of observations on the dependent variable.
- $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of disturbances (unobserved influences).
- $\boldsymbol{\beta}$ is a $k \times 1$ vector of unknown population parameters.

The model has a **systematic component**, $\mathbf{X}\boldsymbol{\beta}$, and a **stochastic component**, $\boldsymbol{\varepsilon}$. The objective is to estimate the unknown population parameters in $\boldsymbol{\beta}$.

1.1 Dimensions

The dimensions are chosen so that the matrix multiplication is valid:

$$(n \times k)(k \times 1) = n \times 1.$$

Thus $\mathbf{X}\boldsymbol{\beta}$ has the same dimension as $\mathbf{y}$ and $\boldsymbol{\varepsilon}$.

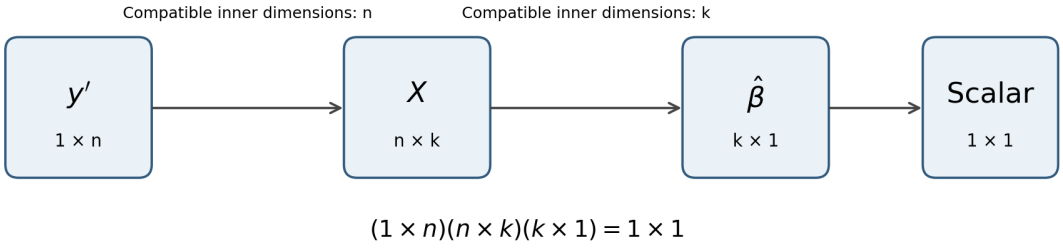

Figure 1: Dimension check for the scalar product $\mathbf{y}'\mathbf{X}\hat{\boldsymbol{\beta}}$.

2 Estimates, Fitted Values, and Residuals

The estimated parameter vector is denoted by $\hat{\boldsymbol{\beta}}$. The fitted (predicted) values are

$$\widehat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}.$$

The residual vector is the difference between observed and fitted values:

$$\boxed{\mathbf{e} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}.} \tag{2.1}$$

A residual is observable because it can be calculated from the sample.

2.1 Disturbances versus residuals

These terms must be distinguished:

	Disturbance ε	**Residual e**
Meaning	Unobserved part of the population model	Observed minus fitted value
Observable?	No	Yes
Role	Part of the true model	Computed after estimating the model

In general, $\boldsymbol{\varepsilon} \neq \mathbf{e}$. Do not infer properties of the unobserved disturbances merely from properties of the residuals.

3 The OLS Criterion: Minimize the Sum of Squared Residuals

The sum of squared residuals (RSS) is

$$\text{RSS} = \mathbf{e}'\mathbf{e}. \tag{3.1}$$

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

$$\text{RSS} = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}). \tag{3.2}$$

The transpose of a product reverses the order:

$$(AB)' = B'A', \qquad (A - B)' = A' - B'.$$

Therefore,

$$(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})' = \mathbf{y}' - (\mathbf{X}\hat{\boldsymbol{\beta}})' = \mathbf{y}' - \hat{\boldsymbol{\beta}}'\mathbf{X}'.$$

Substituting into (3.2) gives

$$\text{RSS} = (\mathbf{y}' - \hat{\boldsymbol{\beta}}'\mathbf{X}')(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \tag{3.3}$$

$$= \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X}\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}. \tag{3.4}$$

The middle expressions are scalars. A scalar is equal to its transpose, so

$$\mathbf{y}'\mathbf{X}\hat{\boldsymbol{\beta}} = (\mathbf{y}'\mathbf{X}\hat{\boldsymbol{\beta}})' = \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y}.$$

Hence the two middle terms are equal, and

$$\boxed{\text{RSS} = \mathbf{y}'\mathbf{y} - 2\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}.} \tag{3.5}$$

3.1 Why is $\mathbf{y}'\mathbf{X}\hat{\boldsymbol{\beta}}$ a scalar?

The dimensions are

$$\underbrace{\mathbf{y}'}_{1\times n}\underbrace{\mathbf{X}}_{n\times k}\underbrace{\hat{\boldsymbol{\beta}}}_{k\times 1}.$$

Multiplying from left to right,

$$(1 \times n)(n \times k) = 1 \times k, \qquad (1 \times k)(k \times 1) = 1 \times 1.$$

A 1×1 matrix contains one number, so it is a scalar.

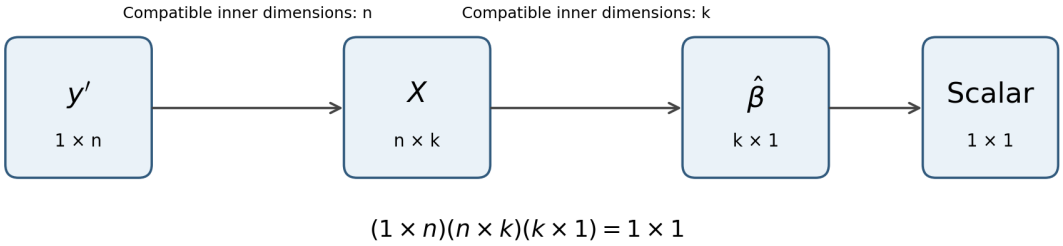

Figure 2: The final dimension of $\mathbf{y}'\mathbf{X}\hat{\boldsymbol{\beta}}$ is 1×1.

4 Differentiating RSS with Respect to the Coefficient Vector

We choose $\hat{\boldsymbol{\beta}}$ to minimize RSS. Differentiate each term in (3.5) with respect to the vector $\hat{\boldsymbol{\beta}}$.

4.1 Useful matrix differentiation rules

For vectors $\mathbf{a}$ and $\mathbf{b}$ of compatible dimensions,

$$\frac{\partial(\mathbf{b}'\mathbf{a})}{\partial \mathbf{b}} = \mathbf{a}.$$

For a square matrix A,

$$\frac{\partial(\boldsymbol{\beta}' A \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = (A + A')\boldsymbol{\beta}. \tag{4.1}$$

If A is symmetric ($A' = A$), this becomes

$$\frac{\partial(\boldsymbol{\beta}' A \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 2A\boldsymbol{\beta}.$$

4.2 Deriving the quadratic-form rule

Let $f(\boldsymbol{\beta}) = \boldsymbol{\beta}' A \boldsymbol{\beta}$. Using the product rule in differential form,

$$df = d\boldsymbol{\beta}' A \boldsymbol{\beta} + \boldsymbol{\beta}' A \, d\boldsymbol{\beta}.$$

The second term is a scalar, so it equals its transpose:

$$\boldsymbol{\beta}' A \, d\boldsymbol{\beta} = (\boldsymbol{\beta}' A \, d\boldsymbol{\beta})' = d\boldsymbol{\beta}' A' \boldsymbol{\beta}.$$

Thus

$$df = d\boldsymbol{\beta}' A \boldsymbol{\beta} + d\boldsymbol{\beta}' A' \boldsymbol{\beta} = d\boldsymbol{\beta}' (A + A') \boldsymbol{\beta}.$$

Therefore the gradient is $(A + A')\boldsymbol{\beta}$, proving (4.1).

4.3 Differentiate the three RSS terms

Starting from

$$\text{RSS} = \underbrace{\mathbf{y}'\mathbf{y}}_{\text{Term 1}} \underbrace{- 2\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y}}_{\text{Term 2}} + \underbrace{\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}}_{\text{Term 3}},$$

we obtain:

$$\begin{aligned}
\frac{\partial(\mathbf{y}'\mathbf{y})}{\partial\hat{\boldsymbol{\beta}}} &= 0, && \text{because } \mathbf{y} \text{ is fixed with respect to } \hat{\boldsymbol{\beta}},\\
\frac{\partial(-2\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y})}{\partial\hat{\boldsymbol{\beta}}} &= -2\mathbf{X}'\mathbf{y}, &&\\
\frac{\partial(\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}})}{\partial\hat{\boldsymbol{\beta}}} &= (\mathbf{X}'\mathbf{X} + (\mathbf{X}'\mathbf{X})')\hat{\boldsymbol{\beta}} &&\\
&= 2\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}, && \text{because } \mathbf{X}'\mathbf{X} \text{ is symmetric.}
\end{aligned}$$

Combining the three derivatives gives

$$\boxed{\frac{\partial\,\mathrm{RSS}}{\partial\hat{\boldsymbol{\beta}}} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}.} \tag{4.2}$$

5 Normal Equations and the OLS Estimator

At an interior minimum, the first derivative (gradient) is zero:

$$-2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = 0.$$

Divide by 2 and rearrange:

$$\boxed{\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}.} \tag{5.1}$$

These are the **normal equations**.

The matrix $\mathbf{X}'\mathbf{X}$ is $k \times k$, square, and symmetric. If it is invertible, pre-multiply both sides of (5.1) by $(\mathbf{X}'\mathbf{X})^{-1}$:

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

Since $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X} = I_k$,

$$\boxed{\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.} \tag{5.2}$$

This is the OLS estimator in matrix form.

5.1 When might the inverse fail to exist?

The inverse of $\mathbf{X}'\mathbf{X}$ may not exist if:

- $n < k$, so there are fewer observations than independent variables; or
- there is perfect multicollinearity, meaning one or more columns of $\mathbf{X}$ are exact linear combinations of other columns.

The full-rank condition rules out these cases.

6 Checking That the Stationary Point Is a Minimum

Differentiate the gradient in (4.2) again with respect to $\hat{\boldsymbol{\beta}}$. The Hessian matrix is

$$\frac{\partial^2\,\mathrm{RSS}}{\partial\hat{\boldsymbol{\beta}}\,\partial\hat{\boldsymbol{\beta}}'} = 2\mathbf{X}'\mathbf{X}.$$

If $\mathbf{X}$ has full column rank, then $\mathbf{X}'\mathbf{X}$ is positive definite. Therefore the Hessian is positive definite, and the stationary point is a minimum.

7 A Fundamental OLS Property: $\mathbf{X}'\mathbf{e} = 0$

Start with the normal equations:

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

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

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

Therefore,

$$\boxed{\mathbf{X}'\mathbf{e} = \mathbf{0}.} \tag{7.1}$$

This says that the residual vector is orthogonal to every column of the regressor matrix.

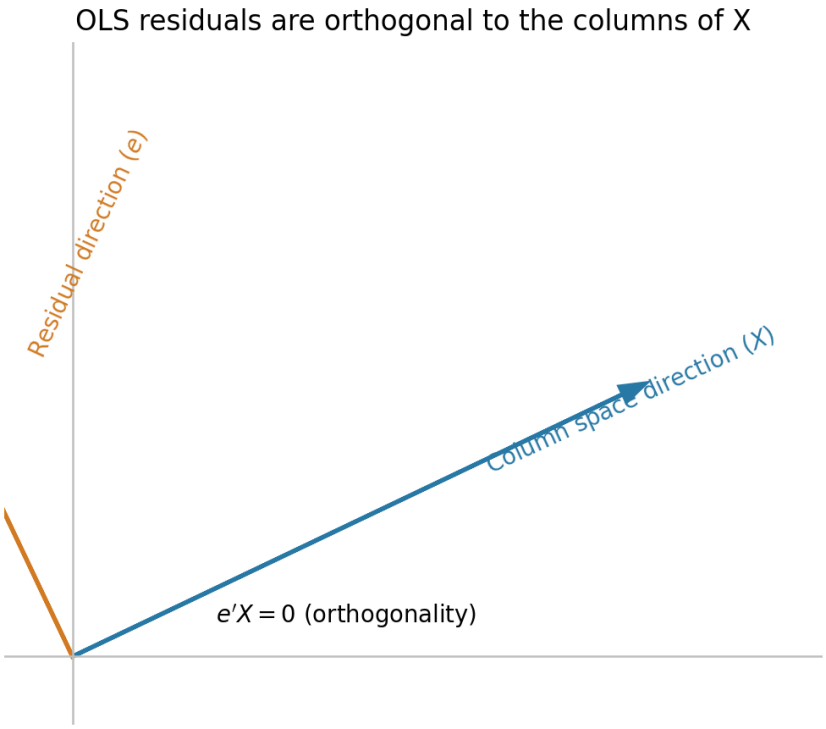

Figure 3: Conceptual illustration of residual orthogonality. The residual direction is perpendicular to the column-space direction.

7.1 Column-by-column interpretation

If $\mathbf{x}_j$ is the jth column of $\mathbf{X}$, then (7.1) implies

$$\mathbf{x}_j'\mathbf{e} = 0 \qquad \text{for every } j = 1, \ldots, k.$$

In sample terms, each regressor has zero sample inner product with the residual vector. With the usual intercept specification, this also corresponds to zero sample covariance/correlation between each regressor and the residuals (provided the relevant variances are nonzero).

This is a property of the **observed regressors and residuals**. It does not by itself establish that regressors are uncorrelated with the unobserved disturbances.

8 Consequences When the Model Includes a Constant

If the regression includes an intercept, one column of $\mathbf{X}$ is a vector of ones:

$$\mathbf{1} = \begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix}'.$$

The corresponding element of $\mathbf{X}'\mathbf{e} = \mathbf{0}$ is

$$\mathbf{1}'\mathbf{e} = 0.$$

Therefore,

$$\boxed{\sum_{i=1}^{n} e_i = 0.} \tag{8.1}$$

The sample mean of the residuals is consequently

$$\boxed{\bar{e} = \frac{1}{n}\sum_{i=1}^{n} e_i = 0.}$$

These two properties require an intercept in the regression.

9 Key Ideas at a Glance

Concept	**Main result**
True model	$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$
Fitted values	$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$
Residuals	$\mathbf{e} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$
RSS	$\text{RSS} = \mathbf{e}'\mathbf{e}$
Expanded RSS	$\mathbf{y}'\mathbf{y} - 2\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}$
RSS gradient	$-2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}$
Normal equations	$\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}$
OLS estimator	$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$
Residual orthogonality	$\mathbf{X}'\mathbf{e} = \mathbf{0}$
With an intercept	$\sum_i e_i = 0$ and $\bar{e} = 0$

10 Practice Questions

1. If $\mathbf{X}$ is $n \times k$, what are the dimensions of $\mathbf{X}'$, $\mathbf{X}'\mathbf{X}$, and $\mathbf{X}'\mathbf{y}$?
2. Why is $\mathbf{y}'\mathbf{X}\hat{\boldsymbol{\beta}}$ a scalar?
3. Explain why $\mathbf{y}'\mathbf{X}\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y}$.
4. Expand $(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$ step by step.
5. Why does the derivative of $\mathbf{y}'\mathbf{y}$ with respect to $\hat{\boldsymbol{\beta}}$ equal zero?
6. State the derivative of $\boldsymbol{\beta}'A\boldsymbol{\beta}$ when A is symmetric.
7. Derive the normal equations from the RSS gradient.
8. Under what condition can the OLS estimator be written using $(\mathbf{X}'\mathbf{X})^{-1}$?
9. Derive $\mathbf{X}'\mathbf{e} = \mathbf{0}$ from the normal equations.
10. Why does including a constant imply that the residuals sum to zero?

Scope of These Notes

These notes cover the material discussed so far from the supplied document: the true model, estimates and residuals, the RSS criterion, its matrix expansion and differentiation, the normal equations, the OLS estimator, the minimum check, and the first residual properties. Later sections of the source document (Gauss–Markov assumptions and theorem, variance-covariance matrix, robust standard errors, partitioned regression, and the Frisch–Waugh–Lovell theorem) are not included in this installment.