Matrix Form and F-tests

Gov 2001: Quantitative Social Science Methods I

Week 11, Lecture 22

Spring 2026

For Today

Required Reading

- Blackwell, Chapter 7 (pp. 139–157)
- Aronow & Miller, §4.2 (pp. 156–170)

Today: Matrix notation for OLS and testing multiple hypotheses.

Roadmap

1. Matrix notation: why and how
2. OLS in matrix form
3. Variance-covariance matrix
4. F-tests for joint hypotheses
5. Testing groups of coefficients

Part I: Matrix Notation

Why Matrix Notation?

With k regressors, the model is:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \varepsilon_i$$

Writing this for each observation gets cumbersome.

Matrix notation:

- More compact
- General formulas (any number of regressors)
- Cleaner derivations of properties
- How software actually computes OLS

Setting Up the Matrices

Outcome vector $(n \times 1)$:

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}$$

Parameter vector $((k+1) \times 1)$:

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}$$

The Design Matrix

Design matrix $(n \times (k+1))$:

$$\mathbf{X} = \begin{pmatrix} 1 & X_{11} & X_{21} & \cdots & X_{k1} \\ 1 & X_{12} & X_{22} & \cdots & X_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{1n} & X_{2n} & \cdots & X_{kn} \end{pmatrix}$$

- First column is all 1s (for the intercept)
- Each row is one observation
- Each column (after the first) is one regressor

The Model in Matrix Form

The regression model:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where:

- $\mathbf{Y}$ is $n \times 1$
- $\mathbf{X}$ is $n \times (k+1)$
- $\boldsymbol{\beta}$ is $(k+1) \times 1$
- $\boldsymbol{\varepsilon}$ is $n \times 1$

This represents all n equations simultaneously!

Part II: OLS in Matrix Form

The OLS Problem

Minimize:

$$SSR = \sum_{i=1}^{n} \hat{e}_i^2 = \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} = (\mathbf{Y} - \mathbf{X}\mathbf{b})'(\mathbf{Y} - \mathbf{X}\mathbf{b})$$

Expand:

$$SSR = \mathbf{Y}'\mathbf{Y} - 2\mathbf{b}'\mathbf{X}'\mathbf{Y} + \mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b}$$

Take derivative with respect to **b** and set to zero:

$$\frac{\partial SSR}{\partial \mathbf{b}} = -2\mathbf{X}'\mathbf{Y} + 2\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{0}$$

The OLS Formula

Normal equations (matrix form):

$$\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{Y}$$

Solve for $\hat{\boldsymbol{\beta}}$:

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

This is THE OLS formula—works for any number of regressors.

Requires: $\mathbf{X}'\mathbf{X}$ must be invertible (no perfect multicollinearity).

Connecting to Simple Regression

For simple regression with one regressor:

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} n & \sum X_i \\ \sum X_i & \sum X_i^2 \end{pmatrix}$$

$$\mathbf{X}'\mathbf{Y} = \begin{pmatrix} \sum Y_i \\ \sum X_i Y_i \end{pmatrix}$$

Solving $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ gives the familiar:

$$\hat{\beta}_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$

Fitted Values and Residuals

Fitted values:

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

where $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is the **hat matrix**.

Residuals:

$$\hat{\boldsymbol{\varepsilon}} = \mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{I} - \mathbf{H})\mathbf{Y} = \mathbf{M}\mathbf{Y}$$

where $\mathbf{M} = \mathbf{I} - \mathbf{H}$ is the **residual maker matrix**.

Properties of Hat and Residual Maker Matrices

Hat matrix H:

- Symmetric: $\mathbf{H}' = \mathbf{H}$
- Idempotent: $\mathbf{HH} = \mathbf{H}$
- $\mathbf{HX} = \mathbf{X}$

Residual maker M:

- Symmetric: $\mathbf{M}' = \mathbf{M}$
- Idempotent: $\mathbf{MM} = \mathbf{M}$
- $\mathbf{MX} = \mathbf{0}$ (residuals orthogonal to regressors)

Part III: Variance-Covariance Matrix

Variance-Covariance Matrix of $\hat{\beta}$

Under homoskedasticity: $\text{Var}(\boldsymbol{\varepsilon}|\mathbf{X}) = \sigma^2\mathbf{I}$

Variance-covariance matrix:

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

This is a $(k+1) \times (k+1)$ matrix:

- Diagonal elements: $\text{Var}(\hat{\beta}_j)$
- Off-diagonal elements: $\text{Cov}(\hat{\beta}_j, \hat{\beta}_\ell)$

Estimating σ^2

Estimator:

$$\hat{\sigma}^2 = \frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{n-k-1} = \frac{SSR}{n-k-1}$$

Degrees of freedom: $n - k - 1$ (sample size minus number of parameters).

Estimated variance-covariance matrix:

$$\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$$

Standard errors are square roots of the diagonal elements.

Part IV: F-tests

Why Test Multiple Coefficients?

t-tests test **one coefficient at a time**:

$$H_0 : \beta_j = 0$$

But sometimes we want to test multiple coefficients jointly:

- Are all coefficients zero? (Overall model significance)
- Are a subset of coefficients zero? (Do these variables matter?)
- Are two coefficients equal? ($\beta_1 = \beta_2$?)

Problem: Testing each separately has multiple testing issues.

The F-test Idea

Compare two models:

Restricted model (imposes H_0): smaller, simpler
Unrestricted model: full model

Key insight:

- If H_0 is true, both models fit similarly
- If H_0 is false, unrestricted model fits much better

F-statistic: Measures improvement in fit from unrestricted model.

The F-Statistic

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

where:

- SSR_R = sum of squared residuals from restricted model
- SSR_U = sum of squared residuals from unrestricted model
- q = number of restrictions (coefficients being tested)
- $n - k - 1$ = degrees of freedom in unrestricted model

Under H_0: $F \sim F_{q,n-k-1}$

F-test Intuition

$$F = \frac{\text{Improvement in fit per restriction}}{\text{Variance estimate from unrestricted model}}$$

- **Large F**: Restricted model fits much worse—reject H_0
- **Small F**: Not much improvement from extra variables—don't reject H_0

Decision rule: Reject H_0 if $F > F_{q,n-k-1,1-\alpha}$
Or: Reject if p-value $< \alpha$.

Example: Overall Model Significance

Test: $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$
(All slope coefficients are zero)

Restricted model: $Y_i = \beta_0 + \varepsilon_i$ (just the mean)
Unrestricted model: $Y_i = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon_i$

F-statistic:

$$F = \frac{(SST - SSR)/k}{SSR/(n-k-1)} = \frac{R^2/k}{(1-R^2)/(n-k-1)}$$

This is the "overall F" reported by regression software.

Example: Testing a Subset of Coefficients

Full model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$$

Test: $H_0 : \beta_3 = \beta_4 = 0$
(Do X_3 and X_4 jointly matter?)

Restricted model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

Procedure:

1. Estimate both models, get SSR_R and SSR_U
2. Compute $F = \frac{(SSR_R - SSR_U)/2}{SSR_U/(n-5)}$
3. Compare to $F_{2,n-5}$ distribution

Relationship Between F and t

Special case: Testing a single coefficient ($q = 1$)

$$H_0 : \beta_j = 0$$

The F-statistic equals the square of the t-statistic:

$$F_{1,n-k-1} = t^2_{n-k-1}$$

For single restrictions, F and t tests give identical results.
F is more general—can test multiple restrictions.

Wald Form of the F-test

For testing $H_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{r}$:

$$F = \frac{(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})/q}{\hat{\sigma}^2}$$

Example: $H_0 : \beta_1 = \beta_2$
Rewrite as: $\beta_1 - \beta_2 = 0$
$\mathbf{R} = (0, 1, -1, 0, \ldots)$, $\mathbf{r} = 0$

The Wald form lets you test any linear combination of coefficients.

Summary

Matrix form:

- $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$
- $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$
- $\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$

F-tests:

- Test multiple coefficients jointly
- Compare restricted vs. unrestricted models
- $F = \frac{(SSR_R - SSR_U)/q}{SSR_U/(n-k-1)}$

Looking Ahead

Next week: Inference for Regression

- Heteroskedasticity
- Robust standard errors
- Clustered standard errors
- When and how to cluster

Matrix notation: $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$

F-tests let us test multiple coefficients at once.

Compare fit of restricted vs. unrestricted models.