Advanced Econometrics
04 OLS Properties

Eduard Brüll

Fall 2025

Lecture 4

Advanced Econometrics

Literature: Greene ch. 2–4, Wooldridge ch. 3–4

4.1.1: FWL Theorem in Equation Algebra

$$y_i = \beta_0 + \beta_2 x_{2,i} + \beta_1 x_{1,i} + \varepsilon_i,$$

where $x_{2,i}$ is the regressor of interest and $x_{1,i}$ is a control.

Frisch-Waugh-Lovell (FWL): The coefficient on $x_{2,i}$ and the residuals from the full model are exactly recovered by either

(a) a regression using **partialled–out variables**:

$$\tilde{y}_i = \beta_0 + \beta_2 \tilde{x}_{2,i} + \varepsilon_i,$$

(b) a regression using **residualized variables**:

$$u_{y,i} = \beta_2 u_{2,i} + \varepsilon_i.$$

Why is the decomposition useful?

The Frisch-Waugh-Lovell theorem is useful because it lets us study the effect of one regressor while <u>controlling for others</u> in a simple way:

- We can visualize the relationship between y_i and $x_{2,i}$ in a two–dimensional scatter plot, once we have partialled out control variables.
- We can partial out high-dimensional controls (e.g. fixed effects) to reduce computation time. This is the principle behind commands such as `reghdfe` in Stata.
- We can separate two sources of variation:
  1. variation in $x_{2,i}$ explained by $x_{1,i}$, and
  2. how y_i responds to the part of $x_{2,i}$ orthogonal to $x_{1,i}$.
- It clarifies where omitted variable bias comes from, by showing exactly how the correlation between $x_{1,i}$ and $x_{2,i}$ matters.

How to partial out $x_{1,i}$

Step 1 (project $x_{2,i}$ on $x_{1,i}$):

$$x_{2,i} = \hat{\gamma}_0 + \hat{\gamma}_1 x_{1,i} + u_{2,i}, \qquad u_{2,i} \perp x_{1,i}.$$

Step 2 (project y_i on $x_{1,i}$):

$$y_i = \hat{\delta}_0 + \hat{\delta}_1 x_{1,i} + u_{y,i}, \qquad u_{y,i} \perp x_{1,i}.$$

Step 3 (define partialled–out variables, keep intercepts):

$$\tilde{x}_{2,i} := \hat{\gamma}_0 + u_{2,i}, \qquad \tilde{y}_i := \hat{\delta}_0 + u_{y,i}.$$

Bivariate regression on adjusted variables: $\boxed{\tilde{y}_i = \beta_0 + \beta_2 \tilde{x}_{2,i} + \varepsilon_i}$

Theorem

For the model

$$y_i = \beta_0 + \beta_2 x_{2,i} + \beta_1 x_{1,i} + \varepsilon_i,$$

the following two bivariate regressions yield the same β_2 *and residuals as the full model:*

$$\tilde{y}_i = \tilde{\beta}_0 + \beta_2 \tilde{x}_{2,i} + \tilde{\varepsilon}_i, \qquad u_{y,i} = \beta_2 u_{2,i} + \varepsilon_i \quad (\textit{no intercept}).$$

Hence, working with partialled–out variables (keeping intercepts) or with residuals (dropping intercepts) is equivalent for estimating β_2 and ε_i.

Partialled-out variables reproduce the full model

Show that

$$y_i = \beta_0 + \beta_2 x_{2,i} + \beta_1 x_{1,i} + \varepsilon_i \tag{1}$$

$$\tilde{y}_i = \beta_0 + \tilde{\beta}_1 \tilde{x}_{2,i} + \tilde{\varepsilon}_i. \tag{2}$$

Plug in the projections

$$y_i = \hat{\delta}_0 + \hat{\delta}_1 x_{1,i} + u_{y,i}, \qquad x_{2,i} = \hat{\gamma}_0 + \hat{\gamma}_1 x_{1,i} + u_{2,i}$$

into equation (1):

$$\begin{aligned}
y_i &= \hat{\delta}_0 + \hat{\delta}_1 x_{1,i} + u_{y,i} \\
&= \beta_0 + \beta_2(\hat{\gamma}_0 + \hat{\gamma}_1 x_{1,i} + u_{2,i}) + \beta_1 x_{1,i} + \varepsilon_i, \\
\tilde{y}_i &= \hat{\delta}_0 + u_{y,i} = \beta_0 + \beta_2(\hat{\gamma}_0 + u_{2,i}) + (\beta_2\hat{\gamma}_1 - \hat{\delta}_1 + \beta_1)x_{1,i} + \varepsilon_i.
\end{aligned}$$

Because we partialled out $x_{1,i}$ using OLS, $x_{1,i}$ is mechanically uncorrelated with $u_{2,i}$ and with $u_{y,i}$. Therefore the regression coefficient on the partialled–out variable $x_{1,i}$ is zero. The equation simplifies with $\tilde{x}_{2,i} = \hat{\gamma}_0 + u_{2,i}$ to

$$\tilde{y}_i = \hat{\delta}_0 + u_{y,i} = \beta_0 + \beta_2(\hat{\gamma}_0 + u_{2,i}) + \varepsilon_i.$$

Partialling out only $x_{2,i}$

If we partial out $x_{2,i}$ but not y_i, then

$$x_{2,i} = \gamma_0 + \gamma_1 x_{1,i} + u_{2,i}, \qquad \tilde{x}_{2,i} = \gamma_0 + u_{2,i}.$$

The regression becomes

$$\begin{aligned} y_i &= \delta_0 + \delta_1 x_{1,i} + u_{y,i} \\ &= (\beta_0 + \delta_1 \bar{x}_1) + \beta_2 \tilde{x}_{2,i} + (\varepsilon_i + \delta_1 x_{1,i}) \\ &= \kappa + \beta_2 \tilde{x}_{2,i} + \epsilon_i. \end{aligned} \tag{1}$$

Here the intercept κ, the residuals ϵ_i, and the standard errors differ from the full model. But the slope β_2 on $\tilde{x}_{2,i}$ is unchanged.

Residualized variables

From the partialled–out form we have

$$\tilde{y}_i = \delta_0 + u_{y,i} \quad = \quad \beta_0 + \beta_2(\gamma_0 + u_{2,i}) + \varepsilon_i.$$

Subtract δ_0:

$$u_{y,i} \quad = \quad \beta_0 - \delta_0 + \beta_2\gamma_0 + \beta_2 u_{2,i} + \varepsilon_i.$$

But by the projection identities,

$$\beta_0 - \delta_0 + \beta_2\gamma_0 = 0,$$

so the constant term cancels.
Thus we obtain the residualized regression:

$$\boxed{u_{y,i} = \beta_2 u_{2,i} + \varepsilon_i}.$$

This is the Frisch–Waugh–Lovell theorem in residualized form: regressing $u_{y,i}$ on $u_{2,i}$ (without intercept) recovers the same β_2 as the full model.

4.1.2: FWL Theorem in Matrix Notation

Why P and M will keep showing up

OLS always decomposes the vector of outcomes $\boldsymbol{y}$ into two orthogonal components:

$$\boldsymbol{y} = \underbrace{\boldsymbol{P}_X \boldsymbol{y}}_{\text{projection onto regressors}} + \underbrace{\boldsymbol{M}_X \boldsymbol{y}}_{\text{orthogonal residuals}},$$

where the matrices

$$\boldsymbol{P}_X = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}', \qquad \boldsymbol{M}_X = \boldsymbol{I} - \boldsymbol{P}_X$$

have the following properties:

- $\boldsymbol{P}_X$ is a **projection matrix** that maps $\boldsymbol{y}$ onto the column space of $\boldsymbol{X}$.
- $\boldsymbol{M}_X$ is a **residual-maker matrix** that removes all variation in $\boldsymbol{y}$ explained by $\boldsymbol{X}$.
- Both are symmetric and idempotent: $\boldsymbol{P}_X' = \boldsymbol{P}_X$, $\boldsymbol{P}_X^2 = \boldsymbol{P}_X$, and similarly for $\boldsymbol{M}_X$.

Key idea for FWL: If we split $\boldsymbol{X}$ into $(\boldsymbol{X}_1, \boldsymbol{X}_2)$, we can first remove the influence of $\boldsymbol{X}_1$ using $\boldsymbol{M}_{X_1}$, then run a regression on the part of $\boldsymbol{X}_2$ that is orthogonal to $\boldsymbol{X}_1$.

Review: Partition of $\boldsymbol{y}$

The OLS model $\boldsymbol{y} = \boldsymbol{X}\hat{\beta} + \boldsymbol{e}$ can be written in matrix form as:

$$\boldsymbol{y} = \hat{\boldsymbol{y}} + \boldsymbol{e} = \boldsymbol{P}_X\boldsymbol{y} + \boldsymbol{M}_X\boldsymbol{y}.$$

This partitions $\boldsymbol{y}$ into two orthogonal pieces:

- $\boldsymbol{P}_X\boldsymbol{y}$: The **fitted part**, spanned by columns of $\boldsymbol{X}$
- $\boldsymbol{M}_X\boldsymbol{y}$: The **residual part**, orthogonal to all columns of $\boldsymbol{X}$

Each term has a clear dimension and meaning:

- $\boldsymbol{y}$: $n \times 1$ vector of data
- $\boldsymbol{P}_X$: $n \times n$ projection matrix
- $\boldsymbol{M}_X$: $n \times n$ residual-maker matrix
- $\boldsymbol{e}$: $n \times 1$ vector of residuals

Orthogonality condition: $\boldsymbol{X}'\boldsymbol{e} = 0$. This is what ensures that OLS residuals are uncorrelated with the regressors.

Decomposing the Normal Equations

OLS minimizes $\|\boldsymbol{y} - \boldsymbol{X}\beta\|^2$, which leads to the normal equations

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

If $\boldsymbol{X}$ is composed of two sets of regressors $(\boldsymbol{X}_1, \boldsymbol{X}_2)$, we can write this in block form:

$$\begin{bmatrix} \boldsymbol{X}_1'\boldsymbol{X}_1 & \boldsymbol{X}_1'\boldsymbol{X}_2 \\ \boldsymbol{X}_2'\boldsymbol{X}_1 & \boldsymbol{X}_2'\boldsymbol{X}_2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_1'\boldsymbol{y} \\ \boldsymbol{X}_2'\boldsymbol{y} \end{bmatrix}.$$

This gives two matrix equations:

$$\boldsymbol{X}_1'\boldsymbol{X}_1\hat{\beta}_1 + \boldsymbol{X}_1'\boldsymbol{X}_2\hat{\beta}_2 = \boldsymbol{X}_1'\boldsymbol{y} \tag{2}$$

$$\boldsymbol{X}_2'\boldsymbol{X}_1\hat{\beta}_1 + \boldsymbol{X}_2'\boldsymbol{X}_2\hat{\beta}_2 = \boldsymbol{X}_2'\boldsymbol{y} \tag{3}$$

Goal: Derive an expression for $\hat{\beta}_2$ that no longer depends on $\hat{\beta}_1$. This is the essence of the FWL theorem in matrix form.

Step 1: Solve for $\hat{\beta}_1$

Starting from Equation (2):

$$\boldsymbol{X}_1'\boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1 + \boldsymbol{X}_1'\boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2 = \boldsymbol{X}_1'\boldsymbol{y}.$$

We isolate $\hat{\boldsymbol{\beta}}_1$:

$$\boldsymbol{X}_1'\boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1 = \boldsymbol{X}_1'\boldsymbol{y} - \boldsymbol{X}_1'\boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2,$$

and multiply by $(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}$:

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

Interpretation: The first term is the coefficient from regressing $\boldsymbol{y}$ on $\boldsymbol{X}_1$ only; the second adjusts for how $\boldsymbol{X}_2$ overlaps with $\boldsymbol{X}_1$.

Step 2: Substitute into the second equation

Plug the expression for $\hat{\boldsymbol{\beta}}_1$ into Equation (3):

$$\boldsymbol{X}_2'\boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1 + \boldsymbol{X}_2'\boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2 = \boldsymbol{X}_2'\boldsymbol{y}.$$

Substitute $\hat{\boldsymbol{\beta}}_1 = (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'(\boldsymbol{y} - \boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2)$:

$$\boldsymbol{X}_2'\boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'(\boldsymbol{y} - \boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2) + \boldsymbol{X}_2'\boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2 = \boldsymbol{X}_2'\boldsymbol{y}.$$

Expand:

$$\boldsymbol{X}_2'\boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{y} - \boldsymbol{X}_2'\boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2 + \boldsymbol{X}_2'\boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2 = \boldsymbol{X}_2'\boldsymbol{y}.$$

Next step: Recognize a familiar projection matrix inside this expression.

Step 3: Identify the projection matrix

The term $\boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'$ is the projection matrix $\boldsymbol{P}_{X_1}$. Use this to rewrite:

$$\boldsymbol{X}_2'\boldsymbol{P}_{X_1}\boldsymbol{y} - \boldsymbol{X}_2'\boldsymbol{P}_{X_1}\boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2 + \boldsymbol{X}_2'\boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2 = \boldsymbol{X}_2'\boldsymbol{y}.$$

Now add and subtract $\boldsymbol{X}_2'\boldsymbol{I}\boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2$ to reveal an $(\boldsymbol{I} - \boldsymbol{P}_{X_1})$ structure:

$$\boldsymbol{X}_2'(\boldsymbol{I} - \boldsymbol{P}_{X_1})\boldsymbol{y} = \boldsymbol{X}_2'(\boldsymbol{I} - \boldsymbol{P}_{X_1})\boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2.$$

Recognize $(\boldsymbol{I} - \boldsymbol{P}_{X_1})$ as the residual-maker matrix $\boldsymbol{M}_{X_1}$:

$$\boldsymbol{X}_2'\boldsymbol{M}_{X_1}\boldsymbol{y} = \boldsymbol{X}_2'\boldsymbol{M}_{X_1}\boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2.$$

Finally, solve for $\hat{\boldsymbol{\beta}}_2$:

$$\hat{\boldsymbol{\beta}}_2 = (\boldsymbol{X}_2'\boldsymbol{M}_{X_1}\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\boldsymbol{M}_{X_1}\boldsymbol{y}.$$

Step 4: Interpretation of the result

We have derived the key matrix formula for the FWL theorem:

$$\boxed{\hat{\boldsymbol{\beta}}_2 = (\boldsymbol{X}_2'\boldsymbol{M}_{X_1}\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\boldsymbol{M}_{X_1}\boldsymbol{y}}.$$

Note that $\boldsymbol{M}_{X_1}$ is symmetric and idempotent:

$$\boldsymbol{M}_{X_1} = \boldsymbol{M}_{X_1}\boldsymbol{M}_{X_1} = \boldsymbol{M}_{X_1}'\boldsymbol{M}_{X_1}.$$

Thus we can rewrite:

$$\hat{\boldsymbol{\beta}}_2 = ((\boldsymbol{M}_{X_1}\boldsymbol{X}_2)'(\boldsymbol{M}_{X_1}\boldsymbol{X}_2))^{-1}(\boldsymbol{M}_{X_1}\boldsymbol{X}_2)'(\boldsymbol{M}_{X_1}\boldsymbol{y}).$$

Interpretation:

- $\tilde{\boldsymbol{X}}_2 = \boldsymbol{M}_{X_1}\boldsymbol{X}_2$: residuals from regressing $\boldsymbol{X}_2$ on $\boldsymbol{X}_1$.
- $\tilde{\boldsymbol{y}} = \boldsymbol{M}_{X_1}\boldsymbol{y}$: residuals from regressing $\boldsymbol{y}$ on $\boldsymbol{X}_1$.

So we can write simply:

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

What FWL tells us about omitted variable bias

FWL gives a clear view of what happens when we omit relevant regressors.

Setup: Partition the true regressor matrix as

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

where X_1 are included and X_2 are omitted variables in the short regression

$$y = X_1 \tilde{\beta}_1 + \tilde{\varepsilon}.$$

By the Frisch–Waugh–Lovell theorem,

$$\tilde{\beta}_1 = (X_1' X_1)^{-1} X_1' y = \beta_1 + (X_1' X_1)^{-1} X_1' X_2 \beta_2$$

Interpretation:

- The bias term $(X_1' X_1)^{-1} X_1' X_2 \beta_2$ arises from projecting X_2 onto X_1.
- Omitted variables X_2 matter for $\tilde{\beta}_1$ only if both:

$$\begin{aligned} X_1' X_2 &\neq 0 \quad \text{(there is correlation between regressors)} \\ \beta_2 &\neq 0 \quad \text{(omitted variables matter for } y\text{).} \end{aligned}$$

Simplified Bivariate Perspective

Consider the true model with one included and one omitted regressor:

$$y_i = \beta_1 x_{1,i} + \beta_2 x_{2,i} + \varepsilon_i.$$

If we omit $x_{2,i}$ and estimate

$$y_i = \tilde{\beta}_1 x_{1,i} + \tilde{\varepsilon}_i,$$

the FWL decomposition implies

$$\tilde{\beta}_1 = \beta_1 + \beta_2 \frac{\mathbf{cov}(x_1, x_2)}{\mathbf{var}(x_1)}.$$

Interpretation:

- The second term is the **omitted variable bias**.
- Bias is positive if x_1 and x_2 move together and both raise y.
- Bias is zero if either:

$$\mathbf{cov}(x_1, x_2) = 0 \quad \text{or} \quad \beta_2 = 0.$$

FWL perspective on OVB: Projections

FWL shows: Bias is just the influence of X_2 transmitted through its correlation with X_1. We can also show it as projection problem.

Step 1: Regress X_2 on X_1:

$$X_2 = P_{X_1} X_2 + M_{X_1} X_2,$$

where $P_{X_1} X_2$ is the part of X_2 explained by X_1.

Step 2: The short regression omits $M_{X_1} X_2$, but keeps $P_{X_1} X_2$ through correlation with X_1.

Implication:

- The bias equals the effect of the omitted variable (β_2) times how strongly X_2 is embedded in X_1.
- When X_1 and X_2 are orthogonal, $P_{X_1} X_2 = 0$ and no bias arises.

4.2: OLS Properties

Finite Sample Properties of OLS (due to A1-A5)

Under Assumptions A1–A5 (linearity, rank, exogeneity, spherical errors, nonstochastic X) the following holds for OLS:

- **Unbiasedness:**
$$E[\hat{\beta}|X] = \beta$$
- **Variance:**
$$\mathrm{Var}[\hat{\beta}|X] = \sigma^2 (X'X)^{-1}$$
- **Gauss–Markov Theorem:** Among all linear unbiased estimators, $\hat{\beta}$ has the smallest variance (**BLUE**).
- **Orthogonality:** $\hat{\beta}$ is uncorrelated with residuals e; fitted values $\hat{y}$ and residuals e are orthogonal.

Distribution under Assumption A6 (Normality)

Recall: With A1–A5 we have

- OLS is unbiased
- Variance formula: $\mathrm{Var}[\hat{\beta}|X] = \sigma^2(X'X)^{-1}$
- OLS is BLUE (Gauss–Markov theorem)

Additional Assumption A6:

$$\varepsilon|X \sim \mathcal{N}(0, \sigma^2 I_n).$$

Implications for finite-sample distribution:

- $\hat{\beta}|X \sim \mathcal{N}(\beta, \sigma^2(X'X)^{-1})$
- *t*- and *F*-statistics have exact finite-sample distributions

Interpretation: A6 is not needed for unbiasedness or efficiency. But it delivers exact finite-sample inference.

4.2.1: Finite Sample Properties

Unbiased Estimation

The least squares estimator can be written as function of the population regression line:

$$\hat{\beta} \;=\; (X'X)^{-1}X'y \;=\; (X'X)^{-1}X'(X\beta + \varepsilon) \;=\; \beta + (X'X)^{-1}X'\varepsilon$$

Now, taking the conditional expectation yields

$$\begin{aligned} E[\hat{\beta}|X] \;&=\; \beta + E[(X'X)^{-1}X'\varepsilon \,|\, X] \;=\; \beta + (X'X)^{-1}X'E[\varepsilon|X] \\ &=\; \beta \quad \text{by assumption A3: Exogeneity} \end{aligned}$$

Applying the law of iterated expectation shows

$$E[\hat{\beta}] \;=\; E[E[\hat{\beta}|X]] \;=\; E[\beta] \;=\; \beta$$

Variance of Least Squares Estimator

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

The conditional variance of $\hat{\beta}$ is

$$\begin{aligned}
\mathrm{Var}[\hat{\beta}\,|\,X] &= E\Big[(\hat{\beta} - E[\hat{\beta}|X])(\hat{\beta} - E[\hat{\beta}|X])' \,\Big|\, X\Big] \\
&= E\Big[(\hat{\beta} - \beta)(\hat{\beta} - \beta)' \,\Big|\, X\Big] \quad \text{(since } E[\hat{\beta}|X] = \beta \text{ by A3)} \\
&= E\Big[\big((X'X)^{-1}X'\varepsilon\big)\big((X'X)^{-1}X'\varepsilon\big)' \,\Big|\, X\Big] \\
&= E\big[(X'X)^{-1}X'\varepsilon\,\varepsilon'X(X'X)^{-1} \,\big|\, X\big] \\
&= (X'X)^{-1}X'\,E[\varepsilon\varepsilon'|X]\,X(X'X)^{-1} \\
&= (X'X)^{-1}X'\,\sigma^2 I_n\,X(X'X)^{-1} \quad \text{by Assumption A4: Homoskedasticity} \\
&= \sigma^2(X'X)^{-1}.
\end{aligned}$$

Law of Total Variance

For any random vector Z and information set X,

$$\mathrm{Var}[Z] = E\big[\mathrm{Var}[Z|X]\big] + \mathrm{Var}\big(E[Z|X]\big).$$

Proof:

$$\begin{aligned}
\mathrm{Var}[Z] &= E\big[(Z - E[Z])(Z - E[Z])'\big] \\
&= E\big[(Z - E[Z|X] + E[Z|X] - E[Z])(Z - E[Z|X] + E[Z|X] - E[Z])'\big] \\
&= E\big[(Z - E[Z|X])(Z - E[Z|X])'\big] \\
&\quad + E\big[(E[Z|X] - E[Z])(E[Z|X] - E[Z])'\big] \\
&\quad + 2E\big[(Z - E[Z|X])(E[Z|X] - E[Z])'\big] .
\end{aligned}$$

The cross term vanishes because $E[Z - E[Z|X] \mid X] = 0$. Thus,

$$\mathrm{Var}[Z] = E\big[\mathrm{Var}[Z|X]\big] + \mathrm{Var}\big(E[Z|X]\big).$$

Interpretation: The total variance equals the average of conditional variances plus the variance of conditional means.

Unconditional Variance of OLS Estimator

We still have to derive the **unconditional** variance of $\hat{\beta}$. By the law of total variance,

$$\begin{aligned}
\mathrm{Var}[\hat{\beta}] &= E\big[\mathrm{Var}[\hat{\beta}|X]\big] + \mathrm{Var}\big(E[\hat{\beta}|X]\big) \\
&= E\big[\sigma^2 (X'X)^{-1}\big] + \mathrm{Var}[\beta] \\
&= \sigma^2 E\big[(X'X)^{-1}\big],
\end{aligned}$$

since population paramter β is nonrandom.

Intuition for OLS Variance

Key idea: The variance of OLS reflects how much $\hat{\beta}$ would change if we drew a new sample.

- Residual variance σ^2 = "background noise" in y.
- Matrix $(X'X)^{-1}$ = "information in X":
  - More spread in $X \Rightarrow (X'X)$ larger $\Rightarrow$ variance of $\hat{\beta}$ smaller.
  - Little variation or multicollinearity $\Rightarrow (X'X)^{-1}$ large $\Rightarrow$ variance of $\hat{\beta}$ large.
- Together:

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

  balances signal in regressors vs. noise in errors.

Analogy: Estimating a mean: more observations $\Rightarrow$ smaller variance. In regression, it's the same idea, but "information" comes from regressor variation.

$$X'X = \begin{bmatrix} n & \sum_i x_{i1} & \sum_i x_{i2} & \cdots & \sum_i x_{iK} \\ \sum_i x_{i1} & \sum_i x_{i1}^2 & \sum_i x_{i1}x_{i2} & \cdots & \sum_i x_{i1}x_{iK} \\ \sum_i x_{i2} & \sum_i x_{i2}x_{i1} & \sum_i x_{i2}^2 & \cdots & \sum_i x_{i2}x_{iK} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_i x_{iK} & \sum_i x_{iK}x_{i1} & \sum_i x_{iK}x_{i2} & \cdots & \sum_i x_{iK}^2 \end{bmatrix}.$$

Intuition:

- $X'X/n$ collects **raw (uncentered) second moments** of the regressors. Although it's not yet in the familiar form, it fully encodes the **variances and covariances** of the X.
- Centering would just adjust by the means.

Sidenote: Structure of $X'y$

$$X'y = \begin{bmatrix} \sum_i y_i \\ \sum_i x_{i1} y_i \\ \sum_i x_{i2} y_i \\ \vdots \\ \sum_i x_{iK} y_i \end{bmatrix}.$$

Intuition:

- $X'y$ collects the **raw (uncentered) cross-moments** between each regressor and the outcome y.
- This mirrors the structure of $X'X$, but for the **relationship between X and y.**
- Centering would only adjust for means, not the underlying covariance structure.

For the bivariate model with intercept

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

we have

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \quad X'X = \begin{bmatrix} n & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{bmatrix}.$$

The inverse is

$$(X'X)^{-1} = \frac{1}{n\sum_i x_i^2 - (\sum_i x_i)^2} \begin{bmatrix} \sum_i x_i^2 & -\sum_i x_i \\ -\sum_i x_i & n \end{bmatrix}.$$

Thus, the covariance matrix of OLS estimates is

$$\mathrm{Var}\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = \sigma^2 (X'X)^{-1} = \frac{\sigma^2}{n\sum_i x_i^2 - (\sum_i x_i)^2} \begin{bmatrix} \sum_i x_i^2 & -\sum_i x_i \\ -\sum_i x_i & n \end{bmatrix}.$$

The Variance–Weights Matrix $(X'X)^{-1}$ in a bivariate case

Entries:

$$\mathrm{Var}[\hat{\beta}_0|X] = \sigma^2 \frac{\sum_i x_i^2}{n\sum_i x_i^2 - (\sum_i x_i)^2},$$

$$\mathrm{Var}[\hat{\beta}_1|X] = \sigma^2 \frac{n}{n\sum_i x_i^2 - (\sum_i x_i)^2},$$

$$\mathrm{Cov}[\hat{\beta}_0, \hat{\beta}_1|X] = -\sigma^2 \frac{\sum_i x_i}{n\sum_i x_i^2 - (\sum_i x_i)^2}.$$

- The denominator $n\sum_i x_i^2 - (\sum_i x_i)^2$ reflects the total variation of x_i.
  - More dispersion in $x_i \Rightarrow$ smaller variances of both estimates.
- The slope variance can be rewritten as $\mathrm{Var}[\hat{\beta}_1|X] = \sigma^2 / \sum_i (x_i - \bar{x})^2$.
  - $\hat{\beta}_1$ is estimated more precisely when x_i are spread out.
- The intercept variance depends on both spread and mean of x_i.
  - If $\bar{x}$ is far from 0, $\mathrm{Var}[\hat{\beta}_0|X]$ increases.
- The covariance term is negative: $\mathrm{Cov}[\hat{\beta}_0, \hat{\beta}_1|X] < 0$.
  - When the slope rises, the intercept tends to fall to fit the same line.

Gauss–Markov Theorem (OLS is BLUE)

We have shown that $\hat{\beta}$ is a conditionally (and unconditionally) unbiased estimator of β. Moreover, $\hat{\beta}$ is a linear estimator, because it is linear in parameters (Assumption A1).

Gauss–Markov Theorem

In the classical linear regression model with regressor matrix *X*, the least squares estimator $\hat{\beta}$ is **efficient** in the class of linear (conditionally) unbiased estimators.

Formally, let b_0 denote any other linear and conditionally unbiased estimator of β. The Gauss–Markov Theorem states that:

$$\mathrm{Var}[b_0|X] - \mathrm{Var}[\hat{\beta}|X] \quad \text{is positive semidefinite.}$$

Interpretation: This means that $\mathrm{Var}[\hat{\beta}|X]$ is the smallest variance matrix in this class. In other words, OLS has **minimal variance** $\Rightarrow$ the Best in **BLUE**.

Gauss–Markov Theorem: Proof (Setup)

Goal: Compare OLS $\hat{\beta} = (X'X)^{-1}X'y$ with any other **linear, unbiased** estimator of β.

Step 1: Write the most general linear estimator.

$$b_0 = Cy, \qquad \text{where } C \text{ is some } (K+1) \times n \text{ matrix of constants.}$$

This is the most general way to express an estimator that is **linear in the data** y. OLS corresponds to the specific choice $C = (X'X)^{-1}X'$.

Step 2: Impose unbiasedness.

$$E[b_0|X] = E[Cy|X] = C\,E[y|X] = C\,X\beta.$$

For b_0 to be unbiased, this must equal β for all possible β:

$$CX\beta = \beta \quad \text{for all } \beta \quad \Rightarrow \quad \boxed{CX = I_{K+1}.}$$

This condition ensures b_0 gives the right average value.

$\Rightarrow$ Any linear, unbiased estimator of β must satisfy $CX = I$. Next, we show that OLS minimizes its variance among all such estimators.

Gauss–Markov Theorem: Proof

Recap so far: We consider any linear unbiased estimator $b_0 = Cy$ satisfying $CX = I_{K+1}$.

Step 3: Compute variances (under A4: spherical errors).

$$\mathrm{Var}[y|X] = \sigma^2 I_n \quad \text{(homoskedastic and uncorrelated errors).}$$

$$\begin{aligned}
\mathrm{Var}[b_0|X] &= C\,\mathrm{Var}[y|X]\,C' \\
&= C\,(\sigma^2 I_n)\,C' \qquad \text{(A4: homoskedasticity \& no autocorrelation)} \\
&= \sigma^2 CC', \\
&= \sigma^2 CC', \\
\mathrm{Var}[\hat{\beta}|X] &= \sigma^2 (X'X)^{-1}.
\end{aligned}$$

Next: Compare these two variances by expressing C as the OLS part plus a "correction" that keeps unbiasedness intact.

Gauss–Markov Theorem: Proof (continued)

Step 4: Express C as OLS part plus deviation.
Let

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

Since both C and $(X'X)^{-1}X'$ satisfy $CX = (X'X)^{-1}X'X = I$, we have

$$DX = 0.$$

$\Rightarrow$ The extra term D does not affect unbiasedness (because it drops out when multiplied by X).

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

Gauss–Markov Theorem: Proof (contd.)

Step 5: Show that the variance difference is positive semidefinite.

$$\begin{aligned} CC' &= \big((X'X)^{-1}X' + D\big)\big((X'X)^{-1}X' + D\big)' \\ &= (X'X)^{-1} + (X'X)^{-1}X'D' + DX(X'X)^{-1} + DD' \\ &= (X'X)^{-1} + DD' \quad (\text{since } DX = 0), \end{aligned}$$

so

$$\mathrm{Var}[b_0|X] - \mathrm{Var}[\hat{\beta}|X] = \sigma^2(CC' - (X'X)^{-1}) = \sigma^2 DD'.$$

Gauss–Markov Theorem: Conclusion

- Any linear, unbiased estimator can be written as

$$b_0 = (X'X)^{-1}X'y + Dy, \quad \text{where } DX = 0.$$

- Its conditional variance is

$$\mathrm{Var}[b_0|X] = \sigma^2(X'X)^{-1} + \sigma^2 DD'.$$

Since $\sigma^2 DD'$ is positive semidefinite:

$$\forall a, \quad a'(\sigma^2 DD')a = \sigma^2 \|D'a\|^2 \geq 0.$$

(Because a squared norm can never be negative.)

Under A1–A4 (classical linear regression model):
$\hat{\beta} = (X'X)^{-1}X'y$ has the smallest variance among all linear unbiased estimators.

OLS is **BLUE** $\Rightarrow$ **Best (minimum variance), Linear, and Unbiased.**

Intuition:

Any other linear unbiased estimator adds a "correction" *Dy* that does not change the mean, but increases the variance by $\sigma^2 DD'$.

Estimating the Error Variance

To compute $\mathrm{Var}[\hat{\beta} \mid X] = \sigma^2 (X'X)^{-1}$ we need an estimate of the unknown error variance σ^2.

Idea: Use the residuals $e = y - \hat{y}$ as proxies for the true errors.

A conditionally unbiased estimator for σ^2 is given by:

$$s^2 = \frac{e'e}{n - (K+1)}.$$

Hence our estimate for $\mathrm{Var}[\hat{\beta} \mid X]$ is

$$\widehat{\mathrm{Var}}[\hat{\beta} \mid X] = s^2 (X'X)^{-1}.$$

Conditional Unbiasedness of s^2

Recall the model:

$$y = X\beta + \varepsilon, \qquad E[\varepsilon \mid X] = 0,\ \mathrm{Var}[\varepsilon \mid X] = \sigma^2 I_n.$$

OLS residuals are

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

Since $y = X\beta + \varepsilon$,

$$e = M(X\beta + \varepsilon) = MX\beta + M\varepsilon = M\varepsilon,$$

because $MX = 0$.

The total squared residuals measure remaining variation:

$$e'e = (M\varepsilon)'(M\varepsilon) = \varepsilon' M' M\varepsilon = \varepsilon' M\varepsilon,$$

since $M'M = M$.

To estimate the unknown variance $\sigma^2 = E[\varepsilon_i^2]$, we average the squared residuals over the $n - (K+1)$ independent directions left after fitting $K+1$ parameters:

$$s^2 = \frac{e'e}{n-(K+1)} = \frac{\varepsilon' M\varepsilon}{n-(K+1)}.$$

Proving Conditional Unbiasedness of s^2

We use two key trace facts for any scalar $a'Ba$:

$$a'Ba = \mathrm{tr}(a'Ba) = \mathrm{tr}(Baa'),$$

where a is $(n \times 1)$ and B is $(n \times n)$. The rule $\mathrm{tr}(AB) = \mathrm{tr}(BA)$ allows cyclic permutation inside the trace.)

Compute the conditional expectation of the residual sum of squares:

$$E[e'e \mid X] = E[\varepsilon' M \varepsilon \mid X] = \mathrm{tr}\big(M\, E[\varepsilon\varepsilon' \mid X]\big).$$

Here:

- ε is $(n \times 1)$: The vector of disturbances,
- M is $(n \times n)$: The residual-maker matrix
- So $\varepsilon' M \varepsilon$ is a scalar

Proving Conditional Unbiasedness of s^2

By Assumption A4 (spherical errors):

$$E[\varepsilon\varepsilon' \mid X] = \sigma^2 I_n,$$

so the conditional covariance of ε is proportional to the identity.

Substitute and use $\mathrm{tr}(M) = n - (K+1)$:

$$E[e'e \mid X] = \mathrm{tr}\big(M\,E[\varepsilon\varepsilon' \mid X]\big) = \sigma^2\,\mathrm{tr}(M) = \sigma^2[n-(K+1)].$$

Therefore,

$$E[s^2 \mid X] = \frac{1}{n-(K+1)} E[e'e \mid X] = \sigma^2.$$

$\boxed{\textbf{Conclusion: } E[s^2 \mid X] = \sigma^2 \;\Rightarrow\; s^2 \text{ is conditionally unbiased.}}$

We now make use of Assumption A6:

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

Theorem (see Greene, Thm. B-103)

If $z \sim \mathcal{N}(\mu, \Sigma)$ then

$$Az + d \sim \mathcal{N}(A\mu + d, A\Sigma A').$$

We apply this theorem to

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

With $A = (X'X)^{-1} X'$ and conditional on X, it follows that

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

And for each element of $\hat{\beta}$:

$$\hat{\beta}_k \,\big|\, X \sim \mathcal{N}\left(\beta_k, \sigma^2 \left[(X'X)^{-1}\right]_{kk}\right).$$

4.2.2 Testing

Testing a Hypothesis about a Coefficient

We want to test

$$H_0 : \beta_k = \beta_{k,0}.$$

Under the normality assumption we can make use of the following test statistic:

$$z_k \; = \; \frac{\hat{\beta}_k - \beta_{k,0}}{\sqrt{\sigma^2 \left(X'X\right)^{-1}_{kk}}}.$$

Conditionally on X, this statistic is standard normal:

$$z_k \mid X \; \sim \; \mathcal{N}(0,1).$$

Problem: We do not observe σ^2. What is the distribution of the test statistic if we replace σ^2 by s^2?

Theorem (see Greene, Thm. B.8)

If $z \sim \mathcal{N}(0, I)$ and A is idempotent, then $z'Az$ has a chi-squared distribution with degrees of freedom equal to the rank of A.

We apply this theorem to

$$\frac{(n-(K+1))s^2}{\sigma^2} = \frac{e'e}{\sigma^2} = \left(\frac{\varepsilon}{\sigma}\right)' M \left(\frac{\varepsilon}{\sigma}\right),$$

which conditionally on X is

$$\chi^2(n-(K+1)).$$

Theorem (see Greene, Thm. 4.4)

If ε is normally distributed, then the least squares estimator $\hat{\beta}$ is statistically independent of the residual vector e and therefore of all functions of e, including s^2.

Independence of Numerator and Denominator

The statistic:

$$t_k = \frac{\hat{\beta}_k - \beta_{k,0}}{\sqrt{s^2 (X'X)^{-1}_{kk}}}.$$

follows a $t-$distribution. We need two ingredients to show this:

1. the numerator and denominator have the right **marginal distributions**, and
2. they are **independent**.

We already know that

$$\frac{(n-(K+1))s^2}{\sigma^2} = \left(\frac{\varepsilon}{\sigma}\right)' M \left(\frac{\varepsilon}{\sigma}\right) \sim \chi^2(n-(K+1)).$$

Next: Are

$$\frac{\hat{\beta}_k - \beta_k}{\sigma\sqrt{(X'X)^{-1}_{kk}}} \quad \text{and} \quad \frac{s^2}{\sigma^2}$$

independent?

Independence of Numerator and Denominator: Express in ε

Start from the linear model

$$y = X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I).$$

Then

$$\hat{\beta} = (X'X)^{-1}X'y = \beta + (X'X)^{-1}X'\varepsilon,$$

and

$$e = My = M\varepsilon, \quad M = I - X(X'X)^{-1}X'.$$

Hence, the random parts of the *t*-statistic are:

$$\frac{\hat{\beta} - \beta}{\sigma} = (X'X)^{-1}X'\frac{\varepsilon}{\sigma}, \quad \frac{e'e}{\sigma^2} = \left(\frac{\varepsilon}{\sigma}\right)' M \left(\frac{\varepsilon}{\sigma}\right).$$

Both are functions of the same random vector ε/σ, so we test independence via their covariance.

Independence of Numerator and Denominator: Covariance Derivation

The denominator s^2 is a quadratic form in ε:

$$s^2 = \frac{1}{n-(K+1)}\,\varepsilon' M \varepsilon.$$

The numerator $(\hat{\beta}-\beta)$ is a linear form in ε:

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

For multivariate normal ε, a standard result says:

Lemma (Greene, Thm. B.12)

If $A\varepsilon$ and $B\varepsilon$ are jointly normal and $E[(A\varepsilon)(B\varepsilon)'] = 0$, then $A\varepsilon$ and $(B\varepsilon)'(B\varepsilon)$ are independent.

Hence we only need to check that the **linear components** generating numerator and denominator are uncorrelated:

$$\operatorname{Cov}\left(M\frac{\varepsilon}{\sigma},\ \frac{\hat{\beta}-\beta}{\sigma}\right) = 0.$$

Substitute $\dfrac{\hat{\beta}-\beta}{\sigma} = (X'X)^{-1}X'\dfrac{\varepsilon}{\sigma}$:

$$E\left[M\frac{\varepsilon}{\sigma}\left(\frac{\hat{\beta}-\beta}{\sigma}\right)'\right] = E\left[M\frac{\varepsilon}{\sigma}\left(\frac{\varepsilon}{\sigma}\right)' X(X'X)^{-1}\right].$$

Using $E[\frac{\varepsilon}{\sigma}(\frac{\varepsilon}{\sigma})'|X] = I$ because $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$,

$$E\left[M\frac{\varepsilon}{\sigma}\left(\frac{\varepsilon}{\sigma}\right)' X(X'X)^{-1} \,\middle|\, X\right] = MX(X'X)^{-1}.$$

Recall $M = I - X(X'X)^{-1}X'$, hence

$$MX = X - X(X'X)^{-1}X'X = X - X = 0.$$

Therefore,

$$MX(X'X)^{-1} = 0 \quad \Rightarrow \quad \mathrm{Cov}\left(M\frac{\varepsilon}{\sigma}, \frac{\hat{\beta} - \beta}{\sigma}\right) = 0.$$

Implication: The vectors $M\varepsilon/\sigma$ and $(\hat{\beta} - \beta)/\sigma$ are uncorrelated, and under normality, this means they are **independent**.

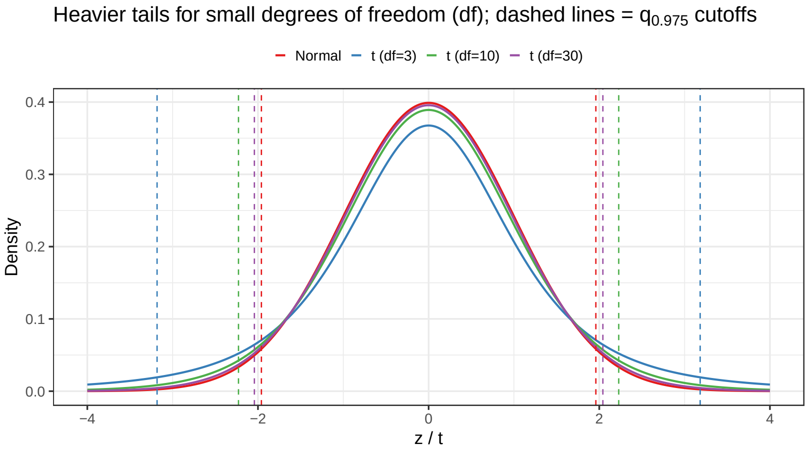
Consequently,

$$\frac{\hat{\beta} - \beta}{\sigma} \text{ is independent of } \frac{e'e}{\sigma^2} = \left(\frac{\varepsilon}{\sigma}\right)' M \left(\frac{\varepsilon}{\sigma}\right).$$

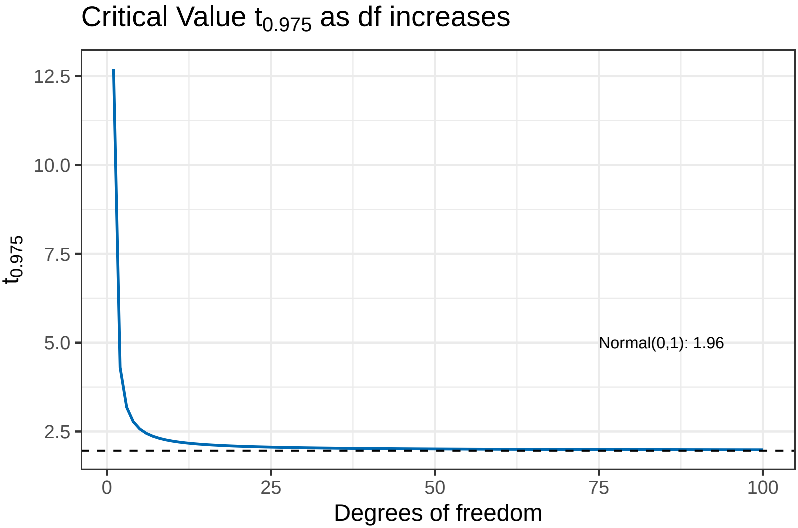
Conclusion: The numerator and denominator of the t−statistic are independent, completing the proof that

$$t_k \sim t_{n-(K+1)}.$$

t-distribution compared to Normal distribution



t-distribution compared to Normal distribution



Data:

$$X = \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 0 \end{bmatrix}, \quad y = \begin{bmatrix} 2.4 \\ 2.7 \\ 2.9 \\ 3.1 \end{bmatrix}.$$

OLS Estimates:

$$\begin{aligned} \hat{\beta}_0 &= 2.9, \\ \hat{\beta}_1 &= -0.25, \\ se(\hat{\beta}_1) &= 0.3202 \end{aligned}$$

Manual computation:

$$t = \frac{\hat{\beta}_1 - 0}{\text{SE}(\hat{\beta}_1)} = \frac{-0.25}{0.3202} = -0.78$$

Decision:

$$|t| = 0.78 < t_{0.975,2} = 4.30$$

$\Rightarrow$ Fail to reject $H_0 : \beta_1 = 0.$

95% Confidence Interval for β_1:

$$\begin{aligned} CI_{95\%}(\beta_1) &= \hat{\beta}_1 \pm t_{0.975,2} \times se(\hat{\beta}_1) \\ &= -0.25 \pm 4.30 \times 0.3202 \\ &= -0.25 \pm 1.38 \end{aligned}$$

$$\Rightarrow \quad CI_{95\%}(\beta_1) = [-1.63, \ 1.13].$$

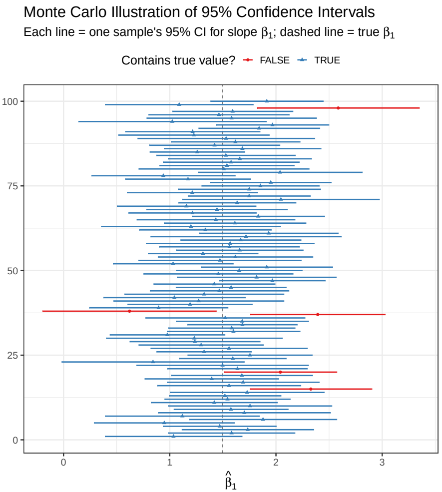
Simualtion of Confidence Intervals across Samples

Simulation setup:

- True model: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
- $\beta_1 = 1.5, \ \varepsilon_i \sim N(0,1)$
- $n = 30$ observations per sample
- Draw 100 Samples

Interpretation:

- Each line: 95% CI for $\hat{\beta}_1$ from one sample
- Dashed line: true β_1
- **Blue:** interval covers β_1
- **Red:** interval misses β_1
- About 95% of CIs contain the truth: Here exactly 5 miss



Testing Multiple Linear Restrictions

Instead of testing a single coefficient, we may want to test

$$H_0 : R\beta = q$$

where

- R is an $r \times K + 1$ restriction matrix of full row rank
- q is an $r \times 1$ vector
- r = number of linear restrictions (e.g., $r = 2$ means testing 2 equations jointly)

Examples:

- $H_0 : \beta_1 = \beta_2 = 0 \;\Rightarrow\; R = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix},\; q = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$
- $H_0 : \beta_1 = \beta_3 \;\Rightarrow\; R = \begin{bmatrix} 0 & 1 & 0 & -1 \end{bmatrix},\; q = \begin{bmatrix} 0 \end{bmatrix}$

(Assuming $\beta = [\beta_0, \beta_1, \beta_2, \beta_3]^\top$ where β_0 is the intercept.)

Testing Multiple Linear Restrictions (contd.)

We want to test r linear restrictions on the regression coefficients:

$$H_0 : R\beta = q, \qquad R \text{ is } r \times K, \quad q \text{ is } r \times 1.$$

Idea:

- Compare model fit between
  1. the unrestricted model (no restrictions), and
  2. the restricted model where $R\beta = q$ holds exactly.
- If H_0 is true, the restricted model should not fit much worse.

The F-statistic formalizes this comparison.

The unrestricted OLS estimator solves

$$\min_{\beta}(y - X\beta)'(y - X\beta) \quad \Rightarrow \quad \hat{\beta}_{UR} = (X'X)^{-1}X'y.$$

Under the classical linear model

$$y = X\beta + \varepsilon, \qquad E[\varepsilon|X] = 0, \quad \mathrm{Var}(\varepsilon|X) = \sigma^2 I_n,$$

we have

$$\hat{\beta}_{UR} \mid X \sim \mathcal{N}\big(\beta, \sigma^2(X'X)^{-1}\big).$$

Residuals:

$$e_{UR} = y - X\hat{\beta}_{UR}.$$

Now impose the restrictions $R\beta = q$ and solve

$$\min_{\beta}(y - X\beta)'(y - X\beta) \quad \text{s.t. } R\beta = q.$$

Use the Lagrangian:

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

where λ is an $r \times 1$ vector of multipliers.

First-order conditions:

$$\begin{aligned} -2X'(y - X\beta) + 2R'\lambda &= 0, \\ R\beta - q &= 0. \end{aligned}$$

Solving for the Restricted Estimator (Step 1)

We start from the first-order conditions under linear restrictions:

$$\begin{bmatrix} X'X & R' \\ R & 0 \end{bmatrix} \begin{bmatrix} \hat{\beta}_R \\ \hat{\lambda} \end{bmatrix} = \begin{bmatrix} X'y \\ q \end{bmatrix}.$$

- The upper block gives the normal equation with Lagrange multipliers:
$$X'X\hat{\beta}_R + R'\hat{\lambda} = X'y.$$
- The lower block encodes the restriction:
$$R\hat{\beta}_R = q.$$
- Substituting $X'y = X'X\hat{\beta}_{UR}$ (from the unrestricted OLS estimator) yields:
$$R'\hat{\lambda} = X'X(\hat{\beta}_{UR} - \hat{\beta}_R).$$

Solving for the Restricted Estimator (Step 2)

From

$$R'\hat{\lambda} = X'X(\hat{\beta}_{UR} - \hat{\beta}_R),$$

post-multiply by $(X'X)^{-1}R'$ and rearrange to isolate $\hat{\lambda}$:

$$\hat{\lambda} = [R(X'X)^{-1}R']^{-1}(R\hat{\beta}_{UR} - q).$$

Substitute this expression back into the first equation to obtain:

$$\boxed{\hat{\beta}_R = \hat{\beta}_{UR} - (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}(R\hat{\beta}_{UR} - q).}$$

- The correction term projects the unrestricted estimate onto the subspace that satisfies $R\beta = q$.
- The matrix $(X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}$ adjusts $\hat{\beta}_{UR}$ just enough to enforce the restrictions.

Under $H_0 : R\beta = q$,

$$R\hat{\beta}_{UR} - q = R(\hat{\beta}_{UR} - \beta) \sim \mathcal{N}\big(0,\, \sigma^2 R(X'X)^{-1}R'\big).$$

Then the quadratic form

$$\frac{1}{\sigma^2}(R\hat{\beta}_{UR} - q)'[R(X'X)^{-1}R']^{-1}(R\hat{\beta}_{UR} - q) \;\sim\; \chi^2_r.$$

Interpretation: This measures how far the sample estimates $R\hat{\beta}_{UR}$ are from the hypothesized values q, scaled by their sampling variance.

The unbiased estimator of σ^2 from the unrestricted model is:

$$s^2 = \frac{e_{UR}'e_{UR}}{n-K}, \qquad \frac{(n-K)s^2}{\sigma^2} \sim \chi^2_{n-K}.$$

Since $X'M = 0$, this χ^2_{n-K} term is **independent** of $(R\hat{\beta}_{UR} - q)$.
Therefore:

$$\begin{aligned} F &= \frac{\left[(R\hat{\beta}_{UR} - q)'[R(X'X)^{-1}R']^{-1}(R\hat{\beta}_{UR} - q)/r\right]}{s^2} \\ &= \frac{\chi^2_r/r}{\chi^2_{n-K}/(n-K)} \; \sim \; F(r, n-K). \end{aligned}$$

Interpretation: The numerator captures the fit loss from imposing $R\beta = q$; the denominator measures the unexplained variance. Under H_0, their ratio follows an F distribution.

Alternative Expression of F-statistic

The F–test can also be written in terms of restricted and unrestricted regression fits:

$$F = \frac{(SSR_R - SSR_{UR})/r}{SSR_{UR}/(n-K)}.$$

where

- SSR_{UR} = sum of squared residuals from unrestricted model
- SSR_R = sum of squared residuals from model estimated under H_0
- r = number of restrictions

Intuition: If restrictions are correct, forcing them should not increase SSR "too much." If SSR_R is much larger than SSR_{UR}, H_0 is rejected.

Special Cases of the F-test

- $r = 1$: $F-$test reduces to the squared $t-$test

$$F(1, n-K) \equiv t^2(n-K).$$

- Joint significance of all slope coefficients:

$$H_0 : \beta_2 = \beta_3 = \cdots = \beta_K = 0.$$

  This is the test of "overall significance" of the regression.

Summary:

- $t-$test: single restriction
- $F-$test: multiple restrictions

4.2.2: OLS in Large Samples

Roadmap for 4.2.2

What we cover (sketches, intuition first; proofs optional):

1. Consistency of OLS: fixed X vs. random X.
2. Asymptotic normality of $\hat{\beta}$.
3. White (heteroskedasticity-robust) variance: the “sandwich”.
4. Homo- vs. heteroskedasticity in large n (what changes?).

Convergence in Probability

Convergence in Probability

A sequence of random variables Z_n **converges in probability** to Z if

$$Z_n \xrightarrow{p} Z \quad \Leftrightarrow \quad \forall \varepsilon > 0: \; P(|Z_n - Z| > \varepsilon) \to 0 \quad \text{as } n \to \infty.$$

Useful Rules for convergence in probability:

If $X_n \xrightarrow{p} a$ and $Y_n \xrightarrow{p} b$, then:

- $X_n + Y_n \xrightarrow{p} a + b$
- $X_n Y_n \xrightarrow{p} ab$
- If $b \neq 0$, then $\frac{X_n}{Y_n} \xrightarrow{p} \frac{a}{b}$
- If $g(\cdot)$ is continuous at a, then $g(X_n) \xrightarrow{p} g(a)$ (Continuous Mapping Thm.)

Why important? Lets us manipulate probability limits just like ordinary limits

Consistency of an Estimator

Definition: An estimator $\hat{\theta}_n$ of parameter θ is consistent if

$$\hat{\theta}_n \xrightarrow{p} \theta.$$

Intuition: As sample size grows, $\hat{\theta}_n$ gets arbitrarily close to the true θ with high probability.

Key ingredients to show consistency:

- Law of Large Numbers (LLN)
- Exogeneity assumptions:
  Errors have mean zero conditional on regressors

Law of Large Numbers (LLN)

(Weak) Law of Large Numbers

If $\{Z_i\}_{i=1}^n$ are IID with $E[Z_i] = \mu$ and $Var(Z_i) < \infty$, then

$$\bar{Z}_n = \frac{1}{n}\sum_{i=1}^{n} Z_i \xrightarrow{p} \mu.$$

Interpretation: The sample average gets arbitrarily close to the population mean as *n* grows.

Examples:

- Toss a coin: $\bar{Z}_n =$ share of heads $\xrightarrow{p} 0.5$.
- Regression context:

$$\tfrac{1}{n}\sum x_i\varepsilon_i \xrightarrow{p} E[x_i\varepsilon_i] = 0.$$

Consistency of OLS (Sketch of a proof)

Recall

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

Rewrite:

$$\hat{\beta} - \beta = \left(\frac{1}{n}X'X\right)^{-1}\left(\frac{1}{n}X'\varepsilon\right).$$

- By LLN: $\frac{1}{n}X'X \xrightarrow{p} Q$ (positive definite).
- By LLN: $\frac{1}{n}X'\varepsilon = \frac{1}{n}\sum x_i\varepsilon_i \xrightarrow{p} 0$ if $E[x_i\varepsilon_i] = 0$ (exogeneity).

Therefore,

$$\hat{\beta} \xrightarrow{p} \beta + Q^{-1} \cdot 0 = \beta.$$

Conclusion: b is a **consistent estimator** of β.

CLT and Convergence in Distribution

Convergence in distribution: $Z_n \xrightarrow{d} Z$ means that the distribution of Z_n approaches that of Z as $n \to \infty$.

Central Limit Theorem (CLT): If $\{Z_i\}$ are IID with $E[Z_i] = 0$, $Var(Z_i) = \sigma^2 < \infty$, then

$$\sqrt{n}\,\overline{Z}_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Z_i \xrightarrow{d} \mathcal{N}(0, \sigma^2).$$

Implication for regression: Sums of random vectors like $\frac{1}{\sqrt{n}} \sum x_i \varepsilon_i$ become approximately normal for large n.

Ingredients for Asymptotic Normality

To study the large-sample behavior of $\hat{\beta}$, we decompose

$$\sqrt{n}(\hat{\beta} - \beta) = \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1} \left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_i \varepsilon_i\right).$$

We need two ingredients for this expression to have a limiting distribution:

1. Regressor matrix (LLN):

$$\frac{1}{n}\sum_{i=1}^{n} x_i x_i' \xrightarrow{p} Q = E[x_i x_i'], \qquad Q \succ 0.$$

2. Score term (CLT):

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_i \varepsilon_i \xrightarrow{d} \mathcal{N}(0, \Sigma), \qquad \text{with } \Sigma = E[\varepsilon_i^2 x_i x_i'].$$

Here $\Sigma = \mathrm{Var}(x_i \varepsilon_i)$ is the variance of the score term.

Moment and exogeneity conditions:

- $E[\varepsilon_i \mid x_i] = 0$ (exogeneity)
- $E[\|x_i\|^2] < \infty, \quad E[\varepsilon_i^2 \|x_i\|^2] < \infty$

These ensure the LLN and CLT apply.

A Quick Reminder: Slutsky's Theorem

Goal: Combine convergence in probability and convergence in distribution.

If

$$X_n \xrightarrow{d} X \quad \text{and} \quad Y_n \xrightarrow{p} c,$$

then

$$Y_n X_n \xrightarrow{d} cX \quad \text{and} \quad X_n + Y_n \xrightarrow{d} X + c.$$

Intuition:

- Random parts (X_n) have limiting distributions.
- Deterministic parts (Y_n) "settle down" to constants.
- Together: stable + random $\Rightarrow$ same limit shape, scaled by the constant.

Here:

$$\underbrace{\left(\frac{1}{n}\sum x_i x_i'\right)^{-1}}_{\xrightarrow{p} Q^{-1}} \underbrace{\left(\frac{1}{\sqrt{n}}\sum x_i \varepsilon_i\right)}_{\xrightarrow{d} \mathcal{N}(0,\Sigma)} \xrightarrow{d} \mathcal{N}(0, Q^{-1}\Sigma Q^{-1}).$$

(Covariance transforms as $C\Sigma C'$ when a normal vector $Z \sim \mathcal{N}(0,\Sigma)$ is multiplied by a matrix C; here $C = Q^{-1}$, hence $Q^{-1}\Sigma Q^{-1}$.)

Why A6 (Normality) is No Longer Needed

Recall A6: $u|X \sim \mathcal{N}(0, \sigma^2 I_n)$ implied exact finite-sample normality of $\hat{\beta}$.

Asymptotics replace A6:

- By LLN: $\frac{1}{n}\sum x_i x_i' \xrightarrow{p} Q$.
- By CLT: $\frac{1}{\sqrt{n}}\sum x_i \varepsilon_i \xrightarrow{d} \mathcal{N}(0, \Sigma)$.
- Slutsky's theorem $\Rightarrow \sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(0, Q^{-1}\Sigma Q^{-1})$.

Key takeaway: Even without normal errors, OLS is asymptotically normal. Exact inference (t, F) requires A6, but robust asymptotic inference does not.

Heteroskedasticity-Robust Variance (White)

Goal:
Do away with homoskedasticity assumption:
Estimate $\mathrm{AVAR}(\hat{\beta}) = \frac{1}{n} Q^{-1} \Sigma Q^{-1}$ without assuming homoskedasticity.

White (HC0) estimator

$$\widehat{\mathrm{Var}}_{\mathsf{rob}}(\hat{\beta}) = (X'X)^{-1} \left(\sum_{i=1}^{n} x_i x_i' e_i^2 \right) (X'X)^{-1}.$$

Variants (finite-sample tweaks): HC1, HC2, HC3.

Sandwich picture:
bread $(X'X)^{-1}$ - toppings $\sum x_i x_i' e_i^2$ - bread $(X'X)^{-1}$

What's in the Sandwich?

Bread: $(X'X)^{-1}$ comes from the usual OLS normal equations

Toppings (the filling):

$$\sum_i x_i x_i' e_i^2$$

- Each observation i contributes $x_i x_i' e_i^2$.
- e_i^2 plays the role of an observation-specific variance.
- $x_i x_i'$ spreads that variance across all covariates according to their values.

Takeaway:
The bread pieces come from the model structure; the filling captures how noisy each observation actually is.

Heteroskedasticity-Robust Variance (White)

Practical note

In applied work, it is common to report robust (heteroskedasticity-consistent) standard errors by default, since the homoskedasticity assumption rarely holds. Variants (HC1–HC3) mainly differ in small-sample adjustments, but all are asymptotically valid.

Extensions for dependent or structured errors:

- **Cluster-robust:** allows arbitrary correlation within clusters (e.g. firms, regions, individuals), but assumes independence across clusters.
- **HAC / Newey–West:** heteroskedasticity- and autocorrelation-consistent, for time series with serial correlation.
- **Spatial-robust:** allows correlation decaying with distance (e.g. Conley standard errors).
- **Panel-robust:** combinations of clustering across two dimensions (e.g. firm and time).

How the Other Sandwiches Look Like

Same recipe, different fillings:
All robust estimators share the general **sandwich form**

$$\widehat{\mathrm{Var}}(\hat{\beta}) = (X'X)^{-1} \left(\sum_{i,j} x_i \, \widehat{\Omega}_{ij} \, x_j' \right) (X'X)^{-1},$$

where $\widehat{\Omega}$ encodes the assumed error covariance structure.

Estimator	**Toppings (middle term)**
White (HC)	$\widehat{\Omega}_{ij} = 0$ if $i \neq j$; $\widehat{\Omega}_{ii} = e_i^2$
Cluster-robust	$\widehat{\Omega} = \mathrm{blockdiag}_g(X_g' e_g e_g' X_g)$
HAC / Newey–West	$\widehat{\Omega}_{ij}$ decays with $\lvert i - j \rvert$ (serial correlation)
Spatial-robust (Conley)	$\widehat{\Omega}_{ij}$ decays with distance d_{ij}
Two-way cluster	Sum of two clustering dimensions minus overlap

Asymptotic t for single coefficients

Null: $H_0 : \beta_k = \beta_{k,0}$. Robust s.e.: $\widehat{\mathrm{se}}_{\mathsf{rob}}(\hat{\beta}_k) = \sqrt{\widehat{\mathrm{Var}}_{\mathsf{rob}}(\hat{\beta})_{kk}}$.

$$t_k^{\mathsf{rob}} = \frac{\hat{\beta}_k - \beta_{k,0}}{\widehat{\mathrm{se}}_{\mathsf{rob}}(\hat{\beta}_k)} \xrightarrow{d} \mathcal{N}(0,1).$$

Interpretation:
Use standard normal critical values asymptotically; in practice, software often reports t with df $n - K$ but based on robust s.e.

Homo- vs. Heteroskedasticity (large n)

	Homoskedasticity	Heteroskedasticity
Consistency of $\hat{\beta}$	Yes	Yes
Asymptotic Var$(\hat{\beta})$	$\sigma^2 Q^{-1}$	$Q^{-1}\Sigma Q^{-1}$
SE to use	classical $(X'X)^{-1}s^2$	**robust** (White/HC)
t/F	classical valid	robust t, Wald/F

References I

Angrist, J. D. and J.-S. Pischke (2009): Mostly Harmless Econometrics: An Empiricist's Companion, Princeton University Press, chapter 3.

Filoso, V. (2013): "Regression Anatomy, Revealed," The Stata Journal, 13, 92–106.

Frisch, R. and F. V. Waugh (1933): "Partial Time Regressions as Compared with Individual Trends," Econometrica, 1, 387–401.

Lovell, M. C. (2008): "A Simple Proof of the FWL Theorem," The Journal of Economic Education, 39, 88–91.