Linear Models Lecture 11: Asymptotic theory for OLS

Robert Gulotty

University of Chicago

February 25, 2026

Today's Claims

- The least squares estimator $\hat{\beta}$ is *consistent* for the projection coefficient β.
- There is an asymptotic normal approximation to the distribution of $\hat{\beta}$ (once normalized).
- Similarly, the error variance estimators $\hat{\sigma}^2$ and s^2 are consistent for σ^2.
- There is a consistent estimator of $\boldsymbol{V}_\beta$, either under homoskedasticity $\hat{\boldsymbol{V}}_\beta^0$, or otherwise.
- We can use z-statistics for asymptotic standard errors, confidence intervals, and prediction intervals.

Strategy: Apply the tools from Lecture 10 — WLLN, CLT, CMT, delta method — to derive the asymptotic properties of OLS.

Today's Claims

- The least squares estimator $\hat{\beta}$ is *consistent* for the projection coefficient β.
- There is an asymptotic normal approximation to the distribution of $\hat{\beta}$ (once normalized).
- Similarly, the error variance estimators $\hat{\sigma}^2$ and s^2 are consistent for σ^2.
- There is a consistent estimator of $\boldsymbol{V}_\beta$, either under homoskedasticity $\hat{\boldsymbol{V}}_\beta^0$, or otherwise.
- We can use z-statistics for asymptotic standard errors, confidence intervals, and prediction intervals.

Strategy: Apply the tools from Lecture 10 — WLLN, CLT, CMT, delta method — to derive the asymptotic properties of OLS.

Target of Estimation

- We are after the projection coefficient $\beta = (\mathbb{E}[\boldsymbol{x}\boldsymbol{x}'])^{-1}\mathbb{E}[\boldsymbol{x}Y]$ from the projection model $Y = \boldsymbol{x}'\beta + e$
- We assume $(Y_i,\ \boldsymbol{x}_i)$ are i.i.d.
- And define $\boldsymbol{Q}_{XX} = \mathbb{E}[\boldsymbol{x}\boldsymbol{x}']$, assume is positive definite.
- We also need moments to exist, including 4th moments for asymptotic normality.

Proof Outline: Two Routes to Consistency

- **Route 1** (Method of Moments):
  - WLLN: sample moments converge to population moments
  - CMT: continuous functions of convergent sequences converge
  - OLS is a continuous function of sample moments
- **Route 2** (Algebraic):
  - Write $\hat{\beta} - \beta$ as a product of sample moments
  - Apply WLLN to each piece, CMT/Slutsky to combine

Both routes use the same tools from Lecture 10. Route 1 generalizes to nonlinear estimators (MLE, GMM); Route 2 gives an explicit formula for the sampling error.

Proof Outline: Two Routes to Consistency

- **Route 1** (Method of Moments):
  - WLLN: sample moments converge to population moments
  - CMT: continuous functions of convergent sequences converge
  - OLS is a continuous function of sample moments
- **Route 2** (Algebraic):
  - Write $\hat{\beta} - \beta$ as a product of sample moments
  - Apply WLLN to each piece, CMT/Slutsky to combine

Both routes use the same tools from Lecture 10. Route 1 generalizes to nonlinear estimators (MLE, GMM); Route 2 gives an explicit formula for the sampling error.

Proof Outline: Two Routes to Consistency

- **Route 1** (Method of Moments):
  - WLLN: sample moments converge to population moments
  - CMT: continuous functions of convergent sequences converge
  - OLS is a continuous function of sample moments
- **Route 2** (Algebraic):
  - Write $\hat{\beta} - \beta$ as a product of sample moments
  - Apply WLLN to each piece, CMT/Slutsky to combine

Both routes use the same tools from Lecture 10. Route 1 generalizes to nonlinear estimators (MLE, GMM); Route 2 gives an explicit formula for the sampling error.

OLS is a function of sample moments

- $\hat{\beta} = \left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i'\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i Y_i\right) = \hat{\boldsymbol{Q}}_{XX}^{-1}\hat{\boldsymbol{Q}}_{XY}$
- $\hat{\boldsymbol{Q}}_{XX}$, $\hat{\boldsymbol{Q}}_{XY}$ are sample moments

OLS is a function of sample moments

- $\hat{\beta} = \left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i'\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i Y_i\right) = \hat{\boldsymbol{Q}}_{XX}^{-1}\hat{\boldsymbol{Q}}_{XY}$
- $\hat{\boldsymbol{Q}}_{XX}$, $\hat{\boldsymbol{Q}}_{XY}$ are sample moments

Apply WLLN

- If $(\boldsymbol{x}_i,\ Y_i)$ are iid, then $\boldsymbol{x}_i\boldsymbol{x}_i'$, $\boldsymbol{x}_i Y_i'$ are iid.
- Recall from Lecture 10: WLLN says that if θ_i are iid with finite variance, $\frac{1}{n}\sum_{i=1}^{n}\theta_i \xrightarrow{p} \mathbb{E}[\theta]$
- $\hat{\boldsymbol{Q}}_{XX} \xrightarrow{p} \boldsymbol{Q}_{XX}, \quad \hat{\boldsymbol{Q}}_{XY} \xrightarrow{p} \boldsymbol{Q}_{XY}$

Apply WLLN

- If $(\boldsymbol{x}_i, \ Y_i)$ are iid, then $\boldsymbol{x}_i \boldsymbol{x}_i'$, $\boldsymbol{x}_i Y_i'$ are iid.
- Recall from Lecture 10: WLLN says that if θ_i are iid with finite variance, $\frac{1}{n}\sum_{i=1}^{n} \theta_i \xrightarrow{p} \mathbb{E}[\theta]$
- $\hat{\boldsymbol{Q}}_{XX} \xrightarrow{p} \boldsymbol{Q}_{XX}, \quad \hat{\boldsymbol{Q}}_{XY} \xrightarrow{p} \boldsymbol{Q}_{XY}$

Apply WLLN

- If $(\boldsymbol{x}_i,\ Y_i)$ are iid, then $\boldsymbol{x}_i\boldsymbol{x}_i'$, $\boldsymbol{x}_i Y_i'$ are iid.
- Recall from Lecture 10: WLLN says that if θ_i are iid with finite variance, $\frac{1}{n}\sum_{i=1}^n \theta_i \xrightarrow{p} \mathbb{E}[\theta]$
- $\hat{\boldsymbol{Q}}_{XX} \xrightarrow{p} \boldsymbol{Q}_{XX}, \quad \hat{\boldsymbol{Q}}_{XY} \xrightarrow{p} \boldsymbol{Q}_{XY}$

Continuous Mapping Theorem (CMT)

- Recall from Lecture 10: if $Z_n \to_p c$ and $g(u)$ is continuous at c, then $g(Z_n) \to_p g(c)$.
- CMT tells us that plims survive continuous functions.
- $\hat{\beta} = g(\hat{\boldsymbol{Q}}_{XX}, \hat{\boldsymbol{Q}}_{XY})$ is a continuous function (matrix inversion is continuous at nonsingular matrices).
- $\hat{\beta} = \hat{\boldsymbol{Q}}_{XX}^{-1}\hat{\boldsymbol{Q}}_{XY} \to_p \boldsymbol{Q}_{XX}^{-1}\boldsymbol{Q}_{XY} = \beta$

Continuous Mapping Theorem (CMT)

- Recall from Lecture 10: if $Z_n \rightarrow_p c$ and $g(u)$ is continuous at c, then $g(Z_n) \rightarrow_p g(c)$.
- CMT tells us that plims survive continuous functions.
- $\hat{\beta} = g(\hat{Q}_{XX}, \hat{Q}_{XY})$ is a continuous function (matrix inversion is continuous at nonsingular matrices).
- $\hat{\beta} = \hat{Q}_{XX}^{-1} \hat{Q}_{XY} \rightarrow_p Q_{XX}^{-1} Q_{XY} = \beta$

Continuous Mapping Theorem (CMT)

- Recall from Lecture 10: if $Z_n \rightarrow_p c$ and $g(u)$ is continuous at c, then $g(Z_n) \rightarrow_p g(c)$.
- CMT tells us that plims survive continuous functions.
- $\hat{\beta} = g(\hat{\boldsymbol{Q}}_{XX}, \hat{\boldsymbol{Q}}_{XY})$ is a continuous function (matrix inversion is continuous at nonsingular matrices).
- $\hat{\beta} = \hat{\boldsymbol{Q}}_{XX}^{-1}\hat{\boldsymbol{Q}}_{XY} \rightarrow_p \boldsymbol{Q}_{XX}^{-1}\boldsymbol{Q}_{XY} = \beta$

Continuous Mapping Theorem (CMT)

- Recall from Lecture 10: if $Z_n \rightarrow_p c$ and $g(u)$ is continuous at c, then $g(Z_n) \rightarrow_p g(c)$.
- CMT tells us that plims survive continuous functions.
- $\hat{\beta} = g(\hat{\boldsymbol{Q}}_{XX}, \hat{\boldsymbol{Q}}_{XY})$ is a continuous function (matrix inversion is continuous at nonsingular matrices).
- $\hat{\beta} = \hat{\boldsymbol{Q}}_{XX}^{-1} \hat{\boldsymbol{Q}}_{XY} \rightarrow_p \boldsymbol{Q}_{XX}^{-1} \boldsymbol{Q}_{XY} = \beta$

Proof of Consistency of OLS: Route 2, Step 1

$$
\begin{aligned}
\hat{\beta} &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} && \text{(Definition)}\\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\beta + \boldsymbol{e}) && \text{(Assumption 1)}\\
&= \beta + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{e} && \text{(Inverse prop.)}\\
&= (\beta + (1/n)n(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{e}) && \\
&= (\beta + (\frac{1}{n}(\frac{1}{n})^{-1}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{e})) && \text{(trick)}\\
\hat{\beta} &= \beta + (\frac{1}{n}\boldsymbol{X}'\boldsymbol{X})^{-1})(\frac{1}{n}\boldsymbol{X}'\boldsymbol{e}) &&
\end{aligned}
$$

Proof of Consistency of OLS: Route 2, Step 1

$$
\begin{aligned}
\hat{\beta} &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} && \text{(Definition)} \\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\beta + \boldsymbol{e}) && \text{(Assumption 1)} \\
&= \beta + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{e} && \text{(Inverse prop.)} \\
&= (\beta + (1/n)n(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{e}) \\
&= (\beta + (\frac{1}{n}(\frac{1}{n})^{-1}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{e})) && \text{(trick)} \\
\hat{\beta} &= \beta + (\frac{1}{n}\boldsymbol{X}'\boldsymbol{X})^{-1})(\frac{1}{n}\boldsymbol{X}'\boldsymbol{e})
\end{aligned}
$$

Proof of Consistency of OLS: Route 2, Step 1

$$\begin{aligned}
\hat{\beta} &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} && \text{(Definition)} \\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\beta + \boldsymbol{e}) && \text{(Assumption 1)} \\
&= \beta + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{e} && \text{(Inverse prop.)} \\
&= (\beta + (1/n)n(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{e}) \\
&= (\beta + (\frac{1}{n}(\frac{1}{n})^{-1}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{e})) && \text{(trick)} \\
\hat{\beta} &= \beta + (\frac{1}{n}\boldsymbol{X}'\boldsymbol{X})^{-1})(\frac{1}{n}\boldsymbol{X}'\boldsymbol{e})
\end{aligned}$$

Proof of Consistency of OLS: Route 2, Step 1

$$
\begin{aligned}
\hat{\beta} &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} && \text{(Definition)} \\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\beta + \boldsymbol{e}) && \text{(Assumption 1)} \\
&= \beta + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{e} && \text{(Inverse prop.)} \\
&= (\beta + (1/n)n(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{e}) \\
&= (\beta + (\frac{1}{n}(\frac{1}{n})^{-1}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{e})) && \text{(trick)} \\
\hat{\beta} &= \beta + (\frac{1}{n}\boldsymbol{X}'\boldsymbol{X})^{-1})(\frac{1}{n}\boldsymbol{X}'\boldsymbol{e})
\end{aligned}
$$

Proof of Consistency of OLS: Route 2, Step 1

$$
\begin{aligned}
\hat{\beta} &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} && \text{(Definition)} \\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\beta + \boldsymbol{e}) && \text{(Assumption 1)} \\
&= \beta + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{e} && \text{(Inverse prop.)} \\
&= (\beta + (1/n)n(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{e}) \\
&= (\beta + (\frac{1}{n}(\frac{1}{n})^{-1}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{e})) && \text{(trick)}
\end{aligned}
$$

$$\hat{\beta} = \beta + (\frac{1}{n}\boldsymbol{X}'\boldsymbol{X})^{-1})(\frac{1}{n}\boldsymbol{X}'\boldsymbol{e})$$

Proof of Consistency of OLS: Route 2, Step 1

$$
\begin{aligned}
\hat{\beta} &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} && \text{(Definition)}\\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\beta + \boldsymbol{e}) && \text{(Assumption 1)}\\
&= \beta + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{e} && \text{(Inverse prop.)}\\
&= (\beta + (1/n)n(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{e})\\
&= (\beta + (\frac{1}{n}(\frac{1}{n})^{-1}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{e})) && \text{(trick)}\\
\hat{\beta} &= \beta + (\frac{1}{n}\boldsymbol{X}'\boldsymbol{X})^{-1})(\frac{1}{n}\boldsymbol{X}'\boldsymbol{e})
\end{aligned}
$$

Proof of Consistency of OLS: Route 2, Step 2

$$\hat{\beta} = \beta + (\frac{1}{n}\boldsymbol{X'X})^{-1})(\frac{1}{n}\boldsymbol{X'e})$$

$$\begin{aligned} \text{plim}\hat{\beta} - \beta &= ((\text{plim}\frac{1}{n}\boldsymbol{X'X})^{-1})\text{plim}\frac{1}{n}\boldsymbol{X'e} && \text{(Slutsky)} \\ &= \boldsymbol{Q}_{XX}^{-1}\mathbb{E}[Xe] \\ &= \boldsymbol{Q}_{XX}^{-1}0 \end{aligned}$$

The least squares estimator $\hat{\beta}$ is *consistent* for the projection coefficient β.

Key condition: Consistency requires $\mathbb{E}[xe] = 0$. When this fails (endogeneity), OLS is inconsistent — motivating instrumental variables (Lecture 13).

Proof of Consistency of OLS: Route 2, Step 2

$$\hat{\beta} = \beta + (\frac{1}{n}\boldsymbol{X'X})^{-1})(\frac{1}{n}\boldsymbol{X'e})$$

$$\text{plim}\hat{\beta} - \beta = ((\text{plim}\frac{1}{n}\boldsymbol{X'X})^{-1})\text{plim}\frac{1}{n}\boldsymbol{X'e} \qquad \text{(Slutsky)}$$

$$= \boldsymbol{Q}_{XX}^{-1}\mathbb{E}[Xe]$$

$$= \boldsymbol{Q}_{XX}^{-1}0$$

The least squares estimator $\hat{\beta}$ is *consistent* for the projection coefficient β.

Key condition: Consistency requires $\mathbb{E}[xe] = 0$. When this fails (endogeneity), OLS is inconsistent — motivating instrumental variables (Lecture 13).

Proof of Consistency of OLS: Route 2, Step 2

$$\hat{\beta} = \beta + (\frac{1}{n}\boldsymbol{X'X})^{-1})(\frac{1}{n}\boldsymbol{X'e})$$

$$\text{plim}\hat{\beta} - \beta = ((\text{plim}\frac{1}{n}\boldsymbol{X'X})^{-1})\text{plim}\frac{1}{n}\boldsymbol{X'e} \qquad \text{(Slutsky)}$$

$$= \boldsymbol{Q}_{XX}^{-1}\mathbb{E}[Xe]$$

$$= \boldsymbol{Q}_{XX}^{-1}0$$

The least squares estimator $\hat{\beta}$ is *consistent* for the projection coefficient β.

Key condition: Consistency requires $\mathbb{E}[xe] = 0$. When this fails (endogeneity), OLS is inconsistent — motivating instrumental variables (Lecture 13).

Proof of Consistency of OLS: Route 2, Step 2

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\frac{1}{n}\boldsymbol{X'X})^{-1})(\frac{1}{n}\boldsymbol{X'e})$$

$$\text{plim}\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = ((\text{plim}\frac{1}{n}\boldsymbol{X'X})^{-1})\text{plim}\frac{1}{n}\boldsymbol{X'e} \qquad \text{(Slutsky)}$$

$$= \boldsymbol{Q}_{XX}^{-1}\mathbb{E}[Xe]$$

$$= \boldsymbol{Q}_{XX}^{-1}0$$

The least squares estimator $\hat{\beta}$ is *consistent* for the projection coefficient β.

Key condition: Consistency requires $\mathbb{E}[\boldsymbol{x}e] = 0$. When this fails (endogeneity), OLS is inconsistent — motivating instrumental variables (Lecture 13).

Asymptotic Normality: Algebraic Setup

Start from the sampling error decomposition:

$$\hat{\beta} - \beta = \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} X_i e_i\right) \tag{1}$$

Multiply both sides by $\sqrt{n}$:

$$\sqrt{n}(\hat{\beta} - \beta) = \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)^{-1} \left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i e_i\right) \tag{2}$$

$$= \hat{Q}_{XX}^{-1} \left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i e_i\right) \tag{3}$$

Now we need the distribution of $\frac{1}{\sqrt{n}}\sum X_i e_i$. This is a sum of mean-zero iid random vectors — exactly the setting of the **multivariate CLT** from Lecture 10.

Asymptotic Normality: Algebraic Setup

Start from the sampling error decomposition:

$$\hat{\beta} - \beta = \left(\frac{1}{n} \sum_{i=1}^{n} X_i X_i' \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} X_i e_i \right) \tag{1}$$

Multiply both sides by $\sqrt{n}$:

$$\sqrt{n}(\hat{\beta} - \beta) = \left(\frac{1}{n} \sum_{i=1}^{n} X_i X_i' \right)^{-1} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i e_i \right) \tag{2}$$

$$= \hat{\boldsymbol{Q}}_{XX}^{-1} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i e_i \right) \tag{3}$$

Now we need the distribution of $\frac{1}{\sqrt{n}} \sum X_i e_i$. This is a sum of mean-zero iid random vectors — exactly the setting of the **multivariate CLT** from Lecture 10.

Asymptotic Normality: Algebraic Setup

Start from the sampling error decomposition:

$$\hat{\beta} - \beta = \left(\frac{1}{n} \sum_{i=1}^{n} X_i X_i' \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} X_i e_i \right) \tag{1}$$

Multiply both sides by $\sqrt{n}$:

$$\sqrt{n}(\hat{\beta} - \beta) = \left(\frac{1}{n} \sum_{i=1}^{n} X_i X_i' \right)^{-1} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i e_i \right) \tag{2}$$

$$= \hat{\boldsymbol{Q}}_{XX}^{-1} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i e_i \right) \tag{3}$$

Now we need the distribution of $\frac{1}{\sqrt{n}} \sum X_i e_i$. This is a sum of mean-zero iid random vectors — exactly the setting of the **multivariate CLT** from Lecture 10.

Asymptotic Normality: Applying CLT and Slutsky

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \underbrace{\hat{\boldsymbol{Q}}_{XX}^{-1}}_{\xrightarrow{p} \boldsymbol{Q}_{XX}^{-1}} \cdot \underbrace{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i e_i}_{\xrightarrow{d} N(0,\Omega)}$$

where $\Omega = \mathbb{E}[X_i e_i e_i' X_i'] = \mathbb{E}[X_i X_i' e_i^2]$.
By Slutsky's theorem (convergent in probability $\times$ convergent in distribution):

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} \boldsymbol{Q}_{XX}^{-1} \cdot N(0, \Omega) \tag{4}$$

$$= N\left(0, \boldsymbol{Q}_{XX}^{-1} \Omega \boldsymbol{Q}_{XX}^{-1}\right) \equiv \boldsymbol{V}_{\beta} \tag{5}$$

Same structure as Probit MLE (Lecture 9): $\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{d} N(0, \mathcal{I}^{-1})$. In both cases, CLT gives the numerator; WLLN/CMT stabilize the denominator; Slutsky combines them.

Asymptotic Normality: Applying CLT and Slutsky

$$\sqrt{n}(\hat{\beta}-\beta) = \underbrace{\hat{\boldsymbol{Q}}_{XX}^{-1}}_{\xrightarrow{p} \boldsymbol{Q}_{XX}^{-1}} \cdot \underbrace{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i e_i}_{\xrightarrow{d} N(0,\Omega)}$$

where $\Omega = \mathbb{E}[X_i e_i e_i' X_i'] = \mathbb{E}[X_i X_i' e_i^2]$.

By Slutsky's theorem (convergent in probability $\times$ convergent in distribution):

$$\sqrt{n}(\hat{\beta}-\beta) \xrightarrow{d} \boldsymbol{Q}_{XX}^{-1} \cdot N(0,\Omega) \tag{4}$$

$$= N\left(0, \boldsymbol{Q}_{XX}^{-1}\Omega\boldsymbol{Q}_{XX}^{-1}\right) \equiv \boldsymbol{V}_\beta \tag{5}$$

Same structure as Probit MLE (Lecture 9): $\sqrt{n}(\hat{\beta}-\beta_0) \xrightarrow{d} N(0,\mathcal{I}^{-1})$. In both cases, CLT gives the numerator; WLLN/CMT stabilize the denominator; Slutsky combines them.

Asymptotic Normality: Applying CLT and Slutsky

$$\sqrt{n}(\hat{\beta}-\beta) = \underbrace{\hat{\boldsymbol{Q}}_{XX}^{-1}}_{\xrightarrow{p} \boldsymbol{Q}_{XX}^{-1}} \cdot \underbrace{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i e_i}_{\xrightarrow{d} N(0,\Omega)}$$

where $\Omega = \mathbb{E}[X_i e_i e_i' X_i'] = \mathbb{E}[X_i X_i' e_i^2]$.
By Slutsky's theorem (convergent in probability $\times$ convergent in distribution):

$$\sqrt{n}(\hat{\beta}-\beta) \xrightarrow{d} \boldsymbol{Q}_{XX}^{-1} \cdot N(0,\Omega) \tag{4}$$

$$= N\left(0, \boldsymbol{Q}_{XX}^{-1}\Omega\boldsymbol{Q}_{XX}^{-1}\right) \equiv \boldsymbol{V}_\beta \tag{5}$$

Same structure as Probit MLE (Lecture 9): $\sqrt{n}(\hat{\beta}-\beta_0) \xrightarrow{d} N(0,\mathcal{I}^{-1})$. In both cases, CLT gives the numerator; WLLN/CMT stabilize the denominator; Slutsky combines them.

Asymptotic Normality: Applying CLT and Slutsky

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \underbrace{\hat{\boldsymbol{Q}}_{XX}^{-1}}_{\xrightarrow{p} \boldsymbol{Q}_{XX}^{-1}} \cdot \underbrace{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i e_i}_{\xrightarrow{d} N(0,\Omega)}$$

where $\Omega = \mathbb{E}[X_i e_i e_i' X_i'] = \mathbb{E}[X_i X_i' e_i^2]$.
By Slutsky's theorem (convergent in probability $\times$ convergent in distribution):

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} \boldsymbol{Q}_{XX}^{-1} \cdot N(0, \Omega) \tag{4}$$

$$= N\left(0, \boldsymbol{Q}_{XX}^{-1} \Omega \boldsymbol{Q}_{XX}^{-1}\right) \equiv \boldsymbol{V}_{\beta} \tag{5}$$

Same structure as Probit MLE (Lecture 9): $\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{d} N(0, \mathcal{I}^{-1})$. In both cases, CLT gives the numerator; WLLN/CMT stabilize the denominator; Slutsky combines them.

Covariance of coefficient estimators under homoskedasticity

- Assume $Y = \beta_1 X_1 + \beta_2 X_2 + e$ with $E[ee'] = \sigma^2 \boldsymbol{I}$.
- Suppose $E[X_1] = 0,\ E[X_2] = 0,\ Var[X_1] = 1,\ Var[X_2] = 1,\ corr[X_1, X_2] = \rho$

$$\boldsymbol{V}_{\beta}^{0} = \sigma^2 \boldsymbol{Q}_{XX}^{-1} = \sigma^2 \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}^{-1} = \frac{\sigma^2}{1-\rho^2} \begin{bmatrix} 1 & -\rho \\ -\rho & 1 \end{bmatrix}$$

- If X_1 and X_2 are positively correlated, $\hat{\beta}_1$ and $\hat{\beta}_2$ are negatively correlated.

Heteroskedastic Covariance Matrix Estimation

- The asymptotic covariance matrix of $\sqrt{n}(\hat{\beta} - \beta)$ is $\boldsymbol{Q}_{XX}^{-1}\Omega\boldsymbol{Q}_{XX}^{-1}$.
- This is the **sandwich form** from Lecture 6 — now with its asymptotic justification.

Estimator:

$$\hat{\Omega} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i'\hat{e}_i^2$$

$$\hat{\boldsymbol{Q}}_{XX} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i'$$

$$\hat{\boldsymbol{V}}_{\beta}^{HC0} = \hat{\boldsymbol{Q}}_{XX}^{-1}\hat{\Omega}\hat{\boldsymbol{Q}}_{XX}^{-1}$$

This is the HC0 estimator. Recall from Lecture 6: HC1, HC2, HC3 apply finite-sample corrections. All are asymptotically equivalent.

Heteroskedastic Covariance Matrix Estimation

- The asymptotic covariance matrix of $\sqrt{n}(\hat{\beta} - \beta)$ is $\boldsymbol{Q}_{XX}^{-1}\Omega\boldsymbol{Q}_{XX}^{-1}$.
- This is the **sandwich form** from Lecture 6 — now with its asymptotic justification.

Estimator:

$$\hat{\Omega} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'\hat{e}_i^2$$

$$\hat{\boldsymbol{Q}}_{XX} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'$$

$$\hat{\boldsymbol{V}}_{\beta}^{HC0} = \hat{\boldsymbol{Q}}_{XX}^{-1}\hat{\Omega}\hat{\boldsymbol{Q}}_{XX}^{-1}$$

This is the HC0 estimator. Recall from Lecture 6: HC1, HC2, HC3 apply finite-sample corrections. All are asymptotically equivalent.

Heteroskedastic Covariance Matrix Estimation

- The asymptotic covariance matrix of $\sqrt{n}(\hat{\beta} - \beta)$ is $\boldsymbol{Q}_{XX}^{-1} \Omega \boldsymbol{Q}_{XX}^{-1}$.
- This is the **sandwich form** from Lecture 6 — now with its asymptotic justification.

Estimator:

$$\hat{\Omega} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' \hat{e}_i^2$$

$$\hat{\boldsymbol{Q}}_{XX} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i'$$

$$\hat{\boldsymbol{V}}_{\beta}^{HC0} = \hat{\boldsymbol{Q}}_{XX}^{-1} \hat{\Omega} \hat{\boldsymbol{Q}}_{XX}^{-1}$$

This is the HC0 estimator. Recall from Lecture 6: HC1, HC2, HC3 apply finite-sample corrections. All are asymptotically equivalent.

Consistency of $\hat{\Omega}$

$$
\begin{aligned}
\hat{\Omega} &= \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' \hat{e}_i^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' \hat{e}_i^2 + \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' e_i^2 - \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' e_i^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' e_i^2 + \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' (\hat{e}_i^2 - e_i^2)
\end{aligned}
$$

$$
\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' e_i^2 \underset{p}{\rightarrow} \mathbb{E}[\boldsymbol{x}\boldsymbol{x}' e^2] = \Omega \qquad \text{(WLLN)}
$$

$$
\hat{\Omega} \underset{p}{\rightarrow} \Omega
$$

Consistency of $\hat{\Omega}$

$$
\begin{aligned}
\hat{\Omega} &= \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' \hat{e}_i^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' \hat{e}_i^2 + \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' e_i^2 - \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' e_i^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' e_i^2 + \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' (\hat{e}_i^2 - e_i^2)
\end{aligned}
$$

$$
\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' e_i^2 \underset{p}{\to} \mathbb{E}[\boldsymbol{x}\boldsymbol{x}' e^2] = \Omega \qquad \text{(WLLN)}
$$

$$
\hat{\Omega} \underset{p}{\to} \Omega
$$

The Sandwich Pattern

The same "bread–meat–bread" structure appears throughout this course:

Model	Bread	Meat (Ω)	Simplification
OLS	Q_{XX}^{-1}	$\mathbb{E}[XX'e^2]$	$\sigma^2 Q_{XX}^{-1}$ if homosk.
Probit MLE	$\mathcal{I}^{-1}$	$\mathbb{E}[s_i s_i']$	$\mathcal{I}^{-1}$ if correct spec.
GMM	$(G'WG)^{-1}G'W$	$\mathbb{E}[g_i g_i']$	—

- Under "ideal" conditions (homoskedasticity / correct specification), the sandwich simplifies.
- Under misspecification, we need the full sandwich — **robust SEs**.
- Same `sandwich::vcovHC()` in R handles all cases.

The Sandwich Pattern

The same "bread–meat–bread" structure appears throughout this course:

Model	**Bread**	**Meat** (Ω)	**Simplification**
OLS	Q_{XX}^{-1}	$\mathbb{E}[XX'e^2]$	$\sigma^2 Q_{XX}^{-1}$ if homosk.
Probit MLE	$\mathcal{I}^{-1}$	$\mathbb{E}[s_i s_i']$	$\mathcal{I}^{-1}$ if correct spec.
GMM	$(G'WG)^{-1}G'W$	$\mathbb{E}[g_i g_i']$	—

- Under "ideal" conditions (homoskedasticity / correct specification), the sandwich simplifies.
- Under misspecification, we need the full sandwich — **robust SEs**.
- Same `sandwich::vcovHC()` in R handles all cases.

Variances: Exact vs Asymptotic

Object	Formula	When to use
Exact variance	$(\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X})(\boldsymbol{X}'\boldsymbol{X})^{-1}$	Fixed X, any n
(homosk.)	$\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$	$+E[ee'] = \sigma^2 I$
Asy. variance	$\boldsymbol{Q}_{XX}^{-1}\Omega\boldsymbol{Q}_{XX}^{-1}$	Random X, n large
(homosk.)	$\sigma^2\boldsymbol{Q}_{XX}^{-1}$	$+E[e^2 \mid X] = \sigma^2$

- The exact variance conditions on X; the asymptotic variance averages over X.
- For large n: $\frac{1}{n}\boldsymbol{X}'\boldsymbol{X} \approx Q_{XX}$, so they agree.
- In practice, we always estimate using the sandwich (or its homoskedastic special case).

Variances: Exact vs Asymptotic

Object	**Formula**	**When to use**
Exact variance	$(\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X})(\boldsymbol{X}'\boldsymbol{X})^{-1}$	Fixed X, any n
(homosk.)	$\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$	$+E[ee'] = \sigma^2 I$
Asy. variance	$\boldsymbol{Q}_{XX}^{-1}\Omega\boldsymbol{Q}_{XX}^{-1}$	Random X, n large
(homosk.)	$\sigma^2\boldsymbol{Q}_{XX}^{-1}$	$+E[e^2\|X] = \sigma^2$

- The exact variance conditions on X; the asymptotic variance averages over X.
- For large n: $\frac{1}{n}\boldsymbol{X}'\boldsymbol{X} \approx Q_{XX}$, so they agree.
- In practice, we always estimate using the sandwich (or its homoskedastic special case).

From F Tests to Wald Tests

- We saw we can calculate an F test by forming a matrix of restrictions.
- The F distribution result depended on the homoskedasticity and normality of the errors.
- In large samples, we can instead rely on the asymptotic behavior of random variables and calculate a **Wald statistic**.
- Many test statistics are reflavored Wald distances. Given a null that $E[\boldsymbol{q}] = \boldsymbol{\theta}$:

$$W = (\boldsymbol{q} - \boldsymbol{\theta})'[\widehat{Var}(\boldsymbol{q} - \boldsymbol{\theta})]^{-1}(\boldsymbol{q} - \boldsymbol{\theta})$$

- $W \sim \chi^2_J$ if $\boldsymbol{q}$ is Normal and $Var(\boldsymbol{q} - \boldsymbol{\theta}) = \Sigma$
- $W \overset{a}{\sim} \chi^2_J$ with CLT on $\boldsymbol{q}$ and consistency for $\hat{\Sigma}$

From F Tests to Wald Tests

- We saw we can calculate an F test by forming a matrix of restrictions.
- The F distribution result depended on the homoskedasticity and normality of the errors.
- In large samples, we can instead rely on the asymptotic behavior of random variables and calculate a **Wald statistic**.
- Many test statistics are reflavored Wald distances. Given a null that $E[\boldsymbol{q}] = \boldsymbol{\theta}$:

$$W = (\boldsymbol{q} - \boldsymbol{\theta})'[\widehat{Var}(\boldsymbol{q} - \boldsymbol{\theta})]^{-1}(\boldsymbol{q} - \boldsymbol{\theta})$$

- $W \sim \chi^2_J$ if $\boldsymbol{q}$ is Normal and $Var(\boldsymbol{q} - \boldsymbol{\theta}) = \boldsymbol{\Sigma}$
- $W \overset{a}{\sim} \chi^2_J$ with CLT on $\boldsymbol{q}$ and consistency for $\hat{\boldsymbol{\Sigma}}$

Mahalanobis Distance: Intuition

- Suppose $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. We want to measure how unusual an observation $\boldsymbol{x}$ is.
- In the scalar case:
$$D = \frac{x - \mu}{\sigma} \sim \mathcal{N}(0, 1) \quad \Rightarrow \quad D^2 \sim \chi^2_1$$
- In the vector case: transform the data to make it standard normal:
$$\boldsymbol{z} = \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{x} - \boldsymbol{\mu}) \sim \mathcal{N}(0, \boldsymbol{I})$$
- Then the squared norm of $\boldsymbol{z}$ gives:
$$\|\boldsymbol{z}\|^2 = (\boldsymbol{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) = D^2$$
- This is the Mahalanobis distance: it tells us how far $\boldsymbol{x}$ is from $\boldsymbol{\mu}$ in standardized units, accounting for correlations and scale.

Wald Statistic as a Distance

- The Wald test generalizes the standardized t-statistic to higher dimensions:

$$\text{Univariate: } \frac{\hat{\beta} - \beta_0}{\text{SE}(\hat{\beta})} \sim \mathcal{N}(0,1) \quad \Rightarrow \quad \left(\frac{\hat{\beta} - \beta_0}{\text{SE}(\hat{\beta})}\right)^2 \sim \chi^2_1$$

- In multivariate form:

$$W = (\boldsymbol{q} - \boldsymbol{\theta})'\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{q} - \boldsymbol{\theta})$$

- If $\boldsymbol{q}$ is consistent and asymptotically normal:

$$\sqrt{n}(\boldsymbol{q} - \boldsymbol{\theta}) \xrightarrow{d} \mathcal{N}(0, \boldsymbol{\Sigma}) \Rightarrow W \overset{a}{\sim} \chi^2_J$$

- Interpretation: Wald tests ask whether the observed $\boldsymbol{q}$ is “far” from $\boldsymbol{\theta}$, accounting for sampling variability.

Matrix form of Restrictions

- Suppose we have k parameters $\boldsymbol{\beta}$ and q restrictions with $q \leq (k)$. (Note $q = J$ in Greene, here k includes the intercept)
- Let $\boldsymbol{R}$ be a $q \times (k)$ matrix.
- We can express our null hypothesis as $H_0: \ \boldsymbol{R}\boldsymbol{\beta} = \boldsymbol{r}$.
- For example in the prior example where $\beta_2 = 0$ and $\beta_3 = \beta_4$

$$\boldsymbol{R} = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 \end{bmatrix}, \quad \boldsymbol{r} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Wald Test for Linear Restrictions

$$W = (\boldsymbol{R}\hat{\beta} - \boldsymbol{r})'[\boldsymbol{R}\hat{\boldsymbol{V}}_{\hat{\beta}}\boldsymbol{R}']^{-1}(\boldsymbol{R}\hat{\beta} - \boldsymbol{r})$$

$$W \overset{a}{\sim} \chi^2_q$$

- If we use $\hat{\boldsymbol{V}}_{\hat{\beta}} = s^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$ (homoskedasticity + normality), then W/q is exactly the F statistic distributed $F_{q,n-k}$.
- Without normality, the CLT tells us W/q is approximately $F_{q,n-k}$.
- Without normality or homoskedasticity, use $\hat{\boldsymbol{V}}^{HC}_{\hat{\beta}}$ (sandwich) and $W \overset{a}{\sim} \chi^2_q$.

The Wald test nests the F test as a special case. With robust SEs, it works without normality or homoskedasticity — only requiring consistency and asymptotic normality.

Wald Test for Linear Restrictions

$$W = (\boldsymbol{R}\hat{\beta} - \boldsymbol{r})'[\boldsymbol{R}\hat{\boldsymbol{V}}_{\hat{\beta}}\boldsymbol{R}']^{-1}(\boldsymbol{R}\hat{\beta} - \boldsymbol{r})$$

$$W \overset{a}{\sim} \chi^2_q$$

- If we use $\hat{\boldsymbol{V}}_{\hat{\beta}} = s^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$ (homoskedasticity + normality), then W/q is exactly the F statistic distributed $F_{q,n-k}$.
- Without normality, the CLT tells us W/q is approximately $F_{q,n-k}$.
- Without normality or homoskedasticity, use $\hat{\boldsymbol{V}}^{HC}_{\hat{\beta}}$ (sandwich) and $W \overset{a}{\sim} \chi^2_q$.

The Wald test nests the F test as a special case. With robust SEs, it works without normality or homoskedasticity — only requiring consistency and asymptotic normality.

Wald Tests in R

```
library(sandwich); library(lmtest); library(car)

mod0 <- lm(prestige ~ income + education + women,
           data = Prestige)

# Wald test with robust SEs (chi-squared)
linearHypothesis(mod0,
    c("education = 0", "women = 0"),
    vcov = vcovHC, test = "Chisq")

# Equivalent: waldtest with nested models
mod1 <- lm(prestige ~ income, data = Prestige)
waldtest(mod1, mod0, vcov = vcovHC, test = "Chisq")
```

Both test $H_0 : \beta_{ed} = \beta_{women} = 0$ using the sandwich covariance.

Wald Tests in R

```
library(sandwich); library(lmtest); library(car)

mod0 <- lm(prestige ~ income + education + women,
           data = Prestige)

# Wald test with robust SEs (chi-squared)
linearHypothesis(mod0,
    c("education = 0", "women = 0"),
    vcov = vcovHC, test = "Chisq")

# Equivalent: waldtest with nested models
mod1 <- lm(prestige ~ income, data = Prestige)
waldtest(mod1, mod0, vcov = vcovHC, test = "Chisq")
```

Both test $H_0 : \beta_{ed} = \beta_{women} = 0$ using the sandwich covariance.

Asymptotic Standard Errors via the Delta Method

- If $\hat{\boldsymbol{V}}_{\hat{\beta}}$ is an estimator of the covariance matrix of $\hat{\beta}$, then the standard errors are the square root of the diagonal elements.
- For a function $\theta = r(\beta)$, recall from Lecture 10 the **delta method**:

Given $\boldsymbol{R} = \frac{\partial r}{\partial \beta}\Big|_{\hat{\beta}}$, the asymptotic standard error is:

$$s(\hat{\theta}) = \sqrt{\boldsymbol{R}' \hat{\boldsymbol{V}}_{\hat{\beta}} \boldsymbol{R}}$$

This works for **any** differentiable function of $\hat{\beta}$ — linear or nonlinear.

Asymptotic Standard Errors via the Delta Method

- If $\hat{\boldsymbol{V}}_{\hat{\beta}}$ is an estimator of the covariance matrix of $\hat{\beta}$, then the standard errors are the square root of the diagonal elements.
- For a function $\theta = r(\beta)$, recall from Lecture 10 the **delta method**:

Given $\boldsymbol{R} = \left.\frac{\partial r}{\partial \beta}\right|_{\hat{\beta}}$, the asymptotic standard error is:

$$s(\hat{\theta}) = \sqrt{\boldsymbol{R}' \hat{\boldsymbol{V}}_{\hat{\beta}} \boldsymbol{R}}$$

This works for **any** differentiable function of $\hat{\beta}$ — linear or nonlinear.

Running Example: Log Wage Regression

- $\log(wage) = \beta_1 \cdot education + \beta_2 \cdot experience + \beta_3 \cdot experience^2/100 + \beta_4 + e$.
- We can calculate:
  - Percentage return to education:

$$\theta_1 = 100\beta_1$$

  - Percentage return to experience for those with 10 years of experience:

$$\theta_2 = 100\beta_2 + 20\beta_3$$

  - Experience level which maximizes expected log wages:

$$\theta_3 = -50\beta_2/\beta_3$$

Each is a function $\theta = r(\beta)$. The first two are linear; the third is nonlinear. All three use the same delta method formula.

Running Example: Log Wage Regression

- $\log(wage) = \beta_1 \cdot education + \beta_2 \cdot experience + \beta_3 \cdot experience^2/100 + \beta_4 + e.$
- We can calculate:
  - Percentage return to education:
  $$\theta_1 = 100\beta_1$$
  - Percentage return to experience for those with 10 years of experience:
  $$\theta_2 = 100\beta_2 + 20\beta_3$$
  - Experience level which maximizes expected log wages:
  $$\theta_3 = -50\beta_2/\beta_3$$

Each is a function $\theta = r(\beta)$. The first two are linear; the third is nonlinear. All three use the same delta method formula.

Example: Return to Education (Linear)

- Percentage return to education:

$$\theta_1 = 100\beta_1$$

- $\boldsymbol{R} = \begin{pmatrix} 100 \\ 0 \\ 0 \\ 0 \end{pmatrix}$
- Asymptotic standard error:

$$s(\hat{\theta}_1) = \sqrt{\boldsymbol{R}'\hat{\boldsymbol{V}}_{\hat{\beta}}\boldsymbol{R}} = 100 \cdot \sqrt{\hat{\text{Var}}(\hat{\beta}_1)}$$

This is just rescaling — the SE scales linearly.

Example: Return to Experience at 10 Years (Linear)

- Percentage return to experience at 10 years:

$$\theta_2 = 100\beta_2 + 20\beta_3$$

- Gradient of θ_2 with respect to $\boldsymbol{\beta}$:

$$\boldsymbol{R} = \begin{pmatrix} 0 \\ 100 \\ 20 \\ 0 \end{pmatrix}$$

- Asymptotic standard error:

$$s(\hat{\theta}_2) = \sqrt{\boldsymbol{R}' \hat{\boldsymbol{V}}_{\hat{\beta}} \boldsymbol{R}}$$

This is a linear combination of two coefficients — the SE depends on their covariance.

Example: Maximum Return to Experience (Nonlinear)

- Experience level which maximizes expected log wages:

$$\theta_3 = -50\frac{\beta_2}{\beta_3}$$

- Gradient of θ_3 with respect to $\boldsymbol{\beta}$:

$$\boldsymbol{R} = \begin{pmatrix} 0 \\ -50/\beta_3 \\ 50\beta_2/\beta_3^2 \\ 0 \end{pmatrix}$$

- Asymptotic standard error:

$$s(\hat{\theta}_3) = \sqrt{\boldsymbol{R}'\hat{\boldsymbol{V}}_{\hat{\beta}}\boldsymbol{R}}$$

This is nonlinear — $\boldsymbol{R}$ itself depends on β, estimated at $\hat{\beta}$. Same logic as the long-run elasticity from Lecture 10.

The z-statistic

- The t-statistic is $T(\theta) = \frac{\hat{\theta}-\theta}{s(\hat{\theta})}$
- When dealing with asymptotic inference, this is called the z-statistic.

$$
\begin{aligned}
T(\theta) &= \frac{\hat{\theta} - \theta}{s(\hat{\theta})} \\
&= \frac{\sqrt{n}(\hat{\theta} - \theta)}{\sqrt{\hat{V}_\theta}} \\
&\to_d \frac{N(0, V_\theta)}{\sqrt{V_\theta}} \\
&= Z \sim N(0, 1)
\end{aligned}
$$

Under $H_0 : \theta = \theta_0$, reject at level α if $|T| > z_{\alpha/2}$. For $\alpha = 0.05$: $z_{0.025} = 1.96$.

The z-statistic

- The t-statistic is $T(\theta) = \frac{\hat{\theta}-\theta}{s(\hat{\theta})}$
- When dealing with asymptotic inference, this is called the z-statistic.

$$\begin{aligned}
T(\theta) &= \frac{\hat{\theta}-\theta}{s(\hat{\theta})} \\
&= \frac{\sqrt{n}(\hat{\theta}-\theta)}{\sqrt{\hat{V}_\theta}} \\
&\to_d \frac{N(0, V_\theta)}{\sqrt{V_\theta}} \\
&= Z \sim N(0,1)
\end{aligned}$$

Under $H_0 : \theta = \theta_0$, reject at level α if $|T| > z_{\alpha/2}$. For $\alpha = 0.05$: $z_{0.025} = 1.96$.

Regression Intervals

- Linear model: $m(x) = x'\beta$
- Define target parameter: $\theta = r(\beta) = x'\beta$
- Estimate: $\hat{m}(x) = \hat{\theta} = x'\hat{\beta}$
- Gradient of $r(\beta)$: $\boldsymbol{R} = \frac{\partial r}{\partial \beta} = x$
- Asymptotic standard error:

$$s(\hat{\theta}) = \sqrt{\boldsymbol{R}'\hat{\boldsymbol{V}}_{\hat{\beta}}\boldsymbol{R}} = \sqrt{\boldsymbol{x}'\hat{\boldsymbol{V}}_{\hat{\beta}}\boldsymbol{x}}$$

- 95% confidence interval for $\mathbb{E}[Y|X = x]$:

$$x'\hat{\beta} \pm 1.96 \cdot \sqrt{\boldsymbol{x}'\hat{\boldsymbol{V}}_{\hat{\beta}}\boldsymbol{x}}$$

Example: Prestige $\sim$ Education

$\beta_{ed} = 5.4$, $\beta_{cons} = -10.7$, $\hat{\boldsymbol{V}}_{\hat{\beta}} = \begin{bmatrix} 13.52 & -1.18 \\ -1.18 & .11 \end{bmatrix}$

at 10 years of education:

$$x'\hat{\beta} \pm 1.96 \cdot \sqrt{\boldsymbol{x}'\hat{\boldsymbol{V}}_{\hat{\beta}}\boldsymbol{x}}$$

$$(1\ \ 10)\begin{pmatrix} -10.7 \\ 5.4 \end{pmatrix} \pm 1.96 \cdot \sqrt{(1\ \ 10)\begin{bmatrix} 13.52 & -1.18 \\ -1.18 & .11 \end{bmatrix}\begin{pmatrix} 1 \\ 10 \end{pmatrix}}$$

$$43.3 \pm 1.96 \cdot \sqrt{0.87}$$

$$43.3 \pm 1.83$$

compare to 22 years of education

$$107 \pm 1.96 \cdot \sqrt{14.8}$$

$$107 \pm 7.5$$

R: Sandwich SEs and Delta Method

```
# Homoskedastic vs robust SEs
coeftest(mod0)                                    # model-based
coeftest(mod0, vcov = vcovHC(mod0, type="HC1")) # sandwich

# Confidence interval for CEF at a point
x0 <- c(1, 5000, 12, 30)  # intercept, income, ed, women
se_x0 <- sqrt(t(x0) %*% vcovHC(mod0) %*% x0)
pred <- sum(x0 * coef(mod0))
cat(pred, "+/-", 1.96 * se_x0)

# Delta method for nonlinear function (car package)
car::deltaMethod(mod0, "income / education")
```

Same tools as Lectures 6, 9, 10 — the sandwich and car packages handle the asymptotic theory automatically.

R: Sandwich SEs and Delta Method

```
# Homoskedastic vs robust SEs
coeftest(mod0)                                        # model-based
coeftest(mod0, vcov = vcovHC(mod0, type="HC1")) # sandwich

# Confidence interval for CEF at a point
x0 <- c(1, 5000, 12, 30)  # intercept, income, ed, women
se_x0 <- sqrt(t(x0) %*% vcovHC(mod0) %*% x0)
pred <- sum(x0 * coef(mod0))
cat(pred, "+/-", 1.96 * se_x0)

# Delta method for nonlinear function (car package)
car::deltaMethod(mod0, "income / education")
```

Same tools as Lectures 6, 9, 10 — the `sandwich` and `car` packages handle the asymptotic theory automatically.

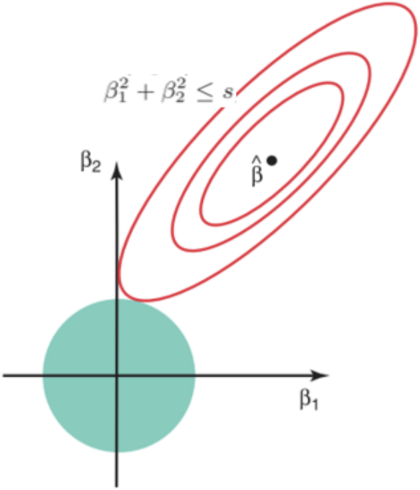
Application: Ridge Regression

- Suppose $Y = X'\beta + e$ with $\mathbb{E}[Xe] = 0$

$$\hat{\beta}_R = \left(\sum_{i=1}^{n} X_i X_i' + \lambda \boldsymbol{I}_k\right)^{-1} \left(\sum_{i=1}^{n} X_i Y_i\right)$$

- Where $\lambda > 0$ is a constant, added to the diagonals of the covariance matrix.
- This is biased, shrinks $\hat{\beta}$.
- Arises from solving $\min_\beta (\boldsymbol{y} - \boldsymbol{X}\beta)'(\boldsymbol{y} - \boldsymbol{X}\beta) + \lambda(\beta'\beta - c)$

Application: Ridge Regression



Ridge is Consistent (via CMT)

$$\text{plim } \hat{\beta}_R = \text{plim}\left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i' + \frac{1}{n}\lambda \boldsymbol{I}_k\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_i Y_i\right)$$

$$= \left(\text{plim}\frac{1}{n}\sum_{i=1}^{n} X_i X_i' + \text{plim}\frac{1}{n}\lambda \boldsymbol{I}_k\right)^{-1}\left(\text{plim}\frac{1}{n}\sum_{i=1}^{n} X_i Y_i\right)$$

$$= \left(\mathbb{E}[X_i X_i']\right)^{-1}\mathbb{E}[X_i Y_i] \quad \text{since } \frac{\lambda}{n} \to 0 \text{ as } n \to \infty$$

- As $n \to \infty$, the Ridge estimator converges in probability to the same plim as OLS.
- Biased in finite samples, consistent asymptotically — the canonical example of a biased-but-consistent estimator (recall Lecture 10's discussion).

Ridge is Consistent (via CMT)

$$\text{plim } \hat{\beta}_R = \text{plim} \left(\frac{1}{n} \sum_{i=1}^{n} X_i X_i' + \frac{1}{n} \lambda \boldsymbol{I}_k \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} X_i Y_i \right)$$

$$= \left(\text{plim} \frac{1}{n} \sum_{i=1}^{n} X_i X_i' + \text{plim} \frac{1}{n} \lambda \boldsymbol{I}_k \right)^{-1} \left(\text{plim} \frac{1}{n} \sum_{i=1}^{n} X_i Y_i \right)$$

$$= \left(\mathbb{E}[X_i X_i'] \right)^{-1} \mathbb{E}[X_i Y_i] \quad \text{since } \frac{\lambda}{n} \to 0 \text{ as } n \to \infty$$

- As $n \to \infty$, the Ridge estimator converges in probability to the same plim as OLS.
- Biased in finite samples, consistent asymptotically — the canonical example of a biased-but-consistent estimator (recall Lecture 10's discussion).

What We Proved Today

The OLS estimator $\hat{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'y$ is:

Property	**Tool (from Lecture 10)**	**Result**
Consistent	WLLN + CMT	$\text{plim } \hat{\beta} = \beta$
Asy. normal	Multivariate CLT + Slutsky	$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, V_\beta)$
Testable	Wald statistic	$W \overset{a}{\sim} \chi^2_q$
Robust	Sandwich estimator	$\hat{V}_\beta^{HC} \xrightarrow{p} V_\beta$

These results hold **without normality** — only requiring iid data with finite 4th moments and $\mathbb{E}[xe] = 0$.

What We Proved Today

The OLS estimator $\hat{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'y$ is:

Property	**Tool (from Lecture 10)**	**Result**
Consistent	WLLN + CMT	$\text{plim } \hat{\beta} = \beta$
Asy. normal	Multivariate CLT + Slutsky	$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, V_\beta)$
Testable	Wald statistic	$W \overset{a}{\sim} \chi^2_q$
Robust	Sandwich estimator	$\hat{V}_\beta^{HC} \xrightarrow{p} V_\beta$

These results hold **without normality** — only requiring iid data with finite 4th moments and $\mathbb{E}[\boldsymbol{x}e] = 0$.

Looking Ahead

What happens when $\mathbb{E}[xe] \neq 0$?

- OLS is **inconsistent**: plim $\hat{\beta} = \beta + Q_{XX}^{-1}\mathbb{E}[Xe] \neq \beta$
- No amount of data can fix this — the bias does not vanish.

Lectures 13–14 (IV and 2SLS): Find instruments Z with $\mathbb{E}[Ze] = 0$ and $\mathbb{E}[ZX'] \neq 0$.

- Replace the OLS moment condition $\mathbb{E}[X(Y - X'\beta)] = 0$ with $\mathbb{E}[Z(Y - X'\beta)] = 0$
- Same asymptotic machinery: WLLN, CLT, Slutsky, sandwich — applied to a different estimating equation
- This is another instance of the GMM framework: $\mathbb{E}[g(W_i, \theta_0)] = 0$

Looking Ahead

What happens when $\mathbb{E}[xe] \neq 0$?

- OLS is **inconsistent**: plim $\hat{\beta} = \beta + Q_{XX}^{-1}\mathbb{E}[Xe] \neq \beta$
- No amount of data can fix this — the bias does not vanish.

Lectures 13–14 (IV and 2SLS): Find instruments Z with $\mathbb{E}[Ze] = 0$ and $\mathbb{E}[ZX'] \neq 0$.

- Replace the OLS moment condition $\mathbb{E}[X(Y - X'\beta)] = 0$ with $\mathbb{E}[Z(Y - X'\beta)] = 0$
- Same asymptotic machinery: WLLN, CLT, Slutsky, sandwich — applied to a different estimating equation
- This is another instance of the GMM framework: $\mathbb{E}[g(W_i, \theta_0)] = 0$

Looking Ahead

What happens when $\mathbb{E}[xe] \neq 0$?

- OLS is **inconsistent**: plim $\hat{\beta} = \beta + Q_{XX}^{-1}\mathbb{E}[Xe] \neq \beta$
- No amount of data can fix this — the bias does not vanish.

Lectures 13–14 (IV and 2SLS): Find instruments Z with $\mathbb{E}[Ze] = 0$ and $\mathbb{E}[ZX'] \neq 0$.

- Replace the OLS moment condition $\mathbb{E}[X(Y - X'\beta)] = 0$ with $\mathbb{E}[Z(Y - X'\beta)] = 0$
- Same asymptotic machinery: WLLN, CLT, Slutsky, sandwich — applied to a different estimating equation
- This is another instance of the GMM framework: $\mathbb{E}[g(W_i, \theta_0)] = 0$