EKT-816 Lecture 5

OLS Consistency and Inference

Jesse Naidoo

University of Pretoria

Preliminaries

- Continuous Mapping Theorem: if $X_n \longrightarrow_p X_0$ and $g(\cdot)$ is continuous, then $g(X_n) \longrightarrow_p g(X_0)$.
- Slutsky's Theorem:
  - if $X_n \longrightarrow_p X_0$ (a constant) and $Y_n \longrightarrow_d Y$ (a nondegenerate distribution), then $X_n + Y_n \longrightarrow_d X_0 + Y$.
  - if $X_n \longrightarrow_p X_0$ (a constant) and $Y_n \longrightarrow_d Y$ (a nondegenerate distribution), then $X_n Y_n \longrightarrow_d X_0 Y$.
- Delta method: if $X_n \longrightarrow_d N(\mu, \Sigma)$, and $g(\cdot)$ is smoothly differentiable, then $g(X_n) \longrightarrow_d N(g(\mu), \nabla g(\mu)\Sigma\nabla g(\mu)')$.
  - here, $\nabla g(x)$ is the gradient of g (recall X can be a vector)
- you can find proofs of these statements in, e.g. Appendix A of Cameron and Trivedi (2005)
  - Ch.3 of Wooldridge (2010) or Ch. 6 of Stachurski (2016) also cover basic asymptotic theory

Preliminaries

- Continuous Mapping Theorem: if $X_n \longrightarrow_p X_0$ and $g(\cdot)$ is continuous, then $g(X_n) \longrightarrow_p g(X_0)$.
- Slutsky's Theorem:
    - if $X_n \longrightarrow_p X_0$ (a constant) and $Y_n \longrightarrow_d Y$ (a nondegenerate distribution), then $X_n + Y_n \longrightarrow_d X_0 + Y$.
    - if $X_n \longrightarrow_p X_0$ (a constant) and $Y_n \longrightarrow_d Y$ (a nondegenerate distribution), then $X_n Y_n \longrightarrow_d X_0 Y$.
- Delta method: if $X_n \longrightarrow_d N(\mu, \Sigma)$, and $g(\cdot)$ is smoothly differentiable, then $g(X_n) \longrightarrow_d N(g(\mu), \nabla g(\mu) \Sigma \nabla g(\mu)')$.
    - here, $\nabla g(x)$ is the gradient of g (recall X can be a vector)
- you can find proofs of these statements in, e.g. Appendix A of Cameron and Trivedi (2005)
    - Ch.3 of Wooldridge (2010) or Ch. 6 of Stachurski (2016) also cover basic asymptotic theory

Preliminaries

- Continuous Mapping Theorem: if $X_n \longrightarrow_p X_0$ and $g(\cdot)$ is continuous, then $g(X_n) \longrightarrow_p g(X_0)$.
- Slutsky's Theorem:
  - if $X_n \longrightarrow_p X_0$ (a constant) and $Y_n \longrightarrow_d Y$ (a nondegenerate distribution), then $X_n + Y_n \longrightarrow_d X_0 + Y$.
  - if $X_n \longrightarrow_p X_0$ (a constant) and $Y_n \longrightarrow_d Y$ (a nondegenerate distribution), then $X_n Y_n \longrightarrow_d X_0 Y$.
- Delta method: if $X_n \longrightarrow_d N(\mu, \Sigma)$, and $g(\cdot)$ is smoothly differentiable, then $g(X_n) \longrightarrow_d N(g(\mu), \nabla g(\mu) \Sigma \nabla g(\mu)')$.
  - here, $\nabla g(x)$ is the gradient of g (recall X can be a vector)
- you can find proofs of these statements in, e.g. Appendix A of Cameron and Trivedi (2005)
  - Ch.3 of Wooldridge (2010) or Ch. 6 of Stachurski (2016) also cover basic asymptotic theory

Preliminaries

- Continuous Mapping Theorem: if $X_n \longrightarrow_p X_0$ and $g(\cdot)$ is continuous, then $g(X_n) \longrightarrow_p g(X_0)$.
- Slutsky's Theorem:
  - if $X_n \longrightarrow_p X_0$ (a constant) and $Y_n \longrightarrow_d Y$ (a nondegenerate distribution), then $X_n + Y_n \longrightarrow_d X_0 + Y$.
  - if $X_n \longrightarrow_p X_0$ (a constant) and $Y_n \longrightarrow_d Y$ (a nondegenerate distribution), then $X_n Y_n \longrightarrow_d X_0 Y$.
- Delta method: if $X_n \longrightarrow_d N(\mu, \Sigma)$, and $g(\cdot)$ is smoothly differentiable, then $g(X_n) \longrightarrow_d N(g(\mu), \nabla g(\mu) \Sigma \nabla g(\mu)')$.
  - here, $\nabla g(x)$ is the gradient of g (recall X can be a vector)
- you can find proofs of these statements in, e.g. Appendix A of Cameron and Trivedi (2005)
  - Ch.3 of Wooldridge (2010) or Ch. 6 of Stachurski (2016) also cover basic asymptotic theory

Preliminaries

- Continuous Mapping Theorem: if $X_n \longrightarrow_p X_0$ and $g(\cdot)$ is continuous, then $g(X_n) \longrightarrow_p g(X_0)$.
- Slutsky's Theorem:
  - if $X_n \longrightarrow_p X_0$ (a constant) and $Y_n \longrightarrow_d Y$ (a nondegenerate distribution), then $X_n + Y_n \longrightarrow_d X_0 + Y$.
  - if $X_n \longrightarrow_p X_0$ (a constant) and $Y_n \longrightarrow_d Y$ (a nondegenerate distribution), then $X_n Y_n \longrightarrow_d X_0 Y$.
- Delta method: if $X_n \longrightarrow_d N(\mu, \Sigma)$, and $g(\cdot)$ is smoothly differentiable, then $g(X_n) \longrightarrow_d N(g(\mu), \nabla g(\mu) \Sigma \nabla g(\mu)')$.
  - here, $\nabla g(x)$ is the gradient of g (recall X can be a vector)
- you can find proofs of these statements in, e.g. Appendix A of Cameron and Trivedi (2005)
  - Ch.3 of Wooldridge (2010) or Ch. 6 of Stachurski (2016) also cover basic asymptotic theory

Preliminaries

- Continuous Mapping Theorem: if $X_n \longrightarrow_p X_0$ and $g(\cdot)$ is continuous, then $g(X_n) \longrightarrow_p g(X_0)$.
- Slutsky's Theorem:
  - if $X_n \longrightarrow_p X_0$ (a constant) and $Y_n \longrightarrow_d Y$ (a nondegenerate distribution), then $X_n + Y_n \longrightarrow_d X_0 + Y$.
  - if $X_n \longrightarrow_p X_0$ (a constant) and $Y_n \longrightarrow_d Y$ (a nondegenerate distribution), then $X_n Y_n \longrightarrow_d X_0 Y$.
- Delta method: if $X_n \longrightarrow_d N(\mu, \Sigma)$, and $g(\cdot)$ is smoothly differentiable, then $g(X_n) \longrightarrow_d N(g(\mu), \nabla g(\mu)\Sigma\nabla g(\mu)')$.
  - here, $\nabla g(x)$ is the gradient of g (recall X can be a vector)
- you can find proofs of these statements in, e.g. Appendix A of Cameron and Trivedi (2005)
  - Ch.3 of Wooldridge (2010) or Ch. 6 of Stachurski (2016) also cover basic asymptotic theory

Preliminaries

- Continuous Mapping Theorem: if $X_n \longrightarrow_p X_0$ and $g(\cdot)$ is continuous, then $g(X_n) \longrightarrow_p g(X_0)$.
- Slutsky's Theorem:
    - if $X_n \longrightarrow_p X_0$ (a constant) and $Y_n \longrightarrow_d Y$ (a nondegenerate distribution), then $X_n + Y_n \longrightarrow_d X_0 + Y$.
    - if $X_n \longrightarrow_p X_0$ (a constant) and $Y_n \longrightarrow_d Y$ (a nondegenerate distribution), then $X_n Y_n \longrightarrow_d X_0 Y$.
- Delta method: if $X_n \longrightarrow_d N(\mu, \Sigma)$, and $g(\cdot)$ is smoothly differentiable, then $g(X_n) \longrightarrow_d N(g(\mu), \nabla g(\mu) \Sigma \nabla g(\mu)')$.
    - here, $\nabla g(x)$ is the gradient of g (recall X can be a vector)
- you can find proofs of these statements in, e.g. Appendix A of Cameron and Trivedi (2005)
    - Ch.3 of Wooldridge (2010) or Ch. 6 of Stachurski (2016) also cover basic asymptotic theory

Preliminaries

- Continuous Mapping Theorem: if $X_n \longrightarrow_p X_0$ and $g(\cdot)$ is continuous, then $g(X_n) \longrightarrow_p g(X_0)$.
- Slutsky's Theorem:
  - if $X_n \longrightarrow_p X_0$ (a constant) and $Y_n \longrightarrow_d Y$ (a nondegenerate distribution), then $X_n + Y_n \longrightarrow_d X_0 + Y$.
  - if $X_n \longrightarrow_p X_0$ (a constant) and $Y_n \longrightarrow_d Y$ (a nondegenerate distribution), then $X_n Y_n \longrightarrow_d X_0 Y$.
- Delta method: if $X_n \longrightarrow_d N(\mu, \Sigma)$, and $g(\cdot)$ is smoothly differentiable, then $g(X_n) \longrightarrow_d N(g(\mu), \nabla g(\mu) \Sigma \nabla g(\mu)')$.
  - here, $\nabla g(x)$ is the gradient of g (recall X can be a vector)
- you can find proofs of these statements in, e.g. Appendix A of Cameron and Trivedi (2005)
  - Ch.3 of Wooldridge (2010) or Ch. 6 of Stachurski (2016) also cover basic asymptotic theory

Asymptotic Distribution of the OLS Estimator

- to build up intuition, think of the single-regressor case:

$$y = x\beta + \varepsilon$$

  with $\varepsilon \perp\!\!\!\perp x$ and $E[x] = E[\varepsilon] = 0$.

  - we know $\widehat{\beta} = \widehat{\text{cov}}(y, x)/\widehat{V}[x] = \sum_i y_i x_i / \sum_i x_i^2$.

- we also know $V[\widehat{\beta}] = \sigma_\varepsilon^2 / V[x]$
  - in the usual picture, this corresponds to the fact that estimating β is "harder" with:
    - more vertical dispersion in y (i.e. larger values of σ_ε^2)
    - less horizontal dispersion in x (i.e. smaller values of $V[x]$)

- now, we are going to extend this result to more complicated settings:
  - multiple regressors
  - unequal variances for ε at different values of x ("heteroskedasticity")
  - correlations between the errors of different observations (serial correlation or clustering)

Asymptotic Distribution of the OLS Estimator

- to build up intuition, think of the single-regressor case:

$$y = x\beta + \varepsilon$$

  with $\varepsilon \perp\!\!\!\perp x$ and $E[x] = E[\varepsilon] = 0$.
  - we know $\widehat{\beta} = \widehat{\text{cov}}(y, x)/\widehat{V}[x] = \sum_i y_i x_i / \sum_i x_i^2$.
- we also know $V[\widehat{\beta}] = \sigma_\varepsilon^2 / V[x]$
  - in the usual picture, this corresponds to the fact that estimating β is "harder" with:
    - more vertical dispersion in y (i.e. larger values of σ_ε^2)
    - less horizontal dispersion in x (i.e. smaller values of $V[x]$)
- now, we are going to extend this result to more complicated settings:
  - multiple regressors
  - unequal variances for ε at different values of x ("heteroskedasticity")
  - correlations between the errors of different observations (serial correlation or clustering)

Asymptotic Distribution of the OLS Estimator

- to build up intuition, think of the single-regressor case:

$$y = x\beta + \varepsilon$$

  with $\varepsilon \perp\!\!\!\perp x$ and $E[x] = E[\varepsilon] = 0$.
  - we know $\widehat{\beta} = \widehat{\text{cov}}(y, x)/\widehat{V}[x] = \sum_i y_i x_i / \sum_i x_i^2$.
- we also know $V[\widehat{\beta}] = \sigma_\varepsilon^2 / V[x]$
  - in the usual picture, this corresponds to the fact that estimating β is "harder" with:
    - more vertical dispersion in y (i.e. larger values of σ_ε^2)
    - less horizontal dispersion in x (i.e. smaller values of $V[x]$)
- now, we are going to extend this result to more complicated settings:
  - multiple regressors
  - unequal variances for ε at different values of x ("heteroskedasticity")
  - correlations between the errors of different observations (serial correlation or clustering)

Asymptotic Distribution of the OLS Estimator

- to build up intuition, think of the single-regressor case:

$$y = x\beta + \varepsilon$$

  with $\varepsilon \perp\!\!\!\perp x$ and $E[x] = E[\varepsilon] = 0$.
  - we know $\widehat{\beta} = \widehat{\text{cov}}(y, x)/\widehat{V}[x] = \sum_i y_i x_i / \sum_i x_i^2$.
- we also know $V[\widehat{\beta}] = \sigma_\varepsilon^2 / V[x]$
  - in the usual picture, this corresponds to the fact that estimating β is "harder" with:
    - more vertical dispersion in y (i.e. larger values of σ_ε^2)
    - less horizontal dispersion in x (i.e. smaller values of $V[x]$)
- now, we are going to extend this result to more complicated settings:
  - multiple regressors
  - unequal variances for ε at different values of x ("heteroskedasticity")
  - correlations between the errors of different observations (serial correlation or clustering)

Asymptotic Distribution of the OLS Estimator

- to build up intuition, think of the single-regressor case:

$$y = x\beta + \varepsilon$$

  with $\varepsilon \perp\!\!\!\perp x$ and $E[x] = E[\varepsilon] = 0$.
  - we know $\widehat{\beta} = \widehat{\text{cov}}(y, x)/\widehat{V}[x] = \sum_i y_i x_i / \sum_i x_i^2$.
- we also know $V[\widehat{\beta}] = \sigma_\varepsilon^2 / V[x]$
  - in the usual picture, this corresponds to the fact that estimating β is "harder" with:
    - more vertical dispersion in y (i.e. larger values of σ_ε^2)
    - less horizontal dispersion in x (i.e. smaller values of $V[x]$)
- now, we are going to extend this result to more complicated settings:
  - multiple regressors
  - unequal variances for ε at different values of x ("heteroskedasticity")
  - correlations between the errors of different observations (serial correlation or clustering)

Asymptotic Distribution of the OLS Estimator

- to build up intuition, think of the single-regressor case:

$$y = x\beta + \varepsilon$$

  with $\varepsilon \perp\!\!\!\perp x$ and $E[x] = E[\varepsilon] = 0$.
    - we know $\widehat{\beta} = \widehat{\text{cov}}(y, x)/\widehat{V}[x] = \sum_i y_i x_i / \sum_i x_i^2$.
- we also know $V[\widehat{\beta}] = \sigma_\varepsilon^2 / V[x]$
    - in the usual picture, this corresponds to the fact that estimating β is "harder" with:
        - more vertical dispersion in y (i.e. larger values of σ_ε^2)
        - less horizontal dispersion in x (i.e. smaller values of $V[x]$)
- now, we are going to extend this result to more complicated settings:
    - multiple regressors
    - unequal variances for ε at different values of x ("heteroskedasticity")
    - correlations between the errors of different observations (serial correlation or clustering)

Asymptotic Distribution of the OLS Estimator

- to build up intuition, think of the single-regressor case:

$$y = x\beta + \varepsilon$$

  with $\varepsilon \perp\!\!\!\perp x$ and $E[x] = E[\varepsilon] = 0$.
  - we know $\widehat{\beta} = \widehat{\text{cov}}(y, x)/\widehat{V}[x] = \sum_i y_i x_i / \sum_i x_i^2$.
- we also know $V[\widehat{\beta}] = \sigma_\varepsilon^2 / V[x]$
  - in the usual picture, this corresponds to the fact that estimating β is "harder" with:
    - more vertical dispersion in y (i.e. larger values of σ_ε^2)
    - less horizontal dispersion in x (i.e. smaller values of $V[x]$)
- now, we are going to extend this result to more complicated settings:
  - multiple regressors
  - unequal variances for ε at different values of x ("heteroskedasticity")
  - correlations between the errors of different observations (serial correlation or clustering)

Asymptotic Distribution of the OLS Estimator

- to build up intuition, think of the single-regressor case:

$$y = x\beta + \varepsilon$$

  with $\varepsilon \perp\!\!\!\perp x$ and $E[x] = E[\varepsilon] = 0$.
  - we know $\widehat{\beta} = \widehat{\text{cov}}(y, x)/\widehat{V}[x] = \sum_i y_i x_i / \sum_i x_i^2$.
- we also know $V[\widehat{\beta}] = \sigma_\varepsilon^2 / V[x]$
  - in the usual picture, this corresponds to the fact that estimating β is "harder" with:
    - more vertical dispersion in y (i.e. larger values of σ_ε^2)
    - less horizontal dispersion in x (i.e. smaller values of $V[x]$)
- now, we are going to extend this result to more complicated settings:
  - multiple regressors
  - unequal variances for ε at different values of x ("heteroskedasticity")
  - correlations between the errors of different observations (serial correlation or clustering)

Asymptotic Distribution of the OLS Estimator

- to build up intuition, think of the single-regressor case:

$$y = x\beta + \varepsilon$$

  with $\varepsilon \perp\!\!\!\perp x$ and $E[x] = E[\varepsilon] = 0$.
  - we know $\widehat{\beta} = \widehat{\text{cov}}(y, x)/\widehat{V}[x] = \sum_i y_i x_i / \sum_i x_i^2$.
- we also know $V[\widehat{\beta}] = \sigma_\varepsilon^2 / V[x]$
  - in the usual picture, this corresponds to the fact that estimating β is "harder" with:
    - more vertical dispersion in y (i.e. larger values of σ_ε^2)
    - less horizontal dispersion in x (i.e. smaller values of $V[x]$)
- now, we are going to extend this result to more complicated settings:
  - multiple regressors
  - unequal variances for ε at different values of x ("heteroskedasticity")
  - correlations between the errors of different observations (serial correlation or clustering)

Asymptotic Distribution of the OLS Estimator

- to build up intuition, think of the single-regressor case:

$$y = x\beta + \varepsilon$$

  with $\varepsilon \perp\!\!\!\perp x$ and $E[x] = E[\varepsilon] = 0$.
  - we know $\widehat{\beta} = \widehat{\text{cov}}(y, x)/\widehat{V}[x] = \sum_i y_i x_i / \sum_i x_i^2$.
- we also know $V[\widehat{\beta}] = \sigma_\varepsilon^2 / V[x]$
  - in the usual picture, this corresponds to the fact that estimating β is "harder" with:
    - more vertical dispersion in y (i.e. larger values of σ_ε^2)
    - less horizontal dispersion in x (i.e. smaller values of $V[x]$)
- now, we are going to extend this result to more complicated settings:
  - multiple regressors
  - unequal variances for ε at different values of x ("heteroskedasticity")
  - correlations between the errors of different observations (serial correlation or clustering)

Asymptotic Distribution of the OLS Estimator

- we want to apply a central limit theorem to $\widehat{\beta}$
- because $\widehat{\beta} = (X'X)^{-1}X'Y = \beta + (X'X)^{-1}X'\varepsilon$, we have

$$\sqrt{N}\left(\widehat{\beta} - \beta\right) \quad = \quad (X'X/N)^{-1}(X'\varepsilon/\sqrt{N}) \tag{1}$$

- we will maintain the assumption that $X'\varepsilon/N \longrightarrow_p 0$
  - an easy sufficient condition is that ε is *mean independent* of X, i.e. $E[\varepsilon|X] = 0$
  - we don't want to go as far as assuming ε is independent of X though
    - why not? Full independence implies no heteroskedasticity
  - if $X'\varepsilon/N \longrightarrow_p 0$, we get that OLS is *consistent* for β
- the simple case is one where $E[\varepsilon\varepsilon'|X] = \sigma^2 I$
  - take variances on both sides of (1) to get that

$$\sqrt{N}\left(\widehat{\beta} - \beta\right) \longrightarrow_d N(0, \sigma^2(X'X)^{-1})$$

  - notice, this is just a multivariable generalization of $V[\widehat{\beta}] = \sigma_\varepsilon^2/V[X]$
  - so, to do inference on the elements of $\widehat{\beta}$ (or functions of them) in practice, we'd use the *asymptotic covariance matrix* $s^2(X'X)^{-1}/N$

Asymptotic Distribution of the OLS Estimator

- we want to apply a central limit theorem to $\widehat{\beta}$
- because $\widehat{\beta} = (X'X)^{-1}X'Y = \beta + (X'X)^{-1}X'\varepsilon$, we have

$$\sqrt{N}\left(\widehat{\beta} - \beta\right) \quad = \quad (X'X/N)^{-1}(X'\varepsilon/\sqrt{N}) \tag{1}$$

- we will maintain the assumption that $X'\varepsilon/N \longrightarrow_p 0$
  - an easy sufficient condition is that ε is *mean independent* of X, i.e. $E[\varepsilon|X] = 0$
  - we don't want to go as far as assuming ε is independent of X though
    - *why not? Full independence implies no heteroskedasticity or clustering*
  - if $X'\varepsilon/N \longrightarrow_p 0$, we get that OLS is *consistent* for β
- the simple case is one where $E[\varepsilon\varepsilon'|X] = \sigma^2 I$
  - take variances on both sides of (1) to get that

$$\sqrt{N}\left(\widehat{\beta} - \beta\right) \longrightarrow_d N(0, \sigma^2(X'X)^{-1})$$

  - notice, this is just a multivariable generalization of $V[\widehat{\beta}] = \sigma_\varepsilon^2/V[X]$
  - so, to do inference on the elements of $\widehat{\beta}$ (or functions of them) in practice, we'd use the *asymptotic covariance matrix* $s^2(X'X)^{-1}/N$

Asymptotic Distribution of the OLS Estimator

- we want to apply a central limit theorem to $\widehat{\beta}$
- because $\widehat{\beta} = (X'X)^{-1}X'Y = \beta + (X'X)^{-1}X'\varepsilon$, we have

$$\sqrt{N}\left(\widehat{\beta} - \beta\right) = (X'X/N)^{-1}(X'\varepsilon/\sqrt{N}) \tag{1}$$

- we will maintain the assumption that $X'\varepsilon/N \longrightarrow_p 0$
  - an easy sufficient condition is that ε is *mean independent* of X, i.e. $E[\varepsilon|X] = 0$
  - we don't want to go as far as assuming ε is independent of X though
    - why not? Full independence implies no heteroskedasticity or clustering
  - if $X'\varepsilon/N \longrightarrow_p 0$, we get that OLS is *consistent* for β
- the simple case is one where $E[\varepsilon\varepsilon'|X] = \sigma^2 I$
  - take variances on both sides of (1) to get that

$$\sqrt{N}\left(\widehat{\beta} - \beta\right) \longrightarrow_d N(0, \sigma^2(X'X)^{-1})$$

  - notice, this is just a multivariable generalization of $V[\widehat{\beta}] = \sigma_\varepsilon^2/V[X]$
  - so, to do inference on the elements of $\widehat{\beta}$ (or functions of them) in practice, we'd use the *asymptotic covariance matrix* $s^2(X'X)^{-1}/N$

Asymptotic Distribution of the OLS Estimator

- we want to apply a central limit theorem to $\widehat{\beta}$
- because $\widehat{\beta} = (X'X)^{-1}X'Y = \beta + (X'X)^{-1}X'\varepsilon$, we have

$$\sqrt{N}\left(\widehat{\beta} - \beta\right) \;\;=\;\; (X'X/N)^{-1}(X'\varepsilon/\sqrt{N}) \tag{1}$$

- we will maintain the assumption that $X'\varepsilon/N \longrightarrow_p 0$
  - an easy sufficient condition is that ε is *mean independent* of X, i.e. $E[\varepsilon|X] = 0$
  - we don't want to go as far as assuming ε is independent of X though
    - why not? Full independence implies no heteroskedasticity or clustering
  - if $X'\varepsilon/N \longrightarrow_p 0$, we get that OLS is *consistent* for β
- the simple case is one where $E[\varepsilon\varepsilon'|X] = \sigma^2 I$
  - take variances on both sides of (1) to get that

$$\sqrt{N}\left(\widehat{\beta} - \beta\right) \longrightarrow_d N(0, \sigma^2(X'X)^{-1})$$

  - notice, this is just a multivariable generalization of $V[\widehat{\beta}] = \sigma_\varepsilon^2/V[X]$
  - so, to do inference on the elements of $\widehat{\beta}$ (or functions of them) in practice, we'd use the *asymptotic covariance matrix* $s^2(X'X)^{-1}/N$

Asymptotic Distribution of the OLS Estimator

- we want to apply a central limit theorem to $\widehat{\beta}$
- because $\widehat{\beta} = (X'X)^{-1}X'Y = \beta + (X'X)^{-1}X'\varepsilon$, we have

$$\sqrt{N}\left(\widehat{\beta} - \beta\right) \;=\; (X'X/N)^{-1}(X'\varepsilon/\sqrt{N}) \tag{1}$$

- we will maintain the assumption that $X'\varepsilon/N \longrightarrow_p 0$
  - an easy sufficient condition is that ε is *mean independent* of X, i.e. $E[\varepsilon|X] = 0$
  - we don't want to go as far as assuming ε is independent of X though
    - why not? Full independence implies no heteroskedasticity or clustering
  - if $X'\varepsilon/N \longrightarrow_p 0$, we get that OLS is *consistent* for β
- the simple case is one where $E[\varepsilon\varepsilon'|X] = \sigma^2 I$
  - take variances on both sides of (1) to get that

$$\sqrt{N}\left(\widehat{\beta} - \beta\right) \longrightarrow_d N(0, \sigma^2(X'X)^{-1})$$

  - notice, this is just a multivariable generalization of $V[\widehat{\beta}] = \sigma^2_\varepsilon/V[X]$
  - so, to do inference on the elements of $\widehat{\beta}$ (or functions of them) in practice, we'd use the *asymptotic covariance matrix* $s^2(X'X)^{-1}/N$

Asymptotic Distribution of the OLS Estimator

- we want to apply a central limit theorem to $\widehat{\beta}$
- because $\widehat{\beta} = (X'X)^{-1}X'Y = \beta + (X'X)^{-1}X'\varepsilon$, we have

$$\sqrt{N}\left(\widehat{\beta} - \beta\right) = (X'X/N)^{-1}(X'\varepsilon/\sqrt{N}) \tag{1}$$

- we will maintain the assumption that $X'\varepsilon/N \longrightarrow_p 0$
  - an easy sufficient condition is that ε is *mean independent* of X, i.e. $E[\varepsilon|X] = 0$
  - we don't want to go as far as assuming ε is independent of X though
    - why not? Full independence implies no heteroskedasticity or clustering
  - if $X'\varepsilon/N \longrightarrow_p 0$, we get that OLS is *consistent* for β
- the simple case is one where $E[\varepsilon\varepsilon'|X] = \sigma^2 I$
  - take variances on both sides of (1) to get that

$$\sqrt{N}\left(\widehat{\beta} - \beta\right) \longrightarrow_d N(0, \sigma^2(X'X)^{-1})$$

  - notice, this is just a multivariable generalization of $V[\widehat{\beta}] = \sigma_\varepsilon^2/V[X]$
  - so, to do inference on the elements of $\widehat{\beta}$ (or functions of them) in practice, we'd use the *asymptotic covariance matrix* $s^2(X'X)^{-1}/N$

Asymptotic Distribution of the OLS Estimator

- we want to apply a central limit theorem to $\widehat{\beta}$
- because $\widehat{\beta} = (X'X)^{-1}X'Y = \beta + (X'X)^{-1}X'\varepsilon$, we have

$$\sqrt{N}\left(\widehat{\beta} - \beta\right) \;=\; (X'X/N)^{-1}(X'\varepsilon/\sqrt{N}) \tag{1}$$

- we will maintain the assumption that $X'\varepsilon/N \longrightarrow_p 0$
  - an easy sufficient condition is that ε is *mean independent* of X, i.e. $E[\varepsilon|X] = 0$
  - we don't want to go as far as assuming ε is independent of X though
    - why not? Full independence implies no heteroskedasticity or clustering
  - if $X'\varepsilon/N \longrightarrow_p 0$, we get that OLS is *consistent* for β
- the simple case is one where $E[\varepsilon\varepsilon'|X] = \sigma^2 I$
  - take variances on both sides of (1) to get that

$$\sqrt{N}\left(\widehat{\beta} - \beta\right) \longrightarrow_d N(0, \sigma^2(X'X)^{-1})$$

  - notice, this is just a multivariable generalization of $V[\widehat{\beta}] = \sigma^2_\varepsilon / V[X]$
  - so, to do inference on the elements of $\widehat{\beta}$ (or functions of them) in practice, we'd use the *asymptotic covariance matrix* $s^2(X'X)^{-1}/N$

Asymptotic Distribution of the OLS Estimator

- we want to apply a central limit theorem to $\widehat{\beta}$
- because $\widehat{\beta} = (X'X)^{-1}X'Y = \beta + (X'X)^{-1}X'\varepsilon$, we have

$$\sqrt{N}\left(\widehat{\beta} - \beta\right) = (X'X/N)^{-1}(X'\varepsilon/\sqrt{N}) \tag{1}$$

- we will maintain the assumption that $X'\varepsilon/N \longrightarrow_p 0$
  - an easy sufficient condition is that ε is *mean independent* of X, i.e. $E[\varepsilon|X] = 0$
  - we don't want to go as far as assuming ε is independent of X though
    - why not? Full independence implies no heteroskedasticity or clustering
  - if $X'\varepsilon/N \longrightarrow_p 0$, we get that OLS is *consistent* for β
- the simple case is one where $E[\varepsilon\varepsilon'|X] = \sigma^2 I$
  - take variances on both sides of (1) to get that

$$\sqrt{N}\left(\widehat{\beta} - \beta\right) \longrightarrow_d N(0, \sigma^2(X'X)^{-1})$$

  - notice, this is just a multivariable generalization of $V[\widehat{\beta}] = \sigma_\varepsilon^2/V[X]$
  - so, to do inference on the elements of $\widehat{\beta}$ (or functions of them) in practice, we'd use the *asymptotic covariance matrix* $s^2(X'X)^{-1}/N$

Asymptotic Distribution of the OLS Estimator

- we want to apply a central limit theorem to $\widehat{\beta}$
- because $\widehat{\beta} = (X'X)^{-1}X'Y = \beta + (X'X)^{-1}X'\varepsilon$, we have

$$\sqrt{N}\left(\widehat{\beta} - \beta\right) \quad = \quad (X'X/N)^{-1}(X'\varepsilon/\sqrt{N}) \qquad (1)$$

- we will maintain the assumption that $X'\varepsilon/N \longrightarrow_p 0$
  - an easy sufficient condition is that ε is *mean independent* of X, i.e. $E[\varepsilon|X] = 0$
  - we don't want to go as far as assuming ε is independent of X though
    - why not? Full independence implies no heteroskedasticity or clustering
  - if $X'\varepsilon/N \longrightarrow_p 0$, we get that OLS is *consistent* for β
- the simple case is one where $E[\varepsilon\varepsilon'|X] = \sigma^2 I$
  - take variances on both sides of (1) to get that

$$\sqrt{N}\left(\widehat{\beta} - \beta\right) \longrightarrow_d N(0, \sigma^2(X'X)^{-1})$$

  - notice, this is just a multivariable generalization of $V[\widehat{\beta}] = \sigma^2_\varepsilon/V[X]$
  - so, to do inference on the elements of $\widehat{\beta}$ (or functions of them) in practice, we'd use the *asymptotic covariance matrix* $s^2(X'X)^{-1}/N$

Asymptotic Distribution of the OLS Estimator

- we want to apply a central limit theorem to $\widehat{\beta}$
- because $\widehat{\beta} = (X'X)^{-1}X'Y = \beta + (X'X)^{-1}X'\varepsilon$, we have

$$\sqrt{N}\left(\widehat{\beta} - \beta\right) = (X'X/N)^{-1}(X'\varepsilon/\sqrt{N}) \quad (1)$$

- we will maintain the assumption that $X'\varepsilon/N \longrightarrow_p 0$
  - an easy sufficient condition is that ε is *mean independent* of X, i.e. $E[\varepsilon|X] = 0$
  - we don't want to go as far as assuming ε is independent of X though
    - why not? Full independence implies no heteroskedasticity or clustering
  - if $X'\varepsilon/N \longrightarrow_p 0$, we get that OLS is *consistent* for β
- the simple case is one where $E[\varepsilon\varepsilon'|X] = \sigma^2 I$
  - take variances on both sides of (1) to get that

$$\sqrt{N}\left(\widehat{\beta} - \beta\right) \longrightarrow_d N(0, \sigma^2(X'X)^{-1})$$

  - notice, this is just a multivariable generalization of $V[\widehat{\beta}] = \sigma_\varepsilon^2/V[X]$
  - so, to do inference on the elements of $\widehat{\beta}$ (or functions of them) in practice, we'd use the *asymptotic covariance matrix* $s^2(X'X)^{-1}/N$

Asymptotic Distribution of the OLS Estimator

- we want to apply a central limit theorem to $\widehat{\beta}$
- because $\widehat{\beta} = (X'X)^{-1}X'Y = \beta + (X'X)^{-1}X'\varepsilon$, we have

$$\sqrt{N}\left(\widehat{\beta} - \beta\right) = (X'X/N)^{-1}(X'\varepsilon/\sqrt{N}) \quad (1)$$

- we will maintain the assumption that $X'\varepsilon/N \longrightarrow_p 0$
  - an easy sufficient condition is that ε is *mean independent* of X, i.e. $E[\varepsilon|X] = 0$
  - we don't want to go as far as assuming ε is independent of X though
    - why not? Full independence implies no heteroskedasticity or clustering
  - if $X'\varepsilon/N \longrightarrow_p 0$, we get that OLS is *consistent* for β
- the simple case is one where $E[\varepsilon\varepsilon'|X] = \sigma^2 I$
  - take variances on both sides of (1) to get that

$$\sqrt{N}\left(\widehat{\beta} - \beta\right) \longrightarrow_d N(0, \sigma^2(X'X)^{-1})$$

  - notice, this is just a multivariable generalization of $V[\widehat{\beta}] = \sigma^2_\varepsilon / V[X]$
  - so, to do inference on the elements of $\widehat{\beta}$ (or functions of them) in practice, we'd use the *asymptotic covariance matrix* $s^2(X'X)^{-1}/N$

Asymptotic Distribution of the OLS Estimator

- we always get that $(X'X/N)^{-1} \longrightarrow_p (E[X'X])^{-1} = M^{-1}$, by the LLN
- so, the key part of the previous argument was characterizing the limiting value of

$$\text{plim}\, X'\varepsilon\varepsilon'X/N = V \text{ , say}$$

- "best practice" in applied micro is *not* to try and explicitly model V
  - you might imagine writing down a parametric model $V(\gamma)$ and trying to estimate γ simultaneously with β
    - or, using the estimated OLS residuals $\hat{\varepsilon}$ as a first-stage input into the estimation of γ, then using $\hat{\gamma}$ to re-estimate β
    - if you specify the model for $V(\gamma)$ correctly, this can yield efficiency gains over OLS
    - BUT, a major disadvantage is that if you get the model for V wrong, you end up losing consistency of $\hat{\beta}$
    - this approach is usually called "generalized least squares" (GLS)
- instead, various data-driven approximations of V are used to get "robust" standard errors

Asymptotic Distribution of the OLS Estimator

- we always get that $(X'X/N)^{-1} \longrightarrow_p (E[X'X])^{-1} = M^{-1}$, by the LLN
- so, the key part of the previous argument was characterizing the limiting value of

$$\text{plim}\, X'\varepsilon\varepsilon'X/N = V \text{ , say}$$

- "best practice" in applied micro is *not* to try and explicitly model V
  - you might imagine writing down a parametric model $V(\gamma)$ and trying to estimate γ simultaneously with β
    - or, using the estimated OLS residuals $\hat{\varepsilon}$ as a first stage input into the estimation of γ, then using $\hat{\gamma}$ to re-estimate β
    - if you specify the model for $V(\gamma)$ correctly, this can yield efficiency gains over OLS
    - BUT, a major disadvantage is that if you get the model for V wrong, you end up losing consistency of $\hat{\beta}$
    - this approach is usually called "generalized least squares" (GLS)
- instead, various data-driven approximations of V are used to get "robust" standard errors

Asymptotic Distribution of the OLS Estimator

- we always get that $(X'X/N)^{-1} \longrightarrow_p (E[X'X])^{-1} = M^{-1}$, by the LLN
- so, the key part of the previous argument was characterizing the limiting value of

$$\text{plim}\, X'\varepsilon\varepsilon'X/N = V \text{ , say}$$

- "best practice" in applied micro is *not* to try and explicitly model V
  - you might imagine writing down a parametric model $V(\gamma)$ and trying to estimate γ simultaneously with β
    - or, using the estimated OLS residuals $\widehat{\varepsilon}$ as a first-stage input into the estimation of γ, then using $\widehat{\gamma}$ to re-estimate β
    - if you specify the model for $V(\gamma)$ correctly, this can yield efficiency gains over OLS
    - BUT, a major disadvantage is that if you get the model for V wrong, you end up losing consistency of $\widehat{\beta}$
    - this approach is usually called "generalized least squares" (GLS)
- instead, various data-driven approximations of V are used to get "robust" standard errors

Asymptotic Distribution of the OLS Estimator

- we always get that $(X'X/N)^{-1} \longrightarrow_p (E[X'X])^{-1} = M^{-1}$, by the LLN
- so, the key part of the previous argument was characterizing the limiting value of

$$\operatorname{plim} X'\varepsilon\varepsilon'X/N = V \text{ , say}$$

- "best practice" in applied micro is *not* to try and explicitly model V
  - you might imagine writing down a parametric model $V(\gamma)$ and trying to estimate γ simultaneously with β
    - or, using the estimated OLS residuals $\widehat{\varepsilon}$ as a first-stage input into the estimation of γ, then using $\widehat{\gamma}$ to re-estimate β
    - if you specify the model for $V(\gamma)$ correctly, this can yield efficiency gains over OLS
    - BUT, a major disadvantage is that if you get the model for V wrong, you end up losing consistency of $\widehat{\beta}$
    - this approach is usually called "generalized least squares" (GLS)
- instead, various data-driven approximations of V are used to get "robust" standard errors

Asymptotic Distribution of the OLS Estimator

- we always get that $(X'X/N)^{-1} \longrightarrow_p (E[X'X])^{-1} = M^{-1}$, by the LLN
- so, the key part of the previous argument was characterizing the limiting value of

$$\operatorname{plim} X'\varepsilon\varepsilon'X/N = V \text{ , say}$$

- "best practice" in applied micro is *not* to try and explicitly model V
  - you might imagine writing down a parametric model $V(\gamma)$ and trying to estimate γ simultaneously with β
    - or, using the estimated OLS residuals $\widehat{\varepsilon}$ as a first-stage input into the estimation of γ, then using $\widehat{\gamma}$ to re-estimate β
    - if you specify the model for $V(\gamma)$ correctly, this can yield efficiency gains over OLS
    - BUT, a major disadvantage is that if you get the model for V wrong, you end up losing consistency of $\widehat{\beta}$
    - this approach is usually called "generalized least squares" (GLS)
- instead, various data-driven approximations of V are used to get "robust" standard errors

Asymptotic Distribution of the OLS Estimator

- we always get that $(X'X/N)^{-1} \longrightarrow_p (E[X'X])^{-1} = M^{-1}$, by the LLN
- so, the key part of the previous argument was characterizing the limiting value of

$$\text{plim}\, X'\varepsilon\varepsilon' X/N = V \text{ , say}$$

- "best practice" in applied micro is *not* to try and explicitly model V
    - you might imagine writing down a parametric model $V(\gamma)$ and trying to estimate γ simultaneously with β
        - or, using the estimated OLS residuals $\widehat{\varepsilon}$ as a first-stage input into the estimation of γ, then using $\widehat{\gamma}$ to re-estimate β
        - if you specify the model for $V(\gamma)$ correctly, this can yield efficiency gains over OLS
        - BUT, a major disadvantage is that if you get the model for V wrong, you end up losing consistency of $\widehat{\beta}$
        - this approach is usually called "generalized least squares" (GLS)
- instead, various data-driven approximations of V are used to get "robust" standard errors

Asymptotic Distribution of the OLS Estimator

- we always get that $(X'X/N)^{-1} \longrightarrow_p (E[X'X])^{-1} = M^{-1}$, by the LLN
- so, the key part of the previous argument was characterizing the limiting value of

$$\text{plim}\, X'\varepsilon\varepsilon'X/N = V \text{ , say}$$

- "best practice" in applied micro is *not* to try and explicitly model V
  - you might imagine writing down a parametric model $V(\gamma)$ and trying to estimate γ simultaneously with β
    - or, using the estimated OLS residuals $\widehat{\varepsilon}$ as a first-stage input into the estimation of γ, then using $\widehat{\gamma}$ to re-estimate β
    - if you specify the model for $V(\gamma)$ correctly, this can yield efficiency gains over OLS
    - BUT, a major disadvantage is that if you get the model for V wrong, you end up losing consistency of $\widehat{\beta}$
    - this approach is usually called "generalized least squares" (GLS)
- instead, various data-driven approximations of V are used to get "robust" standard errors

Asymptotic Distribution of the OLS Estimator

- we always get that $(X'X/N)^{-1} \longrightarrow_p (E[X'X])^{-1} = M^{-1}$, by the LLN
- so, the key part of the previous argument was characterizing the limiting value of

$$\text{plim}\, X'\varepsilon\varepsilon' X/N = V \text{ , say}$$

- "best practice" in applied micro is *not* to try and explicitly model V
  - you might imagine writing down a parametric model $V(\gamma)$ and trying to estimate γ simultaneously with β
    - or, using the estimated OLS residuals $\widehat{\varepsilon}$ as a first-stage input into the estimation of γ, then using $\widehat{\gamma}$ to re-estimate β
    - if you specify the model for $V(\gamma)$ correctly, this can yield efficiency gains over OLS
    - BUT, a major disadvantage is that if you get the model for V wrong, you end up losing consistency of $\widehat{\beta}$
    - this approach is usually called "generalized least squares" (GLS)
- instead, various data-driven approximations of V are used to get "robust" standard errors

Asymptotic Distribution of the OLS Estimator

- we always get that $(X'X/N)^{-1} \longrightarrow_p (E[X'X])^{-1} = M^{-1}$, by the LLN
- so, the key part of the previous argument was characterizing the limiting value of

$$\text{plim}\, X'\varepsilon\varepsilon' X/N = V \text{ , say}$$

- "best practice" in applied micro is *not* to try and explicitly model V
  - you might imagine writing down a parametric model $V(\gamma)$ and trying to estimate γ simultaneously with β
    - or, using the estimated OLS residuals $\widehat{\varepsilon}$ as a first-stage input into the estimation of γ, then using $\widehat{\gamma}$ to re-estimate β
    - if you specify the model for $V(\gamma)$ correctly, this can yield efficiency gains over OLS
    - BUT, a major disadvantage is that if you get the model for V wrong, you end up losing consistency of $\widehat{\beta}$
    - this approach is usually called "generalized least squares" (GLS)
- instead, various data-driven approximations of V are used to get "robust" standard errors

Asymptotic Distribution of the OLS Estimator

- the goal is to obtain estimates of the precision of $\widehat{\beta}$ that are approximately correct under a wide range of assumptions about the exact form of $E[\varepsilon\varepsilon'|X]$
  - after all, we know OLS is consistent (if possibly inefficient)
  - GLS may not even be consistent if we misspecify the model for the error covariances!
- if you carry out the matrix multiplication you will see that

$$X'\varepsilon\varepsilon' X/N = N^{-1}\sum_{i=1}^{N}\sum_{j=1}^{N} X_j X_i' \varepsilon_i \varepsilon_j$$

- there are different choices of "robust" standard errors
  - Newey-West, Eicker-White, HC0, HC1, etc
  - all of these amount to different choices of weights ω_{ij} in a formula like

$$\widehat{V} = \sum_{i=1}^{N}\sum_{j=1}^{N} \omega_{ij} X_j X_i' \widehat{\varepsilon_i}\widehat{\varepsilon_j}$$

Asymptotic Distribution of the OLS Estimator

- the goal is to obtain estimates of the precision of $\widehat{\beta}$ that are approximately correct under a wide range of assumptions about the exact form of $E[\varepsilon\varepsilon'|X]$
  - after all, we know OLS is consistent (if possibly inefficient)
  - GLS may not even be consistent if we misspecify the model for the error covariances!
- if you carry out the matrix multiplication you will see that

$$X'\varepsilon\varepsilon'X/N = N^{-1}\sum_{i=1}^{N}\sum_{j=1}^{N} X_j X_i' \varepsilon_i \varepsilon_j$$

- there are different choices of "robust" standard errors
  - Newey-West, Eicker-White, HC0, HC1, etc
  - all of these amount to different choices of weights ω_{ij} in a formula like

$$\widehat{V} = \sum_{i=1}^{N}\sum_{j=1}^{N} \omega_{ij} X_j X_i' \widehat{\varepsilon_i}\widehat{\varepsilon_j}$$

Asymptotic Distribution of the OLS Estimator

- the goal is to obtain estimates of the precision of $\widehat{\beta}$ that are approximately correct under a wide range of assumptions about the exact form of $E[\varepsilon\varepsilon'|X]$
  - after all, we know OLS is consistent (if possibly inefficient)
  - GLS may not even be consistent if we misspecify the model for the error covariances!
- if you carry out the matrix multiplication you will see that

$$X'\varepsilon\varepsilon' X/N = N^{-1}\sum_{i=1}^{N}\sum_{j=1}^{N} X_j X_i' \varepsilon_i \varepsilon_j$$

- there are different choices of "robust" standard errors
  - Newey-West, Eicker-White, HC0, HC1, etc
  - all of these amount to different choices of weights ω_{ij} in a formula like

$$\widehat{V} = \sum_{i=1}^{N}\sum_{j=1}^{N} \omega_{ij} X_j X_i' \widehat{\varepsilon_i}\widehat{\varepsilon_j}$$

Asymptotic Distribution of the OLS Estimator

- the goal is to obtain estimates of the precision of $\widehat{\beta}$ that are approximately correct under a wide range of assumptions about the exact form of $E[\varepsilon\varepsilon'|X]$
    - after all, we know OLS is consistent (if possibly inefficient)
    - GLS may not even be consistent if we misspecify the model for the error covariances!
- if you carry out the matrix multiplication you will see that

$$X'\varepsilon\varepsilon' X/N = N^{-1}\sum_{i=1}^{N}\sum_{j=1}^{N} X_j X_i' \varepsilon_i \varepsilon_j$$

- there are different choices of "robust" standard errors
    - Newey-West, Eicker-White, HC0, HC1, etc
    - all of these amount to different choices of weights ω_{ij} in a formula like

$$\widehat{V} = \sum_{i=1}^{N}\sum_{j=1}^{N} \omega_{ij} X_j X_i' \widehat{\varepsilon_i}\widehat{\varepsilon_j}$$

Asymptotic Distribution of the OLS Estimator

- the goal is to obtain estimates of the precision of $\widehat{\beta}$ that are approximately correct under a wide range of assumptions about the exact form of $E[\varepsilon\varepsilon'|X]$
  - after all, we know OLS is consistent (if possibly inefficient)
  - GLS may not even be consistent if we misspecify the model for the error covariances!
- if you carry out the matrix multiplication you will see that

$$X'\varepsilon\varepsilon'X/N = N^{-1}\sum_{i=1}^{N}\sum_{j=1}^{N}X_jX_i'\varepsilon_i\varepsilon_j$$

- there are different choices of "robust" standard errors
  - Newey-West, Eicker-White, HC0, HC1, etc
  - all of these amount to different choices of weights ω_{ij} in a formula like

$$\widehat{V} = \sum_{i=1}^{N}\sum_{j=1}^{N}\omega_{ij}X_jX_i'\widehat{\varepsilon}_i\widehat{\varepsilon}_j$$

Asymptotic Distribution of the OLS Estimator

- the goal is to obtain estimates of the precision of $\widehat{\beta}$ that are approximately correct under a wide range of assumptions about the exact form of $E[\varepsilon\varepsilon'|X]$
  - after all, we know OLS is consistent (if possibly inefficient)
  - GLS may not even be consistent if we misspecify the model for the error covariances!
- if you carry out the matrix multiplication you will see that

$$X'\varepsilon\varepsilon' X/N = N^{-1}\sum_{i=1}^{N}\sum_{j=1}^{N} X_j X_i' \varepsilon_i \varepsilon_j$$

- there are different choices of "robust" standard errors
  - Newey-West, Eicker-White, HC0, HC1, etc
  - all of these amount to different choices of weights ω_{ij} in a formula like

$$\widehat{V} = \sum_{i=1}^{N}\sum_{j=1}^{N} \omega_{ij} X_j X_i' \widehat{\varepsilon}_i \widehat{\varepsilon}_j$$

Asymptotic Distribution of the OLS Estimator

- the goal is to obtain estimates of the precision of $\widehat{\beta}$ that are approximately correct under a wide range of assumptions about the exact form of $E[\varepsilon\varepsilon'|X]$
  - after all, we know OLS is consistent (if possibly inefficient)
  - GLS may not even be consistent if we misspecify the model for the error covariances!
- if you carry out the matrix multiplication you will see that

$$X'\varepsilon\varepsilon' X/N = N^{-1}\sum_{i=1}^{N}\sum_{j=1}^{N} X_j X_i' \varepsilon_i \varepsilon_j$$

- there are different choices of "robust" standard errors
  - Newey-West, Eicker-White, HC0, HC1, etc
  - all of these amount to different choices of weights ω_{ij} in a formula like

$$\widehat{V} = \sum_{i=1}^{N}\sum_{j=1}^{N} \omega_{ij} X_j X_i' \widehat{\varepsilon}_i \widehat{\varepsilon}_j$$

Asymptotic Distribution of the OLS Estimator

- for more details about this, see Ch. 4.4 of Cameron and Trivedi (2005) or Ch. 8 of Angrist and Pischke (2008)
    - Cameron and Miller (2015) is a useful reference about clustering, in particular
- next, we turn to a different question: why are we running OLS in the first place?

Asymptotic Distribution of the OLS Estimator

- for more details about this, see Ch. 4.4 of Cameron and Trivedi (2005) or Ch. 8 of Angrist and Pischke (2008)
  - Cameron and Miller (2015) is a useful reference about clustering, in particular
- next, we turn to a different question: why are we running OLS in the first place?

Asymptotic Distribution of the OLS Estimator

- for more details about this, see Ch. 4.4 of Cameron and Trivedi (2005) or Ch. 8 of Angrist and Pischke (2008)
    - Cameron and Miller (2015) is a useful reference about clustering, in particular
- next, we turn to a different question: why are we running OLS in the first place?

OLS as the Best Linear Approximation of $E[Y|X]$

- we can motivate OLS without literally believing $Y = X\beta + \varepsilon$ is the data-generating process
- instead, consider the following problems
  - $\beta^* = \arg\min_b E[(Y - Xb)^2]$, finding the best linear predictor of Y
  - $\beta^{**} = \arg\min_b E[(E[Y|X] - Xb)^2]$, finding the best linear approximation to $E[Y|X]$
- the OLS coefficient is β^* by definition, but these two problems have identical solutions
  - so, we can always think of the OLS coefficient as providing the best linear approximation to the conditional mean $E[Y|X]$, even if it is nonlinear
- of course, these facts tell us *nothing* about causality!
  - the *causal* question "what would happen to Y on average if we manipulated X by one unit" makes no sense without a model!
    - umbrella prevalence predicts rainfall
    - ambulances predict car crashes
  - on the other hand, if you start with a causal model (say from economic theory), knowing that OLS estimates approximate $E[Y|X]$ helps you think about whether you are going to get a good estimate of the causal effect you are trying to measure

OLS as the Best Linear Approximation of $E[Y|X]$

- we can motivate OLS without literally believing $Y = X\beta + \varepsilon$ is the data-generating process
- instead, consider the following problems
  - $\beta^* = \arg\min_b E[(Y - Xb)^2]$, finding the best linear predictor of Y
  - $\beta^{**} = \arg\min_b E[(E[Y|X] - Xb)^2]$, finding the best linear approximation to $E[Y|X]$
- the OLS coefficient is β^* by definition, but these two problems have identical solutions
  - so, we can always think of the OLS coefficient as providing the best linear approximation to the conditional mean $E[Y|X]$, even if it is nonlinear
- of course, these facts tell us *nothing* about causality!
  - the *causal* question "what would happen to Y on average if we manipulated X by one unit" makes no sense without a model!
    - umbrella prevalence predicts rainfall
    - ambulances predict car crashes
  - on the other hand, if you start with a causal model (say from economic theory), knowing that OLS estimates approximate $E[Y|X]$ helps you think about whether you are going to get a good estimate of the causal effect you are trying to measure

OLS as the Best Linear Approximation of $E[Y|X]$

- we can motivate OLS without literally believing $Y = X\beta + \varepsilon$ is the data-generating process
- instead, consider the following problems
  - $\beta^* = \arg\min_b E[(Y - Xb)^2]$, finding the best linear predictor of Y
  - $\beta^{**} = \arg\min_b E[(E[Y|X] - Xb)^2]$, finding the best linear approximation to $E[Y|X]$
- the OLS coefficient is β^* by definition, but these two problems have identical solutions
  - so, we can always think of the OLS coefficient as providing the best linear approximation to the conditional mean $E[Y|X]$, even if it is nonlinear
- of course, these facts tell us *nothing* about causality!
  - the *causal* question "what would happen to Y on average if we manipulated X by one unit" makes no sense without a model!
    - umbrella prevalence predicts rainfall
    - ambulances predict car crashes
  - on the other hand, if you start with a causal model (say from economic theory), knowing that OLS estimates approximate $E[Y|X]$ helps you think about whether you are going to get a good estimate of the causal effect you are trying to measure

OLS as the Best Linear Approximation of $E[Y|X]$

- we can motivate OLS without literally believing $Y = X\beta + \varepsilon$ is the data-generating process
- instead, consider the following problems
    - $\beta^* = \arg\min_b E[(Y - Xb)^2]$, finding the best linear predictor of Y
    - $\beta^{**} = \arg\min_b E[(E[Y|X] - Xb)^2]$, finding the best linear approximation to $E[Y|X]$
- the OLS coefficient is β^* by definition, but these two problems have identical solutions
    - so, we can always think of the OLS coefficient as providing the best linear approximation to the conditional mean $E[Y|X]$, even if it is nonlinear
- of course, these facts tell us *nothing* about causality!
    - the *causal* question "what would happen to Y on average if we manipulated X by one unit" makes no sense without a model!
        - umbrella prevalence predicts rainfall
        - ambulances predict car crashes
    - on the other hand, if you start with a causal model (say from economic theory), knowing that OLS estimates approximate $E[Y|X]$ helps you think about whether you are going to get a good estimate of the causal effect you are trying to measure

OLS as the Best Linear Approximation of $E[Y|X]$

- we can motivate OLS without literally believing $Y = X\beta + \varepsilon$ is the data-generating process
- instead, consider the following problems
  - $\beta^* = \arg\min_b E[(Y - Xb)^2]$, finding the best linear predictor of Y
  - $\beta^{**} = \arg\min_b E[(E[Y|X] - Xb)^2]$, finding the best linear approximation to $E[Y|X]$
- the OLS coefficient is β^* by definition, but these two problems have identical solutions
  - so, we can always think of the OLS coefficient as providing the best linear approximation to the conditional mean $E[Y|X]$, even if it is nonlinear
- of course, these facts tell us *nothing* about causality!
  - the *causal* question "what would happen to Y on average if we manipulated X by one unit" makes no sense without a model!
    - umbrella prevalence predicts rainfall
    - ambulances predict car crashes
  - on the other hand, if you start with a causal model (say from economic theory), knowing that OLS estimates approximate $E[Y|X]$ helps you think about whether you are going to get a good estimate of the causal effect you are trying to measure

OLS as the Best Linear Approximation of $E[Y|X]$

- we can motivate OLS without literally believing $Y = X\beta + \varepsilon$ is the data-generating process
- instead, consider the following problems
    - $\beta^* = \arg\min_b E[(Y - Xb)^2]$, finding the best linear predictor of Y
    - $\beta^{**} = \arg\min_b E[(E[Y|X] - Xb)^2]$, finding the best linear approximation to $E[Y|X]$
- the OLS coefficient is β^* by definition, but these two problems have identical solutions
    - so, we can always think of the OLS coefficient as providing the best linear approximation to the conditional mean $E[Y|X]$, even if it is nonlinear
- of course, these facts tell us *nothing* about causality!
    - the *causal* question "what would happen to Y on average if we manipulated X by one unit" makes no sense without a model!
        - umbrella prevalence predicts rainfall
        - ambulances predict car crashes
    - on the other hand, if you start with a causal model (say from economic theory), knowing that OLS estimates approximate $E[Y|X]$ helps you think about whether you are going to get a good estimate of the causal effect you are trying to measure

OLS as the Best Linear Approximation of $E[Y|X]$

- we can motivate OLS without literally believing $Y = X\beta + \varepsilon$ is the data-generating process
- instead, consider the following problems
  - $\beta^* = \arg\min_b E[(Y - Xb)^2]$, finding the best linear predictor of Y
  - $\beta^{**} = \arg\min_b E[(E[Y|X] - Xb)^2]$, finding the best linear approximation to $E[Y|X]$
- the OLS coefficient is β^* by definition, but these two problems have identical solutions
  - so, we can always think of the OLS coefficient as providing the best linear approximation to the conditional mean $E[Y|X]$, even if it is nonlinear
- of course, these facts tell us *nothing* about causality!
  - the *causal* question "what would happen to Y on average if we manipulated X by one unit" makes no sense without a model!
    - umbrella prevalence predicts rainfall
    - ambulances predict car crashes
  - on the other hand, if you start with a causal model (say from economic theory), knowing that OLS estimates approximate $E[Y|X]$ helps you think about whether you are going to get a good estimate of the causal effect you are trying to measure

OLS as the Best Linear Approximation of $E[Y|X]$

- we can motivate OLS without literally believing $Y = X\beta + \varepsilon$ is the data-generating process
- instead, consider the following problems
  - $\beta^* = \arg\min_b E[(Y - Xb)^2]$, finding the best linear predictor of Y
  - $\beta^{**} = \arg\min_b E[(E[Y|X] - Xb)^2]$, finding the best linear approximation to $E[Y|X]$
- the OLS coefficient is β^* by definition, but these two problems have identical solutions
  - so, we can always think of the OLS coefficient as providing the best linear approximation to the conditional mean $E[Y|X]$, even if it is nonlinear
- of course, these facts tell us *nothing* about causality!
  - the *causal* question "what would happen to Y on average if we manipulated X by one unit" makes no sense without a model!
    - umbrella prevalence predicts rainfall
    - ambulances predict car crashes
  - on the other hand, if you start with a causal model (say from economic theory), knowing that OLS estimates approximate $E[Y|X]$ helps you think about whether you are going to get a good estimate of the causal effect you are trying to measure

OLS as the Best Linear Approximation of $E[Y|X]$

- we can motivate OLS without literally believing $Y = X\beta + \varepsilon$ is the data-generating process
- instead, consider the following problems
  - $\beta^* = \arg\min_b E[(Y - Xb)^2]$, finding the best linear predictor of Y
  - $\beta^{**} = \arg\min_b E[(E[Y|X] - Xb)^2]$, finding the best linear approximation to $E[Y|X]$
- the OLS coefficient is β^* by definition, but these two problems have identical solutions
  - so, we can always think of the OLS coefficient as providing the best linear approximation to the conditional mean $E[Y|X]$, even if it is nonlinear
- of course, these facts tell us *nothing* about causality!
  - the *causal* question "what would happen to Y on average if we manipulated X by one unit" makes no sense without a model!
    - umbrella prevalence predicts rainfall
    - ambulances predict car crashes
  - on the other hand, if you start with a causal model (say from economic theory), knowing that OLS estimates approximate $E[Y|X]$ helps you think about whether you are going to get a good estimate of the causal effect you are trying to measure

OLS as the Best Linear Approximation of $E[Y|X]$

- we can motivate OLS without literally believing $Y = X\beta + \varepsilon$ is the data-generating process
- instead, consider the following problems
  - $\beta^* = \arg\min_b E[(Y - Xb)^2]$, finding the best linear predictor of Y
  - $\beta^{**} = \arg\min_b E[(E[Y|X] - Xb)^2]$, finding the best linear approximation to $E[Y|X]$
- the OLS coefficient is β^* by definition, but these two problems have identical solutions
  - so, we can always think of the OLS coefficient as providing the best linear approximation to the conditional mean $E[Y|X]$, even if it is nonlinear
- of course, these facts tell us *nothing* about causality!
  - the *causal* question "what would happen to Y on average if we manipulated X by one unit" makes no sense without a model!
    - umbrella prevalence predicts rainfall
    - ambulances predict car crashes
  - on the other hand, if you start with a causal model (say from economic theory), knowing that OLS estimates approximate $E[Y|X]$ helps you think about whether you are going to get a good estimate of the causal effect you are trying to measure

OLS as the Best Linear Approximation of $E[Y|X]$

- we can motivate OLS without literally believing $Y = X\beta + \varepsilon$ is the data-generating process
- instead, consider the following problems
  - $\beta^* = \arg\min_b E[(Y - Xb)^2]$, finding the best linear predictor of Y
  - $\beta^{**} = \arg\min_b E[(E[Y|X] - Xb)^2]$, finding the best linear approximation to $E[Y|X]$
- the OLS coefficient is β^* by definition, but these two problems have identical solutions
  - so, we can always think of the OLS coefficient as providing the best linear approximation to the conditional mean $E[Y|X]$, even if it is nonlinear
- of course, these facts tell us *nothing* about causality!
  - the *causal* question "what would happen to Y on average if we manipulated X by one unit" makes no sense without a model!
    - umbrella prevalence predicts rainfall
    - ambulances predict car crashes
  - on the other hand, if you start with a causal model (say from economic theory), knowing that OLS estimates approximate $E[Y|X]$ helps you think about whether you are going to get a good estimate of the causal effect you are trying to measure

More on Causality

- consider the following setup:
    - y_i is output per acre on farm i
    - x_{i1} is an index of soil quality
    - x_{i2} is an index of weather quality
    - x_{i3} is an index of pesticide use
    - e_i is a measure of insect population density
- We know that crop yields are determined as

$$y = x_1\beta_1 + x_2\beta_2 + x_3\beta_3 + e$$

- x_1, x_2, and e are mutually independent with mean zero

More on Causality

- consider the following setup:
  - y_i is output per acre on farm i
  - x_{i1} is an index of soil quality
  - x_{i2} is an index of weather quality
  - x_{i3} is an index of pesticide use
  - e_i is a measure of insect population density
- We know that crop yields are determined as

$$y = x_1\beta_1 + x_2\beta_2 + x_3\beta_3 + e$$

- x_1, x_2, and e are mutually independent with mean zero

More on Causality

- consider the following setup:
  - y_i is output per acre on farm i
  - x_{i1} is an index of soil quality
  - x_{i2} is an index of weather quality
  - x_{i3} is an index of pesticide use
  - e_i is a measure of insect population density
- We know that crop yields are determined as

$$y = x_1\beta_1 + x_2\beta_2 + x_3\beta_3 + e$$

- x_1, x_2, and e are mutually independent with mean zero

More on Causality

- consider the following setup:
  - y_i is output per acre on farm i
  - x_{i1} is an index of soil quality
  - x_{i2} is an index of weather quality
  - x_{i3} is an index of pesticide use
  - e_i is a measure of insect population density
- We know that crop yields are determined as

$$y = x_1\beta_1 + x_2\beta_2 + x_3\beta_3 + e$$

- x_1, x_2, and e are mutually independent with mean zero

More on Causality

- consider the following setup:
  - y_i is output per acre on farm i
  - x_{i1} is an index of soil quality
  - x_{i2} is an index of weather quality
  - x_{i3} is an index of pesticide use
  - e_i is a measure of insect population density
- We know that crop yields are determined as

$$y = x_1\beta_1 + x_2\beta_2 + x_3\beta_3 + e$$

- x_1, x_2, and e are mutually independent with mean zero

More on Causality

- consider the following setup:
  - y_i is output per acre on farm i
  - x_{i1} is an index of soil quality
  - x_{i2} is an index of weather quality
  - x_{i3} is an index of pesticide use
  - e_i is a measure of insect population density
- We know that crop yields are determined as

$$y = x_1\beta_1 + x_2\beta_2 + x_3\beta_3 + e$$

- x_1, x_2, and e are mutually independent with mean zero

More on Causality

- consider the following setup:
  - y_i is output per acre on farm i
  - x_{i1} is an index of soil quality
  - x_{i2} is an index of weather quality
  - x_{i3} is an index of pesticide use
  - e_i is a measure of insect population density
- We know that crop yields are determined as

$$y = x_1\beta_1 + x_2\beta_2 + x_3\beta_3 + e$$

- x_1, x_2, and e are mutually independent with mean zero

More on Causality

- consider the following setup:
  - y_i is output per acre on farm i
  - x_{i1} is an index of soil quality
  - x_{i2} is an index of weather quality
  - x_{i3} is an index of pesticide use
  - e_i is a measure of insect population density
- We know that crop yields are determined as

$$y = x_1\beta_1 + x_2\beta_2 + x_3\beta_3 + e$$

- x_1, x_2, and e are mutually independent with mean zero

More on Causality

- consider two different assumptions about how pesticide use is determined:
  - model A: farmers ignore soil quality (or do not observe it), but they do observe the level of insect populations
    - they set $x_3 = -e/\gamma + \eta$ where η is independent of all the x variables
  - model B: farmers set $x_3 = x_1 - e/\gamma_2 + \varepsilon$, where ε is independent of the x variables
- You have access to data on y, x_1, x_2, and x_3
- do you want to control for pesticide use if your goal is to estimate β_1?
- suppose model A generates the data and you use x_3 in your regression
  - $\widehat{\beta}_1$ will be consistent for β_1 (why?)
  - however, you cannot estimate β_3 consistently:

$$\text{plim}\,\widehat{\beta}_3 = \beta_3 - \frac{\sigma_e^2/\gamma}{\sigma_e^2/\gamma^2 + \sigma_\eta^2} < \beta_3$$

More on Causality

- consider two different assumptions about how pesticide use is determined:
  - model A: farmers ignore soil quality (or do not observe it), but they do observe the level of insect populations
    - they set $x_3 = -e/\gamma + \eta$ where η is independent of all the x variables
  - model B: farmers set $x_3 = x_1 - e/\gamma_2 + \varepsilon$, where ε is independent of the x variables
- You have access to data on y, x_1, x_2, and x_3
- do you want to control for pesticide use if your goal is to estimate β_1?
- suppose model A generates the data and you use x_3 in your regression
  - $\widehat{\beta}_1$ will be consistent for β_1 (why?)
  - however, you cannot estimate β_3 consistently:

$$\text{plim}\,\widehat{\beta}_3 = \beta_3 - \frac{\sigma_e^2/\gamma}{\sigma_e^2/\gamma^2 + \sigma_\eta^2} < \beta_3$$

More on Causality

- consider two different assumptions about how pesticide use is determined:
  - model A: farmers ignore soil quality (or do not observe it), but they do observe the level of insect populations
    - they set $x_3 = -e/\gamma + \eta$ where η is independent of all the x variables
  - model B: farmers set $x_3 = x_1 - e/\gamma_2 + \varepsilon$, where ε is independent of the x variables
- You have access to data on y, x_1, x_2, and x_3
- do you want to control for pesticide use if your goal is to estimate β_1?
- suppose model A generates the data and you use x_3 in your regression
  - $\widehat{\beta}_1$ will be consistent for β_1 (why?)
  - however, you cannot estimate β_3 consistently:

$$\text{plim}\,\widehat{\beta}_3 = \beta_3 - \frac{\sigma_e^2/\gamma}{\sigma_e^2/\gamma^2 + \sigma_\eta^2} < \beta_3$$

More on Causality

- consider two different assumptions about how pesticide use is determined:
  - model A: farmers ignore soil quality (or do not observe it), but they do observe the level of insect populations
    - they set $x_3 = -e/\gamma + \eta$ where η is independent of all the x variables
  - model B: farmers set $x_3 = x_1 - e/\gamma_2 + \varepsilon$, where ε is independent of the x variables
- You have access to data on y, x_1, x_2, and x_3
- do you want to control for pesticide use if your goal is to estimate β_1?
- suppose model A generates the data and you use x_3 in your regression
  - $\widehat{\beta}_1$ will be consistent for β_1 (why?)
  - however, you cannot estimate β_3 consistently:

$$\text{plim}\, \widehat{\beta}_3 = \beta_3 - \frac{\sigma_e^2/\gamma}{\sigma_e^2/\gamma^2 + \sigma_\eta^2} < \beta_3$$

More on Causality

- consider two different assumptions about how pesticide use is determined:
  - model A: farmers ignore soil quality (or do not observe it), but they do observe the level of insect populations
    - they set $x_3 = -e/\gamma + \eta$ where η is independent of all the x variables
  - model B: farmers set $x_3 = x_1 - e/\gamma_2 + \varepsilon$, where ε is independent of the x variables
- You have access to data on y, x_1, x_2, and x_3
- do you want to control for pesticide use if your goal is to estimate β_1?
- suppose model A generates the data and you use x_3 in your regression
  - $\widehat{\beta}_1$ will be consistent for β_1 (why?)
  - however, you cannot estimate β_3 consistently:

$$\text{plim}\, \widehat{\beta}_3 = \beta_3 - \frac{\sigma_e^2/\gamma}{\sigma_e^2/\gamma^2 + \sigma_\eta^2} < \beta_3$$

More on Causality

- consider two different assumptions about how pesticide use is determined:
  - model A: farmers ignore soil quality (or do not observe it), but they do observe the level of insect populations
    - they set $x_3 = -e/\gamma + \eta$ where η is independent of all the x variables
  - model B: farmers set $x_3 = x_1 - e/\gamma_2 + \varepsilon$, where ε is independent of the x variables
- You have access to data on y, x_1, x_2, and x_3
- do you want to control for pesticide use if your goal is to estimate β_1?
- suppose model A generates the data and you use x_3 in your regression
  - $\widehat{\beta}_1$ will be consistent for β_1 (why?)
  - however, you cannot estimate β_3 consistently:

$$\text{plim}\,\widehat{\beta}_3 = \beta_3 - \frac{\sigma_e^2/\gamma}{\sigma_e^2/\gamma^2 + \sigma_\eta^2} < \beta_3$$

More on Causality

- consider two different assumptions about how pesticide use is determined:
  - model A: farmers ignore soil quality (or do not observe it), but they do observe the level of insect populations
    - they set $x_3 = -e/\gamma + \eta$ where η is independent of all the x variables
  - model B: farmers set $x_3 = x_1 - e/\gamma_2 + \varepsilon$, where ε is independent of the x variables
- You have access to data on y, x_1, x_2, and x_3
- do you want to control for pesticide use if your goal is to estimate β_1?
- suppose model A generates the data and you use x_3 in your regression
  - $\widehat{\beta}_1$ will be consistent for β_1 (why?)
  - however, you cannot estimate β_3 consistently:

$$\text{plim}\,\widehat{\beta}_3 = \beta_3 - \frac{\sigma_e^2/\gamma}{\sigma_e^2/\gamma^2 + \sigma_\eta^2} < \beta_3$$

More on Causality

- consider two different assumptions about how pesticide use is determined:
  - model A: farmers ignore soil quality (or do not observe it), but they do observe the level of insect populations
    - they set $x_3 = -e/\gamma + \eta$ where η is independent of all the x variables
  - model B: farmers set $x_3 = x_1 - e/\gamma_2 + \varepsilon$, where ε is independent of the x variables
- You have access to data on y, x_1, x_2, and x_3
- do you want to control for pesticide use if your goal is to estimate β_1?
- suppose model A generates the data and you use x_3 in your regression
  - $\widehat{\beta}_1$ will be consistent for β_1 (why?)
  - however, you cannot estimate β_3 consistently:

$$\text{plim}\,\widehat{\beta}_3 = \beta_3 - \frac{\sigma_e^2/\gamma}{\sigma_e^2/\gamma^2 + \sigma_\eta^2} < \beta_3$$

More on Causality

- consider two different assumptions about how pesticide use is determined:
  - model A: farmers ignore soil quality (or do not observe it), but they do observe the level of insect populations
    - they set $x_3 = -e/\gamma + \eta$ where η is independent of all the x variables
  - model B: farmers set $x_3 = x_1 - e/\gamma_2 + \varepsilon$, where ε is independent of the x variables
- You have access to data on y, x_1, x_2, and x_3
- do you want to control for pesticide use if your goal is to estimate β_1?
- suppose model A generates the data and you use x_3 in your regression
  - $\widehat{\beta}_1$ will be consistent for β_1 (why?)
  - however, you cannot estimate β_3 consistently:

$$\text{plim}\,\widehat{\beta}_3 = \beta_3 - \frac{\sigma_e^2/\gamma}{\sigma_e^2/\gamma^2 + \sigma_\eta^2} < \beta_3$$

References

Angrist, Joshua D, and Jörn-Steffen Pischke. 2008. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Cameron, A Colin, and Douglas L Miller. 2015. "A Practitioner's Guide to Cluster-Robust Inference." *Journal of Human Resources* 50 (2): 317–72.

Cameron, A Colin, and Pravin K Trivedi. 2005. *Microeconometrics: Methods and Applications*. New York: Cambridge University Press.

Stachurski, John. 2016. *A Primer in Econometric Theory*. Cambridge, MA: MIT Press.

Wooldridge, Jeffrey M. 2010. *Econometric Analysis of Cross Section and Panel Data*. 2nd ed. MIT Press Books. The MIT Press.

Table of Contents