EKT-816 Lecture 5

OLS Consistency and Inference

Jesse Naidoo

University of Pretoria

- Continuous Mapping Theorem: if $X_n \longrightarrow_{\rho} X_0$ and $g(\cdot)$ is continuous, then $g(X_n) \longrightarrow_{\rho} g(X_0)$.
- Slutsky's Theorem:
 - if $X_n \longrightarrow_{\rho} X_0$ (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 $g(X_n) \longrightarrow_d N(g(\mu), \nabla g(\mu) \Sigma \nabla g(\mu)')$.
 - ullet here, abla 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

- 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

- 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 $g(X_n) \longrightarrow_d N(g(\mu), \nabla g(\mu) \Sigma \nabla g(\mu)')$.
 - ullet here, abla 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

- 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

- 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

- 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

- 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

- Continuous Mapping Theorem: if $X_n \longrightarrow_{\rho} X_0$ and $g(\cdot)$ is continuous, then $g(X_n) \longrightarrow_{\rho} g(X_0)$.
- 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

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

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

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

 If less horizontal dispersion in x (i.e. smaller values of V)s:
- now, we are going to extend this result to more complicated settings:
 - multiple regressors
 - ullet unequal variances for arepsilon at different values of x ("heteroskedasticity")
 - correlations between the errors of different observations (serial correlation or clustering)

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

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

- 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:
 - less horizontal dispersion in x (i.e. smaller values of V
- now, we are going to extend this result to more complicated settings:
 - multiple regressors
 - ullet unequal variances for arepsilon at different values of x ("heteroskedasticity")
 - correlations between the errors of different observations (serial correlation or clustering)

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

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

- 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 v (i.e. larger values of σ_e^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
 - ullet unequal variances for arepsilon at different values of imes ("heteroskedasticity")
 - correlations between the errors of different observations (serial correlation or clustering)

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

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

- 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 $V[\widehat{\beta}] = \sigma_{\varepsilon}^2/V[x]$
 - \bullet in the usual picture, this corresponds to the fact that estimating β is "harder" with:
 - \blacktriangleright 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)

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

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

- 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 $V[\widehat{\beta}] = \sigma_{\varepsilon}^2/V[x]$
 - \bullet 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
 - ullet unequal variances for arepsilon at different values of x ("heteroskedasticity"
 - correlations between the errors of different observations (serial correlation or clustering)

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

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

- 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 $V[\widehat{\beta}] = \sigma_{\varepsilon}^2/V[x]$
 - ullet in the usual picture, this corresponds to the fact that estimating eta 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
 - ullet unequal variances for arepsilon at different values of x ("heteroskedasticity"
 - correlations between the errors of different observations (serial correlation or clustering)

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

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

- 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 $V[\widehat{eta}] = \sigma_{arepsilon}^2/V[x]$
 - 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)

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

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

- 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)

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

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

- 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 $V[\widehat{eta}] = \sigma_{arepsilon}^2/V[x]$
 - 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)

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

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

- 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)

- \bullet 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. $E[\varepsilon|X]=0$
 - ullet we don't want to go as far as assuming arepsilon is independent of X though
 - why not? Full Independence implies no heteroskedasticity or representation.
 - 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

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

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

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

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

- ullet we will maintain the assumption that $X'arepsilon/N\longrightarrow_{p}0$
 - 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 heteroskedusticity or dustering
 - if $X'\varepsilon/N\longrightarrow_{\rho}0$, 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

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

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

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

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

- we will maintain the assumption that $X'\varepsilon/N \longrightarrow_{p} 0$
 - ullet an easy sufficient condition is that arepsilon is mean independent of X, i.e. E[arepsilon|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
 - ullet if $X'arepsilon/N\longrightarrow_{p} 0$, we get that OLS is *consistent* for eta
- 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 β (or functions of them) in practice, we'd use the asymptotic covariance matrix $s^2(X'X)^{-1}/N$

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

$$\sqrt{N}\left(\widehat{\beta} - \beta\right) = (X'X/N)^{-1}(X'\varepsilon/\sqrt{N})$$
 (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
 - if $X' \varepsilon / N \longrightarrow_{\alpha} 0$, 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

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

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

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

$$\sqrt{N}\left(\widehat{\beta} - \beta\right) = (X'X/N)^{-1}(X'\varepsilon/\sqrt{N})$$
 (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 $X'\varepsilon/N\longrightarrow_{\rho} 0$, we get that OLS is consistent for β
- the simple case is one where $E[\varepsilon \varepsilon' | X] = \sigma^2 I$

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

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

- 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 arepsilon 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 β (or functions of them) in practice we'd use the asymptotic covariance matrix $s^2(X'X)^{-1}/N$

- ullet we want to apply a central limit theorem to \widehat{eta}
- 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. $E[\varepsilon|X]=0$
 - \bullet 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 β (or functions of them) in practice, we'd use the asymptotic covariance matrix $s^2(X'X)^{-1}/N$

- 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

$$\sqrt{N}\left(\widehat{\beta} - \beta\right) = (X'X/N)^{-1}(X'\varepsilon/\sqrt{N})$$
 (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 arepsilon 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 β (or functions of them) in practice, we'd use the asymptotic covariance matrix $s^2(X'X)^{-1}/N$

- ullet we want to apply a central limit theorem to \widehat{eta}
- 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 arepsilon 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 β (or functions of them) in practice, we'd use the asymptotic covariance matrix $s^2(X'X)^{-1}/N$

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

$$\sqrt{N}\left(\widehat{\beta} - \beta\right) = (X'X/N)^{-1}(X'\varepsilon/\sqrt{N})$$
 (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. $E[\varepsilon|X]=0$
 - ullet 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 $V[\widehat{eta}] = \sigma_{arepsilon}^2/V[X]$
- so, to do inference on the elements of β (or functions of them) in practice, we'd use the asymptotic covariance matrix $s^2(X'X)^{-1}/N$

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

$$\sqrt{N}\left(\widehat{\beta} - \beta\right) = (X'X/N)^{-1}(X'\varepsilon/\sqrt{N})$$
 (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 arepsilon 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 $V[\widehat{eta}] = \sigma_{arepsilon}^2/V[X]$
- so, to do inference on the elements of $\widehat{\beta}$ (or functions of them) in practice, we'd use the asymptotic covariance matrix $s^2(X'X)^{-1}/N$

- 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

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

- ullet "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
 - of y, then using $\widehat{\varphi}$ to re-estimate β .

 If you specify the model for $V(\gamma)$ consists, this can yield efficiency gains one of CLS.

 If ELTE, a major disadvantage is that if you get the model for V enemy, you end the up today consistency of $\widehat{\rho}$.

 If is account is usually called "assertations" fact assured (CLS).
- ullet instead, various data-driven approximations of V are used to get "robust" standard errors

- 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

$$\operatorname{\mathsf{plim}} X' arepsilon arepsilon' X/N = V$$
 , say

- ullet "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
 - If you specify the model for V(y) correctly, this can yield efficiency galax over QLS
 BUT, a major disadvantage is that If you get the model for V wrong, you end up to losing consistency of B
- ullet instead, various data-driven approximations of V are used to get "robust" standard errors

- 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

$$\operatorname{\mathsf{plim}} X' arepsilon arepsilon' X/N = V$$
 , 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
 - ightharpoonup or, using the estimated OLS residuals $\hat{\epsilon}$ as a first-stage input into the estimation of γ , then using $\hat{\gamma}$ to re-estimate β
 - ightharpoonup 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 β
 - ightharpoonup this approach is usually called "generalized least squares" (GLS)
- ullet instead, various data-driven approximations of V are used to get "robust" standard errors

- 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

$$\operatorname{\mathsf{plim}} X' arepsilon arepsilon' X/N = V$$
 , 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
 - ${}^{\blacktriangleright}$ 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

- 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

$$\operatorname{\mathsf{plim}} X' arepsilon arepsilon' X/N = V$$
 , 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
 - ${}^{\blacktriangleright}$ 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

- 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

$$\operatorname{\mathsf{plim}} X' arepsilon arepsilon' X/N = V$$
 , 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
 - lacktriangle 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

- 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

$$\operatorname{\mathsf{plim}} X' arepsilon arepsilon' X/N = V$$
 , 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
 - lacktriangle 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

- 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

$$\operatorname{\mathsf{plim}} X' arepsilon arepsilon' X/N = V$$
 , 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
 - lacktriangle 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

- 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

$$\operatorname{\mathsf{plim}} X' arepsilon arepsilon' X/N = V$$
 , 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
 - lacktriangle 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

- 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
 - ullet all of these amount to different choices of weights ω_{ij} in a formula like

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

- 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
 - ullet all of these amount to different choices of weights ω_{ii} in a formula like

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

- 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
 - ullet all of these amount to different choices of weights ω_{ij} in a formula like

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

- 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

$$X' arepsilon arepsilon' X/N = N^{-1} \sum_{i=1}^N \sum_{j=1}^N X_j X_i' arepsilon_i arepsilon_j$$

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

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

- 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

$$X' arepsilon arepsilon' X/N = N^{-1} \sum_{i=1}^N \sum_{j=1}^N X_j X_i' arepsilon_i arepsilon_j$$

- 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

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

- 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

$$X' arepsilon arepsilon' X/N = N^{-1} \sum_{i=1}^N \sum_{j=1}^N X_j X_i' arepsilon_i arepsilon_j$$

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

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

- 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

$$X' arepsilon arepsilon' X/N = N^{-1} \sum_{i=1}^N \sum_{j=1}^N X_j X_i' arepsilon_i arepsilon_j$$

- 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

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

- 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?

- 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?

- 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?

- 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]
- ullet the OLS coefficient is eta^* 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!
 - 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

- 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
 - $\beta^{**}=\arg\min_b\bar{E}[(E[Y|X]-Xb)^2]$, finding the best linear approximation to E[Y|X]
- ullet the OLS coefficient is eta^* 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!
 - E ambulances resolut car cossius
 - 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

- 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]
- ullet the OLS coefficient is eta^* 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!
 - 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

- 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]
- ullet the OLS coefficient is eta^* 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!
 - 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

- 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]
- \bullet 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!
 - 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

- 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]
- \bullet 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 > by one unit" makes no sense without a model!
 - 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

- 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
 - $\beta^{**} = \arg\min_b \bar{E}[(E[Y|X] Xb)^2]$, finding the best linear approximation to E[Y|X]
- ullet the OLS coefficient is eta^* 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!
 - umbrena prevalence predicts rainfair
 - 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

- 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]
- \bullet 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

- 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]
- ullet the OLS coefficient is eta^* 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

- 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]
- ullet the OLS coefficient is eta^* 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

- 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
 - $\beta^{**} = \arg\min_b \bar{E}[(E[Y|X] Xb)^2]$, 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

• consider the following setup:

- y_i is output per acre on farm in
- 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

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

- 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

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

- 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

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

- 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

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

- 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
 - ullet e_i is a measure of insect population density
- We know that crop yields are determined as

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

- 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
 - $oldsymbol{e}_i$ is a measure of insect population density
- We know that crop yields are determined as

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

- 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
 - ei is a measure of insect population density
- We know that crop yields are determined as

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

- 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
 - ei is a measure of insect population density
- We know that crop yields are determined as

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

- 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
 - β_1 will be consistent for β_1 (why?)
 - however, you cannot estimate eta_3 consistently

plim
$$\widehat{eta}_3=eta_3-rac{\sigma_e^2/\gamma}{\sigma_e^2/\gamma^2+\sigma_\eta^2}$$

- 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
 - β_1 will be consistent for β_1 (why?)
 - however, you cannot estimate eta_3 consistently

$$egin{aligned} \widehat{eta}_3 = eta_3 - rac{\sigma_e^2/\gamma}{\sigma_e^2/\gamma^2 + \sigma_n^2} < eta_3 \end{aligned}$$

- 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
 - β_1 will be consistent for β_1 (why?)
 - however, you cannot estimate β_3 consistently

plim
$$\widehat{eta}_3=eta_3-rac{\sigma_e^2/\gamma}{\sigma_e^2/\gamma^2+\sigma_\eta^2}$$

- 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
 - β_1 will be consistent for β_1 (why?)
 - however, you cannot estimate β_3 consistently.



- 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
 - β_1 will be consistent for β_1 (why?)
 - ullet however, you cannot estimate eta_3 consistently



- 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
 - β_1 will be consistent for β_1 (why?)
 - however, you cannot estimate β_3 consistently



- 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 ?
- ullet suppose model A generates the data and you use x_3 in your regression
 - $\hat{\beta}_1$ will be consistent for β_1 (why?)
 - however, you cannot estimate β_3 consistently:

$$\mathsf{plim}\,\widehat{eta}_3 = eta_3 - rac{\sigma_e^2/\gamma}{\sigma_e^2/\gamma^2 + \sigma_\eta^2} < eta$$

- 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 ?
- ullet 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:

plim
$$\widehat{eta}_3=eta_3-rac{\sigma_e^2/\gamma}{\sigma_e^2/\gamma^2+\sigma_\eta^2}$$

- 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 ?
- ullet 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:

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

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

Preliminaries

Asymptotic Distribution of the OLS Estimator

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

More on Causality

References