EKT-816 Lecture 4

Mechanical Properties of OLS

Jesse Naidoo

University of Pretoria

Linear Projections

- let (X_1, X_2) have some joint distribution
- define

$$b = \frac{\text{cov}(X_1, X_2)}{V[X_2]}$$

$$a = E[X_1] - bE[X_2]$$

- and let $X_1^* = X_1 - (a + bX_2)$
- what is
  - $\text{cov}(X_1^*, X_2)$?
  - $E[X_1^*]$
- is X_1^* *independent* of X_2?
- we say $a + bX_2$ is the *linear projection* of X_1 onto X_2
  - X_1^* is the component of X_1 that is orthogonal to X_2

Linear Projections

- let (X_1, X_2) have some joint distribution
- define

$$b = \frac{\text{cov}(X_1, X_2)}{V[X_2]}$$

$$a = E[X_1] - bE[X_2]$$

- and let $X_1^* = X_1 - (a + bX_2)$
- what is
  - $\text{cov}(X_1^*, X_2)$?
  - $E[X_1^*]$
- is X_1^* *independent* of X_2?
- we say $a + bX_2$ is the *linear projection* of X_1 onto X_2
  - X_1^* is the component of X_1 that is orthogonal to X_2

Linear Projections

- let (X_1, X_2) have some joint distribution
- define

$$b = \frac{\text{cov}(X_1, X_2)}{V[X_2]}$$

$$a = E[X_1] - bE[X_2]$$

- and let $X_1^* = X_1 - (a + bX_2)$
- what is
  - $\text{cov}(X_1^*, X_2)$?
  - $E[X_1^*]$
- is X_1^* *independent* of X_2?
- we say $a + bX_2$ is the *linear projection* of X_1 onto X_2
  - X_1^* is the component of X_1 that is orthogonal to X_2

Linear Projections

- let (X_1, X_2) have some joint distribution
- define

$$b = \frac{\text{cov}(X_1, X_2)}{V[X_2]}$$

$$a = E[X_1] - bE[X_2]$$

- and let $X_1^* = X_1 - (a + bX_2)$
- what is
  - $\text{cov}(X_1^*, X_2)$?
  - $E[X_1^*]$
- is X_1^* *independent* of X_2?
- we say $a + bX_2$ is the *linear projection* of X_1 onto X_2
  - X_1^* is the component of X_1 that is orthogonal to X_2

Linear Projections

- let (X_1, X_2) have some joint distribution
- define

$$b = \frac{\text{cov}(X_1, X_2)}{V[X_2]}$$

$$a = E[X_1] - bE[X_2]$$

- and let $X_1^* = X_1 - (a + bX_2)$
- what is
  - $\text{cov}(X_1^*, X_2)$?
  - $E[X_1^*]$
- is X_1^* *independent* of X_2?
- we say $a + bX_2$ is the *linear projection* of X_1 onto X_2
  - X_1^* is the component of X_1 that is orthogonal to X_2

Linear Projections

- let (X_1, X_2) have some joint distribution
- define

$$b = \frac{\text{cov}(X_1, X_2)}{V[X_2]}$$

$$a = E[X_1] - bE[X_2]$$

- and let $X_1^* = X_1 - (a + bX_2)$
- what is
    - $\text{cov}(X_1^*, X_2)$?
    - $E[X_1^*]$
- is X_1^* *independent* of X_2?
- we say $a + bX_2$ is the *linear projection* of X_1 onto X_2
    - X_1^* is the component of X_1 that is orthogonal to X_2

Linear Projections

- let (X_1, X_2) have some joint distribution
- define

$$b = \frac{\text{cov}(X_1, X_2)}{V[X_2]}$$

$$a = E[X_1] - bE[X_2]$$

- and let $X_1^* = X_1 - (a + bX_2)$
- what is
    - $\text{cov}(X_1^*, X_2)$?
    - $E[X_1^*]$
- is X_1^* *independent* of X_2?
- we say $a + bX_2$ is the *linear projection* of X_1 onto X_2
    - X_1^* is the component of X_1 that is orthogonal to X_2

Linear Projections

- let (X_1, X_2) have some joint distribution
- define

$$b = \frac{\text{cov}(X_1, X_2)}{V[X_2]}$$

$$a = E[X_1] - bE[X_2]$$

- and let $X_1^* = X_1 - (a + bX_2)$
- what is
  - $\text{cov}(X_1^*, X_2)$?
  - $E[X_1^*]$
- is X_1^* *independent* of X_2?
- we say $a + bX_2$ is the *linear projection* of X_1 onto X_2
  - X_1^* is the component of X_1 that is orthogonal to X_2

Linear Projections

- let (X_1, X_2) have some joint distribution
- define

$$b = \frac{\text{cov}(X_1, X_2)}{V[X_2]}$$

$$a = E[X_1] - bE[X_2]$$

- and let $X_1^* = X_1 - (a + bX_2)$
- what is
  - $\text{cov}(X_1^*, X_2)$?
  - $E[X_1^*]$
- is X_1^* *independent* of X_2?
- we say $a + bX_2$ is the *linear projection* of X_1 onto X_2
  - X_1^* is the component of X_1 that is orthogonal to X_2

Linear Projections: Example

- suppose the joint distribution of (D, X) is as follows:
  - $P(D = 0, X = 0) = 0.1$
  - $P(D = 0, X = 1) = 0.2$
  - $P(D = 1, X = 0) = 0.5$
  - $P(D = 1, X = 1) = 0.2$
- you can confirm that:
  - $\text{cov}(D, X) = -0.08$ and $V[X] = 0.24$
  - $D^* = D + X/3 - 5/6$, $E[D^*] = 0$, and $\text{cov}(D^*, X) = 0$
  - nevertheless, X and D^* are *not* independent
    - compare $P(D^* = 5/6, X = 0) = 0.1$
    - yet, $P(D^* = 5/6) = 0.1$ and $P(X = 0) = 0.6$
- Ch. 2 - 3 of Stachurski (2016) contains useful linear algebra background for extending these ideas to a contexts where we have many variables
  - we will also need this background later in this lecture, when we discuss the FWL theorem

Linear Projections: Example

- suppose the joint distribution of (D, X) is as follows:
  - $P(D = 0, X = 0) = 0.1$
  - $P(D = 0, X = 1) = 0.2$
  - $P(D = 1, X = 0) = 0.5$
  - $P(D = 1, X = 1) = 0.2$
- you can confirm that:
  - $\text{cov}(D, X) = -0.08$ and $V[X] = 0.24$
  - $D^* = D + X/3 - 5/6$, $E[D^*] = 0$, and $\text{cov}(D^*, X) = 0$
  - nevertheless, X and D^* are *not* independent
    - compare $P(D^* = 5/6, X = 0) = 0.1$
    - yet, $P(D^* = 5/6) = 0.1$ and $P(X = 0) = 0.6$
- Ch. 2 - 3 of Stachurski (2016) contains useful linear algebra background for extending these ideas to a contexts where we have many variables
  - we will also need this background later in this lecture, when we discuss the FWL theorem

Linear Projections: Example

- suppose the joint distribution of (D, X) is as follows:
  - $P(D = 0, X = 0) = 0.1$
  - $P(D = 0, X = 1) = 0.2$
  - $P(D = 1, X = 0) = 0.5$
  - $P(D = 1, X = 1) = 0.2$
- you can confirm that:
  - $\text{cov}(D, X) = -0.08$ and $V[X] = 0.24$
  - $D^* = D + X/3 - 5/6$, $E[D^*] = 0$, and $\text{cov}(D^*, X) = 0$
  - nevertheless, X and D^* are *not* independent
    - compare $P(D^* = 5/6, X = 0) = 0.1$
    - yet, $P(D^* = 5/6) = 0.1$ and $P(X = 0) = 0.6$
- Ch. 2 - 3 of Stachurski (2016) contains useful linear algebra background for extending these ideas to a contexts where we have many variables
  - we will also need this background later in this lecture, when we discuss the FWL theorem

Linear Projections: Example

- suppose the joint distribution of (D, X) is as follows:
  - $P(D = 0, X = 0) = 0.1$
  - $P(D = 0, X = 1) = 0.2$
  - $P(D = 1, X = 0) = 0.5$
  - $P(D = 1, X = 1) = 0.2$
- you can confirm that:
  - $\text{cov}(D, X) = -0.08$ and $V[X] = 0.24$
  - $D^* = D + X/3 - 5/6$, $E[D^*] = 0$, and $\text{cov}(D^*, X) = 0$
  - nevertheless, X and D^* are *not* independent
    - compare $P(D^* = -5/6, X = 0) = 0.1$
    - yet, $P(D^* = -5/6) = 0.1$ and $P(X = 0) = 0.6$
- Ch. 2 - 3 of Stachurski (2016) contains useful linear algebra background for extending these ideas to a contexts where we have many variables
  - we will also need this background later in this lecture, when we discuss the FWL theorem

Linear Projections: Example

- suppose the joint distribution of (D, X) is as follows:
  - $P(D = 0, X = 0) = 0.1$
  - $P(D = 0, X = 1) = 0.2$
  - $P(D = 1, X = 0) = 0.5$
  - $P(D = 1, X = 1) = 0.2$
- you can confirm that:
  - $\text{cov}(D, X) = -0.08$ and $V[X] = 0.24$
  - $D^* = D + X/3 - 5/6$, $E[D^*] = 0$, and $\text{cov}(D^*, X) = 0$
  - nevertheless, X and D^* are *not* independent
    - compare $P(D^* = -5/6, X = 0) = 0.1$
    - with $P(D^* = -5/6) = 0.1$ and $P(X = 0) = 0.6$
- Ch. 2 - 3 of Stachurski (2016) contains useful linear algebra background for extending these ideas to a contexts where we have many variables
  - we will also need this background later in this lecture, when we discuss the FWL theorem

Linear Projections: Example

- suppose the joint distribution of (D, X) is as follows:
  - $P(D = 0, X = 0) = 0.1$
  - $P(D = 0, X = 1) = 0.2$
  - $P(D = 1, X = 0) = 0.5$
  - $P(D = 1, X = 1) = 0.2$
- you can confirm that:
  - $\text{cov}(D, X) = -0.08$ and $V[X] = 0.24$
  - $D^* = D + X/3 - 5/6$, $E[D^*] = 0$, and $\text{cov}(D^*, X) = 0$
  - nevertheless, X and D^* are *not* independent
    - compare $P(D^* = 5/6, X = 0) = 0.1$
    - yet, $P(D^* = 5/6) = 0.1$ and $P(X = 0) = 0.6$
- Ch. 2 - 3 of Stachurski (2016) contains useful linear algebra background for extending these ideas to a contexts where we have many variables
  - we will also need this background later in this lecture, when we discuss the FWL theorem

Linear Projections: Example

- suppose the joint distribution of (D, X) is as follows:
  - $P(D = 0, X = 0) = 0.1$
  - $P(D = 0, X = 1) = 0.2$
  - $P(D = 1, X = 0) = 0.5$
  - $P(D = 1, X = 1) = 0.2$
- you can confirm that:
  - $\text{cov}(D, X) = -0.08$ and $V[X] = 0.24$
  - $D^* = D + X/3 - 5/6$, $E[D^*] = 0$, and $\text{cov}(D^*, X) = 0$
  - nevertheless, X and D^* are *not* independent
    - compare $P(D^* = 5/6, X = 0) = 0.1$
    - yet, $P(D^* = 5/6) = 0.1$ and $P(X = 0) = 0.6$
- Ch. 2 - 3 of Stachurski (2016) contains useful linear algebra background for extending these ideas to a contexts where we have many variables
  - we will also need this background later in this lecture, when we discuss the FWL theorem

Linear Projections: Example

- suppose the joint distribution of (D, X) is as follows:
  - $P(D = 0, X = 0) = 0.1$
  - $P(D = 0, X = 1) = 0.2$
  - $P(D = 1, X = 0) = 0.5$
  - $P(D = 1, X = 1) = 0.2$
- you can confirm that:
  - $\text{cov}(D, X) = -0.08$ and $V[X] = 0.24$
  - $D^* = D + X/3 - 5/6$, $E[D^*] = 0$, and $\text{cov}(D^*, X) = 0$
  - nevertheless, X and D^* are *not* independent
    - compare $P(D^* = 5/6, X = 0) = 0.1$
    - yet, $P(D^* = 5/6) = 0.1$ and $P(X = 0) = 0.6$
- Ch. 2 - 3 of Stachurski (2016) contains useful linear algebra background for extending these ideas to a contexts where we have many variables
  - we will also need this background later in this lecture, when we discuss the FWL theorem

Linear Projections: Example

- suppose the joint distribution of (D, X) is as follows:
  - $P(D = 0, X = 0) = 0.1$
  - $P(D = 0, X = 1) = 0.2$
  - $P(D = 1, X = 0) = 0.5$
  - $P(D = 1, X = 1) = 0.2$
- you can confirm that:
  - $\text{cov}(D, X) = -0.08$ and $V[X] = 0.24$
  - $D^* = D + X/3 - 5/6$, $E[D^*] = 0$, and $\text{cov}(D^*, X) = 0$
  - nevertheless, X and D^* are *not* independent
    - compare $P(D^* = 5/6, X = 0) = 0.1$
    - yet, $P(D^* = 5/6) = 0.1$ and $P(X = 0) = 0.6$
- Ch. 2 - 3 of Stachurski (2016) contains useful linear algebra background for extending these ideas to a contexts where we have many variables
  - we will also need this background later in this lecture, when we discuss the FWL theorem

Linear Projections: Example

- suppose the joint distribution of (D, X) is as follows:
  - $P(D = 0, X = 0) = 0.1$
  - $P(D = 0, X = 1) = 0.2$
  - $P(D = 1, X = 0) = 0.5$
  - $P(D = 1, X = 1) = 0.2$
- you can confirm that:
  - $\text{cov}(D, X) = -0.08$ and $V[X] = 0.24$
  - $D^* = D + X/3 - 5/6$, $E[D^*] = 0$, and $\text{cov}(D^*, X) = 0$
  - nevertheless, X and D^* are *not* independent
    - compare $P(D^* = 5/6, X = 0) = 0.1$
    - yet, $P(D^* = 5/6) = 0.1$ and $P(X = 0) = 0.6$
- Ch. 2 - 3 of Stachurski (2016) contains useful linear algebra background for extending these ideas to a contexts where we have many variables
  - we will also need this background later in this lecture, when we discuss the FWL theorem

Linear Projections: Example

- suppose the joint distribution of (D, X) is as follows:
  - $P(D = 0, X = 0) = 0.1$
  - $P(D = 0, X = 1) = 0.2$
  - $P(D = 1, X = 0) = 0.5$
  - $P(D = 1, X = 1) = 0.2$
- you can confirm that:
  - $\text{cov}(D, X) = -0.08$ and $V[X] = 0.24$
  - $D^* = D + X/3 - 5/6$, $E[D^*] = 0$, and $\text{cov}(D^*, X) = 0$
  - nevertheless, X and D^* are *not* independent
    - compare $P(D^* = 5/6, X = 0) = 0.1$
    - yet, $P(D^* = 5/6) = 0.1$ and $P(X = 0) = 0.6$
- Ch. 2 - 3 of Stachurski (2016) contains useful linear algebra background for extending these ideas to a contexts where we have many variables
  - we will also need this background later in this lecture, when we discuss the FWL theorem

Linear Projections: Example

- suppose the joint distribution of (D, X) is as follows:
  - $P(D = 0, X = 0) = 0.1$
  - $P(D = 0, X = 1) = 0.2$
  - $P(D = 1, X = 0) = 0.5$
  - $P(D = 1, X = 1) = 0.2$
- you can confirm that:
  - $\text{cov}(D, X) = -0.08$ and $V[X] = 0.24$
  - $D^* = D + X/3 - 5/6$, $E[D^*] = 0$, and $\text{cov}(D^*, X) = 0$
  - nevertheless, X and D^* are *not* independent
    - compare $P(D^* = 5/6, X = 0) = 0.1$
    - yet, $P(D^* = 5/6) = 0.1$ and $P(X = 0) = 0.6$
- Ch. 2 - 3 of Stachurski (2016) contains useful linear algebra background for extending these ideas to a contexts where we have many variables
  - we will also need this background later in this lecture, when we discuss the FWL theorem

Linear Projections: Example

- suppose the joint distribution of (D, X) is as follows:
  - $P(D = 0, X = 0) = 0.1$
  - $P(D = 0, X = 1) = 0.2$
  - $P(D = 1, X = 0) = 0.5$
  - $P(D = 1, X = 1) = 0.2$
- you can confirm that:
  - $\text{cov}(D, X) = -0.08$ and $V[X] = 0.24$
  - $D^* = D + X/3 - 5/6$, $E[D^*] = 0$, and $\text{cov}(D^*, X) = 0$
  - nevertheless, X and D^* are *not* independent
    - compare $P(D^* = 5/6, X = 0) = 0.1$
    - yet, $P(D^* = 5/6) = 0.1$ and $P(X = 0) = 0.6$
- Ch. 2 - 3 of Stachurski (2016) contains useful linear algebra background for extending these ideas to a contexts where we have many variables
  - we will also need this background later in this lecture, when we discuss the FWL theorem

Derivation of OLS Formula

- Before we talk about the statistical properties of regression estimates, we need to understand exactly what OLS does mechanically
- given N data points $(Y_i, X_{i1}, \ldots X_{iK})$, consider

$$\min_{\beta_0,\ldots\beta_K} \sum_{i=1}^{N} (Y_i - \beta_0 - \beta_1 X_{i1} \ldots - \beta_K X_{iK})^2$$

  - what are the first-order conditions for this problem?
- we can write the data in matrix form
  - X is the $N \times (K+1)$ matrix of regressors
  - Y is a $N \times 1$ vector of "outcomes"

Derivation of OLS Formula

- Before we talk about the statistical properties of regression estimates, we need to understand exactly what OLS does mechanically
- given N data points $(Y_i, X_{i1}, \ldots X_{iK})$, consider

$$\min_{\beta_0,\ldots\beta_K} \sum_{i=1}^{N} (Y_i - \beta_0 - \beta_1 X_{i1} \ldots - \beta_K X_{iK})^2$$

  - what are the first-order conditions for this problem?
- we can write the data in matrix form
  - X is the $N \times (K+1)$ matrix of regressors
  - Y is a $N \times 1$ vector of "outcomes"

Derivation of OLS Formula

- Before we talk about the statistical properties of regression estimates, we need to understand exactly what OLS does mechanically
- given N data points $(Y_i, X_{i1}, \ldots X_{iK})$, consider

$$\min_{\beta_0,\ldots\beta_K} \sum_{i=1}^{N} (Y_i - \beta_0 - \beta_1 X_{i1} \ldots - \beta_K X_{iK})^2$$

  - what are the first-order conditions for this problem?
- we can write the data in matrix form
  - X is the $N \times (K+1)$ matrix of regressors
  - Y is a $N \times 1$ vector of "outcomes"

Derivation of OLS Formula

- Before we talk about the statistical properties of regression estimates, we need to understand exactly what OLS does mechanically
- given N data points $(Y_i, X_{i1}, \ldots X_{iK})$, consider

$$\min_{\beta_0,\ldots\beta_K} \sum_{i=1}^{N} (Y_i - \beta_0 - \beta_1 X_{i1} \ldots - \beta_K X_{iK})^2$$

  - what are the first-order conditions for this problem?
- we can write the data in matrix form
  - X is the $N \times (K+1)$ matrix of regressors
  - Y is a $N \times 1$ vector of "outcomes"

Derivation of OLS Formula

- Before we talk about the statistical properties of regression estimates, we need to understand exactly what OLS does mechanically
- given N data points $(Y_i, X_{i1}, \ldots X_{iK})$, consider

$$\min_{\beta_0,\ldots\beta_K} \sum_{i=1}^{N} (Y_i - \beta_0 - \beta_1 X_{i1} \ldots - \beta_K X_{iK})^2$$

  - what are the first-order conditions for this problem?
- we can write the data in matrix form
  - X is the $N \times (K+1)$ matrix of regressors
  - Y is a $N \times 1$ vector of "outcomes"

Derivation of OLS Formula

- Before we talk about the statistical properties of regression estimates, we need to understand exactly what OLS does mechanically
- given N data points $(Y_i, X_{i1}, \ldots X_{iK})$, consider

$$\min_{\beta_0,\ldots\beta_K} \sum_{i=1}^{N} (Y_i - \beta_0 - \beta_1 X_{i1} \ldots - \beta_K X_{iK})^2$$

  - what are the first-order conditions for this problem?
- we can write the data in matrix form
  - X is the $N \times (K+1)$ matrix of regressors
  - Y is a $N \times 1$ vector of “outcomes”

Derivation of OLS Formula

- when we do this, the FOC become

$$(X'X)\hat{\beta} = X'Y \Longrightarrow \hat{\beta} = (X'X)^{-1}X'Y$$

- now let
  - $\hat{e} = Y - X\hat{\beta}$ be the residuals
  - $M = I - X(X'X)^{-1}X'$ be the "residual maker" matrix ($N \times N$)
    - notice that M is symmetric ($M' = M$) and idempotent ($M \times M = M$)
    - also notice that $MY = \hat{e}$ and $MX = 0$
    - and, $X'\hat{e} = 0$, by construction
- we can write $Y = X\hat{\beta} + \hat{e}$
  - i.e. decomposition into predicted values + residuals
  - these follow from facts about linear algebra, *not* anything about causality!
  - in fact, you could do this with purely deterministic data: no statistics necessary

Derivation of OLS Formula

- when we do this, the FOC become

$$(X'X)\hat{\beta} = X'Y \Longrightarrow \hat{\beta} = (X'X)^{-1}X'Y$$

- now let
  - $\hat{e} = Y - X\hat{\beta}$ be the residuals
  - $M = I - X(X'X)^{-1}X'$ be the "residual maker" matrix ($N \times N$)
    - notice that M is symmetric ($M' = M$) and idempotent ($M \times M = M$)
    - also notice that $MY = \hat{e}$ and $MX = 0$
    - and, $X'\hat{e} = 0$, by construction
- we can write $Y = X\hat{\beta} + \hat{e}$
  - i.e. decomposition into predicted values + residuals
  - these follow from facts about linear algebra, *not* anything about causality!
  - in fact, you could do this with purely deterministic data: no statistics necessary

Derivation of OLS Formula

- when we do this, the FOC become

$$(X'X)\hat{\beta} = X'Y \implies \hat{\beta} = (X'X)^{-1}X'Y$$

- now let
  - $\hat{e} = Y - X\hat{\beta}$ be the residuals
  - $M = I - X(X'X)^{-1}X'$ be the "residual maker" matrix ($N \times N$)
    - notice that M is symmetric ($M' = M$) and idempotent ($M \times M = M$)
    - also notice that $MY = \hat{e}$ and $MX = 0$
    - and, $X'\hat{e} = 0$, by construction
- we can write $Y = X\hat{\beta} + \hat{e}$
  - i.e. decomposition into predicted values + residuals
  - these follow from facts about linear algebra, *not* anything about causality!
  - in fact, you could do this with purely deterministic data: no statistics necessary

Derivation of OLS Formula

- when we do this, the FOC become

$$(X'X)\hat{\beta} = X'Y \implies \hat{\beta} = (X'X)^{-1}X'Y$$

- now let
  - $\hat{e} = Y - X\hat{\beta}$ be the residuals
  - $M = I - X(X'X)^{-1}X'$ be the "residual maker" matrix ($N \times N$)
    - notice that M is symmetric ($M' = M$) and idempotent ($M \times M = M$)
    - also notice that $MY = \hat{e}$ and $MX = 0$
    - and, $X'\hat{e} = 0$, by construction
- we can write $Y = X\hat{\beta} + \hat{e}$
  - i.e. decomposition into predicted values + residuals
  - these follow from facts about linear algebra, *not* anything about causality!
  - in fact, you could do this with purely deterministic data: no statistics necessary

Derivation of OLS Formula

- when we do this, the FOC become

$$(X'X)\hat{\beta} = X'Y \implies \hat{\beta} = (X'X)^{-1}X'Y$$

- now let
  - $\hat{e} = Y - X\hat{\beta}$ be the residuals
  - $M = I - X(X'X)^{-1}X'$ be the "residual maker" matrix ($N \times N$)
    - notice that M is symmetric ($M' = M$) and idempotent ($M \times M = M$)
    - also notice that $MY = \hat{e}$ and $MX = 0$
    - and, $X'\hat{e} = 0$, by construction
- we can write $Y = X\hat{\beta} + \hat{e}$
  - i.e. decomposition into predicted values + residuals
  - these follow from facts about linear algebra, *not* anything about causality!
  - in fact, you could do this with purely deterministic data: no statistics necessary

Derivation of OLS Formula

- when we do this, the FOC become

$$(X'X)\hat{\beta} = X'Y \Longrightarrow \hat{\beta} = (X'X)^{-1}X'Y$$

- now let
  - $\hat{e} = Y - X\hat{\beta}$ be the residuals
  - $M = I - X(X'X)^{-1}X'$ be the "residual maker" matrix ($N \times N$)
    - notice that M is symmetric ($M' = M$) and idempotent ($M \times M = M$)
    - also notice that $MY = \hat{e}$ and $MX = 0$
    - and, $X'\hat{e} = 0$, by construction
- we can write $Y = X\hat{\beta} + \hat{e}$
  - i.e. decomposition into predicted values + residuals
  - these follow from facts about linear algebra, *not* anything about causality!
  - in fact, you could do this with purely deterministic data: no statistics necessary

Derivation of OLS Formula

- when we do this, the FOC become

$$(X'X)\hat{\beta} = X'Y \implies \hat{\beta} = (X'X)^{-1}X'Y$$

- now let
  - $\hat{e} = Y - X\hat{\beta}$ be the residuals
  - $M = I - X(X'X)^{-1}X'$ be the "residual maker" matrix ($N \times N$)
    - notice that M is symmetric ($M' = M$) and idempotent ($M \times M = M$)
    - also notice that $MY = \hat{e}$ and $MX = 0$
    - and, $X'\hat{e} = 0$, by construction
- we can write $Y = X\hat{\beta} + \hat{e}$
  - i.e. decomposition into predicted values + residuals
  - these follow from facts about linear algebra, *not* anything about causality!
  - in fact, you could do this with purely deterministic data: no statistics necessary

Derivation of OLS Formula

- when we do this, the FOC become

$$(X'X)\hat{\beta} = X'Y \Longrightarrow \hat{\beta} = (X'X)^{-1}X'Y$$

- now let
  - $\hat{e} = Y - X\hat{\beta}$ be the residuals
  - $M = I - X(X'X)^{-1}X'$ be the "residual maker" matrix ($N \times N$)
    - notice that M is symmetric ($M' = M$) and idempotent ($M \times M = M$)
    - also notice that $MY = \hat{e}$ and $MX = 0$
    - and, $X'\hat{e} = 0$, by construction
- we can write $Y = X\hat{\beta} + \hat{e}$
  - i.e. decomposition into predicted values + residuals
  - these follow from facts about linear algebra, *not* anything about causality!
  - in fact, you could do this with purely deterministic data: no statistics necessary

Derivation of OLS Formula

- when we do this, the FOC become

$$(X'X)\hat{\beta} = X'Y \implies \hat{\beta} = (X'X)^{-1}X'Y$$

- now let
  - $\hat{e} = Y - X\hat{\beta}$ be the residuals
  - $M = I - X(X'X)^{-1}X'$ be the "residual maker" matrix ($N \times N$)
    - notice that M is symmetric ($M' = M$) and idempotent ($M \times M = M$)
    - also notice that $MY = \hat{e}$ and $MX = 0$
    - and, $X'\hat{e} = 0$, by construction
- we can write $Y = X\hat{\beta} + \hat{e}$
  - i.e. decomposition into predicted values + residuals
  - these follow from facts about linear algebra, *not* anything about causality!
  - in fact, you could do this with purely deterministic data: no statistics necessary

Derivation of OLS Formula

- when we do this, the FOC become

$$(X'X)\hat{\beta} = X'Y \implies \hat{\beta} = (X'X)^{-1}X'Y$$

- now let
  - $\hat{e} = Y - X\hat{\beta}$ be the residuals
  - $M = I - X(X'X)^{-1}X'$ be the "residual maker" matrix ($N \times N$)
    - notice that M is symmetric ($M' = M$) and idempotent ($M \times M = M$)
    - also notice that $MY = \hat{e}$ and $MX = 0$
    - and, $X'\hat{e} = 0$, by construction
- we can write $Y = X\hat{\beta} + \hat{e}$
  - i.e. decomposition into predicted values + residuals
  - these follow from facts about linear algebra, *not* anything about causality!
  - in fact, you could do this with purely deterministic data: no statistics necessary

Derivation of OLS Formula

- when we do this, the FOC become

$$(X'X)\hat{\beta} = X'Y \Longrightarrow \hat{\beta} = (X'X)^{-1}X'Y$$

- now let
  - $\hat{e} = Y - X\hat{\beta}$ be the residuals
  - $M = I - X(X'X)^{-1}X'$ be the "residual maker" matrix ($N \times N$)
    - notice that M is symmetric ($M' = M$) and idempotent ($M \times M = M$)
    - also notice that $MY = \hat{e}$ and $MX = 0$
    - and, $X'\hat{e} = 0$, by construction
- we can write $Y = X\hat{\beta} + \hat{e}$
  - i.e. decomposition into predicted values + residuals
  - these follow from facts about linear algebra, *not* anything about causality!
  - in fact, you could do this with purely deterministic data: no statistics necessary

Frisch-Waugh-Lovell Theorem

- now, suppose we have $X = [X_1 \ X_2]$ where X_1 contains a constant and X_2 is a single column
  - we want to express the OLS coefficient on X_2 in a different way
  - why we want to do this will become clear later
- let $M_1 = I - X_1(X_1'X_1)^{-1}X_1'$
  - as before, this is symmetric and idempotent
  - further, $M_1X_1 = 0$
  - and, $M_1\hat{e} = \hat{e} - X_1(X_1'X_1)^{-1}X_1'\hat{e} = \hat{e} - 0$
- take our decomposition $Y = X_1\hat{\beta}_1 + X_2\hat{\beta}_2 + \hat{e}$ and premultiply by M_1
  - premultiply again by X_2' (check that the dimensions are appropriate!)
  - then, we get

$$X_2'M_1Y = X_2'M_1X_2\hat{\beta}_2$$

- use the fact that M_1 is symmetric and idempotent to write

$$\hat{\beta}_2 = ((x_2^*)'(x_2^*))^{-1}((x_2^*)'Y)$$

Frisch-Waugh-Lovell Theorem

- now, suppose we have $X = [X_1 \ X_2]$ where X_1 contains a constant and X_2 is a single column
  - we want to express the OLS coefficient on X_2 in a different way
  - why we want to do this will become clear later
- let $M_1 = I - X_1(X_1'X_1)^{-1}X_1'$
  - as before, this is symmetric and idempotent
  - further, $M_1 X_1 = 0$
  - and, $M_1 \hat{e} = \hat{e} - X_1(X_1'X_1)^{-1}X_1'\hat{e} = \hat{e} - 0$
- take our decomposition $Y = X_1\hat{\beta}_1 + X_2\hat{\beta}_2 + \hat{e}$ and premultiply by M_1
  - premultiply again by X_2' (check that the dimensions are appropriate!)
  - then, we get

$$X_2' M_1 Y = X_2' M_1 X_2 \hat{\beta}_2$$

- use the fact that M_1 is symmetric and idempotent to write

$$\hat{\beta}_2 = ((x_2^*)'(x_2^*))^{-1}((x_2^*)'Y)$$

Frisch-Waugh-Lovell Theorem

- now, suppose we have $X = [X_1 \ X_2]$ where X_1 contains a constant and X_2 is a single column
  - we want to express the OLS coefficient on X_2 in a different way
  - why we want to do this will become clear later
- let $M_1 = I - X_1(X_1'X_1)^{-1}X_1'$
  - as before, this is symmetric and idempotent
  - further, $M_1 X_1 = 0$
  - and, $M_1 \hat{e} = \hat{e} - X_1(X_1'X_1)^{-1}X_1'\hat{e} = \hat{e} - 0$
- take our decomposition $Y = X_1\hat{\beta}_1 + X_2\hat{\beta}_2 + \hat{e}$ and premultiply by M_1
  - premultiply again by X_2' (check that the dimensions are appropriate!)
  - then, we get

$$X_2'M_1Y = X_2'M_1X_2\hat{\beta}_2$$

- use the fact that M_1 is symmetric and idempotent to write

$$\hat{\beta}_2 = ((x_2^*)'(x_2^*))^{-1}((x_2^*)'Y)$$

Frisch-Waugh-Lovell Theorem

- now, suppose we have $X = [X_1 \;\; X_2]$ where X_1 contains a constant and X_2 is a single column
  - we want to express the OLS coefficient on X_2 in a different way
  - why we want to do this will become clear later
- let $M_1 = I - X_1(X_1'X_1)^{-1}X_1'$
  - as before, this is symmetric and idempotent
  - further, $M_1X_1 = 0$
  - and, $M_1\hat{e} = \hat{e} - X_1(X_1'X_1)^{-1}X_1'\hat{e} = \hat{e} - 0$
- take our decomposition $Y = X_1\hat{\beta}_1 + X_2\hat{\beta}_2 + \hat{e}$ and premultiply by M_1
  - premultiply again by X_2' (check that the dimensions are appropriate!)
  - then, we get

$$X_2'M_1Y = X_2'M_1X_2\hat{\beta}_2$$

- use the fact that M_1 is symmetric and idempotent to write

$$\hat{\beta}_2 = ((x_2^*)'(x_2^*))^{-1}((x_2^*)'Y)$$

Frisch-Waugh-Lovell Theorem

- now, suppose we have $X = [X_1 \ X_2]$ where X_1 contains a constant and X_2 is a single column
  - we want to express the OLS coefficient on X_2 in a different way
  - why we want to do this will become clear later
- let $M_1 = I - X_1(X_1'X_1)^{-1}X_1'$
  - as before, this is symmetric and idempotent
  - further, $M_1X_1 = 0$
  - and, $M_1\hat{e} = \hat{e} - X_1(X_1'X_1)^{-1}X_1'\hat{e} = \hat{e} - 0$
- take our decomposition $Y = X_1\hat{\beta}_1 + X_2\hat{\beta}_2 + \hat{e}$ and premultiply by M_1
  - premultiply again by X_2' (check that the dimensions are appropriate!)
  - then, we get

$$X_2'M_1Y = X_2'M_1X_2\hat{\beta}_2$$

- use the fact that M_1 is symmetric and idempotent to write

$$\hat{\beta}_2 = ((x_2^*)'(x_2^*))^{-1}((x_2^*)'Y)$$

Frisch-Waugh-Lovell Theorem

- now, suppose we have $X = [X_1 \ X_2]$ where X_1 contains a constant and X_2 is a single column
    - we want to express the OLS coefficient on X_2 in a different way
    - why we want to do this will become clear later
- let $M_1 = I - X_1(X_1'X_1)^{-1}X_1'$
    - as before, this is symmetric and idempotent
    - further, $M_1X_1 = 0$
    - and, $M_1\hat{e} = \hat{e} - X_1(X_1'X_1)^{-1}X_1'\hat{e} = \hat{e} - 0$
- take our decomposition $Y = X_1\hat{\beta}_1 + X_2\hat{\beta}_2 + \hat{e}$ and premultiply by M_1
    - premultiply again by X_2' (check that the dimensions are appropriate!)
    - then, we get

$$X_2'M_1Y = X_2'M_1X_2\hat{\beta}_2$$

- use the fact that M_1 is symmetric and idempotent to write

$$\hat{\beta}_2 = ((x_2^*)'(x_2^*))^{-1}((x_2^*)'Y)$$

Frisch-Waugh-Lovell Theorem

- now, suppose we have $X = [X_1 \; X_2]$ where X_1 contains a constant and X_2 is a single column
  - we want to express the OLS coefficient on X_2 in a different way
  - why we want to do this will become clear later
- let $M_1 = I - X_1(X_1'X_1)^{-1}X_1'$
  - as before, this is symmetric and idempotent
  - further, $M_1 X_1 = 0$
  - and, $M_1 \hat{e} = \hat{e} - X_1(X_1'X_1)^{-1}X_1'\hat{e} = \hat{e} - 0$
- take our decomposition $Y = X_1\hat{\beta}_1 + X_2\hat{\beta}_2 + \hat{e}$ and premultiply by M_1
  - premultiply again by X_2' (check that the dimensions are appropriate!)
  - then, we get

$$X_2'M_1Y = X_2'M_1X_2\hat{\beta}_2$$

- use the fact that M_1 is symmetric and idempotent to write

$$\hat{\beta}_2 = ((x_2^*)'(x_2^*))^{-1}((x_2^*)'Y)$$

Frisch-Waugh-Lovell Theorem

- now, suppose we have $X = [X_1 \ X_2]$ where X_1 contains a constant and X_2 is a single column
  - we want to express the OLS coefficient on X_2 in a different way
  - why we want to do this will become clear later
- let $M_1 = I - X_1(X_1'X_1)^{-1}X_1'$
  - as before, this is symmetric and idempotent
  - further, $M_1X_1 = 0$
  - and, $M_1\hat{e} = \hat{e} - X_1(X_1'X_1)^{-1}X_1'\hat{e} = \hat{e} - 0$
- take our decomposition $Y = X_1\hat{\beta}_1 + X_2\hat{\beta}_2 + \hat{e}$ and premultiply by M_1
  - premultiply again by X_2' (check that the dimensions are appropriate!)
  - then, we get

$$X_2'M_1Y = X_2'M_1X_2\hat{\beta}_2$$

- use the fact that M_1 is symmetric and idempotent to write

$$\hat{\beta}_2 = ((x_2^*)'(x_2^*))^{-1}((x_2^*)'Y)$$

Frisch-Waugh-Lovell Theorem

- now, suppose we have $X = [X_1 \ X_2]$ where X_1 contains a constant and X_2 is a single column
  - we want to express the OLS coefficient on X_2 in a different way
  - why we want to do this will become clear later
- let $M_1 = I - X_1(X_1'X_1)^{-1}X_1'$
  - as before, this is symmetric and idempotent
  - further, $M_1X_1 = 0$
  - and, $M_1\hat{e} = \hat{e} - X_1(X_1'X_1)^{-1}X_1'\hat{e} = \hat{e} - 0$
- take our decomposition $Y = X_1\hat{\beta}_1 + X_2\hat{\beta}_2 + \hat{e}$ and premultiply by M_1
  - premultiply again by X_2' (check that the dimensions are appropriate!)
  - then, we get

$$X_2'M_1Y = X_2'M_1X_2\hat{\beta}_2$$

- use the fact that M_1 is symmetric and idempotent to write

$$\hat{\beta}_2 = ((x_2^*)'(x_2^*))^{-1}((x_2^*)'Y)$$

Frisch-Waugh-Lovell Theorem

- now, suppose we have $X = [X_1 \ X_2]$ where X_1 contains a constant and X_2 is a single column
  - we want to express the OLS coefficient on X_2 in a different way
  - why we want to do this will become clear later
- let $M_1 = I - X_1(X_1'X_1)^{-1}X_1'$
  - as before, this is symmetric and idempotent
  - further, $M_1X_1 = 0$
  - and, $M_1\hat{e} = \hat{e} - X_1(X_1'X_1)^{-1}X_1'\hat{e} = \hat{e} - 0$
- take our decomposition $Y = X_1\hat{\beta}_1 + X_2\hat{\beta}_2 + \hat{e}$ and premultiply by M_1
  - premultiply again by X_2' (check that the dimensions are appropriate!)
  - then, we get

$$X_2'M_1Y = X_2'M_1X_2\hat{\beta}_2$$

- use the fact that M_1 is symmetric and idempotent to write

$$\hat{\beta}_2 = ((x_2^*)'(x_2^*))^{-1}((x_2^*)'Y)$$

Frisch-Waugh-Lovell Theorem

- now, suppose we have $X = [X_1 \;\; X_2]$ where X_1 contains a constant and X_2 is a single column
  - we want to express the OLS coefficient on X_2 in a different way
  - why we want to do this will become clear later
- let $M_1 = I - X_1(X_1'X_1)^{-1}X_1'$
  - as before, this is symmetric and idempotent
  - further, $M_1X_1 = 0$
  - and, $M_1\hat{e} = \hat{e} - X_1(X_1'X_1)^{-1}X_1'\hat{e} = \hat{e} - 0$
- take our decomposition $Y = X_1\hat{\beta}_1 + X_2\hat{\beta}_2 + \hat{e}$ and premultiply by M_1
  - premultiply again by X_2' (check that the dimensions are appropriate!)
  - then, we get

$$X_2'M_1Y = X_2'M_1X_2\hat{\beta}_2$$

- use the fact that M_1 is symmetric and idempotent to write

$$\hat{\beta}_2 = ((x_2^*)'(x_2^*))^{-1}((x_2^*)'Y)$$

Frisch-Waugh-Lovell Theorem

- the OLS coefficient on X_2 is numerically identical to the one we would obtain from:
  - regressing X_2 on X_1 and obtaining the residuals x_2^*
  - regressing Y on x_2^*
- implications:
  - if $X_2 \perp X_1$, we get the same coefficient on X_2 whether we include "controls" for X_1 or not
  - as we will see, richer conditioning sets are *not* always weakly better
  - we can interpret "multicollinearity" as a case where the residual variation in x_2^* is tiny

Frisch-Waugh-Lovell Theorem

- the OLS coefficient on X_2 is numerically identical to the one we would obtain from:
  - regressing X_2 on X_1 and obtaining the residuals x_2^*
  - regressing Y on x_2^*
- implications:
  - if $X_2 \perp X_1$, we get the same coefficient on X_2 whether we include "controls" for X_1 or not
  - as we will see, richer conditioning sets are *not* always weakly better
  - we can interpret "multicollinearity" as a case where the residual variation in x_2^* is tiny

Frisch-Waugh-Lovell Theorem

- the OLS coefficient on X_2 is numerically identical to the one we would obtain from:
  - regressing X_2 on X_1 and obtaining the residuals x_2^*
  - regressing Y on x_2^*
- implications:
  - if $X_2 \perp X_1$, we get the same coefficient on X_2 whether we include "controls" for X_1 or not
  - as we will see, richer conditioning sets are *not* always weakly better
  - we can interpret "multicollinearity" as a case where the residual variation in x_2^* is tiny

Frisch-Waugh-Lovell Theorem

- the OLS coefficient on X_2 is numerically identical to the one we would obtain from:
    - regressing X_2 on X_1 and obtaining the residuals x_2^*
    - regressing Y on x_2^*
- implications:
    - if $X_2 \perp X_1$, we get the same coefficient on X_2 whether we include "controls" for X_1 or not
    - as we will see, richer conditioning sets are *not* always weakly better
    - we can interpret "multicollinearity" as a case where the residual variation in x_2^* is tiny

Frisch-Waugh-Lovell Theorem

- the OLS coefficient on X_2 is numerically identical to the one we would obtain from:
  - regressing X_2 on X_1 and obtaining the residuals x_2^*
  - regressing Y on x_2^*
- implications:
  - if $X_2 \perp X_1$, we get the same coefficient on X_2 whether we include "controls" for X_1 or not
  - as we will see, richer conditioning sets are *not* always weakly better
  - we can interpret "multicollinearity" as a case where the residual variation in x_2^* is tiny

Frisch-Waugh-Lovell Theorem

- the OLS coefficient on X_2 is numerically identical to the one we would obtain from:
  - regressing X_2 on X_1 and obtaining the residuals x_2^*
  - regressing Y on x_2^*
- implications:
  - if $X_2 \perp X_1$, we get the same coefficient on X_2 whether we include "controls" for X_1 or not
  - as we will see, richer conditioning sets are *not* always weakly better
  - we can interpret "multicollinearity" as a case where the residual variation in x_2^* is tiny

Frisch-Waugh-Lovell Theorem

- the OLS coefficient on X_2 is numerically identical to the one we would obtain from:
  - regressing X_2 on X_1 and obtaining the residuals x_2^*
  - regressing Y on x_2^*
- implications:
  - if $X_2 \perp X_1$, we get the same coefficient on X_2 whether we include "controls" for X_1 or not
  - as we will see, richer conditioning sets are *not* always weakly better
  - we can interpret "multicollinearity" as a case where the residual variation in x_2^* is tiny

Wages, Experience and IQ

- suppose we have a sample of workers observed at the same date
  - we observe their wages (w), and the results of an IQ test (x)
  - all workers in the sample were tested at the same date, say 20 years ago
  - we also have their age in years which we encode in a vector of dummies (D)
- suppose we want to estimate

$$\log w = \alpha x + D\gamma + e$$

- suppose age has the following effects:
  - as workers gain experience their productivity rises and employers may pay them more
  - some cohorts differ in ability because of changes in e.g. school quality or environmental factors
- also assume that age at testing affects measured IQ, with older kids doing better on average

Wages, Experience and IQ

- suppose we have a sample of workers observed at the same date
  - we observe their wages (w), and the results of an IQ test (x)
  - all workers in the sample were tested at the same date, say 20 years ago
  - we also have their age in years which we encode in a vector of dummies (D)
- suppose we want to estimate

$$\log w = \alpha x + D\gamma + e$$

- suppose age has the following effects:
  - as workers gain experience their productivity rises and employers may pay them more
  - some cohorts differ in ability because of changes in e.g. school quality or environmental factors
- also assume that age at testing affects measured IQ, with older kids doing better on average

Wages, Experience and IQ

- suppose we have a sample of workers observed at the same date
  - we observe their wages (w), and the results of an IQ test (x)
  - all workers in the sample were tested at the same date, say 20 years ago
  - we also have their age in years which we encode in a vector of dummies (D)
- suppose we want to estimate

$$\log w = \alpha x + D\gamma + e$$

- suppose age has the following effects:
  - as workers gain experience their productivity rises and employers may pay them more
  - some cohorts differ in ability because of changes in e.g. school quality or environmental factors
- also assume that age at testing affects measured IQ, with older kids doing better on average

Wages, Experience and IQ

- suppose we have a sample of workers observed at the same date
    - we observe their wages (w), and the results of an IQ test (x)
    - all workers in the sample were tested at the same date, say 20 years ago
    - we also have their age in years which we encode in a vector of dummies (D)
- suppose we want to estimate

$$\log w = \alpha x + D\gamma + e$$

- suppose age has the following effects:
    - as workers gain experience their productivity rises and employers may pay them more
    - some cohorts differ in ability because of changes in e.g. school quality or environmental factors
- also assume that age at testing affects measured IQ, with older kids doing better on average

Wages, Experience and IQ

- suppose we have a sample of workers observed at the same date
  - we observe their wages (w), and the results of an IQ test (x)
  - all workers in the sample were tested at the same date, say 20 years ago
  - we also have their age in years which we encode in a vector of dummies (D)
- suppose we want to estimate

$$\log w = \alpha x + D\gamma + e$$

- suppose age has the following effects:
  - as workers gain experience their productivity rises and employers may pay them more
  - some cohorts differ in ability because of changes in e.g. school quality or environmental factors
- also assume that age at testing affects measured IQ, with older kids doing better on average

Wages, Experience and IQ

- suppose we have a sample of workers observed at the same date
  - we observe their wages (w), and the results of an IQ test (x)
  - all workers in the sample were tested at the same date, say 20 years ago
  - we also have their age in years which we encode in a vector of dummies (D)
- suppose we want to estimate

$$\log w = \alpha x + D\gamma + e$$

- suppose age has the following effects:
  - as workers gain experience their productivity rises and employers may pay them more
  - some cohorts differ in ability because of changes in e.g. school quality or environmental factors
- also assume that age at testing affects measured IQ, with older kids doing better on average

Wages, Experience and IQ

- suppose we have a sample of workers observed at the same date
  - we observe their wages (w), and the results of an IQ test (x)
  - all workers in the sample were tested at the same date, say 20 years ago
  - we also have their age in years which we encode in a vector of dummies (D)
- suppose we want to estimate

$$\log w = \alpha x + D\gamma + e$$

- suppose age has the following effects:
  - as workers gain experience their productivity rises and employers may pay them more
  - some cohorts differ in ability because of changes in e.g. school quality or environmental factors
- also assume that age at testing affects measured IQ, with older kids doing better on average

Wages, Experience and IQ

- suppose we have a sample of workers observed at the same date
  - we observe their wages (w), and the results of an IQ test (x)
  - all workers in the sample were tested at the same date, say 20 years ago
  - we also have their age in years which we encode in a vector of dummies (D)
- suppose we want to estimate

$$\log w = \alpha x + D\gamma + e$$

- suppose age has the following effects:
  - as workers gain experience their productivity rises and employers may pay them more
  - some cohorts differ in ability because of changes in e.g. school quality or environmental factors
- also assume that age at testing affects measured IQ, with older kids doing better on average

Wages, Experience and IQ

- suppose we have a sample of workers observed at the same date
  - we observe their wages (w), and the results of an IQ test (x)
  - all workers in the sample were tested at the same date, say 20 years ago
  - we also have their age in years which we encode in a vector of dummies (D)
- suppose we want to estimate

$$\log w = \alpha x + D\gamma + e$$

- suppose age has the following effects:
  - as workers gain experience their productivity rises and employers may pay them more
  - some cohorts differ in ability because of changes in e.g. school quality or environmental factors
- also assume that age at testing affects measured IQ, with older kids doing better on average

Wages, Experience and IQ

- does it make sense to adjust the IQ scores for age?
  - it depends: do we want
    - the effect of ability on wages, holding experience constant?
    - or, the effects of experience, holding ability constant?
- notice that all of the following regressions will give the same $\hat{\alpha}$
  - regress $\log w$ on $\alpha x + D\gamma$ (1)
  - regress $\log w$ on $\alpha x^* + D\gamma$ (2)
  - regress $\log w$ on αx^* (3)
- however, the estimated *age* effects will change
  - consider regressing $\log w$ on D alone (4)
  - because $x^* \perp D$ by construction, (4) gives the same $\hat{\gamma}$ as (2)

Wages, Experience and IQ

- does it make sense to adjust the IQ scores for age?
  - it depends: do we want
    - the effect of ability on wages, holding experience constant?
    - or, the effects of experience, holding ability constant?
- notice that all of the following regressions will give the same $\hat{\alpha}$
  - regress $\log w$ on $\alpha x + D\gamma$ (1)
  - regress $\log w$ on $\alpha x^* + D\gamma$ (2)
  - regress $\log w$ on αx^* (3)
- however, the estimated *age* effects will change
  - consider regressing $\log w$ on D alone (4)
  - because $x^* \perp D$ by construction, (4) gives the same $\hat{\gamma}$ as (2)

Wages, Experience and IQ

- does it make sense to adjust the IQ scores for age?
  - it depends: do we want
    - the effect of ability on wages, holding experience constant?
    - or, the effects of experience, holding ability constant?
- notice that all of the following regressions will give the same $\hat{\alpha}$
  - regress $\log w$ on $\alpha x + D\gamma$ (1)
  - regress $\log w$ on $\alpha x^* + D\gamma$ (2)
  - regress $\log w$ on αx^* (3)
- however, the estimated *age* effects will change
  - consider regressing $\log w$ on D alone (4)
  - because $x^* \perp D$ by construction, (4) gives the same $\hat{\gamma}$ as (2)

Wages, Experience and IQ

- does it make sense to adjust the IQ scores for age?
  - it depends: do we want
    - the effect of ability on wages, holding experience constant?
    - or, the effects of experience, holding ability constant?
- notice that all of the following regressions will give the same $\hat{\alpha}$
  - regress $\log w$ on $\alpha x + D\gamma$ (1)
  - regress $\log w$ on $\alpha x^* + D\gamma$ (2)
  - regress $\log w$ on αx^* (3)
- however, the estimated *age* effects will change
  - consider regressing $\log w$ on D alone (4)
  - because $x^* \perp D$ by construction, (4) gives the same $\hat{\gamma}$ as (2)

Wages, Experience and IQ

- does it make sense to adjust the IQ scores for age?
  - it depends: do we want
    - the effect of ability on wages, holding experience constant?
    - or, the effects of experience, holding ability constant?
- notice that all of the following regressions will give the same $\hat{\alpha}$
  - regress $\log w$ on $\alpha x + D\gamma$ (1)
  - regress $\log w$ on $\alpha x^* + D\gamma$ (2)
  - regress $\log w$ on αx^* (3)
- however, the estimated *age* effects will change
  - consider regressing $\log w$ on D alone (4)
  - because $x^* \perp D$ by construction, (4) gives the same $\hat{\gamma}$ as (2)

Wages, Experience and IQ

- does it make sense to adjust the IQ scores for age?
  - it depends: do we want
    - the effect of ability on wages, holding experience constant?
    - or, the effects of experience, holding ability constant?
- notice that all of the following regressions will give the same $\hat{\alpha}$
  - regress $\log w$ on $\alpha x + D\gamma$ (1)
  - regress $\log w$ on $\alpha x^* + D\gamma$ (2)
  - regress $\log w$ on αx^* (3)
- however, the estimated *age* effects will change
  - consider regressing $\log w$ on D alone (4)
  - because $x^* \perp D$ by construction, (4) gives the same $\hat{\gamma}$ as (2)

Wages, Experience and IQ

- does it make sense to adjust the IQ scores for age?
  - it depends: do we want
    - the effect of ability on wages, holding experience constant?
    - or, the effects of experience, holding ability constant?
- notice that all of the following regressions will give the same $\hat{\alpha}$
  - regress $\log w$ on $\alpha x + D\gamma$ (1)
  - regress $\log w$ on $\alpha x^* + D\gamma$ (2)
  - regress $\log w$ on αx^* (3)
- however, the estimated *age* effects will change
  - consider regressing $\log w$ on D alone (4)
  - because $x^* \perp D$ by construction, (4) gives the same $\hat{\gamma}$ as (2)

Wages, Experience and IQ

- does it make sense to adjust the IQ scores for age?
  - it depends: do we want
    - the effect of ability on wages, holding experience constant?
    - or, the effects of experience, holding ability constant?
- notice that all of the following regressions will give the same $\hat{\alpha}$
  - regress $\log w$ on $\alpha x + D\gamma$ (1)
  - regress $\log w$ on $\alpha x^* + D\gamma$ (2)
  - regress $\log w$ on αx^* (3)
- however, the estimated *age* effects will change
  - consider regressing $\log w$ on D alone (4)
  - because $x^* \perp D$ by construction, (4) gives the same $\hat{\gamma}$ as (2)

Wages, Experience and IQ

- does it make sense to adjust the IQ scores for age?
  - it depends: do we want
    - the effect of ability on wages, holding experience constant?
    - or, the effects of experience, holding ability constant?
- notice that all of the following regressions will give the same $\hat{\alpha}$
  - regress $\log w$ on $\alpha x + D\gamma$ (1)
  - regress $\log w$ on $\alpha x^* + D\gamma$ (2)
  - regress $\log w$ on αx^* (3)
- however, the estimated *age* effects will change
  - consider regressing $\log w$ on D alone (4)
  - because $x^* \perp D$ by construction, (4) gives the same $\hat{\gamma}$ as (2)

Wages, Experience and IQ

- does it make sense to adjust the IQ scores for age?
  - it depends: do we want
    - the effect of ability on wages, holding experience constant?
    - or, the effects of experience, holding ability constant?
- notice that all of the following regressions will give the same $\hat{\alpha}$
  - regress $\log w$ on $\alpha x + D\gamma$ (1)
  - regress $\log w$ on $\alpha x^* + D\gamma$ (2)
  - regress $\log w$ on αx^* (3)
- however, the estimated *age* effects will change
  - consider regressing $\log w$ on D alone (4)
  - because $x^* \perp D$ by construction, (4) gives the same $\hat{\gamma}$ as (2)

Wages, Experience and IQ

- does it make sense to adjust the IQ scores for age?
  - it depends: do we want
    - the effect of ability on wages, holding experience constant?
    - or, the effects of experience, holding ability constant?
- notice that all of the following regressions will give the same $\hat{\alpha}$
  - regress $\log w$ on $\alpha x + D\gamma$ (1)
  - regress $\log w$ on $\alpha x^* + D\gamma$ (2)
  - regress $\log w$ on αx^* (3)
- however, the estimated *age* effects will change
  - consider regressing $\log w$ on D alone (4)
  - because $x^* \perp D$ by construction, (4) gives the same $\hat{\gamma}$ as (2)

Wages, Experience and IQ

- how should we interpret the age effects when we age-adjust IQ compared to raw IQ?
  - is it possible to determine whether a given cohort earns more at a given point in time because of
    - higher ability
    - greater experience
    - some combination of the two?
- clearly, without including the IQ measure, the age coefficients pick up both effects
- with IQ as a control, the age effects will be estimated using variation in age that is not predicted by IQ
  - but, is this variation that is orthogonal to cohort ability?
- example: two cohorts
  - one is younger but smarter
  - another is older but less able
  - the two effects could cancel each other out so that IQ $\perp$ cohort

Wages, Experience and IQ

- how should we interpret the age effects when we age-adjust IQ compared to raw IQ?
  - is it possible to determine whether a given cohort earns more at a given point in time because of
    - higher ability
    - greater experience
    - some combination of the two?
- clearly, without including the IQ measure, the age coefficients pick up both effects
- with IQ as a control, the age effects will be estimated using variation in age that is not predicted by IQ
  - but, is this variation that is orthogonal to cohort ability?
- example: two cohorts
  - one is younger but smarter
  - another is older but less able
  - the two effects could cancel each other out so that IQ $\perp$ cohort

Wages, Experience and IQ

- how should we interpret the age effects when we age-adjust IQ compared to raw IQ?
  - is it possible to determine whether a given cohort earns more at a given point in time because of
    - higher ability
    - greater experience
    - some combination of the two?
- clearly, without including the IQ measure, the age coefficients pick up both effects
- with IQ as a control, the age effects will be estimated using variation in age that is not predicted by IQ
  - but, is this variation that is orthogonal to cohort ability?
- example: two cohorts
  - one is younger but smarter
  - another is older but less able
  - the two effects could cancel each other out so that IQ $\perp$ cohort

Wages, Experience and IQ

- how should we interpret the age effects when we age-adjust IQ compared to raw IQ?
  - is it possible to determine whether a given cohort earns more at a given point in time because of
    - higher ability
    - greater experience
    - some combination of the two?
- clearly, without including the IQ measure, the age coefficients pick up both effects
- with IQ as a control, the age effects will be estimated using variation in age that is not predicted by IQ
  - but, is this variation that is orthogonal to cohort ability?
- example: two cohorts
  - one is younger but smarter
  - another is older but less able
  - the two effects could cancel each other out so that IQ $\perp$ cohort

Wages, Experience and IQ

- how should we interpret the age effects when we age-adjust IQ compared to raw IQ?
  - is it possible to determine whether a given cohort earns more at a given point in time because of
    - higher ability
    - greater experience
    - some combination of the two?
- clearly, without including the IQ measure, the age coefficients pick up both effects
- with IQ as a control, the age effects will be estimated using variation in age that is not predicted by IQ
  - but, is this variation that is orthogonal to cohort ability?
- example: two cohorts
  - one is younger but smarter
  - another is older but less able
  - the two effects could cancel each other out so that IQ $\perp$ cohort

Wages, Experience and IQ

- how should we interpret the age effects when we age-adjust IQ compared to raw IQ?
  - is it possible to determine whether a given cohort earns more at a given point in time because of
    - higher ability
    - greater experience
    - some combination of the two?
- clearly, without including the IQ measure, the age coefficients pick up both effects
- with IQ as a control, the age effects will be estimated using variation in age that is not predicted by IQ
  - but, is this variation that is orthogonal to cohort ability?
- example: two cohorts
  - one is younger but smarter
  - another is older but less able
  - the two effects could cancel each other out so that IQ $\perp$ cohort

Wages, Experience and IQ

- how should we interpret the age effects when we age-adjust IQ compared to raw IQ?
  - is it possible to determine whether a given cohort earns more at a given point in time because of
    - higher ability
    - greater experience
    - some combination of the two?
- clearly, without including the IQ measure, the age coefficients pick up both effects
- with IQ as a control, the age effects will be estimated using variation in age that is not predicted by IQ
  - but, is this variation that is orthogonal to cohort ability?
- example: two cohorts
  - one is younger but smarter
  - another is older but less able
  - the two effects could cancel each other out so that IQ $\perp$ cohort

Wages, Experience and IQ

- how should we interpret the age effects when we age-adjust IQ compared to raw IQ?
  - is it possible to determine whether a given cohort earns more at a given point in time because of
    - higher ability
    - greater experience
    - some combination of the two?
- clearly, without including the IQ measure, the age coefficients pick up both effects
- with IQ as a control, the age effects will be estimated using variation in age that is not predicted by IQ
  - but, is this variation that is orthogonal to cohort ability?
- example: two cohorts
  - one is younger but smarter
  - another is older but less able
  - the two effects could cancel each other out so that IQ $\perp$ cohort

Wages, Experience and IQ

- how should we interpret the age effects when we age-adjust IQ compared to raw IQ?
  - is it possible to determine whether a given cohort earns more at a given point in time because of
    - higher ability
    - greater experience
    - some combination of the two?
- clearly, without including the IQ measure, the age coefficients pick up both effects
- with IQ as a control, the age effects will be estimated using variation in age that is not predicted by IQ
  - but, is this variation that is orthogonal to cohort ability?
- example: two cohorts
  - one is younger but smarter
  - another is older but less able
  - the two effects could cancel each other out so that IQ $\perp$ cohort

Wages, Experience and IQ

- how should we interpret the age effects when we age-adjust IQ compared to raw IQ?
  - is it possible to determine whether a given cohort earns more at a given point in time because of
    - higher ability
    - greater experience
    - some combination of the two?
- clearly, without including the IQ measure, the age coefficients pick up both effects
- with IQ as a control, the age effects will be estimated using variation in age that is not predicted by IQ
  - but, is this variation that is orthogonal to cohort ability?
- example: two cohorts
  - one is younger but smarter
  - another is older but less able
  - the two effects could cancel each other out so that IQ $\perp$ cohort

Wages, Experience and IQ

- how should we interpret the age effects when we age-adjust IQ compared to raw IQ?
  - is it possible to determine whether a given cohort earns more at a given point in time because of
    - higher ability
    - greater experience
    - some combination of the two?
- clearly, without including the IQ measure, the age coefficients pick up both effects
- with IQ as a control, the age effects will be estimated using variation in age that is not predicted by IQ
  - but, is this variation that is orthogonal to cohort ability?
- example: two cohorts
  - one is younger but smarter
  - another is older but less able
  - the two effects could cancel each other out so that IQ $\perp$ cohort

Wages, Experience and IQ

- how should we interpret the age effects when we age-adjust IQ compared to raw IQ?
  - is it possible to determine whether a given cohort earns more at a given point in time because of
    - higher ability
    - greater experience
    - some combination of the two?
- clearly, without including the IQ measure, the age coefficients pick up both effects
- with IQ as a control, the age effects will be estimated using variation in age that is not predicted by IQ
  - but, is this variation that is orthogonal to cohort ability?
- example: two cohorts
  - one is younger but smarter
  - another is older but less able
  - the two effects could cancel each other out so that IQ $\perp$ cohort

Wages, Experience and IQ

- now, supposing that IQ $\perp D$, compare
  - regress log w on $\alpha x + D\gamma$ (1)
  - regress log w on $\alpha x^* + D\gamma$ (2)
  - regress log w on $D\gamma$ (4)
  - regress log w on $D^*\gamma$ (5)
- because $x \perp D$ in this example, (4) and (5) give the same estimate of γ
  - further, (1) and (4) give the same estimate of γ for the same reason
  - and, we have seen (2) and (4) always give the same estimate of γ
- we have no way of telling the difference between changes in experience and cohort ability
- the fundamental problem is:
  - we have no information about how measured IQ would change with the age of testing
  - adjusting, normalizing, etc do not solve this problem!
  - age-adjusting IQ scores isolates within-cohort variation in ability, but it does not tell us about which cohorts are smart or dumb

Wages, Experience and IQ

- now, supposing that IQ $\perp D$, compare
  - regress log w on $\alpha x + D\gamma$ (1)
  - regress log w on $\alpha x^* + D\gamma$ (2)
  - regress log w on $D\gamma$ (4)
  - regress log w on $D^*\gamma$ (5)
- because $x \perp D$ in this example, (4) and (5) give the same estimate of γ
  - further, (1) and (4) give the same estimate of γ for the same reason
  - and, we have seen (2) and (4) always give the same estimate of γ
- we have no way of telling the difference between changes in experience and cohort ability
- the fundamental problem is:
  - we have no information about how measured IQ would change with the age of testing
  - adjusting, normalizing, etc do not solve this problem!
  - age-adjusting IQ scores isolates within-cohort variation in ability, but it does not tell us about which cohorts are smart or dumb

Wages, Experience and IQ

- now, supposing that IQ $\perp D$, compare
  - regress $\log w$ on $\alpha x + D\gamma$ (1)
  - regress $\log w$ on $\alpha x^* + D\gamma$ (2)
  - regress $\log w$ on $D\gamma$ (4)
  - regress $\log w$ on $D^*\gamma$ (5)
- because $x \perp D$ in this example, (4) and (5) give the same estimate of γ
  - further, (1) and (4) give the same estimate of γ for the same reason
  - and, we have seen (2) and (4) always give the same estimate of γ
- we have no way of telling the difference between changes in experience and cohort ability
- the fundamental problem is:
  - we have no information about how measured IQ would change with the age of testing
  - adjusting, normalizing, etc do not solve this problem!
  - age-adjusting IQ scores isolates within-cohort variation in ability, but it does not tell us about which cohorts are smart or dumb

Wages, Experience and IQ

- now, supposing that IQ $\perp D$, compare
  - regress $\log w$ on $\alpha x + D\gamma$ (1)
  - regress $\log w$ on $\alpha x^* + D\gamma$ (2)
  - regress $\log w$ on $D\gamma$ (4)
  - regress $\log w$ on $D^*\gamma$ (5)
- because $x \perp D$ in this example, (4) and (5) give the same estimate of γ
  - further, (1) and (4) give the same estimate of γ for the same reason
  - and, we have seen (2) and (4) always give the same estimate of γ
- we have no way of telling the difference between changes in experience and cohort ability
- the fundamental problem is:
  - we have no information about how measured IQ would change with the age of testing
  - adjusting, normalizing, etc do not solve this problem!
  - age-adjusting IQ scores isolates within-cohort variation in ability, but it does not tell us about which cohorts are smart or dumb

Wages, Experience and IQ

- now, supposing that IQ $\perp D$, compare
  - regress $\log w$ on $\alpha x + D\gamma$ (1)
  - regress $\log w$ on $\alpha x^* + D\gamma$ (2)
  - regress $\log w$ on $D\gamma$ (4)
  - regress $\log w$ on $D^*\gamma$ (5)
- because $x \perp D$ in this example, (4) and (5) give the same estimate of γ
  - further, (1) and (4) give the same estimate of γ for the same reason
  - and, we have seen (2) and (4) always give the same estimate of γ
- we have no way of telling the difference between changes in experience and cohort ability
- the fundamental problem is:
  - we have no information about how measured IQ would change with the age of testing
  - adjusting, normalizing, etc do not solve this problem!
  - age-adjusting IQ scores isolates within-cohort variation in ability, but it does not tell us about which cohorts are smart or dumb

Wages, Experience and IQ

- now, supposing that IQ $\perp D$, compare
  - regress log w on $\alpha x + D\gamma$ (1)
  - regress log w on $\alpha x^* + D\gamma$ (2)
  - regress log w on $D\gamma$ (4)
  - regress log w on $D^*\gamma$ (5)
- because $x \perp D$ in this example, (4) and (5) give the same estimate of γ
  - further, (1) and (4) give the same estimate of γ for the same reason
  - and, we have seen (2) and (4) always give the same estimate of γ
- we have no way of telling the difference between changes in experience and cohort ability
- the fundamental problem is:
  - we have no information about how measured IQ would change with the age of testing
  - adjusting, normalizing, etc do not solve this problem!
  - age-adjusting IQ scores isolates within-cohort variation in ability, but it does not tell us about which cohorts are smart or dumb

Wages, Experience and IQ

- now, supposing that IQ $\perp D$, compare
  - regress $\log w$ on $\alpha x + D\gamma$ (1)
  - regress $\log w$ on $\alpha x^* + D\gamma$ (2)
  - regress $\log w$ on $D\gamma$ (4)
  - regress $\log w$ on $D^*\gamma$ (5)
- because $x \perp D$ in this example, (4) and (5) give the same estimate of γ
  - further, (1) and (4) give the same estimate of γ for the same reason
  - and, we have seen (2) and (4) always give the same estimate of γ
- we have no way of telling the difference between changes in experience and cohort ability
- the fundamental problem is:
  - we have no information about how measured IQ would change with the age of testing
  - adjusting, normalizing, etc do not solve this problem!
  - age-adjusting IQ scores isolates within-cohort variation in ability, but it does not tell us about which cohorts are smart or dumb

Wages, Experience and IQ

- now, supposing that IQ $\perp D$, compare
  - regress log w on $\alpha x + D\gamma$ (1)
  - regress log w on $\alpha x^* + D\gamma$ (2)
  - regress log w on $D\gamma$ (4)
  - regress log w on $D^*\gamma$ (5)
- because $x \perp D$ in this example, (4) and (5) give the same estimate of γ
  - further, (1) and (4) give the same estimate of γ for the same reason
  - and, we have seen (2) and (4) always give the same estimate of γ
- we have no way of telling the difference between changes in experience and cohort ability
- the fundamental problem is:
  - we have no information about how measured IQ would change with the age of testing
  - adjusting, normalizing, etc do not solve this problem!
  - age-adjusting IQ scores isolates within-cohort variation in ability, but it does not tell us about which cohorts are smart or dumb

Wages, Experience and IQ

- now, supposing that IQ $\perp D$, compare
  - regress log w on $\alpha x + D\gamma$ (1)
  - regress log w on $\alpha x^* + D\gamma$ (2)
  - regress log w on $D\gamma$ (4)
  - regress log w on $D^*\gamma$ (5)
- because $x \perp D$ in this example, (4) and (5) give the same estimate of γ
  - further, (1) and (4) give the same estimate of γ for the same reason
  - and, we have seen (2) and (4) always give the same estimate of γ
- we have no way of telling the difference between changes in experience and cohort ability
- the fundamental problem is:
  - we have no information about how measured IQ would change with the age of testing
  - adjusting, normalizing, etc do not solve this problem!
  - age-adjusting IQ scores isolates within-cohort variation in ability, but it does not tell us about which cohorts are smart or dumb

Wages, Experience and IQ

- now, supposing that IQ $\perp D$, compare
  - regress log w on $\alpha x + D\gamma$ (1)
  - regress log w on $\alpha x^* + D\gamma$ (2)
  - regress log w on $D\gamma$ (4)
  - regress log w on $D^*\gamma$ (5)
- because $x \perp D$ in this example, (4) and (5) give the same estimate of γ
  - further, (1) and (4) give the same estimate of γ for the same reason
  - and, we have seen (2) and (4) always give the same estimate of γ
- we have no way of telling the difference between changes in experience and cohort ability
- the fundamental problem is:
  - we have no information about how measured IQ would change with the age of testing
  - adjusting, normalizing, etc do not solve this problem!
  - age-adjusting IQ scores isolates within-cohort variation in ability, but it does not tell us about which cohorts are smart or dumb

Wages, Experience and IQ

- now, supposing that IQ $\perp D$, compare
  - regress $\log w$ on $\alpha x + D\gamma$ (1)
  - regress $\log w$ on $\alpha x^* + D\gamma$ (2)
  - regress $\log w$ on $D\gamma$ (4)
  - regress $\log w$ on $D^*\gamma$ (5)
- because $x \perp D$ in this example, (4) and (5) give the same estimate of γ
  - further, (1) and (4) give the same estimate of γ for the same reason
  - and, we have seen (2) and (4) always give the same estimate of γ
- we have no way of telling the difference between changes in experience and cohort ability
- the fundamental problem is:
  - we have no information about how measured IQ would change with the age of testing
  - adjusting, normalizing, etc do not solve this problem!
  - age-adjusting IQ scores isolates within-cohort variation in ability, but it does not tell us about which cohorts are smart or dumb

Wages, Experience and IQ

- now, supposing that IQ $\perp D$, compare
  - regress $\log w$ on $\alpha x + D\gamma$ (1)
  - regress $\log w$ on $\alpha x^* + D\gamma$ (2)
  - regress $\log w$ on $D\gamma$ (4)
  - regress $\log w$ on $D^*\gamma$ (5)
- because $x \perp D$ in this example, (4) and (5) give the same estimate of γ
  - further, (1) and (4) give the same estimate of γ for the same reason
  - and, we have seen (2) and (4) always give the same estimate of γ
- we have no way of telling the difference between changes in experience and cohort ability
- the fundamental problem is:
  - we have no information about how measured IQ would change with the age of testing
  - adjusting, normalizing, etc do not solve this problem!
  - age-adjusting IQ scores isolates within-cohort variation in ability, but it does not tell us about which cohorts are smart or dumb

Wages, Experience and IQ

- now, supposing that IQ $\perp D$, compare
  - regress $\log w$ on $\alpha x + D\gamma$ (1)
  - regress $\log w$ on $\alpha x^* + D\gamma$ (2)
  - regress $\log w$ on $D\gamma$ (4)
  - regress $\log w$ on $D^*\gamma$ (5)
- because $x \perp D$ in this example, (4) and (5) give the same estimate of γ
  - further, (1) and (4) give the same estimate of γ for the same reason
  - and, we have seen (2) and (4) always give the same estimate of γ
- we have no way of telling the difference between changes in experience and cohort ability
- the fundamental problem is:
  - we have no information about how measured IQ would change with the age of testing
  - adjusting, normalizing, etc do not solve this problem!
  - age-adjusting IQ scores isolates within-cohort variation in ability, but it does not tell us about which cohorts are smart or dumb

Wages, Experience and IQ

- now, what if you had instead z, an IQ test administered at the same *age*?
  - would controlling for z help us isolate the experience effects?
- answer: depends whether you adjust z for age.
  - if you use raw z, you are fine
  - the component of ability that predicts cohort gets residualized out by OLS
  - and, our D^* will be variation in experience that is orthogonal to cohort ability
- HOWEVER, if you age-adjust z to form z^*, you undo the benefit of having kids tested at the same age
  - you would then be regressing log wages on $\alpha z^* + D\gamma$
  - because z^* is orthogonal to D by construction, we know the estimated γ will be identical to the one from regressing log wages on D alone
    - and, as we discussed, that reflects both cohort and experience effects

Wages, Experience and IQ

- now, what if you had instead z, an IQ test administered at the same *age*?
  - would controlling for z help us isolate the experience effects?
- answer: depends whether you adjust z for age.
  - if you use raw z, you are fine
  - the component of ability that predicts cohort gets residualized out by OLS
  - and, our D^* will be variation in experience that is orthogonal to cohort ability
- HOWEVER, if you age-adjust z to form z^*, you undo the benefit of having kids tested at the same age
  - you would then be regressing log wages on $\alpha z^* + D\gamma$
  - because z^* is orthogonal to D by construction, we know the estimated γ will be identical to the one from regressing log wages on D alone
    - and, as we discussed, that reflects both cohort and experience effects

Wages, Experience and IQ

- now, what if you had instead z, an IQ test administered at the same *age*?
  - would controlling for z help us isolate the experience effects?
- answer: depends whether you adjust z for age.
  - if you use raw z, you are fine
  - the component of ability that predicts cohort gets residualized out by OLS
  - and, our D^* will be variation in experience that is orthogonal to cohort ability
- HOWEVER, if you age-adjust z to form z^*, you undo the benefit of having kids tested at the same age
  - you would then be regressing log wages on $\alpha z^* + D\gamma$
  - because z^* is orthogonal to D by construction, we know the estimated γ will be identical to the one from regressing log wages on D alone
    - and, as we discussed, that reflects both cohort and experience effects

Wages, Experience and IQ

- now, what if you had instead z, an IQ test administered at the same *age*?
  - would controlling for z help us isolate the experience effects?
- answer: depends whether you adjust z for age.
  - if you use raw z, you are fine
  - the component of ability that predicts cohort gets residualized out by OLS
  - and, our D^* will be variation in experience that is orthogonal to cohort ability
- HOWEVER, if you age-adjust z to form z^*, you undo the benefit of having kids tested at the same age
  - you would then be regressing log wages on $\alpha z^* + D\gamma$
  - because z^* is orthogonal to D by construction, we know the estimated γ will be identical to the one from regressing log wages on D alone
    - and, as we discussed, that reflects both cohort and experience effects

Wages, Experience and IQ

- now, what if you had instead z, an IQ test administered at the same *age*?
  - would controlling for z help us isolate the experience effects?
- answer: depends whether you adjust z for age.
  - if you use raw z, you are fine
  - the component of ability that predicts cohort gets residualized out by OLS
  - and, our D^* will be variation in experience that is orthogonal to cohort ability
- HOWEVER, if you age-adjust z to form z^*, you undo the benefit of having kids tested at the same age
  - you would then be regressing log wages on $\alpha z^* + D\gamma$
  - because z^* is orthogonal to D by construction, we know the estimated γ will be identical to the one from regressing log wages on D alone
    - and, as we discussed, that reflects both cohort and experience effects

Wages, Experience and IQ

- now, what if you had instead z, an IQ test administered at the same *age*?
  - would controlling for z help us isolate the experience effects?
- answer: depends whether you adjust z for age.
  - if you use raw z, you are fine
  - the component of ability that predicts cohort gets residualized out by OLS
  - and, our D^* will be variation in experience that is orthogonal to cohort ability
- HOWEVER, if you age-adjust z to form z^*, you undo the benefit of having kids tested at the same age
  - you would then be regressing log wages on $\alpha z^* + D\gamma$
  - because z^* is orthogonal to D by construction, we know the estimated γ will be identical to the one from regressing log wages on D alone
    - and, as we discussed, that reflects both cohort and experience effects

Wages, Experience and IQ

- now, what if you had instead z, an IQ test administered at the same *age*?
  - would controlling for z help us isolate the experience effects?
- answer: depends whether you adjust z for age.
  - if you use raw z, you are fine
  - the component of ability that predicts cohort gets residualized out by OLS
  - and, our D^* will be variation in experience that is orthogonal to cohort ability
- HOWEVER, if you age-adjust z to form z^*, you undo the benefit of having kids tested at the same age
  - you would then be regressing log wages on $\alpha z^* + D\gamma$
  - because z^* is orthogonal to D by construction, we know the estimated γ will be identical to the one from regressing log wages on D alone
    - and, as we discussed, that reflects both cohort and experience effects

Wages, Experience and IQ

- now, what if you had instead z, an IQ test administered at the same *age*?
  - would controlling for z help us isolate the experience effects?
- answer: depends whether you adjust z for age.
  - if you use raw z, you are fine
  - the component of ability that predicts cohort gets residualized out by OLS
  - and, our D^* will be variation in experience that is orthogonal to cohort ability
- HOWEVER, if you age-adjust z to form z^*, you undo the benefit of having kids tested at the same age
  - you would then be regressing log wages on $\alpha z^* + D\gamma$
  - because z^* is orthogonal to D by construction, we know the estimated γ will be identical to the one from regressing log wages on D alone
    - and, as we discussed, that reflects both cohort and experience effects

Wages, Experience and IQ

- now, what if you had instead z, an IQ test administered at the same *age*?
  - would controlling for z help us isolate the experience effects?
- answer: depends whether you adjust z for age.
  - if you use raw z, you are fine
  - the component of ability that predicts cohort gets residualized out by OLS
  - and, our D^* will be variation in experience that is orthogonal to cohort ability
- HOWEVER, if you age-adjust z to form z^*, you undo the benefit of having kids tested at the same age
  - you would then be regressing log wages on $\alpha z^* + D\gamma$
  - because z^* is orthogonal to D by construction, we know the estimated γ will be identical to the one from regressing log wages on D alone
    - and, as we discussed, that reflects both cohort and experience effects

Wages, Experience and IQ

- now, what if you had instead z, an IQ test administered at the same *age*?
  - would controlling for z help us isolate the experience effects?
- answer: depends whether you adjust z for age.
  - if you use raw z, you are fine
  - the component of ability that predicts cohort gets residualized out by OLS
  - and, our D^* will be variation in experience that is orthogonal to cohort ability
- HOWEVER, if you age-adjust z to form z^*, you undo the benefit of having kids tested at the same age
  - you would then be regressing log wages on $\alpha z^* + D\gamma$
  - because z^* is orthogonal to D by construction, we know the estimated γ will be identical to the one from regressing log wages on D alone
    - and, as we discussed, that reflects both cohort and experience effects

Residual Variation

- the bottom line here is that your choice of specification determines the residual variation used to estimate your coefficient of interest
- many specification choices, e.g.
    - fixed effects
    - measurement error
    - normalizations
    - omitted variables
- the key to thinking clearly about the costs and benefits of these choices is to think about how they change the residual variation "in the denominator"

Residual Variation

- the bottom line here is that your choice of specification determines the residual variation used to estimate your coefficient of interest
- many specification choices, e.g.
    - fixed effects
    - measurement error
    - normalizations
    - omitted variables
- the key to thinking clearly about the costs and benefits of these choices is to think about how they change the residual variation "in the denominator"

Residual Variation

- the bottom line here is that your choice of specification determines the residual variation used to estimate your coefficient of interest
- many specification choices, e.g.
    - fixed effects
    - measurement error
    - normalizations
    - omitted variables
- the key to thinking clearly about the costs and benefits of these choices is to think about how they change the residual variation "in the denominator"

Residual Variation

- the bottom line here is that your choice of specification determines the residual variation used to estimate your coefficient of interest
- many specification choices, e.g.
  - fixed effects
  - measurement error
  - normalizations
  - omitted variables
- the key to thinking clearly about the costs and benefits of these choices is to think about how they change the residual variation "in the denominator"

Residual Variation

- the bottom line here is that your choice of specification determines the residual variation used to estimate your coefficient of interest
- many specification choices, e.g.
  - fixed effects
  - measurement error
  - normalizations
  - omitted variables
- the key to thinking clearly about the costs and benefits of these choices is to think about how they change the residual variation "in the denominator"

Residual Variation

- the bottom line here is that your choice of specification determines the residual variation used to estimate your coefficient of interest
- many specification choices, e.g.
    - fixed effects
    - measurement error
    - normalizations
    - omitted variables
- the key to thinking clearly about the costs and benefits of these choices is to think about how they change the residual variation "in the denominator"

Residual Variation

- the bottom line here is that your choice of specification determines the residual variation used to estimate your coefficient of interest
- many specification choices, e.g.
  - fixed effects
  - measurement error
  - normalizations
  - omitted variables
- the key to thinking clearly about the costs and benefits of these choices is to think about how they change the residual variation “in the denominator”

Conditional Independence and Causal Effects

- When does OLS give us an estimate of the *causal* effect of D?
  - we have already seen one simple case: where D is assigned at random (as in an experiment)
  - in this case, you can compute the difference between treatment and controls by regressing Y on D
  - but, why would we "control" for X then?
    - and, what needs to be true about the joint distribution of (Y_1, Y_0, D, X) for us to recover a causal effect?
- the basic answer is that we need D to be independent of the potential outcomes, conditional on X
  - sometimes people call this the "conditional independence assumption" or CIA
  - you can interpret this to mean: once we have accounted for X, people choose D at random
  - this assumption should make you feel very uncomfortable!

Conditional Independence and Causal Effects

- When does OLS give us an estimate of the *causal* effect of D?
  - we have already seen one simple case: where D is assigned at random (as in an experiment)
  - in this case, you can compute the difference between treatment and controls by regressing Y on D
  - but, why would we "control" for X then?
    - and, what needs to be true about the joint distribution of (Y_1, Y_0, D, X) for us to recover a causal effect?
- the basic answer is that we need D to be independent of the potential outcomes, conditional on X
  - sometimes people call this the "conditional independence assumption" or CIA
  - you can interpret this to mean: once we have accounted for X, people choose D at random
  - this assumption should make you feel very uncomfortable!

Conditional Independence and Causal Effects

- When does OLS give us an estimate of the *causal* effect of D?
  - we have already seen one simple case: where D is assigned at random (as in an experiment)
  - in this case, you can compute the difference between treatment and controls by regressing Y on D
  - but, why would we "control" for X then?
    - and, what needs to be true about the joint distribution of (Y_1, Y_0, D, X) for us to recover a causal effect?
- the basic answer is that we need D to be independent of the potential outcomes, conditional on X
  - sometimes people call this the "conditional independence assumption" or CIA
  - you can interpret this to mean: once we have accounted for X, people choose D at random
  - this assumption should make you feel very uncomfortable!

Conditional Independence and Causal Effects

- When does OLS give us an estimate of the *causal* effect of D?
  - we have already seen one simple case: where D is assigned at random (as in an experiment)
  - in this case, you can compute the difference between treatment and controls by regressing Y on D
  - but, why would we “control” for X then?
    - and, what needs to be true about the joint distribution of (Y_1, Y_0, D, X) for us to recover a causal effect?
- the basic answer is that we need D to be independent of the potential outcomes, conditional on X
  - sometimes people call this the “conditional independence assumption” or CIA
  - you can interpret this to mean: once we have accounted for X, people choose D at random
  - this assumption should make you feel very uncomfortable!

Conditional Independence and Causal Effects

- When does OLS give us an estimate of the *causal* effect of D?
  - we have already seen one simple case: where D is assigned at random (as in an experiment)
  - in this case, you can compute the difference between treatment and controls by regressing Y on D
  - but, why would we "control" for X then?
    - and, what needs to be true about the joint distribution of (Y_1, Y_0, D, X) for us to recover a causal effect?
- the basic answer is that we need D to be independent of the potential outcomes, conditional on X
  - sometimes people call this the "conditional independence assumption" or CIA
  - you can interpret this to mean: once we have accounted for X, people choose D at random
  - this assumption should make you feel very uncomfortable!

Conditional Independence and Causal Effects

- When does OLS give us an estimate of the *causal* effect of D?
  - we have already seen one simple case: where D is assigned at random (as in an experiment)
  - in this case, you can compute the difference between treatment and controls by regressing Y on D
  - but, why would we "control" for X then?
    - and, what needs to be true about the joint distribution of (Y_1, Y_0, D, X) for us to recover a causal effect?
- the basic answer is that we need D to be independent of the potential outcomes, conditional on X
  - sometimes people call this the "conditional independence assumption" or CIA
  - you can interpret this to mean: once we have accounted for X, people choose D at random
  - this assumption should make you feel very uncomfortable!

Conditional Independence and Causal Effects

- When does OLS give us an estimate of the *causal* effect of D?
  - we have already seen one simple case: where D is assigned at random (as in an experiment)
  - in this case, you can compute the difference between treatment and controls by regressing Y on D
  - but, why would we "control" for X then?
    - and, what needs to be true about the joint distribution of (Y_1, Y_0, D, X) for us to recover a causal effect?
- the basic answer is that we need D to be independent of the potential outcomes, conditional on X
  - sometimes people call this the "conditional independence assumption" or CIA
  - you can interpret this to mean: once we have accounted for X, people choose D at random
  - this assumption should make you feel very uncomfortable!

Conditional Independence and Causal Effects

- When does OLS give us an estimate of the *causal* effect of D?
  - we have already seen one simple case: where D is assigned at random (as in an experiment)
  - in this case, you can compute the difference between treatment and controls by regressing Y on D
  - but, why would we "control" for X then?
    - and, what needs to be true about the joint distribution of (Y_1, Y_0, D, X) for us to recover a causal effect?
- the basic answer is that we need D to be independent of the potential outcomes, conditional on X
  - sometimes people call this the "conditional independence assumption" or CIA
  - you can interpret this to mean: once we have accounted for X, people choose D at random
  - this assumption should make you feel very uncomfortable!

Conditional Independence and Causal Effects

- When does OLS give us an estimate of the *causal* effect of D?
  - we have already seen one simple case: where D is assigned at random (as in an experiment)
  - in this case, you can compute the difference between treatment and controls by regressing Y on D
  - but, why would we "control" for X then?
    - and, what needs to be true about the joint distribution of (Y_1, Y_0, D, X) for us to recover a causal effect?
- the basic answer is that we need D to be independent of the potential outcomes, conditional on X
  - sometimes people call this the "conditional independence assumption" or CIA
  - you can interpret this to mean: once we have accounted for X, people choose D at random
  - this assumption should make you feel very uncomfortable!

Example of CIA

- suppose we are interested in the causal effect of education (D) on earnings (Y)
  - people live in different locations, encoded in a vector of location dummies X
  - now, under what conditions can we run a regression of Y on (D, X) and interpret the coefficient on D as causal?
- here is one set of sufficient conditions:
  - suppose $D = bX + u$ with $u \perp\!\!\!\perp X$
  - and, suppose $Y = \alpha + \beta D + \gamma X + \varepsilon$ with $\varepsilon \perp\!\!\!\perp u$
  - finally, assume $\text{cov}(X, \varepsilon) \neq 0$
- under these assumptions you can show that:
  - $\text{cov}(Y_0, D) = \gamma b V[X] + b\,\text{cov}(X, \varepsilon)$
    - so, D is not unconditionally independent of the potential outcomes

Example of CIA

- suppose we are interested in the causal effect of education (D) on earnings (Y)
    - people live in different locations, encoded in a vector of location dummies X
    - now, under what conditions can we run a regression of Y on (D, X) and interpret the coefficient on D as causal?
- here is one set of sufficient conditions:
    - suppose $D = bX + u$ with $u \perp\!\!\!\perp X$
    - and, suppose $Y = \alpha + \beta D + \gamma X + \varepsilon$ with $\varepsilon \perp\!\!\!\perp u$
    - finally, assume $\text{cov}(X, \varepsilon) \neq 0$
- under these assumptions you can show that:
    - $\text{cov}(Y_0, D) = \gamma b V[X] + b \, \text{cov}(X, \varepsilon)$
        - so, D is not unconditionally independent of the potential outcomes

Example of CIA

- suppose we are interested in the causal effect of education (D) on earnings (Y)
  - people live in different locations, encoded in a vector of location dummies X
  - now, under what conditions can we run a regression of Y on (D, X) and interpret the coefficient on D as causal?
- here is one set of sufficient conditions:
  - suppose $D = bX + u$ with $u \perp\!\!\!\perp X$
  - and, suppose $Y = \alpha + \beta D + \gamma X + \varepsilon$ with $\varepsilon \perp\!\!\!\perp u$
  - finally, assume $\text{cov}(X, \varepsilon) \neq 0$
- under these assumptions you can show that:
  - $\text{cov}(Y_0, D) = \gamma b V[X] + b \,\text{cov}(X, \varepsilon)$
    - so, D is not unconditionally independent of the potential outcomes

Example of CIA

- suppose we are interested in the causal effect of education (D) on earnings (Y)
  - people live in different locations, encoded in a vector of location dummies X
  - now, under what conditions can we run a regression of Y on (D, X) and interpret the coefficient on D as causal?
- here is one set of sufficient conditions:
  - suppose $D = bX + u$ with $u \perp\!\!\!\perp X$
  - and, suppose $Y = \alpha + \beta D + \gamma X + \varepsilon$ with $\varepsilon \perp\!\!\!\perp u$
  - finally, assume $\text{cov}(X, \varepsilon) \neq 0$
- under these assumptions you can show that:
  - $\text{cov}(Y_0, D) = \gamma b V[X] + b\,\text{cov}(X, \varepsilon)$
    - so, D is not unconditionally independent of the potential outcomes

Example of CIA

- suppose we are interested in the causal effect of education (D) on earnings (Y)
  - people live in different locations, encoded in a vector of location dummies X
  - now, under what conditions can we run a regression of Y on (D, X) and interpret the coefficient on D as causal?
- here is one set of sufficient conditions:
  - suppose $D = bX + u$ with $u \perp\!\!\!\perp X$
  - and, suppose $Y = \alpha + \beta D + \gamma X + \varepsilon$ with $\varepsilon \perp\!\!\!\perp u$
  - finally, assume $\text{cov}(X, \varepsilon) \neq 0$
- under these assumptions you can show that:
  - $\text{cov}(Y_0, D) = \gamma b V[X] + b\,\text{cov}(X, \varepsilon)$
    - so, D is not unconditionally independent of the potential outcomes

Example of CIA

- suppose we are interested in the causal effect of education (D) on earnings (Y)
    - people live in different locations, encoded in a vector of location dummies X
    - now, under what conditions can we run a regression of Y on (D, X) and interpret the coefficient on D as causal?
- here is one set of sufficient conditions:
    - suppose $D = bX + u$ with $u \perp\!\!\!\perp X$
    - and, suppose $Y = \alpha + \beta D + \gamma X + \varepsilon$ with $\varepsilon \perp\!\!\!\perp u$
    - finally, assume $\text{cov}(X, \varepsilon) \neq 0$
- under these assumptions you can show that:
    - $\text{cov}(Y_0, D) = \gamma b V[X] + b\,\text{cov}(X, \varepsilon)$
        - so, D is not unconditionally independent of the potential outcomes

Example of CIA

- suppose we are interested in the causal effect of education (D) on earnings (Y)
  - people live in different locations, encoded in a vector of location dummies X
  - now, under what conditions can we run a regression of Y on (D, X) and interpret the coefficient on D as causal?
- here is one set of sufficient conditions:
  - suppose $D = bX + u$ with $u \perp\!\!\!\perp X$
  - and, suppose $Y = \alpha + \beta D + \gamma X + \varepsilon$ with $\varepsilon \perp\!\!\!\perp u$
  - finally, assume $\text{cov}(X, \varepsilon) \neq 0$
- under these assumptions you can show that:
  - $\text{cov}(Y_0, D) = \gamma b V[X] + b\,\text{cov}(X, \varepsilon)$
    - so, D is not unconditionally independent of the potential outcomes

Example of CIA

- suppose we are interested in the causal effect of education (D) on earnings (Y)
  - people live in different locations, encoded in a vector of location dummies X
  - now, under what conditions can we run a regression of Y on (D, X) and interpret the coefficient on D as causal?
- here is one set of sufficient conditions:
  - suppose $D = bX + u$ with $u \perp\!\!\!\perp X$
  - and, suppose $Y = \alpha + \beta D + \gamma X + \varepsilon$ with $\varepsilon \perp\!\!\!\perp u$
  - finally, assume $\text{cov}(X, \varepsilon) \neq 0$
- under these assumptions you can show that:
  - $\text{cov}(Y_0, D) = \gamma b V[X] + b\,\text{cov}(X, \varepsilon)$
    - so, D is not unconditionally independent of the potential outcomes

Example of CIA

- suppose we are interested in the causal effect of education (D) on earnings (Y)
  - people live in different locations, encoded in a vector of location dummies X
  - now, under what conditions can we run a regression of Y on (D, X) and interpret the coefficient on D as causal?
- here is one set of sufficient conditions:
  - suppose $D = bX + u$ with $u \perp\!\!\!\perp X$
  - and, suppose $Y = \alpha + \beta D + \gamma X + \varepsilon$ with $\varepsilon \perp\!\!\!\perp u$
  - finally, assume $\text{cov}(X, \varepsilon) \neq 0$
- under these assumptions you can show that:
  - $\text{cov}(Y_0, D) = \gamma b V[X] + b\,\text{cov}(X, \varepsilon)$
    - so, D is not unconditionally independent of the potential outcomes

Example of CIA

- suppose we are interested in the causal effect of education (D) on earnings (Y)
  - people live in different locations, encoded in a vector of location dummies X
  - now, under what conditions can we run a regression of Y on (D, X) and interpret the coefficient on D as causal?
- here is one set of sufficient conditions:
  - suppose $D = bX + u$ with $u \perp\!\!\!\perp X$
  - and, suppose $Y = \alpha + \beta D + \gamma X + \varepsilon$ with $\varepsilon \perp\!\!\!\perp u$
  - finally, assume $\text{cov}(X, \varepsilon) \neq 0$
- under these assumptions you can show that:
  - $\text{cov}(Y_0, D) = \gamma b V[X] + b\,\text{cov}(X, \varepsilon)$
    - so, D is not unconditionally independent of the potential outcomes

Example of CIA

- if you were to run OLS of Y on D you'd get

$$\text{plim}\,\widehat{\beta} = \beta + \gamma b \frac{V[D]}{V[X]} + b \frac{\text{cov}(X, \varepsilon)}{V[D]} \neq \beta$$

- however, the residual variation in D once X has been predicted out (in this case, u) *is* independent of (Y_1, Y_0)
  - thus, controlling for X is *necessary* to obtain the causal effect of D
- now, does the coefficient on X have a causal interpretation?
  - first note that by FWL, the coefficient on X is the same whether we include D^* or not
  - you can show that

$$\frac{\text{cov}(X, Y)}{V[X]} = \beta b + \gamma + \frac{\text{cov}(X, \varepsilon)}{V[X]} \neq \gamma$$

  - beyond the fact that X has an indirect effect via D (the βb term), X is correlated with the potential outcomes because $\text{cov}(X, \varepsilon) \neq 0$
    - thus, we cannot interpret the coefficient on X as causal
    - even though we need to control for it in order to estimate β consistently

Example of CIA

- if you were to run OLS of Y on D you'd get

$$\text{plim}\,\widehat{\beta} = \beta + \gamma b \frac{V[D]}{V[X]} + b\frac{\text{cov}(X,\varepsilon)}{V[D]} \neq \beta$$

- however, the residual variation in D once X has been predicted out (in this case, u) *is* independent of (Y_1, Y_0)
    - thus, controlling for X is *necessary* to obtain the causal effect of D
- now, does the coefficient on X have a causal interpretation?
    - first note that by FWL, the coefficient on X is the same whether we include D^* or not
    - you can show that

$$\frac{\text{cov}(X,Y)}{V[X]} = \beta b + \gamma + \frac{\text{cov}(X,\varepsilon)}{V[X]} \neq \gamma$$

    - beyond the fact that X has an indirect effect via D (the βb term), X is correlated with the potential outcomes because $\text{cov}(X,\varepsilon) \neq 0$

Example of CIA

- if you were to run OLS of Y on D you'd get

$$\operatorname{plim} \widehat{\beta} = \beta + \gamma b \frac{V[D]}{V[X]} + b \frac{\operatorname{cov}(X, \varepsilon)}{V[D]} \neq \beta$$

- however, the residual variation in D once X has been predicted out (in this case, u) *is* independent of (Y_1, Y_0)
  - thus, controlling for X is *necessary* to obtain the causal effect of D
- now, does the coefficient on X have a causal interpretation?
  - first note that by FWL, the coefficient on X is the same whether we include D^* or not
  - you can show that

$$\frac{\operatorname{cov}(X, Y)}{V[X]} = \beta b + \gamma + \frac{\operatorname{cov}(X, \varepsilon)}{V[X]} \neq \gamma$$

  - beyond the fact that X has an indirect effect via D (the βb term), X is correlated with the potential outcomes because $\operatorname{cov}(X, \varepsilon) \neq 0$
    - thus, we cannot interpret the coefficient on X as causal
    - even though we need to control for it in order to estimate β consistently

Example of CIA

- if you were to run OLS of Y on D you'd get

$$\text{plim}\,\widehat{\beta} = \beta + \gamma b \frac{V[D]}{V[X]} + b \frac{\text{cov}(X, \varepsilon)}{V[D]} \neq \beta$$

- however, the residual variation in D once X has been predicted out (in this case, u) *is* independent of (Y_1, Y_0)
  - thus, controlling for X is *necessary* to obtain the causal effect of D
- now, does the coefficient on X have a causal interpretation?
  - first note that by FWL, the coefficient on X is the same whether we include D^* or not
  - you can show that

$$\frac{\text{cov}(X, Y)}{V[X]} = \beta b + \gamma + \frac{\text{cov}(X, \varepsilon)}{V[X]} \neq \gamma$$

  - beyond the fact that X has an indirect effect via D (the βb term), X is correlated with the potential outcomes because $\text{cov}(X, \varepsilon) \neq 0$
    - thus, we cannot interpret the coefficient on X as causal
    - even though we need to control for it in order to estimate β consistently

Example of CIA

- if you were to run OLS of Y on D you'd get

$$\text{plim}\,\widehat{\beta} = \beta + \gamma b \frac{V[D]}{V[X]} + b \frac{\text{cov}(X, \varepsilon)}{V[D]} \neq \beta$$

- however, the residual variation in D once X has been predicted out (in this case, u) *is* independent of (Y_1, Y_0)
  - thus, controlling for X is *necessary* to obtain the causal effect of D
- now, does the coefficient on X have a causal interpretation?
  - first note that by FWL, the coefficient on X is the same whether we include D^* or not
  - you can show that

$$\frac{\text{cov}(X, Y)}{V[X]} = \beta b + \gamma + \frac{\text{cov}(X, \varepsilon)}{V[X]} \neq \gamma$$

  - beyond the fact that X has an indirect effect via D (the βb term), X is correlated with the potential outcomes because $\text{cov}(X, \varepsilon) \neq 0$
    - thus, we cannot interpret the coefficient on X as causal
    - even though we need to control for it in order to estimate β consistently

Example of CIA

- if you were to run OLS of Y on D you'd get

$$\text{plim}\,\widehat{\beta} = \beta + \gamma b \frac{V[D]}{V[X]} + b \frac{\text{cov}(X, \varepsilon)}{V[D]} \neq \beta$$

- however, the residual variation in D once X has been predicted out (in this case, u) *is* independent of (Y_1, Y_0)
  - thus, controlling for X is *necessary* to obtain the causal effect of D
- now, does the coefficient on X have a causal interpretation?
  - first note that by FWL, the coefficient on X is the same whether we include D^* or not
  - you can show that

$$\frac{\text{cov}(X, Y)}{V[X]} = \beta b + \gamma + \frac{\text{cov}(X, \varepsilon)}{V[X]} \neq \gamma$$

  - beyond the fact that X has an indirect effect via D (the βb term), X is correlated with the potential outcomes because $\text{cov}(X, \varepsilon) \neq 0$
    - thus, we cannot interpret the coefficient on X as causal
    - even though we need to control for it in order to estimate β consistently

Example of CIA

- if you were to run OLS of Y on D you'd get

$$\text{plim}\,\widehat{\beta} = \beta + \gamma b \frac{V[D]}{V[X]} + b\frac{\text{cov}(X,\varepsilon)}{V[D]} \neq \beta$$

- however, the residual variation in D once X has been predicted out (in this case, u) *is* independent of (Y_1, Y_0)
  - thus, controlling for X is *necessary* to obtain the causal effect of D
- now, does the coefficient on X have a causal interpretation?
  - first note that by FWL, the coefficient on X is the same whether we include D^* or not
  - you can show that

$$\frac{\text{cov}(X,Y)}{V[X]} = \beta b + \gamma + \frac{\text{cov}(X,\varepsilon)}{V[X]} \neq \gamma$$

  - beyond the fact that X has an indirect effect via D (the βb term), X is correlated with the potential outcomes because $\text{cov}(X,\varepsilon) \neq 0$
    - thus, we cannot interpret the coefficient on X as causal
    - even though we need to control for it in order to estimate β consistently

Example of CIA

- if you were to run OLS of Y on D you'd get

$$\text{plim}\,\widehat{\beta} = \beta + \gamma b \frac{V[D]}{V[X]} + b \frac{\text{cov}(X,\varepsilon)}{V[D]} \neq \beta$$

- however, the residual variation in D once X has been predicted out (in this case, u) *is* independent of (Y_1, Y_0)
    - thus, controlling for X is *necessary* to obtain the causal effect of D
- now, does the coefficient on X have a causal interpretation?
    - first note that by FWL, the coefficient on X is the same whether we include D^* or not
    - you can show that

$$\frac{\text{cov}(X,Y)}{V[X]} = \beta b + \gamma + \frac{\text{cov}(X,\varepsilon)}{V[X]} \neq \gamma$$

    - beyond the fact that X has an indirect effect via D (the βb term), X is correlated with the potential outcomes because $\text{cov}(X,\varepsilon) \neq 0$
        - thus, we cannot interpret the coefficient on X as causal
        - even though we need to control for it in order to estimate β consistently

Example of CIA

- if you were to run OLS of Y on D you'd get

$$\text{plim}\,\widehat{\beta} = \beta + \gamma b \frac{V[D]}{V[X]} + b\frac{\text{cov}(X,\varepsilon)}{V[D]} \neq \beta$$

- however, the residual variation in D once X has been predicted out (in this case, u) *is* independent of (Y_1, Y_0)
  - thus, controlling for X is *necessary* to obtain the causal effect of D
- now, does the coefficient on X have a causal interpretation?
  - first note that by FWL, the coefficient on X is the same whether we include D^* or not
  - you can show that

$$\frac{\text{cov}(X,Y)}{V[X]} = \beta b + \gamma + \frac{\text{cov}(X,\varepsilon)}{V[X]} \neq \gamma$$

  - beyond the fact that X has an indirect effect via D (the βb term), X is correlated with the potential outcomes because $\text{cov}(X,\varepsilon) \neq 0$
    - thus, we cannot interpret the coefficient on X as causal
    - even though we need to control for it in order to estimate β consistently

Example of CIA

- next, we will work through this model of the "data generating process" and ask whether its assumptions make sense in context
- one interpretation is:
  - distribution of education differs across locations for reasons to do with
  - further, earnings differ across people because of
- does it make sense to assume $u \perp\!\!\!\perp \varepsilon$? What does this assumption mean?
  - mechanically, it gives us that $D|X$ is independent of potential outcomes
  - in context, it means location is the *only* thing that affects both earnings and education

Example of CIA

- next, we will work through this model of the "data generating process" and ask whether its assumptions make sense in context
- one interpretation is:
    - distribution of education differs across locations for reasons to do with
        - the characteristics of the location itself (bX, e.g. local education policies)
        - and person-specific factors (u, e.g. ability) which are *independent* of location
        - why might $X \perp\!\!\!\perp u$ fail?
    - further, earnings differ across people because of
        - direct effects of location-specific factors γX, e.g. differences in labor demand
        - education itself (βD)
        - individual-specific factors ε (e.g. individual labor supply preferences, family obligations)
- does it make sense to assume $u \perp\!\!\!\perp \varepsilon$? What does this assumption mean?
    - mechanically, it gives us that $D|X$ is independent of potential outcomes
    - in context, it means location is the *only* thing that affects both earnings and education
        - once we have accounted for it, education is not affected by (e.g.) earnings potential or ability

Example of CIA

- next, we will work through this model of the "data generating process" and ask whether its assumptions make sense in context
- one interpretation is:
  - distribution of education differs across locations for reasons to do with
    - the characteristics of the location itself (bX, e.g. local education policies)
    - and person-specific factors (u, e.g. ability) which are *independent* of location
    - why might $X \perp\!\!\!\perp u$ fail?
  - further, earnings differ across people because of
    - direct effects of location-specific factors γX, e.g. differences in labor demand
    - education itself (βD)
    - individual-specific factors ε (e.g. individual labor supply preferences, family obligations)
- does it make sense to assume $u \perp\!\!\!\perp \varepsilon$? What does this assumption mean?
  - mechanically, it gives us that $D|X$ is independent of potential outcomes
  - in context, it means location is the *only* thing that affects both earnings and education
    - once we have accounted for it, education is not affected by (e.g.) earnings potential or ability

Example of CIA

- next, we will work through this model of the "data generating process" and ask whether its assumptions make sense in context
- one interpretation is:
  - distribution of education differs across locations for reasons to do with
    - the characteristics of the location itself (bX, e.g. local education policies)
    - and person-specific factors (u, e.g. ability) which are *independent* of location
    - why might $X \perp\!\!\!\perp u$ fail?
  - further, earnings differ across people because of
    - direct effects of location-specific factors γX, e.g. differences in labor demand
    - education itself (βD)
    - individual-specific factors ε (e.g. individual labor supply preferences, family obligations)
- does it make sense to assume $u \perp\!\!\!\perp \varepsilon$? What does this assumption mean?
  - mechanically, it gives us that $D|X$ is independent of potential outcomes
  - in context, it means location is the *only* thing that affects both earnings and education
    - once we have accounted for X, education is not self-selected by earnings potential or ability

Example of CIA

- next, we will work through this model of the "data generating process" and ask whether its assumptions make sense in context
- one interpretation is:
  - distribution of education differs across locations for reasons to do with
    - the characteristics of the location itself (bX, e.g. local education policies)
    - and person-specific factors (u, e.g. ability) which are *independent* of location
    - why might $X \perp\!\!\!\perp u$ fail?
  - further, earnings differ across people because of
    - direct effects of location-specific factors γX, e.g. differences in labor demand
    - education itself (βD)
    - individual-specific factors ε (e.g. individual labor supply preferences, family obligations)
- does it make sense to assume $u \perp\!\!\!\perp \varepsilon$? What does this assumption mean?
  - mechanically, it gives us that $D|X$ is independent of potential outcomes
  - in context, it means location is the *only* thing that affects both earnings and education
    - once we have accounted for it, education is not affected by (e.g.) earnings potential or ability

Example of CIA

- next, we will work through this model of the "data generating process" and ask whether its assumptions make sense in context
- one interpretation is:
  - distribution of education differs across locations for reasons to do with
    - the characteristics of the location itself (bX, e.g. local education policies)
    - and person-specific factors (u, e.g. ability) which are *independent* of location
    - why might $X \perp\!\!\!\perp u$ fail?
  - further, earnings differ across people because of
    - direct effects of location-specific factors γX, e.g. differences in labor demand
    - education itself (βD)
    - individual-specific factors ε (e.g. individual labor supply preferences, family obligations)
- does it make sense to assume $u \perp\!\!\!\perp \varepsilon$? What does this assumption mean?
  - mechanically, it gives us that $D|X$ is independent of potential outcomes
  - in context, it means location is the *only* thing that affects both earnings and education
    - once we have accounted for it, education is not affected by (e.g.) earnings potential or ability

Example of CIA

- next, we will work through this model of the "data generating process" and ask whether its assumptions make sense in context
- one interpretation is:
  - distribution of education differs across locations for reasons to do with
    - the characteristics of the location itself (bX, e.g. local education policies)
    - and person-specific factors (u, e.g. ability) which are *independent* of location
    - why might $X \perp\!\!\!\perp u$ fail?
  - further, earnings differ across people because of
    - direct effects of location-specific factors γX, e.g. differences in labor demand
    - education itself (βD)
    - individual-specific factors ε (e.g. individual labor supply preferences, family obligations)
- does it make sense to assume $u \perp\!\!\!\perp \varepsilon$? What does this assumption mean?
  - mechanically, it gives us that $D|X$ is independent of potential outcomes
  - in context, it means location is the *only* thing that affects both earnings and education
    - once we have accounted for it, education is not affected by (e.g.) earnings potential or ability

Example of CIA

- next, we will work through this model of the "data generating process" and ask whether its assumptions make sense in context
- one interpretation is:
  - distribution of education differs across locations for reasons to do with
    - the characteristics of the location itself (bX, e.g. local education policies)
    - and person-specific factors (u, e.g. ability) which are *independent* of location
    - why might $X \perp\!\!\!\perp u$ fail?
  - further, earnings differ across people because of
    - direct effects of location-specific factors γX, e.g. differences in labor demand
    - education itself (βD)
    - individual-specific factors ε (e.g. individual labor supply preferences, family obligations)
- does it make sense to assume $u \perp\!\!\!\perp \varepsilon$? What does this assumption mean?
  - mechanically, it gives us that $D|X$ is independent of potential outcomes
  - in context, it means location is the *only* thing that affects both earnings and education

Example of CIA

- next, we will work through this model of the "data generating process" and ask whether its assumptions make sense in context
- one interpretation is:
    - distribution of education differs across locations for reasons to do with
        - the characteristics of the location itself (bX, e.g. local education policies)
        - and person-specific factors (u, e.g. ability) which are *independent* of location
        - why might $X \perp\!\!\!\perp u$ fail?
    - further, earnings differ across people because of
        - direct effects of location-specific factors γX, e.g. differences in labor demand
        - education itself (βD)
        - individual-specific factors ε (e.g. individual labor supply preferences, family obligations)
- does it make sense to assume $u \perp\!\!\!\perp \varepsilon$? What does this assumption mean?
    - mechanically, it gives us that $D|X$ is independent of potential outcomes
    - in context, it means location is the *only* thing that affects both earnings and education
        - once we have accounted for it, education is not affected by (e.g.) earnings potential or ability

Example of CIA

- next, we will work through this model of the "data generating process" and ask whether its assumptions make sense in context
- one interpretation is:
  - distribution of education differs across locations for reasons to do with
    - the characteristics of the location itself (bX, e.g. local education policies)
    - and person-specific factors (u, e.g. ability) which are *independent* of location
    - why might $X \perp\!\!\!\perp u$ fail?
  - further, earnings differ across people because of
    - direct effects of location-specific factors γX, e.g. differences in labor demand
    - education itself (βD)
    - individual-specific factors ε (e.g. individual labor supply preferences, family obligations)
- does it make sense to assume $u \perp\!\!\!\perp \varepsilon$? What does this assumption mean?
  - mechanically, it gives us that $D|X$ is independent of potential outcomes
  - in context, it means location is the *only* thing that affects both earnings and education
    - once we have accounted for it, education is not affected by (e.g.) earnings potential or ability

Example of CIA

- next, we will work through this model of the "data generating process" and ask whether its assumptions make sense in context
- one interpretation is:
  - distribution of education differs across locations for reasons to do with
    - the characteristics of the location itself (bX, e.g. local education policies)
    - and person-specific factors (u, e.g. ability) which are *independent* of location
    - why might $X \perp\!\!\!\perp u$ fail?
  - further, earnings differ across people because of
    - direct effects of location-specific factors γX, e.g. differences in labor demand
    - education itself (βD)
    - individual-specific factors ε (e.g. individual labor supply preferences, family obligations)
- does it make sense to assume $u \perp\!\!\!\perp \varepsilon$? What does this assumption mean?
  - mechanically, it gives us that $D|X$ is independent of potential outcomes
  - in context, it means location is the *only* thing that affects both earnings and education
    - once we have accounted for it, education is not affected by (e.g.) earnings potential or ability

Example of CIA

- next, we will work through this model of the "data generating process" and ask whether its assumptions make sense in context
- one interpretation is:
  - distribution of education differs across locations for reasons to do with
    - the characteristics of the location itself (bX, e.g. local education policies)
    - and person-specific factors (u, e.g. ability) which are *independent* of location
    - why might $X \perp\!\!\!\perp u$ fail?
  - further, earnings differ across people because of
    - direct effects of location-specific factors γX, e.g. differences in labor demand
    - education itself (βD)
    - individual-specific factors ε (e.g. individual labor supply preferences, family obligations)
- does it make sense to assume $u \perp\!\!\!\perp \varepsilon$? What does this assumption mean?
  - mechanically, it gives us that $D|X$ is independent of potential outcomes
  - in context, it means location is the *only* thing that affects both earnings and education
    - once we have accounted for it, education is not affected by (e.g.) earnings potential or ability

Example of CIA

- next, we will work through this model of the "data generating process" and ask whether its assumptions make sense in context
- one interpretation is:
  - distribution of education differs across locations for reasons to do with
    - the characteristics of the location itself (bX, e.g. local education policies)
    - and person-specific factors (u, e.g. ability) which are *independent* of location
    - why might $X \perp\!\!\!\perp u$ fail?
  - further, earnings differ across people because of
    - direct effects of location-specific factors γX, e.g. differences in labor demand
    - education itself (βD)
    - individual-specific factors ε (e.g. individual labor supply preferences, family obligations)
- does it make sense to assume $u \perp\!\!\!\perp \varepsilon$? What does this assumption mean?
  - mechanically, it gives us that $D|X$ is independent of potential outcomes
  - in context, it means location is the *only* thing that affects both earnings and education
    - once we have accounted for it, education is not affected by (e.g.) earnings potential or ability

Example of CIA

- next, we will work through this model of the "data generating process" and ask whether its assumptions make sense in context
- one interpretation is:
    - distribution of education differs across locations for reasons to do with
        - the characteristics of the location itself (bX, e.g. local education policies)
        - and person-specific factors (u, e.g. ability) which are *independent* of location
        - why might $X \perp\!\!\!\perp u$ fail?
    - further, earnings differ across people because of
        - direct effects of location-specific factors γX, e.g. differences in labor demand
        - education itself (βD)
        - individual-specific factors ε (e.g. individual labor supply preferences, family obligations)
- does it make sense to assume $u \perp\!\!\!\perp \varepsilon$? What does this assumption mean?
    - mechanically, it gives us that $D|X$ is independent of potential outcomes
    - in context, it means location is the *only* thing that affects both earnings and education
        - once we have accounted for it, education is not affected by (e.g.) earnings potential or ability

References

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

Table of Contents

Linear Projections

Derivation of OLS Formula

Frisch-Waugh-Lovell Theorem

Examples

Conditional Independence and Causal Effects