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

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

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

- and let $X_1^* = X_1 - (a + bX_2)$
- what is
  - $\mathrm{cov}(X_1^*, X_2)$?
  - $E[X_1^*]$
- 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

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 $(Y_i, X_{i1}, \dots X_{iK})$, consider

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

  - 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

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

- 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

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

- 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

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

- 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

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

- 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

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

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

- 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, $M_1X_1 = 0$
  - 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, $M_1 X_1 = 0$
  - 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, $M_1 X_1 = 0$
  - 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, $M_1X_1 = 0$
  - 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 effect

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"

Diff-in-Diff Mechanics

- suppose now we have $Y_{it} = \alpha + \beta D_i + \gamma T_t + \theta D_i \times T_t + e$
  - D_i is an indicator for a treatment group that starts off untreated
    - the "control" group with $D_i = 0$ never gets treated
  - T_t is a dummy for the period after treatment is applied
  - what would the data matrix look like?
- the OLS orthogonality conditions tell us that

$$\sum_{D_i=1} \hat{e}_i = 0$$

and similarly for

Diff-in-Diff Mechanics

- suppose now we have $Y_{it} = \alpha + \beta D_i + \gamma T_t + \theta D_i \times T_t + e$
  - D_i is an indicator for a treatment group that starts off untreated
    - the "control" group with $D_i = 0$ never gets treated
  - T_t is a dummy for the period after treatment is applied
  - what would the data matrix look like?
- the OLS orthogonality conditions tell us that

$$\sum_{D_i=1} \hat{e}_i = 0$$

and similarly for

Diff-in-Diff Mechanics

- suppose now we have $Y_{it} = \alpha + \beta D_i + \gamma T_t + \theta D_i \times T_t + e$
  - D_i is an indicator for a treatment group that starts off untreated
    - the "control" group with $D_i = 0$ never gets treated
  - T_t is a dummy for the period after treatment is applied
  - what would the data matrix look like?
- the OLS orthogonality conditions tell us that

$$\sum_{D_i=1} \hat{e}_i = 0$$

and similarly for

Diff-in-Diff Mechanics

- suppose now we have $Y_{it} = \alpha + \beta D_i + \gamma T_t + \theta D_i \times T_t + e$
  - D_i is an indicator for a treatment group that starts off untreated
    - the "control" group with $D_i = 0$ never gets treated
  - T_t is a dummy for the period after treatment is applied
  - what would the data matrix look like?
- the OLS orthogonality conditions tell us that

$$\sum_{D_i=1} \hat{e}_i = 0$$

and similarly for

Diff-in-Diff Mechanics

- suppose now we have $Y_{it} = \alpha + \beta D_i + \gamma T_t + \theta D_i \times T_t + e$
  - D_i is an indicator for a treatment group that starts off untreated
    - the "control" group with $D_i = 0$ never gets treated
  - T_t is a dummy for the period after treatment is applied
  - what would the data matrix look like?
- the OLS orthogonality conditions tell us that

$$\sum_{D_i=1} \hat{e}_i = 0$$

and similarly for

Diff-in-Diff Mechanics

- suppose now we have $Y_{it} = \alpha + \beta D_i + \gamma T_t + \theta D_i \times T_t + e$
  - D_i is an indicator for a treatment group that starts off untreated
    - the "control" group with $D_i = 0$ never gets treated
  - T_t is a dummy for the period after treatment is applied
  - what would the data matrix look like?
- the OLS orthogonality conditions tell us that

$$\sum_{D_i=1} \hat{e}_i = 0$$

and similarly for

Conditional Independence and Causal Effects

References

Table of Contents