Chapter 5: Multivariate Regression

Jonathan Roth

Mathematical Econometrics I
Brown University
Fall 2023

Outline

1. Deriving Multivariate Regression and OLS
2. Regression and Causality
3. Regression Odds and Ends

Moving Beyond One "Regressor"

- So far we've talked about regression as a way of approximating the CEF $E[Y_i|X_i = x] \approx \alpha + x\beta$ for a single scalar X_i
  - We then showed how the estimand (α,β) can be estimated by OLS

Moving Beyond One "Regressor"

- So far we've talked about regression as a way of approximating the CEF $E[Y_i|X_i = x] \approx \alpha + x\beta$ for a single scalar X_i
  - We then showed how the estimand (α, β) can be estimated by OLS

- Next we'll see how this can be generalized to approximate/estimate $E[Y_i|\mathbf{X}_i = \mathbf{x}] \approx \mathbf{x}'\boldsymbol{\beta}$ for a vector $\mathbf{X_i} = (1, X_{i1}, ..., X_{iK})'$
  - Note: As usual, I'll be putting vectors/matrices in bold type-face

Moving Beyond One "Regressor"

- So far we've talked about regression as a way of approximating the CEF $E[Y_i|X_i = x] \approx \alpha + x\beta$ for a single scalar X_i
  - We then showed how the estimand (α, β) can be estimated by OLS
- Next we'll see how this can be generalized to approximate/estimate $E[Y_i|\mathbf{X}_i = \mathbf{x}] \approx \mathbf{x}'\boldsymbol{\beta}$ for a vector $\mathbf{X_i} = (1, X_{i1}, ..., X_{iK})'$
  - Note: As usual, I'll be putting vectors/matrices in bold type-face
- Two main motivations for this:

Moving Beyond One "Regressor"

- So far we've talked about regression as a way of approximating the CEF $E[Y_i|X_i = x] \approx \alpha + x\beta$ for a single scalar X_i
  - We then showed how the estimand (α,β) can be estimated by OLS
- Next we'll see how this can be generalized to approximate/estimate $E[Y_i|\mathbf{X}_i = \mathbf{x}] \approx \mathbf{x}'\boldsymbol{\beta}$ for a vector $\mathbf{X_i} = (1, X_{i1}, ..., X_{iK})'$
  - Note: As usual, I'll be putting vectors/matrices in bold type-face
- Two main motivations for this:

1. We want to use regression to identify causal effects, but conditional unconfoundedness is only plausible with multiple controls
   - In the Brown/URI example, we may want to control for high school GPA, family income, SAT, race ...

Moving Beyond One "Regressor"

- So far we've talked about regression as a way of approximating the CEF $E[Y_i|X_i = x] \approx \alpha + x\beta$ for a single scalar X_i
  - We then showed how the estimand (α, β) can be estimated by OLS

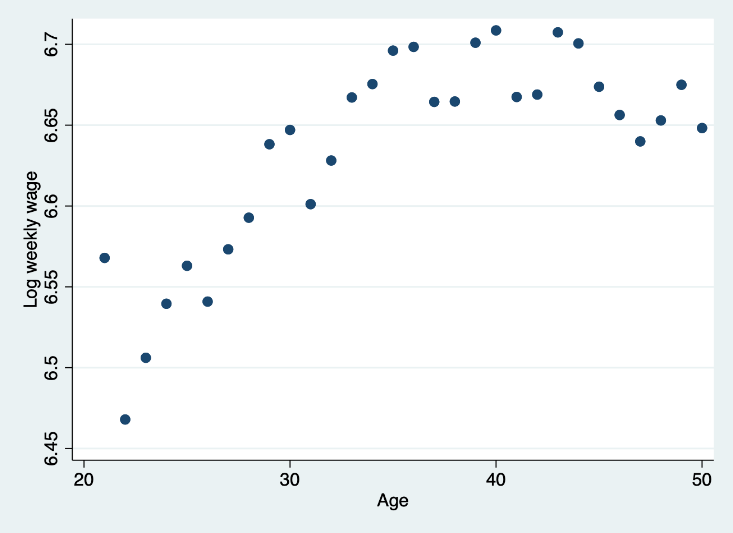
- Next we'll see how this can be generalized to approximate/estimate $E[Y_i|\mathbf{X}_i = \mathbf{x}] \approx \mathbf{x}'\boldsymbol{\beta}$ for a vector $\mathbf{X_i} = (1, X_{i1}, ..., X_{iK})'$
  - Note: As usual, I'll be putting vectors/matrices in bold type-face

- Two main motivations for this:

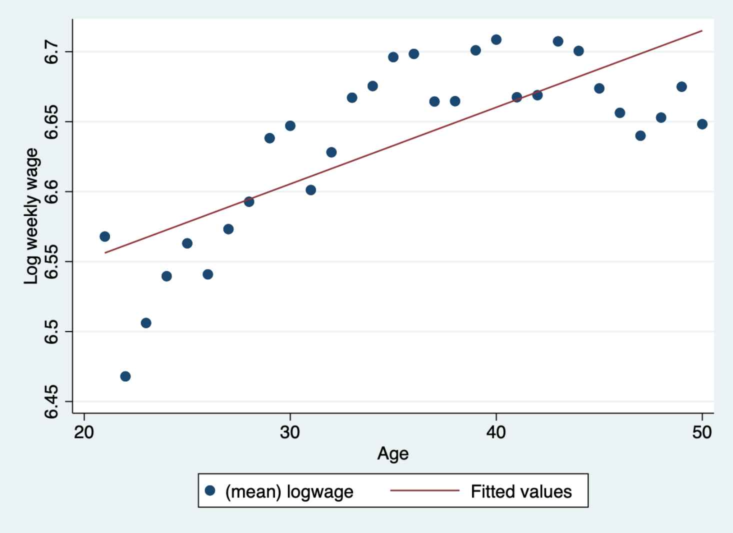
1. We want to use regression to identify causal effects, but conditional unconfoundedness is only plausible with multiple controls
   - In the Brown/URI example, we may want to control for high school GPA, family income, SAT, race ...

2. We want a *nonlinear* CEF approx.: e.g. $E[Y_i \mid X_i] \approx \alpha + X_i\beta + X_i^2\gamma$
   - We can "trick" regression into doing this by setting $\mathbf{X_i} = (1, X_i, X_i^2)'$

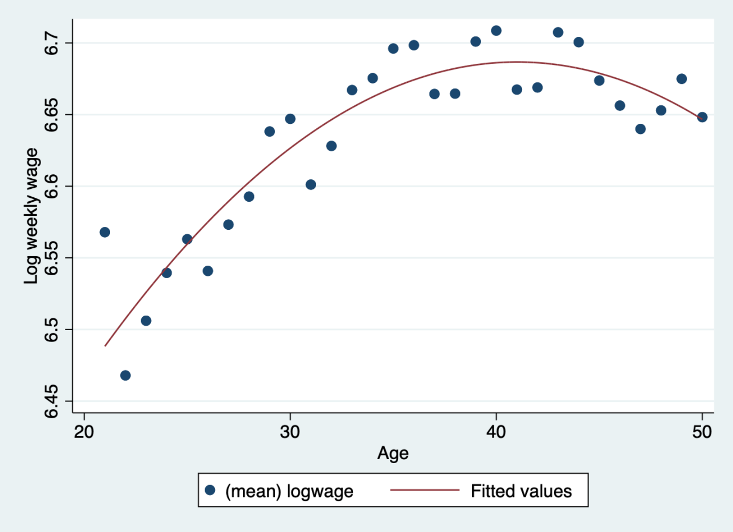
Log Wages by Age



OLS Regression (Linear Fit)



OLS Regression (Quadratic Fit)



Multivariate Regression as a Least-Squares Problem

- Recall with univariate OLS we solved for

$$(\alpha, \beta) = \arg\min_{a,b} E[(Y_i - (a + bX_i))^2]$$

We showed if the CEF is linear, then $E[Y|X] = \alpha + \beta X$; while if not, $\alpha + \beta X$ gave the best non-linear approximation

Multivariate Regression as a Least-Squares Problem

- Recall with univariate OLS we solved for

$$(\alpha, \beta) = \arg\min_{a,b} E[(Y_i - (a + bX_i))^2]$$

We showed if the CEF is linear, then $E[Y|X] = \alpha + \beta X$; while if not, $\alpha + \beta X$ gave the best non-linear approximation

- We will now consider the multi-variate analog:

$$\boldsymbol{\beta} = \arg\min_{\mathbf{b}} E\left[(Y_i - \mathbf{X}_i'\boldsymbol{b})^2\right]$$

- Using similar steps for the univariate case, we can show that if the CEF is linear in $\mathbf{X}$, then $E[Y|\mathbf{X}] = \mathbf{X}'\boldsymbol{\beta}$; if not, then $\mathbf{X}'\boldsymbol{\beta}$ is the MSE-minimizing approximation to the CEF.

Solving for Regression Coefficients

- So the *population regression coefficient* $\boldsymbol{\beta}$ solves least squares problem

$$\boldsymbol{\beta} = \arg\min_{\mathbf{b}} E[(Y_i - \mathbf{X}_i'\mathbf{b})^2]$$

Solving for Regression Coefficients

- So the *population regression coefficient* $\boldsymbol{\beta}$ solves least squares problem

$$\boldsymbol{\beta} = \arg\min_{\mathbf{b}} E[(Y_i - \mathbf{X}_i'\mathbf{b})^2]$$

- To solve for $\boldsymbol{\beta}$, we take the derivative (i.e. gradient) and set it to zero

$$\frac{d}{d\boldsymbol{\beta}} E[(Y_i - \mathbf{X}_i'\boldsymbol{\beta})^2] = \mathbf{0}$$

Solving for Regression Coefficients

- So the *population regression coefficient* $\boldsymbol{\beta}$ solves least squares problem

$$\boldsymbol{\beta} = \arg\min_{\mathbf{b}} E[(Y_i - \mathbf{X}_i'\mathbf{b})^2]$$

- To solve for $\boldsymbol{\beta}$, we take the derivative (i.e. gradient) and set it to zero

$$\frac{d}{d\boldsymbol{\beta}} E[(Y_i - \mathbf{X}_i'\boldsymbol{\beta})^2] = \mathbf{0}$$
$$\Rightarrow E[\frac{d}{d\boldsymbol{\beta}}(Y_i - \mathbf{X}_i'\boldsymbol{\beta})^2] = \mathbf{0}$$

Solving for Regression Coefficients

- So the *population regression coefficient* $\boldsymbol{\beta}$ solves least squares problem

$$\boldsymbol{\beta} = \arg\min_{\mathbf{b}} E[(Y_i - \mathbf{X}_i'\mathbf{b})^2]$$

- To solve for $\boldsymbol{\beta}$, we take the derivative (i.e. gradient) and set it to zero

$$\begin{aligned}
&\frac{d}{d\boldsymbol{\beta}} E[(Y_i - \mathbf{X}_i'\boldsymbol{\beta})^2] = \mathbf{0} \\
\Rightarrow &E[\frac{d}{d\boldsymbol{\beta}}(Y_i - \mathbf{X}_i'\boldsymbol{\beta})^2] = \mathbf{0} \\
\Rightarrow &E[-2\mathbf{X}_i(Y_i - \mathbf{X}_i'\boldsymbol{\beta})] = \mathbf{0}
\end{aligned}$$

Solving for Regression Coefficients

- So the *population regression coefficient* $\boldsymbol{\beta}$ solves least squares problem

$$\boldsymbol{\beta} = \arg\min_{\mathbf{b}} E[(Y_i - \mathbf{X}_i'\mathbf{b})^2]$$

- To solve for $\boldsymbol{\beta}$, we take the derivative (i.e. gradient) and set it to zero

$$\begin{aligned}
&\frac{d}{d\boldsymbol{\beta}} E[(Y_i - \mathbf{X}_i'\boldsymbol{\beta})^2] = \mathbf{0} \\
\Rightarrow &E[\frac{d}{d\boldsymbol{\beta}}(Y_i - \mathbf{X}_i'\boldsymbol{\beta})^2] = \mathbf{0} \\
\Rightarrow &E[-2\mathbf{X}_i(Y_i - \mathbf{X}_i'\boldsymbol{\beta})] = \mathbf{0} \\
\Rightarrow &E[\mathbf{X}_i Y_i] = E[\mathbf{X}_i\mathbf{X}_i']\boldsymbol{\beta}
\end{aligned}$$

Multivariate Regression, in the Population and Sample

- Solving for $\boldsymbol{\beta}$, we obtain an expression involving population means:

$$\boldsymbol{\beta} = E[\boldsymbol{X}_i \boldsymbol{X}_i']^{-1} E[\boldsymbol{X}_i Y_i]$$

  - With a bit of algebra, you can show that this reduces to the bivariate formulas $\beta = \frac{Cov(X_i, Y_i)}{Var(X_i)}$ and $\alpha = E[Y_i] - E[X_i]\beta$ when $\mathbf{X}_i = (1, X_i)'$

Multivariate Regression, in the Population and Sample

- Solving for $\boldsymbol{\beta}$, we obtain an expression involving population means:

$$\boldsymbol{\beta} = E[\boldsymbol{X}_i \boldsymbol{X}_i']^{-1} E[\boldsymbol{X}_i Y_i]$$

  - With a bit of algebra, you can show that this reduces to the bivariate formulas $\beta = \frac{Cov(X_i, Y_i)}{Var(X_i)}$ and $\alpha = E[Y_i] - E[X_i]\beta$ when $\mathbf{X}_i = (1, X_i)'$

- To estimate $\boldsymbol{\beta}$, we can replace population means with sample means.

Multivariate Regression, in the Population and Sample

- Solving for $\boldsymbol{\beta}$, we obtain an expression involving population means:

$$\boldsymbol{\beta} = E[\boldsymbol{X}_i \boldsymbol{X}_i']^{-1} E[\boldsymbol{X}_i Y_i]$$

  - With a bit of algebra, you can show that this reduces to the bivariate formulas $\beta = \frac{Cov(X_i, Y_i)}{Var(X_i)}$ and $\alpha = E[Y_i] - E[X_i]\beta$ when $\mathbf{X}_i = (1, X_i)'$

- To estimate $\boldsymbol{\beta}$, we can replace population means with sample means.

$$\hat{\boldsymbol{\beta}} = \left(\frac{1}{N} \sum_{i=1}^{N} \boldsymbol{X}_i \boldsymbol{X}_i' \right)^{-1} \left(\frac{1}{N} \sum_{i=1}^{N} \boldsymbol{X}_i Y_i \right)$$

Multivariate Regression, in the Population and Sample

- Solving for $\boldsymbol{\beta}$, we obtain an expression involving population means:

$$\boldsymbol{\beta} = E[\boldsymbol{X}_i \boldsymbol{X}_i']^{-1} E[\boldsymbol{X}_i Y_i]$$

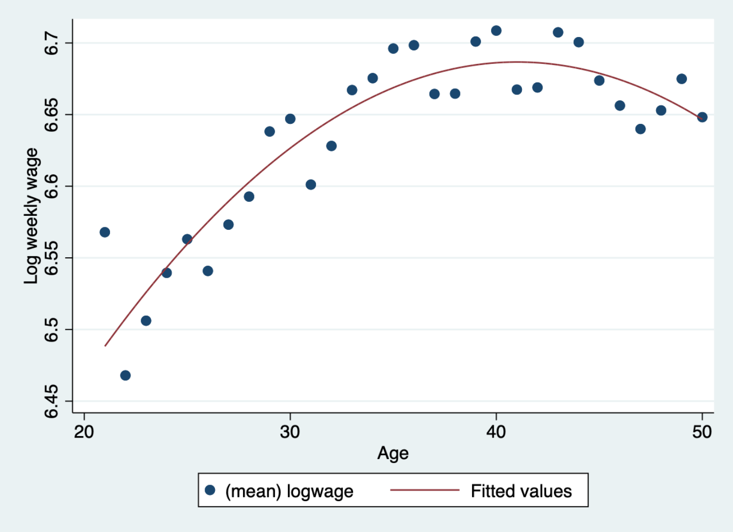
  - With a bit of algebra, you can show that this reduces to the bivariate formulas $\beta = \frac{Cov(X_i, Y_i)}{Var(X_i)}$ and $\alpha = E[Y_i] - E[X_i]\beta$ when $\mathbf{X}_i = (1, X_i)'$

- To estimate $\boldsymbol{\beta}$, we can replace population means with sample means.

$$\hat{\boldsymbol{\beta}} = \left(\frac{1}{N} \sum_{i=1}^{N} \boldsymbol{X}_i \boldsymbol{X}_i' \right)^{-1} \left(\frac{1}{N} \sum_{i=1}^{N} \boldsymbol{X}_i Y_i \right)$$

- We thus now have a general way of estimating $E[Y_i \mid \mathbf{X}_i] \approx \mathbf{X}_i' \boldsymbol{\beta}$ for any vector $\mathbf{X}_i = (1, X_{i1}, \ldots, X_{iK})'$

Quadratic Regression of Log Wages on Age



Constant	5.8591
Age	0.0403
Age^2	-0.0005

Interpreting Quadratic Regression Coefficients

- With a quadratic fit, we have

$$E[Y_i|X_i = x] \approx \hat{\boldsymbol{\beta}}_0 + \hat{\boldsymbol{\beta}}_1 x + \hat{\boldsymbol{\beta}}_2 x^2$$

Interpreting Quadratic Regression Coefficients

- With a quadratic fit, we have

$$E[Y_i|X_i = x] \approx \hat{\boldsymbol{\beta}}_0 + \hat{\boldsymbol{\beta}}_1 x + \hat{\boldsymbol{\beta}}_2 x^2$$

- What is the slope of $E[Y_i|X_i = x]$? Differentiating, we have

Interpreting Quadratic Regression Coefficients

- With a quadratic fit, we have

$$E[Y_i|X_i = x] \approx \hat{\boldsymbol{\beta}}_0 + \hat{\boldsymbol{\beta}}_1 x + \hat{\boldsymbol{\beta}}_2 x^2$$

- What is the slope of $E[Y_i|X_i = x]$? Differentiating, we have

$$\frac{d}{dx} E[Y_i|X_i = x] \approx \hat{\boldsymbol{\beta}}_1 + 2\hat{\boldsymbol{\beta}}_2 x$$

Interpreting Quadratic Regression Coefficients

- With a quadratic fit, we have

$$E[Y_i|X_i = x] \approx \hat{\boldsymbol{\beta}}_0 + \hat{\boldsymbol{\beta}}_1 x + \hat{\boldsymbol{\beta}}_2 x^2$$

- What is the slope of $E[Y_i|X_i = x]$? Differentiating, we have

$$\frac{d}{dx} E[Y_i|X_i = x] \approx \hat{\boldsymbol{\beta}}_1 + 2\hat{\boldsymbol{\beta}}_2 x$$

- The estimated derivative from a multivariate regression, in this case $\hat{\boldsymbol{\beta}}_1 + 2\hat{\boldsymbol{\beta}}_2 x$, is sometimes called the "marginal effect" at x
  - This terminology is a bit unfortunate: this need not be a *causal* effect, just an estimated derivative of the CEF

Interpreting Our Example Coefficients

Constant ($\hat{\beta}_0$)	5.8591
Age ($\hat{\beta}_1$)	0.0403
Age^2 ($\hat{\beta}_2$)	-0.0005

- What is the estimated slope of average log-earnings w.r.t. age?

Interpreting Our Example Coefficients

Constant ($\hat{\beta}_0$)	5.8591
Age ($\hat{\beta}_1$)	0.0403
Age^2 ($\hat{\beta}_2$)	-0.0005

- What is the estimated slope of average log-earnings w.r.t. age?
- $\hat{\boldsymbol{\beta}}_1 + 2\hat{\boldsymbol{\beta}}_2 \cdot Age$

Interpreting Our Example Coefficients

Constant ($\hat{\beta}_0$)	5.8591
Age ($\hat{\beta}_1$)	0.0403
Age^2 ($\hat{\beta}_2$)	-0.0005

- What is the estimated slope of average log-earnings w.r.t. age?
- $\hat{\boldsymbol{\beta}}_1 + 2\hat{\boldsymbol{\beta}}_2 \cdot Age = 0.0403 - 2 \times 0.0005 \cdot Age =$

Interpreting Our Example Coefficients

Constant ($\hat{\beta}_0$)	5.8591
Age ($\hat{\beta}_1$)	0.0403
Age^2 ($\hat{\beta}_2$)	-0.0005

- What is the estimated slope of average log-earnings w.r.t. age?

- $\hat{\boldsymbol{\beta}}_1 + 2\hat{\boldsymbol{\beta}}_2 \cdot Age = 0.0403 - 2 \times 0.0005 \cdot Age = 0.0403 - 0.001 \cdot Age.$

Interpreting Our Example Coefficients

Constant ($\hat{\beta}_0$)	5.8591
Age ($\hat{\beta}_1$)	0.0403
Age^2 ($\hat{\beta}_2$)	-0.0005

- What is the estimated slope of average log-earnings w.r.t. age?
- $\hat{\boldsymbol{\beta}}_1 + 2\hat{\boldsymbol{\beta}}_2 \cdot Age = 0.0403 - 2 \times 0.0005 \cdot Age = 0.0403 - 0.001 \cdot Age.$
- For what age is estimated earnings highest?

Interpreting Our Example Coefficients

Constant ($\hat{\beta}_0$)	5.8591
Age ($\hat{\beta}_1$)	0.0403
Age^2 ($\hat{\beta}_2$)	-0.0005

- What is the estimated slope of average log-earnings w.r.t. age?
- $\hat{\boldsymbol{\beta}}_1 + 2\hat{\boldsymbol{\beta}}_2 \cdot Age = 0.0403 - 2 \times 0.0005 \cdot Age = 0.0403 - 0.001 \cdot Age.$
- For what age is estimated earnings highest?

$$0.0403 - 0.001 \cdot Age = 0$$

Interpreting Our Example Coefficients

Constant ($\hat{\beta}_0$)	5.8591
Age ($\hat{\beta}_1$)	0.0403
Age^2 ($\hat{\beta}_2$)	-0.0005

- What is the estimated slope of average log-earnings w.r.t. age?
- $\hat{\boldsymbol{\beta}}_1 + 2\hat{\boldsymbol{\beta}}_2 \cdot Age = 0.0403 - 2 \times 0.0005 \cdot Age = 0.0403 - 0.001 \cdot Age.$
- For what age is estimated earnings highest?

$$0.0403 - 0.001 \cdot Age = 0 \Rightarrow Age = 0.0403/0.001 = 40.3$$

Re-Writing Multivariate OLS with Matrix Algebra

- We showed that

$$\hat{\boldsymbol{\beta}} = \left(\frac{1}{N} \sum_{i=1}^{N} \boldsymbol{X}_i \boldsymbol{X}_i' \right)^{-1} \left(\frac{1}{N} \sum_{i=1}^{N} \boldsymbol{X}_i Y_i \right)$$

- This formula is often given more compactly with matrix notation.

Re-Writing Multivariate OLS with Matrix Algebra

- We showed that

$$\hat{\boldsymbol{\beta}} = \left(\frac{1}{N} \sum_{i=1}^{N} \boldsymbol{X}_i \boldsymbol{X}_i' \right)^{-1} \left(\frac{1}{N} \sum_{i=1}^{N} \boldsymbol{X}_i Y_i \right)$$

- This formula is often given more compactly with matrix notation.
- Let $\boldsymbol{X}$ be an $N \times K$ matrix w/ X_{ik} giving the element in row i and column k.

Re-Writing Multivariate OLS with Matrix Algebra

- We showed that

$$\hat{\boldsymbol{\beta}} = \left(\frac{1}{N} \sum_{i=1}^{N} \boldsymbol{X}_i \boldsymbol{X}_i' \right)^{-1} \left(\frac{1}{N} \sum_{i=1}^{N} \boldsymbol{X}_i Y_i \right)$$

- This formula is often given more compactly with matrix notation.

- Let $\boldsymbol{X}$ be an $N \times K$ matrix w/ X_{ik} giving the element in row i and column k. Likewise let $\boldsymbol{Y} = (Y_1, ..., Y_N)'$.

Re-Writing Multivariate OLS with Matrix Algebra

- We showed that

$$\hat{\boldsymbol{\beta}} = \left(\frac{1}{N} \sum_{i=1}^{N} \boldsymbol{X}_i \boldsymbol{X}_i' \right)^{-1} \left(\frac{1}{N} \sum_{i=1}^{N} \boldsymbol{X}_i Y_i \right)$$

- This formula is often given more compactly with matrix notation.

- Let $\boldsymbol{X}$ be an $N \times K$ matrix w/ X_{ik} giving the element in row i and column k. Likewise let $\boldsymbol{Y} = (Y_1, ..., Y_N)'$.

- For example, if $\boldsymbol{X}_i = (1, X_i)'$ and $N = 3$, then

$$\boldsymbol{X} = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ 1 & X_3 \end{pmatrix}$$

Re-Writing Multivariate OLS with Matrix Algebra

- We showed that

$$\hat{\boldsymbol{\beta}} = \left(\frac{1}{N} \sum_{i=1}^{N} \boldsymbol{X}_i \boldsymbol{X}_i' \right)^{-1} \left(\frac{1}{N} \sum_{i=1}^{N} \boldsymbol{X}_i Y_i \right)$$

- This formula is often given more compactly with matrix notation.

- Let $\boldsymbol{X}$ be an $N \times K$ matrix w/ X_{ik} giving the element in row i and column k. Likewise let $\boldsymbol{Y} = (Y_1, ..., Y_N)'$.

- For example, if $\boldsymbol{X}_i = (1, X_i)'$ and $N = 3$, then

$$\boldsymbol{X} = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ 1 & X_3 \end{pmatrix} \qquad \boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix}$$

Re-Writing Multivariate OLS with Matrix Algebra

- We showed that

$$\hat{\boldsymbol{\beta}} = \left(\frac{1}{N} \sum_{i=1}^{N} \boldsymbol{X}_i \boldsymbol{X}_i' \right)^{-1} \left(\frac{1}{N} \sum_{i=1}^{N} \boldsymbol{X}_i Y_i \right)$$

- This formula is often given more compactly with matrix notation.

- Let $\boldsymbol{X}$ be an $N \times K$ matrix w/ X_{ik} giving the element in row i and column k. Likewise let $\boldsymbol{Y} = (Y_1, ..., Y_N)'$.

- For example, if $\boldsymbol{X}_i = (1, X_i)'$ and $N = 3$, then

$$\boldsymbol{X} = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ 1 & X_3 \end{pmatrix} \qquad \boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix}$$

- Using this notation, one can show that $\hat{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-1} \boldsymbol{X}' \boldsymbol{Y}$

Asymptotic Properties of Multivariate OLS

- To test hypotheses about the population CEF, we need to derive the asymptotic distribution of $\hat{\boldsymbol{\beta}}$.

Asymptotic Properties of Multivariate OLS

- To test hypotheses about the population CEF, we need to derive the asymptotic distribution of $\hat{\boldsymbol{\beta}}$.
- As with univariate OLS, we can show that $\hat{\boldsymbol{\beta}}$ is consistent and asymptotically normally distributed.

Asymptotic Properties of Multivariate OLS

- To test hypotheses about the population CEF, we need to derive the asymptotic distribution of $\hat{\boldsymbol{\beta}}$.
- As with univariate OLS, we can show that $\hat{\boldsymbol{\beta}}$ is consistent and asymptotically normally distributed.
- The proofs are very similar to those for univariate OLS, so I'll skip them and show you the results!

Asymptotic Properties of Multivariate OLS

- To test hypotheses about the population CEF, we need to derive the asymptotic distribution of $\hat{\boldsymbol{\beta}}$.
- As with univariate OLS, we can show that $\hat{\boldsymbol{\beta}}$ is consistent and asymptotically normally distributed.
- The proofs are very similar to those for univariate OLS, so I'll skip them and show you the results!
- **Consistency:** $\hat{\boldsymbol{\beta}} \to_p \boldsymbol{\beta}$.

Asymptotic Properties of Multivariate OLS

- **Asymptotic normality**:

$$\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \rightarrow_d \mathrm{N}(0, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\Sigma} = E[\boldsymbol{X}_i \boldsymbol{X}_i']^{-1} Var(\boldsymbol{X}_i \varepsilon_i) E[\boldsymbol{X}_i \boldsymbol{X}_i']^{-1}$ and $\varepsilon_i = Y_i - \boldsymbol{X}_i' \boldsymbol{\beta}$

Asymptotic Properties of Multivariate OLS

- **Asymptotic normality**:

$$\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \rightarrow_d \mathrm{N}(0, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\Sigma} = E[\boldsymbol{X}_i \boldsymbol{X}_i']^{-1} Var(\boldsymbol{X}_i \varepsilon_i) E[\boldsymbol{X}_i \boldsymbol{X}_i']^{-1}$ and $\varepsilon_i = Y_i - \boldsymbol{X}_i' \boldsymbol{\beta}$

- We can estimate $\boldsymbol{\Sigma}$ by replacing population means with sample means

$$\hat{\boldsymbol{\Sigma}} = \left(\frac{1}{N} \sum_{i=1}^{N} \boldsymbol{X}_i \boldsymbol{X}_i' \right)^{-1} \left(\frac{1}{N} \sum_{i=1}^{N} \boldsymbol{X}_i \boldsymbol{X}_i' \hat{\varepsilon}_i^2 \right) \left(\frac{1}{N} \sum_{i=1}^{N} \boldsymbol{X}_i \boldsymbol{X}_i' \right)^{-1}$$

where $\hat{\varepsilon}_i = Y_i - \boldsymbol{X}_i' \hat{\boldsymbol{\beta}}$.

Asymptotic Properties of Multivariate OLS

- **Asymptotic normality**:

$$\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \rightarrow_d \mathrm{N}(0, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\Sigma} = E[\boldsymbol{X}_i \boldsymbol{X}_i']^{-1} Var(\boldsymbol{X}_i \varepsilon_i) E[\boldsymbol{X}_i \boldsymbol{X}_i']^{-1}$ and $\varepsilon_i = Y_i - \boldsymbol{X}_i' \boldsymbol{\beta}$

- We can estimate $\boldsymbol{\Sigma}$ by replacing population means with sample means

$$\hat{\boldsymbol{\Sigma}} = \left(\frac{1}{N} \sum_{i=1}^{N} \boldsymbol{X}_i \boldsymbol{X}_i' \right)^{-1} \left(\frac{1}{N} \sum_{i=1}^{N} \boldsymbol{X}_i \boldsymbol{X}_i' \hat{\varepsilon}_i^2 \right) \left(\frac{1}{N} \sum_{i=1}^{N} \boldsymbol{X}_i \boldsymbol{X}_i' \right)^{-1}$$

where $\hat{\varepsilon}_i = Y_i - \boldsymbol{X}_i' \hat{\boldsymbol{\beta}}$.

- Note: $\hat{\boldsymbol{\Sigma}}$ is a *matrix*.
  - The standard error for $\hat{\boldsymbol{\beta}}_j$ is $\sqrt{\hat{\boldsymbol{\Sigma}}_{jj}} / \sqrt{N}$
  - The off-diagonal elements correspond with covariances between $\hat{\boldsymbol{\beta}}_j, \hat{\boldsymbol{\beta}}_k$

Example - Log Earnings by Age

Variable	Coefficient	SE
Constant (β_0)	5.8591	0.1409
Age (β_1)	0.0403	0.0077
Age^2 (β_2)	-0.0005	0.0001

- What is a confidence interval for β_2?

Example - Log Earnings by Age

Variable	Coefficient	SE
Constant (β_0)	5.8591	0.1409
Age (β_1)	0.0403	0.0077
Age^2 (β_2)	-0.0005	0.0001

- What is a confidence interval for β_2?

$$\hat{\beta}_2 \pm 1.96 \times SE_{\beta_2} =$$

Example - Log Earnings by Age

Variable	Coefficient	SE
Constant (β_0)	5.8591	0.1409
Age (β_1)	0.0403	0.0077
Age^2 (β_2)	-0.0005	0.0001

- What is a confidence interval for β_2?

$$\hat{\beta}_2 \pm 1.96 \times SE_{\beta_2} = -0.0005 \pm 1.96 \times 0.0001$$

Example - Log Earnings by Age

Variable	Coefficient	SE
Constant (β_0)	5.8591	0.1409
Age (β_1)	0.0403	0.0077
Age^2 (β_2)	-0.0005	0.0001

- What is a confidence interval for β_2?

$$\hat{\beta}_2 \pm 1.96 \times SE_{\beta_2} = -0.0005 \pm 1.96 \times 0.0001 = [-0.0007, -0.0003]$$

Controlling for Multiple Variables

- In addition to allowing for more flexible functional forms (e.g. quadratic), multivariate OLS allows us to approximate the CEF conditional on multiple variables at once

Controlling for Multiple Variables

- In addition to allowing for more flexible functional forms (e.g. quadratic), multivariate OLS allows us to approximate the CEF conditional on multiple variables at once
- Example: we have data from Texas on each county's presidential vote over the last three elections (2012,2016,2020)

Controlling for Multiple Variables

- In addition to allowing for more flexible functional forms (e.g. quadratic), multivariate OLS allows us to approximate the CEF conditional on multiple variables at once

- Example: we have data from Texas on each county's presidential vote over the last three elections (2012,2016,2020)

- Let $Y_i =$ Biden vote share in county i, $X_{i1} =$ Clinton vote share in county i, and $X_{i2} =$ Obama vote share in county i

Controlling for Multiple Variables

- In addition to allowing for more flexible functional forms (e.g. quadratic), multivariate OLS allows us to approximate the CEF conditional on multiple variables at once

- Example: we have data from Texas on each county's presidential vote over the last three elections (2012,2016,2020)

- Let $Y_i =$ Biden vote share in county i, $X_{i1} =$ Clinton vote share in county i, and $X_{i2} =$ Obama vote share in county i

- We estimate the regression

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$$

Controlling for Multiple Variables

- In addition to allowing for more flexible functional forms (e.g. quadratic), multivariate OLS allows us to approximate the CEF conditional on multiple variables at once

- Example: we have data from Texas on each county's presidential vote over the last three elections (2012,2016,2020)

- Let $Y_i =$ Biden vote share in county i, $X_{i1} =$ Clinton vote share in county i, and $X_{i2} =$ Obama vote share in county i

- We estimate the regression

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$$

- This says that

$$E\left[\text{Biden vote} \middle| \text{Clinton vote}, \text{Obama vote}\right] \approx \beta_0 + \beta_1 \times \text{Clinton vote} + \beta_2 \times \text{Obama vote}$$

OLS Estimates

Variable	Coefficient	SE
Constant	0.05	0.01
Clinton	1.39	0.13
Obama	-0.56	0.13

OLS Estimates

Variable	Coefficient	SE
Constant	0.05	0.01
Clinton	1.39	0.13
Obama	-0.56	0.13

- What is the predicted Biden vote share for a county where Obama got half the vote and Clinton got 60%?

OLS Estimates

Variable	Coefficient	SE
Constant	0.05	0.01
Clinton	1.39	0.13
Obama	-0.56	0.13

- What is the predicted Biden vote share for a county where Obama got half the vote and Clinton got 60%?

$$\hat{\beta}_0 + \hat{\beta}_1 0.6 + \hat{\beta}_2 0.5 =$$

OLS Estimates

Variable	Coefficient	SE
Constant	0.05	0.01
Clinton	1.39	0.13
Obama	-0.56	0.13

- What is the predicted Biden vote share for a county where Obama got half the vote and Clinton got 60%?

$$\hat{\beta}_0 + \hat{\beta}_1 0.6 + \hat{\beta}_2 0.5 = 0.05 + 1.39 \times 0.6 - 0.56 \times 0.5 =$$

OLS Estimates

Variable	Coefficient	SE
Constant	0.05	0.01
Clinton	1.39	0.13
Obama	-0.56	0.13

- What is the predicted Biden vote share for a county where Obama got half the vote and Clinton got 60%?

$$\hat{\beta}_0 + \hat{\beta}_1 0.6 + \hat{\beta}_2 0.5 = 0.05 + 1.39 \times 0.6 - 0.56 \times 0.5 = 0.604$$

OLS Estimates

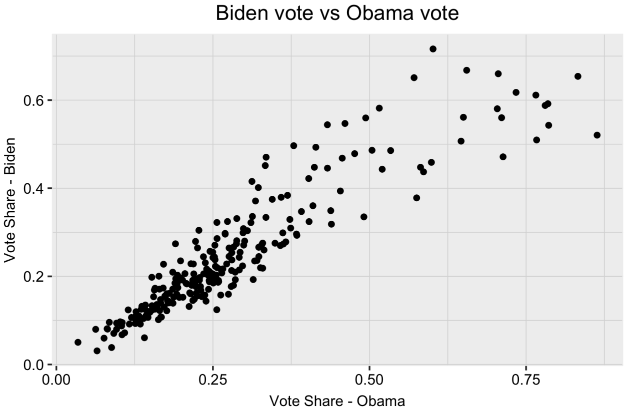
Variable	Coefficient	SE
Constant	0.05	0.01
Clinton	1.39	0.13
Obama	-0.56	0.13

- What is the predicted Biden vote share for a county where Obama got half the vote and Clinton got 60%?

$$\hat{\beta}_0 + \hat{\beta}_1 0.6 + \hat{\beta}_2 0.5 = 0.05 + 1.39 \times 0.6 - 0.56 \times 0.5 = 0.604$$

- Notice that the coefficient on Obama vote share is *negative*

- Wait, does this mean Biden did worse in places that Obama did well?!



- If we look at the data, we see that Biden vote share is highly positively correlated with Obama vote share.
- So what's going on?!

Interpreting Regression Coefficients

- Remember that multivariate OLS is approximating the CEF as

$$E[Y_i|\boldsymbol{X}_i = \boldsymbol{x}] \approx \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2$$

Interpreting Regression Coefficients

- Remember that multivariate OLS is approximating the CEF as

$$E[Y_i|\boldsymbol{X}_i = \boldsymbol{x}] \approx \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2$$

- Thus, β_2 is an estimate of a *partial derivative*,

$$\frac{\partial}{\partial x_{i2}} E[Y_i|\boldsymbol{X}_i = \boldsymbol{x}] \approx \beta_2,$$

  i.e. the change in the CEF from changing X_{i2} *holding* X_{i1} *constant*.

Interpreting Regression Coefficients

- Remember that multivariate OLS is approximating the CEF as

$$E[Y_i|\boldsymbol{X}_i = \boldsymbol{x}] \approx \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2$$

- Thus, β_2 is an estimate of a *partial derivative*,

$$\frac{\partial}{\partial x_{i2}} E[Y_i|\boldsymbol{X}_i = \boldsymbol{x}] \approx \beta_2,$$

i.e. the change in the CEF from changing X_{i2} *holding* X_{i1} *constant*.

- If $\beta_2 < 0$, this means that among places where Clinton had the same vote share, Biden did better in places with lower Obama vote share.

Interpreting Regression Coefficients

- Remember that multivariate OLS is approximating the CEF as

$$E[Y_i|\boldsymbol{X}_i = \boldsymbol{x}] \approx \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2$$

- Thus, β_2 is an estimate of a *partial derivative*,

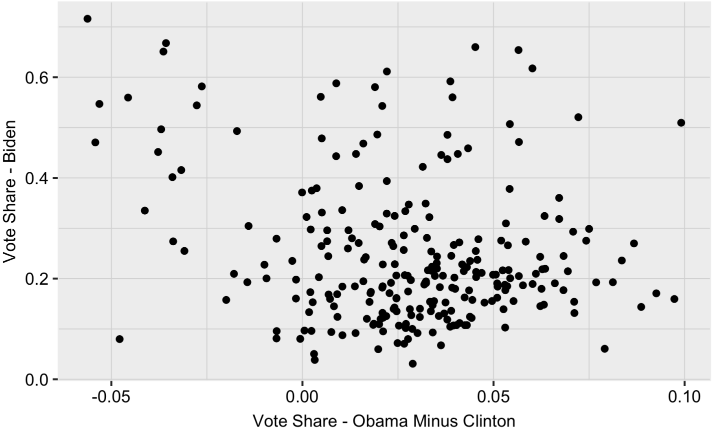
$$\frac{\partial}{\partial x_{i2}} E[Y_i|\boldsymbol{X}_i = \boldsymbol{x}] \approx \beta_2,$$

  i.e. the change in the CEF from changing X_{i2} *holding* X_{i1} *constant*.

- If $\beta_2 < 0$, this means that among places where Clinton had the same vote share, Biden did better in places with lower Obama vote share.

- In other words, Biden did better in places where Democratic vote share was increasing between 2012 and 2016!

Obama vote vs Clinton vote



Generalizing this idea

- The **Frisch-Waugh-Lovell** (FWL) theorem gives us a general way to interpret coefficients in multivariate regression.

Generalizing this idea

- The **Frisch-Waugh-Lovell** (FWL) theorem gives us a general way to interpret coefficients in multivariate regression. Consider the regression

$$Y_i = \beta_0 + X_{i1}\beta_1 + X_{i2}\beta_2 + \varepsilon_i$$

Generalizing this idea

- The **Frisch-Waugh-Lovell** (FWL) theorem gives us a general way to interpret coefficients in multivariate regression. Consider the regression

$$Y_i = \beta_0 + X_{i1}\beta_1 + X_{i2}\beta_2 + \varepsilon_i$$

- FWL says that the OLS coefficient $\hat{\beta}_2$ can be obtained by the following steps:

Generalizing this idea

- The **Frisch-Waugh-Lovell** (FWL) theorem gives us a general way to interpret coefficients in multivariate regression. Consider the regression

$$Y_i = \beta_0 + X_{i1}\beta_1 + X_{i2}\beta_2 + \varepsilon_i$$

- FWL says that the OLS coefficient $\hat{\beta}_2$ can be obtained by the following steps:

- 1) Regress X_{i2} on X_{i1} and a constant:

$$X_{i2} = \gamma_0 + X_{i1}\gamma_1 + u_i$$

Generalizing this idea

- The **Frisch-Waugh-Lovell** (FWL) theorem gives us a general way to interpret coefficients in multivariate regression. Consider the regression

$$Y_i = \beta_0 + X_{i1}\beta_1 + X_{i2}\beta_2 + \varepsilon_i$$

- FWL says that the OLS coefficient $\hat{\beta}_2$ can be obtained by the following steps:

- 1) Regress X_{i2} on X_{i1} and a constant:

$$X_{i2} = \gamma_0 + X_{i1}\gamma_1 + u_i$$

- 2) For each unit, predict X_{i2} using the coefficients obtained in step 1)

$$\hat{X}_{i2} = \hat{\gamma}_0 + X_{i1}\hat{\gamma}_1$$

Generalizing this idea

- The **Frisch-Waugh-Lovell** (FWL) theorem gives us a general way to interpret coefficients in multivariate regression. Consider the regression

$$Y_i = \beta_0 + X_{i1}\beta_1 + X_{i2}\beta_2 + \varepsilon_i$$

- FWL says that the OLS coefficient $\hat{\beta}_2$ can be obtained by the following steps:
- 1) Regress X_{i2} on X_{i1} and a constant:

$$X_{i2} = \gamma_0 + X_{i1}\gamma_1 + u_i$$

- 2) For each unit, predict X_{i2} using the coefficients obtained in step 1)

$$\hat{X}_{i2} = \hat{\gamma}_0 + X_{i1}\hat{\gamma}_1$$

- 3) Obtain $\hat{\beta}_2$ by regressing Y_i on the OLS residual $X_{i2} - \hat{X}_{i2}$:

$$Y_i = \alpha + (X_{i2} - \hat{X}_{i2})\beta_2 + v_i$$

Illustration Using Election Data

- Regress Obama vote share on Clinton vote share

Intercept ($\hat{\gamma}_0$)	0.03
Clinton ($\hat{\gamma}_1$)	0.98

Illustration Using Election Data

- Regress Obama vote share on Clinton vote share

Intercept ($\hat{\gamma}_0$)	0.03
Clinton ($\hat{\gamma}_1$)	0.98

- Predict Obama vote share using Clinton vote share:

$$\hat{X}_{i2} =$$

Illustration Using Election Data

- Regress Obama vote share on Clinton vote share

Intercept ($\hat{\gamma}_0$)	0.03
Clinton ($\hat{\gamma}_1$)	0.98

- Predict Obama vote share using Clinton vote share:

$$\hat{X}_{i2} = 0.03 + 0.98 X_{i1}$$

Illustration Using Election Data

- Regress Obama vote share on Clinton vote share

Intercept ($\hat{\gamma}_0$)	0.03
Clinton ($\hat{\gamma}_1$)	0.98

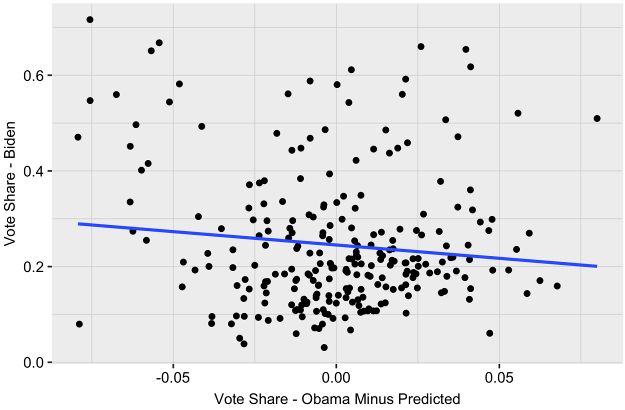
- Predict Obama vote share using Clinton vote share:

$$\hat{X}_{i2} = 0.03 + 0.98 X_{i1}$$

- Regress Biden vote share on $X_{i2} - \hat{X}_{i2}$:

Intercept ($\hat{\alpha}$)	0.25
Obama minus predicted ($\hat{\beta}_2$)	-0.56

- The estimate $\hat{\beta}_2$, -0.56, is exactly what we got before!



- The slope of the best-fit line is precisely $\hat{\beta}_2 = -0.56$.
- FWL generally gives us an easy way to visualize/interpret multivariate regression coefficients

Measures of Model Fit

- Did adding a quadratic term help us improve our approximation to the wage-age CEF? How can we measure this?

Measures of Model Fit

- Did adding a quadratic term help us improve our approximation to the wage-age CEF? How can we measure this?

- One way of measuring model fit is the population R^2: for the regression $Y_i = \boldsymbol{X}_i'\boldsymbol{\beta} + \varepsilon_i$,

$$R^2 = \frac{Var(\boldsymbol{X}_i'\boldsymbol{\beta})}{Var(Y_i)}$$

Measures of Model Fit

- Did adding a quadratic term help us improve our approximation to the wage-age CEF? How can we measure this?

- One way of measuring model fit is the population R^2: for the regression $Y_i = \boldsymbol{X}_i'\boldsymbol{\beta} + \varepsilon_i$,

$$R^2 = \frac{Var(\boldsymbol{X}_i'\boldsymbol{\beta})}{Var(Y_i)}$$

- Intuitively, population R^2 measures the fraction of the variance of Y_i explained by $\boldsymbol{X}_i'\beta$
  - Since $Cov(\mathbf{X}_i'\boldsymbol{\beta}, e_i) = 0$, we also have $R^2 = 1 - \frac{Var(e_i)}{Var(Y_i)}$

Measures of Model Fit

- Did adding a quadratic term help us improve our approximation to the wage-age CEF? How can we measure this?

- One way of measuring model fit is the population R^2: for the regression $Y_i = \boldsymbol{X}_i'\boldsymbol{\beta} + \varepsilon_i$,

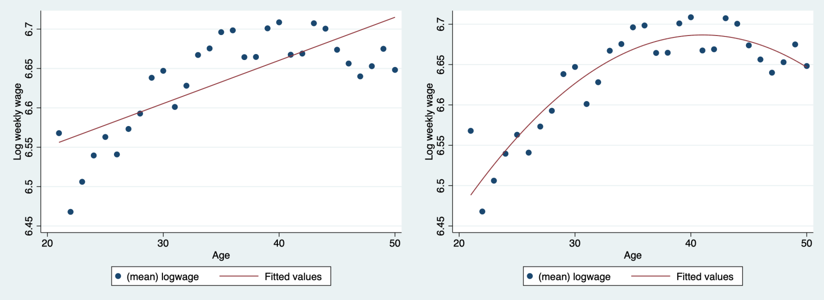
$$R^2 = \frac{Var(\boldsymbol{X}_i'\boldsymbol{\beta})}{Var(Y_i)}$$

- Intuitively, population R^2 measures the fraction of the variance of Y_i explained by $\boldsymbol{X}_i'\beta$
  - Since $Cov(\mathbf{X}_i'\boldsymbol{\beta}, e_i) = 0$, we also have $R^2 = 1 - \frac{Var(e_i)}{Var(Y_i)}$

- To estimate R^2, we replace population values with sample analogs

$$\hat{R}^2 = \frac{\frac{1}{N}\sum_i(\mathbf{X}_i'\hat{\boldsymbol{\beta}} - \bar{\mathbf{X}}_i'\hat{\boldsymbol{\beta}})^2}{\frac{1}{N}\sum_i(Y_i - \bar{Y})^2} = 1 - \frac{\frac{1}{N}\sum_i \hat{\varepsilon}_i^2}{\frac{1}{N}\sum_i(Y_i - \bar{Y})^2}$$

R^2 in the Wage-Age Example



(a) $\hat{R}^2 = 0.44$ (b) $\hat{R}^2 = 0.73$

- The linear fit explains 44% of the variation in average earnings across ages, whereas the quadratic fit explains 73%

Caution about $\hat{R}^2$

- Caution: the sample $\hat{R}^2$ will always increase if you have a more complicated model. Why?

Caution about $\hat{R}^2$

- Caution: the sample $\hat{R}^2$ will always increase if you have a more complicated model. Why?

- The coefficients from a linear fit minimizes

$$\frac{1}{N}\sum_i (\underbrace{Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)}_{\hat{\varepsilon}_{Linear}})^2$$

  While the coefficients in a quadratic fit minimize

$$\frac{1}{N}\sum_i (\underbrace{Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\beta}_2 X_i^2)}_{\hat{\varepsilon}_{Quad}})^2$$

  Since $\hat{\beta}_2 = 0$ is feasible in the quadratic minimization, the minimization will always be weakly lower w/a quadratic term

Caution about $\hat{R}^2$

- Caution: the sample $\hat{R}^2$ will always increase if you have a more complicated model. Why?

- The coefficients from a linear fit minimizes

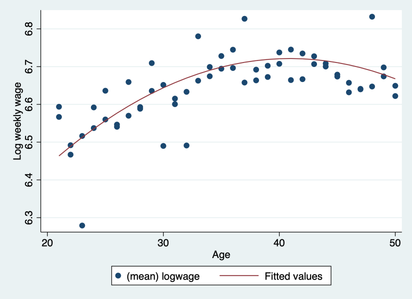
$$\frac{1}{N}\sum_i(\underbrace{Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)}_{\hat{\varepsilon}_{Linear}})^2$$

  While the coefficients in a quadratic fit minimize

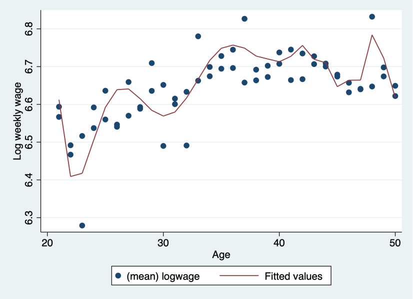
$$\frac{1}{N}\sum_i(\underbrace{Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\beta}_2 X_i^2)}_{\hat{\varepsilon}_{Quad}})^2$$

  Since $\hat{\beta}_2 = 0$ is feasible in the quadratic minimization, the minimization will always be weakly lower w/a quadratic term

- But is a more complicated model always better?

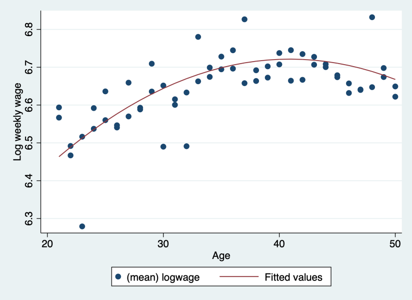


(c) Quadratic, $\hat{R}^2 = 0.44$

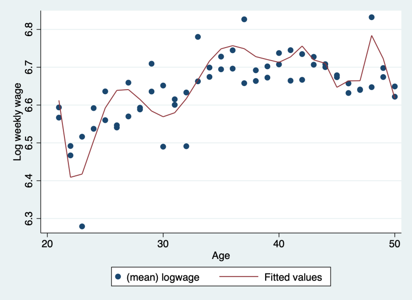


(d) 20th order poly, $\hat{R}^2 = 0.70$

- Suppose we take a sample of size 10,000 and fit a quadratic and a 20th order polynomial

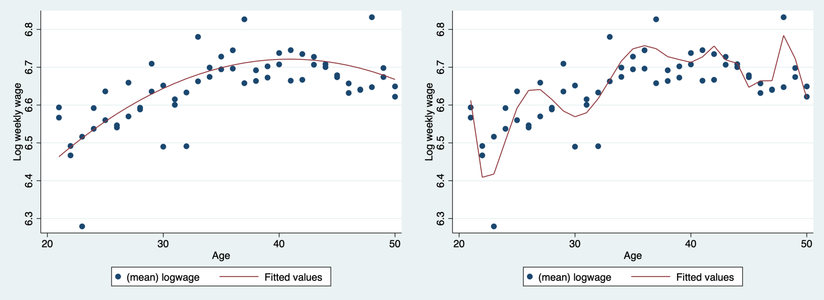


(e) Quadratic, $\hat{R}^2 = 0.44$



(f) 20th order poly, $\hat{R}^2 = 0.70$

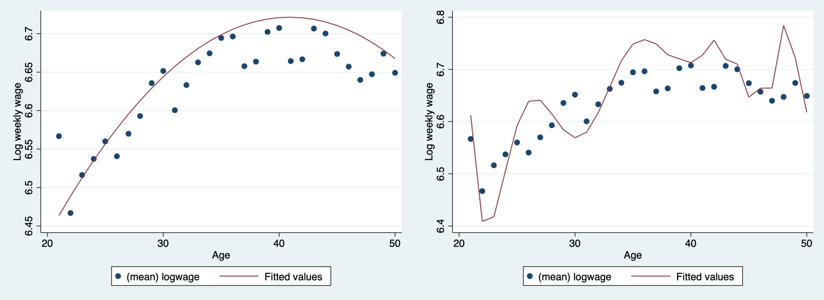
- Suppose we take a sample of size 10,000 and fit a quadratic and a 20th order polynomial
- The 20th order poly has higher R^2, does it look reasonable to you?



(g) Quadratic, $\hat{R}^2 = 0.44$ (h) 20th order poly, $\hat{R}^2 = 0.70$

- Suppose we take a sample of size 10,000 and fit a quadratic and a 20th order polynomial
- The 20th order poly has higher R^2, does it look reasonable to you?
- No, it looks too "squiggly" – it has adapted to fit the exact points in the sample

- Suppose we draw a new sample and test the prediction of our model trained on the first data-set



(i) Quadratic

(j) 20th order poly

- Suppose we draw a new sample and test the prediction of our model trained on the first data-set

- The quadratic fit generalizes pretty well to the new data.

- But the 20th-order polynomial does very poorly. It “overfit” the features of the specific previous sample. This doesn’t generalize well to a new sample

How Can We Avoid Over-Fitting?

- The challenge is to pick a rich enough model to capture the key features of the CEF, but not too rich a model such that we overfit

How Can We Avoid Over-Fitting?

- The challenge is to pick a rich enough model to capture the key features of the CEF, but not too rich a model such that we overfit
- There are several tools available to try to help with this task, none of which is perfect.

How Can We Avoid Over-Fitting?

- The challenge is to pick a rich enough model to capture the key features of the CEF, but not too rich a model such that we overfit
- There are several tools available to try to help with this task, none of which is perfect.
- **Adjusted** $\hat{R}^2$ is a modification to $\hat{R}^2$ that adds a penalty for models with more variables
  - Generally better than $\hat{R}^2$, but model w/highest $\hat{R}^2$ need not be best

How Can We Avoid Over-Fitting?

- The challenge is to pick a rich enough model to capture the key features of the CEF, but not too rich a model such that we overfit
- There are several tools available to try to help with this task, none of which is perfect.
- **Adjusted** $\hat{R}^2$ is a modification to $\hat{R}^2$ that adds a penalty for models with more variables
    - Generally better than $\hat{R}^2$, but model w/highest $\hat{R}^2$ need not be best
- **Cross validation**: choose complexity of the model based on how well it does "out of sample"

How Can We Avoid Over-Fitting?

- The challenge is to pick a rich enough model to capture the key features of the CEF, but not too rich a model such that we overfit

- There are several tools available to try to help with this task, none of which is perfect.

- **Adjusted** $\hat{R}^2$ is a modification to $\hat{R}^2$ that adds a penalty for models with more variables
  - Generally better than $\hat{R}^2$, but model w/highest $\hat{R}^2$ need not be best

- **Cross validation**: choose complexity of the model based on how well it does "out of sample"
  - Split data in two: train model on one half, and see how well it predicts on the second half
  - Choose the complexity of the model based on how well it predicts out of sample

How Can We Avoid Over-Fitting?

- The challenge is to pick a rich enough model to capture the key features of the CEF, but not too rich a model such that we overfit

- There are several tools available to try to help with this task, none of which is perfect.

- **Adjusted** $\hat{R}^2$ is a modification to $\hat{R}^2$ that adds a penalty for models with more variables
  - Generally better than $\hat{R}^2$, but model w/highest $\hat{R}^2$ need not be best

- **Cross validation**: choose complexity of the model based on how well it does "out of sample"
  - Split data in two: train model on one half, and see how well it predicts on the second half
  - Choose the complexity of the model based on how well it predicts out of sample
  - Cross-validation is the basis of modern *machine learning* (ML) methods. ML is very powerful, but how to use ML for causal inference is still being worked out

Avoiding Overfitting in Practice

- The tools described above are useful for deciding between models, but in practice model selection is often done more heuristically

Avoiding Overfitting in Practice

- The tools described above are useful for deciding between models, but in practice model selection is often done more heuristically
- Researchers will typically start with a simple model (e.g. linear or quadratic) that includes what they think are the most important variables

Avoiding Overfitting in Practice

- The tools described above are useful for deciding between models, but in practice model selection is often done more heuristically
- Researchers will typically start with a simple model (e.g. linear or quadratic) that includes what they think are the most important variables
- Then, they will assess the “robustness” of the conclusions to adding/subtracting variables and/or higher-order terms.

Avoiding Overfitting in Practice

- The tools described above are useful for deciding between models, but in practice model selection is often done more heuristically
- Researchers will typically start with a simple model (e.g. linear or quadratic) that includes what they think are the most important variables
- Then, they will assess the "robustness" of the conclusions to adding/subtracting variables and/or higher-order terms.
- Generally, we will be more confident if the model conclusions are not sensitive to tweaks in the model specification.

Outline

1. Deriving Multivariate Regression and OLS✓
2. Regression and Causality
3. Regression Odds and Ends

Regression Meets Causality

- Multivariate regressions are often used to estimate causal effects under conditional unconfoundedness

Regression Meets Causality

- Multivariate regressions are often used to estimate causal effects under conditional unconfoundedness

- Recall that under conditional unconfoundedness,

$$CATE(\boldsymbol{x}) = E[Y_i|D_i = 1, \boldsymbol{X}_i = \boldsymbol{x}] - E[Y_i|D_i = 0, \boldsymbol{X}_i = \boldsymbol{x}]$$

Regression Meets Causality

- Multivariate regressions are often used to estimate causal effects under conditional unconfoundedness

- Recall that under conditional unconfoundedness,

$$CATE(\boldsymbol{x}) = E[Y_i|D_i = 1, \boldsymbol{X}_i = \boldsymbol{x}] - E[Y_i|D_i = 0, \boldsymbol{X}_i = \boldsymbol{x}]$$

- Common to approximate the CEF linearly, as

$$E[Y_i|D_i, \boldsymbol{X}_i] \approx D_i\beta + \boldsymbol{X}_i'\boldsymbol{\gamma}$$

Regression Meets Causality

- Multivariate regressions are often used to estimate causal effects under conditional unconfoundedness

- Recall that under conditional unconfoundedness,

$$CATE(\boldsymbol{x}) = E[Y_i | D_i = 1, \boldsymbol{X}_i = \boldsymbol{x}] - E[Y_i | D_i = 0, \boldsymbol{X}_i = \boldsymbol{x}]$$

- Common to approximate the CEF linearly, as

$$E[Y_i | D_i, \boldsymbol{X}_i] \approx D_i \beta + \boldsymbol{X}_i' \boldsymbol{\gamma}$$

- Then conditional unconfoundedness implies that $CATE(\boldsymbol{x}) \approx \beta$.
    - Doesn't depend on **x**, so also have $\beta \approx ATE$

Regression Meets Causality

- Multivariate regressions are often used to estimate causal effects under conditional unconfoundedness

- Recall that under conditional unconfoundedness,

$$CATE(\boldsymbol{x}) = E[Y_i|D_i = 1, \boldsymbol{X}_i = \boldsymbol{x}] - E[Y_i|D_i = 0, \boldsymbol{X}_i = \boldsymbol{x}]$$

- Common to approximate the CEF linearly, as

$$E[Y_i|D_i, \boldsymbol{X}_i] \approx D_i\beta + \boldsymbol{X}_i'\boldsymbol{\gamma}$$

- Then conditional unconfoundedness implies that $CATE(\boldsymbol{x}) \approx \beta$.
  - Doesn't depend on $\mathbf{x}$, so also have $\beta \approx ATE$

- So if we estimate the multivariate regression

$$Y_i = D_i\beta + \boldsymbol{X}_i'\boldsymbol{\gamma} + \varepsilon_i,$$

we can interpret $\hat{\boldsymbol{\beta}}$ as an estimate of the ATE.

Dale and Krueger

- Dale & Krueger (as you recall) are interested in the effect of attending a more selective college on earnings
  - They have data on earnings and college application information from the College and Beyond (C&B) Survey

Dale and Krueger

- Dale & Krueger (as you recall) are interested in the effect of attending a more selective college on earnings
  - They have data on earnings and college application information from the College and Beyond (C&B) Survey
- The C&B survey covers students who attended 30 colleges for the high school class of 1978; it contains important variables:

Dale and Krueger

- Dale & Krueger (as you recall) are interested in the effect of attending a more selective college on earnings
  - They have data on earnings and college application information from the College and Beyond (C&B) Survey

- The C&B survey covers students who attended 30 colleges for the high school class of 1978; it contains important variables:

- **Earnings** in 1996

- **College application and demographic variables** including SAT scores, class rank, family income, race, etc

- **College application decisions** — i.e. the set of schools students applied to and were admitted

Institution	School-average SAT score in 1978
Barnard College	1210
Bryn Mawr College	1370
Columbia University	1330
Denison University	1020
Duke University	1226
Emory University	1150
Georgetown University	1225
Hamilton College	1246
Kenyon College	1155
Miami University (Ohio)	1073
Northwestern University	1240
Oberlin College	1227
Pennsylvania State University	1038
Princeton University	1308
Rice University	1316
Smith College	1210
Stanford University	1270
Swarthmore College	1340
Tufts University	1200
Tulane University	1080
University of Michigan (Ann Arbor)	1110
University of North Carolina (Chapel Hill)	1080
University of Notre Dame	1200
University of Pennsylvania	1280
Vanderbilt University	1162
Washington University	1180
Wellesley College	1220
Wesleyan University	1260
Williams College	1255
Yale University	1360

Dealing with Selection

- Dale & Krueger assume **conditional unconfoundedness**, i.e. $D_i \perp\!\!\!\perp (Y_i(\cdot))|X_i$ where D_i is the average SAT score for students at your college and X_i is a set of controls

Dealing with Selection

- Dale & Krueger assume **conditional unconfoundedness**, i.e. $D_i \perp\!\!\!\perp (Y_i(\cdot))|X_i$ where D_i is the average SAT score for students at your college and X_i is a set of controls

- They then estimate regressions of the form

$$ln(Y_i) = D_i\beta + \boldsymbol{X}_i'\boldsymbol{\gamma} + \varepsilon_i$$

where Y_i is 1996 earnings

Dealing with Selection

- Dale & Krueger assume **conditional unconfoundedness**, i.e. $D_i \perp\!\!\!\perp (Y_i(\cdot))|X_i$ where D_i is the average SAT score for students at your college and X_i is a set of controls

- They then estimate regressions of the form

$$ln(Y_i) = D_i\beta + \boldsymbol{X}_i'\boldsymbol{\gamma} + \varepsilon_i$$

  where Y_i is 1996 earnings

- If conditional unconfoundedness holds & the regression approx. to the CEF is decent, then β should (approximately) equal the treatment effect of attending a college with higher average SAT scores.

First Pass: SAT Scores and Demographics in $\boldsymbol{X}_i$

Variable	Full sample 1
School-average SAT score/100	0.076 (0.016)
Predicted log(parental income)	0.187 (0.024)
Own SAT score/100	0.018 (0.006)
Female	−0.403 (0.015)
Black	−0.023 (0.035)
Hispanic	0.015 (0.052)
Asian	0.173 (0.036)
Other/missing race	−0.188 (0.119)
High school top 10 percent	0.061 (0.018)
High school rank missing	0.001 (0.024)
Athlete	0.102 (0.025)

- $\hat{\beta} = 0.076$ indicates about an increase in log wages of 7.6 from attending a school with 100 higher SAT points

Do You Buy This Estimate?

- Why might unconfoundedness fail when using these controls?

Do You Buy This Estimate?

- Why might unconfoundedness fail when using these controls?
  - Which colleges you get into may depend on relevant unobserved factors — e.g., students with better application essays may get into more colleges and earn more regardless of where they go

Do You Buy This Estimate?

- Why might unconfoundedness fail when using these controls?
  - Which colleges you get into may depend on relevant unobserved factors — e.g., students with better application essays may get into more colleges and earn more regardless of where they go
- To address this concern, Dale and Kruger have a second analysis: control for the set of colleges that a student applied to / was admitted

Do You Buy This Estimate?

- Why might unconfoundedness fail when using these controls?
  - Which colleges you get into may depend on relevant unobserved factors — e.g., students with better application essays may get into more colleges and earn more regardless of where they go
- To address this concern, Dale and Kruger have a second analysis: control for the set of colleges that a student applied to / was admitted
  - Are able to see this because of the C&B data

Introduction to “Fixed Effects”

- What exactly does it mean that they control for the set of colleges you applied / were admitted to?

Introduction to "Fixed Effects"

- What exactly does it mean that they control for the set of colleges you applied / were admitted to?
- Consider a simplified version where there are three schools, Brown, Yale, and URI, and everyone applies to all 3.

Introduction to "Fixed Effects"

- What exactly does it mean that they control for the set of colleges you applied / were admitted to?
- Consider a simplified version where there are three schools, Brown, Yale, and URI, and everyone applies to all 3.
- Suppose all students in the sample get into URI, but admissions decisions at Brown/Yale differ. There are four possible outcomes:

Introduction to "Fixed Effects"

- What exactly does it mean that they control for the set of colleges you applied / were admitted to?
- Consider a simplified version where there are three schools, Brown, Yale, and URI, and everyone applies to all 3.
- Suppose all students in the sample get into URI, but admissions decisions at Brown/Yale differ. There are four possible outcomes:
  Case 1: Admitted to all schools
  Case 2: Admitted to URI and Brown only
  Case 3: Admitted to URI and Yale only
  Case 4: Admitted to URI only

Introduction to "Fixed Effects"

- What exactly does it mean that they control for the set of colleges you applied / were admitted to?

- Consider a simplified version where there are three schools, Brown, Yale, and URI, and everyone applies to all 3.

- Suppose all students in the sample get into URI, but admissions decisions at Brown/Yale differ. There are four possible outcomes:
  Case 1: Admitted to all schools
  Case 2: Admitted to URI and Brown only
  Case 3: Admitted to URI and Yale only
  Case 4: Admitted to URI only

- We can construct variables where $X_{i1} = 1$ if we're in case 1 and is 0 otherwise, $X_{i2} = 1$ if we're in case 2 and is 0 otherwise, etc.

Introduction to "Fixed Effects"

- What exactly does it mean that they control for the set of colleges you applied / were admitted to?

- Consider a simplified version where there are three schools, Brown, Yale, and URI, and everyone applies to all 3.

- Suppose all students in the sample get into URI, but admissions decisions at Brown/Yale differ. There are four possible outcomes:
  Case 1: Admitted to all schools
  Case 2: Admitted to URI and Brown only
  Case 3: Admitted to URI and Yale only
  Case 4: Admitted to URI only

- We can construct variables where $X_{i1} = 1$ if we're in case 1 and is 0 otherwise, $X_{i2} = 1$ if we're in case 2 and is 0 otherwise, etc.

- The variables $X_{i1}, ..., X_{i4}$ are often called "fixed effects" for the set of schools you were admitted to.

Introduction to "Fixed Effects"

- We can then approximate the CEF as

$$E[Y_i|\boldsymbol{X}_i, D_i] = D_i\beta_D + X_{i1}\beta_1 + X_{i2}\beta_2 + X_{i3}\beta_3 + X_{i4}\beta_4$$

Introduction to "Fixed Effects"

- We can then approximate the CEF as

$$E[Y_i|\boldsymbol{X}_i, D_i] = D_i\beta_D + X_{i1}\beta_1 + X_{i2}\beta_2 + X_{i3}\beta_3 + X_{i4}\beta_4$$

- This allows for a different avg outcome depending on which colleges you were admitted to, and the selectivity of your school (D_i)

Introduction to "Fixed Effects"

- We can then approximate the CEF as

$$E[Y_i|\boldsymbol{X}_i, D_i] = D_i\beta_D + X_{i1}\beta_1 + X_{i2}\beta_2 + X_{i3}\beta_3 + X_{i4}\beta_4$$

- This allows for a different avg outcome depending on which colleges you were admitted to, and the selectivity of your school (D_i)

- Intuitively, β_D represents the average difference from going to an elite school *among* students who got into the same set of schools

Applicant Group	Student	Private			Public			1996 Earnings
		School I	School II	School III	School IV	School V	School VI	
	1		Reject	**Admit**		Admit		110,000
A	2		Reject	**Admit**		Admit		100,000
	3		Reject	Admit		**Admit**		110,000
B	4	**Admit**			Admit		Admit	60,000
	5	Admit			Admit		**Admit**	30,000
C	6		**Admit**					115,000
	7		**Admit**					75,000
D	8	Reject			**Admit**	Admit		90,000
	9	Reject			Admit	**Admit**		60,000

- "Applicant groups" all applied + admitted to the same set of schools
- The school a student actually attended is highlighted

Controlling for Application Decisions in Practice

- In practice, Dale and Krueger don't have that many students who were applied/admitted to the exact same set of schools

Controlling for Application Decisions in Practice

- In practice, Dale and Krueger don't have that many students who were applied/admitted to the exact same set of schools
- As an approximation, they group colleges into bins based on average-SAT rounded to nearest 25 points

Controlling for Application Decisions in Practice

- In practice, Dale and Krueger don't have that many students who were applied/admitted to the exact same set of schools
- As an approximation, they group colleges into bins based on average-SAT rounded to nearest 25 points
- They then control for the set of colleges you applied/were admitted to based on these bins (e.g., X_{i1} might correspond to being rejected at a school w/SAT 1350 and accepted at 2 schools w/SAT 1250)

	Basic model: no selection controls		Matched-applicant model
	Full sample	Restricted sample	Similar school-SAT matches*
Variable	1	2	3
School-average SAT score/100	0.076	0.082	−0.016
	(0.016)	(0.014)	(0.022)
Predicted log(parental income)	0.187	0.190	0.163
	(0.024)	(0.033)	(0.033)
Own SAT score/100	0.018	0.006	−0.011
	(0.006)	(0.007)	(0.007)
Female	−0.403	−0.410	−0.395
	(0.015)	(0.018)	(0.024)
Black	−0.023	−0.026	−0.057
	(0.035)	(0.053)	(0.053)
Hispanic	0.015	0.070	0.020
	(0.052)	(0.076)	(0.099)
Asian	0.173	0.245	0.241
	(0.036)	(0.054)	(0.064)
Other/missing race	−0.188	−0.048	0.060
	(0.119)	(0.143)	(0.180)
High school top 10 percent	0.061	0.091	0.079
	(0.018)	(0.022)	(0.026)
High school rank missing	0.001	0.040	0.016
	(0.024)	(0.026)	(0.038)
Athlete	0.102	0.088	0.104
	(0.025)	(0.030)	(0.039)

- With application controls, we get $\hat{\beta} = -0.016$.

	Basic model: no selection controls		Matched-applicant model
	Full sample	Restricted sample	Similar school-SAT matches*
Variable	1	2	3
School-average SAT score/100	0.076 (0.016)	0.082 (0.014)	$\boxed{-0.016 \; (0.022)}$
Predicted log(parental income)	0.187 (0.024)	0.190 (0.033)	0.163 (0.033)
Own SAT score/100	0.018 (0.006)	0.006 (0.007)	−0.011 (0.007)
Female	−0.403 (0.015)	−0.410 (0.018)	−0.395 (0.024)
Black	−0.023 (0.035)	−0.026 (0.053)	−0.057 (0.053)
Hispanic	0.015 (0.052)	0.070 (0.076)	0.020 (0.099)
Asian	0.173 (0.036)	0.245 (0.054)	0.241 (0.064)
Other/missing race	−0.188 (0.119)	−0.048 (0.143)	0.060 (0.180)
High school top 10 percent	0.061 (0.018)	0.091 (0.022)	0.079 (0.026)
High school rank missing	0.001 (0.024)	0.040 (0.026)	0.016 (0.038)
Athlete	0.102 (0.025)	0.088 (0.030)	0.104 (0.039)

- With application controls, we get $\hat{\beta} = -0.016$.
- This indicates about a -1.6 log wage return to attending a school with 100 higher SAT points (but not significant!)

Evaluating Conditional Unconfoundedness (Again)

- Why might conditional unconfoundedness be violated here?

Evaluating Conditional Unconfoundedness (Again)

- Why might conditional unconfoundedness be violated here?
- A concern is that choice of where to go to college depends on unobservables that are correlated with earnings, even conditional on application choice
  - E.g, students who choose to go to a more selective school could have different career ambitions (e.g. industry versus academia)

Evaluating Conditional Unconfoundedness (Again)

- Why might conditional unconfoundedness be violated here?
- A concern is that choice of where to go to college depends on unobservables that are correlated with earnings, even conditional on application choice
    - E.g, students who choose to go to a more selective school could have different career ambitions (e.g. industry versus academia)
    - Students may choose to go to a lower-ranked school only if it has a particularly good program in what they're interested in

Evaluating Conditional Unconfoundedness (Again)

- Why might conditional unconfoundedness be violated here?
- A concern is that choice of where to go to college depends on unobservables that are correlated with earnings, even conditional on application choice
  - E.g, students who choose to go to a more selective school could have different career ambitions (e.g. industry versus academia)
  - Students may choose to go to a lower-ranked school only if it has a particularly good program in what they're interested in
- It's also important to realize that the schools in the C&B study tend to be selective. These results can at best be interpreted as the causal effect between attending a selective school and highly-selective school

Omitted Variables Bias

- In the C&B example (and many other applications), we might be worried that we didn't condition on all of the necessary variables for conditional unconfoundedness to hold

Omitted Variables Bias

- In the C&B example (and many other applications), we might be worried that we didn't condition on all of the necessary variables for conditional unconfoundedness to hold
- How will the coefficients we estimate be biased if we forget to include some variables?

Omitted Variables Bias

- In the C&B example (and many other applications), we might be worried that we didn't condition on all of the necessary variables for conditional unconfoundedness to hold
- How will the coefficients we estimate be biased if we forget to include some variables?
- To answer this question, we will derive what is called the **omitted variable bias** (OVB) formula

Omitted Variable Bias – Simplest Case

- Suppose that conditional unconfoundedness holds conditional on X_{i1} (e.g. SAT score).

Omitted Variable Bias – Simplest Case

- Suppose that conditional unconfoundedness holds conditional on X_{i1} (e.g. SAT score). We would like to estimate the regression

$$Y_i = \beta_0 + \beta_D D_i + \beta_1 X_{i1} + e_i$$

  The population coefficient β_D approximates the ATE (assuming this is a good approx to the CEF).

- Now, suppose we don't include X_{i1} and just act as if D_i is randomly assigned.

Omitted Variable Bias – Simplest Case

- Suppose that conditional unconfoundedness holds conditional on X_{i1} (e.g. SAT score). We would like to estimate the regression

$$Y_i = \beta_0 + \beta_D D_i + \beta_1 X_{i1} + e_i$$

  The population coefficient β_D approximates the ATE (assuming this is a good approx to the CEF).

- Now, suppose we don't include X_{i1} and just act as if D_i is randomly assigned. What will be the bias of our estimates if we instead estimate the regression $Y_i = \tilde{\beta}_0 + \tilde{\beta}_D D_i + \varepsilon_i$?

Omitted Variable Bias – Simplest Case

- Suppose that conditional unconfoundedness holds conditional on X_{i1} (e.g. SAT score). We would like to estimate the regression

$$Y_i = \beta_0 + \beta_D D_i + \beta_1 X_{i1} + e_i$$

  The population coefficient β_D approximates the ATE (assuming this is a good approx to the CEF).

- Now, suppose we don't include X_{i1} and just act as if D_i is randomly assigned. What will be the bias of our estimates if we instead estimate the regression $Y_i = \tilde{\beta}_0 + \tilde{\beta}_D D_i + \varepsilon_i$?

$$\tilde{\beta}_D = \frac{Cov(Y_i, D_i)}{Var(D_i)}$$

Omitted Variable Bias – Simplest Case

- Suppose that conditional unconfoundedness holds conditional on X_{i1} (e.g. SAT score). We would like to estimate the regression

$$Y_i = \beta_0 + \beta_D D_i + \beta_1 X_{i1} + e_i$$

  The population coefficient β_D approximates the ATE (assuming this is a good approx to the CEF).

- Now, suppose we don't include X_{i1} and just act as if D_i is randomly assigned. What will be the bias of our estimates if we instead estimate the regression $Y_i = \tilde{\beta}_0 + \tilde{\beta}_D D_i + \varepsilon_i$?

$$\tilde{\beta}_D = \frac{Cov(Y_i, D_i)}{Var(D_i)} = \frac{Cov(\beta_0 + \beta_D D_i + \beta_1 X_{i1} + e_i, D_i)}{Var(D_i)}$$

Omitted Variable Bias – Simplest Case

- Suppose that conditional unconfoundedness holds conditional on X_{i1} (e.g. SAT score). We would like to estimate the regression

$$Y_i = \beta_0 + \beta_D D_i + \beta_1 X_{i1} + e_i$$

  The population coefficient β_D approximates the ATE (assuming this is a good approx to the CEF).

- Now, suppose we don't include X_{i1} and just act as if D_i is randomly assigned. What will be the bias of our estimates if we instead estimate the regression $Y_i = \tilde{\beta}_0 + \tilde{\beta}_D D_i + \varepsilon_i$?

$$\tilde{\beta}_D = \frac{Cov(Y_i, D_i)}{Var(D_i)} = \frac{Cov(\beta_0 + \beta_D D_i + \beta_1 X_{i1} + e_i, D_i)}{Var(D_i)}$$
$$= \frac{\beta_D Cov(D_i, D_i) + \beta_1 Cov(X_{i1}, D_i) + Cov(e_i, D_i)}{Var(D_i)}$$

Omitted Variable Bias – Simplest Case

- Suppose that conditional unconfoundedness holds conditional on X_{i1} (e.g. SAT score). We would like to estimate the regression

$$Y_i = \beta_0 + \beta_D D_i + \beta_1 X_{i1} + e_i$$

The population coefficient β_D approximates the ATE (assuming this is a good approx to the CEF).

- Now, suppose we don't include X_{i1} and just act as if D_i is randomly assigned. What will be the bias of our estimates if we instead estimate the regression $Y_i = \tilde{\beta}_0 + \tilde{\beta}_D D_i + \varepsilon_i$?

$$\begin{aligned}
\tilde{\beta}_D &= \frac{Cov(Y_i, D_i)}{Var(D_i)} = \frac{Cov(\beta_0 + \beta_D D_i + \beta_1 X_{i1} + e_i, D_i)}{Var(D_i)} \\
&= \frac{\beta_D Cov(D_i, D_i) + \beta_1 Cov(X_{i1}, D_i) + Cov(e_i, D_i)}{Var(D_i)} \\
&= \beta_D + \beta_1 \frac{Cov(X_{i1}, D_i)}{Var(D_i)}
\end{aligned}$$

where we use $E[e_i] = E[D_i e_i] = 0$ from the FOCs for regression

Evaluating OVB

- Thus, the (population) regression $Y_i = \tilde{\beta}_0 + \tilde{\beta}_D D_i + \varepsilon_i$ yields

$$\tilde{\beta} = \beta_D + \beta_1 \frac{Cov(X_{i1}, D_i)}{Var(D_i)}$$

where β_D is the coefficient we wanted (our approx to the ATE)

Evaluating OVB

- Thus, the (population) regression $Y_i = \tilde{\beta}_0 + \tilde{\beta}_D D_i + \varepsilon_i$ yields

$$\tilde{\beta} = \beta_D + \beta_1 \frac{Cov(X_{i1}, D_i)}{Var(D_i)}$$

  where β_D is the coefficient we wanted (our approx to the ATE)

- The bias depends on two terms, β_1 and $\gamma_D = \frac{Cov(X_{i1}, D_i)}{Var(D_i)}$

Evaluating OVB

- Thus, the (population) regression $Y_i = \tilde{\beta}_0 + \tilde{\beta}_D D_i + \varepsilon_i$ yields

$$\tilde{\beta} = \beta_D + \beta_1 \frac{Cov(X_{i1}, D_i)}{Var(D_i)}$$

  where β_D is the coefficient we wanted (our approx to the ATE)

- The bias depends on two terms, β_1 and $\gamma_D = \frac{Cov(X_{i1}, D_i)}{Var(D_i)}$

- Recall that $Y_i = \beta_0 + \beta_D D_i + \beta_1 X_{i1} + e_i$
  So β_1 is large when X_{i1} is predictive of Y_i, given D_i

Evaluating OVB

- Thus, the (population) regression $Y_i = \tilde{\beta}_0 + \tilde{\beta}_D D_i + \varepsilon_i$ yields

$$\tilde{\beta} = \beta_D + \beta_1 \frac{Cov(X_{i1}, D_i)}{Var(D_i)}$$

  where β_D is the coefficient we wanted (our approx to the ATE)

- The bias depends on two terms, β_1 and $\gamma_D = \frac{Cov(X_{i1}, D_i)}{Var(D_i)}$

- Recall that $Y_i = \beta_0 + \beta_D D_i + \beta_1 X_{i1} + e_i$
  So β_1 is large when X_{i1} is predictive of Y_i, given D_i

- Observe that γ_D is the slope coefficient from regressing X_{i1} on D_i
  So γ_D is large when D_i is strongly correlated with X_{i1}

Evaluating OVB

- Thus, the (population) regression $Y_i = \tilde{\beta}_0 + \tilde{\beta}_D D_i + \varepsilon_i$ yields

$$\tilde{\beta} = \beta_D + \beta_1 \frac{Cov(X_{i1}, D_i)}{Var(D_i)}$$

  where β_D is the coefficient we wanted (our approx to the ATE)

- The bias depends on two terms, β_1 and $\gamma_D = \frac{Cov(X_{i1}, D_i)}{Var(D_i)}$

- Recall that $Y_i = \beta_0 + \beta_D D_i + \beta_1 X_{i1} + e_i$
  So β_1 is large when X_{i1} is predictive of Y_i, given D_i

- Observe that γ_D is the slope coefficient from regressing X_{i1} on D_i
  So γ_D is large when D_i is strongly correlated with X_{i1}

- Hence $\tilde{\beta}$ will be very biased for β_D if the omitted variable X_{i1} is both highly correlated with Y_i and highly correlated with D_i

Evaluating OVB

- Thus, the (population) regression $Y_i = \tilde{\beta}_0 + \tilde{\beta}_D D_i + \varepsilon_i$ yields

$$\tilde{\beta} = \beta_D + \beta_1 \frac{Cov(X_{i1}, D_i)}{Var(D_i)}$$

  where β_D is the coefficient we wanted (our approx to the ATE)

- The bias depends on two terms, β_1 and $\gamma_D = \frac{Cov(X_{i1}, D_i)}{Var(D_i)}$
- Recall that $Y_i = \beta_0 + \beta_D D_i + \beta_1 X_{i1} + e_i$
  So β_1 is large when X_{i1} is predictive of Y_i, given D_i
- Observe that γ_D is the slope coefficient from regressing X_{i1} on D_i
  So γ_D is large when D_i is strongly correlated with X_{i1}
- Hence $\tilde{\beta}$ will be very biased for β_D if the omitted variable X_{i1} is both highly correlated with Y_i and highly correlated with D_i
  - On the flip side, if either $\beta_1 = 0$ or $\gamma_D =$ then we have no OVB!

OVB Formula in Finite Samples

- We just showed that the coefficients from the population regressions

$$Y_i = \beta_0 + \beta_D D_i + \beta_1 X_i + e_i \tag{1}$$

$$Y_i = \tilde{\beta}_0 + \tilde{\beta}_D D_i + \varepsilon_i \tag{2}$$

are related by the OVB formula

$$\tilde{\beta} = \beta_D + \beta_1 \frac{Cov(X_{i1}, D_i)}{Var(D_i)}$$

OVB Formula in Finite Samples

- We just showed that the coefficients from the population regressions

$$Y_i = \beta_0 + \beta_D D_i + \beta_1 X_i + e_i \tag{1}$$
$$Y_i = \tilde{\beta}_0 + \tilde{\beta}_D D_i + \varepsilon_i \tag{2}$$

  are related by the OVB formula

$$\tilde{\beta} = \beta_D + \beta_1 \frac{Cov(X_{i1}, D_i)}{Var(D_i)}$$

- It turns out that the OLS estimates for these two regressions have the same relationship

$$\hat{\tilde{\beta}} = \hat{\beta}_D + \hat{\beta}_1 \frac{\widehat{Cov}(X_{i1}, D_i)}{\widehat{Var}(D_i)}$$

OVB Illustration

- Angrist and Pischke's textbook considers a modified version of Dale and Krueger where D is whether one attends a private college. Suppose X_{i1} is a student's SAT score

OVB Illustration

- Angrist and Pischke's textbook considers a modified version of Dale and Krueger where D is whether one attends a private college. Suppose X_{i1} is a student's SAT score

- Angrist and Pischke estimate the regression

$$Y_i = D_i\beta_D + X_{i1}\beta_1 + \varepsilon_i$$

OVB Illustration

- Angrist and Pischke's textbook considers a modified version of Dale and Krueger where D is whether one attends a private college. Suppose X_{i1} is a student's SAT score
- Angrist and Pischke estimate the regression

$$Y_i = D_i\beta_D + X_{i1}\beta_1 + \varepsilon_i$$

- If conditional unconfoundedness holds conditional on X_{i1} (and the CEF is approximately linear), the coefficient β_D will correspond with the causal effect of attending private school

OVB Illustration

- Angrist and Pischke's textbook considers a modified version of Dale and Krueger where D is whether one attends a private college. Suppose X_{i1} is a student's SAT score

- Angrist and Pischke estimate the regression

$$Y_i = D_i \beta_D + X_{i1} \beta_1 + \varepsilon_i$$

- If conditional unconfoundedness holds conditional on X_{i1} (and the CEF is approximately linear), the coefficient β_D will correspond with the causal effect of attending private school

- Let's think about what would happen if we forgot to control for SAT

- Here are the results that A&P get when controlling for SAT:

Variable	Coefficient	SE
Private school ($\hat{\beta}_D$)	0.095	0.052
SAT score /100 ($\hat{\beta}_1$)	0.048	0.009
Constant	[...]	[...]

- What would happen if we omitted the control for SAT score from the regression? How would $\hat{\beta}_D$ change?

- Here are the results that A&P get when controlling for SAT:

Variable	Coefficient	SE
Private school ($\hat{\beta}_D$)	0.095	0.052
SAT score /100 ($\hat{\beta}_1$)	0.048	0.009
Constant	[...]	[...]

- What would happen if we omitted the control for SAT score from the regression? How would $\hat{\beta}_D$ change?

- To compute this, we need to know the coefficients from the regression of X_{i1} (SAT score/100) on D_i (Private school):

- Here are the results that A&P get when controlling for SAT:

Variable	Coefficient	SE
Private school ($\hat{\beta}_D$)	0.095	0.052
SAT score /100 ($\hat{\beta}_1$)	0.048	0.009
Constant	[...]	[...]

- What would happen if we omitted the control for SAT score from the regression? How would $\hat{\beta}_D$ change?

- To compute this, we need to know the coefficients from the regression of X_{i1} (SAT score/100) on D_i (Private school):

Variable	Coefficient
Private school ($\hat{\gamma}_D$)	.83
Constant	[...]

- Here are the results that A&P get when controlling for SAT:

Variable	Coefficient	SE
Private school ($\hat{\beta}_D$)	0.095	0.052
SAT score /100 ($\hat{\beta}_1$)	0.048	0.009
Constant	[...]	[...]

- What would happen if we omitted the control for SAT score from the regression? How would $\hat{\beta}_D$ change?

- To compute this, we need to know the coefficients from the regression of X_{i1} (SAT score/100) on D_i (Private school):

Variable	Coefficient
Private school ($\hat{\gamma}_D$)	.83
Constant	[...]

- Thus, if we omitted X_{i1} our estimated coefficient on private school would be $\hat{\gamma}_D \times \hat{\beta}_1 = 0.83 \times 0.048 \approx 0.04$ larger.

- Indeed, if we actually run the regression omitting SAT scores, we see that the coefficient on private school is 0.04 larger.

- Results including SAT score:

Variable	Coefficient	SE
Private school (β_D)	0.095	0.052
SAT score /100 (β_1)	0.048	0.009
Constant	[...]	[...]

- Results excluding SAT score:

Variable	Coefficient	SE
Private school ($\tilde{\beta}$)	0.135	0.055
Constant	[...]	[...]

- Indeed, if we actually run the regression omitting SAT scores, we see that the coefficient on private school is 0.04 larger.

- Results including SAT score:

Variable	Coefficient	SE
Private school (β_D)	0.095	0.052
SAT score /100 (β_1)	0.048	0.009
Constant	[...]	[...]

- Results excluding SAT score:

Variable	Coefficient	SE
Private school ($\tilde{\beta}$)	0.135	0.055
Constant	[...]	[...]

- When would omitting SAT score matter more?

- Indeed, if we actually run the regression omitting SAT scores, we see that the coefficient on private school is 0.04 larger.

- Results including SAT score:

Variable	Coefficient	SE
Private school (β_D)	0.095	0.052
SAT score /100 (β_1)	0.048	0.009
Constant	[...]	[...]

- Results excluding SAT score:

Variable	Coefficient	SE
Private school ($\tilde{\beta}$)	0.135	0.055
Constant	[...]	[...]

- When would omitting SAT score matter more?
  - If SAT score were more strongly related to earnings ($\hat{\beta}_1$ larger)

- Indeed, if we actually run the regression omitting SAT scores, we see that the coefficient on private school is 0.04 larger.

- Results including SAT score:

Variable	Coefficient	SE
Private school (β_D)	0.095	0.052
SAT score /100 (β_1)	0.048	0.009
Constant	[...]	[...]

- Results excluding SAT score:

Variable	Coefficient	SE
Private school ($\tilde{\beta}$)	0.135	0.055
Constant	[...]	[...]

- When would omitting SAT score matter more?
  - If SAT score were more strongly related to earnings ($\hat{\beta}_1$ larger)
  - If treatment were more strongly correlated with SAT scores ($\hat{\gamma}_D$ larger)

Omitted Variable Bias Formula - Multiple Variables

- Now, suppose we observe Y_i, D_i, and X_{i1}, but unconfoundedness holds only conditional on $\boldsymbol{X}_i = (X_{i1}, X_{i2})'$.

Omitted Variable Bias Formula - Multiple Variables

- Now, suppose we observe Y_i, D_i, and X_{i1}, but unconfoundedness holds only conditional on $\boldsymbol{X}_i = (X_{i1}, X_{i2})'$. E.g. Y_i could be earnings, D_i college selectivity, X_{i1} HS GPA, and X_{i2} SAT score

Omitted Variable Bias Formula - Multiple Variables

- Now, suppose we observe Y_i, D_i, and X_{i1}, but unconfoundedness holds only conditional on $\boldsymbol{X}_i = (X_{i1}, X_{i2})'$. E.g. Y_i could be earnings, D_i college selectivity, X_{i1} HS GPA, and X_{i2} SAT score

- We would like to estimate the regression $Y_i = \beta_0 + \beta_D D_i + \beta_1 X_{i1} + \beta_2 X_{i2} + e_i$

Omitted Variable Bias Formula - Multiple Variables

- Now, suppose we observe Y_i, D_i, and X_{i1}, but unconfoundedness holds only conditional on $\boldsymbol{X}_i = (X_{i1}, X_{i2})'$. E.g. Y_i could be earnings, D_i college selectivity, X_{i1} HS GPA, and X_{i2} SAT score

- We would like to estimate the regression
  $Y_i = \beta_0 + \beta_D D_i + \beta_1 X_{i1} + \beta_2 X_{i2} + e_i$

- But since we don't observe X_{i1}, we instead estimate the regression
  $Y_i = \tilde{\beta}_0 + \tilde{\beta}_D + \tilde{\beta}_1 X_1 + \varepsilon_i$

Omitted Variable Bias Formula - Multiple Variables

- Now, suppose we observe Y_i, D_i, and X_{i1}, but unconfoundedness holds only conditional on $\boldsymbol{X}_i = (X_{i1}, X_{i2})'$. E.g. Y_i could be earnings, D_i college selectivity, X_{i1} HS GPA, and X_{i2} SAT score

- We would like to estimate the regression $Y_i = \beta_0 + \beta_D D_i + \beta_1 X_{i1} + \beta_2 X_{i2} + e_i$

- But since we don't observe X_{i1}, we instead estimate the regression $Y_i = \tilde{\beta}_0 + \tilde{\beta}_D + \tilde{\beta}_1 X_1 + \varepsilon_i$

- How does $\tilde{\beta}_D$ relate to β_D?

Omitted Variable Bias Formula - Multiple Variables

- Now, suppose we observe Y_i, D_i, and X_{i1}, but unconfoundedness holds only conditional on $\boldsymbol{X}_i = (X_{i1}, X_{i2})'$. E.g. Y_i could be earnings, D_i college selectivity, X_{i1} HS GPA, and X_{i2} SAT score

- We would like to estimate the regression $Y_i = \beta_0 + \beta_D D_i + \beta_1 X_{i1} + \beta_2 X_{i2} + e_i$

- But since we don't observe X_{i1}, we instead estimate the regression $Y_i = \tilde{\beta}_0 + \tilde{\beta}_D + \tilde{\beta}_1 X_1 + \varepsilon_i$

- How does $\tilde{\beta}_D$ relate to β_D?

- Answer:

$$\tilde{\beta}_D = \beta_D + \beta_2 \gamma_D$$

  where γ_D is the coefficient from the regression

$$X_{i2} = \gamma_0 + \gamma_D D_i + \gamma_1 X_{i1} + u_i$$

- This is similar to the OVB formula we had from before!

Evaluating the Bias (Again)

- The multivariate OVB formula is $\tilde{\beta}_D = \beta_D + \beta_2\gamma_D$ where β_D is the coefficient we wanted (if we controlled for X_{i2}).

Evaluating the Bias (Again)

- The multivariate OVB formula is $\tilde{\beta}_D = \beta_D + \beta_2 \gamma_D$ where β_D is the coefficient we wanted (if we controlled for X_{i2}). As before, the bias is the product of two terms: β_2 and γ_D

Evaluating the Bias (Again)

- The multivariate OVB formula is $\tilde{\beta}_D = \beta_D + \beta_2 \gamma_D$ where β_D is the coefficient we wanted (if we controlled for X_{i2}). As before, the bias is the product of two terms: β_2 and γ_D

- Remember that $Y_i = \beta_0 + \beta_D D_i + \beta_1 X_{i1} + \beta_2 X_{i2} + e_i$.
  $\Longrightarrow$ So β_2 is large when X_{i2} is strongly correlated with Y_i, after controlling for X_{i1} and D_{i1}.

Evaluating the Bias (Again)

- The multivariate OVB formula is $\tilde{\beta}_D = \beta_D + \beta_2 \gamma_D$ where β_D is the coefficient we wanted (if we controlled for X_{i2}). As before, the bias is the product of two terms: β_2 and γ_D

- Remember that $Y_i = \beta_0 + \beta_D D_i + \beta_1 X_{i1} + \beta_2 X_{i2} + e_i$.
  $\implies$ So β_2 is large when X_{i2} is strongly correlated with Y_i, after controlling for X_{i1} and D_{i1}.

- Remember that $X_{i2} = \gamma_0 + \gamma_D D_i + \gamma_1 X_{i1} + u_i$
  $\implies$ So γ_D is large when D_i is strongly correlated with X_{i2}, after controlling for X_{i1}

Evaluating the Bias (Again)

- The multivariate OVB formula is $\tilde{\beta}_D = \beta_D + \beta_2 \gamma_D$ where β_D is the coefficient we wanted (if we controlled for X_{i2}). As before, the bias is the product of two terms: β_2 and γ_D

- Remember that $Y_i = \beta_0 + \beta_D D_i + \beta_1 X_{i1} + \beta_2 X_{i2} + e_i$.
  $\Longrightarrow$ So β_2 is large when X_{i2} is strongly correlated with Y_i, after controlling for X_{i1} and D_{i1}.

- Remember that $X_{i2} = \gamma_0 + \gamma_D D_i + \gamma_1 X_{i1} + u_i$
  $\Longrightarrow$ So γ_D is large when D_i is strongly correlated with X_{i2}, after controlling for X_{i1}

- In summary: $\tilde{\beta}_D$ will be very biased for β_D if the omitted variable X_{i2} is both correlated with the outcome Y_i given the treatment D_i and correlated with D_i, both after controlling for X_{i1}

Evaluating the Bias (Again)

- The multivariate OVB formula is $\tilde{\beta}_D = \beta_D + \beta_2\gamma_D$ where β_D is the coefficient we wanted (if we controlled for X_{i2}). As before, the bias is the product of two terms: β_2 and γ_D

- Remember that $Y_i = \beta_0 + \beta_D D_i + \beta_1 X_{i1} + \beta_2 X_{i2} + e_i$.
  $\implies$ So β_2 is large when X_{i2} is strongly correlated with Y_i, after controlling for X_{i1} and D_{i1}.

- Remember that $X_{i2} = \gamma_0 + \gamma_D D_i + \gamma_1 X_{i1} + u_i$
  $\implies$ So γ_D is large when D_i is strongly correlated with X_{i2}, after controlling for X_{i1}

- In summary: $\tilde{\beta}_D$ will be very biased for β_D if the omitted variable X_{i2} is both correlated with the outcome Y_i given the treatment D_i and correlated with D_i, both after controlling for X_{i1}
  - If either correlation is zero, OVB is zero

Evaluating the Bias (Again)

- The multivariate OVB formula is $\tilde{\beta}_D = \beta_D + \beta_2\gamma_D$ where β_D is the coefficient we wanted (if we controlled for X_{i2}). As before, the bias is the product of two terms: β_2 and γ_D

- Remember that $Y_i = \beta_0 + \beta_D D_i + \beta_1 X_{i1} + \beta_2 X_{i2} + e_i$.
  $\implies$ So β_2 is large when X_{i2} is strongly correlated with Y_i, after controlling for X_{i1} and D_{i1}.

- Remember that $X_{i2} = \gamma_0 + \gamma_D D_i + \gamma_1 X_{i1} + u_i$
  $\implies$ So γ_D is large when D_i is strongly correlated with X_{i2}, after controlling for X_{i1}

- In summary: $\tilde{\beta}_D$ will be very biased for β_D if the omitted variable X_{i2} is both correlated with the outcome Y_i given the treatment D_i and correlated with D_i, both after controlling for X_{i1}
  - If either correlation is zero, OVB is zero
  - OVB is positive if and only if the correlations are the same sign

Illustration of Omitted Variable Bias

- Angrist and Pischke's textbook considers a modified version of Dale and Krueger where D_i is whether one attends a private college

Illustration of Omitted Variable Bias

- Angrist and Pischke's textbook considers a modified version of Dale and Krueger where D_i is whether one attends a private college

- Suppose X_{i2} is a student's SAT score, and $\boldsymbol{X}_{i1}$ is a vector containing indicators for the set of colleges you were admitted to.

Illustration of Omitted Variable Bias

- Angrist and Pischke's textbook considers a modified version of Dale and Krueger where D_i is whether one attends a private college
- Suppose X_{i2} is a student's SAT score, and $\boldsymbol{X}_{i1}$ is a vector containing indicators for the set of colleges you were admitted to.
- Angrist and Pischke estimate the regression

$$Y_i = D_i\beta_D + \boldsymbol{X}_{i1}'\boldsymbol{\beta}_1 + X_{i2}\beta_2 + \varepsilon_i$$

Illustration of Omitted Variable Bias

- Angrist and Pischke's textbook considers a modified version of Dale and Krueger where D_i is whether one attends a private college
- Suppose X_{i2} is a student's SAT score, and $\boldsymbol{X}_{i1}$ is a vector containing indicators for the set of colleges you were admitted to.
- Angrist and Pischke estimate the regression

$$Y_i = D_i\beta_D + \boldsymbol{X}_{i1}'\boldsymbol{\beta}_1 + X_{i2}\beta_2 + \varepsilon_i$$

- If conditional unconfoundedness holds conditional on $\boldsymbol{X}_{i1}$ and X_{i2} (and the CEF is approximately linear), the coefficient β_D will correspond to the causal effect of attending private school

Illustration of Omitted Variable Bias

- Angrist and Pischke's textbook considers a modified version of Dale and Krueger where D_i is whether one attends a private college

- Suppose X_{i2} is a student's SAT score, and $\boldsymbol{X}_{i1}$ is a vector containing indicators for the set of colleges you were admitted to.

- Angrist and Pischke estimate the regression

$$Y_i = D_i \beta_D + \boldsymbol{X}_{i1}' \boldsymbol{\beta}_1 + X_{i2} \beta_2 + \varepsilon_i$$

- If conditional unconfoundedness holds conditional on $\boldsymbol{X}_{i1}$ and X_{i2} (and the CEF is approximately linear), the coefficient β_D will correspond to the causal effect of attending private school

- Let's think about what happens if we forgot the control for SAT score

- Here are the results that A&P get when controlling for both SAT score and set of schools you're admitted to:

Variable	Coefficient	SE
Private school ($\hat{\beta}_D$)	0.003	0.039
SAT score /100 ($\hat{\beta}_2$)	0.033	0.007
College admitted ($\hat{\boldsymbol{\beta}}_1$)	[...]	[...]

- What would happen if we omitted the control for SAT score from the regression? How would $\hat{\beta}_D$ change?

- Here are the results that A&P get when controlling for both SAT score and set of schools you're admitted to:

Variable	Coefficient	SE
Private school ($\hat{\beta}_D$)	0.003	0.039
SAT score /100 ($\hat{\beta}_2$)	0.033	0.007
College admitted ($\hat{\boldsymbol{\beta}}_1$)	[...]	[...]

- What would happen if we omitted the control for SAT score from the regression? How would $\hat{\beta}_D$ change?

- To compute this, we need to know the coefficients from the regression of X_{i2} (SAT score/100) on D_i (Private schoool) and $\boldsymbol{X}_{i1}$ (College admitted):

- Here are the results that A&P get when controlling for both SAT score and set of schools you're admitted to:

Variable	Coefficient	SE
Private school ($\hat{\beta}_D$)	0.003	0.039
SAT score /100 ($\hat{\beta}_2$)	0.033	0.007
College admitted ($\hat{\boldsymbol{\beta}}_1$)	[...]	[...]

- What would happen if we omitted the control for SAT score from the regression? How would $\hat{\beta}_D$ change?

- To compute this, we need to know the coefficients from the regression of X_{i2} (SAT score/100) on D_i (Private schoool) and $\boldsymbol{X}_{i1}$ (College admitted):

Variable	Coefficient
Private school ($\hat{\gamma}_D$)	.12
College admitted ($\hat{\boldsymbol{\gamma}}_1$)	[...]

- Here are the results that A&P get when controlling for both SAT score and set of schools you're admitted to:

Variable	Coefficient	SE
Private school ($\hat{\beta}_D$)	0.003	0.039
SAT score /100 ($\hat{\beta}_2$)	0.033	0.007
College admitted ($\hat{\boldsymbol{\beta}}_1$)	[...]	[...]

- What would happen if we omitted the control for SAT score from the regression? How would $\hat{\beta}_D$ change?

- To compute this, we need to know the coefficients from the regression of X_{i2} (SAT score/100) on D_i (Private schoool) and $\boldsymbol{X}_{i1}$ (College admitted):

Variable	Coefficient
Private school ($\hat{\gamma}_D$)	.12
College admitted ($\hat{\boldsymbol{\gamma}}_1$)	[...]

- Omitting X_{i2} leads to a change of $\hat{\gamma}_D \times \hat{\beta}_2$, so if we omitted X_{i2} our estimated coefficient would be

- Here are the results that A&P get when controlling for both SAT score and set of schools you're admitted to:

Variable	Coefficient	SE
Private school ($\hat{\beta}_D$)	0.003	0.039
SAT score /100 ($\hat{\beta}_2$)	0.033	0.007
College admitted ($\hat{\boldsymbol{\beta}}_1$)	[...]	[...]

- What would happen if we omitted the control for SAT score from the regression? How would $\hat{\beta}_D$ change?

- To compute this, we need to know the coefficients from the regression of X_{i2} (SAT score/100) on D_i (Private schoool) and $\boldsymbol{X}_{i1}$ (College admitted):

Variable	Coefficient
Private school ($\hat{\gamma}_D$)	.12
College admitted ($\hat{\boldsymbol{\gamma}}_1$)	[...]

- Omitting X_{i2} leads to a change of $\hat{\gamma}_D \times \hat{\beta}_2$, so if we omitted X_{i2} our estimated coefficient would be $0.033 \times .12 \approx 0.004$ larger.

- Indeed, if we actually run the regression omitting SAT scores, we see that the coefficient on private school is 0.004 larger.

- Results including SAT score:

Variable	Coefficient	SE
Private school (β_D)	0.003	0.039
SAT score /100 (β_2)	0.033	0.007
College admitted ($\boldsymbol{\beta}_1$)	[...]	[...]

- Results excluding SAT score:

Variable	Coefficient	SE
Private school ($\tilde{\beta}_D$)	0.007	0.038
College admitted ($\tilde{\boldsymbol{\beta}}_X$)	[...]	[...]

- When would omitting SAT score matter more?

- Indeed, if we actually run the regression omitting SAT scores, we see that the coefficient on private school is 0.004 larger.

- Results including SAT score:

Variable	Coefficient	SE
Private school (β_D)	0.003	0.039
SAT score /100 (β_2)	0.033	0.007
College admitted ($\boldsymbol{\beta}_1$)	[...]	[...]

- Results excluding SAT score:

Variable	Coefficient	SE
Private school ($\tilde{\beta}_D$)	0.007	0.038
College admitted ($\tilde{\boldsymbol{\beta}}_X$)	[...]	[...]

- When would omitting SAT score matter more?
  - If SAT score were more strongly related to earnings (after controlling for College Admitted), i.e. $\hat{\beta}_2$ were larger

- Indeed, if we actually run the regression omitting SAT scores, we see that the coefficient on private school is 0.004 larger.

- Results including SAT score:

Variable	Coefficient	SE
Private school (β_D)	0.003	0.039
SAT score /100 (β_2)	0.033	0.007
College admitted ($\boldsymbol{\beta}_1$)	[...]	[...]

- Results excluding SAT score:

Variable	Coefficient	SE
Private school ($\tilde{\beta}_D$)	0.007	0.038
College admitted ($\tilde{\boldsymbol{\beta}}_X$)	[...]	[...]

- When would omitting SAT score matter more?
    - If SAT score were more strongly related to earnings (after controlling for College Admitted), i.e. $\hat{\beta}_2$ were larger
    - If treatment were more strongly correlated with SAT score (after controlling for College Admitted), i.e. $\hat{\gamma}_D$ were larger

Modeling Heterogeneous Treatment Effects

- Often we are interested in whether treatment effects are heterogeneous — e.g., is the effect of attending elite college different for students from richer families versus poorer families?

Modeling Heterogeneous Treatment Effects

- Often we are interested in whether treatment effects are heterogeneous — e.g., is the effect of attending elite college different for students from richer families versus poorer families?
- Remember that under conditional unconfoundedness,

$$CATE(\boldsymbol{x}) = E[Y_i|D_i = 1, \boldsymbol{X}_i = \boldsymbol{x}] - E[Y_i|D_i = 0, \boldsymbol{X}_i = \boldsymbol{x}]$$

Modeling Heterogeneous Treatment Effects

- Often we are interested in whether treatment effects are heterogeneous — e.g., is the effect of attending elite college different for students from richer families versus poorer families?

- Remember that under conditional unconfoundedness,

$$CATE(\boldsymbol{x}) = E[Y_i|D_i = 1, \boldsymbol{X}_i = \boldsymbol{x}] - E[Y_i|D_i = 0, \boldsymbol{X}_i = \boldsymbol{x}]$$

- Suppose we approximate

$$E[Y_i|D_i = d, \boldsymbol{X}_i = \boldsymbol{x}] \approx \beta_D d + \beta_{DX}(x_1 \times d) + \boldsymbol{x}'\boldsymbol{\gamma}$$

Modeling Heterogeneous Treatment Effects

- Often we are interested in whether treatment effects are heterogeneous — e.g., is the effect of attending elite college different for students from richer families versus poorer families?

- Remember that under conditional unconfoundedness,

$$CATE(\boldsymbol{x}) = E[Y_i|D_i = 1, \boldsymbol{X}_i = \boldsymbol{x}] - E[Y_i|D_i = 0, \boldsymbol{X}_i = \boldsymbol{x}]$$

- Suppose we approximate

$$E[Y_i|D_i = d, \boldsymbol{X}_i = \boldsymbol{x}] \approx \beta_D d + \beta_{DX}(x_1 \times d) + \boldsymbol{x}'\boldsymbol{\gamma}$$

- Then

$$CATE(\boldsymbol{x}) \approx \beta_D + \beta_{DX} x_1,$$

  So the average treatment effect for someone with $X_{i1} = x_1$ is approximately $\beta_D + \beta_{DX} x_1$.

Modeling Heterogeneous Treatment Effects

- Often we are interested in whether treatment effects are heterogeneous — e.g., is the effect of attending elite college different for students from richer families versus poorer families?

- Remember that under conditional unconfoundedness,

$$CATE(\boldsymbol{x}) = E[Y_i|D_i = 1, \boldsymbol{X}_i = \boldsymbol{x}] - E[Y_i|D_i = 0, \boldsymbol{X}_i = \boldsymbol{x}]$$

- Suppose we approximate

$$E[Y_i|D_i = d, \boldsymbol{X}_i = \boldsymbol{x}] \approx \beta_D d + \beta_{DX}(x_1 \times d) + \boldsymbol{x}'\boldsymbol{\gamma}$$

- Then

$$CATE(\boldsymbol{x}) \approx \beta_D + \beta_{DX} x_1,$$

  So the average treatment effect for someone with $X_{i1} = x_1$ is approximately $\beta_D + \beta_{DX} x_1$.

- If we are interested in heterogeneity by X_{i1}, we can estimate:

$$Y_i = \beta_D D_i + \beta_{DX}(D_i \times X_{i1}) + \boldsymbol{\gamma}'\boldsymbol{X}_i + \varepsilon_i$$

Dale and Krueger - Heterogeneous Effects

- DK estimate a regression of the form:

$$Y_i = \beta_D D_i + \beta_{DX}(D_i \times X_{i1}) + \boldsymbol{\gamma}' \boldsymbol{X}_i + \varepsilon_i$$

where D_i is average-SAT score/100, X_{i1} is log family income, and $\boldsymbol{X}_i$ includes indicators for your attended college + other controls

Dale and Krueger - Heterogeneous Effects

- DK estimate a regression of the form:

$$Y_i = \beta_D D_i + \beta_{DX}(D_i \times X_{i1}) + \boldsymbol{\gamma}' \boldsymbol{X}_i + \varepsilon_i$$

  where D_i is average-SAT score/100, X_{i1} is log family income, and $\boldsymbol{X}_i$ includes indicators for your attended college + other controls

  - The $D_i \times X_{i1}$ regressor is sometimes called an *interaction*

Dale and Krueger - Heterogeneous Effects

- DK estimate a regression of the form:

$$Y_i = \beta_D D_i + \beta_{DX}(D_i \times X_{i1}) + \boldsymbol{\gamma}' \boldsymbol{X}_i + \varepsilon_i$$

  where D_i is average-SAT score/100, X_{i1} is log family income, and $\boldsymbol{X}_i$ includes indicators for your attended college + other controls

  - The $D_i \times X_{i1}$ regressor is sometimes called an *interaction*

- The estimated effect of attending a school with 100 points higher SAT scores for a family with income of 100K is then

Dale and Krueger - Heterogeneous Effects

- DK estimate a regression of the form:

$$Y_i = \beta_D D_i + \beta_{DX}(D_i \times X_{i1}) + \boldsymbol{\gamma}' \boldsymbol{X}_i + \varepsilon_i$$

  where D_i is average-SAT score/100, X_{i1} is log family income, and $\boldsymbol{X}_i$ includes indicators for your attended college + other controls

  - The $D_i \times X_{i1}$ regressor is sometimes called an *interaction*

- The estimated effect of attending a school with 100 points higher SAT scores for a family with income of 100K is then

$$\hat{\beta}_D + \hat{\beta}_{DX} \times ln(100,000)$$

OF MALE AND FEMALE FULL-TIME WORKERS

Variable	Parameter estimates: Basic model: no selection controls	Parameter estimates: Matched-applicant model*
	1	2
School-average SAT score/100	0.701	0.537
	(0.185)	(0.224)
Predicted log(parental income)	0.915	0.819
	(0.212)	(0.247)
Predicted log of parental income ∗ school SAT score/100	−0.063	−0.056
	(0.019)	(0.023)
Own SAT score/100	0.018	−0.011
	(0.006)	(0.007)
High school top 10 percent	0.062	0.080
	(0.019)	(0.026)
High school rank missing	0.005	0.018
	(0.024)	(0.038)
Athlete	0.104	0.105
	(0.025)	(0.040)

OF MALE AND FEMALE FULL-TIME WORKERS

Variable	Parameter estimates: Basic model: no selection controls (1)	Parameter estimates: Matched-applicant model* (2)
School-average SAT score/100	0.701	0.537
	(0.185)	(0.224)
Predicted log(parental income)	0.915	0.819
	(0.212)	(0.247)
Predicted log of parental income * school SAT score/100	−0.063	−0.056
	(0.019)	(0.023)
Own SAT score/100	0.018	−0.011
	(0.006)	(0.007)
High school top 10 percent	0.062	0.080
	(0.019)	(0.026)
High school rank missing	0.005	0.018
	(0.024)	(0.038)
Athlete	0.104	0.105
	(0.025)	(0.040)

- In DK, students at the 10th percentile of family earnings have log parental income of 8.86 ($\approx \$7,000$)
- What is the estimated effect of going to a school w/avg SAT 100 points higher for students at the 10th percentile?

OF MALE AND FEMALE FULL-TIME WORKERS

	Parameter estimates	
	Basic model: no selection controls	Matched-applicant model*
Variable	1	2
School-average SAT score/100	0.701	0.537
	(0.185)	(0.224)
Predicted log(parental income)	0.915	0.819
	(0.212)	(0.247)
Predicted log of parental income * school SAT score/100	−0.063	−0.056
	(0.019)	(0.023)
Own SAT score/100	0.018	−0.011
	(0.006)	(0.007)
High school top 10 percent	0.062	0.080
	(0.019)	(0.026)
High school rank missing	0.005	0.018
	(0.024)	(0.038)
Athlete	0.104	0.105
	(0.025)	(0.040)

- In DK, students at the 10th percentile of family earnings have log parental income of 8.86 ($\approx \$7,000$)
- What is the estimated effect of going to a school w/avg SAT 100 points higher for students at the 10th percentile?

$$\hat{\beta}_D + \hat{\beta}_{DX} x =$$

OF MALE AND FEMALE FULL-TIME WORKERS

Variable	Parameter estimates: Basic model: no selection controls (1)	Parameter estimates: Matched-applicant model* (2)
School-average SAT score/100	0.701	0.537
	(0.185)	(0.224)
Predicted log(parental income)	0.915	0.819
	(0.212)	(0.247)
Predicted log of parental income * school SAT score/100	−0.063	−0.056
	(0.019)	(0.023)
Own SAT score/100	0.018	−0.011
	(0.006)	(0.007)
High school top 10 percent	0.062	0.080
	(0.019)	(0.026)
High school rank missing	0.005	0.018
	(0.024)	(0.038)
Athlete	0.104	0.105
	(0.025)	(0.040)

- In DK, students at the 10th percentile of family earnings have log parental income of 8.86 ($\approx \$7,000$)
- What is the estimated effect of going to a school w/avg SAT 100 points higher for students at the 10th percentile?

$$\hat{\beta}_D + \hat{\beta}_{DX} x = 0.537 - 0.056 \times 8.86 = 0.041$$

OF MALE AND FEMALE FULL-TIME WORKERS

Variable	Parameter estimates: Basic model: no selection controls (1)	Parameter estimates: Matched-applicant model* (2)
School-average SAT score/100	0.701	0.537
	(0.185)	(0.224)
Predicted log(parental income)	0.915	0.819
	(0.212)	(0.247)
Predicted log of parental income * school SAT score/100	−0.063	−0.056
	(0.019)	(0.023)
Own SAT score/100	0.018	−0.011
	(0.006)	(0.007)
High school top 10 percent	0.062	0.080
	(0.019)	(0.026)
High school rank missing	0.005	0.018
	(0.024)	(0.038)
Athlete	0.104	0.105
	(0.025)	(0.040)

- In DK, students at the 50th percentile of family earnings have log parental income of 10.39 ($\approx \$33,000$)
- What is the estimated effect of going to a school w/avg SAT 100 points higher at the median?

$$\hat{\beta}_D + \hat{\beta}_{DX} x =$$

OF MALE AND FEMALE FULL-TIME WORKERS

	Parameter estimates	
	Basic model: no selection controls	Matched-applicant model*
Variable	1	2
School-average SAT score/100	0.701	0.537
	(0.185)	(0.224)
Predicted log(parental income)	0.915	0.819
	(0.212)	(0.247)
Predicted log of parental income * school SAT score/100	−0.063	−0.056
	(0.019)	(0.023)
Own SAT score/100	0.018	−0.011
	(0.006)	(0.007)
High school top 10 percent	0.062	0.080
	(0.019)	(0.026)
High school rank missing	0.005	0.018
	(0.024)	(0.038)
Athlete	0.104	0.105
	(0.025)	(0.040)

- In DK, students at the 50th percentile of family earnings have log parental income of 10.39 ($\approx \$33,000$)

- What is the estimated effect of going to a school w/avg SAT 100 points higher at the median?

$$\hat{\beta}_D + \hat{\beta}_{DX} x = 0.537 - 0.056 \times 10.39 = -0.045$$

Explaining Heterogeneity

- The results above showed that the returns to more selective college are positive for poorer students but negative for richer students

Explaining Heterogeneity

- The results above showed that the returns to more selective college are positive for poorer students but negative for richer students
- What might explain why elite college seems to matter more for students from poor backgrounds?

Explaining Heterogeneity

- The results above showed that the returns to more selective college are positive for poorer students but negative for richer students
- What might explain why elite college seems to matter more for students from poor backgrounds?
- Not entirely clear... networking more important for students who don't have as many family connections?

Linear Combinations of Coefficients

- In the DK example above, the estimated $CATE(x)$ for students with family income x was $\hat{\beta}_D + \hat{\beta}_{DX} x$.

Linear Combinations of Coefficients

- In the DK example above, the estimated $CATE(x)$ for students with family income x was $\hat{\beta}_D + \hat{\beta}_{DX}x$.

- Suppose we want to construct a CI or test hypothese about $CATE(x)$. How can we do that?

Linear Combinations of Coefficients

- We showed several lectures ago that

$$\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \rightarrow_d \mathrm{N}(0, \boldsymbol{\Sigma})$$

Linear Combinations of Coefficients

- We showed several lectures ago that

$$\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \rightarrow_d \mathrm{N}(0, \boldsymbol{\Sigma})$$

- Say we're interested in $\beta_1 + \beta_2 x$.

Linear Combinations of Coefficients

- We showed several lectures ago that

$$\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \rightarrow_d \mathrm{N}(0, \boldsymbol{\Sigma})$$

- Say we're interested in $\beta_1 + \beta_2 x$. Since $g(\boldsymbol{\beta}) = \beta_1 + \beta_2 x$ is continuous, by the continuous mapping theorem we have

Linear Combinations of Coefficients

- We showed several lectures ago that

$$\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \to_d \mathrm{N}(0, \boldsymbol{\Sigma})$$

- Say we're interested in $\beta_1 + \beta_2 x$. Since $g(\boldsymbol{\beta}) = \beta_1 + \beta_2 x$ is continuous, by the continuous mapping theorem we have

$$\sqrt{N}(\hat{\beta}_1 + \hat{\beta}_2 x - (\beta_1 + \beta_2 x)) \to_d \mathrm{N}(0, \sigma_x^2),$$

  where $\sigma_x^2 = \Sigma_{11} + x^2 \Sigma_{22} + 2x\Sigma_{12}$.

Linear Combinations of Coefficients

- We showed several lectures ago that

$$\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \to_d \mathrm{N}(0, \boldsymbol{\Sigma})$$

- Say we're interested in $\beta_1 + \beta_2 x$. Since $g(\boldsymbol{\beta}) = \beta_1 + \beta_2 x$ is continuous, by the continuous mapping theorem we have

$$\sqrt{N}(\hat{\beta}_1 + \hat{\beta}_2 x - (\beta_1 + \beta_2 x)) \to_d \mathrm{N}(0, \sigma_x^2),$$

  where $\sigma_x^2 = \Sigma_{11} + x^2\Sigma_{22} + 2x\Sigma_{12}$.

- In previous classes we derived formulas for $\hat{\boldsymbol{\Sigma}}$, a consistent estimator of $\boldsymbol{\Sigma}$ (plugging in sample analogs)

Linear Combinations of Coefficients

- We showed several lectures ago that

$$\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \to_d \mathrm{N}(0, \boldsymbol{\Sigma})$$

- Say we're interested in $\beta_1 + \beta_2 x$. Since $g(\boldsymbol{\beta}) = \beta_1 + \beta_2 x$ is continuous, by the continuous mapping theorem we have

$$\sqrt{N}(\hat{\beta}_1 + \hat{\beta}_2 x - (\beta_1 + \beta_2 x)) \to_d \mathrm{N}(0, \sigma_x^2),$$

  where $\sigma_x^2 = \Sigma_{11} + x^2\Sigma_{22} + 2x\Sigma_{12}$.

- In previous classes we derived formulas for $\hat{\boldsymbol{\Sigma}}$, a consistent estimator of $\boldsymbol{\Sigma}$ (plugging in sample analogs)

- So we can form a CI for $\beta_1 + \beta_2 x$ using $\hat{\beta}_1 + \hat{\beta}_2 x \pm 1.96\hat{\sigma}_x/\sqrt{N}$, where $\hat{\sigma}_x^2 = \hat{\Sigma}_{11} + x^2\hat{\Sigma}_{22} + 2x\hat{\Sigma}_{12}$.

Example

- Recall that in DK we have $\hat{\beta}_D = 0.537$ and $\hat{\beta}_{DX} = -0.056$.

- Suppose that the part $\hat{\boldsymbol{\Sigma}}/N$ corresponding w/the coefficients of interest is

	$\hat{\beta}_D$	$\hat{\beta}_{DX}$
$\hat{\beta}_D$	0.0050	-0.0025
$\hat{\beta}_{DX}$	-0.0025	0.0005

  [Note: I had to make up the covariance term – not in the paper]

Example

- Recall that in DK we have $\hat{\beta}_D = 0.537$ and $\hat{\beta}_{DX} = -0.056$.

- Suppose that the part $\hat{\mathbf{\Sigma}}/N$ corresponding w/the coefficients of interest is

	$\hat{\beta}_D$	$\hat{\beta}_{DX}$
$\hat{\beta}_D$	0.0050	-0.0025
$\hat{\beta}_{DX}$	-0.0025	0.0005

  [Note: I had to make up the covariance term – not in the paper]

- Then $\hat{\Sigma}_{11}/N =$

Example

- Recall that in DK we have $\hat{\beta}_D = 0.537$ and $\hat{\beta}_{DX} = -0.056$.
- Suppose that the part $\hat{\boldsymbol{\Sigma}}/N$ corresponding w/the coefficients of interest is

	$\hat{\beta}_D$	$\hat{\beta}_{DX}$
$\hat{\beta}_D$	0.0050	-0.0025
$\hat{\beta}_{DX}$	-0.0025	0.0005

  [Note: I had to make up the covariance term – not in the paper]
- Then $\hat{\Sigma}_{11}/N = 0.0050$,

Example

- Recall that in DK we have $\hat{\beta}_D = 0.537$ and $\hat{\beta}_{DX} = -0.056$.

- Suppose that the part $\hat{\mathbf{\Sigma}}/N$ corresponding w/the coefficients of interest is

	$\hat{\beta}_D$	$\hat{\beta}_{DX}$
$\hat{\beta}_D$	0.0050	-0.0025
$\hat{\beta}_{DX}$	-0.0025	0.0005

  [Note: I had to make up the covariance term – not in the paper]

- Then $\hat{\Sigma}_{11}/N = 0.0050$, $\hat{\Sigma}_{22}/N =$

Example

- Recall that in DK we have $\hat{\beta}_D = 0.537$ and $\hat{\beta}_{DX} = -0.056$.
- Suppose that the part $\hat{\mathbf{\Sigma}}/N$ corresponding w/the coefficients of interest is

	$\hat{\beta}_D$	$\hat{\beta}_{DX}$
$\hat{\beta}_D$	0.0050	-0.0025
$\hat{\beta}_{DX}$	-0.0025	0.0005

  [Note: I had to make up the covariance term – not in the paper]
- Then $\hat{\Sigma}_{11}/N = 0.0050$, $\hat{\Sigma}_{22}/N = 0.0005$,

Example

- Recall that in DK we have $\hat{\beta}_D = 0.537$ and $\hat{\beta}_{DX} = -0.056$.
- Suppose that the part $\hat{\mathbf{\Sigma}}/N$ corresponding w/the coefficients of interest is

	$\hat{\beta}_D$	$\hat{\beta}_{DX}$
$\hat{\beta}_D$	0.0050	-0.0025
$\hat{\beta}_{DX}$	-0.0025	0.0005

  [Note: I had to make up the covariance term – not in the paper]
- Then $\hat{\Sigma}_{11}/N = 0.0050$, $\hat{\Sigma}_{22}/N = 0.0005$, ,and $\hat{\Sigma}_{12}/N =$

Example

- Recall that in DK we have $\hat{\beta}_D = 0.537$ and $\hat{\beta}_{DX} = -0.056$.
- Suppose that the part $\hat{\boldsymbol{\Sigma}}/N$ corresponding w/the coefficients of interest is

	$\hat{\beta}_D$	$\hat{\beta}_{DX}$
$\hat{\beta}_D$	0.0050	-0.0025
$\hat{\beta}_{DX}$	-0.0025	0.0005

  [Note: I had to make up the covariance term – not in the paper]
- Then $\hat{\Sigma}_{11}/N = 0.0050$, $\hat{\Sigma}_{22}/N = 0.0005$, ,and $\hat{\Sigma}_{12}/N = -0.0025$.

Example

- Recall that in DK we have $\hat{\beta}_D = 0.537$ and $\hat{\beta}_{DX} = -0.056$.
- Suppose that the part $\hat{\mathbf{\Sigma}}/N$ corresponding w/the coefficients of interest is

	$\hat{\beta}_D$	$\hat{\beta}_{DX}$
$\hat{\beta}_D$	0.0050	-0.0025
$\hat{\beta}_{DX}$	-0.0025	0.0005

  [Note: I had to make up the covariance term – not in the paper]
- Then $\hat{\Sigma}_{11}/N = 0.0050$, $\hat{\Sigma}_{22}/N = 0.0005$, ,and $\hat{\Sigma}_{12}/N = -0.0025$.
- So a CI for $CATE(10.39)$ is

Example

- Recall that in DK we have $\hat{\beta}_D = 0.537$ and $\hat{\beta}_{DX} = -0.056$.

- Suppose that the part $\hat{\mathbf{\Sigma}}/N$ corresponding w/the coefficients of interest is

	$\hat{\beta}_D$	$\hat{\beta}_{DX}$
$\hat{\beta}_D$	0.0050	-0.0025
$\hat{\beta}_{DX}$	-0.0025	0.0005

  [Note: I had to make up the covariance term – not in the paper]

- Then $\hat{\Sigma}_{11}/N = 0.0050$, $\hat{\Sigma}_{22}/N = 0.0005$, ,and $\hat{\Sigma}_{12}/N = -0.0025$.

- So a CI for $CATE(10.39)$ is

$$\beta_D + \beta_{DX}x \pm 1.96\sqrt{\hat{\Sigma}_{11}/N + x^2\hat{\Sigma}_{22}/N + 2x\hat{\Sigma}_{12}/N} =$$

Example

- Recall that in DK we have $\hat{\beta}_D = 0.537$ and $\hat{\beta}_{DX} = -0.056$.
- Suppose that the part $\hat{\boldsymbol{\Sigma}}/N$ corresponding w/the coefficients of interest is

	$\hat{\beta}_D$	$\hat{\beta}_{DX}$
$\hat{\beta}_D$	0.0050	-0.0025
$\hat{\beta}_{DX}$	-0.0025	0.0005

  [Note: I had to make up the covariance term – not in the paper]
- Then $\hat{\Sigma}_{11}/N = 0.0050$, $\hat{\Sigma}_{22}/N = 0.0005$, ,and $\hat{\Sigma}_{12}/N = -0.0025$.
- So a CI for $CATE(10.39)$ is

$$\beta_D + \beta_{DX} x \pm 1.96\sqrt{\hat{\Sigma}_{11}/N + x^2\hat{\Sigma}_{22}/N + 2x\hat{\Sigma}_{12}/N} =$$

$$(0.537 - 0.056 \times 10.39) \pm$$

$$1.96 \times \sqrt{0.0050 + 10.39^2 \times 0.0005 + 2 \times 10.39 \times (-0.0025)} =$$

Example

- Recall that in DK we have $\hat{\beta}_D = 0.537$ and $\hat{\beta}_{DX} = -0.056$.
- Suppose that the part $\hat{\mathbf{\Sigma}}/N$ corresponding w/the coefficients of interest is

	$\hat{\beta}_D$	$\hat{\beta}_{DX}$
$\hat{\beta}_D$	0.0050	-0.0025
$\hat{\beta}_{DX}$	-0.0025	0.0005

  [Note: I had to make up the covariance term – not in the paper]
- Then $\hat{\Sigma}_{11}/N = 0.0050$, $\hat{\Sigma}_{22}/N = 0.0005$, ,and $\hat{\Sigma}_{12}/N = -0.0025$.
- So a CI for $CATE(10.39)$ is

$$\beta_D + \beta_{DX} x \pm 1.96\sqrt{\hat{\Sigma}_{11}/N + x^2\hat{\Sigma}_{22}/N + 2x\hat{\Sigma}_{12}/N} =$$

$$(0.537 - 0.056 \times 10.39) \pm$$

$$1.96 \times \sqrt{0.0050 + 10.39^2 \times 0.0005 + 2 \times 10.39 \times (-0.0025)} =$$

$$[-0.21, 0.12]$$

lincom

- The `lincom` command in Stata generalizes the argument above to test for any linear combination of coefficients, i.e. parameters of the form $a_1\beta_1 + ... + a_k\beta_k$

lincom

- The `lincom` command in Stata generalizes the argument above to test for any linear combination of coefficients, i.e. parameters of the form $a_1\beta_1 + ... + a_k\beta_k$

- Similar (but slightly more complicated) asymptotic arguments can be used to test hypotheses on non-linear combinations of coefficients, e.g. $\beta_1\beta_2 + \beta_3^2$. This can be done using the `nlcom` command in Stata.

Outline

Odds and Ends 1: Multiple Hypothesis Testing

- We've seen how to test for differences in causal effects across groups

Odds and Ends 1: Multiple Hypothesis Testing

- We've seen how to test for differences in causal effects across groups
- We considered one dimension of heterogeneity: poor vs. rich students. But we might be interested in heterogeneity for other groups too
  - Are the effects bigger for men than for woman?

Odds and Ends 1: Multiple Hypothesis Testing

- We've seen how to test for differences in causal effects across groups
- We considered one dimension of heterogeneity: poor vs. rich students. But we might be interested in heterogeneity for other groups too
  - Are the effects bigger for men than for woman?
  - Are the effects bigger for athletes than non-athletes?

Odds and Ends 1: Multiple Hypothesis Testing

- We've seen how to test for differences in causal effects across groups
- We considered one dimension of heterogeneity: poor vs. rich students. But we might be interested in heterogeneity for other groups too
  - Are the effects bigger for men than for woman?
  - Are the effects bigger for athletes than non-athletes?
  - Are the effects bigger for male athletes from rich families than female non-athletes from poor families?

Odds and Ends 1: Multiple Hypothesis Testing

- We've seen how to test for differences in causal effects across groups
- We considered one dimension of heterogeneity: poor vs. rich students. But we might be interested in heterogeneity for other groups too
  - Are the effects bigger for men than for woman?
  - Are the effects bigger for athletes than non-athletes?
  - Are the effects bigger for male athletes from rich families than female non-athletes from poor families?
- There are a large number of hypotheses that we may want to test!

Odds and Ends 1: Multiple Hypothesis Testing

- We've seen how to test for differences in causal effects across groups
- We considered one dimension of heterogeneity: poor vs. rich students. But we might be interested in heterogeneity for other groups too
  - Are the effects bigger for men than for woman?
  - Are the effects bigger for athletes than non-athletes?
  - Are the effects bigger for male athletes from rich families than female non-athletes from poor families?
- There are a large number of hypotheses that we may want to test!
- This can lead to what's called a **multiple hypothesis testing** problem

Multiple Hypothesis Testing (Cont.)

- Remember: we constructed p-values so that (asymptotically), we have $p < 0.05$ only 5% of the time under the null hypothesis of no treatment effect

Multiple Hypothesis Testing (Cont.)

- Remember: we constructed p-values so that (asymptotically), we have $p < 0.05$ only 5% of the time under the null hypothesis of no treatment effect
- Thus, if we test for a significant effect among the entire population, if there is truly no effect we should reject the null only 5% of the time.

Multiple Hypothesis Testing (Cont.)

- Remember: we constructed p-values so that (asymptotically), we have $p < 0.05$ only 5% of the time under the null hypothesis of no treatment effect
- Thus, if we test for a significant effect among the entire population, if there is truly no effect we should reject the null only 5% of the time.
- But suppose we first test for a significant effect among men. And we then also test for a significant effect among women.

Multiple Hypothesis Testing (Cont.)

- Remember: we constructed p-values so that (asymptotically), we have $p < 0.05$ only 5% of the time under the null hypothesis of no treatment effect
- Thus, if we test for a significant effect among the entire population, if there is truly no effect we should reject the null only 5% of the time.
- But suppose we first test for a significant effect among men. And we then also test for a significant effect among women.
- If there is no significant effect among either group, what is the probability that we find at least one significant effect?

- Suppose the samples for men and women are drawn independently.
- Let N_{sig} be the number of significant results.

$$P(N_{sig} \geq 1) =$$

- Suppose the samples for men and women are drawn independently.

- Let N_{sig} be the number of significant results.

$$P(N_{sig} \geq 1) = 1 - P(N_{sig} = 0)$$

- Suppose the samples for men and women are drawn independently.

- Let N_{sig} be the number of significant results.

$$P(N_{sig} \geq 1) = 1 - P(N_{sig} = 0)$$
$$= 1 - P(\text{female insig and male insig})$$

- Suppose the samples for men and women are drawn independently.
- Let N_{sig} be the number of significant results.

$$\begin{aligned} P(N_{sig} \geq 1) &= 1 - P(N_{sig} = 0) \\ &= 1 - P(\text{female insig and male insig}) \\ &= 1 - P(\text{female insig})P(\text{male insig}) \end{aligned}$$

- Suppose the samples for men and women are drawn independently.
- Let N_{sig} be the number of significant results.

$$
\begin{aligned}
&P(N_{sig} \geq 1) = 1 - P(N_{sig} = 0) \\
&= 1 - P(\text{female insig and male insig}) \\
&= 1 - P(\text{female insig})P(\text{male insig}) \\
&= 1 - .95^2
\end{aligned}
$$

- Suppose the samples for men and women are drawn independently.
- Let N_{sig} be the number of significant results.

$$
\begin{aligned}
&P(N_{sig} \geq 1) = 1 - P(N_{sig} = 0) \\
&= 1 - P(\text{female insig and male insig}) \\
&= 1 - P(\text{female insig})P(\text{male insig}) \\
&= 1 - .95^2 = 0.0975
\end{aligned}
$$

- So we'll find at least one significant effect almost 10% of the time if there is no effect for either group

- Suppose the samples for men and women are drawn independently.
- Let N_{sig} be the number of significant results.

$$
\begin{aligned}
&P(N_{sig} \geq 1) = 1 - P(N_{sig} = 0) \\
&= 1 - P(\text{female insig and male insig}) \\
&= 1 - P(\text{female insig})P(\text{male insig}) \\
&= 1 - .95^2 = 0.0975
\end{aligned}
$$

- So we'll find at least one significant effect almost 10% of the time if there is no effect for either group
- By the same argument, if we test the null for k independent groups, and there is no true effect, we will reject the null with probability

- Suppose the samples for men and women are drawn independently.
- Let N_{sig} be the number of significant results.

$$
\begin{aligned}
&P(N_{sig} \geq 1) = 1 - P(N_{sig} = 0) \\
&= 1 - P(\text{female insig and male insig}) \\
&= 1 - P(\text{female insig})P(\text{male insig}) \\
&= 1 - .95^2 = 0.0975
\end{aligned}
$$

- So we'll find at least one significant effect almost 10% of the time if there is no effect for either group
- By the same argument, if we test the null for k independent groups, and there is no true effect, we will reject the null with probability $1 - 0.95^k$

Probability of finding at least one subgroup with a signficant effect with k independent groups (and zero treatment effect):

k	$1-0.95^k$
1	0.05
2	0.0975

Probability of finding at least one subgroup with a signficant effect with k independent groups (and zero treatment effect):

k	$1-0.95^k$
1	0.05
2	0.0975
3	

Probability of finding at least one subgroup with a signficant effect with k independent groups (and zero treatment effect):

k	$1-0.95^k$
1	0.05
2	0.0975
3	0.1426
5	

Probability of finding at least one subgroup with a signficant effect with k independent groups (and zero treatment effect):

k	$1-0.95^k$
1	0.05
2	0.0975
3	0.1426
5	0.2262
10	

Probability of finding at least one subgroup with a signficant effect with k independent groups (and zero treatment effect):

k	$1-0.95^k$
1	0.05
2	0.0975
3	0.1426
5	0.2262
10	0.4013
100	

Probability of finding at least one subgroup with a signficant effect with k independent groups (and zero treatment effect):

k	$1-0.95^k$
1	0.05
2	0.0975
3	0.1426
5	0.2262
10	0.4013
100	0.9941

Simulating Multiple Hypothesis Testing

- Have survey data on average hourly wages from the Current Population Survey
- I generate a fake treatment D_i which is 1 with probability $1/2$

Simulating Multiple Hypothesis Testing

- Have survey data on average hourly wages from the Current Population Survey
- I generate a fake treatment D_i which is 1 with probability $1/2$
- What's the true causal effect of this treatment?

Simulating Multiple Hypothesis Testing

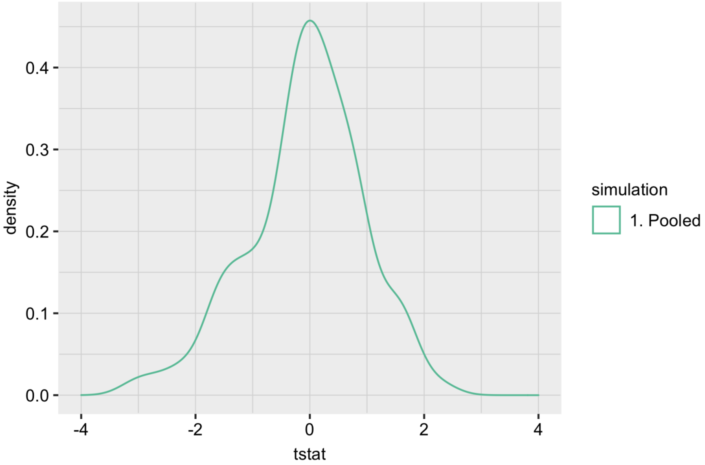
- Have survey data on average hourly wages from the Current Population Survey
- I generate a fake treatment D_i which is 1 with probability $1/2$
- What's the true causal effect of this treatment? 0, of course

Simulating Multiple Hypothesis Testing

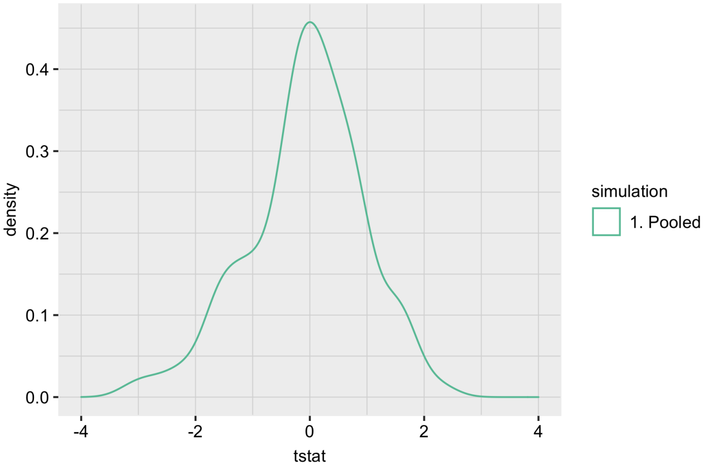
- Have survey data on average hourly wages from the Current Population Survey
- I generate a fake treatment D_i which is 1 with probability $1/2$
- What's the true causal effect of this treatment? 0, of course
- I simulate this fake treament 100 times, and estimate
  1. The treatment effect pooling all states
  2. The individual treatment effect for the first 10 states in the data
  3. The individual treatment effect for all 50 states in the data

Simulating Multiple Hypothesis Testing

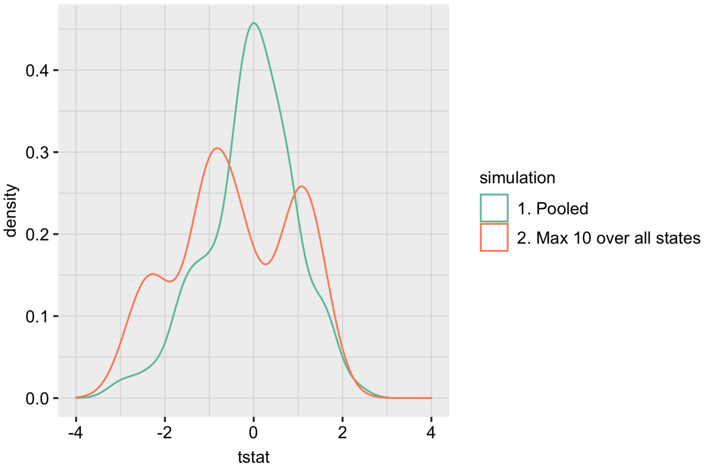
- Have survey data on average hourly wages from the Current Population Survey
- I generate a fake treatment D_i which is 1 with probability $1/2$
- What's the true causal effect of this treatment? 0, of course
- I simulate this fake treament 100 times, and estimate
  1. The treatment effect pooling all states
  2. The individual treatment effect for the first 10 states in the data
  3. The individual treatment effect for all 50 states in the data
- I then calculate the fraction of simulations in which we get at least one significant estimate



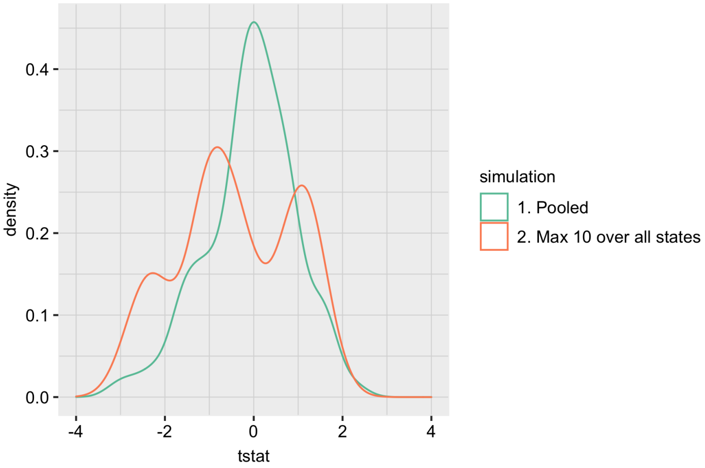
0.4
0.3
0.2
0.1
0.0
density
-4
-2
0
2
4
tstat
simulation
1. Pooled



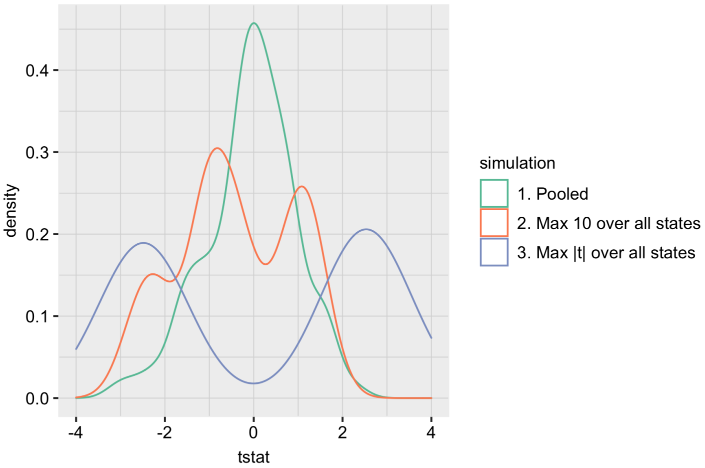
- When testing the pooled effect across all states, we find a significant effect 5% of the time

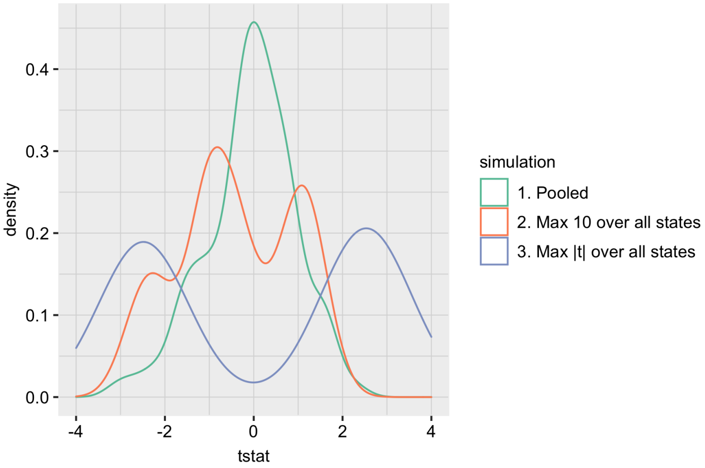


0.4
0.3
0.2
0.1
0.0
density
-4
-2
0
2
4
tstat
simulation
1. Pooled
2. Max 10 over all states



- When testing the pooled effect across all states, we find a significant effect 19% of the time


density
0.4
0.3
0.2
0.1
0.0
-4
-2
0
2
4
tstat
simulation
1. Pooled
2. Max 10 over all states
3. Max |t| over all states



- When testing the pooled effect across all states, we find a significant effect 90% of the time

Correcting for Multiple Hypothesis Testing

- The simplest correction for multiple hypothesis testing is the **Bonferroni** correction

Correcting for Multiple Hypothesis Testing

- The simplest correction for multiple hypothesis testing is the **Bonferroni** correction

- Instead of testing each of the k individual hypotheses at the 5% level, we test each one at the $5/k\%$ level.

Correcting for Multiple Hypothesis Testing

- The simplest correction for multiple hypothesis testing is the **Bonferroni** correction
- Instead of testing each of the k individual hypotheses at the 5% level, we test each one at the $5/k\%$ level.
- Then

$$P(N_{sig} > 0) \leq \sum_k P(\text{k is signficant}) = k(0.05/k) = 0.05$$

Correcting for Multiple Hypothesis Testing

- The simplest correction for multiple hypothesis testing is the **Bonferroni** correction

- Instead of testing each of the k individual hypotheses at the 5% level, we test each one at the $5/k$% level.

- Then

$$P(N_{sig} > 0) \leq \sum_k P(\text{k is signficant}) = k(0.05/k) = 0.05$$

- This works even if the hypotheses are not independent, e.g. when you have overlapping groups (e.g., one group is all states, the second group is Rhode Island)

Correcting for Multiple Hypothesis Testing

- The simplest correction for multiple hypothesis testing is the **Bonferroni** correction

- Instead of testing each of the k individual hypotheses at the 5% level, we test each one at the $5/k$% level.

- Then

$$P(N_{sig} > 0) \leq \sum_k P(\text{k is signficant}) = k(0.05/k) = 0.05$$

- This works even if the hypotheses are not independent, e.g. when you have overlapping groups (e.g., one group is all states, the second group is Rhode Island)

- Downside: if you have lots of hypotheses, power against any one hypothesis can be low (e.g., 99.9% confidence intervals are very wide).

Correcting for Multiple Hypothesis Testing

- The simplest correction for multiple hypothesis testing is the **Bonferroni** correction

- Instead of testing each of the k individual hypotheses at the 5% level, we test each one at the $5/k\%$ level.

- Then

$$P(N_{sig} > 0) \leq \sum_k P(\text{k is signficant}) = k(0.05/k) = 0.05$$

- This works even if the hypotheses are not independent, e.g. when you have overlapping groups (e.g., one group is all states, the second group is Rhode Island)

- Downside: if you have lots of hypotheses, power against any one hypothesis can be low (e.g., 99.9% confidence intervals are very wide). And the test is generally conservative in the sense that we find any significant effect $< 5\%$ of the time

Probability of rejecting at least one hypothesis without and with Bonferroni correction

Simulation	Uncorrected	Corrected
Pooled	0.05	0.05
10 States	0.19	0.00
50 States	0.90	0.04

Odds and Ends 2: Joint Hypotheses

- Sometimes we're happy to know whether there is an effect for *any* subgroup (e.g. any state)

Odds and Ends 2: Joint Hypotheses

- Sometimes we're happy to know whether there is an effect for *any* subgroup (e.g. any state)
- Can test the **joint null** that there is no treatment effect for every subgroup

Odds and Ends 2: Joint Hypotheses

- Sometimes we're happy to know whether there is an effect for *any* subgroup (e.g. any state)
- Can test the **joint null** that there is no treatment effect for every subgroup
- If we reject, we conclude that there is strong evidence that the treatment effect is non-zero for at least one subgroup

Odds and Ends 2: Joint Hypotheses

- Sometimes we're happy to know whether there is an effect for *any* subgroup (e.g. any state)
- Can test the **joint null** that there is no treatment effect for every subgroup
- If we reject, we conclude that there is strong evidence that the treatment effect is non-zero for at least one subgroup
- How do we do this?

- We showed that $\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \rightarrow_d \mathrm{N}(0, \boldsymbol{\Sigma})$

- From the continuous mapping theorem, we get that $X = \sqrt{N}\hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \rightarrow_d$

- We showed that $\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \to_d \mathrm{N}(0, \boldsymbol{\Sigma})$

- From the continuous mapping theorem, we get that $X = \sqrt{N}\hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \to_d \mathrm{N}(0, \boldsymbol{I})$.

- Applying the continuous mapping theorem again, we get that

$$F = \sum_k X_k^2 \to \sum_k Z_k^2, \text{ where } Z \sim \mathrm{N}(0, \boldsymbol{I})$$

- We showed that $\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \to_d \mathrm{N}(0, \boldsymbol{\Sigma})$

- From the continuous mapping theorem, we get that $X = \sqrt{N}\hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \to_d \mathrm{N}(0, \boldsymbol{I})$.

- Applying the continuous mapping theorem again, we get that

$$F = \sum_k X_k^2 \to \sum_k Z_k^2, \text{ where } Z \sim \mathrm{N}(0, \boldsymbol{I})$$

- The sum of k independent normals squared is called a *chi-squared* distribution with k degrees of freedom, denoted $\chi^2(k)$.

- We showed that $\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \rightarrow_d \mathrm{N}(0, \boldsymbol{\Sigma})$

- From the continuous mapping theorem, we get that $X = \sqrt{N}\hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \rightarrow_d \mathrm{N}(0, \boldsymbol{I})$.

- Applying the continuous mapping theorem again, we get that

$$F = \sum_k X_k^2 \rightarrow \sum_k Z_k^2, \text{ where } Z \sim \mathrm{N}(0, \boldsymbol{I})$$

- The sum of k independent normals squared is called a *chi-squared* distribution with k degrees of freedom, denoted $\chi^2(k)$.

- We can compare F to the 95th percentile of the $\chi^2(k)$ distribution, and reject if it is larger.

- We showed that $\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \to_d \mathrm{N}(0, \boldsymbol{\Sigma})$

- From the continuous mapping theorem, we get that $X = \sqrt{N}\hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \to_d \mathrm{N}(0, \boldsymbol{I})$.

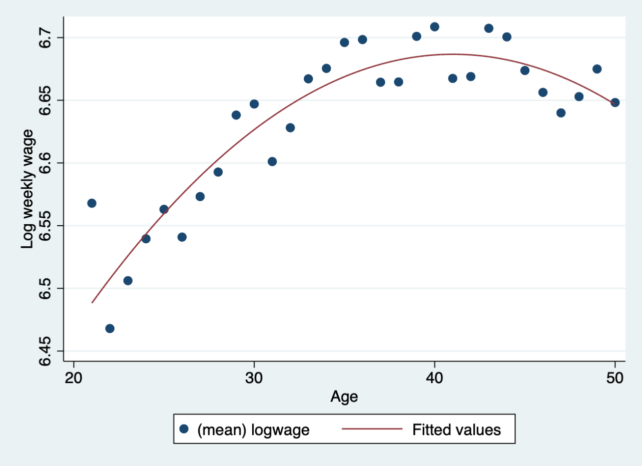
- Applying the continuous mapping theorem again, we get that

$$F = \sum_k X_k^2 \to \sum_k Z_k^2, \text{ where } Z \sim \mathrm{N}(0, \boldsymbol{I})$$

- The sum of k independent normals squared is called a *chi-squared* distribution with k degrees of freedom, denoted $\chi^2(k)$.

- We can compare F to the 95th percentile of the $\chi^2(k)$ distribution, and reject if it is larger.

- This is often called an *F-test*.
  - Sometimes the critical values use what's called an F-distribution, which is a slight modification to the $\chi^2(k)$ that corrects for small sample sizes


Log weekly wage
6.45
6.5
6.55
6.6
6.65
6.7
20
30
40
50
Age
(mean) logwage
Fitted values

```
. reg logwage age agesq, r

Linear regression                               Number of obs     =         30
                                                F(2, 27)          =      30.81
                                                Prob > F          =     0.0000
                                                R-squared         =     0.8412
                                                Root MSE          =     .02622

------------------------------------------------------------------------------
             |               Robust
     logwage | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         age |   .0403444   .0076871     5.25   0.000     .0245718     .056117
       agesq |   -.000492   .0001011    -4.87   0.000    -.0006994   -.0002846
       _cons |   5.859108   .1409008    41.58   0.000     5.570003    6.148212
------------------------------------------------------------------------------

.
. test age agesq

 ( 1)  age = 0
 ( 2)  agesq = 0

       F(  2,    27) =   30.81
            Prob > F =    0.0000
```