Lecture 5: Ordinary Least Squares

Mitch Downey[1]

February 7, 2024

[1]IIES, Stockholm University. Slides heavily based on those developed by Markus Jäntti based on the textbook by Bruce Hansen.

What is OLS?

- Many interpretations
- A linear projection of the n-dimensional vector y onto the k-dimensional space spanned by the k n-dimensional vectors that make up the columns of X
- What is a linear projection?
  <Demonstration with $n = 3$>

Select alternative interpretations

- Linear projection of y onto the space spanned by X
- OLS as a population projection coefficient minimizes the expected squared error
- OLS as a statistical estimator minimizes the sum of squared residuals (residual is an estimated error)

[2] Hansen, Bruce E. "A modern Gauss–Markov theorem." *Econometrica* 90.3 (2022): 1283-1294.

Select alternative interpretations

- Linear projection of y onto the space spanned by X
- OLS as a population projection coefficient minimizes the expected squared error
- OLS as a statistical estimator minimizes the sum of squared residuals (residual is an estimated error)
- Both: Among linear models
    - Why focus on linear models?
        - Historically: Tractability
        - Today: Meaningful, assumptions like $E(X'\varepsilon) = 0$ mean something
    - Gauss-Markov Theorem (Section 4.8): Among linear estimators in the presence of homoskedasticity, OLS is minimum variance unbiased estimator
    - Hansen (2022) reformulation:[2] If the conditional mean is linear and homoskedastic, OLS is minimum variance unbiased estimator

[2] Hansen, Bruce E. "A modern Gauss–Markov theorem." *Econometrica* 90.3 (2022): 1283-1294.

Select alternative interpretations

- Linear projection of y onto the space spanned by X
- OLS as a population projection coefficient minimizes the expected squared error
- OLS as a statistical estimator minimizes the sum of squared residuals (residual is an estimated error)
- Both: *squared* errors
    - Asymmetric loss functions:
        - Minimize: $\sum_{i=1}^{N} \alpha_{pos} \mathbf{1}_{\{\hat{\varepsilon}_i > 0\}} |\hat{\varepsilon}_i| + \alpha_{neg} \mathbf{1}_{\{\hat{\varepsilon}_i < 0\}} |\hat{\varepsilon}_i|$
        - If $\alpha_{pos} = \alpha_{neg}$ then this minimizes sum of absolute errors
    - Squares:
        - $\varepsilon_i = 2$ will be penalized more than $\varepsilon_j = \varepsilon_k = 1$
        - Increases the influence of outliers
        - Better to be slightly wrong often than very wrong once

OLS: Statistical least squares estimator (3.4)

- You observe Y $(n \times 1)$ and X $(n \times k)$
- You want to estimate a linear model so that $Y = X\hat{\beta} + \hat{\varepsilon}$
- You want to minimize $\hat{\varepsilon}'\hat{\varepsilon}$:

$$
\begin{aligned}
\frac{\partial}{\partial\hat{\beta}}\left(Y - X\hat{\beta}\right)'\left(Y - X\hat{\beta}\right) &= \frac{\partial}{\partial\hat{\beta}}\left(Y'Y - \hat{\beta}'X'Y - Y'X\hat{\beta} + \hat{\beta}'X'X\hat{\beta}\right) \\
&= \frac{\partial}{\partial\hat{\beta}}\left(Y'Y - 2\hat{\beta}'X'Y - \hat{\beta}'X'X\hat{\beta}\right) \quad \text{(bec. } Y'X\hat{\beta} \text{ is a scalar)} \\
&= 2X'Y - 2X'X\hat{\beta} \quad = 0
\end{aligned}
$$

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

OLS: Best linear approximation of conditional mean (2.25)

- Which linear function minimizes the average squared distance from the conditional expectations function?
- Choose β to minimize:

$$\int_{\mathbb{R}^k} \left(m(x) - x\beta\right)^2 f_X(x)dx$$

- Similar algebra yields:

$$\beta = \left(E[XX']\right)^{-1} E[XY]$$

  note: X is $k \times 1$

- This β is sometimes called the population projection coefficient
- Use "plug-in" estimator of sample means:

$$\hat{\beta} = \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} X_i Y_i$$

Summary so far

- So far: Why might you do this?
- Next: What do you get out of it?
  - I'll show these for the population linear projection coefficient $\beta = \left(E[XX']\right)^{-1} E[XY]$
  - They hold for the least squares estimator in finite samples too
- Then: What is it and how should you interpret it?

What do you get out of OLS? Properties

$$
\begin{aligned}
E(Xe) &\equiv E\big(X(Y - X'\beta)\big) \\
&= E(XY) - E(XX'\beta) \\
&= E(XY) - E(XX')\beta \\
&= E(XY) - E(XX')\big(E(XX')\big)^{-1}E(XY) \\
&= E(XY) - E(XY) = 0
\end{aligned}
$$

- X is uncorrelated with the error
  - Not the same as saying X is uncorrelated with residual, though that's also true
- Linear projection always does this
- What if this weren't true?
  - If $E(Xe) > 0$ then Y is systematically higher than you'd expect for high X's
  - Then why didn't you estimate a larger β? If you're systematically underestimating Y for high X?
  - $E(Xe)$ means that our linear projection coefficient β has extracted all possible information out of X that is helpful for predicting the mean of Y
    - A non-linear function of X could plausibly do better

What do you get out of OLS? Properties

$$E(X_j e) = 0 \ \ \forall j$$

- The above derivation holds for each component of X
- Unless stated otherwise, we assume this includes a constant
- This means: $E(1e) = E(e) = 0$
- Thus, without a constant, errors need not be mean zero

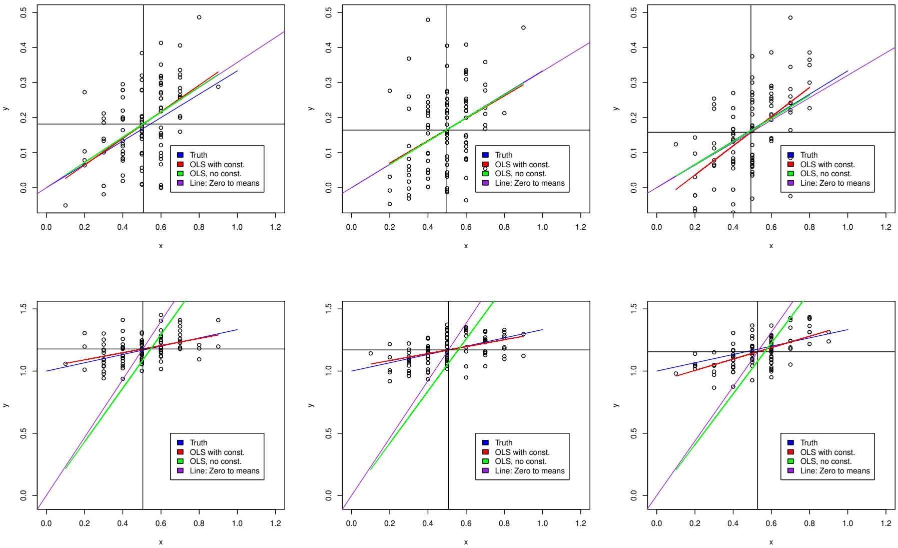
What do you get out of OLS? Properties

$$E(X_j e) = 0 \ \forall j \ \Rightarrow E(e) = 0$$

$$\begin{aligned} E(Y) &= E(X\beta + e) \\ &= E(X)\beta + E(e) = E(X)\beta \end{aligned}$$

- The mean of X multiplied by the OLS coefficient delivers the mean of Y
- OLS "goes through the mean"
- Anchored around the mean of the data

Extra note on problem set 1



What do you get out of OLS? Properties

$$E(X_j e) = 0 \ \forall j \ \Rightarrow E(e) = 0$$

$$\begin{aligned} E(Y) &= E(X\beta + e) \\ &= E(X)\beta + E(e) = E(X)\beta \end{aligned}$$

- The mean of X multiplied by the OLS coefficient delivers the mean of Y
- OLS "goes through the mean"
- Anchored around the mean of the data
- Identified by slope away from the mean

Useful intuition for OLS is "identified by slope away from the mean"

- Suppose we define x not to include the constant, but we keep it in the regression

$$y = x'\beta + \alpha + e \Rightarrow E(y) = E(x'\beta) + E(\alpha) + E(e)$$
$$\mu_y = \mu_x'\beta + \alpha + 0 \Rightarrow \alpha = \mu_y - \mu_x'\beta$$
$$y - \mu_y = (x - \mu_x)'\beta + e \Rightarrow \beta = \left(E[(x - \mu_x)(x - \mu_x)']\right)^{-1} E\left[(x - \mu_x)(y - \mu_y)\right]$$

$$\beta = var(x)^{-1} cov(x, y)$$

- β is just the covariance between x and y scaled by the inverse variance of x
  - Informal: "How x and y move together divided by how much x moves on its own"
- Suppose x is 1-dimensional. You could estimate these two equations:

$$y = \alpha_1 + \beta_1 x + e_1$$
$$x = \alpha_2 + \beta_2 y + e_2$$

- We have an intuition that $\hat{\alpha}_2 = -\hat{\alpha}_1/\hat{\beta}_1$ and $\hat{\beta}_2 = 1/\hat{\beta}_1$
- That isn't true (see the problem set)

Summary so far

- So far: What is OLS?
  - Linear projection of y onto the space spanned by the k column vectors of X
    - Note: A linear projection minimizes the "distance" between the truth and the k-dimensional hyperplane spanned by X
    - "Distance" is the Euclidean norm (i.e., the L^2 norm): It is squared!
  - Scaled covariance of x and y
    - Note: Covariance is the multivariate extension of variance
    - Variance is the second moment: It is squared!
- What about the individual elements of the linear projection coefficient?
  - They are a type of iterated projection
    - What does that mean?
  - They are the conditional covariance
    - Conditional on what?
    - What is covariance actually doing?

Projection matrices: P_X

$$\begin{aligned}\hat{\beta} &= (X'X)^{-1}X'Y \\ \hat{Y} &\equiv X\hat{\beta} \\ &= X(X'X)^{-1}X'Y \\ &\equiv P_X Y\end{aligned}$$

where $P_X = X(X'X)^{-1}X'$ is defined as the projection matrix for X

- P_X is an $n \times n$ matrix
- When you pre-multiply some vector by it, that vector is projected down onto the space spanned by the columns of X
- Intuitively: $P_X X = X(X'X)^{-1}X'X = X$
- Note 1: If you changed X then you'd get a different P_X
- Note 2: P_X is just a deterministic function of X
- Note 3: Hansen calls P_X the "hat matrix" because it's a matrix that gives you fitted values (fitted conditional on X)

Projection matrices: M_X

- P_X is helpful in its own right
- It also helps us define the “annihilator matrix”:

$$M_X = I_n - P_X = I_n - X(X'X)^{-1}X'$$

- M_X just transforms some vector into its residuals from an OLS regression
- What do we know about linear projection and/or minimizing sum of squared errors? The error is orthogonal to the fitted values:

$$\begin{aligned} M_X Y &= \left(I_n - X(X'X)^{-1}X'\right)Y \\ &= Y - X(X'X)^{-1}X'Y \\ &= Y - X\hat{\beta} \\ &= \hat{\varepsilon} \end{aligned}$$

  where $\hat{\varepsilon}$ is the vector of residuals from an OLS regression
- So for any vector Z, $M_X Z$ is just the vector of residuals from an OLS regression regressing Z on X
- Intuitively: $M_X X_k = 0$ for any k that is a column of X

Projection matrices: Summary

- P_X is the projection matrix: When you pre-multiply a vector (or matrix) by it, that vector is projected onto the space spanned by X
- M_X is the annihilator matrix: When you pre-multiply a vector (or matrix) by it, it generates the orthogonal component of that vector, which is orthogonal to the space spanned by X
- These are not "necessary" but useful
- Using them makes it easier to prove a bunch of things about what OLS does and understand which variation drives the results
  - You can prove them other ways, some of which are in Hansen
  - I find that more confusing
  - I encourage you to get an intuitive sense of what the projection and annihilator matrices are, because they make the math and intuition easier to see

Frisch-Waugh-Lovell (FWL) Theorem

- Let's split our $n \times k$ matrix X into an $n \times k_1$ matrix X_1 and an $n \times k_2$ matrix X_2:

$$Y = X\hat{\beta} + \hat{\varepsilon} \quad \Longrightarrow \quad Y = X_1\hat{\beta}_1 + X_2\hat{\beta}_2 + \hat{\varepsilon}$$

  for the same $\hat{\varepsilon}$ (i.e., we're changing nothing but notation)
- Statements: The OLS estimate of β_1 is...
  - ... the linear projection of the part of Y that is orthogonal to X_2 onto the part of X_1 that is orthogonal to X_2
  - ... the (scaled) conditional covariance between Y and X_1 after netting out all of the part of X_1 that is correlated with X_2

Frisch-Waugh-Lovell (FWL) Theorem

- Let's split our $n \times k$ matrix X into an $n \times k_1$ matrix X_1 and an $n \times k_2$ matrix X_2:

$$Y = X\hat{\beta} + \hat{\varepsilon} \quad \Rightarrow \quad Y = X_1\hat{\beta}_1 + X_2\hat{\beta}_2 + \hat{\varepsilon}$$

  for the same $\hat{\varepsilon}$ (i.e., we're changing nothing but notation)
- Statements: The OLS estimate of β_1 is...
  - ... the linear projection of the part of Y that is orthogonal to X_2 onto the part of X_1 that is orthogonal to X_2
  - ... the (scaled) conditional covariance between Y and X_1 after netting out all of the part of X_1 that is correlated with X_2
- Proof:

$$\begin{aligned} M_{X_2}Y &= M_{X_2}X_1\hat{\beta}_1 + M_{X_2}X_2\hat{\beta}_2 + M_{X_2}\hat{\varepsilon} \\ &= M_{X_2}X_1\hat{\beta}_1 + 0\hat{\beta}_2 + M_{X_2}\hat{\varepsilon} \\ &= M_{X_2}X_1\hat{\beta}_1 + \hat{\varepsilon} \end{aligned}$$

  - $M_{X_2}X_2 = 0$ follows from the definition of the annihilator matrix because no part of X_2 orthogonal to X_2
  - $M_{X_2}\hat{\varepsilon} = \hat{\varepsilon}$ follows from the definition of linear projection because all of $\hat{\varepsilon}$ is orthogonal to all of the regressors (including X_2)

Frisch-Waugh-Lovell (FWL) Theorem

- OLS regression 1:

$$Y = X\hat{\beta} + \hat{\varepsilon} \quad \Rightarrow \quad Y = X_1\hat{\beta}_1 + X_2\hat{\beta}_2 + \hat{\varepsilon}$$

- OLS regression 2:

$$M_{X_2}Y = M_{X_2}X_1\hat{\beta}_1 + \hat{\varepsilon}$$

- OLS regression 2 could alternatively be written as:

$$\tilde{Y} = \tilde{X}\hat{\beta}_1 + \hat{\varepsilon}$$

  where $\tilde{Y} = M_{X_2}Y$ is Y residualized of X_2 in some preliminary regression, and $\tilde{X}_1 = M_{X_2}X_1$ is the matrix X_1 residualized of X_2 in k_1 preliminary regressions
- Fact: Both regressions yield identical $\hat{\beta}_1, \hat{\varepsilon}$
  - Standard errors are different
- Statements: The OLS estimate of β_1 is...
  - ... the linear projection of the part of Y that is orthogonal to X_2 onto the part of X_1 that is orthogonal to X_2
  - ... the (scaled) conditional covariance between Y and X_1 after netting out all of the part of X_1 that is correlated with X_2
  - $\hat{\beta}_1 = \left(X_1'M_{X_2}'M_{X_2}X_1\right)^{-1}X_1'M_{X_2}'M_{X_2}Y$

Omitted Variable “Bias” (OV‘B’)

- OLS regression 1:
$$Y = X_1\hat{\beta}_1 + X_2\hat{\beta}_2 + \hat{\varepsilon}$$
- OLS regression 2:
$$Y = X_1\hat{\gamma}_1 + \hat{\nu}$$
(assume that X_1 includes the constant)
- How does $\hat{\gamma}_1$ relate to $\hat{\beta}$?

$$\begin{aligned}
\hat{\gamma}_1 &= (X_1'X_1)^{-1}X_1'Y \\
&= (X_1'X_1)^{-1}X_1'(X_1\hat{\beta}_1 + X_2\hat{\beta}_2 + \hat{\varepsilon}) \\
&= (X_1'X_1)^{-1}X_1'X_1\hat{\beta}_1 + \underbrace{(X_1'X_1)^{-1}X_1'X_2}_{\text{Coef } \hat{\theta} \equiv \text{ Regress } X_1 \text{ on } X_2}\hat{\beta}_2 + \underbrace{(X_1'X_1)^{-1}X_1'\hat{\varepsilon}}_{\text{Coef: Regress } \hat{\varepsilon} \text{ on } X_2} \\
&= \hat{\beta}_1 + \hat{\theta}\hat{\beta}_2
\end{aligned}$$

Omitted Variable “Bias” (OV‘B’)

- OLS regression 1:
$$Y = X_1\hat{\beta}_1 + X_2\hat{\beta}_2 + \hat{\varepsilon}$$
- OLS regression 2:
$$Y = X_1\hat{\gamma}_1 + \hat{\nu}$$
- Relationship:
$$\hat{\gamma}_1 = \hat{\beta}_1 + \hat{\theta}\hat{\beta}_2$$
where $\hat{\theta}$ is a $k_1 \times k_2$ matrix where the i, j element is the coefficient on the j^{th} variable of X_2 that you get when you regress the i^{th} variable on all variables in X_2
- Special case: x_1 and x_2 are scalars:
$$\hat{\theta} = \frac{cov(x_1, x_2)}{var(x_1)}$$
$$\hat{\gamma}_1 = \hat{\beta}_1 + \frac{cov(x_1, x_2)}{var(x_1)}\hat{\beta}_2$$
- When will the “bias” go to zero (i.e., when will $\hat{\gamma}_1$ approach $\hat{\beta}_1$)?
  - When x_2 actually isn’t important for Y ($\hat{\beta}_2 = 0$)
  - When x_2 isn’t related to x_1 $\big(cov(x_1, x_2) = 0\big)$
  - When there’s a lot of identifying variation in x_1 that is not so correlated with x_2 $\big(var(x_1)$ large relative to $cov(x_1, x_2)\big)$

"Bad controls," OVB, and FWL

- Angrist & Pischke coin the term: "Bad controls"
- Example: If you want to estimate causal effect of education on earnings, should you control for occupation?
- OVB: If you don't control for occupation, your estimated projection coefficient includes the projection of earnings onto occupation, multiplied by the projection of occupation onto education
  - Is that bad?
- FWL: If you do control for occupation, you're identified off of whatever variation in education is orthogonal to occupation
  - Is that good?
- For some formal discussion, see these papers:
  - Diegert, Paul, Matthew A. Masten, Alexandre Poirier. "Assessing omitted variable bias when the controls are endogenous." arXiv preprint arXiv:2206.02303 (2023).
  - Masten, Matthew A., and Alexandre Poirier. "The Effect of Omitted Variables on the Sign of Regression Coefficients." arXiv preprint arXiv:2208.00552 (2023).

“Bad controls,” OVB, and FWL

- There’s a reason that I think OLS is worth discussing and teaching *before* ever defining what a causal effect is
- OLS is not only for estimating causal effects
  - But the opposite view (that OLS cannot be used to estimate causal effects) is even more wrong
- “Long regressions” (those which include more controls) are not inherently better for identifying causal effects
- Causal effects are not the main goal of an OLS regression
- Linear projections are
- Key idea: Which variation are you projecting Y onto?

Using the OV'B' formula to understand regressions

- Suppose that there are two groups: college and non-college
- You want to estimate the effect of income on health
- College graduates have higher income, but they also live in different communities and work in different jobs
- Thus, it is plausible that the marginal effect of income on health is different
- This can be written as:

$$H_i = \beta_0 + \beta_1 Y_i + \beta_2 C_i + \beta_3 (Y_i \times C_i) + \varepsilon_i$$

  where H_i, Y_i, C_i are the health, income, and college status (respectively) of individual i
- If $\beta_3 \neq 0$ then the marginal effect of income on health is different for college and non-college people
  - A version of this is if $\bar{Y}$ differs, and the health effects are non-linear (e.g., concave)
  - In this case, the interaction arises from misspecification
  - Feel free to think about that: It corresponds to Figure 2.10 in Section 2.28 ("Limitations of the Best Linear Projection") of Hansen
  - But the point I want to make is more general
  - It also applies when the model is correctly specified, and the heterogeneous treatment effects are "true" features of the real world (not statistical artifacts)

Using the OV'B' formula to understand regressions

- Health as a function of income and education:

$$H_i = \beta_0 + \beta_1 Y_i + \beta_2 C_i + \beta_3 \left(Y_i \times C_i\right) + \varepsilon_i$$

- College slope coefficient: $\beta_1 + \beta_3$
- Non-college slope coefficient: β_1
- If $\beta_3 \neq 0$:
  - Causal interpretation: Marginal effect of income on health is different for college and non-college
  - Descriptive interpretation: The linear projection of health onto income has a different slope for college and non-college
  - I'm fine with a causal or a non-causal interpretation:
    - This example is not about causality or endogeneity
    - This example is not about non-linearities
- Suppose π_C is the college share of the sample

Using the OV'B' formula to understand regressions

- Health as a function of income and education:

$$H_i = \beta_0 + \beta_1 Y_i + \beta_2 C_i + \beta_3 \left(Y_i \times C_i\right) + \varepsilon_i$$

- Suppose π_C is the college share of the sample
- What is the population average "treatment effect"?
  - Recall Law of Total Expectations: $E(X) = P(B_1)E(X|B_1) + P(B_1^c)E(X|B_1^c)$ (B_1^c is complement of B_1)

Using the OV'B' formula to understand regressions

- Health as a function of income and education:

$$H_i = \beta_0 + \beta_1 Y_i + \beta_2 C_i + \beta_3 \big(Y_i \times C_i\big) + \varepsilon_i$$

- Suppose π_C is the college share of the sample
- What is the population average "treatment effect"?
  - Recall Law of Total Expectations: $E(X) = P(B_1)E(X|B_1) + P(B_1^c)E(X|B_1^c)$ (B_1^c is complement of B_1)
- Consider the following limit as $\Delta \longrightarrow 0$

$$\begin{aligned}
E(H|y+\Delta) - E(H|y) =& \pi_C E(H|C=1, y+\Delta) + (1-\pi_C)E(H|C=0, y+\Delta) \\
& - \big[\pi_C E(H|C=1, y) + (1-\pi_C)E(H|C=0, y)\big] \\
=& \pi_C \big[E(H|C=1, y+\Delta) - E(H|C=1, y)\big] \\
& + (1-\pi_C)\big[E(H|C=0, y+\Delta) - E(H|C=0, y)\big] \\
=& \pi_C(\beta_1 + \beta_3) + (1-\pi_C)\beta_1 \\
=& \beta_1 + \pi_C \beta_3
\end{aligned}$$

- Does OLS estimate this?

Using the OV'B' formula to understand regressions

- Health as a function of income and education:

$$H_i = \beta_0 + \beta_1 Y_i + \beta_2 C_i + \beta_3 \left(Y_i \times C_i\right) + \varepsilon_i$$

- Population average "treatment effect": $\beta_1 + \pi_c \beta_3$
- Imagine we estimate regression without the interaction:

$$H_i = \gamma_0 + \gamma_1 Y_i + \gamma_2 C_i + \nu_i$$

- Determine γ_1 by the OV'B' formula:

$$\gamma_1 = \beta_1 + \frac{cov(Y, YC)}{var(Y)} \beta_3$$

- What is $cov(Y, YC)$?

Using the OV'B' formula to understand regressions

- What is $cov(Y, YC)$?
  - Recall that $cov(X, Y) = E(XY) - E(X)E(Y)$
  - For simplicity, let's assume $E(Y|C = 1) = 0$
  - Note that we can always define $\tilde{Y} = Y - E(Y|C = 1)$ which is mean zero, but has the same variance as Y and the same covariance with all other variables that Y does
  - But it simplifies the algebra
  - Let σ_c^2 be $var(Y|C = 1)$
- Also, we'll use Conditional Theorem (Hansen Theorem 2.3): If $E|Y| < \infty$ then

$$E\big(g(X)Y|X\big) = g(X)E(Y|X)$$

Using the OV'B' formula to understand regressions

$$
\begin{aligned}
cov(Y, YC) =& E(Y^2 C) - E(Y)E(YC) \\
=& E(Y^2 C|C = 1)P(C = 1) + E(Y^2 C|C = 0)P(C = 0) \\
& - E(Y)\big(E(YC|C = 1)P(C = 1) + E(YC|C = 0)P(C = 0)\big) \\
& \text{ by Law of Total Expectation} \\
=& E(Y^2|C = 1)(1)P(C = 1) + \underbrace{E(Y^2|C = 0)(0)P(C = 0)}_{=0} \\
& - E(Y)\big(E(Y|C = 1)(1)P(C = 1) + \underbrace{E(Y|C = 0)(0)P(C = 0)}_{=0}\big) \\
& \text{ by Conditioning Theorem} \\
& = E(Y^2|C = 1)\pi_c - E(Y)\underbrace{E(Y|C = 1)}_{=0}\pi_c \\
& = \pi_c\left[\sigma_c^2 + \big(\underbrace{E(Y|C = 1)}_{=0}\big)^2\right] \\
& = \pi_c\sigma_c^2
\end{aligned}
$$

Using the OV'B' formula to understand regressions

- Population average "treatment effect": $\beta_1 + \pi_c \beta_3$
- plim of $\hat{\gamma}_1$:
$$\gamma_1 = \beta_1 + \frac{\sigma_c^2}{\sigma^2} \pi_c \beta_3$$
- OLS does deliver a weighted average of the treatment effects, but it is *not* a population weighted average
- It is a composite population *and* variance weighted average
- Relative to the population average "treatment effect", it is weighted towards whichever group has more variation

OLS as a variance-weighted average

- Population average "treatment effect": $\beta_1 + \pi_c \beta_3$
- plim of $\hat{\gamma}_1$: $\gamma_1 = \beta_1 + \frac{\sigma_c^2}{\sigma^2} \pi_c \beta_3$
  - This is harder to prove when heterogeneity is along some continuous (instead of binary) dimension
  - Still true: OLS disproportionately reflects the slope that exists in the parts of your data where there's the most variation
- Recall FWL: The regression coefficient is identical to the version you'd get if you residualized out any other controls
- This means that the *effective* σ_c^2 is the *residual* variation in Y among college graduates
- If you control for a bunch of stuff that's more strongly correlated with college graduates' income than with non-college income, then that pushes σ_c^2 below σ^2 and your coefficient goes towards β_1
  - This happens even if that other stuff is uncorrelated with health
  - That is, this is a statement about the amount of conditional variance, not about omitted variable "bias"
- Is this a bias? Is that the wrong estimate? Is one of these correct?
- Key question: Where does the identifying variation come from and how does that affect my interpretation?

OLS as a variance-weighted average: Classical measurement error

- The classic (and clearest) way to see the importance of "Where does the identifying variation come from and how does it affect my interpretation?" is **measurement error**
  - Note: Measurement error is substantively important
  - Measurement error is common, and attenuation bias (defined shortly) commonly comes up in seminars
  - But this is usually taught as a substantive concern only, while it actually demonstrates something important about what regressions do
- Suppose x and y are mean zero and x is a scalar
- The true model is given by:
$$y = x\beta + \varepsilon$$
- You cannot estimate this because you don't observe x, you only observe a proxy $\tilde{x}$ that is correlated with x but measured with error: $\tilde{x} \equiv x + v$ where v is a mean zero error term that is uncorrelated with everything
- You can estimate:
$$y = \tilde{x}\hat{\beta} + u$$
- What is $\hat{\beta}$ and what is its relationship to β?

OLS as a variance-weighted average: Classical measurement error

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

$$\tilde{x} = x + \nu$$

$$y = \tilde{x}\hat{\beta} + u$$

$$\begin{aligned}
\hat{\beta} &= \frac{cov(\tilde{x}, y)}{var(\tilde{x})} \\
&= \frac{E\left[(x-\nu)(x\beta+\varepsilon)\right]}{var(x+\nu)} \\
&= \frac{E\left[x^2\beta - \nu x\beta + x\varepsilon - \nu\varepsilon\right]}{var(x) + var(\nu) + 2cov(x,\nu)} \\
&= \frac{E\left[x^2\right]\beta - E\left[\nu x\right]\beta + E\left[x\varepsilon\right] - E\left[\nu\varepsilon\right]}{\sigma_x^2 + \sigma_\nu^2 + 2cov(x,\nu)} \quad (= 0) \\
&= \frac{\sigma_x^2}{\sigma_x^2 + \sigma_\nu^2}\beta < \beta
\end{aligned}$$

OLS as a variance-weighted average: Classical measurement error

$$\hat{\beta} = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_\nu^2}\beta < \beta$$

- Why? $\sigma_x^2/(\sigma_x^2 + \sigma_\nu^2)$ is a signal-to-noise ratio
- How much true variation is there in x, compared to the variation that I see in $\tilde{x}$ (the thing I can actually include in the regression)?
- What's really going on is that $\hat{\beta}$ is a variance-weighted average of the coefficients on its two components: x (which has coefficient β) and ν (which has coefficient 0)
- When most of the identifying variation comes from x, then the coefficient on $\tilde{x}$ will be similar to the coefficient on x
- Note: We said *identifying variation*, which FWL tells us is the variation uncorrelated with other controls
- Classic IO argument in production function estimation:
  - Measured inputs have measurement error
  - Including firm fixed effects exacerbates that but increasing the share of variation due to measurement error
  - Including controls correlated with x but not the measurement error increases attenuation bias

Instrumental variables (Hansen Ch. 12)

- Where does the identifying variation come from and how does that affect my interpretation?
- "Problem": For most interesting X's, the variation comes from all sorts of places
- Which variation is really identifying the coefficient when you have controls?
- How does that change when you do robustness checks?
- What if we could zero in on one source of variation in X?

Instrumental variables (Hansen Ch. 12)

- Let $\hat{X} = P_Z X$
- What is $P_{\hat{X}} Y$?
- Note: This is IV (instrumental variables)/two stage least squares
  - First regress X on Z
  - Save the fitted values
  - Then regress Y on the fitted values
- Traditional motivation is causality
- You will learn that in Econometrics II or Hansen Chapter 12
  - Or Angrist and Pischke's *Mostly Harmless Econometrics*
- That's fine, but I want to emphasize that OLS is incredibly sensitive to non-transparent changes in the source of variation
- IV is always useful as a way of understanding the linear projection onto a specific type of variation in X
- That value does not require IV to isolate a causal effect

Instrumental variables

- Let $\hat{X} = P_Z X$. What is $P_{\hat{X}} Y$?

$$
\begin{aligned}
P_{\hat{X}} Y &= \hat{X}(\hat{X}'\hat{X})^{-1}\hat{X}'Y \\
&= P_Z X(X'P_Z'P_Z X)^{-1}X'P_Z'Y \qquad \text{(because } (AB)' = B'A'\text{)}
\end{aligned}
$$

Note that $\left(A^{-1}\right)' = (A')^{-1} \Rightarrow P_Z' = \left(Z(Z'Z)^{-1}Z'\right)' = P_Z$

$$
\begin{aligned}
P_{\hat{X}} Y &= \underbrace{Z(Z'Z)^{-1}Z'}_{P_Z} X(X' \underbrace{Z(Z'Z)^{-1}Z'}_{P_Z} \underbrace{Z(Z'Z)^{-1}Z'}_{P_Z} X)^{-1} X' \underbrace{Z(Z'Z)^{-1}Z'}_{P_Z} Y \\
&= Z(Z'Z)^{-1}Z'X(X'Z(Z'Z)^{-1}Z'X)^{-1}X'Z(Z'Z)^{-1}Z'Y \\
&= \underbrace{(XX')^{-1}XX'}_{=I_n} Z(Z'Z)^{-1}Z'X(X'Z(Z'Z)^{-1}Z'X)^{-1}X'Z(Z'Z)^{-1}Z'Y \\
&= \underbrace{(XX')^{-1}XX'}_{=I_n} Z(Z'Z)^{-1}Z'Y \\
&= Z(Z'Z)^{-1}Z'Y \\
&= P_Z Y
\end{aligned}
$$

Instrumental variables (Hansen Ch. 12)

- For IV:
  - You first project X onto Z
  - You then project Y onto that projection
- Equivalently, you're just projecting Y onto Z
  - Called the reduced form
- What is the coefficient (simple intuition: Scalar case)?
  - Consider these three regressions

$$
\begin{aligned}
\text{First stage:} \quad & x = \alpha_0 + \alpha z + \varepsilon_1 \\
\text{Second stage:} \quad & y = \beta_0 + \beta \hat{x} + \varepsilon_2 \\
\text{Reduced form:} \quad & y = \gamma_0 + \gamma z + \varepsilon_3
\end{aligned}
$$

  - IV coefficient $\hat{\beta} = \hat{\gamma}/\hat{\alpha}$ (Wald estimator when z, x are binary)
- Why would you do this?
  - Standard answer: Causal inference (Hansen, Econometrics II)
  - X is endogenous, but Z is exogenous and drives some variation in X
    - "drives variation in X" is sometimes called instrument relevance or strength
  - My view: The linear projection of Y onto a known, understandable source of variation is often useful
  - Key thing: This regression is based only on the variation in Z