# Lecture 15

## Quantitative Political Science


Prof. Bisbee

Vanderbilt University


Lecture Date: 2023/10/31


Slides Updated: 2023-12-23

# Agenda

1. Moving from description to inference
2. Unbiasedness
3. OVB

# Inference

- Thus far, $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ is a description of our data
  - $\hat{\beta}_0$ and $\hat{\beta}_1$ are just like the empirical mean or empirical variance
- But we might want to learn about these parameters in the population
  - Just like we use $\bar{Y}$ to learn about $\mu$, we want to find estimators for $\beta_0$ and $\beta_1$
- As before, we want to find **unbiased** and **low variance** estimators

# Unbiasedness

- If we can accept four assumptions, we can use $\hat{\beta}_0$ and $\hat{\beta}_1$ as unbiased estimators for the population parameters

Assumption 1. The relationship between $x$ and $y$ is linear in its parameters, and it is probabilistic

- Relationship is not changing over values of $x$
- True values are defined by $y_i = \beta_0 + \beta_1 x_i + u_i$: **error** $u_i$ means that the relationship between $y$ and $x$ is never **deterministic**. In the population, there is some amount of error.
- Note that $\hat{u}_i$ is the **residual** from our sample, whereas $u_i$ is the inherent error. This relationship is probabilistic.

# Unbiasedness

Assumption 2. sample of $x$ and $y$ is **i.i.d.**

Assumption 3. $VAR(X) \neq 0$

Assumption 4. $u$ has an expected value of zero, no matter what value $x$ takes on

- $E(u|x) = 0$: "zero conditional mean". VERY strong assumption. Requires other things that predict $y$ are **not** correlated with $x$.
- I.e., $income = \beta_0 + \beta_1 education + u$. We know income is predicted by more than education. But in this specification, we are assuming that these other factors are uncorrelated with education.
- Equivalent to writing $cov(u, x) = 0$. But we can't test with $corr(\hat{u}_i, x_i)$ in the sample! This will always be true by construction based on how we calculate $\hat{\beta}_0$ and $\hat{\beta}_1$!

# Unbiasedness of $\hat{\beta}_1$

- $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$
- If $VAR(x) = 0$, this is not defined (hence Assumption 3)
- Note that we can rewrite the numerator as

$$\begin{aligned}
\sum(x_i - \bar{x})(y_i - \bar{y}) &= \sum(x_i - \bar{x})y_i - \sum(x_i - \bar{x})\bar{y} \\
&= \sum(x_i - \bar{x})y_i - [\sum x_i\bar{y} - \sum \bar{x}\bar{y}] \\
&= \sum(x_i - \bar{x})y_i - [n\bar{x}\bar{y} - n\bar{x}\bar{y}] \\
&= \sum(x_i - \bar{x})y_i
\end{aligned}$$

# Unbiasedness of $\hat{\beta}_1$

- So

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum(x_i - \bar{x})y_i}{\sum(x_i - \bar{x})^2} \\ &= \frac{\sum(x_i - \bar{x})(\beta_0 + \beta_1 x_i + u_i)}{SST_x} \\ &= \frac{\beta_0 \sum(x_i - \bar{x}) + \beta_1 \sum(x_i - \bar{x})x_i + \sum(x_i - \bar{x})u_i}{SST_x}\end{aligned}$$

- Note that $\sum(x_i - \bar{x}) = \sum x_i - \sum \bar{x} = n\bar{x} - n\bar{x} = 0$, so the first part of the numerator drops out.

$$\hat{\beta}_1 = \frac{\beta_1 \sum(x_i - \bar{x})x_i + \sum(x_i - \bar{x})u_i}{SST_x}$$

- Let's dig into the second and third parts in order

# Unbiasedness of $\hat{\beta}_1$

$$
\begin{aligned}
\sum(x_i - \bar{x})x_i &= \sum(x_i^2 - x_i\bar{x}) \\
&= \sum x_i^2 - \bar{x}\sum(x_i) \\
&= \sum x_i^2 - n(\bar{x})^2 \\
&= \sum x_i^2 - 2n(\bar{x})^2 + n(\bar{x})^2 \\
&= \sum x_i^2 - 2\bar{x}\sum x_i + \sum(\bar{x})^2 :: \text{since } n\bar{x} = \sum x_i \text{ and } n(\bar{x})^2 = \sum(\bar{x})^2 \\
&= \sum[x_i^2 - 2\bar{x}x_i + (\bar{x})^2] \\
&= \sum(x_i - \bar{x})^2 \\
&= SST_x
\end{aligned}
$$

# Unbiasedness of $\hat{\beta}_1$

- All of this allows us to write $\hat{\beta}_1 = \frac{\beta_1 SST_x + \sum(x_i - \bar{x})u_i}{SST_x}$ which is the same as $\hat{\beta}_1 = \beta_1 + \frac{\sum(x_i - \bar{x})u_i}{SST_x}$
- To find the bias, just take the expectation of $\hat{\beta}_1$
- A trick! **L**aw of **I**terated **E**xpectations (LIE)
    - Expectation of a conditional expectation is just the expectation
    - $E[E[X|Y]] = E[X]$
    - Conditional expectation allows us to treat the condition as a constant
- Use to calculate the expectation of $\hat{\beta}_1$ conditional on $x$: $E[\hat{\beta}_1|x]$

# Unbiasedness of $\hat{\beta}_1$

$$
\begin{aligned}
E(\hat{\beta}_1) &= E[E[\hat{\beta}_1|x]] \\
E[\hat{\beta}_1|x] &= E\left[\beta_1 + \frac{\sum(x_i - \bar{x})u_i}{SST_x} \;\middle|\; x\right] \\
&= E[\beta_1|x] + E\left[\frac{1}{SST_x}\sum(x_i - \bar{x})u_i \;\middle|\; x\right] \\
&= \beta_1 + \frac{1}{SST_x}E[\sum(x_i - \bar{x})u_i|x] \\
&= \beta_1 + \frac{1}{SST_x}\sum(x_i - \bar{x})E[u_i|x]
\end{aligned}
$$

- Assumption 4: $E[u_i|x] = 0$, meaning $E[\hat{\beta}_1|x] = \beta_1$, meaning $E(\hat{\beta}_1) = \beta_1$, meaning unbiased!

# Unbiasedness of $\hat{\beta}_0$

- Recall that $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$
- Note that, since $y_i = \beta_0 + \beta_1 x_i + u_i$, $\bar{y} = \beta_0 + \beta_1 \bar{x} + \bar{u}$

$$
\begin{aligned}
\hat{\beta}_0 &= \beta_0 + \beta_1 \bar{x} + \bar{u} - \hat{\beta}_1 \bar{x} \\
&= \beta_0 + (\beta_1 - \hat{\beta}_1)\bar{x} + \bar{u} \\
E(\hat{\beta}_0) &= E\Bigg[\beta_0 + (\beta_1 - \hat{\beta}_1)\bar{x} + \bar{u}\Bigg] \\
&= E(\beta_0) + E\Bigg[(\beta_1 - \hat{\beta}_1)\bar{x}\Bigg] + E(\bar{u}) \\
&= \beta_0 + \Bigg(E[\beta_1] - E[\hat{\beta}_1]\Bigg)\bar{x} + 0 \\
&= \beta_0 + (\beta_1 - \beta_1)\bar{x} + 0 \\
&= \beta_0
\end{aligned}
$$

# OVB

- What if the true relationship is $y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \nu_i$?
- We don't measure / observe / think about $z$, and model $y_i = \beta_0 + \beta_1 x_i + u_i$
- In practice, we are actually pushing $\beta_2 z_i$ into the error term: $y_i + \beta_0 + \beta_1 x_i + (\beta_2 z_i + \nu_i)$, meaning $u_i = \beta_2 z_i + \nu_i$
- We've just demonstrated that $\hat{\beta}_1 = \beta_1 + \frac{\sum(x_i - \bar{x})u_i}{SST_x}$, but now $u_i = \beta_2 z_i + \nu_i$
- We can calculate the bias as before, with LIE

# OVB

$$
\begin{aligned}
\hat{\beta}_1 &= \beta_1 + \frac{\sum(x_i - \bar{x})(\beta_2 z_i + \nu_i)}{SST_x} \\
E(\hat{\beta}_1) &= E[E(\hat{\beta}_1 | x)] \\
&= E\left[\beta_1 + \frac{\sum(x_i - \bar{x})(\beta_2 z_i + \nu_i)}{SST_x} \;\middle|\; x\right] \\
&= \beta_1 + \frac{\sum(x_i - \bar{x})E[(\beta_2 z_i + \nu_i)]}{SST_x} \\
&= \beta_1 + \beta_2\left[z_i \frac{\sum(x_i - \bar{x})}{SST_x}\right]
\end{aligned}
$$

- Note that $z_i \frac{\sum(x_i - \bar{x})}{SST_x} = \frac{cov(x,z)}{var(x)}$ which is the slope of the coefficient had we regressed $z$ on $x$!

# OVB

- Bias definition: $B(\hat{\theta}) = E(\hat{\theta}) - \theta$
- OVB is just a type of bias:

$$\begin{aligned} B(\hat{\beta}_1) &= E(\hat{\beta}_1) - \beta_1 \\ &= \beta_1 + \beta_2 \frac{cov(x,z)}{var(x)} - \beta_1 \\ &= \beta_2 \frac{cov(x,z)}{var(x)} \end{aligned}$$

- We can **sign** OVB with theory (this is what discussants are always doing)
    - $\beta_2$ is theorized relationship between $z$ and $y$
    - $cov(x,z)$ is theorized relationship between $z$ and $x$

# OVB

- Regress support for Obama ($y$) on Democratic PID ($x$)
  - Omit African-American race ($z$)
- $\beta_2$?
- $cov(x, z)$?
- OVB?