2.4 — OLS: Goodness of Fit and Bias

ECON 480 • Econometrics • Fall 2020

Ryan Safner

Assistant Professor of Economics

safner@hood.edu

ryansafner/metricsF20

metricsF20.classes.ryansafner.com

Outline

Goodness of Fit

Bias: The Sampling Distributions of the OLS Estimators

Bias and Exogeneity

Goodness of Fit

Models

> "All models are wrong. But some are useful." - George Box

Models

> "All models are wrong. But some are useful." - George Box

All of Statistics:

$$Observed_i = \widehat{Model_i} + Error_i$$

Goodness of Fit

- How well does a line fit data? How tightly clustered around the line are the data points?

- Quantify **how much variation in Y_i is "explained" by the model**

$$\underbrace{Y_i}_{Observed} = \underbrace{\widehat{Y_i}}_{Model} + \underbrace{\hat{u}}_{Error}$$

- Recall OLS estimators chosen to minimize **Sum of Squared Errors (SSE)**: $\left(\sum_{i=1}^{n} \hat{u}_i^{\,2}\right)$

Goodness of Fit: R^2

- Primary measure[†] is **regression R-squared**, the fraction of variation in Y explained by variation in predicted values $(\hat{Y})$

$$R^2 = \frac{var(\widehat{Y_i})}{var(Y_i)}$$

[†] Sometimes called the **"coefficient of determination"**

Goodness of Fit: R^2 Formula

$$R^2 = \frac{ESS}{TSS}$$

- **Explained Sum of Squares (ESS)**:[†] sum of squared deviations of *predicted* values from their mean[‡]

$$ESS = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$$

- **Total Sum of Squares (TSS)**: sum of squared deviations of *observed* values from their mean

$$TSS = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

[1] Sometimes called Model Sum of Squares (MSS) or Regression Sum of Squares (RSS) in other textbooks

[2] It can be shown that $\bar{\hat{Y}}_i = \bar{Y}$

Goodness of Fit: R^2 Formula II

- Equivalently, the complement of the fraction of *unexplained* variation in Y_i

$$R^2 = 1 - \frac{SSE}{TSS}$$

- Equivalently, the square of the correlation coefficient between X and Y:

$$R^2 = (r_{X,Y})^2$$

Visualizing R^2

- **Total Variation in Y**: Areas **A** + **C**

$$TSS = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

- **Variation in Y explained by X**: **Area C**

$$ESS = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$$

- **Unexplained variation in Y**: **Area A**

$$SSE = \sum_{i=1}^{n} (\hat{u}_i)^2$$

$$R^2 = \frac{ESS}{TSS} = \frac{C}{A + C}$$

Visualizing R^2

```r
# make a function to calc. sum of sq. devs
sum_sq <- function(x){sum((x - mean(x))^2)}

# find total sum of squares
TSS <- school_reg %>%
  augment() %>%
  summarize(TSS = sum_sq(testscr))

# find explained sum of squares
ESS <- school_reg %>%
  augment() %>%
  summarize(TSS = sum_sq(.fitted))

# look at them and divide to get R^2
tribble(
  ~ESS, ~TSS, ~R_sq,
  ESS, TSS, ESS/TSS
  ) %>%
  knitr::kable()
```

ESS	TSS	R_sq
7794.11	152109.6	0.0512401

$$R^2 = \frac{ESS}{TSS} = \frac{C}{A + C} = 0.05$$

Calculating R^2 in R I

- Recall `broom`'s `augment()` command makes a lot of new regression-based values like:
  - `.fitted`: predicted values ($\hat{Y}_i$)
  - `.resid`: residuals ($\hat{u}_i$)

```
library(broom)
school_reg %>%
  augment() %>%
  head(., n=5) # show first 5 values
```

```
## # A tibble: 5 × 8
##   testscr   str .fitted .resid     .hat .sigma  .cooksd .std.resid
##     <dbl> <dbl>   <dbl>  <dbl>    <dbl>  <dbl>    <dbl>      <dbl>
## 1    691.  17.9    658.   32.7 0.00442   18.5 0.00689        1.76
## 2    661.  21.5    650.   11.3 0.00475   18.6 0.000893       0.612
## 3    644.  18.7    656.  -12.7 0.00297   18.6 0.000700      -0.685
## 4    648.  17.4    659.  -11.7 0.00586   18.6 0.00117       -0.629
## 5    641.  18.7    656.  -15.5 0.00301   18.6 0.00105       -0.836
```

Calculating R^2 in R II

- Or, simpler, can calculate R^2 in `R` as the ratio of variances in model vs. actual

```
# as ratio of variances
school_reg %>%
  augment() %>%
  summarize(r_sq = var(.fitted)/var(testscr)) # var. of *predicted* testscr over var. of *actual* testscr
```

```
## # A tibble: 1 × 1
##      r_sq
##     <dbl>
## 1 0.0512
```

$$R^2 = \frac{var(\hat{Y})}{var(Y)} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \rightarrow \frac{ESS}{TSS}$$

- ESS and TSS are simply the numerators of the variance of $\hat{Y}$ and Y, respectively (i.e. before dividing by $n-1$, which will cancel out).

Goodness of Fit: Standard Error of the Regression

- **Standard Error of the Regression**, $\hat{\sigma}$ or $\hat{\sigma}_u$ is an estimator of the standard deviation of u_i

$$\hat{\sigma}_u = \sqrt{\frac{SSE}{n-2}}$$

- Measures the **average size of the residuals** (distances between data points and the regression line)
  - An average prediction error of the line
  - **Degrees of Freedom correction** of $n-2$: we use up 2 df to first calculate $\hat{\beta}_0$ and $\hat{\beta}_1$!

Calculating SER in R

```
school_reg %>%
  augment() %>%
  summarize(SSE = sum(.resid^2),
            df = n()-2,
            SER = sqrt(SSE/df))
```

```
## # A tibble: 1 × 3
##        SSE    df   SER
##      <dbl> <dbl> <dbl>
## 1 144315.   418  18.6
```

In large samples (where $n - 2 \approx n$), SER $\rightarrow$ standard deviation of the residuals

```
school_reg %>%
  augment() %>%
  summarize(sd_resid = sd(.resid))
```

```
## # A tibble: 1 × 1
##   sd_resid
##      <dbl>
## 1     18.6
```

Goodness of Fit: Looking at R I

- `summary()` command in `Base R` gives:
  - `Multiple R-squared`
  - `Residual standard error` (SER)
  - Calculated with a df of $n-2$

```
# Base R
summary(school_reg)
```

```
##
## Call:
## lm(formula = testscr ~ str, data = CASchool)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -47.727 -14.251   0.483  12.822  48.540
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 698.9330     9.4675  73.825  < 2e-16 ***
## str          -2.2798     0.4798  -4.751 2.78e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.58 on 418 degrees of freedom
## Multiple R-squared:  0.05124,    Adjusted R-squared:  0.04897
## F-statistic: 22.58 on 1 and 418 DF,  p-value: 2.783e-06
```

Goodness of Fit: Looking at R II

```r
# using broom
library(broom)
glance(school_reg)
```

```
## # A tibble: 1 × 12
##   r.squared adj.r.squared sigma statistic    p.value    df logLik   AIC   BIC
##       <dbl>         <dbl> <dbl>     <dbl>      <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1    0.0512        0.0490  18.6      22.6 0.00000278     1 -1822. 3650. 3663.
## # … with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

- `r.squared` is `0.05` $\implies$ about 5% of variation in `testscr` is explained by our model
- `sigma` (SER) is `18.6` $\implies$ average test score is about 18.6 points above/below our model's prediction

```r
# extract it if you want with pull
school_r_sq <- glance(school_reg) %>% pull(r.squared)
school_r_sq
```

```
## [1] 0.0512401
```

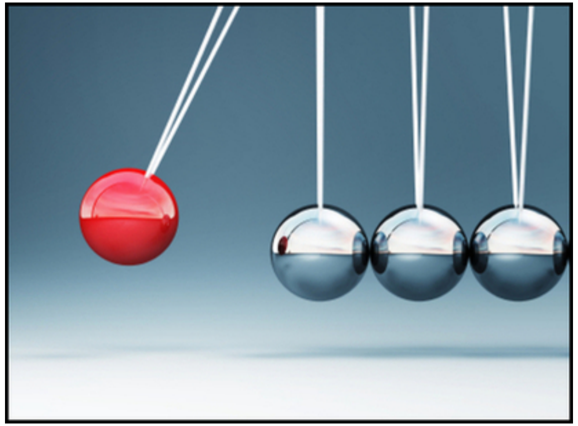
Bias: The Sampling Distributions of the OLS Estimators

Recall: The Two Big Problems with Data

- We use econometrics to **identify** causal relationships and make **inferences** about them

1. Problem for **identification**: **endogeneity**
    - X is **exogenous** if its variation is *unrelated* to other factors (u) that affect Y
    - X is **endogenous** if its variation is *related* to other factors (u) that affect Y

2. Problem for **inference**: **randomness**
    - Data is random due to **natural sampling variation**
    - Taking one sample of a population will yield slightly different information than another sample of the same population



Distributions of the OLS Estimators

- OLS estimators ($\hat{\beta}_0$ and $\hat{\beta}_1$) are computed from a finite (specific) sample of data

- Our OLS model contains **2 sources of randomness**:

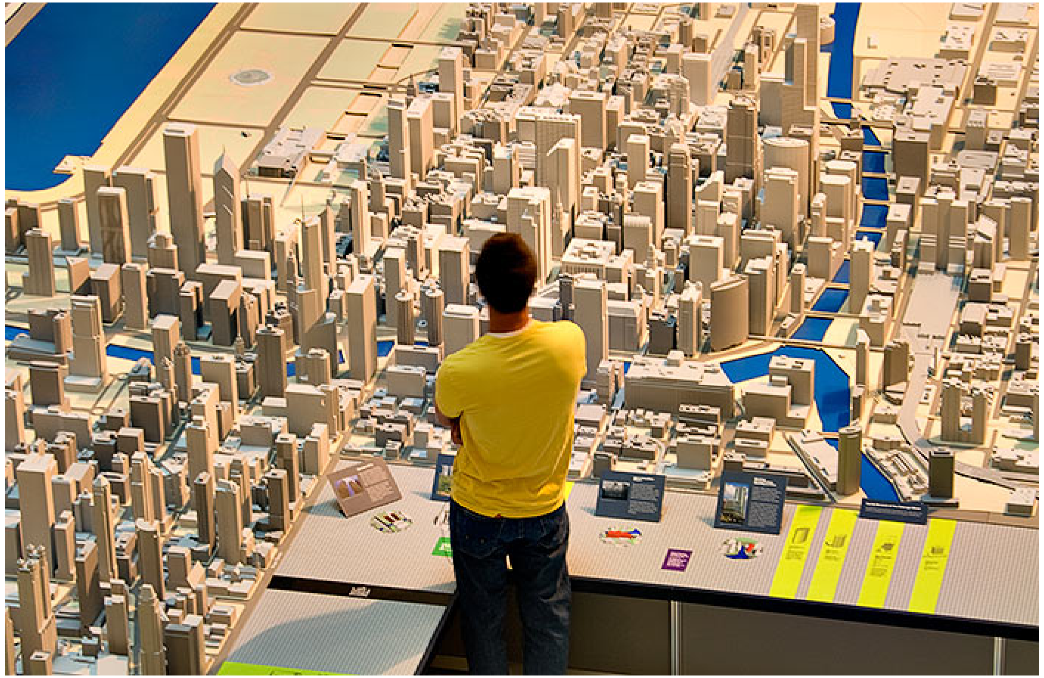
- ***Modeled* randomness**: u includes all factors affecting Y *other* than X

  - different samples will have different values of those other factors (u_i)

- ***Sampling* randomness**: different samples will generate different OLS estimators
  - Thus, $\hat{\beta}_0, \hat{\beta}_1$ are *also* **random variables**, with their own **sampling distribution**

Inferential Statistics and Sampling Distributions

- **Inferential statistics** analyzes a **sample** to make inferences about a much larger (unobservable) **population**
- **Population**: all possible individuals that match some well-defined criterion of interest
  - Characteristics about (relationships between variables describing) populations are called **"parameters"**
- **Sample**: some portion of the population of interest to *represent the whole*
  - Samples examine part of a population to generate **statistics** used to **estimate** population **parameters**



Sampling Basics

Example: Suppose you randomly select 100 people and ask how many hours they spend on the internet each day. You take the mean of your sample, and it comes out to 5.4 hours.

- 5.4 hours is a **sample statistic** describing the sample; we are more interested in the corresponding **parameter** of the relevant population (e.g. all Americans)

- If we take another sample of 100 people, would we get the same number?

- Roughly, but probably not exactly

- **Sampling variability** describes the effect of a statistic varying somewhat from sample to sample
    - This is *normal*, not the result of any error or bias!

I.I.D. Samples

- If we collect many samples, and each sample is randomly drawn from the population (and then replaced), then the distribution of samples is said to be **independently and identically distributed (i.i.d.)**

- Each sample is **independent** of each other sample (due to replacement)

- Each sample comes from the **identical** underlying population distribution

The Sampling Distribution of OLS Estimators

- Calculating OLS estimators for a sample makes the OLS estimators *themselves* random variables:

- Draw of i is random $\implies$ value of each (X_i, Y_i) is random $\implies \hat{\beta}_0, \hat{\beta}_1$ are random

- Taking different samples will create different values of $\hat{\beta}_0, \hat{\beta}_1$

- Therefore, $\hat{\beta}_0, \hat{\beta}_1$ each have a **sampling distribution** across different samples

The Central Limit Theorem

- **Central Limit Theorem (CLT)**: if we collect samples of size n from the same population and generate a sample statistic (e.g. OLS estimator), then with large enough n, the distribution of the sample statistic is approximately normal IF
  1. $n \geq 30$
  2. Samples come from a *known* normal distribution $\sim N(\mu, \sigma)$

- If neither of these are true, we have other methods (coming shortly!)

- One of the most fundamental principles in all of statistics

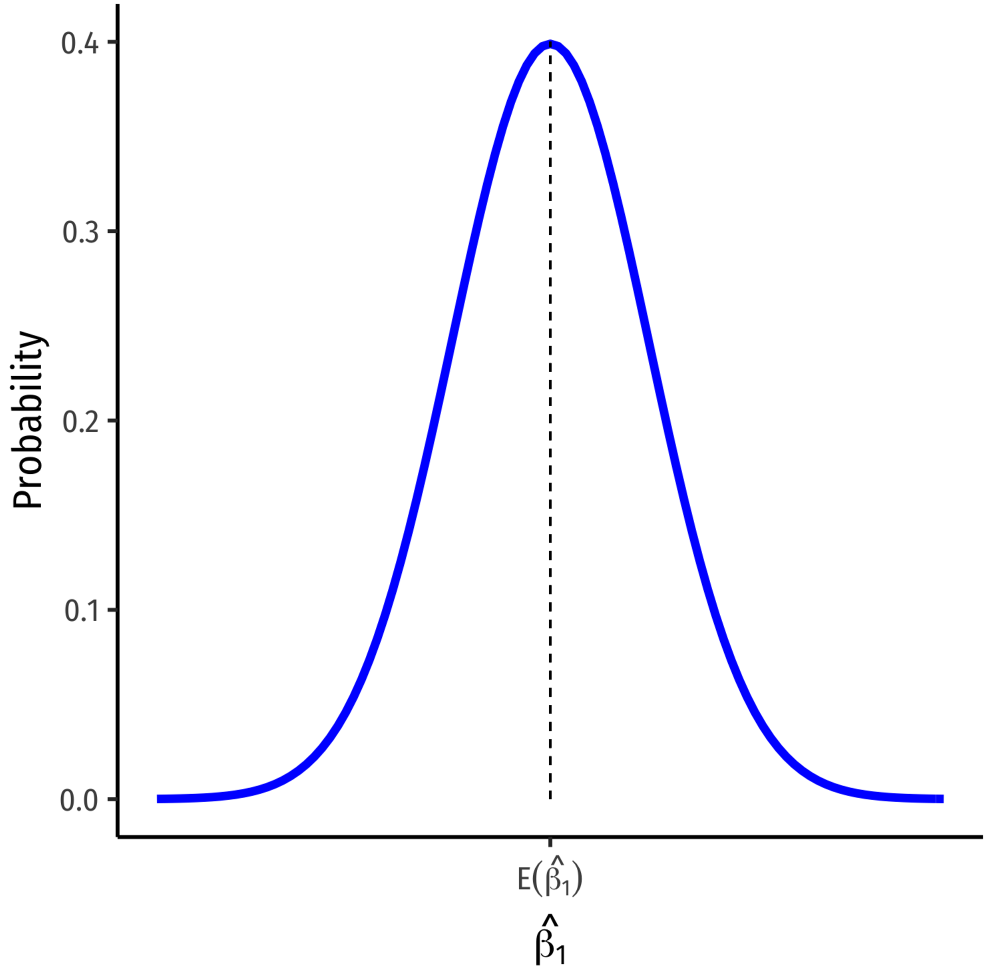
- Allows for virtually all testing of statistical hypotheses $\rightarrow$ estimating probabilities of values on a normal distribution

The Sampling Distribution of $\hat{\beta}_1$ I

- The CLT allows us to approximate the sampling distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$ as normal

- We care about $\hat{\beta}_1$ (slope) since it has economic meaning, rarely about $\hat{\beta}_0$ (intercept)

$$\hat{\beta}_1 \sim N(E[\hat{\beta}_1], \sigma_{\hat{\beta}_1})$$



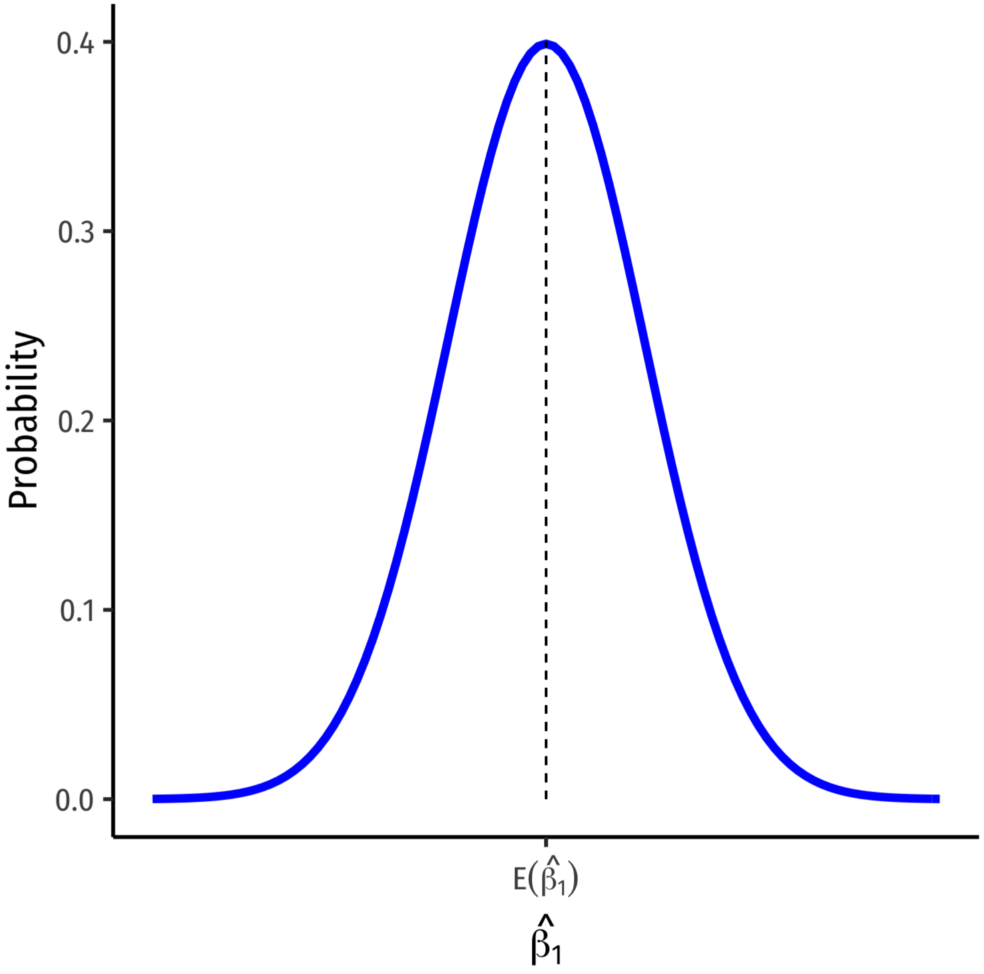
The Sampling Distribution of $\hat{\beta}_1$ II

$$\hat{\beta}_1 \sim N(E[\hat{\beta}_1], \sigma_{\hat{\beta}_1})$$

- We want to know:

1. $E[\hat{\beta}_1]$; what is the **center** of the distribution? (today)

2. $\sigma_{\hat{\beta}_1}$; how **precise** is our estimate? (next class)



Bias and Exogeneity

Assumptions about Errors I

- In order to talk about $E[\hat{\beta}_1]$, we need to talk about u
- Recall: u is a random variable, and we can never measure the error term

Assumptions about Errors II

- We make **4 critical assumptions about u**:

Assumptions about Errors II

- We make **4 critical assumptions about u**:

1. The expected value of the residuals is 0

$$E[u] = 0$$

Assumptions about Errors II

- We make **4 critical assumptions about u**:

1. The expected value of the residuals is 0

$$E[u] = 0$$

2. The variance of the residuals over X is constant:

$$var(u|X) = \sigma_u^2$$

Assumptions about Errors II

- We make **4 critical assumptions about u**:

1. The expected value of the residuals is 0

$$E[u] = 0$$

2. The variance of the residuals over X is constant:

$$var(u|X) = \sigma^2_u$$

3. Errors are not correlated across observations:

$$cor(u_i, u_j) = 0 \quad \forall i \neq j$$

Assumptions about Errors II

- We make **4 critical assumptions about u**:

1. The expected value of the residuals is 0

$$E[u] = 0$$

2. The variance of the residuals over X is constant:

$$var(u|X) = \sigma^2_u$$

3. Errors are not correlated across observations:

$$cor(u_i, u_j) = 0 \quad \forall i \neq j$$

4. There is no correlation between X and the error term:

$$cor(X, u) = 0 \text{ or } E[u|X] = 0$$

Assumptions 1 and 2: Errors are i.i.d.

1. The expected value of the residuals is 0

$$E[u] = 0$$

2. The variance of the residuals over X is constant:

$$var(u|X) = \sigma_u^2$$

- The first two assumptions $\implies$ errors are **i.i.d.**, drawn from the same distribution with mean 0 and variance σ_u^2

Assumption 2: Homoskedasticity

- The variance of the residuals over X is constant:

$$var(u|X) = \sigma_u^2$$

- Assumption 2 implies that errors are **"homoskedastic"**: they have the same variance across X

- Often this assumption is violated: errors may be **"heteroskedastic"**: they do not have the same variance across X

- This *is* a problem for **inference**, but we have a simple fix for this (next class)

Assumption 3: No Serial Correlation

- Errors are not correlated across observations:

$$cor(u_i, u_j) = 0 \quad \forall i \neq j$$

- For simple cross-sectional data, this is rarely an issue

- Time-series & panel data nearly always contain **serial correlation** or **autocorrelation** between errors

- e.g. "this week's sales look a lot like last weel's sales, which look like...etc"

- There are fixes to deal with autocorrelation (coming much later)

Assumption 4: The Zero Conditional Mean Assumption

- No correlation between X and the error term:

$$cor(X,u)=0$$

- **This is the absolute killer assumption, because it assumes exogeneity**

- Often called the **Zero Conditional Mean** assumption:

$$E[u|X]=0$$

> "Does knowing X give me any useful information about u?"
>
> - If yes: model is **endogenous**, **biased** and **not-causal**!

Exogeneity and Unbiasedness

- $\hat{\beta}_1$ is **unbiased** iff there is no systematic difference, on average, between sample values of $\hat{\beta}_1$ and **true population parameter** β_1, i.e.

$$E[\hat{\beta}_1] = \beta_1$$

- Does *not* mean any sample gives us $\hat{\beta}_1 = \beta_1$, only the **estimation procedure** will, *on average*, yield the correct value

- Random errors above and below the true value cancel out (so that on average, $E[\hat{u}|X] = 0$)

Sidenote: Statistical Estimators I

- In statistics, an **estimator** is a rule for calculating a statistic (about a population parameter)

Example: We want to estimate the average height (H) of U.S. adults (population) and have a random sample of 100 adults.

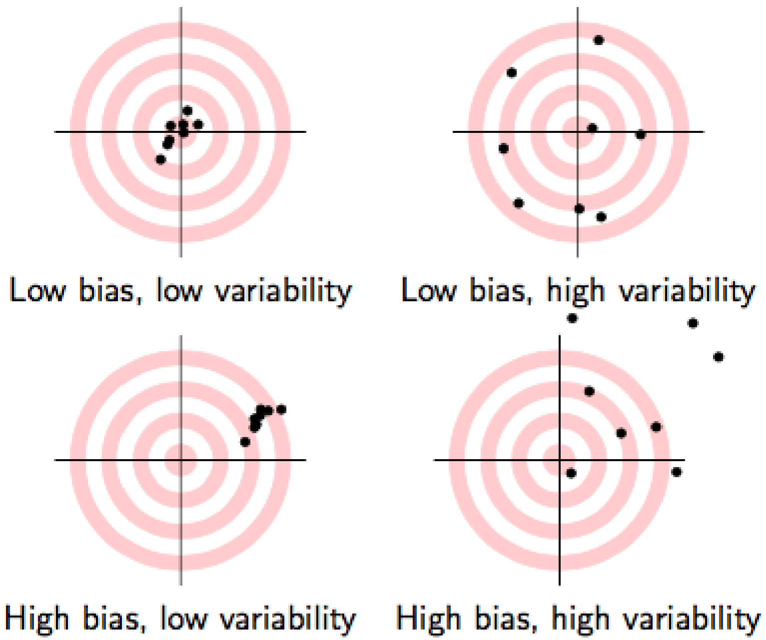
- Calculate the mean height of our sample $(\bar{H})$ to estimate the true mean height of the population (μ_H)
- $\bar{H}$ is an **estimator** of μ_H
- There are many estimators we *could* use to estimate μ_H
  - How about using the first value in our sample: H_1 ?

Sidenote: Statistical Estimators II

- What makes one estimator (e.g. $\bar{H}$) better than another (e.g. H_1)?[†]

1. **Biasedness**: does the estimator give us the true parameter *on average*?

2. **Efficiency**: an estimator with a smaller variance is better



[†] Technically, we also care about **consistency**: minimizing uncertainty about the correct value. The Law of Large Numbers, similar to CLT, permits this. We don't need to get too advanced about probability in this class.

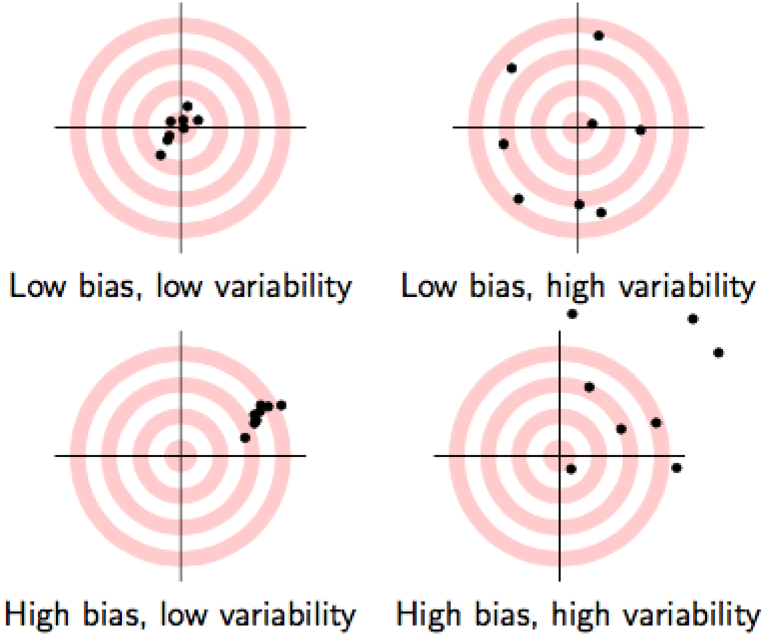
Exogeneity and Unbiasedness I

- $\hat{\beta}_1$ is the **Best Linear Unbiased Estimator (BLUE)** estimator of β_1 **when X is exogenous**[†]

- No systematic difference, on average, between sample values of $\hat{\beta}_1$ and the true population β_1:

$$E[\hat{\beta}_1] = \beta_1$$

- Does *not* mean that each sample gives us $\hat{\beta}_1 = \beta_1$, only the estimation **procedure** will, **on average**, yield the correct value



[†] The proof for this is known as the famous Gauss-Markov Theorem. See today's class notes for a simplified proof.

Exogeneity and Unbiasedness II

- Recall, an **exogenous** variable (X) is unrelated to other factors affecting Y, i.e.:

$$cor(X, u) = 0$$

- Again, this is called the **Zero Conditional Mean Assumption**

$$E(u|X) = 0$$

- For any known value of X, the expected value of u is 0

- Knowing the value of X must tell us *nothing* about the value of u (anything else relevant to Y other than X)

- We can then confidently assert causation: $X \rightarrow Y$

Endogeneity and Bias

- Nearly all independent variables are **endogenous**, they **are** related to the error term u

$$cor(X, u) \neq 0$$

Example: Suppose we estimate the following relationship:

$$\text{Violent crimes}_t = \beta_0 + \beta_1 \text{Ice cream sales}_t + u_t$$

- We find $\hat{\beta_1} > 0$
- Does this mean Ice cream sales $\rightarrow$ Violent crimes?

Endogeneity and Bias: Takeaways

- The true expected value of $\hat{\beta}_1$ is actually:†

$$E[\hat{\beta}_1] = \beta_1 + cor(X,u)\frac{\sigma_u}{\sigma_X}$$

1) If X is exogenous: $cor(X,u) = 0$, we're just left with β_1

2) The larger $cor(X,u)$ is, larger **bias**: $\left(E[\hat{\beta}_1] - \beta_1\right)$

3) We can **“sign”** the direction of the bias based on $cor(X,u)$

- **Positive** $cor(X,u)$ overestimates the true β_1 ($\hat{\beta}_1$ is too high)
- **Negative** $cor(X,u)$ underestimates the true β_1 ($\hat{\beta}_1$ is too low)

† See today's class notes for proof.

Endogeneity and Bias: Example I

Example:

$$wages_i = \beta_0 + \beta_1 education_i + u$$

- Is this an accurate reflection of $education \rightarrow wages$?

- Does $E[u|education] = 0$?

- What would $E[u|education] > 0$ mean?

Endogeneity and Bias: Example II

Example:

$$\text{per capita cigarette consumption} = \beta_0 + \beta_1 \text{State cig tax rate} + u$$

- Is this an accurate reflection of $taxes \rightarrow consumption$?
- Does $E[u|tax] = 0$?
- What would $E[u|tax] > 0$ mean?