ECON4150 - Introductory Econometrics

Lecture 6: OLS with Multiple Regressors

Monique de Haan
(moniqued@econ.uio.no)

Stock and Watson Chapter 6

Lecture outline

- Violation of first Least Squares assumption
- Omitted variable bias
  - violation of unbiasedness
  - violation of consistency
- Multiple regression model
  - 2 regressors
  - k regressors
- Perfect multicollinearity
- Imperfect multicollinearity
- Properties OLS estimators in multiple regression model

Violation of first Least Squares assumption

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

Assumption 1: The conditional mean of u_i given X_i is zero

$$E(u_i|X_i) = 0$$

The first OLS assumption states that:

All other factors that affect the dependent variable Y_i (contained in u_i) are unrelated to X_i in the sense that, given a value of X_i, the mean of these other factors equals zero.

In the class size example:

All the other factors affecting test scores should be unrelated to class size in the sense that, given a value of class size, the mean of these other factors equals zero.

Violation of first Least Squares assumption

Suppose that

- districts with small classes have few immigrants (few English learners)
- districts with large classes have many immigrants (many English learners)

In this case class size is related to percentage of English learners

Students who are still learning English likely have lower test scores

Which implies that percentage of English learners is contained in u_i.

This implies a violation of assumption 1:

$$E\left(u_i|ClassSize_i = small\right) \neq E\left(u_i|ClassSize_i = large\right) \neq 0$$

Omitted variable bias

- The variable measuring the percentage of English learners in a district ($el\ pct_i$) is omitted from the simple regression model

$$TestScore_i = \beta_0 + \beta_1 ClassSize_i + u_i$$

- Omitting a variable from a regression analysis will lead to **omitted variable bias** if:

1. The omitted variable is correlated to the included regressor of interest.
2. The omitted variable is a determinant of the dependent variable.

Omitted variable bias

```
. corr class_size el_pct
(obs=420)
```

	class_~e	el_pct
class_size	**1.0000**	
el_pct	**0.1876**	**1.0000**

```
. corr test_score el_pct
(obs=420)
```

	test_s~e	el_pct
test_score	**1.0000**	
el_pct	**-0.6441**	**1.0000**

Both conditions for omitted variable bias seem to be met

1. The percentage of English learners is correlated with class size
2. The percentage of English learners is correlated with test scores

- If we omit percentage of English learners from regression, $\widehat{\beta}_1^{OLS}$ will not only estimate effect of class size on district average test scores
- but it will also "pick up" the effect of the percentage of English learners in the district on district average test scores
- $\widehat{\beta}_1^{OLS}$ is biased and inconsistent.

Omitted variable bias: violation of unbiasedness

True model : $Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_i + u_i \qquad E(u_i|X_i, W_i) = 0$

Estimated model : $Y_i = \beta_0 + \beta_1 X_i + v_i$

$$E\left[\widehat{\beta}_1\right] = E\left[\frac{\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{\sum_{i=1}^{n}(X_i-\overline{X})(X_i-\overline{X})}\right]$$

substitute for $Y_i, \overline{Y}$ (*true model*!)

$$= E\left[\frac{\sum_{i=1}^{n}(X_i-\overline{X})(\beta_0+\beta_1 X_i+\beta_2 W_i+u_i-(\beta_0+\beta_1\overline{X}+\beta_2\overline{W}+\overline{u}))}{\sum_{i=1}^{n}(X_i-\overline{X})(X_i-\overline{X})}\right]$$

rewrite (β_0 *drops out*)

$$= E\left[\frac{\sum_{i=1}^{n}(X_i-\overline{X})(\beta_1(X_i-\overline{X})+\beta_2(W_i-\overline{W})+(u_i-\overline{u}))}{\sum_{i=1}^{n}(X_i-\overline{X})(X_i-\overline{X})}\right]$$

Omitted variable bias: violation of unbiasedness

$$E\left[\widehat{\beta}_1\right] = E\left[\frac{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)\left(\beta_1(X_i-\overline{X})+\beta_2\left(W_i-\overline{W}\right)+(u_i-\overline{u})\right)}{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)\left(X_i-\overline{X}\right)}\right]$$

rewrite & use expectation rules

$$= \beta_1 + E\left[\frac{\beta_2\sum_{i=1}^{n}\left(X_i-\overline{X}\right)\left(W_i-\overline{W}\right)}{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)\left(X_i-\overline{X}\right)}\right] + E\left[\frac{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)(u_i-\overline{u})}{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)\left(X_i-\overline{X}\right)}\right]$$

put β_2 in front of expectation & use "algebra trick"

$$= \beta_1 + \beta_2 E\left[\frac{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)\left(W_i-\overline{W}\right)}{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)\left(X_i-\overline{X}\right)}\right] + E\left[\frac{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)u_i}{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)\left(X_i-\overline{X}\right)}\right]$$

law of iterated expectations

$$= \beta_1 + \beta_2 E\left[\frac{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)\left(W_i-\overline{W}\right)}{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)\left(X_i-\overline{X}\right)}\right] + E\left[\frac{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)E(u_i|X_i,W_i)}{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)\left(X_i-\overline{X}\right)}\right]$$

Omitted variable bias: violation of unbiasedness

$$E\left[\widehat{\beta}_1\right] = \beta_1 + \beta_2 E\left[\frac{\sum_{i=1}^{n}(X_i-\overline{X})(W_i-\overline{W})}{\sum_{i=1}^{n}(X_i-\overline{X})(X_i-\overline{X})}\right] + E\left[\frac{\sum_{i=1}^{n}(X_i-\overline{X})E(u_i|X_i,W_i)}{\sum_{i=1}^{n}(X_i-\overline{X})(X_i-\overline{X})}\right]$$

$$\textit{by assumption } E\left(u_i|X_i, W_i\right) = 0$$

$$= \beta_1 + \beta_2 E\left[\frac{\sum_{i=1}^{n}(X_i-\overline{X})(W_i-\overline{W})}{\sum_{i=1}^{n}(X_i-\overline{X})(X_i-\overline{X})}\right]$$

- If W_i is unrelated to X_i $\left(E\left[\frac{\sum_{i=1}^{n}(X_i-\overline{X})(W_i-\overline{W})}{\sum_{i=1}^{n}(X_i-\overline{X})(X_i-\overline{X})}\right] = 0\right)$ this implies that $E\left[\widehat{\beta}_1\right] = \beta_1$
- If W_i is no determinant of Y_i ($\beta_2 = 0$) this implies that $E\left[\widehat{\beta}_1\right] = \beta_1$
- The second term is only nonzero if both conditions for omitted variable bias are met
- If the second term is nonzero $\widehat{\beta}_1$ is biased!

Omitted variable bias: violation of consistency

True model : $Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_i + u_i \qquad E(u_i|X_i, W_i) = 0$

Estimated model : $Y_i = \beta_0 + \beta_1 X_i + v_i$

$$Plim\ \widehat{\beta}_1 = \frac{Plim\ \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{Plim\ \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})(X_i-\overline{X})} = \frac{Plim\ s_{XY}}{Plim\ s_X^2}$$

$$= \frac{Cov(X_i, Y_i)}{Var(X_i)}$$

substitute true model for Y_i

$$= \frac{Cov(X_i, \beta_0+\beta_1 X_i+\beta_2 W_i+u_i)}{Var(X_i)}$$

Covariance rules Key concept 2.3

$$= \frac{Cov(X_i, \beta_1 X_i)+Cov(X_i, \beta_2 W_i)+Cov(X_i, u_i)}{Var(X_i)}$$

Omitted variable bias: violation of consistency

$$Plim\ \widehat{\beta}_1 = \frac{Cov(X_i,\beta_1 X_i)+Cov(X_i,\beta_2 W_i)+Cov(X_i,u_i)}{Var(X_i)}$$

$$Cov(X_i, u_i) = 0 \text{ because } E(u_i|X_i, W_i) = 0$$

$$= \beta_1 \frac{Cov(X_i,X_i)}{Var(X_i)} + \beta_2 \frac{Cov(X_i,W_i)}{Var(X_i)}$$

$$Cov(X_i, X_i) = Var(X_i)$$

$$= \beta_1 + \beta_2 \frac{Cov(X_i,W_i)}{Var(X_i)}$$

Omitted variable bias: violation of consistency

$$\widehat{\beta}_1 \quad \xrightarrow{p} \beta_1 + \beta_2 \frac{Cov(X_i, W_i)}{Var(X_i)}$$

- If W_i is unrelated to X_i ($Cov(X_i, W_i) = 0$) this implies that $\widehat{\beta}_1 \xrightarrow{p} \beta_1$
- If W_i is no determinant of Y_i ($\beta_2 = 0$) this implies that $\widehat{\beta}_1 \xrightarrow{p} \beta_1$
- If both omitted variable bias conditions are met $\widehat{\beta}_1$ is inconsistent!

Omitted variable bias: violation of consistency

From the **omitted variable bias formula**

$$\widehat{\beta}_1 \xrightarrow{p} \beta_1 + \beta_2 \frac{Cov\left(X_i, W_i\right)}{Var\left(X_i\right)}$$

we can infer the direction of the bias of $\widehat{\beta}_1$ that persists in large samples

- Suppose W_i has a positive effect on Y_i, then $\beta_2 > 0$
- Suppose X_i and W_i are positively correlated, then $Cov\left(X_i, W_i\right) > 0$
- This implies that $\widehat{\beta}_1$ is upward biased, it converges in probability to a larger number than the true value of β_1

Omitted variable bias: a simulation example

- Lets create a data set with 100 observations
- $W_i \sim N(0,1)$
- We let X_i depend on W_i: $X_i = W_i + \varepsilon_i \quad \varepsilon_i \sim N(0,1)$
- $u_i \sim N(0,1)$
- We define the true population model as:

$$Y_i = 1 + 2X_i + W_i + u_i \quad \beta_1 = 2 \quad \& \quad \beta = 1$$

```
set obs 100
gen w = rnormal()
gen x = w + rnormal()
gen y = 1 + 2*x + w + rnormal()
```

```
. sum y x w
```

Variable	Obs	Mean	Std. Dev.	Min	Max
y	100	1.501122	3.629103	-7.468484	10.27467
x	100	.164158	1.310894	-3.099808	3.644282
w	100	.1819518	1.081655	-2.565364	2.845132

Omitted variable bias: a simulation example

True model : $Y_i = 1 + 2X_i + W_i + u_i \qquad E(u_i|X_i, W_i) = 0$

Estimated model : $Y_i = \beta_0 + \beta_1 X_i + v_i$

```
. regress y x
```

Source	SS	df	MS
Model	**1173.01332**	**1**	**1173.01332**
Residual	**130.855339**	**98**	**1.33525856**
Total	**1303.86866**	**99**	**13.1703905**

Number of obs =	**100**
F(1, 98) =	**878.49**
Prob > F =	**0.0000**
R-squared =	**0.8996**
Adj R-squared =	**0.8986**
Root MSE =	**1.1555**

y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
x	**2.625828**	**.0885926**	**29.64**	**0.000**	**2.450019**	**2.801637**
_cons	**1.070071**	**.116465**	**9.19**	**0.000**	**.83895**	**1.301192**

Omitted variable bias: a simulation example

We can create 999 of these data sets with 100 observations and use OLS to estimate

$$Y_i = \beta_0 + \beta_1 X_i + v_i$$

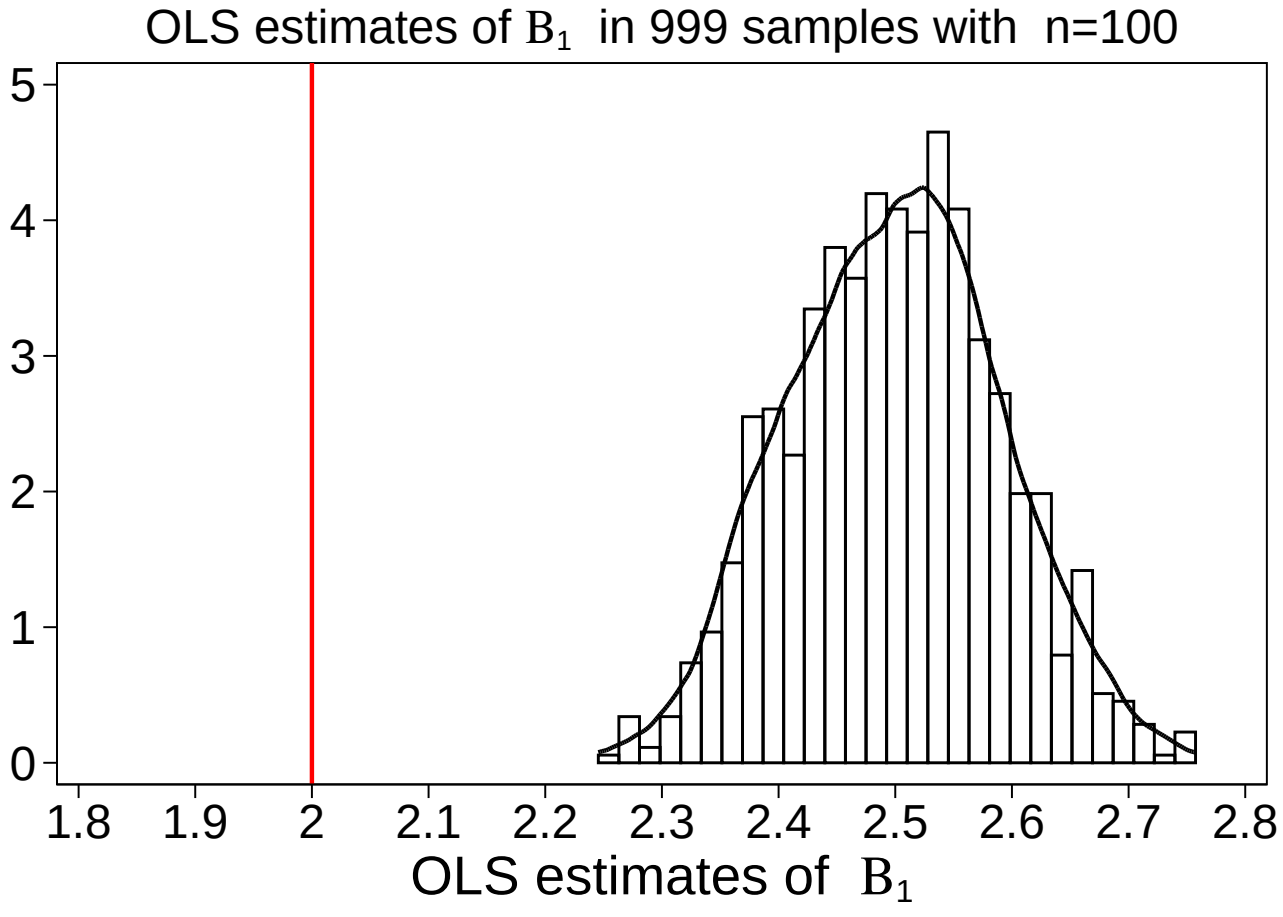
```
1 . program define ols, rclass
    1.  drop _all
    2.  set obs 100
    3.  gen w=rnormal()
    4.  gen x=w+rnormal()
    5.  gen y=1+2*x+w+rnormal()
    6.  regress y x
    7. end

2 .
3 . simulate _b, reps(999)  nodots : ols

      command:  ols


4 . sum
```

Variable	Obs	Mean	Std. Dev.	Min	Max
_b_x	**999**	**2.49988**	**.0897328**	**2.245482**	**2.757368**
_b_cons	**999**	**1.001677**	**.121082**	**.6290014**	**1.383819**

# Omitted variable bias: a simulation example
n=100

True model : $Y_i = 1 + 2X_i + W_i + u_i \qquad E(u_i|X_i, W_i) = 0$

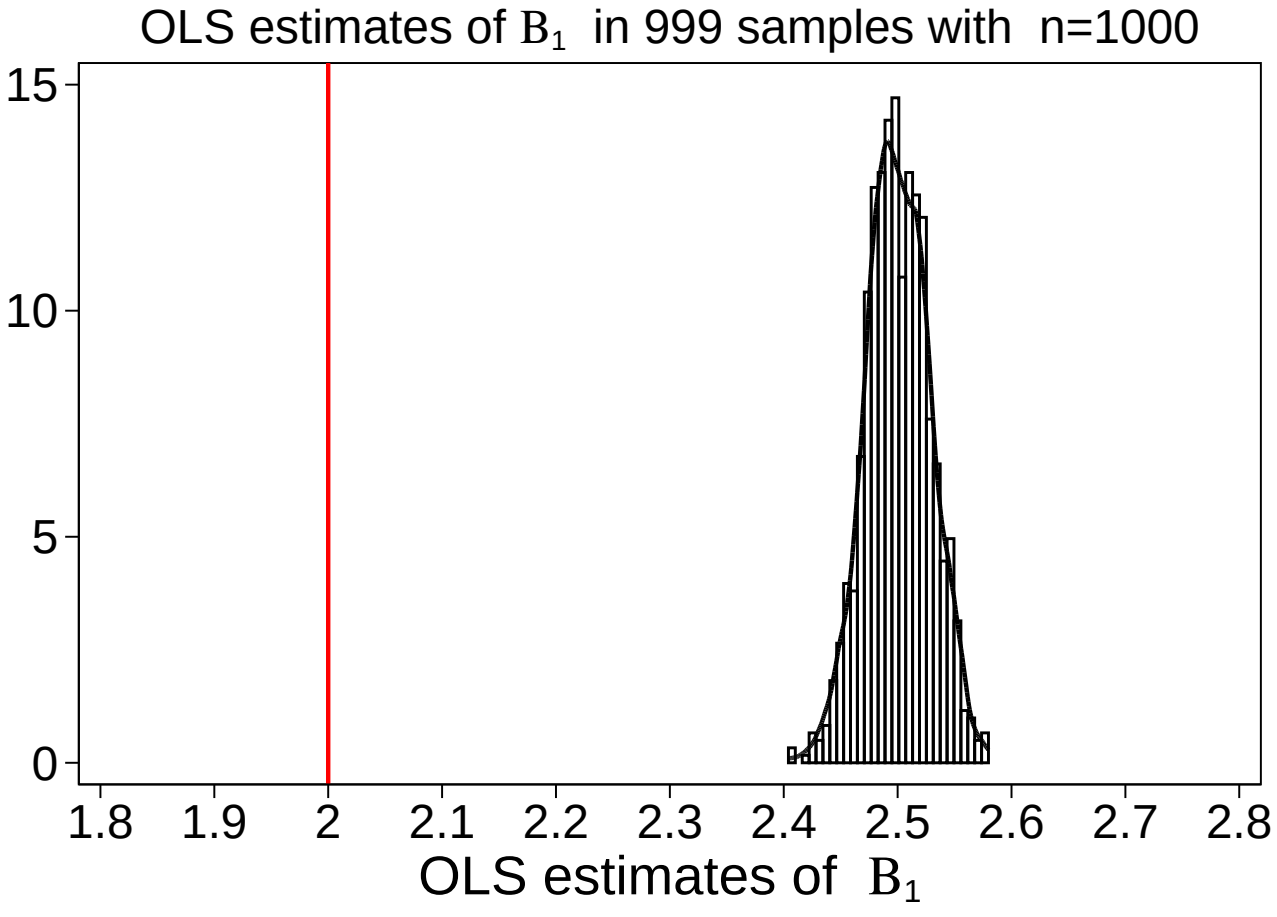
Estimated model : $Y_i = \beta_0 + \beta_1 X_i + v_i$



# Omitted variable bias: a simulation example
n=1000

True model : $Y_i = 1 + 2X_i + W_i + u_i \qquad E(u_i|X_i, W_i) = 0$

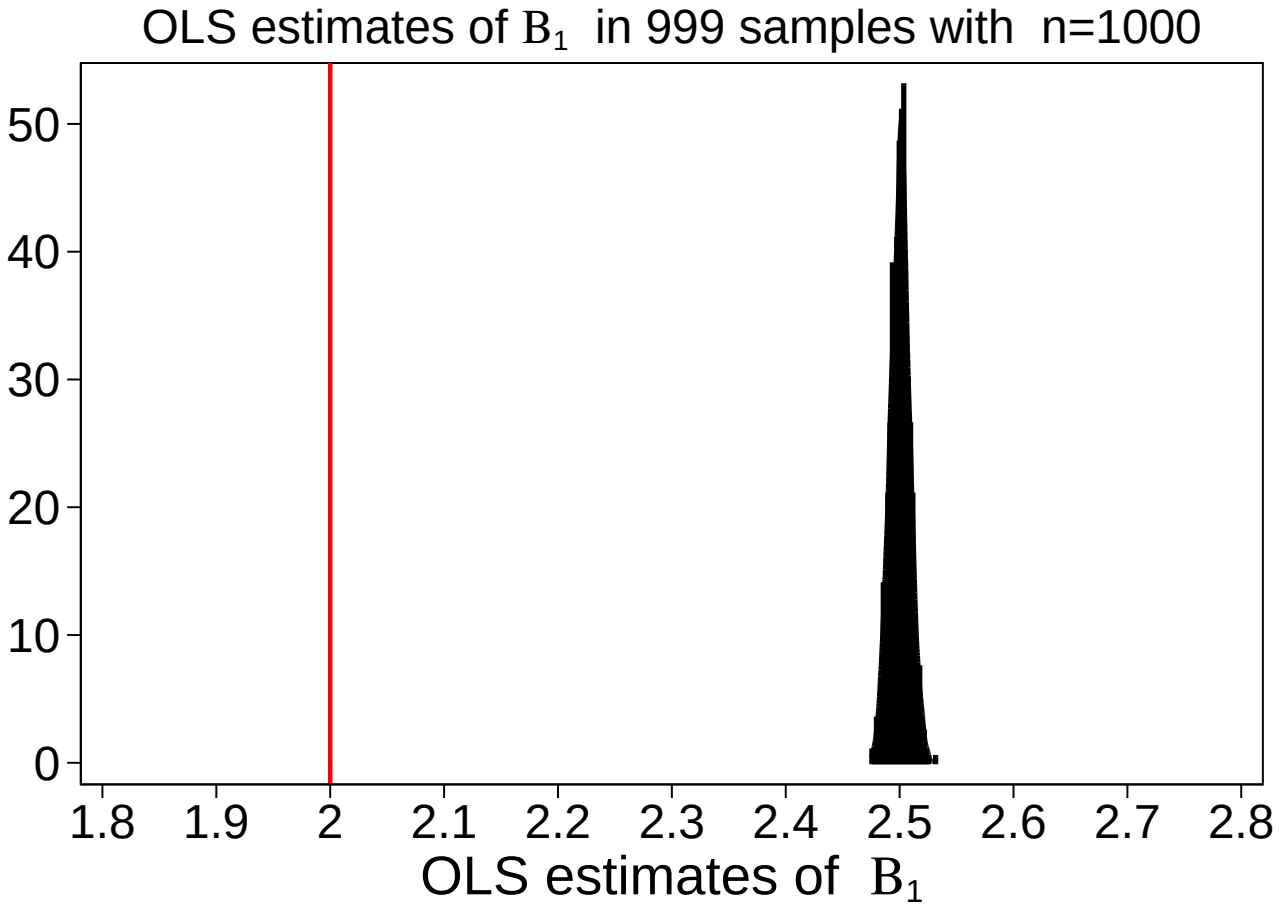
Estimated model : $Y_i = \beta_0 + \beta_1 X_i + v_i$



# Omitted variable bias: a simulation example
n=10000

True model : $Y_i = 1 + 2X_i + W_i + u_i \qquad E(u_i|X_i, W_i) = 0$

Estimated model : $Y_i = \beta_0 + \beta_1 X_i + v_i$

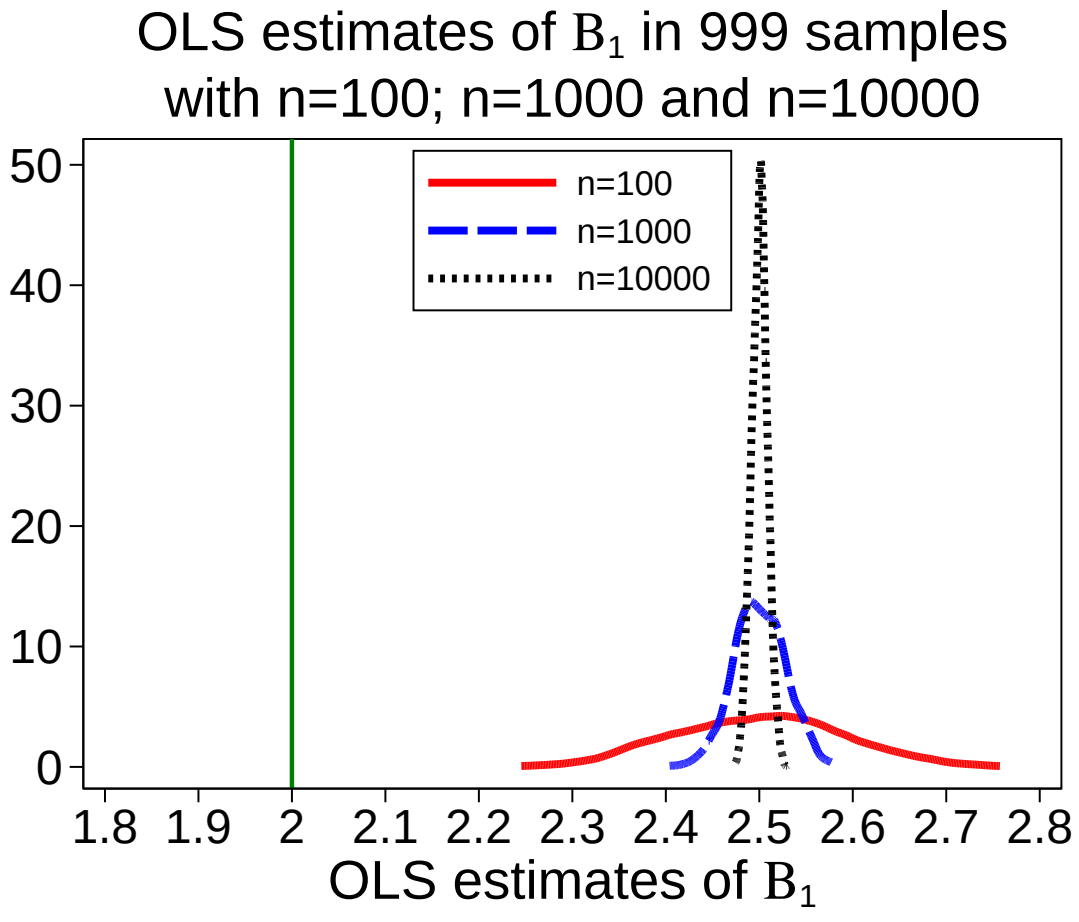


Omitted variable bias: a simulation example

n=100, n=1000, n=10000

True model : $Y_i = 1 + 2X_i + W_i + u_i \qquad E(u_i|X_i, W_i) = 0$

Estimated model : $Y_i = \beta_0 + \beta_1 X_i + v_i$



Including the omitted variable: a simulation example

- Natural solution to omitted variable bias is to include the variable and to estimate a **multiple regression model**.

$$\textit{True model}: \quad Y_i = 1 + 2X_i + W_i + u_i \qquad E(u_i|X_i, W_i) = 0$$

$$\textit{Estimated model}: \quad Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_i + v_i$$

```
. regress y x w
```

Source	SS	df	MS
Model	1531.63416	2	765.817078
Residual	87.0866036	97	.897800037
Total	1618.72076	99	16.3507147

Number of obs	100
F(2, 97)	852.99
Prob > F	0.0000
R-squared	0.9462
Adj R-squared	0.9451
Root MSE	.94752

y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
x	1.968079	.0888201	22.16	0.000	1.791795	2.144362
w	.9801195	.1214549	8.07	0.000	.7390651	1.221174
_cons	1.032288	.095095	10.86	0.000	.8435512	1.221026

Including the omitted variable: a simulation example

We can create 999 of these data sets with 100 observations and use OLS to estimate

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_i + v_i$$

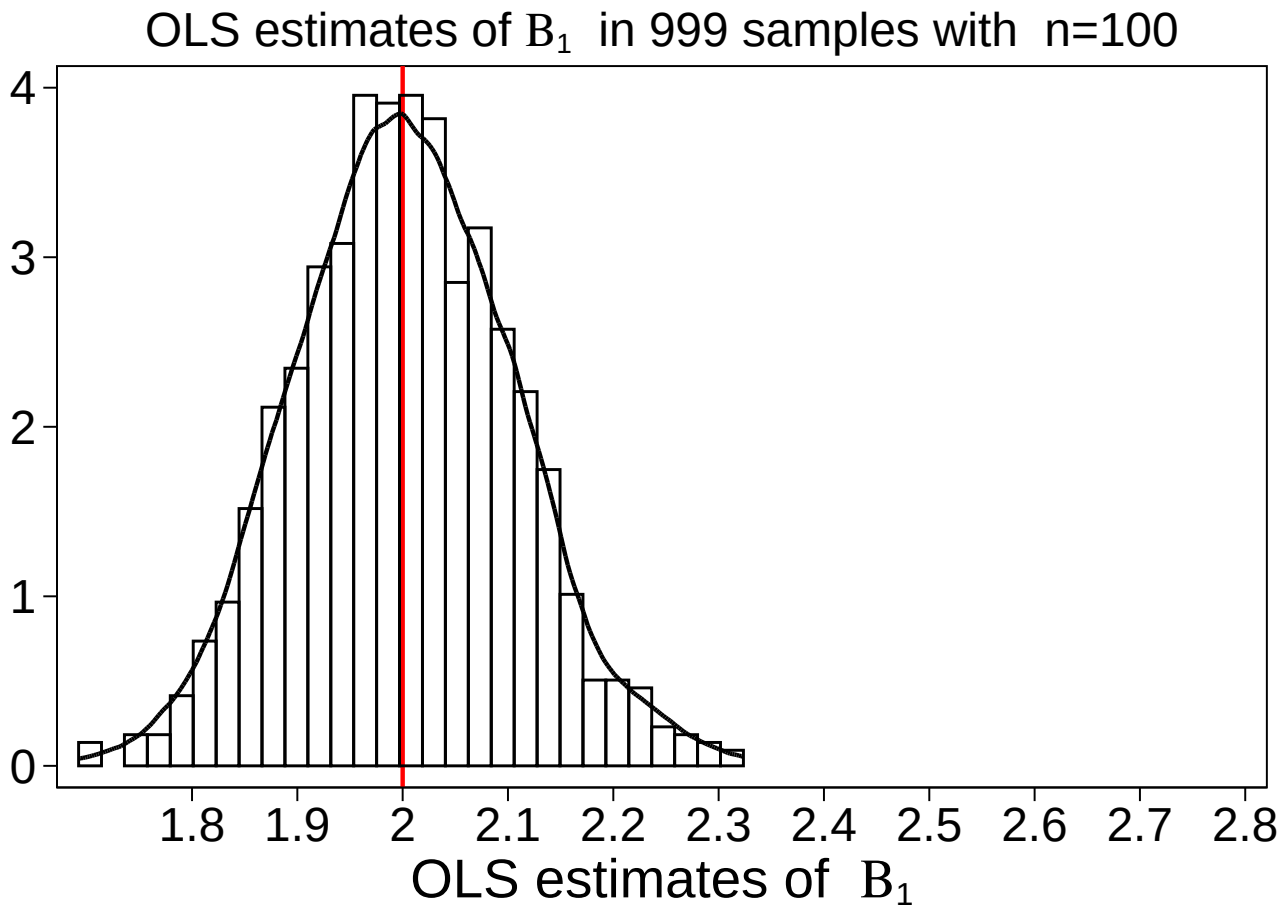
```
1 . program define ols, rclass
    1.  drop _all
    2.  set obs 100
    3.  gen w=rnormal()
    4.  gen x=w+rnormal()
    5.  gen y=1+2*x+w+rnormal()
    6.  regress y x w
    7. end

2 .
3 . simulate _b, reps(999)  nodots : ols

      command:  ols


4 . sum
```

Variable	Obs	Mean	Std. Dev.	Min	Max
_b_x	**999**	**1.997546**	**.1077047**	**1.640418**	**2.342614**
_b_w	**999**	**.9944234**	**.1485172**	**.5168402**	**1.455485**
_b_cons	**999**	**.9994964**	**.0988118**	**.7383428**	**1.301634**

# Including the omitted variable: a simulation example
n=100

True model : $Y_i = 1 + 2X_i + W_i + u_i$ $E(u_i|X_i, W_i) = 0$

Estimated model : $Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_i + v_i$

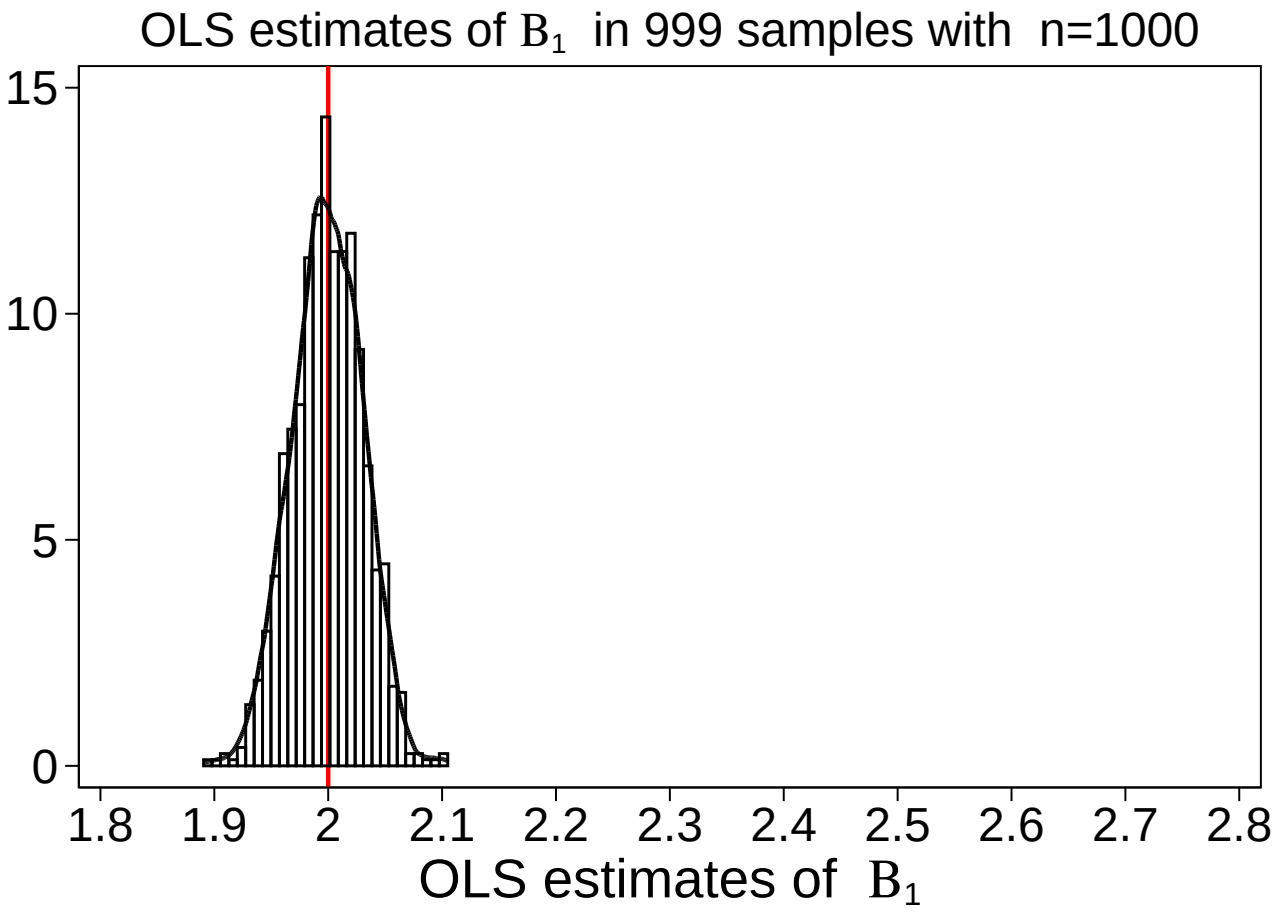


Including the omitted variable: a simulation example

n=1000

True model : $Y_i = 1 + 2X_i + W_i + u_i$ $E(u_i|X_i, W_i) = 0$

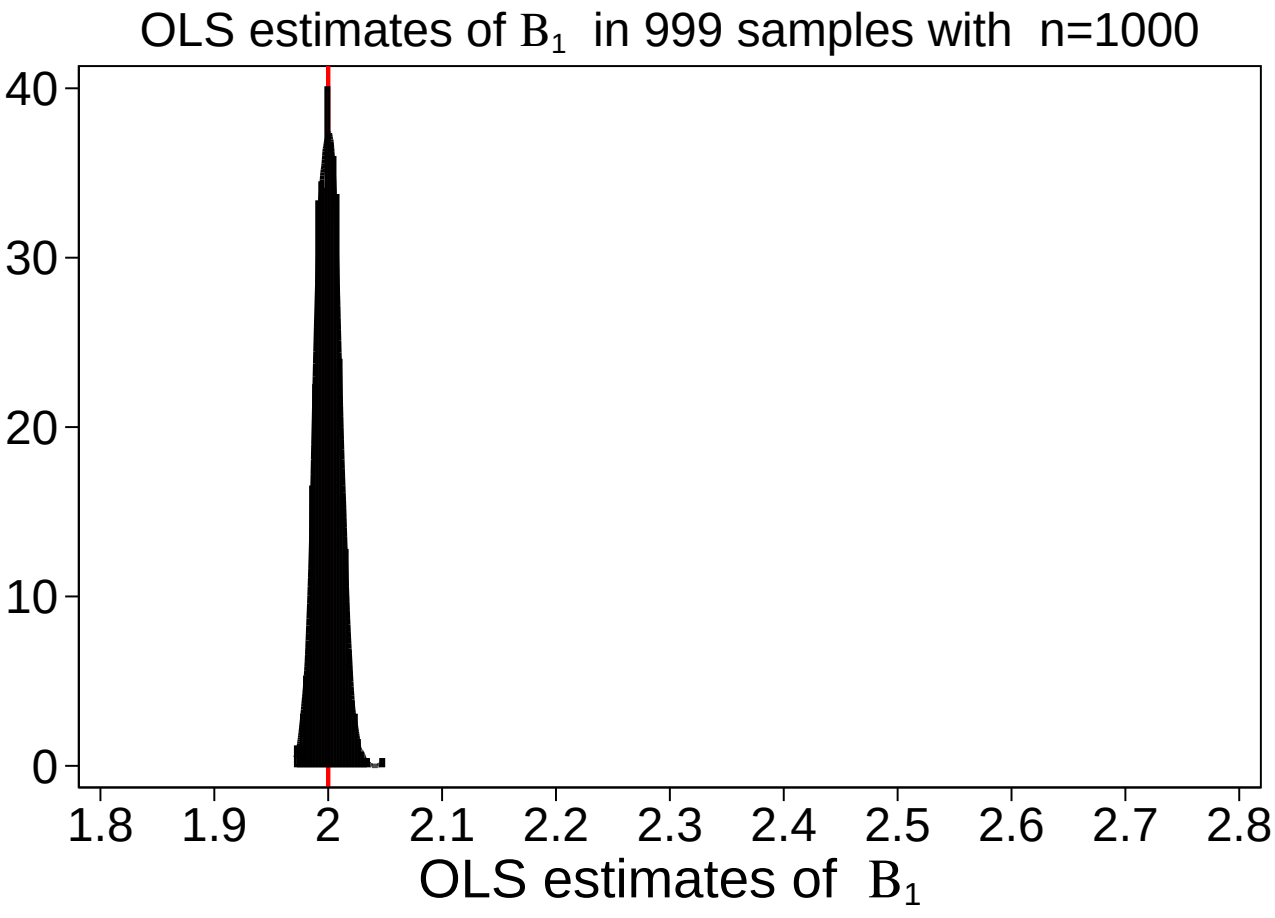
Estimated model : $Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_i + v_i$



# Including the omitted variable: a simulation example
n=10000

True model : $Y_i = 1 + 2X_i + W_i + u_i$ $E(u_i|X_i, W_i) = 0$

Estimated model : $Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_i + v_i$

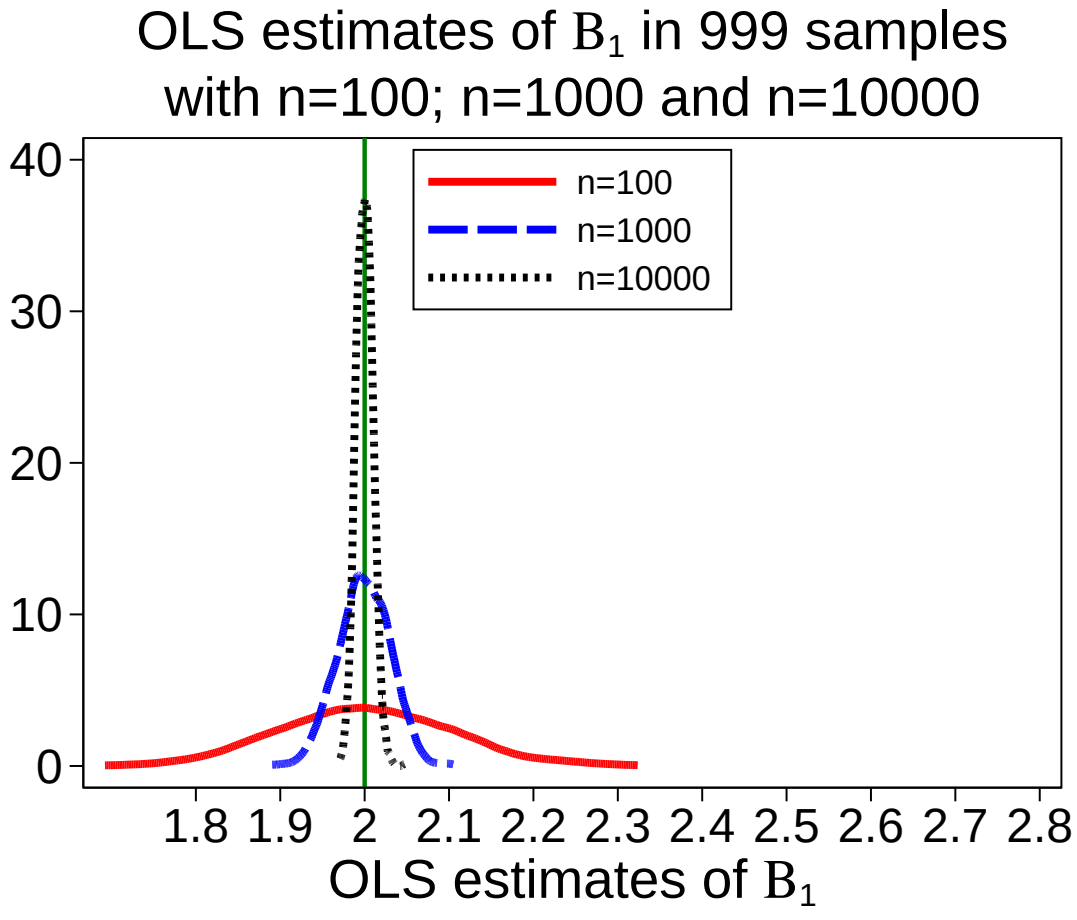


Including the omitted variable: a simulation example

n=100, n=1000, n=10000

True model : $Y_i = 1 + 2X_i + W_i + u_i$ $E(u_i|X_i, W_i) = 0$

Estimated model : $Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_i + v_i$



Multiple regression model with 2 regressors

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

Interpretation of β_1:

- Suppose we would increase X_1 to $X_1 + \triangle X_1$ while keeping X_2 constant.

$$\begin{aligned} E[Y|X_1, X_2] &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 \\ E[Y|(X_1 + \triangle X_1), X_2] &= \beta_0 + \beta_1 (X_1 + \triangle X_1) + \beta_2 X_2 \end{aligned}$$

define $\triangle Y$ as

$$\begin{aligned} \triangle Y &= E[Y|(X_1 + \triangle X_1), X_2] - E[Y|X_1, X_2] \\ &= \beta_1 \triangle X_1 \end{aligned}$$

this implies that β_1 is the expected change in Y due to unit change in X_1 *while keeping X_2 constant!*

Multiple regression model with 2 regressors

Example: The effect of class size on test scores

```
1 . regress test_score class_size, robust

Linear regression                               Number of obs =         420
                                                F(  1,   418) =       19.26
                                                Prob > F      =      0.0000
                                                R-squared     =      0.0512
                                                Root MSE      =      18.581
```

test_score	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
class_size	-2.279808	.5194892	-4.39	0.000	-3.300945	-1.258671
_cons	698.933	10.36436	67.44	0.000	678.5602	719.3057

```
2 . regress test_score class_size el_pct, robust

Linear regression                               Number of obs =         420
                                                F(  2,   417) =      223.82
                                                Prob > F      =      0.0000
                                                R-squared     =      0.4264
                                                Root MSE      =      14.464
```

test_score	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
class_size	-1.101296	.4328472	-2.54	0.011	-1.95213	-.2504616
el_pct	-.6497768	.0310318	-20.94	0.000	-.710775	-.5887786
_cons	686.0322	8.728224	78.60	0.000	668.8754	703.189

Multiple regression model with 2 regressors

Example: The effect of class size on test scores

$$\widehat{test\ score}_i = 686.03 - 1.10 \cdot class\ size_i - 0.65 \cdot el\ pct_i$$

- The expected effect on test scores of increasing class size by 1, while keeping the percentage of English learners constant, equals -1.1 points
- This is about half the size of coefficient estimate when $el\ pct_i$ is omitted from the regression
- Estimated effect of class size in the simple regression model suffers from omitted variable bias
- Omitted variable bias formula already predicted a negative bias.

$$\widehat{\beta}_1 \xrightarrow{p} \beta_1 + \beta_2 \frac{Cov\left(class\ size_i, el\ pct_i\right)}{Var\left(class\ size_i\right)}$$

Multiple regression model with k regressors

General notation for multiple regression model with k regressors:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} + u_i$$

where

- Y_i is the i^{th} observation on the dependent variable
- $X_{1i}, ..., X_{ki}$ are i^{th} observations on the k independent variables or regressors
- β_0 is the intercept of the population regression line (expected value of Y when $X_{1i}, ..., X_{ki} = 0$)
- β_1 is the slope coefficient on X_1; the expected change in Y due to a unit increase in X_{1i} while holding $X_{2i}, ..., X_{ki}$ constant.
- u_i is the error term (all other factors, besides $X_{1i}, ..., X_{ki}$, determining Y_i)

Multiple regression model with k regressors

The OLS estimators $\widehat{\beta}_0, \widehat{\beta}_1, ..., \widehat{\beta}_k$ are obtained by minimizing the sum of squared prediction mistakes:

$$\sum_{i=1}^{n} \left(Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_{1i} - ... - \widehat{\beta}_k X_{ki} \right)^2$$

Similar to the linear regression model with 1 regressor this implies

- taking derivatives w.r.t $\widehat{\beta}_0, \widehat{\beta}_1, ..., \widehat{\beta}_k$
- setting these to zero and solving for $\widehat{\beta}_0, \widehat{\beta}_1, ..., \widehat{\beta}_k$

Formulas for OLS estimators in multiple regression model are best expressed using matrix algebra.

Multiple regression model with k regressors

Least squares assumption for multiple regression model:

Assumption 1: The conditional distribution of u_i given $X_{1i}, ..., X_{ki}$ has mean zero, that is

$$E(u_i|X_{1i}, ..., X_{ki}) = 0$$

Assumption 2: $(Y_i, X_{1i}, ..., X_{ki})$ for $i = 1, ..., n$ are independently and identically distributed $(i.i.d)$

Assumption 3: Large outliers are unlikely

$$0 < E\left(X_{1i}^4\right) < \infty, ..., 0 < E\left(X_{ki}^4\right) < \infty \quad \& \quad 0 < E\left(Y_i^4\right) < \infty$$

Assumption 4: No perfect multicollinearity

Perfect multicollinearity

Perfect multicollinearity arises when one of the regressors is a perfect linear combination of the other regressors

- The other regressors include the regressor on the constant term $X_{0i} = 1 \quad for\ i = 1, ..., n$

$$Y_i = \beta_0 X_{0i} + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} + u_i$$

- If the regressors exhibit perfect multicollinearity, the OLS estimators cannot be computed
- Perfect multicollinearity produces division by zero in the OLS formulas
- Intuitively: you estimate effect of a change in one regressor on Y while holding another regressor, which is a perfect linear combination of the first regressor, constant: This does not make sense!

Perfect multicollinearity

What happens when we include both the percentage of English learners and the share of English learners?

```
. gen el_share=el_pct/100

.
. regress test_score class_size el_pct el_share, robust
note: el_share omitted because of collinearity

Linear regression                               Number of obs =         420
                                                F(  2,   417) =      223.82
                                                Prob > F      =      0.0000
                                                R-squared     =      0.4264
                                                Root MSE      =      14.464
```

test_score	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
class_size	-1.101296	.4328472	-2.54	0.011	-1.95213	-.2504616
el_pct	-.6497768	.0310318	-20.94	0.000	-.710775	-.5887786
el_share	0	(omitted)				
_cons	686.0322	8.728224	78.60	0.000	668.8754	703.189

el pct$_i$ and *el share*$_i$ are perfectly multicollinear

Perfect multicollinearity: Dummy variable trap

What happens when we include both a dummy $SmallClass_i$ (=1 if class size < 20) and a dummy $BigClass_i$ (=1 if class size ≥ 20) and the constant term?

```
. regress test_score SmallClass BigClass, robust
note: BigClass omitted because of collinearity

Linear regression                               Number of obs     =        420
                                                F(1, 418)         =      16.34
                                                Prob > F          =     0.0001
                                                R-squared         =     0.0369
                                                Root MSE          =     18.721
```

test_score	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
SmallClass	**7.37241**	**1.823578**	**4.04**	**0.000**	**3.787884**	**10.95694**
BigClass	**0**	(omitted)				
_cons	**649.9788**	**1.322892**	**491.33**	**0.000**	**647.3785**	**652.5792**

$$BigClass_i = 1 - SmallClass_i = X_{0i} - SmallClass_i$$

(Imperfect) multicollinearity

(Imperfect) multicollinearity means that two or more regressors are highly correlated, but one regressor is NOT a perfect linear function of one or more of the other regressors

- (imperfect) multicollinearity is not a violation of the least squares assumptions
- It does not impose theoretical problem for the calculation of OLS estimators
- If two regressors are highly correlated the the coefficient on at least one of the regressors is imprecisely estimated (high variance)
- With two regressors and homoskedastic errors we have that

$$Var\left(\widehat{\beta}_1\right) = \frac{1}{n}\left(\frac{1}{1-\rho^2_{X_1X_2}}\right)\frac{\sigma^2_u}{\sigma^2_{X_1}}$$

Properties OLS estimators in multiple regression model

If the four least squares assumptions in the multiple regression model hold:

- The OLS estimators $\widehat{\beta}_0, \widehat{\beta}_1, ..., \widehat{\beta}_k$ are unbiased

$$E\left(\widehat{\beta}_j\right) = \beta_j \quad for \quad j = 0, ..., k$$

- The OLS estimators $\widehat{\beta}_0, \widehat{\beta}_1, ..., \widehat{\beta}_k$ are consistent

$$\widehat{\beta}_j \xrightarrow{p} \beta_j \quad for \quad j = 0, ..., k$$

- The OLS estimators $\widehat{\beta}_0, \widehat{\beta}_1, ..., \widehat{\beta}_k$ are normally distributed in large samples

$$\widehat{\beta}_j \sim N\left(\beta_j, \sigma^2_{\widehat{\beta}_j}\right) \quad for \quad j = 0, ..., k$$

Multiple regression model: class size example

```
. regress test_score class_size, robust

Linear regression                               Number of obs     =        420
                                                F(1, 418)         =      19.26
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0512
                                                Root MSE          =     18.581
```

test_score	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
class_size	-2.279808	.5194892	-4.39	0.000	-3.300945	-1.258671
_cons	698.933	10.36436	67.44	0.000	678.5602	719.3057

- Is the average causal effect of class size on test scores equal to -2.27?
- Is there omitted variable bias?

Multiple regression model: class size example

If we add the percentage of English learners as regressor in the regression model we get:

```
. regress test_score class_size el_pct, robust

Linear regression                               Number of obs     =        420
                                                F(2, 417)         =     223.82
                                                Prob > F          =     0.0000
                                                R-squared         =     0.4264
                                                Root MSE          =     14.464
```

test_score	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
class_size	-1.101296	.4328472	-2.54	0.011	-1.95213	-.2504616
el_pct	-.6497768	.0310318	-20.94	0.000	-.710775	-.5887786
_cons	686.0322	8.728224	78.60	0.000	668.8754	703.189

- Is the average causal effect of class size on test scores equal to -1.10?
- Is there omitted variable bias?

Multiple regression model: class size example

If we add the percentage of students eligible for a free lunch as regressor in the regression model we get:

```
. regress test_score class_size el_pct meal_pct, robust

Linear regression                               Number of obs     =        420
                                                F(3, 416)         =     453.48
                                                Prob > F          =     0.0000
                                                R-squared         =     0.7745
                                                Root MSE          =     9.0801
```

test_score	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
class_size	-.9983092	.2700799	-3.70	0.000	-1.529201	-.4674178
el_pct	-.1215733	.0328317	-3.70	0.000	-.18611	-.0570366
meal_pct	-.5473456	.0241072	-22.70	0.000	-.5947328	-.4999583
_cons	700.15	5.56845	125.74	0.000	689.2042	711.0958

- Is the average causal effect of class size on test scores equal to -0.99?
- Is there omitted variable bias?

Multiple regression model: class size example

If we add district average income as regressor in the regression model we get:

```
. regress test_score class_size el_pct meal_pct avginc, robust

Linear regression                               Number of obs     =        420
                                                F(4, 415)         =     467.42
                                                Prob > F          =     0.0000
                                                R-squared         =     0.8053
                                                Root MSE          =     8.4477
```

test_score	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
class_size	-.5603892	.2550641	-2.20	0.029	-1.061768	-.0590105
el_pct	-.1943282	.0332445	-5.85	0.000	-.2596768	-.1289795
meal_pct	-.3963661	.0302302	-13.11	0.000	-.4557895	-.3369427
avginc	.674984	.0837161	8.06	0.000	.5104236	.8395444
_cons	675.6082	6.201865	108.94	0.000	663.4172	687.7992

- Is the average causal effect of class size on test scores equal to -0.56?
- Is there omitted variable bias?

Multiple regression model: class size example

Dependent variable: district average test scores				
	1	2	3	4
Class size	-2.280***	-1.101**	-0.998***	-0.560**
	(0.519)	(0.433)	(0.270)	(0.255)
Percentage of English learners		-0.650***	-0.122***	-0.194***
		(0.031)	(0.033)	(0.033)
Percentage with free lunch			-0.547***	-0.396***
			(0.024)	(0.030)
Average district income				0.675***
				(0.084)
N	420	420	420	420