

ECON4150 - Introductory Econometrics

Lecture 10: Some Repetition & Introduction to Nonlinear Regression Functions

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

Stock and Watson Chapter 2-8

Lecture outline

Simple linear regression model

- OLS assumptions
- A simulation example
- Unbiasedness
- Consistency

When the first OLS assumption is violated

- Omitted variable bias
 - A simulation example
 - Unbiasedness
 - Consistency
- Misspecification of the functional form
 - A simulation example
 - Unbiasedness
 - Consistency

The Simple Linear Regression Model

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

- OLS minimizes sum of squared prediction mistakes:

$$\sum_{i=1}^n \hat{u}_i^2 = \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$

- Step 1:

$$\frac{\partial}{\partial \hat{\beta}_0} \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 = 0$$

- Step 2:

$$\frac{\partial}{\partial \hat{\beta}_1} \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 = 0$$

Step 1: OLS estimator of β_0

$$\begin{aligned}\frac{\partial}{\partial \hat{\beta}_0} \sum_{i=1}^n u_i^2 &= -2 \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0 \\ &= \frac{1}{n} \left(\sum_{i=1}^n Y_i - \sum_{i=1}^n \hat{\beta}_0 - \sum_{i=1}^n \hat{\beta}_1 X_i \right) = 0 \\ &= \bar{Y} - \hat{\beta}_0 - \hat{\beta}_1 \bar{X} = 0\end{aligned}$$

- This gives

$$\widehat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

Step 2: OLS estimator of β_1

$$\begin{aligned}
 \frac{\partial}{\partial \hat{\beta}_1} \sum_{i=1}^n u_i^2 &= -2 \cdot \sum_{i=1}^n -X_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0 \\
 &= \sum_{i=1}^n X_i (Y_i - (\bar{Y} - \hat{\beta}_1 \bar{X}) - \hat{\beta}_1 X_i) = 0 \\
 &= \sum_{i=1}^n X_i (Y_i - \bar{Y}) - \hat{\beta}_1 \sum_{i=1}^n X_i (X_i - \bar{X}) = 0 \\
 &= \sum_{i=1}^n (X_i - \bar{X}) (Y_i - \bar{Y}) - \hat{\beta}_1 \sum_{i=1}^n (X_i - \bar{X}) (X_i - \bar{X}) = 0
 \end{aligned}$$

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})} = \frac{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})} = \frac{s_{xy}}{s_x^2}$$

$$\begin{aligned}
 \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) &= \sum_{i=1}^n X_i Y_i - \sum_{i=1}^n X_i \bar{Y} - \sum_{i=1}^n \bar{X} Y_i + \sum_{i=1}^n \bar{X} \bar{Y} \\
 &= \sum_{i=1}^n X_i Y_i - \sum_{i=1}^n X_i \bar{Y} - n \bar{X} \bar{Y} + n \bar{X} \bar{Y} \\
 &= \sum_{i=1}^n X_i (Y_i - \bar{Y})
 \end{aligned}$$

OLS assumptions & properties

OLS assumptions:

- ① $E[u_i|X_i] = 0$
- ② (X_i, Y_i) for $i = 1, \dots, n$ are i.i.d.
- ③ Large outliers are unlikely $E[X_i^4] < \infty$ & $E[Y_i^4] < \infty$

OLS properties:

- ① Unbiasedness:

$$E[\hat{\beta}_0] = \beta_0 \quad \& \quad E[\hat{\beta}_1] = \beta_1$$

- ② Consistency:

$$plim \widehat{\beta}_1 = \beta_1 \quad \& \quad plim \widehat{\beta}_0 = \beta_0$$

A simulation example

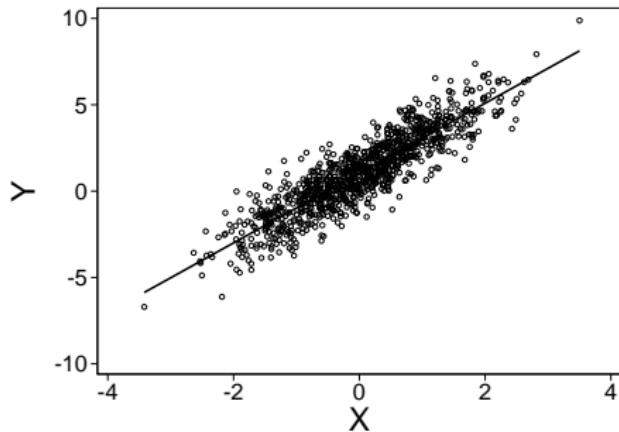
- Lets create a data set with 1000 observations
- $X_i \sim N(0, 1)$
- $u_i \sim N(0, 1)$
- $Y_i = 1 + 2X_i + u_i$

```
set obs 1000
gen x=invnorm(uniform())
gen y=1+2*x+invnorm(uniform())
```

. sum y x

Variable	Obs	Mean	Std. Dev.	Min	Max
y	1000	1.059341	2.276395	-5.694431	9.286
x	1000	.0410602	1.01173	-3.035906	3.878556

A simulation example



```
. regress y x
```

Source	SS	df	MS	Number of obs =	1000
Model	3939.09384	1	3939.09384	F(1, 998) =	4295.91
Residual	915.107165	998	.916941047	Prob > F =	0.0000
Total	4854.201	999	4.85906006	R-squared =	0.8115
				Adj R-squared =	0.8113
				Root MSE =	.95757

	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
x	1.990639	.0303714	65.54	0.000	1.93104 2.050238
_cons	.9850093	.0302894	32.52	0.000	.925571 1.044448

Unbiasedness

True model : $Y_i = 1 + 2X_i + u_i$

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

substitute for Y_i, \bar{Y}

$$= E\left[\frac{\sum_{i=1}^n (X_i - \bar{X})(1 + 2X_i + u_i - (1 + 2\bar{X} + \bar{u}))}{\sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})}\right]$$

rewrite

$$= E\left[\frac{\sum_{i=1}^n (X_i - \bar{X})(2(X_i - \bar{X}) + (u_i - \bar{u}))}{\sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})}\right]$$

$$\bar{u} = 0$$

$$= 2 + E\left[\frac{\sum_{i=1}^n (X_i - \bar{X})u_i}{\sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})}\right]$$

Law of iterated exp.

$$= 2 + E\left[\frac{\sum_{i=1}^n (X_i - \bar{X})E[u_i | X_i]}{\sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})}\right]$$

so

$$E[\hat{\beta}_1] = 2 \quad \text{if} \quad E[u_i | X_i] = 0$$

Consistency

True model : $Y_i = 1 + 2X_i + u_i$, *Estimated model* : $Y_i = \beta_0 + \beta_1 X_i + u_i$

$$\begin{aligned}
 \text{Plim } \hat{\beta}_1 &= \frac{\text{Plim } \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\text{Plim } \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})} \\
 &\quad \text{law of large numbers} \\
 &\quad \text{OLS assumptions 2 and 3} \\
 &= \frac{\text{Cov}(X_i, Y_i)}{\text{Var}(X_i)} \\
 &= \frac{\text{Cov}(X_i, 1 + 2X_i + u_i)}{\text{Var}(X_i)} \\
 &\quad \text{substitute } Y_i \\
 &= \frac{2\text{Var}(X_i) + \text{Cov}(X_i, u_i)}{\text{Var}(X_i)} \\
 &= 2 + \frac{E[(X_i - \mu_x)(u_i - \mu_u)]}{\text{Var}(X_i)} \\
 &= 2 + \frac{E[(X_i - \mu_x)E[u_i | X_i]]}{\text{Var}(X_i)} \\
 &\quad \mu_u = 0
 \end{aligned}$$

so

$$\text{Plim } \hat{\beta}_1 = 2 \quad \text{if} \quad E[u_i | X_i] = 0$$

Unbiasedness vs Consistency

- Unbiasedness & consistency both rely on $E[u_i|X_i] = 0$
- since in the simulation example u_i does not depend on X_i & $u_i \sim N(0, 1)$
 $\Rightarrow E[u_i|X_i] = E[u_i] = 0$
- Unbiasedness implies that $E[\hat{\beta}_1] = \beta_1$ for a given sample size n
- Consistency implies that the sampling distribution becomes more and more tightly distributed around β_1 if the sample size n becomes larger and larger.

A simulation example n=100

```

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

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

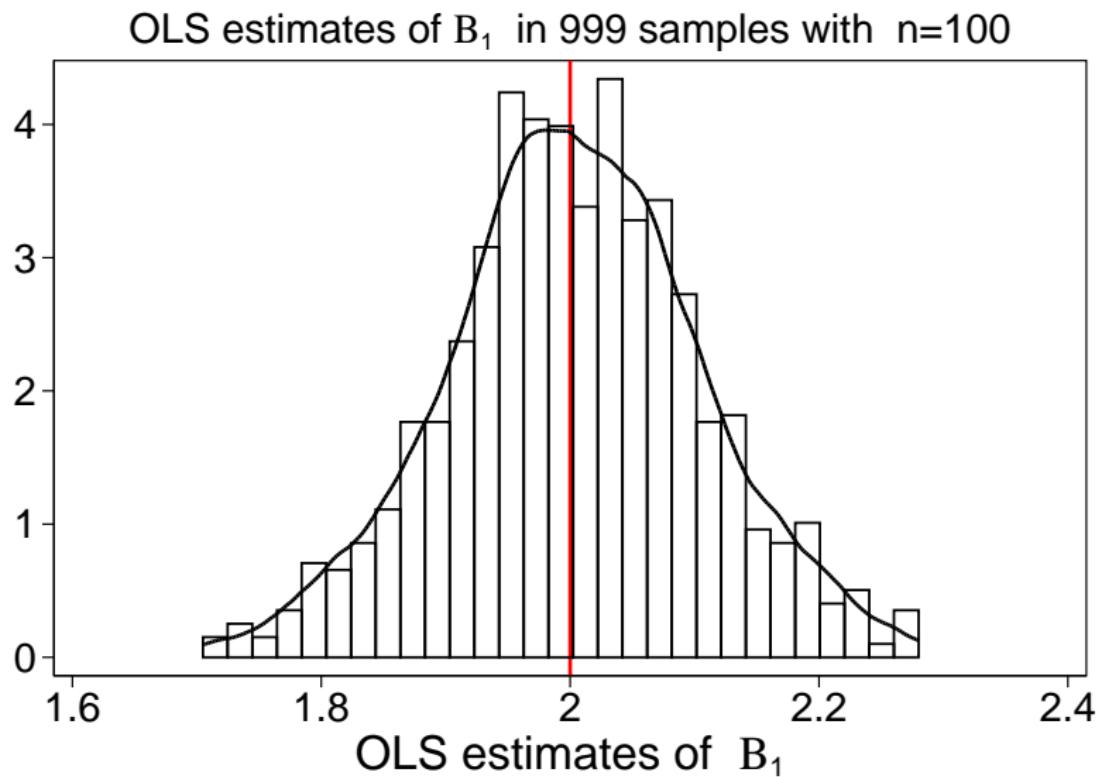
      command:  ols

4 . sum

```

Variable	Obs	Mean	Std. Dev.	Min	Max
_b_x	999	1.997521	.1018595	1.67569	2.308795
_b_cons	999	1.003246	.1019056	.6844429	1.285363

A simulation example n=100



A simulation example n=1000

```
1 . program define ols, rclass
 1. drop _all
 2. set obs 1000
 3. gen x=invnorm(uniform())
 4. gen y=1+2*x+invnorm(uniform())
 5. regress y x
 6. end

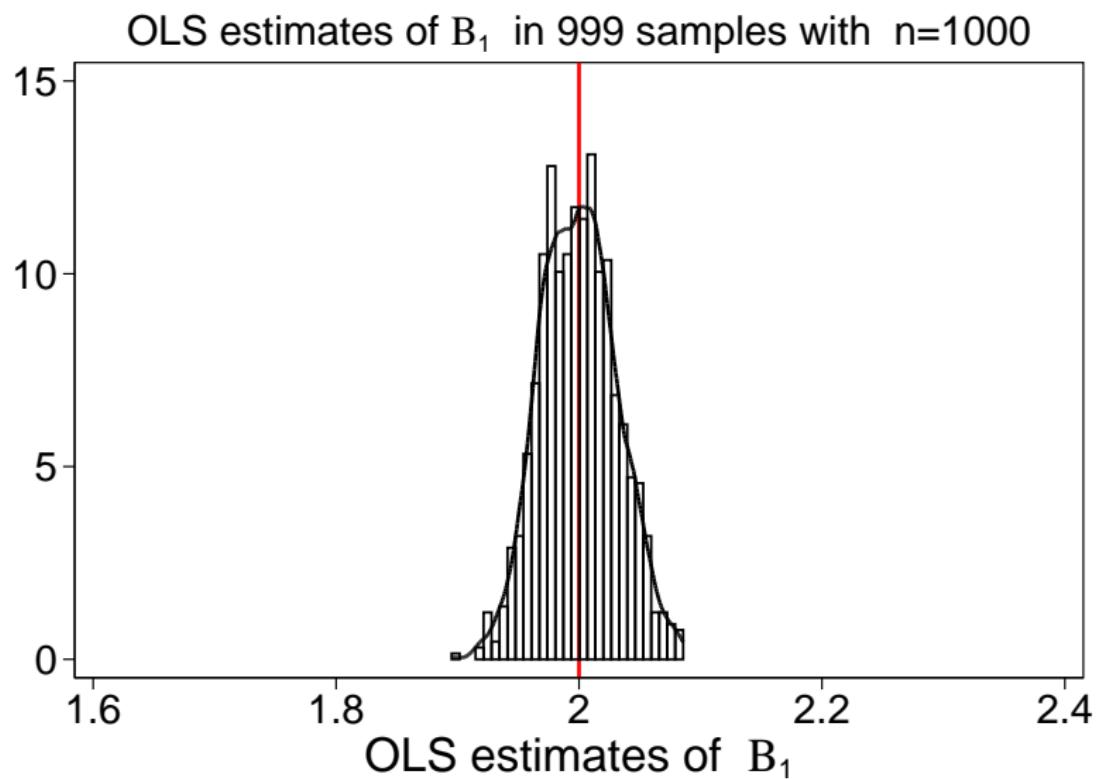
2 .
3 . simulate _b, reps(999)  nodots : ols

      command:  ols
```

```
4 . sum
```

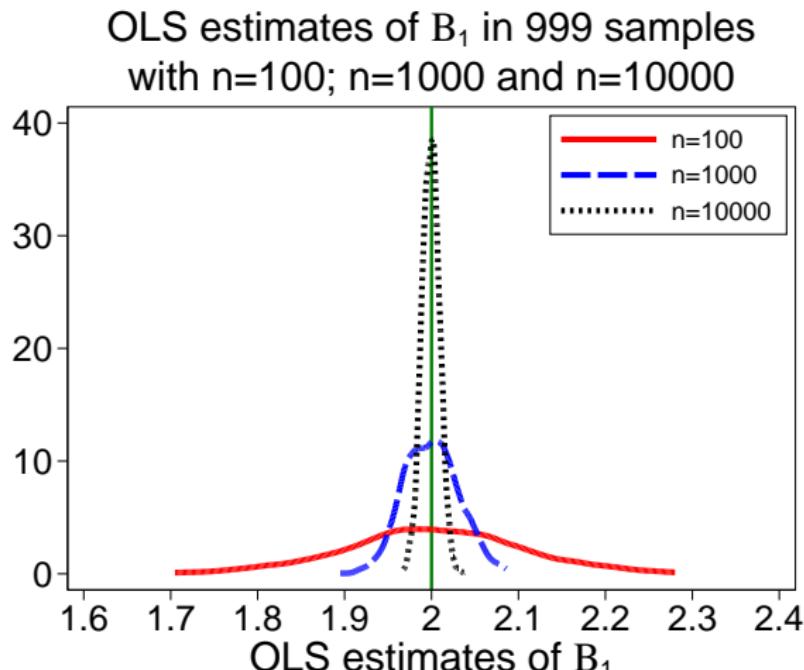
Variable	Obs	Mean	Std. Dev.	Min	Max
_b_x	999	2.000035	.030417	1.908725	2.112585
_b_cons	999	1.000791	.0311526	.8970624	1.088724

A simulation example n=1000



Consistency of the OLS estimator of $\hat{\beta}_1$

True model : $Y_i = 1 + 2X_i + u_i$, Estimated model : $Y_i = \beta_0 + \beta_1 X_i + u_i$



Simulation example with omitted variable bias

- Lets again create a data set with 1000 observations
- $W_i \sim N(0, 1)$
- $X_i = W_i + \varepsilon_i$ with $\varepsilon_i \sim N(0, 1)$
- $u_i \sim N(0, 1)$
- $Y_i = 1 + 2X_i + W_i + u_i$

```
set obs 1000
gen w=invnorm(uniform())
gen x=w+invnorm(uniform())
gen y=1+2*x+w+invnorm(uniform())
```

. sum y x w

Variable	Obs	Mean	Std. Dev.	Min	Max
y	1000	1.222463	3.814473	-11.57748	14.20833
x	1000	.0623473	1.441174	-5.640203	4.454752
w	1000	.0552381	1.004599	-3.461684	3.220981

Simulation example with omitted variable bias

1 . regress y x

Source	SS	df	MS	Number of obs = 1000
Model	13060.2191	1	13060.2191	F(1, 998) = 8834.07
Residual	1475.43573	998	1.47839252	Prob > F = 0.0000
Total	14535.6549	999	14.5502051	R-squared = 0.8985 Adj R-squared = 0.8984 Root MSE = 1.2159

y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
x	2.508858	.0266929	93.99	0.000	2.456477 2.561238
_cons	1.066042	.0384859	27.70	0.000	.9905196 1.141565

2 . regress y x w

Source	SS	df	MS	Number of obs = 1000
Model	13533.3282	2	6766.66411	F(2, 997) = 6730.70
Residual	1002.32664	997	1.00534267	Prob > F = 0.0000
Total	14535.6549	999	14.5502051	R-squared = 0.9310 Adj R-squared = 0.9309 Root MSE = 1.0027

y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
x	1.985033	.0326741	60.75	0.000	1.920915 2.049151
w	1.016837	.0468736	21.69	0.000	.9248553 1.10882
_cons	1.042533	.0317553	32.83	0.000	.980218 1.104848

Simulation example with omitted variable bias

```

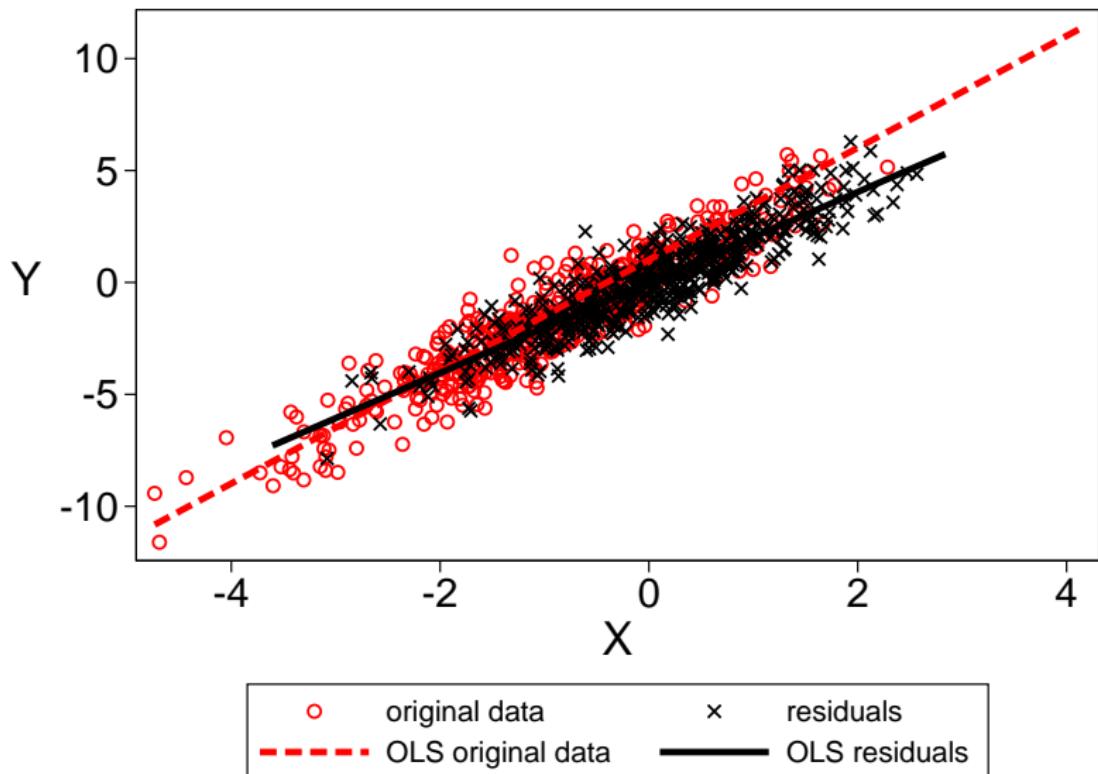
1 . quietly regress y w
2 . predict res_y, residuals
3 . quietly regress x w
4 . predict res_x, residuals
5 . regress res_y res_x

```

Source	SS	df	MS	Number of obs = 1000
Model	4222.82591	1	4222.82591	F(1, 998) = 3945.04
Residual	1068.27433	998	1.07041516	Prob > F = 0.0000
Total	5291.10025	999	5.29639664	R-squared = 0.7981
				Adj R-squared = 0.7979
				Root MSE = 1.0346

res_y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
res_x	1.993204	.0317341	62.81	0.000	1.930931 2.055477
_cons	-2.64e-16	.0327172	-0.00	1.000	-.0642024 .0642024

Simulation example with omitted variable bias



Unbiasedness???

True model : $Y_i = 1 + 2X_i + W_i + u_i$, Estimated model : $Y_i = \beta_0 + \beta_1 X_i + v_i$

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

substitute for Y_i

$$= E\left[\frac{\sum_{i=1}^n (X_i - \bar{X})(1 + 2X_i + W_i + u_i)}{\sum_{i=1}^n (X_i - \bar{X})X_i}\right]$$

rewrite

$$= 2 + E\left[\frac{\sum_{i=1}^n (X_i - \bar{X})W_i + \sum_{i=1}^n (X_i - \bar{X})u_i}{\sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})}\right]$$

Law of it. exp.

$$= 2 + E\left[\frac{\sum_{i=1}^n (X_i - \bar{X})E[W_i|X_i] + \sum_{i=1}^n (X_i - \bar{X})E[u_i|X_i]}{\sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})}\right]$$

so

$$E[\hat{\beta}_1] = 2 \quad \text{if} \quad E[u_i|X_i] = 0 \quad \& \quad E[W_i|X_i] = 0$$

We know $E[W_i|X_i] \neq 0$ so $\hat{\beta}_1$ is biased!

Consistency???

True model : $Y_i = 1 + 2X_i + W_i + u_i$, Estimated model : $Y_i = \beta_0 + \beta_1 X_i + v_i$

$$\begin{aligned}
 P\lim \hat{\beta}_1 &= \frac{P\lim \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{P\lim \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})} = \frac{Cov(X_i, Y_i)}{Var(X_i)} \\
 &= \frac{Cov(X_i, 1 + 2X_i + W_i + u_i)}{Var(X_i)} \quad \text{substitute } Y_i \\
 &= \frac{2Var(X_i) + Cov(X_i, W_i) + Cov(X_i, u_i)}{Var(X_i)} \\
 &= 2 + \frac{E[(X_i - \mu_x)(W_i - \mu_w)]}{Var(X_i)} + \frac{E[(X_i - \mu_x)(u_i - \mu_u)]}{Var(X_i)} \quad \mu_w = \mu_u = 0 \\
 &= 2 + \frac{E[(X_i - \mu_x)E[W_i | X_i]]}{Var(X_i)} + \frac{E[(X_i - \mu_x)E[u_i | X_i]]}{Var(X_i)}
 \end{aligned}$$

so

$$P\lim \hat{\beta}_1 = 2 \quad \text{if} \quad E[u_i | X_i] = 0 \quad \& \quad E[W_i | X_i] = 0$$

We know $E[W_i | X_i] \neq 0$ so $\hat{\beta}_1$ is inconsistent!

A simulation example with omitted variable bias n=1000

```

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

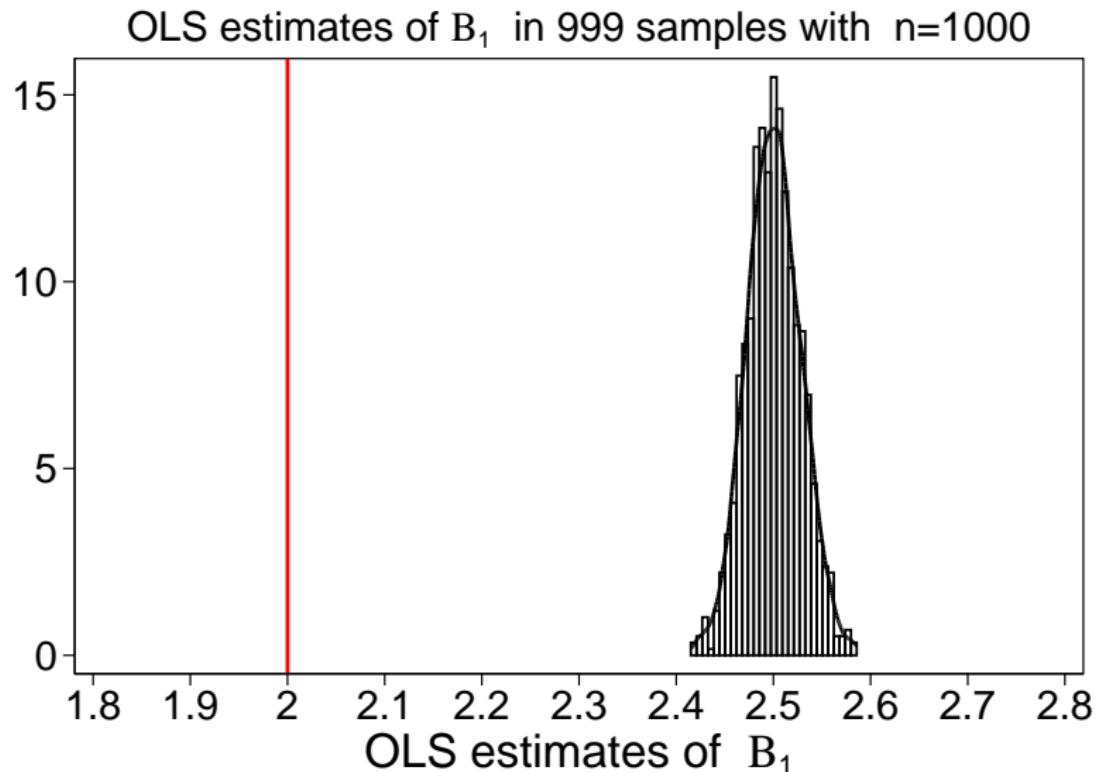
2 .
3 . simulate _b, reps(999)  nodots : ols
      command:  ols

```

4 . sum

Variable	Obs	Mean	Std. Dev.	Min	Max
_b_x	999	2.497618	.088588	2.232391	2.811449
_b_cons	999	1.000449	.1229033	.5489791	1.409755

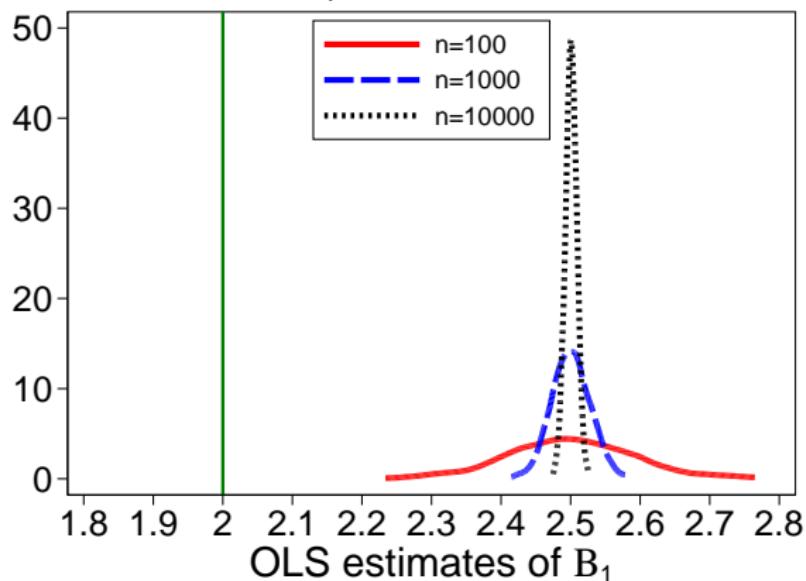
A simulation example with omitted variable bias n=1000



Consistency of the OLS estimator of $\hat{\beta}_1$

True model : $Y_i = 1 + 2X_i + W_i + u_i$, Estimated model : $Y_i = \beta_0 + \beta_1 X_i + v_i$

OLS estimates of B_1 in 999 samples
with $n=100$; $n=1000$ and $n=10000$



Simulation example with misspecification functional form

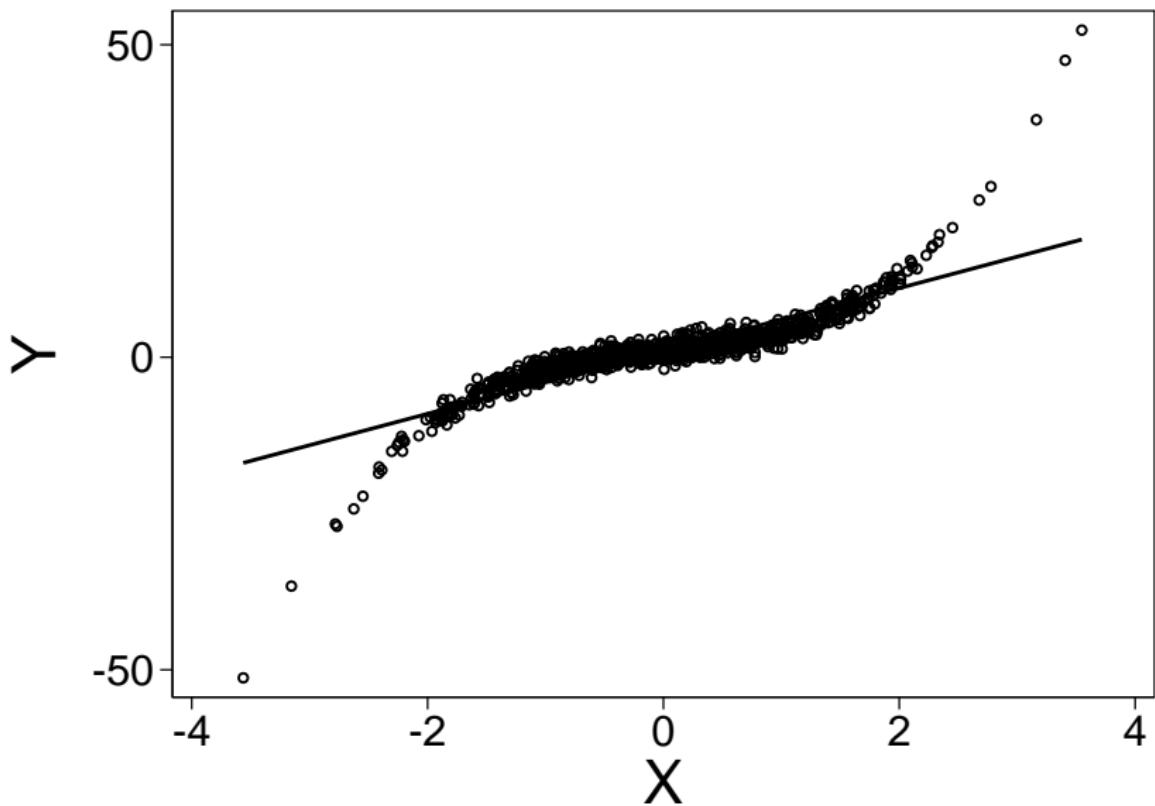
- Lets again create a data set with 1000 observations
- $X_i \sim N(0, 1)$
- $u_i \sim N(0, 1)$
- $Y_i = 1 + 2X_i + X_i^3 + u_i$

```
set obs 1000
gen x = invnorm(uniform())
gen x3 = x^3
gen y=1+2*x+x3+invnorm(uniform())
```

. sum y x x3

Variable	Obs	Mean	Std. Dev.	Min	Max
Y	1000	.9990983	5.388074	-36.55887	37.46202
x	1000	-.0078262	.9799795	-3.185017	3.117949
x3	1000	.0582388	3.642876	-32.30986	30.31146

Simulation example with misspecification functional form



Simulation example with misspecification functional form

1 . regress y x

Source	SS	df	MS	Number of obs = 1000		
Model	22747.8326	1	22747.8326	F(1, 998) =	3629.77	
Residual	6254.47978	998	6.26701381	Prob > F =	0.0000	
Total	29002.3124	999	29.0313438	R-squared =	0.7843	
				Adj R-squared =	0.7841	
				Root MSE =	2.5034	

Y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
x	4.869342	.0808222	60.25	0.000	4.710741	5.027943
_cons	1.037207	.079167	13.10	0.000	.8818536	1.192559

2 . regress y x x3

Source	SS	df	MS	Number of obs = 1000		
Model	27988.5287	2	13994.2643	F(2, 997) =	13762.58	
Residual	1013.78372	997	1.01683422	Prob > F =	0.0000	
Total	29002.3124	999	29.0313438	R-squared =	0.9650	
				Adj R-squared =	0.9650	
				Root MSE =	1.0084	

Y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
x	1.980002	.0517654	38.25	0.000	1.87842	2.081584
x3	.9997269	.0139255	71.79	0.000	.9724001	1.027054
_cons	.9563712	.0319087	29.97	0.000	.8937552	1.018987

Unbiasedness???

True model : $Y_i = 1 + 2X_i + X_i^3 + u_i$, Estimated model : $Y_i = \beta_0 + \beta_1 X_i + v_i$

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

substitute for Y_i

$$= E\left[\frac{\sum_{i=1}^n (X_i - \bar{X})(1 + 2X_i + X_i^3 + u_i)}{\sum_{i=1}^n (X_i - \bar{X})X_i}\right]$$

rewrite

$$= 2 + E\left[\frac{\sum_{i=1}^n (X_i - \bar{X})X_i^3 + \sum_{i=1}^n (X_i - \bar{X})u_i}{\sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})}\right]$$

Law of it. exp.

$$= 2 + E\left[\frac{\sum_{i=1}^n (X_i - \bar{X})E[X_i^3 | X_i] + \sum_{i=1}^n (X_i - \bar{X})E[u_i | X_i]}{\sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})}\right]$$

so

$$E[\hat{\beta}_1] = 2 \quad \text{if} \quad E[u_i | X_i] = 0 \quad \& \quad E[X_i^3 | X_i] = 0$$

We know $E[X_i^3 | X_i] \neq 0$ so $\hat{\beta}_1$ is biased!

Consistency???

True model : $Y_i = 1 + 2X_i + X_i^3 + u_i$, Estimated model : $Y_i = \beta_0 + \beta_1 X_i + v_i$

$$\begin{aligned}
 P\lim \widehat{\beta}_1 &= \frac{P\lim \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{P\lim \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})} = \frac{\text{Cov}(X_i, Y_i)}{\text{Var}(X_i)} \\
 &= \frac{\text{Cov}(X_i, 1 + 2X_i + X_i^3 + u_i)}{\text{Var}(X_i)} \quad \text{substitute } Y_i \\
 &= \frac{2\text{Var}(X_i) + \text{Cov}(X_i, X_i^3) + \text{Cov}(X_i, u_i)}{\text{Var}(X_i)} \\
 &= 2 + \frac{E[(X_i - \mu_x)(X_i^3 - \mu_x^3)]}{\text{Var}(X_i)} + \frac{E[(X_i - \mu_x)(u_i - \mu_u)]}{\text{Var}(X_i)} \quad \mu_x^3 = \mu_u = 0 \\
 &= 2 + \frac{E[(X_i - \bar{X})E[X_i^3 | X_i]]}{\text{Var}(X_i)} + \frac{E[(X_i - \bar{X})E[u_i | X_i]]}{\text{Var}(X_i)}
 \end{aligned}$$

so

$$P\lim \widehat{\beta}_1 = 2 \quad \text{if} \quad E[u_i | X_i] = 0 \quad \& \quad E[X_i^3 | X_i] = 0$$

We know $E[X_i^3 | X_i] \neq 0$ so $\widehat{\beta}_1$ is inconsistent!

A simulation example with misspecification functional form

n=1000

```

1 . program define ols, rclass
    1. drop _all
    2. set obs 1000
    3. gen x=invnorm(uniform())
    4. gen x3=x^3
    5. gen y=1+2*x+x3+invnorm(uniform())
    6. regress y x
    7. end

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

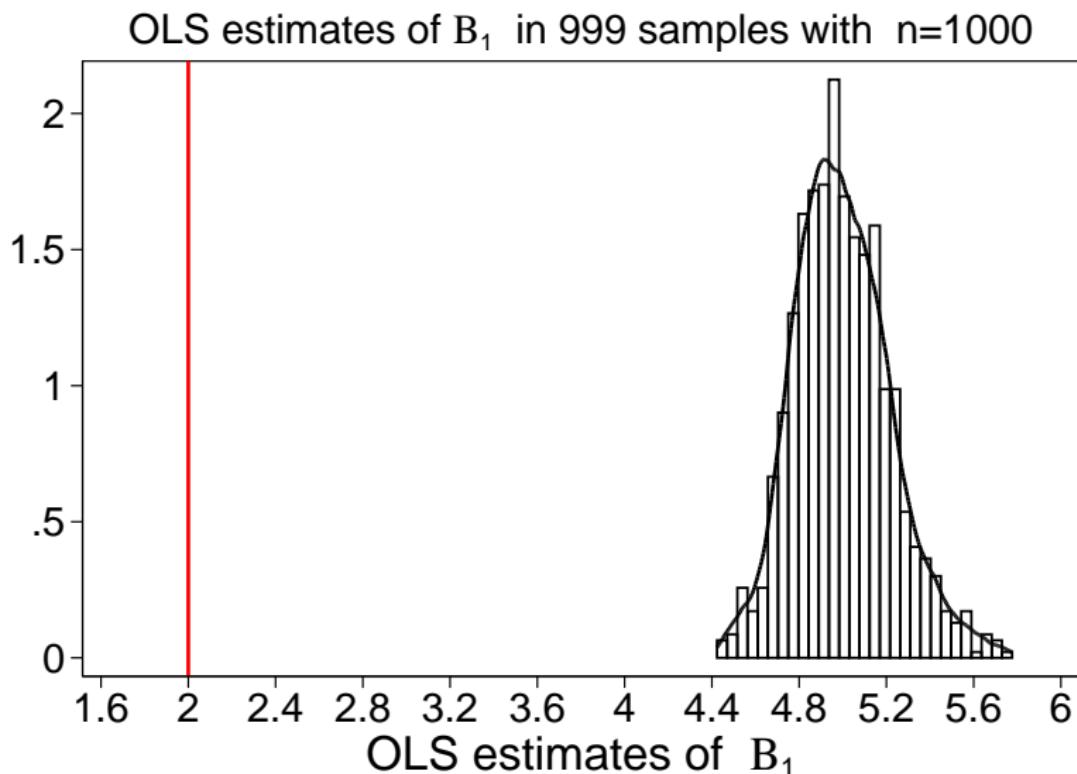
        command:  ols

4 . sum

```

Variable	Obs	Mean	Std. Dev.	Min	Max
_b_x	999	4.998098	.2187471	4.423577	5.77622
_b_cons	999	.996986	.0854847	.7274272	1.292416

A simulation example with misspecification functional form
 $n=1000$



Consistency of the OLS estimator of $\hat{\beta}_1$

True model : $Y_i = 1 + 2X_i + X_i^3 + u_i$, *Estimated model* : $Y_i = \beta_0 + \beta_1 X_i + v_i$

OLS estimates of B_1 in 999 samples
with $n=100$; $n=1000$ and $n=10000$

