POLS 6481. Research Design and Quantitative Methods II
Lecture 6. Omitted Variable Bias
Readings: Wooldridge, *Introductory Econometrics 5e,* 3.2g + 3.3

Outline:

1. Omitted Variable Bias
2. Example: *Temperatures and Shopping*
3. Example: *Water Hardness and Mortality Rates*

POLS 6481. Research Design and Quantitative Methods II
Lecture 6. Omitted Variable Bias
Readings: Wooldridge, *Introductory Econometrics 5e,* 3.2g + 3.3

1. Omitted Variable Bias

Review of the 2-regressor model calculations:

Compare multiple regression… $\widehat{\beta_1} = \frac{cov(x_1,y)}{var(x_1)} - \widehat{\beta_2} \cdot \frac{cov(x_1,x_2)}{var(x_1)}$

…to simple regression $\widetilde{\beta_1} = \frac{cov(x_1,y)}{var(x_1)}$

These will be equal if either (1) $\frac{cov(x_1,x_2)}{var(x_1)} = 0$ i.e., x_1 & x_2 do not covary

or (2) $\widehat{\beta_2} = 0$ i.e., x_2 has no effect on y, after controlling for effect of x_1 on y

(Beware of simultaneity though; you can't replace $\widehat{\beta_2}$ with $\widetilde{\beta_2}$)

If only condition (1) is satisfied, then $\widetilde{\beta_1}$ is unbiased, but including x_2 in the specification may help reduce standard errors, improving model fit if $\frac{cov(x_2,y)}{var(x_2)} \neq 0$.

POLS 6481. Research Design and Quantitative Methods II
Lecture 6. Omitted Variable Bias
Readings: Wooldridge, *Introductory Econometrics 5e,* 3.2g + 3.3

1. Omitted Variable Bias

Review of the 2-regressor model calculations:

Compare multiple regression… $\widehat{\beta_1} = \frac{cov(x_1,y)}{var(x_1)} - \widehat{\beta_2} \cdot \frac{cov(x_1,x_2)}{var(x_1)}$

…to simple regression $\widetilde{\beta_1} = \frac{cov(x_1,y)}{var(x_1)}$

These will be equal if either (1) $\frac{cov(x_1,x_2)}{var(x_1)} = 0$ i.e., x_1 & x_2 do not covary

or (2) $\widehat{\beta_2} = 0$ i.e., x_2 has no effect on y, after controlling for effect of x_1 on y

Suppose $\frac{cov(x_1,x_2)}{var(x_1)} > 0$ and $\widehat{\beta_2} > 0$;

Then $\widehat{\beta_1}$ will be less than $\widetilde{\beta_1}$; if $\widetilde{\beta_1}$ was positive, then it was over-estimating the impact of x_1.

POLS 6481. Research Design and Quantitative Methods II
Lecture 6. Omitted Variable Bias
Readings: Wooldridge, *Introductory Econometrics 5e,* 3.2g + 3.3

1. Omitted Variable Bias

Review of the 2-regressor model calculations:

Compare multiple regression… $\widehat{\beta_1} = \frac{cov(x_1,y)}{var(x_1)} - \widehat{\beta_2} \cdot \frac{cov(x_1,x_2)}{var(x_1)}$

…to simple regression $\widetilde{\beta_1} = \frac{cov(x_1,y)}{var(x_1)}$

These will be equal if either (1) $\frac{cov(x_1,x_2)}{var(x_1)} = 0$ i.e., x_1 & x_2 do not covary

or (2) $\widehat{\beta_2} = 0$ i.e., x_2 has no effect on y, after controlling for effect of x_1 on y

Suppose $\frac{cov(x_1,x_2)}{var(x_1)} > 0$ and $\widehat{\beta_2} < 0$;

Then $\widehat{\beta_1}$ will be greater than $\widetilde{\beta_1}$; if $\widetilde{\beta_1}$ was positive, then it was under-estimating x_1's impact.

POLS 6481. Research Design and Quantitative Methods II
Lecture 6. Omitted Variable Bias
Readings: Wooldridge, *Introductory Econometrics 5e,* 3.2g + 3.3

1. Omitted Variable Bias

Examine Table 3.2 in Wooldridge:

Summary of Bias in $\tilde{\beta}_1$ When x_2 Is Omitted in Estimating Equation (3.40)

	Corr(x_1,x_2) > 0	**Corr(x_1,x_2) < 0**
$\beta_2 > 0$	positive bias	negative bias
$\beta_2 < 0$	negative bias	positive bias

There are *eight* possible scenarios (we are <u>not</u> going to go through all eight), depending on:

whether *y* and x_1 are positively or negatively related

whether *y* and x_2 are positively related (top row) or negatively related (bottom row)

whether x_1 and x_2 are positively related (left column) or negatively related (right column)

POLS 6481. Research Design and Quantitative Methods II
Lecture 6. Omitted Variable Bias
Readings: Wooldridge, *Introductory Econometrics 5e,* 3.2g + 3.3

1. Omitted Variable Bias

Examine Table 3.2 in Wooldridge:

Summary of Bias in $\tilde{\beta}_1$ When x_2 Is Omitted in Estimating Equation (3.40)

	$\text{Corr}(x_1,x_2) > 0$	$\text{Corr}(x_1,x_2) < 0$
$\beta_2 > 0$	positive bias	negative bias
$\beta_2 < 0$	negative bias	positive bias

Suppose y and x_1 are positively related. ↑ pollution → ↑ mortality

Suppose x_2 is positively related to both y and x_1. ↑ % NonWhite → ↑ mortality
↑ % NonWhite ↔ ↑ pollution

Upper-Left Quadrant :

- omitting x_2 would lead us to overestimate the importance of x_1
- including x_2 will decrease the coefficient on x_1

Worst case scenario: $\widetilde{\beta_1}$ is positive, but $\widehat{\beta_1}$ is zero or even negative!!

POLS 6481. Research Design and Quantitative Methods II
Lecture 6. Omitted Variable Bias
Readings: Wooldridge, *Introductory Econometrics 5e,* 3.2g + 3.3

1. Omitted Variable Bias

Examine Table 3.2 in Wooldridge:

Summary of Bias in $\tilde{\beta}_1$ When x_2 Is Omitted in Estimating Equation (3.40)

	$\text{Corr}(x_1,x_2) > 0$	$\text{Corr}(x_1,x_2) < 0$
$\beta_2 > 0$	positive bias	negative bias
$\beta_2 < 0$	negative bias	positive bias

Suppose y and x_1 are positively related. ↑ pollution → ↑ mortality

Suppose x_2 is negatively related to both y and x_1 ↑ Educ → ↓ mortality
↑ Educ ↔ ↓ pollution

Lower-Right Quadrant :

- omitting x_2 would lead us to overestimate the importance of x_1
- including x_2 will decrease the coefficient on x_1

POLS 6481. Research Design and Quantitative Methods II
Lecture 6. Omitted Variable Bias
Readings: Wooldridge, *Introductory Econometrics 5e,* 3.2g + 3.3

1. Omitted Variable Bias

Examine Table 3.2 in Wooldridge:

Summary of Bias in $\tilde{\beta}_1$ When x_2 Is Omitted in Estimating Equation (3.40)

	$\text{Corr}(x_1,x_2) > 0$	$\text{Corr}(x_1,x_2) < 0$
$\beta_2 > 0$	positive bias	negative bias
$\beta_2 < 0$	negative bias	positive bias

Suppose y and x_1 are positively related. ↑ pollution → ↑ mortality

Suppose x_2 is positively related to y ↑ Precip → ↑ mortality
and x_2 is negatively related to x_1 ↑ Precip ↔ ↓ pollution

Upper-Right Quadrant :

- omitting x_2 would lead us to underestimate the importance of x_1
- including x_2 will increase the coefficient on x_1

POLS 6481. Research Design and Quantitative Methods II
Lecture 6. Omitted Variable Bias
Readings: Wooldridge, *Introductory Econometrics 5e,* 3.2g + 3.3

1. Omitted Variable Bias

Examine Table 3.2 in Wooldridge:

Summary of Bias in $\tilde{\beta}_1$ When x_2 Is Omitted in Estimating Equation (3.40)

	$\text{Corr}(x_1,x_2) > 0$	$\text{Corr}(x_1,x_2) < 0$
$\beta_2 > 0$	positive bias	negative bias
$\beta_2 < 0$	negative bias	positive bias

Suppose y and x_1 are positively related. ↑ temperature → ↑ shopping

Suppose x_2 is positively related to y December → ↑ shopping
and x_2 is negatively related to x_1 December ↔ ↓ temperature

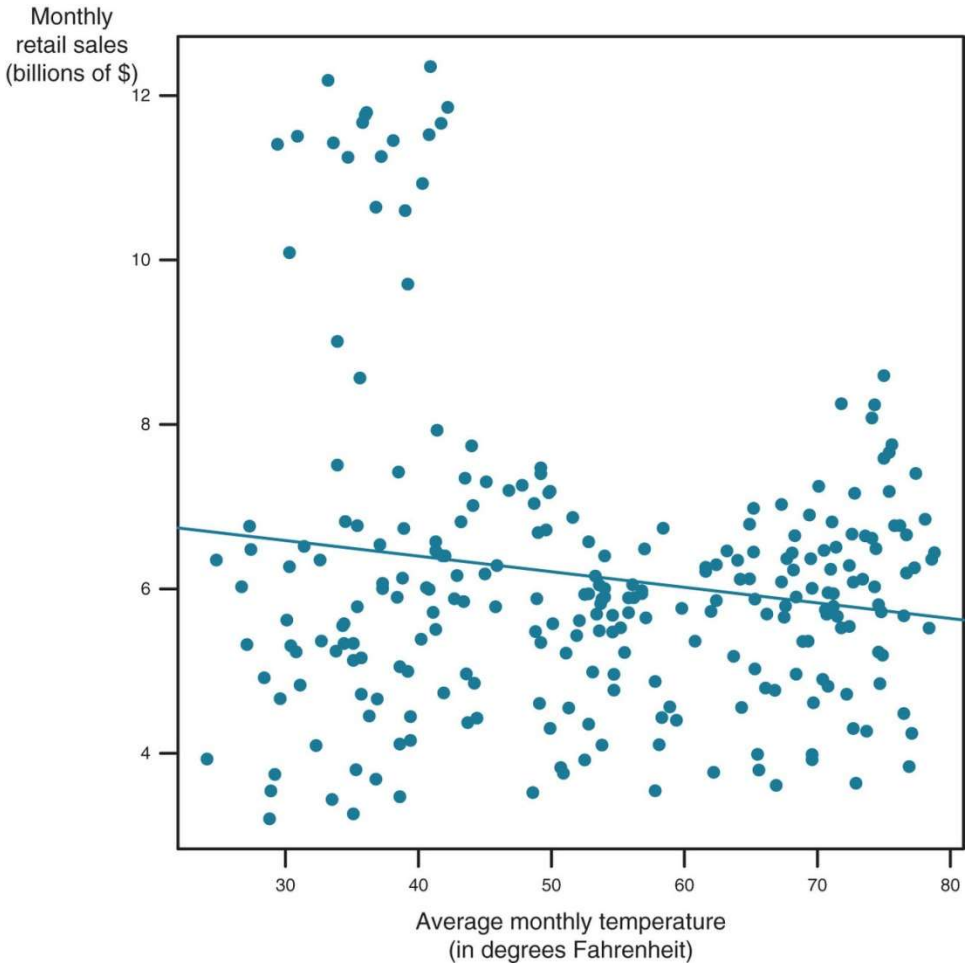
Upper-Right Quadrant :

- omitting x_2 would lead us to underestimate the importance of x_1
- including x_2 will increase the coefficient on x_1

This ought to help recover effects that seemed null ($\widetilde{\beta_1} = 0$) or counterintuitive ($\widetilde{\beta_1} < 0$)

POLS 6481. Research Design and Quantitative Methods II
Lecture 6. Omitted Variable Bias
Readings: Wooldridge, *Introductory Econometrics 5e,* 3.2g + 3.3

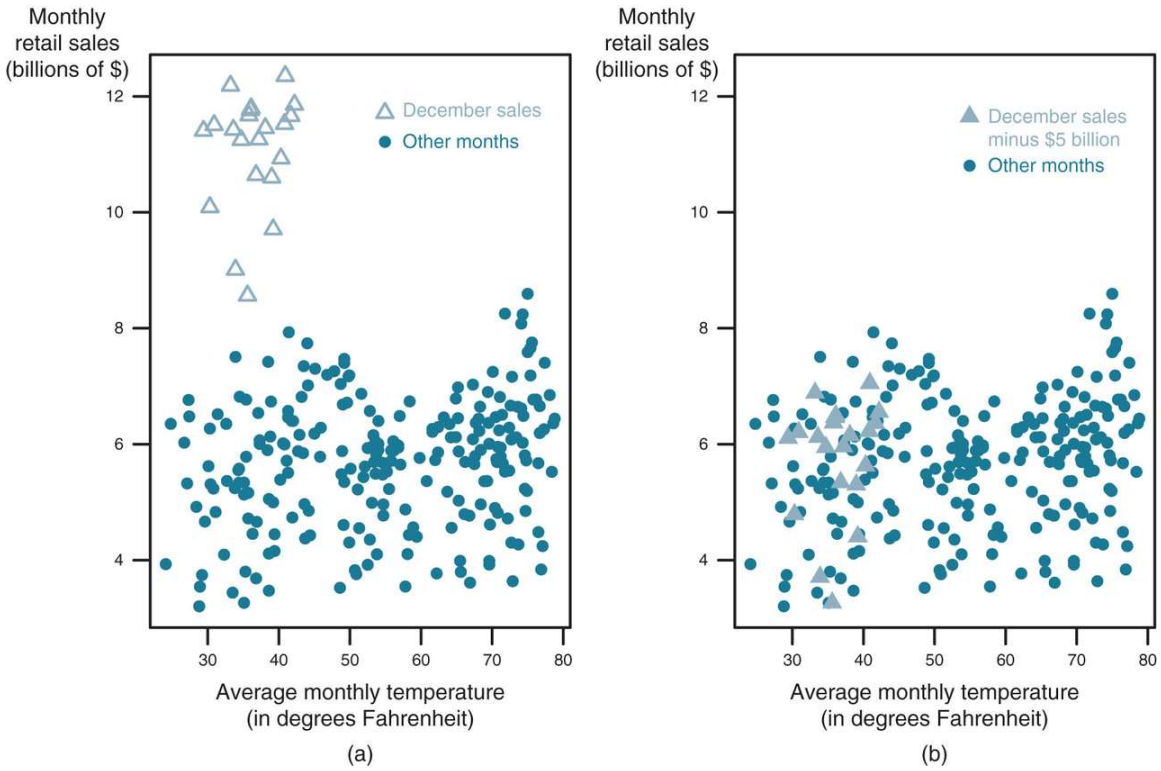
2. Example: *Temperatures and Shopping*



Michael A. Bailey, Real Econometrics

POLS 6481. Research Design and Quantitative Methods II
Lecture 6. Omitted Variable Bias
Readings: Wooldridge, *Introductory Econometrics 5e,* 3.2g + 3.3

2. Example: *Temperatures and Shopping*



POLS 6481. Research Design and Quantitative Methods II
Lecture 6. Omitted Variable Bias
Readings: Wooldridge, *Introductory Econometrics 5e,* 3.2g + 3.3

2. Example: *Temperatures and Shopping*

TABLE 5.1 Bivariate and Multivariate Results for Retail Sales Data

	Bivariate	Multivariate
Temperature	−0.019*	0.014*
	(0.007)	(0.005)
	[$t = 2.59$]	[$t = 3.02$]
December		5.63*
		(0.26)
		[$t = 21.76$]
Constant	7.16*	4.94*
	(0.41)	(0.26)
	[$t = 17.54$]	[$t = 18.86$]
N	256	256
$\hat{\sigma}$	1.82	1.07
R^2	0.026	0.661

Standard errors in parentheses.

* *indicates significance at* $p < 0.05$, *two-tailed.*

POLS 6481. Research Design and Quantitative Methods II
Lecture 6. Omitted Variable Bias
Readings: Wooldridge, *Introductory Econometrics 5e,* 3.2g + 3.3

2. Example: *Temperatures and Shopping*

Simple regression (or difference-of-means test) ____states effect of December ($\widetilde{\beta_2}$ = –+)

Simple regression of retail sales on temperature reveals negative effect ($\widetilde{\beta_1}$ = –0.019)

Multiple regression yields positive effect for December ($\widehat{\beta_2}$ = +5.63)

Multiple regression yields positive estimated effect for temperature ($\widehat{\beta_1}$ = +0.014)

Summary of Bias in $\tilde{\beta}_1$ When x_2 Is Omitted in Estimating Equation (3.40)

	Corr(x_1,x_2) > 0	**Corr(x_1,x_2) < 0**
$\beta_2 > 0$	positive bias	negative bias
$\beta_2 < 0$	negative bias	positive bias

POLS 6481. Research Design and Quantitative Methods II
Lecture 6. Omitted Variable Bias
Readings: Wooldridge, *Introductory Econometrics 5e,* 3.2g + 3.3

3. Example: *Water Hardness and Mortality Rates*

To investigate environmental causes of disease, data were obtained on annual mortality rates per 100,000 males, averaged over the years 1958 – 1964, and the calcium concentration (in parts per million) in the water supply for 61 cities and large towns in England and Wales.

location	town	mortality	hardness	location	town	mortality	hardness	location	town	mortality	hardness
South	Bath	1247	105	North	Grimsby	1379	94	North	Rochdale	1711	13
North	Birkenhead	1668	17	North	Halifax	1742	8	North	Rotherham	1444	14
South	Birmingham	1466	5	North	Huddersfield	1574	9	North	St Helens	1591	49
North	Blackburn	1800	14	North	Hull	1569	91	North	Salford	1987	8
North	Blackpool	1609	18	South	Ipswich	1096	138	North	Sheffield	1495	14
North	Bolton	1558	10	North	Leeds	1591	16	South	Southampton	1369	68
North	Bootle	1807	15	South	Leicester	1402	37	South	Southend	1257	50
South	Bournemouth	1299	78	North	Liverpool	1772	15	North	Southport	1587	75
North	Bradford	1637	10	North	Manchester	1828	8	North	South Shields	1713	71
South	Brighton	1359	84	North	Middlesbrough	1704	26	North	Stockport	1557	13
South	Bristol	1392	73	North	Newcastle	1702	44	North	Stoke	1640	57
North	Burnley	1755	12	South	Newport	1581	14	North	Sunderland	1709	71
South	Cardiff	1519	21	South	Northampton	1309	59	South	Swansea	1625	13
South	Coventry	1307	78	South	Norwich	1259	133	North	Wallasey	1625	20
South	Croydon	1254	96	North	Nottingham	1427	27	South	Walsall	1527	60
North	Darlington	1491	20	North	Oldham	1724	6	South	West Bromwich	1627	53
North	Derby	1555	39	South	Oxford	1175	107	South	West Ham	1486	122
North	Doncaster	1428	39	South	Plymouth	1486	5	South	Wolverhampton	1485	81
South	East Ham	1318	122	South	Portsmouth	1456	90	North	York	1378	71
South	Exeter	1260	21	North	Preston	1696	6				
North	Gateshead	1723	44	South	Reading	1236	101				

POLS 6481. Research Design and Quantitative Methods II
Lecture 6. Omitted Variable Bias
Readings: Wooldridge, *Introductory Econometrics 5e,* 3.2g + 3.3

3. Example: *Water Hardness and Mortality Rates*

To investigate environmental causes of disease, data were obtained on annual mortality rates per 100,000 males, averaged over the years 1958 – 1964, and the calcium concentration (in parts per million) in the water supply for 61 cities and large towns in England and Wales.

26 towns are in the south; their mean mortality rate is 1376.8, with a s.d. of 140.3

35 towns are in the north; their mean mortality rate is 1633.6, with a s.d. of 136.9

Towns in the south also average higher levels of calcium in the water than towns in the north. Here is the variance-covariance matrix:

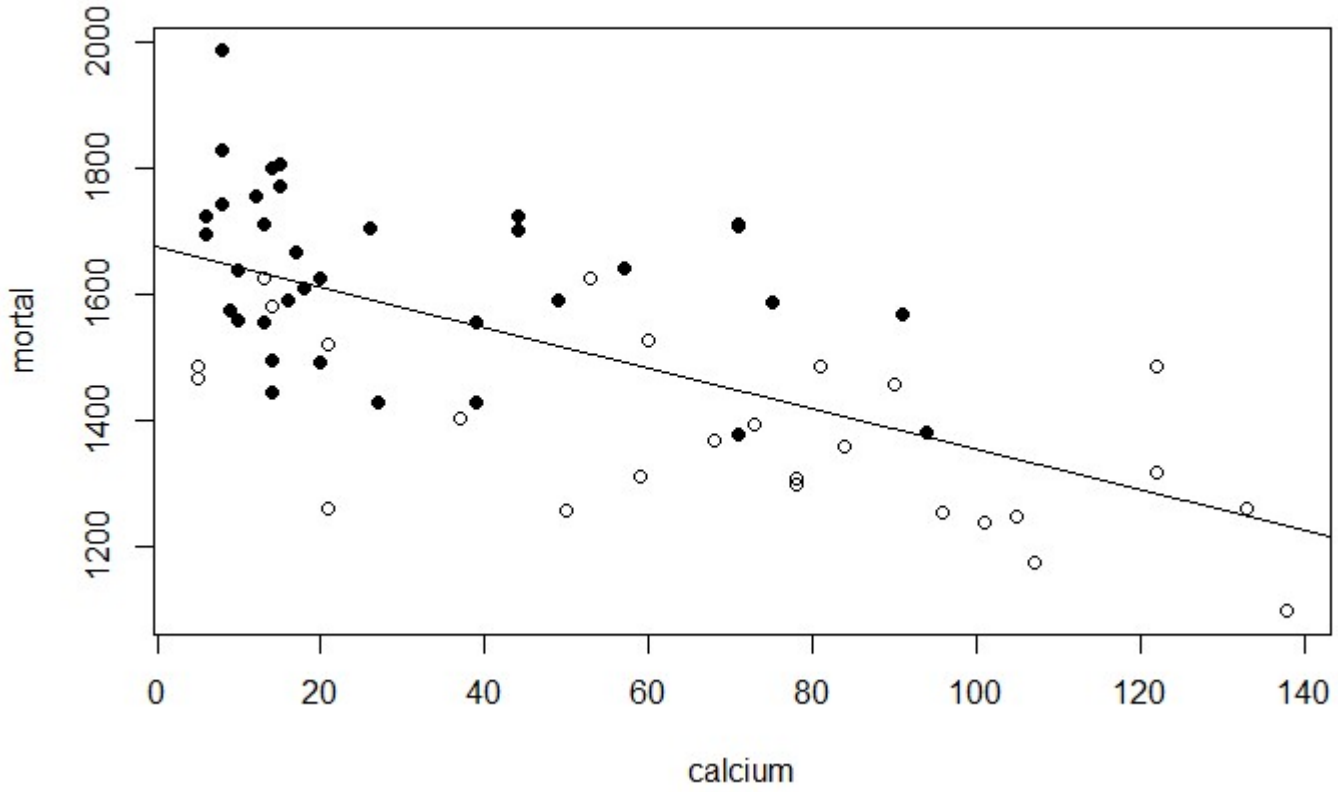
	mortality	calcium	south
mortality	35219.60		
calcium	–4681.54	1451.15	
south	–63.85	9.79	.25

Use this matrix to calculate the correlations between mortality, calcium, and region and to calculate the coefficients for regressions of mortality on calcium and/or region.
Next, use this matrix to explore omitted variable bias (effects of omitting region).

POLS 6481. Research Design and Quantitative Methods II
Lecture 6. Omitted Variable Bias
Readings: Wooldridge, *Introductory Econometrics 5e,* 3.2g + 3.3

3. Example: *Water Hardness and Mortality Rates*

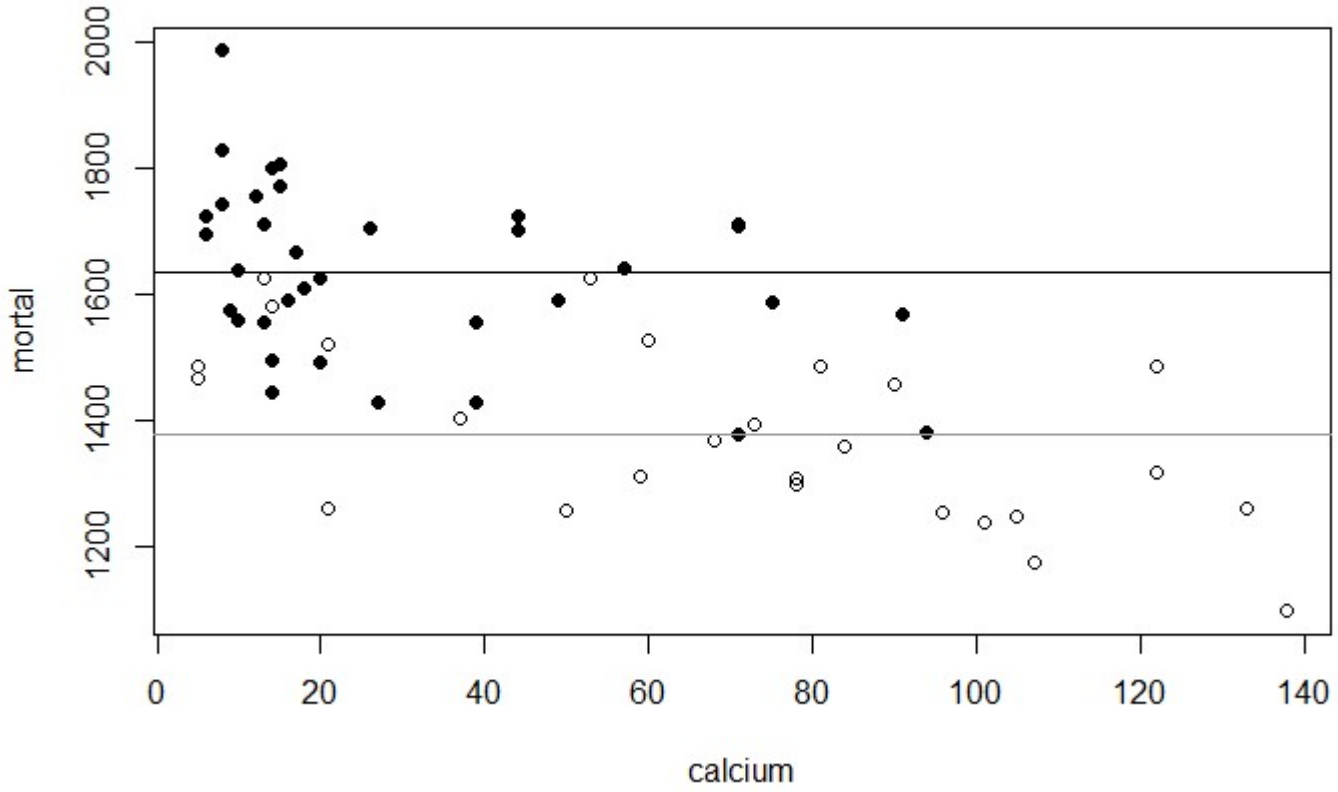
Figure 1. Plot mortality vs. calcium, with simple regression line



POLS 6481. Research Design and Quantitative Methods II
Lecture 6. Omitted Variable Bias
Readings: Wooldridge, *Introductory Econometrics 5e,* 3.2g + 3.3

3. Example: *Water Hardness and Mortality Rates*

Figure 2. Plot mortality vs. calcium, with horizontal lines for regional mean mortality rates

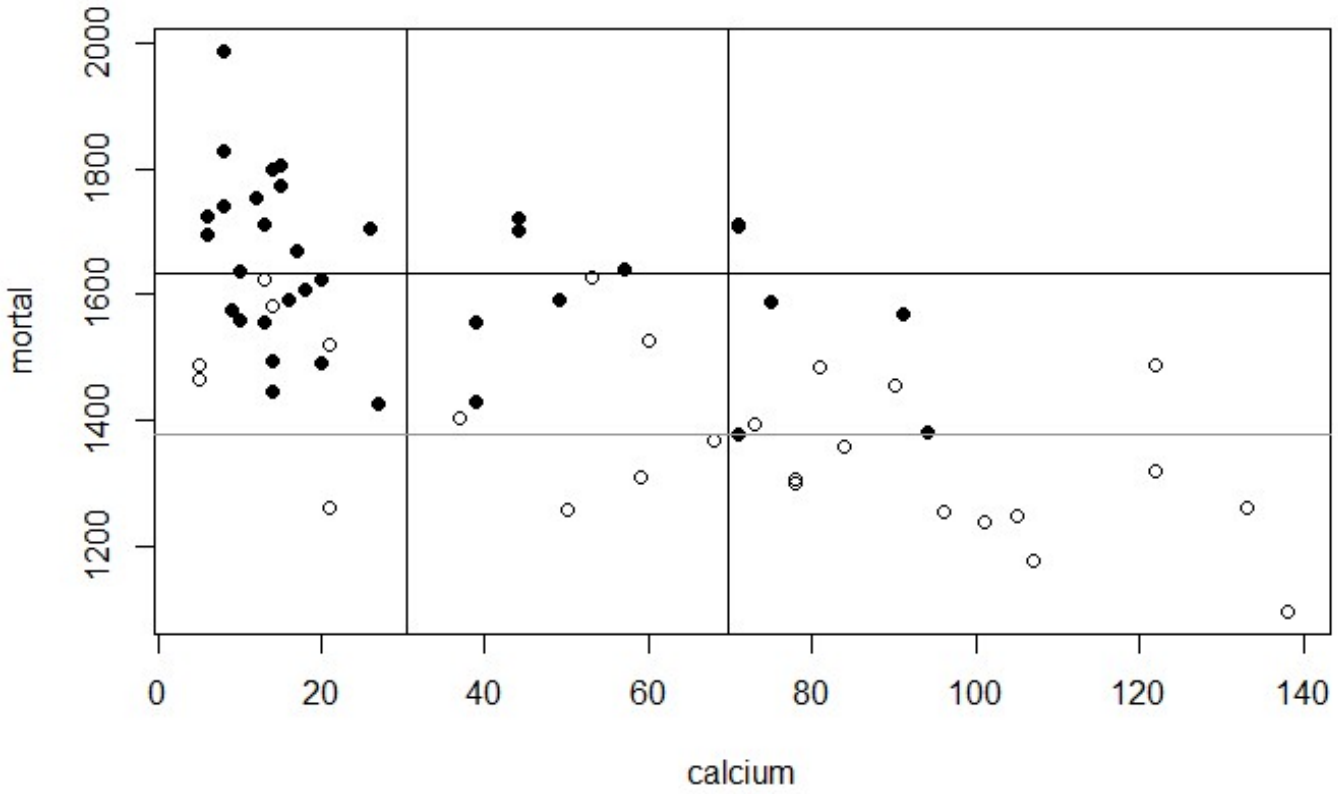


POLS 6481. Research Design and Quantitative Methods II
Lecture 6. Omitted Variable Bias
Readings: Wooldridge, *Introductory Econometrics 5e,* 3.2g + 3.3

3. Example: *Water Hardness and Mortality Rates*

Figure 2. Plot mortality vs. calcium, with horizontal lines for regional mean mortality rates

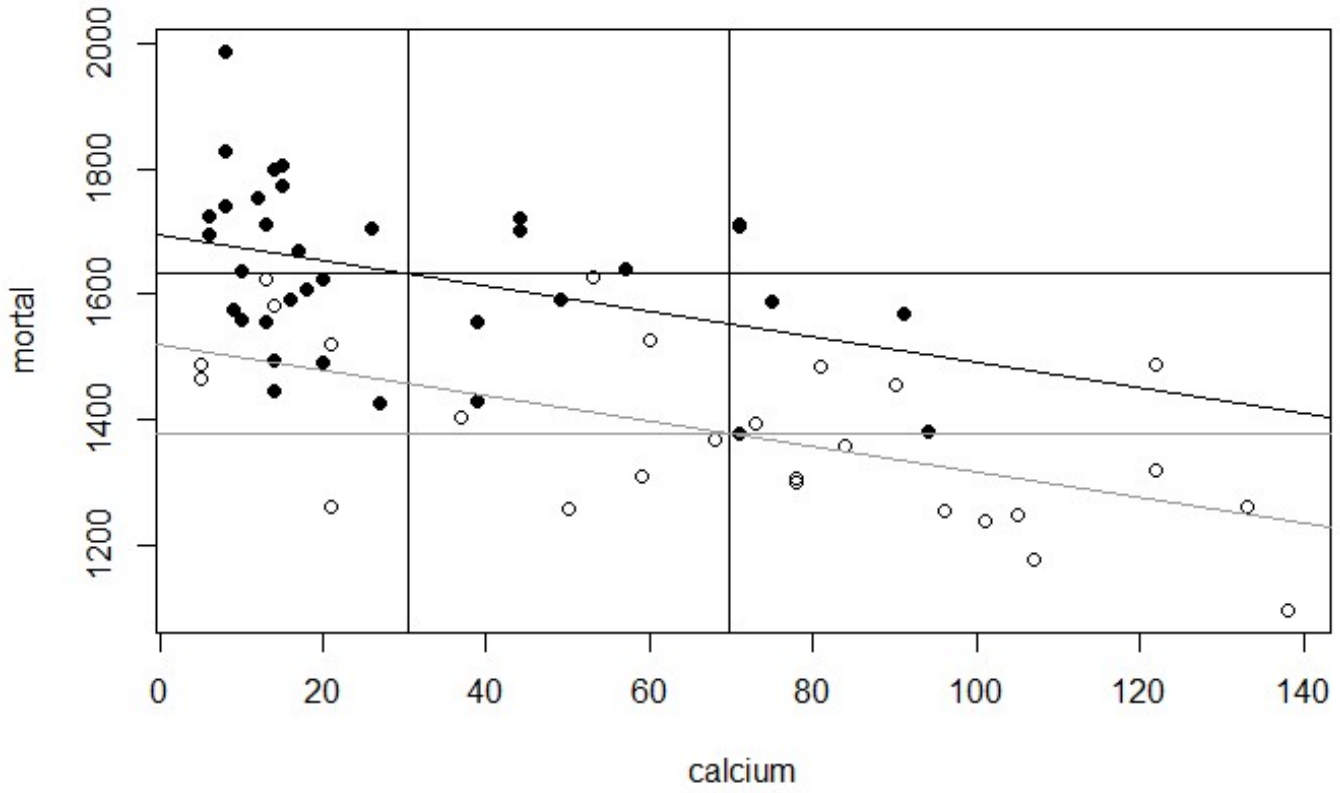
… plus vertical lines for regional mean calcium levels



POLS 6481. Research Design and Quantitative Methods II
Lecture 6. Omitted Variable Bias
Readings: Wooldridge, *Introductory Econometrics 5e,* 3.2g + 3.3

3. Example: *Water Hardness and Mortality Rates*

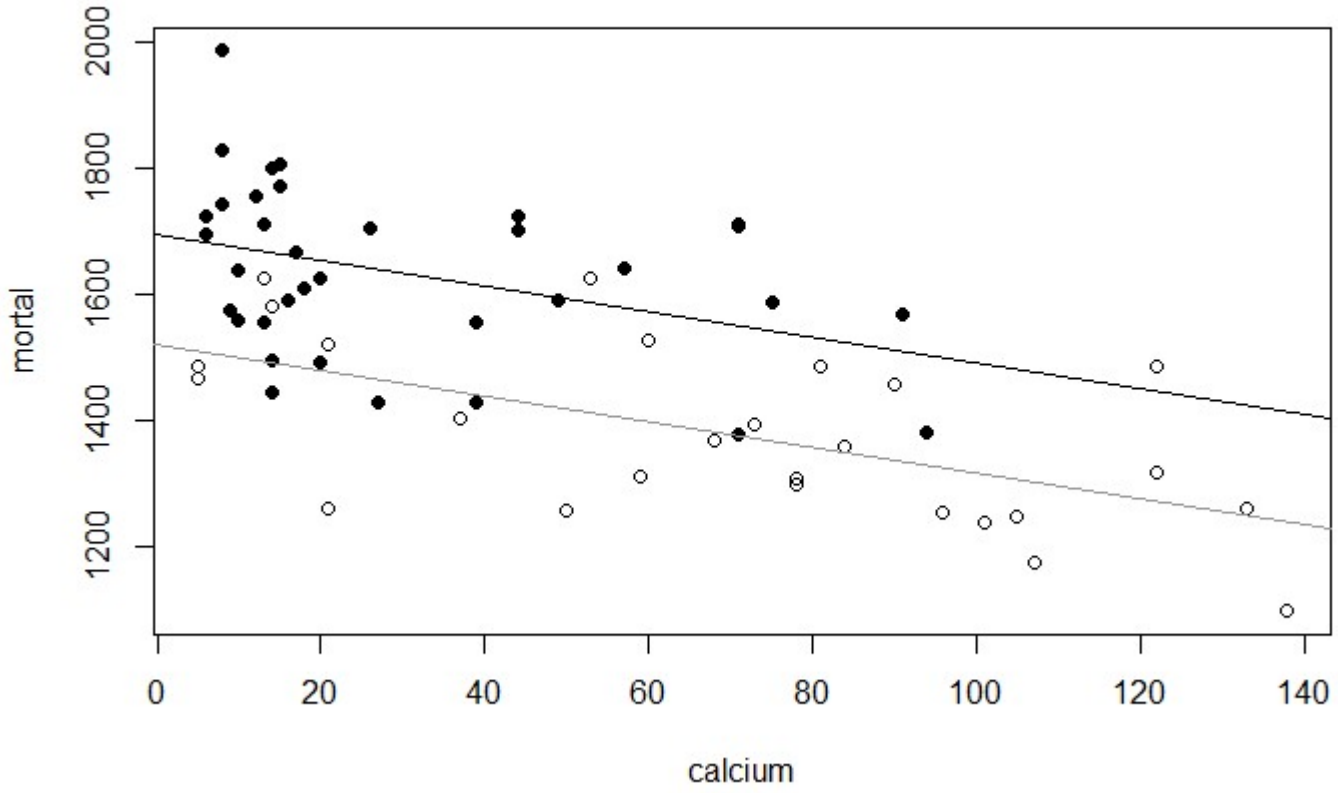
Figure 3. Plot mortality vs. calcium, with multiple regression lines



POLS 6481. Research Design and Quantitative Methods II
Lecture 6. Omitted Variable Bias
Readings: Wooldridge, *Introductory Econometrics 5e,* 3.2g + 3.3

3. Example: *Water Hardness and Mortality Rates*

Figure 3. Plot mortality vs. calcium, with multiple regression lines (region shifts intercept)



POLS 6481. Research Design and Quantitative Methods II
Lecture 6. Omitted Variable Bias
Readings: Wooldridge, *Introductory Econometrics 5e,* 3.2g + 3.3

3. Example: *Water Hardness and Mortality Rates*

Simple regression (or difference-of-means test) overstates effect of region ($\widetilde{\beta_2} = -256.8$)

Simple regression of mortality on calcium overstates effect of calcium ($\widetilde{\beta_1} = -3.226$)

Examine residuals:

- positive residuals tend to come from north (mean residual = 55.3)
- negative residuals tend to come from south (mean residual = –74.5)
- region dummy variable correlates to residuals ($r_{u,x2} = -.456$)

Multiple regression yields smaller estimated impact of region ($\widehat{\beta_2} = -176.7$)

Multiple regression yields smaller estimated impact of calcium ($\widehat{\beta_1} = -2.034$)

Summary of Bias in $\tilde{\beta}_1$ When x_2 Is Omitted in Estimating Equation (3.40)

	Corr(x_1,x_2) > 0	**Corr(x_1,x_2) < 0**
$\beta_2 > 0$	positive bias	negative bias
$\beta_2 < 0$	negative bias	positive bias