

Endogeneity

Exogeneity

We will continue to show how we can weaken the assumptions of the linear regression model

Now, we're going to tackle Assumption 2, the mean-zero errors assumption

This assumption is also known as **exogeneity**

Recall that this assumption holds that, conditional on all of the x 's, $E(\varepsilon_i) = 0$

Also recall that we said that the real assumption here is that the mean of ε doesn't depend on those x 's

In fact, we can restate this assumption as $Cov(\varepsilon_i, x) = 0$ (assuming a univariate regression for simplicity)

Now, we'll investigate the consequences of failures of this assumption

For simplicity, we'll do our analysis in the context of a univariate regression, but our conclusions will carry over to the multivariate case

Endogeneity

Endogeneity is when exogeneity does *not* hold, so that $Cov(x, \varepsilon) \neq 0$

As we will see, the consequences of endogeneity are much more severe than was the case for heteroskedasticity or autocorrelation

The reason for this was that we used the assumption of *exogeneity* when we proved that OLS is unbiased and consistent

Therefore, *endogeneity* will mean that our regressions are not good estimates of the population regression coefficients that we want to know

There are several reasons why we might have endogeneity

1. Omitted-variable bias. One case we've already talked about is when there is omitted-variable bias

Suppose that the true model is

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + u$$

If we exclude x_2 from the regression, then we can write the model as

$$y = \alpha + \beta_1 x_1 + \underbrace{(\beta_2 x_2 + u)}_{\varepsilon} = \alpha + \beta_1 x_1 + \varepsilon$$

If x_2 is correlated with x_1 , the ε will be as well

We already know that the coefficient on x_1 from a model that excludes x_2 is

$$E(b) = \beta_1 + \beta_2 b_{x_2 x_1}$$

where $b_{x_2 x_1}$ is the slope coefficient from a regression of x_2 on x_1

If x_2 and x_1 are correlated, $b_{x_2 x_1} \neq 0$

Hence, OLS will be biased (and also inconsistent)

2. Simultaneity/reverse causality. Suppose that the demand curve is given by

$$Q = a - bP + e$$

and the supply curve is

$$Q = c + dP + u$$

Now suppose that you have data on on prices and quantities from multiple market equilibria (say, from one market over time, or across different markets)

Question: Does a regression of Q on P tell you the demand curve or the supply curve?

Every econ 202 student knows that, in equilibrium, $Q^D = Q^S$, which implies that the equilibrium price satisfies

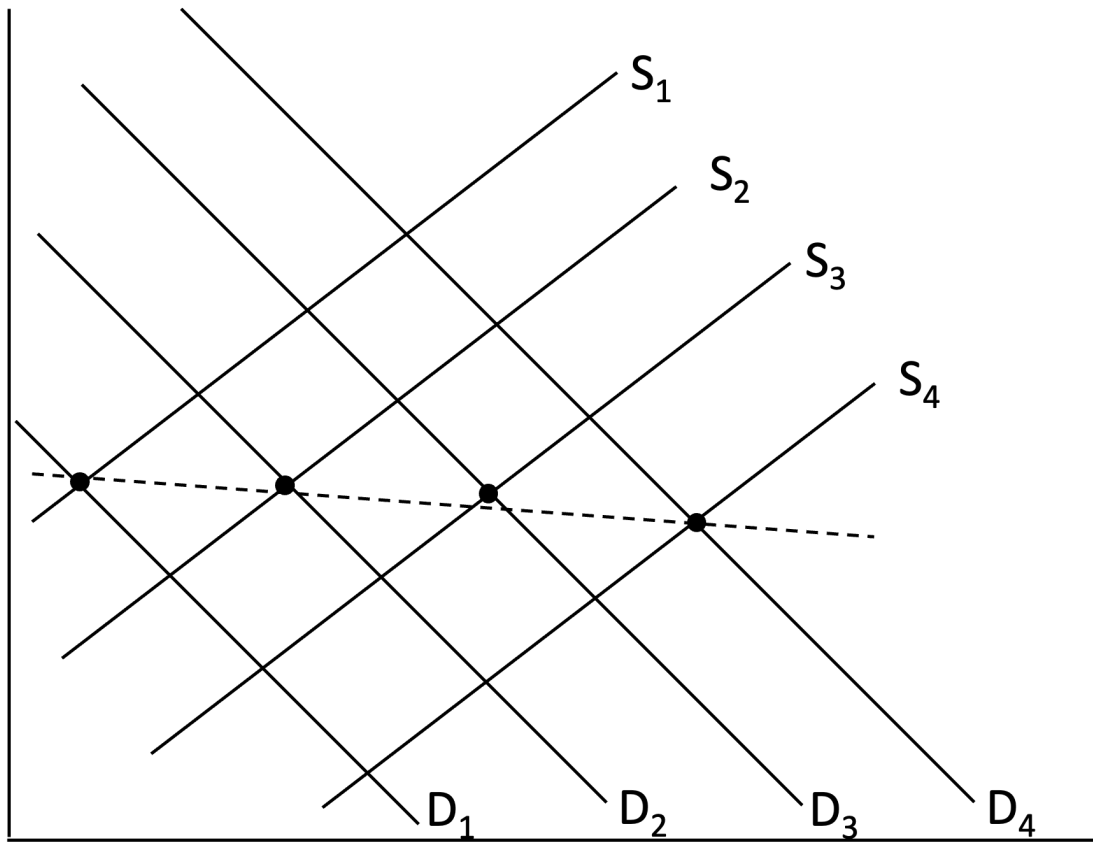
$$P^* = \frac{c - a + u - e}{d - b}$$

But this is a function of the error terms in both the supply and demand curves

Regardless of whether we're trying to estimate supply or demand, the price will be endogenous

Answer: A regression of Q on P doesn't tell us the supply curve or the demand curve

This is best understood graphically



Since P is endogenous, OLS doesn't tell us supply or demand

Consequences of endogeneity

We've already seen intuition for, and examples of, the consequences of endogeneity, but let's take a more formal look

Recall from our discussion of the properties of regression that (assuming deviations from means form)

$$b = \beta + \frac{\sum x_i \varepsilon_i}{\sum x_i^2}$$

Therefore,

$$\begin{aligned} E(b) &= \beta + E \left(\frac{\sum x_i \varepsilon_i}{\sum x_i^2} \right) \\ &= \beta + \frac{\sum_i x_i E(\varepsilon_i)}{\sum_i x_i^2} \end{aligned}$$

When exogeneity held, this last term vanished, so that b was unbiased

Under endogeneity, it won't, so b is biased. Since $Var(b) \rightarrow 0$ as n grows, this also implies that b is inconsistent

Two-stage least squares/instrumental variables

What can we do about endogeneity?

If the endogeneity arises because of omitted variables, one solution is to simply include the omitted variables

But we might not have data on an important variable that we think is omitted

For example, suppose that our model is

$$wage = \alpha + \beta_1 educ + \varepsilon,$$

and we're worried that ambition, which might be correlated with education, is an omitted variable

It could be very difficult to find a dataset that includes any reasonable measure of a person's ambition

Now suppose that we're worried about endogeneity due to simultaneity

In this case, the endogeneity is really caused by the fact that P and Q are determined in equilibrium by the intersection of supply and demand

In this case, the endogeneity doesn't really arise because of an omitted variable, so it's not obvious that we can solve it by just controlling for additional variables

Two-stage least squares is one solution to this problem

Suppose that we have an **instrumental variable** z that satisfies the following conditions:

1. **Exogeneity:** $Cov(z, \varepsilon) = 0$, and
2. **Relevance:** $Cov(z, x) \neq 0$

In other words, z is correlated with x but, unlike x , z is not correlated with ε

Instrumental variables are sometimes called **natural experiments**, because it is as if z is randomly changing x without having any other effect on y

We summarize this by saying “ z only affects y *through* x ”

Now consider the following **instrumental variables** or **two-stage least squares** procedure:

1. Regress x on z to obtain the predicted values \hat{x}
2. Regress y on \hat{x}

Why would we do this?

The predicted value of \hat{x} is clearly related to x

But since the prediction is based on z (which is uncorrelated with ε), \hat{x} is also uncorrelated with ε

We can think of \hat{x} as “the part of x that is uncorrelated with ε ”

Properties of IV/2SLS

If you suspect that regressing y on \hat{x} (instead of x itself) will produce a good estimate of β , you're right

To show this, let's go back to deviations-from-means form, in which case our regressions don't need constant terms

In this case, we can write our predicted value of x as

$$\hat{x}_i = d * z_i,$$

where d is the coefficient from a regression of x on z

The two-stage least squares estimate of β is just the coefficient from a regression of y on \hat{x} , or

$$b^{IV} = \frac{\sum_i \hat{x}_i y_i}{\sum_i \hat{x}_i^2} = \frac{\sum_i dz_i y_i}{\sum_i (dz_i)^2} = \frac{d \sum_i z_i y_i}{d \sum_i dz_i^2} = \frac{\sum_i z_i y_i}{\sum_i (dz_i) z_i}$$

We can write $x_i = dz_i + f_i$, where f_i is the residual from regressing x on z

From our discussion of residual properties, we also know that $\sum_i z_i f_i = 0$, so we can add this to the last line:

$$b^{IV} = \frac{\sum_i z_i y_i}{\sum_i [(dz_i) z_i + z_i f_i]} = \frac{\sum_i z_i y_i}{\sum_i (dz_i + f) z_i} = \frac{\sum_i z_i y_i}{\sum_i x_i z_i}$$

Now, let's use the fact that we don't need a constant in deviations-from-means form to write

$$y_i = \beta x_i + \varepsilon_i$$

Plugging in,

$$\begin{aligned} b^{IV} &= \frac{\sum_i z_i y_i}{\sum_i x_i z_i} = \frac{\sum_i z_i (x_i \beta + \varepsilon_i)}{\sum_i x_i z_i} = \frac{\beta \sum_i z_i x_i}{\sum_i x_i z_i} + \frac{\sum_i z_i \varepsilon_i}{\sum_i x_i z_i} \\ &= \beta + \frac{\sum_i z_i \varepsilon_i}{\sum_i x_i z_i} \end{aligned}$$

This is the home stretch

Rewrite b^{IV} as

$$b^{IV} = \beta + \frac{\sum_i z_i \varepsilon_i}{\sum_i x_i z_i} = \beta + \frac{\frac{1}{n} \sum_i z_i \varepsilon_i}{\frac{1}{n} \sum_i x_i z_i}$$

and recall that the **Law of large numbers** says that sample means converge to populations means as n grows, so

$$b^{IV} \rightarrow \beta + \frac{E(z_i \varepsilon_i)}{E(z_i x_i)}$$

$$b^{IV} \rightarrow \beta + \frac{E(z_i \varepsilon_i)}{E(z_i x_i)}$$

But in deviations-from-means form,

$$Cov(z_i, \varepsilon_i) = E\{[z_i - E(z_i)][\varepsilon_i - E(\varepsilon_i)]\} = E(z_i \varepsilon_i),$$

so the second term is zero, and $b^{IV} \rightarrow \beta$

This shows that IV/2SLS is **consistent**

Finishing up the properties of IV/2SLS, an argument similar to the one we used for OLS (I'll spare you the derivation) shows that (no longer assuming deviations-from-means form)

$$V(b^{IV}) \approx \frac{\sigma^2 \sum_i (z_i - \bar{z})^2}{\sum_i (z_i - \bar{z})(x_i - \bar{x})}$$

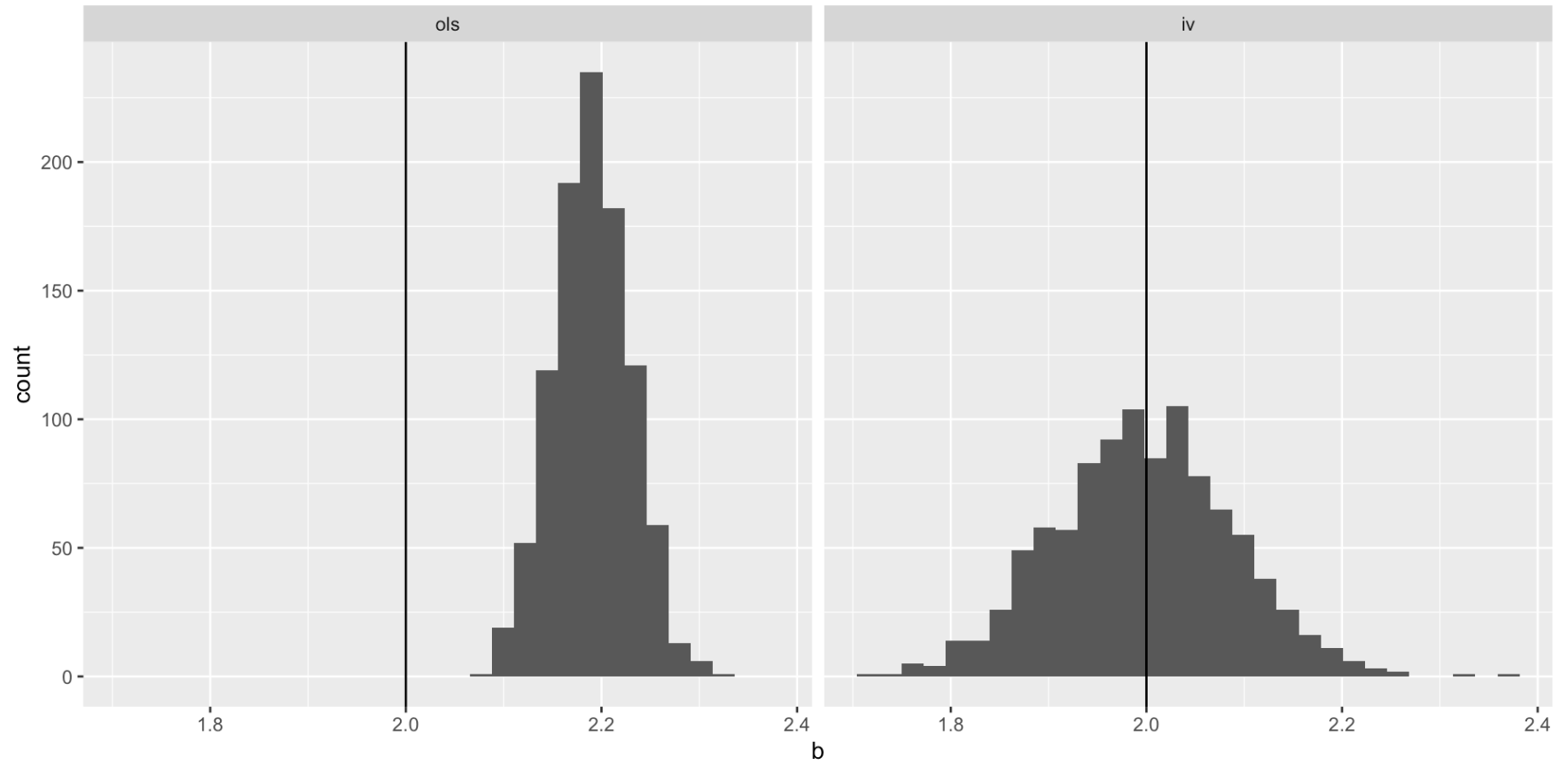
We can estimate σ^2 using

$$s_{IV}^2 = \frac{\sum_i e_i^2}{n - 2} = \frac{\sum_i (y_i - a^{IV} - b^{IV} x_i)^2}{n - 2}$$

Let's do a simulation to see how well this works. We can use the `ivreg` package to do IV regressions in R:

```
1 library(tidyverse)
2 library(ivreg)
3 b <- rep(0,1000)
4 b.iv <- b
5 for (i in 1:1000) {
6   z <- rnorm(1000)
7   u <- rnorm(1000)
8   x <- .5*z + .25*u + rnorm(1000) # x deps on z and u
9   e <- u + rnorm(1000) # e deps on u
10  y <- 1 + 2*x + e
11  b[i] <- lm(y ~ x)$coef[2]
12  b.iv[i] <- ivreg(y ~ x | z)$coefficients[2]
13 }
14 bs <- data.frame(b=c(b, b.iv))
15 bs$est <- factor(c(rep("ols", 1000), rep("iv",1000)), level=c("ols", "iv"))
```

```
1 ggplot(bs, aes(b)) + geom_histogram() + facet_wrap("factor(est)") + geom_vl
```



“IV” vs. “2SLS”

We have been using the terms IV and 2SLS interchangeably

When there is one IV per endogenous variable, IV and 2SLS are the same

However, we can also use our procedure when we have more IVs than endogenous variables

To do this, we obtain \hat{x} by regressing x on *all* of the IVs

In this case, the term “2SLS” is more appropriate

Multivariate IV/2SLS

IV/2SLS carry over to the multivariate case with very little change

Suppose that our model is

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \varepsilon,$$

where we're worried that x_1 is endogenous, but we have an IV z_1

We can do IV/2SLS by regressing x_1 on z_1 and x_2 to obtain \hat{x}_1

If we also thought that x_2 were endogenous, we'd need an additional IV for that variable

“Manual” vs “canned” IV/2SLS

Although we can implement our two-stage procedure manually, it's better to use a pre-programmed IV/2SLS function

The reason for this is that when we 2SLS manually, our software tries to estimate ε as $e_i = y_i - b^{IV} \hat{x}_i$, when what we really want is $y_i - b^{IV} x_i$ (i.e., using the *real*, rather than predicted, x)

A pre-programmed IV function will handle this automatically

Let's illustrate this:

```
1 z <- rnorm(1000)
2 u <- rnorm(1000)
3 x <- .5*z + .25*u + rnorm(1000) # x deps on z and u
4 e <- u + rnorm(1000) # e deps on u
5 y <- 1 + 2*x + e
6 x.hat <- lm(x ~ z)$fitted.values
7 summary(lm(y ~ x.hat))
```

Call:

```
lm(formula = y ~ x.hat)
```

Residuals:

Min	1Q	Median	3Q	Max
-7.9655	-1.7672	-0.0737	1.8105	10.0714

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	0.9590	0.0883	10.86	<2e-16	***
x.hat	1.9736	0.1707	11.56	<2e-16	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.706 on 998 degrees of freedom

```
1 summary(ivreg(y ~ x | z))
```

Call:

```
ivreg(formula = y ~ x | z)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-4.02056	-0.97578	0.06433	0.93118	5.20668

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	0.95905	0.04697	20.42	<2e-16	***
x	1.97359	0.09082	21.73	<2e-16	***

Diagnostic tests:

	df1	df2	statistic	p-value	
Weak instruments	1	998	227.22	< 2e-16	***

The coefficient estimates are the same, but the SEs are a little different

First-stage regression

Remember, a *valid* IV satisfies two properties

1. **Exogeneity:** $Cov(z, \varepsilon) = 0$, and
2. **Relevance:** $Cov(z, x) \neq 0$

Since we don't have data on ε , we can't really test exogeneity¹

1. Technically, there are some roundabout ways to test this, but they make strong assumptions

However, we *can and should* test the relevance of the instrument

To do this, we can examine the **first-stage** regression of x on the instruments to see if they are statistically significant

In the multivariate case, the best way to do this is using the F-test for the joint significance of all of the regressors (i.e., the null hypothesis that the coefficient on *all* of the variables we used to predict x are zero)

“Weak instruments”

You may have noticed that we proved that b^{IV} is consistent, but didn't say anything about unbiasedness

This is because IV is *not* unbiased¹

So consistency is the best we can do, but this is ok — consistency is a good property for an estimator to have

1. In fact, in many cases, the population mean of b^{IV} doesn't even exist!

It turns out that IV tends to be more biased when the correlation between x and z is small (this is known as a **weak instrument**)

Intuitively, this is because our prediction \hat{x} is based on an estimate of the relationship between x and z and not the *true* relationship, so in small samples, \hat{x} is not completely uncorrelated with ε

A common rule of thumb is that if the F-statistic from the **first-stage regression** of x on the instruments is greater than 10, the bias won't be too bad

There are a couple of other bad things that can happen when your instrument is weak

First, the standard errors might be large

Recall that the variance of the IV slope coefficient is

$$V(b^{IV}) \approx \frac{\sigma^2 \sum_i (z_i - \bar{z})^2}{\sum_i (z_i - \bar{z})(x_i - \bar{x})}$$

The denominator is essentially the sample covariance between z and x —when this is low, $V(b^{IV})$ will be large

We care about this because when the SEs are large, we might not be able to reject null hypotheses, even when they're actually false

The second problem is that if the instrument actually turns out to be *endogenous*, IV can be *more biased* than OLS

To see this, recall that we previously showed that

$$b = \beta + \frac{\sum_i x_i \varepsilon_i}{\sum_i x_i^2} \rightarrow \beta + \frac{Cov(x, \varepsilon)}{Var(x)}$$
$$b^{IV} = \beta + \frac{\sum_i z_i \varepsilon_i}{\sum_i x_i z_i} \rightarrow \beta + \frac{Cov(z, \varepsilon)}{Cov(x, z)}$$

When the instrument is weak, a small amount of covariance between z and ε gets divided by a *small* number, resulting in a large bias term

Examples

Let's take a look at some examples of IV in action

We'll start by revisiting the example of simultaneity in a supply and demand model

Our dataset contains information on the price of chicken and per-capita chicken consumption, as well as a few other variables

Recall that, due to simultaneity, a regression of quantity on price doesn't tell us the demand or supply curve

```
1 chx <- read_csv("broiler.csv")
2 summary(chx)
```

obs	Q	Y	PCHICK
Min. :1950	Min. :14.30	Min. : 7863	Min. : 47.40
1st Qu.:1963	1st Qu.:20.98	1st Qu.: 9831	1st Qu.: 53.92
Median :1976	Median :28.40	Median :14684	Median : 77.20
Mean :1976	Mean :31.24	Mean :14850	Mean : 90.99
3rd Qu.:1988	3rd Qu.:39.92	3rd Qu.:19600	3rd Qu.:126.75
Max. :2001	Max. :53.90	Max. :23692	Max. :168.00

PBEEF	PCOR	PF	CPI
Min. : 26.70	Min. : 44.80	Min. : 49.53	Min. : 24.10
1st Qu.: 34.40	1st Qu.: 53.90	1st Qu.: 52.98	1st Qu.: 30.50
Median : 60.60	Median : 78.50	Median :103.04	Median : 55.35
Mean : 76.16	Mean : 81.12	Mean : 97.73	Mean : 77.10
3rd Qu.:113.90	3rd Qu.:101.28	3rd Qu.:127.23	3rd Qu.:119.72
Max. :160.50	Max. :158.50	Max. :174.34	Max. :177.10
		NA's : 12	

What happens when we try to estimate the demand curve by regressing Q on P (let's also include GDP in our regression, since this probably affects demand):

```
1 model.1 <- lm(Q ~ PCHICK + Y, data=chx)
2 summary(model.1)
```

Call:

```
lm(formula = Q ~ PCHICK + Y, data = chx)
```

Residuals:

Min	1Q	Median	3Q	Max
-3.500	-1.459	0.045	1.939	3.729

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-3.7725392	0.9583935	-3.936	0.000261	***
PCHICK	0.0383304	0.0187855	2.040	0.046716	*
Y	0.0021227	0.0001477	14.370	< 2e-16	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Uh oh, the coefficient on P is *positive*, but demand curves are supposed to slope *down*

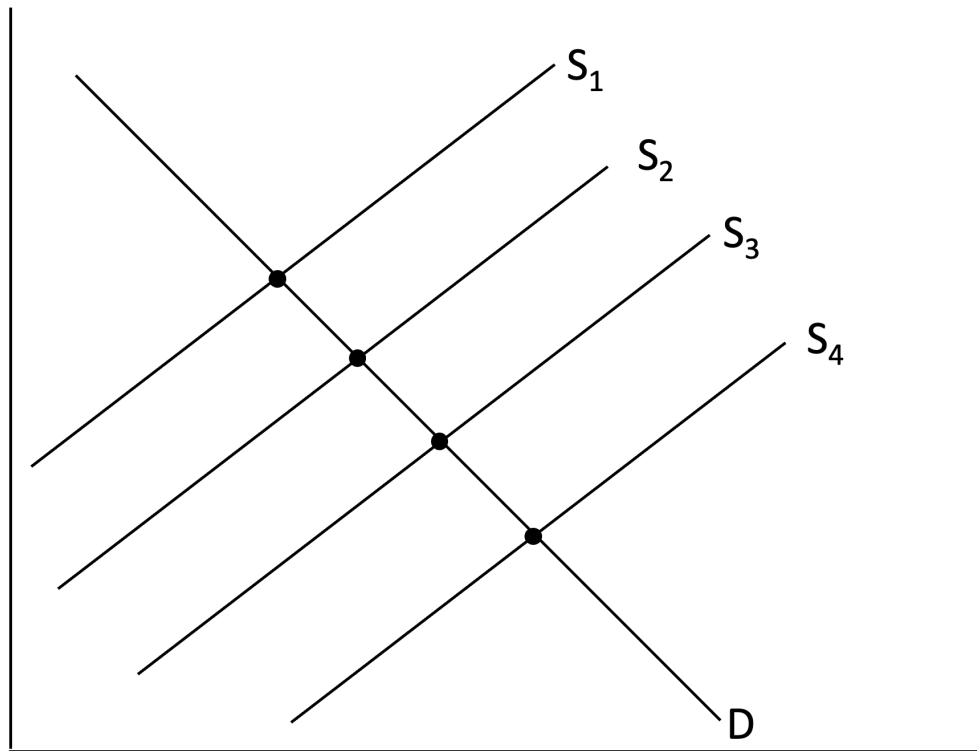
As we saw before, the problem is that the equilibrium price is correlated with the error term in both the demand and supply curves

How can we use IV to solve this problem?

Suppose we have a **supply shifter** that affects the supply curve, but isn't correlated with demand

We can use such a shifter as an IV for the price in the demand equation

Here is a graphical interpretation of this idea:



The supply shifter moves the supply curve without affecting demand, tracing out the demand curve

Our dataset includes the price of chicken feed, which presumably affects the supply of chicken, but not the demand. Let's see what happens when we use this as an IV for the price:

```
1 model.2 <- ivreg(Q ~ PCHICK + Y | PF + Y, data=chx)
2 summary(model.2)
```

Call:

```
ivreg(formula = Q ~ PCHICK + Y | PF + Y, data = chx)
```

Residuals:

Min	1Q	Median	3Q	Max
-4.4966	-2.8526	-0.6623	2.5568	4.9865

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-12.796027	9.392025	-1.362	0.1813
PCHICK	-0.147954	0.166520	-0.889	0.3800
Y	0.003784	0.001555	2.433	0.0199 *

Diagnostic tests:

df1	df2	statistic	p-value
-----	-----	-----------	---------

Now the slope of the demand curve is negative, which is what we'd expect

It's still statistically insignificant, but that might be because we're using such a simple model with few controls

Now let's revisit the effect of education on wages, using a different dataset

```
1 card <- read_csv("card.csv")
2 summary(card)
```

id	nearc2	nearc4	educ
Min. : 2	Min. :0.0000	Min. :0.0000	Min. : 1.00
1st Qu.:1276	1st Qu.:0.0000	1st Qu.:0.0000	1st Qu.:12.00
Median :2541	Median :0.0000	Median :1.0000	Median :13.00
Mean :2582	Mean :0.4409	Mean :0.6821	Mean :13.26
3rd Qu.:3859	3rd Qu.:1.0000	3rd Qu.:1.0000	3rd Qu.:16.00
Max. :5225	Max. :1.0000	Max. :1.0000	Max. :18.00

age	fatheduc	motheduc	weight
Min. :24.00	Min. : 0	Min. : 0.00	Min. : 75607
1st Qu.:25.00	1st Qu.: 8	1st Qu.: 8.00	1st Qu.: 122798
Median :28.00	Median :10	Median :12.00	Median : 365200
Mean :28.12	Mean :10	Mean :10.35	Mean : 321185
3rd Qu.:31.00	3rd Qu.:12	3rd Qu.:12.00	3rd Qu.: 406024
Max. :34.00	Max. :18	Max. :18.00	Max. :1752340
	NA's : 690	NA's : 252	

First, let's run a basic regression on wages on education:

```
1 model.1 <- lm(wage ~ educ, data=card)
2 summary(model.1)
```

Call:

```
lm(formula = wage ~ educ, data = card)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-576.09	-173.36	-34.12	127.82	1686.25

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	183.949	23.104	7.962	2.38e-15	***
educ	29.655	1.708	17.368	< 2e-16	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 250.7 on 3008 degrees of freedom

Here, the wage is measured in cents, so this implies that for every additional year of education, you make about \$.3 more per hour

Of course, we're concerned about omitted variable bias

Maybe people with more motivation or ability, who would have earned higher wages anyway, just tend to get more education

Let's try to control for this by including controls for experience, living in an urban area, and parental education, which might be correlated with both wages and education

```
1 model.2 <- lm(wage ~ educ + exper + smsa + fatheduc + motheduc , data=card)
2 summary(model.2)
```

Call:

```
lm(formula = wage ~ educ + exper + smsa + fatheduc + motheduc,
    data = card)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-647.33	-148.65	-24.97	114.92	1555.55

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-458.236	44.546	-10.287	< 2e-16	***
educ	46.269	2.666	17.357	< 2e-16	***
exper	28.792	1.629	17.673	< 2e-16	***
smsa	97.762	11.416	8.563	< 2e-16	***
fatheduc	4.350	1.822	2.387	0.01702	*

After including these controls, the coefficient on education is now *larger*

If anything, this suggests that our previous coefficient estimate might have been biased *down*. This can happen if educ is *negatively* correlated with the variables that were previously part of the error term

We're still worried that these controls don't completely capture the influence of important omitted variables

Let's try to use proximity to a two- or four-year college as an IV

The theory behind this is that whether you live near a college might not be correlated with factors that affect your wage (i.e., the IV might be exogenous)

But living near a college might make you more likely to attend (i.e., the IV might be relevant)

Before running the IV regression, let's check the first stage to make sure the instrument is actually relevant

```
1 summary(lm(educ ~ nearc2 + nearc4 + exper + smsa + fatheduc + motheduc, dat
```

Call:

```
lm(formula = educ ~ nearc2 + nearc4 + exper + smsa + fatheduc +  
    motheduc, data = card)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-6.7760	-1.3945	-0.0774	1.2437	5.9086

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	13.328239	0.214187	62.227	< 2e-16	***
nearc2	-0.003709	0.081903	-0.045	0.96389	
nearc4	0.234723	0.091983	2.552	0.01078	*
exper	-0.331019	0.010913	-30.334	< 2e-16	***
smsa	0.273393	0.096658	2.828	0.00472	**

The coefficient on `nearc4` is statistically significantly positive, which supports our theory that living near a four-year college increases education (`nearc2` doesn't appear to matter)

Now let's run the IV regression:

```
1 model.3 <- ivreg(wage ~ educ + exper + smsa + fatheduc + motheduc + smsa
2                   | nearc2 + nearc4 + exper + smsa + fatheduc + motheduc,
3                   data=card)
4 summary(model.3)
```

Call:

```
ivreg(formula = wage ~ educ + exper + smsa + fatheduc + motheduc +
      smsa | nearc2 + nearc4 + exper + smsa + fatheduc + motheduc,
      data = card)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-951.14	-187.60	-8.44	175.04	1545.57

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-1731.488	826.224	-2.096	0.03623	*
educ	141.214	61.561	2.294	0.02189	*
exper	60.223	20.452	2.945	0.00327	**
smsa	64.203	26.021	2.467	0.01360	*

The coefficient on education from the IV regression is 141

This implies that each additional year of education increases your wage by \$1.41

The IV estimate is much larger than the OLS estimate

There are two possible explanations for this:

1. The OLS estimate is negatively biased
2. The IV is endogenous too (i.e., living near a college is correlated with ε)

Given the large difference between the OLS and IV estimates, it's likely that the second explanation is at work here

Why might the IV be endogenous? I.e., why might living near a college be correlated with other factors that affect the wage?

One possibility is that people living near a four-year college are more likely to have parents with professional careers

These parents might pass some of their characteristics on to their kids, affecting their wages

Although our model controls for parental education, it's possible that this doesn't capture all of the factors that parents might pass on to their kids

This example illustrates a limitation of IV/2SLS: *it's hard to find good instrumental variables*

Despite this, IV remains an incredibly useful tool for addressing endogeneity

Unfortunately, there is no guarantee that you'll be able to find an IV that works for the specific question you're interested in