Simultaneous Equation Models (SEMs)

Introduction

- At the start of this course, we explained that the two key reasons why causality and correlation differ are: **confounding variables** and **reverse causality**.
- This reverse causality comes about when one of the regressors is **jointly determined** with the outcome variable.
- For example, when looking at how the number of police affects the crime rate, these two variables are jointly determined. There is a relationship going both ways.
- In fact, this is where the term **endogenous** comes from. Something which is endogenous (the opposite of exogenous) is determined within the model, it is not external to the model, so any manipulation of it does not have a simple one-directional effect on the 'outcome variable'.

Introduction

- Again, we turn to Instrumental Variables to solve our endogeneity issues (this was actually the first application of IV).
- We model reverse causality using a system of regressions. Typically it is just two (which we will stick to) but it could be much bigger.
- A typical model looks something like this:

$$\begin{aligned} \mathit{Crime} &= \beta_0 + \beta_1 \mathit{Police} + u_1 \\ \mathit{Police} &= \alpha_0 + \alpha_1 \mathit{Crime} + u_2. \end{aligned}$$

- In this example, we are probably interested in the first equation, we want to know what β_1 is. α_1 is decided through government policy, it is essentially a 'man-made' parameter.

Endogeneity

- The key thing is, we want each equation in the model (or each equation we are interested in) to have a ceteris parabis (causal) interpretation.
- If we naively estimate the first equation by OLS, and look at how Police affects Crime, our results aren't showing the causal effect.
- Our results show the following: we change Police, this changes Crime, but since Crime has changed, Police must also change according to the second equation. This feedback continues back and forth between the equations until we reach a new equilibrium. The final OLS estimate is based on this final outcome.
- In your head, run through the entire feedback loop for the Police/Crime example. Do you think our final estimate from OLS will be too low or too high relative to the real causal effect?

The Role of IV

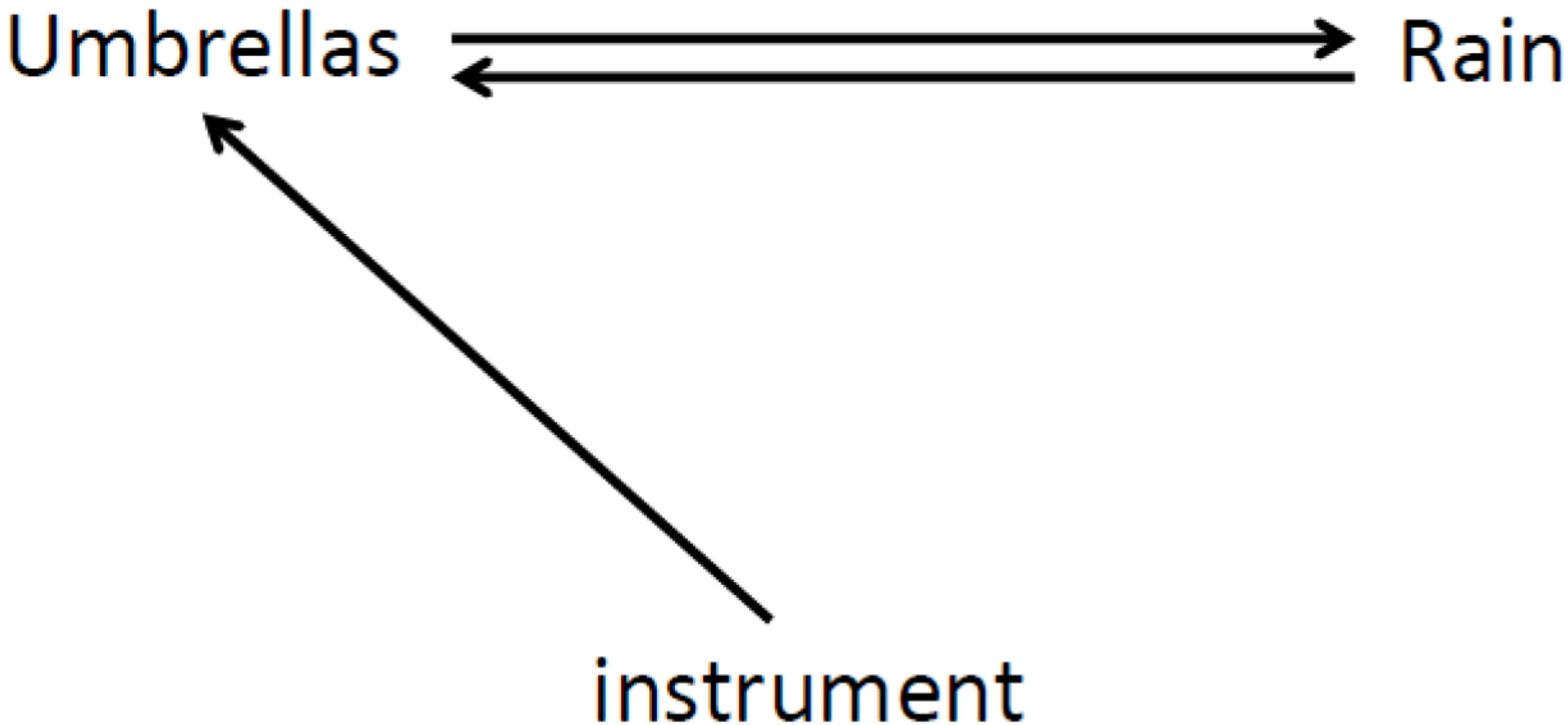
- Consider the Umbrella/Rain example:

$$\begin{aligned} \textit{Rain} &= \alpha_0 + \alpha_1 \textit{Umbrellas} + u_1 \\ \textit{Umbrellas} &= \beta_0 + \beta_1 \textit{Rain} + \beta_3 \textit{Price} + u_2. \end{aligned}$$

- We say that both Umbrellas and Rain are endogenous. An IV in an SEM is a variable which appears in one equation but not both. So the price of umbrellas can be used as an IV for Umbrellas in the first equation.

- Here, the instrument affects umbrellas but not rain directly. Recall, we want our IV to affect the outcome variable only through the endogenous variable. In the first equation, Price affects Rain only through Umbrellas. (Of course, in this silly example, Price doesn't affect Rain because Umbrellas don't affect Rain; $\alpha_1 = 0$).

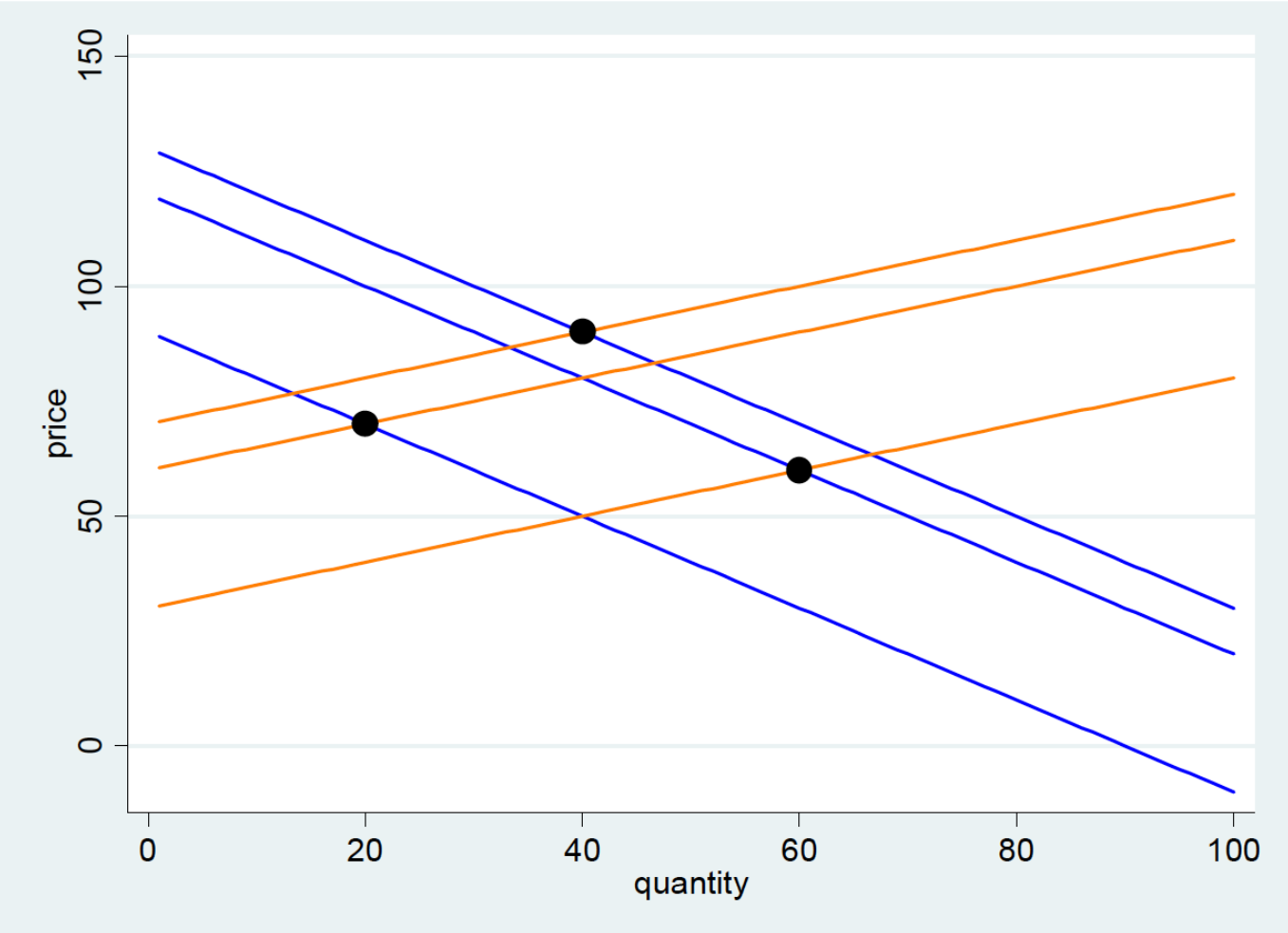
The Role of IV



Supply and Demand

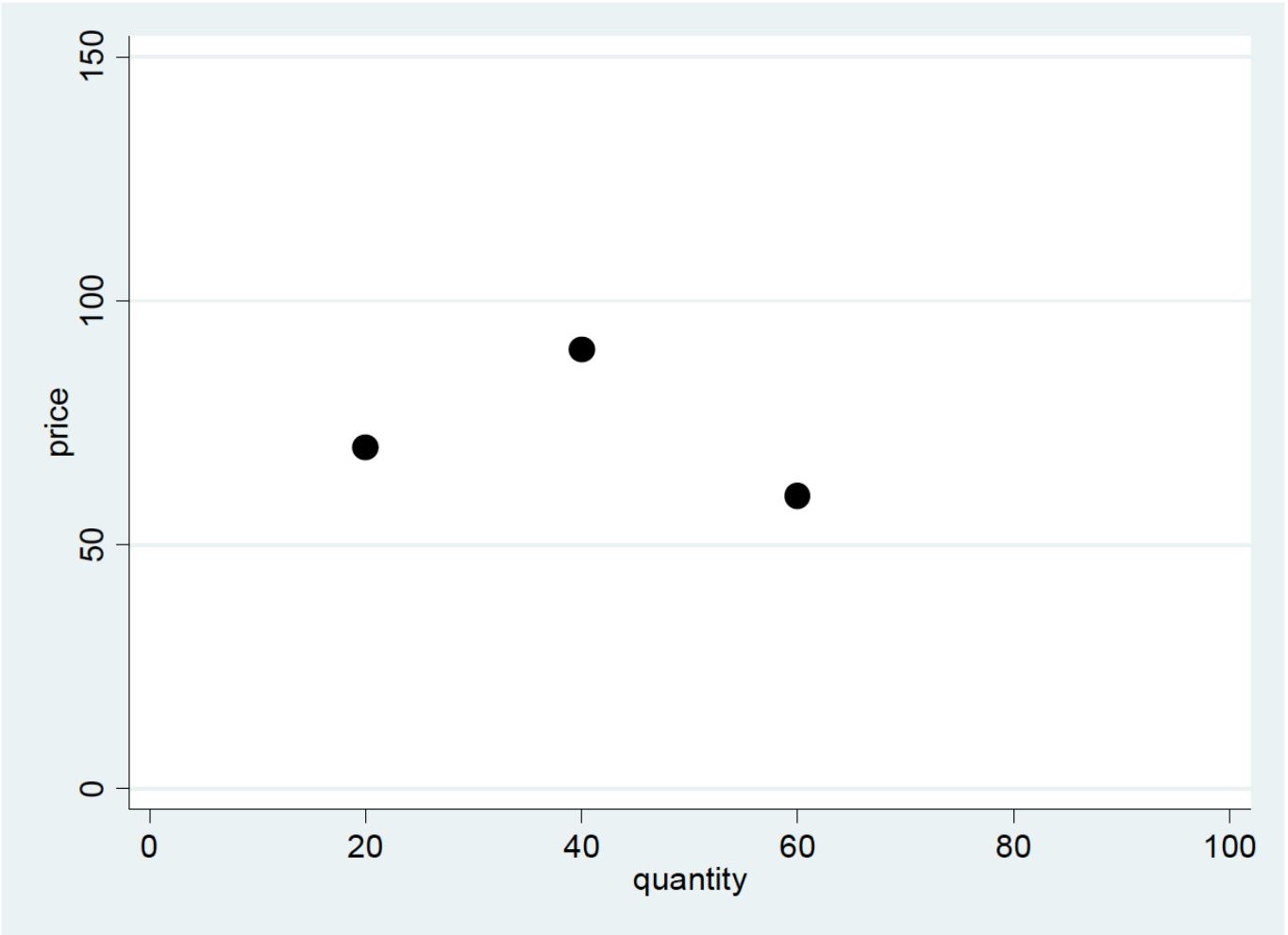
- Supply and demand are probably the most classic of all economic concepts, and dealing with their equilibrium is one of the oldest issues in economics.
- For this reason, the estimation of supply and demand curves are also the canonical example of simultaneous equation models in econometrics.
- Supply and demand interact to determine quantities (Q_i) and prices (P_i) simultaneously.

Supply and Demand in Equilibrium



(Credit: Steve Pischke)

Supply and Demand: the Data



(Credit: Steve Pischke)

Supply and Demand: the Equations

- Typically, we write the system of supply and demand as:

$$\begin{aligned} Q_i^S &= \alpha_S + \beta_S P_i + u_i^S \\ Q_i^D &= \alpha_D + \beta_D P_i + u_i^D \\ Q_i^S &= Q_i^D \end{aligned}$$

- With the final equation, we ensure equilibrium is reached. I then prefer to write the system as:

$$\begin{aligned} P_i &= \beta_0 + \beta_1 Q_i + u_i \\ Q_i &= \alpha_0 + \alpha_1 P_i + e_i \end{aligned}$$

- The first equation is decided by the seller (in effect, the supply equation). The second equation is decided by the buyer (the demand equation).

Example: Demand for Fish

- We will use a classic dataset to estimate a demand equation.
- The Foulton fish market is a large market for fresh fish in New York. Many dealers supply fish to many large buyers (fishmongers, restaurants, etc.), so there should be a good amount of competition.
- We have data from 1992 on daily sales and average daily prices from one dealer.
- For each day, we also have the weather conditions relevant for the daily catch. Our key instrument will be a dummy variable for stormy/mixed weather (as opposed to fair weather). The number of fish caught tends to be lower in bad weather.

Example: Demand for Fish - the Instrument

- Our equations become:

$$\begin{aligned} P_i &= \beta_0 + \beta_1 Q_i + \beta_2 BW_i + u_i \\ Q_i &= \alpha_0 + \alpha_1 P_i + e_i \end{aligned}$$

  where BW is the dummy for bad weather. These are known as the **structural equations** - they give the structure of the underlying model.

- So, why does BW work as an instrument? And what is it an instrument for?

- It is an instrument for P, and will act as an instrument in the second equation (i.e. where P is a regressor).

- We can see that it is correlated with P from the first equation. And since it does not appear in the second equation, it must be that the only way it affects Q is through P.

Example: Demand for Fish - the Instrument

- However, we need to be sure that our variable satisfies the conditions needed for it to be a good IV.
- It is straightforward to see if BW is correlated with P by simply regressing P on BW, this is known as the **first stage regression**. We want to see a significant p-value.
- We then need to argue that BW should not appear in the second equation in its own right.
- Could there be any way for bad weather to affect peoples' demand for fish (outside of its effect on price)?
- Possibly... it might be that people want to eat fish in nice weather. This would cause a problem for us. Maybe we could control for weather on the shore (rather than out at sea)?

Example: Demand for Fish - the Instrument

- Notice that we only have one instrument. There is no instrument for Q in the first equation. This means we **cannot** estimate the first equation.
- If we want to estimate the first equation we would need another variable in the second equation, which does not appear in the first.
- For example, if there was an advertising campaign on certain days. This would affect how much people want to eat fish, but does not directly affect the price (it affects price only through the quantity demanded)
- It's also perfectly fine to have several instruments for one endogenous regressor which can then be used to construct a 2SLS estimator.

Example: Demand for Fish - Endogeneity

- To see why we have an endogeneity problem, it is easiest to write out the **reduced form.**
- The reduced form is just each endogenous variable written as a function of only exogenous variables. P and Q are both endogenous and BW is the only exogenous variable.
- The reduced form is given by

$$\begin{aligned} P_i &= \left(\frac{\beta_0 + \beta_1\alpha_0}{1-\beta_1\alpha_1}\right) + \left(\frac{\beta_2}{1-\beta_1\alpha_1}\right) BW_i + \left(\frac{\beta_1 e_i + u_i}{1-\beta_1\alpha_1}\right) \\ Q_i &= \left(\alpha_0 + \frac{\alpha_1(\beta_0 + \beta_1\alpha_0)}{1-\beta_1\alpha_1}\right) + \left(\frac{\alpha_1\beta_2}{1-\beta_1\alpha_1}\right) BW_i \\ &\quad + \left(\alpha_1 e_i + \frac{\beta_1 e_i + u_i}{1-\beta_1\alpha_1}\right) \end{aligned}$$

- We can see that P is correlated with e, and Q is correlated with u. So P is endogenous in equation 2 and Q is endogenous in equation 1.

Example: Demand for Fish - in R

- The data is contained in 'fish'. We first run the first stage:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.912263   0.210556  -4.333 3.94e-05 ***
speed2      -0.002909   0.011297  -0.257 0.797416
wave2        0.093502   0.026095   3.583 0.000559 ***
speed3       0.001470   0.007594   0.194 0.846982
wave3        0.044131   0.025693   1.718 0.089421 .
mon         -0.014633   0.118362  -0.124 0.901891
tues        -0.013749   0.117643  -0.117 0.907232
wed          0.049052   0.113591   0.432 0.666937
thurs        0.123717   0.112266   1.102 0.273500
t           -0.001169   0.001407  -0.831 0.407999
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3529 on 87 degrees of freedom
Multiple R-squared:  0.3103,    Adjusted R-squared:  0.2389
F-statistic: 4.349 on 9 and 87 DF,  p-value: 0.0001064
```

- We use logs so we get price elasticities.

Example: Demand for Fish - in R

- We want to see if the weather variables are jointly significant:

```
15  # Test to see if the weather variables are jointly significant
16  LM2 = lm(data = data, lavgprc ~ mon + tues + wed + thurs + t)
17  summary(LM2)
18
19  waldtest(LM1, LM2, test = "F")
```

9:1 (Top Level)

Console ~/

```
>
> waldtest(LM1, LM2, test = "F")
Wald test

Model 1: lavgprc ~ speed2 + wave2 + speed3 + wave3 + mon + tues + wed +
    thurs + t
Model 2: lavgprc ~ mon + tues + wed + thurs + t
  Res.Df Df      F    Pr(>F)
1     87
2     91 -4 7.0891 5.473e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- We look good here :)

Example: Demand for Fish - in R

- Now we can run the 2SLS regression:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.27091    0.20978  39.426  < 2e-16 ***
lavgprc     -0.96231    0.38166  -2.521  0.01345 *
mon         -0.32171    0.23352  -1.378  0.17173
tues        -0.68727    0.22983  -2.990  0.00359 **
wed         -0.51973    0.22734  -2.286  0.02460 *
thurs        0.10571    0.22946   0.461  0.64615
t           -0.00290    0.00302  -0.960  0.33945
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7173 on 90 degrees of freedom
Multiple R-Squared: 0.175,      Adjusted R-squared:  0.12
Wald test: 3.635 on 6 and 90 DF,  p-value: 0.002822
```

- Does -0.96 sound reasonable for the price elasticity of demand?

When is an SEM a Good Idea?

- The supply/demand example and the police/crime example share some similar characteristics.
- Both equations (in each example) had an interpretation of their own, independent of the other equation. And each equation was governed by a different person (or entity).
- The supply equation is determined by shopkeepers, etc., the demand equation is determined by buyers.
- The Police/Crime equation is determined by the government, the Crime/Police equation is determined by criminals.
- Each equation should tell its own individual story.

When is an SEM NOT a Good Idea?

- Consider the following:

$$\begin{aligned} \textit{HouseExp} &= \alpha_0 + \alpha_1 \textit{Saving} + u_1 \\ \textit{Saving} &= \beta_0 + \beta_1 \textit{HouseExp} + u_2. \end{aligned}$$

- This may look like a sensible SEM but it's not.
- Both equations are determined by the same household, and neither has a sensible causal interpretation on its own.
- It would be weird to ask: if savings were exogenously increased, what would happen to housing expenditure?
- An easy way to spot a bad use of SEM is if a single entity is deciding some tradeoff between two variables (e.g. the sleep and work regression we saw in a previous problem set). Ask yourself: How many decisions are being made? If it's just one, OLS works fine.

Example: Inflation and Openness

- Romer (1993) proposes a theoretical model which implies that more 'open' countries should have lower inflation. He explains inflation as a function of the proportion of GDP made up by imports (which is his measure of 'openness').
- His model looks like this:

$$\begin{aligned} inf &= \beta_0 + \beta_1 open + \beta_2 ln(pcinc) + u_1 \\ open &= \alpha_0 + \alpha_1 inf + \alpha_2 ln(pcinc) + \alpha_3 log(land) + u_2 \end{aligned}$$

  where *pcinc* is per capita income (assumed exogenous), and *land* is the land area of the country.

- The top equation is the one of interest, and we can estimate it if *land* is deemed a suitable IV.
- The idea is that smaller countries are likely to be more open, but the size of the country does not impact inflation directly.

Example: Inflation and Openness

- The data is contained in 'openness'. First stage:

```
Call:
lm(formula = open ~ lpcinc + lland, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-31.907  -8.843  -3.109   6.057  82.792

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 117.0845    15.8483   7.388 2.97e-11 ***
lpcinc        0.5465     1.4932   0.366    0.715
lland        -7.5671     0.8142  -9.294 1.51e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17.8 on 111 degrees of freedom
Multiple R-squared:  0.4487,    Adjusted R-squared:  0.4387
F-statistic: 45.17 on 2 and 111 DF,  p-value: 4.451e-15
```

- Very significant... Well done Romer!

Example: Inflation and Openness

- 2SLS results:

```
Call:
ivreg(formula = inf ~ lpcinc + open | lpcinc + lland, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-21.686 -10.176  -5.857   2.912 184.875

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  26.8993    15.4012   1.747   0.0835 .
lpcinc        0.3758     2.0151   0.187   0.8524
open         -0.3375     0.1441  -2.342   0.0210 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 23.84 on 111 degrees of freedom
Multiple R-Squared: 0.03088,    Adjusted R-squared: 0.01341
Wald test:  2.79 on 2 and 111 DF,  p-value: 0.06574
```

- *open* has a significant negative effect. For every percentage point increase in the import share of GDP, annual inflation is a third of a percentage point lower.

Summary

- We have seen that we have an endogeneity issue when two variables are jointly determined.
- We have seen how to set up a system of simultaneous equations.
- And we have looked at what we require from our instrument in order to estimate one of the equations
- We have also briefly discussed when an SEM is appropriate and when it is not.