**PLSC 503: "Multivariate Analysis for Political Research"**

**Simultaneity and Endogeneity**
March 22, 2017

# Random Regressors → Endogeneity/Simultaneity

In general, relaxation of the assumption that $\mathbf{X}$ is fixed in repeated samples doesn't have any particularly ill effects on our regression estimates. So long as $\mathbf{X}$ remains independent of $\mathbf{u}$, the estimates of $\boldsymbol{\beta}$ we get from a model in which the $\mathbf{X}$s are unbiased, least-variance, etc.

*Simultaneity* and *endogeneity* go to two of our OLS assumptions: That the $\mathbf{X}$s are fixed in repeated sampling, and that they are uncorrelated with the errors $\mathbf{u}$. As we noted above, relaxing the first of these is a relatively innocuous thing to do; the second, however, raises a whole host of other issues.

## Endogeneity and Simultaneity: Basics

At the outset, let's remember that *endogeneity* and *simultaneity* are essentially the same thing...

- **Endogeneity** is a theoretical/conceptual term referring to the idea that some element or elements of $\mathbf{X}$ "depend on" $\mathbf{Y}$, in a causal sense; the terms "reciprocity" and "nonrecursivity" mean more-or-less the same thing too.
- Likewise, **simultaneity** is a statistical term referring to a two-way relationship between two variables.

Note that simultaneity can arise for a number of reasons, but that we can generally group them into a few categories:

1. Endogeneity can result when *past values of a variable influence present values* – for example, in a time series, when $Y_{t-1}$ influences the value of $Y_t$. This is a relatively common thing, and something we'll discuss at some length a bit later in the course.
2. It can also be the result of *temporal aggregation*. For example, suppose (as we will a bit later) we were estimating a model of the influence of international trade flows on bilateral armed conflict between states. Those data are typically measured annually, but the dynamics of war and trade happen on a scale of hours/days/weeks, rather than years: A conflict in February may reduce trade for the next three months, which in turn could lead to a worsening of hostilities as economic ties between the two countries weaken. Note that here the bi-directional causality is not instantaneous, but happens over time; the simultaneity arises because our data (usually) aren't measured on a fine-grained enough time scale to permit disentangling the effects.

3. Finally, as the classical microeconomics applications tell us, endogeneity can be the result of *rational anticipation* on the part of some actors. In the archetypical supply-demand-price models, for example, prices adjust to supply and demand (and supply adjusts to demand and price, and demand to supply and price) in part because producers and consumers anticipate what the other will do, and adjust their behavior accordingly.

As a practical issue, which of these things is going on can matter a lot; for example, if the endogeneity is due to temporal aggregation, one possible way to fix it is just to collect data at a finer-grained time scale.

## Some Math

The customary way to think about endogeneity is by thinking of a *system of equations*. To make things simple, we'll start with a system of two equations:

$$Y_1 = \mathbf{X}_1\boldsymbol{\beta}_1 + \gamma_1 Y_2 + \mathbf{u}_1 \tag{1}$$

$$Y_2 = \mathbf{X}_2\boldsymbol{\beta}_2 + \gamma_2 Y_1 + \mathbf{u}_2 \tag{2}$$

Equations (1) and (2) are known as the structural equations for $Y_1$ and $Y_2$. Here, the $\mathbf{X}$ variables are referred to as the *exogenous* variables; they may or may not be the same across the two equations. The more important thing here is that $Y_1$ depends on $Y_2$, but the reverse is also true; they are the *endogenous* variables.

Now, notice what happens if we "unpack" the variation in Equation (1) by substituting (2) in for $Y_2$ in (1):

$$\begin{aligned}
Y_1 &= \mathbf{X}_1\boldsymbol{\beta}_1 + \gamma_1[\mathbf{X}_2\boldsymbol{\beta}_2 + \gamma_2 Y_1 + \mathbf{u}_2] + \mathbf{u}_1 \\
&= \mathbf{X}_1\boldsymbol{\beta}_1 + \gamma_1(\mathbf{X}_2\boldsymbol{\beta}_2) + \gamma_1\gamma_2 Y_1 + \gamma_1\mathbf{u}_2 + \mathbf{u}_1 \\
Y_1 - \gamma_1\gamma_2 Y_1 &= \mathbf{X}_1\boldsymbol{\beta}_1 + \gamma_1(\mathbf{X}_2\boldsymbol{\beta}_2) + \gamma_1\mathbf{u}_2 + \mathbf{u}_1 \\
(1 - \gamma_1\gamma_2)Y_1 &= \mathbf{X}_1\boldsymbol{\beta}_1 + \gamma_1(\mathbf{X}_2\boldsymbol{\beta}_2) + \gamma_1\mathbf{u}_2 + \mathbf{u}_1 \\
Y_1 &= \mathbf{X}_1\left(\frac{1}{1-\gamma_1\gamma_2}\boldsymbol{\beta}_1\right) + \mathbf{X}_2\left(\frac{\gamma_1}{1-\gamma_1\gamma_2}\boldsymbol{\beta}_2\right) + \left(\frac{\gamma_1\mathbf{u}_2 + \mathbf{u}_1}{1-\gamma_1\gamma_2}\right)
\end{aligned} \tag{3}$$

where it is sometimes convenient to write

$$Y_1 = \Delta_1\mathbf{X}_1 + \Delta_2\mathbf{X}_2 + \mathbf{e} \tag{4}$$

where $\Delta_1 = \frac{\boldsymbol{\beta}_1}{1-\gamma_1\gamma_2}$, $\Delta_2 = \frac{\gamma_1\boldsymbol{\beta}_2}{1-\gamma_1\gamma_2}$, and $\mathbf{e} = \frac{\gamma_1\mathbf{u}_2+\mathbf{u}_1}{1-\gamma_1\gamma_2}$ is a composite error term.

Equation (3) is known as the *reduced form* of Equation (1); we could write a similar reduced form equation for Eq. (2). In a nutshell, it is the rearrangement of (1) so that all of the

endogenous variables are on the left-hand side, and all the exogenous ones on the right-hand side.

The reduced form equations can be very useful for a number of things. In particular, they are valuable for calculating *comparative statics* – the overall level of responsiveness in $Y_1$ and $Y_2$ to changes in the exogenous factors. One can usually do this by taking first derivatives of the reduced-form model with respect to the variable of interest. So, for example, Equation (3) provides a very direct indication of the overall marginal change in $Y_1$ associated with a one-unit change in a particular element $X_\ell$ in $\mathbf{X}_1$:

$$\frac{\partial Y_1}{\partial X_\ell} = \frac{\beta_\ell}{1 - \gamma_1\gamma_2} \tag{5}$$

This would be a very hard thing to figure out directly from (1), because, in that equation, we can't disentangle the "direct" effect of $X_\ell$ on $Y_1$ from its "indirect" effect (that is, the effect of $X_\ell \rightarrow Y_1 \rightarrow Y_2 \rightarrow Y_1$).

The problem with the reduced-form equation in (3) is that the parameters estimated – call them $\hat{\Delta}_1$ and $\hat{\Delta}_2$ – are usually not the parameters of interest. Those belong to the structural models in (1) and (2), and (as we'll see below) we can't get "good" estimates of those directly using OLS. It is possible to derive unique estimates of the $\boldsymbol{\beta}$s and $\gamma$s from the $\hat{\Delta}$s – a method called "indirect least squares" – but that is not usually how systems of equations are estimated.

**Simultaneity Bias**

The endogeneity in (1) and (2) is a problem because it can be shown that estimating either of these models in isolation – say, by estimating (1) by OLS and treating $Y_2$ as exogenous when it really is not – leads to correlation between the error term and the endogenous covariate. Intuitively, this is because variation ("shocks") in the error term for either equation lead to variation in *both* $Y_1$ and $Y_2$.

As an example, suppose that, for a particular observation $i$, some purely random thing leads to a big error for that observation $u_{1i}$. Now:

- a "big" value of $u_{1i}$ will lead to a "big" value of $Y_{1i}$ as well. However,
- according to Equation (2) (and assuming, for simplicity, that $\gamma_2 > 0$), the resulting "big" value of $Y_{1i}$ also leads directly to a "big" value of $Y_{2i}$. That's a problem, because
- by Equation (1) (and again assuming $\gamma_1 > 0$), a "big" value of $Y_{2i}$ will lead to a "big" value of $Y_{1i}$.
- In other words, the effect of $Y_2$ on $Y_1$ is confounded with the error term, and cannot be estimated accurately. Moreover,

- This problem is not one that will go away as $N \to \infty$; we are stuck with a biased OLS estimator.

Statistically, the key condition for unbiasedness in an OLS estimator is that $\mathrm{E}(\mathbf{X}, \mathbf{u}) = 0$. In the context of (1), it can be shown that this is not the case with respect to the regressor $Y_2$; in fact, with endogeneity of the sort described above,

$$\mathrm{E}(Y_2, \mathbf{u}) = \frac{\gamma_2}{1 - \gamma_1\gamma_2}\sigma^2_{\mathbf{u}}$$

and that, as a result, the OLS estimate of $\gamma_1$ will necessarily be biased, and inconsistent. Notice, however, that the correlation between the regressor and the error goes away if $\gamma_2 = 0$ – that is, if $Y_2$ doesn't depend on $Y_1$. In that case, $Y_2$ is just like any other (exogenous) regressor, and OLS works just fine.

## What to Do?

There is a lot written on the topic of simultaneity – far, far too much for me to cover it in a single class session, or (arguably) even in a single class. Part of the reason for – and the result of – all that work on the subject is that there are lots of different ways one can approach simultaneity. Kennedy has a nice overview of them; we'll talk about a few of the simpler ones, and then work through an example.

### OLS

One alternative – frankly, it's always an alternative – is just to ignore the simultaneity and use OLS. A few things bear mentioning about such an approach:

- This is by far the most common thing people actually do. As a practical matter, nearly all analysis of observational data has at least the potential for endogeneity; and, most of the time, analysts go about running their regressions as if nothing were the matter. Often, this is because there are other, potentially more serious problems that also need addressing (perhaps autocorrelation, or something similar) that are hard to deal with using a non-OLS approach.

- While it's hard to be too concrete, the extent to which OLS will lead to simultaneity bias is generally a function of the degree of the simultaneity. That in turn means that if your simultaneity isn't very serious – that is, your $Y_2$ doesn't depend on $Y_1$ very strongly – then the degree of bias will be small, and OLS won't be a bad choice. (Remember: OLS, even when inconsistent, is still a minimum-variance estimator; that means that in such situations it may still be the "best" estimator according to a mean-squared-error criterion).

- Relatedly, OLS is quite robust – it is generally less sensitive to violations of its canonical assumptions than are other, more specialized estimators. That means that, if you have

multiple "problems" (say, endogeneity plus autocorrelation plus multicollinearity) OLS might win out over other alternatives by virtue of the fact that it is "not too bad" on any of those three dimensions.

### Lagged Variables

If your data have a temporal component (i.e., are repeated measurements over time on the same observations), then you can take advantage of that fact to remedy potential endogeneity. The simplest way of doing this is to consider using lagged values of (possibly endogenous) covariates; such a model might look like:

$$Y_{it} = \mathbf{X}_{it}\boldsymbol{\beta}_1 + \gamma_1 Y_{i,t-1} + u_{it} \tag{6}$$

Models like this can take on two forms. If data are a single series, they are *time-series* data; we'll talk (a lot) more about such data in the next few weeks. Data that have variation both across units and over time are called *panel* data (or *time-series cross-sectional* data). In addition to dealing with endogeneity, panel/TSCS data can be valuable for a number of other reasons – take my longitudinal data analysis course to find out more.

### Instrumental Variables / 2SLS

The standard way of dealing with endogeneity – and, frankly, any problem in which one or more covariates are correlated with the disturbances – is through the use of *instrumental variables*.[1] The intuition of IV estimation is to find a variable or set of variables (called "instruments," and nearly always denoted $\mathbf{Z}$) that are simultaneously highly correlated with the endogenous $\mathbf{X}$, but uncorrelated with the disturbance term $\mathbf{u}$.

Mathematically, consider

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}. \tag{7}$$

Recall that the OLS estimator for (7) converges to

$$\hat{\boldsymbol{\beta}}_{OLS} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u} \tag{8}$$

where the second term goes to zero iff $\text{Cov}(\mathbf{X}, \mathbf{u}) = \mathbf{0}$. If $\text{Cov}(\mathbf{X}, \mathbf{u}) \neq \mathbf{0}$, $\hat{\boldsymbol{\beta}}_{OLS}$ is biased. However, if we have instrument(s) $\mathbf{Z}$ such that $\text{Cov}(\mathbf{Z}, \mathbf{u}) = \mathbf{0}$, we can use those instruments

[1] Full disclosure: For years, when I taught this class, I didn't even mention IV estimation. Teaching it to political scientists always seemed to me to be akin to giving a loaded pistol to a three-year-old to protect him/her self: it was almost certainly more likely to lead to bad outcomes than good. I still have that opinion, more or less, but I've decided to teach it anyway to prevent you all from having any major holes in your understanding of "econometrics."

to construct a version of $\mathbf{X}$ that is uncorrelated with $\mathbf{u}$ (and will therefore give us unbiased estimates of $\boldsymbol{\beta}$). The new estimator (call it $\hat{\boldsymbol{\beta}}_{IV}$) is:

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{IV} &= (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{Y} \\ &= (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'(\mathbf{X}\boldsymbol{\beta} + \mathbf{u}) \\ &= \boldsymbol{\beta} + (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{u} \end{aligned} \tag{9}$$

Since (by assumption or construction) $\text{Cov}(\mathbf{Z}, \mathbf{u}) = \mathbf{0}$, this new estimator will be consistent. Of course, how "good" an estimator it will be depends critically on the covariation between $\mathbf{X}$ and $\mathbf{Z}$; we'll talk more about this in a bit.

IV estimation encompasses a number of different approaches, and can be used (at least, in principle) any time that we have a problem with our disturbances being correlated with our covariates. A widely-used variant of IV estimation used specifically to deal with endogeneity is "two-stage least squares" (abbreviated 2SLS). In the context of simultaneous equations, the intuition of 2SLS is to

1. regress any endogenous $\mathbf{X}$ variables on the full set of exogenous covariates (both those in $\mathbf{Z}$ and any exogenous variables in $\mathbf{X}$) to get a set of predicted values for the endogenous $\mathbf{X}$s (say, $\hat{\mathbf{X}}$), and then
2. regress $Y$ on $\hat{\mathbf{X}}$ to get our 2SLS estimates of $\boldsymbol{\beta}$.

So, if we consider our original models in (1)-(2), where we had two endogenous variables:

$$Y_1 = \mathbf{X}_1\boldsymbol{\beta}_1 + \gamma_1 Y_2 + \mathbf{u}_1$$

$$Y_2 = \mathbf{X}_2\boldsymbol{\beta}_2 + \gamma_2 Y_1 + \mathbf{u}_2$$

then the 2SLS approach is to estimate:

$$Y_2 = \mathbf{X}_1\Theta_1 + \mathbf{X}_2\Theta_2 + \epsilon \tag{10}$$

in the "first stage," then use our estimates from that model to generate predicted values:

$$\hat{Y}_2 = \mathbf{X}_1\hat{\Theta}_1 + \mathbf{X}_2\hat{\Theta}_2.$$

By construction, these estimated values of $\hat{Y}_2$ are uncorrelated with the error term in the original equation $\mathbf{u}_1$ (intuitively, this is because they use information from the exogenous variables, which we also assume to be uncorrelated with $\mathbf{u}_1$, and because estimated $u$s in a regression equation are uncorrelated with the regressors). Moreover, the estimates $\hat{Y}_2$ optimally combine information on the exogenous variables to produce the "best" possible estimate of $\mathbf{Y}_2$ given the data.

Once we've done this, our second-stage model is just the structural model for the variable of interest, but with the endogenous covariate(s) replaced with its instrument:

$$Y_1 = \mathbf{X}_1\boldsymbol{\beta}_1 + \gamma_1\hat{Y}_2 + \mathbf{v}_1 \tag{11}$$

This will now give us an (asymptotically) unbiased estimate of $\gamma_1$. Note a few things about 2SLS:

- If the number of instruments is the same as the number of variables being instrumented (say, we have only two endogenous variables in two equations) – that is, when the system is *exactly identified* – then 2SLS gives the same results as if we had estimated the two associated reduced-form equations.
- When the system is *overidentified* – that is, when we have additional exogenous variables with which to identify the system – 2SLS optimally uses the information in those regressors.
- The reported $R^2$ from a 2SLS is typically an indication of the fit of the *instruments* to $Y$, *not* an estimate of the fit of the original model to the data; that is, it is based on $\hat{\mathbf{v}}$ rather than $\hat{\mathbf{u}}$. The fit of the second-stage model is often lousy, in part because the imperfect correlation between the endogenous variable and the instruments introduces random error into the equation. To get a "real" $R^2$, we need to replace the actual (data) values of the covariates $\mathbf{X}$ and $\mathbf{Y}_2$ into (11):

  $$\hat{\mathbf{u}} = Y_1 - \mathbf{X}_1\hat{\boldsymbol{\beta}}_1 + \hat{\gamma}_1\hat{Y}_2$$

  and use those residuals as a basis for the calculation of the RSS and $R^2$. (Note that most software packages will do this for you).
- Similarly 2SLS also requires that one correct one's standard error estimates too. While it's a bit more than I care to go into here, suffice it to say that – as with $R^2$ – we want to use $\hat{\mathbf{u}}$ rather than $\hat{\mathbf{v}}$ when calculating $\hat{\sigma}^2$. Again, most software packages that estimate 2SLS models do this more-or-less automatically.

I should note that 2SLS was a lot more widely used in political science, say, 30 or 40 years ago than it is now (and it's now making something of a "comeback," thanks to the influence of the causal inference mafia). That is in part due to fashion, in part because people came to realize it has some drawbacks; more on those later...

### Systems Methods

In addition to multi-stage approaches, in which we estimate a series of single equations in sequence, it is also possible to estimate the entire system of equations (e.g., both structural equations in (1) and (2)) using a systems-of-equations approach. Doing uses *all* the information in the system to estimate *all* the parameters thereof, and so has the advantage of being

more asymptotically efficient than the single-stage methods. There are two broad methods in this category:

- *Three-stage least squares* is an analogue to 2SLS; it essentially consists of:
  1. Estimating the equations for all relevant endogenous variables using 2SLS,
  2. Using those parameter estimates to generate estimates of the disturbances **u** from the structural equation(s),
  3. Use those estimates to get consistent estimates of the cross-equation variance-covariance matrix of the parameters, and then
  4. Using GLS to reestimate the system of equations.
- *Likelihood-based methods* specify the entire conditional distribution of the data, and then estimate the parameters by iteratively maximizing some objective function (which might be a full-information likelihood, a partial likelihood, etc.).

## Simultaneity and 2SLS: A Simple Simulation

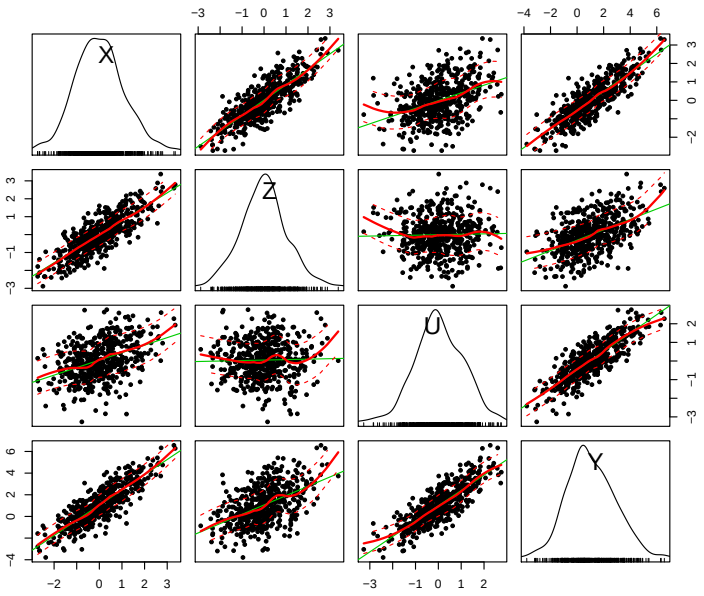
It's easy to demonstrate how IV can correct for simultaneity / specification error bias. Consider a "best case" scenario, where we have a covariate $X$ that is moderately correlated with the residuals $u$ (say, $r = 0.4$), and an instrument $Z$ that is strongly correlated with $X$ ($r = 0.8$) and completely uncorrelated with $u$. We can simulate this using `mvrnorm` in the `MASS` package:

```
library(MASS)
library(sem)
library(car)

seed<-1337
set.seed(seed)

mu<-c(0,0,0) # <== X, Z, U
Sigma<-matrix(c(1,0.8,0.4,0.8,1,0,0.4,0,1),nrow=3,byrow=TRUE) # Cor(X,Y)=0.8, etc.
Vars<- mvrnorm(500,mu,Sigma)
colnames(Vars)<-c("X","Z","U")
Vars<-data.frame(Vars)
```

Now, generate $Y$ such that it varies with $X$ and $u$, setting $\beta_0 = \beta_1 = 1$:

```
Vars$Y<- 1 + Vars$X + Vars$U

scatterplotMatrix(Vars)
```

Figure 1: Scatterplot Matrix of IV Simulation




OLS regression of $Y$ on $X$ yields a biased estimate of $\beta_1$:

```
> OLS<- lm(Y~X,data=Vars)
> summary(OLS)

Call:
lm(formula = Y ~ X, data = Vars)

Residuals:
    Min      1Q  Median      3Q     Max
-3.3809 -0.6058 -0.0102  0.6320  2.9470

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.04770    0.04209   24.89   <2e-16 ***
X            1.40254    0.04005   35.02   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.9413 on 498 degrees of freedom
Multiple R-squared:  0.7112,Adjusted R-squared:  0.7106
F-statistic:  1226 on 1 and 498 DF,  p-value: < 2.2e-16
```

By contrast, 2SLS "fixes" the estimate:

```
> TSLS<-tsls(Y~I(X),data=Vars,instruments=~Z)
> summary(TSLS)

 2SLS Estimates

Model Formula: Y ~ I(X)

Instruments: ~Z

Residuals:
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-3.29300 -0.68210 -0.06139  0.00000  0.76270  2.70300

             Estimate Std. Error  t value   Pr(>|t|)
(Intercept) 1.0491828  0.0456017 23.00754 < 2.22e-16 ***
I(X)        1.0302012  0.0536909 19.18763 < 2.22e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.0196738 on 498 degrees of freedom
```

Of course, this is a single simulation; if you wanted to demonstrate that IV estimation "works" more generally, you would loop over many such simulations.

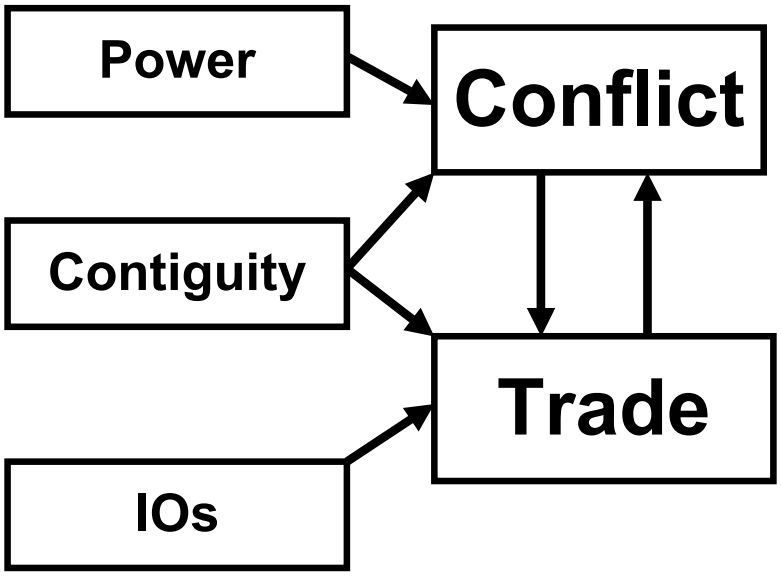
## Simultaneity and 2SLS: A Little Real-Data Example

For our "real-data" example, we'll examine the relationship between interstate wars and trade, a topic more than a few people have spent a good bit of time on. Assume for the moment that we've posited a simple model of war and trade. In this model, both war and trade are endogenous, while IOs, contiguity, and power (here, military capabilities) are exogenous. Moreover, we have some theories that suggest that power imbalances don't (directly) affect trade, and that IOs don't (directly) affect conflict.[2] The theory is illustrated schematically in Figure 2:

The data we'll use to examine this system are measured at the dyad level, on all "politically-relevant" dyads, and are aggregated over the 1950-1985 period; they are a stripped-down version of the data used in Oneal, Russett, and Davis's 1998 *International Organization* article. They consist of five variables:

[2]The Neo-Kanto-Russettonian liberals among you are undoubtedly seething at this crazy characterization of the relationships among these variables. For now, deal with it: I'm doing this for pedagogical purposes, not to be the next big thing in international relations.

Figure 2: Trade and War: A Simple System




- `logdisputes` is the log of the number of disputes between the two countries in the dyad during the 1950-1985 period,
- `logtrade` is the log of the mean level of trade dependence between the two countries in the dyad during the 1950-1985 period,
- `capratio` is the mean capability ratio (calculated according to the COW definition) between the two countries in the dyad during the 1950-1985 period,
- `contiguity` is an indicator of whether (=1) or not (=0) the two countries are contiguous, and
- `IOs` is the average number of joint international organization memberships the two countries in the dyad shared during the 1950-1985 period.

The data look like this:

```
> summary(IRData)
     dyadid        logdisputes         logtrade            IOs
 Min.   :  2020   Min.   :-0.6931   Min.   :-0.6931   Min.   : 4.579
 1st Qu.:135155   1st Qu.:-0.6931   1st Qu.: 2.4079   1st Qu.:19.500
 Median :220484   Median :-0.6931   Median : 5.5786   Median :27.704
 Mean   :275526   Mean   :-0.2627   Mean   : 4.6518   Mean   :30.891
 3rd Qu.:385710   3rd Qu.: 0.0000   3rd Qu.: 7.1248   3rd Qu.:39.289
 Max.   :900920   Max.   : 3.4965   Max.   :11.5037   Max.   :93.700
   contiguity        capratio          GDPgrowth
 Min.   :0.0000   Min.   :   1.081   Min.   :-9.0800
 1st Qu.:0.0000   1st Qu.:   4.849   1st Qu.:-0.2923
 Median :0.0000   Median :  26.577   Median : 0.8363
 Mean   :0.3207   Mean   : 196.310   Mean   : 0.5097
 3rd Qu.:1.0000   3rd Qu.: 144.035   3rd Qu.: 1.7106
 Max.   :1.0000   Max.   :7451.982   Max.   : 7.0460
```

### OLS vs. 2SLS in a Texas Cage Match

Suppose at the outset that we're interested in the disputes variable. A natural place to start would be to estimate an OLS regression of the structural model for disputes implied by our "theory:"

```
> OLSWar<-lm(logdisputes~logtrade+contiguity+capratio)
> summary(OLSWar)

Residuals:
     Min       1Q   Median       3Q      Max
-0.82840 -0.32644 -0.26860 -0.08972  3.45504

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.253e-01  6.020e-02  -7.065 3.46e-12 ***
logtrade     8.558e-03  1.057e-02   0.809   0.4185
contiguity   4.623e-01  7.124e-02   6.489 1.50e-10 ***
capratio    -1.296e-04  6.467e-05  -2.003   0.0455 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.853 on 813 degrees of freedom
Multiple R-squared: 0.08301,Adjusted R-squared: 0.07962
F-statistic: 24.53 on 3 and 813 DF,  p-value: 3.345e-15
```

If we believe these results, we'd likely say that:

1. contiguous countries fight more often,
2. countries with greater power imbalances between them fight less often, and
3. trade has no effect, one way or the other, on international conflict (the point estimate is even positive, but is very imprecisely estimated).

Both (1) and (2) are consistent with theory and expectations; depending on who you talk to, however, (3) is not. After hearing today's lecture, though, we now understand that, if `logtrade` is, in fact endogenous (as our theory implies), it is likely that there is bias in (at a minimum) our estimate of its effects on `logdisputes`.

To attempt to correct for that, we can estimate a 2SLS model that "instruments" the `logtrade` variable. There are a number of R packages and routines that will estimate systems of endogenous equations via 2SLS (and, for that matter, by other means as well). Probably the most straightforward to use is the `tsls` command, which is part of the `sem` package (for estimating structural equation models):

```
> library(sem)
> TwoSLSWar<-tsls(logdisputes~contiguity+capratio+I(logtrade),
   instruments=~contiguity+capratio+IOs)
> summary(TwoSLSWar)

2SLS Estimates

Model Formula: logdisputes ~ contiguity + capratio + I(logtrade)

Instruments: ~contiguity + capratio + IOs

Residuals:
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-1.21e+00 -5.24e-01 -2.26e-01 -7.44e-17 -2.10e-02  3.65e+00

              Estimate Std. Error t value  Pr(>|t|)
(Intercept) -0.1515180  8.562e-02  -1.770 7.717e-02
contiguity   0.6263774  8.111e-02   7.722 3.353e-14
capratio    -0.0002664  7.252e-05  -3.674 2.543e-04
I(logtrade) -0.0558374  1.769e-02  -3.157 1.652e-03

Residual standard error: 0.8723 on 813 degrees of freedom
```

These results are quite a bit different. Note, for example, that:

- The effects of both `contiguity` and `capratio` – while retaining in the (expected) signs from the OLS model – are both significantly larger and not particularly less precisely estimated than in the OLS model.
- The effect of `logtrade` is now large, negative (as most neoliberal theory would suggest), and statistically distinguishable from zero with a high degree of confidence.

Also:

- `tsls` automatically includes the "other" exogenous variable (`capratio`) in the first-stage equation for `logtrade`. It does so because not to do so would cause biases in the instrument, which would in turn lead to a biased and inconsistent estimator in the second stage.
- R does not report an $R^2$ from the 2SLS model. In fact, the RMSE is a better indicator of model fit, and suggests that – at least in terms of prediction – the 2SLS model is doing more-or-less as well at prediction as is the OLS estimator (which is not so terrible).

We can see what's going on "inside" the `tsls` routine by reestimating the 2SLS model "by hand":

```
> ITrade<-lm(logtrade~contiguity+IOs+capratio)
> summary(ITrade)

Residuals:
    Min      1Q  Median      3Q     Max
-6.0385 -1.7666  0.4139  1.6154  7.6029

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.7319793  0.1912570   3.827 0.000140 ***
contiguity   1.3386037  0.1816041   7.371 4.17e-13 ***
IOs          0.1218373  0.0055313  22.027  < 2e-16 ***
capratio    -0.0013913  0.0001626  -8.555  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.239 on 813 degrees of freedom
Multiple R-squared: 0.5535,Adjusted R-squared: 0.5519
F-statistic: 335.9 on 3 and 813 DF,  p-value: < 2.2e-16

> IVWarByHand<-lm(logdisputes~capratio+contiguity+(ITrade$fitted.values))
> summary(IVWarByHand)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-1.0055 -0.3618 -0.2782 -0.0492  3.5301

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)          -1.515e-01  8.323e-02  -1.821 0.069050 .
capratio             -2.664e-04  7.049e-05  -3.780 0.000168 ***
contiguity            6.264e-01  7.884e-02   7.944 6.49e-15 ***
ITrade$fitted.values -5.584e-02  1.719e-02  -3.248 0.001210 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.8479 on 813 degrees of freedom
Multiple R-squared: 0.09402,Adjusted R-squared: 0.09068
F-statistic: 28.12 on 3 and 813 DF,  p-value: < 2.2e-16
```

### What Happens When Your Instruments Are...Um...Not Good

Now suppose we want to do the same thing for the `logtrade` variable, estimating the influences on it using 2SLS. The basic, structural regression is:

```
> OLSTrade<-lm(logtrade~logdisputes+contiguity+IOs)
> summary(OLSTrade)

Call:
lm(formula = logtrade ~ logdisputes + contiguity + IOs)

Residuals:
    Min      1Q  Median      3Q     Max
-6.2467 -2.2067  0.4275  1.6659  6.1264

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.191111   0.182875   1.045    0.296
logdisputes 0.408116   0.095067   4.293 1.98e-05 ***
contiguity  1.357557   0.193109   7.030 4.38e-12 ***
IOs         0.133778   0.005614  23.831  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.312 on 813 degrees of freedom
```

```
Multiple R-squared: 0.5241,Adjusted R-squared: 0.5223
F-statistic: 298.4 on 3 and 813 DF,  p-value: < 2.2e-16
```

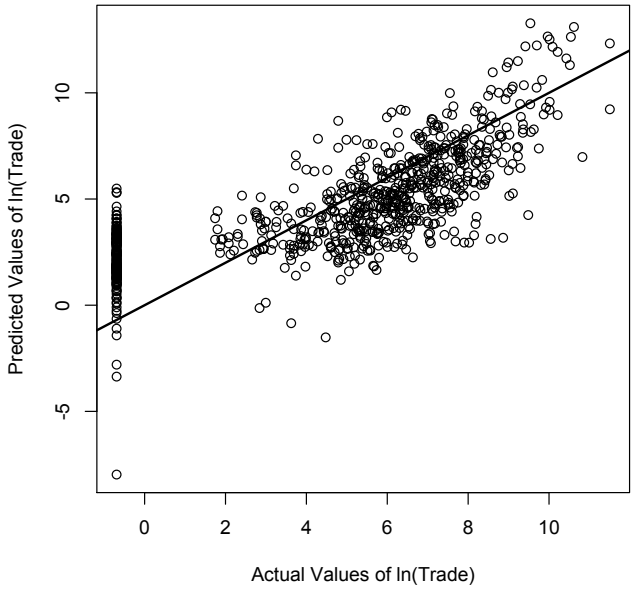
We can estimate the 2SLS model in exactly the same way as before:

```
> TwoSLSTrade<-tsls(logtrade~contiguity+IOs+I(logdisputes),
+   instruments=~contiguity+capratio+IOs)
> summary(TwoSLSTrade)

2SLS Estimates

Model Formula: logtrade ~ contiguity + IOs + I(logdisputes)

Instruments: ~contiguity + capratio + IOs

Residuals:
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-2.57e+01 -1.46e+00  1.36e+00  2.84e-14  4.00e+00  1.09e+01

               Estimate Std. Error t value  Pr(>|t|)
(Intercept)       2.150    0.85122   2.526 1.173e-02
contiguity       -2.728    1.52615  -1.787 7.427e-02
IOs               0.172    0.02045   8.408 2.220e-16
I(logdisputes)    7.371    2.45198   3.006 2.727e-03

Residual standard error: 6.3721 on 813 degrees of freedom
```

Here, we'd actually be led to believe that – in addition to war leading (instantaneously) to *more* trade – countries that are contiguous with each other actually trade *less* than those that are not. In addition to common sense, this goes against something like 200 years of macroeconomic theory...

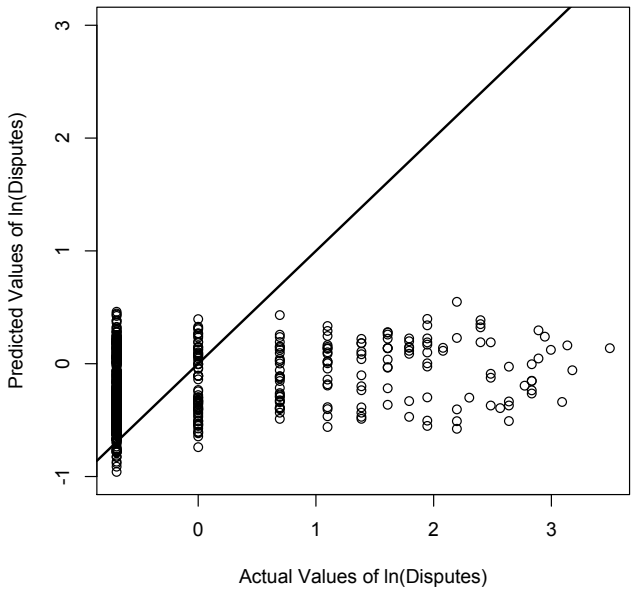
One key to the difference between these models is the strength of the instruments: in the model of war, the trade variable was instrumented fairly well ($R^2 = 0.55$), while in the trade model the instrument for war was lousy ($R^2 = 0.09$). We can see just how good / bad they were in Figures 3 and 4:

Figure 3: A Pretty Good Instrument: Trade




Note: Line is a 45-degree line (actual vs. actual).

Figure 4: A Crappy Instrument: Wars




Note: Line is a 45-degree line (actual vs. actual).

A second issue, however, is model specification. No one ought to believe that the models given here are fully, correctly specified models of war, or trade, or even IO membership. For example, we'd probably (minimally) want to include a variable for the dyad's GDP (or some other indicator of the economies' size) in a complete model of trade. Similarly, we'd probably also want a measure of (say) joint democracy, since we know all about the democratic peace. Of course, to the extent that GDP is endogenous to both trade and war, democracy is endogenous to GDP, and so forth, things can get *very* complicated, very quickly...