SGPE QM Lab 5: Endogeneity and Instrumental Variables

Mark Schaffer
version of 23.10.2011

Introduction

In the first part of this lab, we consider a particular violation of the assumptions of the classical model, namely endogeneity, i.e., E(Xu)≠0. We first look at Monte Carlos illustrating the consequences of the failure of this assumption. We then look in detail at a well-known paper by Acemoglu-Johnson-Robinson (2001) that used IV methods to address the endogeneity problem.

Preparation

Created a folder on your network drive called "M:\QM\Lab5". Go to the coursework folder where we keep central copies of QM files and copy the files `lab5assignment1.do` and `lab5assignment2.do` to your M:\QM\Lab5 folder. Also copy `Lab5_AJR_v01.do` and `maketable8.dta`. The last file is an AJR data file.

Review of Assignment 1: Monte Carlo 1

The first part of the MC do file defines the program `mysim`:

```
program define mysim, rclass
      drop _all
      set obs 100
      gen t = _n
      tsset t

* Generate independent variable x and error u
      gen x = rnormal()
      gen u = rnormal()

* Generate the dependent variable y according to the
* following true model: y = b0 + b1*x + u
* b0 (the constant)   = 1
* b1 (the coeff on x) = 1
      gen y = 1 + 1*x + u

* Estimate the model using OLS
      reg y x

* Return the estimated b1 and the estimated SE(b1).
      return scalar b1=_coef[x]
      return scalar se1=_se[x]
end
```

The program `mysim` will be called by the Stata `simulate` command. Each time `mysim` is called, it generates a dataset with 100 observations and estimates an OLS regression using the `regress` command. It then reports back to `simulate` what the estimate coefficient `b1` is, and what the estimated standard error for `b1` is.

The **simulate** command calls **mysim**, collecting the estimated **b1** (coefficients) and **se1** (standard error of **b1**). It runs 5,000 times, so when it is done running there will be 5,000 different coefficients and SEs, representing 5,000 different uses of OLS on 5,000 different datasets:

```
simulate b1 = r(b1) se1 = r(se1), reps(5000): mysim
```

The data generation process (DGP) is as follows:

$$\begin{aligned}
&\boldsymbol{y}_i = \beta_0 + \beta_1 \boldsymbol{x}_i + \boldsymbol{u}_i \\
&\boldsymbol{x} \sim \boldsymbol{N}(0,1) \\
&\boldsymbol{u} \sim \boldsymbol{N}(0,1) \\
&\beta_0 = 1 \\
&\beta_1 = 1
\end{aligned}$$

We use this MC to investigate the possible bias in estimates of β_1 (the coefficient on x), and whether there are signs of any size distortions in tests of the null hypothesis of H_0: β_1=1. (Recall the definition of a Type I error: incorrectly rejecting the null when it is actually true. A correctly-sized test will incorrectly reject the null 5% of the time when the chosen significance level is 5%, and similarly for other significance levels.)

Questions for Monte Carlo 1:

What is the evidence that the OLS estimate of β_1 is biased or unbiased?

The summarize command

```
sum b1, detail
```

produces the following output:

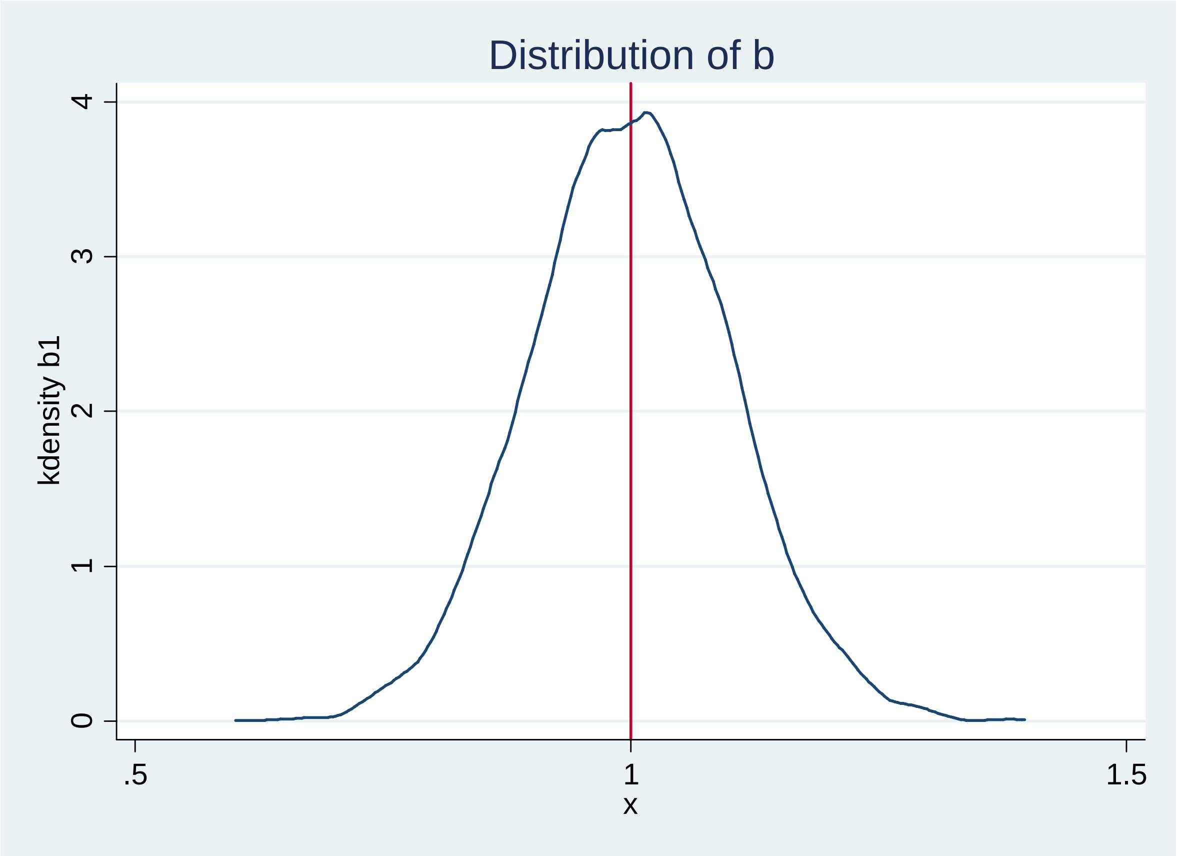
```
                            r(b1)
-------------------------------------------------------------
      Percentiles      Smallest
 1%     .7651131       .6012219
 5%     .8356454       .6562059
10%     .8712333       .6585693       Obs                5000
25%     .9334866       .6613736       Sum of Wgt.        5000

50%      1.00048                      Mean           .9999794
                        Largest       Std. Dev.       .100346
75%     1.067853       1.312769
90%     1.126495       1.375645       Variance       .0100693
95%     1.164968        1.37572       Skewness       .0321371
99%     1.236093        1.39743       Kurtosis       3.090021
```

You can see that the 5,000 different estimates of β_1 had a mean of approximately 1. In other words, the MC suggests that the OLS estimator is approximately unbiased for this DGP. This is easy to see visually using the code:

```
twoway (kdensity b1), xlabel(0.5 1 1.5) xline(1)            ///
      title(Distribution of b) name(bias, replace)
```

which generates the following graph:



It appears unbiased, but checking the actual mean is important – you could easily miss a small bias in either direction if you just relied on looking at the graph.

What is the evidence that 2-tailed tests of the null hypothesis H_0: $\beta_1=1$ at the 5% significance level are correctly sized, i.e., that they wrongly reject the null 5% of the time?

The following code in the do file generated the test necessary statistic:

```
gen t=(b1-1)/se1
```

and this can be checked against the critical values of the Normal distribution, -1.96 and 1.96, using the `count` command:

```
. count if t<-1.96
  124

. count if t> 1.96
  139
```

H_0 was rejected 124 times because t<-1.96, and 139 times because t>1.96.[1] There were 5,000 tests of H_0, the null being tested is true, we are using the 5% significance level, so we would expect to incorrectly reject the null 5% * 5,000 = 250 times. We conclude we have MC evidence that tests using classical OLS are correctly sized.

[1] In fact, because we set the seed for the pseudo-random-number generator, the number of rejections should be the same each time – 124 and 139.

What is the evidence that 2-tailed tests of the null are correctly sized across the full range of possible significance levels?

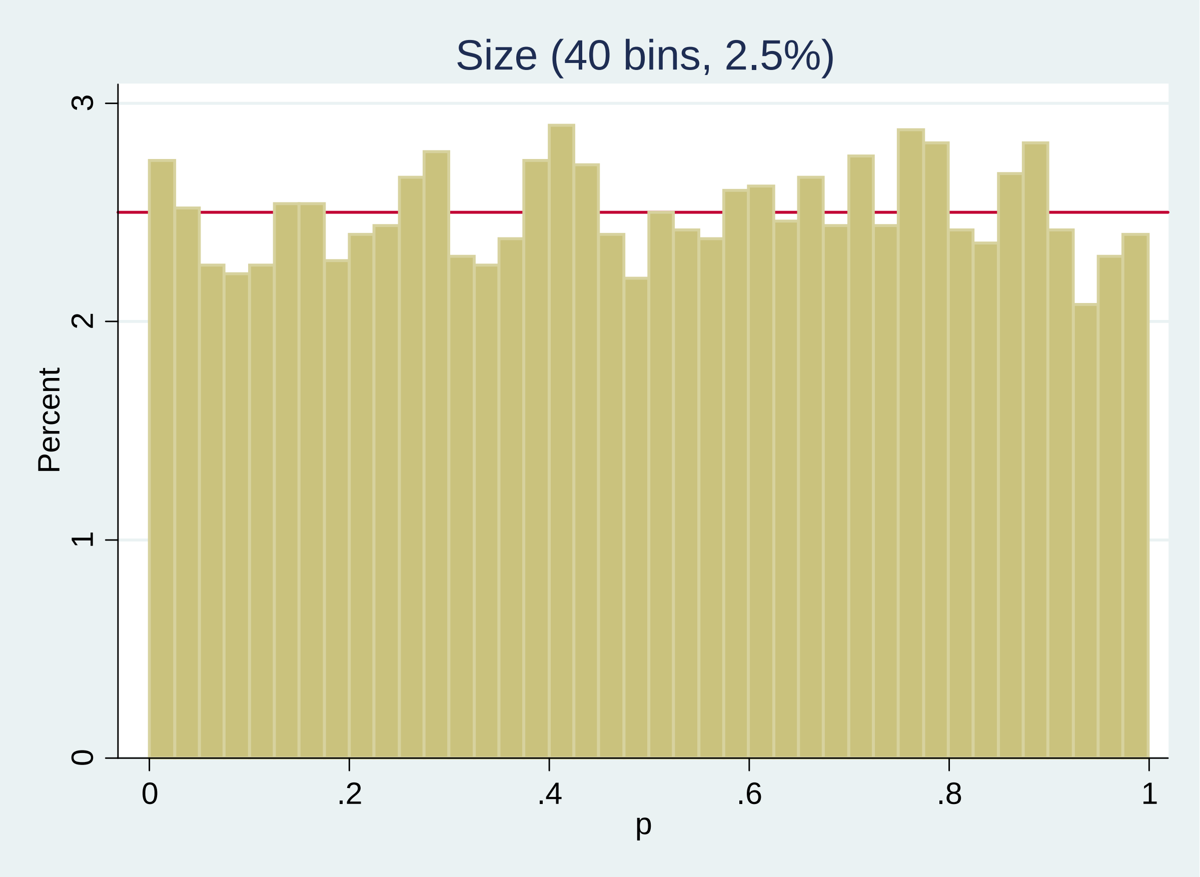
If the test is correctly sized, the distribution of p-values should have the uniform distribution. (Remember: the estimated coefficient is a random variable because it depends on the dataset at hand; the estimated SE is a random variable as well; so is the test statistic; and so, therefore, is the calculation of the p-value for the test statistic.)

A p-value is the probability of observing (if the null is true) a test statistic at least as extreme as the one we have, and "extreme" means either very large (positive) or very small (negative). For a 2-tailed test we work with the absolute value of the test statistic. The distribution of the absolute value of a Normal variable is called the *Half-Normal*. The CDF of the Half-Normal is simply F=2Φ(a)-1 where Φ(a) is the CDF for the Normal.

The Stata function `normal(t)` is the Normal CDF, i.e., the area under the curve and to the LEFT of `t`. The area under the Half-Normal CDF to the LEFT of abs(t) is `2*normal(abs(t))-1`. The p-value for `t` is therefore the area under the Half-Normal CDF to the RIGHT of abs(t) = `1-(2*normal(abs(t))-1)` = `2*(1-normal(abs(t)))`. The last command graphs the data.

```
gen p = 2*(1-normal(abs(t)))
hist p, bin(40) percent yline(2.5) title("Size (40 bins, 2.5%)")  ///
      name(size, replace)
```

You should have obtained the following graph:



There are 40 bins, so if the p-values were uniformly distributed we should observe 2.5% of the distribution in each bin. This is indeed roughly what we see. For this DGP, tests using the OLS estimates of the parameter and its standard error are properly sized.

Based on the above, in your view is OLS a good estimator to use for such a DGP?

Yes. The estimator is approximately unbiased, and tests using the estimator and the estimated SEs are correctly sized.

Review of Assignment 2: Monte Carlo 2

The second Monte Carlo is in `lab5assignment2.do`. It is almost identical to the first MC, but with a slightly different DGP. The motivation is that the explanatory variable in the DGP is *x** - the "true" *x* – but researcher can't observe the true x. Instead, the researcher has to work with x measured with error. This is the case of *measurement error* in the explanatory variable. The DGP is:

$$
\begin{array}{ll}
\boldsymbol{y}_i = \beta_0 + \beta_1 \boldsymbol{x}_i^* + \boldsymbol{u}_i & \text{(True DGP)} \\
\boldsymbol{x}_i = \boldsymbol{x}_i^* + \boldsymbol{e}_i & (\boldsymbol{x}_i \text{ is } \boldsymbol{x}_i^* \text{ observed with error)} \\
\boldsymbol{y}_i = \beta_0 + \beta_1 \boldsymbol{x}_i + \eta_i & \text{(Researcher estimates using observed } \boldsymbol{x}_i \text{, not true } \boldsymbol{x}_i^* \text{)} \\
\boldsymbol{x} \sim \boldsymbol{N}(0,1) & \\
\boldsymbol{u} \sim \boldsymbol{N}(0,1) & \\
\boldsymbol{e} \sim \boldsymbol{N}(0,0.4) & \text{(SD of measurement error=0.4)} \\
\beta_0 = 1 & \\
\beta_1 = 1 &
\end{array}
$$

And this is implemented in the `mysim` program by:

```
* Generate independent variable xstar and error u
     gen xstar = rnormal()
     gen u = rnormal()

* Generate the dependent variable y according to the
* following true model: y = b0 + b1*xstar + u
* b0 (the constant)   = 1
* b1 (the coeff on x) = 1
     gen y = 1 + 1*xstar + u

* But we don't observe xstar, we observe x.
* x is xstar but measured with error.
* Generate measurement error e and observed x.
     gen e = 0.4*rnormal()
     gen x = xstar + e

* Estimate the model using OLS
     reg y x
```

The DGP is the classic case of *measurement error in the explanatory variable.* The dependent variable y is determined by x*, but we don't observe the true x*, we have only the noisy but observed measure x. If we use the observed x in an OLS regression, we will get biased and inconsistent estimates of β_1. The reason is that measurement error causes x to be *endogenous*: x is correlated with u, i.e., E(Xu)≠0.

The consequence of measurement error for OLS estimates is sometimes called "attenuation bias". The reason is that the OLS estimate of β_1 is *biased towards zero*; it is "attenuated".

Questions for Monte Carlo 2:

What is the evidence that the OLS estimate of β_1 is biased or unbiased?

The summarize command

```
      sum b1, detail
```

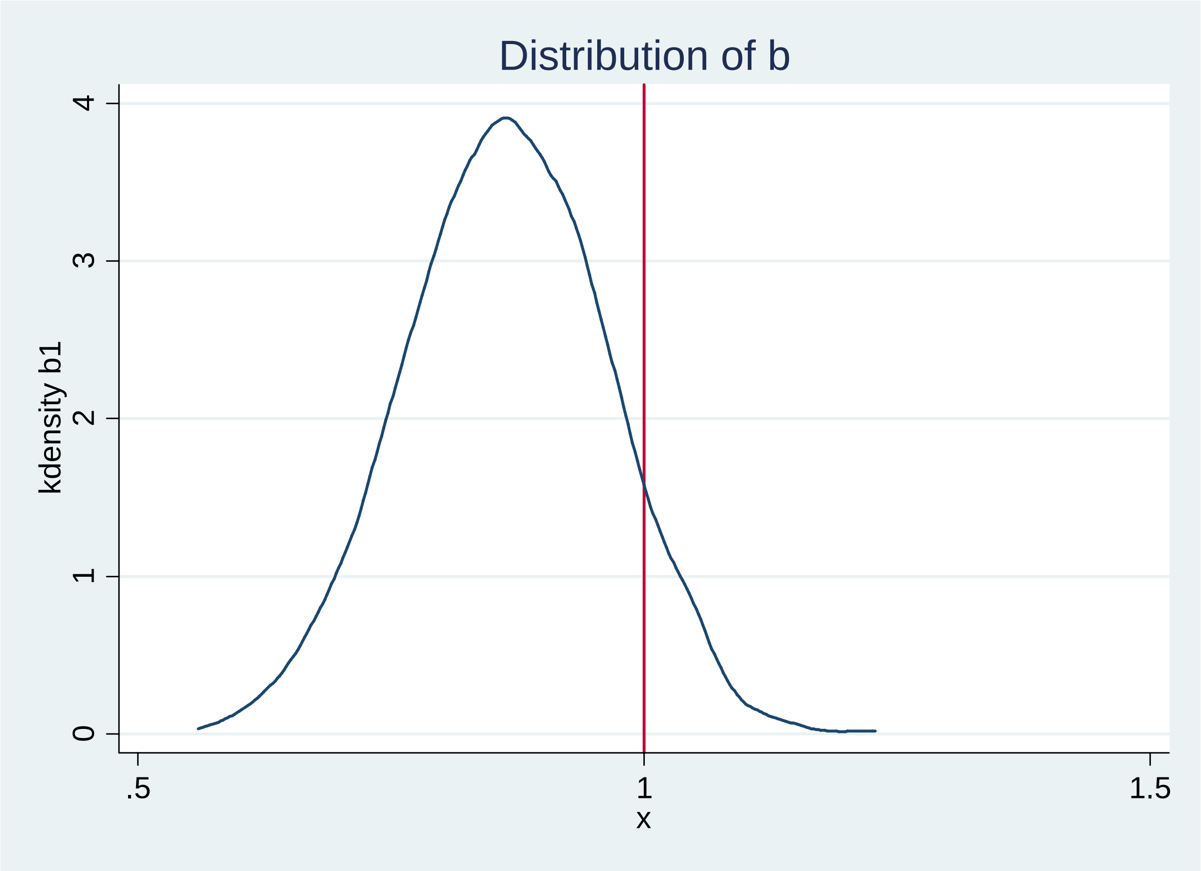
generates the following output:

```
                            r(b1)
-------------------------------------------------------------
      Percentiles      Smallest
 1%     .6312414      .5594656
 5%     .6967987      .5667847
10%     .7356413      .5701028       Obs                5000
25%     .7961353      .5719769       Sum of Wgt.        5000

50%      .865135                     Mean            .863896
                        Largest       Std. Dev.      .0998257
75%     .9324819       1.21109
90%     .9921023      1.217089       Variance       .0099652
95%     1.029745      1.222988       Skewness      -.0077368
99%     1.088056      1.228306       Kurtosis       2.895613
```

The true value is 1, but the 5,000 different estimates of β_1 had a mean of approximately 0.86, a noticeable downward bias of about 0.14. This is what we expect from measurement error: attenuation bias, i.e, bias towards zero.

This is very apparent in the distribution of the estimator:



What is the evidence that 2-tailed tests of the null hypothesis H_0: $\beta_1=1$ at the 5% significance level are correctly sized, i.e., that they wrongly reject the null 5% of the time?

Now if we use the **`count`** command,

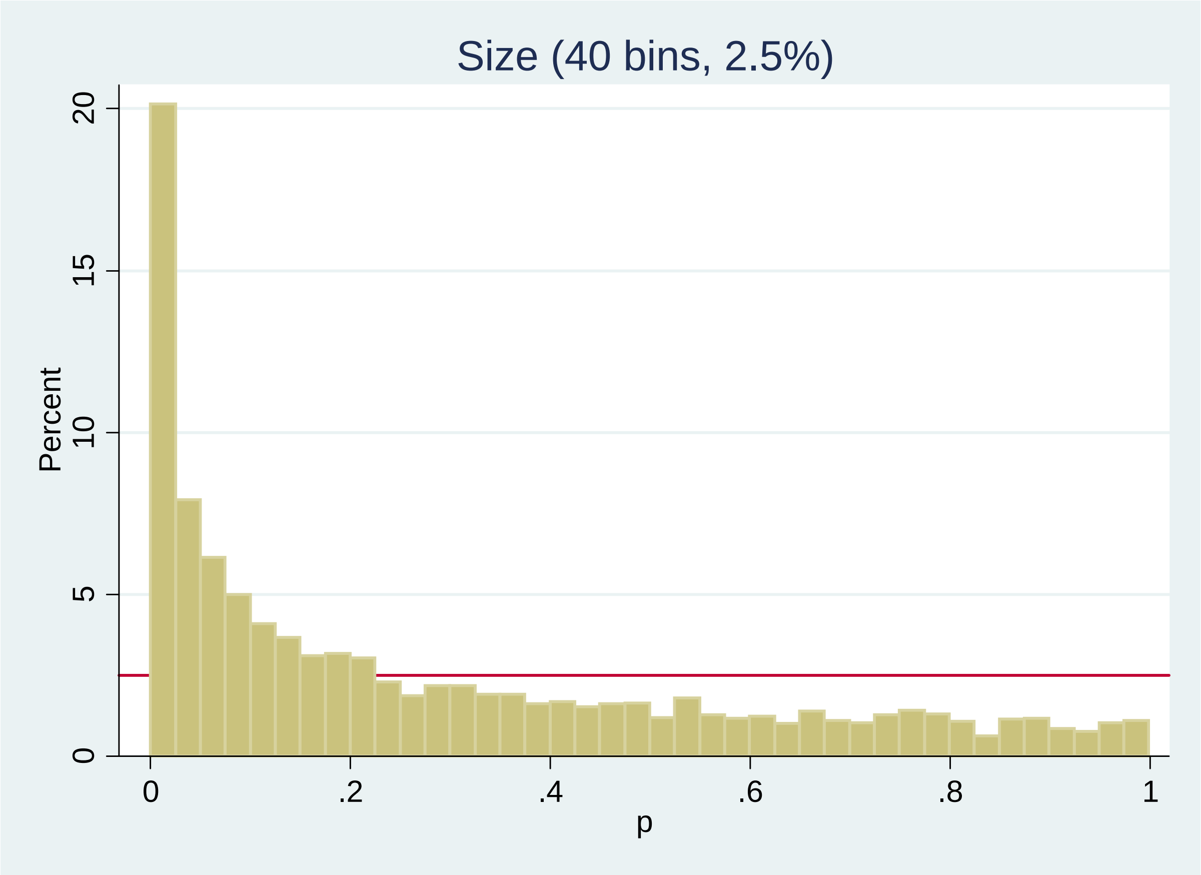
```
. count if t<-1.96
 1402

. count if t> 1.96
    1
```

we find that there are almost no rejections because of a large test statistic (t>1.96), which is not surprising giving the graph – there are very large estimates of β_1. But the number of rejections because of a small test statistic (t<-1.96) is very large – about 1,400 times out of the 5,000 total replications, i.e., 25-30% of the time. We should be rejecting a correct null only about 5% of the time. Tests using OLS are therefore badly sized.

What is the evidence that 2-tailed tests of the null are correctly sized across the full range of possible significance levels?

The graph of the p-values is very far from uniform:



The piling-up of very small p-values is telling us that we are very likely to over-reject the correct null hypothesis. The true value of β_1 is 1, but very often we will conclude this is wrong (and, based on the estimated β_1, be mislead into thinking the true value is lower).

Based on the above, in your view is OLS a good estimator to use for such a DGP?

No!

Review of Assignment 3: Background work for Acemoglu-Johnson-Robinson (2001)

"The Colonial Origins of Comparative Development" is a 2001 AER paper by Daron Acemoglu, Simon Johnson and James Robinson (hereafter AJR). The AJR paper is an attempt to tease out the strength of the relationship between the log of GDP per capita (our dependent variable) and the quality of institutions (our independent variable of interest). Economists have long believed that there should be a positive relationship between institutional quality and GDP per capita, and the main purpose of the paper is to quantify that relationship.

The AJR paper can be found at http://economics.mit.edu/files/4123, and we will also put a copy in the Lab5 coursework folder (where the do files are).

The three main variables of interest are:

y_i `logpgp95` Log PPP GDP per capita in 1995, World Bank (note: PPP denotes that the measure is based on purchasing power parity rather than nominal exchange rates).

x_i `avexpr` Average protection against expropriation risk. This is measured on a 1 to 10 scale, where a low score denotes a bad institutions (the government is likely to steal your property) and a high score indicates good institutions (the government is *not* likely to steal your property).

$\mathrm{z_i}$ `logem4` Log settler mortality. This is based on the earliest available record of the mortality rate of European settlers living in the country of interest, measured as deaths per thousand per year.

Two additional variables we will use in the lab:

`lat_abst` Abs(latitude)/90. This captures distance from the equator.

`euro1900` Percentage of the population descended from European settlers, as of the year 1900.

The model to be estimated is (ignoring all other explanatory variables):

$$\boldsymbol{y_i = \beta x_i + u_i}$$

Questions for Assignment 3:

Why is it a problem for OLS estimation if x is correlated with u?

The answer to this is the same as we saw in the MC assignment: if E(`avexpr`*u)≠0, x is *endogenous*, and the OLS estimate of β will be biased an inconsistent.

Why are AJR worried that this could be the case here?

AJR's primary concern is reverse causality. What they want to explore is the impact of institutions (proxied by `avexpr`) on living standards (measured by `logpgp95`, log GDP per capita).

The problem is that the causality may run the other way: rich countries can afford to pay for good institutions, and poor countries cannot. That is, not only does `avexpr`→`logpgp95`, but also `logpgp95`→`avexpr`. This will cause `avexpr` to be endogenous, i.e., E(`avexpr`*u)≠0. This will lead to an *upward bias* in β: our estimated β would be misleadingly large. We would think we found that better institutions lead to higher living standards, whereas we would really have found that countries with higher incomes spend some of it on better institutions.

A second reason for endogeneity that we should worry about is measurement error in the proxy for institutional quality, `avexpr`. Measuring "good institutions" is not straightforward, and if this measure is noisy (which is likely), we will have a measurement error problem. Note that this will lead to a *downward bias* in β (towards 0). In other words, in this application, reverse causality bias and measurement error bias will go in opposite directions. Which is larger, and by how much, we cannot say in advance.

What is the AJR argument for why z (death rates of European settlers in their colonies in the 19th century and earlier) should be correlated with x (quality of institutions today)?

The AJR argument is that the European colonial powers followed one of two different strategies with their colonies: either set up "little Europes", complete with European settlers and European institutions (laws, business codes, etc.); or exploit the colony by extracting raw materials, commodities, etc., and without regard for institution building.

If the colony was a safe place for Europeans to live, the colonial power would send settlers and build little Europes; if the colony was a dangerous place for Europeans to live because of diseases (malaria, for example), the colonial power exploit the colony instead.

Death rates of European settlers are a measure of how attractive a colony was for building little Europes. Low death rates = attractive, build institutions; high death rates = not attractive, exploit.

This makes the variable `logem4` potentially a good *instrument* for IV estimation. It is correlated with institutional quality today **`avexpr`** because former colonies that were good places for Europeans to live were also places where European institutions were implanted. Former colonies that were exploited instead were also places where these European institutions were not left behind when the colonial powers left.

What is the AJR argument for why z should NOT be correlated with u, i.e., why death rates of European settlers should have no effect living standards (GDP per capita) today other than x?

The concern here is that the instrument `logem4` must *also* be exogenous, i.e., E(`logem4`*u)=0. Why might this fail? Say that the disease environment is a determinant of living standards today, `logpgp95`. This sounds plausible: countries with high rates of malaria, say, could be poor countries partly *because* of the disease environment. This would make any measure of the disease environment (such as settler mortality) correlated with the error term u, because it would be correlated with something that causes `logpgp95` but isn't in the regression (and hence is in the error term).

AJR argue that this is unlikely because these diseases (malaria and yellow fever, mostly) were almost always fatal to Europeans (who had no immunity) but were much less serious for native inhabitants (who had some immunity). NB: some other authors are not convinced of this argument.

Preparation

AJR have made their data and do files available on their website. We will work with the AJR data and modified versions of their do files. Load the do file `Lab5_AJR.do`. Execute the following lines at the beginning:

```
use maketable8, clear
keep if baseco==1
```

This loads ones of the AJR datasets and limits the sample to the "base case" countries.

Task 1: Examine the data

We are going to do some basic summary statistics. First, type

```
hist logpgp95
```

Notice the distribution of the log of GDP per capita. To see the distribution of the level of GDP per capita, we must generate a new variable by exponentiating the logs:

```
gen pgp95=exp(logpgp95)
hist pgp95
```

Make sure you understand what we do when we take natural logs of GDP per capita; type

```
list shortnam logpgp95 pgp95
```

Take notice of the values for `NGA` and `NIC`. Notice that the unlogged GDP per capita of Nicaragua is about double that of Nigeria. By how much do the logs differ? Now take note of the values for `ETH`. Note that the unlogged GDP per capita of Nigeria is about double that of Ethiopia. By how much do the logs differ?

Now we're going to look at a scatterplot from the paper. To see how log GDP per capita relates to institutional quality (i.e. `avexpr`, our main independent variable), execute the following line. Note that by using the `mlabel` option we can see which country is which.

```
twoway scatter logpgp95 avexpr, mlabel(shortnam)
```

Stare at the plot for a while. Is it what you expected? Argue with the other people at your table about whether or not it seems reasonable. Notice that the slope of the line is roughly 0.5. What does this mean?

Next, we're going to replicate two of the OLS results from the AJR paper. The first estimation is a simple bivariate regression; in the second, AJR include latitude as an additional explanatory variable:

```
reg logpgp95 avexpr
reg logpgp95 lat_abst avexpr
```

Interpret coefficient on `avexpr` in light of what you've already done today.

But remember that we can't trust these results, because we have reason to think that OLS is biased and inconsistent.

Task 2: IV estimation, part 1 – is the instrument z correlated with x?

Recall the two requirements of a good ("valid") instrument z: it should be correlated with the regressor z, and uncorrelated with the error u. We can't (yet) check that the instrument `logem4` is uncorrelated with u, but we *can* check if it's correlated with `avexpr`.

First, eyeball the data:

```
twoway (scatter avexpr logem4, mlab(shortnam)) (lfit avexpr logem4)
```

Is the relationship what you expected? Why is the slope negative?

Now estimate the following regression:

```
reg avexpr logem4
```

This is the “first-stage” regression for the simple bivariate model; it is called this because it is the first stage of 2-stage least squares (2SLS), another name for IV. What is the F-stat for the coefficient on `logem4`? A rule of thumb is that the F-stat should be at least 10 for an instrument be “valid” in the sense of correlated enough with the endogenous regressor. If the F-stat is less than 10, we would call it “weak”. Is `logem4` “valid” or “weak”?

Repeat the exercise for the model including latitude. (Remember: any additional regressors that appear in the main equation must also appear in the first stage regression.) Do your conclusions change?

Task 2: IV estimation, part 2 – estimate using IV

Now estimate using IV. You could do this the old-fashioned way using 2SLS (get the fitted values from the first stage, etc.), but it’s much easier to use an IV estimator.

You can either use Stata’s built-in IV estimator **`ivregress`**, or you can use the **`ivreg2`** add-in. You’ll also need to install the **`ranktest`** add-in. To get these, use the **`findit`** command. Be sure to install from http://fmwww.bc.edu/RePEc.

```
findit ivreg2
findit ranktest
```

To estimate using IV:

```
ivreg2 logpgp95 (avexpr=logem4)
ivregress 2sls logpgp95 (avexpr=logem4)
```

Note that with the **`ivreg2`** add-in the first-stage F-stat is reported in the footer as a “weak identification test” or “Cragg-Donald F statistic”.

How do your conclusions change compared to the OLS results? Repeat for the specification with latitude as a regressor.

Task 3: IV estimation, part 3 – is IV necessary?

The qualitative conclusions from IV and OLS estimation appear similar. Are the coefficients quantitatively similar? Are they statistically similar enough that we can use OLS instead of IV? After all, OLS is more efficient.

It turns out that this is the same question as, “Is **`avexpr`** actually endogenous after all?” The Durbin-Wu-Hausman (DWH) endogeneity test is a test of whether a regressor is endogenous or not. It works by comparing the coefficient from IV and the coefficient from OLS. Under the null that the regressor is exogenous, there shouldn't be any difference, because both IV and OLS are consistent. If they *are* different, it must be because the regressor is endogenous.

We can implement this very simply using the `endog` option of `ivreg2`:

```
ivreg2 logpgp95 (avexpr=logem4), endog(avexpr)
```

What do you find? Repeat for the specification with latitude. Do your conclusions change?

Task 4: IV estimation, part 3 – overidentification

What if we have more than one instrument available? We can obtain more efficient estimates this way: two instruments are better than one in an efficiency sense. But it also means we can do another check of the specification: an *overidentifying restrictions test*. (This test is sometimes called an overidentification test, a Sargan test, a Sargan-Hansen test, a Basmann test, etc.)

The main reason that an overid test is useful is that it provides a test, albeit a limited one, of the assumption that the instruments are exogenous. That is, it is a test of E(Zu)=0.

AJR have several alternative candidates for instruments. We will use one of them, `euro1900`, a measure (%) of how much of the population in 1900 was descended from European settlers. The rationale is similar to that of `logem4`; more European settlers in 1900 implies more European institutions. But we also require that `euro1900`, like `logem4`, is uncorrelated with the error term. Put another way, if these two variables have an impact on GDP per capita today, it should be *only* via institutions `avexpr`.

The intuition behind an overid test is simple, and similar to that of the DWH endogeneity test (in fact, they are both special cases of GMM distance tests, but more about that in Adv QM). Under the null, both `euro1900` and `logem4` are exogenous and hence valid instruments. If you estimated using just `euro1900` as an instrument, you should get about the same results as if you estimated using just `logem4`, and these should be about the same as using both `euro1900` and `logem4` as instruments. If, on the other hand, the results were different, you have reason to suspect at least one of your instruments is correlated with the error term u.

When using more than one instrument, of course, you should also check the first-stage regression and confirm that all the instruments are correlated with the endogenous regressor. There's no point using a variable as an instrument if it's just noise.

First, consider euro1900 as an instrument on its own:

```
twoway (scatter avexpr euro1900, mlab(shortnam)) (lfit avexpr euro1900)
reg avexpr euro1900
test euro1900
```

Next, compare the estimations using the two instruments separately. The `if e(sample)` ensures that the same estimation sample (only 63 obs) are used in the two regressions. Are the estimated coefficients on `avexpr` similar? What does that predict for your overidentification test later?

```
ivreg2 logpgp95 (avexpr=euro1900)
ivreg2 logpgp95 (avexpr=logem4) if e(sample)
```

The following will produce all the results you need for the simple bivariate model:

```
ivreg2 logpgp95 (avexpr=logem4 euro1900), first
```

Note the use of the `first` option; this will generate the first-stage regression results for you.

Check the first-stage results. Is `logem4` still related to `avexpr`? What about the new instrument, `euro1900`?

What is the first-stage F statistic for the joint significance of `euro1900` and `logem4`? Is it greater than 10 (the "rule of thumb")?

What is the value of the Sargan overidentification statistic? Do you reject the null that the two instruments are exogenous?

What is the estimated coefficient on `avexpr`? Compare this to the estimated coefficients you obtained using the two instruments separately. Can you relate this comparison to the result of the Sargan test?