Problem Set 4

Nonstationarity, Causality, Instrumental Variables

EC 421: Introduction to Econometrics

Due *before* midnight (11:59pm) on Saturday, 16 March 2019

DUE Your solutions to this problem set are due *before* midnight on Sunday, 03 March 2019. Your files must be uploaded to Canvas—including (1) your responses/answers to the question and (2) the R script you used to generate your answers. Each student must turn in her/his own answers.

OBJECTIVE This problem set has three purposes: (1) reinforce econometrics topics from class; (2) build your R toolset; (3) strengthen your intuition on causality and time series.

Problem 1: Nonstationarity—the Basics

1a. Define stationarity.

Note: You can define it using math or words (or both).

1b. If our disturbance term u_t follows a random walk, *i.e.*,

$$u_t = u_{t-1} + \varepsilon_t$$

then it's variance is $\mathrm{Var}(u_t) = t\sigma_\varepsilon^2$. Explain how this expression of its variance shows that the disturbance is nonstationary (*i.e.*, it violates stationarity).

1c. We previously discussed autocorrelated distrubances, *e.g.*, an AR(1) process such that

$$u_t = \rho u_{t-1} + \varepsilon_t$$

Under which circumstances would this AR(1) process become a random walk?

Hint: Consider the values of ρ.

Problem 2: Nonstationarity—the Simulation

In this problem, we are going to create two independent, **nonstationary** time series. Specifically, we'll create two random walks. Then, we'll regress the first random walk on the second random walk.

Hint: Generating random walks is *nearly* identical to generating AR(1) processes, as you did in lab.

2a. Generate the first 30-period random walk. We will name it `v`.

$$v_t = v_{t-1} + \varepsilon_t$$

where ε_t comes from a normal distribution with mean 0 and standard deviation 1.

Here is some `R` to help.

```
# Set a seed (so your results stay the same)
set.seed(123)
# Generate the initial number, (this will be v[1])
v ← rnorm(1, mean = 0, sd = 1)
# For loop to create the random walk
for (t in 2:30) {
  # Create the 'next' observation
     ...
}
```

while you're filling in the `for` loop, keep in mind (**1**) our equation for the random walk at the beginning of this question (meaning v_t depends upon v_{t-1} and ε_t) and (**2**) the fact that you can reference different observations in `R`, *e.g.*,

- `v[t]` refers to the t^{th} observation
- `v[t-1]` refers to the $(t-1)^{\text{th}}$ observation
- `v[3]` refers to the 3^{rd} observation

If you need more help on for loops, don't forget there are lab materials on Canvas and resources online (*e.g.*, datamentor.io and datacamp.com have lots of resources).

2b. Generate a second 30-period random walk called `w`. This part is exactly the same as (2a), but you **use a different seed** (*i.e.*, `set.seed(456)`) and **name the variable** `w`.

2c. We independently generated these two time series. Ideally (from a statistical point of view), should we find a statistically significant relationship between the two series? Explain.

2d. Regress `w` on `v`. Report the results from the t test. Do they match your expectations from (2c)?

Problem 3: Causality

Following the Rubin causal model, imagine that we observe the following data (which would be impossible observe in real life):

Table: Imaginary dataset

i	**Trt.**	**y_1**	**y_0**
1	0	12	8
2	0	7	5
3	1	5	1
4	1	6	4

3a. Calculate the treatment effect **for each individaul** (*i.e.*, τ_i).

3b. [T/F] The treatment effect is constant across individuals.

3c. Calculate the **average treatment effect**.

3d Estimate the average treatment effect by comparing the **mean of the treatment group** to the **mean of the control group**.

3e. Should we expect our estimator in (3d) to provide unbiased estimates? **Explain.**

3f. Why would it be impossible to actually observe all of the data in the table (in real life)?

3g. How does your answer in (3f) relate to *the fundametal problem of causal inference*?

Problem 4: Instrumental Variables

Let's return to our question of the returns to education. Specifically, we will use the dataset `wages.csv`, which .[†]

We're interested in estimating β_1 in

$$\text{Wage}_i = \beta_0 + \beta_1 \text{Education}_i + u_i$$

but we have a problem with omitted-variable bias. Instrumental variables can potentially help.

4a. Load and inspect the dataset.

4b. What are the two requirements for a valid instrument?

4c. As we've discussed, we need an instrument for (endogenous) education. Do you think the variable `n_kids`—the number of children—would be a valid instrument? Explain why it passes/fails ech of the two requirements for a valid instrument.

4d. We can test the *relevance* of our instrument by estimating the first stage, *i.e.*, regressing our endogenous variable $\texttt{education}$ on our (potential) instrument $\texttt{n\_kids}$.

Do it.

Is there evidence that our potential instrument is relevant? Explain using a statistical test and interpret the coefficient.

4e. Let's assume *number of children* is a valid instrument for education.

Using the number of children (`n_kids`) as an instrument for education (`education`), estimate the returns to education via instrumental variables (IV).

Interpret the coefficient that gives the returns to education and its significance.

Hint: Recall that we can use `iv_robust(y ~ x | z, data)` from the `estimatr` package to get IV/2SLS estimates of the effect of `x` on `y` with the instrument `z` (and dataset `data`).

4f. How do your estimates of the returns to education from instrumental variables (IV) compare to estimates using plain ordinary least squares (OLS)?

Hint: You'll need to estimate the model using OLS.

4g. Extra credit: Explain which estimates you would trust more (or why you distrust both).

These data come from `wage1` in the `wooldridge` package. I took a subset of variables and renamed them.