Problem Set 4, Solutions

Nonstationarity, Causality, Instrumental Variables

EC 421: Introduction to Econometrics

Due *before* midnight (11:59pm) on Wednesday, 05 June 2019

DUE Your solutions to this problem set are due *before* midnight on Wednesday, 05 June 2019. Your files must be uploaded to Canvas.

IMPORTANT Your submission must include (1) **your responses/answers to the question in a PDF, Word, or similar file** and (2) the R script you used to generate your answers. **The R script is just for your code. To receive credit, your answers/figures/*etc.* must be in the PDF/Word document.** 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).

Answer: *Stationarity* provides a concept of "well-behaved" time-series processes. We want our data to be *weakly persistant*, meaning periods that are far apart in time do not have a strong relationship. Stationarity formalizes this requirement. Specifically, stationarity means

1. The **mean** of our variable is independent of time, (*i.e.*, $\boldsymbol{E}[x_t] = \boldsymbol{E}[x_{t-k}]$ for any k)
2. The **variance** of the variable is independent of time (*i.e.*, $\text{Var}(x_t) = \text{Var}(x_{t-k})$ for all k)
3. The **covariance** between two periods is independent of time (*i.e.*, $\text{Cov}(x_t,\, x_{t-k}) = \text{Cov}(x_s,\, x_{s-k})$ for any s, t, and k)

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 $\text{Var}(u_t) = t\sigma_\varepsilon^2$. Explain how this expression of its variance shows that the disturbance is nonstationary (*i.e.*, it violates stationarity).

Answer: The variance of a random walk clearly depends upon time—meaning it is **not** independent of time. In other words: As time increases, the variance increases.

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.

Answer: If $\rho = 1$, then this AR(1) process becomes a random walk.

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 50-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(1234)
# 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 2:50) {
  # 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).

Answer: Here is R code for our first random walk...

```
# Set the seed
set.seed(1234)
# 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 2:50) {
  # Create the 'next' observation
  v[t] ← v[t-1] + rnorm(1, mean = 0, sd = 1)
}
```

2b. Generate a second 50-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`.

Answer: Here is R code for our second random walk...

```
# Set the seed
set.seed(5678)
# Generate the initial number, (this will be v[1])
w ← rnorm(1, mean = 0, sd = 1)
# For loop to create the random walk
for (t in 2:50) {
  # Create the 'next' observation
  w[t] ← w[t-1] + rnorm(1, mean = 0, sd = 1)
}
```

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.

Answer: If two variables are generated independently, then we ideally would not find a statistically significant relationship between them.

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

Answer: Regressing `w` on `v`

```
# Regress w on v
reg_2d ← lm(w ~ v)
# 'tidy' results
reg_2d %>% tidy()
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)   -2.55     0.676      -3.78 0.000438
## 2 v             -0.215    0.0613     -3.50 0.00101
```

We estimate a coefficient of -0.21, which has a p-value of approximately 0.001. Thus, we find statistically significant evidence of a relationship (at the 5-percent level) despite the fact that there is no true relationship.

As we discussed in class, random walks—and other nonstationary processes—can lead to a higher probability of finding a spurious relationship.

Note Depending on the random numbers that you draw, you might not find evidence of a statistically significant relationship.

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	25	17
2	0	15	11
3	1	11	3
4	1	13	9

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

Answer: The treatment effects for the individuals are 8, 4, 8, and 4.

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

Answer: False: the treatment effect varies across individuals.

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

Answer: The average treatment effect is (8 + 4 + 8 + 4)/4 = 6.

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

Answer: Our estimate of the average treatment effect is (17 + 11)/2 - (11 + 13)/2 = 2.

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

Answer: No! Unless we have a reason to believe that treatment was randomly distributed (or as-good-as randomly distributed), there is likely selection bias. Here, we can see that selection bias is very present: the y_0 values for the treatment group are very different from the y_0 values for the control group.

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

Answer: We cannot observe the same individual (*i*) simultaneously receiving treatment and control. Thus, we will either observe y.[0] or y.[1]—not both.

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

Answer: (3f) pretty much depicts the fundamental problem of causal inference: We cannot observe the same person with treatment and without treatment.

Problem 4: Instrumental Variables

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

Answer: A valid instrument must be:

1. **Relevant**, *i.e.*, the instrument affects our endogenous variable x
2. **Exogenous**, *i.e.*, the instrument only affects our outcome variable y through the endogenous variable x ***and*** the instrument is uncorrelated with the disturbance u

4b. 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.

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

Answer: *Number of kids* is probably not a valid instrument.

While a person's number of children seems reasonably **relevant** for the person's educational level, it seems likely to be correlated with other omitted variables that are in the disturbance—*e.g.*, age, experience, parents' income. In addition, it even seems plausible that the number of children could directly affect a person's weekly wage and the available jobs, as number of children may affect the of hours a person is available to work. Thus, number of children is probably not **exogenous**.

4c. Which estimates would you trust more—OLS or IV, where number-of-children is your instrument? Explain.

Answer: It's hard to know which would we should trust more. We probably don't want to put too much faith in either estimate: the OLS-based estimate is almost definitely suffers from omitted-variable bias, and the IV-based estimate is likely inconsistent due to the fact the *number of kids* is probably not a valid instrument.