Problem Set 3

EC 421: Introduction to Econometrics

Solutions

DUE Upload your answer on Canvas *before* midnight on Friday, 05 March 2021.

IMPORTANT! You must submit **two files**:

1. your typed responses/answers to the question (in a Word file or something similar)
2. the R script you used to generate your answers. Each student must turn in her/his own answers.

If you are using RMarkdown, you can turn in one file, but it must be an HTML or PDF that includes your responses and R code (not just the RMD file).

If we ask you to create a figure or run a regression, then the figure or the regression results should be in the document that you submit (not just the code—we want the actual figure or regression output with coefficients, standard errors, *etc.*).

OBJECTIVE In this problem set, we want to (1) reinforce key topics to time-series econometrics; (2) continue to build your R toolset; (3) keep building your intuition about causality and inference within econometrics/regression.

INTEGRITY If you are suspected of cheating, then you will receive a zero. We may report you to the dean.

Theory/review

Q01. First, let's review some of the key concepts and results of time-series econometrics. For all of the sub-questions, you can assume that all of our assumptions are satisfied **unless we explicitly say otherwise**.

A. How does autocorrelation affect **static models**? Consider unbiasedness and consistency.
B. How does autocorrelation affect **dynamic models with lagged explanatory variables**? Consider unbiasedness and consistency.
C. How does autocorrelation affect **dynamic models with lagged outcome variables**? Consider unbiasedness and consistency.
D. With no autocorrelation, can **dynamic models with lagged outcome variables** be unbiased and/or consistent?
E. With no autocorrelation, can **dynamic models with lagged explanatory variables** be unbiased and/or consistent?
F. Why do we care about **nonstationarity**?

Answers:

A: In the presence of autocorrelation, static models are **unbiased** and **consistent**.
B: In the presence of autocorrelation, dynamic models with lagged explanatory variables are **unbiased** and **consistent**.
C: In the presence of autocorrelation, dynamic models with lagged outcome variables are **biased** and **inconsistent**.
D: Without autocorrelation, dynamic models with lagged outcome variables are **biased** and **consistent**.
E: Without autocorrelation, dynamic models with lagged explanatory variables are **unbiased** and **consistent**.
F: Nonstationarity is our definition of a "well-behaved" time-series variable. If the variable is nonstationary, then we could be in danger of finding spurious results.

Crime and policing

In this problem set, we're going to compare the historical time series data for crime, population, and policing (specifically the size of the police force) in Illinois between 1985 and 2019.

Q02. Load your packages. You'll probably going to need/want `tidyverse` and `here` (among others). Now load the data (`003-data.csv`). The last page of this problem set describes the variables.

Important: Please note that most of the variables (excluding `year` and `unemployment`) are in tens of thousands. For example, in 1985 the population of Illinois (the state from which we have data) was `1139.981` **times 10,000** (which is 11,399,806).

Answer:

```
# Load packages
library(pacman)
p_load(tidyverse, patchwork, broom, scales, magrittr, here)
# Load the data
crime_df = read_csv("003-data.csv")
```

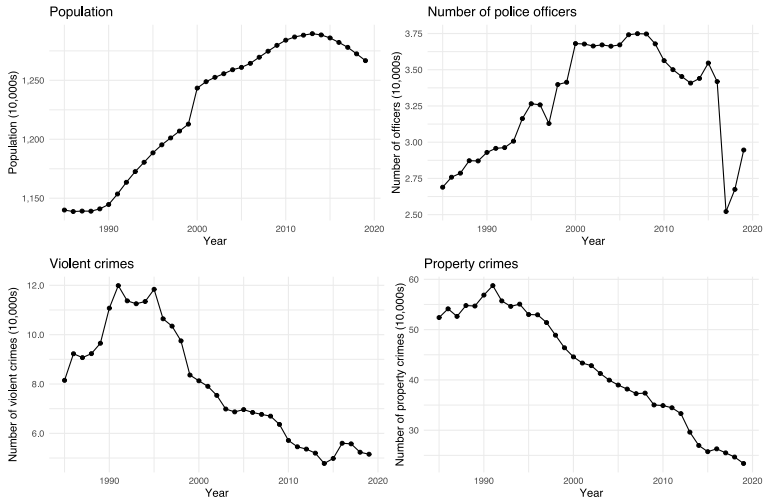
Q03. Create a time-series graph (time, which is `year` here, should be in the *x* axis) for each of the following four variables: population (`pop`, in 10,000s), number of employed police officers (`n_officers`, in 10,000s), and both types of crime (violent: `n_crime_violent`; property: `n_crime_property`; both are in 10,000s).

Don't forget to label your figures.

Answer:

```
# Population
p1 = ggplot(data = crime_df, aes(x = year, y = pop)) +
geom_line() + geom_point() +
scale_x_continuous("Year") +
scale_y_continuous("Population (10,000s)", labels = comma) +
ggtitle("Population") +
theme_minimal()
# Officers
p2 = ggplot(data = crime_df, aes(x = year, y = n_officers)) +
geom_line() + geom_point() +
scale_x_continuous("Year") +
scale_y_continuous("Number of officers (10,000s)", labels = comma) +
ggtitle("Number of police officers") +
theme_minimal()
# Violent crimes
p3 = ggplot(data = crime_df, aes(x = year, y = n_crime_violent)) +
geom_line() + geom_point() +
scale_x_continuous("Year") +
scale_y_continuous("Number of violent crimes (10,000s)", labels = comma) +
ggtitle("Violent crimes") +
theme_minimal()
# Property crimes
p4 = ggplot(data = crime_df, aes(x = year, y = n_crime_property)) +
geom_line() + geom_point() +
scale_x_continuous("Year") +
scale_y_continuous("Number of property crimes (10,000s)", labels = comma) +
ggtitle("Property crimes") +
theme_minimal()
```

```
# Combine figures
(p1 + p2) / (p3 + p4)
```



Q04. Based upon your figures in **03**, which, if any, of your variables appear to be positively autocorrelated? Which, if any, appear to be negatively autocorrelated?

Answer: Each of the variables appears to be positively autocorrelated: years that are near each other tend to take on similar values.

Q05. Based upon your time-series figure for property crimes in **03**: Does the variable appear to be stationary? If so: Explain how it satisfies each of the three requirements for stationarity. If not: Explain how it violates stationarity.

Answer: The variable appears to violate stationarity, since it is trending downward (pretty strongly) for almost all of the sample period. It is difficult to say whether this is a violation of mean stationarity (trending downward) or variance stationarity (like a random walk).

Q06. Let's start with a static model. Regress the **log** of property crimes (*i.e.*, `log(n_crime_property)`) on an intercept and on the **log** of the number of police officers (*i.e.*, `log(n_officers)`).

Include your regression results, interpret the coefficient on police officers, and comment on the significance.

Answer:

```
# Estimate the regression
reg06 = lm(log(n_crime_property) ~ log(n_officers), data = crime_df)
# Output results
reg06 %>% tidy()
```

```
#> # A tibble: 2 x 5
#>   term            estimate std.error statistic  p.value
#>   <chr>              <dbl>     <dbl>     <dbl>    <dbl>
#> 1 (Intercept)        4.30      0.491      8.75 4.07e-10
#> 2 log(n_officers)   -0.494     0.413     -1.20 2.40e- 1
```

The relationship between property crimes and police officers is negative, but it is not statistically significant. The interpretation for the (non-significant) coefficient is: a 1-percent increase the number of police officers is associated with a 0.49% **decrease** in property crimes.

Q07. We're currently using a log-log model. Explain why this might make sense in this setting, compared to a log-linear model.

Answer: We probably want a log-log model, rather than a log-linear model, because it likely make more sense to link **percent changes** in crime to **percent changes** in police employment (rather than linking percent changes in crime to **level changes** in policing).

Q08. Now add **log** population (*i.e.*, `log(pop)`) into your regression model—you now are regressing the **log** of the number of property crimes on the number of police officers and population.

Include your regression results. Interpret the coefficient on population, and comment on its significance.

Answer:

```
# Estimate the regression
reg08 = lm(log(n_crime_property) ~ log(n_officers) + log(pop), data = crime_df)
# Output results
reg08 %>% tidy()
```

```
#> # A tibble: 3 x 5
#>   term            estimate std.error statistic  p.value
#>   <chr>              <dbl>     <dbl>     <dbl>    <dbl>
#> 1 (Intercept)        54.2      2.83      19.2  4.07e-19
#> 2 log(n_officers)     1.30     0.163      7.96 4.37e- 9
#> 3 log(pop)           -7.32     0.414    -17.7  4.39e-18
```

The relationship between property crimes and population is negative and strongly statistically significant. The interpretation for the coefficient on population: a one-percent increase in population is associated with a 7.3% decrease in property crimes (holding all else constant).

Q09. When you added the log of population (moving from the regression in **06** to the regression in **08**), the coefficient on the log number of police officers should have changed signs. Give an explanation for why this happened.

Hint: Think about the lecture on signing the bias from omitted variables.

Answer: If population affects crime, and if population is correlated with the size of the police force, then our coefficient on police officers will be biased. Including population will remove this bias.

Q10. These previous models have all been static models. Explain why you think a static model in this setting could be appropriate or inappropriate.

Answer: Lots of options here. The basic idea is that a static model assumes the outcome in unaffected by previous values of the explanatory variables and previous values of the outcome. If we think the last year's policing affects this current crime—or if we believe last year's crime level affects this year's crime—then we want a dynamic model.

Q11. Time for a dynamic model. Add the lags of your two explanatory variables (the log of police officers and the log of population). Note: You should have four regressors now (plus the intercept)—the contemporaneous variables *and* their lags. Everything is still logged.

Include your regression results. Interpret the coefficient on the lag of police officers, and comment on its significance.

Answer:

```
# Estimate the regression
reg11 = lm(
  log(n_crime_property) ~
    log(n_officers) + log(pop) +
    lag(log(n_officers)) + lag(log(pop)),
  data = crime_df
)
# Output results
reg11 %>% tidy()
```

```
#> # A tibble: 5 x 5
#>   term                 estimate std.error statistic  p.value
#>   <chr>                   <dbl>     <dbl>     <dbl>    <dbl>
#> 1 (Intercept)             53.7      2.80     19.1   5.46e-18
#> 2 log(n_officers)          0.484    0.214     2.26  3.12e- 2
#> 3 log(pop)                 2.25     3.04      0.742 4.64e- 1
#> 4 lag(log(n_officers))     0.737    0.215     3.42  1.86e- 3
#> 5 lag(log(pop))           -9.48     2.81     -3.37  2.13e- 3
```

The relationship between property crimes and lagged (log) number of officers is positive and statistically significant. The interpretation for the coefficient on the lag of log number of officers: a one-percent increase in the number of police officers **in the previous year** is associated with a 0.74% increase in property crimes (holding all else constant).

Q12. Based upon your regression in **11**, what is the *total effect* of a 1-percent increase in the number of police officers (on property crime)?

Answer: We sum the coefficients on the contemporaneous and lagged number of police officers to find that the total "effect" of a 1-percent increase in the size of the police force is an increase in property crime of approximately 1.22%.

Q13. Should we be worries about reverse causality here? Explain your answer.

Answer: We should be very worried about reverse causality here: It is quite likely that the number of police officers is affected by the amount of crime (and also could affect the amount of crime).

Q14. Use the residuals from your regression model in **11** to test for first-order autocorrelation in the disturbance. Include the steps of your test and clearly state what you conclude from your test.

Hint: You test for autocorrelation in dynamic models with lagged *explanatory variables* just like you test for autocorrelation in static models.

Another hint: Don't forget to add an `NA` when adding the residuals to your dataset (see slide 49/64 from the autocorrelation notes if you don't remember).

Answer:

```
# Add the residuals to our dataset
crime_df$e_reg11 = c(NA, residuals(reg11))
# Regress residuals on their lag (no intercept)
reg14 = lm(e_reg11 ~ -1 + lag(e_reg11), data = crime_df)
# Output regression results
reg14 %>% tidy()
```

```
#> # A tibble: 1 x 5
#>   term          estimate std.error statistic p.value
#>   <chr>            <dbl>     <dbl>     <dbl>   <dbl>
#> 1 lag(e_reg11)    0.461     0.159      2.89 0.00678
```

We find strong, statistically significant evidence of first-order autocorrelation with a p-value of approximately 0.0068, which rejects the null hypothesis of "no first-order autocorrelation" at the 5% level.

Q15. If we have autocorrelation in the regression in **11**, what "problems" does it cause? What are our options for "living with" (or "fixing") the issues caused by autocorrelation?

Answer: Because our model does not include a lagged **outcome** variable, OLS is still unbiased and consistent for the **coefficients** in the presence of autocorrelation. However, the standard errors will be biased (messing up our inference), and OLS will be inefficient. We can attempt to fix the specification, use autocorrelation-robust standard errors, and/or try FGLS.

Q16. Let's make sure our inference in **11** is robust to autocorrelation.

Step 1: Load the `lmtest` and `sandwich` packages, *i.e.*, `p_load(lmtest, sandwich)` (you must have loaded the `pacman` packge to use `p_load()`).
Step 2: Combine the `coeftest()` and `NeweyWest()` functions with your regression output from **11** to get **autocorrelation-robust standard errors** (also called *Newey West* standard errors).

For example, if your regression output in **11** is called `reg11` (the output of `lm()`), then you should run

```
# Load the packages
library(pacman)
p_load(lmtest, sandwich)
# Autocorrelation-robust standard errors
coeftest(reg11, NeweyWest(reg11))
```

Output your results. Does anything change?

Answer: Our coefficients are unchange—this exercise is for our standard errors (and the rest of our inference). The standard errors on the number of police officers decrease (increasing in statistical significance), while the standard errors on the population increase (decreasing in statistical significance).

```
# Load the packages
library(pacman)
p_load(lmtest, sandwich)
# Autocorrelation-robust standard errors
coeftest(reg11, NeweyWest(reg11))
```

```
#>
#> t test of coefficients:
#>
#>                       Estimate Std. Error t value Pr(>|t|)
#> (Intercept)             53.651      3.200   16.76  < 2e-16 ***
#> log(n_officers)          0.484      0.100    4.83  4.0e-05 ***
#> log(pop)                 2.251      5.024    0.45    0.657
#> lag(log(n_officers))     0.737      0.157    4.70  5.8e-05 ***
#> lag(log(pop))           -9.480      4.806   -1.97    0.058 .
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Q17. One way the disturbance can have autocorrelation is when we've omitted a variable that is autocorrelated. Let's now specify our model as an ADL(1,1) (include the lag of the outcome and the lag for both explanatory variables). Keep everything in logs.

Adjust your standard errors to be robust to autocorrelation as we did above in **16**.

Include your regression results. Interpret the coefficient on the lag of logged property crime and comment on its significance.

Note: This regression should have an intercept two explanatory variables, the lags of the two explanatory variables, *and* the lag of the outcome variable.

Answer:

```
# Estimate the regression
reg17 = lm(
  log(n_crime_property) ~
    log(n_officers) + log(pop) +
    lag(log(n_officers)) + lag(log(pop)) +
    lag(log(n_crime_property)),
  data = crime_df
)
# Output results
coeftest(reg17, NeweyWest(reg17))
```

```
#>
#> t test of coefficients:
#>
#>                              Estimate Std. Error t value Pr(>|t|)
#> (Intercept)                   11.6635     2.7585    4.23  0.00023 ***
#> log(n_officers)                0.0610     0.0419    1.46  0.15649
#> log(pop)                      -1.2479     0.3150   -3.96  0.00047 ***
#> lag(log(n_officers))           0.1939     0.0717    2.71  0.01149 *
#> lag(log(pop))                 -0.3561     0.5214   -0.68  0.50025
#> lag(log(n_crime_property))    0.8443     0.0414   20.39  < 2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We find a strong, statistically significant relationship between the log of property crimes and its lag. Specifically, the coefficient tells us that a 1-percent increase in property crimes **last year** is associated with a 0.84% increase in property crime this year.

Q18. Now that we've included the lag of our outcome variable (in **17**), use the residuals for evidence of autocorrelation. Remember that testing models with a lagged outcome variable differs from testing models without a lagged outcome variable.

Report your results/conclusions from the test.

Answer:

```
# Grab the residuals from the regression in 17
crime_df$e_reg17 = c(NA, residuals(reg17))
# The regression for autocorrelation
reg18 = lm(
  e_reg17 ~
    lag(e_reg17) + log(n_officers) + log(pop) +
    lag(log(n_officers)) + lag(log(pop)) +
    lag(log(n_crime_property)),
  data = crime_df
)
coeftest(reg18)
```

```
#>
#> t test of coefficients:
#>
#>                             Estimate Std. Error t value Pr(>|t|)
#> (Intercept)                -0.44200    4.56605   -0.10     0.92
#> lag(e_reg17)               -0.01029    0.22563   -0.05     0.96
#> log(n_officers)            -0.00597    0.11092   -0.05     0.96
#> log(pop)                    0.19371    1.43004    0.14     0.89
#> lag(log(n_officers))        0.00149    0.10871    0.01     0.99
#> lag(log(pop))              -0.13336    1.59095   -0.08     0.93
#> lag(log(n_crime_property))  0.00453    0.08782    0.05     0.96
```

We no longer find statistically significant evidence of autocorrelation: The coefficient on the lag of the residual is small in magnitude (-0.0103) and not statistically significant (p-value of approximately 0.964). We cannot reject the null hypothesis of "no autocorrelation."

Q19. Based upon everything you've seen in this assignment, what is your conclusion? Does a larger police force increase or decrease property crime? Or does it do nothing? Or do you need more data/information? Explain your answer.

Answer: Lots of options here. Looking for a reasonable justification of the specification, concerns about omitted variables, issues with dynamic models, and/or concerns about nonstationarity.

Description of variables and names

Variable	Description
`year`	The year
`pop`	The population of Illinois (in 10,000s)
`unemployment`	The unemployment rate (between 0 and 1)
`n_officers`	The number of police officers employed in Illinois (in 10,000s)
`n_crime_violent`	The number of property crimes in Illinois (in 10,000s)
`n_crime_property`	The number of violent crimes in Illinois (in 10,000s)