Problem Set 3 Solutions

Time Series, Autocorrelation, and Consistency

EC 421: Introduction to Econometrics

Due *before* midnight (11:59pm) on Friday, 29 May 2020

Conceptual Questions

1. Remember that we've discussed three types of time-series models: (**1**) static models, (**2**) dynamic models with lagged explanatory variables, (**3**) dynamic models with lagged outcome variables.

1a. If the disturbance u_t **is not autocorrelated**, for which of the 3 types of models is OLS **unbiased**? If any of the models are biased, explain why.

Answer: If the disturbance **is not autocorrelated**, then OLS is

- **unbiased** for static models
- **unbiased** for dynamic models with **lagged explanatory variables**
- **biased** for dynamic models with **lagged outcome variables** because they violate exogeneity.

1b. If the disturbance u_t **is not autocorrelated**, for which of the 3 types of models is OLS **consistent**? If any of the models are inconsistent, explain why.

Answer: If the disturbance **is not autocorrelated**, then OLS is

- **consistent** for static models
- **consistent** for dynamic models with **lagged explanatory variables**
- **consistent** for dynamic models with **lagged outcome variables**

Note: We need contemporaneous exogeneity for consistency.

1c. If the disturbance u_t **is autocorrelated**, for which of the 3 types of models is OLS **unbiased**? If any of the models are biased, explain why.

Answer: If the disturbance **is not autocorrelated**, then OLS is

- **unbiased** for static models
- **unbiased** for dynamic models with **lagged explanatory variables**
- **biased** for dynamic models with **lagged outcome variables** because they violate exogeneity.

1d. If the disturbance u_t **is autocorrelated**, for which of the 3 types of models is OLS **consistent**? If any of the models are inconsistent, explain why.

Answer: If the disturbance **is not autocorrelated**, then OLS is

- **consistent** for static models
- **consistent** for dynamic models with **lagged explanatory variables**
- **inconsistent** for dynamic models with **lagged outcome variables** because they violate contemporaneous exogeneity

2. In our time-series lecture, we discussed how static time-series models are a pretty restrictive and simplistic way to model time-series data.

2a. Explain why static time-series models are generally restrictive and simplistic.

Answer: Static models assume (1) that all explanatory variables only affect our outcome for exactly one period (the current period) and (2) the outcome variable in the current period is not affected by the outcomes in previous periods. In other words: We are saying that all variables have immediate effects and then no future effects.

This approach to modeling is restrictive because many variables likely have effects for many periods and some outcome variables are affected by previous outcomes.

2b. Give an example of a reasonable **static** time-series model. By *reasonable* we mean that it would be reasonable to model the relationship as a static relationship. Explain why it is reasonable to model the relationship as static rather than dynamic—and make sure you tell us what t would represent (*e.g.*, days, months, years).

Note: The model should look something like $\text{Births}_t = \beta_0 + \beta_1 \text{Income}_t + u_t$

Answer: You have a lot of options here. One example:

$$\text{Yield}_t = \beta_0 + \beta_1 \text{Temperature}_t + \beta_2 \text{Precipitation}_t + \beta_3 (\text{Soil Quality})_t + u_t$$

Imagining that t represents years, we might expect that one year's crop yield mainly depends upon the conditions during its growing season. That said, we still *may* want to consider dynamics even for this model...

2c. Give an example of a reasonable **dynamic** time-series model. By *reasonable* we mean that it would be reasonable to model the relationship as a dynamic relationship. Explain why this relationship should be modeled as a dynamic relationship. Make sure you tell us what t would represent (*e.g.*, days, months, years).

Note: The model should look something like $\text{Births}_t = \beta_0 + \beta_1 \text{Income}_t + \text{Income}_{t-1} + u_t$

Answer: Again, you have a lot of options here. Here's one example of a clearly dynamic model:

$$\text{Population}_t = \beta_0 + \beta_1 \text{Population}_{t-1} + \beta_2 \text{Births}_t + \beta_3 \text{Deaths}_t + \beta_4 \text{Migration}_t + u_t$$

We definitely want population in year t to depend upon the previous year's population—plus the effects of births, deaths, and net migration in the current year.

Note: This model is a bit silly, as it is essentially an accounting exercise: β_0 should be zero, and the rest of the coefficients should be 1. The randomness in the model mainly comes from measurement error.

3. Time-series models frequently include the lag of a variable, *e.g.*, x_{t-1}. Explain why we usually do not use lags in cross-sectional models, *e.g.*, x_{i-1}.

Answer: We typically do not include lags in cross-sectional models because that would say that individual i's outcome (y_i) depends upon individual $i-1$'s explanatory variable $\left(x_{i-1}\right)$. There *are* situations where one person's outcome depends on other people's explanatory variables, but we often ignore this possibility.

In addition: The dataset must be organized in a manner so that individual i is affected by $i-1$'s explanatory variable. Often i is a meaningless index for "individual".

Some Real Data

4. Load packages and your dataset `003-data.csv`.

Answer:

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

5. Which dates does the dataset cover (what are the start and end dates)? How many months?

Answer: The data cover 174 months—starting with 2005-01-01 and ending with 2019-06-01.

6. How many plants retired during this sample?

Answer: 1,051 plants retired during the sample period. `gen_df$cumulative_retirements %>% tail(1)`

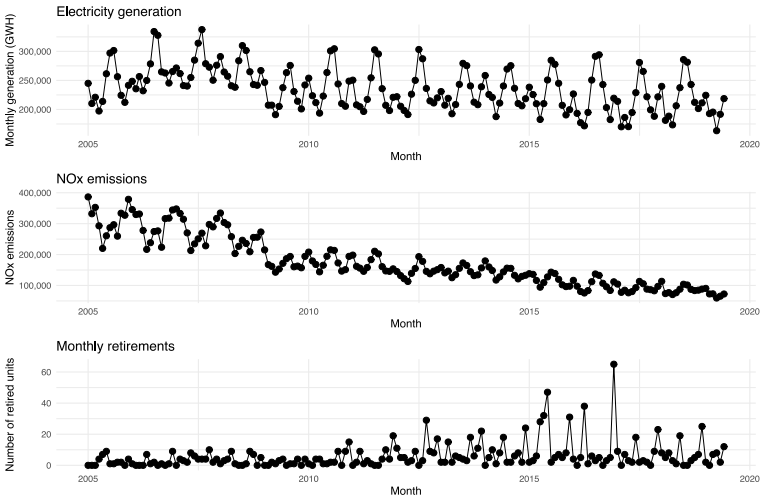
7. Create (and include) **three figures**: (**1**) the time series of total monthly generation (`generation_gwh`), (**2**) the time series of NO_X (Nitrogen Oxide) emissions (`emissions_nox`), and (**3**) the time series for the number of electricity generators who retired in the given month (`n_retirements`).

Hint: A time-series graph has time on the x axis and a variable on the y axis. Your x axis can have either time `t` (time relative to the beginning of the sample) or date (`month`).

Answer:

```
# Figure 1
f7.1 = ggplot(data = gen_df, aes(x = month, y = generation_gwh)) +
geom_line(size = 0.3) +
geom_point(size = 2.5) +
scale_x_date("Month") +
scale_y_continuous("Monthly generation (GWH)", labels = scales::comma) +
ggtitle("Electricity generation") +
theme_minimal()
# Figure 2
f7.2 = ggplot(data = gen_df, aes(x = month, y = emissions_nox)) +
geom_line(size = 0.3) +
geom_point(size = 2.5) +
scale_x_date("Month") +
scale_y_continuous("NOx emissions", labels = scales::comma) +
ggtitle("NOx emissions") +
theme_minimal()
# Figure 3
f7.3 = ggplot(data = gen_df, aes(x = month, y = n_retirements)) +
geom_line(size = 0.3) +
geom_point(size = 2.5) +
scale_x_date("Month") +
scale_y_continuous("Number of retired units", labels = scales::comma) +
ggtitle("Monthly retirements") +
theme_minimal()
# Plot together
f7.1 / f7.2 / f7.3
```

Figures on next page.



8. For each of the three time-series graphs in **7**, explain whether the variable appears to be positively autocorrelated, negatively autocorrelated, or *not* autocorrelated. Make sure you explain your reasoning.

Answer: Each of the time series appears to have positive autocorrelation—especially electricity generation and NO_X emissions. It is likely positive because the level in one month is typically close to the level in the previous month.

9. Estimate a **static** time-series model where monthly *NO_X emissions* (`emissions_nox`) are the outcome variable and our two explanatory variables are the *number of retirements* in the month (`n_retirements`) and the amount of electricity generation in the month (`generation_gwh`).

Report your coefficient estimates and their statistical significance.

Answer:

```
# Estimate the model
model09 = lm(emissions_nox ~ n_retirements + generation_gwh, data = gen_df)
# The results
model09 %>% tidy()
```

```
#> # A tibble: 3 x 5
#>   term            estimate std.error statistic  p.value
#>   <chr>              <dbl>     <dbl>     <dbl>    <dbl>
#> 1 (Intercept)    -73519.   34319.        -2.14 3.36e- 2
#> 2 n_retirements   -1657.     587.        -2.82 5.33e- 3
#> 3 generation_gwh      1.10     0.140      7.81 5.56e-13
```

Based upon the coefficient on `n_retirements`, this simple static model suggests that an additional retirement typically reduced NO_X emissions by 1,657 tons, on average, holding all else constant. The second coefficient suggests that an additional GWh of generation is associated with a 1 ton increase in NO_2 emissions, on average, holding all else constant. Both effects are statistically significant at the 5% level.

10. Now estimate a **dynamic** model in which you include the first lag for each of your explanatory variables (number of retirements and amount of electricity generation). *Note:* You still want the non-lagged version of the variables too—*i.e.*, include x_t *and* x_{t-1}. Interpret the coefficient on the lagged number of retirements.

Answer:

```
# Estimate the model
model10 = lm(
  emissions_nox ~
  n_retirements + lag(n_retirements) + generation_gwh + lag(generation_gwh),
  data = gen_df
)
# The results
model10 %>% tidy()
```

```
#> # A tibble: 5 x 5
#>   term                  estimate std.error statistic     p.value
#>   <chr>                    <dbl>     <dbl>     <dbl>       <dbl>
#> 1 (Intercept)         -78742.    37030.       -2.13  0.0349
#> 2 n_retirements        -1213.      580.       -2.09  0.0382
#> 3 lag(n_retirements)   -1601.      567.       -2.83  0.00529
#> 4 generation_gwh           0.949     0.187     5.08  0.000000980
#> 5 lag(generation_gwh)      0.193     0.195     0.988 0.324
```

The coefficient on the lagged number of retirements—*i.e.*, `lag(n_retirements)`—says that an additional retirement **in the previous month** is associated with reduced NO_x emissions of 1 tons, on average, holding all else constant.

11. Why might it make sense to include lags of the variable *number of retirements*? In other words: Why might we want a dynamic model with lagged explanatory variables in this setting?

Answer: We may want a dynamic model with the lagged number of retirements because the effect of a retirement is likely a long-term, sustained effect: once a plant retires, those emissions may be gone (to some extent) forever.

12. If the disturbance is autocorrelated, what problems does it cause for OLS regression estimates in **10**?

Answer: If **10** has an autocorrelated disturbance, then OLS is inefficient and has biased standard-error estimates.

13. Use the residuals from the regression in **10** to test for first-order autocorrelation in your disturbance. Report the results from the hypothesis test.

Hint: Don't forget about the missing values due to lags (see lecture notes).

Answer:

```
# Add residuals to dataset
gen_df$e10 = c(NA, residuals(model10))
# Regress residuals on their first lag
model13 = lm(e10 ~ -1 + lag(e10), data = gen_df)
# Results
tidy(model13)
```

```
#> # A tibble: 1 x 5
#>   term     estimate std.error statistic  p.value
#>   <chr>       <dbl>     <dbl>     <dbl>    <dbl>
#> 1 lag(e10)    0.870    0.0352      24.7 2.16e-58
```

Our test finds highly significant evidence of first-order autocorrelation.

14. Now estimate a dynamic model (still with NO_X emissions as the outcome variable) with **0, 1, 2, and 3** lags of the **number of retirements** and also the current month's electricity generation (no lags). Interpret the coefficient on the third lag of the number of retirements.

Answer:

```
# Estimate the model
model14 = lm(
  emissions_nox ~
  n_retirements + lag(n_retirements) + lag(n_retirements, 2) +
  lag(n_retirements, 3) + generation_gwh,
  data = gen_df
)
# The results
model14 %>% tidy()
```

```
#> # A tibble: 6 x 5
#>   term                    estimate std.error statistic  p.value
#>   <chr>                      <dbl>     <dbl>     <dbl>    <dbl>
#> 1 (Intercept)           -48727.   30888.        -1.58 1.17e- 1
#> 2 n_retirements          -1135.     519.        -2.18 3.03e- 2
#> 3 lag(n_retirements)     -1398.     513.        -2.72 7.15e- 3
#> 4 lag(n_retirements, 2)  -1362.     514.        -2.65 8.88e- 3
#> 5 lag(n_retirements, 3)  -1591.     509.        -3.13 2.09e- 3
#> 6 generation_gwh             1.07     0.123      8.73 2.63e-15
```

The coefficient on the third lag of the number of retirements tells us that an additional retirement **three months ago**, holding all else constant, is associated with a 1,591 ton reduction in NO_X emissions. This effect is statistically significant at the 5% level.

15. Based upon your estimates in **14**, what is the *total* effect of a retirement on NO_X emissions?

Answer: Based upon the model in **15**, the total effect of a retirement on NO_X emissions, holding all else constant, is a reduction of 5,487 tons of NO_X emissions (the sum of the coefficients).

Note: This estimate essentially assumes that the effect is gone after four months, which is not likely.

16. Now estimate an ADL(1,1) model with NO_X emissions as the outcome and with *number of retirements* and *electricity generation* as the explanatory variables. Report/interpret the coefficient on the lag of NO_X emissions.

Hint: Your regression should have an intercept plus five more terms.

Answer:

```
# Estimate the model
model16 = lm(
  emissions_nox ~ lag(emissions_nox) +
  n_retirements + lag(n_retirements) + generation_gwh + lag(generation_gwh),
  data = gen_df
)
# The results
model16 %>% tidy()
```

Output on next page

```
#> # A tibble: 6 x 5
#>   term                  estimate  std.error statistic  p.value
#>   <chr>                    <dbl>      <dbl>     <dbl>    <dbl>
#> 1 (Intercept)         -19355.    11249.       -1.72   8.72e- 2
#> 2 lag(emissions_nox)       0.925     0.0226   41.0    3.78e-89
#> 3 n_retirements          -13.6     177.       -0.0767 9.39e- 1
#> 4 lag(n_retirements)     -87.9     175.       -0.504  6.15e- 1
#> 5 generation_gwh           0.645     0.0567   11.4    1.47e-22
#> 6 lag(generation_gwh)     -0.512     0.0612   -8.37   2.28e-14
```

The coefficient on the lag of NO_X emissions tells us that a one-ton increase in NO_X emissions in the previous month is associated with a 0.925-ton increase in NO_X emissions in the current month. This relationship is very statistically significant. The relationship says that our outcome is strongly correlated with itself in time.

17. Does it make sense to regress current NO_X emissions on the previous month's emissions? Explain your answer.

Answer: Probably not... though it's difficult. One reason **not** to do it is that last month's emissions are unlikely to actually affect this month's emissions. The emissions are coming from generating electricity—not from last month's emissions.

18. If the disturbance is autocorrelated, then OLS is not consistent for the coefficients in **16**. Explain how you could test for an autocorrelated disturbance using the model from **16**.

Note: You do not actually need to run this test.

Answer: To test for an autocorrelated disturbance in **16**, we want to run a Breusch-Godfrey test, which regresses the residuals from **16** on their lags **and** on the explanatory variables (RHS variables) from **16.**. It then tests the significance of the coefficients on the lagged residuals.

19. Try to find the "best" model for explaining the relationship between monthly NO_X emissions (your outcome variable) and retirements. Include lags, other variables, interactions, logs—whatever you want. Report your final model and explain why you chose it.

Answer: Lots of options here. We're looking for explanation and effort.

20. Return to your figures in **7**: Do any of the three figures suggest a violation of mean stationarity? Explain.

Answer: NO_X emissions appear to violate mean stationarity: The mean is decreasing over time.

Description of Variables

Variable	Description
`t`	Time, relative to the first month of the sample (1, 2, ...)
`month`	Month of the sample (e.g., 2015-12-01)
`generation_gwh`	Total monthly electricity generation (Gigawatt hours, GWh)
`emissions_so2`	Total monthly emissions of SO_2 (in tons)
`emissions_nox`	Total monthly emissions of NO_x (in tons)
`n_plants`	Number of unique electricity-generating units (EGUs) operating in the month
`n_retirements`	Number of retired electricity generating units in the month
`cumulative_retirements`	Cumulative number of retirements (through the given month)
`i_cair`	Binary indicator for months during the Clean Air Interstate Rule (CAIR)
`i_csapr`	Binary indicator for months during the Cross-State Air Pollution Rule (CSAPR)