Problem Set 3

Time Series and Autocorrelation

EC 421: Introduction to Econometrics

Due *before* midnight (11:59pm) on Sunday, 23 May 2021

DUE Your solutions to this problem set are due *before* midnight on Sunday, 23 May 2021. 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 the econometrics topics we reviewed in class; (2) build your R toolset; (3) start building your intuition about causality and time series within econometrics.

Problem 1: Time Series

Imagine that we are interested in estimating the effect of monthly propane prices on monthly natural gas prices. Propane is often used in rural areas where natural gas is difficult to pipe - and so often times energy demand between more remote vs. less remote regions leads to these variables to covary. Let's investigate this phenomenon.

The dataset `ps03_data.csv` contains these prices and also the price of oil—the monthly average oil price (the price in dollars per barrel of *Brent Crude oil*, as measured by the US EIA) and the monthly average price of natural gas (dollars per million BTUs for natural gas at the *Henry Hub*, recorded by the US EIA) and the price of propane(dollars per gallon for propane, recorded by the US EIA)

The table on the last page describes the variables in this dataset.

1a. First, we consider the possibility that P_t^{Gas} (the price of natural gas in month t) only depends upon a constant β_0, P_t^{Propane} (the price of propane in month t), and a random disturbance u_t.

$$P_t^{\text{Gas}} = \beta_0 + \beta_1 P_t^{\text{Propane}} + u_t \tag{1a}$$

If model $(1a)$ is the true model, should we expect OLS to be consistent for β_1? **Explain.**

1b. Read `ps03_data.csv` and estimate model $(1a)$ with OLS. Interpret your estimate for β_1 and comment on its statistical significance.

1c. In (1b), you should have found that the coefficient on P_t^{Propane} is statistically significant. Does this finding also mean that the price of propane explains a lot of the variation in the price of natural gas?

Hint: What is the R^2? (In R, you can find R^2 using `summary()` applied to a model you estimated with `lm()`.)

1d. The model that we estimated in (1a) is a static model—meaning it does not allow previous periods' prices to affect the current price of natural gas. Suppose we think believe that the previous two months' propane prices also affect the price of natural gas,along with current period oil *i.e.*,

$$P_t^{\text{Gas}} = \beta_0 + \beta_1 P_t^{\text{Oil}} + \beta_2 P_t^{\text{Propane}} + \beta_3 P_{t-1}^{\text{Propane}} + \beta_4 P_{t-2}^{\text{Propane}} + u_t \tag{1d}$$

Estimate this model and compare your new estimate for β_2 to your previous estimate (β_1 from model 1a).

Hint: Use the function `lag(x, n)` from the `dplyr` package to take the `n`th lag of variable `x`.

1e. Interpret your estimated coefficients for β_1 and β_3. Are they statistically significant?

1f. Has the amount of variation that we can explain increased very much? Compare the R^2 values for model (1a) and (1d). Also consider the *adjusted* R^2.

continued on next page

1g. Formally test model $(1a)$ vs. model $(1d)$ using an F test.

Hint: You can test one model against another model in R using the `waldtest()` function from the `lmtest` package. For example,

```
# OLS model of y on x and two lags
est_model ← lm(y ~ x + z + lag(x) + lag(x, 2), data = example_df)
# Jointly test the coefficients on z and lag(x, 2)
waldtest(est_model, c("z", "lag(x, 2)"), test = "F")
```

calculates an F test for the coefficients on `z` and `lag(x, 2)` in the model `est_model`.

Note: For some reason, `lag(x, n)` needs to have a space between the comma (`,`) and `n` when you use `waldtest` to test lags.

1h. If model $(1d)$ is the true model, should we expect OLS to be consistent for β_1? **Explain.**

1i. Suppose we now think that the actual model includes the current price of propane *and* the previous two months' prices of propane and the previous month of natural gas prices, *i.e.*,

$$P_t^{\text{Gas}} = \beta_0 + \beta_1 P_t^{\text{Propane}} + \beta_2 P_{t-1}^{\text{Propane}} + \beta_3 P_{t-2}^{\text{Propane}} + \beta_4 P_{t-1}^{\text{Gas}} + u_t \tag{1i}$$

Estimate this model. Interpret the coefficients on β_1 and β_3. How has your estimate on β_1 changed?

1j. Compare the R^2 from model $(1i)$ to the R^2s of the previous models. Explain what happened.

1k. If we assume u_t in $(1i)$ **A** follows our assumption of *contemporaneous exogeneity* and **B** is not autocorrelated, should we expect OLS to produce consistent estimates for the βs in this model? **Explain.**

2a. After starting to estimate these time-series models, you remember that autocorrelation affects OLS. For each of the three models above (1a, 1d, and 1i), explain how autocorrelation will affect OLS.

Hint: It will affect two of the models the same way and one of them differently.

2b. Add the residuals from your estimate of model $(1i)$ to your dataset.

Important: Don't forget that you will need to tell R that you have a missing observation (since we have a lag in our model).

```
# Add residuals from our estimated model in 1i to dataset 'price_df'
price_df$e_1i ← c(NA, NA, residuals(ols_1i))
```

Here, I'm adding a new column to the dataset `price_df` for the residuals from the model I saved as `ols_1i`. The first observation is missing, because our model `ols_1i` includes a single lag.

2c. Construct two plots with the residuals from $(1i)$: **1** plot the residuals against the time variable (`t_month`) and **2** plot the residuals against their lag. Do you see any evidence of autocorrelation? What would autocorrelation look like?

I strongly encourage you to use `ggplot2` for these graphs.

2d. Add the residuals from the model in $(1d)$ to your dataset. See below (we have to keep track of missing observations due to lags).

```
# Residuals from the model in 1a
price_df$e_1a ← residuals(ols_1a)
# Residuals from the model in 1d
price_df$e_1d ← c(NA, NA, residuals(ols_1d))
```

2e. Repeat the plots from above—**1** plot the residuals against the time variable (`t_month`) and **2** plot the residuals against their lag, *i.e.*, for the residuals from $(1d)$. You should end up with two graphs for this part. Interpret your graphs and comment on whether you think there may be some autocorrelation for this model.

2f. Why do you think the residuals from $(1d)$ appear to have autocorrelation, while the residuals in $(1i)$ show much less evidence of autocorrelation?

Hint: Think back to our discussion of the ways we can work/live with autocorrelation.

2g. Following the steps for the Breusch-Godfrey test that we discussed in class, test the residuals from the model in $(1i)$ for second-order autocorrelation.

Hint: You can use the `waldtest()` from the `lmtest` package, as shown in the lecture slides.

2h. If we assume u_t is **not** autocorrelated, then can we trust OLS to be consistent for its estimates of the coefficients in model $(1i)$? **Explain.**

Description of variables and names

Variable	**Description**
`month_year`	The observation's month and year (`character`)
`price_gas`	The average (Henry Hub) price of natural gas, $ per 1MM BTU (`numeric`)
`price_oil`	The average (Brent Crude) price of oil, $ per barrel
`price_prop`	The average Retail/Resale price of propane, $ per gallon (`numeric`)
`month`	Month of Observation (`numeric`)
`year`	Year of Observation (`numeric`)
`t_month`	Time, measured by months in the dataset (`numeric`)
`t`	Time, approximately by fractions of years (`numeric`)