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.**

ANSWER The model in (1a) is a *static time-series* model—there are no lags of the explanatory or dependent variables. OLS provides consistent estimates for the β_j in static time-series models.

Note: We're ignoring nonstationarity.

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

ANSWER

```r
# Load packages
library(pacman)
p_load(tidyverse, broom, here)
# Load data
price_df <- read_csv("ps03_data.csv")
# Estimate model 1a with OLS
ols_1a <- lm(price_gas ~ price_prop, data = price_df)
# Results
tidy(ols_1a)
```

```
#> # A tibble: 2 x 5
#>   term        estimate std.error statistic  p.value
#>   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
#> 1 (Intercept)     1.04     0.379      2.75 6.60e- 3
#> 2 price_prop      1.66     0.203      8.16 5.09e-14
```

Our estimate for β_1 is approximately 1.656, and it is statistically significant at the 5-percent level. *Interpretation:* Holding all else constant, if the price of propane increases by 1 dollar, we expect the price of natural gas to increase by 1.656 in that same period.

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()`.)

ANSWER Our model in (1a) has an R^2 of approximately 0.267, which suggests that the price of oil explains a moderate amount of the variation in the price of natural gas, despite the fact that the correlation between the two variables is statistically significant. Statistical significance does not tell us whether the variable explains a substantial amount of variation.

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`.

ANSWER

```
# Estimate model 1a with OLS
ols_1d ← lm(
  price_gas ~ price_oil + price_prop + lag(price_prop, 1) + lag(price_prop, 2),
  data = price_df
)
# Results
tidy(ols_1d)
```

```
#> # A tibble: 5 x 5
#>   term               estimate std.error statistic p.value
#>   <chr>                 <dbl>     <dbl>     <dbl>   <dbl>
#> 1 (Intercept)          1.18       0.422     2.80  0.00571
#> 2 price_oil           -0.0970     0.448    -0.217 0.829
#> 3 price_prop           2.49       1.02      2.46  0.0150
#> 4 lag(price_prop, 1)   1.56       1.36      1.15  0.252
#> 5 lag(price_prop, 2)  -2.36       0.937    -2.52  0.0126
```

After controlling for the the first and second lags of the price of propane, our estimate for β_1 is now approximately 2.494 (we previously estimated 1.656). The point estimate is now a bit larger (but no longer statistically significant).

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

ANSWER Our estimates for β_1 and β_3 are -0.097 and 1.564, respectively. *Interpretation:* $\hat{\beta}_1$ implies that when the contemporaneous price of oil is increased by 1 dollar, the price of natural gas increases by approximately -0.097 . $\hat{\beta}_3$ suggests that when last months' propane price increased by one dollar, we expect this month's price for natural gas to increase by approximately 1.564 (holding all else constant). Neither estimate is statistically significant at the 5-percent level.

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

ANSWER Nope—we still explaining a moderate amount of variation in the price of natural gas. The R^2 has only increased slightly (it's now 0.296), as has adjusted R^2 (now 0.28). Technically though, we cannot use R^2 to compare these models (see below).

In cases like these, R^2 can actually decrease between models when we add more lagged variables. This may seem a bit weird given what you know about R^2, but, REMEMBER, when we include lags in a model we actually throw away a few observations in the first few time periods (because we never see $x_{t=-1}$), which means the two values for R^2 are not technically comparable since they are run on marginally different datasets.

1g. Formally test model $(\mathrm{1a})$ vs. model $(\mathrm{1d})$ 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.

ANSWER The test via `waldtest`...

```
# Load 'lmtest'
p_load(lmtest)
# F test
waldtest(ols_1d, c("price_oil", "lag(price_prop, 1)", 'lag(price_prop, 2)'))
```

```
#> Wald test
#>
#> Model 1: price_gas ~ price_oil + price_prop + lag(price_prop, 1) + lag(price_prop,
#>     2)
#> Model 2: price_gas ~ price_prop
#>   Res.Df Df    F Pr(>F)
#> 1    178
#> 2    181 -3 2.83   0.04 *
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The F test comparing the two models rejects the null hypothesis at the 5% level with a *p*-value of approximately 0.040. In this test, H$_0$ is $\beta_1 = 0$, $\beta_3 = 0$ and $\beta_4 = 0$ (for model (1d)). Thus, we find statistically significant evidence that the price of oil contemporaneously affects the current price of natural gas, and lagged price of propane also has a significant effect on the price of gas (after controlling for the current price of propane and the twice-lagged price of propane).

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

ANSWER The model in (1d) only includes lags of the explanatory variable, which means we can expect OLS to be consistent for β_1 (even if u_t is autocorrelated).

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?

ANSWER

```
# Estimate model 1a with OLS
ols_1i <- lm(
  price_gas ~ price_prop + lag(price_prop, 1) + lag(price_prop,2)+ lag(price_gas, 1),
  data = price_df
)
# Results
tidy(ols_1i)
```

```
#> # A tibble: 5 x 5
#>   term               estimate std.error statistic  p.value
#>   <chr>                 <dbl>     <dbl>     <dbl>    <dbl>
#> 1 (Intercept)           0.206    0.176       1.17 2.43e- 1
#> 2 price_prop            3.12     0.410       7.61 1.50e-12
#> 3 lag(price_prop, 1)   -1.64     0.624      -2.63 9.37e- 3
#> 4 lag(price_prop, 2)   -1.33     0.414      -3.21 1.58e- 3
#> 5 lag(price_gas, 1)     0.878    0.0332     26.5  1.19e-63
```

Our estimate for β_1 is now approximately 3.118, which is statistically significant at the 5-percent level. This value is almost exactly what we estimated in (1d). The interpretation of this effect is that we expect a 1-dollar increase in the current month's price of propane to increase the the price of natural gas in the current month by approximately 3.118—holding all else constant.

Our estimate for β_3 is approximately -1.329, which is also statistically significant at the 5-percent level. The interpretation of this effect is that we expect a 1-dollar increase in the previous month's price of natural gas to increase the the price of natural gas in the current month by approximately -1.329—holding all else constant.

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

ANSWER The R^2 in the current model (1i) is now approximately 0.8574, which is **much** higher than the R^2 values we saw in the previous two models. It appears as though the price of natural gas is very strongly correlated with the previous month's price of natural gas: once we control for one lag of the price of natural, we are able to account for a substantial amount of the variation in the price of natural gas.

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.**

ANSWER Yes. OLS is consistent for models with lagged dependent variables as long as the disturbances follow our assumptions of contemporaneous exogeneity and no autocorrelation.

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.

ANSWER For models (1a) and (1d), autocorrelated disturbances will cause OLS to be inefficient with biased standard errors, but OLS will still be unbiased and consistent for the coefficients in (1a) and (1d).

In models like (1i), autocorrelation causes a violation of our contemporaneous exogeneity assumption, which causes OLS to be biased and inconsistent for estimating the coefficients in the model.

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.

ANSWER Done in hint.

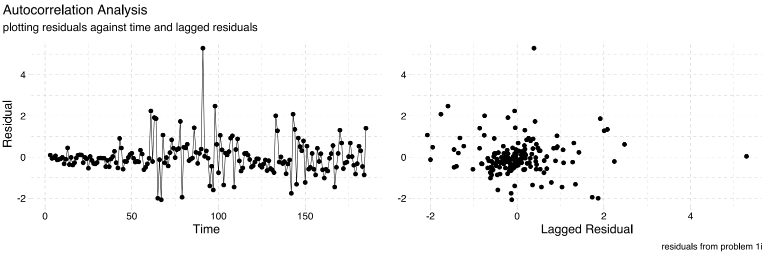
2c. Construct two plots with the residuals from (1i): **1** plot the residuals against the time variable (`t_month`) and **2** plot the residuals against their lag. Do you see any evidence of autocorrelation? What would autocorrelation look like?

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

ANSWER First, let's plot the residuals over time. Then, we'll plot the residuals against their lags.

```
# Load 'ggplot2' and 'ggthemes' packages
p_load(ggplot2, ggthemes, patchwork)
# Plot 1: Residuals over time
plot_1 = ggplot(data = price_df, aes(x = t_month, y = e_1i)) +
geom_path(size = 0.1) +
geom_point() +
xlab("Time") + ylab("Residual") +
theme_pander()
# Plot 2: Residuals against their lags
plot_2 = ggplot(data = price_df, aes(x = lag(e_1i), y = e_1i)) +
geom_point() +
xlab("Lagged Residual") + ylab("") +
theme_pander()

plots = plot_1 + plot_2
plots + plot_annotation(title = 'Autocorrelation Analysis', subtitle = 'plotting residuals aga
```



Autocorrelation Analysis

plotting residuals against time and lagged residuals

residuals from problem 1i

Neither figure really bears strong evidence of autocorrelation. In the first figure, we would expect to see larger residuals followed by other large residuals. We might see a little of this trend, but it isn't obvious. In the second figure, autocorrelation would look show up with residuals forming some sort of line. No line emerges.

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))
```

ANSWER Done in hint.

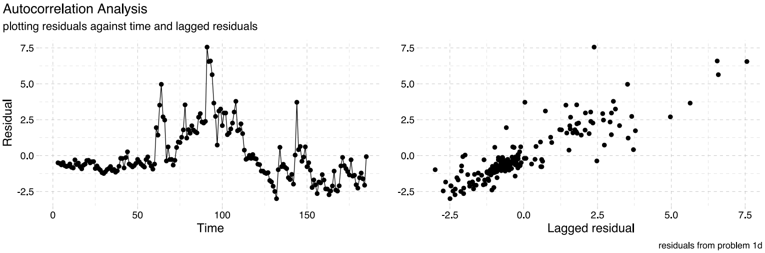
2e. Repeat the plots from above—**1** plot the residuals against the time variable (`t_month`) and **2** plot the residuals against their lag, *i.e.*, for the residuals from (1d). You should end up with two graphs for this part. Interpret your graphs and comment on whether you think there may be some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

ANSWER First, let's examine the residuals from model (1a)

let's consider the residuals from (1d)

```
# Plot 1d 1: Residuals over time
plot1 = ggplot(data = price_df, aes(x = t_month, y = e_1d)) +
geom_path(size = 0.3) +
geom_point() +
xlab("Time") + ylab("Residual") +
theme_pander()
# Plot 1d 2: Residuals against their lags
plot2 = ggplot(data = price_df, aes(x = lag(e_1d), y = e_1d)) +
geom_point() +
xlab("Lagged residual") + ylab("") +
theme_pander()

plots = plot1 + plot2
plots + plot_annotation(title = 'Autocorrelation Analysis', subtitle = 'plotting residuals aga
```



All of these figures are strongly suggestive of autocorrelation in our residuals—specifically of positive autocorrelation, as one period's residual tends to be very similar to the residuals in the previous and next periods. We also see strong linear relationships when plotting residuals against their lags.

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.

ANSWER Model misspecification can cause autocorrelation in the disturbance if an omitted variable is, itself, autocorrelated. In this case, if the current price of natural gas depends strongly on the previous period's price of natural gas, then if we fail to control/include the previous period's price of natural gas (as we do in (1a) and (1d)), then the previous period's price of natural gas shows up in the disturbance/residual, which is likely causing at least some of the 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.

ANSWER Because (1i) includes a lagged outcome variable, we use the **Breusch-Godfrey** test. We already completed the **first step** (estimating the model with OLS) and the **second step** (recording the residuals).

The **third step** involves regressing the residuals on the original explanatory variables and lags of the residuals (here: 2 lags).

```
# Regress residuals on explanatory variables and two lags of residuals
price_df$e_1i <- c(NA, NA, residuals(ols_1i))

bg_2g <- lm(
  e_1i ~ price_prop + lag(price_prop, 1) + lag(price_prop,2) + lag(price_gas, 1) + lag(e_1i, 1)
  data = price_df
)
# F test
waldtest(bg_2g, c("lag(e_1i, 1)", "lag(e_1i, 2)"))
```

```
#> Wald test
#>
#> Model 1: e_1i ~ price_prop + lag(price_prop, 1) + lag(price_prop, 2) +
#>     lag(price_gas, 1) + lag(e_1i, 1) + lag(e_1i, 2)
#> Model 2: e_1i ~ price_prop + lag(price_prop, 1) + lag(price_prop, 2) +
#>     lag(price_gas, 1)
#>   Res.Df Df    F Pr(>F)
#> 1    174
#> 2    176 -2 0.72   0.49
```

The **fourth step** invovles an F test for the two lags. The F test above has a *p*-value of approximately 0.49, which means we fail to reject H_0 as the 5-percent level.

In the **fifth step**, we make our conclusion. Here, H_0 is "no autocorrelation". Thus, we fail to reject "no autocorrelation"—meaning we did not find statistically significant evidence of autocorrelation for model (1i).

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.**

ANSWER Because model (1i) has a lagged outcome variable, we can trust OLS to consistently estimate the coefficients in (1i) *if* there is not autocorrelation in the disturbances u_t (as long as there are no other violations of our assumptions).

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`)