# Problem Set 3

## Time Series and Autocorrelation

**EC 421:** Introduction to Econometrics

Due *before* midnight (11:59pm) on Saturday, 03 March 2019

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 oil prices on monthly natural gas prices. The dataset `gas_oil.csv` contains these prices—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).

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

**1a.** First, we consider the possibility that $P_t^{\text{Gas}}$ (the price of natural gas in month $t$) only depends upon a constant $\beta_0$, $P_t^{\text{Oil}}$ (the price of oil in month $t$), and a random disturbance $u_t$.

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

If model (1a) is the true model, should we expect OLS to be consistent for $\beta_1$? **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 $\beta_j$ in static time-series models.

*Note:* We're ignoring nonstationarity.

**1b.** Read `gas_oil.csv` and estimate model (1a) with OLS. Interpret your estimate for $\beta_1$ and comment on its statistical significance.

**ANSWER**

```
# Load packages
library(pacman)
p_load(tidyverse, broom, here)
# Load data
price_df ← read_csv("gas_oil.csv")
# Estimate model 1a with OLS
ols_1a ← lm(price_gas ~ price_oil, 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)     3.34    0.264      12.6  6.08e-29
## 2 price_oil     0.0175  0.00396       4.41 1.50e- 5
```

Our estimate for $\beta_1$ is approximately 0.017, and it is statistically significant at the 5-percent level. *Interpretation:* Holding all else constant, if the price of oil increases by 1 dollar, we expect the price of natural gas to increase by 0.017.

**1c.** In (1b), you should have found that the coefficient on $P_t^{\text{Oil}}$ is statistically significant. Does this finding also mean that the price of oil 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 0.0689, which suggests that the price of oil explains a fairly small 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' oil prices also affect the price of natural gas, *i.e.*,

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

Estimate this model and compare your new estimate for $\beta_1$ to your previous estimate (from model 1a).

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

**ANSWER**

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

```
## # A tibble: 4 x 5
##   term              estimate std.error statistic  p.value
##   <chr>                <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)        3.37        0.269    12.6    1.52e-28
## 2 price_oil          0.0260      0.0267    0.975  3.31e- 1
## 3 lag(price_oil, 1)  0.00232     0.0437    0.0532 9.58e- 1
## 4 lag(price_oil, 2) -0.0115      0.0267   -0.431  6.67e- 1
```

After controlling for the the first and second lags of the price of oil, our estimate for $\beta_1$ is now approximately 0.026 (we previously estimated 0.017). The point estimate is now a bit larger (but no longer statistically significant).

**1e.** Interpret your estimated coefficients for $\beta_2$ and $\beta_3$. Are they statistically significant?

**ANSWER** Our estimates for $\beta_2$ and $\beta_3$ are 0.002 and -0.011, respectively. *Interpretation:* $\hat{\beta}_2$ suggests that when last months' oil price increased by one dollar, we expect this month's price for natural gas to increase by approximately 0.002 (holding all else constant). Similarly, $hat\beta_3$ suggests that when 2 months' prior price of oil increased by one dollar, we expect this month's price for natural gas to increase by approximately -0.011 (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* $R^2$.

**ANSWER** Nope—we still are not explaining much variation in the price of natural gas. The $R^2$ has actually decreased slightly (it's now 0.068), as has adjusted $R^2$ (now 0.058).

(At first the fact that $R^2$ actually decreased when we added more variables seem a bit weird, but we lost a few observations due to the lags in model (1d), which means the two values for $R^2$ are not technically comparable.)

**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 + lag(x) + lag(x, 2), data = example_df)
# Jointly test the coefficients on lag(x) and lag(x, 2)
waldtest(est_model, c("lag(x)", "lag(x, 2)"), test = "F")
```

calculates an $F$ test for the coefficients on `lag(x)` 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("lag(price_oil, 1)", "lag(price_oil, 2)"))
```

```
## Wald test
##
## Model 1: price_gas ~ price_oil + lag(price_oil, 1) + lag(price_oil, 2)
## Model 2: price_gas ~ price_oil
##   Res.Df Df      F Pr(>F)
## 1    259
## 2    261 -2 0.2342 0.7914
```

The $F$ test comparing the two models fails to reject the null hypothesis at the 5% level with a *p*-value of approximately 0.79. In this test, $H_0$ is $\beta_2 = 0$ and $\beta_3 = 0$ (for model (1d)). Thus, we fail to find statistically significant evidence that the first and second lags of oil price affects the current price of natural gas (after controlling for the current price of oil).

**1h.** If model (1d) is the true model, should we expect OLS to be consistent for $\beta_1$? **Explain.**

**ANSWER** The model in (1d) only includes lags of the explanatory variable, which means we can expect OLS to be consistent for $\beta_1$ (even if $u_t$ is autocorrelated).

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

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

Estimate this model. Interpret the coefficients on $\beta_1$ and $\beta_3$. How has your estimate on $\beta_1$ changed?

**ANSWER**

```
# Estimate model 1a with OLS
ols_1i ← lm(
  price_gas ~ price_oil + lag(price_oil, 1) + lag(price_gas, 1),
  data = price_df
)
# Results
tidy(ols_1i)
```

```
## # A tibble: 4 x 5
##   term             estimate std.error statistic   p.value
##   <chr>               <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)        0.220     0.118      1.87 6.30e-  2
## 2 price_oil          0.0263   0.00866     3.03 2.65e-  3
## 3 lag(price_oil, 1) -0.0254   0.00865    -2.94 3.54e-  3
## 4 lag(price_gas, 1)  0.937    0.0216     43.3  6.11e-121
```

Our estimate for $\beta_1$ is now approximately 0.026, 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 oil to increase the the price of natural gas in the current month by approximately 0.026—holding all else constant.

Our estimate for $\beta_3$ is approximately 0.937, 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 0.937—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 0.8868018, 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 $\beta$s 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.

# Problem 2: 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, 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)
# Plot 1: Residuals over time
ggplot(data = price_df, aes(x = t_month, y = e_1i)) +
geom_path(size = 0.3) +
geom_point() +
xlab("Time") + ylab("Residual (1i)") +
theme_pander()
```
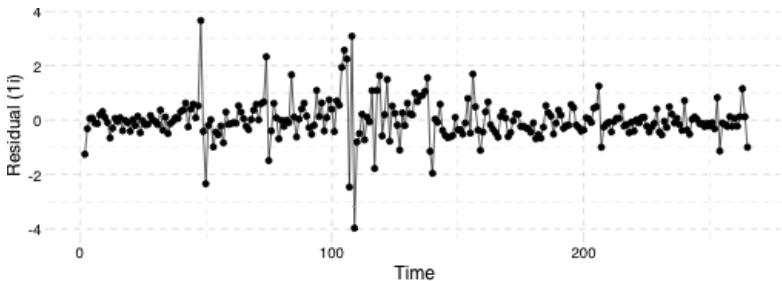


```
# Plot 2: Residuals against their lags
ggplot(data = price_df, aes(x = lag(e_1i), y = e_1i)) +
geom_point() +
xlab("Lagged residual (1i)") + ylab("Residual (1i)") +
theme_pander()
```
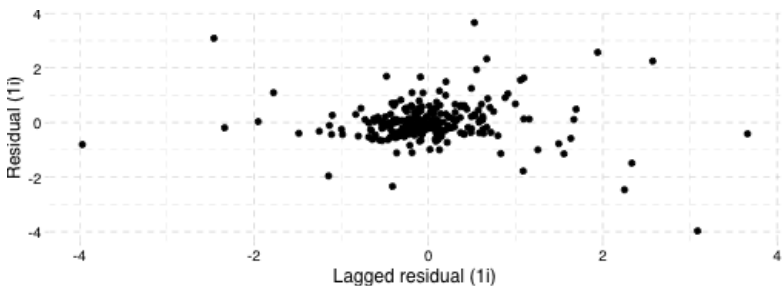


Neither figure really bears strong evidence of autocorrelation. In the first figure, we would expect to see larger residuals followed by other large residuals. We might see a little of this trend, but it isn't obvious. In the second figure, autocorrelation would look show up with residuals forming some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

**2d.** Add the residuals from the models in (**1a**) and (**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—for both sets of residuals, *i.e.*, for the residuals from (**1a**) and for the residuals from (**1d**). You should end up with four graphs for this part.

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

```
# Plot 1a 1: Residuals over time
ggplot(data = price_df, aes(x = t_month, y = e_1a)) +
geom_path(size = 0.3) +
geom_point() +
xlab("Time") + ylab("Residual (1a)") +
theme_pander()
```
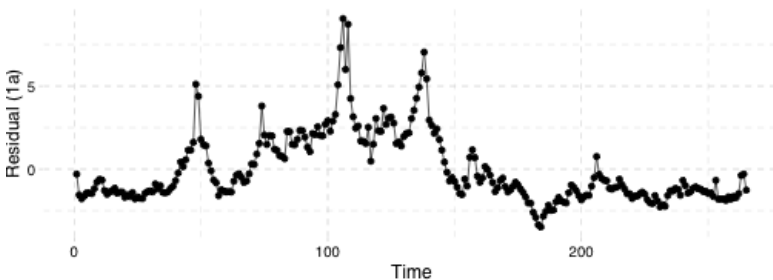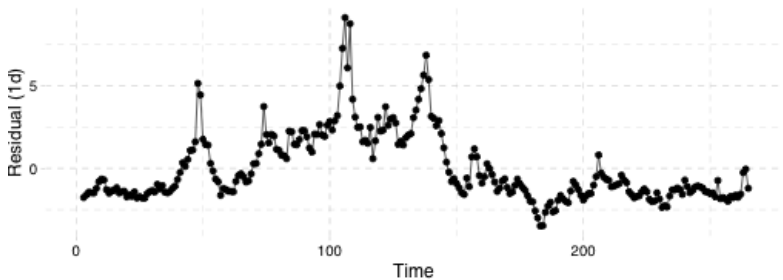


```
# Plot 1a 2: Residuals against their lags
ggplot(data = price_df, aes(x = lag(e_1a), y = e_1a)) +
geom_point() +
xlab("Lagged residual (1a)") + ylab("Residual (1a)") +
theme_pander()
```
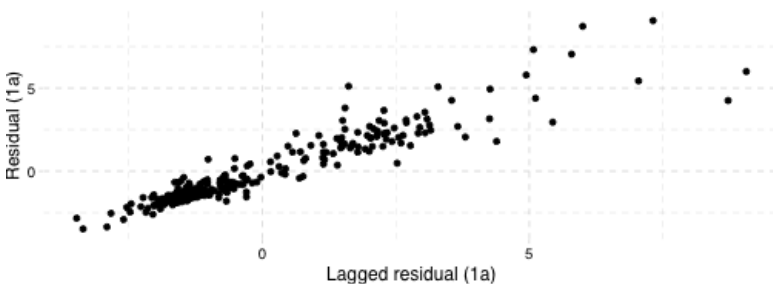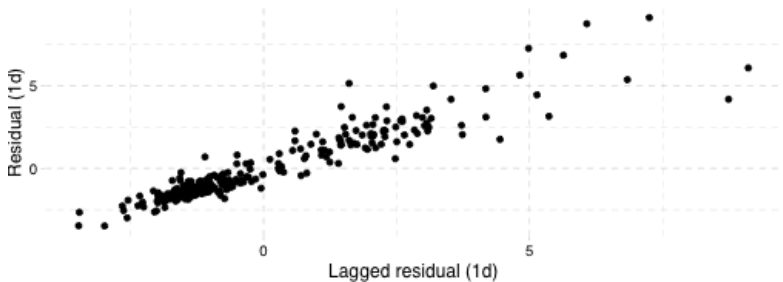
Now, let's consider the residuals from (1d)

```
# Plot 1d 1: Residuals over time
ggplot(data = price_df, aes(x = t_month, y = e_1d)) +
geom_path(size = 0.3) +
geom_point() +
xlab("Time") + ylab("Residual (1d)") +
theme_pander()
```



```
# Plot 1d 2: Residuals against their lags
ggplot(data = price_df, aes(x = lag(e_1d), y = e_1d)) +
geom_point() +
xlab("Lagged residual (1d)") + ylab("Residual (1d)") +
theme_pander()
```



*Not necessary for full points:* 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 pervious and next periods. We also see strong linear relationships when plotting residuals against their lags.

**2f.** Why do you think the residuals from (1a) and (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
bg_2g ← lm(
  e_1i ~ price_oil + lag(price_oil, 1) + lag(price_gas, 1) + lag(e_1i, 1) + lag(e_1i, 2),
  data = price_df
)
# F test
waldtest(bg_2g, c("lag(e_1i, 1)", "lag(e_1i, 2)"))
```

```
## Wald test
##
## Model 1: e_1i ~ price_oil + lag(price_oil, 1) + lag(price_gas, 1) + lag(e_1i,
##      1) + lag(e_1i, 2)
## Model 2: e_1i ~ price_oil + lag(price_oil, 1) + lag(price_gas, 1)
##   Res.Df Df      F Pr(>F)
## 1    256
## 2    258 -2 0.1814 0.8342
```

The **fourth step** invovles an $F$ test for the two lags. The $F$ test above has a *p*-value of approximately 0.83, 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).

**2i.** Should we interpret our estimates from (1i) as causal? **Explain.**

**ANSWER** Probably not. It is not even clear which way the causal relationship would go—do natural gas prices influence oil prices, do oil prices influence natural gas prices, or do they both influence each other? There are definitely omitted variables—variables that affect the prices of both natural gas and oil. (We also probably want to think about nonstationarity.)

# Description of variables and names

| Variable | Description |
|----------|-------------|
| month_year | The observation's month and year (`character`) |
| month | The month (`numeric`) |
| year | The year (`numeric`) |
| 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 (`numeric`) |
| t_month | Time, measured by months in the dataset (`numeric`) |
| t | Time, approximately by fractions of years (`numeric`) |