Problem Set 3 Solutions

Time Series and Autocorrelation

EC 421: Introduction to Econometrics

Due *before* midnight (11:59pm) on Wednesday, 29 May 2019

`DUE` Your solutions to this problem set are due *before* midnight on Wednesday, 29 May 2019. Your files must be uploaded to Canvas.

`IMPORTANT` Your submission must include (1) **your responses/answers to the question in a PDF, Word, or similar file** and (2) the `R` script you used to generate your answers. **The `R` script is just for your code. To receive credit, your answers/figures/*etc.* must be in the PDF/Word document.** 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 oil prices on monthly natural gas prices. The dataset `ps03_data.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^{Oil} (the price of oil in month t) only depends upon a constant β_0, P_t^{Gas} (the price of natural gas in month t), and a random disturbance u_t.

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

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 is consistent for estimating the β_j in static time-series models. (As long as there are no omitted variables, which may be assumed by the term "true model".)

Note: We're ignoring nonstationarity.

1b. Read `ps03_data.csv` and summarize them.

How many observations do you have? Which months/years do they cover? (*Hint:* use `nrow()`, `head()`, and `tail()`).

Now estimate model (1a) with OLS. Interpret your estimate for β_1 and comment on its statistical significance.

Answer

```
# Load packages
library(pacman)
p_load(tidyverse, broom, here)
# Load data
price_df <- read_csv("ps03_data.csv")
# Number of observations
nrow(price_df)
```

```
## [1] 268
```

```
# Summary of data
summary(price_df)
```

```
##    month_year           price_gas        price_oil          month
##  Min.   :1997-01-01   Min.   : 1.720   Min.   :  9.82   Min.   : 1.00
##  1st Qu.:2002-07-24   1st Qu.: 2.815   1st Qu.: 28.20   1st Qu.: 3.00
##  Median :2008-02-15   Median : 3.695   Median : 54.73   Median : 6.00
##  Mean   :2008-02-15   Mean   : 4.331   Mean   : 58.13   Mean   : 6.44
##  3rd Qu.:2013-09-08   3rd Qu.: 5.415   3rd Qu.: 77.34   3rd Qu.: 9.00
##  Max.   :2019-04-01   Max.   :13.420   Max.   :132.72   Max.   :12.00
##       year         t_month             t
##  Min.   :1997   Min.   :  1.00   Min.   :1997
##  1st Qu.:2002   1st Qu.: 67.75   1st Qu.:2003
##  Median :2008   Median :134.50   Median :2008
##  Mean   :2008   Mean   :134.50   Mean   :2008
##  3rd Qu.:2013   3rd Qu.:201.25   3rd Qu.:2014
##  Max.   :2019   Max.   :268.00   Max.   :2019
```

```
# Estimate model 1a with OLS
ols_1a <- lm(price_oil ~ price_gas, 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)    41.3      4.31       9.58 7.18e-19
## 2 price_gas       3.89     0.888      4.38 1.71e- 5
```

We have 268 observations, staring in January of 1997 and running until April of 2019.

Our estimate for β_1 in equation (1a) is approximately 3.891, and it is statistically significant at the 5-percent level. *Interpretation:* Holding all else constant, if the price of natural gas increases by 1 dollar, we expect the price of oil to increase by 3.891.

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

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 0.0673, 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 oil. Suppose we think believe that the previous two months' natural gas prices also affect the price of oil, *i.e.*,

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

Estimate this model and compare your new estimate for β_1 to your previous estimate (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 1d with OLS
ols_1d <- lm(
  price_oil ~ price_gas + lag(price_gas, 1) + lag(price_gas, 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)          40.9       4.44     9.21  1.05e-17
## 2 price_gas             1.69      2.62     0.644 5.20e- 1
## 3 lag(price_gas, 1)     1.23      3.58     0.343 7.32e- 1
## 4 lag(price_gas, 2)     1.10      2.61     0.422 6.73e- 1
```

After controlling for the the first and second lags of the price of natural gas, our estimate for β_1 is now approximately 1.688 (we previously estimated 3.891). The point estimate is smaller and no longer statistically significant.

1e. Interpret your estimated coefficients for β_2 and β_3. Are they statistically significant?

Answer Our estimates for β_2 and β_3 are 1.231 and 1.101, respectively.

Interpretations: $\hat{\beta}_2$ suggests that when last months' natural gas price increased by one dollar, we expect this month's price for oil to increase by approximately 1.231 (holding all else constant). Similarly, $\hat{\beta}_3$ suggests that when 2 months' prior price of natural gas increased by one dollar, we expect this month's price of oil to increase by approximately 1.101 (holding all else constant). However, 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 **increased** very slightly (it's now 0.068; it was 0.067). The adjusted R^2 has decreased (now 0.058; was 0.064).

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 Wald test...

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

```
## Wald test
##
## Model 1: price_oil ~ price_gas + lag(price_gas, 1) + lag(price_gas, 2)
## Model 2: price_oil ~ price_gas
##   Res.Df Df      F Pr(>F)
## 1    262
## 2    264 -2 0.4621 0.6305
```

The F test comparing the two models fails to reject the null hypothesis at the 5% level with a p-value of approximately 0.63. 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 natural gas affects the current price of oil (after controlling for the current price of natural gas).

1h. If model $(1d)$ 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. (Again, we also must assume no omitted variables, which may be assumed by the term "true model".)

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

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

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

Answer

```
# Estimate model 1i with OLS
ols_1i <- lm(
  price_oil ~ price_gas + lag(price_gas, 1) + lag(price_oil, 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)          0.680    0.829      0.820 4.13e-  1
## 2 price_gas            1.27     0.428      2.97  3.21e-  3
## 3 lag(price_gas, 1)   -1.18     0.429     -2.74  6.48e-  3
## 4 lag(price_oil, 1)    0.984    0.0101    97.5   1.78e-208
```

Our estimate for β_1 is now approximately 1.272, which is statistically significant at the 5-percent level. This value is a bit smaller than what we estimated in (1d)—and much smaller than the estimate from (1a). The interpretation of this effect is that we expect a 1-dollar increase in the current month's price of natural gas to increase the the price of oil in the current month by approximately 1.272—holding all else constant.

Our estimate for β_3 is approximately 0.984, 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 oil to increase the the price of oil in the current month by approximately 0.984—holding all else constant.

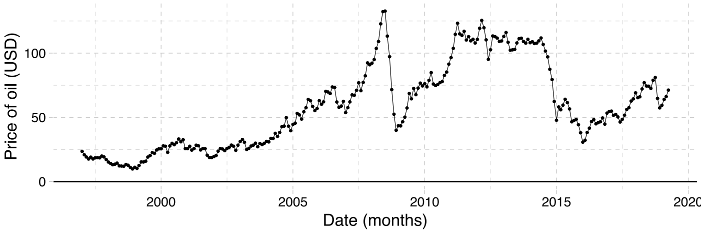
1j. Compare the R^2 from model (1i) to the R^2s of the previous models. Explain what happened.

Answer The R^2 in the current model (1i) is now approximately 0.9749, which is **much** higher than the R^2 values we saw in the previous two models. It appears as though the price of oil is very strongly correlated with the previous month's price of oil: 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 oil.

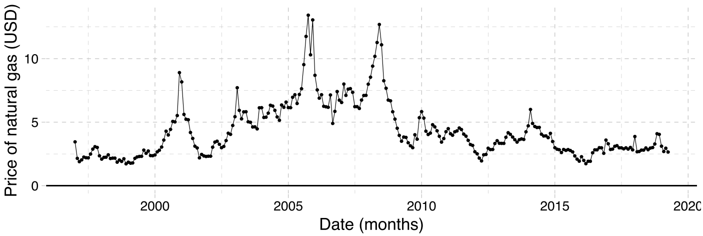
1k. Plot the prices against time. Does it look like we should be concerned about nonstationarity? Explain.

Answer

```r
# Load 'ggplot2' and 'ggthemes' packages
p_load(ggplot2, ggthemes)
# Plot the price of oil over time
ggplot(data = price_df, aes(x = month_year, y = price_oil)) +
geom_line(size = 0.2) +
geom_point(size = 0.5) +
geom_hline(yintercept = 0) +
xlab("Date (months)") +
ylab("Price of oil (USD)") +
theme_pander()
```




```r
# Plot the price of natural gas over time
ggplot(data = price_df, aes(x = month_year, y = price_gas)) +
geom_line(size = 0.2) +
geom_point(size = 0.5) +
geom_hline(yintercept = 0) +
xlab("Date (months)") +
ylab("Price of natural gas (USD)") +
theme_pander()
```



It looks like the price of oil may be nonstationary, as its mean, $E[P_t^{\text{Oil}}]$ tends to increase with time.

There may also be a violation of variance stationarity. The variance appears to increase in the middle period of both time series.

1l. 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.

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.

ANSWER For models (1a) and (1d), autocorrelated disturbances will cause OLS to (1) be inefficient and (2) have 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.

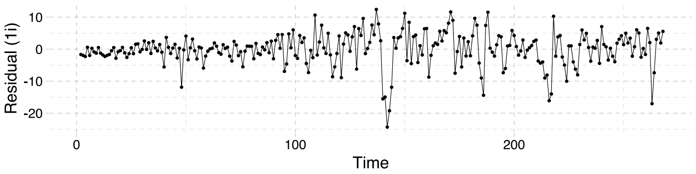
Answer Done in hint.

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?

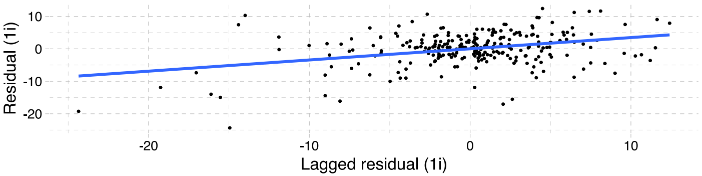
I strongly encourage you to use `ggplot2` for these graphs.

Answer

```
# Plot 1: Residuals over time
ggplot(data = price_df, aes(x = t_month, y = e_1i)) +
geom_path(size = 0.2) +
geom_point(size = 0.5) +
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(size = 0.5) +
geom_smooth(method = lm, se = F, alpha = 0.5) +
xlab("Lagged residual (1i)") + ylab("Residual (1i)") +
theme_pander()
```



This figures might suggest autocorrelation.

In the first figure, we're looking for residuals to closely follow the residual that preceded them—for example, larger residuals followed by other large residuals. This seems to be the case—especially in the middle and end of the the time series.

In the second figure, autocorrelation would look show up with residuals forming some sort of line. There *may* be a positive relationship between the current residual and its lag (the **blue line**).

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).

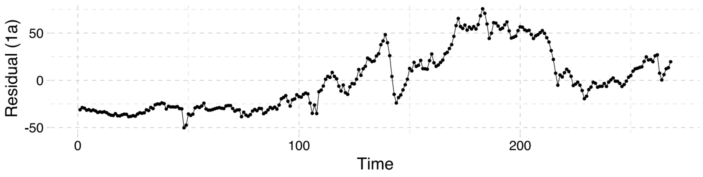
```
# Residuals from the model in 1a
price_df$e_1a ← residuals(ols_1a)
# Residuals from the model in 1d
price_df$e_1d ← c(NA, NA, residuals(ols_1d))
```

Answer Done in hint.

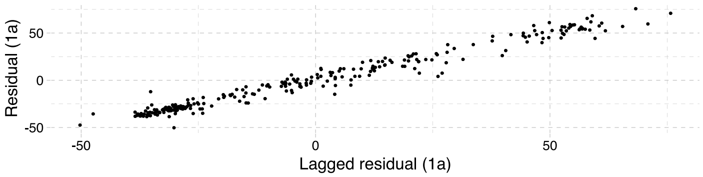
2e. Repeat the plots from above—**1** plot the residuals against the time variable (`t_month`) and **2** plot the residuals against their lag—for both sets of residuals, *i.e.*, for the residuals from (1a) and for the residuals from (1d). You should end up with four graphs for this part.

Answer

```
# Plot 1a 1: Residuals over time
ggplot(data = price_df, aes(x = t_month, y = e_1a)) +
geom_path(size = 0.2) +
geom_point(size = 0.5) +
xlab("Time") + ylab("Residual (1a)") +
theme_pander()
```

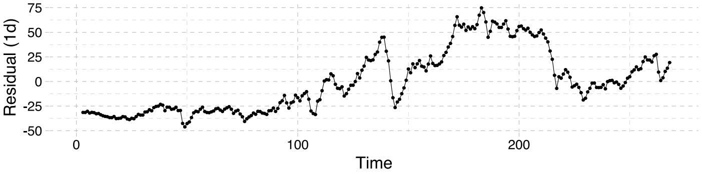



```
# Plot 1a 2: Residuals against their lags
ggplot(data = price_df, aes(x = lag(e_1a), y = e_1a)) +
geom_point(size = 0.5) +
xlab("Lagged residual (1a)") + ylab("Residual (1a)") +
theme_pander()
```

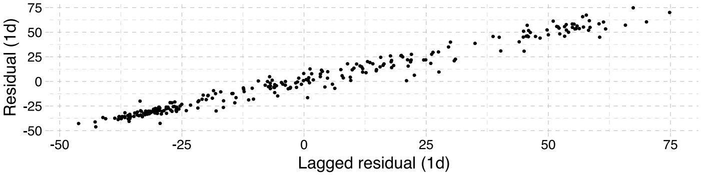


Answer, continued

```r
# Plot 1d 1: Residuals over time
ggplot(data = price_df, aes(x = t_month, y = e_1d)) +
geom_path(size = 0.2) +
geom_point(size = 0.5) +
xlab("Time") + ylab("Residual (1d)") +
theme_pander()
```




```r
# Plot 1d 2: Residuals against their lags
ggplot(data = price_df, aes(x = lag(e_1d), y = e_1d)) +
geom_point(size = 0.5) +
xlab("Lagged residual (1d)") + ylab("Residual (1d)") +
theme_pander()
```



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 oil depends strongly on the previous period's price of oil, then if we fail to control/include the previous period's price of oil (as we do in (1a) and (1d)), then the previous period's price of oil 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_gas + lag(price_gas, 1) + lag(price_oil, 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_gas + lag(price_gas, 1) + lag(price_oil, 1) + lag(e_1i,
##     1) + lag(e_1i, 2)
## Model 2: e_1i ~ price_gas + lag(price_gas, 1) + lag(price_oil, 1)
##   Res.Df Df      F    Pr(>F)
## 1    259
## 2    261 -2 18.038 4.631e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The **fourth step** invovles an F test for the two lags. The F test above has a p-value of approximately 0.0000000463, which means we reject H_0 at the 5-percent level.

In the **fifth step**, we make our conclusion. Here, H_0 is "no autocorrelation". Thus, we reject "no autocorrelation"—meaning we find statistically significant evidence of autocorrelation for model (1i) at the 5-percent level.

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 (and 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. Plus we've found evidence of autocorrelation and we have a lagged dependent variable, so the estimate is potentially biased/inconsistent. (We also probably want to think about nonstationarity.)

Description of variables and names

Variable	Description
`month_year`	The observation's month and year (`character`)
`price_gas`	The month (`numeric`)
`price_oil`	The year (`numeric`)
`month`	The average (Henry Hub) price of natural gas, $ per 1MM BTU (`numeric`)
`year`	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`)