Problem Set 4 Solutions

Autocorrelation, Nonstationarity, and Causality

EC 421: Introduction to Econometrics

Due *before* midnight (11:59pm) on Sunday, 08 March 2020

DUE Your solutions to this problem set are due *before* midnight on Sunday, 08 March 2020. Your files must be uploaded to Canvas.

IMPORTANT You must submit **two files**:

1. your typed responses/answers to the question (in a Word file or something similar)
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: Concepts

1a. Define the term **standard error**.

Answer The standard error is the standard deviation of an estimator's distribution.

1b. If the *p*-value of a hypothesis test is 0.14, what do we conclude about the null hypothesis?

Answer With a *p*-value of 0.14, we would **fail to reject** the null hypothesis—concluding there is not sufficient statistical evidence to reject the null hypothesis being true.

1c. If the *p*-value of a hypothesis test is 0.01, what do we conclude about the null hypothesis?

Answer With a *p*-value of 0.01, we **reject** the null hypothesis—concluding there is statistically significant evidence it is not true.

1d. Define **mean stationarity** in your own words.

Answer Mean stationarity means that a random variable's mean is independent of time.

1e. If the variance of a random variable increases over time, is that variable variance stationary? Explain.

Answer Variance stationarity requires that a variable's variance is constant through time. Thus, a variable whose variance increases through time would violate variance stationary (it is not variance stationary).

1f. Define a **random walk** and show why/how it violates **variance stationarity**.

Answer A random walk is defined as

$$u_t = u_{t-1} + \varepsilon_t$$

and, as defined in class and the notes, its variance can be written as $\text{Var}(u_t) = t\sigma_\varepsilon^2$. Thus, its variance depends upon time, which directly violated variance stationarity.

1f. How does the task of *understanding causality* differ from the task of *prediction*?

Answer Causality is concerned with understanding how one variable (or a set of variables) influences our outcome—we care about the βs, whereas prediction is focused on predicting (estimating) the outcomes (the ys).

1g. Define the fundamental problem of causality in your own words.

Answer The fundamental problem of causality is that we cannot observe the same individual both with and without treatment—if we observe a person with treatment, then we cannot observe her without treament.

1h. Define **covariance stationarity** in your own words.

Answer Covariance stationarity means that a variable's covariance between two time periods (*e.g.*, k periods apart) does not change over time.

Problem 2: Empirics

Load the energy-price dataset from the previous problem set (`004-data.csv`). The variables are explained on the last page.

```r
# Setup
library(pacman)
p_load(tidyverse, broom, magrittr, here)
# Load data
energy_df = here("004-data.csv") %>% read_csv()
```

```
## Parsed with column specification:
## cols(
##   t = col_double(),
##   date = col_date(format = ""),
##   year = col_double(),
##   month = col_double(),
##   price_electricity = col_double(),
##   price_coal = col_double(),
##   price_gas = col_double()
## )
```

2a. Write out a regression model that has

- outcome variable: the price of electricity
- explanatory variables: the price of natural gas and its first **two** lags (plus an intercept)

Answer $\left(\text{Price of electricity}\right)_t = \beta_0 + \beta_1 \left(\text{Price of nat. gas}\right)_t + \beta_2 \left(\text{Price of nat. gas}\right)_{t-1} + u_t$

2b. Estimate the model in **2a** and report your results. Interpret the coefficients and comment on their significance.

Answer We now regress the price of electricity on the price of natural gas and its first lag.

```
# Estimate the regression
est_2b = lm(price_electricity ~ price_gas + lag(price_gas) + lag(price_gas, 2), data = energy_
# Tidy results
est_2b %>% tidy()
```

```
## # A tibble: 4 x 5
##   term              estimate std.error statistic   p.value
##   <chr>                <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)        13.9        0.211    65.9    7.67e-103
## 2 price_gas          -0.466      0.133    -3.51   6.16e-  4
## 3 lag(price_gas)      0.208      0.178     1.17   2.46e-  1
## 4 lag(price_gas, 2)  -0.0709     0.114    -0.622  5.35e-  1
```

Our coefficient on the current price of natural gas is statistically significant and suggests that an additional dollar increase in the price of natural gas reduces the price of electricity by approximately 0.47 dollars (holding everything else constant).

Neither of the lags are statistically significant. The first lag tells us how the price of natural gas *last month* affects electricity prices *this month* (holding everything else constant). The second lag tells us how the price of natural gas *two months ago* affects the price of electricity *this month* (holding everything else constant).

2c. If the disturbance in the model in **2a** is autocorrelated, will OLS be unbiased and/or consistent for the coefficients? Briefly explain your answer.

Answer In the presence of autocorrelation, ordinary least squares (OLS) is unbiased and consistent when estimating the coefficients in static models and in dynamic models with only lagged explanatory variables. Our model in **2a** only has lagged explanatory variables, so OLS will is unbiased and consistent in estimating the coefficients.

2d. What issues does OLS have if the disturbance in the model in **2a** is autocorrelated?

Answer In the presence of autocorrelation, our OLS estimator of the standard error is biased, which will "mess up" our inference—causing our standard errors, confidence intervals, and hypothesis tests to be incorrect.

2e. Following the methods described in class and in the notes, use your residuals (from **2b**) to test the disturbance for first-order autocorrelation.

Answer We want to regress the residuals from **2b.** on their first lag.

```r
# Add residuals to the dataset
energy_df %<>% mutate(resid_2b = c(NA, NA, residuals(est_2b)))
# Estimate the regression
est_2e = lm(resid_2b ~ -1 + lag(resid_2b), data = energy_df)
# Tidy results
est_2e %>% tidy()
```

```
## # A tibble: 1 x 5
##   term          estimate std.error statistic  p.value
##   <chr>            <dbl>     <dbl>     <dbl>    <dbl>
## 1 lag(resid_2b)    0.824    0.0464      17.8 4.92e-37
```

Our null hypothesis is H$_0$: $\rho = 0$, which means *no autocorrelation*.

Our alternative hypothesis is H$_A$: $\rho \neq 0$, which means we have first-order autocorrelation.

The coefficient on the first lag of the residuals is statistically significant at the five-percent level (the *p*-value is less than 0.001), which means we reject the null hypothesis of "no relationship between the residuals and their lags". In other words: We have found statistically significant evidence of first-order autocorrelation in our disturbance.

2f. Now add two lags of the price of electricity to your model in **2a**—so your explanatory variables should be:

- the price of natural gas
- the first two lags of the price of natural gas
- the first two lags of the price of electricity
- an intercept

Write out this model.

Answer

$$\begin{aligned}(\text{Price of electricity})_t = \beta_0 + \beta_1(\text{Price of nat. gas})_t + \beta_2(\text{Price of nat. gas})_{t-1} \\ + \beta_3(\text{Price of electricity})_{t-1} + \beta_4(\text{Price of electricity})_{t-2} + u_t\end{aligned}$$

2g. Is OLS unbiased and/or consistent for the model you wrote out in **2f**? Briefly explain.

Answer This model has a lagged dependent variable, so OLS will not be unbiased (it is biased) for the coefficients in **2f**—it mechanically violates exogeneity.

If u_t **is not autocorrelated**, then OLS will still be consitent for the coefficients in **2f**, as it still satisfies contemporaneous exogeneity. However, if u_t **is autocorrelated**, then OLS will be inconsistent for the coefficients, as it violates contemporaneous exogeneity.

2h. Now estimate the model that you wrote out in **2f**. Interpret the coefficient on the lags of the price of electricity and comment on all of the variables' significance—how have they changed from the first model?

Answer

```
# Estimate the regression
est_2h = lm(
  price_electricity ~
  price_gas + lag(price_gas) + lag(price_gas, 2) +
  lag(price_electricity) + lag(price_electricity, 2),
  data = energy_df
)
# Tidy results
est_2h %>% tidy()
```

```
## # A tibble: 6 x 5
##   term                      estimate std.error statistic  p.value
##   <chr>                        <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)                 2.09      0.536      3.90  1.53e- 4
## 2 price_gas                  -0.0700    0.0593    -1.18  2.40e- 1
## 3 lag(price_gas)              0.0519    0.0766     0.678 4.99e- 1
## 4 lag(price_gas, 2)          -0.0371    0.0489    -0.759 4.49e- 1
## 5 lag(price_electricity)      1.17      0.0824    14.2   2.73e-28
## 6 lag(price_electricity, 2)  -0.320     0.0824    -3.88  1.63e- 4
```

Both of the lags on the price of electricity are statistically significant. The first lag tells us how the price of electricity *last month* affects electricity prices *this month*—suggesting a one-dollar increase in last month's electricity price will, on average, increase the price of electricity this month by 1.17 dollars (holding everything else constant). The second lag tells us how the price of electricity *two months ago* affects the price of electricity *this month*—suggesting a one-dollar increase two months ago will lower electricity prices by 0.32 dollars this month (holding all else constant).

The price of natural gas is no longer significant, and its coefficients are quite small.

2i. Using the residuals from **2h**, test the disturbance for second-order autocorrelation. What do you conclude? What do your conclusions imply for our coefficient estimates?

Hint: Make sure you include the **full** set of explanatory variables in the test (the notes were unclear in class but have since been updated).

Answer We now want to regress the residuals from **2h** on their two lags and eacho of the explanatory variables in the original regression.

```
# Add residuals to dataset
energy_df %<>% mutate(resid_2h = c(NA, NA, residuals(est_2h)))
# Estimate the regression
est_2i = lm(
  resid_2h ~
  lag(resid_2h) + lag(resid_2h, 2) +
  price_gas + lag(price_gas) + lag(price_gas, 2) +
  lag(price_electricity) + lag(price_electricity, 2),
  data = energy_df)
# Tidy results
est_2i %>% tidy()
```

```
## # A tibble: 8 x 5
##   term                      estimate std.error statistic  p.value
##   <chr>                        <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)                -5.29      1.75      -3.02  0.00304
## 2 lag(resid_2h)              -2.33      0.671     -3.48  0.000700
## 3 lag(resid_2h, 2)           -0.724     0.250     -2.89  0.00447
## 4 price_gas                  -0.0173    0.0602    -0.288 0.774
## 5 lag(price_gas)              0.139     0.0967     1.44  0.154
## 6 lag(price_gas, 2)           0.0169    0.0545     0.311 0.756
## 7 lag(price_electricity)      2.22      0.650      3.41  0.000886
## 8 lag(price_electricity, 2)  -1.84      0.537     -3.43  0.000803
```

The coefficients on the lagged residuals are individually statistically significant, but we want to test them jointly.

(Answer continues on the next page.)

Answer, continued We now jointly test the two lagged residuals

```
# Wald test
p_load(lmtest)
waldtest(est_2i, c("lag(resid_2h)", "lag(resid_2h, 2)"))
```

```
## Wald test
##
## Model 1: resid_2h ~ lag(resid_2h) + lag(resid_2h, 2) + price_gas + lag(price_gas) +
##     lag(price_gas, 2) + lag(price_electricity) + lag(price_electricity,
##     2)
## Model 2: resid_2h ~ price_gas + lag(price_gas) + lag(price_gas, 2) + lag(price_electricity) +
##     lag(price_electricity, 2)
##   Res.Df Df      F   Pr(>F)
## 1    126
## 2    128 -2 6.3418 0.002376 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This *p*-value suggests that the two lagged residuals are jointly statistically significant.

We reject the null hypothesis (no autocorrelation) and conclude that there is statistically significant evidence of second-order autocorrelation (or autocorrelation consistent with first- or second-order autocorrelation).

Problem 3: Causality

Following the Rubin causal model, imagine that we observe the following dataset.

Table: Imaginary dataset

i	Trt.	y_1	y_0
1	0	34	22
2	0	19	13
3	1	13	1
4	1	13	10

3a. Which numbers would be impossible to observe in real life?

Answer We cannot observe the treated outcome (y_1) for control individuals (for $i = 1$ and 2). Similarly, we cannot observe the untreated outcome (y_0) for treated individuals (for $i = 3$ and 4).

3b. Define the term **average treatment effect**.

Answer The averate treatment effect is literally the average (mean) effect of treatment across the individuals in the population.

3c. Based upon this dataset, calculate the **true average treatment effect**.

Answer The individual treatment effects are: 12, 6, 12, 3.

Thus, the average treatment effect is 8.25.

3d. Based upon this dataset, **estimate** the average treatment effect by taking the difference between the mean of the treatment group and the mean of the control group.

Answer The "estimated" average treatment effect is (13 + 13)/2 - (22 + 13)/2 = -4.5.

3e. Under what conditions would the estimate in **3d** be unbiased? Explain.

Answer Our estimate in **3d** will be unbiased when there is no selection bias—for example, when we have randomized a treatment.

3f. Is the treatment effect heterogeneous? Explain your answer.

Answer Yes! The treatment effect is heterogeneous because the individual treatment effects are not constant (*i.e.*, they differ across some individuals).

Description of variables and names

Variable	**Description**
`t`	Time, measured by months in the dataset (`numeric`)
`date`	The observation's month and year (`character`)
`year`	The year (`numeric`)
`month`	The month (`numeric`)
`price_electricity`	The average residential electricity price in USD (`numeric`)
`price_coal`	The average price of coal, $ per short ton (`numeric`)
`price_gas`	The average price of natural gas, $ per cubic ft (`numeric`)