Pol Sci 630: Problem Set 13 Solutions: Autocorrelation, Clustered SE

Prepared by: Anh Le (anh.le@duke.edu)

December 2, 2016

1 Diagnosing autocorrelation

1.1 Generating autocorrelated data

Generate data (i.e. e, X, Y) that follow an AR(2) process, described below. It's an AutoRegressive(2) process because the error term is correlated with itself up to two periods, i.e. $e(t) = a_1 e(t-1) + a_2 e(t-2) + v(t)$. (This is also described in slide 4 of your lecture note).

$$v(t) \sim N(0,1) \tag{1}$$

$$e(t) = a_1 e(t-1) + a_2 e(t-2) + v(t) \qquad \text{Important: } a_1 + a_2 < 1 \tag{2}$$

$$Y(t) = X(t) + e(t) \tag{3}$$

Let arbitrarily pick $a_1 = 0.4$, $a_2 = 0.2$, and T, the number of time periods, $= 100$

Hint: How do we simulate the vector $e(t)$? Start with $e(1) = rnorm(1), e(2) = a_1 * e(1) + rnorm(1)$, then calculate $e(3), e(4), \ldots, e(100)$ according to the DGP above. You may want to use a loop for this.

Solution

```
T <- 100 # Num of time periods

# Generate autocorrelated e
e <- vector(mode = 'numeric', length = T)
e[1] <- rnorm(1)
e[2] <- 0.4 * e[1] + rnorm(1)
for (t in 3:T) {
  e[t] <- 0.4 * e[t - 1] + 0.2 * e[t - 2] + rnorm(1)
}

X <- rnorm(T)
Y <- X + e
```

1.2 Diagnostics: Visual

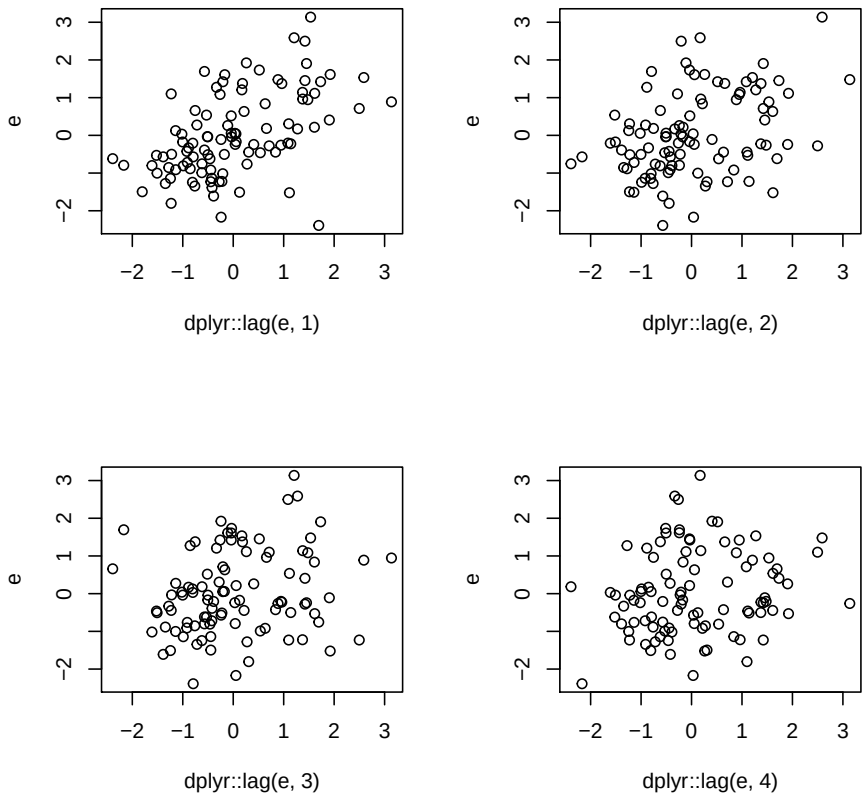
Plot residual against time and against lagged , up to 4 lags (e.g. residual lag-1 residual, residual lag-2 residual, etc. up to 4 plots) How does the correlation look across the four plots?

Hint: to create a lagged vector, you can use `lag` in package `dplyr`. In other cases, i.e. generating lags in a data frame / panel, probably best to use `DataCombine` like in the lab tutorial.

Solution

```
m_auto <- lm(Y ~ X)
e <- resid(m_auto)

par(mfrow = c(2, 2))
plot(dplyr::lag(e, 1), e)
plot(dplyr::lag(e, 2), e)
plot(dplyr::lag(e, 3), e)
plot(dplyr::lag(e, 4), e)
```



We see that he error autocorrelation diminishes the further the lag is (See how the relationship is very strong in the first plot, but not so much in the fourth?).

1.3 Diagnostics: Hypothesis testing

Regress residuals against X and lag1 and lag2 residuals, and then doing an F test for joint significance in the lagged residuals. (This is described in slide 4 of your lecture note).

You can use `linearHypothesis` function in package `car` to conduct a joint F test.

Solution

```
lag1_e <- dplyr::lag(e, 1)
lag2_e <- dplyr::lag(e, 2)

# Reg residual against X and lagged residuals
```

```r
m_autotest <- lm(e ~ X + lag1_e + lag2_e)

# Doing an F test
library(car) # to run F-test
linearHypothesis(m_autotest, c("lag1_e", "lag2_e"))

## Linear hypothesis test
##
## Hypothesis:
## lag1_e = 0
## lag2_e = 0
##
## Model 1: restricted model
## Model 2: e ~ X + lag1_e + lag2_e
##
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1     96 118.292
## 2     94  85.157  2    33.135 18.288 1.957e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We reject the null of no autocorrelation. Notice how we can do the F-test on more than just one lag to detect autocorrelation. In real research, you would use your judgement to guess the autocorrelation structure (i.e. how far back does the autocorrelation go?) and test it.

2 Clustered errors

2.1 Conceptual: Why is it a problem?

Note: Just me explaining here, nothing for you to do.

Among the 5 assumptions of classical linear regression, which one was violated by clustered standard errors?

Clustered errors violate the Spherical Error assumption, specifically the No Autocorrelation part, i.e. $E[\epsilon_i \epsilon_j | X] = 0 for i \neq j$. Indeed, clustered errors mean that two observations within the same clusters have correlated error. In other words, $E[\epsilon_i \epsilon_j | X] \neq 0$ if i, j are in the same cluster.

This violation leads to higher standard errors, i.e. higher uncertainty in our model estimates. Intuitively, if clustered errors are present, units within a cluster are correlated. Thus, each unit does not present a 100% new piece of information. In this case, we have less information to estimate our model, and we are less certain about our estimate.

For example, our cluster is a classroom, and our unit is a student. Imagine the extreme case when every student is a clone of one person. The correlation

between units within a classroom is 1 (its maximum). Even if we may have 100-student classroom, there is in fact only one person, one piece of information.

2.2 Clustered errors and diff-in-diff design

Describe a diff-in-diff design (briefly, an equation + a few sentences should be enough). Explain why clustered errors is a problem for diff-in-diff.

Solution

A diff-in-diff design:

$$y_{i,t} = \beta_0 + \beta_1 time_t + \beta_2 treat_i + \beta_3 time_t \times treat_i + \epsilon_{i,t}$$

where i denotes the unit, and t denotes the time. We have clustered errors here because $\epsilon_{i,t}$ are very likely to be correlated within the same unit i. This can happen if I fail to control for an important covariate, thus systematically under- or over-estimate the outcome in a unit i.

2.3 Dealing with clustered errors

Given the discussion above, it's now common practice to use clustered standard errors with a diff-in-diff design. Indeed, Eddy does in his own work, "The Impact of Recentralization on Public Services: A Difference-in-Differences Analysis of the Abolition of Elected Councils in Vietnam." We'll replicate the first 2 columns in table 2, his main findings.

Put results in a stargazer table, including BOTH non-clustered and clustered SE. The final table should 4 columns. How do clustered errors change the size of the coefficient and the SEs?

How to do it? You have two choices:

1. "We do these things not because they are easy but because they are hard." Here is the link to Eddy's replication data and code. Figure out yourself how to replicate.

Easier choice after the break. Resist the easier choice.

2. You can follow these steps. The main findings code is in `APSR_30indicators_0810.do`

- Load data `panel_commune_2008_2010.dta`
- Drop Vietnam's Central Highland (`reg8 == 6`) from dataset
- Create a city dummy, which equals 1 if `tinh==1 | tinh==31 | tinh==48 | tinh==92 | tinh==79`
- Run regression. Dependent vars: goodroadv, transport. Ind vars: time + treatment + time:treatment + lnarea + lnpopden + city + regional fixed effects. (reg8 is the region ID).
- Calculate clustered standard errors with multivcov and coeftest. Note that Eddy clustered both by tinh and huyen at the same time. (Tinh and Huyen are province and district in VNese).

Solution

```
library(haven) ; library(dplyr) ; library(multiwayvcov) ; library(lmtest)
d <- read_dta("panel_commune_2008_2010.dta")

*** Drop Central Highland drop if reg8==6
d <- d %>% filter(reg8 != 6)

# gen city=(tinh==1 | tinh==31 | tinh==48 | tinh==92 | tinh==79)
d$city <- d$tinh == 1 | d$tinh == 31 | d$tinh==48 | d$tinh==92 | d$tinh==79

# global inde time treatment time_treatment lnarea lnpopden city i.reg8
# xi: ivreg2 goodroadv $inde, cluster(tinh huyen)
m_road <- lm(goodroadv ~ time + treatment + time:treatment + lnarea +
               lnpopden + city + factor(reg8), data = d)
vcov_road <- cluster.vcov(m_road, cluster = d[, c("tinh", "huyen")])

m_transport <- lm(transport ~ time + treatment + time:treatment + lnarea +
               lnpopden + city + factor(reg8), data = d)
vcov_transport <- cluster.vcov(m_transport, cluster = d[, c("tinh", "huyen")])
```

```
library(stargazer)

##
## Please cite as:
##  Hlavac, Marek (2015).  stargazer:  Well-Formatted Regression and
Summary Statistics Tables.
##  R package version 5.2.  http://CRAN.R-project.org/package=stargazer

stargazer(m_road, coeftest(m_road, vcov_road),
          m_transport, coeftest(m_transport, vcov_transport),
          omit = "factor",
          order = c("time$", "treatment$", "time:treatment"))
```

Table 1:

	Dependent variable:			
	goodroadv		transport	
	OLS	*coefficient test*	*OLS*	*coefficient test*
	(1)	(2)	(3)	(4)
time	−0.007	−0.007	−0.065***	−0.065***
	(0.012)	(0.013)	(0.014)	(0.018)
treatment	−0.104***	−0.104**	−0.089***	−0.089*
	(0.025)	(0.051)	(0.028)	(0.052)
time:treatment	0.086**	0.086**	0.102***	0.102*
	(0.035)	(0.037)	(0.038)	(0.053)
lnarea	0.015	0.015	0.102***	0.102***
	(0.014)	(0.023)	(0.016)	(0.018)
lnpopden	0.098***	0.098***	0.116***	0.116***
	(0.012)	(0.016)	(0.013)	(0.022)
city	−0.031	−0.031	0.020	0.020
	(0.024)	(0.032)	(0.027)	(0.040)
Constant	0.352***	0.352***	−0.068	−0.068
	(0.094)	(0.126)	(0.104)	(0.156)
Observations	4,126		4,126	
R^2	0.140		0.146	
Adjusted R^2	0.138		0.144	
Residual Std. Error (df = 4113)	0.370		0.410	
F Statistic (df = 12; 4113)	56.022***		58.819***	
Note:				*p<0.1; **p<0.05; ***p<0.01

Clustered SEs do not change coefficient size, but inflate the SE. Also, Eddy is an upstanding scholar whose results replicate.