# Pol Sci 630: Problem Set 13: 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.

### 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.

### 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).

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

# 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.

(Side note: be precise in using the word error and standard errors. Error is the unobserved part in our DGP. Standard error is the standard deviation of our coefficient estimates. Clustered error is the fact that units in the same cluster have correlated error. Clustered standard error is a techique to adjust our standard error, reflecting our decreased certainty in our estimate in the face of clustered error.)

## 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.

## 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 have 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).