MIT 14.32/14.320 **Spring 2023**
J. Angrist

Lecture Note 10

Standard Standard Error Issues

LN7 outlines the tools of regression inference for the carefree world of random samples. We elaborate here on two complications, both arising in samples in which observations are dependent. The first complication is *serial correlation;* the second is *clustering.* Serial correlation is seen most often in time series data, when one observation is correlated with the next. For example, the unemployment rate this month is likely to be similar to the unemployment rate last month. Clustering arises when when observations within groups are similar, but observations are independent across groups. In a randomized trial involving school-age children, for example, those in the same school or classroom are likely to have similar outcomes.

Big picture:

- Dependent observations carry less information than independent observations. Standard errors should reflect this information downgrade
- Practical fix-ups for serial correlation and clustering are straightforward to implement, but can be extraordinarily consequential
- Our SE fix-ups work only in large samples and can be misleading otherwise

1 Serial Correlation in Time Series

1.1 Defining Serial Correlation

- Consider a model relating US quarterly GDP growth (Y_t) to the federal funds rate (X_t), an important component of the Federal Reserve's monetary policy toolkit. For this time series regression, we write:

$$Y_t = \alpha + \beta X_t + \varepsilon_t\,;\; t = 1, \ldots, T \tag{1}$$

- In this case, it's likely that

$$C(\varepsilon_t, \varepsilon_s) = E[\varepsilon_t \varepsilon_s] \neq 0 \text{ for } s \neq t.$$

  In particular, we expect $E[\varepsilon_t \varepsilon_s] > 0$, that is, positive *serial correlation*
- Many economic time series are positively serially correlated. This reflects the fact that macro variables like unemployment rates, GDP levels and growth, financial variables, aggregate consumption, interest rates, and public policy variables like government spending and some interest rates are highly *persistent.*

Consequences of serial correlation

- When residuals are positively serially correlated, conventional regression SEs are usually too small. Robust SEs do not fix this.
- Even if classical assumptions (like homoskedasticity) are otherwise satisfied, OLS isn't BLUE
- OLS estimates are consistent and may still be unbiased (the latter holds assuming regressors are fixed or the CEF is linear)

1.2 Serial Correlation Fix-ups

Modeling serial correlation

- Serial correlation is often described using *autoregressive models*. The simplest is a first-order autoregression or $AR(1)$:

  $$\varepsilon_t = \rho\varepsilon_{t-1} + \nu_t;\ -1 < \rho < 1, \tag{2}$$

  where the error term in this equation is assumed to satisfy:

  - $E[\nu_t] = 0$ for all t (not a restriction since $E[\varepsilon_t] = 0$)
  - $E[\nu_t\nu_s] = 0$ for any $s \neq t$ (serially uncorrelated leftovers)

- We require $|\rho| < 1$ so that the time series error process is *stationary* (the variance of a non-stationary time series process is infinite)

  - Because economic data are persistent, we expect $\rho \geq 0$

- An $AR(2)$ puts two lags on the RHS of the model for serial correlation:

  $$\varepsilon_t = \rho_1\varepsilon_{t-1} + \rho_2\varepsilon_{t-2} + \nu_t;\ -1 < \rho < 1 \tag{3}$$

  This allows residuals to cycle systematically up and down

Modern Times: Newey-West Standard Errors

- A simple fix for serially correlated data uses Newey-West standard errors, named after our colleague, Whitney Newey, who invented these with Ken West in his youth
- Newey-West SEs are also called HAC standard errors (*Heteroskedasticity and Autocorrelation Consistent*)
- HAC SEs allow for unrestricted serial correlation and heteroskedasticity, with the former not limited to $AR(1)$
  - HAC generalizes robust SEs
  - Stata can HAC it (but you must pick the lag length)

Old School: Generalized Least Squares

- Old-timers transform a model with $AR(1)$ residuals into something with serially uncorrelated residuals
  - Write the model of interest and ρ-times-the-lagged-model:

  $$Y_t = \alpha + \beta X_t + \varepsilon_t \tag{4}$$

  $$\rho Y_{t-1} = \rho\alpha + \rho\beta X_{t-1} + \rho\varepsilon_{t-1} \tag{5}$$

  - Subtract (5) from (4):

  $$\begin{aligned}(Y_t - \rho Y_{t-1}) &= (1-\rho)\alpha + \beta(X_t - \rho X_{t-1}) + \varepsilon_t - \rho\varepsilon_{t-1} \\ &= (1-\rho)\alpha + \beta(X_t - \rho X_{t-1}) + \nu_t\end{aligned} \tag{6}$$

- *Quasi-differenced* equation (6) has a serially uncorrelated error term (assumed homoskedastic in the old-school framework)

- OLS on the transformed equation is a version of *generalized least squares* (GLS), a strategy that transforms an original model with problematic residuals into one with residuals that have classical properties (provided υ_t is homoskedastic)
- And, provided υ_t is homoskedastic, conventional standard errors for estimates of equation (6) should not be misleading. We therefore estimate:

$$Y_t^* = \alpha^* + \beta X_t^* + \nu_t \tag{7}$$

where

$$\begin{aligned} Y_t^* &= (Y_t - \hat{\rho} Y_{t-1}) \\ X_t^* &= (X_t - \hat{\rho} X_{t-1}) \\ \alpha^* &= (1 - \hat{\rho})\alpha, \end{aligned}$$

and $\hat{\rho}$ is a consistent estimate of ρ, computed from OLS residuals

  - Does it matter that we quasi-difference using $\hat{\rho}$ rather than ρ? Not in asymptopia!
  - Stata automates quasi-differencing using a command called `Prais`, named for Prais-Winsten (1954), which introduces a version of the procedure. You'll use a variation known as Cochrane-Orcutt (CORC) on Pset 5.

The Durbin-Watson Test

- How to decide whether serial correlation is a problem? Simplest is to look at the regression of $\hat{\epsilon}_t$ on $\hat{\epsilon}_{t-1}$, a direct estimate of ρ
- Among the Lost Tribes of Macroeconomia, isolated as they are from modern applied microeconomics, it's customary to report the Durbin-Watson (DW) statistic:

$$\begin{aligned} DW &= \sum_{t=2}^{T} (\hat{\varepsilon}_t - \hat{\varepsilon}_{t-1})^2 / \sum_{t=1}^{T} \hat{\varepsilon}_t^2 \\ &= \sum_{t=2}^{T} (\hat{\varepsilon}_t^2 - 2\hat{\varepsilon}_t\hat{\varepsilon}_{t-1} + \hat{\varepsilon}_{t-1}^2) / \sum_{t=1}^{T} \hat{\varepsilon}_t^2 \\ &= \left\{ \sum_{t=2}^{T} \hat{\varepsilon}_t^2 / \sum_{t=1}^{T} \hat{\varepsilon}_t^2 \right\} - 2 \left\{ \sum_{t=2}^{T} \hat{\varepsilon}_t\hat{\varepsilon}_{t-1} / \sum_{t=1}^{T} \hat{\varepsilon}_t^2 \right\} + \left\{ \sum_{t=2}^{T} \hat{\varepsilon}_{t-1}^2 / \sum_{t=1}^{T} \hat{\varepsilon}_t^2 \right\} \\ &\approx 1 - 2\hat{\rho} + 1 = 2(1 - \hat{\rho}) \end{aligned}$$

- Moral: Look for DW that is "close to 2" when testing $H_0 : \rho = 0$. How close? Stata computes DW p-values

1.3 Something Fishy in the Data (from Graddy 1995)

The Fulton Fish Market. which moved from lower Manhattan to the South Bronx in 2005, is the second largest wholesale fish market in the world (Tokyo's Tsukiji is the largest)

- Graddy (1995) looks for evidence of non-competitive behavior at Fulton by comparing the prices of fish paid by buyers of different ethnicities (she observes a single white seller)
- Does the law of one price hold for fish - or is there something fishy about Fulton fish prices? An economic rationale for the Asian-white price difference (for the same fish) is *price discrimination*, as a consequence of which more elastic Asian buyers pay less

As luck would have it, the fish of interest in this study is called *whiting*.

```
. /*******************************************************************************
> *Title:         Serial Correlation Fix-ups
> *Author:        JA revised 03-13-23
> *******************************************************************************/
.
. cd "/Users/joshangrist/Documents/teaching/14.32/SP2023/notes/LN10/newfishformat/"
/Users/joshangrist/Documents/teaching/14.32/SP2023/notes/LN10/newfishformat

.
. use fish.dta, clear

. *** data start wide, with asians and whites as columns, rows are days
. *** generate a unique id identifier (i)
. gen t = _n

.
. *** reshape to panel format
.
. reshape long price_ qty_, i(t) j(race) string
(j = a w)

Data                               Wide   ->   Long
-----------------------------------------------------------------------------
Number of observations               97   ->   194
Number of variables                  15   ->   14
j variable (2 values)                     ->   race
xij variables:
                        price_a price_w   ->   price_
                            qty_a qty_w   ->   qty_
-----------------------------------------------------------------------------

.
. list race t price* in 1/10

     +---------------------+
     | race   t     price_ |
     |---------------------|
  1. |    a   1   .6222222 |
  2. |    w   1   .7666667 |
  3. |    a   2   .9722222 |
  4. |    w   2      1.175 |
  5. |    a   3   1.233333 |
     |---------------------|
  6. |    w   3      1.475 |
  7. |    a   4   1.928571 |
  8. |    w   4      1.625 |
  9. |    a   5    .803125 |
 10. |    w   5   .8642857 |
     +---------------------+

.
. /**********                     Regression Analysis (Time Series)         *********************/
.
. gen ln_price = log(price)

. *** create asian dummy in long file format
. gen asian = race == "a"

.
. reg ln_price asian day* wave*

      Source |       SS           df       MS      Number of obs   =       194
-------------+----------------------------------   F(7, 186)       =     12.10
       Model |  10.0637107         7  1.43767296   Prob > F        =    0.0000
    Residual |  22.0908542       186  .118768034   R-squared       =    0.3130
-------------+----------------------------------   Adj R-squared   =    0.2871
       Total |   32.154565       193  .166603964   Root MSE        =    .34463

------------------------------------------------------------------------------
    ln_price | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
       asian |  -.1006769   .0494856    -2.03   0.043    -.1983021   -.0030516
        day1 |  -.0298537   .0794494    -0.38   0.708    -.1865915    .1268841
        day2 |  -.0222711   .0782519    -0.28   0.776    -.1766465    .1321043
        day3 |   .0551473   .0779863     0.71   0.480    -.0987041    .2089987
        day4 |   .1095638   .0774423     1.41   0.159    -.0432144    .2623419
       wave2 |   .0961695   .0148811     6.46   0.000     .0668121    .1255269
       wave3 |   .0506641   .0138567     3.66   0.000     .0233276    .0780005
       _cons |  -.9542108   .1037614    -9.20   0.000    -1.158911   -.7495102
------------------------------------------------------------------------------

.
. *** tell stata the file is a panel to avoid a bad lag over the asian-white seam
. tsset asian t

Panel variable: asian (strongly balanced)
 Time variable: t, 1 to 97
         Delta: 1 unit

.
```

```
. *** newey2 does HAC with panels
. 
. newey2 ln_price asian day* wave2 wave3, lag(1)


Regression with Newey-West standard errors          Number of obs  =       194
maximum lag : 1                                     F(  7,   186)  =     14.44
                                                    Prob > F       =    0.0000

------------------------------------------------------------------------------
             |             Newey-West
    ln_price | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
       asian |  -.1006769    .061994    -1.62   0.106    -.2229786    .0216249
        day1 |  -.0298537   .0703021    -0.42   0.672    -.1685456    .1088382
        day2 |  -.0222711   .0841825    -0.26   0.792    -.1883463    .1438041
        day3 |   .0551473   .0777887     0.71   0.479    -.0983143     .208609
        day4 |   .1095638   .0596327     1.84   0.068    -.0080796    .2272071
       wave2 |   .0961695    .013967     6.89   0.000     .0686153    .1237237
       wave3 |   .0506641   .0122038     4.15   0.000     .0265884    .0747397
       _cons |  -.9542108   .1071439    -8.91   0.000    -1.165584   -.7428372
------------------------------------------------------------------------------

. newey2 ln_price asian day* wave2 wave3, lag(2)


Regression with Newey-West standard errors          Number of obs  =       194
maximum lag : 2                                     F(  7,   186)  =     14.64
                                                    Prob > F       =    0.0000

------------------------------------------------------------------------------
             |             Newey-West
    ln_price | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
       asian |  -.1006769   .0686142    -1.47   0.144     -.236039    .0346853
        day1 |  -.0298537   .0668651    -0.45   0.656    -.1617651    .1020577
        day2 |  -.0222711   .0784429    -0.28   0.777    -.1770232     .132481
        day3 |   .0551473   .0746567     0.74   0.461    -.0921354    .2024301
        day4 |   .1095638   .0550637     1.99   0.048     .0009341    .2181935
       wave2 |   .0961695   .0145235     6.62   0.000     .0675176    .1248214
       wave3 |   .0506641   .0121467     4.17   0.000      .026701    .0746271
       _cons |  -.9542108   .1124171    -8.49   0.000    -1.175987   -.7324342
------------------------------------------------------------------------------


. 
. *** oldschool: quasi-diff
. *** stata prais command does corchrane-orcutt quasi-differencing, use option "corc twostep" to prevent iteration
. *** (corc will drop the first obs in each time series sequence)
. 
. prais ln_price asian day* wave2 wave3, corc twostep

Number of gaps in sample = 1   (gap count includes panel changes)
note: computations for rho restarted at each gap.

Iteration 0:   rho = 0.0000
Iteration 1:   rho = 0.5792

Cochrane-Orcutt AR(1) regression with twostep estimates

      Source |       SS           df       MS      Number of obs   =       192
-------------+----------------------------------   F(7, 184)       =      5.09
       Model |  2.64091725         7  .377273893   Prob > F        =    0.0000
    Residual |  13.6427548       184  .074145407   R-squared       =    0.1622
-------------+----------------------------------   Adj R-squared   =    0.1303
       Total |   16.283672       191  .085254827   Root MSE        =     .2723

------------------------------------------------------------------------------
    ln_price | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
       asian |  -.0963775   .0933951    -1.03   0.303    -.2806405    .0878854
        day1 |   .0051914   .0499291     0.10   0.917    -.0933156    .1036985
        day2 |   -.014794   .0560661    -0.26   0.792    -.1254091    .0958211
        day3 |   .0604477   .0562757     1.07   0.284    -.0505809    .1714764
        day4 |   .1013305   .0473123     2.14   0.034     .0079862    .1946749
       wave2 |   .0605689   .0126723     4.78   0.000     .0355673    .0855705
       wave3 |   .0435863   .0128231     3.40   0.001     .0182871    .0688855
       _cons |  -.7280926   .1216098    -5.99   0.000    -.9680215   -.4881636
-------------+----------------------------------------------------------------
         rho |   .5791789
------------------------------------------------------------------------------
Durbin-Watson statistic (original)    = 0.820225
Durbin-Watson statistic (transformed) = 1.639846

. 
. log close
      name:  <unnamed>
       log:  /Users/joshangrist/Documents/teaching/14.32/SP2023/notes/LN10/newfishformat/newfish.smcl
  log type:  smcl
 closed on:  13 Mar 2023, 14:14:02
------------------------------------------------------------------------------------------------------------------------------
```

2 Clustering in Data with a Group Structure

Many econometrically interesting samples have a group structure. For example, data on K-12 test scores come from samples of children grouped into classes and schools. An econometric challenge in such settings arises from the fact that observations within groups are correlated.

- Regressions for data with a group structure can be written like this:

$$\text{Y}_{ig} = \beta_0 + \beta_1 x_g + \varepsilon_{ig} \tag{8}$$

  - Y_{ig} is the dependent variable for individual i in group g
  - Regressor x_g varies only at the group level

- Do small classes enhance student learning? Krueger (1999) and Angrist and Lavy (1999) study the effects of class size on test scores. These studies analyze samples of children grouped into schools and classes.
  - Children in the same classroom have much in common; they have, for example, the same teacher. This makes their test scores dependent or *clustered.*
  - Clustering is often a big deal: clustered standard errors take the shine off many a bright idea!

- Data from the STAR experiment analyzed by Krueger (1999) consist of Y_{ig}, the test score of student i in class g, and class size, x_g. Scores of students in the same class are probably correlated.
  - This intra-class correlation means that observations on children in the same classroom are less informative about class size effects than are data drawn from other classrooms. To see why, imagine that children in the same class are so similar that they have the same values of Y_{ig}. We then learn about the entire class by observing a single student, while information on other students in the same class is worthless.

- In econometric discussions of clustering problems, the correlation between residuals in a group is called an *intra-class correlation coefficient*, even when the groups of interest are not actually classrooms

2.1 Random-Effects and the Moulton Factor

- A *random-effects model* postulates:

$$\varepsilon_{ig} = v_g + \eta_{ig} \tag{9}$$

  - v_g is an error component or *random effect* specific to class g, assumed to be uncorrelated across classes:

$$E[v_g v_h] = 0;\ \ g \neq h$$
$$E[v_g^2] = \sigma_\nu^2$$

  - η_{ig} is a student-level error component assumed to uncorrelated within and between groups:

$$E[\eta_{ig}\eta_{jh}] = 0;\ \ i \neq j$$
$$E[\eta_{ig}^2] = \sigma_\eta^2$$

  - These error components are assumed to be homoskedastic (we're focusing on dependence within clusters) and defined so that they're uncorrelated with each other (η_{ig} is the resid from a regression of ε_{ig} on group dummies).

- Random effects residuals are said to be *equicorrelated* because, within groups, any two observations are equally correlated (contrast this with autoregressive time series models, in which correlation diminishes over time)
  - In the random-effects model, the intra-class correlation coefficient in residuals becomes:

$$\rho = \frac{\sigma_v^2}{\sigma_v^2 + \sigma_\eta^2} \tag{10}$$

    (show this)

Clustering is consequential

- Let $V_c(\widehat{\beta}_1)$ be the classical OLS sampling variance and let $V(\widehat{\beta}_1)$ be the variance that accounts for random-effects clustering as modeled above. In a random-effects model with regressors constant within groups and groups of equal size, n, we have:

$$\frac{V(\widehat{\beta}_1)}{V_c(\widehat{\beta}_1)} = 1 + (n-1)\rho \tag{11}$$

- The square root of this is called the *Moulton factor*, after Moulton (1986); a simple cluster fix-up is to multiply conventional SEs by the Moulton factor.
- In Angrist and Lavy (2009), a randomized evaluation of high school achievement awards, 4000 students are grouped in 40 schools, so, on average, $n = 100$. The regressor of interest is a treatment dummy indicating schools that offered large cash awards to students who pass a matriculation exam. The intra-class residual correlation in this study is around 0.10. Applying formula ((11)), the Moulton factor is over 3!
- The general Moulton formula is:

$$\frac{V(\widehat{\beta}_1)}{V_c(\widehat{\beta}_1)} = 1 + \left[\frac{V(n_g)}{\bar{n}} + (\bar{n} - 1)\right]\rho_x\rho, \tag{12}$$

  where n_g is the size of group g, $\overline{n}$ is average group size, and ρ_x is the intra-class correlation of the regressor, x_{ig} (this is less than 1 when regressors vary within groups)
- The Moulton worst-case scenario is when the regressor of interest is fixed within groups ($\rho_x = 1$), as it is for a regressor like class size

Better Head Back to Tennessee, Jed

- In the Krueger (1999) data, a regression of Tennessee STAR Kindergartners' percentile score on class size yields an estimate of -0.62 with a robust (HC_1) standard error of 0.09
- In this case, $\rho_x = 1$ because class size is fixed within classes while $V(n_g)$ is positive because classes vary in size ($V(n_g) = 17.1$)
- The intra-class correlation coefficient for residuals is .31 and the average class size is 19.4
- These numbers give a value of about 7 for $\frac{V(\widehat{\beta}_1)}{V_c(\widehat{\beta}_1)}$, so that conventional standard errors should be multiplied by a factor of $2.65 = \sqrt{7}$ to adjust for clustering

2.2 Other Cluster Fix-Ups

- GLS for random-effects models is a kind of quasi-differencing procedure similar to that used for serial correlation, but this approach to clustering is now rarely seen
- The Stata cluster option generalizes robust SEs to a clustered setting - this has become the default approach in applied microeconometrics
  - Stata cluster works by treating entire clusters as the sampling unit instead of individual data
  - With few clusters, clustered SEs are unreliable (they might, for instance, be smaller rather than larger than unclustered SEs; this is usually a symptom of bias)

MHE Table 8.2.1 compares standard-error fix-ups in Tennessee:

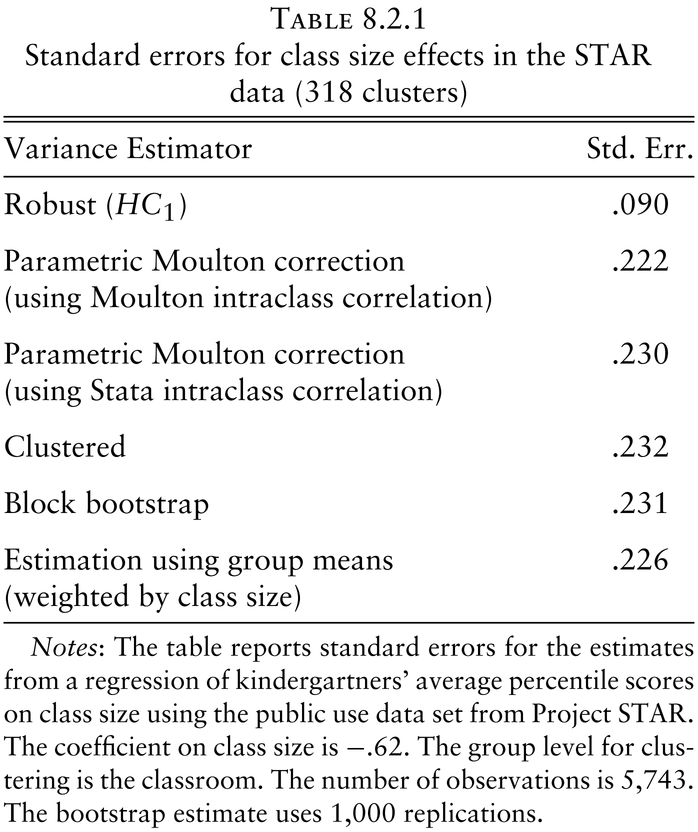

Table 8.2.1
Standard errors for class size effects in the STAR data (318 clusters)

Variance Estimator	Std. Err.
Robust (HC_1)	.090
Parametric Moulton correction (using Moulton intraclass correlation)	.222
Parametric Moulton correction (using Stata intraclass correlation)	.230
Clustered	.232
Block bootstrap	.231
Estimation using group means (weighted by class size)	.226

Notes: The table reports standard errors for the estimates from a regression of kindergartners' average percentile scores on class size using the public use data set from Project STAR. The coefficient on class size is −.62. The group level for clustering is the classroom. The number of observations is 5,743. The bootstrap estimate uses 1,000 replications.

- The SE generated by running regression (8) on 318 group means instead of 5,743 students is close to the clustered standard error
- In clustered samples, the effective sample size is typically closer to the number of clusters than to the number of people

- Note that the group-mean estimates above are "weighted by class size." A weighted least squares regression using J grouped observations, weighted by group size n_j, minimizes a weighted sum of squared errors:

  $$WSSE(a,b) = \sum_{j=1}^{J} n_j(\bar{y}_j - a - bx_j)^2$$

  There are many reasons for weighting. In this context, we might worry that regression models for group averages are especially likely to have heteroskedastic residuals. We can, of course, fix this using robust standard errors (Stata `cluster` also corrects for heteroskedasticity). An alternative (somewhat old-fashioned, but still sensible approach) is to weight the grouped data by group size. In particular, weighting by class size corrects for the fact that class-level means are noisier in smaller classes.

- MHE Section 3.4.1 discusses weighted least squares.