CHAPTER 9: Serial Correlation

MODELLING SERIAL CORRELATION

Serial Correlation: Violation of $Cov(u_t u_{t-s}) = E(u_t u_{t-s}) = 0$ for all $t \neq s$

Often observed in time series data, not in cross-section, but also in panel data.
Hence, it is a rule rather than an exception

Sources: a) Intrinsic serial correlation
b) Model misspecification: Growth in variables (existence of a trend, omitted variables, non-linerarity, measurement errors etc.)

Example on Intrinsic serial correlation: Permanent Income Hypothesis

$Y_t = \beta X_t^* + \varepsilon_t$ where Y_t is consumption and X^*_t is unobserved permanent income. How to estimate X^*_t?

Behavioral Assumption: $X_t^* = X_t + pX_{t-1}^*$ where X_t is current income and p is weight for past unobserved permanent income. Also, note $E(\varepsilon_t \varepsilon_{t-s}) = 0$ and $E(\varepsilon_t^2) = \sigma^2_\varepsilon$

Transformation: lag the model one period: $Y_{t-1} = \beta X_{t-1}^* + \varepsilon_{t-1}$ and multiply by p and subtract from this equation.

One gets: $Y_t - pY_{t-1} = \beta(X_t^* - pX_{t-1}^*) + (\varepsilon_t - p\varepsilon_{t-1})$

$$Y_t - pY_{t-1} = \beta X_t + (\varepsilon_t - p\varepsilon_{t-1})$$

Notice this is a function of observed current income, X_t and hence, is estimable provided we know p. However, the residuals, say, $u_t = (\varepsilon_t - p\varepsilon_{t-1})$ has non-zero covariance:

$$E(u_t u_{t-1}) = E[(\varepsilon_t - p\varepsilon_{t-1}).(\varepsilon_{t-1} - p\varepsilon_{t-2})] = E[(\varepsilon_t\varepsilon_{t-1} - p\varepsilon_t\varepsilon_{t-2} - p\varepsilon^2_{t-1} - p^2\varepsilon_{t-1}\varepsilon_{t-2}]$$
$$= E[(0 - p.0 - p\varepsilon^2_{t-1} - p^2 0] = -p\ \sigma^2_\varepsilon \neq 0$$

Hence, the needed transformation to convert the model into an estimable form generates intrinsic SC in the residuals with $E(u_t u_{t-1}) \neq 0$

Diagnosis of Model Specification: a) Look at the residual plot ($\hat{u}_t$), this may tell you whether you have a non-linear model as a source of SC. Functional form of your model may not be linear, and this may cause SC in the residuals b) Explore if you may have omitted variables in your model, again it may create SC in the residuals.

Disturbances with AR(p) (autoregressive of order p) structure

Suppose SC is present in the following AR(1) form in the residuals such that

$Y_t = \alpha + \beta X_t + u_t$ where $u_t = \rho u_{t-1} + \varepsilon_t$ and ε_t is white-noise iid with

$E(\varepsilon_t \varepsilon_{t-s}) = 0$ and $E(\varepsilon_t^{\,2}) = \sigma^2_{\varepsilon}$, and $1 < \rho < -1$ but

$Cov(u_t u_{t-1}) = E(u_t u_{t-1}) = E[(\rho u_{t-1} + \varepsilon_t) u_{t-1}] = \rho \sigma_u^{\,2} \neq 0$ if $\rho > 0$ then there is positive SC, if not negative SC.

In general, $Cov(u_t u_{t-s}) = \rho^s \sigma_u^{\,2}$

Proofs that a) $E(u_t) = 0$, b) $Var(u_t) = \sigma_u^{\,2} = \dfrac{\sigma_{\varepsilon}^{\,2}}{1-\rho^2}$ and

c) $Cov(u_t u_{t-s}) = \rho^s \sigma_u^{\,2}$

Note that we can write $u_t = \rho u_{t-1} + \varepsilon_t$ as

$$u_t = \rho u_{t-1} + \varepsilon_t = \varepsilon_t + \rho(\varepsilon_{t-1} + \rho u_{t-2}) = \varepsilon_t + \rho \varepsilon_{t-1} + \rho^2(\varepsilon_{t-2} + \rho u_{t-3})$$

or

$= \varepsilon_t + \rho \varepsilon_{t-1} + \rho^2 \varepsilon_{t-2} + \rho^3 \varepsilon_{t-3} \ldots\ldots$ since $E(\varepsilon_t = \ldots\ldots = \varepsilon_{t-s}) = 0$, we have

$E(u_t) = 0$ (end of proof)

Since $u_t = \varepsilon_t + \rho \varepsilon_{t-1} + \rho^2 \varepsilon_{t-2} + \rho^3 \varepsilon_{t-3} \ldots\ldots$ then

$$Var(u_t) = \sigma_u^{\,2} = Var(\varepsilon_t) + \rho^2 Var(\varepsilon_{t-1}) + \rho^4 Var(\varepsilon_{t-2}) \ldots$$

which can be written as $Var(u_t) = \sigma_u^{\,2} = \sigma_{\varepsilon}^{\,2}(1 + \rho^2 + \rho^4 \ldots) = \dfrac{\sigma_{\varepsilon}^{\,2}}{1-\rho^2}$

(end of proof)

Notice that an infinite series will only sum to a finite number iff $|\rho| < 1$

Finally, Proof that $Cov(u_t u_{t-s}) = \rho^s \sigma_u^{\ 2}$ Since

$Cov(u_t u_{t-s}) =$

$$E[(\varepsilon_t + \rho\varepsilon_{t-1} + \rho^2\varepsilon_{t-2} + \rho^3\varepsilon_{t-3} \ldots)(\varepsilon_{t-s} + \rho\varepsilon_{t-s-1} + \rho^2\varepsilon_{t-s-2} + \rho^3\varepsilon_{t-s-3} \ldots)]$$

Multiplying and taking expectations, we get

$$E[\rho^s \varepsilon^2_{t-s} + \rho^{s+2}\varepsilon^2_{t-s-1} + \rho^{s+4}\varepsilon^2_{t-s-2} + \ldots]$$

$$= \rho^s \sigma_\varepsilon^{\ 2}(1 + \rho^2 + \rho^4 + \ldots) = \frac{\rho^s \sigma_\varepsilon^{\ 2}}{1-\rho^2} = \rho^s \sigma_u^{\ 2}$$ (end of proof)

Consequences of Ignoring Serial Correlation

---OLS coefficients are still unbiased and consistent but inefficient (if no lagged dependent on the RHS as an explanatory variable, if present, OLS is biased and inconsistent)

---Forecasts inefficient (again if lagged dependent variable on the RHS, biased also)

---Variances of coefficients biased and tests are invalid

---Rsq will overestimate the fit, indicating a better fit than actually present, and t values imply significance when in essence insignificant coefficients.

TESTING for SERIAL COORELATION

1) **Durbin Watson Statistic, d:** provides a test of $H_o : \rho = 0$ (No AR(1)) in the following specification for the error terms, $u_t = \rho u_{t-1} + \varepsilon_t$. If the test is rejected, there is evidence for AR(1) or *first-order serial correlation* (auto-regressive process of order 1). After your regression, issue the command dwstat to obtain the durbin-watson statistic. By checking the DW table for critical values, you can test for the above hypothesis.

Remarks: a) If d=2, no serial correlation. If d<2, there is positive serial correlation and if d>2, there is negative serial correlation.

This is because $d = \dfrac{\sum(\hat{u}_t - \hat{u}_{t-1})^2}{\sum \hat{u}_t^2}$ and $\hat{\rho} = \dfrac{\sum(\hat{u}_t . \hat{u}_{t-1})}{\sum \hat{u}_t^2}$ (estimated serial correlation coefficient) and $d \approx 2(1-\hat{\rho})$

If there is no serial correlation, $\hat{\rho} = 0$ then $d \approx 2$

If there is positive serial correlation, i.e. $\hat{\rho} > 0$ then $d < 2$

If there is negative serial correlation, i.e. $\hat{\rho} < 0$ then $d > 2$

b) DW test is not valid if there are lagged values of the dependent variable on the right hand side of the equation (in this case use Breusch-Godfrey LM test or Durbin's h-Test).
c) Not valid for higher order serial correlation.

TESTING WITH DW STATISTIC

a) Testing for positive AR(1)

Test $H_o: \rho = 0$ against $H_A: \rho > 0$ (There is + SC)
Look the Table in the Appendix (A.5) for critical values, d_L and d_u at α of 1%, 5% or 10%, and note k' is the number of coefficients in your regression excluding the constant. a) Reject the null if $d \le d_L$, b) if $d \ge d_U$, do not reject the null, c) if $d_L < d < d_U$, the test is inconclusive (Use the LM test below)

b) Testing for negative AR(1)

Test $H_o: \rho = 0$ against $H_A: \rho < 0$ (There is - SC)

Compute $4-d$
a) Reject the null if $4-d \le d_L$, b) if $4-d \ge d_U$, do not reject the null, c) if $d_L < 4-d < d_U$, the test is inconclusive (Use the LM test below)

c) Two tailed test on AR(1)

Test $H_o: \rho = 0$ against $H_A: \rho \neq 0$ (There is – or + SC)
a) Reject the null if $d \leq d_L$ or $4 - d \leq d_L$, b) if $d_U \leq d \leq 4 - d_U$, do not reject the null, c) if $d_L < d < d_U$ or $4 - d_U < d < 4 - d_L$, the test is inconclusive (Use the LM test below)

Warning: (*) You can not use this test when there is a lagged dependent variable on the RHS (*)

Example: $y_t = \alpha + \beta_1 y_{t-1} + \beta_2 x_t + u_t$ with lagged dependent variable on the RHS.

If this is the case, use either durbin's h test (issue **durbina** after regression) or **bgodfrey** for testing for models with lagged dependent variables on the RHS.

2) **Breusch-Godfrey Serial Correlation LM test:** can be used for AR(1) and higher orders of serial correlation like AR(2), AR(3) etc.

Example: $u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \rho_3 u_{t-3} + \varepsilon_t$ ➔AR(3) structure

STATA: After running your regression (with or without the lagged dependent on RHS), issue the command **bgodfrey**, **lags (1 2 3)** to test for higher order serial correlation.

The null is $H_o: \rho_1 = \rho_2 = \rho_3 = 0$ if you are testing for AR(3) structure.

3) **Correlograms and Q-Statistics:** This is a combination of visual and direct test of serial correlation which gives you an idea as to the *order of serial correlation* as well as whether there exists serial correlation in your regression equation.

Stata: Use Corrgram (variable list) command in Stata. Autocorrelation (AC) and partial autocorrelations (PAC) along with Q-statistic and its associated p-value will be displayed. If there is no serial correlation, AC and PAC at all lags should be equal to zero and Q-stat should be insignificant with large p-values.

4) **Durbin's h-test (Stata command: durbina):** can be used also when lagged dependent variables exists, but only for AR(1), not for higher order SC. Steps involved in this test are as follows:

Step 1: Estimate the model by OLS and obtain the residuals, $\hat{u}_t$.

Step 2: Estimate $\hat{\rho}$ from $(2-d)/2 = \hat{\rho}$ relationship.

Step 3: Construct the following statistic, called Durbin's h-statistic,

$$h = \hat{\rho}.\left[\frac{n'}{1-n'.s^2_{\hat{\beta}}}\right]^{1/2}$$ where n' is the number of observations-1 and $s^2_{\hat{\beta}}$ is the variance of the coefficient in front of the lagged dependent variable. In large samples, this statistic has a normal distribution and hence reject the $H_o : \rho = 0$ against $H_A : \rho \neq 0$ when $|h| > z^*$, the critical value of z.

ESTIMATING AR/SC MODELS (AR(p))

Before you specify your model, taking into account serial correlation, make sure that the source of serial correlation is not misspecification! This is because a misspecified model is the most common source of serial correlation and this can simply be corrected by taking logs, specifying non-linear models etc.

Once you make sure that SC is inherent in the residuals, you should estimate your model with the help of Stata.

Examples: Suppose you wish to regress Consumption on GDP in the following model: $CS_t = \alpha + \beta GDP_t + u_t$ where $u_t = \rho u_{t-1} + \varepsilon_t$ with AR (1).

In Stata, you should specify: newey Y X1 X2, lag(3) if AR(3) is present. The estimates will be consistent and unbiased.

Previous tests and correlograms will help you determine whether you should use only AR(4) or other lags as well.

Remark: Simple OLS with AR(.)s is OK provided that regressors (independent variables) are not correlated. If regressors are correlated (as in most data) and there is evidence for serial correlation at different orders, you should use the TSLS (Two-stage least squares) option in estimation. Will be discussed later.

Example (from the book, pg. 406, fifth edition): Data 9-3 has quarterly data to model the consumption of electricity by residential customers served by the San Diego Gas and Electric Company with the following variables:

1) RSKWH=Electricity sales to residential customers
2) NOCUST=Number of residential customers
3) PRICE=Average price for the single-family rate tariff
4) CPI=San Diego Consumer Price Index (1982-84=100)
5) INCM=County's total personal income
6) CDD=Cooling degree rates
7) HDD=Heating degree days

8) POP=County's population

Here, a double-log model is estimated (implies that coefficients are constant elasticities) with the following necessary transformation of variables:

. generate float lkwh=log(reskwh/nocust) ➔log of electricity sales per residential customer
. generate float ly=log(100*incm/cpi*pop)➔per-capita income in constant 82-84 dollars
. generate float lprice=log(100*price/cpi)➔Price of electricity in real constant dollars

Weather is one of the most important determinant of electricity consumption so CDD and HDD will be also included (see book on the details of their computation) and are expected to have positive effect on electricity consumption.

<u>The Basic Model</u>

$$lkwh_t = \alpha + \beta_1 ly + \beta_2 lprice + \beta_3 cdd + \beta_4 hdd + u_t$$

Testing for AR(4)➔ with quarterly data, most appropriate, with

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \rho_3 u_{t-3} + \rho_4 u_{t-4} + \varepsilon_t$$

Since DW test is invalid for higher orders of SC, an LM test is most appropriate here.

Stata: Obtain a regression of the basic model and then issue The following output is obtained and there is strong evidence for the presence of a serious serial correlation of order 4 based on the following output. Also check the correlogram: Q-statistics are all significant and there are spikes on all four lags.

reg lkwh ly lprice cdd hdd

```
     Source |       SS       df       MS              Number of obs =      87
-------------+------------------------------           F(  4,    82) =   45.90
      Model |  .387886229     4  .096971557           Prob > F      =  0.0000
   Residual |  .173234745    82  .002112619           R-squared     =  0.6913
-------------+------------------------------           Adj R-squared =  0.6762
      Total |  .561120974    86  .006524662           Root MSE      =  .04596

------------------------------------------------------------------------------
        lkwh |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          ly |  -.0333977   .0131665    -2.54   0.013    -.0595901   -.0072053
      lprice |  -.0855861   .0262276    -3.26   0.002    -.1377612    -.033411
         cdd |   .0002652   .0000331     8.01   0.000     .0001993     .000331
         hdd |   .0003569   .0000288    12.40   0.000     .0002996    .0004141
       _cons |   .8853698   .2190091     4.04   0.000      .449691    1.321049
------------------------------------------------------------------------------
```

tsset time (enter numbers from 1 to 87 for the time variable in your data set)

time variable: time, 1 to 87

dwstat➔ generates the Durbin Watson statistic

bgodfrey, lags(1 2 3 4)

Breusch-Godfrey LM test for autocorrelation

lags(p)	chi2	df	Prob > chi2
1	11.868	1	0.0006
2	25.700	2	0.0000
3	36.579	3	0.0000
4	51.464	4	0.0000

H0: no serial correlation

Clearly, you should reject the null in favor of AR(4) based on the B-G LM test results.

Estimation with AR(4): Regression with Newey standard errors under serial correlation (to gain efficiency!)

newey lkwh ly lprice cdd hdd, lag(4)

Regression with Newey-West standard errors   Number of obs =   87
maximum lag: 4   F(4,   82) =   93.00
Prob > F   =   0.0000

lkwh	Coef.	Newey-West Std. Err.	t	P>|t|	[95% Conf. Interval]	
ly	-.0333977	.015277	-2.19	0.032	-.0637886	-.0030068
lprice	-.0855861	.0260454	-3.29	0.001	-.1373987	-.0337735
cdd	.0002652	.0000329	8.07	0.000	.0001998	.0003306
hdd	.0003569	.0000213	16.77	0.000	.0003146	.0003992
_cons	.8853698	.2779385	3.19	0.002	.3324615	1.438278

prais lkwh ly lprice cdd hdd➔ regression with AR(1) disturbances (Cochrane-Orcutt iterative producedure). Use prais in such a case. Output is below.

Iteration 0: rho = 0.0000
Iteration 1: rho = 0.3212
Iteration 2: rho = 0.3270
Iteration 3: rho = 0.3271
Iteration 4: rho = 0.3271
Iteration 5: rho = 0.3271

Prais-Winsten AR(1) regression -- iterated estimates

Source | SS   df   MS   Number of obs =   87

```
-------------+------------------------------           F(  4,    82) =   61.04
       Model |  .459925271     4  .114981318           Prob > F      =  0.0000
    Residual |  .154475835    82  .001883852           R-squared     =  0.7486
-------------+------------------------------           Adj R-squared =  0.7363
       Total |  .614401106    86  .007144199           Root MSE      =   .0434

------------------------------------------------------------------------------
        lkwh |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          ly |  -.0299038   .0181739    -1.65   0.104    -.0660574    .0062498
      lprice |  -.0836747   .0355569    -2.35   0.021    -.1544087   -.0129408
         cdd |   .0002654   .0000265    10.01   0.000     .0002127    .0003181
         hdd |   .0003627   .0000237    15.33   0.000     .0003157    .0004098
       _cons |   .8215591   .3006255     2.73   0.008     .2235191    1.419599
-------------+----------------------------------------------------------------
         rho |   .3271167
------------------------------------------------------------------------------
Durbin-Watson statistic (original)    1.316929
Durbin-Watson statistic (transformed) 1.758268
```

Modeling Structural Change: The electricity price has significantly changed over the sample period and hence, we define three dummies, to reduce this incidence of misspecification,

D74=1 for 1974.1 onward, 0 otherwise
D79=1 for 1979.1 onward, 0 otherwise
D83=1 for 1983.3 inward, 0 otherwise

and interactive dummies with existing variables. First, a general model with all of the variables were specified and then, insignificant ones were eliminated one by one to arrive at the following model (after testing for serial correlation with LM test)

FINAL Model

newey lkwh lyd79 lyd83 lprd79 lprd83 hddd83, lag(4)

```
Regression with Newey-West standard errors            Number of obs  =        87
maximum lag: 4                                        F(  5,    81)  =     19.37
                                                      Prob > F       =    0.0000

------------------------------------------------------------------------------
             |             Newey-West
        lkwh |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       lyd79 |   .0387181   .0084782     4.57   0.000     .0218492     .055587
       lyd83 |  -.0392496   .0095129    -4.13   0.000    -.0581773   -.0203219
      lprd79 |  -.2837981   .0590021    -4.81   0.000    -.4011937   -.1664024
      lprd83 |   .2427546   .0664567     3.65   0.000     .1105267    .3749825
      hddd83 |   .0001448    .000018     8.02   0.000     .0001089    .0001807
       _cons |   .3569166   .0096886    36.84   0.000     .3376393    .3761939
------------------------------------------------------------------------------
```

Time Series Operators

generate float lagly=l.ly➔ generates ly_{t-1}
generate float Dly=d.ly➔ generates ly_t-ly_{t-1}

Use L2.ly to generate ly_{t-2}

generate float gnplag2=l2.gnp➔ generates gnp_{t-2}