Introductory Econometrics

Serial Correlation

Monash Econometrics and Business Statistics

2022

Recap

The multiple regression model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + u_i, \;\; i = 1, 2, ...n.$$

A1 model is linear in parameters: $y = X\beta + u$.

A2 columns of X are linearly independent.

A3 conditional mean of errors is zero: $E(u|X) = 0$.

A4 homoskedasticity and no serial correlation: $Var(u|X) = \sigma^2 I_n$.

A5 errors are normally distributed: $u|X \sim N(0, \sigma^2 I_n)$.

No serial correlation

The multiple regression model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + u_i,\ \ i = 1, 2, ...n.$$

A4 homoskedasticity and no serial correlation: $Var(u|X) = \sigma^2 I_n$.

$$Var(u|X) = \begin{pmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{pmatrix}$$

Lecture Outline

1. Definition of serial correlation

A4 homoskedasticity and no serial correlation: $Var(u|X) = \sigma^2 I_n$.

A4(a) homoskedasticity: $Var(u_1|X) = ... = Var(u_n|X) = \sigma^2$.

A4(b) no serial correlation: $Cov(u_{i,} u_j|X) = 0$ for all $i \neq j$.

When A4(b) does not hold, the error terms in u are serially correlated:

$$Var(u|X) = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{12} & \sigma_2^2 & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1n} & \sigma_{2n} & \cdots & \sigma_n^2 \end{pmatrix}.$$

2. Causes of serial correlation

Time series data is very likely to show serial correlation.

Example:

- Number of confirmed covid cases:
  The number of confirmed covid cases today is very much correlated with the number of confirmed covid cases yesterday.

- A simple model with serial correlation in the error term:

$$
\begin{aligned}
y_t &= \beta_0 + \beta_1 x_{t1} + \ldots\ldots + \beta_k x_{tk} + u_t, \\
u_t &= \phi_1 u_{t-1} + e_t, \quad e_t \sim i.i.d(0, \sigma^2),
\end{aligned}
$$

  where the subscript t rather than i indicates time series data.

- Let the errors in the linear regression model be generated by:

$$u_t = \phi_1 u_{t-1} + e_t, \quad e_t \sim i.i.d(0, \sigma^2).$$

- It can be shown that

$$Cov(u_t, u_{t-j}|X) = \frac{\phi_1^j \sigma^2}{1 - \phi_1^2} \neq 0 \text{ if } \phi_1 \neq 0.$$

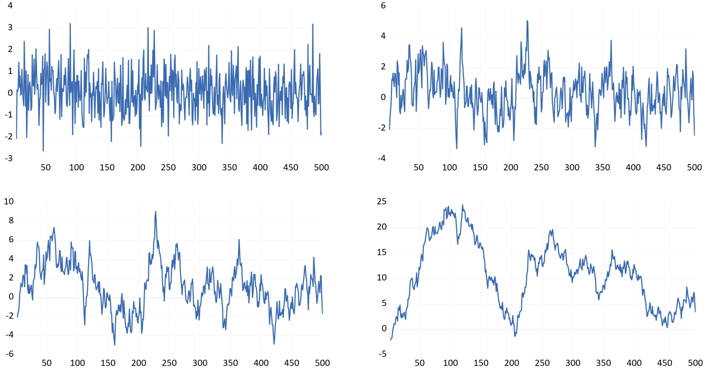
- This violates A4(b):

$$\text{no serial correlation: } Cov(u_{i,} u_j|X) = 0 \text{ for all } i \neq j.$$

- Let the errors in the linear regression model be generated by:

$$u_t = \phi_1 u_{t-1} + e_t, \quad e_t \sim i.i.d(0, \sigma^2).$$

- Line graph of the errors with $\phi_1 = \{0.00, 0.70, 0.90, 0.99\}$:



3. Consequences of serial correlation

- Serial correlation does not affect A1-A3:
  - the OLS estimator remains unbiased.
- Serial correlation violates A4:
  - the OLS estimator is no longer BLUE.
  - $Var(\hat{\beta}) \neq \sigma^2(X'X)^{-1}$.
    - default standard errors are incorrect.
      - default t and F tests are incorrect.

4. Detecting serial correlation

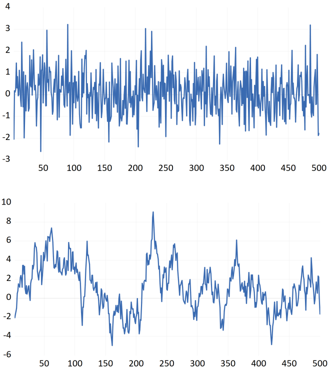
4.1 The line graph of the residuals

4.2 The correlogram of the residuals

4.3 The Breusch-Godfrey test for serial correlation

4.1 The line graph of the residuals

- Example:



- But we cannot observe the actual errors from a linear regression!

4.1 The line graph of the residuals

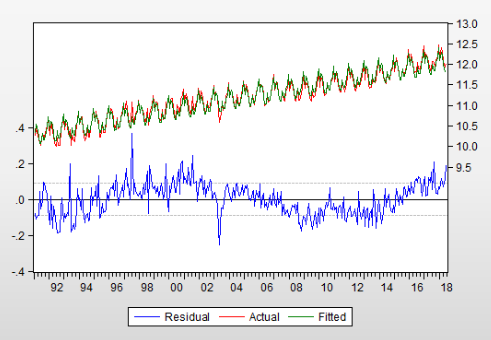
- We are interested in whether $\{u_t\}$ is serially correlated.
- We cannot observe the errors from a linear regression.
- We can observe the residuals from the estimated regression.
- We use the observed residuals as proxies for the unobserved errors.
- Inspect the line graph of the residuals to assess serial correlation.

Example:

- Consider the linear regression equation

$$\log(Vic_t) = \beta_0 + \beta_1 time_t + \sum_{i=1}^{11} \alpha_i Q_{ti} + u_t,$$

  where $\log(Vic)$ is the natural logarithm of monthly international tourist arrivals in Victoria, time is a time trend and $Q_i, \;\; i = 1, 2, ..., 11$ is a set of monthly dummy variables.

- 

4.2 The correlogram of the residuals

The correlogram shows the estimated autocorrelations of a time series.

- Autocorrelations are the correlations with its own lags j.
- Suppose that we estimate the linear regression equation

  $$y_t = \beta_0 + \beta_1 x_t + u_t,$$

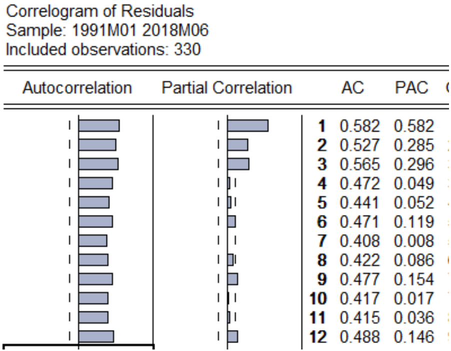
  and obtain the OLS residuals

  $$\widehat{u}_t = y_t - \widehat{\beta}_0 - \widehat{\beta}_1 x_t.$$

- The correlogram of the residuals shows $Corr(\widehat{u}_t, \widehat{u}_{t-j})$.

Example:

- Monthly international visitor arrivals in Victoria



Autocorrelation	Partial Correlation		AC	PAC
		1	0.582	0.582
		2	0.527	0.285
		3	0.565	0.296
		4	0.472	0.049
		5	0.441	0.052
		6	0.471	0.119
		7	0.408	0.008
		8	0.422	0.086
		9	0.477	0.154
		10	0.417	0.017
		11	0.415	0.036
		12	0.488	0.146

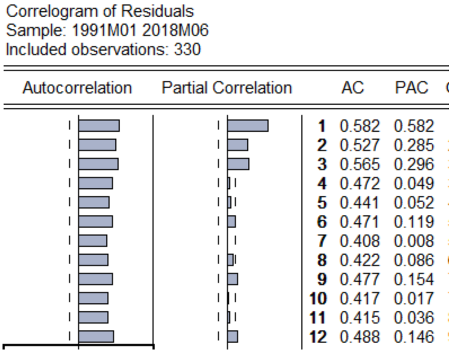
- Column 3: Autocorrelation (AC) $\hat{\rho}_j = Corr(\widehat{u}_t, \widehat{u}_{t-j})$.
- Column 1: Bar charts $\hat{\rho}_j$ with 95% confidence bands.
    - If $\hat{\rho}_j$ outside the bands, reject $H_0 : \rho_j = Corr(u_t, u_{t-j}) = 0$.
- Column 4: Partial autocorrelation coefficients (PAC):
    - Coefficient estimates final lagged error terms:

$$
\begin{aligned}
u_t &= \boldsymbol{\phi}_1 u_{t-1} + e_t, \\
u_t &= \phi_1 u_{t-1} + \boldsymbol{\phi}_2 u_{t-2} + e_t, \\
u_t &= \phi_1 u_{t-1} + \phi_2 u_{t-2} + \boldsymbol{\phi}_3 u_{t-3} + e_t.
\end{aligned}
$$

- Column 2: Bar charts $\hat{\phi}_j$ with 95% confidence bands.
    - If $\hat{\phi}_j$ outside the bands, reject $H_0 : \phi_j = 0$.

Example:

- Monthly international visitor arrivals in Victoria



Example:

- All $\widehat{\rho}_j s$ are outside their confidence bands, so reject

$$H_0 : \rho_j = 0,\ j = 1, 2, ..., 12.$$

- This suggests serially correlated errors in the linear regression

$$\log(Vic_t) = \beta_0 + \beta_1 time_t + \sum_{i=1}^{11} \lambda_i Q_{ti} + u_t.$$

- The first three $\widehat{\phi}_j s$ are outside their confidence bands, so reject

$$H_0 : \phi_j = 0,\ j = 1, 2, 3.$$

- This suggests an AR(3) process of the form

$$u_t = \phi_0 + \phi_1 u_{t-1} + \phi_2 u_{t-2} + \phi_3 u_{t-3} + e_t.$$

4.3 The Breusch-Godfrey test for serial correlation

Consider the linear regression equation

$$y_t = \beta_1 + \beta_2 x_{t2} + ... + \beta_k x_{tk} + u_t,$$

and assume that the errors are autoregressive of order q:

$$u_t = \phi_1 u_{t-1} + \phi_2 u_{t-2} + ... + \phi_q u_{t-q} + e_t,\ \ e_t \sim i.i.d(0, \sigma^2).$$

- The null and the alternative of the test can be written as:

$$H_0 : \phi_1 = \phi_2 = ... = \phi_q = 0,$$
$$H_1 : \phi_j \neq 0 \text{ for at least one } j = 1, 2, ..., q.$$

- Determine q with reference to the frequency of the data (annual 1 or 2, quarterly 4, ...).

4.3 The Breusch-Godfrey test for serial correlation

1. Obtain the OLS residuals $\widehat{u}_t$ for $t = 1, \dots, n$ from the model:

$$y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \cdots + \beta_k x_{tk} + u_t, \; t = 1, \dots, n.$$

2. Obtain the R-squared $R^2_{\widehat{u}}$ from the auxiliary regression:

$$\widehat{u}_t = \alpha_1 + \alpha_2 x_{t2} + ... + \alpha_k x_{tk} + \phi_1 \widehat{u}_{t-1} + ... + \phi_q \widehat{u}_{t-q} + e_t.$$

3. Under $H_0 : \phi_1 = \phi_2 = ... = \phi_q = 0$, we have the test statistic:

$$BG = (n-q) R^2_{\widehat{u}} \overset{asy}{\sim} \chi^2(q).$$

4. Reject H_0 in favor of $H_1 : \phi_j \neq 0$ for at least one $j = 1, 2, ..., q$, if

$$BG_{calc} > \chi^2_{crit}(q).$$

4.3 The Breusch-Godfrey test for serial correlation

- An alternative way to conduct the BG test is to estimate

  $$\widehat{u}_t = \alpha_1 + \alpha_2 x_{t2} + ... + \alpha_k x_{tk} + \phi_1 \widehat{u}_{t-1} + ... + \phi_q \widehat{u}_{t-q} + e_t.$$

  and perform a standard F test of $H_0 : \phi_1 = \phi_2 = ... = \phi_q = 0$.

- Remember that we must choose the value of q for the BG test.

Example:

- Monthly international visitor arrivals in Victoria

1. Obtain the OLS residuals $\widehat{u}_t$ for $t = 1, \ldots, 330$ from the model:

$$\log(Vic_t) = \beta_0 + \beta_1 time_t + \sum_{i=1}^{11} \lambda_i Q_{ti} + u_t, \;\; t = 1, \ldots, n.$$

2. Obtain the R-squared $R^2_{\widehat{u}} = 0.504$ from the auxiliary regression:

$$\widehat{u}_t = \alpha_1 + \alpha_2 time_t + \sum_{i=1}^{11} \gamma_i Q_{ti} + \phi_1 \widehat{u}_{t-1} + \phi_2 \widehat{u}_{t-2+\ldots} .. + \phi_{12} \widehat{u}_{t-12} + e_t.$$

3. Under $H_0 : \phi_1 = \phi_2 = ... = \phi_{12} = 0$, we have the test statistic:

$$BG = (330 - 12) R^2_{\widehat{u}} \overset{asy}{\sim} \chi^2(12).$$

4. Reject H_0 in favor of $H_1 : \phi_j \neq 0$ for at least one $j = 1, 2, ..., 12$, if

$$BG_{calc} = 318 \times 0.504 = 160.27 > \chi^2_{crit}(12) = 21.03.$$

- Note we use $n - q$ to compute the BG test statistic.
- The reason is that we lose q observations when we form q lags.
- Suppose we have 5 observation on the time series $\{\widehat{u}_t\}$.
- Each time we lag $\{u_t\}$ one time period, we lose an observation:

Table 1			
t	$\{\widehat{u}_t\}$	$\{\widehat{u}_{t-1}\}$	$\{\widehat{u}_{t-2}\}$
1	$\widehat{u}_1$	-	-
2	$\widehat{u}_2$	$\widehat{u}_1$	-
3	$\widehat{u}_3$	$\widehat{u}_2$	$\widehat{u}_1$
4	$\widehat{u}_4$	$\widehat{u}_3$	$\widehat{u}_2$
5	$\widehat{u}_5$	$\widehat{u}_4$	$\widehat{u}_3$

- Some software packages compute the BG test statistic differently.
- EViews replaces all missing values in lags with zero.
- Eviews uses $BG = nR^2_{\widehat{u}}$ instead of $BG = (n-q)R^2_{\widehat{u}}$.
- So both n and R^2 of the auxiliary regression are different.

Table 2			
t	$\{\widehat{u}_t\}$	$\{\widehat{u}_{t-1}\}$	$\{\widehat{u}_{t-2}\}$
1	$\widehat{u}_1$	0	0
2	$\widehat{u}_2$	$\widehat{u}_1$	0
3	$\widehat{u}_3$	$\widehat{u}_2$	$\widehat{u}_1$
4	$\widehat{u}_4$	$\widehat{u}_3$	$\widehat{u}_2$
5	$\widehat{u}_5$	$\widehat{u}_4$	$\widehat{u}_3$

Breusch-Godfrey Serial Correlation LM Test:

F-statistic	24.98731	Prob. F(12,305)	0.0000
Obs*R-squared	163.5945	Prob. Chi-Square(12)	0.0000

Test Equation:
Dependent Variable: RESID
Method: Least Squares
Sample: 1991M01 2018M06
Included observations: 330
Presampple missing value lagged residuals set to zero.

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.003345	0.014325	-0.233480	0.8155
T	2.63E-05	3.85E-05	0.683692	0.4947
@MONTH=1	-0.000505	0.017920	-0.028165	0.9775
@MONTH=2	0.000294	0.017922	0.016424	0.9869
...				
...				
RESID(-1)	0.279408	0.056746	4.923856	0.0000
RESID(-2)	0.121981	0.058959	2.068922	0.0394
...				
RESID(-11)	0.006485	0.059677	0.108665	0.9135
RESID(-12)	0.162810	0.057558	2.828621	0.0050

R-squared	0.495741	Mean dependent var	-3.40E-16
Adjusted R-squared	0.456061	S.D. dependent var	0.090042

5. HAC standard errors

Recall that the two consequences of heteroskedasticity are:

- The OLS estimator of β is no longer BLUE.
- The standard t and F tests are no longer valid.

So we cannot conduct reliable hypothesis tests anymore!

- Whitney Newey and Kenneth West, proposed alternative hypothesis tests which are valid in large samples, even when serial correlation is present.

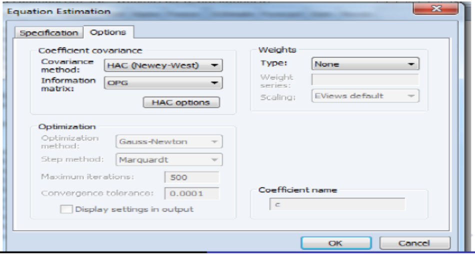
5. HAC standard errors

Hypothesis tests proposed by Newey and West use

- a different formula for the estimated variance matrix of $\widehat{\beta}$.
- therefore different standard errors for each $\widehat{\beta}_j$.
- HAC (heteroskedasticity and autocorrelation consistent) standard errors instead.
- t and F tests based on HAC standard errors
  - which are reliable in large samples, even in the presence of heteroskedasticity and autocorrelation.

Example

- Consider Victoria's international tourist arrivals again.
- Newey-West HAC estimate of variance can be chosen in EViews estimation window under 'Options'



Dependent Variable: LOG(VIC)
Method: Least Squares
Sample: 1991M01 2018M06
Included observations: 330

Variable	Coefficient	Std. Error	t-Statistic
C	10.66171	0.019773	539.2010
T	0.005401	5.30E-05	101.8685
@MONTH=1	-0.317922	0.024743	-12.84874
@MONTH=2	-0.135544	0.024743	-5.478088

Dependent Variable: LOG(VIC)
Method: Least Squares
Sample: 1991M01 2018M06
Included observations: 330
HAC standard errors & covariance (Bartlett kernel, Newey-West fixed bandwidth = 6.0000)

Variable	Coefficient	Std. Error	t-Statistic
C	10.66171	0.019836	537.5007
T	0.005401	0.000104	51.79185
@MONTH=1	-0.317922	0.013028	-24.40314
@MONTH=2	-0.135544	0.015520	-8.733471

Summary

- Serial correlation in the error term of a linear regression model:
- How to define serial correlation
- What implications does the existence of serial correlation have on the properties of the OLS estimator
- How to detect it (Breusch-Godfrey test)
- How to correct for it: HAC standard errors