POLS 6481. Research Design and Quantitative Methods II

Lecture 14. Time Series Data II: Serial Correlation

Wooldridge, *Introductory Econometrics*, 11.1 + 12.1 + 12.2 + 12.3 + 12.4

Outline

1. First-Differencing a Non-Stationary Series
2. Example of First-Differencing: *Fertility Rates*
3. Serial Auto-correlation
4. The Durbin-Watson Test for Serial Auto-correlation
5. Quasi-Differencing: Cochrane-Orcutt and Prais-Winsten Estimation
6. Example of Quasi-Differencing: *Divorce Rates*

POLS 6481. Research Design and Quantitative Methods II
Lecture 14. Time Series Data II: Serial Correlation
Wooldridge, *Introductory Econometrics*, 11.1 + 12.1 + 12.2 + 12.3 + 12.4

1. First-Differencing a Non-Stationary Series

Review your notes from Lecture 13 on stationarity and how to test for a unit root.

"simple transformations are available that render a unit root process weakly dependent"

"provided the time series we use are weakly dependent, usual OLS inference procedures are valid…"

Unit root processes… are said to be **integrated of order one**; the **first difference** of the process is said to be **integrated of order zero**. A time series that is I(1) is often said to be a **difference-stationary process**.

POLS 6481. Research Design and Quantitative Methods II
Lecture 14. Time Series Data II: Serial Correlation
Wooldridge, *Introductory Econometrics*, Examples 10.4 and 10.8 and 11.6

1½. Example of Unnecessarily* First Differencing a Non-Stationary Series: *Manatees*

Data: manatees.csv
Script: Lecture 13 manatees.R

N = 33 years from 1977 to 2009

kills is the number of manatees killed in a boating accident in a calendar year

boats is the number of boats registered in the state of Florida in a calendar year (x 1,000)

Static model:

$Kills_t = -43.2 + 0.13\ Boats_t$ $\quad (R^2 = .905)$

First-differenced model:

$\Delta\ Kills_t = 1.74 + 0.05\ \Delta\ Boats_t$ $\quad (R^2 = .011)$

Quasi-differenced model (Cochrane-Orcutt estimates, then Prais-Winsten estimates)

$Kills_t = -42.9 + 0.13\ Boats_t$ $\quad$ (*DW* before = 1.74; *DW* after = 1.89)

$Kills_t = -43.3 + 0.13\ Boats_t$ $\quad$ (*rho* = .092)

* Plot of residuals, correlations between residuals, etc., show there's no serial autocorrelation

POLS 6481. Research Design and Quantitative Methods II
Lecture 14. Time Series Data II: Serial Correlation
Wooldridge, *Introductory Econometrics*, Examples 10.4 and 10.8 and 11.6

2. Example of First Differencing a Non-Stationary Series: *Fertility Rates*

Data: FERTIL3.dta
Script: Lecture 14 fertility without dyn.R

N = 72 years from 1913 to 1984

gfr is the number of children per 1000 women of childbearing age
pe is the real value of personal tax exemptions – as exemptions increase, families might be incentivized to have larger families;
ww2 is a dummy variable for the years 1941–1945; and
pill is a dummy variable for the years 1963–1984, after birth control became available

Static model:

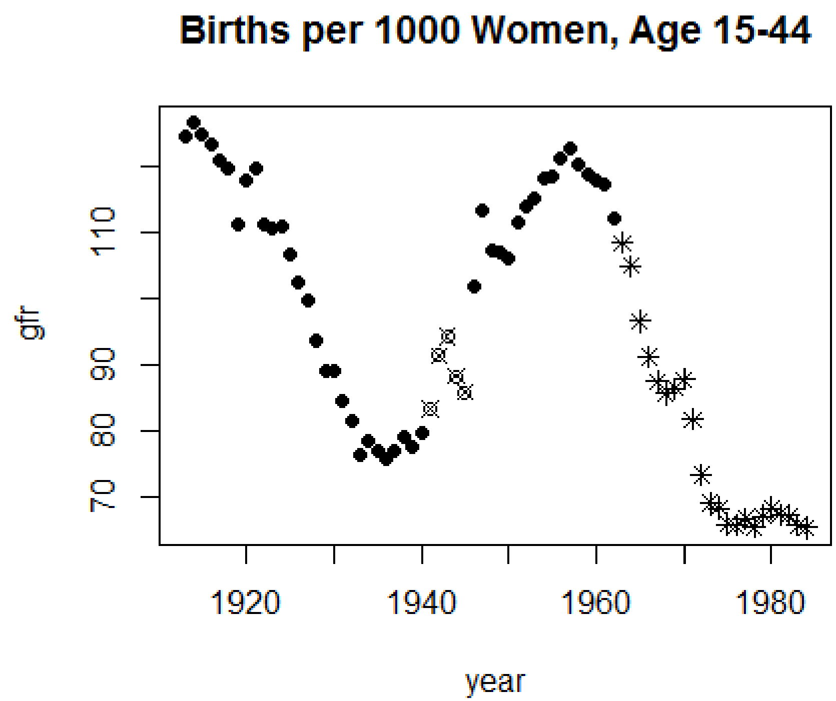
$$\widehat{gfr}_t = \underset{(3.68)}{98.68} + \underset{(.030)}{.083}\, pe_t - \underset{(7.46)}{24.24}\, ww2_t - \underset{(4.08)}{31.59}\, pill_t \qquad R^2 = .473, \bar{R}^2 = .450$$

L.A. Whittington, J. Alm, & H.E. Peters (1990) "Fertility and the Personal Exemption: Implicit Pronatalist Polity in the United States" *American Economics Review* 80: 545–556

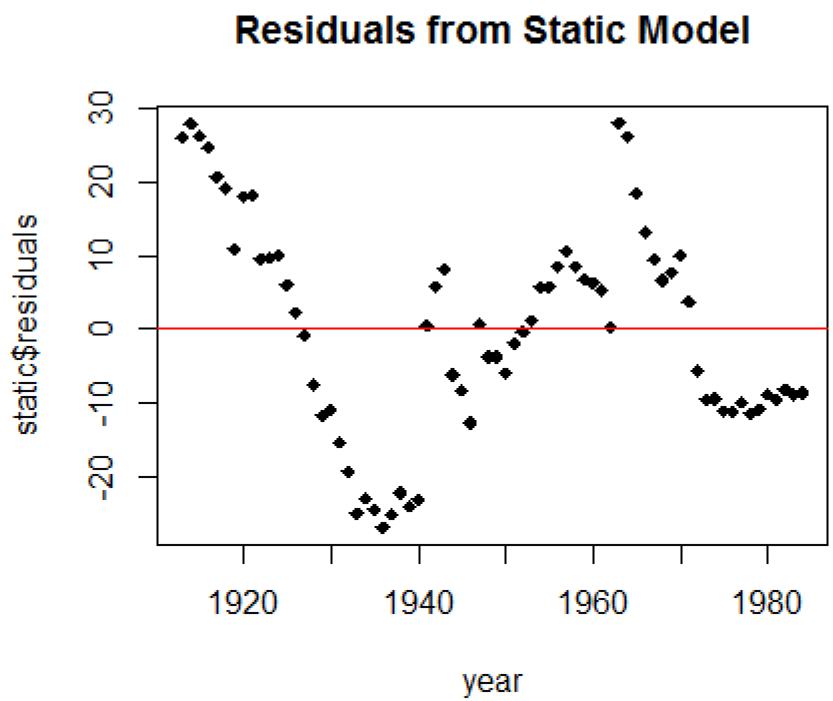
POLS 6481. Research Design and Quantitative Methods II
Lecture 14. Time Series Data II: Serial Correlation
Wooldridge, *Introductory Econometrics*, Examples 10.4 and 10.8 and 11.6

2. Example of First Differencing a Non-Stationary Series: *Fertility Rates*

plot of dependent variable over time



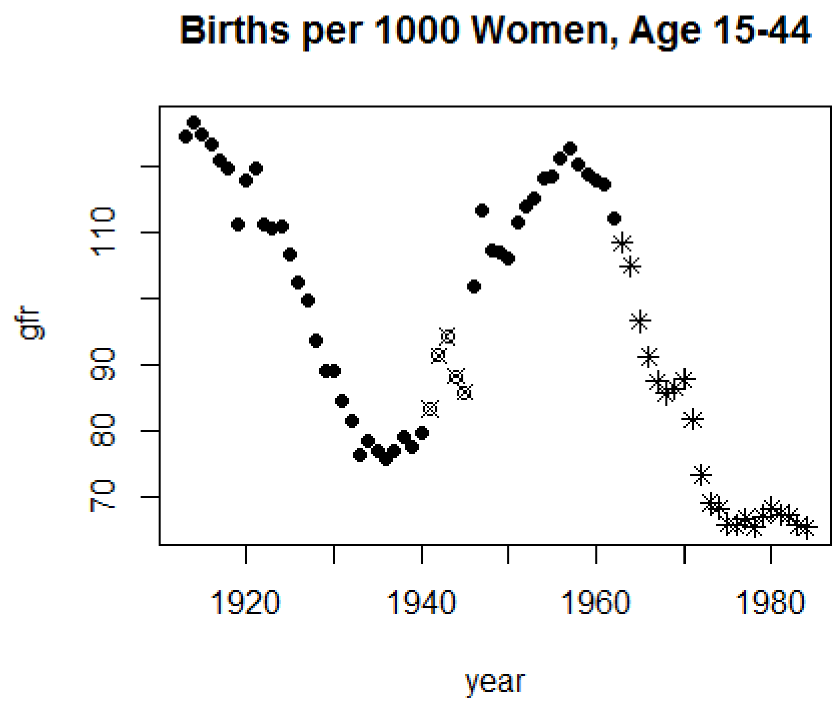
plot of model's residuals over time



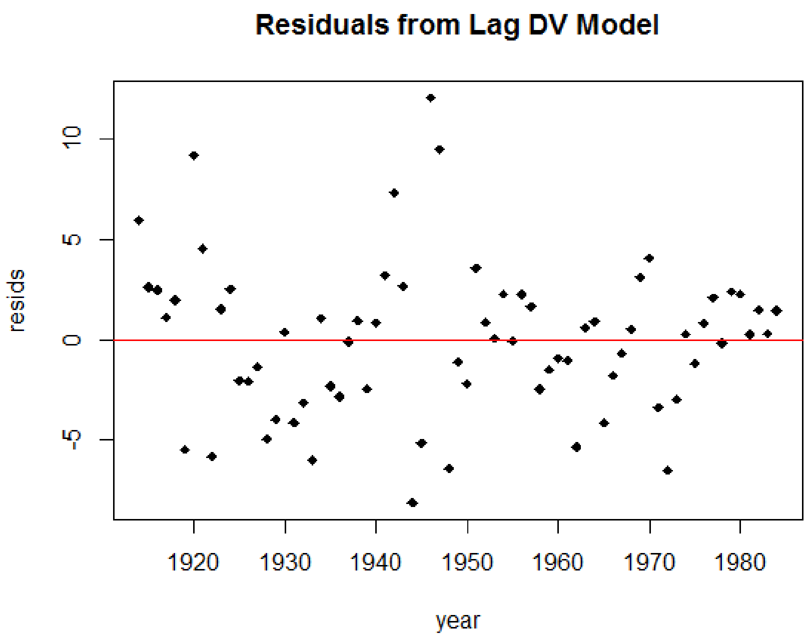
Data tend to stay above the line for many consecutive observations, then stay below the line for many consecutive observations; hence there are few 'runs' compared to wine example

2. Example of First Differencing a Non-Stationary Series: *Fertility Rates*

plot of dependent variable over time



plot of model's residuals over time
with lagged y variable



One option is to add the lagged value of the dependent variable as a regressor; this requires a package and/or a few lines of code if it isn't already in the dataset.

POLS 6481. Research Design and Quantitative Methods II
Lecture 14. Time Series Data II: Serial Correlation
Wooldridge, *Introductory Econometrics*, Examples 10.4 and 10.8 and 11.6

2. Example of First Differencing a Non-Stationary Series: *Fertility Rates*

Data: FERTIL3.dta
Script: Lecture 14 fertility without dyn.R

Example 11.6 in Wooldridge 5th Ed. (p. 397–398)

y is the number of children per 1000 women of childbearing age (*gfr*)
x is the real value of personal tax exemptions (*pe*)

Both the fertility series and the series of the personal tax exemption are highly persistent:

$$\hat{\rho}_{gfr} = .977, \;\; \hat{\rho}_{pe} = .964$$

Because they are not stationary we can estimate the equation in first differences:

$$\widehat{\Delta gfr_t} = -.785 - \; .043 \; \Delta pe_t$$
$$\quad\quad (.502) \quad (.028)$$

$n = 71$; $R^2 = .032$

The sign on *pe* is opposite our expectations… maybe there is a lag in responses?

POLS 6481. Research Design and Quantitative Methods II
Lecture 14. Time Series Data II: Serial Correlation
Wooldridge, *Introductory Econometrics*, Examples 10.4 and 10.8 and 11.6

2. Example of First Differencing a Non-Stationary Series: *Fertility Rates*

Data: FERTIL3.dta
Script: Lecture 14 fertility without dyn.R

Example 11.6 in Wooldridge 5^{th} Ed. (p. 397–398)

y is the number of children per 1000 women of childbearing age (*gfr*)
x is the real value of personal tax exemptions (*pe*)

Both the fertility series and the series of the personal tax exemption are highly persistent:

$$\hat{\rho}_{gfr} = .977, \;\; \hat{\rho}_{pe} = .964$$

Because they are not stationary we can estimate the equation in first differences:

$$\Delta\widehat{gfr} = \underset{(.468)}{-\;.964} - \underset{(.027)}{.036}\;\Delta pe - \underset{(.028)}{.014}\;\Delta pe_{-1} + \underset{(.027)}{.110}\;\Delta pe_{-2}$$

n = 69; R^2 = .233

We might come back to the lags of *pe* a little later… standard errors are large! But at least with two lags the results make sense – after a delay due to behavioral and biological factors…

3. Serial Auto-correlation

"Auto-correlation" is **not** the same thing as auto-regression:

- auto-regression refers to the relationship between y_t and y_{t-1}
- auto-correlation refers to the relationship between u_t and u_{t-1}

(Note that "spatial auto-correlation" also exists, but the ordering is not temporal)

A *lot* of econometrics books treat autocorrelation like a heteroscedasticity… also a *lot* of econometrics books deal with serial autocorrelation first, and auto-regression only later.

Assumption TS.5: $\text{corr}(u_t, u_s) = 0$, for all $t \neq s$
where $\{u_t: t = 1, \ldots, n\}$ is the sequence of disturbances.

Consequences of serial autocorrelation: standard errors are too small

Causes of serial autocorrelation: "often, serial correlation in the errors of a dynamic model simply indicates that the dynamic regression function has not been completely specified."

My interpretation: you might find *both* auto-regression and auto-correlation, until you correct for auto-regression (by augmenting the model specification or first-differencing).

4. The Durbin-Watson Test

DW test statistic is approximately $2(1-\hat{\rho})$, where $\hat{\rho}$ is coefficient in regression of $\hat{u}_t$ on $\hat{u}_{t-1}$.

Conduct test of whether DW = 2 by examining whether $\hat{\rho} = 0$

Null hypothesis: $H_0 : \rho = 0$; Alternative hypothesis: $H_1 : \rho > 0$

$$y_t = \beta_0 + \beta_1 x_{t1} + \cdots + \beta_k x_{tk} + u_t$$

$$u_t = \rho u_{t-1} + e_t$$

$$\hat{u}_t = \rho \hat{u}_{t-1} + error$$

$$DW = \sum_{t=2}^{n} (\hat{u}_t - \hat{u}_{t-1})^2 / \sum_{t=2}^{n} \hat{u}_t^2 \approx 2(1 - \hat{\rho})$$

POLS 6481. Research Design and Quantitative Methods II
Lecture 14. Time Series Data II: Serial Correlation
Wooldridge, *Introductory Econometrics*, 11.1 + 12.1 + 12.2 + 12.3

4. The Durbin-Watson Test

DW test statistic is approximately $2(1-\hat{\rho})$, where $\hat{\rho}$ is coefficient in regression of $\hat{u}_t$ on $\hat{u}_{t\text{-}1}$.

Conduct test of whether DW = 2 by examining whether $\hat{\rho} = 0$

Null hypothesis: $H_0 : \rho = 0$; Alternative hypothesis: $H_1 : \rho > 0$

$$y_t = \beta_0 + \beta_1 x_{t1} + \cdots + \beta_k x_{tk} + u_t$$

$$u_t = \rho u_{t-1} + e_t$$

$$\hat{u}_t = \boxed{\alpha_0 + \alpha_1 x_{t1} + \cdots + \alpha_k x_{tk} +} \rho\, \hat{u}_{t-1} + error$$

$$DW = \sum_{t=2}^{n} (\hat{u}_t - \hat{u}_{t-1})^2 / \sum_{t=2}^{n} \hat{u}_t^2 \;\approx\; 2(1 - \hat{\rho})$$

5. Quasi-Differencing: Cochrane-Orcutt and Prais-Winsten Estimation

Under the assumption of AR(1) errors, you can transform the data to ensure OLS is BLUE; the process is to "quasi-difference" *y* and all *x*'s:

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

$$\rho y_{t-1} = \rho\beta_0 + \rho\beta_1 x_{t-1} + \rho u_{t-1}$$

$$\Rightarrow \boxed{y_t - \rho y_{t-1}} = \beta_0 \boxed{(1-\rho)} + \beta_1 \boxed{(x_t - \rho x_{t-1})} + \boxed{u_t - \rho u_{t-1}}$$

$$u_t = \rho u_{t-1} + e_t \Leftrightarrow \boxed{u_t - \rho u_{t-1}} = e_t$$

5. Quasi-Differencing: Cochrane-Orcutt and Prais-Winsten Estimation

Under the assumption of AR(1) errors, you can transform the data to ensure OLS is BLUE; the process is to "quasi-difference" y and all x's:

Typically ρ is unknown; replace it with $\hat{\rho}$ which is estimated from the data; use FGLS – but the downside of this is that FGLS requires strict exogeneity (TS.3) and AR(1) errors

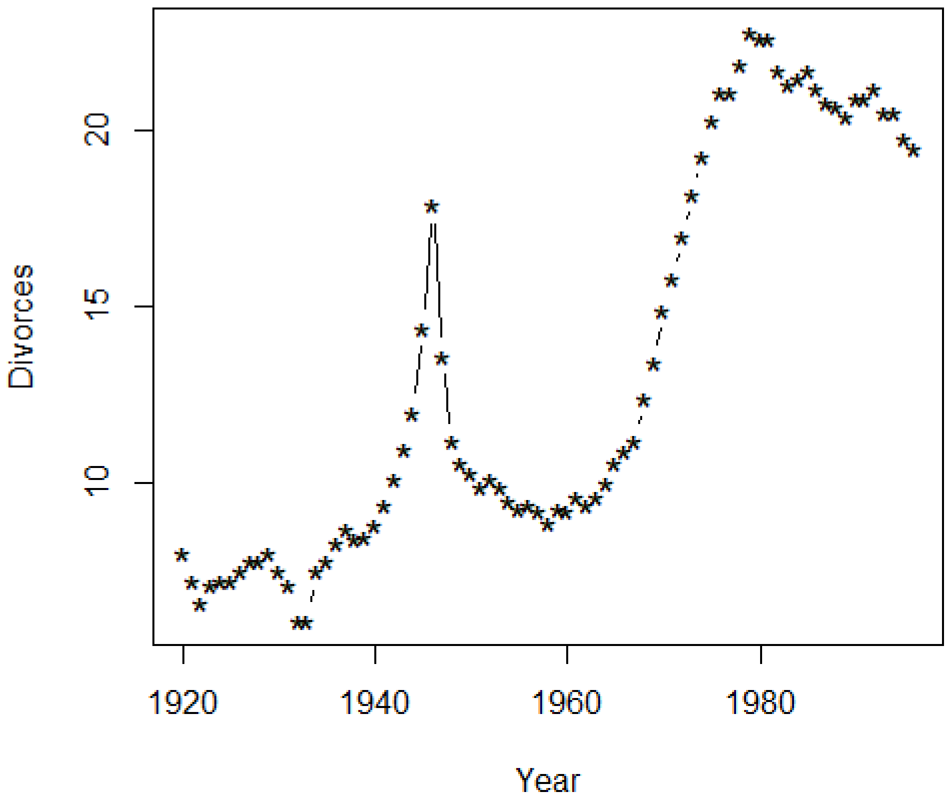
Cochrane-Orcutt omits the first observation;
Prais-Winsten adds a transformed first observation (it is more efficient in small samples)

(12.5 discusses Newey-West standard errors, which correct for serial correlation without transforming the data… "most useful when we have doubts about some of the explanatory variables being strictly exogenous"; the very last paragraph of 12.6 combines this with FGLS controls for heteroskedasticity too…)

6. Example of Quasi-Differencing: *Divorce Rates*

Data: divusa in faraway package
Script: Lecture 14 divorce.R

N = 77 yearly observations on *divorce* rate (1920 –1996).



6. Example of Quasi-Differencing: *Divorce Rates*

Data: divusa in faraway package
Script: Lecture 14 divorce.R

N = 77 yearly observations on *divorce* rate (1920 –1996). In this time interval, the US experienced the Second World War, which impacted the *marriage* and *birth* rates, how many people served in the *military*, and female participation in the workforce (*femlab*) – which also experienced a long-term trend.

Other variables include the unemployment rate (*unemployed*).

Julian J. Faraway, *Linear Models with R*, 1ed, Exercises 1.5, 4.5, 5.3, 6.2, …

- Estimate static model with five variables on RHS
- Check for serial autocorrelation (Durbin-Watson test, plot residuals over time, runs test)
- Carry out Prais-Winsten estimation; examine *rho*
- Check for autoregression / unit root in dependent variable (Dickey-Fuller test)
- Re-estimate model with first differences on LHS and RHS
- Re-check for serial autocorrelation in static model (Durbin-Watson test, run test)