Applied Econometrics for Health Policy

Time Series Analysis

Li-Lin Liang, Ph.D.

Institute of Public Health, National Yang Ming Chiao Tung University

Contents

1. The nature of time series data
2. Finite distributed lag (FDL) models
3. Finite sample properties of OLS under classical assumptions
4. Using dummy explanatory variables in time series
5. Modelling a linear time trend and seasonality in time series
6. Stationary and weakely dependent time series
7. Autoregressive and moving average processes
8. Random walks
9. Transformations on highly persistent time series

Contents

Basic Regression Analysis with Time Series Data

- **The nature of time series data**
  - Temporal ordering of observations; may not be arbitrarily reordered
  - Typical features: serial correlation/nonindependence of observations
  - How should we think about the randomness in time series data?
    - The outcome of economic variables (e.g. GNP, Dow Jones) is uncertain; they should therefore be modeled as random variables
    - Time series are sequences of r.v. (= stochastic processes)
    - Randomness does not come from sampling from a population
    - "Sample" = the one realized path of the time series out of the many possible paths the stochastic process could have taken

Basic Regression Analysis with Time Series Data

- **Example: US inflation and unemployment rates 1948-2003**

TABLE 10.1 Partial Listing of Data on U.S. Inflation and Unemployment Rates, 1948–2003

Year	Inflation	Unemployment
1948	8.1	3.8
1949	−1.2	5.9
1950	1.3	5.3
1951	7.9	3.3
.	.	.
.	.	.
.	.	.
1998	1.6	4.5
1999	2.2	4.2
2000	3.4	4.0
2001	2.8	4.7
2002	1.6	5.8
2003	2.3	6.0

Here, there are only two time series. There may be many more variables whose paths over time are observed simultaneously.

<u>Time series analysis focuses on modeling the dependency of a variable on its own past, and on the present and past values of other variables.</u>

Basic Regression Analysis with Time Series Data

- **Examples of time series regression models**

- **Static models**
  - In static time series models, the **current value** of one variable is modeled as the result of the **current values** of explanatory variables

- **Examples for static models**

$$inf_t = \beta_0 + \beta_1 unem_t + u_t$$

There is a contemporaneous relationship between unemployment and inflation (= Phillips curve).

$$mrdrte_t = \beta_0 + \beta_1 convrte_t + \beta_2 unem_t + \beta_3 yngmle_t + u_t$$

The current murder rate is determined by the current conviction rate, unemployment rate, and the fraction of young males in the population.

Basic Regression Analysis with Time Series Data

- **Finite distributed lag (FDL) models**
  - In finite distributed lag models, the explanatory variables are allowed to influence the dependent variable with a time lag
- **Example for a finite distributed lag model**
  - The fertility rate may depend on the tax value of a child, but for biological and behavioral reasons, the effect may have a lag

$$gfr_t = \alpha_0 + \delta_0 \boxed{pe_t} + \delta_1 \boxed{pe_{t-1}} + \delta_2 \boxed{pe_{t-2}} + u_t$$

Children born per 1,000 women in year t

Tax exemption in year t

Tax exemption in year t - 1

Tax exemption in year t - 2

Basic Regression Analysis with Time Series Data

- **FDL of order 2**

$$y_t = \alpha_0 + \delta_0 \boxed{z_t} + \delta_1 z_{t-1} + \delta_2 z_{t-2} + u_t$$

- **Suppose the increase in z is temporary**

$$\ldots, z_{t-2} = c, z_{t-1} = c, \boxed{z_t} = c + 1, z_{t+1} = c, z_{t+2} = c, \ldots.$$

- **We set the error term in each time period to zero. Then,**

$$\begin{aligned}
y_{t-1} &= \alpha_0 + \delta_0 c + \delta_1 c + \delta_2 c, \\
y_t &= \alpha_0 + \delta_0 \boxed{(c + 1)} + \delta_1 c + \delta_2 c, \\
y_{t+1} &= \alpha_0 + \delta_0 c + \delta_1 \boxed{(c + 1)} + \delta_2 c, \\
y_{t+2} &= \alpha_0 + \delta_0 c + \delta_1 c + \delta_2 \boxed{(c + 1)}, \\
y_{t+3} &= \alpha_0 + \delta_0 c + \delta_1 c + \delta_2 c,
\end{aligned}$$

Immediate change

$$y_t - y_{t-1} = \delta_0$$

Impact propensity/ multiplier

$$\delta_1 = y_{t+1} - y_{t-1}$$

$$\delta_2 = y_{t+2} - y_{t-1}$$

$$y_{t+3} = y_{t-1}.$$

Basic Regression Analysis with Time Series Data

- **FDL of order 2**

$$y_t = \alpha_0 + \delta_0 \boxed{z_t} + \delta_1 z_{t-1} + \delta_2 z_{t-2} + u_t$$

- **Suppose the increase in z is permenant**

$$z_s = c,\ s < t \text{ and } z_s = c + 1,\ s \geq t.$$

- **We set the error term in each time period to zero. Then,**

$$y_{t-1} = \alpha_0 + \delta_0 c + \delta_1 c + \delta_2 c,$$
$$y_t = \alpha_0 + \delta_0 \boxed{(c + 1)} + \delta_1 c + \delta_2 c,$$
$$y_{t+1} = \alpha_0 + \delta_0 (c + 1) + \delta_1 \boxed{(c + 1)} + \delta_2 c,$$
$$y_{t+2} = \alpha_0 + \delta_0 (c + 1) + \delta_1 (c + 1) + \delta_2 \boxed{(c + 1)},$$

After 1 period: y increased by $\delta_0 + \delta_1$

After 2 periods: y increased by $\boxed{\delta_0 + \delta_1 + \delta_2}$

(No further changes in y after two periods)

Long-run propensity/multiplier

Basic Regression Analysis with Time Series Data

- **Interpretation of the effects in finite distributed lag models**

$$y_t = \alpha_0 + \delta_0 z_t + \delta_1 z_{t-1} + \ldots + \delta_q z_{t-q} + u_t$$

- **Effect of a past shock on the current value of the dep. variable**

$$\boxed{\frac{\Delta y_t}{\Delta z_{t-s}} = \delta_s}$$

$$\boxed{\frac{\Delta y_t}{\Delta z_{t-q}} + \ldots + \frac{\Delta y_t}{\Delta z_t} = \delta_1 + \ldots + \delta_q}$$

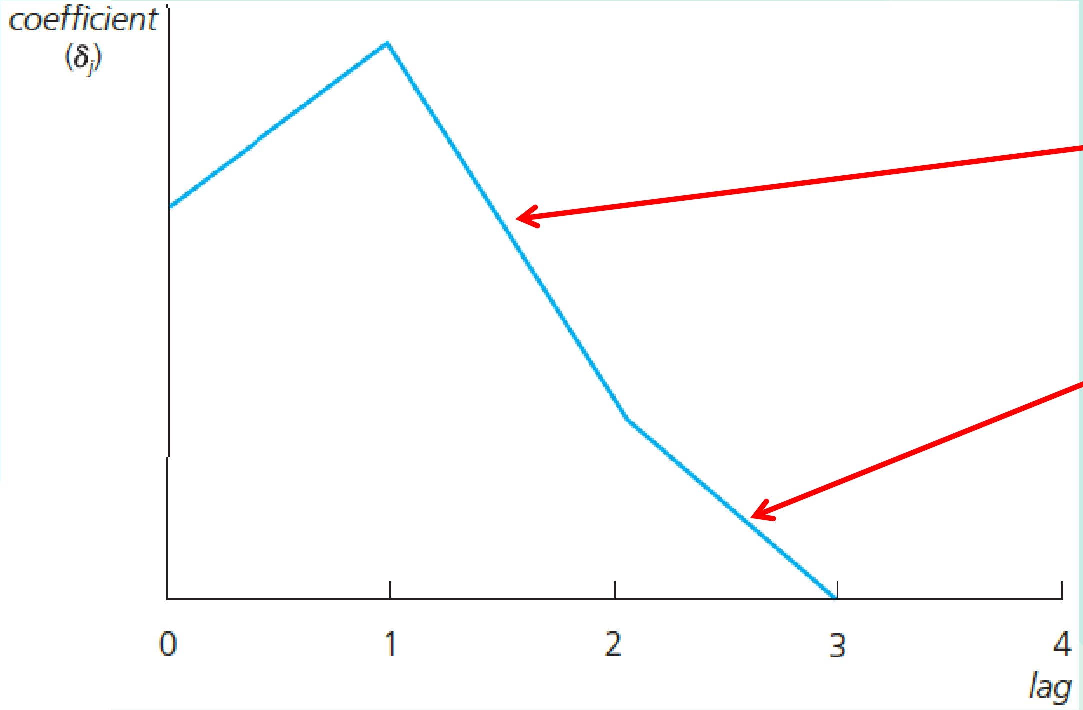
Effect of a transitory shock:
If there is a one time shock in a past period, the dep. variable will change temporarily by the amount indicated by the coefficient of the corresponding lag.

Effect of permanent shock:
If there is a permanent shock in a past period, i.e. the explanatory variable permanently increases by one unit, the effect on the dep. variable will be the cumulated effect of all relevant lags. This is a long-run effect on the dependent variable.

Basic Regression Analysis with Time Series Data

- **Graphical illustration of lagged effects**

A lag distribution with two nonzero lags. The maximum effect is at the first lag.



For example, the effect is biggest after a lag of one period. After that, the effect vanishes (if the initial shock was transitory).

The long run effect of a permanent shock is the cumulated effect of all relevant lagged effects. It does not vanish (if the initial shock is a permanent one).

Basic Regression Analysis with Time Series Data

- **Finite sample properties of OLS under classical assumptions**

- **Assumption TS.1 (Linear in parameters)**

$$y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \ldots + \beta_k x_{tk} + u_t \qquad t = 1,2,\ldots, n \text{ (time period)}$$

The time series involved obey a linear relationship. The stochastic processes y_t, x_{t1},…, x_{tk} are observed, the error process u_t is unobserved. The definition of the explanatory variables is general, e.g. they may be lags or functions of other explanatory variables.

- **Assumption TS.2 (No perfect collinearity)**

"In the sample (and therefore in the underlying time series process), no independent variable is constant nor a perfect linear combination of the others."

Basic Regression Analysis with Time Series Data

- **Notation X: the collection of all independent variables for all t**

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ \vdots & \vdots & & \vdots \\ \boxed{x_{t1} \quad x_{t2} \quad \cdots \quad x_{tk}} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}$$

This matrix collects all the information on the complete time paths of all the explanatory variables

The values of all the explanatory variables in period number t

- **Assumption TS.3 (Zero conditional mean)**

$$E(u_t|\mathrm{X}) = 0, t = 1, 2, \ldots, n$$

The mean value of the unobserved factors is uncorrelated to the values of the explanatory variables in all periods

Basic Regression Analysis with Time Series Data

- **Discussion of assumption TS.3**

Exogeneity: $E(u_t|\mathbf{x}_t) = 0$ — The mean of the error term is uncorrelated to the explanatory variables of the same period

Strict exogeneity: $E(u_t|\mathbf{X}) = 0$ — The mean of the error term is uncorrelated to the values of the explanatory variables of all periods

- **Strict exogeneity is stronger than contemporaneous exogeneity**
  - TS.3 rules out feedback from the dep. variable on future values of the explanatory variables; this is often questionable esp. if explanatory variables "adjust" to past changes in the dependent variable
  - If the error term is related to past values of the explanatory variables, one should include these values as contemporaneous regressors

Basic Regression Analysis with Time Series Data

- **Theorem 10.1 (Unbiasedness of OLS)**

$$TS.1 - TS.3 \quad \Rightarrow \quad E(\hat{\beta}_j) = \beta_j, \quad j = 0, 1, \ldots, k$$

- **Assumption TS.4 (Homoskedasticity)**

$$Var(u_t|\mathrm{X}) = Var(u_t) = \sigma^2, t = 1, 2, \ldots, n$$

The volatility of the errors must not be related to the explanatory variables in any of the periods

  - A sufficient condition is that the volatility of the error is independent of the explanatory variables and that it is constant over time
  - In the time series context, homoskedasticity may also be easily violated, e.g. if the volatility of the dep. variable depends on policy regime changes

Basic Regression Analysis with Time Series Data

- **Assumption TS.5 (No serial correlation)**

$$Corr(u_t, u_s|\mathbf{X}) = 0, \ t \neq s$$

Conditional on the explanatory variables, the un-observed factors must not be correlated over time

- **Discussion of assumption TS.5**
  - Why was such an assumption not made in the cross-sectional case?
  - The assumption may easily be violated if, conditional on knowing the values of the indep. variables, omitted factors are correlated over time
  - The assumption may also serve as substitute for the random sampling assumption if sampling a cross-section is not done completely randomly
  - In this case, given the values of the explanatory variables, errors have to be uncorrelated across cross-sectional units (e.g. states)

Basic Regression Analysis with Time Series Data

- **Theorem 10.2 (OLS sampling variances)**

Under assumptions TS.1 – TS.5:

The same formula as in the cross-sectional case

$$Var(\hat{\beta}_j|\mathbf{X}) = \frac{\sigma^2}{SST_j(1-R_j^2)}, \quad j = 1, \ldots, k$$

The conditioning on the values of the explanatory variables is not easy to understand. It effectively means that, in a finite sample, one ignores the sampling variability coming from the randomness of the regressors. This kind of sampling variability will normally not be large (because of the sums).

SST_j is the total sum of squares of x_{tj}

R^2_j is the R-squared from the regression of x_j on the other independent variables.

- **Theorem 10.3 (Unbiased estimation of the error variance)**

$$TS.1 - TS.5 \quad \Rightarrow \quad E(\hat{\sigma}^2) = \sigma^2$$

Basic Regression Analysis with Time Series Data

- **Theorem 10.4 (Gauss-Markov Theorem)**
  - Under assumptions TS.1 – TS.5, the OLS estimators have the minimal variance of all linear unbiased estimators of the regression coefficients
  - This holds conditional as well as unconditional on the regressors

- **Assumption TS.6 (Normality)**

This assumption implies TS.3 – TS.5

$u_t \sim \text{Normal}(0, \sigma^2)$ independently of $\mathbf{X}$

- **Theorem 10.5 (Normal sampling distributions)**
  - Under assumptions TS.1 – TS.6, the OLS estimators have the usual normal distribution (conditional on $\mathbf{X}$). The usual F- and t-tests are valid.

Basic Regression Analysis with Time Series Data

- **Example: Static Phillips curve**

$$\widehat{inf}_t = \underset{(1.72)}{1.42} + \underset{(.289)}{.468}\ unem_t$$

Contrary to theory, the estimated Phillips Curve does not suggest a tradeoff between inflation and unemployment

$$n = 49, R^2 = .053, \bar{R}^2 = .033$$

- **Discussion of CLM assumptions**

The error term contains factors such as monetary shocks, income/demand shocks, oil price shocks, supply shocks, or exchange rate shocks

<u>TS.1:</u> $inf_t = \beta_0 + \beta_1 unem_t + u_t$

<u>TS.2:</u> A linear relationship might be restrictive, but it should be a good approximation. Perfect collinearity is not a problem as long as unemployment varies over time.

Basic Regression Analysis with Time Series Data

- **Discussion of CLM assumptions (cont.)**

TS.3: $E(u_t|unem_1, \ldots, unem_n) = 0$ ← Easily violated

$unem_{t-1} \uparrow \rightarrow \ u_t \downarrow$ ← For example, past unemployment shocks may lead to future demand shocks which may dampen inflation

$u_{t-1} \uparrow \rightarrow \ unem_t \uparrow$ ← For example, an oil price shock means more inflation and may lead to future increases in unemployment

TS.4: $Var(u_t|unem_1, \ldots, unem_n) = \sigma^2$ ← Assumption is violated if monetary policy is more "nervous" in times of high unemployment

TS.5: $Corr(u_t, u_s|unem_1, \ldots, unem_n) = 0$ ← Assumption is violated if exchange rate influences persist over time (they cannot be explained by unemployment)

TS.6: $u_t \sim \text{Normal}(0, \sigma^2)$ ← Questionable

Basic Regression Analysis with Time Series Data

- **Using <u>dummy explanatory variables</u> in time series**

Children born per 1,000 women in year t

Tax exemption in year t

Dummy for World War II years (1941-45)

Dummy for availabity of contraceptive pill (1963-present)

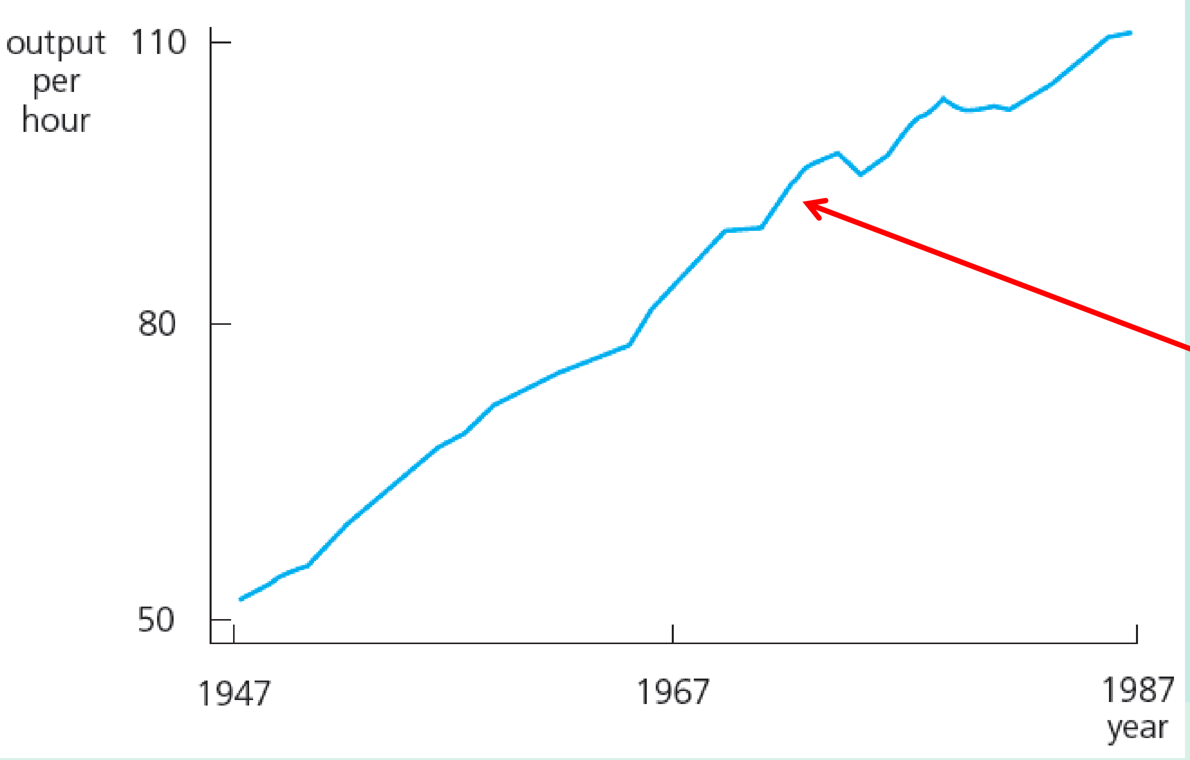
$$\widehat{gfrt} = \underset{(3.21)}{98.68} + \underset{(.030)}{.083}\, pe_t - \underset{(7.46)}{24.24}\, \boxed{ww2_t} - \underset{(4.08)}{31.59}\, \boxed{pill_t}$$

$$n = 72, R^2 = .473, \bar{R}^2 = .450$$

- **Interpretation**
  - During World War II, the fertility rate was temporarily lower
  - It has been permanently lower since the introduction of the pill in 1963

Basic Regression Analysis with Time Series Data

- **Time series with trends**



Example for a time series with a linear upward trend

Basic Regression Analysis with Time Series Data

- **Modelling <u>a linear time trend</u>**

$$y_t = \alpha_0 + \alpha_1 t + e_t \quad \Leftrightarrow \quad E(\Delta y_t) = E(y_t - y_{t-1}) = \alpha_1$$

$$\Delta y_t / \Delta t = \alpha_1$$

Abstracting from random deviations, the dependent variable increases by a constant amount per time unit

$$E(y_t) = \alpha_0 + \alpha_1 t$$

Alternatively, the expected value of the dependent variable is a linear function of time

- **Modelling <u>an exponential time trend</u>**

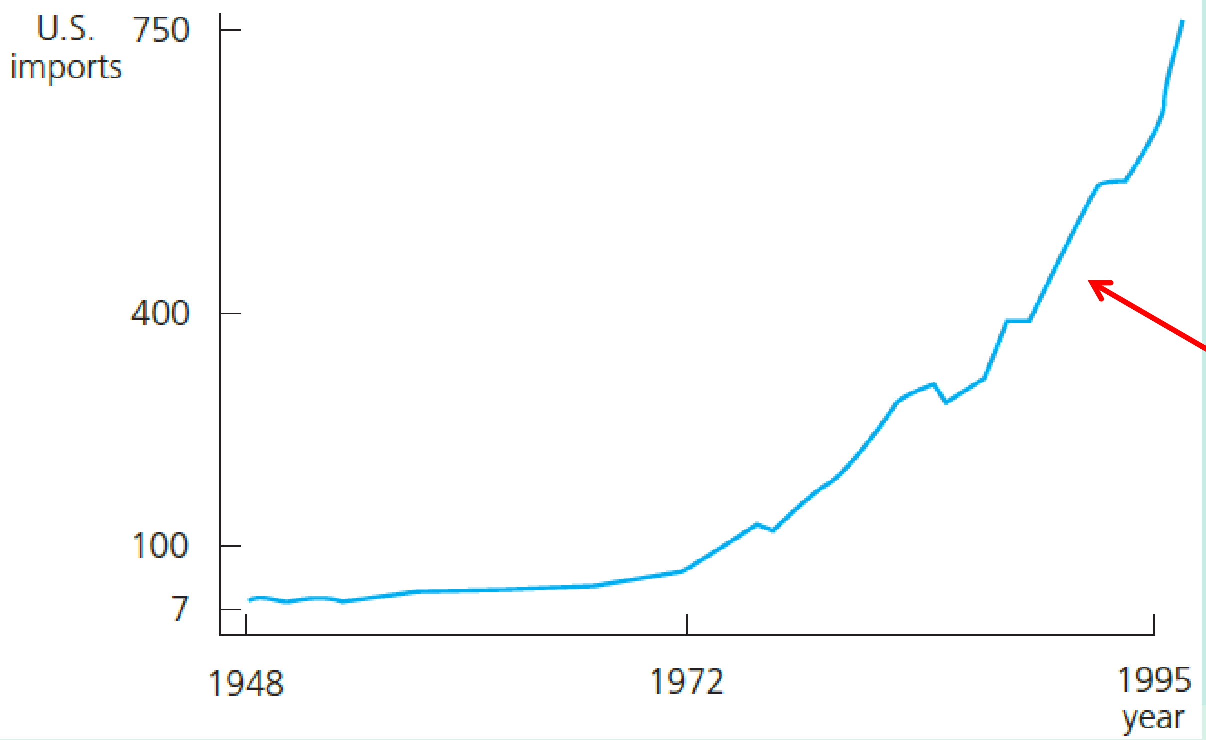
$$\log(y_t) = \alpha_0 + \alpha_1 t + e_t \quad \Leftrightarrow \quad E(\Delta \log(y_t)) = \alpha_1$$

$$(\Delta y_t / y_t) / \Delta t = \alpha_1$$

Abstracting from random deviations, the dependent variable increases by <u>a constant percentage</u> per time unit

Basic Regression Analysis with Time Series Data

- **Example for a time series with an exponential trend**



Abstracting from random deviations, the time series has a constant growth rate

Basic Regression Analysis with Time Series Data

- **Using trending variables in regression analysis**
  - If trending variables are regressed on each other, a spurious relationship may arise if the variables are driven by a common trend
  - In this case, it is important to include a trend in the regression
- **Example: Housing investment and prices**

Per capita housing investment

Housing price index

$$\widehat{\log}(invpc) = -\underset{(.043)}{.550} + \underset{(.382)}{1.241} \log(price)$$

It looks as if investment and prices are positively related

$$n = 42, R^2 = .208, \bar{R}^2 = .189$$

Basic Regression Analysis with Time Series Data

- **Example: Housing investment and prices (cont.)**

$$\widehat{\log}(invpc) = -.913 - .381 \log(price) + \boxed{.0098\, t}$$
$$(1.36) \quad (.679) \qquad\qquad (.0035)$$

$$n = 42, R^2 = .341, \bar{R}^2 = .307$$

There is no significant relationship between price and investment anymore

- **When should a trend be included?**
  - If the dependent variable displays an obvious trending behaviour
  - If both the dependent and some independent variables have trends
  - If only some of the independent variables have trends; their effect on the dep. var. may only be visible after a trend has been substracted

Basic Regression Analysis with Time Series Data

- **A detrending interpretation of regressions with a time trend**
  - It turns out that the OLS coefficients in a regression including a trend are the same as the coefficients in a regression without a trend but where all the variables have been detrended before the regression
  - This follows from the general interpretation of multiple regressions

When we regress y_t on x_{t1}, x_{t2}, and t, we obtain the fitted equation

$$\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 x_{t1} + \hat{\beta}_2 x_{t2} + \hat{\beta}_3 t.$$

(i) Regress each of y_t, x_{t1}, and x_{t2} on a constant and the time trend t and save the residuals, say, $\ddot{y}_t$, $\ddot{x}_{t1}$, $\ddot{x}_{t2}$, $t = 1, 2, \ldots, n$. For example,

$$\ddot{y}_t = y_t - \hat{\alpha}_0 - \hat{\alpha}_1 t.$$

(ii) Run the regression of

$$\ddot{y}_t \text{ on } \ddot{x}_{t1}, \ddot{x}_{t2}.$$

Basic Regression Analysis with Time Series Data

- **Computing R-squared when the dependent variable is trending**
  - Due to the trend, the variance of the dep. var. will be overstated
  - It is better to first detrend the dep. var. and then run the regression on all the indep. variables (plus a trend if they are trending as well)

$$\ddot{y}_t \text{ on } x_{t1}, x_{t2}, \text{ and } t.$$

  - The R-squared of this regression is a more adequate measure of fit

The R-squared from this regression is

$$1 - \frac{\text{SSR}}{\sum_{t=1}^{n} \ddot{y}_t^2},$$

where SSR is identical to the sum of squared residuals from $\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 x_{t1} + \hat{\beta}_2 x_{t2} + \hat{\beta}_3 t.$

Basic Regression Analysis with Time Series Data

- **Modelling seasonality in time series**

- **A simple method is to include a set of seasonal dummies:**

$$y_t = \beta_0 + \delta_1 \boxed{feb_t} + \delta_2 \boxed{mar_t} + \delta_3 \boxed{apr_t} + \ldots + \delta_{11} \boxed{dec_t}$$

$$+ \beta_1 x_{t1} + \beta_2 x_{t2} + \ldots + \beta_k x_{tk} + u_t$$

= 1 if obs. from december
= 0 otherwise

- **Similar remarks apply as in the case of deterministic time trends**
  - The regression coefficients on the explanatory variables can be seen as the result of first deseasonalizing the dep. and the explanat. variables
  - An R-squared that is based on first deseasonalizing the dep. var. may better reflect the explanatory power of the explanatory variables

Further Issues in Using OLS with Time Series Data

- **The assumptions used so far seem to be too restricitive**
  - Strict exogeneity, homoskedasticity, and no serial correlation are very demanding requirements, especially in the time series context
  - Statistical inference rests on the validity of the normality assumption
  - Much weaker assumptions are needed if the sample size is large
  - A key requirement for large sample analysis of time series is that the time series in question are stationary and weakly dependent
- **Stationary time series**
  - Loosely speaking, a time series is stationary if its stochastic properties and its temporal dependence structure do not change over time

Further Issues in Using OLS with Time Series Data

- **Stationary stochastic processes**

A stochastic process $\{x_t : t = 1, 2, \ldots\}$ is stationary, if for every collection of indices $1 \leq t_1 < t_2 < \cdots < t_m$ the joint distribution of $(x_{t_1}, x_{t_2}, \ldots, x_{t_m})$ is the same as that of $(x_{t_1+h}, x_{t_2+h}, \ldots, x_{t_m+h})$ for all integers $h \geq 1$.

- **Covariance stationary processes**

A stochastic process $\{x_t : t = 1, 2, \ldots\}$ is covariance stationary, if its expected value, its variance, and its covariances are constant over time:
1) $E(x_t) = \mu$, 2) $Var(x_t) = \sigma^2$, and 3) $Cov(x_t, x_{t+h}) = f(h)$

Further Issues in Using OLS with Time Series Data

- **Weakly dependent time series**

A stochastic process $\{x_t : t = 1, 2, \ldots\}$ is weakly dependent , if x_t is "almost independent" of x_{t+h} if h grows to infinity (for all t).

- **Discussion of the weak dependence property**
  - An implication of weak dependence is that the correlation between x_t and x_{t+h} must converge to zero if *h* grows to infinity
  - For the LLN and the CLT to hold, the individual observations must not be too strongly related to each other; in particular their relation must become weaker (and is fast enough) the further they are apart
  - Note that a series may be nonstationary but weakly dependent

Further Issues in Using OLS with Time Series Data

- **Examples for weakly dependent time series**

- **Moving average process of order one [MA(1)]**

The process is a short moving average of an i.i.d. series e_t

$$x_t = e_t + \alpha_1 e_{t-1},\ t = 1, 2, \ldots$$

The process is weakly dependent because observations that are more than one time period apart have nothing in common and are therefore uncorrelated.

- **Autoregressive process of order one [AR(1)]**

$$y_t = \rho_1 y_{t-1} + e_t,\ t = 1, 2, \ldots$$

The process carries over to a certain extent the value of the previous period (plus random shocks from an i.i.d. series e_t)

$$\Rightarrow \quad Corr(y_t, y_{t+h}) = \rho_1^h$$

If the stability condition $|\rho_1| < 1$ holds, the process is weakly dependent because serial correlation converges to zero as the distance between observations grows to infinity.

Further Issues in Using OLS with Time Series Data

- **Asymptotic properties of OLS**
- **Assumption TS.1' (Linear in parameters)**
  - Same as assumption TS.1 but now the dependent and independent variables are assumed to be stationary and weakly dependent
- **Assumption TS.2' (No perfect collinearity)**
  - Same as assumption TS.2
- **Assumption TS.3' (Zero conditional mean)**
  - Now the explanatory variables are assumed to be only contemporaneously exogenous rather than strictly exogenous, i.e.

$$E(u_t|\mathrm{x}_t) = 0$$

The explanatory variables of the same period are uninformative about the mean of the error term

Further Issues in Using OLS with Time Series Data

- **Theorem 11.1 (Consistency of OLS)**

$$TS.1'\text{–}TS.3' \quad \Rightarrow \quad plim\ \widehat{\beta}_j = \beta_j, \quad j = 0, 1, \ldots, k$$

Important note: For consistency it would even suffice to assume that the explanatory variables are merely contemporaneously *uncorrelated* with the error term.

- **Why is it important to relax the strict exogeneity assumption?**
  - Strict exogeneity is a serious restriction because it rules out all kinds of dynamic relationships between explanatory variables and the error term
  - In particular, it rules out feedback from the dep. var. on future values of the explanat. variables (which is very common in economic contexts)
  - Strict exogeneity precludes the use of lagged dep. var. as regressors

Further Issues in Using OLS with Time Series Data

- **Why do lagged dependent variables violate strict exogeneity?**

$$y_t = \beta_0 + \beta_1 y_{t-1} + u_t$$

This is the simplest possible regression model with a lagged dependent variable

<u>Contemporanous exogeneity:</u> $E(u_t | y_{t-1}) = 0$

<u>Strict exogeneity:</u> $E(u_t | y_0, y_1, \ldots, y_{n-1}) = 0$

Strict exogeneity would imply that the error term is uncorrelated with all y_t, t=1, … , n-1

This leads to a contradiction because:

$$Cov(y_t, u_t) = \beta_1 Cov(y_{t-1}, u_t) + Var(u_t) > 0$$

- **OLS estimation in the presence of lagged dependent variables**
  - Under contemporaneous exogeneity, OLS is consistent but biased

Further Issues in Using OLS with Time Series Data

- **Assumption TS.4�much (Homoskedasticity)**

$$Var(u_t|\mathbf{x}_t) = Var(u_t) = \sigma^2$$

The errors are contemporaneously homoskedastic

- **Assumption TS.5‵ (No serial correlation)**

$$Corr(u_t, u_s|\mathbf{x}_t, \mathbf{x}_s) = 0, \;\; t \neq s$$

Conditional on the explanatory variables in periods t and s, the errors are uncorrelated

- **Theorem 11.2 (Asymptotic normality of OLS)**
  - Under assumptions TS.1‵ – TS.5‵, the OLS estimators are asymptotically normally distributed. Further, the usual OLS standard errors, t-statistics, F-statistics, and LM statistics are asymptotically valid.

Further Issues in Using OLS with Time Series Data

- **Example: Efficient Markets Hypothesis (EMH)**

The EMH in a strict form states that information observable to the market prior to week t should not help to predict the return during week t. A simplification assumes in addition that only past returns are considered as relevant information to predict the return in week t.This implies that

$$E(return_t | return_{t-1}, return_{t-2}, \ldots) = E(return_t)$$

A simple way to test the EMH is to specify an AR(1) model. Under the EMH, assumptionTS.3' holds so that an OLS regression can be used to test whether this week's returns depend on last week's.

$$\underset{}{\widehat{return}_t} = \underset{(.081)}{.180} + \boxed{\underset{(.038)}{.059}}\, return_{t-1}$$

$$n = 689, R^2 = .0035, \bar{R}^2 = .0020$$

There is no evidence against the EMH. Including more lagged returns yields similar results.

Further Issues in Using OLS with Time Series Data

- **Using highly persistent time series in regression analysis**
  - Unfortunately many economic time series violate weak dependence because they are highly persistent (= strongly dependent)
  - In this case OLS methods are generally invalid (unless the CLM hold)
  - In some cases transformations to weak dependence are possible

Further Issues in Using OLS with Time Series Data

- **Random walks**

The process is called random walk because it wanders from the previous position y_{t-1} by an i.i.d. random amount e_t

$$\boxed{y_t = y_{t-1} + e_t,}\ t = 1, 2, \ldots$$

$$\Rightarrow\ y_t = (y_{t-2} + e_{t-1}) + e_t = \ldots = e_t + e_{t-1} + \ldots + e_1 + y_0$$

The value today is the accumulation of all past shocks plus an initial value. This is the reason why the random walk is highly persistent: The effect of a shock will be contained in the series forever.

$$E(y_t) = E(y_0)$$

$$Var(y_t) = \sigma_e^2 t$$

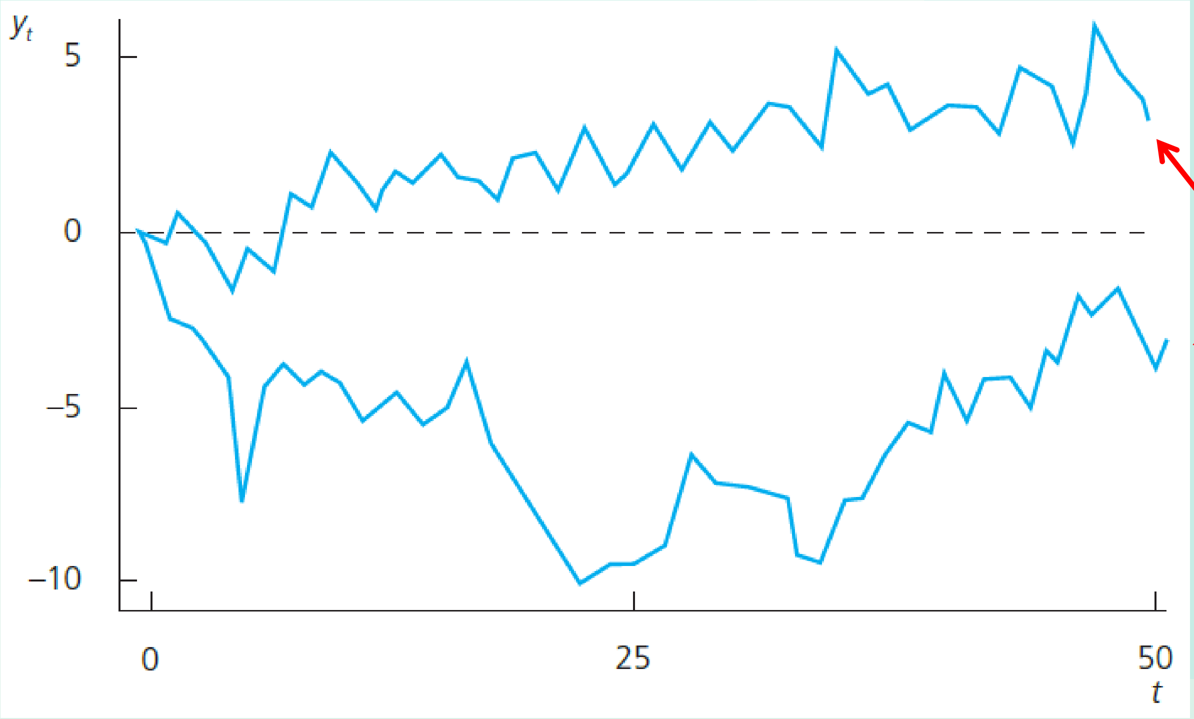
$$Corr(y_t, y_{t+h}) = \sqrt{t/(t+h)}$$

The random walk is <u>not covariance stationary</u> because its variance and its covariance depend on time.

It is also <u>not weakly dependent</u> because the correlation between observations vanishes very slowly and this depends on how large t is.

Further Issues in Using OLS with Time Series Data

- **Examples for random walk realizations**



The random walks wander around with no clear direction

Further Issues in Using OLS with Time Series Data

- **A random walk is a special case of a unit root process.**
- Unit root processes are defined as the random walk but e_t may be an arbitrary weakly dependent process.
- From an economic point of view it is important to know whether a time series is highly persistent.

1) In highly persistent time series, shocks or policy changes have permanent effects
2) In weakly dependent processes their effects are transitory.

Further Issues in Using OLS with Time Series Data

- **Random walks with drift**

In addition to the usual random walk mechanism, there is a deterministic increase/decrease (= drift) in each period

$$y_t = \boxed{\alpha_0} + y_{t-1} + e_t,\ t = 1, 2, \ldots$$

$$\Rightarrow \quad y_t = \boxed{\alpha_0 t} + e_t + e_{t-1} + \ldots + e_1 + y_0$$

This leads to a linear time trend around which the series follows its random walk behaviour. As there is no clear direction in which the random walk develops, it may also wander away from the trend.

$$E(y_t) = \boxed{\alpha_0 t} + E(y_0)$$

$$Var(y_t) = \sigma_e^2 t$$

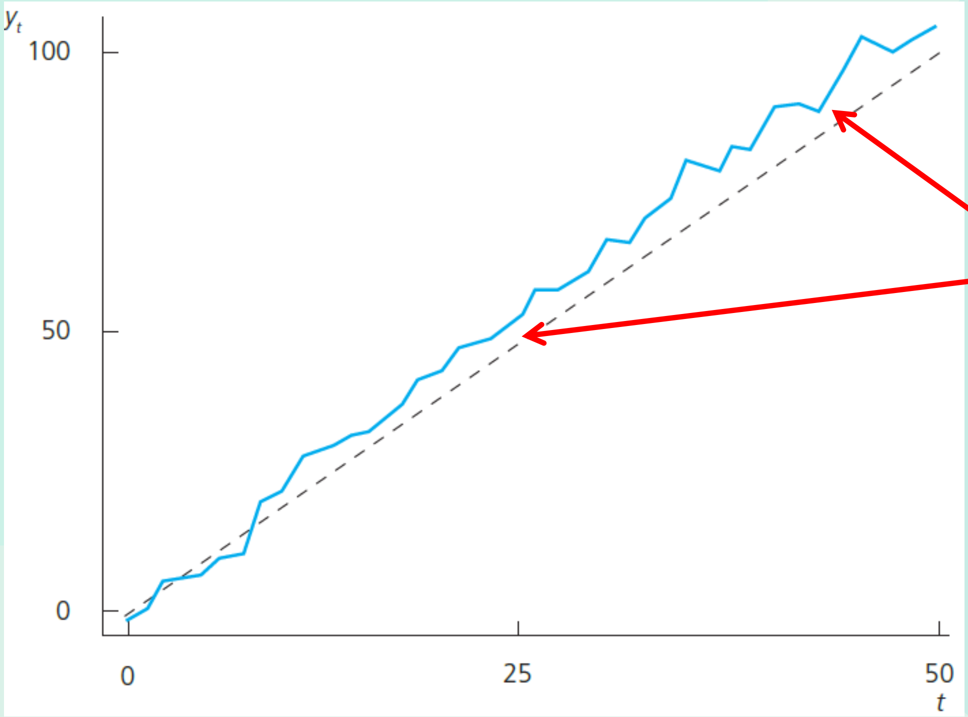
$$Corr(y_t, y_{t+h}) = \sqrt{t/(t+h)}$$

Otherwise, the random walk with drift has similar properties as the random walk without drift.

Random walks with drift are not covariance stationary and not weakly dependent.

Further Issues in Using OLS with Time Series Data

- **Sample path of a random walk with drift**



Note that the series does not regularly return to the trend line.

Random walks with drift may be good models for time series that have an obvious trend but are not weakly dependent.

Further Issues in Using OLS with Time Series Data

- **Transformations on highly persistent time series**

- **Order of integration**
  - Weakly dependent time series are integrated of order zero (= I(0))
  - If a time series has to be differenced one time in order to obtain a weakly dependent series, it is called integrated of order one (= I(1))

- **Examples for I(1) processes**

After differencing, the resulting series are weakly dependent (because e_t is weakly dependent).

$$y_t = y_{t-1} + e_t \;\Rightarrow\; \Delta y_t = y_t - y_{t-1} = e_t,\; t = 2, 3 \ldots$$

$$\Delta \log(y_t) \approx (y_t - y_{t-1})/y_{t-1}$$

- **Differencing is often a way to achieve weak dependence**

Further Issues in Using OLS with Time Series Data

- **Deciding whether a time series is I(1)**
  - There are statistical tests for testing whether a time series is I(1) (= **unit root tests**)
  - Alternatively, look at the sample first order autocorrelation:

$$\hat{\rho}_1 = \widehat{Corr}(y_t, y_{t-1})$$

Measures how strongly adjacent times series observations are related to each other.

  - If the sample first order autocorrelation is close to one, this suggests that the time series may be highly persistent (= contains a unit root)
  - Alternatively, the series may have a deterministic trend
  - Both unit root and trend may be eliminated by differencing

Further Issues in Using OLS with Time Series Data

- **Example: Fertility equation**

$$gfr_t = \alpha_0 + \delta_0 pe_t + \delta_1 pe_{t-1} + \delta_2 pe_{t-2} + u_t$$

This equation could be estimated by OLS if the CLM assumptions hold. These may be questionable, so that one would have to resort to large sample analysis. For large sample analysis, the fertility series and the series of the personal tax exemption have to be stationary and weakly dependent. This is questionable because the two series are highly persistent:

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

It is therefore better to estimate the equation in first differences. This makes sense because if the equation holds in levels, it also has to hold in first differences:

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

$$n = 69, R^2 = .233, \bar{R}^2 = .197$$

Estimate of δ_2

Further Issues in Using OLS with Time Series Data

- **Example: Wages and productivity**

Include trend because both series display clear trends.

$$\widehat{\log}(hrwage) = \underset{(.37)}{-5.33} + \underset{(.09)}{1.64} \log(outphr) - \underset{(.002)}{.018}\, t$$

$$n = 41, R^2 = .971, \bar{R}^2 = .970$$

The elasticity of hourly wage with respect to output per hour (=productivity) seems implausibly large.

It turns out that even after detrending, both series display sample autocorrelations close to one so that estimating the equation in first differences seems more adequate:

$$\Delta\widehat{\log}(hrwage) = \underset{(.0042)}{-.0036} + \underset{(.173)}{.809}\, \Delta \log(outphr)$$

$$n = 40, R^2 = .364, \bar{R}^2 = .348$$

This estimate of the elasticity of hourly wage with respect to productivity makes much more sense.

Serial Correlation in Time Series Regressions

- **Properties of OLS with serially correlated errors**
  - OLS still unbiased and consistent if errors are serially correlated
  - Correctness of R-squared also does not depend on serial correlation
  - OLS standard errors and tests will be invalid if there is serial correlation
  - OLS will not be efficient anymore if there is serial correlation

Serial Correlation in Time Series Regressions

- **Testing for serial correlation**
- **Testing for AR(1) serial correlation <u>with strictly exog. regressors</u>**

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

AR(1) model for serial correlation (with an i.i.d. series e_t)

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

Replace true unobserved errors by estimated residuals

Test $H_0 : \rho = 0$ in $\hat{u}_t = \rho \hat{u}_{t-1} + error$

- **Example: Static Phillips curve (see above)**

$$\hat{\rho} = .573, t = 4.93, p\text{-}value = .000$$

Reject null hypothesis of no serial correlation

Serial Correlation in Time Series Regressions

- **The Durbin-Watson test** **under classical assumptions**
  - Under assumptions TS.1 – TS.6, the Durbin-Watson test is an exact test (whereas the previous t-test is only valid asymptotically).

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

$H_0 : \rho = 0$ vs. $H_1 : \rho > 0$

Unfortunately, the Durbin-Watson test works with a lower and an upper bound for the critical value. In the area between the bounds the test result is inconclusive.

Reject if $DW < d_L$, fail to reject if $DW > d_U$

- **Example: Static Phillips curve (see above)**

$DW = .80 < d_L = 1.32$ Reject null hypothesis of no serial correlation

Serial Correlation in Time Series Regressions

- **Testing for AR(1) serial correlation with general regressors**
  - The t-test for autocorrelation can be easily generalized to allow for the possibility that the explanatory variables are not strictly exogenous:

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

The test now allows for the possibility that the strict exogeneity assumption is violated.

Test for $H_0 : \rho = 0$

  - The test may be carried out in a heteroskedasticity robust way

- **General Breusch-Godfrey test for AR(q) serial correlation**

$$\widehat{u}_t = \alpha_0 + \alpha_1 x_{t1} + \ldots + \alpha_k x_{tk} + \boxed{\rho_1}\,\widehat{u}_{t-1} + \ldots + \boxed{\rho_q}\,\widehat{u}_{t-q} + \ldots$$

Test $H_0 : \rho_1 = \cdots = \rho_q = 0$

Serial Correlation in Time Series Regressions

- **Correcting for serial correlation with strictly exog. regressors**
  - Under the assumption of AR(1) errors, one can transform the model so that it satisfies all GM-assumptions. For this model, OLS is BLUE.

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

Simple case of regression with only one explanatory variable. The general case works analogusly.

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

Lag and multiply by ρ

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

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

The transformed error satisfies the GM-assumptions.

  - Problem: The AR(1)-coefficient is not known and has to be estimated

Serial Correlation in Time Series Regressions

- **Correcting for serial correlation (cont.)**
  - Replacing the unknown ρ by $\hat{\rho}$ leads to a FGLS-estimator
  - There are two variants:
    - Cochrane-Orcutt estimation omits the first observation
    - Prais-Winsten estimation adds a transformed first observation
  - In smaller samples, Prais-Winsten estimation should be more efficient