Basic Regression Analysis: Time Series Data

CIVL 7012/8012

Nature of time series data

- Temporal ordering of observations; may not be arbitrarily reordered
- Time series data has a separate observation for each time period –
  - e.g. annual traffic volume on a corridor,
  - census observations over multiple decades
  - Population of a city over multiple years

Some notes on time series (1)

- One observation repeated over time
- Past can affect future, not vice versa
- Randomness?
  - Not drawn from population like cross-sectional
  - Not drawn randomly (outcome is not foreknown)
  - So can be viewed as random variable
- Formally a sequence of random variables are defined as “stochastic” or “time series process”

Some notes on time series (2)

- When we collect time series data
  - We collect possible outcomes of stochastic data
  - (We can’t go back in time and repeat the process)
- Population
  - All the elements of the stochastic process
- Sample
  - Only some periods of data is used ad avaibale

Data features

- Time periods to consider
  - Daily, Weekly, Monthly, Quarterly, Annually, Quinquennially (every five years), Decennially (every 10 years)
- Since not a (purely) random sample, different problems to consider
  - Trends and seasonality will be important

Data issues

- Stationary issue
  - Loosely speaking a time series is stationary if its mean and standard deviation does not vary systematically over time
- 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

Example data

- 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.

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.

Example of time series regression model

- **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**

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

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

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

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

Finite distributed lag models

- **Finite distributed lag models**
  - In finite distributed lag models, the explanatory variables are allowed to influence the dependent variable with a time lag

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

- **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

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

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

Interpretation of coefficients: finite distributed lag models

- Interpretation of the effects in finite distributed lag models

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

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

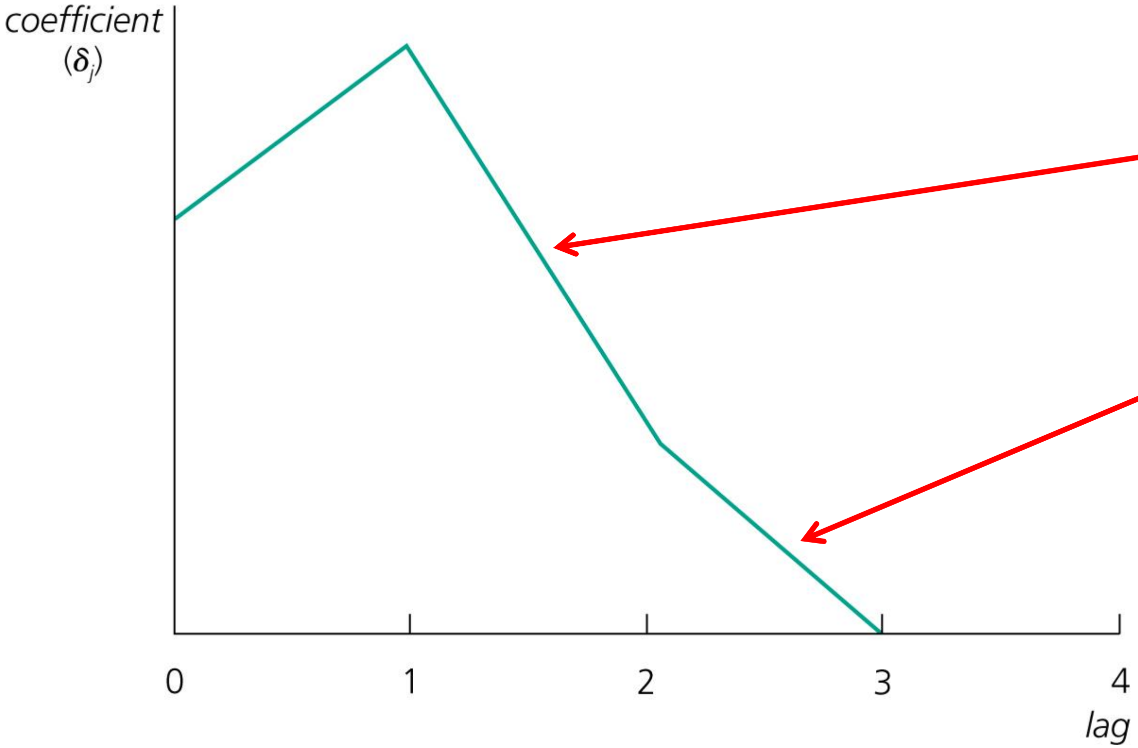
$$\frac{\partial y_t}{\partial z_{t-s}} = \delta_s$$

$$\frac{\partial y_t}{\partial z_{t-q}} + \cdots + \frac{\partial y_t}{\partial z_t} = \delta_1 + \cdots + \delta_q$$

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

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.

Lagged effects



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).

Example-1

- **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 (t-stat: 1.62)

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

- **Discussion of CLM assumptions**

(CLM: Classical Linear Model)

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

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

<u>Assumption.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.

Example-2

- Example: Effects of inflation and deficits on interest rates

Interest rate on 3-months T-bill

Government deficit as percentage of GDP

$$\widehat{i3}_t = \underset{(0.43)}{1.73} + \underset{(.082)}{.606}\, inf_t + \underset{(.118)}{.513}\, def_t$$

$$n = 56, R^2 = .602, \bar{R}^2 = .587$$

The error term represents other factors that determine interest rates in general, e.g. business cycle effects

Assumption.1: $i3_t = \beta_0 + \beta_1 inf_t + \beta_2 def_t + u_t$

Assumption.2: A linear relationship might be restrictive, but it should be a good approximation. Perfect collinearity will seldomly be a problem in practice.

Example-3

- **Using dummy explanatory variables in time series**

Children born per 1,000 women in year t → $\widehat{gfr}_t$

Tax exemption in year t → pe_t

Dummy for World War II years (1941-45) → $ww2_t$

Dummy for availabity of con-traceptive pill (1963-present) → $pill_t$

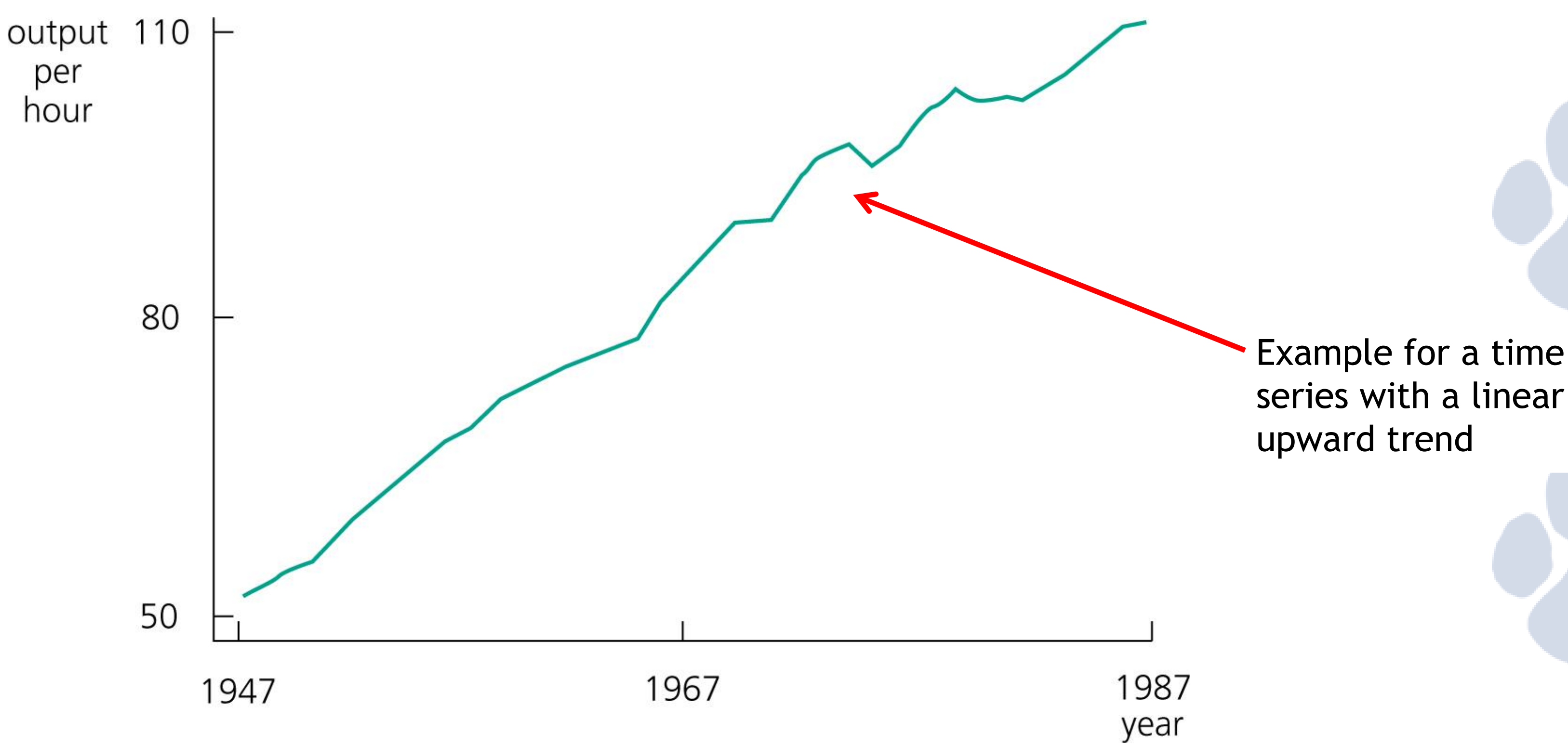
$$\underset{(3.68)}{\widehat{gfr}_t = 98.68} + \underset{(.030)}{.083}\, pe_t - \underset{(7.46)}{24.24}\, ww2_t - \underset{(4.08)}{31.59}\, pill_t$$

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

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

Time series with trends

- Plot of labor productivity (output per hour of work) in the United States for the years 1947 through 1987



Modeling a linear time trend

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

$$\partial y_t / \partial t = \alpha_1$$

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

Modeling an exponential time trend

- Example

$$\log(y_t) = \beta_0 + \beta_1 t + e_t,\ t = 1, 2, \ldots.$$

$$y_t = \exp(\beta_0 + \beta_1 t + e_t).$$

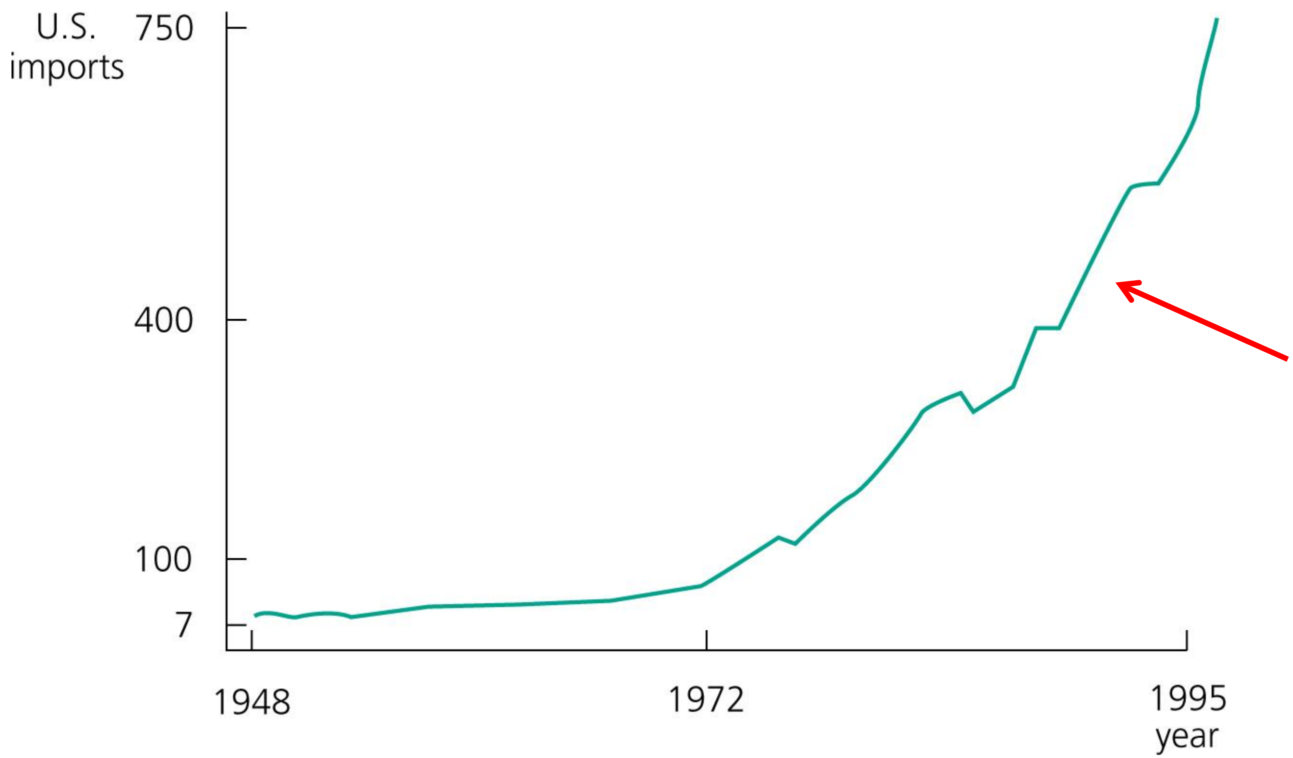
- Representation

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

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

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

Example of a time series with exponential trend



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

Interpretation in exponential trend

- How to interpret b_1 in $\log(y_t) = \beta_0 + \beta_1 t + e_t, t = 1, 2, \ldots.$
- For small changes in y (test at home).

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

- RHS is **growth rate** in *y* from period *t*-1 to *t*.
- To turn the growth rate into a percentage, we simply multiply by 100
- For example, if *t* denotes year and *b*_1 is .027, then y_t grows about 2.7% per year on average

More complex trends

- Quadratic time trend

$$y_t = \alpha_0 + \alpha_1 t + \alpha_2 t^2 + e_t.$$

- If *a*1 and *a*2 are +ve, then the slope of the trend is increasing, as is easily seen by computing the approximate slope (holding e_t fixed)

$$\frac{\Delta yt}{\Delta t} \approx \alpha_1 + 2\alpha_2 t.$$

- Trend will be different if *a*1 and *a*2 are of opposite sign

Using trending variables

- **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 → $invpc$; Housing price index → $price$

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

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

It looks as if investment and prices are positively related (Counter-intuitive showing spurious relationship)

Using trending variables (2)

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

The time trend is statistically significant, and its coefficient implies an approximate 1% increase in *invpc* per year, on average

$$\widehat{\log(invpc)} = -.913 - .381\, log(price) + .0098\, t$$

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

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
  - <u>If only some of the independent variables have trends</u>; their effect on the dep. var. may only be visible after a trend has been subtracted

So far…

- So far we have learned
  - Simple time series
  - Finite distributed lag model
  - Linear trending model
  - Exponential trending model
  - Polynomial trending model
  - Detrnding models can be possible for all the above
- Solution approach: we can still use OLS (all the time series assumption so far is BLUE)

Modeling seasonality in time series

- Modelling seasonality in time series
- Data might have trends (season, week, day, hour, etc.)
- A simple method is to include a set of seasonal dummies:

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

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

=1 if obs. from december
=0 otherwise

Further exploration (not covered in class)

- Problem of stationarity
- Auto regressive models
- Moving average models
- Autoregressive and moving average models
- Other advanced time series models