Time Series Models

Nir Kaldero
University of California, Berkeley

General Assembly – Data Science 6 – San Francisco

Time Series Data

- Intuition – why time series?
  - Big Data
  - From Static to Dynamic analysis
  - Dependency (x today depends on the past/future)
  - Capturing Trends over Time
  - Seasonality (cleaning, controlling)
  - Smoothing (policies, more realistic approach)
- Another main advantage:
  - A way to deal with endogeneity ($Cov(x_t, e_t) \neq 0$)
    - Lag Effect – FDL – Finite Distributed lag across time
    - Instrumental Variables (IV)

Time Series Data

- What have we learned so far?
  - Leveraging *static* data into predictive modeling power
  - Our variables are *random* –
    - they are part of a stochastic process ('realization')
      - Stochastic process: a collection of random values often used to represent the evolution of some random variable, or system, over time
  - We are able to make prediction and forecast
- What about time across observations?
  - Time seems to play a major role – why?
  - Observations are dependent across time
    - Intuitions:
      - The chronological time ('The Order') is important (easy)
      - The past might affect the future!
      - But what about Expectations for the Future? (How to accumulate?)

Static Model - The Main Problem

- A static mode: $y_{i,t} = \beta_0 + \beta_1 x_{i,t} + \beta_i x_{i,t} + e_{i,t}$
  $t = t..T$
  - This model takes into consideration time across observations – in a unique way (pay attention!)

- Problems:
  - All variables affect y ('the outcome') **at the same point in time**
  - There is no dynamic
  - There is no dependency across time!
- Results:
  - Bias (in some cases) in the estimators
  - Overshooting
  - Inaccuracy – less predictive power

Static Model - Thoughts

- We already saw that the static model is problematic
  - It is inconsistent with our prior intuition
    - Time matters
    - Sometimes – the model is "overshooting", over forecasting
      - Nothing to smooth/moderate effects across time (e.g.: interest rate)
    - There is no dependency
    - There is no dynamic
- However, the [Real] main problem is – endogeneity
  - When we want to control for time:

    → from Static to Dynamic settings
  - There are some implications especially in the assumptions

The Exogenous Problem in Time Series

- Let's think about the exogenous problem for a second…
  - In the static model we argued that:

$$Cov(x_i, e_i) = 0$$

Which was "easy to digest"…

  - However, in a dynamic setting (time series), the exogenous assumption argues that:

$$Cov(x_{i,t}, e_{i,t}) = 0 \; \forall \; i \in N \mid x \in X \mid t \in T$$

  - This is hard to digest and believe…
    - Do you believe that **for-all** observations across time there is not correlation between the **observed** information and the **unobserved ones?**
    - We already thought about that – of course not!
- We need to think how to overcome this problem!

Some [big] Open Questions..

- Do we believe in the assumptions?
  - And what can we do in order to relax these assumptions?
- What about the past?
- What about the future – expectations?
- What about correlation across $y_{i,t}$ over time?
- What about correlations across $e_{i,t}$ over time?
- What about over-shooting/forecasting?
- What is a good identification strategy?

- Let's try to solve/answer each of these questions..

Get ready, fasten your seat belt!!! :P

Relaxing the Endogeneity Assumption

- Having a Time Series data-set can be helpful
  - …if you will use it correctly!!! ☺
- Main advantage: We have lots of observations to play with!
- Most of the time – it is hard to believe that: $Corr(x_i, e_i) = 0$
  - even harder to believe that $Corr(x_{i,t}, e_{i,t}) = 0 \ \forall \ t$
- How can we relax this problematic assumption?
  - We can use the past data that we have
    - We are already using it – but under a 'static' settings
  - We can create lag effects – for the variables → Dynamic Model
    - In this way we "take out" important information from the error term (e) and incorporate it in the model
    - How come?

Dynamic Model

- Lag Effect - Moving from a Static to a Dynamic Model
  - If in the Static Model:

$$y_{i,t} = \beta_0 + \beta_1 x_{i,t} + \beta_i x_{i,t} + e_{i,t} (\exists\ x_{i,t-1}, x_{i,t-2} . .)$$

And $x_{i,t}$ is correlated with $x_{i,t-1}$ - we cannot trust the results

  - They are biased!

  - But if we "take out" $x_{i,t-1}$ and/or $x_{i,t-2}$ from the error-term?

$$y_{i,t} = \beta_0 + \beta_1 x_{i,t} + \beta_2 x_{i,t-1} + \beta_3 x_{i,t-2} + \beta_i x_{i,t-s} + e_{i,t}$$

$$y_{i,14} = \beta_0 + \beta_1 x_{i,14} + \beta_2 x_{i,13} + \beta_3 x_{i,12} + \beta_i x_{i,t-s} + e_{i,14}$$

  - We overcome the problem of endogeneity across time
    - But since $x_{i,t}$, $x_{i,t-1}, x_{i,t-2}$ are highly correlated – Multicollinearity
      - Try to check their **joint** significance (and not exclusive!)

- Creating Lag Variables ("Lag Effect") - in Lab!

What about correlation across the errors?

- Time series models raise another problem:
  - Correlation across the information in the error term (we Do Not observe) over time
    - Or in other words – correlation between the errors-term

$$Corr(e_{i,t}, e_{i,t-1}) \neq 0$$
$$Corr(e_{i,t-1}, e_{i,t-s}) \neq 0$$

- In fact, you already know how to deal with this issue
  - Arima model (Alessandro ☺)
  - Use it! – it is a useful strategy and tool

What about lagged effect of the outcome?

- In many series (especially in Macro) there is correlation between y_{today} and y_{past}
  - It is very useful to include in the model the lags of y (outcome)
  - Let's look at a real example:
- Forecasting the Interest rate
  - The Federal Reserve Committee forecasts and determines the interest rate – each month
  - It is crucial to forecast the interest rate:
    - why?
      - Investments decisions (NPV)
      - Calculating some Indexes
      - Affect unemployment, inflation, etc
      - Bond, stocks, etc
    - Who?
      - Investors – can be you!
      - Companies (calculating net present value of their project and assets)
      - etc

Time Series – Forecasting Interest Rate

- I chose this example because:
  - Implementing lag effects of the explanatory variables
  - Incorporating lag effect of the dependent variables
  - How to avoid over-shooting/forecasting – Smoothing
    - Before we will go to the model
- What is smoothing?
  - Many times our models are over forecasting
    - They give us estimators that are larger than the real effect
  - In some real predictive models we want to avoid "shocks" for the outcome
    - What is the intuition behind it?
      - The central bank wants to smooth the interest rate path
      - We want to avoid radical changes (investment, etc) in the market
      - Could you think of a strategy about how to incorporate this idea in the model?

Smoothing

- You already know the answer!
  - We want to link past and present
  - You can include in the model – the lag (past) effect of the interest rate (or the outcome you care about)
    - In this way, we "smooth", we consider and link the past rate in the model
    - Our estimators will be more "modest" (moderated)
  - Smoothing is a useful strategy in real world interfaces
  - Use it – it makes more sense
  - It is also a great method to deal with overshooting

- Now let's go back to forecasting the interest rate!

Forecasting the Interest Rate

- The well-known (simple) model is:

$$int_t = \beta_0 + \beta_1(infl_t - infl^T) + \beta_2 outputgap_t + \beta_3 int_{t-1} + \beta_4 int_{t-2} + e_t$$

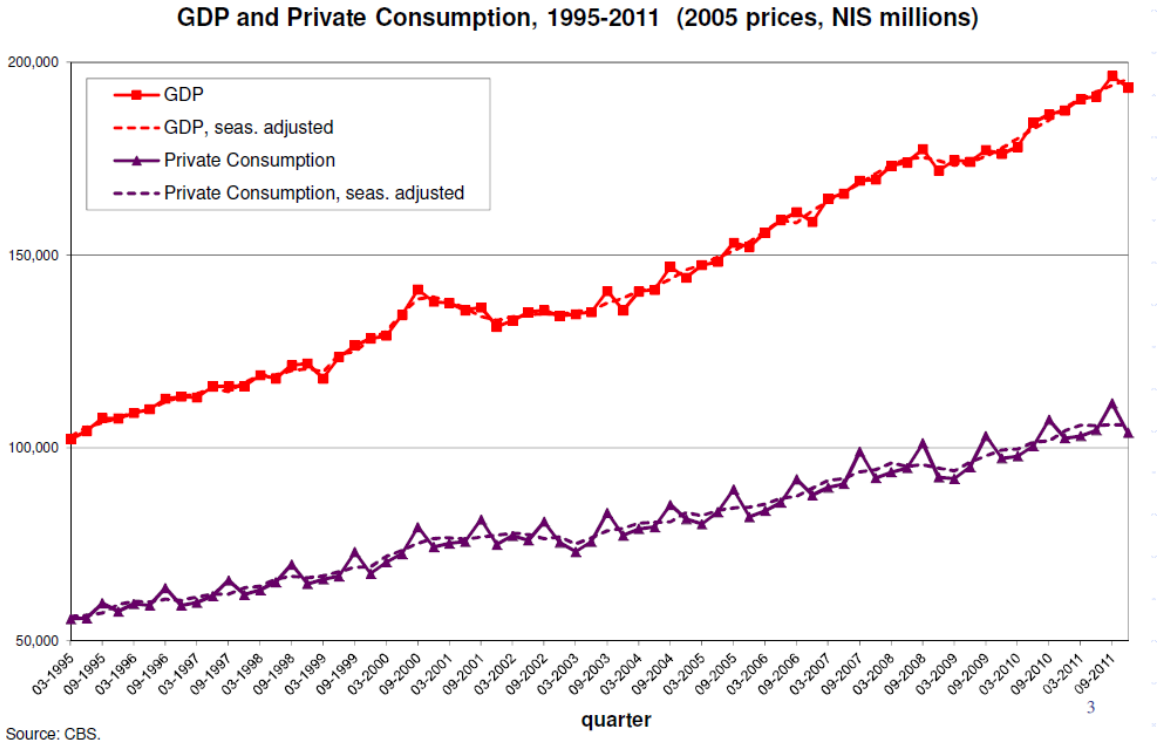
- We can re-write the model since $infl^t = 2\%\ (OECD)$ is constant

$$int_t = (\beta_0 - \beta_1 infl^t) + \beta_1(infl_t) + \beta_2 outputgap_t + \beta_3 int_{t-1} + \beta_4 int_{t-2} + e_t$$

- Any problems? (sign/significant) – Lab
  - You will have to come up with a strategy to deal with it…
  - What about output_gap?
  - Try to incorporate more lags (also for other variables)
    - Any difference?
- Now you know how to forecast the interest rate!
  - Hooray ! (sure..)

Let's move on – Trend over Time

- Another useful advantage of having a data set across time is that we have the luxury to observe/capture trends
  - What are the trends?



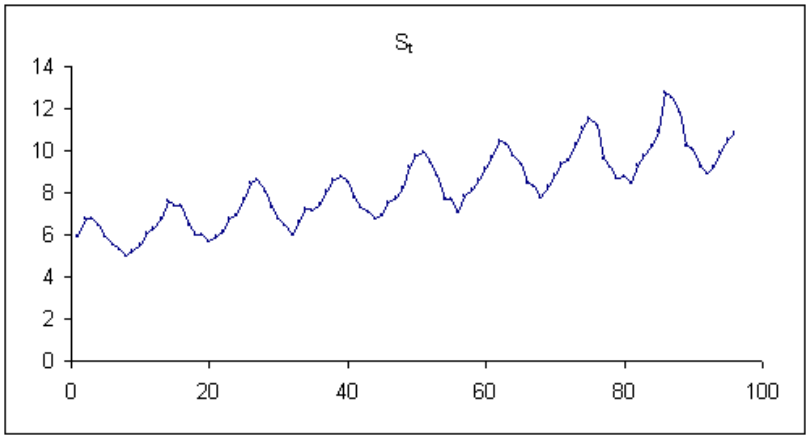
Trend over Time

- If we see such a pattern in the data – we have a problem!
  - Something in the background effects this curve (the trend)
  - If you will not control for it – your results will be biased
  - Larger sample can help

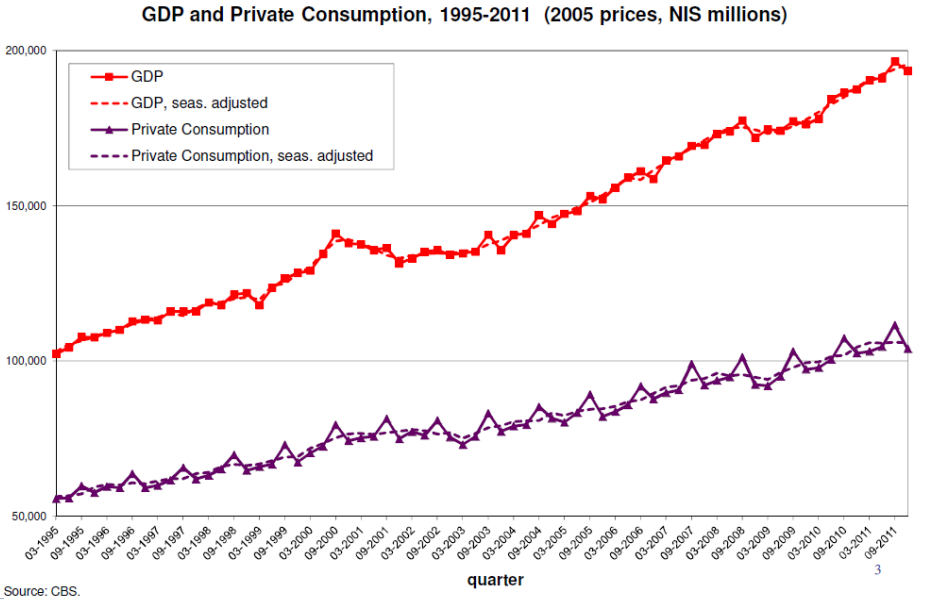
- How to fix this problem?
  - If you know what might create the trend – control for it!
    - Most of the time it will be hard to say..
  - If you do not know:
    - You can calculate the trend – and normalize it (do not do it..)
    - You can control for the time horizon over the tend/sample
      - Create dummy variables for each point in time (year, quarter, day..) and control for them in your analysis
      - It should help
      - The dummy variables (time variables) will 'normalize' the trend over your sample

Seasonality

- Many people mix between Trend and Seasonality
  - You will not! ☺
- This is seasonality:



- We spot a **cycle** over time



Source: CBS.

Seasonality

- Think about seasonality as holiday's shopping sales
- Again, If you know what causes the cycle – you are safe
  - Otherwise, your predictions are biased
    - Larger sample can help

- How to overcome this issue? (very similar to trends)
  - Assume you do not know what causes the curve to be cycled
    - First option:
      - Calculate the direct effect of the season at each point in time
        (you really do not want to do it now..)
    - Second option (recommended):
      - Create dummy variables for each pick in time, for each season
      - Control for these dummy variables

Dealing with [big] Open Questions..

- So far we went over the big open questions
  - Yes I know it was fast!
  - I just wanted to give you a taste from the theory
  - You can always go back home – and read more
- We are only left with one big open question:
  - What is a good identification strategy?
- Let's try to conquer this one as well

Identification Strategy – Instrument Variable

- Identification – It's all that matters!
  - If you come up with a good identification strategy on how to elicit the direct effect and provide some good arguments – you are a rock-star!
- Why do we need to deeply think about Identification Strategy (IS)? (besides your desire to be a rock-star..)
  - A good IS will lead to more accurate results
    - Not speaking about un-biased estimators
  - Your model will do better than others!
  - You have more predictive power! – use it (worth lot of money)
  - People will tend to consider/believe in your results

Instrumental Variable

- IV is a common method to overcome the main problem with endogeneity
  - Recall: $corr(x_{i,t}, e_{i,t}) \neq 0$
    - We already suggested and walk-through several methods
  - The least obvious and practical one – is to use an Instrumental Variable (z_i)
- What is a good instrumental variable?
  - A good instrumental variable MUST satisfy these conditions:
    1. $Corr(x_i, z_i) \gg 0$ (you can check it!)
    2. $Corr(z_i, e_i) = 0$
       - The problem with this condition – you cannot check it
       - You can only argue (with some theory) or assume

Instrumental Variable

- Once you find a good IV
  - You can just "replace" the endogenous variable (x_i) with z_i
    - In this way – you overcome the issue with endogeneity in the model (biased estimators)
- We are not fully replacing x_i with z_i
  - We are calculating the model in two steps
    - From here the name - Two Steps Least Squares - 2SLS
  - The first step is to check the first condition – $Corr(x_i, z_i) \gg 0$
    - The model exhausts the correlation between x_i, z_i → ω_i
    - Homework
  - The second step is to run the model with z_i but with respect to ω_i

$$y_i = (\beta_0 + \beta_1 z_1 + \beta_2 k_2 + \beta_3 k_3, .., + e_i \mid \omega_i(x_i))$$

    - You can even use several IVs
- This models called – IV model or 2SLS – Homework!

Let's take a real example

- IV models are extremely important
  - You can replace x_{today} with x_{past} if the conditions hold (did it)
  - You can replace x with z – if you come up with a good theory
- Let's look at an interesting example:
  - Important question:
    How does the # of police officers effect crime rate in a certain region?
    - Levitt (1997): Estimate the Effect of Police on Crime with an interesting IV approach
  - The basic model:

$$crimerate_t = \beta_0 + \beta_1 police_t +, \dots, + e_t$$

- What do you expect?

Levitt (1977) - IV

- "Shocking Results" – the effect of police (β_1) on Crime rate was **positive**
  - Inconsistent with our intuition!
  - Something is definitely wrong
- What could be wrong?
  - Suggestions?

Levitt (1977) - IV

- The problem: endogeneity
  - $Corr(police_t, e_t) \neq 0$
  - Our estimators are biased
  - There are some underline factors that distorted the results
- How can we overcome this issue?
  - As we learned:
    - We can use lags (y, x)
    - We can use Arima (for e)
  - Still – nothing has changed, the effect is still positive
- The real problem - we have a serious identification problem
  - # of Police officers are too correlated with the error term and we cannot use it in order to predict crime rate
- In a paper from 1997 Levitt came up with an interesting identification strategy for this question

Levitt (1977) - IV

- Steven Levitt simply recall the option to use IV
  - He came up with a theory:
  - He argues that there is an **electoral cycle** in police hiring, with faster hiring in **election years** and slower hiring in other years
  - He then uses **elections** as an instrument for police hiring to estimate the causal effect of police on crime

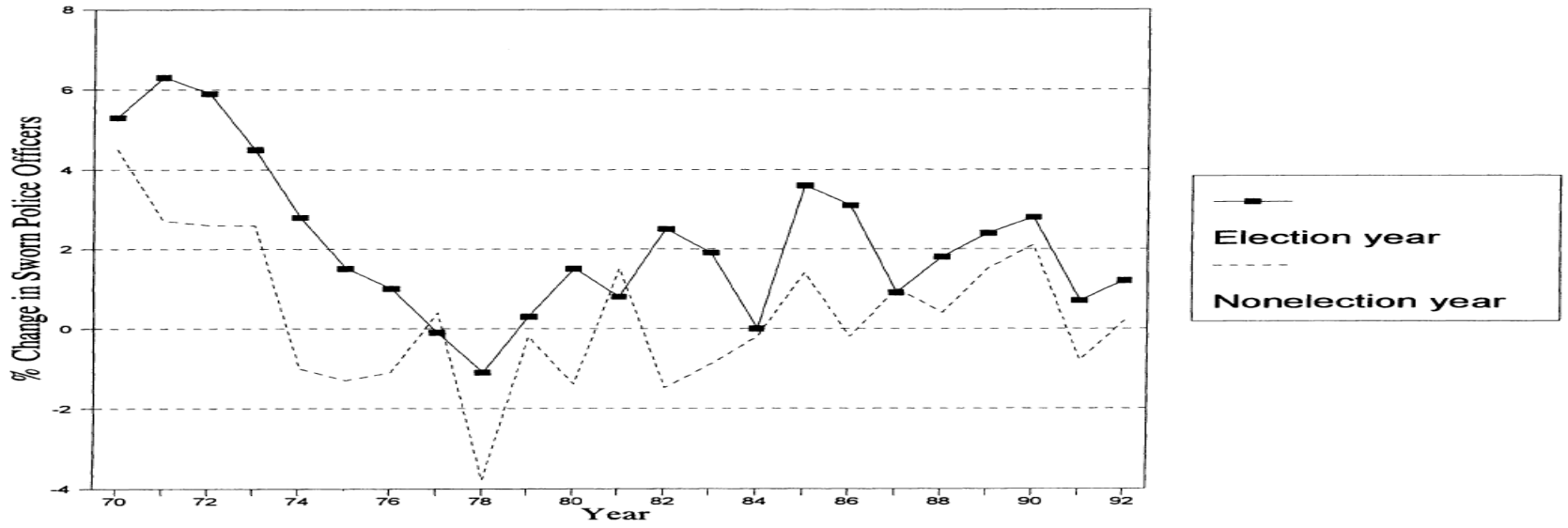


FIGURE 2. YEARLY CHANGES IN SWORN POLICE (ELECTION YEARS VERSUS NONELECTION YEARS)

Levitt (1977) - IV

- Instead of using the number of police officers – use Election Years as a proxy to the number of police officers in a region
  - Genius Strategy!
    - You just need to be creative..
- He firstly checked if: $Corr(police_t, election_t) \neq 0$
  - This came up as true (highly correlated)
    - Intuition
  - Afterwards, he argued, in his paper, that elections are "random"
    - It has nothing to do with un-observed information $Corr(elec_t, e_t) = 0$
      - Elections – are exogenous
      - While – num_police – endogenous!
- After validating his arguments – he used the 2SLS model (IV)

Levitt (1977) - IV

- This time – the results met with our prior intuition
  - The coefficient for election ←proxy→ police – was negative

- What can we learn from all of these?
  - We can try to relax the assumptions (lag, Arima, more variables, time series, etc)
  - We can control for seasonality and trends
  - We can do lots of manipulations in our data
  - But the most important aspect –
    come up with a good identification strategy!

Learning from experience

- My advice to you –
  - Invest time in exploring your data set
    - "it is like walking in Paris by foot" vs. "driving" in Paris
      - You really need to "feel the data"
  - Be sure you are familiar with all the variables you have
  - Read some papers about the question you are trying to answer
  - Be creative! Open your mind! (you do not need to be a mathematician..)
  - Always trust your instinct and do not underestimate your initial intuition (BUT DO NOT TRUST IT BLINDLY)
  - Try multiple models for robustness checks