Some Hints on Writing an Empirical Work[1]

Gabe Lee
University of Regensburg, Germany

for UEH, School of Economics Presentation

December 19, 2019

[1] Adopted from my old supervisor John Cochrane's notes on PhD writing.

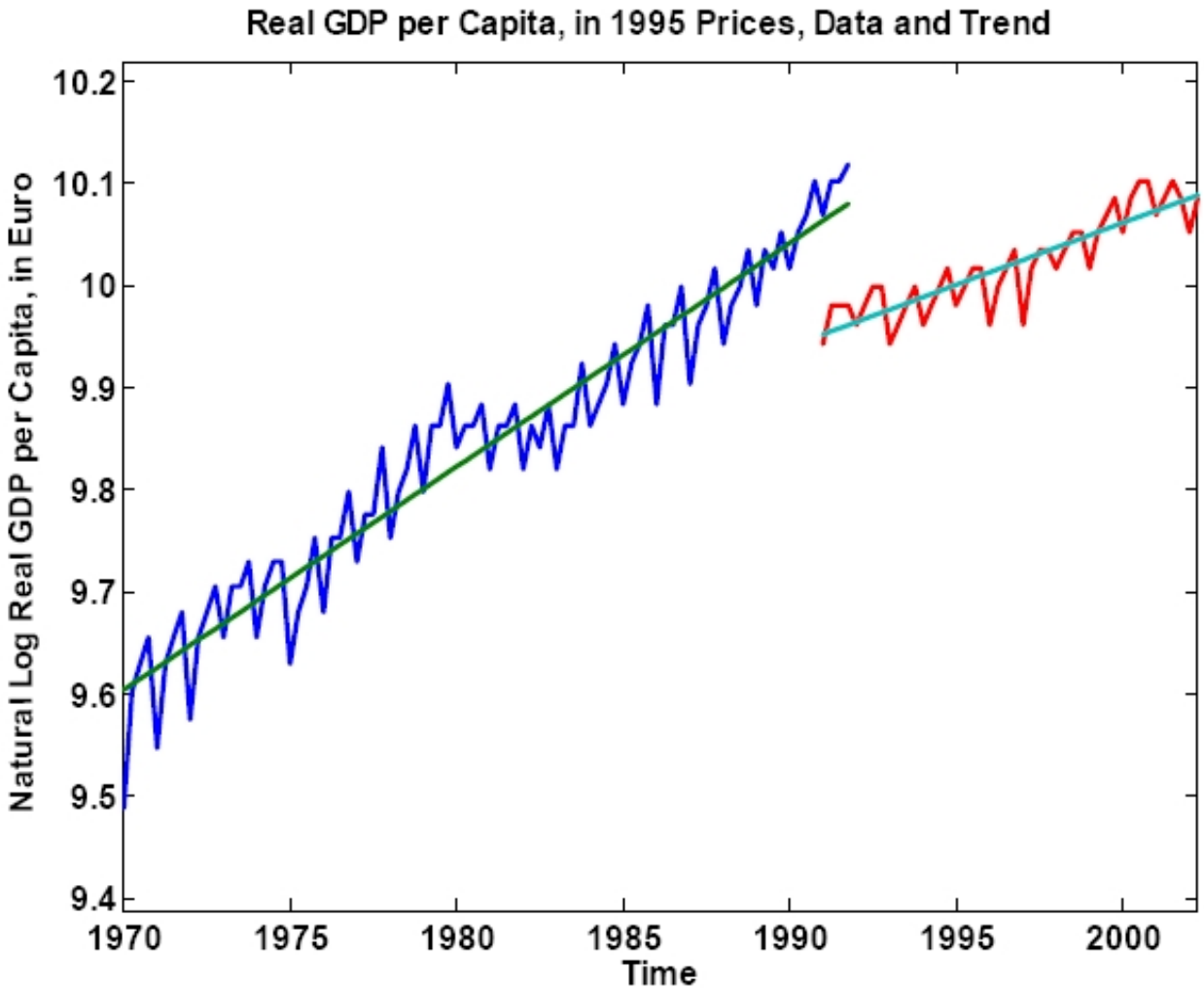
General Comments on Writing an Empirical Paper

- Three most important things in Empirical work:
  1. Identification
  2. Identification
  3. Identification!
- Describe your identification strategy clearly: Method of estimating those relationships and parameters.
  - Most of empirical work boils down to a claim that "A causes B," usually documented by some sort of regression.
  - Explain how the causal effect you think you see in the data is identified.

Some words on Data

- Before talking about the trends, cycles, etc, we need to mention about the data itself.
- Any empirical work should be based on a "good" set of data.
  - Large sample size. (e.g. Central Limit Thm, Law of Large Numbers, etc)
  - No clear structural breaks.
    - This is one of the main reasons why most of the U.S. data are from 1947 onwards.
    - German data: any structural breaks? or any other "problems"?
- "Clear" explanations as to the data itself. e.g. Consistency and the construction of the data should be clear.

- The reason why we want to avoid a "structural break" data in aggregate. Which country could this be?



Some words on Data

- U.S. aggregate data (e.g. GDP, Consumption, Investment, Labor, etc) satisfies all the above requirements.
- The frequency of the data in analyzing business cycles is usually in quarterly and in "real" term (seasonally adjusted).

Empirical Results

- Try to start with the main result.
  - Do not do warmup exercises, extensive data description (especially of well-known datasets), preliminary estimates, replication of others' work.
  - Do not motivate the specification that worked with all your failures. If any of this is really important, it can come afterwards or in an appendix.

Empirical Results

- If you can't follow it, at least **do not** put anything before the main result that a reader does not need to know in order to understand the main result.
- Follow the main result with graphs and tables that give intuition, showing how the main result is a robust feature of compelling stylized facts in the data.
- Try to put robustness checks, etc in the appendix.

Empirical Results

1. Describe what economic mechanism caused the dispersion in your right hand variables.
2. Describe what economic mechanism constitutes the error term. What things other than your right hand variable cause variation in the left hand variable?
3. Hence, explain why you think the error term is uncorrelated with the right hand variables in economic terms.
    - There is no way to talk about this crucial assumption unless you have done items 1 and 2!
4. Explain the economics of why your instruments are correlated with the right hand variable and not with the error term.

Empirical Results

1. Do you understand the difference between an instrument and a control?
   - In regressing y on x, when should z be used as an additional variable on the right hand side and when should it be an instrument for x?
   - $E(xz) \neq 0$ and $E(z \cdot error) = 0$
2. Describe the source of variation in the data that drives your estimates, for every single number you present.
   - For example, the underlying facts will be quite different as you add fixed effects in a panel regression. With firm fixed effects, the regression coefficient is driven by how the variation over time within each firm.
   - Without firm fixed effects, the coefficient is (mostly) driven by variation across firms at a moment in time.

1. Are you sure you're looking at a demand curve, not a supply curve? As one way to clarify this question, ask "whose behavior are you modeling?"

- Classical examples of simultaneous equations: Supply and demand, e.g.
  - labor: hours worked (Q) on wage (P)
  - housing: investment on price (hedonic price)

  The demand function is given by

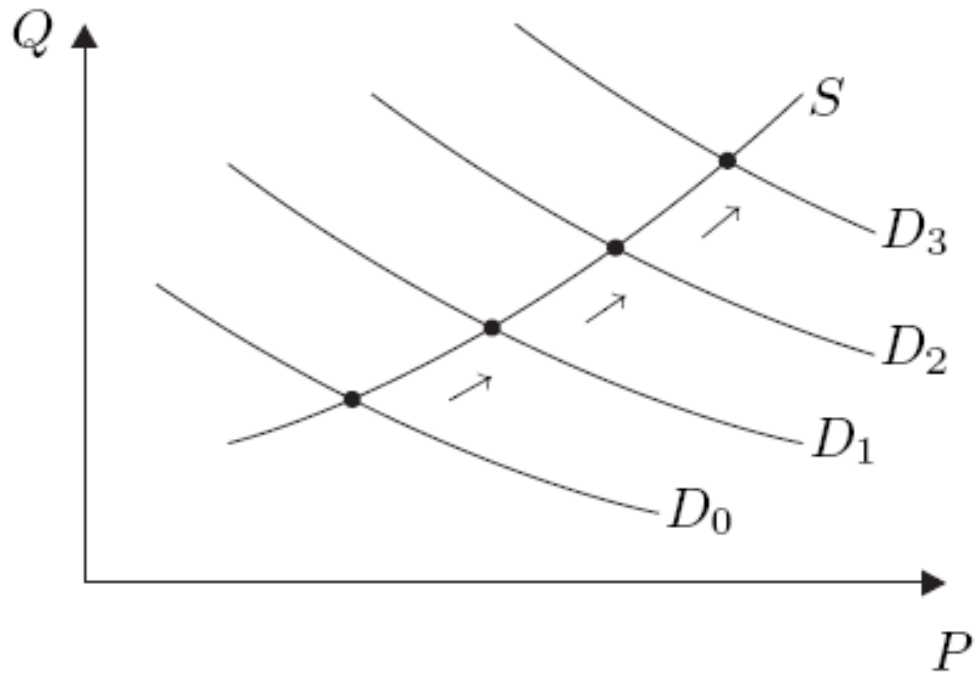
$$Q_t^d = \alpha_1 + \alpha_2 P_t + u_{1t}$$

  and the supply function is

$$Q_t^s = \beta_1 + \beta_2 P_t + u_{2t}$$

  These equations are called "structural" equations, the parameters α_1, α_2, β_1, β_2 are called "structural" parameters.
- Market equilibrium: $Q_t^d = Q_t^s$

- What are the shocks (u_{1t} and u_{2t}) for Q_t^d and Q_t^s?
  - u_{1t}: income, taste, demographics...



  - u_{2t}: input cost, strikes, weather...
  - This tells you that the extra information of income is going to let you identify the supply function:

$$Q_t^s = \beta_1 + \beta_2 P_t + u_{2t} \qquad \text{[using income as an IV]}$$

then $\hat{\beta}_2 > 0$.

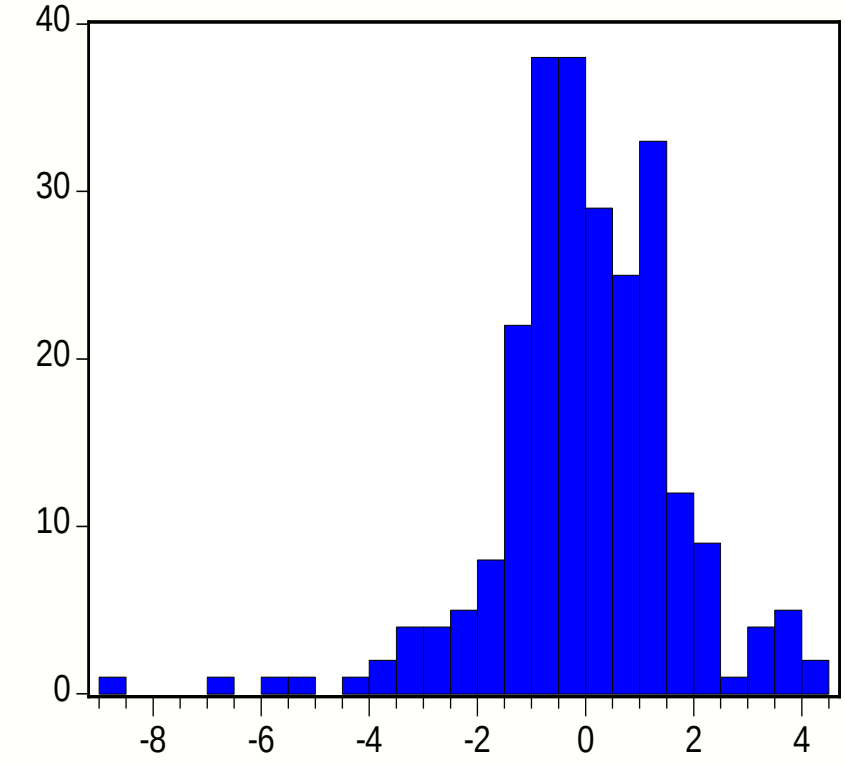
1. Are you sure causality doesn't run from y to x, or from z to y and x simultaneously?
    - Think of the obvious reverse-causality stories.
    - Does price changes supply, or vice versa?
2. Consider carefully what controls should and should not be in the regression.
    - Most papers have far too many right hand variables. You do not want to include all the "determinants" of y on the right hand side.

1. High R^2 is usually **bad sign** — it means you ran left shoes $= \alpha + \beta$ right shoes $+\gamma$price $+$ error.
   - Right shoes should not be a control!
2. Don't run a regression like wage $= a + b$ education $+c$ industry $+$ error.
   - Of course, adding **industry** helps raise the R^2, and industry is an important other determinant of wage
   - But the whole point of getting an **education** is to help people move to better industries, not to move from assistant burger-flipper to chief burger-flipper.

Empirical Results

- Once the data has been transformed (e.g. filtered, logged and stationary), we could perform various descriptive statistics to analyze our data.
- The next few slides are some of the examples of statistics that we macro economists who look at business cycles often employ.
    - General descriptive statistics
    - standard deviations
    - correlations
    - lead - lag structures

Some Stats



Series: PERCENTLGDPHP	
Sample 1947Q1 2008Q2	
Observations 246	
Mean	6.67e-12
Median	-0.038928
Maximum	4.202952
Minimum	-8.507424
Std. Dev.	1.695162
Skewness	-0.849519
Kurtosis	6.533985
Jarque-Bera	157.6017
Probability	0.000000

- Autocorrelation of HP filtered of real GDP
  - A(1) =0.829, A(2) =0.543, A(3) =0.204, A(4) =-0.087, A(5) =-0.299, A(6) =-0.385
  - highly persistent!

Matching the Second Moments of some aggregate variables: Standard Deviations

Variables	Data (1948 - 2001)
Volatility of σ_ω	
Output (GDP)	2.26 (base)
Private consumption (PCE)	0.78 (to GDP)
Labor (N)	1.01 (to GDP)
Nonresidential investment (i_c)	2.3 (to GDP)
Residential investment (i_d)	**5.04** (to GDP)
House price (p_h)	**1.37** (to GDP)

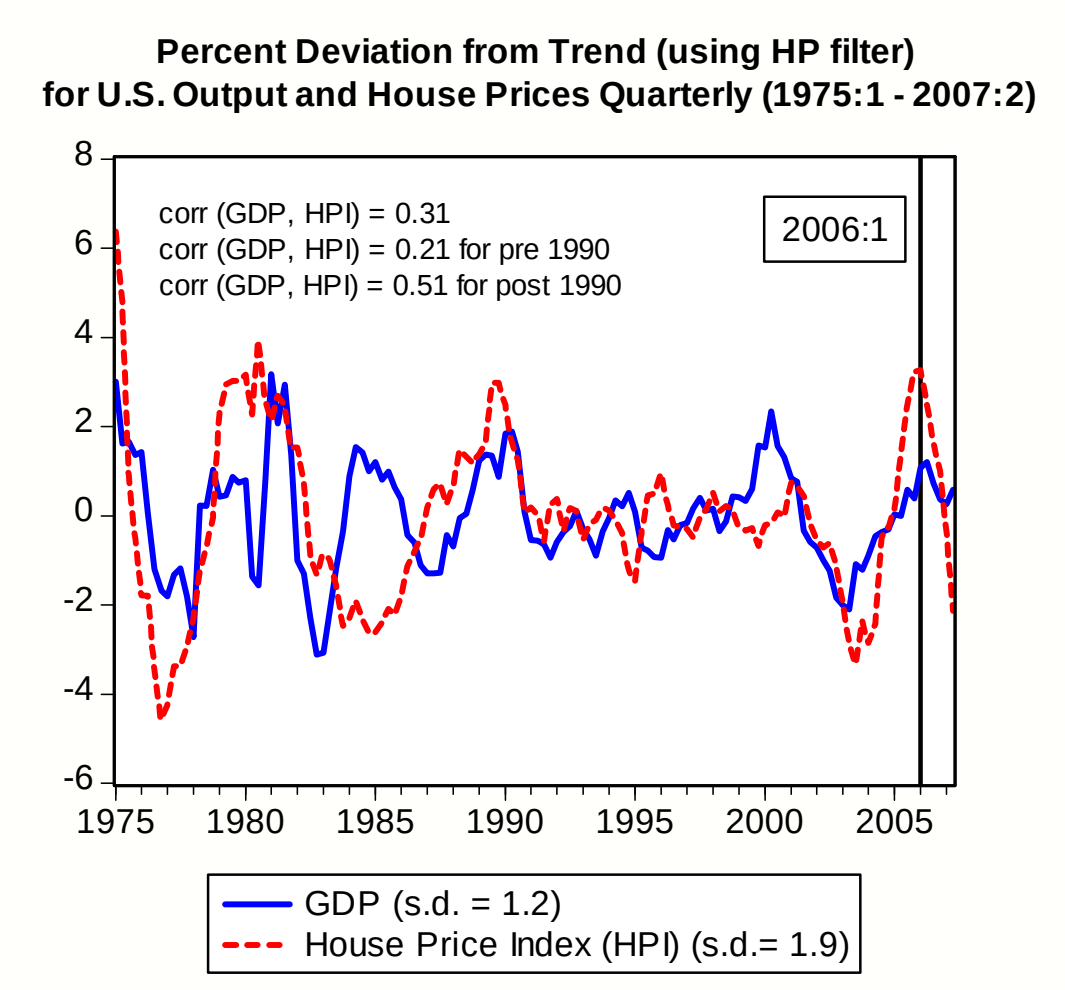
Matching the Second Moments: Correlations

Variables	Data (1948 - 2001) Correlations
(GDP, PCE)	0.8
(GDP, p_h)	0.65
(i_c, PCE)	0.61
(i_d, PCE)	0.66
(i_c, i_d)	0.25
(i_d, p_h)	**0.34**

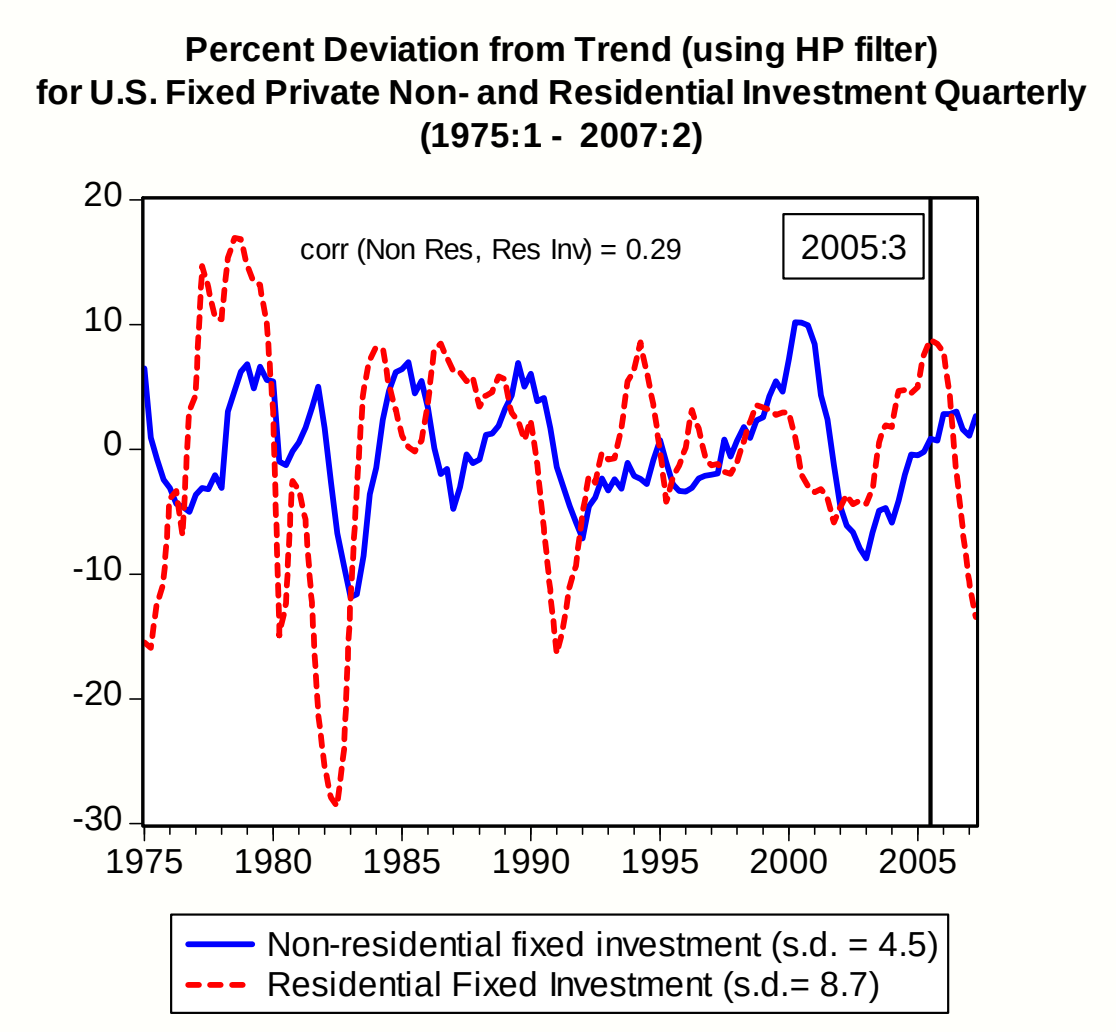
The Lead - Lag Patterns

Variables	Data (1948 - 2001)
$(i_{non-resid}[-1], GDP[0])$	0.25
$(i_{non-resid}[0], GDP[0])$	0.75
$(i_{non-resid}[1], GDP[0])$	0.48
$(i_{resid}[-1], GDP[0])$	**0.52**
$(i_{resid}[0], GDP[0])$	0.47
$(i_{resid}[1], GDP[0])$	-0.22

Fact 1: Housing Price is More Volatile than Output



Fact 2: Residential Investment is more volatile than Non-residential Investment.



Results: Matching the Second Moments

- After the stylized facts in the data, show how your estimation/calibration results are related to those facts.
- For a good example,

	S.D. relative to GDP		
	Model		Data
Loan to Value (LTV)	$m = 0.1$	$m = 0.85$	
House Price	**2.07**	**2.07**	$2.08 \left(\substack{+0.44 \\ -} \right)$
Resid. Investment	2.56	2.55	$5.04 \left(\substack{+0.90 \\ -} \right)$
House Output	2.50	2.51	$5.59 \left(\substack{+0.1 \\ -} \right)$

Some Last Remarks

- Explain the economic significance of your results.
- Explain the economic magnitude of the central numbers, not just their statistical significance.
- Especially in large panel data sets: even the tiniest of effects is “statistically significant”,but does it really matter, economically?