Identification: The problem that just won't die

Monkey Valley, RSA

Peter Berck

Oct. 2013

IDENTIFICATION EXPLAINED

Identification

- I have a set of parameters, β, that I want to estimate.
- My estimate of β maximizes likelihood ,
  - L(β given data)
- When two or more estimates of β maximize L, we say that our model is not identified.

Example: No variation in the data

- $y = \alpha x + \beta z + e$ and $x = kz$. mle is
- $(\propto + k\beta) = (x'x)^{-1}x'y$
  - k is a number.
  - A common problem, actually.
  - Suppose that the shelf price of soda never varies and you estimate a demand function for soda. Now price and the constant term are just multiples of each other.
- So people correctly say that the model wasn’t identified because there wasn’t enough variation in the data.

Example: Time for Event

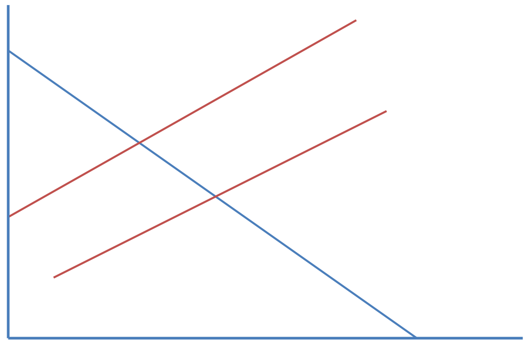
- $y = \alpha x + \beta z + e$
  - y is electric consumption in AM in Sydney in Aug and Sept. 2000
  - x is 1 during DST and zero otherwise during Sept.
  - z is 1 during the Olympics and zero otherwise during Sept.
- Both events happen at the same time. Daylight savings time was extended so the Olympics would look better on TV (less shadow). So z= x. Unidentified.
- There are ALWAYS other things that happen at the same time as interesting events. So timing alone never gives identification. You must make the untestable assertion that everything was not caused by something else, like a

Example: Classic lack of identification

- lemons make lemonade.
- good weather, z, increases lemon supply.
- good weather increases the drinking of lemonade.
- Supply $q = \alpha\ p + \beta z$
- Demand $q = \delta p + \gamma z$
- Means also
- $q = p(\alpha+\delta)/2 + z(\beta + \gamma)/2$
- So I regress q on p. I get a coefficient for z, but is it γ or β or even $(\beta + \gamma)/2$? Can’t tell. Technique won’t help, either.

Classic Identification by exclusion

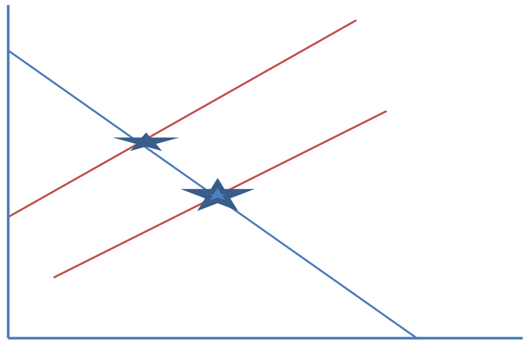
- Supply $q = p\alpha + \beta z$
- Demand $q = \delta p + 0z$
- Now if I regress (by suitable technique) q on p with no z in the equation, I should be able to find δ.



z, weather, shifts only supply and we see two points on the demand curve.

Classic Identification by exclusion

- Supply $q = p\alpha + \beta z$
- Demand $q = \delta p + 0z$
- We just observe the two points and draw a line through them and that is demand.



z, weather, shifts only supply and we see two points on the demand curve.

Example: Unidentified Treatment

- Adopts new stove: $q = T\alpha + \beta z$
- Given discount coupon, treatment, T
- $T = \gamma z$
- Here z is some personal characteristic, like lives near the road where the researchers do their work. Add the two equations together.
- $q = T(\alpha\text{-}1) + (\beta + \gamma)z$
- So regressing q on T doesn’t tell you what the coupon does, unless you KNOW something else.

Identification: x not correlated with the error term e.

- $y = x\beta + e$ is the true equation.
- $b = (x'x)^{-1}x'y$ is the MLE estimate of β
- $b = \beta + (x'x)^{-1}x'e$
  - $\text{plim } (x'x/T)^{-1} = Q$
  - plim $x'e/T$ is possibly not zero, call it ρ
  - $\text{plim } b = \beta + Q\rho$
  - And there we are again, can't know b without knowing ρ

  - Layman's definition of plim $w = w^*$: If we had a whole lot of data, T big, the answer would be almost exactly w^*.

$\text{plim } b = \beta + Q\rho$

- Is the last really new equation you will see this hour, because….
- NO PAGE in Green or Wooldridge gives an econometric trick that will estimate an unidentified model.

Typical Study: Olden Times

- Do small farms have lower value added?
- Cross section 549 farms. Value added (VA), and farm labor, land, schooling, age, etc.
- Two equations:
  - 1. VA regressed on land, labor (L) , personal characteristics, (S).
  - 2. Labor on land, personal characteristics. Factor demand equation.
  - Land is fixed.

Things we can’t see show up in the error term.

- M is motivation.
- e1 and e2 are vectors of other random things
- So the error terms are M + e
  - Cause we can’t see either e or M.

And what is wrong?

- $L = \propto \; S + \ldots + (M + e1)$.
- $VA = \beta * L + \gamma \, S + \ldots + (M + e2)$
- Can we estimate this…
  - $\text{plim} \; L(M + e2)/T = \text{plim}(\; M \; M/T \;)$which isn’t zero.
  - NO.

Unobserved M works two ways.

- Unmeasured motivation drives the farmer to go to work many hours on the farm.
- Motivation also makes them read books about farming and contributes directly to VA.
- So we see labor increase and VA increase but we can’t tell if VA increased because labor increased or if M increasing caused them both to go up.

Classical Solution Technique, IV

- Schooling predicts Labor and schooling is at least predetermined.
- We run the regression, L = b*S, and use it to *predict L, call this L'. L'* depends only on S and so it is NOT correlated with e2 *or M.*
- *VA = b*L' +…(M+e2) works!*
- But is S really independent of motivation, even though it happened earlier?

Incredible

- It is extremely difficult to find an *instrument,* a variable like S, that is truly
  - 1. correlated with our endogenous variable (L)
  - 2. uncorrelated with our error term

How to get Identification

- 1.Random Assignment to treatment and control.
- 2. Take away fixed effects
- 3. Almost random assignment, usually plus fixed effects.
  - Use treated place
  - Use cut off scores
  - Use multiple comparisons
- 4. Use a truly exogenous variable, perhaps weather.

RANDOM ASSIGNMENT

The Ideal: austerity and placebo pills

Why the ideal works

- x is blue or red, austerity or placebo
- y is gdp
- $y = x \beta + e$
- e are all the things we didn’t model
- Blue or red is decided by a coin flip
- The many RSA’s get independent treatments
- x and e are independent.
- We are done.

Random Assignment is Best. Conditional Income Transfer Program

- From a set of rural communities (Mexico) in the same geographic region, localities were randomly selected for participation in **PROGRESA** (treatment localities), while the rest were introduced into the program 2 or more years later (control localities). As the randomization was adequately done (Behrman and Todd, 1999), there is only a small known probability that the differences between treatment and control groups are due to unobserved factors.

Experiments

- Agronomic: vary P K N application
- Field: Give some improved stoves, some not.
- Lab: Dictator games, etc.

GET RID OF INDIVIDUAL EFFECTS

The panel data solution to unobserved individual characteristics.

- Suppose we observe each farm twice or more.
- We let c_i and d_i be dummy terms for farmer i.
- $VA_{it} = c_i + b \text{ Lab}_{it} + \ldots$
- $\text{Lab}_{it} = d_i + e\ S_{it}$ ……
- M is constant for farmer i and so a constant term for farmer i, called a fixed effect, absorbs his M.
- Now e1 and e2 no longer have motivation and every other thing under the sun that is **constant over time** for a single farmer. All works nicely.

So, did we solve the problem?

- Suppose we measure the farms a decade apart.
- We find that farmer Clara has all the same observables but spends more time on her farm.
- Were her personal fixed effects constant over time?
- Suppose her wealth was unmeasured and went up, so she could spend more time at her farming hobby and give up her programming work.
- Or it went down and she needed to eat.

USING ALMOST RANDOMNESS:
CLASSIC DIFFERENCE IN DIFFERENCE

‘good as random’ will have to do

- Does increased wealth lead to less inbreeding?
- Need something random that increases wealth for some and not others.

Consanguinity and Wealth Shock

- A flood embankment in Bangladesh increases wealth only the side of the river it protects.
- Observe 52,000 marriage decisions between 1982 and 1996. Who marries a cousin?
- Use data at village level to create a **panel**. (people get married once, so we don’t see them repeatedly.)

Design

- Two groups (two sides of river) are the same but for chance.
- Two times, before and after dam.
- We see change in consanguinity as people on the good side of river get protected.
- We see change in consanguinity as people on bad side of river don’t get protected.
- The difference in the change is the effect of the embankment.

Marry a cousin equation

- m_{it} % marry cousins in village i at time t.
- d_i dummy variable for village i.
- $em_i = 1$ if village i is on the good side.
- $n_t = 1$ if time is after the embankment is built

- $m_{it} = d_i + \alpha\, em_i + \beta\, n_t + \gamma\, n_t\, em_i + e_{it}$

What have we controlled for

- Village fixed effects
- Just being on the good side, all the time.
- Just getting married after the dam is built, regardless of side.

What did we need to make this work?

- Villages are randomly assigned to one or the other side of the river.
- Best we can do is see that the two groups of villages have matched characteristics.
- Test for same age, wealth, and a whole raft of other things.
- Turns out they differ a tiny bit in wealth.

DISCONTINUITY FOR RANDOM SELECTION

Do Loans Cause College?

- Chile gives college loans to those who score well on a test.
- Does getting a loan change the probability of going to college?
- We have a cross-section. People choose to go to college once. No fixed effects for us.

Enter The Test

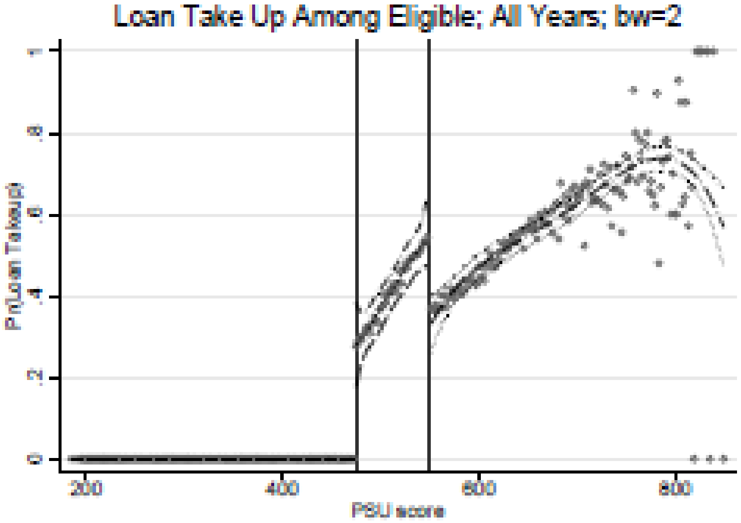
- Can’t get a loan unless you get a 475 on the college test (and are low income.)
- 95% take the test.
- Scores distribution looks like a bell curve. No bunching at 475.

Good as random

- Income, parents ed, etc., etc., etc. are all nice smooth functions of score. No funny stuff at 475.
- So what is the difference between someone who gets 475 and 475 +- 44? Pure, dumb luck.
- The people just below and above are different through randomness.

1/3 who get 475+, take loan

Figure 3: Loan take up. Probability of taking up a college tuition loan among preselected eligible students.



At 475 students become eligible for a college loan and at 550 for a scholarship. The chart shows percent who take the loan.

Eligible for loan, go to college

- Look only at people near the cutoff, so eligibility for loan is RANDOM.
- go to college = $\propto$ eligible for loan $+$ constant
- Very simple regression, because eligible is good as random.
- Result: Loans double college enrollment, particularly for lowest 3 income quintiles

Sloping Land Studies

- Would discontinuity work?
- Sloping land slopes.
  - Projects protect land from soil erosion
  - Just too little slope and just too much slope to be included in the project is ‘good as random’

COMBINING DISCONTINUITY AND PANEL TO OVERCOME ENDOGENEITY

Do customers respond to average or marginal price?

- Electric prices depend on usage. In CA, the more you use, the higher your price. I pay 20 cents per kwh on margin and about 15 cents average.
- People with big houses and pools pay 40 cents on margin and perhaps 30 on average.
- Some people pay as low as 20 cents on average and 40 on margin.
- So does average or margin drive electric use?

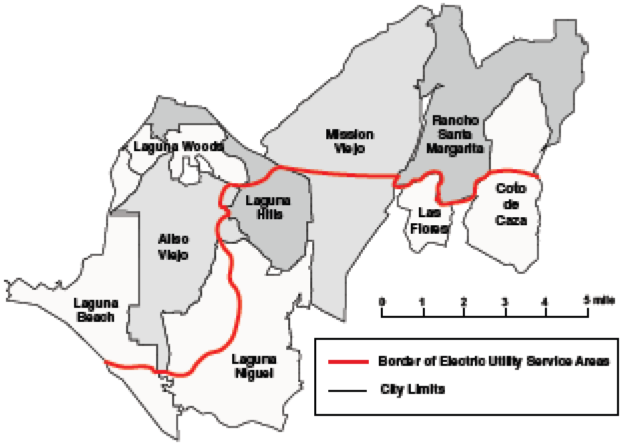
The Identification Problem

- Q = α price + …
- price = f(Q, electric rates) + ….
  - f is nonlinear. Depends on usage and rates.
- We want an instrument for price, that is we want to regress price on exogenous things.
- Could use rates except that everyone faces the same rates, or do they…..

Two electric companies

- SDGE and SOCAL have a border that splits towns.

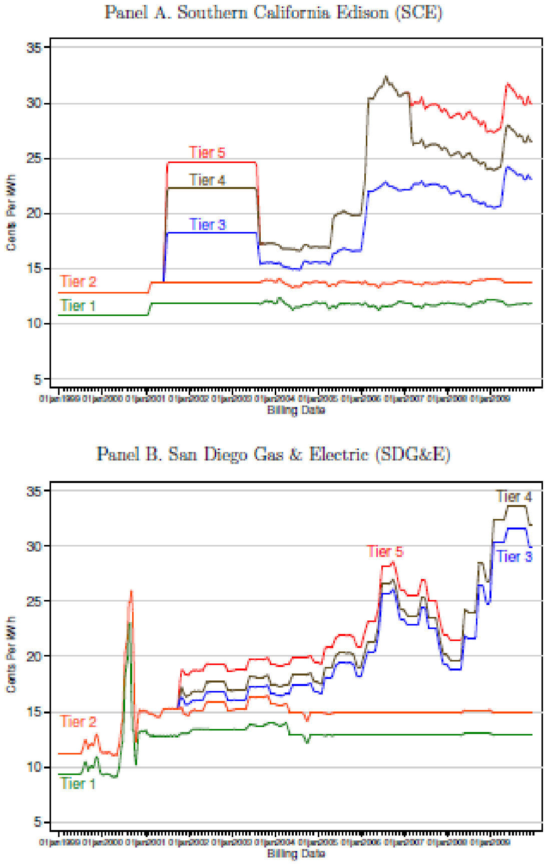
Figure 2: Border of Electricity Service Areas in Orange County, California



Notes: The border of electricity service areas lies within the city limits in six cities. SCE serves the north side of the border and SDG&E serves the south side of the border.

Companies change prices at different times.

Figure 4: Time-Series Price Variation in Nonlinear Electricity Pricing



Panel A. Southern California Edison (SCE)

Panel B. San Diego Gas & Electric (SDG&E)

Beautiful design

- Customers share towns, weather, school systems and everything else, except
- Electricity prices, which DO change over time and DO change differently for the two companies.
- Koichiro Ito uses this to show that people change their usage based on average, not marginal price.

RANDOMIZATION BY MATCHING

Does energy efficiency pay

- In commercial buildings?
- We want a sample of more and less energy efficient buildings and their sale price or rental rates.
- More efficient ones should sell for more.

Create your own control group.

- Building codes include energy efficiency.
- Find buildings built just before the code is strengthened.
- Find buildings built just after the code is strengthened.
- Create a set of matched pairs of buildings, one built before and one built after, that have nearly the same location (1km or so), same size etc.
- Compare their prices.

Yes indeed,

- Stronger energy efficiency in commercial buildings does increase sales price, and more or less by the right amount, too.

COMPARE TO MORE THAN ONE TYPE OF CONTROL

Extended Day Light Savings Time Saves Energy?

- Melbourne Olympics extends daylight savings for just one year and not in all states of Australia.

DST Before, during, after, Olympics

	’99	2000	’01-’05
Victoria	Oct-March	27Aug-March	Oct-March
South Aus.	Oct-March	Oct-March	Oct-March

Victoria had extra daylight savings to make the Olympic games work better.
South Australia did not.
Power consumption was measured every half hour throughout period.

Treated place or not creates DiffnDiff

- So far, Dif = change in usage dst vs reg time.
- DiffnDiff = diff(Vic) – diff(South Aus)
- But the noon hour usage should be unaffected by DST relative to ST.
- DiffnDiffnDiff = DiffnDiff(morning) – DiffnDiff(noon)
- **Treated hour or not creates 3diff.**

What did they see?

- That in DST people use their lights in the morning, so extending DST causes more morning usage.
- They went the extra step and found the error in the engineering-economic model that says DST extension saves power. The code didn’t take account of extra morning usage.
- Americans get up in the dark for no good reason at all.

Taxing water bottles

- The here again, gone again tax.
- Washington state: no taxes, then taxes, then no taxes
- Oregon: no taxes, no taxes, no taxes
- Water gets the Tax, but Juice does not.
- Another triple diff!
  - After or before v. during
  - Washington v. Oregon
  - juice v. water

Results

- 9% tax on bottles gives about 3% decrease in quantity.
- It will take one heck of a tax to make these bottles go away.

EXOGENOUS VARIABLES

Rain, sun, etc

- Rain, sun, cold, etc. are ‘good as random,’ so everything having to do with agriculture and weather ought to be doable.
- Maybe.

Yield

- Yield per acre is a function of weather.
  - fixed effects for plot to account for soil (or use soil and other location variables)
  - fixed effect for year to account for prices (or use prices for inputs and outputs and need an IV)
  - and weather, which changes by year and place and so doesn’t get taken out by the fixed effects.

Revenue and Climate

- $Y_{cst}=\alpha_c+\beta_{st}+\gamma\ \text{weather}_{ct}+e$
  - Y could be profits, corn revenue, etc.
  - Data by county c in state s in census year t.
  - Dummy for county c
  - State x year dummy. (s,t)
- Weather looks like a coin toss, plus the fixed effects for (states and time) and counties.

So what got measured.

- All input and output prices got rolled up into the (state by year) dummy.
- All soil variables in the county dummy.
- Average weather by state in the (state by year) dummy.
- So weather is what is left, variation of weather within a state. But most of the variation in weather is between states. Perhaps not ideal.

Response to temperature…

- Temperature doesn’t do much in this setup, because
- 1. weather effect is attenuated.
- 2. welfare includes consumers and this only looks at producers.

Demand and Supply of Calories with changing weather

- Supply depends on summer weather, planting time price, …
- Demand depends on harvest time price …
- Planting time price depends on all past weather, because carry in depends on past weather..
- Harvest time price depends on all past weather plus summer weather.

The use of different slices of weather

- Gives instruments for price planting and harvest time.
- All is well.
- And hotter weather really reduces calories in this model.

Conclusion: Three Ways to get estimates

- Kill off unobservables, like motivation.
- Find contrasts between things that were and were not treated.
- Finding exogenous variables that affect what we care about, though perhaps indirectly.