

Moving Beyond 10%: Specification Issues in Comparative Research

Nathaniel Beck (with the help of many friends)

Department of Politics, NYU, New York, NY 10012, nathaniel.beck@nyu.edu

Prepared for Lecture at Essex Summer School, August 15, 2007

Introduction

- Estimation issues - important but small
- SO GET THEM RIGHT, but that is just the start
- That is, this is not an excuse to either get them wrong or ignore them
- E.g. PCSEs, heteroskedasticity
- Specification Issues - large but hard
- Distinction not hard and fast
- E.g. Instrumental variables, dynamics, fixed effects (all big!)
- When correctly viewed hard estimation issues are specification issues
- Important critiques of qualitative analysts
- Or an excuse to recycle some old ideas and think about some interesting issues
- Could have chosen lots of other important things, but they are not things I have worked on
- Prefer to not just do a long laundry list

Platitudes - nothing new

- New and better variables
- Better and more precise theory
- Dealing with endogeneity/selection - critical but I have nothing much to add to Heckman and Rubin
- Spatial stuff - really important, but leave to Gleditsch
- Better measurement
- Combining theory and measure (e.g. new work using item response theory, Jackman and others)
- Focus here on specification and what data can say about specification
- Focus on comparative politics (broadly defined, includes IR, but excludes behavior)



Easy example: count data

- Suppose that y_i is count data
- Generate 300 Poisson random draws, y_i
- Let X be uniform on the unit interval
- Let $E(y_i) = \lambda_i = f(x_i)$
- DGP1: $\lambda_i = e^{\beta x_i + \alpha}$, $\beta = 1$, $\alpha = 2$
- SEs in parentheses
- Using Stata Poisson command, $\hat{\beta} = 1.03(.06)$, $\hat{\alpha} = 1.99(.04)$
- Quite good
- If foolishly regress y on x , get $\hat{\beta} = 12.9(.7)$, $\hat{\alpha} = 6.3(.4)$
- But this is model with WRONG MEAN (linear, not exponential)
- Since for this data $y_i > 0$, can get the mean “right” by
- regressing $\ln y$ on x , get $\hat{\beta} = 1.04(.05)$, $\hat{\alpha} = 1.94(.03)$
- Pretty close
- By all means do Poisson, but the linear with the right mean does just about as well



Poisson can make bad assumption too

- Linear with badly misspecified performs poorly
- When fit a Poisson model with ml (as in Stata) IMPLICIT assumption is that the mean is exponential
- Guarantees that EV of event count is positive, which is good
- But who said that GOD (Generator Omniscient of Data) chose this nice simple form
- Suppose DGP2: $\lambda_i = \beta x_i + \alpha \cdot \beta = 10, \alpha = 3$
- Big parameters to make $\lambda > 0$
- Using Stata Poisson command, $\hat{\beta} = 1.32(.07), \hat{\alpha} = 1.32(.05)$
- Using linear regression, $\hat{\beta} = 10.15(.53), \hat{\alpha} = 2.75(.31)$
- Now which one is bad?
- Cannot save Poisson by exponentiating, get $\lambda = 3.7x^{1.3}$
- So if know mean function, and use it, old linear regression about as good as Poisson, implicit mean function in Poisson is bad if wrong
- Can fix, but have to write own ML code



More generally

- Cannot be a sane world when we
- Do Poisson (or negative binomial) because DV is number of signs rather than pounds spend on signs
- But would do OLS for amount spend on signs
- Why constant returns for money but increasing returns for count
- Or logit on whether or not have a sign
- Maximal returns here when probability is .5
- Cf. Alt, King and Signorino, PA
- Cannot choose models based on how some analyst chose to measure the DV
- All different methods should be commensurate



Slightly harder example: fixed effects

- Thinking in terms of specification, fixed effects avoid omitted variable bias
- But it changes the question to how much does x affect y to
- how much do changes in x over time IN A UNIT affect changes in y
- So a big issue, not just a bias issues
- Will not talk about this for continuous dv's since you have Plümper and Troeger

FEs with duration data

- If we have continuous time durations (that is, for each unit one observation which is has date of death)
- Would never dream of putting in FE's since would explain perfectly
- But if put in discrete yearly form, could put in FE's
- As before, is eliminating any interest in cross-sectional variation
- Again, as in AKS, changing the data format should not change our modeling of quantities of interest
- But we lose all countries, say, that are always dictatorships
- Question of what makes some countries become democratic is very different from
- Amongst countries which moved from autocracy to democracy, what led some to do it sooner
- Two different questions, be sure of which one you want to answer
- So FEs are not about estimation, but specification
- For IR, with lots of censored states which are dropped by FEs DO NOT use FEs

Various issues

- ① Getting the mean MORE right
- ② Complicated interactions - non linearities - causal complexity
- ③ Dynamics/State dependence
- ④ Choice of cases (and time period) - unit heterogeneity - causal thinking again

Drop linearity but keep additivity

- Linear model $E(y_i) = \mathbf{x}_i\beta$
- Additive model $E(y_i) = \sum_{j=1}^k f_j(x_i^j)$
- Where f_j is some smooth function
- We often think about simple f 's, say $f(x) = \ln(x)$, $f(x) = x^2$
- First imposes decreasing returns, second increasing
- What if do not know?
- Sensitive to range of x ; how handle $\ln(0)$?

Box-Cox

- Well known Box-Cox transform at least asks data about returns
- Transform is $f(x) = \frac{x^\lambda - 1}{\lambda}$, $x > 0$, $\lambda \neq 0$
- If $\lambda = 0$, $f(x) = \ln(x)$
- λ is a parameter that is ESTIMATED, not assumed
- Subsumes linear, logarithmic, quadratic, etc.
- Almost never used. Not sure why?
- Why assume what you can EASILY test?

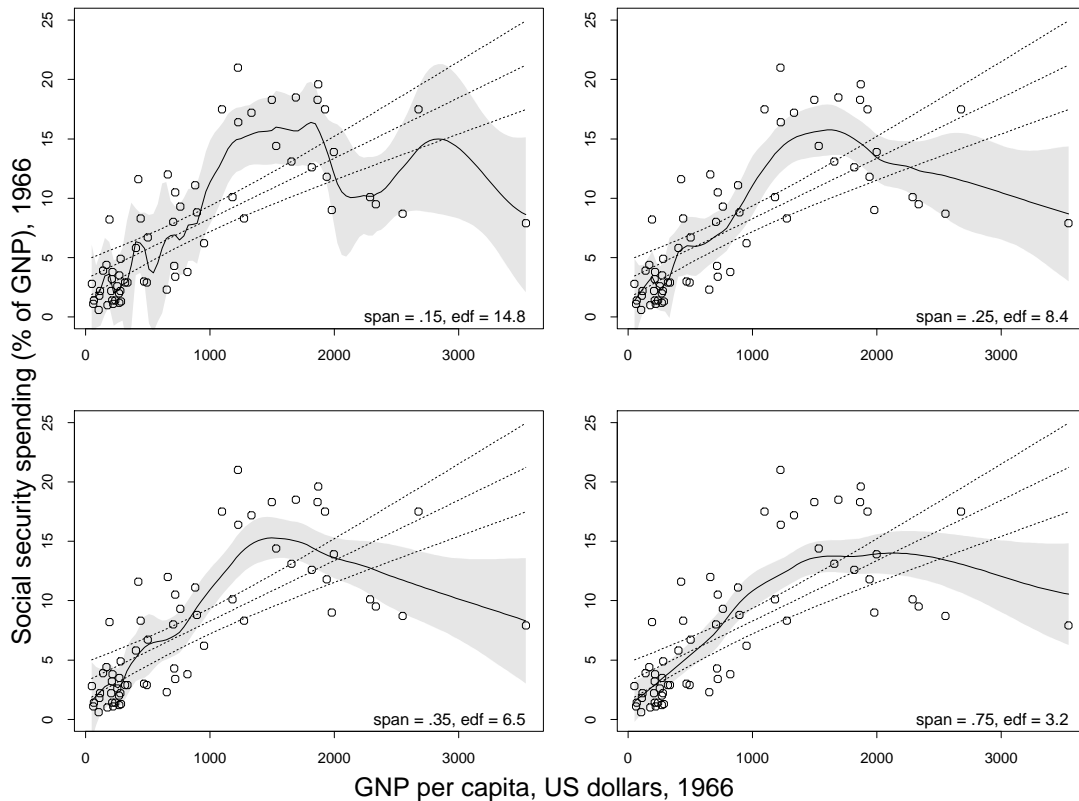


GAM

- Assume mean is simply additive function of smooth functions
- $f(x)$ is some spline or other smooth function (LOESS)
- For expository work in PS, see Beck and Jackman
- Almost never used. Is software too hard? (bad reason!)
- Not only deals with non-linearity, but also influential points
- At almost no real cost (other than learning R)
- For single variable (welfare spending as proportion of GDP against GDP/Capita)
- Loess Smooth



4 different smooths- from Beck and Jackman



Beck (NYU)

Specification

Essex

13 / 31

Causal complexity

Necessary and sufficient conditions

- Qual folks have stressed complicated N&S stuff
- But they do in a binary world where N&S is easy
- What about with continuous data
- Suppose some level of GDP is necessary for democracy
- How do? Without knowing what level, what do we do
- Can do by theory or assertion
- E.g. Politically relevant (large power or contiguous) dyads
- Claim: relevancy NECESSARY for conflict
- Good idea but alas, probability of conflict in irrelevant dyads only about one fourth of PRDs, not zero



Beck (NYU)

Specification

Essex

14 / 31

Basic issue: causal heterogeneity

- Problem is that causal effects may vary from unit to unit, so that the average causal effect may mask a few good sized effects combined with many small ones
- Imaging that there are ten kinds of stomach cancers
- We cannot differentiate them empirically
- We do a standard trial on a group with stomach cancer
- We find no significant impact
- But perhaps this is an average that the drug is great for one or two types of cancer and doesn't work for the others
- How can we deal with this?
- Similar issue for conflict, coups, other things
- Can see this as there are necessary conditions for something to work
- But no easy way to know A PRIORI what those necessary conditions are EXACTLY (so may want to estimate)



Possible solutions

- With dummy variables interactions easy
- With continuous can multiply but MAY be far off (is rigid and sensitive to scaling)
- Can do 2 dimensional gams
- Alas curse of dimensionality!!
- So no simple way to do non-parametric multidimensional models
- Sometimes two dimensional smoothes work
- How does democracy of two nations go together to change conflict?
- But hard
- Does not generalize to higher dimensions - curse of dimensionality



Random coefficient (hierarchical/multilevel) models

- Let data speak about units being heterogenous
- RCM can be used to nicely model heterogeneity
- If just have $\beta_i = \beta + \nu_i$
- get heterogeneity but in uninteresting way
- But if make $\beta_i = \beta + \gamma z_i + \nu_i$
- get serious modeled heterogeneity
- Much better way of thinking of many interactions
- But even if just estimate the random β_i , if one is far away from others, indicates heterogeneity
- For replicated data, this model comes at pretty low cost
- If no heterogeneity, RCM seems to say this
- Why not do more? Not sure Maybe new software (Stata, MCMC (Gelman book) will make more common



More extreme empirical solutions

- Neural nets (Beck, King and Zeng)
- Small effects found in conflict studies MAY average a few large effects and many tiny ones
- Generalize idea of political relevance beyond all or nothing
- Let data find complex high order interactions via neural nets
- Use good Bayesian nets, with a strong prior for simplicity
- Heavy use of out of sample validation
- In one try seemed to work okay
- Another possible method is CART
- Classification and regression trees
- Trying to find high order interactions
- Seems underutilized



Ignore standard time series issues today

- (Often much) more than 10%, but orthogonal to talk today
- Getting TS wrong is BIG
- Ignoring Serial Correlation is BIG
- Ignoring whether series are integrated is BIGGER
- Spurious regressions
- We know (somewhat) how to deal with this
- For TS think about specification implications
- That is, impulse response functions
- So note diff between lagged dependent variable and serially correlated error specifications
- But no more here



Markov Transition Matrices

- A Markov process (first order) assumes that whether or not you are 0 or 1 at time t is a function only of where you were at time $t - 1$ and covariates.
- Thus would estimate two different probits (or logits) depending on prior state

$$P(y_{i,t} = 1 | y_{i,t-1} = 0) = \text{Probit}(\mathbf{x}_{i,t}\boldsymbol{\beta}) \quad (1)$$

$$P(y_{i,t} = 1 | y_{i,t-1} = 1) = \text{Probit}(\mathbf{x}_{i,t}\boldsymbol{\alpha}) \quad (2)$$

which can be written more compactly as

$$P(y_{i,t} = 1) = \text{Probit}(\mathbf{x}_{i,t}\boldsymbol{\beta} + y_{i,t-1}\mathbf{x}_{i,t}\boldsymbol{\gamma}) \quad (3)$$

where

$$\boldsymbol{\gamma} = \boldsymbol{\alpha} - \boldsymbol{\beta}. \quad (4)$$

- Test hypothesis that prior state does not matter by testing $\boldsymbol{\gamma} = \mathbf{0}$



Comparison of Markov and LDV model

- An alternative that some have suggested is to use a lagged dependent variable

$$Pr(y_{i,t} = 1 | y_{i,t-1} = 1) = \text{logit}(\mathbf{x}_{i,t}\beta + \rho)$$

whereas

$$P(y_{i,t} = 1 | y_{i,t-1} = 0) = \text{logit}(\mathbf{x}_{i,t}\beta)$$

so the two logit equations are parallel (in the latent space).

- Thus the only thing that differs by prior state is the intercept, the effect of all the iv's on the dv is the same regardless of whether past state was 0 or 1.
- Note this is just a strong restriction on the Markov transition model
- and could be tested by first estimating the full Markov model and then testing the null that all coefficients (other than the intercept) are the same in both probits



Should it matter if state is measured continuously or discretely

- If, like Przeworski, we measure DEMOC as binary
- Then Markov transition model is natural
- But suppose we measure as continuous (in the sense of pol sci) say with Polity
- Now we typically just add Polity score to equation and do not assume that different model for nations at top and bottom of scale
- But how can a choice of how to measure sanely determine our model?



Transitions from DEMOC to AUTOC and Vice-versa 1951-1990 - Przeworski, et al.

- 135 countries
- Spells of democracy
 - 1683 country years
 - 72 spells of DEMOC
 - 38 spells of DEMOC end in AUTOC
 - 34 spells of DEMOC right censored
- Spells of autocracy
 - 2530 country years item 101 spells of AUTOC
 - 49 spells end in DEMOC
 - 52 spells right censored



Beck (NYU)

Specification

Essex

23 / 31

Dynamics - Path Dependence

Democracy/Autocracy

Variable	ALL		DEMLAG		From AUT		From DEM	
	b	SE	b	SE	b	SE	b	SE
GDPLAG	.33	.01	.16	.02	.12	.03	.22	.05
GDPLAG%	-.57	.35	-.18	.69	-1.97	.85	3.96	1.38
DEMLAG			3.75	.10				
C	-1.32	.04	-2.41	.08	-2.30	.10	1.12	.14
N	4126		3991		2407		1584	



Beck (NYU)

Specification

Essex

24 / 31

Choice of cases

- Often we take what is given to us by data collection organization
- But as EU enlarges, and collects data on E. Europe, does that mean that the causal process (equation) governing those new countries is the same as the W. European countries
- We have our own choice of what units to study
- Qualitative analysts often accuse quantitative analysts of “concept stretching” or ignoring “causal heterogeneity”
- Not totally inaccurate
- But also an issue relevant already in quant research
- How might we proceed without limiting ourselves to case studies?



Matching, overlap

- Will not talk about literature of causality/selection
- Except insofar as deals with what units to compare
- One of the great insights of labor economics matching is that to study the effect of a labor market training program
- We do not include rich people who are far from all who got the program
- If we did, we would have to believe that we could extrapolate a lot from rich people to folks eligible for labor market programs
- If just put income in a linear regression, same assumption, but now a BIG LINEAR extrapolation

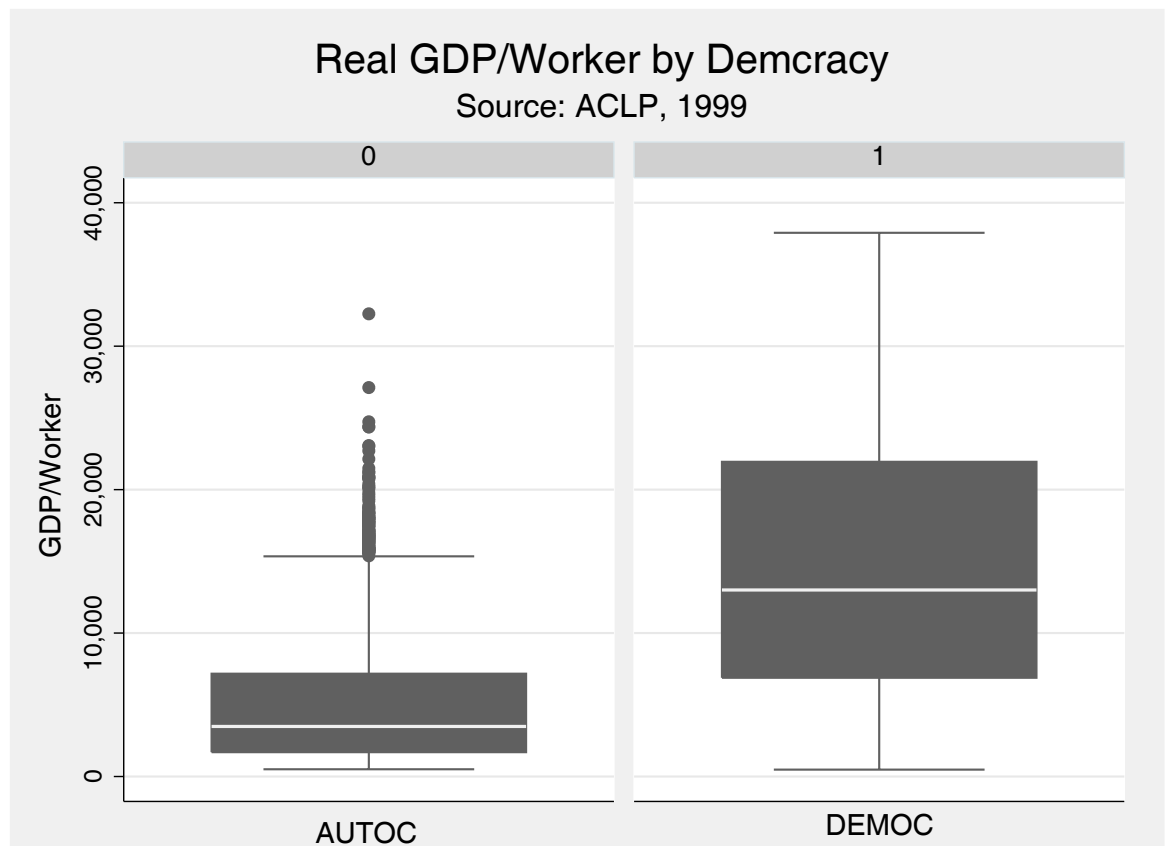


- Similar to idea of King and Zeng about how counterfactual is your counterfactual
- Thus, even in a regression, say with DEMOC, we might think about excluding all DEMOCs that are very far from all AUTOCS
- Else we are leaning heavily on linear extrapolation
- Example below is not the worst case
- Rich AUTOCS are oil exporters
- Would need balance on lots of other things
- Even if doing standard statistics, balancing via propensity scores first is good thing (see Ho, et al., PA)
- This just slightly formalizes the idea that we do not want to study transitions to democracy adding in the OECD nations

Ideas from case control designs

- Case control studies in medicine - have a whole bunch of cases (maybe all of them) for rare condition
- What to use as controls?
- Expensive to collect data on controls
- If use large random sample, few will be similar to cases, so expensive
- Instead, find a group of controls relatively similar to cases and collect data on them
- Often use people in hospital without that disease
- Can be seen as a way of allowing for efficient data collection in this difficult situation
- But also a way to do intuitive matching
- f relatively rare cases (coups, MIDs), might think about a case-control type setup
- Even if already have the universe of controls might want to limit analysis to controls that are similar to cases
- Same idea as before, different words

Simple example - real life is worse with more variables



Choice of units - cross-validation

- Do not just take all units (countries) as given by data gathering org.
- Do they all follow same equation?
- Are any really different?
- Cross validation
- Leave out one unit at a time
- Estimate on others
- Compute “forecast” error for left out unit
- Do some units forecast much worse?
- Is there reason to expect them to differ for good reason?
- Is difference in forecast error large enough?
- Do NOT keep dropping units that do not fit!
- Do not drop units unless there is some argument they do not fit

Conclusion

- This talk is a beginning, no conclusion
- Or the conclusion is get the easy trivial things right
- But think about the really big issues
- Or this talk needs no conclusion