Observing the Unobservable in a Stepwise Approach: Teasing Out Causal Effects through Panel Data

Sinisa Hadziabdic
Department of Sociology
University of Geneva

Methods & Research Meetings, Lausanne
6.10.2015

GENEVA SCHOOL
OF SOCIAL SCIENCES
Department of Sociology

UNIVERSITÉ
DE GENÈVE

Outline

- The Concept of "Causal Effect"
- Living with Observational Data
- Endogeneity
- Panel Data and Unobserved Heterogeneity
- Time-Invariant and Time-Varying Endogeneity
- Applying It All...
- Conclusions and Critical Regard

The Concept of "Causal Effect" (1)

- An empirical example: the causal effect of union membership ("Non-member" or "Member") on:
  - Job attitude: satisfaction with working conditions
  - Political attitude: interest in politics
- Swiss Household Panel data between 1999 and 2011

The Concept of “Causal Effect” (2)

- Causality ⇔ causal mechanism identifiable
- Counterfactual definition of causality:
  - For a single unit:

$$\Delta \text{attitude}_i = (\text{attitude}_i \mid \text{member}_i) - (\text{attitude}_i \mid \text{non} - \text{member}_i)$$

  - Average Treatment Effect (ATE):

$$E[\Delta \text{attitude}_i] = E[\text{attitude}_i \mid \text{member}_i] - E[\text{attitude}_i \mid \text{non} - \text{member}_i]$$

The Concept of “Causal Effect” (3)

- The “Fundamental Problem of Causal Inference”:

$$\mathbf{E[\Delta attitude_i] = E[E[\Delta attitude_{i\,\epsilon\,member}] + E[\Delta attitude_{i\,\epsilon\,non-member}]] =}$$

$$\mathbf{E[(E[attitude_{i\,\epsilon\,member} \mid member_i] - \boldsymbol{E[attitude_{i\,\epsilon\,member} \mid non-member_i]})}$$

$$\mathbf{+ (\boldsymbol{E[attitude_{i\,\epsilon\,non-member} \mid member_i]} - E[attitude_{i\,\epsilon\,non-member} \mid non-member_i])]}$$

⇒ four terms, only two observed ones

Living with Observational Data (1)

- Substitute the Average Treatment Effect with the observed difference between treatment and control group:

$$\mathbf{E}[\boldsymbol{\Delta}\mathbf{attitude}_{\mathbf{i}\,\epsilon\,\mathbf{observed}}] =$$

$$\mathbf{E}[\mathbf{attitude}_{\mathbf{i}\,\epsilon\,\mathbf{member}} \mid \mathbf{member}_{\mathbf{i}}] - \mathbf{E}[\mathbf{attitude}_{\mathbf{i}\,\epsilon\,\mathbf{non-member}} \mid \mathbf{non-member}_{\mathbf{i}}]$$

- Two sources of bias may lead the observed difference to differ from the “true”, counterfactual Average Treatment Effect:

$$\mathbf{E}[\boldsymbol{\Delta}\mathbf{attitude}_{\mathbf{i}\,\epsilon\,\mathbf{observed}}] = \mathbf{E}[\boldsymbol{\Delta}\mathbf{attitude}_{\mathbf{i}}] +$$

$$\{E[attitude_{i\,\epsilon\,member} \mid non-member_{i}] - \mathbf{E}[\mathbf{attitude}_{\mathbf{i}\,\epsilon\,\mathbf{non-member}} \mid \mathbf{non-member}_{\mathbf{i}}]\}$$

$$+ \{(1-\pi) *$$

$$((\mathbf{E}[\mathbf{attitude}_{\mathbf{i}\,\epsilon\,\mathbf{member}} \mid \mathbf{member}_{\mathbf{i}}] - E[attitude_{i\,\epsilon\,member} \mid non-member_{i}])$$

$$-(E[attitude_{i\,\epsilon\,non-member} \mid member_{i}] - \mathbf{E}[\mathbf{attitude}_{\mathbf{i}\,\epsilon\,\mathbf{non-member}} \mid \mathbf{non-member}_{\mathbf{i}}]))\}$$

Living with Observational Data (2)

- The selection bias:

  $$\{E[attitude_{i\,\epsilon\,member}\mid non-member_i] - \mathrm{E}[\mathrm{attitude}_{\mathrm{i\,\epsilon\,non-member}}\mid \mathrm{non-member_i}]\}$$

- Baseline difference in the outcome variable between treatment and control group in the non-treated status (violation of the “conditional independence” assumption): correlation between the chances of becoming a union member and the outcome variable
- Solutions:
  - Random assignment of the treatment variable
  - Selection on observables: matching techniques and regression analysis
- Example: job satisfaction:
  - Union members: 6 as non-members (unobserved) and 7 as members (observed)
  - Non-members: 8 as non-members (observed) and 9 as members (unobserved)

    ⇒ “true” causal effect = + 1 vs. observed difference = -1
  - Selection on observables, e.g. income:

  $$E[attitude_{i\,\epsilon\,member}\mid non-member_i, income_i] = \mathrm{E}[\mathrm{attitude}_{\mathrm{i\,\epsilon\,non-member}}\mid \mathrm{non-member_i}, income_i]$$

Living with Observational Data (3)

- The «heterogeneous effects» bias:

$$\{(1-\pi) * ((\mathbf{E}[\mathbf{attitude}_{\mathbf{i} \,\epsilon\, \mathbf{member}} \,|\mathbf{member}_{\mathbf{i}}] - E[attitude_{i \,\epsilon\, member} \,|\, non-member_{i}]) - (E[attitude_{i \,\epsilon\, non-member} \,|member_{i}] - \mathbf{E}[\mathbf{attitude}_{\mathbf{i} \,\epsilon\, \mathbf{non-member}} \,|\, \mathbf{non-member}_{\mathbf{i}}]))\}$$

- The «true» causal effect differs in treatment and control group
- A negligible bias? An external validity issue
- Composition issue vs. underlying causal mechanisms
- Solutions:
  - Randomization
  - No real solution: re-estimation of the effect on different segments of the population

Endogeneity (1)

- Switch to a regression framework:

$$\mathbf{attitude_i = \alpha + \beta\, union_i + \xi_i, \qquad for\ i = 1, 2, \ldots, N}$$

- Under which conditions does β equal the «true» causal effect on the treated?

  $\Leftrightarrow E[\varepsilon_i] = E[union_i|\varepsilon_i] = 0$ (exogeneity)

- Conditional independence assumption equivalent to absence of selection bias

Endogeneity (2)

- Sources of endogeneity: relevant omitted variables and reverse causality (measurement error and functional form miss-specification)
- Solutions in a cross-sectional setting:
  - Selection on observables: control variables -> unobserved variables?
  - Instrumental variables -> difficult to find valid instruments, inefficiency, external validity (generalizable only to the unknown population of compliers, unless homogeneity of the effect is supposed)
- Example: job satisfaction

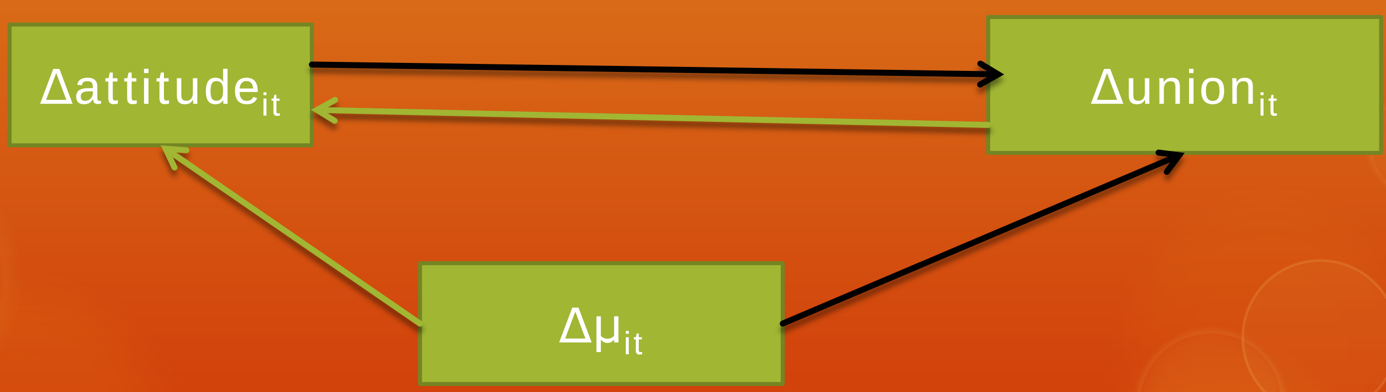
Panel Data and Unobserved Heterogeneity (1)

- Panel data setting:

$$\mathbf{attitude}_{it} = \beta\,\mathbf{union}_{it} + \nu_i + \mu_{it}\,, \qquad \text{for } i = 1, 2, \ldots, N \text{ and } t = 1, 2, \ldots, T$$

- Model variations instead of original levels:

$$\Delta\mathbf{attitude}_{it} = \beta\,\Delta\mathbf{union}_{it} + \Delta\mu_{it}\,, \qquad \text{for } i = 1, 2, \ldots, N \text{ and } t = 1, 2, \ldots, T$$



Panel Data and Unobserved Heterogeneity (2)

- Same endogeneity issues, but potential sources of bias much less numerous
- Problems by modeling variations instead of levels:
  - Time-invariant invariables excluded (even though remedies exist)
  - Efficiency: need enough within variation
  - External validity: is the effect homogeneous across individuals with high and low within variation? Individuals with high within variation contribute more to the estimated Average Treatment Effect on the treated
- Assumptions on regressors and disturbances ⇒ consider different types of variations:
  - Fixed effects: variations with respect to the individual mean (within transformation): $\Delta_{FE}x_{it} = x_{it} - \bar{x}_i$
  - First-differencing: variations with respect to the level in the previous year: $\Delta_{FD}x_{it} = x_{it} - x_{i,t-1}$
  - More exotic alternatives: adjusted differencing procedure

Panel Data and Unobserved Heterogeneity (3)

- Fixed effects vs. first-differencing:
  - Exogeneity requirements:
    - Strict exogeneity $E[x_{it} \mid \mu_{is}] = 0$ vs. $E[\Delta_{FD}x_{it} \mid \Delta_{FD}\mu_{it}] = 0$ for all i, t, s
  - Exogeneity not satisfied :
    - Bias is O(1/T) vs. better suited to deal with non-stationarity (in micro panels T is small)
    - Instrumental variables need to satisfy strict exogeneity and have to be time-varying (need to transform them too) vs. weaker exogeneity assumption and no restrictions on type of instruments (levels of variables also allowed)
  - Efficiency:
    - No serial correlation vs random walk (truth usually somewhere in-between…)
  - Fixed effects can be derived in an "elegant" way through the results of the partitioned regression theorem and works better with data gaps
- Upshot: goal is the estimation of a causal effect ⇒ first-differencing preferred because: weaker exogeneity assumption, more "intuitive" and easier to implement an IV estimation, but…

Panel Data and Unobserved Heterogeneity (4)

- Intuition behind the adjusted differencing procedure:
  - An individual is observed for ten years with data on union membership and job satisfaction
  - The individual is marked as union member only in the sixth year and the goal is to estimate the causal effect of the episode of membership by exploiting the counterfactual definition of causal effect:

$$\Delta \mathbf{attitude_i} = (\mathbf{attitude_i} \mid \mathbf{member_i}) - (\mathbf{attitude_i} \mid \mathbf{non-member_i})$$

  - Which one of the remaining years could be used to substitute the job satisfaction level in a counterfactual setting?

Panel Data and Unobserved Heterogeneity (5)

- Goal is to approximate the counterfactual definition of causal effect: for an individual becoming member, which reference point to use to represent the counterfactual setting in which he is a non-member? The most recent year in which he was non-member ⇒ model variations with respect to the most recent year in which the individual is observed in the non-treated status
- Adjusted differencing procedure:

$$\Delta_{adj.}x_{it} \equiv \begin{cases} x_{i,t} - x_{i,t-1} & \textit{if } union_{it-1} = 0 \textit{ and } union_{it} = 0 \\ x_{i,t} - x_{i,t-1} & \textit{if } union_{it-1} = 0 \textit{ and } union_{it} = 1 \\ x_{i,t} - x_{i,t-p} & \textit{if } union_{it-p} = 0, union_{it-(p-1)} = 1, \; union_{it-(p-2)} = 1, \ldots, union_{it-1} = 1 \text{ and } union_{it} = 1 \end{cases}$$

$$\text{for } i = 1, 2, \ldots, N \text{ and } t = 2, \ldots, T$$

⇒ first-difference observations representing years as non-member, first-difference observations representing first year of membership, second-difference observations representing second year of membership, third-difference observations representing third year of membership,… Do not consider transitions from “Member” to “Non-member” because of the likely durable effect of union membership

Panel Data and Unobserved Heterogeneity (6)

- Effect may vary with the duration of union membership and the inclusion of leads and/or lags is not always desired ⇒ estimate average effect of all durations of membership: first-differencing is not well suited, while fixed effects and adjusted differencing procedure are ok

year	union	attitude	Δ_{FD}union	Δ_{FD}attitude	$\Delta_{adj.}$union	$\Delta_{adj.}$attitude	Δ_{FE}unioin	Δ_{FE}attitude
1	0	0					-0,5	-0,75
2	0	0	0	0	0	0	-0,5	-0,75
3	1	1	1	1	1	1	0,5	0,25
4	1	2	0	1	1	2	0,5	1,25
OLS estimate	1.5		0.5		1.5		1.5	

Panel Data and Unobserved Heterogeneity (7)

- Get an estimator "centered" around a single variable ⇒ need to assume that the effect of the remaining control variables does not vary with duration
- Combine advantages of first-differencing and fixed effects:
  - Exogeneity requirement: $E[\Delta_{adj.}x_{it} \mid \Delta_{adj.}\mu_{it}] = 0$ for all i, t
  - Intuitive
  - No restrictions on type of instruments (levels of variables also allowed)
  - Get average effect of all years of membership
  - Ok with gaps

Time-Invariant and Time-Varying Endogeneity (1)

- Time-invariant endogeneity ⇒ time-varying endogeneity? Not necessarily, even when relevant time-varying omitted variables and/or time-varying reverse causality seem to constitute an issue
- Possible reasons:
  - Not much variation over time
  - The two types of variations may be “susbstantively” different
- Example: job attitudes vs. political attitudes

Time-Invariant and Time-Varying Endogeneity (2)

- If time-varying endogeneity represents a problem, need to apply an instrumental variable estimator on the transformed model:
  - Much easier to find valid instruments after having got rid off all time-invariant unobserved heterogeneity
  - Exogeneity assumption for instrument z:

$$\mathbf{cov}(\mathbf{z}, \Delta_{\mathbf{adj.}}\boldsymbol{\mu}) = \mathbf{0}$$

or

$$\mathbf{cov}(\Delta_{\mathbf{adj.}}\mathbf{attitude}, \mathbf{z} \mid \Delta_{\mathbf{adj.}}\mathbf{union}, \Delta_{\mathbf{adj.}}\mathbf{controls}) = \mathbf{0}$$

- Examples of instruments for job satisfaction (and interest in politics):
  - Consider validity, strength and “locality” of the instruments
  - Union density by canton and occupation (level form), number of associations an individual is member of (differenced form)

Applying It All… (1)

- Following the previous discussion, a strategy in four steps allowing to stepwise exclude the possible sources of exogeneity:
  - Pooled OLS without control variables: purely descriptive results
  - Pooled OLS with control variables: cross-sectional regression analysis
  - OLS on difference data: get rid of all potential sources of bias related to time-invariant unobserved heterogeneity between control and treatment group
  - 2SLS on differenced data: get rid of the bias related to time-varying endogeneity
- Heterogeneous effects bias: re-estimate the models on different segments of the treatment group
  - Variations across different sub-populations informative also regarding the underlying causal mechanism(s)
- Functional form and standard errors:
  - Linearity (OLS)
  - Cluster robust standard errors

Regression models estimating the effect of union membership on the satisfaction with working conditions

Source: Swiss Household Panel (SHP)

	Pooled OLS without control variables		Pooled OLS with control variables		OLS on differenced data with control variables		2SLS on differenced data with control variables	
Main effect								
Estimate – Standard error	-0.28***	(0.031)	-0.24***	(0.033)	-0.14***	(0.046)	0.15	(0.19)
Duration of membership								
First year								
Estimate – Standard error	-0.22***	(0.033)	-0.18***	(0.036)	-0.022	(0.042)	0.52+	(0.29)
Second year								
Estimate – Standard error	-0.33***	(0.048)	-0.28***	(0.052)	-0.21**	(0.077)	0.82	(0.79)
Third year or more								
Estimate – Standard error	-0.31***	(0.044)	-0.29***	(0.046)	-0.38***	(0.10)	0.13	(0.48)
Type of membership								
Active								
Estimate – Standard error	-0.29***	(0.046)	-0.26***	(0.049)	-0.17**	(0.064)	0.60	(0.57)
Passive								
Estimate – Standard error	-0.27***	(0.034)	-0.24***	(0.037)	-0.13*	(0.055)	0.17	(0.22)
Occupation								
Full-time								
Estimate – Standard error	-0.19***	(0.038)	-0.20***	(0.041)	-0.13*	(0.058)	0.11	(0.24)
Part-time								
Estimate – Standard error	-0.38***	(0.048)	-0.29***	(0.051)	-0.16*	(0.076)	0.35	(0.31)
Sex								
Man								
Estimate – Standard error	-0.21***	(0.041)	-0.20***	(0.044)	-0.21***	(0.062)	-0.33	(0.26)
Woman								
Estimate – Standard error	-0.33***	(0.047)	-0.26***	(0.048)	-0.064	(0.069)	0.60*	(0.28)
Sector								
Private								
Estimate – Standard error	-0.27***	(0.043)	-0.22***	(0.044)	-0.15*	(0.071)	-0.075	(0.35)
Public								
Estimate – Standard error	-0.29***	(0.044)	-0.27***	(0.046)	-0.15*	(0.061)	0.27	(0.21)

*Level of statistical significance : *** < 0.001, ** < 0.01, * < 0.05, + < 0.10*

Regression models estimating the effect of union membership on the interest in politics

Source: Swiss Household Panel (SHP)

	Pooled OLS without control variables		Pooled OLS with control variables		OLS on differenced data with control variables		2SLS on differenced data with control variables	
Main effect								
Estimate – Standard error	0.96***	(0.058)	0.48***	(0.057)	0.12*	(0.053)	0.2	(0.16)
Duration of membership								
First year								
Estimate – Standard error	0.66***	(0.053)	0.39***	(0.053)	0.056	(0.047)	0.049	(0.25)
Second year								
Estimate – Standard error	0.96***	(0.074)	0.45***	(0.077)	0.075	(0.084)	0.066	(0.47)
Third year or more								
Estimate – Standard error	1.22***	(0.081)	0.56***	(0.081)	0.30*	(0.13)	0.38	(0.34)
Type of membership								
Active								
Estimate – Standard error	1.16***	(0.083)	0.63***	(0.083)	0.14+	(0.082)	0.066	(0.31)
Passive								
Estimate – Standard error	0.85***	(0.063)	0.39***	(0.061)	0.11+	(0.058)	0.11	(0.21)
Occupation								
Full-time								
Estimate – Standard error	0.95***	(0.072)	0.53***	(0.071)	0.11	(0.068)	0.088	(0.23)
Part-time								
Estimate – Standard error	0.97***	(0.088)	0.48***	(0.086)	0.15+	(0.084)	0.32	(0.24)
Sex								
Man								
Estimate – Standard error	0.65***	(0.076)	0.39***	(0.076)	0.041	(0.074)	0.048	(0.24)
Woman								
Estimate – Standard error	1.11***	(0.087)	0.62***	(0.085)	0.21**	(0.075)	0.40+	(0.24)
Sector								
Private								
Estimate – Standard error	0.75***	(0.082)	0.46***	(0.079)	0.0086	(0.078)	-0.019	(0.046)
Public								
Estimate – Standard error	0.98***	(0.077)	0.51***	(0.074)	0.21**	(0.070)	0.25*	(0.11)

*Level of statistical significance : *** < 0.001, ** < 0.01, * < 0.05, + < 0.10*

Conclusions and Critical Regard

- Teasing out causal effects through observational data is a thorny enterprise
- Take into account the possible presence of a selection bias, but also of the potential heterogeneity of the effect
- Panel data offer a clear advantage when the dependent variable is affected by a large set of unobserved factors
- Different ways to get rid of the unobserved heterogeneity
- Think about the relationship between time-invariant and time-varying endogeneity on a substantive level
- Finding valid instruments without worrying about the correlation with the unobserved heterogeneity is much easier than in a cross-sectional setting

References

- Angrist, Joshua D. and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton: Princeton University Press.
- Hadziabdic, Sinisa. 2015. *Malleable Minds? Teasing Out the Causal Effect(s) of Union Membership on Job Attitudes and Political Outcomes*. Master's Thesis. University of Geneva
- Winship, Christopher and Stephen L. Morgan. 1999. "The Estimation of Causal Effects from Observational Data." *Annual Review of Sociology* 25:659–706.
- Wooldridge, Jeffrey M. 2010. *Econometric Analysis of Cross Section and Panel Data*. Cambridge, Massachusetts: The MIT Press.
- Wooldridge, Jeffrey M. 2013. *Introductory Econometrics*. Andover: Cengage Learning.

- Thank you for your attention
- Questions/remarks?