Path analysis for discrete variables: The role of education in social mobility

Jouni Kuha[1] John Goldthorpe[2]

[1]London School of Economics and Political Science

[2]Nuffield College, Oxford

ESRC Research Methods Festival, Oxford
8.7.2010

Methodology Institute LSE

Outline

- Example: Analysis of social mobility
- Reminder: Linear path analysis
- Path analysis for general variables: definition
- Estimation of the effects and their standard errors
- Interpretation of the effects in the path analysis
    - in causal terms
    - in non-causal terms
- Example: Analysis of UK mobility data

(For more, see Kuha, J. and Goldthorpe, J. (2010). Path analysis for discrete variables: The role of education in social mobility. *JRSS A* **173**, 351–369.)

Example: Intergenerational social mobility

- Five variables will be considered today:
- Social class:
  - Origin class (O): Person's *father's* class
  - Destination class (D): Person's *own* class

  ...classified using a 3-class version of the Goldthorpe class schema:
  - "Salariat" (S)
  - "Intermediate" (I)
  - "Working" (W)
- Education (E), with seven ordered levels
- + Analysis stratified by Sex and Period

Methodology Institute LSE

Today's data

- Data from the British General Household Survey (GHS), as used by Goldthorpe and Mills (2004; in Breen (ed.), *Social Mobility in Europe*)
- Consider separately men and women, from the 1973 and 1992 surveys
- Respondents aged 25–59
- Sample sizes:

	Men	Women
1973	6276	6882
1992	4835	5284

Distributions of D given O: Mobility tables

- Example: Women in the 1992 survey

	Destination		
Origin	Sal.	Int.	Work
Salariat	759	508	228
Intermediate	519	503	342
Working	558	893	974

Associations of O and D: Odds ratios

- For example, the 3 "diagonal" (log) odds ratios:

		D		
		S	I	W
	S	○	○	
O	I	○	○	
	W			

- E.g. "I–S" odds ratio calculated from frequencies in cells ○
- "W–I" and "W–S" associations similarly

Diagonal log odds ratios in the GHS data

	1973		1992	
log-OR	Men	Women	Men	Women
I-S	.87	.42	.95	.37
W-I	.74	.65	.74	.47
W-S	2.00	2.19	1.85	1.76

Methodology Institute LSE

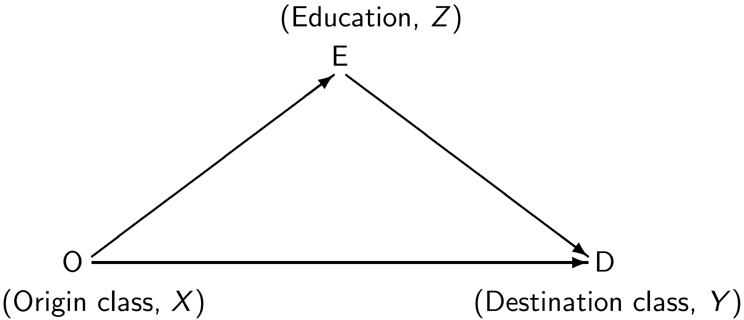
Path analysis of social mobility

- Association between O and D describes (lack of) social mobility between generations
- This is the "total effect" of O on D discussed below
- Try to partition the total effect into...
- Indirect effect $O \longrightarrow E \longrightarrow D$
  - $O \longrightarrow E$: Class inequalities in educational attainment (and opportunity?)
  - $E \longrightarrow D$: Dependence of class position on educational qualitifications
- Direct effect $O \longrightarrow D$ not via E
  - Class inequalities in social networks, living conditions, social capital?
- How to assess relative sizes of these?
  - In particular, is the indirect effect dominant, as has been claimed in UK?

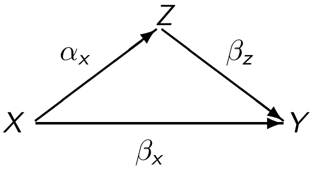
Path analysis of social mobility

In pictures:

...elaborated into...



Reminder: Linear path analysis



$$\begin{aligned} E(Y|X,Z) &= \beta_0 + \beta_x X + \beta_z Z \\ E(Z|X) &= \alpha_0 + \alpha_x X \end{aligned}$$

$$E(Y|X) = \int E(Y|X,Z)\, p(Z|X)\, dZ = \beta_0^* + \beta_x^* X$$

where

$$\beta_x^* = \beta_x + \beta_z \alpha_x$$

i.e.

$$\text{Total effect} = \text{Direct effect} + \text{Indirect effect}$$

Path analysis for discrete variables

- How to define and estimate direct and indirect effects when Z and/or Y are categorical variables, and modelled as such?
- Here, multinomial logistic models for both
  - Education given Origin (Z given X)
  - Destination given Origin and Education (Y given X and Z)

(Re)defining the effects for non-linear models

- Let Y_l be an indicator for $Y = l$
  - Thus $\mathrm{E}(Y_l) = P(Y = l)$
- Consider (any) two values X_1 and X_2 of X
- The total effect of X on Y is described in terms of comparisons of

$$\mathrm{E}(Y_l|X_j) = \int \mathrm{E}(Y|X_j, Z)\, p(Z|X_j)\, dZ$$

e.g. a mean difference $\mathrm{E}(Y_l|X_2) - \mathrm{E}(Y_l|X_1)$ or a log-OR

$$\log\left[\frac{\mathrm{E}(Y_m|X_2)}{\mathrm{E}(Y_l|X_2)}\right] - \log\left[\frac{\mathrm{E}(Y_m|X_1)}{\mathrm{E}(Y_l|X_1)}\right]$$

(Re)defining the effects for non-linear models

- For a direct effect, define

$$\mathrm{E}^{D}_{(12)}(Y_l|X_j) = \int \mathrm{E}(Y_l|X_j, Z)\, p_{(12)}(Z)\, dZ$$

where

$$p_{(12)}(Z) = \frac{p(Z|X_1) + p(Z|X_2)}{2}$$

and compare

$$\mathrm{E}^{D}_{(12)}(Y_l|X_1) \quad \text{vs.} \quad \mathrm{E}^{D}_{(12)}(Y_l|X_2)$$

(Re)defining the effects for non-linear models

- For an indirect effect, define

$$E^{I}_{(12)}(Y_l|X_j) = \int E_{(12)}(Y_l|Z)\, p(Z|X_j)\, dZ$$

where

$$E_{(12)}(Y_l|Z) = \frac{E(Y_l|X_1, Z) + E(Y_l|X_2, Z)}{2}$$

and compare

$$E^{I}_{(12)}(Y_l|X_1) \quad \text{vs.} \quad E^{I}_{(12)}(Y_l|X_2)$$

Decompositions of total effects

- These quantities provide an exact partitioning of a total mean difference:

$$\begin{aligned} \mathrm{E}(Y_l|X_2) - \mathrm{E}(Y_l|X_1) &= [\mathrm{E}^D_{(12)}(Y_l|X_2) - \mathrm{E}^D_{(12)}(Y_l|X_1)] \\ &\quad + [\mathrm{E}^I_{(12)}(Y_l|X_2) - \mathrm{E}^I_{(12)}(Y_l|X_1)] \end{aligned}$$

- For log odds ratios, corresponding additive decomposition is approximate but typically quite accurate

Calculating the estimated effects

- First, need to specify models for $\mathrm{E}(Y|X,Z)$ and $p(Z|X)$
  - Estimates of these are obtained in standard ways
- Second, the estimated effects are functions of estimates of $\mathrm{E}(Y|X,Z)$ and $p(Z|X)$
  - For example, when intermediate variable Z is discrete, this involves only summation, e.g.

$$\hat{\mathrm{E}}^{D}_{(12)}(Y_l|X_j) = \frac{1}{2}\sum_{k}\sum_{t=1,2}\hat{\mathrm{E}}(Y_l|X_j,Z_k)\,\hat{p}(Z_k|X_t)$$

- Third, standard errors of the estimated effects can be derived, ultimately from the standard errors of estimated parameters of $\mathrm{E}(Y|X,Z)$ and $p(Z|X)$

Methodology Institute LSE

Causal interpretations: Total effects

- Consider the counterfactual (potential outcomes) framework of formal causal inference
- Define potential outcomes (dropping subscript from Y):
  - $Y(x)$: value of Y for a single subject when X has value x
- *Total effect* of changing from $X = 1$ to $X = 2$ is defined in terms of comparisons of $Y(1)$ and $Y(2)$
- E.g. the mean difference (average treatment effect)

$$\mathrm{E}\{Y(2)\} - \mathrm{E}\{Y(1)\}$$

where expectation is over all subjects in a population

- analogously for odds ratios etc.

Causal interpretations: Direct and indirect effects

- Define potential outcomes $Z(x)$ and $Y(x,z)$ similarly
  - Total effect can be expressed as

$$\mathrm{E}\{Y[2, Z(2)]\} - \mathrm{E}\{Y[1, Z(1)]\}$$

- Natural direct effect of changing from $X = 1$ to $X = 2$ is

$$NDE(1 \rightarrow 2) = \mathrm{E}\{Y[2, Z(1)]\} - \mathrm{E}\{Y[1, Z(1)]\}$$

and natural indirect effect is defined as either

$$\begin{aligned} NIE(1 \rightarrow 2) &= \mathrm{E}\{Y[2, Z(2)]\} - \mathrm{E}\{Y[2, Z(1)]\} \quad \text{or} \\ NIE(1 \rightarrow 2) &= \mathrm{E}\{Y[1, Z(2)]\} - \mathrm{E}\{Y[1, Z(1)]\} \end{aligned}$$

e.g. Pearl (2001), Robins (2003), and [in a different framework] Geneletti (2007)

Causal interpretations: Direct and indirect effects

- Estimates of the effects/associations defined in terms of $\mathrm{E}(Y|X, Z)$ and $p(Z|X)$ above can be thought of as estimates of the following averages of natural effects:
  - For direct effect:

$$\frac{1}{2}\left[NDE(1 \rightarrow 2) + NDE(2 \rightarrow 1)\right]$$

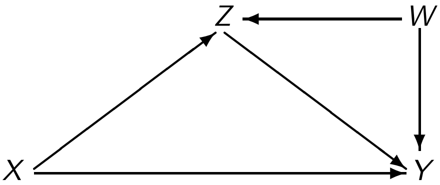
  - For indirect effect:

$$\frac{1}{2}\left[NIE(1 \rightarrow 2) + NIE(2 \rightarrow 1)\right]$$

- ... at least under some fairly strict assumptions...

Conditions for causal interpretation

- Essentially, there should be no unmeasured confounders (common causes) of the relationships of X, Z and Y
- Particularly problematic are confounders of the relationship of Z and Y:



  - Such confounders should be controlled for in the estimation

Interpretation as associations: Total effects

- A more cautious interpretation than a causal one
  - ...and most that we can claim in the mobility example
- Consider first two groups:

	Group 1	Group 2
Distribution of X	X_1 for all	X_2 for all
Distribution of Z	$p(Z\|X_1)$	$p(Z\|X_2)$

  - i.e. observations with $X = X_1$ and with $X = X_2$, exactly as observed
- $\text{E}(Y|X_1)$ and $\text{E}(Y|X_2)$ are average expected values of Y in these groups, when $\text{E}(Y|X,Z)$ is as observed
- The total association is a comparison of these expected values

Interpretation as associations: Direct and indirect effects

- The direct-effect association is what would be observed when comparing average expected values of Y between these two groups:

	Group 1	Group 2
Distribution of X	X_1 for all	X_2 for all
Distribution of Z	$[p(Z\|X_1) + p(Z\|X_2)]/2$	

  - i.e. groups which differ in X but have the same distribution of Z

- Indirect-effect association analogously, comparing groups which differ in $p(Z|X_j)$ but have the same (even) mixture of X_1 and X_2 in both

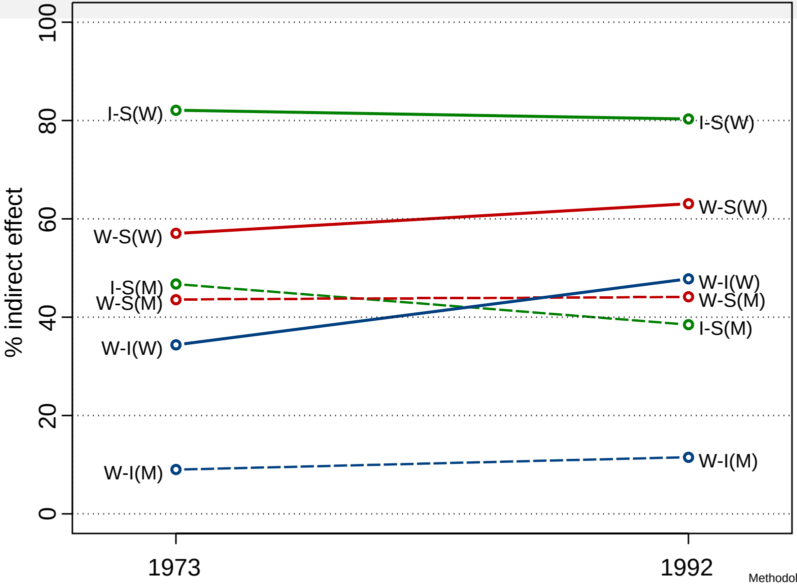
Mobility example: Women in 1992

- Estimated (symmetric) log-odds ratios: total, direct and indirect

	I–S	W–I	W–S
Observed total effect	.37 (.08)	.47 (.08)	1.76 (.09)
Direct + Indirect effect	.37 (.08)	.47 (.08)	1.72 (.07)
Direct effect	.07 (.08)	.25 (.08)	.63 (.08)
Indirect effect	.30 (.03)	.22 (.02)	1.08 (.03)
% Indirect effect	80* (18)	48 (9)	63 (7)

* Consistent with 100% indirect effect.

% of indirect (education) effect of total log-OR



Methodology Institute LSE

Future work

- Application to more recent British mobility data (1946, 1958 and 1970 birth cohort studies)
- Analysis with more detailed class classification
- Extensions to cases with more intervening variables

References

Kuha, J. and Goldthorpe, J. H. (2010). Path analysis for discrete variables: The role of education in social mobility. *JRSS A*, 173, 351–369.

Geneletti, S. (2007). Identifying direct and indirect effects in a non-counterfactual framework. *JRSS B*, 69, 199–215.

Goldthorpe, J. H. and Mills, C. (2004) Trends in intergenerational class mobility in Britain in the late twentieth century. In *Social Mobility in Europe* (Ed. R. Breen), pp. 195–224. OUP.

Pearl, J. (2001). Direct and indirect effects. In *Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence*, pp. 411–420. Morgan Kaufmann.

Robins, J. M. (2003). Semantics of causal DAG models and the identification of direct and indirect effects. In *Highly Structured Stochastic Systems* (Eds. P. Green, N. Hjort and S. Richardson), pp. 70–81. OUP.