

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

Jouni Kuha

Methodology Institute & Dept. of Statistics

London School of Economics

and Political Science

John Goldthorpe

Nuffield College

University of Oxford

“Causality in Statistical Investigations”

Royal Statistical Society

27 May 2009



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**

Example: Intergenerational social mobility

- Data from General Household Survey (GHS), as used by Goldthorpe and Mills (2004)
- Consider separately men and women, and the 1973 and 1992 surveys
- Respondents aged 25–59
- Sample sizes: 4835–6882

Distributions of D given O : Mobility tables

- Example: Women in the 1992 survey

Origin	Destination		
	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
O	S	○	○	
	I	○	○	
	W			

- E.g. “I–S” odds ratio calculated from frequencies in cells marked with ○
- “W–I” and “W–S” associations similarly

Diagonal log odds ratios in the GHS data

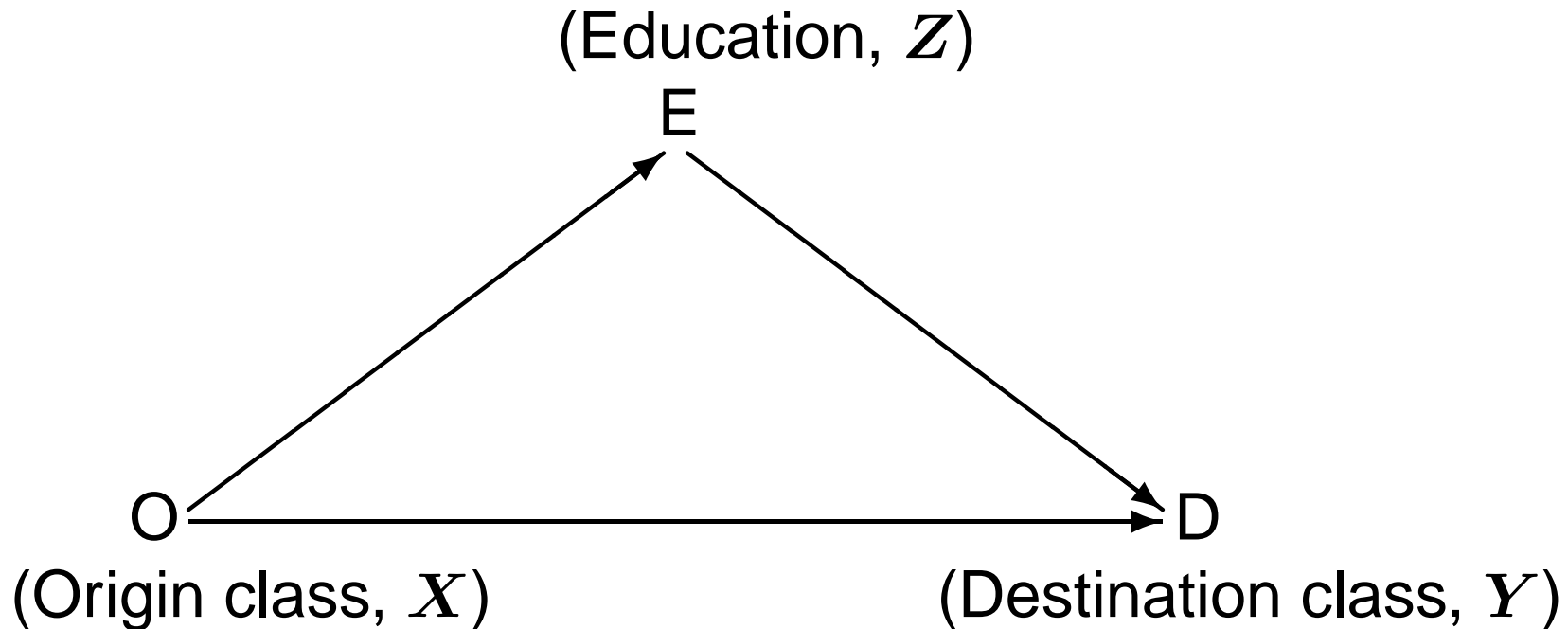
log-OR	1973		1992	
	Men	Women	Men	Women
I-S	.87	.42	.95	.37
W-I	.74	.65	.74	.47
W-S	2.00	2.19	1.85	1.76

- These associations between O and D describe (lack of) social mobility between generations
- This is the “total association” or “total effect” of O on D discussed below

Path analysis of social mobility



...elaborated into...



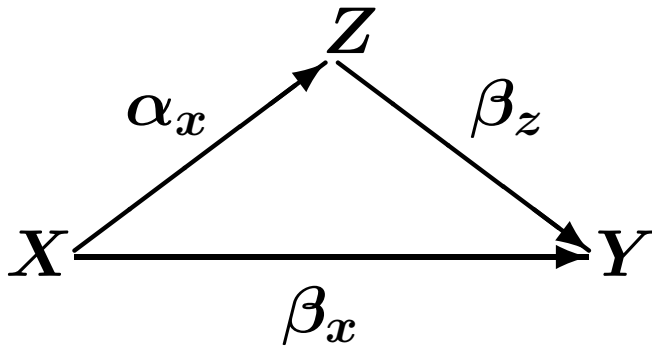
Path analysis of social mobility

- Aim: to define and estimate strengths of...
- **Indirect effect** (association) $O \longrightarrow E \longrightarrow D$
 - $O \longrightarrow E$: Class inequalities in educational attainment (and opportunity?)
 - $E \longrightarrow D$: Dependence of class position on educational qualifications
- **Direct effect** (association) $O \longrightarrow D$, not via E
 - Class inequalities in social networks, social capital?

Path analysis of social mobility

- How to assess magnitudes and relative sizes of these?
 - In particular, is the indirect effect dominant, as has been claimed in UK?
- How to do this when E and/or D are categorical variables, and modelled as such?
- Here, **multinomial logistic models** for both
 - Education given Origin
(saturated, so just a two-way cross-tabulation)
 - Destination given Origin and Education

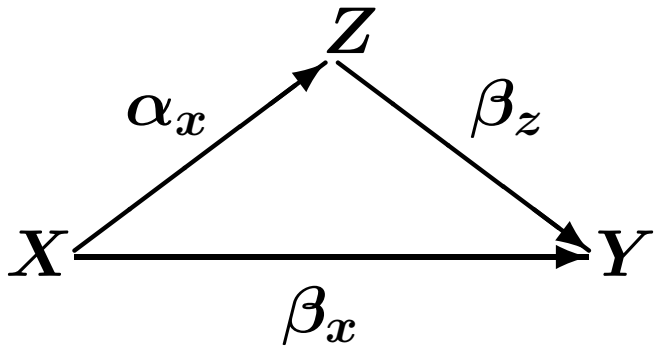
Reminder: Linear path analysis



$$E(Y|X, Z) = \beta_0 + \beta_x X + \beta_z Z$$

$$E(Z|X) = \alpha_0 + \alpha_x X$$

Reminder: Linear path analysis

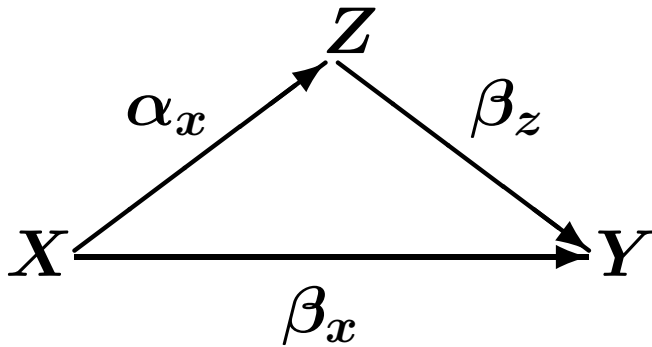


$$E(Y|X, Z) = \beta_0 + \beta_x X + \beta_z Z$$

$$E(Z|X) = \alpha_0 + \alpha_x X$$

$$E(Y|X) = \int E(Y|X, Z) p(Z|X) dZ$$

Reminder: Linear path analysis

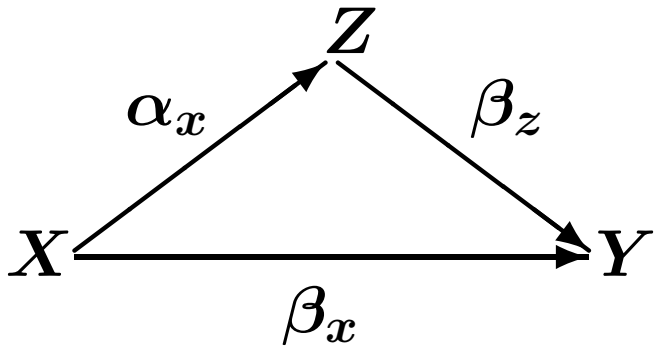


$$E(Y|X, Z) = \beta_0 + \beta_x X + \beta_z Z$$

$$E(Z|X) = \alpha_0 + \alpha_x X$$

$$\begin{aligned} E(Y|X) &= \int E(Y|X, Z) p(Z|X) dZ \\ &= \beta_0^* + \beta_x^* X \end{aligned}$$

Reminder: Linear path analysis



$$E(Y|X, Z) = \beta_0 + \beta_x X + \beta_z Z$$

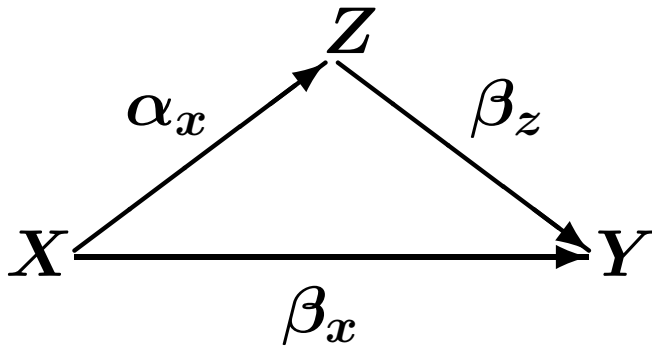
$$E(Z|X) = \alpha_0 + \alpha_x X$$

$$\begin{aligned} E(Y|X) &= \int E(Y|X, Z) p(Z|X) dZ \\ &= \beta_0^* + \beta_x^* X \end{aligned}$$

where

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

Reminder: Linear path analysis



$$E(Y|X, Z) = \beta_0 + \beta_x X + \beta_z Z$$

$$E(Z|X) = \alpha_0 + \alpha_x X$$

$$\begin{aligned} E(Y|X) &= \int E(Y|X, Z) p(Z|X) dZ \\ &= \beta_0^* + \beta_x^* X \end{aligned}$$

where

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

i.e.

Total effect = **Direct effect** + **Indirect effect**

(Re)defining the effects

- Let Y_l be an indicator for $Y = l$
 - Thus $E(Y_l) = P(Y = l)$
- 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

$$E(Y_l|X_j) = \int E(Y_l|X_j, Z) p(Z|X_j) dZ$$

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

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

- For a *direct effect*, define

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

where

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

and compare

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

- For an *indirect effect*, define

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

where

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

and compare

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

Decompositions of total effects

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

$$\begin{aligned} E(Y_l|X_2) - E(Y_l|X_1) &= [E_{(12)}^D(Y_l|X_2) - E_{(12)}^D(Y_l|X_1)] \\ &\quad + [E_{(12)}^I(Y_l|X_2) - E_{(12)}^I(Y_l|X_1)] \end{aligned}$$

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

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)

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

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

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

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

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

and **natural indirect effect** is defined as either

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

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

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

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

- For direct effect:

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

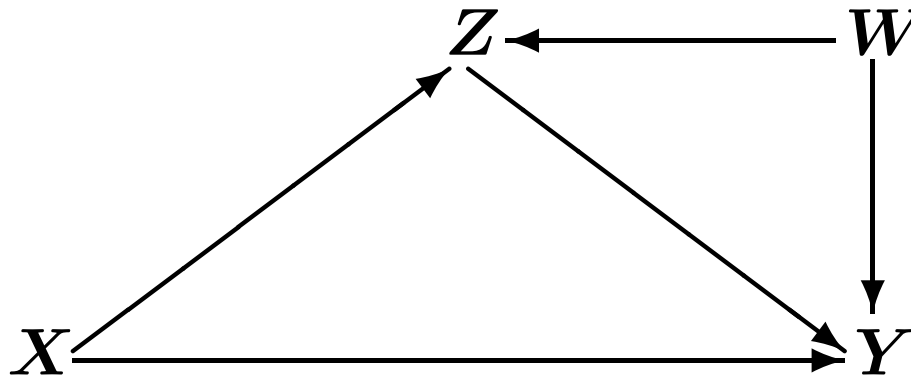
- For indirect effect:

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

- ... 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

- A more cautious interpretation than a causal one
- 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
- $E(Y|X_1)$ and $E(Y|X_2)$ are average expected values of Y in these groups, when $E(Y|X, Z)$ is as observed
- The **total** association is a comparison of these expected values

- 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

Calculating the estimated effects

- First, need to specify models for $p(Z|X)$ and $E(Y|X, Z)$
- Here:

- Saturated model for

$$P(Z = k|X = j) = \pi_{jk}$$

- i.e. Education given Origin
- Multinomial logistic model for

$$E(Y_l|X, Z) = P(Y = l|X, Z)$$

- i.e. Destination given Origin and Education
- Estimates and their standard errors for these are obtained in standard ways

- Second, estimated effects are functions of the fitted values $\hat{E}(Y|X, Z)$ and $\hat{p}(Z|X)$
- When intermediate variable Z is discrete, this involves only summation, e.g.

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

- When Z is continuous, simple Monte Carlo integration can be used to approximate the integral over its distribution

Standard errors of the estimated effects

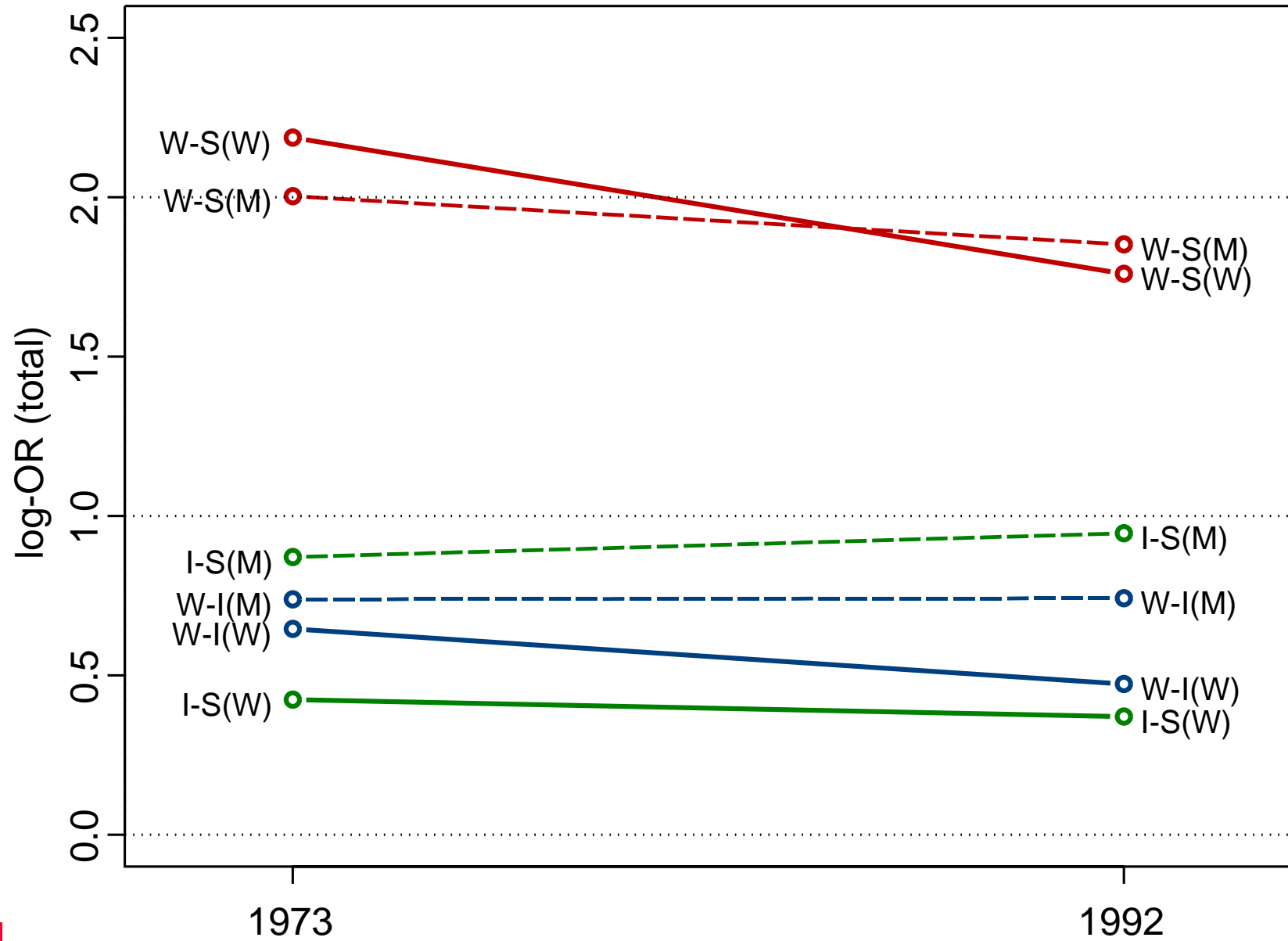
- All of the effect quantities are functions of the parameters of the models $E(Y_l|X, Z)$ and $p(Z|X)$
- Standard estimates of the (asymptotic) variances of these *parameters* are easily available
- Standard errors of the estimated *effects* are obtained through standard statistical techniques
 - i.e. repeated application of the delta method

Mobility example: Women in 1992

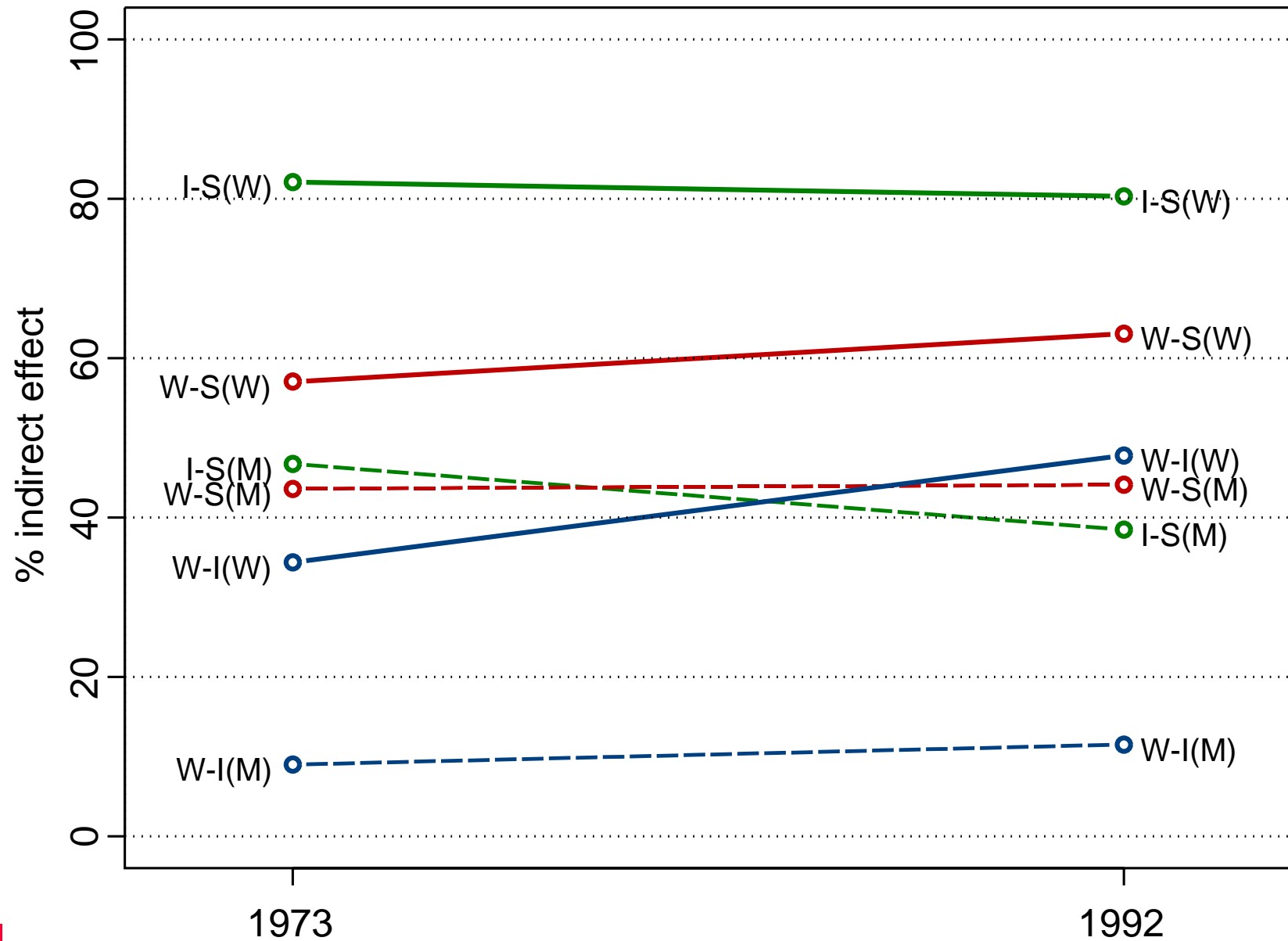
	I-S	W-I	W-S
Observed total effect	.37 (.08)	.47 (.08)	1.76 (.09)
Direct + Indirect effect	.37 (.08)	.47 (.08)	1.72 (.07)
Direct effect	.07 (.08)	.25 (.08)	.63 (.08)
Indirect effect	.30 (.03)	.22 (.02)	1.08 (.03)
% Indirect effect	80* (18)	48 (9)	63 (7)

* Consistent with 100% indirect effect.

Total log-ORs



% of indirect (education) association



References

Kuha, J. and Goldthorpe, J. H. (2009). Path analysis for discrete variables: The role of education in social mobility. (Under revision for *JRSS A*)

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.