Stata: a short history
viewed through epidemiology

Bianca L De Stavola

UCL Great Ormond Street Institute of Child Health

b.destavola@ucl.ac.uk

UK Stata Conference, 12-13 September 2024

A personal perspective

- This talk is a personal reflection on 35+ years of applied research in epidemiology
- Aims:
  - Pay tribute to influential contributors
  - Share some highlights
  - Offer reflections

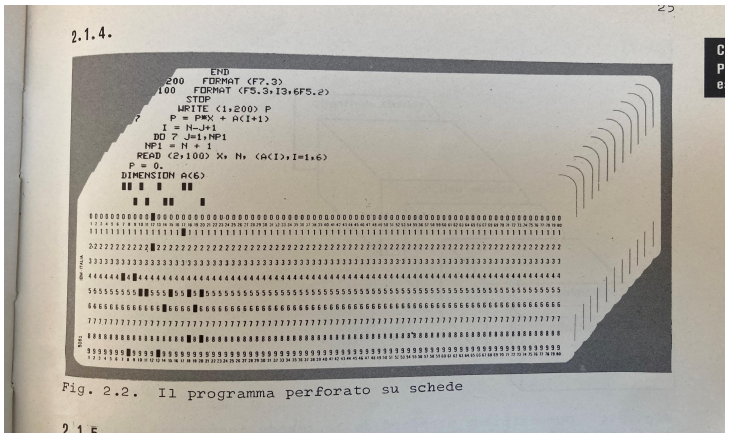
Overview

Life before Stata

Life before Stata

Fortran and the first statistical software

```
      DIMENSION KA(10,10), KB(10,10),KC(10,10)
      EQUIVALENCE (KB(1,1),KC(1,1))
      DEFINE FILE 1(10,10,U,N), 2(10,10,U,L)
      DO 10 I=1,10
      READ (2,100) (KA(I,M),M=1,10)
10    WRITE (1'I) (KA(I,M),M=1,10)
      DO 20 I=1,10
20    READ (1'I)  (KA(I,M),M=1,10)
      DO 30 I=1,10
      DO 30 J=1,10
30    KC(I,J) = KA(J,I)
      DO 40 I=1,10
40    WRITE (2'I) (KB(I,M),M=1,10)
      DO 50 I=1,10
      READ (1'I) (KA(I,M),M=1,10)
      READ (2'I) (KB(I,M),M=1,10)
50    WRITE (1,200) (KA(I,M),M=1,10), (KB(I,M),M=1,10)
      STOP
100   FORMAT (10I3)
200   FORMAT (10X,10I5,10X,10I5)
      END
```



2.1.4.

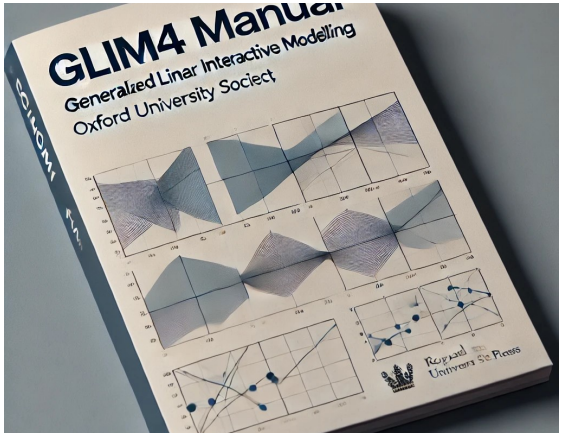
```
      END
  200 FORMAT (F7.3)
  100 FORMAT (F5.3,I3,6F5.2)
      STOP
      WRITE (1,200) P
    7 P = P*X + A(I+1)
      I = N-J+1
      DO 7 J=1,NP1
      NP1 = N + 1
      READ (2,100) X, N, (A(I),I=1,6)
      P = 0.
      DIMENSION A(6)
```

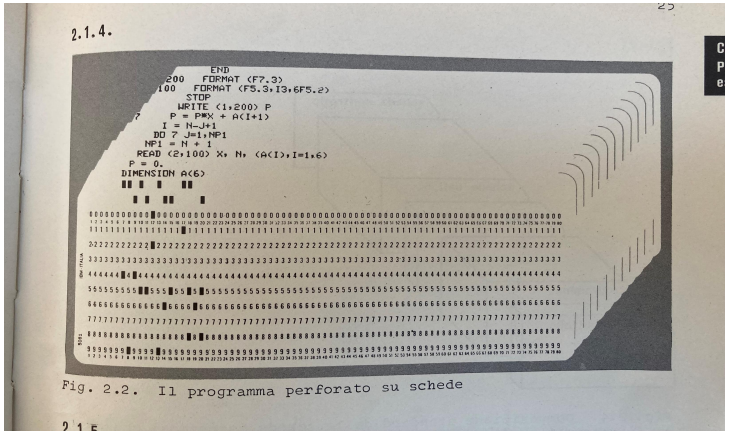
Fig. 2.2. Il programma perforato su schede

Life before Stata

Fortran and the first statistical software

```
      DIMENSION KA(10,10), KB(10,10),KC(10,10)
      EQUIVALENCE (KB(1,1),KC(1,1))
      DEFINE FILE 1(10,10,U,N), 2(10,10,U,L)
      DO 10 I=1,10
      READ (2,100) (KA(I,M),M=1,10)
10    WRITE (1'I) (KA(I,M),M=1,10)
      DO 20 I=1,10
20    READ (1'I)  (KA(I,M),M=1,10)
      DO 30 I=1,10
      DO 30 J=1,10
30    KC(I,J) = KA(J,I)
      DO 40 I=1,10
40    WRITE (2'I) (KB(I,M),M=1,10)
      DO 50 I=1,10
      READ (1'I) (KA(I,M),M=1,10)
      READ (2'I) (KB(I,M),M=1,10)
50    WRITE (1,200) (KA(I,M),M=1,10), (KB(I,M),M=1,10)
      STOP
100   FORMAT (10I3)
200   FORMAT (10X,10I5,10X,10I5)
      END
```




2.1.4.

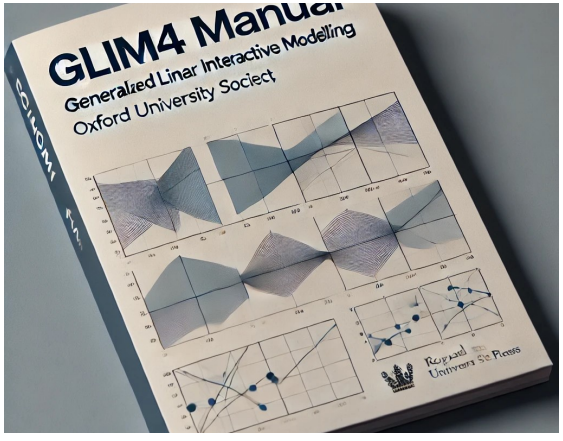
```
      END
  200 FORMAT (F7.3)
  100 FORMAT (F5.3,I3,6F5.2)
      STOP
      WRITE (1,200) P
    7 P = P*X + A(I+1)
      I = N-J+1
      DO 7 J=1,NP1
      NP1 = N + 1
      READ (2,100) X, N, (A(I),I=1,6)
      P = 0.
      DIMENSION A(6)
```

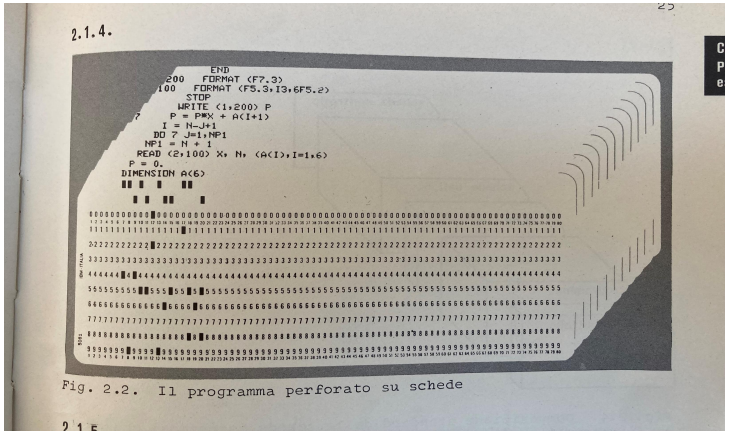
Fig. 2.2. Il programma perforato su schede

Life before Stata

Fortran and the first statistical software

```
      DIMENSION KA(10,10), KB(10,10),KC(10,10)
      EQUIVALENCE (KB(1,1),KC(1,1))
      DEFINE FILE 1(10,10,U,N), 2(10,10,U,L)
      DO 10 I=1,10
      READ (2,100) (KA(I,M),M=1,10)
10    WRITE (1'I) (KA(I,M),M=1,10)
      DO 20 I=1,10
20    READ (1'I)  (KA(I,M),M=1,10)
      DO 30 I=1,10
      DO 30 J=1,10
30    KC(I,J) = KA(J,I)
      DO 40 I=1,10
40    WRITE (2'I) (KB(I,M),M=1,10)
      DO 50 I=1,10
      READ (1'I) (KA(I,M),M=1,10)
      READ (2'I) (KB(I,M),M=1,10)
50    WRITE (1,200) (KA(I,M),M=1,10), (KB(I,M),M=1,10)
      STOP
100   FORMAT (10I3)
200   FORMAT (10X,10I5,10X,10I5)
      END
```





2.1.4.

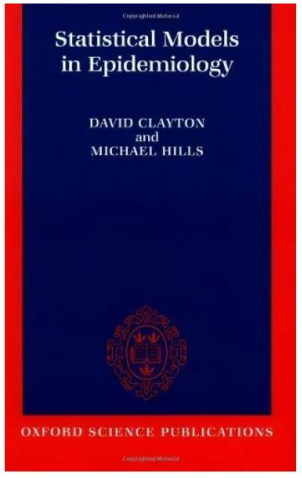
```
      END
  200 FORMAT (F7.3)
  100 FORMAT (F5.3,I3,6F5.2)
      STOP
      WRITE (1,200) P
    P = P*X + A(I+1)
  I = N-J+1
  DO 7 J=1,NP1
  NP1 = N + 1
  READ (2,100) X, N, (A(I),I=1,6)
  P = 0.
  DIMENSION A(6)
```

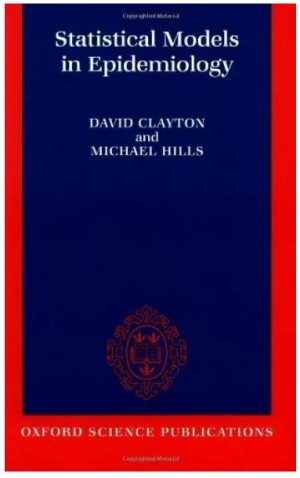
Fig. 2.2. Il programma perforato su schede

- Michael Hills and David Clayton

The 1990s

Michael Hills and David Clayton





Acknowledgments

The original version of `strate` was written by David Clayton (retired) of the Cambridge Institute for Medical Research and Michael Hills (1934–2021) of the London School of Hygiene and Tropical Medicine.

Acknowledgments

`stsplit` and `stjoin` are extensions of `lexis` by David Clayton (retired) of the Cambridge Institute for Medical Research and Michael Hills (1934–2021) of the London School of Hygiene and Tropical Medicine (Clayton and Hills 1995). The original `stsplit` and `stjoin` commands were written by Jeroen Weesie of the Department of Sociology at Utrecht University, The Netherlands (Weesie 1998a, 1998b), as was the revised `stsplit` command.

Acknowledgments

We thank David Clayton (retired) of the Cambridge Institute for Medical Research and Michael Hills (1934–2021) of the London School of Hygiene and Tropical Medicine, who wrote the original versions of `mhodds` and `tabodds`.

- London School of Hygiene and Tropical Medicine
- European Education Program in Epidemiology in Florence

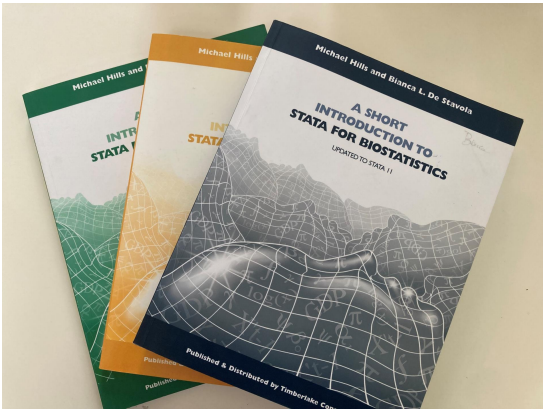
Michael in Florence

Ana Timberlake

- London School of Hygiene and Tropical Medicine
- European Education Program in Epidemiology in Florence

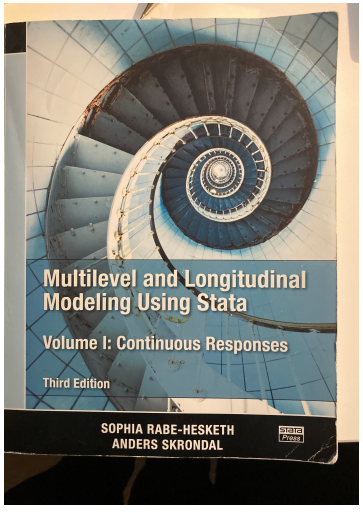
The Stata manuals . . .

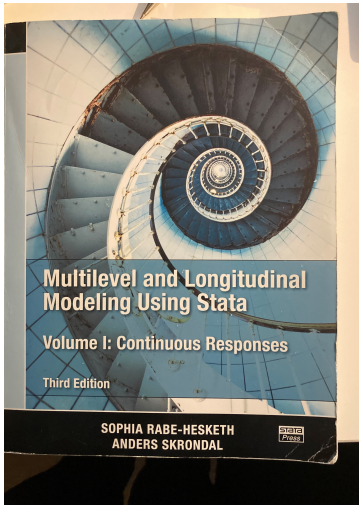
Michael's version!



The 2000s

- Mixed effects models
- Missing data


Multilevel and Longitudinal Modeling Using Stata
Volume I: Continuous Responses
Third Edition
SOPHIA RABE-HESKETH
ANDERS SKRONDAL
STATA Press

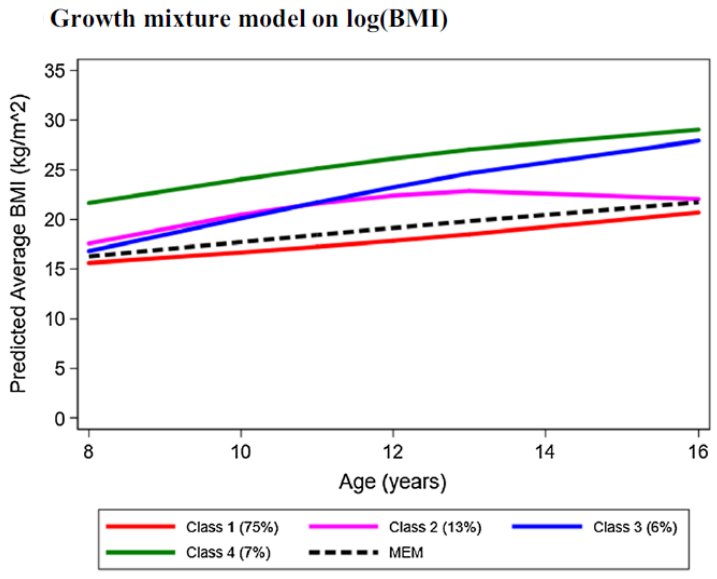


gllamm — Generalized linear and latent mixed models

Description Remarks and examples References Also see

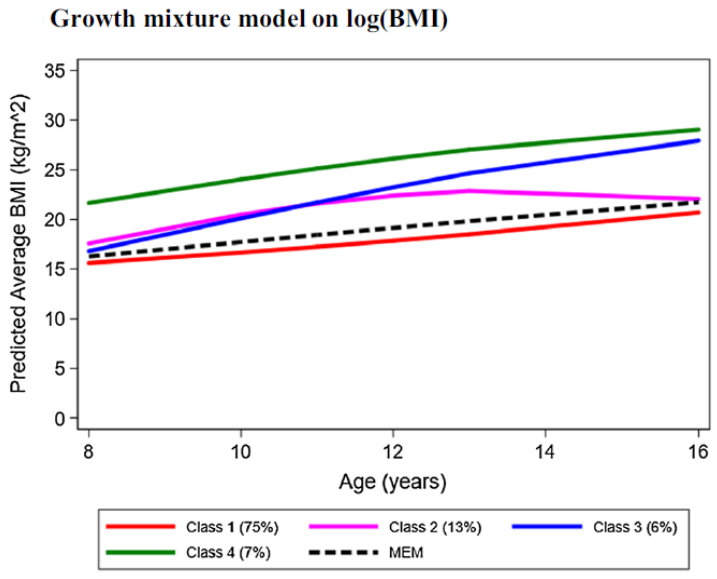
Description

GLLAMM stands for generalized linear latent and mixed models, and `gllamm` is a Stata command for fitting such models written by Sophia Rabe-Hesketh (University of California–Berkeley) as part of joint work with Anders Skrondal (Norwegian Institute of Public Health) and Andrew Pickles (King's College London).



Using *mixed* and *gllamm*

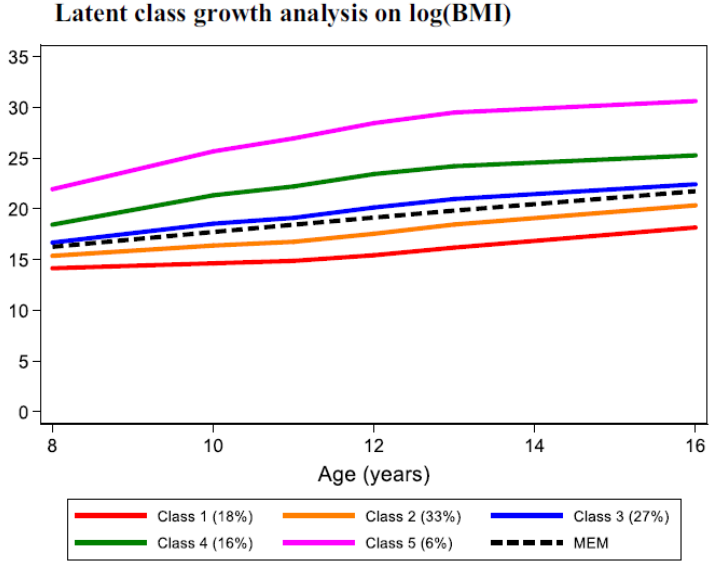
[Herle *et al.* EJE 2021]



Using *mixed* and *gllamm*

[Herle *et al.* EJE 2021]

Using *mixed* and *traj* (Jones and Nagin, 2013)



- Increasing awareness of bias from ignoring missing data bias
- Rubin's Multiple Imputation approach and van Buuren's Multiple Imputation by Chained Equations were starting to gain traction

- Increasing awareness of bias from ignoring missing data bias
- Rubin's Multiple Imputation approach and van Buuren's Multiple Imputation by Chained Equations were starting to gain traction

ice, mim

The Stata Journal (2008)
8, Number 1, pp. 49–67

A new framework for managing and analyzing multiply imputed data in Stata

John B. Carlin
Clinical Epidemiology & Biostatistics Unit
Murdoch Children's Research Institute &
University of Melbourne
Parkville, Australia
john.carlin@mcri.edu.au

John C. Galati
Clinical Epidemiology & Biostatistics Unit
Murdoch Children's Research Institute &
University of Melbourne
Parkville, Australia

Patrick Royston
Cancer and Statistical Methodology Groups
MRC Clinical Trials Unit
London, UK

- Causal inference

- The currently dominant approach in biostatistics and epidemiology relies on potential outcomes (POs) [Rubin, 1974; Robins, 1986; Pearl, 1995]
- Adopting this approach, we are concerned with questions formulated as contrasts of outcomes that would occur under hypothetical interventions on the exposure:

"Would the outcome of an individual differ if they had/not had that exposure?"

- Robins proposed solutions for estimation of POs*:
  (a) inverse probability weighting (IPW) (of marginal structural models)
  (b) the g-computation formula
  (c) g-estimation (of structural nested models)
- `teffects` implements (a) and (b) for time-fixed exposures

*Under assumptions of: no interference & consistency (i.e. SUTVA) and conditional exchangeability

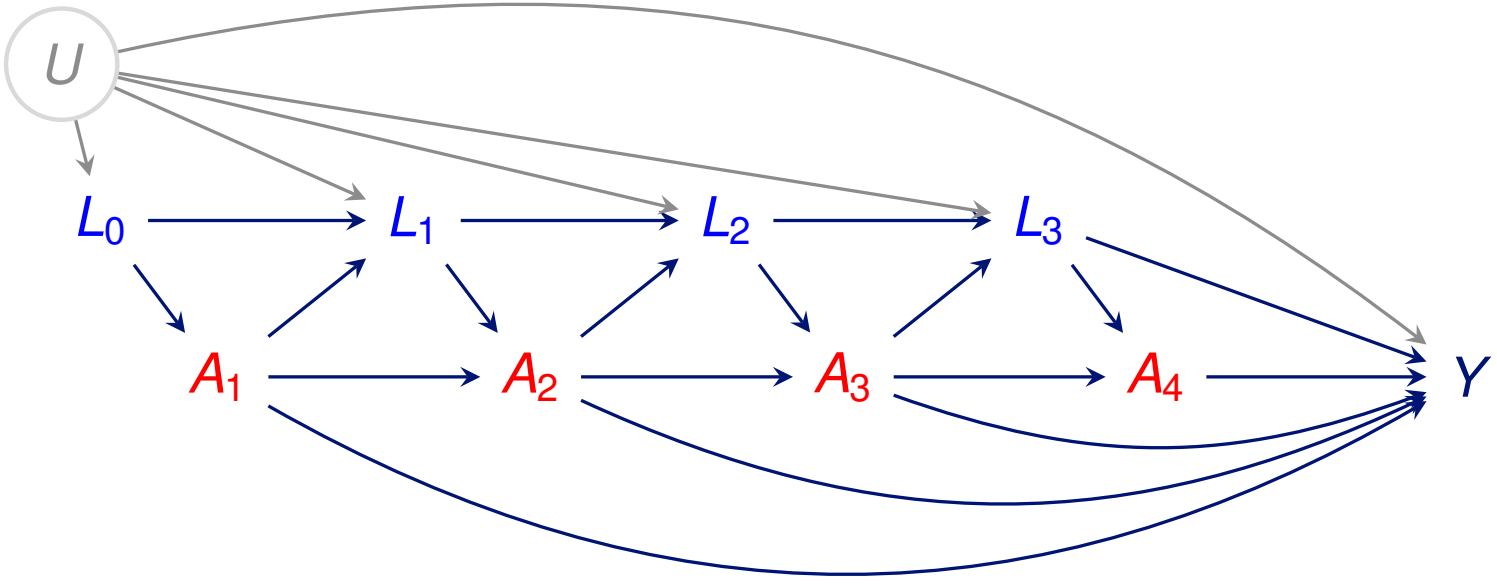
- The currently dominant approach in biostatistics and epidemiology relies on potential outcomes (POs) [Rubin, 1974; Robins, 1986; Pearl, 1995]
- Adopting this approach, we are concerned with questions formulated as contrasts of outcomes that would occur under hypothetical interventions on the exposure:
"Would the outcome of an individual differ if they had/not had that exposure?"
- Robins proposed solutions for estimation of POs*:
  (a) inverse probability weighting (IPW) (of marginal structural models)
  (b) the g-computation formula
  (c) g-estimation (of structural nested models)
- `teffects` implements (a) and (b) for time-fixed exposures

*Under assumptions of: no interference & consistency (i.e. SUTVA) and conditional exchangeability

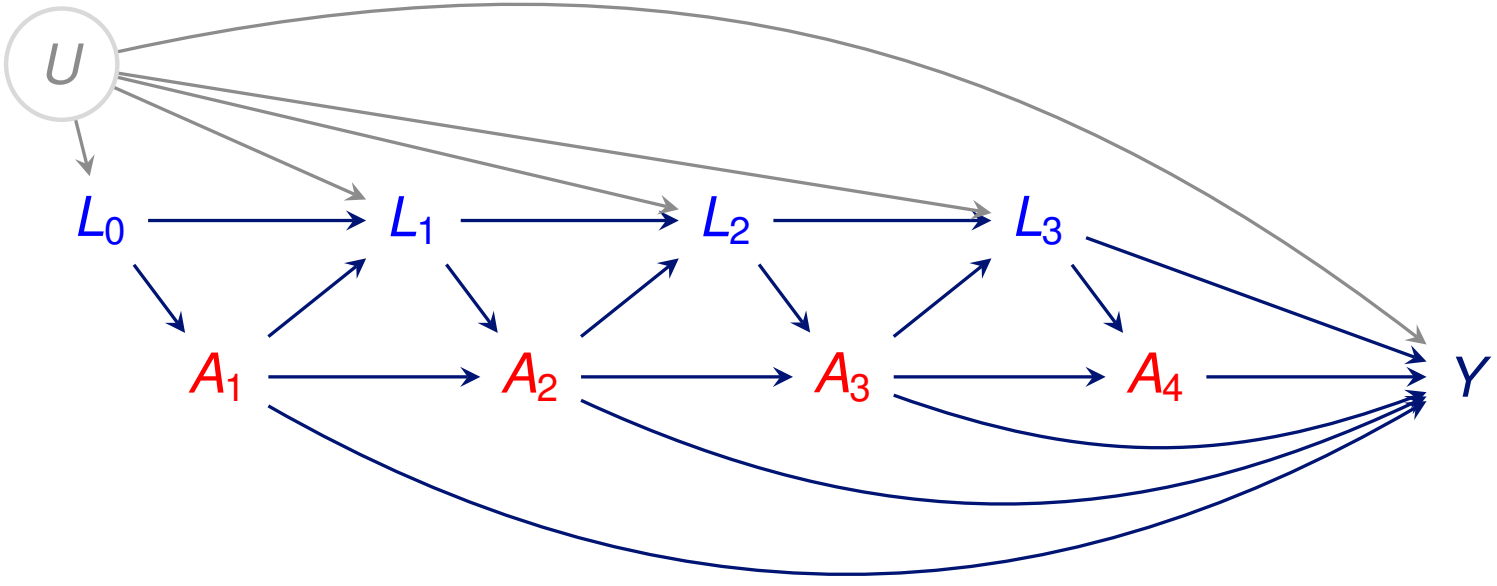
- The currently dominant approach in biostatistics and epidemiology relies on potential outcomes (POs) [Rubin, 1974; Robins, 1986; Pearl, 1995]
- Adopting this approach, we are concerned with questions formulated as contrasts of outcomes that would occur under hypothetical interventions on the exposure:
"Would the outcome of an individual differ if they had/not had that exposure?"
- Robins proposed solutions for estimation of POs*:
  (a) inverse probability weighting (IPW) (of marginal structural models)
  (b) the g-computation formula
  (c) g-estimation (of structural nested models)
- `teffects` implements (a) and (b) for time-fixed exposures

*Under assumptions of: no interference & consistency (i.e. SUTVA) and conditional exchangeability

We often deal with scenarios with time-varying confounding of the effect of a time-varying exposure A by a time-varying confounder L:



We often deal with scenarios with time-varying confounding of the effect of a time-varying exposure A by a time-varying confounder L:



Here the total causal effect of A involves L_1, L_2, L_3, although these are also confounders for A_2, A_3,A_4: standard regression modelling does not work!

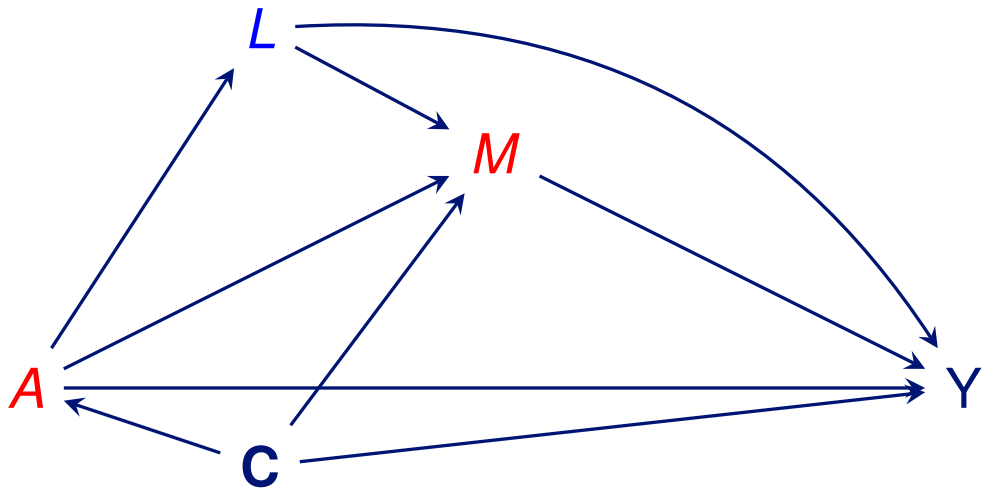
The Stata Journal (2011)
11, Number 4, pp. 479–517

gformula: Estimating causal effects in the presence of time-varying confounding or mediation using the g-computation formula

Rhian M. Daniel
Centre for Statistical Methodology
London School of Hygiene and Tropical Medicine
London, UK
rhian.daniel@lshtm.ac.uk

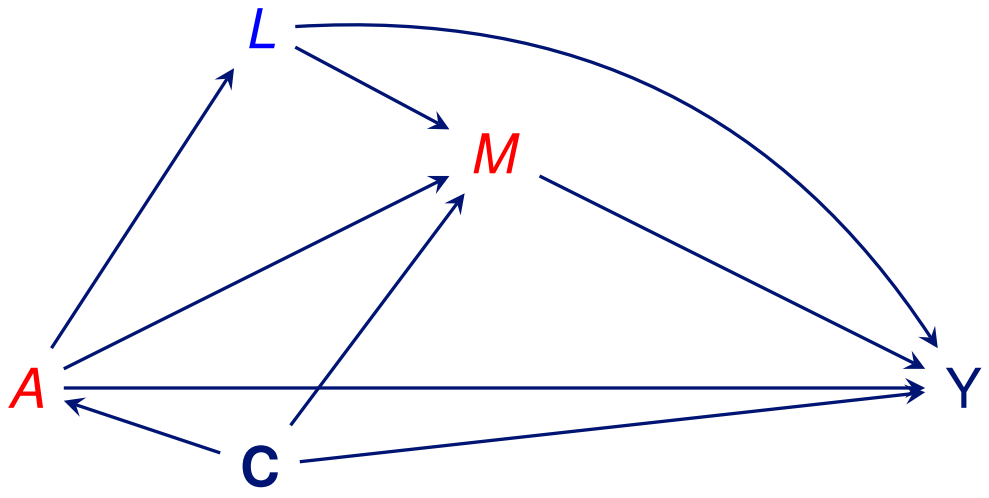
Bianca L. De Stavola
Centre for Statistical Methodology
London School of Hygiene and Tropical Medicine
London, UK

Simon N. Cousens
Centre for Statistical Methodology
London School of Hygiene and Tropical Medicine
London, UK



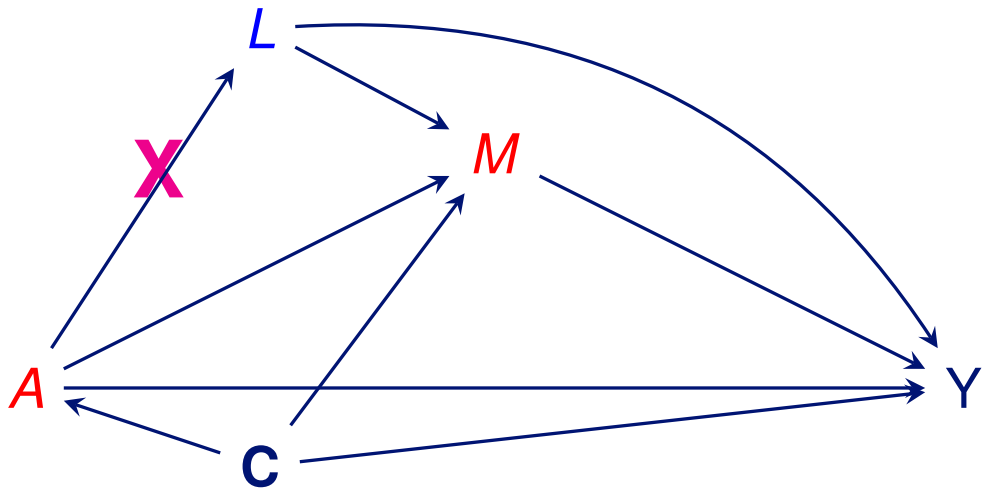
- `gformula` can be used to estimate natural and interventional effects
- `medeff` (Hicks and Tingley, 2011) and `paramed` (Emsley and Liu, 2013)† can only be used when *L* is not an intermediate confounder

†Now incorporated in version 18



- `gformula` can be used to estimate natural and interventional effects
- `medeff` (Hicks and Tingley, 2011) and `paramed` (Emsley and Liu, 2013)† can only be used when *L* is not an intermediate confounder

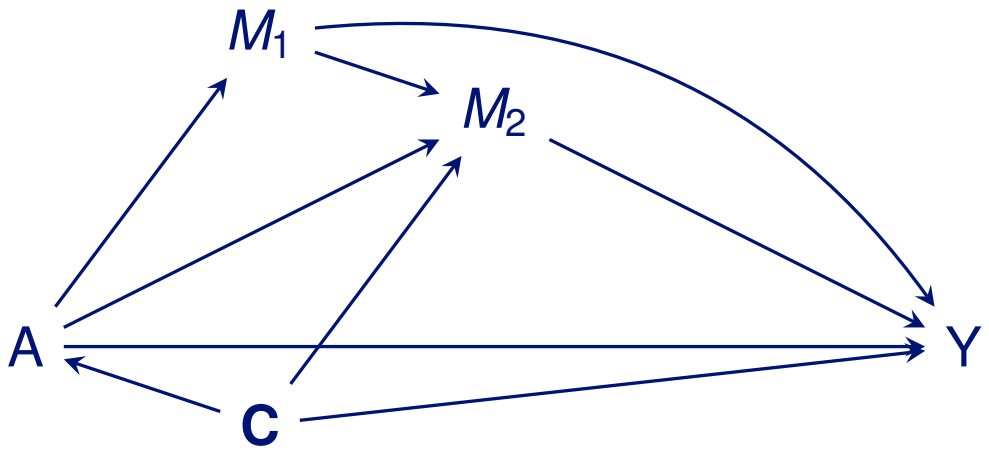
†Now incorporated in version 18



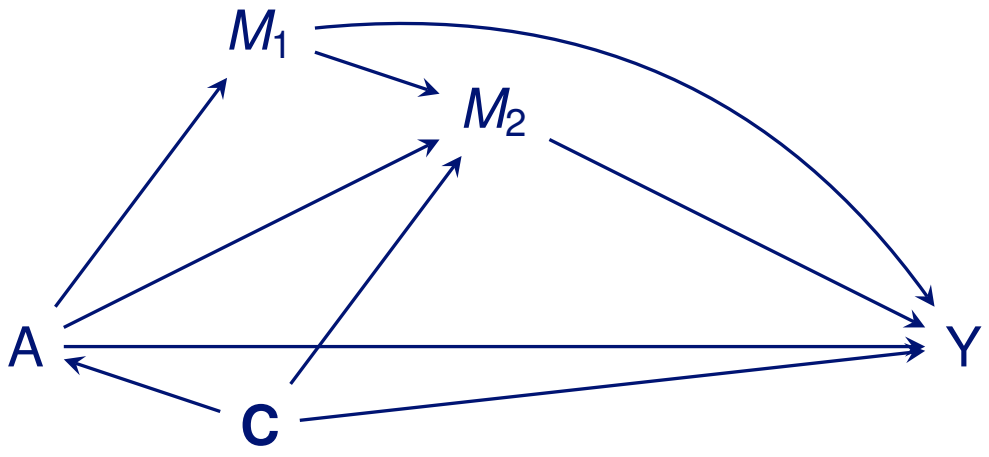
- `gformula` can be used to estimate natural and interventional effects
- `medeff` (Hicks and Tingley, 2011) and `paramed` (Emsley and Liu, 2013)† can only be used when *L* is not an intermediate confounder

†Now incorporated in version 18

Vansteelandt & Daniel “Interventional effects for mediation analysis with multiple mediators”, *Epidemiology* 2017



Vansteelandt & Daniel "Interventional effects for mediation analysis with multiple mediators", *Epidemiology* 2017



Micali *et al.* "Maternal Prepregnancy Weight Status and Adolescent Eating Disorder Behaviors", *Epidemiology* 2018

A: Prepregnancy maternal BMI

Y: Binge eating score at 13/14y

M_1: Childhood growth 8-12y

M_2: Maternal food avoidance at 8y

Effect of Maternal overweight	Mean difference	95% CI
Total	0.25	0.18, 0.32
Direct	-0.02	-0.08, 0.05
Indirect via growth	0.28	0.23, 0.33
Indirect via environment	-0.02	-0.04, -0.01

- Administrative databases
- High-dimensional covariates

- Linked administrative data sources increasingly available for:
  - comparative effectiveness research
  - policy evaluations
- Recognition of biases potentially affecting such research:
  - Confounding and measurement error
  - Selection bias
  - Lack of positivity
  - Immortal time bias
  - High dimensionality
- Advantages in emulating the design principles of experimental studies to avoid some of these biases (*"target trial emulation"*)

- Linked administrative data sources increasingly available for:
  - comparative effectiveness research
  - policy evaluations
- Recognition of biases potentially affecting such research:
  - Confounding and measurement error
  - Selection bias
  - Lack of positivity
  - Immortal time bias
  - High dimensionality
- Advantages in emulating the design principles of experimental studies to avoid some of these biases (*"target trial emulation"*)

- Linked administrative data sources increasingly available for:
  - comparative effectiveness research
  - policy evaluations
- Recognition of biases potentially affecting such research:
  - Confounding and measurement error
  - Selection bias
  - Lack of positivity
  - Immortal time bias
  - High dimensionality
- Advantages in emulating the design principles of experimental studies to avoid some of these biases (*"target trial emulation"*)

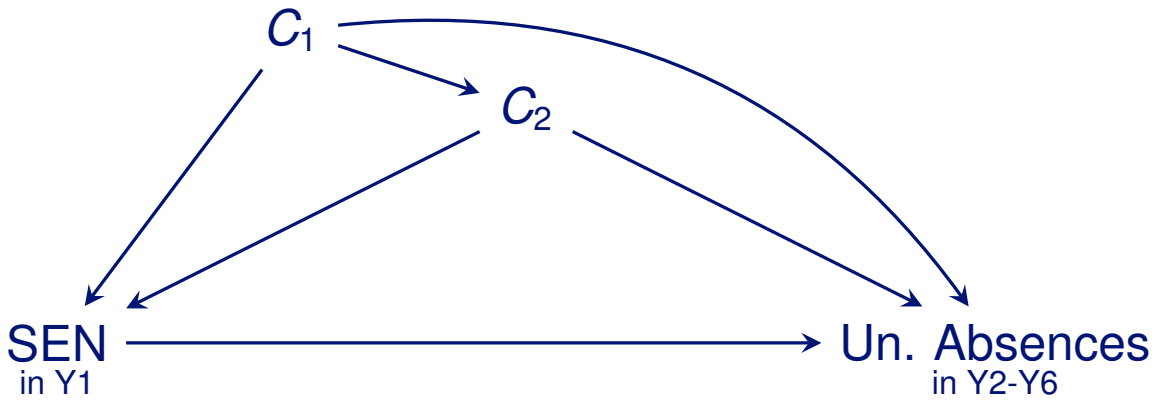
- Linked administrative data sources increasingly available for:
  - comparative effectiveness research
  - policy evaluations
- Recognition of biases potentially affecting such research:
  - Confounding and measurement error
  - Selection bias
  - Lack of positivity
  - Immortal time bias
  - High dimensionality
- Advantages in emulating the design principles of experimental studies to avoid some of these biases (*"target trial emulation"*)

- Background: Special educational needs (SEN) provision is designed to help pupils with additional educational, behavioural or health needs
- Aim: assess the impact of SEN provision on an educational outcomes during primary education for children with a certain congenital abnormality
- Data: ECHILD, linked educational and health records across England
- Results with/without (correct) lasso selection (using `telasso`)[‡]:

[‡] As developed by Chernozhukov (2018); Code to be deposited in GitHub

Linked administrative data: Nguyen *et al.* 2024

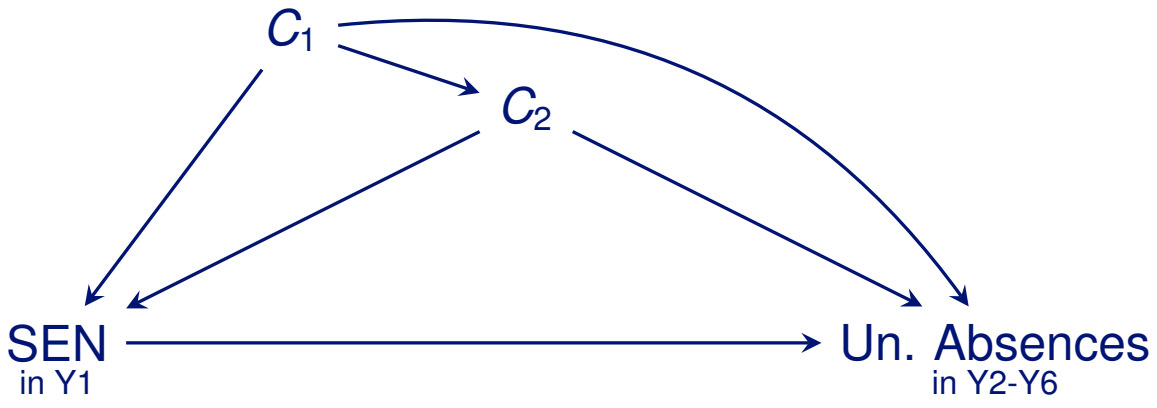
- Background: Special educational needs (SEN) provision is designed to help pupils with additional educational, behavioural or health needs
- Aim: assess the impact of SEN provision on an educational outcomes during primary education for children with a certain congenital abnormality
- Data: ECHILD, linked educational and health records across England



- Results with/without (correct) lasso selection (using `telasso`)[‡]:

[‡] As developed by Chernozhukov (2018); Code to be deposited in GitHub

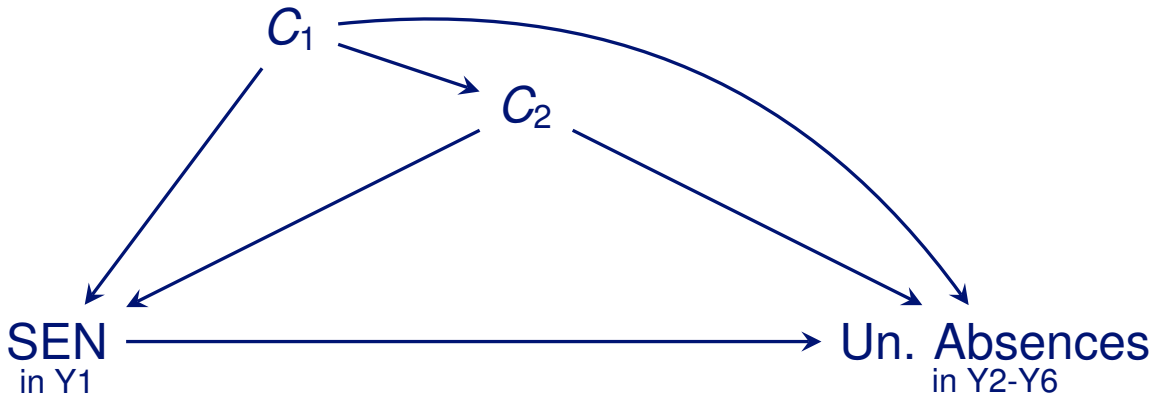
- Background: Special educational needs (SEN) provision is designed to help pupils with additional educational, behavioural or health needs
- Aim: assess the impact of SEN provision on an educational outcomes during primary education for children with a certain congenital abnormality
- Data: ECHILD, linked educational and health records across England



- Results with/without (correct) lasso selection (using `telasso`)[‡]:

[‡] As developed by Chernozhukov (2018); Code to be deposited in GitHub

- Background: Special educational needs (SEN) provision is designed to help pupils with additional educational, behavioural or health needs
- Aim: assess the impact of SEN provision on an educational outcomes during primary education for children with a certain congenital abnormality
- Data: ECHILD, linked educational and health records across England



- Results with/without (correct) lasso selection (using `telasso`)[‡]:

	Effect of SEN in Y1	
	Rate Ratio	95% CI
Crude	1.22	1.11, 1.34
IPW	0.86	0.76, 0.97
G-computation	0.98	0.86, 1.09
AIPW-lasso with int.	0.80	0.66, 0.95

[‡] As developed by Chernozhukov (2018); Code to be deposited in GitHub

Final thoughts . . .

Final thoughts . . .

Positives

- Wonderful Stata community
- Cross-pollination with econometricians
- Results increasingly reproducible

Final thoughts . . .

Positives

- Wonderful Stata community
- Cross-pollination with econometricians
- Results increasingly reproducible

Future challenges

- Access to Stata within secure environments: only via Google Notebooks and/or Python

Final thoughts . . .

Positives

- Wonderful Stata community
- Cross-pollination with econometricians
- Results increasingly reproducible

Future challenges

- Access to Stata within secure environments: only via Google Notebooks and/or Python

Thank you for listening!