Stata: a short history
viewed through epidemiology

Bianca L De Stavola

UCL Great Ormond Street Institute of Child Health

b.destavola@ucl.ac.uk

2025 Stata Biostatistics and Epidemiology Virtual Symposium

A personal perspective

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

A personal perspective

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

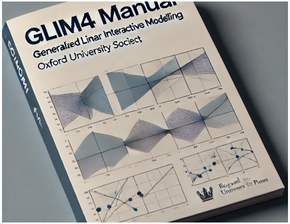
Overview

1. Some history
   - Before Stata
   - The 1990s
   - The 2000s
   - The 2010s
   - The 2020s
2. ...

Life before Stata

- Dominant software: GLIM

(*Generalised Linear Interactive Modelling*

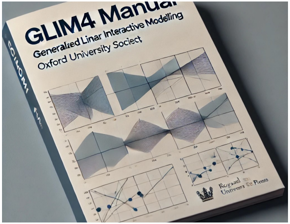


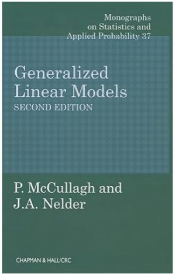
- Stemmed from a Royal Statistical Society WP
- Linked to the seminal book by McCullogh and Nelder
- Accessible on mainframe computers

Life before Stata
Statistical software in the UK

- Dominant software: GLIM

(*Generalised Linear Interactive Modelling*

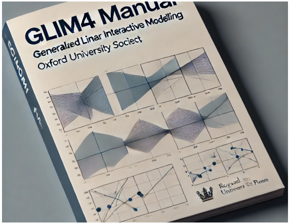


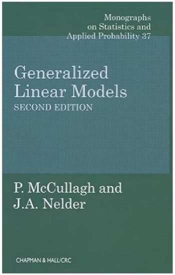


- Stemmed from a Royal Statistical Society WP
- Linked to the seminal book by McCullogh and Nelder
- Accessible on mainframe computers

- Dominant software: GLIM

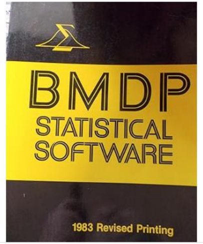
(*Generalised Linear Interactive Modelling*





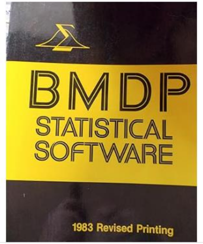
- Stemmed from a Royal Statistical Society WP
- Linked to the seminal book by McCullogh and Nelder
- Accessible on mainframe computers

- Biostatistics the US: BMDP ((*Biomedical Data Package*)



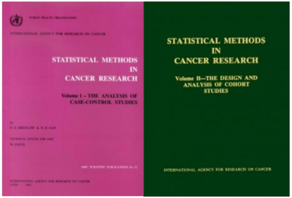
- Accessible (mostly) on mainframe computers

- Biostatistics the US: BMDP ((*Biomedical Data Package)*



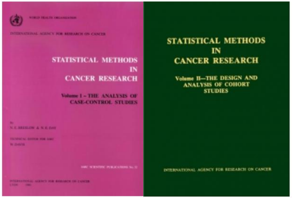
- Accessible (mostly) on mainframe computers

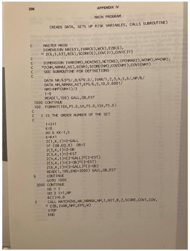
- Reference books in epidemiology:



- Included several Fortran programs specific for analyses of epi studies (*e.g.* conditional logistic regression)

- Reference books in epidemiology:





298 APPENDIX IV

MAIN PROGRAM

(READS DATA, SETS UP RISK VARIABLES, CALLS SUBROUTINE)

```
C     MASTER MAIN
      DIMENSION NR(63),IVAR(6),W(6),EZB(6),
     * Z(6,5,63),B(6),SCORE(6),COV(21),COVI(21)
C
C     DIMENSION IVAR(NM),NCA(NS),NCT(NS),IP(NMAX2),W(NM),WW(NM),
C    *Z(NM,NRMAX,NS),B(NM),SCORE(NM),COV(NM1),COVI(NM1)
C     SEE SUBROUTINE FOR DEFINITIONS
C
      DATA NR/63*5/,B/6*0.0/,IVAR/1,2,3,4,5,6/,NP/6/
      DATA NM,NRMAX,NIT,EPS/6,5,10,0.0001/
      NM1=NM*(NM+1)/2
      I-0
      READ(1,100) GALL,OB,EST
 1000 CONTINUE
  100 FORMAT(10X,F5.0,5X,F5.0,15X,F5.0)
C
C   I IS THE ORDER NUMBER OF THE SET
C
      I=I+1
      K=0
      DO 6 KK-1,5
      K=K+1
      Z(1,K,I)=2-GALL
      IF (OB.EQ.9)  OB=2
      Z(2,K,I)=2-OB
      Z(3,K,I)=2-EST
      Z(4,K,I)=(2-GALL)*(2-EST)
      Z(5,K,I)=(2-OB)*(2-EST)
      Z(6,K,I)=(2-GALL)*(2-OB)
      READ(1,100,END=2000) GALL,OB,EST
 6    CONTINUE
      GOTO 1000
 2000 CONTINUE
      NS = I
      DO 2 I=1,NP
      B(I)=0.0
 2    CALL MATCH(NS,NR,NRMAX,NM,I,NIT,B,Z,SCORE,COVI,COV,
     * EZB,IVAR,NM1,EPS,W)
      STOP
      END
```

- Included several Fortran programs specific for analyses of epi studies (*e.g.* conditional logistic regression)

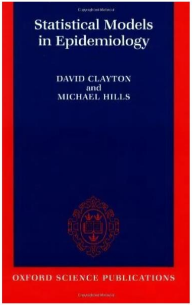
1984

- Stata arrives!

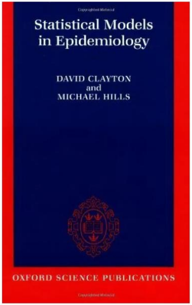
Commands in Stata 1.0 and Stata 1.1

append	dir	infile	plot	spool
beep	do	input	query	summarize
by	drop	label	regress	tabulate
capture	erase	list	rename	test
confirm	exit	macro	replace	type
convert	expand	merge	run	use
correlate	format	modify	save	
count	generate	more	set	
describe	help	outfile	sort	

- Developed for personal computers


Statistical Models
in Epidemiology
DAVID CLAYTON
and
MICHAEL HILLS
OXFORD SCIENCE PUBLICATIONS

The 1990s
Michael Hills and David Clayton and Stata



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

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

Michael in Florence

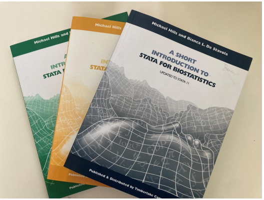
Ana Timberlake

The 1990s
Teaching with Michael Hills

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

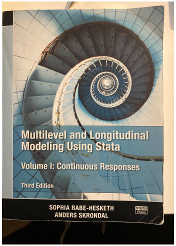
The Stata manuals . . .

Michael's version!

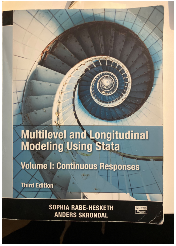


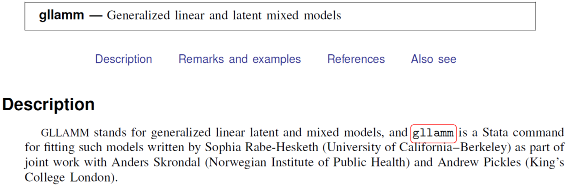
The 2000s

- Mixed effects models
- Missing data



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

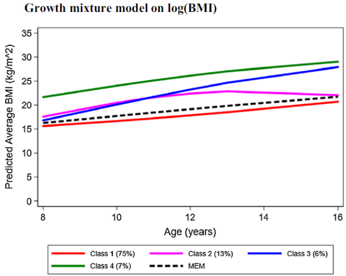

Multilevel and Longitudinal Modeling Using Stata
Volume I: Continuous Responses
Third Edition
SOPHIA RABE-HESKETH
ANDERS SKRONDAL


gllamm — Generalized linear and latent mixed models

Description Remarks and examples References Also see

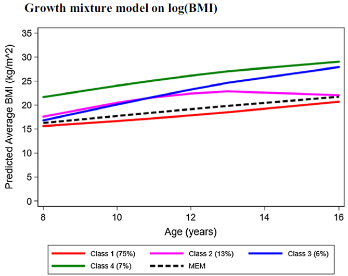
Description

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`

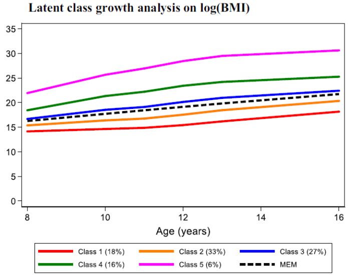
[Herle *et al.* EJE 2021]



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

The 2000s
Missing data

- 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

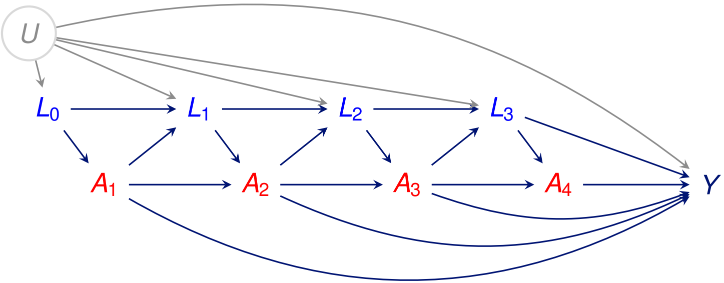
- The currently dominant approach in biostatistics and epidemiology relies on potential outcomes (POs) [Rubin, 1974; Robins, 1986; Pearl, 1995]
- Adopting this approach, we are concerned with questions formulated as contrasts of outcomes that would occur under hypothetical interventions on the exposure: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

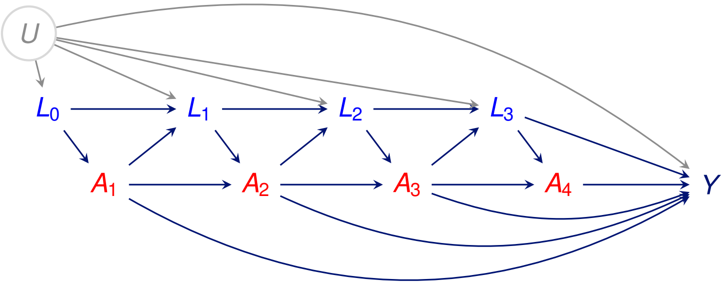
- `teffects` implements (a) and (b) for time-fixed exposures

* 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!

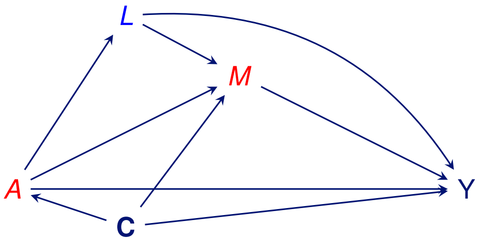
The Stata Journal (2011)
11, Number 4, pp. 479–517

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

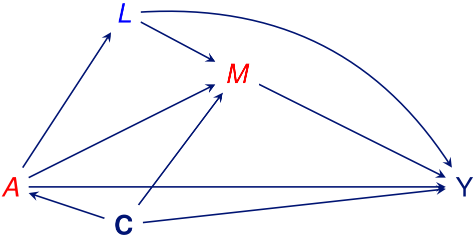
Bianca L. De Stavola
Centre for Statistical Methodology
London School of Hygiene and Tropical Medicine
London, UK

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

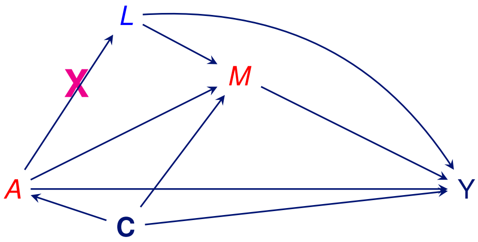
† Now incorporated in version 18



- `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

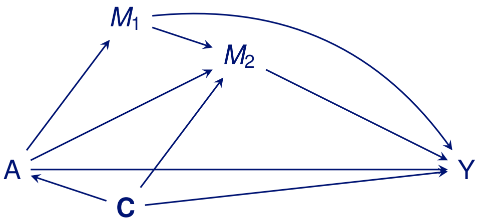
† Now incorporated in version 18



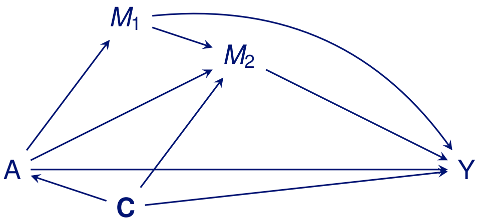
- `gformula` can be used to estimate natural and interventional effects
- `medeff` (Hicks and Tingley, 2011) and `paramed` (Emsley and Liu, 2013)† can only be used when L is not an intermediate confounder

† 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

- 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

- Background: Special educational needs (SEN) provision: policy designed to help pupils with additional educational, behavioural or health needs
- Aim: assess the impact of SEN provision on an unauthorised absences for children with a certain health needs
- Data: ECHILD, linked educational and health records across England
- Many challenges including high-dimensionality of confounders
- Results with/without (correct) lasso selection (using `telasso`)‡:

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

The 2020s

Linked administrative data: Nguyen *et al.* 2024

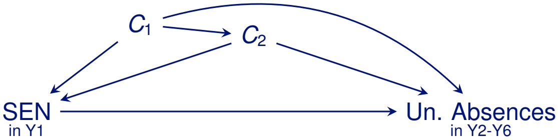
- Background: Special educational needs (SEN) provision: policy designed to help pupils with additional educational, behavioural or health needs
- Aim: assess the impact of SEN provision on an unauthorised absences for children with a certain health needs
- Data: ECHILD, linked educational and health records across England

- Many challenges including high-dimensionality of confounders
- 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: policy designed to help pupils with additional educational, behavioural or health needs
- Aim: assess the impact of SEN provision on an unauthorised absences for children with a certain health needs
- Data: ECHILD, linked educational and health records across England



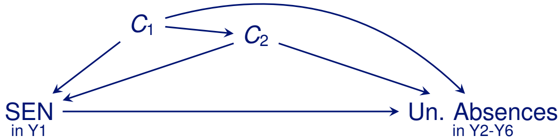
- Many challenges including high-dimensionality of confounders
- Results with/without (correct) lasso selection (using `telasso`)‡:

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

The 2020s

Linked administrative data: Nguyen *et al.* 2024

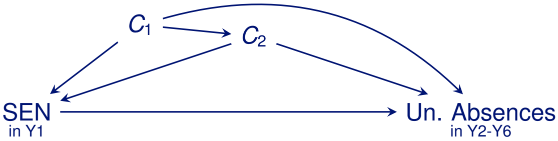
- Background: Special educational needs (SEN) provision: policy designed to help pupils with additional educational, behavioural or health needs
- Aim: assess the impact of SEN provision on an unauthorised absences for children with a certain health needs
- Data: ECHILD, linked educational and health records across England



- Many challenges including high-dimensionality of confounders
- 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: policy designed to help pupils with additional educational, behavioural or health needs
- Aim: assess the impact of SEN provision on an unauthorised absences for children with a certain health needs
- Data: ECHILD, linked educational and health records across England



- Many challenges including high-dimensionality of confounders
- Results with/without (correct) lasso selection (using `telasso`)[‡]:

Effect of SEN in Y1	Rate Ratio	95% CI
Crude	1.22	1.11, 1.34
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 of biostatisticians and econometricians
- Results increasingly reproducible

Final thoughts . . .

Positives

- Wonderful Stata community
- Cross-pollination of biostatisticians and econometricians
- Results increasingly reproducible

Future challenges

- Access to Stata within secure environments

Final thoughts . . .

Positives

- Wonderful Stata community
- Cross-pollination of biostatisticians and econometricians
- Results increasingly reproducible

Future challenges

- Access to Stata within secure environments

Thank you for listening!