Two countries, twenty models and one secret weapon:

CANCELLED!

Comparing change over time in convictions trajectories between Scotland and Queensland

Ben Matthews, University of Edinburgh

Two countries, lots of models and one secret weapon:

Two applications of the “many models” workflow

Ben Matthews, University of Edinburgh

Overview

- The many models workflow
- Why do this to yourself?
    - The secret weapon
        - Change over time in criminal careers
        - One latent class growth model per cohort
    - Visualizing uncertainty
        - How many latent classes?
        - One latent class growth model per bootstrap
- What have we learned?

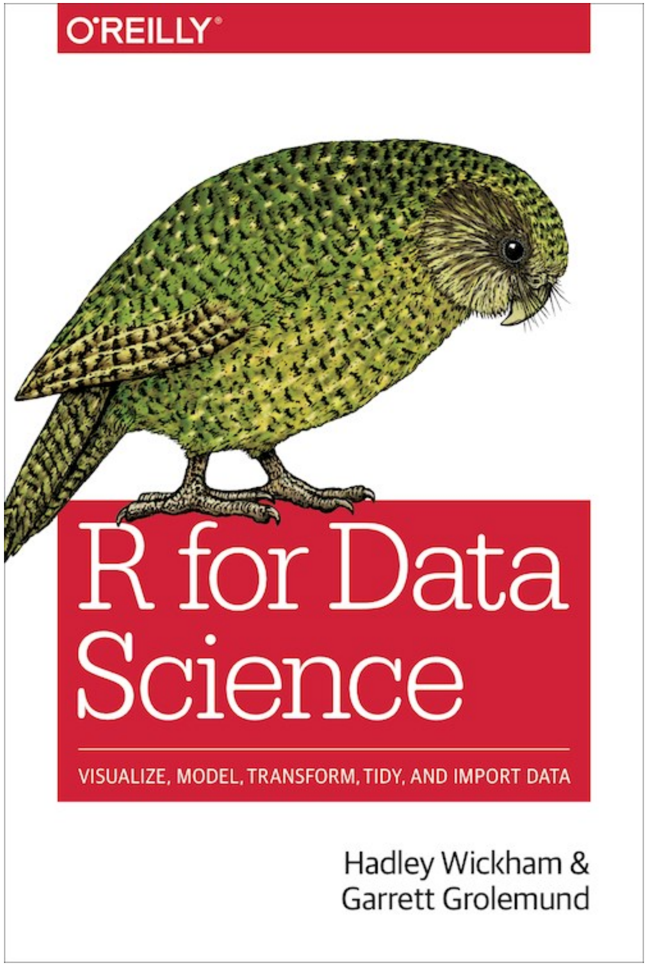
Not going to cover

- How latent class growth models work really
- Substantively what the results mean

Many models

Many models

1. Use many simple models to better understand complex datasets.
2. Use list-columns to store arbitrary data structures in a data frame.
3. Turn models into tidy data so you can work with model results like any other dataset.



https://r4ds.had.co.nz/many-models.html

List-columns?

- "Arbitrary data" can include, well anything!
- A dataset, a model call, model results…

```
# A tibble: 18 x 4
   birth_year data                     mplus_command     k_2
        <int> <list>                   <list>            <list>
 1       1973 <tibble [1,114 x 12]> <S3: mplusObject> <S3: mplusObject>
 2       1974 <tibble [1,102 x 12]> <S3: mplusObject> <S3: mplusObject>
 3       1975 <tibble [1,072 x 12]> <S3: mplusObject> <S3: mplusObject>
...
18       1990 <tibble [770 x 12]>   <S3: mplusObject> <S3: mplusObject>
```

List-columns and iteration

- For each one of a group of things do some thing to them
- For each birth year, fit a two class model to that cohort's data using the appropriate Mplus command

```
# A tibble: 18 x 4
   birth_year data                    mplus_command     k_2
        <int> <list>                  <list>            <list>
 1       1973 <tibble [1,114 x 12]> <S3: mplusObject> <S3: mplusObject>
 2       1974 <tibble [1,102 x 12]> <S3: mplusObject> <S3: mplusObject>
 3       1975 <tibble [1,072 x 12]> <S3: mplusObject> <S3: mplusObject>
...
18       1990 <tibble [770 x 12]>   <S3: mplusObject> <S3: mplusObject>
```

Iteration in R

- I used R's `purrr` package to iterate (there are other ways too)
- Lots of great resources to learn about iteration using `purrr`:
  - https://github.com/cwickham/purrr-tutorial/blob/master/slides.pdf
  - https://speakerdeck.com/jennybc/purrr-workshop
  - https://github.com/jenniferthompson/RLadiesIntroToPurrr

Why do this to yourself?

“The Secret Weapon”

Statistical Modeling, Causal Inference, and Social Science

HOME BOOKS BLOGROLL SPONSORS

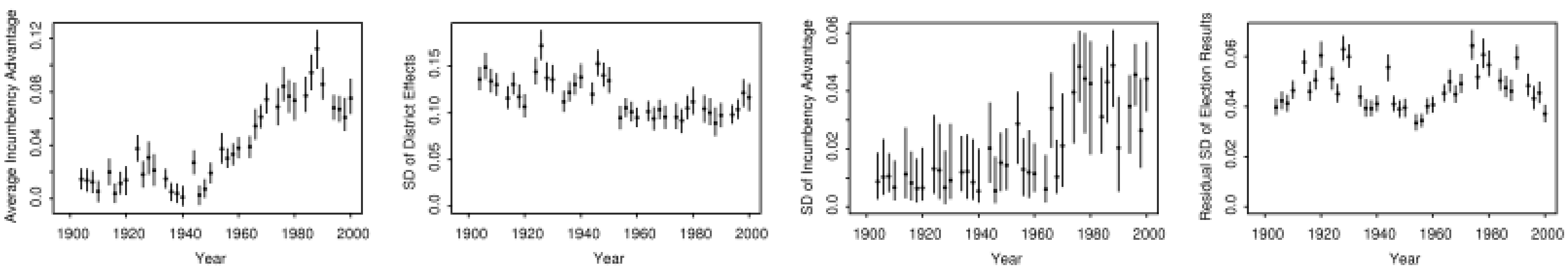
« Meritocracy won't happen: the problem's with the "ocracy"   Still more on R software for matching for causal inference »

The secret weapon

Posted by Andrew on 7 March 2005, 12:52 am

An incredibly useful method is to fit a statistical model repeatedly on several different datasets and then display all these estimates together. For example, running a regression on data on each of 50 states (see here as discussed here), or running a regression on data for several years and plotting the estimated coefficients over time.

Here's another example:



The idea is to fit a separate model for each year, or whatever, and then to look at all these estimates together to see trends. This can be considered as an approximation to multilevel modeling, with the partial pooling done by eye on the graphs rather than using a full statistical model.

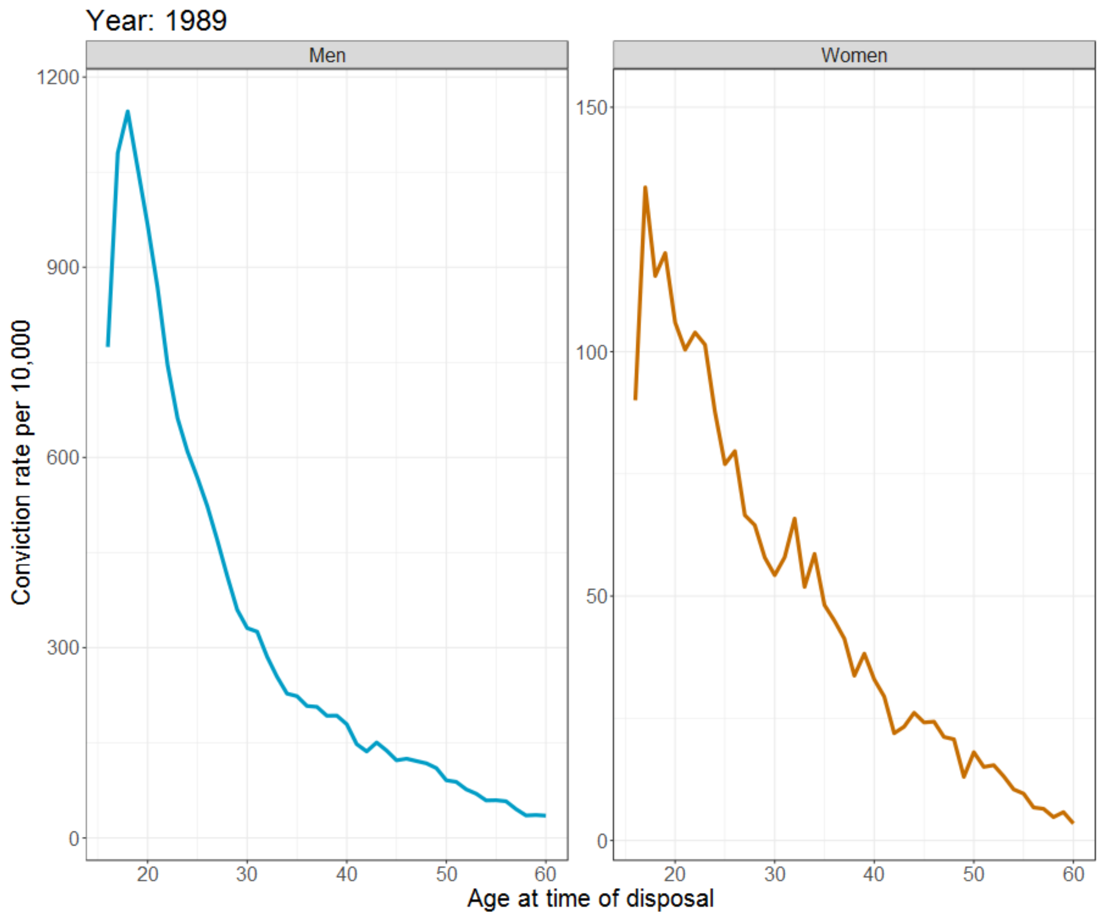
https://statmodeling.stat.columbia.edu/2005/03/07/the_secret_weap/

Understanding Inequalities

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

Changing Criminal Careers

- Well-established relationship between age and crime as far back as Quetelet (1832)
- Significant change in last 25 or so years in Scotland (Matthews and Minton 2018)
- Fit a model to successive cohorts in the Scottish Offenders Index and compare results!


Year: 1989
Men
Women
1200
900
600
300
0
150
100
50
0
20
30
40
50
60
20
30
40
50
60
Conviction rate per 10,000
Age at time of disposal

A core issue in the criminal careers debate is the shape of the age-crime curve at the level of the individual. Is it single peaked, as argued by Gottfredson and Hirschi, or age invariant, as argued by Blumstein and Cohen? The results reported in Table 4 support the Gottfredson and Hirschi position. The debate over the shape of the age-crime curve, however, neglects a possibility that fundamentally alters the nature of the phenomenon to be explained—individuals may not only consistently differ in their rate of offending at any given age but may also have distinctive trajectories of offending over age.

Nagin and Land (1993:346) *Age, Criminal Careers, and Population Heterogeneity: Specification and Estimation of a Nonparametric, Mixed Poisson Model,* Criminology, *Vol. 31, No. 3, pp. 327:362*

Identifying conviction trajectories

- Latent Class Growth Mixture Model (LCGM, Nagin and Land 1993), makes groups out of data based on a repeated measures dependent variable
- Each group gets a trajectory (in our case, probability of conviction in a given year)
- LCGMs are very popular (if divisive) in criminology (Nagin and Land 2005a, 2005b; Sampson and Laub 2005)

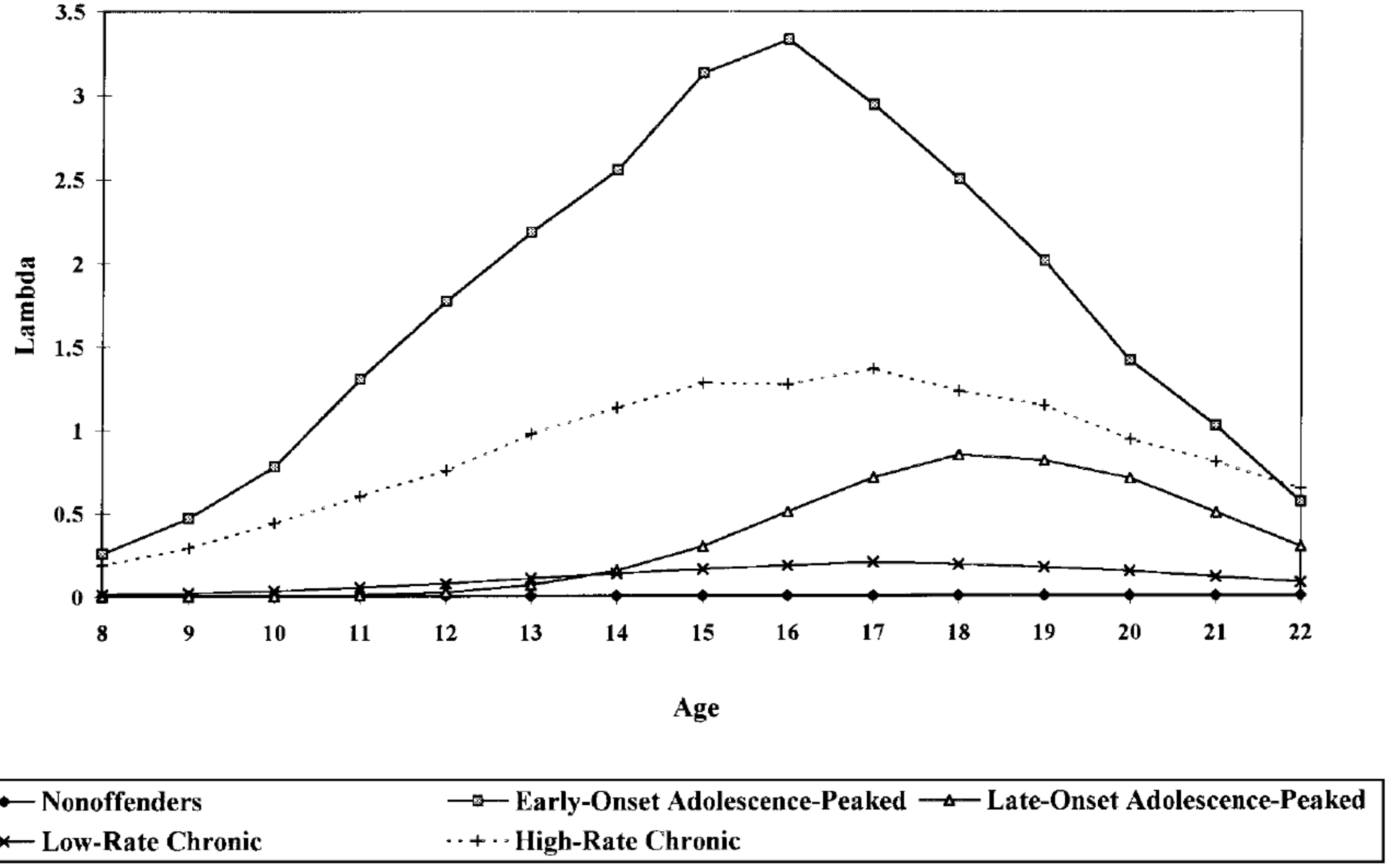


FIG. 5.—Predicted misdemeanor/felony rates by age for the Racine 1955 five-class model

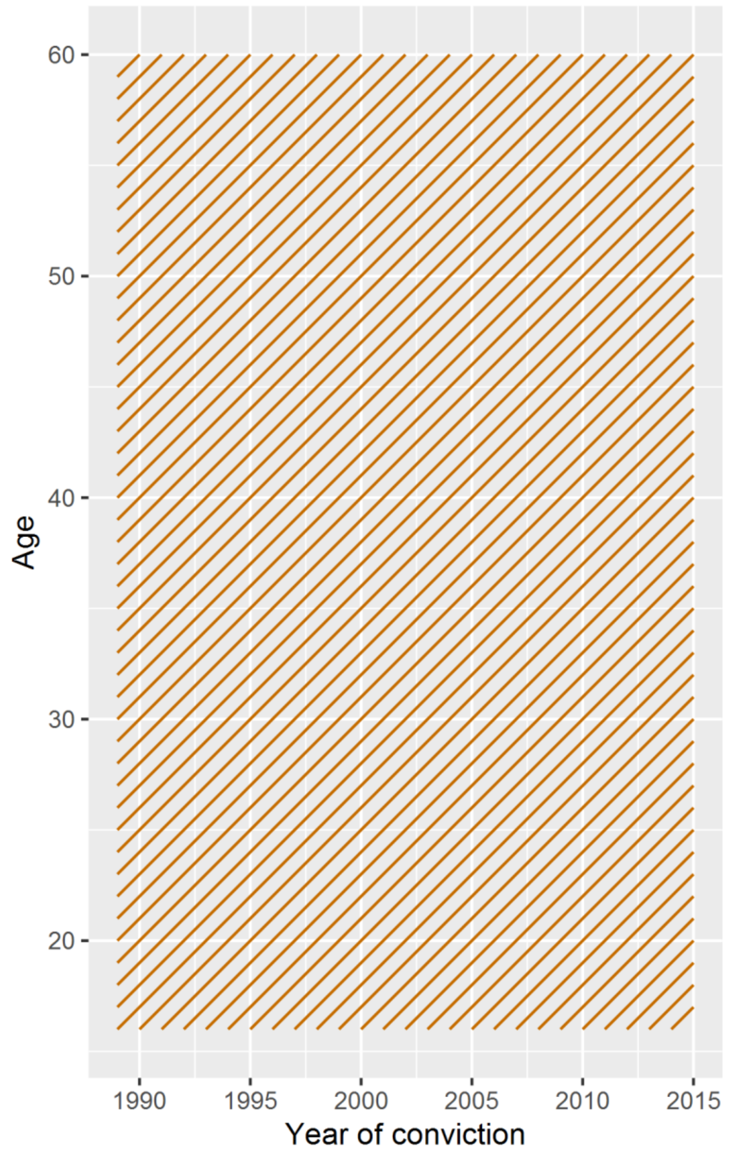
D'Unger et al. (1998), *How Many Latent Classes of Delinquent/Criminal Careers? Results from Mixed Poisson Regression Analyses,* American Journal of Sociology, *Vol. 103, No. 6, pp. 1593-1630*

Understanding Inequalities

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

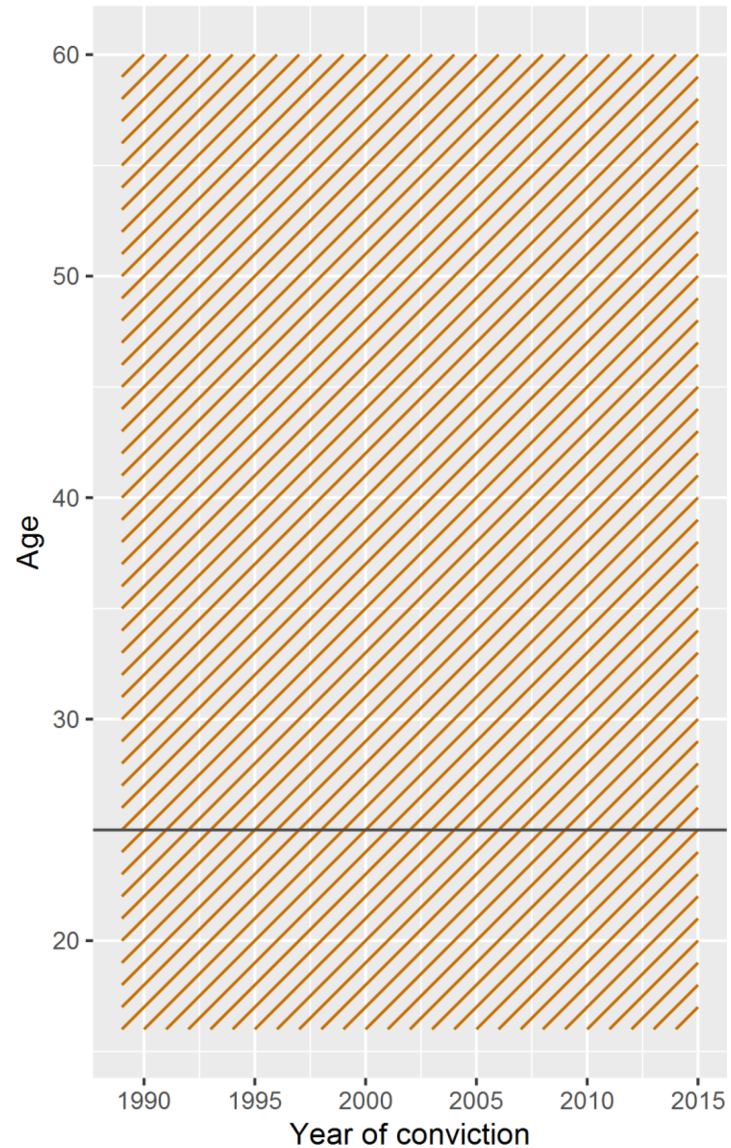
Many LCGMs

For each cohort,
fit LCGM



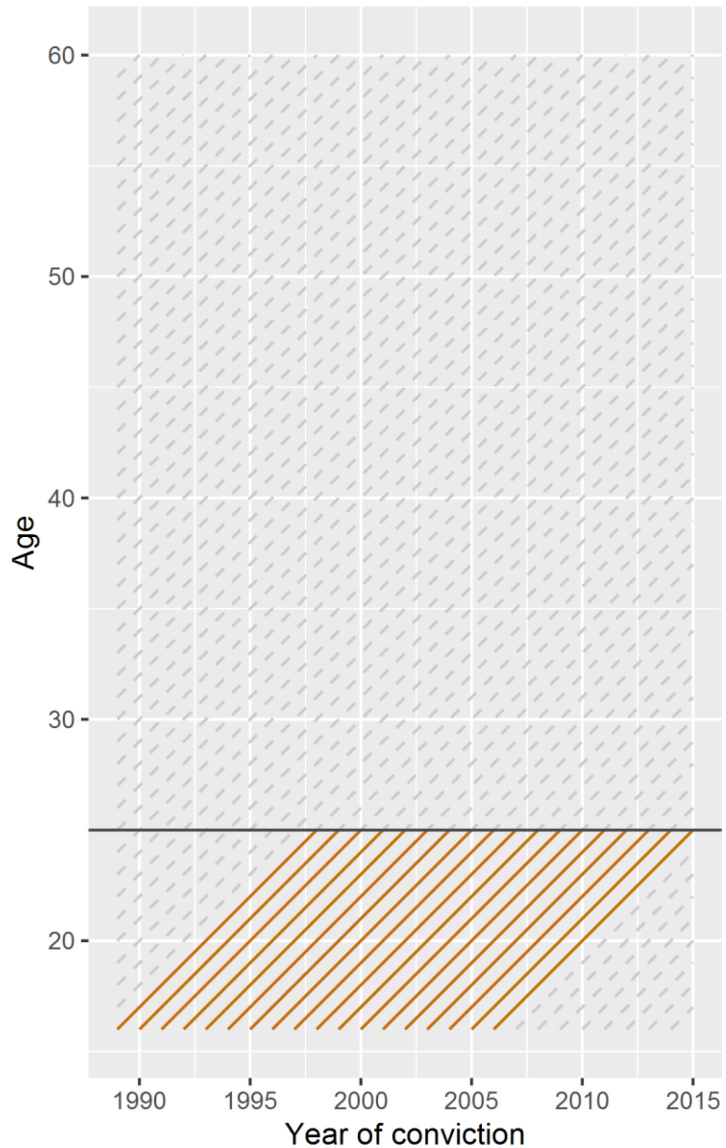
Many LCGMs

`For each cohort,`
`fit LCGM`



Many LCGMs

`For each cohort,`
`fit LCGM`



Understanding Inequalities

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

Modelling strategy

- Models fit in Mplus 8.2 using the `MplusAutomation` (Hallquist and Wiley 2018) R package to call Mplus, fit the models and then pull the results back into R
- This lets you use R's standard iteration tools with Mplus

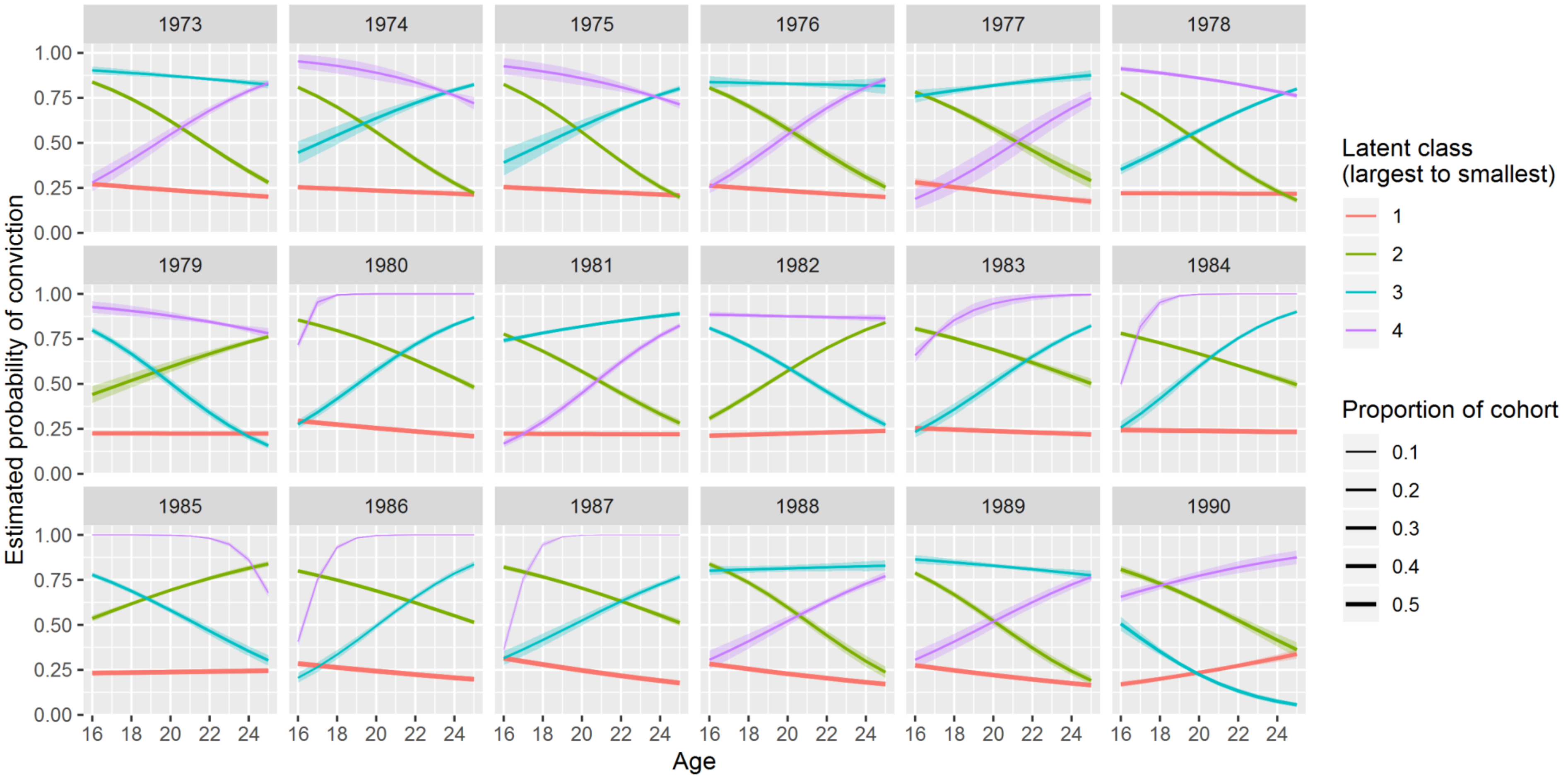
Which results?

- Shape of trajectories
- Number of classes
- Proportion of people assigned to each class

How many latent classes?

Model selection process:

1. AIC/BIC/aBIC
2. (VLMR test)
3. (BLRT test)
4. Substantive interpretation
5. “Norris method” (see Norris, 2009)



Understanding Inequalities

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

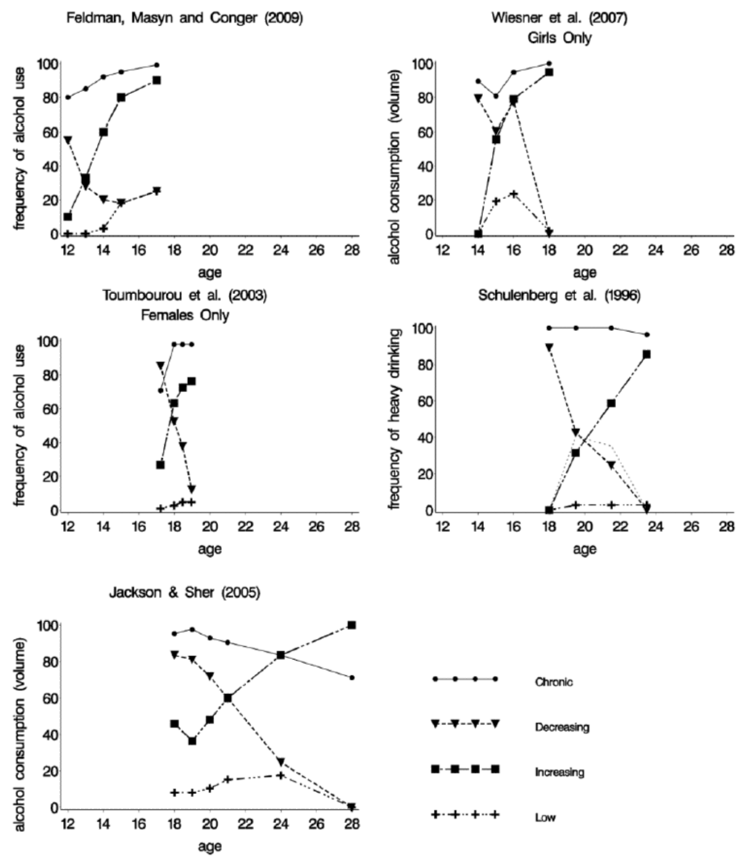


Figure 1.
Trajectories of various measures of alcohol involvement spanning ages 13-29 illustrating the "soldier's bed cat's cradle" phenomenon in five studies. All data have beeen scaled so that lowest reported value is a 0 and the highest reported value is 100 in each study.

Sher et al. (2011), *Alcohol Use Trajectories and the Ubiquitous Cat's Cradle: Cause for Concern?,* J Abnorm Psychol, *Vol. 120, No. 2, pp. 322-335*

Results

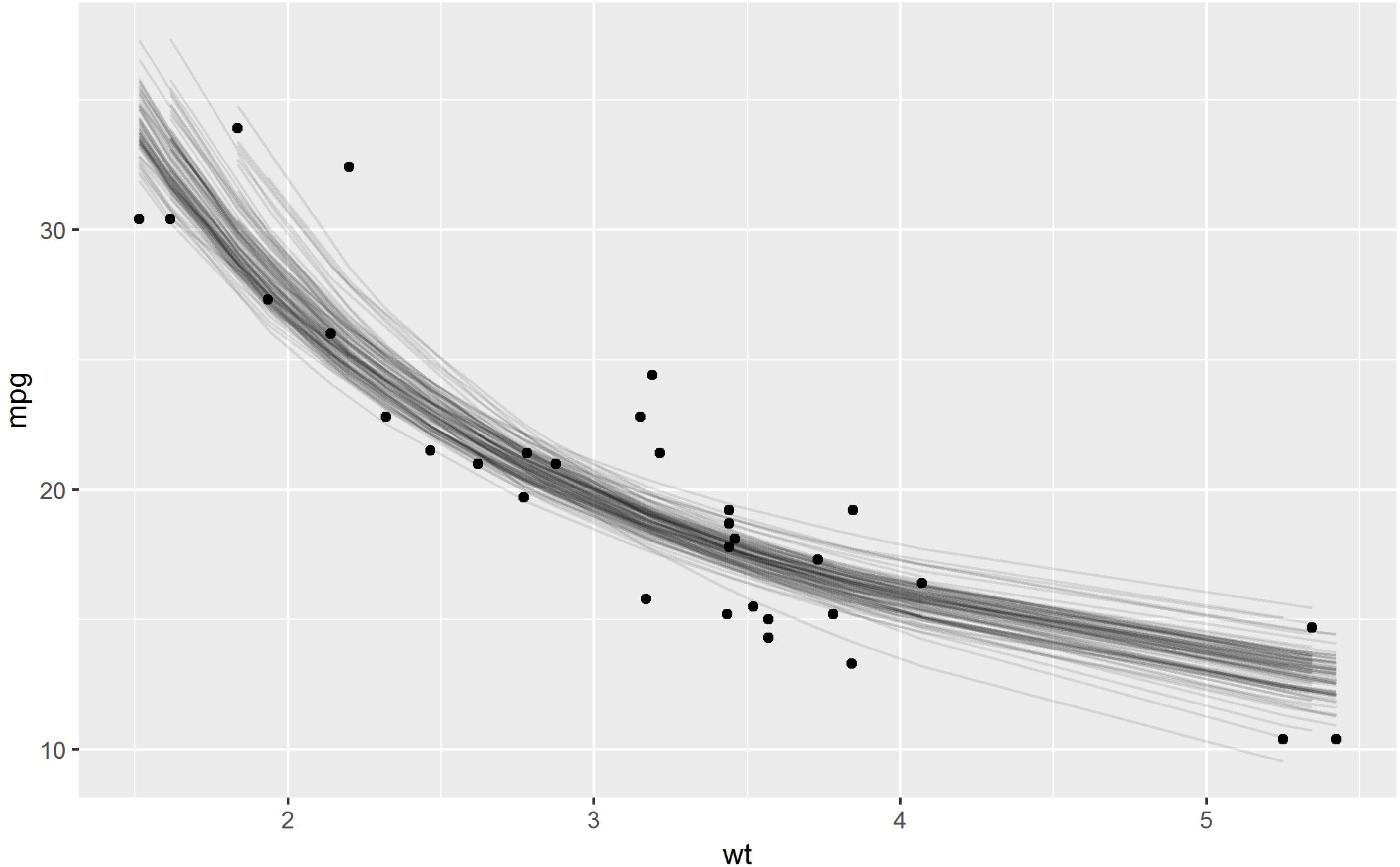
- Does look like there's maybe some differences between cohorts in the trajectory shapes, but cat's cradle means it's hard to tell
- … but are these differences meaningful? Many models approach means we don't know statistically
- One model per cohort means class definitions change – can't really compare proportion assigned
- Different number of classes – selection is very subjective!

Visualizing uncertainty in LCGMs

“Point estimates should be accompanied by information on precision!!!”

- Vernon Gayle

Motivation 1: I wanted this -



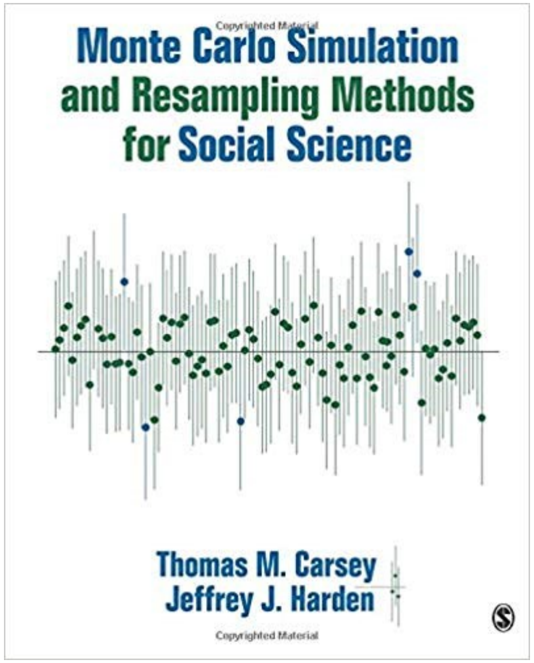
https://cran.r-project.org/web/packages/broom/vignettes/bootstrapping.html

Motivation 2: What do these groups **mean**?

- Direct (LCGM groups are real!) and indirect interpretations (LCGM groups are statistical shorthand!) (Bauer 2007:779)
- In some cases, aetiology of groups is theorized **after** they have been identified by LCGM (Krohn et al. 2013:194)
- Are we just chasing noise?

Many LCGMs (again)

- For each bootstrap replication, fit LCGM
- Sample with replacement 25 times from original dataset using rsample package
- Work with bootstrap replications as with any (nested/list-column) dataset
- See Carsey and Harden (2014) on bootstrapping methods



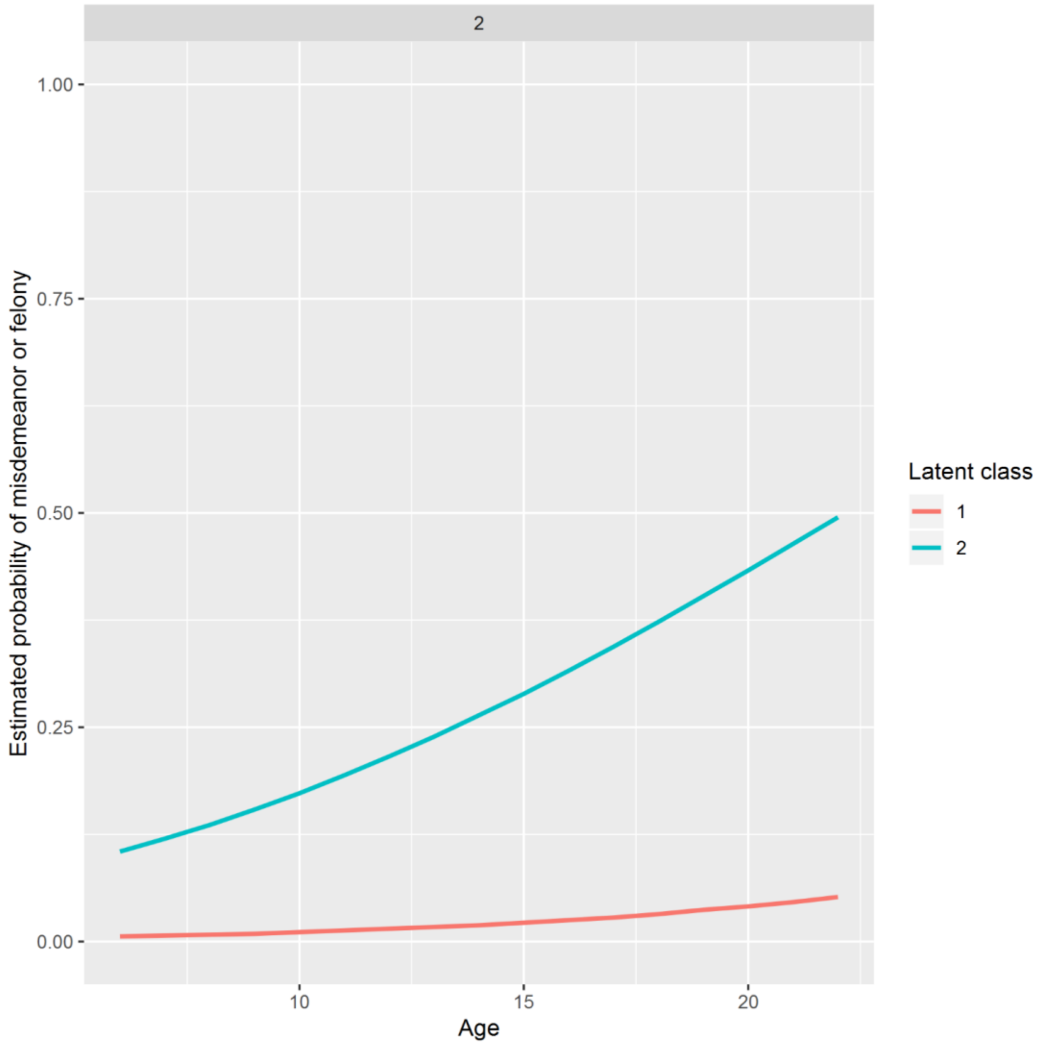
Many LCGMs (again)

- Data from Racine, Wisonsin 1955 birth cohort (downloaded from https://www.icpsr.umich.edu/icpsrweb/)
- Remember: substantive interest in the **shape** of the different trajectories

How many latent classes?

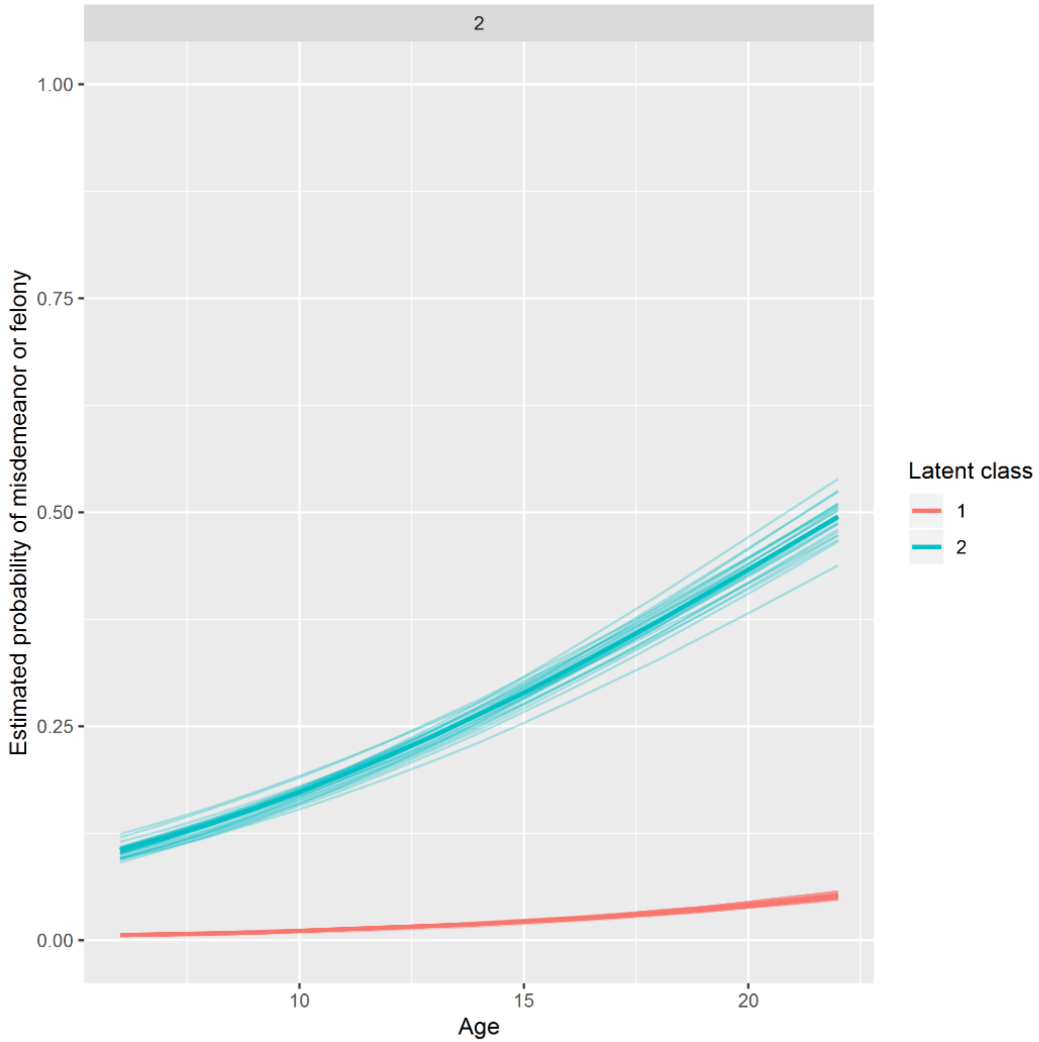
Model selection process:

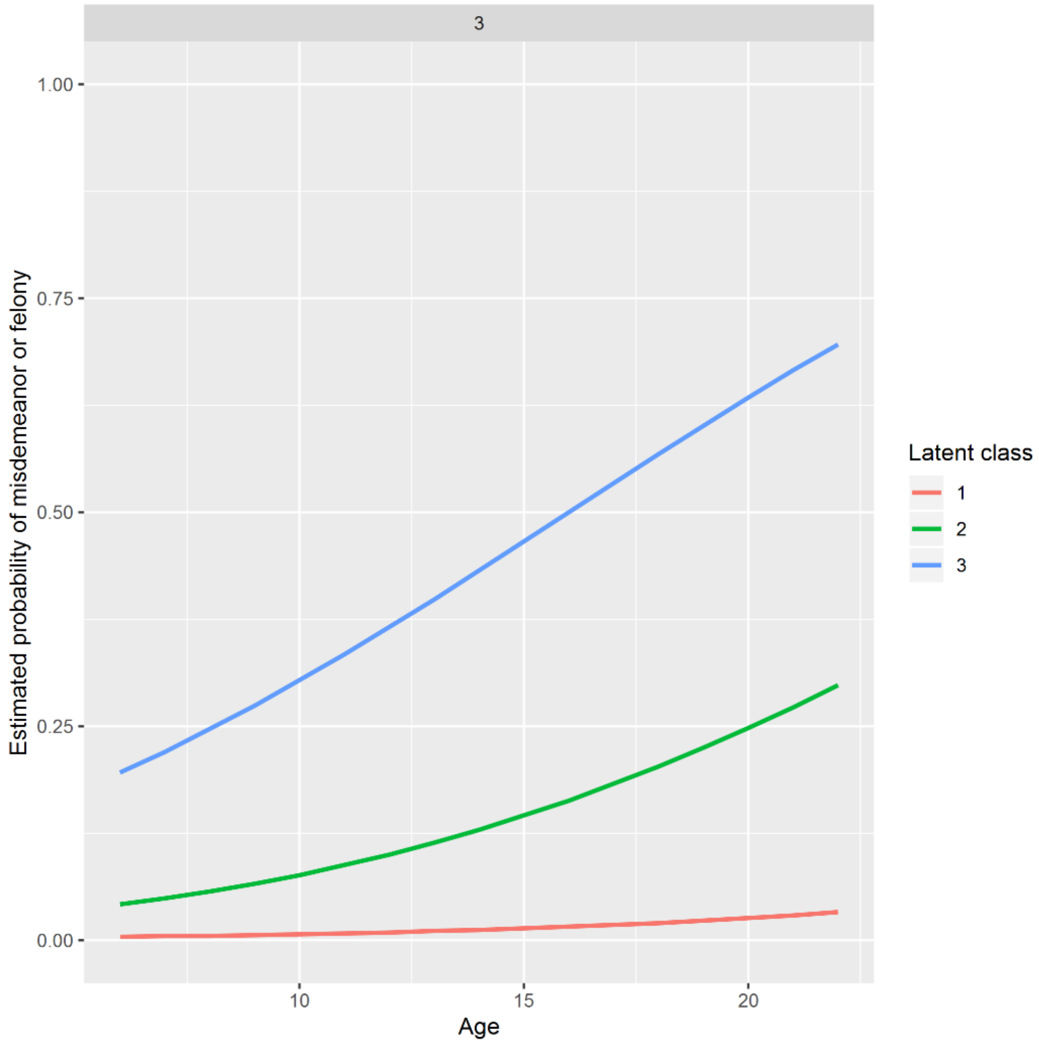
1. AIC/BIC/aBIC
2. (VLMR test)
3. (BLRT test)
4. Substantive interpretation
5. “Norris method” (see Norris, 2009)
6. ... Consistency of trajectory shapes across bootstrap replications?



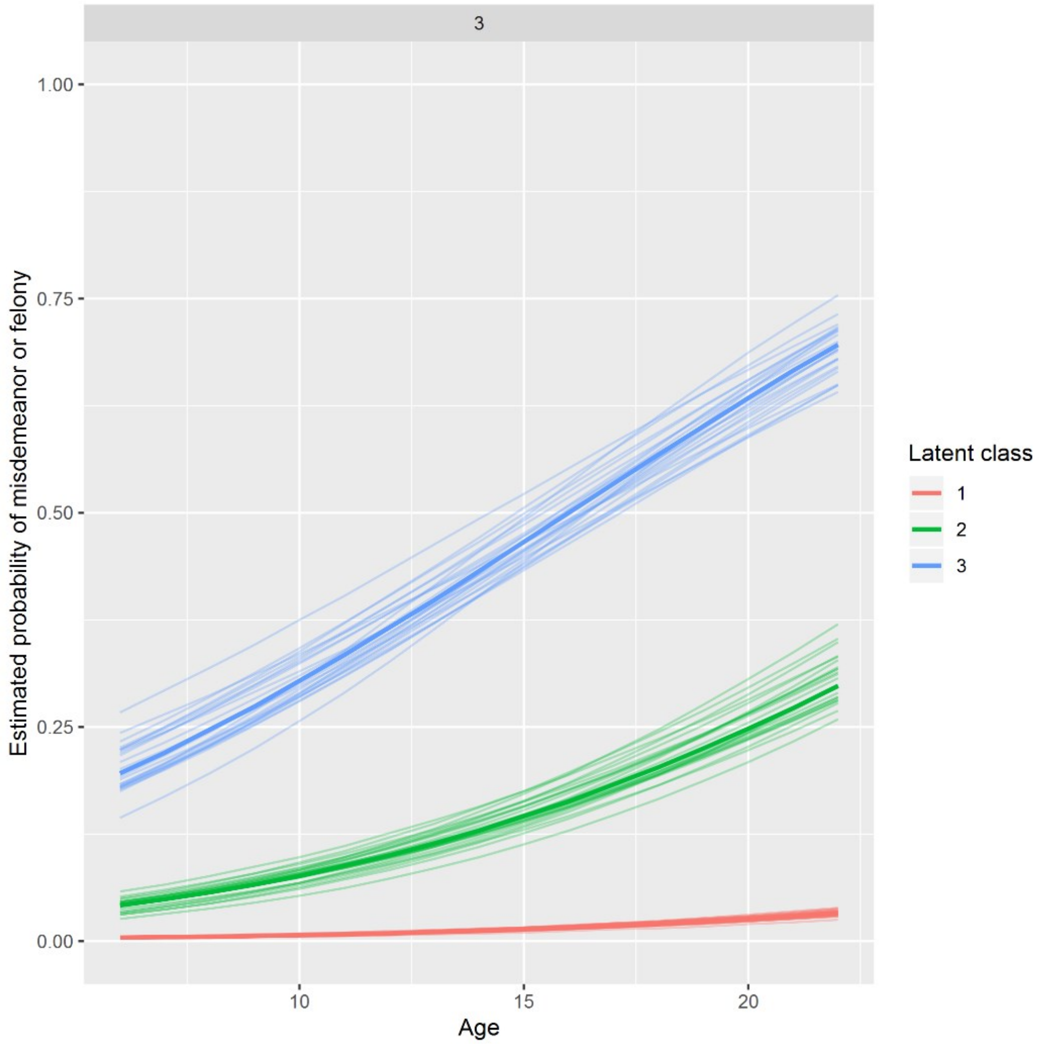
Understanding Inequalities

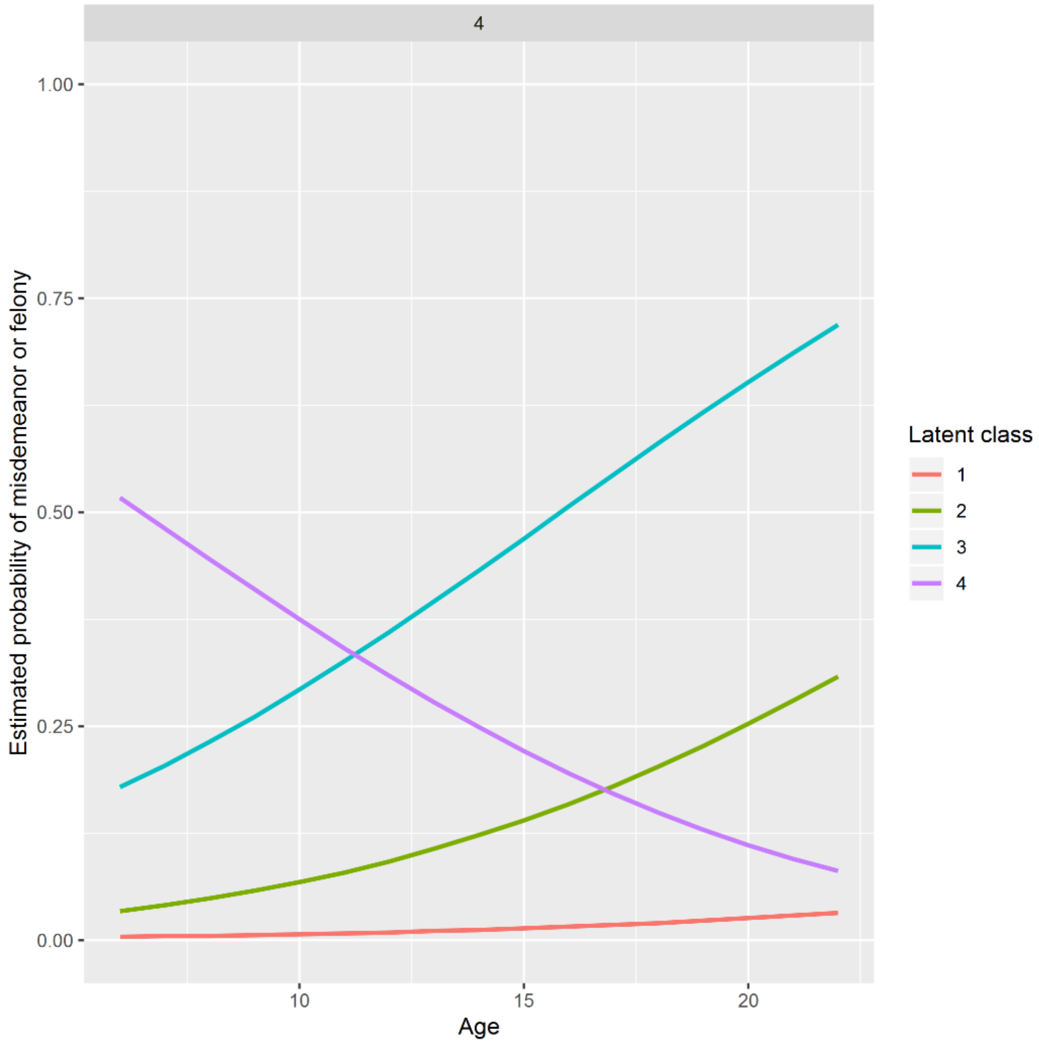
E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

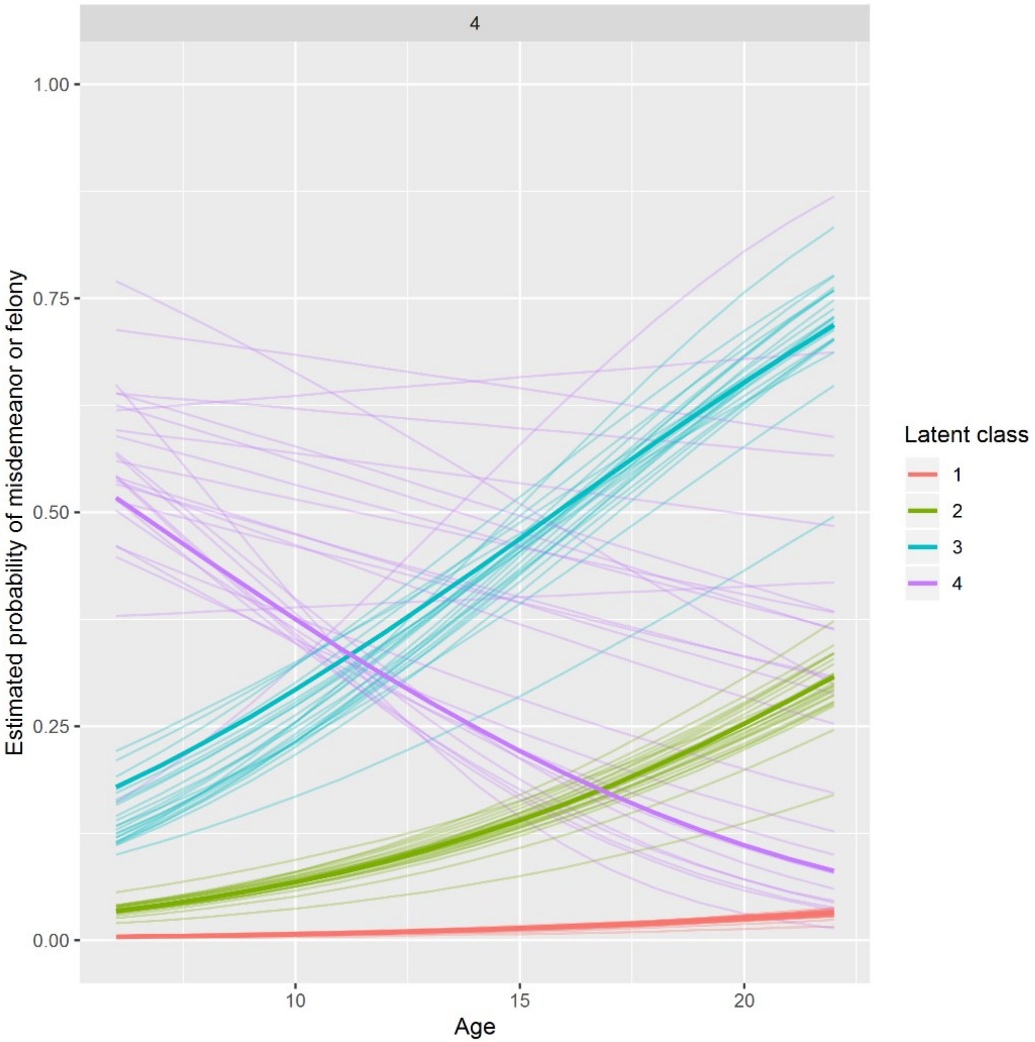

2
1.00
0.75
0.50
0.25
0.00
Estimated probability of misdemeanor or felony
10
15
20
Age
Latent class
1
2



3
1.00
0.75
0.50
0.25
0.00
Estimated probability of misdemeanor or felony
10
15
20
Age
Latent class
1
2
3

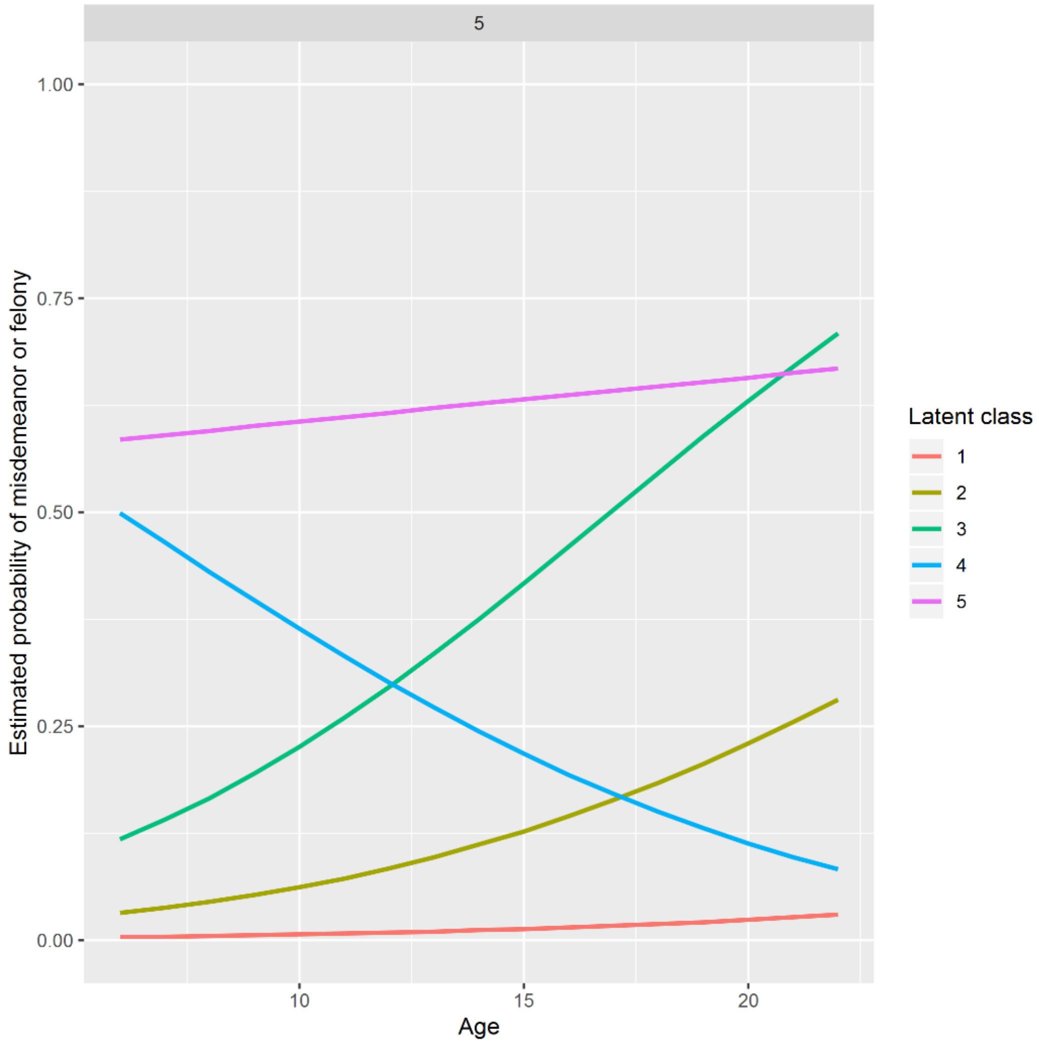

3
Estimated probability of misdemeanor or felony
1.00
0.75
0.50
0.25
0.00
10
15
20
Age
Latent class
1
2
3

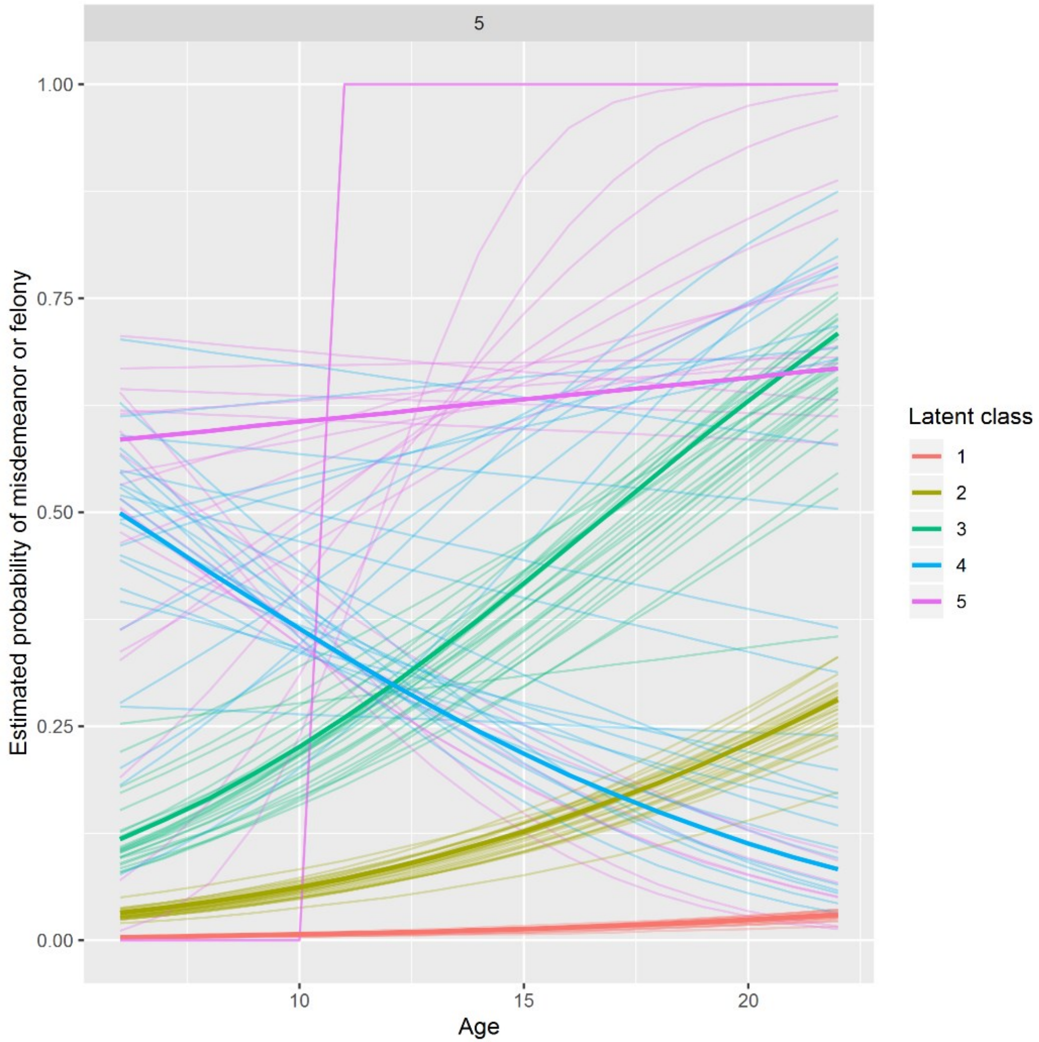

4
Estimated probability of misdemeanor or felony
1.00
0.75
0.50
0.25
0.00
10
15
20
Age
Latent class
1
2
3
4

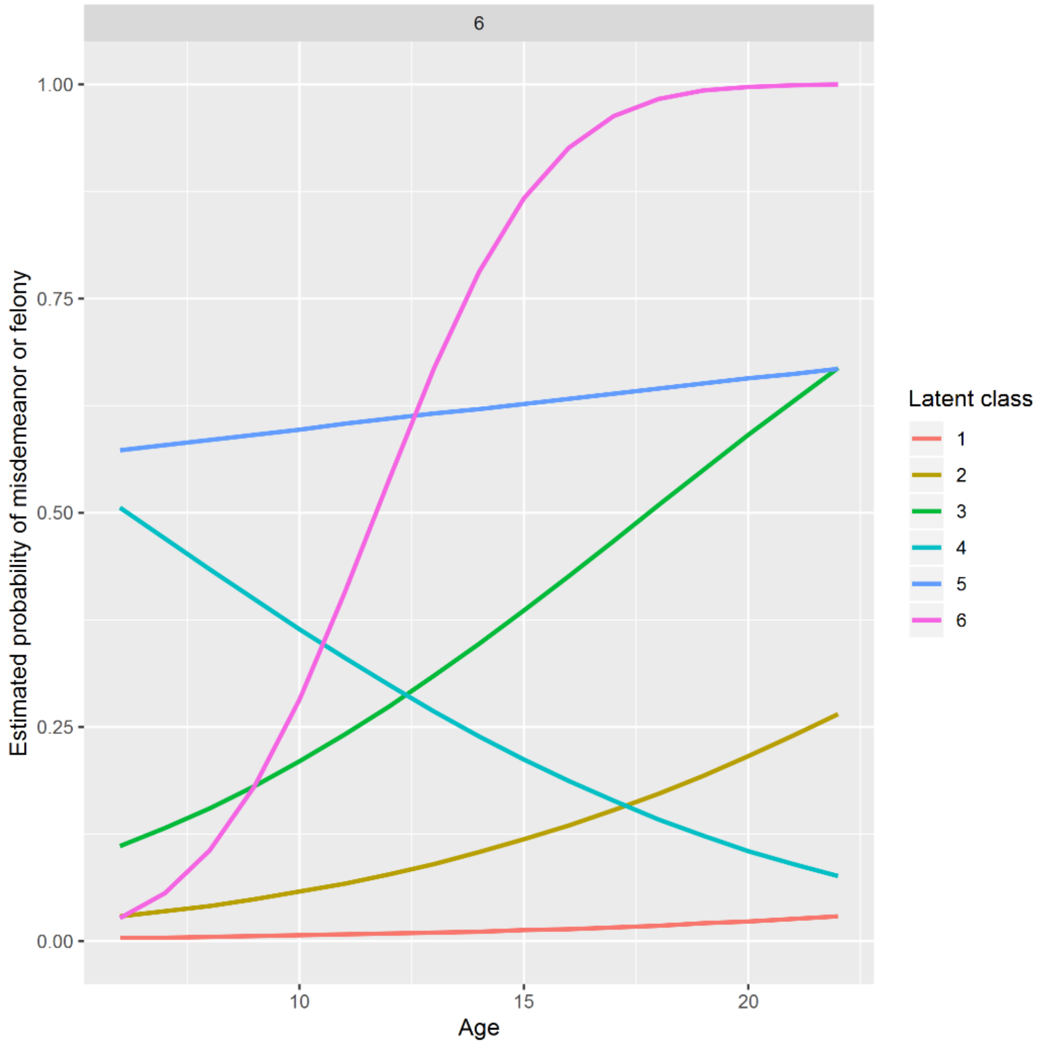


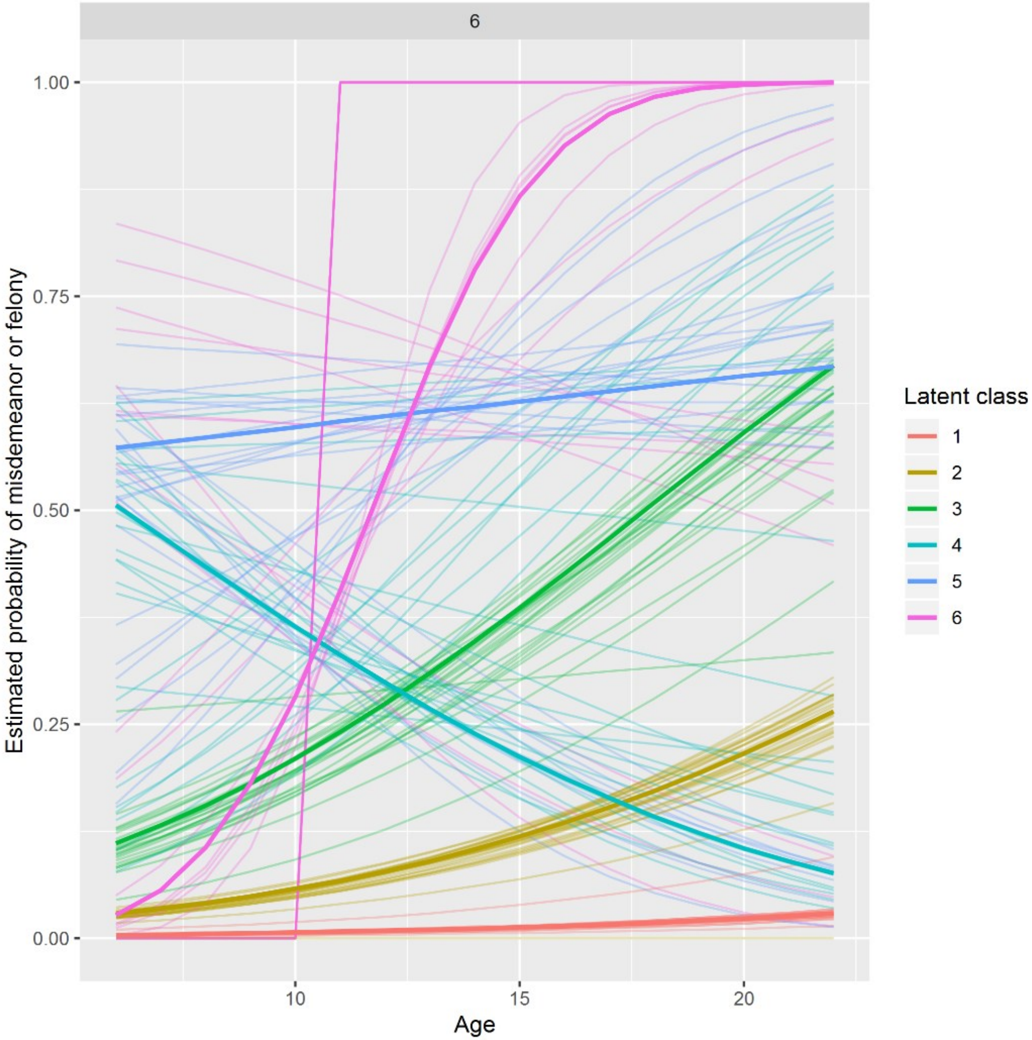
Understanding Inequalities

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL


5
1.00
0.75
0.50
0.25
0.00
Estimated probability of misdemeanor or felony
10
15
20
Age
Latent class
1
2
3
4
5


5
1.00
0.75
0.50
0.25
0.00
Estimated probability of misdemeanor or felony
10
15
20
Age
Latent class
1
2
3
4
5


6
Estimated probability of misdemeanor or felony
1.00
0.75
0.50
0.25
0.00
10
15
20
Age
Latent class
1
2
3
4
5
6



6
1.00
0.75
0.50
0.25
0.00
Estimated probability of misdemeanor or felony
10
15
20
Age
Latent class
1
2
3
4
5
6

Results

- These figures would suggest maybe a three or *possibly* a four-class model?
- Trajectory shapes from five- and six-class models aren't very stable across replications
- … But is it appropriate to move from visualizing uncertainty to model selection?

2001) which more generally fits latent variable models from the structural equation modeling approach. Other free programs include macros `HETMIXED` (Komárek and Verbeke 2002) and `HETNLMIXED` (Spiessens, Verbeke, Komárek, and Fieuws 2004) in SAS (SAS Institute Inc. 2003) that are numerically limited (Proust and Jacqmin-Gadda 2005), and the free Fortran 90 program `HETMIXSURV` (Proust-Lima 2015; Proust-Lima *et al.* 2016) that might be faster but is not user-friendly. To our knowledge, three packages can be used to fit latent class mixed models in R (R Core Team 2017): function `GLMM_MCMC` (Komárek 2009) of package **mixAK** (Komárek and Komárková 2014) fits latent class generalized linear mixed models with possibly multivariate longitudinal data based on MCMC estimation; package **flexmix** (Leisch 2004; Grün and Leisch 2007, 2008) also proposes the estimation of classes of mixture models including the latent class linear mixed model with function `FLXMRlmer`; and **mixtools** (Benaglia, Chauveau, Hunter, and Young 2009) includes a functionality to fit latent class linear models with random effects. In Stata (StataCorp. 2015), the program `GLLAMM` fits some latent class and latent variable models (Rabe-Hesketh, Skrondal, and Pickles 2004). Finally, when one is interested in classification of trajectories, exploratory methods such as the latent class growth analysis (Nagin 1999) are usually preferred. They have especially become very popular in psychology (Bongers, Koot, Van der Ende, and Verhulst 2004) and more recently in public health (Gill, Gahbauer, Han, and Allore 2010). It should be stressed that the latent class growth analysis as implemented in SAS with `PROC TRAJ` (Jones, Nagin, and Roeder 2001) is a specific case of latent class mixed models in which no random effect is included. As such, it assumes that given a specific latent class, the repeated measures of the same subject are independent. Although of possible interest in an exploratory analysis, the assumption that repeated measures are independent given a restricted number of latent groups is very strict and unlikely so inference based on this approach is usually impossible.

Proust-Lima et al. (2017:51), *Estimation of Extended Mixed Models Using Latent Classes and Latent Processes: The R Package lcmm,* Journal of Statistical Software, *Vol. 78, No. 2, pp. 1-56*

https://www.jstatsoft.org/article/view/v078i02/v78i02.pdf

What have we learned?

Summary

- Many models workflow has positives and negatives
- Not so sure the ‘secret weapon’ was useful for me
- It was useful for re-fitting models to bootstrap replications to see what the LCGM was doing

Questions for you

- Is this a useful approach to look at change over time in the LCGMs?
- Is bootstrapping to select number of classes a valuable technique?
  - If so, how many bootstrap replications do you need?
    - For hypothetical outcome plots, ~20
    - To calculate CIs, 999 or 1999
  - Limitation – time and processing power!

Thank you!

@benmatthewsed | ben.matthews@ed.ac.uk

These slides and Racine code at : https://github.com/benmatthewsed/qstep-march-2019

SOI code: email me