Modeling Long-Term Trajectories

Scott M. Lynch
Dept. of Sociology
Dept. of Family Medicine & Community Health (DUMC)
Center for Population Health and Aging (DUPRI)
Duke University

2020 RCCN Workshop

Introduction

- Trajectories are patterns of change over time
- Variety of methods exist for modeling longitudinal data, many of which are sometimes called "trajectory methods."
- But today: focus on methods for repeated measures of same variable (vs. models for transitions or states)
- Two general approaches: Growth Modeling methods (GM) and Latent Class Methods (LC)
- Highlight basic ideas and discuss similarities/differences
- Dispell some myths about the methods

Nomenclature

- GMs and LCs go by a variety of names, some overlapping
  - GM: trajectory models, latent trajectory models, latent curve models, multilevel growth models, random coefficient models, parametric trajectory models, random effects growth models
  - LC: trajectory models, latent trajectory models, group based trajectory methods, nonparametric trajectory modeling, finite mixture models
- Different names emerge from different statistical origins (e.g., “latent” from SEMs; “multilevel” from HLMs)
- So...usually can’t tell what someone has done without looking at methods section.

Data for Examples

- Data are from the Health and Retirement Study (HRS)

  Nationally representative panel study with replenishment of $\sim$35k persons from 1992-

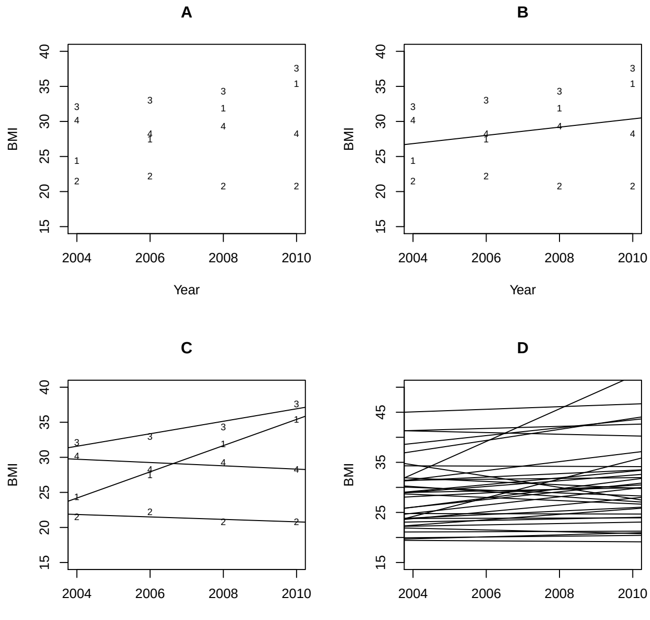
- BMI data on 1951 birth cohort collected in ‘04, ‘06, ‘08, ‘10.

  Restrict to survivors with no missing data ($n = 353$)

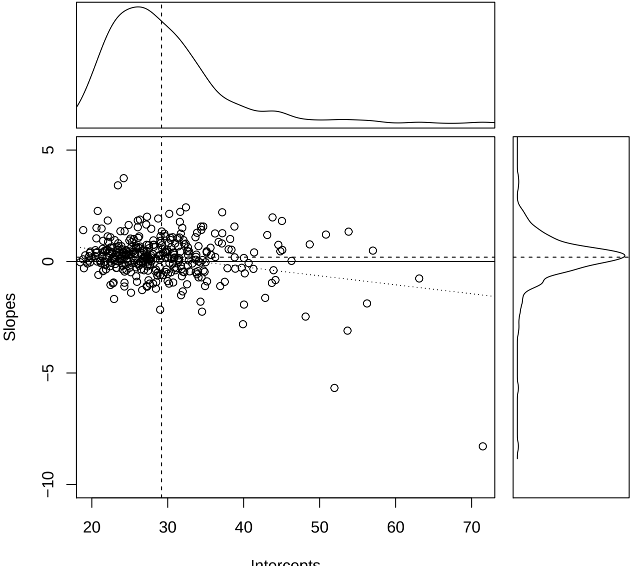
- BMI is kg/m^2; normal$<$25; 25$<$overweight$<$30; 30$<$obese

  Some examples treat BMI as continuous; some as categorical/dichotomous

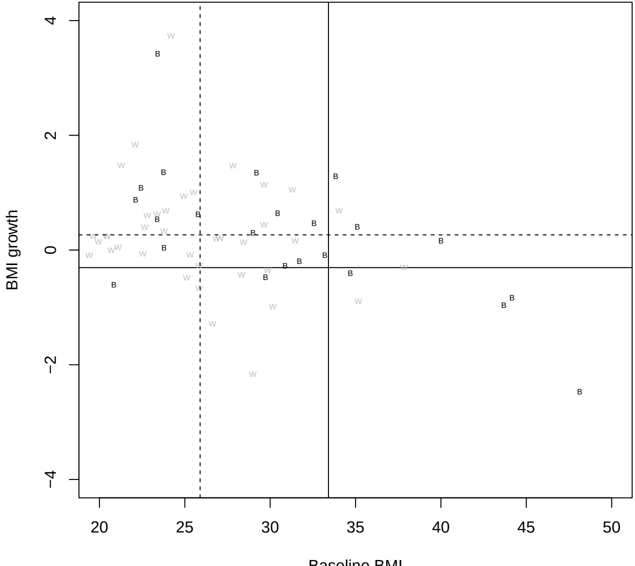
GM Idea: Intercepts (I) and Slopes (S) for All Individuals



Bivariate Distribution of I and S



Black/White Differences in I and S



Basic GC

- Basic OLS model with longitudinal data:

$$
\begin{aligned}
\text{Level 1:} \quad y_{it} &= b_{0i} + b_{1i} t_{it} + b_2 x_{it} + e_{it} \\
\text{Level 2:} \quad b_{0i} &= \gamma_0 + \gamma_1 z_i + u_i \\
b_{1i} &= \delta_0 + \delta_1 z_i + v_i
\end{aligned}
$$

$$
\begin{gathered}
e \sim N(0, \sigma^2) \\
[u, v] \sim MVN(0, \tau)
\end{gathered}
$$

- Reduced form (insert Level 2 into Level 1):

$$
\begin{aligned}
y_{it} &= (\gamma_0 + \gamma_1 z_i + u_i) + (\delta_0 + \delta_1 z_i + v_i) t_{it} + b_2 x_{it} + e_{it} \\
&= b_0 + b_1 z_i + b_2 t_{it} + b_3 z_i t_{it} + b_4 x_{it} + (u_i + v_i t_{it} + e_{it})
\end{aligned}
$$

- So, OLS will work but produce bad s.e.'s because of heteroscedasticity and non-independence of errors

Basic GC, continued

- Can be estimated in hierarchical/random effects framework with data in “long” format (via Stata “mixed” or HLM software)

  In that context, sometimes called a variance components model because of the Level 1 and (Level 2) random effects variances

- ...Or as a multivariate model in an SEM framework, with the random effects as latent variables (hence “latent” growth)

  Data in that framework is in “wide” format

- Missing data handled implicitly in long format, but must be handled via FIML or other means in wide format

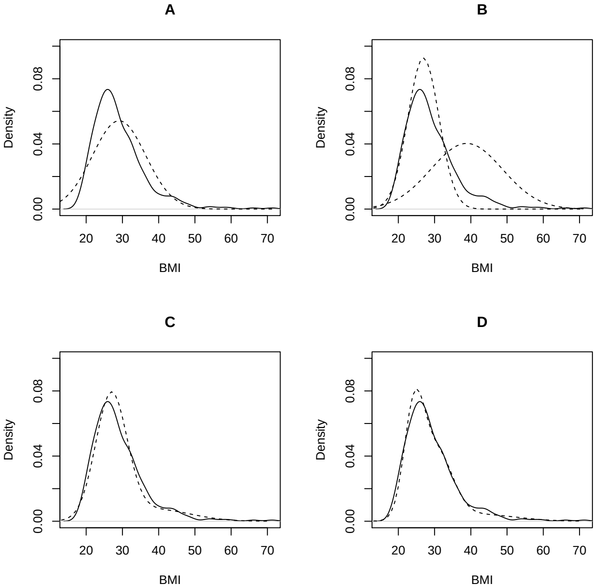
GC Example

Unconditional Model	(Random) Intercept		(Random) Slope
Mean	29.16		.196
Variance	50.5		.40
Correlation (I,S)		-.21	
Conditional Model			
Intercept	34.9*		.13
Male	.71		-.07
Black	2.24*		.003
South	-1.2		-.13
Education	-.46*		.01
RE Vars	$48.1(R^2 = .048)$		$.397(R^2 = .013)$
Correlation (I,S)		-.21	

Latent Class Modeling

- GM (1) assumes parametric trajectory shape and estimates an "average" one, (2) estimates (smooth) variance around it, and (3) estimates covariate effects on trajectory components
- LC (1) does not (necessarily) assume a parametric form for trajectories, and (2) does not assume a smooth distribution of intercepts/slopes, were trajectories parameterized
- Instead: LC assumes population consists of multiplie discrete, (relatively) homogenous "classes"
- Goal is to identify how many distinct classes (using BIC) and use covariates to predict membership in them

Basic Idea of LC: Finite Mixture Distribution



Basic Idea of LC, continued

- Involves finite mixture modeling and can handle more than 1 measure/repeated measure:

$$f(y_{it}) = \sum_{k=1}^{K} f(y_{it}|c_k)p(c_k)$$

with $\sum_{k=1}^{K} c_k = 1$. (proportion of population in each class)

- $f(y_{it}|c_k)$ can be specified to be parametric wrt time, or not. And, the distribution can be anything. e.g.:

$$p(y_{it} = 1|c_k) \sim Bernoulli(p_{kt})$$

or

$$f(y_{it}|c_k) \sim N(\mu_{kt}, \sigma^2_{kt})$$

Bernoulli Ex. (obese/not at each t; $2^4 = 16$ "trajectories")

Classes and Proportions in Each

Class	p(obese1)	p(o2)	p(o3)	p(o4)	% in c_k
"Stable nonO"	.018	.009	.000	.037	57%
"Variable"	.321	.605	.655	.642	13%
"Stable O"	.990	.986	1.00	.982	30%

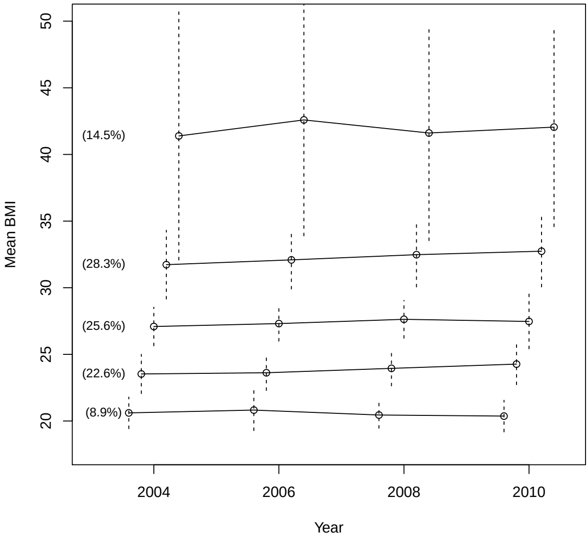
Individual Trajectories and Assigned Classes

n	Sequence	$p(i \in c_1)$	$p(i \in c_2)$	$p(i \in c_3)$	Assumed class
190	0000	.992	.008	.000	1
10	0001	.727	.273	.000	1
4	1000	.824	.176	.000	1
105	1111	.000	.036	.964	3

Bernoulli Example, cont'd

n	Sequence	$p(i \in c_1)$	$p(i \in c_2)$	$p(i \in c_3)$	Assumed class
2	0010	0	1	0	2
7	0011	0	.997	.003	2
4	0100	.422	.578	0	2
3	0101	.016	.984	0	2
5	0110	0	.996	.004	2
9	0111	0	.883	.117	2
2	1001	.092	.908	0	2
4	1011	0	.625	.375	2
3	1100	.027	.973	0	2
1	1101	.001	.999	0	2
4	1110	0	.535	.465	2

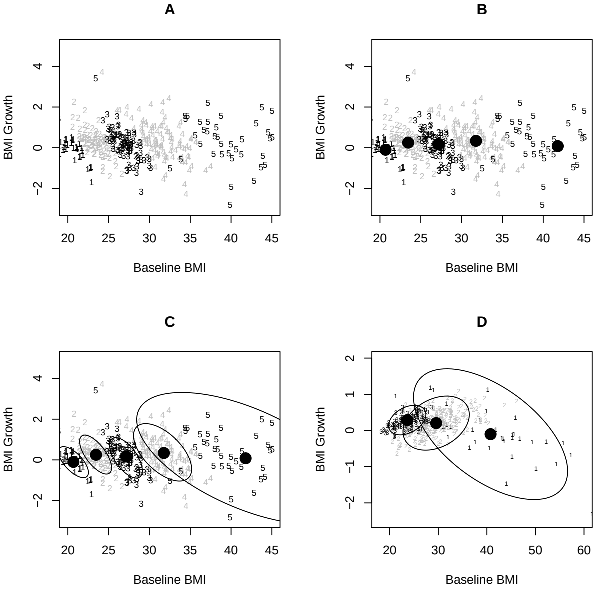
Normal Example: Suggests flat, linear trajectories



Issues to Consider

- Issue 1: Once classes are assigned, usually use multinomial logit to predict membership...
  - ...but there is clear uncertainty in class membership
  - ...but classes aren't necessarily "latent" (e.g., body weight subpopulations are determined by sex—there's nothing latent there). So, should covariates be considered WHILE estimating classes?
- If one assumes parametric shape for trajectories, decision between latent class and growth model is fundamentally whether one believes distribution of intercepts and slopes is smooth or "lumpy" (LCGA vs. GM)
- Key LC assumption is that there is no heterogeneity within classes (unrealistic)
- Assumption can be relaxed, but it becomes dicey. (GMM)

Growth? Latent Class? How Many Classes?



Conclusions

- GM and LC are two main approaches to modeling trajectories in social science aging research
- Neither is fundamentally superior to the other, but they rely on different assumptions about the nature of the population (parametric with noise vs. smooth or lumpy distributions of parametric patterns)
- Growth modeling requires fewer assumptions but may be less satisfying than a "crisp" categorization
- Changing LC assumptions, though, can lead to radically different conclusions, so caution is warranted