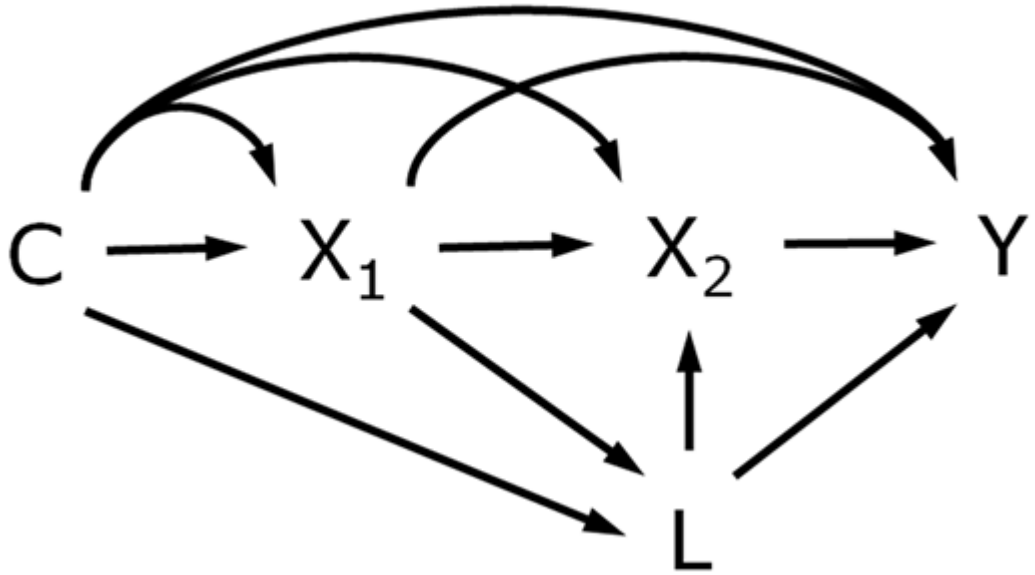
Parametric g-formula for decomposing social life-course processes

- In life-course social epidemiology, people typically start their papers by establishing a DAG grounded in theory. Kind of like the below, but sometimes really big and complicated.



-
    - Demographers and epidemiologists talk about "accumulating risk" and "critical periods." However, they typically then go on to not actually estimate their full DAG in a satisfying way. They fit some sort of GLM, maybe with some latent random effect structure, and interpret the significance of some (in my opinion) confusing and convoluted marginal effects to determine whether different lagged/unlagged exposures are "important."
    - Some people do try to estimate their whole DAG more precisely. They fit a lot of models and multiply coefficients together, which is not generalizable and runs into tons of problems depending on the complexity and collapsibility of your specific model form. Some people use structural equation models (SEMs), which rely on assumptions of linearity and no interactions. All these traditional methods rely on analytic solutions.
- The parametric g-formula allows complex non-linear models, any type of data/exposure, and time-varying confounding by relying on bootstrapping and simulation (MCMC) to fit all parameters.
    - Demographers tend to use simulation methods (multi-state life-tables) to try to model these sorts of processes. I think of these as just very simple mini g-formulas because they treat all the transition probabilities as independent processes across the life-course. Ignoring the potential for causal inference and mediation analyses, I see the parametric g-formula as just the logical extension of multi-state life-table simulations by accounting for the entire conditional probability space defining an individual's transitions throughout their life.
- For now, forget the mediation analyses. The "natural course" is the most important thing to fit correctly. From there, we can just examine the direct/indirect decomposition of effects. In this way, we are essentially just doing really rigorous associational decomposition that I feel pretty comfortable with because we're basically just doing what everyone else does (or posits to do),

but way better using a generalizable simulation approach that we can validate empirically in a lot of different ways. We can spend the whole introduction of different analyses establishing a solid theoretical foundation for why we think our DAG (and thus fitted g-formula simulation) approximates the real-world life-course process. I like this approach because it relies on explicitly specifying a theory-driven model of how you think the life-course unfolds, and then precisely fitting that model rather than simplifying it by employing some more typical, simple analysis.

    - Doing mediation analysis (counterfactuals) with the resulting g-formula fit is still very sensitive to specifying the "correct" model. Just fitting the empirical data with your natural course does not guarantee your model is correct. The g-formula approach was developed in clinical epidemiology to study dynamic treatment regimes *in RCTs controlling for known confounders*. In that setting, I'm way more likely to believe that the model and the parametric relationships are specified more or less correctly and you can interpret the direct and indirect effects of your treatment in a causal way. Right now, I kind of think interpreting these "causal" counterfactual effects for a selective, social process may potentially distract from all the other great contributions of this approach.
    - For our application, I don't even think it's necessary to say "What is the direct/indirect effect of increasing education by 2 years?" I think this is potentially confusing and the results still a bit sketchy. I just want to be able to say "*Given* that our hypothesized DAG is theoretically sound and our fitted g-formula describing that DAG approximates the empirical data very closely, how are changes decomposed into the direct/indirect effects of various predictors like income, education, geography, etc.?"
    - I think of this as a rigorous, generalizable method for doing the type of big multi-state life-table simulation approaches that demographers tend to prefer.
- Of course, it's easy to fit various counterfactuals after we have the natural course fit. But in my thinking right now, this method is potentially super valuable even without doing that.
- We can also start building in many more models to the g-formula framework (i.e. spatial models).