Stat238: Lab 8

October 28

- This lab will be a discussion of model specification for two applied problems provided by class auditors, Melissa (problem 1) and Harm (problem 2) (who will be present for the discussion).
- In class we'll have an opportunity to ask questions of the problem presenters and then I'll have you discuss ideas for model specification in small groups. I'll ask you to focus on
  - the likelihood
  - random effects / latent process specification including any regression-style structure
  - complications that would require a more sophisticated model

Problems

1. The data for this project come from a rural Zimbabwean community. We have tracked four different kinds of changes in households: counts of births, deaths, people moving into a household, and people moving out of a household, for each sample household. We have changes from three time intervals (1986 through 1992, 1992 through 1999, and 1999 through 2010), and the size of each household at the beginning of each interval. Sample households are a subset of the households in each village, and the data are from a number of different villages. We also have the geographical locations of each household. We are interested in the degree to which demographic change (birth/death) versus mobility contributes to change in households over time.

   We also have individual level data (individual age, gender, household membership) for each person in each sampled household. There is an additional wrinkle in that sometimes an individual leaves one household in the sample and goes to a different household in the sample (so those two changes are not independent).

2. I am trying to build a model for predicting media slant for financial news. The prediction is that investors might have heterogeneous preferences for what financial news they want to read. For various reasons people might prefer to read news in accordance with their prior beliefs. There is some empirical support for this happening with political news, where studies found pretty significant media slant in news articles about the same congressional hearings, where the direction of the slant (to the left or right) could be well predicted by the political demography of the zip code where the newspaper circulated. The mechanism is quite clear: If the news market is competitive, then what should happen is that news will target different audiences, audiences that are basically clusters across some preference dimensions (like preference for left or right ideology). Whether this happens for financial news is not quite clear however, because we are not sure to what extent people have heterogeneous preferences for financial news. To analyze these questions, I managed to get hold a large dataset of online news about different company earnings announcements by different news sources.

   What I would like to do is estimate the financial sophistication and degree of optimism for the target audiences for each news source where these quantities are not directly observable but could impact the language of the news posts and whether a given earnings announcement is covered by a post. The observable variables I can construct are noisy measures of writing style and tone of the news posts. On top of that we can observe a bunch of company characteristics which should in theory also affect news worthiness and interact with the sophistication and optimism dimensions of target groups. And I have the text of the earnings announcement as published by the company itself as benchmark for slant. So, the goal of the exercise is to derive some estimate of each news source's unobserved choice of target group in terms of their average sophistication and optimism.