BST 701: Bayesian Modeling in Biostatistics
Breheny

Assignment 4
Due: Tuesday, April 9

1. The data set `cd4.txt` contains information from a study of disease progression in HIV-positive children. HIV infection leads to depletion of a certain kind of lymphocyte known as a CD4+ helper T-cell; the progressive depletion of these cells eventually compromises the immune system. The concentration of these cells, which I will call "CD4 cells" for short, is often used as a marker for the severity of the illness, as is the fraction of total lymphocytes that are CD4 cells. This latter quantity is recorded in the data set and named `CD4Pct`. Typically, HIV-negative individuals have a CD4 percentages of about 40 percent, while HIV-infected people's CD4 percentage are around 25 percent, but can be much lower depending on the progression of the disease.

   The data set contains repeated measurements on a sample of 251 children, each followed up for a period of one and a half years. In addition to `CD4Pct`, the data set also contains the following variables:

   - `ID`: Each patient is assigned a unique ID. Note that there are several observations (rows) per patient.
   - `Visit`: Ideally, patients were supposed to have visits at 1, 4, 7, 10, 13, 16, and 19 months, with 1 month being the initial visit. As is always the case in hospital data like this, the "16-month" visit did not take place exactly 15 months after the "1-month" visit, but `Visit` provides a rough approximation to the elapsed time between visits. Note that not all patients have an observation for each visit.
   - `Date`: The exact date of each visit.
   - `ARV`: Whether the patient was taking any antiretroviral drugs to combat the HIV infection[1]
   - `VisAge`: The age of the patient (in years) at each visit
   - `Treatment`: Whether the patient received Zinc (`Treatment=2`) or not (`Treatment=1`). Some studies have indicated that the addition of zinc and other micronutrients to the diet may reduce immune system problem and slow the progression of HIV infection[2].
   - `BaseAge`: The age of the patient at the initial visit (baseline).

   As a normalizing transformation, Gelman & Hill recommend taking the square root of `CD4Pct` prior as the outcome in a hierarchical linear model. Fit the following four models:

   - A varying-intercept model, with time as an individual-level predictor
   - A varying-intercept, varying-slope model
   - A varying-intercept, varying-slope model with treatment as a group-level predictor

[1] At least, I *think* that is what this means; the documentation for this data set was not as complete as I would have liked

[2] See the above footnote

- Your choice. Modify one of the above models in a way that seems interesting and meaningful to you. For example, you might try to include age at baseline or `ARV` as covariates, or you could try investigating progression models that do not assume a linear trend (for example, estimating separate means at each visit without requiring that they progress linearly).

Write a report containing your analyses of this data. The report should contain at least four sections:

- **Models**: Briefly write out each model in mathematical notation and comment on how each model differs from the previous one.
- **Predictions**: Find a child in your data set that is missing an observation for a visit (or multiple visits). For each model, provide its posterior estimate, interval, and prediction interval for that patient's CD4 count at the missing visit(s).
- **Progression**: Describe the inferences that each model provides in terms of the progression of disease. There are many ways to communicate this, including but not limited to: providing estimated progression lines for a few representative cases, plotting the varying intercept and slopes (if the slopes vary), and forest plots (as we have done in class with the `CIplot` function). Grouping (*e.g.*, having different colors represent different models) and conditioning (*e.g*, having panels in which the same basic figure structure is repeated across several panels, with, say, the subject or the model varying across the panels) are useful tools in communicating these results. Be sure to present inferences from each model and to communicate the ways in which the inferences from each model differ (or how they remain the same).
- **Effect of treatment**: Does Zinc supplementation impact disease progression? In what way?
- **Other (if necessary)**: This depends on your "Model 4". For example, if your model includes baseline age, then you need a section describing its effect (or lack thereof). On the other hand, if your model investigates nonlinear progressions of the disease, then you wouldn't necessarily need a new section, as this would extend the Progression section.