STAT/BIOSTAT 571: Homework 8

To be handed in on Weds March 2nd, in class. Please see 'Chapter 0' of the slides for a summary of how to answer questions appropriately, and the guidelines from 570. Where solutions require use of `R`, summarize your findings in a written answer, and append your **annotated** code, to show what you did. For each question, write up your solution on your own, using **full sentences**.

1. [**Random slopes**] This question uses the models fit on slide 3.67, and variations of them
   (a) For the model with random intercepts and slopes where their covariance is unconstrained (`lme2b`) is the fitted correlation between the random intercepts and slopes random positive or negative? Without using any technical language, interpret the sign of this correlation; what two subject-level characteristics about reaction time and sleep deprivation are more likely to occur together?
   (b) Using the fitted model with independent random slopes and intercepts (`lme2a`) make a graph showing the marginal mean and 2.5% and 97.5% quantiles of possible observations at each of the time points recorded. How do these quantities change if we fit the model where `Days` variable is not centered, but everything else is identical? Based on the data – and assuming everything else about the models is specified correctly – do you think the centered or uncentered version is more appropriate?

2. [**Coverage of Emprical Bayes intervals**] As seen in class, we obtain the predicted random effects $\tilde{\boldsymbol{b}}_i$ as approximately the conditional means (or modes) of the true $\boldsymbol{b}_i$, conditional on the outcomes and covariates. As noted on slide 3.68, using the same approximations we can also obtain the conditional variances of the true random effects. When fitting mixed models using `lmer()` or `glmer()` from the `lme4` package, this approximate conditional standard deviations can be obtained straightforwardly using the `se.ranef()` function in the `arm` package.
   (a) Using simulation, evaluate the proportion P of the n b_i's that are covered, using intervals of the form $\tilde{b}_i \pm 1.96 \times \tilde{s}_i$, where $\tilde{s}_i$ denotes the approximate conditional standard deviation above. Use the LMM where
   $$\begin{aligned} b_i &\sim N(0, \sigma_b^2) \\ Y_{ij}|b_i &\sim N(b_i, \sigma_Y^2), \end{aligned}$$
   with $n = 1000$, $n_i = 5$, $\sigma_b = 1$, $\sigma_y = 6$. As well as stating the average of P over many replicates – which gives the average coverage for all b_i in all experiments – you should also describe the distribution of P, again over many replicates.
   (b) Keen people: repeat a) but change the values of $n_i, \sigma_b^2, \sigma_Y^2$, and report what you find

3. [**Fitting GLMMs**] Do Q9.4 from Jon's book, noting the following;
   - The data is available on the course site – the `epicalc` package is no longer available
   - To code the sine and cosine terms, for time recorded in quarters use `sin(0.5*pi*(time+1))` and `cos(0.5*pi*(time+1))`
   - I have provided `age` in years, not months. As in the book, this variable has also been roughly centered – so negative 'ages' are not data-entry errors
   - As well as implementing particular models, the 'interpret' parts of this question are very important. In these parts, you should describe the parameters being estimated in clear but accurate language

4. [**Diagnostics**] Give diagnostics for the mean model in the GLMM fit in Q3. Say what you are plotting and what, if any, violation of the assumptions is indicated