Announcements

- Homework due as a pdf on Blackboard.
- Reading in Gelman & Hill (G&H):
  - G&H, Chapters 11–13. Please read and try out the ideas in these chapters.
  - We will return to ideas in these chapters (as well as Ch’s 14-15) throughout the mini.
  - In Ch 13, skip material on inverse-Wishart distribution (pp 286–287) for now—premature for us!
- This is a lot of material; you won’t be able to learn it in detail. Try to find the main ideas, and get a sense of where information is in the chapters, in case you need to refer to them later.
- This assignment and the associated data sets are in the “hw02” area of the class website.
- For several exercises below, you may find that you will need the “foreign”, “arm”, “lme4” and “ggplot2” libraries in R.
- This assignment is a little long—sorry! But I think the things you will be doing here are worth thinking about.

Exercises

1. Consider the following multilevel model for data $y_i, i = 1, \ldots, n$, arranged into $J$ groups, $j = 1, \ldots, J$, where each group $j$ has $n_j$ observations:

$$
\begin{align*}
  y_i &= \alpha_j + \epsilon_i, \quad \epsilon_i \sim \text{id} N(0, \sigma^2) \\
  \alpha_j &= \beta_0 + \eta_j, \quad \eta_j \sim \text{id} N(0, \tau^2)
\end{align*}
$$

Prove the following assertions:

(a) If $i \neq i'$ and $j[i] \neq j[i']$, then Corr ($y_i, y_{i'}$) = 0.

(b) If $i \neq i'$ but $j[i] = j[i']$, then Corr ($y_i, y_{i'}$) = $\frac{\tau^2}{\sigma^2}$.

(c) Let $\bar{y}_j = \frac{1}{n_j} \sum_{i \in j} y_i$, the average of all observations in group $j$. Then Var ($\bar{y}_j$) = $\tau^2 + \sigma^2/n_j$.

(d) Suppose we exactly replicate the experiment generating new data $y^*_i$ following the model

$$
\begin{align*}
  y^*_i &= \alpha_j + \epsilon^*_i, \quad \epsilon^*_i \sim \text{id} N(0, \sigma^2) \\
  \alpha_j &= \beta_0 + \eta_j, \quad \eta_j \sim \text{id} N(0, \tau^2)
\end{align*}
$$

so that the group level $\alpha$’s and $\eta$’s (and $\beta_0$) are the same between (*) and (**) [the conditions we are measuring didn’t change] but the new set of $\epsilon$’s are independent of $\eta$’s and $\epsilon$’s [we re-measured, and so we have new measurement error on each observation]. Form the group averages $\bar{y}^*_j$, analogous to $\bar{y}_j$. Then

$$
\text{Corr} (\bar{y}_j, \bar{y}^*_j) = \frac{\tau^2}{\tau^2 + \sigma^2/n_j}.
$$

(This is another interpretation of the reliability coefficient $\frac{\tau^2}{\tau^2 + \sigma^2/n_j}$.)

In all four parts, carefully state any assumptions that you need.
2. G&H Chapter 11, #4.

3. G&H Chapter 11, #2. This problem uses the same CD4 dataset as Chapter 11, #4. Also do
   (e) Extend the model in part (b) to allow for varying slopes for the time predictor, if possible. Compare the
   results of fitting this model with the results in part (b).

   R has a function `read.csv()` that should prove useful. Examples of analyses are available at http://www.stat.columbia.edu/~gelman/arm/examples, but please do not look at these other analyses until you turn this assignment in.

5. G&H Chapter 12, #6. This problem uses the “beauty” data again.

6. G&H Chapter 13, #1. Again, with the “beauty” data set...

7. A certain researcher has been collecting data on waste products from fracking in Pennsylvania. She measured
   various inputs and outputs from roughly 1/10 of the state’s fracking wells over the course of one year; her data
   look like this:
   - \( y_i \): number of gallons of wastewater (which can contain various toxic and/or radioactive compounds) from
     fracking disposed of locally at each fracking well \( i, i = 1, \ldots, 1214 \).
   - \( j[i] \): The identity of the county \( j, j = 1, \ldots, 67 \) that the \( i^{th} \) well is in.
   - \( \text{county}_j \): the name of the \( j^{th} \) county.
   - \( x_{1i} \): the number of wells operated by the company that owns well \( i \) (a measure of experience in fracking)
   - \( x_{2i} \): the amount of natural gas produced by well \( i \)
   - \( z_{1j} \): the number of wells in county \( j \).

   The following `lmer()` model was proposed for this data:
   \[
   y \sim 1 + x_{1i} + x_{2i} + z_{1j} + (1 + x_{2i} | \text{county})
   \]
   (a) Write the model as a variance-components model, using standard mathematical notation for statistical
       models as in class or as in G&H.
   (b) Write the model as a multilevel model (hierarchical linear model), again using notation like that in class or
       in G&H.
   (c) Write the model as a hierarchical Bayes model as in class (treating the level 1 model as a likelihood and
       level 2 as a prior; ignore priors on level 1 and level 2 variances (\( \sigma^2 \) and \( \tau^2 \)'s) and on any fixed effects.