Likelihood Theory and Extensions (BIOS:7110)
Breheny

Assignment 12
Due: Wednesday, December 7

1. *Marginal likelihood for linear mixed models.* This question deals with the mixed model discussed in class during the "Marginal likelihood" lecture, in which paired observations share a random intercept α_i, with the assumption that $\alpha_i \overset{\text{iid}}{\sim} \text{N}(\mu, \tau^2)$.

   (a) Given estimates $\hat{\mu}$ and $\widehat{\beta}$, derive estimators for σ^2 and τ^2. Provide specific formulas, introducing notation as needed. If you introduce a symbol, please define it clearly. The estimators $\hat{\sigma}^2$ and $\hat{\tau}^2$ can be the maximum likelihood estimators, but do not have to be.

   (b) Write a function, `paired_lmm()`, that fits this linear mixed model. For the purposes of this assignment, your function may assume that consecutive observations are paired. The function should work as follows (some code to simulate paired data is provided):

   ```
   n <- 100
   x <- runif(n*2)
   u <- rep(rnorm(n), each=2)
   y <- rnorm(n*2, x+u)
   paired_lmm(x, y)
   ```

   The function should return

   - $\hat{\mu}$
   - $\widehat{\beta}$
   - $\hat{\sigma}^2$
   - $\hat{\tau}^2$
   - The information matrix for the fixed effects (μ and β)

   Hint: To set up a block diagonal matrix, you may wish to use the `bandSparse` function from the `Matrix` package. For the specific covariance structure in this problem, the following code works:

   ```
   V <- bandSparse(n*2, k=0:1,
                   diagonals=list(rep(sig^2+tau^2, n*2),
                                  c(rep(c(tau^2, 0), n), 1)),
                   symmetric=TRUE)
   ```

2. *Logistic regression for case control studies.* Let Y denote a binary outcome and $\mathbf{x}$ a vector of predictors that are thought to affect the probability of Y. Our goal is to estimate odds ratios of the form

$$\text{OR} = \frac{\mathbb{P}(Y = 1|\mathbf{x})/\mathbb{P}(Y = 0|\mathbf{x})}{\mathbb{P}(Y = 1|\mathbf{x}_0)/\mathbb{P}(Y = 0|\mathbf{x}_0)},$$

   where $\mathbf{x}_0$ is a reference individual. For a logistic regression model applied to a prospective simple random sample, these odds ratios are of the form

$$\text{OR} = \exp\{(\mathbf{x} - \mathbf{x}_0)^\top \boldsymbol{\beta}\}.$$

If the data come from a case-control study, however, the contributions to the likelihood are

$$L_i = \begin{cases} p(\mathbf{x}_i|y_i = 1, s_i = 1) & \text{case} \\ p(\mathbf{x}_i|y_i = 0, s_i = 1) & \text{control,} \end{cases}$$

where S_i is a variable indicating whether or not the subject was sampled. Below, let $\tau_1 = \mathbb{P}(s = 1|y = 1)$ and $\tau_0 = \mathbb{P}(s = 1|y = 0)$ denote the sampling fractions.

(a) An implicit assumption in the above setup is that τ_1 and τ_0 are constants that do not depend on the covariates $\mathbf{x}$. In other words, it is important that $p(s = 1|\mathbf{x}, y = 1) = p(s = 1|y = 1) = \tau_1$ and $p(s = 1|\mathbf{x}, y = 0) = p(s = 1|y = 0) = \tau_0$. Comment on this assumption and provide a **specific, realistic** example of a situation in which this assumption would *not* hold as well as what kind of bias it might introduce.

(b) Assuming that the logistic regression model is correct and that the assumption in part (a) holds, show that

$$p(y = 1|\mathbf{x}, s = 1) = \frac{\exp(\tilde{\eta})}{1 + \exp(\tilde{\eta})},$$

where $\tilde{\eta} = \log(\tau_1/\tau_0) + \mathbf{x}^\top \boldsymbol{\beta}$. In other words, using the prospective likelihood for a case-control study is correct up to the factor $\log(\tau_1/\tau_0)$, which only affects the intercept and therefore does not affect the odds ratio estimates. Note that the prospective likelihood here is a pseudo-likelihood, in the sense that it does not represent the actual likelihood of the experiment.

(c) Suppose the ratio τ_1/τ_0 were known. Given the pseudo-MLE of the intercept, $\widetilde{\beta}_0$, from fitting the pseudo-likelihood model, how can we estimate the true intercept β_0 (and thereby estimate probabilities in addition to odds ratios)?

(d) Suppose that a case-control study of a single exposure was carried out with an equal number of cases and controls, and that the study obtained the estimates $\widehat{\beta}_0 = -1/2$, $\widehat{\beta}_1 = 1$. In the actual population, however, the prevalence of the disease is known to be 1%. Based on this study and its results, provide two sets estimates for the probability of disease for exposed and unexposed individuals – the raw/naive predictions from the model itself and the adjusted estimates you would obtain using your derivation from part (c).