Likelihood Theory and Extensions (BIOS:7110)
Breheny

Assignment 12
Due: Wednesday, December 11

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 $\alpha_i \overset{\text{iid}}{\sim} \mathrm{N}(\mu, \tau^2)$.

   (a) Given estimates $\hat{\mu}$ and $\widehat{\boldsymbol{\beta}}$, 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(X, y)`, 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; here $\boldsymbol{\beta}$ has dimension 2 but your function should work for any dimension):

```
n <- 100
X <- cbind(runif(n*2), runif(n*2))
u <- rep(rnorm(n), each=2)
y <- rnorm(n*2, X[,1]+u)
paired_lmm(X, y)
```

   The function should return a named list of:

   - `mu`: $\hat{\mu}$
   - `beta`: $\widehat{\boldsymbol{\beta}}$
   - `sigma`: $\hat{\sigma}$
   - `tau`: $\hat{\tau}$
   - `info`: 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. *Penalized logistic regression.* Suppose we have trios of binary outcomes Y_{i1}, Y_{i2}, Y_{i3} and we wish to fit the probability model

$$\log \frac{\pi_{ij}}{1 - \pi_{ij}} = \alpha_i + x_{ij}\beta,$$

where $\pi_{ij} = \mathbb{P}(Y_{ij} = 1)$; in other words, each trio has its own intercept, but we assume a common effect for the covariate x. We have seen that maximum likelihood with so many parameters is often inaccurate; thus, consider the penalized logistic regression model in which we estimate $\boldsymbol{\alpha}, \beta$ by minimizing the negative log-likelihood from the traditional logistic regression model plus the penalty

$$p(\boldsymbol{\alpha}) = \frac{\lambda}{2} \sum_{i=1}^{n} \alpha_i^2;$$

in other words, we apply a ridge penalty to the $\{\alpha_i\}_{i=1}^n$ parameters, but no penalty to β.

(a) Derive the penalized score and information (i.e., the first and second derivatives of the objective function).

(b) Describe an algorithm to obtain the penalized MLE $\widehat{\beta}$. There are multiple acceptable answers, but please keep your answers consistent; whatever algorithm you describe in (b), you should implement in (c).

(c) Implement this algorithm to obtain both the penalized MLE and Wald test for $H_0 : \beta = 0$ using the penalized information. Generate data using this code:

```
set.seed(1)
n <- 30
m <- 3
x <- runif(n*m)
u <- rep(rnorm(n, sd=2), each=m)
y <- rbinom(m*n, size=1, prob=binomial()$linkinv(x + u))
```

You do not need to turn in your code – just report the estimate and p-value when $\lambda = 1$.

(d) Carry out a simulation in which you generate $N = 1,000$ independent data sets according to the mechanism given above. For each data set, fit a penalized logistic regression model these 51 different values of λ:

```
lam <- exp(seq(log(0.25), log(3), len=51))
```

Finally, produce a plot with λ on the horizontal axis and the mean squared error of $\widehat{\beta}(\lambda)$ on the vertical axis. Comment on the figure and what it tells you about how penalized logistic regression might compare to other methods for this analyzing this data.