BIODS215 Problem Set 1 due Tuesday, May 9, 2016

Regression: The Linear Model

This is about the linear model applied to data from wearable biosensors from Li, Dunn, Salins, et al. (2017) PLoS Genetics ("Digital Health: Tracking Physiomes and Activity Using Wearable Biosensors Reveals Useful Health-Related Information"). We will subset the data to study the relationship between activity-related parameter, including: acceleration forces caused by movement and their impact on physiological response variable skin temperature.

We will work on building a Bayesian Linear Mixed Model from a fixed effects model to a varying intercepts model and finally to a varying intercepts, varying slopes model.

a) Load data (parta.txt) using RStan (visit biods215/biods215.github.io/PSET1 in Github).

b) **Fixed Effects Model (Simple Linear Model)**. We begin by making the assumption that the physiological response variable **skin temperature** is approximately normally distributed. It has some unknown grand mean β_0 . The mean of the normal distribution of skin temperature is the sum of β_0 and adjustments $\beta_1 \times \text{acceleration magnitude}$.

i) Express the model.

ii) Assume that ϵ_i are independently and identically distributed as a normal distribution with mean zero and unknown standard deviation σ_e. Stan parameterizes the normal distribution by the mean and standard deviation, and we follow that convention here, by writing the distribution of ϵ as $\mathcal{N}(0, \sigma_e)$ (the standard notation in statistics is in terms of mean and variance). Independence implies that there should be no correlation between the errors—this is not the case in the data, since we have multiple measurements from each subject.

iii) Fit the fixed effects model using RStan.

iv) Evaluate model convergence and summarize the results.

c) **Varying Intercepts Mixed Effects Model**. The Fixed Effects Model in b) is inappropriate for the data when we have multiple measurements for each subject. As mentioned above, these multiple measurements lead to a violation of the independence of errors assumption. Moreover, the fixed effects coefficients β_0 and β_1 represent means over all subjects and neasurements, ignoring the fact that some subjects will have higher skin temperatures and some lower skin temperatures than average.

We now express the skin temperature, which was produced by subjects $j = 1, \ldots, J$; measurement number $k = 1, \ldots, K$; by adding adjustment terms u_{0j} to take the by-subject variability into account. We assume that these adjustments are normally distributed around zero with unknown standard deviation: $u_0 \sim \mathcal{N}(0, \sigma_u)$. We now should have two sources of variance in this model: the standard deviation of the errors σ_e, and the standard deviation of the by-subject random intercepts σ_u, which we can refer to as variance components.

Notice that we are now using a slightly different way to describe the model, compared to the fixed effects model. We are using indices for subject and measurement to identify unique rows of the data.

Repeat Steps i, iii, and iv in b).

d) **Varying Intercepts, Varying Slopes Mixed Effects Model.** The first change is to let the size of the effect for the predictor so vary by subject. The goal here is to express that some subjects exhibit greater effects for some predictors than others. We let effect size vary by subject by including in the model by-subject varying slopes which adjust the fixed slope β_1 in the same way that the by-subject adjust the fixed intercept β_0. This adjustment of the slope by subject is expressed by adjusting β_1 by adding a term u_{1j}.

Repeat Steps i, iii, and iv in b).

Sampling algorithms for inference

Implement importance sampling to draw 100 (weighted) samples from a mixture of two Gaussians: $0.5\mathcal{N}(0, 1) + 0.5\mathcal{N}(1, 1)$. What is the proposal distribution you sample from? Choose some functions f(x). How close is the importance weighted sample estimate to the actual $\mathbb{E}[f(x)]$?

Permutation testing and monte carlo for p-value and FDR computation

Consider a gene with three exons of equal length L enumerated 1,2,3. Suppose there are exactly 2 isoforms of the gene, isoform 1 contains exons 1,2,3 and isoform 2 contains exons 1,3 skipping exon 2. Suppose single-end sequencing reads are used to estimate the expression of the isoforms, and the read length is $L > r > 0$.

For $i = 1, 2, 3$, define n_i as the number of reads from exon i, and n_{ij} as the number of reads observed that map to the junctional sequence between eoxn i and j.

Assume that each RNA isoform generates a sequenced read at each position across the isoform uniformly at random. Assume reads from isoform i are observed with parameter θ_i. Write the likelihood and solve for the MLE of θ_1, θ_2 when n_{ij} are ignored in the likelihood and when they are included.

Simulation approach using the code here: https://pachterlab.github.io/kallisto/download (https://pachterlab.github.io/kallisto/download)

If you prefer, design a simulation of the above sequencing experiment: choose a fixed value θ_1, θ_2 and simulate Poisson reads using a simulation package or your own scripts from these two isoforms according to the model above. For 1000 simulations, run kalisto and output the distribution of point estimates of of θ_1 and θ_2 as a function of four read lengths.