Design Document for Relevance-Weighting: R Package glmm

Sydney Benson

April 7, 2019

Abstract

This design document will give an overview of the changes made to the R package `glmm` with respect to a relevance-weighted likelihood method. We use relevance-weighting to better reflect the real-world occurrence of more or less informative observations.

1 Introduction

This project is meant to enable the user of the `glmm` function in the `glmm` R package to include an optional relevance-weighting scheme. A common assumption of linear models is that each observation in a data set is equally informative and trustworthy; however, in real-world data sets, this is rarely the case. Thus, the optional relevance-weighting scheme will allow the user to place a heavier weight on the more informative and/or trustworthy observations in their data set so that those data points that are less informative affect the model to a lesser degree.

2 The Process

First, the function will need to establish whether the user has supplied a proper weighting scheme. Next, the weighting scheme will need to be applied in the `el.C` function. After defining the weighting vector, the remainder of this section will illustrate how this weighting scheme will be applied.

2.1 The Weighting Vector

This vector, called Λ, must be a vector with the same length as the response vector and must contain all positive values or zeros.

2.2 Weighted Log-Likelihood

We begin by defining the canonical link function as

$$g(\mu) = \eta = X\beta + Zu \tag{1}$$

Additionally, the likelihood is defined as

$$l_m(\theta|y) = \log\left(\frac{1}{m}\sum_{k=1}^{m}\frac{f_\theta(u_k,y)}{\tilde{f}(u_k)}\right) \tag{2}$$

and $f_\theta(u_k, y) = f_\theta(y|u_k)\tilde{f}_\theta(u_k)$. We then define $f_\theta(y|u_k)$ as $f_\theta(y|u_k) = \exp\left(\sum_i y_i\eta_i - c(\eta_i)\right)$. Thus,

$$f_\theta(u_k, y) = \exp\left(\sum_i y_i\eta_i - c(\eta_i)\right)\tilde{f}_\theta(u_k) \tag{3}$$

and

$$\Lambda f_\theta(u_k, y) = \Lambda \exp\left(\sum_i y_i\eta_i - c(\eta_i)\right)\tilde{f}_\theta(u_k) \tag{4}$$

$$= \lambda_i \exp\left(\sum_i y_i\eta_i - c(\eta_i)\right)\tilde{f}_\theta(u_k) \tag{5}$$

2.3 Integrating the Weighted Log-Likelihood

As shown above, the weighting scheme must be accounted for in $f_\theta(u_k, y)$. Thus, the weighting scheme must be applied within the `el.C` function eventually. However, before we get there, we must follow these steps:

1. Write the test for the weighting scheme using the Booth Hobert data set.
2. Code a general version of the weighting scheme.
3. Re-code the general version of the implementation of the weighting scheme in C.