Design Document for Relevance-Weighting: R Package glmm

Sydney Benson

April 11, 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

The weights must be in the form of a vector and the vector must be the same length as the response vector, y. All weights must be numeric and non-negative. There can be no missing weights.

2.2 The Relevance-Weighted Log Density

Following Hu and Zidek (1997), the relevance-weighted likelihood is defined as

$$\text{REWL}(\theta) = \prod_{i=1}^{n} f(y_i|\theta)^{\lambda_i} \tag{1}$$

where n is the length of the response vector, y. Since we define $\log f_\theta(y|u_k)$ as $\sum_i y_i\eta_i - c(\eta_i)$, where $\eta = X\beta + Zu$, $\log f_\theta(y_i|u_k) = y_i\eta_i - c(\eta_i)$ and $f_\theta(y_i|u_k) = \exp(y_i\eta_i - c(\eta_i))$. So, the relevance-weighted log density becomes

$$\text{RWLD}(\theta) = \prod_{i=1}^{n} \left[\exp\left(y_i\eta_i - c(\eta_i)\right)\right]^{\lambda_i} \tag{2}$$

$$= \prod_{i=1}^{n} \exp\left[\lambda_i\left(y_i\eta_i - c(\eta_i)\right)\right] \tag{3}$$

Then,

$$\log \text{RWLD}(\theta) = \log \prod_{i=1}^{n} \exp\left[\lambda_i\left(y_i\eta_i - c(\eta_i)\right)\right] \tag{4}$$

$$= \sum_{i=1}^{n} \log \exp\left[\lambda_i\left(y_i\eta_i - c(\eta_i)\right)\right] \tag{5}$$

$$= \sum_{i=1}^{n} \lambda_i\left(y_i\eta_i - c(\eta_i)\right) \tag{6}$$

2.3 The First Derivative

Remember that the derivative of the log density of the data with respect to one component, η_j, is

$$\frac{\partial}{\partial \eta_j} \log f_\theta(y|u) = y_j - c'(\eta_j). \tag{7}$$

Consequently,

$$\frac{\partial}{\partial \eta_j} \log \text{RWLD}(\theta) = \lambda_j\left[y_j - c'(\eta_j)\right]. \tag{8}$$

and the derivative of the component η_j with respect to one of the fixed effect predictors, β_j, is

$$\frac{\partial \eta_j}{\partial \beta_l} = X_{jl} \tag{9}$$

Then, we can use the chain rule to find that

$$\frac{\partial}{\partial \beta_l} \log \mathrm{RWLD}(\theta) = \lambda \; X \left[y - c'(\eta) \right] \tag{10}$$

2.4 The Second Derivative

Similar to the first derivative, we find that the second derivative of the relevance-weighted log density of the data is

$$\frac{\partial^2}{\partial \beta_l^2} \log \mathrm{RWLD}(\theta) = \lambda \; X' \left[-c''(\eta) \right] X \tag{11}$$