

Classification using a joint model of longitudinal data and binary outcomes based on the SAEM algorithm

*Cristian Meza*¹, *Maritza Márquez*², *Rolando de la Cruz*³, *Claudio Fuentes*⁴

¹cristian.meza@uv.cl, INGEMAT-CIMFAV, Universidad de Valparaíso, Chile

²maritza.marquez@edu.uai.cl, Faculty of Engineering and Sciences, Universidad Adolfo Ibáñez, Chile

³rolando.delacruz@uai.cl, Faculty of Engineering and Sciences, Universidad Adolfo Ibáñez, Chile

⁴fuentesc@oregonstate.edu, Department of Statistics, Oregon State University, USA

Despite the pervasiveness of binary outcomes in the vast scientific literature, joint models for longitudinal and binary data are not standard. The idea of jointly modeling these data, which share common parameters or covariates that may link them, is because they provide a practical way of evaluating the association between these two types of data, which, in clinical studies, are usually collected jointly on a series of individuals. In this work, we propose a new estimation method based on a stochastic approximation version of the EM algorithm, the so-called SAEM algorithm, to estimate the parameters of a joint model based on an (non)linear mixed effects model for the longitudinal part and a generalized linear model (GLM) for the primary response of interest. We applied our method to two different sets of data from the monitoring of pregnant women. The first dataset that drove our research proposal corresponds to a follow-up study carried out on 173 young pregnant women during the first trimester of pregnancy, in which the concentration values of the β -hCG hormone were measured. At the time of birth, the women were classified into two groups: a normal group, in which those women who had a normal delivery were considered, and an abnormal group, in which only those women who had any complication that would result in a non-terminal delivery along with the loss of the fetus were considered. According to France et al. (1996), they assured that the hormone concentration values vary from women who have normal pregnancies with terminal deliveries to women who have spontaneous abortions or other types of adverse pregnancy outcomes. This association has made it possible to predict (with some uncertainty) the pregnancy outcomes. Then, for this dataset, we analyze the β -hCG hormone levels measured in early pregnancy that can be used to predict normal versus abnormal pregnancy outcomes. The resulting joint model allows us to carry out a classification process that, from this point of view, improves on previous studies. In the second case, we are interested in the predictions of post-molar gestational trophoblastic neoplasia (GTN) studied by Dandis et al. (2020). We want to provide updated predictions of postmolar GTN based on age and repeated measurements of human chorionic gonadotropin (hCG). A model for a binary outcome is required since it is unknown when GTN development began. The joint model considered here uses as predictors in a logistic regression model with the status of GTN as the outcome the random effects from the first model, which models the hCG values via a mixed effects model. For this dataset, the use of the longitudinal data improves the predictive ability of the logistic regression model compared to the reduced model using only the age of the pregnant woman as a baseline covariate.

Keywords: Joint model; Mixed effects models; SAEM algorithm