# AUC estimation in logistic regression with missing data: a case study

*Susana Rafaela Martins*[1], *Jacobo de Una-Alvarez*[2], *Maria del Carmen Iglesias-Perez* [3]

[1]srgm@estg.ipvc.pt, Escola Superior de Desporto e Lazer, Instituto Politecnico de Viana do Castelo & SiDOR research group, Universidade de Vigo

[2] jacobo@uvigo.es, Department of Statistics and OR & CINBIO, Universidade de Vigo

[3] mcigles@uvigo.es, Department of Statistics and OR & CINBIO, Universidade de Vigo

The obesity is defined by the excessive accumulation of fat and the consequent excess weight, which translate into a risk to the individual's health. According to the WHO, the levels of overweight and obesity continue to increase, particularly in young people and children between 5 and 19 years of age. Globally, in 2016 the prevalence levels of overweight and obesity were around 18%. This problem has a solution, however its prevention is the best option. In this sense, and taking advantage of data from a previously existing study, we decided to study the possibility of building predictive models that allow monitoring obesity levels. The aforementioned study was carried out with children from the municipality of Viana do Castelo, in Portugal. The initial data was collected in 1997, and annually until the year 2000, with a subsequent collection in 2007. Naturally, one of the problems that arise in this type of study is the existence of missing data.

In this work that we present, we try to monitor the levels of childhood obesity and overweight using the logistic model to predict them through covariates related to performance in physical test in addition to gender and previous levels of overweight. The importance of this study is related to the fact that the results of the physical tests can be easily collected by any sports teacher in a school context and, consequently, it can be easy to screen children who may be at risk of having overweight or obesity. To study the predictive capacity of the logistic model, the area under the ROC curve (AUC) was used.

In this work we investigate the issue of estimating the AUC in presence of missing data. A simulation study is carried out to compare the performance of several approaches: Complete Case Analysis, Inverse Probability Weighting and Multiple Imputation. We also take into account the problem of the optimistic estimation of the AUC.

**Keywords:** Obesity, prediction, ROC curve