Model selection for the estimation of prevalence of smokers in small areas in Galicia

Esther López Vizcaíno[1], *María José Lombardía Cortiña*[2]

[1]esther.lopez@ige.eu, Instituto Galego de Estatística [2]maria.jose.lombardia@udc.es, Departamento de Matemáticas, Universidade da Coruña

In Galicia, like in other European countries, the estimation of some socioeconomic indicators (unemployed, health, poverty,...) is made by means of surveys that most municipalities and another local areas are not represented in the sample and many of them are present with a very small sample size. In this situation the sample size could be enlarged but this can cause delays in obtaining results and the impact of non-sampling errors. Then, the increase of the sample size is not always advisable and even sometimes unfeasible from an economic point of view. But frequently, auxiliary variables exists that are correlated with the variable of interest and several estimators can make use of auxiliary information. The small area estimation techniques deal with this kind of situations.

On the other hand, the question of model selection has received a lot of attention in the literature in the past (starting with the well-known paper by Akaike (1973)) and also in recent years due, among other reasons, to the increasing complexity of modeling approaches. In particular, the question have received considerable attention in the context of linear mixed models although not so much in small area problems. One of the most popular approaches to model selection is to use the AIC (Akaike Information Criteria).

The objective of this work is to compare several models to estimate the prevalence of smokers in small areas in Galicia. We use different models that go from linear mixed models to non parametric models.The domains of interest are the 53 counties of Galicia and the biggest municipalities.

Keywords: small area estimation, AIC, model selection.