**A new statistical model to assess the burden of misreported epidemiological data**

David Moriña(1), Amanda Fernández-Fontelo(2), Alejandra Cabaña(2), Argimiro Arratia(3), Pedro Puig(2).

(1) Departament d’Econometria, Estadística i Economia Aplicada, Universitat de Barcelona; (2) Departament de Matemàtiques, Universitat Autònoma de Barcelona; (3) Department of Computer Science, Universitat Politècnica de Catalunya.

**Background/Goal**

The Covid-19 pandemic that is hitting the world since late 2019 has made evident that having quality data is essential in any decision-making chain. Several methodological approaches have been recently proposed although there is a lack of flexible continuous time series models capable of dealing with misreporting. A typical limitation of this kind of models is the computational effort needed to properly estimate the parameters, so the main goal of this work is to propose a new statistical model capable of detecting misreported observations in Spanish Covid-19 data, incorporate the impact of covariates such as the impact of the vaccination strategy and nevertheless with an efficient parameter estimation method.

**Methods**

AutoRegressive Conditional Heteroskedasticity (ARCH) models are a well-known approach to fit time series data where the variance error is believed to be serially correlated. We consider the actual unobservable Covid-19 incidence as an AR(1) process with ARCH(1) error structure, and the reported process is a part of it, expressed as , where is the frequency of underreporting and represents its intensity.

**Results**

All Spanish Autonomous Communities (CCAAs) reported weekly cases underestimating the real incidence of the disease in almost all the considered time period (23/2/2020-27/2/2022), with differences in the intensity of this underreporting issue among them. For instance, only 44% of the actual cases were reported in Aragón while 69% of the actual cases within the study period were reported in Comunidad Valenciana. The impact of vaccination on the Covid-19 incidence was also unequal, although a significant reduction was detected in all CCAAs after reaching 50% of population with at least one dose (22/6/2021).

**Conclusions**

Although it is very common in biomedical and epidemiological research to use data from disease registries, there is a concern about their reliability, and there have been some recent efforts to standardize the protocols seeking to improve their accuracy. However, it is not always possible to implement these recommendations in a proper way. The proposed methodology can deal with misreported data in a very natural and straightforward way and is able to reconstruct the most likely hidden process, providing public health decision-makers with a valuable tool able to predict the evolution of the disease under different scenarios.