Size Estimation - Statistical Models for Underreporting

Gerhard Neubauer[1,*], **Gordana Djuraš**[1], **Herwig Friedl**[2]

1. Institute of Applied Statistics, JOANNEUM RESEARCH, Graz, Austria
2. Institute of Statistics, Technical University, Graz, Austria

* Contact author: gerhard.neubauer@joanneum.at

Keywords: Binomial Model, conditional Poisson models, regression

Underreporting is a problem in data collection, when events are counted and for some reason errors occur. The most prominent example is crime reporting, where crimes associated with shame are likely not to be reported to the police, just as theft of low value goods. The same holds for traffic accidents with minor damage. And also counting infectious diseases like HIV may be subject to underreporting.

As a consequence the mean of the observed counts is smaller than the true mean λ. Using a Binomial model the mean of the observed counts is $\mu = \lambda\pi$, with π the reporting probability, and both parameters to be estimated. Neubauer & Friedl (2006) introduced a regression approach for the Binomial model and - to adopt for overdispersion - also for a Beta-Binomial model. The Binomial and a Beta-Binomial regression model are suited for a wide range of applications. However, if the sample variance is larger than the sample mean the binomial approach fails to give reasonable estimates. For this kind of data Neubauer & Djuraš (2008) proposed a regression model based on the Generalized Poisson distribution. This model allows to handle Poisson under- and overdispersion, as it covers the binomial, Poisson and the negative binomial case. Recently Neubauer & Djuraš (2009) proposed a further extension of the binomial approach leading to a Beta-Poisson regression model.

A second approach to underreporting builds upon conditional Poisson models, i.e. $Y|L \sim Poisson(L)$ or $Y|L \sim Poisson(L\pi)$. Here the limit $\pi \to 1$ does not cause a problem as with the binomial approach, where $Y \to \lambda$. Using different distributional assumptions for L we obtain a variety of models for possibly perfect reporting systems.

Inference in all cases is based upon maximum likelihood estimation. The scope of the R implementation in package `sizEst` is to cover all mentioned models in a framework, where estimation, testing and model selection is enabled. The approach is illustrated with examples from real data.

References

Neubauer, G. and Djuraš, G. (2008). A Generalized Poisson Model for Underreporting. In: Proceedings of the 23rd International Workshop on Statistical Modelling, 7-11 July, 2008, Utrecht, Netherlands.

Neubauer, G. and Djuraš, G. (2009). A Beta-Poisson Model for Underreporting (Tech. Rep. No.1). Graz: Joanneum Research.

Neubauer, G. and Friedl, H. (2006). Modelling sample sizes of frequencies. In: Proceedings of the 21st International Workshop on Statistical Modelling, 3-7 July 2006, Galway, Ireland.