# Subsampling based variable selection for generalized linear models

**Marinela Capanu[1], Mihai Giurcanu[2], Colin B. Begg[1], and Mithat Gönen[1]**

[1]*Memorial Sloan Kettering Cancer Center, New York, USA*
[2] *Department of Public Health Sciences, University of Chicago, Illinois, USA*

## Abstract

In earlier work we introduced a novel variable selection method for low dimensional linear models involving repeatedly splitting the data to establish an optimal variable selection cutoff. Building on this approach, in this article we adapt our strategy for the generalized linear model setting. We propose repeatedly subsampling the data, minimizing the Akaike's Information Criterion (AIC) over a sequence of nested models for each subsample, and including in the final model those predictors selected in the minimum AIC model in a large fraction of the subsamples. We name this novel approach AIC OPTimization via Subsampling (OPTS-AIC). We also introduce new techniques which involve optimization of the screening threshold over repeated subsamples. In an extensive simulation study examining a variety of proposed variable selection methods we show that, although no single method uniformly outperforms the others in all the scenarios considered, OPTS-AIC enjoys superior performance compared to candidate methods in many settings. We illustrate the methods by applying them to logistic and Poisson regressions and discuss extensions to the high-dimensional setting.

## Keywords

AIC, regression, screening threshold, subsampling, variable selection.

## References:

Capanu, M., Giurcanu, M., Begg, C. B. and Gönen, M. (2020). Optimized variable selection via repeated data splitting. *Statistics in Medicine. 39*, 2167-2184.