Design and analysis of follow-up studies with genetic component

Juha Karvanen
National Public Health Institute, Helsinki, Finland

In gene-disease association studies, the cost of genotyping makes it economical to use a two-stage design where only a subset of the cohort is genotyped. At the first-stage, the follow-up data along with some risk factors or non-genetic covariates are collected for the cohort and a subset of the cohort is then selected for genotyping at the second-stage. The case-cohort design and the nested case-control design are examples of two-stage designs that are commonly used in epidemiological follow-up studies. The data from a two-stage study can be analyzed as a missing data problem where the genotype data are missing by design for the majority of the cohort. The parameters of the data model, typically logistic model or proportional hazards model, can be estimated by maximizing the full likelihood of the data, which in general case becomes an integral over the missing observations. When dealing with single nucleotide polymorphism (SNP) data, the integrals are replaced by sums over the possible genotypes. As a consequence, the likelihood can be directly maximized by numerical optimization, e.g. by R function `optim`.

The straightforward implementation of full likelihood analysis makes it possible to consider alternative designs for the second stage. One such alternative is the extreme selection where cases and non-cases are selected for genotyping starting from those with largest and smallest covariate values. Another alternative is the D-optimal design, which maximizes the determinant of the Fisher information matrix of the parameters. The determination of the D-optimal design requires the use of heuristic algorithms, which is illustrated in Figure 1.

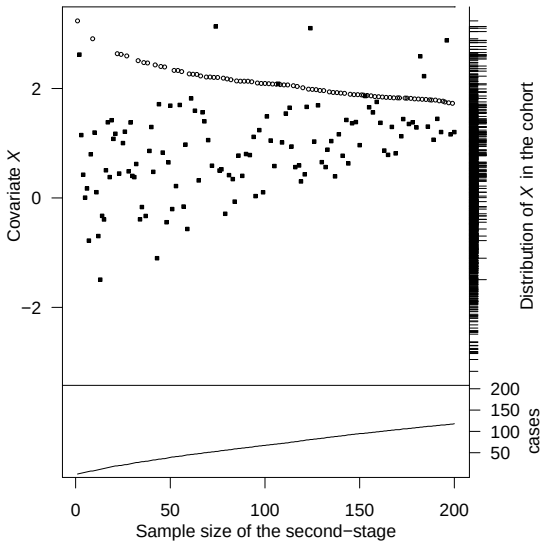

Figure 1: Sequential selection of observations to be included in the second-stage when the response is time-to-event. In the upper panel, sample size on the x-axis indicates the order in which the observations are included. Non-cases are marked by circles and cases are marked by squares. The y-axis on the left presents the covariate values of the selected observations. The tick-marks on the y-axis on the right present the distribution of the covariate values in the whole cohort. The longer tick-marks correspond to cases and the shorter tick-marks correspond to non-cases. In the lower panel, the number of cases selected is shown as a function of the second-stage sample size.

References

J. Karvanen, S. Kulathinal, D. Gasbarra (2008). Optimal designs to select individuals for genotyping conditional on observed binary or survival outcomes and non-genetic covariates. *Computational Statistics & Data Analysis*, doi:10.1016/j.csda.2008.02.010.