Design and analysis of follow-up studies with genetic component

Juha Karvanen

Department of Health Promotion and Chronic Disease Prevention
National Public Health Institute
Finland

Problem definition

- Genotyping is expensive.
- In large epidemiological cohort studies it is economical to genotype only a subset of the cohort.
    1. First stage: some non-genetic covariates and the disease outcome are recorded for a cohort.
    2. Second stage: a subset of the cohort is genotyped.
- How the individuals for genotyping should be selected?

Study designs for two-stage studies

- simple random sampling
- case-control design
- nested case-control design
- case-cohort design
  - select all cases (rare disease assumption) and a random sample of the cohort (subcohort).
- extreme selection
  - Individuals with highest and lowest covariate values are selected.
  - For example, select 100 old cases, 100 old controls, 100, young cases and 100 young controls.
  - Optimal under linear regression model (Elfving, 1952)
- D-optimal design

Inference and missing data

- Genotyping only a part of the cohort can be understood as a missing data problem (missing by design).
- Sampling distribution inference
  - Observations with complete data represent the whole cohort when appropriately weighted.
- Full likelihood inference
  - All observations are included. Likelihood is an integral over the missing data.
  - "*When making direct-likelihood or Bayesian inferences about θ, it is appropriate to ignore the process that causes missing data if the missing data are missing at random and the parameter of the missing data process is 'distinct' from θ.*" (Rubin, 1976)
  - may be computationally demanding.

Statistical analysis

- At the first-stage we have measured the covariate $x(i)$ and the disease outcome $y(i)$ for the whole cohort $i \in C = \{1, 2, \ldots, N\}$.
- At the second-stage, the genetic covariate of interest $g(i)$ is measured for a subset of the cohort but is missing for the most of the cohort.
- The model parameters can be estimated by directly maximizing the likelihood

$$L(\psi, \theta) \propto \prod_{j=1}^{n} p_{\boldsymbol{\theta}}(g(j)) p_{\boldsymbol{\psi}}(x(j) \mid g(j)) p_{\boldsymbol{\theta}}(t(j), \delta(j) \mid g(j), x(j))$$
$$\prod_{j=n+1}^{N} \sum_{g} p_{\boldsymbol{\theta}}(G(j) = g) p_{\boldsymbol{\psi}}(x(j) \mid g) p_{\boldsymbol{\theta}}(t(j), \delta(j) \mid g, x(j)),$$

  where $Y = (t, \delta)$ and G is observed for individuals $j = 1, \ldots, n$ and not observed for individuals $j = n + 1, \ldots, N$. The possible dependence between g and x need to be modeled.

Statistical analysis in R

- Maximum likelihood analysis is a general approach but requires flexible tools $\Rightarrow$ use R.
- The missing genetic variable is discrete $\Rightarrow$ integration reduces to summation.
- The likelihood function can be written in closed form and maximized using the R function `optim`.
- Variances are estimated from Hessian returned by `optim`.

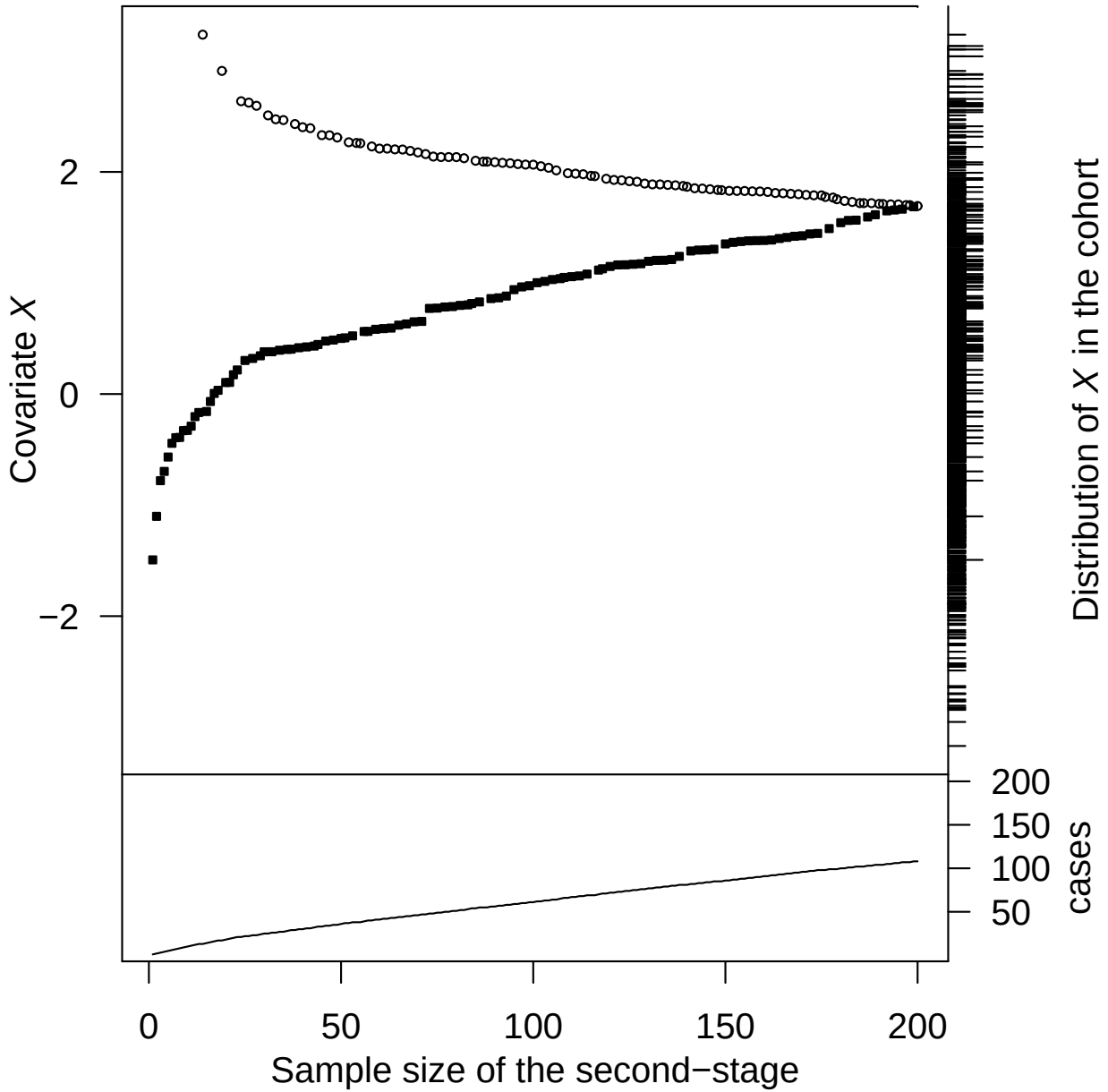
D-optimal design

- D-optimal design maximizes the determinant of Fisher information matrix.
- We used observed information matrix and derived the D-criterion to be maximized under logistic regression and proportional hazards models.
- Equations are given in Karvanen, J., Kulathinal S., Gasbarra D., 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.
- D-optimal designs are found by heuristic search. Greedy method works well: the individuals are selected sequentially one by one so that the D-criterion D_n for n individuals is maximized on the condition that $n-1$ individuals have been already selected.

Simulation example: Rare disease

- Follow-up data for 2000 individuals are generated.
- The event times of a rare disease follow the Weibull regression model where the covariates are a normally distributed phenotype x (regression coefficient $a = 1$) and a genetic indicator variable g (regression coefficient $b = 0.5$, allele frequency $\pi = 0.4$).
- Phenotype x is generated from the distribution $N(\mu + \gamma g, \sigma^2)$, where $\mu = 0$, $\sigma^2 = 1$ and $\gamma = 0.3$.
- Simple random sampling (SRS), case-cohort design (CC), extreme selection and D-optimal design are compared when logistic regression model or proportional hazards model are fitted to the data.
- Selection under D-optimal design is illustrated graphically.

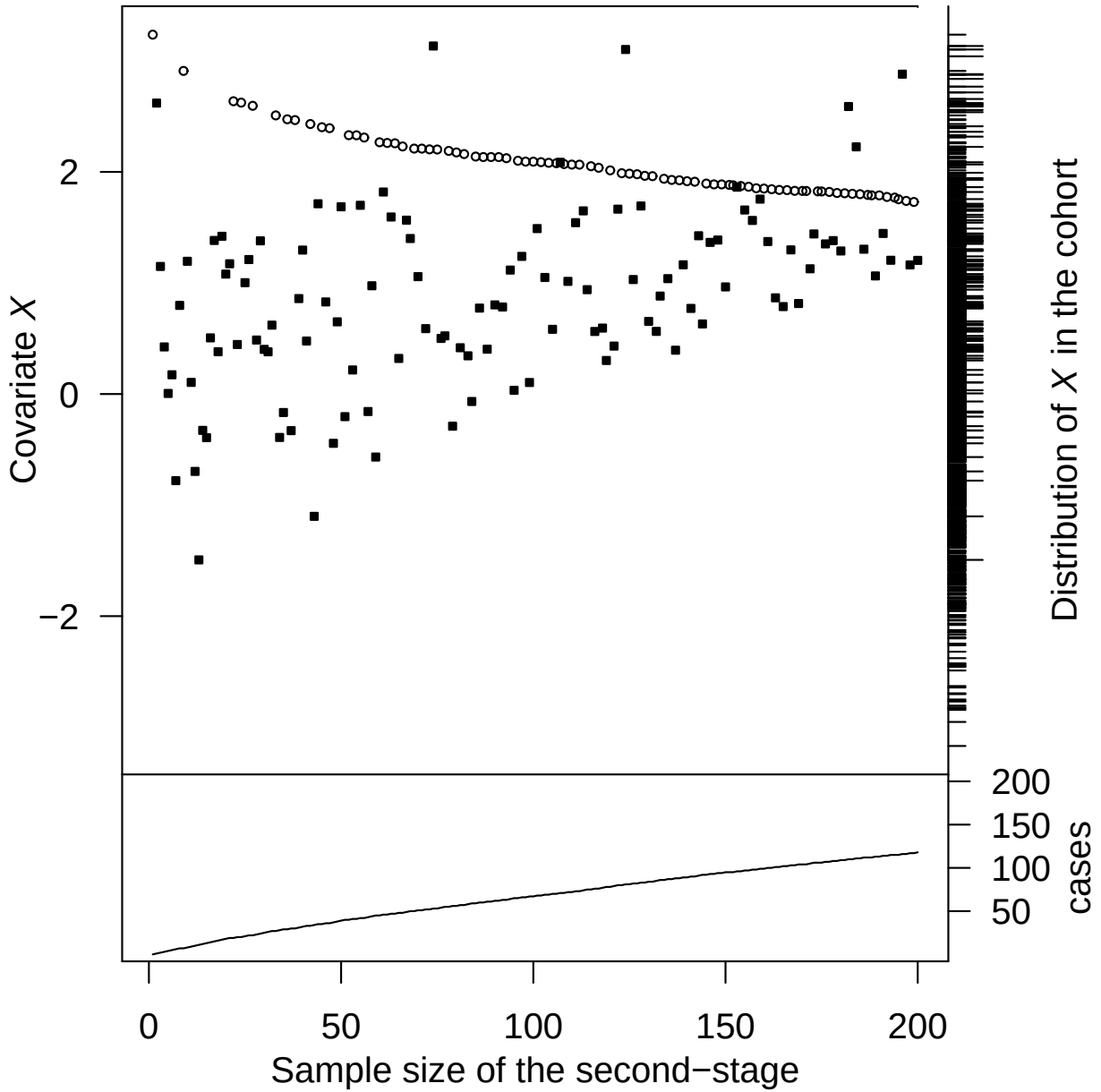
Rare disease: logistic model



Rare disease: logistic model

parameter	design	n=100		n=200		n=500	
		estim.	SE	estim.	SE	estim.	SE
a	SRS	1.08	0.11	1.08	0.11	1.06	0.11
	CC	1.08	0.11	1.07	0.10	1.07	0.10
	extreme	1.08	0.11	1.07	0.10	1.07	0.10
	D-optimal	1.08	0.10	1.07	0.10	1.07	0.10
b	SRS	0.63	0.85	0.65	0.61	0.64	0.38
	CC	0.52	0.40	0.54	0.29	0.59	0.22
	extreme	0.58	0.37	0.60	0.27	0.59	0.20
	D-optimal	0.57	0.30	0.57	0.24	0.56	0.20
c	SRS	-3.54	0.46	-3.53	0.35	-3.53	0.24
	CC	-3.46	0.23	-3.48	0.19	-3.49	0.17
	extreme	-3.52	0.23	-3.52	0.19	-3.49	0.17
	D-optimal	-3.49	0.20	-3.50	0.17	-3.49	0.17
$\pi = 0.4$	SRS	0.40	0.041	0.40	0.032	0.40	0.021
	CC	0.40	0.049	0.40	0.039	0.40	0.023
	extreme	0.40	0.046	0.40	0.037	0.40	0.023
	D-optimal	0.40	0.046	0.40	0.035	0.40	0.021

Rare disease: proportional hazards model



Rare disease: proportional hazards model

parameter	design	n=100		n=200		n=500	
		estim.	SE	estim.	SE	estim.	SE
$a = 1$	SRS	0.99	0.095	0.99	0.093	0.98	0.087
	CC	1.00	0.088	0.99	0.082	0.99	0.077
	extreme	0.99	0.087	0.98	0.082	0.99	0.076
	D-optimal	1.00	0.085	0.99	0.080	0.99	0.076
$b = 0.5$	SRS	0.50	0.24	0.52	0.23	0.53	0.20
	CC	0.50	0.22	0.54	0.19	0.52	0.16
	extreme	0.50	0.21	0.52	0.18	0.52	0.15
	D-optimal	0.51	0.20	0.50	0.17	0.52	0.15
$\pi = 0.4$	SRS	0.40	0.041	0.40	0.032	0.40	0.021
	CC	0.40	0.045	0.40	0.037	0.40	0.023
	extreme	0.41	0.045	0.40	0.035	0.40	0.022
	D-optimal	0.41	0.045	0.40	0.035	0.40	0.021

Conclusions and remarks

- On the basis of the simulation results, extreme selection may be recommended as a practical study design.
  - does not require initial estimates
  - easy to implement
  - gives relatively good results compared to D-optimal design
  - probably possible improve the results of extreme selection further by specifying the ratio of cases and non-cases according to some suitable criterion
- D-optimality and other criteria based on Fisher information provide the theoretical background for efficient study design and serve as benchmarks for the ad-hoc designs.

Conclusions and remarks

- One should be aware that if the data are analyzed using the full likelihood, also extreme selection may be sensitive to wrong distributional assumptions. This was seen in another simulation example where the covariate x was generated from a non-normal distribution but modeled by normal distribution and as result, especially the estimates of the genotype effect were clearly biased. Fortunately, the empirical distribution of x is observed and we have a possibility to check our distributional assumptions.

Conclusions and remarks

- D-optimal design and extreme selection may be applied also in situations where the number of genetic or non-genetic covariates is greater than one.
- For a vector $\mathbf{X}$ of non-genetic covariates we may consider the linear combination $z = \mathbf{ax}$, where $\mathbf{a}$ is a vector of initial parameter estimates, and proceed as in the case of a single non-genetic covariate.
- When there are several genetic covariates of interest, extreme selection can be applied without modifications and for D-optimal design we may compute the optimal design for a typical genetic covariate or alternatively define the selected subset as a union of the optimal designs computed separately for each genetic covariate.

References

[1] Elfving, G., 1952. Optimum allocation in linear regression theory. The Annals of Mathematical Statistics 23 (2), 255–262.

[2] Karvanen, J., Kulathinal S., Gasbarra D., 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.

[3] Rubin, D. B., 1976. Inference and missing data. Biometrika 63 (3), 581–592.