

Genetic analysis of age-at-onset traits based on case-control family data

Benjamin H Yip¹, Tron Anders Moger², Yudi Pawitan¹

¹Department of Medical Epidemiology and Biostatistics,
Karolinska Institutet, 17177 Stockholm, Sweden

²Institute of Basic Medical Sciences, Department of Biostatistics,
University of Oslo, 0317 Oslo, Norway

Contact: yudi.pawitan@ki.se

August 20, 2009

YP – ISCB August 2009

Background and Motivation

- Recent findings: familial concordance in cancer survival (Lindström L et al, Lancet Oncol. 2007 Nov;8(11):961-2).
- How much is due to genetic (G) or environmental (E) effect?
- Separation of G and E requires random-effect modelling for family survival data (at least nuclear family)
- Standard approach based on frailty models (e.g. Hougaard P, 2000) mostly for exchangeable structure (one frailty term), **very** complicated for nuclear family structure (4 frailty terms just for trios).

Illustration: Swedish melanoma data

- melanoma: expect both genetic (e.g., skin type or propensity to sun-burn) and environmental (e.g., sun exposure) factors to be important
- melanoma is rare cancer (4% of all cancers), need large population-based cohort
- Linkage of the Multi-Generation Register (MGR), the Swedish Cancer Register, the Death Register and the Migration Register
- Follow-up period: 1961-2001
- 125,739 families with 2 oldest children: almost 500K individuals

- The distribution of the number of melanomas within the families:

Number of melanomas	0	1	2	3	> 3
Number of families	94,297	31,075	360	7	0

- Full data too large for direct analyses. Measurement of covariates too costly.
- Sub-sampling to get family-based case-control data:
 - case families: at least 2 affected members
 - control families: other families

MAFT model: mixed accelerated failure time model

- MAFT specifies a direct relationship between survival time and covariates including fixed and random effects.
- T_{ij} be the survival time for the j th member of the i th family, for $i = 1, \dots, N$ and $j = 1, \dots, n_i$.
- left truncation time L_{ij} and the right censoring time F_{ij} , independent of T_{ij} .
- Let $y_{ij} \equiv \min(\log T_{ij}, \log F_{ij})$, and $\delta_{ij} \equiv I(\log T_{ij} \leq \log F_{ij})$ the event indicator. We observe y_{ij} only if $y_{ij} \geq \log L_{ij}$, i.e., the person is not truncated.

- The MAFT model

$$\begin{aligned}\log T_{ij} &= \mu_{ij} + e_{ij}, \\ \mu_{ij} &= x_{ij}^T \beta + g_{ij} + a_{ij} + c_{ij},\end{aligned}\tag{1}$$

where $x_{ij} = (x_{ij1}, \dots, x_{ijp})$ is a vector of known fixed covariates, β a p-vector of regression parameters, $g_{ij} \sim N(0, \sigma_g^2)$, $a_{ij} \sim N(0, \sigma_a^2)$, $c_{ij} \sim N(0, \sigma_c^2)$ and $e_{ij} \sim N(0, \sigma_e^2)$.

- Let $b_{ij} = g_{ij} + a_{ij} + c_{ij}$. Assume:

$$\begin{aligned}\text{var}(b_{ij}) &= \sigma_g^2 + \sigma_a^2 + \sigma_c^2 + \sigma_e^2 \\ \text{cov}(b_{\text{siblings}}) &= \frac{1}{2}\sigma_g^2 + \sigma_c^2\end{aligned}$$

$$\begin{aligned}\text{cov}(b_{parent}, b_{child}) &= \frac{1}{2}\sigma_g^2 \\ \text{cov}(b_{spouses}) &= \sigma_a^2,\end{aligned}$$

- The dependencies between effects can also be summarized as

$$D_i = \sigma_g^2 R_g + \sigma_a^2 R_a + \sigma_c^2 R_c,$$

where the relationship matrices R_g , R_a and R_c are implied by the covariances between members in one family given above.

H-likelihood procedure

- Given $(y_{ij}, \delta_{ij}, x_{ij})$ s, and putting all the random effects into an array v , the h-likelihood is

$$h \equiv h(\beta, \theta, v) = \sum_{ij} \ell_{1ij} + \sum_i \ell_{2i}, \quad (2)$$

where

$$\ell_{1ij} = -\frac{1}{2}\delta_{ij} \{ \log(2\pi\sigma_e^2) + (m_{ij})^2 \} + (1-\delta_{ij}) \log \{ 1 - \Phi(m_{ij}) \} - \log \{ 1 - \Phi(m_{ij}^*) \},$$

where $m_{ij} \equiv (y_{ij} - \mu_{ij})/\sigma_e$, $m_{ij}^* \equiv (\log L_{ij} - \mu_{ij})/\sigma_e$, and $\Phi(\cdot)$ is the standard normal distribution function.

- For the second term, use the matrix form:

$$\mu_i \equiv X_i\beta + Z_iv_i,$$

then

$$\ell_{2i} = -\frac{1}{2}\{\log |2\pi D_i| + v_i^T D_i^{-1} v_i\}.$$

Computation

- The h-likelihood estimation procedure is summarized in the following steps:
 - A.** Given θ , estimate (β, v) by maximizing h in (2), where $v \equiv (v_1, \dots, v_N)$. This is given explicitly below.
 - B.** Given the maximizer $(\hat{\beta}, \hat{v})$ from step 1, estimate θ by maximizing the adjusted-profile likelihood

$$p_v(h) = h(\hat{\beta}, \theta, \hat{v}) - \frac{1}{2} \log |I(\hat{v}) / (2\pi)|, \quad (3)$$

where

$$I(\hat{v}) = \sigma_e^{-2} Z_i^T \text{diag}(w_{ij}) Z_i + D_i^{-1},$$

using

$$\begin{aligned}w_{ij} &\equiv \delta_{ij} + (1 - \delta_{ij})\psi(m_{ij}) - \psi(m_{ij}^*) \\ \psi(x) &\equiv h(x)\{h(x) - x\} \\ h(x) &\equiv \phi(x)/\{1 - \Phi(x)\}.\end{aligned}$$

Ascertainment and pseudo-h-likelihood

- Let $S = \{1, 2, \dots, N\}$ be the set of all families, divide S into disjoint subsets: $S = S_1 \cup \dots \cup S_K$. We then sample the families from each set S_k with probability p_k .
- define the pseudo-h-likelihood

$$h_p = \sum_{k=1}^K \frac{1}{p_k} \sum_{i \in A_k} \left\{ \sum_j \ell_{1ij} + \sum_i \ell_{2i} \right\}, \quad (4)$$

with the same ℓ_{1ij} and ℓ_{2i} as before. The previous algorithm only needs a small modification by applying the sampling probability weights.

Simulation study: nuclear families

- we simulate nuclear family data by taking some features of the real melanoma data.
- Total 5000 families. Use the model to generate true survival times T_{ij} , using gender as a fixed covariate.
- truncation rate is around 20% and the censoring rate around 90%
- All results are based on 100 replications.
- Case-control data:

- CC1: $n(\text{control families}) = n(\text{case families})$
- CC2: $n(\text{control families}) = 2n(\text{case families})$.

		β_0	β_1	σ_g^2	σ_a^2	σ_c^2	σ_e^2
True		4.3	0.1	0.1	0.04	0.02	0.06
Full data	Mean	4.243	0.096	0.089	0.037	0.017	0.046
	SD	0.027	0.014	0.003	0.004	0.003	0.005
	AveSE	0.010	0.015	0.003	0.003	0.004	0.002
CC1	Mean	4.221	0.095	0.086	0.035	0.018	0.042
	SD	0.033	0.050	0.006	0.003	0.006	0.004
	AveSE	0.119	0.041	0.006	0.003	0.006	0.004
CC2	Mean	4.230	0.094	0.088	0.036	0.017	0.044
	SD	0.031	0.034	0.005	0.004	0.004	0.005
	AveSE	0.118	0.031	0.005	0.003	0.005	0.003

Table 1: *Summary of the family survival data simulation.*

- with the full data, the h-likelihood procedure gives accurate estimates of the parameters of the MAFT model,
- the variance component parameters can be well estimated using the pseudo-h-likelihood procedure applied to the case-control data, but potentially with some loss of precision.
- For variance component estimation, this loss can be reduced by increasing the number of control families.

Application to Melanoma data

- There is a total of 125,739 families,

Family size		2	2	3	3	4
Number of children		1	2	1	2	2
Number of affected	0	250	1	24,027	399	69,620
	1	57	2	7,981	92	22,943
	2	2	0	64	1	293
	3	0	0	2	0	5

Table 2: *Distribution of families according family size, number of children and the number of melanoma cases within the family.*

- 367 families with at least 2 affected members are defined as case families.

- obtain two case-control datasets, by sampling 2 and 4 control families for every case family, called CC2 and CC4 datasets respectively.

Variable	Full data	Full data	CC2	CC4
<i>Fixed-regression parameter</i>				
Constant	5.183 (0.006)	5.052 (0.003)	4.850 (0.018)	5.043
Gender	-0.040 (0.005)	-0.036 (0.004)	-0.033 (0.021)	-0.016
Generation	0.041 (0.006)	-0.090 (0.004)	-0.175 (0.022)	-0.078
<i>Variance components</i>				
σ_g^2		0.089 (0.001)	0.097 (0.001)	0.088
σ_a^2		0.051 (0.001)	0.046 (0.001)	0.051
σ_c^2		0.027 (0.001)	0.003 (0.001)	0.028
σ_e^2	0.603 (0.007)	0.235 (0.001)	0.165 (0.001)	0.234

Table 3: *Summary of the melanoma data analysis.*

- The results from the ascertained data, particularly CC4, are close to those from the full data.
- The standard errors (SEs) of the regression estimates from the ascertained data are larger compared to full-data SEs
- SEs of the variance component estimates are comparable. This is consistent with simulation results.

Conclusions

- Mixed accelerated failure time (MAFT) model is useful for family survival data of complex structure
- Case-control modification makes the method applicable for large datasets.