Computational Statistics and Data Analysis 93 (2016) 324–335

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

The liability threshold model for censored twin data

Klaus K. Holst [a,*], Thomas H. Scheike [a], Jacob B. Hjelmborg [b]

[a] Department of Biostatistics, University of Copenhagen, Denmark

[b] Department of Epidemiology, Biostatistics and Biodemography, University of Southern Denmark, Denmark

ARTICLE INFO

Article history:
Received 30 April 2014
Received in revised form 18 January 2015
Accepted 22 January 2015
Available online 31 January 2015

Keywords:
Liability-threshold
Random effects
Probit model
Cumulative incidence
Right censoring
Competing risks
Polygenic model
Twins
Heritability

ABSTRACT

Family studies provide an important tool for understanding etiology of diseases, with the key aim of discovering evidence of family aggregation and to determine if such aggregation can be attributed to genetic components. Heritability and concordance estimates are routinely calculated in twin studies of diseases, as a way of quantifying such genetic contribution. The endpoint in these studies are typically defined as occurrence of a disease versus death without the disease. However, a large fraction of the subjects may still be alive at the time of follow-up without having experienced the disease thus still being at risk. Ignoring this right-censoring can lead to severely biased estimates. The classical liability threshold model can be extended with inverse probability of censoring weighting of complete observations. This leads to a flexible way of modelling twin concordance and obtaining consistent estimates of heritability. The method is demonstrated in simulations and applied to data from the population based Danish twin cohort to describe the dependence in prostate cancer occurrence in twins.

1. Introduction

Family studies provide an important tool for understanding etiology of diseases, with the key aim of discovering evidence of family aggregation and to determine if such aggregation can be attributed to genetic components. Heritability and concordance estimates are routinely calculated in twin studies of diseases, as a way of quantifying such genetic contribution. As a key paper for studying heritability of cancer, Lichtenstein et al. (2000) reported heritability estimates for prostate cancer of 0.42 (95% confidence limits 0.29–0.50) and casewise concordance of 0.21 in monozygotic (MZ) twins and 0.06 in dizygotic (DZ) twins based on combined cohorts of 44,788 twin pairs from the Nordic twin registries. This suggests a considerable genetic contribution to the development of prostate cancer. A polygenic liability threshold model, i.e., a Probit variance component model, was used to quantify the heritability on the liability scale from the classification of subjects as cancer cases or non-cancer cases (died without cancer). However, a large fraction of the twin-pairs were still alive at the end of follow-up but treated as non-cancer case. This corresponds to treating this part of the population as immune to cancer, suggesting that the estimates of the targeted population parameters in this study could be severely biased. The censoring mechanism has largely been ignored in the epidemiological literature of family studies, which unfortunately makes reported estimates of both heritability, and other population parameters of interest such as concordance probabilities, very difficult to interpret.

The key to solving this problem is to consider the event times in the analysis. Standard techniques for correlated survival data are not appropriate here, due to the competing risk of death. Dependence on the hazard scale while taking possible dependence between causes into account has been considered by Ripatti et al. (2003) and Gorfine and Hsu (2011). Scheike

* Corresponding author. Tel.: +45 35327901.
E-mail address: k.k.holst@biostat.ku.dk (K.K. Holst).

http://dx.doi.org/10.1016/j.csda.2015.01.014
0167-9473/

et al. (2014a) considered dependence on the probability scale via random effects models and Scheike et al. (2014b) examined non-parametric estimates of the concordance function, i.e., the probability of both twins experiencing cancer before a given time point. These methods yield constructive ways of analysing twin data of disease status, however, care in correctly specifying the dependence structure over time via the random effects structure has to be taken. Furthermore, none of the approaches provide heritability estimates that are comparable with the classical definition of heritability on the liability scale given by Falconer (1967). In the following we will define a simple estimator which gives consistent concordance estimates and estimates of heritability on the liability scale under independent right-censoring.

The paper is structured as follows. In Section 2 we review basic concepts in quantitative genetics and define heritability with the aim of estimating the degree of association due to genes and environmental factors through random effects modelling. In particular, we note that dependence on the probability scale is something quite different from dependence on the normal scale. We introduce the competing risks framework and present the inverse probability of censoring weighted estimating equations in Section 3. The method is demonstrated in simulations in Section 4. A worked example based on the Danish twin registry is presented in Section 5 followed by a general discussion.

2. Polygenic models

The basic idea of family-studies of a quantitative trait is to exploit that stronger phenotypic resemblance will be seen between closely related family members when the trait is genetically determined. In particular, for twin studies we may exploit that monozygotic (MZ) twins in principle are genetic copies whereas dizygotic (DZ) twins genetically on average resembles ordinary siblings. This allows us under appropriate genetic assumptions to decompose the trait into genetic and environmental components, $Y = Y_{\text{gene}} + Y_{\text{envir}}$, which may be modelled using random effects. Assuming independence between genetic and environmental effects the *broad-sense heritability* may then be quantified as the fraction of the total variance due to genetic factors.

The theoretic foundation in modern quantitative genetics was laid out in the pioneering work of Fisher (1918) who formally described the above genetic decomposition in terms of additive and dominant genetic effects. Familial resemblance may be defined from the *kinship-coefficient* Φ_{jk} which is the probability that two randomly selected alleles from the same locus of relatives k and j are *identical by descent*, i.e., the alleles are physical copies of the same gene carried by a common ancestor. Under assumptions of random mating (no inbreeding), linkage equilibrium, no gene–environment interaction and epistasis, and parents do not transmit their environmental effects to their children, this leads to a covariance between the observed phenotypes Y_k and Y_j for the relatives given by

$$\mathbb{C}\text{ov}(Y_k, Y_j) = 2\Phi_{kj}\sigma_A^2 + \Delta_{7kj}\sigma_D^2 + \sigma_C^2,$$

where the identity coefficient Δ_{7kj} describes the probability that at a given loci both alleles for the two relatives are identical by descent (Lange, 2002). The variance components σ_A^2 describe the additive genetic effects, σ_D^2 the dominant genetic effects and σ_C^2 describes variance of shared environmental effects for the two relatives.

This can be captured in a random effects model where the polygenic phenotype Y_{ij} may be modelled as

$$Y_{ij} = \beta^T X_{ij} + \eta_{ij}^A + \eta_i^C + \eta_{ij}^D + \varepsilon_{ij}, \tag{1}$$

for family $i = 1, \ldots, n$ and family member $j = 1, \ldots, K$ with covariates X_{ij}. Here we assume that there is the same shared environmental effect for all family members. All the random effects are assumed to be independent and normally distributed which in general may be reasonable for polygenic traits (Lange, 1997):

$$(\eta_{ij}^A, \eta_i^C, \eta_{ij}^D, \varepsilon_{ij})^T \sim \mathcal{N}\left(0, \text{diag}(\sigma_A^2, \sigma_C^2, \sigma_D^2, \sigma_E^2)\right).$$

The residual terms ε_{ij} are assumed to be i.i.d. normal and the variance component σ_E^2 may be interpreted as the variance of the unique environmental effects. The (broad-sense) heritability may then be defined as

$$H^2 = \frac{\sigma_A^2 + \sigma_D^2}{\sigma_A^2 + \sigma_C^2 + \sigma_D^2 + \sigma_E^2}.$$

For MZ twins we have $\Phi_{kj}^{\text{MZ}} = \frac{1}{2}$ and $\Delta_{7kj}^{\text{MZ}} = 1$ and for DZ twins $\Phi_{kj}^{\text{DZ}} = \Delta_{7kj}^{\text{DZ}} = \frac{1}{4}$, hence

$$\mathbb{C}\text{ov}(Y_{i1}^{\text{MZ}}, Y_{i2}^{\text{MZ}}) = \begin{pmatrix} \sigma_A^2 + \sigma_C^2 + \sigma_D^2 + \sigma_E^2 & \sigma_A^2 + \sigma_C^2 + \sigma_D^2 \\ \sigma_A^2 + \sigma_C^2 + \sigma_D^2 & \sigma_A^2 + \sigma_C^2 + \sigma_D^2 + \sigma_E^2 \end{pmatrix},$$

$$\mathbb{C}\text{ov}(Y_{i1}^{\text{DZ}}, Y_{i2}^{\text{DZ}}) = \begin{pmatrix} \sigma_A^2 + \sigma_C^2 + \sigma_D^2 + \sigma_E^2 & \frac{1}{2}\sigma_A^2 + \sigma_C^2 + \frac{1}{4}\sigma_D^2 \\ \frac{1}{2}\sigma_A^2 + \sigma_C^2 + \frac{1}{4}\sigma_D^2 & \sigma_A^2 + \sigma_C^2 + \sigma_D^2 + \sigma_E^2 \end{pmatrix}.$$

Note that one consequence of the model is that MZ and DZ twins follow the same marginal distribution. Unfortunately, the classic twin design does not allow identification of all variance components. Further inclusion of other family members or twin-adoptives can remedy this problem, but may further complicate assumptions regarding shared/non-shared environmental effects across different family members. The pragmatic solution is typically to report results from the most biologically relevant model, i.e., for certain traits the shared environmental effect may be known to be negligible, or to choose