Cohort sampling in epidemiological studies

Ørnulf Borgan
Department of Mathematics, University of Oslo
P.O.Box 1053 Blindern, N-0316 Oslo, Norway
borgan@math.uio.no

1. Introduction

Cox regression is much used in epidemiology to assess the influence of exposure variables and other covariates on mortality or morbidity. Estimation in Cox's model requires ascertainment of covariate values for all individuals in a cohort even when only a small fraction of these get diseased or die ("fail"). This may be very costly, or even logistically impossible. Cohort sampling techniques, where covariate information is collected for all failing individuals ("cases"), but only for a sample of the non-failing ones ("controls"), offer useful alternatives. Two common cohort sampling designs are nested case-control and case-cohort sampling. The purpose of this paper is to review, discuss and compare these two sampling designs with emphasis on recent developments and challenges for further research. We restrict our attention to situations which involve only a single disease.

2. Model and inference for cohort data

We first review Cox regression for cohort data. Consider a cohort of n individuals, and let $\lambda_i(t)$ be the hazard for the ith individual with covariates $\mathbf{x}_i(t) = (x_{i1}(t), \ldots, x_{ip}(t))'$. Here t may be age, time since employment, or some other time-scale relevant to the problem at hand. The hazard is related to the covariates through the relation

$$\lambda_i(t) = \lambda_0(t) \exp\{\boldsymbol{\beta}' \mathbf{x}_i(t)\} \tag{1}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is a vector of parameters, and $\lambda_0(t)$ is a non-negative baseline hazard.

The individuals may be followed over different periods of time, i.e., our observations may be left-truncated and/or right censored. We denote by $t_1 < t_2 < \cdots < t_d$ the times when failures are observed, and, assuming no tied failures, denote by i_j the individual who fails at t_j. The risk set $\mathcal{R}(t_j)$ consists of all individuals who are under observation at t_j, including the case i_j. Then $\boldsymbol{\beta}$ in (1) is estimated by maximizing Cox's partial likelihood. This is the special case of

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{d} \frac{\exp\{\boldsymbol{\beta}' \mathbf{x}_{i_j}(t_j)\} w_{i_j}(t_j)}{\sum_{k \in \mathcal{S}(t_j)} \exp\{\boldsymbol{\beta}' \mathbf{x}_k(t_j)\} w_k(t_j)} \tag{2}$$

where $\mathcal{S}(t_j) = \mathcal{R}(t_j)$ and $w_k(t_j) = 1$. Since in this case, (2) is a partial likelihood, the usual large sample likelihood methods apply.

3. Sampling designs

In their original forms, the nested case-control and case-cohort designs use simple random sampling without replacement. Recently modifications have been developed which use stratified random sampling. Such stratified designs may be advantageous when a surrogate measure of exposure is available for everyone and can be used to classify the individuals into strata.

For the simple *nested case-control design* (Thomas, 1977), one selects, for each t_j, a random sample of $\widetilde{m}-1$ controls from the $n(t_j)-1$ non-failing individuals in $\mathcal{R}(t_j)$. The sampled risk set $\widetilde{\mathcal{R}}(t_j)$ consists of the case i_j and these $\widetilde{m}-1$ controls. Covariate values are ascertained for the cases and their sampled controls, but are not needed for the other individuals in the cohort. A stratified version of this design, called counter-matching, is as follows (Langholz and Borgan, 1995). For each t_j one selects at random $\widetilde{m}_s$ controls from the $n_s(t_j)$ individuals at risk in stratum s, except for the case's stratum where only $\widetilde{m}_s - 1$ controls are sampled. The case i_j is included in the sampled risk set, so $\widetilde{\mathcal{R}}(t_j)$ contains a total of $\widetilde{m}_s$ from each stratum.

For the simple *case-cohort design* (Prentice, 1986), one selects at random a subcohort $\mathcal{C}$ of size m from the full cohort. Covariate values are ascertained for all individuals in $\mathcal{C}$ as well as for cases occurring outside the subcohort. When the individuals in the cohort may be classified into a number of distinct strata, with n_s individuals in stratum s, one may alternatively select a random sample of m_s individuals to $\mathcal{C}$ from each stratum s (Borgan *et al.*, 2000).

Control sampling in a nested case-control study are from those at risk at the cases' failure times, while in a case-cohort study the subcohort is selected at the outset of the study without consideration of at risk status. This difference in the way sampling is performed, creates two limitations for nested case-control studies which are avoided for case-cohort studies (Barlow *et al.*, 1999): (i) ascertainment of covariate values for the controls has to wait until failures occur, and (ii) choice of time-scale for the analysis has to be decided before the controls are selected. The relevance of these limitations depends on the situation at hand, but they are most likely to be of importance for prospective studies like disease prevention trials.

Important questions for the stratified designs are (i) how one should classify the individuals into strata, and (ii) how many individuals one should choose from each stratum. Answers to (ii) can be based on large sample variances; see Borgan *et al.* (2000) for a discussion related to case-cohort studies. To answer question (i), one has, e.g., to determine the number of strata to be used and how to categorize a continuous surrogate measure used for creating the strata.

4. Inference and computing

Both for nested case-control and case-cohort studies, estimation of $\boldsymbol{\beta}$ is based on (2) with particular choices of the sets $\mathcal{S}(t_j)$ and the weights $w_k(t_j)$. For the *nested case-control designs* one has $\mathcal{S}(t_j) = \widetilde{\mathcal{R}}(t_j)$, while $w_k(t_j) = 1$ for simple random sampling and $w_k(t_j) = n_s(t_j)/\widetilde{m}_s$, when k belongs to stratum s, for stratified random sampling. In any case, (2) is a partial likelihood, and it may be shown, using counting processes and martingales, that the usual large sample likelihood methods apply (Borgan *et al.*, 1995). For computing one may use standard software treating the label j of the sampled risk sets as a stratification variable in the Cox regression and including $\log w_k(t_j)$ as an "offset" in the model for the counter-matched design.

For the *case-cohort design* with simple random sampling, Prentice (1986) proposed to estimate $\boldsymbol{\beta}$ from (2) with $\mathcal{S}(t_j) = (\mathcal{C} \cap \mathcal{R}(t_j)) \cup \{i_j\}$ and $w_k(t_j) = 1$. For stratified sampling, the weights are $w_k(t_j) = n_s/m_s$, when k belongs to stratum s (Borgan *et al.*, 2000; estimators I and III). In these cases, (2) is not a partial likelihood, but rather a pseudo likelihood. Therefore martingale results cannot be used, and for large sample studies one has to combine convergence results for cohort data with finite population convergence results (Self and Prentice, 1988).

For many years, the analysis of case-cohort studies has been hampered by lack of suitable software and the fact that Prentice's (1986) original variance estimator is computationally very demanding. It is now realized that standard software can be accommodated to estimate the regression parameters and their (large sample) standard errors; Barlow *et al.* (1999) and Therneau and Li (1999) give details for the simple case-cohort design. Nevertheless, since usual large sample likelihood methods do *not* apply, the analysis of data from case-cohort studies is more cumbersome than the analysis of nested-case control data.

5. Relative efficiencies and alternative analysis methods

The efficiency of a simple nested case-control study relative to a full cohort study is $(\widetilde{m} - 1)/\widetilde{m}$ when $\boldsymbol{\beta} = \mathbf{0}$, independent of censoring and covariate distributions (Goldstein and Langholz, 1992). For the simple case-cohort design, it does not seem possible to derive a similar general and simple result (Self and Prentice, 1986). Although published results are somewhat conflicting (e.g., Langholz and Thomas, 1991; Barlow, *et al.*, 1999), the relative efficiencies of nested case-control and case-cohort studies seem to be about the same when they involve the same number of individuals for whom covariate values have to be ascertained. Stratified sampling may give an appreciable improvement in statistical efficiency both for nested case-control studies (Langholz and Borgan, 1995) and case-cohort studies (Borgan *et al.*, 2000), and the gain seems to be of comparable size for the two designs.

The analysis methods described in the previous section, include the cases in the denominator of (2) only at their failure times. Methods have been developed which use information from the cases whenever they are at risk; see Samuelsen (1997) for nested case-control studies and Kalbfleisch and Lawless (1988), Chen and Lo (1999), and Borgan *et al.* (2000; estimator II) for case-cohort studies. For these alternative methods, estimation of $\boldsymbol{\beta}$ is based on pseudo likelihoods of the form (2), where the cases and controls/subcohort members are weighted differently. When the disease under study is not too rare, these alternative methods may perform better than the ones presented above, in particular for case-cohort data. However, cohort sampling is mainly suited for rare diseases, and then it is not at all clear that the alternative methods are to be preferred, even for the analysis of case-cohort studies.

REFERENCES

[1] Barlow, W. E., Ichikawa, L., Rosner, D., and Izumi, S. (1999). Analysis of case-cohort designs. *J. Clin. Epidemiol.*, **52**, 1165-72.

[2] Borgan, Ø., Goldstein L., and Langholz, B. (1995). Methods for the analysis of sampled cohort data in the Cox proportional hazards model. *Ann. Stat.*, **23**, 1749-78.

[3] Borgan, Ø., Langholz, B., Samuelsen, S. O., Goldstein L., and Pogoda, J. (2000). Exposure stratified case-cohort designs. *Lifetime Data Anal.*, **6**, 39-58.

[4] Chen, K. and Lo, S.-H. (1999). Case-cohort and case-control analysis with Cox's model. *Biometrika*, **86**, 755-64.

[5] Goldstein L. and Langholz, B. (1992). Asymptotic theory for nested case-control sampling in the Cox regression model. *Ann. Stat.*, **20**, 1903-28.

[6] Kalbfleisch, J. D. and Lawless, J. F. (1988). Likelihood analysis of multi-state models for disease incidence and mortality. *Stat. Med.*, **7**, 149-60.

[7] Langholz, B. and Borgan, Ø. (1995). Counter-matching: A stratified nested case-control sampling method. *Biometrika*, **82**, 69-79.

[8] Langholz, B. and Thomas, D. C. (1991). Efficiency of cohort sampling designs: some surprising results. *Biometrics*, **47**, 1563-71.

[9] Prentice, R. L. (1986). A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika*, **73**, 1-11.

[10] Samuelsen, S. O. (1997). A pseudolikelihood approach to analysis of nested case-control studies. *Biometrika*, **84**, 379-94.

[11] Self, S. G. and Prentice, R. L. (1988). Asymptotic distribution theory and efficiency results for case-cohort studies. *Ann. Stat.*, **16**, 64-81.

[12] Therneau, T. M. and Li, H. (1999). Computing the Cox model for case-cohort designs. *Lifetime Data Anal.*, **5**, 99-112.

[13] Thomas, D. C. (1977). Addendum to: "Methods of cohort analysis: appraisal by application to asbestos mining," by F. D. K. Liddell, J. C. McDonald and D. C. Thomas. *J. Roy. Stat. Soc. A*, **140**, 469-91.

RÉSUMÉ

La régression de Cox est une méthode souvent utilisée en épidémiologie pour estimer l'influence de variables d'expositions et autres covariables sur la mortalité ou la morbidité. Elle nécessite la connaissance des valeurs des covariables pour tous les individus d'une cohorte, même si seulement un petit nombre d'entre eux décède ou devient malade. Ceci peut être extrêmement coûteux, voire logistiquement impossible. Des techniques d'échantillonnage au sein de cohorte, avec un recueil des valeurs des covariables pour l'ensemble des individus décédés ou devenus malades et sur un échantillon des individus non-atteints, constituent une alternative intéressante. Le présent article présente, discute et compare deux techniques de sélection de cohorte classiques, en insistant sur les dévelopements récents et les défis posés à la recherche.