Methods for the Analyses of Case-Cohort Studies

ALEXANDER VOLOVICS
Department of Methodology and Statistics
Maastricht University
The Netherlands

PIET A. VAN DEN BRANDT
Department of Epidemiology
Maastricht University
The Netherlands

Summary
Case-cohort and nested case-control sampling methods have recently been introduced as a means of reducing cost in large cohort studies. The asymptotic distribution theory results for relative rate estimation based on Cox type partial or pseudolikelihoods for case-cohort and nested case-control studies have been accounted for. However, many researchers use (stratified) frequency table methods for a first or primary summarization of the most important evidence on exposure-disease or dose-response relationships, i.e. the classical Mantel-Haenszel analyses, trend tests and tests for heterogeneity of relative rates. These can be followed by exponential failure time regression methods on grouped or individual data to model relationships between several factors and response. In this paper we present the adaptations needed to use these methods with case-cohort designs, illustrating their use with data from a recent case-cohort study on the relationship between diet, life-style and cancer. We assume a very general setup allowing piecewise constant failure rates, possible recurrent events per individual, independent censoring and left truncation.

Key words: Case-cohort; Piecewise constant failure rates; Relative rate regression; Right censoring; Left truncation.

1. Introduction

The germs of the ideas of nested case-control and case-cohort sampling from a cohort can be found scattered through the statistical and epidemiological literature of the 60's and 70's. In the failure time context Thomas, in LIDDELL, MCDONALD and THOMAS (1977) appears to have first comprehensively formulated the nested case-control approach while PRENTICE (1986a) proposed the case-cohort design as a more efficient solution, in some situations, to sampling cohort subjects than the nested case-control sampling design. The general approach to nested case-control sampling involves the selection of a random sample without replacement of sub-
jects at risk but without disease (control subjects) at each distinct failure time (every time a case is observed). Selected controls remain eligible for control selection at other failure times. The control selection procedures at distinct failure times are statistically independent. Relative rate estimation can be based on a Cox type partial likelihood approach. Curiously, this approach has only recently been theoretically justified by Goldstein and Langholz (1992). When studying a range of disease endpoints this design can pose logistical and administrative problems as a different random sample of controls has to be selected each time a specific endpoint is observed. Furthermore, collecting and processing of covariate information on controls can only start at the time of the first failure and all covariate information will only be available at the end of the study period.

The case-cohort design avoids these problems by selecting a subcohort randomly from the entire cohort which then provides a comparison group at each disease occurrence time, i.e., those subjects in the subcohort still at risk for the disease under study at a given failure time function as controls for the occurring failure whether that failure occurs 'inside' or 'outside' the subcohort. This design allows the comparison group to be selected in advance of cohort follow-up, a distinct advantage since the subcohort can then be used, for example, to monitor the achievement of intervention goals and the collection and processing of covariate information for all controls (the subcohort members) can be started immediately on inception. Also, in contrast to the control sample in the time-matched nested case-control design, the subcohort provides a natural comparison group for a range of disease endpoints. Prentice (1986b) provided heuristic justification for relative rate regression analysis based on a Cox type pseudo partial likelihood approach. Full theoretical justification for the proposed methods was presented in Self and Prentice (1988). Readers interested in further methodological details regarding the above and other similar designs can consult Lubin and Gail (1984), Prentice (1986a, b), Robins, Gail and Lubin (1986), Wacholder and Boitvin (1987), Robins, Prentice and Blevins (1989), Langholz and Thomas (1990), Wacholder, Gail, Pee and Brookmeyer (1991).

We will concentrate here on procedures for the analyses for case-cohort studies. The organization of this paper is as follows: §2 presents a heuristic introduction, §3 contains the adaptations for exponential failure time regression, §4 contains an illustration of the procedures in a case-cohort analysis of smoking and lung cancer and proofs are presented in an appendix. All formulas were programmed in GLIM-code; these macros are available upon request.

2. Heuristic Introduction to Case-cohort Analyses

The elementary analysis of cohort studies and Poisson regression for grouped data are comprehensively described in Breslow and Day (1987), chapters 2, 3, 4 and 5, to which we refer for standard formulas and practice. 'Poisson' regression for individual data is described in Aitkin et al. (1989), chapter 6.
The results of a cohort and case-cohort study with only two exposure groups can be summarized as in Table 1, a and b.

The notation $t_{di}$, $t_{ci}$, $t_{sc}$ in Table 1 indicates person years for failures, for censored individuals and for censored individuals in the subcohort only, respectively. Censoring is due for example to ending of follow-up, withdrawals or competing causes of failure. Assuming a constant failure rate during the study period the failure rate $\lambda_i$ for exposure group $i$ in the cohort study can be estimated as $\hat{\lambda}_i = d_i/t_i$, the number of observed cases divided by the total person years of exposure in group $i$, $i = 1, 2$. The relative rate $\psi$ of exposure 2 versus exposure 1 can be estimated as $\hat{\psi} = \hat{\lambda}_2/\hat{\lambda}_1 = d_2 t_1/d_1 t_2$. It will be intuitively clear that the analogous statistic $d_2 t_1/d_1 t_2$ in the case-cohort set-up should in general not be used to estimate the relative rate. The person years of the failures is the same in both studies but the person years of the censored individuals in the case-cohort study is a random sample from the censoring times in the cohort study. If the sampling fraction is $q = m/n$ with $m$ denoting subcohort and $n$ cohort size, we can expect that $t_{sc} \approx q t_{ci}$. Thus $d_2 t_1/d_1 t_2 \approx d_2 (t_{d1} + q t_{c1})/d_1 (t_{d2} + q t_{c2})$ and this will in general not be close to $d_2 t_1/d_1 t_2$.

Two solutions present themselves:

a. 'Blow-up' the person years of the censored individuals in the case-cohort study by a factor $q^{-1}$, that is to say use $t_{di} + q^{-1} t_{sc} \approx t_i$ instead of $t_i$. Using the 'blow-up' method we estimate $\psi$ by

$$\hat{\psi}^B = d_2 (t_{d1} + q^{-1} t_{c1})/d_1 (t_{d2} + q^{-1} t_{c2}) \approx \hat{\psi}.$$  

b. 'Shrink' the person years of the failures by a factor $q$, i.e. use $Q t_{di} + t_{sc}$ instead of $t_i$. The same effect could of course be accomplished by using only the total person years of subcohort members, namely $t_{si} = t_{sdi} + t_{sc}$ where $t_{sdi}$ is the observed person years for failures in the subcohort only: $t_i \approx Q t_i$. Using the 'shrink' method we estimate $\psi$ by $\hat{\psi}^S = d_2 t_{s1}/d_1 t_{s2} \approx d_2 t_1/d_1 t_2 = \hat{\psi}$.

It is important to note that the 'blow-up' method gives direct estimates of the exposure specific failure rates, namely $\hat{\lambda}_i^B = d_i/(t_{di} + q^{-1} t_{sc}) \approx \hat{\lambda}_i$. With the
'shrink' method we cannot use the analogous estimator for the exposure specific rates because $d_i/t_i \approx q^{-1}d_i/t_i \approx q^{-1}\hat{\lambda}_i$. The obvious corrected estimator for $\hat{\lambda}_i$ is $q\hat{d}_i/t_i$. We shall also say that the 'shrink' method estimates blown-up exposure specific rates $\mu_i = q^{-1}\hat{\lambda}_i$ with estimators $\hat{\mu}_i = d_i/t_i$; to estimate the $\hat{\lambda}_i$ we have to 'shrink' $\hat{\mu}_i$ by the sampling fraction: $\hat{\lambda}_i = q\hat{\mu}_i$.

Confidence intervals for $\bar{\psi}$ are usually calculated using estimated asymptotic variances of $\log \hat{\psi}$. Concentrating for the moment on the 'shrink' approach we have:

$$\log \hat{\psi} = \log \left\{ \frac{d_2t_2/d_1t_2}{d_2t_1/d_1t_2} \right\} = \log \left\{ \frac{d_2t_1/d_1t_2}{d_2t_1/d_1t_2} \right\} + \left\{ \log (t_2/n) - \log (t_2/m) \right\} - \left\{ \log (t_1/n) - \log (t_1/m) \right\} = \log \hat{\psi} + D^5, \quad \text{say}.$$ 

This equality illustrates 2 things.

Firstly, if a suitably normalized version of $D^5$ converges in distribution to a nondegenerate random variable, then as was to be expected, $\hat{\psi}$ will be an inefficient 
estimate of $\psi$ compared to $\bar{\psi}$, having a larger asymptotic variance (AV): namely $AV(\log \hat{\psi}) = AV(\log \bar{\psi}) + AV(D^5)$. (It can be shown that $\bar{\psi}$ and $D^5$ are asymptotically independent).

Secondly, given the $n$ individual failure times constituting the $t_i$, $i = 1, 2$, the variance of $D^5$ is due to random sampling of $m$ failure times from the finite population of cohort failure times and in estimating the variance of $\log \hat{\psi}$ this extra finite sampling variability will have to be accounted for. This means that 'naive' estimators for $AV(\log \hat{\psi})$ calculated analogously to the estimator for $AV(\log \bar{\psi})$ will underestimate the true AV and should not be used. Analogous reasoning can be used to show comparable results for $\hat{\psi}^B$.

The classical elementary methods all assume (at least implicitly) an underlying relative or multiplicative rate model: $\theta_k = \lambda_j \psi_k$ for $j = 1$ to $J$ strata and $k = 1$ to $K$ exposure classes. $\psi_1 = 1$, so that $\hat{\lambda}_j = \theta_k$, the rate for individuals in the baseline or nonexposed category ($k = 1$).

Assuming this relative rate model it is easily seen that the above mentioned (the $J = 1$, $K = 2$ case) 'blown-up' and 'shrink' estimators are maximum likelihood estimators for $\psi = \psi_2$ based on a 'blown-up' or 'shrink' pseudo partial likelihood

$$L^B(\lambda, \psi) = \lambda^{d_1} e^{-\lambda t_1} (\lambda \psi)^{d_2} e^{-\lambda \psi t_2},$$

$$L^S(\mu, \psi) = \mu^{d_1} e^{-\mu t_1} (\mu \psi)^{d_2} e^{-\mu \psi t_2}$$

where $t_k$ stands for $t_{dk} + q^{-1}t_{bk}$ in the $L^B$ case and for $t_{dk} + t_{bk}$ in the $L^S$ case, $k = 1, 2$, and we have written $\lambda(\mu)$ for $\lambda_1(\mu_1)$.

For the $J > 1$ strata, $K > 2$ exposure classes case the obvious generalization of such a pseudo partial likelihood is:

$$L(\theta_1, \ldots, \theta_J) = \prod_{j=1}^{J} \prod_{k=1}^{K} \theta_{jk}^{d_{jk}} e^{-\theta_{jk} t_{jk}}$$
where $\theta_{jk}$ is either $\lambda_{jk}$ ('blow-up') or $\mu_{jk}$ ('shrink'), $d_{jk}$ stands for the number of failures in stratum $j$ and exposure class $k$ and where $t_{jk}$ is either $t_{djk} + Q^{-1}t_{eijk}$ ('blow-up') or $t_{adjk} + t_{eijk}$ ('shrink').

Elementary, summary table based methods for case-cohort analyses can be derived from the above likelihood.

Of these methods based on summary tables, described in Breslow and Day (1987), the Mantel-Haenszel test, the test for trend and the test for heterogeneity of relative risk are score tests based on models like the likelihood presented above and as such are easily available from the results in § 3. The formulas for the Mantel-Haenszel estimator and the other chi-square tests mentioned in Breslow and Day (1987) remain the same but the formula for the variance of the Mantel-Haenszel estimator has to be adapted. Of course 'shrink' or 'blow-up' person years must always be substituted. See Volovics and Van den Brandt (1995) for details.

The classical setup assumes a Poisson distribution for the number of failures in each cell. As already indicated in Breslow and Day (1987) this is not necessary and we shall here assume a piecewise constant failure rate, possible recurrent events (multiple 'failures') per individual, independent censoring (Andersen, Borgan, Gill and Keiding, 1993) and left truncation. Left truncation, in the simple form of only including subjects conditionally on being alive (not having failed) at a certain given calendar time or age, is quite common to cohort studies.

3. Regression Methods for Grouped and Individual Case-cohort Data

Regression analyses of case-cohort data on the individual level offers a more versatile approach to modelling exposure-response relationships than the above described elementary techniques and at the same time enables one to compute these elementary, or practically equivalent methods, much more efficiently.

Grouped data analyses, such as the elementary methods described above, are easily obtained after defining for each individual in the study dummy variables or scores indicating whether this individual belongs to a certain stratum or belongs to a certain exposure group characterized by a mean or summary score for the exposure group. As with the elementary methods we shall concentrate on the multiplicative rate model as this seems to be the model that is by far the most frequently used.

The multiplicative rate model for regression assumes individual specific rate functions:

$$\lambda_i(u) = \exp \left( \sum_{a=1}^{A} \beta_0 a 1(\tau_a - 1 < u \leq \tau_a) + \beta_1 Z_i(u) \right)$$

where $\beta_1$ and $Z_i(u)$ are $p$-dimensional vectors. We have written the vector of co-variables as $Z_i(u) = (Z_{i1}(u), \ldots, Z_{ip}(u))'$ to indicate that one or more of the $p$
components \( Z_j(u), 1 \leq j \leq p \), could be time dependent. The notation \( 1(\tau_{a-1} < u \leq \tau_a) \) is used for the indicator function with value 1 if \( \tau_{a-1} < u \leq \tau_a \) and 0 for \( u \) elsewhere.

The \( \tau_a \) are a set of time points with \( \tau_0 < \tau_1 < \tau_2 \ldots < \tau_{A-1} < \tau_A = \tau \), \( \tau < \infty \) the maximal possible observation time for any individual under study. If all components of \( Z(u) \) are fixed, i.e. \( Z(u) \equiv Z_t \) on \((0, \tau] \), such a failure rate describes a piecewise exponential distribution for the failure times: on each interval \((\tau_{a-1}, \tau_a]\) we model an individual hazard function \( \lambda_i \) as a constant \( \lambda_{0a} \). Given a multiplicative rate model this means defining a piecewise constant baseline failure rate common to all subjects, i.e. \( \lambda_{0a} = \exp (\beta_{0a}' Z_i) \) on \( \tau_{a-1} < u \leq \tau_a, a = 1, 2, \ldots, A \). Then we have as individual failure rates \( \lambda_{ia} = \lambda_{0a} \exp (\beta_{0a}' Z_i) = \exp (\beta_{0a}' Z_i) \exp (\beta_{ia}' Z_i) \).

Note that the ‘shrink’ approach estimates ‘blown-up’ baseline failure rates on each interval \( \tau_{a-1} < u \leq \tau_a \), namely \( \mu_{0a} = \exp (\beta_{0a}' Z_i) = \exp (\beta_{0a} - \log \varphi) = \exp (\alpha_{0a}) \), say, and when modelling the ‘shrink’ approach we have to consider individual failure rates:

\[
\mu_i(u) = \exp \left( \sum_{a=1}^{A} \alpha_{0a} 1(\tau_{a-1} < u \leq \tau_a) + \beta_{ia}' Z_i(u) \right), \quad 0 < u \leq \tau.
\]

When convenient we shall write \( \beta \) for both \( (\beta_{0a}', \beta_{ia}') \) and \( (\alpha_{0a}', \beta_{ia}') \) and subsume the indicators \( 1(\tau_{a-1} < u \leq \tau_a) \) into the covariates vector \( Z_i(u) \) so that both \( \lambda_i(u) \) and \( \mu_i(u) \) can be written as \( e^{\beta Z_i(u)} \).

We shall assume that for each individual \( i \) we study quadruples \((Z_i(\cdot), V_i, X_i, C_i)\) where \( V_i \) is a positive left truncation time, \( X_i \) a failure time, \( C_i \) a right censoring time and \( Z_i(\cdot) \) a \( p \)-vector of possibly time dependent covariates. The \((Z_i(\cdot), V_i, X_i, C_i)\), \( 1 \leq i \leq u \), are independently and identically distributed. For each individual \( i \) selected in the study we observe \( V_i, T_i = X_i \wedge C_i, D_i = 1(X_i \leq C_i) \) and \( Z_i(u) \), for \( V_i < u \leq T_i \), conditionally on \( V_i < T_i \), i.e. we observe \( X_i \) and \( Z_i(\cdot) \) only `through' the 'filter' \( 1(V_i < u \leq C_i) \).

We shall not here discriminate between first sampling the cohort and then observing the 'filtered' process or first observing the filtered process and then sampling the filtered individuals. In both cases \( \varphi = m/n \) will denote the sampling fraction where \( n = \) cohort size and \( m = \) subcohort size. Details regarding these matters will be specified in the appendix.

Relative rate regression parameter estimates can be based on a pseudo partial likelihood which we present separately for the ‘shrink’ and ‘blow-up’ approaches.

\[
L^S(\beta) = \prod_{i=1}^{n} \left( \mu_i(T_i) \right)^{D_i} e^{-\int_{1(V_i < u \leq T_i)} S_i \lambda_i(u) \, du}.
\]

\[
L^B(\beta) = \prod_{i=1}^{n} \left[ \lambda_i(T_i) \right] e^{-\int_{1(V_i < u \leq T_i)} \lambda_i(u) \, du} [\exp \left( -\int_{1(V_i < u \leq T_i)} \lambda_i(u) \, du \right) S_i e^{-1(1-D_i)}].
\]

Here \( S_i \) stand for a subcohort sampling indicator, \( S_i = 1(0) \) indicating that \( i \) was sampled (not sampled) in the subcohort and \( P(S_i = 1) = \varphi \), the sampling fraction.
In the appendix we prove that the maximum likelihood statistic \( \hat{\beta} \) is a consistent and asymptotically normally distributed estimator of \( \beta \) with covariance matrix \( \sum \), say, and that \( \sum \) can be consistently estimated by \( I_n(\hat{\beta})^{-1} + I_n(\hat{\beta})^{-1} \Delta_n(\hat{\beta}) I_n(\hat{\beta})^{-1} \).

\( I_n(\hat{\beta}) \) stands for the \( (A + p) \times (A + p) \) dimensional observed information matrix at \( \hat{\beta} \) with \( j,k \)-th element \( -\frac{\partial^2}{\partial \beta_j \partial \beta_k} \log L(\beta) \) where for \( L(\beta) \) either \( L^S(\beta) \) or \( L^B(\beta) \) can be substituted.

\( \Delta_n(\hat{\beta}) \) stands for the \( (A + p) \times (A + p) \) dimensional empirical, finite population sampling, covariance matrix to account for the extra variability introduced by sampling from the cohort.

\[
\Delta_n(\hat{\beta}) = nq^{-1}(1-q) \int_0^1 \int_0^1 [Q(\hat{\beta}, u, v) - S_1(\hat{\beta}, u) S_1(\hat{\beta}, v)] \, du \, dv
\]

where for \( \Delta_n, Q \) and \( S_1 \) we must substitute \( \Delta_n^S(\Delta_n^B), Q^S(Q^B) \) and \( S_1^S(S_1^B) \) depending on whether the ‘shrink’ or ‘blow-up’ approach has been used.

\[
Q^S(\beta, u, v) = \frac{1}{m} \sum_{i=1}^n S_i Y_i(u) Y_i(v) Z_i(u) Z_i(v) e^{\beta' Z_i(u)} e^{\beta' Z_i(v)},
\]

\[
S_1^S(\beta, u) = \frac{1}{m} \sum_{i=1}^n S_i Y_i(u) Z_i(u) e^{\beta' Z_i(u)},
\]

and

\[
Q^B(\beta, u, v) = \frac{1}{m} \sum_{i=1}^n S_i(1 - D_i) Y_i(u) Y_i(v) Z_i(u) Z_i(v) e^{\beta' Z_i(u)} e^{\beta' Z_i(v)},
\]

\[
S_1^B(\beta, u) = \frac{1}{m} \sum_{i=1}^n S_i(1 - D_i) Y_i(u) Z_i(u) e^{\beta' Z_i(u)},
\]

where \( Y_i(u) = 1(V_i < u \leq T_i) \) and \( S_i \) is the indicator for subcohort selection. When the failure rate is constant on the whole observation interval and all covariates are time independent \( \Delta_n^S(\Delta_n^B) \) is easily calculated from:

\[
\Delta_n = nq^{-1}(1-q) [\tilde{Q} - \tilde{S}_1 \tilde{S}_1^T]
\]

with

\[
\tilde{Q}^S = \frac{1}{m} \sum_{i=1}^n S_i Z_i Z_i'(e^{\beta' Z_i} W_i)^2,
\]

\[
\tilde{S}_1^S = \frac{1}{m} \sum_{i=1}^n S_i Z_i e^{\beta' Z_i} W_i,
\]

\[
\tilde{Q}^B = \frac{1}{m} \sum_{i=1}^n S_i(1 - D_i) Z_i Z_i'(e^{\beta' Z_i} W_i)^2,
\]

\[
\tilde{S}_1^B = \frac{1}{m} \sum_{i=1}^n S_i(1 - D_i) Z_i e^{\beta' Z_i} W_i
\]

where \( W_i = T_i - V_i \).
Any exponential failure time regression, Poisson regression or nonlinear regression program, can be used to obtain the necessary maximum likelihood estimates \( \beta \) and \( I_0(\hat{\beta}) \). Obtaining estimates of \( \hat{Q} - \hat{S}_1 \hat{S}_1' \) is also easy, it is actually a sample covariance matrix and all statistical package include a routine to calculate such a matrix given a data matrix of observations. Of course, the observation have to be calculated beforehand by multiplying each \( Z_i \) with \( S_i e^{\hat{\beta}' X_i} W_i \), etc. and the covariance matrix must possibly be adjusted to get the correct divisor, i.e. m.

Likelihood ratio tests for \( \beta \) are available in most programs. Score and Wald tests will, on account of the adjusted covariance matrix, have to be calculated by 'hand'.

If it is suspected that the failure rate is not approximately constant for the study duration it is advisable to fit a piecewise exponential distribution to the failure times, i.e. \( \lambda_i = \sum_{a=1}^{A} 1(\tau_{a-1} < u \leq \tau_a) \lambda_{ia} \). If we consider failure for each interval \( (\tau_{a-1}, \tau_a] \) separately, then the \( i \)th subject experience a sequence of censorings at \( \tau_1, \tau_2, \ldots \) until final censoring or failure at \( t_i \) defined to fall in the \( A_i \)th interval, so that \( \tau_{A_i-1} < t_i \leq \tau_{A_i} \). Define for every subject a sequence of failure indicators \( d_{1i}, d_{2i}, \ldots, d_{Ai} \) with \( d_{ai} = 0 \) for \( 1 \leq a < A_i \) and \( d_{Ai} = 1 \) or 0 depending on whether \( i \) failed or was censored at \( t_i \) and a sequence of failure (exposure) times

\[
\begin{align*}
t_{1i}, t_{2i}, \ldots, t_{Ai} \quad \text{with} \quad t_{ai} = \tau_a - \tau_{a-1} \quad \text{for} \quad 1 \leq a < A_i
\end{align*}
\]

and

\[
\begin{align*}
t_{Ai} = t_i - \tau_{A_i-1}.
\end{align*}
\]

To fit such piecewise exponential models the data matrix must be augmented to contain \( A_i \) rows for each individual (see Attkin et al. 1989). This means that the total number of data matrix rows equals

\[
\begin{align*}
n_N \quad \text{where} \quad n_N = \sum_{i=1}^{n} A_i.
\end{align*}
\]

When fitting such a piecewise constant failure rate model we treat each row of this augmented data matrix as an independent subject and proceed as if fitting a constant failure rate model. \( D_n(\hat{\beta}) \) is calculated by applying the above formulas to this augmented data matrix.

For further details on fitting piecewise exponential distribution models and also on fitting time-dependent covariates and multiple failures (which follows the same procedure) see Attkin et al. (1989) sections 6.15, 6.16, 6.17 and 6.22.

Of course, this procedure treats a time dependent covariate as a step function, i.e. as piecewise constant. It is also possible to fit continuously varying time dependent covariates using for example non-linear regression programs, see Petersen (1986). However, the estimation of \( A \) becomes increasingly complex.

When using the 'shrink' approach failure times for individuals failing outside the subcohort must be set to zero or a (very) small positive number (depending on
how the regression program treats 0 (for example log (0) = 0)). When using GLIM or comparable facilities in S-plus, SAS, the ‘blow-up’ approach can be fit using an offset log \( q_n^{-1} \) or by blowing up the failure times of censored individuals in the subcohort beforehand with a factor \( q_n^{-1} \). When using the ‘blow-up’ approach with a piecewise constant failure rate just use an 'offset' with the augmented matrix or blow up each failure subtime \( t_{ia} \), i.e. \( q_n^{-1}t_{ia} \), \( 1 \leq a \leq A_i \).

4. Illustrative Analyses

The data for this illustration come from the Netherlands Cohort Study (NLCS) on diet, life-style and cancer that was started in 1986. The NLCS-cohort included 58,279 men and 62,573 women aged 55–69 years at the start of the study. At baseline, cohort members completed a self-administered questionnaire on dietary habits, potential confounders and other independent risk factors for cancer such as smoking habits, occupation and education. Following the case-cohort approach, a subcohort of 3,500 subjects was randomly sampled from the cohort after the baseline exposure measurement. The subcohort has been followed up biennially for vital status information in order to estimate the accumulated person-time in the cohort (VAN DEN BRANDT et al., 1990a). Incident cancer cases occurring in the cohort have been identified by record linkage to cancer registries and a pathology register (VAN DEN BRANDT et al., 1990b). This illustrative analysis pertains to the lung cancer incidence in the recently completed 3.3 year follow-up period. In this period a total of 617 cases of lung cancer were detected in the total cohort of 120,852 subjects. After excluding incident cases with in situ carcinoma, cases whose diagnosis was not microscopically confirmed and cases who reported a history of cancer other than skin cancer in the baseline questionnaire, 550 incident cases with lung carcinoma were available for analysis. After excluding prevalent cancer cases other than skin cancer from the subcohort of 3500 as well, 3346 subjects remained in this group.

For this illustration, the relationship between smoking habits (categorized as never/ex/current smokers) and lung cancer risk was analyzed with the proposed methods of case-cohort analysis using GLIM macros. To be concise, we only present results on relative rate estimates, confidence intervals, Mantel-Haenszel tests and test for trend in relative rates across exposure categories. In the stratified analysis, we stratified for gender and age (in three five-year categories). This was followed by relative rate regression analysis using the individual data, while adjusting for gender and age (again in 3 categories for reasons of comparability with the stratified analysis). Analyses were conducted with GLIM using the ‘shrink’ method and the ‘blow-up’ method of estimating the person years.

Recently, the software program EPICURE was released (PRESTON et al., 1993) which contains the survival analysis module PEANUTS that allows for case-cohort designs. The program is based on the likelihood equations for proportional
models in case-cohort studies as published by Prentice (1986a). We compared the regression results obtained with the GLIM-program to the results obtained by using EPICURE in our analyses when analyzing the smoking-lung cancer association.

In Table 2 (panel A), the results of the stratified analyses are shown for both the shrink and the blow-up approach. As expected, the relationship between smoking and lung cancer is very strong. Compared to never-smokers, the Mantel-Haenszel relative rate estimates for ex-smokers and for current smokers are 3.78 and 10.79, respectively, using the shrink method. The RR estimates obtained with the blow-up method are virtually identical. The uncorrected and corrected estimate of var (log $RR_{MH}$) differed slightly: for the contrast between ex- and never smokers the two estimates of var (log $RR_{MH}$) were 0.0839 and 0.0878, respectively, while for the contrast between current and never smokers these variance estimate were 0.0786 and 0.0820, respectively (using the shrink method). Both relative rate estimates were significantly different from 1 and the $\chi^2$-test for trend was also highly significant. The variance estimates and $\chi^2$-test values were again similar when using the blow-up method instead of the shrink method.

Table 2 (panel B) also shows the results of the relative rate regression analysis with the individual data. When an exponential distribution of failure times is assumed (i.e., a constant hazard), the association between smoking status and lung cancer is estimated essentially similar to the stratified analyses. It should be mentioned that we compare here the Mantel-Haenszel relative rate of the stratified analysis with the maximum likelihood estimate of the regression analysis. Also, the tests for trend are results from the score test and the likelihood ratio test, respectively. The uncorrected standard error estimates of the regression coefficients are 0.2477 and 0.2396 for the respective exposure contrasts (with the shrink method). Again, there is virtually no difference in relative rate estimates, confidence intervals and trend tests between the shrink and blow-up approach. When a piece-wise exponential distribution is assumed (with constant hazard rates per year of follow-up in the 3.3 year period), the results are essentially similar to the situation where an exponential distribution is assumed. Thus, the assumption of a constant hazard during the 3.3 years of follow-up is justified. Finally, in Table 2 (B3) and Table 3 the EPICURE-results using a proportional hazards model are presented. Again the results are similar to those obtained by our GLIM regression programs as far as estimates and standard errors are concerned. The $\chi^2$-value for the test for trend differs between the programs. This might be the result of the use of different likelihoods in the respective programs. Although in both situations the test for trend is highly significant, additional work is needed to evaluate the origin of difference between GLIM and EPICURE.

Together with the results of the programs we also listed in Table 2 the computing time needed to run the programs. Our GLIM analyses were run on a VAX 8650 VMS computer, while EPICURE-PEANUTS (version 1.8w) analyses were run on a DOS-PC with a 486 DX2 66 MHz processor. Although the two
<table>
<thead>
<tr>
<th>Type of analysis</th>
<th>Shrink method</th>
<th>Blow-up method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Never smoked</td>
<td>Ex-smoker</td>
</tr>
<tr>
<td>No. of cases in cohort</td>
<td>20</td>
<td>157</td>
</tr>
<tr>
<td>Person years in subcohort</td>
<td>3832</td>
<td>3588</td>
</tr>
<tr>
<td><strong>A</strong>. Stratified analysis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RR</td>
<td>1.00</td>
<td>3.78</td>
</tr>
<tr>
<td>95% CI</td>
<td>2.11–6.76</td>
<td>6.16–18.92</td>
</tr>
<tr>
<td>Trend χ² (p-value)</td>
<td>199.79 (&lt;0.001)</td>
<td>201.63 (&lt;0.001)</td>
</tr>
<tr>
<td>CPU-time¹</td>
<td>00.05:43</td>
<td></td>
</tr>
</tbody>
</table>

B. RR regression analysis

B1. Exponential model

<table>
<thead>
<tr>
<th>Beta</th>
<th>Shrink method</th>
<th>Blow-up method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Never smoked</td>
<td>Ex-smoker</td>
</tr>
<tr>
<td>RR</td>
<td>1.2948</td>
<td>2.2805</td>
</tr>
<tr>
<td>S.E. (beta)</td>
<td>0.2541</td>
<td>0.2454</td>
</tr>
<tr>
<td>RR</td>
<td>1.00</td>
<td>3.65</td>
</tr>
<tr>
<td>95% CI</td>
<td>2.22–6.04</td>
<td>6.05–15.82</td>
</tr>
<tr>
<td>Trend χ² (p-value)</td>
<td>217.12 (&lt;0.001)</td>
<td>216.87 (&lt;0.001)</td>
</tr>
<tr>
<td>CPU-time¹</td>
<td>00.01:45</td>
<td></td>
</tr>
</tbody>
</table>

B2. Piecewise exponential model

<table>
<thead>
<tr>
<th>Beta</th>
<th>Shrink method</th>
<th>Blow-up method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Never smoked</td>
<td>Ex-smoker</td>
</tr>
<tr>
<td>RR</td>
<td>1.2862</td>
<td>2.2707</td>
</tr>
<tr>
<td>S.E. (beta)</td>
<td>0.2367</td>
<td>0.2282</td>
</tr>
<tr>
<td>RR</td>
<td>1.00</td>
<td>3.62</td>
</tr>
<tr>
<td>95% CI</td>
<td>2.28–5.76</td>
<td>6.19–15.15</td>
</tr>
<tr>
<td>Trend χ² (p-value)</td>
<td>215.89 (&lt;0.001)</td>
<td>216.87 (&lt;0.001)</td>
</tr>
<tr>
<td>CPU-time¹</td>
<td>00.05:35</td>
<td></td>
</tr>
<tr>
<td>Type of analysis</td>
<td>Shrink method</td>
<td>Blow-up method</td>
</tr>
<tr>
<td>------------------</td>
<td>---------------</td>
<td>----------------</td>
</tr>
<tr>
<td></td>
<td>Never smoked</td>
<td>Ex-smoker</td>
</tr>
<tr>
<td>B3. Prop. hazards model (EPICURE)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Beta</td>
<td>1.295</td>
<td>2.280</td>
</tr>
<tr>
<td>S.E. (beta)</td>
<td>0.2552</td>
<td>0.2468</td>
</tr>
<tr>
<td>RR</td>
<td>1.00</td>
<td>3.65</td>
</tr>
<tr>
<td>95% CI</td>
<td>2.21–6.02</td>
<td>6.02–15.85</td>
</tr>
<tr>
<td>Trend $\chi^2$ (p-value)</td>
<td>291.13 (&lt;0.001)</td>
<td></td>
</tr>
<tr>
<td>PC-time $^d$</td>
<td>61.33:25</td>
<td></td>
</tr>
</tbody>
</table>

* GLIM-programs

$^d$ Using GLIM 3.77 version on VAX 8650

$^d$ Run on DOS-PC, 486 DX2, 66 MHz processor with 16 Mb memory (time was 34.51:35 on a DOS-PC, Pentium Processor, 66 MHz)
Table 3
Comparison of GLIM and EPICURE results on smoking and lung cancer in subgroup analysis

<table>
<thead>
<tr>
<th>Subgroup</th>
<th>No. of cases</th>
<th>Person years in subcohort</th>
<th>GLIM (shrink)</th>
<th>EPICURE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>(\beta)</td>
<td>s.e.</td>
</tr>
<tr>
<td>Women, 60—64* years</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Never smoked</td>
<td>6</td>
<td>1150</td>
<td>0.8191</td>
<td>0.6545</td>
</tr>
<tr>
<td>Ex-smoker</td>
<td>4</td>
<td>338</td>
<td>1.7695</td>
<td>0.5101</td>
</tr>
<tr>
<td>Current smoker</td>
<td>12</td>
<td>392</td>
<td>13.46 (&lt;0.001)</td>
<td>0.00:45</td>
</tr>
<tr>
<td>Trend-(\chi^2) (p-value)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Men, 65—69 years</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Never smoked</td>
<td>6</td>
<td>136</td>
<td>0.6829</td>
<td>0.4585</td>
</tr>
<tr>
<td>Ex-smoker</td>
<td>63</td>
<td>722</td>
<td>1.6204</td>
<td>0.4511</td>
</tr>
<tr>
<td>Current smoker</td>
<td>124</td>
<td>556</td>
<td>51.51 (&lt;0.001)</td>
<td>0.00:49</td>
</tr>
<tr>
<td>Trend-(\chi^2) (p-value)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age as continuous variable</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Women, 60—64* years</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Never smoked</td>
<td>6</td>
<td>1150</td>
<td>0.7962</td>
<td>0.6566</td>
</tr>
<tr>
<td>Ex-smoker</td>
<td>4</td>
<td>338</td>
<td>1.7376</td>
<td>0.5150</td>
</tr>
<tr>
<td>Current smoker</td>
<td>12</td>
<td>392</td>
<td>17.74 (&lt;0.001)</td>
<td>0.00:49</td>
</tr>
<tr>
<td>Trend-(\chi^2) (p-value)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Men, 65—69 years</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Never smoked</td>
<td>6</td>
<td>136</td>
<td>0.6814</td>
<td>0.4603</td>
</tr>
<tr>
<td>Ex-smoker</td>
<td>63</td>
<td>722</td>
<td>1.7026</td>
<td>0.4544</td>
</tr>
<tr>
<td>Current smoker</td>
<td>124</td>
<td>556</td>
<td>58.10 (&lt;0.001)</td>
<td>0.00:52</td>
</tr>
<tr>
<td>Trend-(\chi^2) (p-value)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* There were no female lung cancer cases in the age group 55—69 years who had never smoked.
computers are not directly comparable, the difference between required CPU time of the VAX and run time of the (stand-alone) PC is striking for this analysis: 00:01:45 versus 61:33:25 (hh:mm:ss), respectively. Comparable analyses 'by hand', i.e. not using functions or macros, in S-Plus and SAS-IML could be done in less than 6 minutes on a Pentium 60 MHz OS/2-PC.

We also evaluated the results of the GLIM and EPICURE programs in subgroup analyses. These subgroups were women aged 60–64 years (representing the situation with few cases \( n = 22 \) relative to the subcohort size \( n = 583 \)), and men aged 65–69 years (a subgroup with a large number of cases \( n = 195 \) compared to the subcohort \( n = 442 \)). In both situations almost identical estimates of the regression coefficients and standard errors were obtained, but again the test for trend in EPICURE yielded a larger \( \chi^2 \)-value as in GLIM. The required computing time by EPICURE was dramatically reduced in the female subgroup and resembled that of GLIM. In the male subgroup the required time by EPICURE was 25 minutes versus 49 seconds by GLIM. This indicates that the number of risk sets is the primary determinant of the required time, leading to long computing times in EPICURE when a moderately large number of cases is involved (Table 3).

In conclusion, the well-known strong positive dose-response relationship between smoking and lung cancer is reproduced in this case-cohort analysis. With two covariates in the model, the stratified analysis and the regression analyses yield essentially similar results. The advantage of the relative rate regression model using individual data is that it can easily be extended to more covariates. The regression coefficients and standard errors of the GLIM-model show great similarity with EPICURE results.

5. Discussion

The results presented in this paper were essentially (minus left-truncation) derived in 1990 and GLIM macros written to facilitate the analysis of the NLCS-cohort study on diet, lifestyle and cancer. The 'Poisson regression' approach was chosen over implementing the Prentice case-cohort model because it was judged to be less work than writing a SAS macro to calculate the necessary adapted covariance matrix and because the elementary stratified methods possible with this approach were familiar tools for the epidemiologists working on the NLCS-study. The results were not submitted for publication at the time.

In the mean time the program Epicure with facilities for case-cohort studies analyses has become available and a paper by Schouten et al. (1993) presenting a similar idea has been published. The Epicure program does not seem to be very efficiently coded given the extremely long run times with quite simple models. The approach presented in this paper is easily implemented with standard software and has short run times, even with large data sets, while the estimates are comparable with those obtained by Epicure. The results presented in Schouten et al.
(1993) were independently derived and are quite similar but they calculate the covariance matrix of the parameters differently using an estimating function approach. However, they derive the results only in the less general classical Poisson regression situation.

Considering the computational superiority over Epicure, the fact that we have in the meantime derived the results for very general filtering models and the fact that potential users of the GLIM macros would need an introduction to the implementations, we have decided to justify and present the results in this paper.

Acknowledgements

The authors thank Marion de Leeuw who efficiently coded the GLIM macros and Patrice Florax for typing the many versions of the manuscript with enormous drive and dedication.

Appendix

We refer to Andersen et al. (ABGK) (1993), chapters III, VI and VII for justification of the technical details of the following heuristic introduction.

We consider left-truncated and right-censored duration data observed on a finite interval which we take to be $[0, 1]$. That is we have lifetimes $X_1, \ldots, X_n$ and time-dependent $p$-dimensional covariate vectors $Z_1(\cdot), \ldots, Z_n(\cdot)$ on $[0, 1]$ such that the $X_i$ have hazard rate functions

$$h_i(t; \beta) = \exp(\beta Z_i(t)) = \exp(\beta_i^1 Z_{1i}(t) + \beta_i^2 Z_{2i}(t))$$

where

$$Z_{1i}(t) = Z_i(t) = (1(0 \leq t \leq \tau_1), 1(\tau_1 < t \leq \tau_2), \ldots, 1(\tau_{p_i-1} < t \leq 1))^T$$

for certain known $p_i \geq 1, 0 < \tau_1 < \ldots < \tau_{p_i-1} < 1$, and $p = p_1 + p_2$.

Exp$\{\beta_i^1 Z_{1i}(t)\}$ represents the (piecewise) constant baseline hazard rate and exp$\{\beta_i^2 Z_{2i}(t)\}$ the relative rate for individual $i$.

The $X_i$ are observed conditionally on $X_i > V_i$ for each $i$ where $V_1, \ldots, V_n$ are truncation times. These left-truncated durations are then subject to right-censoring at the censoring times $U_1, \ldots, U_n$. The sample of left-truncated, right-censored lifetimes and covariates consists of quadruples $(Z_i(\cdot), V_i, T_i, D_i)$ where $T_i = X_i \wedge U_i > V_i$ and $D_i = 1(T_i = X_i)$. The process $N_i(t) = 1(T_i \leq t, D_i = 1)$ is a counting process with intensity process $h_i(t; \beta) Y_i(t)$ where $Y_i(t) = 1(V_i < t \leq T_i)$.

The partial likelihood for this setup is proportional to

$$L(\beta) = \prod_{i=1}^n [\exp(\beta Z_i(T_i))]^{D_i} \exp\left(-\int_{V_i}^{T_i} \exp(\beta Z_i(u)) \, du\right).$$
An observation scheme that gives rise to a formally equivalent partial likelihood is obtained by observing \( n \) lifetimes \( X_1, \ldots, X_n \), and time-dependent covariate vectors \( Z_1(\cdot), \ldots, Z_n(\cdot) \) with hazard rate functions \( h_i(t; \beta) = \exp(\beta^T Z_i(t)) \) (\( Z_i \) and \( \beta \) as above), only on the intervals \( [V_i, U_i] \) determined by filtering processes \( Y_i(t) = 1(V_i < t \leq U_i), \quad V_i < U_i \). The filtered counting process \( N_i(t) = 1(V_i < X_i \leq t, U_i \geq X_i) \) again has intensity process \( h_i(t; \beta) Y_i(t) \) and the partial likelihood is formally equivalent to \( L(\beta) \).

The actual observed number of observations \( K \) under this scheme, is a random variable: \( \sum_{i=1}^{n} Y_i(X_i) \).

Individuals with \( X_i \leq V_i \) are not observed and no failure times \( X_i \) or covariate values are available for such individuals.

To evade problems with conditional probability spaces and in the light of the above formal equality, results were proved using a filtering observation scheme and not a (conditional) truncation scheme. This is also done in Lai and Ying (1989) and Gross and Huber-Carol (1992).

We assume given triples \( W_i = (Z_i, Y_i, N_i) \), \( i = 1, \ldots, n \), i.i.d. replicates of \( W = (Z, Y, N) \) where \( Y \) and \( Z \) are right continuous with left hand limits. The properties of \( W^{(n)} = (W_1, \ldots, W_n) \), as a process, are with respect to a filtration \((\mathcal{F}^{(n)}(t) : t \in [0, 1])\) on the \( n \)th sample space \((\Omega^{(n)}, \mathcal{F}^{(n)}, P^{(n)})\). Probability models \( P^{(n)} \) are parametrized by \( \beta \in \mathbb{R}^p \) and possible nuisance parameters \( \Phi \). \( Y \) is a filtering observation process as described above, i.e. \( Y(t) = 1(V < t \leq U) \), \( U > V \), and we assume independent (ABGK, page 163) but not necessarily noninformative (ABGK, page 165) filtering. We henceforth everywhere drop the superscript \((n)\).

The filtered counting process \((N(t), t \in [0, 1])\) has intensity process \( Y(t) \exp(\beta_0^T Z(t)) \) for a fixed \( p \)-dimensional column vector \( \beta_0 \).

Let us denote the log partial likelihood process evaluated at \( t = 1 \) by

\[
L_n(\beta, 1) = \sum_{i=1}^{n} \int_{0}^{1} \beta^T Z_i(u) \ dN_i(u) - \sum_{i=1}^{n} \int_{0}^{1} Y_i(u) e^{\beta^T Z_i(u)} \ du .
\]

In the case-cohort setup a random sample, the subcohort, of size \( m \) is drawn from the cohort of size \( n \) and \((Z_i, Y_i, N_i)\) are observed for all \( i \) in the subcohort and all failures in the cohort. Let \( S_i = 1 \) or 0 be a subcohort membership indicator, i.e. \( S_i = 1 \) indicating that \( i \) was sampled and \( P(S_i = 1) = m/n \).

The log pseudo partial likelihood for the case-cohort setup can then be formulated:

\[
\bar{L}_n(\beta, 1) = \sum_{i=1}^{n} \int_{0}^{1} \beta^T Z_i(u) \ dN_i(u) - \sum_{i=1}^{n} \int_{0}^{1} Y_i(u) e^{\beta^T Z_i(u)} \ du .
\]

With this pseudo likelihood we estimate 'blown-up' versions of the piecewise constant hazard rate \( \exp(\beta_{0i}^T Z_i(t)) \).
Write $U_n(\beta, t)$ for the first derivative $DL_n(\beta, t)$, of $L_n(\cdot \cdot t)$ at $\beta$, idem $\bar{U}_n(\beta, t)$
for $D\bar{L}_n(\beta, t)$, and $\bar{U}_n(\beta, t)$ for $D^2L_n(\beta, t)$ the second derivative. Assuming for
the moment, a constant hazard rate $\exp(\bar{\beta}_0) = \exp(\bar{\beta}_0 Z(t))$ and assuming the
existence of the expectation $E$ (relative to probability measures parametrized by $\beta_0$)
we have

$$E\bar{U}_n(\beta, 1) = E[\bar{U}_n(\beta, 1) - U_n(\beta_0, 1)].$$

It can easily be shown that

$$E\bar{U}_n(\beta, 1) = 0$$

when

$$\beta_1 = \beta_{01} - \log \frac{m}{n} \left( \text{or } \exp(\beta_1) = \frac{n}{m} \exp(\beta_0) \right) \quad \text{and} \quad \beta_2 = \beta_{02}$$

where we write $\beta = (\beta_1, \beta_2)'$ to match $\beta_0 = (\beta_{01}, \beta_{02})'$.

We shall therefore write the log pseudo partial likelihood as $(Q_n = m/n)$

$$\bar{L}_n(\beta, 1) = \sum_{i=1}^{n} \int_0^1 \beta' Z_i(u) \, dN_i(u) - \sum_{i=1}^{n} \int_0^1 S_i Y_i(u) Q_n^{-1} e^{\beta' Z_i(u)} \, du$$

to indicate that the parameter being estimated is actually $\beta_{01} - \log Q_n$ and to obtain
an immediate bias corrected estimate of $\beta_{01}$.

The log pseudo partial likelihood formulated above is based on first sampling
the subcohort and then observing the filtered processes $(W_1, \ldots, W_n)$. In practice
the filtered observations could be given and a subcohort of size $k$ would have to be
drawn from the $K = \sum_{i=1}^{n} Y_i(X_i)$ observed filtered cohort members where $K$, and
thus $k$, is a random number.

The log pseudo partial likelihood for the case-cohort setup in this second case

$$\bar{L}_n(\beta, 1) = \sum_{i=1}^{n} \int_0^1 \beta' Z_i(u) \, dN_i(u) - \sum_{i=1}^{n} \int_0^1 \bar{S}_i Y_i(u) Q_k^{-1} e^{\beta' Z_i(u)} \, du.$$}

Here $\bar{S}_i$ again denotes with 1 or 0 whether $i$ was sampled, but here a random
sample, without replacement, is drawn from the $K$ observed filtered individuals
and $P(\bar{S}_i = 1 \mid K) = k/K (= Q_k)$ and $\bar{S}_i \equiv 0$ if $Y_i(X_i) = 0$.

Here too we write $\bar{L}_n(\beta, 1)$ for $D^2\bar{L}_n(\beta, 1)$.

Taking 3 sets of necessary assumptions from BORGAN (1984, page 14) and 5
sets of slightly adapted assumptions from SELF and PRENTICE (1988, page 68) the
three propositions below can be proved for the $\bar{L}_n(\beta, 1)$ setup. Exact formulations
of the assumptions and sketches of the proofs can be found in VOLOVICS and VAN
DEN BRANDT (1995). Formulations and proofs for the $\bar{L}_n(\beta, 1)$ setup follow self-
evidently.
To formulate the propositions we will need the following notation and assumptions:

\[
\tilde{S}_n^1(\beta, u) = \sum_{i=1}^{n} \tilde{S}_i Z_i(u) Y_i(u) e^{\beta Z_i(u)}, \\
\tilde{Q}_n(\beta, u, v) = \sum_{i=1}^{n} \tilde{S}_i Z_i(u) Z_i(v)' Y_i(u) Y_i(v) e^{\beta Z_i(u)} e^{\beta Z_i(v)}.
\]

For the analogous \( \tilde{S}_n^1(\beta, u) \) and \( \tilde{Q}_n(\beta, u, v) \) in the \( L_n(\beta, 1) \) setup substitute \( \tilde{S}_i \) for \( \tilde{S}_i \) on the right hand side.

To obtain \( S_n \), \( Q_n \) in the full cohort setting remove the \( \tilde{S}_i \) on the right hand side of the above formulas.

\[
P(Y(t) = 1 \text{ for some } t \in [0, 1]) = \pi > 0,
\]

\[
\sup_{\beta \in B_0, t \in [0, 1]} \left| \frac{1}{n} S_n^1(\beta, t) - s^1_t(\beta, t) \right| \to 0, \quad \text{as } n \to \infty, \quad \text{for } 1 \leq j \leq p,
\]

\[
\sup_{\beta \in B_0, (u, v) \in [0, 1]^2} \left| \frac{1}{n} Q_n(\beta, u, v) - q_{\beta}(u, v) \right| \to 0,
\]

as \( n \to \infty \), for \( 1 \leq j, k \leq p \)

for \( s^1 \) defined and bounded on \( B_0 \times [0, 1] \), \( q \) defined and bounded on \( B_0 \times [0, 1]^2 \) where \( B_0 \) is a neighbourhood of \( \beta_0 \).

**Proposition 1.** With probability tending to 1, the pseudo partial likelihood equation \( \tilde{U}_n(\beta) = 0 \) has exactly one consistent solution \( \hat{\beta}_K \).

**Proposition 2.**

\[
\sqrt{n} \left( \hat{\beta}_K - \beta_0 \right) \overset{D}{\to} N \left( 0, \pi \sum_0^{-1} + \pi \sum_0^{-1} A_0 \pi \sum_0^{-1} \right)
\]

as \( n \to \infty \)

where

\[
A_0 = \alpha^{-1}(1 - \alpha) \frac{1}{1} \int_0^1 \left[ \pi^{-1} q(\beta_0, u, v) - \pi^{-1} s^1(\beta_0, u, v) \right] du dv.
\]

**Proposition 3.** \( -\frac{1}{K} \tilde{I}_n(\hat{\beta}_K, 1) \) is a consistent estimator of \( \pi^{-1} \sum_0 \) and

\[
\tilde{L}_n(\hat{\beta}_K) = \int_0^1 \int_0^1 \tilde{C}_K(\hat{\beta}_K, u, v) du dv \quad \text{is a consistent estimator of } A_0
\]

where

\[
\tilde{C}_K(\hat{\beta}_K, u, v) = \frac{K}{k} \left( 1 - \frac{k}{K} \right)
\]

\[
\times \left[ \frac{1}{k} \tilde{Q}_n(\hat{\beta}_K, u, v) - \frac{1}{k} \tilde{S}_n^1(\hat{\beta}_K, u) \frac{1}{k} \tilde{S}_n^1(\hat{\beta}_K, v)' \right].
\]
The equivalent propositions for the $\bar{L}_n(\beta, 1)$ setup are obtained by substituting $\bar{S}^1_n$ for $S^1_n$, $\bar{Q}_n$ for $Q_n$, $n$ for $K$, $m$ for $k$ and $1$ for $\pi$.

From these propositions asymptotic results for the Wald, score and likelihood ratio tests can be derived in the usual way. Asymptotics for the elementary methods based on summary tables mentioned in § 3 can be derived using the occurrence/exposure rate models of BORGAN (1984), pp. 9–12 and procedures as described in WITTING and NÖLLE (1970), pp. 85–92.

Equivalent assumptions and propositions can be formulated and proved for the ‘blow-up’ version.

References


Alexander Volovics
Department of Methodology and Statistics
Maastricht University
P.O. Box 616
6200 MD Maastricht
The Netherlands

Piet A. van den Brandt
Department of Epidemiology
Maastricht University
P.O. Box 616
6200 MD Maastricht
The Netherlands

Received, September 1995
Revised, December 1996
Accepted, December 1996