PARTIAL LIKELIHOOD RATIO TESTS FOR THE COX MODEL UNDER COMPLEX SAMPLING

THOMAS LUMLEY AND ALASTAIR SCOTT

ABSTRACT. We develop an analogue of the likelihood ratio test for Cox proportional hazards models fitted to sample survey data. We look at methods for computing the asymptotic distribution and at ways of improving the small sample performance. The methods are illustrated with examples using data from the National Health and Nutrition Examination Survey (NHANES) and from a stratified case-cohort study.

Keywords: proportional hazards; sample surveys; case-cohort design; weighting; Rao-Scott tests

1. Background

Time-to-event data are important in many areas, particularly the medical and social sciences, and complex probability samples are an increasingly common source of such data. These include multistage national surveys such as the National Health and Nutrition Examination Surveys' (NHANES) Linked Mortality Files[1, 2, 3, 4, 5], as well as stratified two-phase epidemiologic studies using case-cohort [6, 7, 8, 9] and complex case-control [10] designs. The wide range of applications of survival analysis to survey data is illustrated by a Google Scholar search that found more than 17,000 items containing both “NHANES” and “survival analysis”.

What do researchers analyzing data sets like NHANES want?

If the data had come from a simple random sample, there would be no question - they would use a standard statistical package to fit a proportional hazards model, or whatever other technique that they thought appropriate. When the data is collected via a complex survey, there are problems with the technical details of the analysis. However, the population remains unchanged and researchers still want to answer the same questions about it. Thus, they want to fit the same models and to estimate the same quantities as they would with a simple random sample from the same population. Moreover, in our experience, they want to do this using methods and software that mimic familiar standard methods and software as closely as possible.

Software to fit the Cox model to survey-sampled data is now widely available, using the weighted partial likelihood approach of Binder [11] and generalizations that make use of whole-cohort or whole-population information [12, 13, 14, 15]. One major gap in this software is the likelihood-ratio test and related quantities. Although SAS and SUDAAN both calculate and report a weighted log partial likelihood ratio, the only reference distribution available for it has been the $\chi^2$ distribution that would be appropriate under simple random sampling. The main purpose of this paper is to fill this gap by developing analogues of the partial likelihood ratio test for survey data.
The sampling distribution of a modified log likelihood-ratio statistic with survey data has been developed and found useful in other settings. Rao and Scott [16, 17] developed a widely-used set of tests in contingency tables for data from complex surveys. Although the primary aim for the Rao–Scott tests was a simple design-effect correction for tests that ignored the nature of the sampling, the tests were subsequently found to be more reliable in small samples than the Wald-type tests[18]. Rotnitzky and Jewell [19] computed the asymptotic distribution of the working-model deviance in GEE estimation for longitudinal data. Vuong [20] gave general results for misspecified models under independent sampling. In all these cases the asymptotic distribution of the likelihood-ratio statistic when comparing nested models is a linear combination of \( \chi^2_1 \) random variables. The coefficients in the linear combination are the eigenvalues of the product of the information matrix and the asymptotic covariance matrix of the parameters. For correctly-specified models under i.i.d. sampling, this product reduces to the identity matrix and the distribution for testing that \( q \) of the parameters are zero reduces to the familiar \( \chi^2_q \). It turns out that very similar results hold for the Cox model.

In section 2, we derive the asymptotic sampling distribution of the partial likelihood ratio test statistic under complex sampling and discuss computational issues and some novel small-sample approximations. The accuracy of these approximations in data with small numbers of degrees of freedom is examined in Section 3, showing that the likelihood ratio test has slightly better performance than a Wald test and that the small-sample adjustments further improve performance. Section 4 contains two worked examples, a real example from the 3rd National Health and Nutrition Examination Survey (NHANES III) and an artificial example of a two-phase design using real data from the National Wilms Tumor Study Group trials.

2. Some Theory

2.1. Basic set-up. We want to fit a Cox proportional hazards model

\[
h(t \mid z) = h_0(t)e^{\beta^T z(t)},
\]

for the failure time, \( T \), associated with a vector of (possible time-varying) covariates, \( z \). Here \( h_0(t) \) is an unspecified baseline hazard function and \( \beta \) is a vector of unknown parameters. Typically, the failure time is subject to right censoring. Let \( C \) denote the censoring time, \( \hat{T} = \min(T, C) \), \( \Delta = I(T \leq C) \) and \( Y(t) = I(\hat{T} \geq t) \).

Suppose that we have finite population of \( N \) units with measurements \( (\hat{T}_i, \Delta_i, z_i(.)) \) associated with the \( i \)th unit \( (i = 1, \ldots, N) \). For simplicity, we shall follow Lin [21] and assume that the values \( (\hat{T}_i, \Delta_i, z_i(.)): i = 1, \ldots, N \) are drawn independently from some superpopulation in which the proportional hazards model (1) applies. Then if we had data from every unit in the population, we would estimate \( \beta \) by maximizing the population partial log-likelihood

\[
\ell_{\text{pop}}(\beta) = \sum_{i=1}^N \Delta_i \left[ \beta^T z_i(\hat{T}_i) - \log S(\beta, \hat{T}_i) \right],
\]
where \( S(\beta, t) = \sum_{j=1}^{N} Y_j(t) e^{\beta^T z_j(t)} \).

FIG 1 ABOUT HERE

Of course, most real populations do not look much like a set of independent observations from a superpopulation, at least at first glance. For example, Figure 1 shows stroke death rates across counties in the US (the population from which NHANES is drawn). These rates show clear evidence of spatial auto-correlation which persists to some degree even after adjusting for measured risk factors such as smoking and diet. In fact, the spatial structure of stroke death rates is more complex and detailed, since the smoothing replaces detail with a general appearance of auto-correlation. We can accommodate such patterns to a surprising extent within the basic independence model by augmenting the measurements \((\tilde{T}_i, \Delta_i, z_i(.))\) with other variables correlated with \(T_i\) (for example, longitude and latitude) and arranging the population by these extra variables, and this is how our simulations are constructed.

Moreover, we note that \( \hat{\beta}_{\text{pop}} \), the value that maximizes \( \ell_{\text{pop}}(\beta) \), will give a consistent estimator of \( \beta \) under even more general assumptions (basically, provided the finite population is a self-weighting sample from some superpopulation with the right margins, cf using estimating equations with a “working independence model” for longitudinal data as in Liang & Zeger, 1986).

In practice, we do not have data for the whole population but only from a subset, \( s \), of units chosen using some probability sampling scheme. For any \( i \in s \), let \( \pi_i \) be the probability of selecting the \( i \)th unit with the given scheme and let \( w_i = 1/\pi_i \) be the associated sampling weight. (In most large surveys, these quantities will be adjusted in various ways to account for such things as missing data and frame errors.) To make our output look more like the output that we would get from a standard survival package with simple random sampling, we shall re-standardize the weights so that \( \sum_{i \in s} w_i = n \). Then \( \beta_0 \), the parameter for which \( \hat{\beta}_{\text{pop}} \) is consistent, can be estimated from the sample by maximising a weighted log partial likelihood

\[
\ell_W(\beta) = \sum_{i \in \text{sample}} w_i \Delta_i \left[ \beta^T z_i(\tilde{T}_i) - \log S_W(\beta, \tilde{T}_i) \right]
\]

where

\[
S_W(\beta, t) = \sum_{j \in \text{sample}} w_j Y_j(t) e^{\beta^T z_j(t)}.
\]

Note that \( \beta_0 \) is the same parameter that anyone would be estimating if they had a simple random sample and used a standard survival program. If this is a sensible target with a simple random sample, then presumably it is equally sensible with a more complex sample. The resulting weighted partial likelihood estimator, \( \hat{\beta} \), is now standard, and is widely available in statistical packages (in stcox in Stata[22] or svycoxph in the R package survey, for example[23, 24]). Since \( \ell_W(\beta) \), and hence \( \hat{\beta} \), is a function of estimated totals, we can produce an estimate of \( \text{Var}\{\hat{\beta}\} \) using
standard linearization arguments [11, 21]. This is available routinely in all the programs above.

Now write \( \beta = (\alpha, \gamma) \), with \( \alpha \) being of dimension \( p \) and \( \gamma \) of dimension \( q \), and suppose that we want to test the hypothesis that \( \gamma = 0 \). The proposed test statistic has the form

\[
D_W = -2 \left[ \ell_W(\hat{\beta}_R) - \ell_W(\hat{\beta}) \right]
\]

where \( \ell_W(\beta) \) is the weighted log partial likelihood given in (2), \( \hat{\beta} \) is its unrestricted maximizer as above, and \( \hat{\beta}_R \) is the value that maximizes \( \ell_W(\beta) \) under the restriction \( \gamma = 0 \). There is no significant loss of generality in considering tests of the form \( \gamma = 0 \), as other tests of nested hypotheses can be reparametrized as testing for zero coefficients.

2.2. Asymptotics. We shall adopt the asymptotic framework used in Lin [21] in which complex samples are drawn from a sequence of superpopulations obtained by i.i.d. sampling. As we tried to explain above, this can result in superpopulations that exhibit a substantial amount of spatial auto-correlation. Alternatively, we could use a more complicated generating process that incorporates clustering. However, as explained in Graubard and Korn [25], provided that the number of clusters in the population is large, as it is in NHANES and similar surveys, the asymptotics are driven by the sample size rather than the underlying population size and the asymptotic results are largely unaffected. Thus we shall assume that a sequence of populations, indexed by \( \nu \), is generated by drawing i.i.d. observations, \( \{ \bar{T}_i, \Delta_i, z_i(\cdot) : i = 1, \ldots, N_\nu \} \), from a superpopulation in which the Cox model (1) applies. Then a sample of size \( n_\nu < N_\nu \) is drawn from the \( \nu \)th finite population using some complex probability sampling design. As \( \nu \to \infty \), we assume that \( n_\nu \to \infty, N_\nu \to \infty \), and \( \lim \sup n_\nu/N_\nu < 1 \).

The sequence of normalising constants will depend on how the sampling is done, but we shall assume here that the sequence of sampling designs is such that \( \sqrt{n_\nu}(\hat{\beta} - \beta_0) \) converges to a non-degenerate multivariate Normal, say \( N(0, V_\beta) \) and that \( n_\nu^{-1} \ell(\beta_0) \) converges in probability to a non-zero constant as \( \nu \to \infty \). The necessary weak convergence has so far been rigorously verified only in a limited number of settings[21, 26, 27], but the experience in simulations is that approximate normality of \( \hat{\beta} \) obtains more generally. (For simplicity, we omit the subscript \( \nu \) from here on.)

Let \( J_W = -\partial^2 \ell_W(\beta)/\partial^2 \beta \) denote the observed information matrix. It follows from the assumptions above that \( n^{-1} \text{E} \{ J_W \} \) converges to some matrix, say \( I = \begin{pmatrix} I_{\alpha\alpha} & I_{\alpha\gamma} \\ I_{\gamma\alpha} & I_{\gamma\gamma} \end{pmatrix} \), which we assume to be non-singular. Note that under simple random sampling, the limit of var \( \{ \sqrt{n}(\hat{\beta} - \beta) \} \), which we shall denote by \( V_\beta^* = \begin{pmatrix} V_{\alpha\alpha}^* & V_{\alpha\gamma}^* \\ V_{\gamma\alpha}^* & V_{\gamma\gamma}^* \end{pmatrix} \), is equal to \( I^{-1} \) so that, using standard results for the inverse of a partitioned matrix, \( V_{\gamma\gamma}^* \) can be written in the form \( V_{\gamma\gamma}^* = (I_{\gamma\gamma} - I_{\gamma\alpha} I_{\alpha\alpha}^{-1} I_{\alpha\gamma})^{-1} \). We write \( V_{\gamma\gamma} \) for the corresponding \( q \times q \) block of \( V_\beta \).
**Theorem.** Under the null hypothesis $H_0 : \gamma = 0$, $D_W = -2 \left[ \ell_W(\hat{\beta}_R) - \ell_W(\hat{\beta}) \right]$ converges in distribution to a linear combination of $q$ independent $\chi^2_1$ random variables. The coefficients in the asymptotic linear combination are the eigenvalues of

$$
\Gamma = (I_{\gamma\gamma} - I_{\gamma\alpha}I_{\alpha\gamma}) V_{\gamma\gamma} = V_{\gamma\gamma}^{-1} V_{\gamma\gamma}.
$$

□

The derivation of the asymptotic sampling distribution follows the same lines as in the classical i.i.d. setting. A Taylor series expansion is used to show that the statistic is asymptotically equivalent to a quadratic form in the parameters, and the result then follows from classical linear-model theory on the distribution of quadratic forms. A sketch of the details is given in the Appendix. Obviously, $\Gamma$ is equal to the identity matrix under simple random sampling and the asymptotic null distribution reduces to the usual $\chi^2_q$ distribution.

An estimate of $V_\beta$ is already computed by all software for survey data that implements the Cox model, and we can estimate $I(\beta_0)$ by $J_W(\hat{\beta})/n$. This enables us to estimate $\Gamma$, and hence its eigenvalues, and we can use these to estimate $p$-values and percentage points of the null distribution of $D_W$. Note that in contrast to the estimate of $\Gamma$, which is a simple mean and governed by the law of large numbers, the estimate of $Var \{ \hat{\beta} \}$ uses an empirical variance that is often based on relatively few degrees of freedom, even in very large surveys. For example, a two-year cycle of the current continuous NHANES survey has approximately 15 degrees of freedom. In these situations it is important to take account of the variability of the estimated eigenvalues. We explore this in more detail in later sections.

We have assumed that the proportional hazards model is true in the superpopulation but none of our results depend on this assumption; under mild regularity conditions $\hat{\beta}$ is still asymptotically Normal and converges to $\beta^* = \left( \begin{array}{c} \alpha^* \\ \gamma^* \end{array} \right)$, the limit of the so-called census parameter $\hat{\beta}_\text{pop}$, the value that maximizes $\ell_\text{pop}(\beta)$ [21]. The partial likelihood ratio test is now of the null hypothesis $\gamma^* = 0$; whether this is a useful hypothesis to test depends on the degree of misspecification and the purpose of the test.

2.3. **Case-cohort studies.** Case–cohort designs are a special case of the complex sampling approach we have described, but they can also be analyzed using martingale-based estimators[28, 13] rather than with sampling weights. These estimators maximize a weighted pseudolikelihood, and Therneau and Li [29] have shown how to obtain these estimators from Cox regression software, constructing the pseudolikelihood as a partial likelihood for a modified and weighted data set. As our derivation of the asymptotic sampling distribution of the log likelihood ratio depends only on asymptotic Normality for the regression parameter estimates, it also applies to the pseudolikelihood in the martingale-based case–cohort estimators. In fact, since the derivation of the asymptotic sampling distribution only requires asymptotic Normality and the ability to estimate the variance, this approach can be applied much more generally to pseudo-likelihood estimation, and potentially to composite likelihood[30].

The two matrices needed to compute the coefficients of the linear combination are the variance matrix of $\hat{\gamma}$, which is part of the standard output of a case–cohort
analysis, and the second derivative of the pseudolikelihood, which is a byproduct of
the algorithm of Therneau and Li [29], the quantity they denote by \( I \), equivalent
to \( J_W \) in the notation of our Appendix. Note that the martingale-based estimators
are not Horvitz–Thompson estimators, and it is not clear whether the results under
misspecified models also extend to such estimators.

2.4. Power. For single-parameter hypotheses the partial likelihood ratio test is
asymptotically equivalent to the Wald test based on \( \chi^2 = n\hat{\gamma}^T V_{\gamma\gamma}^{-1}\hat{\gamma} \), as is true
for independently-sampled data. When testing multiparameter hypotheses, the
partial likelihood ratio weights departures from the null differently than the Wald
test does. The partial likelihood ratio test weights the departures the same way as
they would be weighted in a simple random sample or with complete population
data. The Wald test gives less weight to departures in directions where the survey
sample is relatively less informative.

Giving more weight to the same alternatives that we would with complete pop-
ulation data is a useful general rule, reflecting the philosophy that analysis of data
from a survey sample should mirror the analysis for simple random sample from
the same population as closely as possible. However, neither test dominates the
other and neither set of weightings is uniformly preferable. In some situations the
differences in relative precision between population and survey estimates reflect the
priorities of the original survey designers, such as the higher precision given to con-
trasts between racial and ethnic minorities by oversampling in NHANES, and the
Wald test might be appropriate here. In other situations the differences are sim-
ply a matter of feasibility, such as the lower precision for coefficients of variables
that are strongly correlated within cluster, and the likelihood-ratio test might be
preferred.

2.5. Evaluating the null distribution. The linear combination of \( p \) \( \chi^2 \) ran-
don variables required for the asymptotic distribution is not one of the standard
distributions provided in statistical software. The usual approximation in survey
statistics is a Satterthwaite approximation that matches the mean and variance of
the distribution to a scaled \( \chi^2 \) distribution with \( k \leq p \) degrees of freedom. \( \tau^2 \chi^2_k \).
In many large surveys, the Satterthwaite approximation is adequate as far out in
the tail as the asymptotics apply. In some genomic case-cohort studies, however,
large sample sizes and massive multiple testing make it possible and necessary to
compute smaller \( p \)-values more accurately.

The asymptotic distribution can be obtained by numerically integrating the
characteristic function[31] or summing an infinite series of \( F \) percentiles[32]. A
saddlepoint approximation[33] is also available, giving higher accuracy than the
Satterthwaite approximation with easier implementation than the integration and
infinite-series approaches.

FIG 2 ABOUT HERE

Figure 2 shows that the Satterthwaite approximation can be seriously anticon-
servative for small \( p \)-values. In Figure 2 the saddlepoint approximation is almost
indistinguishable from the infinite series computation up to the level of accuracy
of the series computation. The infinite series is for the left tail of the distribution,
with the right tail being obtained by subtraction and thus having an absolute error of at least machine epsilon, approximately $2 \times 10^{-16}$ on current hardware. Saddlepoint approximations often have approximately constant relative error even far out in the tails, as seen here. The saddlepoint approximation is likely to be especially useful for large-scale multiple testing, for example in genetic and genomic studies where data from complex survey designs are being used increasingly.

### 2.6. Small-sample improvements

In designs with small numbers of degrees of freedom, such as the NHANES design, the variance matrix of $\hat{\beta}$ is not very accurately estimated and so we would expect the test statistic to be overdispersed relative to its asymptotic distribution. In linear regression we would correct this overdispersion by comparing the test statistic to an $F$ distribution with denominator degrees of freedom $m$ based on the design. Motivated by these $F$ tests we propose dividing the $\chi^2_1$ variables in the asymptotic distribution by a common $\chi^2_m$ variable representing the uncertainty in the estimated variance. Although the likelihood ratio statistic will not have exactly an $F$ distribution it is reasonable to hope for an improvement in small-sample performance, as has been seen in other contexts in survey statistics[18].

Applying this adjustment is trivial for the Satterthwaite approximation, where the $\tau^2 \chi^2$ null distribution is simply replaced by $\frac{\tau^2}{m} F_{k,m}$. The adjustment is still straightforward for the saddlepoint and characteristic function approaches, since the linear combination of $F$ variables can be transformed to a linear combination of $\chi^2$, as long as negative multipliers are allowed (see Appendix C).

### 2.7. Implementation

The series and integration approaches to the linear combination of $\chi^2$ distributions are implemented in the R package CompQuadForm[34], using C++ translations of the Applied Statistics algorithms AS155 and AS204. With the default settings these have absolute error close to $10^{-7}$ and $10^{-9}$ respectively. The error analysis in AS204 does not appear to include rounding error, and it will give inaccurate results regardless of input settings when the true $p$-value is in the vicinity of machine epsilon. AS155 will correctly report non-convergence if used for very small $p$-values. The R package survey[35, 23] uses CompQuadForm and implementations of the Satterthwaite and saddlepoint approximations to provide functions pchisqsum and pFsum for tail probabilities for the sum of $\chi^2$ and sum of $F$ distributions respectively.

The regTermTest function in the survey package for R provides both Wald and likelihood ratio tests for terms or sets of terms in a generalized linear model or Cox model, defaulting to the saddlepoint approximation and $F$-based tests.

### 3. Simulations

We conducted two sets of simulations, using scenarios loosely based on the sample designs used in the examples of the next section.

The first set is an attempt to mimic a large national survey with a relatively small number of large clusters. The code used to generate population values is given in Appendix B. There are 4 potential predictor variables, $\{z_1, z_2, z_3, z_4\}$, survival times are generated according to a Cox model with constant baseline hazard and $\beta^T = (0.25, 0, 0, 0)$ (so that only the first variable, $z_1$, affects survival times), and there is approximately 33% Type 1 censoring. Finite populations of $N = 100,000$ units are split into 1000 clusters of 100 units, with clusters grouped by the value
of $z_1$ and the latent survival time. Clusters are then sampled with probabilities proportional to their mean latent survival times. Design effects (inflation of variance compared to simple random sampling) for the means of the three null variables vary widely, working out at 1, 13 and 18 for $z_2$, $z_3$, and $z_4$ respectively. The eigenvalues, however are much less variable: 1.56, 1.05, and 0.96. The coefficient of variation of the sampling weights was 0.41.

We compared the observed significance levels of four different procedures for testing the null hypothesis that variables $z_2$, $z_3$, and $z_4$ have no effect, using two sample sizes: $c = 32$ clusters ($n = 3200$) and $c = 60$ clusters ($n = 6000$). Two of the procedures were variants of the proposed likelihood-ratio test, using the saddlepoint approximation to linear combinations of $\chi^2_1$ and $F_1,c-4$ random variables as reference distributions. The other two were based on the Wald test, using $\chi^2_3$ and $3F_3,c-4$ as reference distributions. Results for $c = 32$ are shown in Figure 3 and for $c = 60$ in Figure 4.

FIGURES 3 and 4 ABOUT HERE

In large studies such as NHANES, the number of observations is not the same in every cluster. We repeated the simulation using 32 clusters with a population of 700 clusters of size 50 and 325 clusters of size 200 to judge the effect of varying cluster sizes keeping the average cluster size approximately the same. The results were almost identical to those shown in Figure 3 and are not shown.

We see that the likelihood-ratio tests give uniformly better control of test size than their Wald counterparts, and that using an F-based, instead of a $\chi^2$-based, approximation for the null distribution gives a substantial improvement in the performance of both tests. The observed significance levels of the $F$-based likelihood-ratio test are very close to the nominal values, even at very small levels and with a small number of clusters. The Wald $\chi^2$ test is very anti-conservative with 32 clusters and still reasonably anti-conservative even with 60 clusters. Both the other tests are somewhat anti-conservative with 32 clusters but perform reasonably well with 60 clusters.

One possible explanation for the improved performance of the likelihood ratio test is that, when the variance matrix is not well estimated, the estimated eigenvalues tend to be more dispersed than the true eigenvalues, leading to a reference distribution with heavier tails. For example, the median estimated eigenvalues for the 32-cluster simulation are 1.62, 1.03, and 0.65, and for the 60-cluster simulation are 1.48, 0.98, and 0.70. Another contributing factor is that the Wald test requires inverting $V_{\gamma\gamma}$ where the LRT requires inverting $V^*_\gamma\gamma$, and the latter matrix is much more accurately estimated.

The second set of simulations is based on a stratified case-cohort design. The code used to generate the data is given in Appendix B. There are three, correlated Normal, potential predictor variables $\{x_1, x_2, z\}$ with $x_1$ and $x_2$ measured on the whole cohort (ie, at phase one) and $z$ measured only on a subsample. The subsample consists of all the cases and a stratified reference subcohort, stratified on $I(x_1 > 0)$ and $I(x_2 > 0)$. The reference subcohort includes 100 observations from three of the strata and all available observations from the remaining, smaller, stratum. The analysis used the Horvitz-Thompson estimator, not the martingale-based estimator.
Table 1. Survival by race/ethnicity, in data from NHANES III

<table>
<thead>
<tr>
<th>Race/ethnicity by age interaction</th>
<th>( \chi^2 ) Wald</th>
<th>( \chi^2 ) LRT</th>
<th>( F ) Wald</th>
<th>( F ) LRT</th>
</tr>
</thead>
<tbody>
<tr>
<td>plus main effect of race/ethnicity</td>
<td>3.0 \times 10^{-12}</td>
<td>1.3 \times 10^{-6}</td>
<td>2.5 \times 10^{-7}</td>
<td>2.6 \times 10^{-5}</td>
</tr>
</tbody>
</table>

of Borgan et al. [13]. The coefficient of variation of the weights was approximately 0.9.

The design degrees of freedom are large enough that the difference between \( \chi^2 \) and \( F \) distributions is negligible, so only the \( \chi^2 \)-based tests were done. Figure 5 shows a quantile-quantile plot of the \(-\log_{10}p\)-values from the Wald test and the partial likelihood ratio test, based on 50,000 simulations. Both tests perform well; there is a slight suggestion that the likelihood ratio test may be less conservative.

FIGURE 5 ABOUT HERE

4. Examples

4.1. A large multistage survey. The National Health and Nutrition Examination Surveys (NHANES) carried out by the US National Center for Health Statistics are multistage stratified samples of the US civilian, non-institutionalized population. The design has a relatively small number of primary sampling units, due to the need to provide a mobile clinical examination center in each sampled area. NHANES includes a health interview, a clinical exam, and a large collection of biochemical measurements. To capture survival information, the NHANES data are linked to the National Death Index. In 2008, the National Center for Health Statistics started producing public-use versions of the linked mortality data, using data perturbation techniques to preserve confidentiality[36]. We will analyse data from NHANES III, which was carried out from 1988 to 1994.

The NHANES III public-use data set has 49 pseudo-strata, with two primary sampling units per stratum, giving 49 design degrees of freedom. The coefficient of variation of the weights is 1.25. We analysed the data on race/ethnicity, age, rural/urban living and survival. NHANES deliberately oversamples minority subpopulations, so the design effect for the number or proportion of non-Hispanic whites (DMARETHN=1) is much larger than for non-Hispanic blacks (DMARETHN=2) or for Mexican-Americans (DMARETHN=3). There is also substantial sampling bias even for the Cox model coefficients: an unweighted analysis will underestimate the mortality disadvantage of both non-Hispanic blacks and Mexican-Americans. Table 1 shows p-values from tests for interaction between race/ethnicity and age, and for the combined main effect of race/ethnicity and its interaction with age, all adjusted for urban/rural place of living.

After fitting a model with 7 parameters there are 42 denominator degrees of freedom for the \( F \) tests, a small enough number to produce a visible difference between the \( \chi^2 \)-based and \( F \)-based tests. The Wald test is more highly significant than the LRT. Some of the difference may be due to anti-conservatism of the Wald test, but
most is because the LRT gives more weight than the Wald test to contrasts between non-Hispanic whites and the minority categories and less weight to contrasts within the minority categories.

4.2. A stratified case-cohort study. Wilms’ Tumor is a rare childhood tumor of the kidney. In the majority of cases it is effectively treatable, but several rare subtypes have a much worse prognosis. The histological classification of Wilms’ Tumor is difficult, and histologic classifications from the Wilms’ Tumor Study Group central lab are substantially more predictive than those from typical hospital labs. De-identified data sets on histology, stage, and survival in Wilms’ Tumor[37, 38] have been published by Norman Breslow with the permission of the National Wilms Tumor Study Group and have been used extensively in methodological research on measurement error and two-phase sampling. This analysis uses a version of the data set that is included in the R survey package, with records for 4028 cases of Wilms’ Tumor.

In this data set, an artificial case-cohort study has been created by selecting a random subsample of 668 for the reference subcohort and adding to it the 486 deaths or recurrences that were observed in people outside the subcohort. We artificially assume that the central-lab histology classification is available only for the case-cohort subsample, while stage, age, time to relapse, and the local institution’s histology classification are available for all participants. The analysis in our example tests for an interaction between disease stage (I-IV) and histology (favorable or unfavorable), in a model also including age at diagnosis. The test is performed using simple inverse-probability weights, and also using estimated (calibrated) weights that use information on the joint distribution of relapse, stage, and local-institution histology for the whole cohort[15, 39]. An unweighted analysis of these data would dramatically underestimate the effects of stage and histology, because of the outcome-dependent sampling. The coefficient of variation of the weights is 0.70 before calibration and 0.72 after calibration.

As there are 1154 individuals in 2 strata in the case-cohort subsample, the denominator degrees of freedom in a model with $p$ parameters would be $1154 - 2 - p$, large enough so that the difference between $F$ and $\chi^2$ null distributions will be negligible. We thus examine only the $\chi^2$ distribution for the Wald test and the linear combination of $\chi^2$ for the likelihood ratio test. Table 2 shows the test statistics for the calibrated and uncalibrated analyses. The interaction parameter estimates are correlated and have relatively large standard errors because there are relatively few deaths in the reference categories; the evidence for interaction is much clearer from the overall test statistics.

As seen in other analyses of these data, there is a substantial increase in the value of the test statistic with the calibrated analysis, reflecting more efficient use of information from the whole cohort. The $p$-values are smaller for the likelihood ratio test than for the Wald test, because the likelihood ratio test places more weight on the “Stage IV × unfavorable histology” coefficient, which turns out to be where the greatest departure from the null is seen.

5. Discussion

We have shown that the partial likelihood ratio test is useful in survival analysis even when data come from a complex sample. Its small-sample performance appears better than that of the Wald test, consistent with what has been found
Table 2. Estimates and tests for stage by histology interaction in Wilms’ Tumor

<table>
<thead>
<tr>
<th></th>
<th>uncalibrated</th>
<th>calibrated</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\hat{\beta}$</td>
<td>se</td>
</tr>
<tr>
<td>Stage II: unfavorable</td>
<td>-0.03</td>
<td>0.42</td>
</tr>
<tr>
<td>Stage III: unfavorable</td>
<td>-0.01</td>
<td>0.40</td>
</tr>
<tr>
<td>Stage IV: unfavorable</td>
<td>1.07</td>
<td>0.49</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Test statistic</th>
<th>$p$</th>
<th>Test statistic</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wald test</td>
<td>7.16</td>
<td>0.067</td>
<td>10.17</td>
<td>0.017</td>
</tr>
<tr>
<td>LR test</td>
<td>8.90</td>
<td>0.032</td>
<td>12.73</td>
<td>0.008</td>
</tr>
</tbody>
</table>

previously for loglinear models. Although the null sampling distribution is not a standard one, code is available to compute it and there are straightforward approximations that give good enough accuracy for most practical purposes. Using a linear combination of $F$ distributions with denominator degrees of freedom given by the sampling design appears to give modest but useful improvements in controlling the size of the test. Implementing these tests is not difficult, and we have provided an implementation in the survey package for R.
APPENDIX A. DETAILS OF ASYMPTOTIC EXPANSION

As in Section 2.1, we assume that our sequence of sampling designs is such that \( \sqrt{n} (\hat{\beta} - \beta_0) \overset{D}{\sim} N(0, \mathbf{V}_\beta) \) and \( \frac{1}{n} \mathcal{J}_W(\beta) \overset{D}{\rightarrow} \mathcal{I}(\beta) \) as \( n, N \rightarrow \infty \). We need a preliminary result on the relationship between \( \hat{\beta} \) and the restricted estimator \( \hat{\beta}_R \).

**Lemma.** If \( \gamma_0 = 0 \) then \( \hat{\beta} - \hat{\beta}_R = A \hat{\gamma} + o_p(n^{-1/2}) \) with \( A = \left( \begin{array}{c} \mathcal{I}_{\alpha\alpha}^{-1} \mathcal{I}_{\alpha\gamma} \\ \mathbf{I}_q \end{array} \right) \), where \( \mathbf{I}_q \) is the \( q \times q \) identity matrix.

**Proof.** We first expand \( U(\beta) = \frac{1}{n} \partial \ell_W / \partial \beta \) about \( \beta_0 \), noting that \( U(\hat{\beta}) = 0 \) and that \( -\partial U / \partial \beta^T = \frac{1}{n} \mathcal{J}_W(\beta) = \mathcal{I}(\beta) + o_p(1) \):

\[
0 = U(\hat{\beta}) = U(\beta_0) - \mathcal{I}(\beta_0) (\hat{\beta} - \beta_0) + o_p(n^{-1/2}).
\]

In particular, if \( U_1 = \frac{1}{n} \partial \ell_W / \partial \alpha \) is the first component of \( U \),

\[
U_1(\beta_0) = \mathcal{I}_{\alpha\alpha}(\hat{\alpha} - \alpha_0) + \mathcal{I}_{\alpha\gamma}(\hat{\gamma} - \gamma_0) + o_p(n^{-1/2})
\]

Similarly, expanding \( U_1(\hat{\beta}_R) \) about \( \beta_0 \) and assuming that \( \gamma_0 = 0 \):

\[
0 = U_1(\hat{\beta}_R) = U_1(\beta_0) - \mathcal{I}_{\alpha\alpha}(\hat{\alpha}_R - \alpha_0) + o_p(n^{-1/2}).
\]

Combining the two expressions for \( U_1(\beta_0) \) leads to \( \hat{\alpha} - \hat{\alpha}_R = \mathcal{I}_{\alpha\alpha}^{-1} \mathcal{I}_{\alpha\gamma} \hat{\gamma} + o_p(n^{-1/2}) \) when \( \gamma = 0 \). The lemma then follows immediately. \( \square \)

**Theorem.** If \( \gamma = 0 \), \( D_W = -2 \left[ \ell_W(\hat{\beta}_R) - \ell_W(\beta) \right] \sim \sum_{i=1}^q \lambda Z_i^2 \), where \( Z_1^2, \ldots, Z_q^2 \) are independent \( \chi_1^2 \) random variables and \( \lambda_1, \ldots, \lambda_q \) are the eigenvalues of \( (\mathcal{I}_{\gamma\gamma} - \mathcal{I}_{\alpha\gamma} \mathcal{I}_{\alpha\alpha}^{-1} \mathcal{I}_{\alpha\gamma}) \mathbf{V}_{\gamma\gamma} \).

**Proof.** Expand \( \ell_W(\hat{\beta}_R) \) about \( \beta \), noting that \( \partial \ell_W(\beta) / \partial \beta = 0 \) at \( \beta = \beta \):

\[
\ell_W(\hat{\beta}_R) = \ell_W(\beta) - \frac{1}{2} (\hat{\beta}_R - \hat{\beta}_R^T) \mathcal{J}_W(\beta) (\hat{\beta}_R - \hat{\beta}) + o_p(n \| (\hat{\beta}_R - \hat{\beta}) \|^2)
\]

Thus, using the results of the lemma,

\[
D_W = -2[\ell_W(\hat{\beta}_R) - \ell_W(\hat{\beta})] = n \hat{\gamma}^T A^T \mathcal{I} A \hat{\gamma} + o_p(1).
\]

Note that \( A^T \mathcal{I} A = \mathcal{I}_{\gamma\gamma} - \mathcal{I}_{\alpha\gamma} \mathcal{I}_{\alpha\alpha}^{-1} \mathcal{I}_{\alpha\gamma} \). The theorem then follows from standard results on quadratic forms of asymptotically normal random variables. \( \square \)

APPENDIX B. SIMULATION MODELS

Code in R for simulating a population from the superpopulation model with true effect of \( z_1 \) and no effect of \( z_2, z_3, z_4 \). Cluster sampling probability is proportional to the sum of \( tt \) in the cluster.

```R
z1<-rnorm(100000)
z2<-rnorm(100000)
z3<-z1+rnorm(100000)
z4<-sin(z1)+rnorm(100000)/2+z3/2
tt<-rexp(100000,rate=exp(z1/4))
time<-pmin(tt,1)
status<-tt<1
gg<-z1+tt+rnorm(100000)
cluster<-(rank(gg)-1) %% 100
```
Code in R for simulating a two-phase (case-cohort) design, with phase-one variables $x_1$, $x_2$, and phase-two variable $z$. There is a true effect of $x_1$ but not of $x_2$ or $z$. Sampling depends on case status and on strata defined by the signs of $x_1$ and $x_2$.

```r
library(MASS)
library(survey)
x <- mvrnorm(4000, mu = c(0, 0, 0), Sigma = cbind(c(1, .8, .5), c(0.8, 1, .8), c(0.5, .8, 1))))
t0 <- rexp(4000, rate = exp(x[, 1]/5))
status <- as.numeric(t0 < 0.05)
time <- pmin(t0, 0.25)
xstrats <- interaction(x[, 1] > 1, x[, 2] > 0)
n10 <- sum(xstrats == "TRUE.FALSE")
phase2 <- stratsample(xstrats, c(FALSE.FALSE = 100, TRUE.FALSE = n10, FALSE.TRUE = 100, TRUE.TRUE = 100))
phase2 <- unique(c(phase2, which(status == 1)))
in.phase2 <- (1:4000) %in% phase2
strats <- as.factor(ifelse(status == 1, "CASE", as.character(xstrats)))
df <- data.frame(x1 = x[, 1], x2 = x[, 2], z = x[, 3], strats = strats, in.phase2 = in.phase2, status = status, time = time)
```

**Appendix C. Computing linear combinations of $F$ statistics**

Let $A \sim \sum \lambda_i \chi^2$ and $B \sim \chi^2_m$. Then

\[
\frac{A}{B/m} > t \iff A > Bt/m \iff A - Bt/m > 0
\]

so the tail probability at $t$ for the linear combination of $F$ distributions is the tail probability at 0 for a linear combination of $p+1$ $\chi^2$ variables with multipliers $(\lambda_1, \lambda_2, \ldots, \lambda_p, -t/m)$ and degrees of freedom $(1, 1, \ldots, 1, m)$ respectively. The infinite-series computation[32] is not valid for negative multipliers, but the denominator adjustment can still be applied by integrating out the denominator numerically

\[
P\left(\frac{A}{B/m} > t\right) = \int_0^\infty P\left(\frac{A}{b/m} > t\right) f_B(b) \, db.
\]

**References**


Figure 1. Smoothed map of stroke death rates in US white males, adjusted to age 70[40, p73]
Figure 2. Quantile-quantile plot of saddlepoint approximation (x-axis) against other approximations to the p-value, for \( \lambda = (2.7, 1.0, 0.3) \). The y-axis shows Farebrother’s infinite series algorithm with a range of tolerances for error, and the Satterthwaite approximation.
Figure 3. Quantile–quantile plots of $-\log_{10}(p)$-values from simulated large survey with 32 PSUs. Points above the line indicate an anti-conservative test. The top row are the Wald-type tests, the bottom row are the partial likelihood ratio tests. The left column uses a $\chi^2$ or linear combination of $\chi^2$, the right column uses an $F$ or linear combination of $F$ with 28 denominator degrees of freedom.
Figure 4. Quantile–quantile plots of $-\log_{10}(p)$-values from simulated large survey with 60 PSUs. Points above the line indicate an anti-conservative test. The top row are the Wald-type test, the bottom row are the partial likelihood ratio tests. The left column uses a $\chi^2$ or linear combination of $\chi^2$, the right column uses an $F$ or linear combination of $F$ with 56 denominator degrees of freedom.
Figure 5. Quantile-quantile plots of $-\log_{10}(p)$-values from simulated two-phase design. Points above the line indicate an anti-conservative test. The left panel is the Wald test, the right panel is the partial likelihood ratio test.