Exact inference in the proportional hazard model: possibilities and limitations

Sven Ove Samuelsen
Department of Mathematics, University of Oslo,
P.O.Box 1053 Blindern, N-0316 Oslo 3, Norway,
email: osamuels@math.uio.no

December 11, 2001

Abstract
It is suggested that inference under the proportional hazard model can be carried out by programs for exact inference under the logistic regression model. Furthermore a different type of exact inference is developed under Type II censoring. Performance of logistic exact and exact inference is compared to large sample Wald, score and likelihood inference by coverage and power calculations. The logistic exact confidence intervals have coverage well above the nominal level in most computation. The exact inference under Type II censoring turn out to be less conservative but numerically very complex. Large sample methods works well with remarkably small data but score and likelihood ratio methods are preferable to inference by Wald statistics.

Key words: Cox regression; Survival analysis, Exact inference, Coverage, Power.

1 Introduction
Cox-regression (Cox (1972)) is without question the most often applied regression technique for survival data. Inference for the Cox estimator is almost exclusively based on asymptotic results (Andersen and Gill (1982)). The validity of these large sample properties have been evaluated in simulation studies, see for instance Peace and Flora (1978), Johnson et al. (1982) and Lee et al. (1983), and have been found acceptable with moderately large sample sizes, moderate amount of censoring and balanced covariate distributions.
However it frequently occurs that covariates have very skew distributions, for instance when only a small fraction of the subjects are exposed to a risk factor. It is also common that a large fraction of the subjects are censored. In for instance a large cohort study of the effect of a rare exposition on a rare disease the number of exposed cases may be quite small. One may then question the validity of inference based on large sample results.

Inference for logistic regression and conditional logistic regression is also mainly carried out by large sample methods even though the size of data often makes the validity of the large sample methods questionable. However for the logistic regression models there has been developed methods that do not rely on large sample results (Hirji et al. (1987)) and commercial software, for instance the program LogXact (Metha and Patel (1996)), is available to carry out this exact inference. The basis of the exact inference is the multiparameter exponential class representation of data from a logistic regression models.

Similar methods have not been derived for the proportional hazard model, although a number of authors (see Davidov & Zelen (1998) for references) starting with Cox (1972) have discussed the relation between the partial likelihood score test and permutation tests. Davidov & Zelen (1998) discuss the analogue of urn sampling and some extensions. In a different direction Broström & Nilsson (2000) view the partial likelihood as a likelihood for a series of binary experiments and derive an exact test for the proportionality assumption within this context. Inference suggested in this paper is related to their approach, but focus is on regression parameters.

The Cox partial-likelihood is on a similar form as the likelihood for conditional logistic regression. The initial idea for this paper was to apply this similarity to construct inference with good small sample properties under the proportional hazard model. It will be demonstrated that the LogXact program can produce such logistic exact inference on a cohort of Norwegian children where the outcome is cancer death and the primary covariate is low birth weight. Both outcome and exposure are rare and validation of large sample inference is prudent.

A problem with this approach however is that the partial likelihood does not, in general, have a direct probabilistic interpretation (Kalbfleisch & Prentice, 1980, pg. 77). Thus it does not follow, at least not directly, that attained levels of tests and confidence intervals are equal to nominal levels. The performance of the logistic exact inference thus requires further study.

There is actually one situation where the Cox-likelihood does has a direct probabilistic interpretation, namely Type II censoring when follow-up ends after a predetermined number of events (Kalbfleisch & Prentice, 1980, pg. 72). Then the Cox-likelihood corresponds to the distribution of the censored rank statistic. With one covariate this becomes a one-parameter distribution. A suggestion for obtaining exact inference is then to order the ranks according to the estimators of the regression parameter they give rise to and invert this distribution.
The paper presents results on coverage of confidence intervals for rate ratios and power of tests for no difference between two groups under different situations. For some very small situations coverage and power are calculated directly from the distribution of the ranks. For larger situation coverage and power have been obtained by simulation. Comparisons are made between large sample inference by Wald, score and likelihood ratio methods and small sample inference by logistic exact and when possible by the exact method based on the distribution of the ranks.

In the next section we review the Cox partial likelihood and large sample inference based on it. The relation between the partial likelihood and conditional logistic regression is described in Section 3 and it is pointed out how LogXact may be used to obtain small sample inference. It is also shown how exact inference can be made under Type II censoring and discussed to what extent this procedure may be generalized. In Section 4 the cohort of childhood cancer mortality is reanalyzed with logistic exact inference. Section 5 presents direct computations and simulations of coverage and power in some small data size situations. The paper ends with a short discussion.

2 Review of large sample inference for Cox-regression

Assume that survival times $T_i$ follow a proportional hazard model $\lambda_i(t) = \exp(\beta' z_i) \lambda_0(t)$ where $\lambda_0(t)$ is a baseline hazard, $\beta = (\beta_1, \ldots, \beta_p)'$ a $p$-dimensional regression parameter and $z_i = (z_{i1}, \ldots, z_{ip})'$ $p$-dimensional covariates. The $RR_k = \exp(\beta_k)$ have interpretations as relative risks or rate ratios. We observe censored survival times $x_i = \min(T_i, c_i)$ and indicators of events $d_i = I(x_i = T_i)$. Here $c_i$ are the censoring times. Let $\mathcal{R}(x) = \{j : x_j \geq x\}$ be the risk set at time $x$ and $d. = \sum d_i$ the total observed events. We will order indices so that the first $d_i$ correspond to the subjects with $d_i = 1$. The Cox estimator $\hat{\beta}$ is then obtained by maximizing the partial likelihood

$$L(\beta) = \prod_{i=1}^{d} \frac{\exp(\beta' z_i)}{\sum_{j \in \mathcal{R}(x_i)} \exp(\beta' z_j)}.$$

Let further $U(\beta) = \partial \log(L(\beta))/\partial \beta$ be the score function, $I(\beta) = -\partial U(\beta)/\partial \beta$ the information matrix and $\hat{\Sigma} = I(\hat{\beta})^{-1} = (\hat{\sigma}_{jk})$ the estimated covariance matrix of $\hat{\beta}$. Let also $se_k = (\hat{\sigma}_{kk})^{0.5}$ be the standard error of $\hat{\beta}_k$.

The usual variants of large sample inference are Wald, score and likelihood ratio tests and confidence intervals. The Wald confidence interval for $\beta_k$ with confidence level $1 - \alpha$ is given by

$$< \hat{\beta}_k - z_{\alpha/2} se_k, \hat{\beta}_k + z_{\alpha/2} se_k >$$
where \( z_{\alpha/2} \) is the \((1 - \alpha/2)\) 100\% percentile in the standard normal distribution. Similarly the Wald test of \( \beta_k = 0 \) is obtained by comparing \( \hat{\beta}_k / s\varepsilon_k \) by the standard normal distribution.

Let \( \beta^* \) be the value of \( \beta \) that maximizes \( L(\beta) \) when \( \beta_k = \beta_k^0 \), thus \( \beta^* \) depend on \( \beta_k^0 \). Let \( U_k^* \) be the k-th component of \( U(\beta) \) and \( I_{kk}^* \) be the k-th element of the diagonal of \( I(\beta) \) both evaluated at \( \beta = \beta^* \). The score test of the null hypothesis of \( \beta_k = \beta_k^0 \) is then given by comparing \( U_k^*/(I_{kk}^*)^{0.5} \) by a standard normal distribution and the score confidence interval with confidence level \((1 - \alpha)\) is determined by

\[
\{ \beta_k^0 : (U_k^*)^2 / I_{kk}^* < \gamma_\alpha \}
\]

where \( \gamma_\alpha \) is the \((1 - \alpha)\) 100 \% percentile of the \( \chi^2 \) distribution with one degree of freedom.

Similarly likelihood ratio tests of \( \beta_k = \beta_k^0 \) are obtained by comparing \( 2(\log(L(\hat{\beta})) - \log(L(\beta^*))) \) by the \( \chi^2 \) distribution with one degree of freedom and the \((1 - \alpha)\) level likelihood ratio confidence interval is given by

\[
\{ \beta_k^0 : 2(\log(L(\hat{\beta})) - \log(L(\beta^*))) < \gamma_\alpha \}.
\]

Usually it is more meaningful to present confidence intervals for the rate ratios \( RR_k \). These are obtained by taking anti-logs of the confidence intervals for \( \beta_k \).

Most programs for Cox-regression will only present the Wald confidence intervals. As discussed by Ternau & Grambsch (2000) it is however easy to obtain likelihood ratio intervals with programs that allow for ”offset”-terms in the model. Score intervals are also fairly easy to calculate.

Under certain data structures \( L(\beta) \) is maximized for \( \hat{\beta}_k = \pm \infty \). In such cases programs for Cox-regression will cease iteration at some large value of \( |\beta_k| \) and an typically even larger value of \( s\varepsilon_k \). In practical terms this means that the Wald interval degenerates to the real line. Ternau & Grambsch (2000) however claimed that likelihood ratio and score intervals may still be valid. In this paper that claim is investigated.

3 Suggestions towards exact inference in the proportional hazard model

3.1 Logistic exact inference

In this subsection the framework for conditional logistic regression is presented and it pointed out how exact inference can be carried out in that framework. The connection with Cox-regression is discussed. This leads to a suggestion for carrying out small sample inference under the proportional hazard model by a ”logistic exact” method. Notation for conditional logistic regression is kept similar to notation for Cox-regression.
Suppose we observe binary outcomes $Y_{ij}$ for $j = 1, \ldots, n_i$ subjects in $i = 1, \ldots, d$ groups. Associated with the outcomes $Y_{ij}$ are $p$-dimensional covariates $z_{ij}$. The subjects are independent and we assume a logistic regression model

$$P(Y_{ij} = 1) = \frac{\exp(u_i + \beta'z_{ij})}{1 + \exp(u_i + \beta'z_{ij})}.$$ 

Often sample size increase with the number of groups $d$, and each $Y_i = \sum_{j=1}^{n_i} Y_{ij}$ is small. Thus there are many nuisance parameters $u_i$. It is well known that full likelihood estimation may then lead to severe bias in the estimates of the interest parameter $\beta$. However it is a standard result that

$$p_i(y_{i1}, \ldots, y_{in_i}; \beta) = P(Y_{i1} = y_{i1}, \ldots, Y_{in_i} = y_{in_i} | \sum_{j=1}^{n_i} Y_{ij} = Y_i) = \frac{\exp(\beta' \sum_{j=1}^{n_i} z_{ij} y_{ij})}{\sum_s \exp(\beta' \sum_{j=1}^{n_i} z_{ij} y_{ij}^s)}$$ 

where $\sum_s$ means the sum over all permutations of $Y_i$ outcomes of $Y_{ij} = 1$ from $n_i$ possible and $y_{ij}^*$ is the values assigned to $Y_{ij}$ in each permutation, thus $\sum_{j=1}^{n_i} y_{ij}^* = Y_i$ for all permutations. In particular the $p_i(y_{i1}, \ldots, y_{in_i}; \beta)$ do not depend on the $u_i$. We obtain a conditional likelihood

$$L_{\text{cond}}(\beta) = \prod_{i=1}^{d} p_i(Y_{i1}, \ldots, Y_{in_i}; \beta)$$

with usual large sample likelihood properties.

The joint distribution $\prod_{i=1}^{d} p_i(y_{i1}, \ldots, y_{in_i}; \beta)$ of the $Y_{ij}$ conditional on $\sum_{j=1}^{n_i} Y_{ij} = Y_i$ can be written on form of a multi-parameter exponential family with sufficient statistics $T_k = \sum_{i=1}^{d} \sum_{j=1}^{n_i} z_{ijk} Y_{ij}$. Here $z_{ijk}$ is the $k$-th component of $z_{ij}$. From classical theory for exponential families (see for instance Lehmann, 1986) it then follows that the distribution of $T_k$ given $T_l = t_l$ for all $l \neq k$ only depend on the $k$-th component $\beta_k$ of $\beta$. A confidence interval for $\beta_k$ with level $1 - \alpha$ is thus given as $<\hat{\beta}_{kL}, \hat{\beta}_{kU}>$ determined by

$$P(T_k \geq t_k | T_l = t_l \text{ for all } l \neq k; \hat{\beta}_{kL}) = \alpha/2$$

and

$$P(T_k \leq t_k | T_l = t_l \text{ for all } l \neq k; \hat{\beta}_{kU}) = \alpha/2.$$ 

Here $t_k$ and the $t_l$ are the observed values of $T_k$ and the $T_l$.

This conditional distribution of $T_k$ is computationally very complex, but software, in particular the LogXact program (Metha and Patel (1996)), is available to obtain such confidence intervals and perform tests of $\beta_k = 0$. 

5
Suppose now that all $Y_i = 1$. In a survival data context this corresponds to one death at each event time. Then with $y_{ik} = 1$ we have that $p_k(y_{i1}, \ldots, y_{im}; \beta)$ simplifies to

$$\frac{\exp(\beta'z_{ik})}{\sum_{j=1}^{n} \exp(\beta'z_{ij})}.$$  

Thus $L_{\text{cond}}(\beta)$ is on the same form as a Cox partial likelihood. In particular we obtain $L(\beta)$ if $n_i$ equals the number of subjects in risk set $\mathcal{R}(x_i)$ and if there is a 1-1 correspondence between \{z_{i1}, \ldots, z_{im}\} and \{z_j : j \in \mathcal{R}(x_i)\}.

A motivation behind the Cox-regression technique consists in looking at probabilities of who failed at $x_i$ among the members of $\mathcal{R}(x_i)$ at different times $x_i$. One may thus think of the construction as a series of binary experiments (Broström & Nilsson, 2000) at each event time.

Given this perspective a promising idea becomes to interpret $L(\beta)$ as the likelihood stemming from a “distribution”

$$f(d_1, \ldots, d_n; \beta) = \prod_{i=1}^{n} \left[ \frac{\exp(\beta'z_i)}{\sum_{j \in \mathcal{R}(x_i)} \exp(\beta'z_j)} \right]^{d_i}$$

which is formally on the form as the conditional distribution of the $Y_{ij}$ and then carry out the “exact” inference under this interpretation based on “sufficient” statistics $T_k = \sum_{i=1}^{n} z_{ik}d_i$. This kind of inference is referred to as “logistic exact”. The main advantages of the approach are that multivariate proportional hazard models can be handled and that software is available.

Unfortunately the partial likelihood $L(\beta)$ does not allow such a direct probabilistic interpretation (Kalbfleisch & Prentice, 1980). One may however hope that the analogy is close enough to make the method useful. It is a main purpose of the paper to investigate this.

Application of LogXact is possible after calculating at each event time $x_i$ the number at risk with each covariate value $z_i$ and determining the covariate value of the case at $x_i$. As an example consider Table 1 summarizing an imaginary data set with two binary covariates $z_1$ and $z_2$ and with $d_i$ = 100 deaths. Initially there were 600 unexposed ($z_1 = z_2 = 0$) subjects, 200 subjects exposed to $z_1$, but not $z_2$ and both 100 subjects exposed to $z_2$ alone and exposed to both factors. The first event occurred for a subject with $z_1 = 0$ and $z_2 = 1$. For later events the number at risk will decrease due to censoring and events, but could increase with left truncated data or time-dependent covariates. In LogXact one would enter the first column as stratum, the fourth as the group size (weight) variable and the fifth as response. One may fit models with either and both covariates included.
Table 1 Structure of input file to LogXact of survival data set with two binary covariates $z_1$ and $z_2$ and $d. = 100$ event times (artificial example).

<table>
<thead>
<tr>
<th>Event no.</th>
<th>$z_1$</th>
<th>$z_2$</th>
<th>No. at risk</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>600</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>200</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>570</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>1</td>
<td>193</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
<td>93</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>96</td>
<td>0</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>100</td>
<td>0</td>
<td>0</td>
<td>182</td>
<td>0</td>
</tr>
<tr>
<td>100</td>
<td>0</td>
<td>1</td>
<td>45</td>
<td>0</td>
</tr>
<tr>
<td>100</td>
<td>1</td>
<td>0</td>
<td>11</td>
<td>0</td>
</tr>
<tr>
<td>100</td>
<td>1</td>
<td>1</td>
<td>23</td>
<td>1</td>
</tr>
</tbody>
</table>

3.2 Exact inference under type II censoring

Type II censoring means that observation is continued until a predetermined number of events $d.$ have been observed. Failure time data without censoring is of course a special case with $d. = n$. In this section an exact method valid under Type II censoring with one covariate is presented. This method is computationally so complex that it may be of little practical use. The reason for introducing it here is mainly to point out that there is a special case where an exact method is possible and that this method differ from the logistic exact method of Section 3.1.

With the previous notation $x_1,\ldots, x_d$ are the $d.$ smallest survival times. Then under assumption of a proportional hazard model we have the probability that these $d.$ subjects are the first to fail and fail in this particular order equals (Kalbfleisch & Prentice, 1980)

$$p(R = r; \beta) = \prod_{i=1}^{d} \frac{\exp(\beta z_i)}{\sum_{j \in R(z_i)} \exp(\beta z_j)},$$

where $R$ denotes the random censored rank statistic and $r$ an outcome of $R$. In particular the partial likelihood $L(\beta)$ becomes the marginal likelihood of the ranks. Let $\hat{\beta}(r)$ be the Cox-estimator with ranks $r$. 

7
Now with only one covariate in the model we have a one-parameter distribution. The ranks are naturally represented as a \( n \)-dimensional object and there is no unique way of ordering them. Here however we suggest ordering the ranks according to the estimator they give rise to, thus if \( r \) and \( r' \) are two censored rank statistics we say that \( r \) is smaller than \( r' \) if \( \tilde{\beta}(r) < \tilde{\beta}(r') \). It is possible that \( \tilde{\beta}(r) = \tilde{\beta}(r') \) for \( r \neq r' \), for instance with binary covariates, but then one should consider different sequences of covariates rather than different sequences of subjects. In principle it may be possible to have \( \tilde{\beta}(r) = \tilde{\beta}(r') \) even for non-equivalent \( r \) and \( r' \), however this eventuality has not turned up in calculations so far.

With this ordering we may speak of a cumulative distribution for \( R \), \( P(R \leq r; \beta) = \sum_{r' \leq r} p(r'; \beta) \) and with an observation \( R = r \) it is possible to find a confidence interval with exact confidence level \( 1 - \alpha \) as \( \beta^*_L, \beta^*_U > \) where

\[
P(R \geq r ; \beta^*_L) = \alpha/2
\]

and

\[
P(R \leq r ; \beta^*_U) = \alpha/2.
\]

To calculate such intervals it is necessary to determine all rank statistics \( r \) (or rather all different sequences of covariates) and the corresponding estimates \( \tilde{\beta}(r) \). This will require substantial computational power and in practice only be feasible for rather small data sets. Performance of the exact method could only be studied under the very small data situations in Section 5.1.

Another problem with the exact method is that it is derived only under Type II censoring and only with one covariate. However \( p(R = r; \beta) \) is the distribution of the censored rank statistic also under progressive Type II censoring (Kalbfleisch & Prentice, 1980). It is also possible to extend the method to fixed censoring at a common censoring time \( c \). The inference then is based on conditioning on the \( d \) events in \( [0, c] \) and conditionally we have Type II censoring. The unconditional confidence level is then at least \( 1 - \alpha \). To extend the method to several categorical covariates note that a proportional hazard model \( \lambda_{z_1, z_2}(t) = \exp(\beta z_1 + \gamma z_2) \lambda_0(t) \) is a special case of a stratified proportional hazard model \( \lambda_{z_1, z_2}(t) = \exp(\beta z_1) \lambda_{z_2}(t) \). Similar to a suggestion by Davidov & Zelen (1998) we may perform the exact inference under the stratified model. With known censoring times taking on a finite set of values it is possible to do the inference stratified on censoring times as well. This stratification technique may lead to substantially loss in precision with many small strata, but may work well otherwise.

Thus the major limitation of the method is the numerical complexity involved in enumerating all orderings of covariates. Monte Carlo techniques that require enumeration of only a sample of the orderings may be the best promise for application of this method.
4 A data example

A recent paper (Samuelsen et al. (1998)) describe effects of low birth weight on mortality in childhood. The study was based on the Norwegian Birth Registry linked to the National Cause of Death Registry. Although the data set was very large, about 1.25 million subjects were included, both the outcome (death in childhood) and the exposure (low birthweight) were rare. The large number of subjects thus does not appropriately describe the size of the data. A better measure of the size of such a data set would be the smallest number in the two-by-two table of outcome by exposure.

For cancer mortality a somewhat surprising protective effect was observed. The rate ratio of cancer mortality for children with birthweight below 2500 grams (g) was only 0.27 compared with children with birthweight above 2500 g. This number was however based on only 6 low birth weight cancer deaths out of a total number of 687 cancer deaths. A researcher should in this case be wary of relying on large sample inference. Even worse when considering mortality in separate age groups there were 4 low birth weight cancer deaths in age 1-5 years and only one low birth weight cancer death in each of the age intervals 6-10 and 11-15 years.

The data were analyzed by Poisson regression. In Table 2 cancer mortality is reanalyzed by Cox-regression. Confidence intervals for the rate ratios are given, in addition to the usual Wald interval also score and likelihood ratio intervals are computed. Furthermore data were processed so that logistic exact confidence intervals could be obtained from LogXact.

Both crude and adjusted rate-ratios are given for the effect of low birth weight. In the previous paper this effect was adjusted for a number of confounders. In LogXact however it proved possible to obtain adjusted confidence interval for only a few covariates. It was chosen to adjust only for gestational age coded as categories pregnancy of 37 or more weeks, less than 37 weeks and unknown length of pregnancy. Gestational age was the only confounder that influenced the effect of low birth weight to any noticeable degree.

For the age intervals 1-15 years and also 1-5 years the different intervals are in fairly good agreement both for the crude and adjusted analyses. One does however observe the tendency that the upper limit of LR intervals are lower than the other upper limits. Lower limits of the LR and logistic exact intervals are on the other hand in better agreement than the Wald and score intervals. A similar tendency is also found in age intervals 6-10 and 11-15 years, but differences are larger, in particular the upper limits of the LR intervals are lower than the other upper limits. The different types of tests of no effect of low birth weight at a 5% percent level thus generally gave the same result. The two only exception were the unadjusted Wald test for age interval 6-10 that did not reject and the adjusted likelihood ratio test for age interval 11-15 that rejected the null-hypothesis of no difference between the groups with a 5 % level.

In Samuelsen et al. (1998) exact intervals for the crude analyses were presented.
These intervals were based on the assumption that the number of events in the two groups were Poisson distributed. This of course leads to a binomial distribution for the number of low birth weight deaths conditional on the total number of deaths and the intervals were obtained by the standard inversion technique. It is interesting to note that the crude logistic exact interval reported above are almost identical to the intervals in Samuelsen et al. (1998).

However the usual justification for Poisson regression on survival data (Holford, 1980) relies on the likelihood only being proportional to a Poisson likelihood. The data does not stem from a Poisson distribution which is a requirement for such exact inference. It is generally possible to construct exact inference under log-linear Poisson models since the Poisson distributions constitute an exponential family. Application of such inference to survival data will then not be strictly correct.

**Table 2** Rate ratios (RR) of cancer mortality between low birth weight children (≤ 2500 g) and normal birth weight children (> 2500 g) with Wald, score, likelihood ratio (LR) and logistic exact ("Exact") 95% confidence intervals in age intervals 1-15, 1-5, 6-10 and 11-15 years. (a) Crude estimates with low birth weight as only covariate. (b) Estimates adjusted for normal, short and unknown gestational age.

<table>
<thead>
<tr>
<th>Covariate</th>
<th>RR</th>
<th>Wald interval</th>
<th>Score interval</th>
<th>LR interval</th>
<th>&quot;Exact&quot; interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>a) Crude</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age 1-15</td>
<td>0.275</td>
<td>0.123 - 0.614</td>
<td>0.126 - 0.602</td>
<td>0.109 - 0.560</td>
<td>0.101 - 0.603</td>
</tr>
<tr>
<td>Age 1-5</td>
<td>0.351</td>
<td>0.131 - 0.941</td>
<td>0.136 - 0.908</td>
<td>0.109 - 0.822</td>
<td>0.095 - 0.907</td>
</tr>
<tr>
<td>Age 6-10</td>
<td>0.160</td>
<td>0.022 - 1.142</td>
<td>0.028 - 0.910</td>
<td>0.009 - 0.711</td>
<td>0.004 - 0.902</td>
</tr>
<tr>
<td>Age 11-15</td>
<td>0.241</td>
<td>0.034 - 1.723</td>
<td>0.042 - 1.372</td>
<td>0.014 - 1.075</td>
<td>0.006 - 1.367</td>
</tr>
<tr>
<td>b) Adjusted</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age 1-15</td>
<td>0.269</td>
<td>0.116 - 0.614</td>
<td>0.117 - 0.645</td>
<td>0.104 - 0.563</td>
<td>0.095 - 0.609</td>
</tr>
<tr>
<td>Age 1-5</td>
<td>0.511</td>
<td>0.184 - 1.417</td>
<td>0.188 - 1.486</td>
<td>0.164 - 1.245</td>
<td>0.134 - 1.417</td>
</tr>
<tr>
<td>Age 6-10</td>
<td>0.103</td>
<td>0.014 - 0.761</td>
<td>0.018 - 0.728</td>
<td>0.006 - 0.485</td>
<td>0.003 - 0.629</td>
</tr>
<tr>
<td>Age 11-15</td>
<td>0.193</td>
<td>0.026 - 1.462</td>
<td>0.032 - 1.519</td>
<td>0.011 - 0.943</td>
<td>0.005 - 1.233</td>
</tr>
</tbody>
</table>
5 Comparisons in some small data situations

5.1 Exact computations for very small data

This section presents calculations of coverage of confidence intervals and power of tests of difference between two groups of exposed and non-exposed subjects. To perform these calculations data sizes need to be very small. Censoring is not present, thus the Cox-partial likelihood has the interpretation as the probability of the observed ranks and the exact confidence intervals of Section 3.2 can be obtained.

Let \( n \) be the total sample size, \( n_E \) the number of exposed subjects. Then there are \( M = \binom{n}{n_E} \) data different data configuration (each of which correspond to \( n_E!(n - n_E)! \) orderings of subjects) resulting in \( M \) different estimates and \( M \) different confidence intervals of each type.

Let now \( \text{CI}_j, j = 1, \ldots, M \) be the \( M \) different confidence intervals of a specific type, say Wald intervals. The coverage probability at a parameter value \( \beta \) is then given as

\[
C(\beta) = \sum_{j=1}^{M} I(\beta \in \text{CI}_j)q_j(\beta)
\]

where \( q_j(\beta) = P(R = r; \beta)n_E!(n - n_E)! \) is the probability of data configuration \( j \) with the parameter \( \beta \) and \( r \) is a rank of subjects corresponding to data configuration \( j \). This way we may produce figures similar to for instance Vollset (1993) of the coverage as functions of \( \beta \).

Figure 1 gives all the 95\% intervals (Figure 1(f)) and coverage functions for \( n = 4 \) and \( n_E = 2 \) which is perhaps the smallest situation Cox-regression may even be contemplated. There are then \( M = 6 = \binom{4}{2} \) different data-configurations resulting in maximum partial likelihood estimates of \( \beta \) of \( \pm \infty, \pm 0.941 \) and \( \pm 0.481 \). With \( \hat{\beta} = \pm \infty \) the intervals end or start in \( \pm \infty \). The Wald interval degenerates to the real line, but the other intervals have finite upper or lower limits.

The coverage curves \( C(\beta) \) are discontinuous at the endpoints of the confidence intervals. Note that since all intervals in Figure 1(f) contain \( \beta = 0 \) we have \( C(0) = 1 \).

Both the exact and the logistic exact intervals have coverage of at least 97.5\%, but the logistic exact interval is even more conservative than the exact. Both the Wald and the LR intervals generally have coverage of 95\% or more, but the minimum is well below 95\%, for the LR interval as low as 85\% . The score interval coverage is often well below 95\%, but averaging over a range of \( \beta \) values from -6 to +6 gives an average coverage of 94.8 \% closer to the nominal level than any of the other intervals.

A coverage curve does not give a full picture of performance of an inference procedures. A method may well have satisfactory coverage and at the same time low power, examples can be found in this paper particularly for Wald intervals. As an addition to
Figure 1: Coverage and confidence intervals with \( n = 4 \) and \( n_F = 2 \). Coverage of (a) Wald-interval, (b) score interval, (c) likelihood ratio interval, (d) logistic exact interval and (e) exact interval for \( \beta \). Averaged coverage is over interval \( \beta \in [-6, 6] \). Figure (f) shows the different intervals for the 6 data-configurations.
the coverage curves one may then present power curves for the null-hypothesis of no
difference between exposed and non-exposed subjects given as

$$\text{Power}(\beta) = \sum_{j=1}^{M} I(\beta = 0 \notin \text{CI}_j) q_j(\beta).$$

Figure 2 depicts coverage curves and power of 5% level tests of the null-hypothesis
$$\beta = 0$$ for $$n = 8$$ and $$n_E = 4$$. The power curves were not given in Figure 1 since
for $$n = 4$$ since all confidence intervals contain $$\beta = 0$$, thus no tests will reject the
null-hypothesis and the power equals zero for all values of $$\beta$$. The same phenomena
occurs for the Wald test for $$n = 8$$. It also turns out that the power of the score
and likelihood ratio tests were identical. This happens since score and likelihood ratio
confidence intervals for each of the \( \binom{8}{4} = 70 \) data configurations either contain or do
not contain $$\beta = 0$$. Of the same reason the logistic exact and exact tests have identical
power.

Similarly to Figure 1 both coverage of the logistic exact and the exact confidence
intervals are well above the nominal level of 95%, but the exact confidence interval is
clearly less conservative than the logistic exact. The Wald interval is also generally con-
servative, but minimum coverage is well below the nominal level. The conservatism of
the Wald intervals is to a degree accounted for by degenerated intervals when $$\hat{\beta} = \pm \infty$$. However a closer inspection of the confidence intervals reveals that for configurations
that give very low values of $$\hat{\beta}$$ the upper Wald interval limit is closer to the cor-
responding exact and logistic exact limits than to the likelihood ratio and the score limits. Of
course the same holds for lower limits with very large $$\hat{\beta}$$. This phenomena contributes to
the conservatism of the Wald interval. The likelihood ratio interval has a conservative
average coverage of 96.3% over the interval $$[-6,6]$$, however over the interval $$[-2,2]$$ the
average is only 93.1% and minimum coverage is as low as 82.2% for $$\beta = \pm 1.15$$ which
are the endpoints of the likelihood ratio interval when $$\hat{\beta} = \pm \infty$$. The score interval has
an average coverage of 94.7% close to the nominal level, however in the interval $$\pm 0.69$$
corresponding to relative risks between 0.5 and 2 it is clearly below 95%.

Figure 3 depicts coverage and power curves when $$n = 12$$ and $$n_E = 6$$. The shapes
of the coverage curves are similar to Figure 2. There is a tendency of convergence to
the nominal level for all types of intervals although the average over the interval $$[-6,6]$$
does not change much. In particular the exact interval coverage lies between 95.0% and
95.5% for $$-0.95 < \beta < 0.95$$. The logistic exact interval has a minimum coverage
of 96.0% and a maximum coverage of 97.7% in the same interval. The likelihood ratio
interval has coverage between 92.6% and 94.9% for $$-1.22 < \beta < 1.22$$, but close to
$$\beta = \pm 2$$ it fall well below 85%. The score and Wald intervals have coverage in rough
agreement with the nominal level for all but very extreme $$\beta$$ although the Wald coverage
in general is higher than 95%.

The power of the tests are naturally increased compared with a sample size of $$n = 8$$
and no test now has zero power. The likelihood ratio and the score test have the highest
Figure 2: Coverage and power curves with $n = 8$ and $n_{E} = 4$. Coverage of (a) Wald-interval, (b) score interval, (c) likelihood ratio interval, (d) logistic exact interval and (e) exact interval for $\beta$. Averaged coverage is over interval $\beta \in [-6, 6]$. Figure (f) shows the power of tests of no difference between exposed and not exposed groups for $\beta \in [-3, 3]$. 
Figure 3: Coverage and power curves with $n = 12$ and $n_F = 6$. Coverage of (a) Wald-interval, (b) score interval, (c) likelihood ratio interval, (d) logistic exact interval and (e) exact interval for $\beta$. Averaged coverage is over interval $\beta \in [-6, 6]$. Figure (f) shows the power of tests of no difference between exposed and not exposed groups for $\beta \in [-3, 3]$. 

15
power, but the exact test fares not much worse. The power of the logistic exact test
is however somewhat lower. The power of the Wald tests tends to zero for large $|\beta|$. 
One might think this is due to degenerate intervals for $\hat{\beta} = \pm \infty$ only. However it
turns out that of the $M = 924$ Wald intervals as many as 878 contain $\beta = 0$. The
corresponding number for the score intervals is 93, for the likelihood ratio intervals 96,
for the exact intervals 104 and for the logistic exact intervals 114. Thus there seem
to be some inherent conservatism of Wald intervals and tests with very small sample
sizes.

Figures comparable to Figures 1, 2 and 3 were produced for $n = 6$ and $n = 10$. These
gave results in between what has been shown. Exact computation of coverage
for larger sample sizes was however not feasible with available computer power. In the
next subsection results for larger sample sizes are addressed by means of simulation.

5.2 Simulations

The first set of simulations continues the situations in Section 4 where there was no
censoring and half the subjects were either exposed or not exposed. Indeed the first
row in Figure 5.1 were $n = 12$ is simply repeating the results of Figure 3 but with
simulated data. This figure is only included to give an impression of the variation due
to simulation. Simulations are in all cases repeated 5000 times giving a standard error
of 0.3% for a coverage of 95%. In all simulations coverage and power are presented for
the Wald, score, likelihood ratio and logistic exact procedure. The exact method of
Section 3.2 could have been computed in Figure 4a), but for larger data situations it
was not feasible to obtain all enumerations. In contrasts with Section 5.1 the results
are presented for $-3 \leq \beta \leq 3$ corresponding to rate ratios between 0.05 and 20.

The most marked change when sample size is increased is that the $\beta$ values where
the likelihood ratio coverage drops to about 85% moves away from zero. As before
these $\beta$ values are the endpoints of the confidence intervals when $\hat{\beta} = \pm \infty$. For $n = 40$
the endpoint are below -3 and above +3. The Wald coverage is above or close to 95%,
but the power of the Wald test is lower than the other tests for large $|\beta|$. The score
intervals also have coverage close to 95%, but somewhat lower than Wald intervals.
The power of the score tests are close to the power of the likelihood ratio tests.

The logistic exact interval coverage is consistently above 95% in contrast with the
large sample intervals. Increasing the sample size make the coverage of the logistic
exact interval move only slowly toward the nominal level. This conservativeness is
reflected in somewhat lower power than that of the score and likelihood ratio tests.
Figure 4: Simulated coverage and power with \( n = 12, 20, 40, n_E = n/2 \) and no censoring. Results from 5000 simulations for each \( n \) and \( \beta = -3.0, -2.9, \ldots, 3.0 \). Figure (a) gives coverage and (b) power with \( n = 6 \), figure (c) coverage and (d) power with \( n = 20 \) and figure (e) coverage and (f) power with \( n = 40 \).
Figure 5: Simulated coverage and power with 90% fixed censoring, \( n = 100 \text{ and } n_E = 50 \) (figures (a) and (b) respectively) and \( n = 250 \text{ and } n_E = 50 \) (figures (c) and (d) respectively). Results from 5000 simulations for each \( n \) and \( \beta = -3.0, -2.9, \ldots, 3.0 \).

In the second set of simulations presented in Figure 5 the survival times are censored at a fixed common time. The proportion of uncensored observations equals 10%. The covariate \( z \) is binary with either 50% or 20% of the \( z = 1 \). With 50% of subjects exposed the sample size was \( n = 100 \) and with 20% exposed it was increased to 250. Thus with \( \beta = 0 \) there will in both situations be 5 expected uncensored and exposed subjects. The simulations are repeated 5000 times for each of \( \beta = -3.0, -2.9, \ldots, 3.0 \).

The Wald intervals are also in this situation conservative and the corresponding power may be very low. In particular with the skew distributed covariate (20% exposed) the Wald test is practically useless for negative \( \beta \). This is in clear contrast with the other tests. The likelihood ratio intervals again attain the minimum coverage and the corresponding test the maximum power. The score coverage is quite close to the nominal level, but with skew covariate the power is not much better than that of the logistic exact test.
Figure 6: Simulated coverage and power with $n = 10000$, $n_E = 310$, fixed censoring and expected number of uncensored subjects equal to $d = 200$ figure (a) coverage and (b) power and $d = 700$ figure (c) coverage and (d) power. Results from 5000 simulations for each $d$ and $\beta = -3.0, -2.9, \ldots, 3.0$.

The logistic exact intervals are also in this situation very conservative. Taking this into account the power of the logistic exact tests are however not all that low and it is in general much better than the power of the Wald test.

The third set of simulations mimics situations in the data example of Section 4. Of course it was not possible to do extensive simulations from a set of 1.25 million subject so total sample size is set to $n = 10000$. The proportion exposed is 3.1% close to the proportion of low birth weight children. In Figure 6 (a) and (b) the expected number of deaths are 200 close to the number of deaths in the age interval 6-10 years. In Figures 6 (c) and (d) the expected number of deaths are 700 close to the number of deaths in the age interval 1-15 years. Censoring is however at a common fixed time contrary to the real data.
The results from these simulations are in fairly good agreement with what has been presented before. Coverage curves are not symmetric around $\beta = 0$ because the covariate distribution is skew. The logistic exact method has too high coverage especially when the expected number of deaths equals 200 and when $\beta$ is small, but with 700 events and $\beta = 3$ it actually drops to 94.8%. A new simulation in this situation repeated 25000 times was carried out with a resulting coverage of 95.2%. The power is however close to the power of the score test and not all that much lower than the likelihood ratio test. The likelihood ratio interval again attains the minimum of all coverage values. The Wald and score coverage curves have peculiar oscillations for small values of $\beta$ when expected number of deaths equals 200. On average the Wald method is somewhat conservative whereas the score method is close to the nominal level. However the power of the Wald test renders it useless with 200 events and $\beta$ negative.

Another perspective on small sample performance of the Cox-estimator is bias. Usually bias is assessed by taking the average of estimated coefficients minus the true value. In simulations of small samples however the $\hat{\beta}$ frequently take on infinite values. A standard fix for this is comparing the average of the finite estimates $\overline{\beta}$ with the true value.

By this comparison $\hat{\beta}$ is often severely biased in the present simulations. A closer inspection shows however that $\overline{\beta} + \log(1 - r_-)$ is in much better agreement with $\beta$ when $r_-$ is the proportion of simulations with $\hat{\beta} = -\infty$ as long as the proportion $r_+$ with $\hat{\beta} = +\infty$ is small. This suggests that bias of $\overline{RR} = \exp(\overline{\beta})$ may be small when $\overline{RR} = 0$ are not ignored. Indeed the possible bias of $\overline{RR}$ is more relevant than bias of $\hat{\beta}$. Unfortunately $\overline{RR}$ may also be infinite and taking the averages of rate ratios are therefore neither a useful way of discussing bias. Averages of $\overline{RR}$ are not presented even when they could be calculated as only $\overline{\beta}$'s were stored in the simulations.

Figure 7 presents $\overline{\beta}$ and $\overline{\beta} + \log(1 - r_-) - \log(1 - r_+)$ as functions of $\beta$. The modification $-\log(1 - r_+)$ takes care of $\overline{\beta} = \infty$. In most simulations either $r_- > 0$ or $r_+ > 0$, but when both were positive they were generally small and so the modification did not differ much from $\overline{\beta}$. The modified average is quite close to $\beta$ suggesting that bias of $\overline{RR}$ is small.

The modified average is presented only to show that bias of the Cox-estimator may be very small. It is not intended as a bias correction method. With a finite $\hat{\beta}$ there is no need to adjust and with $\hat{\beta} = \pm \infty$ it does not work.

It is however suspected that researchers may suppress results when $\hat{\beta} = \pm \infty$ and Figure 7 shows that such a practice may lead to misinterpretation of data. It is of course awkward to put relative risks of zero or infinity in a table, but one could argue that confidence intervals only with finite upper (or lower) limits make sense.
Figure 7: Average of finite $\hat{\beta}$ an modified average for simulation with (a) with $n = 12$ and $n_E = 6$ and no censoring, (b) $n = 20$ and $n_E = 10$ and no censoring, (c) $n = 40$ and $n_E = 20$ and no censoring, (d) $n = 100$ and $n_E = 50$ and 90% censoring, (e) $n = 250$ and $n_E = 50$ and 90% censoring and (f) $n = 10000$ and $n_E = 310$ and 98% censoring. Results from 5000 simulations for each $n$ and $\beta = -3.0, -2.9, \ldots, 3.0$. 
6 Discussion

This paper has investigated adaptations of exact inference methods for logistic regression to survival data under the proportional hazard model. It proves possible to perform this type of inference with standard software for exact logistic regression with a few confounders. Simulations and direct calculations of coverage of such confidence intervals have in all situations had at least the nominal level. Whether this holds generally is however an open question.

Actually the coverage of exact logistic intervals are often considerably higher than the nominal level and from this perspective asymptotical methods may be preferable. However the power of the exact logistic tests were in some instances not all that far from the power of the score and likelihood tests and in some instances much better than the power of the Wald test. It may be good advise to also carry out the logistic exact inference as an extra precaution when there is doubt about the validity of large sample methods.

Recently Blaker (2000) and Blaker & Spjøtvoll (2000) have developed a method that improves on the usual inversion technique while retaining an assured confidence level. Application of this technique in exact inference for logistic regression should improve performance and at the same time not increase numerical complexity. Adaptation to proportional hazard models will then follow and could make the logistic exact confidence interval more appealing.

Somewhat surprisingly confidence intervals and tests from Wald type inference were in several examples quite conservative. In particular power of the Wald test in some examples turned out to be very low compared to all other methods. Partly this is due to degenerate tests and confidence intervals with $\hat{\beta} = \pm \infty$, but in Section 5.1 it was found that a much larger proportion of Wald intervals contained $\beta = 0$, thus it appears that the degenerate intervals is not the full explanation.

Perhaps likelihood ratio is the most intuitively appealing of the large sample methods. However the likelihood ratio coverage also showed the most extreme departure from nominal level in several examples. This appears partly to be due to too low (high) interval limits when $\hat{\beta} = -\infty (+\infty)$, but may be a more general problem. In particular in the data example it was seen that the upper limit of likelihood ratio is considerably smaller than the other upper limits even though $\hat{\beta} > -\infty$. However the likelihood ratio test generally showed the highest power.

Thus of the large sample based methods the score method seems to be the best choice. It generally had coverage not far from the nominal level, although there have been shown examples to the contrary, and the power of the score tests were generally good.

A discussion of performance of large sample methods will easily focus on shortcomings, as the above remarks do. Such a focus does not give fair credit to these methods. Indeed the performance of the large sample methods can be considered remarkably
good. It seems that advise that sample size should be \( n = 50 \) with no censoring and balanced covariate distributions before reliance can be put on the large sample inference is more cautious than necessary. On the other hand when faced with data that in some sense may be considered small sample it seems to be good advise to check the results with at least all three large sample method.

A limitation with this study is that coverage and power have only been evaluated with one covariate. Similar multivariate evaluations would require source code for logistic exact inference and could in any case only be carried out in very limited situations due to numerical complexity. Still it would be interesting to study whether the good coverage properties found here are valid in more general situations.

References


