FITTING COX'S PROPORTIONAL HAZARDS MODELS FROM SURVEY DATA

David A. Binder, Statistics Canada
11-F R.H Coats Building, Ottawa, Canada K1A 0T6

KEY WORDS: Specification error, Variance estimation, Complex surveys design, Taylor linearization

1. Proportional Hazards Models

The Cox proportional hazards model assumes that lifetime data are independently distributed with hazard function given by

$$h\{(t|\boldsymbol{x}(t)\} = h_0(t)\exp\{\boldsymbol{x}(t)'\boldsymbol{\beta}\},$$

where x(t) is a vector of (possibly time dependent) covariates, assumed to be non-stochastic, and $h_0(t)$ is the hazard function at x(t) = 0. This model has been very popular in the fitting of failure data, both in biometric and in reliability applications. One of the reasons for its popularity is that the unknown parameter, ß, can be estimated without putting a parametric structure on h_0. This is due to the fact that the conditional distribution of the lifetimes, given the failure times of the sample, does not depend on h_0 explicitly. Therefore, a partial likelihood analysis is possible.

However, this model is certainly not assumption-free. The true model can differ from the classical proportional hazards model in a number of ways. Some of these are:

- non-linearity of the exponential argument,
- missing variables,
- hazard functions not proportional,
- observations not independent.

As well, when the data are right-censored due to withdrawal from the study, it is usually assumed that the mechanism generating the withdrawal is unrelated to the censored portion of the hazard function.

Many of the consequences of misspecifying the model in the case of independent observations have been dealt with in the literature; see Lin and Wei (1989) for some recent references. Of particular note are Struthers and Kalbfleisch (1986), where the asymptotic biases under a wide class of models are explicitly derived, and Lin and Wei (1989) who propose a method for estimating the covariance matrix of the estimated parameters, when the model is misspecified.

We consider here a slightly different setting. Suppose the sample has been drawn from a population using a complex design, such as is commonplace for large-scale population-based surveys. For example, consider a large-scale health survey, with a stratified multi-stage design. Some of the design parameters may be related to the true hazard function, but are not explicitly part of the model being fitted. This could include, for example, environmental factors which are related to the geographic regions being controlled in the sample design, but which are not explicitly measured. As well, many of the important covariates may not be included in the survey. For example, in a study on cancer mortality, not all the risk factors are even known. Linearity of some of the covariates could also be violated for many of the quantitative factors, such as number of cigarettes smoked per day. Independence may be violated. For example, for family data, risk of heart disease may be correlated within families. In general, the model being fitted is at best an approximation. All these will affect the biases, variances and estimates of variance of the parameter estimates.

These same comments also apply to non-survey contexts. Many studies which include family data do not explicitly recognize the possible impact of the intra-family correlations. Nested designs can suffer from correlations at each level of nesting.

However, even when the model assumptions have been violated, fitting the proportional hazards model can have both descriptive and analytic value. Every effort should be made by the analyst to avoid violation of the assumptions, so that the model is at least approximately true. Scott and Wild (1989) quantify this approximation error in the case of logistic regression. A similar definition may be used here, so that the approximation error would be the difference (or ratio) of the true hazard given x(t) to the hazard implied by the model. The hazard implied by the model in the case of a finite population is the value of the hazard function at ß = B, where B is the estimated parameter when all the finite population has been observed (with the appropriate censoring).

When the approximation error is small, fitting the wrong model will still be useful analytically. Even, when the approximation error is not small, some users may find that the model has descriptive value as being the best fit to some simple model in what may be a very complex setting. For example, the parameter value can give an indication of the change in relative risk when a covariate is changed. For some data analysts, it may be easiest to think in terms such as: (Approximately) how much reduction of a risk factor is required in order to reduce the hazard by 50%? These types of statements are used for public education of the risk of smoking or of high blood pressure.

We will describe a design-based procedure to estimate the parameters of the proportional hazards model and their estimates of variance. The estimation of variance employs methods not previously discussed in the literature. We also give the results of a simulation study, comparing the proposed procedure with other estimators when certain model assumptions have been violated.

2. Parameter Definition and Estimation

For a sample of size N, the partial likelihood function, conditional on the observed failure and censoring times is given by:

$$\prod_{i=1}^{N}\left[\frac{h\{t_i|\boldsymbol{x}_i(t_i)\}}{\sum_{j=1}^{N} Y_j(t_i)\,h\{t_i|\boldsymbol{x}_j(t_i)\}}\right]^{\delta_i},$$

where $t_1, \ldots, t_N$; t_i is the failure time of the i-th unit;

$$\delta_i = \begin{cases} 1 & \text{if the i-th unit is an observed failure,} \\ 0 & \text{if the i-th unit is censored;} \end{cases}$$

$$Y_j(t) = \begin{cases} 1 & \text{if } t \le t_j \\ 0 & \text{if } t > t_j. \end{cases}$$

To maximize the partial likelihood function we determine B such that

$$\sum_{i=1}^{N} \delta_i \left[\boldsymbol{x}_i(t_i) - \frac{\boldsymbol{S}^{(1)}(t_i, \boldsymbol{B})}{S^{(0)}(t_i, \boldsymbol{B})} \right] = \mathbf{0}, \qquad (2.1)$$

where

$$S^{(0)}(t, \boldsymbol{B}) = \frac{1}{N} \sum_{i=1}^{N} Y_i(t) \exp\{\boldsymbol{x}_i(t)'\boldsymbol{B}\}, \qquad (2.2)$$

$$\boldsymbol{S}^{(1)}(t, \boldsymbol{B}) = \frac{1}{N} \sum_{i=1}^{N} Y_i(t) \boldsymbol{x}_i(t) \exp\{\boldsymbol{x}_i(t)'\boldsymbol{B}\}, \qquad (2.3)$$

The solution to (2.1) defines the finite population parameter, B, which depends on the $Y_i(t)$'s and the δ_i's. We consider B to be the finite population parameter of interest, based on N observations, and which we wish to estimate from a sample of size n. The estimation procedure replaces the summations in (2.1), (2.2) and (2.3) by weighted sums, using the sampling weights, $\{w_i\}$. We scale the weights so that $\Sigma\, w_i = 1$. The weights are constructed so that the weighted sums are approximately unbiased and consistent estimates of the corresponding means over the finite population. In particular, the estimating equations for b, the estimator of B, are

$$\sum_{i=1}^{n} w_i \delta_i \left[\boldsymbol{x}_i(t_i) - \frac{^{(1)}(t_i, \boldsymbol{b})}{^{(0)}(t_i, \boldsymbol{b})} \right] = \mathbf{0}, \qquad (2.4)$$

where

$$^{(0)}(t, \boldsymbol{b}) = \sum_{i=1}^{n} w_i Y_i(t) \exp\{\boldsymbol{x}_i(t)'\boldsymbol{b}\}, \qquad (2.5)$$

$$^{(1)}(t, \boldsymbol{b}) = \sum_{i=1}^{n} w_i Y_i(t) \boldsymbol{x}_i(t) \exp\{\boldsymbol{x}_i(t)'\boldsymbol{b}\}, \qquad (2.6)$$

The estimator b which is the solution to (2.4) is referred to as the pseudo-maximum likelihood estimator; see Skinner (1989).

3. Variance Estimation

We are interested in obtaining the design-based variance of b - B. In Binder (1983), a general method is given for deriving the variance of parameter estimators which satisfy estimating equations of the form:

$$\hat{\boldsymbol{U}}(\boldsymbol{b}) = \sum_{i=1}^{n} w_i \boldsymbol{u}_i(\boldsymbol{b}) = \mathbf{0}. \qquad (3.1)$$

Using Taylor linearization, the design-based variance of b in (3.1) is given as:

$$\boldsymbol{J}^{-1} \boldsymbol{V}_{\boldsymbol{U}} \boldsymbol{J}^{-1}, \qquad (3.2)$$

where

$$\boldsymbol{V}_{\boldsymbol{U}} = \mathrm{Var}\{\hat{\boldsymbol{U}}(\boldsymbol{B})\}, \qquad (3.3)$$

$$\boldsymbol{J} = \frac{\partial \hat{\boldsymbol{U}}(\boldsymbol{B})}{\partial \boldsymbol{B}}, \qquad (3.4)$$

$$\boldsymbol{U}(\boldsymbol{B}) = \sum_{i=1}^{N} \boldsymbol{u}_i(\boldsymbol{B}) = \mathbf{0}. \qquad (3.5)$$

Binder (1983) gave conditions under which (3.2) is the asymptotic variance for b. For example, Chambless and Boyle (1985) derived the design-based variance for the discrete proportional hazards model suggested by Prentice and Gloeckler (1978) using this approach.

In our case, however, the form of $\hat{U}(b)$ does not conform to the expression given in (3.1), since the u_i's of the pseudo-maximum likelihood equation are functions of weighted sums, where the sums depend on the t_i's. Therefore, we need to adapt the derivation for the variance of b. We shall still consider variances based on Taylor series expansions, as opposed to other methods of variance estimation; see, for example Rust (1985).

The method we use is to find an alternative expression to the partial likelihood equations, given by (2.1), which has the form given by (3.5), and which is asymptotically equivalent to (2.1). Once this is found, the variance given by (3.2) may be used. In fact, Lin and Wei (1989) derived such an expression in the case of independent sampling. We summarize their results for the pseudo-maximum likelihood equation given by (2.4).

Expression (2.4) may be rewritten as

$$\sum_{i=1}^{n} w_i \delta_i \boldsymbol{x}_i(t_i) - \int_0^{\infty} \left[\frac{^{(1)}(t, \boldsymbol{b})}{^{(0)}(t, \boldsymbol{b})} \right] d\hat{G}(t) = \mathbf{0}, \qquad (3.6)$$

where

$$(t) = \sum_{i=1}^{n} w_i G_i(t),$$

$$G_i(t) = \begin{cases} 1 & \text{if } t \ge t_i \text{ and } \delta_i = 1 \\ 0 & \text{otherwise.} \end{cases}$$

We define

$$G(t) = \sum_{i=1}^{N} G_i(t).$$

and take a Taylor series expansion of (3.6) around $\hat{S}^{(0)} = S^{(0)}$, $\hat{S}^{(1)} = S^{(1)}$ and $\hat{G} = G$. The first order Taylor expansion of the left-hand side of (3.6) is

$$\sum_{i=1}^{n} w_i \delta_i \boldsymbol{x}_i(t_i)$$

$$- \int_0^{\infty} \left[\frac{\hat{S}^{(1)}(t,\boldsymbol{b})}{S^{(0)}(t,\boldsymbol{b})} \right] dG(t)$$

$$+ \int_0^{\infty} \left[\frac{\hat{S}^{(0)}(t,\boldsymbol{b}) S^{(1)}(t,\boldsymbol{b})}{S^{(0)}(t,\boldsymbol{b})^2} \right] dG(t)$$

$$- \int_0^{\infty} \left[\frac{S^{(1)}(t,\boldsymbol{b})}{S^{(0)}(t,\boldsymbol{b})} \right] d\hat{G}(t) \; . \tag{3.7}$$

Since, $\hat{S}^{(0)}$, $\hat{S}^{(1)}$ and $\hat{G}$ are consistent estimates of $S^{(0)}$, $S^{(1)}$ and G, respectively, the remainder terms are negligible. Therefore, expression (3.7) is asymptotically equivalent to $\hat{U}(b)$ in (3.1), where $u_i(b)$ is

$$\delta_i \left[\boldsymbol{x}_i(t_i) - \frac{S^{(1)}(t_i,\boldsymbol{b})}{S^{(0)}(t_i,\boldsymbol{b})} \right]$$

$$- \int_0^{\infty} \left[\frac{Y_i(t) \boldsymbol{x}_i(t) \exp\{\boldsymbol{x}_i(t)'\boldsymbol{b}\}}{S^{(0)}(t,\boldsymbol{b})} \right] dG(t)$$

$$+ \int_0^{\infty} \left[\frac{Y_i(t) S^{(1)}(t,\boldsymbol{b}) \exp\{\boldsymbol{x}_i(t)'\boldsymbol{b}\}}{S^{(0)}(t,\boldsymbol{b})^2} \right] dG(t) \; . \tag{3.8}$$

To estimate the variance of $\hat{U}(b)$, it is necessary to substitute estimates of $S^{(0)}$, $S^{(1)}$ and G into (3.8). This results in $\hat{u}_i(b)$ defined as

$$\delta_i \left[\boldsymbol{x}_i(t_i) - \frac{\hat{S}^{(1)}(t_i,\boldsymbol{b})}{\hat{S}^{(0)}(t_i,\boldsymbol{b})} \right]$$

$$- \sum_{j=1}^{n} \delta_j w_j \left[\frac{Y_i(t_j) \boldsymbol{x}_i(t_j) \exp\{\boldsymbol{x}_i(t_j)'\boldsymbol{b}\}}{\hat{S}^{(0)}(t_j,\boldsymbol{b})} \right]$$

$$+ \sum_{j=1}^{n} \delta_j w_j \left[\frac{Y_i(t_j) \hat{S}^{(1)}(t_j,\boldsymbol{b}) \exp\{\boldsymbol{x}_i(t_j)'\boldsymbol{b}\}}{\hat{S}^{(0)}(t_j,\boldsymbol{b})^2} \right] . \tag{3.9}$$

Using these values for $\hat{u}_i(b)$, V_U in expression (3.3) can be estimated using design-based methods. The matrix J is estimated by taking the derivative of (2.4) with respect to b. The variance of b is then given by (3.2).

4. Simulation Study

To assess the performance of the variance estimator, we performed a simulation study and computed the coverage properties of the implied confidence interval, assuming approximate normality of b - B. Note, we have shown that $\hat{U}$ is asymptotically equivalent to the weighted sum of the u_i's, so that asymptotic normality would result from conditions similar to that in Binder (1983).

We let z_1, z_2 and z_3 be independent standard normal random variables. Also bin_1 and bin_2 are independent 0-1 random variables with p=0.5. We generated a population of N=5000 lifetimes using hazard function

$$h(t) = \exp\{z_1\} + 0.5 bin_1 .$$

The largest 2500 observations were right-censored. We notice that this model is not a proportional hazards model.

The population was then stratified using size variable $\exp\{z_1 + 0.5z_2\}$. Five strata were obtained and samples of size n=500 were drawn using stratified random sampling.

Two methods of stratification and sample allocation were used. In the first case, the sample allocation was based on Neyman allocation, and the stratification was chosen to minimize the variance of the estimated mean of the stratification variable. The resulting strata sizes were 2319, 1473, 798, 342 and 68, with corresponding sample sizes 107, 105, 110, 110 and 68, respectively.

In the second case, five equally-sized strata were delineated, based on the ranking of the stratification variable. The sample allocation was equal within strata.

We fit the proportional hazards model using the two variables: $z_1 + 0.5z_3$, and bin_2. For each case, we generated 2000 stratified samples. We computed three variances for each sample:

- the design based variances given in Section 3;
- the usual model-based variance using unweighted estimates and the model-based information matrix;
- the "robust" variance suggested by Lin and Wei (1989).

This last estimator uses the same linearization as our method, except that it is based on unweighted estimates and assumes simple random sampling with replacement in the variance calculation.

The population parameter values for the N=5000 units of the population were B_1 = 0.622 and B_2 = 0.041. Table 1 gives the average values of the parameter estimates over the 2000 simulations.

TABLE 1. Expected value of parameter estimates based on 2000 simulations

	Unequal Allocation		Equal Allocation	
	B_1	B_2	B_1	B_2
Weighted	0.625	0.046	0.624	0.037
Unweighted	0.679	-0.011	0.624	0.037

The standard deviations of the sampling distribution of the parameters estimates were also calculated. These are given in Table 2.

TABLE 2. Standard deviation of parameter estimates based on 2000 simulations

	Unequal Allocation		Equal Allocation	
	B_1	B_2	B_1	B_2
Weighted	0.071	0.145	0.059	0.012
Unweighted	0.045	0.094	0.059	0.012

From Table 1, we notice that the unweighted estimates exhibit larger biases, as would be expected. From Table 2, we see that the variances of the parameter estimates are smaller with an equal allocation design. This shows that optimal stratification and allocation with an imprecise stratification variable can lead to poorer results.

For each of the 2000 samples, we also computed the z-score based on b - B using the three different variance estimation methods. P-P plots of the are given in Figures 1 to 6. Figures 1 to 3 are for the unequal allocation case and Figures 4 to 6 are for the equal allocation case. The z-scores for both fitted variables are given on the same plot.

Figures 1 to 3 show the superiority of the design-based methods in the case of unequal allocation. This is due to the biases inherent in the unweighted procedures. For the equal allocation case, where the estimators are all unweighted, the model-based method (Figure 5) is inferior to the other two methods. It appears that in this case the model-based variance estimates tend to be too large, giving a relatively conservative test. This is possibly due to the fact that the stratification has reduced the variance of the estimators.

5. Summary

This simulation study has shown that the design-based approach to estimating variances, using Taylor linearization methods may perform at least as well as the model-based methods. For the case of equal allocation, the procedure suggested by Lin and Wei (1989) performed as well as the design-based method, for the case of the simulation study. This could be due to the fact that there was no intra-cluster correlation present, so the independence assumption was valid, even though there were missing variables in the model. It seems, though, that the design-based approach can be used effectively in a wide set of circumstances.

The method suggested for obtaining the linearization shows that the general procedure suggested by Binder (1983) can be extended to handle models such as we have discussed.

Acknowledgments

I am grateful to Michael Hidiroglou and Georgia Roberts for their insightful comments to an earlier draft of this paper.

References

Binder, D.A. (1983), "On the Variances of Asymptotically Normal Estimators from Complex Surveys," International Statistical Review, 51, 279-292.

Chambless, L.E. and K.E. Boyle (1985), "Maximum Likelihood Methods for Complex Sample Data: Logistic regression and Discrete Proportional Hazards Models," Communications in Statistics - Theory and Methods, 14, 1377-1392.

Lin, D.Y. and L.J. Wei (1989), "The Robust Inference for the Cox Proportional Hazards Model," Journal of the American Statistical Association, 84, 1074- 1078.

Prentice, R.L. and L.A. Gloeckler (1978), "Regression Analysis of Grouped Survival Data with Application to Breast Cancer Data," Biometrics, 34, 57-67.

Rust, K. (1985), "Variance Estimation for Complex Estimators in Sample Surveys," Journal of Official Statistics, 1, 381-397.

Scott, A.J. and C.J. Wild (1989), "Selection Based on the Response Variable in Logistic Regression," Analysis of Complex Surveys, (C.J. Skinner, D.Holt and T.M.F. Smith, eds.) Wiley, 191-205.

Skinner, C.J. (1989), "Domain Means, Regression and Multivariate Analysis," Analysis of Complex Surveys, (C.J. Skinner, D.Holt and T.M.F. Smith, eds.) Wiley, 59-87.

Struthers, C.A. and J.D. Kalbfleisch (1986), "Misspecified Proportional Hazard Models," Biometrika, 73, 363-369.

FIGURE 1. Design-Based Methods

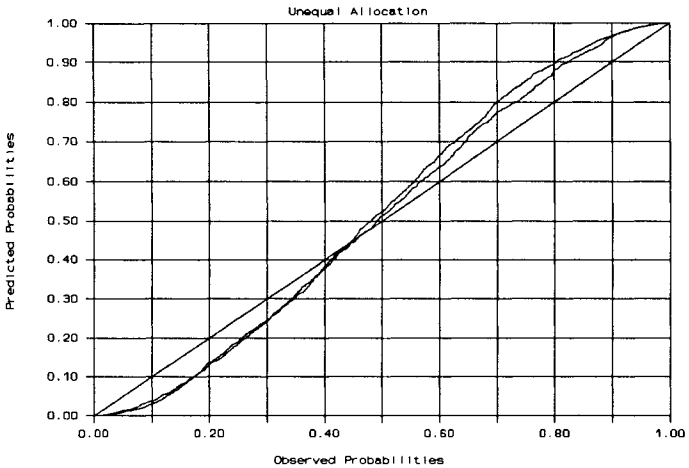

FIGURE 2. Model-Based Methods

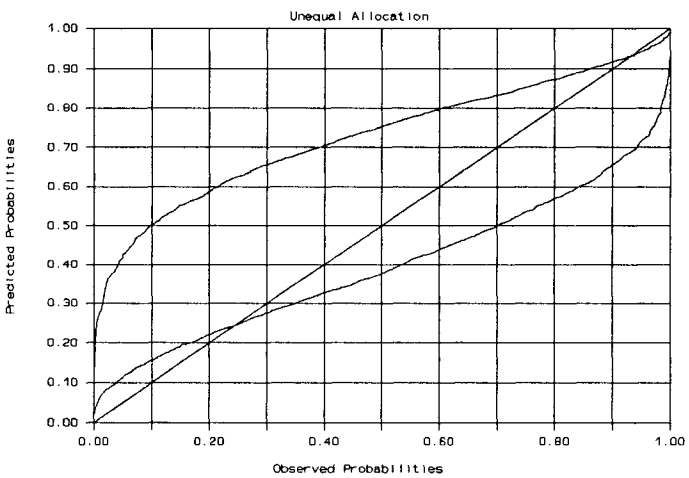

FIGURE 3. "Robust" Methods

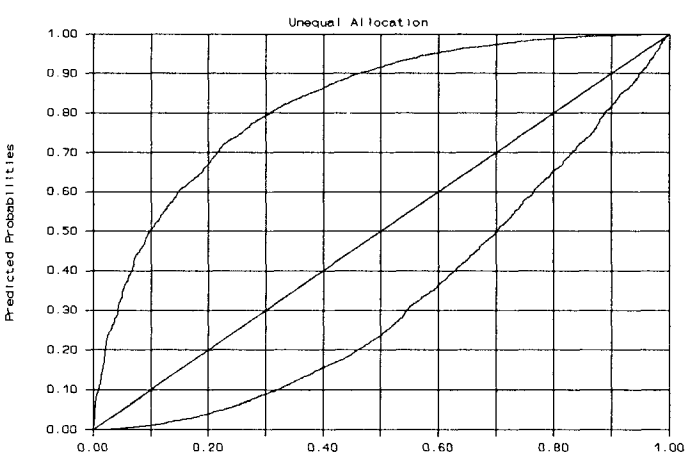

FIGURE 4. Design-Based Methods

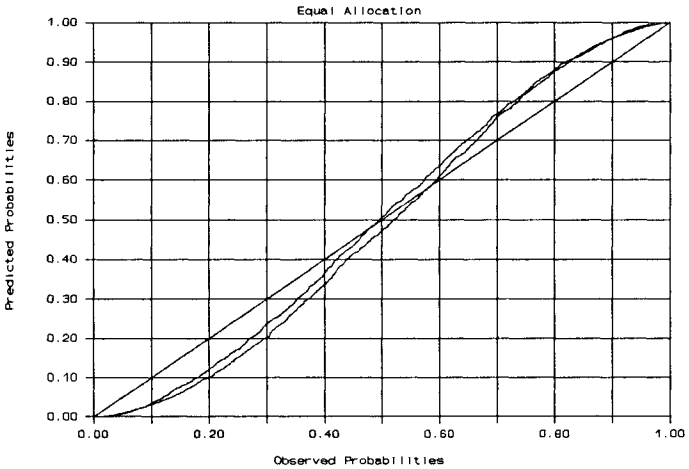

FIGURE 5. Model-Based Methods

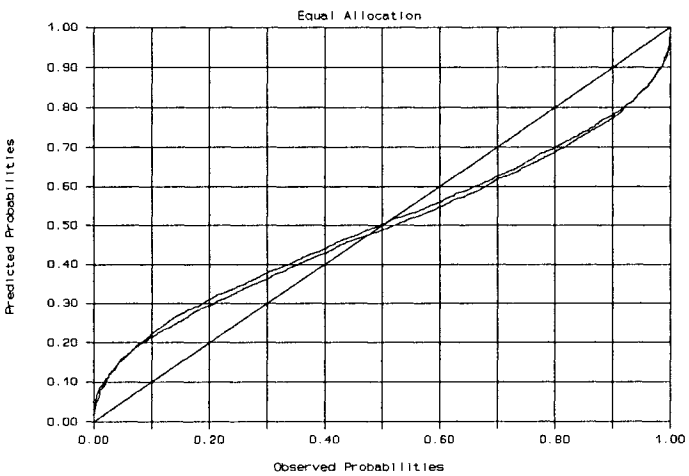

FIGURE 6. "Robust" Methods