Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

Firth's penalized likelihood for proportional hazards regressions for complex surveys

by Pushpal K. Mukhopadhyay

Release date: December 15, 2020

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

Email at STATCAN.infostats-infostats.STATCAN@canada.ca

Telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-514-283-9350

Depository Services Program

- Inquiries line 1-800-635-7943
- Fax line 1-800-565-7757

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "Standards of service to the public."

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Published by authority of the Minister responsible for Statistics Canada

An HTML version is also available.

Cette publication est aussi disponible en français.

Survey Methodology, December 2020
Vol. 46, No. 2, pp. 215-241
Statistics Canada, Catalogue No. 12-001-X

Firth's penalized likelihood for proportional hazards regressions for complex surveys

Pushpal K. Mukhopadhyay[1]

Abstract

This article proposes a weight scaling method for Firth's penalized likelihood for proportional hazards regression models. The method derives a relationship between the penalized likelihood that uses scaled weights and the penalized likelihood that uses unscaled weights, and it shows that the penalized likelihood that uses scaled weights have some desirable properties. A simulation study indicates that the penalized likelihood using scaled weights produces smaller biases in point estimates and standard errors than the biases produced by the penalized likelihood using unscaled weights. The weighted penalized likelihood is applied to estimate hazard rates for heart attacks by using a public-use data set from the National Health and Epidemiology Followup Study (NHEFS). SAS® statements to estimate hazard rates using data from complex surveys are given in the appendix.

Key Words: Monotone likelihood; Delete-one jackknife; Weight scaling.

1 Introduction

The Cox proportional hazards regression model (Cox, 1972) is widely used to analyze survival data. It is a semiparametric model that explains the effect of explanatory variables on hazard rates. The model assumes a linear form for the effect of the explanatory variables but allows an unspecified form for the underlying survivor function. The parameters of the model are estimated by maximizing a partial likelihood (Cox, 1972, 1975).

For estimating canonical parameters in the exponential family distributions, Firth (1993) suggested multiplying the likelihood by the Jeffreys prior to obtain a maximum likelihood estimate that is first-order unbiased. The penalized likelihood is of the form

$$L_p(\boldsymbol{\beta}) = L(\boldsymbol{\beta})|I(\boldsymbol{\beta})|^{0.5}$$

where $L(\boldsymbol{\beta})$ is the unpenalized likelihood, I is the information matrix, and $\boldsymbol{\beta}$ is a vector of regression parameters. Firth's penalized likelihood is a very useful technique in practice, not only to reduce bias but also to correct for monotone likelihoods.

Proportional hazards regression models often suffer from monotone likelihoods, in which the likelihood converges to a finite value but at least one parameter diverges (Heinze, 1999). Firth's penalized likelihood is also used to correct monotone likelihoods and to obtain parameter estimates that converge (Heinze, 1999; Heinze and Schemper, 2001; Heinzel, Rüdiger and Schilling, 2002).

Although Firth's penalized likelihood is useful for reducing biases and for obtaining estimates from monotone likelihoods, the penalized likelihood is not studied for complex surveys involving unequal

1. Pushpal K. Mukhopadhyay, Advanced Analytics Division, SAS Institute Inc., Cary, NC, U.S.A. E-mail: pushpal.mukhopadhyay@sas.com.

weights. It is reasonable to use a weighted likelihood for complex surveys to compensate for unequal weighting (Fuller, 1975; Binder and Patak, 1994). Survey data sets commonly include design weights or analysis weights for which the sum of the weights is an estimator of the population size. However, these unscaled weights will not appropriately scale the information matrix that is used in the penalty term. It is desirable for proportional hazards regression parameters for survey data to have the following two properties:

- *Invariance:* The point estimates and standard errors for the regression parameters should be invariant to the scale of the weights.
- *Closeness:* The Taylor linearized variance for the estimated regression parameters should be close to the delete-one jackknife variance.

In this article, we first show that if the Firth correction is not used, then both the invariance and closeness are satisfied; but if the Firth correction is used with the unscaled weights, then the point estimates and the standard errors are not invariant to the scale of the weights. That is, if the weights are multiplied by a constant and the Firth correction is used, then the point estimates and standard errors will be different. We then propose a commonsense weight scaling method and demonstrate that the Firth correction using the scaled weights has both properties. The only difference between the scaled and unscaled weights is that the sum of the scaled weights is equal to the sample size, but the sum of the unscaled weights is an estimator of the population size.

1.1 Example that uses unscaled weights

We used a data set from a study of 65 myeloma patients who were treated with alkylating agents (Lee, Wei and Amato, 1992) to demonstrate the properties of Firth's penalized likelihood that uses unscaled weights. Survival times in months were recorded for each patient. Patients who were alive after the study period were considered to be censored. The following variables were available for each patient:

- Time: Survival time in months,
- Vstatus: Patient status, zero or one, indicating whether the patient was alive or dead, respectively,
- LogBUN: Log of blood urea nitrogen level,
- HGB: Blood hemoglobin level.

To create a monotone likelihood, we added a new explanatory variable, Contrived, such that its value at all event times is the largest of all values in the risk set (see the example "Firth's Correction for Monotone Likelihood" in "The PHREG Procedure" in SAS Institute Inc. (2018)). The variable Contrived has the value 1 if the observed survival time is less than or equal to 65; otherwise it has the value 0.

To demonstrate the effect of weights in Firth's penalized likelihood, we created three weight variables, w1, w3, and w5, with the values of 1, 1,000, and 100,000 for each observation, respectively. Proportional hazards regression parameters are estimated by maximizing a weighted likelihood as described in Section 1.2. Because $w1$ has the value 1 for all observations, using $w1$ in the analysis is equivalent to performing the unweighted analysis.

We fitted the following two proportional hazards models using the PHREG procedure in SAS/STAT® (see “The PHREG Procedure” in SAS Institute Inc. (2018)):

$$\lambda(t, \mathbf{Z}) = \lambda_0(t)\exp(\beta_1\text{LogBUN} + \beta_2\text{HGB})$$

$$\lambda(t, \mathbf{Z}) = \lambda_0(t)\exp(\beta_1\text{LogBUN} + \beta_2\text{HGB} + \beta_3\text{Contrived})$$

where $\lambda(t)$ and $\lambda_0(t)$ are the hazard function and the baseline hazard function, respectively. Firth’s penalized likelihood is not required in order to fit the first model without Contrived (the likelihood converged in three iteration steps), but the second model containing the variable Contrived does not converge without the Firth penalty in the likelihood. Table 1.1 displays the value of the likelihood and the three regression coefficients for 14 iterations. Although the objective function and the coefficients for LogBun and HGB converge to a finite value after the fourth iteration, the coefficients for Contrived diverges. This is an example of a monotone likelihood for the variable Contrived. Because of this monotonicity, Firth’s penalized likelihood must be used to fit the second model containing Contrived.

Table 1.1
Maximum likelihood iteration history showing a monotone likelihood for the variable Contrived

Iteration Number	Likelihood Value	LogBUN	HGB	Contrived
1	-140.693405	1.994882	-0.084319	1.466331
2	-137.784163	1.679468	-0.109068	2.778361
3	-136.971190	1.714061	-0.111564	3.938095
4	-136.707893	1.718174	-0.112273	5.003054
5	-136.616426	1.718755	-0.112370	6.027436
6	-136.583520	1.718829	-0.112382	7.036445
7	-136.571515	1.718839	-0.112384	8.039764
8	-136.567113	1.718841	-0.112384	9.040985
9	-136.565495	1.718841	-0.112384	10.041434
10	-136.564900	1.718841	-0.112384	11.041600
11	-136.564681	1.718841	-0.112384	12.041660
12	-136.564601	1.718841	-0.112384	13.041683
13	-136.564571	1.718841	-0.112384	14.041691
14	-136.564560	1.718841	-0.112384	15.041694

If Contrived is not used as an explanatory variable, then all three sets of weights produce the same point estimates and Taylor linearized variance estimates (Table 1.2). The delete-one jackknife variance estimates are also the same for all three sets of weights. Thus, the point estimates and the standard errors are invariant to the scale of the weights when the Firth correction is not used.

Table 1.2
Parameter estimates and standard errors without the Firth correction for all three sets of weights

	Estimate	Std. Err.
LogBUN	1.674	0.583
HGB	-0.119	0.060

However, if the unscaled weights are used, then the point estimates for Contrived are not invariant to the scale of the weights. Table 1.3 displays the parameter estimates for three sets of weights when Contrived is used as an explanatory variable (and Firth's penalized likelihood is applied). Because the likelihood is not monotone (Table 1.1) for LogBun and HGB, the point estimates for these two coefficients are not affected by the scale of the weights.

Table 1.3
Parameter estimates with the Firth correction and unscaled weights

	Weight $w1$		**Weight $w3$**		**Weight $w5$**	
	Estimate	**Std. Err.**	**Estimate**	**Std. Err.**	**Estimate**	**Std. Err.**
LogBUN	1.722	0.584	1.719	1.85E-2	1.719	1.85E-3
HGB	-0.112	0.061	-0.112	1.93E-3	-0.112	1.93E-4
Contrived	3.815	1.558	10.629	1.38	14.633	1.02

If Contrived is not used as an explanatory variable, then the ratio of jackknife standard errors to Taylor linearized standard errors is 1.13 and 1.10 for all three sets of weights for the variables LogBUN and HGB, respectively. Thus the ratio of the jackknife variance to the Taylor linearized variance for the unpenalized likelihood is invariant to the scale of weights, and it is reasonable to expect the ratio to be invariant when the penalized likelihood is used.

1.2 A brief review of point and variance estimates for regression parameters for finite populations

Before we discuss the weight scaling method, we briefly review point and variance estimates for regression parameters for proportional hazards regression of complex surveys involving unequal weights. Lin and Wei (1989); Binder (1990, 1992); Lin (2000); and Boudreau and Lawless (2006) discussed pseudo-maximum likelihood estimation of proportional hazard regression parameters for survey data. For a more general description for estimating regression parameters for complex surveys, see Kish and Frankel (1974); Godambe and Thompson (1986); Pfeffermann (1993), Korn and Graubard (1999, Chapter 3), Chambers and Skinner (2003, Chapter 2), and Fuller (2009, Section 6.5). Wolter (2007) described different variance estimation techniques for survey data.

Let $\mathcal{U}_N = \{1, 2, \ldots, N\}$ be the set of indices and let $\mathcal{F}_N$ be the set of values for a finite population of size N. The survival time of each member of the finite population is assumed to follow its own hazard function, $\lambda_i(t)$, expressed as

$$\lambda_i(t) = \lambda(t; \mathbf{Z}_i(t)) = \lambda_0(t) \exp\left(\mathbf{Z}_i'(t)\, \boldsymbol{\beta}\right)$$

where $\lambda_0(t)$ is an arbitrary and unspecified baseline hazard function, $\mathbf{Z}_i(t)$ is a vector of size P of explanatory variables for the i^{th} unit at time t, and $\boldsymbol{\beta}$ is a vector of unknown regression parameters.

The partial likelihood function introduced by Cox (1972, 1975) eliminates the unknown baseline hazard $\lambda_0(t)$ and accounts for censored survival times. If the entire population is observed, then this partial likelihood function can be used to estimate $\boldsymbol{\beta}$. Let $\boldsymbol{\beta}_N$ be the desired estimator.

Assuming a working model with uncorrelated responses, $\boldsymbol{\beta}_N$ is obtained by maximizing the partial log likelihood,

$$l_N(\boldsymbol{\beta}) = \sum_{i \in \mathcal{U}_N} \log\{\mathrm{L}(\boldsymbol{\beta}; \mathbf{Z}_i(t), t_i)\}$$

with respect to $\boldsymbol{\beta}$, where $\mathrm{L}(\boldsymbol{\beta}; \mathbf{Z}_i(t), t_i)$ is Cox's partial likelihood function.

Assume that a probability sample A_N is selected from the finite population $\mathcal{U}_N$. Let π_i be the selection probability and $w_i\,(=\pi_i^{-1})$ be the sampling weight for unit i. Further assume that explanatory variables $\mathbf{Z}_i(t)$ and survival time t_i are available for every unit in sample A_N. A design unbiased estimator for the finite population log likelihood is

$$l(\boldsymbol{\beta}) = \sum_{i \in A_N} \pi_i^{-1} \log\{\mathrm{L}(\boldsymbol{\beta}; \mathbf{Z}_i(t), t_i)\} = \sum_{i \in A_N} w_i \log\{\mathrm{L}(\boldsymbol{\beta}; \mathbf{Z}_i(t), t_i)\}.$$

A sample-based estimator $\hat{\boldsymbol{\beta}}_N$ for the finite population quantity $\boldsymbol{\beta}_N$ can be obtained by maximizing the partial pseudo-log likelihood $l(\boldsymbol{\beta}; \mathbf{Z}_i(t), t_i)$ with respect to $\boldsymbol{\beta}$. The design-based variance for $\hat{\boldsymbol{\beta}}_N$ is obtained by assuming that the set of finite population values $\mathcal{F}_N$ is fixed.

The weighted Breslow likelihood can be expressed as

$$L(\boldsymbol{\beta}) = \prod_{k=1}^{K} \frac{\exp\left(\boldsymbol{\beta}' \sum_{\mathcal{D}_k} w_i \mathbf{Z}_i(t)\right)}{\left\{\sum_{\mathcal{R}_k} w_i \exp(\boldsymbol{\beta}' \mathbf{Z}_i(t))\right\}^{\sum_{\mathcal{D}_k} w_i}}$$

where $\mathcal{R}_k$ is the risk set just before the k^{th} ordered event time $t_{(k)}$, $\mathcal{D}_k$ is the set of individuals who fail at the $t_{(k)}$, and K is the number of distinct event times.

The point estimates for $\boldsymbol{\beta}$ are obtained by maximizing $l(\boldsymbol{\beta}) = \log[L(\boldsymbol{\beta})]$.

Although the weights are sufficient for estimating regression coefficients for the finite population, stratification and clustering information must also be used to estimate sampling variability. In order to estimate sampling variability, you can use either the Taylor series linearization method or a replication method.

1.2.1 Analytic variance estimator using the Taylor series linearization method

The Taylor series linearization method uses a sum of squares of the weighted score residuals to estimate the sampling variability.

Define $\bar{\mathbf{Z}}(\boldsymbol{\beta}, t) = \frac{\mathbf{S}^{(1)}(\boldsymbol{\beta}, t)}{S^{(0)}(\boldsymbol{\beta}, t)}$, where

$$S^{(0)}(\boldsymbol{\beta}, t) = \sum_{A_N} w_i I(t_i \geq t) \exp(\boldsymbol{\beta}' \mathbf{Z}_i(t))$$

and

$$\mathbf{S}^{(1)}(\boldsymbol{\beta}, t) = \sum_{A_N} w_i I(t_i \geq t) \exp(\boldsymbol{\beta}' \mathbf{Z}_i(t)) \mathbf{Z}_i(t).$$

The score residual for the i^{th} subject is

$$\begin{aligned}\mathbf{u}_i(\boldsymbol{\beta}) = {} & \Delta_i\left\{\mathbf{Z}_i(t_i) - \overline{\mathbf{Z}}(\boldsymbol{\beta}, t_i)\right\} \\ & - \sum_{j \in A_N}\left[\Delta_j \frac{w_j I(t_i \geq t_j)\exp(\boldsymbol{\beta}'\mathbf{Z}_i(t_j))}{S^{(0)}(\boldsymbol{\beta}, t_j)}\left\{\mathbf{Z}_i(t_j) - \overline{\mathbf{Z}}(\boldsymbol{\beta}, t_j)\right\}\right]\end{aligned}$$

where Δ_i is the event indicator.

Then the Taylor linearized variance estimator is

$$\hat{\mathbf{V}}(\hat{\boldsymbol{\beta}}) = \mathcal{I}^{-1}(\hat{\boldsymbol{\beta}})\, \mathrm{G}\mathcal{I}^{-1}(\hat{\boldsymbol{\beta}})$$

where $\mathcal{I}(\hat{\boldsymbol{\beta}})$ is the observed information matrix and the $p \times p$ matrix $\mathbf{G}$ is defined as

$$\mathbf{G} = \sum_{i,\ j \in A_N:\, i<j} \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}}\left(\frac{\hat{\mathbf{u}}_i}{\pi_i} - \frac{\hat{\mathbf{u}}_j}{\pi_j}\right)'\left(\frac{\hat{\mathbf{u}}_i}{\pi_i} - \frac{\hat{\mathbf{u}}_j}{\pi_j}\right)$$

where π_{ij} are the joint inclusion probabilities for units i and j.

In particular, for stratified cluster designs in which the PSUs are selected by using a simple random sample without replacement, the $p \times p$ matrix $\mathbf{G}$ reduces to

$$\mathbf{G} = \sum_{h=1}^{H}\frac{n_h(1 - f_h)}{n_h - 1}\sum_{i=1}^{n_h}(\mathbf{e}_{hi+} - \bar{\mathbf{e}}_{h\cdot\cdot})'(\mathbf{e}_{hi+} - \bar{\mathbf{e}}_{h\cdot\cdot})$$

where $\mathbf{e}_{hi+}$ is the weighted sum of the score residuals, $\hat{\mathbf{u}}_{hij}$, in stratum h and PSU i; $\bar{\mathbf{e}}_{h\cdot\cdot}$ is the mean of $\mathbf{e}_{hi+}$; n_h is the number of PSUs; and f_h is the sampling fraction in stratum h.

These estimators are well studied in the sample survey literature. For example, Binder (1992) and Lin (2000) provide conditions under which $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{V}}(\hat{\beta})$ are consistent. Chambless and Boyle (1985) derived the design-based variance and asymptotic normality for discrete proportional hazards models.

1.2.2 Replication variance estimator using the delete-one jackknife method

The jackknife method is a commonly used replication variance estimation method for complex surveys. To create replicates, it deletes (assigns a zero weight to) one PSU at a time from the full sample. In each replicate, the sampling weights of the remaining PSUs are modified by the jackknife coefficient α_r. The modified weights are called *replicate weights*.

Let PSU i_r in stratum h_r be omitted from the r^{th} replicate; then the replicate weights and jackknife coefficients are given by

$$w_{hij}^{(r)} = \begin{cases} 0 & i = i_r \text{ and } h = h_r \\ w_{hij} / \alpha_r & i \neq i_r \text{ and } h = h_r \\ w_{hij} & h \neq h_r \end{cases}$$

and $\alpha_r = \frac{n_{h_r} - 1}{n_{h_r}}$, respectively, for all observation units j in stratum h and PSU i. The number of PSUs in stratum h_r is n_{h_r}.

The jackknife method can be applied to estimate variances for the estimated regression parameters for Cox's model because the model parameters are solutions of a set of estimating equations that are smooth functions of totals (the corresponding score functions are given in Section 2). Properties of jackknife variance estimators for proportional hazard regression models are discussed in Shao and Tu (1995, Section 8.3).

To apply the jackknife method, model parameters are estimated by using the full sample and by using every replicate sample. Let $\hat{\boldsymbol{\beta}}$ be the estimated proportional hazards regression coefficients from the full sample, and let $\hat{\boldsymbol{\beta}}_r$ be the estimated regression coefficients from the r^{th} replicate. Then the covariance matrix of $\hat{\boldsymbol{\beta}}$ is estimated by

$$\hat{\mathbf{V}}(\hat{\boldsymbol{\beta}}) = \sum_{r=1}^{R} \alpha_r (\hat{\boldsymbol{\beta}}_r - \hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}}_r - \hat{\boldsymbol{\beta}})'.$$

If the sampling fractions are not ignorable, then the covariance matrix of $\hat{\boldsymbol{\beta}}$ is estimated by

$$\hat{\mathbf{V}}(\hat{\boldsymbol{\beta}}) = \sum_{r=1}^{R} \alpha_r (1 - f_r)(\hat{\boldsymbol{\beta}}_r - \hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}}_r - \hat{\boldsymbol{\beta}})'$$

where $f_r = \frac{n_{h_r}}{N_{h_r}}$ is the sampling fraction in stratum h_r.

In practice, both Taylor linearized variance and jackknife variance estimates are used to construct Wald t confidence intervals with $R - H$ degrees of freedom, where R is the number of PSUs (or the number of replicates) and H is the number of strata.

It is straightforward to show that the jackknife variance estimator is algebraically equivalent to the Taylor linearized estimator for design linear estimators. But for design nonlinear estimators, such as the regression coefficients for proportional hazards regression models, the jackknife method tends to produce slightly higher variance estimates than the Taylor linearized method (Fuller, 2009).

Note that if the full sample estimate suffers from a monotone likelihood, then it is very likely that most replicate samples will also suffer from monotone likelihoods. This will results in many "unusable" replicate estimates.

Survey data analysis procedures in SAS/STAT support both Taylor linearized and replication variance estimation methods (Mukhopadhyay, An, Tobias and Watts, 2008).

2 Weight scaling

Let w_i be the weight for unit i. We propose to use $\tilde{w}_i = \left(\sum_{A_N} 1 \big/ \sum_{A_N} w_i\right) w_i = \left(n \big/ \sum_{A_N} w_i\right) w_i$ as the scaled weight. By construction, the scaled weights are invariant to the scale of the weight. That is, $\tilde{w}_i^* = \left(n \big/ \sum_{A_N} \gamma w_i\right) \gamma w_i = \left(n \big/ \sum_{A_N} w_i\right) w_i = \tilde{w}_i$ for all $\gamma \neq 0$.

Firth's penalized likelihood is given by $L_p(\boldsymbol{\beta}) = L(\boldsymbol{\beta}) |\mathcal{I}(\boldsymbol{\beta})|^{0.5}$, where $L(\boldsymbol{\beta})$ and $\mathcal{I}(\boldsymbol{\beta})$ are the unpenalized likelihood and information matrix, respectively. The penalized log likelihood is

$$l_p(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) + 0.5 \log(|\mathcal{I}(\boldsymbol{\beta})|).$$

In particular, when the scaled weights are used, the Breslow unpenalized log partial likelihood (Breslow, 1974) is

$$l(\boldsymbol{\beta}) = \sum_{k=1}^{K} \left\{ \boldsymbol{\beta}' \sum_{i \in \mathcal{D}_k} \tilde{w}_i \mathbf{Z}_i(t_k) - \left(\sum_{i \in \mathcal{D}_k} \tilde{w}_i \right) \log \sum_{i \in \mathcal{R}_k} \tilde{w}_i \exp(\boldsymbol{\beta}' \mathbf{Z}_i(t_k)) \right\}$$

where w_i is the unscaled weight for unit i.

Denote

$$\mathbf{S}_k^{(a)}(\boldsymbol{\beta}) = \sum_{i \in \mathcal{R}_k} \tilde{w}_i \exp(\boldsymbol{\beta}' \mathbf{Z}_i(t_k)) [\mathbf{Z}_i(t_k)]^{\otimes a}$$

where k is the k^{th}-ordered event time, $a = 0, 1, 2$, $[\mathbf{Z}_i(t_k)]^{\otimes 0}$ is 1, $[\mathbf{Z}_i(t_k)]^{\otimes 1}$ is the vector $\mathbf{Z}_i(t_k)$, and $[\mathbf{Z}_i(t_k)]^{\otimes 2}$ is the matrix $[\mathbf{Z}_i(t_k)][\mathbf{Z}_i(t_k)]'$.

Then the score function is given by

$$\begin{aligned} \mathbf{U}(\boldsymbol{\beta}) &\equiv (U(\beta_1), \ldots, U(\beta_p))' \\ &= \frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \\ &= \sum_{k=1}^{K} \left\{ \sum_{i \in \mathcal{D}_k} \tilde{w}_i \mathbf{Z}_i(t_k) - \sum_{i \in \mathcal{D}_k} \tilde{w}_i \frac{\mathbf{S}_k^{(1)}(\boldsymbol{\beta})}{S_k^{0}(\boldsymbol{\beta})} \right\} \end{aligned}$$

and the Fisher information matrix is given by

$$\begin{aligned} \mathcal{I}(\boldsymbol{\beta}) &= -\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2} \\ &= \sum_{k=1}^{K} \sum_{i \in \mathcal{D}_k} \tilde{w}_i \left\{ \frac{\mathbf{S}_k^{(2)}(\boldsymbol{\beta})}{S_k^{(0)}(\boldsymbol{\beta})} - \left[\frac{\mathbf{S}_k^{(1)}(\boldsymbol{\beta})}{S_k^{(0)}(\boldsymbol{\beta})} \right] \left[\frac{\mathbf{S}_k^{(1)}(\boldsymbol{\beta})}{S_k^{(0)}(\boldsymbol{\beta})} \right]' \right\}. \end{aligned}$$

Denote

$$\mathbf{Q}_{kp}^{(a)}(\boldsymbol{\beta}) = \sum_{i \in \mathcal{R}_k} \tilde{w}_i \exp(\boldsymbol{\beta}' \mathbf{Z}_i(t_k)) Z_{i,p}(t_k) [\mathbf{Z}_i(t_k)]^{\otimes a}$$

where $a = 0, 1, 2$; $p = 1, \ldots, P$; and $\mathbf{Z}_i(t) = (Z_{i,1}(t), \ldots, Z_{i,p}(t))$. Then

$$\begin{aligned} \frac{\partial \mathcal{I}(\boldsymbol{\beta})}{\partial \beta_p} = \sum_{k=1}^{K} \sum_{i \in \mathcal{D}_k} \tilde{w}_i \Bigg\{ & \left[\frac{\mathbf{Q}_{kp}^{(2)}(\boldsymbol{\beta})}{S_k^{(0)}(\boldsymbol{\beta})} - \frac{\mathbf{Q}_{kp}^{(0)}(\boldsymbol{\beta})}{S_k^{(0)}(\boldsymbol{\beta})} \frac{\mathbf{S}_k^{(2)}(\boldsymbol{\beta})}{S_k^{(0)}(\boldsymbol{\beta})} \right] \\ & - \left[\frac{\mathbf{Q}_{kp}^{(1)}(\boldsymbol{\beta})}{S_k^{(0)}(\boldsymbol{\beta})} - \frac{\mathbf{Q}_{kp}^{(0)}(\boldsymbol{\beta})}{S_k^{(0)}(\boldsymbol{\beta})} \frac{\mathbf{S}_k^{(1)}(\boldsymbol{\beta})}{S_k^{(0)}(\boldsymbol{\beta})} \right] \left[\frac{\mathbf{S}_k^{(1)}(\boldsymbol{\beta})}{S_k^{(0)}(\boldsymbol{\beta})} \right]' \\ & - \left[\frac{\mathbf{S}_k^{(1)}(\boldsymbol{\beta})}{S_k^{(0)}(\boldsymbol{\beta})} \right] \left[\frac{\mathbf{Q}_{kp}^{(1)}(\boldsymbol{\beta})}{S_k^{(0)}(\boldsymbol{\beta})} - \frac{\mathbf{Q}_{kp}^{(0)}(\boldsymbol{\beta})}{S_k^{(0)}(\boldsymbol{\beta})} \frac{\mathbf{S}_k^{(1)}(\boldsymbol{\beta})}{S_k^{(0)}(\boldsymbol{\beta})} \right]' \Bigg\} \end{aligned}$$

where $p = 1, \ldots, P$.

Point estimates and Taylor linearized standard errors for the penalized likelihood are obtained from the score functions and the Hessian as described in Section 1.2. The jackknife standard errors are obtained by maximizing the penalized likelihood in every replicate sample.

Appendix 1 shows that under certain regularity conditions, the point estimators obtained by maximizing Firth's penalized likelihood are design-consistent.

2.1 Penalized likelihoods and the scale of weights

In this section, we derive a relationship between the penalized log likelihood that uses scaled weights and the penalized log likelihood that uses unscaled weights, and we demonstrate that Firth's penalized likelihood using unscaled weights does not have the invariance property.

Let $l(\boldsymbol{\beta}_{\tilde{w}};\, \tilde{w})$ be the log likelihood using weights $\tilde{w}$, and let $l(\boldsymbol{\beta}_{w};\, w)$ be the log likelihood using weights w, where $\tilde{w}_i = \alpha w_i$ for all i and $\alpha \neq 0$. The Breslow log likelihood can be written as

$$\begin{aligned}
l(\boldsymbol{\beta}_{\tilde{w}};\, \tilde{w}) &= \sum_{k=1}^{K}\left\{\boldsymbol{\beta}'_{\tilde{w}}\sum_{i\in\mathcal{D}_k}\tilde{w}_i\mathbf{Z}_i(t_k) - \left(\sum_{i\in\mathcal{D}_k}\tilde{w}_i\right)\log\sum_{i\in\mathcal{R}_k}\tilde{w}_i\exp\left(\boldsymbol{\beta}'_{\tilde{w}}\mathbf{Z}_i(t_k)\right)\right\}\\
&= \sum_{k=1}^{K}\left\{\boldsymbol{\beta}'_{\tilde{w}}\alpha\sum_{i\in\mathcal{D}_k}w_i\mathbf{Z}_i(t_k) - \left(\alpha\sum_{i\in\mathcal{D}_k}w_i\right)\log\sum_{i\in\mathcal{R}_k}\alpha w_i\exp\left(\boldsymbol{\beta}'_{\tilde{w}}\mathbf{Z}_i(t_k)\right)\right\}\\
&= \alpha\sum_{k=1}^{K}\left\{\boldsymbol{\beta}'_{\tilde{w}}\sum_{i\in\mathcal{D}_k}w_i\mathbf{Z}_i(t_k) - \left(\sum_{i\in\mathcal{D}_k}w_i\right)\log\sum_{i\in\mathcal{R}_k}w_i\exp\left(\boldsymbol{\beta}'_{\tilde{w}}\mathbf{Z}_i(t_k)\right)\right\}\\
&\quad - \sum_{k=1}^{K}\left(\alpha\sum_{i\in\mathcal{D}_k}w_i\right)\log\alpha\\
&= \alpha l(\boldsymbol{\beta}_{\tilde{w}};\, w) - \sum_{k=1}^{K}\left(\alpha\sum_{i\in\mathcal{D}_k}w_i\right)\log\alpha.
\end{aligned}$$

Because the second term on the right-hand side does not contain $\boldsymbol{\beta}$, the derivative and the Hessian of the log likelihood are only a multiplier of α and the parameter estimates and standard errors are invariant to the scale of the weights.

However, the following relation shows that the point estimates that are obtained by maximizing the penalized log likelihood are not invariant to the scale of the weights:

$$\begin{aligned}
l_p(\boldsymbol{\beta}_{\tilde{w}};\, \tilde{w}) &= l(\boldsymbol{\beta}_{\tilde{w}};\, \tilde{w}) + 0.5\log|I(\boldsymbol{\beta}_{\tilde{w}};\, \tilde{w})|\\
&= \alpha l(\boldsymbol{\beta}_{\tilde{w}};\, w) + 0.5\log|\alpha I(\boldsymbol{\beta}_{\tilde{w}};\, w)| - \sum_{k=1}^{K}\left(\alpha\sum_{i\in\mathcal{D}_k}w_i\right)\log\alpha\\
&= \alpha l(\boldsymbol{\beta}_{\tilde{w}};\, w) + 0.5\{\log|I(\boldsymbol{\beta}_{\tilde{w}};\, w)| + p\log\alpha\} - \sum_{k=1}^{K}\left(\alpha\sum_{i\in\mathcal{D}_k}w_i\right)\log\alpha\\
&= \alpha\{l(\boldsymbol{\beta}_{\tilde{w}};\, w) + 0.5\log|I(\boldsymbol{\beta}_{\tilde{w}};\, w)|\} - \sum_{k=1}^{K}\left(\alpha\sum_{i\in\mathcal{D}_k}w_i\right)\log\alpha\\
&\quad + 0.5\{p\log\alpha + (1-\alpha)\log|I(\boldsymbol{\beta}_{\tilde{w}};\, w)|\}\\
&= \alpha l_p(\boldsymbol{\beta}_{\tilde{w}};\, w) - \sum_{k=1}^{K}\left(\alpha\sum_{i\in\mathcal{D}_k}w_i\right)\log\alpha\\
&\quad + 0.5\{p\log\alpha + (1-\alpha)\log|I(\boldsymbol{\beta}_{\tilde{w}};\, w)|\}.
\end{aligned}$$

The additional term in the right hand side of the preceding equation involves the regression parameters. Thus the point estimates and the standard errors are not invariant to the scale of the weights.

By construction, point estimates that use the penalized log likelihood and the scaled weights are invariant to the scale of the weights.

2.2 Example that uses scaled weights

Consider the myeloma study described in Section 1.1. We refit the same proportional hazards regression model using LogBUN, HGB, and Contrived as explanatory variables, but now we use scaled weights in constructing Firth's penalized likelihood.

Table 2.1 displays point estimates and standard errors from Firth's penalized likelihood using scaled weights and the Taylor linearized variance estimator. These statistics are invariant to the scale of the weights.

Table 2.1
Parameter estimates and their standard errors using the Taylor linearized method with the Firth correction and scaled weights

	Weight $w1$		Weight $w3$		Weight $w5$	
	Estimate	**Std. Err.**	**Estimate**	**Std. Err.**	**Estimate**	**Std. Err.**
LogBUN	1.722	0.564	1.722	0.564	1.722	0.564
HGB	-0.112	0.064	-0.112	0.064	-0.112	0.064
Contrived	3.815	0.458	3.815	0.458	3.815	0.458

Standard errors using jackknife replicates are also invariant to the scale of the weights. For replicate variance estimation methods, every set of replicate weights must be scaled using the same scaling factor that is used to scale the full sample weights. Table 2.2 displays point estimates and standard errors from Firth's penalized likelihood using scaled weights and the jackknife replicate variance estimator.

Table 2.2
Parameter estimates and their standard errors using jackknife replicates with the Firth correction and scaled weights

	Weight $w1$		Weight $w3$		Weight $w5$	
	Estimate	**Std. Err.**	**Estimate**	**Std. Err.**	**Estimate**	**Std. Err.**
LogBUN	1.722	0.653	1.722	0.653	1.722	0.653
HGB	-0.112	0.074	-0.112	0.074	-0.112	0.074
Contrived	3.815	0.642	3.815	0.642	3.815	0.642

Estimates from the penalized log likelihood using the scaled weights also have the closeness property. The ratios of jackknife standard errors to Taylor linearized standard errors are 1.16, 1.17, and 1.40 for all three sets of weights for the variables LogBUN, HGB, and Contrived, respectively (Tables 2.1 and 2.2).

3 Applications in complex surveys

Data from complex surveys frequently contain unequal weights, strata, and clusters. It is recommended that the weights and other design features be used in the analysis stage. Weighted data provide a better representation of the study population than unweighted data. In this section, we compare the scaled and unscaled weights to estimate proportional hazards regression coefficients through a simulation study and apply the Firth penalized likelihood using the scaled weights to estimate survival times from a data set from NHEFS.

3.1 A simulation study

We performed a small simulation study to compare biases in parameter estimates and standard errors for scaled and unscaled weights using Firth's penalized likelihood. We used two sampling methods to select samples from a fixed finite population: a simple random sample without replacement (SRS) in which each observation unit gets an equal weight; and a probability proportional to size (PPS) without replacement sample in which the sampling weight for an observation unit depends on the value of a size measure associated with the hazard function for the unit. For the purpose of finite population inference, we treat the estimated proportional hazards regression parameters in the finite population as the "true" parameter values. Biases are measured from these true values.

Finite populations of size 10,000 are generated as follows:

- $\mathrm{Z}_1, \mathrm{Z}_2, \ldots, \mathrm{Z}_{10} \sim \text{Bernoulli}(0.75)$,
- $h = \exp(-0.69Z_1 - 0.69Z_2 - \cdots - 0.69Z_{10})$,
- $u \sim \text{uniform}(0, 1)$,
- $t = \log(u)/h$,
- $c \sim \text{Bernoulli}(v)$,
- $m \sim \text{uniform}(10h, 10h + 0.1)$

where h is the hazard function, t is the survival time, c is a censoring indicator, and m is a size measure for each unit. Six finite populations are generated by using different censoring values ($v = 0.1$, 0.3, 0.5, 0.7, 0.8, 0.9). See Bender, Augustin and Blettner (2005) for methods of generating survival times. Ten regressors $(\mathrm{Z}_1, \mathrm{Z}_2, \ldots, \mathrm{Z}_{10})$ are generated using Bernoulli distributions to create monotone likelihoods, especially when the sample size is small and the censoring rate is high.

Samples are selected from each finite population by using two sampling methods: simple random samples without replacement; and probability proportional to size samples without replacement, where the variable m is used as the size measure. Four sample sizes are used for each sampling method: 50, 100, 500, and 1,000. Sampling weights for all units for SRS depend only on the sample size, but the sampling weight for a unit for PPS depends on both the sample size and the observed value of the variable m for that corresponding unit. To ensure the same distribution of the censored observations in the sampled data as in the population, samples are selected independently from censored and uncensored units in the population.

Finally, the regression parameters from the proportional hazards regression model

$$\lambda(t, \mathbf{Z}) = \lambda_{\mathbf{0}}(\mathbf{t}) \exp(\beta_{\mathbf{1}}\mathbf{Z}_{\mathbf{1}} + \beta_{\mathbf{2}}\mathbf{Z}_{\mathbf{2}} + \cdots + \beta_{\mathbf{10}}\mathbf{Z}_{\mathbf{10}})$$

are estimated from each sampled data set, where $\lambda_0(t)$ is the baseline hazard, t is the survival time, and c is the censored indicator. The regression parameters are estimated by maximizing the weighted Firth penalized likelihood. Note that the unpenalized likelihood does not converge in most cases because of the monotonicity of the likelihood in the simulated data. When the likelihood is not monotone, we found that the point estimates obtained by using the penalized likelihood are very close to the point estimates obtained by using the unpenalized likelihood. Heinze and Schemper (2001) reported similar findings for unweighted data.

We compare relative biases in point estimates and standard errors using the jackknife method for scaled and unscaled weights. The relative biases (RBs) are defined below (Sitter, 1992).

Let $\hat{\beta}_s$ be the point estimate and $\hat{v}_s$ be the variance estimate for one component of $\boldsymbol{\beta}$ from data set s. Define the following:

Relative bias for point estimates, $\hat{\beta}$,

$$\mathrm{RB}(\hat{\beta}) = S^{-1}\sum_{s=1}^{S}\frac{|(\hat{\beta}_s - \beta_T)|}{|\beta_T|}.$$

Relative bias for variance estimates, $\hat{v}$

$$\mathrm{RB}(\hat{v}) = S^{-1}\sum_{s=1}^{S}\frac{|(\hat{v}_s - \mathrm{MSE}_T)|}{\mathrm{MSE}_T}$$

where the true MSE is

$$\mathrm{MSE}_T(\hat{\beta}) = S^{-1}\sum_{s}(\hat{\beta}_s - \beta_T)^2$$

and β_T is the "true" parameter value obtained by fitting the proportional hazards regression model using all units in the finite population. The ratio of RBs is defined as the ratio of the RB using the unscaled weights to the RB using the scaled weights.

The median of ratios of RBs over 5,000 repetitions is displayed in this section. We report the median because there are some "bad" samples in which convergences are questionable even with the Firth correction. These "bad" samples produce few estimates with very large biases. Because of these large biases, the mean of the ratio of RBs is a more unstable statistic than the median. Without the "bad" replicates, the mean and medians are very close. We also noticed that the penalized log likelihood using the unscaled weights produces more of these "false" convergences.

Results for all regressors $\mathrm{Z}_1, \mathrm{Z}_2, \ldots, \mathrm{Z}_{10}$ are similar. For simplicity, we display results for only two regressors, Z_3 and Z_8.

Ratios of RBs in parameter estimates for unscaled and scaled weights for the variables Z_3 and Z_8 are displayed in Figures 3.1, 3.2, 3.3, and 3.4. For small sample sizes and a large number of censored observations, RBs using scaled weights are much smaller than RBs using unscaled weights. For large sample sizes, RBs from both weights are similar primarily because the Firth option is not necessary, since the convergence is not an issue with large data sets.

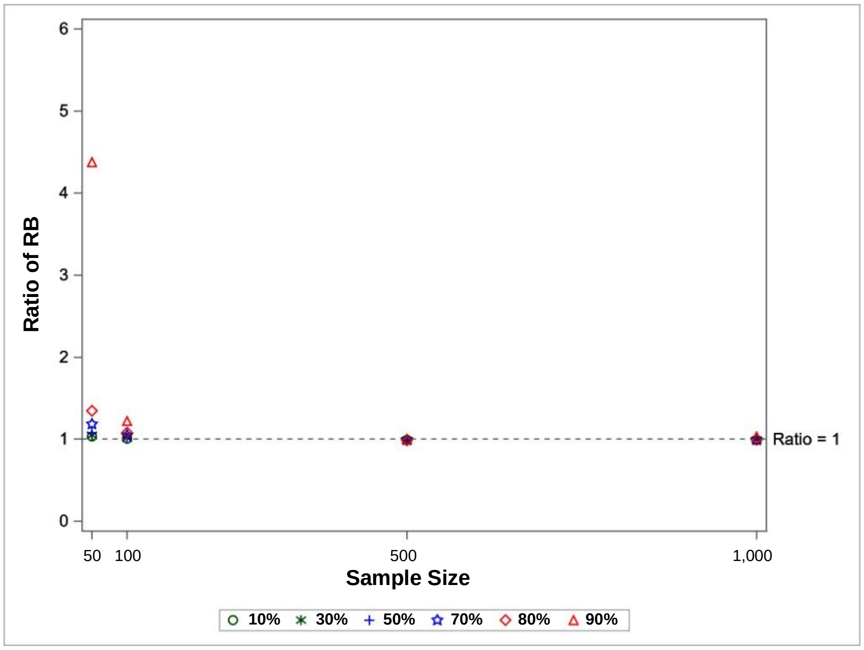

Figure 3.1 Ratio of relative biases in parameter estimates for SRS samples for Z3.

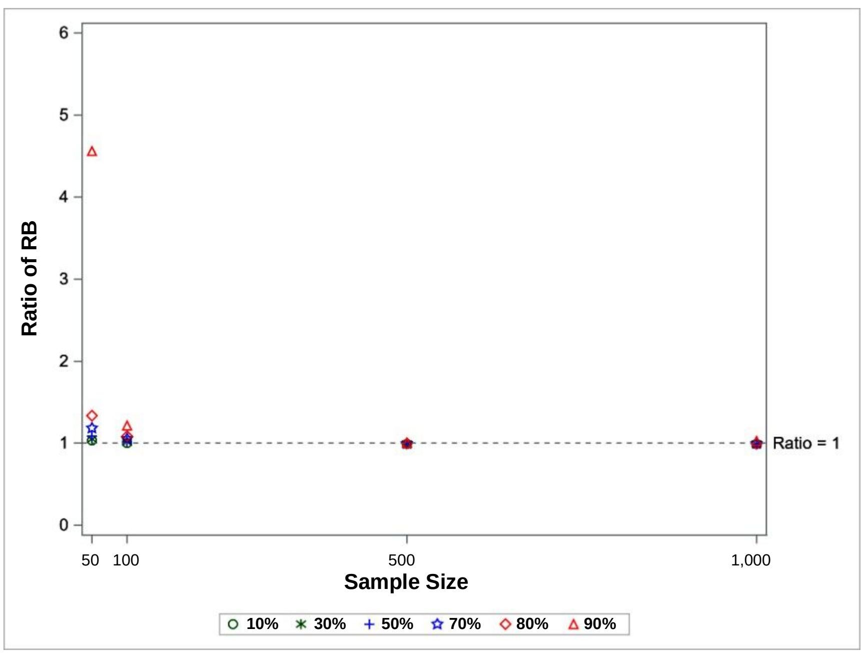

Figure 3.2 Ratio of relative biases in parameter estimates for SRS samples for Z8.

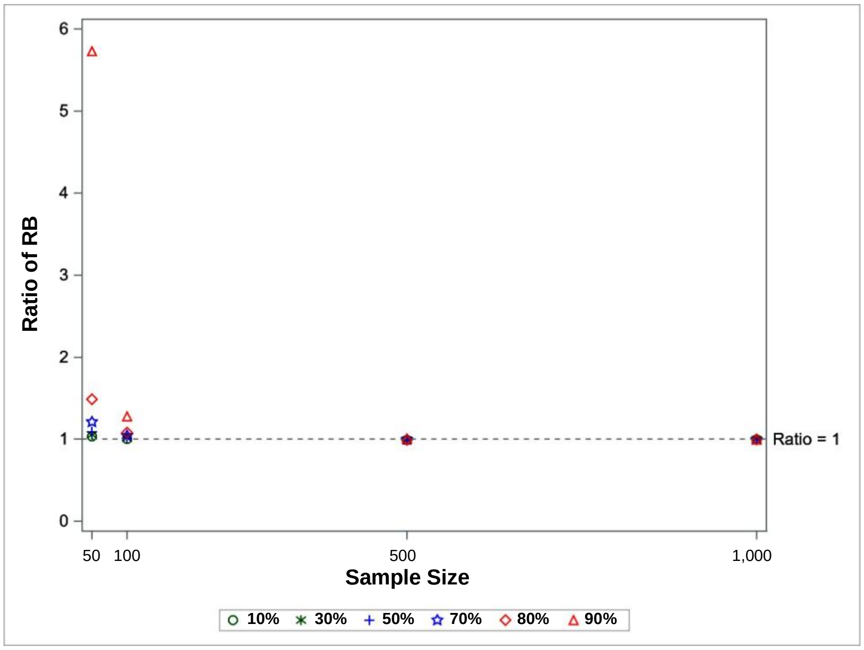

Figure 3.3 Ratio of relative biases in parameter estimates for PPS samples for Z3.

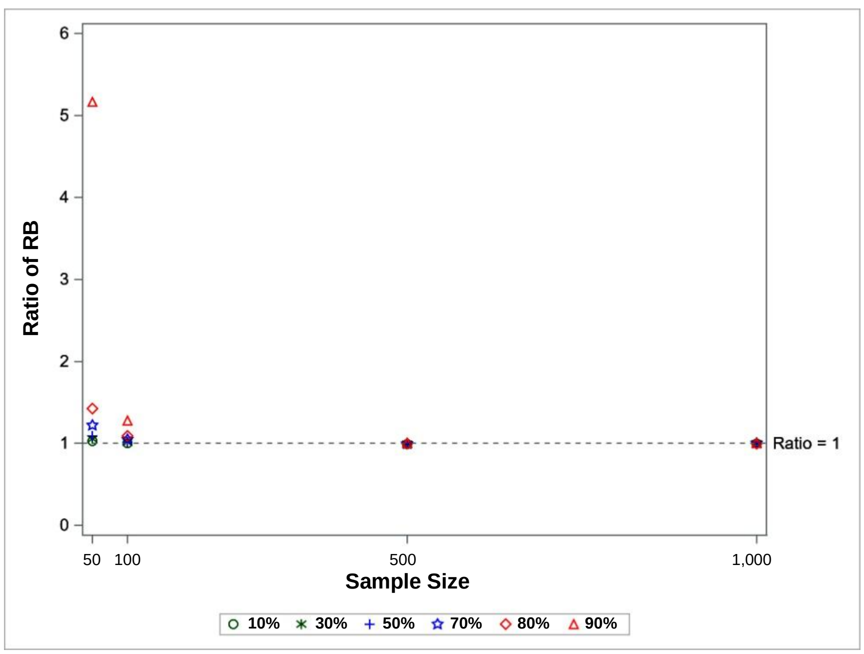

Figure 3.4 Ratio of relative biases in parameter estimates for PPS samples for Z8.

Ratios of RBs in standard errors for unscaled and scaled weights for the variables Z_3 and Z_8 are displayed in Figures 3.5, 3.6, 3.7, and 3.8. RBs for standard errors follow the same trend as RBs for point estimates. However, RBs for standard errors are higher than RBs for point estimates. For small sample sizes and a large number of censored observations, RBs using scaled weights are much smaller than RBs using unscaled weights. For large sample sizes, RBs from both scaled and unscaled weights are similar primarily because the Firth option is not necessary, since the convergence is not an issue with large data sets.

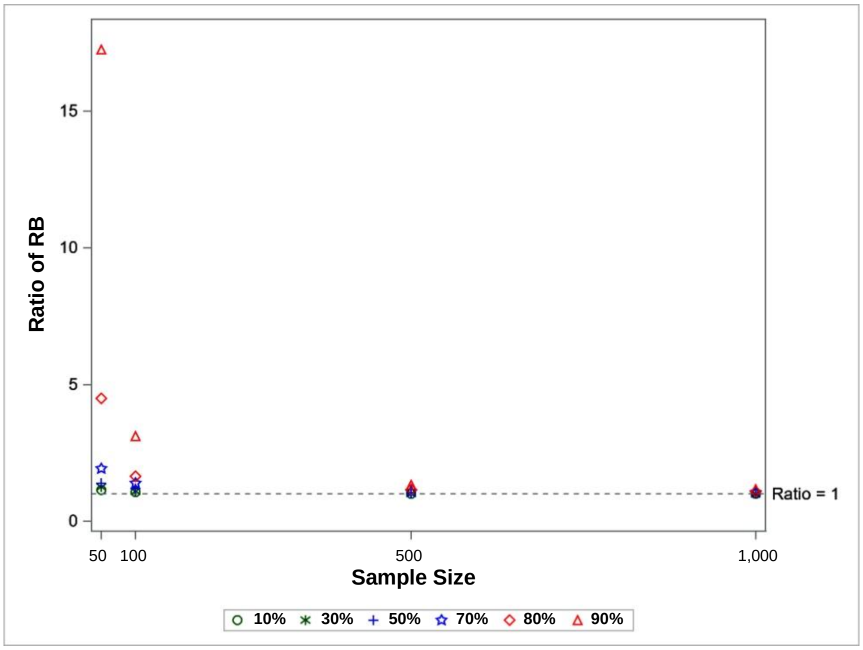

Figure 3.5 Ratio of relative biases in standard errors for SRS samples for Z3.

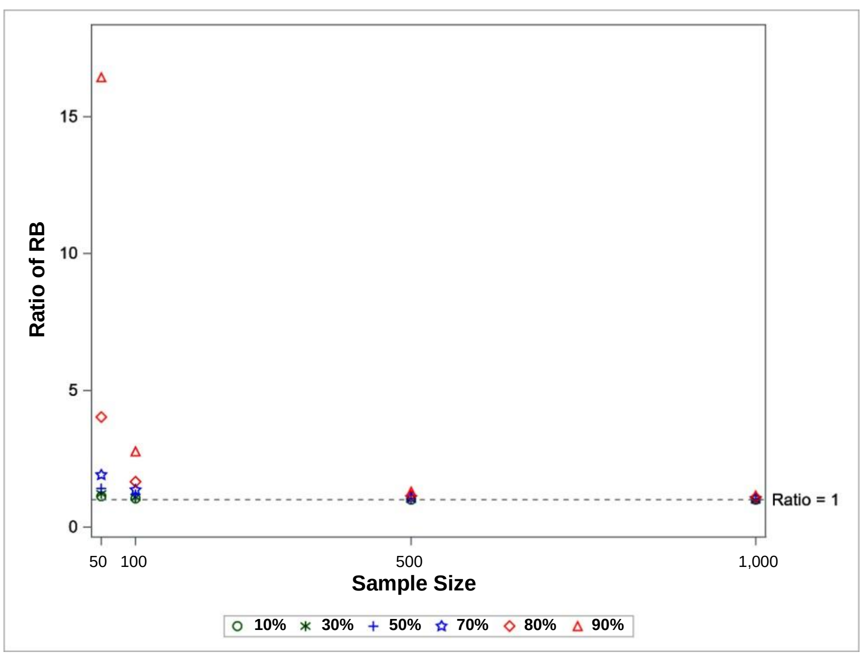

Figure 3.6 Ratio of relative biases in standard errors for SRS samples for Z8.

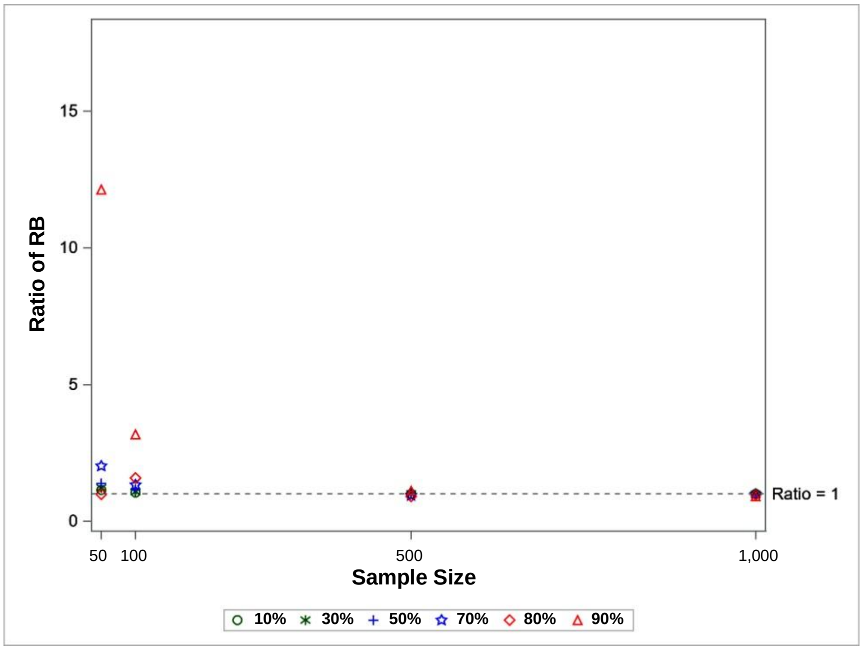

Figure 3.7 Ratio of relative biases in standard errors for PPS samples for Z3.

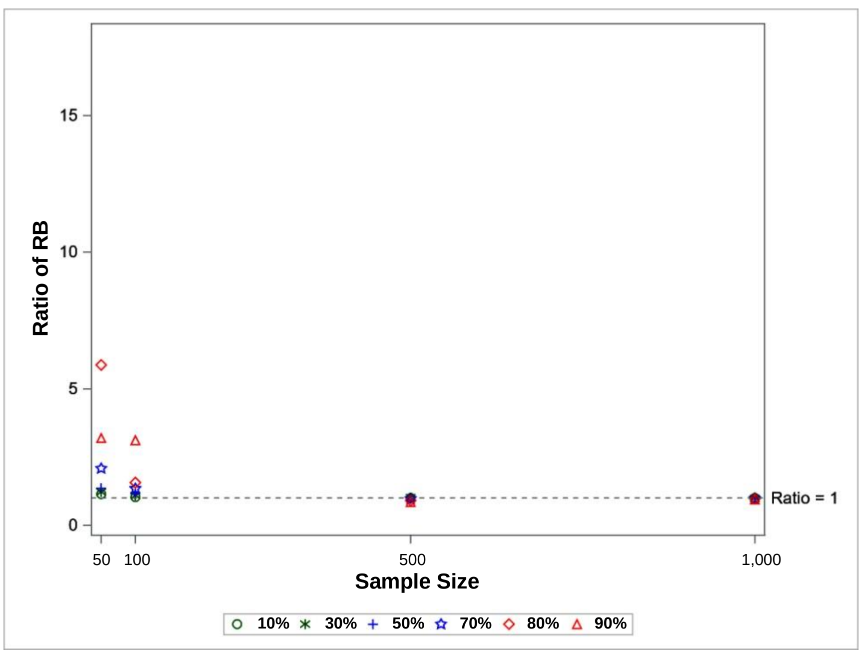

Figure 3.8 Ratio of relative biases in standard errors for PPS samples for Z8.

Table 3.1 displays the first quartile, median, and third quartile for ratio of RBs in point estimates and standard errors for sample size 50. Results for variable Z_3 for 10% and 90% censoring are reported in the table. We observed that for all variables, the first and third quartiles for ratio of RBs do not contain 1 when the sample size is small and the percentage of censoring is high. However, as expected, for large samples and a small number of censored observations, the difference in RBs between the scaled and unscaled weights is small.

Table 3.1
Ratio of RBs in point estimates and standard errors for sample size 50 (variable Z_3)

Design	Ratio of RBs in Point Estimates					
	90% Censored			10% Censored		
	First Quartile	Median	Third Quartile	First Quartile	Median	Third Quartile
SRS	1.81	4.38	7.37	1.00	1.03	1.06
PPS	3.36	5.73	11.54	0.99	1.03	1.08
	Ratio of RBs in Standard Errors					
Design	**90% Censored**			**10% Censored**		
	First Quartile	**Median**	**Third Quartile**	**First Quartile**	**Median**	**Third Quartile**
SRS	9.03	17.26	40.87	1.03	1.15	1.33
PPS	5.57	12.13	29.92	1.00	1.15	1.33

3.2 An application using NHEFS

We studied the time to get a heart attack and its relation to blood cholesterol and smoking using a data set from NHEFS.

The NHEFS is a national longitudinal survey in the United States that is used to determine the relationships between clinical, nutritional, and behavioral factors; to determine hospital utilizations; and to monitor changes in risk factors for an initial cohort that represents the NHANES I population. A cohort of size 14,407 was selected for the NHEFS. Vital and tracing status data, interview data, health care facility stay data, and mortality data from 1987 are available for public use. For more information about the survey and the data sets used in this section, see the Centers for Disease Control and Prevention’s website (https://www.cdc.gov/).

We used 4,673 observations from 1987 NHEFS public-use interview data to study the occurrence of first heart attack for the 1987 survey population and its relation to blood cholesterol and smoking. The following variables are used:

- Stratum, the stratum identification.
- ObservationWeight, the sampling weight associated with each observation unit.
- PSU, the primary sampling unit identification.
- Age, the event-time variable, defined as follows:
    - age of the subject when the first heart attack was reported for subjects who reported a heart attack,
    - age of the subject as reported in the interview for subjects who never reported a heart attack.
- HeartAttack, the heart attack indicator (1 = heart attack reported).

- Income, household income standardized to mean zero.
- HighBloodChol, the indicator that a subject has a high or low blood cholesterol level.
- Smoker, subject's smoking habit (1 = current, 2 = former, –1 = non-smoker).
- Race, the race of the subject (1 = black, 2 = white, 3 = other).
- Gender, the gender of the subject.

The SURVEYPHREG procedure in SAS/STAT (Mukhopadhyay, 2010) is used to fit a proportional hazards regression model for age on income, blood cholesterol, smoking habit, race, gender, and race and gender interaction. Heart attack is used as the censored indicator. Observation weights range from 1,164 to 121,040 with mean 16,036.51, median 12,321, and coefficient of variation 74.35. The subjects are divided into 644 clusters and 35 strata.

PROC SURVEYPHREG is used in this section instead of PROC PHREG because the NHEFS uses complex survey design involving stratification, clustering, and unequal weights. PROC SURVEYPHREG supports STRATA, CLUSTER, and WEIGHT statements to account for stratification, clustering, and unequal weights, respectively. In addition, PROC SURVEYPHREG supports both Taylor series linearization and jackknife variance estimation methods for survey data (Mukhopadhyay, 2010). We used the jackknife variance estimation method for this study. SAS statements to fit this model are given in Appendix 2.

The 4,673 subjects in the sample represent almost 74.9 million individuals in the 1987 study population. Among all the subjects, 213 subjects reported at least one heart attack, and the other 4,460 subjects are considered to be censored. The 213 event observations in the sample represent an estimated 3.2 million population units, and the 4,460 censored observations in the sample represent an estimated 71.7 million population units. There are 95.44% observations in the sample that have not reported a heart attack which estimates 95.68% individuals in the population (Table 3.2) without a heart attack.

Table 3.2
Number of censored and uncensored observations and their sum of weights

	Total	Event	Censored	Percent Censored
Number of Observations	4,673	213	4,460	95.44
Sum of Weights	74,938,614	3,239,653	71,698,961	95.68

Without the Firth penalty, the Newton-Raphson optimization converges by satisfying the relative gradient convergence criterion (GCONV = 1E–8), but coefficients for the variables Smoker and Race do not converge. The coefficients for Smoker = 2 are 7.47, 10.87, and 11.83; and the coefficients for Race = 1 are 7.55, 10.95, and 11.17 in the last three iterations, respectively. This phenomenon is very common when you have a monotone likelihood (see Table 1.1). Among 644 replicate samples

(= 644 PSUs), monotone likelihood is observed in 542 replicates. Firth's penalized likelihood is a good alternative when you encounter monotone likelihoods.

We use the FIRTH option in PROC SURVEYPHREG (see "The SURVEYPHREG Procedure" in SAS Institute Inc. (2018)) to maximize Firth's penalized likelihood. The FIRTH option in PROC SURVEYPHREG uses the scaled weights. The penalized likelihood optimization converges with GCONV = 1E–8, along with reasonable convergence in all coefficients. Convergence is also achieved in all 644 replicate samples with the Firth penalty.

Table 3.3 displays the estimated hazards ratios along with their 95% Wald confidence intervals for blood cholesterol levels and smoking. In the 1987 study population, the estimated hazard of having a heart attack for a subject with low blood cholesterol is 0.6 times the estimated hazard of having a heart attack for a subject with high blood cholesterol. Because the 95% confidence interval does not contain 1, it is reasonable to conclude that the hazard of having a heart attack for a subject with low blood cholesterol is significantly lower than the hazard of having a heart attack for a subject with high blood cholesterol after adjusting for smoking, race, and other regressors in the 1987 study population.

The estimated hazard ratios for nonsmokers, current smokers, and former smokers are 0.59, 0.64, and 1.1, respectively. The estimated hazard for nonsmokers to have a heart attack is lower than the estimated hazard for current or former smokers. However, we do not have enough evidence to conclude that hazard ratios for smoking are significantly different at the 95% level after adjusting for blood cholesterol, race, and other regressors in the 1987 study population.

Table 3.3
Hazard ratios for blood cholesterol and smoking, and their 95% Wald confidence intervals

		Confidence Limit	
	Point Estimate	**Lower**	**Upper**
HighBloodChol 0 vs 1	0.643	0.469	0.882
Smoker -1 vs 1	0.590	0.259	1.345
Smoker -1 vs 2	0.641	0.361	1.140
Smoker 1 vs 2	1.087	0.359	3.290

4 Summary

Firth's penalized likelihood is useful for obtaining maximum likelihood estimates from a monotone likelihood from proportional hazards regression models. We proposed a weight scaling method and demonstrated that Firth's penalized likelihood using the scaled weights have some desirable properties for complex surveys. A simulation study shows that estimated biases in point estimates and standard errors using the scaled weights are lower than estimated biases using the unscaled weights. Although Firth's penalized likelihood produces "good" estimates in most simulated data sets, there are a few data sets in which the Firth penalty failed to produce "good" convergences. The Firth penalized likelihood that uses scaled weights successfully corrected for a monotone likelihood when we estimated hazard rates for heart attacks using a data set from the NHEFS. Although the numeric results are quite encouraging, further research is needed to derive asymptotic distributions of the estimators obtained by using Firth’s penalized likelihood.

We recommend the unpenalized likelihood when convergence is not an issue, but we recommend Firth’s penalized likelihood using the scaled weights when a monotone likelihood is encountered in fitting proportional hazards regression models for complex surveys.

Acknowledgements

I am grateful to Ying So, Randy Tobias, and Ed Huddleston at SAS Institute Inc. for their valuable assistance in the preparation of this article. I would also like to thank the two anonymous referees and the associate editor for their constructive suggestions.

Appendix 1

Consistency of the Firth penalized likelihood estimator

The estimators in Section 2 are defined as the solution to a system of equations that are constructed by using the score functions from proportional hazards regression models. In this appendix, we show that under certain regularity conditions these estimators are design consistent. Properties of estimators that are solutions to a set of estimating equations are well studied in the survey literature. For example, see Binder (1983), Godambe and Thompson (1986), and Fuller (2009, Section 1.3.4).

However, the estimating equations for proportional hazards regression models are more complex than the estimating equations for generalized linear models because the score functions involve weighted sums over the sampled units. Binder (1992) and Lin (2000) showed that the estimators obtained by solving the estimating equations for proportional hazards regression models are consistent. In this appendix, we follow arguments similar to those of Lin (2000) and Andersen and Gill (1982).

Several technical assumptions are necessary to show that the point estimates are consistent. We need assumptions about the estimating equations, the finite population, and the sample design – to whit:

- The functions defining the estimating equations should be smooth and convex.
- The finite population should be such that the moments for population quantities that are used in defining the estimating equations exists.
- The sample design should be such that the Narain-Horvitz-Thompson (NHT) estimators (Rao, 2005) for the population totals are well behaved.

All these assumptions are common in the sample survey literature; for example, see Fuller (2009). The score functions for proportional hazards regression models involve ratios of means of exponential functions that are infinitely differentiable.

Let $\mathcal{U}_N$ and $\mathcal{F}_N$ denote, respectively, the index set and values for the N^{th} finite population in a sequence of populations indexed by $N,$ and let A_N be a sample of size n from $\mathcal{U}_N$. To study large sample properties for sample-based estimators, we assume sequences of population and samples such that $N \to \infty$ and $(N - n) \to \infty,$ keeping the sampling fraction, $\frac{n}{N},$ fixed.

Assume $\mathcal{F}_N = \{(t_i, \Delta_i, \mathrm{Z}_i(\cdot))\}_{i=1}^N$ is an independent random sample of size N from the joint distribution of $(T, \Delta, \mathrm{Z}(\cdot)),$ where t is the failure time or the censoring time, whichever is less; $\Delta = 1$ if the failure time is less than the censoring time and 0 otherwise; and $Z(\cdot)$ is a vector of possibly time-varying explanatory variables.

Let $\boldsymbol{\beta}$ be a set of regression parameters for the superpopulation that is defined by the joint distribution of $(T, \Delta, Z(\cdot))$. Let $\boldsymbol{\beta}_N$ be a set of finite population parameters obtained by solving the estimating equations when all N units in the population are observed, and let $\hat{\boldsymbol{\beta}}_N$ be an estimator of $\boldsymbol{\beta}_N$ that is obtained by solving the weighted estimating equations by using only the sampled units. Our objective is to show that $\hat{\boldsymbol{\beta}}_N$ approaches $\boldsymbol{\beta}_N$ and that they both approach $\boldsymbol{\beta}$ as the sample size and population size increase.

Consider the estimating equations that correspond to Firth's penalized likelihood described in Section 2. For simplicity, we write these equations when there are no tied events. To further simplify notation, we write each component of the estimating equations separately. The finite population parameters, $\boldsymbol{\beta}_N,$ are a solution to the penalized partial likelihood score function, $U_N(\boldsymbol{\beta}) = (U_{N,1}(\boldsymbol{\beta}), \ldots, U_{N,P}(\boldsymbol{\beta}))',$ where

$$
\begin{aligned}
U_{N,p}(\boldsymbol{\beta}) = N^{-1} \sum_{i \in \mathcal{U}_N} \Delta_i \Bigg[\mathbf{Z}_i(t_i) - \frac{S_p^{(1)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)} \\
+ 0.5 \operatorname{tr} \Bigg(\Bigg[N^{-1} \sum_{i \in \mathcal{U}_N} \Delta_i \Bigg\{ \frac{S^{(2)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)} - \left(\frac{\mathbf{S}^{(1)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)} \right) \left(\frac{\mathbf{S}^{(1)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)} \right)' \Bigg\} \Bigg]^{-1} \\
\Bigg\{ \left(\frac{Q_p^{(2)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)} - \frac{Q_p^{(0)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)} \frac{S^{(2)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)} \right) \\
- \left(\frac{\mathbf{Q}_p^{(1)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)} - \frac{Q_p^{(0)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)} \frac{\mathbf{S}^{(1)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)} \right) \left(\frac{\mathbf{S}^{(1)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)} \right)' \\
- \left(\frac{\mathbf{S}^{(1)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)} \right) \left(\frac{\mathbf{Q}_p^{(1)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)} - \frac{Q_p^{(0)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)} \frac{\mathbf{S}^{(1)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)} \right)' \Bigg\} \Bigg) \Bigg]
\end{aligned}
$$

where

$$S^{(a)}(\boldsymbol{\beta}, t) = N^{-1} \sum_{i \in \mathcal{U}_N} I(t_i \geq t) \exp(\boldsymbol{\beta}' \mathbf{Z}_i(t)) [\mathbf{Z}_i(t)]^{\otimes a}$$

$$Q_p^{(a)}(\boldsymbol{\beta}, t) = N^{-1} \sum_{i \in \mathcal{U}_N} I(t_i \geq t) \exp(\boldsymbol{\beta}' \mathbf{Z}_i(t)) Z_{i,p}(t) [\mathbf{Z}_i(t)]^{\otimes a}$$

and where $a = 0, 1, 2$; $Z_i(t) = (Z_{i,1}(t), \ldots, Z_{i,p}(t))'$; $\mathbf{S}^{(1)}(\boldsymbol{\beta}, t) = (S_1^{(1)}(\boldsymbol{\beta}, t), \ldots, S_P^{(1)}(\boldsymbol{\beta}, t))'$; $\operatorname{tr}(\cdot)$ denotes the trace of a matrix; $I(\cdot)$ denotes the indicator function; $p = 1, 2, \ldots, P$; and P is the number of regression parameters. Note that $S^{(a)}(\boldsymbol{\beta}, t)$ and $Q_p^{(a)}(\boldsymbol{\beta}, t)$ depend on N, although the notation does not reflect this for reasons of simplicity.

In defining the score function for the penalized likelihood, we assume that the information matrix for the finite population, $I_N(\boldsymbol{\beta}, t)$, is always positive definite.

However, in any realistic situation, not all units in the finite population are available. Let a sample A_N be selected by using a probability design that assigns a nonzero selection probability, π_i, to every unit in the population. Let $w_i = \pi_i^{-1}$ be the design weight. A sample-based estimator, $\hat{\boldsymbol{\beta}}_N$, is obtained by solving the estimated penalized partial likelihood score equations. Assuming that N is known, a sample-based estimator for $U_{N,p}(\boldsymbol{\beta})$ is

$$\begin{aligned}
\hat{U}_{N,p}(\boldsymbol{\beta}) = N^{-1} \sum_{i \in A_N} w_i \Delta_i \Bigg[& \mathbf{Z}_i(t_i) - \frac{\hat{S}_p^{(1)}(\boldsymbol{\beta}, t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta}, t_i)} \\
& + 0.5 \operatorname{tr}\Bigg(\Bigg[N^{-1} \sum_{i \in A_N} w_i \Delta_i \Bigg\{ \frac{\hat{S}^{(2)}(\boldsymbol{\beta}, t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta}, t_i)} - \left(\frac{\hat{\mathbf{S}}^{(1)}(\boldsymbol{\beta}, t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta}, t_i)} \right) \left(\frac{\hat{\mathbf{S}}^{(1)}(\boldsymbol{\beta}, t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta}, t_i)} \right)' \Bigg\} \Bigg]^{-1} \\
& \Bigg\{ \left(\frac{\hat{Q}_p^{(2)}(\boldsymbol{\beta}, t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta}, t_i)} - \frac{\hat{Q}_p^{(0)}(\boldsymbol{\beta}, t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta}, t_i)} \frac{\hat{S}^{(2)}(\boldsymbol{\beta}, t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta}, t_i)} \right) \\
& - \left(\frac{\hat{\mathbf{Q}}_p^{(1)}(\boldsymbol{\beta}, t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta}, t_i)} - \frac{\hat{Q}_p^{(0)}(\boldsymbol{\beta}, t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta}, t_i)} \frac{\hat{\mathbf{S}}^{(1)}(\boldsymbol{\beta}, t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta}, t_i)} \right) \left(\frac{\hat{\mathbf{S}}^{(1)}(\boldsymbol{\beta}, t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta}, t_i)} \right)' \\
& - \left(\frac{\hat{\mathbf{S}}^{(1)}(\boldsymbol{\beta}, t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta}, t_i)} \right) \left(\frac{\hat{\mathbf{Q}}_p^{(1)}(\boldsymbol{\beta}, t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta}, t_i)} - \frac{\hat{Q}_p^{(0)}(\boldsymbol{\beta}, t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta}, t_i)} \frac{\hat{\mathbf{S}}^{(1)}(\boldsymbol{\beta}, t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta}, t_i)} \right)' \Bigg\} \Bigg) \Bigg]
\end{aligned}$$

where

$$\hat{S}^{(a)}(\boldsymbol{\beta}, t) = N^{-1} \sum_{i \in A_N} w_i I(t_i \geq t) \exp(\boldsymbol{\beta}' \mathbf{Z}_i(t)) [\mathbf{Z}_i(t)]^{\otimes a}$$

$$\hat{Q}_p^{(a)}(\boldsymbol{\beta}, t) = N^{-1}\sum_{i \in A_N} w_i I(t_i \geq t)\exp(\boldsymbol{\beta}'\mathbf{Z}_i(t)) Z_{i,\,p}(t)[\mathbf{Z}_i(t)]^{\otimes a}$$

are the NHT estimators for $S^{(a)}(\boldsymbol{\beta}, t)$ and $Q_p^{(a)}(\boldsymbol{\beta}, t)$, respectively.

Because $\hat{S}^{(a)}(\boldsymbol{\beta}, t)$ and $\hat{Q}_p^{(a)}(\boldsymbol{\beta}, t)$ use weighted sums over sampled units, we need techniques defined in Lin (2000) to study large sample properties of these estimators. Define $G_i(t) = \Delta_i I(t_i \leq t)$, $G(t) = N^{-1}\sum_{i \in \mathcal{U}_N} G_i(t)$, and $\hat{G}(t) = \sum_{i \in A_N} w_i G_i(t)$. Then the finite population score functions can be written using stochastic integration,

$$\begin{aligned} U_{N,\,p}(\boldsymbol{\beta}) = N^{-1}\sum_{i \in \mathcal{U}_N}\int_0^\infty \Bigg[& \mathbf{Z}_i(t_i) - \frac{S_p^{(1)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)} \\ & + 0.5\mathrm{tr}\Bigg(\Bigg[N^{-1}\sum_{i \in A_N}\int_0^\infty \Bigg\{\frac{S^{(2)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)} - \left(\frac{\mathbf{S}^{(1)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)}\right)\left(\frac{\mathbf{S}^{(1)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)}\right)'\Bigg\} dG(t_i)\Bigg]^{-1} \\ & \Bigg\{\left(\frac{Q_p^{(2)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)} - \frac{Q_p^{(0)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)}\frac{S^{(2)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)}\right) \\ & - \left(\frac{\mathbf{Q}_p^{(1)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)} - \frac{Q_p^{(0)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)}\frac{\mathbf{S}^{(1)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)}\right)\left(\frac{\mathbf{S}^{(1)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)}\right)' \\ & - \left(\frac{\mathbf{S}^{(1)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)}\right)\left(\frac{\mathbf{Q}_p^{(1)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)} - \frac{Q_p^{(0)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)}\frac{\mathbf{S}^{(1)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)}\right)'\Bigg\}\Bigg)\Bigg] dG_i(t) \end{aligned}$$

and the sample-based score functions are

$$\begin{aligned} \hat{U}_{N,p}(\boldsymbol{\beta}) = N^{-1}\sum_{i \in \mathcal{U}_N}\int_0^\infty I(i \in A_n) w_i \Bigg[& \mathbf{Z}_i(t_i) - \frac{\hat{S}_p^{(1)}(\boldsymbol{\beta}, t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta}, t_i)} \\ & + 0.5\mathrm{tr}\Bigg(\Bigg[N^{-1}\sum_{i \in \mathcal{U}_N}\int_0^\infty I(i \in A_N) w_i \\ & \Bigg\{\frac{\hat{S}^{(2)}(\boldsymbol{\beta}, t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta}, t_i)} - \left(\frac{\hat{\mathbf{S}}^{(1)}(\boldsymbol{\beta}, t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta}, t_i)}\right)\left(\frac{\hat{\mathbf{S}}^{(1)}(\boldsymbol{\beta}, t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta}, t_i)}\right)'\Bigg\} dG_i(t)\Bigg]^{-1} \\ & \Bigg\{\left(\frac{\hat{Q}_p^{(2)}(\boldsymbol{\beta}, t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta}, t_i)} - \frac{\hat{Q}_p^{(0)}(\boldsymbol{\beta}, t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta}, t_i)}\frac{\hat{S}^{(2)}(\boldsymbol{\beta}, t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta}, t_i)}\right) \\ & - \left(\frac{\hat{\mathbf{Q}}_p^{(1)}(\boldsymbol{\beta}, t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta}, t_i)} - \frac{\hat{Q}_p^{(0)}(\boldsymbol{\beta}, t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta}, t_i)}\frac{\hat{\mathbf{S}}^{(1)}(\boldsymbol{\beta}, t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta}, t_i)}\right)\left(\frac{\hat{\mathbf{S}}^{(1)}(\boldsymbol{\beta}, t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta}, t_i)}\right)' \\ & - \left(\frac{\hat{\mathbf{S}}^{(1)}(\boldsymbol{\beta}, t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta}, t_i)}\right)\left(\frac{\hat{\mathbf{Q}}_p^{(1)}(\boldsymbol{\beta}, t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta}, t_i)} - \frac{\hat{Q}_p^{(0)}(\boldsymbol{\beta}, t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta}, t_i)}\frac{\hat{\mathbf{S}}^{(1)}(\boldsymbol{\beta}, t_i)}{\hat{S}^{(0)}(\boldsymbol{\beta}, t_i)}\right)'\Bigg\}\Bigg)\Bigg] dG_i(t). \end{aligned}$$

Note that the quantities $S^{(a)}$ and $Q_p^{(a)}$ are simply means over finite population quantities. Define the limits of these means as follows:

$$\begin{aligned}
\mathbf{s}^{(a)}(\boldsymbol{\beta}, t) &:= \lim_{N\to\infty} \mathbf{S}^{(a)}(\boldsymbol{\beta}, t) \\
&= \lim_{N\to\infty} N^{-1} \sum_{i\in\mathcal{U}_N} I(t_i \geq t) \exp(\boldsymbol{\beta}'\mathbf{Z}_i(t))[\mathbf{Z}_i(t)]^{\otimes a} \\
q_p^{(a)}(\boldsymbol{\beta}, t) &:= Q_p^{(a)}(\boldsymbol{\beta}, t) \\
&= N^{-1} \sum_{i\in\mathcal{U}_N} I(t_i \geq t) \exp(\boldsymbol{\beta}'\mathbf{Z}_i(t)) Z_{i,p}(t)[\mathbf{Z}_i(t)]^{\otimes a} \\
g(t) &:= \lim_{N\to\infty} N^{-1} \sum_{i\in\mathcal{U}_N} G_i(t) \\
\alpha &:= \lim_{N\to\infty} N^{-1} \sum_{i\in\mathcal{U}_N} \int_0^\infty \mathbf{Z}_i(t)\, dG_i(t).
\end{aligned}$$

Thus the finite population score function, $U_{N,p}(\boldsymbol{\beta})$, converges to the superpopulation score function $u_{N,p}(\boldsymbol{\beta})$, where

$$\begin{aligned}
u_{N,p}(\boldsymbol{\beta}) = \alpha &- \int_0^\infty \frac{s_p^{(1)}(\boldsymbol{\beta}, t)}{s^{(0)}(\boldsymbol{\beta}, t)}\, dg(t) \\
&+ 0.5\mathrm{tr}\left(\int_0^\infty \left[\int_0^\infty \left\{ \frac{s^{(2)}(\boldsymbol{\beta}, t)}{s^{(0)}(\boldsymbol{\beta}, t)} - \left(\frac{\mathbf{s}^{(1)}(\boldsymbol{\beta}, t)}{s^{(0)}(\boldsymbol{\beta}, t)}\right)\left(\frac{\mathbf{s}^{(1)}(\boldsymbol{\beta}, t)}{s^{(0)}(\boldsymbol{\beta}, t)}\right)' \right\} dg(t) \right]^{-1} \right. \\
&\qquad \left\{ \left(\frac{q_p^{(2)}(\boldsymbol{\beta}, t)}{s^{(0)}(\boldsymbol{\beta}, t)} - \frac{q_p^{(0)}(\boldsymbol{\beta}, t)}{s^{(0)}(\boldsymbol{\beta}, t)} \frac{s^{(2)}(\boldsymbol{\beta}, t)}{s^{(0)}(\boldsymbol{\beta}, t)} \right) \right. \\
&\qquad - \left(\frac{\mathbf{q}_p^{(1)}(\boldsymbol{\beta}, t)}{s^{(0)}(\boldsymbol{\beta}, t)} - \frac{q_p^{(0)}(\boldsymbol{\beta}, t)}{s^{(0)}(\boldsymbol{\beta}, t)} \frac{\mathbf{s}^{(1)}(\boldsymbol{\beta}, t)}{s^{(0)}(\boldsymbol{\beta}, t)} \right)\left(\frac{\mathbf{s}^{(1)}(\boldsymbol{\beta}, t)}{s^{(0)}(\boldsymbol{\beta}, t)} \right)' \\
&\qquad \left. \left. - \left(\frac{\mathbf{s}^{(1)}(\boldsymbol{\beta}, t)}{s^{(0)}(\boldsymbol{\beta}, t)} \right)\left(\frac{\mathbf{q}_p^{(1)}(\boldsymbol{\beta}, t)}{s^{(0)}(\boldsymbol{\beta}, t)} - \frac{q_p^{(0)}(\boldsymbol{\beta}, t)}{s^{(0)}(\boldsymbol{\beta}, t)} \frac{\mathbf{s}^{(1)}(\boldsymbol{\beta}, t)}{s^{(0)}(\boldsymbol{\beta}, t)} \right)' \right\} dg(t) \right).
\end{aligned}$$

Now assume that the population quantities, $\mathbf{Z}_i$, that are used to define the score functions have finite moments and the sequence of sample designs is such that any smooth functions of NHT estimators are consistent. Because $U_N(\boldsymbol{\beta})$ is a smooth function of population totals, and each total is estimated by using a NHT estimator, $\hat{U}_N(\boldsymbol{\beta})$ is design-consistent for $U_N(\boldsymbol{\beta})$. That is, $\left(U_N(\boldsymbol{\beta}) - \hat{U}_N(\boldsymbol{\beta})\right) \big| \mathcal{F}_N = o(1)$. Therefore, by using arguments similar to Lin (2000) and Andersen and Gill (1982), it can be shown that $\boldsymbol{\beta}_N$ and $\hat{\boldsymbol{\beta}}_N$ converge to the same limit.

Because n/N is fixed, $\sum_{A_N} w_i$ is the NHT estimator for N, and $\hat{U}(\boldsymbol{\beta})$ is a consistent estimator (not necessarily unbiased) of 0, both $\hat{U}(\boldsymbol{\beta})$ and $(n/N)\hat{U}(\boldsymbol{\beta})$ converge to the same limit with the same order of convergence. It is straightforward to show that $(n/N)\hat{U}(\boldsymbol{\beta})$ and $\left(n\big/\sum_{A_N} w_i\right)\hat{U}(\boldsymbol{\beta})$, the estimating equations that use the scaled weights, have the same expectation.

Appendix 2

SAS program to obtain the Firth penalized likelihood estimates

The SAS statements at the end of this section fit a proportional hazards regression model using the scaled weights in Firth's penalized likelihood. The PROC statement invokes the procedure, and the VARMETHOD = JK option requests the jackknife variance estimation method. You can also specify VARMETHOD = TAYLOR, VARMETHOD = BRR, or VARMETHOD = BOOT to request the Taylor series linearized, balanced repeated replication, or bootstrap replication variance estimation method, respectively. The DETAILS sub-option of the VARMETHOD = JK option prints estimates from each replicate sample along with the convergence status. The WEIGHT statement specifies the sampling weights, the STRATA statement specifies the strata, and the CLUSTER statement specifies the PSUs. The MODEL statement specifies the analysis model. The FIRTH option in the MODEL statement requests Firth's penalized likelihood. The two HAZARDRATIO statements requests hazard ratios for blood cholesterol and smoking, respectively. The ODS OUTPUT statement stores replicate estimates and convergence status from each replicate in the SAS data set RepEstimatesFirth. This data set is useful for checking the convergence status of every replicate sample.

```
proc surveyphreg data = NHEFS varmethod=jk (details);
      class        Gender HighBloodChol Race Smoker;
      weight       ObservationWeight;
      strata       Stratum;
      cluster      PSU;
      model        EventTime*HeartAttack(2)  =  Income  HighBloodChol
                   Smoker Race Gender Race*Gender / firth;
      hazardratio HighBloodChol;
      hazardratio Smoker;
      ods output   repestimates=RepEtimatesFirth;
run;
```

References

Andersen, P.K., and Gill, R.D. (1982). Cox's regression model counting process: A large sample study. *Annals of Statistics*, 10, 1100-1120.

Bender, R., Augustin, T. and Blettner, M. (2005). Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine*, 24, 1713-1723.

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.

Binder, D.A. (1990). Fitting Cox’s proportional hazards models from survey data. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 342-347.

Binder, D.A. (1992). Fitting Cox’s proportional hazards models from survey data. *Biometrika*, 79, 139-147.

Binder, D.A., and Patak, Z. (1994). Use of estimating functions for estimation from complex surveys. *Journal of the American Statistical Association*, 89(427),1035-1043.

Boudreau, C., and Lawless, J.F. (2006). Survival analysis based on the proportional hazards model and survey data. *Canadian Journal of Statistics*, 34, 203-216.

Breslow, N.E. (1974). Covariance analysis of censored survival data. *Biometrics*, 30, 89-99.

Chambers, R.L., and Skinner, C.J. (2003). *Analysis of Survey Data*. Chichester, UK: John Wiley & Sons, Inc.

Chambless, L.E., and Boyle, K.E. (1985). Maximum likelihood methods for complex sample data: Logistic regression and discrete proportional hazards models. *Communications in Statistics - Theory and Methods*, 14, 1377-1392.

Cox, D.R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society, Series B*, (with discussion), 34, 187-220.

Cox, D.R. (1975). Partial likelihood. *Biometrika*, 62, 269-276.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, 80, 27-38.

Fuller, W.A. (1975). Regression analysis for sample survey. *Sankhyā*, *Series C*, 37, 117-132.

Fuller, W.A. (2009). *Sampling Statistics*. Hoboken, NJ: John Wiley & Sons, Inc.

Godambe, V.P., and Thompson, M.E. (1986). Parameters of superpopulation and survey population: Their relationships and estimation. *International Statistical Review*, 54, 127-138.

Heinze, G. (1999). *The Application of Firth’s Procedure to Cox and Logistic Regression*. Technical Report 10, updated January 2001, Department of Medical Computer Sciences, Section of Clinical Biometrics, University of Vienna.

Heinze, G., and Schemper, M. (2001). A solution to the problem of monotone likelihood in Cox regression. *Biometrics*, 51, 114-119.

Heinzel, G., Rüdiger, A. and Schilling, R. (2002). *Spectrum and Spectral Density Estimation by the Discrete Fourier Transform (DFT), Including a Comprehensive List of Window Functions and Some New Flat-Top Windows*. Tech. rep., Max-Planck-Institut für Gravitationsphysik (Albert-Einstein-Institut), Teilinstitut Hannover.

Kish, L., and Frankel, M.R. (1974). Inference from complex samples. *Journal of the Royal Statistical Society, Series B*, 36, 1-37.

Korn, E.L., and Graubard, B.I. (1999). *Analysis of Health Surveys*. New York: John Wiley & Sons, Inc.

Lee, E.W., Wei, L.J. and Amato, D.A. (1992). Cox-type regression analysis for large numbers of small groups of correlated failure time observations. *Survival Analysis: State of the Art*, (Eds., J.P. Klein and P.K. Goel), Dordrecht, Netherlands: Kluwer Academic, 237-247.

Lin, D.Y. (2000). On fitting Cox's proportional hazards models to survey data. *Biometrika*, 87, 37-47.

Lin, D.Y., and Wei, L.J. (1989). The robust inference for the proportional hazards model. *Journal of the American Statistical Association*, 84, 1074-1078.

Mukhopadhyay, P.K. (2010). Not hazardous to your health: Proportional hazards modeling for survey data with the SURVEYPHREG procedure. *Proceedings of the SAS Global Forum 2010 Conference*. Cary, NC: SAS Institute Inc. http://support.sas.com/resources/papers/proceedings10/254-2010.pdf.

Mukhopadhyay, P.K., An, A.B., Tobias, R.D. and Watts, D.L. (2008). Try, try again: Replication-based variance estimation methods for survey data analysis in SAS 9.2. *Proceedings of the SAS Global Forum 2008 Conference*. Cary, NC: SAS Institute Inc. http://www2.sas.com/proceedings/forum2008/367-2008.pdf.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review*, 61, 317-337.

Rao, J.N.K. (2005). Interplay between sample survey theory and practice: An appraisal. *Survey Methodology*, 31, 2, 117-138. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2005002/article/9040-eng.pdf.

SAS Institute Inc. (2018). *SAS/STAT 15.1 User's Guide*. Cary, NC: SAS Institute Inc. http://go.documentation.sas.com/?docsetId=statug&docsetTarget=titlepage.htm&docsetVersion=15.1&locale=en.

Shao, J., and Tu, D. (1995). *The Jackknife and Bootstrap*. New York: Springer-Verlag.

Sitter, R.R. (1992). Comparing three bootstrap methods for survey data. *Canadian Journal of Statistics*, 20, 135-154.

Wolter, K.M. (2007). *Introduction to Variance Estimation*. 2nd Ed. New York: Springer.