

FRACTIONAL LOGISTIC REGRESSION FOR CENSORED SURVIVAL DATA

SHAOWU TANG

Roche Molecular Systems, Inc., 4300 Hacienda Drive, Pleasanton, CA 94588, USA

Email: shaowu.tang@roche.com

JONG-HYEON JEONG*

*Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh
130 DeSoto Street, Pittsburgh, PA 15261, USA*

Email: jjeong@pitt.edu

CHI SONG

*Division of Biostatistics, College of Public Health, Ohio State University
1841 Neil Ave, 280E Cunz Hall, Columbus, OH 43210, USA*

Email: song.1188@osu.edu

SUMMARY

In the analysis of time-to-event data, e.g. from cancer studies, the group effect of main interest such as treatment effect of a chemo-therapy often needs to be adjusted by confounding factors (possibly continuous) such as hormonal receptor status, age at diagnosis, and pathological tumor size, when the study outcome is affected by their imbalanced distributions across the comparison groups. The median, or quantile, is a popular summary measure for censored survival data due to its robustness. In this paper, first the logistic regression is extended to fractional responses transformed from censored survival data, which can *directly* predict conditional survival probabilities beyond a fixed time point given covariates. As a special case, we construct a median test for censored survival data that can be used to assess a group effect adjusting for the potentially multiple confounding factors. A quasi-likelihood-based inference procedure is adopted to construct the test statistic. Simulation studies show empirical type I error probabilities and powers for the adjusted two-sample median test are reasonable. The method is illustrated with a breast cancer dataset.

Keywords and phrases: Adjusted median test, Censoring, Fractional response, Median failure time, Quantile

AMS Classification: 62N

* Corresponding author

© Institute of Statistical Research and Training (ISRT), University of Dhaka, Dhaka 1000, Bangladesh

1 Introduction

In clinical studies that generate time-to-event data, or survival data, a popular summary measure to compare the outcome data has been the median, or quantile, event time due to its robustness. For example, in a breast cancer treatment study (B-14) performed by National Surgical Adjuvant and Bowel Project (NSABP), breast cancer patients with positive estrogen receptors were assigned to either a hormonal therapy (tamoxifen) or placebo group (Fisher *et al.*, 1989, 2002). The usual study design for this type of study is to detect a reduction in the hazard rate of the events of interest, but investigators might also be interested in comparing the median times to the events of interest, say, recurrence of the original cancer, to evaluate an effect of the hormonal therapy. This concept of the quantile would provide a straightforward interpretation about the drug effect in terms of extending times to the events of interest.

In time-to-event data analysis, true probabilities of surviving beyond a fixed time point can be simply dichotomized as long as events occurs at any time point or censoring happens after the fixed time point. A case where the survival probability has to be stochastically determined would be when the observed survival time is censored before the fixed time point. This process would transform the time-to-event responses into the fractions between 0 and 1, the limits inclusive, which will be referred to as “fractional responses” for the rest of this paper. Therefore the inference procedures developed for the ordinary logistic regression model cannot be directly applied to this case. In econometrics, Papke and Wooldridge (1996) proposed a regression model on simple fractional responses such as proportions, but, to the best of our knowledge, it has never been extended to a biomedical setting where the fractional responses can be obtained by transforming censored time-to-event data.

When two or more groups are compared in terms of the median times to events of interest, Brookmeyer and Crowley (1982), Wang and Hettmansperger (1990), and Tang and Jeong, (2012), among others, proposed various median test statistics without adjusting for confounding factors. In the fractional logistic regression setting described above, if the fixed time point is set to be the true median of the population under the null hypothesis of equal medians, any significant difference among comparison groups in probabilities of surviving beyond the pooled sample median would lead to a rejection of the null hypothesis. Therefore this particular application of the fractional logistic regression based on the transformed time-to-event data would be an extension of the existing median tests to one accommodating covariate adjustment.

The rest of the paper is organized as follows: In Section 2, the fractional logistic regression model is reviewed and extended to the right censoring case. In Section 3, the quasi-likelihood-based inference procedure is reviewed for the regression coefficients from the fractional logistic regression model, and the new median test is constructed. In Section 4, simulation studies are performed to assess performance of the proposed median test with and without covariate adjustment, and compared with existing methods. In Section 5, the proposed test is applied to a breast cancer dataset, and Section 6 discusses some advantages of the proposed test.

2 Model

Suppose that for the i^{th} subject ($i = 1, 2, \dots, n$), X_i is the true event time and C_i is the true censoring time, and define the observed survival time as $T_i = \min(X_i, C_i)$. Also define the event indicator as $\delta_i = I(X_i < C_i)$. Furthermore suppose that we are interested in inferring the survival probability beyond a fixed time point t_0 , and define an indicator $Y_i = I(X_i > t_0)$. Given partially or fully observable survival data (T_i, δ_i, z_i) , where z_i is a vector of covariates, we extend the following regression model (Papke and Wooldridge, 1996) for censored survival data:

$$E(Y_i|z_i) = G(z_i'\beta), \tag{2.1}$$

where the function $G(\cdot)$ can take a nonlinear form and β is a vector of regression coefficients. Note that without censoring, the outcome Y_i is fully observable as a Bernoulli random variable once a time point is fixed at t_0 . Even under censoring, Y_i is observable as long as T_i is greater than t_0 or T_i is less than t_0 as an event. However, when T_i is less than t_0 as censored, Y_i is not observable any longer, so that $E(Y_i|T_i < t_0, \delta_i = 0)$ can be replaced with the estimate of $\Pr(X_i > t_0|T_i < t_0, \delta_i = 0)$. We adopt the estimation procedure from Brookmeyer and Crowley (1982) as follows:

$$q_i(t_0) = \Pr(X_i > t_0|T_i < t_0, \delta_i = 0) = \Pr(X_i > t_0|X_i > T_i) = \frac{\Pr(X_i > t_0)}{\Pr(X_i > T_i)} = \frac{S(t_0)}{S(T_i)}, \tag{2.2}$$

where the true survival function $S(\cdot)$ can be estimated by the Kaplan-Meier estimator (Kaplan and Meier, 1958). Heuristically, from the uniform consistency of the Kaplan-Meier estimator (Gill, 1983; Wang, 1987), the estimate $\hat{q}_i(t_0)$ would be also consistent for a large sample. Therefore, conditional on a fixed time point t_0 , the time-to-event outcome can be transformed into the fractional outcome consisting of any numbers between 0 and 1. Under model (2.1), if t_0 is the median failure time under the null hypothesis of equal medians between two groups, testing for the regression coefficient for a group effect would be equivalent to testing for the equal medians adjusted for other covariates.

3 Adjusted Median Test Statistic

In model (2.1), note that the left hand side term is basically the conditional survival probability beyond t_0 given covariates, so that a reasonable choice for $G(\cdot)$ would be a cumulative distribution function, such as the logistic function, i.e. $G(u) = e^u/(1 + e^u)$. Under model (2.1) with the logistic link function, the Bernoulli pseudo-loglikelihood function for a subject i is given by (TABLE II, Gourieroux *et al.*, 1984)

$$l(\beta) = \sum_{i=1}^n y_i \log\{G(z_i'\beta)\} + (1 - y_i) \log\{1 - G(z_i'\beta)\}. \tag{3.1}$$

Under some regularity conditions, Gourieroux *et al.* (1984) showed that the pseudo-maximum likelihood estimator of β from (3.1) is asymptotically consistent and efficient through the quasi-likelihood inference only by specifying the mean and variance functions.

To find the asymptotic variance of the Bernoulli pseudo-maximum likelihood estimator, define $g(u) := dG(u)/du$, $\hat{G}_i := G(z_i'\hat{\beta}) = \hat{y}_i$ and $\hat{g}_i := g(z_i'\hat{\beta})$. Note that with $G(u) = e^u/(1 + e^u)$, it holds

$$g(u) = \frac{dG(u)}{du} = \frac{e^u}{(1 + e^u)^2} = G(u) \cdot (1 - G(u)). \quad (3.2)$$

Therefore, the observed information matrix is

$$\hat{\mathcal{J}} := \mathcal{J}(\hat{\beta}) = -\nabla\nabla^T l(\beta)|_{\beta=\hat{\beta}} = \sum_{i=1}^N \hat{G}_i(1 - \hat{G}_i)z_i'z_i. \quad (3.3)$$

A naive standard error of $\hat{\beta}_j$ can be estimated from the square root of the j^{th} diagonal element of $\hat{\mathcal{J}}^{-1}$. However, as Papke and Wooldridge (1996) pointed out, this would not be a consistent estimator of the true asymptotic standard error because the outcomes are fractions, not binary. They proposed a robust estimator of the asymptotic variance of $\hat{\beta}$ against model misspecification, i.e. the Huber sandwich estimator, as

$$\hat{\mathcal{J}}^{-1}\hat{B}\hat{\mathcal{J}}^{-1}, \quad (3.4)$$

where

$$\hat{B} = [\nabla^T l(\beta) \cdot \nabla l(\beta)]|_{\beta=\hat{\beta}} = \sum_{i=1}^N \frac{\hat{u}_i^2 \hat{g}_i^2 z_i'z_i}{[\hat{G}_i(1 - \hat{G}_i)]^2} = \sum_{i=1}^N \hat{u}_i^2 z_i'z_i$$

with

$$\hat{u}_i := y_i - G(z_i'\hat{\beta}).$$

The consistent standard error estimates of $\hat{\beta}_j$'s can be obtained as the square roots of the diagonal elements of equation (3.4). Therefore, the Wald-type median test statistic adjusted for other covariates would be the estimated regression coefficient associated with the group indicator as a covariate divided by its corresponding standard error from (3.4).

4 Simulation Studies

In this section, we consider the following model to test for a difference in the medians between the two groups adjusted for the covariate vector $z_i = (1, z_{i1}, \dots, z_{im})'$, where z_{i1} is a group indicator and the others are confounding factors:

$$E(Y_i|z_i) = \Pr(X_i > \hat{\theta}_{pooled}|z_i) = \frac{\exp(z_i'\beta)}{1 + \exp(z_i'\beta)}, \quad (4.1)$$

where $\beta = (\beta_0, \beta_1, \dots, \beta_m)'$ is a corresponding vector of the regression coefficients. Here $\hat{\theta}_{pooled}$ is the estimated median under the Kaplan-Meier estimates based on the pooled

sample (T_i, δ_i) , i.e. under the null hypothesis of $\beta_1 = 0$, where β_1 is the regression coefficient associated with z_{i1} . Failure to reject the null hypothesis of $\beta_1 = 0$ would imply that the population medians being compared are equivalent adjusted for other covariates.

Once the data are observed, the fractional response Y_i can be estimated under the null hypothesis as follows (Brookmeyer and Crowley, 1982):

$$\Pr(X_i > \hat{\theta}_{pooled}) = \begin{cases} 1, & \text{if } T_i > \hat{\theta}_{pooled} \text{ and } \delta_i = 1, \\ 1, & \text{if } T_i \geq \hat{\theta}_{pooled} \text{ and } \delta_i = 0, \\ 0, & \text{if } T_i \leq \hat{\theta}_{pooled} \text{ and } \delta_i = 1, \\ \hat{q}_i(\hat{\theta}_{pooled}), & \text{if } T_i < \hat{\theta}_{pooled} \text{ and } \delta_i = 0, \end{cases}$$

where $\hat{q}_i(\cdot)$ was defined in (2.2).

4.1 Simulation Studies with a Single Binary Covariate

First we performed simulation studies to evaluate the proposed method for two-sample cases without covariate adjustment. When two groups are compared, the group indicator can be included in the regression model as a single binary covariate. With n_k being the number of subjects in group k ($k = 1, 2$), denote T_{kj} ($j = 1, \dots, n_k$) for a random variable for failure times for a subject j in group k . Therefore in the combined sample, we have

$$\begin{aligned} T &= (T_{11}, T_{12}, \dots, T_{1,n_1}, T_{21}, T_{22}, \dots, T_{2,n_2})', \\ \delta &= (\delta_{11}, \delta_{12}, \dots, \delta_{1,n_1}, \delta_{21}, \delta_{22}, \dots, \delta_{2,n_2})', \text{ and} \\ Y &= (Y_{11}, Y_{12}, \dots, Y_{1,n_1}, Y_{21}, Y_{22}, \dots, Y_{2,n_2})'. \end{aligned}$$

Let us denote the i^{th} element of the vectors T , δ , and Y as T_i , δ_i , and Y_i , respectively, and the covariate vector for subject i as $z_i = (1, z_{i1})'$.

Extensive simulations have been performed to assess type I error probabilities and powers of the proposed approach at various nominal significance levels (α), sample sizes per group (n_k), and censoring proportions. Two scenarios are used to generate failure times to investigate type I error probabilities: (1) $S_1(t) = S_2(t) = \exp(-t)$ and (2) $S_1(t) = \exp(-t)$ and $S_2(t) = 1 - \Phi(\log(1.44t))$. In case (1), both samples are taken from the same distribution, so that the medians are equal. In case (2), two samples are taken from different distributions but with the same medians down to two decimal places. Censoring times are generated from $\text{Uniform}(0, c_k)$, where c_k determines a different censoring proportion for group k ($k = 1, 2$). The censoring parameter c_k 's are chosen so that the censoring proportions are similar to those reported in Su and Wei (1993) for comparison. With sample sizes of 30, 50 and 100 for each group and various censoring proportions, the proportions of cases that were significantly different in median failure times between two samples are reported in Table 1 under scenario (1) and Table 2 under scenario (2) at various nominal levels of 5%, 10%, 15% and 20%. One can notice that the proposed method controls the type I error probabilities reasonably well.

Table 1: Fractional logistic regression: empirical type I error probabilities for case I with $S_1(t) = S_2(t) = \exp(-t)$

n_k	α	Mean Censoring Proportion					
		.43, .43	.28, .28	.1, .1	.01, .01	.1, .28	.1, .43
30	0.05	0.051	0.06	0.046	0.036	0.044	0.045
	0.1	0.116	0.096	0.107	0.089	0.107	0.116
	0.15	0.154	0.159	0.161	0.155	0.143	0.151
	0.2	0.198	0.208	0.202	0.228	0.221	0.207
50	0.05	0.046	0.046	0.045	0.068	0.051	0.056
	0.1	0.105	0.097	0.084	0.089	0.103	0.101
	0.15	0.137	0.141	0.137	0.135	0.155	0.159
	0.2	0.197	0.220	0.218	0.20	0.193	0.188
100	0.05	0.041	0.052	0.052	0.049	0.056	0.054
	0.1	0.109	0.077	0.098	0.086	0.105	0.085
	0.15	0.165	0.156	0.164	0.154	0.142	0.140
	0.2	0.198	0.206	0.192	0.193	0.203	0.183

Power analysis at a significance level of 0.05 has also been performed for various censoring proportions by increasing the median differences with different sample sizes of 50, 100 and 200. Data were generated similarly as before from two distributions $S_1(t) = \exp(-t)$ and $S_2(t) = \exp(-t + t^*)$, where t^* is the median difference between the two exponential distributions. Tables 3 summarizes the proportion of the proposed test statistic rejecting the null hypothesis of equal medians, which quickly increases with larger median differences and sample sizes, as expected.

We have also compared the powers of the proposed regression-type median test (Fractional Logistic) with existing two-sample median tests, i.e. Brookmeyer and Crowley's test (Brookmeyer and Crowley, 1982) and the Chi-square test (Tang and Jeong, 2012) (see Figure 1). The proposed method seems to perform better than the Chi-square test, and performs similar to Brookmeyer and Crowley's approach. In the next subsections, simulation studies will be performed under the accelerated failure time (AFT) model to evaluate the proposed procedure to adjust for an additional continuous covariate.

Table 2: Fractional logistic regression: empirical type I error probabilities for case II with $S_1(t) = \exp(-t), S_2(t) = 1 - \Phi(\log(1.44t))$

n_k	α	Mean Censoring Proportion					
		.43,.43	.28,.28	.1,.1	.01,.01	.1,.28	.1,.43
30	0.05	0.05	0.059	0.046	0.035	0.05	0.055
	0.1	0.118	0.092	0.103	0.097	0.112	0.114
	0.15	0.153	0.153	0.165	0.158	0.152	0.157
	0.2	0.208	0.211	0.207	0.240	0.216	0.227
50	0.05	0.052	0.051	0.042	0.073	0.050	0.054
	0.1	0.115	0.102	0.081	0.090	0.105	0.102
	0.15	0.136	0.142	0.140	0.140	0.147	0.160
	0.2	0.203	0.215	0.213	0.195	0.192	0.193
100	0.05	0.041	0.055	0.055	0.049	0.056	0.047
	0.1	0.122	0.085	0.104	0.092	0.110	0.096
	0.15	0.160	0.159	0.162	0.148	0.147	0.135
	0.2	0.205	0.208	0.197	0.193	0.196	0.190

Table 3: Empirical powers for $S_1(t) = 1 - \exp(-t)$ and $S_2(t) = 1 - \exp(-t + t^*)$, where t^* is the median difference; significance level=0.05

n_k	t^*	Mean Censoring Proportion					
		.43,.43	.28,.28	.1,.1	.01,.01	.1,.28	.1,.43
50	0.1	0.134	0.145	0.132	0.127	0.131	0.117
	0.2	0.287	0.292	0.283	0.253	0.278	0.270
	0.3	0.503	0.474	0.455	0.426	0.450	0.499
	0.4	0.697	0.660	0.637	0.617	0.654	0.698
	0.5	0.855	0.789	0.800	0.782	0.805	0.825
100	0.1	0.209	0.183	0.167	0.199	0.184	0.189
	0.2	0.465	0.443	0.409	0.407	0.423	0.440
	0.3	0.776	0.705	0.693	0.659	0.706	0.720
	0.4	0.932	0.905	0.885	0.877	0.900	0.923
	0.5	0.986	0.986	0.965	0.961	0.968	0.980
200	0.1	0.297	0.279	0.258	0.239	0.282	0.278
	0.2	0.741	0.690	0.645	0.621	0.696	0.697
	0.3	0.934	0.942	0.915	0.908	0.920	0.937
	0.4	0.999	0.994	0.992	0.986	0.992	0.994
	0.5	1.0	1.0	0.999	1.0	1.0	1.0

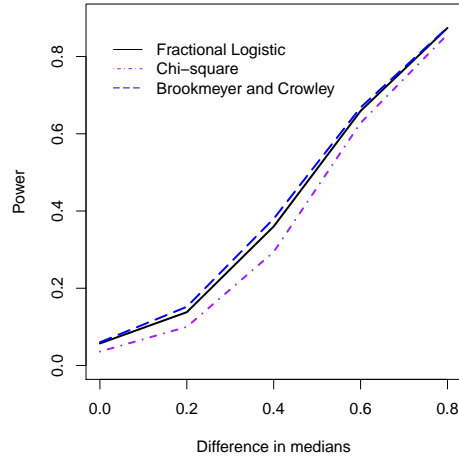


Figure 1: Power comparison among the proposed method (Fractional Logistic), the Chi-square test, and Brookmeyer and Crowley’s test.

4.2 Simulation Studies with an Additional Adjusting Covariate

In this section, we simulate survival data from a log-linear model, to demonstrate practical applicability of the proposed test when covariate adjustment is needed. The log-linear model can be specified as

$$\log X_i = z_i' \beta + \sigma W, \tag{4.2}$$

where W is a random variable for an error distribution and $z_i = (1, z_{i1}, z_{i2})'$, z_{i1} being a binary covariate and z_{i2} being a continuous covariate. It is well known that the model (4.2) is equivalent to the accelerated failure time (AFT) model $S(x|z_i) = S_0(xe^{-z_i' \beta})$, noting that the sign of the vector of regression coefficients are reversed. In our simulation studies, we assume that W follows (i) a standard extreme value distribution with the cumulative distribution function (CDF) of $F_W(w) = 1 - e^{-e^w}$ and (ii) a standard logistic distribution with the CDF of $F_W(w) = e^w / (1 + e^w)^2$. It is also well known that under the extreme value error structure, the true failure time X follows a Weibull distribution, which is closed under both AFT and proportional hazards models, and the logistic error structure gives the log-logistic distribution for X , which is closed under the proportional odds model (Bennett, 1983). The simulation scenario was as follows:

1. $\sigma = 0.5$.
2. $z_i' \beta = \beta_0 + \beta_1 z_{i1} + \beta_2 z_{i2}$ where $\beta_0 = -0.5, \beta_2 = 1$ and $\beta_1 = 0$ for type I error and $\beta_1 = 1$ for power analysis.

3. $z_{i1} \sim \text{Bernoulli}(1, 0.5)$ and $z_{i2} \sim N(1, 1)$.
4. Censoring times were generated from an exponential distribution with a fixed rate parameter that produced the corresponding censoring proportions in Table 4.
5. Sample sizes per group; 60, 100 and 200.
6. For each simulation, 1000 simulated datasets were generated to assess type 1 error probabilities and powers at the nominal significance level 0.05.

Type 1 error probabilities and powers from the proposed model are summarized for the extreme value error and logistic error models in Table 4. One can notice that type I error probabilities were controlled quite well under both models, and powers increase as the sample size increases, but not much affected by heavy censoring.

5 Application to NSABP Data

We applied the proposed method to a dataset from the NSABP B-14 study described in the Introduction Section. In this study, total 2,892 patients with primary cancer, negative axillary nodes and estrogen receptor positive tumors were originally randomized to receive either tamoxifen (a hormonal therapy) or placebo following a surgery. In this example, a cohort of 2,852 eligible patients with follow-up and known pathological tumor size (1,283 patients from tamoxifen group and 1,299 patients from placebo group) has been analyzed. The censoring proportion in this dataset was about 66% when the long-term follow-up was closed. The median follow-up was about 16 years. For our analysis here, we have included three covariates; treatment indicator (z_{i1}), pathological tumor size (z_{i2}), and age at diagnosis (z_{i3}), which give the model at a fixed time point t_0

$$\Pr(X_i > t_0 | z_i) = \frac{\exp(\beta_0 + \beta_1 z_{i1} + \beta_2 z_{i2} + \beta_3 z_{i3})}{1 + \exp(\beta_0 + \beta_1 z_{i1} + \beta_2 z_{i2} + \beta_3 z_{i3})}. \quad (5.1)$$

The primary endpoint was overall survival (OS), defined as time to any death.

Since the survival estimates in the pooled data haven't reached 0.5 yet as shown in Figure 2, we have compared the lower quantile failure times between the two treatment groups, i.e. 0.1-, 0.2-, 0.3-, and 0.4-quantile failure times, rather the median failure time.

Recall that after replacing the fixed time point t_0 in model (5.1) with the median failure time from the pooled data, testing the null hypothesis of $\beta_1 = 0$ would be equivalent to a test for the equal median failure times between two treatment groups adjusted for the other two covariates, i.e. tumor size and age at diagnosis. The test results for different quantile failure times are given in Table 6.

The results indicate that there is no statistically significant difference up to the .2-quantile failure time between the two treatment groups after adjusting for tumor size and age at diagnosis, which changes the direction starting from the .25 quantile. This was also reflected in the Kaplan-Meier plot without the covariate adjustment in Figure 2. As the

Table 4: Type 1 error probabilities ($\beta_1 = 0$) and powers ($\beta_1 = 1$) of the proposed test under the extreme value error and logistic error models

		Censoring Proportion		Type I Error/Power	
Sample Size		Extreme value	Logistic	Extreme value	Logistic
$\beta_1 = 0$	$n = 60$	0.23	0.24	0.058	0.059
		0.32	0.36	0.050	0.05
		0.47	0.56	0.055	0.063
	$n = 100$	0.23	0.24	0.048	0.041
		0.32	0.36	0.048	0.044
		0.47	0.56	0.060	0.051
	$n = 200$	0.23	0.24	0.051	0.052
		0.32	0.36	0.047	0.044
		0.47	0.56	0.051	0.056
$\beta_1 = 1$	$n = 60$	0.30	0.33	0.381	0.765
		0.41	0.48	0.374	0.716
		0.56	0.67	0.301	0.583
	$n = 100$	0.30	0.33	0.595	0.951
		0.40	0.47	0.578	0.936
		0.55	0.67	0.526	0.798
	$n = 200$	0.30	0.33	0.882	0.999
		0.40	0.48	0.853	0.994
		0.55	0.67	0.820	0.978

Table 5: Estimates of regression coefficients and associated standard errors and p -values for various quantiles of the distribution of NSABP B-14 data

Quantile	$\hat{\beta}_0$			$\hat{\beta}_1$			$\hat{\beta}_2$			$\hat{\beta}_3$		
	Est.	P-val	S.E.	Est.	P-val	S.E.	Est.	P-val	S.E.	Est.	P-val	S.E.
.10	2.84	$< 10^{-4}$.41	.19	.15	.13	-.03	$< 10^{-4}$.004	-.002	.78	.007
.20	2.40	$< 10^{-4}$.31	.16	.11	.10	-.025	$< 10^{-4}$.004	.01	.07	.005
.30	2.44	$< 10^{-4}$.28	.25	.0034	.085	-.02	$< 10^{-4}$.0034	-.023	$< 10^{-4}$.0046
.40	2.28	$< 10^{-4}$.23	.20	.005	.07	-.02	$< 10^{-4}$.003	-.028	$< 10^{-4}$.004

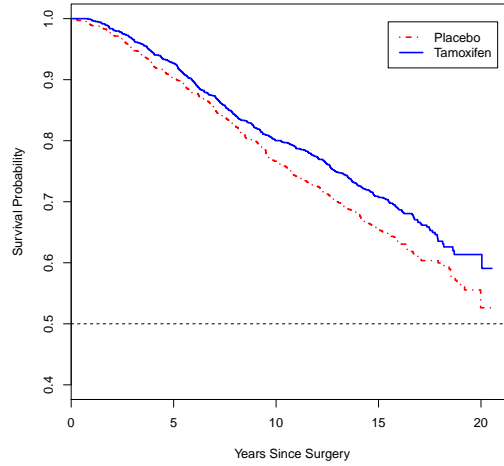


Figure 2: Kaplan-Meier estimates in NSABP B-14 data.

byproduct of the results presented in Table 5, we can also estimate the survival probability beyond the .4-quantile failure time in the pooled sample (=18.6 years) for a woman who developed her breast cancer at the age of 50, was assigned to tamoxifen group, and her tumor size was 20mm, from the following model:

$$\Pr(X_i > 18.6|z_i) = \frac{\exp(2.28 + 0.20 - 0.02 \times 20 - 0.028 \times 50)}{1 + \exp(2.28 + 0.20 - 0.02 \times 20 - 0.028 \times 50)}, \quad (5.2)$$

which gives 67%. Cox's proportional hazards model (Cox, 1972) also provided a close predicted survival probability of 68% at year 18.6. In fact, the predicted probabilities from the two models were very similar for the 0.1-, 0.2-, and 0.3-quantiles as well; 92%, 83%, and 76% from the fractional logistic regression model vs. 93%, 85%, and 77% from Cox's model, respectively, even though the underlying assumptions are different between these two models. It is well known that Cox's model needs to satisfy the proportional hazards assumption and the logit should be linear in covariates under the logistic regression model.

6 Discussion

In this paper, we have extended the fractional logistic regression to censored survival data. This work contributes in two folds; (i) it provides a simple and direct model to predict survival probabilities for right censored data and (ii) a special case allows for constructing a median test adjusting for confounding factors. A major advantage of the latter would be that it could be applied easily to any time point including various quantiles as well as the mean

of the pooled failure time distribution, where the mean can be evaluated by integrating the Kaplan-Meier estimates of the pooled distribution. Another advantage would be that the proposed approach does neither require estimation of the probability density function nor a complicated variance formula to infer the quantiles. The third advantage would be that the proposed procedure uses the quasi-likelihood-based inference, which does not require any distributional assumption for the error distribution and can be easily implemented by using an existing software such as the `glm` procedure in R. R codes used for the simulation studies and real data analysis are available from the corresponding author.

Acknowledgments

Dr. Jeong's research was supported in part by National Health Institute (NIH) grants 5-U10-CA69974-09 and 5-U10-CA69651-11.

References

- Bennett, S. (1983), "Analysis of survival data by the proportional odds model," *Statistics in Medicine*, 2, 273–278.
- Brookmeyer, R. and Crowley, J. (1982), "A K -sample median test for censored data," *Journal of the American Statistical Association*, 77, 433–440.
- Cox, D. R. (1972), "Regression models and life-tables (with discussion)," *Journal of the Royal Statistical Society: Series B*, 34, 187–220.
- Fisher, B., Costantino, J., Redmond, C. *et al.* (1989), "A randomized clinical trial evaluating tamoxifen in the treatment of patients with node-negative breast cancer who have estrogen-receptor-positive tumors," *New England Journal of Medicine*, 320, 479–484.
- Fisher, B., Jeong, J., Anderson, S. *et al.* (2002), "Twenty-five year findings from a randomized clinical trial comparing radical mastectomy with total mastectomy and with total mastectomy followed by radiation therapy," *New England Journal of Medicine*, 347, 567–575.
- Gourieroux, C., Monfort, A., and Trognon, A. (1984), "Pseudo maximum likelihood methods: theory," *Econometrika*, 52, 681–700.
- Gill, R. D. (1983), "Large sample behaviour of the product limit estimator on the whole line," *Annals of Statistics*, 11, 49–58.
- Kaplan, E. L. and Meier, P. (1958), "Nonparametric estimation from incomplete observations," *Journal of the American Statistical Association*, 53, 457–481.

- Papke, L. and Wooldridge, J. M. (1996), “Econometric methods for fractional response variables with an application to 401 (K) plan participation rates,” *Journal of Applied Econometrics*, 11, 619–632.
- Su, J. Q. and Wei, L. J. (1993), “Nonparametric estimation for the difference or ratio of median failure times,” *Biometrics*, 49, 603–607.
- Tang, S. and Jeong, J. (2012), “Median tests for censored survival data; a contingency table approach,” *Biometrics*, 68, 983–989.
- Wang, J.-G. (1987), “A note on the uniform consistency of the Kaplan-Meier estimator,” *Annals of Statistics*, 15, 1313–1316.
- Wang, J.-L. and Hettmansperger, T. P. (1990), “Two-sample inference for median survival times based on one-sample procedures for censored survival data,” *Journal of the American Statistical Association*, 85, 529–536.

Received: November 15, 2017

Accepted: January 5, 2018