

# A homoscedasticity test for the accelerated failure time model

Lili Yu<sup>1</sup>  · Liang Liu<sup>2</sup> · Ding-Geng Chen<sup>3</sup>

Received: 14 September 2017 / Accepted: 14 September 2018 / Published online: 27 September 2018

## Abstract

The semiparametric accelerated failure time (AFT) model is a popular linear model in survival analysis. AFT model and its associated inference methods assume homoscedasticity of the survival data. It is shown that violation of this assumption will lead to inefficient parameter estimation and anti-conservative confidence interval estimation, and thus, misleading conclusions in survival data analysis. However, there is no valid statistical test proposed to test the homoscedasticity assumption. In this paper, we propose the first novel quasi-likelihood ratio test for the homoscedasticity assumption in the AFT model. Simulation studies show the test performs well. A real dataset is used to demonstrate the usefulness of the developed test.

**Keywords** Accelerated failure time model · Homoscedasticity test · Quasi-likelihood ratio test · Right censoring · Survival analysis

**Electronic supplementary material** The online version of this article (<https://doi.org/10.1007/s00180-018-0840-9>) contains supplementary material, which is available to authorized users.

---

✉ Lili Yu  
lyu@georgiasouthern.edu

Liang Liu  
lliu@uga.edu

Ding-Geng Chen  
dinch@email.unc.edu

<sup>1</sup> Department of Biostatistics, Jiann-Ping Hsu college of Public Health, Georgia Southern University, Statesboro, GA 30460, USA

<sup>2</sup> Department of Statistics, University of Georgia, Athens, GA 30602, USA

<sup>3</sup> School of Social Work, UNC-Chapel Hill, Chapel Hill, NC 27599, USA

## 1 Introduction

The Cox proportional hazards model (Cox 1972) is the most popular method for modeling and estimation of parameters in survival analysis because of the availability of the efficient inference procedures implemented in all statistical softwares. However, as noted by Cox (Reid 1994, p. 450), the classical accelerated failure time (AFT) model (Kalbfleisch and Prentice 1980) is more appealing in many ways due to its direct physical interpretation. The classical AFT model regresses the logarithm of survival time  $T$ , say  $Y \equiv \log T$ , over the covariates  $\mathbf{X}_i$  as follows,

$$Y_i \equiv \log(T_i) = \mu_i + \epsilon_i^\#, \quad i = 1, \dots, n \quad (1)$$

where  $\mu_i = \mathbf{X}_i^T \boldsymbol{\beta}$  is the mean of  $Y_i$  in which  $\mathbf{X}_i$  is a nonrandom  $p$ -dimensional vector of covariates and  $\boldsymbol{\beta}$  is a vector of parameters;  $\epsilon_i^\#, i = 1, \dots, n$ , are independently and identically distributed with unspecified distribution function  $F^\#$  and an assumed constant variance (homoscedasticity). Note the classical AFT model serves as a counterpart of the classical linear model for survival data. The differences between these two models include data transformation and censorship. For data transformation, the classical AFT model uses  $\log$  transformation to make the data symmetric and homoscedastic, while the classical linear model uses some transformations (not just  $\log$  transformation). For the censorship of the survival data, it adds complexity to the inference of the AFT model. The rank method (Lai and Ying 1992; Robins and Tsiatis 1992; Ying 1993; Lin and Ying 1995; Jin et al. 2003; Zhou 2005) and the least squares method (Buckley and James 1979; Ritov 1990; Lai and Ying 1991; Jin et al. 2006) are proposed for the AFT model by accommodating the censorship of survival data. Both of them assume the log-transformed survival data are homoscedastic.

The next logic question is whether this log-transformation can really make the data homoscedastic as assumed in the classical AFT model. It is well known that if the model assumption is not satisfied, the inference may not be valid. For example, in the classical linear model, if data is heteroscedastic, the least squares estimate is inefficient and the estimated variance of the least squares estimate is not a consistent estimator anymore (Greene 2008). For the AFT model, Yu (2011), Yu et al. (2012, 2013) and Yu and Peace (2012) showed if survival data violate the homoscedasticity assumption, the statistical inference of survival based on the AFT model and its associated methodologies (e.g. the rank method and the least squares method) will lead to inefficient estimation. And most importantly, the variance estimates of the parameter estimators are mostly under-estimated with the homogeneous assumption of variance. Consequently, the 95% confidence intervals are anti-conservative and their coverage probabilities are much less than the nominal level 0.95. This will lead to inaccurate conclusions. Therefore, to handle the heteroscedastic data, Yu et al. (2013) proposed a weighted least squares method for the AFT model. They showed the weighted least squares method was more efficient for heteroscedastic data than either the rank method or the least squares method. Moreover, the coverage probabilities of parameter estimators' 95% confidence intervals obtained from the weighted least squares method are closer to the nominal level 0.95.

As an example, we estimated the conditional mean of the *log* survival time (in days), given age (in years) and sex for the NCCTG Lung Cancer Data in R *survival* package. In Sect. 5, we analyzed this data using both the methods that do not consider the heteroscedasticity of the data (i.e., the rank method and the least squares method) and the method that considers the heteroscedasticity of the data (i.e., the weighted least squares method). The results from these methods (shown in Sect. 5) are substantially different. In order to choose an appropriate method to obtain efficient estimation and accurate conclusions, we needed to know if the data is homoscedastic or not. Therefore, a clear need exists for a formal statistical test to investigate the homoscedasticity assumption in the AFT model.

There are many homoscedasticity tests for complete data in the literature as seen in Breusch and Pagan (1980), Cook and Weisberg (1983), Muller and Zhao (1995), Cai et al. (1998), Wang and Zhou (2005, 2007), among others. For censored data, Zhou et al. (2012) proposed an empirical likelihood ratio test method to estimate the coefficients  $\beta$  for both correlation and regression AFT models, but they assumed equal variance of the data. To the best of our knowledge, no homoscedasticity test has yet been proposed for the AFT model, probably due to the presence of a censoring mechanism which poses major theoretical challenges in the semiparametric test of the homoscedasticity in the AFT model. This paper is then aimed to fill this methodological gap by proposing a powerful quasi-likelihood ratio test for the homoscedasticity assumption in the AFT model for survival data.

The paper is organized as follows. In Sect. 2, we introduce the new test. Simulation studies are conducted in Sect. 3 followed by a real data analysis in Sect. 4. Discussions and conclusions are given in Sect. 5.

## 2 Quasi-likelihood ratio test for the AFT model

### 2.1 The novel quasi-likelihood ratio test

Let  $C$  be the *log* of the censoring time. Define  $Z = Y \wedge C$  and  $\delta = I(Y \leq C)$ , where  $\wedge$  is the minimum operator and  $I(\cdot)$  is the indicator function. Then the observed data are denoted by  $\{Z_i, \mathbf{X}_i, \delta_i\}$ ,  $i = 1, \dots, n$ . The censoring time is assumed to be independent of survival time given the covariates  $\mathbf{X}_i$ .

In order to accommodate heteroscedasticity in the AFT model, we use the general linear model for right-censored data (Yu et al. 2013),

$$Y_i \equiv \log(T_i) = \mu_i + \sigma(\mu_i)e_i, \quad i = 1, \dots, n \quad (2)$$

where  $e_i$ ,  $i = 1, \dots, n$ , are independent and identically distributed with unspecified distribution function  $F$  which has mean 0 and variance 1;  $\sigma^2(\mu_i)$  is the variance of  $Y_i$ . Let  $\epsilon_i = \sigma(\mu_i)e_i$ . Note the variances of  $Y_i$  in model (2) are individual-specific which represents the heteroscedasticity of the data. Additionally, this model (2) reduces to the classical AFT model (1) when  $\sigma(\mu_i)$  are all equal. Therefore, the homoscedasticity test for the AFT model tests whether the variance function  $\sigma^2(\mu_i)$  in (2) is constant; that is,

$$\begin{aligned}
H_0 : \sigma^2(\mu_i) &= \sigma^2 \quad \text{for } i = 1, \dots, n \\
H_1 : \sigma^2(\mu_i) &\neq \sigma^2 \quad \text{for at least one } i = 1, \dots, n
\end{aligned}$$

for some unknown positive constant  $\sigma^2$ .

In developing our proposed homoscedasticity test, we used quasi-likelihood (Wedderburn 1974) defined as follows:

$$Q^\#(\boldsymbol{\mu}, \mathbf{Y}, \sigma^2(\boldsymbol{\mu})) = \frac{1}{n} \sum_{i=1}^n \int_{Y_i}^{\mu_i} \frac{Y_i - a}{\sigma^2(a)} da. \quad (3)$$

Because the quasi-likelihood has similar properties to the *log*-likelihood, the test statistic can be constructed as

$$QLR^\# = -2 \left( Q(\tilde{\boldsymbol{\mu}}, \tilde{\mathbf{Y}}, \tilde{\sigma}^2) - Q^\#(\hat{\boldsymbol{\mu}}, \hat{\mathbf{Y}}, \hat{\boldsymbol{\Sigma}}) \right)$$

where  $Q(\tilde{\boldsymbol{\mu}}, \tilde{\mathbf{Y}}, \tilde{\sigma}^2)$  is the quasi-likelihood function under  $H_0$ , in which  $\tilde{\boldsymbol{\mu}}$  is a vector with elements  $\tilde{\mu}_i = \mathbf{X}_i^T \tilde{\boldsymbol{\beta}}$  and  $\tilde{\boldsymbol{\beta}}$ , the least squares estimator (Buckley and James 1979), satisfies

$$\tilde{\boldsymbol{\beta}} = \left( (\mathbf{X} - \bar{\mathbf{X}})^T (\mathbf{X} - \bar{\mathbf{X}}) \right)^{-1} (\mathbf{X} - \bar{\mathbf{X}})^T \tilde{\mathbf{Y}}$$

where  $\bar{X}$  is the sample mean of  $X$ ;  $\tilde{\sigma}^2 = 1/n \sum_{i=1}^n (\tilde{Y}_i - \tilde{\mu}_i)^2$ .  $\tilde{\mathbf{Y}}$  is a vector of synthetic data with elements  $\tilde{Y}_i = Y_i \delta_i + \hat{E}^\#(Y_i | Y_i > C_i)(1 - \delta_i)$  and  $\hat{E}^\#(Y_i | Y_i > C_i)$  is an estimator of  $E(Y_i | Y_i > C_i)$  based on the Kaplan–Meier estimate  $\hat{F}^\#(\epsilon^\#)$ . Let  $r^\#(i)$  be the ordered  $r_i^\# = Z_i - \mathbf{X}_i^T \tilde{\boldsymbol{\beta}}$ , then

$$\hat{F}^\#(\epsilon^\#) = 1 - \prod_{\{i: r^\#(i) \leq \epsilon^\#\}} ((n - i)/(n - i + 1))^{\delta_i}.$$

If the largest observation is censored, we treat it as an uncensored observation (Meier 1975).

$Q^\#(\hat{\boldsymbol{\mu}}, \hat{\mathbf{Y}}, \hat{\boldsymbol{\Sigma}})$  is the quasi-likelihood function under  $H_1$ , in which  $\hat{\boldsymbol{\Sigma}}$  is a  $n \times n$  variance-covariance matrix with diagonal elements  $\hat{\sigma}_n^2(\hat{\mu}_i)$  as the estimators of the  $\sigma^{*2}(\mu_i)$ , variances of synthetic data  $Y_i^* = Y_i \delta_i + E(Y_i | Y_i > C_i)(1 - \delta_i)$ , and  $\hat{\boldsymbol{\mu}}$  is a vector with elements  $\hat{\mu}_i = \mathbf{X}_i^T \hat{\boldsymbol{\beta}}$ . The  $\hat{\sigma}_n^2(\hat{\mu}_i)$  and  $\hat{\boldsymbol{\beta}}$  are values maximizing the quasi-likelihood  $Q^\#(\hat{\boldsymbol{\mu}}, \hat{\mathbf{Y}}, \hat{\boldsymbol{\Sigma}})$  under  $H_1$ . We will describe how to calculate them in detail below.  $\hat{\mathbf{Y}}$  is composed of the elements  $\hat{Y}_i = Y_i \delta_i + \hat{E}(Y_i | Y_i > C_i)(1 - \delta_i)$ , the estimators of the synthetic data  $Y_i^*$ . Here  $\hat{E}(Y_i | Y_i > C_i)$  is an estimator of  $E(Y_i | Y_i > C_i)$ . To be specific, let  $r_i = (Z_i - \hat{\mu}_i)/\hat{\sigma}_n(\hat{\mu}_i)$ . If  $r(i)$  is the ordered  $r_i$ , then

$$\hat{F}(e) = 1 - \prod_{\{i: r(i) \leq e\}} ((n - i)/(n - i + 1))^{\delta_i}$$

When the largest observation is censored, we still treat it as uncensored (Meier 1975). Then  $\hat{E}(Y_i | Y_i > C_i) = \hat{\mu}_i + \hat{\sigma}_n(\hat{\mu}_i) \hat{E}(e_i | e_i > r_i)$  based on the Kaplan–Meier

estimate  $\hat{F}(e)$ . Note  $\hat{E}(Y_i | Y_i > C_i)$  depends on  $C_i$  through  $r_i = (Z_i - \hat{\mu}_i)/\hat{\sigma}_n(\hat{\mu}_i)$ , because when the observation is censored,  $Z_i = C_i$ .

To calculate  $Q(\tilde{\mu}, \tilde{\mathbf{Y}}, \tilde{\sigma}^2)$  under  $H_0$  (i.e., the data are homoscedasticity), iterative procedure as in Buckley and James (1979) is used to obtain  $\tilde{\beta}$  and  $\tilde{\mathbf{Y}}$ . Then we calculate  $\tilde{\mu}_i$  and  $\tilde{\sigma}^2$ . Therefore, the quasi-likelihood under  $H_0$  can be calculated as

$$Q(\tilde{\mu}, \tilde{\mathbf{Y}}, \tilde{\sigma}^2) = \frac{1}{n} \sum_{i=1}^n \int_{\tilde{Y}_i}^{\tilde{\mu}_i} \frac{\tilde{Y}_i - a}{\tilde{\sigma}^2} da.$$

To calculate  $Q^\#(\hat{\mu}, \hat{\mathbf{Y}}, \hat{\Sigma})$  under  $H_1$  (i.e., the data are heteroscedasticity), we estimate  $\beta$  by maximizing the quasi-likelihood  $Q^\#(\hat{\mu}, \hat{\mathbf{Y}}, \hat{\Sigma})$  under  $H_1$ , which is equivalent to solving the corresponding quasi-score function  $U_{\hat{\beta}} = o_p(1)$ , where

$$\begin{aligned} U_{\hat{\beta}} &= \mathbf{X}^T \hat{\Sigma}^{-1} (\hat{\mathbf{Y}} - \mathbf{X}^T \hat{\beta}) \\ &= \sum_{i=1}^n \frac{\mathbf{X}_i}{\hat{\sigma}_n(\hat{\mu}_i)} \left\{ \frac{\hat{Y}_i}{\hat{\sigma}_n(\hat{\mu}_i)} - \frac{\mathbf{X}_i^T}{\hat{\sigma}_n(\hat{\mu}_i)} \hat{\beta} \right\}, \\ &= \sum_{i=1}^n \mathbf{X}_{i,new} \{ \hat{Y}_{i,new} - \mathbf{X}_{i,new}^T \hat{\beta} \}. \end{aligned} \quad (4)$$

Here we define  $\hat{Y}_{i,new} = \hat{Y}_i/\hat{\sigma}_n(\hat{\mu}_i)$ ;  $\mathbf{X}_{i,new} = \mathbf{X}_i/\hat{\sigma}_n(\hat{\mu}_i)$ . Note  $U_{\hat{\beta}}$  is not monotone and is not continuous, so  $U_{\hat{\beta}} = 0$  may fail to have a solution. We are looking for the solution to make  $U_{\hat{\beta}}$  closest to zero. Therefore, we use  $U_{\hat{\beta}} = o_p(1)$ , not  $U_{\hat{\beta}} = 0$  (Ritov 1990).

To obtain  $\hat{\sigma}_n^2(\cdot)$ , we use the polynomial spline method to smooth the ‘observed’ variance values of  $\hat{\mathbf{Y}}$  over the mean. Specifically, we consider the variance function model,

$$\hat{\epsilon}_i^2 = \sigma^2(\hat{\mu}_i) + \tau_i, \quad i = 1, \dots, n,$$

where  $\hat{\epsilon}_i^2 = (\hat{Y}_i - \mathbf{X}_i^T \hat{\beta})^2$  is the ‘observed’ variance value,  $\tau_i$  is an error term with  $E(\tau_i) = 0$ . We use  $\hat{Y}_i$  instead of  $Z_i$  to compute the residual because  $\hat{Y}_i$  is the estimated complete data, i.e., when the observation is uncensored,  $\hat{Y}_i$  is equal to the true survival time; when the observation is censored,  $\hat{Y}_i$  is the expected true survival time for this observation. However,  $Z_i$  is the observed data, i.e., when the observation is uncensored,  $Z_i$  is equal to the true survival time; when the observation is censored,  $Z_i$  is the censored time, not the true survival time. Then the  $\sigma^2(\hat{\mu}_i)$  is expanded using B-spline basis functions as

$$\hat{\epsilon}_i^2 = \sum_{j=1}^J B_j(\hat{\mu}_i) \gamma_j + \tau_i, \quad i = 1, \dots, n,$$

where  $B_j$ ,  $j = 1, \dots, J$ , are the B-spline basis functions (De Boor 1978) and  $\gamma_j$ ,  $j = 1, \dots, J$ , are unknown spline coefficients. In the algorithm, we use a cubic B-spline

and use the *bs* function in *R* software to choose the quantiles as knots for a given number of knots. The number of knots can vary from 1 to 15 because it is adequate for a smooth function (Huang 2006). We select the number that minimizes the Akaike-like (AIC-like) criterion (Yu and Peace 2012) as follows

$$AIC - like = \sum_{i=1}^n \frac{\hat{Y}_i - \hat{\mu}_i}{\hat{\sigma}_n^2(\hat{\mu}_i)} + 2(J + 2 + d_1 - 1).$$

We use this criteria because Yu and Peace (2012) showed through simulations that the AIC-like criterion works well for the estimation of  $\beta$  and  $\sigma^2(\mu_i)$ . Now least squares method is applied to obtain  $\hat{\gamma}_j$ , estimator of  $\gamma_j$ . Hence, the  $\hat{\sigma}_n^2(\cdot)$  can be calculated by  $\hat{\sigma}_n^2(\cdot) = \sum_{j=1}^J B_j(\hat{\mu}_i)\hat{\gamma}_j$ .

The following iterative procedure is used to obtain  $\hat{\beta}$ ,  $\hat{Y}$  and  $\hat{\sigma}_n^2(\hat{\mu}_i)$  in the quasi-likelihood under  $H_1$ :

1. Initialize  $\beta$  with the least squares estimator  $\hat{\beta}^{(0)} = \tilde{\beta}$  proposed by Buckley and James (1979). The initial estimators of the variances are  $\hat{\sigma}_n^{2(0)}(\hat{\mu}_i) = 1, i = 1, \dots, n$  and  $\hat{\mu}_i = \mathbf{X}_i^T \hat{\beta}^{(0)}$ .
2. At the  $m$ th step, calculate the estimator of  $\hat{Y}_i$  as

$$\hat{Y}_i = \delta_i Y_i + (1 - \delta_i) \left( \mathbf{X}_i^T \hat{\beta}^{(m)} + \hat{\sigma}_n^{(m)}(\hat{\mu}_i) \int_{\hat{r}_i}^{\infty} \frac{td\hat{F}(t)}{1 - \hat{F}(\hat{r}_i)} \right)$$

where  $\hat{F}(e) = 1 - \prod_{\{i:\hat{r}(i) \leq e\}} \left( \frac{n-i}{n-i+1} \right)^{\delta_i}$ ,  $\hat{\mu}_i = \mathbf{X}_i^T \hat{\beta}^{(m)}$  and  $\hat{r}(i)$  are the ordered  $\hat{r}_i = \left( Z_i - \mathbf{X}_i^T \hat{\beta}^{(m)} \right) / \hat{\sigma}_n^{(m)}(\hat{\mu}_i)$ .

3. Update the variance estimator  $\hat{\sigma}_n^{2(m+1)}(\hat{\mu}_i)$  by smoothing the ‘‘observed’’ variance values  $\hat{\epsilon}_i^2 = \left( \hat{Y}_i - \mathbf{X}_i^T \hat{\beta}^{(m)} \right)^2$  over the estimated mean using polynomial spline method.
4. When variances are estimated, the coefficient estimators are updated by solving Eq. (4) using the least squares method:

$$\hat{\beta}^{(m+1)} = (\mathbf{X}_{new}^T \mathbf{X}_{new})^{-1} \mathbf{X}_{new}^T \hat{Y}_{new}$$

where  $\mathbf{X}_{new}$  is a matrix with elements  $\mathbf{X}_{i,new}$  and  $\hat{Y}_{new}$  is a vector with elements  $\hat{Y}_{i,new}$  defined in (4).

5. Repeat Steps (2)–(4) till  $|\hat{\beta}^{(m+1)} - \hat{\beta}^{(k)}| < d|\hat{\beta}^{(m+1)}|$  for any  $k \in \{0, 1, \dots, m\}$  and  $d$  is the prespecified convergence criterion. This stopping rule considers possible oscillation among iterations due to the discrete estimating function. The oscillation of the estimator means the estimator oscillating between two or more points, not converging to one point (Jin et al. 2006).

The values in the last step are used to calculate quasi-likelihood  $Q^\#(\hat{\boldsymbol{\mu}}, \hat{\mathbf{Y}}, \hat{\boldsymbol{\Sigma}})$  under  $H_1$  (i.e., the data are heteroscedasticity) as

$$Q^\#(\hat{\boldsymbol{\mu}}, \hat{\mathbf{Y}}, \hat{\boldsymbol{\Sigma}}) = \frac{1}{n} \sum_{i=1}^n \int_{\hat{Y}_i}^{\hat{\mu}_i} \frac{\hat{Y}_i - a}{\hat{\sigma}_n^2(a)} da. \quad (5)$$

However, we observe that the variance  $\hat{\sigma}_n^2(a)$  in the integral in (5) depends on  $a$  with the range in the integral from  $\hat{Y}_i$  to  $\hat{\mu}_i$ . Because we use nonparametric smoothing to estimate  $\hat{\sigma}_n^2(a)$ , it would be time-consuming in computation at each  $a$  value. Therefore, we propose a new form of quasi-likelihood as

$$Q(\boldsymbol{\mu}, \mathbf{Y}, \boldsymbol{\Sigma}) = \frac{1}{n} \sum_{i=1}^n \int_{Y_i}^{\mu_i} \frac{Y_i - a}{\sigma_n^2(\mu_i)} da. \quad (6)$$

Note the first derivative of (6) is (4), same as that in (3). In Theorem 1 in Wedderburn (1974), it was proved that (3) has properties similar to those of *log* likelihoods based on its score function. Now because (6) has the same score function as (3), (6) has similar properties to those of *log*-likelihood as well. Hence, the parameter estimators by maximizing (6) are same as those by maximizing (3). In addition, (6) has the same value as (3) under  $H_0$ . Now we use the quasi-likelihood under  $H_1$  based on (6) as follows

$$Q(\hat{\boldsymbol{\mu}}, \hat{\mathbf{Y}}, \hat{\boldsymbol{\Sigma}}) = \frac{1}{n} \sum_{i=1}^n \int_{\hat{Y}_i}^{\hat{\mu}_i} \frac{\hat{Y}_i - a}{\hat{\sigma}_n^2(\hat{\mu}_i)} da. \quad (7)$$

Then the quasi-likelihood ratio test statistic is

$$\begin{aligned} QLR &= -2 \times \frac{1}{n} \left( \sum_{i=1}^n \int_{\tilde{Y}_i}^{\tilde{\mu}_i} \frac{\tilde{Y}_i - a}{\tilde{\sigma}^2} da - \sum_{i=1}^n \int_{\hat{Y}_i}^{\hat{\mu}_i} \frac{\hat{Y}_i - a}{\hat{\sigma}_n^2(\hat{\mu}_i)} da \right) \\ &= -2 \times \frac{1}{n} \left( \frac{1}{\tilde{\sigma}^2} \sum_{i=1}^n \int_{\tilde{Y}_i}^{\tilde{\mu}_i} (\tilde{Y}_i - a) da - \sum_{i=1}^n \frac{1}{\hat{\sigma}_n^2(\hat{\mu}_i)} \int_{\hat{Y}_i}^{\hat{\mu}_i} (\hat{Y}_i - a) da \right) \\ &= \frac{1}{n} \left( \frac{1}{\tilde{\sigma}^2} \sum_{i=1}^n (\tilde{Y}_i - \tilde{\mu}_i)^2 - \sum_{i=1}^n \frac{1}{\hat{\sigma}_n^2(\hat{\mu}_i)} (\hat{Y}_i - \hat{\mu}_i)^2 \right) \\ &= \frac{1}{n} \sum_{i=1}^n \left( 1 - \frac{\hat{\epsilon}_i^2}{\sum_{j=1}^n \hat{a}_{ij} \hat{\epsilon}_j^2} \right) \\ &= \frac{1}{n} \sum_{i=1}^n \frac{\sum_{j=1}^n \hat{a}_{ij} \hat{\epsilon}_j^2 - \hat{\epsilon}_i^2}{\sum_{j=1}^n \hat{a}_{ij} \hat{\epsilon}_j^2} \end{aligned}$$

### Remarks:

1. We use same estimates of  $\boldsymbol{\beta}$  and  $\hat{\sigma}_n^2(\cdot)$  as in (5) to calculate (7).

2. Using (7) instead of (5) will change the power of the test. Note the quasi-likelihood under  $H_0$  is always equal to one. The difference between (7) and (5) is the integrand. (5) uses  $(\hat{Y}_i - a)/\hat{\sigma}_n^2(a)$  and (7) uses  $(\hat{Y}_i - a)/\hat{\sigma}_n^2(\hat{\mu}_i)$ . For different relationship between  $\hat{\sigma}_n^2(a)$  and  $\hat{\sigma}_n^2(\hat{\mu}_i)$ , (7) can be either larger or smaller than (5), in other words, (7) can be either closer to or further away from the quasi-likelihood value under  $H_0$  than (5), and hence either higher or lower test power.
3. Using (7) instead of (5) dramatically reduces the computational time. One reason is that we use the variances at  $\hat{\mu}_i$ , to avoid variations in its value in the integral. Another reason is that unlike (5), the integration of (7) has closed form, so we do not need to approximate the integration in the logarithm.

## 2.2 Bootstrap procedure for the critical values of the test

It is well known that for nonparametric smoothing test, the bootstrap method often provides more accurate approximation to the distribution of the test statistic than the asymptotic normal theory does when the sample size is small or moderate (Hardle and Mammen 1993). On this basis, we propose a bootstrap procedure in order to approximate the critical values of the test in a practical application of the QLR. Note we cannot generate bootstrap samples from the residuals  $\hat{\epsilon}_i$  and the estimated means  $\hat{\mu}_i$  from the main observed sample  $(Z_i, X_i, \sigma_i)$ . The reason is that if the data is heteroscedastic, the residuals  $\hat{\epsilon}_i, i = 1, \dots, n$  have different variances. Therefore, if we generate the bootstrap samples from the residuals  $\hat{\epsilon}_i$ , we are generating heteroscedastic data. This data cannot approximate the distribution of the test statistic under the null hypothesis, which assumes data is homoscedastic. We generated the bootstrap sample using the same way as in Wang and Zhou (2005), which consists of the following steps.

Denote  $B$  as the sufficient large number of bootstrapping (we used  $B = 200$  in the simulation study and the real data analysis since it performed very satisfactory). Therefore for  $b = 1, \dots, B$ ,

1. For  $i = 1, \dots, n$ ,

- Let  $Y_{i,b}^* = \mathbf{X}_i^T \tilde{\boldsymbol{\beta}} + \epsilon_i^*$ , where  $\epsilon_i^*$  is a bootstrap sample from centered  $\tilde{\epsilon}_i = (\tilde{Y}_i - \tilde{\mu}_i)$ ,
- For the *log* of the censoring time  $C_i, i = 1, \dots, n$ , some  $C_i$  are not observed when  $C_i > Y_i$ . We use a similar approach for  $Y_{i,b}^*$  to estimate the unknown  $C_i$ . To be specific, we use the model

$$C_i = \mathbf{X}_i^T \boldsymbol{\beta}_c + \epsilon_{ci}$$

where  $\epsilon_{ci}, i = 1, \dots, n$ , are independent and identically distributed with unspecified distribution. Then we use the least squares estimator (Buckley and James 1979) described in Sect. 2.2 to obtain the estimator  $\tilde{\boldsymbol{\beta}}_c$  for  $\boldsymbol{\beta}_c$ , and  $\tilde{C}_i$  for  $C_i$ . Then  $C_{i,b}^* = \mathbf{X}_i^T \tilde{\boldsymbol{\beta}}_c + \epsilon_{ci}^*$ , where  $\epsilon_{ci}^*$  is a bootstrap sample from centered  $\tilde{\epsilon}_{ci} = \tilde{C}_i - \mathbf{X}_i^T \tilde{\boldsymbol{\beta}}_c$

- Let  $Z_{i,b}^* = \min(Y_{i,b}^*, C_{i,b}^*)$  and  $\delta_{i,b}^* = I(Y_{i,b}^* \leq C_{i,b}^*)$ .



**Table 1** Empirical size of the tests

n	Censoring (%)	Normal	Extreme
100	10	0.057	0.042
100	30	0.047	0.046
200	10	0.055	0.047
200	30	0.053	0.048

2. With the resultant bootstrap sample  $\{(\mathbf{X}_i, Z_{i,b}^*, \delta_{i,b}^*), i = 1, \dots, n\}$ , calculate  $QLR_b$  for each bootstrapping sample following the procedures in Sect. 2.2.
3. Let  $QLR_\eta$  be the  $\eta$ th order statistic of  $QLR_1, \dots, QLR_B$ , then  $QLR_{|(1-\alpha)B|}$  approximates the  $(1 - \alpha)$ -quantile of the distribution of  $QLR$  under the null hypothesis.

### 3 Simulation studies

#### 3.1 Evaluate size (type-I error) of the test

We generated homoscedastic right censored data under  $H_0$  from the following model (8) to investigate the size of the test,

$$\log(T_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + 1/2 e_i \quad (8)$$

where  $\beta_0 = 0, \beta_1 = \beta_2 = 1$  and  $e_i$  are generated from the standard normal distribution or standard extreme value distribution with mean 0 and variance  $\pi^2/6$ .  $x_{i1}, i = 1, \dots, n$ , are simulated from the standard normal distribution and  $x_{i2}, i = 1, \dots, n$ , are simulated from the Binomial distribution with a success rate of 0.5. The  $\log$  censoring times  $C_i$  are simulated from uniform distribution  $U(0, \text{upper})$  and the upper is determined to yield 10% or 30% censoring observations. The sample size  $n$  is 100 or 200. The test is calculated with 1000 simulation runs and nominal level 0.05. We use  $B=200$  bootstrap samples per run to obtain the critical values. The results are summarized in Table 1 and it can be seen from this table that the proposed test satisfactorily maintains the specified nominal level.

#### 3.2 Evaluate power of the test

We generate heteroscedastic right censored data under  $H_1$  from the following models to investigate the power of the test. Let  $\mu_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$ ,

$$\begin{aligned} \text{Alternative I: } \log(T_i) &= \mu_i + 1/2(\mu_i)e_i \\ \text{Alternative II: } \log(T_i) &= \mu_i + 1/2(1/\mu_i)e_i \\ \text{Alternative III: } \log(T_i) &= \mu_i + 1/2(\mu_i + \mu_i^2)e_i \end{aligned}$$

The terms in the Alternatives are generated the same way as in (8), and we use the same procedure as for (8) to obtain the critical values. Figure 1 shows the variance

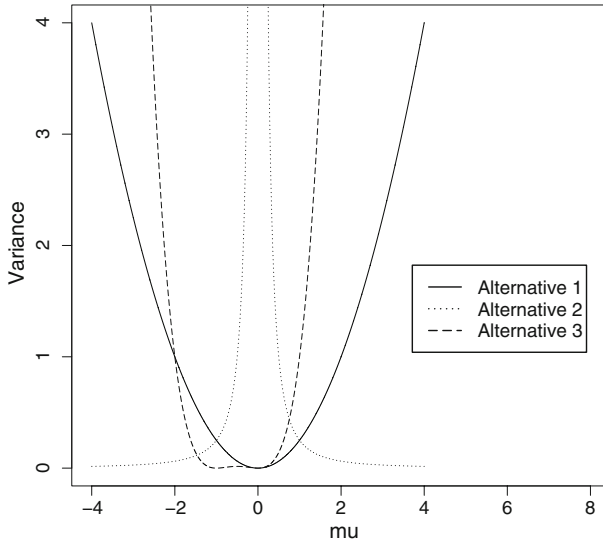


Fig. 1 The variance functions in Alternative hypotheses

Table 2 Empirical power of the tests for Alternative I

n	Censoring (%)	Normal	Extreme
100	10	0.625	0.497
100	30	0.202	0.275
200	10	0.861	0.677
200	30	0.354	0.325

functions in the Alternatives. The least heteroscedastic variance function is that in the Alternative I. The results are summarized in Table 2 (Alternative I), Table 3 (Alternative II) and Table 4 (Alternative III). These results show the test has satisfactory power, especially for large sample sizes and less censoring percentages. We observed that for the Alternative I (Table 2), the powers of the test decrease a lot when the censoring percentage is high (30%). Even when the sample size is 200 and censoring percentage is 30% and hence there are 140 uncensored observations, the power of the test is lower than the setting when sample size is 100 and censoring percentage is 10% and hence there are only 90 uncensored observations. The result indicates that when the data is closer to homoscedastic, the power of the test is lower for larger censoring percentage, even when the data has more information. Further research can be conducted to investigate how close to homoscedasticity and how large censoring percentage will have significant effect on the power of the test.

#### 4 Real data analysis

We applied the newly developed test to NCCTG Lung Cancer Data in R *survival* package. This dataset reflects the survival of patients with advanced lung cancer from

**Table 3** Empirical power of the tests for Alternative II

n	Censoring (%)	Normal	Extreme
100	10	0.634	0.568
100	30	0.597	0.529
200	10	0.698	0.672
200	30	0.686	0.666

**Table 4** Empirical power of the tests for Alternative III

n	Censoring (%)	Normal	Extreme
100	10	0.727	0.618
100	30	0.666	0.552
200	10	0.860	0.698
200	30	0.742	0.619

the North Central Cancer Treatment Group. There are totally 228 observations with 63 right censored.

We are interested in estimating the conditional mean of the *log* survival time (in days), given age (in years) and sex. Then the model is

$$Y_i \equiv \log(T_i) = \beta_0 + \beta_1 \times \text{age} + \beta_2 \times \text{sex} + \epsilon_i.$$

In order to analyze this data appropriately, we should first test the assumption of homoscedasticity of the data. We applied the test as proposed in Sect. 2 to this dataset and found that the associated *p* value was  $< 0.01$  indicating heterogeneity of variance. Graphically, we plotted the estimated variances versus the estimated means  $\mu$  (see in Fig. 2). It clearly shows heteroscedasticity in the data. Therefore we concluded that the homoscedasticity assumption was not satisfied for this dataset. Then the classical AFT model and associated methodologies (e.g. least squares method and rank method) might provide inaccurate conclusions. We should apply a method that considers the heteroscedasticity of the data, such as weighted least squares method (Yu et al. 2013).

We compare and report the results from the weighted least squares method (WLSE), the least square method (LSE) and the rank method (Rank) for the point estimates, their standard errors (SE) and 95% confidence intervals (Table 5). The point estimates from WLSM differ from those obtained using the LSM and rank method. The SEs from LSM and rank method are smaller than SEs from WLSE. Based on 95% confidence intervals, LSE results indicate that both age and sex are significant factors; rank results indicate that sex is the only significant factor; and WLSE results indicate that neither age and sex are significant. From the simulation study in Yu et al. (2013), when the data are heteroscedastic, both the LSM and rank method might have small SEs and the coverage probabilities of 95% confidence intervals are much lower than the nominal level 0.95. These findings indicate that these two methods might provide inaccurate conclusions.

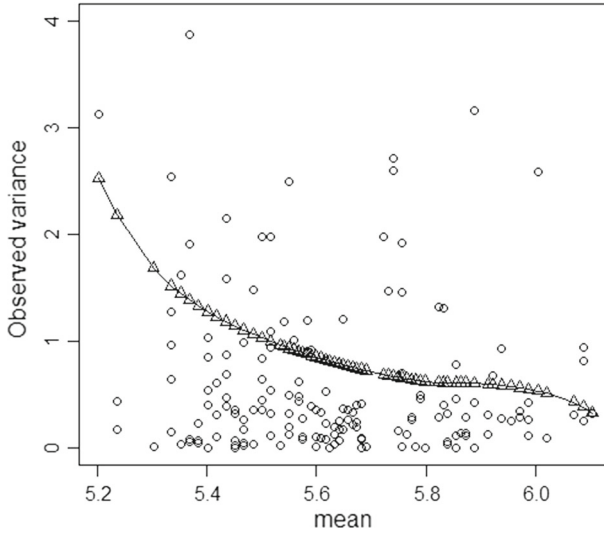


Fig. 2 Illustration of estimated variances versus estimated means: the curve is the estimated variance

Table 5 Real data analysis

	$\beta_1$ (age)			$\beta_2$ (sex)		
	Estimate	Variance	95% CI	Estimate	Variance	95% CI
LSM	-0.0249	0.00008	(-0.0424, -0.0074)	0.3652	0.0297	(0.0274, 0.7030)
Rank	-0.0122	0.00006	(-0.0274, 0.0029)	0.3308	0.0207	(0.0488, 0.6128)
WLSM	-0.0165	0.00008	(-0.0340, 0.0010)	0.2556	0.0316	(-0.0928, 0.6040)

## 5 Discussions and conclusions

This paper proposed the first homoscedasticity test for censored data for the AFT model. The simulation results showed the proposed test has satisfactory finite sample performance. The censoring has effect on the power of the test and indicates that the greater the percentage of censoring, the greater the number of observations needed to achieve high power. The real data analysis demonstrated the importance of the homoscedasticity test and that this test is necessary for accurate conclusions for the AFT model for survival data.

This novel test is a powerful quasi-likelihood ratio test for the homoscedasticity assumption in the AFT model for survival data. Quasi-likelihood was introduced by Wedderburn (1974) to describe a function that has similar properties as the log-likelihood function. It only requires assumptions on the first two moments which are much easier to postulate than the entire distribution of the data. In addition, we further simplified the quasi-likelihood by using fixed variance values in the integral. Given these properties, the quasi-likelihood ratio test has the potential to handle the complex-

ity posed by a censoring mechanism. This project opens a new research direction for the AFT model that will make the inferences based on the AFT model more accurate.

As emphasized, using (7) instead of (5) can increase or decrease the power of the test. The key advantage of using (7) is the substantial reduction in computational time. Future research should investigate the effects of using (7) on the size, power and computational time with different types of data.

If the test rejects the null hypothesis, we can find the pattern of the variances by plotting the estimated variances versus the mean. If it shows that the variance is a function of the mean, we can use weighted least squares method (Yu et al. 2013) to analyze the data. If it shows sub-group of subjects who do not follow the constant variance assumption, such as outliers, we may use least absolute deviations which is robust to the outliers. The ability to show the patterns of the variances is another advantage of this test.

The test we have proposed assumes variance is a function of the mean of the data. This assumption is reasonable because many distributions for survival data satisfy this assumption, such as exponential distribution, Weibull distribution, and extreme value distribution. Future research can relax this assumption to propose tests for settings of general nature. Moreover, this test can also be extended to the AFT model with nonlinear regression functions. In such a case, we need to first estimate the mean and variance in the quasi-likelihood based on the nonlinear function model. Then construct the corresponding quasi-likelihood ratio test. The goal of this paper was to fill the methodological gap by providing the first method and a new direction for homoscedasticity tests for the AFT model.

**Acknowledgements** We are grateful to the editor, associate editor, and two referees for their insightful comments, which significantly improved this manuscript.

## References

- Breusch TS, Pagan AR (1980) A simple test for heteroscedasticity and random coefficient variation. *Econometrica* 47:1287–1294
- Buckley J, James I (1979) Linear-regression with censored data. *Biometrika* 66(3):429–436
- Cai ZW, Hurvich CM, Tsai CL (1998) Score tests for heteroscedasticity in wavelet regression. *Biometrika* 85(1):229–234
- Cook RD, Weisberg S (1983) Diagnostics for heteroscedasticity in regression. *Biometrika* 70(1):1–10
- Cox DR (1972) Regression models and life-tables. *J R Stat Soc Ser B* 34:187–202
- De Boor C (1978) A practical guide to splines. Springer, New York
- Greene W (2008) *Econometric analysis*, 6th edn. Prentice Hall, Upper Saddle River
- Hardle W, Mammen E (1993) Comparing nonparametric versus parametric regression fits. *Ann Stat* 21(4):1926–1947
- Huang JZ (2006) Polynomial spline estimation and inference of proportional hazards regression models with flexible relative risk form. *Biometrics* 62:793–802
- Jin ZZ, Lin DY, Wei LJ, Ying ZL (2003) Rank-based inference for the accelerated failure time model. *Biometrika* 90(2):341–353
- Jin ZZ, Lin DY, Ying ZL (2006) On least-squares regression with censored data. *Biometrika* 93(1):147–161
- Kalbfleisch DJ, Prentice LR (1980) *The statistical analysis of failure time data*. Wiley, New York
- Lai TL, Ying ZL (1991) Large sample theory of a modified Buckley–James estimator for regression-analysis with censored-data. *Ann Stat* 19(3):1370–1402

- Lai TL, Ying ZL (1992) Linear rank statistics in regression-analysis with censored or truncated data. *J Multivar Anal* 40(1):13–45
- Lin DY, Ying ZL (1995) Semiparametric inference for the accelerated life model with time-dependent covariates. *J Stat Plan Inference* 44(1):47–63
- Meier P (1975) Estimation of a distribution function from incomplete observations. *Perspectives in probability and statistics*. Academic Press, London, pp 67–87
- Muller HG, Zhao PL (1995) On a semiparametric variance function model and a test for heteroscedasticity. *Ann Stat* 23(3):946–967
- Reid N (1994) A conversation with sir david cox. *Stat Sci* 9:439–455
- Ritov Y (1990) Estimation in a linear regression model with censored data. *Ann Stat* 18:303–328
- Robins J, Tsiatis AA (1992) Semiparametric estimation of an accelerated failure time model with time-dependent covariates. *Biometrika* 79(2):311–319
- Wang L, Zhou XH (2005) A fully nonparametric diagnostic test for homogeneity of variances. *Can J Stat* 33(4):545–558
- Wang L, Zhou XH (2007) Assessing the adequacy of variance function in heteroscedastic regression models. *Biometrics* 63(4):1218–1225
- Wedderburn RWM (1974) Quasi-likelihood functions, generalized linear-models, and Gauss–Newton method. *Biometrika* 61(3):439–447
- Ying ZL (1993) A large-sample study of rank estimation for censored regression data. *Ann Stat* 21(1):76–99
- Yu LL (2011) Nonparametric quasi-likelihood for right censored data. *Lifetime Data Anal* 17(4):594–607
- Yu LL, Peace KE (2012) Spline nonparametric quasi-likelihood regression within the frame of the accelerated failure time model. *Comput Stat Data Anal* 56(9):2675–2687
- Yu LL, Yu R, Liu L, Chen D (2012) Extended quasi-likelihood with fractional polynomials in the frame of the accelerated failure time model. *Stat Med* 31(13):1369–1379
- Yu LL, Liu L, Chen D (2013) Weighted least-squares method for right-censored data in accelerated failure time model. *Biometrics* 69(2):358–365
- Zhou M (2005) Empirical likelihood analysis of the rank estimator for the censored accelerated failure time model. *Biometrika* 92(2):492–498
- Zhou M, Kim M, Bathke A (2012) Empirical likelihood analysis for the heteroscedastic accelerated failure time model. *Stat Sin* 22:295–316