Quantile Regression for Residual Life and Empirical Likelihood

Mai Zhou  email: mai@ms.uky.edu
Department of Statistics, University of Kentucky, Lexington, KY 40506-0027, USA

Jong-Hyeon Jeong  email: jeong@nsabp.pitt.edu
Department of Biostatistics, University of Pittsburgh, Pittsburgh, PA 15261, USA

Mi-Ok Kim,  email:
Division of Biostatistics and Epidemiology, Cincinnati Children’s Medical Center, Cincinnati, OH 45229

Abstract

This approach has several advantages: (1) there is no need to estimate the variance/covariance at all, which may become prohibitively complicated for other procedures that requires the estimation of such. (2) When inverting the tests to obtain confidence regions/_intervals, this procedure inherits all the good properties of a likelihood ratio test. (3) Free software implementation of the test is readily available.

Key words: Censored data; Chi square distribution; Wilks theorem; Confidence interval.

1 Introduction

Jeong, Jung and Costantino in a recent Biometrics paper (2008) proposed a score type test for the median residual life time. They argue that “the need for such estimates is becoming more critical in breast cancer research as long-term courses of secondary therapies are now being considered for patients who remain recurrence free after several years of initial treatment”.

There are also a couple of recent papers dealing with the mean residual life time by Qin and Zhao (2007); Zhao and Qin (2006). The empirical likelihood approach proposed in this paper can also be used to give statistical inference for mean residual time without estimating any variances. Qin and Zhao, on the other hand, used a different approach: they used an estimating function that involves nuisance parameters. The version of empirical likelihood they proposed do not have a regular chi squared limiting distribution under null hypothesis, while ours does.
2 Median Residual Life

For a given CDF $F(t)$, the median residual lifetime at age $t_0$ is, by definition, the number $\theta = \theta(t_0)$ that solves the following equation:

$$\frac{1 - F(t_0 + \theta)}{1 - F(t_0)} = 0.5.$$ 

Other quantiles of the residual life distribution can be defined similarly. The development below can be easily modified to test a quantile other than median. But we shall focus on the median in most places of the paper.

Clearly $\theta$ is also the solution to $1 - F(t_0 + \theta) = 0.5[1 - F(t_0)]$. After easy calculation we see that $\theta$ is the solution to the equation $0 = F(t_0 + \theta) - 0.5F(t_0) - 0.5$.

Define a function $g_b(t)$ as

$$g_b(t) = I[t \leq (t_0 + b)] - 0.5I[t \leq t_0] - 0.5$$

then the hypothesis $H_0 : \theta(t_0) = b$ can be tested by testing

$$H_0 : \int_0^\infty g_b(t)dF(t) = 0.$$ 

This, in turn, can be accomplished by an empirical likelihood ratio test, given data from the distribution $F(t)$, see Owen (2001).

We now discuss the definition of the sample estimator, $\hat{\theta}_n$, of median residual life in the case where $n$ iid observations of lifetimes are available without censoring. Suppose we have an iid sample of lifetimes $Y_1, \ldots, Y_n \sim F(t)$. The estimating equation for the median residual time is, in view of the (1) above,

$$0 = \sum_{i=1}^n g_b(Y_i) = \sum_{i=1}^n \{I[Y_i \leq (t_0 + b)] - 0.5I[Y_i \leq t_0] - 0.5\}$$

In other words, the sample estimator of the median residual time based on the $n$ observations is $b = \hat{\theta}_n$ that solves this equation.

The above estimating equation can be simplified to (aside from a negative sign)

$$0 = \sum_{i=1}^n I[Y_i > t_0]\{0.5 - I[Y_i < (t_0 + b)]\}$$

(2)
or as a minimization problem

\[
\min_b \frac{1}{2} \sum_{i=1}^{n} I[Y_i > t_0]|Y_i - (t_0 + b)|.
\tag{3}
\]

Let \( \hat{F}_n(t) \) denote the empirical distribution based on \( Y_1, \ldots, Y_n \). The above estimating equation can also be written as

\[
0 = \int I[t > t_0]|0.5 - I[t < (t_0 + b)]d\hat{F}_n(t). \tag{4}
\]

**Remark:** For the purpose of \((\xi\text{-th})\) quantile regression, Koenker and Basset (1978) defined a so called “check function”,

\[
\rho_\xi(t) = t(\xi - I[t < 0]).
\]

We note that when \( \xi = 0.5 \) this is just the absolute value function: \( \rho_{0.5}(t) = 0.5|t| \). They also used its derivative

\[
\psi_\xi(t) = (\xi - I[t < 0]) \tag{5}
\]

to define a quantile regression estimating equation.

We point out that the \( g \) function we defined above is closely related to \( \psi \):

\[
-g_b(t) = I[t > t_0] \psi_{\xi=0.5}(t - (t_0 + b)) \cdot
\]

Obviously, the estimating equation for the \( \xi^{th} \) quantile for the residual life is

\[
0 = \sum_{i=1}^{n} I[Y_i > t_0]|\xi - I[Y_i < (t_0 + b)]| \text{ or } 0 = \int I[t > t_0]|\xi - I[t < (t_0 + b)]|d\hat{F}_n(t).
\]

Next suppose we do not observe all the \( Y_i \) but rather have right censored data \( U_i = \min(Y_i, C_i), \delta_i = I[Y_i \leq C_i] \) instead of \( Y_i \). The only modification to the above estimating equation (4) is to replace the empirical distribution \( \hat{F}_n(t) \) with the Kaplan-Meier estimator \( \hat{F}_{KM}(t) \) based on \((U_i, \delta_i)\) (Kaplan and Meier 1958). This motivates our definition of residual life quantile regression later.

### 3 Quantile Regression Models for Residual Lifetimes

We study in this section a Quantile Regression model with the residual lifetimes. The advantages of a general quantile regression model is discussed in Koenker (2006). We would like to
point out here that with censored data, the mean is often not well defined due to censoring in
the right tail of the distribution (long term survivors). Median (or other quantile) regression
is a particularly attractive choice in these cases.

We fix a time \( t_0 \) and focus on the residual life after \( t_0 \). Use all the notation of JJB (2008).

The model adopted in this paper is the following: we suppose the median of the conditional
distribution of \((Y_i - t_0)\) given \( Y_i > t_0 \) is \( \exp(\beta^T Z_i) \) i.e. the model
\[
P(Y_i - t_0 > \exp(\beta^T Z_i)|Y_i \geq t_0) = 0.5.
\]
This is equivalent to say \( P(\log(Y_i - t_0) > \beta^T Z_i|Y_i \geq t_0) = 0.5 \) or
\[
\text{median of } (\log(Y_i - t_0)|Y_i \geq t_0) = \beta^T Z_i
\]
where \( Z_i \) is the covariates associated with \( Y_i \).

But the above formulation is not quite enough. We need to be more specific on the error
structure. If we assume the error is additive for the log residual lifetimes, we define a residual
life quantile regression model as follows. Given \( Y_i > t_0 \), we have
\[
\log(Y_i - t_0) = \beta^T Z_i + \epsilon_i , \quad \text{given } Y_i > t_0 \quad (6)
\]
where \( \epsilon_i \) are independent random variables that has zero median (or \( \xi \)th quantile).

If the error is additive on the original residual lifetime, the model should then be modified
to, given that \( Y_i > t_0 \),
\[
Y_i - t_0 = \exp(\beta^T Z_i) + e_i , \quad i = 1, 2, \cdots n \quad (7)
\]
where \( e_i \) are independent random variables with zero \( \xi \)th quantile. (Do we need this para-
graph??)

Note here, \( \beta \) can depend on \( t_0 \), and can also depend on the particular quantile we are
modeling, \( \beta = \beta_{t_0}(\xi) \). We fix a time \( t_0 \) and only look at median, so here \( \beta \) is a constant. We
simply write \( \beta_{t_0}(\xi) = \beta \).

The above is our model. But the observations we have is the censored \( Y_i \): We observe
\( U_i = \min(Y_i, C_i) \), \( \delta_i = I[Y_i \leq C_i] \) along with the covariates \( Z_i \).

The rest of the discussion will be using median regression model. We define the estimator
\( \hat{\beta} \) for the model (8) by a minimization of the sample function
\[
\min_b \frac{1}{2} \sum_{i=1}^n w_i I[U_i > t_0] \mid \log(U_i - t_0) - b^T Z_i \mid \quad (8)
\]
where the weights \( w_i \) are the so-called inverse probability of censor weights (IPCW). See Bathke, Kim and Zhou (2008) for more discussion of this weights. Notice here whenever \( U_i \leq t_0 \) then the \( i^{th} \) term is zero due to the zero indicator function. In other words, \( I[U_i > t_0] \log(U_i - t_0) \) should be understood as zero whenever \( U_i \leq t_0 \).

**Discussion:** The IPCW \( w_i \) takes care of the censoring. The weights is zero for the censored observations. The median regression estimation of Bathke, Zhou and Kim (2008) is same as above except without the indicator \( I[U_i > t_0] \).

The minimization quantity for the second model (without log transformation) when defining the estimator is

\[
\min_b \frac{1}{2} \sum_{i=1}^{n} w_i I[U_i > t_0] | (U_i - t_0) - \exp(b^\top Z_i) | .
\]

This is closely related to the median residual life estimator we defined in section two.

However, the regression models traditionally take some transformation like the log, so we shall work with the log transformed model, (8), for the rest of the paper.

Next we formulate the equivalent estimating equation for the \( \hat{\beta} \). We first define the functions

\[
\psi_i^*(s, b, Z_i) = I[s > t_0] \{0.5 - I[\log(s - t_0) < b^\top Z_i]\}
\]

Then the estimating equation for the residual median regression is

\[
0 = V(b) = \sum_{i=1}^{n} w_i Z_i \psi_i^*(U_i, b, Z_i) = \sum_{i=1}^{n} w_i Z_i I[U_i > t_0] \{0.5 - I[\log(U_i - t_0) < b^\top Z_i]\}
\]

where the weights \( w_i \) are the IPCW discussed above. This can be considered as the formal derivative with respect to \( b \) of the function we minimize in (10).

We can use the general weighted quantile regression program in the R package `quantreg` of Koenker to solve the above equation, with the weights \( w_i I[U_i > t_0] \).

### 3.1 Empirical Likelihood Ratio Test

Next we discuss the empirical likelihood inference for the parameter \( \beta \).

The empirical likelihood ratio tests, first proposed by Thomas and Grunkemeier (1975) and Owen (1988), attracted many research attentions since then. The empirical likelihood methods developed in the last 20 years has emerged as a very competitive nonparametric test.
procedure for quite general settings, including the test of parameters defined by integration \( \int g(t) dF(t) \) or by estimating equations \( 0 = \int g(t, \theta) dF(t) \) with censored survival data. It parallels the theory of the parametric likelihood ratio test, except the parametric likelihood is replaced by a nonparametric one. The book of Owen (2001) summarized many of the results (Chapter 6 in particular). Other relevant papers include Murphy and van der Vaart (1997), Pan and Zhou (2002), Zhou (2005). The following is an adaptation/summary of the relevant results from above sources suitable for our applications.

Suppose \( Y_i, i = 1, 2, \ldots, n \) are iid survival times of interest with a distribution \( F(t) \). Due to censoring, we only observe a censored sample \( U_i = \min(Y_i, C_i) \) and \( \delta_i = I[Y_i \leq C_i] \). We assume the censoring \( C_i \) is independent of the survival time.

Let \( p_i \) denote the probability mass put on the observation \( U_i \), then the empirical likelihood for the censored data \( U_i \delta_i \), is defined as

\[
EL(p) = \prod_{i=1}^{n} \left\{ p_i \right\}^{\delta_i} \left\{ \sum_{U_j > U_i} p_j \right\}^{1-\delta_i} .
\]  

(12)

The maximization of the above EL with respect to \( p_i \), subject to \( p_i \geq 0 \) and \( \sum_{i=1}^{n} p_i = 1 \), is well known to be achieved by (the jumps of) the Kaplan-Meier estimator computed from \((U_i, \delta_i)\). Let us denote the maximum empirical likelihood value achieved as \( EL(KM) \).

In order to form the likelihood ratio, we also need to maximize the EL above with respect to \( p_i \) under an extra constraint equation:

\[
\sum_{i=1}^{n} g(U_i)p_i = \theta \quad \text{or} \quad 0 = \sum_{i=1}^{n} g(U_i, \theta)p_i
\]

(13)

where \( g(t) \) is a given function such that \( 0 < \text{Var} \ g(Y) < \infty \) and the \( \theta \) is the value we wish to test.

Assume

\[
\int \frac{g^2(t)}{1 - G(t)} dF(t) < \infty ,
\]

(14)

where \( G(\cdot) \) is the distribution of the censoring variable \( C_i \). This is to guarantee the variance of \( \int g(t) d\hat{F}_{KM}(t) \) is finite.

The Empirical Likelihood Theorem asserts that under the null hypothesis, \( H_0: \theta = \mathbb{E}g(T) \), then -2 log empirical likelihood ratio has an asymptotic chi squared distribution.
**Theorem** Consider the right censored data and its empirical likelihood defined above. Suppose \( \mathbb{E}_g(Y) = \theta \). Assume also that condition (6) holds. Then we have

\[
W(\theta) = -2 \log \frac{\max EL}{EL(KM)} \overset{D}{\to} \chi^2_1 \quad \text{as} \quad n \to \infty,
\]

where the maximum in the numerator is carried out over all probabilities \( p_i \) that satisfy (3).

A publicly downloadable software implementation of the empirical likelihood ratio tests with censored data is **emplik**. It is an extension package to be used with the R software (R Development Core Team 2008, http://cran.wustl.edu). In particular, the function **el.cen.EM2** inside the package **emplik** carries out the above test.

Due to the discrete nature of the quantile function, some have advocated for a smoothed quantile empirical likelihood ratio test (Chen and Hall 2001). One way of smoothing the residual median is to replace the indicator functions in (1) by a smoothed version.

A 90% confidence interval can be obtained as those \( \theta \) values that results a p-value larger then 0.1, etc.

**Remark 1**: Since **el.cen.EM2( )** can handle doubly censored data as well, the same test procedure outlined above can test median residual lifetime with doubly censored data. Left truncated and right censored data can be treated similarly. But we will have to use another function **emplikH2.test( )** inside **emplik** package and reformulate the hypothesis in terms of cumulative hazard.

**Remark 2**: The advantage of a likelihood ratio based confidence region, i.e. range respecting and transform invariant, is inherited in the empirical likelihood ratio inference.

For the residual median regression model, the constraint equation to consider when forming the empirical likelihood ratio (15) is then (in view of estimating equation (11))

\[
0 = \sum_{i=1}^{n} p_i Z_i \psi^*_i(U_i, b, Z_i).
\]

(15)

This is the so called AFT correlation model by Bathke, Kim and Zhou (2008). We have chi square limit theorem for the empirical likelihood ratio.

The Empirical Likelihood Theorem for the median residual lifetimes asserts that under the null hypothesis, \( H_0 : \beta = b \), then \(-2 \log \) empirical likelihood ratio has an asymptotic chi squared distribution.
Theorem Consider the right censored data and its empirical likelihood defined above. Suppose $\beta = b$. Assume also that $\int \frac{dF(t)}{1-G(t^-)} < \infty$. Then we have

$$W(\theta) = -2 \log \max \frac{EL}{EL(KM)} \overset{d}{\rightarrow} \chi_q^2 \quad \text{as} \quad n \rightarrow \infty,$$

where $q = \text{dim}(\beta)$ and the numerator $\max$ is carried out over all probabilities $p_i$ that satisfy (12).

Two special cases: (median regression) if the time $t_0 = 0$ (or it is the lower bound of $Y_i$) in the above definition of $\psi_i$, than this is a (censored) median regression. The empirical likelihood analysis of it were studied by Bathke, Kim and Zhou (2008) with the log transformation.

(residual median): If $Z_i \equiv 1$ (or constant), then the estimating equation corresponding to minimization (9) reverts back to the median residual life estimator we discussed in section 2 if the weights $w_i \equiv 1$ (no censor).

4 Examples and Simulations

4.1 Simulations

A simulation study was performed to compare the two sample testing procedure from Jeong et al. and one based on the empirical likelihood approach. For both of the two groups simulated, failure times were generated identically from the Weibull distribution with censoring proportions of 0%, 10%, 20% and 30% similarly as in Jeong et al. (2008). For a fair comparison, the non-smoothed version of the empirical likelihood ratio test was considered. The proportion of rejecting the null hypothesis of the equality of the two medians were compared for different sample sizes at various time points. Table 1 summarizes the results. In general, the results from Jeong et al.’s method tend to be slightly conservative than ones from the EL approach, especially when the sample size becomes large.
Table 1 Empirical coverage probabilities for the Jeong et al. (2008) and Empirical Likelihood method.

<table>
<thead>
<tr>
<th>$n$</th>
<th>$t_0$</th>
<th>0%</th>
<th>10%</th>
<th>20%</th>
<th>30%</th>
<th>0%</th>
<th>10%</th>
<th>20%</th>
<th>30%</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>1</td>
<td>.978</td>
<td>.978</td>
<td>.981</td>
<td>.976</td>
<td>.976</td>
<td>.975</td>
<td>.981</td>
<td>.979</td>
</tr>
<tr>
<td>50</td>
<td>2</td>
<td>.974</td>
<td>.974</td>
<td>.973</td>
<td>.977</td>
<td>.976</td>
<td>.975</td>
<td>.976</td>
<td>.974</td>
</tr>
<tr>
<td>50</td>
<td>3</td>
<td>.984</td>
<td>.986</td>
<td>.977</td>
<td>.979</td>
<td>.985</td>
<td>.989</td>
<td>.982</td>
<td>.985</td>
</tr>
<tr>
<td>100</td>
<td>0</td>
<td>.971</td>
<td>.970</td>
<td>.971</td>
<td>.977</td>
<td>.969</td>
<td>.969</td>
<td>.967</td>
<td>.972</td>
</tr>
<tr>
<td>100</td>
<td>1</td>
<td>.971</td>
<td>.973</td>
<td>.976</td>
<td>.979</td>
<td>.968</td>
<td>.969</td>
<td>.972</td>
<td>.977</td>
</tr>
<tr>
<td>100</td>
<td>2</td>
<td>.974</td>
<td>.976</td>
<td>.976</td>
<td>.978</td>
<td>.971</td>
<td>.973</td>
<td>.975</td>
<td>.975</td>
</tr>
<tr>
<td>100</td>
<td>3</td>
<td>.979</td>
<td>.981</td>
<td>.981</td>
<td>.982</td>
<td>.975</td>
<td>.976</td>
<td>.981</td>
<td>.981</td>
</tr>
<tr>
<td>500</td>
<td>0</td>
<td>.965</td>
<td>.966</td>
<td>.966</td>
<td>.968</td>
<td>.953</td>
<td>.952</td>
<td>.947</td>
<td>.957</td>
</tr>
<tr>
<td>500</td>
<td>1</td>
<td>.964</td>
<td>.966</td>
<td>.968</td>
<td>.968</td>
<td>.954</td>
<td>.957</td>
<td>.955</td>
<td>.958</td>
</tr>
<tr>
<td>500</td>
<td>2</td>
<td>.969</td>
<td>.969</td>
<td>.967</td>
<td>.970</td>
<td>.962</td>
<td>.961</td>
<td>.956</td>
<td>.960</td>
</tr>
<tr>
<td>500</td>
<td>3</td>
<td>.974</td>
<td>.972</td>
<td>.973</td>
<td>.969</td>
<td>.960</td>
<td>.958</td>
<td>.962</td>
<td>.959</td>
</tr>
</tbody>
</table>

4.2 Examples

We take the data set cancer from the R package survival. It contains 228 survival times of lung cancer patients from Mayo Clinic with 63 right censored observations.

We shall fit a median regression on residual lifetime after one year (365.25 days), using the age and sex of the patient as covariates.

We consider a lung cancer data set (Maksymiuk et al., 1994) that has been analyzed by Ying et al. (1995) using median regression, and by Huang et al. (2005) using a least absolute deviations method in the accelerated failure time (AFT) model. In this study, 121 patients with limited-stage small-cell lung cancer were randomly assigned to one of two different treatment sequences A and B, with 62 patients assigned to A and 59 patients to B. Each death time was either observed or administratively censored, and the censoring variable did not depend on the covariates treatment and age. Denote $X_{1i}$ the treatment indicator variable, and $X_{2i}$ the entry age for the $i$th patient, where $X_{1i} = 1$ if the patient is in group B. Let $Y_i$ be the base 10 logarithm of the $i$th patient’s failure time. The average survival time is ?? days

The data set is available in the R package rankreg.

When inverting the empirical likelihood ratio tests to get the confidence intervals, it is often very helpful to know where is the ‘center’ of that confidence interval, i.e. when testing
for this ‘center’ value, you should get a P-value of one. For the empirical likelihood ratio tests described in previous sections, the ‘center’ is given by the nonparametric maximum likelihood estimator based on the Kaplan-Meier estimator. This is given by the function `MMRtime` of the `emplik` package.

After some computation (see appendix) we find the 90% confidence interval for the median residual time is $[184.75, 321.75]$. Notice due to the discrete nature of the quantile function, we do not get exactly a P-value of 0.1. Smoothing the indicator function in (1) would enable us to always get P-values of 0.1. Another benefit of smoothing is (potentially) a more accurate P-value, as argued by Chen and Hall (1991). If we use the linear smooth function [bandwidth $1/20$], we get a 90% confidence interval $[184.74, 321.71]$. If we use the cubic smoother function [with bandwidth $1/20$] we get the 90% confidence interval $[184.77, 321.73]$. These intervals are practically the same as before. These are also very similar to the confidence interval obtained by the score test: $[184.75, 321.74]$.

The second example comes from a breast cancer study (NSABP Protocol B-04) as described in Jeong et al. (2008). The data includes ?? node positive patients and ?? node-negative patients. In this example, we first estimate the median residual lifetimes among node-positive and node-negative patients separately by using the empirical likelihood approach and then statistically compare them by using the 95% confidence intervals of the ratio estimated from both Jeong et al.’s (J) and the empirical likelihood (EL) ratio method. From Table 2 we see that the two approaches provided almost identical results for the 95% confidence intervals for the ratio of the two medians.

<table>
<thead>
<tr>
<th>$t_0$</th>
<th>Median Residual Lifetime</th>
<th>Ratio</th>
<th>95% CI</th>
<th>J</th>
<th>EL</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Node-Negative 12.46 (11.2,13.5) 6.87 (6.4,7.4) 0.55 (0.49, 0.63)</td>
<td>(0.49, 0.63)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Node-Positive 12.44 (11.2,13.6) 6.93 (5.9,12.1) 0.56 (0.47, 0.70)</td>
<td>(0.47, 0.70)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Node-Negative 13.05 (11.8,14.8) 8.24 (6.8,8.7) 0.63 (0.49, 0.81)</td>
<td>(0.49, 0.81)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Node-Positive 13.40 (12.5,14.3) 8.75 (7.7,10.6) 0.65 (0.54, 0.81)</td>
<td>(0.54, 0.81)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Node-Negative 12.91 (11.9,13.8) 10.19 (8.8,11.6) 0.79 (0.66, 0.93)</td>
<td>(0.66, 0.93)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Node-Positive 12.48 (11.2,13.7) 9.66 (8.2,11.8) 0.77 (0.62, 1.00)</td>
<td>(0.62, 1.00)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Node-Negative 11.85 (10.6,13.0) 9.66 (7.5,12.6) 0.82 (0.63, 1.08)</td>
<td>(0.63, 1.08)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3. Estimated median residual lifetimes in node-negative and node-positive groups, the ratios, and 95% confidence intervals for the ratios (NSABP B-04 data)
5 Discussion

6 Appendix

We list here R code for some of the computation in the examples.

Acknowledgements: Kim and Zhou’s Research supported in part by NSF grant DMS-0604920.

REFERENCES
