NON-PARAMETRIC LEAST SQUARE ESTIMATION OF DISTRIBUTION FUNCTION

Chai Genxiang¹  Hua Hong¹  Shang Hanji²

Abstract. By using the non-parametric least square method, the strong consistent estimations of distribution function and failure function are established, where the distribution function $F(x)$ after logit transformation is assumed to be approximated by a polynomial. The performance of simulation shows that the estimations are highly satisfactory.

§ 1 Introduction

Suppose that $X_1,X_2,\ldots,X_n$ is a sample from a population with distribution function $F(x)$. If $F(x)$ is assumed to have a known functional form with some unknown parameters $\theta_1,\theta_2,\ldots,\theta_r$, then the moment method can be used, but such an estimation is not effective approximately. This leads to a mixed moment method studied by Soong[1].

When $F(x)$ is completely unknown, the most directly and most extensively used estimation is the empirical distribution function $(F_n(x))$, which has given many famous large sample results.

When the functional form of $F(x)$ is unknown, but a polynomial on $[0, 1]$ such as $\theta_1x^1 + \ldots + \theta_rx^r$, can be used to approximate it, a mixed moment estimation method has been proposed by Zheng[2], and the strong consistency of estimation has been established. But the assumption in Zheng's method that the support set of $F(x)$ is an interval $[0, 1]$ narrows its applicable fields heavily.

It is assumed in this paper that the functional form of $F(x)$ is unknown, but the logit transformation of $F(x)$, that is $\log \frac{F(x)}{1-F(x)}$, can be approximated by a polynomial. Consequently, it need not to assume that the support set of $F(x)$ is $[0, 1]$, as is the case in [2].

The background for this paper arises from actuaries. In a mutual aid insurance for the aged, it is necessary to estimate the survival functions respectively when the ability of the
individual to provide for oneself is separated into the following three states: be able completely, be partly able and be completely unable. Samples are only available after age 60, but survival functions for different states before age 60 have to be estimated. Since the admissible sample intervals and the field of definition of survival functions are not coincide, hence the usual empirical distribution method may not be applicable.

Let 
\[ a_0 = \sup \{ x; F(x) = 0 \}, \quad b_0 = \inf \{ x; F(x) = 1 \}. \]

Hereafter it is assumed that \( F(x) \) is continuous everywhere, and \( \log \frac{F(x)}{1-F(x)} \) can be approximated by a polynomial on \( (a,b) \), where \( (a,b) \subseteq (a_0,b_0) \). The value of \( F(x) \), for \( x \in (a,b) \), is to be estimated based on sample \( X_1, X_2, \ldots, X_n \). It is known that if \( [a,b] \) is a finite close interval, \( \log \frac{F(x)}{1-F(x)} \) can be approximated by a polynomial. Therefore, the assumptions in this paper do not bring about much limitation.

If there exists a density function \( f(x) \) for \( F(x) \), the failure function \( \lambda(x) = \frac{f(x)}{1-F(x)} \) will be an important index for reliability. In this paper, we pay attention to be estimation of \( F(x) \), and at the same time the non-parametric estimation of \( \lambda(x) \) can also be obtained. The paper is organized as follows: we will discuss estimation methods in \( \S \ 2 \) and establish strong consistency for estimations in \( \S \ 3 \), while simulated results are offered in \( \S \ 4 \).

\section*{§ 2 Estimation method}

In this paper, we use the following basic assumptions:

(i) For \( F(x) \), there exists a density function \( f(x) \);

(ii) There exists a \( r \geq 1 \) (\( r \) being unknown) such that

\[
\log \frac{F(x)}{1-F(x)} = a_0 + a_1 x + \ldots + a_r x^r = J_r(x), \quad x \in (a,b),
\]

(1)

where \( (a,b) \subseteq (a_0,b_0) \).

As is asserted by Zheng\(^{(2)}\), the fact that \( r \) is unknown does not affect the execution of the estimation method.

Let \( X_{(1)} \ldots X_{(n)} \) denote the order statistics of \( X_1, X_2, \ldots, X_n \), then according to assumption (i), it is easy to assert

\[
EF(X_{(i)}) = \frac{i}{n+1}, \quad i = 1, \ldots, n.
\]

Let

\[
Y_i = \log \left( \frac{i}{n+1} \right), \quad i = 1, 2, \ldots, n,
\]

then \( Y \) may be seen as an approximation of \( \log \frac{F(X_{(i)})}{1-F(X_{(i)})} \), and its error sum of squares is