ASM Sc. J., 12, Special Issue 5, 2019 for ICoAIMS2019, 184-189

Rank Regression for Modeling Bus Dwell Time in the Presence of Censored Observations

Mostafa Karimi[1*] and Noor Akma Ibrahim[1]

[1]*Institute for Mathematical Research, Universiti Putra Malaysia, Malaysia*

Bus dwell time estimation is very important for public transport planners and bus operators. Modeling bus dwell time is challenging, both theoretically and computationally, in the presence of censored observations. Common linear regression models are parametric models that involve assumptions that are difficult to satisfy in applications. Rank regression based on the accelerated failure time model is a semiparametric model that does not involve assumptions about the model variables or the model error terms. Hence, this paper proposes rank estimators for modeling bus dwell time on the basis of Gehan and log-rank weight functions. An iterative algorithm is introduced that involves a monotone estimating function of the model parameter, and its minimization is a computationally simple optimization problem. A resampling technique is used for estimating the distribution of the rank estimator through its empirical distribution. The proposed methodology is performed on a real data set to assess the efficiency of the rank estimators in applications. The results illustrate that the proposed parameter estimators are fairly unbiased and censored observations do not significantly impact the performance of the rank estimators.

Keywords: bus dwell time; rank regression; censored data; accelerated failure time model

I. INTRODUCTION

The bus dwell time is defined as the time that is consumed by a bus at a scheduled bus stop without moving, which includes the time that the bus spends for passengers boarding and alighting and the time of opening and closing bus doors [1]. The bus dwell time is greatly important since it is essential for estimating bus station capacity [2], and it has been regarded as a significant component of bus travel time [3-5]. Moreover, the dwell time functions are fundamental in the analysis of the transit network reliability [6-9] as well as the transit assignment models [10-11]. Therefore, modeling and estimating bus dwell time is substantial for public transport designers as well as bus operators [12].

The earliest research on the estimation of bus dwell time was conducted by [13]. He modeled the bus dwell time by using the linear regression approach and considered the number of boarding and alighting passengers and the consumed time during the opening and closing bus doors as two primary contributing factors. Since then, a number of case studies were implemented to estimate the bus dwell time with respect to some secondary contributing factors. For instance, the relationship between the bus fare payment system and bus dwell time was investigated by [14]. The influence of bus floor types on the bus dwell time was examined by [15]. The impact of platform walking on bus rapid transit (BRT) stations on the bus dwell time was studied by [1]. The influence of fare collection technology in city bus services was analyzed by [5]. It has been well established by [16-17] that the most significant contributing factors to the bus dwell time are the number of boarding and alighting passengers and other parameters are the secondary contributing factors.

Since the dwell time for a bus in a bus stop is a time interval that begins once the bus arrives at the bus stop and ends once the bus departs the bus stop, the bus dwell time data could be considered as time-to-event data. In particular, the event of interest in such case is the bus departure from the bus stop, and the bus dwell time is the time-to-departure. Censored observations are very common in analyzing time-to-event data. A bus dwell time is censored if the actual bus departure time is unknown since the observation period ended before the bus departure time. Modeling time-to-event data through

*Corresponding author's e-mail: mostafa.karimi.ir@gmail.com

commonly used linear models and parametric approaches is quite challenging in the presence of censored observations.

Accelerated failure time (AFT) models are very suitable for modeling time-to-event data with censored observations [18]. The greatest advantage of AFT models that makes them appealing to researchers is that they are easy to interpret since they relate a set of independent variables to the logarithm of dwell time [19-20]. In order to estimate the AFT model parameters, Prentice [21] proposed rank estimators as a semiparametric inference procedure. Rank estimators based on the weighted log-rank statistics are computationally simple since they do not involve complex assumptions about the distribution of the model variables and error terms [22-25]. The asymptotic properties of the rank estimators were widely studied by [26-28] among others.

In the present paper, simple and authentic methods is provided for modeling bus dwell time through the AFT models by using the rank estimators. The proposed inference procedure is based on two well-known estimating functions for the rank estimators, which are the Gehan and the log-rank estimating functions. Solving the estimating equations with the Gehan weight function can be readily done through regular optimization techniques, since the Gehan estimating function is a monotone function of the model parameters. However, the log-rank estimating function is a step function of the model parameters and its corresponding estimating equation potentially has multiple roots. An iterative algorithm is used to solve the estimating equations that are corresponding to a class of monotone weighted log-rank estimating functions. Each iteration of the algorithm involves a standard optimization technique and it yields a consistent root of the estimating equations. Moreover, a resampling method is used to approximate the distribution of the rank estimators which leads to calculating the Wald statistics and constructing the confidence intervals. The proposed methodology is applied to a real data set to illustrate the efficiency and usefulness of the rank estimators and the resampling technique for modeling the bus dwell time.

II. ACCELERATED FAILURE TIME MODEL AND RANK ESTIMATORS

Let random variable T_i denote the bus dwell time and Z_i denote the $P \times 1$ associated vector of independent variables for the ith subject, where $i = 1,2,\dots,n$. Subjects are assumed to be independent. The accelerated failure time model is

$$\log(T_i) = \beta_0' Z_i + \varepsilon_i \ (i = 1,\dots,n) \quad (1)$$

In this regression model, β_0 is a $P \times 1$ vector of unknown regression parameters and ε_i's are independent random error terms with an unspecified distribution function. Let C_i denote the censoring time for the ith bus dwell time T_i. The data consist of $(\tilde{T}_i, \Delta_i, Z_i)$, where $\tilde{T}_i = \min(T_i, C_i)$, and $\Delta_i = 1$ when $T_i \le C_i$ and $\Delta_i = 0$ otherwise.

Define $\tilde{e}_i(\beta) = \log(\tilde{T}_i) - \beta' Z_i$, $\tilde{N}_i(t;\beta) = \Delta_i I\{\tilde{e}_i(\beta) \le t\}$ and $\tilde{Y}_i(t;\beta) = I\{\tilde{e}_i(\beta) \ge t\}$ when the indicator function $I\{A\}$ specifies whether condition A is true or not by the values 1 and 0, respectively. Define $S^{(0)}(t;\beta) = n^{-1}\sum_{i=1}^{n} \tilde{Y}_i(t;\beta)$ and $S^{(1)}(t;\beta) = n^{-1}\sum_{i=1}^{n} \tilde{Y}_i(t;\beta) Z_i$.

Regarding to the weighted log-rank estimating function for the AFT model [19], the estimating function for β_0 is given by

$$U_\phi(\beta) = \sum_{i=1}^{n} \Delta_i \phi(\tilde{e}_i(\beta);\beta)\{Z_i - \bar{Z}(\tilde{e}_i(\beta);\beta)\} \quad (2)$$

where ϕ is a specified weight function and $\bar{Z} = S^{(1)}/S^{(0)}$. The estimating function $U_\phi(\beta)$ is called the log-rank estimating function if $\phi = 1$ [29], and it is called the Gehan estimating function if $\phi = S^{(0)}$ [30]. The estimator of β_0, denoted by $\hat{\beta}_\phi$, is the solution of the estimating equation $U_\phi(\beta) = 0$. The random vector $n^{\frac{1}{2}}(\hat{\beta}_\phi - \beta_0)$ has asymptotic normal distribution with mean zero [27-28].

The main difficulty with solving the equation $U_\phi(\beta) = 0$ in case of log-rank weight function is that $U_\phi(\beta)$ is a step function of β and this non-monotonicity may result multiple solutions to the equation, especially with high-dimensional β. Such problems do not arise in case of the Gehan weight function, since the Gehan estimating function is a monotone function of the model parameter [31].

The Gehan estimating function for the model parameter takes the form of

$$U_G(\beta) = n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n} \Delta_i\{Z_i - Z_j\} I\{\tilde{e}_i(\beta) \le \tilde{e}_j(\beta)\} \quad (3)$$

Clearly $U_G(\beta)$ is the gradient in β of

$$L_G(\beta) = n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n} \Delta_i\{\tilde{e}_j(\beta) - \tilde{e}_i(\beta)\} I\{\tilde{e}_i(\beta) \le \tilde{e}_j(\beta)\} \quad (4)$$

Thus, the minimizer of $L_G(\beta)$ is equivalent to the solution of the equation $U_G(\beta) = 0$.

On the basis of the Gehan estimating function, [19] proposed a general weight estimating function. Consider the following modification of $L_\phi(\beta)$:

$$\tilde{L}_\phi(\hat{\beta};\beta) = n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n}\psi\left(\tilde{e}_i(\hat{\beta});\hat{\beta}\right)\Delta_i\{\tilde{e}_j(\beta) - \tilde{e}_i(\beta)\}I\{\tilde{e}_i(\beta) \leq \tilde{e}_j(\beta)\} \quad (5)$$

where $\psi = \phi/S^{(0)}$ and $\hat{\beta}$ is a preliminary estimator of β_0, such as $\hat{\beta}_G$.

Estimating the AFT model parameter β_0 could be implemented through an iterative algorithm. starting with an initial value of $\hat{\beta}^{(0)} = \hat{\beta}_G$, at the rth iteration of the, for $r \geq 1$

$$\hat{\beta}^{(r)} = \operatorname{Arg}\min_{\beta}\tilde{L}_\phi\left(\beta;\hat{\beta}^{(r-1)}\right) \quad (6)$$

For approximating the distribution of the rank estimator $\hat{\beta}$ a resampling technique similar to those of [32-33] is suitable to approximate the distribution of $\hat{\beta}$. To be specific, define a new function

$$\tilde{L}^*_\phi(\hat{\beta};\beta) = n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n}\psi\left(\tilde{e}_i(\hat{\beta});\hat{\beta}\right)\Delta_i\{\tilde{e}_j(\beta) - \tilde{e}_i(\beta)\}I\{\tilde{e}_i(\beta) \leq \tilde{e}_j(\beta)\}W_i \quad (7)$$

where W_i for $i = 1, \dots, n$ are independent random variables with positive values and $E(W_i) = Var(W_i) = 1$. The corresponding iterative algorithm is given by

$$\hat{\beta}^{*(r)} = \operatorname{Arg}\min_{\beta}\tilde{L}^*_\phi\left(\beta;\hat{\beta}^{*(r-1)}\right) \quad (8)$$

The distribution of $\hat{\beta}$ can be approximated through the empirical distribution of $\hat{\beta}^*$. To be precise, the random sample $(W_1, \dots, W_n)$ must be repeatedly generated while holding the data $(\tilde{T}_i, \Delta_i, Z_i)$ at their observed values for $i = 1, \dots, n$ to obtain a large number of realizations of $\hat{\beta}^*$. The limiting distribution of $n^{\frac{1}{2}}(\hat{\beta} - \beta_0)$ can be approximated by the limiting distribution of $n^{\frac{1}{2}}(\hat{\beta}^* - \beta_0)$ [33]. Wald statistics for hypothesis testing and the confidence intervals for $\hat{\beta}$ can be readily obtained from the empirical distribution of $\hat{\beta}^*$.

III. A CASE STUDY

The primary data for evaluating this study was collected at route U32, located at Kuala Lumpur City Centre, Malaysia, during October 2015 and November 2015 for buses that were operating under RapidKL public transport corporate. Figure 1 depicts the layout of route U32. This route is a high-frequency route with high passenger demand that passes across the most congested sections of the Kuala Lumpur City Centre. Bus dwell times were recorded through site observation. The data collection period was categorized into two categories including peak hour which was 6 AM to 9 AM and 4 PM to 7 PM, and non-peak hour which was 10 AM to 12 PM and 7 PM to 9 PM. A set of bus stops were selected along the route as the data collection sites and the observers were asked to record passenger movement, boarding, and alighting activities as well as on-board passengers.

Figure 1. Route U32 layout

The data set did not include censored observations. To demonstrate the proposed methodology and assess its performance in applications in the presence of censored observations, censored versions of the data set were generated. Specially, for generating the data sets with 10%, 20%, or 30% censored observations the bus dwell time values greater than or equal to 90, 80, or 70 quantiles were considered as censored, respectively. The summary of the data and the corresponding descriptive statistics are presented in Table 1.

The accelerated failure time models were fit for the bus dwell time with respect to peak hour, on-board passengers, and boarding passengers as independent variables. The regression model is given by

$$\log(T) = \beta_1 \times \text{Peak hour} + \beta_2 \times \text{on-board} + \beta_3 \times \text{Boarding} + \varepsilon \quad (9)$$

The accuracy of 0.0001 between successive estimates was set as the convergence criterion.

The standard error of the estimators, p-values, and confidence intervals were estimated by using the Wald statistics based on 100 resamples with standard exponential random variables satisfying $E(W) = Var(W) = 1$.

Table 2 summarizes the results of the rank regression model for the bus dwell time. In all cases the algorithm converged in less than 5 iterations. According to Table 2, for both Gehan and log-rank weight functions the parameter estimates are positive and since all the p-values are less than 0.05 all the parameter estimates are significant.

Table 1. Summary of descriptive statistics of the bus dwell time data

Censoring	Variable	Peak hour	N	Mean	SD
0%	Dwell time	AM & PM	180	115.197	45.276
		AM	90	127.970	49.749
		PM	90	102.425	36.303
	On-board	AM & PM	180	21.189	11.843
		AM	90	19.133	11.753
		PM	90	23.244	11.637
	Boarding	AM & PM	180	14.356	6.812
		AM	90	17.422	7.576
		PM	90	11.289	4.111
10%	Dwell time	AM & PM	162	105.942	37.163
		AM	77	115.226	41.224
		PM	85	97.532	30.968
	On-board	AM & PM	162	19.019	10.031
		AM	77	15.948	8.885
		PM	85	21.800	10.246
	Boarding	AM & PM	162	13.272	6.032
		AM	77	16.026	6.928
		PM	85	10.776	3.613
20%	Dwell time	AM & PM	144	98.345	31.986
		AM	63	103.400	35.883
		PM	81	94.414	28.199
	On-board	AM & PM	144	17.847	9.613
		AM	63	13.667	7.447
		PM	81	21.099	9.886
	Boarding	AM & PM	144	12.625	5.610
		AM	63	15.206	6.602
		PM	81	10.617	3.625
30%	Dwell time	AM & PM	126	91.249	27.601
		AM	51	92.906	31.664
		PM	75	90.123	24.621

Table 2. Accelerated failure time analysis for the bus dwell time

Censoring	Weight	Parameter	Estimate	SE	P-value
0%	Gehan	Peak hour	0.1707	0.0513	0.0004
		On-board	0.0205	0.0015	0.0000
		Boarding	0.0197	0.0045	0.0000
	Log-rank	Peak hour	0.2712	0.0163	0.0000
		On-board	0.0235	0.0009	0.0000
		Boarding	0.0106	0.0017	0.0000
10%	Gehan	Peak hour	0.1688	0.0552	0.0001
		On-board	0.0234	0.0023	0.0000
		Boarding	0.0236	0.0054	0.0000
	Log-rank	Peak hour	0.2924	0.0206	0.0000
		On-board	0.0276	0.0013	0.0000
		Boarding	0.0132	0.0026	0.0000
20%	Gehan	Peak hour	0.1694	0.0470	0.0002
		On-board	0.0244	0.0019	0.0000
		Boarding	0.0268	0.0061	0.0000
	Log-rank	Peak hour	0.3097	0.0226	0.0000
		On-board	0.0301	0.0018	0.0000
		Boarding	0.0152	0.0027	0.0000
30%	Gehan	Peak hour	0.1629	0.0420	0.0001
		On-board	0.0261	0.0028	0.0000
		Boarding	0.0317	0.0043	0.0000
	Log-rank	Peak hour	0.2982	0.0272	0.0000
		On-board	0.0319	0.0031	0.0000
		Boarding	0.0200	0.0025	0.0000

This means that for AM peak hour and for the bus stops with smaller number of on-board and boarding passengers the bus dwell time tends to be shorter, and for PM peak hour and for the bus stops with larger number of on-board and boarding passengers the bus dwell time tends to be longer. Although for larger percentage of censoring the parameter estimates are slightly greater than the uncensored data estimates, the difference is very small and negligible. Therefore, censoring does not have considerable impacts on the efficiency of the parameter estimators, since the estimated values are fairly close before and after censoring.

IV. CONCLUSION

Accelerated failure time models are suitable for modeling bus dwell time with respect to independent variables, specifically in the presence of censored observations. While parametric approaches are challenging due to their assumptions about the distribution of the variables and model error term, semiparametric methods are computationally simple alternatives with significantly reliable performance in applications. In this paper, rank regression was introduced as a semiparametric approach for estimating bus dwell time. Rank estimators were

considered with both Gehan and log-rank weight functions. On the basis of general weighted estimating functions, an iterative algorithm was proposed for estimating the model parameters that involves a simple optimization problem. A resampling technique was implemented to estimate the distribution of the proposed parameter estimators. The results of the real data analysis illustrated that the proposed methodology is completely reliable and efficient in applications. The parameter estimators were fairly unbiased, and the censored observations did not have considerable influences on the performance of the proposed estimators.

V. ACKNOWLEDGEMENT

This research was supported by Institute for Mathematical Research, Universiti Putra Malaysia. The authors are grateful to the reviewer's comments and suggestions.

VI. REFEREENCES

[1] Jaiswal, S., Bunker, J. & Ferreira, L., 2010, 'Influence of Platform Walking on BRT Station Bus Dwell Time Estimation: Australian Analysis', *Journal of Transportation Engineering*, 136(12): 1173–1179.

[2] Gu, W., Li, Y., Cassidy, M. J. & Griswold, J. B., 2011, 'On the capacity of isolated, curbside bus stops', *Transportation Research Part B: Methodological*, 45(4): 714–723.

[3] Hadas, Y. & Ceder, A. A., 2010, 'Optimal coordination of public-transit vehicles using operational tactics examined by simulation', *Transportation Research Part C: Emerging Technologies*, 18(6): 879–895.

[4] Lin, W. & Bertini, R. L., 2004, 'Modeling schedule recovery processes in transit operations for bus arrival time prediction', *Journal of Advanced Transportation*, 38(3): 347–365.

[5] Tirachini, A. & Hensher, D. A., 2011, 'Bus congestion, optimal infrastructure investment and the choice of a fare collection system in dedicated bus corridors', *Transportation Research Part B: Methodological*, 45(5): 828–844.

[6] Szeto, W. Y., Solayappan, M. & Jiang, Y., 2011, 'Reliability - Based Transit Assignment for Congested Stochastic Transit Networks', *Computer - Aided Civil and Infrastructure Engineering*, 26(4): 311 – 326.

[7] Szeto, W. Y., Jiang, Y., Wong, K. I. & Solayappan, M., 2013, 'Reliability-based stochastic transit assignment with capacity constraints: Formulation and solution method', *Transportation Research Part C: Emerging Technologies*, 35: 286–304.

[8] Yan, Y., Meng, Q., Wang, S. & Guo, X., 2012, 'Robust optimization model of schedule design for a fixed bus route', *Transportation Research Part C: Emerging Technologies*, 25: 113–121.

[9] Yan, Y., Liu, Z., Meng, Q. & Jiang, Y., 2013, 'Robust optimization model of bus transit network design with stochastic travel time', *Journal of Transportation Engineering*, 139(6): 625–634.

[10] Kepaptsoglou, K. & Karlaftis, M., 2009, 'Transit route network design problem: review', *Journal of transportation engineering*.

[11] Lam, W. H. K. & Bell, M. G. H., 2002, Advanced modeling for transit operations and service planning'

[12] Ceder, A., 2007, *Public transit planning and operation: theory, modeling and practice*, Elsevier, Butterworth-Heinemann.

[13] Levinson, H. S., 1983, *Analyzing transit travel time performance*, (No. 915).

[14] Guenthner, R. P. & Hamat, K., 1988, 'Transit dwell time under complex fare structure', *Journal of Transportation Engineering*, 114(3): 367–379.

[15] Levine, J. C. & Torng, G.-W., 1994, 'Dwell-time effects of low-floor bus design', *Journal of transportation engineering*, 120(6): 914–929.

[16] Milkovits, M. N., 2008, 'Simulating Service Reliability of a High Frequency Bus Route Using Automatically Collected Data by Master of Science in Transportation'.

[17] Tirachini, A., 2013, 'Bus dwell time: the effect of different fare collection systems, bus floor level and age of passengers', *Transportmetrica A: Transport Science*, 9(1): 28–49.

[18] Kalbfleisch, J.D. & Prentice, R.L., 2011, *The*

statistical analysis of failure time data, (Vol. 360). John Wiley & Sons.

[19] Jin, Z., Lin, D.Y., Wei, L.J. & Ying, Z., 2003, 'Rank - based inference for the accelerated failure time model', *Biometrika*, 90(2), pp.341-353.

[20] Peng, L. & Fine, J.P., 2007, 'Regression modeling of semicompeting risks data', *Biometrics*, 63(1), pp.96-108.

[21] Prentice, R.L., 1978, 'Linear rank tests with right censored data', *Biometrika*, 65(1), pp.167-179.

[22] Wang, Y.G. & Fu, L., 2011, 'Rank regression for accelerated failure time model with clustered and censored data', *Computational Statistics & Data Analysis*, 55(7), pp.2334-2343.

[23] Zhang, J. & Peng, Y., 2012, 'Semiparametric estimation methods for the accelerated failure time mixture cure model', *Journal of the Korean Statistical Society*, 41(3), pp.415-422.

[24] Chung, M., Long, Q. & Johnson, B.A., 2013, 'A tutorial on rank-based coefficient estimation for censored data in small-and large-scale problems', *Statistics and computing*, 23(5), pp.601-614.

[25] Chiou, S., Kang, S. & Yan, J., 2015, 'Rank - based estimating equations with general weight for accelerated failure time models: an induced smoothing approach', *Statistics in medicine*, 34(9), pp.1495-1510.

[26] Ritov, Y., 1990, 'Estimation in a linear regression model with censored data', *The Annals of Statistics*, pp.303-328.

[27] Tsiatis, A.A., 1990, 'Estimating regression parameters using linear rank tests for censored data', *The Annals of Statistics*, pp.354-372.

[28] Ying, Z., 1993, 'A large sample study of rank estimation for censored regression data', *The Annals of Statistics*, pp.76-99.

[29] Mantel, N., 1966, 'Evaluation of survival data and two new rank order statistics arising in its consideration', *Cancer Chemother. Rep.*, 50, pp.163-170.

[30] Gehan, E.A., 1965, 'A generalized Wilcoxon test for comparing arbitrarily singly-censored samples', *Biometrika*, 52(1-2), pp.203-224.

[31] Fygenson, M. & Ritov, Y.A., 1994, 'Monotone estimating equations for censored data', *The Annals of Statistics*, pp.732-746.

[32] Rao, C.R. & Zhao, L.C., 1992, 'Approximation to the distribution of M-estimates in linear models by randomly weighted bootstrap', *Sankhyā: The Indian Journal of Statistics*, Series A, pp.323-331.

[33] Parzen, M.I., Wei, L.J. & Ying, Z., 1994, 'A resampling method based on pivotal estimating functions', *Biometrika*, 81(2), pp.341-350.