Communications for Statistical Applications and Methods
2019, Vol. 26, No. 3, 325–336

https://doi.org/10.29220/CSAM.2019.26.3.325
Print ISSN 2287-7843 / Online ISSN 2383-4757

Application of discrete Weibull regression model with multiple imputation

Hanna Yoo[1,a]

[a]Department of Computer Software, Busan University of Foreign Studies, Korea

Abstract

In this article we extend the discrete Weibull regression model in the presence of missing data. Discrete Weibull regression models can be adapted to various type of dispersion data however, it is not widely used. Recently Yoo (*Journal of the Korean Data and Information Science Society*, **30**, 11–22, 2019) adapted the discrete Weibull regression model using single imputation. We extend their studies by using multiple imputation also with several various settings and compare the results. The purpose of this study is to address the merit of using multiple imputation in the presence of missing data in discrete count data. We analyzed the seventh Korean National Health and Nutrition Examination Survey (KNHANES VII), from 2016 to assess the factors influencing the variable, 1 month hospital stay, and we compared the results using discrete Weibull regression model with those of Poisson, negative Binomial and zero-inflated Poisson regression models, which are widely used in count data analyses. The results showed that the discrete Weibull regression model using multiple imputation provided the best fit. We also performed simulation studies to show the accuracy of the discrete Weibull regression using multiple imputation given both under- and over-dispersed distribution, as well as varying missing rates and sample size. Sensitivity analysis showed the influence of mis-specification and the robustness of the discrete Weibull model. Using imputation with discrete Weibull regression to analyze discrete data will increase explanatory power and is widely applicable to various types of dispersion data with a unified model.

Keywords: discrete count data, discrete Weibull regression model, missing data, multiple imputation, KNHANES

1. Introduction

Discrete count data arise in many fields of study including the social sciences, medical, and industrial fields. It refers to the number of times an event occurs within a certain period. A Poisson regression model is the ideal model for analyzing discrete count data. However, the Poisson regression model assumes the equality of the mean and variance, which makes its use limited with real data.

With real data, the dispersion is mostly over-dispersed or under-dispersed relative to the Poisson distribution. A negative Binomial regression is widely used to address over-dispersion, as well as the Poisson-inverse Gaussian model (Willmot, 1987). However, these models cannot deal with data that is highly skewed or zero-inflated data. In the case of under-dispersed distribution, a generalized Poisson regression model (Consul and Fanmoye, 1992) and the Conway-Maxwell Poisson models have been used (Sellers and Shmueli, 2010); more recently, Saez-Castillo and Conde-Sanchez (2013)

[1] Department of Computer Software, Busan University of Foreign Studies, 65 Geumsaem-ro 485 beon-gil, Geumjeong-gu, Busan 46234, Korea. E-mail: pinkcan78@bufs.ac.kr

Published 31 May 2019 / journal homepage: http://csam.or.kr

proposed a hyper-Poisson regression Poisson model to address non-normal, under-dispersed distribution. However, these models, which are modifications of Poisson regression models, are complex and computationally intensive (Chanialidis *et al.*, 2018).

Klakattawi *et al.* (2018) showed how a discrete Weibull regression model can be adapted in a simple way to address over-dispersed and under-dispersed data relative to a Poisson regression. Discrete Weibull distribution was first introduced by Nakagawa and Osaki (1975). There are only few papers that deal with discrete Weibull distributions. Khan *et al.* (1989) and Kulasekera (1994) showed parameter estimation of the discrete Weibull distribution and Englehardt and Li (2011) used discrete Weibull regression for correlated counts with confirmation for microbial counts in water. Moreover, Peluso and Vinciotti (2018) introduced a generalised additive discrete Weibull regression model and showed that it can provide a simple and unified framework to capture different levels of dispersion in the data.

In this paper we extend the discrete Weibull regression model to a case where the count data has missing values. Missing data is a common problem and is often encountered in many applications. In an example of count data with missing values, Pahel *et al.* (2011) performed multiple imputation methods to analyze dental caries data using a zero-inflated Poisson (ZIP) regression model. Moreover, Kleinke and Reinecke (2013) proposed a multiple imputation (MI) method for zero-inflated count data based on a Bayesian regression approach. In addition, Saffari and Adnan (2010) proposed a maximum likelihood method using the ZIP model with right-censored count data. However, there are not many studies that use a discrete Weibull distribution to address the issue of missing data. The merit of a discrete Weibull regression model is that it is appropriate for both over and under-dispersion count data. Imputing the missing data under a Weibull regression method enables the use of the entire dataset and provides significant empirical advantages to alternative methods, which require the omission of missing values.

Recently Yoo (2019) adapted the discrete Weibull regression in the presence of missing data using single imputation (SI). They compared the discrete Weibull regression with ZIP using SI. In the simulation result, using SI yield better results compared to the complete case (CC) analysis. In this paper, we extend the methods used in Yoo (2019) by employing MI method to impute the missing values. We also conduct a simulation study and compare the results using MI with SI and also with the results from a CC analysis, as well as a sensitivity analysis. We also set two different dispersion types and investigate the performance of MI using discrete Weibull regression. In this paper for comparison, two models (Negative Binomial, Poisson regression model) which are used frequently in discrete count data are also added. The remainder of the paper is organized as follows. In Section 2, we introduce the discrete Weibull regression model and briefly describe the MI method. In Section 3, we used the data from the 2016 seventh Korean National Health and Nutrition Examination Survey (KNHANES VII) and present the results. The results of the simulation study are summarized in Section 4. We then conclude the study with a brief discussion in Section 5.

2. Discrete Weibull regression model with multiple imputation

2.1. Discrete Weibull regression model

In this section we introduce the discrete Weibull regression model and its properties.

Let random variable Y follow a type 1 discrete Weibull distribution; we denote Y as

$$Y \sim \mathrm{DW}(q, \beta),$$

where $0 < q < 1$ and $\beta > 0$. The cumulative distribution function is given by

$$F(y; q, \beta) = \begin{cases} 1 - q^{(y+1)^\beta}, & y = 0, 1, 2, 3, \ldots, \\ 0, & y < 0. \end{cases}$$

The probability mass function is $f(y; q, \beta) = q^{y^\beta} - q^{(y+1)^\beta}$, for $y = 0, 1, 2, 3, \ldots$.

Since $f(0) = 1 - q$, parameter q is the probability that random variable Y has a value other than 0. Parameter β reflects the distribution skewness and controls the range of values of Y. Parameter β is especially connected to the variance ratio (VR) which is the ratio between the observed variance from the data and the theoretical variance from the model. If VR > 1, the data is said to be over-dispersed; if VR < 1, the data is said to be under-dispersed. If VR = 1, the data is said to be of equal dispersion. The data is over-dispersed if $0 < \beta \leq 1$ and under-dispersed when $\beta \geq 3$, regardless of the value q. If $1 < \beta < 3$, the data can be either over-dispersed or under-dispersed depending on the value of q. In the presence of covariates, consider a discrete Weibull regression model with linked functions of parameters q and β. Let random variable Y follow a discrete Weibull distribution with covariates $X_1, X_2, \ldots, X_p$ as

$$Y|X \sim \mathrm{DW}(q(X), \beta(X)),$$

where $X = (1, X_1, X_2, \ldots, X_p)$ is a vector with p covariates. The covariates are linked through $q(X)$ and $\beta(X)$. Peluso and Vinciotti (2018) conducted a simulation study linking parameters q and β to inspect a discrete Weibull regression model's level of flexibility. In this paper, we linked the covariates only through $q(X)$. The cumulative distribution function with covariates can be denoted as

$$F(y; q(X), \beta) = \begin{cases} 1 - q(X)^{(y+1)^\beta}, & y = 0, 1, 2, 3, \ldots, \\ 0, & y < 0. \end{cases}$$

and the probability mass function is given by

$$f(y; q(X), \beta) = q(X)^{y^\beta} - q(X)^{(y+1)^\beta}, \quad \text{for } y = 0, 1, 2, 3, \ldots,\ 0 < q(X) < 1,\ \beta > 0.$$

Considering the covariates, the discrete Weibull regression model can be denoted as follows:

$$\log(-\log(q(X)) = X'\boldsymbol{\theta}, \qquad \log(\beta) = \vartheta,$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)'$ is a vector composed of p regression coefficients and ϑ is the parameter of $\boldsymbol{\theta}$. In addition to the $\log(-\log)$ link function of $q(X)$, the logit link function can also be used. The log-likelihood is given as

$$L(y, x, \boldsymbol{\theta}, \vartheta) = \Pi_{i=1}^{n} f(y_i|x_i) = \Pi_{i=1}^{n} \left(q(x_i)^{y_i^\beta} - q(x_i)^{(y_i+1)^\beta} \right).$$

MLEs for the unknown parameters are obtained by maximizing the log-likelihood as follows:

$$l(y, x, \theta, \vartheta) = \sum_{i=1}^{n} \log \left(q(x_i)^{y_i^\beta} - q(x_i)^{(y_i+1)^\beta} \right).$$

Given the estimated parameters, the conditional expectation of $E(Y|X)$ can be obtained using a truncated support (Barbiero, 2015) through numerical approximations since there is no closed form. However, the conditional τ quantile can be easily derived given the covariates. The conditional τ quantile

of a DW distribution is the smallest integer $\mu_{(\tau)}(x)$ for which $P(Y \leq \mu_{(\tau)}(x)) = 1 - q(x)^{(y+1)^\beta} \geq \tau$. It can be derived as follows:

$$\mu_{(\tau)}(x) = \left\lceil \left(\frac{\log(1-\tau)}{\log(q(x))} \right)^{\frac{1}{\beta}} - 1 \right\rceil,$$

where $\lceil \cdot \rceil$ is the ceiling function. The log of the conditional τ quantile can be denoted as

$$\log\left(\mu_{(\tau)}(x) + 1\right) = \frac{1}{\beta}\log(-\log(1-\tau)) - \frac{1}{\beta}\log(-\log(q(x))).$$

Consider the special case where $\tau = 0.5$, $\mu_{(0.5)}(x)$ is the discrete Weibull distribution's conditional median, denoted as follows:

$$\mu_{(0.5)}(x) = \left\lceil \left(-\frac{\log(2)}{\log(q(x))} \right)^{\frac{1}{\beta}} - 1 \right\rceil.$$

Using the log transformation, the log of the conditional median can be denoted as

$$\log\left(\mu_{(0.5)}(x) + 1\right) = \frac{1}{\beta}\log(\log(2)) - \frac{1}{\beta}x'\boldsymbol{\theta}.$$

Thus, $-\theta_p/\beta$, $p = 1, \ldots, P$, is interpreted as the change in the median of the response variable given a one unit change of x_p holding all other covariates constant. Also $(1/\beta)\log(\log(2)) - (1/\beta)\theta_0$ is interpreted as the conditional median when all covariates are set to zero. Using a log transformation of the median is similar to the Poisson and negative Binomial models which also use the log of the mean.

2.2. Multiple imputation

The MI method is an attractive method for addressing missing data. It was first introduced by Rubin (1987) and involves a three-step approach in estimating a regression model with incomplete data. The first step is to impute the missing values using an appropriate model, which incorporates random variation. Usually, five to ten imputed datasets are created. The second step is to analyze the imputed data using standard methods similar to those used for a complete dataset. The last step is to combine the results for each imputed dataset and obtain pooled estimates. Compared with SI, a MI method makes it possible for the researcher to obtain approximately unbiased estimates of the parameters, and it can minimize standard error and increase efficiency of estimates. However, they are more difficult to perform.

In this study, for the MI method, MI-MICE, is used; it is an imputation method using Multivariate Imputation by Chained Equations (MICE). MICE is a software program used for imputing incomplete multivariate data by fully conditional specification. It appeared in the R package library in 2001. The fully conditional method specifies the multivariate imputation model on a variable-by-variable basis using a set of conditional densities, one for each incomplete variable (van Buuren, 2007). Using MICE, a Gibbs sampler is constructed from specified conditionals under the assumption that a multivariate distribution exists from conditional distributions. This method has different names, including regression switching (van Buuren *et al.*, 1999), variable-by-variable imputation (Brand, 1999), and chained equations. We compared the MI with predictive mean matching method which is one of the SI method and also with the CC analysis in the real data analysis and in the simulation study.

Table 1: Covariates' basic descriptions and missing values rates for the KHANES data

Variables	Categories	*n* (%) or mean (SD)	Missing rate
Sex	Male	3665 (45)	0.0%
	Female	4485 (55)	
Age		41.81 (22.98)	0.0%
Region	Town	6604 (81.0)	0.0%
	Country	1546 (19.0)	
Household income	Low	1407 (17.3)	0.4%
	Mid-low	2047 (25.1)	
	Mid-high	2316 (28.4)	
	High	2346 (28.8)	
Education level	Under elementary	2665 (32.7)	8.4%
	Middle School	820 (10.1)	
	High School	1881 (23.1)	
	University	2101 (25.8)	
Employment type	Management, Expert, Service etc.	2134 (26.2)	25.1%
	Agriculture, Assembler, etc.	1328 (16.3)	
	No job	2638 (32.4)	
Health status	Good	2868 (35.2)	7.8%
	Moderate	3399(41.7)	
	Bad	1274 (15.6)	
Alcohol	Yes	2237(27.4)	5.7%
	No	5443(66.8)	

3. Application to the seventh Korean National Health and Nutrition Examination Survey data (KNHANES VII, 2016)

To demonstrate the approach with real data, we used the most recent data from the seventh Korean National Health and Nutrition Examination Survey (KNHANES VII). KNHANES is a national program that provides statistics on the health and nutritional status of adults and children in Korea. It is based on the National Health Promotion Act, and the surveys have been conducted by the Korea Centers for Disease Control and Prevention (KCDC). Since 1988, the KNHANES has collected data via direct physical examination, clinical and laboratory tests, and related measurement procedures. These data are used widely to design health programs and services. For example, Lee *et al.* (2010) compared health behavior and health services of private medical insurance companies using the KNHANES data. In addition, Kim *et al.* (2008) utilized the KNHANES data to assess risk factors for predicting medical service use. There are many studies that use the KNHANES data, including empirical studies with discrete count dependent variables where Poisson regression is usually used. However, the assumptions that the Poisson regression model requires are difficult to satisfy using real data. In this study, we analyzed the KNHANES data to investigate the covariates that affect the 1month hospital stay using discrete Weibull regression.

Recently Yoo (2019) analyzed the same data, however in this paper we use different dependent variable which is the 1month hospital stay. We wanted to investigate which covariates affect the 1month hospital stay and we also added two more variables (health status, alcohol) which are known as important covariates that can affect hospital stays. Several of variables in the KNHANES contains missing values. The covariates we considered are sex, age, region, household income, education level, employment type, health status, and alcohol. Table 1 shows the basic description of each covariate and its missing values rate. Sex, age, and region had no missing values and household income had only a 0.4% missing rate. However, alcohol, education level and health status had relatively higher missing rates which were 5.7%, 8.4%, and 7.8%, respectively. Employment type showed the highest missing

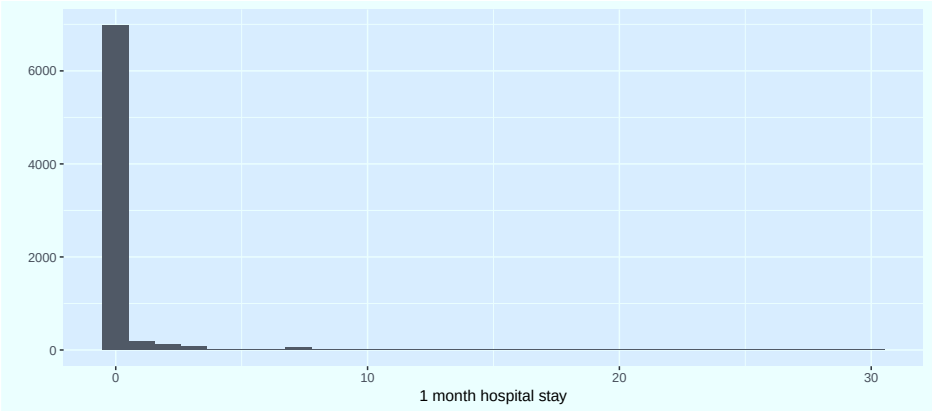


Figure 1: *Distribution of 1month hospital stay in KNHANES data.*

Table 2: Comparison of the maximum likelihood estimates

	Variables	DW_MI	DW_SI	Poi	NB	ZIP
Sex	Female	−0.1601*	−0.1692*	0.5484*	0.5444*	0.1067
Age		0.0019	0.0019	0.0024	0.0028	0.0109*
Region	Country	−0.0318	−0.0499	0.0124	0.1493	0.0047
Household income	Mid-low	0.2095*	0.2190*	−0.4955*	−0.4448	−0.0555
	Mid-high	0.2648*	0.2680*	−1.0357*	−0.8858*	−0.5641*
	High	0.2057*	0.2053*	−0.5665*	−0.5119	−0.1331
Education	Middle School	0.0744	0.0706	−0.4582*	−0.3931	−0.2930*
	High School	0.0475	0.0746	−0.1862	−0.3991	−0.0466
	University	0.0438	0.0188	−0.2551	−0.3691	−0.1412
Employment type	Agriculture, Assembler etc.	−0.0091	−0.0478	0.2732	−0.0721	0.3254
	No job	−0.1246*	−0.1265*	0.6955*	0.5321*	0.5785*
Alcohol	Yes	−0.0621	−0.0812	0.3578	0.1754	0.2372*
Health status	Moderate	−0.1674*	−0.1594*	0.6780*	0.3495*	0.2839*
	Bad	−0.6128*	−0.5874*	2.0956*	2.1038*	0.7936*
Other		$\beta = 0.3117^*$	$\beta = 0.3078^*$			
AIC		6252.49	6453.21	14586.28	6311.67	7940.39

$*$: p-value < 0.05. AIC = Akaike information criterion.
AIC of the KNHANES data using discrete Weibull with multiple imputation (DW_MI), discrete Weibull with single imputation (DW_SI), Poisson (Poi) and negative Binomial (NB) and zero-inflated Poisson (ZIP) regression model.

rate of 25.1%, and the 1month hospital stay had an 8.1% missing rate. The five-number summary of the 1month hospital stay variable was $(0, 0, 0, 0, 30)$; considering only the patients without missing values, 85.6% of the patients did not stay in a hospital during the 1month period.

Figure 1 shows the distribution of the 1month hospital stay. We can see the distribution is highly skewed to the right. Without imputing the missing values, statistical analysis is usually done with a list-wise deletion method; thus, due to missing values in this data, only 74.3% of patients could be included in the analysis. Thus, we imputed the missing values using MI and SI, applied the discrete Weibull regression model, and investigated which covariate affects the 1month hospital stay.

We compared the results from applying the discrete Weibull regression with the results from the Poisson, negative Binomial, and ZIP regression. All models were used using MI and in addition we also compared the result of discrete Weibull regression using SI. The maximum likelihood estimates and the Akaike information criterion (AIC) for each of the five different models are shown in Table 2.

The parameters of the discrete Weibull regression are reported with the parameterization linked to the log of the median; however, the other three regression models use the log of the mean of the response variable, so the interpretation of each model's results differs from those of the discrete Weibull regression. Thus, we compared the significance of each covariate of the five regression models and compared the AIC value to select the best model. The covariate's significance varies among the five regression models.

According to the AIC, the discrete Weibull model using MI provides the best fit, while the Poisson regression model produced the largest AIC. Based on the discrete Weibull model with MI, the median of the variable, 1month hospital stay, was significantly higher among women than men. Moreover, the results indicate that respondents of, low household income, who were unemployed, and those with moderate/bad health tend to stay in the hospital longer than other groups. Further, $\beta = 0.3117$ shows evidence of over-dispersion of the data compared to the Poisson regression.

4. Simulation studies

We conducted a simulation study to investigate the performance of discrete Weibull regression using MI and SI under various sample sizes and missing rates. We compared the accuracy of each setting by using the root mean squared error (RMSE) of the conditional quantile as below:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} \left(\hat{\mu}_i(\tau) - \mu_{i(\tau)}\right)^2}{n}},$$

where $\mu_{i(\tau)}$ is the true quantile and $\hat{\mu}_i(\tau)$ is the estimated quantile from the specified model.

We compared the results of the discrete Weibull model with three different models (Poisson, negative Binomial, ZIP) using MI and also compared the results using the SI (DW_SI) and CC analysis with the discrete Weibull model (DW_CC). Discrete count data were sampled from the discrete Weibull distribution, $Y \sim \text{DW}(q(X), \beta)$. We considered one covariate X with uniform distribution $(0, 2)$. We set $\log(-\log(q(X))) = -1.8 - 0.5X$ with $\beta = 2.5$ for over-dispersion and set $\log(-\log(q(X))) = -2.8 + 0.8X$ with $\beta = 3.7$ for under-dispersion. We considered sample sizes of $N = 50$ and 100 with missing rates of 10% and 40%. We assumed the missing mechanism as MAR. Therefore, the missing values of covariates only depend on observed values. Table 3 and Table 4 show the result of RMSE for three different quantiles, $\tau = 0.25, 0.5, 0.75$, with sample size $N = 50, 100$, and missing rates 10%, 40% for the over-dispersed and under-dispersed data, respectively.

Considering the case of over-dispersed data, DW_MI showed the smallest RMSE in all sample sizes and missing rates. Comparing DW_MI and DW_SI, as the sample size decreases and missing rate increases the DW_SI showed rather poor performance. Also when comparing the DW under MI, SI and CC, DW_CC always showed the worst performance. This result suggests that even if the underlying model is correct, deleting the missing values can give inefficient estimates. Moreover, the results show that the RMSE of DW_CC is larger than that of the mis-specified models (Poisson, NB, ZIP) in most cases and reveals the importance of imputing missing values.

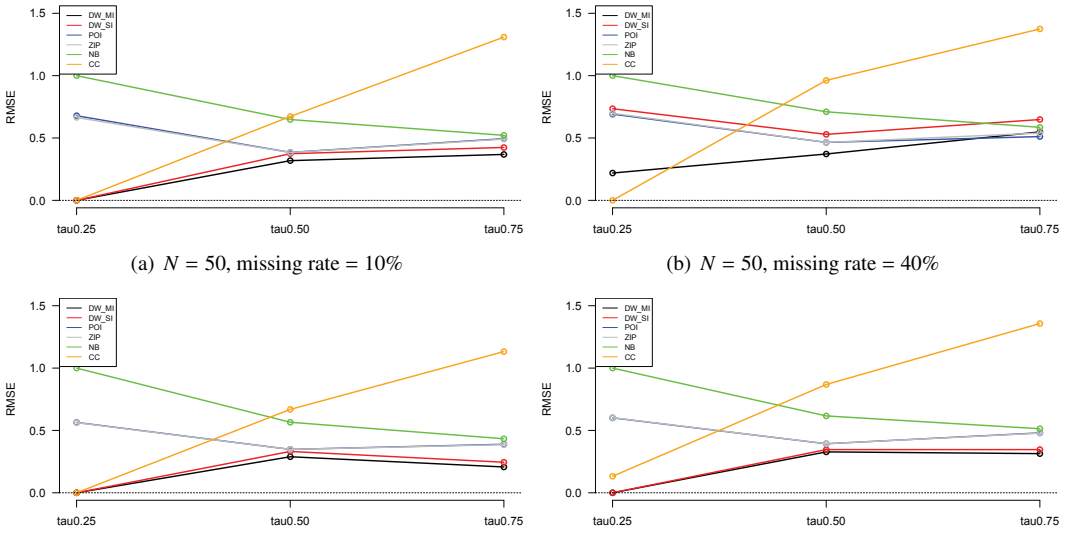
Figure 2 and Figure 3 graphically summarize the results from Table 3 and Table 4, respectively. As the figures show, the RMSE decreases as the sample size increases and the missing rate decreases.

The simulation of the under-dispersed case produced similar results to those of the over-dispersed scenario; that is, DW with MI outperforms the other models. We can see that DW_CC performs poorly compared to DW_MI or DW_SI and also has relatively bigger RMSE compared to other mis-specified models (Poisson, NB, ZIP).

Table 3: Comparison of different models in terms of RMSE on over-dispersed data simulated from DW model with $N = 50, 100$ and missing rates 10%, 40%

Sample size	Missing rate	Quantile	DW_MI	DW_SI	Poisson	ZIP	NB	DW_CC
$N = 50$	10%	$\tau = 0.25$	0.0000	0.0000	0.6787	0.6646	1.0000	0.0000
		$\tau = 0.50$	0.3184	0.3741	0.3847	0.3839	0.6480	0.6725
		$\tau = 0.75$	0.3682	0.4242	0.4947	0.4909	0.5214	1.3093
	40%	$\tau = 0.25$	0.2190	0.7348	0.6914	0.6981	1.0000	0.0000
		$\tau = 0.50$	0.3716	0.5291	0.4647	0.4647	0.7101	0.9623
		$\tau = 0.75$	0.5489	0.6480	0.5113	0.5396	0.5856	1.3744
$N = 100$	10%	$\tau = 0.25$	0.0000	0.0000	0.5644	0.5626	1.0000	0.0000
		$\tau = 0.50$	0.2889	0.3316	0.3480	0.3480	0.5657	0.6686
		$\tau = 0.75$	0.2070	0.2449	0.3891	0.3859	0.4334	1.1324
	40%	$\tau = 0.25$	0.0000	0.0000	0.6010	0.6026	1.0000	0.1324
		$\tau = 0.50$	0.3278	0.3464	0.3932	0.3936	0.6164	0.8686
		$\tau = 0.75$	0.3141	0.3464	0.4804	0.4782	0.5136	1.3572

RMSE = root mean squared error; DW_MI = discrete Weibull with multiple imputation; DW_SI = discrete Weibull with single imputation; ZIP = zero-inflated Poisson; NB = negative Binomial; DW_CC = complete case analysis with the discrete Weibull model.



(a) $N = 50$, missing rate = 10%

(b) $N = 50$, missing rate = 40%

(c) $N = 100$, missing rate = 10%

(d) $N = 100$, missing rate = 40%

Figure 2: *RMSE of six different models given over dispersion with* $N = 50, 100$ *and missing rates 10%, 40%.*

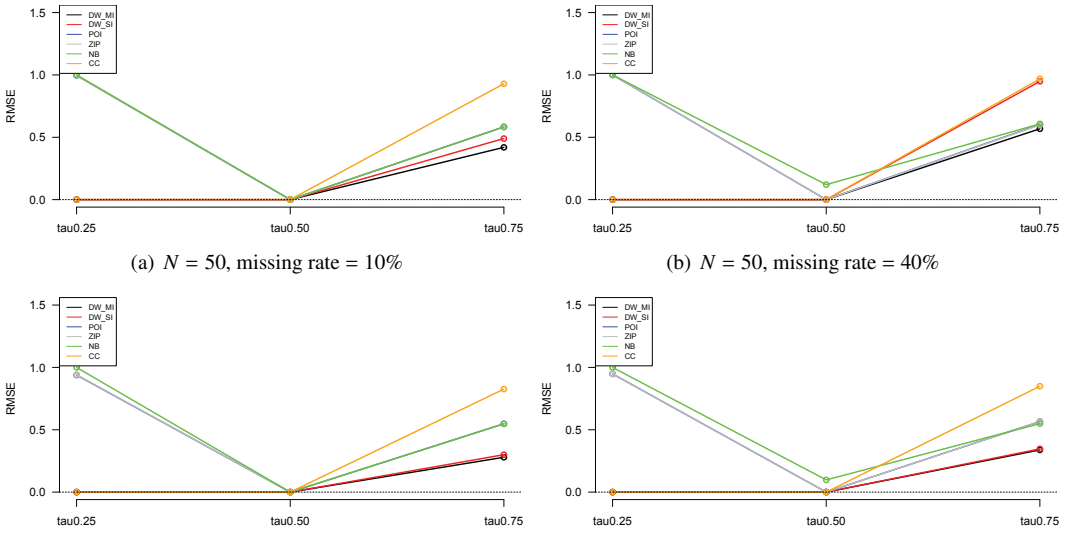
To further check the robustness of DW against mis-specified distributions, a sensitivity analysis was conducted. We simulated data with Poisson distribution with mean (μ) linked to the covariates through $\log(\mu(X)) = -0.2 + 0.4X$ with $X \sim \text{Uniform}(0, 1)$. Sample size was set to $N = 50, 100$ and missing rate as 10% and 40%. The RMSE are shown in Table 5 and summarized in Figure 4 for the discrete Weibull, Poisson, negative Binomial and ZIP models.

The results show that the RMSE of DW_MI is similar, and even lower in some simulated settings, compared with the Poisson model. DW_SI also shows favorable results however comparing with the MI, DW_MI shows better performance. This suggests the usefulness of the DW regression model and that it is robust to mis-specified distributions. Given real data, it is difficult to satisfy certain conditions

Table 4: Comparison of different models in terms of RMSE on under-dispersed data simulated from DW model with $N = 50, 100$ and missing rates 10%, 40%

Sample size	Missing rate	Quantile	DW_MI	DW_SI	Poisson	ZIP	NB	DW_CC
$N = 50$	10%	$\tau = 0.25$	0.0000	0.0000	0.9959	0.9959	1.0000	0.0000
		$\tau = 0.50$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
		$\tau = 0.75$	0.4193	0.4898	0.5830	0.5830	0.5830	0.9293
	40%	$\tau = 0.25$	0.0000	0.0000	1.0000	1.0000	1.0000	0.0000
		$\tau = 0.50$	0.0000	0.0000	0.0000	0.0000	0.1200	0.0000
		$\tau = 0.75$	0.5681	0.9486	0.6022	0.5995	0.6074	0.9682
$N = 100$	10%	$\tau = 0.25$	0.0000	0.0000	0.9387	0.9410	1.0000	0.0000
		$\tau = 0.50$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
		$\tau = 0.75$	0.2792	0.3000	0.5477	0.5477	0.5477	0.8260
	40%	$\tau = 0.25$	0.0000	0.0000	0.9478	0.9478	1.0000	0.0000
		$\tau = 0.50$	0.0000	0.0000	0.0000	0.0000	0.0978	0.0000
		$\tau = 0.75$	0.3381	0.3464	0.5656	0.5656	0.5490	0.8488

RMSE = root mean squared error; DW_MI = discrete Weibull with multiple imputation; DW_SI = discrete Weibull with single imputation; ZIP = zero-inflated Poisson; NB = negative Binomial; DW_CC = complete case analysis with the discrete Weibull model.



(a) $N = 50$, missing rate = 10%

(b) $N = 50$, missing rate = 40%

(c) $N = 100$, missing rate = 10%

(d) $N = 100$, missing rate = 40%

Figure 3: *RMSE of six different models given under dispersion with* $N = 50, 100$ *and missing rates 10%, 40%.*

using a specific model. In such cases, a DW regression model can give desirable results.

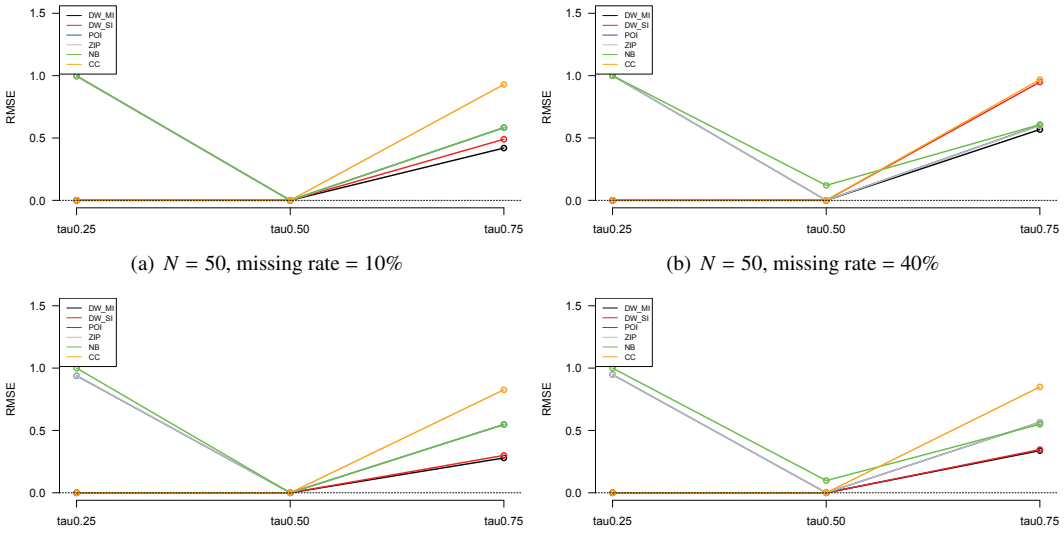
5. Discussion

In this article we extend the discrete Weibull regression model using MI in the presence of missing values in discrete count data. The principle merit of using a discrete Weibull regression model is that it can capture different levels of dispersion in the data. We adopt this discrete Weibull model to the situation where missing values occur. We used multiple and SI to impute the missing values and compared the performance of the discrete Weibull model with the Poisson, negative-Binomial,

Table 5: Comparison of different models in terms of RMSE simulated from Poisson model with $N = 50, 100$ and missing rates 10%, 40%

Sample size	Missing rate	Quantile	DW_MI	DW_SI	Poisson	ZIP	NB
$N = 50$	10%	$\tau = 0.25$	0.0000	0.0000	0.0000	0.0000	0.0000
		$\tau = 0.50$	0.3438	0.4242	0.4536	0.5547	0.8528
		$\tau = 0.75$	0.5599	0.5656	0.5669	0.5273	0.6179
	40%	$\tau = 0.25$	0.0000	0.0000	0.0000	0.0000	0.0000
		$\tau = 0.50$	0.3780	0.5477	0.7245	0.7245	1.0000
		$\tau = 0.75$	0.7348	0.7348	0.7348	0.7348	0.7428
$N = 100$	10%	$\tau = 0.25$	0.0000	0.0000	0.0000	0.0000	0.0000
		$\tau = 0.50$	0.0000	0.0000	0.0000	0.0000	0.7665
		$\tau = 0.75$	0.4266	0.4472	0.4184	0.4613	0.4330
	40%	$\tau = 0.25$	0.0000	0.0000	0.1378	0.1315	0.0000
		$\tau = 0.50$	0.4889	0.5000	0.4917	0.5489	0.7983
		$\tau = 0.75$	0.4934	0.5196	0.4952	0.4935	0.5298

RMSE = root mean squared error; DW_MI = discrete Weibull with multiple imputation; DW_SI = discrete Weibull with single imputation; ZIP = zero-inflated Poisson; NB = negative Binomial; DW_CC = complete case analysis with the discrete Weibull model.



(a) $N = 50$, missing rate = 10%

(b) $N = 50$, missing rate = 40%

(c) $N = 100$, missing rate = 10%

(d) $N = 100$, missing rate = 40%

Figure 4: *RMSE of five different models under Poisson model with* $N = 50, 100$ *and missing rates 10%, 40%.*

and ZIP, which are common models that are frequently used in count data. We applied the discrete Weibull regression model to the KNHANES VII data after imputing the missing values; we also compared the model's adequacy with other models and showed that the DW model with MI shows good performance.

Based on the simulation results of this study, the discrete Weibull regression model showed its wide applicability to modeling over- and under-dispersed count data; it also proved to be robust to mis-specified distributions. To assess the effect of imputation, we compared the results of the discrete Weibull model both with and without using imputation. The results showed that even though the underlying model is correct, deleting the missing values can lead to inefficient estimates. We also

compared the results of multiple and SI. SI with discrete Weibull regression model showed rather good performance compared to other models however when compared to MI, DW_MI always showed better performance in all sample sizes and missing rates.

While Poisson, negative-Binomial and ZIP regression are the most widely used models for analyzing count data, the discrete Weibull regression model can be a more attractive model producing more efficient and robust results.

In this paper we used MICE as a MI method. However, the models' performance can be affected by different imputation methods, missing rates, distribution of variables, and the missing-data mechanism. Thus, further study will be needed extending the simulation settings. Also we only considered predictive mean matching for SI method. We can expand this to several SI methods and compare the results with several MI methods. In addition, the KNHANES VII data used in this study contains complex survey data, where participants are selected according to stratified multistage cluster sampling. This characteristic of complex design should have been considered in our model to obtain accurate results; however, in this article we only focused on the application of the discrete Weibull model. Extension of this model to complex survey data is an opportunity for further study.

Acknowledgement

This work was supported by the research grant of the Busan University of Foreign Studies in 2019.

References

Barbiero A (2015). Discrete Weibull: Discrete Weibull Distributions (Type 1 and 3). Available from: http://CRAN.R-project.org/package=DiscreteWeibull. R package version 1.0.1

Brand JJPL (1999). *Development, Implementation and Evaluation of Multiple Imputation Strategies for the Statistical Analysis of Incomplete Data Sets*, Erasmus University, Rotterdam.

Chanialidis C, Evers L, Neocleous T, and Nobile A (2018). Efficient Bayesian inference for COM-Poisson regression models, *Statistics and Computing*, **28**, 595–608.

Consul P and F Famoye (1992). Generalized Poisson regression model, *Communications in Statistics-Theory and Methods*, **21**, 89–109.

Englehardt JD and R Li (2011). The discrete Weibull distribution: an application for correlated counts with confirmation for microbial counts in water, *Risk Analysis*, **31**, 370–381.

Khan MA, Khalique A, and Abouammoh A (1989). On estimating parameters in a discrete Weibull distribution, *IEEE Transactions on Reliability*, **38**, 348–350.

Kim TI, Choi YY, and Lee KH (2008). Analysis on the differences in medical service usage in terms of income Levels, *Korean Social Security Studies*, **24**, 53–75.

Klakattawi HS, Vinciotti V, and Yu K (2018). A simple and adaptive dispersion regression model for count data, *Entropy*, **20**, 142.

Kleinke K and Reinecke J (2013). Multiple imputation of incomplete zero-inflated count data, *Statistica Neerlandica*, **67**, 311–336.

Kulasekera K (1994). Approximate MLE's of the parameters of a discrete Weibull distribution with type 1 censored data. Microelectron, *Reliab*, **34**, 1185–1188.

Lee YC, Im BH, and Park YH (2010). The determinants and comparison of health behavior and health service by private medical insurance on National Health-Nutrition Survey, *Journal of the Korea Contents Association*, **10**, 190–204.

Nakagawa T and Osaki S (1975). The discrete Weibull distribution, *IEEE Transactions on Reliability*, R-24.

Pahel BT, Presisser JS, Stearns SC, and Rozier RG (2011). Multiple imputation of dental caries data using a zero inflated Poisson regression model, *Journal of Public Health Dental*, **71**, 71–78.

Peluso A and Vinciotti V (2018). Discrete weibull generalised additive model: an application to count fertility data, *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, arXiv:1801.0790.

Rubin DB (1987). *Multiple Imputation for Nonresponse in Surveys*, John Wiley & Sons, New York.

Saez-Castillo A and Conde-Sanchez A (2013). A hyper-Poisson regression model for overdispersed and underdispersed count data, *Computational Statistics & Data Analysis*, **61**, 148–157.

Saffari SE and Adnan R (2010). Zero-inflated Poisson regression models with right censored count data, *Mathematika*, **27**, 21–29.

Sellers KF and Shmueli G (2010). A flexible regression model for count data, *Annals of Applied Statistics*, **4**, 943–961.

van Buuren S (2007). Multiple imputation of discrete and continuous data by fully conditional specification, *Statistical Methods in Medical Research*, **16**, 219–242.

van Buuren S, Boshuizen HC, and Knook DL (1999). Multiple imputation of missing blood pressure covariates in survival analysis, *Statistics in Medicine*, **18**, 681–694.

Willmot GE (1987). The Poisson-inverse Gaussian distribution as an alternative to the negative Binomial, *Scandinavian Actuarial Journal*, 113–127.

Yoo H (2019). A study of discrete Weibull regression model with missing data, *Journal of the Korean Data and Information Science Society*, **30**, 11–22.

Received March 20, 2019; Revised April 23, 2019; Accepted April 29, 2019