Modeling veterans' health benefit grants using the expectation maximization algorithm

Tatjana Miljkovic*
Department of Statistics
North Dakota State University
Fargo, North Dakota 58108-6050
USA

Nikita Barbanov †
Department of Mathematics
North Dakota State University
Fargo, North Dakota 58108-6050
USA

Abstract

A novel application of the Expectation Maximization (EM) algorithm is proposed for modeling right censored multiple regression. Parameter estimates, variability assessment, and model selection are summarized in a multiple regression settings assuming a normal model. The performance of this method is assessed through a simulation study. New formulas for measuring model utility and diagnostics are derived based on the EM algorithm. They include reconstructed coefficient of determination and influence diagnostics based on one-step deletion method. A real data set, provided by North Dakota Department of Veterans Affairs (ND DVA) is modeled using the proposed methodology. Empirical findings should be of benefit to government policy makers.

Key Words:linear models, regression, censoring, EM algorithm, veterans
JEL Classification: 62N01, 62N02, 62J05

1 Background

According to the U.S. Census Bureau [4], there were 22.5 million living veterans in the United States as of 2010, representing 7.3% of the total population. Veterans are eligible for a number of federal and state benefit programs and services offered by the Department of Veterans Affairs (DVA)[29]. The uninsured rate of veterans decreased from 7.6% in 2000 to 7.2% in 2009 [4]. As federal and state medical health benefits are available to eligible veterans, the number of veterans 18 years and older using these programs increased from 50% in 2000 to 60% in 2009. The availability of these

*Corresponding author: tatjana.miljkovic@ndsu.edu
†nikita.barabanov@ndsu.edu

programs is critical for veterans who live below poverty level. Poverty rate among veterans, defined as income below 100% of poverty threshold [5], has increased over the past decade, and it was reported at 6.3% in 2009 compared to 5% in 2000 [26]. The Bureau of Labor Statistics reported that in 2007, 11.8% of North Dakota's population was living below the poverty level. The national average for the same period was 13% [27].

State benefit programs for veterans vary from state to state. In the state of North Dakota, ND DVA [19], working under the supervision of the Administrative Committee of Veterans Affairs (ACOVA), administers various state benefit programs available to low income veterans and their families. The Hardship Grants Program provides aid to veterans for unmet medical needs and encompasses medical benefits for the following categories: dental, denture, hearing, optical, and special. The cost of this program is underwritten by the Veterans Post War Trust Fund (VPWTF). The State Treasurer is the trustee of this fund, as provided for in the state constitution. This fund relies on its investments in the financial market in order to grow and generate annual income for use in grant programs that will benefit veterans. The ND DVA is responsible for the administration of these programs. The policy and guidelines of these programs are set by the ACOVA whose board is made up of veterans. In order to prudently manage the fund and budget Hardship Grants Program, it is important to evaluate the medical benefit needs of veterans in North Dakota so that appropriate decisions are made at the state level to generate sufficient funds to pay eligible veterans and their families in future years. This study provides statistical models and tools which can be applied in the financial assessment of the medical benefit needs for veterans in North Dakota and may be used in any other U.S. state where similar programs exist. Government and policy makers may also be interested in this study as they want to make decisions and provide sound investments for future public policies.

2 Introduction

Censoring has been extensively discussed as a part of survival analysis and a large volume of literature is generated in this area. Good information on these topics can be found in books by Klein and Moeschberger [11] and Lee [12]. An observation is right censored at a censoring point if when it is above the censoring point, it is recorded as being equal to the censoring point, but when it is below the censoring point, it is recorded as its observed value. In medical statistics, right censoring is analyzed from the data of patients who are still alive at the end of the study and those who terminated the study due to surrender as stated by Miller [16]. Right censoring in insurance industry was discussed by Guiahi [9]. Some policies are structured in such a way that the policy limits serve as a restricted amount of payment on a given loss. For a loss below or equal to the policy limit, payment is made in the amount equal to the loss. If the loss exceeds the policy limit, payment is imputed at the policy limit.

Linear regression models are commonly used in many applications to analyze the

functional relationship between a response variable and other explanatory variables that are perceived to be related to the response variable. Typically, a normal distribution is assumed for the underlying assumption of the error structure. However, these models have limitations when the response variable is right censored since they may yield fitted values of the variable of interest to exceed its upper or lower bound when the censoring is ignored.

Tobit models [22] were popular for some time since they allowed for the response variable to be latent (i.e. unobservable) in the regression settings (e.g., [6], [21], [23], [8]). The observable response variable is equal to the latent variable whenever the latent variable is above zero and zero otherwise. The interpretation of the coefficient is not the same as that used in the ordinary regression. For example, the interpretation may look at the change of the response variable of those above the limit weighted by the corresponding probability of being above the limit [14]. Also basic Tobit model uses one censoring level (threshold) that is constant across all observations. Some critiques of the use of Tobit model were raised by Maddala [13]. He suggested that this method is appropriate only when the bunching of the y values in a regression occurs because of some exogenous mechanism (e.g., the way in which data were collected or recorded) and not in other situations.

Censored, Sample Selected, or Truncated Data were nicely summarized in a book by Breen [3]. The book includes many examples of censored regressions which apply the concepts and relate a reader to the applications in non-experimental social sciences. The book also emphasizes the advantage of using the maximum likelihood approach in parameter estimation. Breen [3] makes a note of caution that large sample size is important for the desirable properties of the estimators. However, none of the methods presented in the book are based on the EM algorithm.

Early studies on parametric methods for right censored regression were dated in the 1970s. An iterative procedure known as the EM Algorithm was proposed by Dempster *et al.* [7]. The EM algorithm has been extensively used for missing data or data containing missing values. Good information on the EM methodology and the applications can be found in the book published by McLachlan and Krishnan [15]. More recent significant developments in using EM algorithm in right censored modeling problem are presented in papers by Wei and Tanner [24] and Ng *et al.* [18].

Aitkin [1] analyzed data on electrical insulation in 40 motorettes tested at four different temperature settings. The time until failure in hours of each motorette is recorded. Observations were right censored if the motorettes were still on test without failure at the indicated time. Aitkin used a simple linear regression model and showed that the parameter estimates for the same data (40 motorettes) can be obtained by maximum likelihood using the EM algorithm. In the E-step, censored observations were replaced with their conditional expectations given the observed data and the current parameter estimates. Then in the M-step, the new parameter estimates were computed by the maximum likelihood method based on the complete data.

This article extends the Aitkin's idea and offers another application of the EM algorithm in right censored multiple regression settings by providing parameter estimates,

variability assessment, model selection, and measures of model utility and influence. A novel application of this methodology is demonstrated on financial benefit data set provided by ND DVA.

The organization of this paper is as follows. Section 3 defines problem settings, parameter estimates, variability assessment, and normality assumptions based on the EM algorithm. Section 4 introduces new formulas for measuring model utility and diagnostics based on one-step deletion. Simulation study is provided with parameter estimates and model validation in Section 5. Section 6 includes the analysis and discussion of the ND DVA data incorporating methodology presented in this article. Concluding remarks are given in Section 7.

3 Right Censored Regression

3.1 Problem Setting

Consider the traditional form of the multiple regression model:

$$y = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon} \sim N(0, \sigma^{\mathbf{2}}\mathbf{I})$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^{'}$ is a vector of unknown parameters. The matrix, referred to as the design matrix, $\boldsymbol{X}$ is of size $n \times (p+1)$ and is assumed to have rank equal to $p+1$ (full column rank). The goal of traditional multiple regression is to estimate the parameter vector $\boldsymbol{\psi} = (\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \ldots, \ \beta_p, \sigma^2)^{'}$. This can be accomplished through minimization of a suitable cost functional, for example, by Least-Squares method, which minimizes the sum of squares of deviations for the n observed responses, y_i , from their fitted values,$(\hat{y}_i)$.

Now, consider the linear regression model with censored observations. Assume $\boldsymbol{y}$ and $\boldsymbol{z}$ are n_1 - and n_2-vectors of uncensored and censored observations respectively; $n = n_1 + n_2$.

Denote by $\tilde{z}$ the vector of unknown values which are censored to vector $\boldsymbol{z}$. Denote by $\boldsymbol{y}^* = \begin{pmatrix} y \\ \tilde{z} \end{pmatrix}$. The linear regression model has a form

$$\mathbf{y}^* = \mathbf{X}^*\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon} \sim N(0, \boldsymbol{I}\sigma^2)$, and $\boldsymbol{X}^*$ is a design matrix. Then $\boldsymbol{X}^*$ may be partitioned into two parts: $\boldsymbol{X}^* = \begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{pmatrix}$ corresponding to the uncensored and censored observations. As above the goal is to get values of the set of parameters $\boldsymbol{\psi} = (\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \ldots, \ \beta_p, \sigma^2)^{'}$.

3.2 Expectation Maximization Method

We can use the EM method when we have two sets of unknown quantities: parameters of model (like coefficients of regression model and variation of additive noise) and data

which provide an incomplete information about some of observations (for example, censored observations).

Every iteration of EM algorithm consists of two steps, which are usually called E-step and M-step. On E-step we try to restore the values of incomplete observations having the parameters of model fixed. In many cases these restored values are taken equal to corresponding conditional expectations of these values given available information about these observations and parameters of the model.

On the next M-step the parameters of model are recomputed based on new values of observations found on the E-step. To this end the method of maximization of the likelihood function may be used.

On each iteration both sets of unknown quantities are changed. In many cases (and in the case considered in this article) it is proved that the iterations converge to certain limits. The stopping criterion is based on when the relative increase in the likelihood function is no bigger than some small pre-specified tolerance value.

3.3 Parameter Estimates

The complete likelihood function, based on the complete information for censored regression, is defined as follows:

$$L_c(\boldsymbol{\psi}, \boldsymbol{y}, \tilde{\boldsymbol{z}}) = (2\pi)^{-n/2}(\sigma^2)^{-n/2} \exp\left\{-\frac{[(\boldsymbol{y}-\boldsymbol{X}_1\boldsymbol{\beta})'(\boldsymbol{y}-\boldsymbol{X}_1\boldsymbol{\beta}) + (\tilde{\boldsymbol{z}}-\mathbf{X}_2\boldsymbol{\beta})'(\tilde{\boldsymbol{z}}-\boldsymbol{X}_2\boldsymbol{\beta})]}{2\sigma^2}\right\}$$

The logarithm of function L_c, known as the complete-data loglikelihood function l_c is given by

$$l_c(\boldsymbol{\psi}, \mathbf{y}, \tilde{\mathbf{z}}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2)$$

$$-\frac{[\mathbf{y}'\mathbf{y} - 2\boldsymbol{\beta}'\mathbf{X}_1'\mathbf{y} + \boldsymbol{\beta}'\mathbf{X}_1'\mathbf{X}_1\boldsymbol{\beta} + \tilde{\mathbf{z}}'\tilde{\mathbf{z}} - 2\boldsymbol{\beta}'\mathbf{X}_2'\tilde{\mathbf{z}} + \boldsymbol{\beta}'\mathbf{X}_2'\mathbf{X}_2\boldsymbol{\beta}]}{2\sigma^2}$$

The conditional expectation of l_c given the observed data $(\mathbf{y}, \mathbf{z})$ and $\boldsymbol{\psi}$ is defined as Q-function, given by

$$Q(\boldsymbol{\psi}, \boldsymbol{y}, \boldsymbol{z}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2)$$

$$-\frac{[y'y - 2\boldsymbol{\beta}'\boldsymbol{X}_1'\mathbf{y} + \boldsymbol{\beta}'\boldsymbol{X}_1'\mathbf{X}_1\boldsymbol{\beta} + B - 2\boldsymbol{\beta}'\boldsymbol{X}_2'\boldsymbol{A} + \boldsymbol{\beta}'\boldsymbol{X}_2'\boldsymbol{X}_2\boldsymbol{\beta}]}{2\sigma^2} \tag{3.1}$$

Here, n_2-vector $\boldsymbol{A}$ and a number B are calculated in the E-step as the first and second moments of the conditional expectation for censored observations, given that their values are above the censoring point. It is straightforward to show that

$$\boldsymbol{A} = E(\tilde{\mathbf{z}} \mid \tilde{z} > z, \boldsymbol{\beta}, \sigma^2) = \mathbf{X}_2\boldsymbol{\beta} + \sigma f(\frac{z - \boldsymbol{X}_2\boldsymbol{\beta}}{\sigma}),$$

$$B = E(\tilde{z}^{'}\tilde{z} \mid \tilde{z} > \mathbf{z}, \boldsymbol{\beta}, \sigma^2) = \|\boldsymbol{X}_2\boldsymbol{\beta}\|^2 + \sigma(\boldsymbol{X}_2\boldsymbol{\beta} + z)' f(\frac{z - \boldsymbol{X}_2\boldsymbol{\beta}}{\sigma}) + n_2\sigma^2,$$

where $f(x) = \frac{\varphi(x)}{\Phi(-x)}$, $\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$, $\Phi(-x) = \int_{-\infty}^{-x} \varphi(s)\, ds$, and $\|\cdot\|$ is the Euclidean norm of vectors.

The E-step consists of computing $\boldsymbol{A}$ and B. During the next step, the M-step, we maximize the Q-function with respect to parameters $\boldsymbol{\beta}$ and σ using the values $\boldsymbol{A}$ and B. The maximized value of the Q-function will lead to the maximum likelihood estimates (MLEs) for the model. Finding the maximum amounts to finding the solutions to the following equations:

$$\frac{\partial Q}{\partial \boldsymbol{\beta}} = 0 \qquad \text{and} \qquad \frac{\partial Q}{\partial \sigma^2} = 0.$$

From this, we have

$$\frac{\partial Q(\boldsymbol{\psi}, \boldsymbol{y}, \boldsymbol{z})}{\partial \boldsymbol{\beta}} = \frac{\boldsymbol{X}_1^{'} y - \boldsymbol{X}_1^{'}\boldsymbol{X}_1\boldsymbol{\beta} + \boldsymbol{X}_2^{'}\boldsymbol{A}\boldsymbol{X}_2^{'}\boldsymbol{X}_2\boldsymbol{\beta}}{\sigma^2}$$

and therefore

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^{'}\boldsymbol{X})^{-1}(\boldsymbol{X}_1^{'} y + \boldsymbol{X}_2^{'}\boldsymbol{A}).$$

Similarly,

$$\frac{\partial Q(\boldsymbol{\psi}, \boldsymbol{y}, \boldsymbol{z})}{\partial \sigma^2} = \frac{[y^{'} y - 2\boldsymbol{\beta}^{'}\boldsymbol{X}_1 y + \boldsymbol{\beta}^{'}\mathbf{X}_1^{'}\boldsymbol{X}_1\boldsymbol{\beta} + B - 2\boldsymbol{\beta}^{'}\boldsymbol{X}_2^{'}\boldsymbol{A} + \boldsymbol{\beta}^{'}\boldsymbol{X}_2^{'}\boldsymbol{X}_2\boldsymbol{\beta}] - n\sigma^2}{2(\sigma^2)^2}$$

and

$$\hat{\sigma}^2 = \frac{y^{'} y + B + \boldsymbol{\beta}^{'}(\mathbf{X}_1^{'}\mathbf{X}_1 + \mathbf{X}_2^{'}\mathbf{X}_2)\boldsymbol{\beta} - 2\boldsymbol{\beta}^{'}(\mathbf{X}_1^{'}\mathbf{y} + \mathbf{X}_2^{'}\mathbf{A})}{n}$$

Here $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ are MLEs of parameters $\boldsymbol{\beta}$ and σ^2, respectively. Using norms notation, the equation above can be expressed as

$$\hat{\sigma}^2 = \frac{\|\mathbf{y} - \mathbf{X}_1\hat{\boldsymbol{\beta}}\|^2 + \|\mathbf{X}_2\hat{\boldsymbol{\beta}} - \boldsymbol{A}\|^2 + B - \|\boldsymbol{A}\|^2}{n}.$$

Calculation of parameter estimates $\hat{\boldsymbol{\beta}}^{(k+1)}$ and $(\hat{\sigma}^2)^{(k+1)}$ in each (k + 1) step can be obtained as follows:

$$\hat{\boldsymbol{\beta}}^{(k+1)} = (\mathbf{X}^{'}\boldsymbol{X})^{-1}(\boldsymbol{X}_1^{'} y + \boldsymbol{X}_2^{'}\boldsymbol{X}_2\hat{\boldsymbol{\beta}}^{(k)} + \sigma^{(k)}\boldsymbol{X}_2^{'} f(\frac{z - \boldsymbol{X}_2\boldsymbol{\beta}^{(k)}}{\sigma^{(k)}}))$$

$$(\hat{\sigma}^2)^{(k+1)} = \frac{(\|y - \boldsymbol{X}_1\hat{\boldsymbol{\beta}}^{(k)}\|^2 + \sigma^{(k)}(z - \mathbf{X}_2\hat{\boldsymbol{\beta}}^{(k)})' f(\frac{\boldsymbol{z} - \boldsymbol{X}_2\boldsymbol{\beta}^{(k)}}{\sigma^{(k)}}) + n_2(\sigma^2)^{(k)})}{n}$$

3.4 Variability Assessment

McLachlan and Peel [17] defined an approach that can be employed for the variability assessment of all parameter estimates. The empirical observed information matrix serves as an estimate of the corresponding observed information matrix and is obtained by

$$\boldsymbol{I}_e(\hat{\boldsymbol{\psi}}) = \sum_{i=1}^{n} \bigtriangledown \boldsymbol{q}_i(\hat{\boldsymbol{\psi}}) \bigtriangledown \boldsymbol{q}_i(\hat{\boldsymbol{\psi}})'$$

,

where $\hat{\boldsymbol{\psi}} = (\hat{\boldsymbol{\beta}}, \hat{\sigma}^2)$ represents the vector of parameter estimates, or MLEs, and $\bigtriangledown \mathbf{q}_i(\hat{\boldsymbol{\psi}})$ is the gradient vector of the conditional expectation of the complete data log likelihood function constructed on the ith observation and evaluated at $\hat{\boldsymbol{\psi}}$. Note that: $Q(\hat{\boldsymbol{\psi}}) = \sum_{i=1}^{n} \mathbf{q}_i(\hat{\boldsymbol{\psi}})$. For each i, $\bigtriangledown \mathbf{q}_i(\hat{\boldsymbol{\psi}})$ is a vector of length (p + 2) defined by

$$\bigtriangledown \boldsymbol{q}_i(\hat{\boldsymbol{\psi}}) = ((\frac{\partial \boldsymbol{q}_i(\boldsymbol{\psi})}{\partial \boldsymbol{\beta}})', (\frac{\partial \boldsymbol{q}_i(\boldsymbol{\psi})}{\partial \sigma}))'$$

.

Consider a vector $d = (d_1, \ldots, d_n)$ of length n, where $d_j = 1$ if jth observation is censored and $d_j = 0$ if jth observation is uncensored, $j = 1, \ldots, n$. Denote by $\boldsymbol{x}_{1i}$, $\boldsymbol{x}_{2i}$ the ith rows of matrices $\boldsymbol{X}_1$, $\boldsymbol{X}_2$ respectively, which have been introduced in Section 3.1. Denote by y_i^* the ith component of vector $\boldsymbol{y}^*$ of uncensored and censored observations. It follows that

$$\frac{\partial \boldsymbol{q}_i(\boldsymbol{\psi})}{\partial \boldsymbol{\beta}} = \frac{\boldsymbol{x}_{1i}' y_i^* (1 - d_i) - \boldsymbol{x}_{1i}' \boldsymbol{x}_{1i} (1 - d_i) \boldsymbol{\beta} + \boldsymbol{x}_{2i}' d_i E(y_i^*) - \boldsymbol{x}_{2i}' \boldsymbol{x}_{2i} d_i \boldsymbol{\beta}}{\sigma^2},$$

and

$$\frac{\partial \mathbf{q}_i(\boldsymbol{\psi})}{\partial \sigma} = -\frac{1}{\sigma} + \frac{(y_i^*)^2 (1 - d_i) - 2\boldsymbol{\beta}' \mathbf{x}_{1i}' y_i^* (1 - d_i) + \boldsymbol{\beta}' \mathbf{x}_{1i}' \mathbf{x}_{1i} \boldsymbol{\beta} (1 - d_i)}{\sigma^3}$$

$$+ \frac{E((y_i^*)^2) d_i - 2 - 2\boldsymbol{\beta}' \mathbf{x}_{2i}' E(y_i^*) d_i + \boldsymbol{\beta}' \mathbf{x}_{2i}' \mathbf{x}_{2i} \boldsymbol{\beta} d_i}{\sigma^3}$$

.

These partial derivatives will be used to assemble the covariance matrix. This covariance matrix of the MLEs, which is obtained by taking the inverse of $\boldsymbol{I}_e(\hat{\boldsymbol{\psi}})$ can be directly employed for testing various hypotheses and finding confidence intervals for the parameters of the model.

3.5 Model Selection

The Akaike Information Criterion (AIC) is a popular model selection procedure proposed by Akaike [2]. The AIC considers the negative log-likelihood plus a penalty term

that reflects the number of free parameters (M) in the model. The form of the AIC is given by

$$AIC = -2l(\hat{\boldsymbol{\psi}}) + 2M,$$

where $l(\hat{\boldsymbol{\psi}})$ is defined as follows:

$$l(\hat{\boldsymbol{\psi}}) = -\frac{(n-m)}{2}\log(2\pi) - \frac{(n-m)}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(y - \boldsymbol{X}_1\boldsymbol{\beta})'(y - \boldsymbol{X}_1\boldsymbol{\beta})$$

$$+ \sum_{i=n-m+1}^{n} \log P(y_i^* > z_i),$$

$$l(\hat{\boldsymbol{\psi}}) = -\frac{(n-m)}{2}\log(2\pi) - \frac{(n-m)}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(y - \boldsymbol{X}_1\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}_1\boldsymbol{\beta})$$

$$+ \sum_{i=n-m+1}^{n} \log[1 - \Phi(\frac{z_i - \boldsymbol{x}_{2i}\boldsymbol{\beta}}{\sigma})].$$

The model with the minimum AIC is selected as the best model to fit the data.

Another commonly used method in model selection was proposed by Schwarz [20] and is known as Bayesian Information Criterion (BIC). Similar to AIC, the BIC approach adjusts the log-likelihood $l(\hat{\boldsymbol{\psi}})$ by a penalty term which considers the number of observations (n) in the sample in addition to the number of parameters in the model:

$$BIC = -2l(\hat{\boldsymbol{\psi}}) + M\ \log(n)$$

.

The model with the minimum BIC is chosen as the best model to fit the data.

3.6 Normality assumption

In our model the term ϵ in the linear regression model is supposed to have normal distribution. This assumption is certainly important in deriving the iteration formulas for parameters (β, σ^2) of this distribution. If the error term ϵ has different distribution, then the whole approach remains valid, but the formulas on each step of the EM algorithm take different form. In many cases the M-step requires solution of implicit equations which is an additional computational burden. For such cases it is necessary to derive efficient methods to find maximum with respect to parameters of distribution of the conditional expectation of the complete-data loglikelihood function l_c. This can be a subject of future investigations.

4 Model Validation and Diagnostics

4.1 Measuring Model Utility

It is a standard approach for modeling multiple regression to consider the coefficient of determination R^2 as a useful measure of how well the model fits the data. The R^2 is defined as the proportion of total response variation that is explained by the model. Higher R^2 indicating better model fit. However, R^2 alone does not indicate whether the model is appropriate. The R^2 for ordinary regression is defined as

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y_i})^2}{\sum_{i=1}^{n}(y_i - \bar{y_i})^2} = 1 - \frac{SSE}{TSS},$$

where SSE represents the sum of squares for error and TSS is the total sum of squares. The TSS measures the variability in the model relative to the horizontal line $\bar{y}$. The SSE measures the variability in the response y from the fitted line $\hat{y}$. For ordinary regression, the best fitted model is defined based on the principle of least squares which minimizes the sum of squares of errors SSE.

For right-censored regression using the EM algorithm, there is no comparable measure developed by researchers. The least squares method cannot be applied due to the presence of censored data. The following proposed R^2 calculation is based on the idea of maximizing the Q-function (3.1) given optimal values of the parameters relative to the maximization of the same function assuming the intercept term only.

Assume p is the number of independent variables in the model. Define the following objective function based on the Q-function

$$J(\beta_0, \beta_1, \ldots \beta_p) = \|\mathbf{y} - \mathbf{X}_1\boldsymbol{\beta}\|^2 + \|\mathbf{y} - \mathbf{X}_1\boldsymbol{\beta}\|^2 + B - \|\mathbf{A}\|^2$$

Next, define

$$J_{lin}(p) = min_{\beta_0,\beta_1,\ldots,\beta_p} J(\beta_0, \beta_1, \ldots, \beta_p),$$

$$J_{const} = min_{\beta_0} J(\beta_0, 0, \ldots, 0) = J_{lin}(0, \ldots, 0).$$

J_{lin} is the optimal value of the objective function J if we use the whole design partition matrix $\mathbf{X}$. J_{const} is the optimal value of the same function if we use only the first column of $\mathbf{X}$.

The proposed R-squared is defined as the reconstructed coefficient of determination:

$$R_c^2 = 1 - J_{lin}(p)/J_{const}. \tag{4.1}$$

The $R_c^2(p)$ does not have a closed form solution compared to an ordinary coefficient of determination.

Theorem 4.1. *The following statements about $R_c^2(p)$ are true:*
(i) $0 \leq R_c^2(p) \leq 1$
(ii) Function $R_c^2(p)$ is non-decreasing with respect to p.

Proof. (i) By definition of J_{lin} and J_{const}, we have $0 \leq J_{lin} \leq J_{const}$. Therefore $0 \leq R_c^2(p) \leq 1$. (ii) By definition, $J_{lin}(p)$ is nonincreasing in p. Therefore $R_c^2(p)$ is nondecreasing with respect to p.

4.2 Measuring the Influence by One-step Deletion Method

For assessing the influence of a single observation on the parameter estimates in censored regression, one of the popular methods is one-step deletion method. The one-step deletion method measures the change in parameter estimates when the ith data point is deleted from the sample. Weissfeld and Schneider [25] studied this method but our formula for one-step deletion based on the EM algorithm is differently developed and produces different results.

Consider the following model

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}\hat{y}^{*}$$

where $\hat{y}^{*} = \begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{A} \end{pmatrix}$ is a vector of uncensored and reconstructed censored observations. Now, assume that the ith observation is omitted. Then, instead of using matrix $\boldsymbol{X}$ we have to use matrix $\boldsymbol{X}_{(i)}$, which is the matrix $\boldsymbol{X}$ with the ith row omitted. For this problem, the optimal model has optimal parameters which can be found using a similar formula:

$$\hat{\boldsymbol{\beta}}_{(i)} = (\boldsymbol{X}_{(i)}^{'}\boldsymbol{X}_{(i)})^{-1}\boldsymbol{X}_{(i)}^{'}\tilde{y}_{(i)}^{*},$$

where $\tilde{y}_{(i)}^{*}$ is the vector of uncensored and reconstructed censored observations based on all available observations except for the ith observation.

Denote by $\hat{y}_i^{*}$ the ith component of the vector of observations $\hat{y}^{*}$, based on all available observations including the ith. Denote by $\hat{y}_{(i)}^{*}$ the vector $\hat{y}^{*}$ with the ith observation $\hat{y}_i^{*}$ omitted. Notice that vectors $\tilde{y}_{(i)}^{*}$ and $\hat{y}_{(i)}^{*}$ have the same number of components; the former vector is based on all observations except the ith one while the latter vector is based on all observations including this ith observation.

Denote by x_i the ith row of the matrix $\boldsymbol{X}$ which is omitted in $\boldsymbol{X}_{(i)}$. Then, $\boldsymbol{X}^{'}\boldsymbol{X} = \boldsymbol{X}_{(i)}^{'}\boldsymbol{X}_{(i)} + x_i^{'}x_i$ and $\boldsymbol{X}^{'}\hat{y}^{*} = \boldsymbol{X}_{(i)}^{'}\hat{y}_{(i)}^{*} + x_i^{'}\hat{y}_i^{*}$. Thus,

$$(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}(\boldsymbol{X}_{(i)}^{'}\boldsymbol{X}_{(i)} + x_i^{'}x_i) = \boldsymbol{I}.$$

Multiplying this equation by $(\boldsymbol{X}_{(i)}^{'}\boldsymbol{X}_{(i)})^{-1}$ we obtain:

$$(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}(\boldsymbol{I} + x_i^{'}\ x_i(\boldsymbol{X}_{(i)}^{'}\boldsymbol{X}_{(i)})^{-1}) = (\boldsymbol{X}_{(i)}^{'}\boldsymbol{X}_{(i)})^{-1}.$$

Next, if we multiply each side of equation by x_i from the left and regroup the terms, we have

$$x_i(\boldsymbol{X}^{'}\boldsymbol{X})^{-1} + x_i(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}x_i^{'}x_i(\boldsymbol{X}_{(i)}^{'}\boldsymbol{X}_{(i)})^{-1} = x_i(\boldsymbol{X}_{(i)}^{'}\boldsymbol{X}_{(i)})^{-1}$$

$$x_i(\boldsymbol{X}^{'}\boldsymbol{X})^{-1} = (1 - \ x_i(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}x_i^{'})x_i(\boldsymbol{X}_i^{'}\boldsymbol{X}_i)^{-1}$$
$$x_i(\boldsymbol{X}_{(i)}^{'}\boldsymbol{X}_{(i)})^{-1} = (1 - x_i(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}x_i^{'})^{-1}x_i(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}.$$

Substituting into the appropriate part of we get

$$(\mathbf{X}^{'}\boldsymbol{X})^{-1}(\boldsymbol{I} + \frac{x_i^{'}x_i(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}}{1 - x_i(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}x_i^{'}}) = (\boldsymbol{X}_{(i)}\boldsymbol{X}_{(i)})^{-1}.$$

Then, we have

$$\triangle\boldsymbol{\beta}^{EM} = \hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)} = (\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}\hat{y}^{*} - (\boldsymbol{X}_{(i)}^{i}\boldsymbol{X}_{(i)})^{-1}\boldsymbol{X}_{(i)}^{'}\tilde{y}_{(i)}^{*}$$
$$= (\boldsymbol{X}^{'}\boldsymbol{X})^{-1}[\boldsymbol{X}_{(i)}^{'}\hat{y}_{(i)}^{*} + x_i^{'}\hat{y}_i^{*} - (\boldsymbol{I} + \frac{x_i^{'}x_i(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}}{1 - x_i(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}x_i^{'}})\boldsymbol{X}_{(i)}^{'}\tilde{y}_{(i)}^{*}]$$
$$= \frac{(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}}{1 - x_i(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}x_i^{'}}[(1 - x_i(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}x_i^{'})\boldsymbol{X}_{(i)}^{'}(\hat{y}_{(i)}^{*} - \tilde{y}_{(i)}^{*})$$
$$+x_i^{'}\hat{y}_i^{*}(1 - x_i(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}x_i^{'}) - x_i^{'}x_i(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}_{(i)}^{'}\tilde{y}_{(i)}^{*}].$$

Finally, we obtain $\triangle\boldsymbol{\beta}^{EM}$ as

$$\triangle\boldsymbol{\beta}^{EM} =$$
$$\frac{(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}}{1 - x_i(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}x_i^{'}}[[(1 - x_i(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}x_i^{'})\boldsymbol{I} + x_i^{'}x_i(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}]\boldsymbol{X}_{(i)}^{'}(\hat{y}_{(i)}^{*} - \tilde{y}_{(i)}^{*}) + x_i^{'}(\hat{y}_i^{*} - x_i\hat{\boldsymbol{\beta}})].$$

By comparing this formula to that developed by Weissfeld and Schneider we observe that they are different and coincide if $\hat{y}_{(i)}^{*} - \tilde{y}_{(i)}^{*} = 0$. However, if the difference $\hat{y}_{(i)}^{*} - \tilde{y}_{(i)}^{*} = 0$ is not equal to zero, then in general, the formulas produce different results.

In order to eliminate the influence of an observation due to its position on the interval of x values, the vector $\triangle\boldsymbol{\beta}^{EM}$ in the formula can be divided by the vector $(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}x_i$ component-vise providing a valuable measure of sensitivity of parameters with respect to the coefficients of the linear model. Thus the normalized version of the formula is defined as

$$[\triangle\boldsymbol{\beta}_{NOR}^{EM}]_j = \frac{[\triangle\boldsymbol{\beta}^{EM}]_j}{[(\mathbf{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{x}_i]_j}, \qquad j = 1, 2, \ldots, p.$$

Relatively large values of this statistic indicate the most influential observations on the coefficient estimates of the model. This issue was not addressed by Weissfeld and Schneider for the one-step deletion based on the EM algorithm.

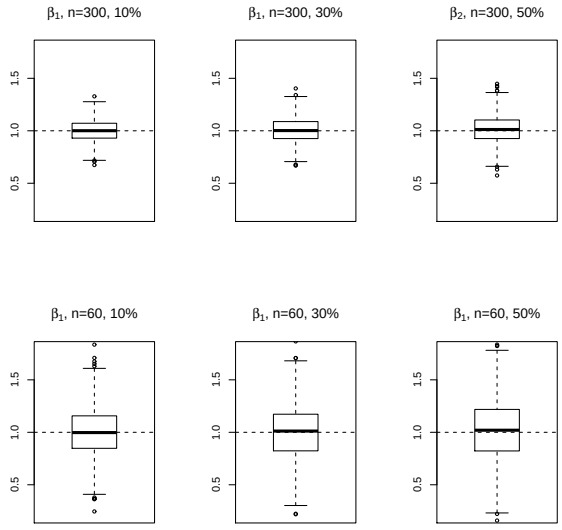

Figure 1: Simulation results for β_1 for sample size (300, 60) with 10%, 30%, and 50% censored data.

5 Simulation Study

A simulation study was conducted to assess the performance of the proposed method for different sample size and amount of censoring. The data were simulated from the model

$$y = 2 + X_1 + X_2 + \epsilon$$

where X_1 is designed such that $x_i = i/n, i = 1, 2, ..., n$ where n is the sample size, $X_2 \sim Binomial(n, 0.5)$, and $\epsilon \sim N(0, 0.2)$. For this model $\beta_0 = 2$, $\beta_1 = 1$, and $\beta_2 = 1$. Different simulation settings were created by manipulating the sample size (300, 60) and the percentage of points censored (10%, 30%, 50%) in order to validate performance of the algorithm. Once the data were generated, censored points were selected at random on the entire interval. Their values were compared to a censoring level randomly drawn from U(1, 4). If a value of a selected data point was above the censoring level, it was trimmed at the censoring level, otherwise it remained uncensored. This procedure was repeated until the desired censoring amount was achieved.

The results of the 1000 runs for each setting of simulation are summarized in Table 1, which shows the average parameter estimates and their corresponding mean square errors (MSE). For visual illustration, a box plot summary of estimated β_1 for sample size (300, 60) with 10%, 30% and 50% censoring is shown in Figure 1.

The results show that the proposed parameter estimators have very small bias and

Table 1: Simulation Results

Parameter	Sample Size	Censoring (%)	$\hat{\beta}$ (*MSE*)
β_0	60	10	1.9992 *(0.0035)*
		30	1.9982 *(0.0044)*
		50	2.0051 *(0.0056)*
	300	10	1.9993 *(0.0007)*
		30	2.0011 *(0.0008)*
		50	2.0010 *(0.0011)*
β_1	60	10	1.0008 *(0.0088)*
		30	1.0055 *(0.0118)*
		50	0.9993 *(0.0147)*
	300	10	1.0014 *(0.0018)*
		30	0.9997 *(0.0022)*
		50	1.0030 *(0.0031)*
β_2	60	10	1.0021 *(0.0028)*
		30	1.0021 *(0.0039)*
		50	1.0023 *(0.0053)*
	300	10	1.0004 *(0.0005)*
		30	1.0002 *(0.0007)*
		50	1.0043 *(0.0010)*
σ	60	10	0.1932 *(0.0004)*
		30	0.1921 *(0.0005)*
		50	0.1905 *(0.0007)*
	300	10	0.1987 *(0.0001)*
		30	0.1984 *(0.0001)*
		50	0.1979 *(0.0001)*

the mean square error. Therefore, the proposed EM method works very well in a multiple regression setting assuming a normal model. Sample size and the amount of censoring have impact on the parameter estimates and their corresponding $MSEs$. It can be noticed from Table 1 that an increase in MSE is observable for an increase in sample size and amount of censoring.

6 Application to North Dakota DVA Data

6.1 Data

Data used in this study were provided by ND DVA. Typically, categories of health benefits available to veterans are capped (right censored) or limited at certain level. The censoring points change over time, as they are subject to review and state approval, and they may vary across different categories. For any claim, if the expense exceeds the amount granted, it will be reimbursed at the value of granted amount.

Medical grants are subject to a limit and the annual amount of benefits is capped (right censored). The data provided consist of payment amounts granted to each applicant for years 2000 through 2010. Table 2 shows the variables provided and their descriptions.

About half of the variables listed in Table 2 were of interest to our project. The difference between application year and birth year was used to determine the applicant's age. Year when the application was approved was extracted from the approved date.

Table 2: The ND DVA Data Summary

Variable	Description
VaucherDate	Day/Month/Year when the benefit payment is made
Gender	Male(0) or Female (1)
ApplicationDate	Day/Month/Year when the application was filed
ApprovedDate	Day/Month/Year when the application was approved
BirthDate	Birth date of each applicant
AmountGranted	Amount granted by the grant program
Category	Category of benefits (dental, denture,hearing, optical, and special)
ApplicantTB	Applicant's unique non-identifiable ID
Status	Status of a person receiving benefits (v-primary beneficiary(veteran), vs-spouse of a living veteran, and w-widow/widower).
NoIndependents	Family size including applicant(seven levels:1,2,3,4,5,6,7)
AmountPaid	Benefit amount paid
ZipCode	5-digit postal code of the applicant address
County	County code of the applicant address
CountyName	County name of the applicant address

An applicant is given only 90 days to use the grant. In this case approved date and voucher date are only three months apart, and the data are available only for those applicants who actually used the grants. Dates for others who have not managed to use the grant were provided as cancelations and were ignored in this study. The amount of money granted as well as the amount of money given from 2000 to 2010 by ND DVA is adjusted for inflation using the Consumer Price Index (CPI) published by Bureau of Labor Statistics, U.S. Department of Labor [4].

Historically, benefit categories carry different benefit caps (limits) on an annual basis. Dental benefits started with a $500 cap as of 12/2004, then increased to $750 as of 1/2006, and finally reached $1000 as of 11/2007. Dental services sometime require more than one appointment; in this case applicants receive several payments during the year. Therefore, the data for dental category were aggregated by year and applicant. The data for dentures, hearing, optical, and special categories of benefits were excluded since they contained significantly lower number of records and as such they may not be reliable.

The ND DVA uses monthly income level and family size to determine if an applicant meets benefit eligibility criteria. Each income level corresponds to a certain family size. For example, a family of two earning less than $1400 per month, or a family of eight earning less than $2600 per month, would be eligible for benefits. Many records were missing family size but had income level provided. For this reason, we used income level only and ignored family size as these two variables seem to be correlated.

Table 3: Department of Human Services-Poverty Guidelines

Variable	**Description**						
Household Size	1	2	3	4	5	6	7
Annual Income	$10,400	$14,000	$17,600	$21,200	$24,800	$28,400	$32,000

Table 4: Eligibility Requirements Set by ACOVA

Variable	**Description**						
Household Size	1	2	3	4	5	6	7
Annual Income	$14,400	$16,800	$19,200	$21,600	$25,200	$28,800	$31,200

Dental records show that the applicant' age varies from 24 to 94 with 84% of the individuals being older than 50. Men represent 287 applicants compared to 81 women. Based on status, 26 applicants are spouses of living veterans and 33 applicants are widows or widowers. Living veterans represent 309 individuals or 84% of the sample. It is observed that 34 individuals or 9.2% of the sample reported zero income. The highest income reported is $2600 per month for a large family. Thus, most of these people live below the poverty level. The poverty guidelines are issued each year in the Federal Register by the Department of Health and Human Services (HHS)[28]. The 2008 income threshold by family size, reported by HHS, for the 48 contiguous states is summarized in Table 3.

North Dakota had 11.8% of its total population living below the poverty level in 2007 compared to the national average of 13% reported for the same period. Poverty guidelines determined by ACOVA on the basis of national statistics are reported in Table 4. These poverty tables are analyzed periodically by ACOVA and they are used to adjust eligibility criteria for benefits as well as to modify limits on benefits.

There were 575 annual aggregate applications for dental benefits used by 368 different individuals for years 2000-2010. We identified 274 (48%) applications with a paid amount in benefits equal to or higher than the amount granted. These policies represent right censored data. For uncensored data records, paid amount in benefits was greater than zero and less that the defined limit (cap or censoring point).

Finally, the following variables were selected for inclusion in the modeling of dental benefits: year, age, gender, amount granted, censored amount, income level, and applicant's status. Application year, age, gender, income level, and applicant's status represented explanatory variables while the amount paid (adjusted for inflation) was used as a response variable in the model.

6.2 Analysis

The EM algorithm was applied to illustrate the modeling of veterans' health benefits with a special focus at dental benefit category. Statistical computing was performed

Table 5: Parameter Estimates for the Full EM Model.

Parameters	**Estimates**	95% CI	p-**value**
Intercept	329.60	(116.36, 542.83)	0.0024
Application Year	58.37	(42.66, 74.07)	0.0000
Age	-0.31	(-3.30, 2.68)	0.8391
Gender	88.19	(-34.30, 210.69)	0.1582
Income Level	0.05	(-0.03, 0.14)	0.2422
Spouse	-54.12	(-223.15, 114.91)	0.5303
Widow/er	−157.45	(-328.90, 13.99)	0.0718

in R software version 3.01. First, the right censored regression model was considered with all explanatory variables. That is:

$$E(BenefitPaid) = \beta_0 + \beta_1(Applicationyear) + \beta_2(Age) + \beta_3(Gender) +$$

$$\beta_4(IncomeLevel) + \beta_5(Spouse) + \beta_6(Widow/er). \tag{6.1}$$

The EM algorithm, employed in modeling parameter estimates and variability assessments, indicated that gender, age, income level, and spouse were not significant predictors of the paid benefits. Application year and widow /er were significant predictors with the possibility of application year entering the model as a quadratic term. The parameter estimates (and their significance) of this model are shown in Table 5.

Table 6: Summary of Different Criteria Used in the Model Selection

Model	**Log-likelihood**	**AIC**	**BIC**
Model-1	-2126.44	4266.88	4272.19
Model-2	-2129.39	4262.78	4264.29
Model-3	-2128.37	4262.74	4265.19
Model-4	-2129.14	4262.28	4263.79
Model-5	-2128.12	4262.23	4264.50
Model-6	-2120.93	4249.86	4252.89

Model-1: Full model per (6.1)
Model-2:E(Benefit Paid)= $\beta_0 + \beta_1$(Application year)
Model-3:E(Benefit Paid)= $\beta_0 + \beta_1$(Application year) + β_6(Widow/er)
Model-4:E(Benefit Paid)= $\beta_0 + \beta_1$(Application year)^2
Model-5:E(Benefit Paid)= $\beta_0 + \beta_1$(Application year)^2 + β_6(Widow/er)
Model-6:E(Benefit Paid)= $\beta_0 + \beta_1$(Application year)^2 + β_3(Gender) + β_6(Widow/er)

Table 7: Parameter Estimates for the EM Model-6

Parameters	**OLS Estimates**	**EM Estimates**	**EM** 95% CI	p**-value**
Intercept	375.76	522.45	(457.54, 587.36)	0.0000
(Application year)^2	4.93	4.41	(3.25, 5.58)	0.0000
Gender	34.39	71.00	(-23.25, 165.26)	0.1300
Widow/er	-86.38	-143.42	(-286.90, 0.05)	0.0500

In subsequent model selections, five additional models were examined. Additionally, we also considered models with the interaction terms but none of the interaction terms were significant. Summary of results for six selected models includes the log likelihood value, AIC, and BIC and it is shown in Table 6. The minimum values of AIC and BIC are reported for Model-6, which is proposed to be the best model.

Parameter estimates for Model-6 with their confidence intervals and corresponding p-values are summarized in Table 7. If we consider the same portfolio of applicants, the total dental benefit needs of ND veterans for the period 2003-2009, calculated based on the EM algorithm, was $407,562 compared to the amount of $333,472 actually spent. The difference of $74,090 can be used to help ACOVA increase the cap on benefits in the future and suggest to the State Treasurer that additional investments were needed in funding this grant program.

Model-6 is the best model based on AIC and BIC criteria even though the gender is not significant variable. According to this model, widowers generate $143.42 less in benefit payments on average compared to a living veteran or a spouse of a living veteran. On average, female applicants require $71 more in benefits compared to a male applicant. While there is a larger proportion of a male veterans compared to female veterans or dependents, it seems that females are using benefits more than

males. Benefits are also a function of money that is available in the state budget for that purpose. When more money is available in state budgets more needy veterans will potentially benefit.

The data show that in more recent years, a higher amount of money was available for spending even when the benefits are adjusted for inflation. The intercept coefficient provides us with a fixed cost per person for running this program. In other words, the veterans spent, based on Model-6, an amount of about $522.45 per applicant/ per year irrespective of the number of applicants and their characteristics. We observe that income level is an insignificant predictor of benefits used. If the overall veteran population was considered in the analysis, one might expect that the lower income veterans are the most likely to use the benefits. However, most veterans eligible for benefits have income below the 100% poverty threshold. Hence the income level is very low and it does not segregate people further into subgroups. Age is another insignificant variable in Model-6 suggesting that benefits are used across all age groups 23-94.

The results of Model-6 are compared to those generated by ordinary least square when censoring is ignored and an improvement is observed. The ordinary least square produces a lower mean and the intercept of the model compared to the EM method. By employing the EM method, not only that we are able to estimate parameters more accurately in presence of censored data, but we are also able to find the conditional execrated values of those censored observations (the value above censoring level). Using the ordinary least square would results in an underestimation [3] and under-prediction of future benefit needs. For example, the ordinary least square generates fixed cost of dental expenses of $375.76 for running this program compared to $522.45 estimated by the EM method. The latter one is more reasonable considering the cost of dental services for the time period considered in the analysis.

The results of the EM method are more informative to the policy holders who make decisions about ND DVA program.This analysis helps our understanding of what are the determinants of the distribution of the available benefit funds. It also helps us determine the total benefit need of the veteran population in ND.

The reconstructed coefficient of determination for Model-6 is 10.75%, lower than the coefficient of determination of 25.74% for the same model when censoring is ignored. The overall fit is relatively low but this is due to the large variability observed in the data set and the large proportion of censored points.

In addition, the reconstructed values for the censored observations can be used to validate the reasonability of the existing benefit caps. Based on the selected model, one can obtain more information about the average amount of expenses in excess of the existing cap.

Six uncensored outliers (1% of the total number of observations) were found in the data. These outliers had t-values above the critical value of 1.96 used for their detection. After careful inspection of the data, it was found that these observations reported extremely low amounts of benefits in the range of $31 to $75. Without additional knowledge as to whether these observations are results of errors or true benefit values, it was decided that they should not be removed.

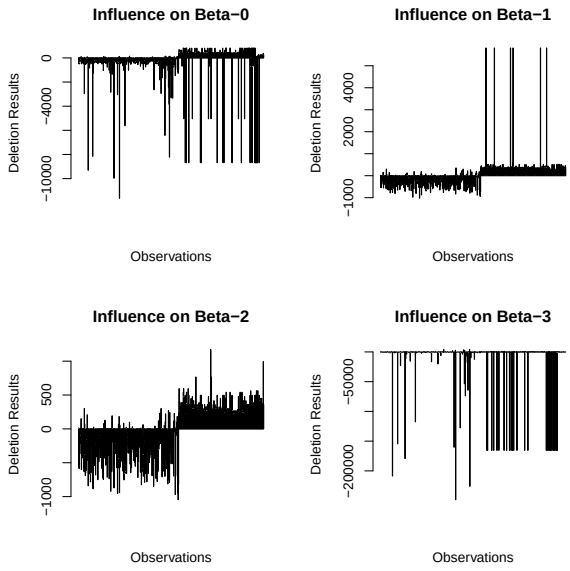

Figure 2: The influence based on one-step deletion of $\beta 0$, $\beta 1$, $\beta 2$, and $\beta 3$.

Influence diagnostics based on the one-step deletion method were applied to the ND DVA dental data. Formula (4.1), proposed in Section 4.2, was used in these calculations. The results for the four parameters from Model-6, based on all 575 data points, are plotted in Figure 2.

The biggest spikes correspond to the most influential points. By careful inspection, it was found that these influential points correspond to most of the censored data reported for years 2006 and 2007 as well as uncensored outliers from these years. If we recall that the cap on dental benefits increased from \$500 to \$750 as of 1/2006 and further increased from \$750 to \$1000 as of 11/2007, these results are expected. The jumps in the censoring levels as well as several uncensored outliers explain the high influence of the corresponding observations on parameter estimates.

6.3 Discussion

Data shows that majority of widows and widowers veterans in North Dakota have age range 49 to 91 with a half of them being older than 75. They are scattered around the state of North Dakota far from Cass County where ND DVA office is located. These counties includes: Morton, Emmons, Grant, McIntosh, McKenzie, Stark, Tower, Ward, Wells, and Renville. In our analysis, we found that widows acquire fewer benefits, on average, compared to living veterans. This could be due to their age and immobility considering the distance from the main ND DVA office in Cass County. The DVA agency may consider different ways of reaching out to this veteran's population segment.

The amount of money granted seems to vary by year. In most recent years, more money was available for benefits. The trend in number of veterans is increasing. Even though caps by category of benefits have been evaluated periodically, the latest caps are still low compared to their expected level generated by the EM algorithm.

The veterans who reported zero income should be evaluated for other benefit opportunities.

7 Conclusion

This paper provides a novel application of the EM algorithm for modeling the right censored multiple regression. The right censored, response variable represents the amount of benefits received by the low-income veterans population in ND as a function of individual characteristics such as gender, age, income, application year, marital status, family size, etc. The EM algorithm was employed for finding the parameter estimates of the censored multiple regression model. Simulation study showed that the proposed method performers well under different simulation settings. New formulas for reconstructed coefficient of determination and influence diagnostics based on one-step deletion were derived using the objective functions of the EM algorithm. Application of this model to North Dakota veterans' data set showed that significant predictors of veterans benefits are: application year, marital status, and gender. On average, widowers acquire significantly less benefits than a living veteran. Application year is another significant predictor of benefits as the money available from the state may vary from year to year. Female veterans spend more money then male veterans although number of female veterans is significantly lower than male veterans. The influence diagnostics formula based on one-step deletion allowed us to easily detect those observations that have great influence on the parameter estimates such as changes made in the censoring level from year to year. This model can also be used to asses appropriates of benefit caps. The reconstructed value of the censored observations can be easily obtain from the EM model and used when decisions are made to increase the benefit caps. These results and findings should be beneficial to both North Dakota policy makers and policy makers in other states with similar programs.

Acknowledgements

The authors are grateful to Kelly Schmidt, State Treasurer of North Dakota, and Lonnie Wangen, Commissioner of North Dakota Department of Veterans Affairs, for providing the data used in this project. We also appreciate helpful discussion with Dr. Volodymyr Melnykov during the initial stage of this project. Finally, the authors thank the editor and two anonymous reviewers whose detailed comments and suggestions greatly improved the quality of this paper.

References

[1] M. Aitkin, *A note on Regression Analysis of Censored Data.*, Technometrics. 23(2)(1981), pp. 161-163.

[2] H. Akaike, *A new look at the statistical modelling identification*, IEEE Transactions on Automatic Control 19(1974), pp.716–723.

[3] R. Breen, *REGRESSION MODELS censored, Sample-Selected, or Truncated Data*, Sage Publication. Thousand Oaks.(1996).

[4] Bureau of Economic Analysis, Bureau of Labor Statistics, U.S. Census Bureau. Available at http:www.fedstats.gov/qf/states/38000.html (Last Accessed August 10, 2012).

[5] Bureau of Economic Analysis (BEA); US Department of Commerce. Available at www.bea.gov (Last Accessed August 2012).

[6] S. Chib *Bayes inference in the Tobit censored regression model*, Journal of Econometrics, 51(1-2)(1992), pp.79-99.

[7] A. P. Dempster, N.M. Laird, and D. B Rubin *Maximum Likelihood from Incomplete data Using EM Algorithm*, Journal of the Royal Statistical Society, Series B, 39 (1977)pp. 1-38.

[8] M. Golder, *Explaining variation in the success of extreme right parties in Western Europe*,Comparative Political Studies (2003), vol 36(4), pp. 432-466.

[9] F. Guiahi *Fitting Loss Distributions in the Presence of Rating Variables*, Journal of Actuarial Practice (2001), Vol.9, pp. 97–129.

[10] J. D. Kalbfleisch, R. L. Prentice *The Statistical Analysis of Failure Time Data*, John Wiley & Sons (2011). 2nd Ed.

[11] J. P. Klein, and M. L. Moeschberger *Survival Analysis*, New York: Springer (2003).

[12] C. T. Lee *Applied Survival Analysis*, New York: John Wiley and Sons, Inc. (1997).

[13] G. S. Maddala em Censored Data Models, In J. Eatwell, M. Milgate, & P. Newman (eds.), The New Palgrave Econometrics, London: Macmillan (1992), pp. 54-57.

[14] J. F. McDonald, and R. A. Moffitt *The Uses of Tobit Analysis*, The Review of Economics and Statistics. 62(2)(1980), pp. 318-321.

[15] G. J. McLachlan, and T. Krishnan *EM Algorithm and Extensions*, New York: John Wiley and Sons Inc. (2007).

[16] R. G. Miller *Least Squares Regression with Censored Data*, Biometrica. 63(3)(1976), pp. 449-464.

[17] G. McLachlan. P. Peel *Finite Mixture Models*, New York: John Wiley and Sons, Inc. (2000).

[18] H.K.T. Ng, P.S. Chan, N. Balakrishnan *Estimation of parameters from progressively censored data using EM algorithm*, Computational Statistics & Data Analysis (2002), vol 39(4), pp. 371-386.

[19] North Dakota Department of Veterans Affairs. Available at www.nd.gov/veterans/

[20] G. Schwarz *Estimating the dimension of model*, The Annals of Statistics. 6 (1978), pp. 461–464.

[21] R. J. Smith, R. W. Blundell *An exogeneity test for a simultaneous equation Tobit model with an application to labor supply*, Econometrica. 54(3) (1986). pp.679-685.

[22] J. Tobin *Estimation of relationships for limited depended variables*, Econometrica. 26(1958), pp. 24-36.

[23] J. S. Thomas *A methodology for linking customer acquisition to customer retention*, Journal of Marketing Research. 38(2)(2001), pp. 262-268.

[24] G. C. G. Wei, M.A. Tanner *A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms*, Journal of the American Statistical Association 85 (1990), pp. 699-704.

[25] L.A. Weissfeld, and H. Schneider *Influence Diagnostics for the Normal Linear Model With Censored Data*. Austal. J. Stat. 32(1) (1990), pp. 262-268.

[26] U.S. Census Bureau, Department of Commerce, National Security and Veterans Affairs. Available at http://www.census.gov (Last Accessed 20, 2012).

[27] U.S. Census Bureau, Current Population Surveys, ASEC 2000 to 2009.
Available at https://www.census.gov/hhes/www/povery/publications (Last Accessed August 20, 2012).

[28] U.S. Department of Health and Human services. Available athttp://aspe.hhs.gov/poverty/08poverty.shtml (Last Accessed August 20, 2012).

[29] U.S. Department of Veterans Affairs. Available at www.va.gov (Last Accessed August 20, 2012).