COMMUNICATIONS IN STATISTICS—SIMULATION AND COMPUTATION®
2017, VOL. 0, NO. 0, 1–14
https://doi.org/10.1080/03610918.2016.1255969

An evaluation of the reconstructed coefficient of determination and potential adjustments

Tatjana Miljkovic[a] and Megan Orr[b]

[a]Department of Statistics, Miami University, Oxford, Ohio, USA; [b]Department of Statistics, North Dakota State University, Fargo, North Dakota, USA

ABSTRACT

Previously, a method was proposed for calculating a reconstructed coefficient of determination in the case of right-censored regression using the expectation–maximization (EM) algorithm. This measure is assessed via simulation study for the purpose of evaluating the utility of model fit. Further, several reconstructed adjusted coefficients of determination are proposed and compared via simulation study for the purpose of model selection. The application of these proposed measures is illustrated on a real dataset.

ARTICLE HISTORY
Received 11 May 2016
Accepted 21 October 2016

KEYWORDS
Coefficient of determination; EM algorithm; Linear models; Regression; Right censoring

MATHEMATICS SUBJECT CLASSIFICATION
62J05; 62J99; 62P25

1. Introduction

Right-censored regression assuming normally distributed errors is commonly used in economics, social science applications, and environmental studies to investigate the relationship between a dependent (or response) variable and one or more independent (or predictor) variables. In this type of regression, the values of the dependent variable are allowed to be right-censored. These models are known as TOBIT models, named after Tobin (1958), or censored regression models. Assumptions of TOBIT models are that the censored dependent variable is of Type I, implying that the censoring level is known in advance, the sample size is fixed, and the number of censored observations is a random variable. In practice, when multiple censoring levels are common, the estimation of the parameters in the right-censored regression can be obtained using the expectation-maximization (EM) algorithm (Miljkovic and Barabanov, 2015). Early work on the least-square regression with censored data can be found in the articles published by Miller (1976) and Miller and Halpern (1982). Good information about censored data and their applications is provided in books by Kalbfleisch and Prentice (2011), Klein and Moeschberger (2003), Breen (1996), and Le (1997).

The problem of assessing model fit in regression analysis is an old one. The coefficient of determination, R^2, is one of the most extensively used measures of goodness of fit for ordinary least-square (OLS) regression models (Draper and Smith, 1998). There appears to be a general consensus on the use of R^2 in the case of a quantitative dependent variable (Menard 2000), and as such, has become a standard part of the regression output produced by all statistical software and packages. However, if data with censored responses are analyzed using these tools and the censoring is ignored (i.e., censored responses are treated as uncensored), the estimated regression coefficients as well as the R^2 estimate will be biased and inconsistent.

CONTACT Tatjana Miljkovic miljkot@miamioh.edu Department of Statistics, Miami University, Oxford, OH 45056, USA.

There is currently no consensus on how a corresponding measure of the strength of association between the dependent variable and a set of predictors should be calculated in the case of right-censored regression. However, numerous pseudo-R^2 measures, analogs of OLS-R^2, have been proposed for measuring the goodness of fit for some common limited categorical dependent variable models, such as logistic regression models (see Tjur, 2012). The most popular pseudo-R^2 is one proposed by McFadden (1973), which is based on the ratio of the log-likelihoods of the full model and the intercept-only model. The characteristics of pseudo-R^2 and its interpretability are similar to that of OLS-R^2 as the pseudo-R^2 also measures the level of improvement in the fit that the full model has over the intercept-only model; the higher the value of the pseudo-R^2, the more improvement the full model has over the intercept-only model. Veall and Zimmermann (1996) pointed out that these goodness-of-fit measures should not be used in cases when the limited dependent variable is continuous since the McFadden's pseudo-R^2 may result in a value greater than 1 when the values of the log-likelihood functions have the opposite sign. Miljkovic and Barabanov (2015) proposed a reconstructed $R_c^2(p)$, an OLS-R^2 analog, that can be used in right-censored regression when the EM algorithm is employed in parameter estimation. This measure of goodness of fit maintains the same properties as the OLS-R^2; as such, its value does not go above 1 when the dependent variable is continuous. However, because the main goal of Miljkovic and Barabanov (2015) was in estimating the model coefficient parameters, the performance of $R_c^2(p)$ in estimating OLS-R^2 was not evaluated.

For the OLS model, a degrees-of-freedom-adjusted R^2 has been developed to provide a penalty as the number of predictor variables increases. Similarly, an adjusted McFadden's pseudo-R^2 has also been developed to adjust for the number of parameters and penalize models with a large number of predictors. These measures were developed to assist in model selection, that is, to determine which subset of independent variables are most useful in predicting the dependent variable. Other likelihood-based measures of fit include Cox and Snell (1971) and Uhler and Cragg (1971). For these two measures, no adjustments have been proposed to take into account the number of predictors; therefore, they cannot be used for model selection. Statistical software such as Stata 14 (2015) and SAS include these four measures as part of the TOBIT regression output.

While the reconstructed R^2, $R_c^2(p)$, proposed by Miljkovic and Barabanov (2015), maintains many favorable characteristics of R^2, it also increases as more predictors are added to the model, regardless of whether the additional predictors are good at explaining the variation in the responses, making $R_c^2(p)$ an unfavorable tool in model selection.

There are two objectives of this article, both of which will be assessed through simulation. First, we will evaluate the performance of $R_c^2(p)$ in estimating OLS-R^2 for right-censored regression. Miljkovic and Barabanov (2015) showed, through simulation, that the estimated regression coefficients obtained using the EM algorithm have small empirical bias, but no simulations were done to evaluate $R_c^2(p)$. Second, motivated by McFadden's adjusted pseudo measures and $R_c^2(p)$, we will propose and compare multiple measures of adjusted coefficients of determination for the purpose of model selection for right-censored regression. All of these measures attempt to balance model fit with model complexity.

This article is organized as follows. In Section 2, we provide the background development of $R_c^2(p)$ as well as the new proposed formulas for $Ra_c^2(p)$, which we call "reconstructed" adjusted coefficients of determination. Section 3 includes the simulation study to evaluate the performance of $R_c^2(p)$ and the proposed $Ra_c^2(p)$ measures. The proposed new measures are tested on a real dataset in Section 4. Concluding remarks are given in Section 5.

2. Methodology

Consider the traditional form of the multiple regression model:

$$\mathbf{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{2.1}$$

where $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ is the vector of responses, $\boldsymbol{\epsilon} \sim N(0, \sigma^2\mathbf{I})$ is the vector of error terms, and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$ is the vector of unknown parameters. The matrix $\boldsymbol{X}$ of size $n \times (p+1)$ is known as the design matrix and assumed to have rank equal to $p+1$ (full column rank). Let $\hat{y}_i$ denote the fitted (estimated) value of y_i based on the vector of estimated regression coefficients, $\hat{\boldsymbol{\beta}}$, and the values x_{ij} of the matrix $\boldsymbol{X}$ (for $j = 0, 1, \ldots, p$ and $i = 1, 2, \ldots, n$). Furthermore, let $\bar{y}_i$ denote the sample mean of y_i. Then, the following formula for the coefficient of determination, R^2, appears throughout the literature:

$$\boldsymbol{R}^2 = 1 - \frac{\text{unexplained variation}}{\text{total variation}}.$$

Now, consider the linear regression model where the response (depended variable) is allowed to be right-censored. Assume $\mathbf{y}$ and $\boldsymbol{z}$ are n_1- and n_2-vectors of uncensored and censored observations, respectively; $n = n_1 + n_2$. Denote by $\tilde{\mathbf{z}}$ the vector of unknown values, which are censored in vector $\mathbf{z}$, and let $\mathbf{y}^* = \binom{\mathbf{y}}{\tilde{\mathbf{z}}}$. The linear regression model now has the form

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

where $\boldsymbol{\epsilon}$ and $\boldsymbol{\beta}$ are the same as in (2.1), and $\boldsymbol{X}^*$ is the design matrix partitioned into two parts, $\boldsymbol{X}^* = \binom{\boldsymbol{X}_1}{\boldsymbol{X}_2}$ corresponding to the uncensored ($\boldsymbol{X}_1$) and censored ($\boldsymbol{X}_2$) observations.

As proposed by Miljkovic and Barabanov (2015), the reconstructed coefficient of determination, $R_c^2(p)$, is derived from the idea of optimizing the following objective function, based on the conditional expectation of the complete data log-likelihood given the observed values and current parameter estimates:

$$J(\beta_0, \beta_1, \ldots \beta_p) = \|\mathbf{y} - \boldsymbol{X}_1\boldsymbol{\beta}\|^2 + \|\boldsymbol{A} - \boldsymbol{X}_2\boldsymbol{\beta}\|^2 + \boldsymbol{B} - \|\boldsymbol{A}\|^2, \tag{2.2}$$

where

$$\boldsymbol{A} = E(\tilde{\mathbf{z}} \mid \tilde{\mathbf{z}} > \mathbf{z}, \boldsymbol{\beta}, \sigma^2) = \boldsymbol{X}_2\boldsymbol{\beta} + \sigma f\left(\frac{\mathbf{z} - \boldsymbol{X}_2\boldsymbol{\beta}}{\sigma}\right), \tag{2.3}$$

$$\boldsymbol{B} = E(\tilde{\mathbf{z}}'\tilde{\mathbf{z}} \mid \tilde{\mathbf{z}} > \mathbf{z}, \boldsymbol{\beta}, \sigma^2) = \|\boldsymbol{X}_2\boldsymbol{\beta}\|^2 + \sigma(\boldsymbol{X}_2\boldsymbol{\beta} + \mathbf{z})' f\left(\frac{\mathbf{z} - \boldsymbol{X}_2\boldsymbol{\beta}}{\sigma}\right) + n_2\sigma^2, \tag{2.4}$$

$f(x) = \frac{\varphi(x)}{\Phi(-x)}$, $\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$, and $\Phi(-x) = \int_{-\infty}^{-x} \varphi(s)\, ds$. $\|\cdot\|$ in (2.2) and (2.4) above represents the Euclidean norm of vectors. Finally, $R_c^2(p)$ is computed as

$$R_c^2(p) = 1 - J_{\text{lin}}(p)/J_{\text{const}}, \tag{2.5}$$

where

$$J_{\text{lin}}(p) = \min_{\beta_0, \beta_1, \ldots, \beta_p} J(\beta_0, \beta_1, \ldots, \beta_p) \tag{2.6}$$

and

$$J_{\text{const}} = \min_{\beta_0} J(\beta_0, 0, \ldots, 0) = J_{\text{lin}}(0). \tag{2.7}$$

Here, the optimal value of the objective function J, if we use the whole design partition matrix $\boldsymbol{X}^*$ is defined as J_{lin} in (2.6), while the optimal value of the J function, if we use only the first column of $\boldsymbol{X}^*$, is denoted as the J_{const} in (2.7). The idea behind building $R_c^2(p)$ is similar to McFadden-R^2 with the exception that the complete-data log-likelihood function is used rather than the observed log-likelihood function. The $R_c^2(p)$ is computed based on the "reconstructed" values of censored points found by imputation. The proposed work can be easily extended to other forms of censoring such as left censoring or interval censoring. For these extensions, the first and second moments of the conditional expectation of censored data (2.3) and (2.4) should be adjusted to reflect these new situations.

The $R_c^2(p)$ maintains many properties of a "good" R^2 proposed by Kvalseth (1985). For example: $R_c^2(p)$ is independent of the units of measurement (unitless), $0 \leq R_c^2(p) \leq 1$, and the function $R_c^2(p)$ is non-decreasing with respect to p.

As previously mentioned, it is inappropriate to use $R_c^2(p)$ for model selection because $R_c^2(p)$ increases as the number of predictors increases. We propose three adjustments to $R_c^2(p)$ and two alternative coefficients of determination that use the likelihood function in order to take into account the number of predictors so that the model fit and model complexity can be balanced. For simplicity, we will refer to these proposed measures as "reconstructed" adjusted coefficients of determination. The following formulas summarize these measures and discuss the rationale for each. Their performance is evaluated through a simulation study discussed in Section 3.

When a dataset contains all uncensored responses, the uncensored coefficient of determination, $R^2(p)$, is adjusted to account for the number of covariates in the model. The formula for this adjustment is

$$Ra^2(p) = 1 - (1 - R^2)\frac{n-1}{n-p-1}, \tag{2.8}$$

where n is the sample size and p is the number of covariates in the regression model.

We will first consider three adjustments to $R_c^2(p)$ similar to (2.8). The first adjustment is equivalent to the uncensored coefficient of determination, but replacing R^2 with $R_c^2(p)$:

$$Ra_n^2(p) = 1 - (1 - R_c^2(p))\frac{n-1}{n-p-1}. \tag{2.9}$$

The second adjustment excludes observations with censored data points from the sample size due to the censoring of the response for these observations resulting in incomplete information. Therefore, the number of observations with uncensored responses, n_1, is used in the reconstructed adjusted coefficient determination instead of the total sample size, n:

$$Ra_{n_1}^2(p) = 1 - (1 - R_c^2(p))\frac{n_1-1}{n_1-p-1}. \tag{2.10}$$

The third adjustment uses what we call the "effective" sample size, n_e, in the formula for the reconstructed adjusted coefficient of determination:

$$Ra_{n_e}^2(p) = 1 - (1 - R_c^2(p))\frac{n_e-1}{n_e-p-1}. \tag{2.11}$$

The value of n_e will fall between n_1 and n and is determined by summing the weights computed for each observation. The weight for an observation quantifies the amount of information we have about the value of the response for that observation. Therefore, a weight of one (the highest possible value for a weight) is given to each observation with an uncensored

response because we have complete information about the value of the response. The weight for a censored data point is computed using the formula

$$w_i = \Phi\left(\frac{z_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}}{s}\right), \tag{2.12}$$

for $i = 1, \ldots, n_2$, where $\Phi(\cdot)$ represents the cumulative density function of a standard normal random variable, z_i is the censoring value for the ith censored response, $\mathbf{x}_i$ is the vector of covariates for the ith censored response, $\hat{\boldsymbol{\beta}}$ is the vector of estimated parameter values, and s is the estimate of σ. In this formula, $\hat{\boldsymbol{\beta}}$ and s are estimates of $\boldsymbol{\beta}$ and σ, respectively, for the regression model with p covariates determined using the EM algorithm.

An assumption of the model in (2.1) is that $y_i \sim N(\mathbf{x}_i'\boldsymbol{\beta}, \sigma^2)$. Thus, the weight in (2.12) is the estimated cumulative density function evaluated at the censoring level z_i and has a value between 0 and 1 for the ith censored observation. The higher the value of z_i falls above the line, the more information we have about the response value for the ith censored observation, the higher value of w_i.

The final two reconstructed adjusted coefficients of determination are related to McFadden's approach to the adjusted pseudo-R^2. Both measures use an estimate of the following log likelihood function:

$$\begin{aligned} l(\mathbf{y}, \mathbf{z}|\boldsymbol{\beta}, \sigma) = {} & \frac{-n_1 \log(2\pi)}{2} - \frac{n_1 \log(\sigma^2)}{2} - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}_1\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}_1\boldsymbol{\beta}) \\ & + \sum_{i=1}^{n_2}\left(1 - \Phi\left(\frac{z_i - \mathbf{x}_i'\boldsymbol{\beta}}{\sigma}\right)\right), \end{aligned} \tag{2.13}$$

where $\mathbf{z} = (z_1, z_2, \ldots, z_{n_2})'$. The first reconstructed adjusted coefficient of determination similar to McFadden's approach is

$$Ra^2_{\text{EM}}(p) = 1 - \frac{l(\mathbf{y}, \mathbf{z}|\hat{\boldsymbol{\beta}}, s) - (p+1)}{l(\mathbf{y}, \mathbf{z}|\hat{\boldsymbol{\beta}}_0, s_0)}. \tag{2.14}$$

In this formula, $\hat{\boldsymbol{\beta}}$ and s are estimates of $\boldsymbol{\beta}$ and σ, respectively, for the regression model with p covariates determined by the EM algorithm. The values $\hat{\boldsymbol{\beta}}_0$ and s_0 are also estimated using this EM algorithm but for the intercept-only model.

The final reconstructed adjusted coefficient of determination is

$$Ra^2_l(p) = 1 - \frac{l(\mathbf{y}, \mathbf{z}|\hat{\boldsymbol{\beta}}_l, s_l) - (p+1)}{l(\mathbf{y}, \mathbf{z}|\hat{\boldsymbol{\beta}}_{0,l}, s_{l,0})}, \tag{2.15}$$

where $\hat{\boldsymbol{\beta}}_l$ and s_l are the estimates of $\boldsymbol{\beta}$ and σ determined by directly maximizing the log likelihood function for the regression model with p covariates given in (2.13) and where $\hat{\boldsymbol{\beta}}_{0,l}$ and $s_{0,l}$ are the estimates that directly maximize the function in (2.13) but for the intercept-only model.

The first three measures ($Ra^2_n(p)$, $Ra^2_{n_1}(p)$, and $Ra^2_{n_e}(p)$) have properties similar to the uncensored adjusted coefficient of determination, $Ra^2(p)$. Most notably, each of these measures will be less than $R^2_c(p)$ as long as the corresponding sample size used (n, n_1, and n_e, respectively) is larger than $p + 1$. Additionally, it is possible for one or all of these measure to be negative if $R^2_c(p)$ is close to zero.

The final two measures ($Ra^2_{\text{EM}}(p)$ and $Ra^2_l(p)$) have properties similar to McFadden's adjusted pseudo-R^2. Although these values are expected to be lower than the first three measures for many datasets, it is possible for $Ra^2_{\text{EM}}(p)$ (or $Ra^2_l(p)$) to be greater than one if the numerator of the fraction in (2.14) (or (2.15)) is positive and the denominator is negative. It is also possible for both the numerator and denominator to be positive, resulting in a decrease in the reconstructed adjusted coefficient of determination as the model fit improves (Veall and Zimmermann, 1996). In such cases, the researcher should be aware of these limitations and proceed with caution when considering the use of these measures.

The methodology developed in this article is implemented using the statistical computing environment R (R Core Team, 2015). To determine the parameter estimates corresponding to $Ra^2_l(p)$, the `optim` function from the base package **stats** in R is used.

3. Simulation studies

First, we evaluate the performance of $R^2_c(p)$ for different sample size and percent of censoring. The simulation study considers the regression model

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

where the value of X_1 from the ith observation is $x_i = i/n$, $i = 1, 2, \ldots, n$, and n is the sample size. The value of X_2 for the ith observation was a randomly selected value in the sequence $\{i/n;\ i = 1, 2, \ldots, n\}$. Thus, the variables X_1 and X_2 are uncorrelated and for this model $\boldsymbol{\beta} = (2, 1, 1)$ and $\epsilon \sim N(0, 0.5^2)$. Different simulation settings were then created by manipulating the sample size, $n = (600, 60)$, and the percentage of points censored, $l_c = (10\%, 30\%, 50\%)$. In each simulation setting, first, all the values of the dependent variable y are generated based on the parameter vector $\boldsymbol{\beta}$ and the distribution of the error terms previously specified. Then, two ordinal levels were created in a two-stage process based on the $(1 - l_c/2)$th quantile of the values of y. These ordinal levels were treated as censoring levels at which the observations above these levels were trimmed. The ordinal nature of the dependent variable is designed to mimic the way censored data are observed in practice, usually related to methodological limitations or confidentiality reasons involved in the data collection (see McKelvey and Zavoina, 1975). For each simulation setting, 10,000 datasets were randomly generated, and the following quantities were computed from each dataset:

- R^2: The coefficient of determination calculated from the original simulated data before trimming/censoring. This is the true value of R^2 based on complete information in the data.
- $R^2_c(2)$: The coefficient of determination after trimming that is computed based on the reconstructed values of the censored observations.
- $\frac{\Delta R^2 - R^2_c(2)}{R^2}$: The difference between R^2 and $R^2_c(2)$ relative to R^2.

The boxplot in Figure 1 shows the distribution of relative differences between R^2 and $R^2_c(2)$ for each combination of sample size and percent of censoring. The first setting corresponds to a large sample size and small percent of censoring and it is clearly seen that the empirical bias of the estimator $R^2_c(2)$ is negligible; however, the empirical bias slightly increases as the percent of censoring increases for the same sample size. The same is observed with a small sample size with larger magnitude of empirical bias and variability corresponding to a large percent of censoring. Overall, from our selected simulation settings, we observe that the best performance of $R^2_c(2)$ is achieved when sample size is 600 with 10% censoring while the most variable results are observed for sample size 60 and 50% censoring.

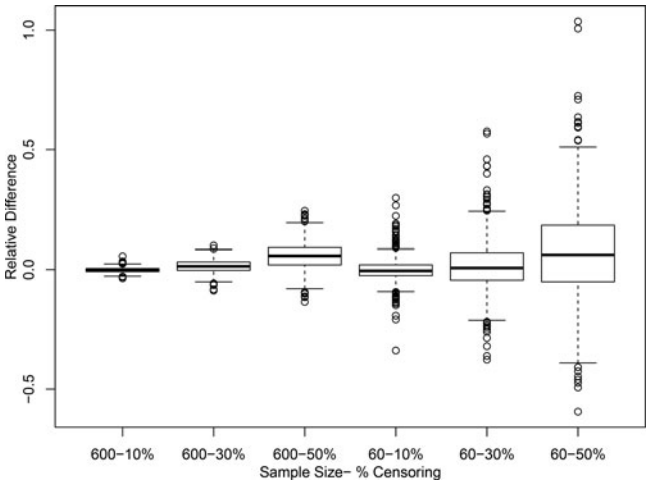

Figure 1. Results of the simulation study for $R^2_c(p)$.

Four additional simulation studies were performed to evaluate the performance of $Ra^2_n(p)$, $Ra^2_{n_1}(p)$, $Ra^2_{n_e}(p)$, $Ra^2_{\mathrm{EM}}(p)$, and $Ra^2_l(p)$ in terms of model selection under different simulation settings. In these studies, the model assumption of normally distributed error terms was met. Similar to the previous simulation studies, the sample size ($n = 600$, 60) and percentage of points censored (10%, 30%, and 50%) were varied to create the different simulation settings. Additionally, different magnitudes of the parameters in $\boldsymbol{\beta}$ values were used. Three models were considered in these simulation studies: a model with 2 potential predictors of the response variable, a model with 4 potential predictors of the response variable, and a model with 8 potential predictors of the response variable. For each model selection method for each simulation setting, the percentage of simulations in which the correct model was selected was determined. Similarly to the previous simulation study already described, 10,000 datasets were randomly generated for each simulation setting.

The first simulation study involved the analysis of data simulation from the model

$$y = 2 + \beta_1 X_1 + \beta_2 X_2 + \epsilon, \tag{3.1}$$

in which $\beta_1 \neq 0$ and $\beta_2 = 0$. The value of $\beta_1 = 0.5$, 1, or 2 (depending on the simulation setting). The value of X_1 for the ith observation was $x_{1i} = i/n$, $i = 1, 2, \ldots, n$, and the value of X_2 for the ith observation was a randomly selected value in the sequence $\{i/n;\ i = 1, 2, \ldots, n\}$. Thus, the variables X_1 and X_2 are uncorrelated. Finally, $\epsilon \sim N(0, 0.5^2)$. Because $\beta_2 = 0$, a model selection criterion correctly selects a model if X_1 is chosen as the sole predictor of the response variable for the model.

The second and third simulation studies used data simulated from the model

$$y = 2 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon.$$

The second simulation study simulated data from this model where $\beta_1 = \beta_2 = 0.5$, 1, or 2 (depending on the simulation setting) and $\beta_3 = \beta_4 = 0$. The values in X_1 and X_2 were simulated using the same methods as in the first simulation study. The values in X_3 were simulated using the same method employed to simulate the values in X_2, making X_3 uncorrelated with the other three predictor variables. The ith value in X_4 was drawn from an $N(x_{1i}, 0.15^2)$ distribution, resulting in X_1 and X_4 having an approximate correlation of 0.90. As with the model in (3.1), $\epsilon \sim N(0, 0.5^2)$.

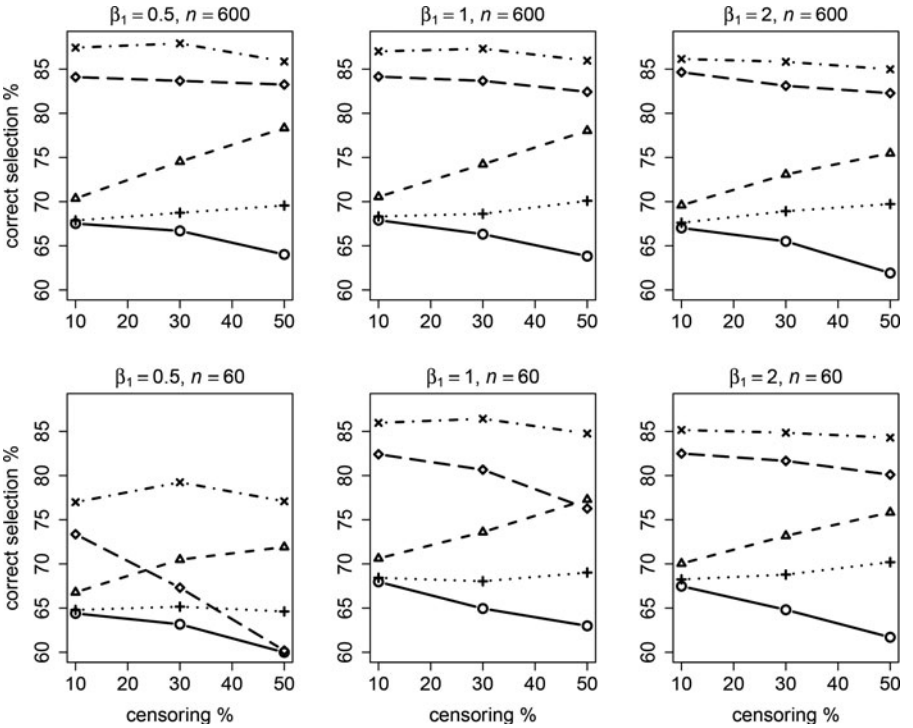

Figure 2. Results of first simulation study comparing model selection for different reconstructed adjusted coefficients of determination. The percentage of times each method resulted in the selection of the correct model at each censoring level is presented in each plot. The symbol for each method are as follows: $Ra_n^2(2)$—circle (O); $Ra_{n_1}^2(2)$—triangle (△); $Ra_{n_e}^2(2)$—plus sign (+); $Ra_{\text{EM}}^2(2)$—cross (×); $Ra_I^2(2)$—diamond (◇).

The third, simulation study simulated data in a similar fashion to the second simulation study with the exception that $\beta_1 = 2\beta_2$.

For both the second and third simulation studies, because $\beta_3 = \beta_4 = 0$, a model selection criterion correctly selects a model if X_1 and X_2 are chosen as the predictors of the response variable.

In order to evaluate the performance of each method when there are a higher number of potential predictors, a fourth, and final, simulation study simulated data from the model

$$y = 2 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \epsilon,$$

where $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 1$ or 2 with the remaining parameters equal to zero. The value of X_1 for the ith observation was $x_{1i} = 1/n$, $i = 1, 2, \ldots, n$. The values in X_2 through X_8 were randomly selected from the sequence $\{i/n;\ i = 1, 2, \ldots, n\}$. Because the first four parameter values are non-zero, a model selection criterion that selects X_1, X_2, X_3, and X_4 as the true predictors selects the correct model.

Figures 2 through 5 present the results of the simulation studies performed to evaluate the different measures of the adjusted coefficient of determination in terms of model selection.

Figure 2 presents the results of the first simulation study. The top three plots in this figure, corresponding to the simulation settings with $n = 600$, all have similar patterns. $Ra_{\text{EM}}^2(2)$ performs best, followed by $Ra_I^2(2)$. Both of these measures decrease slightly in correct model selection percentage as the censoring level increases. The method with the next best performance is $Ra_{n_1}^2(2)$, with a correct model selection percentage that increases as the censoring

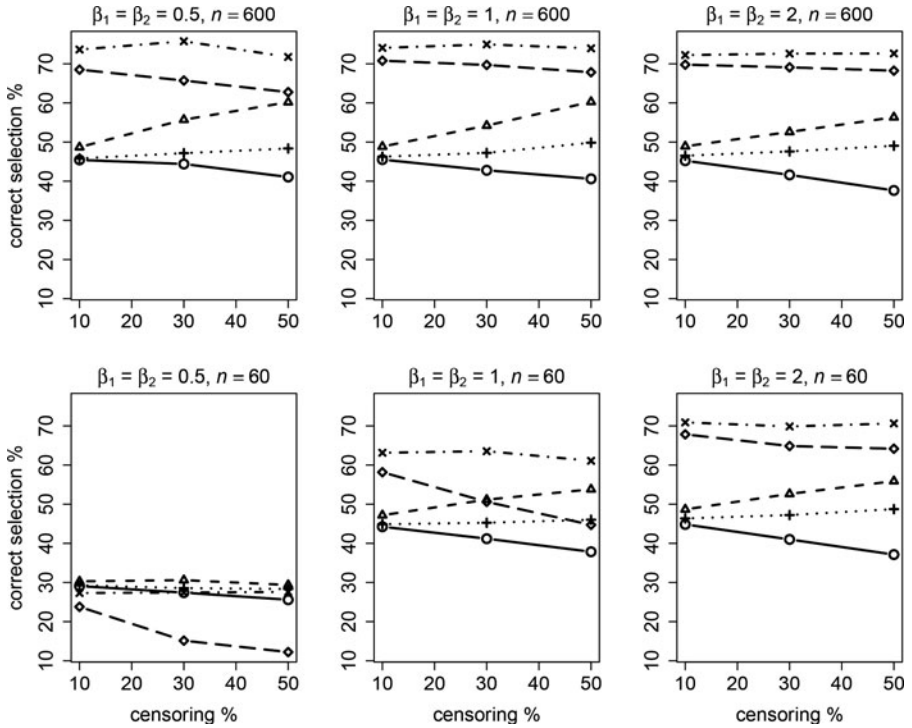

Figure 3. Results of second simulation study comparing model selection for different reconstructed adjusted coefficients of determination. The percentage of times each method resulted in the selection of the correct model at each censoring level is presented in each plot. The symbol for each method are as follows: $Ra_n^2(2)$—circle (O); $Ra_{n_1}^2(2)$—triangle (△); $Ra_{n_e}^2(2)$—plus sign (+); $Ra_{\text{EM}}^2(2)$—cross (×); $Ra_l^2(2)$—diamond (◇).

level increases. $Ra_{n_e}^2(2)$ performs fourth best (second worst) with a correct model selection percentage that slightly increases as the censoring level increases. Finally, $Ra_n^2(2)$ performs the worst with a correct model selection percentage that decreases as the censoring level increases.

We see a similar pattern in the two rightmost bottom plots of Fig. 2. The one exception is that $Ra_{n_1}^2(2)$ outperforms $Ra_l^2(2)$ for the simulation setting when $\beta_1 = 1, n = 60$, and the censoring level is 50%. For the simulation settings with $\beta_1 = 0.5$ and $n = 60$, represented by the leftmost bottom plot, the correct model selection percentage is generally lower compared to the simulation settings represented by the other five plots. However, the relative performance among the five measures remains the same, with the exception of $Ra_l^2(2)$, which is outperformed by both $Ra_{n_1}^2(2)$ and $Ra_{\text{EM}}^2(2)$ for the simulation setting with a 30% censoring level. $Ra_l^2(2)$ is also outperformed by all measures except $Ra_n^2(2)$ for the simulation setting with a 50% censoring level.

Figure 3 shows the results of the second simulation study in which data were simulated from a regression model with four predictors, two of which are true predictors with equal coefficient parameters. Although the correct model selection percentages are lower than in the first simulation study (see Figure 2), we see the same relative pattern in the performances of the five measures for the simulation settings with $n = 600$ as well as the settings with $\beta_1 = \beta_2 = 2$ and $n = 60$. $Ra_{\text{EM}}^2(4)$ performs best, followed by $Ra_l^2(4)$, $Ra_{n_1}^2(4)$, $Ra_{n_e}^2(4)$, and finally $Ra_n^2(4)$. The plot presenting the results of the simulation settings with $\beta_1 = \beta_2 = 1$ and $n = 60$ also has a similar pattern with the exception of the performance of $Ra_l^2(4)$, which decreases more

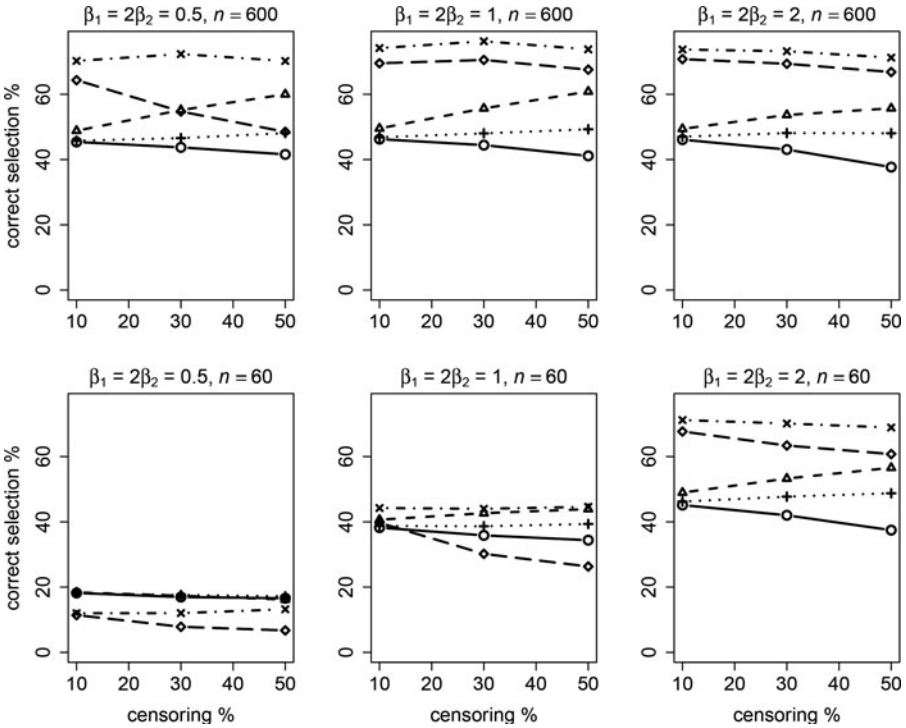

Figure 4. Results of third simulation study comparing model selection for different reconstructed adjusted coefficients of determination. The percentage of times each method resulted in the selection of the correct model at each censoring level is presented in each plot. The symbol for each method are as follows: $Ra_n^2(4)$—circle (O); $Ra_{n_1}^2(4)$—triangle (△); $Ra_{n_e}^2(4)$—plus sign (+); $Ra_{\text{EM}}^2(4)$—cross (×); $Ra_l^2(4)$—diamond (◇).

sharply as the censoring level increases compared to the other settings. The simulation settings with $\beta_1 = \beta_2 = 0.5$ and $n = 60$ show the worst performance by far. $Ra_n^2(4)$, $Ra_{n_1}^2(4)$, $Ra_{n_e}^2(4)$, and $Ra_{\text{EM}}^2(4)$ perform similarly, with $Ra_{n_1}^2(4)$ performing marginally better and $Ra_l^2(4)$ clearly performing the worst.

Figure 4 shows the results of the third simulation study in which data were simulated from a regression model with four predictors, two of which are true predictors with the coefficient parameter for one of these predictors twice the coefficient parameter of the other predictor. The two rightmost upper plots, corresponding to the simulation settings with the two largest parameter coefficient values and $n = 600$, and the rightmost lower plot, corresponding to the simulation settings with the largest parameter coefficient values and $n = 60$, show similar patterns to many of the plots in the previous figures. $Ra_{\text{EM}}^2(4)$ performs the best, followed by $Ra_l^2(4)$, $Ra_{n_1}^2(4)$, $Ra_{n_e}^2(4)$, and finally $Ra_n^2(4)$. The leftmost upper plot, corresponding to the simulation settings with the smallest parameter coefficient values and $n = 600$, shows a similar pattern, except that $Ra_l^2(4)$ decreases more sharply than in the three plots previously mentioned. Similar to what we observed in Fig. 3, the leftmost bottom plot, corresponding to the simulation settings with the smallest parameter coefficient values and $n = 60$, shows the worst performance in model selection for all five measures. $Ra_{n_1}^2(4)$, $Ra_{n_e}^2(4)$, and $Ra_n^2(4)$ perform similarly, followed by $Ra_{\text{EM}}^2(4)$ and $Ra_l^2(4)$. Finally, the middle bottom plot, corresponding to $\beta_1 = 2\beta_2 = 1$ and $n = 60$, shows a different pattern. $Ra_{\text{EM}}^2(4)$ and $Ra_{n_1}^2(4)$ perform relatively similarly, followed by $Ra_{n_e}^2(4)$, $Ra_n^2(4)$, and finally $Ra_l^2(4)$.

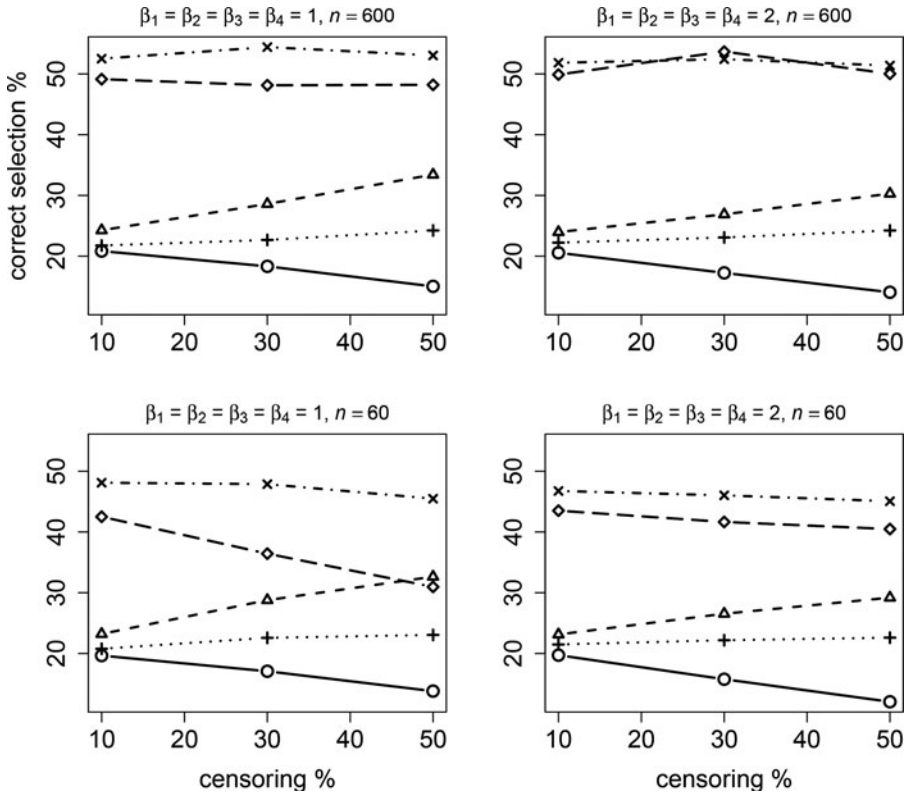

Figure 5. Results of fourth simulation study comparing model selection for different reconstructed adjusted coefficients of determination. The percentage of times each method resulted in the selection of the correct model at each censoring level is presented in each plot. The symbol for each method are as follows: $Ra^2_n(4)$—circle (O); $Ra^2_{n_1}(4)$—triangle (△); $Ra^2_{n_e}(4)$—plus sign (+); $Ra^2_{\text{EM}}(4)$—cross (×); $Ra^2_l(4)$—diamond (◇).

Figure 5 shows the results of the final simulation study involving eight potential predictors. The results of this study show the following general ranking of model selection criteria: $Ra^2_{\text{EM}}(8)$, $Ra^2_l(8)$, $Ra^2_{n_1}(8)$, $Ra^2_{n_e}(8)$, and $Ra^2_n(8)$. These results are similar to what was observed in the third simulation study in terms of relative performance of the measures.

4. Real data

We examine the veterans' health benefits grants dataset for the years 2000 to 2010, provided by the North Dakota Department of Veterans Affairs and first analyzed by Miljkovic and Barabanov (2015). The dataset includes 575 applications with 48% of censored data. The authors provide a thorough discussion about this dataset. The censored dependent variable, y, under consideration, in the right-censored regression setting, is the amount of health benefit paid to low income veterans which was capped at $750 in year 2006 and $1,000 in year 2010. The independent variables under consideration are: application year (x_1), age of the applicant (x_2), gender (x_3), income level (x_4), spousal status (x_5), marital status (x_6). We consider the following models for which we compute $Ra^2(p)$ (ignoring censoring), $Ra^2_n(p)$, $Ra^2_{n_1}(p)$, $Ra^2_{n_e}(p)$, $Ra^2_{\text{EM}}(p)$, and $Ra^2_l(p)$:

Model-1: $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6$

Model-2: $E(y) = \beta_0 + \beta_1 {x_1}^2 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6$

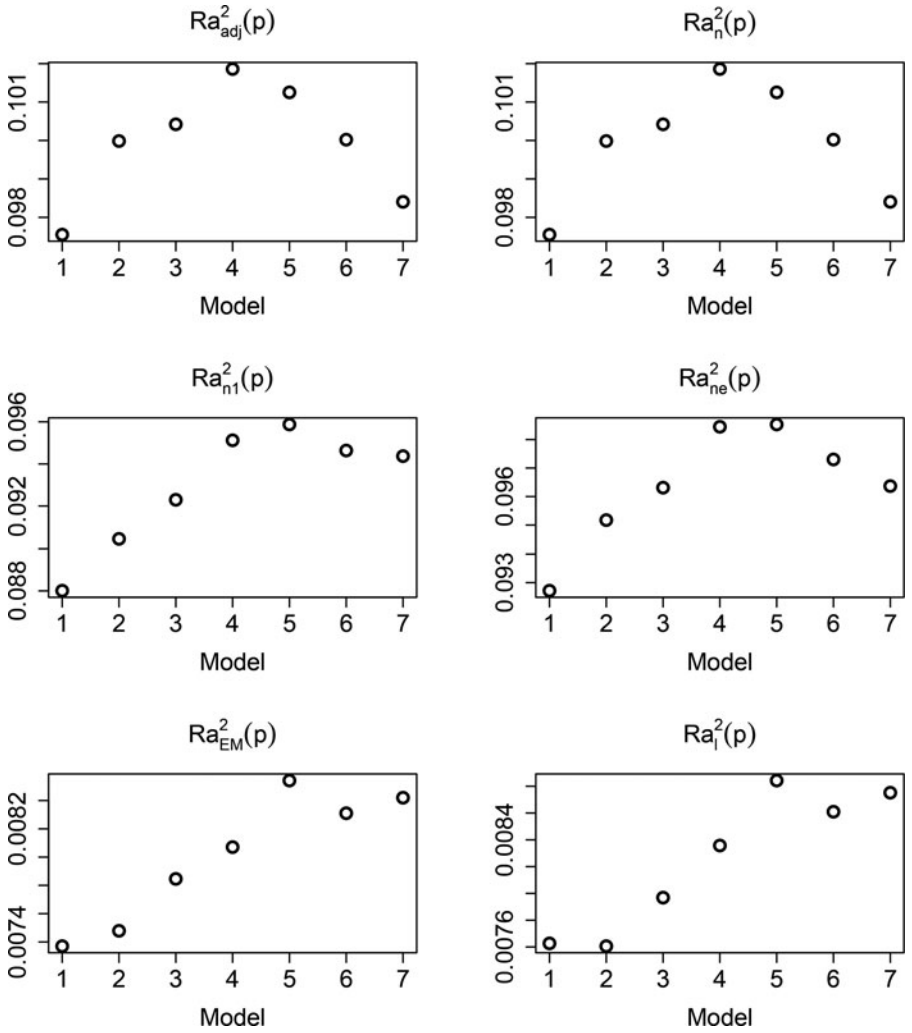

Figure 6. Optimal model selection. Top left is Ra^2(OLS); top right is $Ra^2_n(p)$; middle left is $Ra^2_{n_1}(p)$; middle right is $Ra^2_{ne}(p)$; bottom left is $Ra^2_{\text{EM}}(p)$; bottom right is $Ra^2_l(p)$.

Model-3: $E(y) = \beta_0 + \beta_1 x_1{}^2 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_6$
Model-4: $E(y) = \beta_0 + \beta_1 x_1{}^2 + \beta_2 x_2 + \beta_3 x_3 + \beta_5 x_6$
Model-5: $E(y) = \beta_0 + \beta_1 x_1{}^2 + \beta_2 x_3 + \beta_5 x_6$
Model-6: $E(y) = \beta_0 + \beta_1 x_1{}^2 + \beta_2 x_4 + \beta_5 x_6$
Model-7: $E(y) = \beta_0 + \beta_1 x_1{}^2 + \beta_5 x_6$.

Figure 6 shows the summary of the results for these six measures used in the model selection. The optimal model is Model-5 according to $Ra^2_{n_1}(p)$, $Ra^2_{n_e}(p)$, $Ra^2_{\text{EM}}(p)$, and $Ra^2_l(p)$ with the values 0.09588, 0.09853, 0.00854, and 0.00884, respectively. Model-4 is selected as the optimal model when using $Ra^2_{n_e}(p)$. For this data, the maximum value of the adjusted coefficient of determination is achieved when using Formula (2.11) that includes the penalty term based on the effective sample size, n_e. This result alone indicates that the goodness of fit for Model-5 is improved by 9.59% over the intercept model. Therefore, the same percent represents the amount of variability explained when using this model.

Miljkovic and Barabanov (2015) selected Model-5 based on the highest value of $R_c^2(p)$. If censoring is ignored, the OLS-based adjusted coefficient of determination of 0.2541 is the highest for Model-4. The model would not only be incorrectly selected using this criteria but also the magnitude of the adjusted coefficient is more than twice higher compared to that based on the EM algorithm. While we observe that the value of these results are low, Veall and Zimmermann (1996) noted that for some type of applications R^2 will be low. For example, it is expected for a regression using microdata on labor supply to have an R^2 of 0.1 while a regression of macroecomometric variable using county data may have R^2 of 0.4. Low R^2 values are also observed in social and behavioral sciences. This phenomenon is explained by that fact that due to the complexity of the problems studied in these areas, we do not expect the models to include all relevant predictors to explain the response, many other variables may not be accounted for in the model specification. In other fields of study where the R^2 is higher, it is easier to specify complete and well-defined models. Examples of very low R^2 values can be found in the study of intertemporal asset pricing by Shanken (1990). Further discussion about R^2 measure of explanatory power can be found in the book by Neter et al. (1996).

5. Recommendations and conclusion

In this article, we evaluated the performance of $R_c^2(p)$ in estimating OLS-R^2 for right-censored regression, proposed by Miljkovic and Barabanov (2015). It was found that the overall performance of $R_c^2(p)$ depends on the samples size and percent of censoring. A slight increase in the empirical bias and noticeable increase in the variability is observed as the sample size tends to decrease with an increase in the proportion of censored data. However, the $R_c^2(p)$ performs well for a large sample size (e.g., $n = 600$) and small percent of censored data (e.g., 10%).

We also evaluated five proposed reconstructed adjusted coefficients of determination used for model selection in right-censored regression. Based on the results of the simulation study, we generally recommend the use of $Ra^2_{\text{EM}}(p)$ for model selection as this measure performed the best in the majority of the simulation settings among the five methods tested . However, as already mentioned, there are limitations to $Ra^2_{\text{EM}}(p)$. More specifically, this measure can be greater than one if the numerator and denominator have opposing signs in the fraction in (2.14) or this measure can decrease as the model fit improves if both the values of the numerator and denominator in (2.14) are positive. Thus, the researcher must pay attention to the value of the estimated likelihoods in (2.14) to verify that $Ra^2_{\text{EM}}(p)$ is valid for a specific dataset. If the researcher decides not to use $Ra^2_{\text{EM}}(p)$ for model selection, then $Ra^2_{n_1}(p)$ is recommended because it performed the best among the three measures that adjusted the reconstructed coefficient of determination ($Ra^2_n(p)$, $Ra^2_{n_1}(p)$, and $Ra^2_{n_e}(p)$) in the overwhelming majority of simulation settings.

We illustrated the performance of these measures to a real dataset related to veterans' benefits. For this dataset the values of estimated likelihoods used in (2.14) and (2.15) are the same, so the $Ra^2_{\text{EM}}(p)$ is consistent with $Ra^2_{n_1}(p)$ when used in the model selection. These results are expected based on the findings from our simulations study, and they are in line with our recommendations.

Also note, as already mentioned, that the proposed work can be easily extended to other forms of censoring such as left censoring or interval censoring. For these extensions, the first and second moment of the conditional expectation of censored data (2.3) and (2.4) will need to be adjusted to reflect these new situations. However, all of the goodness-of-fit

measures surveyed in this article would remain the same. While we recognize that some applications of interval censoring and left censoring exist, we focused exclusively on right censoring as we believe this situation occurs more often in the applications that we considered in this article.

References

Breen, R. (1996). *Regression Models: Censored, Sample-Selected, or Truncated Data*. Thousand Oaks: Sage Publication.

Cox, D., Snell, E. (1971). On test statistics calculated from residuals. *Biometrika* 58(3):589–594.

Draper, N. R., Smith, H. (1998). *Applied Regression Analysis*. 3rd ed. New York: John Wiley.

Kalbfleisch, J. D., Prentice, R. L. (2011). *The Statistical Analysis of Failure Time Data*. 2nd ed. New York: John Wiley & Sons.

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

Kvalseth, T. O. (1985). Cautionary note about R^2. *The American Statistician* 39(4):279–285.

Le, C. T. (1997). *Applied Survival Analysis*. New York: John Wiley and Sons, Inc.

McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior. In: Zarembka, P., ed. *Frontiers in Econometrics*. New York: Academic Press, pp. 105–142.

McKelvey, R. D., Zavoina, W. (1975). A statistical model for the analysis of ordinal level dependent variable. *Journal of Mathematical Sociology* 4:103–120.

Menard, S. (2000). Coefficients of determination for multiple logistic regression analysis. *The American Statistician* 54(1):17–22.

Miljkovic, T., Barabanov, N. (2015). Modeling veterans' health benefit grants using the EM algorithm. *Journal of Applied Statistics* 42(6):1166–1182.

Miller, R. G. (1976). Least squares regression with censored data. *Biometrica* 63(3):449–464.

Miller, R. G., Halpern, J. (1982). Regression with censored data. *Biometrica* 69(3):521–531.

Neter, J., Kutner, M. H., Nachtsheim, C. J., Wasserman, W. (1996). *Applied Linear Statistical Models*. Chicago: Irwin.

Shanken, J. (1990). Intertemporal asset pricing: An empirical investigation. Journal of Econometrics 45(1–2):99–120.

R Core Team, (2015). R: A Language and Environmentfor Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. Available at: https://www.R-project.org/.

Tjur, T. (2012). Coefficients of determinations in logistic regression models- A new proposal: The coefficient of discrimination. *The American Statistician* 63(4):366–372.

Tobin, J. (1958). Estimation of relationship for limited dependent variables. *Econometrica* 26:24–36.

Uhler, R. S., Cragg, J. G. (1971). The structure of the asset portfolios of households. *The Review of Economic Studies* 38(3):341–357.

Veall, M. R., Zimmermann, K. F. (1996). Pseudo-R^2 measures for some common limited dependent variable models. *Journal of Economic Surveys* 10(3):241–259.