A Regression-Based Approach for Mediation Analysis with Censored Data

N A Salehah\textsuperscript{1} and Suhartono\textsuperscript{1}  
Department of Statistics, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia  
e-mail: noviajeng.s@gmail.com

Abstract. The regression method is one of the most applied approaches to estimate the direct and indirect effect of the causal variable \((X)\) to an outcome \((Y)\) with any mediating variable \((M)\) in their mechanism. Censored or limited data on any one of the inputs, mediators, and output often observed in a randomized trial with a specific follow-up time or with certain conditions and limitations caused by a particular setting. In this study, several regression strategies were employed to estimate each different path parameter, i.e., the effect of \(X\) on \(M\) \((a)\), the effect of \(M\) on \(Y\) \((b)\), and the effect of \(X\) on \(Y\) \((c')\) for a single mediator model. A simulation study was conducted under various conditions to assess the performance of the proposed strategy while dealing with a censored mediator. A continuous mediator and outcome are observed in this study so that the product method can be applied to understand the estimation of the indirect effect. Based on the results, it is clearly shown that a higher censoring percentage in the mediating variable yields higher bias estimate. By applying Tobit model in \(M\)-regression can minimize the bias of estimate effect with higher censoring percentage. Furthermore, it can conclude that the proposed strategy can effectively perform the estimation of indirect effect with a censored mediator.

Keyword: Regression, mediation analysis, censored data

1. Introduction
In the last few decades, many studies have tried to address the question about mediation \cite{1,2}. Mediation analysis is a statistical approach that widely used to examine the direct and indirect effect of exposure to the outcome by including one or more mediating variables.

\[
\begin{align*}
E[Y | M] &= \delta_0 + cY + \epsilon_0 \\
E[M | A] &= \delta_M + aX + \epsilon_M
\end{align*}
\]

\textbf{Figure 1.} Single Mediator Model

Figure 1 illustrates a single mediator model that can be specified using the following regression.
\[ E[Y \mid A, M] = \delta_Y + c'X + bM + \varepsilon_Y \] (3)

In this single mediator case, \( X \) denotes the initial or causal variable, \( M \) denotes the mediating variable, \( Y \) denotes the outcome variable, and \( \varepsilon_0, \varepsilon_M, \varepsilon_Y \) denote the error term of the model. If one did not consider any mediating variable then the first equation will depict that \( X \) is associated with \( Y \) (path \( c \)). The second equation shows that \( X \) is associated with \( M \) (path \( a \)) and the third equation shows that \( M \) affect \( Y \) (path \( b \)) while controlling for \( X \) (path \( c' \)). In mediation analysis, the effect of \( X \) to \( Y \) in the mechanism is decomposed into direct and indirect effect. From these models, \( c' \) represents the direct effect of \( A \) to \( Y \) and \( ab \) represents the indirect effect of \( A \) to \( Y \) through \( M \). This is well-known as product method approach \([3]\).

Mediation analysis has been applied in various research fields, such as epidemiology \([4,5]\), psychology and social science \([6,7]\). In these researches, the investigators might observe censored or limited data instead of complete data \([8]\). The phenomenon of censored data might happen due to certain settings with any interfering policy or any applied threshold. In the setting of a continuous mediator and continuous outcome, OLS approach can be applied to estimate the value of each path, but if there are any one of the mediator and output is censored, traditional OLS may produce bias estimate of \( a, b \) or \( c' \).

It may cause severe consequences of effect estimates and the conclusion of the mediation analysis. Many general approaches have been proposed and applied to deal with censored data \([9,10]\). Censored regression model or also known as Tobit regression is one of regression approach that is widely used to analyze such case. Tobit model is widely applied in many advances, such as econometrics \([11,12]\), public health \([13]\), and social demography \([14]\).

Among the many existing methods for mediation analysis with censored data, the complete case approach is the easiest one to apply, which done by merely omitting the observations with censored data. However, such an analysis potentially yields a biased and inefficient estimate because it does not utilize all observed observation \([15,16]\). Imputation approach was also necessarily popular due to the missingness information in the data \([17]\). Wang and Zhang \((2011)\) introduce the use of Tobit mediation model to deal with censored data with continuous outcome \([8]\). Many approaches then further developed to deal with many conditions of data, i.e. binary mediator and binary outcome by utilizing an advance computational setting \([16,18]\). However, in some fields, more interpretable methods, i.e. regression model is more preferable and easier to understand.

The aim of this study is to investigate the estimation of each path using several regression scenarios to evaluate mediated effects for a single mediator model, through a simulation study. This study will focus on right-censored mediator and use complete observations on other variables. Tobit regression will be employed to estimate the model parameter in the setting of censored data. This study is organized as follows. Section 2 describes the general overview of causal mediation analysis and Tobit model. Section 3 describes the methodology and simulation study designs, materials, scenarios. Section 4 performs the analytical results of the simulation study. The conclusion and discussion of this study will be performed in Section 5.

2. Methodology

2.1. Mediation Analysis

The mediated or indirect effect is the effect of \( X \) which indirectly affect \( Y \) through the mediator (the \( a \) and \( b \) path in Figure 1). Moreover, there is also an effect in which \( X \) directly affects \( Y \) in the model (the \( c' \) path in Figure 1), called as the direct effect. In the case in which both \( M \) and \( Y \) are continuous variables, the sum of the indirect effect and the direct effect is the total effect (the \( c \) path in Figure 1). There are two different ways the indirect effect can be quantified, 1) the product of coefficients estimate \( a \times b \) (known as \( ab \)) and 2) the difference in coefficients estimate: \( c - c' \). These two quantities is equal \( (ab = c - c') \) only when \( M \) and \( Y \) are both continuous variables and the sample size for estimating the parameters for \( M \)-regression and \( Y \)-regression equations are the same \([3]\). If the mediator and/or the outcome are non-continuous variables (e.g., a categorical variable, a survival variable) then the return
value of product method and difference method may not be same. In this paper, the product of coefficients $(ab)$ estimate is chosen over the difference in coefficient $c - c'$ because this study investigates the effect of censored mediator in the mediation analysis. Besides that, the statistical tests of the $ab$ estimate performed better than the $c - c'$ estimation \[19\].

### 2.2. Tobit Model

Let $y'$ be the latent true scores and $y$ be the observed scores. Suppose $c_L$ and $c_R$ be constant which denotes the left threshold and right threshold, respectively. Tobin (1958) formalized the mechanism of assigning censored data as [20],

$$y = \begin{cases} c_L, & \text{if } y' \leq c_L \\ y', & \text{otherwise} \end{cases} \quad (4)$$

which is known as left-censoring, while right censoring case illustrated as,

$$y = \begin{cases} c_R, & \text{if } y' \geq c_R \\ y', & \text{otherwise} \end{cases} \quad (5)$$

Censoring arise when cases which have value at or above some threshold (left-censoring) and at or below some threshold (right-censoring), all take on the value of that threshold so that the true value might be equal to the threshold, but it might also be higher (left-censoring) or lower (right-censoring).

Censored regression models or often called as Tobit model is one type of regression model that widely used to illustrate the linear relationship between censored dependent variables and independent variables. The general Tobit regression model can be stated as,

$$y' = \beta^* x + \varepsilon$$

with $\varepsilon$ is normally distributed with mean 0 and variance $\sigma^2$ [21].

### 3. Simulation Study

The aim of this simulation study is to investigate the performance of different regression strategies of evaluating a mediation effect (i.e., indirect effect) in a single mediator model with censored mediator. This study focuses on right-censoring data as an example.

The single-mediator model consists of fitting two regression models: the $M$-regression and the $Y$-regression. Based on Figure 1, it is shown that the $a$-path coefficient is estimated from the $M$-regression and the $b$-path and $c'$-path coefficients are estimated from the $Y$-regression. Because this study focuses on a continuous mediator and continuous outcome, OLS regression can be employed for $M$-regression and $Y$-regression. Besides that, this study also investigates the use of Tobit regression for estimating $a$-path when $M$ is censored.

Four different factors were manipulated for data generation. The four factors were 1) the size of $a$-path (0.2 and 0.6), 2) the size of $b$-path (0.15 and 0.5), 3) the censoring percentage ($cp$) of $M$ (0.1, 0.3, 0.5), and 4) the sample size (100 and 500). In total, there are 24 combinations of factor used in this study. The data was generated with following conditions:

$$X \sim U(\min_x, \max_x) \quad (7)$$

$$M = \delta_M + aX + \varepsilon_M, \varepsilon_M \sim N(0, \sigma_M^2) \quad (8)$$

$$M = \begin{cases} C, & \text{if } M > C \\ M, & \text{if } M \leq C \end{cases} \quad (9)$$

$$Y = \delta_Y + c'X + bM + \varepsilon_Y, \varepsilon_Y \sim N(0, \sigma_Y^2) \quad (10)$$

$C$ denotes the cut point where $M$ is censored. This cut point was decided by the value of $(1 - cp)$ th percentile of $M$. The value of $\delta_M$, $\delta_Y$, $\sigma_M^2$, and $\sigma_Y^2$ were set into certain constant. The data was replicated 1000 times. The illustration of right-censored mediator with different censoring percentage was illustrated in Figure 2. It can be seen that if the data is censored, the distribution of the observed data
becomes skewed (in light blue coloured) and does not represent the actual data distribution (black coloured histogram).

**Figure 2.** Illustration of Data with a) 10%, b) 30%, and c) 50% Censoring

There are three approaches applied to analyze the mediation analysis with the censored mediator which are,
- Using OLS for \( Y \)-regression with censored \( M \)
  
  By using original data containing censored mediators, OLS regression is employed in \( M \)-regression and \( Y \)-regression.
- Using OLS for \( Y \)-regression using only uncensored \( M \) (Complete-Case Analysis)
  
  This approach will assume the censored observation as missing data. Different missing data mechanism may determine different missing data techniques used [22]. By using MCAR assumption, listwise deletion (a.k.a. complete case analysis) can be used. Thus, the observations with the censored mediator will be discarded from the study. Note that the number of used sample may differ (reduced) from the original sample size.
- Using Tobit Regression for \( M \)-regression and OLS for \( Y \)-regression
  
  In the third approach, \( a \)-path in \( M \)-regression will be estimated using Tobit regression, while \( b \) and \( c' \) paths in \( Y \)-regression will be estimated using OLS.

Three different methods were employed to evaluate the indirect effect of each approach, i.e. 1) raw bias, 2) relative bias, and 3) MSE. The formula of each method can be written as follows:

\[
\hat{\text{bias}} = \theta - \hat{\theta}
\]

(1)

\[
\text{relative bias} = \frac{\theta - \hat{\theta}}{\theta}
\]

(2)

\[
MSE = \frac{1}{R} \sum_{i=1}^{R} (\theta - \hat{\theta})^2
\]

(3)

where \( R \) is the number of replication.

4. Results and Analysis

The censored data were expected to affect the estimated value of the mediated effect a.k.a. indirect effect which is the product of \( a \)-path and \( b \)-path (\( ab \)). From 1000 replications with 24-factor combinations, the estimates of \( a \), \( b \), and \( c' \) were all in a reasonable range. There is no noticeable difference in estimated value of \( c' \) or direct effect which indicate no effect of any changes in the values of \( a \) or \( b \) on \( c' \). Besides that, there were no non-convergence or improper solutions while fitting the model to the datasets.

As illustrated in Figure 3, when the size of \( a \)-path is manipulated, variations in the estimated results of \( a \)-path and \( b \)-path become narrower whereas when the size of \( b \)-path is manipulated, there is no significant change in the estimated value of \( a \) or \( b \). This result shows that when the influence of \( X \) on \( M \) is getting bigger, both the OLS and the Tobit model will produce more consistent estimate of \( a \).
Figure 3. Estimate parameter of a-path and b-path using each Approach with: a) $a = 0.2; b = 0.15$, b) $a = 0.2; b = 0.5$, c) $a = 0.6; b = 0.15$, and d) $a = 0.6; b = 0.5$
The censoring percentage (cp) of M has an essential role in the estimation of OLS regression parameters, especially the estimated values for a-path and the variations of estimated b-path. By applying OLS in approach 1 (raw data) and approach 2 (complete-case analysis), higher censoring percentage will lead to more bias estimate value of a. The OLS method tends to produce underestimate value of a. Also, it can be noted that higher the cp greater the diversity of the estimated value b. Moreover, when the effect of X on M(a-path) is higher then the complete-case analysis is able to produce the estimated parameters of the M-regression model better than OLS using original data (censored M) when the censoring percentage increases. This shows that a stronger association between X and M will minimize the bias in estimation using complete case analysis.

Least square regression that uses all observed values regardless of the censoring status (approach 1) will produce bias estimation of model parameter because the censored data will affect the regression parameter, i.e., the regression slope [23]. Whereas in the complete case analysis (approach 2), the parameter estimate will be biased due to omitting process of observations with censored M, so that the sample size is reduced and does not imply the whole population, but only observations with uncensored M.

Based on the results, it is shown that censored data will affect the value of indirect effects, especially those caused by bias in estimation a-path. The Tobit model can produce unbiaes estimated values of a even though the cp value is higher. When the values of a and b are manipulated, the Tobit model is able to produce estimates that are consistently unbiased. This shows that the Tobit approach can effectively perform the estimation of a, and minimize the possibility of biased indirect effect estimation. Generally, the sample size affects the estimated value and standard error of a and b. If the sample size is larger, the estimated parameter will be more consistent and closes to the true value.

Several measurements were employed to evaluate the performance of each approach used in this study. The average raw bias, average relative bias, and MSE of each scenario were illustrated in Figure 4. The raw bias was calculated using the difference between the true indirect effect (θ) and the estimated indirect effect (θ̂). The average raw bias and average relative bias show how accurately the model parameters recover the true values for a given condition while the MSE captures the variance as well as the bias of an estimate.

Based on Figure 4, in general, a bigger sample size tends to minimize the raw bias, relative bias, and MSE. The underestimate value of a had an impact on the higher and positively signed value of raw bias, relative bias, and MSE. Under many different conditions, the censoring percentage is necessarily affecting the bias rate of estimated IE. Among all censored data approaches, utilizing Tobit model in M-regression produces the smallest bias and MSE. Besides the result of estimated a, b and c’ paths, these results show that the proposed regression strategy by applying Tobit model for censored M can effectively perform the estimation of indirect effect.

5. Discussions
This study investigates three different strategies to estimate indirect effect when M is censored based on simulation study in several different conditions. A regression approach is chosen as the focus of study because it is easy to use and more interpretable. The first approach applies OLS regression on both Y-regression and M-regression using raw data (data with censored M), while the second approach applied the same regression approach but only using observation with uncensored M, known as complete-case analysis. Due to the popularity of Tobit model while handling censored data, this study proposed new strategy by applying Tobit regression in M-regression and utilizing OLS in Y-regression. Based on the simulations, it is clearly shown that the censoring percentage has an essential effect on estimated a and b paths, which correspond to the estimated value of indirect effect. The effect of the size of a-path, b-path, and the sample size were also clearly discussed. Least square regression that uses all observations without considering the censoring status and the reduction of sample in the complete-case analysis may lead to produces bias estimate when the censoring percentage is higher [23,24]. However, among all
three approaches, Tobit model yields the smallest raw bias, relative bias, and MSE. It shows that the proposed regression strategy can be adapted to estimate the indirect effect on mediation analysis.

Figure 4. Average Raw Bias, Relative Bias and MSE of Estimate IE using Each Approach

References
[1] Baron R M and Kenny D A 1985 Journal of Personality and Social Psychology 51(6) 1173-1182.
[2] Pearl J 2001 In: Proceedings of the seventeenth conference on uncertainty and artificial intelligence 411-420.
[3] MacKinnon D P 2008 Introduction to statistical mediation analysis. New York: Routledge.
[4] VanderWeele T J and Vansteelandt S 2010 American Journal of Epidemiology 172(12) 1339–1348.
[5] Lange T, Hansen K W, Sørensen R, and Galatius S 2017 Epidemiol Health 39.
[6] Buffardi L E and Campbell W K 2008 Personality and Social Psychology Bulletin 34.
[7] Rucker D D, Preacher K J, Tormala Z L, and Petty R E 2011 Social and Personality Psychology Compass 359-371.
[8] Wang L and Zhang Z 2011 *Structural Equation Modeling: A Multidisciplinary Journal* **18**(1) 18-34
[9] Miller R and Halpern J 1982 *Biometrika* **69**(3) 521–531.
[10] Muthén B O 1989 *British Journal of Mathematical and Statistical Psychology* **42** 241–250.
[11] McDonald J F and Moffitt R A 1980 *The Review of Economics and Statistics* **62**(2) 318-321
[12] Moffitt R A 1999 *Handbook of Labor Economics* 1367–1397.
[13] Austin P C, Escobar M, and Kopec J A 2000 *Quality of Life Research* **9**(8) 901–910.
[14] Tansel A and Bircan F 2006 *Economics of Education Review* **25**(3) 303–313.
[15] White I R and Carlin J B 2010 *Statistics in Medicine* **29**(28) 2920-2931.
[16] Wang J and Shete S 2017 *Stat Methods Med Res* **27**(10) 3010-3025.
[17] Vansteelandt S, Bekaert M, and Lange T 2012 *Epidemiologic Methods* **1**(1)
[18] Wang L, Ning J, and Shete S 2018 *Statistics in Medicine* 1-17
[19] MacKinnon D P, Lockwood C M, Hoffman J M, West S G, and Sheets V 2002 *Psychological Methods* **7**(1) 83-104.
[20] Tobin J 1958 *Econometrica* **26** 24-36.
[21] Greene W H 1990 *Econometric Analysis* New York: Macmillan Publishing Company.
[22] Shafer J L and Graham J W 2002 *Psychological Methods* **7**(2) 147-177.
[23] Long J S 1997 *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks, CA: Sage.
[24] Zhang Z and Wang L 2012 *Psychometrika* **78**(1) 154–184.