International Journal of Science and Mathematics Education
Vol. 1, No.4 December 2024
e-ISSN: 3062-7214; p-ISSN: 3062-7206, Hal 08-17
DOI ; https://doi.org/10.62951/ijsme.v1i4.32
Available online at: https://international.arimsi.or.id/index.php/IJSME

Lasso Tobit Principal Component Regression With An Application

Hameedah Naeem Melik *
University of Al- Qadisiyah, Iraq
Email:*hameedah.naeem@qu.edu.iq**

Abstract, *One of the most crucial subjects in the analysis of statistical models is the identification of important variables. Therefore, the search for best variable selection methods is a good in obtaining best estimators. The Lasso method is considered the most effective approach for variable selection and parameter estimation in building statistical models with high explanatory power in representing the studied phenomenon. Therefore, using the Lasso method to estimate the parameters of a regression model that contains a dependent variable with data that is censored at zero can be achieved through the use of Lasso tobit principal component regression, it has attractive properties in estimating the parameters of this model. The our proposed method is illustrated via simulation scenario and a new real data .*

Keywords: *Lasso ,component selection ,tobit principal component regression, variable selection.*

1. INTRODUCTION

Since the first work of (Tobin ,1958),tobit regression model has extended In many field of knowledge, Such as: demand for medical care(Duan, N et al.1983)), Medical Sciences (Amemiya, T. (1984)), health status (Austin, P. C et al.(2000)), econometrics(Maddala, G. S. (1993), and so on. For the purpose of estimating coefficients with a censored response variable at the zero censored point, the tobit regression model is crucial(Amemiya, T. (1984)).Therefore , The dependent variable of tobit model has a mixed distribution, with a continuous distribution for the non-limit data and a discontinuous distribution at the lower limit (Amore et al, M. D(2021)). he Tobit model take care of estimation a linear relationship between the response variable and the explanatory variables. Nevertheless, only the non-limit observations show this linear relationship. At the lower limit, the dependent variable is suppressed for limit observations, such as zeros. But sometimes, there are situations when there is a significant degree of overlap between the independent variables, leading to a strong linear or non-linear relationship between them, in this case tobit regression model unable provided a good parameters estimation or Or misleading estimators. This is due to the fact that the existence of these strong correlations between the independent variables will have a significant impact on the information matrix(Dawoud, Issam, et al,2022).therefore, tobit regression model with Multicollinearity problem Estimates of coefficients become erratic (large standard errors), also challenge in identifying each independent variable's unique influence, to overcome this problem . We will proposed a new methods Lasso tobit principal component regression, it have

Received: September 16,2024; Revised: September 30,2024; Accepted:October 13,2024; Online Available : October 17, 2024

some objectives, first: it is an effective method of resolving the Tobit regression model's Multicollinearity issue via principal component and lasso methods. Second: it is an effective method for achieving component selection and variables selection by two stages.

The our paper is organized as follows: in section two tobit principal component regression has been shown. In section three, we proposed lasso tobit principal component regression model. In section four ,we study the performance our proposed method via simulation scenarios.We shall display the more concentrated conclusions in section five.

2. TOBIT PRINCIPAL COMPONENT REGRESSION

The kind of data at hand has a significant influence on the regression model that should be chosen, and choosing the best analytical strategy requires carefully weighing the features of the data. The kind of data used for the dependent variable determines which regression model is best. For instance, models like polynomial or multiple linear regression would be more acceptable if the dependent variable was of the continuous kind. However, more specialized models like logistic regression, ordered (logit , probit), or multinomial (logit , probit) would be more appropriate if the dependent variable is discrete or categorical , such as binary, ordinal, or nominal. It is essential to handle censored data appropriately to prevent biased conclusions and produce precise estimations of the model parameters. The exact kind of censoring that is present in the data determines which regression model is best. For the analyzing of censored data at zero point, the tobit model would be more acceptable as shown :

$$y_i = \begin{cases} y_i^* = a + \beta x_i^T + u_i & if\ y_i^* > 0 \\ 0 & if\ y_i^* \le 0 \end{cases} \qquad (i = 1,2,\dots n) \qquad (1)$$

where $u_i \sim N(0, \sigma^2 I)$

y_i is the censored dependent variable has limited information equivalent to zero at the censored point. y_i^* is the latent variable has free information equivalent to quantitative data (Greene, W. (1999)). Finally, Assessing the relationship between a group of independent variables and a censored dependent variable is the focus of the Tobit model. Unfortunately, when there is a strong or complete relationship correlation between two or more of the independent variables, the tobit model becomes Not useful its estimates become misleading and inaccurate due to the problem of Multicollinearity, to overcome this problem tobit principal component regression has been used (Fadel Hamid and Meshal Harbi(2016)) By transforming the correlated independent variables in a Tobit model into principal components, which are a linear combination of the independent variables within each component that are uncorrelated

as following $F = X\varphi$ (F is the matrix components matrix derived from rank (n, q) , φ is corresponds to the eigenvalue in the information matrix and is the orthogonal matrix from the eigenvector. After a series of derivations, the following can be obtained the following $X = F\varphi^T$(Alhusseini, F. H.,2016)

$$y_i = \begin{cases} y_i^* = a + F\varphi^T\beta + u_i & if\ y_i^* > 0 \\ 0 & if\ y_i^* \leq 0 \end{cases} \quad (i = 1,2, \dots n) \qquad (2)$$

when it are φ^T is the orthogonal matrix ,lets assume $\varphi^T\beta = \omega$, Consequently, equation (2) takes on the following form:

$$y_i = \begin{cases} y_i^* = a + F\omega + u_i & if\ y_i^* > 0 \\ 0 & if\ y_i^* \leq 0 \end{cases} \quad (i = 1,2, \dots n) \qquad (3)$$

From the equation above, we will obtain a mathematical model that aims to estimate the effect of the relationship between a censored dependent variable and q from principal component(Tobit principal component regression) (Fadel Hamid and Meshal Harbi(2016)) ,from using minimizing the equation (3) ,we will obtained coefficients estimation of principal components in our model as following :

$$\min_{a,\beta} \sum_{i=1}^{n} (y_i - max\{0, a + F^T\omega\}) \qquad (4)$$

From the equation above, we can estimate the parameters of the principal components for the Tobit regression model of the principal components, and then use the feedback to obtain the regression of the independent variables on the censored dependent variable.

3. LASSO TOBIT PRINCIPAL COMPONENT REGRESSION MODEL

Mixing between Lasso and tobit principal component regression means applying a regularization technique in two stages firstly important components selection and secondly stage variables selection important that can help with issues like Multicollinearity . As following form

$$\min_{,\beta} \sum_{i=1}^{n} (y_i - max\{0, F^T\omega\}) + \lambda \parallel \beta_j \parallel \qquad (5)$$

where , $\lambda \parallel \beta_\theta \parallel$ is penalty lasso and λ, ($(\lambda \geq 0)$ is the tuning parameter**,** we will use especial function to achieving parameters estimation for principal component . It is highly likely that some of the estimated coefficients in a principal component regression model with Lasso regularization will be exactly equal to zero. Lasso automatically identifies the most

significant principal components and eliminates unnecessary ones from the model by setting coefficients to zero. This is a key feature of Lasso that enables effective principal component selection and leads to more interpretable, parsimonious of principal component regression models. When it comes to determining the most crucial elements, the approach mentioned above is thought to be novel compared to earlier methods. Through the orthogonal property of the eigenvalue matrix and the principal components, where each principal component is a linear combination of the original variables, we will obtained on coefficients estimation for original independent variable .

$$\varphi^T\beta = \omega$$

Where φ is orthogonal matrix($\varphi^T\varphi = \varphi\varphi^T = I$) , therefore the coefficients estimation for original independent variable

$$min_{\beta} \sum_{i=1}^{n} (y_i - max\{0, X^T\beta\}) + \lambda \parallel \beta_j \parallel \qquad (6)$$

where $\beta = \varphi\hat{\omega}$

The equation (6) is consider secondly stage for coefficients estimation and variables selection in tobit principal component regression,

The above equation proceeds to estimate each unknown parameter , via proposed a new algorithm that it is run 12000 iterations, first 2000 is excluded as burn in .

4. SIMULATION APPROACH

In this section , Simulation studies are used in this paper to assess the effectiveness of the proposed our method lasso tobit principal component regression (L.T.P.C.R.Model) . To compare with our method (L.T.P.C.R.Model), we will employ two additional methods. Firstly, tobit regression model (T.R.Model) as reported in Tobin, James (1958),secondly, topit principal component regression (T.P.C.R .Model) as reported in Alhusseini, F. H. H, Odah ,M.H(2016) . All method under investigation is assessed using the median of mean absolute deviations, often known as (MMAD) and standard deviation often known as (SD). Two simulation scenarios were employed in this study .

First simulation example (sparse case)

In this simulation example, We demonstrate how well the proposed method works with sparse models. Specifically, we take into account the following true model:

$$y_i = \max(0, y_i^*)$$

$$y_i^* = a + \beta x_i^T + u_i \qquad where\ i = 1,2,\dots.n$$

Seven independent variables were generated using a multivariate normal with mean zero and$Cov = (x_i, x_j) = (0.5)^{i+j}$. Consequently, the accurate predictor coefficients, which include the intercept term are $a = 0.5, \beta = (1,0,3,0,0,0,1)$. In this study, we will used five sample size (25,75,125,175, and 225)respectively. Also, with this simulation scenario, Six different pattern of correlation coefficients between independent variables will be employed.($\rho = 0.40$,$\rho = 0.50$,$\rho = 0.60$,$\rho = 0.70$,$\rho = 0.80$ and $\rho = 0.90$).

Table (1) the MMAD and SD for First simulation example

	N=25					
	$\rho = 0.40$	$\rho = 0.50$	$\rho = 0.60$	$\rho = 0.70$	$\rho = 0.80$	$\rho = 0.90$
Comparison Methods	MMAD	MMAD	MMAD	MMAD	MMAD	MMAD
T.R.Model	0.977 (0.705)	0.821 (0.478)	0.768(0.422)	0.741(0.211)	0.921(0.467)	0.945(0.615)
T.P.R.Model	0.855 (0.623)	0.722 (0.371)	0.624(0.683)	0.662(0.347)	0.641(0.442)	0.624(0.475)
L.T.P.C.R.Model	0.764 (0.560)	0.555 (0.354)	0.614 (0.562)	0.432 (0.067)	0.412 (0.394)	0.455 (0.163)
N=75						
T.R.Model	0.845 (0.396)	0.894 (0.382)	0.827(0.289)	0.845(0.490)	0.942(0.623)	0.745(0.205)
T.P.R.Model	0.734 (0.471)	0.639 (0.330)	0.582(0.195)	0.742(0.363)	0.571(0.545)	0.424(0.105)
L.T.P.C.R.Model	0.634 (0.505)	0.582 (0.330)	0.452(0.087)	0.364 (0.647)	0.391 (0.560)	0.255 (0.064)
N=125						
T.R.Model	0.712(0.467)	0.761 (0.361)	0.934(0.771)	0.883 (0.426)	0.949 (0.721)	1.059 (0.911)
T.P.R.Model	0.612 (0.399)	0.623 (0.474)	0.691 (0.383)	0.496 (0.326)	0.465 (0.516)	0.518 (0. 216)
L.T.P.C.R.Model	0.407 (0.250)	0.463 (0.230)	0.462 (0.271)	0.428 (0.154)	0.365 (0.517)	0.421 (0.194)
N=175						
T.R.Model	0.627 (0.260)	0.791 (0.573)	0.798 (0. 562)	0.696 (0.423)	0.564 (0.464)	0. 708 (0.483)
T.P.R.Model	0.525 (0.304)	0. 770 (0.402)	0.638 (0.266)	0.448 (0.412)	0.346 (0.406)	0.463 (0.239)
L.T.P.C.R.Model	0.337 (0.058)	0.406 (0.515)	0.439 (0. 296)	0.425 (0.451)	0.318 (0.325)	0.421 (0.282)
N=225						

T.R.Model	0.606 (0. 372)	0.525 (0.259)	0.487 (0.207)	0.486 (0.299)	0.424 (0. 268)	0.403 (0.238)
T.P.R.Model	0.498 (0.277)	0.456 (0.373)	0.429 (0.253)	0.416 (0.275)	0.420 (0.260)	0.302 (0.107)
L.T.P.C.R.Model	0.429 (0.231)	0.407 (0.219)	0.319 (0.190)	0.305 (0.118)	0.237 (0.035)	0.328 (0.009)

Note: In the parentheses are SD

Based on the results presented in the above table, We find that our proposed method performs much better than the two comparative methods in terms of parameter estimation and variable selection in the Multicollinearity-prone Tobit model. due to the fact that our proposed method's computed (MMAD) values are substantially smaller than the other two ways' computed (MMAD) values. Additionally, we note that these outcomes are consistent with the various sample sizes and correlation coefficients that were employed in the simulation experiment. The below figure show the coefficients estimation and variable selection of our proposed method (Lasso tobit principal component regression) via five sample size (25,75,125,175, and 225) with correlation ($\rho = 0.60$).

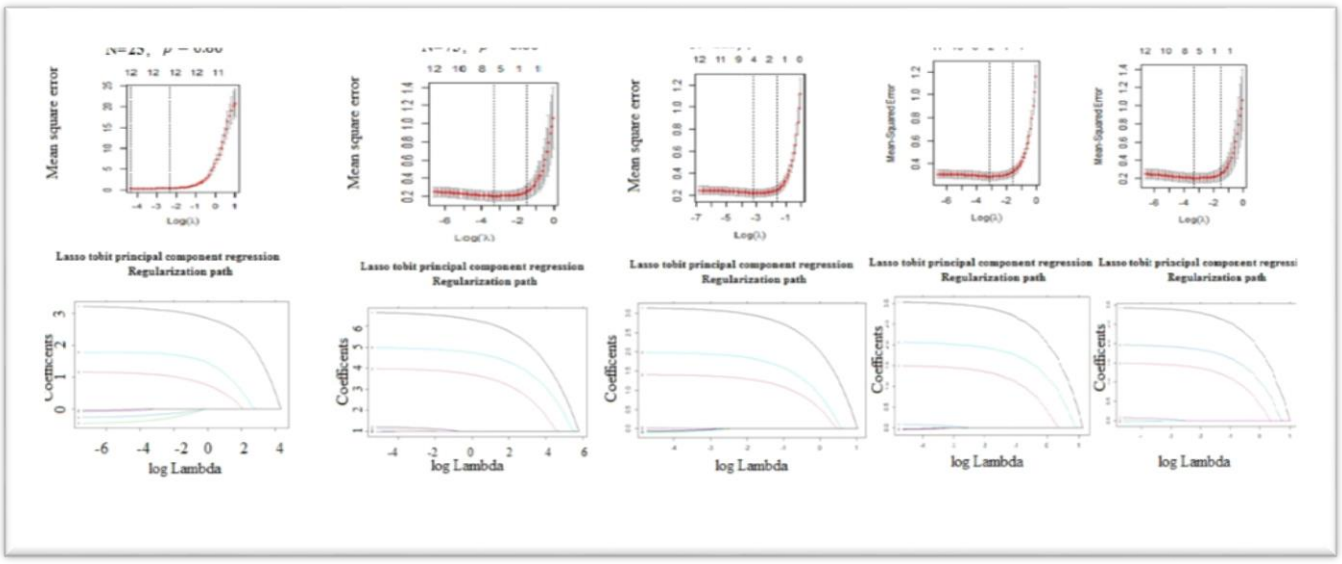

Figure -1- show Cross Validation and coefficients estimation of our proposed method

Second Simulation Example (Very Sparse Case)

In this simulation example, We demonstrate how well the proposed method works with very sparse models. Specifically, we take into account the following true model:

$$y_i = \max(0, y_i^*)$$

$$y_i^* = a + \beta x_i^T + u_i \qquad where\ i = 1,2, \ldots . n$$

Seven independent variables were generated using a multivariate normal with mean zero and$Cov = (x_i, x_j) = (0.5)^{i+j}$. Consequently, the accurate predictor coefficients, which include the intercept term are $a = 0.5, \beta = (1,0,0,0,0,0,0)$. In this study, we will used five sample size (25,75,125,175, and 225)respectively. Also, with this simulation scenario, Six different pattern of correlation coefficients between independent variables will be employed.($\rho = 0.40$,$\rho = 0.50$,$\rho = 0.60$,$\rho = 0.70$,$\rho = 0.80$ and $\rho = 0.90$).

Table (2) the MMAD and SD for second simulation example

	N=25					
	$\rho = 0.40$	$\rho = 0.50$	$\rho = 0.60$	$\rho = 0.70$	$\rho = 0.80$	$\rho = 0.90$
Comparison Methods	MMAD	MMAD	MMAD	MMAD	MMAD	MMAD
T.R.Model	0.819 (0.693)	0.877 (0.605)	0.841 (0.638)	0.852 (0.679)	0.803(0.598)	0.779 (0.572)
T.P.R.Model	0.807 (0.694)	0.759 (0.523)	0.715 (0.573)	0.699 (0.409)	0.514(0.352)	0.425(0.283)
L.T.P.C.R.Model	0.710 (0.544)	0.514 (0.360)	0.517 (0.377)	0.543 (0.385)	0.472(0.202)	0.414(0.256)
N=75						
T.R.Model	0.809 (0.667)	0.721 (0.578)	0.728 (0.540)	0.765 (0.516)	0.7691(0.562)	0.794 (0.506)
T.P.R.Model	0.760 (0.567)	0.622 (0.471)	0.686 (0.441)	0.565 (0.317)	0.472(0.281)	0.428 (0.249)
L.T.P.C.R.Model	0.646 (0.402)	0.555 (0.384)	0.569 (0.347)	0.416 (0.299)	0.392(0.221)	0.330 (0.144)
N=125						
T.R.Model	0.859 (0.720)	0.674 (0.472)	0.543(0.266)	0.692 (0.465)	0.694 (0.406)	0.689 (0.419)
T.P.R.Model	0.764 (0.750)	0.653 (0.475)	0.539 (0.396)	0.513(0.368)	0.508 (0.349)	0.456 (0.215)
L.T.P.C.R.Model	0.447 (0.760)	0.552 (0.390)	0.336 (0.126)	0.472(0.223)	0.390 (0.144)	0.331 (0.153)
N=175						
T.R.Model	0.629 (0.467)	0.627 (0.422)	0.683 (0.438)	0.635(0.493)	1.886(0.534)	1.892(0.497)
T.P.R.Model	0.562 (0.394)	0.639 (0.426)	0.402 (0.227)	0.507(0.334)	1.673(0.561)	1.563(0.673)
L.T.P.C.R.Model	0.559 (0.348)	0.426 (0.276)	0. 338 (0.113)	1.935(0.493)	1.886(0.534)	1.892(0.497)
N=225						
T.R.Model	0.565 (0.399)	0.577 (0.322)	0.448 (0.336)	0.471 (0.218)	0.539 (0.371)	0.527 (0.346)

T.P.R.Model	0.432 (0.310)	0.420 (0.215)	0.418 (0.319)	0.429 (0.241)	0.381 (0.197)	0.374 (0.205)
L.T.P.C.R.Model	0.239 (0.374)	0.366 (0.120)	0.369 (0.141)	0.351 (0.183)	0.223 (0.095)	0.218 (0.093)

Note: In the parentheses are SD

Based on the results presented in the above table, We find that our proposed method performs much better than the two comparative methods in terms of parameter estimation and variable selection in the Multicollinearity-prone Tobit model. due to the fact that our proposed method's computed (MMAD) values are substantially smaller than the other two ways' computed (MMAD) values. Additionally, we note that these outcomes are consistent with the various sample sizes and correlation coefficients that were employed in the simulation experiment. The below figure show the factor (MPSRF) to our proposed method (Lasso tobit principal component regression) via one sample size (125 with one correlation case ($\rho = 0.60$). As demonstrated in Figure 1, the factor MPSRF stabilizes and approaches 1 after around 1000 iterations . Because of this, our proposed method (Lasso tobit principal component regression) has a very quick convergence and good stationary .

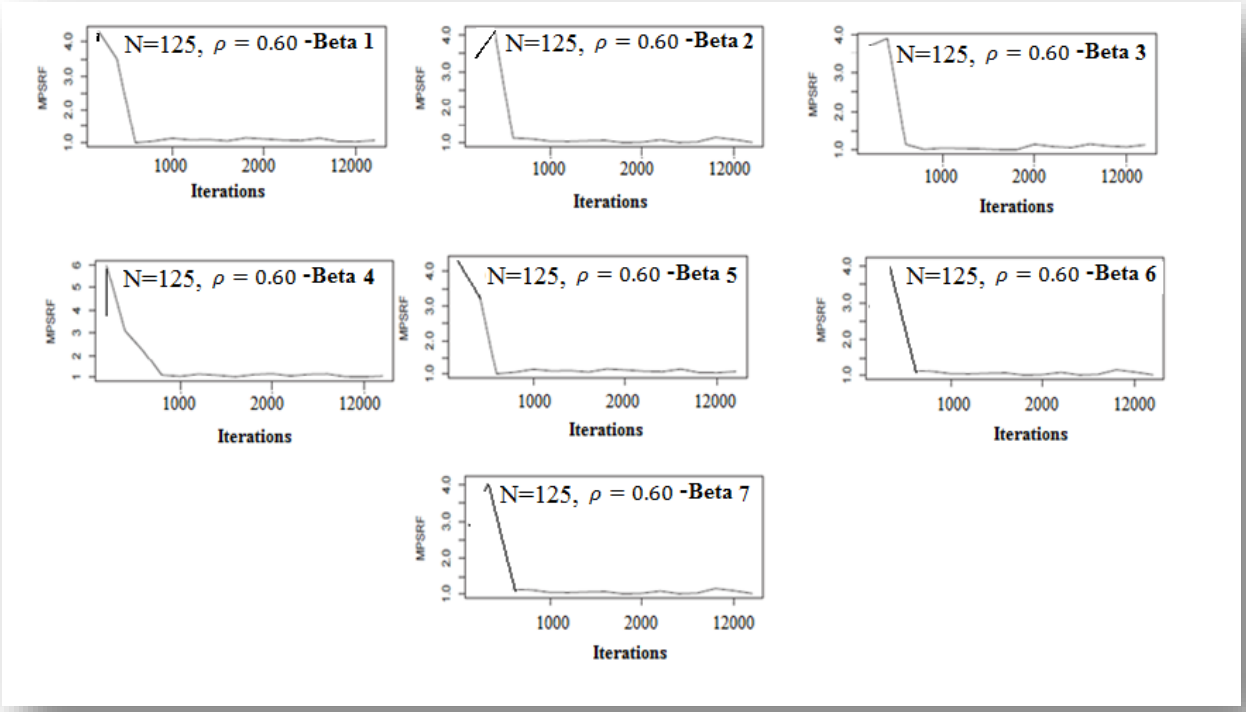

Figure -2- Show factor ((MPSRF) for coefficients estimation in our proposed method with (N=125, $\rho = 0.60$)

5- Brief Conclusions and recommendation

Our proposed method is considered a distinctive approach in achieving the technique of variable selection. On the one hand, it is a robust method against the problem of Multicollinearity. In the variable selection process, it can reduce the important principal components and zero out the less important principal components. It is an unbiased and accurate method in selecting the effective principal components. As for the second stage, it can select the effective variables and exclude the ineffective variables in building the tobit principal component regression model. As a result, we advise using our proposed method in cases when the tobit regression model's independent variables exhibit strong correlation . To ensure that the selection of important principal components and important original variables happens simultaneously and in an automated manner. As a result, the current study can be expanded to include other regularization methods such as the Adaptive Lasso method or the Elastic Net method, with the Tobit principal component regression model and others. Also, The current study can be further extended by estimating and selecting variables in the tobit principal regression model using a Bayesian approach, etc.

REFERENCES

Alhusseini, F. H. H. (2016). Bayesian Tobit principal component regression with application. *American Review of Mathematics and Statistics*, *4*(2), 63-73.

Alhusseini, Fadel Hamid Hadi, and Meshal Harbi Odah. "Principal component regression for tobit model and purchases of gold." *Proceedings of the international management conference*. Vol. 10. No. 1. 2016.

Amemiya, T. (1984). Tobit models: A survey. Journal of Econometrics, 24(1-2), 3-61.

Amore, M. D., & Murtinu, S. (2021). Tobit models in strategy research: Critical issues and applications. *Global Strategy Journal*, *11*(3), 331-355.

Austin, P. C., Escobar, M., & Kopec, J. A. (2000). The use of the Tobit model for analyzing measures of health status. *Quality of Life Research*, *9*, 901-910.

Dawoud, Issam, et al. "A new Tobit Ridge-type estimator of the censored regression model with multicollinearity problem." *Frontiers in Applied Mathematics and Statistics* 8 (2022): 952142.

Duan, N., Manning, W. G., Morris, C. N., & Newhouse, J. P. (1983). A comparison of alternative models for the demand for medical care. Journal of Business & Economic Statistics, 1(2), 115-126.

Greene, W. (1999). Marginal effects in the censored regression model.*Economics Letters*, *64*(1), 43-49.

Maddala, G. S. (1983). Limited-dependent and qualitative variables in econometrics. Cambridge University Press.

Tobin, James (1958). "Estimation of Relationships for Limited Dependent Variables". Econometrica. 26 (1): 24–36. doi:10.2307/1907382.