PP: 82-96 *The Scientific Journal of Cihan University – Sulaimanyia*
Volume (2), Issue (2), December 2018
ISSN 2520-7377 (Online), ISSN 2520-5102 (Print)

Limited Dependent Variable Modelling (Truncated and censored Regression models) with Application

Asst.Prof: Dr. Nawzad. M. Ahmed
Statistics & Informatics Department - College of Administration & Economy - University of Sulaimani-Iraq
nawzadstat@yahoo.com

Abstract

this study is about fitting a such regression model for Tobit (**Truncated), and (censored)** data regression models for a sample of persons (n=500), with their sugar rate per person as a response variable (Y), and explanatory variables (X_1: Age , X_2: Rate of cholesterol.gm/dl, and X_3: Triglyceride.gm/dl). (Data of this study were obtained through the follow-up of diabetic patients directly from the laboratories of analysis of blood diseases and diabetes in several laboratories with experience in the city of Sulaymaniyah) . The persons whom their sugar rate is above 120gm/dl only remained and the others are censored as a zero values (left censored), also the same set of data was truncated from below, then in truncated sample, only the cases under risk of diabetes (greater than the sugar rate=120gm/dl) are recorded, and the others are omitted from the data set. With remaining explanatory with their actual records, and then after fitting censored and truncated models also in order to detect the unbiased, and consistency of the estimated (Parameters) model, the (*marginal effects for explanatory*) are calculated and comparing these two models with multiple regression model estimation by an ordinary least square method (OLS).

Keywords:
Censored Regression Model (CRM), Truncated Regression Model (TRM), Maximum Likelihood Estimator (MLE), Marginal Effect (ME), Ordinary Least Square (OLS),

الملخص

هذه الدراسة تتعلق بنماذج (توبت) وخصوصا أنموذجي أنحدار (المراقبة و المبتور)والمتعلق بالمتغيرات الستجابة المحدودة, ونمذجة البينات لعينة البحث بهذين الأنموذجين . عينة البحث متكون من (n=٥٠٠) شخص من ثم تسجيل معدل السكر في الدم كمتغير استجابة(Yمعتمد), وثلاث متغيرات تفسيرية وهم (X1 : العمر ومقاسة بالسنوات, X2: معدل الكولسترول في الدم ومقاسة بالملغم / ديساليتر, و X3: معدل الدهون الثلاثية ومقاسة أيضا بالملغم / ديساليتر). أن الأشخاص الذين هم يتجاوز لديهم معدل السكر في الدم عن ١٢٠ ملغم هم هدف النمذجة لذلك تسجل قيم استجاباتهم (الأستجابة المحدوة) وتصفرالمتبقي من العينة من قيم الاستجابة في الأنموذج المراقبة من اليسار في حين مسح هذه القيم (دون ١٢٠ ملغم) عند نمذجة البيانات بالأنحدار المبتورمن الأسفل.

أذن بشكل عام عملية النمذجة (مراقبة والمبتور) في هذه الدراسة هما للاشخاص الذين هم تحت مخاطرة الاصابة بمرض السكر. وبعد أيجاد النماذج للبيانات تحت الدراسة و من ثم أختبار معالمهم وأختيار أفضل نموذج بينهما و لمعرفة التأثيرات الجزئية للمتغيرات التفسيرية سابقة الذكر على معدلات السكر للاشخاص ,تم أستخدام (التاثيرات الجزئية) لتقديرات معالم الأنموذج الأفضل.علما بأن تقديرات النماذج المقدرة هم (تقديرات الأمكان الأعظم المشروط), و هي تقديرات غير متحيزة و متسقة في ان واحد للبينات المراقبة والمبتورة ,مقارنة بتقديرات لأنموذج الأنحدار المتعدد بطريقة المربعات الصغرى للبيانات ذاتها والتي تكون عادة متحيزة و غير متسقة وبالتالي عدم أمكانية الاعتماد عليها في تفسير النموذج بسبب فقدانها لخاصية العمومية للنماذج.

DOI: http://dx.doi.org/10.25098/2.2.32

پوختە

ئەم توێژینەوەیە تایبەتە بە مودیلەکانی(توبت Tobit)، بەتایبەتی هەردوو مودیلی (سێنسەرد، بڕدراو (censored, and truncated)کە بەکاردەهینریت بو بەمودیلکردنی پەرچە کۆڕاوی سنوردار(limited dependent variable) . سامپڵی توێژینەوە بریتیە لە (٥٠٠کەس) کە ڕێژەی شەکر لە خوینیان تۆمارکراوە وەك پەرچە گۆڕاو (dependent variable) لەگەڵ سێ گۆڕاوی تەفسیری (explanatory) کە بریتین لە (X1: تەمەن بە پێوەری ساڵ، X2: ڕێژەی کولسترۆڵ- ملغم/دیسالتر , X3:جەوری سیانی –ملغم /دیسالتر). ئەو کەسانەی ڕێژەی شەکر لە خوێنیاندا زیاترە لە ١٢٠ ملغم/دیسالتر بە ئامانج کراون بۆ بەمۆدێل کردن، لەبەر ئەوە تومارکردنی دیاری کراویان بۆ کراوە لە سەروو ١٢٠ ملغم وەهاوکات بە(سفر)کردنی ڕێژەی شەکر بۆ ئەکەسانەی خوار ئەو ڕێژەیەن بۆ مۆدێل کردنیان بە جۆری (مۆدێلی لێژبونەوەی چاودێری کراو-لای چەپ)، هەروەها جارێکی تر سڕینەوەی تۆمارەکانی ئەم گۆڕاوە dependent بۆ کەسانی خاوەن ڕێژەی شەکری خوار ١٢٠ ملغم بۆ بەمۆدێلکردنیان بە جۆری (مۆدێلی بڕدراو-لای خوارەوە).

بەشێوەیەکی گشتی پرۆسەی بەمۆدێلکردن (چاودیری-بڕدراو)تەنها ئەوکەسانە دەگرێتەوە کە مەترسی توشبونیان لەسەرە بە نەخۆشی شەکرە . دوای دەرهێنانی مۆدێلەکان و تاقیکردنەوەی پارامیتەرەکانیان و هەڵبژاردنی باشترین مۆدێل و بۆ زانینی کاریگەری تەنیا یان سەربەست (marginal effects) بوگۆڕاوەکانی کارتێکەر(X1, X2, X3) هەریەکە بە تەنیا بۆسەر ڕێژەی شەکر لە خوێندا، پێوەری (marginal effects) بەکارهێنراوە بۆ باشترین مۆدێل دواى هەڵبژاردنی . نرخی خەمڵێنراوی پارەمیتەرەکانی مۆدێلەکانی (سێنسەردو بڕدراو) بە ڕێگای لایکلیهودی مەرجدار(conditional likelihood (estimation method)کە نرخەکانیان بێ لایەن و لەهەمان کاتدا کونسستن (unbiased &consistent) بە بەراورد(لەگەڵ نرخی خەمڵێنراوی پارەمیتەرەکانی مۆدێلی هەمەلێژبونەوە (multiple regression model)بۆ هەمان داتای توێژینەوەکەمان کە بە ڕێگای بچوکترین دوجای ئاسایی (ordinary least square: OLS) کە خەسڵتی (بێ لایەنی و کونسست) وون دەکات وسیفەت و توانای گەشتاندنی(generalization property) تێدا نامێنێت و ناتوانێت بە باشی لێکدانەوەو شیکردنەوەی داتاکان بە باشی ئە نجام بدات کاتێك سانسۆر لەسەر داتاکان هەبێت وەك لەم توێژینەوەیەدا ئاماژەی پێکراوە.

Introduction [(1), (5)]*:*

A limited dependent variable is a variable whose range of possible values is 'restricted in some important way'. In advanced micro-econometrics term is often used when estimation of the relationship between the limited dependent variable of interest and other variables requires methods that take this restriction into account. This may arise when the variable of interest is constrained to lie between (0, and 1), as in the case of a probability, or is constrained to be generally positive lie between (0, ∞), as in the case of a required age for a special job is greater or equal than (20 years) as an average. Then a limited dependent variable means there is a limit or boundary on its values and some of the observations hit the limit such as the quantity of product consumed is zero for some consumers, and positive amount for others, or as a capacity issues, the demand for tickets for a game is censored at the hall, or state capacity which is a maximum ticket that can be sold. Limited dependent variable models include: **a**-Censoring, where for some individuals in a data set, some data are missing but other data are present, **b**-Truncation, where some individuals are systematically excluded from observation. **c**- Discrete outcomes, such as binary or qualitative data restricted to a small number of categories.

DOI: http://dx.doi.org/10.25098/2.2.32

1-1 The different between (Censored and Truncated data) [8]:

this difference can be detected, as an example, for a sample consumers of coffee, the censored sample included consumers who consumes "zero" quantity. But truncated sample, are consumers whom choose positive quantities of coffee. (Both types are only described the values of the dependent observations value). It is difficult to modulate these types of data, this because the marginal effects for the factors or what so called (Explanatory) were varied if they compared with each other's in those models of regression that interesting this behaviours with other that can't able to recognize censored or truncated dependent variables (limited dep. variables may defined as(latent variable)).This because the censored data sample includes (zero values) introduced during mean value computing, but truncated one, these (zero values) omitted from the sample, and then the mean value of censored one is less than the truncated for same data set, this because of the sample size(n) which greater for censored than truncated for the same sample[6].

2- Truncated, censored data & regression models:

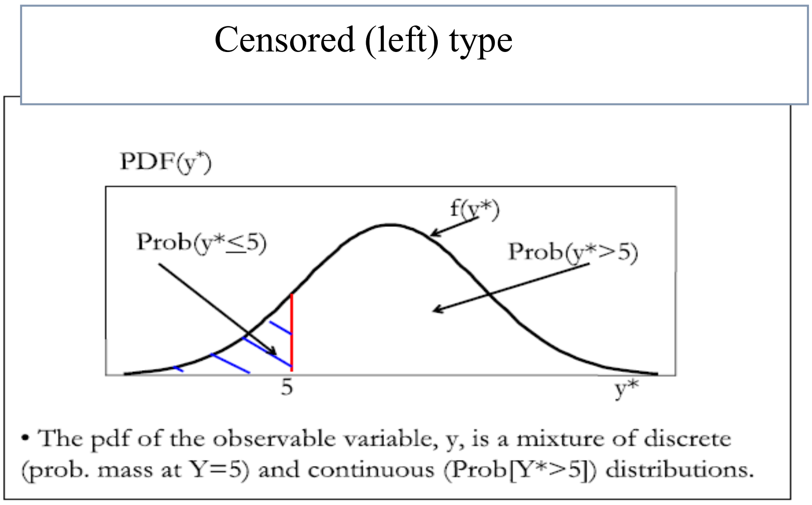
2-1 Truncated, and censored data [7]:

The data described as a censored or truncated are defined logically as follow:

(Y) is ***censored*** when we observe (X's) for all observations, but we only know a true value of (Y) for a restricted range of observations. Values of (Y) in a certain range are reported as a single value or there is significant clustering around a value, say (0).

If Y=k, or Y>k (***k is a threshold***) for all Y, then, (Y^*) is censored from below or (***left-censored***). If Y=k, or Y<k for all Y, then (Y^*) is censored from above or (***right-censored***).

(Y) is ***truncated*** when only observe(X)for observations where(Y) wouldn't be censored, then we don't have a full sample for (Y, X), we exclude observations based on characteristics of (Y). Also, if the excluded observations (Y) are above the threshold (k), then (Y^*) is ***truncated above***, and (Y^*) ***truncated below*** if the excluded observations (Y) are below the threshold (k). And if there are two thresholds (k_1, and k_2) from above, and below then (Y^*) is said to be ***truncated from both*** (***above, and below***). The following graphs explain each case above.

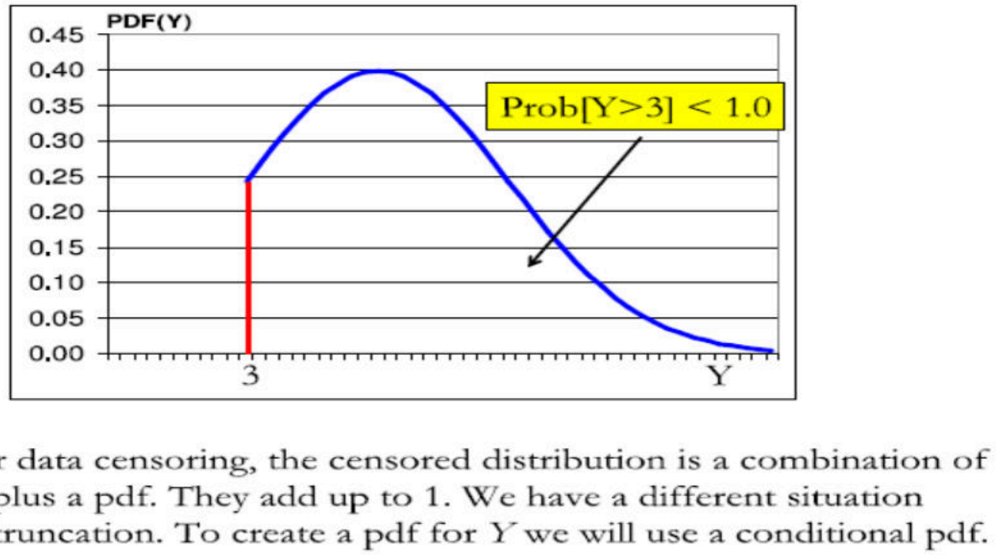


• The pdf of the observable variable, y, is a mixture of discrete (prob. mass at Y=5) and continuous (Prob[Y*>5]) distributions.

Figure (1): Censored from below with the probability distribution explanation (threshold =5) [3]

DOI: http://dx.doi.org/10.25098/2.2.32

Normal Truncated



Under data censoring, the censored distribution is a combination of a pmf plus a pdf. They add up to 1. We have a different situation under truncation. To create a pdf for *Y* we will use a conditional pdf.

Figure (2): Truncated from below with the probability distribution explanation (threshold =3) [(4)]

2-2 Truncated, and censored, regression models *[(4,(9)]:*

Truncated regression is different from censored with the following points: *Censored regressions:* the dependent variable may be censored, but you can introduce the censored observations in the regression. But *truncated regressions:* A subset of observations are dropped, and only the remainder data after truncation are available in the regression. It is very important to say that the (OLS) estimates were (biased) when the sample is determined by the value of (Y-variable) in both truncated, and censored data, this is because of the correlated sample selection with the random variable (ε_i).

Consider the regression model: $y_i = \beta_o + \beta_1 x_i + \varepsilon_i$, for applying truncated reg. and instead of using all observations, we use a sub-sample. Let(a_i) de an indicator, if $\varepsilon_i$then observation (i) included in the regression, and if $a_i = 0$, then it doesn't and dropped from the data. Since in(OLS) all the data including in regression then for all observations $a_i = 1$, the case will return to the above reg. model and then the sample selection are uncorrelated with the error term (ε_i) , so their estimates are unbiased, but if the selection sample at only the obs. with $a_i = 1$ were used, that's we run the following regression $a_i y_i = a_i \beta_o + \beta_1 a_i x_i + a_i \varepsilon_i$,now $(a_i x_i)$ is the independent variable and ($u_i = a_i \varepsilon_i$)is the error term. This means that *(OLS) estimators are unbiased* if $E(u_i = a_i \varepsilon_i / a_i x_i)$ =0, but let now we have a such threshold(k) for truncation, then now (a_i) depends on (y_i), and (ε_i). It cannot be dropped out from the conditional set as we did previously, then $E(\varepsilon / x, y) \neq E(\varepsilon / x) = 0$ means that *(OLS) estimators are biased.*

DOI: http://dx.doi.org/10.25098/2.2.32

2-3 Truncated Regression Model Fitting [4], [7]:

When the data truncation is based on the (Y-variable), suppose we have the following general linear regression satisfied all properties and assumptions of linear model:

$$y_i = x_i'\beta + \varepsilon_i\,, \qquad \varepsilon_i \sim N(0,\sigma^2) \qquad \text{-----------} \ (1)$$

The underline sample is only if ($y_i < k_{i,}$), observation dropped if ($y_i \geq k_i$) , truncated below , and we know the exact value of (k_i) for each individual. Here the truncated model that produces unbiased estimate is based on (Maximum likelihood: ML) Estimation. The difference of truncated regression, and biased regression (OLS), is explained from the figure below:

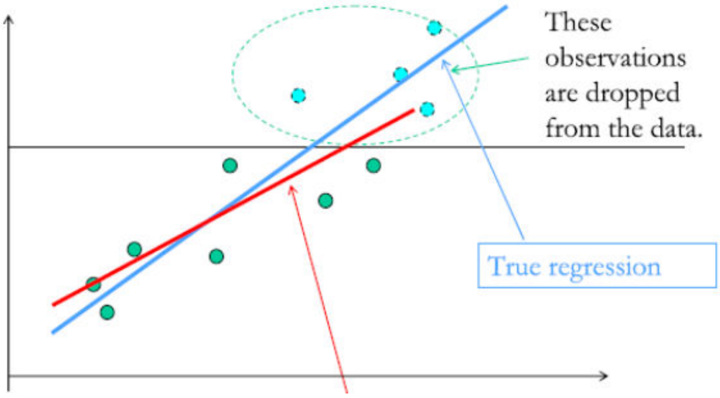


Figure (3): The difference of truncated regression, and biased regression (OLS)[8]:

2-4 Normal Truncated Regression. Conditional Distribution [4], [6]:

Given the normality assumption, $\varepsilon_i \sim N(0,\sigma^2)$, and estimating the truncated model, then the ML method is capable to apply this case, is for each $\varepsilon_i = y_i - x_i'\beta$ the likelihood contribution is the probability density function (p.d.f) for (ε_i), given by f(ε_i), but we remember that we select sample only if $(y_i < c_i)$, then we have to use f(ε_i) conditional to $(y_i < c_i)$: then:

$$f\ (\varepsilon_i / y_i < c) \ = f(\varepsilon_i / \varepsilon_i < c_i - x_i'\beta) = \frac{f(\varepsilon_i)}{pr(\varepsilon_i < c_i - x_i'\beta)}$$

$$= \frac{f(\varepsilon_i)}{pr(\frac{\varepsilon_i}{\sigma} < \frac{c_i - x_i'\beta}{\sigma})} = \frac{f(\varepsilon_i)}{\Phi(\frac{c_i - x_i'\beta}{\sigma})}$$

$$= \frac{1}{\Phi(\frac{c_i - x_i'\beta}{\sigma})} \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{\varepsilon_i^2}{2\sigma^2})$$

$$f(\varepsilon_i \ / \ y_i < k) = \frac{\frac{1}{\sigma}\phi(\frac{\varepsilon_i}{\sigma})}{\Phi(\frac{c_i - x_i'\beta}{\sigma})} \qquad \text{------------} \ (2)$$

$\Phi(.), and\ \phi(.)\ are\ the\ cummulative\ density\ function, and\ (p.d.f)\ for\ normal\ distribution\ respectiv$

DOI: http://dx.doi.org/10.25098/2.2.32

2-5 Truncated Normal Moments: (2),(8)

Let: y^* be a (truncated variable),

$$y^*N(\mu^*,\sigma^2)\ ,and\ \alpha = (c-\mu^*)/\sigma \quad and\ \ \lambda(\alpha) = \frac{\emptyset(\alpha)}{1-\Phi(\alpha)}$$

Then the *first moment* is

$$E[y^*/y > c] = \mu^* + \sigma\lambda(\alpha) \qquad \text{-------------} \ (3)$$

This is a truncated regression. If $\mu^* > 0$ and the truncation is from below (this means $\lambda(\alpha)$>0) then the mean of the truncated variable is greater than the original mean, remember that for the standard normal distribution $\lambda(\alpha)$ is the mean of the truncated distribution. And the second moment is given by:

$$var[y^*/y > c] = \sigma^2[1-\delta(\alpha)] \quad where,\ \ \delta(\alpha) = \lambda(\alpha)[\lambda(\alpha)-\alpha]$$

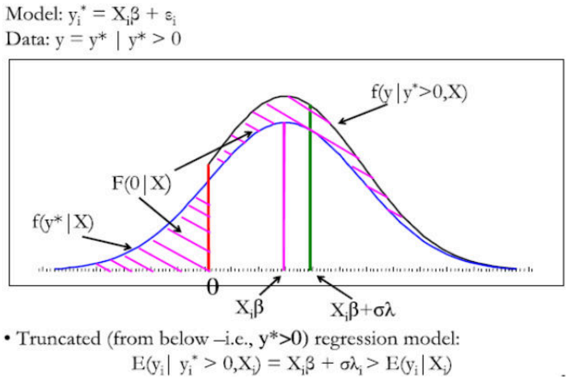
-------------- (4)

$\lambda(\alpha)$, is called (inverse Mill's ratio which defined as a ratio of the probability density function $\emptyset(\alpha)$ to the [1- $\Phi(\alpha)$] Where $\Phi(\alpha)$ is the cumulative normal distribution function. Its use is often motivated by the following property of the truncated normal distribution. If (*X*) is a random variable having a normal distribution with mean μ and variance σ^2, then:

$$E(y/\ y > \alpha) = \mu + \sigma\frac{\emptyset(\frac{\alpha-\mu}{\sigma})}{1-\Phi(\frac{\alpha-\mu}{\sigma})}$$, and also

$$E(y/\ y < \alpha) = \mu + \sigma\frac{-\emptyset(\frac{\alpha-\mu}{\sigma})}{\Phi(\frac{\alpha-\mu}{\sigma})}$$, Where (α) is constant.

A common application of the inverse Mills ratio (sometimes also called "non-selection hazard") arises in regression analysis to take account of a possible selection bias. If a dependent variable is censored (i.e., not for all observations a positive outcome is observed) it causes a concentration of observations at zero values. This problem was first acknowledged by Tobin (1958), who showed that if this is not taken into consideration in the estimation procedure, an ordinary least squares estimation will produce biased parameter estimates. With censored dependent variables there is a violation of the Gauss distribution assumption of zero correlation between independent variables and the error term. This moment gave a general result for the truncated regression that it reduces the variance when it is applied to upper or lower truncation that's because of the range of $\delta(\alpha)$, such that: $0 \leq \delta(\alpha) \leq 1$. The following figure explains clearly the truncated normal model:



Model: $y_i^* = X_i\beta + \varepsilon_i$
Data: $y = y^* \mid y^* > 0$

• Truncated (from below –i.e., y*>0) regression model:
$E(y_i|\ y_i^* > 0,X_i) = X_i\beta + \sigma\lambda_i > E(y_i|X_i)$

Figure (4): Truncated Normal Distribution (9), (10) The likelihood contribution for ith observation is given by:

$$L_i = \frac{\frac{1}{\sigma}\emptyset\left\{\frac{y_i - x_i'\beta}{\sigma}\right\}}{\Phi\left(\frac{x_i'\beta}{\sigma}\right)}$$

The likelihood function is given by

$$logL(\boldsymbol{\beta},\sigma) = \sum_{i=1}^{N} logL_i = -\frac{N}{2}[\log(2\pi) + \log(\sigma^2)]$$

$$-\frac{1}{2\sigma^2}\sum_{i=1}^{N}\varepsilon_i{}^2 - \sum_{i=1}^{N} log\left[\Phi\left(\frac{x_i'\beta}{\sigma}\right)\right] \qquad \text{-------------- (5)}$$

$\left(-\frac{N}{2}[\log(2\pi)+\log(\sigma^2)] - \frac{1}{2\sigma^2}\sum_{i=1}^{N}\varepsilon_i{}^2\right)$, is the (log function) of the joint density of (N) values of (y*: truncated), and $-\sum_{i=1}^{N} log\left[\Phi\left(\frac{x_i'\beta}{\sigma}\right)\right]$ is a (log function for joint probability of(y* > 0). The value of $(\boldsymbol{\beta},\sigma)$ that maximizes (logL) are the ML estimators of the truncated regression.

2-6 The Marginal Effects [(1), (4)]:

The estimated $(\boldsymbol{\beta_k})$ **vector** shows the effect of (x_{ki}) on (y_i). Thus coefficients and marginal effects of a variable is the effect of an unit change of this variable on the probability P(Y = 1|X = x), given that all other variables are constant, The slope parameter of the linear regression model measures directly the marginal effect of the variable on the other variable:

$$\frac{\partial E(y_i/X_i, y_i^* > 0)}{\partial X_{k,i}} = \boldsymbol{\beta_k} + \frac{\partial E(\varepsilon_i \,/\, y_i^* > 0)}{\partial X_{k,i}} \qquad \text{……… (6)}$$

$$\frac{\partial E(y_i/X_i, y_i^* > 0)}{\partial X_{k,i}} = \boldsymbol{\beta_k} + \sigma\frac{\partial\lambda}{\partial X_{k,i}} = \boldsymbol{\beta_k} + \sigma(\lambda_i^2 - \alpha_i\lambda_i)\left(-\frac{\beta_k}{\sigma}\right)$$

$$= \boldsymbol{\beta_k}(1 - \lambda_i^2 - \alpha_i\lambda_i) = \boldsymbol{\beta_k}(1 - \delta_i)$$

Where $\delta_i = \lambda(\alpha_i)[\lambda(\alpha_i) + \alpha_i]$ and $0 < \delta_i < 1$ ……………. (7)

2-6: Censored Regression Model [(5), (6)] :

This type of regression was also is another type of (Tobit) model, which is arises when the(y) variable is limited (or **censored**) from right, or left. If we are estimating sales (y) of a particular good, we would probably observe a large number of observation where (y = 0). The sales variable would be censored at y=0 There are numerous other examples where the dependent variable in a regression equation is similarly limited.

The censored (Tobit) Model structure:

$y_i^* = \beta_1 + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + u_i$ (Standard Tobit model)

--------- (8)

DOI: http://dx.doi.org/10.25098/2.2.32

y_i^* : is unobservable but $y_i = \begin{Bmatrix} 0 \ \ if \ \ y_i^* < 0 \\ y_i^* \ if \ y_i^* \geq 0 \end{Bmatrix}$

Censored Regression model Case where all (y_i, x_i) are observed it is just that when y_i "passes" the truncation point, y_i is recorded as the truncation point. As in the truncation model you can have left-censoring, right-censoring, or upper and lower-censoring. The Tobit reg. model can be represented also as follow [(2)]:

Censored regression model is given by: $y_i^* = X_i'\beta + \varepsilon_i$

$$y_i = 0 \qquad \text{if} \qquad y_i^* = X_i'\beta + \varepsilon_i \leq 0$$

$$y_i = X_i'\beta + \varepsilon_i \ \text{if} \qquad y_i^* = X_i'\beta + \varepsilon_i > 0 \qquad \text{-----------} \ (9)$$

$E(y_i \mid y_i > 0, X_i)$: Is given above in the truncated regression model discussion (where X_i is a scalar explanatory variable). The unconditional expectation of y is:

$$E(y_i \mid X) = (1 - \Phi(-X_i\beta/\sigma))X_i\beta + \sigma\phi(-X_i\beta/\sigma) \qquad \text{-----------} \ (10)$$

The total effect of a change in the k-th explanatory variable $x_{k,i}$ on the expectation of y_i is $\dfrac{\partial E[y_i \mid X_i]}{\partial x_{k,i}} = (1 - \Phi(-X_i\beta/\sigma))\beta_k$ ---------- (11)

In the standard Tobit model (Tobin 1958), we have a dependent variable y that is left-censored at zero:

$$y_i^* = x_i'\beta + \varepsilon_i \ , \ y_i = \begin{Bmatrix} 0 & if \ y_i^* \leq 0 \\ y_i^* & if \ y_i^* > 0 \end{Bmatrix} \qquad \text{----------} \ (12)$$

Here the subscript (i = 1, …N) indicates the observation, y_i^* is an unobserved Variable,(x$_i$) is a vector of explanatory variables, β is a vector of unknown parameters, and $(\varepsilon_i)$is a disturbance term. The censored regression model is a generalisation of the standard (Tobit) model. The dependent variable can be either left-censored, right-censored, or both left-censored and right-censored, where the Lower (and/or) upper limit of the dependent variable can be any number:

$$y_i^* = x_i'\beta + \varepsilon_i$$

$$y_i = \begin{Bmatrix} a & if \ y_i^* \leq a \\ y_i^* & if \ \ a < y_i^* < b \\ b & if \ y_i^* > b \end{Bmatrix} \qquad \text{-----------} \ (13)$$

Here (a) is the lower limit and (b) is the upper limit of the dependent variable.

DOI: http://dx.doi.org/10.25098/2.2.32

2-6-1 Estimation Method [(4),(5),(9)]:

Censored regression models (including the standard Tobit model) are usually estimated by the Maximum Likelihood (ML) method. Assuming that the disturbance term (ε) follows a normal distribution with mean (0) and variance (σ^2) , the log-likelihood function is:

$$logL = \sum_{i=1}^{N}\begin{bmatrix} I_i^a log\Phi\left(\frac{a-x_i'\beta}{\sigma}\right) + I_i^b log\Phi\left(\frac{x_i'\beta-b}{\sigma}\right) + (1-I_i^a-I_i^b) \\ (log\phi\left(\frac{y_i-x_i'\beta}{\sigma}\right) - \log\sigma) \end{bmatrix} \qquad \text{---------} \ (14)$$

Where φ(.), and Φ(.) denote the probability density function and the cumulative distribution function, respectively, of the standard normal distribution, and I_i^a and I_i^b are indicator functions with:

$$I_i^a = \begin{Bmatrix} 1 \ \ if \ \ y_i = a \\ 0 \ \ if \ \ y_i > a \end{Bmatrix} \quad , \quad \text{and} \quad I_i^a = \begin{Bmatrix} 1 \ \ if \ \ y_i = b \\ 0 \ \ if \ \ y_i < b \end{Bmatrix}$$

Note that the standard Tobit model presented above can be written as a combination of two familiar models. The first model is called (Probit) model, which determines whether the (y_i) variable is (zero or positive) and the second model is a Truncated Regression model for the positive values of (y_i). The (Tobit I) model assumes that the parameters for the effect of the explanatory variables on the probability that an observation is censored and the effect on the conditional mean of the non-censored observations are the same.

3-Application:

Table (1): A sample taken from the data as an example (100 obs.) out of total cases (n) = 500, left censored, (Y*=Y > 120).

ID	X2: Age	X2 :CHOLESTROL(mg/dl)	X3:TRIGLYCERID(mg/dl)	Y = Sugar(mg/dl)
1	39	232	222	0
2	57	226	169	0
3	40	168	91	0
4	54	213	453	0
5	65	230	134	0
6	46	152	143	0
7	50	207	257	0
8	45	238	202	0
9	72	101	96	0
10	48	247	137	0
.	.	.	.	.
.	.	.	.	.

DOI: http://dx.doi.org/10.25098/2.2.32

57	50	303	299	141
58	65	206	203	146
59	50	179	149	147
60	60	152	157	147
61	45	195	249	149
62	64	280	125	156
63	65	165	265	164
.	.	.	.	.
.	.	.	.	.
98	49	235	134	123
99	57	202	177	206
100	34	352	381	164

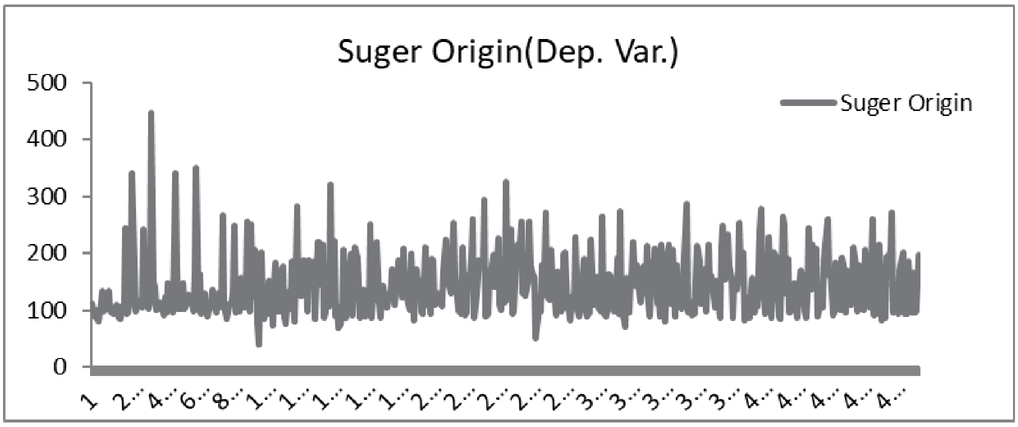


Figure (5): Original Sugar rate (Dependent Variable: Yi)

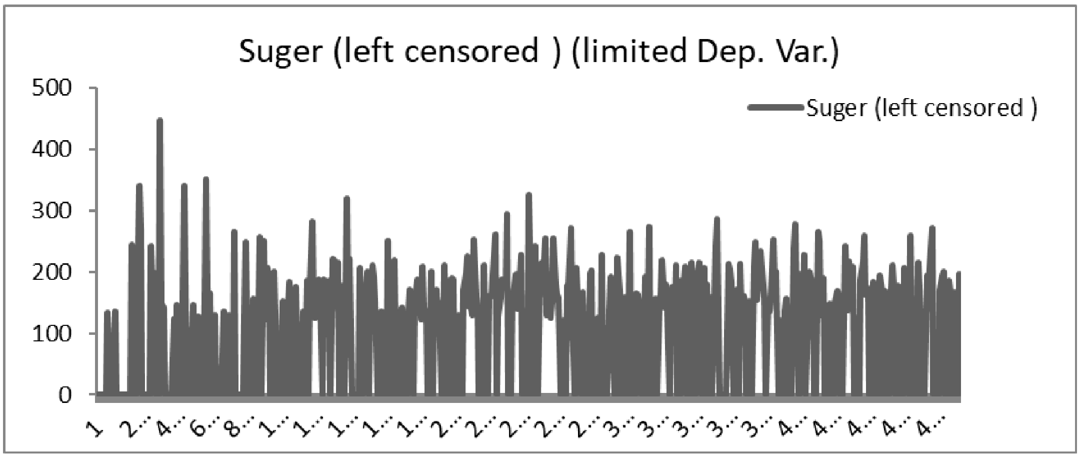


Figure (6): Sugar rate (left censored: Y^*: limited Dependent Variable: Y^*>120)

DOI: http://dx.doi.org/10.25098/2.2.32

3-1 Fitting (censored, and truncated) Regression Models Results:

Using the statistical package (R 3.0.3 for Statistical Computing).

3-1-1: Censored Regression

(formula = y ~ x, left = 0, right = Infinity) Observations: Total (500 observations, Left-censored= 221 observations, Uncensored =279 observations, left censored (Y < 120 then Y* = 0: observation). The results of censored regression are as follow:

Table (2): Results of left censored regression model:

Coefficients:	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.2053853	0.3020629	-3.991	6.59e-05 ***
Age	0.0257398	0.0046099	5.584	2.36e-08 ***
Cholesterol	-0.0014870	0.0011273	-1.319	0.187
Triglycerides	0.0028108	0.0005348	5.256	1.47e-07 ***

Mean square error= 624.94 on (496) degrees of freedom Wald-statistic: 68.03 on 3 Df, p-value:1.1294e-14, No. of iterations=4, AIC=632.94 ,$AIC = \{-2(\text{log.likelihood}) + 2K\}$, Where: K :the number of model parameters (the number of variables in the model plus the intercept). Log-likelihood is a measure of model fit, the higher the number, the better the fit. And minimum AIC is the score for the best model.

3-1-2: Marginal Effects (ME) for censored regression:

Table (3): results of Marginal effects for explanatory variables:

	Marg. Eff.	Std. Error	t value	Pr(>|t|)
Age	2.142491	0.352280	6.0818	2.379e-09 ***
Cholesterol	-0.124643	0.088041	-1.4157	0.1575
Triglycerides	0.220005	0.037824	5.8166	1.079e-08 ***

3-1-3: Truncated (below) regression model results:

Table (4): Results for truncated regression model:

Coefficients:	Estimate	Std. Error	t-value	Pr(>|t|)
(Intercept)	-422.78466	102.56397	-4.1222	3.753e-05 ***
Age	6.61668	1.31048	5.0491	4.440e-07 ***
Cholesterol	-0.75165	0.33338	-2.2546	0.02416
Triglycerides	0.74587	0.14152	5.2705	1.360e-07 ***

DOI: http://dx.doi.org/10.25098/2.2.32

Log-Likelihood: -2774.9 on 5 Df , , LR = -2256.359 with df = 4. AIC=5549.8

3-1-4: Multiple Regression Model (OLS):

Table (5): Results for multiple regression model with (OLS) method for origin sugar rate:

Model	Unstandardized Coefficients		Standardized Coefficients	T	Sig.	Collinearity Statistics	
	B	Std. Error	Betas			Tolerance	VIF
(Constant)	73.260	13.568		5.399	.000		
Age	1.254	.202	.259	6.201	.000	.996	1.004
Cholestrol	-.091	.052	-.076	-1.762	.079	.942	1.061
TRI	.144	.023	.275	6.414	.000	.945	1.058

Table (6): Multiple Regression Model Summary, and Analysis of Variance:

Model Summary

odel	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.373[a]	.139	.134	59.657	1.915

ANOVA

Model	Sum of Squares	Df	Mean Square	F	Sig.
Regression	285256.938	3	95085.646	26.717	.000[a]
Residual	1765261.214	496	3558.994		
Total	2050518.152	499			

DOI: http://dx.doi.org/10.25098/2.2.32

Log-Likelihood: -2774.9 on 5 Df , , LR = -2256.359 with df = 4. AIC=5549.8

3-1-4: Multiple Regression Model (OLS):

Table (5): Results for multiple regression model with (OLS) method for origin sugar rate:

Model	Unstandardized Coefficients		Standardized Coefficients	T	Sig.	Collinearity Statistics	
	B	Std. Error	Betas			Tolerance	VIF
(Constant)	73.260	13.568		5.399	.000		
Age	1.254	.202	.259	6.201	.000	.996	1.004
Cholestrol	-.091	.052	-.076	-1.762	.079	.942	1.061
TRI	.144	.023	.275	6.414	.000	.945	1.058

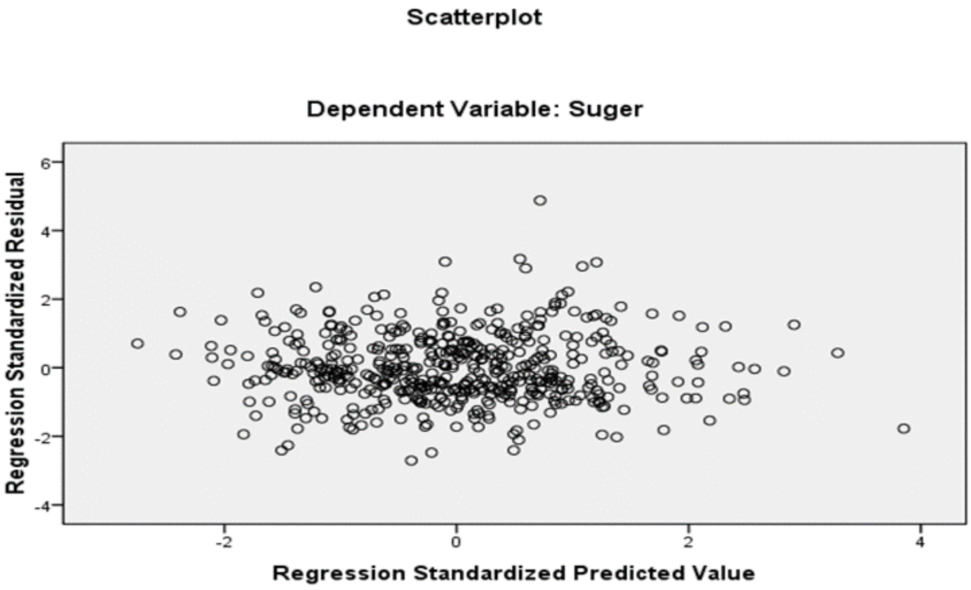


Figure (7): Standardized Residuals for Multiple regression model using (OLS)

DOI: http://dx.doi.org/10.25098/2.2.32

$AIC = nlog(Mean\ Square\ Error) + 2k$, is an alternative formula for least squares regression type analyses for normally distributed errors, minimum AIC is the score for the best least square model: AIC= 1775.66

4-Results Analysis and Conclusion:

After applying censored and truncated regression models, also multiple regression with least square method on the sample data (n=500) for persons whom their rates are exceed 120gm/dl under the risk of injury with diabetes, by taking the hypothesis that the blood sugar (Y), depends on the explanatory (Age: X_1, Cholesterol: X_2 gram/decilitre, Triglycerides: X_3 gram/decilitre), and comparing their results, the following important points are introduced.

1- The censored regression model for sample under consideration was more applicable than the other regression models (truncated, and multiple regression), this result was detected by comparing their AIC, and also Log-likelihood values (the minimum AIC, the best fitted model, also the maximum log-likelihood value the best fitted model).

2- The two explanatory (Age, and Triglycerides have highly significant effects on the blood sugar, but not cholesterol which appeared has no significant effects in data sample.

3- Recall the table (3) above, concerned with the marginal effects as defined theoretically. The change in age one year makes significant increasing in blood sugar by %2 as a mean effect for each person in the sample with standard error (0.35) by remaining the other explanatory fixed. And also one unit of triglycerides causes significant increases by %0.22 approximately with slandered error 0.03.

4- The results of multiple regression model which was estimated by ordinary least square method (tables 5, and 6), indicates that this model is not an adequate candidate model with maximum AIC value if it compares with the reminder regression in this application, especially with censored candidate regression model because the ordinary least square estimates(unconditional estimates) are biased , and moreover it deals with the origin data (without censoring or truncating) of the dep. variable (blood sugar) that contents (n=500) in fitting multiple regression model, so they are not consists , in other hand the likelihood estimates for the censored regression are conditional estimates,(the condition is that all dependent variable’s observation where exceeds (120 gm/dl) are all zeros), makes these estimates unbiased and consists.

DOI: http://dx.doi.org/10.25098/2.2.32

References:

1- Alferd. D, and Bowling, G. “Regression with Social Data Modelling Continues and Limited Response Variables”,(2004). State University, Dept. of Sociology, Ohio, John Wiley & Sons, Inc., Publication, Hoboken, New Jersy.

2- Franses, P. H and Richard. P,(2001). “Quantitative models in Marketing Research “Cambridge University Press.

3- Green, W. H, “Econometric Analysis” (2003). Fifth Edition, Upper Saddle River, NJ, Prentice Hall.

4- Long, J. S. “Regression Models for Categorical and Limited Dependent Variables” (1997). Thousand Oaks, CA: Saga Publication.

5- Maddala, G. S. “Limited Dependent and Qualitative Variables in Econometrics” (1983). Cambridge University Press, Cambridge, UK. ISBN 0-521-33825-5. OCLC 25207809.

6- Richard Breen. “Regression Models Censored Sample Selected or Truncated Data”, (1996). Saga Publication Ltd. California 9130, Email:order@sagpub.com, ISBN 0-8039-5710-6.

7- Stock, James H.; Watson, Mark W. “Introduction to Econometrics” (2003). Addison-Wesley, Bosten. ISBN 0-201-71595-3.

8- Tobin, J. “Estimation of relationships for limited dependent variables” (1958), Econometrica. 26(1): P:(24-36).

9- Toomet. O, Henningsen. A” Tools for Maximum Likelihood Estimation”. (2010). R package version 0.7, http://CRAN.R-project.org/package=maxLik. maxLik.

10- Wooldridge, J. M. “Econometric Analysis of Cross Section and Panel Data” (2002). MIT Press, Cambridge. ISBN 0-262-23219-7. OCLC 47521388.

DOI: http://dx.doi.org/10.25098/2.2.32