American Review of Mathematics and Statistics
December 2016, Vol. 4, No. 2, pp. 63-73
ISSN: 2374-2348 (Print), 2374-2356 (Online)

Published by American Research Institute for Policy Development
DOI: 10.15640/arms.v4n1a7
URL: https://doi.org/10.15640/arms.v4n1a7

Bayesian Tobit Principal Component Regression with Application

Fadel Hamid Hadi Alhusseini[1]

Abstract

Tobit is considered an important statistical modeling, as it has become common in numerous practical applications, such as econometrics, biological sciences, finance, and medicine. The Tobit regression model is special case of censored regression model at zero point. In some cases, Tobit regression model is exposed to econometrics problems, including the multicollinearity problem. Most of the independent variables from economic, social, and medical field are overlapping with each other, thus, estimating the parameters of the Tobit model may lead to are biased and inconsistent estimation. In order to overcome this issue, a set of methods can be deployed, such as principal component. In this paper, the Bayesian technique will be used for estimating the parameters of the model. The case study included in the paper is the medical problem of the abortion within Iraqi women. The estimation was performed by using Bayesian Tobit principal component regression model through building algorithm in programing language (R).

Keywords: Bayesian Technique, Principal Component Regression, Tobit Model, Abortion.

1. Introduction

Spontaneous abortion or miscarriage is a natural method of expulsion of the fetus from the utero, in case there are detected abnormalities, like diseases or fetus death. Spontaneous abortion occurs for 15% to 20% of all cases of pregnancy and often the pregnancies are lost in the first three months of gestation. There is a set of biological, social, and economic variables that can affect the spontaneous abortion. The phenomenon under study contains one response variable that has two parts: the number of abortions that equal zero (censored part) and the number of abortion that equal one or are higher than one (uncensored part). The Tobit model is based on data of the phenomena under study. The censored part takes cumulative distribution function (c.d.f) for normal distribution and the uncensored part takes probability density function (p.d.f) for normal distribution.

As a result, the Tobit model is mixture between (c.d.f) and (p.d.f) for normal distribution. In this study, the response variable is affected by a set of independent variables. After testing the model, the results show that the model suffers from multicollinearity problem, according to Farrar-Glauber test. In order to overcome this issue, the principal components method will be used by transforming the original linked variables into unlinked new variables, representing the principal components. Each principal component will contain original variables. For estimating the parameters for Tobit principal components regression model, the Bayesian technique will be employed, after filtering the model from multicollinearity problem, through using principal components method. This parameter will be used for estimating the original coefficients for Tobit model through building algorithm in programing (R).

[1] Department of Statistics and Economic Informatics, University of Craiova, Romania.

The rest of this paper is organized as follows: Section 2 includes a study of the Tobit regression model, Section 3 illustrates the methodology of principal component regression**,** Section 4 includes the Bayesian Tobit principal components regression, Section 5 details the estimation of original parameters, Section 6 shows a sample of the case study, Section 7 includes the analysis of the case study results and Section 8 highlights the conclusions.

2. Tobit regression model

The Tobit model was proposed in the middle of the last century by researcher James Tobin (1958). The Tobit model considers special a case of censored regression model, where the censored point is equal to zero (a=0). The censored regression model general formula is as follows:

$$Y = \begin{cases} a if y_i^* \le a \\ y_i^* \quad if y_i^* > a \end{cases} \qquad 1)$$

$$where y_i^* = x_i^T \beta + \varepsilon_i, \qquad i = 1, \dots \dots, n$$

whereT_i is an $(n \times 1)$ vector of latent variable, x_i^T is a $1 \times k$ vector of explanatory variables, β is a $(k \times 1)$ vector of regression parameters and ε_i is an $(n \times 1)$ vector of random error, where $\varepsilon_i \sim N(0, \sigma^2)$.

If (a=0), the censored regression model will become Tobit regression model as follows:

$$Y = \begin{cases} 0 \qquad if y_i^* \le 0 \\ y_i^* \quad if y_i^* > 0 \end{cases} \qquad 2)$$

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

Whereas, latent variable distributes normal distribution for mean $(x_i^T \beta)$and variance (σ^2), then the probability density function ofT_i takes the formula as follows, if $pro(Y = y_i^*)\ if\ pro(y_i^* > 0)$:

$$p(y_i^*) = \frac{1}{\sqrt{2\pi\sigma^2}} Exp\left(\frac{\left(y_i^* - x_i^T \beta\right)^2}{2\sigma^2}\right) \qquad 3)$$

The probability density function in (3) belongs to continuous part (uncensored part). Equation (3) can be rewritten as follows:

$$p(Y) = \frac{1}{\sigma} \phi\left(\frac{y_i^* - (x_i^T \beta)}{\sigma}\right) \qquad 4)$$

and the censored part will take (c.d.f) for normal distribution as follows, if

$$pro(Y = 0) if pro(T_i \le 0) \longrightarrow \Phi\left(\frac{Y - (x_i^T \beta)}{\sigma}\right) = \Phi\left(\frac{0 - (x_i^T \beta)}{\sigma}\right) = \Phi\left(\frac{-(x_i^T \beta)}{\sigma}\right) = 1 - \Phi\left(\frac{(x_i^T \beta)}{\sigma}\right) \qquad 5)$$

From equations (4) and (5), the Tobit model is mixed function between probability density function and cumulative distribution function for normal distribution. $\phi(.)$ is a probability density function (p.d.f) and $\Phi(.)$ is a cumulative distribution function (c.d.f).

$$p(Y) = \left[\frac{1}{\sigma}\phi\left(\frac{y_i^* - (x_i^T\beta)}{\sigma}\right)\right]\left[1 - \Phi\left(\frac{(x_i^T\beta)}{\sigma}\right)\right] \quad 6)$$

The function of the Tobit model in (6) is mixed between censored and uncensored parts (see William H. Greene, 2007). For estimating parameters of Tobit regression model may be depended on maximum likelihood or (O.L.S), through numerical method. Estimating parameters of Tobit model has become a routine operation because of the numerous available packages. Nevertheless, in this paper, the Tobit model suffers from multicollinearity problem. Therefore, the issue must be solved before estimating its parameters.

3. tobit principal component regression method

Each regression models is exposed to a set of econometrics problems. One of these problems is the multicollinearity problem. This problem appears when the independent variables are linked linearly or when the independent variables have perfect relationship. This perfect relationship leads to violation of the rank condition (rank(X)<p, where p is the number of independent variables). In this case, we cannot find the inverse information matrix (X^tX). Therefore, we cannot estimate the parameters model. In some case, the determinant of the information matrix can be relative for zero ($|X^tX| \approx 0$). In this case, the parameters model can be estimated, but these estimations will be inaccurate, as the variance of the parameters will be very large. Therefore, test (t) gives fake information about the independent variables confidence. A reasonable method to overcome this problem must be identified. For solving the multicollinearity problem, various sets of methods are available. Each method has properties and characteristics. In this paper, we will use the principal components method in the treatment of the multicollinearity problem in Tobit model. The principal components are orthogonal linearity compotation from independent variables (X) as follows:

$$Z = X\Gamma \quad 7)$$

where Z is the matrix of principal components from rank $(n * q)$, Γ is the orthogonal matrix from eigenvector corresponding to eigenvalue in the information matrix (X^tX) with rank $(q * q)$, elements$\gamma_{ij} i = 1, \dots \dots n$and column Γ_j(j=1,2,…..q). This matrix makes from information matrix (X^tX) orthogonal matrix under the assume eigenvalue for matrix(X^tX). For $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_q$ the composition Tobit regression model in equation (3) on principal components $Z$where the latent variable $T_i$as function of orthogonal principal component instead of original independent variables (X):

$$\begin{gathered}(Z = X\Gamma) * \Gamma^t \\ Z\Gamma^t = X\Gamma\Gamma^t \text{ where } \Gamma\Gamma^t = \Gamma^t\Gamma = I_q\end{gathered} \quad 8)$$

$$X = Z\Gamma^t \quad 9)$$

After compensation of the equation (10) in equation (3) the Tobit principal components regression model will be as follows:

$$Y = \begin{cases} 0 & if y_i^* \leq 0 \\ y_i^* & if y_i^* > 0 \end{cases} \quad 10)$$

Where $T_i = Z\Gamma^t\beta + \varepsilon_i$,

Let's assume $\Gamma^t\beta = \theta$, therefore the equation (11) becomes as follows:

$$Y = \begin{cases} 0 & if y_i^* \leq 0 \\ y_i^* if y_i^* > 0 \end{cases} \qquad 11)$$
$$y_i^* = Z\theta + \varepsilon_i$$

The known Tobit model is a mixture model between uncensored observation ($Y = 0$, if$y_i^* \leq 0$) and the censored observation ($Y = y_i^* \quad if y_i^* > 0$). Therefore, the Tobit principal components regression model takes the following form:

$$p(Y) = \left[\frac{1}{\sigma}\phi\left(\frac{y_i^* - (Z\theta)}{\sigma}\right)\right]\left[1 - \Phi\left(\frac{(Z\theta)}{\sigma}\right)\right] \qquad 12)$$

$$L = \prod_{i=1}^{N}\left[\frac{1}{\sigma}\phi\left(\frac{y_i^* - (Z\theta)}{\sigma}\right)\right]\left[1 - \Phi\left(\frac{(Z\theta)}{\sigma}\right)\right] \qquad 13)$$

We can use the maximum likelihood method for estimating the parameters of the Tobit principal component regression. In this paper, the Bayesian technique for estimating parameters of Tobit principal components regression will be used.

4. **Bayesian Tobit principal components regression**

The Bayesian technique is considered an advanced method in estimating parameters of the Tobit principal components regression. It has a set of features, such as Bayesian technique that is able to estimate parameters even if the sample size is small (see Alhamzawi, R., K. Yu, and Benoit, D. F, 2012). The Bayesian technique also updates the parameters through prior distribution. From the equation (13),the Tobit principal component regression is the following:

$$Y = \begin{cases} 0 & if y_i^* \leq 0 \\ y_i^* if y_i^* > 0 \end{cases}$$
$$y_i^* = \theta Z + \varepsilon_i,$$

where Y is the response variable,y_i^* is the latent variable, Z is principal components,ε_i is the random error term distributed according to normal distribution with mean (zero) and variance (σ^2), therefore the Bayesian hierarchicalmodel given by:

$$Y = \begin{cases} 0 & if y_i^* \leq 0 \\ y_i^* if y_i^* > 0 \end{cases}$$

where $\Rightarrow Y = max(y_i^*, 0)$anthor formula for Tobit model

$$y_i^* | \sigma^2, \beta, Z \sim N(Z\theta, \sigma^2)$$

The probability density function for latent variable$T_i$is:

$$f(y_i^* | \sigma^2, \beta, Z) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{(y_i^* - Z\theta)^2}{\sigma^2}} \qquad 14)$$

The joint distribution of $y_i^* = (y_i^*, \dots, y_i^*)^T \; given \; Z = (Z_1, \dots, Z_n)^T, \; \beta = (\theta_1, \dots . \theta_k)^T$ is the following:

$$f(y_i^*|\sigma^2, \beta, Z) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{\left(y_i^* - Z_j\theta\right)^2}{\sigma^2}} \quad 15)$$

$$f(y_i^*|\sigma^2, \beta, Z) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{\frac{\sum_{i=1}^{n}\left(y_i^* - Z_j\theta\right)^2}{\sigma^2}} \quad 16)$$

For finding the estimation of parameters(θ, σ^2) prior distribution will be used. According to Claiming yu and Mooyd (2011), the researchers can use any prior distribution for parameters, but selecting suitable prior distribution leads to good results for posterior distribution.

4.1 Prior Distribution Specification

The coefficients of the regression model have a range from$(-\infty, \infty)$. Therefore, the prior of coefficient regression takes normal distribution. By assigning zero, the normal prior distribution for θ takes the following form:

$$p(\theta|\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-\frac{\theta^2}{2\sigma^2}\right\}, \quad 17)$$

The parameter σ^2 has range $(0, \infty)$. Inverse gamma can be used for parameter σ^2, with prior taking the form of:

$$p(\sigma^2|a, b) = \frac{a^b}{\Gamma(b)} (\sigma^2)^{-b-1} e^{\frac{-a}{\sigma^2}}, \quad 18)$$

Where a, b are hyper parameters.

4.2 Posterior Computation Inferences

In order to obtain the posterior distribution of the coefficient regression model, multiple joint functions will be used in equation (16) by the prior distributions in equations (17) and (18), as follows:

$$f(\theta, \sigma^2|y_i^*, Z_i) = f(y_i^*|\sigma^2, \beta, Z_i) * p(\theta|\sigma^2) * p(\sigma^2|a, b)$$

$$f(\theta, \sigma^2|, y_i^*, Z_i) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{\frac{\sum_{i=1}^{n}\left(T_{i-Z_i\theta}\right)^2}{\sigma^2}} * \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-\frac{\theta^2}{2\sigma^2}\right\} * \frac{a^b}{\Gamma(b)} (\sigma^2)^{-b-1} e^{\frac{-a}{\sigma^2}} \quad 19)$$

Posterior distribution for parameterθ:

$$f(\theta, \sigma^2|y_i^*, Z) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{\frac{\sum_{i=1}^{n}(y_i^* - Z\theta)^2}{\sigma^2}} * \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-\frac{\theta^2}{2\sigma^2}\right\}$$

The posterior distribution θ is distributed according to the distribution of normal with mean$(\sum_{i=1}^{n} Z_i \, T_i)(\sum_{i=1}^{n} {Z_i}^2 - 1)^{-1}$and variance$\sigma^2(\sum_{i=1}^{n} {Z_i}^2 - 1)^{-1}$.

Posterior distribution for parameterσ^2:

$$f(\sigma^2|\theta, y_i^*, Z_i) = (\frac{1}{\sqrt{2\pi\sigma^2}})^n e^{\frac{\sum_{i=1}^{n}(y_i^* - Z_j\theta)^2}{\sigma^2}} * \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-\frac{\theta^2}{2\sigma^2}\right\} \quad (20) * \frac{a^b}{\Gamma(b)}(\sigma^2)^{-b-1}e^{\frac{-a}{\sigma^2}}$$

The posterior distribution of parameter$\sigma^2$is distributed according to inverse gamma distribution with Shape parameter$\frac{n-1}{2}$ + $b$and scale parameter$(\sum_{i=1}^{n}(y_i^* - Z_j\theta)^2/_2 - \theta^2/_2 - b)$. Tobit principal components regression model can be estimated through buildingGibbs samplers that are derived for all posteriordistributions, after taking the mean for thousands iterations for all posterior distributions.

5. Estimation of original parameters

We can find the coefficients$\hat{\beta}$of the original variables (X_i), through using the relationship between coefficients$\hat{\beta}$and the coefficients of the principal components$\hat{\theta}$:

$$\begin{gathered}\Gamma^t\hat{\beta} = \hat{\theta} \\ \Gamma\Gamma^t = 1 \\ \Gamma\Gamma^t\hat{\beta} = \Gamma\hat{\theta} \\ \hat{\beta} = \Gamma\widehat{\theta}\end{gathered} \quad 21)$$

The coefficients$\hat{\beta}$ are distributed according to the normal distribution by mean($\Gamma\hat{\theta}$)and variance$\sigma^2\sum_{i=1}^{n}\frac{\gamma_{ij}{}^2}{\lambda_j}$. Then,$\hat{\beta} \sim N(\Gamma\widehat{\theta}, \sigma^2\sum_{i=1}^{n}\frac{\gamma_{ij}{}^2}{\lambda_j})$.Also, the variance of parameter$\hat{\beta}$depends on the Eigen value for the information matrix(X^tX). Therefore, the variance of parameter $\hat{\beta}$ will be affected by a small Eigen values. Which leads to inflated parameters variance.

The principal component is become overlapping with various fields form statistics such as, it is overlap with fuzzy technique (seeGeorgescu, V (1996)). Also the principal component used to overcome on multicollinearity problem, the small Eigen values will be canceled corresponding to last principal components in the informationmatrix(X^tX).For reducing the total variance of the parameters, the study will focus on principal components corresponding to large Eigen values.

There is a set of criteria for excluding non-dominant principal components for analysis. The principal components with eigenvalues less than (1)will exclude from analysis, were considered (see Jeffers (1967), Chatterje and price (1991)). In the same topic, Morrison (1976), mentions that depending on principal components, 75% of the total variance can be explained. For building a Tobit regression model on dominant principal components and exclude non-dominant principal components, the number of principal components is (q), thither dominant principal components are r and non-dominant principal components is (q-r). Therefore, (q-r) is excluded from the analysis, using only r as principal components. By existence of the orthogonal feature, the values of the estimated parameters (θ) are not different, whether using all the principal components or part of them.

$$\theta_r = \Lambda_r^{-1}Z^t{}_rY \quad 22)$$

where $\Lambda_r^{-1}$is diagonal matrix: $\Lambda_r^{-1} = diag(\lambda_1, \lambda_2, \ldots, \lambda_r)$.

The parameter vector θ_r will result through partitioning the matrix$\Gamma = [\Gamma_r, \Gamma_{q-r}]$ where: $\Gamma_r = [\Gamma_1, \Gamma_2, \dots \dots \Gamma_r]$, $\Gamma_1 = \gamma_{11}, \gamma_{12}, \gamma_{13}, \dots\dots \gamma_{1r}$.

In this case, the principal component may be reduced as follows: $Z = [Z_r, Z_{q-r}]$. Therefore, the principal components in the analysis are: $Z = [Z_1, Z_2, \dots \dots, Z_r]and\theta_r = [\theta_r, \theta_{q-r-1}]$.

After calculating the results on estimators for the parametersθ_r, they will be used for estimating the original parameter of the Tobit model, as follows:

$$\hat{\hat{\beta}}_\iota = \sum_{i=1}^{r} \Gamma_r \hat{\theta}_r \qquad 23)$$

The estimation of the original parameter for Tobit model will result from the parameters dominant principal components:

$$\hat{\hat{\beta}}_\iota = \sum_{i=1}^{r} \gamma_{ij} \hat{\theta}_r, i = 1,2,3, \dots . q \qquad 24)$$

Through the equation (22) the Tobit regression model parameter can be estimated after building the algorithm in programming language {R}.

6. Sample of the case study

Data for the current study was collected from the Al-Shamia hospital in Iraq with the sample size consisting of 300 observations. The sample of study has one response variable, which is the number of spontaneous abortions at women. Some of the respondents have had several abortions, while some of them reported none. The response variable of this study is censored at zero point. The censored observations number is (215), by percentage - (64%). The uncensored observations number is (85), by percentage - (36%). This study contains 10 independent variables described below:

X_1- Mother's age: abortion number may be influenced by the age of the mother; from medical point of view, there are biological changes in the mother's womb when the mother is above 18 years of age

X_2- Mother's weight: the high or low weight of the mother may influence the occurrence of spontaneous abortion

X_3– Mother's blood pressure: the elevated or low blood pressure of the mother may influence the occurrence of spontaneous abortion

X_4– mother's blood sugar:the blood sugar levels of the mother may influence the occurrence of spontaneous abortion; from medical point of view, the mother certain blood sugar levels cause increased weight of the fetus

X_5- Number of births: frequent births may influence the occurrence of spontaneous abortion

X_6- Monthly income of the family: abortion may be affected by the type of food available to the mother during the pregnancy; the quality of the food depends on the family income

X_7- Working hours of the mother: there is a negative relationship between the number of working hours of the mother and the occurrence of spontaneous abortion

X_8 – Progesterone: low levels of progesterone hormone of the mother may influence the occurrence of spontaneous abortion

X_9 – Misuse of medicine: may cause problems for the fetus and this lead to spontaneous abortion

X_{10}– Toxoplasmosis is considered one of the most common diseases in Iraq, and in some cases this disease causes spontaneous abortion.

After collecting all data, the algorithm for the data analysis was built using {R}.

7. Analysis of study Results

The independent variables suffer from multicollinearity problem, as can be seen in Table no. 1.

Table no. 1 - Test of Farrar-Glauber

$x^2_{\text{Calculated}}$	$x^2_{\text{Tabulated}}$
x^2_{C} = 12543.213	x^2_{T} = 122.34

From Table no. 1, x^2_{C} is greater than x^2_{T} ,therefore, the conclusion is that the model suffers from multicollinearity problem and the parameters of Tobit regression model cannot be estimated due to this issue, as the estimation will be abnormal. In order to overcome this issue, the principal component method will be used.

Table no. 2 - Eigen Values and Eigen Vector for Principal Component

	Principle components									
	Z_1	Z_2	Z_3	Z_4	Z_5	Z_6	Z_7	Z_8	Z_9	Z_{10}
	Eigen value λ									
	11.56	11.32	11.27	11.19	11.1	11.00	00.82	00.74	00.62	00.20
	Reduction of variance %									
	221.23	119.34	117.1	113.39	110.87	66.54	44.45	33.7	22	11.38
	collecting of variance %									
Vvariables	221.23	440.57	557.67	771.06	881.93	888.47	992.92	996.62	998.62	1100
	γ_{i1}	γ_{i2}	γ_{i3}	γ_{i4}	γ_{i5}	γ_{i6}	γ_{i7}	γ_{i8}	γ_{i9}	γ_{i10}
X_1	00.543	00.324	-0.232	00.076	-0.122	-0.023	00.300	00.729	00.046	00.088
X_2	00.234	00.823	-0.112	-0.001	00.056	00.196	-0.617	-0.365	00.138	00.017
X_3	00.630	00.345	00.232	00.295	00.163	-0.292	-0.070	00.482	-0.047	00.006
X_4	00.034	-0.295	00.063	00.146	00.230	-0.451	-0.147	-0.318	00.366	-0.681
X_5	00.427-	00.462	00.431	-0.383	-0.165	-0.340	00.266	00.289	00.329	00.023
X_6	00.754-	00.002	-0.094	-0.077	-0.337	00.374	00.091	-0.024	-0.002	00.716
X_7	00.231	00.438	00.287	-0.129	-0.493	-0.046	00.289	00.338	00.288	-0.017
X_8	-0.212	-0.325	00.342	00.597	00.255	00.091	-0.024	00.012	00.015	-0.006
X_9	00.234	-0.234	00.098	00.076	00.041	-0.118	00.338	00.647	-0.002	-0.152
X_{10}	00.110	00.234	00.032	-0.245	-0.255	-0.141	00.440	-0.030	-0.338	0.482

As can be seen in Table no. 2 the dominant principal components of this study are six, where the extraction variance for these five principal components is 84% from total variance. Therefore, the remaining four principal components corresponding to Eigen value lower than 1 which increase the variance coefficient of regression, will be excluded.

Excluding on-dominant principal components has no effect on building the principal components regression for Tobit model because any principal component contains all independent variables. As a first step, the response variable (Number of abortions per woman) was composed on dominant principal components as shown in following table.

Table no. 3: Regression Model for Number of Abortions per Woman on dominant Principal Components

Principal component	Regression coefficient	t-value	p-value
Intercept	0.040	10.34	0.020
Z_1	-0.124	-3.74	0.023
Z_2	0.164	10.78	0.040
Z_3	0.207	13.90	0.000
Z_4	-0.066	-16.30	0.000
Z_5	0.032	5.04	0.000
Z_6	0.453	2.432	0.000

By using estimators of regression model (Number of abortions per woman) on dominant principal components, the regression model will have the following results.

Table no. 4 -Tobit Principal Component Regression Results Coefficients of Regression (Number of abortions per woman) on Original Variables

Variables	Regression coefficients	t-value	p-value
Intercept	-0.0263	-0.325	0.000
X_1	-0.0003	-7.435	0.023
X_2	0.0197	4.216	0.083
X_3	0.0170	36.187	0.000
X_4	0.0512	3.345	0.567
X_5	-0.1061	-0.180	0.756
X_6	0.008-	0.132	0.004
X_7	0.2147	4.436	0.034
X_8	0.0618	11.435	0.045
X_9	0.5012	2.934	0.000
X_{10}	0.1024	2.122	0.000

Table no. 4 shows that the independent variables either positive relationship, either inverse relationship, as follows:

X_1(Mother's age) is significant from statistical point of view; this variable is in inverse relationship with the number of abortions –if the variable (mother's age) increases, the number of abortion decreases

X_2 (Mother's weight) is not significant from statistical point of view; this variable is in positive relationship with the number of abortions: if the variable (mother's weight) increases over the limit, the probability of spontaneous abortion occurrence increases

X_3 (Mother's blood pressure) is significant from statistical point of view; this variable is in positive relationship with the number of abortions – if the variable increases (mother's blood pressure), the number of abortions increases.

X_4(Mother's blood sugar) is non-significant from statistical point of view.

X_5(Number of births) is not significant from statistical point of view.

X_6 (Monthly income of the family) is significant from statistical point of view; this variable (monthly income of the family) is in inverse relationship with the number of abortions – if the monthly income of the family increases, the number of the abortions decreases.

X_7(Working hours of the mother) is significant from statistical point of view; t in the variable (working hours of the mother) is in positive relationship with the number of abortions.

X_8(Progesterone) is significant from statistical point of view; the variable (progesterone) is in inverse relationship with the number of abortions – if the level of progesterone decreases, the number of abortions increases.

X_9(Misuse of medicine) is significant from statistical point of view; the variable (misuse of medicine) is in positive relationship with the number of abortions – if the variable decreases, the number of abortions decreases.

8. Conclusion

The model under study suffers from multicollinearity problem, whichis obvious in the Farrar-Glauber test , as shown in the Table no.1.

$$\sum_{i=1}^{p} \lambda_{i=1}^{-1} = 14.11 \quad , p number of independent variables$$

Where the value of ($\sum_{i=1}^{p} \lambda_{i=1}^{-1}$) is greater than the number of independent variables($p = 10$), meaning that the model has multicollinearity problem.

The number of dominant principal components is six out of ten. The dominant principal components can explain 88.47% of the total variance, meaning that the dominant principals component have explanatory power.

According to the phenomena under study, the significant variables on the number of abortions are the following:

X_1 (Mother's age) – is in inverse relationship with the number of abortions.
X_3 (Mother's blood pressure) – is in positive relationship with the number of abortions.
X_6 (Monthly income of the family) – is in inverse relationship with the number of abortions.
X_7 (Working hours of the mother) – is in positive relationship with the number of abortions.
X_9 (Misuse of medicine) - is in positive relationship with the number of abortions.
X_{10}(Toxoplasmosis)- is in positive relationship with number of abortions.

References

Alhamzawi, R., K. Yu, and D. F. Benoit (2012). Bayesian adaptive Lasso quantile regression.*Statistical Modelling 12*, 279–297. 5, 16.

Intrilligator, M. D.,(1996)" Econometrics Models", Techniques and Applications", Prentice Hall.

Farrar, D. E. and Glauber, R. R. (1967) Multicollinearity in regression analysis: the problem revisited, Review ofEconomics and Statistics, **49**, 92-107.

Georgescu, V (1996). A fuzzy generalization of principal components analysis and hierarchical clustering. In *Proceedings of the Third Congress of SIGEF*, Paper 2.25. Buenos, Aires.

JEFFERSJ., N. R. *(1967).*Two case studies in the application of principal component analysis. *Appl. Statist.,* 16, *225-236.*

Tobin, James (1958)"Estimation of Relationships for limited dependent Variables " Econometrica, January 26, pp24-36 .

Yan, Xin, Gang Su, Xiao, 2009: "Linear Regression Analysis, Theory and Computing". puplished by world scientific puplishing. Co. Pte. Ltd, London.

Yu, K. and R. A. Moyeed (2001). Bayesian quantile regression. *Statistics &ProbabilityLetters 54*, 437–447. 4, 6.

Reference to a chapter in an edited book:

Morrison (1976)MULTIVARIATE. STATISTICAL METHODS. The Wharton School. University of Pennsylvania. McGraw-Hill, Inc. New York St.

Chatterjee, S. and Price, B. (1991), "Regression Diagnostics", New York: John Wiley.

william,H, Greene. (2007). Econometric analysis. Book. New York University. Seven edition.