

PRINCIPAL COMPONENT REGRESSION FOR TOBIT MODEL AND PURCHASES OF GOLD

Fadel Hamid Hadi ALHUSSEINI¹
Meshal Harbi ODAH²

ABSTRACT

This study focuses on Tobit principal component regression model in the analysis of studied data when the response variable is censored at zero point. The studied model focuses on the gold quantity purchase by individuals. When examining the model, it suffers from multicollinearity problem. Which is considered a risky problem for the stability of the model and affects the accuracy of parameters estimation. Estimating regression model coefficients under multicollinearity problem will produce inaccurate estimations. Therefore, in order to obtain a viable estimation, the multicollinearity problem must first be handled. There are various methods for treating this problem, one of them being the method of principal component. In this paper, there are five dominant principal components selected based on the Eigenvalue greater than one. There are also six important independent variables in Quantities of gold purchased (gender of the respondent, monthly income of the respondent (in US dollars), (price of gold during the period of the study, age of the respondent, region of residence and marital status of the respondent). The rest of the independent variables was unimportant in Quantities of gold purchased, with more details presented in section four.

KEYWORDS: *Tobit regression model, principal components, gold purchase, Eigen value, Eigen vector*

JEL CLASSIFICATION: *C24, C36, A13*

1. INTRODUCTION

Gold is considered a very precious metal and is used as a monetary unit by many states. Also it is used in jewelry and gems industry. Economically, the price of gold tends to rise when the individual's confidence is shaken in financial markets, because gold is used as commodity in crisis. Often, global events affect the demand on gold because it is considered a safety source. During economic or political turmoil, gold purchasing is considered a safe way against inflation and currency devaluation. The value of gold on long term is more stable compared to currencies, and it is considered as an investment without risk, being strong in economic crisis. As a result individuals feel encouraged to purchase gold, when the value of currencies are expected to drop. Another important factor that influences the gold price is the value of U.S. dollar - when the dollar is strong, the gold price will be weak, and vice versa.

The bank's collapse has made gold purchasing seem like a safe alternative to other investments. The United States and several European states have large reserves in gold. More recently, these states bought more gold for their reserves. When central banks started to purchase higher quantities in gold, the gold price started to increase. More than half of the gold demand is for the jewelry

¹ University of Craiova (Department of Statistics and Economic Informatics), Romania, fadhelfadhel222@yahoo.com

² The Bucharest University of Economic Studies (Department of Statistics and Econometrics), Romania, Mesheal_11@yahoo.com

industry. China, India and the United States have the highest gold demands, including golds used in industrial applications, as electronic devices, computers and medical devices.

In this paper the relationship between response variable (amount of gold purchased by Iraqi individual) and a set of independent variables will be discussed. The response variable is represented by the amount of gold purchase, which varies from one individual to another. Some people reach quite high purchase levels, while others have none. Through description of the response variable, it is censored at zero point, and free in the other part. Therefore, the Tobit regression model is an appropriate model for this study.

The data was collected through a questionnaire and the sample size was of 250 respondents. After examination of this model, it was observed that it suffers from multicollinearity problem. In order to overcome this problem, the principal component method for Tobit model was employed. For data analysis, an efficient algorithm in programming (R) was built. The rest of this paper is organized as follows: Section 2 presents the methodology of Tobit principal component regression model, Section 3 illustrates a sample of the study, Section 4 includes an analysis of the study results and Section 5 highlights the conclusions of the paper.

2. METHODOLOGY OF TOBIT PRINCIPAL COMPONENT REGRESSION MODEL

The regression model is a tool that shows the relationship between the response variable and a set of independent variables (Yan and Gang, 2009). This regression models is exposed to a set of econometrics problems, one of them being the multicollinearity problem. This problem appears when independent variables are correlate. The parameters estimation with multicollinearity problem will cause quite extensive issues because the parameters estimation is inaccurate. Therefore, the problem must be treated. The research about a suitable method to solve multicollinearity problem will be reflected on the accuracy of parameters estimation. A classical method used to overcome this problem is to identify the variables which caused the multicollinearity problem and eliminate them. Nevertheless, in some cases, these variables may prove important in building the model and omitting them will cause other issues.

In order to overcome the multicollinearity problem, there a set of methods is available, such as principal component method for overcoming this problem. The philosophy of this method is that ,the correlated independent variables will be transformed into new uncorrelated variables, as principal components. The Tobit regression is one of the regression models, affected by multicollinearity problem. Therefore ,the principal component method will be used to treat the multicollinearity problem in Tobit regression model. Then, a new Tobit model containing censored response variable and principal component as independent variable(new variable) and its parameters will be estimated by using the estimation method of Tobit regression model. Afterwards, these estimations will be used in estimating the parameters of original independent variables. The Tobit model has specific properties for the censored response variable, being different than the rest of regression models. The response variable in the Tobit regression is a censored - variable at point called censored point. In Tobit model the censored point is zero (Tobin, 1958). The important formula for the Tobit model is:

$$Y = \begin{cases} 0 & \text{if } T_i \leq 0 \\ T_i & \text{if } T_i > 0 \end{cases} \quad (1)$$

where $T_i = X_i\beta + U$, $i = 1, \dots, n$

T_i is a latent variable and U is the random error term, which is distributed according to normal distribution by mean equal zero and constant variance σ^2 $U \sim N(0, \sigma^2)$. x_i^T is a set of original independent variables. The Tobit model is affected by a set of economic problems, including

multicollinearity problem. One of the methods available for handling this problem is the principal component method. It is an orthogonal linear combination from independent variable (X).

$$F = X\Lambda \tag{2}$$

where F represents the matrix of principal component from rank ($n * q$), Λ is the orthogonal matrix from Eigen vector corresponding to the Eigen value in the information matrix ($X^t X$) with rank ($q * q$), elements θ_{ij} ($i = 1, \dots, n$) and columns Λ_j ($j=1,2,\dots,q$). This matrix is composed from the information matrix as diagonal matrix under the assumption of the Eigen value for matrix ($X^t X$) $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_q$. For the composition of the Tobit regression model in the equation (1) on principal component F , the variable T_i is considered the function orthogonal principal component instead of the correlated independent variables (X). The matrix Λ is orthogonal, meaning that $\Lambda\Lambda^t=1$.

$$\begin{aligned} (F = X\Lambda) * \Lambda^t \\ F\Lambda^t = X\Lambda\Lambda^t \\ X = F\Lambda^t \end{aligned} \tag{3}$$

From the above formulations in (1) and (3) , the Tobit model will be as follows:

$$Y = \begin{cases} 0 & \text{if } T_i \leq 0 \\ T_i & \text{if } T_i > 0 \end{cases} \tag{4}$$

where $T_i = F\Lambda^t\beta + U$,

Assuming $\Lambda^t\beta = \varphi$, the equation (4) becomes as follows:

$$Y = \begin{cases} 0 & \text{if } T_i \leq 0 \\ T_i & \text{if } T_i > 0 \end{cases} \tag{5}$$

where $T_i = F\varphi + U$.

The equation (5) is a formula of Tobit regression model with principal component as new independent variables for the Tobit principal component regression model. From known the Tobit model is a mixture model between the censored observation ($Y = 0$ if $T_i \leq 0$) and the not censored observation ($Y = T_i$ if $T_i > 0$). Thus, the function of the Tobit model can be expressed as follows:

$$p(Y) = \left[\frac{1}{\sigma} \phi \left(\frac{Y - (X_i\beta)}{\sigma} \right) \right] \left[1 - \Phi \left(\frac{(X_i\beta)}{\sigma} \right) \right] \tag{6}$$

where $\phi(\cdot)$ is the probability density function (p.d.f) and $\Phi(\cdot)$ is the cumulative distribution function (c.d.f). Therefore, in equation (6) the principal component can be used as independent variable , then , the function of the Tobit model becomes as follows:

$$p(Y) = \left[\frac{1}{\sigma} \phi \left(\frac{Y - (F\varphi)}{\sigma} \right) \right] \left[1 - \Phi \left(\frac{(F\varphi)}{\sigma} \right) \right] \tag{7}$$

$$L = \prod_{i=1}^N \left[\frac{1}{\sigma} \phi \left(\frac{Y - (F\varphi)}{\sigma} \right) \right] \left[1 - \Phi \left(\frac{(F\varphi)}{\sigma} \right) \right] \tag{8}$$

$$\ln L = \sum_{i=1}^N \left[-\ln(\sigma) + \ln \left(\phi \left(\frac{Y - (F\varphi)}{\sigma} \right) \right) + \ln \left(1 - \Phi \left(\frac{(F\varphi)}{\sigma} \right) \right) \right] \quad (9)$$

$$\ln L = -N \ln(\sigma) + \sum_{T_i > 0} \ln \left(\phi \left(\frac{Y - (F\varphi)}{\sigma} \right) \right) + \sum_{T_i \leq 0} \ln \left(1 - \Phi \left(\frac{(F\varphi)}{\sigma} \right) \right) \quad (10)$$

for estimating the parameters of Tobit principal component regression model through using likelihood method after using numerical methods. Tobit estimating has become routine, through using statistics programming. In this paper we build tractable algorithm through (censReg) function in programming language (R). After estimating the vector parameters of Tobit principal component regression, we must account the output variance for each principal component according forum below:

$$\left(\frac{\lambda_j}{\sum_{j=1}^p \lambda_j} \right) * 100\% \quad (11)$$

where λ_j is Eigen value and p is original independent variable

Through the definition of the principal component it results that a small Eigen value, which contributes to increasing the variance of parameters regression model, always corresponds to the last principal component for matrix $(X^t X)$. In order to reduce the total variance for the parameters, the principal components corresponding to small Eigen value will be excluded, according to the philosophy of the principal component method. There are several methods for excluding this principal component from the analysis, and for the present study, the principal component corresponding to Eigen value lower than 1 will be excluded.

In same topic, Morrison (1976) mentioned that the principal components are reliable for analysis, as they explain 75% of the independent variables variance. For illustrating how to build a Tobit principal components regression model for the response variable (y) on the rest of the principal components, after excluding the weak principal components, the number (p) from Eigen value for matrix $(X^t X)$ is assumed. There are large (r) Eigen values let (r), while the number of $(p-r)$ from the Eigen value is small. Therefore, $(p-r)$ can be excluded from the principal component (F). The Tobit principal components regression model is applied to the rest of the principal components for the composition of the response variable (Y).

By orthogonal property, the estimated values of the parameters φ are not different whether using all the principal components or part of them. Therefore, the Tobit estimation calculated is the following:

$$\varphi_r = \vartheta^{-1}_r F^t_r Y \quad (13)$$

where $\vartheta_r = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_r)$

A diagonal matrix resulted, through dividing the matrix φ_r for parameters $\Lambda = [\Lambda_r, \Lambda_{q-r}]$.

$\Lambda_r = [\Lambda_1, \Lambda_2, \dots, \Lambda_r]$, $\theta_{12}, \theta_{13}, \dots, \theta_{1r}$.

In this case, the principal components can be reduced according to $F = [F_r, F_{q-r}]$, where the principal components which are important for the analysis are $[F_1, F_2, \dots, F_r]$. Also, the range of vector of the principal components parameters is determined according to the principal components which will be included in the analysis: $\varphi_r = [\varphi_r, \varphi_{q-r-1}]$. After obtaining the estimated vector of parameters φ_r ,

After parameters estimation of Tobit principal component regression model(φ_r). Therefore, we will estimate the original parameters $\hat{\beta}$ for the original independent variable X_i through using the relationship between the parameters $\hat{\beta}$ and Tobit principal component regression model $\hat{\varphi}$

$$\begin{aligned} \Lambda^t \hat{\beta} &= \hat{\varphi} \\ \Lambda \Lambda^t &= 1 \\ \Lambda \Lambda^t \hat{\beta} &= \Lambda \hat{\varphi} \\ \hat{\beta} &= \Lambda \hat{\varphi} \end{aligned} \tag{14}$$

where the parameter $\hat{\beta}$ will be distributed according to normal distribution by mean $\Lambda \hat{\varphi}$ and variance $\vartheta^{-1} \Lambda \sigma^2$ $\hat{\beta} \sim N(\Lambda \hat{\varphi}, \vartheta^{-1} \Lambda \sigma^2)$. For estimating the original parameter to Tobit model takes the following formula:

$$\hat{\beta}_r = \Lambda_r \hat{\varphi}_r \tag{15}$$

$$\hat{\beta}_r = \sum_{j=1}^r \theta_{ij} \hat{\varphi}_j \tag{16}$$

The regression coefficients of the independent variables by Tobit model will result from using the rest of the coefficients of the principal component regression. From equation (16) the parameter for the Tobit regression will result after solving the multicollinearity problem by using the principal components method. The estimation of the parameter of the Tobit regression model was made by building an algorithm in language (R) after depending on (censReg) function.

3. SAMPLE STUDY

The data was collected through applying a questionnaire on 250 individuals, both male and female. This questionnaire contains a set of questions. One response variable is represented by the quantity of gold purchased, measured in grams. This variable is censored at zero point, as some individuals indicated a zero gram purchase, while others a certain quantity. The quantity differs from one individual to another. This study contains 89 censored observations (a rate of 36%), 161 uncensored observations (a rate of 64%) and 9 independent variables, classified according to the following:

- (a) X_1 - gender of the respondent - females tend to purchase more gold than males; because the females use gold as jewelry, and for savings, while males buy gold for protecting their money from the possible financial changes
- (b) X_2 - age of the respondent
- (c) X_3 - monthly income of the respondent (in US dollars)
- (d) X_4 - education level of the respondent (elementary, preparatory, university and high certificate)
- (e) X_5 - marital status of the respondent (single, wedded, divorced, widowed)
- (f) X_6 - number of family members
- (g) X_7 - purpose of the gold purchase (jewelry, savings, financial assets)
- (h) X_8 - region of residence (countryside, city)
- (i) X_9 - price of gold during the period of the study (in US dollars)

After transforming the data from the questionnaire in order to obtain statistical data structure, the nominal variables were transformed into quantiles. Also the analysis of the Tobit regression model requires the data are distribution according to normal distribution. where, based on the central limit theorem if the sample size is greater than 30 observations the data variables will be approximately

in normal distribution . An algorithm was built in programming language (R) for the analysis of the phenomenon data.

4. ANALYSIS OF STUDY RESULTS

In order to determine the variables influencing the gold purchase at Iraqi individual level, the Tobit regression model will be composed for the response variable, which represents the quantities of purchased gold, for the independent variables as listed above. The regression model was tested on the studied phenomenon in order to identify the econometrics problems, more precisely the multicollinearity problem. The model under study suffers from multicollinearity problem, identified through using the Farrar-Glauber test:

Table 1. Farrar-Glauber test results

$x^2_{\text{Calculated}}$	$x^2_{\text{Tabulated}}$
$x^2_{\text{C}} = 14849.519$	$x^2_{\text{T}} = 124.34$

Table 1. shows that, by comparing the value of x^2_{C} with x^2_{T} , x^2_{C} is greater than x^2_{T} ,proving that the model suffers from multicollinearity problem. Therefore, the parameters of the Tobit regression model cannot be estimated directly, because the estimation will be inaccurate. First, the issue must be treated by using the principal component method.

Table 2. Eigen values and Eigen vector for principal component

Principal component										
Variables	F_1	F_2	F_3	F_4	F_5	F_6	F_7	F_8	F_9	
	Eigen value λ									
	1.96	1.23	1.19	1.13	1.04	0.98	0.62	0.45	0.40	
	Reduction of variance %									
	24.45	21.94	16.67	12.34	9.45	5.15	4.67	3.95	1.38	
	Collecting of variance %									
	24.45	46.39	63.06	75.4	84.85	90	94.67	98.62	100	
	θ_{i1}	θ_{i2}	θ_{i3}	θ_{i4}	θ_{i5}	θ_{i6}	θ_{i7}	θ_{i8}	θ_{i9}	
X_1	-0.313	0.342	0.231	0.434	-0.232	0.231	0.345	0.131	0.045	
X_2	-0.271	-0.453	0.341	0.123	0.534	0.354	0.645	-0.190	0.320	
X_3	0.363	0.231	0.652	-0.432	-0.235	0.443	0.341	-0.233	0.650	
X_4	0.261	0.034	-0.103	0.723	0.324	0.423	0.347	-0.035	0.023	
X_5	0.256	-0.578	-0.321	-0.124	0.323	0.342	-0.323	0.673	0.532	
X_6	0.079	0.436	0.093	-0.341	-0.342	0.345	-0.236	0.099	0.234	
X_7	0.258	-0.021	0.346	0.211	0.342	0.453	-0.623	-0.034	0.328	
X_8	-0.398	-0.056	0.172	0.329	0.237	0.323	0.003	-0.233	0.239	
X_9	-0.235	0.341	-0.453	-0.343	0.043	0.673	-0.021	0.001	0.004	

From the results in Table 2., there are five important principal components in this phenomenon and the extraction variance for these five principal components is 84% from the total variance. Therefore, the rest of the principal components corresponding to Eigen value less than 1 (four components) will be excluded (see Figure (-1-)) , as they caused increase of the variance coefficients of regression. Excluding the unimportant principal components, will not affect the building of the Tobit principal component regression model, because the principal components contain all the independent variables.

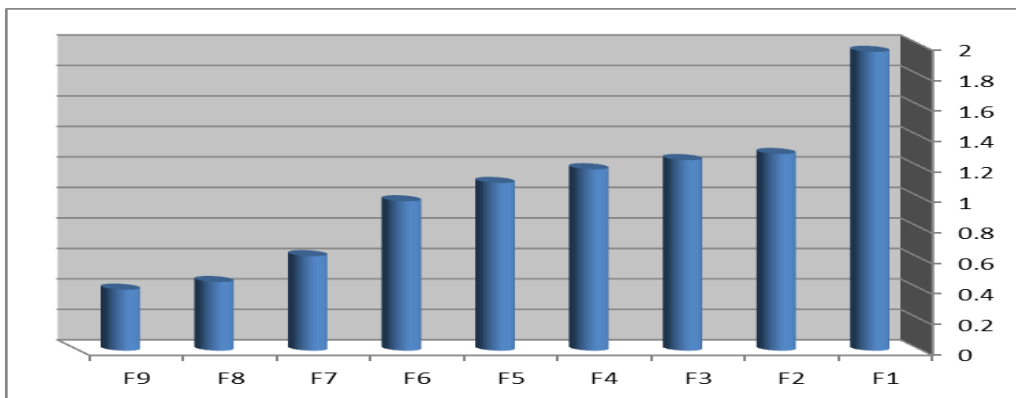


Figure (-1-) Eigen value with principal component

From Figure (-1-), the principal components with Eigenvalue greater than one are (F1,F2,F3,F4,F5), and the principal components with Eigenvalue lower than one are (F6,F7,F8,F9). Therefore the first five principal components are important.

Table 3. Regression model on important principal components

Principal component	Regression coefficient	t-value	p-value
Intercept	0.339	50.056	0.000
<i>F₁</i>	-0.153	-28.34	0.000
<i>F₂</i>	-0.092	-10.34	0.000
<i>F₃</i>	-0.040	-3.74	0.000
<i>F₄</i>	-0.124	-10.78	0.000
<i>F₅</i>	-0.164	-13.90	0.000

From the results in Table (-3-) we found all important principal component have significant relationship with the response variable. But from the results in Table (-3-) ,we will obtain the original parameters estimation for independent variables without multicollinearity problem This is shown in Table 4.

Table 4. Regression coefficients on original variables by principal component regression

Variables	Regression coefficients	t-value	p-value
Intercept	0.0024	0.429	0.047
<i>X₁</i>	0.1024	1.693	0.005
<i>X₂</i>	0.0002	26.439	0.000
<i>X₃</i>	0.0415	5.085	0.000
<i>X₄</i>	-0.0264	-3.840	0.231
<i>X₅</i>	-0.007	-0.67	0.042
<i>X₆</i>	-0.0003	-0.028	0.092
<i>X₇</i>	0.0197	3.627	0.240
<i>X₈</i>	0.0170	1.920	0.013
<i>X₉</i>	-0.5012	-10.584	0.001

From the results in Table 4., the independent variables have positive or inverse relationship with the response variable(see Figure (-2-)), as follows:

(a) X_1 (gender of the respondent) has positive relationship with (Quantities of purchased gold), and it is a significant variable from statistical point of view in response variable; this variable is identical to logic; the gender of the individual clearly influences the quantity of gold purchase, females tend to purchase more gold, as jewelry, and males buy gold in order to keep monetary assets, especially if there is a negative fluctuation in the monetary markets.

(b) X_2 (age of the respondent) is a significant variable from statistical point of view in response variable; the relationship between the age of the individual and the amount of purchased gold is a positive relationship - if the age of the individual increases by one unit, the purchased gold quantity increases by 0.0002.

(c) X_3 (monthly income of the respondent (in US dollars)) is a significant variable from statistical point of view in response variable – the purchased gold quantity and the monthly individual income are in positive relationship: if the monthly income increases by one unit, the quantity of purchased gold increases by 0.0415.

(d) X_4 (education level of the respondent) is a non-significant variable from statistical point of view in the response variable.

(e) X_5 (marital status of the respondent) is a significant variable from statistical point of view in the response variable.

(f) X_6 (number of family members) is a non-significant variable from statistical point of view in the response variable.

(g) X_7 (the purpose of the gold purchase) is a non-significant variable from statistical point of view in the response variable.

(h) X_8 (region of residence) is a significant variable from statistical point of view in the studied model; the quantity of purchased gold and the region of residence are in positive relationship.

(i) X_9 (price of gold during the period of the study) is a significant variable from statistical point of view in the response variable; the quantity of purchased gold and the price of gold are in inverse relationship – when the price decreases, the quantity increases and vice versa, corresponding to the economic logic.

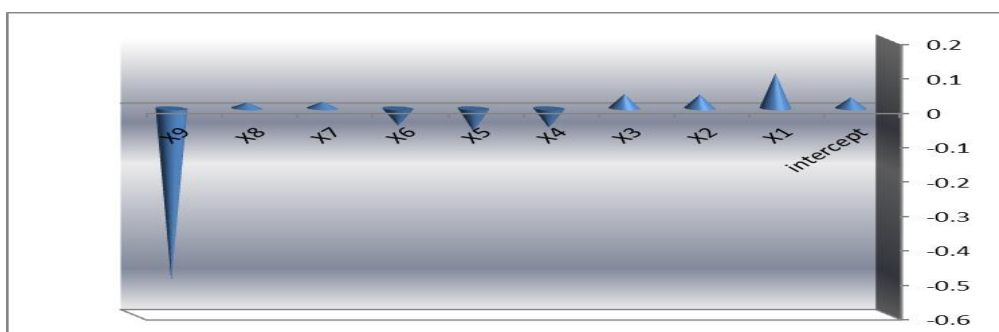


Figure (2) the relationship between original independent variable and response variable

As shown in Figure (-2-), we found that the original independent variables (X_1 , X_2 , X_3 , X_7 , X_8) have a positive relationship with the response variable (Quantities of purchased gold). And rest of the original independent variables (X_4 , X_5 , X_6 , X_9) have an inverse relationship with response variable (Quantities of purchased gold).

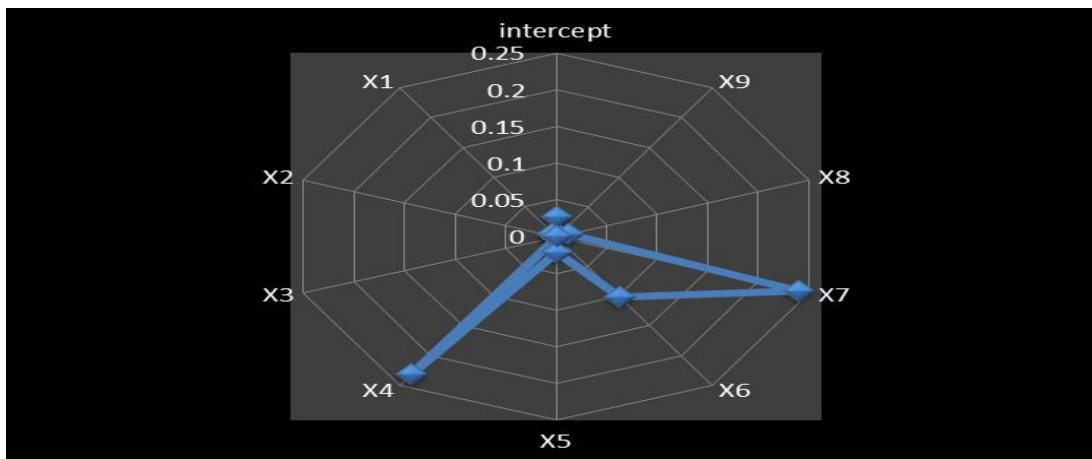


Figure (3) Original independent variables with p-value

From known p-value is a measurement of statistical significance: ,if p-value lower than (0.05),then it is significant and if p-value is greater than (0.05),then it is insignificant. From Figure (-3-) we see that the original independent variables (X_1 , X_2 , X_3 , X_5 , X_8 , X_9) have significant effect on the responses variable (Quantities of purchased gold). The original independent variables (X_4 , X_6 , X_7) have insignificant effect on the responses variable (Quantities of purchased gold).

5. CONCLUSIONS

From (**the Farrar-Glauber test**), the value of $x^2_{\text{Calculated}}$ is greater than $x^2_{\text{Tabulated}}$. This signifies that the model suffers from the multicollinearity problem. Also the value of the summation of inverse Eigen value is considered another test for the model is equal to $\sum_{i=1}^p \lambda_{i=1}^{-1} = 11.37$, where p is the number of independent variables. The result is higher than the number of independent variables ($q = 9$), which leads to multicollinearity problem between independent variables.

Also, the important principal component has a clear explanation force through ratio of extraction variance which reaches 84.85% from the total variance. This shows that the indicator is elucidated for forcing the principal components in explaining the phenomenon under study. Each principal components are significant in the purchased gold quantity.

According to the phenomenon under study, there are six significant variables on the quantity of the purchased of gold arranged according to priority and importance

- X_1 (gender of the respondent) is in positive relationship with the purchased gold quantity if the gender of the individual increases by one unit, the purchased gold quantity increases by 0.1024.
- X_3 (monthly income of the respondent (in US dollars)) is in positive relationship with the purchased gold quantity if the monthly income of the individual increases by one unit, the purchased gold quantity increases by 0.0415.
- X_9 (price of gold during the period of the study) is in reverse relationship with the purchased gold quantity if the price of gold of the individual increases by one unit, the purchased gold quantity decreases by -0.5012.
- X_2 (age of the respondent) is in positive relationship with the purchased gold quantity if the age of the individual increases by one unit, the purchased gold quantity increases by 0.0002.
- X_8 (region of residence) is in positive relationship with the purchased gold quantity if the region of residence of the individual increases by one unit, the purchased gold quantity increases by 0.0170.
- X_5 (marital status of the respondent) inverse relationship with the purchased gold quantity if the region of residence of the individual increases by one unit, the purchased gold quantity decreases by -0.007.

REFERENCES

- Chatterjee, S. & Price, B. (1991). *Regression diagnostics*. New York, John Wiley.
- Farrar, D. E. & Glauber, R. R. (1967). Multicollinearity in regression analysis: the problem revisited. *Review of Economics and Statistics*, 49, 92-107.
- Greene, W. H., (2007). *Econometric analysis*, 7th ed. New York University.
- Intrilligator, M. D. (1996). *Econometrics models, techniques and applications*, Prentice Hall.
- Jeffers, N. R. (1967). Two case studies in the application of principal component analysis. *Appl. Statist.*, 16, 225-236.
- Morrison (1976). *Multivariate. statistical methods*. The Wharton School, University of Pennsylvania. McGraw-Hill, Inc. New York St.
- Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica*, 26, 24-36.
- Yan, X. & Gang, S. X., (2009). *Linear regression analysis, theory and computing*. London, World Scientific Publishing. Co. Pte. Ltd.