CAUCHY –Jurnal Matematika Murni dan Aplikasi
Volume 5(3) (2018), Pages 80-87
p-ISSN: 2086-0382; e-ISSN: 2477-3344

Restricted Maximum Likelihood Method as An Alternative Parameter Estimation in Heteroscedastic Regression

Dwi Masrokhah, Loekito Adi Soehono, Suci Astutik

Department of Statistics, Faculty of Mathematics and Natural Sciences, Brawijaya University, Malang, Indonesia

Email: dwi.masrokhah@gmail.com

ABSTRACT

Students are part of the community who have an income. The income of student is pocket money, scholarships, part-time jobs and so forth. They are trying to become trendsetter for their dress style. The consumption patterns are very influential in the behavior of saving. If the savings increases, not only the public funds will increase but also the investment. If the investment increases, the economic growth will also increase. The purpose of this research is to estimate multiple regression parameters using REML methods in modeling the student's saving in Faculty of Mathematics and Natural Science, Brawijaya University. The variables used were: the student's age, the amount of income of student's parent, the amount of student's pocket money, the amount of student's additional income, the amount of student's consumption and the amount of student's saving. REML method can overcome heteroscedasticity of error variance and provide unbiased estimator. The model of student's saving using REML method is as follows:

$$\hat{Y}_i = -1609 + 112\,X_1 + 0.0088X_2 + 0.0504\,X_3 + 0.4706\,X_4 - 0.636X_5$$

Student's saving is affected significantly by: student's age (X_1), the amount of student's additional income (X_4), and the amount of student's consumption (X_5).

Keywords: Student's Saving, REML, regression, assumptions, Heteroscedasticity

INTRODUCTION

The regression analysis is used to create a functional model of the data to explain or predict a natural phenomenon based on the other phenomena. Regression analysis was introduced by Sir Francis Galton in 1822-1911. The purpose of regression analysis is for prediction based on the relationship between the predictor variables and the response variables [1]. Based on the shape of the relationship, the regression analysis can be divided into linear regression and non-linear regression. Linear regression is an approach for modeling the relationship between a dependent variable *y* and one or more explanatory variables (or independent variables) denoted by *X.* Parameter estimation methods which are often used in multiple linear regression is Ordinary Least Squares (OLS). The OLS method minimizes the sum squared of residuals (error). The OLS method require some classical assumptions in order to achieve estimator which is Best Linear Unbiased Estimator (BLUE). The assumptions related to errors that is generated from that model. The assumptions that must be met, namely the normality of error, non-autocorrelation, homoscedasticity, and non-multicollinearity.

Homoscedasticity is one of the important assumptions in the regression analysis, where the variance of the error term is constant otherwise heteroscedasticity. The effect of heteroscedasticity will give much weight to a small subset of data (namely the subset where the error variance is largest) when estimating regression parameters. Restricted Maximum Likelihood (REML) is known as an unbiased parameter estimation method. REML method can be applied to models that have a normal experimental error,

Submitted: 18 Nopember 2016 Reviewed: 15 March 2018 Accepted: 30 November 2018
DOI: http://dx.doi.org/10.18860/ca.v5i3.3777

interrelated and different variance. The used of REML variance component estimation can be done even if the data did not meet the assumptions of analysis of variance [2].

Economics is one of social field that is often use regression analysis to make decisions. One of the developed economic theory is consumption theory. Consumption theory states that any individual who has income, is assumed to set aside their part of revenue after being deducted by consumption [3].

The consumption pattern is significantly affecting the saving's behavior. Indonesian society is known as a consumer society, it could lead to the low motivation of savings. The benefits of savings are degrading consumerist patterns, practicing thrift and as a reserve fund. If the savings increases, not only the public funds will increase but also the investment [4]. Students are part of the community who have an income.

The purpose of this research to estimate multiple regression parameters using REML methods in modeling student's saving at the Faculty of Mathematics and Natural Science, Brawijaya University.

METHODS

The parameters used in this study consisted of a response variable and five predictor variables. The response variable is student's saving (Y). Five predictor variables that affect student savings and used in the research are: X_1 = The student's age (years), X_2 = The amount of income of student's parent (thousand rupiah), X_3 = The amount of student's pocket money (thousand rupiah), X_4 = The amount of student's additional income (thousand rupiah) and X_5 = The amount of student's consumption (thousand rupiah).

Linear regression analysis is a statistical method that is useful to model the relationship between the response variable and predictor variables. The relationships model derived from regression analysis can be used as a description of the phenomenon of data. The regression model can also be used for predicting the values of the response variable. The concept of predicting in the regression analysis can only be done in the data range of the predictor variables used to establish the regression model [5]. The response variable is also called dependent variable and denoted by Y. Predictor variables are called independent variables and denoted by X.

Multiple linear regression model is a model where one response variable is determined as a function of more than one predictor variable (p):

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi} + \varepsilon_i \quad (1)$$

where:

i = 1,2, ⋯ , n
Y_i = response variable
$X_{1i}, X_{2i}, \cdots, X_{pi}$ = predictor variables
$\beta_0, \beta_1, \cdots, \beta_p$ = regression coefficients
ε_i = error
p = number of predictor variables.

Equation (1) has $(p + 1)$ unknown parameters, with $\{x_{1i}, \ldots, x_{pi}, i = 1, \cdots, n\}$ is assumed fix and $\{\varepsilon_i\}$ assumed variables are independent, normal distribution with average 0 and variance σ^2:

Using a matrix of the equation (1) can be denoted:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \dots & x_{pi} \\ 1 & x_{21} & \dots & x_{2k} \\ 1 & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_n \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

or

$$\mathbf{Y} = \mathbf{X\beta} + \boldsymbol{\varepsilon} \quad (2)$$

where

$\mathbf{Y}$ = respons vector of size $(n \times 1)$

$\mathbf{X}$ = predictor matrix size $(n \times (p + 1))$

$\boldsymbol{\beta}$ = regression coefficient measuring ((p + 1) × 1)

$\boldsymbol{\varepsilon}$ = error vector size (n × 1)

The steps of data analysis are as follows:

1. Estimate the parameters by using the OLS.

Ordinary Least Squares method is one of the parameter estimations in regression analysis by minimizing the sum of squared errors. By using OLS, the obtained estimators for the parameter β is $\hat{\beta}$. Based on the model (2), it is obtained:

$$\boldsymbol{\varepsilon} = \mathbf{Y} - \mathbf{X\beta} \quad (3)$$

So that

$$S(\beta) = \sum_{i=1}^{n} \varepsilon_i^2 = \boldsymbol{\varepsilon}^{\boldsymbol{T}}\boldsymbol{\varepsilon} = (\mathbf{Y} - \mathbf{X\beta})^{\mathbf{T}}(\mathbf{Y} - \mathbf{X\beta})$$
$$= \mathbf{Y^T Y} - \boldsymbol{\beta}^{\mathbf{T}}\mathbf{X}^{\mathbf{T}}\mathbf{Y} - \mathbf{Y}^{\mathbf{T}}\mathbf{X\beta} + \boldsymbol{\beta}^{\mathbf{T}}\mathbf{X}^{\mathbf{T}}\mathbf{X\beta}$$
$$= \mathbf{Y^T Y} - 2\boldsymbol{\beta}^{\mathbf{T}}\mathbf{X}^{\mathbf{T}}\mathbf{Y} + \boldsymbol{\beta}^{\mathbf{T}}\mathbf{X}^{\mathbf{T}}\mathbf{X\beta}$$

By using the properties of the inverse matrix, $\boldsymbol{\beta}^{\mathbf{T}}\mathbf{X}^{\mathbf{T}}\mathbf{Y} = \mathbf{Y}^{\mathbf{T}}\mathbf{X\beta}$ is a scalar, then the least squares estimators must meet:

$$\frac{\partial S}{\partial \beta} | \hat{\beta} = -2\mathbf{X}^{\mathbf{T}}\mathbf{Y} + 2\mathbf{X}^{\mathbf{T}}\mathbf{X}\widehat{\boldsymbol{\beta}} = 0$$

be simplified,

$$\mathbf{X}^{\mathbf{T}}\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}^{\mathbf{T}}\mathbf{Y} \quad (4)$$

Multiply the final form of the matrix equation (4) both sides with $(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-\mathbf{1}}$, produces the least squares estimator for β is:

$$(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-\mathbf{1}}\mathbf{X}^{\mathbf{T}}\mathbf{X}\widehat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-\mathbf{1}}\mathbf{X}^{\mathbf{T}}\mathbf{Y}$$
$$\mathbf{I}\widehat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-\mathbf{1}}\mathbf{X}^{\mathbf{T}}\mathbf{Y}$$
$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-\mathbf{1}}\mathbf{X}^{\mathbf{T}}\mathbf{Y} \quad (5)$$

2. Test the classical assumption of multiple linear regression analysis.

The model derived from multiple regression analysis must meet the assumptions of the classical regression analysis. The assumptions include: error normally distributed error, homoscedasticity of error variance, non-autocorrelation and non-multicollinearity. The normality assumption of error is an error value (ε_i) obtained from the regression model should follow the normal distribution. One of the methods to detect normality of error is Shapiro-Wilk [6] .

The hypotheses tested:

H_0 : ε_{ij} is normally distributed

H_1 : ε_{ij} is not normally distributed

If H_0 is true, Shapiro-Wilk test statistics:

$$G = b_n + c_n \ln\left[\frac{T_3 - d_n}{1 - T_3}\right] \sim Z(0,1) \quad (6)$$

where

$$T_3 = \frac{1}{D}\left[\sum_{i=1}^{n} a_i \left(X_{(n-i+1)} - X_i\right)\right]^2$$

$D = \sum_{i=1}^{n} (\varepsilon_i - \bar{\varepsilon})^2$

G Value can be approximated by the normal distribution as the Z value is the value of the coefficient counting. The value of a_i is Shapiro-Wilk's value with certain n. Value b_n, c_n, and $d_n$is the conversion value Shapiro-Wilk statistical approaches a normal distribution for n (many observations), If G value less than the critical value of Z distribution, then it can be decided to accept H_0, which means that the experimental error is normally distributed [7].

One of the assumptions of classical regression model is homoscedasticity [8]. If the variance is not constant, is expressed as heteroscedasticity. One of the methods to detect the presence of heteroscedasticity is by using Glejser test. After getting e_i from regression with OLS method, Glejser suggest regressing the absolute e_i as a response to the predictor variables based on the hypotheses:

H_0 : $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_j^2 = \sigma^2$; $\sigma_j^2 = \sigma^2$

H_1 : At least one j where $\sigma_j^2 \neq \sigma^2$

If H_0true, the test statistic

$$\frac{MS_{regression}}{MS_{error}} \sim F_{(p,(n-(p+1)))} \quad (7)$$

where:

$$MS_{regression} = \mathbf{\beta X'Y} - n\bar{y}^2 \Big/ p$$

$$MS_{error} = \mathbf{Y'Y} - \mathbf{\beta X'Y} \Big/ (n - (p+1))$$

If the test statistic is less than the critical point $F_{\alpha(p,(n-p-1))}$, then it is decided to accept H_0, which means that the error variance is homogeneous [8]

Autocorrelation is the correlation between members of a series of observations which are sorted by time (time series) or space (data cross-section). To detect the presence of autocorrelation, the Durbin Watson's test was used based on:

H_0: $\rho = 0$ (Error are independent)

H_1:: $\rho \neq 0$ (Error are not independent)

Statistical test:

$$d = \frac{\sum_{i=2}^{n} (e_i - e_{i-1})^2}{\sum_{i=1}^{n} {e_i}^2} \quad (8)$$

where:

d : Durbin Watson statistic

e_i : the $i-th$ error value

e_{i-1} : the $(i-1)$ error value

H_0 rejected if $d < dL \; or \; d > 4 - dL$

H_0 acceptable if $dU < d < 4 - dU$

No decision if dL <d <dU or 4-dU <d <4-dL

One of the assumptions that must be met in the establishment of regression model with multiple variables predictor is non-multicollinearity. Multicollinearity is the presence of high linear relationship between the predictor variables. Multicollinearity could be detected by using Variance Inflation Factor (VIF) [9], based on hypotheses:

H_0: No multicollinearity between variables

H_1 : There multicollinearity between variables

$$VIF = \frac{1}{1-R_j^2} \quad (11)$$

$$R_j^2 = \frac{JK_{regresi}}{JK_{total}} = \frac{\mathbf{\beta X'Y} - n\bar{y}^2}{\mathbf{Y'Y} - n\bar{y}^2}$$

where R_j^2 is coefficient of determination of auxiliary regression, regressing between X_j and $(p-1)$ the other predictor variable. Auxiliary regression equation is

$$X_{ji} = \gamma_0 + \sum_{k=1|k\neq j}^{p-1} \gamma_j X_{ki} + u_i; k = j = 1,2,\cdots,p \quad (9)$$

If VIF is less than 10 then H_0 accepted, so the assumption of non-multicollinearity is met.

3. If the assumptions are not met homoscedasticity, then if is followed by suspected regression parameter using REML.

Restricted Maximum Likelihood is an alternative variance estimation parameter derived from the Maximum Likelihood Method (MLM)[2]. The parameters obtain from REML estimators is divided into two parts, namely fixed effects parameter by parameter β and σ^2. As an example of a random sample that has a normal distribution, then:

$Y - \mu = (Y - \bar{Y}) + (\bar{Y} - \mu)$

With the μ=Xβ so:

$$\mathbf{Y} - \mathbf{X\beta} = \left(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right) + \left(\mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{X\beta}\right) = \left(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right) + \mathbf{X}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)$$

Then the likelihood function is:

$$(\mathbf{Y} - \mathbf{X\beta})^{\mathbf{T}}(\mathbf{Y} - \mathbf{X\beta}) = \left[\left(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right) + \mathbf{X}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\right]^{\mathrm{T}}\left[\left(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right) + \mathbf{X}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\right]$$
$$= \left(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)^{\mathrm{T}}\left(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right) - 2\left(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)^{\mathrm{T}}\mathbf{X}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) + \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)$$

Based on these equations obtained

$$2\left(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)^{\mathbf{T}}\mathbf{X}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) = 2\left(\mathbf{X}^{\mathbf{T}}\mathbf{Y} - \mathbf{X}^{\mathbf{T}}\mathbf{X}\widehat{\boldsymbol{\beta}}\right)^{\mathbf{T}}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)$$

by completing $\mathbf{X}^{\mathrm{T}}\mathbf{Y} - \mathbf{X}^{\mathrm{T}}\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{0}$ to (p + 1) parameters, resulted

$$(\mathbf{Y} - \mathbf{X\beta})^{\mathrm{T}}(\mathbf{Y} - \mathbf{X\beta}) = (\mathbf{Y} - \mathbf{Xb})^{\mathrm{T}}\left(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right) + \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \quad (10)$$

Equation (6) is a function of (p + 1) parameters (β). For this function free of the parameter β can be written:

$$\left(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)^{\mathbf{T}}\left(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right) = (\mathbf{Y} - \mathbf{X}(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-\mathbf{1}}\mathbf{X}^{\mathbf{T}}\mathbf{Y})^{\mathbf{T}}(\mathbf{Y} - \mathbf{X}(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-\mathbf{1}}\mathbf{X}^{\mathbf{T}}\mathbf{Y})$$
$$= \mathbf{Y}^{\mathbf{T}}(\mathbf{I} - \mathbf{X}(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-\mathbf{1}}\mathbf{X}^{\mathbf{T}})^{\mathbf{T}}(\mathbf{I} - \mathbf{X}(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-\mathbf{1}}\mathbf{X}^{\mathbf{T}})\mathbf{Y}$$

Since, $\mathbf{I} - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$ is symmetric and idempotent, then:

$$\left(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)^{\mathrm{T}}\left(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right) = \mathbf{Y}^{\mathrm{T}}(\mathbf{I} - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}})\mathbf{Y}$$

The log likelihood of multiple linear regression is:

$$logLikelihood\ \mathbf{Y} = -\tfrac{\mathrm{p+1}}{2}\ln(2\pi) - \tfrac{\mathrm{p+1}}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) - \tfrac{\mathrm{n-1-p}}{2}\ln(2\pi) -$$
$$\tfrac{\mathrm{n-1-p}}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\mathbf{Y}^{\mathrm{T}}(\mathbf{I} - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}})\mathbf{Y}$$

With minimum variance the obtained results estimator σ^2

$$\sigma^2 = \frac{\left(\mathbf{Y}-\mathbf{X}\widehat{\boldsymbol{\beta}}\right)^{\mathrm{T}}\left(\mathbf{Y}-\mathbf{X}\widehat{\boldsymbol{\beta}}\right)}{\mathrm{n-1-p}} = \frac{\mathbf{Y}^{\mathrm{T}}\left(\mathbf{I}-\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\right)\mathbf{Y}}{\mathrm{n-1-p}} \quad (11)$$

RESULTS AND DISCUSION

Estimation of Regression Parameter Using OLS

Linear regression analysis is a statistical method that is useful to model the relationship between response variable and predictor variables[5]. The relationships derived from regression analysis that can be used as a descript of the phenomenon of data. In this study, the model obtained using Ordinary Least Squares (OLS) Method is:

$$\hat{Y}_i = -1855 + 121{,}5X_1 + 0{,}0098X_2 + 0.0524\,X_3 + 0.559\,X_4 - 0.585X_5$$

Testing of The Classical Assumption of Multiple Linear Regression Analysis

Testing the assumption of normality of errors using the Shapiro-Wilk test based on the hypothesis:

H_0 : An error is normally distributed

H_1 : An error is not normally distributed

The p-value of 0.294 for the test and the Shapiro Wilk test statistic G of 0.9784. Based on testing criteria, because the value of p $>$ α and statistic's test G $<$ critical point Z (1,96) , H_0 accepted and concluded that the error distributes normally with a confidence level of 95%.

Detection of autocorrelation using Durbin Watson test[10]. Hypotheses were tested:

H_0 : ρ = 0 (error is independent)

H_1 : $\rho \neq 0$ (error is not independent)

D test statistic of 1.928.

According to the Durbin Watson table then obtained a value of 1.464 dL and dU value of 1.768. The test of the statistic is between the value d and 4-dU dU then H_0 accepted, can be conclude non-autocorrelation assumptions are met.

The Variance Inflation Factor (VIF) is one of the values used to detect the presence of multicollinearity [9]. Hypotheses were tested:

H_0 : Non multicollinierity

H_1 : Multicollinierity

Table 1. VIF value of each predictor variable

Predictor	Value of VIF	Information
X_1	1,04	H_0 accepted
X_2	1,13	H_0 accepted
X_3	1,05	H_0 accepted
X_4	1,15	H_0 accepted
X_5	1,06	H_0 accepted

Table 1 showed VIF value of each predictor variable <10 so H_0 is accepted, non-multicollinearity assumptions are met.

An error variance assumption test homoscedasticity using Glejser Test [8]. After getting e_i from OLS method, Glejser suggest regressing the absolute error and predictor variables. Hypothesis were tested:

H_0 : $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_j^2 = \sigma^2$; $\sigma_j^2 = \sigma^2$

H_1 : At least one j where $\sigma_j^2 \neq \sigma^2$

P value of Glejser test 0,006, then H_0 rejected and concluded that error variance not homogen.

Estimation of Regression Parameter Using REML

Restricted Maximum Likelihood (REML) is an alternative variance estimation parameter derived from the Maximum Likelihood Method (MLM) [2]. The outlinesof the parameters REML estimators into two parts, namely fixed effects parameter by parameter β and σ^2.

The model obtained using Restricted Maximum Likelihood (REML) Method is:

$\hat{Y}_i = -1609 + 112\,X_1 + 0.0088X_2 + 0.0504\,X_3 + 0.4706\,X_4 - 0.636X_5$

The results of testing each parameter using Wald test is shown in Table 4.2

Table 2. The result of partial test use REML method

Parameter	Value	Information
X1	0,001	H_0 rejected
X2	0,615	H_0 accepeted
X3	0,218	H_0 accepted
X4	0,017	H_0 rejected
X5	0,017	H_0 rejected

Based on Table 2, variables that affect student's saving are student's age (years), the amount of student's additional income (thousand rupiah), and the amount of student's consumption (thousand rupiah).

Model Validation

To find out whether the model obtained in the study is in accordance with the actual conditions in the field, the model validation is performed. Then, the comparison between the predicted values from REML method with the actual value of observations is tested using paired t test. The summary result of paired t test is presented in Table 4.3.

Table 3. Paired t Test predicted value from REML method and actual value

Data	N	Average	Standard Deviation
Y Observation	29	408,62	364,76
Y Prediction	29	399,19	184,22
Difference	29	9,43	182,10
p-*value* = 0,904			

The result of paired t test provided in Table 3 decided to accept H_0 because p value is larger than 0.05, lead to conclusion that the model of REML method can be used to predict student's saving.

CONCLUSION

The model of student's saving using OLS method as follows:

$$\hat{Y}_i = -1855 + 121{,}5X_1 + 0{,}0098X_2 + 0.0524\,X_3 + 0.559\,X_4 - 0.585X_5$$

Student's saving is affected significantly by: student's age (X_1), the amount of student's additional income (X_4), and the amount of student's consumption (X_5). REML method can overcome heteroscedasticity error variance and producing unbiased estimator.

The model of student saving using REML method as follows:

$$\hat{Y}_i = -1609 + 112\,X_1 + 0.0088X_2 + 0.0504\,X_3 + 0.4706\,X_4 - 0.636X_5$$

Student's saving is affected significantly by: student's age (X_1), the amount of student's additional income (X_4), and the amount of student's consumption (X_5). Based on the results and the discussion it can be concluded that: REML method can overcome heteroscedasticity error variance and unbiased estimator.

REFERENCES

[1] Chatterjee, S., & Hadi, A. S. (2006). *Regression Analysis by Example Fourth Edition.* New York: John Wiley and Sons, Inc.

[2] O' Neill, M. (2010). *Anova & REML: A Guide to Linier Mixed Models in an Experimental Design Context.* Statsitical Advsory Training Service Pty Ltd.

[3] Browning, M., & Lusardi, A. (1996). Household Saving: Micro Theories and Micro Facts. *Journal of Ecomomic Literature*.

[4] Dupas, P., & Robinson, S. (2009). Savings constraints and microenterprise development: Evidence from a field experiment in Kenya. *National Bureau Research Working Paper*.

[5] Draper, N. R., & Smith, N. R. (1998). *Applied Regression Analysis. Third Edition.* New York: John Wiley and Sons.

[6] Shapiro, S. S., & Wilk, M. B. (1965). An Anlytics of Variance Test for Normality. *Biometrika*.

[7] Razali, N., & Wah, Y. P. (2011). Power Comparisons of Shapiro Wilk, Kolmogorov Smirnov, Lilliefors and Anderson Darling. *Journal of Statistical Modeling and Analytics*.

[8] Gujarati, D. (2004). *Basic Econometrics Fourth Edition.* New York: The McGraw-Hill.

[9] Alaudin, M., & Nghiem, H. S. (2010). Do Instructional Attributes Pose Multicollinierity Problems? Am Empirical Exploration. *Economic Analysis and Policy*.

[10] Gujarati, D., & Porter, D. C. (2012). *Dasar-dasar Ekonometrika (Jilid 2) Terjemahan Raden Carlos Mangunsong.* Jakarta: Salemba Empat.