A BEST MODEL ON MULTIPLE LINEAR REGRESSION

B. Pratikno [#1], I.P. Sulaeman [#2], D. Sopanti [#3] and Supriyono [#4]
[#] First-Third
Department of Mathematics
Faculty of Mathematics and Natural Science
Jenderal Soedirman University, Purwokerto, Indonesia.
*Second Company
[#]Fourth
Open University, UPBJJ Purwokerto, Indonesia
[1] bpratikto@gmail.com
[2] indahpns7@gmail.com
[3] diahsopanti.ds@gmail.com
[4] supriyono@ecampus.ut.ac.id

***Abstract*—We study the selection variables (predictors) on multiple linear regression model (MLRM). The best subset (C_p Mallow) and the stepwise (forward selection and backward elimination) methods are used to identify a best model on the MLRM. A simulation study is conducted using secondary data on *Y* (percentage of poverty line) and four predictors X_1 (jobless), X_2 (population growth rate), X_3 (life expectancy), and X_4 (length of study). The result showed that the best model of the forward selection and backward elimination are similar, X_1 and X_3, are significant, with the model $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_3 x_3$, but the forward selection (with 2 steps) is more efficient than backward elimination (with 3 steps). Unfortunately, the best subset method has a different selection variables, here X_2 is also significant, so the model is $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3$. Finally, we conclude that the best model is $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_3 x_3$. This is due to the both variables X_1 and X_3 are available on the three tests, namely (1) correlation testing (partial test), (2) stepwise method, and (3) best subset methodFor your paper to be published in the conference proceedings, you must use this document as both an instruction set and as a template into which you can type your own text. If your paper does not conform to the required format, you will be asked to fix it.**

Keyword - Best subset, *Cp* , MLRM, and stepwise.

I. INTRODUCTION

Generally, inferences about population parameter can be drawn from sample, and increasing a number of sample will affect significantly to the quality of inferences (Soejoeti, 2010). More detail about this, Bancrof [21] already studied to improve the inferences population using non-sample prior information (NSPI) from trusted sources. Therefore, we need sampling to get the eligible and representative sample (*n)* from population *(N)*, with small error on level of significance (α), $\alpha = 0.01, 0.05$ and or 0.10. Following, Bhatacharya and Johnson [12], Walpole and Myers [19] and Bluman [1], the sample size (*n*) is then formulated by $n \geq \left(Z_{\alpha/2}\sigma \big/ d \right)^2$, where *d* is determined by researcher, $Z_{\alpha/2}$ is from standard normal distribution, and σ^2 is unknown variance. Furthermore, from the pair of data sample $(x_i, y_i), i = 1, 2, \cdots, n$, we then estimate the linier regression model of the multiple regression model (MLRM) as $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$, where x_i is a predictor and y_i is a response. Note that the general model of the MLRM is given as $y_i = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + e_i$. To ensure that the model is significant, we then test the significant model of the $\hat{y}_i$ to be recommended to users. Moreover, the model ($\hat{y}_i$) is generally tested using three criteria, namely (1) the coefficient determination (R^2), (2) analysis of variance (Anova) with *F* test, and or (3) partial (Draper and Smith, [21]). In this case, some assumptions of the regression model are also tested, such as (1) normality test and random error of $e_i \sim N(0, \sigma^2)$, (2) autocorrelation of the error, (3) heteroscedasticity and (4) multicollinearity problem among predictors.

In addition, the paper is focused on the best subset and stepwise methods in getting the best model on the MLRM. Here, we noted many authors have been studied about the selection variables on regression model, such as Draper and Smith [21], Gujarati [7], Gujarati [8], Ghozali [13], Kutner et al.[15], Montgomery [5], Montgomery [6], Bhattacharyya and Johnson [12], Greene [22], Baltagi [2], Baltagi [3], Mendenhall [23], Mendenhall and Sincich [24], Bluman [1], Huggins and Staudte [20], Iriawan and Astuti [16], Graybil and Iyer [10], Walpole and Myers [18], Walpole et al. [19] and Pratikno [4]. Here, Pratikno [4] have been studied testing of the hypothesis tests on regression models using power and size of the hypothesis testing.

In this paper, the introduction is presented in Section 1. The regression model, best subset and stepwise methods are given in Section 2. A simulation is then obtained in Section 3. Section 4 described the conclusion of the research. document is a template.

II. REGRESSION MODEL, BEST SUBSET AND STEPWISE METHODS

2.1. Multiple Regression Model

For an ***n*** pair of observations on ***p*** independent variables ($X_1,...,X_p$) and one dependent variable (Y), (X_{ij}, Y_i), for i=1,2,...,n and j = 1,2,...,p, the multiple regression model (in matrix) is given by

$$Y = X\beta + e \tag{1}$$

Here, $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)'$ is a $(p+1)$ dimensional column vector of unknown regression parameters, $\boldsymbol{Y} = (y_1, ..., y_n)'$ is $(n \times 1)$ vector of response variables, $\boldsymbol{X}$ is a $n \times (p+1)$matrix of know fixed values of the independent variables an $\boldsymbol{e}$ is the error term which is assumed to be identically and independently distributed as $N_n(0, \sigma^2 \boldsymbol{I}_n)$. Here, $\boldsymbol{I_n}$ is the identity matrix of order n and σ^2 is the common variance of the error variables. Following Montgomery [5] and Graybil and Iyer [10], the estimate coefficient regression model in matrix is then given as

$$\hat{\beta} = (X^t X)^{-1} X^t Y \tag{2}$$

Furthermore, we test the hypothesis testing of the, $H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0$. Here, Anova *(F test)* is used, and it is presented in Table 1.

Table 1. Anova of the MLRM with p independent variables

Source of Variance	Degree of freedom (df)	Sum Square (SS)	Mean Square (MS)	*F**
Regression	p	SSR	MSR	$F^* = \frac{MSR}{MSE} > F_{\alpha;p;n-(p+1)}$ W e then reject H_0
Error	$n-(p+1)$	SSE	MSE = s^2	
Total	$n-1$	SST		

Note that *SSE* is sum square error, *SST* is sum square total and *MSE* is mean square error. Following Montgomery [5], we must test the assumptions of the MLRM, such as (1) the **m**ulticollinearity is tested using variance inflaction factor (*VIF*), that is *VIF* > 10, or *F* test is significant but *t* is not. We guarantee that there is no autocorelation of the error term and there is no heteroscedasticity. Moreover, Gujarati [7] showed that if the plot of error versus X_i going to be large (see Figure 1.), then X_i and Y_i must be divided by X_i in order to get the eligible data.

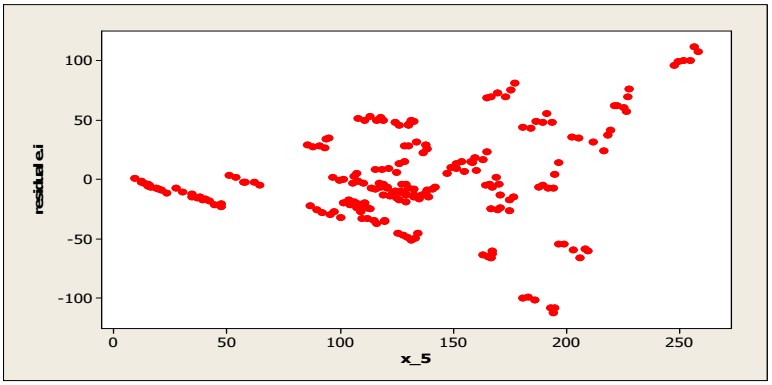

Figure 1. Plot between *error* e_i and weighted X_i

2.2. Stepwise Regression Method

The stepwise regression method generally consists of forward selection and backward elimination (Soejoeti, [25]). Here, we want to choose a small subset from the larger set (large set of candidate predictor variables) so that the resulting regression model is good predictive ability. In this method, we enter and remove predictors until there is no justifiable reason to enter or remove more. First step, we fit each of the one-predictor models, that is, regress y on x_1, regress y on x_2,…, regress y on x_p. The first predictor is the predictor that has the smallest t-test (or high correlation (r) or significant in F test). Similarly, 2nd step, we suppose x_1 was the "best" one predictor, then we fit each of the two-predictor models with x_1 in the model, that is, regress y on (x_1, x_2), regress y on (x_1, x_3),…, and y on (x_1, x_p). Again, the second predictor is the predictor that has the smallest t-test (or high correlation (r) or significant in F test). But, we must consider to remove one of them if the model on the 2nd step, $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$, is not significant. For example, if the $\beta_1 = 0$ has become not significant, remove x_1. The procedure is stopped when adding an additional predictor has no significant ***t*-**test more**.** Following, [20], the first step of the backward elimination procedure is to allow us to fit the full model of the MLRM. We then eliminate one by one using correlation (r) criteria (and or t or F test) in testing hypothesis of the coefficient regression parameters. Here, we need many steps in getting the eligible coefficient regression parameters model of the MLRM.

2.3. Best Subset Method

The best subset method is used to find the eligible predictors (X) in the MLRM model, with $n > p$. Here, we select the subset of predictors that do the best at meeting some objective criterion. Following Draper and Smith [21], we follow several steps to get the eligible predictor in the model, that are:

(1) consider and choose the highest R^2: $R^2 = 1 - \dfrac{SSE}{SST}$,

(2) determine the maximum R^2_{adj}: $R^2_{adj} = 1 - \dfrac{(n-1)SSE}{(n-p)SST}$,

(3) choose the small C_p and C_p is close to p, where $C_p = \dfrac{SSE}{MSE} - n + 2p$, and

(4) finally, we must choose the small s, where s is square root of MSE .

III. A SIMULATION STUDY

Following Subsections 2.1 to 2.3, we then simulate the model of response Y (percentage of poverty line) and four predictors X_1 (jobless), X_2 (population growth rate), X_3 (life expectancy), and X_4 (length of study). The full model of the data simulation and their Pearson correlation (Table 2.) are then given as, respectively,

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_4 = 108 + 0.5x_1 - 0.8x_2 - 1.6x_3 + 0.9x_4, \text{ and}$$

Table 2. Pearson correlation

	X_1	X_2	X_3	X_4
Y	0.8	0.5	-0.8	-0.5.

From Table 2, it is clear that X_1 and X_3 have high correlation, so they are significant and eligible to the model. To make sure that model is really good, we then analysis it using both methods, namely forward and backward selection methods as follow.

The output of the procedure of the stepwise forward selection and backward elimination are presented in Table 3. and Table 4. (Suleman, [14]).

Table 3. Stepwise Regression (2 steps): Y versus X_1, X_2, X_3, X_4 with $\alpha = 0.05$

Step	1	2
Constant	123.7	89.7
X_3	-1.7	-1.2
T-Value	-3.8	-5.2
p-Value	0.005	0.001
X_1		0.37
T-Value		5.2
p-Value		0.001
s	0.63	0.31
R-sq	64.5	92.6
R-sq (adj)	60.0	90.4
Mallow-C_p	25.3	2.5

Table 4. Stepwise Regression (3 steps): Y versus X_1, X_2, X_3, X_4 with $\alpha = 0.05$

Step	1	2	3
Constant	108.0	89.7	89.9
X_1	0.46	0.48	0.39
T-Value	4.0	4.3	5.2
p-Value	0.011	0.005	0.001
X_2	-0.8	-0.60	
T-Value	-1.2	-1.1	
p-Value	0.27	0.31	
X_3	-1.6	-1.2	-1.2
T-Value	-2.6	-5.3	-5.2
p-Value	0.049	0.002	0.001
X_4	0.9		
T-Value	0.7		
p-Value	0.55		
s	0.32	0.30	0.31
R-sq	94.3	93.8	92.6
R-sq (adj)	89.8	90.8	90.4
Mallow-C_p	5.0	3.4	2.5

Table3. and Table 4. showed that the eligible model is similar, namely $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_3 x_3 = 90 + 0.4x_1 - 1.2x_3$. Here, we got two steps on the forward selection model but not in backward elimination model (3 steps). We then conclude that forward selection method is more efficient than backward elimination method.

Moreover, we then analysis the data using the best subsets method using Minitab software, and the result is given in the Table 5. (Suleman, [14]).

Table 5. Best Subsets Procedure

Vars	R-Sq	R-Sq(adj)	Mallows Cp	S	X_1	X_2	X_3	X_4
1	64.5	60.0	25.3	0.633			x	
1	63.8	59.2	25.9	0.639	x			
2	92.6	90.4	2.5	0.310	x		x	
2	86.7	82.8	7.7	0.415	x			x
3	93.8	90.8	3.4	0.304	x	x	x	
3	92.6	88.9	4.5	0.334	x		x	x
4	94.3	89.8	5.0	0.320	x	x	x	x

Here, we also presented the output of the best subset method in the case of poverty line (Y) and X_2 (population growth rate), X_3 (life expectancy), and X_4 (length of study) on different data (see Table 5., Sopanti, [9]).

Table 6. Output Best Subset Model

Vars	R-Sq	R-Sq (adj)	Mallows Cp	S	X_2	X_3	X_4
1	97,9	97,7	4,4	0,2882	x		
1	63,7	60,0	209,6	1,2072		x	
2	98,6	98,3	2,1	0,2460	x		x
2	98,5	98,2	3,0	0,2594	x	x	
3	98,7	98,2	4,0	0,2588	x	x	x

From Table 6., we see that X_2 and X_3 are the best significant predictors to the model of the MLRM. This is due to the C_p close to $p=3$ *(*even it is not the smallest*)*, and the $R^2 = 98.5$ (2nd highest), maximum $R_{adj}{}^2 = 98.2$, $C_p = 3.0$ close to p ($p = 3$) and small $s = 0.25943$. In this case, we note that the variables name are: X_2 is population growth rate, X_3 is life expectancy, and X_4 is length of study. We then conclude that the best model is $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3$.

IV. CONCLUSION

The paper studied the selection variables (predictors) in multiple regression model (MLRM) using best subset and stepwise regression methods. The result showed that the best model of the stepwise (forward and backward) method are similar (X_1 and X_3, are significant, and the model is $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_3 x_3)$, but the forward selection (with 2 steps) is more efficient than backward elimination (with 3 steps). Unfortunately, the best subset method has a different selection variables, where X_2 is also significant, so the model is $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3$. Here, we note that the output of the $R^2 = 93.8$ (2nd highest), maximum $R_{adj}{}^2 = 90.8$, $C_p = 3.4$ close to p (p=4) and $s=0.30445$, are the suitable indicators of the criteria. We therefore (finally) conclude that the best model is, $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_3 x_3$. This is due to the X_1 and X_3 variables are available on the three tests, namely (1) correlation testing (partial test), (2) stepwise method, and (3) best subset method.

ACKNOWLEDGMENT

I thankfully to the colleague for providing me the data

REFERENCES

[1] A. G. Bluman, 1997. Elementary Statistics: A Step by Step Approach. 3rd edition, Mc Graw-Hill.

[2] B. H. Baltagi. 2005. Econometric Analysis of Panel Data 3^{th} edition. England: John Willey and Sons, Ltd.

[3] B. H. Baltagi. 2008. Econometrics 4^{th} edition. Verlag Berlin Heidelberg: Springer.

[4] B. Pratikno. 2012, Tests Of Hypothesis For Linear Regression Models With Non Sample Prior Information, Dissertation, University of Shouthern Queensland, Australia.

[5] D. C. Montgomery. 1996. Regression Analysis. John and Wiley Son, USA

[6] D. C. Montgomery, E. A. Peck and G. G. Vining. 2012. Introduction to Linear Regression Analysis, 5th Edition.Canada: John Wiley and Sons, Inc.

[7] D.N. Gujarati. 2004. Basic Econometrics 4^{th} edition. New York: McGraw-Hill
[8] D.N. Gujarati. 2016. Dasar-Dasar Ekonometrika terjemahan. Jakarta: Erlangga.
[9] D. Sopanti. 2019. Analisis Faktor Yang Berpengaruh Terhadap Persentase Kemiskinan Di Provinsi Jawa Barat Dengan Metode Regresi Linier Berganda, Laporan Kerja Praktek (Unpblished)
[10] F.A. Graybil, and H.K. Iyer. 1992. Regression Analysis. John & Wiley Son, USA.
[11] G.A.F. Seber and J. L. Alan (2003). Linear Regression Analysis, Second Edition. New Jersey: John Wiley dan Sons, Inc.
[12] G.K. Bhattacharyya and R.A. Johnson. 1977. Statistical Concepts and Methods, John and Wiley Son, USA.
[13] I. Ghozali. (2016). Aplikasi Analisis Multivariete Dengan Program IBM SPSS 23, Ed. 8 Cetakan ke VIII. Semarang : Badan Penerbit Universitas Diponegoro.
[14] I.P. Suleman. 2019. Regresi Linier Berganda Dan Aplikasinya Pada Pemodelan Data Kemiskinan Di Kabupaten Tasikmalaya, Laporan Kerja Praktek (Unpblished)
[15] M.H. Kutner, C.J. Nachtsheim, J. Neter and W. Li. 2005. Aplied Linear Statistical Models, 5th Edition. New York: McGraw-Hill Companies, Inc.
[16] N. Iriawan and S.P. Astuti. 2006. Mengolah Data Statistik dengan Mudah Menggunakan Minitab 14. Yogyakarta: ANDI.
[17] N. R. Draper and H. Smith. 1998. Applied Regression Analysis. Jhon Wiley & Sons, USA
[18] R.E. Walpole and R.H. Myers. 1997. Probability and Statistics for Engineers and Scientist, Mc. Milan, USA.
[19] R.E. Walpole, R.H. Myers, S.L. Myers and K. Ye. 2012. Probability and Statistics for Engineers and Scientist. Pearson, USA.
[20] R. Huggins and R. Staudte. 2002. Biostatistics. La Trobe University, Melbourne.
[21] T. A. Bancrof. 1944. On biases in estimation due to the use of the preliminary tests of significance, Annals of Mathematical Statistics, 15(1944), 190-204.
[22] W.H. Greene. 2004. Econometric Analysis. Prentice Hall: New Jersey
[23] W. Mendenhall. 2012. Regression Analysis (7th ed). Amerika Serikat : University of Florida.
[24] W. Medenhall and T. Sincich. 2004. A Second Course in Statistic: Regression Analisis, Edisi ketujuh. New York: Prentice Hall.
[25] Z. Soejoeti. 2010. Metode Statisik I, Universitas Terbuka, Jakarta.