VARIABLE SELECTION IN ANALYZING LIFE INFANT BIRTH IN INDONESIA USING GROUP LASSO AND GROUP SCAD

Ita Wulandari[1], Khairil Anwar Notodiputro[2], Bagus Sartono[3]
{ita.wulandari@stis.ac.id[1], khairil@apps.ipb.ac.id[2],bagusco@apps.ipb.ac.id[3]}

Statistics Departement, Polytechnic of Statistic STIS, Jl. Otista No. 64C, Jakarta 13330,Indonesia[1],
Statistics Departement, Bogor Agricultural University, Jl.Raya Dramaga Bogor, 16680, Indonesia[2,3]

Abstract. Regression analysis often requires a selection of explanatory variables X_1, X_2, ... X_p so shrinkage coefficients can occur that can facilitate the interpretation of the regression equation obtained. In this context, the explanatory variable often has a grouping structure so that a more relevant problem is how to choose groups rather than individuals. Group LASSO and group SCAD are techniques for selecting groups of variables which in many works of literature appear to have advantages over LASSO. In this study, the percentage of live born children in the province of Bali, East Nusa Tenggara and other Indonesia provinces were analyzed and linked to the explanatory variables using the group LASSO and group SCAD methods. The classification of available explanatory variables is grouped based on the theory and results of previous studies. The results show that the best model is the group SCAD method with the smallest AIC, BIC and GCV values. Factors included in the model for Bali province are demographic factors, women's status, and autonomy and the economy. For East Nusa Tenggara province the factors that enter the model are demographics and economics, while generally for Indonesia the factors that are included in the model are demography, women's status, and autonomy and family planning.

Keywords: LASSO, group selection, group LASSO, group SCAD, life infant birth

1 Introduction

In regression modeling, some explanatory variables are often considered as one group. The grouping structure appears in many statistical problem modeling. This method proposes to select the variables that concern to the grouping structure. Grouping arises for various reasons and different modeling objectives. Grouping also utilizes scientifically meaningful prior knowledge[17]. For example in the analysis of gene expression, the same biological pathway genes are considered as one group. Considerations in grouping variables depend on the context of the research. For example, if variables grouping is more meaningful in research, the groupingof the variables is better than choosing individual variables vice versa.

Grouping in the modeling process must consider the benefits obtained like model accuracy and ease of interpretation. These advantages are important to set the case for high-dimensional data of which sparsity and variables selection roles are significant for estimation accuracy.

ICSA 2019, August 02-03, Bogor, Indonesia

DOI 10.4108/eai.2-8-2019.2290485

Many researchers have considered group selection problems in various statistical modeling problems. Bakin [2] proposed the Group LASSO and algorithm computation and was further developed by Yuan and Lin [30]. Other researchers also studied the theoretical nature of the Group LASSO that showed the Group LASSO was a consistent group selection method in a random design model for fixed *p*. Nardi and Rinaldo [23] recommends Group LASSO as its selection consistency in unrepresented conditions, false prediction, and estimation with limited eigenvalue.

Commonly, Group LASSO does not have a consistent characteristic in group selection. It is due to the penalties of the Group LASSO that do not change according to the group coefficients size and lead to biased estimation. However, Yuan and Lin [30] stated the Group LASSO penalty function which is between penalties L_1 and L_2, an extension of the LASSO method, enables reduction of inconsistencies and asymptotic bias.

Fan and Li [10] stated that the estimator produced by the penalty function must have three characteristics which are unbiasedness, sparsity, and continuity. Fan and Li [10] also proposed the SCAD method that has those three characteristics. Fan and Peng [11] and Kim et al. [19] studied the three characteristics in the SCAD method in high dimensional data. Wang et al. and Guo et al. [13] proposed a Group SCAD approach for group selection with time-varying coefficients. Zeng and Xie [31] studied the strength of Group SCAD combined with L_2 in-group effect selection.

The LASSO and SCAD Group methods have been widely applied in various fields, such as health, economics, social, weather and in various fields. In addition, both methods have been applied in the case of simulation data.

Population growth is the main problem facing developing countries, one of which is Indonesia. This is especially due to the high fertility rate and if left unchecked it will affect various aspects of life. Increased population growth at this time is due to a decrease in mortality rates earlier and faster than fertility. In 2017, the highest Total Fertility Rate (TFR) occurred in Bali Province at 2.1 while the lowest in East Nusa Tenggara (NTT) Province was 3.4. Considering these conditions, the case study in this study aims to analyze the factors that affect fertility in both provinces at once in Indonesia by using the percentage of live born children as a response variable. The method applied for modeling live born children in Bali Province, NTT Province and Indonesia is the Group LASSO and Group SCAD, which then select the best model from both methods using values Akaike Information Criterion (AIC), the minimum Bayesian Information Criterion (BIC) and Generalized Cross Validation (GCV).

1.1 Least Absolute Shrinkage and Selection Operator (LASSO)

Least Absolute Shrinkage and Selection Operator (LASSO)proposed by Tibshirani [28] is a regression analysis method that performs variable selection and regularization to improve the accuracy of the prediction and interpretation of the resulting statistical model. The basic idea of this method comes from Non-negative Garrotte[5] and began to become a concern after developing the Least Angle Regression (LAR) algorithm by Efron, Hastie, Johnstone, and Tibshirani [9].

In the modeling data set, for example $(x_1, y_1), \dots, (x_n, y_n)$ are n pairs of explanatory variables, y_i is the response variable with $i = 1,2,\dots,n$ and $\hat{\beta}^{lasso}$ is LASSO coefficient estimators that minimize the number of equal squares with constraints L_1[28], then the estimator of the coefficient on LASSO is defined as follows:

$$\hat{\beta}^{lasso} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2 \tag{1}$$

with constraints:

$$\sum_{j=1}^{p} |\beta_j| \leq t, t \geq 0$$

The estimated coefficient using LASSO becomes:

$$\hat{\beta}^{lasso} = \operatorname{argmin}_{\beta} \left[\sum_{j=1}^{N} (y_i - x_i'\beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right] \tag{2}$$

In equation (2) LASSO regression coefficients cannot be obtained in a closed form as in the Smallest Squares Method or Regression Ridge, but must be in the form of squares [28]. The LASSO coefficient is controlled by the parameter λ which is set by the constraint $\sum_{j=1}^{p} |\beta_j| \leq t$. The small value of t causes a number of regression coefficients to be zero[14] so that the variables that have a large influence on the selected model and selected small influential variables are zero. This results in a smaller variety of estimators and efficient models.

1.2 Group LASSO

Group LASSO is a generalization of LASSO to select group variables [30]. The grouping of variables aims to facilitate the selection of variables that have similar characteristics. This method selects large group variables with a number that is not the same and gives better results than the LASSO method [16], [21].

In linear regression models with p predictors divided into groups and not overlapping, the model can be written as follows:

$$y = \sum_{j=1}^{J} X_j \beta_j + \varepsilon \tag{3}$$

With y being the vector of the response variable size $n \times 1$, X_j is the predictor matrix variable of d_j sized $n \times d_j$ in the jth group, $\beta_j = \left(\beta_{j1}, \beta_{j2}, \ldots, \beta_{jd_j}\right)' \in \mathbb{R}^{d_j}$ are vectors of the regression coefficients in the jth group sized $d_j \times 1$ dan Ɛ is the error vector. The Group LASSO estimation is defined as follows:

$$\frac{1}{2} \left\| \boldsymbol{y} - \sum_{j=1}^{J} \boldsymbol{X_j} \beta_j \right\|^2 + \lambda \sum_{j=1}^{J} \left(\beta_j^T K \beta_{j_j} \right)^{1/2} \tag{4}$$

With $\lambda \geq 0$ is the tuning parameter and $K_1, \ldots, K_j$ is a positive definite matrix of sized $p_j \times p_j$. Group LASSO estimators are consistently asymptotic even when the complexity of the model increases with an increasing sample size [23].

1.3 Group Smoothly Clipped Absolute Deviation (SCAD)

Fan and Li [10] state that the estimator produced by the penalty function must have three properties, namely unbiasedness, and continuity, then propose the SCAD method which has three traits. Furthermore, Fan and Peng [11], Kim et al. [19] studied these three properties in the SCAD method with high dimensional data. Next, Wang et al. [24] propose group SCAD to select groups with coefficients that vary in time. Furthermore, Zeng and Xie [31] used SCAD with L_2 penalties in selecting groups.

By considering linear regression models which have variable predictors as many as d divided into p groups and not overlapping, the equations are as follows:

$$Y = \sum_{j=1}^{p} X_j \beta_j^* + \varepsilon \qquad (5)$$

With Y being the vector of the response variable sized $n \times d_j$ in the jth group, $\sum_{j=1}^{p} d_j = d$. $\beta_j^* = \left(\beta_{j1}^*, \dots, \beta_{jd_j}^*\right)^T \in R^{d_j}$ is a regression coefficient vector in the jth group, ε is an error vector. Suppose $X = \left(X_1^T, \dots, X_p^T\right)$ with $X_j = \left(X_{j1}^T, \dots, X_{jd_j}^T\right)^T$, then the Group SCAD estimation is as follows:

$$Q(\beta) = \frac{1}{2n} \|Y - X\beta\|_2^2 + \sum_{j=1}^{p} p_{\lambda_n}\left(\|\beta_j\|_2\right) \qquad (6)$$

with $p_\lambda^{'}(t) = \lambda \left\{ I(t \leq \lambda) + \frac{(a\lambda - t)_+}{(a-1)\lambda} I(t > \lambda) \right\}$ for $a > 2$ and $t > 0$, with λ is the tuning parameter. In equation (6) it can be seen that $Q(\beta)$ consists of two parts, namely square loss function and Group SCAD penalty. Group SCAD penalties are midway between L_1 penalties and SCAD penalties which can easily select groups.

1.4 Cross-Validation

Cross-validation divides data into two parts, namely training data and testing data. Training data is used to determine the value or to compile a model, while testing data is used to test goodness. The cross-validation values obtained are predictors of prediction errors [12]. One method of cross-validation type is k-fold [8]. The cross-validation that should be used is 5-fold and 10-fold cross validation because it produces a Cross Validation (CV) value with a high bias but low variability [4]. In this study, 10-fold cross validation was used to get all three models of children born alive.

2 Materials and Methods

This study used secondary data derived from Indonesia's 2017 Demographic and Health Survey (IDHS 2017). IDHS data is part of an international Demographic and Health Survey (DHS) program designed to collect fertility, family planning, and maternal and child health data. In line with the research objectives, the samples used in SDKI were 500 people for Bali, 1,327 people for East Nusa Tenggara (NTT), and 31,940 for Indonesia with the Fertile Age Women (WUS) as the observation unit. The response variable used was the logarithm of the percentage of live born children calculated by dividing the number of live born childrenwith the total number of children born to a woman. The result of the calculation transformed in the form of logarithms. The explanatory variables used are grouped based on the theories of Davis and Black [7], Rusli [26], and Upadhyay and Hindin [29]. The explanatory variables in this study were 28 which were then formed into 4 groups. Group 1: Demographic factors consist of 6 variables, Group 2: Factors of female status and autonomy consist of 6 variables, Group 3: Economic factors consisting of 9 variables and Group 4: Family Planning Factors consist of 7 variables.

The steps to get the best model on the Group LASSO and Group SCAD methods for cases of live-born children in Bali, NTT, and Indonesia were as follows:

1. Grouping explanatory variables based on theory or information in previous studies.
2. Standardizing response variables and explanatory variables so that they have zero average and a variety of one.

3. Initiating the variable coefficient value with zero.
4. Finding the optimum value of λ with λ is the controller parameter in which value is selected from the minimum CVE. In the group SCAD method, the value of a is 4.
5. Performing an iterative process to obtain a convergent $\hat{\beta}$ value.
6. Choosing the best model with minimum AIC, BIC and GCV values.

3 Results and Discussion

3.1 Stages of Forming a Model Using Group LASSO and Group SCAD

Grouping variables can utilize previous knowledge or theories that are scientifically meaningful [17]. In this study 28 predictor variables used were grouped into four groups based on previous theories. Davis and Black [7] state that fertility is influenced by social, economic and cultural factors. Mantra [22] states that factors that can affect fertility are demographic factors: composition of age, marital status, age of first marriage, and non-demographic factors: economy, women's status, urbanization, and industrialization. Rusli [26] in his research stated that fertility is influenced by social, economic, demographic, biological and contraceptive use factors. Other studies underlying the grouping of variables is Upadhyay and Hindin [29]. The four groups formed are demographic factors, factors of women's status and autonomy, economic factors and family planning factors.

The group LASSO and the group SCAD want all coefficients in one group to be nonzero (or zero) simultaneously. When in our study we have qualitative variables among predictor variables, then based on Yuan and Lin [30] these variables are formed into dummy variables which will then be selected or not simultaneously in the group selection process.

As emphasized in Simon and Tibshirani [27], the standardization process is just as important for the group LASSO as LASSO was found first. For this reason, standardization of the LASSO and group SCAD was carried out in the response variable and predictor variable before modeling.

3.2 Selection of The Lambda Values

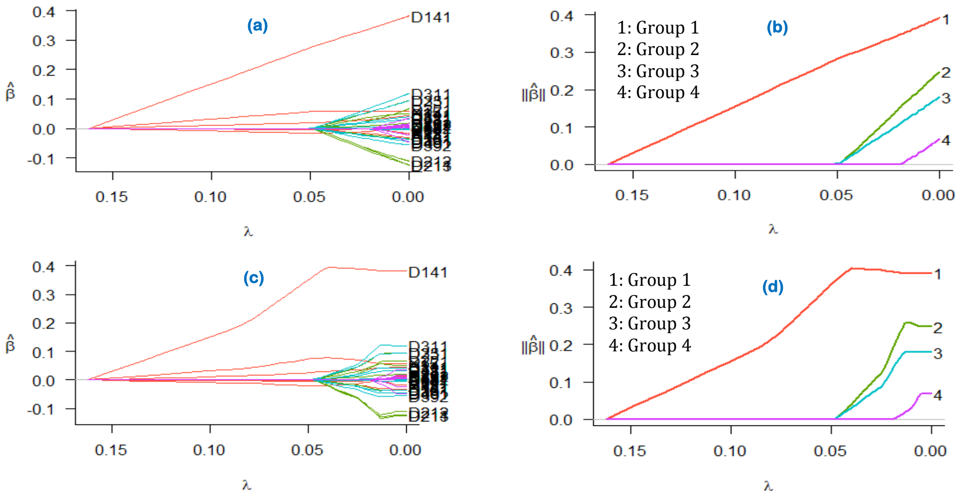
Estimators of group LASSO and group SCAD coefficients are obtained from the optimal selection of λ. his value is obtained from the minimum CVE (Cross Validation Error). Table 1 shows the optimal λ selected for the group LASSO and group SCAD methods for each model. It is seen that in the NTT and Indonesia models produce lambda and CVE values that tend to be almost the same in each of the LASSO and SCAD group methods. In the Bali model, the selected lambda value for the group LASSO method is 0.0441 with the CVE value (0.92), while the SCAD group method is 0.0304 with the CVE value (0.91). The group that enters the model is seen in column 5, while for the number of variables that enter the model in column 6. For the Bali and NTT models groups, 1,2 and 3 belong to the LASSO and SCAD group models, except for the NTT model for the group SCAD selected only groups 1 and 3. The Lasso and SCAD group models for Indonesia’s selected groups are groups 1, 2 and 4.

Table 1. The Predictor Variables Selection Result and Group of Group LASSO and Group SCAD in Bali, NTT, and Indonesia Model

Model	Method	Lambda Selected	CVE	Significant Group	Number of Variables
Bali	*Group LASSO*	0.04410	0.92	g1,g2,g3	24
	Group SCAD	0.03039	0.91	g1,g2,g3	24
NTT	*Group LASSO*	0.02367	0.94	g1,g2,g3	24
	Group SCAD	0.02367	0.94	g1 ,g3	16
Indonesia	*Group LASSO*	0.00674	0.90	g1,g2,g4	23
	Group SCAD	0.00675	0.90	g1,g2,g4	23

3.3 Steps of Group Selection

Estimation of the group LASSO and group SCAD coefficients is done in stages by setting the initial coefficients all to zero. In the pictures (a) and (c) shows the variable predictors selected by the group LASSO and group SCAD methods based on the group. While images (b) and (d) show the stages of the group that entered the model.



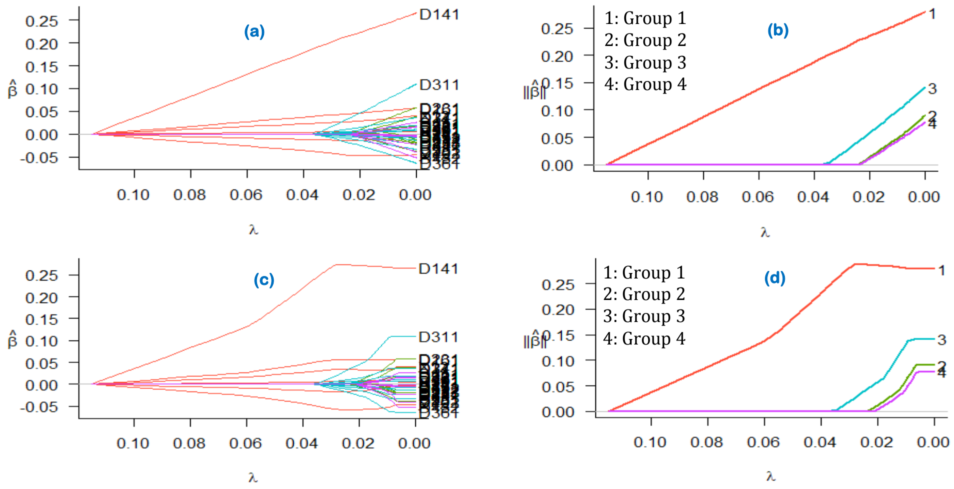
Description: (a) and (b) : LASSO Group; (c) and (d) : SCAD Group

Figure 1. Selection Paths for Predictors Variables and Groups for Bali Model

From Figure 1 for the group LASSO and group SCAD models, group 1 is the first group to be followed by group 2, group 3 and the last is group 4. Of the four groups, group 4 is selected from the model because it has a zero coefficient, p. this can be seen in Figures 3 (b) and 3 (d). Based on this, it can be concluded that demographic factors (group 1) consisting of variables of residence, wife's age, age at first marriage, history of pregnancy, age of husband,

and marital status have a greater influence than female status and autonomy factors (group 2) and economic factors (group 3), while family planning factors (group 4) did not affect the model of childbirth in Bali Province. This condition is possible because there are several opinions and research that state that the family planning program with two children is considered to be incompatible with Balinese customs, culture, and culture. The family planning program is considered to have broken the generation of Bali which has been around for a long time, namely the breakup of generations of Nyoman and Ketut. The results of this study are supported previous studies thatstate that although family planning programs have been introduced throughout the province, success rates differ in each region. This depends on changes in the social, economic and political situation and policies in the area.

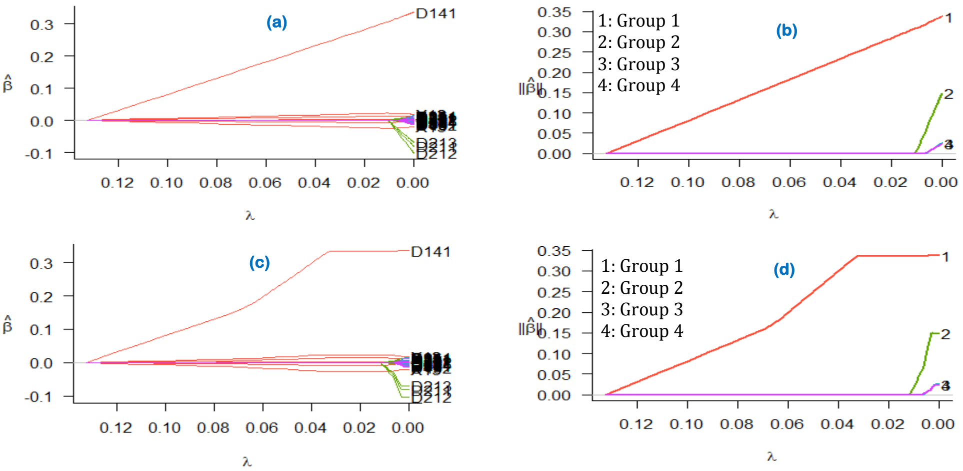
The Selection of groups in both methods for the NTT model obtained different results. The sequences of groups that enter the model in the group LASSO method are groups 1, 3, and 2. In the group of group SCAD method, 2 and 4 do not enter the model. Seen in Figure 2 (d) the lines for the two groups coincide in a straight line at zero, which means that all the coefficients in the group are zero. In this case, it can be seen that the group SCAD method is more stringent in selecting groups compared to the group LASSO method. The same thing happened in NTT province, the family planning factor did not affect the model. Based on previous research, in addition to the level of education, culture greatly influences the number of children born. This is because there is a culture that boys are the successors of the fam (family), besides that they want girls to get belies (mahr), so that if they only have sons, they still want to get a daughter.



Description: (a) and (b) : *LASSO Group*; (c) and (d) : *SCAD Group*

Figure 2. Selection Paths for Predictors Variables and Groups for NTT Model

The model of the percentage of live born children in Indonesia is influenced by 3 factors, namely demographic factors, factors of women's status and autonomy and family planning factors. This can be seen from Figures 3 (b) and 3 (d) where groups 1, 2 and 4 enter the model, while group 3 has zero coefficients so that they are removed from the model.



Description: (a) and (b) : *LASSO Group*; (c) and (d) : *SCAD Group*

Figure 3. Selection Paths for Predictors Variables and Groups for the Indonesia Model

3.4 Selection of the Best Model

The best stages of model selection are done by determining the minimum AIC, BIC and GCV values for the three models of live born children in Bali, NTT and Indonesia. In Table 2 it can be seen that the best method for the model of childbirth in Indonesia is the SCAD group method with the values of AIC, BIC, and GCV in a row of 87,086,710; 87,184,620 and 2,728. The selection of the best model in the other two models using the values of AIC, BIC, and GCV looks a little confusing. This is because when viewed from the AIC value, the SCAD group is the best method for the Bali and NTT models, but when viewed from the BIC and GCV values, the opposite results are obtained on the two models. Therefore, in this study, we will see a plot of all AIC, BIC and GCV values obtained based on the results of the 10-fold cross validation used in this study.

Table 2. The Best Model Criteria for the Group LASSO and Group SCAD on the Models of Bali, NTT and Indonesia

Model	Method	Best Model Criteria		
		AIC	BIC	GCV
Bali	*Group LASSO*	1,350.254	**1,385.346**	**2.758**
	Group SCAD	**1,345.871**	1,408.124	2.795
NTT	*Group LASSO*	3,669.953	**3,722.783**	**2.793**
	Group SCAD	**3,667.557**	3,726.040	2.794
Indonesia	*Group LASSO*	87,098.220	87,190.810	2.728
	Group SCAD	**87,086.710**	**87,184.620**	**2.728**

Description: Bold numbers are the smallest value

The plot of the AIC, BIC and GCV values for the group LASSO and group SCAD methods can be seen in Figure 4. It can be seen that in the nine plots presented in Figure 4, the AIC, BIC and GCV values in the group SCAD tend to be below the group LASSO especially in the area around the value CVE is the minimum, in other words the values produced in the area tend to be smaller than the group LASSO method. The greater AIC, BIC and GCV values for the group SCAD method occur when the lambda value is between 0.00 to 0.02. The values of AIC, BIC, and GCV tend to be the same for the two methods in the area after the minimum point and increasingly increase as the lambda value increases. So based on the 2017 IDHS data, the SCAD group method is a better method than the LASSO group method.

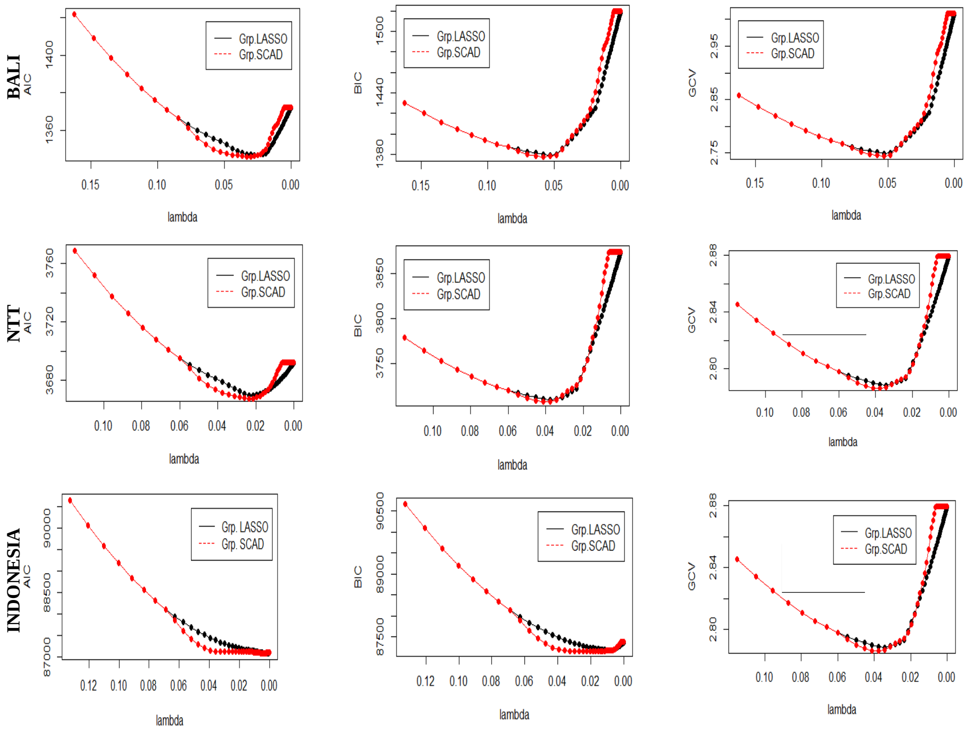


Figure 4. Plots of AIC, BIC and GCV Values in the Group LASSO and Group SCAD for the Bali, NTT, and Indonesia Models

3.5 Born Children Models in Bali Province, East Nusa Tenggara Province, and Indonesia

The selection of optimum lambda in the group LASSO and group SCAD causes the parameter coefficient estimation to shrink so that groups that are important or influential on the model are still included in the model, while the less important groups will be depreciated to zero and selected from the model so the model becomes more efficient. In the case of live born childrenin the province of Bali, NTT province and Indonesia, the group SCAD is a method that can be said to be better than the group LASSO. Estimates of the parameter coefficients in the group SCAD model for the percentage of live-born children in the province of Bali, NTT province and Indonesia are presented in Table 3.

Total Fertility Rate (TFR) is a measuring tool to see fertility rates in a region. The NTR province's TFR is always above the national figure from 1970 and is the highest national figure since 2010. Contrary to this, the province of Bali is always below the national figure and includes provinces with low TFR.

Table 3. The estimator of the Group SCAD Parameter Coefficient for Models of Bali, NTT and Indonesia

Group	Variables	Bali Model	NTT Model	Indonesia Model
Group 1: Demographic factor	*(Intercept)*	3.62×10^{-16}	-9.40×10^{-16}	3.10×10^{-15}
	D_{111}	0.033663	-0.016351	-0.005961
	X_{12}	-0.004746	0.034200	0.022138
	X_{13}	-0.015446	0.007032	-0.024876
	D_{141}	0.391968	0.272513	0.333590
	X_{15}	0.073522	-0.058015	0.013678
	D_{161}	-0.013840	0.056063	0.002410
Group 2: Factors of female status and autonomy	D_{211}	-0.050092	0.000000	-0.033678
	D_{212}	-0.048125	0.000000	-0.042455
	D_{213}	-0.046302	0.000000	-0.028570
	D_{221}	0.013984	0.000000	0.003784
	D_{231}	-0.013381	0.000000	-0.001583
	D_{241}	0.021934	0.000000	-0.002317
	D_{251}	0.032817	0.000000	0.005376
	D_{261}	0.007112	0.000000	0.003278
Group 3: Economic factors	D_{311}	0.040201	0.033604	0.000000
	D_{321}	0.013879	0.006532	0.000000
	D_{331}	0.034852	-0.002275	0.000000
	D_{341}	-0.000960	-0.006090	0.000000
	D_{351}	0.002193	-0.010992	0.000000
	D_{361}	-0.016926	-0.019065	0.000000
	D_{371}	-0.001896	0.010762	0.000000
	D_{381}	0.011804	0.003396	0.000000
	D_{391}	-0.018304	0.002527	0.000000
	D_{392}	-0.026638	-0.003269	0.000000

Group	Variables	Bali Model	NTT Model	Indonesia Model
Group 4: **Family Planning Factors**	D_{411}	0.000000	0.000000	0.000278
	D_{421}	0.000000	0.000000	0.000031
	D_{422}	0.000000	0.000000	0.000031
	D_{431}	0.000000	0.000000	-0.000036
	D_{432}	0.000000	0.000000	-0.000344
	D_{441}	0.000000	0.000000	-0.000168
	D_{451}	0.000000	0.000000	-0.000007
	D_{461}	0.000000	0.000000	-0.000333
	D_{471}	0.000000	0.000000	0.000106

The demographic factor consists of six predictor variables. In general, from the three models, it can be said that demographic factors have a positive influence on children born alive. This can be seen from the positive sign that tends to be more on the coefficient. In the Balinese model, for example, the area of residence, pregnancy history, and age of the husband have a positive influence, which means that women of childbearing age who live in urban areas do not have pregnancy complications and increase their age, more children will be born alive. The history of pregnancy appears to be the most influential variable in the model because it has the greatest coefficient value. This is in line with researchAbdullah et al [1] which states that the main cause of infant mortality in the first week of life is a complication of pregnancy and childbirth.

The second group also had a positive effect on the models of living children in the provinces of Bali and Indonesia. The better the status and autonomy of women, the more the percentage of live born childrenincreases. One of the things that need to be studied in this study is that education has a negative effect on the model, which means that the higher the education of women will reduce the percentage of live born childrenwho have.

Nine variables are used in group 3, some of which have a positive influence on the model. Four variables have a positive effect on the Bali model, while for the NTT model five variables have a positive effect. Based on this, the researcher concludes that economic factors have a positive influence on both models.

The use of contraceptives, the desire to get more children and family planning resources from magazines in group 4 have a positive influence on the model of live born childrenin Indonesia. In other words, women of childbearing age who use contraceptives, tend to not want many children, and family planning sources from magazines will increase the percentage of live born childrenthey have. The four variables that have a negative influence on the model are the desire of the husband to get children, the ideal number of children, the source of family planning from Radio and Television.

5 Conclusion

The formation of groups in this study is based on theory and previous research. Four groups were formed consisting of 28 predictor variables. The first group as demographic factors, the second group as a factor of women's status and autonomy, the third group as an economic factor and the last is the Family Planning (KB) group.

The method applied to the case of the percentage of live-born children in the province of Bali, East Nusa Tenggara Province and Indonesia using the 2017 IDHS data is the group LASSO and group SCAD with 10 fold-cross validation. Based on the minimum AIC, BIC and GCV values, the group SCAD can be said to be better than the group LASSO. The factors that influence the percentage of live born childrenin the province of Bali are demographic factors, factors of women's status and autonomy and economic factors. For models in NTT province, only two factors are influential, namely demographic factors and economic factors. While the factors that influence the percentage of live born childrenin Indonesia are demographic factors, factors of women's status and autonomy and family planning factors.

References

[1] Abdullah.Z., Andin., Naiem, F.M., and Mahmud.U.N.: Faktor Resiko Kematian Neonatal Dini di Rumah Sakit Bersalin. Jurnal Kesehatan Masyarakat Nasional. Vol. 6, pp. 283-288 (2012)

[2] Bakin, S.: Adaptive Regression and Model Selection in Data Mining Problems. Australian National University, Canberra (1999)

[3] Balzere, Melanie., Loubes, J.M., and Gamboa,F.: Group Lasso for Generalized Linear Models in High Dimension. HAL archives-ouvertes. (2014)

[4] Bien, J, Taylor, J., and Tibshirani, R.: A Lasso for Hierarchical Interactions. The Annals of Statistics. Vol. 41.No.3, pp.1111-1141 (2013)

[5] Breiman, L.: Better Subset Regression Using The Nonnegative Garrote. Technometrics, Vol. 37, pp. 373-384 (1995)

[6] Chen, H., and Xiang, Y.: The Study of Credit Scoring Model Based on Group LASSO. Procedia Computer Science. Elsevier. Vol.122, pp. 677-684. (2017)

[7] Davis, K., and Blake, J.: Socail Structure and Fertility: An Analytic Framework . Economic Development and Cultural Change. Vol.4. No.. 3, pp 211-235. (1956)

[8] Efron, B., and Tibshirani, R.J.: An Introduction to The Bootstrap. Monographs on Statistics and Applied Probability 57. Springer US (1993)

[9] Efron B, Hastie T, Johnstone I, and Tibshirani R.: Least Angle Regression. The Annals of Statistics. Vol. 32. No. 2, pp. 407-451. (2004)

[10] Fan, J., and Li, R.: Variable Selection via nonconcave Penalized Likelihood and its Oeacle Properties. Journal of the American Statistical Association December 2001, Vol.96, No.456, Theory and Methods (2001)

[11] Fan, J., and Peng, H.: Nonconcave Penalized Likelihood with a Diverging Number of Parameters. Ann.Statist.Vo.32.pp 928-961 (2004)

[12] Friedman,J., Hastie, T., Hofling, H., and Tibshirani, R.: Pathwise Coordinate Optimization. Ann.Appl.Stat. pp. 302-332 (2007)

[13] Guo, X., Zhang, H., Wang, Y., and Wu, JL.: Model Selection and Estimation in High Dimensional Regression Models With Group SCAD. Statistics and Probability Letters. Elsevier. Vol. 103, pp. 86-92. (2015)

[14] Hastie, T., Tibshirani, R., and Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Edisi Kedua. Springer : New York. (2008)

[15] Hastie,T., Tibshirani, R., and Walnwrright, M.: Statistical Learning with Sparcity. The LASSO and Generalizations. CRC Press. Taylor and Francis Group. (2016)

[16] Huang J., and Zhang T.: The Benefit of Group Sparsity. Ann Statist.Vol. 38, pp. 5277-5286. (2009)

[17] Huang, J., Breheny, P., and Ma, S.: A Selective Review of Group Selection in High-Dimensional Models. Statistical Scince. Vol.27.No.4, pp. 481-499 (2012)

[18] James, G., Witten D., Hastie, T., and Tibshirani, R.: An Introduction to Statistical Learning with Applications in R. Springer : New York (2013)

[19] Kim, Y., Choi, H., and Oh, H.S.: Smoothy Clipped Absolute Deviation on High Dimension. Journal of The American Statistical Association. Vol.103.No. 484 (2008)

[20] Lim, M.: The Group LASSO : Two Novel Applications. Disertation. Standford University (2013)

[21] Lounici, K., Pontil,M., Van, G.S., and Tsybakov, AB.: Oracle Inequalities and Optimal Inference Under Group Sparsity. Ann Statist. Vol. 39, pp. 2164-2204 (2011)

[22] Mantra, I.B.: Demografi Umum. Yogyakarta.Pustaka Pelajar (2000)

[23] Nardi, Y., and Rinaldo, A.: One the Asymtotic Properties of The Group LASSO Estimator for Linear Models. Electronic Journal of Statistics. ISSN. pp. 1935-7524 (2008)

[24] Wang, L., Chen, G., and Li, H.: Group SCAD Regression Analysis for Microarray Time Course Gene Expression Data. Bioinformatics. Vol 23. No.12, pp 1486-1497 (2007)

[25] Wei. F., and Huang, J. : Consisten Group Selection in High-Dimensional Linear Regression. Bernaulli. pp. 1369-1384 (2010)

[26] Rusli.: Pengantar Ilmu Kependudukan. Jakarta.LP3S (1996)

[27] Simon, N., and Tibshirani, R.: Standarization and The Group Lasso Penalty. HHS Public access. Statistica Sinica (2012)

[28] Tibshirani R.: Regression Shrinkage and Selection via the LASSO. Journal of the Royal Statistical Society. Series B. Vol. 58. No. 1, pp. 267-288 (1996)

[29] Upadhyay, U.D., and Hindin M.J.: Do Higher Status and More Autonomous Women Have Longer Birth Intervals. Results from Cebu, Philippines. Soc. Sci. Med. Vol. 60, pp. 2641–2655. (2005)

[30] Yuan, M., and Lin, Y.: Model Selection and Estimation in Regression with Grouped Variables. Journal of the Royal Statistical Society Series B Vol. 68. No. 1, pp. 49-67 (2006)

[31] Zeng, L., and Xie, J.: Group Variable Selection via SCAD-L2. Statistics: A Journal of Theoretical and Applied Statistics. pp 1-18 (2012)

Appendix

Table 4. The Explanatory Variables Used in This Study

Variables	Initial Notation	Notation in The Model	Description
Group 1 : Demographic factors			
Residential area	X_{11}	D_{111}	0 : Rural ; 1: Urban
Age	X_{12}	X_{12}	-
Age of first marriage	X_{13}	X_{13}	-
Pregnancy history	X_{14}	X_{14}	0 : Complications ; 1 : No complications
Husband/partner's age	X_{15}	D_{151}	-
Marital status	X_{16}	D_{161}	0 : No married ; 1 : Married/Never married
Group 2 : Factors of female status and autonomy			
Education Level	X_{21}	D_{211}, D_{212}, D_{213}	0 : No education; 1 : Primary; 2: Secondary; 3 : Higher
Working status	X_{22}	D_{221}	0 : No; 1: Yes
Decision making : health problems	X_{23}	D_{231}	0 : Doesn't involve a wife ; 1 : Involve wife
Decision making : large household purchases	X_{24}	D_{241}	0 : Doesn't involve a wife ; 1 : Involve wife
Decision making : Visits to family or relatives	X_{25}	D_{251}	0 : Doesn't involve a wife ; 1 : Involve wife
Decision making : Money husband/ partner earns	X_{26}	D_{261}	0 : Doesn't involve a wife ; 1 : Involve wife
Group 3 : Economic factors			
Household has : electricity	X_{31}	D_{311}	0 : No; 1 : Yes
Household has : radio	X_{32}	D_{321}	0 : No; 1 : Yes
Household has : television	X_{33}	D_{331}	0 : No; 1 : Yes
Household has : refrigerator	X_{34}	D_{341}	0 : No; 1 : Yes
Household has : bicycle	X_{35}	D_{351}	0 : No; 1 : Yes
Household has : motorcycle/scooter	X_{36}	D_{361}	0 : No; 1 : Yes
Household has : car/truk	X_{37}	D_{371}	0 : No; 1 : Yes
Has an account in bank /financial institution	X_{38}	D_{381}	0 : No; 1 : Yes
Wealth Indkes	X_{39}	D_{391}, D_{392}	0 : Poor ; 1 : Middle ; 2: Rich

Variables	Initial Notation	Notation in The Model	Description
Group 4 : Family Planning Factors			
Current contraceptive method	X_{41}	D_{411}	0 : No ; 1 : Yes
Desire for more childern	X_{42}	D_{421}, D_{422}	0 : Wants ; 1 : Undecided ; 2 : Wants no more
Husband's desire for children	X_{43}	D_{431}, D_{432}	0 : Husband wants more ; 1 : Both wants same; 2 : Husband wants fewer
Ideal number of children	X_{44}	D_{441}	0 : > 2 children ; 1 : <= 2 children
Family Planning : radio	X_{45}	D_{451}	0 : No ; 1 : Yes
Family Planning : television	X_{46}	D_{461}	0 : No ; 1 : Yes
Family Planning : newspaper/magazine	X_{47}	D_{471}	0 : No ; 1 : Yes

Description: Nottion used: $X_{ij,}$ D_{ijk} : (i: ith Group; j: jth Predictor Variable; k: kth dummy)