Subset Best Method Regression Analysis with Cp Mallows Statistics on Factors Affecting Life Expectancy

by Hardian Bimanto

Submission date: 19-Aug-2022 11:29AM (UTC+0800)
Submission ID: 1884202064
File name: artikel-Subset_Best_Method.pdf (151.95K)
Word count: 3547
Character count: 17620

DOI: http://dx.doi.org/10.33846/hn60306
http://heanoti.com/index.php/hn

RESEARCH ARTICLE
URL of this article: http://heanoti.com/index.php/hn/article/view/hn60306

Subset Best Method Regression Analysis with Cp Mallows Statistics on Factors Affecting Life Expectancy

Hardian Bimanto[1(CA)], Hari Basuki Notobroto[2], Soenarnatalina Melaniani[3]
[1(CA)]Magister Program, Department of Biostatistics and Population, Faculty of Public Health, Universitas Airlangga, Indonesia; hardian.bimanto-2019@fkm.unair.ac.id (Corresponding Author)
[2]Department of Biostatistics and Population, Faculty of Public Health, Universitas Airlangga, Indonesia; harinotobroto@gmail.com
[3]Department of Biostatistics and Population, Faculty of Public Health, Universitas Airlangga, Indonesia; soenarnatalina.m@fkm.unair.ac.id

ABSTRACT

One method used in the multiple linear regression model is the best subset using Cp Mallows statistics. The best subset begins by combining independent variables to describe the dependent variable. Select a combination with a high coefficient of determination and a low Cp value. The purpose of this study was to apply the best subset method of Cp Mallows statistics to obtain multiple linear regression models of factors that affect life expectancy. This survey was a secondary data survey using data from Health Profiles published in 2016 in East Java. Data from 38 cities with dependent variables were life expectancy and independent variables such as diarrhea prevalence, dengue prevalence, healthy homes, clean and healthy living behavior, and average school hours. The results showed that the coefficient of determination was 69.7% and the Cp value was 3.9 for the three combinations of diarrhea prevalence, healthy family, and average school hours.

Keywords: life expectancy; subset best method; regression; Cp Mallows

INTRODUCTION

Background

Simple linear regression is used when there is only one independent variable, but multiple regression is used for multiple independent variables [(1)]. Regression model selection is the selection of independent variables (independent variables) contained in the regression model to explain the dependent variables (dependent variables) [(2)]. When determining the model, you can enter the independent variables together or individually [(3)]. Still, all the variables in the regression model make the model the most appropriate, efficient, and theoretical way to explain the dependent variables, it cannot be rational [(4)]. One of the methods used in modeling is best subset regression [(5)]. The best subset regression allows any combination of independent variables that can explain the dependent variable [(6)].

There are several conditions for choosing the right model for Best Subset Regression. One of them uses Cp Mallow's statistics [(7)]. The best subset regression statistics using Cp Mallows start with choosing the simplest model, a one-variable model. Then move to another variable one at a time [(8)]. In the best subset approach, predictive modeling uses all independent variables. All independent variables are considered important for obtaining the correct model using the best subset method using Mallows' Cp statistics and building all possible combinations of variables. Then make a selection based on a small error and a small Cp value in the appropriate regression model [(9)].

Purpose

The purpose of this study was to apply the best subset method of Cp Mallows statistics to obtain multiple linear regression models of factors that affect life expectancy.

METHODS

Health is affected by several factors, including environmental factors, morbidity, and level of education. A study entitled Factors Affecting Life Expectancy in Gembel Regency [(10)] shows that clean and healthy lifestyles affect life expectancy in Gembel Regency. A study [(11)] suggests that one of the factors contributing to the

development of diarrhea is caused by a lack of a clean and healthy lifestyle. The variables investigated are the factors that affect your health. Henrik L. Bloom's theory explains that health is affected by four factors: environmental factors, behavior, health care, and heredity. One of the environmental factors that can affect health is the prevalence of diarrhea and dengue [12]. Life expectancy, high or low, depends on the factors that influence it.

This survey was a secondary data search and is the population of all districts/cities in East Java. The sample consists of 38 districts and cities taken from the Central Statistics Bureau of East Java in 2016. The data used East Java Health Profile In this study; life expectancy was a dependent variable in East Java in 2016. The influencing factors were the prevalence of diarrhea, dengue fever, clean and healthy living behavior, healthy living, and independent variables.

To obtain mathematical models that can predict the factors that influence life expectancy in East Java, they need to be analyzed statistically. In this case, multiple linear regression analysis is used because the analysis is primarily used to address the problem of life expectancy and helps determine the effect of the independent variable on the dependent variable. This study uses multiple regression using the best statistical subset of the Cp Mallows regression method. Cp Mallow's statistics can provide the accuracy of possible models constructed from all independent variables [13].

RESULTS

Mallows Cp statistics are used to evaluate the suitability of regression models. When you select **a** model in this way, all variables in the model are considered important**,** many equations form all possible combinations of variables**,** and then one selection is made**.** Combination of Cp and R^2_{adj} value criteria.

Table 1. Variable combination of Cp and R^2_{adj} value criteria

Group		Variable combination	R^2 (%)	R^2_{adj} (%)	Cp	S^2
A	1	X_1	6.70	4.10	70.30	2.01
	2	X_2	0.00	0.00	77.80	2.08
	3	X_3	40.70	39.10	32.30	1.60
	4	X_4	20.00	17.80	55.40	1.86
	5	**X_5**	53.50	52.20	18.00	1.42
B	1	X_1,X_3	47.60	44.60	26.50	1.53
	2	**X_1,X_5**	65.20	63.30	6.90	1.24
	3	X_2,X_5	54.80	52.20	18.60	1.42
	4	X_3,X_5	59.20	56.80	13.60	1.35
	5	X_4,X_5	57.30	54.80	15.80	1.38
C	1	X_1,X_2,X_5	65.70	62.70	8.30	1.25
	2	**X_1,X_3,X_5**	69.70	67.00	3.90	1.18
	3	X_1,X_4,X_5	67.00	64.10	6.90	1.23
	4	X_2,X_3,X_5	61.60	58.20	13.00	1.33
	5	X_3,X_4,X_5	60.90	57.50	13.70	1.34
D	1	X_1,X_2,X_3,X_4	51.00	45.00	26.80	1.52
	2	**X_1,X_2,X_3,X_5**	70.90	67.40	4.50	1.17
	3	X_1,X_2,X_4,X_5	67.40	63.40	8.40	1.24
	4	X_1,X_3,X_4,X_5	70.30	66.70	5.20	1.18
	5	X_2,X_3,X_4,X_5	62.90	58.40	13.40	1.32
E	1	**X_1,X_2,X_3,X_4,X_5**	71.40	66.90	6.00	1.18

Choose a combination of Independent variables on the condition that the value of Cp is small and R^2_{adj} is close to 100%.

Table 2. Variable Combinations Formed

Variable combination	R^2	R^2_{adj} (%)	Cp	S^2
X_5	53.5	52.20	18.00	1.42
X_1,X_5	65.2	63.30	6.90	1.24
X_1,X_3,X_5	69.7	67.00	3.90	1.18
X_1,X_2,X_3,X_5	70.9	67.40	4.50	1.17
X_1,X_2,X_3,X_4,X_5	71.4	66.90	6.00	1.18

Description:

X_1 = Diarrhea prevalence

X_2 = Dengue Hemorrhagic fever prevalence

X_3 = Healthy house

X_4 = Clean and healthy life behavior

X_5 = Average length of time in school

The next step is to select the appropriate combination of simultaneous and regression coefficient test variables for each independent variable. The test is first run from most variable combinations, that is, five independent variable combinations. If you want to use an alpha of 0.05, use the Anova hypothesis as follows:
H_0: There is no significant common effect between the independent variables of the dependent variable
H_1: There is a significant joint effect between the independent variables of the dependent variable

Table 3. Simultaneous test of regression model with independent variables

Model	DF	SS	MS	F-value	p
Regresion	5	112.107	22.421	15.960	0.000
Eror	32	44.962	1.405		
Total	37	157.069			

In Table 3, $p < 0.05$ (H_0 was rejected). The value of R2 or the dependent variable with a coefficient of determination of 71.4% has important effects such as the independent variables X1, X2, X3, X4, X5. In addition, run a regression coefficient test to see if each independent variable can explain the dependent variable. All independent variables need to be important to get a good model. Below is the regression coefficient hypothesis
H_0: No significant effect of the independent variable on the dependent variable
H_1: Independent variables have a large effect on dependent variables

Table 4. Regression coefficient test with independent variables X1, X2, X3, X4, X5

Predictor	Coefficient	T	p
Constant	64.33	57.15	0.000
Diarrhea prevalence (X_1)	-0.0557	-3.07	0.004
DHF prevalence (X_2)	0.0416	1.10	0.280
Healthy house (X_3)	0.0279	2.11	0.043
Clean and healthy life behavior (X_4)	0.0108	0.72	0.476
Average length of time in school (X_5)	0.709	4.78	0.000

In table 4, three independent variables are significant or meet the criteria for H_o rejection, namely the prevalence of diarrhoea, healthy homes, and average years of schooling. Still, most dengue hemorrhagic fever and clean and healthy living behaviour do not meet the criteria for rejection of H_o. Hence, this combination is not A good model then tests a variety of four variables.

Table 5. Simultaneous Test of Regression Model with Independent Variables X_1, X_2,X_3, X_5

Model	DF	SS	MS	F	p
Regresion	4	111.377	27.844	20.11	0.000
Eror	33	45.692	1.385	-	-
Total	37	157.069	-	-	-

Table 5 shows the results that $p < 0.05$, which means that there is a significant influence together, namely the Independent variables X_1, X_2, X_3, X_5 on the dependent variable with a value of R^2 or a coefficient of determination of 70.9%

Table 6. Regression coefficient test with independent variables X_1, X_2, X_3, X_5

Predictor	Coefficient	T	p
Constant	64.59	60.99	0.000
Diarrhea Prevalence	-0.0578	-3.26	0.003
DHF Prevalence	0.0438	1.17	0.250
	0.0306	2.43	0.021
	0.727	5.00	0.000

Table 6 appears that the factors of the runs predominance, solid domestic and average length of tutoring are littler than alpha 0.05. In contrast, the predominance of dengue hemorrhagic fever is more prominent than alpha 0.05, which implies that the combination of four isn't an outstanding demonstration.

Table 7. Simultaneous test of regression model with independent variables X_1, X_3, X_5

Model	DF	SS	MS	F	p
Regresion	Healthy house	109.479	36.493	26.07	0.000
Eror	Average length of time in school	47.591	1.400		
Total	37	157.069			

Table 7 appears the comes about that $p < 0.05$, which implies that there's a noteworthy impact together, specifically the Autonomous factors X_1, X_3, X_5 on the Subordinate variable with R^2 esteem or the coefficient of assurance of 69.7%

Table 8. Relapse coefficient test with autonomous factors X_1, X_3, X_5

Predictor	Coefficient	T	p
Constant	65.156	68.79	0.000
Diarrhea Prevalence	-0.0608	-3.44	0.002
Healthy house	0.0279	2.24	0.032
Average length of time in school	0.728	4.98	0.000

Table 8 shows that the runs predominance, solid domestic and standard length of tutoring are noteworthy since p <0.05 implies that all combinations with three autonomous elements influence the subordinate variable. So the combination could be a great demonstrate

Cp Esteem and Coefficient of Assurance in All Combinations

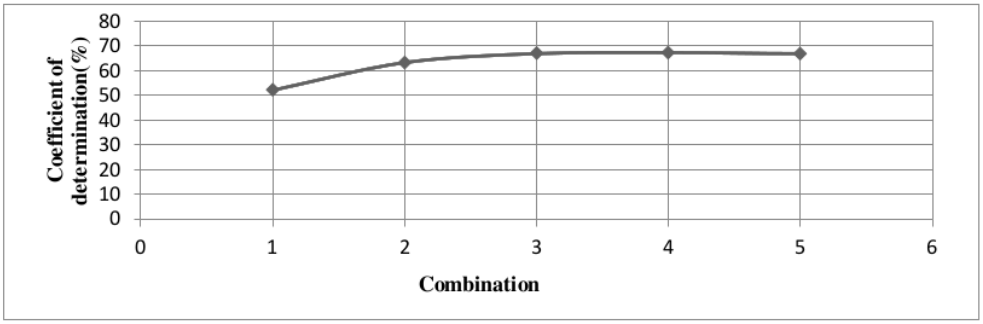

Figure 1. Coefficient of determination

It three varieties of the runs predominance, solid homes and standard length of tutoring meet the criteria in clarifying life anticipation with a coefficient of assurance of 69.7% and a Cp esteem of 3.9

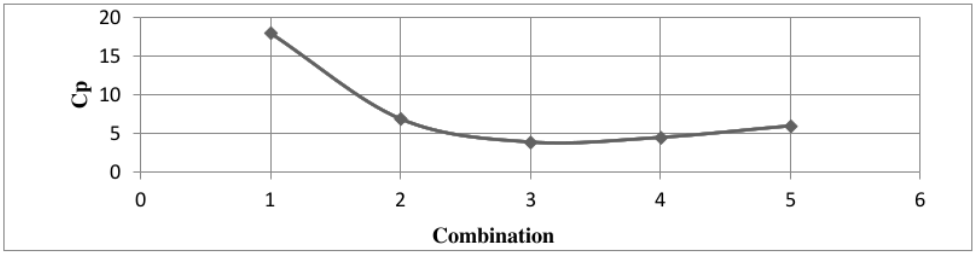

Figure 2. Cp Mallow's value

DISCUSSION

Best Subset strategy may be a strategy to induce the most excellent demonstration in relapse [14]. This strategy considers all autonomous variables essential to utilize. Another, make openings by combining all the Autonomous factors to clarify the Dependent variable [15]. Usually necessary because the more Free elements used in making the relapse demonstrate, the superior since it'll be depicted as an entirety in predicting the circumstance within the populace [16]. After making a combination of Autonomous factors that are shaped with one combination, two combinations, three combinations, and so on, select the most excellent combination with fundamental criteria, to be specific by considering the tall coefficient of assurance and Mallows Cp esteem [17].

The coefficient of assurance is the amount utilized to degree the fit within the relapse show. Estimation is done by calculating the differences of autonomous factors to explain the subordinate variable within the demonstration. Agreeing to Melisa, 2009 The more noteworthy the coefficient of assurance within the demonstrate or condition, the more prominent the level of certainty within the show in clarifying the subordinate variable [18].

The Cp Mallows esteem is the assessed esteem to induce the remaining or mistake within the combination by comparing the cruel square error between factors. Concurring to Hanum, 2011 the contrast within the affection of Cp to the variety of free factors will be utilized as a degree of mistake within the demonstration to ensure that

the chosen combination of Independent variables will have a minor error in clarifying the subordinate variable within the show [19]. The inadequacy of the Most excellent Subset of Mallows Cp factual strategy requires a long and careful step in making conceivable combinations and calculating the Cp esteem as a condition for choosing a great combination [20].

Based on the results of research that has been carried out, it is found that the prevalence of diarrhoea affects life expectancy. The results of the significant value of the partial statistical test of 0.002, which is less than 0.05, and the negative value of the regression coefficient is 0.0608, which means that every decrease in the prevalence of diarrhoea by one case per 1000 population can increase life expectancy by 0.0608. year. This is in line with Hendrik L. Blum's theory which proves that environmental factors are factors that influence the health status [21].

Based on information from Riskedas in 2013 [22], East Java Territory was within the 11th position with the highest predominance of the runs out of 33 territories in Indonesia. If numerous individuals have diarrhoea in a region, it'll directly influence their well-being [23]. It was evident from this study that the predominance of loose bowels is conversely relative to life anticipation, which suggests that the higher the frequency of loose bowels, the lower the life hope of an individual [24].

A sound domestic is one of the variables that can influence life anticipation. Based on the comes about of the think about, an excellent domestic influences life anticipation. That can be proven by a halfway real test of 0.032, less than 0.05. The esteem of the regression coefficient is positive at 0.0279, which implies that each one per cent increase in sound homes will increment life anticipation 0.0279 a long time. Typically in line with individuals who live in stable environments have greater well-being than individuals who live in undesirable situations [25]. The sound environment alluded to here may be a house or territory that has a clean water supply, squander administration, and transfer of faeces. Typically related to this investigation, a clean and sound domestic can decrease the hazard of malady and increment life anticipation [26].

The average length of time in school is one of the components that can influence life hope—typically demonstrated by the comes about of a halfway real test of 0.000 which is less than 0.05. The esteem of the regression coefficient is positive at 0.728, which suggests that each increment in one year of tutoring can increment life anticipation by 0.728 years. On the off chance that somebody has tall information, the insight of one's well-being will increase. The average length of tutoring depicts a person's instruction which implies that education is required to make strides in the quality of human life; moving forward, wellbeing status isn't as it were sound but broad-minded.

CONCLUSION

The Leading Subset strategy makes a combination shaped from all the Autonomous factors. Making these combinations decides the conditions in a populace that cannot be known as an entire. So that it can be utilized to play down show mistakes in expectations. The more free factors, the populace forecasts will be portrayed entirely. Variables that influence life hope in East Java Territory in 2016 are the predominance of the runs, solid homes and the normal length of the school.

REFERENCES

1. Schneider A, Hommel G, Blettner M. Lineare Regressions Analyse - Teil 14 der serie zur bewertung wissenschaftlicher publikationen. Dtsch Arztebl. 2010;107(44):776–82.
2. Ratner B. Variable Selection Methods in Regression: Ignorable Problem, Outing Notable Solution. J Targeting, Meas Anal Mark. 2010;18(1):65–75.
3. Austin PC, Steyerberg EW. The Number of Subjects Per Variable Required in Linear Regression Analyses. J Clin Epidemiol. 2015;68(6):627–36.
4. Alexopoulos EC. Introduction to Multivariate Regression Analysis. Hippokratia [Internet]. 2005;14(1):23–8.
5. Felix N. A New Method for Regression Model Selection. 2020;8(12):1858–99.
6. Maxwell O. Comparison of Some Variable Selection Techniques in Regression Analysis. Am J Biomed Sci Res. 2019;6(4):281–93.
7. Darnius O, Tarigan G. Simulation Method of Model Selection based on Mallows' Cp Criteria in Linier Regression. J Phys Conf Ser. 2018;1116(2).
8. Zhang Z. Variable Selection with Stepwise and Best Subset Approaches. Ann Transl Med. 2016;4(7):1–6.
9. Chowdhury MZI, Turin TC. Variable Selection Strategies and Its Importance in Clinical Prediction Modelling. Fam Med Community Heal. 2020;8(1).
10. Chen Z, Ma Y, Hua J, Wang Y, Guo H. Impacts from Economic Development and Environmental Factors on Life Expectancy: A Comparative Study Based on Data from Both Developed and Developing Countries from 2004 to 2016. Int J Environ Res Public Health. 2021;18(16):1–18.
11. Autoridad Nacional del Servicio Civil. 済無No Title No Title No Title. Angew Chemie Int Ed 6(11), 951–952. 2021;(January):2013–5.
12. WHO. Comprehensive Guidelines for Prevention and Control of Dengue and Dengue Haemorrhagic Fever. WHO Regional Publication SEARO. 2011;159–168

13. Zuccaro C. Mallows' Cp Statistic and Model Selection in Multiple Linear Regression. Mark Res Soc Journal. 1992;34(2):1–10.
14. Rohsenow DJ. Cognitive Behavioral Therapy for Substance Use Disorders. Encycl Ment Heal Second Ed. 2016;33(3):307–16.
15. Kusurkar RA, Ten Cate TJ, Van Asperen M, Croiset G. Motivation as an Independent and a Dependent Variable in Medical Education: A Review of the Literature. Med Teach. 2011;33(5).
16. Substance Abuse and Mental Health Services Administration. Brief Interventions and Brief Therapies for Substance Abuse. Br Interv Br Ther Subst Abus. 2012;105–21.
17. Wulff SS. A First Course in Design and Analysis of Experiments. The American Statistician. 2003;57:66-67.
18. Gerbing DW. Campbell and Stanley for Undergraduates. Vol. 29, Contemporary Psychology: A Journal of Reviews. 1984. 333–333 p.
19. del Campo-Albendea L, Muriel-García A. Ten Common Statistical Mistakes to Watch Out for when Writing or Reviewing a Manuscript. Enferm Intensiva. 2021;32(1):42–4.
20. Miyashiro R, Takano Y. Subset Selection by Mallows' C P. Expert Syst Appl. 2015;42(1):325–31.
21. Wibisono FA, Kurniawati ER. Modeling the Number of Multibacillary Leprosy Using Negative Binomial Regression to Overcome Overdispersion in Poisson Regression. J Biometrika dan Kependud. 2020;9(2):153.
22. Linder FE. National Health Survey. Science (80-). 1958;127(3309):1275–9.
23. WHO. Preventing Diarrhoea through Better Water, Sanitation and Hygiene. Geneva: World Health Organization; 2014.
24. Sweetser S. Evaluating the Patient with Diarrhea: A Case-based approach. Mayo Clin Proc. 2012;87(6):596–602.
25. VALENCIA A. No se ve un incentivo para el sector agropecuario en la reforma: Fenavi. Caracol Radio. 2016;51:1–16.
26. Martuzzi M, Tickner J a. The Precautionary Principle: Protecting Public Health, the Environment and the Future of Our Children. Geneva: WHO; 2004.

Subset Best Method Regression Analysis with Cp Mallows Statistics on Factors Affecting Life Expectancy

ORIGINALITY REPORT

11%	9%	7%	0%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Repository.Unej.Ac.Id Internet Source	2%
2	"1st Annual Conference of Midwifery", Walter de Gruyter GmbH, 2020 Publication	2%
3	repository.ub.ac.id Internet Source	2%
4	www.ijicc.net Internet Source	1%
5	www.heanoti.com Internet Source	1%
6	www.coursehero.com Internet Source	1%
7	es.scribd.com Internet Source	<1%
8	mafiadoc.com Internet Source	<1%

9	Suroso, Dede Nadhilah, Ardiansyah, Edvin Aldrian. "Drought detection in Java Island based on Standardized Precipitation and Evapotranspiration Index (SPEI)", Journal of Water and Climate Change, 2021 Publication	<1 %
10	condor.ucr.edu Internet Source	<1 %
11	fmch.bmj.com Internet Source	<1 %
12	Ediu Carlos da Silva Junior, Lúcia Helena de Oliveira Wadt, Kátia Emídio da Silva, Roberval Monteiro Bezerra de Lima et al. "Geochemistry of selenium, barium, and iodine in representative soils of the Brazilian Amazon rainforest", Science of The Total Environment, 2022 Publication	<1 %

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	On		

Subset Best Method Regression Analysis with Cp Mallows Statistics on Factors Affecting Life Expectancy

GRADEMARK REPORT

FINAL GRADE

/0

GENERAL COMMENTS

Instructor

PAGE 1

PAGE 2

PAGE 3

PAGE 4

PAGE 5

PAGE 6