Predictive Model on Determinants of Child Mortality Using Multiple Regression Analysis

M A Sonawane\textsuperscript{1,2} and A Y Tayade\textsuperscript{1}
\textsuperscript{1}Department of Statistics, Dr. Babasaheb Ambedkar Marathwada university, Aurangabad – 431004. (M.S)-INDIA
\textsuperscript{2}Corresponding author’s e-mail: manju.stat1311@gmail.com

Abstract. Child mortality is one of the important factors to reflect sustainable development for any nation. It is common sense now that child mortality rate depends on medical services platform and its quality. To predict model under-five child mortality rates in India, multiple linear regression analysis was used. Identifying the factors that affects the under-five child mortality. To examine the relationship between all the variables along with recognize the problem of multicollinearity in the variables. In this study our fitted multiple linear regression model found that under five child mortality rates in India are influenced by tuberculosis case detection rate, measles death rate and hepatitis immunization rate. To choose appropriate model ACI score is used.

Keywords: child mortality rate; multiple linear regression; variable selection method; AIC score.

1. Introduction
The under-five child mortality is a leading indicator of health of children and overall development of the economy of region or country [1]. UNICEF reported that the welfare of child’s has raised, and the under-five child mortality rates reduce in the most of nations. Thirty year ago (1990) the word made a commitment to protect and fulfil children’s rights, the important right is the right of every child to survive. There is improvement in under-five child survival rate but still defeat to completely reach that responsibility today. These children died largely of treatable and preventable causes namely infection disease and injuries, it is intolerable [2].

Recently UNICEF reported that the universal overall under five child mortality rates decrease by 59\% from 93\% in 1990 to 39\% in 2018 and In India under five child mortality rate decreases by 83\% from 126\% in 37 to 37 in 2018. There is the global progress in reducing the child mortality rates but, in few regions, still have high mortality rates. Most of the universal burden of under-five child mortality rate bear by two regions. The under-five child death in sub-Saharan Africa was 52\% died and in central and South Asia 29\% died. The two regions Sub-Saharan Africa and central & south Asia account for more than 80\% under five mortality in 2018, while these region account for 52 percent of the universal under-five population. In year 2018 five countries: India, Nigeria, Pakistan, the Democratic Republic of the Congo and Ethiopia occurred 50 percent under five child mortality all the under-five child mortality. In India the major causes of under-five child mortality are infection diseases [2]. According on that reality the aim of this research was to predict model for under five child mortality rates in India and identify the factors that affects the under-five child mortality. In this study factors are carbon dioxide emission, percentage of Hepatitis B immunization coverage among one-year-old, percentage of case detection rate for all forms of tuberculosis rate, number of reported death on HIV, number of reported deaths on one-year-old immunization against measles.

2. Reviews of Literature
Based on UNICEF (2012) child mortality is high in some region. By nation report UNISCAP (2013) child mortality rate highest decreases in East Asia and the lowest in Central Asia. Based on that fact Anwar et al. (2016) study was correlation between under-five child mortality rates and its related factors
of Asia region and they found that the under-five child mortality rate affected by quality of drinking water [3].

United nation children’s fund (2013) reported that the highest child mortality shown in two region sub-Saharan Africa 41% and south Asia 34% [4]. Black et. al (2003) conducted a research in year 2000 for estimating the under-five child mortality and its various causes for 42 countries using predictive model [5]. The results of predictive model were compared with World Health Organization figures and understanding vital causes of under-five child mortality.

Aiken et al. (2003) was conducted research on application of multiple linear regression analysis and examining the relationship between one dependent variable and more than one independent variables. They also talk about problem of analysis of variance in multiple interrelationship [6]. The multiple linear regression analysis was applied to study the contribution of few explanatory variables to a response variable according to Draper and Smith (1981) [7].

Marrow- Howell (1994) discussed about multicollinearity: in multiple regression analysis if two or more explanatory variables are related to each other as well related to dependent variable this situation refers the multicollinearity. Variable Influence Factor (VIF) was used to estimating the multicollinearity. VIF evaluate how much the variance of an explanatory variable is inflated by its correlation with the other independent variable [8].

In (1974) for the purpose model/variable selection Akaike introduce the Akaike information criterion (AIC). AIC based on the Kullback-Leiber discrepancy. Akaike Information Criterion is the most used model selection for time series data [9].

(2012) Ashok Chourasia have conducted study about model selection criteria, he discussed that complete data in applied statistics. They explore model selection by Akiake Information Criteria in the multiple regression arrangement with ignoring missing value consider for along multiple imputation [10].

3. Data and methodology

3.1 Data and source of data
For the present research, under-five child mortality rates data was used form year 2007 to 2014. This data is collected from the annual report Register General and Census of India, United Nation Economic and Social Commission for Asia and the Pacific (2015) [11], United Nation Children’s Fund (2019) [12] and World Health Organization (2015) [13].

The under-five child mortality rate data are obtained from the Register General and Census of India, for carbon dioxide emissions, the % of children immunized against measles are taken from the UNESCAP (2015), coverage of hepatitis, death on HIV, immunized against measles, tuberculosis case detection rate, deaths on measles, were collected from the WHO (2015).

3.2 Research methodology
The multiple regression model use for estimating relationship between a response variable and more than one explanatory variable by using Ordinary Least Square method [14].

The multiple linear regression model was given

\[ y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon \]  

The unknown parameter \( \beta_j \), \( j = 0, 1, 2, \ldots, n \) are regression coefficient

Correlation coefficient is the method used to measure strength of the linear association between variables also say that use to estimate the dependency between the two variables.

The Pearson correlation coefficient obtained by

\[ \text{corr}(x,y) = \rho_{x,y} = \frac{\text{cov}(x,y)}{\sigma_x \sigma_y} \]  

Where,
\[ \sigma_x = \text{the standard deviation of } x \]
\[ \sigma_y = \text{the standard deviation of } y \]

The correlation coefficient is always taken value between -1 to +1,
Where,
\[ +1 = \text{Two variables are perfectly positive related} \]
\[ -1 = \text{Two variables are perfectly negative related} \]
\[ 0 = \text{No correlation between the two variables} \]

Testing of significance of the correlation

\[ \rho_{x,y} = 0 \] (There is not a significant linear relationship between two variables)

\[ \rho_{x,y} \neq 0 \] (There is a significant linear relationship between the two variables)

For examining the association between two variables, tabulated t value of correlation coefficient was compared with to the critical t value at 5% level of significance.
If \( t_{\text{cal}} > t_{\text{tab}} \), then reject \( H_0 \). There is no linear correlation between variable.

Multicollinearity is a condition that arises in multiple regression analysis when two or more explanatory variables are highly linearly related. If there is multicollinearity in variables that can create inaccurate estimates of the regression coefficients [15]. The presence of multicollinearity adversely affects regression result. The appearance of multicollinearity was checked using variance influence factor (VIF). The variable influence factor (VIF) calculated based on tolerance

\[ \text{Tolerance} = 1 - R_j^2 \]

Where \( R_j^2 \) is the coefficient determination of a regression of equation. Following is the formula of VIF is written as

\[ VIF = \frac{1}{\text{tolerance}} \]

The VIF measures how much a variable is contributing to the standard error in the regression. The value of VIF is 10 or more than 10 that indicates there is presence of multicollinearity [16]. The ordinary least square method was applied to data set for best characterizes as well as to estimate the parameters. For this study, the independent variable \( y = \text{child under age five mortality rate} \) and the independent variables were \( x_1 = \text{carbon dioxide emission (metric tons per capita)} \), \( x_2 = \text{immunized against measles} \), \( x_3 = \text{coverage of hepatitis B immunization} \), \( x_4 = \text{case detection rate of tuberculosis} \), \( x_5 = \text{number of death on HIV} \), \( x_6 = \text{number of deaths on measles} \).

Akaike information criteria for variable selection
AIC model is given as follows:

\[ AIC = n \ln(\hat{\theta})^2 + 2p \]

AIC uses a model’s maximum likelihood estimation as a measure fit. Choose the model with the lowest score as the best [17].

4. Results and Discussion
In this research under-five child mortality is response variable was expected to have relationship with explanatory variables.

| Table 1. Linear correlation coefficient between variables |
|----------------|----------------|----------------|----------------|----------------|----------------|
| y               | x_1            | x_2            | x_3            | x_4            | x_5            | x_6            |
| y               | 1              | -0.97203       | 1              | -0.9088        | 0.87397        | 1              |
| x_1             | -0.97203       | 1              |                |                |                |                |
| x_2             | -0.9088        | 0.87397        | 1              |                |                |                |
Table 1 display the correlation matrix between variables for the under-five child mortality rates. On basis of table 1, there is strong positive relationship between $x_1$ and $x_2$, $x_1$ and $x_3$, $x_1$ and $x_4$, $x_2$ and $x_3$, $x_4$ and $x_5$, $x_4$ and $x_6$, $x_5$ and $x_6$. This situation recommended may have multicollinearity problem. Already we discussed in above section, if the VIF > 10, then possibility of multiple correlation.

Table 2. The under-five child mortality rate model

| variable | parameter estimate | standard Error | t   | p     | VIF |
|----------|--------------------|----------------|-----|-------|-----|
| intercept| 69.968             | 7.107          | 9.84| 0.064 | 0   |
| $x_1$    | 2.682              | 2.132          | 1.26| 0.428 | 62.8|
| $x_2$    | -0.29948           | 0.06202        | -4.83| 0.13  | 58.5|
| $x_3$    | -0.06875           | 0.0146         | -4.71| 0.133 | 62.4|
| $x_4$    | -0.11201           | 0.01916        | -5.85| 0.108 | 7.1 |
| $x_5$    | 0.00009733         | 0.00003935     | 2.47 | 0.245 | 699 |
| $x_6$    | 0.00007345         | 0.00005142     | 1.43 | 0.388 | 355.5|

The table 2 display the full model of multiple regression for the under-five child mortality with estimated regression coefficient and VIF for each explanatory variable. Based on 2 table variables $x_1$, $x_2$, $x_3$, $x_5$ and $x_6$ have VIF > 10 that shows the appearance of multicollinearity in the model. Remove the variables from the model with a high VIF and continue this process up to VIF getting less than 10. Among these 5 variables $x_5$ number of reported death on HIV showed the highest VIF (699), and therefor this variable was remove for new selection. After removing variable $x_5$ outcome was shown in table 3.

Table 3. Model of the under-five child mortality rate study after omitting $x_5$ (measles)

| Variable | parameter estimate | standard Error | t   | p     | VIF |
|----------|--------------------|----------------|-----|-------|-----|
| Intercept| 86.544             | 4.464          | 19.39| 0.003 | 0   |
| $x_1$    | 5.64               | 3.329          | 1.69 | 0.232 | 43  |
| $x_2$    | -0.44423           | 0.03872        | -11.47| 0.008 | 6.4 |
| $x_3$    | -0.09247           | 0.02077        | -4.45 | 0.047 | 35.5|
| $x_4$    | -0.13864           | 0.0299         | -4.64 | 0.044 | 4.9 |
| $x_6$    | 0.00019647         | 0.00002327     | 8.44 | 0.014 | 20.6|

The table 3 shows that variables $x_1$, $x_3$ and $x_6$ have VIF > 10 which shows that the f multicollinearity in the model. Out of these 3 variables $x_4$ total carbon dioxide emission showed the highest VIF (43), and therefor this variable was remove for remaining selection. After removing variable $x_1$ the outcome was display in table 4.

Table 4. model of the under-five child mortality rate study after omitting $x_1$ (carbon dioxide emission)


The table 4 shows that variables $x_3$ and $x_6$ have VIF > 10 which show the multicollinearity problem in the model. Out of these 2 variables $x_6$ no of death on measles showed the highest VIF (20), and therefor this variable was omitted for remaining selection. After removing variable $x_1$ the outcome was shown in table 5.

Table 5. model of the under-five child mortality rate study after omitting $x_6$ immunization against measles

| variable | parameter estimate | standard Error | t | p | VIF |
|----------|--------------------|----------------|---|---|-----|
| intercept | 118.740            | 10.31          | 11.52 | 0.000 | 0.0 |
| $x_2$    | -0.52910           | 0.129          | -4.1 | 0.015 | 3.6 |
| $x_3$    | -0.14012           | 0.033483       | -4.02 | 0.016 | 5.1 |
| $x_4$    | -0.27978           | 0.08336        | -3.36 | 0.028 | 1.9 |

According to backword elimination criteria for the variable selection, which was $\alpha = 0.15$, not all variable contributes to under-five child mortality rate. The explanatory variable with highest p value omitted from the model. The procedure is continuing up to p value of variables is less than no alpha value. All p values in table 5 is less than $\alpha$ value.

The predicted regression model for the under-five child mortality rates was written as:

$$ \hat{y} = 119 - 0.529x_2 - 0.140x_3 - 0.280x_4 $$

The predicted model found variables which is affecting the child mortality rate, $x_2$ immunized against measles, $x_3$coverage of hepatitis B immunization, $x_4$ case detection rate of tuberculosis. Another method AIC was used to select variables, based on this selection criteria the pick model with smallest AIC score value as best.

Table 6. AIC selected variable

| Variable | Df | SS   | RSS  | AIC   |
|----------|----|------|------|-------|
| None     | 1  | 0.0151 | -36.178 |
| $x_1$    | 1  | 0.0239 | 0.039 | -30.589 |
| $x_6$    | 1  | 0.03103 | 0.04613 | -29.245 |
| $x_5$    | 1  | 0.02941 | 0.10751 | -22.477 |
| $x_3$    | 1  | 0.33496 | 0.35007 | -13.033 |
| $x_2$    | 1  | 0.35223 | 0.36733 | -12.647 |
| $x_4$    | 1  | 0.51608 | 0.53118 | -9.697 |
Based on the for variable selection was found variables which affecting the under-five child mortality rate, \( x_2 \) immunized against measles, \( x_3 \) coverage of hepatitis B immunization, \( x_4 \) case detection rate of tuberculosis.

5. Conclusion
In applied research model selection is an important problem. We used multiple regression analysis and AIC (Akaike information criteria) to predict the model for under-five child mortality rate in India. In this work multicollinearity is one of the important problems. Multicollinearity resolved by using variable selection method and for choosing appropriate model AIC is used. Our study reveals that variable selection method and Akaike information criteria both methods give the same results. This study finds that the main factors which affecting under-five mortality rate in India are immunized against measles, coverage of hepatitis B immunization, case detection rate of tuberculosis.

References
[1] Health Status Statistics: Mortality. 2011 World Health Organization, Health statistics and health information systems
[2] United nations children’s fund (UNICEF) 2010. UNICEF annual report 2019
[3] Fitrianto A, Hanafi I and Chui T L 2016. Modeling aria’s child mortality rate: a thinking of human development in Asia. procedia economics and finance. Elsevier 35 (2016) 249-255.
[4] United Nations Children’s Fund (UNICEF), 2013. UNICEF Annual Report 2012. New York: UNICEF 2013.
[5] Black R E, Morris S S, and Bryce J, 2003. Where and Why are 10 Million Children Dying Every Year, PubMed The Lancet. 2003 Jun 28 361, 2226-2234.
[6] Aiken L S, West S G and Pitts S C, 2003. Multiple Linear Regression in Handbook of Psychology. Schinka J A, Velicer, WF and Weiner, I B Canada: John Wiley & Sons Inc.
[7] Draper N R and Smith H, 1981. Applied Regression Analysis 2nd Edition 135 New York: John Wiley and Sons.
[8] Morrow-Hawell, N. 1994. The M word: Multicollinearity in multiple regression. Social work research, 18, 247-215
[9] Allan D R, Quarrie M C, Tsai C L, 1923. Regression and time series model selection. 2 Word scientific publishing, Singapore.
[10] Chaurasia A and Harel O 2012 Using AIC in multiple linear regression framework with multiply imputed data. Health serv outcomes Res Methodol. 2012 12(2-3) 219-233.
[11] United Nations Economic and Social Commission for Asia and the Pacific (UNESCAP), 2015. Statistical Yearbook for Asia and the Pacific 2014. Bangkok: United Nations Publication.
[12] United Nations Children’s Fund (UNICEF), 2019. UNICEF Annual Report 2018. New York: UNICEF 2019.
[13] World Health Organization (WHO), 2019. World Health Statistics 2018.
[14] Uyanik G K, Guler N (2013) A study on multiple regression analysis. Elsevier Procedia social and behavioural science 106 (2013) 234-240.
[15] Chen G J, 2012. A Simple Way to Deal with Multicollinearity, Taylor and Francis Journal of Applied Statistics, 39, 1893-1909.
[16] Akinwande M O, Dikko H G, Samson A 2015. Variable inflation factor: as a condition for the inclusion of suppressor variables in regression analysis. Open journal of statistics 2015, 5, 754-767
[17] Rao C, Wu Y 2001. Model selection. Lecture notes-monograph series. 38: 1-64