Statistical Analysis on Multiple Linear Regression model building on Child Health data

Abhishek Padalkar

x19221576
MSc in Data Analytics
National College of Ireland

I. Motivation and Statement of Objectives

The crux of this statistical report is building a multiple regression model, applying statistical methods to best fit the model, and perform diagnostics to confirm its generalizability over child mortality health data. The aim is to achieve the best linear unbiased estimator by checking all the required linear model assumptions are met to make it generalized over the total population data based on the available data. To achieve it, a combination of variables was considered and the process was back and forth of adding and removing the variables to settle on a final generalized model. This project is performed in R studio in its entirety and based on the statistical concepts.[1]

The child mortality is the chosen topic of interest in this project. Infant and child mortality has been a big issue over a few decades now, which is still present. With the sustainable development plans by WHO and UN, the rate is seen to be decreasing in the past 10 years. [2] In this project, we work on the data related to Child Mortality with aiming for building a best possible multiple linear regression model on that data.
Objectives in this statistical analysis and model building:

1. Selection of the sample data and variable descriptions.
2. Perform a descriptive graphical analysis of the data to get a high-level understanding of the data and linearity between the dependent variable and other independent variables, and their correlations.
3. Build a multiple linear regression model by measuring the performance of the model using adjusted R^2 and AIC metric. Settle on a model that has the best R^2 value.
4. Perform diagnostics on the model and check if it meets the assumptions of the Gauss Markov and other required assumptions on Linear regression of Ordinary Least Squared method. For any assumption not satisfied, it is rectified and a new model is finalized. All the diagnostics are again performed on the new model until the model meets all the criteria and is generalized.
5. Interpret the final model, its coefficients, and summarize.

II. Data and Description of Variables

As mentioned in the previous section, with the motivation of selecting a child mortality topic, the sample data was researched and gathered from the World Health Organization(WHO) website. The Child Health data is based on the year 2014.[2] It consists of a total of seven variables. The dependent variable, Child Mortality rate, was first collected, and then five independent variables that may have any potential effect on the mortality rate were targeted. Selection of attributes is not a concern in this project, thus the task was performed being cognizant of the fact that there can be other significant variables that can have a greater impact on the mortality rate.

A. Description of the variables

1) Country: This attribute consists of all the countries where the other variables have been measured. Based on these measures, child mortality is predicted by multiple linear regression.[2]

2) Mortality_Rate: This variable is a probability measure of a child dying in a given country under age 5 per 1000 live births.[2]

3) Health_Service_Coverage: Universal health service coverage is an index of health coverage in a country from a scale of 0 to 100 based on a mean of 14 indicators in 4 areas as follows: a. Reproductive, Maternal, newborn, and child health. b. Infectious diseases. c. Non-communicable diseases. d. Service capacity and access. [2] This value focus on services availability in a country related to newborns and maternity and in disease indicators which are also a cause of child mortality. Thus, this variable was chosen.

4) Early_initial_breastfeed: Provides information on the proportion of the children born within 24months who were breastfed within an hour of their births. This indicator was chosen because according to WHO, a child breastfed within an hour of birth is less prone to mortality or getting any disease or infections.[2]

5) Underweight_children: This indicator is a percentage measure of weight less than 2 Standard deviations, underweight proportion, of all of the children weight aged below 5 in a given country. It represents the prevalence of underweights in a country under age 5 among all children. This variable was selected as it can give an estimate of malnourished children.[2]

6) Post_birth_contact_health_provider: This factor is a percentage of newborns who received a health checkup from health personnel within 2 days of delivery. According to WHO, it is seen that if a mother and a newborn have been health checked, there is an opportunity to improve the health of both, and their survival for any complications. Because of this reason, this variable was considered.[2]

7) Infants_provided_minimum_accepted_diet: This variable gives a percentage of children under age 2 who were provided with foods from more than equal to four out of seven food groups in a day. This indicator may have an impact on child mortality due to malnutrition and thus considered to be included in the dataset.[2]

These variables were individually collected and joined to form a single dataset of child health data for year 2014 based on country using python language. For variables that had missing values, were replaced by their respective mean values. Thus, above independent variables, 3 to 7, were selected as potential to predict our dependent variable Child Mortality Rate in a country.

III. Descriptive Analysis

A descriptive analysis and graphical plots are performed to first get an understanding of our data and, on high level, what kind of relationships exists between our independent variables and dependent variables may have.

By the descriptive statistics given below, the dataset has total of 78 instances. Thus, we have a reasonable amount of sample data out of all the countries in the world. For mortality rate, it shows that in 2014, minimum rate was 2.3 % and maximum was 133.1%. With the mean of 33.35%, it is clear that majority of the mortality rate is towards lower side of the spectrum and lies in a range of +-31.92% standard deviations. It is positively skewed, with skewness 1.12. Similarly looking for other independent variables, it is seen that variable Underweight children is also positively skewed and mean percent of underweight children in all the countries in interest is only 7.68%. Rest variables can be said to be reasonably normally distributed.

```
> describe(child_data)
                                      vars  n  mean    sd median trimmed   mad  min   max range  skew kurtosis   se
Country*                                 1 78 39.50 22.66   39.5   39.50 28.91  1.0  78.0  77.0  0.00    -1.25 2.57
Health_Service_Coverage                  2 78 43.41  7.70   43.0   43.20  8.90 29.0  70.0  41.0  0.42     0.49 0.87
Early_initial_breastfeed                 3 78 51.69 11.00   51.7   51.71  0.00 23.0  82.5  59.5  0.10     1.80 1.25
Post_birth_contact_health_provider       4 78 59.19 20.00   59.2   60.00  0.00  4.5  98.5  94.0 -0.41     0.94 2.26
Underweight_children                     5 78  7.68  9.58    2.9    6.09  4.30  0.0  34.6  34.6  1.16     0.43 1.08
Infants_provided_minimum_accepted_diet   6 78 46.43 25.19   46.4   46.03 27.21  0.0  96.2  96.2  0.10    -0.84 2.85
Mortality_Rate                           7 78 33.35 31.92   17.4   28.93 18.68  2.3 133.1 130.8  1.12     0.25 3.61
```

Fig. 1: Descriptive statistics

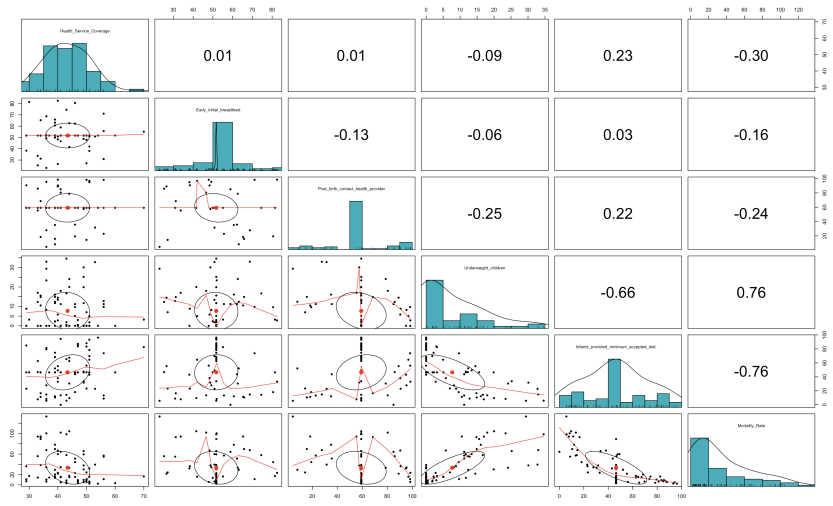


Fig. 2: Correlation and Matrix Scatter plot

A matrix scatterplot and correlation was performed to understand linearity and relationship between all the individual variables with each other and with dependent variable, Mortality Rate. The variables in this diagram below are in the following order: 1. Health_Service_Coverage, 2. Early_initial_breastfeed, 3. Post_birth_contact_health_provider, 4. Underweight_children, 5. Infants_provided_minimum_accepted_diet, 6. Mortality_Rate. First, we look at relationship between our dependent variable and all the independent variables. It is seen that as percentage of Underweight children increase and as the percent of infants provided with minimum food variety decrease, the mortality rate increases. These two variables have above average degree of correlation, of 0.76 and -0.76 respectively, with Mortality rate. By the scatterplot, we can see that Underweight_children has a almost linear relationship and Infants_provided_minimum_accepted_diet has a curvi-linear relationship with Mortality Rate respectively. It is seen that these two independent variables also follows a slight linear pattern and have a reasonable degree of correlation of -0.66. Other variables show a small negative pattern, negative correlation, with increase in mortality rate. There is very low correlation between these variables and do not have any collinearity between them. We can see that, because of higher missing values in 2nd and 3rd variable in the chart, the density of the plot is shifted. Thus, we are cognizant of the fact that it can

affect the linearity and our model mortality prediction. With this understanding, we now can expect that underweight children percentage and infants provided with minimum accepted diet can be a strong predictor of our dependent variable and other 3 variables may have a significant impact on Mortality rate.

IV. Model Building

In this section, we finalize our multiple linear model which can be a best predictor model based on the performance measure of adjusted R^2 statistic. Our only aim here is to raise the performance of the model as high as possible to have a best prediction possible.

We first start with simple linear model to check individual variable impact on our response variable. We measure the performance of simple linear model using R^2 statistic as here we only focus on individual variable. It is given as follows:

1. Early_initial_breastfeed: This variable gives a poor performance model of R^2 0.024 and is insignificant to our response variable alone with p-value of 0.172. This p-value is a hypothesis test value of whether the coefficient of the prediction variable is zero or not. If it is greater than 0.05 then it is zero and thus the variable is insignificant to the response variable.
2. Health_service_coverage: This variable also is not a good predictor as the R^2 value of the model is just 0.089, but it is shown to be significant predictor to our response variable with p-value of 0.0078.
3. Infants_provided_minimum_accepted_diet: It gives a sign of a good predictor as the R^2 value is 0.571 and it also is a strong significant variable with p-value less than 0.000.
4. Post_birth_contact_health_provider: The performance of the model is poor with R^2 value 0.059, but is also significant predictor variable with p-value of 0.031.
5. Underweight_children: This a good predictor to our response variable with 0.570 as R^2 value of the model and is strong significant variable with p-value less than 0.000.

By performing 5 simple linear regression with each independent variable, we now know that, as expected, 3. and 5. are significant and good predictors to our response variable mortality rate. The other three variables do not contribute well to the variability of the dependent variable but 2. and 4. are significant as predictors to our response variable.

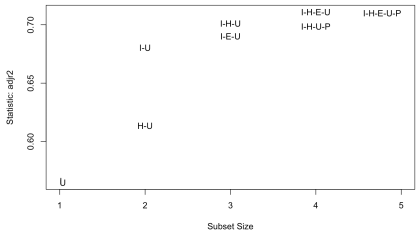


Fig. 3: Best Subset

It is important to know which set of variables can actually contribute to the variability of the response variable. Thus, best possible subset selection of the independent variables is carried to know which combinations can collectively create a best model that can predict best possible values and has high adjusted R^2 value. To get this best subset of variables, we use best subset selection forward approach. Interpreting the best subset graph given below, it is seen that, combination of 4 variables(I-H-E-U) creates equally or more high performance model than including all the variables with adjusted R^2 roughly around 0.71. All the other combinations gives adjusted R^2 value less than 0.71, thus we reject those combinations and move forward with these 4 independent variables by dropping Post_birth_contact_health_provider variable. The performance of the current model is: adjusted R^2 = 0.7109, Residual Standard Error = 17.17 in terms of mortality rate. All our variables are significant except for Early_initial_breastfeed which has F-statistic p-value 0.0665, slightly higher than 0.05. For now, we keep that variable in our best subset model and drop the remaining variable in our next step of model development.

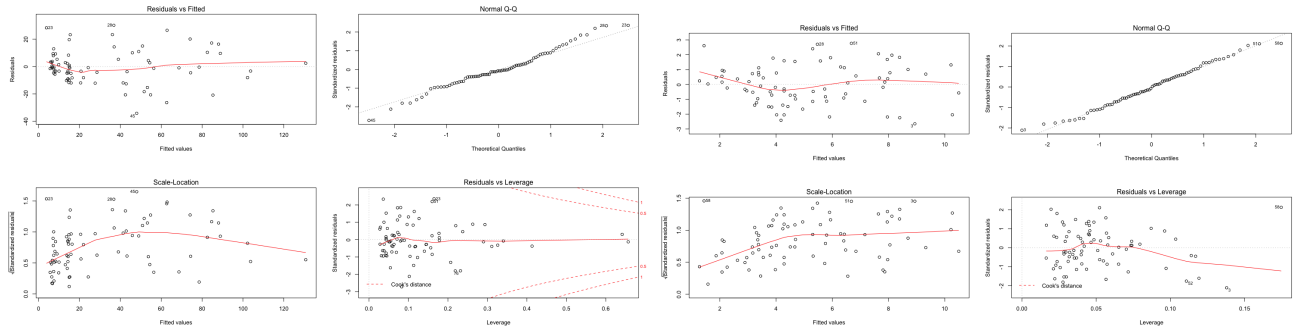
We check if there are any interactions between our independent variables which can improve our model prediction even better. By including it we can know if the response variable is not just based on additive nature of the independent variables. But before that, we check for multicollinearity between these individual variables using VIF(variance inflation factor) metric which shows all the independent variables have vif around 1 which is below 5. Thus, all our 4 predictor variables do not present multicollinearity. We are cognizant of the fact that including interactions may lead to a problem of multicollinearity, but our current aim is to build best possible model. We use step function in R to get a best possible set of variables which will provide minimum AIC metric value. By performing it, the new model consists of 12 variables with possible interactions between each other considered. This is the best combination of the variables with the performance adjusted R^2 value 0.7735. But, there are many insignificant variables according to the p-values. We remove the insignificant variables one by one by decreasing order of their p-value significance and also having a check on how far is it affecting the adjusted R^2 value of the overall model. We perform 6 variable removal steps to then come to our final model having total of 6 variables with all having significant p-value. The adjusted R^2 is 0.7684 of the model by including two interaction terms between Early_initial_breastfeed and Underweight_children, and Health_service_coverage and Underweight_children.

Going back to our descriptive analysis section, we saw that there is a curvi-linear relationship between our dependent variable and our independent variables. Thus, to check if the performance of the model can be increased or not we include polynomials of the 4 variables. By doing so, we see that the new model has the adjusted R^2 of 0.8284. Also, Health_service_coverage squared term is not significant with p-value 0.82. Thus, we discard that variable to get our final model with total of 9 variables, 4 base independent variables, 2 interaction variables and 3 polynomial variables. This final model has the highest performance so far of 0.83 and is good enough to settle on this model for our prediction purposes. But, including interactions and polynomials we are cognizant of the fact that it might increase the chances of multicollinearity.

V. Diagnostics

Multiple Linear Regression model based on Ordinary least squares method has to follow a total of seven assumptions, of which 4 are Gauss Markov assumptions and 3 other assumptions, to make it called to be generalized and best fit model over the available variables. We plot the residual plots for our final model to check the assumptions and is shown in the figure (a). In this section, we discuss and check each assumption by referring to this figure and by checking certain metrics for our final model, and see if it can be rectified.

1) Linearity: This is a Gauss Markov(GM) assumption where the linear model is assumed to have linear relationships between the response variable and all the predictor variables. This can be checked by using Residuals vs Fitted plot in the figure(a). As the plot has a linear fitted red line, it shows that the residuals don't follow any curvi-linear pattern and our model is following the linearity assumption.
2) Homoscedasticity: This is another GM assumption which assumes that the linear model must have a constant variance in the residuals. The residual plot should be randomly scattered and not follow a funnel-in or fan-out pattern. This can be checked in the Scale Location plot in the figure(a). We see that our model has a funnel-in pattern, which means our errors are biased because of which some statistically insignificant variables are shown as significant in our model. And we confirm the presence of heteroscedasicity by performing ncvTest(Non-constant Variance Test). We get a p-value of 0.0469 which is less than 0.05. To rectify this we use square root function to our dependent variable and we get a new model which has a better linear Scale Location plot. Also, now by performing ncvTest, we get p-value of our new model as 0.834 which is greater than 0.05.
3) No Autocorrelation: This is a GM assumption which checks if the error terms in our model are correlated or not. If they are correlated then the errors follow an upward or downward trend or both. This can be checked by performing a durbinWatson test. If the test value is near 2, that generally means that there is no autocorrelation present in the errors. We perform this test on our new, heteroscedasticity corrected, model and the test statistic value is 2.03. Thus, we confirm no autocorrelation.
4) Predictor variables must be independent of Error Terms: This is a GM assumption which assumes that the model must include all the predictor variables which can predict the response variable without any errors. This assumption is an ideal theoretical case which is difficult to practically implement. Thus, we ignore this assumption in our project.
5) Error terms must follow normal distribution: This assumption focus on the distribution of the error terms and checks if it meets the criteria of normal distribution. This is checked by the Normal Q-Q plot in our figure (a). It is seen that the current model follows normality as all the error terms follow the 45 degree line. Thus, we confirm normal distribution.
6) No influential data points: This assumption assumes that the linear model is not influenced from any outlier data point and is a generalized model. This assumption is checked by using cook's distance, which is the difference of distance of a data point when it is present and when it is not present in the model. If the cook's distance is more than 1, then it is considered as high leverage data point. We can check this in the Residuals vs Leverage plot in the figure (a). No data point has cook's distance greater than 1, thus there is no influential data point in our model.
7) No multicollinearity: This assumption checks if the independent variables in the linear model are correlated or possible functions of each other. For a linear model, independent variables must be not correlated or functions of each other. This can be checked by using Variance Inflation Factor(VIF) which checks the commonness in variance between the all the predictors. If the VIF score is less than 10 then there is no presence of multicollinearity. But in our current model, 5 out of 9 predictors have the VIF score more than 10. This was expected in the model building stage, but at that stage our aim was to build a best performing model. Now, to rectify this issue we remove each variable with highest VIF one by one by checking VIF at each stage. By doing so, we again come back to a model with **all the initial four variables** with no interactions and polynomial variables in it. Now, the VIF for each predictor variable is back to values around 1.



(a) Before Diagnostics

(b) After Diagnostics

After performing these iterations for removing, heteroscedasticity, and multicollinear variables, the current model is consist of square root of Mortality Rate as our dependent variable and 4 initial best subset independent variables. But, it is seen that in this model, variable "Early_initial_breastfeed" is insignificant with p-value=0.212. Thus, we drop that variable. On this current model, we again check for linearity, homoscedasticity, no autocorrelation, normality and no influential

```
> ncvTest(final_model)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 3.196522, Df = 1, p = 0.073795
```

(a) NCV Test

```
> durbinWatsonTest(final_model)
 lag Autocorrelation D-W Statistic p-value
   1       -0.146533      2.288967   0.212
 Alternative hypothesis: rho != 0
```

(b) Durbin Watson Test

```
> vif(final_model)
Infants_provided_minimum_accepted_diet              Health_Service_Coverage
                              1.865198                             1.065160
                    Underweight_children
                              1.778687
```

(c) VIF for predictors

points assumptions. The tests scores are: ncvTest=0.073, durbinWatson=2.28, cook's distances are less than 0.5. We can see from the figure that all the assumptions are met with our final generalized model.

VI. Summary of the Final Model and Conclusion

After a number of iterations, addition and removal of variable terms, and checking for all the assumptions, the multiple linear regression model is finalized. This model is our generalized model meeting all the assumptions of linearity and the least square method. By performing all the diagnostics and iterations in the variable selection, the final model is said to be best linear unbiased estimator based on the current available data. Also, all from 5 total independent variables at the start of the model is brought down to 3 variables with high performance value. This also meets the principle of parsimony.

Summary and interpretation of the generalized multi-linear regression model shown in the figure 6:

```
Residuals:
     Min       1Q   Median       3Q      Max
-2.65048 -0.94925 -0.03637  0.87656  2.77027

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                         8.680386   0.950215   9.135 9.25e-14 ***
Infants_provided_minimum_accepted_diet -0.046076   0.008354  -5.515 4.87e-07 ***
Health_Service_Coverage             -0.055022   0.020641  -2.666  0.00943 **
Underweight_children                 0.128710   0.021451   6.000 6.72e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.352 on 74 degrees of freedom
Multiple R-squared:  0.7495,    Adjusted R-squared:  0.7394
F-statistic: 73.82 on 3 and 74 DF,  p-value: < 2.2e-16
```

Fig. 6: Final Model Summary

A. Variable Coefficients interpretation

1) Infants_provided_minimum_accepted_diet: For a unit change, the square root of Mortality rate decreases by 0.046 units when the other variables are constant.

2) Health_Service_Coverage: For a unit change, the square root of Mortality rate decreases by 0.055 units when the other variables are constant.

3) Underweight_children: For a unit change, the square root of Mortality rate increases by 0.128 units when the other variables are constant.

B. Performance

The final generalized model has the adjusted R^2 value of 0.7394, i.e. goodness of fit of 73.94%, which conveys 73.94% of variablilty of the dependent variable is caused due to our selected set of independent variables. The residual standard error of the model is 1.35, that means for every estimation an average of 1.35 units of error is likely to occur in our predicted response variable. All the individual variables are significant in our generalized model.

C. Conclusion

Based on the back and forth of variable selection process and meeting all the assumptions of the Ordinary least square method of linear regression, we conclude that, based on the available data, this build of multi-linear regression model is a generalized best fit model for predicting child mortality with the given set of predictor variables

References

[1] LIND, D. A., MARCHAL, W. G., WATHEN, S. A. (2008). Statistical techniques in business economics. Boston, McGraw-Hill/Irwin.
[2] "World Health Data Platform - WHO," World Health Organization. [Online]. Available: https://www.who.int/data.