Applications of Multiple Linear Regression in Life Expectancy and Mortality

Elise Dixon and Adam Heiss

Department of Statistics, Florida State University

April 27, 2023

Abstract

Utilizing multiple linear regression on data collected from the United Nations and the World Health Organization, we investigate relationships between health status, socioeconomic factors, and morbidity toward the purpose of determining life expectancy using data from countries around the world. Significant predictors toward each category (respectively) included thinness at specific ages, income composition, and cardiovascular disease.

Though we initially intended to produce a single model with respect to level of development, development status led to such significant impacts that it was necessary to develop a model for each stratification, developed and developing. Though results of analysis were initially unstable and difficult to interpret, we saw significant improvement upon splitting the data and treating each model separately.

1 Introduction

As health science enters a new epoch of rapid evolution, perhaps the most remarkable development seen throughout recent years is the continuously accelerating increase in overall life expectancy. The most common tools for regression analysis on health-related data are linear regression, ridge regression, LASSO regression, and polynomial regression. We seek to identify qualities of health, socioeconomic status, and mortality that have strongly significant impacts on life span.

Utilizing multiple linear regression with variable selection methods such as significance level and BIC, performing diagnostic analysis on residuals, and conducting necessary variable transformations, we implement a multiple linear regression analysis on the life expectancy of

countries to detect the most significant predictors. Countries are classified by the WHO according to status, indicating their level of development, toward producing the most prevalent results. Predictors of life expectancy were found to be significant across wide-ranging topics; morbidity and mortality was primarily represented by cause of death, aspects of socioeconomic status were accounted for in predictors like income composition, total expenditure on health care, and schooling, and underlying health factors such as BMI.

1.1 Data Collection

Two datasets were utilized; one is produced by the World Health Organization (WHO) and United Nations (UN) and accessed from Kaggle (https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who). This data was collected throughout 2000-2015. Though collection was performed in the original dataset for all countries, missing data is present for lesser-known countries, some of which were ultimately excluded. The WHO tracks data related to health, economic aspects, and other associated factors through the Global Health Observatory (GHO) data repository.

Another set is utilized from Our World in Data and likewise accessed from Kaggle (https://www.kaggle.com/datasets/iamsouravbanerjee/cause-of-deaths-around-the-world). This set is comprised of data related to mortality and morbidity from 1990-2019. Data includes counts of wide-ranging causes of death by country and year. The data quantifies mortality and morbidity by DALYs (Disability Adjusted Life Years), the scale of which corresponds to the equivalent of one year of life, measuring the impact of 'lost health' and disease by years of life lost, allowing for quantifiable comparison of the impacts of disease across different demographic factors.

1.2 Preliminary Data Processing

We compiled the data with a cross section of overlapping countries and years which are included in both sets. All predictors were initially investigated through a correlation matrix with the purpose of determining strong linear relationships, detecting multicollinearity between predictors, and nonlinear relationships warranting further investigation. After preliminary visual inspection, variables of interest toward model selection were chosen based on their suspected significance regarding life expectancy impacts. The data was primarily numeric, so little cleaning was necessary besides concatenation by country and year.

After significance of potential predictors was assessed, predictors of interest with nonlinear relationships to the response (or other model violations) were transformed to fit model assumptions. Many variables were skewed right, a phenomenon common to health data analysis, and were corrected with a log transformation; these variables included adult mortality, alcohol, drowning, HIV/AIDS, liver disease, malaria, maternal disorders, Parkinson's, poisonings, tuberculosis, thinness from 5-9 ears, and deaths under 5 years. A value of '1' was added to these variables to account for missing values to adjust for the log-scale:

$$x_i = \log(x_i + 1)$$

Diphtheria, Hepatitis B, and Polio were raised to the fourth power and divided by 10,000,000 to combat left-skewness:

$$x_i = \frac{{x_i}^4}{10{,}000{,}000}$$

HIV/AIDS Prevalence was the only variable whose relationship could not be linearized; this variable was omitted from the analysis and replaced by a different HIV/AIDS metric. Finally, countries were classified by status with respect to their level of development. Status was selected as the classification measure over level of development since observations were much more evenly distributed across groups.

2 Methodology

2.1 Variable Selection Criteria

The original data consisted of many predictor variables and various standard selection methods initially produced unstable results. Since results were highly unstable at first, it was difficult to discern the impact of each predictor on the model and detect a best fit. To correct for this, we first determined that development status ('developed' vs. 'developing') was impacting variable selection so significantly that it would be difficult to produce relevant results for both classifications with one model. After splitting the data and creating a different model for each of the two classifications, variable selection became much more consistent (although still somewhat variable).

To further combat these concerns and produce the best fit possible, we conducted 1000 replications to our initial variable selection process for the most accurate final predictor selection with significance level as the entry criteria (≤ 0.01). The stopping criteria utilized was the Bayesian Information Criterion (BIC, Schwartz, 1978). The BIC is a criterion for model selection which utilizes aspects of Bayesian inference by suggesting a model which uses information from the given data. The goal is to select a model which minimizes BIC; it decreases inversely with likelihood but is penalized with the addition of predictors to counterbalance model overfitting.

The k-BIC algorithm, enumerated in the StrandNGS Reference Manual by Streamlining NGS Data Management and Analysis, attempts to balance the number of observations (n) against the number of parameters (k) by maximizing the likelihood function $L(\theta)$. It is easy to increase the log-likelihood with the addition of parameters (predictors), but this leads to biases resulting from over-fitting the model. The information criteria adjusts the likelihood function with the addition of a penalty to prevent an overfit. The general form of the IC is:

$$IC(n,p) = -2\log(likelihood)\left(x\middle|\hat{\theta}_p\right) + \alpha(n,p)$$

where $\hat{\theta}_p$ is the maximum likelihood estimates of the parameters, and α is the penalty function. For BIC, this penalty term is:

$$\alpha(n,p) = BIC(n,p) = plog(n)$$

The number of parameters is $p = 2k$, consisting of the k normal means $\mu_1, \dots, \mu_k$, k-1 proportions of elements in each cluster, $p_1, \dots, p_j$ with $\sum_{j=1}^{k} p_j = 1$, and standard deviation σ. The likelihood is:

$$\prod_{i=1}^{n} f(x_i; \Theta) = \prod_{i=1}^{n} \prod_{j=1}^{k} p_j \phi(x; \mu_i, \sigma)$$

where $\phi(x; \mu_i, \sigma)$ is the probability density.

In the BIC algorithm, means of each cluster, their proportions, and a common standard deviation are calculated. The log-liklihood is computed with sample means $\bar{x}_i$, sample standard deviations:

$$l(k) = \sum_{m=1}^{n} \log[\sum_{i=1}^{k} p_i f_i(x_m; \bar{x}_i, s)]$$

BIC(k) is then calculated as

$$BIC(k) = -2l + 2kln(n)$$

where the optimal number of clusters is

$$\hat{k} = -2\ln(\hat{\theta}_k | y) + kln(n)$$

("k-BIC", StrandNGS Reference Manual).

The BIC is computed for each candidate model and the model with the smallest BIC is selected (Neath, 2011). BIC was used for selection as opposed to AIC as the large number of predictors included in the original data required a higher degree of selectivity. The final model selected was the one with the lowest validation average squared error, as can be seen by figures (2.1) and (2.2) for developed and developing models, respectively.

Figure (2.1)

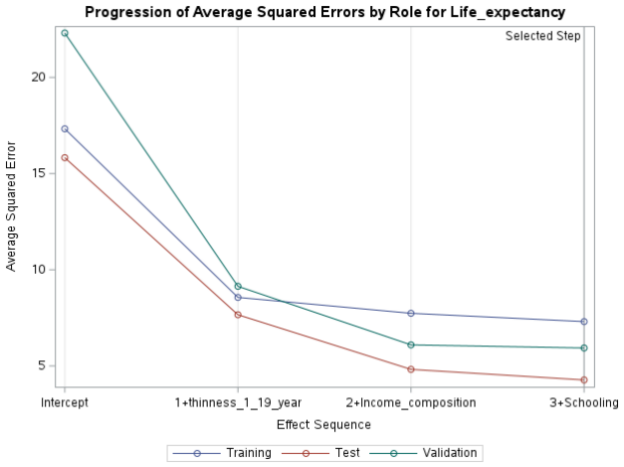
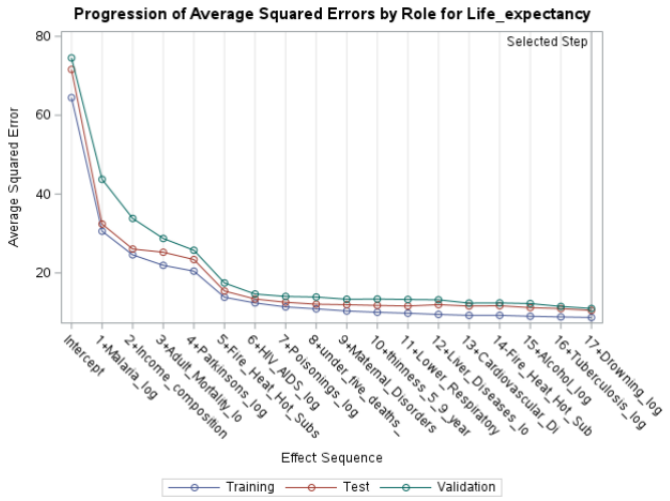


Figure (2.2)



2.2 Model Fitting

The model utilized is the generalized multiple linear regression model with response Y, life expectancy in years, and predictors $X_1, \ldots, X_n$:

$$E[Y|X = x] = B_0 + B_1 x_1 + \cdots + B_k x_k.$$

Assumptions include a linear mean function, constant variance, and independence of errors. Linearity was met by transformations on predictors noted previously, and visual inspection of preliminary plots indicated independence of errors. After data transformations to ensure model assumptions were met, the approach taken toward initial variable selection and model fitting was to perform a significance-level selection process using BIC as the stopping criteria with replications. This process was conducted separately for each status 'developed' or 'developing' and resulted in a model fit for each.

Bootstrapping is a method toward producing the most accurate coefficient estimates toward minimizing prediction error by resampling existing data to create several samples (Shao, 1996). Again with 1000 replications, the coefficient estimates for each model were bootstrapped to further minimize the variance observed in initial variable selection and model fitting. The assumption of constant variance was somewhat relaxed by the application of bootstrapping methods to coefficient estimation. The coefficient progression for each model can be seen in figures (2.3) for developed countries, and (2.4) for developing countries.

The final predictors suggested for the model resulting from this procedure were:
Model A (Developed): Thinness (from 1-19 years), income composition, and schooling
Model B (Developing): Malaria, income composition, adult mortality, Parkinson's, HIV/AIDS, poisonings, deaths under 5 years (per 1000), maternal disorders, thinness from 5-9 years, lower respiratory infection, liver disease, and cardiovascular disease.

Figure (2.3)

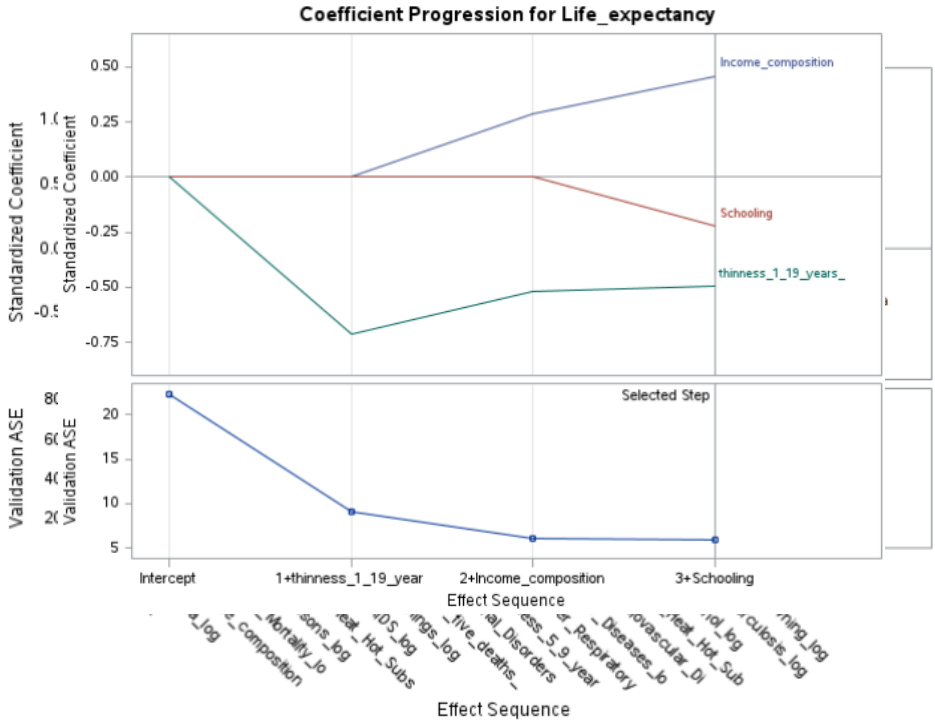


Figure (2.4)

2.3 Validation

To further stabilize predictor coefficients and selection, out-of-sample holdout model validation was utilized in combination with bootstrapping to produce a model that was as robust as possible. The data was partitioned into three groups: 50% for training, 25% for validation, and 25% for testing. The training set is a subset of the complete data in which relationships between the predictors and response are initially investigated, the validation set measures the accuracy with which the model produced from the training data assessed relationships, and the holdout data is the final test of accuracy for the developed model.

The proportion of data, or folds, used in each group was randomized on each replication, and coefficient estimates were bootstrapped and compared. In cross-validation, a model developed with all other folds is tested on the holdout information which hasn't been processed yet, and all folds are tested once, as seen in figure (2.5) ("Holdout Data and Cross-Validation"). The goal of model and coefficient validation was to produce the most robust results possible toward achieving accuracy and consistency. In addition, we observed several data values that seemed improbable at best, many of which were high-leverage outliers; cross-validation helps minimize the impacts of the bad data, and any of their associated, underlying trends.

Figure (2.5) exhibits the impact of the validation process for the developed-country model (A), as even with the more restrictive selection criteria BIC, suggested models included too many variables leading to an over-fit, which is then corrected by the ASE validation.

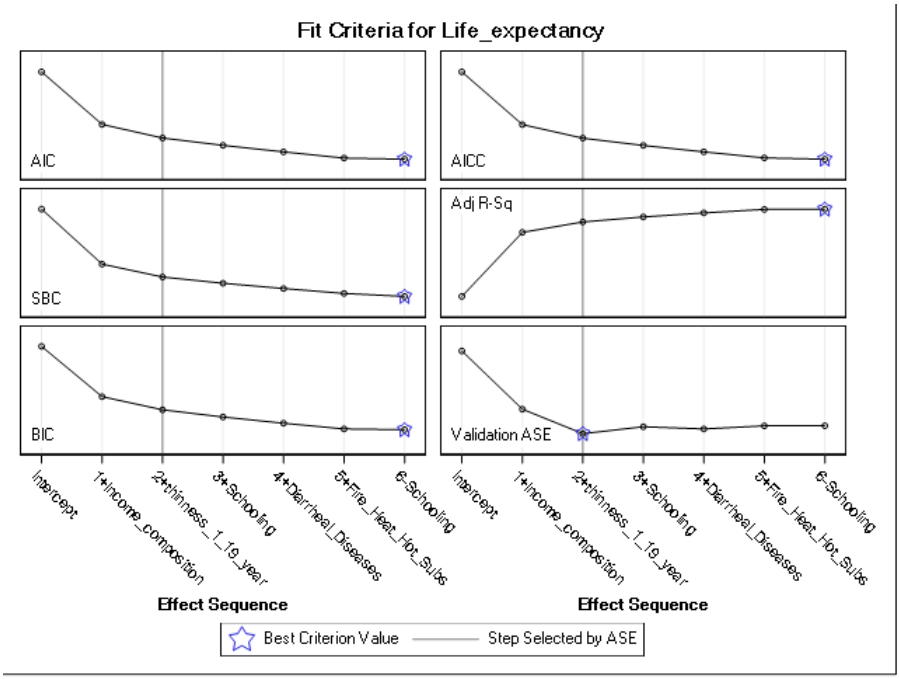


Figure (2.5)

3 Results and Interpretation

The final models for each development status are as follows:

Model A:

Model B:

We can utilize these models toward interpreting the strength of the impacts that health status, lifestyle choices, and socioeconomic status have on life expectancy. In addition, these models can be used in predicting general life expectancy based on the development status of a country, and understanding the differences between the major health concerns of various regions.

The coefficient estimate was negative for most factors, which was expected as many of the factors are related to morbidity and health concerns. The estimate for income composition was positive for both models, also as expected, as life expectancy increases with income. However, the coefficient estimate for Parkinson's(log) in the developing country model had an unexpectedly high positive coefficient, 4.71. After investigating the potential for redundancy verses real suppression by comparing coefficient estimates and the adjusted R-squared of models with the removal of Parkinson's(log) to the original model, we determined that Parkinson's(log) was not redundant or a real suppression, as neither the signs of coefficient estimates for other predictors nor the R-squared changed. Thus, we concluded that the unexpected positive coefficient estimate for Parkinson's(log) results from the high average age of diagnosis (60); as the life expectancy for developing countries is much lower than that of developed, it is possible

that if a subject survives long enough to be diagnosed with Parkinson's disease, they are already above the average life expectancy.

Tables (3.1) and (3.2) summarize the regression outputs for final models selected.

Analysis of Fit Criteria for Developed Countries								
Label1	N Obs	Mean	Std Dev	Minimum	Lower Quartile	Median	Upper Quartile	Maximum
ASE (Test)	1000	6.59	1.59	2.93	5.46	6.54	7.62	12.81
ASE (Train)	1000	5.95	0.99	3.00	5.27	5.96	6.62	9.44
ASE (Validate)	1000	6.62	1.62	2.34	5.43	6.50	7.70	12.69
Adj R-Sq	1000	0.66	0.05	0.51	0.63	0.66	0.69	0.82
BIC	1000	221.19	24.81	130.36	204.54	221.69	237.14	292.70
R-Square	1000	0.67	0.05	0.52	0.64	0.67	0.70	0.82
Root MSE	1000	2.47	0.20	1.77	2.34	2.48	2.61	3.10

Table (3.1)

Analysis of Fit Criteria for Developing Countries								
Label1	N Obs	Mean	Std Dev	Minimum	Lower Quartile	Median	Upper Quartile	Maximum
ASE (Test)	1000	11.52	1.70	8.03	10.38	11.31	12.48	21.18
ASE (Train)	1000	10.75	1.58	7.31	9.62	10.51	11.73	18.02
ASE (Validate)	1000	11.59	1.74	7.96	10.42	11.39	12.50	21.90
Adj R-Sq	1000	0.84	0.02	0.74	0.83	0.85	0.86	0.89
BIC	1000	1620.57	98.52	1362.52	1553.31	1610.54	1686.07	1992.91
R-Square	1000	0.84	0.02	0.74	0.83	0.85	0.86	0.89
Root MSE	1000	3.30	0.23	2.74	3.13	3.27	3.45	4.27

Table (3.2)

4 Conclusion

To study the impacts of predictors from a wide range of topics on life expectancy, we apply multiple linear regression with variable selection based on significance level as selection criteria and BIC as stopping criteria. In producing the model, we transformed predictors to fit assumptions and replicated variable selection utilizing BIC, and utilized methods of bootstrapping and cross-validation to further adjust parameter estimates to minimize the impact

of bad data values, missing data, and outliers. Bootstrapping also aided in combatting nonconstant variance. By applying methods of bootstrapping and model validation, we attempted to produce the most consistent, effective, and robust model possible toward accurate analysis of the data.

Of the initial 53 predictors, 3 were selected for the developed country model and 12 were selected for the developing country model. The adjusted R-square, or coefficient of determination, was 66% for the developed country model and 84% for the developing country model. Due to the high number of initial predictors and the potential for health-related data to contain multicollinearity, we were quite restrictive in our variable selection; however, the higher adjusted R-squared value for the developing country model suggests that many of these variables do contain important information. Since development status had such significant impacts, future analysis points toward further investigation into the effects of region, level of development, and socioeconomic status on health impacts and life expectancy.

References

"Holdout Data and Cross-Validation." Qlik Help, Qlik International AB, https://help.qlik.com/en-US/cloud-services/Subsystems/Hub/Content/Sense_Hub/AutoML/holdout-crossvalidation.htm [Article, Photo].

Neath, Andrew. "The Bayesian Information Criterion: Background, Derivation, and Applications." Wiley Interdisciplinary Reviews, WIREs Computational Statistics, Vol. 4 no. 2, Wiley Online Library, 14 Dec. 2011, https://wires.onlinelibrary.wiley.com/doi/full/10.1002/wics.199.

Shao, Jun. "Bootstrap Model Selection." Journal of the American Statistical Association, Vol. 91, No. 434, Jun. 1996, https://www.jstor.org/stable/2291661.

"k-BIC." StrandNGS Reference Manual, Streamlining NGS Data Management & Analytics, Strand Life Sciences Private Limited, https://www.strand-ngs.com/files/manual/reference/appx_kbic.html.

Appendix

Model A: Developed

Analysis Variable : Estimate								
Parameter	N Obs	Mean	Std Dev	Minimum	Lower Quartile	Median	Upper Quartile	Maximum
Cardiovascular_Disea	16	-3.45	1.69	-7.21	-4.71	-2.80	-2.15	-1.81
Chronic_Kidney_Disea	35	1.53	0.76	0.34	0.49	1.71	2.01	2.75
Chronic_Respiratory_	1	1.87	.	1.87	1.87	1.87	1.87	1.87
Dementias_log	13	2.92	1.36	1.26	1.44	3.10	3.42	6.16
Diarrheal_Diseases_l	86	0.76	0.35	0.33	0.40	0.83	1.06	1.49
Digestive_Diseases_l	1	-2.62	.	-2.62	-2.62	-2.62	-2.62	-2.62
Drug_Use_Disorders_l	2	0.56	0.02	0.55	0.55	0.56	0.57	0.57
Fire_Heat_Hot_Substa	2	-0.83	0.00	-0.83	-0.83	-0.83	-0.82	-0.82
Forces_of_Nature_log	41	0.56	0.08	0.45	0.52	0.56	0.60	0.76
HIV_AIDS_log	4	0.49	0.21	0.36	0.37	0.40	0.61	0.81
Heat_Cold_Exposure_l	3	-1.14	0.30	-1.47	-1.47	-1.07	-0.87	-0.87
Hepatitis_B2	1	0.34	.	0.34	0.34	0.34	0.34	0.34
Income_composition	990	38.00	8.63	3.16	32.59	37.35	43.87	59.43
Intercept	1000	54.72	6.38	40.32	50.92	53.99	57.13	104.96
Interpersonal_Violen	42	-1.15	0.38	-2.31	-1.14	-1.06	-0.90	-0.73
Maternal_Disorders_l	1	-1.00	.	-1.00	-1.00	-1.00	-1.00	-1.00
Measles_log	1	0.26	.	0.26	0.26	0.26	0.26	0.26
Neoplasms_log	1	-5.87	.	-5.87	-5.87	-5.87	-5.87	-5.87
Nutritional_Deficien	1	-1.05	.	-1.05	-1.05	-1.05	-1.05	-1.05
Parkinsons_log	2	5.99	2.00	4.58	4.58	5.99	7.40	7.40
Poisonings_log	1	0.56	.	0.56	0.56	0.56	0.56	0.56
Polio2	3	-0.45	0.02	-0.47	-0.47	-0.45	-0.43	-0.43
Population_log	1	0.32	.	0.32	0.32	0.32	0.32	0.32
Protein_Energy_Malnu	1	-1.05	.	-1.05	-1.05	-1.05	-1.05	-1.05
Road_Injuries_log	18	-1.38	0.29	-1.71	-1.61	-1.45	-1.18	-0.70
Schooling	341	-0.54	0.11	-0.97	-0.60	-0.53	-0.48	0.02
thinness_1_19_years_	985	-5.62	0.87	-9.95	-6.15	-5.73	-5.19	-1.62
thinness_5_9_years_l	9	-5.17	0.50	-6.05	-5.54	-5.13	-4.82	-4.52

Analysis Variable : nValue1								
Label1	N Obs	Mean	Std Dev	Minimum	Lower Quartile	Median	Upper Quartile	Maximum
AIC	1000	344.67	30.01	236.06	324.70	344.58	364.54	441.42
AICC	1000	345.13	29.94	236.71	325.14	345.04	364.87	441.59
ASE (Test)	1000	6.59	1.59	2.93	5.46	6.54	7.62	12.81
ASE (Train)	1000	5.95	0.99	3.00	5.27	5.96	6.62	9.44
ASE (Validate)	1000	6.62	1.62	2.34	5.43	6.50	7.70	12.69
Adj R-Sq	1000	0.66	0.05	0.51	0.63	0.66	0.69	0.82
BIC	1000	221.19	24.81	130.36	204.54	221.69	237.14	292.70
C(p)	1000	66.92	33.62	0.45	43.09	63.37	83.93	266.95
Dependent Mean	1000	78.68	0.27	77.60	78.50	78.68	78.86	79.56
R-Square	1000	0.67	0.05	0.52	0.64	0.67	0.70	0.82
Root MSE	1000	2.47	0.20	1.77	2.34	2.48	2.61	3.10
SBC	1000	231.64	24.26	145.44	216.16	232.19	248.38	299.39

Model B: Developing

Analysis Variable : Estimate								
Parameter	N Obs	Mean	Std Dev	Minimum	Lower Quartile	Median	Upper Quartile	Maximum
Adult_Mortality_log	636	-0.47	0.12	-1.07	-0.55	-0.45	-0.38	-0.26
Alcohol_Use_Disorder	61	0.70	0.21	0.10	0.61	0.73	0.83	1.08
Alcohol_log	465	-1.07	0.19	-1.72	-1.20	-1.07	-0.92	-0.62
Cardiovascular_Disea	600	-2.37	0.46	-3.97	-2.66	-2.37	-2.03	-1.24
Chronic_Kidney_Disea	54	1.48	0.39	0.50	1.17	1.59	1.78	2.05
Chronic_Respiratory_	2	-0.86	0.01	-0.87	-0.87	-0.86	-0.85	-0.85
Dementias_log	259	2.26	0.97	0.60	1.56	2.14	2.96	4.55
Diabetes_Mellitus_lo	1	-0.87	.	-0.87	-0.87	-0.87	-0.87	-0.87
Diarrheal_Diseases_l	2	0.38	0.02	0.37	0.37	0.38	0.40	0.40
Digestive_Diseases_l	13	-3.55	0.74	-5.04	-3.81	-3.62	-3.19	-1.96
Diphtheria2	18	0.18	0.03	0.12	0.16	0.19	0.21	0.23
Drowning_log	399	1.18	0.21	0.58	1.02	1.17	1.31	1.86
Drug_Use_Disorders_l	2	0.64	0.01	0.63	0.63	0.64	0.65	0.65
Fire_Heat_Hot_Substa	873	-1.20	0.69	-3.21	-1.69	-0.88	-0.67	-0.03

Analysis Variable : Estimate								
Parameter	N Obs	Mean	Std Dev	Minimum	Lower Quartile	Median	Upper Quartile	Maximum
Forces_of_Nature_log	48	0.35	0.12	0.15	0.27	0.35	0.42	0.73
GDP_log	19	-0.72	0.11	-0.90	-0.80	-0.67	-0.63	-0.57
HIV_AIDS_log	972	-0.74	0.15	-1.25	-0.84	-0.73	-0.63	-0.35
Heat_Cold_Exposure_l	186	-0.81	0.20	-1.31	-0.93	-0.81	-0.68	-0.29
Hepatitis_B2	1	0.13	.	0.13	0.13	0.13	0.13	0.13
Income_composition	883	5.01	1.28	2.26	4.11	4.84	5.67	12.03
Intercept	1000	70.68	5.74	54.89	66.14	71.58	75.28	83.90
Interpersonal_Violen	40	-0.60	0.11	-0.88	-0.66	-0.58	-0.53	-0.41
Liver_Diseases_log	389	1.70	0.51	0.89	1.42	1.65	1.86	4.77
Lower_Respiratory_In	317	-1.31	0.21	-1.94	-1.46	-1.29	-1.15	-0.83
Malaria_log	930	-0.27	0.11	-0.70	-0.35	-0.25	-0.18	0.11
Maternal_Disorders_l	307	1.05	0.25	0.42	0.86	1.02	1.22	1.71
Measles_log	59	-0.19	0.05	-0.34	-0.22	-0.18	-0.15	-0.10
Meningitis_log	248	-1.61	0.62	-3.78	-1.79	-1.45	-1.18	-0.51
Neonatal_Disorders_l	4	0.12	0.99	-0.81	-0.74	0.11	0.97	1.05
Neoplasms_log	19	-2.15	0.58	-3.50	-2.36	-2.04	-1.84	-1.21
Nutritional_Deficien	2	0.63	0.38	0.36	0.36	0.63	0.90	0.90
Parkinsons_log	906	4.71	1.07	0.85	4.10	4.86	5.52	7.28
Poisonings_log	483	-0.92	0.21	-1.99	-1.04	-0.92	-0.80	-0.24
Polio2	1	0.14	.	0.14	0.14	0.14	0.14	0.14
Protein_Energy_Malnu	1	0.33	.	0.33	0.33	0.33	0.33	0.33
Road_Injuries_log	2	0.83	0.20	0.69	0.69	0.83	0.97	0.97
Schooling	35	0.20	0.07	0.12	0.15	0.19	0.24	0.43
Self_harm_log	13	-0.84	0.22	-1.23	-1.00	-0.78	-0.71	-0.53
Total_expenditure	17	-0.21	0.04	-0.28	-0.24	-0.21	-0.19	-0.16
Tuberculosis_log	614	-1.05	0.31	-2.88	-1.20	-1.08	-0.90	-0.26
infant_deaths_log	210	3.77	0.68	2.35	3.24	3.67	4.15	6.33
percentage_expenditu	113	0.42	0.23	0.22	0.29	0.33	0.41	1.20
thinness_1_19_years_	70	-0.94	0.17	-1.28	-1.06	-0.97	-0.80	-0.60
thinness_5_9_years_l	69	-1.02	0.24	-1.63	-1.19	-1.04	-0.83	-0.57
under_five_deaths_lo	694	-2.75	1.84	-7.43	-4.79	-1.85	-1.42	-0.76

Analysis Variable : nValue1								
Label1	N Obs	Mean	Std Dev	Minimum	Lower Quartile	Median	Upper Quartile	Maximum
AIC	1000	2299.88	108.95	2004.46	2225.02	2292.17	2365.23	2718.07
AICC	1000	2300.48	108.73	2006.01	2225.40	2292.91	2365.81	2718.15
ASE (Test)	1000	11.52	1.70	8.03	10.38	11.31	12.48	21.18
ASE (Train)	1000	10.75	1.58	7.31	9.62	10.51	11.73	18.02
ASE (Validate)	1000	11.59	1.74	7.96	10.42	11.39	12.50	21.90
Adj R-Sq	1000	0.84	0.02	0.74	0.83	0.85	0.86	0.89
BIC	1000	1620.57	98.52	1362.52	1553.31	1610.54	1686.07	1992.91
C(p)	1000	181.35	108.69	19.86	97.08	159.35	238.36	682.30
Dependent Mean	1000	67.80	0.22	67.09	67.65	67.79	67.95	68.46
R-Square	1000	0.84	0.02	0.74	0.83	0.85	0.86	0.89
Root MSE	1000	3.30	0.23	2.74	3.13	3.27	3.45	4.27
SBC	1000	1676.51	87.61	1441.28	1617.22	1670.39	1731.13	2013.39