Project 3:
Examining Life Expectancy of Males Using Multiple Regression

Chad Chapnick

May 2, 2016

Introduction

Increasing life expectancy ranks as one of society's greatest achievements during the 20th century. This progressive increase in survival can be attributed to a number of factors, including a reduction in infant mortality rate, improved living styles and education, as well as advancements in health, medicine and nutrition. Although life expectancy at birth has steadily increased globally, not all regions have shared these improvements. One example of this exception is decline in life expectancy in many parts of Africa due the HIV/AIDS epidemic. According to the WHO, sub–Saharan Africa is the most affected region, with 25.8 million people living with HIV in 2014 (World Health Organization, 2015).

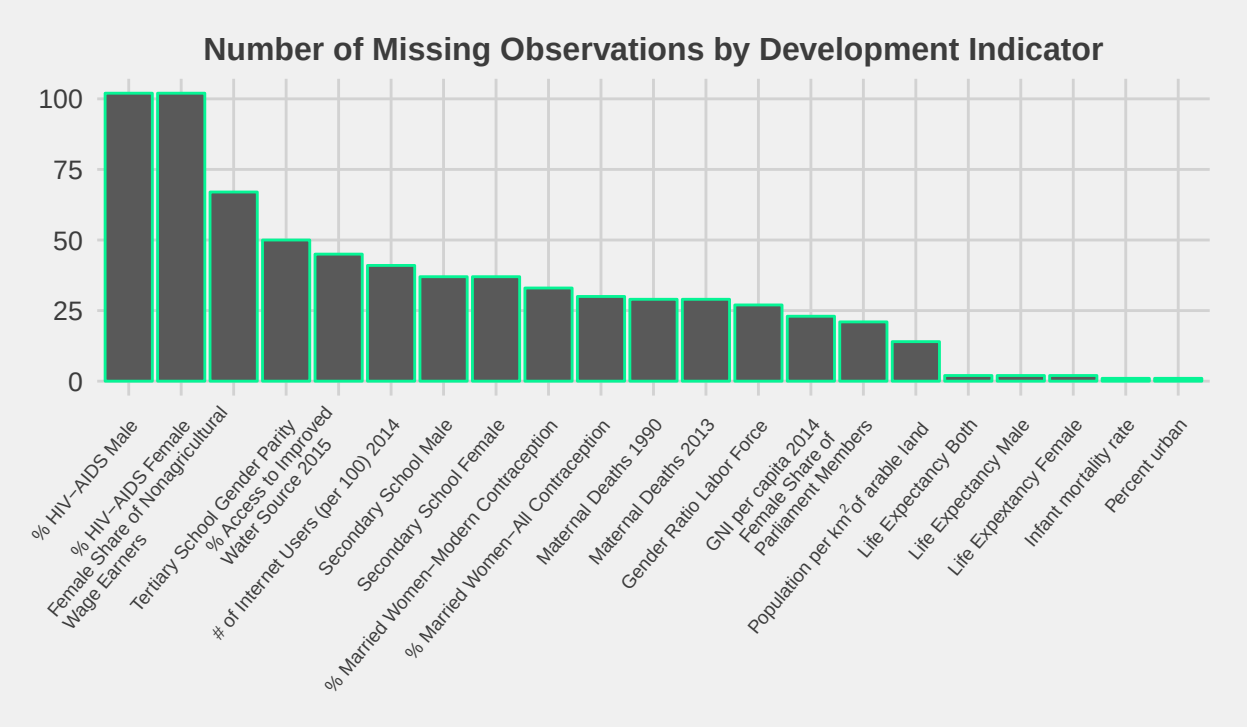
In this analysis, we consider male life expectancy at birth, a preferred health indicator in epidemiological studies (Arias, 2014; World Health Organization, 2014). Using regression analysis, we attempt to build a statistical model to understand the relationship between life expectancy and a number of environmental, economic and social variables at a global level. In order to compare these countries, internationally standardized data were obtained from the Population Reference Bureau and the World Bank. These organizations are known for delivering the most current and accurate global development data from officially-recognized sources such as the United Nations, the U.S. Census Bureau, and the Organization for Economic Co-operation and Development. Featured below are five randomly sampled countries (out of 210) and the first five columns of the data set:

Table 1: Sample of the Data

	COUNTRY	REGION	CONTINENT	LifeExp_Both	LifeExp_Male
3	libya	NORTHERN AFRICA	AFRICA	71	69
72	curacao	CARIBBEAN	NORTH AMERICA	78	75
103	iraq	WESTERN ASIA	ASIA	69	67
144	japan	EAST ASIA	ASIA	83	80
145	korea north	EAST ASIA	ASIA	70	66
206	samoa	OCEANIA	AUSTRALIA	74	73

Missing Data

Before beginning the regression analysis, it is important to acknowledge that data availability varies by country, and there is a substantial amount of missing data in the data set. This fact raises the question of whether or not certain groups are more likely to have missing values. Moreover, could missing values be a helpful explanatory variable in modeling the life expectancy of males at birth? One method of dealing with missing observations in regression analysis is to remove all incomplete observations and perform the oridnary-least squares (OLS) technique to estimate the coefficients of the explanatory variables using the remaining observations (Haitovsky, 1968). In fact, classical regression in `R` excludes all observations which have missing values for any variable. For this data set, there are 160 countries out of the 210 in the data frame that contain missing values for some variable. Featured below is a plot showing the amout of missing values in each variable.



In the era of "big data", missing values are a common problem in almost any statistical analysis. By performing classical regression with the current data set, the model would be built using less than 25% of the countries and render the quality of the model inconclusive. Due to the hard work of many brilliant researchers, there are a number of imputation algorithms to replace missing values in a data set with some plausible values. However, when employing imputation algorithms we must have an understanding of the missing data mechanism in order to make reasonable assumptions. In particular, we want to know whether the missing values depend on unobserved values, or if the missingness of some observation X may depend on the observed part of X, but not the unobserved part. Little and Rubin define these *missingness mechanisms* as Missing at Random (MAR) and Missing Not at Random (MNAR), respectively (Little & Rubin, 1987). Although it is impossible to determine whether a data set is MNAR through the data itself, the sample can be divided into two groups—those with missing values and those without—then a hypothesis test can be performed for a difference in the mean of the predictor variable for the two groups (Soley-Bori, 2013).

Accordingly, a Wilcoxon rank sum test was performed in order to determine if there was a significant difference in the average life expectancy for countries which had missing data and for those that did not. This test was chosen so as to relax the assumtion that samples were drawn from normally distributed populations with unknown population means. The results of the test are shown in table 2. The following figure is a boxplot of life expectancy of males at birth for the two groups.

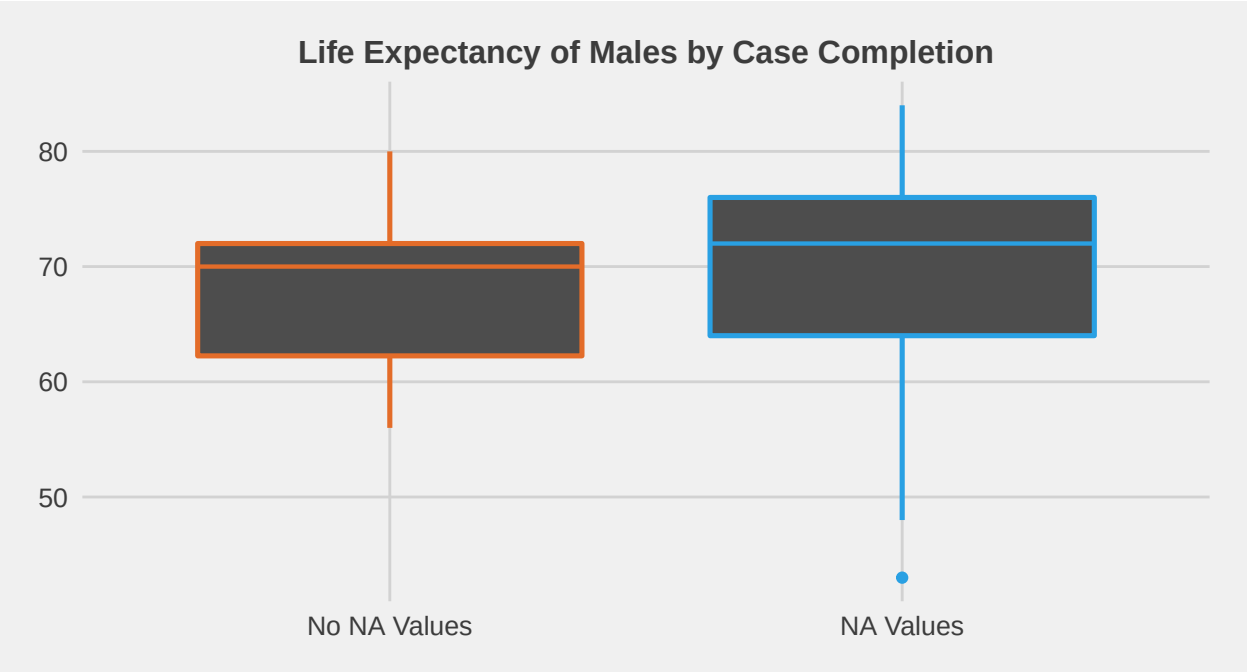


Table 2: Wilcoxon rank sum test with continuity correction: `LifeExp_Male` by `ISNA`

Test statistic	P value	Alternative hypothesis
3294	0.07711	two.sided

Lesotho, a country in Southern Africa, was identified as an outlier in the group which contained NA values, with a life expectancy of 43. This country has higher infant mortality, higher death rates, and lower population growth rates due to excessive proliferation of the AIDS epidemic (Central Intelligence Agency, 2015). The results of the wilcoxon test gave a p–value of 0.077 which, at the $\alpha = 0.05$ level, indicates that we fail to reject the null hypothesis and there is not a significant differnece in the mean of the two groups. Thus, it is reasonable to conclude that the MAR assumption is mostly fulfilled. One thing to note is that there were 50 countries without missing values and 160 countries which had one or more missing values. This may have ultimately affected the results of the test, and will be kept in mind as the analysis continues.

Another aspect which could contribute to male life expectancy at birth is geography. To compare the average male life expectancy across the six continents in the data set, a pairwise Wilcoxon rank sum test was performed for male life expectancy by continent.

Table 3: Pairwise Wilcoxon Rank Sum Tests

	AFRICA	ASIA	AUSTRALIA	EUROPE	NORTH AMERICA
ASIA	1.276e-10	NA	NA	NA	NA
AUSTRALIA	0.00008221	1	NA	NA	NA
EUROPE	9.573e-14	0.0007138	0.03387	NA	NA
NORTH AMERICA	6.621e-09	1	1	0.1026	NA
SOUTH AMERICA	0.0001084	1	1	0.02251	1

From our hypothesis tests, we can conclude that at the $\alpha = 0.05$ level, geography plays a significant role in determining life expectancy in most countries. This is not surprising considering the mean life expectancy in Africa is drastically lower.

K-Nearest Neighbor Imputation

To ameliorate the issue of missing data, we employ the k-nearest neighbor algorithm as an imputation method. The basic idea behind this method is to classify a new object, with input vector x, by examining the k closest data set points to x and assigning the object to the class that has the majority of points among these k (Hand, Mannila, & Smyth, 2001). In the context of missing-value imputation, there are many ways to use the observed values of the k-nearest neighbors. It is commmon to use a weighted average of the values of the neighbors, where the weights are given by $e^{-d(k,x)}$ where $d(k, x)$ is the euclidean distance between the case (x) with NAs and the neighbour k (Torgo, 2010). An important drawback of this approach is that the values derived from the model are usually more orderly than the actual values. This is because the missing values are predicted from a set of attributes, so the values are likely to be consistent with the particular set of attributes. One of the benefits of k-nearest neighbors is that it is a non parametric lazy learning algorithm, meaning that it does not make any assumptions on the underlying data distribution.

The main question which arises is how to determine the parameter k. Steven Buechler, the chair of the Department of Applied and Computational Mathematics and Statistics at the University of Notre Dame suggests that a rule of thumb in kNN is to pick k near the square root of the size of the training set (Buechler, 2014). Following this notion, a reasonable k value for this data set should be close to $\sqrt{210} \approx 14.5$. With this knowledge, the `knnImputation()` function in the DMwR package was used to fill in all NA values using weighted averages and a k value of 15. In general, larger k values such as this are less susceptible to noise in the data set when compared to single digit k values.

Before imputing the missing values, log-transformations were performed on a number of explanatory variables to reduce nonlinear relationships with our predictor variable. The relevant varaibles are listed below and a graph of male life expectancy vs. each variable (original and transformed) is shown in the supplementary material.

- Percent of Population with Age > 65
- GNI per Capita in 2014 (current \$US)
- Maternal Deaths in 1990 and in 2013

Calibration

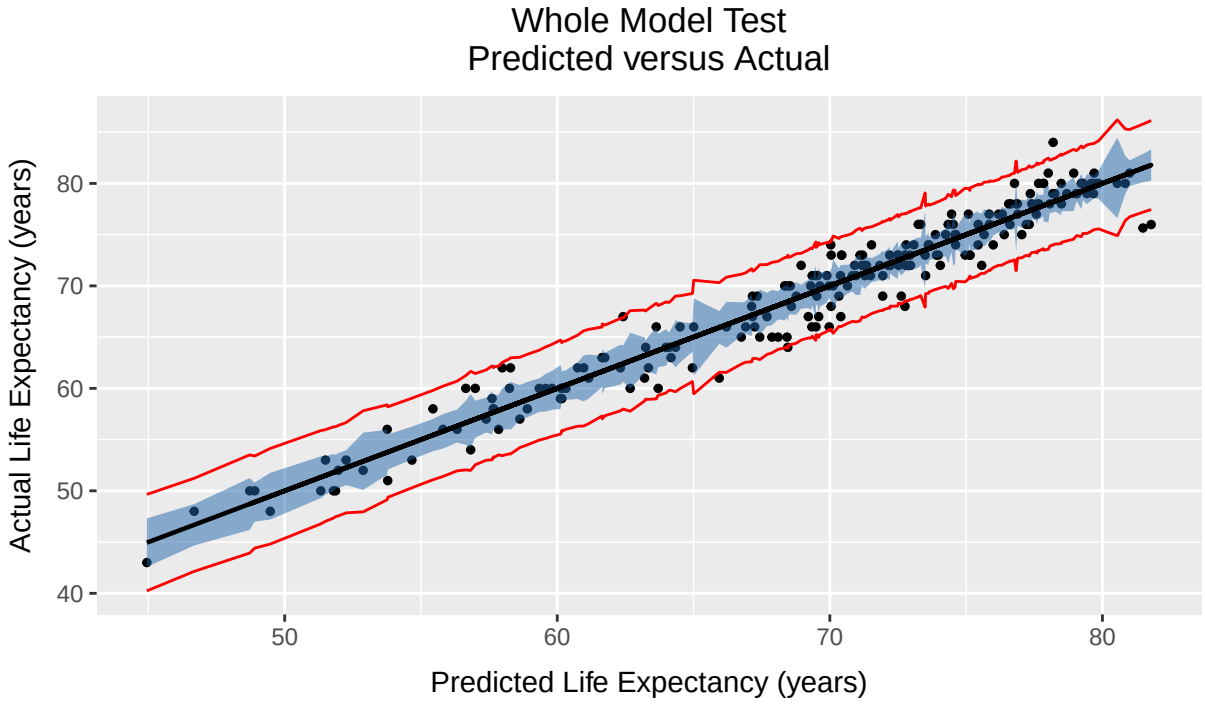
We began our regression anaylsis with explanatory variables which reflect the overall health, mortality, economic development and infrastructre of a country. To begin investigating the relationship between male life expectancy at birth an our explanatory variables, a linear regression model was built using 27 of the variables in the data set (excluding continent and region). The structural model can be written as:

$$E(Y|x_1, x_2, \ldots, x_k) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon, \epsilon \sim Normal(0, \sigma^2)$$

which is read as the expected value of the response variable given the values of the explanatory variables x_1 through x_k (Seltman, 2015). The ordinary least squares method was used to estimate the intercept, β_0 and the slopes, β_i, corresponding to each explanatory variable.

The summary statistics of the fitted linear model are shown in tables 5-7 in the appendix. Below is a plot of actual values of life expectancy versus the values predicted by the linear model and the corresponding 95% prediction and confidence bands.

In regards to the residuals, the median is near zero, and the minimum and maximum are roughly equal in absolute value which offers support to the assumption of a normally distributed error term. The multiple R-squared value is 0.95 indicating that the model reasonably explains the variability of male life expectancy at birth about its mean. With respect to the coefficients estimation (table 6) it is difficult to determine the accuracy of the results. This is due in part to relationships between varaiables, known as collinearity, and to inherent uncertainty when examining multiple variables resulting from type I errors (ie. incorrectly reject a null hypothesis that is true).



Refining the Model

Variance Inflation Factors

The absence of collinearity is a crucial assumption in multiple regression. If collinearity is ignored, the results of the anaylsis will often appear to have low significance since it increases the probability of type II errors (ie. failure to reject the null hypothesis when it is true) (Alain F. Zuur, Ieno, & Elphick, 2010). We mitigated this issue by considering variance inflation factors (VIFs) for the linear model. For some intuition behind this method, consider the expression for the variances of the coefficients β_i (J Fox, 2008):

$$Variance(\beta_i) = \frac{1}{1 - R_i^2} \cdot \frac{\sigma^2}{(n-1)S_i^2}$$

What's important here is the first term contatining R_i^2, which is the variance inflation factor. The R_i^2 value is the R^2 from a linear regression model where the covariate X_i is the response variable and all other covariates are used as explanatory variables (Alain F. Zuur et al., 2010). If the magnitude of R^2 is large, this means that the variation in X_i is well accounted for by other covariates and they are highly correlated. For this analysis we used a variant of the usual VIF, called the generalized varaiance inflation factor (GVIF), which was proposed by Fox and Monette. The GVIF is essentially the VIF adjusted for the degrees of freedom (df) of the predictor varaible and is given by (John Fox & Monette, 1992; Samuel-Rosa et al., 2015):

$$GVIF = VIF^{1/(2 \cdot df)}$$

In this analysis, multicollinearity was reduced using the strategy suggested by Zuur, Ieno and Elphick in which the covariate with the highest VIF is sequentially dropped from the model, then the VIFs are recalculated and the process repeats until all VIFs are smaller than a predetermined threshold (Alain F. Zuur et al., 2010). However, in this case only the GVIFs were considered.

Using a cutoff value of 5, this procedure removed four variables from the model:

- Population mid-2030 (millions)
- Population mid-2050 (millions)
- Births per 1,000 Population
- Secondary School Enrollment Ratio – Females

The summary statistics for the reduced model are shown in tables 8–10.

Stepwise Regression

Next, stepwise model selection was performed to determine which explanatory variables should be retained in the model and which could be dropped. In particular, we used both "backward simplification" and "forward selection" search algorithms. For backward simplification, the basic idea is to find explanatory variables whose removal does not significantly reduce the quality of the model, then those terms are removed from the formula resulting in a simplified model (Calcagno & Mazancourt, 2010). In contrast, forward selection sequentially adds variables to the model which have the most significant effects. This process is repeated until all explanatory variables included in the formula are signifcant. The stepwise-selected model was created using the `step()` function from the `stats` library in R. This function uses Akaike information criterion (AIC) to measure the quality of the model and choose the model with the best score.

The stepwise variable selection algorithm removed the following variables:

- Maternal Deaths in 2013
- Percent Males in the population with HIV-AIDS
- Secondary School Male
- Tertiary School Gender Parity
- Female Share of Parliament Members
- Population mid 2015 (millions)
- Percent of Population with Age < 15
- GNI per Capita 2014
- Percent Urban
- Population per Km^2 of Arable Land
- Percent of Married Women Using All Contraception
- Percent of Population with Access to Improved Water sources (2015)

By performing this step, we are making the hypothesis that the omitted coefficients are not statistically significant from zero. It is important to determine if this is in fact true. This can be done by comparing the sum squared error from the reduced model with that of the full model using `R`'s anova function with nested linear models. The results of the hypothesis test are shown below in table 4 where the null and alternative hypotheses are respectively $H_0 : \beta = 0$ and $H_a : \beta \neq 0$ for all the omitted terms. The results of the test show a p–value of approximately 0.788 indicating that we fail to reject the null hypothesis and the omitted coefficients are not statistically signifcant from zero.

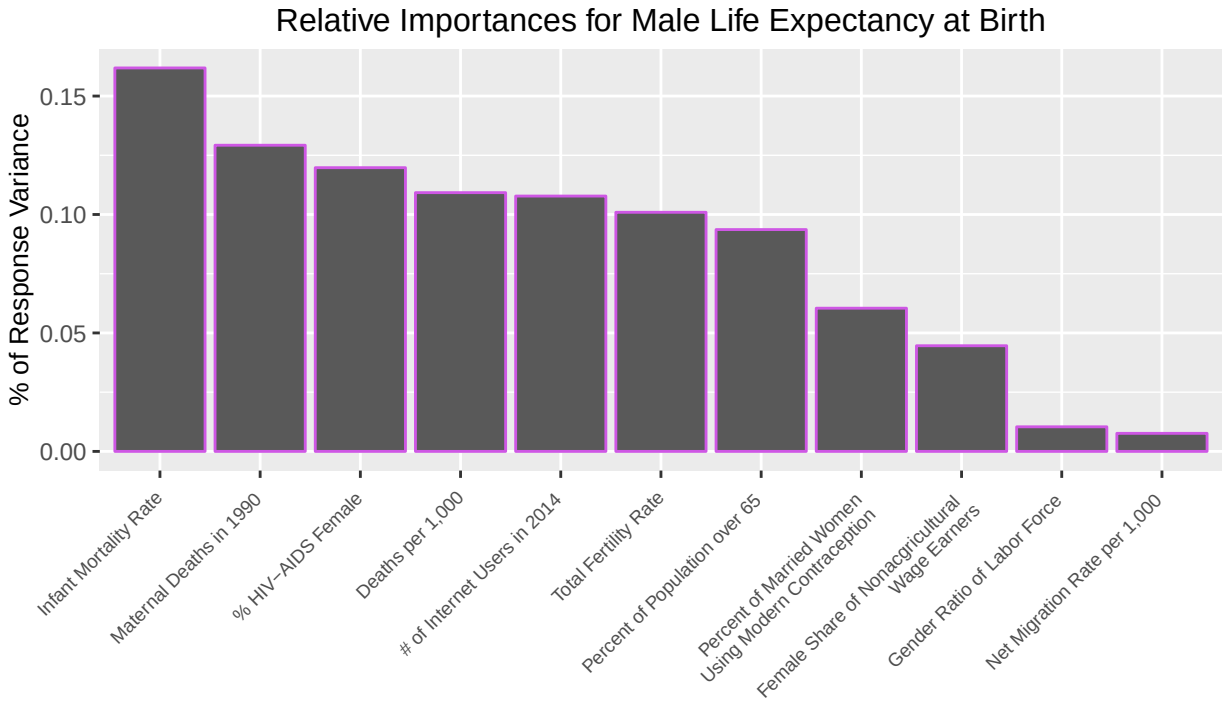
Table 4: Analysis of Variance Table

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
174	739.4	NA	NA	NA	NA
194	801.7	-20	-62.34	0.7335	0.7877

Further Variable Selection

A linear regression model was created using the 11 remaining variables. The summary statistics are shown in tables 11–13 in the appendix. In general, the results indicate that we should reject the null hypothesis, $H_0 : \beta_i = 0$ for the explanatory variables (x_is) included in the mixed stepwise regression model.

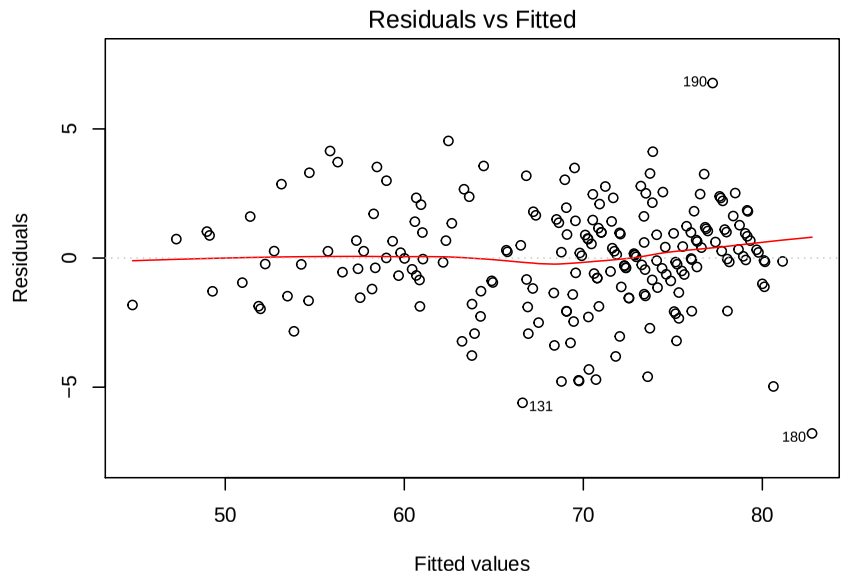
In order to find the explanatory variables which best explain male life expectancy at birth at a gloabl scale, the "lmg" metric (named after its authors) was calculated for the remaining variables using the `calc.relimp()` function from the `relaimpo` package in R. This metric gives an averaging of the R^2 contribution obtained from all possible orderings of the predictors and has recently been adopted by many researchers (Grömping, 2006).

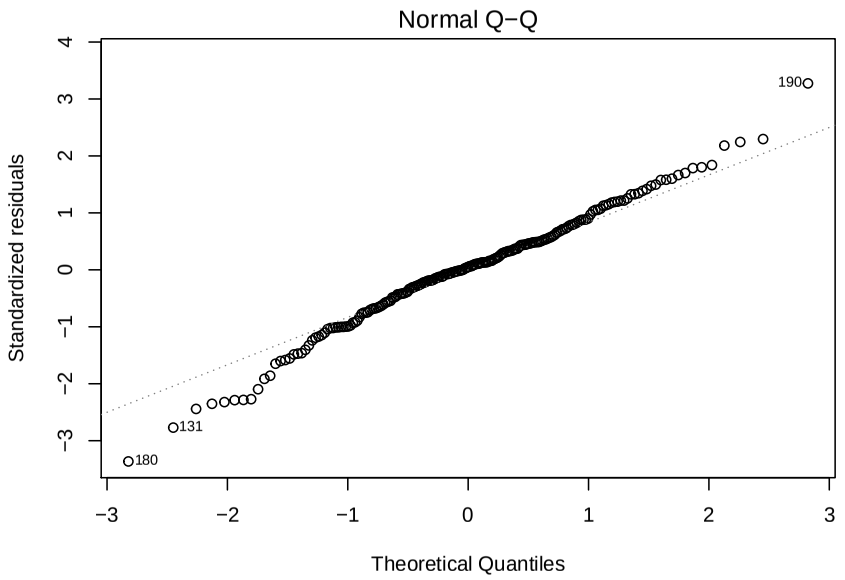


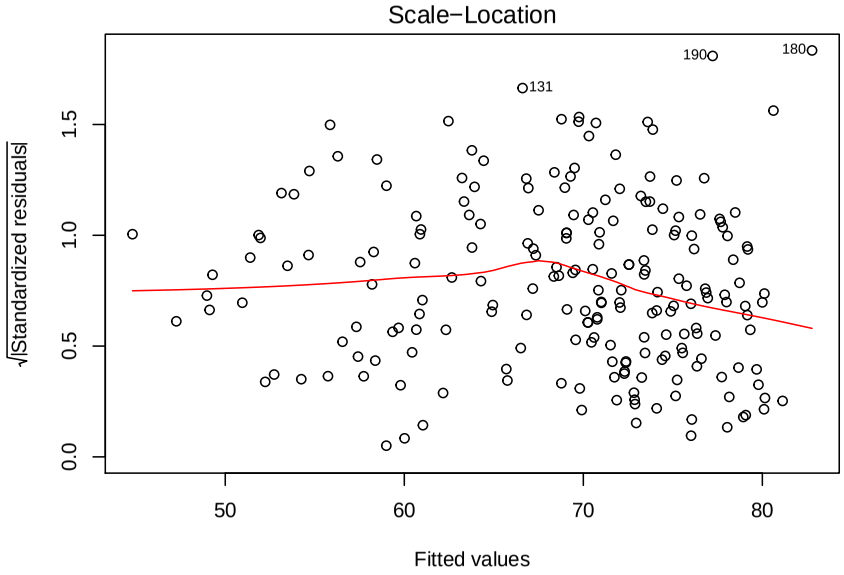
Explanatory variables with relative contributions less than 10% to the coefficient of determination (R^2) were removed from the model in order to focus on variables most relevant to the predictor variable. One exception here was the percent of the population over 65, with an lmg value of 0.0936. The choice to retain this variable was due to it's very low p–value in the stepwise model summary (shown in table 12) which strongly suggests that it has a nonzero coefficient β. In constrast, the p–values for the variables removed in the step were all greater than 0.05, with the exception of the female share of nonagricultural wage earners, which was slightly less than the α value. Thus, the variables included in the final regression model were:

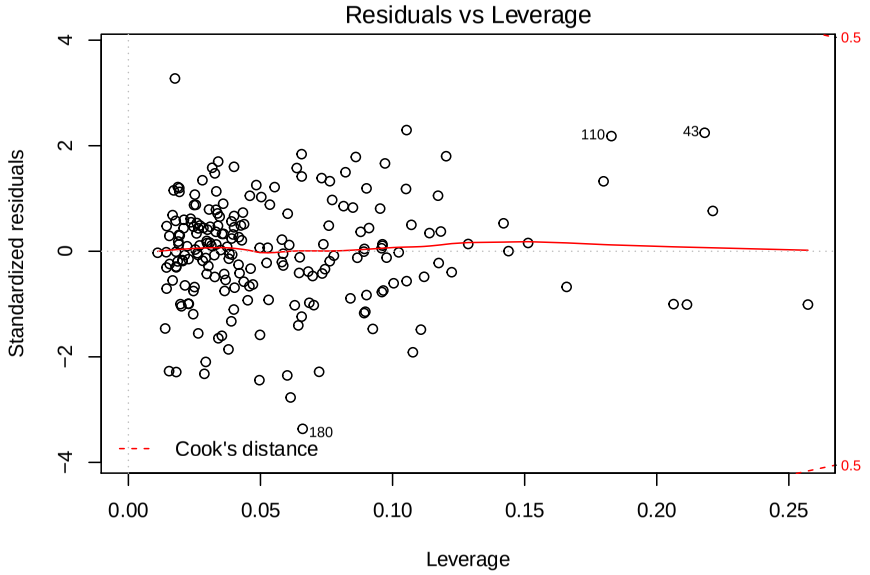
- Maternal Deaths in 1990
- Percent HIV–AIDS Female
- Deaths per 1,000 Individuals
- Infant Mortality Rate
- Percent of Population over 65
- Total Fertility Rate
- Number of Internet Users in 2014

Intuitively, these explanatory variables seem reasonable and the sign of the coefficients (as shown in table 15) offer further support to their logical interpretation. For example, there is a negative relationship between the number of deaths per 1,000 individuals and the life expectancy of males at birth, a straightforward result. The variable with the largest coefficient is the percent of the population over 65. It is conceivable that if a country has a high proportion of older individuals this would reveal a great deal about the country as a whole. Aside from the promise of a longer life, an elderly population could help raise the younger generation and allow middle age adults to focus on raising their family and improving living conditions.


Residuals vs Fitted
Residuals
Fitted values
190
131
180


Normal Q−Q
Standardized residuals
Theoretical Quantiles
190
131
180


Scale−Location
√|Standardized residuals|
Fitted values
190
180
131


Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
0.5
110
43
180

Residuals

One useful metric to asses the quality of the regression model is to check the residuals. These are defined as the deviation of an outcome from the predicated mean value for all observations with the same value for the explanatory variable (Seltman, 2015).

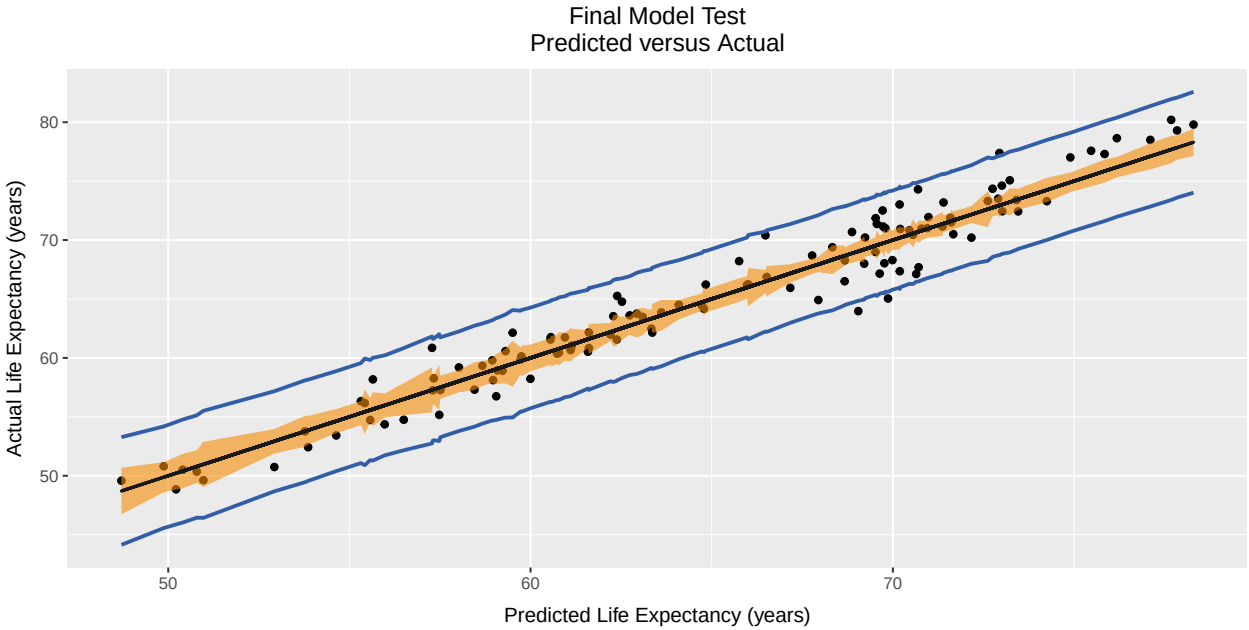
The plot of residuals vs. fitted gives some information about the linearity and equal variance assumptions. If there is a distinct U–shape, this could mean there is a lurking variable, higher order terms are needed in the model, or perhaps the assumption of constant variance is violated. With repsect to the residuals vs. fitted plot on the previous page, there are no obvious issues. It is worth mentioning that Andorra, San Marino, and Cambodia have relatiely high residuals.

Next, the normal Q-Q plot can help check the assumption that the error term ϵ is normally distributed. For our model, the normal–quantile plot appears to have slightly heavy tails and Andorra, San Marino, and Cambodia again do not follow the trend. Based on this plot, the residuals seem to be overall normally distributed.

The scale–location plot helps to check the assumption of a constant variance, σ^2. If this plot has a distinct linear trend, this would indicate variance increases for the fitted values and the validity of the p–values is low. For our model, the square root of the standardized residuals appears to be approximately constant, however as the fitted values increase there is a slightly negative linear trend and we still observe the same three outliers. Due to the fact that this trend only occurs for a small range of values near the upper extreme this is not of deep concern.

Lastly, we consider the standardized residuals vs leverage plot. This plot allows one to see if there are any data points which exert a coefficient–altering effect on the model. The intuition behind this comes from the nature of OLS which seeks to minimize the vertical distances between the data and the line. If there is an extreme value in the sample, the squared penalty term will result in that point having greater leverage and the regression line will be fit such that it is closer to that point. The metric we used for this analysis is Cook's distance, which is essentially a measure of how far the predicted values would move if the data point in question were removed. The general rule of thumb is that if a point has a value of Cook's distance greater than 0.5, it is said to have high leverage. For our model, all points have Cook's distance values less than 0.5, and thus have relatively low leverage.

Prediction



A crucial point in determining the quality of our model was assessing its ability to accurately predict values for male life expectancy at birth given empirical data from several random nations. Supplementary data were obtained from the World Bank collection of development indicators. This international organization is known for delivering the most current and accurate global development data from officially-recognized sources (World Bank Group, 2016). The data were downloaded using a Python script, for which the code is available, which implements the `wbdata` module from the Python package index. One important difference here is that the data obtained for *percent HIV–AIDS female*, *deaths per 1,000 individuals*, *infant mortality rate*, *total fertility rate*, and *percent of population over 65* were for 2014 as opposed to 2015. In addition, the data obtained for the number of internet users was for 2013 to stay consistent and maternal deaths were kept for the year 1990, since data from previous years were not available. It is important to note that the units of measurement were the same and logarithmic transformations were performed on the relevant variables, so valid comaprisons could be made.

Using the final linear regression model and the supplementary data, estimates for the fitted values were obtained using the `predict()` function in R. In addition, prediction bands and confidence bands of 95% were constructed for these predicted values to see how well the model holds and to help determine if in fact we can predict the average life expectancy of an citizen based on the characteristics of the nation. The above figure shows the predicted life expectancy values versus the actual life expectancy values for the data obtained from the World Bank. In general, the predicted values fall within the 95% prediction intervals, with the exception of the Syrian Arab Republic which fell slightly outside. On the whole, these results further highlight the ability of our regression model to forecast life expectancy given the seven parameters.

Conclusion

In the final regression model the seven remaining explanatory variables were: *percent HIV–AIDS female*, *deaths per 1,000 individuals*, *infant mortality rate*, and *percent of population over 65*, *number of internet users*, *maternal deaths in 1990*, and *total fertility rate*. All of these indicators were significant (shown in table 15) and accounted for approximately 94% of the variance in male life expectancy at birth at a global level. This underscores the fact that life expectancy is meaningfully determined by infrastructure, social development, and illness—and that to improve the life expectancy in developing regions we must assist them in these critical issues.

Appendix

Note: some values (p-values in particular) are altered due to formatting issues

Table 5: Whole Model Residuals

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-5.84	-0.956	0.171	0	1.24	5.81

Table 6: Whole Model Coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7.49e+01	5.62e+00	1.33e+01	0.00e+00
Maternal_Deaths_1990	-6.22e-01	3.22e-01	-1.93e+00	5.50e-02
Maternal_Deaths_2013	-4.23e-01	3.38e-01	-1.25e+00	2.13e-01
`Percent_HIV-AIDS_Male`(0.09,0.15]	7.50e-02	5.28e-01	1.41e-01	8.88e-01
`Percent_HIV-AIDS_Male`(0.15,0.25]	8.62e-01	5.27e-01	1.64e+00	1.04e-01
`Percent_HIV-AIDS_Male`(0.25,0.5]	5.38e-01	6.63e-01	8.12e-01	4.18e-01
`Percent_HIV-AIDS_Male`(0.5,1]	6.56e-01	1.31e+00	4.99e-01	6.18e-01
`Percent_HIV-AIDS_Male`(1,10]	5.21e-01	1.45e+00	3.60e-01	7.20e-01
`Percent_HIV-AIDS_Female`(0.099,0.25]	-1.36e+00	5.35e-01	-2.54e+00	1.20e-02
`Percent_HIV-AIDS_Female`(0.25,0.5]	-1.25e+00	8.92e-01	-1.40e+00	1.63e-01
`Percent_HIV-AIDS_Female`(0.5,1]	-2.02e+00	1.01e+00	-2.00e+00	4.70e-02
`Percent_HIV-AIDS_Female`(1,5]	-2.97e+00	1.51e+00	-1.96e+00	5.10e-02
`Percent_HIV-AIDS_Female`(5,50]	-5.59e+00	1.79e+00	-3.12e+00	2.00e-03
Secondary_School_Male	-1.00e-02	3.00e-02	-3.18e-01	7.51e-01
Secondary_School_Female	8.00e-03	2.80e-02	2.87e-01	7.74e-01
Tertiary_School_Gender_Parity	-1.07e-01	3.99e-01	-2.69e-01	7.88e-01
Gender_Ratio_Labor_Force	2.64e+00	1.27e+00	2.08e+00	3.90e-02
Female_Share_of_Nonagricultural_Wage_Earners	-9.40e-02	3.10e-02	-3.07e+00	3.00e-03
Female_Share_of_Parliament_Members	-2.00e-03	1.60e-02	-1.06e-01	9.16e-01
Population_mid2015_mill	5.00e-03	6.70e-02	6.80e-02	9.46e-01
Births_per_100k_Population	-2.24e-01	9.40e-02	-2.38e+00	1.80e-02
Deaths_per_100k_Population	-1.06e+00	9.10e-02	-1.16e+01	0.00e+00
Net_migration_rate_per_100k	2.60e-02	2.60e-02	1.02e+00	3.10e-01
Pop_mid2030_mill	-8.00e-03	1.00e-01	-7.70e-02	9.39e-01
Pop_mid2050_mill	3.00e-03	3.80e-02	6.70e-02	9.47e-01
Infant_mortality_rate	-7.50e-02	1.80e-02	-4.13e+00	0.00e+00
Total_fertility_rate	1.79e+00	5.34e-01	3.35e+00	1.00e-03
Percent_of_pop_under_15	2.00e-03	7.00e-02	3.50e-02	9.72e-01
Percent_of_pop_over_65	4.24e+00	6.27e-01	6.76e+00	0.00e+00
GNI_per_capita_2014	1.63e-01	3.66e-01	4.46e-01	6.56e-01
Percent_urban	1.20e-02	1.00e-02	1.26e+00	2.08e-01
Population_per_Square_kilom_of_arable_land	0.00e+00	0.00e+00	6.40e-02	9.49e-01
Percent_of_married_women_using_all_contraception	-5.00e-03	2.10e-02	-2.45e-01	8.07e-01
Percent_of_married_women_using_modern_contraception	2.80e-02	1.80e-02	1.55e+00	1.23e-01
INET_USRS_2014	1.50e-02	1.40e-02	1.05e+00	2.96e-01
IMPROVED_H20_SRC_2015	-5.00e-03	2.80e-02	-1.90e-01	8.49e-01

Table 7: Whole Model Summary Continued

Continued
Residual standard error: 2.06 on 174 degrees of freedom
Multiple R-squared: 0.95, Adjusted R-squared: 0.94
F-statistic: 93.9 on 35 and 174 DF, p-value: <2e-16

Table 8: Reduced VIF Model Residuals

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-6.05	-0.917	0.0523	0	1.19	6.06

Table 9: Reduced VIF Model Coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7.40e+01	5.59e+00	1.33e+01	0.00e+00
Maternal_Deaths_1990	-6.76e-01	3.22e-01	-2.10e+00	3.70e-02
Maternal_Deaths_2013	-3.92e-01	3.36e-01	-1.17e+00	2.45e-01
`Percent_HIV-AIDS_Male`(0.09,0.15]	8.50e-02	5.28e-01	1.62e-01	8.72e-01
`Percent_HIV-AIDS_Male`(0.15,0.25]	9.56e-01	5.25e-01	1.82e+00	7.00e-02
`Percent_HIV-AIDS_Male`(0.25,0.5]	6.02e-01	6.57e-01	9.16e-01	3.61e-01
`Percent_HIV-AIDS_Male`(0.5,1]	6.43e-01	1.31e+00	4.90e-01	6.25e-01
`Percent_HIV-AIDS_Male`(1,10]	2.78e-01	1.45e+00	1.92e-01	8.48e-01
`Percent_HIV-AIDS_Female`(0.099,0.25]	-1.32e+00	5.24e-01	-2.52e+00	1.30e-02
`Percent_HIV-AIDS_Female`(0.25,0.5]	-1.16e+00	8.51e-01	-1.36e+00	1.76e-01
`Percent_HIV-AIDS_Female`(0.5,1]	-2.12e+00	9.89e-01	-2.15e+00	3.30e-02
`Percent_HIV-AIDS_Female`(1,5]	-2.66e+00	1.50e+00	-1.77e+00	7.90e-02
`Percent_HIV-AIDS_Female`(5,50]	-5.42e+00	1.78e+00	-3.04e+00	3.00e-03
Secondary_School_Male	-2.00e-03	1.30e-02	-1.22e-01	9.03e-01
Tertiary_School_Gender_Parity	9.60e-02	3.61e-01	2.65e-01	7.91e-01
Gender_Ratio_Labor_Force	2.51e+00	1.26e+00	1.99e+00	4.90e-02
Female_Share_of_Nonagricultural_Wage_Earners	-9.70e-02	3.00e-02	-3.25e+00	1.00e-03
Female_Share_of_Parliament_Members	-2.00e-03	1.60e-02	-1.33e-01	8.94e-01
Population_mid2015_mill	-1.00e-03	1.00e-03	-8.29e-01	4.08e-01
Deaths_per_100k_Population	-1.09e+00	9.00e-02	-1.21e+01	0.00e+00
Net_migration_rate_per_100k	2.70e-02	2.60e-02	1.04e+00	3.01e-01
Infant_mortality_rate	-8.00e-02	1.70e-02	-4.61e+00	0.00e+00
Total_fertility_rate	9.29e-01	3.86e-01	2.41e+00	1.70e-02
Percent_of_pop_under_15	-6.00e-02	6.40e-02	-9.42e-01	3.48e-01
Percent_of_pop_over_65	4.78e+00	5.83e-01	8.21e+00	0.00e+00
GNI_per_capita_2014	1.25e-01	3.65e-01	3.41e-01	7.33e-01
Percent_urban	8.00e-03	9.00e-03	8.62e-01	3.90e-01
Population_per_Square_kilom_of_arable_land	0.00e+00	0.00e+00	8.20e-02	9.35e-01
Percent_of_married_women_using_all_contraception	-6.00e-03	2.00e-02	-3.19e-01	7.50e-01
Percent_of_married_women_using_modern_contraception	2.80e-02	1.80e-02	1.55e+00	1.23e-01
INET_USRS_2014	2.20e-02	1.40e-02	1.56e+00	1.21e-01
IMPROVED_H20_SRC_2015	-5.00e-03	2.90e-02	-1.77e-01	8.60e-01

Table 10: Reduced VIF Model Summary Continued

Continued
Residual standard error: 2.07 on 178 degrees of freedom
Multiple R-squared: 0.948, Adjusted R-squared: 0.939
F-statistic: 105 on 31 and 178 DF, p-value: <2e-16

Table 11: Stepwise Model Residuals

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-6.19	-1.09	-0.0063	0	1.18	6.14

Table 12: Stepwise Model Coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7.39e+01	1.75e+00	4.23e+01	0.00e+00
Maternal_Deaths_1990	-9.50e-01	2.32e-01	-4.10e+00	0.00e+00
`Percent_HIV-AIDS_Female`(0.099,0.25]	-9.91e-01	4.60e-01	-2.15e+00	3.30e-02
`Percent_HIV-AIDS_Female`(0.25,0.5]	-7.18e-01	7.03e-01	-1.02e+00	3.08e-01
`Percent_HIV-AIDS_Female`(0.5,1]	-1.73e+00	7.94e-01	-2.17e+00	3.10e-02
`Percent_HIV-AIDS_Female`(1,5]	-2.30e+00	7.79e-01	-2.95e+00	4.00e-03
`Percent_HIV-AIDS_Female`(5,50]	-5.82e+00	1.12e+00	-5.19e+00	0.00e+00
Gender_Ratio_Labor_Force	2.08e+00	1.14e+00	1.83e+00	6.90e-02
Female_Share_of_Nonagricultural_Wage_Earners	-8.50e-02	2.70e-02	-3.11e+00	2.00e-03
Deaths_per_100k_Population	-1.05e+00	7.70e-02	-1.36e+01	0.00e+00
Net_migration_rate_per_100k	3.90e-02	2.30e-02	1.68e+00	9.50e-02
Infant_mortality_rate	-8.40e-02	1.60e-02	-5.29e+00	0.00e+00
Total_fertility_rate	4.89e-01	2.21e-01	2.21e+00	2.80e-02
Percent_of_pop_over_65	4.81e+00	4.78e-01	1.01e+01	0.00e+00
Percent_of_married_women_using_modern_contraception	1.70e-02	1.10e-02	1.63e+00	1.05e-01
INET_USRS_2014	3.10e-02	1.20e-02	2.57e+00	1.10e-02

Table 13: Stepwise Model Summary Continued

Continued
Residual standard error: 2.03 on 194 degrees of freedom
Multiple R-squared: 0.946, Adjusted R-squared: 0.941
F-statistic: 224 on 15 and 194 DF, p-value: <2e-16

Table 14: Final Regression Model Residuals

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-6.79	-1.14	0.101	0	1.15	6.78

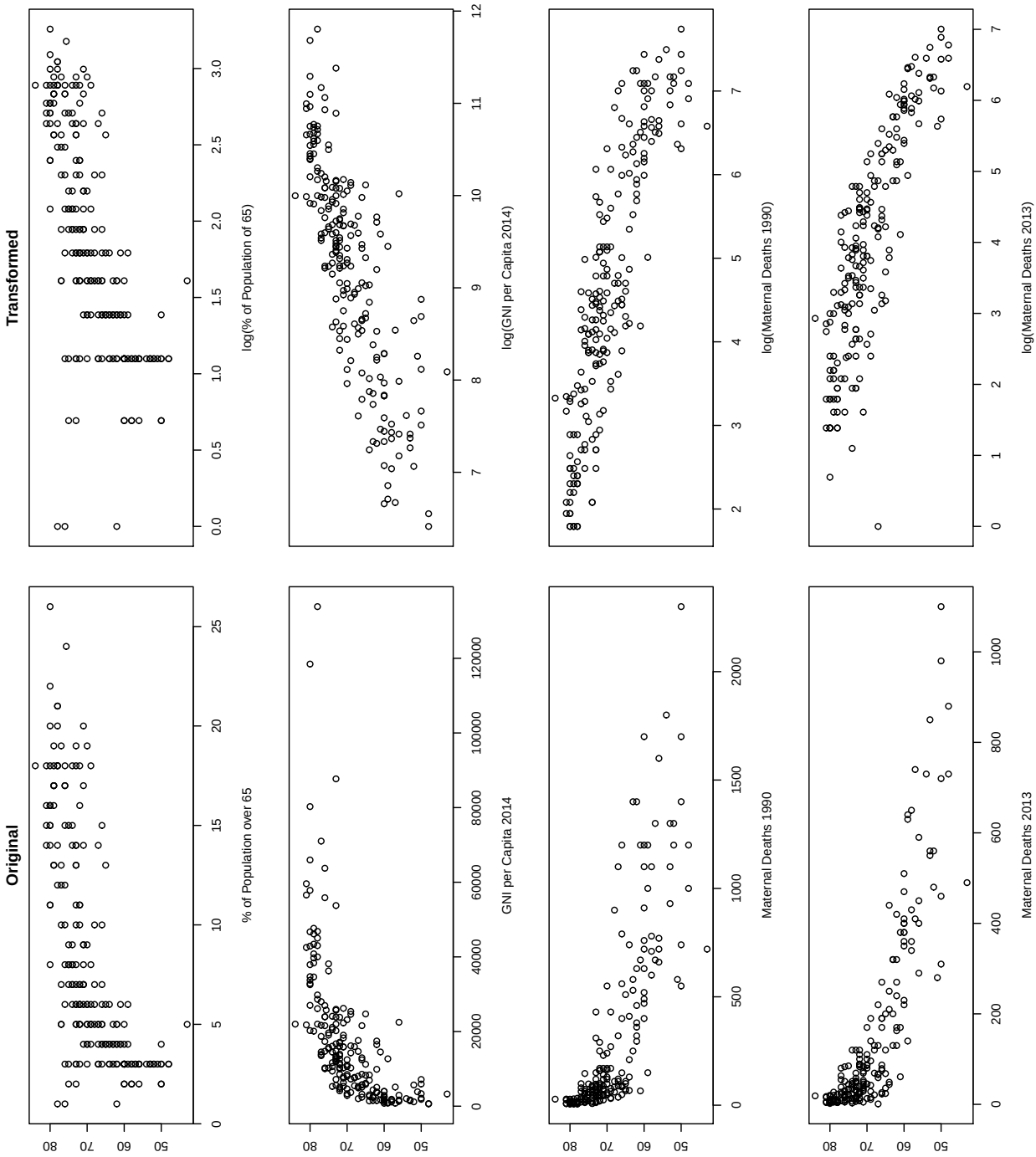
Table 15: Final Regression Model Coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7.32e+01	1.68e+00	4.35e+01	0.00e+00
Maternal_Deaths_1990	-8.92e-01	2.36e-01	-3.77e+00	0.00e+00
`Percent_HIV-AIDS_Female`(0.099,0.25]	-1.16e+00	4.30e-01	-2.69e+00	8.00e-03
`Percent_HIV-AIDS_Female`(0.25,0.5]	-5.94e-01	6.83e-01	-8.70e-01	3.86e-01
`Percent_HIV-AIDS_Female`(0.5,1]	-1.50e+00	7.78e-01	-1.93e+00	5.60e-02
`Percent_HIV-AIDS_Female`(1,5]	-2.10e+00	7.32e-01	-2.86e+00	5.00e-03
`Percent_HIV-AIDS_Female`(5,50]	-5.50e+00	1.06e+00	-5.20e+00	0.00e+00
Deaths_per_100k_Population	-1.09e+00	7.40e-02	-1.46e+01	0.00e+00
Infant_mortality_rate	-7.00e-02	1.60e-02	-4.41e+00	0.00e+00
Total_fertility_rate	4.04e-01	2.21e-01	1.83e+00	6.90e-02
Percent_of_pop_over_65	4.43e+00	4.23e-01	1.05e+01	0.00e+00
INET_USRS_2014	3.90e-02	1.10e-02	3.47e+00	1.00e-03

Table 16: Final Regression Model Summary Continued

Continued
Residual standard error: 2.09 on 198 degrees of freedom Multiple R-squared: 0.941, Adjusted R-squared: 0.938 F-statistic: 289 on 11 and 198 DF, p-value: <2e-16

Supplementary Material



References

Alain F. Zuur, Ieno, E. N., & Elphick, C. S. (2010). A protocol for data exploration to avoid commonstatistical problems. Methods in Ecology and Evolution.

Arias, E. (2014). United states life tables, 2009. *National Vital Statistics Reports*, *62*. Journal Article. Retrieved from http://www.cdc.gov/nchs/data/nvsr/nvsr62/nvsr62_07.pdf

Buechler, S. (2014). Refining a k-nearest-neighbor classification. Retrieved from https://www3.nd.edu/~steve/computing_with_data/17_Refining_kNN/refining_knn.html

Calcagno, V., & Mazancourt, C. de. (2010). Glmulti: An r package for easy automated model selection with (generalized) linear models. Journal of Statistical Software. Retrieved from https://www.jstatsoft.org/article/view/v034i12

Central Intelligence Agency. (2015). Lesotho. Retrieved from https://www.cia.gov/library/publications/the-world-factbook/geos/lt.html

Fox, J. (2008). Applied regression analysis and generalized linear models. Sage Publications.

Fox, J., & Monette, G. (1992). Generalized collinearity diagnostics. Retrieved from http://www.jstor.org/stable/2290467

Grömping, U. (2006). Relative importance for linear regression in r: The package relaimpo. Journal of Statistical Software. Retrieved from https://www.jstatsoft.org/article/view/v017i01

Haitovsky, Y. (1968). Missing data in regression analysis. Retrieved from http://facweb.cs.depaul.edu/sjost/csc423/documents/missing_values.pdf

Hand, D., Mannila, H., & Smyth, P. (2001). MIT Press. Retrieved from ftp://gamma.sbin.org/pub/doc/books/Principles_of_Data_Mining.pdf

Little, R., & Rubin, R. (1987). Statistical analysis with missing data. Retrieved from https://www.jstor.org/stable/1165119?seq=1#page_scan_tab_contents

Samuel-Rosa, A., Anjos, L., Vasques, G., Heuvelink, G., Olsen, T., Kincaid, T., ... Xu, R. (2015). Pedometrics: Pedometric tools and techniques. Retrieved from https://cran.r-project.org/web/packages/pedometrics/index.html

Seltman, H. J. (2015). Experimental design and analysis. Retrieved from http://www.stat.cmu.edu/~hseltman/309/Book/Book.pdf

Soley-Bori, M. (2013). Dealing with missing data:Key assumptions andmetho ds for applied analysis. Retrieved from ttp://www.bu.edu/sph/files/2014/05/Marina-tech-report.pdf

Torgo, L. (2010). Data mining using r: Learning with case studies. CRC Press. Retrieved from http://www.dcc.fc.up.pt/~ltorgo/DataMiningWithR

World Bank Group. (2016). World development indicators. Retrieved from http://data.worldbank.org/data-catalog/world-development-indicators

World Health Organization. (2014). An overarching health indicator for the post-2015 development agenda. Retrieved from http://www.who.int/healthinfo/indicators/hsi_indicators_SDG_TechnicalMeeting_December2015_BackgroundPaper.pdf

World Health Organization. (2015). HIV/AIDS. Online. Retrieved from http://www.who.int/mediacentre/factsheets/fs360/en/