World Health Data Report

DTSC13-302

Muiz Murad
Charles Liebenberg
Ricky Soh Ray Kee
Kabilan Cholan
Thomas Watkins

Table of Contents

Executive Report

Executive Summary

A country's life expectancy is the culmination of its performance as a vehicle of service to its citizens. It can be used then, as a measure of success for a government. Therefore, the identification of key target areas that are most efficient in improving life expectancy can prove invaluable. The purpose of this report is to investigate these factors that affect life expectancy, quantify them and their effects, and ultimately allow for efficient maximization of life expectancy.

Detailed Investigation

After investigation into the data set, a linear fit was determined to model a county's average life expectancy quite effectively. The model can quantify the relationship between average life expectancy and other national variables, as such this model can be used to answer the research questions that were considered before endeavoring on this assignment.

What factors influence a country's Life Expectancy?

The variables which were found to have a significant influence on a country's average life expectancy are adult mortality (variable which describes mortality rate of both sexes amongst 15-60 year old), infant deaths (variable which describes infant deaths per 1000 population), income composition of resources (variable which describes human development index in terms of income composition of resources), total expenditure (variable which describes proportion of government spending on health), Hepatitis B (variable which describes HepB immunization coverage amongst one-year old), Polio (variable which describes Pol3 immunization coverage amongst one-year old), thinness prevalence amongst 5-9 years (variable which describes prevalence of thinness amongst children aged 5 to 9) and HIV/AIDS (variable which describes deaths per 1000 live births).

These variables are used as predictors for average life expectancy, but they do not necessarily directly influence the average life expectancy. In fact, most of these variables are not influential on their own (within the context of the model), rather the interactions between several of these variables provides predictive value. It is further noted that there is considerable correlation between these variables and other underlying factors. These variables that have been identified, then, may act as proxy variables and represent underlying immeasurable things such as overall quality of a health system.

It was found that while hepatitis immunization extremely important, it is relatively meaningless without a focus on polio immunization. It can be inferred that this could be because if hepatitis B does not kill someone, polio is likely to get that person instead. Therefore, it is only with systematic focus on both variables that life expectancy will meaningfully increase. Furthermore, while these diseases are extremely prevalent, they may also act as proxy for a countries underlying health system, and as such the system as a whole should remain a priority. It is also noted that the effect of this polio prevention is drastically increased with a focus on decreasing malnourishment in younger people and income disparity. If these factors are not addressed in tandem, the realized effects will be less than desired and less cost effective.

Additionally, it was concluded that while the decrease of adult mortality is extremely important in increasing life expectancy (quite intuitively), its effects must be amplified through a focus on reducing infant deaths and income disparity. It is not enough for a country to focus its efforts on purely the reduction of adult mortality, it must address the root and stem of the problem. This intuitively implies that adult mortality is heavily dependent on income disparity and infant mortality and that a country must focus its efforts on these three areas in tandem for best results.

Technical Report

Initial Data Cleaning

The first task was to ensure reproducibility of the analysis. To do this, a seed was set. This seed means that every random function will produce the same output on reruns of the report (the internal number generator begins at the same value each time). This is extremely important for both verification of the findings and for use in further analysis at a later point.

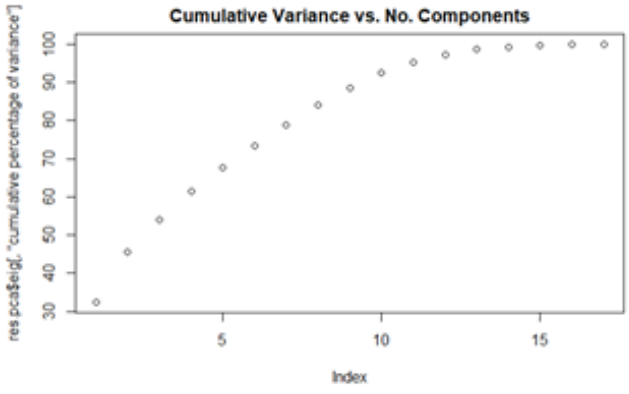
Initially the dataset contained variables mostly categorised as factors. To obtain a regression output that describes the effects of a unit increase in each variable, this data needed to be converted to numeric. This reflects the inherent nature of the data and will also decrease the amount of time required to run the analysis.

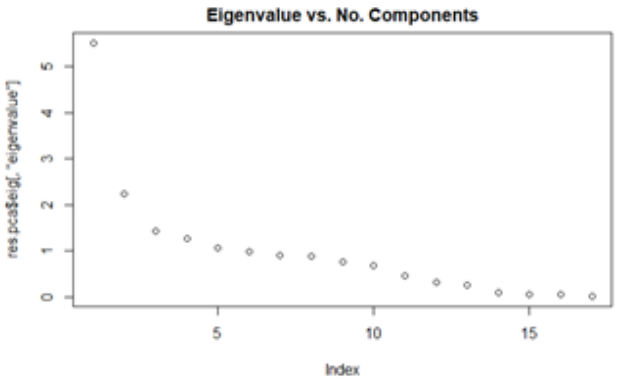
Next, through investigation of the data, the presence of multiple NA's within the dataset was found. It was established that all the null values were contained within only 25 datapoints. It was deemed that the removal of these values would not affect the analysis significantly and thus they were removed to allow for the development of the models.

Finally, a choice was made to not split the data into training and test partitions. This is because a limited amount of data was available, and as such better predictions can be made with more data. However, this choice does necessitate the use of metrics that approximate external metrics later in the analysis.

Exploratory Data Analysis

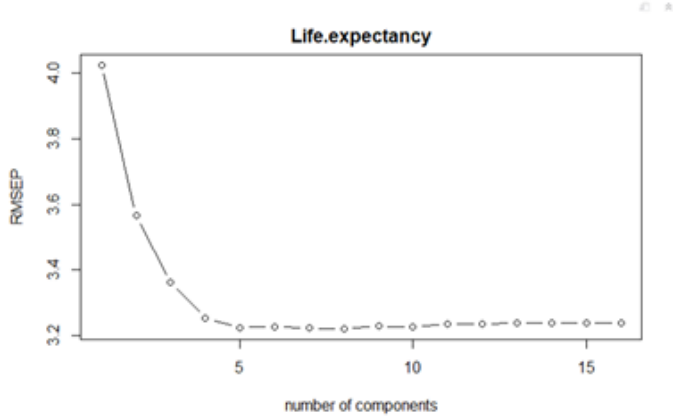
Initially a principal components analysis was performed. PCA is used for exploratory data analysis and allows for dimensionality reduction. In essence, it attempts to capture the variance of the regressors through components consisting of these regressors. This process is substantially sensitive to scaling, and thus, the data was scaled and PCA was performed. This resulted in the following graphs:





Here it can be seen that more than ten components are required to adequately explain the variance in the dataset. Furthermore, the Kaiser Rule dictates that all components that have an Eigenvalue > 1 should be used. The second graph shows that this rule would suggest that at least 8 components are required. Therefore, it can be concluded that principal components analysis has not meaningfully decreased the number of variables for the analysis.

After it was observed that principal components analysis was not useful in reducing dimensionality, partial least squares (PLS) was investigated for its dimensionality reduction properties. PLS differs from principal components analysis in that instead of trying to explain the most variance with components, it fits a linear model through projection of the predicted variables and the observable variables. When partial least squares was performed, the following results were observed:

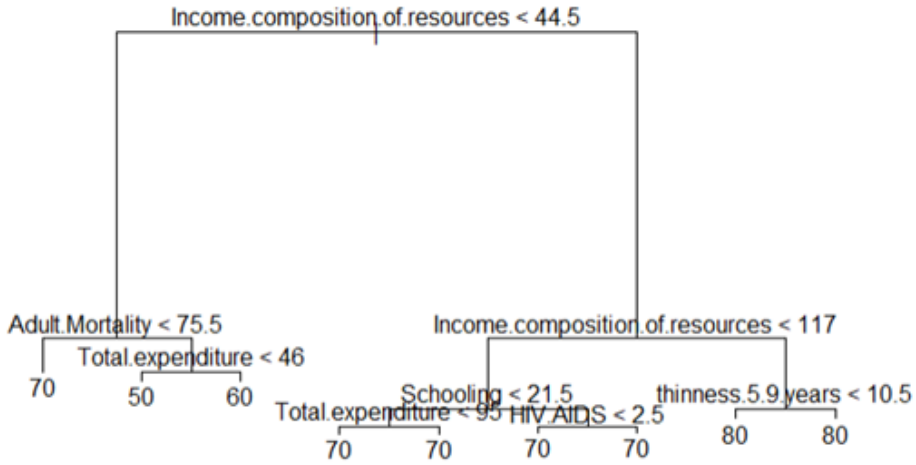


This graph demonstrates that Root Mean Square Error of Prediction is minimised after 5 components. This represents a significant reduction in dimensionality and therefore the results of this analysis was saved for later analysis.

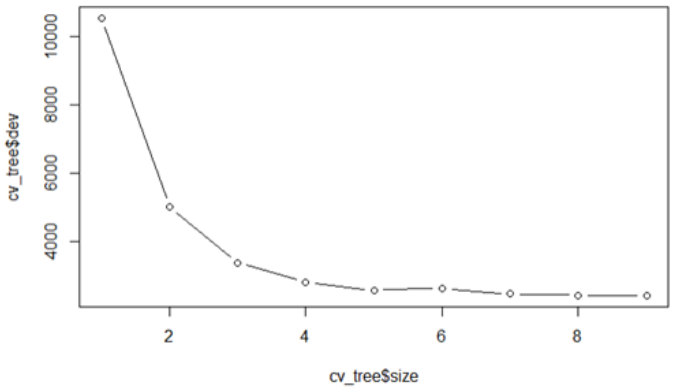
Recursive Partitioning

Decision Tree

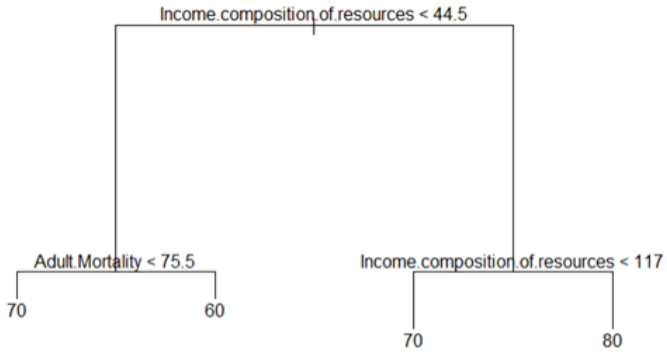
Decision tree is a classification model in form of a tree structure, it generally breaks down dataset into smaller branches amongst predictors. In theory, decision tree is a better classification tool than PCA and is the most popular tool for variable selection for data analysis. In summary, it is a machine learning algorithm that partitions data into subsets which starts with a binary split then further splits can be made. Also, the splits are determined through minimising deviance or Gini Index which maximises use of variables until the complexity parameter (CP) starts to converge which means that more variables do not improve the interpretability of the data according to the dependent variable. In this case, decision trees' splits are equivariant to monotonic transformation and useless variables are automatically ignored. Below is the full decision tree (overfitting tree) without any pruning:



However, the tree does need to be pruned, so it does not overfit and subsequently uses only the most significant variables. Cross validation will be used to decide on the size of the tree, allowing for a choice less likely to overfit from because of not using a train-test split. The cross validated deviance can be plotted against size and an informed choice can be made from the graph as follows:



Around size 4 the deviance starts to flatten out. Using pruning techniques, the following decision tree has been created:

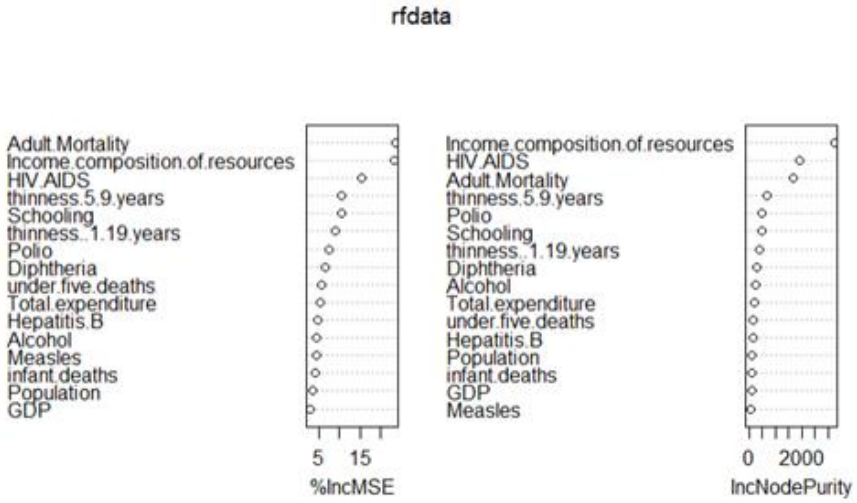


From the tree above the significant variables can be found. Income composition of resources is clearly the most significant variable here. It interacts with Adult Mortality and interacts itself at different level. This is clearly shown in the first decision tree itself as variables such as Schooling and thinness.5.9 years are not significant because it does not matter if Schooling or thinness.5.9 years have less than or over values.

Random Forest

Like decision trees, random forest also automatically chooses the best predictors. Random forest is based on generating many decision trees to identify classification consensus by selecting the most common output. It will try to classify at nodes to maximize information gain until all the nodes are exhausted or there is no further information gain. However, this is not a great predictive tool which is why they are referred as weak leaner.

From the random forest selection, it can be seen that Adult Mortality, Income composition of resources and HIV.AIDS are the most important variables regarding Life expectancy.



Summary

From both the decision tree and random forest, it is interesting to see the significance of the variables in relation to the life expectancy of a country. For instance, it can be seen that income composition, HIV.AIDS and adult mortality would have the most impact to life expectancy while all other variables including population and GDP would not contribute too much to life expectancy. It is logical that income composition of resources, mortality rate and HIV.AIDS would have the most impact on life expectancy as income composition of resources is the measure of human development index that if a country utilizes its resources efficiently then it would be most likely that life expectancy will be longer. On the other hand, mortality rate and HIV.AIDS would have the negative effect on life expectancy as higher mortality rate and HIV.AIDS death would decrease in life expectancy. Therefore, it is through these initial variable exploratory techniques that an initial understanding of the interrelationships has been gathered.

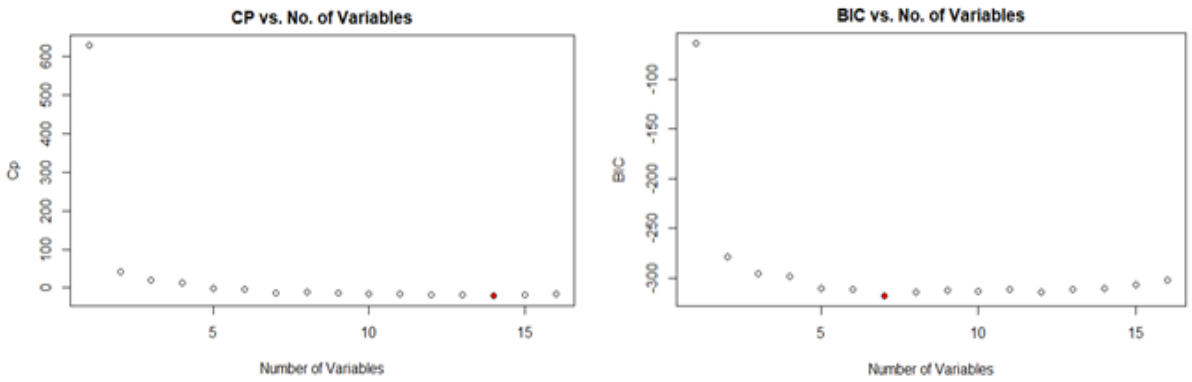
Linear Regression

Variable Selection

To confirm the previously established exploratory variable analysis, a backward stepwise approach was taken. In essence, this method will step backwards starting with the largest model choice and remove the variable with the least predictive capability. The function will then output calculated metrics that will allow for quick identification of the most significant variables/interactions and additionally, for a more informed choice on model size.

The metrics used to evaluate model performance were carefully considered. It might seem reasonable to minimise residual sum of squares or maximise R^2, but these always improve when the number of predictors is increased (the larger model will always explain more of the variance, even if it is noise). Some other common metrics which do not essentially improve with number of predictors used are Adjusted-R^2, Mallows CP (referred to as CP in analysis) and Schwartz's information criterion (referred to as BIC in analysis).

Once these metrics have been determined, we find the model size that either maximises or minimises (depending on which metric is being used) the metric. During the analysis, both Mallow's CP and BIC was considered.



As the graphs show, Mallow's CP suggested the use of 14 variables and BIC suggested the use of 7. The internal function performing this backward stepwise regression uses CP to make the choices. Therefore, it was decided that perhaps BIC was a better choice in approximating external success. Furthermore, it can be observed that after the 7th variable, Mallow's CP does not improve substantially. Therefore the 7 best predictors were chosen.

The seven variables/interactions chosen for the creation of the linear model were: Adult.Mortality, Adult.Mortality:infant.deaths,Adult.Mortality:Income.composition.of.resources,infant.deaths:Total.expenditure, Hepatitis.B:Polio, Polio:thinness.5.9.years and HIV.AIDS:Income.composition.of.resources

After this variable/interaction selection and analysis, a linear model was created (using the variables previously outlined). This basic linear model would serve as a baseline for the rest of the linear analysis.

Penalty Functions

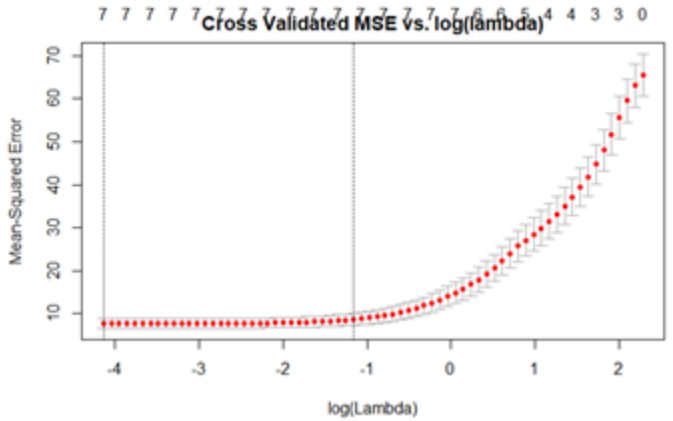
In regression, the bias-variance trade-off can be addressed through regularisation. The bias variance trade-off is the concept by which bias and variance are internally linked. A decreased variance will result in increased bias and vice-versa. In situations where it is beneficial to have increased consistency of predictions, it is often a good idea to trade this for some bias. For this report, it was decided that this concept would be trialed through regularisation to allow for increased confidence in predictions.

A regularisation method used to address this trade-off is ridge regression in which overfitting is reduced through placing constraints on large β coefficients. By using ridge regression, the variance of coefficient reduces as it limits the space of the parameter vector β with a constraint on squared deviance.

Another regularisation method used is named lasso regression. Like Ridge regression, Lasso regression also trades off an increase in bias to decrease variance. It differs in that it uses absolute deviance as the constraint for the coefficients.

Both of these methods were used in the analysis. This is because it was believed that this would allow for the benefits of both. Ridge regression would place greater emphasis on greater deviances (by virtue of the square function), and lasso regression would allow for the removal of overly variable regressors (the absolute function has a tendency to decrease coefficients to 0).

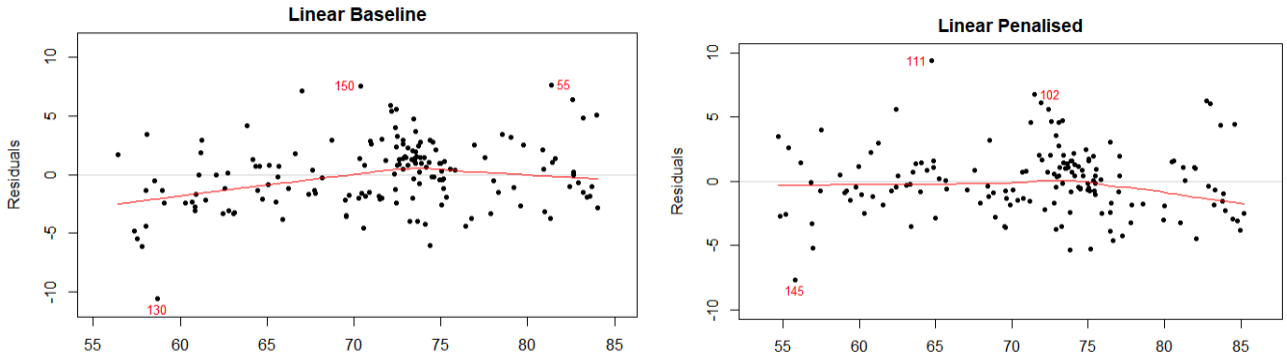
The next task was to identify an optimal value for lambda (and moderate the degree of penalisation). To choose this value, two graphs were plotted. The first is coefficient value against lambda. This allows for visualisation of the effects of the penalty function. The second is the mean squared error against lambda. This will allow for visualisation of the bias variance tradeoff.



The plot demonstrates that at a value of –1 for log lambda the bias variance trade-off seems optimised. The first shows that at this value the coefficients have begun to decrease, but not massively to completely reduce it to zero.

Model Diagnostics

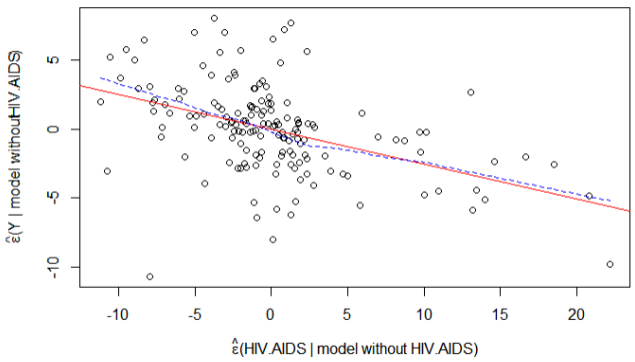
The two newly created linear models, one with the use of elastic-net (penalized) regression and one without, were then compared with the use of residual plots. These are as follows:

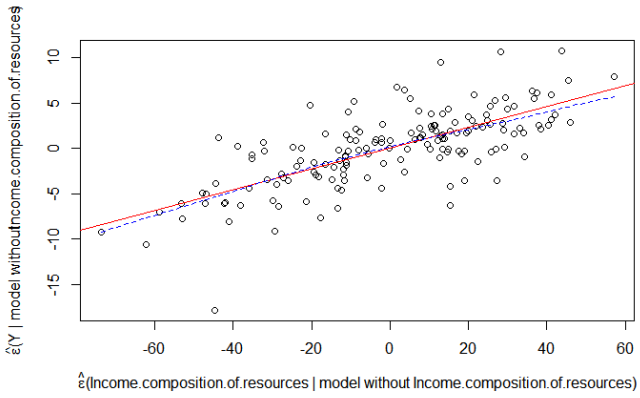


As can be seen, both models demonstrated reasonably random, homoscedastic residuals. However, it can be observed that there is some slight overdispersion in the left area of the plot with slightly more of this dispersion in the penalized model.

This overdispersion could possibly be due to three things: incorrect variable selection, incorrect variable links or incorrect link functions.

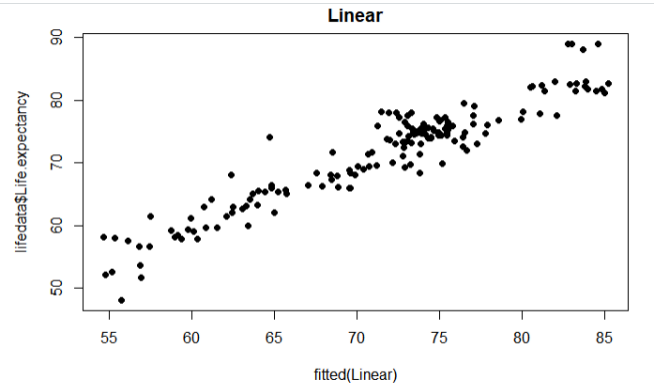
Initially, it was thought that perhaps the predictor links were incorrect. To investigate this, partial residual plots were plotted. Partial residual plots allow for the investigation of the validity of the predictor links through removing the influence of all other variables in the residuals and then plotting that against life expectancy.

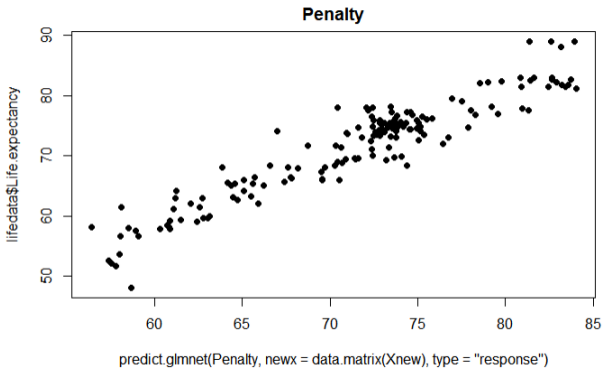




As shown in these graphs, no overdispersion or heteroskedasitity was present. The residuals are slightly more spreadout in some places, however this is just a reflection of the increased number of residuals in these sections. It was concluded that the predictor links were not obviously incorrect. The next potential reason for overdispersion investigated was Life Expectancy's link.

The independent variable link can be investigated through observing the relationship between the independent variable (and its associated transformed link if relevant) and the model's fitted values. When plotted, the relationship should be strictly linear and homoscedastic. When the linear model link residual graph was plotted, the following was observed:





To the human eye, these look quite good. It is difficult to objectively determine which graph looks more linear and homoscedastic, however it was thought that the Linear model had slightly better residuals in the lower fitted values. It was also concluded that this was not perfect, and thus there was some use in investigating the use of generalized linear models which would allow for some natural heteroskedasticity and different link functions.

Generalised Linear Regression

Once it was observed that there may be some problems with the assumption of homoscedasticity in the linear model, a generalised linear model was considered. This allows for the use of different distribution families and their associated links.

The family chosen for this analysis is the gamma family. This is because of its implicit assumption of a little bit of heteroskedasticity and also its similarity to the normal distribution.

Variable Selection

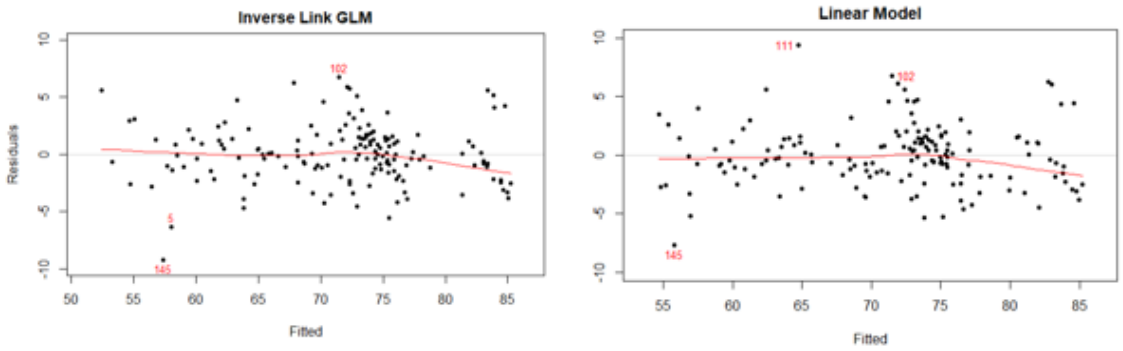
Once a general structure has been established for the generalised linear model, variable selection had to be performed again. This is because there is no guarantee that the most significant variables for linear regression are as significant when fitting to a gamma distribution.

Similar to linear regression, a backward stepwise approach was taken to the selection of variables. However, in this instance, a new stepwise analysis was performed for each link function trialed. The four link functions trialed, in an attempt to reduce the fanning problem previously described in the linear regression analysis, were Inverse, Squareroot, Logarithm and Identity.

The observed BIC and deviance for the inverse link generalised linear model was lowest and as such that was the model chosen and created

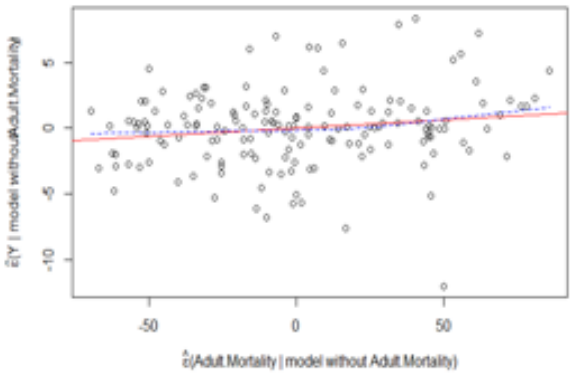
Model Diagnostics

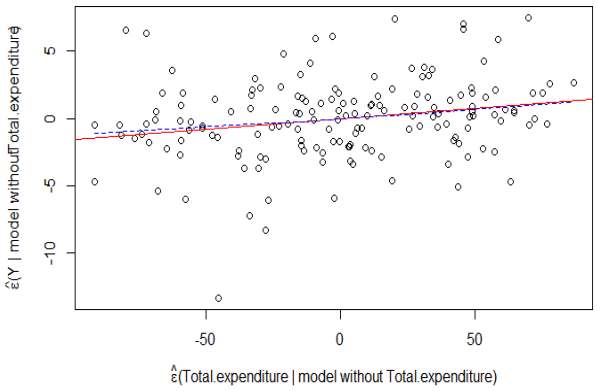
After the model was created some diagnostics were performed. The residual plots for the inverse link GLM and linear model were compared as follows:



These seemed quite similar with some fanning in the left of the graphs and some problems with outliers. This suggests that either the link is wrong, the predictor links need to be changed or incorrect selection of predictors. It was thought that perhaps this was a consequence of incorrect choice of predictor links.

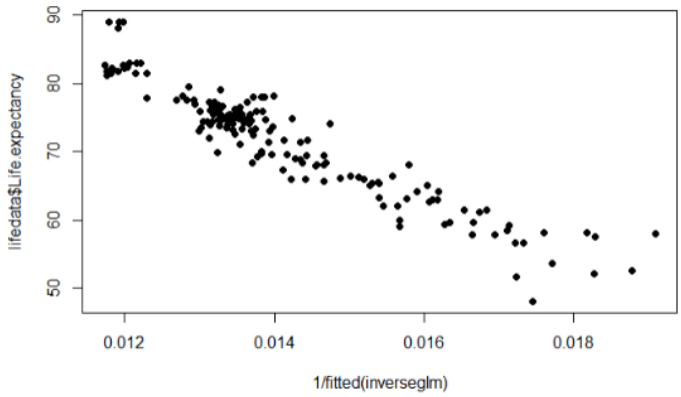
It was thought that perhaps this was a consequence of incorrect choice of predictor links. Therefore, in investigating this, partial residual plots were created. Partial residual plots are residual plots where only the influence of one predictor is present. Two are shown here for reference.





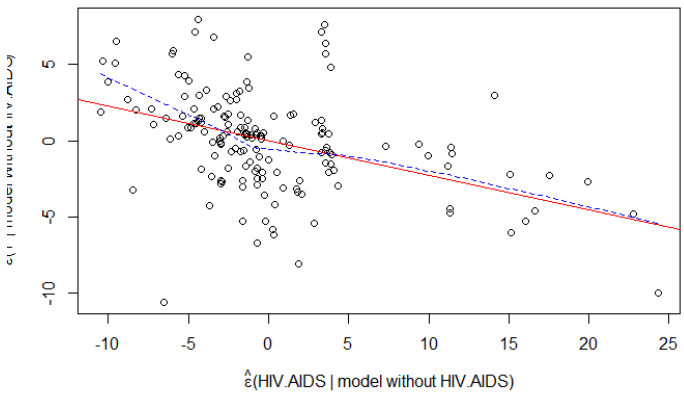
All of the partial residual plots demonstrated linear, homoscedastic residuals. This indicates that there was no problem with the x variable links. The next potential culprit was then investigated: the link function.

To investigate the validity of the link function, the transformed(inversed) fitted values were plotted against life expectancy. This should have been a straight homoscedastic line:

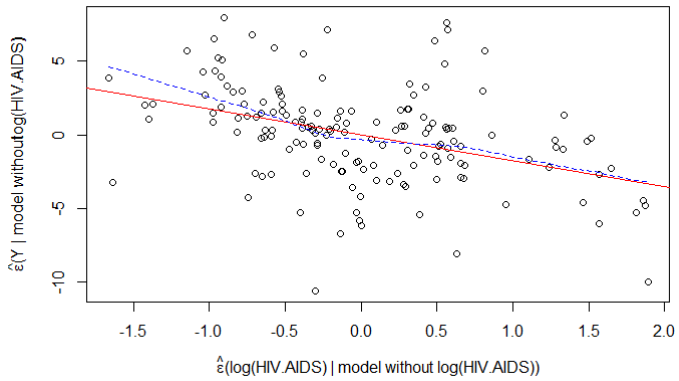


The graph demonstrates significantly more variance in the larger fitted values, and as such this choice is clearly not homoscedastic and it could be concluded that the chosen inverse link function was incorrect.

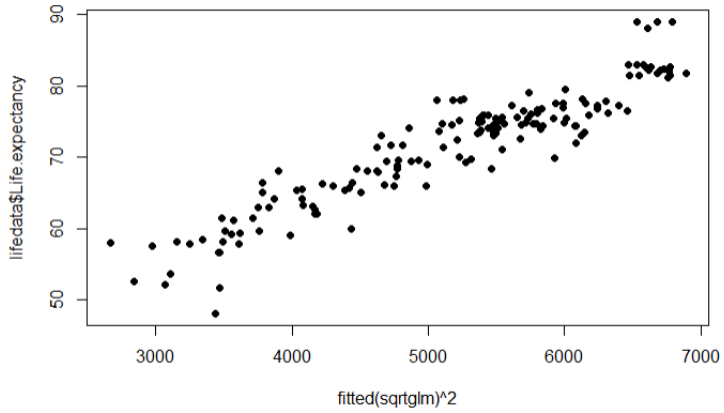
Therefore, to rectify this, the next best link function as used (as chosen by BIC and deviance): the square root link. In contrast to the inverse link gamma GLM, the square root gamma GLM exhibited a problem with a predictor link. The HIV.AIDS variable seemed to require a log link:



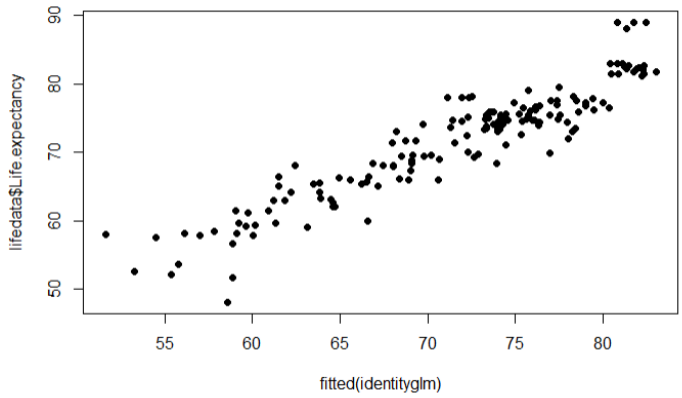
This was rectified, and when replot, the partial residual plot looked much better:



Lastly, the link had to be checked again. When the transformed fitted values were plotted against life expectancy, heteroskedasticity was observed. Again, there seemed to be increased variance in some regions of the transformed fitted values.



The link function had to be changed again. The exact same process was recommenced with the identity link. It was thought that the identity link would be a better choice because so far the linear model had presented the best diagnostic plots. Therefore, an identity link-generalised linear model would replicate the link but not the assumed homoscedasticity. This time when partial residuals were plotted, HIV.AIDS did not require a log link and yet, similarly to the other generalized link models, the link function was incorrect:



It was concluded that the use of a generalised linear model was incorrect, and that the problem could be more accurately modelled with the previously described gaussian linear regression.

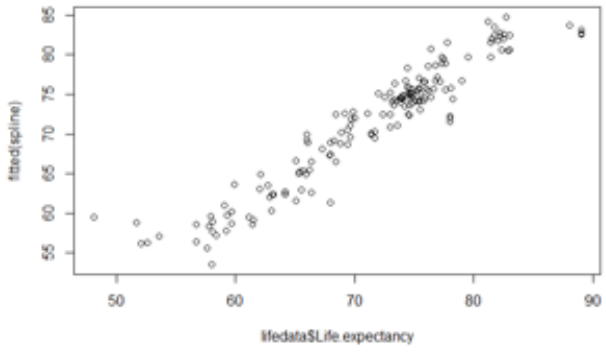
Generalized Additive Model and Splines

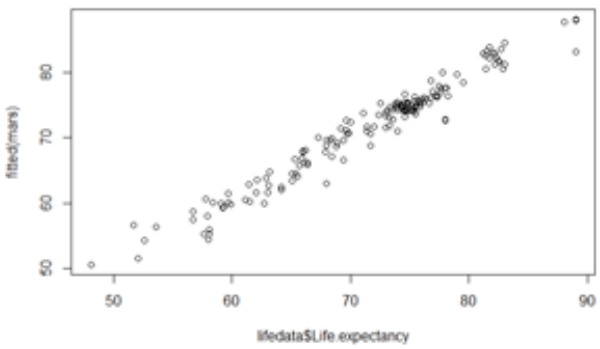
Another possibility is that the data set has some non-linear characteristics which would need to be modeled accordingly. To achieve this, we can implement splines and a generalized additive model (GAM) routine. By giving the spline fits a large degree of freedom, we can observe whether the variables have a non-linear relationship with life expectancy.

It was also considered that a multivariate adaptive regression splines model should be used despite the coefficients being uninterpretable.

Regular polynomial fits are too wild, so for this analysis smoothing spline terms were used, thus trading bias off for better variance. The variables that were selected for the splines were adult mortality, total expenditure, diphtheria, HIV/AIDS, thinness prevalence 5-9 years and income composition of resources. The variables were modeled as smoothing terms.

The two splines were fit, and diagnostic plots created, It was observed that these models fit the data better than the classic regression models as demonstrated by the following homoscedastic plots.





This is merely a reflection of the fact that the models have more flexibility in fitting this data and accordingly will fit the data better. When the fitted splines were observed, however, each demonstrated strictly linear curves. This is a demonstration that even when a model is given massively increased levels of degrees of

freedom, it still reverts to a linear form. This verifies that a linear model is a good choice for the final model, and corroborates earlier decisions.

Using MAE and R^2 as metrics, the smoothing spline GAM routine was found to be a better fit to the data set than a standard linear model using all variables as predictors and the MARS model was found to fit better than the GAM. This follows from the earlier point made that the splines have a greater degree of freedom in fitting to the data.

	Earth <dbl>	Smooth <dbl>	Linear <dbl>
MAE	1.2285341	1.7903115	1.992028
R.squared	0.9589052	0.9013828	0.895075

However, despite this better predictive capability, this report is interested in directly applicable results and interpretations. Splines will not allow for this, and having demonstrated that they merely perform a slightly more flexible linear regression, they can be discarded.

Final Comparison and Results

	InvGLM <dbl>	SqrtGLM <dbl>	Linear <dbl>	PCA <dbl>
AIC	788.171165	815.215195	774.904983	805.566705
BIC	831.135824	839.766428	802.525121	827.049034
MAE	1.916504	2.099253	1.992028	2.099253

The choice to not use training/testing split has resulted in the necessity of performance measures which approximate external measures. As previously explained, AIC and BIC approximate external measures. Mean absolute error, on the other hand, does not approximate an external measure, and thus using it as the only measure of success would be misleading and lead to a choice of a model which has overfit to noise rather than signal.

The best model that can be seen from the diagnostics table above was concluded to be the linear model. The linear model presented the lowest AIC and BIC value however it had the second highest MAE value with the inverse-link GLM model having the lowest MAE, but it also demonstrated a higher BIC and AIC.

It can be concluded that when presented with new data (external data) the basic linear model will demonstrate the best predictive performance. As such, the following results have been found:

	coef.Linear. <dbl>
(Intercept)	72.0872204428
Adult.Mortality	-0.1557812492
Adult.Mortality:infant.deaths	-0.0009879729
Adult.Mortality:Income.composition.of.resources	0.0017956967
infant.deaths:Total.expenditure	0.0006974918
Hepatitis.B:Polio	0.0026758215
Polio:thinness.5.9.years	-0.0011943425
Income.composition.of.resources:HIV.AIDS	-0.0035889174

It can be seen that adult mortality and its interactions with infant deaths and income composition is relevant for estimating a country's life expectancy. This inherently implies that the effect of an increase in adult mortality on life expectancy is different for different levels of infant deaths and income composition.

In particular, a 1 unit increase in adult mortality ceteris paribus (no change in any other variables), would result in a decrease in life expectancy by 0.15 units and a further 0.00099 decrease for each unit in infant deaths. Furthermore, the effect of adult mortality on expected life expectancy will increase by 0.0018 years for each unit in income composition (disparity).

This implies that, while the decrease of adult mortality is extremely important in increasing life expectancy (quite intuitively), its effects must be amplified through a focus on reducing infant deaths and income disparity. It is not enough for a country to focus its efforts on purely the reduction of adult mortality, it must address the root and stem of the problem. This intuitively implies that adult mortality is heavily dependent on income disparity and also infant mortality and that a country must focus its efforts on these three areas in tandem for best results.

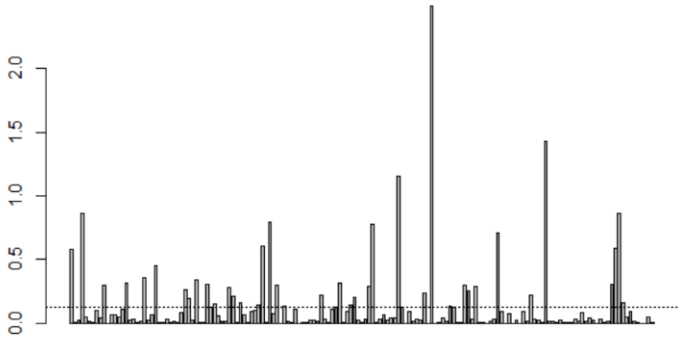
Furthermore, it is observed that various medical factors are relevant for estimation of life expectancy. These include the interaction of infant deaths and total expenditure, the interaction of hepatitis B immunization and polio prevalence, the interaction of polio prevalence and thinness in young children and finally the interaction between income composition and HIV/AIDS prevalence.

In particular, a 1 percent change in the immunization coverage for hepatitis B will result in a 0.00267 year increase in life expectancy for every percentage coverage of Polio prevalence, ceteris paribus. Similarly, a 1 percent change in polio prevalence will decrease life expectancy by 0.0012 years for each percentage in thinness in young children. Finally, it can be seen that for every unit increase in income composition (disparity), life expectancy will decrease for each percentage point of HIV prevalence.

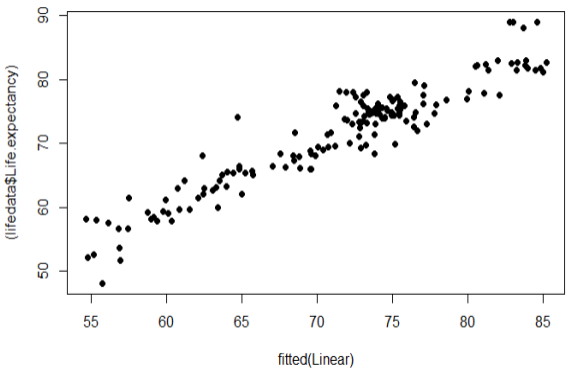
This implies that not only is hepatitis immunization extremely important, it is relatively meaningless without a focus on polio immunization. It can be inferred that this could be due to the fact that if hepatitis B does not kill someone, polio is likely to get that person instead. Therefore, it is only with systematic focus on both of these variables that life expectancy will meaningfully increase. It is also noted that the effect of this polio prevention is drastically increased with a focus on decreasing malnourishment in younger people and income disparity. If these factors are not addressed in tandem, the realized effects will be less than desired and less cost effective.

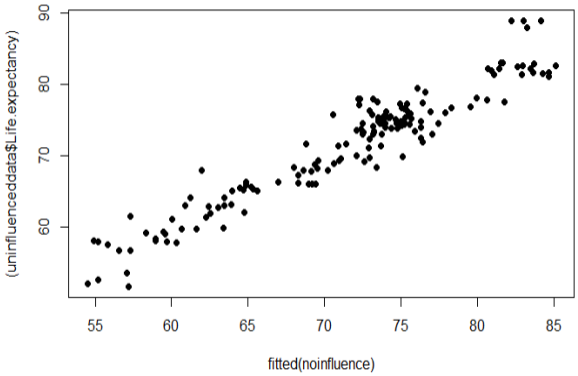
Influence Diagnostics

In order to ascertain the influence that one or more observations have on the model, a technique called Cook's Distance. The purpose of Cook's distance is to identify influential outliers within a set of predictor variables, this is done by plotting a chart of the Cook's distance of each observation in the predictor set:



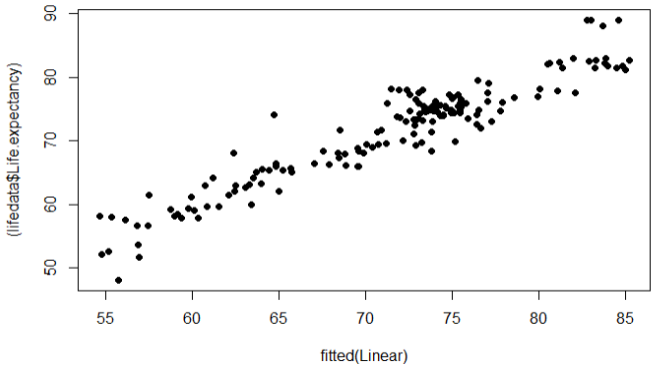
Observations which achieve Cook's distance greater than 0.5 should arouse some suspicion, as such these observations should be investigated for being outliers, but this is not necessarily the case. If an observation has Cook's distance which is greater than 1, it is usually a good assessment to classify this observation as being an outlier. By this assessment, one outlier can be clearly identified as observation 99 which exceeds the upper limit for acceptable non-outliers by more than double. Additionally, there are two further observations which have Cook's distance greater than one at observation 90 and observation 130. By removing these three outlier observations, their influence can be visualized by directly comparing the linear model with the influential observations to the one without.

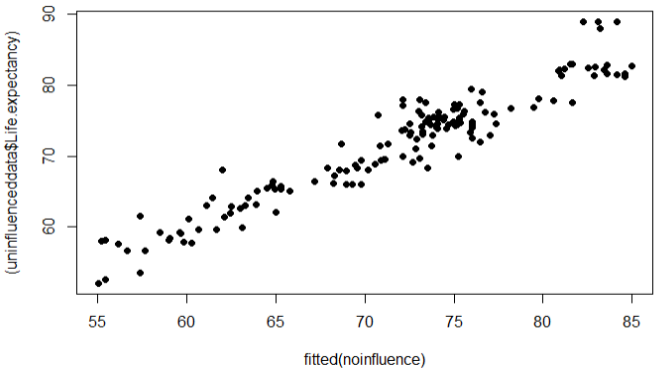




A very subtle difference, but it can be seen the fitted linear model values for life expectancy have a larger intercept and maybe a marginally less steep of a gradient, which is slightly harder to measure visually, but there is a definite upward shift in fitted values on the set without influential observations compared to the set without.

There are also two additional observations that have Cook's distance greater than 0.5, but not greater than 1 which could reasonably be deemed influential. Observations 4 and 150 are both fairly close to attaining a Cook's distance of 1 and stick out from the other observations quite noticeably, for this reason it is fair to assume these observations as outliers and remove them to measure their influence.





It does not appear as though removing these two additional observations has really had a significant influence on the fitted values compared to just removing the three initial influence variables. It might be that it was overzealous to call these observations outliers based on their Cook's distance.