Fertility, Adult Morality, Internet Usage

By Spencer Leo, Trevor Press, Aaron Tam

Introduction

In this analysis, we study the determining factors of general development for countries and consider possible policy implications on quality-of-life development dynamics. Health and economic development for a country is essential to human happiness and well-being for its citizens. Discerning complex links and variables between general health and the economy can help impact different outcomes like poverty or infant mortality rates as health and economic performance are often interlinked. Understanding these relationships can help countries achieve sustained growth for economic and health impacts.

We examine the correlations and relationships between a country's internet availability, mortality rate, and fertility rate and its significance to general development. As part of our analysis, we investigate the relationship amongst these variables using different clustering methods and creating multiple regression models.

Data

For our analysis, we use two sets of country level data from the United Nations Database, one from the World Bank Database, and one from Wikipedia as shown below. We merge the datasets into a singular set by each matching country and examine the following variables:

Fertility

- Fertility rates, total (births per woman) per country in 2010
- Collected from census reports and other statistical publications from national statistical offices
- Retrieved from the World Bank Database

Mortality

- Adult mortality rate (over 35 years of age) per country in 2010
- Deaths per 100 of population 35 years of age and over
- Collected from most recent health statistics for WHO member states
- Retrieved from the United Nations Database

Internet

- Internet usage rate, percent of country's population with access to internet per country in 2010
- Collected from telecommunication regulatory agencies and/or ministries and national statistical offices by means of an annual questionnaire, and subsequently verified, harmonized and complemented by International Telecommunications Union
- Retrieved from the United Nations Database

Continent

- List of countries by continent from Wikipedia.

Methods

Clustering:

For our analysis we implemented several clustering techniques including partitioning, agglomerative and divisive hierarchical clustering, and density-based clustering. We used all three of our variables (internet usage, fertility rate, and adult mortality rate) when clustering the data. Using these variables, we determined that 4 clusters allowed partitioning to have no outliers, therefore, we clustered countries into 4 groups for all clustering techniques.

Partitioning:

For a partitioning method we used partitioning around medoids (PAM). This clustering technique finds the most centrally located point and determines clusters based on which points are more closely related to that centrally located point. This centrally located point is representative of the cluster. Using the PAM method is more robust in the presence of outliers than k-means clustering which instead of using a point uses means of all points in a cluster to be representative of the cluster.

Hierarchical clustering:

Hierarchical clustering finds all possible ways that the data can be clustered individually and in subgroups following a tree approach. For our research we used both agglomerative and divisive methods. Agglomerative methods start by clustering individual data points and slowly increasing the size of clusters. Divisive methods work backwards, starting with everything in the same cluster and then breaking up clusters into smaller, more descriptive groups. Both of these methods are affected by outliers.

Density Based Clustering:

The final method we used for grouping our data was density-based clustering. Density based clustering works by specifying a minimum number of cases that must exist within a given space to be considered a cluster. This allows for specifying clusters that are less clear than with other methods. Density-based clustering allows outliers to be left out of any cluster which makes it the most effective method that we used in our analysis.

Regression Modeling:

For our regression models we test the following hypotheses:

- Hypothesis 1: Live births per woman will increase as mortality rate increases and internet usage decreases
- Hypothesis 2: Live births per woman will increase as mortality rate increases and internet usage decreases controlling for continent
- Hypothesis 3: Live births per woman will increase at a decreasing rate as mortality rate increases and internet usage decreases
- Hypothesis 4: Live births per woman will increase at a decreasing rate as mortality rate increases and internet usage decreases controlling for continent

Hypothesis 3 and 4 are sensitivity analyses to test the assumption that mortality rate may actually have a quadratic relationship with fertility rate as opposed to a linear relationship. Thus, we used a squared term for mortality rate in these regression equations. Hypothesis 2 and 4 control for the effect of continent on fertility rate as we assume that countries on the same continent may hold similar patterns for their fertility rates.

We take two approaches to the regression analysis. The first approach is an explanatory approach where we run a general linearized regression model on the 4 hypotheses and find the best fitting model. We then test the fit of that model with the data and test the assumptions of that model in mathematical and graphical form. The second approach is a predictive approach where we train a machine learning algorithm on a subset of the data to predict the fertility rate given our strongest model. We then see the coefficients of the machine learning algorithm afterwards and see how well it fits the data.

Results

Clustering:

Of our supervised clustering techniques, we found divisive hierarchical clustering to be the most representative of the data. This was the only unsupervised clustering method that yielded no negative silhouette widths (see figures 1, 2, & 3). A great majority of the cases were in a single group with 98 out of 158 cases in cluster 1. Clusters 2 through 4 had 35, 13, and 12 cases respectively. Each cluster had a width between .42 and .68.

Density-based clustering yielded similar results. 125 of the cases were in group one, 8 in group two, 4 in group three, and 8 in group four. When using density based clustering we found 13 outlier countries. It appears that the largest cluster represented low-income countries with fairly undeveloped health care systems. Group one, on average, has high fertility, high mortality, and low internet usage. Group two, on average, has slightly higher fertility rate than groups two and three, low mortality, and the highest internet usage. Group three, on average, has the lowest fertility rate, similar but slightly higher mortality rate to groups two and four, and internet usage between that of groups two and four. Group four, on average, has slightly higher fertility rate than group three, the lowest mortality rate, and fairly low internet usage. Average fertility rates, adult mortality rates, and percent internet usage can be seen in figure 7. Due to the breakdowns of averages for our variables, it appears that clusters represent differing levels of development with the majority of countries having less development.

Regression modeling:

In all of the regressions, internet usage had no significant relationship with fertility rates at the 5% level of significance except for model 1 which did not include continent as a control variable or mortality rate as a quadratic variable. The p-value for internet usage was .00317 in model 1, but the coefficient was -0.02370 which showed a very minimal influence. In model 2 and 4, where continent was controlled for, the models had higher R-squared goodness of fit values. You can see that the results of model 2 and 4 are pretty similar in figure 8, but model 4 fits better (with an R-squared value of 0.7657).

The coefficients for model 4 indicate that mortality rate generally increases with fertility rate (17 live births per woman with one more death per 100 people); however, it has less of an effect the greater the mortality rate is. The coefficients also indicate that in comparison to Africa, most continents have lower fertility rates, and the countries in the EU have the lowest fertility rate in comparison with an average of -1.7 less live births per woman. The trained predictive model using model 4 had very similar results to the explanatory version, and it had an R-squared value of 0.8337. See figure 16 for the results of the model.

Model 2 was demonstrated to not pass the assumption of normality with a p-value of 0.00244 for the Shapiro-Wilk normality test. It also did not pass the homoskedascity test with a p-value that was 3.065e-05 for the studentized Breusch-Pagan test. See figure 9 and 10 to see model 2's results. Unlike model 2, model 4 passed the normality test (see figure 11). It did not pass the homoskedascity test, but the p-value for the studentized Breusch-Pagan test was much lower at 0.01645 (see figure 12). This demonstrates that the use of a quadratic term for the mortality rate resulted in a more accurate regression model. Other tests on the assumptions of model 4 were conducted such as checking the linearity between the dependent variable and predictors, checking correlation between predictors, and analyzing the effect of whether outliers are influential. See the test results in figures 11-15. The tests confirmed that there is linearity

between the dependent variable and predictors, that there is little correlation between the predictors, and the outliers have less influence.

Discussion

Using different clustering techniques, we categorized countries on varying levels of development. Our analysis shows using internet usage had no significant relationship with fertility rates for most of our models. Quality-of-life indicators like internet usage may not necessarily be the best predictor of population dynamics in countries. The internet was still a novel technology in 2010 and did not have high usage even for the most developed countries. Other indicators like general income, inequality, education, type of government and healthcare and religious views toward conceptratives can play larger roles on fertility rates. In addition, external events like political upheaval and civil wars can drastically affect results. Generally, it is understood there is an inverse relationship between development and total fertility rates within nations. Demographic transitions lead to a change in the supply and demand of labor, thus affecting the labor market.

Assuming this relationship, the goal of decreasing fertility rates may best be served by focusing on contraceptive education, family planning education, and improved life expectancy instead of focusing on improvements in economic well-being like access to the internet. Of course, additional research and other variables should be investigated for a better understanding of development in relation to fertility rates.

Appendix

Figure 1: Partitioning Silhouettes

Figure 2: Agglomerative Hierarchical Clustering Silhouettes

Figure 3: Divisive Hierarchical Clustering Silhouettes

Figure 4: Supervised Clustering Comparison

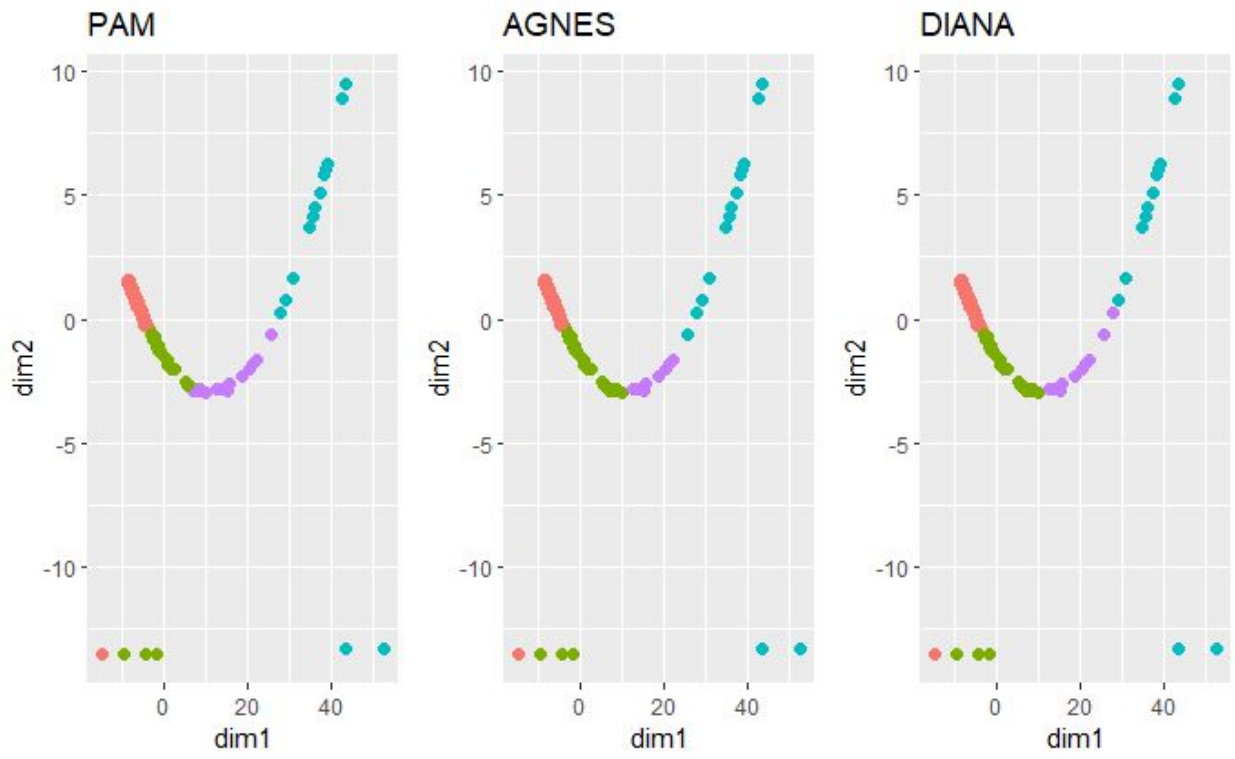

Figure 5: Density Based Clustering

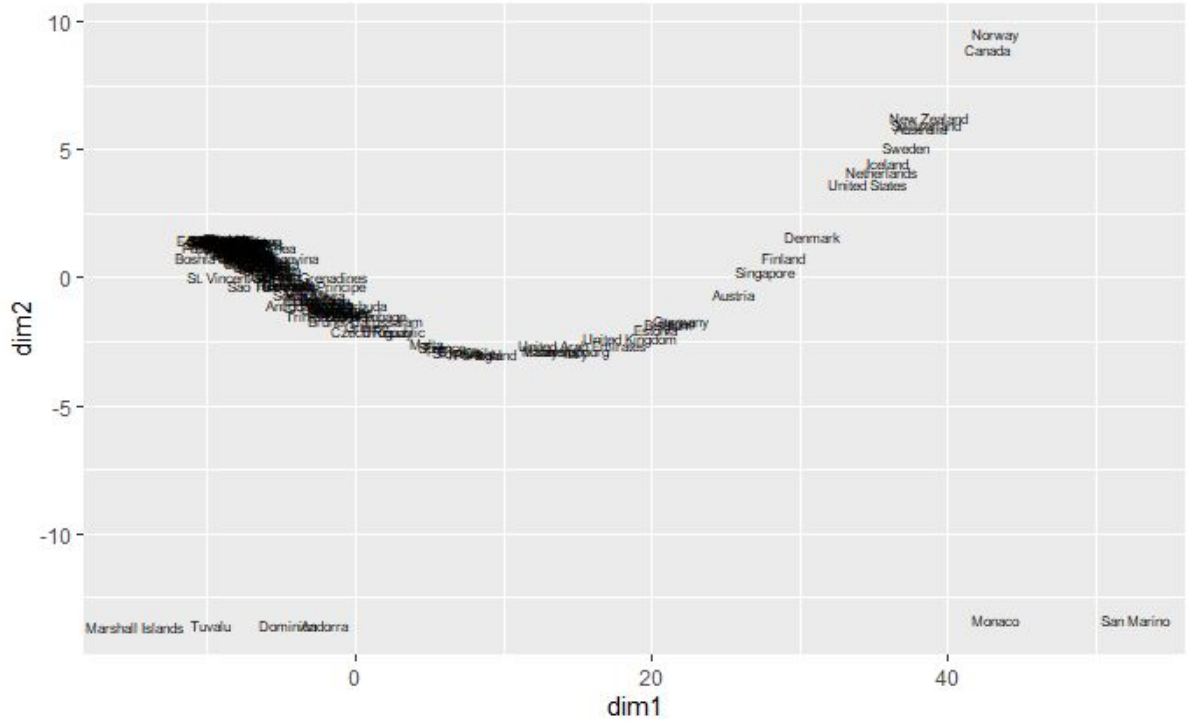

Figure 6: Outliers from Density Based Clustering

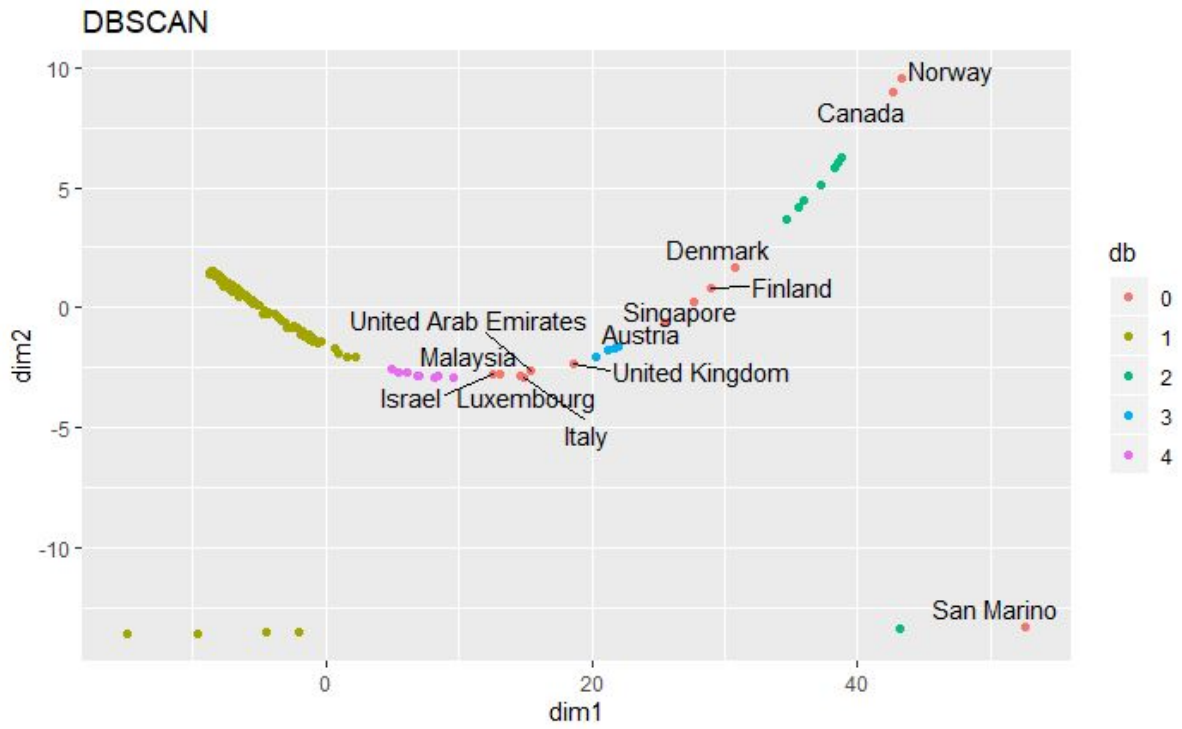

Figure 7: Group Averages

Groups	Fertility Rate	Adult Mortality Rate	Internet Usage
0 (Outliers)	1.903000	0.06623077	33.611869
1	3.654769	0.21788800	2.417138
2	1.804571	0.06250000	45.080279
3	1.442250	0.07500000	29.553933
4	1.657625	0.06200000	15.286648

Figure 8

Hypothesis 2 Results:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 3.85351 0.33674 11.444 < 2e-16 ***
MortalityRate1 3.85245 0.85469 4.507 1.35e-05 ***
X.InternetUsage -0.01328 0.00728 -1.825 0.070098 .
ContinentCodeAS -1.57149 0.28508 -5.512 1.59e-07 ***
ContinentCodeEU -2.39473 0.32253 -7.425 9.08e-12 ***
ContinentCodeNA -1.50730 0.32489 -4.639 7.78e-06 ***
ContinentCodeOC -0.62563 0.38352 -1.631 0.105014
ContinentCodeSA -1.44093 0.37249 -3.868 0.000165 ***
R-squared 0.699724

Hypothesis 4 Results:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.872137 0.427779 4.376 2.31e-05 ***
MortalityRate1 17.450006 2.240804 7.787 1.26e-12 ***
MortalityRate2 -19.223314 2.982669 -6.445 1.66e-09 ***
X.InternetUsage 0.004692 0.007010 0.669 0.50435
ContinentCodeAS -1.074269 0.263381 -4.079 7.49e-05 ***
ContinentCodeEU -1.715911 0.303758 -5.649 8.43e-08 ***
ContinentCodeNA -1.051267 0.295589 -3.557 0.00051 ***
ContinentCodeOC -0.474914 0.339593 -1.398 0.16413
ContinentCodeSA -0.919237 0.338857 -2.713 0.00749 **
R squared 0.7656866

Figure 9
Model 2: Shapiro-Wilk normality test
W = 0.97061, p-value = 0.002436

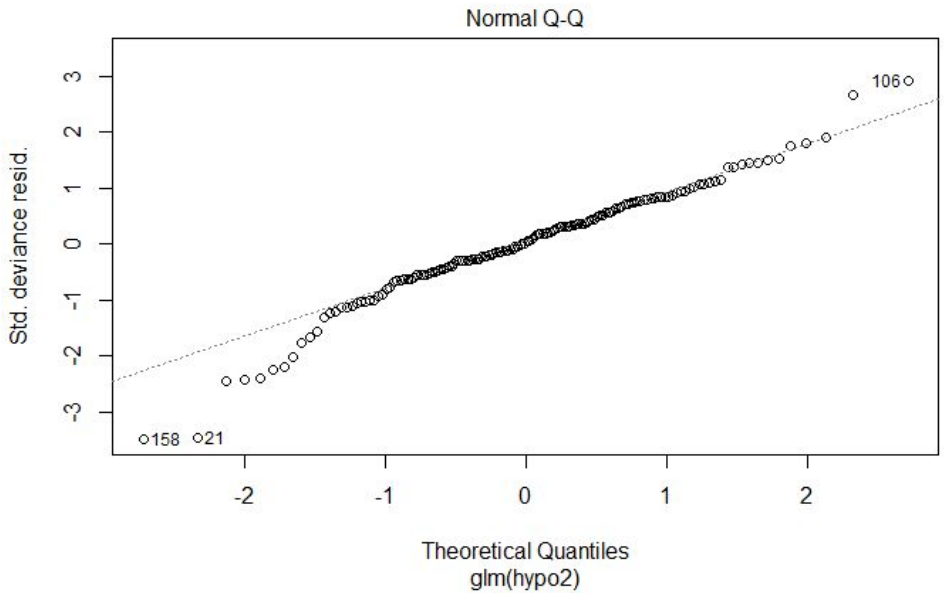

Figure 10
Model 2: studentized Breusch-Pagan test
BP = 32.659, df = 7, p-value = 3.065e-05

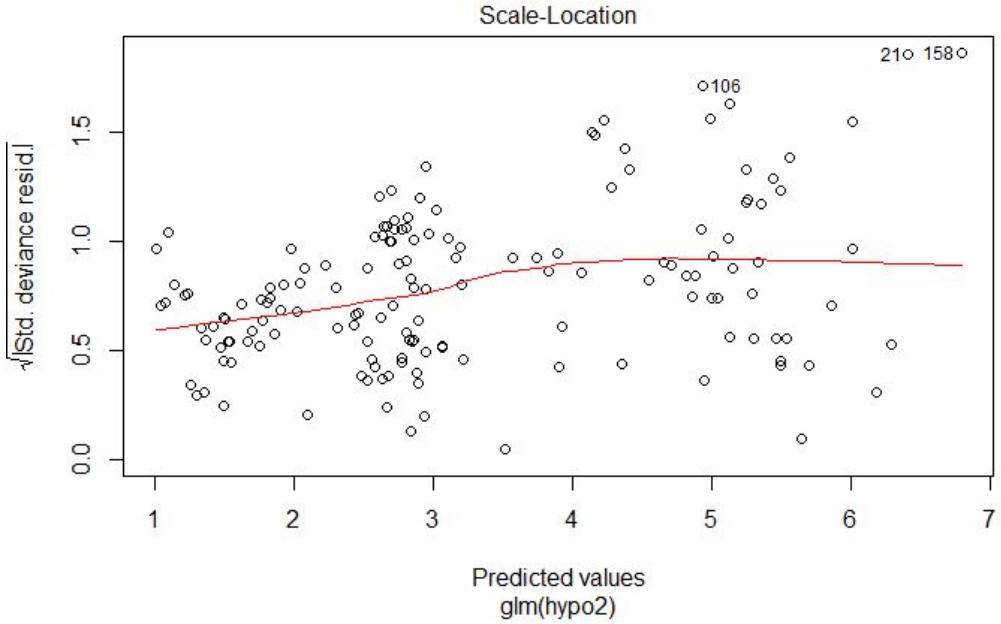

Figure 11
Model 4: Shapiro-Wilk normality test
W = 0.99307, p-value = 0.6783

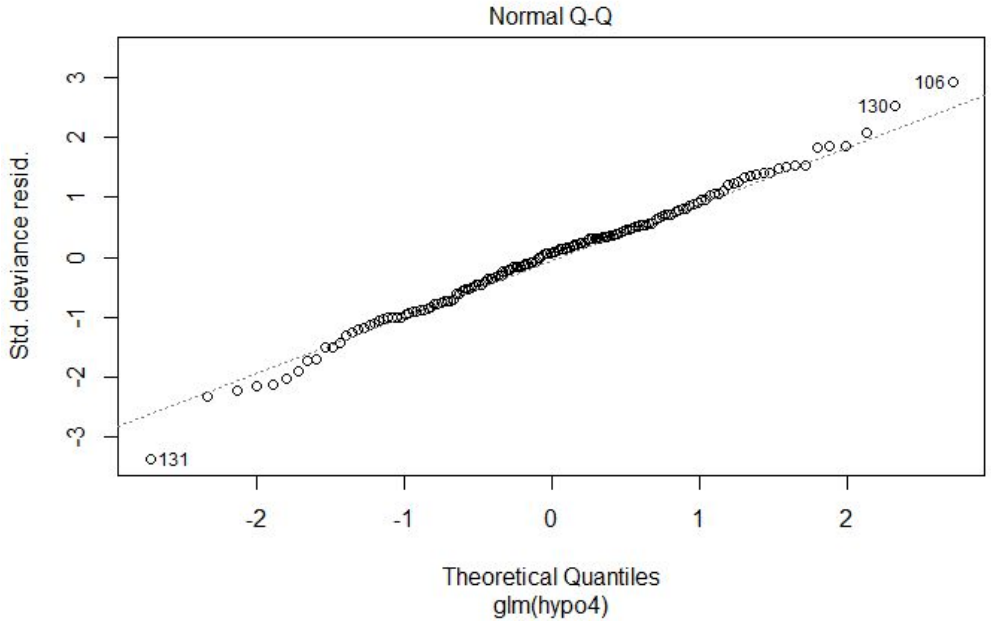

Figure 12
Model 4: studentized Breusch-Pagan test

BP = 18.716, df = 8, p-value = 0.01645

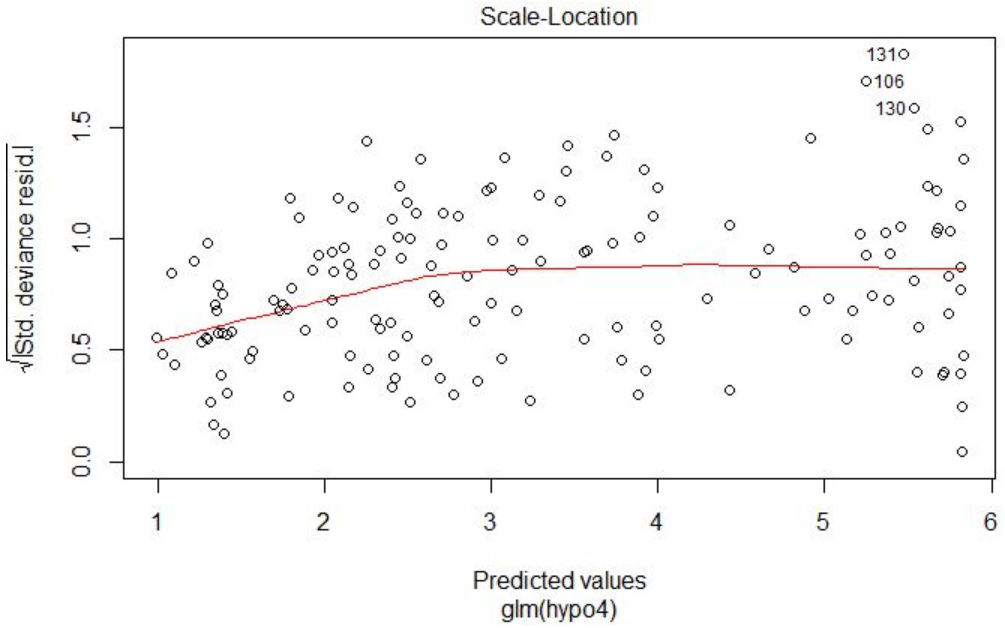

Figure 13
Model 4: Looking for linearity between dependent variable and predictors

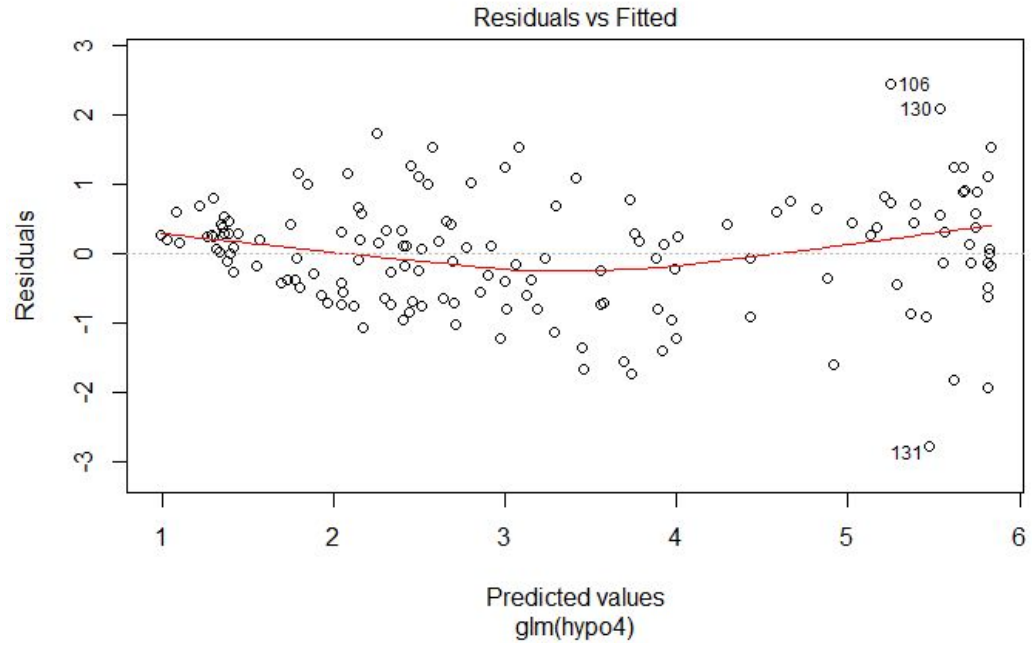

Figure 14
Model 4: Seeing the influence of outliers

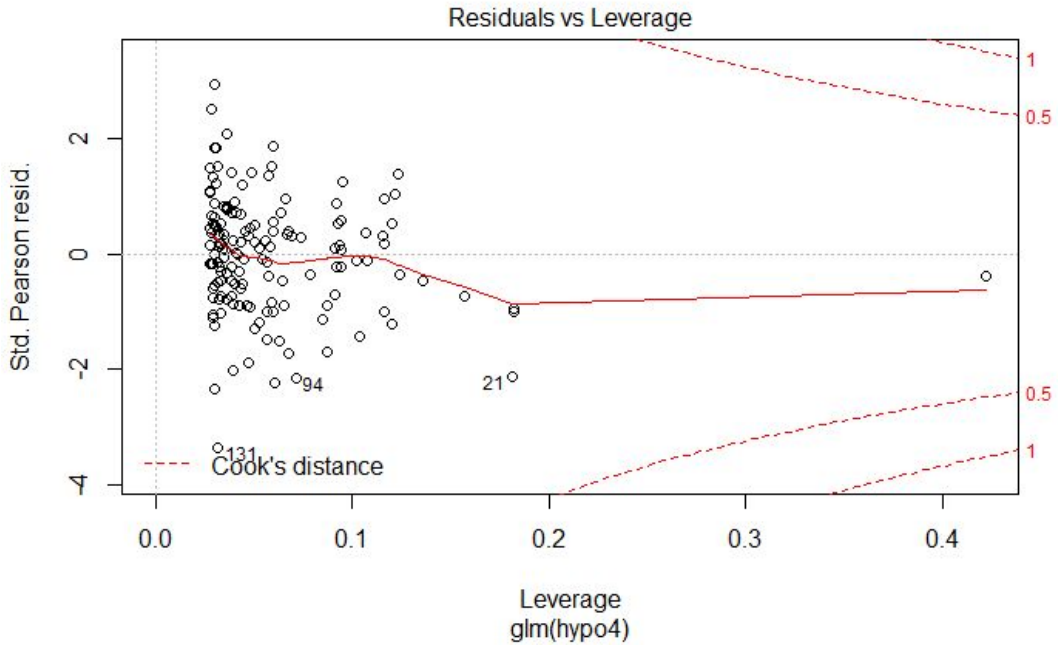

Figure 15
Model 4: Plotting the coefficients of model 4

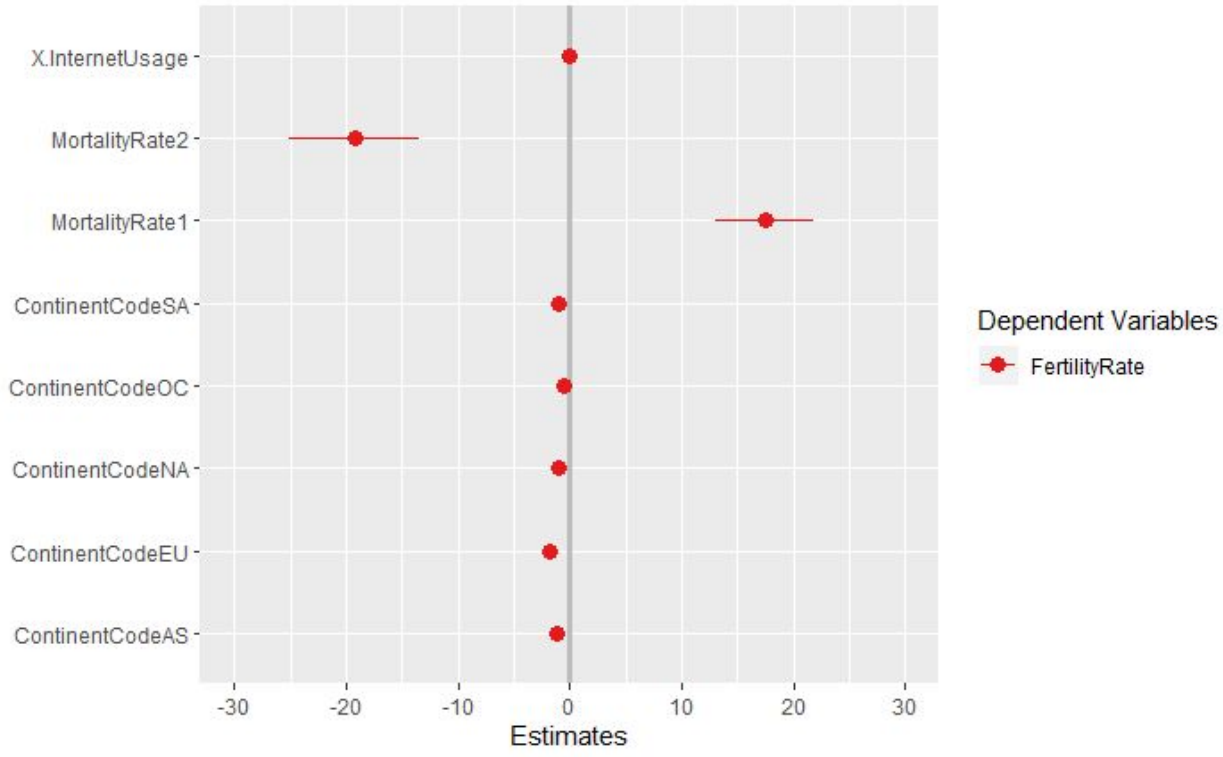

Figure 16
Predictive approach using model 4 results

Predictive Approach Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	2.154889	0.519485	4.148	6.74e-05	***
MortalityRate1	15.873550	2.604365	6.095	1.76e-08	***
MortalityRate2	-17.654878	3.312427	-5.330	5.50e-07	***
X.InternetUsage	0.002775	0.008507	0.326	0.744910	
ContinentCodeAS	-1.096045	0.315731	-3.471	0.000748	***
ContinentCodeEU	-1.854045	0.365461	-5.073	1.66e-06	***
ContinentCodeNA	-1.079514	0.343678	-3.141	0.002177	**
ContinentCodeOC	-0.443848	0.442612	-1.003	0.318224	
ContinentCodeSA	-1.195706	0.414742	-2.883	0.004762	**

RMSE	Rsquared	MAE
0.8337601	0.8029709	0.6536