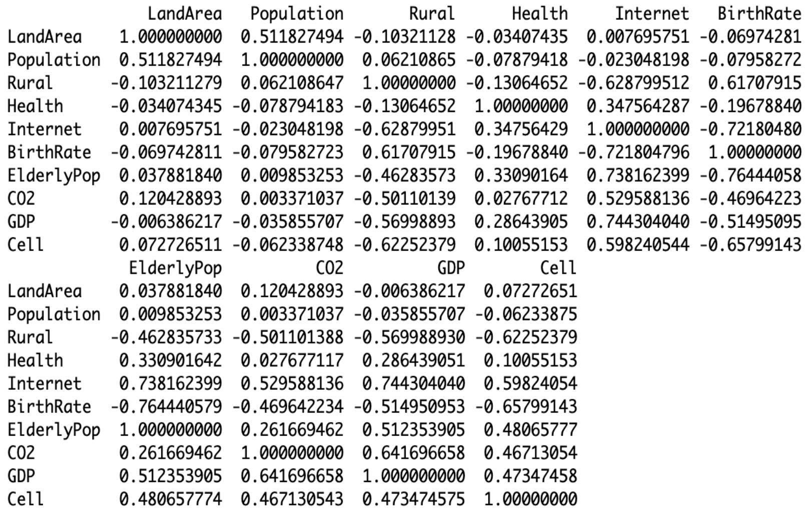
Alex Garcia
Brandon Sparks

Executive summary: This report explains the process of model selection of a multiple regression model, with life expectancy as the outcome variable and all the other data as predictor variables using data from the data set "country.csv". The goal was to find a model that fits the most significant variables that affect life expectancy the most. Using model selection and model building we find a model that best fits the most significant variables that affect life expectancy the most. After undergoing model selection using Stepwise, AIC, backwards, and forward selection, and using Adjusted r squared and mallows cp as creation, we obtained a model with predictors Birth Rate, GDP, Internet, Cell, and Land Area. The model was a reasonable fit, but it is highly probable that there is a model(s) that is a better fit based on the residual plots and behavior of the model.

INTRODUCTION:

This report intends to assess if there is a linear relationship between life expectancy and the following variables; Land area, population, percentage of population living in rural areas, percentage of government expenditures directed towards health care, percentage of population with internet access, Births per 1000 people, percentage of population at least 65 years old, Average life expectancy in years, CO2 emissions in metric tons per capita, Gross Domestic Product per capita, Cell phone subscriptions per 100 people. Using the Data set country.csv, the report will show the results of model building and model selection about a multivariable linear regression model of life expectancy on the predictor variables, residual analysis and diagnostics on said model, any possible transformation for variables, any violation of model assumptions, if there is multicollinearity between the predictors, and finally select a valid model with the most significant variables that predict life expectancy

Process:Undergoing the preliminary steps of multiple linear regression, we first look at the simple graphs of each predictor variable, we will omit the name and code of each country since they are not a quantitative variable through this entire process. Simple variable descriptive graphs and linear regression for each predictor variable Looking at the scatter plots of all the predictor variables it seems that Rural, Birth Rate, Health, and Cell seem to have the strongest linear behavior and most linear relation in terms of Life expediency. While Land area, Population, Internet, Elderly population, CO2 emissions, and GDP all have weaker and less linearity. We also assess multicollinearity, obtaining the multicollinearity matrix for the multiple regression model. We see that the Internet, Birth Rate, and GDP have a correlation level of around .7 which is a notable observation that we continue as we go through the process



```
              LandArea   Population       Rural      Health     Internet   BirthRate
LandArea    1.000000000  0.511827494 -0.10321128 -0.03407435  0.007695751 -0.06974281
Population  0.511827494  1.000000000  0.06210865 -0.07879418 -0.023048198 -0.07958272
Rural      -0.103211279  0.062108647  1.00000000 -0.13064652 -0.628799512  0.61707915
Health     -0.034074345 -0.078794183 -0.13064652  1.00000000  0.347564287 -0.19678840
Internet    0.007695751 -0.023048198 -0.62879951  0.34756429  1.000000000 -0.72180480
BirthRate  -0.069742811 -0.079582723  0.61707915 -0.19678840 -0.721804796  1.00000000
ElderlyPop  0.037881840  0.009853253 -0.46283573  0.33090164  0.738162399 -0.76444058
CO2         0.120428893  0.003371037 -0.50110139  0.02767712  0.529588136 -0.46964223
GDP        -0.006386217 -0.035855707 -0.56998893  0.28643905  0.744304040 -0.51495095
Cell        0.072726511 -0.062338748 -0.62252379  0.10055153  0.598240544 -0.65799143
             ElderlyPop          CO2          GDP        Cell
LandArea    0.037881840  0.120428893 -0.006386217  0.07272651
Population  0.009853253  0.003371037 -0.035855707 -0.06233875
Rural      -0.462835733 -0.501101388 -0.569988930 -0.62252379
Health      0.330901642  0.027677117  0.286439051  0.10055153
Internet    0.738162399  0.529588136  0.744304040  0.59824054
BirthRate  -0.764440579 -0.469642234 -0.514950953 -0.65799143
ElderlyPop  1.000000000  0.261669462  0.512353905  0.48065777
CO2         0.261669462  1.000000000  0.641696658  0.46713054
GDP         0.512353905  0.641696658  1.000000000  0.47347458
Cell        0.480657774  0.467130543  0.473474575  1.00000000
```

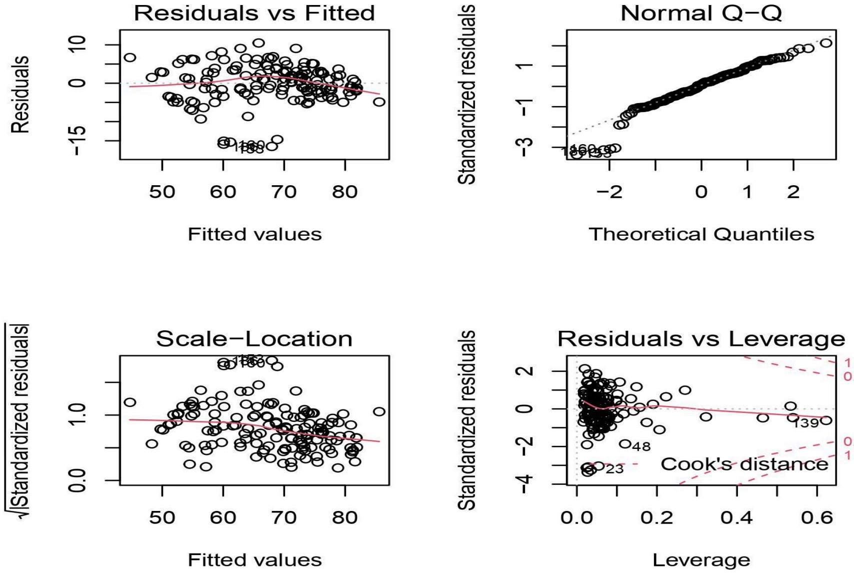
.

We also obtain a vif for the predictor variables and see that there is no correlation level bigger than 5, thus the predictor variables are not that correlated with each other

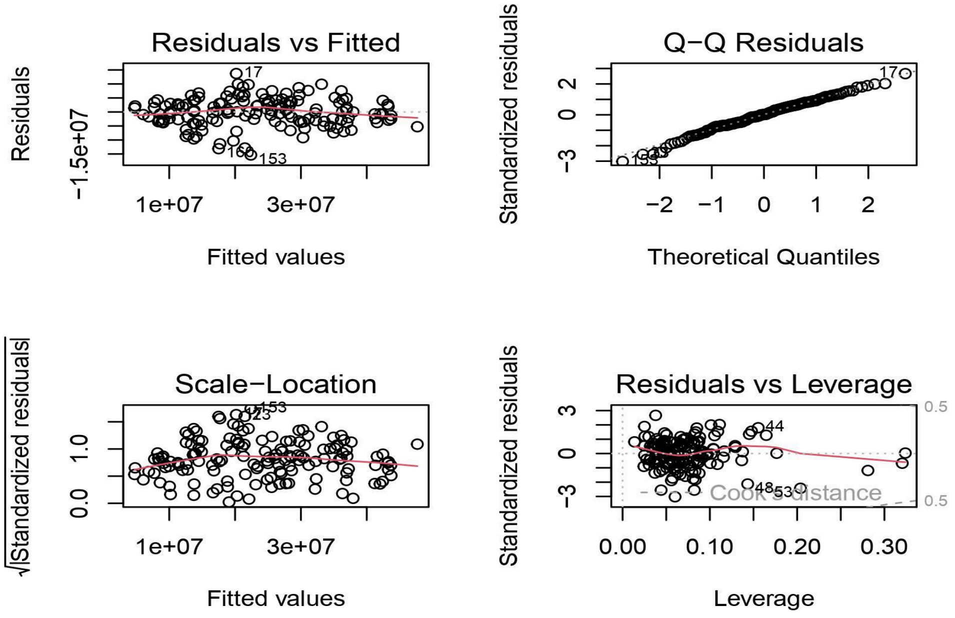
```
  LandArea Population      Rural     Health   Internet  BirthRate ElderlyPop        CO2        GDP 
  1.470218   1.483266   2.235445   1.248738   4.434731   4.056270   3.531392   2.199200   3.022116 
      Cell 
  2.166401 
```

.

Upon fitting an original multiple regression model with all the untransformed predictor variables and performing diagnostics on the model, we see that there is some constant error and normality is violated at the tails

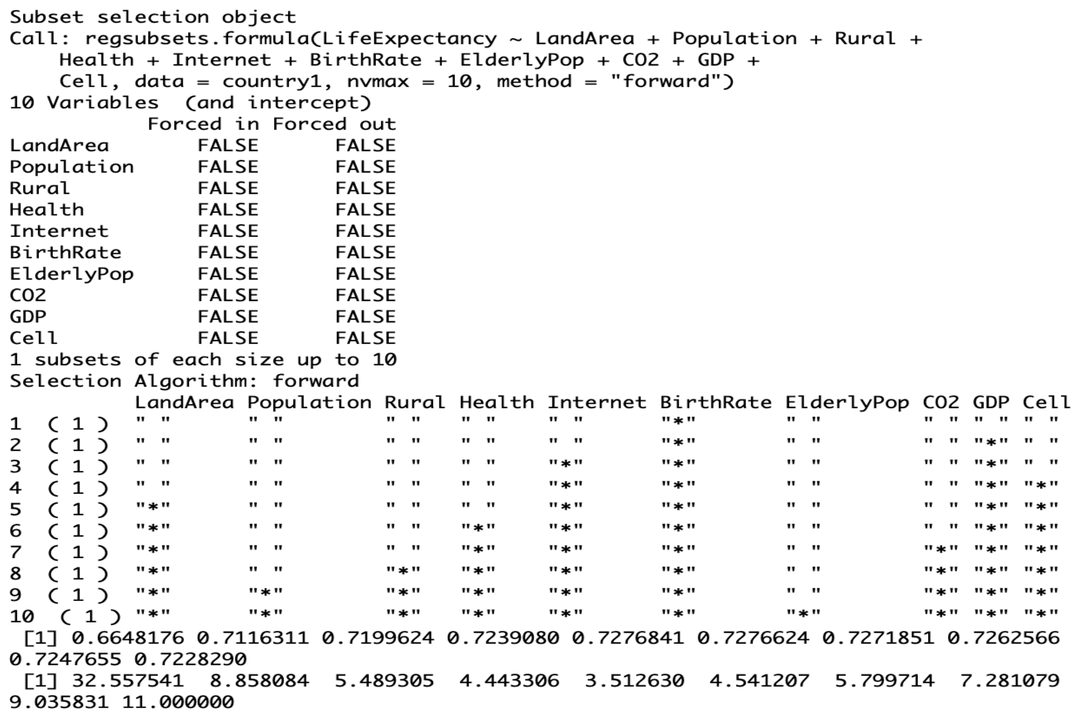


so we perform a Boxcox on the original multiple regression model for life expectancy. Which produces a lambda of 4, Which suggests to raise life expectancy to the 4th power so we then raise life expectancy to the 4th power in the multiple regression model.. Next We then look at the simple linear regression model to assess possible transformation for each predictor variable. Upon executing residual analysis for all simple linear regression models we obtain possible transformation for each predictor variable except rural and health since they do not violate linearity in their residual plots or scatter plots. We then fit these transformed variables including the box cox transformation on life expectancy, into a new data frame called "country1". And use these transformed variables into a new multiple regression model for life expectancy using the data from the data frame "country1" with all of the transformed variables, including the box cox transformation on life expectancy. Looking at the residual plots of the new model, the residual plots do look more linear and follow normality and linearity better than the original model,



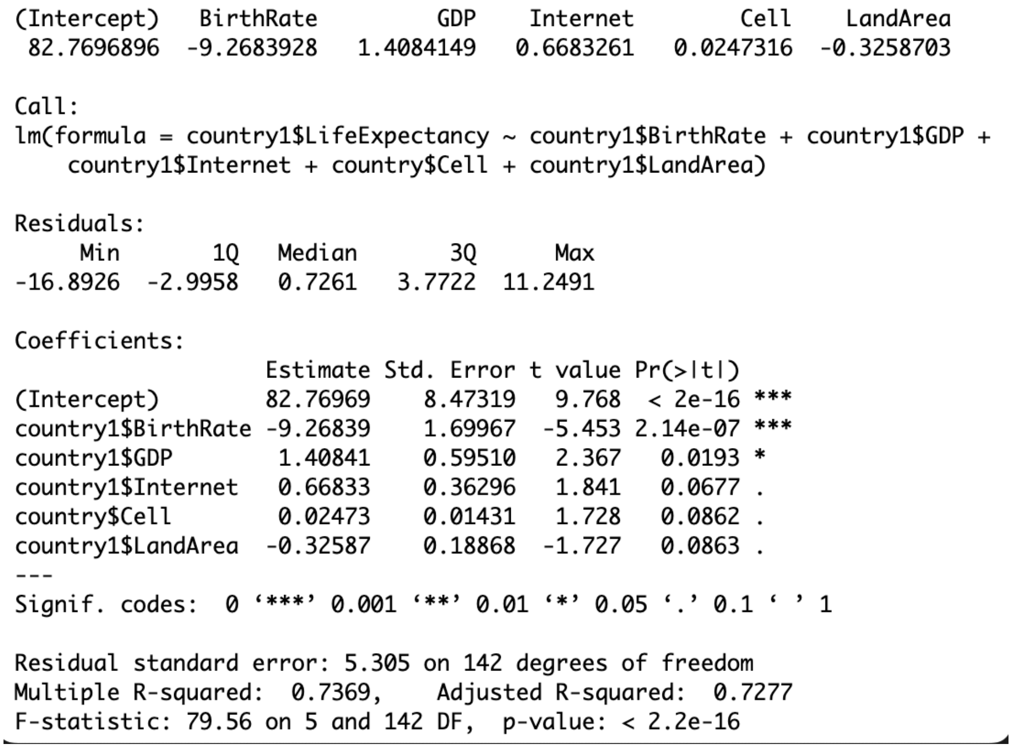
Now we use this transformed multiple regression model for model selection and building. The first step we took in model building is using the step function which executes stepwise model selection using AIC, first we use the step function with forward selection, then we use the same step function again but using backwards selection, both methods produce the same model , with 5 predictor variables of Birthrate, GDP, Internet, Cell, and Land Area. We use this model to compare with the models that are obtained in the next step.

Next we use the regsubsets function from leaps to find the models with the most significant variables. Executing the function for both forwards and backwards selection and using adjusted r squared and Mallows Cp as criterion to judge our models, leaps provides us with the same two models. Using adjusted r-squared as criterion we obtain a model with 7 predictor variables, using mallows cp we obtain a model with 9 predictor variables with the first line of output code at the bottom the adjusted r squared values and the second line of output code the mallows cp values.



```
Subset selection object
Call: regsubsets.formula(LifeExpectancy ~ LandArea + Population + Rural +
    Health + Internet + BirthRate + ElderlyPop + CO2 + GDP +
    Cell, data = country1, nvmax = 10, method = "forward")
10 Variables  (and intercept)
           Forced in Forced out
LandArea       FALSE      FALSE
Population     FALSE      FALSE
Rural          FALSE      FALSE
Health         FALSE      FALSE
Internet       FALSE      FALSE
BirthRate      FALSE      FALSE
ElderlyPop     FALSE      FALSE
CO2            FALSE      FALSE
GDP            FALSE      FALSE
Cell           FALSE      FALSE
1 subsets of each size up to 10
Selection Algorithm: forward
          LandArea Population Rural Health Internet BirthRate ElderlyPop CO2 GDP Cell
1  ( 1 )  " "      " "        " "   " "    " "      "*"       " "        " " " " " " 
2  ( 1 )  " "      " "        " "   " "    " "      "*"       " "        " " "*" " " 
3  ( 1 )  " "      " "        " "   " "    "*"      "*"       " "        " " "*" " " 
4  ( 1 )  " "      " "        " "   " "    "*"      "*"       " "        " " "*" "*" 
5  ( 1 )  "*"      " "        " "   " "    "*"      "*"       " "        " " "*" "*" 
6  ( 1 )  "*"      " "        " "   "*"    "*"      "*"       " "        " " "*" "*" 
7  ( 1 )  "*"      " "        " "   "*"    "*"      "*"       " "        "*" "*" "*" 
8  ( 1 )  "*"      " "        "*"   "*"    "*"      "*"       " "        "*" "*" "*" 
9  ( 1 )  "*"      "*"        "*"   "*"    "*"      "*"       " "        "*" "*" "*" 
10  ( 1 ) "*"      "*"        "*"   "*"    "*"      "*"       "*"        "*" "*" "*" 
 [1] 0.6648176 0.7116311 0.7199624 0.7239080 0.7276841 0.7276624 0.7271851 0.7262566
0.7247655 0.7228290
 [1] 32.557541  8.858084  5.489305  4.443306  3.512630  4.541207  5.799714  7.281079
9.035831 11.000000
```

Finally we are left with three models



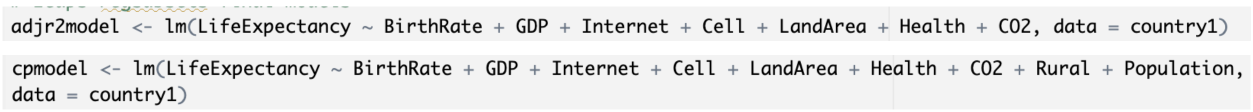
```
(Intercept)   BirthRate         GDP    Internet        Cell    LandArea 
 82.7696896  -9.2683928   1.4084149   0.6683261   0.0247316  -0.3258703 

Call:
lm(formula = country1$LifeExpectancy ~ country1$BirthRate + country1$GDP + 
    country1$Internet + country$Cell + country1$LandArea)

Residuals:
     Min       1Q   Median       3Q      Max 
-16.8926  -2.9958   0.7261   3.7722  11.2491 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)        82.76969    8.47319   9.768  < 2e-16 ***
country1$BirthRate -9.26839    1.69967  -5.453 2.14e-07 ***
country1$GDP        1.40841    0.59510   2.367   0.0193 *  
country1$Internet   0.66833    0.36296   1.841   0.0677 .  
country$Cell        0.02473    0.01431   1.728   0.0862 .  
country1$LandArea  -0.32587    0.18868  -1.727   0.0863 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.305 on 142 degrees of freedom
Multiple R-squared:  0.7369,	Adjusted R-squared:  0.7277 
F-statistic: 79.56 on 5 and 142 DF,  p-value: < 2.2e-16
```

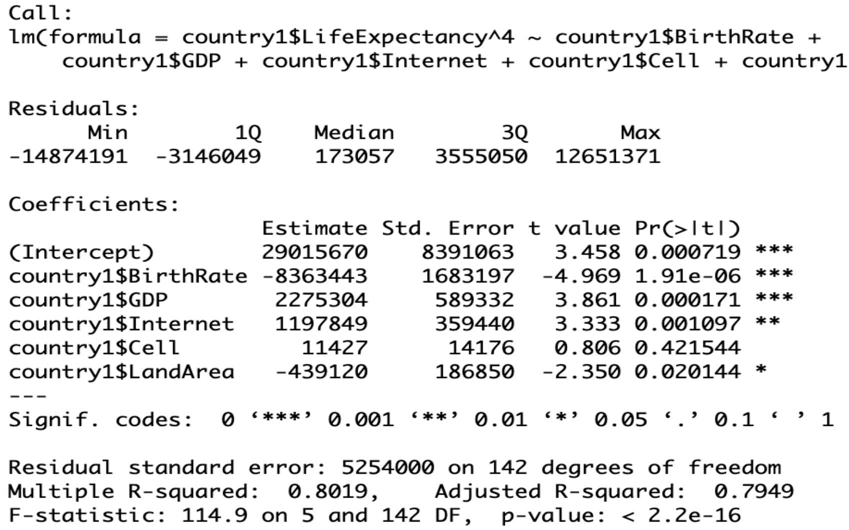


```
adjr2model <- lm(LifeExpectancy ~ BirthRate + GDP + Internet + Cell + LandArea + Health + CO2, data = country1)
```

```
cpmodel <- lm(LifeExpectancy ~ BirthRate + GDP + Internet + Cell + LandArea + Health + CO2 + Rural + Population,
data = country1)
```

Using adjusted r squared as our criteria, the best model with the most significant variables that affect life expectancy is provided by the step function with 5 variables, with a slightly larger adjusted r squared of .7277 Running diagnostics on this final model, there appears to be some non constant error and heavy tails in the qq plot, but executing the box cox transformation on this final model we obtain a lambda of 4, So we transform life expectancy in the final model by raising it to the 4, giving us our final model.

Results:
Upon going through the model selection process using regsubsets function from leaps, as well as using the step function using both backwards and forwards selection with the variables, we end up with a multiple linear regression using the transformed data from "country1", with Births per 1000 people, GDP per capita , Percentage of population with internet access, Cell phone subscriptions per 100 people, and Land Area that theoretically affect life expectancy the most.

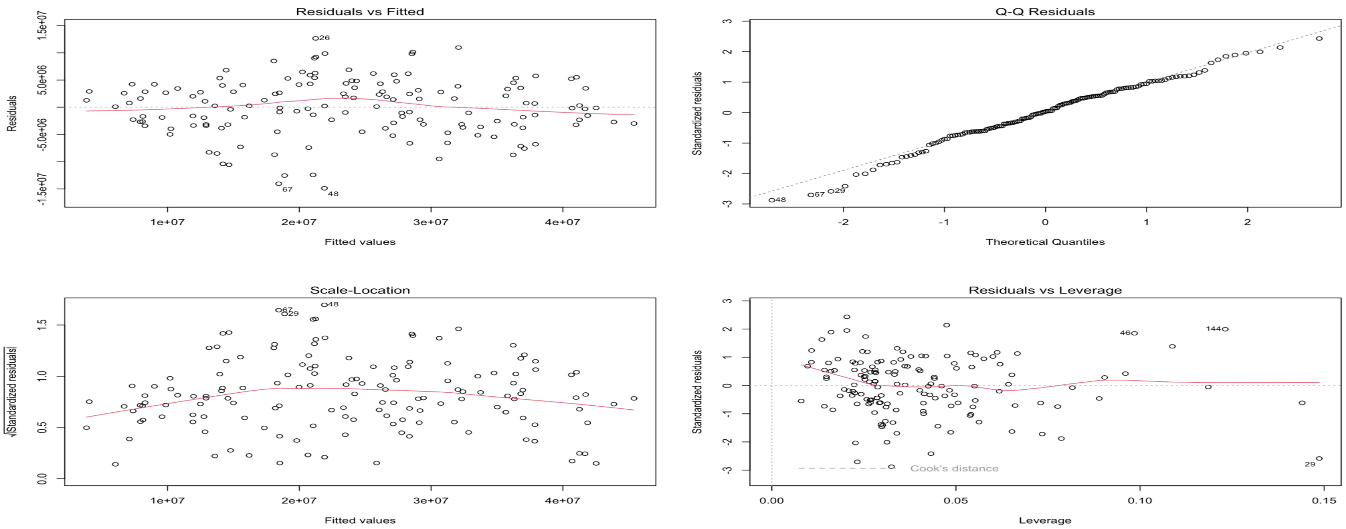


```
Call:
lm(formula = country1$LifeExpectancy^4 ~ country1$BirthRate +
    country1$GDP + country1$Internet + country1$Cell + country1$LandArea)

Residuals:
      Min        1Q    Median        3Q       Max
-14874191  -3146049    173057   3555050  12651371

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        29015670    8391063   3.458 0.000719 ***
country1$BirthRate -8363443    1683197  -4.969 1.91e-06 ***
country1$GDP        2275304     589332   3.861 0.000171 ***
country1$Internet   1197849     359440   3.333 0.001097 **
country1$Cell         11427      14176   0.806 0.421544
country1$LandArea   -439120     186850  -2.350 0.020144 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5254000 on 142 degrees of freedom
Multiple R-squared:  0.8019,    Adjusted R-squared:  0.7949
F-statistic: 114.9 on 5 and 142 DF,  p-value: < 2.2e-16
```



With Land Area and Birth Rate decreasing per change of unit of life expectancy, while GDP, Internet,and Cell all increasing per unit of change of life expectancy. Overall the model is an ok fit for life expectancy but there is a high probability that there are models that fit better, based on the residual plots of the final model and logical induction, such that, Land area and cell, should not be that much of a significant variable to predict life expectancy.