QF 112

LifeExpectancy Model

By: Drew Carranti, Noah Porcelain, Matthew Rutledge, Nicholas Suppiah, Alicia Zajac

April 21st, 2022

Initial Testing to Identify Important Variables

In our initial approach to the problem, we decided to use R's regsubsets() function to identify which significant variables would help predict. Since our final model would be able to use all the data provided for training, our initial testing using regsubsets() was conducted on all the data provided.

In running this function on the data, we noticed that using Year and Country significantly increased the run time of the function and complicated regsubsets' optimization of variable combinations, so when running regsubsets, we did not include those variables to allow for the function to run in a reasonable time. However, upon later analysis, we realized that adding country and year significantly increased the accuracy of our model, so the analysis from running regsubsets was made somewhat irrelevant by a change in the variables to choose from. However, we figured identifying essential variables in our original test would still be relevant as it still compares the variables to life expectancy.

In our initial conduction of regsubsets(), we found the best (lowest adjusted) number of variables (excluding Country and Year) was 8. Schooling, HIVAIDS, Adult Mortality, Income Composition of Resources, Percentage Expenditure, Infant Deaths, Under Five Deaths, and BMI were identified as the critical 8 variables that should be used to achieve the lowest adjusted R^2 . Also, considering regsubsets runs to test up to 8 variables in a combination, and that amount of variables was identified as the lowest possible adjusted R^2 , it is possible that the addition of other variables (Country & Year) could still improve the adjusted R^2 , even though we could not directly test that using regsubsets. After identifying these important variables, we continued to run regsubsets on various combinations of these variables while also adding in other less-significant variables to see if they could account for the remaining variability that the robust variables, earlier identified, could not account for. These substitution decisions were in large part made based on our correlation analyses.

````
```{r}
subsets = regsubsets(LifeExpectancy ~ LifeExpectancy + AdultMortality + infantDeaths + Alcohol +
percentageExpenditure + HepatitisB + Measles + BMI + underfiveDeaths + Polio + TotalExpenditure +
Diphtheria + HIVAIDS + GDP + Population + thinness119Years + thinness59Years +
IncomeCompositionOfResources + Schooling, data = data)
summary(subsets)
adjr2 = summary(subsets)$adj
best = which.max(adjr2)
best
```
````

```
1 subsets of each size up to 8
Selection Algorithm: exhaustive
         AdultMortality infantDeaths Alcohol percentageExpenditure HepatitisB Measles BMI
1  ( 1 ) " "            " "          " "     " "                   " "        " "     " "
2  ( 1 ) " "            " "          " "     " "                   " "        " "     " "
3  ( 1 ) "*"            " "          " "     " "                   " "        " "     " "
4  ( 1 ) "*"            " "          " "     " "                   " "        " "     " "
5  ( 1 ) "*"            " "          " "     "*"                   " "        " "     " "
6  ( 1 ) "*"            " "          " "     "*"                   " "        " "     "*"
7  ( 1 ) "*"            "*"          " "     "*"                   " "        " "     " "
8  ( 1 ) "*"            "*"          " "     "*"                   " "        " "     "*"
         underfiveDeaths Polio TotalExpenditure Diphtheria HIVAIDS GDP Population
1  ( 1 ) " "             " "   " "              " "        " "     " " " "
2  ( 1 ) " "             " "   " "              " "        "*"     " " " "
3  ( 1 ) " "             " "   " "              " "        "*"     " " " "
4  ( 1 ) " "             " "   " "              " "        "*"     " " " "
5  ( 1 ) " "             " "   " "              " "        "*"     " " " "
6  ( 1 ) " "             " "   " "              " "        "*"     " " " "
7  ( 1 ) "*"             " "   " "              " "        "*"     " " " "
8  ( 1 ) "*"             " "   " "              " "        "*"     " " " "
         thinness119Years thinness59Years IncomeCompositionOfResources Schooling
1  ( 1 ) " "              " "             " "                          "*"
2  ( 1 ) " "              " "             " "                          "*"
3  ( 1 ) " "              " "             " "                          "*"
4  ( 1 ) " "              " "             "*"                          "*"
5  ( 1 ) " "              " "             "*"                          "*"
6  ( 1 ) " "              " "             "*"                          "*"
7  ( 1 ) " "              " "             "*"                          "*"
8  ( 1 ) " "              " "             "*"                          "*"
[1] 8
```

Working with Training and Testing Data

In order to begin testing our linear regression models, we broke our data up into training and testing data to test the accuracy of our model on data it was not trained on. We began by setting a seed, which determined the random number selection of the samples from which our model made predictions.

We chose to train our data on approximately 900 randomly selected data points for this testing, which meant testing it on the remaining 100. We repeated this process with different seeds multiple times to ensure that the accuracy of our model based on our testing measurement wasn't just an artifact of the first random sample found by setting the seed once and leaving it. This also allowed us to build and test our model on various different subsets of the original data to ensure the model performed well with different train and test data situations.

Several factors contributed to our decision to break our data up into testing and training sections. When testing a larger testing dataset than previously described, such as 300, we ran into an issue when testing our model. As previously mentioned, we selected country and year to be a factor.

Upon testing with an increased testing set and, therefore, smaller training set, we noticed that, more commonly, the training data would not include data points for a particular country and, in rare cases, even specific years. This is why we later transitioned into using smaller testing sets to test our data.

Although we experienced this issue when dividing our data into training and testing sets, this error will not be an issue in the final project evaluation, as all of the data we have been working with will be the "training" data, and predictions will be tested on a new data set.

For variable adjustment testing, after the preliminary stage, we used the mean squared error calculation to test our model's predictive ability. We did this by training our model on our training set of 900 and then used the predict function to predict life expectancy values based on the input variables in our separate test set. We then used the mean squared error calculation to compare our model's predictions against the actual life expectancy values in our testing set to identify the error in our model's predictive power.

```
set.seed(235687)
index = sample(1:nrow(data), 100, replace=F)
train = data[-index,]
test = data[index,]
```

```
test_pred = predict(model, test)
test_RSS = sum((test_pred-test$LifeExpectancy)^2)
test_MSE = train_RSS / nrow(test)
test_MSE
```

Correlation Analysis

Highest Correlation >.50	
LifeExpectancy,Adult Mortality	-70%
LifeExpectancy, BMI	55%
LifeExpectancy,HIVAIDS	-59%
LifeExpectancy,IncomeCompositionOfResources	72%
LifeExpectancy, Schooling	73%
AdultMortality, HIVAIDS	56%
infantDeaths, Measles	54%
infantDeaths, underfiveDeaths	100%
infantDeaths, Population	68%
Alcohol, IncomeCompositionofResources	56%
Alcohol, Schooling	62%
percentageExpenditurem, GDP	96%
HepatitisB, Diphtheria	58%
Measles, underfiveDeaths	52%
BMI, thinness119Years	-54%
BMI, thinness59Years	-55%
BMI, IncomeCompositionOfResources	53%
BMI, Schooling	56%
underfiveDeaths, Population	67%
Polio, Diphtheria	61%
thinness119Years thinness59Years	92%
IncomeComposition of Resources, Schooling	79%

High Correlation .50< x >.40	
LifeExpectancy, Alcohol	40%
LifeExpectancy, percentageExpenditure	42%
LifeExpectancy, GDP	45%
LifeExpectancy, thinness119Years	-46%
LifeExpectancy, thinness59Years	-46%
AdultMortality, IncomeCompResources	-44%
AdultMortality, Schooling	-42%
infantDeaths, thinness119Years	47%
infantDeaths, thinness59Years	46%
Alcohol, percentageExpend	42%
Alcohol,GDP	46%
Alcohol, thinness119Years	-41%
percentExped, incomeCompResources	41%
percentExped, Schooling	43%
HepatitisB, Polio	48%
underfiveDeaths, thinness119Years	47%
underfiveDeaths, thinness59Years	47%
GDP, IncomeCompResources	46%
GDP, Schooling	48%
thinness119Years, IncomeCompResources	-46%
thinness119Years, Schooling	-49%
thinness59Years, IncomeCompResources	-45%
thinness59Years, Schooling	-48%

High Correlation .40< x >.30	
LifeExpectancy, Polio	34%
LifeExpectancy, Diphtheria	34%
AdultMortality, BMI	-36%
Alcohol, BMI	36%
Alcohol, thinness59Years	-39%
Measles, Population	31%
Polio, IncomeCompResource	30%
Polio, Schooling	34%
Diphetheria, IncomeCompResource	35%
Diphetheria, Schooling	35%

Medium Correlation .30< x >.20	
AdultMortality, percentageExpen	-24%
AdultMortality, Polio	-24%
AdultMortality, Diphtheria	-21%
AdultMortality, GDP	-26%
AdultMortality, 119Years	26%
AdultMortality, 59Years	27%
infantDeaths,HepB	-26%
infantDeaths,BMI	-23%
infantDeaths,Schooling	-21%
Alcohol, Polio	23%
Alcohol, TotalExpenditure	22%
Alcohol, Diptheria	23%
percentageExpenditure, BMI	25%
percentageExpenditure, Total Expend	20%
percentageExpenditure, 119Years	-26%
percentageExpenditure, 59Years	-26%
HepatitisB, underfiveDeaths	-26%
HepatitisB, Schooling	21%
BMI, underfiveDeaths	-24%
BMI, Polio	20%
BMI, TotalExpend	22%
BMI, HIVAIDS	-21%
BMI, GDP	27%
underfiveDeaths, Schooling	-22%
TotalExpenditure, GDP	21%
TotalExpenditure, 119Years	-22%
TotalExpenditure, 59Years	-23%
TotalExpenditure, Schooling	26%
HIVAIDS, IncomeCompResources	-25%
HIVAIDS, Schooling	-21%
GDP, 119Years	-29%
GDP, 59Years	-29%
Population, 119Years	26%
Population, 59Years	26%

Correlation by Variable

Life Expectancy	
LifeExpectancy, Schooling	73%
LifeExpectancy,IncomeCompositionOfResources	72%
LifeExpectancy,Adult Mortality	-70%
LifeExpectancy,HIVAIDS	-59%
LifeExpectancy, BMI	55%
LifeExpectancy, thinness59Years	-46%
LifeExpectancy, thinness119Years	-46%
LifeExpectancy, GDP	45%
LifeExpectancy, percentageExpenditure	42%
LifeExpectancy, Alcohol	40%
LifeExpectancy, Polio	34%
LifeExpectancy, Diphtheria	34%

Schooling	
IncomeComposition of Resources, Schooling	79%
Alcohol, Schooling	62%
BMI, Schooling	56%
thinness119Years, Schooling	-49%
thinness59Years, Schooling	-48%
GDP, Schooling	48%
percentExped, Schooling	43%
AdultMortality, Schooling	-42%
Diphetheria, Schooling	35%
Polio, Schooling	34%

IncomeCompositionOfResources	
IncomeComposition of Resources, Schooling	79%
Alcohol, IncomeCompositionofResources	56%
BMI, IncomeCompositionOfResources	53%
GDP, IncomeCompResources	46%
thinness119Years, IncomeCompResources	-46%
thinness59Years, IncomeCompResources	-45%
AdultMortality, IncomeCompResources	-44%
percentExped, incomeCompResources	41%
Diphetheria, IncomeCompResource	35%
Polio, IncomeCompResource	30%

AdultMortality	
AdultMortality, HIVAIDS	56%
AdultMortality, IncomeCompResources	-42%
AdultMortality, Schooling	-36%
AdultMortality, BMI	-36%

HIVAIDS	
AdultMortality, HIVAIDS	56%

BMI	
BMI, Schooling	56%
BMI, thinness59Years	-55%
BMI, thinness119Years	-54%
BMI, IncomeCompositionOfResources	53%
AdultMortality, BMI	-36%
Alcohol, BMI	36%

thinness59Years	
thinness119Years thinness59Years	92%
BMI, thinness59Years	-55%
thinness59Years, Schooling	-48%
underfiveDeaths, thinness59Years	47%
infantDeaths, thinness59Years	46%
thinness59Years, IncomeCompResources	-45%
Alcohol, thinness59Years	-39%

thinness119Years	
thinness119Years thinness59Years	92%
BMI, thinness119Years	-54%
thinness119Years, Schooling	-49%
underfiveDeaths, thinness119Years	47%
infantDeaths, thinness119Years	47%
thinness119Years, IncomeCompResources	-46%
Alcohol, thinness119Years	-41%

GDP	
percentageExpenditurem, GDP	96%
GDP, Schooling	48%
GDP, IncomeCompResources	46%
Alcohol,GDP	46%

percentageExpenditure	
percentageExpenditurem, GDP	96%
percentExped, Schooling	43%
Alcohol, percentageExpend	42%
percentExped, incomeCompResources	41%

Alcohol	
Alcohol, Schooling	62%
Alcohol, IncomeCompositionofResources	56%
Alcohol,GDP	46%
Alcohol, percentageExpend	42%
Alcohol, thinness119Years	-41%
Alcohol, thinness59Years	-39%
Alcohol, BMI	36%

Polio	
Polio, Diphtheria	61%
HepatitisB, Polio	48%
Polio, Schooling	34%
Polio, IncomeCompResource	30%

Diphtheria	
Polio, Diphtheria	61%
HepatitisB, Diphtheria	58%
Diphetheria, IncomeCompResource	35%
Diphetheria, Schooling	35%

We ran correlation analysis on the given variables in Excel to analyze the relationship between each variable and Life Expectancy to understand which variables would make up the core of our model.

We then ran the same correlation analysis between each variable to discern the relationship between our variables. We found this necessary because we not only wanted to create the most accurate model, but we wanted to build a model using the least amount of variables and to avoid multicollinearity due to highly correlated input variables.

Some of our data popped out to us and ultimately made its way into our final model. Standout variables included Schooling, Income Composition Of Resources, Adult Mortality, HIVAIDS, and BMI. All of which had correlations coefficients relative to Life Expectancy of greater than .50, and ultimately, our top 5 variables correlated to Life Expectancy ended up making our final model.

In summary, we used correlation analysis to make adjustment decisions for our model using the variables with high correlation to Life Expectancy but low correlation with the other variables used.

Testing Substitution Variables

After our initial analyses, we began basing our linear regression model on the most important variables found using both regsubsets and independent correlation.

Since we were unable to run regsubsets using all the variables provided, we broke our process down into two different elements: one including all variables except Country and Year, and another including all variables except Country.

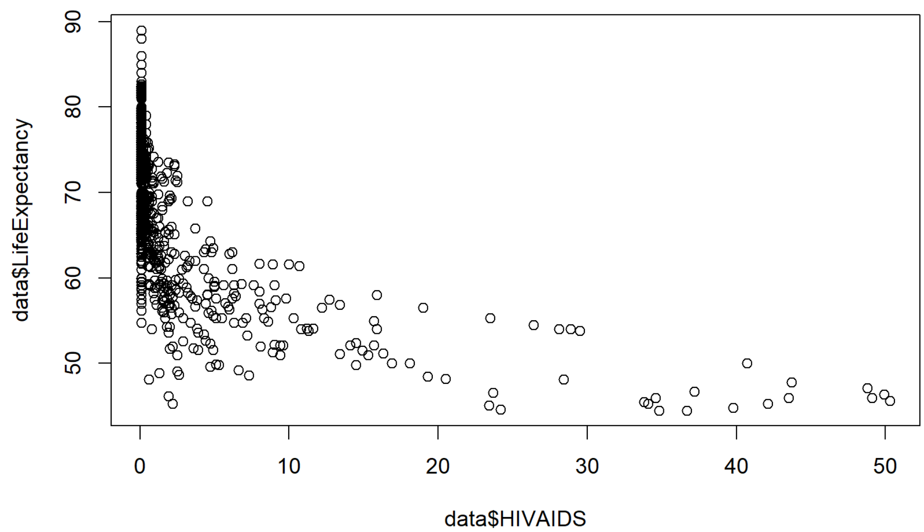
Our findings gave us a strong base of essential variables to start with, and from there, we started to take out and replace variables based on their correlations using the excel sheets provided above. For variables that were highly correlated to Life Expectancy that our previous analyses didn't find as significant input variables, we tested the effect of adding them. For variables already in place in our model that were highly correlated to one another, we tested the effect of their removal as we wanted to avoid multicollinearity.

Every time we changed the variables, we would test our mean squared error to see if it had improved our model's predictive power, and if it made it worse, we would disregard that addition or removal and revert back to our previous model.

Identifying Non-Linear Relationships Between Independent and Dependent

The plot() function in R enabled us to input two columns from the data frame and plot them against each other. The independent variable is plotted along the X-Axis and the dependent is plotted along the Y-Axis. The output of the plot() function is a scatter plot that we can use to determine the relationship between the two variables. Since we were building a model to predict Life Expectancy, we tested several variables as the independent variable while keeping Life Expectancy as the dependent variable.

Bellow is the output of : plot(data$HIVAIDS, data$LifeExpectancy)



After plotting several variables against Life Expectancy we found logarithmic relationships, exponential decay relationships and linear relationships. Since the linear model function lm() carries out regression using linear relationships, we were hoping that we could manipulate the input data of the independent variables so that they would be linearized, and therefore could be predicted with more accuracy using linear regression.

Unfortunately, while testing different methods (using various polynomials (square and cube), logarithmic functions, and exponential functions) these alterations to the input data did not improve the fit of our model and often times hurt it, so we were unable to actually employ any of our nonlinear relationship observations in our final model despite our efforts to achieve a more accurate fit.

Final Model Construction

Ultimately, our final regression model included the variables: "AdultMortality + infantDeaths + percentageExpenditure + BMI + underfiveDeaths + HIVAIDS + IncomeCompositionOfResources + Schooling + Country + Year".

This variable combination resulted in the smallest MSE both when trained and tested on all the data provided, as well as, when training and testing on various subsets from the original data. While we wanted to maintain a low amount of variables to avoid over-fitting, because the regsubset model suggested eight variables (excluding Country and Year) we figured our final model including eight variables, plus the additional two not included in the regsubsets test, would be acceptable for our model, especially considering they consistently resulted in a lower MSE's.

For our final function for submission, we ran our model on all the provided data and then saved it so that you may input the data we will be assessed on and it will output our predictions for said data. One important note: we loaded the Country and Year data in as factors so it is important that they are loaded in as factors prior to the running of the predicting function.