Ian Dang

Life Expectancy Report

The project is to take an 80% subset of the data set countries.csv to do the following:

1. Build a model with LifeExpectancy as the outcome and any of the remaining variable as predictors
2. Carry out a residual analysis to identify
    a. Deviations from linearity in any of the predictors
    b. Possible transformations of predictors
    c. Possible transformation of the outcome variable
3. Assess the potential for multicollinearity
4. Identify which variables are predictors of LifeExpectancy using suitable model selection algorithms.

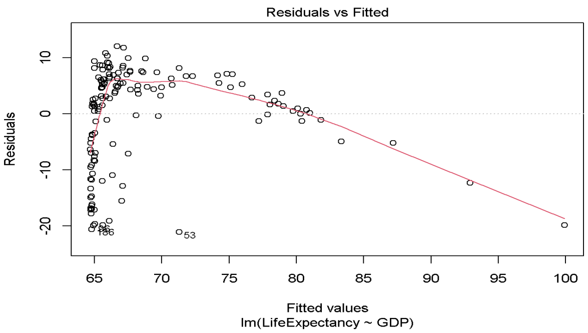
Result:

For 1 I chose GDP as the predictor and my result is :

```
Residuals:
    Min      1Q  Median      3Q     Max
-21.102  -4.992   2.972   6.286  12.010

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.460e+01  8.648e-01  74.708  < 2e-16 ***
GDP         3.349e-04  4.009e-05   8.354  4.7e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.615 on 146 degrees of freedom
Multiple R-squared:  0.3234,	Adjusted R-squared:  0.3188
F-statistic: 69.79 on 1 and 146 DF,  p-value: 4.702e-14
```

In this output, we can ignore the intercept as it might not be significant in this context. The coefficient for GDP represents the estimated change in Life Expectancy for a one-unit increase in GDP. In this case, a one-unit increase in GDP is associated with an increase of 0.0003349 in Life Expectancy. The p-value is very small (4.7e-14), indicating that GDP is a significant predictor of LifeExpectancy. Residual standard error is an estimate of the standard deviation of the errors, lower value indicates better fit. Multiple R-squared represents the proportion of variance in the dependent variable that is explained by the independent variable, which is LifeExpectancy and GDP; in this case, approximately 32.34% of the variability in LifeExpectancy is explained by GDP. The F-statistic tests the overall significance on the model, in this case the F-stat is 69.79 with a low p-value indicating the overall model is significant. Overall, the model suggests that there is a significant linear relationship between GDP and LifeExpectancy.

Ian Dang



I then check for linearity by plotting a residuals vs fitted model

The red line indicates the trend in the residuals and ideally, it should be horizontal at zero. In this case, the red line shows a clear downward trend which suggests that as the fitted values increase, the residuals tend to be more negative. In this context, it means that the model is systematically underpredicting the response variable (Life Expectancy) for higher values of the predictor (GDP). Overall, the plot indicates that this is not linear.

I then transformed the predictor because there is a non-linear relationship between the predictor and the outcome.

```
Residuals:
    Min      1Q  Median      3Q     Max 
-26.342  -2.124   1.328   4.245  13.046 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   22.698      3.190   7.116 4.62e-11 ***
log(GDP)       5.437      0.371  14.656  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.663 on 146 degrees of freedom
Multiple R-squared:  0.5954,	Adjusted R-squared:  0.5926 
F-statistic: 214.8 on 1 and 146 DF,  p-value: < 2.2e-16
```

The log(GDP) of 5.437 represents the estimated change in LifeExpectancy for a 1 unit increase in the log(GDP). In this case, a 1 unit increase in the log(GDP) means an increase of 5.437 in LifeExpectancy. The small p-calues means the transformation of GDP is statistically significant and the *** next to the p-value emphasize the high level of significance. The multiple R-squared in the context is approximately 59.54% of the variability in LifeExpectancy can be explained by log(GDP). Adjusted R-squared penalizes the inclusion of irrelevant predictors. The residual standard error is an estimate of the standard deviation of error which is 6.663 in this context. The F-stat is 214.8 with an extremely low p-value which indicates that the overall model is highly significant.

I also transformed the outcome variable using natural logarithm.

```
Residuals:
     Min       1Q   Median       3Q      Max 
-0.37076 -0.07097  0.05284  0.09963  0.17836 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 4.157e+00  1.415e-02 293.801  < 2e-16 ***
GDP         4.937e-06  6.559e-07   7.527 4.94e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1409 on 146 degrees of freedom
Multiple R-squared:  0.2795,	Adjusted R-squared:  0.2746 
F-statistic: 56.65 on 1 and 146 DF,  p-value: 4.942e-12
```

The intercept is the predicted value of the natural lof of LifeExpectancy when GDP is zero. Since the outcome variable is log-transformed, this values is interpreted as the natural of the LifeExpectancy at GDP equal to 0. The coefficient for GDP means that for a 1 unit increase in GDP there is an increase of 4.937e-06 in the natural log of LifeExpectancy. The p-value is really small which means that the transformation of GDP is statistically significant. Multiple R-squared in this case means that approximately 27.95% of the variability in the natural lof of LifeExpectacy is explained by GDP. Adjusted R-squared penalizes the inclusion of irrelevant predictors. Residual standard error estimates the standard deviation of the errors. The F-stat is 56.65 with a very low p-value which indicates that the model is highly significant.

Next, I check for multicollinearity and got the result:

```
                 LandArea    Population       Rural      Health    Internet
 LandArea      1.00000000  0.4363475398 -0.11682358 -0.02544283  0.02809784
Population     0.43634754  1.0000000000  0.09158706 -0.13508055 -0.07456124
Rural         -0.11682358  0.0915870565  1.00000000 -0.20919022 -0.67645508
Health        -0.02544283 -0.1350805505 -0.20919022  1.00000000  0.34369631
 Internet      0.02809784 -0.0745612384 -0.67645508  0.34369631  1.00000000
 BirthRate    -0.07798648 -0.0594990212  0.60678535 -0.21078599 -0.71735260
ElderlyPop     0.05608752 -0.0185184494 -0.55181192  0.33849505  0.74785302
LifeExpectancy  0.01370876 -0.0009462724 -0.62074962  0.32149592  0.72621637
CO2            0.14905457 -0.0365404226 -0.41239994  0.07994246  0.59773658
 GDP           0.02876334 -0.0771296782 -0.58731609  0.34011165  0.79790626
Cell           0.03460372 -0.0794335196 -0.58790667  0.10398609  0.59978372
                BirthRate   ElderlyPop LifeExpectancy         CO2         GDP
LandArea      -0.07798648  0.05608752   0.0137087583  0.14905457  0.02876334
Population    -0.05949902 -0.01851845  -0.0009462724 -0.03654042 -0.07712968
 Rural         0.60678535 -0.55181192  -0.6207496182 -0.41239994 -0.58731609
Health        -0.21078599  0.33849505   0.3214959166  0.07994246  0.34011165
Internet      -0.71735260  0.74785302   0.7262163706  0.59773658  0.79790626
 BirthRate     1.00000000 -0.76428797  -0.8510090005 -0.47953301 -0.51978039
ElderlyPop    -0.76428797  1.00000000   0.6471227778  0.33343227  0.60210835
 LifeExpectancy -0.85100900  0.64712278   1.0000000000  0.45614162  0.56868549
 CO2          -0.47953301  0.33343227   0.4561416230  1.00000000  0.56436765
 GDP          -0.51978039  0.60210835   0.5686854883  0.56436765  1.00000000
 Cell         -0.67793204  0.50417943   0.6575410865  0.46769781  0.45750380
                     Cell
 LandArea      0.03460372
Population    -0.07943352
 Rural        -0.58790667
Health         0.10398609
Internet       0.59978372
BirthRate     -0.67793204
ElderlyPop     0.50417943
LifeExpectancy  0.65754109
CO2            0.46769781
GDP            0.45750380
Cell           1.00000000
```

This output is a correlation matrix that shows the correlation coefficients between different variables. The values in the matrix range from -1 to 1. A value of 1 indicates a perfect positive linear relation, a value of -1 indicates a perfect negative linear relationship, and a value of 0 indicates no linear relationship. For example, LandArea has a moderate positive correlation with population(0.44).

Next, I identified which variables are predictors of LifeExpectancy using a suitable model selection
algorithm. I chose stepwise regression as the model method and the final model in the output shows that Health, Internet, Birthrate, Ederlypop, and Cell are significant predictors of LifeExpectancy.

```
Step:  AIC=481.34
LifeExpectancy ~ Health + Internet + BirthRate + ElderlyPop +
    Cell

             Df Sum of Sq    RSS    AIC
<none>                    3527.8 481.34
+ Rural       1     38.58 3489.2 481.71
+ LandArea    1     24.72 3503.1 482.30
+ CO2         1     22.59 3505.2 482.39
+ GDP         1     20.11 3507.7 482.49
- Cell        1     90.25 3618.0 483.08
+ Population  1      0.11 3527.7 483.33
- ElderlyPop  1    154.10 3681.9 485.67
- Health      1    260.58 3788.4 489.89
- Internet    1    294.77 3822.5 491.22
- BirthRate   1   2393.23 5921.0 555.98

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 78.46295    3.08989  25.393  < 2e-16 ***
Health       0.32952    0.10175   3.239 0.001495 **
Internet     0.09036    0.02623   3.445 0.000752 ***
BirthRate   -0.70993    0.07233  -9.815  < 2e-16 ***
ElderlyPop  -0.35753    0.14355  -2.491 0.013906 *
Cell         0.02557    0.01342   1.906 0.058673 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.984 on 142 degrees of freedom
Multiple R-squared:  0.7797,	Adjusted R-squared:  0.772
F-statistic: 100.5 on 5 and 142 DF,  p-value: < 2.2e-16
```

Using stepwise, the goal is to find the best subset of the predictor variables that minimzes the AIC. The AIC with the variables Health, Internet, Birthrate, Ederlypop, and Cell has the lowest AIC of 481.34. Using the model coefficient we can interpret the predictors and the out come in the context. For example, the coefficient for Health is 0.32952 which suggests that on average, for a 1 unit increase in the percentage of government expenditures directed towards healthcare, the life expectancy is estimated to increase by 0.32952 years. In summary, the final model's variables suggest that they are significant predictors of LifeExpectancy in this context. Furthermore, the adjusted R-squared of 0.772 indicates a good fit of the model to the data.

Model Building:

Ian Dang

I chose stepwise regression as my model because it provides an automated and systematic way to select variables based on statistical criteria which is the AIC in this case. Stepwise regression also tends to generate simpler models by selecting a subset of predictors. Furthermore, the simpler models generated by stepwise regression are easier to interpret and may generalize better to new data. Also, the AIC criterion balances the fit with the complexity of the model, helping to avoid overlifting.

Summary:

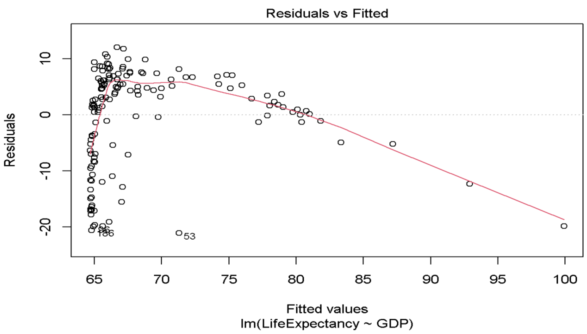
In summary, I created a model with LifeExpectancy as the outcome and GDP as the predictors. I then checked for linearity and there were none so I carried out steps to check for possible transformation of predictors and outcomes. Next, I checked for multicollinearity which shows the pairwise correlation coefficients between the different variables. Strong correlations can indicate potential dependencies between variables. Finally, I wrote an algorithm using stepwise regression modelto identify which variables are predictors of LifeExpectancy and it is Health, Internet, Birthrate, Ederlypop, and Cell which are significant predictors of LifeExpectancy.

Appendix:

```
countries = read.csv('C:\\Users\\Benny\\Downloads\\countries.csv')
n <- nrow(countries)
set.seed(108)
subset_id = sample(n, 0.8*n)
countries_subset = countries[subset_id, ]

head(countries_subset)

model <- lm(LifeExpectancy ~ GDP, data = countries_subset)

summary(model)


#check for linearity

plot(model)
```

Ian Dang

```
#possible transformation of predictor

model_transformed <- lm(LifeExpectancy ~ log(GDP), data = countries_subset)

summary(model_transformed)

#Possible transformation of the outcome variable
model_transformed_outcome <- lm(log(LifeExpectancy) ~ GDP, data = countries_subset)

summary(model_transformed_outcome)

#Assess the potential for multicollinearity

variables <- c('LandArea', 'Population', 'Rural', 'Health', 'Internet', 'BirthRate', 'ElderlyPop',
'LifeExpectancy', 'CO2', 'GDP', 'Cell')

cor_matrix <- cor(countries_subset[, variables])
print(cor_matrix)


library(MASS)

simple_model <- lm(LifeExpectancy ~ LandArea + Population + Rural + Health + Internet +
BirthRate + ElderlyPop + CO2 + GDP + Cell, data = countries_subset)


# Stepwise Regression for model selection
stepwise_model <- stepAIC(simple_model, direction = "both")

# Summary of stepwise model
summary(stepwise_model)
```


Residuals vs Fitted
Residuals
Fitted values
lm(LifeExpectancy ~ GDP)