Darian Lee
STA 108
12/8/24

EXECUTIVE SUMMARY:

The goal of this analysis was to create a well fitting linear regression model for predicting life expectancy using various data based on real world countries. The final model contains 6 predictors; namely land area, health, internet, birth rate, cell, and gdp. This model was built by exploring the individual predictors, building a preliminary model using criterion based model selection, performing necessary transformations of the predictors and response variable, and lastly, using a criterion based approach again to finalize the model by removing and adding certain variables to improve fit. It was concluded that the final model has good fit with R^2 of .8326 and follows the assumptions of linear regression, however the moderate multicollinearity present in GDP may need to be reassessed depending on the specific application.

INTRODUCTION:

This project was motivated by an interest in what factors have the greatest influence on the average life expectancy of a country, and the objective was to produce a well fitting regression to model this relationship using data collected from a random sample of 145 countries. It was found that the top 3 factors most predictive of average life expectancy are the GDP, birth rate, and health*. As birth rate increases, average life expectancy in the country tends to decrease, whereas as health and/or GDP increases, life expectancy is believed to increase. Other less significant predictors included in the model were internet, cell, and land area. Internet and cell both show possessive correlation with life expectancy, meaning that people living in areas with high internet and cell coverage tend to live longer, whereas land area was slightly negatively correlated, meaning that larger countries tend to have shorter life expectancies. It is worth noting that the actual model uses the natural log of GDP and life expectancy raised to the 4th power in order to be more aligned with the assumptions of linear regression, but this detail will not affect the interpretation of the model mentioned above.

METHODS

After initial data observation of each of the predictor variables did not hint towards any errors in data collection, the package regsubsets was used to determine the best predictors to include in the preliminary model. The function chose the best model for every possible number of predictors based on mallows CP, and outputted those ten models. Of those 10 models, the models with the top 3 highest R^2 values, adjusted R^2 values and lowest mallows CP values were obtained, and after holistically examining each, the one that performed the best comparatively across all three categories was selected.

This model included health, internet, birth rate, elderly population, CO2, GDP, and cell. Thus this model differed from the final model in the inclusion of elderly population and CO2 as well as the exclusion of land area. This initial model had an R^2 of 0.8034, Adjusted R^2 of 0.7936, and mallows CP of 6.987.

While this mallows CP value was impressive since it was very close to the number of predictors, the model showed

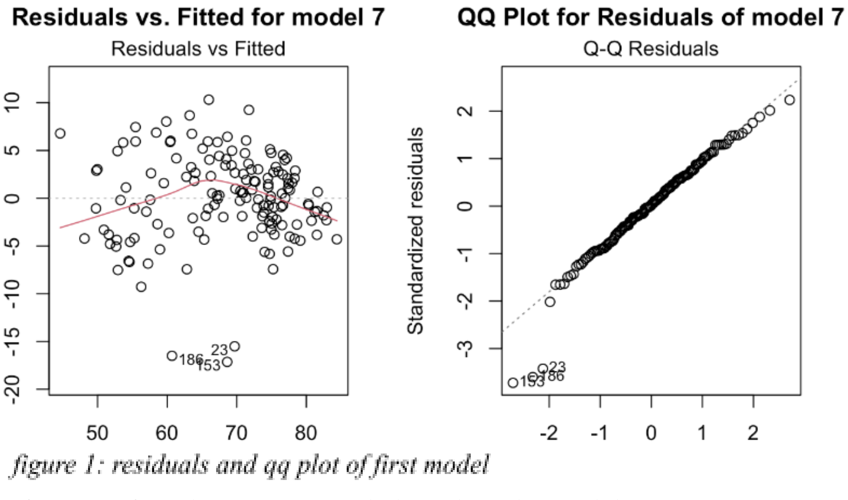


figure 1: residuals and qq plot of first model

*Note: Statements regarding the relative significance of predictors are made based on the model coefficients, which may be inflated or deflated in part by slight differences in the scaling of each variable notable deviations from linearity and a few residuals that were non normal and had non constant error with the rest of the model (figure 1). There were three points in particular that were outliers on both plots; That is, Botswana, South Africa, and Zimbabwe, labeled 23, 153, 186 respectively. These points were analyzed across every variable, however no evidence was found of measurement error or these points being an outlier in any predictor, and thus they were left in the model. For more information on this analysis, refer to scatter_plots_with_highlights.pdf and the Rcode in the appendix.

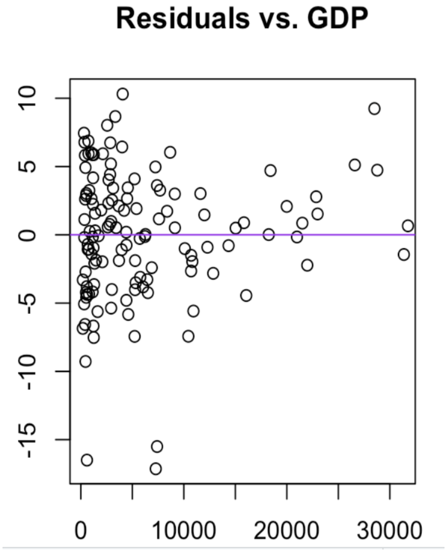


figure 2: residuals plotted against GDP

In order to address the non linearity, the residuals were plotted against each of the predictors. It was noted that the GDP residual plot showed a slight u shape (figure 2), so this variable was transformed using the natural log. This transformation improved the linearity of the residuals in the full model (figure 3: left), but it made the normality of errors slightly worse (figure 3: right)

To improve the normality of errors as well as address issues with non constant error due to the outliers, a box cox analysis was performed on life expectancy, and it was determined that the best transformation of y would be to square it. While this improved normality and constant error slightly by drawing the outliers closer to the rest of the points on both the qq plot and the residual plot (figure 4: previous model) (figure 5: previous model), it did not completely fix the problem, so the boxcox transformation was used again on the new values of life expectancy squared. In the second boxcox analysis, it was determined that the optimum transformation was to square y again, causing life expectancy to be raised to the 4th power. This slightly improved the normality of the errors (figure 4: new model) and got rid of one of the outliers in terms of normality and constant residuals, but the main reason why this transformation was kept in the final model is because linearity (figure 5: new model) and R² significantly improved as a beneficial side effect. Each of the individual predictors were plotted against

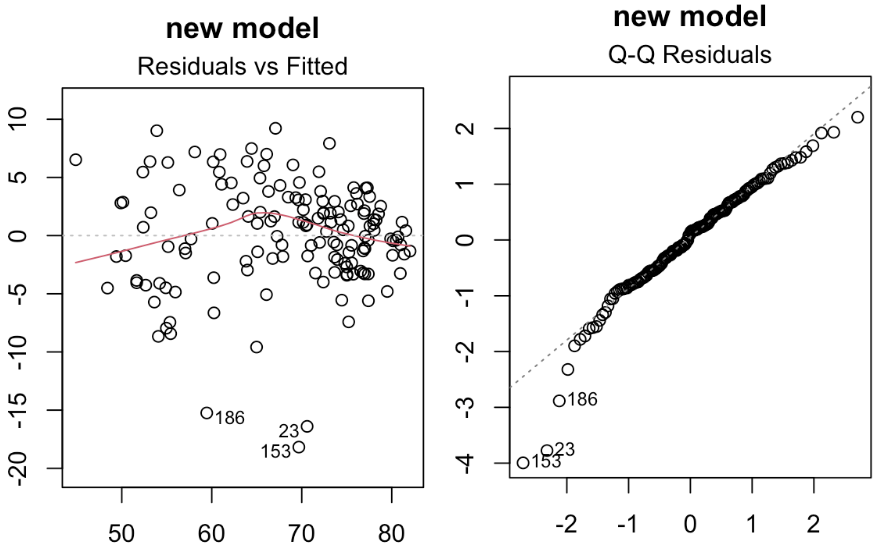


figure 3: residual and qq plot after transforming GDP

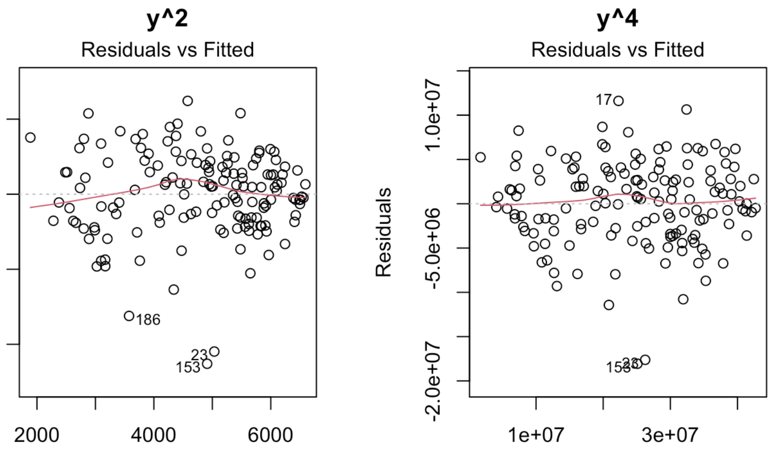


figure 4: residuals of y^2 (left) vs y^4 (right)

the residuals again to see if there were any new violations of linearity, and none were found.

The final step of the model selection process involved running regsubsets again to determine which predictors should be removed and added in order to maximize R^2 and adjusted R^2, and minimize mallows CP.

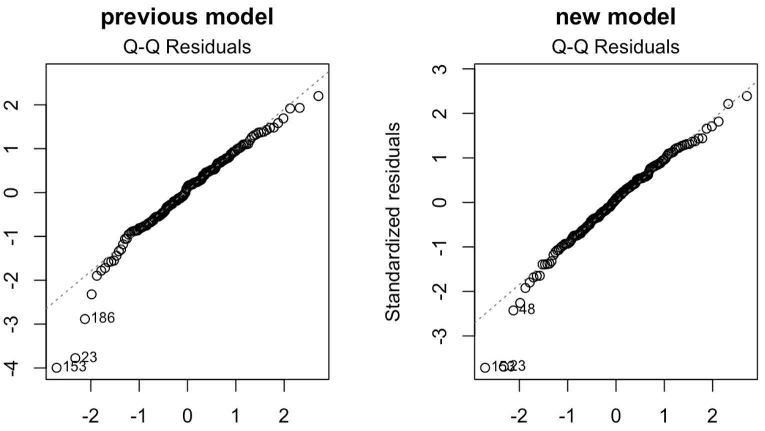


figure 5: qq plot of y^2 (left) vs y^4 (right)

It was found that removing CO2 and elderly population and adding land area resulted in a model that was maximized across all three parameters, with an R^2 of .832569, mallows cp of 5.418374, and adjusted R^2 of 0.8254443 (figure 7).

RESULTS:

The final model uses land area, health, internet, birth rate, cell, and the log of GDP to predict life expectancy, which has been raised to the 4th power as explained in methods. The residuals have constant error and are centered around 0, with 2 outliers at 153 and 23 (Botswana and South Africa respectively). The errors also exhibit normality as shown on the qq plot, with the only exception being the 2 outliers as mentioned earlier. (Note: earlier models had 3 outliers, but this number was reduced to 2).

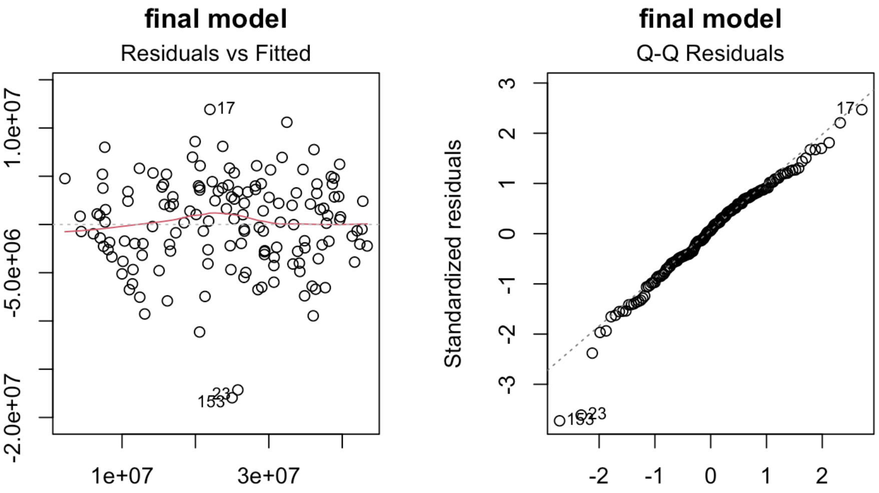


figure 6: final model residual and qq plot

The most significant variable in my final model is the log of GDP, which means that the log of GDP is highly predictive of life expectancy*. It is expected that when all the other predictors are held constant, a one unit change in the log of GDP would cause an increase in the value of life expectancy raised to the 4th power of approximately 2.64e+6. The other predictors rank as follows from most to least important: birth rate (-4.24e-5), health (2.06e+5), internet (6.46e+4), cell (2.37e+4), and land area (-4.2e+1)*. Of these predictors, GDP, health, internet, and cell have a positive influence on the predicted value of life expectancy, and birth rate and land area have a negative influence.

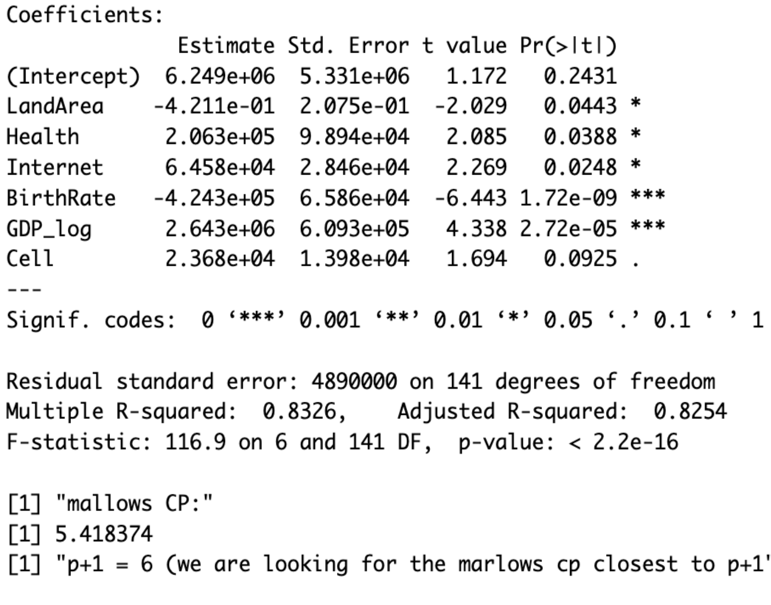
All of the beta parameter estimates for the predictor variables are statistically significant at a 90% confidence level, and all of the beta parameter estimates excluding that for Cell are statistically significant at a 95% confidence level (figure 7). It is important to note that the beta estimates shown in figure 7 are

**Note: Statements regarding the relative significance of predictors are made based on the model coefficients, which may be inflated or deflated in part by slight differences in the scaling of each variable* very large because the response variable has been raised to the 4th power. It should be understood that while this model is useful in seeing how this combination of variables correlates with life expectancy, they cannot be considered to have any causational effect on life expectancy based on this model. Additionally, this model may be inaccurate in estimating values that are outside the range of the data used in building the model.

The high R² value and adjusted R² value of .8326 and .8254 (figure 7) indicate that the model explains a significant amount of variation in the response variable, and the low mallows cp value of 5.418 (figure 7) indicates that our model has a very good fit considering its level of simplicity.

DISCUSSION:

Although the final model fits the data well, the fit is inhibited by the presence of two outliers on the residual plot, namely Botswana and South Africa. The scatter plots indicated that these are countries where medium levels of birthrate and cell coverage correspond to lower than expected life expectancies, thus perhaps our model is limited in accurately predicting the life expectancy of countries with this characteristic.
The deviation in constant error shown in the lower left hand side of the residual plot could also be a potential concern (figure 6 left).



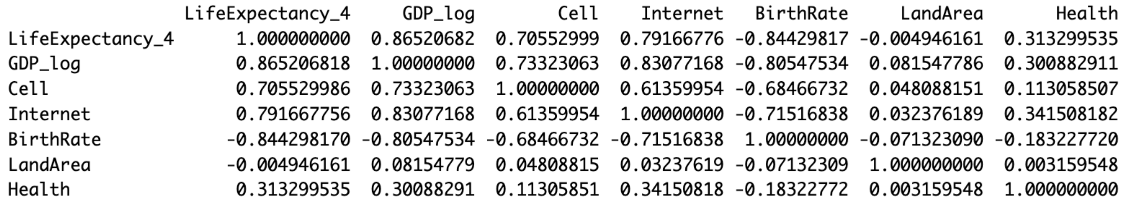
```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  6.249e+06  5.331e+06   1.172   0.2431    
LandArea    -4.211e-01  2.075e-01  -2.029   0.0443 *  
Health       2.063e+05  9.894e+04   2.085   0.0388 *  
Internet     6.458e+04  2.846e+04   2.269   0.0248 *  
BirthRate   -4.243e+05  6.586e+04  -6.443 1.72e-09 ***
GDP_log      2.643e+06  6.093e+05   4.338 2.72e-05 ***
Cell         2.368e+04  1.398e+04   1.694   0.0925 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4890000 on 141 degrees of freedom
Multiple R-squared:  0.8326,	Adjusted R-squared:  0.8254 
F-statistic: 116.9 on 6 and 141 DF,  p-value: < 2.2e-16

[1] "mallows CP:"
[1] 5.418374
[1] "p+1 = 6 (we are looking for the marlows cp closest to p+1"
```

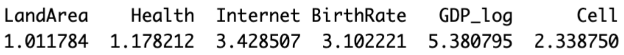
figure 7: final model measures of fit

Another limitation of my model is possible concerns with multicollinearity. The log of gdp, cell, and internet, were all shown to correlate with each other in the correlation matrix (figure 8).
When the multicollinearity’s effect on the variance of a regression coefficient was measured using the VIF function, all the values were under 4 except for GDP, which had a value of 5.38 (figure 9). Typically values above 4 are considered moderate, so it should be noted that in situations where it is very important to have accurate predictions for the relationship between each individual predictor and life expectancy, this model may need to be re-evaluated due to potential issues with multicollinearity. However, since removing one of the correlated variables caused the model to violate the assumptions of linear regression and decreased the fit of the model, it was decided that the moderate multicollinearity present in the current model was acceptable for the purposes of this assignment.



```
                 LifeExpectancy_4     GDP_log        Cell    Internet   BirthRate     LandArea       Health
LifeExpectancy_4      1.000000000  0.86520682  0.70552999  0.79166776 -0.84429817 -0.004946161  0.313299535
GDP_log               0.865206818  1.00000000  0.73323063  0.83077168 -0.80547534  0.081547786  0.300882911
Cell                  0.705529986  0.73323063  1.00000000  0.61359954 -0.68466732  0.048088151  0.113058507
Internet              0.791667756  0.83077168  0.61359954  1.00000000 -0.71516838  0.032376189  0.341508182
BirthRate            -0.844298170 -0.80547534 -0.68466732 -0.71516838  1.00000000 -0.071323090 -0.183227720
LandArea             -0.004946161  0.08154779  0.04808815  0.03237619 -0.07132309  1.000000000  0.003159548
Health                0.313299535  0.30088291  0.11305851  0.34150818 -0.18322772  0.003159548  1.000000000
```

figure 8: correlation matrix



```
LandArea    Health  Internet BirthRate   GDP_log      Cell 
1.011784  1.178212  3.428507  3.102221  5.380795  2.338750 
```

figure 9: VIF values