Statistical Analysis of the Gapminder Dataset

Ryan Jackson

21/10/2021

1 Introduction

As would be expected, the average life expectancy of humans around the world can vary wildly depending on where they are situated. For example, some continents are much more developed than others and as such have greater access to higher quality medicines and medical facilities meaning things such as infections & diseases may have a harder time spreading leading to a potentially higher life expectancy. For this analysis, we examine a subset of the Gapminder dataset which is readily available through the gapminder library in R and it contains data on Life Expectancy, Population and GDP per Capita for each country from 1952-2007 in 5 year intervals however for this analysis, we examine the years 1952, 1982 & 2007. The aim for our analysis is to assess how the life expectancy for each continent changes over time and what effect population, GDP per Capita and time may have had on the life expectancy. Section 2 explores the data in detail with some visualisations followed by Section 3 which contains the statistical model and our assumptions, lastly Section 4 has our concluding remarks on the data and our model.

2 Exploratory Analysis

The Gapminder dataset contains 1704 rows of 6 variables namely, Country, Continent, Year, Life Expectancy, Population and GDP/Cap. After taking our subset of the years 1952, 1982 and 2007 we have 426 rows left over.

2.1 Data Cleaning

One of the first steps in our analysis is to esure that we take care of any missing values in our data should they exist. Table 1 below shows the count of the missing values in each column of our dataset and as we can see there are no missing values to deal with.

Table 1: Count of Missing Values for each Variable.

	Country	Continent	Year	LifeExp	Population	GDP.Cap
No. of Missing Values	0	0	0	0	0	0

Before moving onto the visualisations in Section 2.2 below, we need to make a slight adjustment to some of the values in the Continent column in the dataset as currently the values we have are Africa, Americas, Asia, Europe, Oceania but instead we'd like to split the Americas into North and South America depending on the country.

Table 2 below shows the count of countries in each continent and demonstrates how the Americas continent has been split into North and South America.

Table 2: Count of Countries in each Continent.

Continent	Frequency
Africa	156
Americas	75
Asia	99
Europe	90
Oceania	6

Continent	Frequency
Africa	156
Asia	99
Europe	90
North America	42
Oceania	6
South America	33

2.2 Data Visualisation

Now its time to explore the data visually in greater detail and we begin by a very simple comparison of the life expectancy for each continent from 1952-2007,

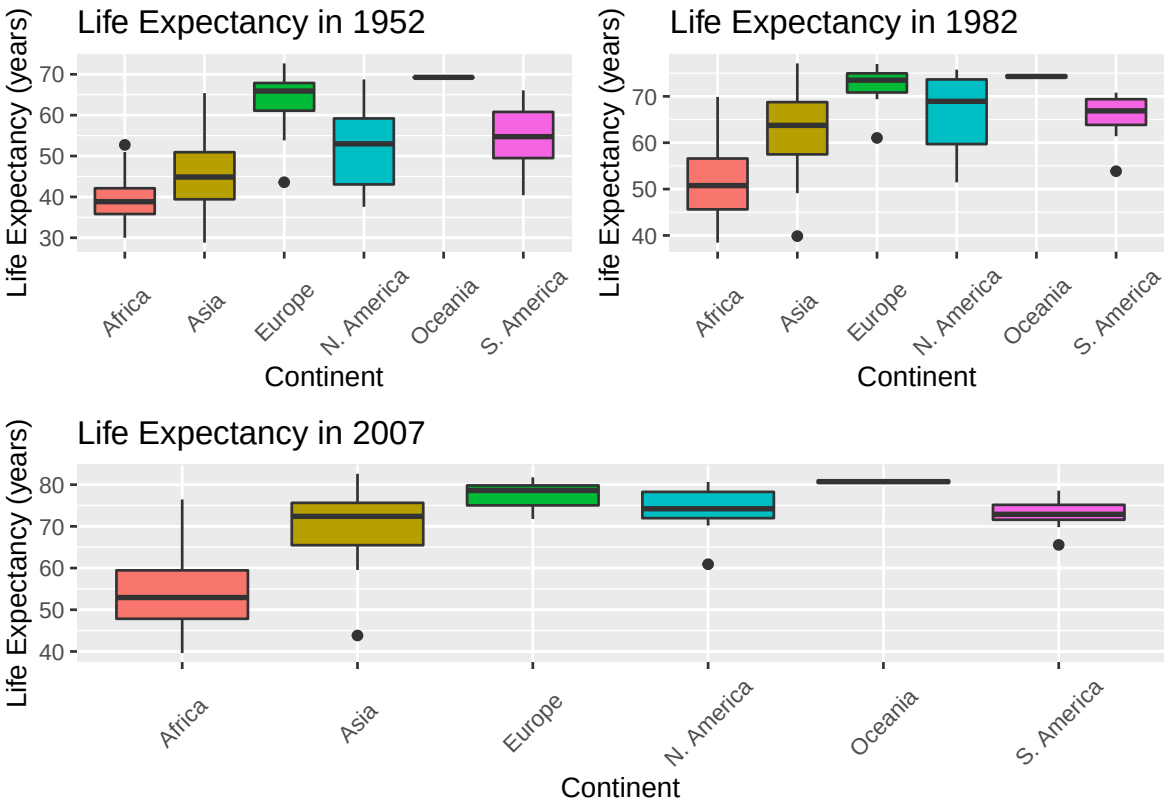

Figure 1: Change in Life Expectancy in each Continent in 1952(Top Left), 1982(Top Right) and 2007(Bottom).

In Figure 1 we see boxplots of the life expectancy for each of the continents in the years 1952, 1982 and 2007. Looking at 1952 on the top left we see that overall Africa appears to have the lowest median life expectancy followed by Asia whereas on the opposite end of the scale we have a much greater median life expectancy for both Europe and Oceania. An interesting observation is the spread of both Asia and Oceania with Asia having a significantly large spread whereas Oceania has little to no spread at all, however a likely explanation is that Oceania only contains 2 countries - Australia and New Zealand whereas Asia contains 33 countries.

Moving to the next plot along of 1982, we see a rise across all continents in life expectancy which would be expected due to technological advancements and advancements in medical knowledge & treatments however it does appear that Africa is lagging behind the other continents by a bigger margin compared to 1952 with Asia and North America in particular taking huge leaps. We also appear to see more outliers compared to 1952 suggesting some countries aren't keeping pace with the continents as a whole however, there doesn't appear to be any continent that is heavily skewed. Lastly we come to 2007 on the bottom of Figure 1 and this is really where the huge gulf in life expectancy begin to show with the middle 50% of countries in Africa lagging significantly behind the rest of the world having seen no substantial increase in life expectancy from 1982. The other continents appear to all be roughly similar to one another with Asia ever so slightly behind and Oceania ever so slightly ahead.

Next we can look to how population sizes for each of the continents has changed across time with Figure 2 below showing boxplots of the log population changes for each continent from 1952 in the top left to 2007 on the bottom

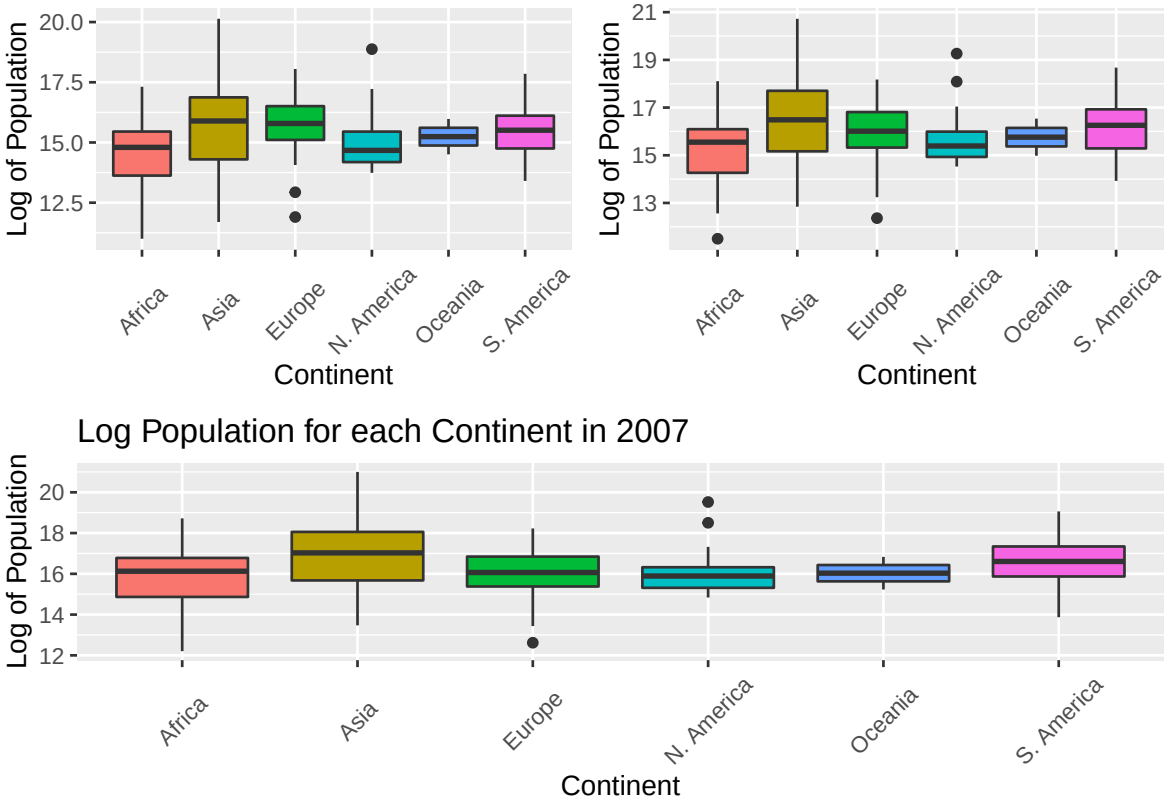

Figure 2: Change in Population size in each Continent in 1952(Top Left), 1982(Top Right) and 2007(Bottom).

From Figure 2 above we can see that in 1952, the median population of each continent appears to be roughly similar with Asia and Europe sitting slightly higher than South America and Oceania followed by Africa and then suprisingly North America. The spread doesn't appear to be too large for Europe, North & South America and Oceania however, for Africa and Asia in particular we see pretty substantial standard deviations. As is well documented around the world today, the Earth's population is ever increasing and we see this demonstrated in both 1982 and 2007 where in 1982 we see median increases across the board with the standard deviations in Africa and Asia appearing to decrease slightly and no real increase in that of the other continents. Similarly in 2007, we see slight rises again with the biggest median increase coming from Asia and it does appear that the spread for both Africa and Asia is still significantly greater than that of

any other continent.

The final visualisations we'll take a look at for this section are shown below in Figure 3 where we have boxplots of the log GDP per Capita of each continent from 1952-2007. We use the log of the GDP per Capita here similarly to the log of the population in Figure 2 above due to how skewed the plots are without it.

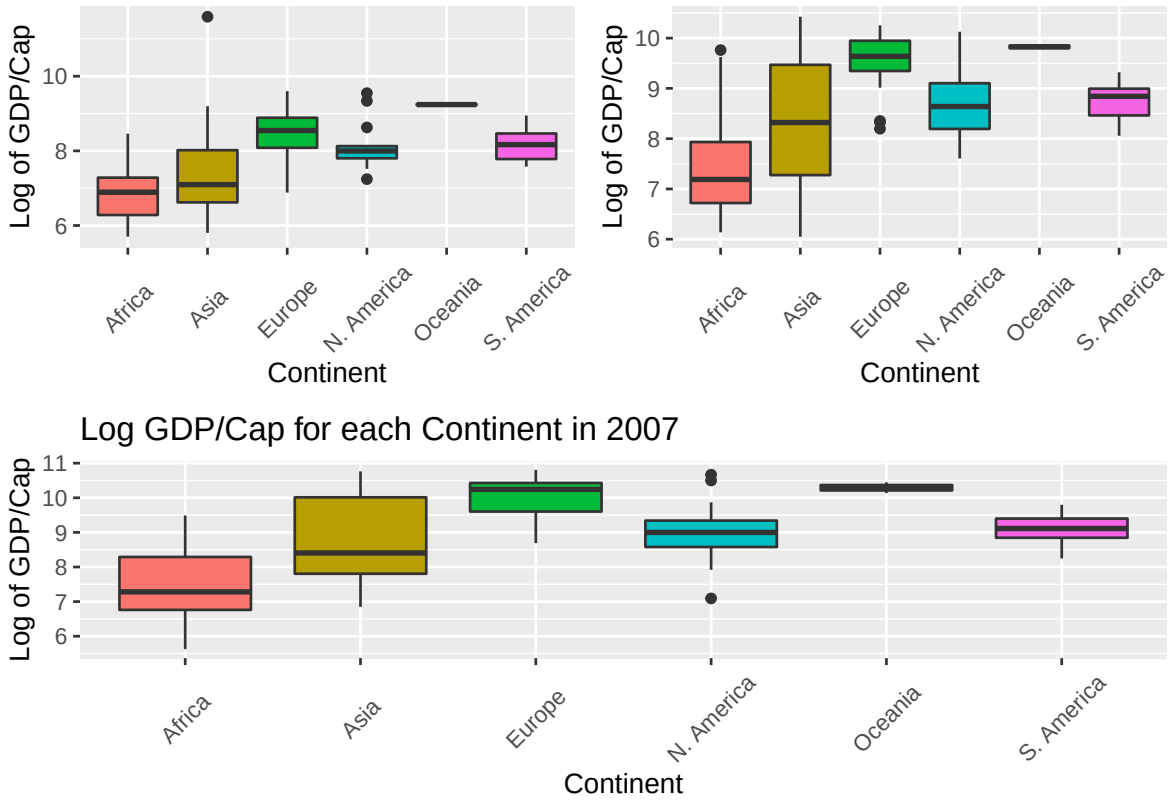

Figure 3: Change in GDP/Cap in each Continent in 1952(Top Left), 1982(Top Right) and 2007(Bottom).

From looking at the year 1952 we see that the continents with the lowest median GDP per Capita appear to be Africa and Asia followed by North America and South America with Oceania being the highest followed by Europe. We also do appear to see a number of outliers from Asia and North America which is what we would expect. From the plot of 1982, we see that more variability in the data begins to appear particularly in Asia as technological advancements occur and some countries are left behind the others. We see in 1982 that Europe and Oceania appear to have the highest GDP per Capita with Asia now on par with North and South America but Africa hasn't really made any significant progress from 1952. Lastly, examining the plot in 2007, we see that again there appears to have been no real progress in GDP per Capita growth for Africa and we also see that boxplot for Asia is becoming slightly more skewed and this is likely due to countries like China beginning to rapidly expand economically. Europe and Oceania again appear to be pretty level still as are North and South America.

3 Statistical Model

In this section we will fit a multiple linear regression model to the data in order to assess whether continent, population and GDP per Capita are good predictors of life expectancy in the years 1952, 1982 and 2007.

We begin by considering the full model for each year which has Continent, Population and GDP per Capita as our explanatory variables.

In order to find the optimal model for the data we use stepwise regression which tells us what variables to include in the final model. We use this procedure to compare models using the Akaike Information Criterion (AIC) with both forward selection and backwards elimination and our final model will be the one which produces the lowest AIC score.

After running the procedure, the final model for each of the years included the variables Continent and GDP per Capita as our final significant predcitors meaning our model is given by:

$$y_i = \alpha + \beta_{Asia} \cdot \mathbb{I}_{\text{Asia}}(x) + \beta_{Eur} \cdot \mathbb{I}_{\text{Eur}}(x) + \beta_{N.Am} \cdot \mathbb{I}_{\text{N.Am}}(x) + \beta_{Ocn} \cdot \mathbb{I}_{\text{Ocn}}(x) + \beta_{S.Am} \cdot \mathbb{I}_{\text{S.Am}}(x) + \beta_{GDP/Cap} \cdot x_{1i} + \epsilon_i$$

where

- α is the mean life expectancy for baseline continent Africa;
- $\beta_{\text{Continent}}$ is the difference in mean life expectancy of given continent relative to baseline continent Africa;
- $\beta_{GDP/Cap}$ is the term added for the GDP/Cap;
- $\epsilon_i \sim N(0, \sigma^2)$ is the error term; and
- $\mathbb{I}_{\text{Continent}}(x)$ is an indicator function such that:

$$\mathbb{I}_{\text{Continent}}(x) = \begin{cases} 1 & \text{if country is in continent,} \\ 0 & \text{Otherwise.} \end{cases}$$

Thus, from our model above we obtain our regression equations of:

$$\widehat{y}_{1952} = 38.9 + 6.46 \cdot \mathbb{I}_{\text{Asia}}(x) + 24.47 \cdot \mathbb{I}_{\text{Eur}}(x) + 13 \cdot \mathbb{I}_{\text{N.Am}}(x) + 28.46 \cdot \mathbb{I}_{\text{Ocn}}(x) + 14.42 \cdot \mathbb{I}_{\text{S.Am}}(x) \quad (1)$$

$$\widehat{y}_{1982} = 50.14 + 8.12 \cdot \mathbb{I}_{\text{Asia}}(x) + 13.52 \cdot \mathbb{I}_{\text{Eur}}(x) + 11.56 \cdot \mathbb{I}_{\text{N.Am}}(x) + 13.28 \cdot \mathbb{I}_{\text{Ocn}}(x) + 11.86 \cdot \mathbb{I}_{\text{S.Am}}(x) \quad (2)$$

$$\widehat{y}_{2007} = 53.74 + 12.67 \cdot \mathbb{I}_{\text{Asia}}(x) + 15.23 \cdot \mathbb{I}_{\text{Eur}}(x) + 16.14 \cdot \mathbb{I}_{\text{N.Am}}(x) + 16.66 \cdot \mathbb{I}_{\text{Ocn}}(x) + 15.96 \cdot \mathbb{I}_{\text{S.Am}}(x) \quad (3)$$

So taking our equations above, if we were to take someone from Europe in 1952, 1982 and 2007 then their average life expectancy would be 63.4 years, 63.7 years and 69 years respectively. It is important to note here that the term for GDP per Capita isn't included in the regression equations above because the model estimate for that term in each of the years was given as very close to 0 and as such had no real impact on the equations.

3.1 Assessing Model Fit/ Assumptions

We can now move on to determining whether our model is an appropriate fit for the data by checking the 5 main assumptions that need to hold true. These are that the residuals have mean 0, are normally distributed and are independent, we also require that the scale of the variability of the residuals is constant at all values of the explanatory variables and that the values of the explanatory variables are recorded without error. Note that we are unable to assess the assumptions of the residuals being independent and that the values of the explanatory variables are recorded without error so assume these to be true.

We first plot the residuals against the explanatory variable GDP per Capita by Continent in Figure 4 below to assess the first assumption of the residuals having mean 0. From the plots we can see that the residuals

do appear to be randomly scattered above and below the zero line for each of the years and as such can say this assumption holds.

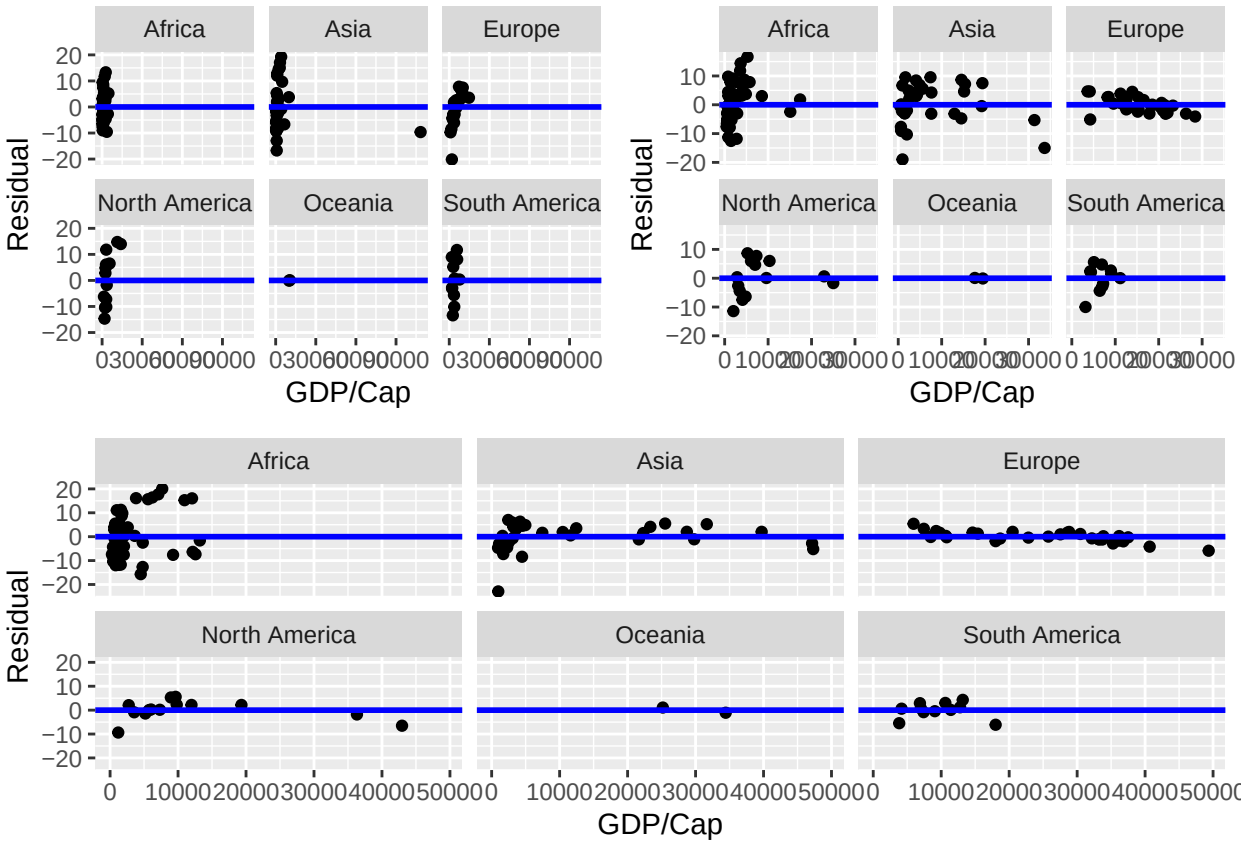

Figure 4: Residuals against GDP/Cap by continent for 1952(Top L), 1982(Top R), 2007(Bottom)

We next use the plots below in Figure 5 of the residuals against the fitted values from each of the models to assess the assumption of constant variance. As we can see, it would be fair to say this would hold for the year 1952 but it becomes slightly more dubious as we move to 1982 and 2007 which is what we'd expect given the inequalities that have emerged over time between developing and developed nations within the continents.

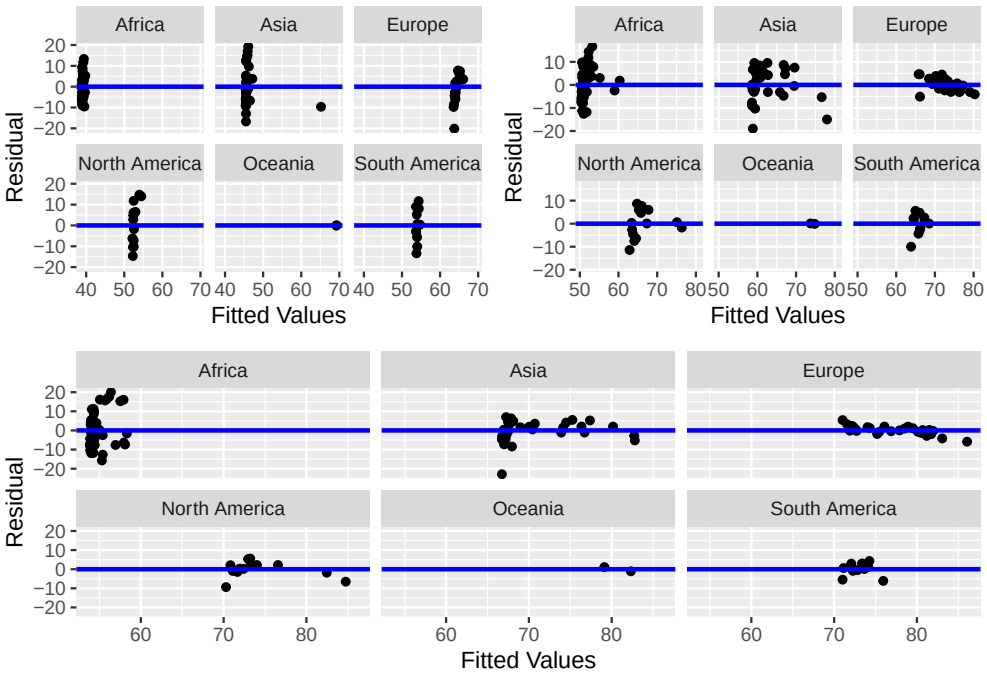

Figure 5: Residuals against fitted values by continent for 1952(Top L), 1982(Top R), 2007(Bottom)

We now lastly move on to assess whether the residuals are normally distributed and we do so by examining histograms of the residuals below in Figure 6,

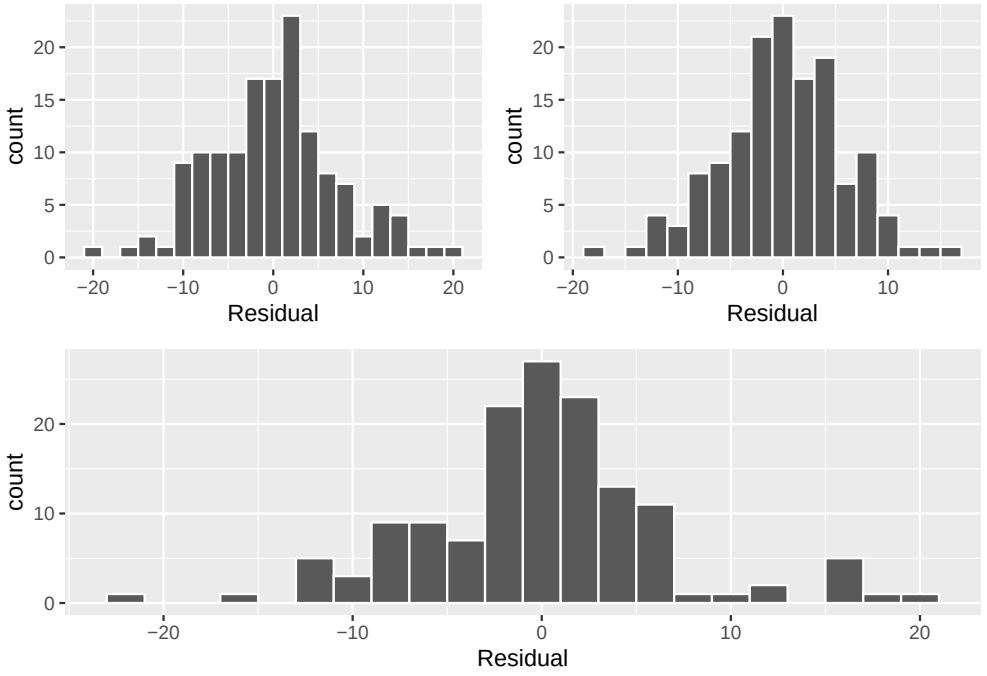

Figure 6: Histograms of the Residuals for 1952(Top L), 1982(Top R), 2007(Bottom)

From Figure 6 above we see that for each of the years, the histogram of the residuals appears to be roughly

symmetrical and bell-shaped suggesting we do indeed have normally distributed residuals. There does appear to be some outlying values at the extremes but nothing significant and there is no extreme skew for any of the years either.

4 Conclusion

Overall, it appears this analysis has backed-up the logical way of thinking when it comes to life expectancy around the world in that as time progresses one would likely believe that the average life expectancy of humans would increase due to medical, technological and dietary advancements that humans have made over decades and centuries and from examining the 3 years 1952, 1982 and 2007 we can see that this seems to be the case. Despite this however, the issue of health inequalities arises across the world with these advancements as some continents become more and more well-off, others are left behind and the gap is only likely to increase as we have seen here with comparing Africa in particular to the other four continents. Imporvements could be made to this analysis in the future by taking into account socio-economic factors such as education and income to really assess how inequalities are developing across the globe instead of only looking at factors such as life expectancy, population and GDP per Capita