TOPICS IN APPLIED STATISTICS – SPRINT 2019
Li Ai (uniwander@gmail.com) and Carlos Sathler (cssathler@gmail.com)

Mini Project 1 – Part 2

Topic: Life Expectancy

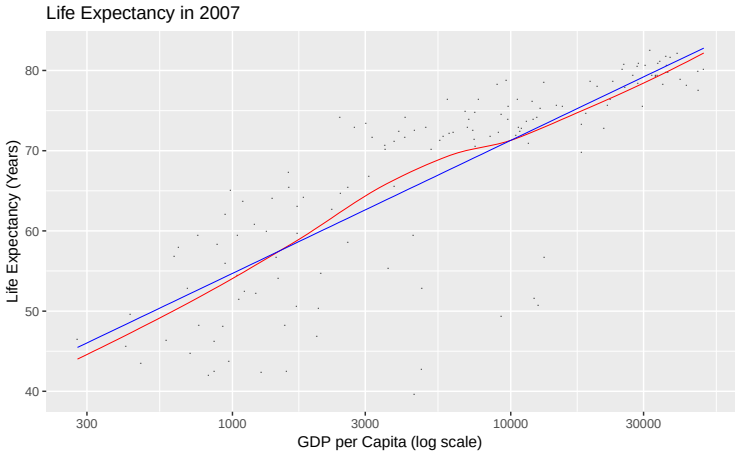
Major research question - Can the increase in life expectancy since World War 2 be largely explained by increases in GDP per capita?

Data - R "gapminder" dataset contained in package of the same name. For detailed description of dataset see https://cran.r-project.org/web/packages/gapminder/README.html#what-is-gapminder-good-for

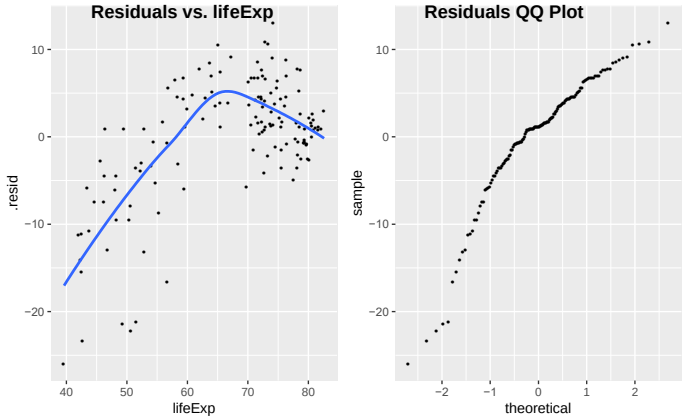
GDP and life expectancy in 2007

We show below the relationship between life expectancy and log of GDP per capita in 2007. The blue line on the plot is the linear model fit to the data. The red line is the loess fit, whose shape curves to capture the non-linear patterns in the data. The fact the two lines are closely located indicates there is evidence that the relationship between life expectancy and log GDP per capita might be liner in 2007. A simple linear model would describe the trend of the data as follows:

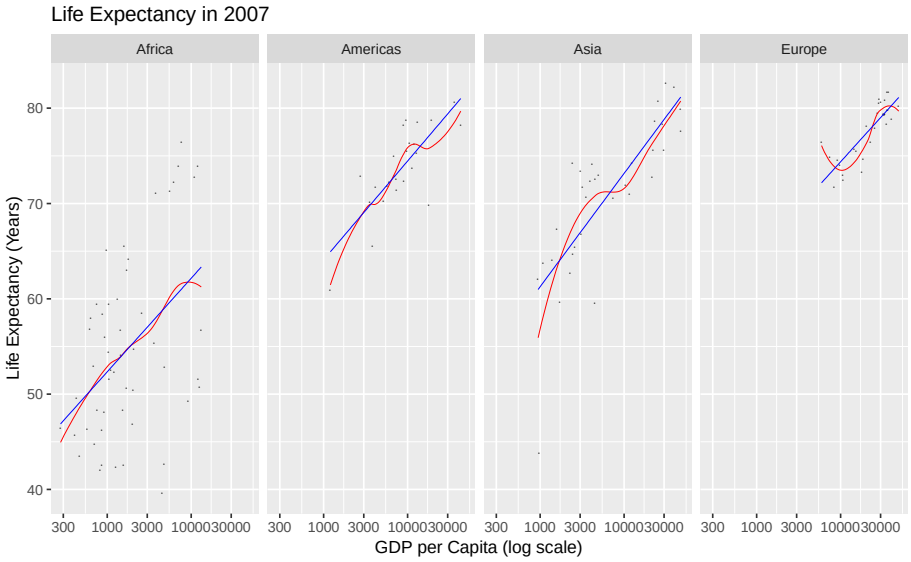
$$\text{LifeExp} = 4.95 + 7.2 \log(\text{GDPperCap}).$$



However, because residuals of a simple linear regression fit to the data do not show independence to life expectancy, and are not normally distributed (see figure below), we question the accuracy of the coefficients above (p-value from R lm() is questionable) as the standard error is underestimated. Additionally, we should not make probabilistic statements about point location. It is also worth noting that some points are scattered at a distance from the fitted line, suggesting that a large part of the variance in life expectancy is not explained by GDP per capita only.

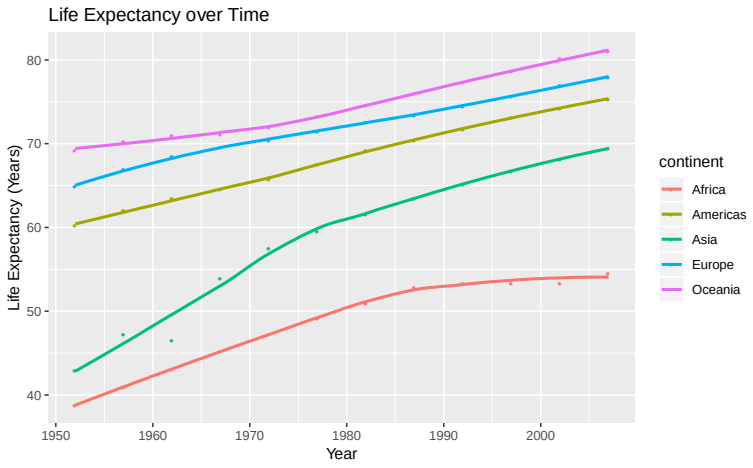


While the relationship between life expectancy and log of GDP per capita is different in different continents, they all seem to be less linear than the overall relationship as shown above. Below we show the first plot broken down by continent (we omit Oceania because it does not have enough data for the loess regression). Notice the loess red line is not so close to the linear blue line in most continents. Europe, in fact, shows a sinusoidal trend pattern, rather than a linear one. Additionally, several continents display decreasing life expectancy for certain GDP intervals. For example, both Africa and Europe show a decreasing trend at the end of the GDP range by loess fitting, indicating that there is significant non-linear relationship.



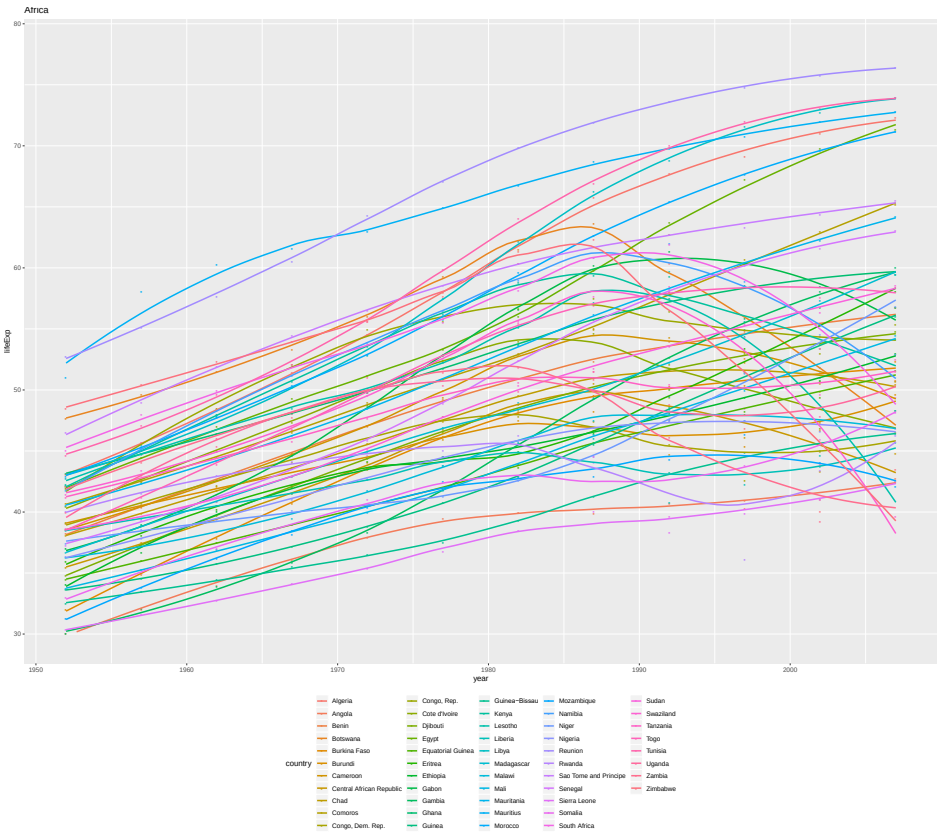
Life expectancy over time by continent

Life expectancy has improved over time for all continents, except Africa, which hasn't improved much since the late 90's. Americas, Europe and Oceania show a (mostly) linear increase over time. Asia life expectancy was improving faster than other continents until the mid 60's. After that, the rate of improvement started to match that of Americas, Europe and Oceania. In 1952 the life expectancy gap between Asia and the Americas was about 15 years. The gap was down to about 10 years in 2007, so Asia has partially caught up with the Americas.

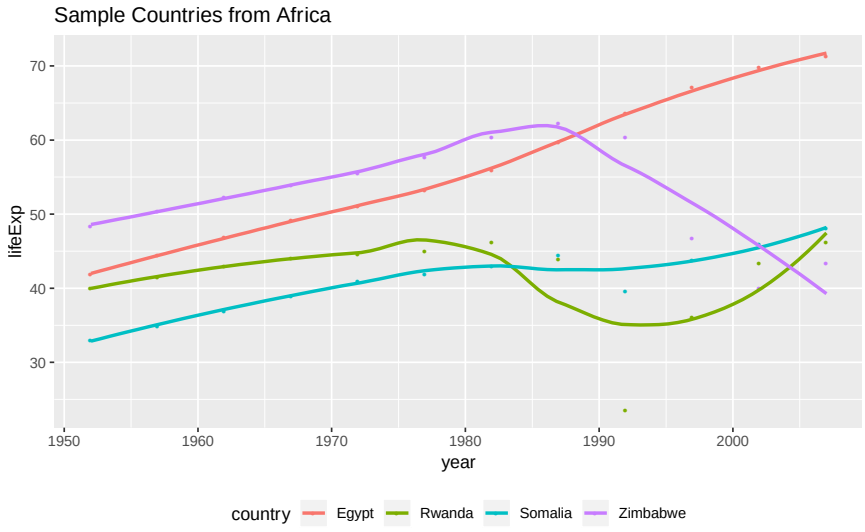


The analysis of life expectancy at the country level, by continent, show variations of life expectancy which explain the trends we see in the plot above. If we analyze Africa in more detail (next page) we notice that all countries show similar improvement rate in the period between 1952 and 1960. After that, many countries go through a period of decreasing life expectancy. One of the countries shows a saddle pattern.

Li Ai (uniwander@gmail.com) and Carlos Sathler (cssathler@gmail.com)



Looking closer at some of these countries we can make sense of the average numbers for Africa. Several countries in Africa were impacted by wars, such as Zimbabwe and Rwanda, or famine, such as Somalia. Contrast the trend of life expectancy for these countries with that of Egypt, which has enjoyed relative stability since 1952.



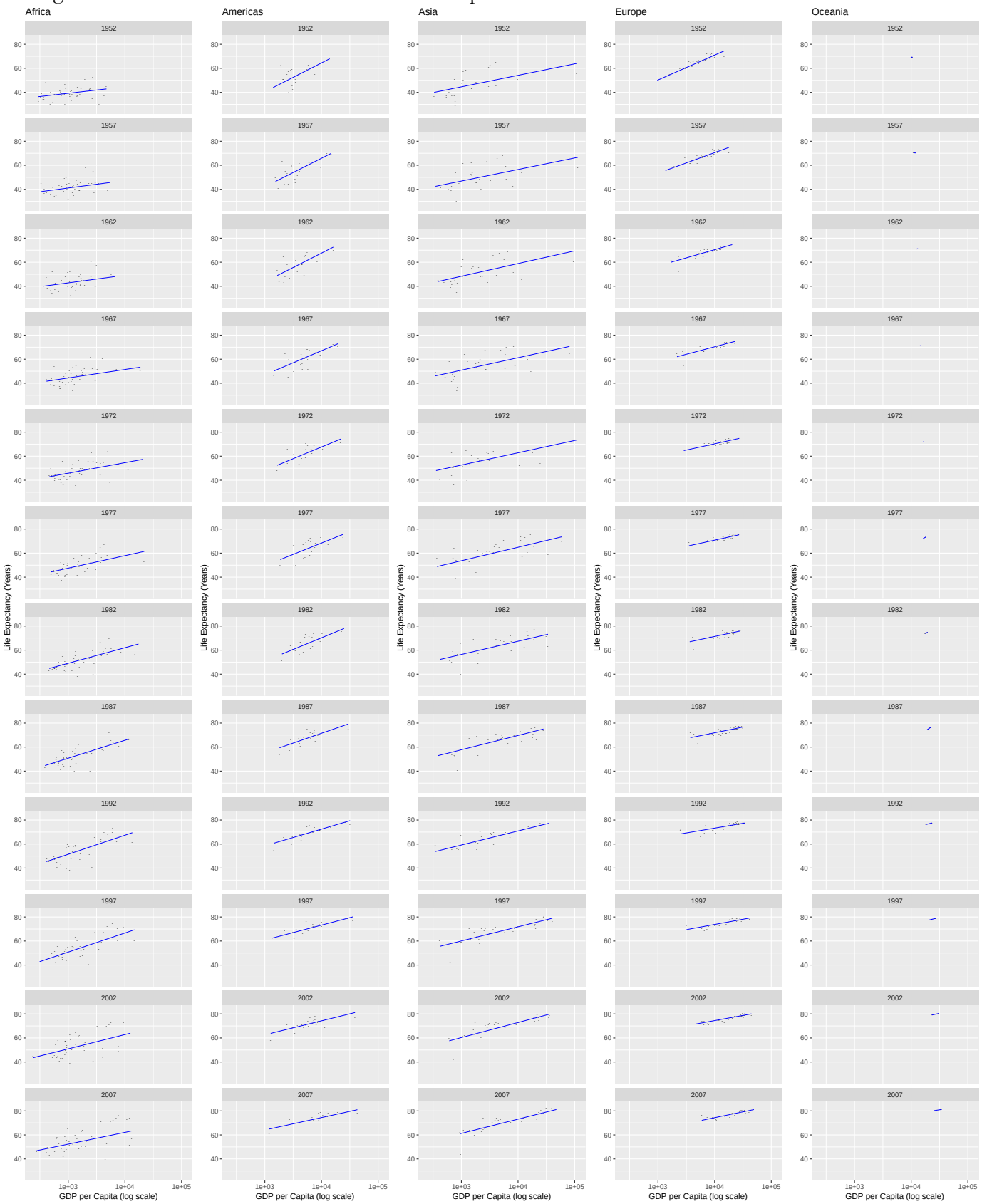
Similar patterns can be observed for countries in Asia (e.g. Cambodia and Iraq), Americas (El Salvador) and even Europe (Montenegro), which also went through wars. However, no continent went through as many wars as Africa, so while many countries in that continent show consistent growth in life expectancy, on average the continent significantly lags the others. (We omit additional plots in this report due to space limitations.)

TOPICS IN APPLIED STATISTICS – SPRINT 2019

Li Ai (uniwander@gmail.com) and Carlos Sathler (cssathler@gmail.com)

Changes in the relationship between GDP and life expectancy over time

The figure and table below summarize the relationship in each continent over time.



TOPICS IN APPLIED STATISTICS – SPRINT 2019

Li Ai (uniwander@gmail.com) and Carlos Sathler (cssathler@gmail.com)

Continent	Evolution of relationship between life expectancy and GDP per capita (log)
Oceania	The data points for the 2 countries in this continent clearly move from the bottom left to the top right of the yearly plots, indicating a pattern of increasing life expectancy as time and GDP increase in the continent.
Europe	The data points for Europe also move from bottom left to the top right of the yearly plots, indicating similar trend observed for Oceania. We also observe that the slope of the regression line capturing the strength of the relationship between life expectancy and GDP per capita decreases in the period from 1952 and 1992, when it seems to stabilize. Additionally, the continent shows smaller variance for both life expectancy and GDP over time, suggesting a reduction in income and social inequality.
Asia	In this continent, we see the data points moving up over time, more so than they move to the right, from 1952 through 1992. That means the increase in life expectancy over time in this continent is more pronounced than the increase of GDP over the same period. After 1992 we start to see the data points also moving right, indicating both life expectancy and GDP increased from 1992 and 2007. The slope of the regression line remains mostly constant from 1952 to 1972, after which it appears to increase ever so slightly until 2007, indicating a strengthening of the relationship between life expectancy and GDP per capita.
Americas	In the Americas there is a clear trend of increasing life expectancy and GDP per capita from 1952 and 2007. The slope of the regression line showing the strength of the relationship between life expectancy and GDP per capita decreases, for the most part steadily, from 1952 through 2007. The spread of the GDP per capita data increases consistently after 1982, suggesting an increase in income inequality among countries in this continent.
Africa	The data clearly moves right and up between 1952 and 1992, indicating increase of both life expectancy and GDP per capita during this period. After that, the increase continues, but is less pronounced from 1992 through 2007. The slope of regression line showing the positive correlation between life expectancy and GDP per capita increases until 1992 after which it seems to decrease slightly over time.

Below we show the correlation coefficients between (1) life expectancy and time, (2) life expectancy and GDP per capita and, (3) time and GDP per capita. The coefficients express the strength of the relationship between these variables. Clearly, changes in life expectancy are not entirely explained by changes in GDP per capita. Time definitely has an effect on life expectancy in addition to GDP, as seen in the first column of the table, particularly for Europe and Oceania. We speculate that advances in medicine and public health systems, over time, explain why time positively impacts the increase in life expectancy.

	Correlations		
Continent	Life Expectancy and Year	Life Expectancy and Log GDP per capita	Year and Log GDP per capita
Africa	0.547	0.426	0.160
Americas	0.680	0.558	0.306
Asia	0.660	0.382	0.137
Europe	0.706	0.781	0.609
Oceania	0.977	0.956	0.926
All Continents	0.436	0.584	0.227

When comparing the plots for each continent, except for Africa, we notice a "convergence" across the continents towards higher life expectancy. We observe, in general, the data points tend to form a flatter pattern at the top of the plots, suggesting by 2007 we may have reached a limit to life expectancy, around 80 years. That limit doesn't seem to be affected by either time, or increases in GDP, since the richest countries in all continents (excluding Africa) don't see higher life expectancy, as they get richer over time.

Li Ai (uniwander@gmail.com) and Carlos Sathler (cssathler@gmail.com)

Conclusion

Life expectancy is largely explained by increases in GDP per capita however time is also a factor. We speculate that advances in medicine and public health systems explain why time has a positive effect on life expectancy, in addition to prosperity expressed as GDP per capita.

Below we offer 5 linear models showing life expectancy (Y) as a function of log GDP per capita (X1), time (X2) and sometimes second-order polynomials. The P-values of the factors chosen are all equal or less than 0.05 (please refer to the Rmd file for more information if interested) so that they are considered statistically significant in this case. These models are imperfect, but they provide some explanatory power to describe the relationship between these variables. We show the R-squared score for each model. R-squared reflects the percentage change in the dependent variable Y that is explained by changes in the independent variables X1 and X2.

Continent	**Model**	**R-squared**
Africa	Y = – 16,390 + 4.753X1 + 16.33X2 – 0.004X2^2	0.511
Americas	Y = – 653.668 + 38.904X1 – 1.801X1^2 + 0.262X2	0.725
Asia	Y = – 642.22 + 4.823X1 + 0.335X2	0.687
Europe	Y = – 172 + 4.94X1 + 0.1X2	0.796
Oceania	Y = – 341 + 0.21X2	0.952
All Continents	Y = – 636 + 18.29X1 – 0.641X1^2 + 6.177X2 – 0.002X2^2	0.723

The comparison of the R-squares confirms what we have observed from the various plots in the previous sections: GDP per capita, when combined with time (year), can explain the change in life expectancy from 1952 to 2007 in most continents fairly well except for Africa, where the life expectancy has been affected by other important factors and thus shows a larger variation and sometimes a different pattern.

Note: Please refer to the appendix for sample residue plots for one of the continents (Americas). The plots reveal imperfections and violations of key assumptions of the linear model, thus suggesting that more factors and/or other non-linear relationships would need to be considered should one want to improve the model we selected.

TOPICS IN APPLIED STATISTICS – SPRINT 2019

Li Ai (uniwander@gmail.com) and Carlos Sathler (cssathler@gmail.com)

APPENDIX

Sample residual and QQ plot for regression model for the Americas continent:

$$Y = -653.668 + 38.904X1 - 1.801X1^2 + 0.262X2$$

Issues:

- Residuals do not show independence to life expectancy
- Residuals do not follow normal distribution