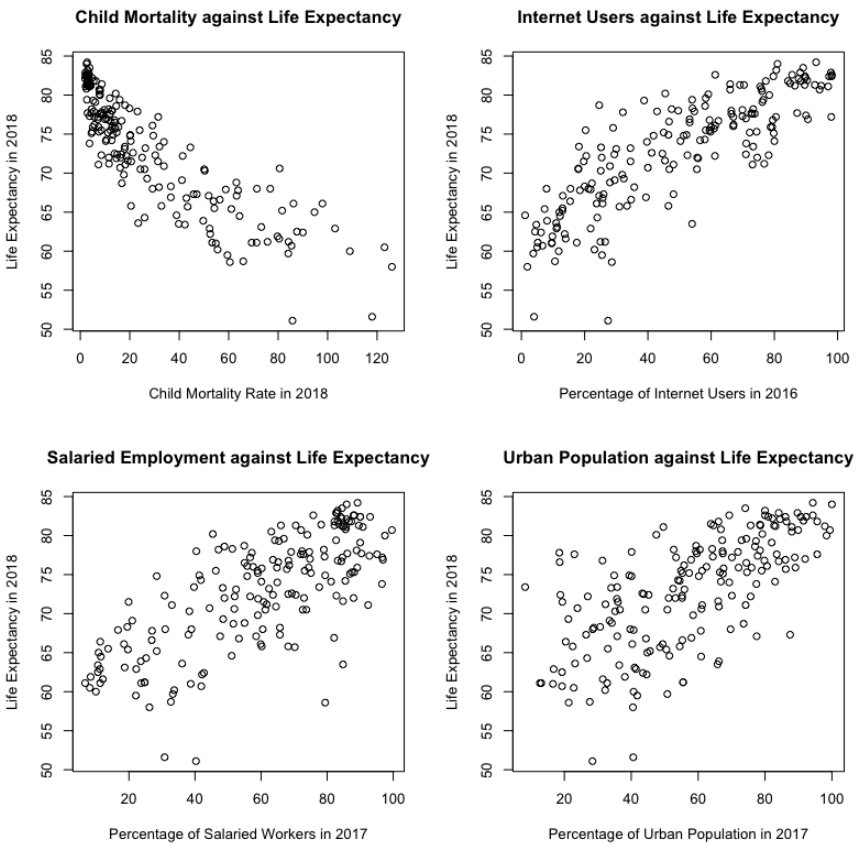
Figure 1: Scatterplots that show substantial linearity



Life expectancy in 2018 (lifeexp2018) is defined as the average number of years that a baby born in 2018 is expected to live. There are 187 different countries recorded in the data. The average life expectancy is 72.73 years. Only 8 out of 187 countries have life expectancy below 60. The lowest is 51.1 years in Lesotho, Sub-Saharan Africa and the highest is 84.2 years in Japan. Our task is to investigate the main drivers behind lifeexp2018. In the data, we have economic (agriculture, income, inflation and employment), technological (mobile phones, broadband subscribers and internet users), demographical (child mortality, children per woman, urban population and population density), and geographical variables (four regions and world bank regions) which are used as potential covariates. By looking at the scatterplots of each of them against lifeexp2018, we observe that some may contain important information as they show substantial linearity with lifeexp2018 in Figure 1. These variables are child mortality rate in 2018 (child_mort_2018), number of children per women in 2018 (child_per_woman_2018), percentage of internet users in 2016 (internet_2016), percentage of self-employed in 2018 (self_employed_2018), percentage of salaried workers in 2017 (sl_emp_2017) and percentage of urban population in 2017 (unban_pop_2017). In the plots of broadband subscribers in 2016 (broadband_2016) and income per person in 2018 (income_pp_2018), they seem to show a nonlinear relationship with lifeexp_2018.

The variable four_regions is the 'general version' of variable worldbankregion. Countries can be divided into four regions or seven world bank regions. Africa as a whole has the lowest average life expectancy but in world bank regions, Middle East & North Africa has the third-highest average life expectancy while Sub-Saharan Africa has the lowest. If we include worldbankregion instead of four_regions in our model, more information could be added to better explain the distribution of life expectancy in different regions. Therefore, in our final model, worldbankregion is included instead of four_regions.

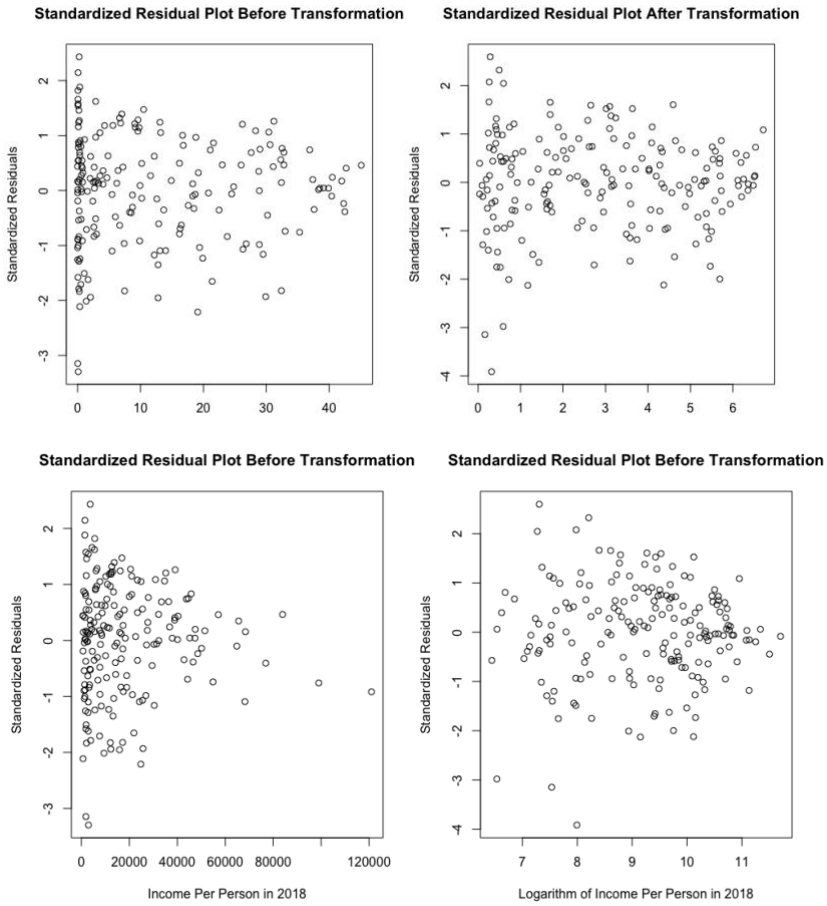
We continue to look at the data of potential covariates. The five-number summary statistics (FNSS) for annual percentage inflation in 2017 are -1.11, 1.64, 34.06, 6.82, 5220.00. The greatest inflation occurs in Venezuela which is 100 times more than the average. FNSS for population density in 2018 is 2.01, 33.85, 200.24, 183.00 and 8270.00. The highest value occurs in Singapore and is 40 times more than the average. The variations in inflation and population density across most of the countries are very little. Variables with little variation contain not much useful information in explaining the response variable. In our final model, inflation and population density are not included.

Intuitively, we would suspect some (or even many) of the potential covariates to be correlated. For example, the percentage of self-employed (self_employed_2018) and the percentage of the salaried workers (sl_emp_2017) would be expected to have high correlation because, in essence, they are complementary variables of each other in describing the working population. Although these two variables are in different years, with a one-year difference, the data should not alter too much. A calculation of the Pearson correlation coefficient (PCC), which is approximately -1, and the scatterplot between these two confirmed our speculation. Also, we would expect child mortality and the average number of children per woman to be correlated. Logical reasoning to have more children per woman could be that the country has a high child mortality rate and is usually less developed with relatively low hygiene and medical care standard. For more developed countries with lower child mortality rate, women tend to be more educated

and thus have a more career-focused lifestyle and choose to have fewer children. Our speculation is further confirmed with the scatterplot and PCC (0.868). We chose to omit one of each pair of correlated variables to avoid collinearity issue. Therefore, self_employed_2018 and number of children per woman are not included in our final model.

In the process of looking at the plots between each pair of possible covariates, we observe that the variable, internet_2016, shows close linearity with many other variables. For example, its PCCs with broadband_2016, child_mort_2018, income_pp_2018 and sl_emp_2017 are all about 0.8. This made us speculate that internet_2016 is itself a variable affected by many other factors. We compared two models with and without internet_2016. In the model including Internet_2016 and other variables that show correlation with internet_2016, we found out that internet_2016 was unable to provide much information for lifeexp2018 and seemed redundant. We also tried to keep internet_2016 and see if removing some of the correlated variables was a good idea using the F-test. The answer is no. This shows that internet_2016 contains similar information as the other correlated variables but could not explain the variation in response as well as those correlated variables. We confirmed our previous speculation by building linear regression using internet_2016 as the response and the correlated variables with internet_2016 as covariates. About 89 per cent of the variation in internet_2016 could be explained. This shows indeed internet_2016 is an underlying variable collectively affected by other factors. Therefore, internet_2016 is not included in our final model.

Continuing from our previous analysis, we analyzed the remaining variables. We decided to delete agri_2016 from our model. In the output of the model summary, the p-value for agri_2016 is 0.29 (fairly large) which suggests that in hypothesis testing, the removal of agri_2016 from the model is not rejected. In the scatterplot of agri_2016 against lifeexp2018, it indeed shows little linearity. These all suggest that agri_2016 makes trivial improvement in explaining the variation in response. After removing agri_2016, variables mobile_2016 and sl_emp_2017 show moderately large p values (0.099 and 0.076). Using ANOVA F-test, we compared models with and without these two variables. Residual sum of squares is reduced by 43.56 and the p-value is 0.086 (quite large). This shows that the two variables provide little explanation for lifeexp2018. Furthermore, the technological and economic aspects of these two could be supplemented by the remaining variables, so we decided to drop them. Our model has broadband_2016, child_mort_2018, income_pp_2018, urban_pop_2017 and worldbankregion as our covariates.



By analysing the plots of standardized residuals (SRs) against each covariate, we observed some systematic patterns as shown in Figure 2 before using transformation. The variance of SRs seems to decrease as broadband_2016 and income_pp_2018 increase. These are indications of departure from the assumptions of linearity and homoscedasticity. A log or square root transformation on the covariates could mitigate the situation. We decided to use square root (sqrt) on broadband_2018 and logarithm (log) on income_pp_2018 as they make randomize the points at our best knowledge. After transformation, in the plot of SRs against fitted values, the points seem pretty random along the horizontal axis. The assumption of homoscedasticity and linearity after transformation seems adequate here. The plot in Figure 3 is the normal probability plot of the SRs which give indications about any departure from the assumption of normality of the error terms. Most of the points lie on or close to the y=x line so there

Figure 2: Standardized residual plot before and after transformation

seems no issue with the assumption of normality. Finally, we want to check for the assumption of independence of observations. In the scatterplot of SRs after transformation, we observed that the points are random and show no obvious pattern. The source of the data is believed to be reliable and collected across the world so the assumption of independence of errors is not greatly worried in our model.

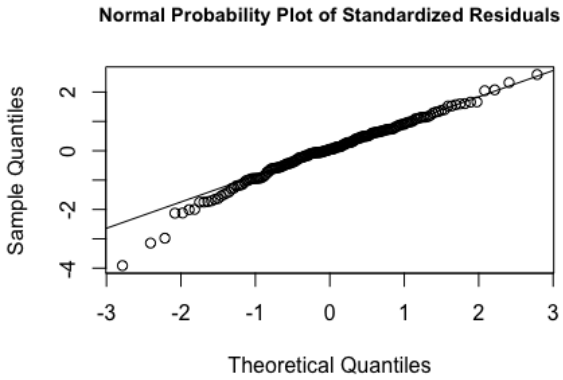


Figure 3: Normal probability plot

We observed three outlying points in the plot of SRs against fitted values. Afghanistan has an unexpectedly low life expectancy of 58.7 years, a very low percentage of broadband subscribers at 0.0254 and a quite high child mortality at 65.9. Although it is in South Asia supposedly with the highest average life expectancy, it is a country constantly in warfare which explains its low life expectancy. Lesotho in Sub-Saharan Africa has an unexpectedly low life expectancy. As we mentioned before, Lesotho has the lowest life expectancy of 51.1 years. The child mortality rate of Lesotho is at the 11th highest. However, its income per person is not as low as expected which could be the result of wealth disparity in the country. Timor-Leste has a very low percentage of broadband subscribers, income and below-average salaried employment but it shows an unexpectedly high life expectancy of 73.3 years. The extreme observations seem not to be incorrectly recorded but are merely rare situations. Deletion of the three observations does not alter estimates of the coefficients too much. Inclusion of them increases the credibility of our model as it takes rare situations into account. Thus, we did not drop these extreme observations in our model.

Our final model shown in Table 1 provides a reasonable fit for the data. It can explain about 83% of the variation in life expectancy which is quite good for a real dataset. We have checked using the plots of SRs against dropped variables to see if any important variables are omitted accidentally. The result is negative. There may be other unknown factors affecting the response. As we can see in Table 1, life expectancy is positively affected by the percentage of broadband subscribers, income per person and the percentage of the urban population. Among these three, an increase in income seems to increase life expectancy the most. Life expectancy is negatively affected by child mortality rate which agrees with common sense. In our reference category, Sub-Saharan Africa has the lowest life expectancy. East Asia & Pacific, Middle East & North Africa, Europe & Central Asia, North America, Latin America & Caribbean and South Asia have ascending order of expected life expectancy.

Coefficient	Estimate	Std. Error	t-value	p-value
Intercept	57.87	3.24	17.85	<2e-16
Sqrt(Broadband_2016)	0.59	0.23	2.54	0.012
Log(Income_pp_2018)	1.23	0.37	3.34	0.001
Child_mort_2018	-0.10	0.02	-6.23	3.33e-09
Urban_pop_2017	0.06	0.01	4.13	5.68e-05
East Asia & Pacific	1.07	0.91	1.17	0.242
Middle East & North America	1.42	1.01	1.40	0.163
Europe & Central Asia	1.63	1.01	1.61	0.108
North America	1.64	2.35	0.70	0.484
Latin America & Caribbean	2.15	2.15	2.36	0.019
South Asia	3.97	1.23	3.23	0.001

Table 1: Summary of our final linear model

As reliable as our model can be, there are limitations to it. One limitation of our model is that we are not quite satisfied with the plot of SRs against the square root of broadband_2016 in Figure 2. However, we have applied our best knowledge at hand. I think this is a limitation using transformation as a method to moderate the effect of departure from linearity and homoscedasticity. We may not always find a transformation that does a perfect job in randomizing the points. Another possible limitation of the data could be that most of the data are recorded in 2016 and 2017 which is close to the year we wish to investigate 2018. However, the impact of some covariates may take longer to emerge. For example, the impact of internet accessibility in the 21st century on income and subsequently on life expectancy may emerge in later years. The renders our model less resistant across time.