Bridging the Gap
Life Expectancy Analysis Between Developing and Non-Developing Countries

Understanding the influence of healthcare, economic and social factors on life expectancy around the turn of the millennium

Introduction

For this study a WHO dataset was used, with a total of 2938 rows and 22 columns. After the preprocessing step a total of 20 columns and 183 records were considered.
Using this data we aim to answer the following questions.

1. What are the significant factors that affect life expectancy?
2. Is there a significant difference in life expectancy between developing and non-developing countries regarding the influence of all other significant factors?

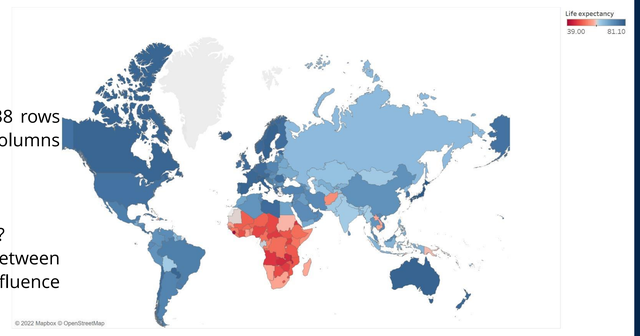

Fig 1.: Life expectancy per country on the year 2000

Methodology

In a first moment, a multiple linear regression model with all features available for the year of 2000 was built, with Life Expectancy as the dependent variable. (Fig 2). As a way to access the significance of each independent variable the p-value was studied, considering values below 0.3 significant, since there is a high probability of substantial changes in its value due to transformations conducted to the variables later.
The correlations between variables was also plotted (Fig. 3), as a high correlation between independent features results in redundant information that will affect the coefficient values, failing to assure condition 3 and so invalidating our model. On top of that VIF (Variance inflation factor) values were taken into account, as it shows how well a single independent variable can be explained by a linear regression of the remaining. A value above five is considered highly multicollinear. To ensure quality, the variable with the higher VIF was dropped before calculating the VIF values again to verify if there was any need to omit another feature. This process was repeated until all values were below 5. This step lead to taking infant.deaths and thinness.1.19. years as omitted variables.

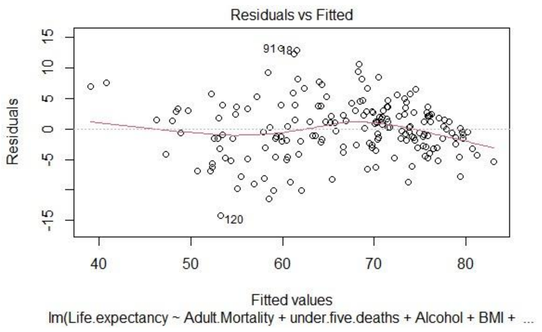

Fig 2.: Plot of the residuals (ordinate) against the fitted values (abscissa) for the linear model taking the Life expectancy as the dependent variable. The independent variables are named in Table 1.

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	5.107e+01	2.224e+00	22.964	< 2e-16	***
Adult.Mortality	-1.503e-02	3.253e-03	-4.621	7.74e-06	***
infant.deaths	1.964e-01	4.427e-02	4.437	1.67e-05	***
Alcohol	2.085e-01	1.238e-01	1.684	0.0940	.
percentage.expenditure	4.284e-05	6.520e-04	0.066	0.9477	
Hepatitis.B	9.737e-03	1.728e-02	0.564	0.5738	
Measles	1.423e-05	2.802e-05	0.508	0.6122	
BMI	1.144e-01	2.453e-02	4.664	6.43e-06	***
under.five.deaths	-1.430e-01	3.227e-02	-4.431	1.71e-05	***
Polio	3.425e-02	1.759e-02	1.947	0.0533	.
Total.expenditure	6.677e-03	1.816e-01	0.037	0.9707	
Diphtheria	3.757e-02	1.557e-02	2.412	0.0170	*
HIV.AIDS	-4.369e-01	6.255e-02	-6.984	6.91e-11	***
GDP	1.190e-04	1.001e-04	1.188	0.2366	
Population	-3.654e-08	1.666e-08	-2.193	0.0297	*
thinness..1.19.years	-3.624e-01	2.319e-01	-1.563	0.1200	
thinness.5.9.years	5.072e-01	2.299e-01	2.207	0.0287	*
Income.composition.of.resources	-1.097e+00	1.859e+00	-0.590	0.5558	
Schooling	7.824e-01	1.367e-01	5.724	4.88e-08	***
Developed_Dummy	1.153e-01	1.330e+00	0.087	0.9310	

Table 1.: Summary of the ML Regression Model using all available features, with a residual standard error of 4.5 and a multiple R^2 of 0.83 and adjusted R^2 of 0.8108. Highlighting the variables with no significance for our dependent variable.

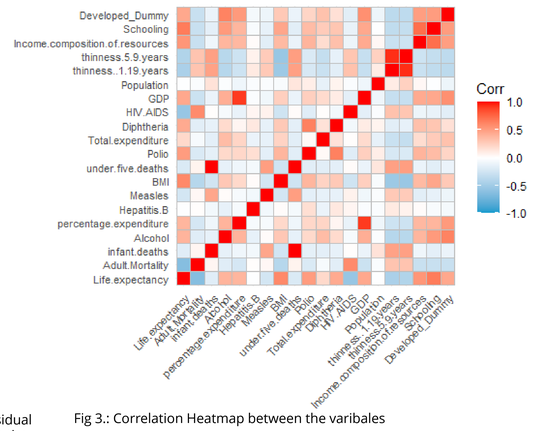

Fig 3.: Correlation Heatmap between the varibales

Assumption 4 states that non linear functions of the explanatory variables should not add any significance to the model if it is well specified. To verify this assumption the Reset-Test was applied, in which we added the squared and cubed fitted values to the model and applied F-statistics. If H_0 was rejected a non linear function of the independent variables was added to the model and the test repeated. This was done until we failed to reject the null hypothesis, thus the model specification was improved.

Assumption 5 states that the model must have homoscedasticity. If this is not given the OLS standard errors and test statistics are no longer valid. To review our model two tests were applied: The Breusch-Pagan Test and Special White Test. The first test uses the R^2 of the squared error, u^2, described as a linear regression of the independent variables then performs an F-statistics from which the p-value is retrieved. The second test is less restrictive, as its null Hypothesis (Heteroscedasticity is present) can be rejected if the p-value of the regression on the estimated fitted and the squared estimated fitted values is below 0.05. (See. Wooldridge 2018, pp.272)

Test	p-value	H_0	Result
RESET	0.1058	Well specified	Fail to reject with 5% (significance) level. Well specified.
Special White	0.0004	Homoskedasticity	Reject with 5% (significance) level. Heteroskedasticity.
Breusch-Pagan	8.15e-06	Homoskedasticity	Reject with 5% (significance)level. Heteroskedasticity.

Table 2.: Resume of the results of the final tests

As shown in Table 2, our model is under the presence of heteroskedasticity, therefore the robust standard errors of the coefficient estimators will be used for future interpretations. The other Assumptions for Multiple Linear Regression are assumed to be given. After taking the logarithm of our dependent variable and applying multiple transformations to the independent variables the final model was defined. It is plotted in Fig 4. and the coefficients for the independent variables are shown in Table 3.

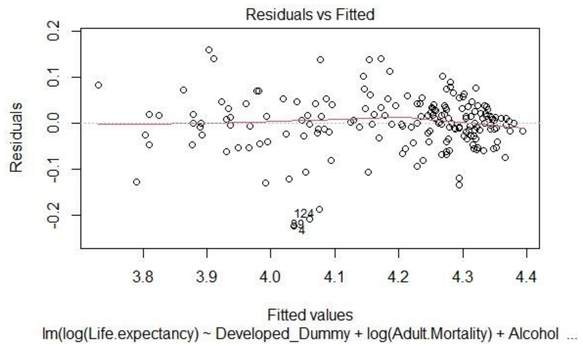

Fig 4.: Plot of the residuals (ordinate) against the fitted values (abscissa) for the linear model taking the log of Life expectancy as the dependent variable. The independent variables are shown in Table 3.

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	3.8421e+00	4.9047e-02	78.3337	< 2.2e-16	***
Developed_Dummy	2.0789e-02	1.4860e-02	1.3990	0.163634	
log(Adult.Mortality)	2.6558e-02	5.6495e-03	4.7009	5.307e-06	***
Alcohol	6.9622e-03	3.3472e-0	2.0800	0.039014	*
BMI	4.4833e-03	1.1871e-03	3.7768	0.000219	***
HIV.AIDS	-1.5546e-02	6.9526e-03	-2.2359	0.026649	*
Schooling	9.3880e-03	2.7014e-03	3.4752	0.000647	***
I(Adult.Mortality * HIV.AIDS)	2.7976e-05	1.2846e-05	2.1779	0.030786	*
I(Alcohol * BMI)	-1.3838e-04	6.3249e-05	-2.1879	0.030036	*
I(BMI * Diphtheria)	-3.5285e-05	1.2119e-05	-2.9116	0.004075	**
I(Adult.Mortality^2)	-1.3725e-06	2.1253e-07	-6.4580	1.059e-09	***
I(Diphtheria^2)	2.0938e-05	5.1040e-06	4.1022	6.320e-05	***

Table 3.: Summary of the final ML Regression Model after adjusting to respect the assumptions. With a R^2 of 0.88 and adjusted R^2 of 0.87

Results

Comparing the plots in Fig. 2 and 4, it's possible to detect an increase of linearity on the final model, since the red line is closer to the dashed line, which shows that the mean
Despite this increase of quality the heteroskedasticity seems to remain in both models, as suspected from the tests, this is visible by analyzing the distribution along the x-axis, it is clearly decreasing with a higher x, instead of the desired uniform spread.
For a final analysis it's also noted that there are some outliers in the final model.
Besides this plot study the R squared was also analyzed since it's a measure of how much of the variance can be explained by the model, as shown in Table 4 there was an increase of these values supporting the conclusion that there was an growth of the quality of the model.

Model	R-Squared	Adjusted R-Squared
First Model	0.8305	0.8108
Final Model	0.8810658	0.873415

Table 4.: Resume of the R-squared per model

Conclusion

From this study we retrieve our answers from Table 3, for the first question the factors that had a significant impact on the life expectancy on the year 2000 are the independent variables listed on Table 3 – with the exception of *Developed_Dummy*, those with a negative value decreased the expectancy and the ones with a positive increased it. For the particular case of HIV this result is no surprise since it was around 2000 that treatments for this infection started to appear.
To answer the second point the variable *Developed_Dummy* is studied, as the name indicates the binary value is 1 if the country is developed, this means that holding all factors constant there is only 2% difference on the target value with the change on the dummy and with a p-value of 0.16 it is not significant at all. This raised some concerns about the reliability of the model so an extra one was made not considering this feature, despite some minor changes on the coefficients it doesn't invalidate the model presented. (See notebook)
From this analysis it's hoped to help countries better understand which factors are relevant to its life expectancy in an effort to increase it, for example an investment in school and vaccination is highly recommended, other features like adult mortality and alcohol aren't straightforward to draw conclusions as so deeper research should be done in the future.

Bibliography

- Wooldridge, J.M. (2013) *Introductory econometrics: A modern approach*. Mason: Cengage Learning
- Unaids.org (2022) *HIV integration*, *UNAIDS*. UNAIDS. Available at: https://www.unaids.org/en/keywords/hiv-integration (Accessed: December 17, 2022).
- Kumar Rajarshi (2018) *Life expectancy (WHO)*, *Kaggle*. Available at: https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who
- *4.2 - residuals vs. fits plot* (no date) *4.2 - Residuals vs. Fits Plot | STAT 462*. Available at: https://online.stat.psu.edu/stat462/node/117/.

Authors: Lukas Stark 20220626 David Halder 20220632 Felix Gayer 20220320 Rita Soares 20220616