\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Factors Influencing Life Expectancy**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**The problem statement I addressed**

Life expectancy is one of the metrics used to measure health of a certain population. Life expectancy is not constant, it can change under the effect of some factors. I believe it would be beneficial to identify some factors influencing life expectancy as the interest in a longer healthier life has increased. Even though there are a lot of factors potentially influencing life expectancy, for this specific analysis variables “Schooling”, “Alcohol”, “BMI”, “Population”, “Expenditure on health” were chosen. Number of years spent on education (variable “Schooling”) could have a relationship with life expectancy. The more educated people are, the more information they have about the possible choices and opportunities of effectively living their lives and improving overall wellbeing. Can a relationship between education and life expectancy be found? Even though potential health benefits from moderate alcohol consumption are claimed to exist, as well as the benefits of alcohol helping boost social interaction, there are quite a few health risks associated with alcohol consumption. Are the risks significant enough to play a part in impacting life expectancy? High values of BMI are known to play a role in the increased risks of a variety of diseases but what is the exact relationship of BMI with life expectancy? Densely populated areas are known to affect the quality of the environment and availability of certain resources needed to maintain an improved wellbeing. Is there a relationship between population density and life expectancy? Would people enjoy a longer life if the government spent more on health?

**How I addressed the problem statement**

The data. The data set I used for this project had been compiled using the economic data retrieved from the United Nation website and the rest of the data retrieved from the World Health Organization repository. The data represents information for the time period of 2000-2015. After much thought and consideration, the variable “Population” was removed from the analysis due to a high number of missing values and invalid data points present.

Multiple Linear Regression. This project examined the hypotheses about the relationships between factors of education, BMI, population density, expenditure on health, alcohol consumption and life expectancy with the help of multiple regression analysis.

**Analysis**

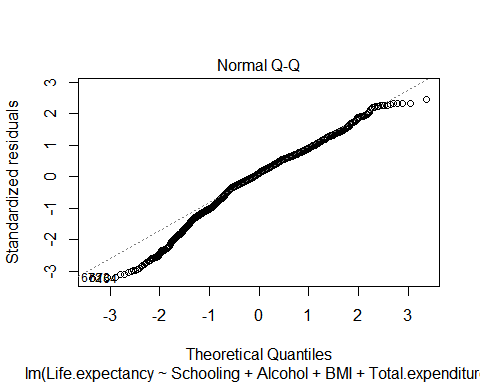
Assumptions of Multiple Linear Regression model were tested. Assumption of Independence and assumption of normality of residuals’ distribution were violated for the first regression model:

lm(formula = Life.expectancy ~ Schooling + Alcohol + BMI + Total.expenditure,   
data = newdataset2)

The Durbin-Watson test showed a value bellow 1, which indicated a positive correlation between adjacent residuals, and also a case of concern.

lag Autocorrelation D-W Statistic p-value  
 1 0.8607623 0.272879 0  
 Alternative hypothesis: rho != 0

Q-Q plot of Regression Standardized Residuals:



Not all of the points on the Q-Q plot lie on the straight line, which shows abnormal distribution of the residuals.

Solution

After performing some research on the possible causes of the independence violation, the decision to test the hypotheses including the data for only one year was made. The data for the same country for each subsequent year with 16 total years seemed to be similar and could have potentially caused the serial correlation affecting the first analysis.

The second model with observations for the year 2014 showed the following results for the estimated coefficients:

Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 41.76533 3.48571 11.982 < 2e-16 \*\*\*  
BMI 0.15755 0.05704 2.762 0.00716 \*\*   
Schooling 1.81364 0.31882 5.689 2.14e-07 \*\*\*  
Alcohol -0.02183 0.25298 -0.086 0.93147   
Total.expenditure -0.11451 0.26433 -0.433 0.66607

The variables BMI and Schooling are statistically significant. *p*-values for the coefficients of Alcohol and Total.expenditure are above .05, which indicates that there is no sufficient evidence to determine that effect exists for these variables at the population level.

Since it is pretty common to use *p*-values of the coefficients to decide which values should be in the final model and which shouldn’t, I only used BMI and Schooling in the final model:

lm(formula = Life.expectancy ~ BMI + Schooling, data = newdataset\_LED2)

Standardized betas

BMI Schooling   
0.2571045 0.5420347

*Schooling.* Standardized beta is 0.542. This value shows that when Schooling increases by one standard deviation (2.3), life expectancy increases by 0.542 standard deviations. Standard deviation for Life.expectancy is 7.65, so the change that will be observed is about 4.1463

*BMI.* Standardized beta is 0.257. This value shows that when BMI increases by one standard deviation (12.81), life expectancy increases by 0.257 standard deviations. Standard deviation for Life.expectancy is 7.65, so the change that will be observed is 1.96605.

Confidence intervals

2.5 % 97.5 %  
(Intercept) 35.28388475 47.4233148  
BMI 0.04557858 0.2617836  
Schooling 1.20135112 2.4044434

Estimated coefficients for easier comparison of the results:

Estimate Std. Error t value Pr(>|t|)  
(Intercept) 41.3535998 3.05001085 13.558509 1.975326e-22  
BMI 0.1536811 0.05432115 2.829121 5.897901e-03  
Schooling 1.8028972 0.30227485 5.964430 6.350759e-08

None of the confidence intervals cross 0 which means that the predictors are related to the outcome. The confidence intervals are small, which is an indicator of a good model.

Multicollinearity

VIF:

BMI Schooling   
 1.309279 1.309279

None of the VIF are greater than 10.

Average of VIF:

1.309279

The average of VIF values is not substantially greater than 1, which helps make a conclusion that there is no collinearity within the data.

Tolerance:

BMI Schooling   
0.7637793 0.7637793

None of the values have tolerance that is lower than 0.2, which indicates absence of collinearity.

Assumption of no multicollinearity has been met.

Independence

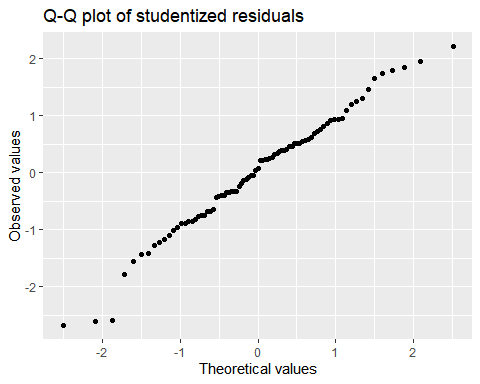
lag Autocorrelation D-W Statistic p-value  
 1 -0.1799472 2.345003 0.092  
 Alternative hypothesis: rho != 0

DW statistic is close to 2 and *p*-value is greater than .05, which indicates that assumption of independence has been met.

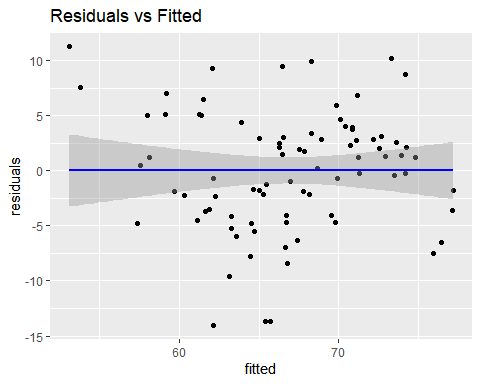
Outliers and Influential Cases

After performing case wise diagnostics and analyzing values with large residuals for cases of concern when looking at cooks.distance, leverage and covariance ratios, no cases of concern were found.

Residuals



It appears that the Studentized residuals are normally distributed, the scatters for the most part are on or very close to the normal distribution line.



Looking at the scatterplot of Residuals vs Fitted values I can tell that there is no evidence of nonlinear pattern to the residuals.

**Cross Validation of the Model**

Residual standard error: 5.505 on 80 degrees of freedom  
Multiple R-squared: 0.4954, Adjusted R-squared: 0.4828   
F-statistic: 39.27 on 2 and 80 DF, p-value: 1.315e-12

The F-statistic *p*-value of 1.315e-12 is highly significant, which implies at least one of the independent variables is significantly related to the dependent variable. R squared statistic explains how much variability in the dependent variable is accounted for by the independent variables. In this case it is 49.5%. The *adjusted* R squared at 48.3% is pretty close to the value of *R squared*. The difference is small, about 1.2%. The shrinkage means that the model derived from a population and not a sample would account for 1.2% less variance in the outcome. Using Stein’s formula, I get the version of R squared that equals to 0.463. The difference between this value and the value of unadjusted R squared is 0.495-0.463= 0.032, or about 3.2% percent. This value helps me with cross-validation of the model. The value is similar to the observed value of R squared, which is a good indicator of a cross validity of the model.

After cross-validating the final model with the help of Stein’s formula, and assessing the assumptions, I can tell that the model looks good without violations of assumptions, which makes me believe the results would be more accurate than for the model I originally built.

**Implications**

The two factors that seem to have a significant positive relationship with life expectancy are “*Schooling”* which belongs to the social factor, and “*BMI*” which is associated with the health factor. The variables could be used for predictions of life expectancy with the help of the model.

Complexity of BMI*.* The results of the estimated coefficients that I got with the help of my model indicate that there is a positive relationship between BMI and life expectancy. The discrepancy between high values of BMI and gains in life expectancy at first seemed confusing because my initial belief about BMI is that the greater, the worse the effects might be for the life expectancy. After reading a few studies (see the references), it was evident that the variable BMI is more complicated than it seems. There are a few main ranges for BMI (underweight, normal, overweight and obese), and for each of these ranges the effects on the life expectancy could differ. For the underweight (especially for malnourished, or those whose weight loss happened due to effect of diseases) the increase in BMI might mean longer life expectancy. For normal BMI range, according to the studies the BMI change that stays within the normal range is known not to affect life expectancy significantly (a small increase has been mentioned). It is the overweight and obese categories that might benefit from the decrease of BMI since increase of weight affects the likelihood of getting diseases associated with increased mortality rates, at the same time the increased effective treatments of the diseases known to be caused by increased weight in the obesity range also changes the overall picture of the effect of BMI on life expectancy. The quality of life might be significantly worse for people with the BMI in the overweight range but their life expectancy is known to have benefitted from the advanced treatments. After researching these findings with the help of the study [Obesity and Trends in Life Expectancy](https://www.hindawi.com/journals/jobe/2012/107989/) I had a better understanding why I got the results that I didn’t expect. It also makes sense to suggest to look at different ranges of BMI separately to get a clear picture of the findings and more possible reasons for the findings I see.

No significant effect of the variables *Alcohol* and *Total.expenditure* was discovered. I can’t conclude if they positively or negatively affected life expectancy from the analysis I performed since the results were not statistically significant.

The data were not letting me use “Population” where I wouldn’t be concerned about incorrect results. There were a lot of missing values, and a lot of the values were invalid. The analysis for this variable wasn’t performed in the end.

**Limitations**

The project has some limitations. Even though it provides results for effects of BMI and Schooling, it doesn’t explain other factors that could be influencing life expectancy. Considering performing the analyzes of other variables from the data set can allow for discovery of additional factors influencing life expectancy.

Incorrect data points present in the data set and missing values significantly reduced the amount of observations available thus reducing the possibility of accounting for all of the information for the variables. Finding the original data set and verifying the values would help with substituting the incorrect values of the secondary data set analyzed in the project.

The project only covered the values for the year of 2014. There are 15 more years which the project hasn’t analyzed. Further analysis of the rest of the years and comparison of the results could help with confirming or questioning the validity of the results obtained so far, or identifying the trends of effects on life expectancy over the given time frame.

Looking at different ranges of BMI separately could help with getting a clearer picture of the findings and more possible reasons for explanation of the results of this project.

**Conclusion**

This project examined the hypotheses about the relationships between factors of education, BMI, expenditure on health, alcohol consumption and life expectancy with the help of multiple regression analysis. Since the effect of only two of the variables I analyzed turned out to be statistically significant, I could only suggest that if we increase the number of years spent on education, our life expectancy might increase. The same is true for BMI, but the increase seen is a lot smaller than that for Schooling. Also, as mentioned above BMI is a complicated variable and its effect might depend on the specific range of BMI we are looking at, which makes Schooling the variable whose results seem to be the most influential and reliable. It is important to remember that the relationships discovered do not imply causation. Also, limitations of the project should be taken into consideration when reviewing the results.

References

1. [Association of All-Cause Mortality With Overweight and Obesity Using Standard Body Mass Index Categories](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4855514/)

2. [How Much Should We Weigh for a Long and Healthy Life Span? The Need to Reconcile Caloric Restriction versus Longevity with Body Mass Index versus Mortality Data](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4115619/)

3. [Body-mass index and cause-specific mortality in 900 000 adults: collaborative analyses of 57 prospective studies](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2662372/)

4. [Obesity and Trends in Life Expectancy](https://www.hindawi.com/journals/jobe/2012/107989/)

5. [The impact of increasing education levels on rising life expectancy: a decomposition analysis for Italy, Denmark, and the USA](https://genus.springeropen.com/articles/10.1186/s41118-019-0055-0)