ISTA 321 Midterm Project Report

Claire Baker and Sherif Shawashen

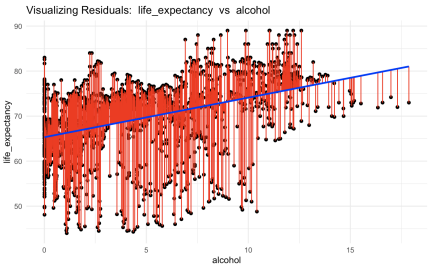
Introduction

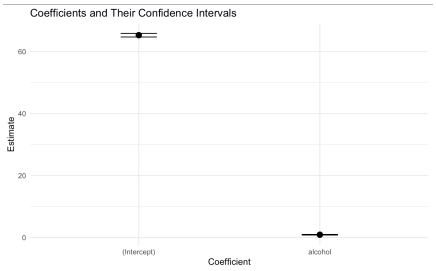
For our Midterm Project, we were assigned to analyze and apply data mining techniques to the dataset, “Life Expectancy (WHO)”, a statistical analysis of factors influencing life expectancy. The dataset was compiled from the World Health Organization and contains 2938 observations of 22 variables. The columns of this dataset are country, status (developing or developed), year (2000-2015), life expectancy, adult mortality, infant deaths, alcohol consumption, percentage expenditure, hepatitis b immunization coverage, measles cases, BMI, deaths under five-years-old, polio immunization coverage, total expenditure, diphtheria immunization coverage, deaths from HIV/AIDS, GDP, population, prevalence of thinness in adolescents aged 10-19, prevalence of thinness in adolescents aged 5-9, income composition of resources, and years of schooling. In order to clean this dataset, we aimed to make the variable names more cohesive and to omit “NA” values. To maintain clarity, we changed all of the variable names to lowercase and changed “.” to “_” in the variable names to access them. We also used the R function “na.omit” to clean the dataset of any “NA” values. Also, we created functions in order to perform our data analysis more efficiently. We created functions for visualizing the residual plots, visualizing the linear regression, and visualizing the confidence interval to maintain cohesion within our output. It is also important to note that our dataset contains a column for year, so this could affect some of our data. There could be trends or changes over time that are not accounted for in our models that could affect the relationships between our variables.

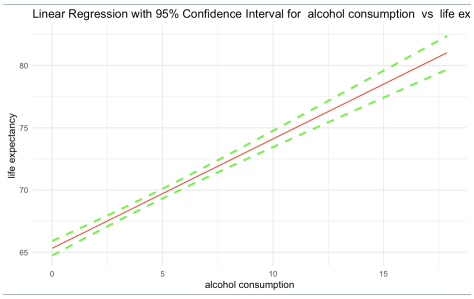
Research Question #1

The first research question that we used to build a linear regression model was “What is the relationship between consumption of alcohol and life expectancy?”. We assumed that there would be a positive relationship before testing out the assumption since increased alcohol consumption is bad for your health. We created a linear regression model comparing the variables “life expectancy” and “alcohol” and calculated the summary statistics for our data. Based on our output, the intercept showed us that when the alcohol consumption is 0, life expectancy is 64.76 years. This statistic was interesting because the mean life expectancy from the whole dataset is 69.22 years, and we assumed that no alcohol would make life expectancy longer than the average. However, our assumption was proved by the output because we found that for each unit increase in alcohol consumption, life expectancy increases by 0.87925 years, holding all other factors are constant. The p values for both the intercept and alcohol are less than 0.05, which indicates a relationship between alcohol consumption and life expectancy and we can reject the null hypothesis. Also, both the intercept and slope have very small standard errors, indicating the coefficient estimates are precise. In terms of the residual standard error, the output shows that it is 8.054, which indicates that the regression model did not fit the dataset very well. The degree of freedom was 1647, which shows us that there were 1647 data points left over after fitting the model. The model's multiple r-squared is 0.1622, which means that the model explains 16.22% of the variance in life expectancy based on the independent variable of alcohol. We attempted to overfit the model using a polynomial, however, this was not helpful as our multiple r-squared values did not have much of an effect

on it. In Figure 1, we visualized the linear regression for life expectancy versus alcohol. As you can see from the visualization, the blue regression line has a positive slope, indicating that there is a positive correlation between alcohol consumption and life expectancy in this dataset. However, this does not necessarily imply causation, because the data points are widely scattered around the regression line. While the visualization shows a positive relationship between the variables, the wide scatter of points implies that there is a weaker relationship between the two, and other factors are more significant determinants of life expectancy. The next visualization shown in Figure 2 shows the distribution of residuals for life expectancy and alcohol. The distribution of residuals is widely spread and varies greatly from the regression line. This suggests that the model is not the best fit for this data and suggests heteroscedasticity from the uneven distribution. We also made a model to visualize the confidence intervals for our model, shown in Figures 3 and 4. We can see that in Figure 3, the confidence interval for the intercept, life expectancy, is between 64.7308102 and 65.9022146 which means that 95% of life expectancy years lie in this range. The same for the confidence interval for alcohol which is between 0.7826617 and 0.9758293, meaning that 95% of the alcohol consumption lies in this range. The confidence interval range for both life expectancy and alcohol does not contain a zero, which implies that we can reject the null hypothesis and there is a statistically significant relationship between alcohol consumption and life expectancy. In Figure 4, the graph shows the green dashed line which represents the 95% confidence level around the regression line. The confidence interval tells us that 95% of the model's predictions fall within this range and that the predictions are reasonably reliable. The model diagnostics for this model can give us more information on whether it fits the linear assumptions. We created a Q-Q plot, residuals vs. fitted plot, scale-location plot, and Cook's distance plot to test these assumptions. The Q-Q plot shows that the residuals follow the red line fairly closely, representing a normal distribution, which is an assumption of linear regression. The residuals vs. fitted plot shows that the fitted values (predicted values of the dependent variable, in this case, life expectancy), are randomly scattered around 0, which means that the model shows homoscedasticity, the constant variance of residuals which is another assumption of linear regression. The scale-location plot shows an almost horizontal line with equally spread points above and below it. This also indicates homoscedasticity. Our Cook's distance plot shows that none of the observation numbers has



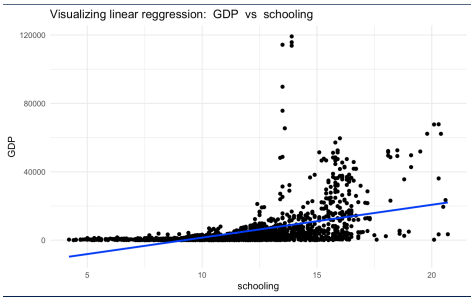


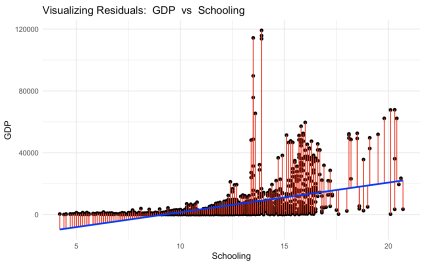


a Cook's distance of more than 1, so there are no significant observations in our data that influence the model.

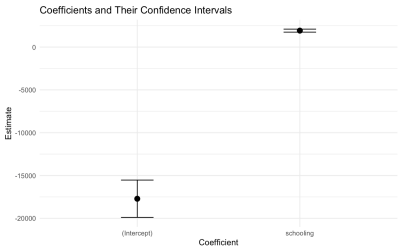
Research Question #2

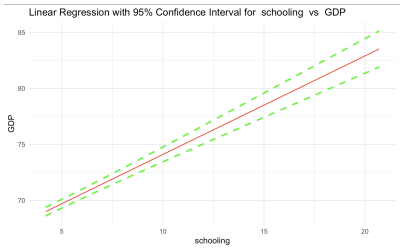
The second research question that we chose to build a linear regression model was, “What is the relationship between the number of years in schooling and GDP per capita?”. We wanted to find out if there was a relationship between the average number of years of schooling that a country’s population has and its gross domestic product. Before conducting the data analysis, we assumed that the more years of schooling that a country has, the higher its GDP would be, and the higher a country’s GDP, the more years of schooling it would have. In order to test this, we created a linear regression model comparing the variables “gdp” and “schooling”. We found that from our summary statistics, the intercept shows that when schooling is 0, GDP is -$17717.00 dollars. This indicates to us that when a country has no years of schooling, its GDP level is lower. The other statistic for schooling is 1921.1, meaning that for every unit increase in schooling, GDP increases by $1921.10. The p-values for both GDP and schooling are less than 0.05, which indicates a relationship between schooling and GDP, and we can reject the null hypothesis that there is no relationship between years of schooling and GDP. Both the intercept and slope have big standard errors, especially the intercept, which is 1111.9, indicating the coefficient estimates are not precise. In terms of the residual standard error, the output shows that it is 10140, which indicates to us that the residual spread is very large. The degree of freedom was 1647, which shows us that there were 1647 data points left over after fitting the model. The model's multiple r-squared is 0.219, which means that the model only explains 21.9% of the variance in GDP, meaning that there are other factors that could explain GDP better than years of schooling. In Figure 5, we visualized the relationship between GDP and schooling. This model shows us that while schooling and GDP have a positive linear relationship, the regression line is almost flat and there are a lot of observations that do not fall close to the regression line. We can also see that from the years of schooling from 0-10 years, there is little increase in the country’s GDP, and the GDP only increases by a significant amount after 10 years of schooling and on. The next visualization in Figure 6 shows the the distribution of residuals for GDP and years of schooling. The distribution of residuals is widely spread and varies greatly from the regression line. The residuals represent the difference between he actual data points shown in black, and the predicted values on the blue line. The vertical red lines connecting the data points to the blue line are the residuals. Many of the residuals towards the right side of the plot appear to be large, indicating that the linear model might not be the best representation for the data. In the years of schooling on the left side of the plot, the GDP values are relatively low, and the residuals are smaller. As the years of schooling increases, the GDP values become more spread out, and the residuals increase significantly, suggesting that GDP becomes more variable for higher education levels. In order to visualize the confidence intervals for our model, we created two plots





shown in Figures 7 and 8. In Figure 7, we can see that the confidence interval for the intercept ranges from -19897.975 to -15536.084. This means that we can be 95% confident that the true value of the intercept lies within this range. In this case, the intercept is GDP. Similarly, the confidence interval for schooling is between 1745.715 and 2096.409, indicating that we can be 95% confident that the true effect of schooling on GDP falls within this interval. Since the confidence intervals for both the intercept and schooling do not include zero, we can reject the null hypothesis. This suggests that there is a statistically significant relationship between schooling and GDP. Figure 8 visualizes the confidence interval in terms of the linear regression model for GDP and years of schooling. The plot shows the green dashed line which represents the 95% confidence level around the regression line. The confidence interval tells us that 95% of the model's predictions for the relationship between GDP and years of schooling fall within this range and that the predictions are reasonably reliable. The green dashed line is fairly narrow to the regression line, indicating that the estimates have a higher certainty. Similarly to our other research question, we created a Q-Q plot, residuals vs. fitted plot, scale-location plot, and Cook's distance plot to test these assumptions with model diagnostics. In our Q-Q plot, the residuals deviate from the Q-Q line, indicating a violation of normality and that the model does not meet the assumptions of linear regression. The residuals vs. fitted model shows heteroskedasticity because the red line is not horizontal and the residuals are not evenly scattered along the line. Heteroskedasticity refers to a situation where the variance of the residuals is unequal over a range of measured values. Since heteroskedasticity exists, GDP and schooling contain unequal variance, and the analysis results may be invalid. In the scale-location plot, the residuals are not spread equally across the range of fitted values. The red line should be horizontal with the points equally spread around it, and this does not. This also indicates heteroscedasticity, since the residuals are not equally distributed across the range of fitted values. In our Cook's distance plot, none of the observation numbers have a cook's distance of more than 1, so there are no significant observations in our data that is influencing the model.

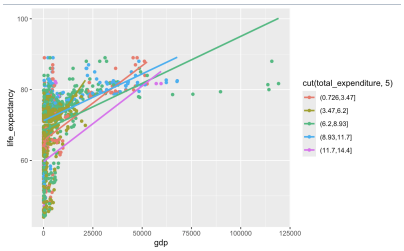




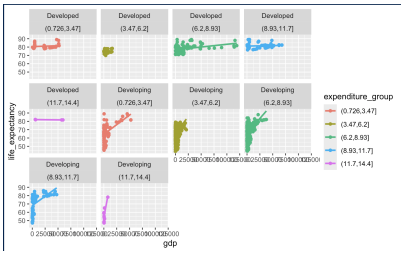
Research Question #3

Our final research question for our dataset is, "Does a country's GDP, years of schooling, status, or total expenditure affect life expectancy?". We were trying to find out if one fo these predictors had a significant effect on life expectancy and used a multiple regression to interpret our model. First, we created a multiple regression model to evaluate the summary statistics for our predictor variables. The baseline for our model was that the country has a status of "Developed". For the intercept, the model predicts that when all variables, GDP, schooling, and total expenditure are zero and status is Developed (baseline), life expectancy equals 45.67 years. Also, the intercept's p-value is less than 0.05, which indicates that this estimate is statistically significant. The second coefficient shows that for each unit increase in GDP, life expectancy increases by 8.576e-05 years, holding all other factors constant and the country's status is developed. The p-value is less than 5% indicates a statistically significant relationship between GDP and life expectancy and we can reject null hypothesis, which was that there is no

relationship between GDP and life expectancy. The coefficient “statusDeveloping” measured the difference in the intercept between "Developed" and "Developing", holding all other variables GDP, schooling, and total expenditure as zero. The output shows that developing countries have life expectancies that are about 1.47 years lower than developed countries. The p-value is 0.0041, which shows this estimate is statistically significant. The model also shows that for each additional year in schooling, life expectancy increases by 2.041e+00 years, holding all other factors constant and the country’s status is developed. The p-value is less than 0.05, indicating a statistically significant relationship between life expectancy and years of schooling. For total expenditure, the estimate is 5.722e-02, which means that the model predicts a decrease in life expectancy with an increase in total expenditure. However, the p-value of 0.3857 is higher than 0.05, which means that this effect is not statistically significant we can not reject null hypothesis and total expenditure is not significant in predicting life expectancy. The multiple r-squared statistic is 0.5449, which means that about 54.49% of the variance in life expectancy is explained by our predictor variables. This is a moderate estimate, showing us that there are other factors that could predict life expectancy. The residual standard error is 5.945, and since this is a measure of the variability of the residuals, it means that the residuals are not spread out in the model. Our p-value is less than 2.2e-16, meaning that the overall model is highly statistically significant. This shows that overall, the predictors have a statistically significant relationship with life expectancy. As we can see from these statistics, schooling has the most significant impact on life expectancy, and total expenditure is the least significant out of our predictors. In order to get a better picture of our predictor variables, we also tested for multicollinearity. First, we converted our “status” column to numeric values for better classification and to be able to use in within our function. We ran a correlation test to see if any of our predictors were correlated and would influence our results. From the correlation matrix, we can see that GDP and years of schooling are correlated as the correlation estimate is 0.4679470. This suggests that as GDP increases, years of schooling tends to increase. Because of this correlation, we decided to remove schooling from our next iteration of model testing. Also, the correlation coefficient of GDP and status is 0.4848010, which shows that there is a correlation that countries with a higher GDP tend to be under a developed status. Since these predictors also show correlation, we removed status from the next model as well. After removing schooling and status from our predictor variables, we ran another test to see how the summary statistics changed. According to our output, the multiple r-squared went down to 0.2041 compared to 0.5449 before removing removing the correlated factors. This means that the model explains 20.41% of the variance in life expectency, wich is not very high. We also used a regression model to analyze the interaction between GDP and total expenditure to predict life expectancy. The most important data that we found from this test was that the interaction term was -1.795e-05, which means that as GDP increases, the positive effect of total expenditure on life expectancy slightly decreases. This is statistically significant, as the p-value is less than 0.05. In terms of our predictors, increasing the total expenditure in countries with a higher GDP does not lead to as large an increase in life expectancy as it would in countries with lower GDP. We then ran an ANOVA test to evaluate our hypotheses and found that both total expenditure and GDP have significant effects on life expectancy. There is a significant interaction effect between total expenditure and GDP, meaning that the impact of total expenditure on life expectancy varies depending on the level of GDP. These results all suggest that we can reject the null hypothesis that GDP and Total Expenditure do not significantly affect life expectancy. In Figure 9, we visualized this interaction and divided the total



expenditure levels into five groups to better understand how the regression lines changed. As shown, all of the regression lines have positive slopes, indicating that as gdp increases, at all levels of total expenditure, life expectancy also increases. The total expenditure level of 6.2-8.93 has the longest regression line because it contains the observations with the highest GDP's. Most of the observations are clustered around the left side of the plot, showing that our observations mostly contain values of GDP from 0-25000. In Figure 10, we explored this relationship in terms of if a country was developed vs. developing. As shown in the plots, the observations on the developed plots all begin at a life expectancy above or equal to about 70 years of age, while the developing country plots show much lower life expectancy levels. In order to check our assumptions for non-linearity, we ran model diagnostics on our model for total expenditure, gdp, and life expectancy. We found that in our Q-Q plot, it shows that the residuals follow the red line fairly closely, representing a normal distribution, which is an assumption of linear regression. In the residuals vs. fitted plot, the red line is not flat, the red line is curved, which suggests that the relationship between the independent and dependent variables might not be well-captured by the model, meaning there could be non-linearity. The scale-location plot shows the red line is upward slanting and the residuals are clustered at one end of the plot and become more spread out as the plot moves along. The scale-location plot suggests heteroscedasticity because the red line is not flat but shows a distinct trend, indicating that the variance of the residuals is not constant across fitted values. In the Cook's distance plot, none of the observations reach a Cook's distance of 1, so there are no significant observations that are influencing the model. However, it is important to note the spike of observations at around 800, because this data could be affecting the model because of the volume of observation numbers.



Conclusion

In conclusion, this report depicts our findings using data mining techniques to better understand the "Life Expectancy (WHO)" dataset with three guiding research questions. The first research question "What is the relationship between consumption of alcohol and life expectancy?", showed us that there is a positive relationship between alcohol consumption and life expectancy and that the relationship is statistically significant. However, our model diagnostics showed us that the model could not be the best fit in predicting life expectancy because of the spread of residuals and low r-squared value. For our second research question, "What is the relationship between the number of years in schooling and GDP per capita?", the linear regression model showed a statistically significant relationship between years of schooling and GDP per capita. As the number of years in schooling increased, the GDP per capita also increased. Similarly to our first question, we found that the presence of heteroscedasticity suggested that schooling is not the strongest predictor of GDP. The third research question, "Does a country's GDP, years of schooling, status, or total expenditure affect life expectancy?", was altered to "Does a country's GDP or total expenditure affect life expectancy?" after finding correlation between some of the predictor variables. We found that both GDP and total expenditure have a statistically significant effect on life expectancy, and that as GDP increases, the positive effect of total expenditure on life expectancy slightly decreases. This report gave us knowledge of important factors that influence the variables of life expectancy, GDP, years of schooling, alcohol consumption, and total expenditure. In the future, we would further test more predictors for life expectancy to find more predictable models.