The World Health Organization: Factors that Influence Life Expectancy

Lily Geiser, Hannah Ashburn, Angel Baeza

2023-12-16

I. Introduction

The World Health Organization (WHO) is a specialized agency under the United Nations, its parent organization. Established on April 7, 1948, the WHO merged the assets, personnel, and duties of the League of Nations Health Organization with Paris's Office International d'Hygiène Publique, which included the International Classification of Diseases (ICD).

While their primary objective is to help ensure the global population attains the highest level of health possible, the WHO has also been actively involved in gathering a wide range of data from member countries. This data, most of which is publicaly available, has been used to conduct a variety of research projects in hopes of solving the world's most pressing global health challenges. The metrics collected by the WHO's Global Health Observatory (GHO) include a broad range of health indicators such as mortality, disease prevalence, health system performance, and other social determinants of health. They also collect more qualitative data through the World Health Survey, which collects voluntarily submitted insights into people's perceived access to healthcare, health-related behaviors, and other personal information that would not be available otherwise.

By analyzing this data, not only can we add to a growing body of research, we can contribute our findings towards creating a better understanding of health trends and disparities across the globe. Furthermore, any remarkable trends identified in this data could have significant implications on policy development, resource allocation, and disease prevention and control– all of which are vastly important areas. Such findings could bolster and facilitate efforts to increase the global population's quality of life and overall life expectancy, generating a significant impact on society as we know it.

PART I: Predicting Life Expectancy

II. Model Creation and Analysis: Full Model

The first model we created was a multiple regression model using data from only the year 2015. Our goal in creating this model is to see how life expectancy is correlated with the other variables in our dataset. We used data from only 2015 so that the different life expectancies reported varied by country alone, not by country and year. Further, this model could then be applied to other years to verify that its accuracy; if this model is not accurate from year to year, this could imply to us that there are other confounding variables not included in the data, or that the explanatory factors could have varying impacts by year.

The full model we created it shown below. As discussed, the dependent variable is life expectancy; all other variables in the dataset are used, excluding country (as each life expectancy comes from a different country, therefore this variable could predict the life expectancy with 100% accuracy), year (as all data is from 2015), and Economy_status_Developing (as this is simply the reverse of Economy_status_Developed, which is included). Our null hypothesis associated with this model is that none of the explanatory variables have any correlation with life expectancy (i.e. that their coefficients equal 0). Our alternative hypothesis is that at least one of the explanatory variables has a coefficient that does not equal zero. For this model and all models used throughout this report, we will be using a significance level of 0.05.

```r
#importing data and creating 2015 dataset
data <- read.csv("cleandata.csv")
data.2015 <- data[data$Year == 2015, ]

#creating regression model and outputting summary data
reg_2015 <- lm(Life_expectancy~as.factor(Region)+Infant_deaths+Under_five_deaths+
                 Adult_mortality+Alcohol_consumption+Hepatitis_B+Measles+BMI+Polio+
                 Diphtheria+Incidents_HIV+GDP_per_capita+Population_mln+
                 Thinness_ten_nineteen_years+Thinness_five_nine_years+Schooling+
                 as.factor(Economy_status_Developed), data=data.2015)
print(summary(reg_2015))
```

```
##
## Call:
## lm(formula = Life_expectancy ~ as.factor(Region) + Infant_deaths +
##     Under_five_deaths + Adult_mortality + Alcohol_consumption +
##     Hepatitis_B + Measles + BMI + Polio + Diphtheria + Incidents_HIV +
##     GDP_per_capita + Population_mln + Thinness_ten_nineteen_years +
##     Thinness_five_nine_years + Schooling + as.factor(Economy_status_Developed),
##     data = data.2015)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9233 -0.8085  0.0111  0.6475  3.6337
##
## Coefficients:
##                                              Estimate Std. Error t value
## (Intercept)                                 8.428e+01  2.365e+00  35.632
## as.factor(Region)Asia                       2.464e-01  4.255e-01   0.579
## as.factor(Region)Central America and Caribbean  1.799e+00  4.834e-01   3.722
## as.factor(Region)European Union             -7.480e-01  6.876e-01  -1.088
## as.factor(Region)Middle East                1.204e-01  5.340e-01   0.226
## as.factor(Region)North America              -1.769e-01  9.619e-01  -0.184
## as.factor(Region)Oceania                    -1.078e+00  5.891e-01  -1.831
## as.factor(Region)Rest of Europe             1.407e-01  5.500e-01   0.256
## as.factor(Region)South America              1.849e+00  5.292e-01   3.494
## Infant_deaths                               -3.149e-02  3.852e-02  -0.817
## Under_five_deaths                           -6.452e-02  2.631e-02  -2.452
## Adult_mortality                             -4.996e-02  2.992e-03 -16.698
## Alcohol_consumption                         -1.137e-02  4.914e-02  -0.231
## Hepatitis_B                                 -2.112e-02  2.444e-02  -0.864
## Measles                                     1.168e-02  8.017e-03   1.456
## BMI                                         -1.414e-01  8.248e-02  -1.715
## Polio                                       -4.576e-03  2.242e-02  -0.204
## Diphtheria                                  1.564e-02  2.792e-02   0.560
## Incidents_HIV                               2.071e-01  9.027e-02   2.294
## GDP_per_capita                              2.476e-05  8.902e-06   2.782
## Population_mln                              -1.112e-04  7.092e-04  -0.157
## Thinness_ten_nineteen_years                 -1.251e-01  1.127e-01  -1.109
## Thinness_five_nine_years                    1.078e-01  1.124e-01   0.959
## Schooling                                   7.727e-02  7.408e-02   1.043
## as.factor(Economy_status_Developed)1        2.679e+00  6.394e-01   4.189
##                                             Pr(>|t|)
```

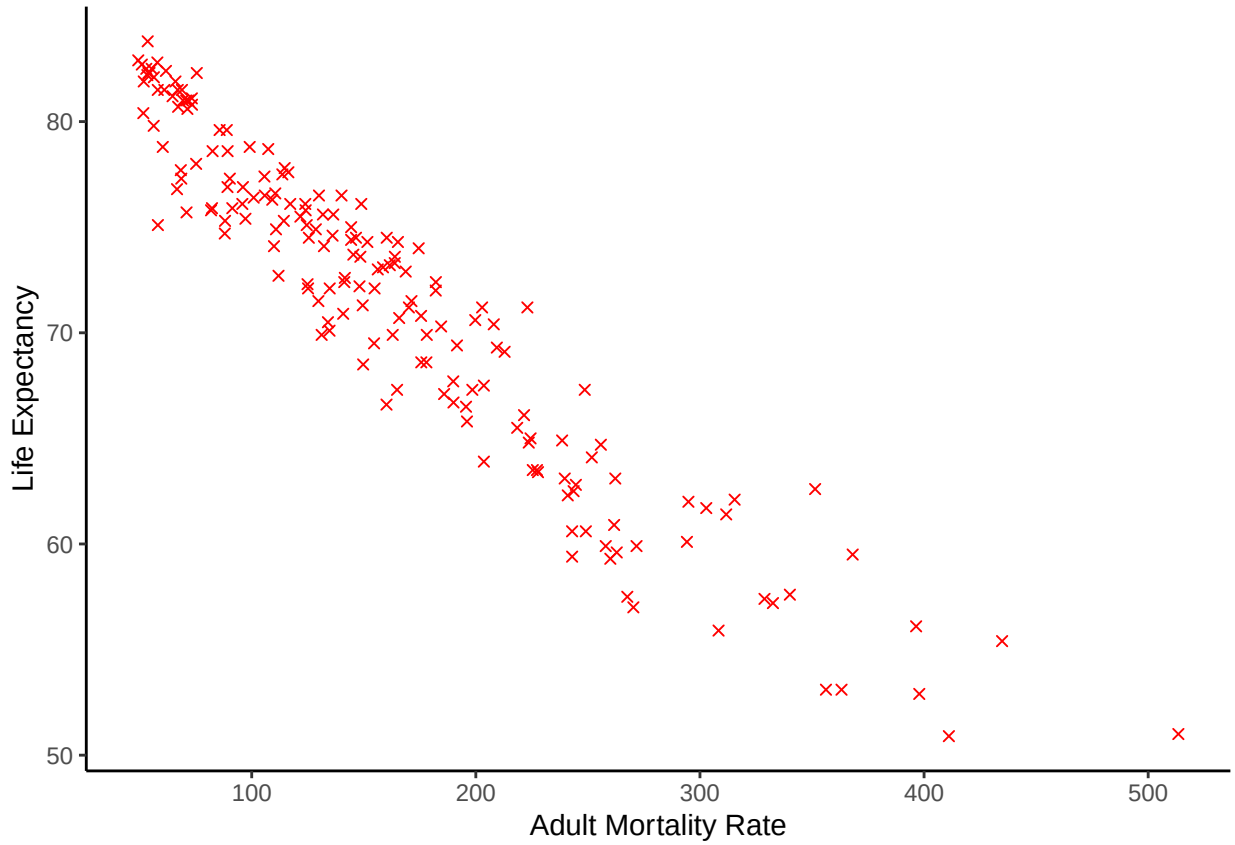
```
## (Intercept)                                      < 2e-16 ***
## as.factor(Region)Asia                           0.563373
## as.factor(Region)Central America and Caribbean 0.000277 ***
## as.factor(Region)European Union                 0.278369
## as.factor(Region)Middle East                    0.821848
## as.factor(Region)North America                  0.854297
## as.factor(Region)Oceania                        0.069058 .
## as.factor(Region)Rest of Europe                 0.798480
## as.factor(Region)South America                  0.000621 ***
## Infant_deaths                                   0.414921
## Under_five_deaths                               0.015324 *
## Adult_mortality                                  < 2e-16 ***
## Alcohol_consumption                             0.817268
## Hepatitis_B                                     0.388847
## Measles                                         0.147345
## BMI                                             0.088415 .
## Polio                                           0.838578
## Diphtheria                                      0.576235
## Incidents_HIV                                   0.023123 *
## GDP_per_capita                                  0.006083 **
## Population_mln                                  0.875556
## Thinness_ten_nineteen_years                     0.268963
## Thinness_five_nine_years                        0.338990
## Schooling                                       0.298511
## as.factor(Economy_status_Developed)1            4.69e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.247 on 154 degrees of freedom
## Multiple R-squared:  0.9781, Adjusted R-squared:  0.9746
## F-statistic: 286.1 on 24 and 154 DF,  p-value: < 2.2e-16
```

There are several insights we can obtain from the above model. Firstly, we can say that this model appears to be a good fit and does an excellent job of predicting life expectancy. We can observe that the p-value associated with the overall model is extremely close to 0, at less than 2.2e-16. This tells us that we can certainly reject our null hypothesis and accept the alternative hypothesis. Further, we can see that the multiple R-squared value is 0.9781; this is a value incredibly close to 1, which tells us that nearly all of the variation in life expectancy can be explained by the variation in the explanatory variables.

If we look at the coefficients associated with the variables, we will observe that multiple of them are statistically significant; specifically, the regions Central America and the Caribbean, and South America; the under five mortality rate; adult mortality; HIV incidence; GDP per capita; and whether the country is classified as Developed or Developing all have statistically significant coefficients. The coefficient associated with the intercept is also significant. This tells us there is evidence that these variables specifically have nonzero coefficents; there is not sufficient evidence to come to this conclusion for the other variables, as they all have associated p-values of greater than 0.05.

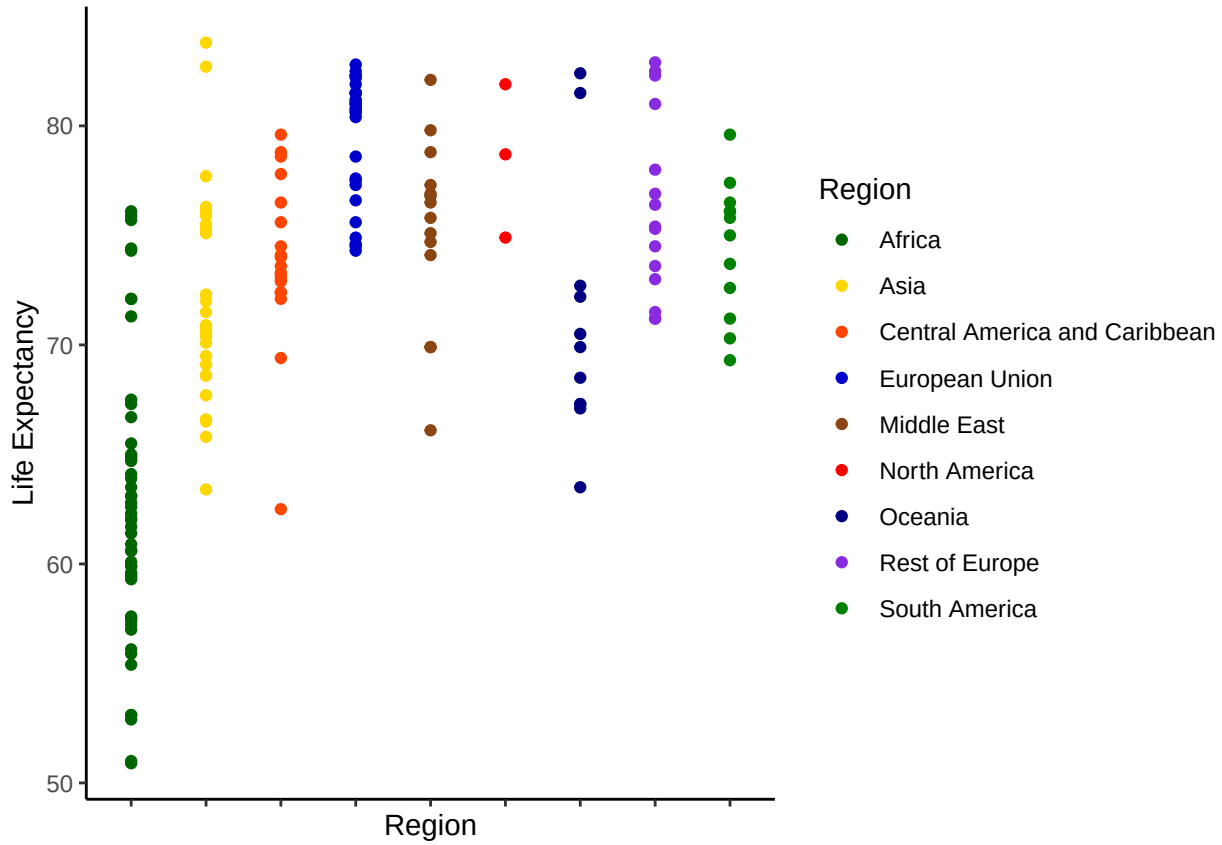
The significance of these variables can be corroborated by observing some of their marginal scatterplots. For instance, adult mortality appears to have a very clear linear relationship with life expectancy:

```
library(ggplot2)
ggplot(data.2015, aes(x = Adult_mortality, y = Life_expectancy)) +
  geom_point(size=1.5, color = "red", shape = 4) + theme_classic() +
  labs(x = "Adult Mortality Rate", y = "Life Expectancy")
```



We can also observe regional differences with the graph below:

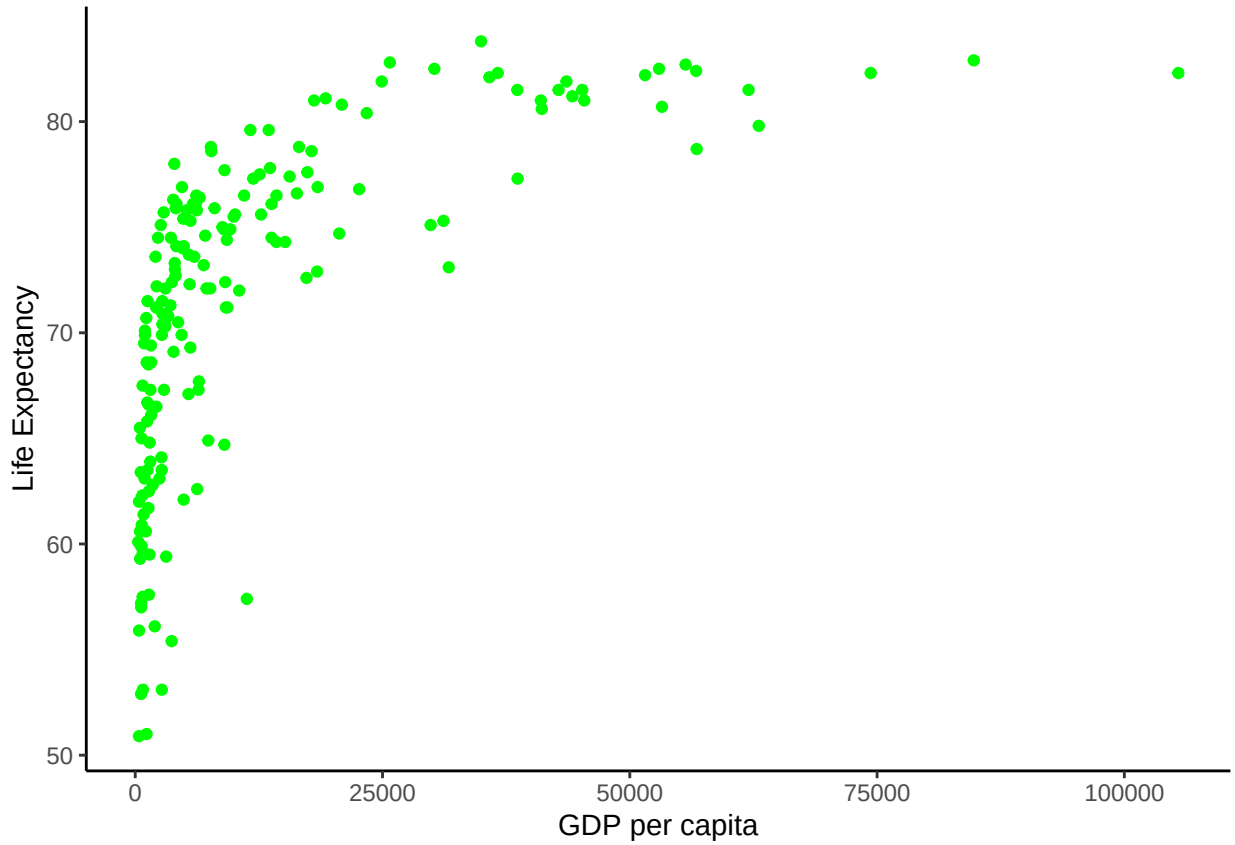
```
ggplot(data.2015, aes(x = Region, y = Life_expectancy, color = Region)) +
  geom_point() + theme_classic() + labs(x = "Region", y = "Life Expectancy") +
  theme(axis.text.x = element_blank()) +
  scale_color_manual(values = c("Africa" = "#006400", "Asia" = "#FFD700", "Central America and Caribbean
```



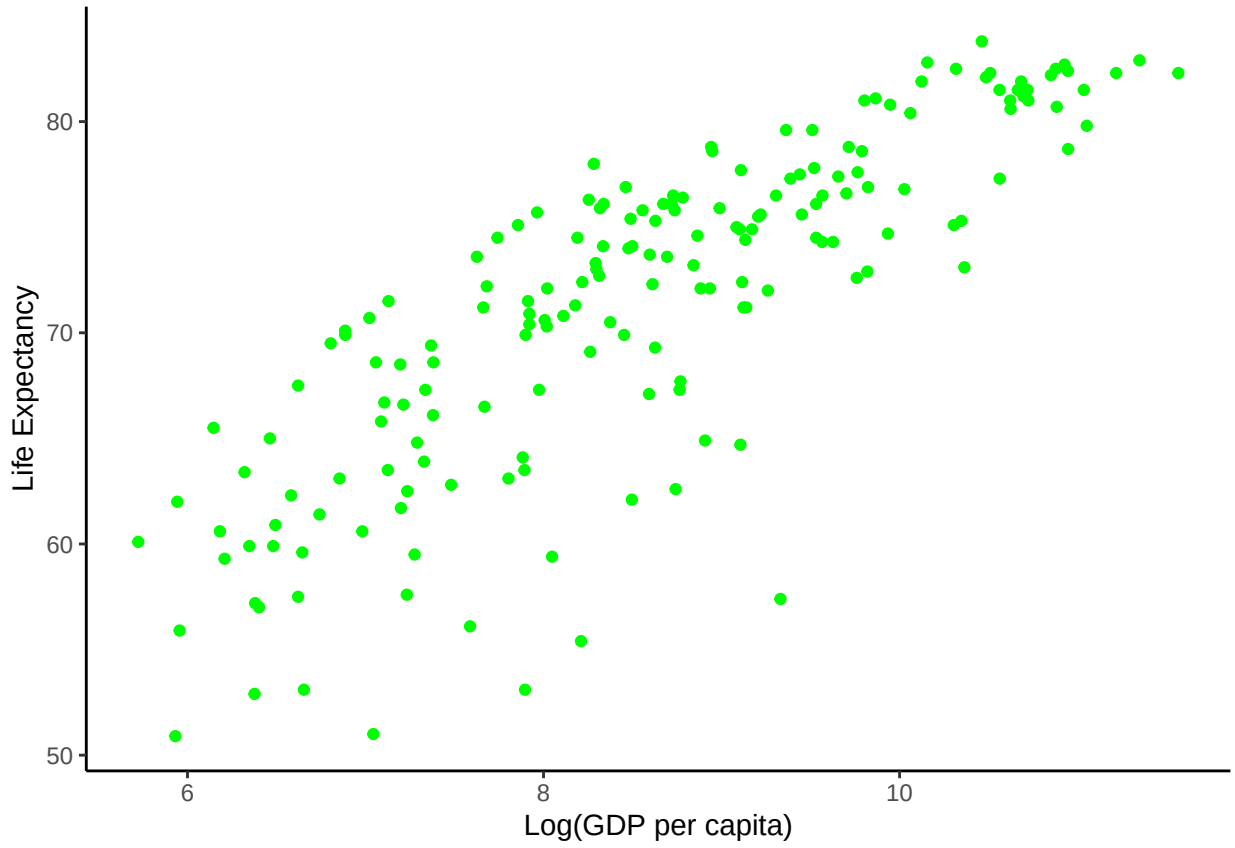
Interestingly, although it seems clear that South America would have substantially different life expectancies than other countries, Central America and the Caribbean do not appear to be as substantially different than our model suggested. This tells us that there is likely another confounding factor going on here (for instance, there may be some multicollinearity).

Additionally, by graphing GDP v. life expectancy, we also see a slightly different trend than what our model predicts. As shown below, it appears to have a logarithmic relationship with life expectancy rather than a linear one. When we graph log(GDP) rather than simply GDP, we observe this more linear relationship:

```
ggplot(data.2015, aes(x = GDP_per_capita, y = Life_expectancy)) +
  geom_point(size=1.5, color = "green") + theme_classic() +
  labs(x = "GDP per capita", y = "Life Expectancy")
```




```
ggplot(data.2015, aes(x = log(GDP_per_capita), y = Life_expectancy)) +
  geom_point(size=1.5, color = "green") + theme_classic() +
  labs(x = "Log(GDP per capita)", y = "Life Expectancy")
```



III. Explanatory Variable Correlation and a Reduced Model

While our model above appears to be an incredibly good fit, we wanted to see if this model could be improved upon with a reduced model. Firstly, we wanted to see whether any of the variables above were highly correlated. To accomplish this, we first obtained the associated VIF values:

```
library(car)
```

```
## Loading required package: carData
```

```
print(vif(reg_2015))
```

```
##                                GVIF Df GVIF^(1/(2*Df))
## as.factor(Region)         78.924227  8        1.313940
## Infant_deaths             78.408752  1        8.854872
## Under_five_deaths         82.240022  1        9.068628
## Adult_mortality            8.291875  1        2.879562
## Alcohol_consumption        3.872895  1        1.967967
## Hepatitis_B               13.718816  1        3.703892
## Measles                    1.927141  1        1.388215
## BMI                        3.739444  1        1.933764
## Polio                      9.762729  1        3.124537
## Diphtheria                19.264854  1        4.389175
## Incidents_HIV              2.451655  1        1.565776
```

```
## GDP_per_capita                       2.847603  1        1.687484
## Population_mln                       1.235429  1        1.111499
## Thinness_ten_nineteen_years         24.635592  1        4.963425
## Thinness_five_nine_years            25.470867  1        5.046867
## Schooling                            6.219399  1        2.493872
## as.factor(Economy_status_Developed)  7.715681  1        2.777712
```

As we can see, some of these variables have incredibly high VIF values; specifically, Region, Infant_deaths, Under_five_deaths, Hepatitis_B, Diphtheria, Thinness_ten_nineteen_years, and Thinness_five_nine_years all have GVIF values above 10. It is important to note that the adjusted GVIF values found in the righthand column are signficantly lower (for example, due to the multiple degrees of freedom in the Region variable, the adjusted GVIF is substantially lower). However, our goal in creating the reduced model was to reduce multicollinearity as much as possible.

Our next step was to determine which specific variables are closely related to each other. The code to accomplish this is below; the output of this code is which cells of the covariance matrix have values greater than 0.8, and the associated variables. Categorical variables were excluded from the generated covariance matrix.

```r
#create covariance matrix
matrix <- cor(data.2015[, sapply(data.2015, is.numeric)])
```

```
## Warning in cor(data.2015[, sapply(data.2015, is.numeric)]): the standard
## deviation is zero
```

```r
#find high correlation values and their corresponding row and column names
high_correlation_indices <- which(matrix > 0.8 & matrix < 1, arr.ind = TRUE)
high_correlation_values <- matrix[high_correlation_indices]
row_names <- rownames(matrix)[high_correlation_indices[, 1]]
col_names <- colnames(matrix)[high_correlation_indices[, 2]]

#export information to a datafram and print in order of correlation
result_df <- data.frame(
  Row = row_names,
  Column = col_names,
  Correlation = high_correlation_values
)
result_df <- result_df[order(-result_df$Correlation), ]
print(result_df)
```

```
##                            Row                    Column Correlation
## 1            Under_five_deaths             Infant_deaths   0.9899859
## 3                Infant_deaths         Under_five_deaths   0.9899859
## 13    Thinness_five_nine_years Thinness_ten_nineteen_years   0.9731355
## 14 Thinness_ten_nineteen_years    Thinness_five_nine_years   0.9731355
## 8                   Diphtheria               Hepatitis_B   0.9438458
## 11                Hepatitis_B                Diphtheria   0.9438458
## 10                  Diphtheria                     Polio   0.9296290
## 12                       Polio                Diphtheria   0.9296290
## 7                        Polio               Hepatitis_B   0.8886645
## 9                 Hepatitis_B                     Polio   0.8886645
## 2              Adult_mortality             Infant_deaths   0.8537298
## 5                Infant_deaths          Adult_mortality   0.8537298
```

```
## 4              Adult_mortality           Under_five_deaths   0.8422934
## 6            Under_five_deaths             Adult_mortality   0.8422934
```

As we can see above, there are multiple variables which are very strongly correlated. For example, Under_five_deaths and Infant_deaths have a near perfect correlation, with a covariance value of nearly one.

Based on the above information, we created the reduced model coded below:

```
reg_2015_red <- lm(Life_expectancy~Adult_mortality
                   +Alcohol_consumption+Measles+BMI+
                   Diphtheria+Incidents_HIV+GDP_per_capita+Population_mln+
                   Thinness_ten_nineteen_years+Schooling+
                   as.factor(Economy_status_Developed), data=data.2015)
print(summary(reg_2015_red))
```

```
##
## Call:
## lm(formula = Life_expectancy ~ Adult_mortality + Alcohol_consumption +
##     Measles + BMI + Diphtheria + Incidents_HIV + GDP_per_capita +
##     Population_mln + Thinness_ten_nineteen_years + Schooling +
##     as.factor(Economy_status_Developed), data = data.2015)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.646 -1.098 -0.022  1.058  5.239
##
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          7.816e+01  2.691e+00  29.044  < 2e-16 ***
## Adult_mortality                     -7.208e-02  2.898e-03 -24.868  < 2e-16 ***
## Alcohol_consumption                  1.450e-01  5.478e-02   2.647 0.008904 **
## Measles                              7.219e-03  1.064e-02   0.678 0.498413
## BMI                                 -6.087e-02  8.871e-02  -0.686 0.493592
## Diphtheria                           2.746e-02  1.105e-02   2.485 0.013938 *
## Incidents_HIV                        3.777e-01  1.090e-01   3.465 0.000673 ***
## GDP_per_capita                       1.505e-05  1.103e-05   1.365 0.174239
## Population_mln                       1.448e-04  9.423e-04   0.154 0.878047
## Thinness_ten_nineteen_years         -7.239e-02  4.377e-02  -1.654 0.100011
## Schooling                            3.443e-01  8.269e-02   4.164 4.99e-05 ***
## as.factor(Economy_status_Developed)1 3.622e-02  5.487e-01   0.066 0.947452
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.712 on 167 degrees of freedom
## Multiple R-squared:  0.9552, Adjusted R-squared:  0.9522
## F-statistic: 323.4 on 11 and 167 DF,  p-value: < 2.2e-16
```

We can see based on the above that this model is still an incredibly good fit; the multiple R-squared value is still extremely close to 1 at 0.9552, and the p-value associated with the overall model is still extremely small. This model also has a significantly reduced amount of multicollinearity, as shown below:

```
library(car)
print(vif(reg_2015_red))
```

```
##                       Adult_mortality                 Alcohol_consumption
##                              4.127313                            2.552720
##                               Measles                                 BMI
##                              1.800868                            2.295263
##                             Diphtheria                       Incidents_HIV
##                              1.600795                            1.896535
##                          GDP_per_capita                      Population_mln
##                              2.320639                            1.157087
##          Thinness_ten_nineteen_years                          Schooling
##                              1.970545                            4.111272
## as.factor(Economy_status_Developed)
##                              3.014722
```

All of the VIF values above are quite low, demonstrating that we have managed to reduce collinearity.

This reduced model also provides some different insights into the data. For example, alcohol consumption, Diphtheria, and Schooling now all have significant p-values associated with their coefficients. This could either be because they truly do have a significant relationship with life expectancy, which was masked in the full model, or because we removed too many variables and they are confounding the information we have above.

To determine whether we should use the full model or reduced model, we will perform an ANOVA test:

```
anova(reg_2015_red, reg_2015, test = "F")
```

```
## Analysis of Variance Table
##
## Model 1: Life_expectancy ~ Adult_mortality + Alcohol_consumption + Measles +
##     BMI + Diphtheria + Incidents_HIV + GDP_per_capita + Population_mln +
##     Thinness_ten_nineteen_years + Schooling + as.factor(Economy_status_Developed)
## Model 2: Life_expectancy ~ as.factor(Region) + Infant_deaths + Under_five_deaths +
##     Adult_mortality + Alcohol_consumption + Hepatitis_B + Measles +
##     BMI + Polio + Diphtheria + Incidents_HIV + GDP_per_capita +
##     Population_mln + Thinness_ten_nineteen_years + Thinness_five_nine_years +
##     Schooling + as.factor(Economy_status_Developed)
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    167 489.59
## 2    154 239.52 13    250.07 12.368 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The output above tells us that, although we had initially seen that the reduced model is still a pretty good fit, our full model is in fact substantially better than the reduced model. The p-value associated with this ANOVA test is $<$ 2.2e-16, which tells us there is a statistically significant difference in the performance of the two models. Further, we can see that the RSS associated with the reduced model is over twice the value of the full model (489.59 v. 239.52). Despite the good fit of the reduced model, we can therefore conclude based on this that the full model should give us more accurate results.

IV. Testing our Model on 2014 Data

To further confirm that our full model is more accurate than our reduced model, we are going to use these two models to predict the life expectancy for the different countries in 2014, and see which yields more accurate results. The code for this trial can be found below:

```r
#creating 2014 data
data.2014 <- data[data$Year == 2014, ]

#predicting 2014 life expectancy with different models
predict_full <- predict(reg_2015, data.2014)
predict_red <- predict(reg_2015_red, data.2014)

#find MSE for both models
mse_full <- mean((predict_full - data.2014$Life_expectancy)^2)
mse_red <- mean((predict_red - data.2014$Life_expectancy)^2)
print(mse_full)
```

```
## [1] 1.354277
```

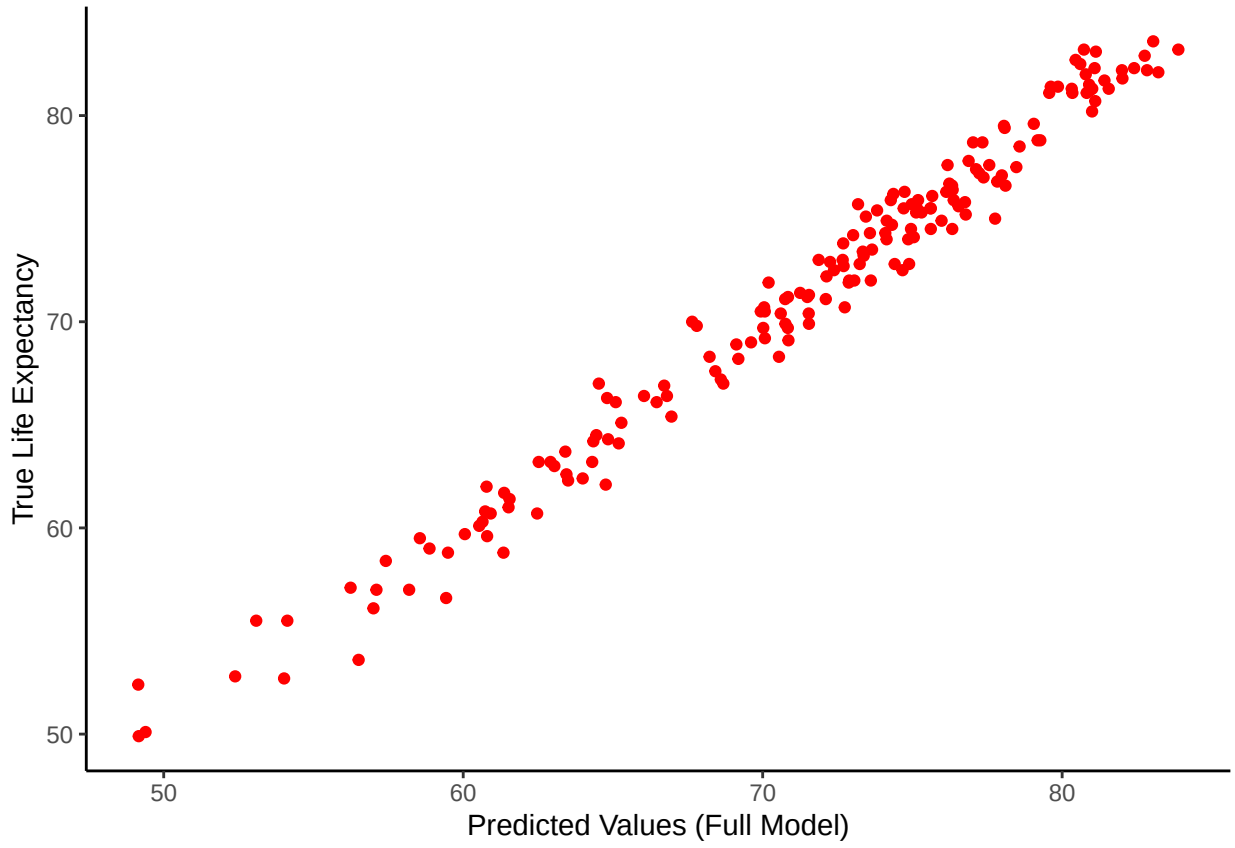
```r
print(mse_red)
```

```
## [1] 2.834837
```

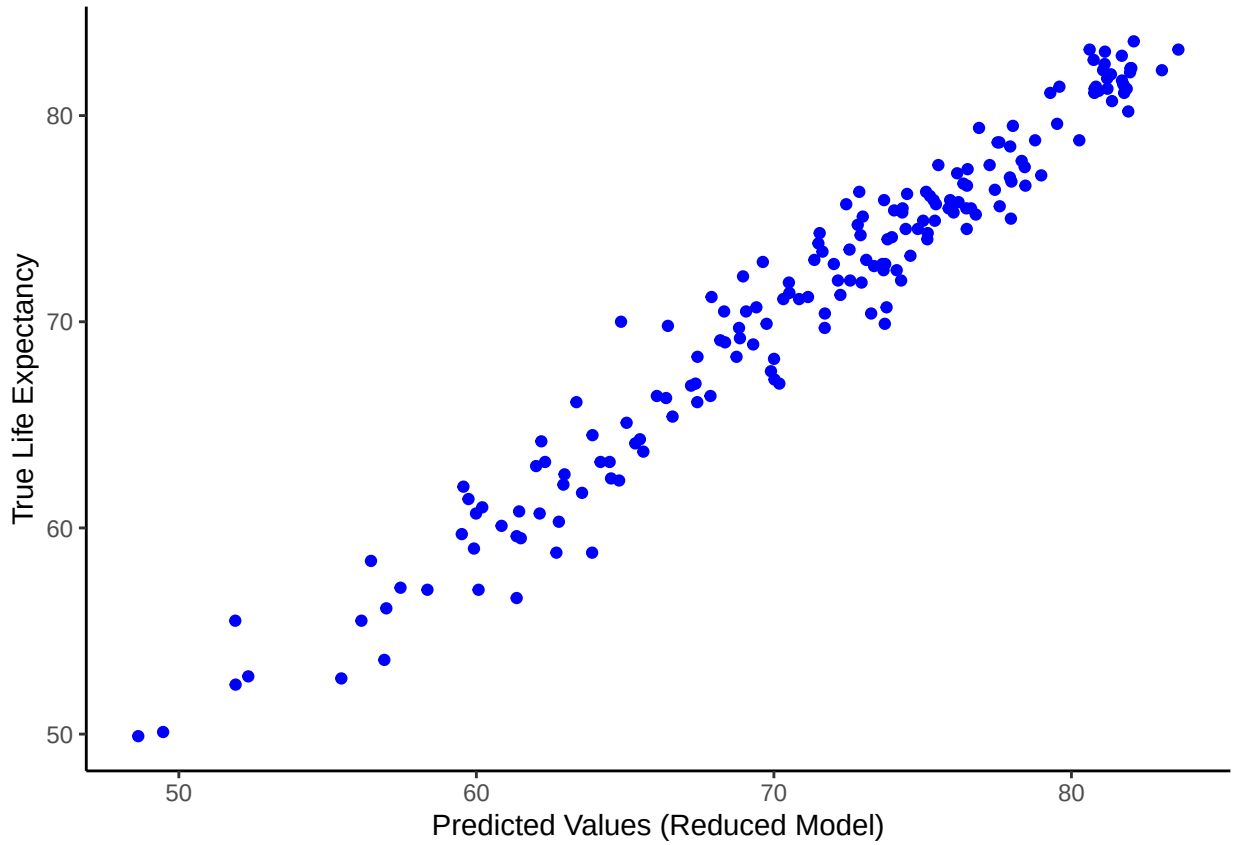
We can see based on the above that our test MSE for the full model is approximately 1.35, while the test MSE for the reduced model is about 2.83. As the full model produces a lower MSE, we can once again conclude that the full model produces a more accurate prediction of life expectancy than the reduced model.

Despite this, both of these MSEs are quite low, considering the range of the data and the number of data points we are looking at. As we can see by the below graphs, the predicted values for both line up extremely closely with the true life expectancies:

```r
#creating dataframes for graphs
pred_vals_full <- data.frame(y = data.2014$Life_expectancy, x = predict_full)
pred_vals_red <- data.frame(y = data.2014$Life_expectancy, x = predict_red)

ggplot(pred_vals_full, aes(x = x, y = y)) +
  geom_point(size=1.5, color = "red") + theme_classic() +
  labs(x = "Predicted Values (Full Model)", y = "True Life Expectancy")
```




```
ggplot(pred_vals_red, aes(x = x, y = y)) +
  geom_point(size=1.5, color = "blue") + theme_classic() +
  labs(x = "Predicted Values (Reduced Model)", y = "True Life Expectancy")
```



PART II: A Closer Look at Our Explanatory Variables

V. Testing Individual Diseases

Out of curiosity, we wanted to check if life expectancy had higher correlation with the disease variables, something we all presume to have a great impact on one's health. Similarly to our previous models, the dependent variable is life expectancy and our predictor variables are Hepatitis_B, Measles, Diphtheria, Polio, and Incidents_HIV. Using 2015 as the year, this was the model we created:

```
reg_dis <- lm(Life_expectancy~Hepatitis_B+Measles+Polio+Diphtheria+Incidents_HIV,
              data=data.2015)
print(summary(reg_dis))
```

```
##
## Call:
## lm(formula = Life_expectancy ~ Hepatitis_B + Measles + Polio +
##     Diphtheria + Incidents_HIV, data = data.2015)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.4288  -3.3484  -0.0704   3.9339  11.9551
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   41.83976    2.79276  14.982  < 2e-16 ***
## Hepatitis_B   -0.29104    0.08274  -3.517 0.000557 ***
## Measles        0.14915    0.02875   5.188 5.91e-07 ***
```

```
## Polio          0.31633    0.08188   3.863 0.000158 ***
## Diphtheria     0.18237    0.09893   1.843 0.066978 .
## Incidents_HIV -1.55190    0.24182  -6.418 1.28e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.101 on 173 degrees of freedom
## Multiple R-squared:  0.5877, Adjusted R-squared:  0.5758
## F-statistic: 49.32 on 5 and 173 DF,  p-value: < 2.2e-16
```

Looking at the results of the model, we can see that the model itself isn't the absolute best fit compared to our previous models, with the multiple R-squared 0.5877, however, the p-value is significantly low, at $<$ 2.2e-16. One problem that does from this model is the amount of multicollinearity with some variables, as seen below:

```
##   Hepatitis_B       Measles         Polio    Diphtheria Incidents_HIV
##      9.399759      1.481452      7.779646     14.455149      1.051429
```

It seems as if Diphtheria is the most correlated with another variable in the model (with a VIF of 14.455149) and Hepatitis_B being the second (with a VIF of 9.399759). We attempted at making a reduced model riddening the two diseases with the highest VIFs, trying to eliminate the multicollinearity, producing this new one:

```
##
## Call:
## lm(formula = Life_expectancy ~ Measles + Polio + Incidents_HIV,
##     data = data.2015)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.5835  -3.6143  -0.1134   4.3366  11.2113
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   40.62808    2.82127  14.401  < 2e-16 ***
## Measles        0.15223    0.02960   5.143 7.19e-07 ***
## Polio          0.22242    0.03648   6.096 6.75e-09 ***
## Incidents_HIV -1.65444    0.24763  -6.681 3.04e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.261 on 175 degrees of freedom
## Multiple R-squared:  0.5564, Adjusted R-squared:  0.5488
## F-statistic: 73.18 on 3 and 175 DF,  p-value: < 2.2e-16

##       Measles         Polio Incidents_HIV
##      1.476316      1.452278      1.036742
```

This model produces a model that has less multicollinearity, however, the fitness is reduced, having multiple R-squared of 0.5564, 0.0313 lower than the first disease model. Although, we did have lower VIFs for each of the variables that we did keep.

From this curiosity attempt at making sense of the impact of disease on life expectancy, we can take that diseases, themselves, are not the sole predictor variables for life expectancy, and that we get a better fit

model when we include other predictor variables, such as BMI and Alcohol_consumption. We can also potentially assume that, although some diseases are highly correlated with others, they still are important in creating a model that looks at the impact of diseases on life expectancy.

VI. Effects of Alcohol Consumption by Region

Though the results from the models previously discussed provide general insight into the many factors that influence life expectancy, the regional organization of the data set further piqued our interest, prompting us to investigate potential differences in the way that alcohol consumption impacts population longevity between regions. Alcohol has long been an integral part of social rituals and traditions, with its patterns of consumption intricately woven into the fabric of diverse societies. The frequency, context, and method by which alcohol is consumed are heavily influenced by cultural practices, physical location, and historical legacies. Since it is well known that different cultures have different practices regarding, and attitudes towards alcohol, we questioned whether the relationship between alcohol consumption and life expectancy differs significantly across geographic regions.

While numerous studies have examined the health implications of alcohol consumption, few have undertaken a comprehensive analysis that considers the influence of culture and location. While this area of research is not the focus of this project, it represents a field of potential interest that could be explored further. Here, the premise is that the impact of alcohol on life expectancy is not uniform and is significantly shaped by the unique cultural and geographic contexts in which it occurs. While our data set categorizes countries into one of nine geographical regions, it does not contain data regarding each country's unique drinking culture– only a data point quantifying "alcohol consumption". As such, this section only aims to highlight instances where the impact of alcohol consumption on life expectancy (and adult mortality) is positive for some regions whilst being negative for others.

The two models used in this section are single regression models using data from the fifteen year period between 2000 and 2015. Our goal in creating these models is to illustrate the regional differences in how life expectancy is correlated with life expectancy and adult mortality.

The first model we created is shown below. As discussed, the dependent variable is life expectancy, the independent variable is alcohol consumption, and the significance level is 0.05.

```
#importing data and creating regional datasets

data <- read.csv("Life-Expectancy-Data-Updated.csv")
unique(data$Region)
```

```
## [1] "Middle East"                   "European Union"
## [3] "Asia"                          "South America"
## [5] "Central America and Caribbean" "Rest of Europe"
## [7] "Africa"                        "Oceania"
## [9] "North America"
```

```
data.Asia <- data[data$Region == "Asia", ]
data.Rest_of_Europe <- data[data$Region == "Rest of Europe", ]
data.Africa <- data[data$Region == "Africa", ]
data.SA <- data[data$Region == "South America", ]
data.Cen_America_Caribbean <- data[data$Region == "Central America and Caribbean", ]
data.Oceania <- data[data$Region == "Oceania", ]
data.EU <- data[data$Region == "European Union", ]
data.ME <- data[data$Region == "Middle East", ]
data.NA <- data[data$Region == "North America", ]
```

```r
#creating single regression models

Africa_results <- lm(Life_expectancy~Alcohol_consumption, data = data.Africa)
Rest_Europe_results <- lm(Life_expectancy~Alcohol_consumption, data = data.Rest_of_Europe)
Asia_results <- lm(Life_expectancy~Alcohol_consumption, data = data.Asia)
SA_results <- lm(Life_expectancy~Alcohol_consumption, data = data.SA)
CenA_Car_results <- lm(Life_expectancy~Alcohol_consumption, data = data.Cen_America_Caribbean)
Oceania_results <- lm(Life_expectancy~Alcohol_consumption, data = data.Oceania)
EU_results <- lm(Life_expectancy~Alcohol_consumption, data = data.EU)
ME_results <- lm(Life_expectancy~Alcohol_consumption, data = data.ME)
NA_results <- lm(Life_expectancy~Alcohol_consumption, data = data.NA)

#outputting summary data
summary(Africa_results)
```

```
##
## Call:
## lm(formula = Life_expectancy ~ Alcohol_consumption, data = data.Africa)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.0938  -5.7545  -0.6313   3.8739  19.1679
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         59.21686    0.40533 146.096  < 2e-16 ***
## Alcohol_consumption -0.45948    0.09792  -4.692 3.17e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.034 on 814 degrees of freedom
## Multiple R-squared:  0.02634,    Adjusted R-squared:  0.02514
## F-statistic: 22.02 on 1 and 814 DF,  p-value: 3.165e-06
```

```r
summary(Rest_Europe_results)
```

```
##
## Call:
## lm(formula = Life_expectancy ~ Alcohol_consumption, data = data.Rest_of_Europe)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.001 -3.452 -0.620  4.023  8.960
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          75.5874     0.8569  88.211   <2e-16 ***
## Alcohol_consumption  -0.1401     0.1058  -1.324    0.187
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.672 on 238 degrees of freedom
```

```
## Multiple R-squared:  0.007312,   Adjusted R-squared:  0.003141
## F-statistic: 1.753 on 1 and 238 DF,  p-value: 0.1867
```

```
summary(Asia_results)
```

```
##
## Call:
## lm(formula = Life_expectancy ~ Alcohol_consumption, data = data.Asia)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.5352  -4.1260  -0.6656   3.9756  13.6374
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          67.8352     0.3584 189.279  < 2e-16 ***
## Alcohol_consumption   0.6744     0.1043   6.467 2.72e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.328 on 430 degrees of freedom
## Multiple R-squared:  0.08864,    Adjusted R-squared:  0.08652
## F-statistic: 41.82 on 1 and 430 DF,  p-value: 2.721e-10
```

```
summary(SA_results)
```

```
##
## Call:
## lm(formula = Life_expectancy ~ Alcohol_consumption, data = data.SA)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.216 -1.946  0.058  2.849  5.301
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          68.0654     0.8973  75.856  < 2e-16 ***
## Alcohol_consumption   0.8207     0.1506   5.451 1.54e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.305 on 190 degrees of freedom
## Multiple R-squared:  0.1352, Adjusted R-squared:  0.1307
## F-statistic: 29.72 on 1 and 190 DF,  p-value: 1.541e-07
```

```
summary(CenA_Car_results)
```

```
##
## Call:
## lm(formula = Life_expectancy ~ Alcohol_consumption, data = data.Cen_America_Caribbean)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -14.5205  -1.6941  -0.1391   2.6076   8.1599
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          70.1325     0.5784 121.261  < 2e-16 ***
## Alcohol_consumption   0.4099     0.0930   4.408 1.45e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.306 on 302 degrees of freedom
## Multiple R-squared:  0.06044,    Adjusted R-squared:  0.05733
## F-statistic: 19.43 on 1 and 302 DF,  p-value: 1.455e-05
```

```
summary(Oceania_results)
```

```
##
## Call:
## lm(formula = Life_expectancy ~ Alcohol_consumption, data = data.Oceania)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.6816 -1.9067 -0.0916  2.0388  5.6965
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         64.87576    0.28289   229.3   <2e-16 ***
## Alcohol_consumption  1.61162    0.06447    25.0   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.831 on 174 degrees of freedom
## Multiple R-squared:  0.7822, Adjusted R-squared:  0.781
## F-statistic:   625 on 1 and 174 DF,  p-value: < 2.2e-16
```

```
summary(EU_results)
```

```
##
## Call:
## lm(formula = Life_expectancy ~ Alcohol_consumption, data = data.EU)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.1775 -1.8788  0.9471  2.3002  5.6005
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         82.90972    0.76226 108.769  < 2e-16 ***
## Alcohol_consumption -0.48137    0.06929  -6.947 1.39e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.075 on 430 degrees of freedom
```

```
## Multiple R-squared:  0.1009, Adjusted R-squared:  0.09882
## F-statistic: 48.26 on 1 and 430 DF,  p-value: 1.388e-11
```

```
summary(ME_results)
```

```
##
## Call:
## lm(formula = Life_expectancy ~ Alcohol_consumption, data = data.ME)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.7666  -1.8371   0.3151   2.2412   5.3751
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          71.2509     0.3041  234.33   <2e-16 ***
## Alcohol_consumption   3.0816     0.2473   12.46   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.163 on 222 degrees of freedom
## Multiple R-squared:  0.4116, Adjusted R-squared:  0.4089
## F-statistic: 155.3 on 1 and 222 DF,  p-value: < 2.2e-16
```

```
summary(NA_results)
```

```
##
## Call:
## lm(formula = Life_expectancy ~ Alcohol_consumption, data = data.NA)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.50604 -0.99282 -0.05498  1.22563  3.09817
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          70.6896     0.8640  81.817  < 2e-16 ***
## Alcohol_consumption   1.0140     0.1185   8.557 4.53e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.516 on 46 degrees of freedom
## Multiple R-squared:  0.6142, Adjusted R-squared:  0.6058
## F-statistic: 73.22 on 1 and 46 DF,  p-value: 4.526e-11
```

There are several insights we can obtain from the above models and marginal scatter plots. For simplicity's sake, we will discuss each model's goodness of fit, including the p-values and R-squared values, as well as the significance of alcohol consumption on life expectancy.

Africa: R-Squared: 0.02634 p-value: 3.165e-06 Coefficient: -0.45948 at 3.17e-06

Comments: The associated p-value is extremely close to 0 (3.17e-06). This tells us that we can reject the null hypothesis and accept the alternative hypothesis. However, the R-squared value is 0.02634; this is a

value incredibly close to 0, which tells us that almost none of the variation in Africa's life expectancy can be explained by the variation in alcohol consumption. Regarding the coefficient associated with alcohol consumption, we observe that it is statistically significant at an alpha level of 0.05. Overall, this model is not a good fit.

Asia: R-Squared: 0.08864 p-value: 2.721e-10 Coefficient: 0.6744 at 2.72e-10

Comments: The associated p-value is extremely close to 0 (2.72e-10). This tells us that we can reject the null hypothesis and accept the alternative hypothesis. However, the R-squared value is 0.08864; this is a value incredibly close to 0, which tells us that almost none of the variation in Asia's life expectancy can be explained by the variation in alcohol consumption. Regarding the coefficient associated with alcohol consumption, we observe that it is statistically significant at an alpha level of 0.05. Overall, this model is not a good fit.

Central America & Caribbean: R-Squared: 0.06044 p-value: 1.455e-05 Coefficient: 0.4099 at 1.45e-05

Comments: The associated p-value is extremely close to 0 (1.45e-05). This tells us that we can reject the null hypothesis and accept the alternative hypothesis. However, the R-squared value is 0.06044; this is a value incredibly close to 0, which tells us that almost none of the variation in Central America and the Carribean's life expectancy can be explained by the variation in alcohol consumption. Regarding the coefficient associated with alcohol consumption, we observe that it is statistically significant at an alpha level of 0.05. Overall, this model is not a good fit.

European Union: R-Squared: 0.1009 p-value: 1.388e-11 Coefficient: -0.48137 at 1.39e-11

Comments: The associated p-value is extremely close to 0 (1.388e-11). This tells us that we can reject the null hypothesis and accept the alternative hypothesis. However, the R-squared value is 0.1009; this is a value incredibly close to 0, which tells us that appx. 10% of the variation in the Eurupean Union's life expectancy can be explained by the variation in alcohol consumption. Regarding the coefficient associated with alcohol consumption, we observe that it is statistically significant at an alpha level of 0.05. Overall, this model is not a good fit.

Rest of Europe: R-Squared: 0.007312 p-value: 0.1867 Coefficient: -0.1401 at 0.187

Comments: The associated p-value is not extremely close to 0 (0.1867). This tells us that we fail to reject the null hypothesis. However, the R-squared value is 0.007312; this is a value incredibly close to 0, which tells us that almost none of the variation in the Rest of Europe's life expectancy can be explained by the variation in alcohol consumption. Regarding the coefficient associated with alcohol consumption, we observe that it is not statistically significant at an alpha level of 0.05. Overall, this model is not a good fit.

Oceania: R-Squared: 0.7822 p-value: <2e-16 Coefficient: 1.61162 at <2e-16

Comments: The associated p-value is extremely close to 0 (<2e-16). This tells us that we can reject the null hypothesis and accept the alternative hypothesis. However, the R-squared value is 0.7822; this is a value relatively close to 1, which tells us that appx. 78% of the variation in Oceania's life expectancy can be explained by the variation in alcohol consumption. Regarding the coefficient associated with alcohol consumption, we observe that it is statistically significant at an alpha level of 0.05. Overall, this model is an okay fit.

Middle East: R-Squared: 0.4116 p-value: < 2.2e-16 Coefficient: 3.0816 at <2e-16

Comments: The associated p-value is extremely close to 0 (<2e-16). This tells us that we can reject the null hypothesis and accept the alternative hypothesis. However, the R-squared value is 0.4116; this is a value almost halfway between 0 and 1, which tells us that appx. 41% of the variation in the Middle East's life expectancy can be explained by the variation in alcohol consumption. Regarding the coefficient associated with alcohol consumption, we observe that it is statistically significant at an alpha level of 0.05. Overall, this model is not a great fit.

North America: R-Squared: 0.6142 p-value:4.526e-11 Coefficient: 1.0140 at 4.53e-11

Comments: The associated p-value is extremely close to 0 (4.526e-11). This tells us that we can reject the null hypothesis and accept the alternative hypothesis. However, the R-squared value is 0.6142; this is a

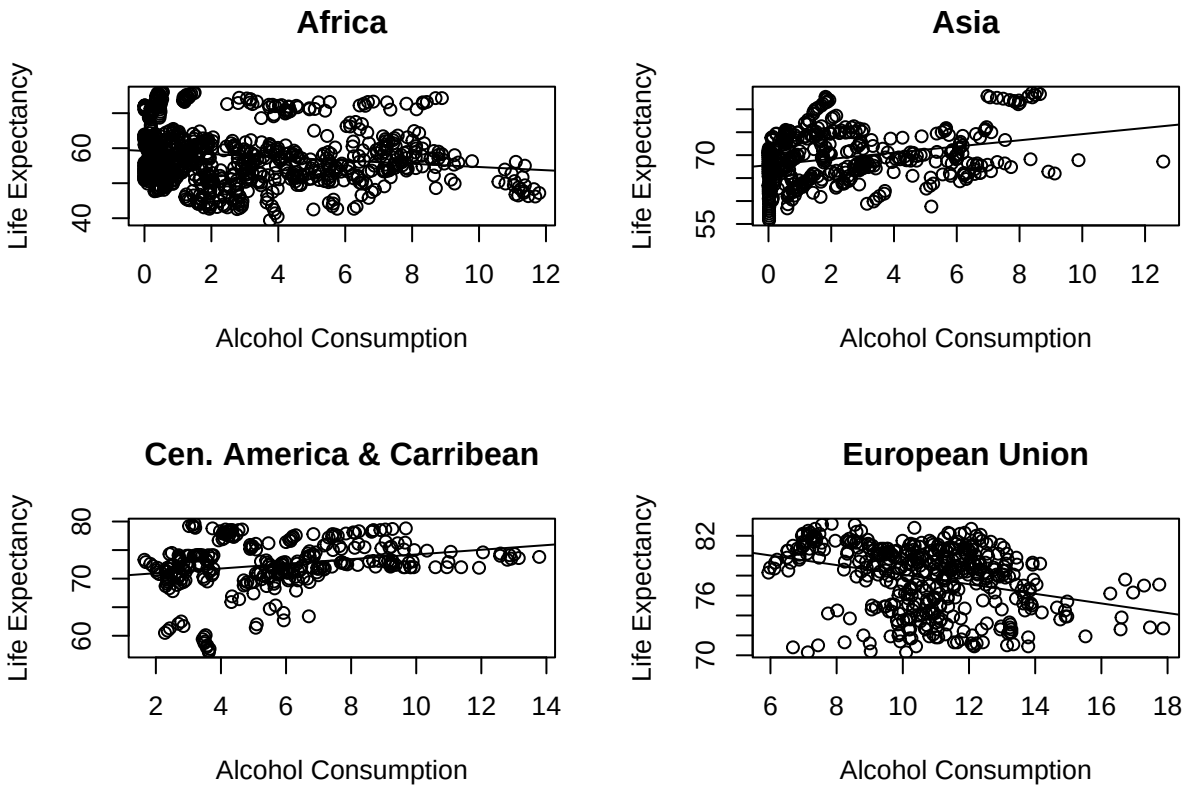
value almost halfway between 0 and 1, which tells us that appx. 61% of the variation in North America's life expectancy can be explained by the variation in alcohol consumption. Regarding the coefficient associated with alcohol consumption, we observe that it is statistically significant at an alpha level of 0.05. Overall, this model is an okay fit, but less so than the model for Oceania.

South America: R-Squared: 0.2139 p-value: 0.1301 Coefficient: 0.8999 at 0.13

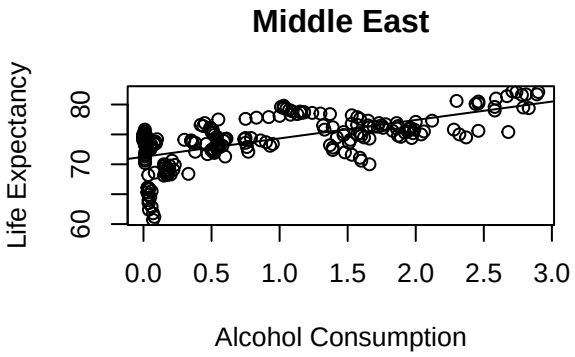
Comments: The associated p-value is not extremely close to 0 (0.13). This tells us that we fail to reject the null hypothesis. However, the R-squared value is 0.2139; this is a value close to 0, which tells us that appx. 21% of the variation in South America's life expectancy can be explained by the variation in alcohol consumption. Regarding the coefficient associated with alcohol consumption, we observe that it is not statistically significant at an alpha level of 0.05. Overall, this model is not a good fit.

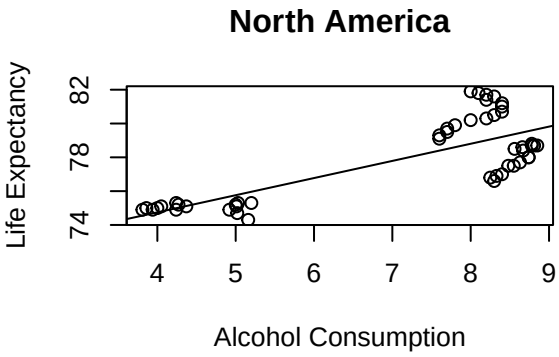
Marginal Scatter Plots by Region:

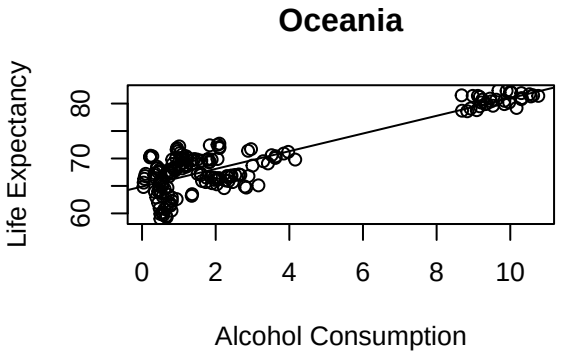
```
#generating marginal scatter plots of the data for visual comparison
par(mfrow=c(2,2))
plot(data.Africa$Alcohol_consumption, data.Africa$Life_expectancy, xlab = "Alcohol Consumption", ylab =
abline(Africa_results)

plot(data.Asia$Alcohol_consumption, data.Asia$Life_expectancy, xlab = "Alcohol Consumption", ylab = "Li
abline(Asia_results)

plot(data.Cen_America_Caribbean$Alcohol_consumption, data.Cen_America_Caribbean$Life_expectancy, xlab =
abline(CenA_Car_results)

plot(data.EU$Alcohol_consumption, data.EU$Life_expectancy, xlab = "Alcohol Consumption", ylab = "Life E
abline(EU_results)
```

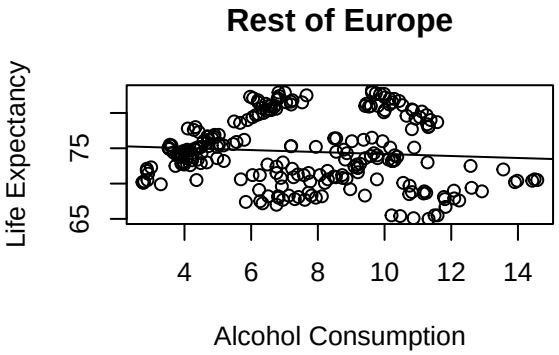


```r
plot(data.ME$Alcohol_consumption, data.ME$Life_expectancy, xlab = "Alcohol Consumption", ylab = "Life E
abline(ME_results)

plot(data.NA$Alcohol_consumption, data.NA$Life_expectancy, xlab = "Alcohol Consumption", ylab = "Life E
abline(NA_results)

plot(data.Oceania$Alcohol_consumption, data.Oceania$Life_expectancy, xlab = "Alcohol Consumption", ylab
abline(Oceania_results)

plot(data.Rest_of_Europe$Alcohol_consumption, data.Rest_of_Europe$Life_expectancy, xlab = "Alcohol Consu
abline(Rest_Europe_results)
```

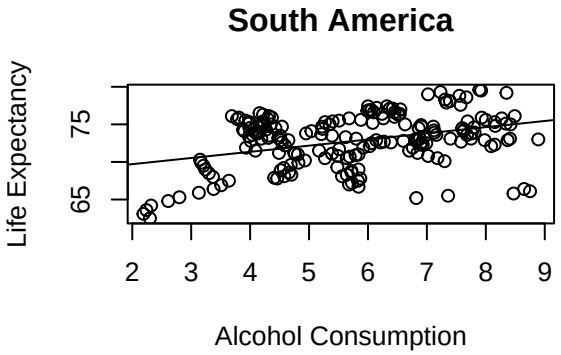









```r
plot(data.SA$Alcohol_consumption, data.SA$Life_expectancy, xlab = "Alcohol Consumption", ylab = "Life E
abline(SA_results)
```



Here we can see that while there is not a strong linear correlation between alcohol consumption and life expectancy among all regions, the best models appear to be North America, Oceania, and generiously, the Middle East and South America.

Because "life expectancy" accounts for those of all ages, we also thought it was important to consider the effects of alcohol on adult mortality only. As adults are likely the age group with the largest influence on the alcohol consumption data point, we hoped that replacing the dependent variable of life expectancy with adult mortality in the single linear regression would yeild better fitting models.

```
#creating single regression models

Africa_results <- lm(Adult_mortality~Alcohol_consumption, data = data.Africa)
Rest_Europe_results <- lm(Adult_mortality~Alcohol_consumption, data = data.Rest_of_Europe)
Asia_results <- lm(Adult_mortality~Alcohol_consumption, data = data.Asia)
SA_results <- lm(Adult_mortality~Alcohol_consumption, data = data.SA)
CenA_Car_results <- lm(Adult_mortality~Alcohol_consumption, data = data.Cen_America_Caribbean)
Oceania_results <- lm(Adult_mortality~Alcohol_consumption, data = data.Oceania)
EU_results <- lm(Adult_mortality~Alcohol_consumption, data = data.EU)
ME_results <- lm(Adult_mortality~Alcohol_consumption, data = data.ME)
NA_results <- lm(Adult_mortality~Alcohol_consumption, data = data.NA)

#outputting summary data
summary(Africa_results)
```

```
##
## Call:
```

```
## lm(formula = Adult_mortality ~ Alcohol_consumption, data = data.Africa)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -224.62  -63.81  -15.39   68.55  406.15
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          289.466      6.097  47.476  < 2e-16 ***
## Alcohol_consumption    9.933      1.473   6.744 2.93e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 120.9 on 814 degrees of freedom
## Multiple R-squared:  0.05291,    Adjusted R-squared:  0.05175
## F-statistic: 45.48 on 1 and 814 DF,  p-value: 2.925e-11
```

```
summary(Rest_Europe_results)
```

```
##
## Call:
## lm(formula = Adult_mortality ~ Alcohol_consumption, data = data.Rest_of_Europe)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -104.163  -49.765   -7.616   43.843  157.834
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)           82.977     12.187   6.808 7.97e-11 ***
## Alcohol_consumption    7.336      1.505   4.874 1.99e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 66.45 on 238 degrees of freedom
## Multiple R-squared:  0.09077,    Adjusted R-squared:  0.08695
## F-statistic: 23.76 on 1 and 238 DF,  p-value: 1.994e-06
```

```
summary(Asia_results)
```

```
##
## Call:
## lm(formula = Adult_mortality ~ Alcohol_consumption, data = data.Asia)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -120.341  -41.572    7.086   43.620  139.109
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          171.726      3.971  43.242   <2e-16 ***
## Alcohol_consumption   -0.233      1.155  -0.202     0.84
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 59.04 on 430 degrees of freedom
## Multiple R-squared:  9.459e-05,  Adjusted R-squared:  -0.002231
## F-statistic: 0.04068 on 1 and 430 DF,  p-value: 0.8403
```

```
summary(SA_results)
```

```
##
## Call:
## lm(formula = Adult_mortality ~ Alcohol_consumption, data = data.SA)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -50.240 -27.478  -3.713  20.213  97.410
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          205.734      9.173  22.428  < 2e-16 ***
## Alcohol_consumption   -8.782      1.539  -5.706 4.38e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 33.79 on 190 degrees of freedom
## Multiple R-squared:  0.1463, Adjusted R-squared:  0.1418
## F-statistic: 32.56 on 1 and 190 DF,  p-value: 4.378e-08
```

```
summary(CenA_Car_results)
```

```
##
## Call:
## lm(formula = Adult_mortality ~ Alcohol_consumption, data = data.Cen_America_Caribbean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -84.970 -23.027   3.686  23.561 136.441
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         179.8976     5.5712  32.291  < 2e-16 ***
## Alcohol_consumption  -2.9016     0.8959  -3.239  0.00133 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 41.48 on 302 degrees of freedom
## Multiple R-squared:  0.03357,    Adjusted R-squared:  0.03037
## F-statistic: 10.49 on 1 and 302 DF,  p-value: 0.001334
```

```
summary(Oceania_results)
```

```
##
## Call:
```

```
## lm(formula = Adult_mortality ~ Alcohol_consumption, data = data.Oceania)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -64.891 -26.197   0.423  20.661  81.181
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         202.0308     3.2173   62.80   <2e-16 ***
## Alcohol_consumption -13.5905     0.7332  -18.54   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 32.2 on 174 degrees of freedom
## Multiple R-squared:  0.6638, Adjusted R-squared:  0.6619
## F-statistic: 343.6 on 1 and 174 DF,  p-value: < 2.2e-16
```

```
summary(EU_results)
```

```
##
## Call:
## lm(formula = Adult_mortality ~ Alcohol_consumption, data = data.EU)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -55.09 -27.71 -15.13  24.44 155.95
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          30.0795    10.2460   2.936  0.00351 **
## Alcohol_consumption   6.8866     0.9314   7.394 7.53e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 41.33 on 430 degrees of freedom
## Multiple R-squared:  0.1128, Adjusted R-squared:  0.1107
## F-statistic: 54.67 on 1 and 430 DF,  p-value: 7.526e-13
```

```
summary(ME_results)
```

```
##
## Call:
## lm(formula = Adult_mortality ~ Alcohol_consumption, data = data.ME)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -78.662 -31.986  -1.577  17.871 118.247
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          136.799      4.074  33.581  < 2e-16 ***
## Alcohol_consumption  -28.999      3.313  -8.753 5.32e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 42.37 on 222 degrees of freedom
## Multiple R-squared:  0.2566, Adjusted R-squared:  0.2532
## F-statistic: 76.61 on 1 and 222 DF,  p-value: 5.324e-16
```

```
summary(NA_results)
```

```
##
## Call:
## lm(formula = Adult_mortality ~ Alcohol_consumption, data = data.NA)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.755 -18.992   2.659  16.000  21.531
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           145.162     10.114  14.353  < 2e-16 ***
## Alcohol_consumption    -6.305      1.387  -4.545 3.98e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.75 on 46 degrees of freedom
## Multiple R-squared:  0.3099, Adjusted R-squared:  0.2949
## F-statistic: 20.66 on 1 and 46 DF,  p-value: 3.975e-05
```

Africa: R-Squared: 0.05291 p-value: 2.925e-11 Coefficient: 9.933 at 2.93e-11

Comments: The associated p-value is extremely close to 0 (2.93e-11). This tells us that we can reject the null hypothesis and accept the alternative hypothesis. However, the R-squared value is 0.05291; while this value is slightly better than the life expectancy model, it is still incredibly close to 0, which tells us that almost none of the variation in Africa's adult mortality can be explained by the variation in alcohol consumption. Regarding the coefficient associated with alcohol consumption, we observe that it is statistically significant at an alpha level of 0.05. Overall, this model is not a good fit.

Asia: R-Squared: 9.459e-05 p-value: 0.8403 Coefficient: -0.233 at 0.84

Comments: The associated p-value is not close to 0 (0.8403). This tells us that we fail to reject the null hypothesis. However, the R-squared value is 9.459e-05; this is a value incredibly close to 0, which tells us that almost none of the variation in Asia's adult mortality can be explained by the variation in alcohol consumption. Regarding the coefficient associated with alcohol consumption, we observe that it is not statistically significant at an alpha level of 0.05. Overall, this model is not a good fit.

Central America & Caribbean: R-Squared: 0.03357 p-value: 0.001334 Coefficient: -2.9016 at 0.00133

Comments: The associated p-value is extremely close to 0 (0.00133). This tells us that we can reject the null hypothesis and accept the alternative hypothesis. However, the R-squared value is 0.03357; this is a value incredibly close to 0, which tells us that almost none of the variation in Central America and the Caribbean's adult mortality can be explained by the variation in alcohol consumption. Regarding the coefficient associated with alcohol consumption, we observe that it is statistically significant at an alpha level of 0.05. Overall, this model is not a good fit.

European Union: R-Squared: 0.1128 p-value: 7.526e-13 Coefficient: 6.8866 at 7.53e-13

Comments: The associated p-value is extremely close to 0 (7.526e-13). This tells us that we can reject the null hypothesis and accept the alternative hypothesis. However, the R-squared value is 0.1128; this is a value

incredibly close to 0, which tells us that appx. 11% of the variation in the European Union's adult mortality can be explained by the variation in alcohol consumption. Regarding the coefficient associated with alcohol consumption, we observe that it is statistically significant at an alpha level of 0.05. Overall, this model is not a good fit.

Rest of Europe: R-Squared: 0.09077 p-value: 1.994e-06 Coefficient: 7.336 at 1.99e-06

Comments: The associated p-value is extremely close to 0 (1.994e-06). This tells us that we can reject the null hypothesis and accept the alternative hypothesis. However, the R-squared value is 0.09077; this is a value incredibly close to 0, which tells us that almost none of the variation in the Rest of Europe's adult mortality can be explained by the variation in alcohol consumption. Regarding the coefficient associated with alcohol consumption, we observe that it is statistically significant at an alpha level of 0.05. Overall, this model is not a good fit.

Oceania: R-Squared: 0.6638 p-value: $<$ 2.2e-16 Coefficient: -13.5905 at $<$2e-16

Comments: The associated p-value is extremely close to 0 ($<$ 2.2e-16). This tells us that we can reject the null hypothesis and accept the alternative hypothesis. However, the R-squared value is 0.7822; this is a value relatively close to 1, which tells us that appx. 66% of the variation in Oceania's adult mortality can be explained by the variation in alcohol consumption. Regarding the coefficient associated with alcohol consumption, we observe that it is statistically significant at an alpha level of 0.05. Overall, this model is an okay fit.

Middle East: R-Squared: 0.2566 p-value: 5.324e-16 Coefficient: -28.999 at 5.32e-16

Comments: The associated p-value is extremely close to 0 (5.324e-16). This tells us that we can reject the null hypothesis and accept the alternative hypothesis. However, the R-squared value is 0.4116; this is a value almost halfway between 0 and 1, which tells us that appx. 25% of the variation in the Middle East's adult mortality can be explained by the variation in alcohol consumption. Regarding the coefficient associated with alcohol consumption, we observe that it is statistically significant at an alpha level of 0.05. Overall, this model is not a great fit.

North America: R-Squared: 0.3099 p-value: 3.975e-05 Coefficient: -6.305 at 3.98e-05

Comments: The associated p-value is extremely close to 0 (3.975e-05). This tells us that we can reject the null hypothesis and accept the alternative hypothesis. However, the R-squared value is 0.3099; this is a value a third of the way between 0 and 1, which tells us that appx. 31% of the variation in North America's adult mortality can be explained by the variation in alcohol consumption. Regarding the coefficient associated with alcohol consumption, we observe that it is statistically significant at an alpha level of 0.05. Overall, this model is not a good fit.

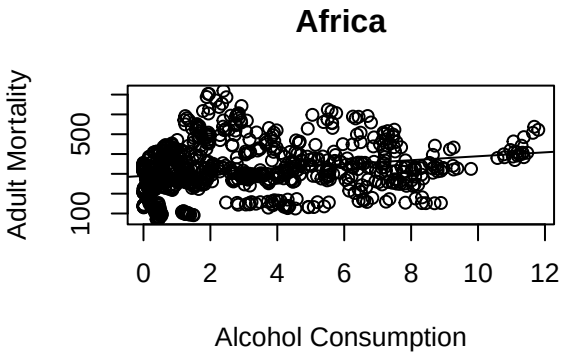
South America: R-Squared: 0.2702 p-value: 0.0832 Coefficient: -11.048 at 0.083204

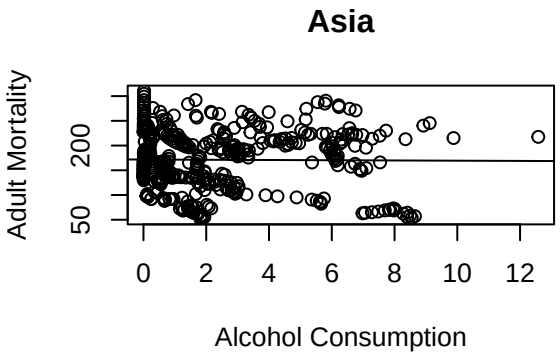
Comments: The associated p-value is not extremely close to 0 (0.0832). This tells us that we fail to reject the null hypothesis. However, the R-squared value is 0.2702; this is a value close to 0, which tells us that appx. 27% of the variation in South America's adult mortality can be explained by the variation in alcohol consumption. Regarding the coefficient associated with alcohol consumption, we observe that it is not statistically significant at an alpha level of 0.05. Overall, this model is not a good fit.

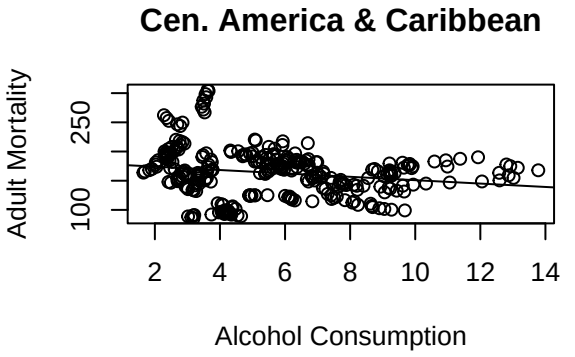
Marginal Scatter Plots by Region:

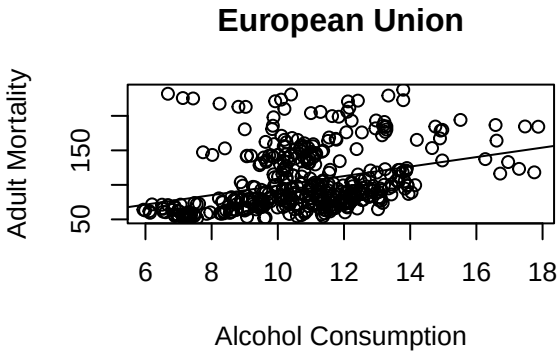
```
#generating marginal scatter plots of the data for visual comparison
par(mfrow=c(2,2))
plot(data.Africa$Alcohol_consumption, data.Africa$Adult_mortality, xlab = "Alcohol Consumption", ylab =
abline(Africa_results)

plot(data.Asia$Alcohol_consumption, data.Asia$Adult_mortality, xlab = "Alcohol Consumption", ylab = "Adu
abline(Asia_results)

plot(data.Cen_America_Caribbean$Alcohol_consumption, data.Cen_America_Caribbean$Adult_mortality, xlab =
```

```r
abline(CenA_Car_results)

plot(data.EU$Alcohol_consumption, data.EU$Adult_mortality, xlab = "Alcohol Consumption", ylab = "Adult M
abline(EU_results)
```

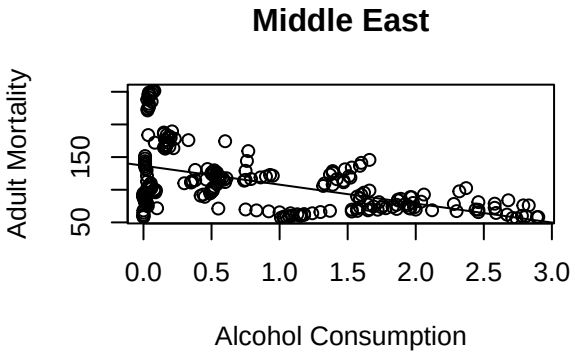


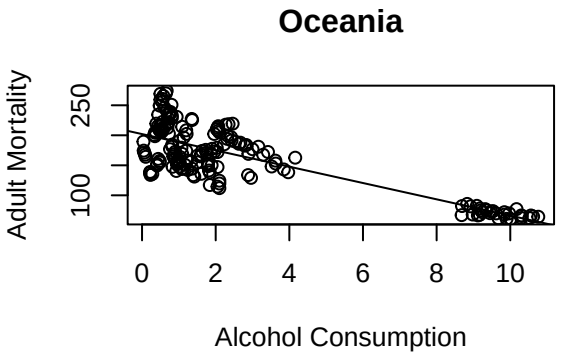







```r
plot(data.ME$Alcohol_consumption, data.ME$Adult_mortality, xlab = "Alcohol Consumption", ylab = "Adult M
abline(ME_results)

plot(data.NA$Alcohol_consumption, data.NA$Adult_mortality, xlab = "Alcohol Consumption", ylab = "Adult M
abline(NA_results)

plot(data.Oceania$Alcohol_consumption, data.Oceania$Adult_mortality, xlab = "Alcohol Consumption", ylab
abline(Oceania_results)

plot(data.Rest_of_Europe$Alcohol_consumption, data.Rest_of_Europe$Adult_mortality, xlab = "Alcohol Consu
abline(Rest_Europe_results)
```






```r
plot(data.SA$Alcohol_consumption, data.SA$Adult_mortality, xlab = "Alcohol Consumption", ylab = "Adult M
abline(SA_results)
```

Here we can see that while there is not a strong linear correlation between alcohol consumption and adult mortality among all regions, the best model appears to be Oceania.

While it was hypothesized that the linear regression models of alcohol consumption's effects on adult mortality would be stronger than the previous models that used life expectancy as the dependent variable, we see that this is actually not the case; while the adult mortality models show more regions having a statistically significant coefficient estimate for alcohol consumption, the R-squared values of these models are not better than those modeling life expectancy across the board.

Though the results of these single linear regression models did not yeild as strong of correlations as we had hoped, that is likely due to inconsistencies in the data set, particularly regarding the unequal number of data points for each region. While our research here does not discover anything groundbreaking, it raises an important question about the impact of alcohol on life expectancy across various geographical regions.

As alcohol consumption is a multifaceted behavior deeply intertwined with cultural norms, societal contexts, and individual habits, a more comprehensive and focused data set is needed should one want to conduct meaningful and significant research on the topic.