MA317 Final Assignment

Jasper Yang

12/6/2022

## Question 1

1. Exploratory Data Analysis: describe the data and deal with missing values
2. Analyse using descriptive statistics (both graphical and numerical representations) and R the Life Expectancy data1.csv dataset. [12 marks]

## [1] 217 27

##   
## FALSE TRUE   
## 198 19

##   
## FALSE TRUE   
## 0.9124424 0.0875576

## [1] 85.07805

## [1] "Hong Kong SAR, China"

## [1] 53.283

## [1] "Central African Republic"

## vec  
## character   
## 27

## Table by Continents

| Continent | Number of Countries | Life Expectancy at Birth (years) | Primary School Completion Rate (total of % by age) | Health Expenditure (current international $) | Population Density (people per sq. km) | GDP per Capita (current international $) |
| --- | --- | --- | --- | --- | --- | --- |
| Africa | 54 | 64.1 (5.9) | 79.8 (15.4) | 128.8 (173.4) | 104.8 (133.2) | 2630.1 (3096.2) |
| Asia | 50 | 74.6 (5.1) | 97.7 (9.8) | 721.8 (991.3) | 949.5 (3069.1) | 15116.8 (19922.9) |
| Australia/Oceania | 19 | 73.5 (6.0) | 94.1 (15.4) | 1164.2 (1679.5) | 145.2 (152.8) | 14783.1 (16508.8) |
| Europe | 48 | 79.3 (3.6) | 98.4 (5.5) | 2929.8 (2382.9) | 650.7 (2819.2) | 40646.5 (40399.5) |
| North America | 34 | 76.2 (3.8) | 92.3 (10.5) | 1293.0 (2334.5) | 281.6 (295.5) | 24082.3 (27212.6) |
| South America | 12 | 75.1 (3.2) | 97.3 (6.8) | 675.5 (450.6) | 24.7 (18.4) | 8495.9 (4220.4) |
| Total | 217 | 72.9 (7.5) | 93.0 (12.9) | 1143.7 (1838.1) | 446.0 (1996.6) | 18605.5 (27774.1) |

## Table of Variables with Missingness over 50%

| Variables | Missingness |
| --- | --- |
| Children newly infected with HIV | 58.5% |
| Educational attainment, primary | 83.4% |
| Educational attainment, bachelor's | 82.5% |
| Literacy rate | 88.5% |
| Poverty Ratio | 89.9% |

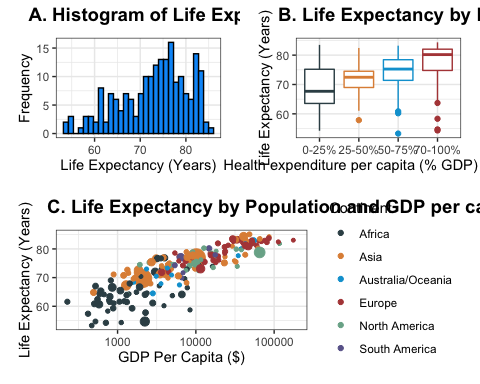
## Some graphs for out variable of interest: LE

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 19 rows containing non-finite values (stat\_bin).

## Warning: Removed 9 rows containing non-finite values (stat\_boxplot).

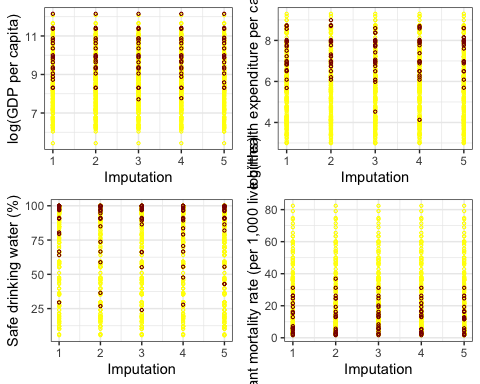
## Warning: Removed 28 rows containing missing values (geom\_point).



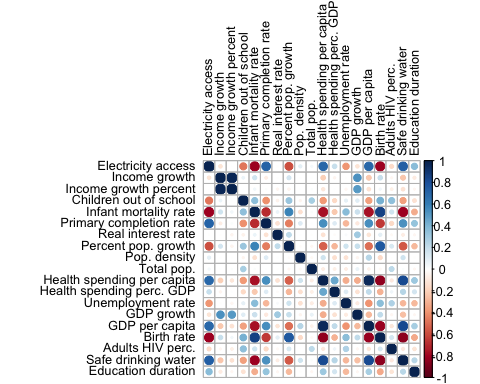
* The variable of primary interest is life expectancy. We show some graphs describing the life expectancy data for the 217 countries.
* We then select a subset of the 29 variables and showed the mean and standard deviation by continent. We selected one education, healthcare, financial, population
* We also observe that there is quite a bit of missingness here. We show the missingness rates for some variables here.

1. Many predictors in the dataset contain missing values. Is deleting predictor variables with many missing values an appropriate method to deal with missing values? Choose a method to deal with the missing values and then employ this method to the life expectancy data. Justify your choice. Additionally, there are some countries (cases) where the value of Life expectancy is missing. Explain how you will handle this problem. [13 marks]

## Warning: Number of logged events: 626



1. Collinearity increases the variance of the estimators and hence, reduces the adequacy of the model. When collinearity is present, how do you solve this problem? Investigate collinearity between the predictor variables in the LifeExpectancyData1.csv dataset. [12 marks]



## electricity\_access income\_growth income\_growth\_pc   
## 5.056682 2003.765408 1966.604947   
## children\_out\_of\_school infant\_mortality\_rate primary\_completion\_rate   
## 2.181388 5.809056 2.787419   
## real\_interest\_rate pop\_growth\_perc\_annual pop\_dens   
## 1.485066 82.468159 1.585584   
## pop\_total health\_expend\_pc health\_expend\_perc\_gdp   
## 1.411182 153.537923 10.396736   
## unemployment\_rate gdp\_growth gdp\_pc   
## 1.961158 2.376803 115.501396   
## birth\_rate adults\_HIV safe\_drinking\_water   
## 12.646941 1.531246 6.348283   
## ed\_duration   
## 1.398885

## income\_growth\_pc children\_out\_of\_school infant\_mortality\_rate   
## 1.599452 1.893906 4.622000   
## primary\_completion\_rate real\_interest\_rate pop\_growth\_perc\_annual   
## 2.305675 1.382186 1.878223   
## pop\_dens pop\_total health\_expend\_perc\_gdp   
## 1.148092 1.350169 1.295054   
## unemployment\_rate gdp\_growth gdp\_pc   
## 1.760251 1.894910 4.443390   
## adults\_HIV safe\_drinking\_water ed\_duration   
## 1.436581 5.028125 1.301247

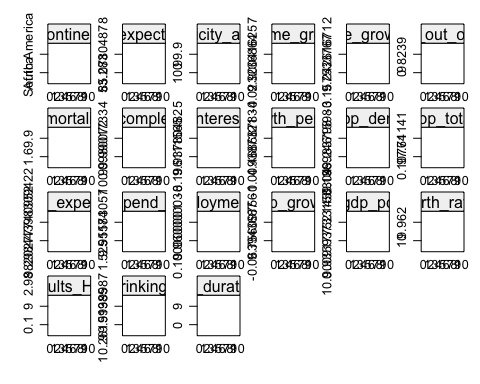
### VIF table  
temp <- data.frame("Feature" = names(vif\_output\_means\_after),  
 "VIF After Dropping Vars" = vif\_output\_means\_after)  
  
tableS1\_df <- data.frame("Feature" = colnames(cor\_df),  
 "VIF" = sprintf("%.2f",vif\_output\_means))  
tableS1\_df <- dplyr::left\_join(tableS1\_df, temp, by = "Feature")%>%  
 dplyr::mutate(VIF.After.Dropping.Vars =   
 ifelse(is.na(VIF.After.Dropping.Vars),  
 " ",  
 sprintf("%.2f", VIF.After.Dropping.Vars)))  
  
names(tableS1\_df)[3] <- "VIF After Dropping Vars"  
  
  
tableS1\_table <- flextable(tableS1\_df)  
  
tableS1\_table <- align(tableS1\_table,j =1, align = "left", part = "all")  
tableS1\_table <- align(tableS1\_table,j =2:3, align = "right", part = "all")  
tableS1\_table <- width(tableS1\_table, width = c(1.4,1.0, 1.0))  
#table1\_table <- height\_all(table1\_table,height = 0.30)  
tableS1\_table <- hrule(tableS1\_table, rule = "at least", part = "all")  
tableS1\_table

| Feature | VIF | VIF After Dropping Vars |
| --- | --- | --- |
| Electricity access | 5.06 |  |
| Income growth | 2003.77 |  |
| Income growth percent | 1966.60 |  |
| Children out of school | 2.18 |  |
| Infant mortality rate | 5.81 |  |
| Primary completion rate | 2.79 |  |
| Real interest rate | 1.49 |  |
| Percent pop. growth | 82.47 |  |
| Pop. density | 1.59 |  |
| Total pop. | 1.41 |  |
| Health spending per capita | 153.54 |  |
| Health spending perc. GDP | 10.40 |  |
| Unemployment rate | 1.96 |  |
| GDP growth | 2.38 |  |
| GDP per capita | 115.50 |  |
| Birth rate | 12.65 |  |
| Adults HIV perc. | 1.53 |  |
| Safe drinking water | 6.35 |  |
| Education duration | 1.40 |  |

We see a number of variables that have high VIFs and correlations with similar values with absolute values > 0.8. Based on these observations, we decide to remove national net income growth because of its high correlation with national net income growth per capita (r = 0.979), birth rate because of its high positive correlation with population growth rate (r = 0.76), infant mortality rate (r = 0.86) and large negative correlation with safe drinking water rate (r = - 0.85), and healthcare expenditure per capita because of its high correlation with GDP per capita (r = 0.81) and healthcare expenditure as percent of GDP (r = 0.51). We also decide to remove electricity access because of its high VIF and large correlations with infant mortality rate (r = 0.81), primary completion rate (r = 0.71), and safe drinking water rate (r = 0.72). Removing these helps us address multicollinearity, as evidenced by the larrge drop in VIF (Table S1), which makes our coefficients and their standard errors more appropriate.

1. To understand better life expectancy and the factors that affect it, suggest the best linear model which predicts life expectancy in 2020. Interpret and evaluate the suggested model. [25 marks]

stripplot(imputations, pch = 20, cex = 1.2)



#xyplot(imputations, bmi ~ chl | .imp, pch = 20, cex = 1.4)  
  
#Define the analysis you want to perform at Stage 2  
feature.selection1 <- expression(null.model1 <- lm(life\_expectancy ~ 1), model2 <- step(null.model1, scope = ~ income\_growth\_pc + children\_out\_of\_school + infant\_mortality\_rate + primary\_completion\_rate + real\_interest\_rate + pop\_growth\_perc\_annual + pop\_dens + pop\_total + health\_expend\_perc\_gdp + unemployment\_rate + gdp\_growth + gdp\_pc + adults\_HIV + safe\_drinking\_water + ed\_duration, direction = "forward"))  
  
#feature.selection1 <- expression(null.model1 <- lm(life\_expectancy ~ 1), model2 <- step(null.model1, scope = ~ income\_growth\_pc + infant\_mortality\_rate + pop\_growth\_perc\_annual + pop\_dens + pop\_total + health\_expend\_pc + health\_expend\_perc\_gdp + gdp\_growth + gdp\_pc + ed\_duration))  
  
# Redefine list of imputations to remove life expectancy for countries  
# with NA life expectancy originally.   
countries\_with\_LE <- dplyr::filter(LE\_data,  
 !(is.na(life\_expectancy)))$Country.Name  
  
  
imp\_filt <- mice::filter(imputations,   
 Country.Name %in% countries\_with\_LE)

Stepwise feature selection with our imputed datasets yields the following table showing the frequency of feature selection:

## feature.frequency  
## gdp\_pc health\_expend\_perc\_gdp infant\_mortality\_rate   
## 10 10 10   
## pop\_dens safe\_drinking\_water adults\_HIV   
## 10 10 8   
## pop\_growth\_perc\_annual income\_growth\_pc unemployment\_rate   
## 8 5 5   
## real\_interest\_rate pop\_total children\_out\_of\_school   
## 4 2 1   
## gdp\_growth primary\_completion\_rate   
## 1 1

We observe that five features appear in at least 80% of the final models after stepwise selection, so we decide to build a model with only these five variables.

## List of 10  
## $ :List of 12  
## ..- attr(\*, "class")= chr "lm"  
## $ :List of 12  
## ..- attr(\*, "class")= chr "lm"  
## $ :List of 12  
## ..- attr(\*, "class")= chr "lm"  
## $ :List of 12  
## ..- attr(\*, "class")= chr "lm"  
## $ :List of 12  
## ..- attr(\*, "class")= chr "lm"  
## $ :List of 12  
## ..- attr(\*, "class")= chr "lm"  
## $ :List of 12  
## ..- attr(\*, "class")= chr "lm"  
## $ :List of 12  
## ..- attr(\*, "class")= chr "lm"  
## $ :List of 12  
## ..- attr(\*, "class")= chr "lm"  
## $ :List of 12  
## ..- attr(\*, "class")= chr "lm"

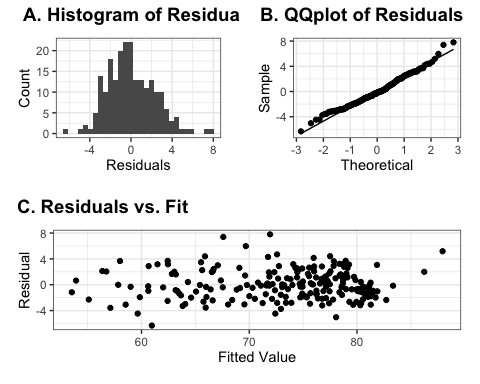
## term estimate std.error statistic df  
## 1 (Intercept) 64.3235583144 2.0588808256 31.242002 105.06635  
## 2 health\_expend\_perc\_gdp 0.1719842795 0.0778533048 2.209081 114.12325  
## 3 infant\_mortality\_rate -0.2481012959 0.0167783762 -14.786967 126.45356  
## 4 safe\_drinking\_water 0.0206619179 0.0137692438 1.500585 51.78524  
## 5 pop\_dens 0.0003418472 0.0001239502 2.757939 99.21297  
## 6 gdp\_pc 1.2622025589 0.2585648119 4.881571 75.82206  
## p.value  
## 1 0.000000000000  
## 2 0.029166357685  
## 3 0.000000000000  
## 4 0.139535840143  
## 5 0.006925697436  
## 6 0.000005693592

## est lo 95 hi 95 fmi  
## adj R^2 0.8973307 0.8657457 0.9218149 0.03397423

## health\_expend\_perc\_gdp infant\_mortality\_rate safe\_drinking\_water   
## 1.113815 3.092219 4.182744   
## pop\_dens gdp\_pc   
## 1.116424 3.685890

| Feature | Estimate | Standard Error | df | P-value | Proportion of Best Subsets Appeared in | VIF |
| --- | --- | --- | --- | --- | --- | --- |
| Intercept | 64.3236 | 2.05888 | 105.07 | < 0.0001 |  |  |
| Health Expenditure (% GDP) | 0.1720 | 0.07785 | 114.12 | 0.0292 | 1 | 1.11381537067822 |
| Infant mortality rate | -0.2481 | 0.01678 | 126.45 | < 0.0001 | 1 | 3.09221879971042 |
| % drinking safe water | 0.0207 | 0.01377 | 51.79 | 0.1395 | 1 | 4.18274432283142 |
| Population density (per sq. km) | 0.0003 | 0.00012 | 99.21 | 0.0069 | 1 | 1.11642350413605 |
| log(GDP per capita) | 1.2622 | 0.25856 | 75.82 | < 0.0001 | 1 | 3.68589014054934 |

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



The figure above shows evidence that the residuals produced by our model predictions are approximately normally distributed without a clear pattern. This evidence is a sign that our model is good.

## New names:  
## \* NA -> ...1  
## \* NA -> ...2  
## \* NA -> ...3  
## \* NA -> ...4  
## \* NA -> ...5  
## \* ...

| Country | Predicted life expectancy |
| --- | --- |
| American Samoa | 77.11 |
| Andorra | 80.19 |
| British Virgin Islands | 74.92 |
| Cayman Islands | 81.04 |
| Curacao | 77.11 |
| Dominica | 70.31 |
| Gibraltar | 78.36 |
| Greenland | 79.11 |
| Isle of Man | 80.55 |
| Marshall Islands | 72.36 |
| Monaco | 87.94 |
| Nauru | 73.08 |
| Northern Mariana Islands | 77.53 |
| Palau | 76.96 |
| San Marino | 80.88 |
| Sint Maarten (Dutch part) | 75.77 |
| St. Kitts and Nevis | 76.55 |
| Turks and Caicos Islands | 78.35 |
| Tuvalu | 75.57 |

1. Using the same dataset and using the additional feature Continent, employ an appropriate experimental design to study differences of average life expectancies across the continents: Asia, Europe, North America, South America, Africa, Australia/Oceania. Justify your choice of experimental design and methods. [13 marks]

## Df Sum Sq Mean Sq F value Pr(>F)   
## as.factor(Continent) 5 6797 1359.4 58.84 <0.0000000000000002 \*\*\*  
## Residuals 211 4874 23.1   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## Anova Table (Type III tests)  
##   
## Response: life\_expectancy  
## Sum Sq Df F value Pr(>F)   
## (Intercept) 221946 1 9607.306 < 0.00000000000000022 \*\*\*  
## as.factor(Continent) 6797 5 58.844 < 0.00000000000000022 \*\*\*  
## Residuals 4874 211   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

To study the differences in life expectancy across continents (a factor variable), we employ an ANCOVA experimental design. ANCOVA is designed specifically for testing whether there is a difference in means for a continuous response variable (in our case, life expectancy) across a categorical predictoor (continent), and it has the additional benefit of controlling for covariates that are independent of the predictor. In our experimental design, we perform an ANCOVA with % of GDP spent on health as a covariate.

Due to the fact that the references and software for performing ANCOVA and related post-hoc analyses on multiply imputed data are limited, we use single imputation by chained equations (multiple imputation with m = 1) for this experiment. To combat the limitations of only using one imputation, we will also perform sensitivity analyses with different random seeds given to the imputation algorithm to ensure that was not do to the randomness of the single imputation.

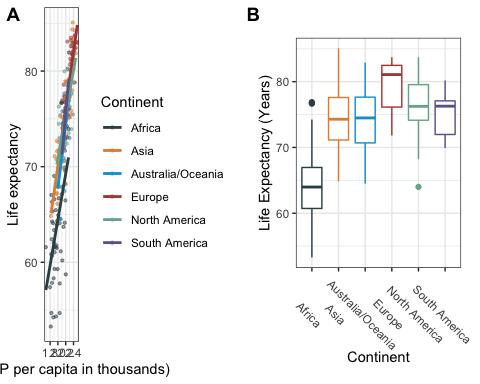
## Df Sum Sq Mean Sq F value Pr(>F)   
## log(gdp\_pc) 1 8322 8322 692.75 < 0.0000000000000002 \*\*\*  
## Continent 5 827 165 13.78 0.0000000000122 \*\*\*  
## Residuals 210 2523 12   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##   
## Simultaneous Tests for General Linear Hypotheses  
##   
## Multiple Comparisons of Means: Tukey Contrasts  
##   
##   
## Fit: aov(formula = life\_expectancy ~ log(gdp\_pc) + Continent, data = anova\_data)  
##   
## Linear Hypotheses:  
## Estimate Std. Error t value Pr(>|t|)   
## Asia - Africa == 0 5.80144 0.75882 7.645 <0.001 \*\*\*  
## Australia/Oceania - Africa == 0 4.45432 1.00605 4.428 <0.001 \*\*\*  
## Europe - Africa == 0 6.52681 0.92947 7.022 <0.001 \*\*\*  
## North America - Africa == 0 5.42194 0.90230 6.009 <0.001 \*\*\*  
## South America - Africa == 0 5.50317 1.17334 4.690 <0.001 \*\*\*  
## Australia/Oceania - Asia == 0 -1.34711 0.93602 -1.439 0.694   
## Europe - Asia == 0 0.72538 0.75770 0.957 0.928   
## North America - Asia == 0 -0.37949 0.78526 -0.483 0.997   
## South America - Asia == 0 -0.29827 1.11549 -0.267 1.000   
## Europe - Australia/Oceania == 0 2.07249 0.96683 2.144 0.261   
## North America - Australia/Oceania == 0 0.96762 0.99694 0.971 0.923   
## South America - Australia/Oceania == 0 1.04885 1.27800 0.821 0.962   
## North America - Europe == 0 -1.10487 0.78892 -1.400 0.718   
## South America - Europe == 0 -1.02364 1.14282 -0.896 0.945   
## South America - North America == 0 0.08123 1.16778 0.070 1.000   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
## (Adjusted p values reported -- single-step method)

Before proceeding with the statistical tests, we check the assumptions of ANCOVA which are (potentially show diagnostic plots in appendix only):

* Homogeneity of regression slopes.
* The relationship between the response and the covariate is linear.

## `geom\_smooth()` using formula 'y ~ x'

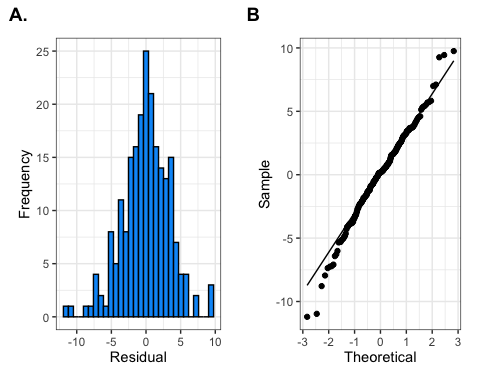


## Levene's Test for Homogeneity of Variance (center = median)  
## Df F value Pr(>F)   
## group 5 2.2035 0.05517 .  
## 211   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

* Normality of residuals
* Equal variances between the different continents

# Look at residuals  
anova\_data$residuals <- le.aov$residuals  
p1 <- ggplot(anova\_data, aes(residuals)) +  
 geom\_histogram(color = "#000000", fill = "#0099F8") +  
 theme\_bw()+  
 xlab("Residual")+  
 ylab("Frequency")+   
 theme(plot.margin = unit(c(1,0.5,0.5,0.5), "cm"))  
p2 <- ggplot(anova\_data, aes(sample = residuals)) +  
 stat\_qq() +  
 stat\_qq\_line() +   
 theme\_bw()+   
 theme(plot.margin = unit(c(1,0.5,0.5,0.5), "cm")) +  
 xlab("Theoretical") +  
 ylab("Sample")  
  
plot\_grid(p1, p2, labels = c("A.", "B"))

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



shapiro.test(anova\_data$residuals)

##   
## Shapiro-Wilk normality test  
##   
## data: anova\_data$residuals  
## W = 0.98869, p-value = 0.0848

(Null hypothesis and alternative hypothesis here)