Title

Final Data-analysis Project STAT 420 by

Saitejas Mopuri, Kimberly Martin, Woo Jeon

Appendix

Introduction

Methods

Loading Data

```
#read data
library(readr)
life_expectancy_data_full = read_csv("Life Expectancy Data.csv")
```

```
## Rows: 2938 Columns: 22
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr  (2): Country, Status
## dbl (20): Year, Life expectancy, Adult Mortality, infant deaths, Alcohol, pe...
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
#remove rows with empty vals
life_expectancy_data = na.omit(life_expectancy_data_full)
View(life_expectancy_data)

#convert Status (levels: Developing, Developed) to factor variable
life_expectancy_data$Status = as.factor(life_expectancy_data$Status)

#convert Country to factor variable
life_expectancy_data$Country = as.factor(life_expectancy_data$Country)
#levels(life_expectancy_data$Country)

#check structure of dataset to ensure correct factors
str(life_expectancy_data)
```

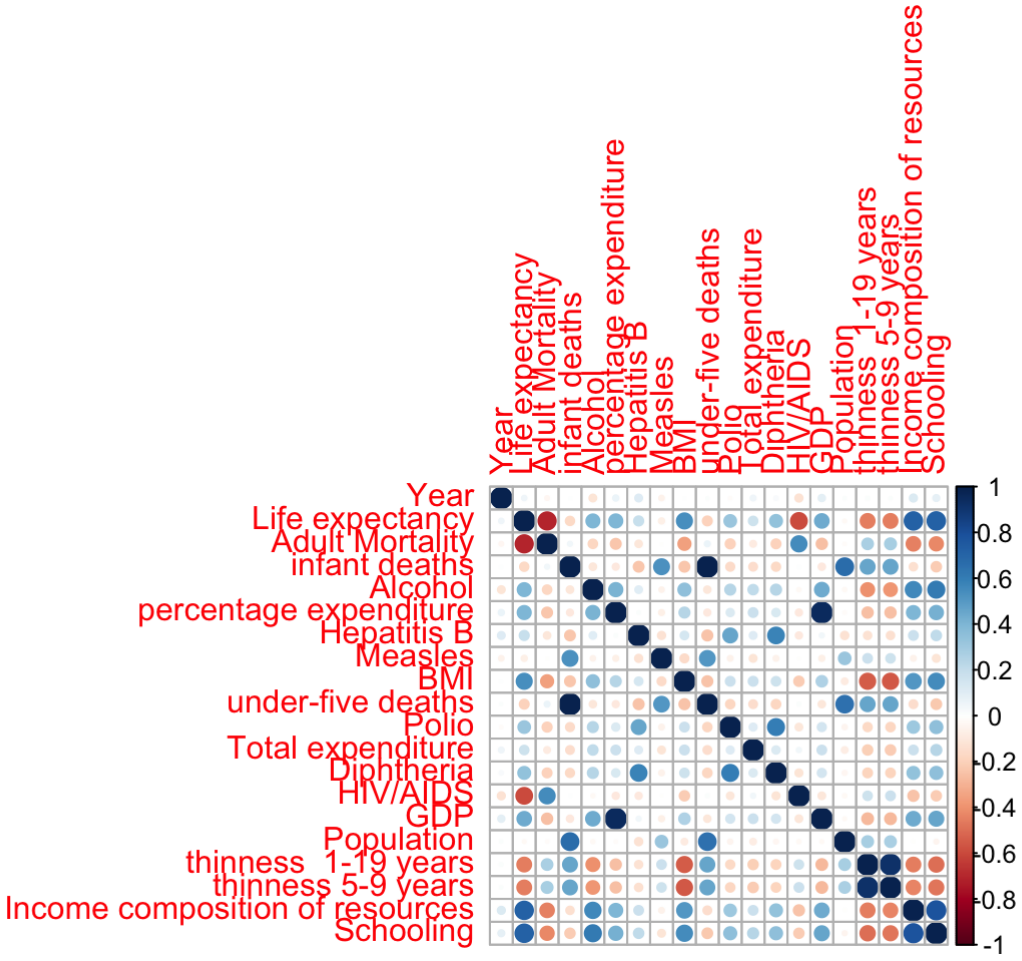
```
## tibble [1,649 × 22] (S3: tbl_df/tbl/data.frame)
##  $ Country                        : Factor w/ 133 levels "Afghanistan",..: 1 1 1 1 1
1 1 1 1 1 ...
##  $ Year                           : num [1:1649] 2015 2014 2013 2012 2011 ...
##  $ Status                         : Factor w/ 2 levels "Developed","Developing": 2 2
2 2 2 2 2 2 2 ...
##  $ Life expectancy                : num [1:1649] 65 59.9 59.9 59.5 59.2 58.8 58.6 58.
1 57.5 57.3 ...
##  $ Adult Mortality                : num [1:1649] 263 271 268 272 275 279 281 287 295
295 ...
##  $ infant deaths                  : num [1:1649] 62 64 66 69 71 74 77 80 82 84 ...
##  $ Alcohol                        : num [1:1649] 0.01 0.01 0.01 0.01 0.01 0.01 0.01
0.03 0.02 0.03 ...
##  $ percentage expenditure         : num [1:1649] 71.3 73.5 73.2 78.2 7.1 ...
##  $ Hepatitis B                    : num [1:1649] 65 62 64 67 68 66 63 64 63 64 ...
##  $ Measles                        : num [1:1649] 1154 492 430 2787 3013 ...
##  $ BMI                            : num [1:1649] 19.1 18.6 18.1 17.6 17.2 16.7 16.2 1
5.7 15.2 14.7 ...
##  $ under-five deaths              : num [1:1649] 83 86 89 93 97 102 106 110 113 116
...
##  $ Polio                          : num [1:1649] 6 58 62 67 68 66 63 64 63 58 ...
##  $ Total expenditure              : num [1:1649] 8.16 8.18 8.13 8.52 7.87 9.2 9.42 8.
33 6.73 7.43 ...
##  $ Diphtheria                     : num [1:1649] 65 62 64 67 68 66 63 64 63 58 ...
##  $ HIV/AIDS                       : num [1:1649] 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1
0.1 ...
##  $ GDP                            : num [1:1649] 584.3 612.7 631.7 670 63.5 ...
##  $ Population                     : num [1:1649] 33736494 327582 31731688 3696958 297
8599 ...
##  $ thinness  1-19 years           : num [1:1649] 17.2 17.5 17.7 17.9 18.2 18.4 18.6 1
8.8 19 19.2 ...
##  $ thinness 5-9 years             : num [1:1649] 17.3 17.5 17.7 18 18.2 18.4 18.7 18.
9 19.1 19.3 ...
##  $ Income composition of resources: num [1:1649] 0.479 0.476 0.47 0.463 0.454 0.448
0.434 0.433 0.415 0.405 ...
##  $ Schooling                      : num [1:1649] 10.1 10 9.9 9.8 9.5 9.2 8.9 8.7 8.4
8.1 ...
##  - attr(*, "na.action")= 'omit' Named int [1:1289] 33 45 46 47 48 49 58 59 60 61 ...
##   ..- attr(*, "names")= chr [1:1289] "33" "45" "46" "47" ...
```

Exploratory Analysis

```
#Pairs plot is very large, so we create a Correlogram to view correlations
#pairs(life_expectancy_data)

#find pairwise correlations between all numeric variables in dataset
numeric_cols =  unlist(lapply(life_expectancy_data, is.numeric))
life_expectancy_data_numeric = life_expectancy_data[ , numeric_cols]
View(life_expectancy_data_numeric)

correlations = cor(life_expectancy_data_numeric)

#plot correlogram (positive correlations are blue, negative correlations are red)
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
corrplot(correlations)
```



Initial Additive Model

```
#create initial additive model with all predictors

additive_model = lm(life_expectancy_data$`Life expectancy` ~ ., data = life_expectancy_d
ata)
#summary(additive_model)
```

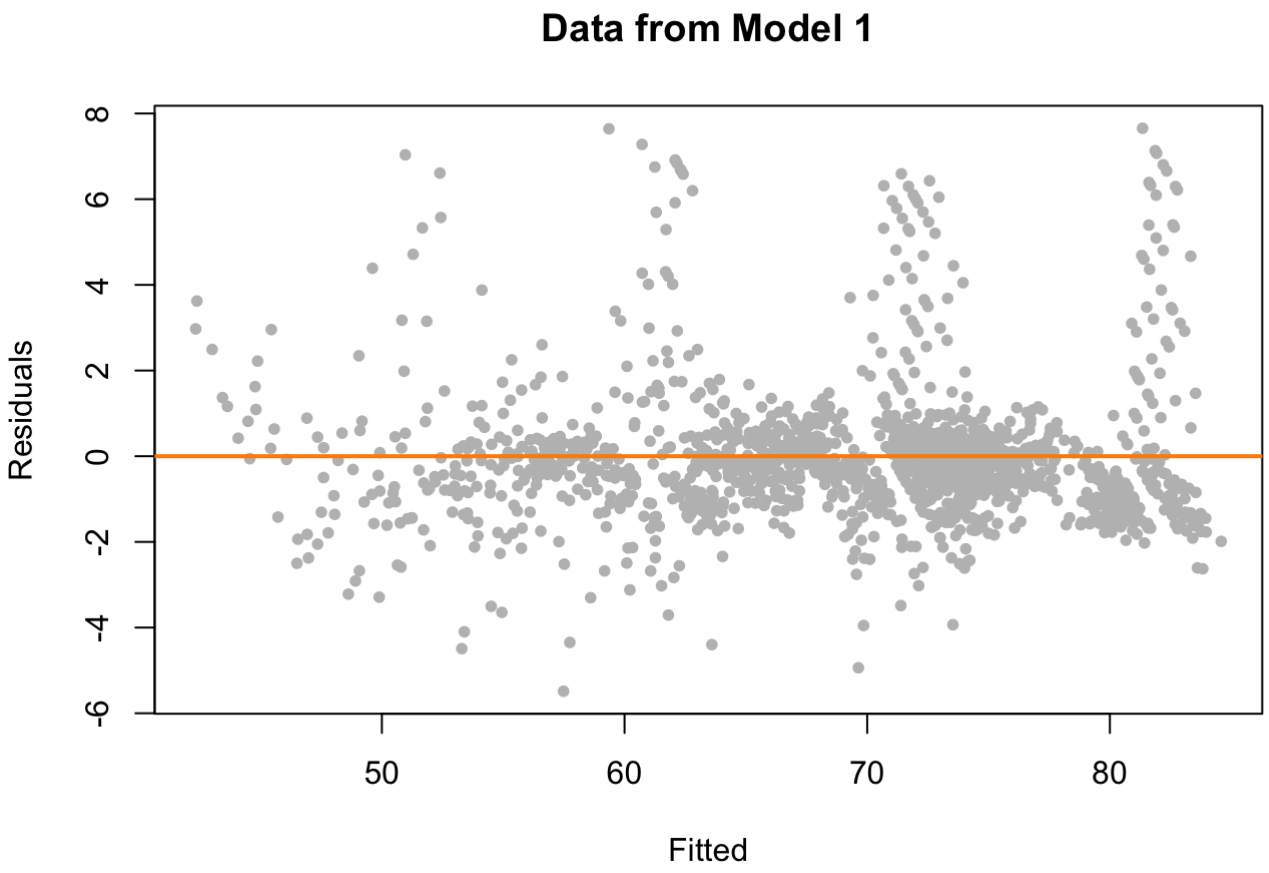
Analyze Model

```
#Check adjusted R squared
additive_model_adjr2 = summary(additive_model)$adj.r.squared

#Calculate RMSE
additive_model_rmse = sqrt(mean(resid(additive_model) ^ 2))

#Calculate LOOCV RMSE
calc_loocv_rmse = function(model) {
  sqrt(mean((resid(model) / (1 - hatvalues(model))) ^ 2))
}

additive_model_loocv_rmse = calc_loocv_rmse(additive_model)

#Calculate AIC
additive_model_AIC = extractAIC(additive_model)
#it is very large (bad)!
additive_model_AIC
```

```
## [1]  152.000 1824.977
```

Check model assumptions

```
#------- Check model assumptions -------------

#Fitted vs Residuals Plot

plot(fitted(additive_model), resid(additive_model), col = "grey", pch = 20,
     xlab = "Fitted", ylab = "Residuals", main = "Data from Model 1")
abline(h = 0, col = "darkorange", lwd = 2)
```

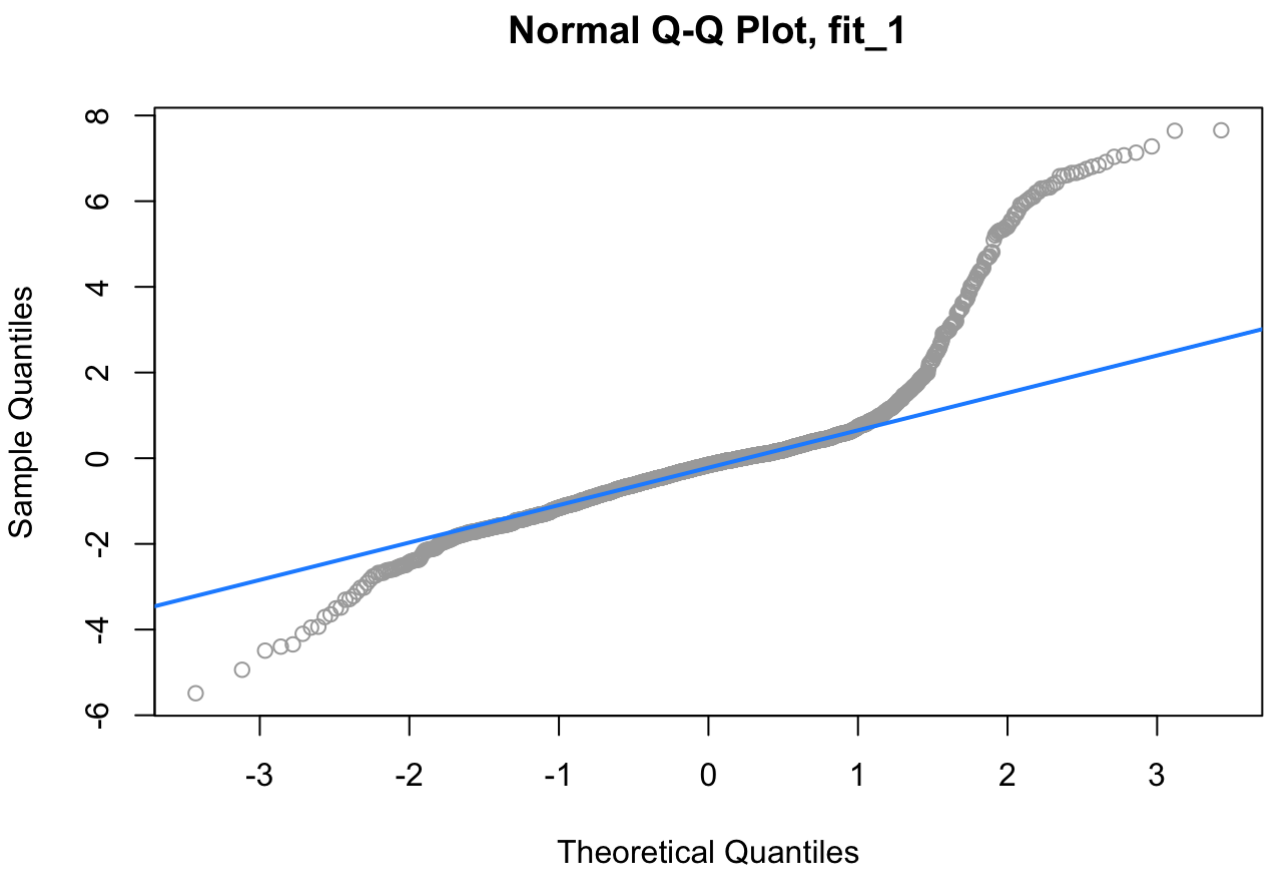



```
#Q-Q Plot

qqnorm(resid(additive_model), main = "Normal Q-Q Plot, fit_1", col = "darkgrey")
qqline(resid(additive_model), col = "dodgerblue", lwd = 2)

#confirm issues via more formal tests

library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```




```
bptest(additive_model)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  additive_model
## BP = 283.35, df = 151, p-value = 4.012e-10
```

```
#result shows data violates constant variance assumption!

shapiro.test(resid(additive_model))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(additive_model)
## W = 0.82427, p-value < 2.2e-16
```

```
#result shows data violates normality assumption!
```

```
#---------- Next step ------------------------------

#We will attempt to fix these issues by removing highly correlated predictors
#under 5 death & infant deaths : cor = 0.997
#thinness 1-19 years & thinness 5-9 years : cor = 0.928
#We'll remove one from each of the above pairs of predictors and reassess model



#---------- Next step ------------------------------
#We will attempt to further improve the model by log transforming certain predictor vari
ables (Population & GDP??)



#---------- Next step ------------------------------

#We will use AIC and BIC backwards search to attempt to further improve the model
```

Results

Discussions