Appendix

2022-12-05

# Appendix A: Data Cleaning

<https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who>

library(tidyverse)  
library(skimr)  
library(GGally)  
library(car)  
library(broom)  
library(leaps)  
library(gt)

df <- read\_csv(  
 paste('https://raw.githubusercontent.com/',  
 'stinalindaa/git-life-expectancy/main/Life.csv',   
 sep = ""))  
df <- df |>  
 rename(country = Country,  
 year = Year,  
 status = Status,  
 life = `Life expectancy`,  
 mort = `Adult Mortality`,  
 inf = `infant deaths`,  
 alc = Alcohol,  
 exp.p = `percentage expenditure`,  
 hep = `Hepatitis B`,  
 meas = Measles,   
 bmi = BMI,   
 under5 = `under-five deaths`,  
 polio = Polio,  
 exp.t = `Total expenditure`,  
 dip = Diphtheria,  
 hiv = `HIV/AIDS`,  
 gdp = GDP,  
 pop = Population,  
 thin1.19 = `thinness 1-19 years`,  
 thin5.9 = `thinness 5-9 years`,  
 comp = `Income composition of resources`,  
 school = Schooling) |>  
 mutate(year = factor(year))  
head(df)

## # A tibble: 6 × 22  
## country year status life mort inf alc exp.p hep meas bmi under5  
## <chr> <fct> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 Afghanist… 2015 Devel… 65 263 62 0.01 71.3 65 1154 19.1 83  
## 2 Afghanist… 2014 Devel… 59.9 271 64 0.01 73.5 62 492 18.6 86  
## 3 Afghanist… 2013 Devel… 59.9 268 66 0.01 73.2 64 430 18.1 89  
## 4 Afghanist… 2012 Devel… 59.5 272 69 0.01 78.2 67 2787 17.6 93  
## 5 Afghanist… 2011 Devel… 59.2 275 71 0.01 7.10 68 3013 17.2 97  
## 6 Afghanist… 2010 Devel… 58.8 279 74 0.01 79.7 66 1989 16.7 102  
## # … with 10 more variables: polio <dbl>, exp.t <dbl>, dip <dbl>, hiv <dbl>,  
## # gdp <dbl>, pop <dbl>, thin1.19 <dbl>, thin5.9 <dbl>, comp <dbl>,  
## # school <dbl>

# Appendix B: Exploratory Data Analysis

## B.1: Skim

skim\_without\_charts(df)

Data summary

|  |  |
| --- | --- |
| Name | df |
| Number of rows | 2938 |
| Number of columns | 22 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| character | 2 |
| factor | 1 |
| numeric | 19 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: character**

| skim\_variable | n\_missing | complete\_rate | min | max | empty | n\_unique | whitespace |
| --- | --- | --- | --- | --- | --- | --- | --- |
| country | 0 | 1 | 4 | 52 | 0 | 193 | 0 |
| status | 0 | 1 | 9 | 10 | 0 | 2 | 0 |

**Variable type: factor**

| skim\_variable | n\_missing | complete\_rate | ordered | n\_unique | top\_counts |
| --- | --- | --- | --- | --- | --- |
| year | 0 | 1 | FALSE | 16 | 201: 193, 200: 183, 200: 183, 200: 183 |

**Variable type: numeric**

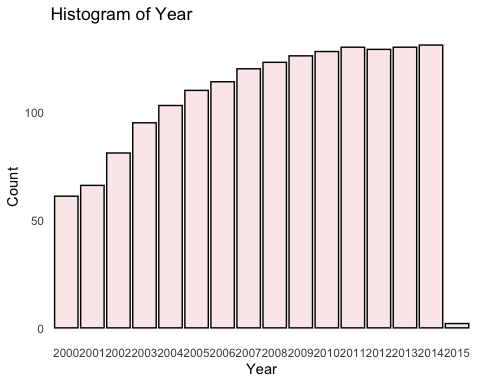
| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| life | 10 | 1.00 | 69.22 | 9.52 | 36.30 | 63.10 | 72.10 | 75.70 | 8.900000e+01 |
| mort | 10 | 1.00 | 164.80 | 124.29 | 1.00 | 74.00 | 144.00 | 228.00 | 7.230000e+02 |
| inf | 0 | 1.00 | 30.30 | 117.93 | 0.00 | 0.00 | 3.00 | 22.00 | 1.800000e+03 |
| alc | 194 | 0.93 | 4.60 | 4.05 | 0.01 | 0.88 | 3.76 | 7.70 | 1.787000e+01 |
| exp.p | 0 | 1.00 | 738.25 | 1987.91 | 0.00 | 4.69 | 64.91 | 441.53 | 1.947991e+04 |
| hep | 553 | 0.81 | 80.94 | 25.07 | 1.00 | 77.00 | 92.00 | 97.00 | 9.900000e+01 |
| meas | 0 | 1.00 | 2419.59 | 11467.27 | 0.00 | 0.00 | 17.00 | 360.25 | 2.121830e+05 |
| bmi | 34 | 0.99 | 38.32 | 20.04 | 1.00 | 19.30 | 43.50 | 56.20 | 8.730000e+01 |
| under5 | 0 | 1.00 | 42.04 | 160.45 | 0.00 | 0.00 | 4.00 | 28.00 | 2.500000e+03 |
| polio | 19 | 0.99 | 82.55 | 23.43 | 3.00 | 78.00 | 93.00 | 97.00 | 9.900000e+01 |
| exp.t | 226 | 0.92 | 5.94 | 2.50 | 0.37 | 4.26 | 5.76 | 7.49 | 1.760000e+01 |
| dip | 19 | 0.99 | 82.32 | 23.72 | 2.00 | 78.00 | 93.00 | 97.00 | 9.900000e+01 |
| hiv | 0 | 1.00 | 1.74 | 5.08 | 0.10 | 0.10 | 0.10 | 0.80 | 5.060000e+01 |
| gdp | 448 | 0.85 | 7483.16 | 14270.17 | 1.68 | 463.94 | 1766.95 | 5910.81 | 1.191727e+05 |
| pop | 652 | 0.78 | 12753375.12 | 61012096.51 | 34.00 | 195793.25 | 1386542.00 | 7420359.00 | 1.293859e+09 |
| thin1.19 | 34 | 0.99 | 4.84 | 4.42 | 0.10 | 1.60 | 3.30 | 7.20 | 2.770000e+01 |
| thin5.9 | 34 | 0.99 | 4.87 | 4.51 | 0.10 | 1.50 | 3.30 | 7.20 | 2.860000e+01 |
| comp | 167 | 0.94 | 0.63 | 0.21 | 0.00 | 0.49 | 0.68 | 0.78 | 9.500000e-01 |
| school | 163 | 0.94 | 11.99 | 3.36 | 0.00 | 10.10 | 12.30 | 14.30 | 2.070000e+01 |

There are a lot of missing values for some variables, so the first thing we want to do is drop NA’s from the data set. Next, we will explore the data while dropping NA’s.

common\_theme = theme\_minimal() +   
 theme(panel.grid.minor.y = element\_blank(),  
 panel.grid.minor.x = element\_blank(),  
 panel.grid.major.y = element\_blank(),  
 panel.grid.major.x = element\_blank())

## B.2: Year

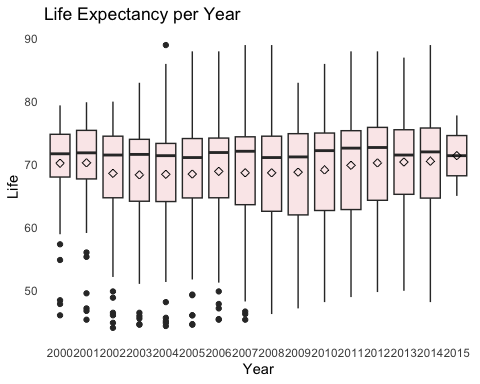
df |>  
 drop\_na() |>  
 ggplot(aes(x = year)) +  
 geom\_histogram(stat = "count", fill = "#FBEAEB", color = "black") +  
 labs(x = "Year", y = "Count", title = "Histogram of Year") +  
 common\_theme



Each year has progressively more data per year when null values are excluded. After 2011, the count flattens out. 2015 is not complete, so it should be excluded.

df |>  
 drop\_na() |>  
ggplot(mapping = aes(x = year, y = life)) +  
 geom\_boxplot(fill = "#FBEAEB") +  
 stat\_summary(fun.y=mean, geom="point", shape=23, size=2) +  
 labs(x = "Year", y = "Life", title = "Life Expectancy per Year") +  
 common\_theme

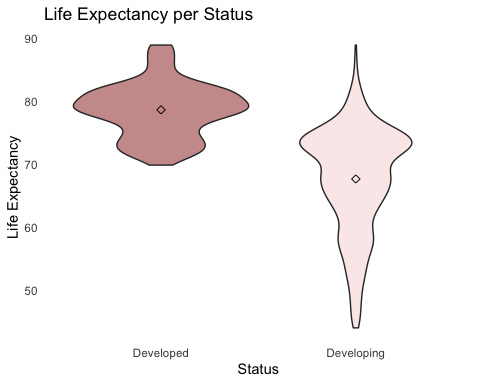
## Warning: The `fun.y` argument of `stat\_summary()` is deprecated as of ggplot2 3.3.0.  
## ℹ Please use the `fun` argument instead.



The data should be subset for a recent year, to be most applicable. 2014 is the most recent year with complete data. The boxplot is similar to other recent years, so there is no concern with choosing 2014 as our year to subset the data.

## B.3: Status

df |>  
 drop\_na() |>  
ggplot(aes(status, life, fill = status)) +   
 geom\_violin() +  
 stat\_summary(fun.y=mean, geom="point", shape=23, size=2) +  
 scale\_fill\_manual(values = c("rosybrown3", "#FBEAEB")) +  
 labs(x = "Status", y = "Life Expectancy", title = "Life Expectancy per Status") +  
 common\_theme + theme(legend.position = "none")



There is a wider range of life expectancy in developing countries. Also, the research interest is in understanding how some developing can have higher life expectancy than others. As such, the data with be further subset by filtering for developing countries only.

## B.4: Subset

df1 <- df |>  
 filter(year %in% "2014",  
 status %in% "Developing") |>  
 select(-country, -year, -status) |>  
 drop\_na()  
head(df1)

## # A tibble: 6 × 19  
## life mort inf alc exp.p hep meas bmi under5 polio exp.t dip hiv  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 59.9 271 64 0.01 73.5 62 492 18.6 86 58 8.18 62 0.1  
## 2 77.5 8 0 4.51 429. 98 0 57.2 1 98 5.88 98 0.1  
## 3 75.4 11 21 0.01 54.2 95 0 58.4 24 95 7.21 95 0.1  
## 4 51.7 348 67 8.33 24.0 64 11699 22.7 101 68 3.31 64 2   
## 5 76.2 118 8 7.93 847. 94 1 62.2 9 92 4.79 94 0.1  
## 6 74.6 12 1 3.91 296. 93 13 54.1 1 95 4.48 93 0.1  
## # … with 6 more variables: gdp <dbl>, pop <dbl>, thin1.19 <dbl>, thin5.9 <dbl>,  
## # comp <dbl>, school <dbl>

skim\_without\_charts(df1)

Data summary

|  |  |
| --- | --- |
| Name | df1 |
| Number of rows | 112 |
| Number of columns | 19 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| numeric | 19 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| life | 0 | 1 | 68.74 | 7.80 | 48.10 | 62.98 | 69.60 | 74.62 | 8.800000e+01 |
| mort | 0 | 1 | 174.47 | 111.86 | 2.00 | 109.50 | 156.50 | 242.75 | 5.220000e+02 |
| inf | 0 | 1 | 33.30 | 107.60 | 0.00 | 1.00 | 5.50 | 25.50 | 9.570000e+02 |
| alc | 0 | 1 | 2.23 | 3.23 | 0.01 | 0.01 | 0.01 | 4.12 | 1.394000e+01 |
| exp.p | 0 | 1 | 471.47 | 885.33 | 0.44 | 43.33 | 145.74 | 618.03 | 6.739680e+03 |
| hep | 0 | 1 | 80.54 | 24.17 | 2.00 | 77.00 | 89.00 | 96.00 | 9.900000e+01 |
| meas | 0 | 1 | 2376.57 | 10614.90 | 0.00 | 0.00 | 8.50 | 326.50 | 7.956300e+04 |
| bmi | 0 | 1 | 38.23 | 19.91 | 2.00 | 22.70 | 35.35 | 57.12 | 7.710000e+01 |
| under5 | 0 | 1 | 44.60 | 141.09 | 0.00 | 1.00 | 6.50 | 36.75 | 1.200000e+03 |
| polio | 0 | 1 | 81.45 | 22.00 | 8.00 | 75.75 | 91.00 | 96.00 | 9.900000e+01 |
| exp.t | 0 | 1 | 5.82 | 2.38 | 1.21 | 4.34 | 5.66 | 7.22 | 1.373000e+01 |
| dip | 0 | 1 | 81.86 | 22.99 | 2.00 | 77.75 | 91.00 | 96.25 | 9.900000e+01 |
| hiv | 0 | 1 | 0.93 | 1.66 | 0.10 | 0.10 | 0.20 | 0.70 | 9.400000e+00 |
| gdp | 0 | 1 | 4382.11 | 6704.13 | 25.45 | 528.15 | 1665.99 | 5551.47 | 4.295524e+04 |
| pop | 0 | 1 | 25640759.09 | 125975736.58 | 41.00 | 288257.25 | 1458733.50 | 13813606.25 | 1.293859e+09 |
| thin1.19 | 0 | 1 | 5.22 | 4.53 | 0.10 | 1.90 | 4.15 | 7.03 | 2.680000e+01 |
| thin5.9 | 0 | 1 | 5.50 | 4.64 | 0.10 | 1.90 | 5.20 | 7.32 | 2.740000e+01 |
| comp | 0 | 1 | 0.64 | 0.14 | 0.34 | 0.51 | 0.66 | 0.74 | 9.100000e-01 |
| school | 0 | 1 | 12.07 | 2.42 | 5.30 | 10.38 | 12.35 | 13.60 | 1.730000e+01 |

There are 112 observations in our subset of data to analyze. All variables are now numeric because we have filtered to 1 value each for 2 of the categorical variables (year and status). Country was dropped for most of our analysis because each data point represents one country, so it’s not useful in setting up regression.

# Appendix C: Which developing countries had the highest life expectancy in 2014?

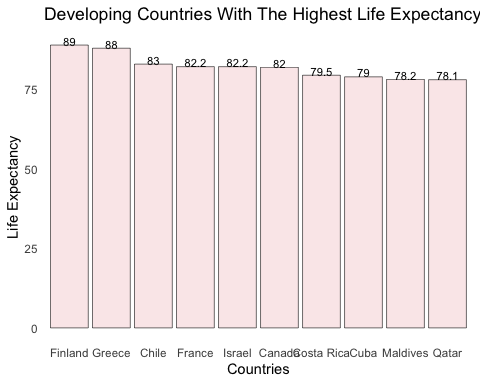
Bring country back into the data set for visualizations:

df2 <- df |>  
 filter(year %in% "2014",  
 status %in% "Developing") |>  
 select(life, mort, exp.t, hiv, comp, country) |>  
 drop\_na()

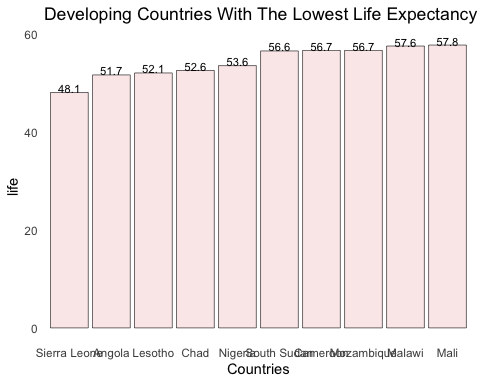
Top and bottom countries:

df2 |>  
 arrange(desc(life)) |>  
 head(10) |>  
 ggplot(aes(x = reorder(country,-life), y = life)) +  
 geom\_bar(stat = "identity", fill = "#FBEAEB", color = "black", size = 0.2) +  
 geom\_text(aes(label=life), vjust=0, color="black", size=3) +  
 labs(x = "Countries", y = "Life Expectancy",  
 title = "Developing Countries With The Highest Life Expectancy in 2014") +  
 common\_theme

## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.  
## ℹ Please use `linewidth` instead.



df2 |>  
 arrange(life) |>  
 head(10) |>  
 ggplot(aes(x = reorder(country,life), y = life)) +  
 geom\_bar(stat = "identity", fill = "#FBEAEB", color = "black", size = 0.2) +  
 geom\_text(aes(label=life), vjust=0, color="black", size=3) +  
 labs(x = "Countries",  
 title = "Developing Countries With The Lowest Life Expectancy in 2014") +  
 common\_theme



# Appendix D: Which variables best predict life expectancy?

## D.1 Stepwise Regression Variable Selection

n = nrow(df1)  
mod0 = lm(life ~ 1, data = df1)  
mod.all = lm(life ~., data = df1)  
step(mod0, scope = list(lower = mod0, upper = mod.all))

## Start: AIC=461.2  
## life ~ 1  
##   
## Df Sum of Sq RSS AIC  
## + comp 1 5072.3 1685.7 307.68  
## + school 1 3921.8 2836.3 365.96  
## + mort 1 3868.9 2889.2 368.03  
## + hiv 1 2721.6 4036.4 405.48  
## + bmi 1 2219.4 4538.7 418.61  
## + gdp 1 1805.1 4953.0 428.40  
## + exp.p 1 1487.9 5270.1 435.35  
## + alc 1 1321.4 5436.6 438.83  
## + thin5.9 1 888.0 5870.1 447.42  
## + thin1.19 1 779.6 5978.4 449.47  
## + polio 1 682.6 6075.5 451.27  
## + dip 1 501.9 6256.2 454.56  
## + hep 1 408.7 6349.3 456.21  
## + exp.t 1 359.2 6398.9 457.08  
## + under5 1 279.0 6479.1 458.48  
## + inf 1 198.6 6559.5 459.86  
## <none> 6758.1 461.20  
## + meas 1 0.5 6757.5 463.19  
## + pop 1 0.0 6758.1 463.20  
##   
## Step: AIC=307.68  
## life ~ comp  
##   
## Df Sum of Sq RSS AIC  
## + mort 1 517.6 1168.1 268.60  
## + hiv 1 485.2 1200.5 271.67  
## + exp.t 1 66.6 1619.1 305.17  
## + hep 1 43.2 1642.5 306.77  
## + polio 1 42.6 1643.1 306.82  
## <none> 1685.7 307.68  
## + dip 1 22.7 1663.0 308.16  
## + alc 1 17.1 1668.7 308.54  
## + under5 1 13.7 1672.0 308.77  
## + school 1 12.0 1673.8 308.89  
## + exp.p 1 8.2 1677.5 309.14  
## + inf 1 7.6 1678.1 309.18  
## + bmi 1 1.8 1684.0 309.57  
## + thin1.19 1 1.7 1684.1 309.57  
## + meas 1 1.0 1684.8 309.62  
## + pop 1 0.6 1685.1 309.64  
## + thin5.9 1 0.1 1685.7 309.68  
## + gdp 1 0.0 1685.7 309.68  
## - comp 1 5072.3 6758.1 461.20  
##   
## Step: AIC=268.6  
## life ~ comp + mort  
##   
## Df Sum of Sq RSS AIC  
## + hiv 1 145.15 1023.0 255.74  
## + exp.t 1 90.67 1077.5 261.55  
## + exp.p 1 37.23 1130.9 266.97  
## + hep 1 21.56 1146.6 268.51  
## <none> 1168.1 268.60  
## + dip 1 18.10 1150.0 268.85  
## + gdp 1 13.40 1154.7 269.31  
## + polio 1 9.47 1158.7 269.69  
## + bmi 1 7.96 1160.2 269.83  
## + under5 1 6.61 1161.5 269.97  
## + alc 1 4.15 1164.0 270.20  
## + thin5.9 1 3.84 1164.3 270.23  
## + inf 1 3.81 1164.3 270.23  
## + thin1.19 1 1.18 1166.9 270.49  
## + pop 1 0.49 1167.6 270.55  
## + school 1 0.06 1168.1 270.60  
## + meas 1 0.03 1168.1 270.60  
## - mort 1 517.63 1685.8 307.68  
## - comp 1 1721.08 2889.2 368.03  
##   
## Step: AIC=255.74  
## life ~ comp + mort + hiv  
##   
## Df Sum of Sq RSS AIC  
## + exp.t 1 87.80 935.17 247.69  
## + exp.p 1 44.60 978.37 252.75  
## <none> 1022.97 255.74  
## + gdp 1 18.00 1004.97 255.75  
## + hep 1 15.16 1007.81 256.07  
## + dip 1 14.49 1008.48 256.14  
## + under5 1 9.56 1013.41 256.69  
## + thin5.9 1 9.10 1013.87 256.74  
## + inf 1 6.77 1016.20 257.00  
## + bmi 1 4.10 1018.88 257.29  
## + thin1.19 1 3.43 1019.54 257.36  
## + polio 1 2.65 1020.32 257.45  
## + meas 1 2.42 1020.55 257.48  
## + alc 1 1.81 1021.16 257.54  
## + school 1 0.59 1022.38 257.68  
## + pop 1 0.09 1022.88 257.73  
## - hiv 1 145.15 1168.12 268.60  
## - mort 1 177.57 1200.55 271.67  
## - comp 1 1612.58 2635.55 359.73  
##   
## Step: AIC=247.69  
## life ~ comp + mort + hiv + exp.t  
##   
## Df Sum of Sq RSS AIC  
## <none> 935.17 247.69  
## + exp.p 1 12.75 922.42 248.15  
## + hep 1 8.33 926.85 248.69  
## + dip 1 7.12 928.05 248.83  
## + bmi 1 4.06 931.12 249.20  
## + under5 1 3.81 931.37 249.23  
## + gdp 1 2.80 932.37 249.35  
## + thin5.9 1 2.60 932.58 249.38  
## + inf 1 2.16 933.01 249.43  
## + school 1 1.04 934.14 249.57  
## + thin1.19 1 0.99 934.18 249.57  
## + alc 1 0.63 934.54 249.61  
## + meas 1 0.45 934.73 249.64  
## + pop 1 0.14 935.03 249.67  
## + polio 1 0.01 935.17 249.69  
## - exp.t 1 87.80 1022.97 255.74  
## - hiv 1 142.28 1077.45 261.55  
## - mort 1 191.52 1126.70 266.56  
## - comp 1 1461.68 2396.85 351.10

##   
## Call:  
## lm(formula = life ~ comp + mort + hiv + exp.t, data = df1)  
##   
## Coefficients:  
## (Intercept) comp mort hiv exp.t   
## 48.47452 34.50445 -0.01754 -0.89398 0.37833

The variables selected by the AIC workflow are comp, mort, hiv, exp.t

mod.aic <- lm(life ~ comp + mort + hiv + exp.t, data = df1)  
summary(mod.aic)

##   
## Call:  
## lm(formula = life ~ comp + mort + hiv + exp.t, data = df1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -10.918 -1.702 0.069 1.881 7.823   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 48.474517 2.122438 22.839 < 2e-16 \*\*\*  
## comp 34.504450 2.668108 12.932 < 2e-16 \*\*\*  
## mort -0.017539 0.003747 -4.681 8.4e-06 \*\*\*  
## hiv -0.893979 0.221573 -4.035 0.000103 \*\*\*  
## exp.t 0.378334 0.119368 3.169 0.001992 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.956 on 107 degrees of freedom  
## Multiple R-squared: 0.8616, Adjusted R-squared: 0.8564   
## F-statistic: 166.6 on 4 and 107 DF, p-value: < 2.2e-16

R-squared for the AIC model is 0.8616, so 86% of variation is explained by the AIC model.

## D.2 Best Subset Regression Variable Selection

xmat = df1 |>  
select(-life) |>  
select\_if(is.numeric)  
dim(xmat)

## [1] 112 18

There are 18 numeric variables up for selection

mod = regsubsets(xmat, df1$life, nvmax = 18)  
summary.mod = summary(mod)  
summary.mod$which

## (Intercept) mort inf alc exp.p hep meas bmi under5 polio exp.t  
## 1 TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## 2 TRUE TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## 3 TRUE TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  
## 4 TRUE TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE TRUE  
## 5 TRUE TRUE FALSE FALSE TRUE FALSE FALSE FALSE FALSE FALSE TRUE  
## 6 TRUE TRUE TRUE FALSE FALSE FALSE FALSE FALSE TRUE FALSE TRUE  
## 7 TRUE TRUE TRUE FALSE TRUE FALSE FALSE FALSE TRUE FALSE TRUE  
## 8 TRUE TRUE TRUE FALSE TRUE FALSE FALSE FALSE TRUE FALSE TRUE  
## 9 TRUE TRUE TRUE FALSE TRUE TRUE FALSE FALSE TRUE FALSE TRUE  
## 10 TRUE TRUE TRUE FALSE TRUE TRUE FALSE FALSE TRUE FALSE TRUE  
## 11 TRUE TRUE TRUE FALSE TRUE TRUE FALSE FALSE TRUE FALSE TRUE  
## 12 TRUE TRUE TRUE FALSE TRUE TRUE TRUE FALSE TRUE FALSE TRUE  
## 13 TRUE TRUE TRUE FALSE TRUE TRUE TRUE FALSE TRUE TRUE TRUE  
## 14 TRUE TRUE TRUE FALSE TRUE TRUE TRUE FALSE TRUE TRUE TRUE  
## 15 TRUE TRUE TRUE FALSE TRUE TRUE TRUE TRUE TRUE TRUE TRUE  
## 16 TRUE TRUE TRUE FALSE TRUE TRUE TRUE TRUE TRUE TRUE TRUE  
## 17 TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE  
## 18 TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE  
## dip hiv gdp pop thin1.19 thin5.9 comp school  
## 1 FALSE FALSE FALSE FALSE FALSE FALSE TRUE FALSE  
## 2 FALSE FALSE FALSE FALSE FALSE FALSE TRUE FALSE  
## 3 FALSE TRUE FALSE FALSE FALSE FALSE TRUE FALSE  
## 4 FALSE TRUE FALSE FALSE FALSE FALSE TRUE FALSE  
## 5 FALSE TRUE FALSE FALSE FALSE FALSE TRUE FALSE  
## 6 FALSE TRUE FALSE FALSE FALSE FALSE TRUE FALSE  
## 7 FALSE TRUE FALSE FALSE FALSE FALSE TRUE FALSE  
## 8 FALSE TRUE TRUE FALSE FALSE FALSE TRUE FALSE  
## 9 FALSE TRUE TRUE FALSE FALSE FALSE TRUE FALSE  
## 10 FALSE TRUE TRUE FALSE TRUE FALSE TRUE FALSE  
## 11 FALSE TRUE TRUE TRUE TRUE FALSE TRUE FALSE  
## 12 FALSE TRUE TRUE TRUE TRUE FALSE TRUE FALSE  
## 13 FALSE TRUE TRUE TRUE TRUE FALSE TRUE FALSE  
## 14 FALSE TRUE TRUE TRUE TRUE FALSE TRUE TRUE  
## 15 FALSE TRUE TRUE TRUE TRUE TRUE TRUE FALSE  
## 16 FALSE TRUE TRUE TRUE TRUE TRUE TRUE TRUE  
## 17 FALSE TRUE TRUE TRUE TRUE TRUE TRUE TRUE  
## 18 TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE

summary.mod$rsq #check R^2

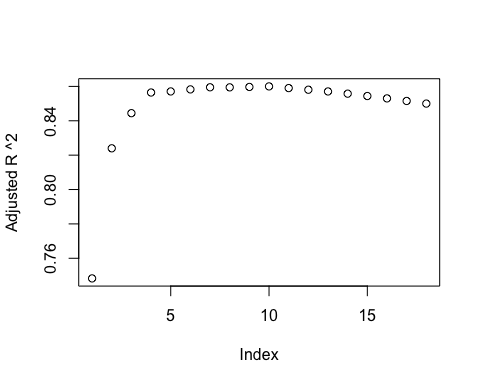
## [1] 0.7505579 0.8271519 0.8486295 0.8616211 0.8635083 0.8659193 0.8683494  
## [8] 0.8695786 0.8710567 0.8726119 0.8729950 0.8734140 0.8738031 0.8739667  
## [15] 0.8740737 0.8742070 0.8742664 0.8743277

R-squared levels out at the model with 4 predictors

summary.mod$adjr2 #check adjusted R^2

## [1] 0.7482903 0.8239804 0.8444247 0.8564480 0.8570700 0.8582575 0.8594883  
## [8] 0.8594488 0.8596793 0.8599992 0.8590245 0.8580702 0.8570627 0.8557764  
## [15] 0.8543977 0.8530208 0.8515274 0.8500041

`Adjusted R ^2` <- summary.mod$adjr2  
plot(`Adjusted R ^2`)

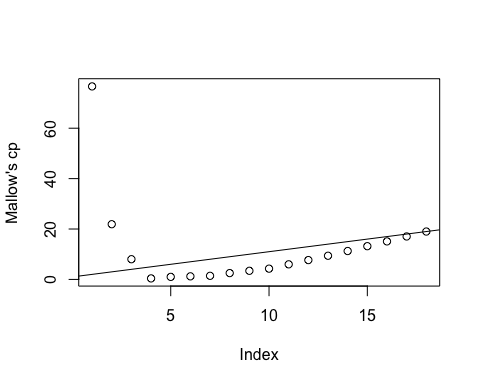


Adjusted R-squared levels out at the model with 4 predictors

summary.mod$cp

## [1] 76.5921736 21.9110780 8.0172448 0.4031971 1.0066214 1.2224209  
## [7] 1.4241162 2.5144551 3.4206729 4.2697585 5.9862349 7.6762262  
## [13] 9.3882377 11.2671548 13.1880066 15.0893862 17.0453657 19.0000000

`Mallow's cp` <- summary.mod$cp  
plot(`Mallow's cp`)   
abline(1,1)



The model with 4 variables is the simplest model with a Cp value (0.403) lower than p (19). The model with 4 variables includes comp, mort, hiv, and exp.t. So the model selected by best subset regression is the same as the model selected by stepwise regression.

# Appendix E: Are there interactions between variables used to predict life expectancy?

## E.1 Check interactions

add1(mod.aic, ~.+comp\*mort+comp\*hiv+comp\*exp.t, test = 'F')

## Single term additions  
##   
## Model:  
## life ~ comp + mort + hiv + exp.t  
## Df Sum of Sq RSS AIC F value Pr(>F)   
## <none> 935.17 247.69   
## comp:mort 1 11.548 923.63 248.30 1.3253 0.252230   
## comp:hiv 1 89.125 846.05 238.47 11.1663 0.001151 \*\*  
## comp:exp.t 1 70.193 864.98 240.95 8.6019 0.004116 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## E.2 Update model

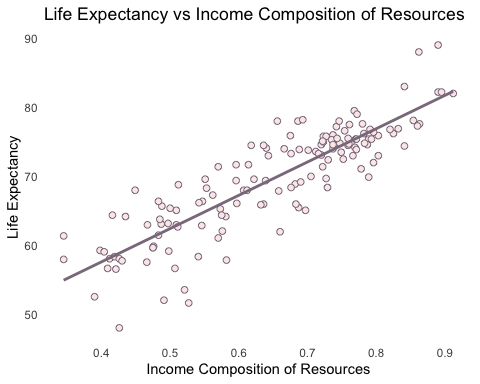
mod = update(mod.aic, ~.+comp:exp.t+comp:hiv)  
summary(mod)

##   
## Call:  
## lm(formula = life ~ comp + mort + hiv + exp.t + comp:exp.t +   
## comp:hiv, data = df1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -8.4418 -1.6289 0.1997 1.8930 6.0646   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 54.239712 3.547274 15.291 < 2e-16 \*\*\*  
## comp 25.258188 5.367988 4.705 7.76e-06 \*\*\*  
## mort -0.017148 0.003489 -4.915 3.29e-06 \*\*\*  
## hiv 2.436576 1.096983 2.221 0.02849 \*   
## exp.t -0.946437 0.503578 -1.879 0.06296 .   
## comp:exp.t 2.085222 0.773651 2.695 0.00819 \*\*   
## comp:hiv -6.470091 2.069247 -3.127 0.00229 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.745 on 105 degrees of freedom  
## Multiple R-squared: 0.8829, Adjusted R-squared: 0.8762   
## F-statistic: 132 on 6 and 105 DF, p-value: < 2.2e-16

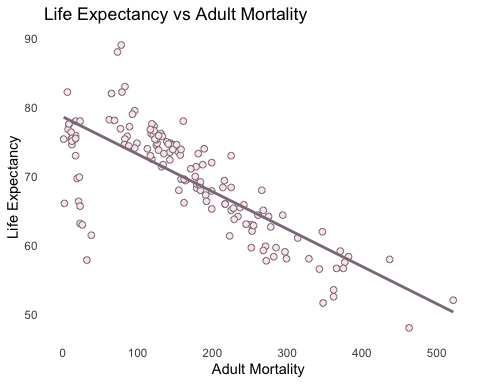
## E.3 Model assumptions

Check linearity:

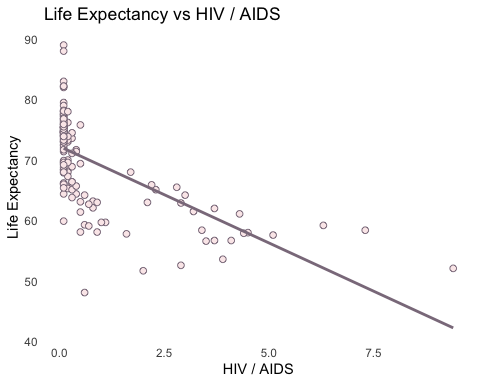
df2 |>  
 ggplot(aes(x = comp, y = life, label = country)) +  
 geom\_point(shape = 21, color = "thistle4", fill = "#FBEAEB", size = 2) +  
 geom\_smooth(method = "lm", se = FALSE, color = "thistle4") +   
 labs(x = "Income Composition of Resources", y = "Life Expectancy",   
 title = "Life Expectancy vs Income Composition of Resources") +  
 common\_theme



df2 |>  
 ggplot(aes(x = mort, y = life, label = country)) +  
 geom\_point(shape = 21, color = "thistle4", fill = "#FBEAEB", size = 2) +  
 geom\_smooth(method = "lm", se = FALSE, color = "thistle4") +   
 labs(x = "Adult Mortality", y = "Life Expectancy",   
 title = "Life Expectancy vs Adult Mortality") +  
 common\_theme



df2 |>  
 ggplot(aes(x = hiv, y = life, label = country)) +  
 geom\_point(shape = 21, color = "thistle4", fill = "#FBEAEB", size = 2) +  
 geom\_smooth(method = "lm", se = FALSE, color = "thistle4") +   
 labs(x = "HIV / AIDS", y = "Life Expectancy", title = "Life Expectancy vs HIV / AIDS") +  
 common\_theme

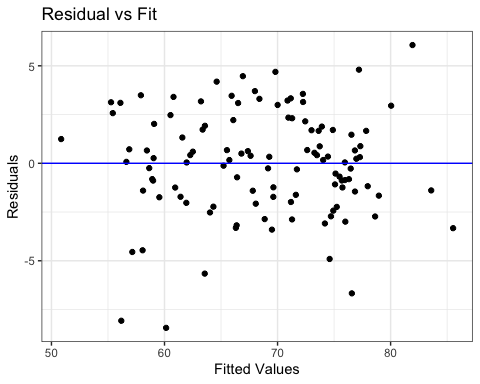


df2 |>  
 ggplot(aes(x = exp.t, y = life, label = country)) +  
 geom\_point(shape = 21, color = "thistle4", fill = "#FBEAEB", size = 2) +  
 geom\_smooth(method = "lm", se = FALSE, color = "thistle4") +   
 labs(x = "Total Expenditure", y = "Life Expectancy",   
 title = "Life Expectancy vs Total Expenditure") +  
 common\_theme



Check variance and linearity:

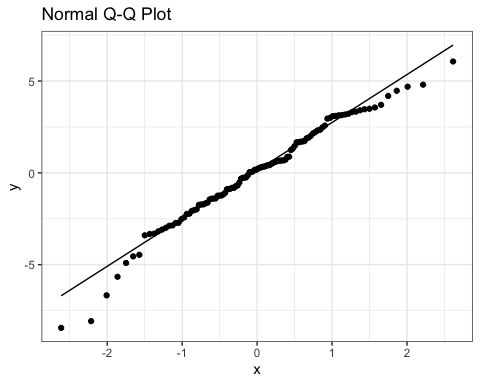
model.table = augment(mod)  
ggplot(model.table, aes(x = .fitted, y = .resid)) +  
 geom\_point() +  
 geom\_hline(yintercept = 0, colour = 'blue') +  
 labs(x = 'Fitted Values', y = 'Residuals') +  
 ggtitle('Residual vs Fit') +  
 theme\_bw()



There are no issues with the variance or linearity assumption.

Check normality:

ggplot(model.table, aes(sample = .resid)) +  
 stat\_qq() +  
 stat\_qq\_line() +  
 ggtitle('Normal Q-Q Plot') +  
 theme\_bw()

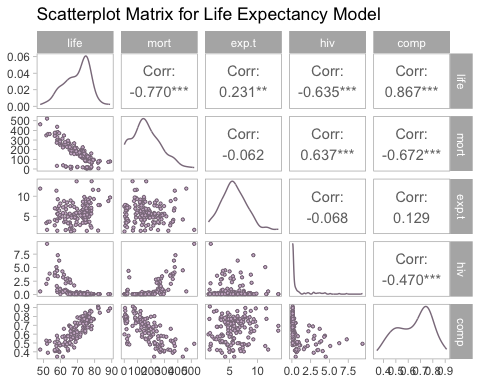


shapiro.test(resid(mod))

##   
## Shapiro-Wilk normality test  
##   
## data: resid(mod)  
## W = 0.97829, p-value = 0.06495

# Appendix F: Are the predictors positive or negatively correlated to life expectancy?

df2 |>  
 select(-country) |>  
 ggpairs(  
 lower = list(continuous = wrap("points", shape = 21, fill = "thistle3",   
 color = "thistle4", size = 0.8)),  
 diag = list(continuous = wrap("densityDiag", color = "thistle4")),  
 title = "Scatterplot Matrix for Life Expectancy Model") +  
 theme\_light() + theme(panel.grid.minor.y = element\_blank(),  
 panel.grid.minor.x = element\_blank(),  
 panel.grid.major.y = element\_blank(),  
 panel.grid.major.x = element\_blank())



# Appendix G: Make a 95% point prediction for the life expectancy of a country with predictor values as the mean response of each predictor.

new <- df2 |>  
 summarize(mort = mean(mort),  
 hiv = mean(hiv),  
 comp = mean(comp),  
 exp.t = mean(exp.t)) |>  
 data.frame()  
pi = predict(mod, new, interval = "prediction", level = 0.95)  
pi

## fit lwr upr  
## 1 68.79102 63.31005 74.27199

# Appendix H: Summary Table

df2 |>  
 select(country, life, mort,hiv,comp,exp.t) |>  
 rename("Country" = country, "Life" = life, "Mort" = mort,  
 "Comp" = comp, "Exp.T" = exp.t) |>  
 arrange(desc(Life)) |>  
 head(10) |>  
 gt(rowname\_col = "Country") |>  
 tab\_header(title = md("Summary of \*\*Life Expectancy\*\*")) |>  
 tab\_stubhead(label = md("Country"))

Table 1: Summary of <w:rPr><w:b w:val="true"></w:b></w:rPr>Life Expectancy<w:rPr><w:b w:val="false"></w:b></w:rPr>

| Country | Life | Mort | hiv | Comp | Exp.T |
| --- | --- | --- | --- | --- | --- |
| Finland | 89.0 | 78 | 0.1 | 0.890 | 9.68 |
| Greece | 88.0 | 73 | 0.1 | 0.862 | 8.80 |
| Chile | 83.0 | 83 | 0.1 | 0.841 | 7.79 |
| France | 82.2 | 79 | 0.1 | 0.890 | 11.54 |
| Israel | 82.2 | 6 | 0.1 | 0.895 | 7.81 |
| Canada | 82.0 | 65 | 0.1 | 0.912 | 1.45 |
| Costa Rica | 79.5 | 96 | 0.1 | 0.768 | 9.31 |
| Cuba | 79.0 | 93 | 0.1 | 0.772 | 11.60 |
| Maldives | 78.2 | 62 | 0.1 | 0.693 | 13.73 |
| Qatar | 78.1 | 69 | 0.1 | 0.854 | 2.19 |