Linear Regression of Life Expectancy

By:

Christina Lopez
(Section 01)

Andres Gonzalez
(Section 01)

Katya Mora
(Section 01)

I. Introduction

Health-related data from the World Health Organization repository was merged with economic-related data from the United Nations website. The data was made public in order to help countries improve their citizens' life expectancy. The data set contains 22 variables to summarize data for 193 countries between the years of 2000 and 2015. The variables include life expectancy (the response), country and status (identifying information), health-related data, and economic-related data (potential predictors). The analysis within this report will focus on developing countries in the year 2014, which results in a subset of data containing 112 observations.

The research interests of this report are: identifying countries with the highest life expectancy, building a model for life expectancy using variable selection procedures, understanding the effects of interaction, understanding the correlation between predictors and response, and predicting life expectancy using the model. It will be shown that there are 4 variables and 2 interaction terms that can adequately predict life expectancy in the context of this data set. The following regression model is obtained:

$$life\ = 54.24 + 25.26comp - 0.017mort + 2.44hiv - 0.95exp.t$$
$$+2.09comp{:}\,exp.t - 6.47comp{:}\,hiv$$

Where

- life = Life Expectancy: Age the general population is expected to live to
- comp = Income composition of resources: Human development index from 0 - 1
- mort = Adult mortality: Probability of dying between 15 - 60 years per 1,000 adults
- hiv = HIV/AIDS: Deaths per 1,000 live births HIV/AIDS 0-4 years
- exp.t = Total expenditure: Expenditure on health as a % of total expenditure)
- comp:exp.t = interaction between income composition with total expenditure
- comp:hiv = interaction between income composition with HIV/AIDS

II. Questions of Interest

1. Which developing countries had the highest life expectancy in 2014?
2. Which variables best predict life expectancy?
3. Are there interactions between variables in the model?
4. Are the predictors positive or negatively correlated to life expectancy?
5. Make a 95% point prediction for the life expectancy of a country with predictor values as the mean response of each predictor.

III. Regression Method

To answer the questions of interest, the data set was ordered from highest to lowest life expectancy. A bar chart displaying the first 10 countries in the ordered list provided a useful visual for displaying the top countries in terms of life expectancy. Next, the original un-ordered data set was used in the workflow for stepwise regression variable selection. It was proven that the results of best subset regression variable selection resulted in the same predictor variables. Once the variables for the regression model were defined, an F test was conducted to decide whether adding interactions to the model had a significant effect or not.

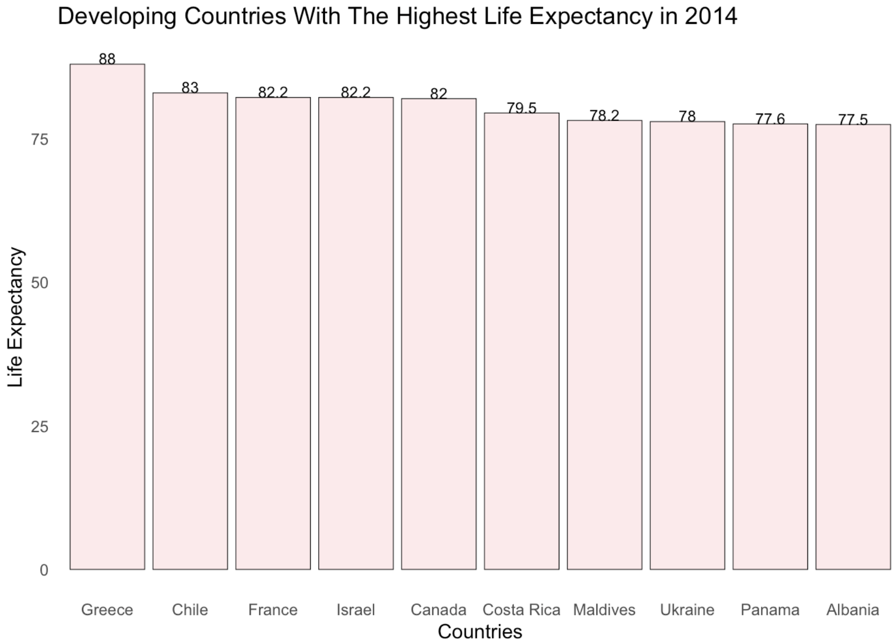
All variables involved in the model were then graphed on a scatterplot matrix with correlation values to determine the sign of correlation for each predictor with the response. Finally, a prediction interval on the final model resulted in the point prediction of interest.

IV. Regression Analysis, Results and Interpretation

Question 1

Which developing countries had the highest life expectancy in 2014?

The country with the highest life expectancy was Greece, which had a life expectancy of 88 years of age. To find this, the data set was arranged by the variable 'Life Expectancy' in descending order. The bar chart below shows the resulting top 10 countries by highest life expectancy.



Question 2

Which variables best predict life expectancy?

To start variable selection, the variables of interest must be identified. Recall from the introduction that we start with 22 variables. In the context of the subset of data chosen to analyze, the year was held constant as 2014 and the status was held constant as “developing”, so these two variables are not of interest. Each data point represents data for a unique country, so the name of the country is also not of interest. Lastly, life expectancy is the response, so it is not a variable of interest. Therefore, there were 18 variables of interest to use in a variable selection procedure. To determine which variables best predict life expectancy, it will be shown that stepwise regression variable selection was conducted on all variables of interest.

As a result of AIC workflow for stepwise regression in R (Appendix D.1), life expectancy can be predicted by using 4 of the 18 available variables: comp (Income Composition of Resources), mort (Adult Mortality), hiv (HIV/AIDS) and exp.t (Total Expenditure). These 4 variables in that particular order produced the lowest AIC value of 247.69. To understand the significance of the variables, a summary of the model was produced. With R’s summary of the AIC model in Appendix D.1, a t-test was conducted on each variable to test the following hypotheses with α = 0.05:

$H_0: \beta_{comp} = 0$ vs $H_1: \beta_{comp} \neq 0$. The t statistic is 12.932 and the p-value is < 2E-16.

$H_0: \beta_{mort} = 0$ vs $H_1: \beta_{mort} \neq 0$. The t-statistic is -4.681 and the p-value is 8.4E-6.

$H_0: \beta_{hiv} = 0$ vs $H_1: \beta_{hiv} \neq 0$. The t-statistic is -4.035 and the p-value is 1.03E-4.

$H_0: \beta_{exp.t} = 0$ vs $H_1: \beta_{exp.t} \neq 0$. The t-statistic is 3.169 and the p-value is 1.99E-3.

It can be seen that the p-values of all predictors in the model are < α = 0.05. Therefore, all predictors are significant and all were included in the model. Lastly, it is noted that the R-squared value is .8616, which shows that 86% of the variation in life expectancy is explained by these 4 predictors.

To verify if these 4 variables would result in the best model, the results of the stepwise regression were compared to the results of best subset regression. The regsubsets function in the leaps package resulted in 18 models to analyze. It can be seen in Appendix D.2 that the model selected by best subset regression is the same model as the one selected by stepwise regression.

Question 3

Are there interactions between variables used to predict life expectancy?

Yes, there are interaction terms added into the model used to predict life expectancy. The variable comp was shown to be the most significant predictor, so it was used for interaction comparison. The add1 function in R was used to conduct the following F tests on adding each interaction term to the AIC model:

H_0: $\beta_{comp:mort} = 0$ vs H_1: $\beta_{comp:mort} \neq 0$. The-F statistic is 1.325 and the p-value is 0.252.

H_0: $\beta_{comp:hiv} = 0$ vs H_1: $\beta_{comp:hiv} \neq 0$. The F-statistic is 11.166 and the p-value is 0.001.

H_0: $\beta_{comp:exp.t} = 0$ vs H_1: $\beta_{comp:exp.t} \neq 0$. The F-statistic is 8.602 and the p-value is 0.004.

Therefore, only two interaction terms were significant: comp:hiv and comp:exp.t. This shows that the effect of income composition of resources on life expectancy depends on HIV/AIDS and total expenditure. Therefore, the two interaction terms were included in the model.

The last step is to check the significance of each predictor in the model after adding the interaction terms. R conducts the following t-tests on each of the predictors:

H_0: $\beta_{comp} = 0$ vs H_1: $\beta_{comp} \neq 0$. The t statistic is 4.702 and the p-value is 7.76E-6.

H_0: $\beta_{mort} = 0$ vs H_1: $\beta_{mort} \neq 0$. The t-statistic is -4.915 and the p-value is 3.29E-6.

H_0: $\beta_{hiv} = 0$ vs H_1: $\beta_{hiv} \neq 0$. The t-statistic is 2.221 and the p-value is 0.028.

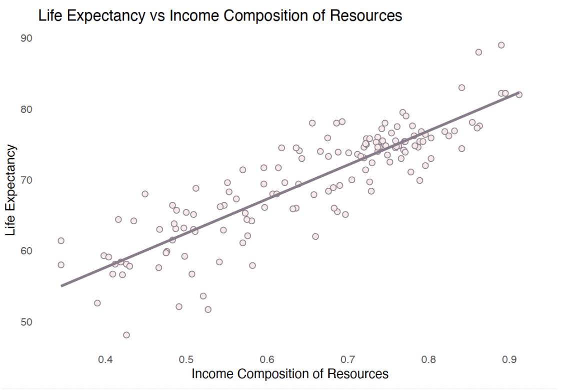
H_0: $\beta_{exp.t} = 0$ vs H_1: $\beta_{exp.t} \neq 0$. The t-statistic is -1.879 and the p-value is 0.0630.

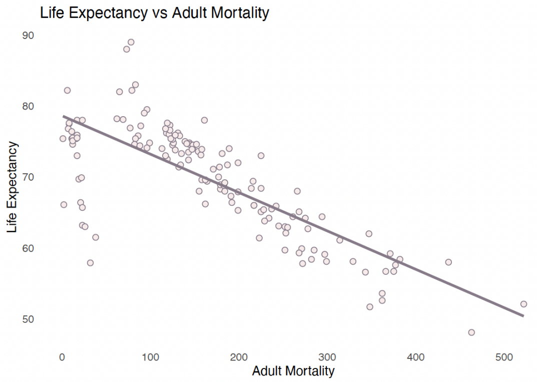
H_0: $\beta_{comp:exp.t} = 0$ vs H_1: $\beta_{comp:exp.t} \neq 0$. The t-statistic is 2.695 and the p-value is 0.008.

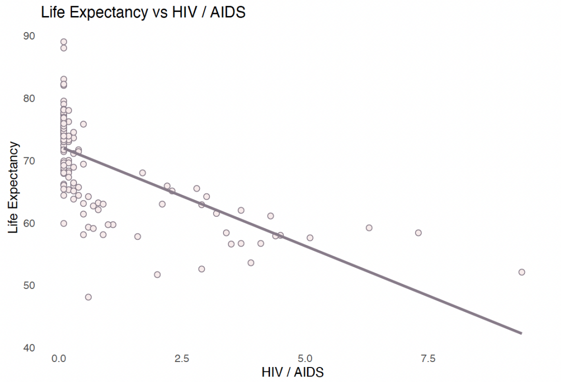
H_0: $\beta_{comp:hiv} = 0$ vs H_1: $\beta_{comp:hiv} \neq 0$. The t-statistic is -3.127 and the p-value is 0.002.

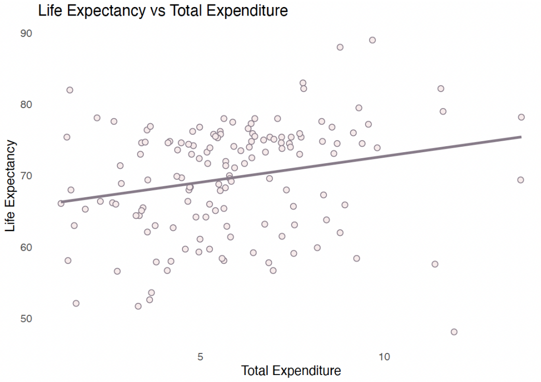
It can be noted that all predictors and interaction terms are significant. The predictor exp.t stands out by having a p - value of 0.063, which is greater than the usual α = 0.05. However, for stepwise regression variable selection, α = 0.15 can be used, so the variable was included in the model. The coefficient of determination for the full model was 0.8829. This shows that 88.29% of variation in a country's life expectancy is explained by variation in the country's income composition, adult mortality, HIV/AIDS, total expenditure, the interaction between income composition and total expenditure, and the interaction between income composition and HIV/AIDS.

Once the full model was defined, LINE conditions must be checked. To start with, the following 4 plots show that there are linear relationships between life expectancy and each of the predictors:

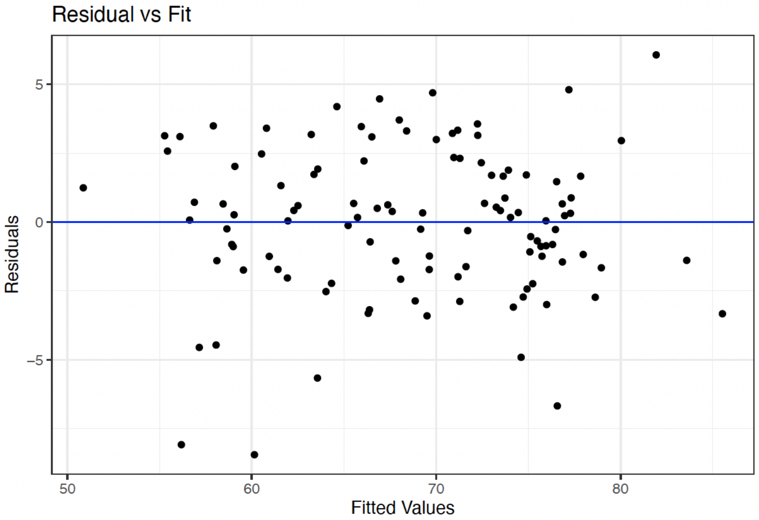




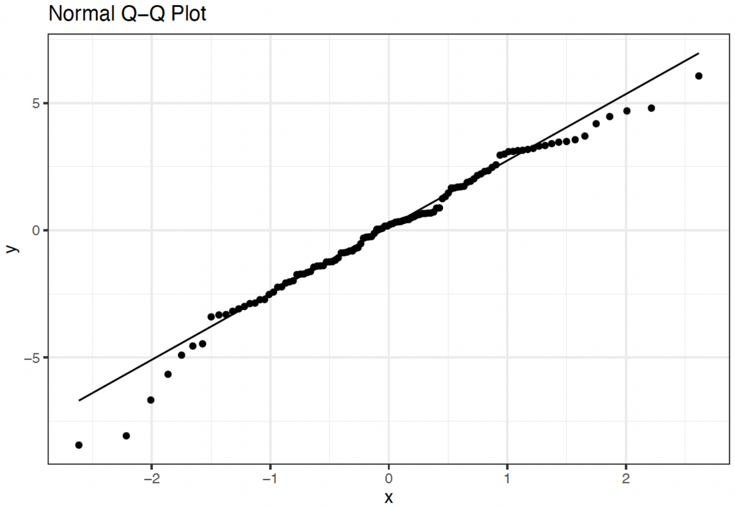




Next, the equal variance and linearity assumptions of the full model were checked by analyzing the residual vs fit plot shown below. The plot is well behaved, so there are no concerns with equal variance or linearity. As such, no transformations were required for the model at this point.



Next, independence and normality assumptions were checked. It can reasonably be assumed that the data is independent because each data point corresponds to a unique country, and the year has been held constant as 2014. Normality was checked with the following Q-Q plot.



The plot shows long tails on both sides, which might suggest non-normality. However, a Shapiro-Wilk test was conducted to test the null hypothesis that the residuals are normally distributed, against the alternative hypothesis that they are not. The p-value for the Shapiro test was 0.0649, which is > α = 0.05. Therefore, the null hypothesis cannot be rejected and it can be concluded with 0.05 significance that the residuals are normally distributed. Therefore, no transformations are required for the model.

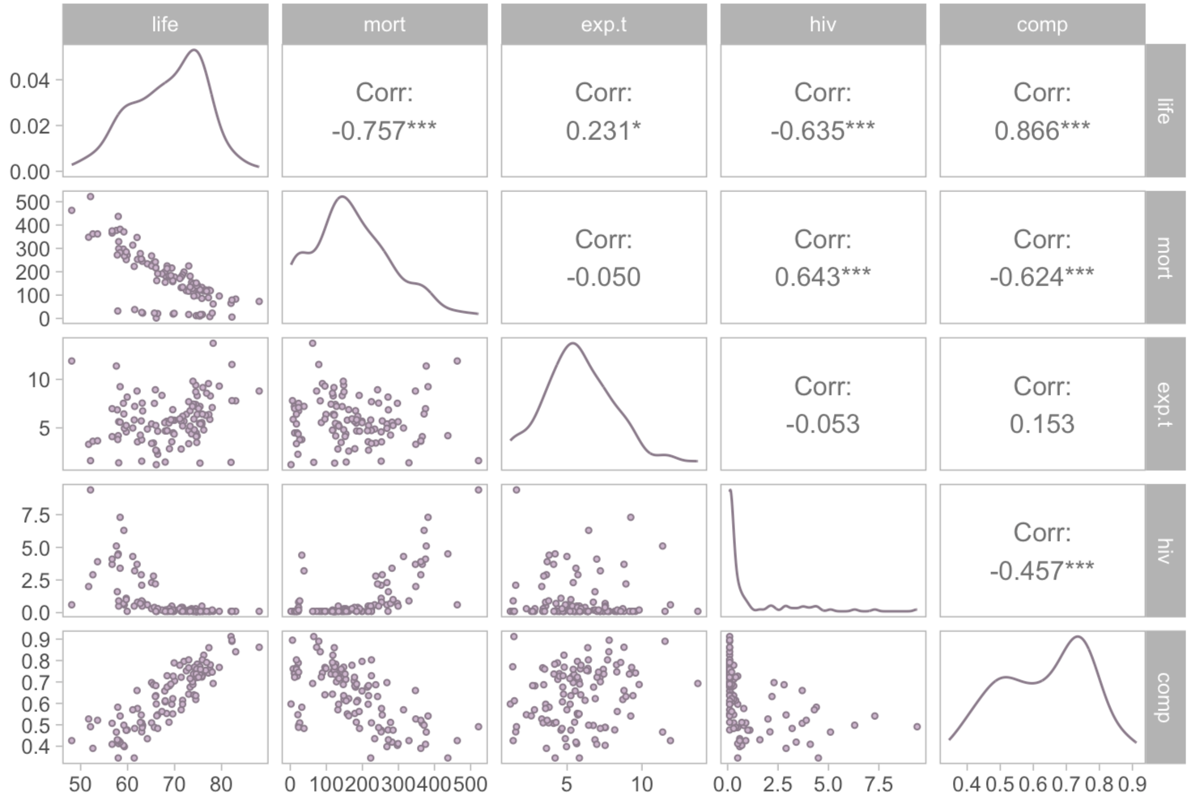
In conclusion, there were 4 variables and 2 interaction terms that adequately predicted life expectancy of developing countries in 2014v(summarized in Appendix E.2). It can be seen that for every unit increase in adult mortality, life expectancy decreases by 0.017 years. For every unit increase in HIV/AIDS, life expectancy increases by 2.44 years. However, this is offset by the comp:hiv coefficient, since the impact of income composition on life expectancy decreases by 6.47 years, based on the results of HIV/AIDS. Lastly, for every unit increase in total expenditure, life expectancy decreases by 0.946. However, this is offset by the comp:exp.t coefficient, since the impact of income composition on total expenditure increases by 2.08 years, based on the results of total expenditure.

Question 4

Are the predictors positive or negatively correlated to life expectancy?

To show which predictors are positively or negatively correlated to life expectancy a scatterplot matrix was created. The first row in this matrix shows how life expectancy is correlated to the 4 predictors. Mortality and HIV/AIDS have negative correlations with values -0.770 and -0.635 respectively. While Expenditure and Income composition of resources have positive correlations with values 0.231 and 0.867 respectively.

Scatterplot Matrix for Life Expectancy Model



<u>Question 5</u>

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

To start with, mean values for each predictor in the model were calculated. The mean value for mort was 174.473, and that for hiv was 0.930. The mean of comp was 0.636, and lastly the mean of exp.t was 5.823.

R calculated the prediction interval using these values along with the full model with interaction (Appendix G). It can be seen from the results that such a country is predicted to have a life expectancy of 67.9 years, with a 95% prediction interval of life expectancy being between 62.5 years and 73.5 years.

V. Conclusion

It is now known that the developing country with the highest life expectancy in 2014 was Greece. Additionally, it was determined that life expectancy can be significantly predicted in developing countries in 2014 using data from the WHO repository and United Nations website. It was found that not all variables in the data set were significant with the response of interest.

Predictors were obtained and interaction terms were shown to significantly increase the effectiveness of the model. The 4 variables that predict life expectancy are income composition of resources, adult mortality, HIV/AIDS, and total expenditure. Additionally, the effect of income composition on life expectancy was shown to depend on the values for HIV/AIDS and total expenditure. As such, about 88% of the variation in life expectancy is explained by these 4 predictors and 2 interaction terms. Furthermore, 2 of the predictors are economic-related variables, for which life expectancy will increase as the variables increase. In addition, 2 of the variables are health-related data, for which life expectancy will decrease as the variables increase. Importantly, we did not need to transform the model since the LINE assumptions were met. Based on the model obtained, life expectancy is expected to be 67.9 years for a country that has the mean of each predictor variable.

The results of the model can be used by developing countries to determine which sections of their economy and healthcare they should focus on to improve the life expectancy of its citizens. However, the data was from 8 years ago and was from before the global pandemic and major supply chain disruption in recent years. The analysis could be improved and made more applicable by including data from more recent years.

VI. Appendix

Contents:

Appendix

2022-12-05

Appendix A: Data Cleaning

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

```
library(tidyverse)
library(skimr)
library(GGally)
library(readr)
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 x 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)
```

Table 1: 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

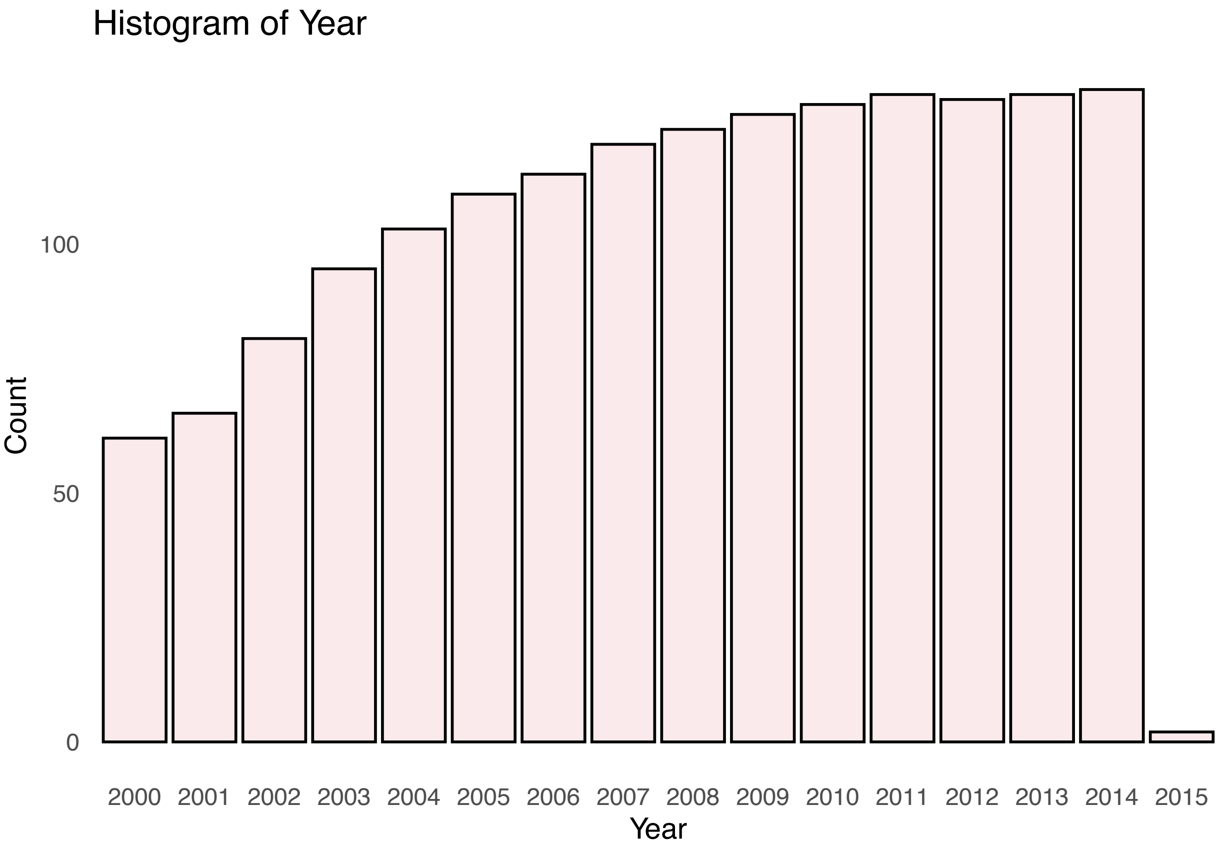
skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
exp.p	0	1.00	738.25	1987.91	0.00	4.69	64.91	441.53	1.947991e+04
hep	553	0.81	80.94	25.07	1.00	77.00	92.00	97.00	9.900000e+01
meas	0	1.00	2419.59	11467.27	0.00	0.00	17.00	360.25	2.121830e+05
bmi	34	0.99	38.32	20.04	1.00	19.30	43.50	56.20	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.

```r
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

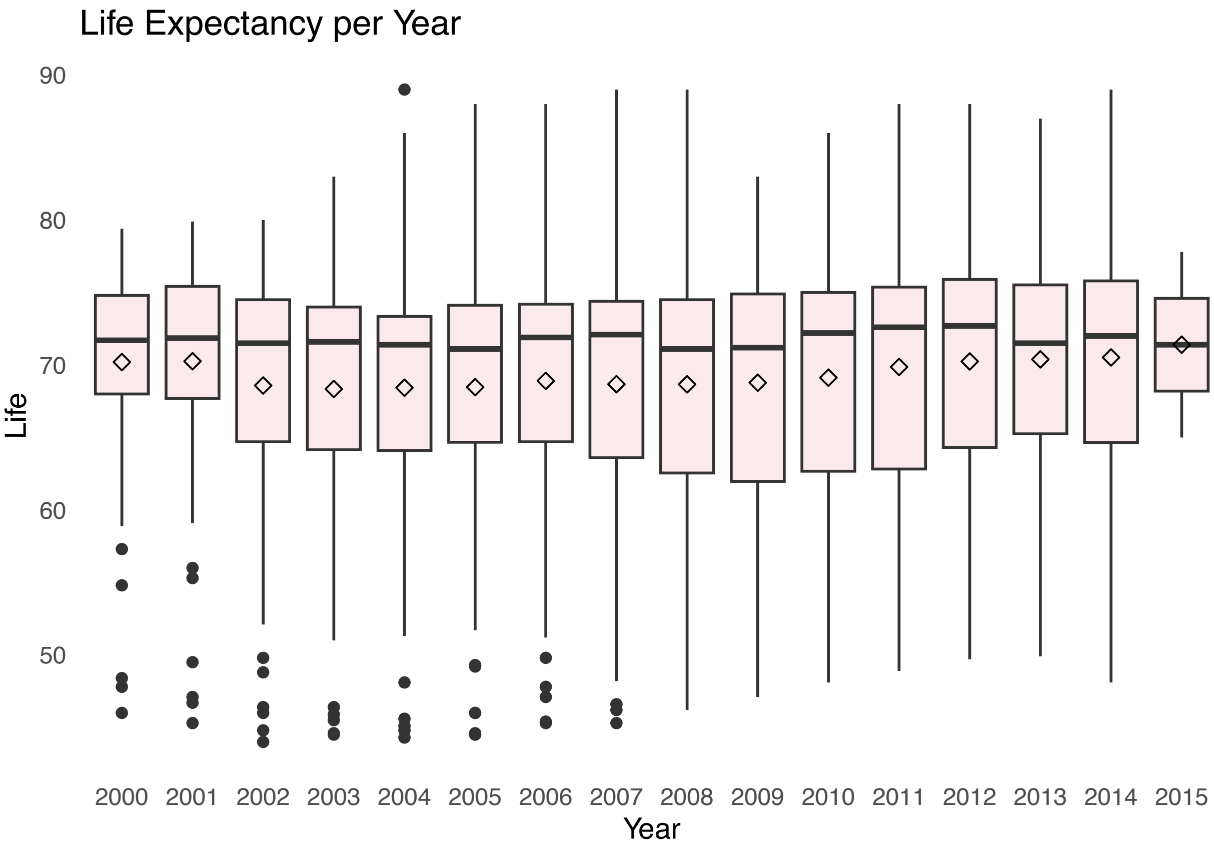
```r
df |>
  drop_na() |>
  ggplot(aes(x = year)) +
  geom_histogram(stat = "count", fill = "#FBEAEB", color = "black") +
  labs(x = "Year", y = "Count", title = "Histogram of Year") +
  common_theme
```



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
```

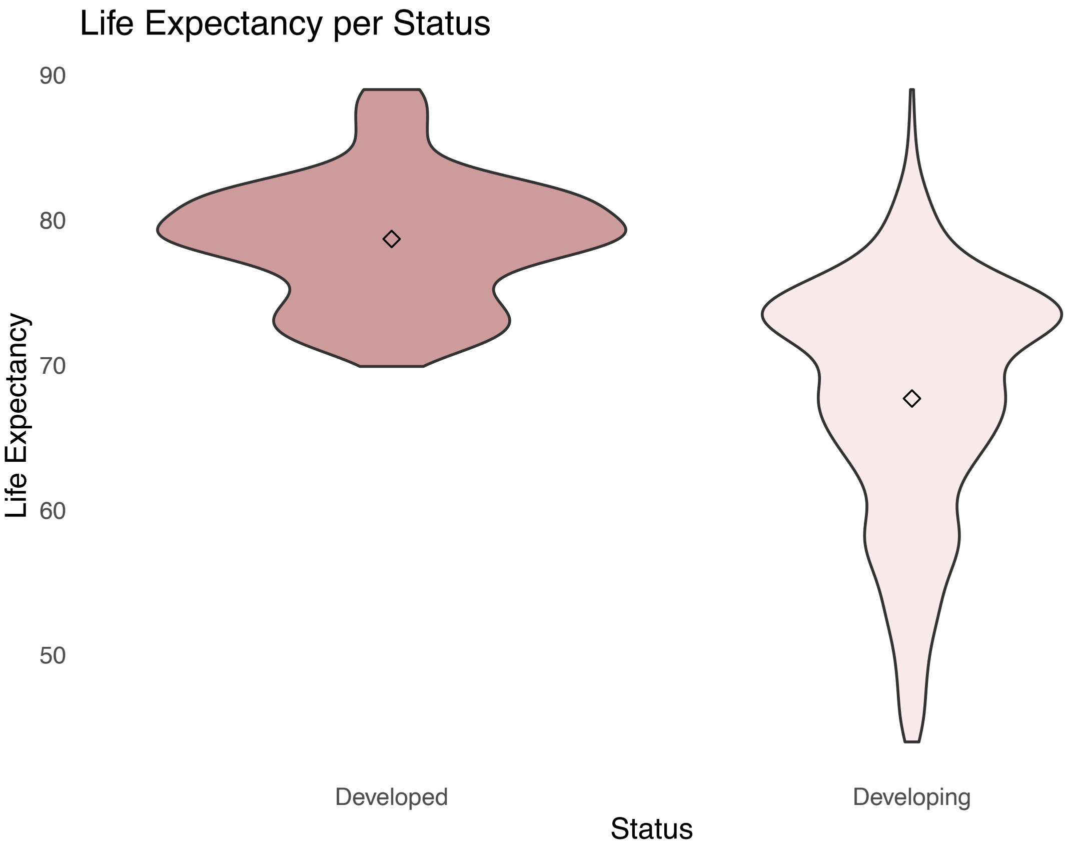
```
## Warning: The `fun.y` argument of `stat_summary()` is deprecated as of ggplot2 3.3.0.
## i Please use the `fun` argument instead.
```



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") |>
  drop_na() |>
  select(-country, -year, -status)


head(df1)
```

```
## # A tibble: 6 x 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)
```

Table 5: 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?

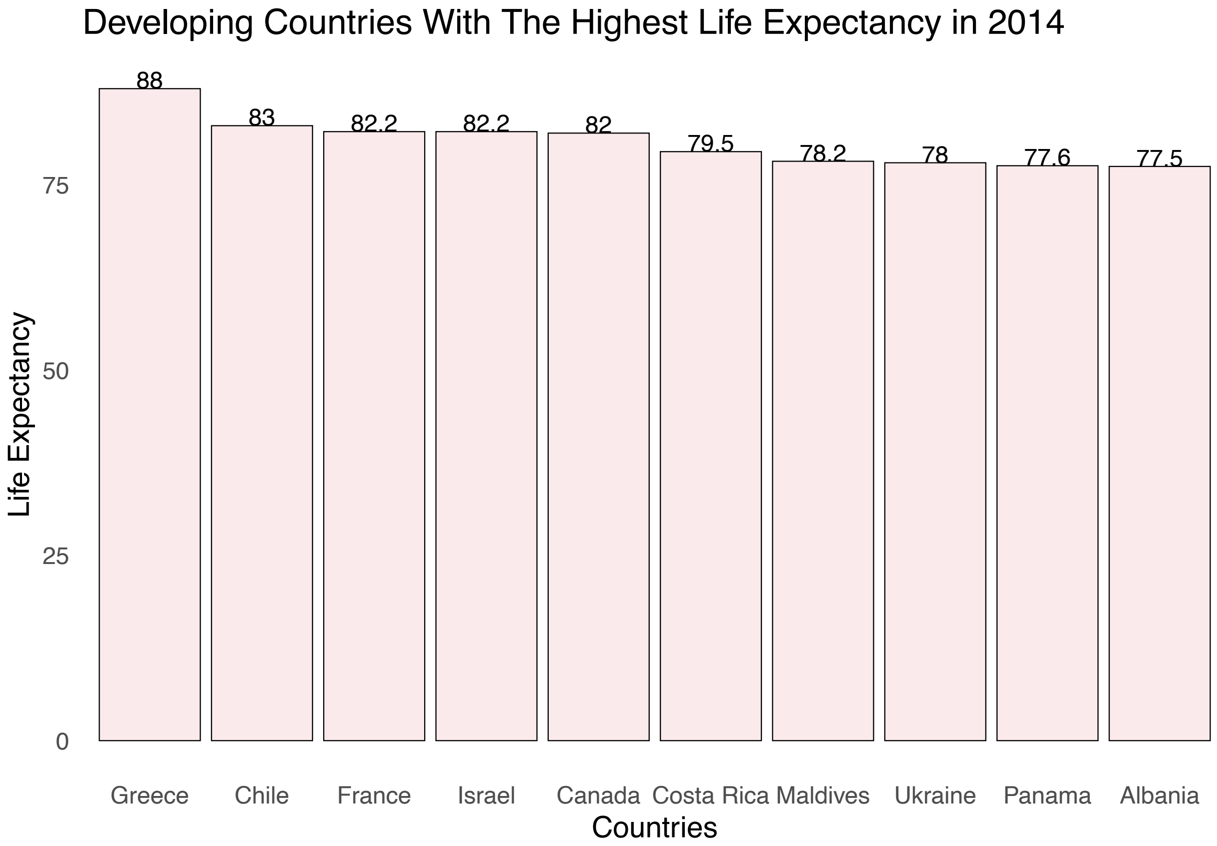
Bring country back into the data set for visualizations:

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

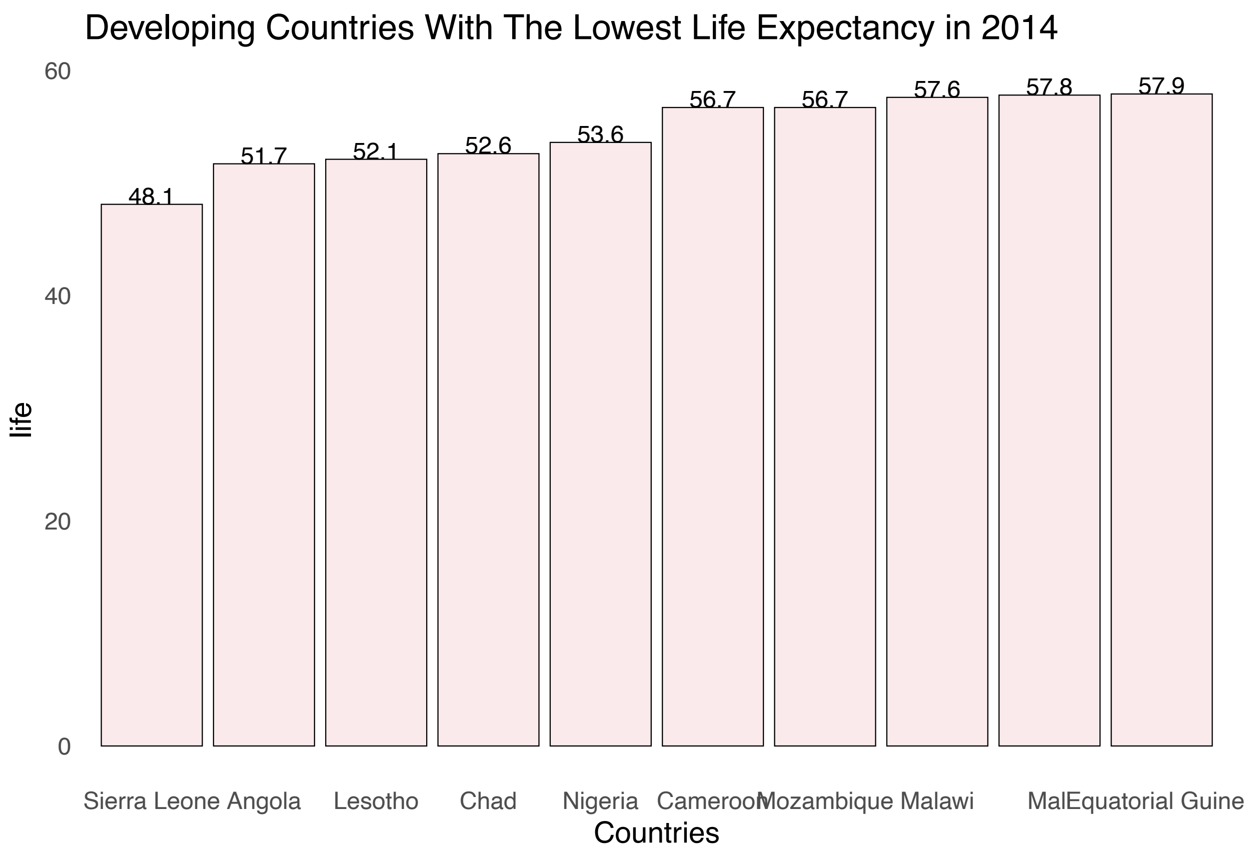
Top and bottom countries:

```
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

```r
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
```

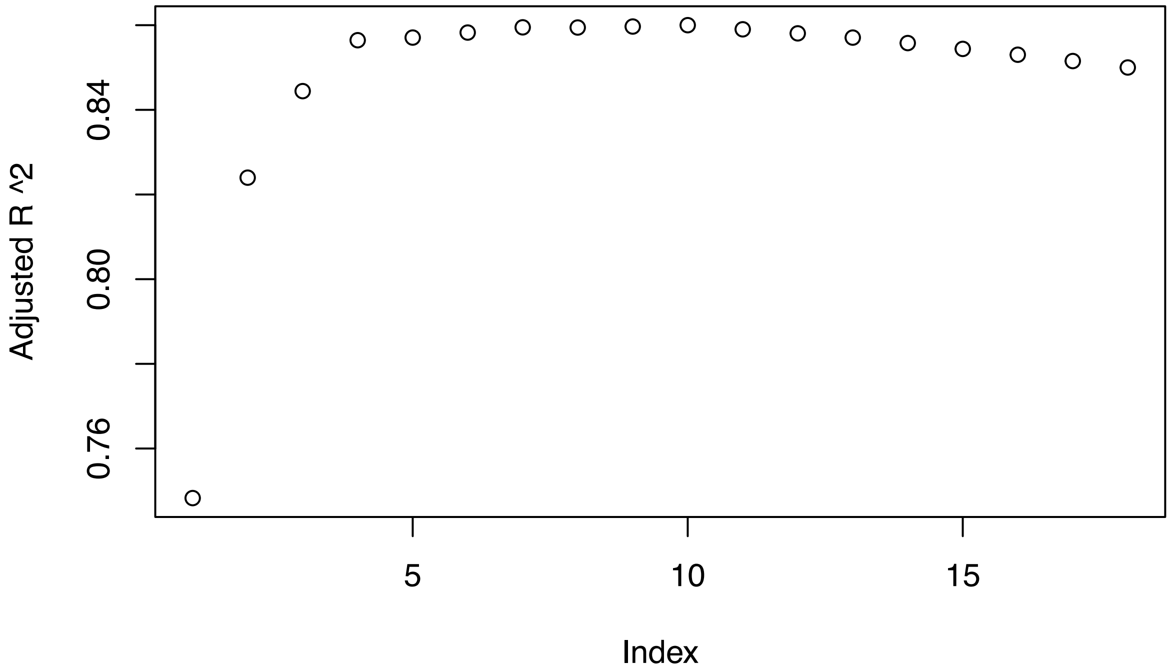
```
##  [1] 0.7505579 0.8271519 0.8486295 0.8616211 0.8635083 0.8659193 0.8683494
##  [8] 0.8695786 0.8710567 0.8726119 0.8729950 0.8734140 0.8738031 0.8739667
## [15] 0.8740737 0.8742070 0.8742664 0.8743277
```

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`)
```

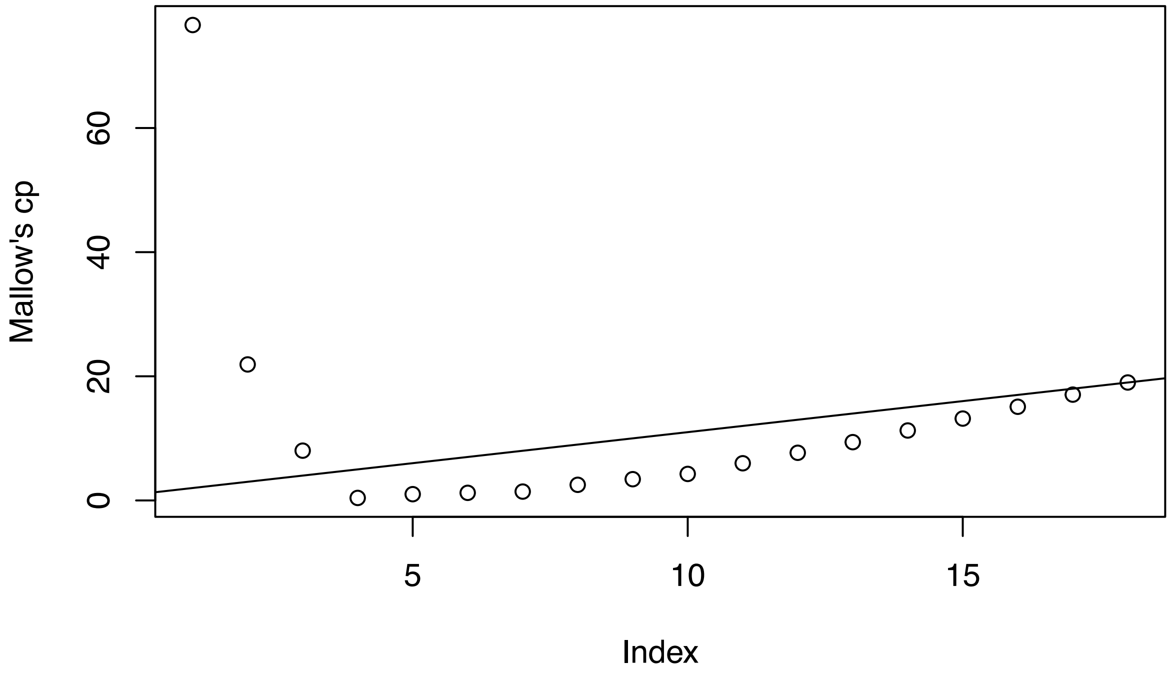


Adjusted R-squared levels out at the model with 4 predictors

```
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

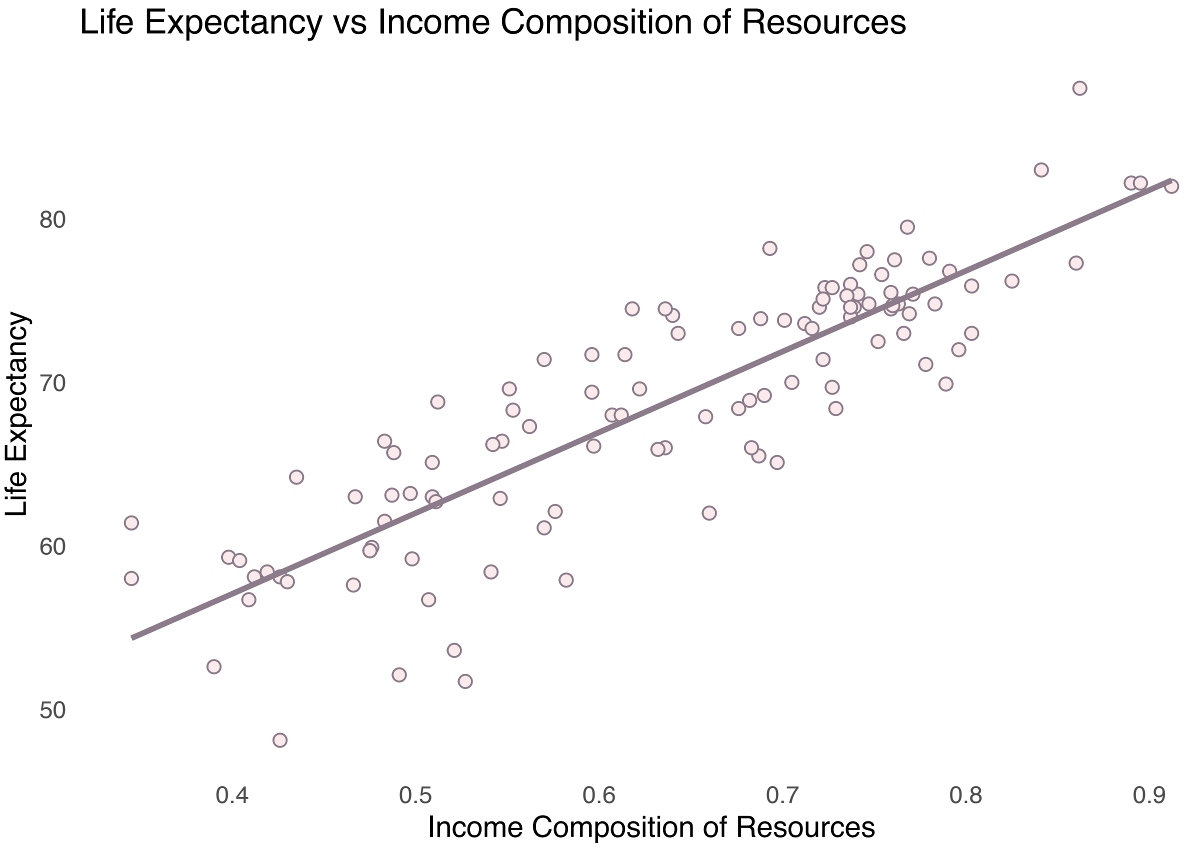
```
mod = update(mod.aic, ~.+comp:exp.t+comp:hiv)
summary(mod)
```

```
##
## 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
```

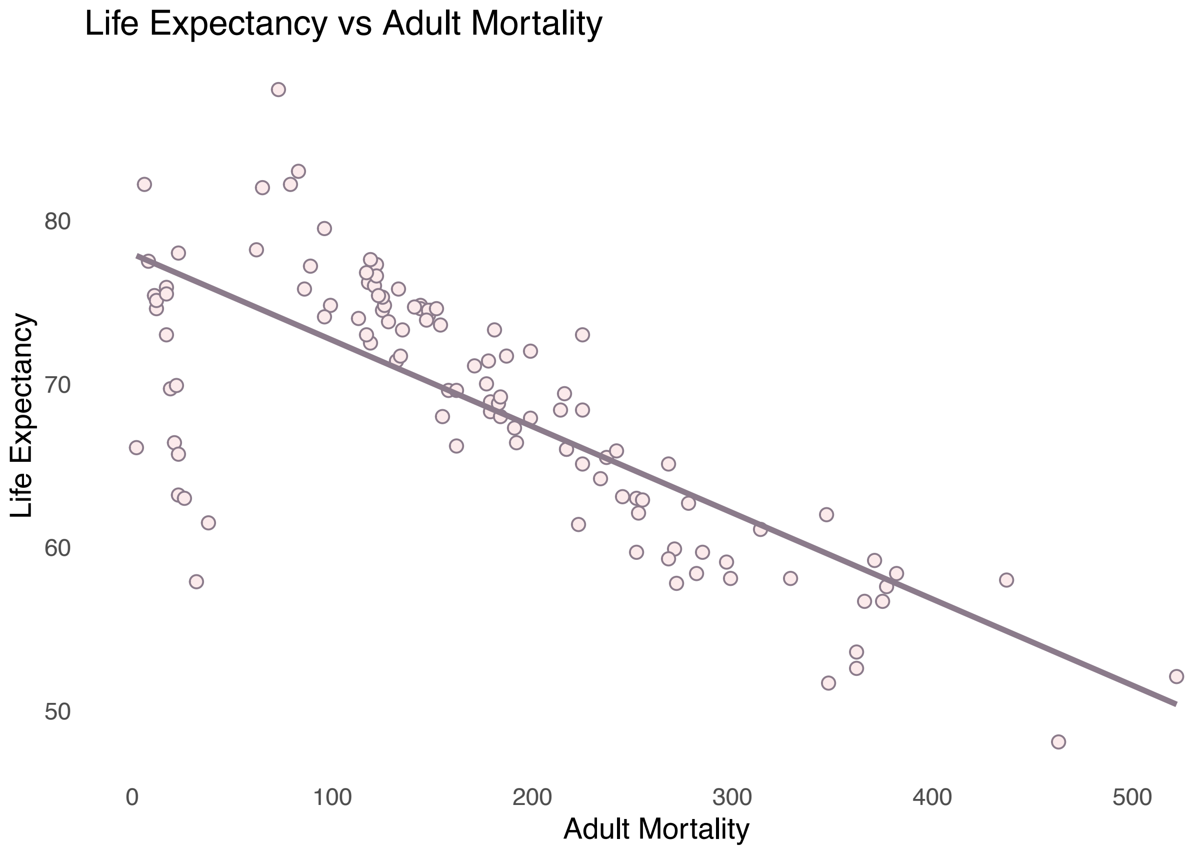
E.3 Model assumptions

Check linearity:

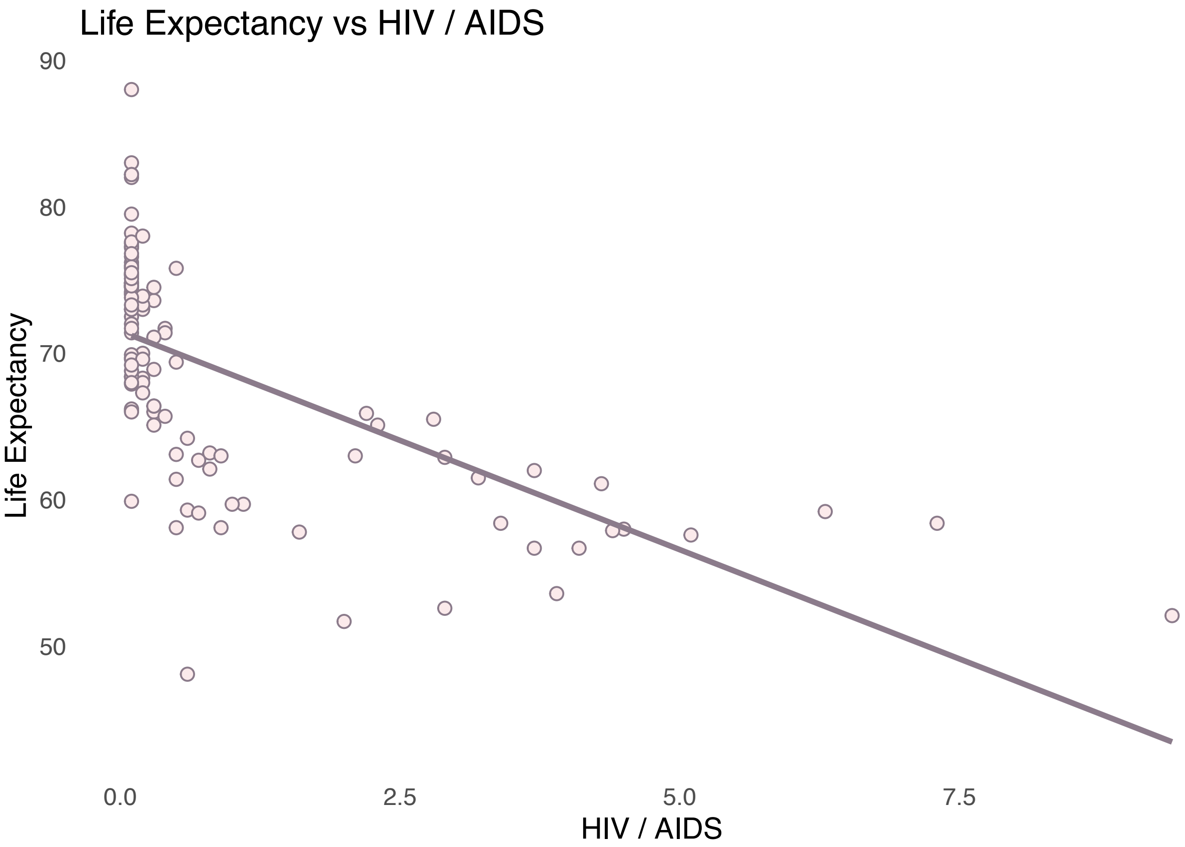
```
df2 |>
  ggplot(aes(x = comp, y = life, label = country)) +
  geom_point(shape = 21, color = "thistle4", fill = "#FBEAEB", size = 2) +
  geom_smooth(method = "lm", se = FALSE, color = "thistle4") +
  labs(x = "Income Composition of Resources", y = "Life Expectancy",
       title = "Life Expectancy 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
```

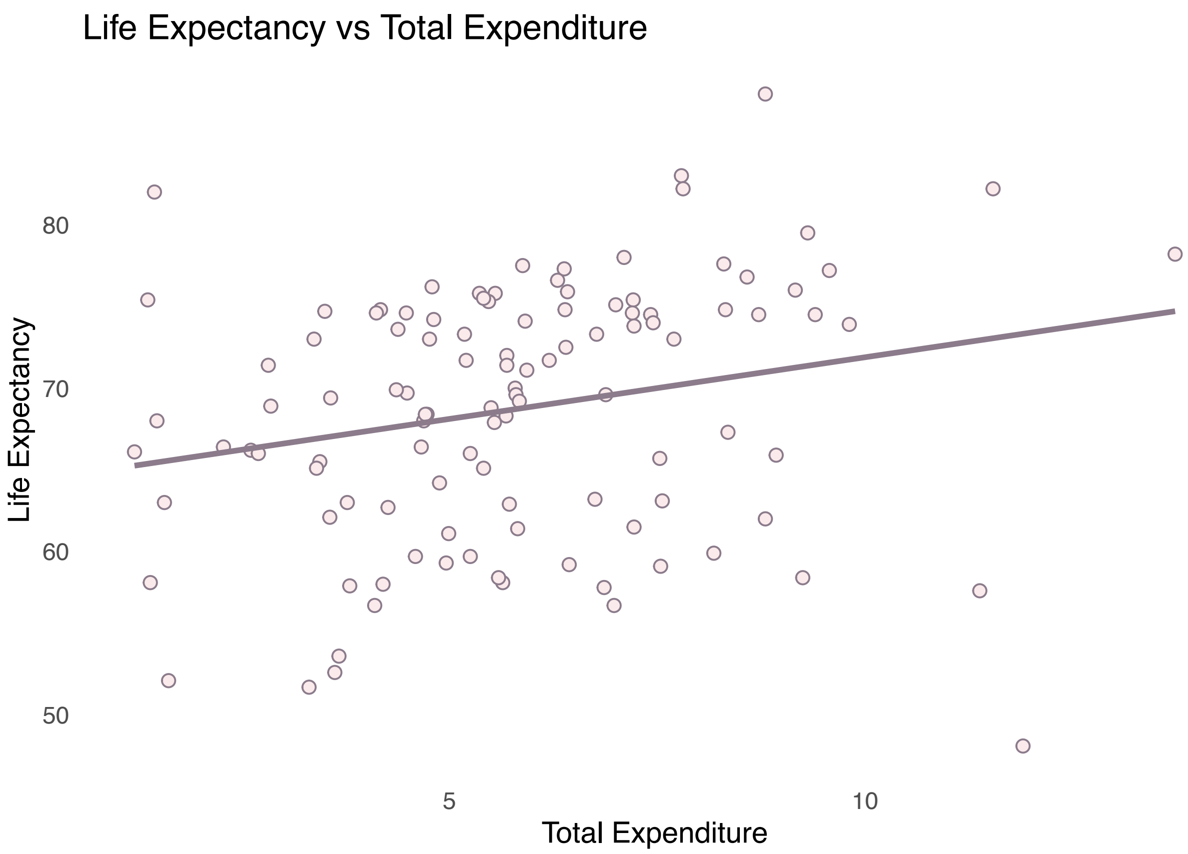


```
df2 |>
  ggplot(aes(x = hiv, y = life, label = country)) +
  geom_point(shape = 21, color = "thistle4", fill = "#FBEAEB", size = 2) +
  geom_smooth(method = "lm", se = FALSE, color = "thistle4") +
  labs(x = "HIV / AIDS", y = "Life Expectancy", title = "Life Expectancy vs HIV / AIDS") +
  common_theme
```



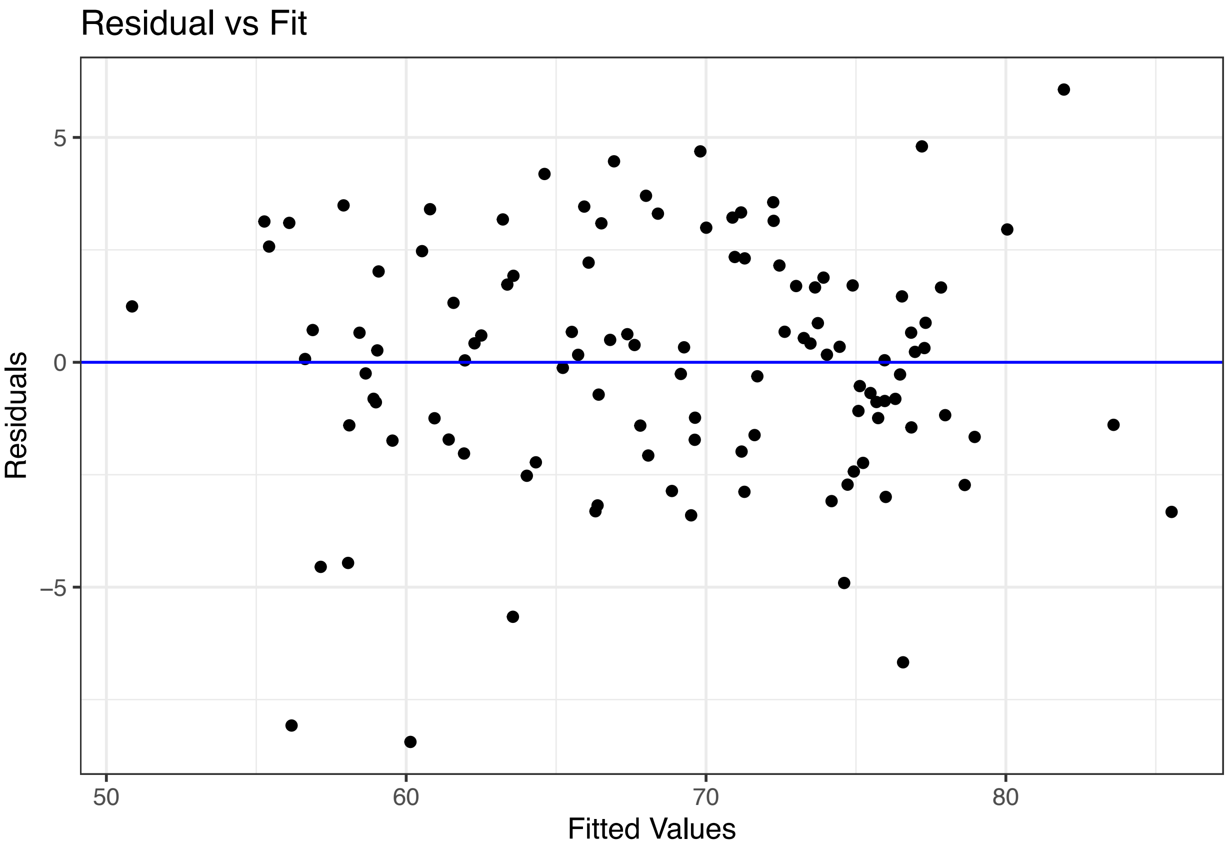
Life Expectancy vs HIV / AIDS

```
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:

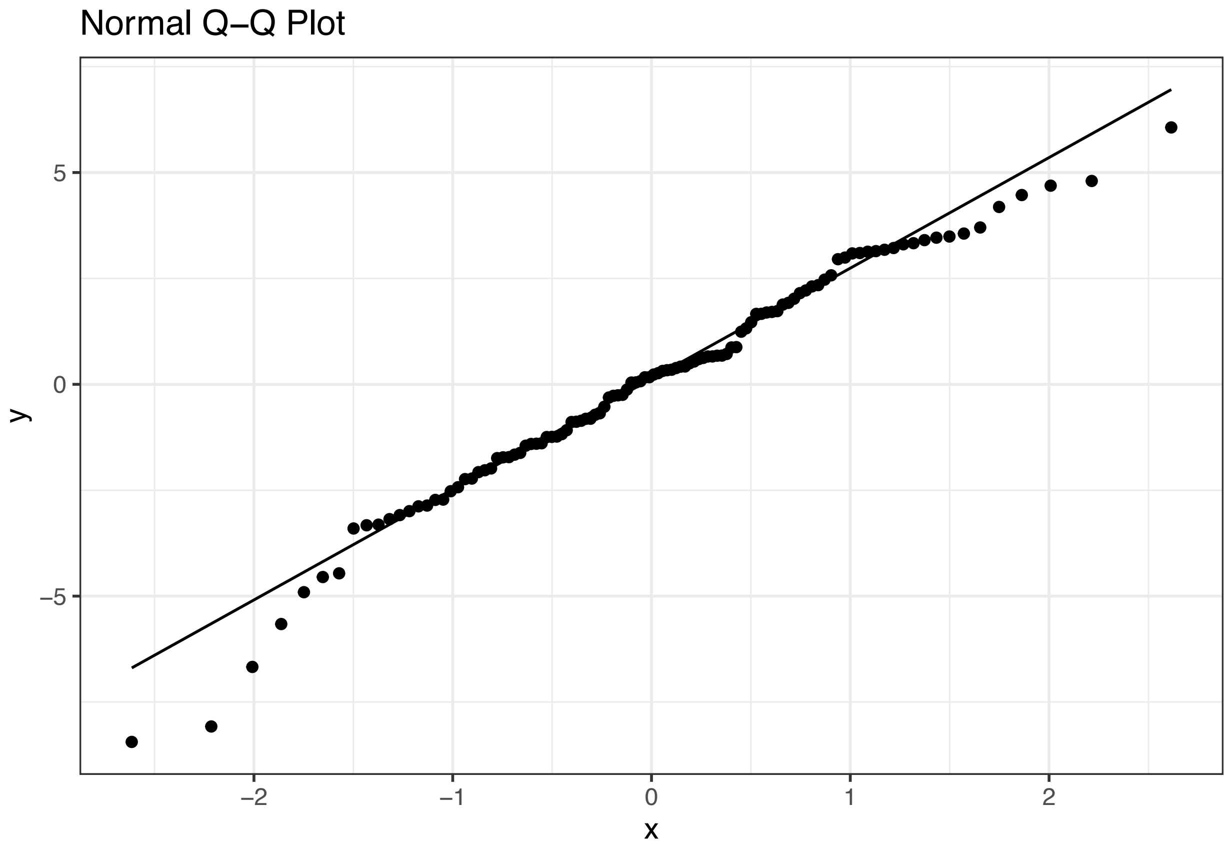
```
model.table = augment(mod)
ggplot(model.table, aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0, colour = 'blue') +
  labs(x = 'Fitted Values', y = 'Residuals') +
  ggtitle('Residual vs Fit') +
  theme_bw()
```



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()
```

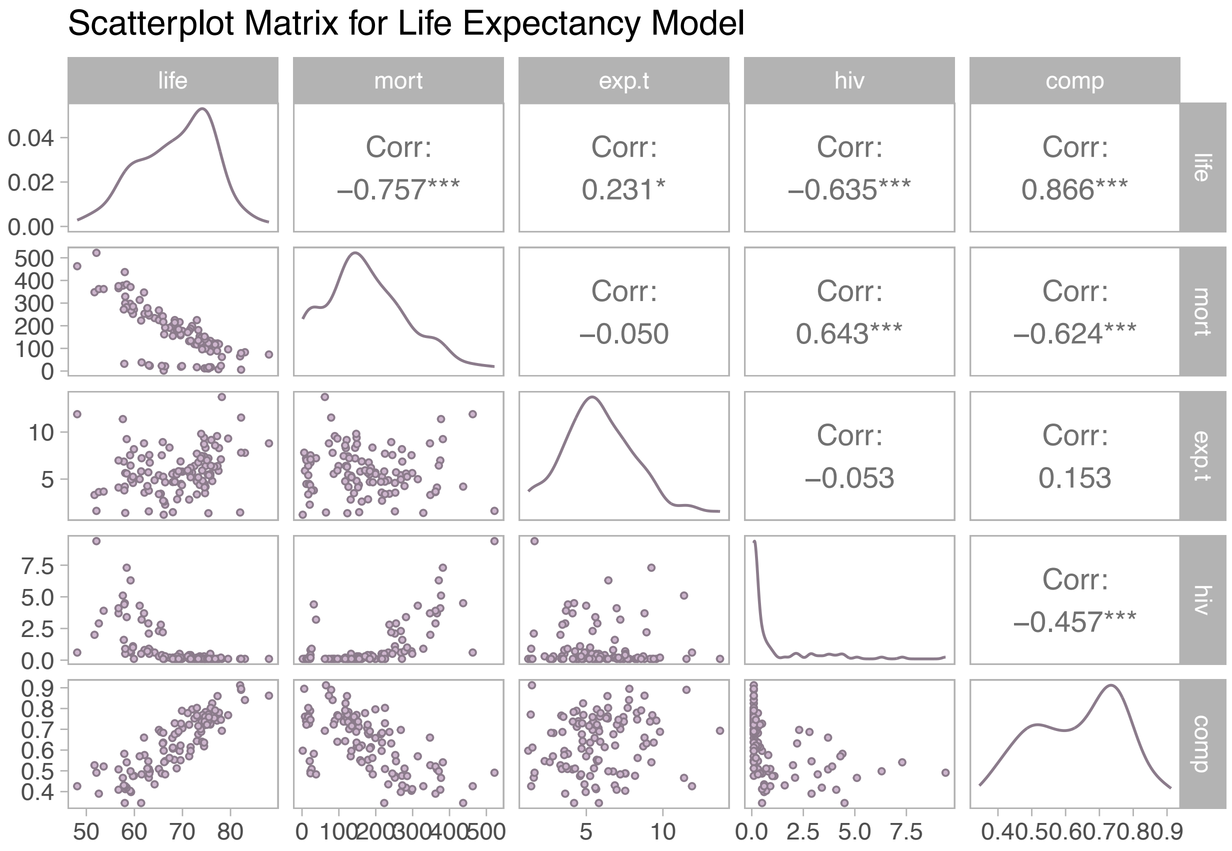


```
shapiro.test(resid(mod))
```

```
##
##  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()

print(new)
```

```
##       mort       hiv      comp    exp.t
## 1 174.4732 0.9303571 0.6362054 5.822768
```

```
pi = predict(mod, new, interval = "prediction", level = 0.95)
pi
```

```
##        fit      lwr      upr
## 1 67.96833 62.48465 73.45202
```

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"))
```

Summary of **Life Expectancy**

Country	Life	Mort	hiv	Comp	Exp.T
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
Maldives	78.2	62	0.1	0.693	13.73
Ukraine	78.0	23	0.2	0.746	7.10
Panama	77.6	119	0.1	0.780	8.30
Albania	77.5	8	0.1	0.761	5.88