Gapminder HW

Bilal Gilani

3/24/2020

Part 1

1.

```
if(!file.exists("./data")) {dir.create("./data")}

fileURLs <- c("https://docs.google.com/spreadsheet/pub?key=0AkBd6lyS3EmpdHo5S0J6ekhVOF9QaVhod05QSGV4T3c&
              "https://docs.google.com/spreadsheet/pub?key=phAwcNAVuyj2tPLxKvvnNPA&output=xlsx",
              "https://docs.google.com/spreadsheet/pub?key=tSUr_yZVbM6a3AGJEq_Z2Pw&output=xlsx",
              "https://docs.google.com/spreadsheet/pub?key=0ArfEDsV3bBwCdHBzUVVSMDlTX1ZCUnNJQ3ZFdkFXVFE&
              "https://docs.google.com/spreadsheet/pub?key=phAwcNAVuyj0XOoBL_n5tAQ&output=xlsx" )

var_names <- c("GDP","life_expectancy", "alt_GDP", "blood press", "population")
```

```
library(readxl)
library(tidyr)
```

```
## Warning: package 'tidyr' was built under R version 3.6.3
```

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 3.6.3

## -- Attaching packages ---------------------------------------------------------------------------

## v ggplot2 3.2.1     v dplyr   0.8.3
## v tibble  2.1.3     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
## v purrr   0.3.3

## Warning: package 'purrr' was built under R version 3.6.3

## -- Conflicts ------------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
get_clean <- function(url_in, var_name) {
  download.file(url_in, destfile = "./data/tmp.xslx", mode = "wb")
  output <- read_excel("./data/tmp.xslx")
  names(output)[[1]] <- "country"
  output <- output %>%
    pivot_longer(-country, names_to = "year", values_to = var_name) %>%
    mutate(year = as.numeric(year)) %>%
    drop_na()
}
out1 <- get_clean(fileURLs[1], var_names[1])
head(out1)
```

```
## # A tibble: 6 x 3
##   country  year   GDP
##   <chr>   <dbl> <dbl>
## 1 Albania  1980 1061.
## 2 Albania  1981 1100.
## 3 Albania  1982 1111.
## 4 Albania  1983 1101.
## 5 Albania  1984 1065.
## 6 Albania  1985 1060.
```

```
all_data <- map2(fileURLs, var_names, get_clean)
```

```
## New names:
## * `` -> ...1
```

2.

```
joined_data <- all_data %>%
  reduce(full_join, by = c("country", "year"))
```

3.

```
library(countrycode)
```

```
## Warning: package 'countrycode' was built under R version 3.6.3
```

```
continent <- countrycode(sourcevar = joined_data$country, origin = "country.name", destination = "conti
```

```
## Warning in countrycode(sourcevar = joined_data$country, origin = "country.name", : Some values were r
```

```
## Warning in countrycode(sourcevar = joined_data$country, origin = "country.name", : Some strings were
```

```r
new_gapminder <- cbind(joined_data, continent)

no_continent <- new_gapminder[is.na(new_gapminder$continent),]
unique(no_continent$country)
```

```
##  [1] "Channel Islands"       "Kosovo"
##  [3] "Serbia and Montenegro" "Central African Rep."
##  [5] "Akrotiri and Dhekelia" "Cocos Island"
##  [7] "Czechoslovakia"        "East Germany"
##  [9] "Eritrea and Ethiopia"  "St. Martin"
## [11] "North Yemen (former)"  "South Yemen (former)"
## [13] "Yugoslavia"
```

```r
## These countries were not given values in the continent column because they no longer exist
## or are islands such as in the examples of the Channel Islands and St. Martin, or territories of
## another country such as "Akrotiri and Dhekelia" which is a British Overseas Territory and considered
## part of the continent however is not part of the continent shelf or Cocos Island which is an Austral
## territory but is considered to be a part of Asia. Another possibility is that they were considered d
## such a in the case of "Eritrea and Ethiopia", both of which are now individual countries.

new_gapminder$continent[which(new_gapminder$country == "Channel Islands")] <- "Europe"
new_gapminder$continent[which(new_gapminder$country == "Kosovo")] <- "Europe"
new_gapminder$continent[which(new_gapminder$country == "Serbia and Montenegro")] <- "Europe"
new_gapminder$continent[which(new_gapminder$country == "Central African Rep.")] <- "Africa"
new_gapminder$continent[which(new_gapminder$country == "Akrotiri and Dhekelia")] <- "Europe"
new_gapminder$continent[which(new_gapminder$country == "Cocos Island")] <- "Asia"
new_gapminder$continent[which(new_gapminder$country == "Czechoslovakia")] <- "Europe"
new_gapminder$continent[which(new_gapminder$country == "East Germany")] <- "Europe"
new_gapminder$continent[which(new_gapminder$country == "Eritrea and Ethiopia")] <- "Africa"
new_gapminder$continent[which(new_gapminder$country == "St. Martin")] <- "Americas"
new_gapminder$continent[which(new_gapminder$country == "North Yemen (former)")] <- "Asia"
new_gapminder$continent[which(new_gapminder$country == "South Yemen (former)")] <- "Asia"
new_gapminder$continent[which(new_gapminder$country == "Yugoslavia")] <- "Europe"

which(is.na(new_gapminder$continent))
```

```
## integer(0)
```

```r
## no missing values in the continent column.

new_gapminder <- new_gapminder %>%
  arrange(country, year)
```

4.

```r
a <- new_gapminder[,c(1,2,3,4,7,8)]
a <- a %>% fill(population)
plot_data <- a %>% distinct()
```

```
library(ggplot2)
library(gganimate)
```

```
## Warning: package 'gganimate' was built under R version 3.6.3
```

```
library(gifski)
```

```
## Warning: package 'gifski' was built under R version 3.6.3
```

```
plot1 <- ggplot(plot_data, aes(GDP, life_expectancy)) +
  geom_point(alpha = 0.7, aes(size = population, colour = continent)) +
  scale_size(range = c(2, 12)) +
  scale_x_log10() +
  theme(legend.position = "right") +
  labs(title = "Year: {frame_time}", x = "GDP", y = "life expectancy") +
  transition_time(year) +
  ease_aes("linear")
```

Part 2

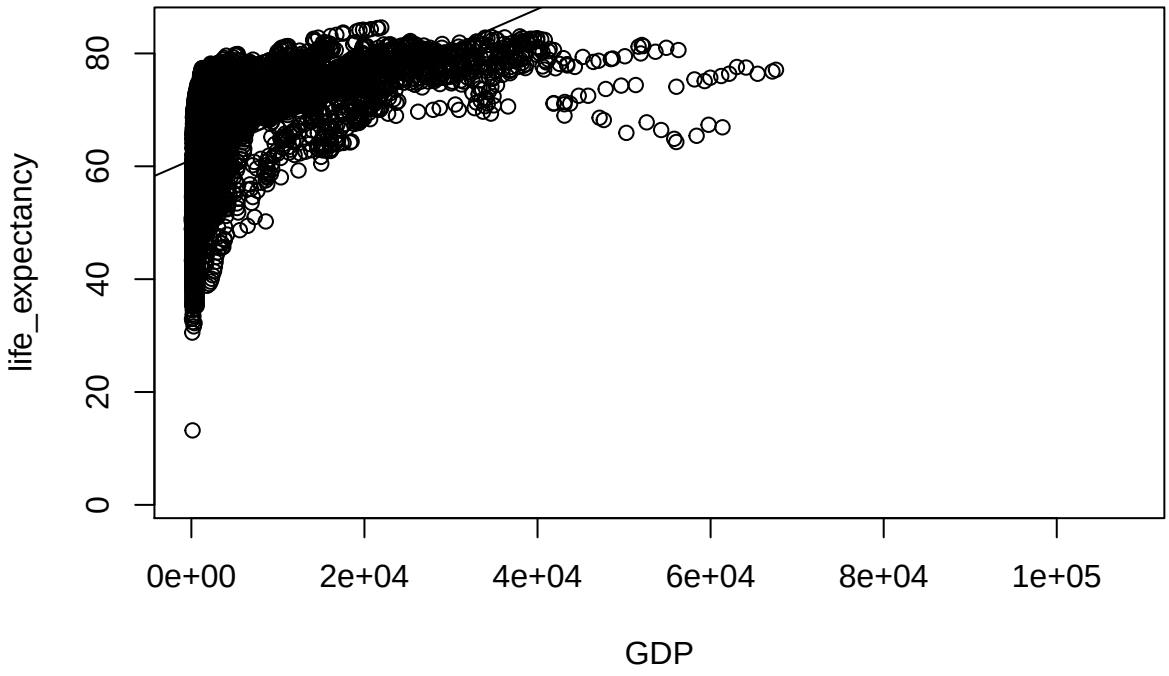
1.

```
attach(plot_data)
```

```
## The following object is masked _by_ .GlobalEnv:
##
##     continent
```

```
## The following object is masked from package:tidyr:
##
##     population
```

```
plot(GDP, life_expectancy) +
  abline(lm(life_expectancy ~ GDP))
```



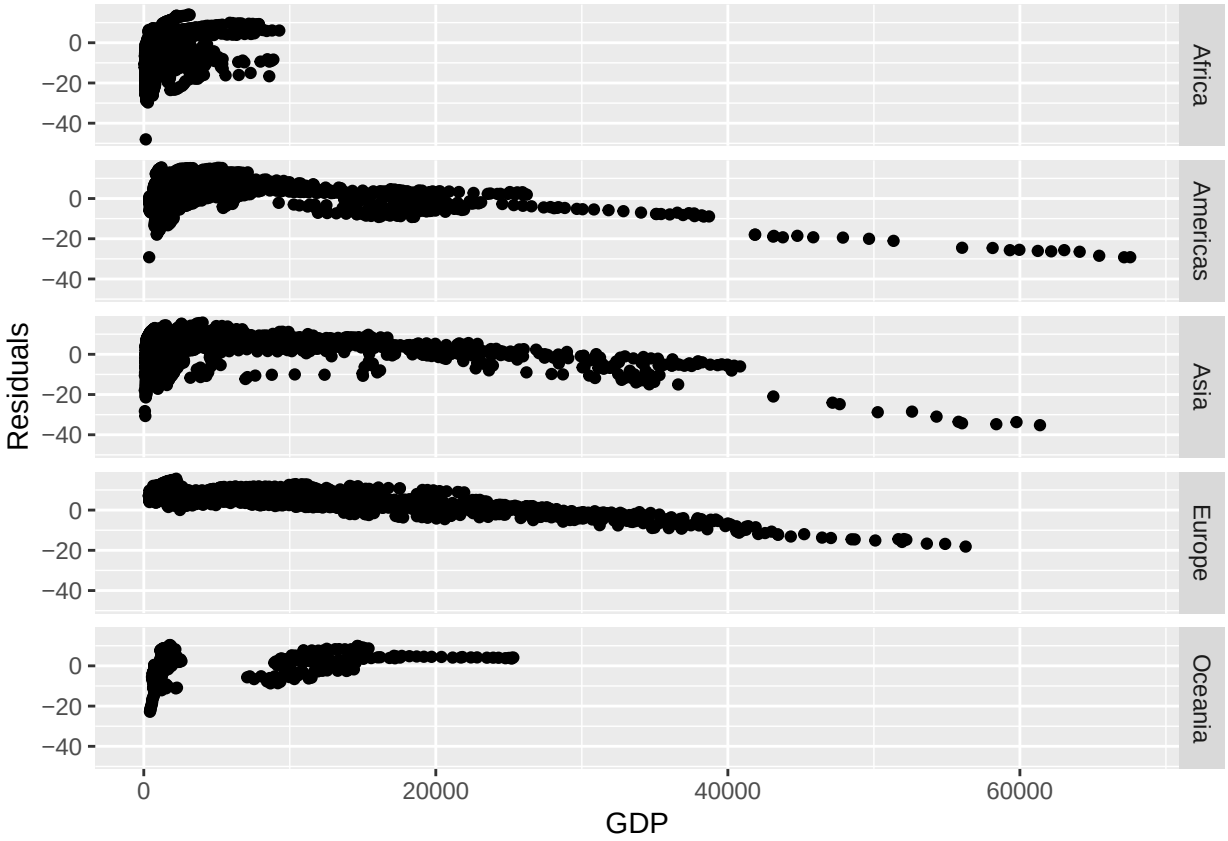

```
## integer(0)
```

2.

```
reg <- lm(life_expectancy ~ GDP, data = plot_data)
res.lm <- resid(reg)
NoNA <- na.omit(plot_data)
NoNA <- cbind(NoNA, res.lm)

plot2 <- ggplot(NoNA, aes(GDP, res.lm)) +
  geom_point() +
  xlab("GDP") +
  ylab("Residuals") +
  facet_grid(rows = vars(continent))

plot2
```

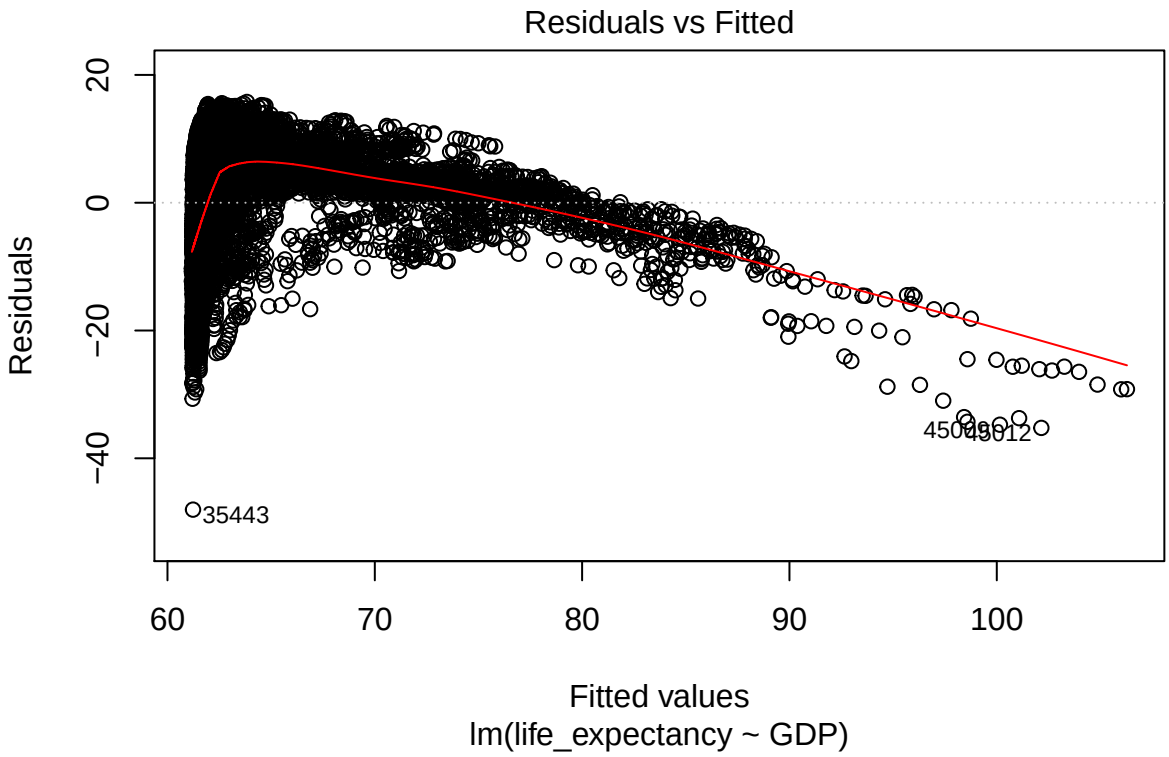


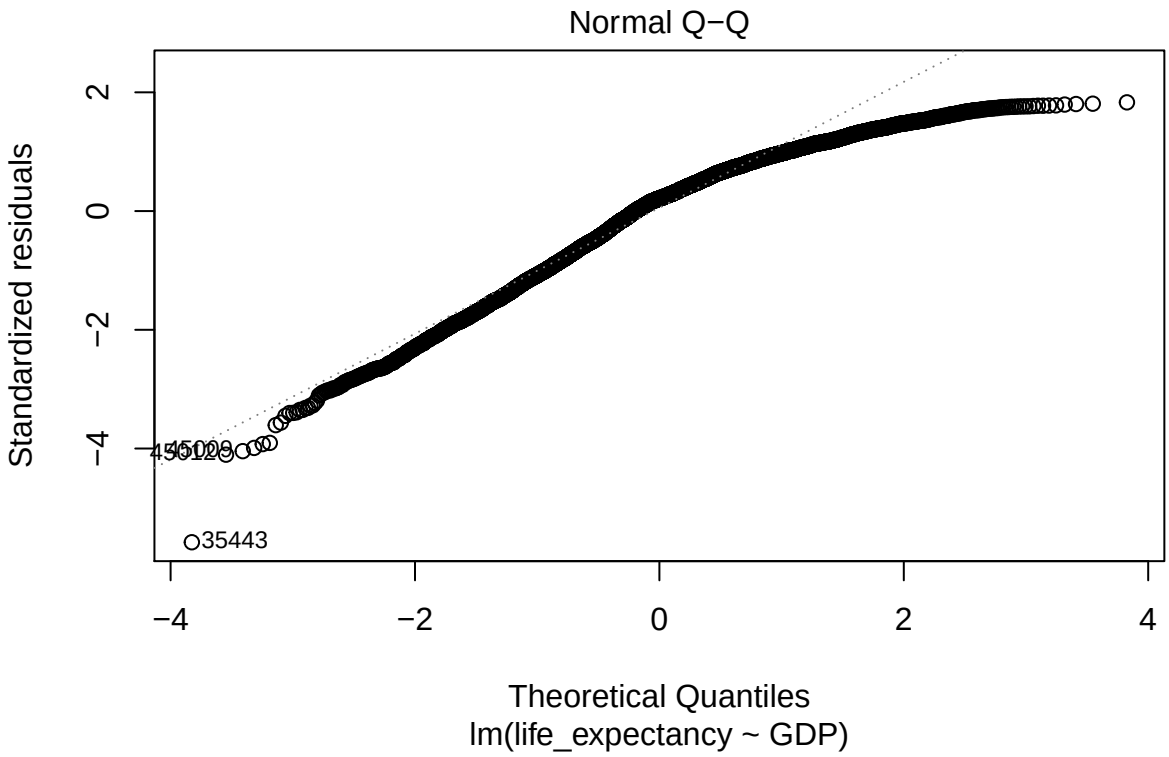
3.

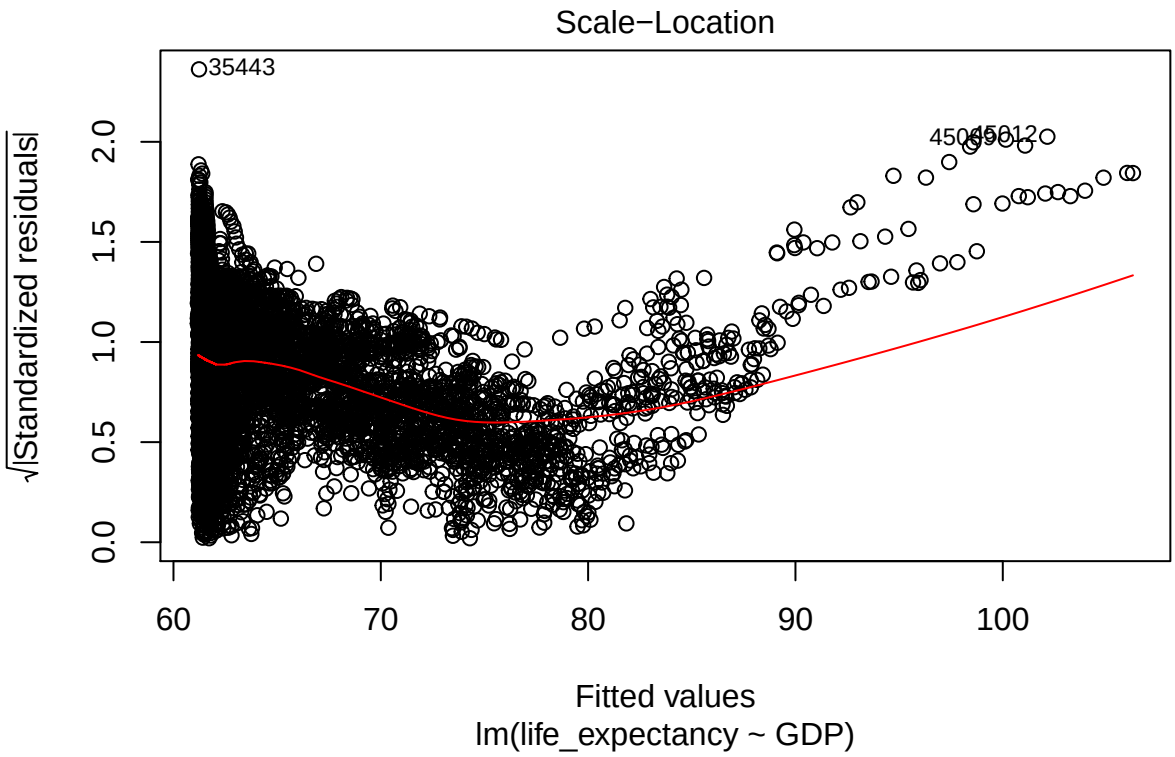
```
library(broom)
glance(reg)
```

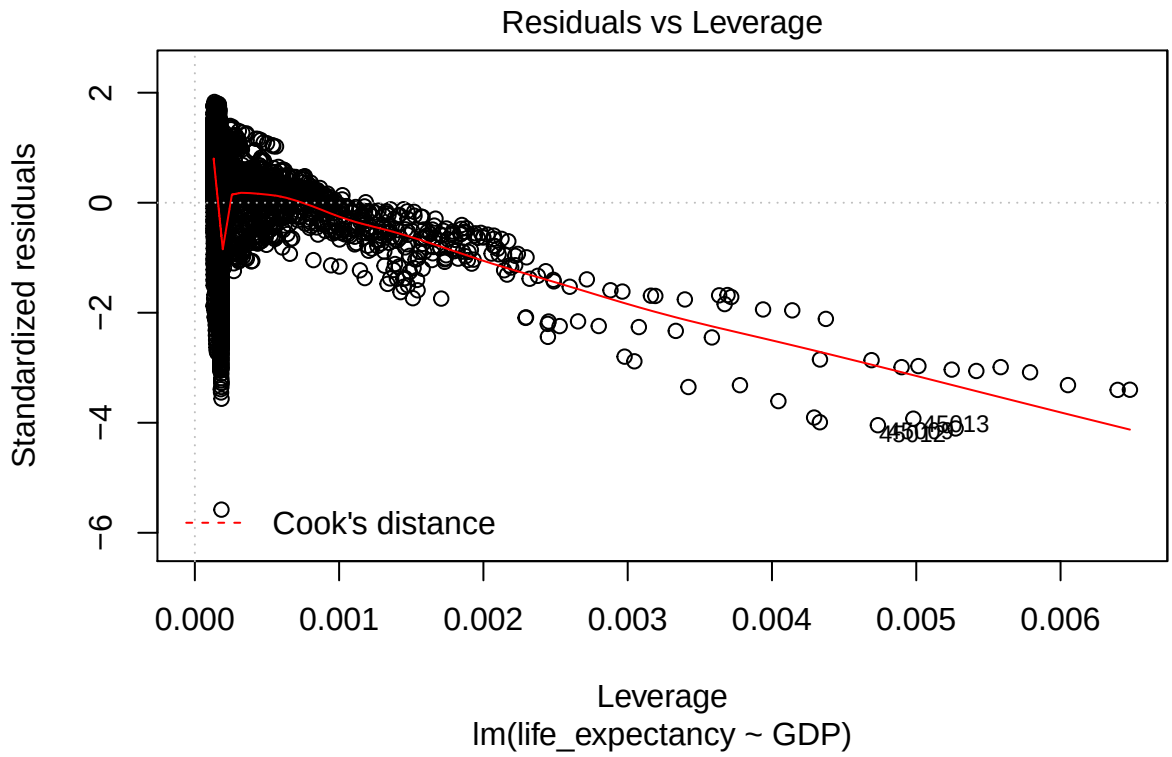
```
## # A tibble: 1 x 11
##   r.squared adj.r.squared sigma statistic p.value    df  logLik    AIC
##       <dbl>         <dbl> <dbl>     <dbl>   <dbl> <int>   <dbl>  <dbl>
## 1     0.320         0.320  8.61     3612.       0     2 -27438. 54882.
## # ... with 3 more variables: BIC <dbl>, deviance <dbl>, df.residual <int>
```

```
plot(reg)
```


Residuals vs Fitted
Residuals
20
0
−20
−40
60
70
80
90
100
35443
Fitted values
lm(life_expectancy ~ GDP)


Normal Q−Q
Standardized residuals
2
0
−2
−4
45009
45012
35443
−4
−2
0
2
4
Theoretical Quantiles
lm(life_expectancy ~ GDP)


Scale−Location
35443
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
60
70
80
90
100
Fitted values
lm(life_expectancy ~ GDP)




```
## The outliers are in rows 35,443 , 45,009 , and 45,012.

outlier.reg <- plot_data %>%
  slice(35443, 45009, 45012)

head(outlier.reg)
```

```
##                country year       GDP life_expectancy population continent
## 1               Rwanda 1994   139.501           13.20    5995987    Africa
## 2 United Arab Emirates 1977 58384.480           65.41     722849      Asia
## 3 United Arab Emirates 1980 61374.755           66.91    1016789      Asia
```

The outliers in this model are Rwanda in 1994 which can be attributed to the Rwandan genocide. The other two outliers is the UAE in both 1977 and 1980. I could not find significant events as to why the UAE stuck out in the particular years.

Part 3

```
library(gapminder)
```

```
## Warning: package 'gapminder' was built under R version 3.6.3
```

```r
data("gapminder")
```

a.

```r
gapminder %>%
  mutate(year = year - mean(year)) ->
  gapminder

mean(gapminder$year)
```

```
## [1] 0
```

b.

```r
attach(gapminder)
```

```
## The following object is masked _by_ .GlobalEnv:
##
##     continent

## The following objects are masked from plot_data:
##
##     continent, country, year
```

```r
year2 <- (year)^2
gapminder <- cbind(gapminder, year2)

quad <- lm(lifeExp ~ year + year2, data = gapminder)
quad
```

```
##
## Call:
## lm(formula = lifeExp ~ year + year2, data = gapminder)
##
## Coefficients:
## (Intercept)         year        year2
##   60.511805     0.325904    -0.003482
```

```r
summary(quad)
```

```
##
## Call:
## lm(formula = lifeExp ~ year + year2, data = gapminder)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -40.443  -9.594   1.441  10.280  23.754
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 60.511805   0.423931 142.740  < 2e-16 ***
## year         0.325904   0.016277  20.022  < 2e-16 ***
## year2       -0.003482   0.001066  -3.268  0.00111 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.6 on 1701 degrees of freedom
## Multiple R-squared:  0.1948, Adjusted R-squared:  0.1939
## F-statistic: 205.8 on 2 and 1701 DF,  p-value: < 2.2e-16
```

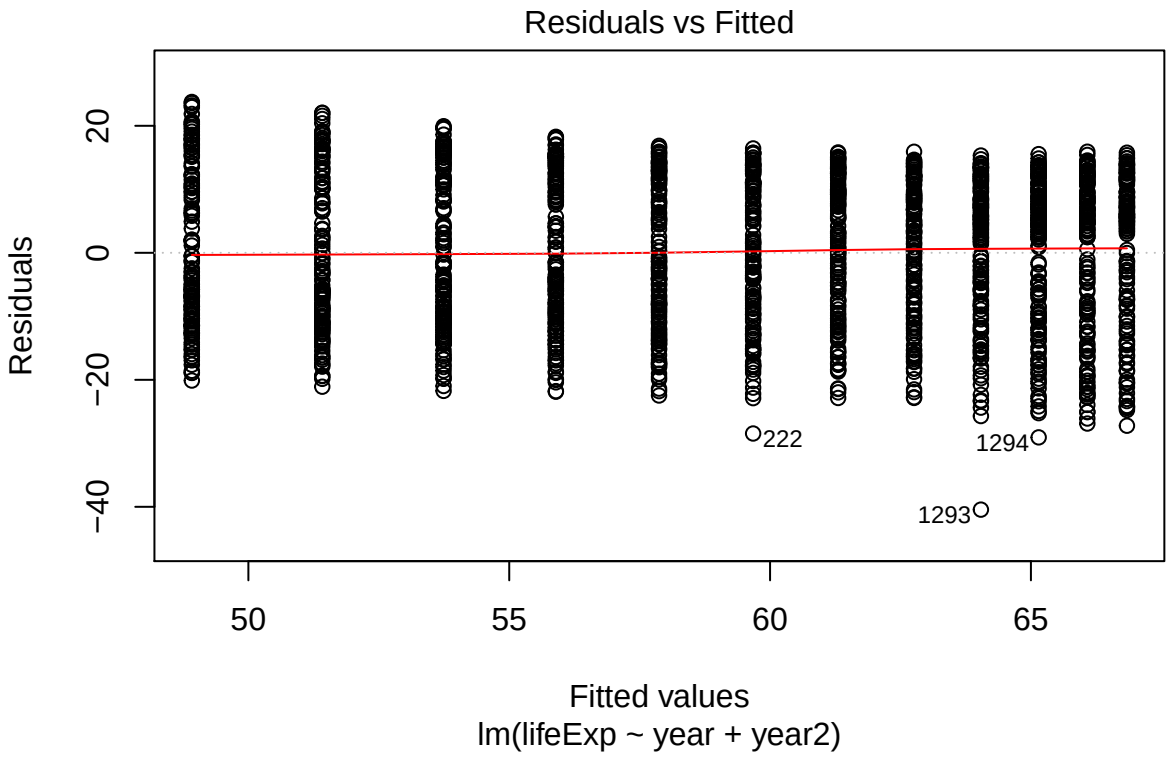
The quadratic formula is Y = 60.512 + 0.33X - 0.003Xˆ2 with X being the year and Y being life expectancy.

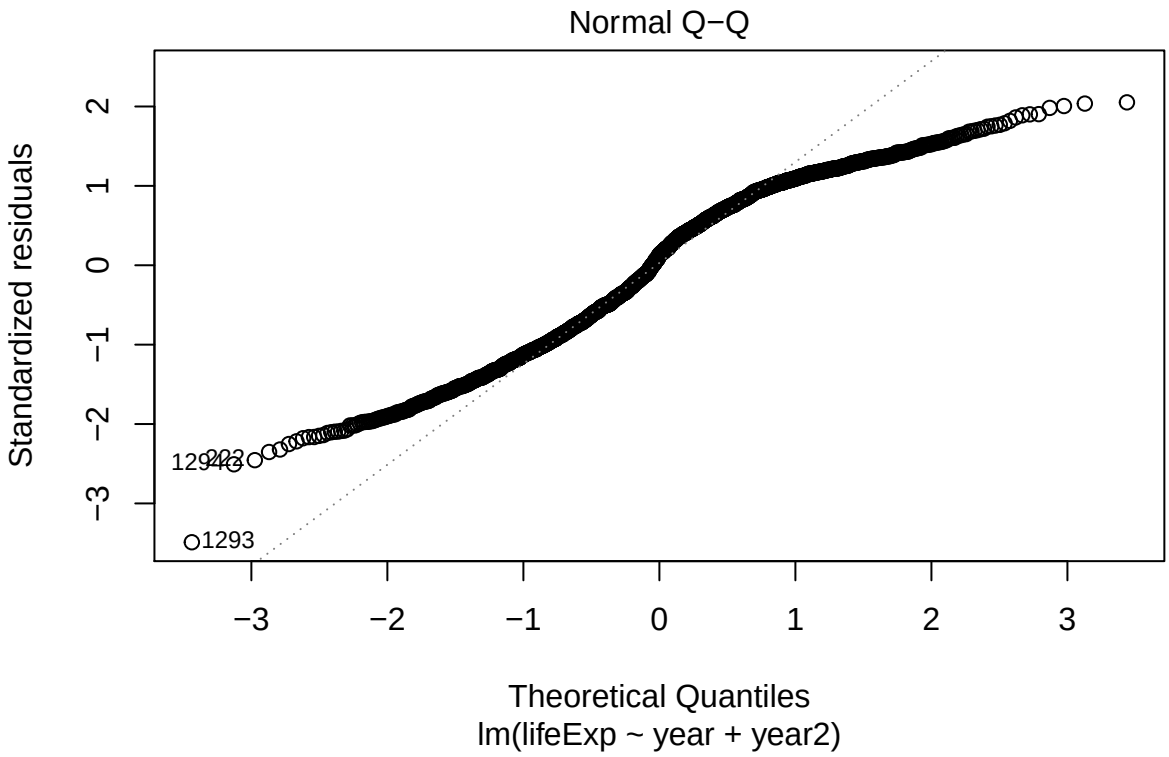
c.

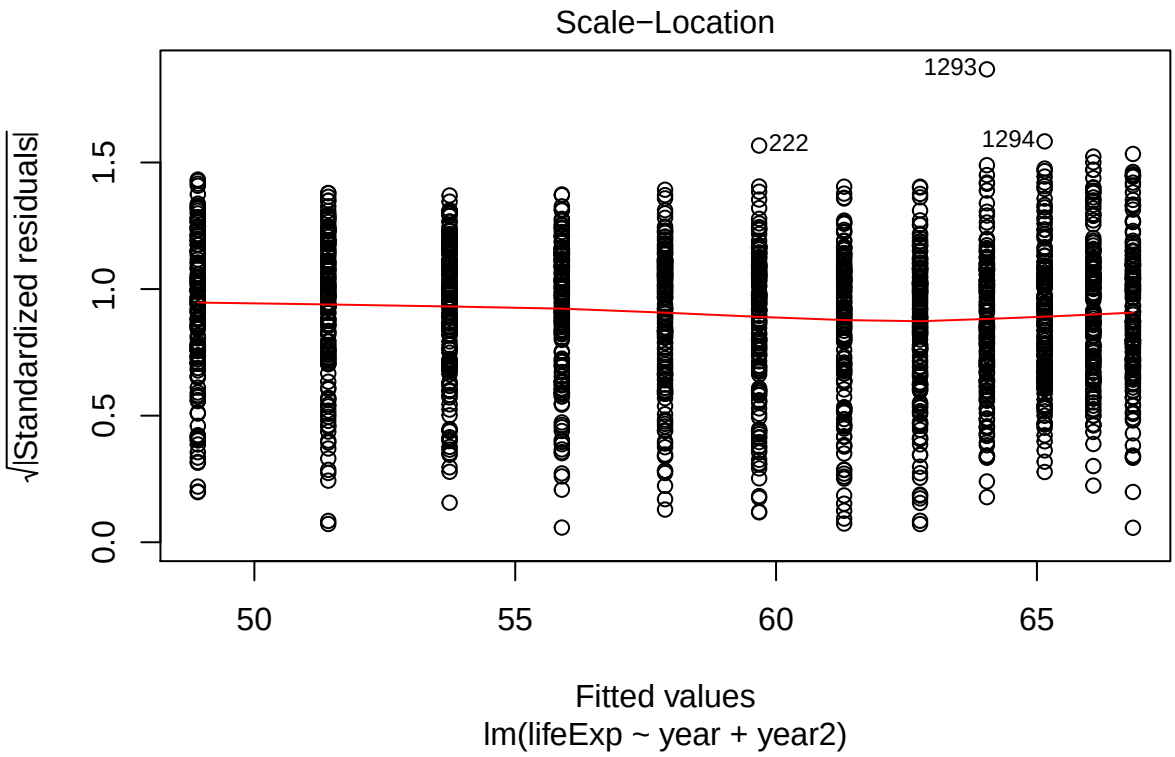
```
glance(quad)
```

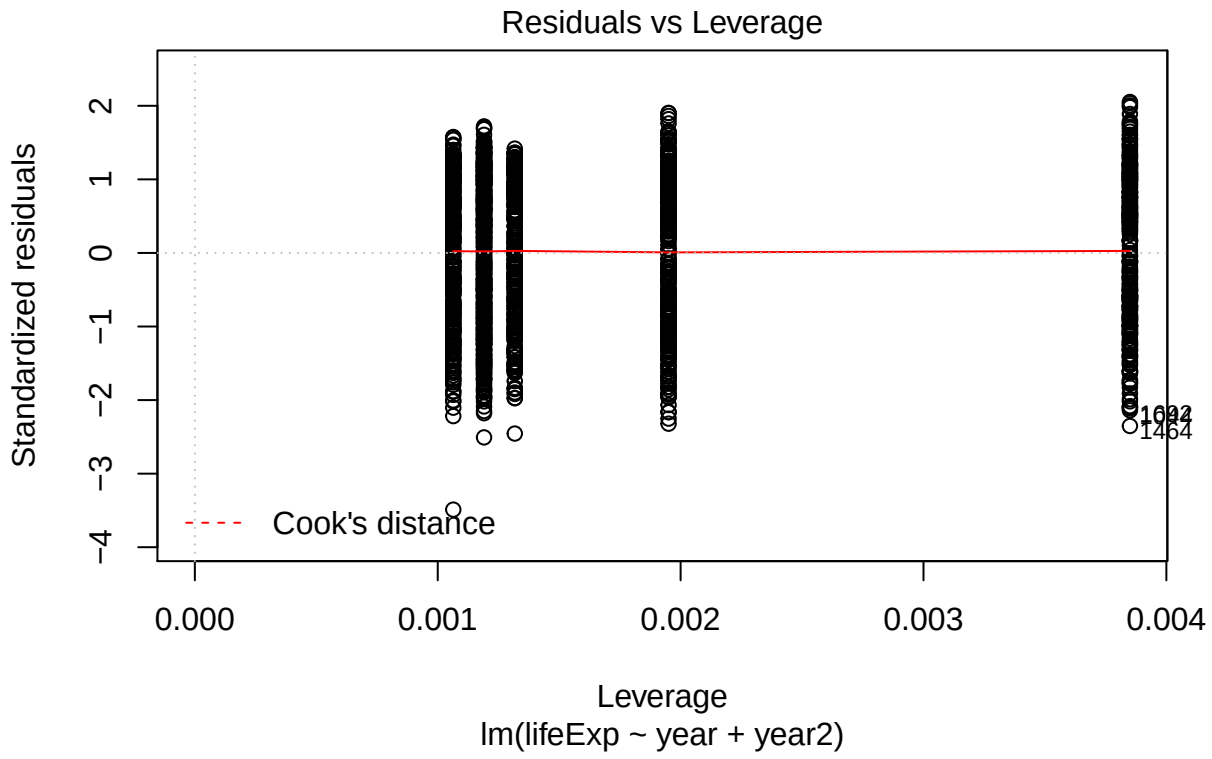
```
## # A tibble: 1 x 11
##   r.squared adj.r.squared sigma statistic  p.value    df logLik    AIC
##       <dbl>         <dbl> <dbl>     <dbl>    <dbl> <int>  <dbl>  <dbl>
## 1     0.195         0.194  11.6      206. 9.24e-81     3 -6593. 13193.
## # ... with 3 more variables: BIC <dbl>, deviance <dbl>, df.residual <int>
```

```
plot(quad)
```


Residuals vs Fitted
Residuals
20
0
−20
−40
222
1294
1293
50
55
60
65
Fitted values
lm(lifeExp ~ year + year2)


Normal Q−Q
Standardized residuals
2
1
0
−1
−2
−3
1294
223
1293
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(lifeExp ~ year + year2)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
1293
222
1294
50
55
60
65
Fitted values
lm(lifeExp ~ year + year2)




```
## The outliers are in rows 222 , 1,293 , and 1,294.

outlier.quad <- gapminder %>%
  slice(222, 1293, 1294)

head(outlier.quad)
```

```
##    country continent year lifeExp     pop gdpPercap  year2
## 1 Cambodia      Asia -2.5  31.220 6978607  524.9722   6.25
## 2   Rwanda    Africa 12.5  23.599 7290203  737.0686 156.25
## 3   Rwanda    Africa 17.5  36.087 7212583  589.9445 306.25
```

Once again, Rwanda is an outlier, which can be attributed to the genocide. Cambodia in the year 1977 is an outlier in this model. Upon further research I found that there was a genocide in Cambodia in 1977, which could be the reason for its outlier status.