Mini Project 1: Life Expectancy

Ashwed Patil, Vatsal Jatakia, Saniya Ambavanekar, Akshay Naik (Team Hong Kong)

February 14th, 2018

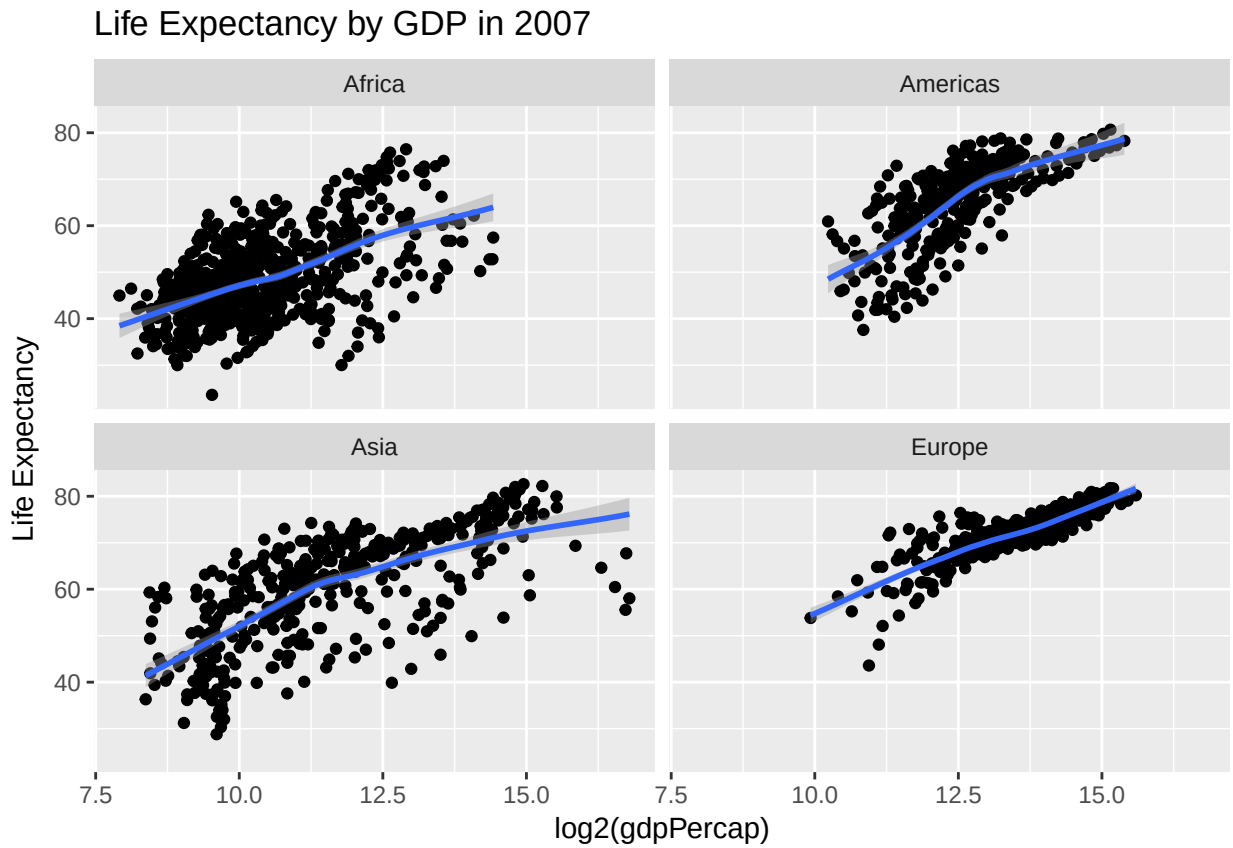
Question 1: GDP and Life Expectancy in 2007

We decide to drop Oceania from our data because there's data for only two countries from that continent and it won't be possible to fit complex regression models on such data (Also, it doesn't make sense as to why fit a model on just two points).

```
gap = gapminder[year == "2007", ]
gap = gapminder[continent != "Oceania", ]
```

We plotted a Q-Q Plot of the GDP Per Capita (See Appendix for the plot) and found that the distribution is significantly right skewed. Hence we decided to do a log2 transformation of the said variable for regression. Let's see if a LOESS curve provides a good fit for our data or not. Loess (local regression) is a non parametric regression method where we try to fit the best possible curve for every datapoint and generally works for both linear and non-linear data.

```
ggplot(gap, aes(log2(gdpPercap), lifeExp)) + geom_point() + geom_smooth(method = "loess",
    method.args = list(degree = 1)) + facet_wrap(~continent) +
    ggtitle("Life Expectancy by GDP in 2007") + ylab("Life Expectancy")
```



We can see that for all the continents, life expectancy increases with the log of gdp per capita. Thus, people living in richer countries are likely to live longer on an average than those in poorer countries. This is most

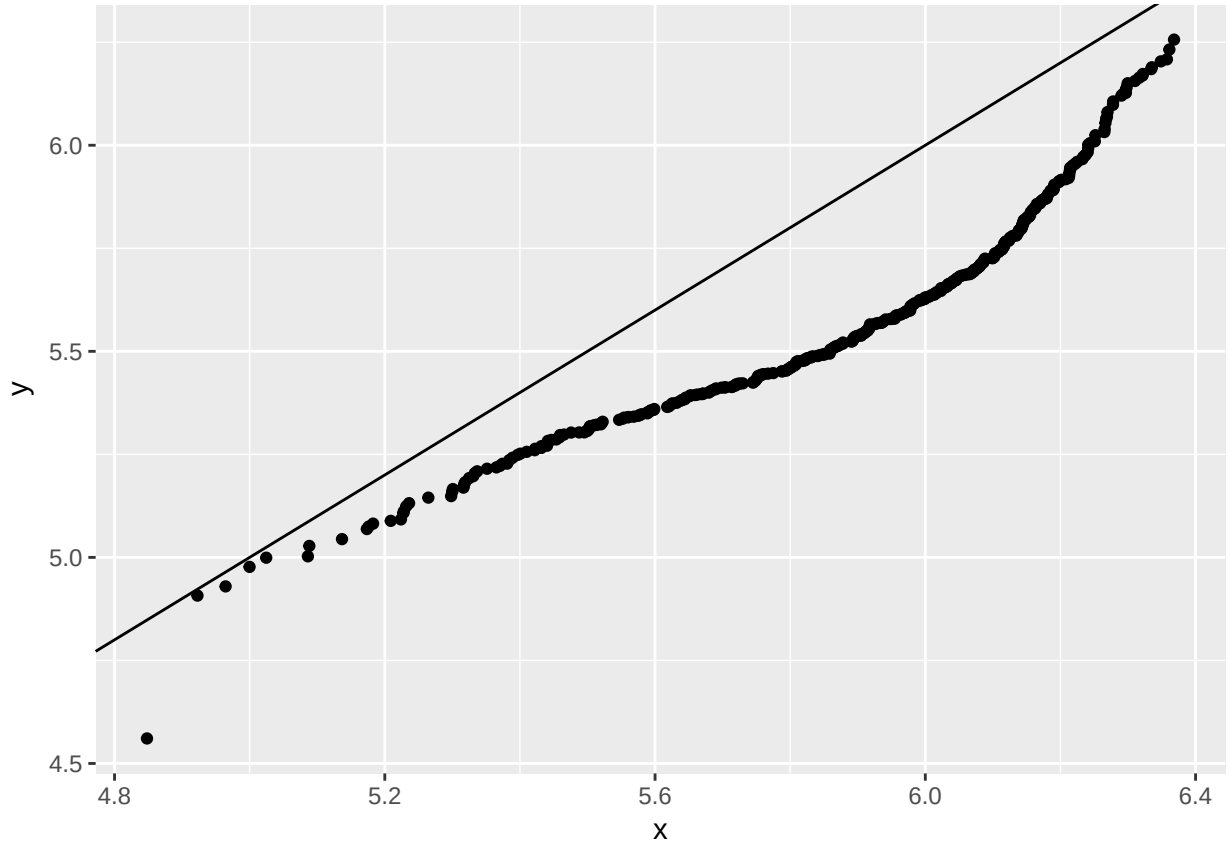
pronounced in Asia and Africa. We fitted a LOESS Model and checked the residual plot (see Appendix), the curve wiggles around zero which makes the model plausible. We also looked at the spread-location plot and found that the residuals are heteroskedastic which isn't a problem. Finally we computed the R^2 value to see how much variation the model captures.

```
ls1 = loess(lifeExp ~ log2(gdpPercap), gap)
df1 = augment(ls1)
var(df1$.fitted)/var(df1$lifeExp)
```

```
## [1] 0.6509769
```

The model captures 65% of the variation in the data which isn't bad. Because the variation in life expectancy by gdp per capita is the most in Asia and Africa, let's see if we can describe the difference using a multiplicative shift or not.

```
asia_life = log2(gap$lifeExp[gap$continent == "Asia"])
afr_life = log2(gap$lifeExp[gap$continent == "Africa"])
qq.df = as.data.frame(qqplot(asia_life, afr_life, plot.it = F))
ggplot(qq.df, aes(x = x, y = y)) + geom_point() + geom_abline()
```

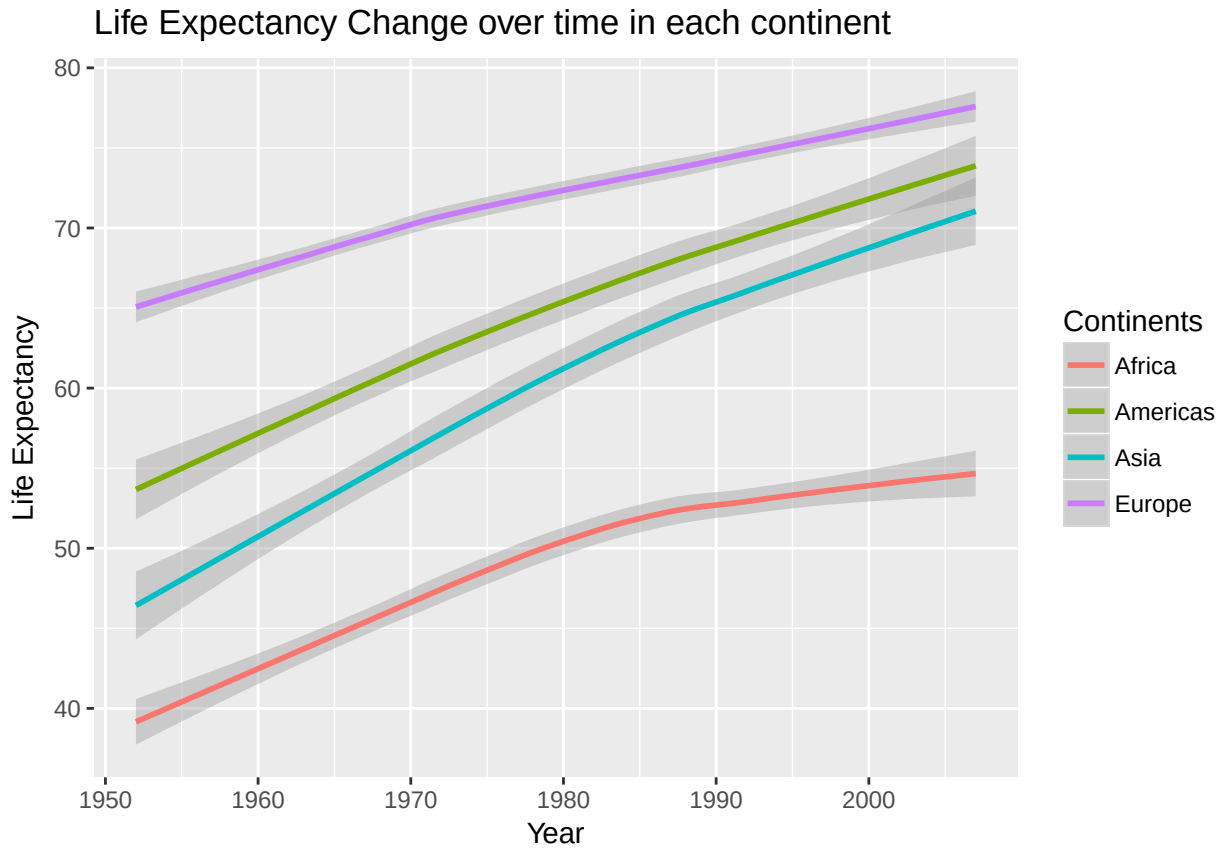


Thus, the data is fairly well described by a straight line below y = x incidicating a multiplicative shift.

Question 2: Life Expectancy Over Time By Continent

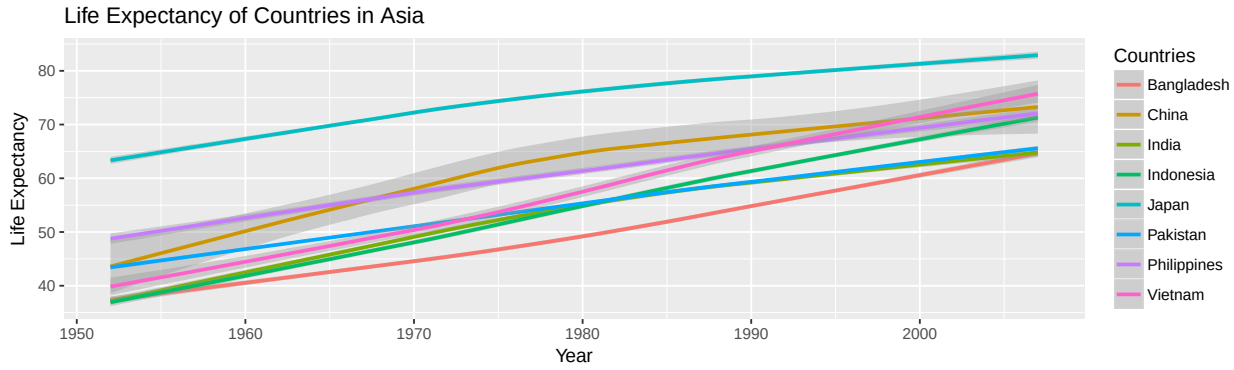
```
data = subset(gapminder, gapminder$continent != "Oceania")
first_data = data[c("continent", "year", "lifeExp")]
```

```
ggplot(first_data, aes(x = first_data$year, y = (first_data$lifeExp),
    color = first_data$continent)) + geom_smooth(method.args = list(degree = 1)) +
    xlab("Year") + ylab("Life Expectancy") + ggtitle("Life Expectancy Change over time in each continen
    labs(colour = "Continents")
```



From the above graph we can see that Asia's overall life expectancy has increased significantly from around 46 to 72 over the period 1950-2007.Alongside, Americas and Africa have also shown significant improvement in their life expectancies.

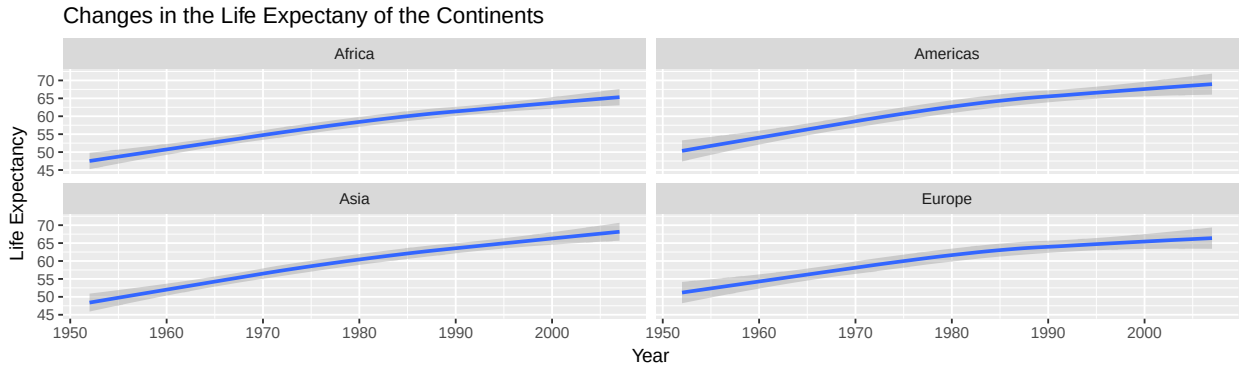
Let us inspect the life expectancy growth in Asia in more depth.Asia is noted for its size and huge population, but this is characterized by some countries with a very high population while many of them being sparsely populated.Hence we decided to have a population threshold (50 million) and observe life expectancies of only those countries with very high population as these are ones which will have the most interesting observations for life expectancy.



From the above graph we can see that there are 8 countries that are above our population threshold and all of the have shown a significant increase in their life expectancies.

How are the changes in the life Expectancy of the Continents(Fast/Slow,Linear)

```
third_data = data[c("continent", "year", "lifeExp", "gdpPercap")]
ggplot(third_data, aes(x = third_data$year, y = third_data$lifeExp)) +
    geom_smooth(method.args = list(degree = 1)) + facet_wrap(~third_data$continent) +
    xlab("Year") + ylab("Life Expectancy") + ggtitle("Changes in the Life Expectany of the Continents")
```

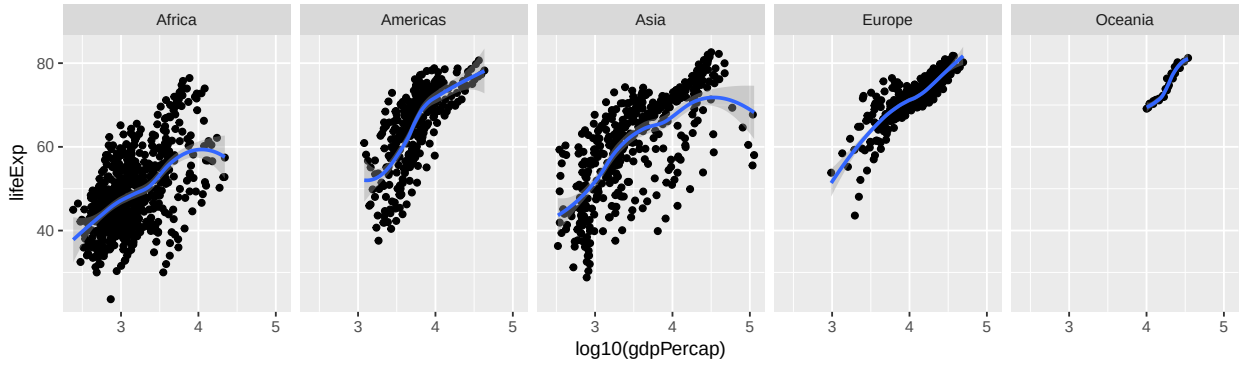


From the above plots we can observe that the life expectancies of all the continents have shown a somewhat linear increase over time. It seems that the continent doesn't matter when it comes to investigating life expectancies as it used to earlier since a lot of developing countries, primarily in Asia and Africa have shown a significant increase in their life expectancies over time and hence other aspects must be looked upon.

Question 3: Changes in relationship between GDP and life expectancy over time

If we see the counts of the continent in the summary, there is a variation among the continents with Africa having the highest number of observations with 624 and Oceania having the lowest with 24. Hence, it will make more sense to take the weighted average instead of just normal averages. Also, the skewness of the gdpPercap makes it important to transform the data.

```
ggplot(gapminder, aes(log10(gdpPercap), lifeExp)) + geom_point() +
    geom_smooth() + facet_grid(~continent)
```



Observing the above plot, a general trend observed is steep increase in lifeExp with increase in log10 of GDP. We will now try the loess model to further evaluate the relationship between the 2 variables and also check the time effect on lifeExp.

```
lm1 = loess(lifeExp ~ log10(gdpPercap), data = gapminder)
df1 = augment(lm1)
var(df1$.fitted)/var(df1$lifeExp)
```

```
## [1] 0.6573748
```

On fitting the loess model, we can see that the r-square value is 0.69 which means that the almost 70% of the variance in the lifeExp can be explained by GDP which is a good amount of variance but we still cannot entirely explain the variations which means there is scope for improvement in the modeling.

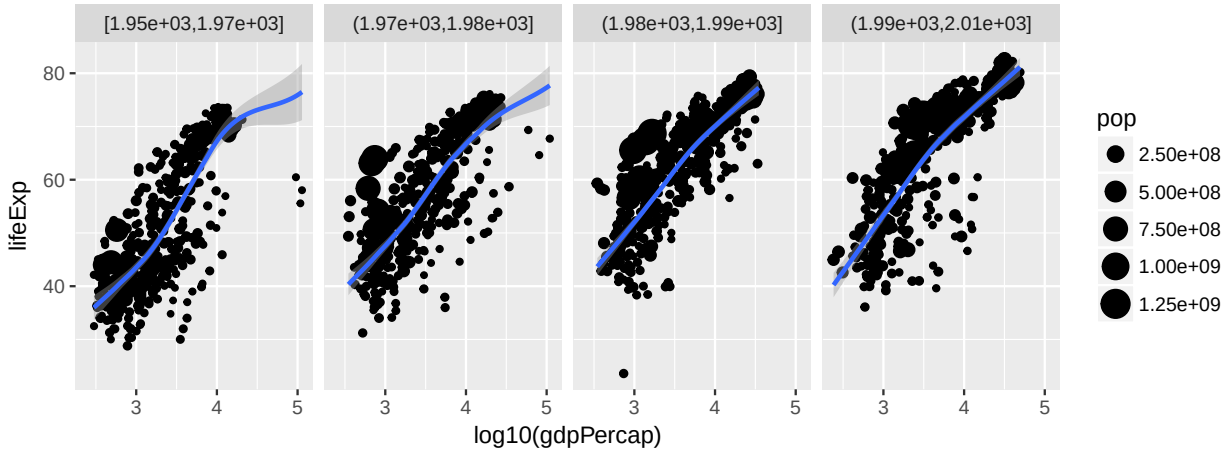
Checking the time effect in addition to the GDP effect.

```
lm2 = loess(lifeExp ~ log10(gdpPercap) + year, data = gapminder, degree = 1)
df2 = augment(lm2)
var(df2$.fitted)/var(df2$lifeExp)
```

```
## [1] 0.7176335
```

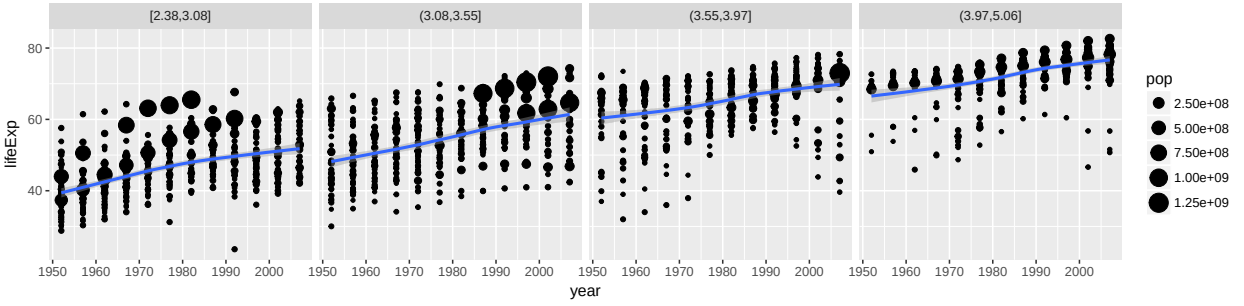
On adding the time effect, we can see that the R-square value increases to 0.85 which means that the 85% of the variance can now be explained which is an improvement of almost 15% over the previous model.

```
ggplot(gapminder, aes(x = log10(gdpPercap), y = lifeExp)) + geom_point(aes(size = pop)) +
    geom_smooth(method.args = list(degree = 1)) + facet_grid(~cut_number(year,
    n = 4))
```



The plots above confirm the variation of data with respect to time, we see that for different timeframes, the GDP shows an increase in the lifeExp.

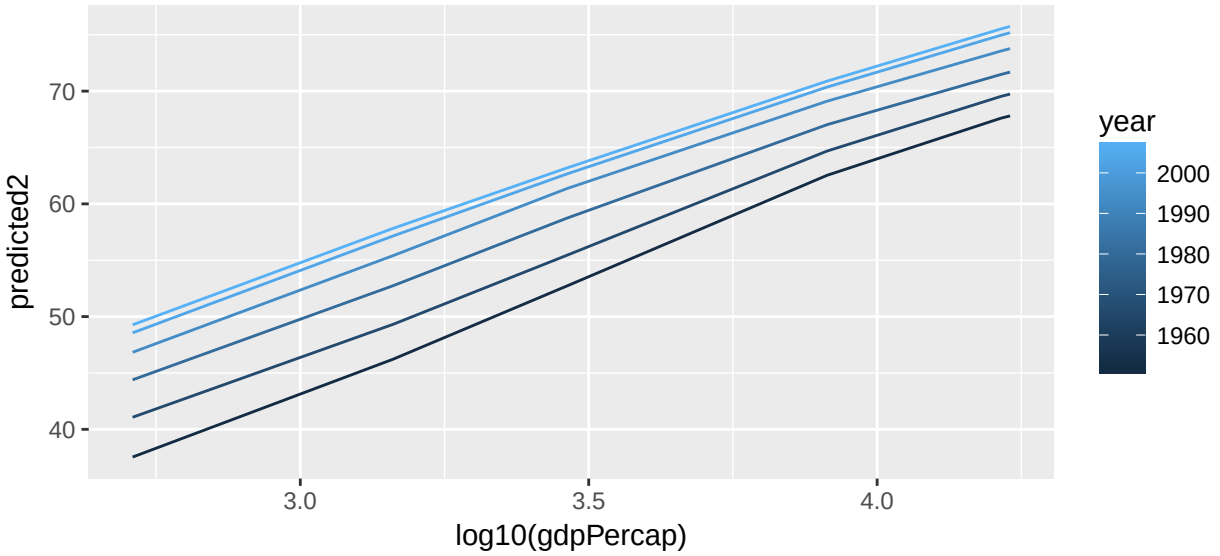
```
ggplot(gapminder, aes(x = year, y = lifeExp)) + geom_point(aes(size = pop)) +
    geom_smooth(method.args = list(degree = 1)) + facet_grid(~cut_number(log10(gdpPercap),
    n = 4))
```



Here, for the different ranges of GDP as time is increasing the lifeExp is also increasing gradaully which is indicated by the small slope. We will now check for convergence.

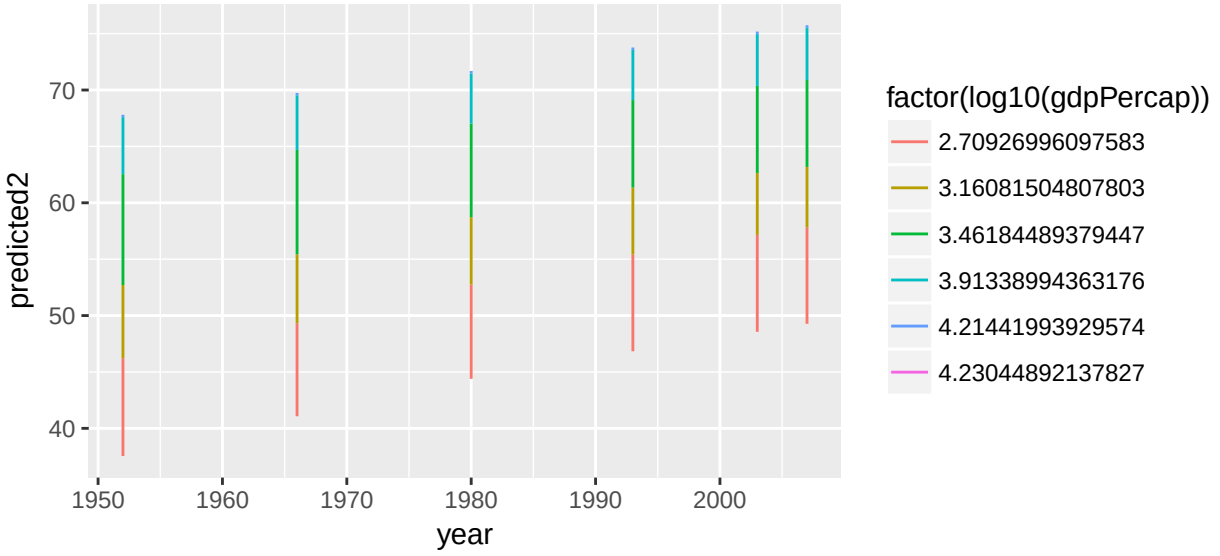
```
grid1 = expand.grid(year = c(1952, 1966, 1980, 1993, 2003, 2007),
    gdpPercap = c(512, 1448.155, 2896.309, 8192, 16384, 17000))
pred1 = predict(lm2, grid1)
```

```
pred2 = data.frame(grid1, predicted2 = as.vector(pred1))
ggplot(pred2, aes(x = log10(gdpPercap), y = predicted2, group = year,
    color = year)) + geom_line()
```



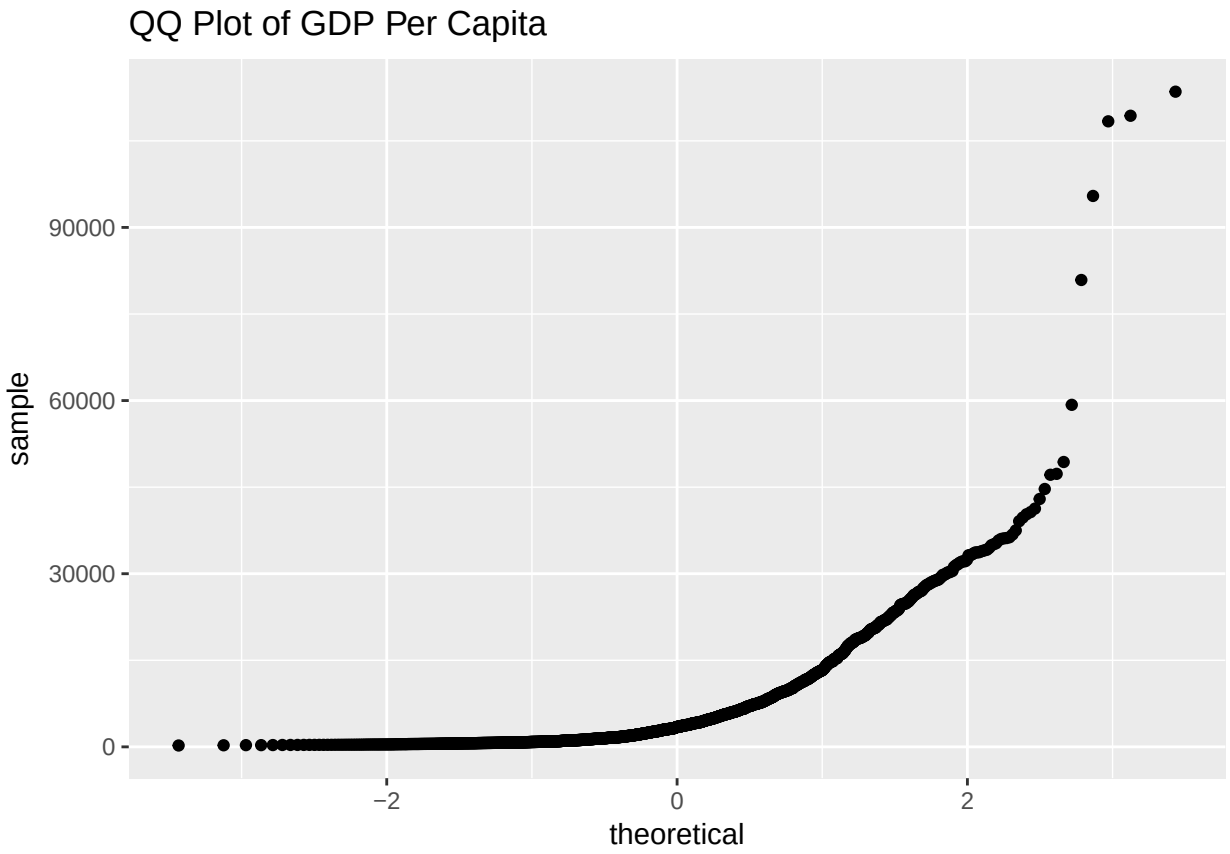
Here the different shaded lines seem to be parallel to each other and hence convergence is not a reasonable observation for the different timeframes.

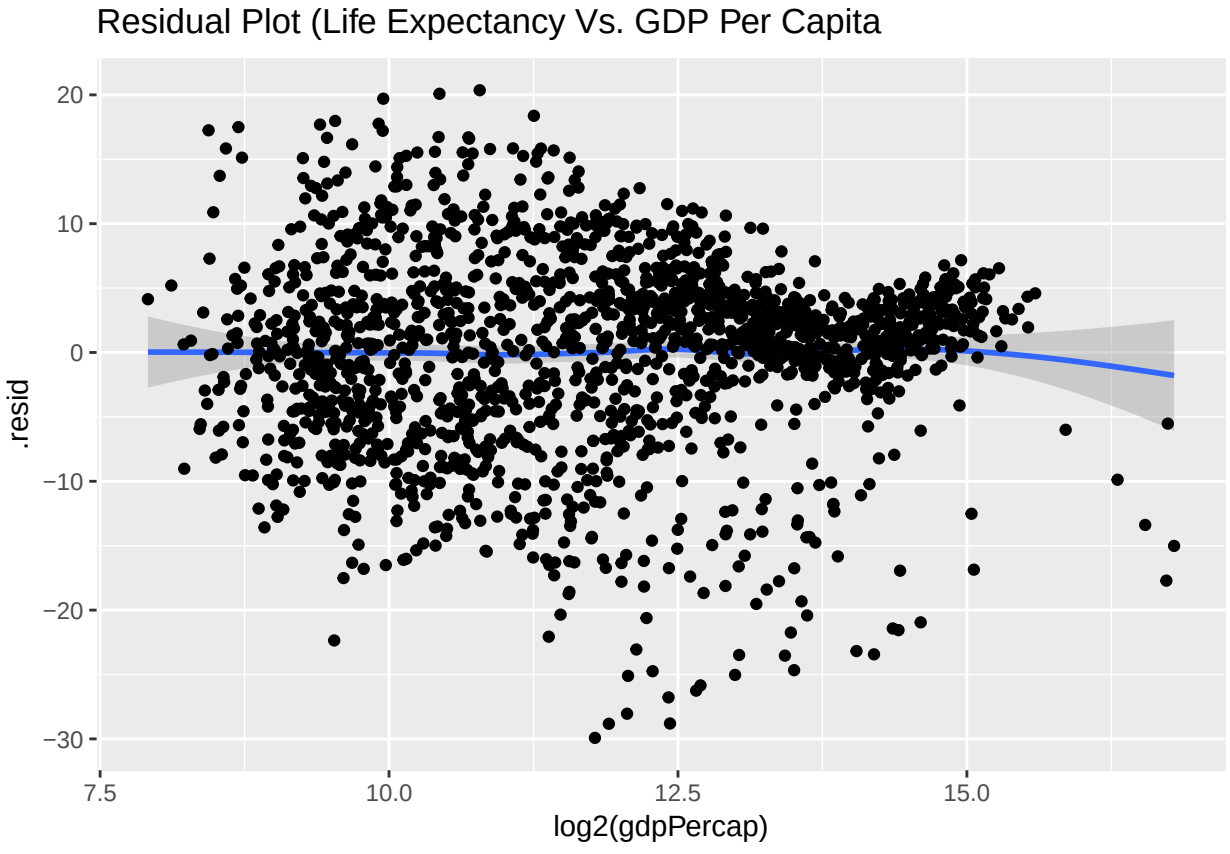
```
grid1 = expand.grid(year = c(1952, 1966, 1980, 1993, 2003, 2007), gdpPercap = c(512,
    1448.155, 2896.309, 8192, 16384, 17000))
pred1 = predict(lm2, grid1)
pred2 = data.frame(grid1, predicted2 = as.vector(pred1))
ggplot(pred2, aes(x = year, y = predicted2, group = year, color = factor(log10(gdpPercap)))) +
    geom_line()
```

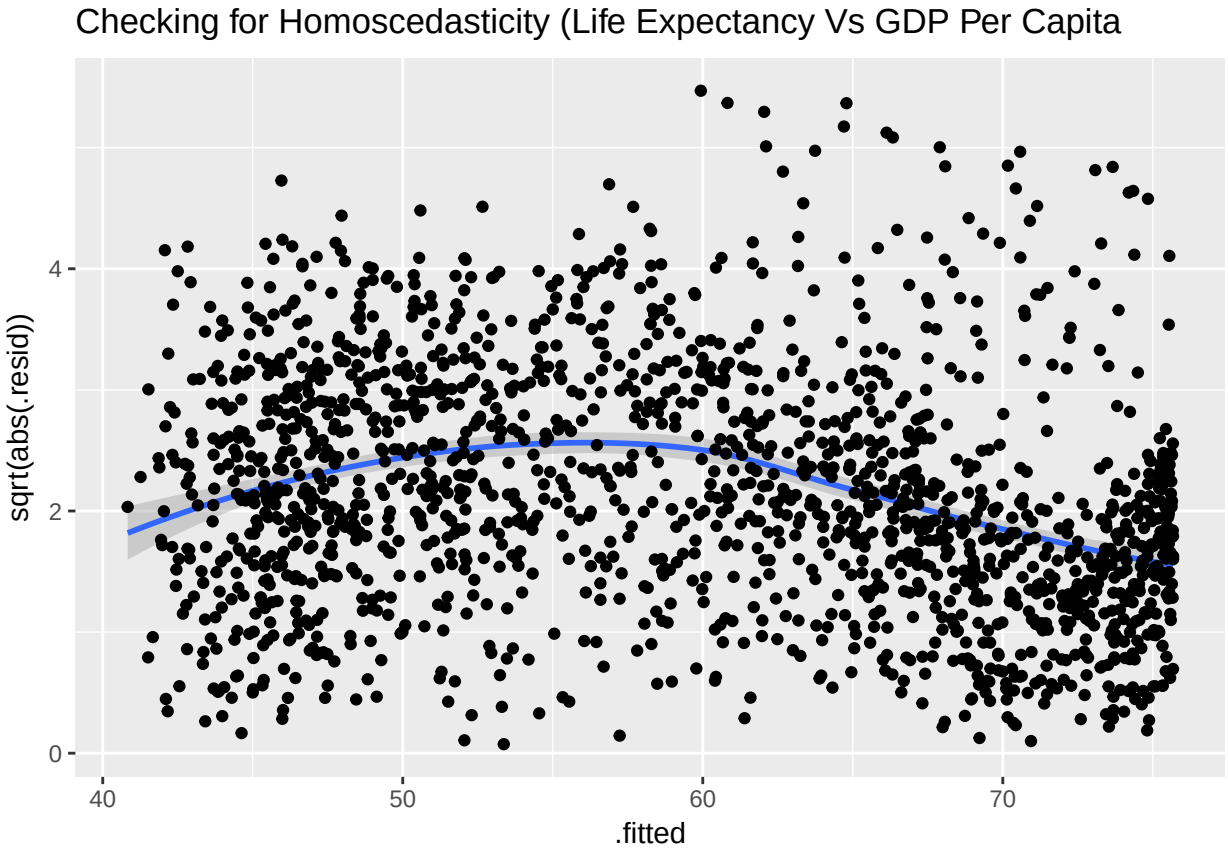


On observing the above plot, we see a gradual increase in the predicted lifeExp, but here again we cannot say there is convergence of any sort. I don't see any exceptions to the general pattern as we are getting parallel lines for both the above plots for the different ranges of years and GDP.

APPENDIX


QQ Plot of GDP Per Capita
sample
90000
60000
30000
0
−2
0
2
theoretical


Residual Plot (Life Expectancy Vs. GDP Per Capita
.resid
20
10
0
−10
−20
−30
7.5
10.0
12.5
15.0
log2(gdpPercap)


Checking for Homoscedasticity (Life Expectancy Vs GDP Per Capita
sqrt(abs(.resid))
0
2
4
40
50
60
70
.fitted