Final Project: Factors affecting Life Expectancy

Team Gapminder

Chinmayee, Chinmayee, Jordan, Nitesh

Introduction:

Our research is concerned with discovering relationships between average life expectancy by country and various explanatory variables for each country. The explanatory variables we are specifically interested in for each country are per-capita gross domestic product (GDP), per-capita healthcare spending, income inequality as measured by the gini coefficient, and continent.

We know that per-capita GDP is a very strong predictor of average life expectancy and that this prediction is even stronger when the continent of the country is considered. That said, the goal of our research is to discern whether the gini coefficient or per-capita healthcare spending explain variation in average life expectancy beyond what per-capita GDP and continent can explain. Using datasets from Gapminder, a non-profit organization dedicated to providing free datasets for purposes like ours, we sought to answer the following questions about the relationships between average life expectancy by country and our explanatory variables:

- What is the relationship, if there is any, between gini coefficient and average life expectancy after controlling for per-capita GDP and continent?
  - Is this relationship different if we also control for per-capita healthcare spending?
  - Is this relationship different for different continents or different time periods?
- What is the relationship, if there is any, between per-capita healthcare spending and average life expectancy after controlling for per-capita GDP and continent?
  - Is this relationship different if we also control for the gini coefficient?
  - Is this relationship different for different continents or different time periods?
- Does the relationship between gini coefficient and average life expectancy depend on the country's per capita GDP, continent, or per-capita healthcare spending?
- Does the relationship between per-capita healthcare spending and average life expectancy depend on the country's per capita GDP, continent, or gini coefficient?

To answer these questions, we decided to fix most of our analysis in one particular year to control for the effect that time might have on our relationships, and we chose the year 2009 for data-availability related reasons. After preprocessing our data, we're left with a dataset that consists of 5 measurements for each of our observations where the measurements are our 4 explanatory variables and average life expectancy, and each observation is an individual country.

Data Preparation

The Data used for this analysis is taken from the Gapminder package - (https://www.gapminder.org/data/). We used 4 datasets each containing yearly observations of a variable by country The four variables we are interested in: 1. Average Life Expectancy in Years - Data Files/GM-Life Expectancy- Dataset - v11.xlsx 2. Real GDP per Capita in US Dollars - Data Files/gdp_percap.csv 3. Dispersion of income throughout

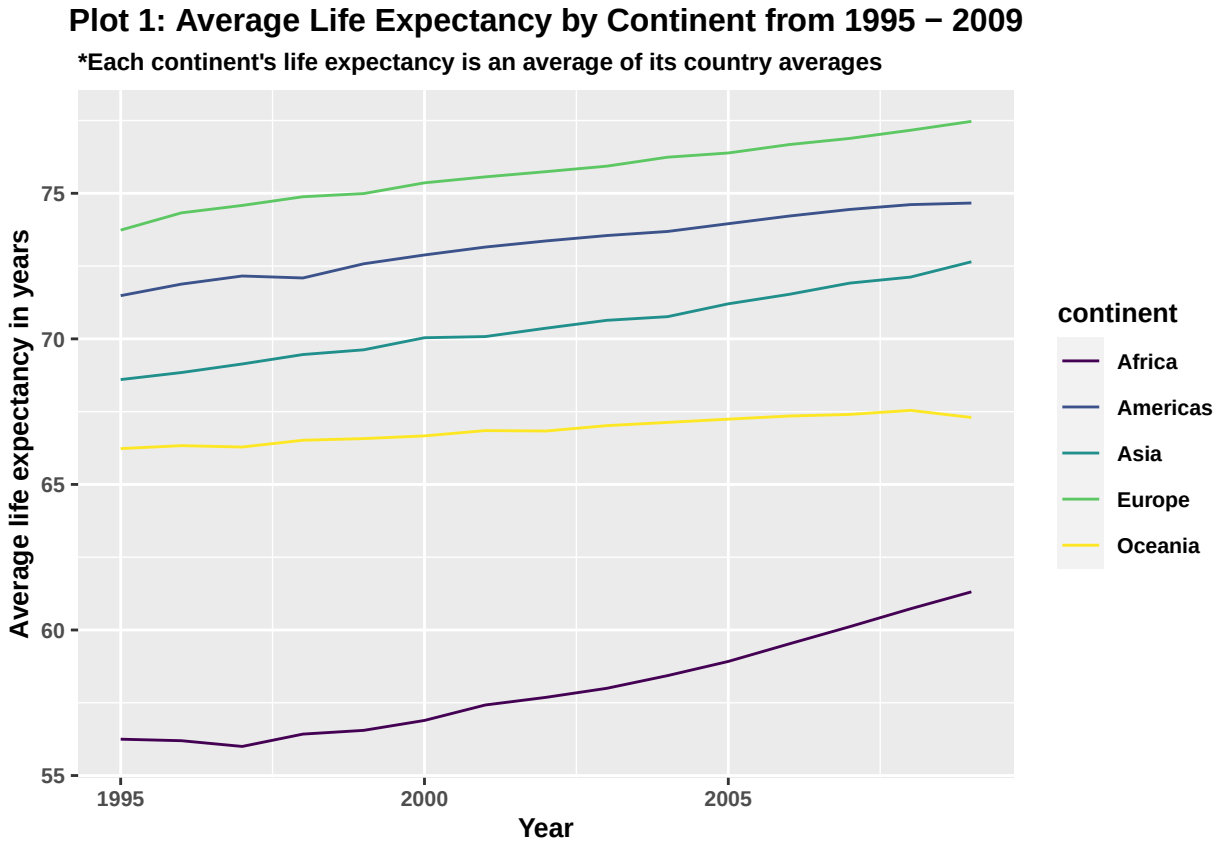
the population as measured by the gini coefficient - Data Files/__Gini Data - v3 - by Gapminder.xlsx 4. Healthcare spending in US dollars per capita - Data Files/HealthcareSpending.xlsx

After merging the 4 datasets we ended up with 175 countries for the period 1995-2009, this is the overlap between all datasets. We found that the data in Healthcare spending and GDP is extrapolated for the future/missing years and the Gini coefficient data has 40 as default for missing values. To avoid contamination of data and its ripple effect on modeling we decided to use the valid and actual data as much as we could.

To help visualize and draw insights better for our higher level analysis we grouped the countries by Continent for each year and aggregated the variable values. Based on the top level analysis we created our models but for modeling purposes we used country wise data without any aggregation, we also factored in the variable continent in our model because we believe that life expectancy can also be influenced by the living conditions, resources, geography etc of the continent. After the pre-processing we ended up with a total of 2585 data points in our master dataset.

Additionally, we excluded Oceania from our faceted graphs since there are not many countries in Oceania to make any conclusion.

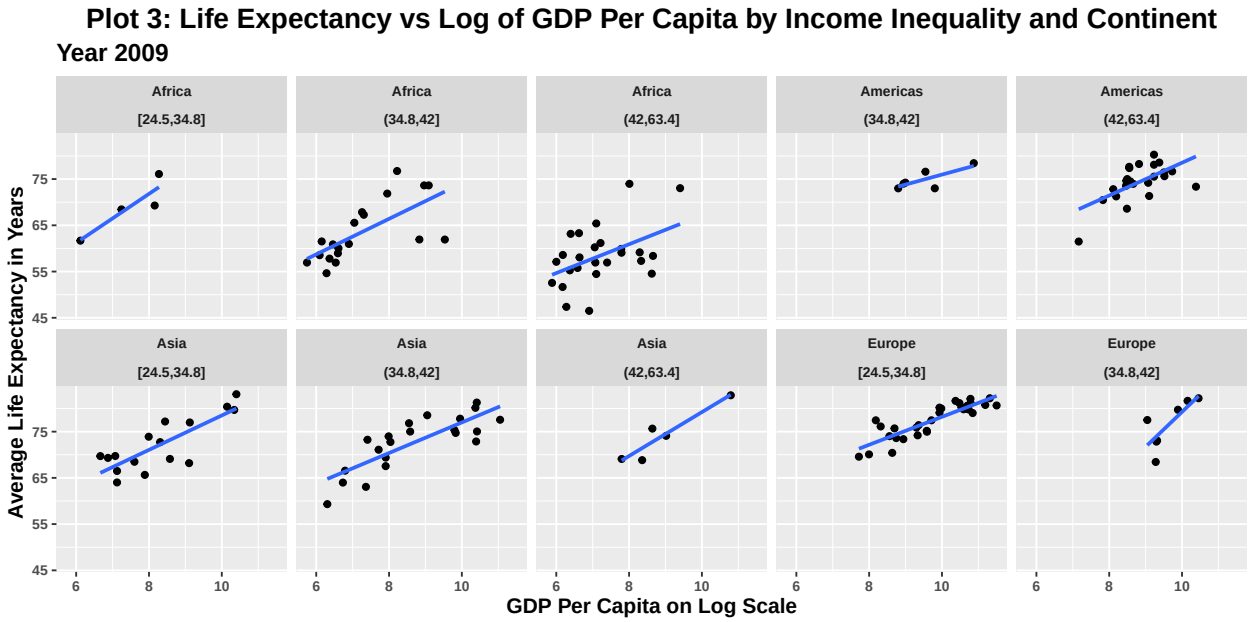
Data Analysis



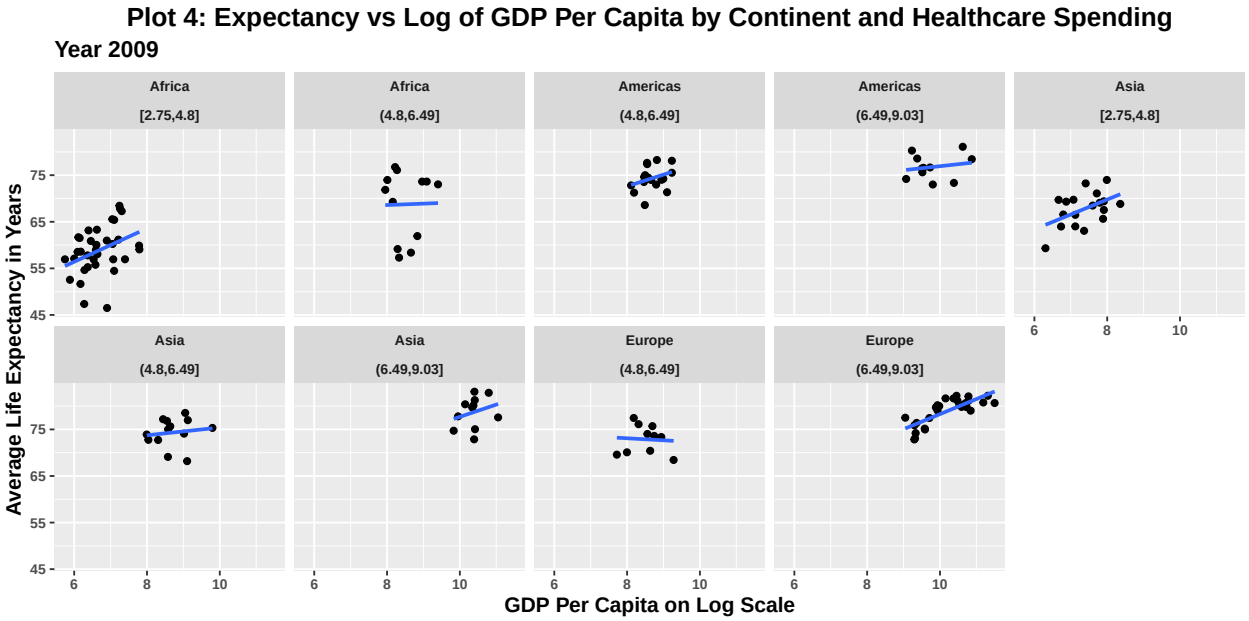
Average life expectancy has been increasing in the period 1995-2009 across continents. However, we can see a considerable difference between the 5 continents, with Africa having the lowest life expectancy and Europe the highest. Life expectancy is known to be directly dependent upon the development status of a country. It is considered that developed countries have more resources and advanced infrastructure to provide proper healthcare, which directly or indirectly leads to higher average life expectancies than those of economically weak/developing countries. One of the key indicators for a country's development is gross domestic product per capita. The more the GDP per capita of a country, the developed a country is assumed to be. However, in this project, we will look into economic aspects such as income inequality and government spending in

healthcare besides GDP per capita that explains higher or lower life expectancy in a particular country.

As expected, we see a direct relationship between life expectancy and GDP per capita. Additionally, we can observe lower life expectancy in African and American countries with high income inequality and vice versa. Also, there is a direct relationship between life expectancy and total health spending per capita, i.e., countries with high total health spending per capita have higher life expectancy than countries that spend less $ per capita on healthcare. Nevertheless, we cannot ignore the influence of GDP per capita in the latter relationship. As for the same proportion of health spending of GDP, a country with high GDP per capita has higher total health spending than a country with low GDP per capita. Hence, it become imperative to look into other plots to understand these relationships.



A few points to note here are that there is no country in Europe with gini more than 42 (high inequality group), and no country in the Americas has gini less than 34.7 (low inequality group) except Canada (removed). It is interesting to see that African countries with higher income inequality have lower life expectancy than countries with lower income inequality, whereas the trend is non-conclusive for other continents. Also, in our analysis, we did not see any significant difference in the trend between the years 1995 (appendix) and 2009. Hence, we checked this relationship by fitting a model for the year 2009 alone.

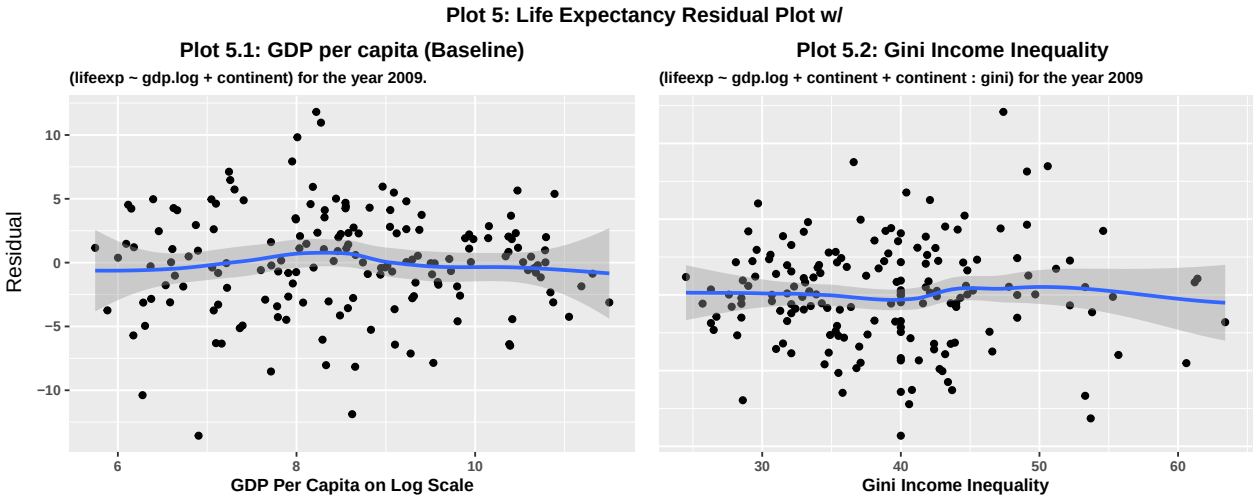


Countries with higher health spending per capita seem to have higher life expectancy than the countries with lower spending per capita on healthcare. However, we can notice that the countries with higher per-capita healthcare spending also have higher per-capita GDP which makes sense since countries that produce more econmic output per person tend to spend more money per person. While it does seem that per-capita healthcare spending does positively relate to average life expectancy, we believe this is probably only due to its relationship with per-capita GDP which appears to have more explanatory power. Moreover, our findings regarding this matter do not change significantly for the years between 1995 and 2009.

Modeling

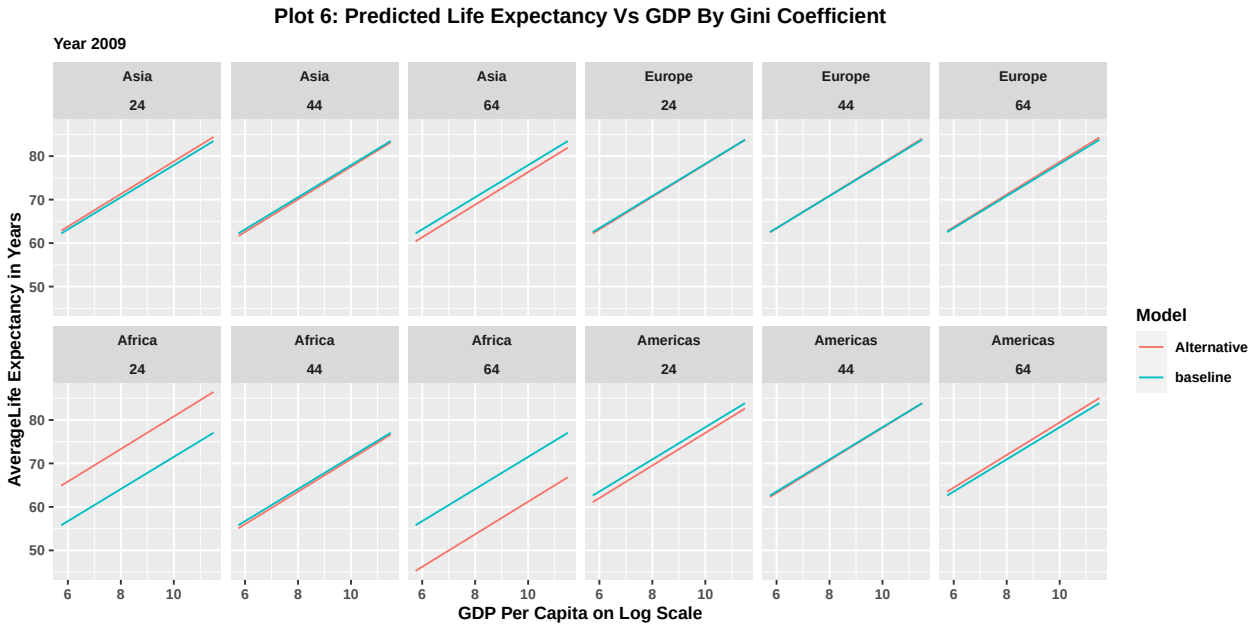
In order to more formally evaluate the explanatory power of per-capita healthcare spending and income inequality, we will first develop a baseline model using only per-capita GDP and continent as predictors, and see if we can improve on that model by adding terms that include per-capita healthcare spending or gini coefficient.

The best model we could fit that can include the gini coefficient and per-capita healthcare spending as predictors was a linear model that includes the main effects of per-capita GDP on a log scale and continent as well as an interaction between continent and gini coefficient. Moreover, none of the models we considered changed significantly when adding per-capita healthcare spending was considered which is consistent with our findings from plot 4. We decided to use a linear model because the non-parametric models we tested which included GAMs and Loess models did not offer significant imporvements in prediction, and linear models are the most interpretable. Below are residual plots of the baseline and alternative models which show that the models' errors are pretty randomly scattered about the line $y = 0$.



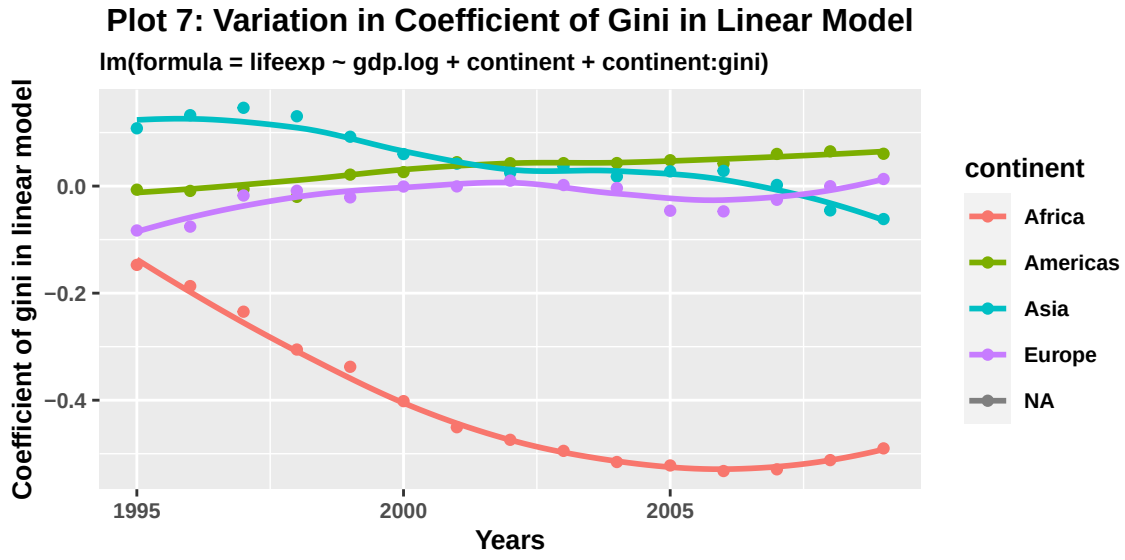

```
## lm(formula = lifeexp ~ gdp.log + continent + continent:gini,
##     data = gapminder.2009)
##                         coef.est coef.se
## (Intercept)              55.10     3.32
## gdp.log                   3.75     0.27
## continentAmericas       -17.01     6.14
## continentAsia           -12.30     5.05
## continentEurope         -14.75     5.47
## continentOceania         12.63    14.02
## continentAfrica:gini     -0.49     0.07
## continentAmericas:gini    0.06     0.11
## continentAsia:gini       -0.06     0.11
## continentEurope:gini      0.01     0.14
## continentOceania:gini    -0.83     0.35
## ---
## n = 171, k = 11
## residual sd = 3.50, R-Squared = 0.84
```

As you can see from the model display above, our model appears to be explaining a lot of variation in average life expectancy by country. As expected we see a positive coefficient next to per-capita GDP. From our analysis so far, we know that African countries have lower life expectancies than Asian, European, and American countries. However, for the continent which is a factor variable, Africa is the baseline factor, and all the coefficients next to each continent are negative besides Oceania which we can ignore since there are so few countries in that continent. This is the case because the relationship between the gini coefficient and average life expectancy is, assuming our model is perfect, negative in Africa after controlling for per-capita GDP. Our model suggests that this isn't the case in other continents, and so it doesn't decrease its prediction in non-African countries depending on its gini coefficient. Nevertheless, the interaction term between the gini coefficient and Africa is very interesting considering the interaction terms for Asia, Europe, and the Americas are insignificant. This would suggest that our alternative model only differs in prediction with the baseline model if the country the models are predicting on is in Africa. We see this to be case when we plot the models below:



As you can see from the plot above, our model is only making predictions that are significantly different for countries in Africa. As the gini coefficient increases across the 3 different values it is fixed for in the plot above, you can see that for Africa, the models predictions shift downward significantly. It does appear that a higher gini coefficient does predict lower average life expectancy but only in Africa. In other countries, it appears that the baseline model and the new model making almost exactly the same predictions which suggests that the gini coefficient doesn't explain any variation in average life expectancy across countries beyond what per-capita GDP and continent already explain. Assuming our model isn't over-fitting, this suggests that, yes, the gini coefficient adds explanatory power, but only in Africa. To test our model for over-fitting, we can see how the coefficients of the model change as we train it on data in different fixed years. This approach leaves our interpretation of the importance of the gini coefficient and how that changes over time susceptible to temporal autocorrelation, but seeing how our model differs between years can yeild some valuable insight nevertheless:

Over the years coefficient trend.



As you can see from the plot above, the model term for the gini coefficient is close to 0 and somewhat constant

over the time period from 1995 for the Americas, Europe, and Asia. More interestingly, we see that the coefficient for that term in Africa starts negative but close to 0 and decreases significantly, or moves further away from 0, over the time period from 1995 to 2009. This suggests that, assuming our model isn't similarly over-fitting the data in African countries for each year from 1995 to 2009, the gini coefficient's ability to explain variation in average life expectancy in African countries has increased dramatically from 1995 to 2009. Our model comparisons in 2009 and over the 1995 to 2009 time period yield very interesting results for African countries, but these results shouldn't be interpreted to mean there's a causal relationship between income inequality and average life expectancy. The dramatic increase in the gini coefficient's importance in predicting average life expectancy over the time period 1995 to 2009 suggests that the gini coefficient probably has some explanatory power but it's more likely the result of confounding involving complicated relationships among other variables that are outside the scope of our research.

Conclusion

We found that per-capita healthcare spending does not explain any variation in average life expectancy beyond what per-capita GDP already explains. As one might expect, per-capita healthcare spending and per-capita GDP are very closely related since countries that produce more per person tend to spend more per person. Per-capita healthcare spending does explain variation in life expectancy, but it appears to explain the same variation that per-capita GDP explains. However, we found per-capita GDP to be a stronger explanatory variable so per-capita healthcare spending does not add any explanatory power. These findings remain the same after also controlling for gini coefficient.

More interestingly, we found that the gini coefficient does explain some variation in average life expectancy by country beyond what per-capita gdp and continent can explain. However, the gini coefficient appears to only add explanatory power in African countries. In Europe, Asia, and the Americas, we found that the gini coefficient doesn't tell us any information about average life expectancy in a particular country beyond what the main effects of per-capita GDP and continent already tell us. However, in Africa, there appears to be a negative relationship between the gini coefficient and average life expectancy even after controlling for per-capita GDP. This relationship was very weak and perhaps insignificant in 1995 but grew more strongly negative over the time period from 1995 to 2009.

It does seem plausible that greater income inequality would imply lower average life expectancy because if two countries generate the same amount of wealth and resources over a particular time frame but one country allocates more resources to smaller subsets of people, that country probably has more people with less access to basic human needs which would probably result in a lower average life expectancy. However, we do not feel confident making this assertion since after adjusting for the main effects of per-capita GDP and continent, we only saw this relationship in Africa. Moreover, we believe it is possible that there's some variable or variables that are only relevant in Africa and are related to the gini coefficient that confound the relationship between the gini index and average life expectancy in African countries.

While our findings regarding the importance of the gini coefficient in predicting average life expectancy for a particular country in Africa don't yield any concrete conclusions about the precise nature of how income inequality impacts average life expectancy in a particular country, we believe that they serve as a basis for a more thorough investigation of factors that impact life expectancy. More thorough research could possibly explain why and how variation in the gini coefficient explains variation in average life expectancies in African countries 2009, or perhaps why our findings were so different or African and non-African countries, or perhaps why the gini coefficient's explanatory power beyond the main effects of per-capita GDP and continent. Future work in this area would probably involve the consideration of cultural, political, and geographic variables that could possibly expand our understanding of the relationships that seem apparent in our data anaysis and modeling.

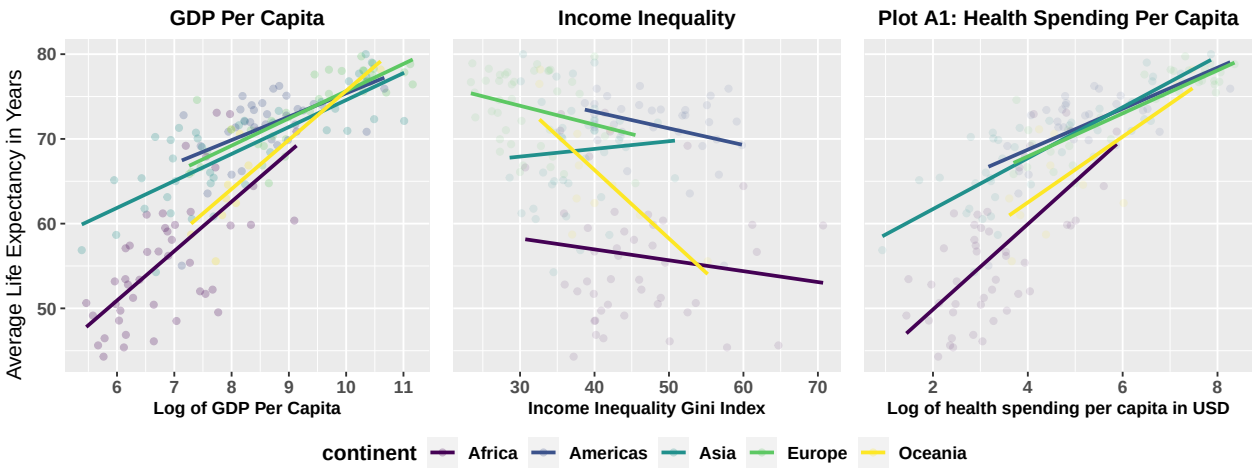
Our findings are interesting but come with the following limitations:

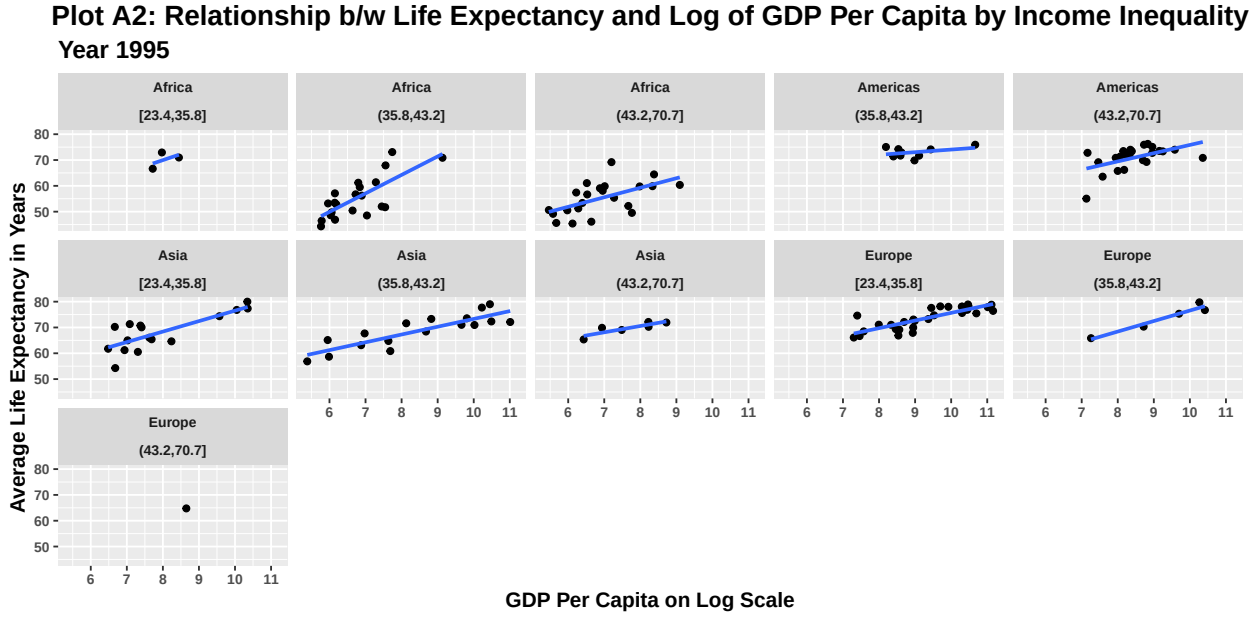
- The relationship we found between the gini coefficient and average life expectancy in Africa appear to have become more strongly negative from 1995 to 2009, but with the data we have it's impossible to say whether this relationship is real or whether there are some confounding variables.

- There are very different amounts of data in different areas of the feature space. For instance, typical African and American countries have greater income equality than Asian and European countries. Furthermore, typical African countries have lower per-capita GDPs than typical countries in the other continents. This sub-optimal distribution of observations across the feature space limits our ability to draw conclusions from our findings.
- We evaluated the explanatory power of many variables at once, and we only have about 170 countries to use as observations in any particular year. That said, it's difficult to distinguish between signal and noise when constructing plots and models.
- Our findings assume the data provided by Gapminder is perfectly accurate which is not a reasonable assumption in our case. It's entirely possible that average life expectancies or gini coefficients are, in actuality, very different than Gapminder's estimations.

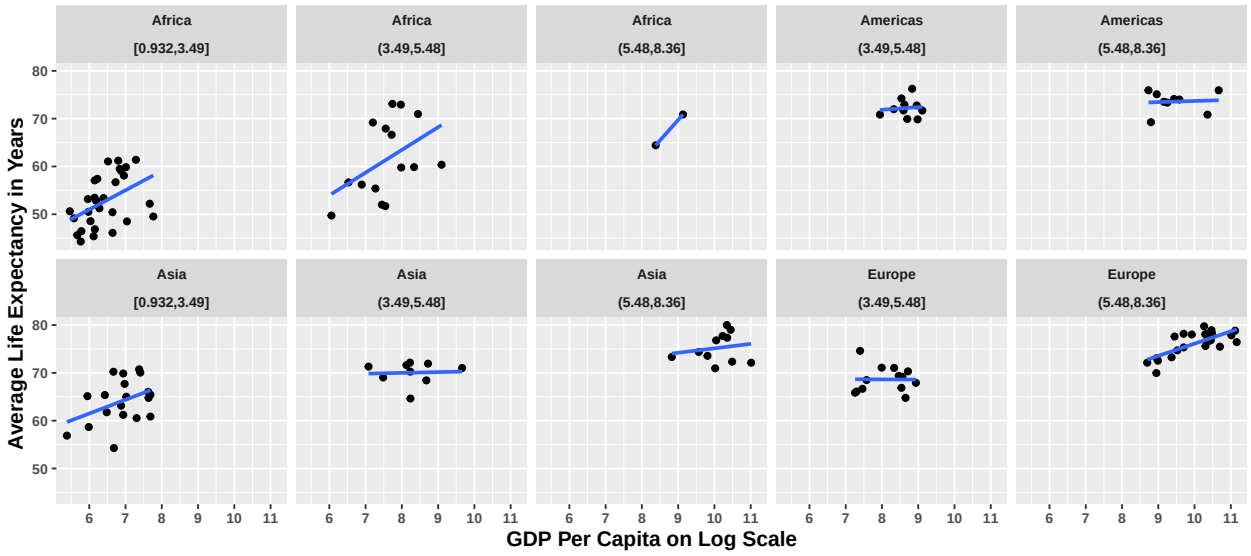
Appendix

Plot A1 : Relation of Life Expectancy w/





Plot A3: Relationship b/w Life Expectancy and Log of GDP Per Capita by Healthcare Spending Year 1995



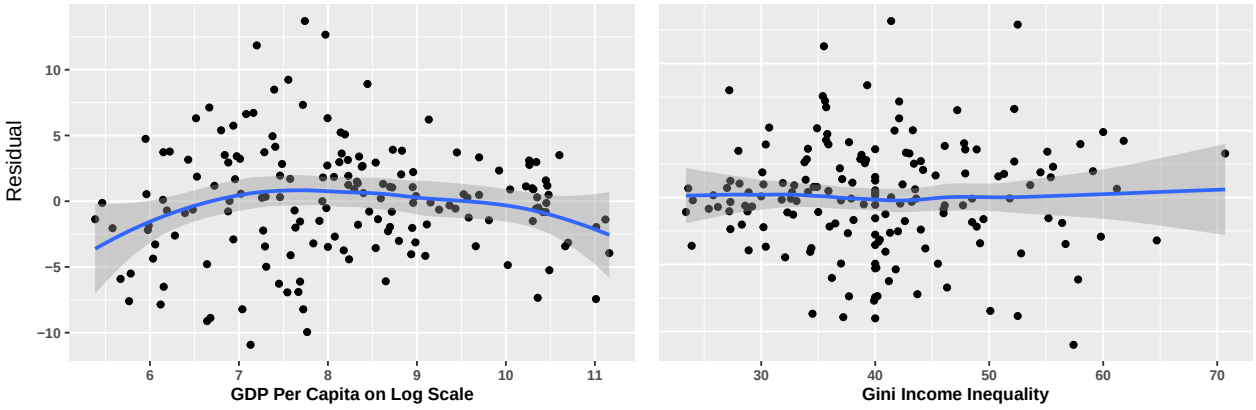
Plot A4: Life Expectancy Residual Plot w/

Plot A4.1: GDP per capita (Baseline)

(lifeexp ~ gdp.log + continent) for the year 1995

Plot A4.2: Gini Income Inequality

(lifeexp ~ gdp.log + continent + continent : gini) for the year 1995




```
## lm(formula = lifeexp ~ gdp.log + continent + continent:gini,
##     data = gapminder.1995)
##                       coef.est coef.se
## (Intercept)           37.40     4.03
## gdp.log                3.70     0.31
```

```
## continentAmericas       2.66     6.66
## continentAsia          -2.87     6.37
## continentEurope         4.11     5.59
## continentOceania       15.81     9.41
## continentAfrica:gini   -0.15     0.08
## continentAmericas:gini -0.01     0.11
## continentAsia:gini      0.11     0.14
## continentEurope:gini   -0.08     0.14
## continentOceania:gini  -0.45     0.21
## ---
## n = 167, k = 11
## residual sd = 4.28, R-Squared = 0.79
```