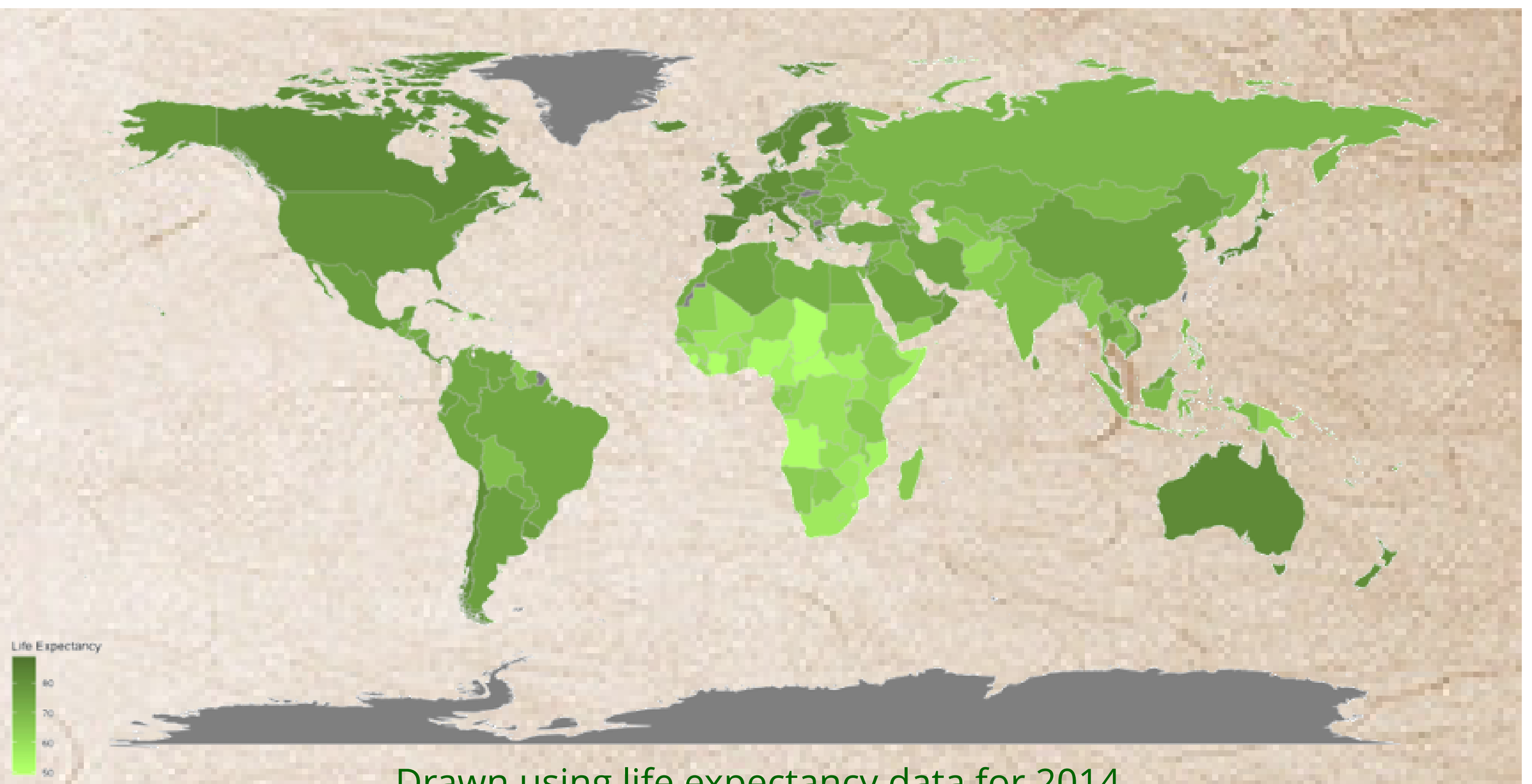
Life Expectancy Analysis and Prediction

Xuehan Zhao, Xinyue Li | Supervised by Professor Ranjini Grove
Department of Statistics, University of Washington



Drawn using life expectancy data for 2014

We found that some countries, such as the United States, spend 2-5 times more on healthcare than other developed countries while having shorter life expectancies. Does this imply that higher expenditure on healthcare do not necessarily lead to longer life expectancy? Our group is curious about this phenomenon and is interested in figuring out some potential factors that might affect the life expectancy of people and predicting how long a baby born in a particular country can expect to live.

PROBLEM

What are some potential factors that might affect the life expectancy at birth of people?

INTRODUCTION

We began by coming up with some factors that we think may have large impacts on life expectancy. We have considered factors from economics, environment, healthcare, diseases prevention, and food aspects. Then, we searched to see whether there are available datasets for these variables. The variables we used are described in the table below:

name	unit	meaning
Life expectancy	yr	Life expectancy at birth indicates the number of years a newborn infant would live if prevailing patterns of mortality at the time of its birth were to stay the same throughout its life. in years
health	$	Government expenditure on health per capita.
water	%	Improved water source (% of population with access to)
GDP	$	GDP per capita
measles	%	Child immunization against measles (% of children ages 12-23 months)
PM2.5	%	PM 2.5 air pollution (population exposed to levels exceeding WHO guideline value, in % of total)
food		Gross per capita Food Production Index Number

METHODS

1. Collinearity, Variance Inflation Factor (VIF)
2. Multiple regression
3. Model selection (AIC, BIC, Cp and Adjusted R-squared as criteria)
4. ANOVA F-test
5. Prediction
6. Model diagnostics

RESULT

1. Collinearity

High correlation exists between many pairs of regressors, especially between GDP and health expenditure. So we double there could be a multicollinearity problem in the analysis. A check of Variance Inflation Factor (VIF) values shows that the VIF values of log(health) and log(GDP) are greater than 5, indicating the presence of multicollinearity.

2. Multiple Regression

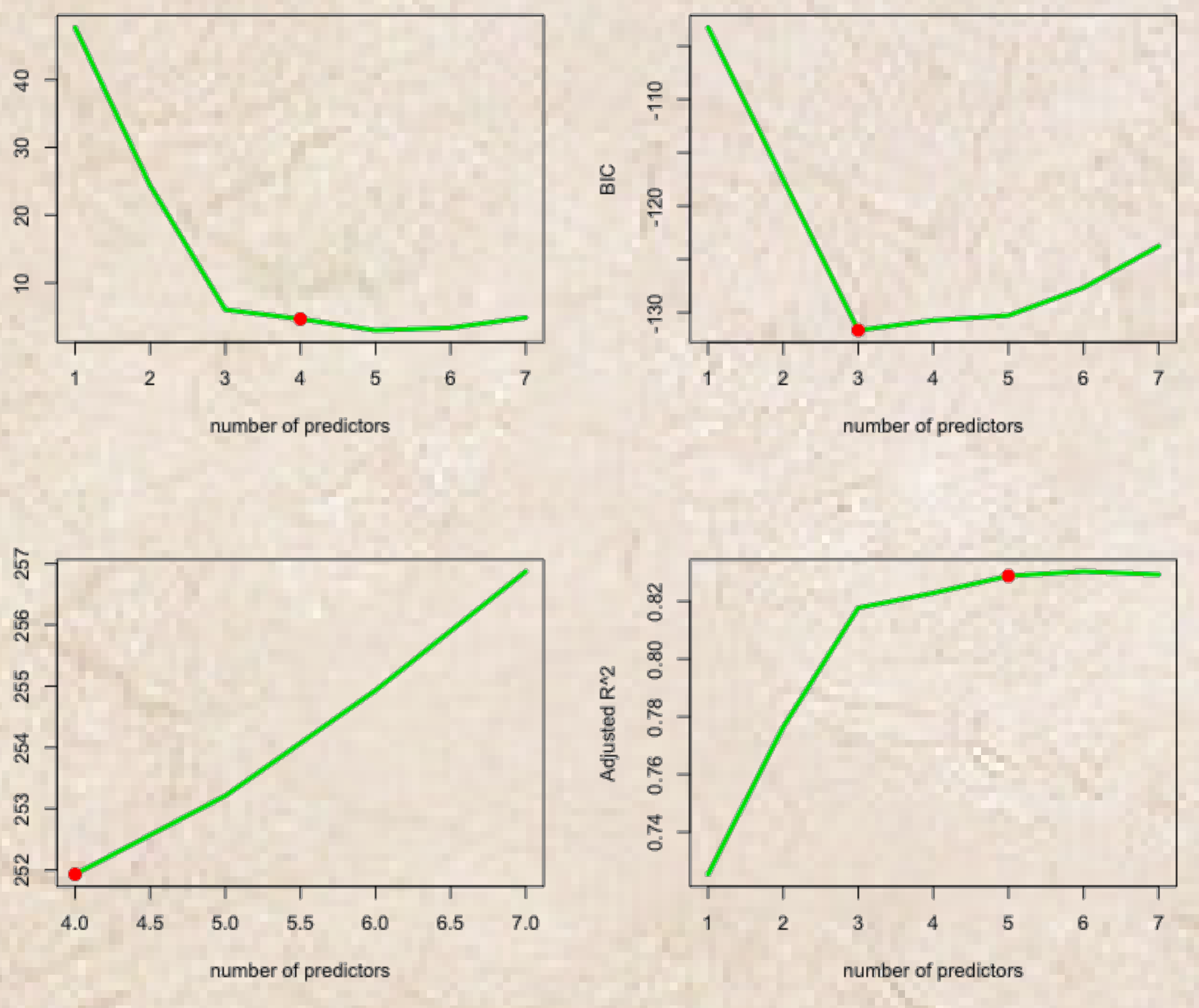
Overall Model:

$$\begin{aligned}\textit{Life Expectancy} &= \beta_0 + \beta_1 \times log(health) + \beta_2 \times \frac{water^9-1}{9} + \beta_3 \times log(GDP) \\ &+ \beta_4 logit(PM2.5) + \beta_5 food + \gamma_1 M_1 + \gamma_2 M_2 + \delta_1 R_1 + \delta_2 R_2 + \delta_3 R_3 \\ &+ \delta_4 R_4 + \delta_5 R_5 + \ldots\ldots + \varepsilon\end{aligned}$$

Note: M1 = (50,75], M2 = (75,100], R1 = Europe, R2 = North America, R3 = South America, R4 = Africa, R5 = Oceania; 2. "……" represents all interaction terms.

3. Model Selection

- Criteria: AIC, BIC, Cp and adjusted R-Square.
- We want to find the model that leads to
  - smallest AIC,
  - smallest BIC,
  - Cp that is closest to number of variables in the model plus one
  - or relatively stable adjusted R-Square.
- The best model would be the one including four predictors: health, water, measles(75,100] and regoin5. For the predictor measles(75,100] and region5
- We think that if one level of measles and region is significant, the variables measles and region should be considered significant. Therefore, we use health, water, measles and region to fit the regression.



* We then check the VIF values again. With all VIF < 5, there is no serious multicollinearity problem.

4. ANOVA F-Test

Before performing a multiple regression, it is necessary to check if any interaction terms are needed.

Full Model:

$$\begin{aligned}\textit{Life Expectancy} &= \beta_0 + \beta_1 \times log(health) + \beta_2 \times \frac{water^9-1}{9} + \gamma_1 M_1 + \gamma_2 M_2 + \delta_1 R_1 \\ &+ \delta_2 R_2 + \delta_3 R_3 + \delta_4 R_4 + \delta_5 R_5 + \textit{all interaction terms} + \varepsilon\end{aligned}$$

Reduced Model:

$$\begin{aligned}\textit{Life Expectancy} &= \beta_0 + \beta_1 \times log(health) + \beta_2 \times \frac{water^9-1}{9} + \gamma_1 M_1 + \gamma_2 M_2 + \delta_1 R_1 \\ &+ \delta_2 R_2 + \delta_3 R_3 + \delta_4 R_4 + \delta_5 R_5 + \varepsilon\end{aligned}$$

(Note: M_1 = (50,75], M_2 = (75,100], R_1 = Europe, R_2 = North America, R_3 = South America, R_4 = Africa, R_5 = Oceania)

The null hypothesis is that all coefficients of interaction terms are zero. The P-value = 0.8388 < 0.05. So, we fail to reject the null hypothesis. As a result, we decided to drop all the interaction terms and use the reduced model.

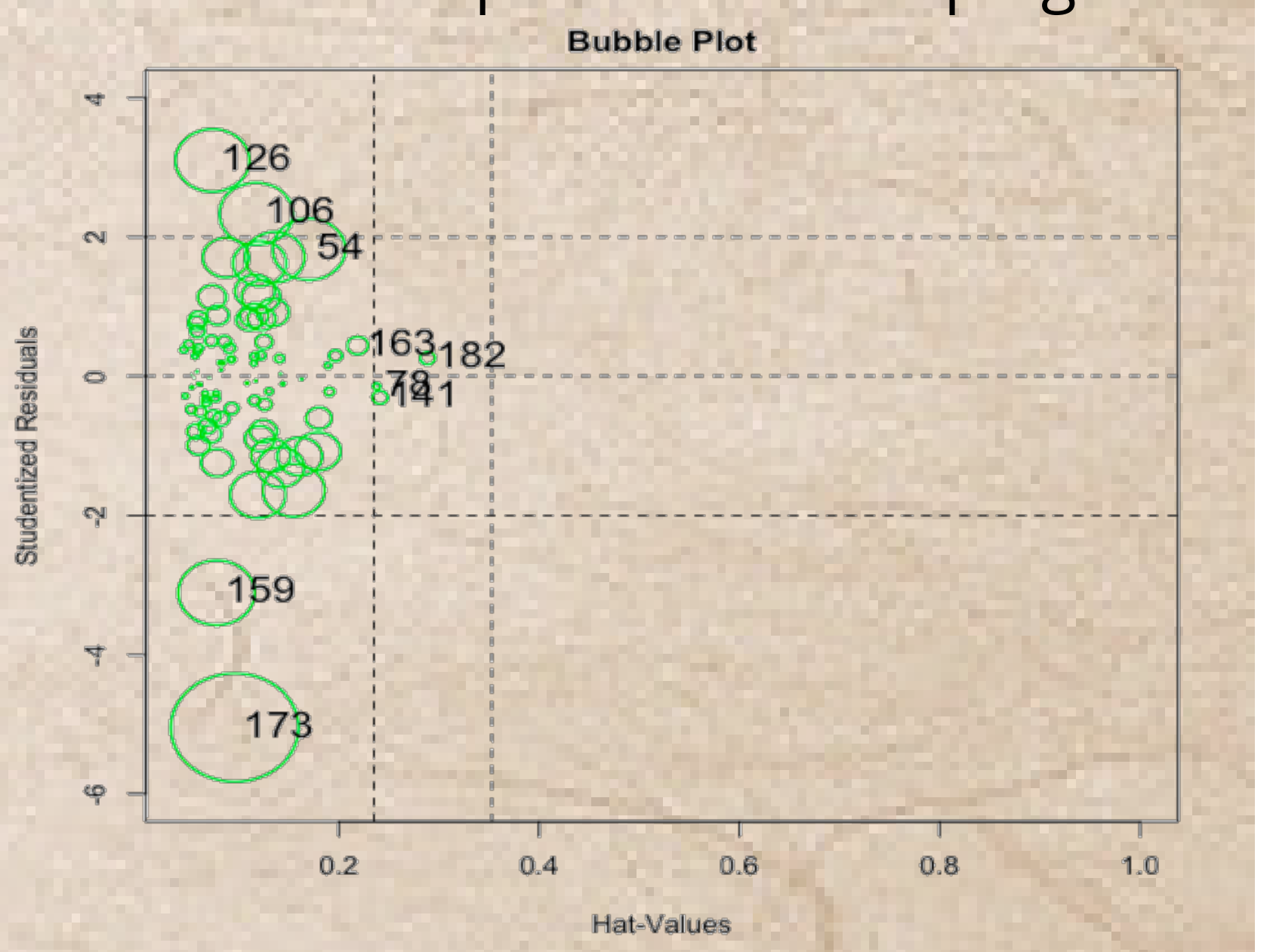
5. Prediction

Using the model, the expected life expectancy for:

- the United Kingdom is 80.52 while the true value is 81.10;
- Thailand is 74.53 while the true value is 74.40;
- Ethiopia is 56.42 while the true value is 64.00.

So, the model does a good job in the predictions for developed and developing countries.

6. Model Diagnostics



- Obs 173 (Swaziland), 159 (Sierra Leone), 126 (Morocco) and 106 (Libya) are outliers.
- Obs 173 (Swaziland), 159 (Sierra Leone) have high cook's distances.
- Obs 182 (Tonga), 78 (Haiti) and 141 (Papua New Guinea) have high leverages.
- Overall, obs 173 (Swaziland, the country with the lowest life expectancy) and 106 (Libya) appear to have the most influence on the results.

CONCLUSION

The project aims to find the factors that affect the life expectancy at birth. Based on the statistical analysis above, we conclude that health expenditure per capita, percentage of population with access to improved water source, percentage of children (age 12-23 months) got immunized against measles and the continent the country belongs to clearly have some impacts on the life expectancy at birth.

REFERENCES

UN Data. Vers. v0.14.6. 2017. United Nations Statistics Division. 10 Mar 2018.

DataBank World Development Indicator. 2017. The World Bank Group. 10 Mar 2018.

Food and Agriculture Organization of the United Nations. 2017. 10 Mar 2018.