Do alcohol consumption and immunization coverage affect life expectancy? A cross-sectional study based on samples of countries in 2014

Yihan Chen

December 20, 2022

Introduction

The influence of vaccines on public health has proved to be profound, and it is one of the most potent tools against diseases. A study on the 2001 birth cohort in the US showed that regular child vaccines prevented around 33000 deaths and were very cost-effective (Zhou et al., 2005). During the ongoing global COVID-19 pandemic, vaccines have also been crucial in preventing contagion. A noticeable phenomenon during COVID-19 is vaccine hesitancy, which refers to delay and even refusal of vaccines. A study showed that more than 30% of the public was reluctant to receive COVID-19 vaccination based on a national sample in 2020 in the US (Callaghan et al., 2021).

Another health concern that becomes even more salient during COVID-19 is alcohol consumption. A cross-sectional survey in Canada reveals that about 12% of the population increased their drinking frequency during COVID-19 (Thompson et al., 2021). Several studies have shown a negative correlation between life expectancy and alcohol consumption in Russia and Nordic countries (Inna et al., 2021; Olof et al., 2019).

Currently, the research on both topics tends to focus on samples from middle and high-income countries, with limited studies on low-income countries. Hence, this study aims to study the effect of alcohol consumption and immunization coverage on life expectancy based on samples from both developed and developing countries in 2014. So that countries can make more informed choices on health policies to improve life expectancy.

Methods

Study sample

This report is based on the Life Expectancy (WHO) dataset by KUMARRAJARSHI from https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who. The data were acquired from the WHO data repository and the United Nations website for 193 countries from 2000 to 2015. Six variables measure a country's economic level, and fourteen variables depict countries' demographic characteristics and public health levels. We will limit our data to only from 2014, so a multiple linear regression model is appropriate. 2014 is chosen because it is the dataset's most recent data with few missing values.

Assumption checking

This report aims to assess the effect of alcohol consumption and immunization coverage on life expectancy by fitting a multiple linear regression model on the dataset. Our model will have *Life.expectancy* as the dependent variable, which measures the life expectancy of a country. As for independent variables, we have *Alcohol*, *Polio*, *Diphtheria*, and *Hepatitis. B*. *Alcohol* measures the alcohol consumption in liters per capita. *Polio*, *Diphtheria*, *Hepatitis. B* are all measures of vaccine coverage among one-year-olds in percentage. Apart from these variables, we also account for confounding variables *GDP* and *Population* in our initial model. Among all variables, these two are likely to be correlated with both independent and dependent variables. Also, research shows evidence of adverse health effects due to income inequality based on a sample of 148 countries from 1970 to 2010 (Linden et al., 2017). After deciding on our initial model, we then fit the model based on our dataset.

We check the assumptions for our model:

- Construct a residual versus fitted plot to check linearity by assessing if there is any non-linear pattern. If there is a sign of violation, we should consider box-cox transformation.
- Check the independence assumption by assessing if our observations are correlated with each other.
- Plot the square root of the standardized residuals versus the fitted values to evaluate the homoscedasticity assumption. A box-cox transformation may be needed if the plot has a non-horizontal linear pattern.
- Check the normality assumption with a normal quantile-quantile plot of the standardized residuals. If the relationship is not one-to-one, we will apply the box-cox transformation to adjust.

It is crucial for us to also recheck the assumptions after transformation. However, since box-cox transformation affects our model's interpretability, it would not be appropriate for us to apply it based on our goal.

Identifying influential observations and multicollinearity

We then identify any influential observations for our model; it is vital to acknowledge their existence since they affect our regression line noticeably. Nevertheless, we should only remove them with a contextual reason. The main criteria are the cook's distance, DFFITS, and DFBETAS. The cook's distance measures the effect of individual observation on all fitted values by checking if the cook's distance is greater than the 50th percentile of the $F_{p+1,n-p+1}$. Where p is the number of predictors and n is the number of observation. In terms of DFFITS, it measures the effect of each observation on its fitted value by checking if the absolute value of test statistics is greater than $2\sqrt{(p+1)/n}$. On the other hand, DFBETAS focus on the change in estimated coefficients when one observation is removed. If the absolute value of the test statistics is greater than $2/\sqrt{n}$, we record the observation as influential. By assessing influential points using various criteria, we have a more comprehensive view of them.

Another potential issue to check is multicollinearity. We can identify multicollinearity by checking if any variance inflation factor is greater than 5. If multicollinearity occurs, we should respecify our model. However, we should keep all variables we are interested in based on our goal.

Model selection and validation

In order to improve the predictive power of our model, we consider variable selection to find the best set of predictors. This report will apply stepwise selection based on AIC, BIC, and LASSO. We then consider adding extra predictors into our model to improve the fit by assessing if the adjusted R^2 of the new model is higher than the initial model. After having the final model, we check all assumptions, influential observations, and multicollinearity again.

Finally, we estimate the prediction error of our final model using cross-validation. Since we have a small dataset, we split our dataset into three parts and fit the model with two parts. The remaining part is used as a test set to assess prediction errors. By using all three sets as test sets, we can use a calibration plot to assess the prediction accuracy of our model.

Results

Sample characteristics

The health and economic characteristics of 131 countries in 2014 are shown in Table 1. The average life expectancy is relatively high at 70.51. The average alcohol consumption per capita is 3.06 liters. In terms of immunization coverage, Polio, Diphtheria, and Hepatitis. B vaccine has similar mean coverage of about 80%. We also see large standard deviations for all these variables, indicating the variability across countries.

Table 1: Characteristics of 131 countries in year 2014

Variable names	Mean	Standard Deviation
Life expectancy	70.52	8.61
Alcohol consumption per capita	3.06	4.09
Polio vaccine coverage among one-year-olds	83.50	20.97
Diphtheria vaccine coverage among one-year-olds	83.89	21.84
Hepatitis.B vaccine coverage among one-year-olds	81.71	23.76
GDP per capita	7256.85	14741.40
Population	22269096.43	116699866.41
Adult mortality rate of both sexes	160.37	110.14
Health expenditure as a percentage of total expenditure	6.11	2.53
Deaths per 1000 due to HIV/AIDS	0.81	1.56
Human Development Index	0.67	0.15

Initial Model

Table 2: Association between life expectancy, alcohol consumption and immunization coverage based on two linear regression models

	Dependent variable:	
	Life.expectancy	
	Initial Model	Final Model
	(1)	(2)
Constant	57.293^{***} (2.582)	48.118^{***} (2.623)
Alcohol	0.764^{***} (0.168)	0.042 (0.091)
Polio	0.091^{**} (0.042)	−0.015 (0.021)
Diphtheria	0.059 (0.069)	0.021 (0.033)
Hepatitis.B	−0.031 (0.057)	0.003 (0.027)
GDP	0.0001^{***} (0.00005)	0.00001 (0.00002)
Population	−0.000 (0.000)	−0.000 (0.000)
Adult.Mortality		-0.019^{***} (0.004)
Total.expenditure		0.343^{***} (0.117)
HIV.AIDS		-0.862^{***} (0.239)
Income.composition.of.resources		34.395^{***} (3.286)
Observations	131	131
R^2	0.409	0.876
Adjusted R^2	0.381	0.866
Residual Std. Error	6.773 (df = 124)	3.151 (df = 120)
F Statistic	14.310^{***} (df = 6; 124)	84.977^{***} (df = 10; 120)
Note:		*p<0.1; **p<0.05; ***p<0.01

The initial model we have takes the form of:

$$
\begin{aligned}
y_{life\ expectancy} = \beta_0 &+ \beta_1 x_{Alcohol\ Consumption} + \beta_2 x_{Polio} + \beta_3 x_{Diphetheria} \\
&+ \beta_4 x_{Hepatitis.B} + \beta_5 x_{GDP} + \beta_6 x_{Population} + \epsilon
\end{aligned}
$$

Table 2 shows that the coefficient of *Alcohol* is 0.764, which is significant at a p-level of 0.01. Similarly, we see a positive association between *Polio*, *Diphtheria*, and *Life.expectancy.* Whereas the *Hepatitis. B* has a negative relationship with *Life.expectancy.* Among the three variables measuring vaccine coverage, only the coefficient for *Polio* is significant at a p-level of 0.05. The model has an adjusted R^2 of 0.381.

Based on Fig. 3 in the appendix, we conclude that linearity and homoscedasticity are satisfied. Since our data comes from independent countries, the independence assumption is satisfied. We may have a mild violation of the normality assumption. No transformation is applied due to the goal of the report.

Table 3: Variables selected based on different criteria

Selection Method	Variables Selected
AIC based selection	Adult.Mortality, infant.deaths, under.five.deaths, Total.expenditure, HIV.AIDS, Income.composition.of.resources
BIC based selection	Adult.Mortality, Total.expenditure, HIV.AIDS, Income.composition.of.resources
LASSO based selection	Adult.Mortality, Total.expenditure, HIV.AIDS, Income.composition.of.resources

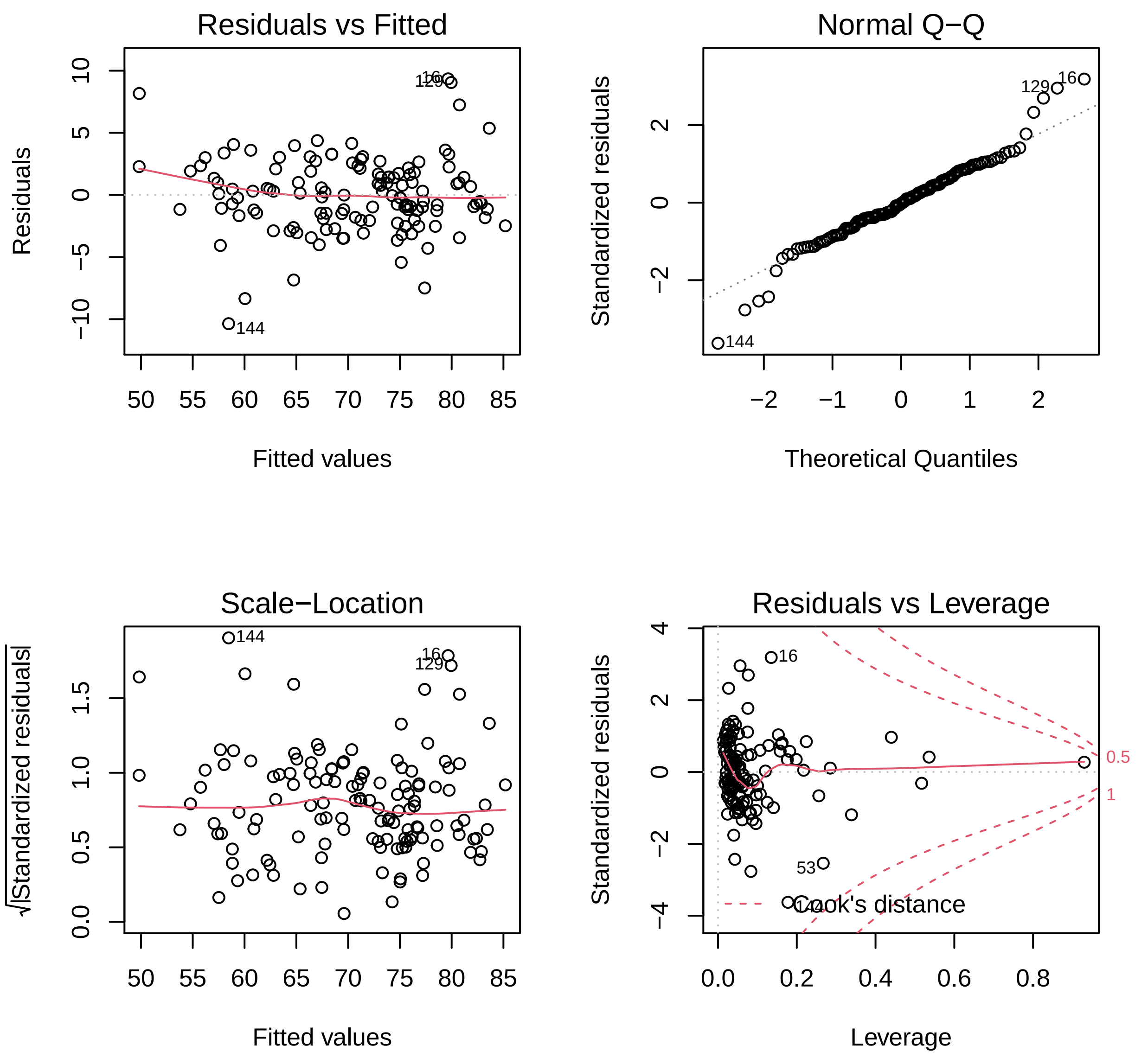
Final Model

Since the adjusted R^2 is relatively low for the initial model, we need to improve our model's fit via variable selection. Table. 3 summarizes the variables chosen based on different criteria through stepwise variable selection. As our goal is to identify the relationship, we should have fewer variables. Thus, these variables appearing in all three criteria were chosen as part of our final model:

$$\begin{aligned} y_{life\ expectancy} = \beta_0 + \beta_1 x_{Alcohol\ Consumption} + \beta_2 x_{Polio} + \beta_3 x_{Diphetheria} \\ + \beta_4 x_{Hepatitis.B} + \beta_5 x_{GDP} + \beta_6 x_{Population} + \beta_7 x_{Adult\ Mortality} + \beta_8 x_{Total\ expenditure} \\ \beta_9 x_{HIV Rate} + \beta_{10} x_{HDI} + \epsilon \end{aligned}$$

After fitting the final model, we see that the final model outperforms the initial model since the adjusted R^2 more than doubled. Our interpretation will be based on the final model since it fits the data better.

Figure.1 Diagnostic plots for final model



According to Fig.1, all assumptions are satisfied for our final model except for a minor violation of normality. Again, we will not apply any transformation.

Table. 4 in the appendix shows only the VIF of *Diphtheria* and *Hepatitis. B* is slightly above 5 for both models, which indicates some multicollinearity. Since both predictors are essential for the report's goal, we will keep them in our model.

Table. 5 in the appendix shows the number of influential observations for each model based on different criteria. We will keep them in the dataset since we do not have a contextual reason to remove them.

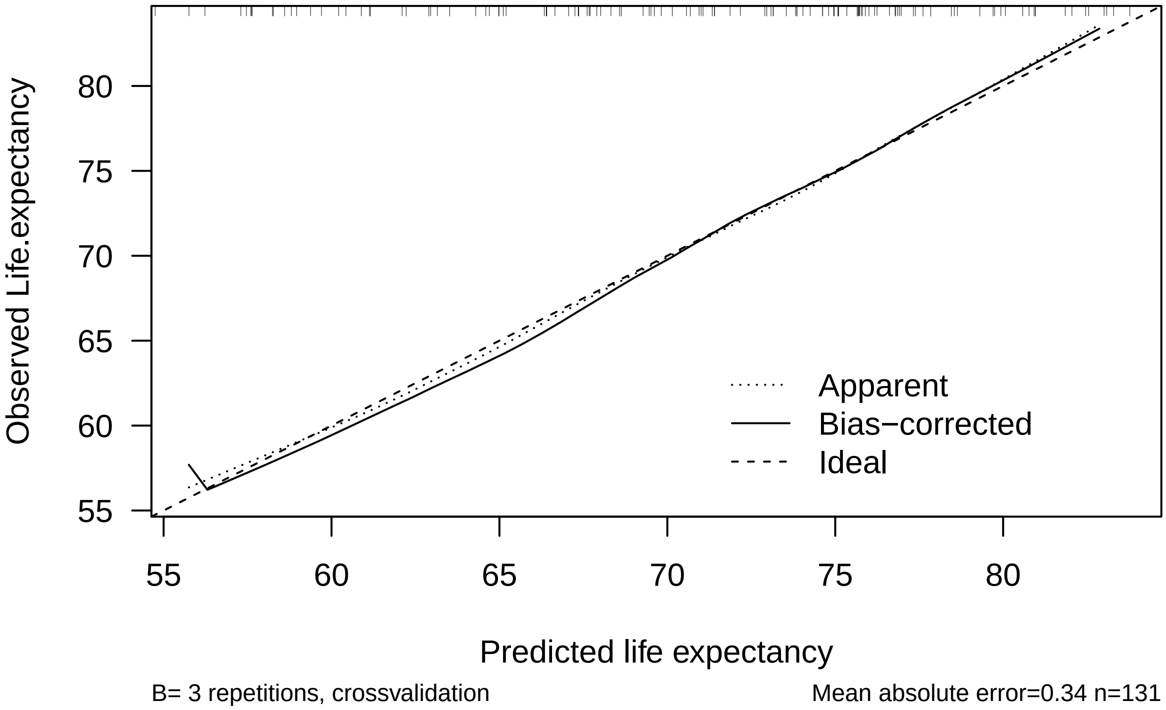


Fig. 2 Calibration plot for final model

Cross−Validation calibration

Based on Fig. 2, the bias-corrected curve is very close to the ideal curve; our final model has outstanding good prediction accuracy.

Discussion

Interpretation of final model

Based on the estimated coefficients of the final model in table 2, a one-liter increase in alcohol consumption per capita will increase life expectancy by 0.042. In terms of the Polio vaccine, a one percent increase in coverage decreases life expectancy by 0.015. Moreover, a one percent increase in Diphtheria coverage increases life expectancy by 0.015. Lastly, a one percent increase in Hepatitis. B coverage increases life expectancy by only 0.003. All these interpretations are made when holding other variables at a constant. However, none of these effects are statistically significant. Instead, life expectancy is more influenced by adult mortality, percentage expenditure on health, HIV-caused deaths per 1000, and the HDI index of a country. In other words, the model shows that the effect of alcohol consumption and immunization coverage on life expectancy is not significant. Governments should consider other health policies to improve life expectancy based on the result.

Limitations

A major limitation of this report is the model specification; a multiple linear regression may not capture the proper relationship between alcohol consumption, immunization coverage, and life expectancy. Also, we still have multicollinearity and mild violation of normality in the final model; our coefficient estimates may be biased. In the report, we did not apply any transformation as the interpretability of our model will suffer.

In terms of data, we only fit the model on the data from 2014, so multiple linear regression is appropriate. The estimated effects may differ if we fit a different model over the years.

References

Zhou, F., Santoli, J., Messonnier, M. L., Yusuf, H. R., Shefer, A., Chu, S. Y., Rodewald, L., & Harpaz, R. (2005). Economic Evaluation of the 7-Vaccine Routine Childhood Immunization Schedule in the United States, 2001. Archives of Pediatrics & Adolescent Medicine, 159(12), 1136–1144. https://doi.org/10.1001/archpedi.159.12.1136

Callaghan, T., Moghtaderi, A., Lueck, J. A., Hotez, P., Strych, U., Dor, A., Fowler, E. F., & Motta, M. (2021). Correlates and disparities of intention to vaccinate against COVID-19. Social Science & Medicine (1982), 272, 113638–113638. https://doi.org/10.1016/j.socscimed.2020.113638

Thompson K, Dutton DJ, MacNabb K, Liu T, Blades S, Asbridge M. Changes in alcohol consumption during the COVID-19 pandemic: exploring gender differences and the role of emotional distress. Health Promot Chronic Dis Prev Can. 2021;41(9):254-63. https://doi.org/10.24095/hpcdp.41.9.02

Danilova, I., Shkolnikov, V. M., Andreev, E., & Leon, D. A. (2020). The changing relation between alcohol and life expectancy in Russia in 1965–2017. Drug and Alcohol Review, 39(7), 790–796. https://doi.org/10.1111/dar.13034

Östergren O, Martikainen P, Tarkiainen L, et al Contribution of smoking and alcohol consumption to income differences in life expectancy: evidence using Danish, Finnish, Norwegian and Swedish register data J Epidemiol Community Health 2019;73:334-339.

Linden, M., & Ray, D. (2017). Aggregation bias-correcting approach to the health–income relationship: Life expectancy and GDP per capita in 148 countries, 1970–2010. Economic Modelling, 61, 126–136. https://doi.org/10.1016/j.econmod.2016.12.001

Appendix

Figure.3 Diagnostic plots for initial model

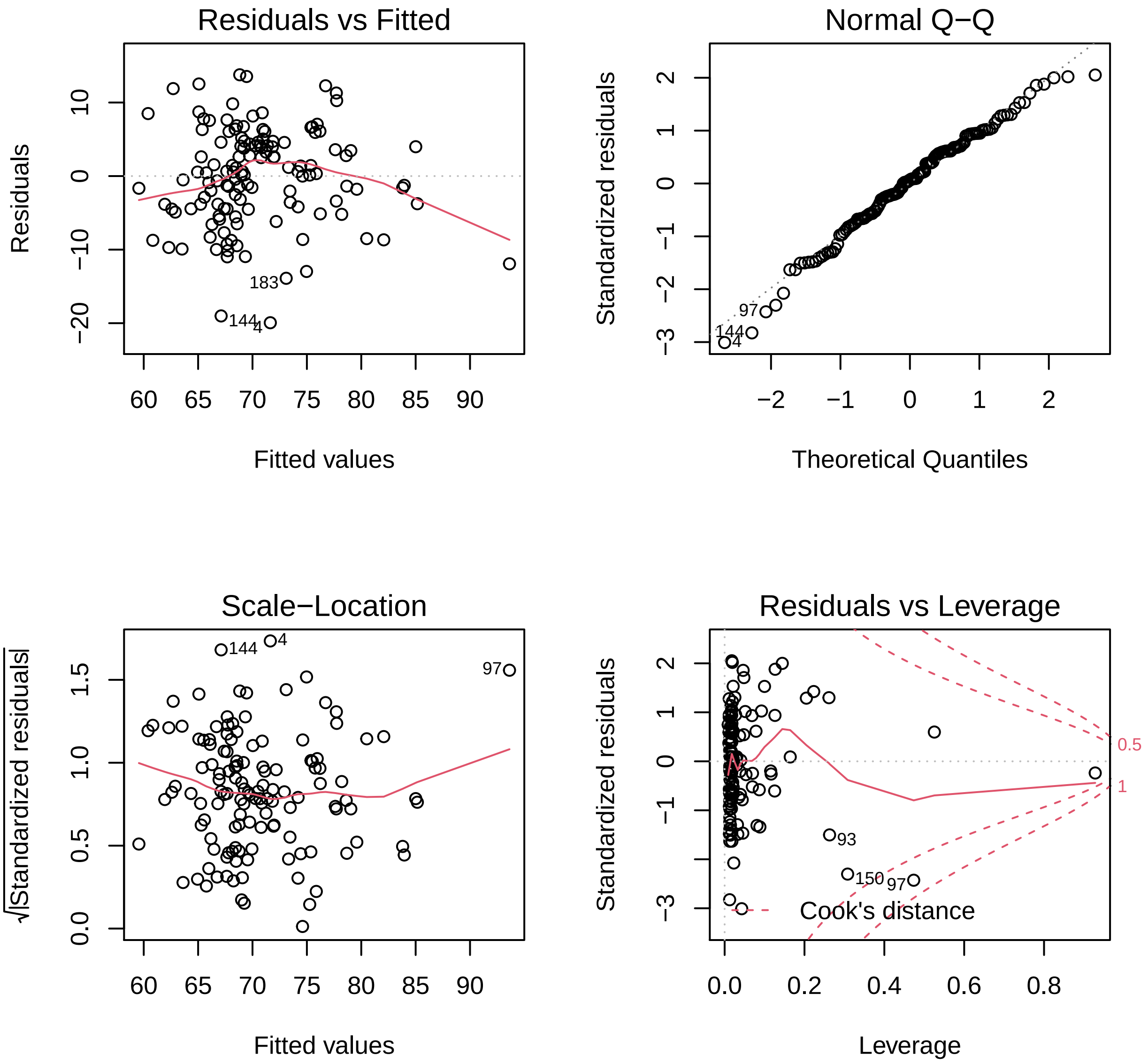


Table 4: Variance inflation factor for each variables based on two models

	VIF based on initial model	VIF based on final model
Alcohol	1.34	1.81
Polio	2.21	2.43
Diphtheria	6.49	6.90
Hepatitis.B	5.16	5.38
GDP	1.27	1.43
Population	1.01	1.03
Adult.Mortality	0.00	2.57
Total.expenditure	0.00	1.15

	VIF based on initial model	VIF based on final model
HIV.AIDS	0.00	1.83
Income.composition.of.resources	0.00	3.24

Table 5: Number of influential observations for each model based on different criteria

Criteria	Number of Influential observations for initial model	Number of Influential observations for final model
Cook's distance	0	0
DFFITS	12	9
DFBETAS	10	11