Life expectancy*

Yufei Liu

November 1, 2023

This paper

1 Introduction

ashdjk asdhkja askdhka

ahdjkajks

adkajhk

akjdha

2 Data

We used `R` to do the analysis in this paper (R Core Team 2022). We used packages `tidyverse`(Wickham et al. 2019), `janitor`(Firke 2021), and `here`(Müller 2020) to clean and load the data as well as create figures, and `knitr`(Xie 2014) and `modelsummary`(**citeModelsummary?**) to generate tables. The color style of the figures was created referring to a R colors cheet-sheet (**citeRcolor?**).

*Code and data are available at: https://github.com/Florence-Liu/life_expectancy

2.1 Data description

The datasets used in this paper was obtained from World Health Organization (WHO) Global Health Observatory data repository, and is publicly available from WHO website (**citeWHO?**). We utilized 3 datasets: life expectancy and Health life expectancy by WHO region (**citeLife?**), SDG Target 3.5 Substance abuse by WHO region (**citeAlcohol?**), and SDG Target 3.a Tobacco control by WHO region (**citeTobacco?**). The life expectancy dataset contains life expectancy and healthy life expectancy at birth and at age 60 for different sexes and WHO regions in year 2000, 2010, 2015, and 2019. The life expectancy values were estimated based on mortality data from civil registration. The substance abuse dataset contains total alcohol per capita (aged 15+) consumption (total APC) for different sexes and WHO regions in year 2000, 2005, 2010, 2015, and 2019. The tobacco control dataset contains the percentage of the population aged 15+ who currently use tobacco products based on population-based surveys for different sexes and WHO regions in year 2000, 2005, 2010, 2015, 2018, 2019, and 2020.

We specifically selected life expectancy at age 60 since it would be more associated with either tobacco or alcohol usage than life expectancy at birth. Then we merged the three datasets into one by region, sex, and year. Our cleaned data for analysis contains variables:

- region: WHO regions
- year: Year data collected
- sex: Sex at birth
- life_expectancy: The average number of years that a person of age 60 could expect to live
- alcohol_consumption: Total alcohol per capita (aged 15+) consumption in litre
- prevalence_of_tobacco: The percentage of the population aged 15+ who currently use any tobacco products.

2.2 Data Visualization

Figure 1 shows how prevalence of tobacco use relates to life expectancy at age 60 for two sexes, male and female. The scatterplot clearly demonstrates an overall decreasing trend that when the prevalence of tobacco use increases, the mean life expectancy at age 60 decreases. For the two sex groups, we could discover that there are clear clusters of the points, indicating the mean prevalence of tobacco may be different in the two groups. The two dashed lines show the linear fitted line for each group and the green line represents the linear fitted line for both group as a whole. It shows that the two variables interact and we should consider an interaction term between sex and prevalence of tobacco in our model.

Figure 2 shows how total alcohol per capita consumption relates to life expectancy at age 60 for male and female separately. The scatterplot demonstrates that when the total alcohol per capita consumption increases, the life expectancy also increases for both groups. But by how

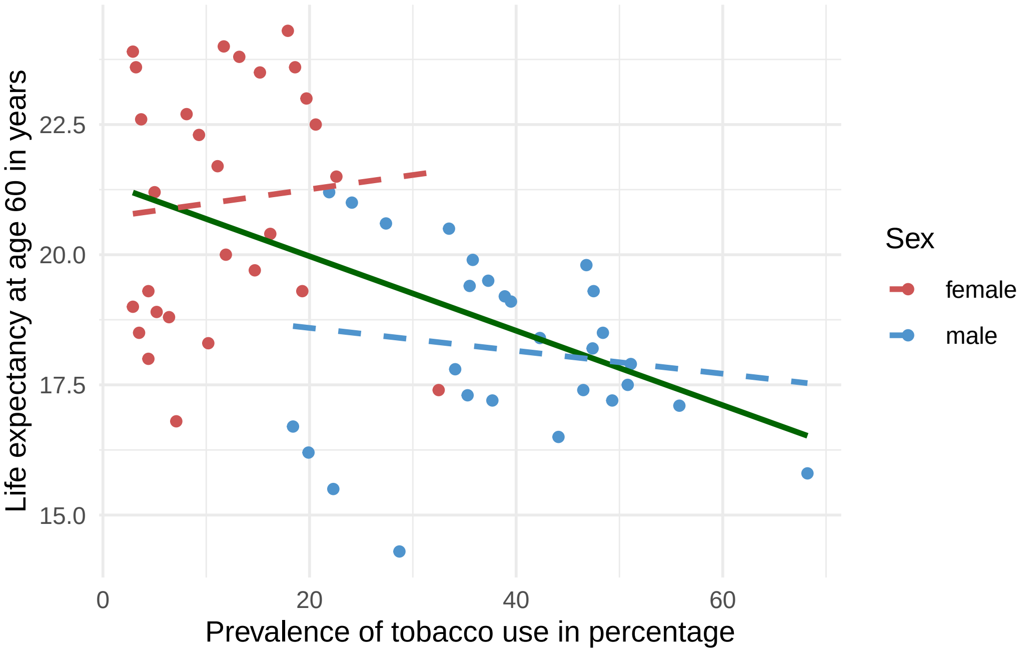


Figure 1: The effect of prevalence of tobacco on life expectancy at age 60 for different sex

much the life expectancy increases is different for male and female. It indicates that we may need interaction terms in our model for sex and alcohol consumption since the difference of sex matters how alcohol consumption affects life expectancy. The two dashed lines showing the linear fitted line for each group and the green line representing the linear fitted line for both group as a whole also agree with adding interaction terms in the model given that the direction of correlation for two sex groups and total is opposite.

Figure 3 shows life expectancy at age 60 for different years and different sexes. In general, life expectancy at age 60 increased from 2000 to 2019 for both sexes. The boxplots also shows that in each year, the mean life expectancy at age 60 for female is much higher than that for male along with larger range and variance. Also, the distribution of life expectancy for both sexes are skewed with one whisker longer than the other, indicating more observations centered at higher values. There is no outliers for all the boxplots.

Table 1: Summary of average life expectancy, prevalence of tobacco, and alcohol consumption across years

Year	Mean life expectancy (year)	Mean prevalence of tobacco (%)	Mean alcohol consumption (litre)
2000	18.3	31.2	5.2
2010	19.5	25.2	5.6

Year	Mean life expectancy (year)	Mean prevalence of tobacco (%)	Mean alcohol consumption (litre)
2015	20.1	22.9	5.6
2019	20.5	21.4	5.3

Table 1 and Table 2 shows summary tables for the mean values of life expectancy at age 60, prevalence of tobacco, and alcohol consumption across years and regions. We could see that the mean life expectancy increased from 2000 to 2019, which is consistent with Figure 3, while the mean prevalence of tobacco decreased across years and the mean alcohol consumption is relatively stable with minor changes. The potential inverse relationship between prevalence of tobacco and life expectancy is consistent with what we found in Figure 1. For different WHO regions, the three values differ a lot. Africa has the lowest mean life expectancy at age 60 while Americas has the highest mean life expectancy at age 60. However, Africa also has the lowest prevalence of tobacco, which contradicts our previous findings. Also, the mean alcohol consumption has a large range with the lowest value at 0.3 for Eastern Mediterranean and highest value at 10.2 for Europe. These findings in region difference could be further investigated since there might be potential cultural and socio-economic reasons behind.

Table 2: Summary of average life expectancy, prevalence of tobacco, and alcohol consumption for different WHO regions

Region	Mean life expectancy (year)	Mean prevalence of tobacco (%)	Mean alcohol consumption (litre)
Africa	16.9	13.4	4.9
Americas	22.0	21.2	7.8
Eastern Mediterranean	18.0	22.2	0.3
Europe	21.2	29.2	10.2
Global	20.2	26.6	5.5
South-East Asia	18.2	37.6	3.2
Western Pacific	20.8	26.0	5.9

3 Model

Based on the data visualization, there are clear trends between prevalence of tobacco, alcohol consumption, sex, year, region, and life expectancy. So we will fit multiple linear regression models to discover how these factors contribute to life expectancy at age 60.

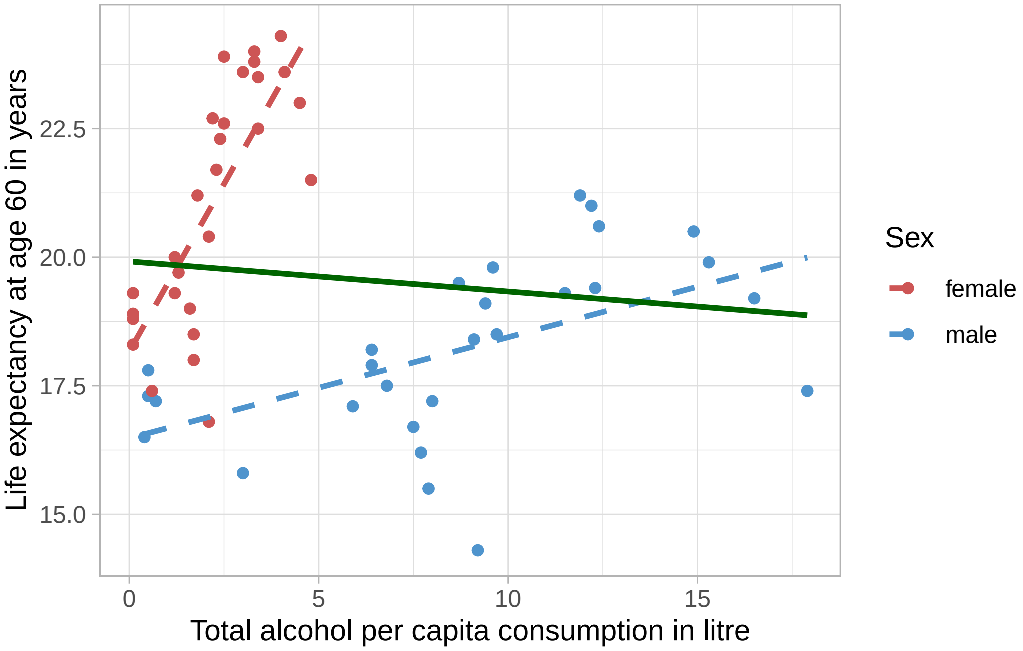


Figure 2: The effect of total alcohol per capita consumption on life expectanct at age 60 for different sex

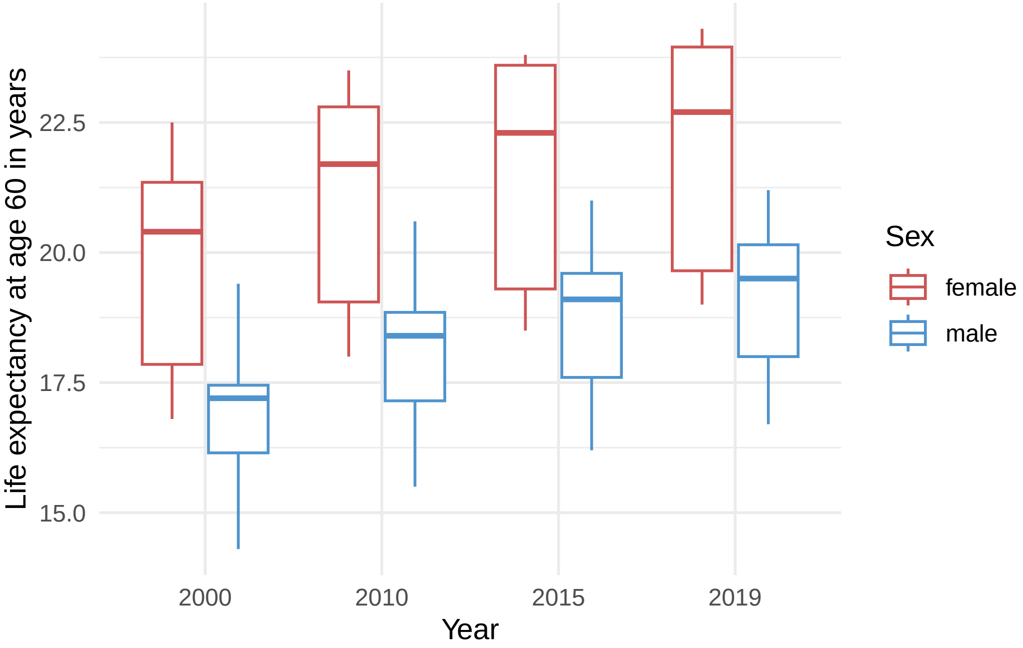


Figure 3: Life expectancy at age 60 across years for different sex

To justify our choice of including interaction terms based on observation, we will fit two models, model 1 contains interaction terms and model 2 does not include interaction terms. The full model is shown as

$$\begin{aligned} Y = \beta_0 + \beta_1 X_{tobacco} + \beta_2 D_{male} + \beta_3 X_{alcohol} + \beta_4 D_{2010} + \beta_5 D_{2015} + \beta_6 D_{2019} + \\ \beta_7 D_{Americas} + \beta_8 D_{EM} + \beta_9 D_{Europe} + \beta_{10} D_{Global} + \beta_{11} D_{SeA} + \beta_{12} D_{WP} + \\ \beta_{13} X_{tobacco} D_{male} + \beta_{14} X_{alcohol} D_{male} + \epsilon \end{aligned}$$

where

- Y is the dependent varianle life expectancy at age 60
- β_0 represents the intercept of the model, which is the expected life expectancy when all other variables are zero
- β_1 represents the change in expected life expectancy for a one-unit change in the variable prevalence of tobacco $X_{tobacco}$ when other variables are held constant
- β_2 represents the difference in expected life expectancy between male and female when other variables are held constant
- β_3 represents the change in expected life expectancy for a one-unit change in the variable alcohol consumption $X_{alcohol}$ when other variables are held constant
- β_4, β_5, and β_6 represent the difference in expected life expectancy between each year and year 2000
- β_7, β_8, β_9, β_{10}, β_{11}, and β_{12} represent the difference in expected life expectancy between each region and Africa
- β_{13} represents the difference in the change of expected life expectancy for a one-unit change in $X_{tobacco}$ for male and female
- β_{14} represents the difference in the change of expected life expectancy for a one-unit change in $X_{alcohol}$ for male and female
- $X_{tobacco}$ is the variable prevalence of tobacco use
- D_{male} is a dummy variable with 1 for male and 0 for female
- $X_{alcohol}$ is the variable total alcohol per capita consumption
- D_{2010}, D_{2015}, and D_{2019} are dummy variables with 1 for the specific year and 0 if it is not the specific year
- $D_{America}$, D_{EM}, D_{Europe}, D_{Global}, D_{SeA}, and D_{WP} are dummy variables with 1 for the specific region and 0 if it is not the specific region. EM represents Eastern Mediterranean region, SeA represents South-east Asia region, and WP represents Western Pacific region.
- ϵ is the random error

For the reduced model, we just removed the interaction terms, that is the predictors with coefficient β_{13} and $\beta 14$ in the full model.

The linear model will generate the best estimates for parameters β_i and D_j that minimize the residual sum of squares (RSS). After getting the best fit of the model, we need to implement model validation to make sure the assumptions for the model hold, that is linearity, homoscedasticity of errors, independence of errors, and influential observations. The model validation is done in the Section 7. It seems that assumptions for both models holds.

4 Result

Table 3 shows a summary for both models with the listed values representing the estimates of parameters and values in brackets representing the standard error for the estimate of the parameter. The two models produced different but close estimates of parameters. It is noticeable that the direction of correlation between alcohol assumption and life expectancy is different in two models, positive correlation in the full model and negative in the reduced model. This is consistent with data visualization in Figure 2 that fitted dash lines within two sex groups show an increasing trend but green fitted line for the whole data shows a decreasing trend.

For the full model, according to the p-values for each parameters, the variable `sex` is not significant at significance level $\alpha = 0.05$ since it has a p-value of 0.15, indicating the variable `sex` does not have a significant effect on the expected life expectancy at age 60. However, since the interaction term between age and alcohol consumption is statistically significant with p-value smaller than the significance level $\alpha = 0.05$, we should still include the separate parameter `sex`.

For the reduced model, according to the p-values for each parameters, the variable `sex` is also insignificant with p-value equal 0.3, larger than that in the full model. This may be due to the interaction effect. Also, it is noticeable that the dummy variable for Eastern Mediterranean region is also statistically insignificant with p-value equal to 0.08 at significance level $\alpha = 0.05$, however, if we set the significant level at $\alpha = 0.1$, then the predictor becomes significant. It all depends on our choice of significance level.

To justify our choice of interaction terms and choose a better model, we will compare AIC, BIC, RMSE, and adjusted R^2 values for both models. We found that the rounded AIC for the full model is 45 with 64 for the reduced model, the rounded BIC for the full model is 77 with 92 for the reduced model, the RMSE for the full model is 0.27 with 0.33 for the reduced model, and the adjusted R^2 for the full model is 0.98 with 0.97 for the reduced model. Since we want AIC, BIC, and RMSE to be as small as possible while adjusted R^2 to be as large as possible, the full model with interaction terms could be a better fit for the data. It has higher predictive performance without loss of general interpretability. Generally, the prevalence of tobacco use has a negative linear relationship with the life expectancy at age 60, the alcohol per capita consumption has a positive linear relationship with the life expectancy at age 60 while sex itself does not have a significant effect on the life expectancy at age 60.

Table 3: Summary of two linear regression models

	With interaction	Without interaction
(Intercept)	16.569	17.523
	(0.351)	(0.189)
prevalence_of_tobacco	−0.059	−0.035
	(0.013)	(0.010)
sexmale	−0.536	−0.436
	(0.368)	(0.406)
alcohol_consumption	0.432	−0.225
	(0.164)	(0.031)
as.factor(year)2010	0.956	1.104
	(0.133)	(0.154)
as.factor(year)2015	1.404	1.591
	(0.142)	(0.163)
as.factor(year)2019	1.761	1.879
	(0.147)	(0.172)
regionAmericas	5.349	6.037
	(0.288)	(0.235)
regionEastern Mediterranean	1.542	0.408
	(0.349)	(0.233)
regionEurope	4.798	6.047
	(0.455)	(0.328)
regionGlobal	3.761	3.884
	(0.201)	(0.234)
regionSouth-East Asia	2.386	1.768
	(0.278)	(0.290)
regionWestern Pacific	4.268	4.590
	(0.209)	(0.235)
prevalence_of_tobacco × sexmale	0.024	
	(0.012)	
sexmale × alcohol_consumption	−0.509	
	(0.123)	
Num.Obs.	56	56
R2	0.988	0.982
R2 Adj.	0.984	0.977
AIC	45.0	63.9
BIC	77.4	92.2
Log.Lik.	−6.510	−17.937
F	238.350	192.716
RMSE	0.27	0.33

5 Discussion

5.1 Brief summary

In this paper, we have conducted a analysis of life expectancy at age 60 with several important predictors including prevalence of tobacco use, alcohol per capita consumption, sex at birth, year in 2000, 2010, 2015, 2019, and WHO regions. We have found potential linear relationships between each variable and life expectancy through data visualization. To quantitatively understand how these factors affect life expectancy at age 60, we have utilized multiple linear regression models. We have fitted two model, a full model with interaction effect between sex and prevalence of tobacco as well as the interaction effect between sex and alcohol consumption, and a reduced model with five independent variables but no interaction terms. The results show that the full model have a better fit with smaller root mean squared errors (RMSE), AIC, and BIC, and a larger adjusted R^2.

5.2 Findings

According to Table 3 and Figure 1, when other variables are held constant, the increase of the prevalence of tobacco use will decrease the life expectancy at age 60. This is consistent with the fact that tobacco is a health risk factor for cardiovascular and respiratory diseases and thus people with a habit of use tobacco products not only cigarettes will have a lower life expectancy at age 60(**citeTobaccouse?**). Combined with Table 1 and Table 2, we could find that although the prevalence of tobacco in general decreases across years, it has an uneven region distribution and notably for Africa, it has the lowest prevalence of tobacco but relatively low life expectancy.

Also for the alcohol per capita consumption, the model suggested that the alcohol consumption is associated with increased life expectancy at age 60. Other than tobacco products, the effect of alcohol on health reveals a complex relationship. Modest drinking has been approved to be protective against some diseases, but in most cases, alcohol could be seen as a health risk factor (**citeDrinking?**). In our model, we did not specify alcohol consumption level, so it could be further investigated by creating a new dummy variable indicating different drinking levels.

According to Table 1 and Table 2, the values that are not consistent with our model result indicate that there might be other factors that influence the life expectancy at age 60 for a specific year or region. For different regions, the difference in economic development, medical treatments, and environmental conditions could be potential influencing factors. However, if we introduced these factors into our model, we have to consider the problem of multicollinearity that the independent variables correlates since these socio-economic factors and environmental factors could also contribute to the prevalence of tobacco and alcohol consumption.

5.3 Weakness and future work

One weakness of this paper is about data quality. For the life expectancy dataset, the limitation comes from lack of complete and reliable mortality data for some civil registrations. This may cause the estimation of life expectancy biased since additional model estimations were used instead of real data. For the prevalence of tobacco dataset, one limitation is that some countries or regions do not have reliable data and estimation using Bayesian models held many assumption that may not actually satisfy in real cases. Another limitation is the data source. Since the estimation was made by population-based surveys, self-report biases exist that people may tend to hide some smoking habits if they think it is “not good”. For alcohol consumption dataset, same problem for missing several data points for some regions or in a specific year. Additionally, the unrecorded consumption calculation and tourists consumption were calculated from model estimations and several assumption were made during the process that may not be actually valid.

Another weakness is about the model. Since we decide to employ linear regression models based on data visualization, we assume linear relationships between variables, which may ignore some non-lineaer relationship.

For further studies, we could increase our sample size and split into training and testing data to better assess the model performance. We could also use cross validation to find a more robust estimate. Also, additional factors could be considered as well as more complex models that could capture potential non-linear relationships

6 Reference

Firke, Sam. 2021. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://CRAN.R-project.org/package=janitor.

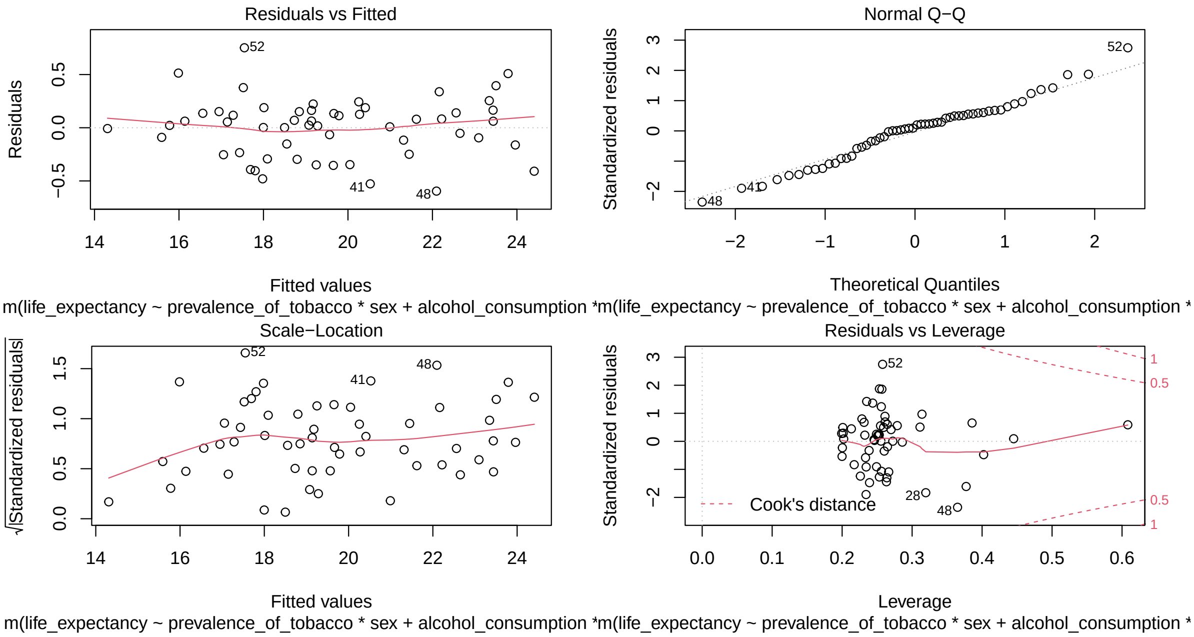
Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

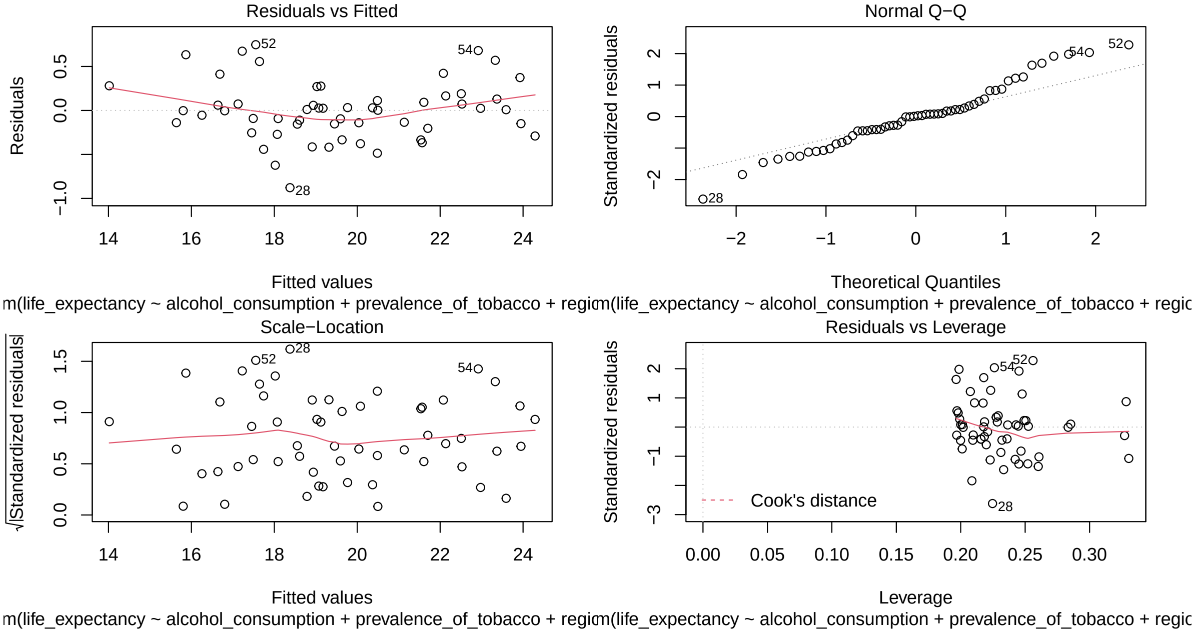
R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC. http://www.crcpress.com/product/isbn/9781466561595.

7 Appendix


Residuals vs Fitted
Residuals
Fitted values
m(life_expectancy ~ prevalence_of_tobacco * sex + alcohol_consumption *
52
41
48
−0.5
0.0
0.5
14
16
18
20
22
24
Normal Q−Q
Standardized residuals
Theoretical Quantiles
m(life_expectancy ~ prevalence_of_tobacco * sex + alcohol_consumption *
52
41
48
−2
0
1
2
3
−2
−1
0
1
2
Scale−Location
√|Standardized residuals|
Fitted values
m(life_expectancy ~ prevalence_of_tobacco * sex + alcohol_consumption *
52
41
48
0.0
0.5
1.0
1.5
14
16
18
20
22
24
Residuals vs Leverage
Standardized residuals
Leverage
m(life_expectancy ~ prevalence_of_tobacco * sex + alcohol_consumption *
Cook's distance
52
28
48
1
0.5
0.5
1
−2
0
1
2
3
0.0
0.1
0.2
0.3
0.4
0.5
0.6


Residuals vs Fitted
Residuals
Fitted values
m(life_expectancy ~ alcohol_consumption + prevalence_of_tobacco + regio
52
54
28
−1.0
0.0
0.5
14
16
18
20
22
24
Normal Q−Q
Standardized residuals
Theoretical Quantiles
m(life_expectancy ~ alcohol_consumption + prevalence_of_tobacco + regic
52
54
28
−2
0
1
2
−2
−1
0
1
2
Scale−Location
√|Standardized residuals|
Fitted values
m(life_expectancy ~ alcohol_consumption + prevalence_of_tobacco + regio
28
52
54
0.0
0.5
1.0
1.5
14
16
18
20
22
24
Residuals vs Leverage
Standardized residuals
Leverage
m(life_expectancy ~ alcohol_consumption + prevalence_of_tobacco + regic
Cook's distance
52
54
28
−3
−1
1
2
0.00
0.05
0.10
0.15
0.20
0.25
0.30