

LSTAT2120 - Project 2022

Linear models of life expectancy

Rousseau Mathieu, 67001800

Noiset Sorenza, ???



UCLouvain
Belgium
23/12/2022

1 Introduction

We are working with a dataset containing different *health factors* collected from **WHO** (World Health Organization) as well as *economic factors* collected from **ONU** for almost every countries in the world and years between 2000 and 2015. Some countries are not present in this dataset because they had too much missing data.

Since we are dealing with panel data, we chose to only work with data from the year 2012. We removed observations with missing values and we modified the ‘adult mortality’ continuous variable into a qualitative variable. So the final dataset has ??? **observations** and 20 **variables**. For our analysis, we separated the dataset into a training set and a testing set containing respectively 80% and 20% of the observations. This separation is random (*i.e. the dataset is shuffled before separating it*).

2 Research question

We want to understand how different factors affect positively or negatively the life expectancy. We would like to be able to predict the mean life expectancy (response variable : *life.expectancy*) for a given country based on different health, economic and social factors.

Firstly, we will start by doing a descriptive analysis of the different variables. Then we will try different linear models and select the best one based on different relevant criterions. We will check if the classical hypothesis are respected as well as nonlinearity, influential observations. If some hypotheses are not respected, we will fix that. We will finish by making prediction on a testing set with our model.

3 Descriptive statistic

We have 20 variables in our dataset of which 2 are qualitative. The *status* indicates if the country is developed or developing and the *adult.mortality* feature categorize the probability of dying between 15 and 60 years old into five levels : very low, low, middle, high, very high.

More generally, we can classify the different variables into several categories : *economic* (country status, expenditure on health, gdp, hdi), *social* (total population of each country, number of years of schooling), *mortality* (adult mortality, infant death, under five death, under four death because of HIV/AIDS, thinness) and *immunization* factors (immunization of hepatitis b, polio, diphtheria as well as number of reported cases of measles). We will only describe some variables, the curious reader can find a complete description of these in the appendix.

The *hepatitis.b*, *polio* and *diphtheria* variables are respectively the immunization coverages against hepatitis B, polio and DPT3 (diphtheria tetanus toxoid and pertussis) among the 1 year olds and are given in percentage.

The *alcohol* variable is the consumption of alcohol per capita (of 15 years old or more) in litres of pure alcohol.

3.1 Qualitative variables

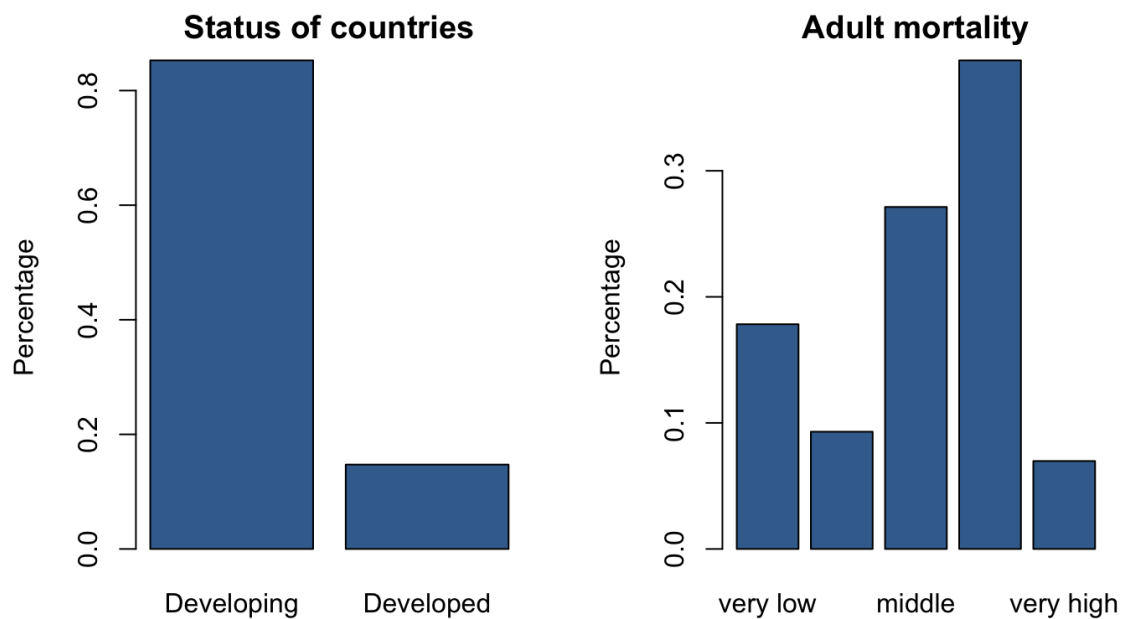


FIGURE 1 – barplot of the qualitative variables

3.2 Quantitative variables

Let's take a look to the table of the 4 moments (*mean*, *standard deviation*, *skewness* and *kurtosis*) for each of the quantitative variables.

4 moments of quantitative variables				
variable	mean	std_dev	skewness	kurtosis
life.expectancy	70.24	8.51	-0.37	2.41
infant.deaths	31.28	112.66	7.52	66.92
alcohol	3.87	4.24	0.76	2.43
percentage.expenditure	900.80	1931.89	3.92	19.17
hepatitis.b	80.57	26.54	-1.94	5.66
measles	967.93	2794.70	4.14	21.96
bmi	38.03	20.93	-0.12	1.63
polio	82.64	25.40	-2.17	6.60
total.expenditure	6.18	2.48	0.21	2.72
diphtheria	84.53	22.58	-2.51	8.70
hiv.aids	1.08	2.15	2.83	11.13
gdp	7460.10	12333.66	2.90	11.53
population	13580015.34	35183192.47	4.32	23.81
thinness.10.19.years	4.76	4.57	1.92	7.77
thinness.5.9.years	4.74	4.46	1.96	8.48
income.composition.of.resources	0.66	0.15	-0.25	2.04
schooling	12.54	2.73	-0.07	2.89

FIGURE 2 – table of moments (mean, standard deviation, skewness, kurtosis) for the quantitative variables

The **life expectancy** has a mean of roughly 70 years with a standard deviation of 8.6. It is slightly negatively skewed which indicates that some countries have low life expectancy. The kurtosis is less than 3 so the distribution is a little bit flattened.

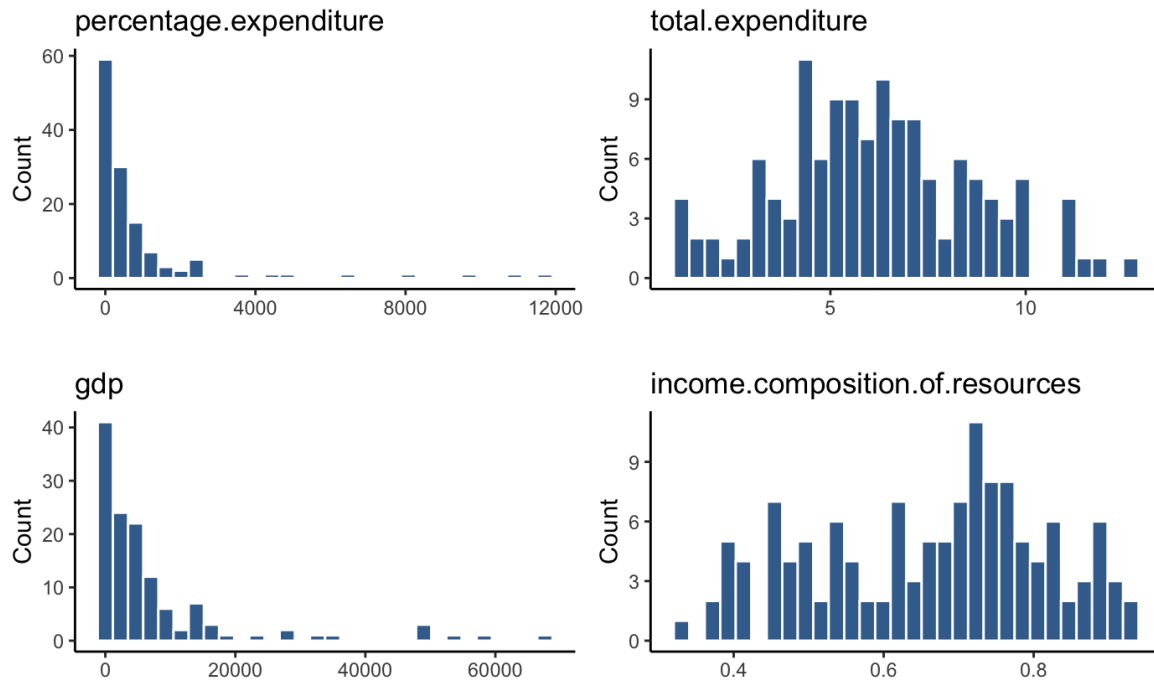


FIGURE 3 – Histogram of economic features

The **infant death** has a mean of 31/1000 but has a huge standard deviation (112.6). We see that heavily tailed with a kurtosis of 66.9. The same kind of conclusion can be made for the **deaths under five year old**.

In average, countries spend 9 times the GDP¹ per capita on health services but the standard deviation is huge (1931.8) which indicates the presence of outliers far away from the mean. Therefore, we expect that some countries spend much less than that on health services.

Furthermore, in average, countries spend 6% of their total budget on health services. The distribution of this variable seems to follow a normal distribution given the skewness and kurtosis.

3.3 Correlation matrix

We look at the correlation matrix in order to see if there are any highly correlated variables. Indeed, it could be a sign of multicollinearity

1. Gross Domestic Product

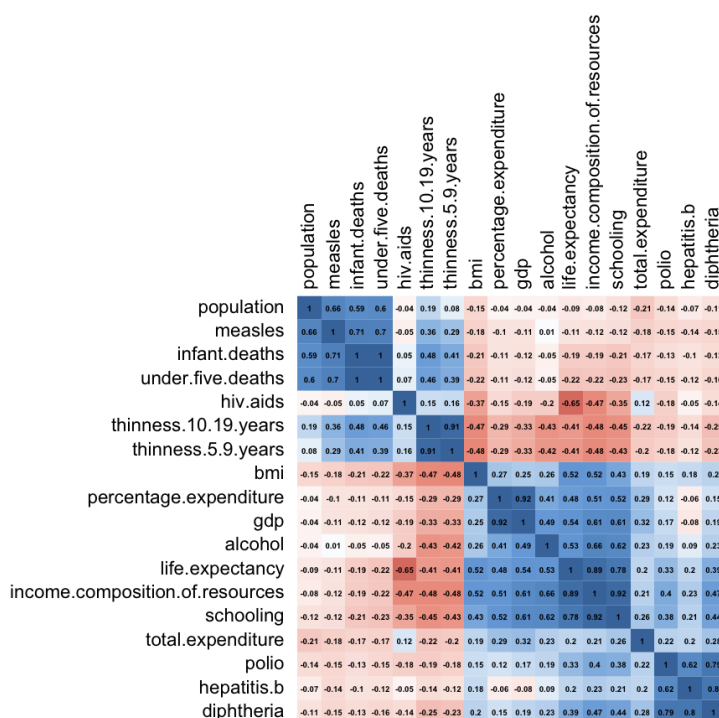


FIGURE 4 – correlation matrix of dataset

We see that several variables are highly correlated :

- The **infant death** is perfectly correlated with the **death under five year old** : these two features are redundant. As a consequence, we're gonna remove *under.five.death* feature from our dataset.
- The **thinness from 5 to 9 years old** with the **thinness from 10 to 19 years old** : we could suspect that extreme thinness comes from a problem of access to food which implies that thinness doesn't stop at 10 years old but continue throughout the teenage.
- The **number of measles cases per 1000 inhabitants** with the **infant death** : indeed, measles hits essentially children and youth and can (often) lead to death.
- The **percentage of expenditure made on health services** with the **GDP per capita**.
- The **hepatitis B** and **polio** with **diphtheria**.
- The **life expectancy** with the **HDI in terms of income composition of resources** and **schooling**.

4 Model selection

4.1 Multicollinearity on full model

Before doing any model selection, we first check if we have any multicollinearity problem. We saw in the last part that some variables were highly correlated. These high pairwise correlations could lead to multicollinearity problem but it is not always the case.

Let the full model be,

$$Y = X\beta + \varepsilon \quad (1)$$

To verify that, we can use the **Variance inflation factor (VIF)** that is defined by for the coefficient $\hat{\beta}_k$,

$$\text{VIF}_k = \frac{1}{1 - R_k^2}, \quad k = 1, \dots, p-1 \quad (2)$$

where R_k^2 is the coefficient of determination of a regression of X_k on $X_1, \dots, X_{k-1}, X_{k+1}, \dots, X_{p-1}$. Multicollinearity leads to an "ill-conditioned" matrix X . As a consequence, this matrix is numerically instable and becomes difficult to invert. Therefore, the OLS estimator $\hat{\beta}$ cannot be "correctly" computed. We have multicollinearity problem if the **VIF** is greater than 10 and if the **average VIF** is much greater than 1. We can also check the tolerance which is $1 - R_k^2$.

In the appendix (??), you can find a table with the VIF and tolerance for each variable of the full model. We notice a multicollinearity problem for the variables *gpd* (with VIF = 13.10) and *income.composition.of.resources* (with VIF = 12.72). These two have tolerance of 0.08. Furthermore, the average VIF is 4.67.

In order to fix this multicollinearity problem, we can perform a **Ridge regression** that is similar to the OLS in the sense it minimizes the SSR but it also adds a constraint on the L_2 norm (euclidean) for β . The constraint is,

$$\|\beta\|_2^2 \leq t \quad (3)$$

where $t \in \mathbb{R}$ is a parameter to be determined.

To improve our model accuracy, we need to carefully select variables without adding too much to avoid overfitting. To do that, we will perform model selection using *AIC*, *BIC* and R_a^2 criterions. Since we have a lots a variables in our full model, we will only consider model selection of type II and III.

4.2 Type II : Backward stepwise selection

The goal of a backward stepwise selection is to start from the full model then gradually removing variables one at a time. At each iteration, we remove the one that yields the lowest accuracy in prediction when added the pool of selected variables. We can measure the accuracy using different criterions, for this project, we will focus on *AIC* and *BIC* criterions.

4.2.1 Akaike Information Criterion (AIC)

We want to minimize the AIC that is,

$$\text{AIC} = -2 \ln L(\hat{\beta}) + 2k \quad (4)$$

where $L(\hat{\beta})$ is the maximum of the likelihood function and p is the number of estimated parameters in the model.

4.2.2 Bayes Information Criterion (BIC)

The problem with AIC is that it tends to select too much variables and therefore overfit the model. We can then use another criterions that penalizes more severely complex models and tends to favorize simpler models. This criterion we want to minize is the BIC and is given by,

$$\text{BIC} = -2 \ln L(\hat{\beta}) + (2k \cdot \ln(n)) \quad (5)$$

where n is the number of observations in our training dataset.

4.3 Type III : LASSO

The LASSO estimator is similar to the OLS (it minimizes the SSR) but it adds a constraint on the L_1 norm (Manhattan) for β . The constraints is,

$$\sum_{j=1}^p |\beta_j| \leq t \quad (6)$$

where $t \in \mathbb{R}$ is a parameter to be determined.

This change allows some coefficients to be shrunk exactly to zero.

4.4 Interactions between variables

We should then try to add interactions between variables to our chosen model. We decide to test the following interactions and compare the resulting models with our chosen one,

- *infant.deaths* with *measles*
- *adult.mortality.high* with *alcohol*
- *adult.mortality.very_high* with *alcohol*
- *total.expenditure* with *adult.mortality.low*
- *total.expenditure* with *infant.deaths*

4.5 Verifying underlying hypotheses

After selecting a model, we want to check the underlying hypothesis of the linear model. We will verify the 3 main hypothesis : homoskedasticity, independance of observations and normality of the residuals. we will also check for outliers, autocorrelation and nonlinearity.

4.5.1 Nonlinearity

We can check for nonlinearity by looking at the scatterplot of the residuals (e_i) versus the explanatory variables.

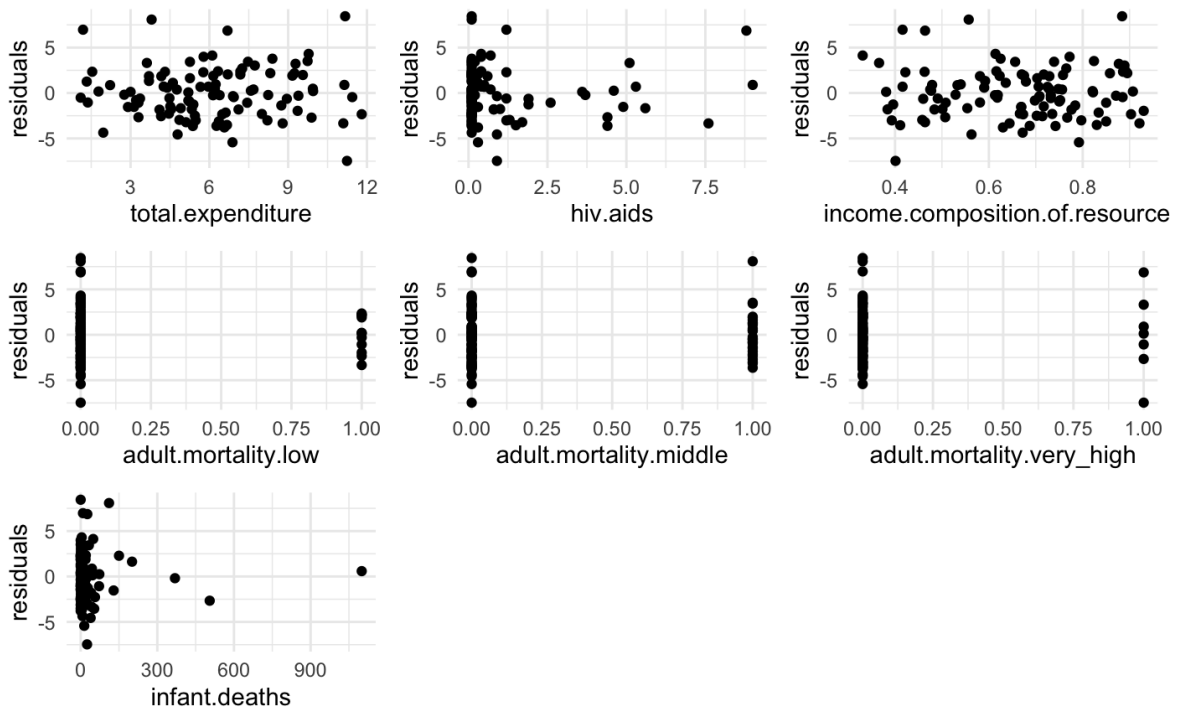


FIGURE 5 – scatterplots of residuals versus explanatory variables for the selected model

We do not see clear nonlinear patterns in the different plots above. Every variable has more or less a linear relation with its residuals. The continuous variables *hiv.aids* and *infant.deaths* have a strange pattern but the nonlinear seems too complicated to infer and would add a lot of complexity to our model. Therefore, we do not take remedial actions.

4.5.2 Outliers and influential observations

a) outliers with respect to the explanatory variables

We first try to identify outliers with respect to the explanatory variables X_{ij} .

We need to study the leverages that are the diagonal elements of the hat matrix $H = X(X^T X)^{-1} X^T$. We can say that X_i is an outlier if,

$$h_{ii} > \frac{2p}{n} \approx 0.155$$

We find 12 outliers that are the following rows of our dataset : 9, 11, 12, 23, 35, 40, 42, 49, 58, 65, 82, 97.

We want then know if the outliers found are influentials for the their fitted values. We use the DFFITS criterion for the i -th observation,

$$\text{DFFITS}_i = d_i^* \sqrt{\frac{h_{ii}}{1 - h_{ii}}} \quad (7)$$

where d_i^* are the standardized deleted residuals,

$$d_i^* = e_i \sqrt{\frac{n - p - 1}{\text{SSE}(1 - h_{ii}) - e_i^2}} \quad (8)$$

We have a criterion for DFFITS. For $n > 30$, the i -th observation is influential for its fitted value if,

$$|\text{DFFITS}_i| > 2\sqrt{\frac{p}{n}} \approx 0.557 \quad (9)$$

We find that 8 observations are influential for the fitted values : 2, 9, 35, 42, 49, 58, 95, 96. Therefore, the observations 9, 35, 42, 49 are outliers and influential for the fitted values.

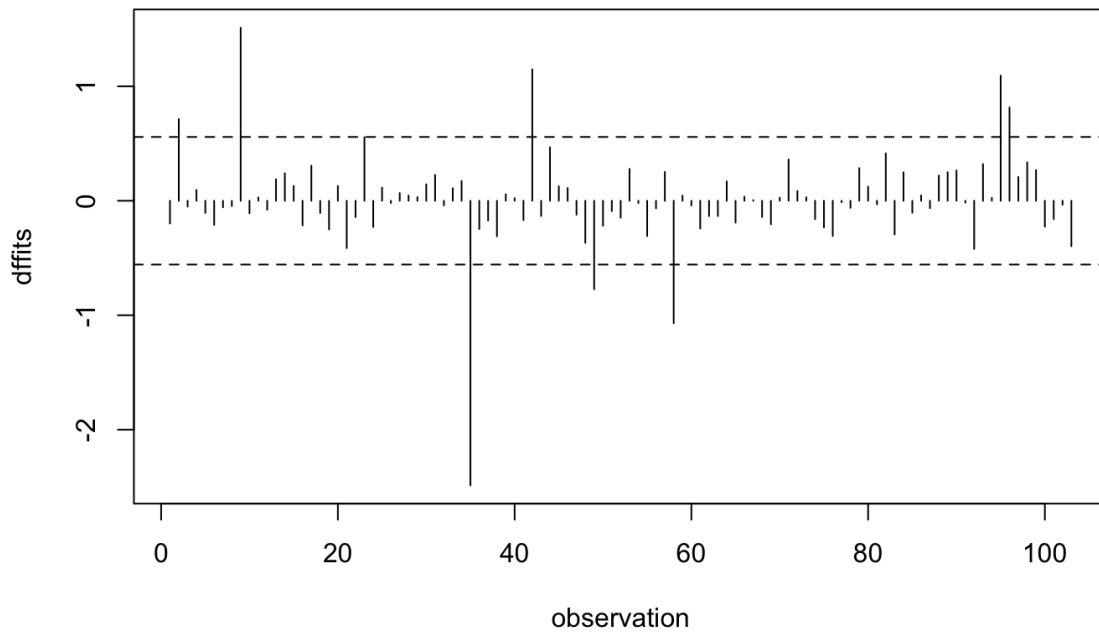


FIGURE 6 – *outside the two lines are outliers influential for their fitted values*

Now, are they also influential for the regression coefficients? Let's have a look at the Cook's distance.

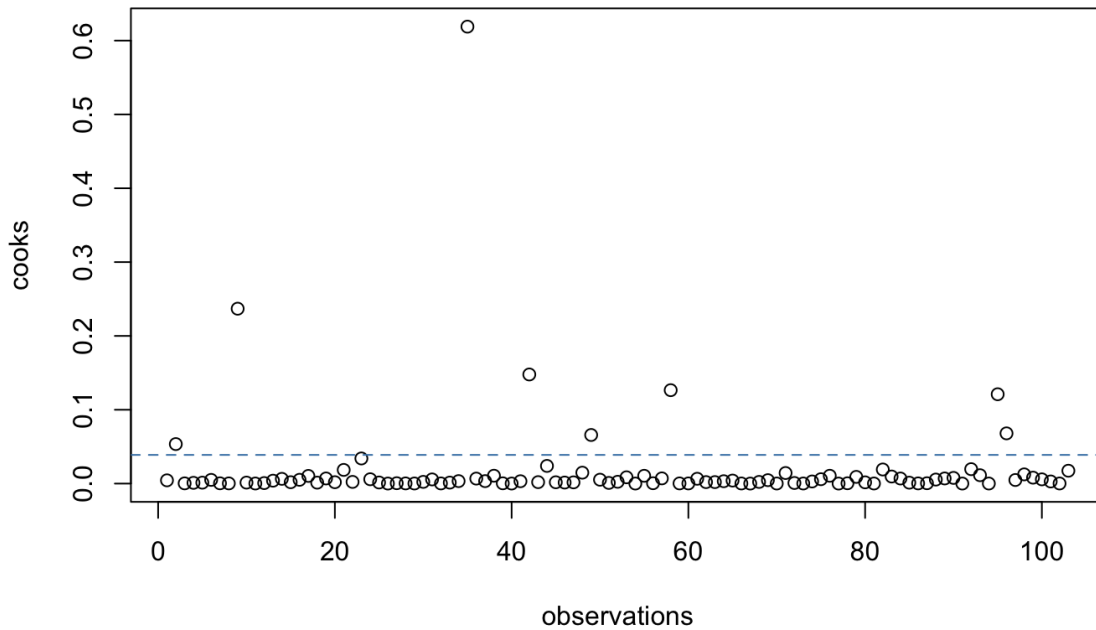


FIGURE 7 – *Cook's distance*

We find that the same 8 observations influential for the fitted values are also influential for the

The i -th observation is influential if $D_i > F_{p,n-p;1-\alpha}$ where D_i is the Cook's distance

b) outliers with respect to the response variable

We then identify outliers with respect to the response variable Y_i . Y_i is an outliers if $d_i^* > t_{n-p-1;1-\frac{\alpha}{2}}$. For a level of significance of $\alpha = 0.05$, we have $t_{94;0.975}$ and we find that the observations 2, 9, 95 and 96 are outliers for the response variable.

4.5.3 Multicollinearity

Variation inflation factor (VIF)		
final model		
Variables	Tolerance	VIF
total.expenditure	0.76	1.32
hiv.aids	0.55	1.82
income.composition.of.resources	0.49	2.04
adult.mortality.low	0.64	1.55
adult.mortality.middle	0.74	1.36
adult.mortality.very_high	0.59	1.70
infant.deaths	0.02	40.50
total.expenditure:infant.deaths	0.03	39.78

FIGURE 8 – VIF for the selected model

4.5.4 Heteroskedasticity

We can check for heteroskedasticity by looking at the plot of the residuals versus the fitted values.

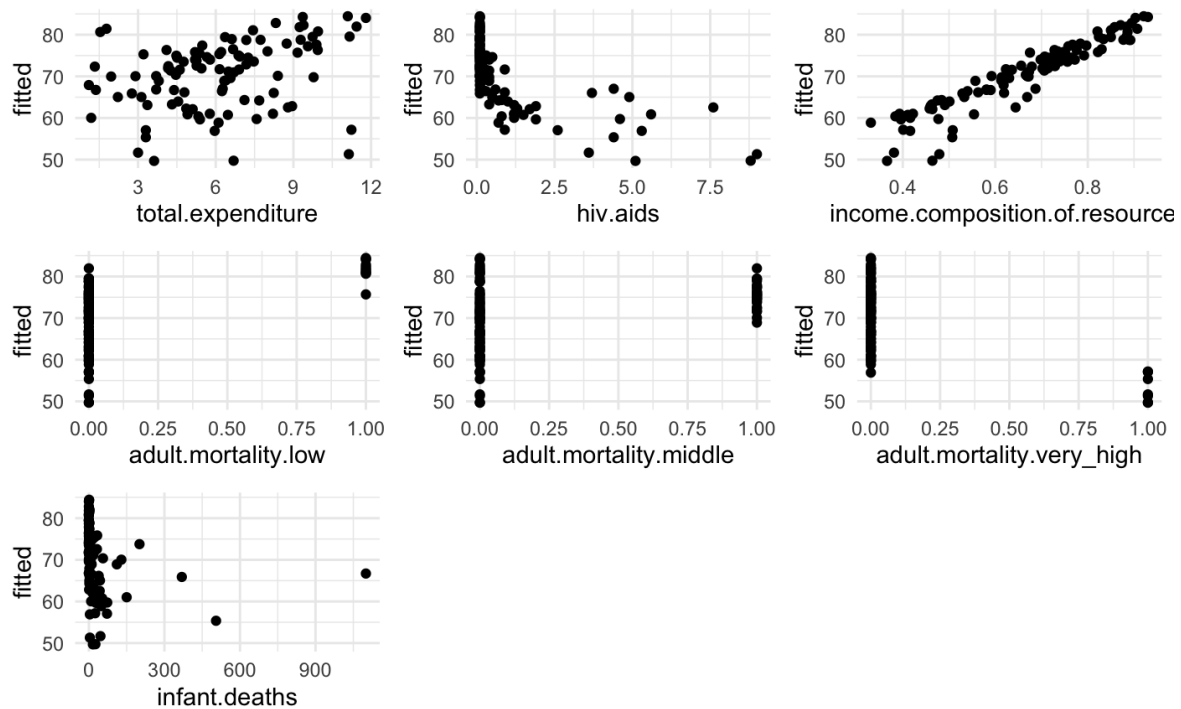


FIGURE 9 – scatterplots of residuals versus fitted values for the selected model

However, the best way to determinate if there is heteroskedasticity is to perform the **White test**. The hypothesis are,

$$H_0 : \text{there is homoscedasticity}$$

$$H_1 : \text{there is heteroskedasticity}$$

The p-value of the **White test** is 0.01729. At a significance level of $\alpha = 0.05$, we reject the null hypothesis and therefore we can conclude there is **heteroskedasticity**.

4.5.5 Autocorrelation

We can check for autocorrelation by performing the **Breusch-Godfrey test**. The hypotheses are,

$$H_0 : \text{there is no autocorrelation}$$

$$H_1 : \text{there is autocorrelation}$$

The p-value of the **Breusch-Godfrey test** is 0.938. At a significance level of $\alpha = 0.05$, we fail to reject the null hypothesis and therefore we can conclude there is **no autocorrelation**.

4.6 Normality of the residuals

Considering the Normal Q-Q plot of the residuals comparing the quantiles of the residuals versus the quantiles of a normal distribution. If the residuals are normal, the points on the following plot should follow the straight line,

Looking at this plot, we see that globally the points follow the straight line so the residuals are normals.

To ensure there is no a problem of normality for the residuals, we can perform a **Jarque Bera test**.

Let S be the skewness and κ of the residuals. The **Jarque Bera test** is given by,

$$JB = \frac{n}{6} \left(S^2 + \frac{(\kappa - 3)^2}{4} \right) \quad (10)$$

The hypothesis are,

$$H_0 : JB \sim X_2^2$$

$$H_1 : \text{residuals are not normally distributed}$$

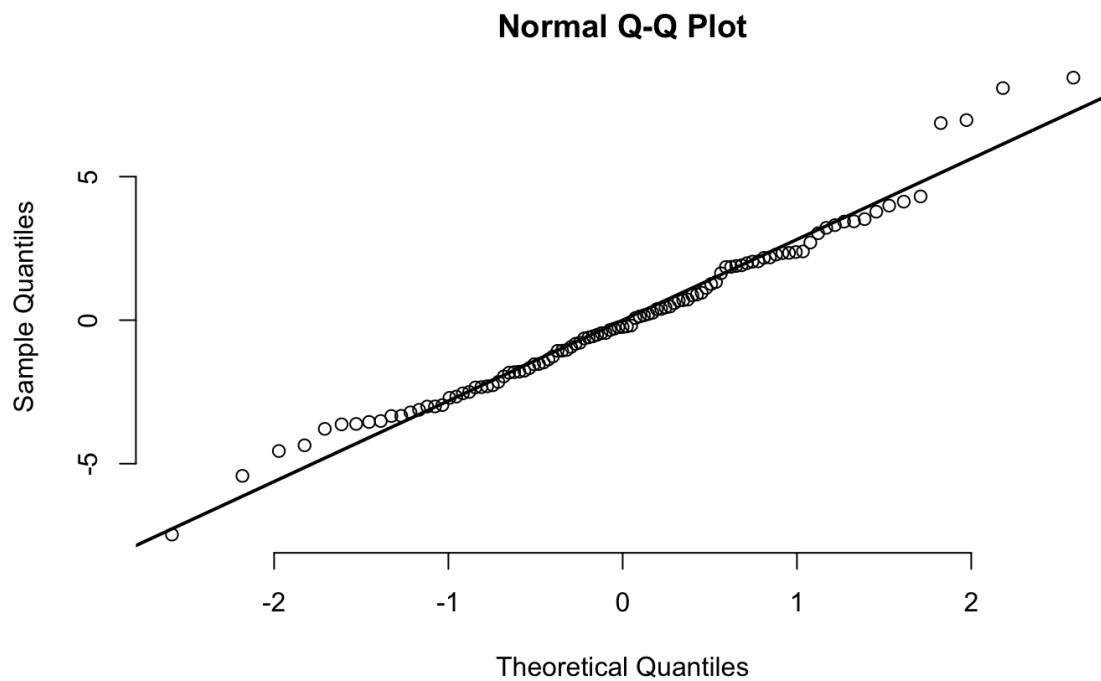


FIGURE 10 – *Normal QQ plot*

The p-value of the **Jarque Bera test** is 0.9287. At a significance level of $\alpha = 0.05$, we fail to reject the null hypothesis and therefore we can conclude the **residuals are normally distributed**.

Appendix

The **response variable** is *life.expectancy*.

The **explanatory variables** are the following :

Economic factors :

- *status* : is the country **developping** or **developped**.
- *percentage.expenditure* : expenditure on health as percentage of gross domestic product (PIB in french) per capita (%).
- *total.expenditure* : general government expenditure on health as a percentage of total government expenditure (%).
- *gdp* : gross domestic product per capita (\$).
- *income.composition.of.resources* : human development index (HDI) in terms of income composition of resources ($\in [0, 1]$).

Social factors

- *population* : total population of the country.
- *schooling* : number of years of schooling

Mortality factors :

- *adult.mortality* : probability of dying between 15 and 60 years old (**very low, low, middle, high, very high**).
- *under.five.deaths* : number of under five deaths per 1000 population.
- *infant.deaths* : number of infant deaths per 1000 population.
- *hiv.aids* : death per 1000 live births HIV/AIDS (between 0 and 4 years old).
- *thinness.5.9.years* : prevalence of thinness among children for age 5 to 9 years old.
- *thinness.10.19.years* : prevalence of thinness among children and adolescents for age 10 to 19 years old.

Immunization factors

- *hepatitis.b* : Hepatitis B immunization coverage among 1 year olds (%).
- *measles* : number of reported cases of measles per 1000 population.
- *polio* : Polio immunization coverage among 1 year olds (%).
- *diphtheria* : Diphtheria tetanus toxoid and pertussis (DPT3) immunization coverage among 1 year olds (%)

Other factors

- *alcohol* : consumption of alcohol (15 years old or more) per capita (**in litres of pure alcohol**).
- *bmi* : average BMI of entire population.

VIF for full model

Variation inflation factor (VIF)		
full model		
Variables	Tolerance	VIF
infant.deaths	0.35	2.89
alcohol	0.32	3.10
percentage.expenditure	0.12	8.10
hepatitis.b	0.20	4.97
measles	0.34	2.91
bmi	0.42	2.36
polio	0.33	3.03
total.expenditure	0.68	1.47
diphtheria	0.14	7.21
hiv.aids	0.48	2.09
gdp	0.08	13.10
population	0.41	2.46
thinness.10.19.years	0.17	5.83
thinness.5.9.years	0.17	5.81
income.composition.of.resources	0.08	12.72
schooling	0.14	7.40
status.developed	0.37	2.71
adult.mortality.low	0.42	2.36
adult.mortality.middle	0.41	2.43
adult.mortality.high	0.36	2.75
adult.mortality.very_high	0.43	2.33