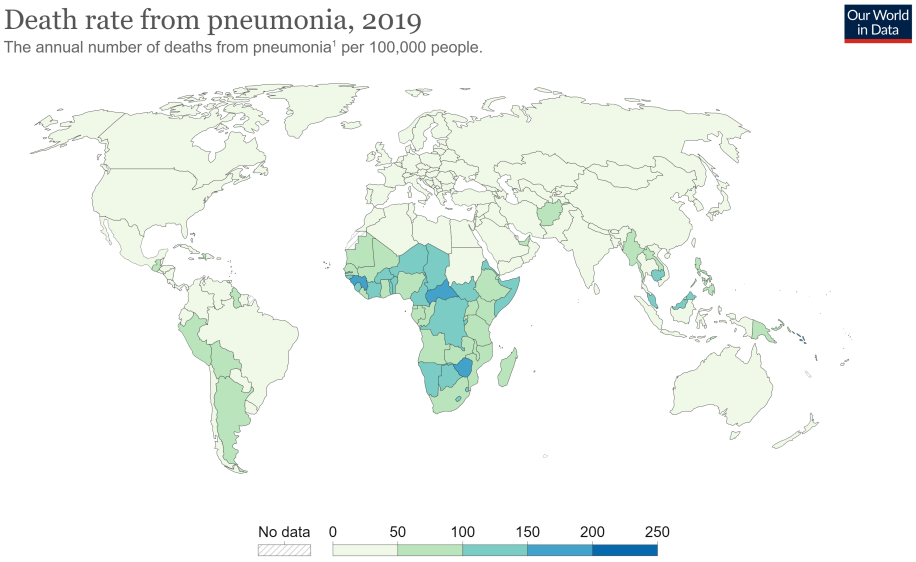
Research project: Pneumonia in 2019

Elisa Degara Emanuele Marino

1 Introduction

'Pneumonia is a form of acute respiratory infection that is most commonly caused by viruses or bacteria. It can cause mild to life-threatening illness in people of all ages.'
(Pneumonia - World Health Organization (WHO)).

In 2019 about 2.5 million people died from pneumonia. Even though this number is slowly falling through years, pneumonia remains a serious illness with a high death rate.
The aim of our statistical research is to analyze different factors that may lead to death from pneumonia and verify that there is a higher risk of death for poorer countries. For this reason, we divide the chosen factors into three macro-categories: socio-economical, lifestyle and environmental risks.
Indeed, by looking at the following map that represents the number of deaths due to pneumonia per year per 100,000 individuals, we can understand that the death rate is higher in Sub-Saharan Africa and Southeast-Asia; clearly, there is a big difference between richer and poorer countries.



2 Constructing the data set

For our project we create a data set merging together data collected from multiple sources (section 7). The response variable in our study is '*deaths from pneumonia*', while all the explanatory variables, with the relative abbreviations used in the R code, are listed below.
As already said, we divide them into three main categories, corresponding to different types of risk factor. The data set consists of 138 entries, 12 columns in total.

Abbreviations:

gea - geographic area
coun - countries
dep - number of deaths from pneumonia per 100,000 individuals, from under-5 to 70+ years old people

Socio-economical risks

gdp - GDP per capita (measured in current American $)
old - old age dependency (ratio of the number of people older than 64 relative to the number of people in the working age population (15-64 years))
deh - number of deaths attributed to a lack of access to hand-washing facilities per 100,000 people
dew - number of deaths attributed to unsafe water sources per 100,000 people
dem - number of deaths from protein-energy malnutrition per 100,000 people

Lifestyle risks

smo - share of men and women (15+ years) who smoke any tobacco product on a daily / non-daily basis
psmo - number of deaths from second hand smoking (aka passive smoking)

Environmental risks

air - air pollution (concentrations of fine particulate matter (PM2.5))
inair - household (or indoor) air pollution from solid fuels

3 At first sight

First of all, we explore the data set by plotting each explanatory variable against the response to get a feeling of their correlation. We are using the *ggplot2* library to make nicer plots, dividing countries by geographic areas. We also restrict the domain of **psmo** and **inair** to have a wider view of the scatterplots. We insert them in section 9 to have a better display.
There seems to be no direct correlation with **gdp**, while visually there is some correlation with **old**, even tough it is the opposite of what you would expect: **coun** with lower **old** have higher **dep**.
It seems that **deh**, **dem** and **dew** are all correlated to **dep** in a similar way. **deh**, **dem** and **dew** have higher values in Africa's **coun**, that is the poorest region of the world according to the Multidimensional Poverty Measure (MPM) (see section 7). Indeed, in those countries living expectation is much lower, so **old** is very small, therefore the correlation we noticed for **old** starts to make sense. In addition, in poorer regions living standards are very low, so it is reasonable that the values of **deh**, **dem** and **dew** are high. Overall we can say that there seems to be a correlation between poverty and **dep**.
Regarding **air** and **inair**, we cannot find an evident observable correlation. The same applies for **smo** and **psmo**, but we can notice that **smo** behaves as **old**: African **coun** have lower **smo** and higher **dep**.

4 Multiple regression

Let's perform a multiple linear regression to understand the relation between the covariates and the response and explain part of the variance of **dep**.

```
Residuals:
    Min      1Q  Median      3Q     Max
-501.03 -145.65  -32.09   68.65  744.19

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.293e+02  8.871e+01   8.222 1.94e-13 ***
gdp         -6.308e-04  9.918e-04  -0.636    0.526
old         -1.279e+01  2.653e+00  -4.821 3.97e-06 ***
deh          1.829e+01  4.105e+00   4.457 1.80e-05 ***
dew          6.683e+00  2.965e+01   0.225    0.822
dem          1.791e+00  3.998e+00   0.448    0.655
smo         -2.785e+00  2.322e+00  -1.200    0.233
psmo         8.353e-05  1.020e-03   0.082    0.935
air         -4.545e+00  1.824e+00  -2.492    0.014 *
inair       -2.968e-04  7.567e-04  -0.392    0.696
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 233.8 on 128 degrees of freedom
Multiple R-squared:  0.704,     Adjusted R-squared:  0.6832
F-statistic: 33.83 on 9 and 128 DF,  p-value: < 2.2e-16
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.98838 -0.31941  0.03016  0.26341  1.38484

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.720e+00  1.924e-01  34.920  < 2e-16 ***
gdp         -2.615e-06  2.152e-06  -1.215  0.22643
old         -3.402e-02  5.755e-03  -5.910 2.91e-08 ***
deh          1.800e-02  8.905e-03   2.021  0.04538 *
dew          5.485e-02  6.432e-02   0.853  0.39531
dem          6.076e-03  8.674e-03   0.700  0.48494
smo         -1.077e-02  5.037e-03  -2.138  0.03440 *
psmo        -1.168e-06  2.212e-06  -0.528  0.59843
air         -1.073e-02  3.956e-03  -2.713  0.00759 **
inair        4.848e-07  1.642e-06   0.295  0.76823
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5071 on 128 degrees of freedom
Multiple R-squared:  0.6461,    Adjusted R-squared:  0.6212
F-statistic: 25.97 on 9 and 128 DF,  p-value: < 2.2e-16
```

We first consider all the covariates we mentioned above (*mul_regression* in the R code). Looking at the adjusted R-squared, the model explains about 68% of the variance and **old**, **deh** and **air** are significant at 0.05 level.

It is worth noticing that if we perform simple linear regressions, covariates like **dew** and **dem** are significant, contrary to what happens in *mul_regression*. That's reasonable, since we noticed that probably these covariates have some kind of correlation between them, as we discussed in section 3. Moreover, looking at the p-value, this model is better than having no model at all.

Let's check the assumptions of normality and homoscedasticity of the residuals to understand whether the model is statistically significant. Plotting the residuals against the fitted values, we see that there is a fairly strong violation of homoscedasticity, and while the histogram and the Q-Q plot seem ok, the performed Shapiro Wilk Test fails, so we cannot assume normality neither homoscedasticity.

To fix this problem we try another multiple linear regression, using log(**dep**) as response (*mul_regression2* in the R code). This model explains about 62% of the variance of **dep**. We get that **old**, **deh** and **air** are still relevant and that **smo** is now significant too (at level 0.05). As to normality we have again a reasonable histogram and Q-Q plot and now the Shapiro Wilk test does not fail: we do not have evidence against normality. Plotting the residuals vs the fitted values, we find that there is still a milder violation. Therefore we have to take into account the fact that homoscedasticity is not completely satisfied, so our analysis may not be so solid after all. (Plots in section 9)
In the next section, we try to understand whether it actually makes sense to use *mul_regression2*.

5 Selecting the model

Starting from *mul_regression*, we perform stepwise model selection by AIC: the *stepAIC* function selects the best model using AIC as stopping criterion.
The output includes **old**, **deh** and **air** as covariates (*model1*).
We repeat the same procedure starting from *mul_regression2*: as expected, the output includes **old**, **deh**, **air** and **smo** as covariates (*model2*).

Now we want to understand whether it's a good idea to perform a multiple linear regression using log(**dep**) as response. We compare the BIC of *model1* and *model2*. The BIC of *model2* is much smaller, therefore we conclude that it provides a better fit: the final selected model is *model2*.

```
> BIC(model1, model2)
       df       BIC
model1  5 1913.9923
model2  6  227.7291
```

6 Results and limitations of the study

Let's analyze the regression *model2*. The model explains about 62% of the variance of **dep**, so it's pretty

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.11728 -0.35982  0.01022  0.29402  1.45506

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.727581   0.169009  39.806  < 2e-16 ***
old         -0.037386   0.005186  -7.209 3.82e-11 ***
deh          0.028945   0.003797   7.624 4.17e-12 ***
smo         -0.008875   0.004855  -1.828  0.06975 .
air         -0.011672   0.003583  -3.257  0.00143 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5054 on 133 degrees of freedom
Multiple R-squared:  0.6348,    Adjusted R-squared:  0.6239
F-statistic: 57.81 on 4 and 133 DF,  p-value: < 2.2e-16
```

good. All the covariates but **smo** are significant at level 0.05, while they are all relevant at level 0.1.
The final p-value suggests us that the model is better than no model at all. The residual standard error is 0.5054, which should be good.
Let's check the assumptions of normality and homoscedasticity of the residuals on the chosen model. We obtain reasonable histogram and Q-Q plot, the output of the Shapiro Wilk test is 0.3, clearly better than the one of *mul_regression*: normality is reasonable. Plotting the residuals against the fitted values, we encounter the same milder violation as before, so homoscedasticity is not completely satisfied. Again, our analysis may not be flawless. (Plots in section 9)

In the final selected model, we see that the most significant covariates include **old**, **deh**, **air**:

- **old**: as we pointed out in section 3, poorest regions have low **old** and high **dep**
- **deh**: poorest regions have high **deh** and high **dep**
- **air**: poorest regions have high **air** and high **dep**

(again, we refer to Africa as the poorest area of the world, according to the Multidimensional Poverty Measure (MPM)).
This seems to confirm the hypothesis of our statistical research: there is some correlation between **dep** and 'poverty', so it can be considered a significant factor of risk.

Despite the quite satisfactory result, there are some limitations in the way we conducted our research.

The first thing to point out is that we selected only some of the many possible covariates, but surely we did not include all the relevant factors. Another problem arises from the fact that pneumonia is commonly caused by viruses and bacteria, so it is not easy to determine how other factors come into play.

Moreover, as we already underlined in section 2, there may be some correlation between the covariates we have chosen, therefore not considering the interaction between them is a simplification.

Furthermore, we said that the hypothesis of a correlation between poverty and risk of death from pneumonia is confirmed, but we have to take into account that we decided to use a limited number of factors to represent 'poverty'. This is a simplification, since the phenomenon of poverty is more complex and it depends on many other aspects.

The size of our dataset is another important limitation: since data is not available for all countries, we kept only the ones with all data points of interest, thus we considered just 138 of them. We also decided to use only the data of the year 2019, therefore our research may not be very accurate.

7 These are the references you are looking for

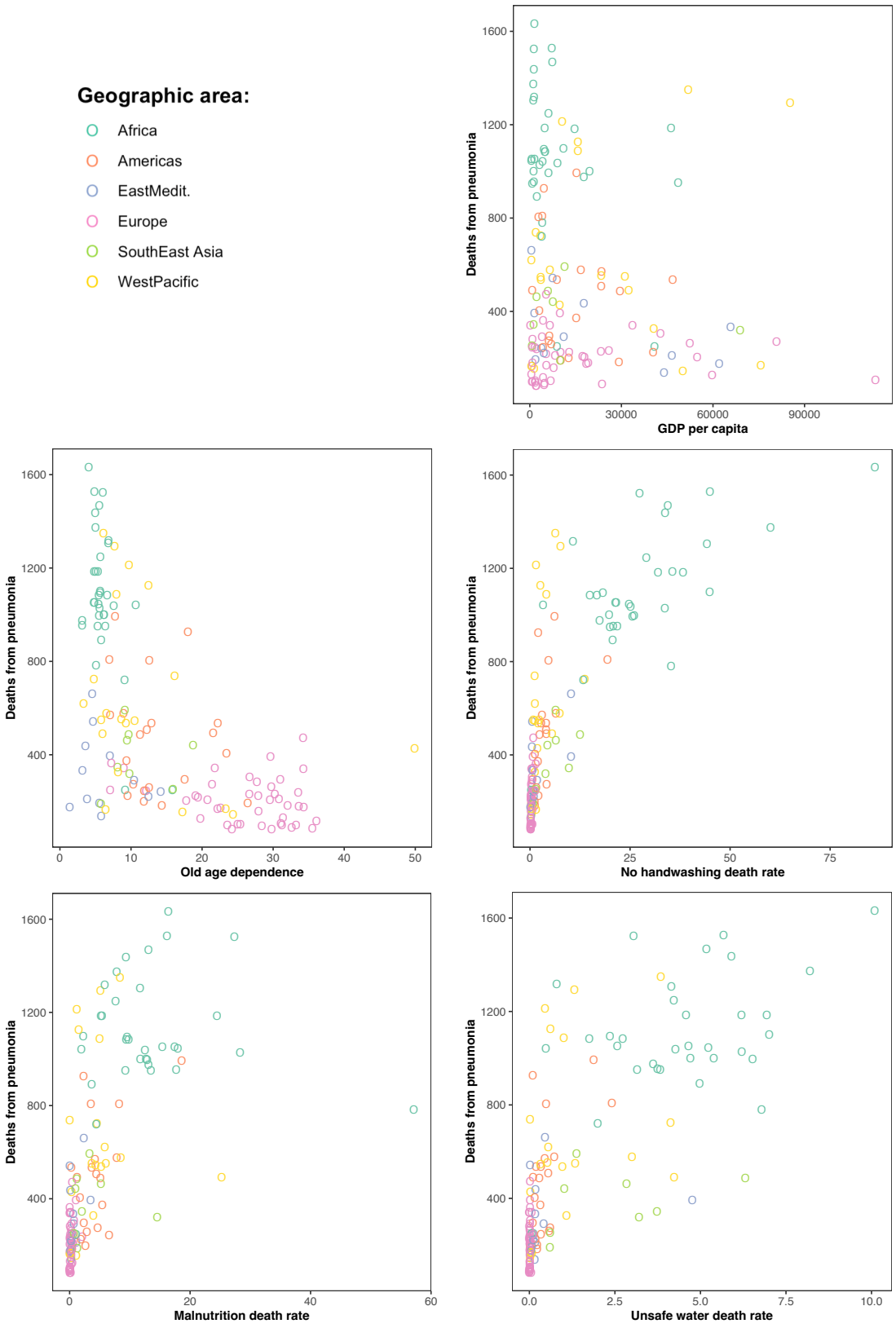
- *Map section 1*: https://ourworldindata.org/grapher/pneumonia-death-rates-age-standardized
- *Deaths from pneumonia*: https://ourworldindata.org/grapher/pneumonia-mortality-by-age
- *GDP per capita*: https://data.worldbank.org/indicator/NY.GDP.MKTP.CD
- *Old-age dependency ratio*: https://ourworldindata.org/grapher/age-dependency-ratio-old
- *Deaths from lack of hand-washing facilities*: https://ourworldindata.org/grapher/number-of-deaths-by-risk-factor
- *Deaths from unsafe water sources*: https://ourworldindata.org/grapher/share-deaths-unsafe-water
- *Deaths from malnutrition*: https://ourworldindata.org/grapher/malnutrition-death-rates
- *Share of smokers*: https://ourworldindata.org/grapher/share-of-adults-who-smoke
- *Deaths from second hand smoking*: https://ourworldindata.org/grapher/number-of-deaths-by-risk-factor?time=latest
- *Concentration of PM2.5*: https://www.who.int/data/gho/data/themes/air-pollution/who-air-quality-database
- *Indoor air pollution*: https://ourworldindata.org/grapher/number-of-deaths-by-risk-factor?time=latest
- *Multidimensional Poverty Measure (circa 2018), World Bank, Washington, DC.*: https://www.worldbank.org/en/topic/poverty/brief/multidimensional-poverty-measure

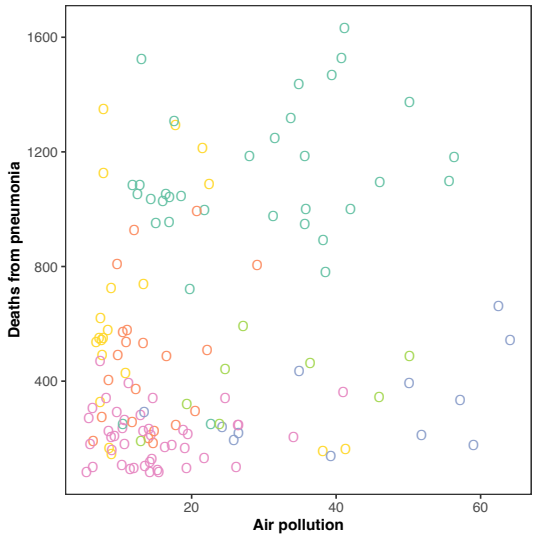
8 R code

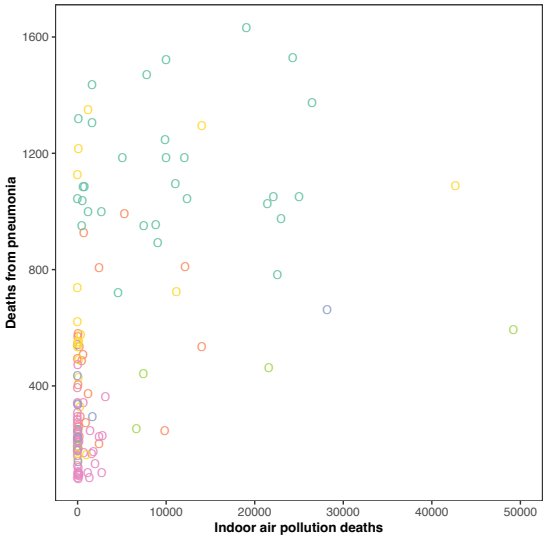
The R code and all relative data sets are provided in a separate folder.

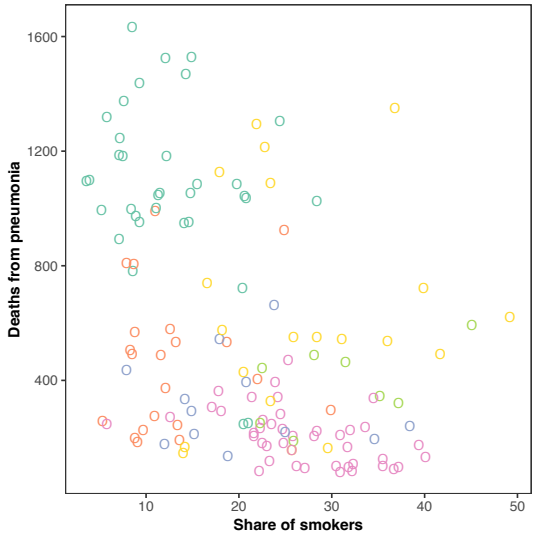
9 Visual complements

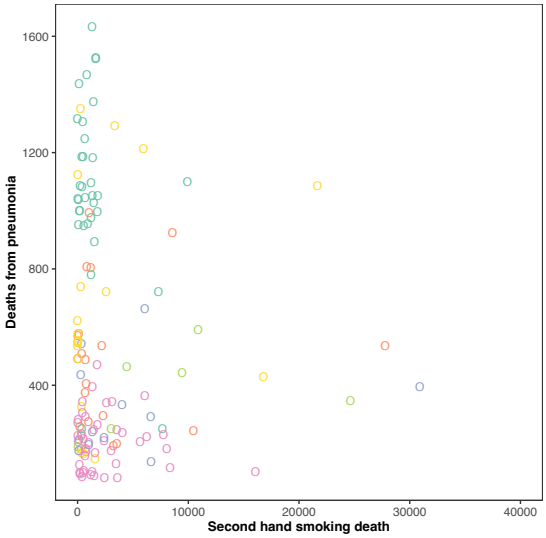
3 At first sight




Deaths from pneumonia
1600
1200
800
400
20
40
60
Air pollution

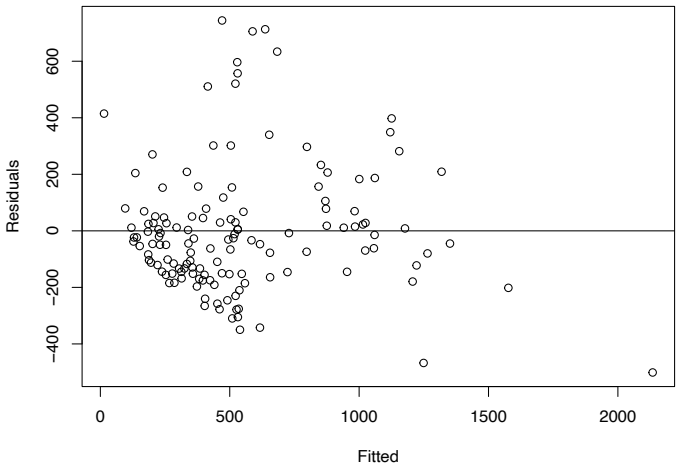

Deaths from pneumonia
1600
1200
800
400
0
10000
20000
30000
40000
50000
Indoor air pollution deaths

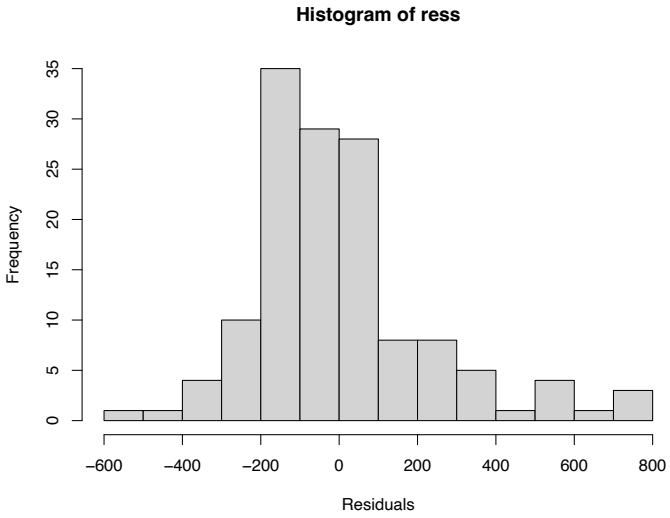

Deaths from pneumonia
1600
1200
800
400
10
20
30
40
50
Share of smokers

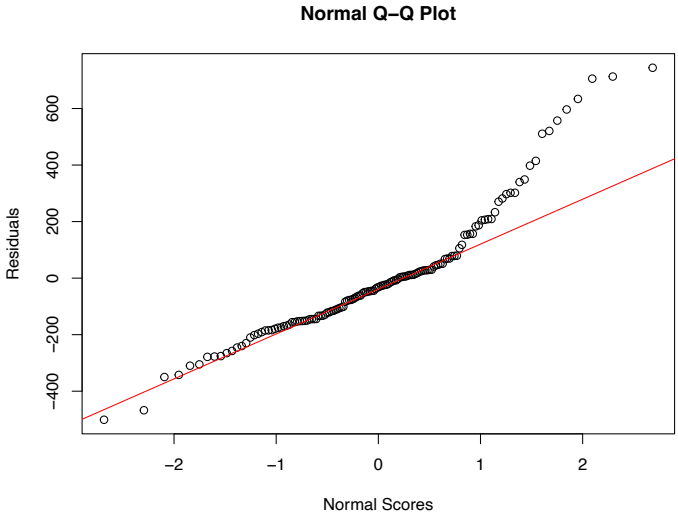

Deaths from pneumonia
1600
1200
800
400
0
10000
20000
30000
40000
Second hand smoking death

4 Multiple regression

mul_regression: checking normality and homoscedasticity

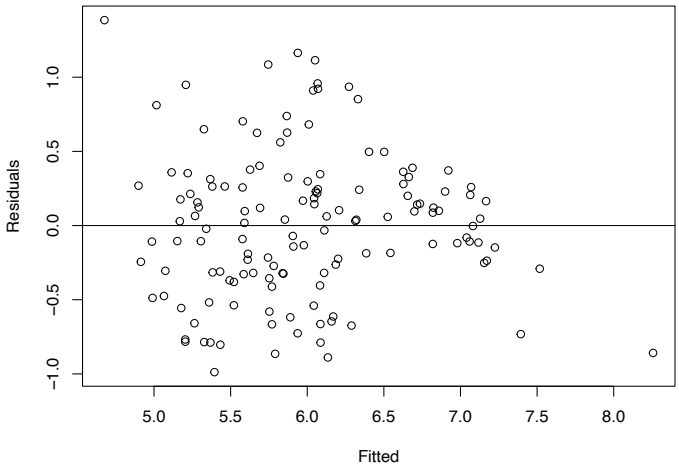


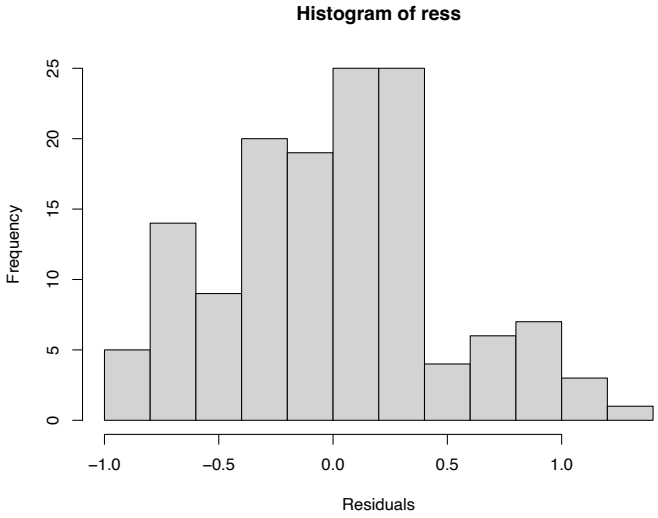


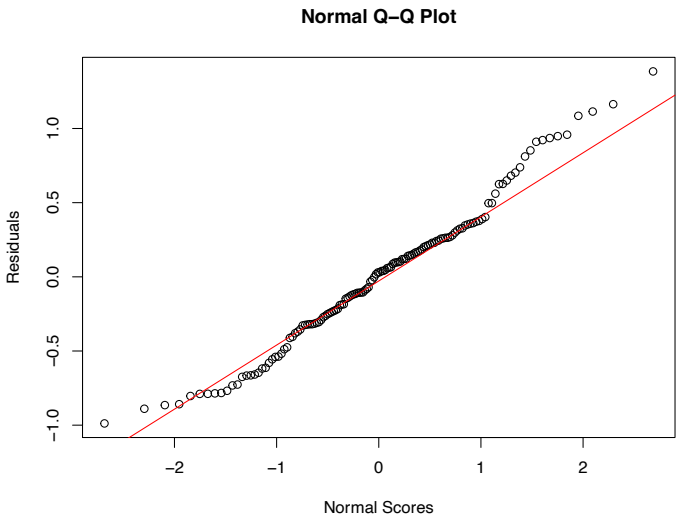



```
        Shapiro-Wilk normality test

data:  residuals(mul_regression)
W = 0.91379, p-value = 2.287e-07
```

mul_regression2: checking normality and homoscedasticity



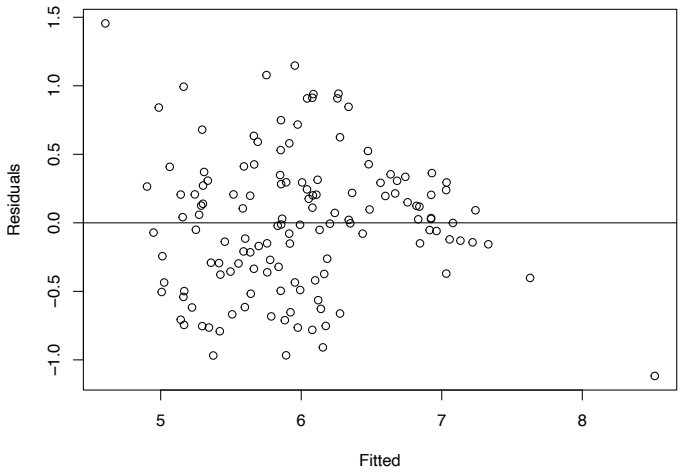


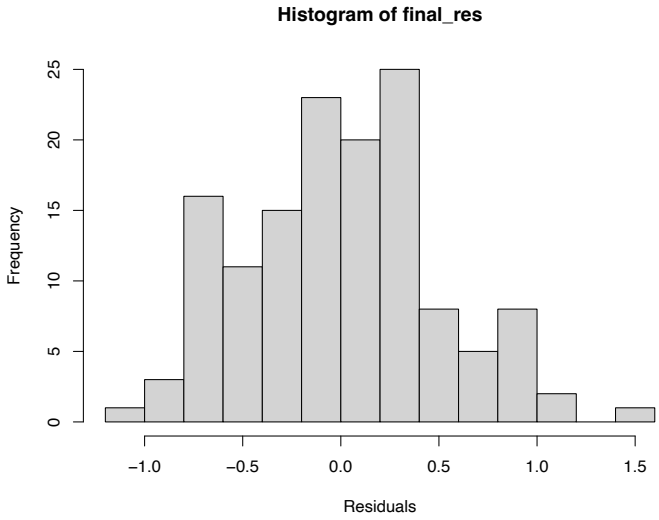


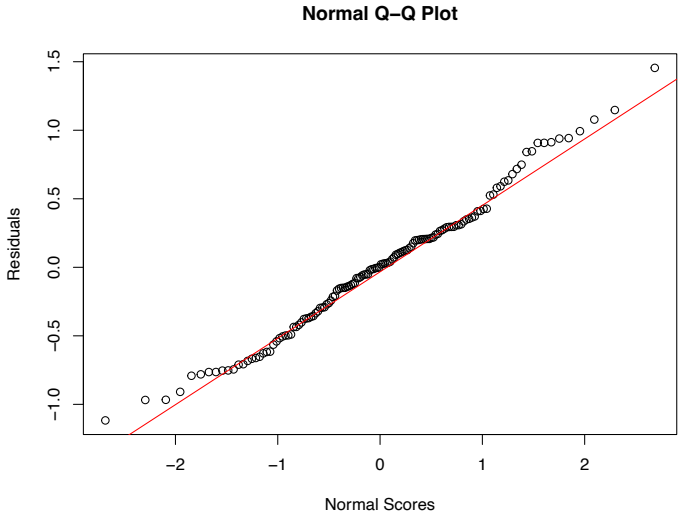

```
        Shapiro-Wilk normality test

data:  residuals(mul_regression2)
W = 0.98242, p-value = 0.07292
```

6 Results

model2: checking normality and homoscedasticity








```
        Shapiro-Wilk normality test

data:  final_res
W = 0.98836, p-value = 0.3002
```