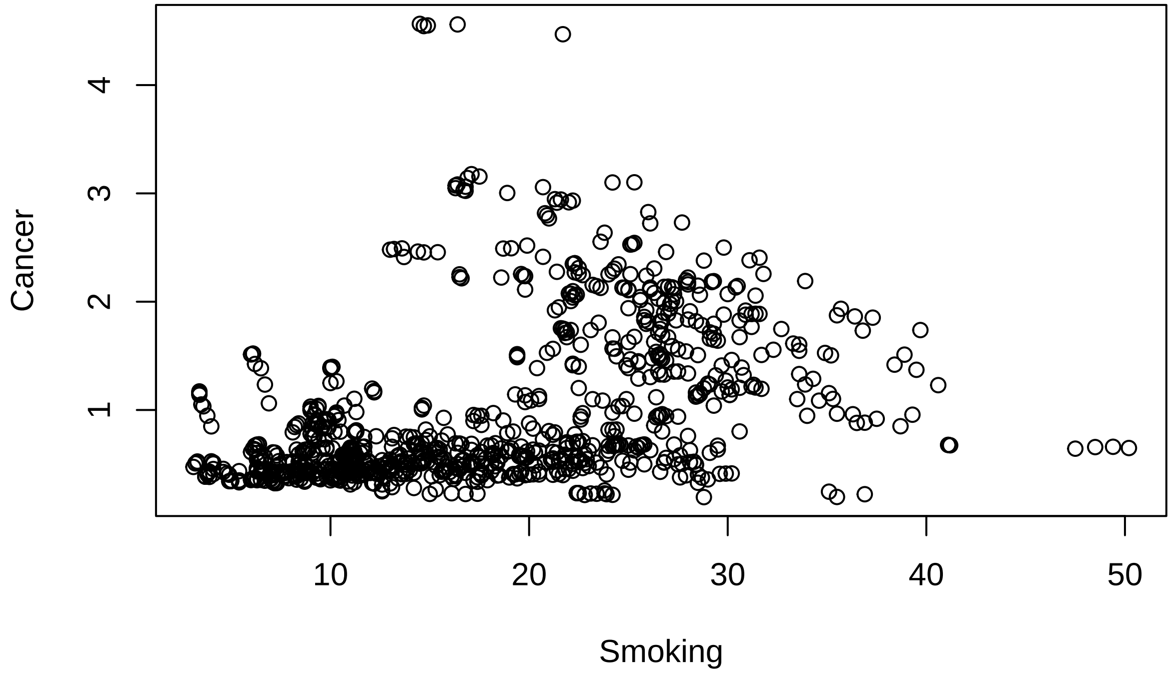
Multivariate Linear Regression.Rmd

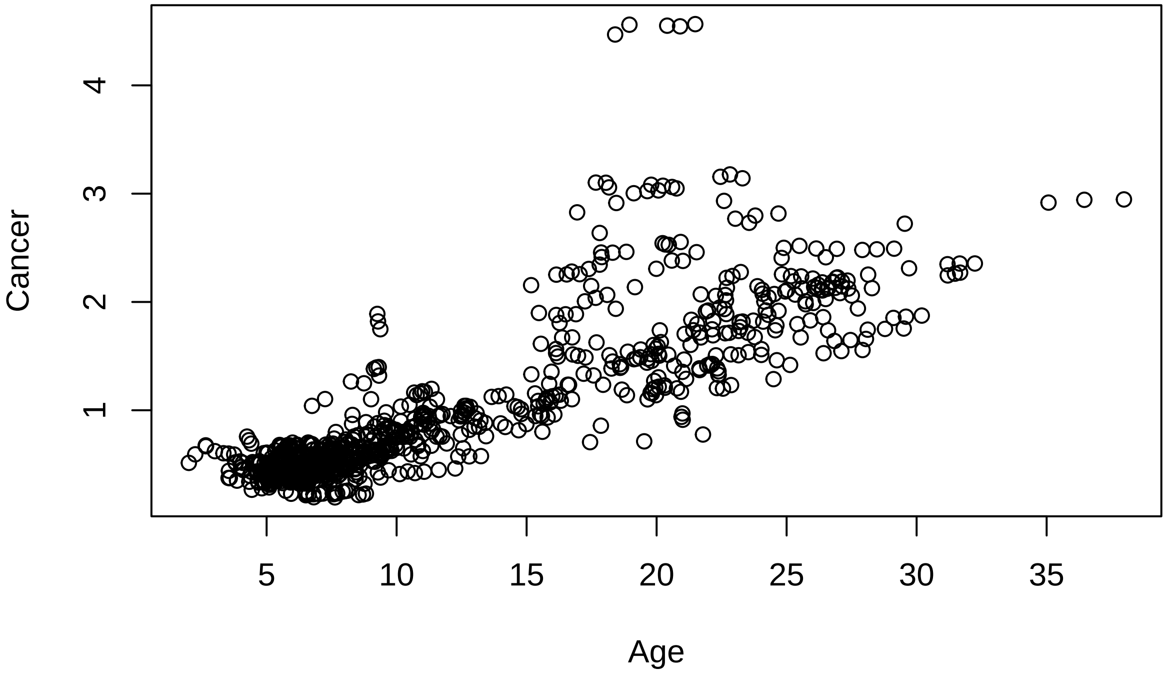
R Markdown

We are now going to perform a multivariate linear regression, trying to examine the relationship between some explanatory variable and one response variable. Regarding our case, we are going to take in consideration as explanatory variables some factors belonging to various categories that we considered interesting to analyze and to help us to find some interesting correlation with the number of people affected by cancer worldwide. We decided to use data of approximately 150 countries, for approximately 20 years. In this way we will be able to do not miss some trend or to do not being deceived from some outliers years or random trends. Coming back to the explanatory variables they are divided in 3 categories, each one with some subcategories: environment(Air Pollution), lifestyle(Smoking, Obesity, Alcohol), socioeconomic/demographic condition (GDP of the country and Age). We are trying then to see which of these variables explain better the data that we have. Before performing the actual linear regression, let's analyze the data in a visual way, plotting the cancer cases against each of the covariates.

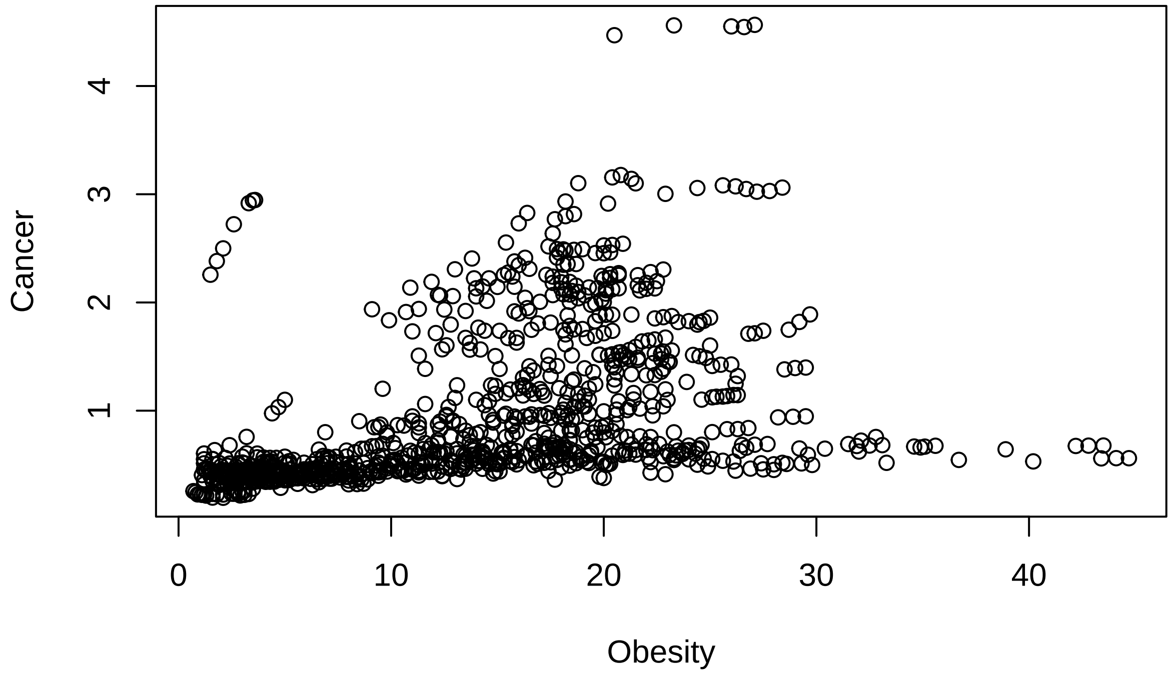
```
mydata = read.csv("Cancer_final.csv")
data3 = na.omit(mydata)

plot(Cancer ~ Smoking, data=data3)
```



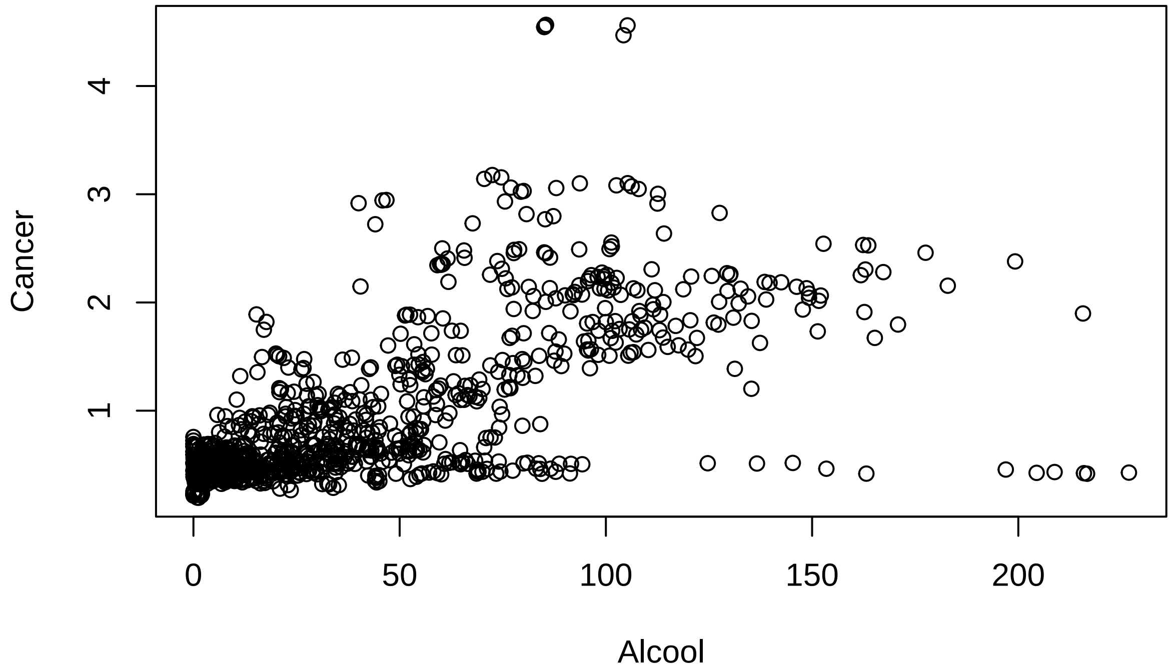
```
plot(Cancer ~ Age, data=data3)
```



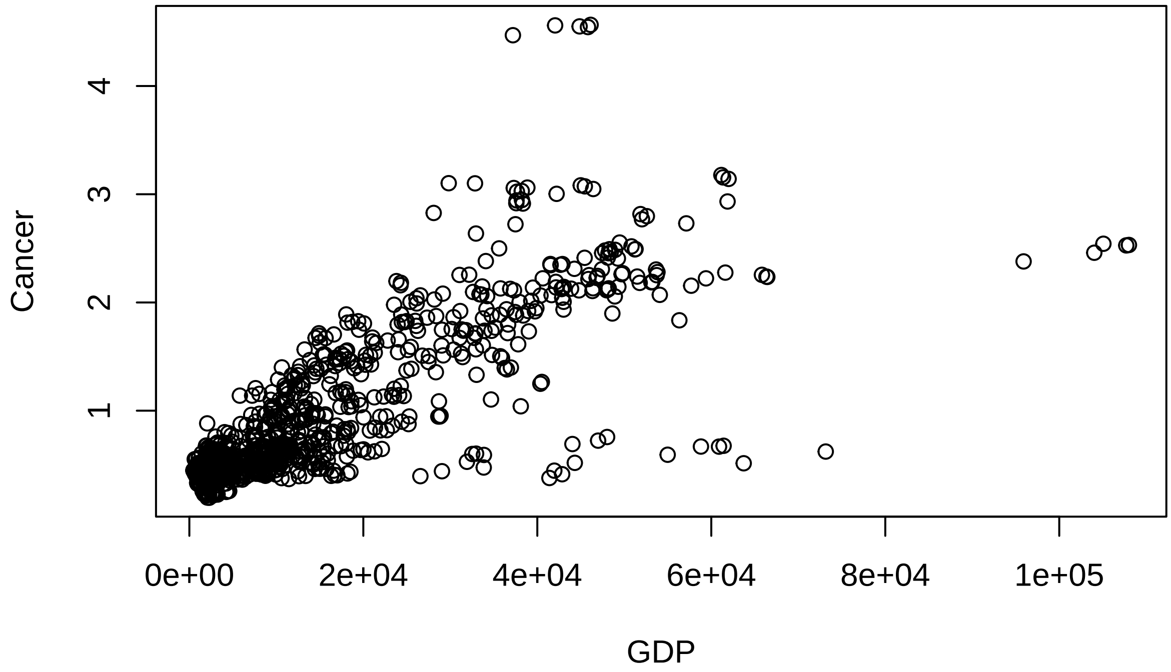
```
plot(Cancer ~ Obesity, data=data3)
```



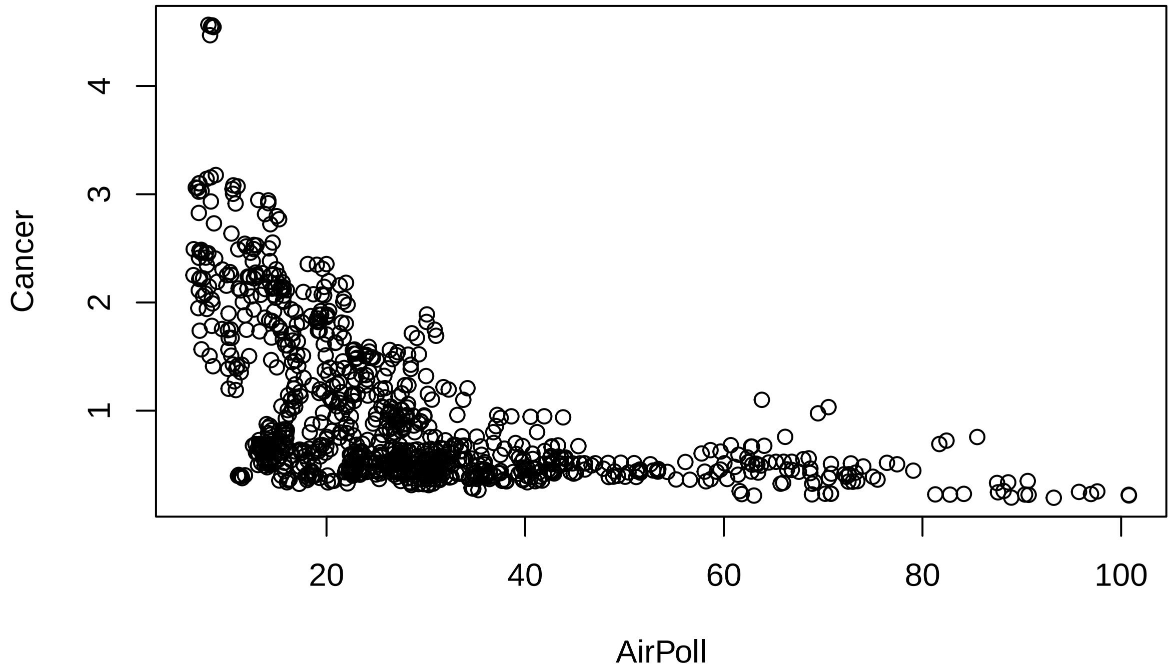
```
plot(Cancer ~ Alcool, data=data3)
```



```
plot(Cancer ~ GDP, data=data3)
```



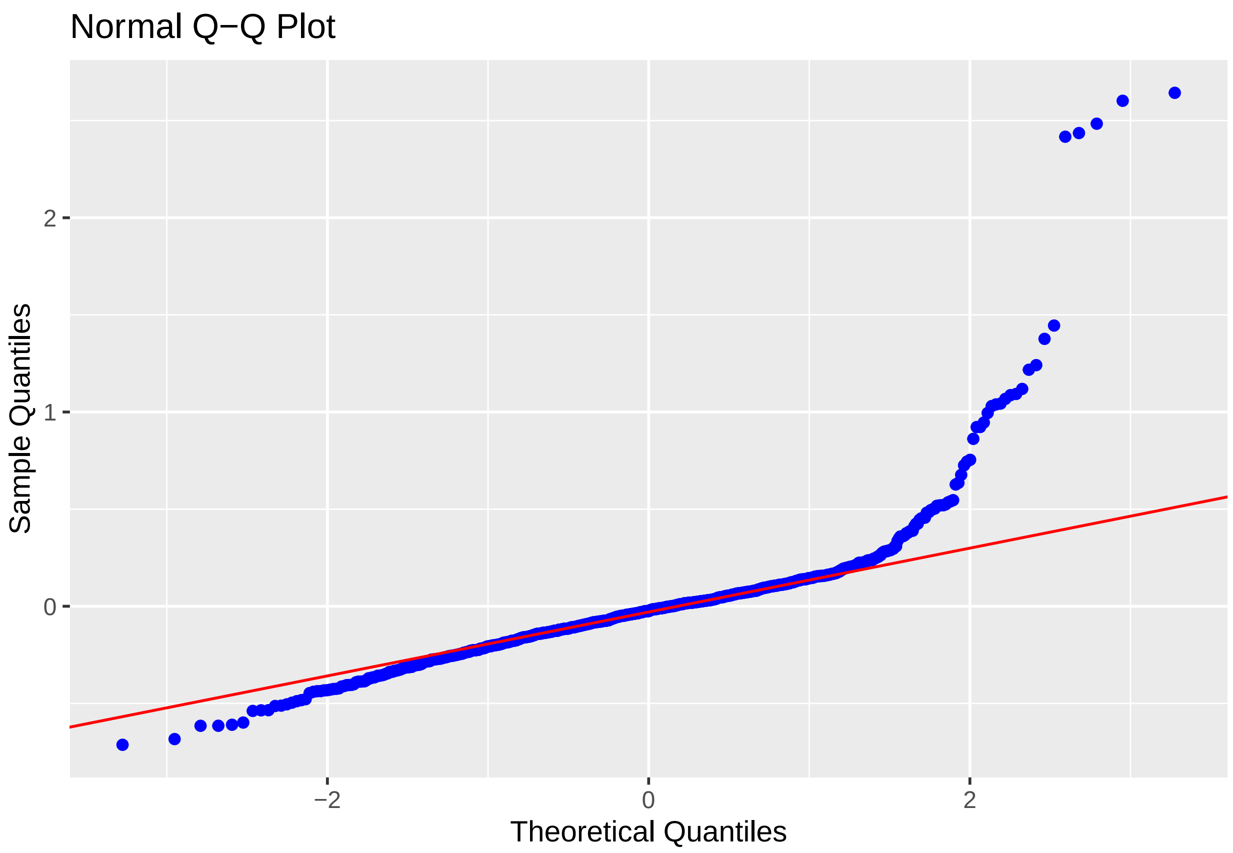
```
plot(Cancer ~ AirPoll, data=data3)
```



```
#plot(Age ~ GDP, data=data3)
```

From this simple data visualization, we can see that we can identify some sort of relation between cancer cases and some of our explanatory variables, still as not strong as we imagined before doing those plots. We can see that age is the factor which presents a stronger correlation visually. We also have some outliers, which are Canada and US, but we will analyze this further in the limitation of our study. Now, let's proceed with the last step before performing the actual multiple linear regression: we will see if the assumptions on the data are met. Let's check for the normality and for the homoscedasticity of the residuals.

```
#Normality of the residuals
model = lm( Cancer~ Smoking + Obesity + GDP + Age + AirPoll + Alcool, data = data3)
ols_plot_resid_qq(model)
```



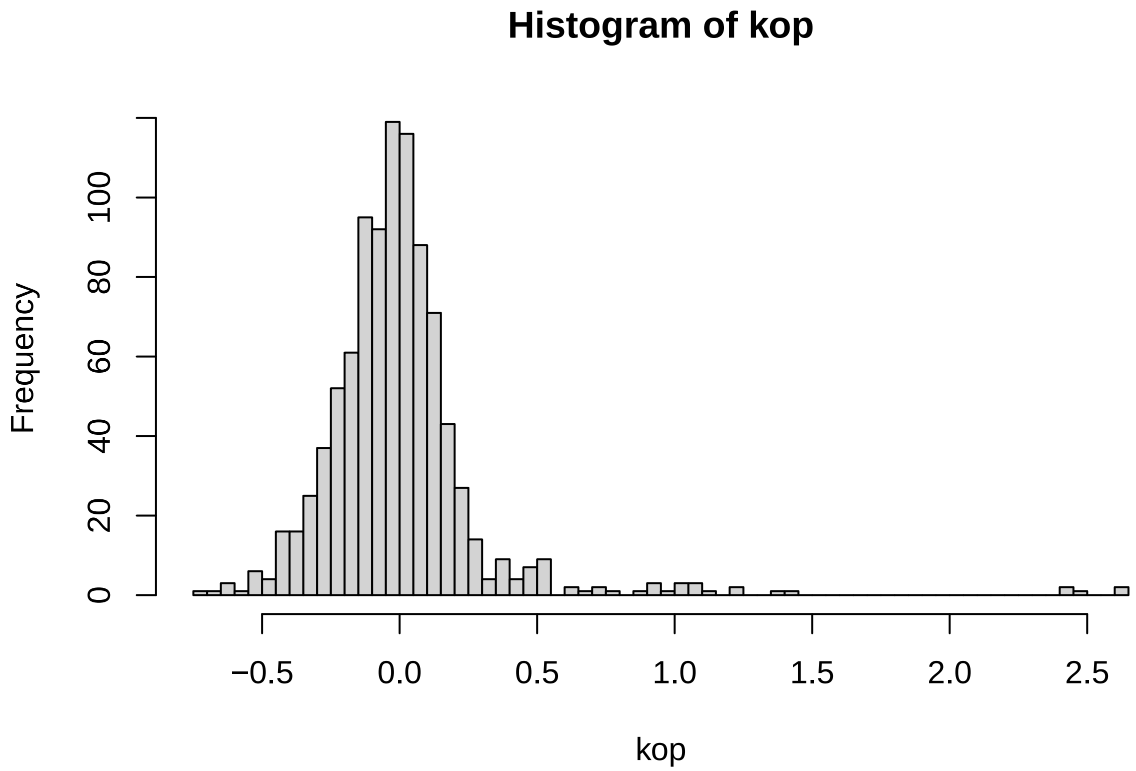
```r
ols_test_normality(model)
```

```
## -----------------------------------------------
##        Test             Statistic       pvalue
## -----------------------------------------------
## Shapiro-Wilk              0.722          0.0000
## Kolmogorov-Smirnov        0.1671         0.0000
## Cramer-von Mises        216.6397         0.0000
## Anderson-Darling         46.5541         0.0000
## -----------------------------------------------
```

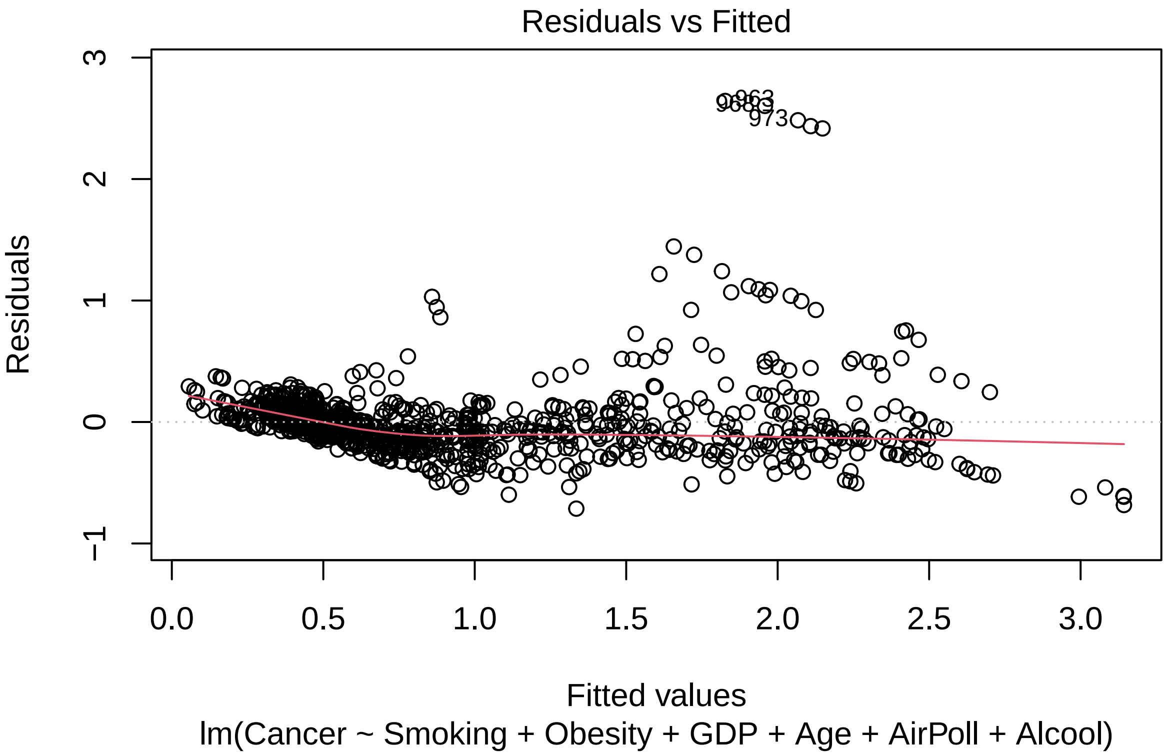
```r
ols_test_correlation(model)
```

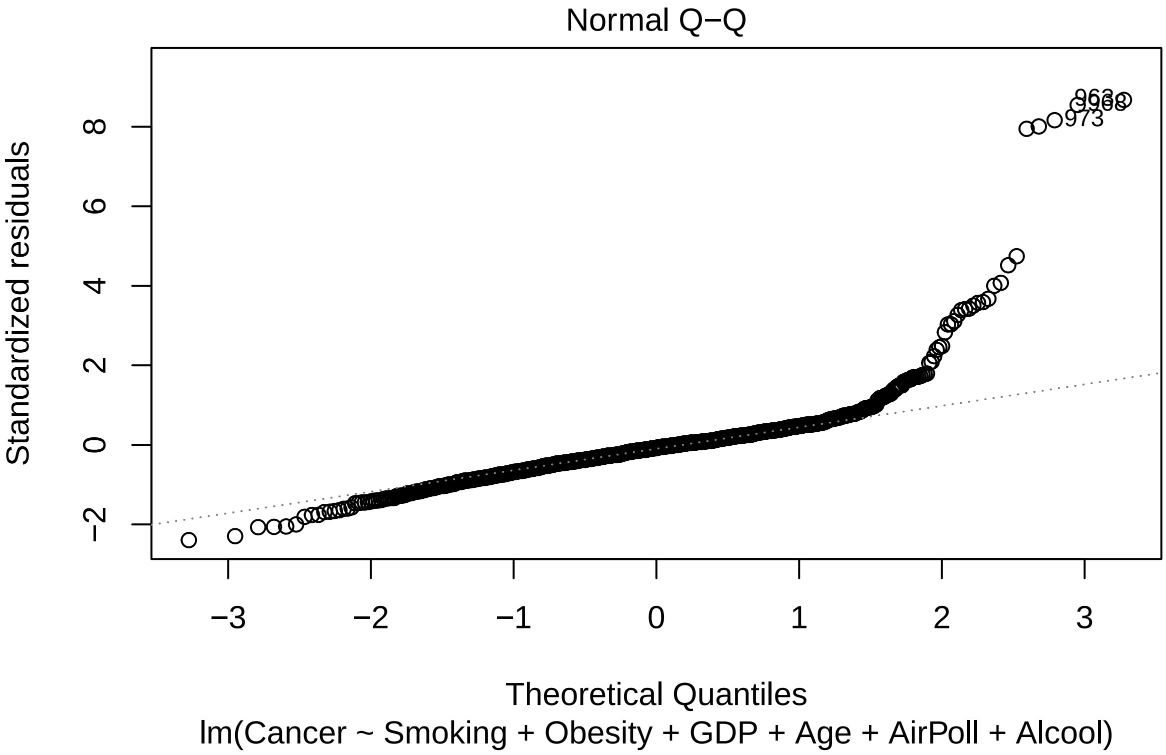
```
## [1] 0.8482072
```

```r
#ols_plot_resid_hist(model)
kop = residuals.lm(model)
hist(kop, breaks = 50)
```

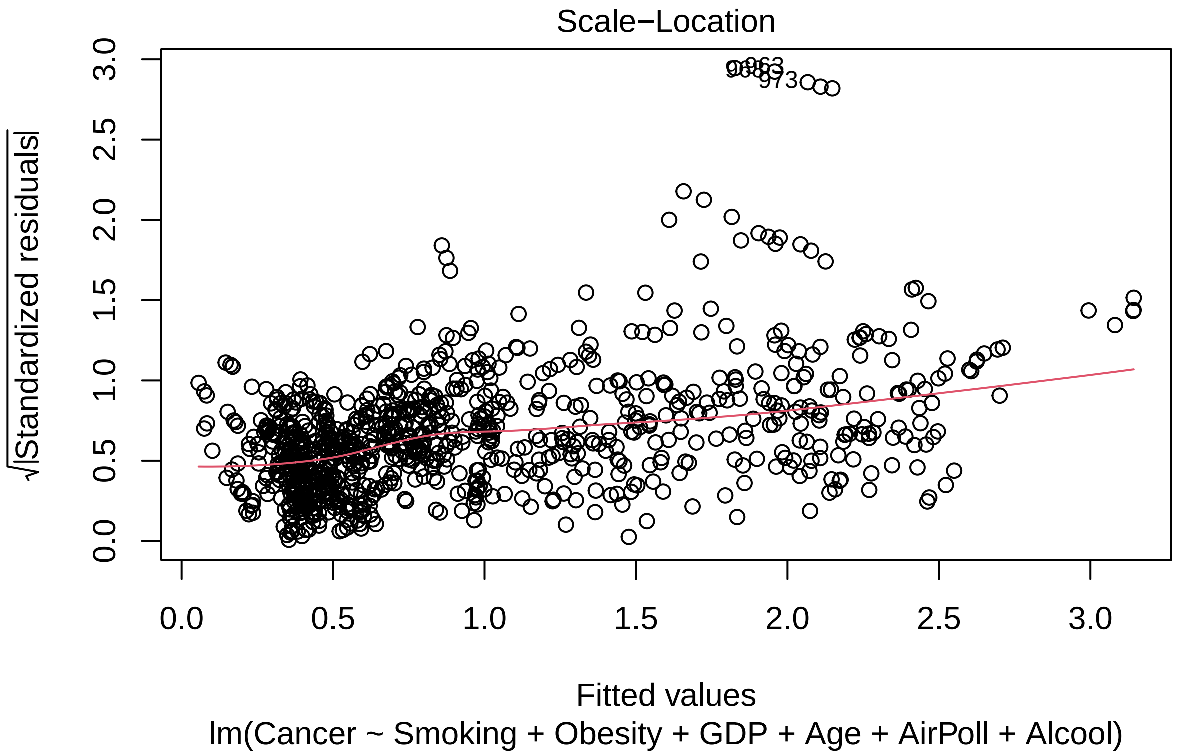


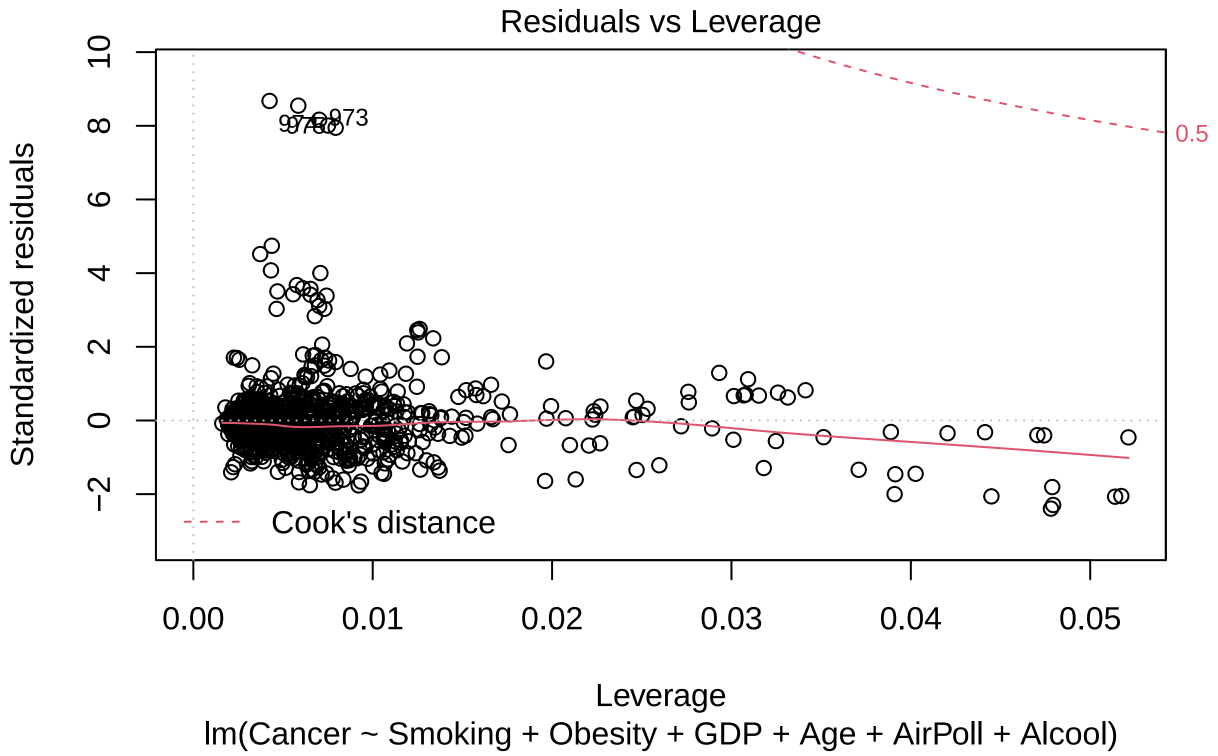
```
#homoscedasticity of the residuals
plot(model)
```


Residuals vs Fitted
963
968
973
Residuals
3
2
1
0
−1
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Fitted values
lm(Cancer ~ Smoking + Obesity + GDP + Age + AirPoll + Alcool)



Normal Q−Q
Standardized residuals
−2
0
2
4
6
8
−3
−2
−1
0
1
2
3
963
968
973
Theoretical Quantiles
lm(Cancer ~ Smoking + Obesity + GDP + Age + AirPoll + Alcool)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
3.0
963
968
973
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Fitted values
lm(Cancer ~ Smoking + Obesity + GDP + Age + AirPoll + Alcool)



Let's analyze each of the test we performed above, and let's see which conclusion we can take from what we observed. About the model, we can see that the p-value of the tests is satisfying for all the tests, thus we can reject the null hypothesis that the errors are not normally distributed. Also, for what concern the correlation between observed residuals and expected residuals under normality, the value is satisfactory to assume a strong correlation. The qq-plot shows that the data fit pretty well the diagonal line, exceptions made for a little deviation near the top. Lastly, to check the homoscedasticity we see the plots produced by the function plot(model). All the plots presents a satisfactory result, thus we can conclude that the assumption about homoscedasticity of the result is respected.

Let's now perform and interpret the result of the multiple linear regression.

```
mydata = read.csv("Cancer_final.csv")
data3 = na.omit(mydata)

#plot(AirPoll~Obesity, data=data3)
fit = lm( Cancer ~  Age + Smoking + GDP + Obesity + Alcool , data = data3)
fit
```

```
##
## Call:
## lm(formula = Cancer ~ Age + Smoking + GDP + Obesity + Alcool,
##     data = data3)
##
## Coefficients:
## (Intercept)          Age      Smoking          GDP      Obesity       Alcool
##   6.334e-03    5.387e-02   -3.101e-03    1.643e-05    4.417e-03    1.487e-03
```

```r
summary(fit)
```

```
##
## Call:
## lm(formula = Cancer ~ Age + Smoking + GDP + Obesity + Alcool,
##     data = data3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.81870 -0.14309 -0.02455  0.09149  2.68228
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.334e-03  2.640e-02   0.240  0.81041
## Age          5.387e-02  2.231e-03  24.144  < 2e-16 ***
## Smoking     -3.101e-03  1.399e-03  -2.217  0.02686 *
## GDP          1.643e-05  9.189e-07  17.884  < 2e-16 ***
## Obesity      4.417e-03  1.416e-03   3.120  0.00186 **
## Alcool       1.487e-03  3.283e-04   4.530 6.64e-06 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.3117 on 942 degrees of freedom
## Multiple R-squared:  0.8167, Adjusted R-squared:  0.8158
## F-statistic: 839.7 on 5 and 942 DF,  p-value: < 2.2e-16
```

From just these results, we can see various things: the value of the Adjusted R-Squared is satisfying, and then this model fits pretty well the data, telling us there is a quite good linear relationship. Also, all the variables seems to explain pretty well the number of cancer case: all of the p-values are statistically significant! The value that more explain our variable is the Age(not surprisingly I would say). On the other hand, if we try for example just to consider data coming from just one year, the results are pretty surprising

```r
data4 = subset(data3, Year == "2005")
fit2 = lm( Cancer ~ Alcool + GDP + Obesity + Age + Smoking + AirPoll, data = data4)
fit2
```

```
##
## Call:
## lm(formula = Cancer ~ Alcool + GDP + Obesity + Age + Smoking +
##     AirPoll, data = data4)
##
## Coefficients:
## (Intercept)       Alcool          GDP      Obesity          Age      Smoking
##   2.535e-01    1.691e-03    1.624e-05    2.328e-03    4.987e-02   -4.751e-03
##     AirPoll
##  -5.075e-03
```

```r
summary(fit2)
```

```
##
## Call:
## lm(formula = Cancer ~ Alcool + GDP + Obesity + Age + Smoking +
##     AirPoll, data = data4)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.70899 -0.15859 -0.02667  0.07779  2.56728
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.535e-01  1.156e-01   2.194   0.0300 *
## Alcool       1.691e-03  9.420e-04   1.795   0.0749 .
## GDP          1.624e-05  2.353e-06   6.902 1.90e-10 ***
## Obesity      2.328e-03  4.335e-03   0.537   0.5921
## Age          4.987e-02  6.652e-03   7.497 8.23e-12 ***
## Smoking     -4.751e-03  4.283e-03  -1.109   0.2693
## AirPoll     -5.075e-03  2.004e-03  -2.532   0.0125 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.3345 on 133 degrees of freedom
## Multiple R-squared:  0.8125, Adjusted R-squared:  0.804
## F-statistic: 96.03 on 6 and 133 DF,  p-value: < 2.2e-16
```

```
#fin = subset(mydata, select = c(Entity,Year,Cancer,Smoking))
#fan = na.omit(fin)
#fit2 = lm (Cancer ~ Smoking, data= fin)
#fit2
#summary(fit2)
#plot(Cancer ~ Smoking, data=fin)
#abline(fit2, col="blue", lwd=2)
```

In this case it seems that some of the variables do not correlate well! We can try the conclusion that the subset of the dataset we are considering (just one year) is not satisfying to take some satisfactory conclusion. This makes perfectly sense considering the context we are working on, since we should take in consideration also the time dependence. ‘Let’s take an extreme case as an example: if all of the population of a country starts to smoke instantaneously, the cancer cases for some years will remain quite constant, and if analyzing the data relating to just a short period of time, you would see that there is not a positive correlation between cancer and smoking, which does not make any sense! This test that we just carried out was just a demonstration of the fact that we should be very careful to draw conclusions instinctively.