Research Project: Describing Cancer

Abdel Sadek Karim, Tarricone Lorenzo
5/1/2022 ~ Bocconi University

1 - Introduction and research question

Life-Expectancy has constantly increased over the last centuries, and it has never been so high as nowadays. Medicine has taken great strides forward, eradicating and finding cures for a lot of mortal diseases. Having said that, there are still other big challenges that we have to face in medicine: one of those is for sure cancer. Cancer is a leading cause of death all over the world, accounting for almost 10 million deaths in 2020. This is the reason why we thought that would be very interesting to analyze the principal factors that, according to state-of-the-art scientific research, cause this disease. We also wanted to divide these factors into some macro-groups of variables: environmental, lifestyle-related, and socio-economical/demographic.

Our analysis (purely statistical) aims to verify if it is indeed possible to use data about the selected risk factors to predict the share of people with cancer in a given year and a given country. Moreover, we try to improve our all-inclusive model by finding the best model out of the above-mentioned factors.

2 - Dataset, Data Cleaning and Data Visualization

2.1 - Source

Our research started by searching out on some reliable websites what are the leading causes of cancer and dividing them into the macro-categories. To do that we adopted the WHO official website and we selected the following risk factors:

ENVIRONMENTAL: Air pollution
LIFESTYLE RELATED: Alcohol consumption, Tobacco consumption, Obesity
SOCIO-ECONOMICAL/DEMOGRAPHIC: GDP, Share of old people
(IMPORTANT DISCLAIMER: of course, the factors defined under "Socio-Economical/Demographic" cannot be considered to **cause** cancer, but as we will see

they turn out to be very useful to model cancer all over the world and overtime)

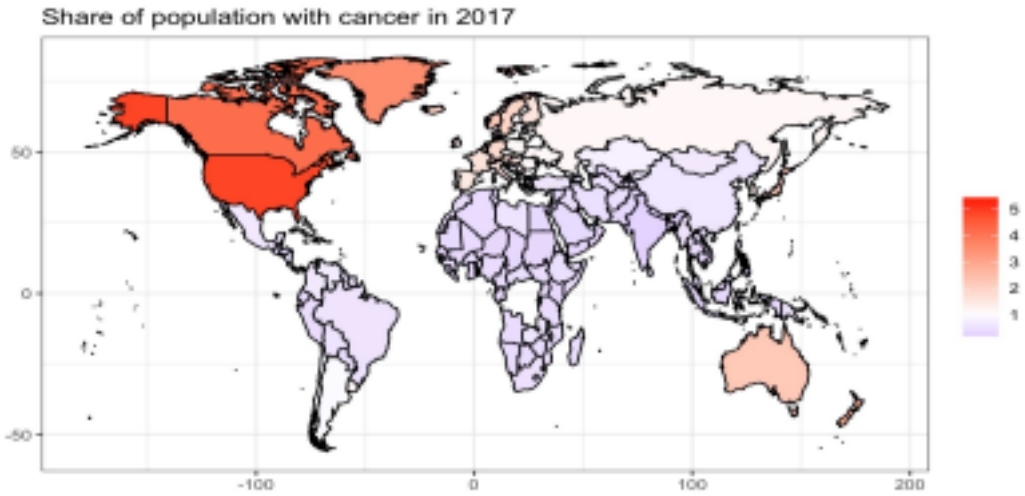
2.2 - Constructing the table

We started by loading all the different CSV files. Then we took the column we were interested in and we glued them together. After some other little manipulation (see the R code for that) the complete table with the N.A.s results in a dimension of 6468x11

2.3 - Data Visualization

Now we will proceed with some plotting. Again, for the detailed process for making these graphs please refer to the R script. You may also want to check our file in Html format, in which we made some nice animation to be able to display the evolution in time of the variables.

We start by displaying the variable we are interested in, that is the share of the population with any form of cancer. In this static image, we chose to plot the most recent year in our data frame, that is 2017. As we can clearly see from the plot below the countries which are suffering more (in percentage) of cancer are the USA, Canada, Australia and Greenland. These states are followed by European countries. Asian and African countries are the ones with relatively fewer cases of cancer in percentage. Here we could argue that a big downside of this picture is the fact that different countries have different abilities in the process of screening. As the WHO writes "Late-stage presentation and lack of access to diagnosis and treatment are common, particularly in low- and middle-income countries". We will further discuss this limitation in the dedicated section. (Countries for which data were not available are omitted from the map, but those are luckily few).



Now we are going to explore the dataset by plotting each selected factor against the Cancer one to start having an idea of some possible relation. Because we selected them as "risk factors" we would expect to see some positive correlation going on. To be consistent with what we have done above we again plot the graphs relative to the year 2017. We also used some libraries to make plots nicer by coloring the points according to the continent and by changing the size according to the population of the country. The result is the following:

Our hypothesis is confirmed: evidently, there's some sort of positive correlation, but still not strong as we imagined beforehand. We can see that age is the factor that presents a stronger correlation visually. The only problem is the one related to Air pollution. Our explanation for that is that will be provided in the section "Limitations of the study". We also have some outliers, among which we can find the US (The orange point that on each graph is above the others). The reasons behind that are complicated, but we can suppose that these two countries perform a more intense screening activity. The exact study of this phenomenon goes beyond the scope of this project. We will just delete these two countries from our data set. Having said that, the result that follows will be not significantly changed by this action.

3 - Multivariate Linear Regression

We are now going to perform a multivariate linear regression, trying to examine the relationship between some explanatory variables and one response variable. As mentioned before, we are going to take into consideration as explanatory variables all the factors selected to help us try to explain part of the variance of the share of people affected by cancer worldwide. The table used is considering approximately 150 countries for approximately 20 years. In this way, we will be able to do not miss some trend or to do not be deceived by some outliers years/countries or random trends.

Now, let's proceed with performing the actual multiple linear regression. After that, we want to see if the assumptions required to state the statistical significance are met: we'll check for the normality and the homoscedasticity of the residuals.

Analyzing each of the tests we performed above we see which conclusion we can take from what we observed. For what concerns normality of the residuals, we can see that the p-values of the tests we performed are below the 0.05 significance level for all of them, thus we can reject the null hypothesis that the errors are normally distributed. This is exactly as we imagined: the data set is made by more or less 1000 data points, thus violating this assumption was something that we expected. This is due to the fact that with this many data points it's easy for the testing methods to find evidence against normality. Trying to see the situation graphically, the setting seems a bit better. The QQ-plot shows that the data fit not so badly the diagonal line, exceptions made for a little

deviation near the top right. Furthermore, analyzing another graph we see that plotting the residuals against the fitted values approximates more or less a straight line along with zero. The analysis of the residuals is enough satisfying, but if we want to be strict, we will say that we are not able to take out any significant conclusions from the experiments we will carry on. We now perform and interpret the result of the multiple linear regression itself:

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.020e-01  4.013e-02   5.035 5.74e-07 ***
## Smoking      -2.872e-03  1.371e-03  -2.095 0.036418 *
## Obesity      2.647e-03  1.414e-03   1.871 0.061588 .
## GDP          1.666e-05  9.009e-07  18.497  < 2e-16 ***
## Age          5.045e-02  2.250e-03  22.419  < 2e-16 ***
## AirPoll      -4.296e-03  6.737e-04  -6.377 2.83e-10 ***
## Alcool       1.150e-03  3.259e-04   3.528 0.000439 ***
## ---
##
## Residual standard error: 0.3053 on 941 degrees of freedom
## Multiple R-squared:  0.8243, Adjusted R-squared:  0.8232
## F-statistic:   736 on 6 and 941 DF,  p-value: < 2.2e-16
```

From just this result, we can see various things: the value of the Adjusted R-Squared is very satisfying and this model fits pretty well the data, telling us there is a quite good relationship with the covariates we indicated: our factors explain an 82% variance in the share of the population with cancer! We are actually considering the Adjusted R-Squared since we are performing a multivariate linear regression, thus the number of factors we include in our model is relevant and we want to penalize bigger models. In fact, we know that the value of R-Squared can just increase when adding more factors, even if those are completely random ones. In addition, all the variables seem to explain pretty well the share of cancer cases: almost all of the p-values are statistically significant! The only exception is given by the "Obesity" explanatory variable.

4 - Model Selection (Testing)

Now that we discussed the linear model and the underlying characteristics, we want to do some basic model selection to see if there are other models (possibly simpler) that can explain the dependent variable well enough. To do that we will exploit two different approaches to model selection. The first is the one provided by testing and the other the one provided by information theory. We will compare the results and see whether they coincide or not. We start by importing the data we will need for the model selection and loading some useful libraries. We will then create our linear model using all the covariates and the dataset without the missing values (we will use the same model obtained above with the regression).

We begin by trying three different testing methods, called respectively "Step-up/Forward", "Step down/Backward" and "Step-both". The first acts in a dual way with respect to the second and the last is in a way able to overcome the downsides of the first two models (I'll explain later what I mean by that). Without entering too much into the details this greedy methodology starts with no(all) predictors in the model and iteratively adds (subtract) the one that is more statistically significant in contributing to the model. The procedure automatically stops when no other variables can be added to gain statistical significance. In practice what we do at each step is to perform an F-test (that reduces to a t-test) for each candidate variable and see which one has the lowest p-value (if there is one). In this setting, we kept a significance level of 0.05 both for

entering and for exiting the model. The Step-up method yields:

```
##                             Selection Summary
## --------------------------------------------------------------------------
##        Variable                     Adj.
## Step   Entered       R-Square       R-Square      C(p)          AIC     RMSE
## --------------------------------------------------------------------------
## 1      Age           0.7156         0.7153  579.6198       896.8510       0.3875
## 2      GDP           0.8108         0.8104  71.6307 512.4933       0.3162
## 3      AirPoll       0.8210         0.8204  19.0720 462.0425       0.3078
## 4      Alcool        0.8231         0.8224         9.5664  452.6010       0.3061
## --------------------------------------------------------------------------
```

The selection led us to add to the model the "Age", "GDP", "Airpoll" and "Alcool". We now try the symmetrical method. The "Step-down" method yields a specular result: we rejected the "Obesity" and the "Smoking" cofactor. Notice that even if the methods seem completely specular, it's not granted that they'll lead us to the same outcome.

To do one last cross-check using testing methods we use a mixed-method to allow variables that were added in previous steps to be excluded in succeeding steps (this wasn't allowed in the simple forward/step-up method). The "Step-both" method yields the same result like the one we got with the step-forward method. As well as before, the fact that we obtained the same result with all the three approaches is very likable. This seems to suggest a model of four variables that are cutting "Smoking" and "Obesity", given that these two risk factors are removed by all three of our greedy methods.

We now try all together different methods based on Information criteria. These work in such a way that they penalize big models that otherwise would be selected by methods such as LSE and MLE. The math underneath this approach is absolutely nontrivial and explaining it goes beyond the scope of this research. To implement this we will use the function "best_subset", which is returning the best model containing from one to six covariates. We, therefore, limit our comparison to the comparison between all the best subsets models that we can have.

This seems to suggest a different result. All of these criteria are suggesting that the best model is the one with all six explanatory variables except for the SBC method, which again is suggesting the model with four variables as before. This discrepancy can be explained by the fact that, at least for the Akaike Information Criteria, that we studied in class, the underlying constant C could dominate. (Notice also the difference in AIC is very low: $\Delta AIC = 2.59$). With small models AIC can easily prefer the complete model as the gain in "explained variance" obtained by adding a variable is more than the penalty introduced by that extra parameter. We conclude this section by saying that the best model after this analysis is the one with four explanatory variables, obtained with the three step-wise methods.

5 - Limitations of the study and Conclusions

Based on the results of our study, we can conclude that our last model explains quite well the variance in the data and that we collected about cancer and that some factors are more influential than others in explaining the share of people suffering from cancer in each country. Relatively older and richer countries (the two things often overlap) present many more cancer cases in percentage than the others and the same we can conclude of countries with consumption of alcohol and cigarettes above the mean.

However, our study has a lot of limitations that we must indicate. First, as we had to construct the data set on our own, we encountered some difficulties: we joined some data together into a table, but obviously, we had some missing data. It's not easy even just to collect the data in the same way for all the countries for more than 25 years for all those factors! Secondly, we have to consider the difference in data availability that is present among the countries. To give an example: it's much easier to find a data set about a given topic regarding the USA with respect to one regarding Burkina Faso. Besides that, we have taken into considerations just six factors to try to explain cancer incidence, but as we know it is a much more difficult research question: genetical heritage, other factors concerning lifestyle, other environmental elements, access to the healthcare system, prevention and treatment offered by the country and many more things play a role in developing cancer or not. Moreover, it is not an 'immediate' disease: factors of risk sum up over the years and so do progress in medicine. This ends up in a complicated and intricate situation to analyze. Diving into some details, we also had a quite surprising result: the correlation between cancer cases and air pollution seems strongly negative, as the plot at the beginning shows. It is difficult to explain such a result, but it may be that since the worst quality of air is on average in less developed countries (e.g. Nigeria, Pakistan, India…), it may be that other factors as the demographic distribution of the population or GDP, that typically obtain low values in those countries, have a stronger influence on the actual number of cancer cases, ending in this counter-intuitive result. The division that we made at the beginning seems to suggest that the "non-risk factors" are the ones better suited to describe the distribution of cases worldwide, but again that's what we wanted for this research: a merely descriptive approach. We conclude by saying that another interesting improvement to this research could be to try to repeat the same analysis but just focus on everyday habits and maybe by including the interaction effects within the possible covariates that we will choose in order to provide tips for a healthier routine.

Appendix

Sitography and References

The sources used for the data are:

- Cancer -> Global Burden of Disease Collaborative Network. Global Burden of Disease Study 2017 (GBD 2017) Results. Seattle, United States: Institute for Health Metrics and Evaluation (IHME), 2018
- Air_pollution -> World Development Indicators - World Bank
- Alcool -> Food and Agriculture Organization of the United Nations (FAO) (2017)
- GDP_per_capta -> World Development Indicators - World Bank
- Obesity -> World Health Organization (WHO)
- Old_age_dependency -> World Development Indicators - World Bank
- Smoking -> Nationally representative sources, survey data

Description of the variables

We built our dataset selecting by merging different CSV files all taken from the same source, that is "Our world in data". This is by itself collecting datasets from verified sources and the precise reference for each factor is given in the sitography at the end of the paper. Here we provide a complete description of the variable we decided to use in our model.

Cancer: Share of the total population with any form of cancer, measured as the age-standardized percentage. This share has been age-standardized assuming a constant age structure to compare prevalence between countries and through time.

Air_pollution: Population-weighted average level of exposure to concentrations of suspended particles measuring less than 2.5 microns in diameter (PM2.5). Exposure is measured in micrograms of PM2.5 per cubic meter ($\mu g/m^3$).

Alcohol: Average per capita consumption of alcoholic beverages, measured in kilograms per year. Data is based on per capita food supply at the consumer level but does not account for food waste at the consumer level

GDP per capita: Measured in constant international-$

Obesity: Obesity is defined as having a body-mass index (BMI) equal to or greater than 30. BMI is a person's weight in kilograms divided by his or her height in meters squared.

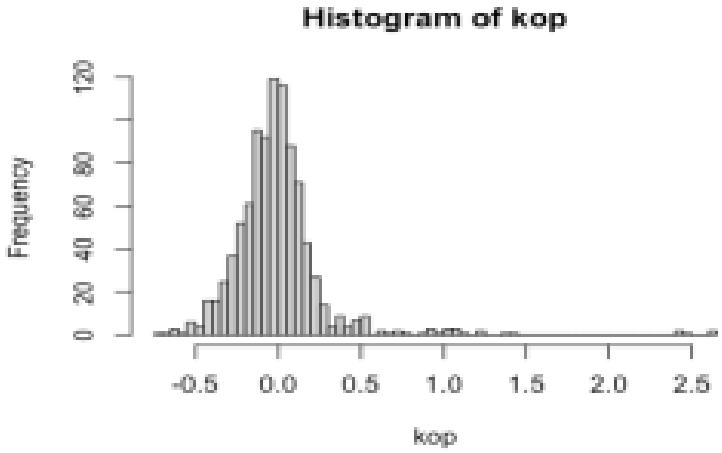
Old age Dependency: This is the ratio of the number of people older than 64 relative to the number of people in the working-age (15-64 years). Data are given as the proportion of dependents per 100 working-age population.

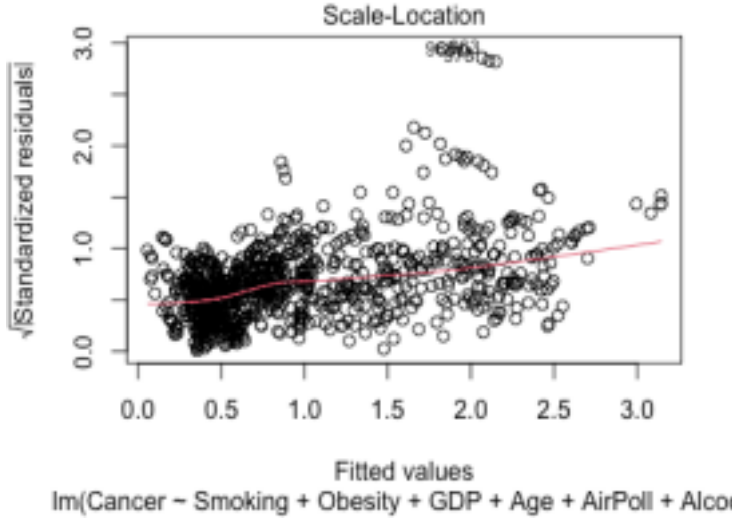
Smoking: Share of population who smoke every day.

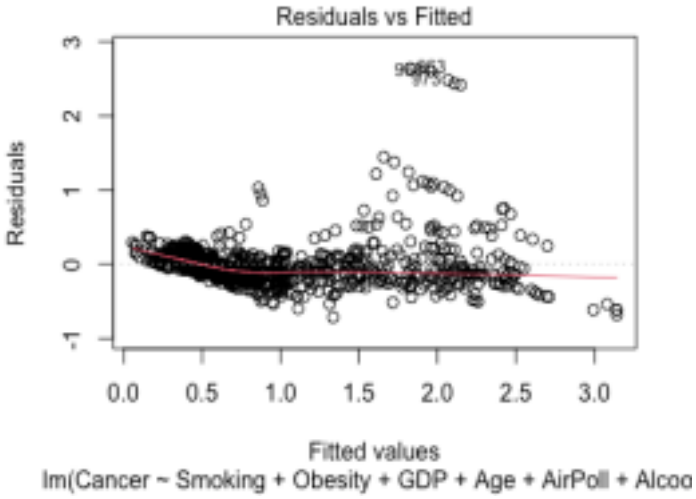
Images and missing code output (Linear regression)

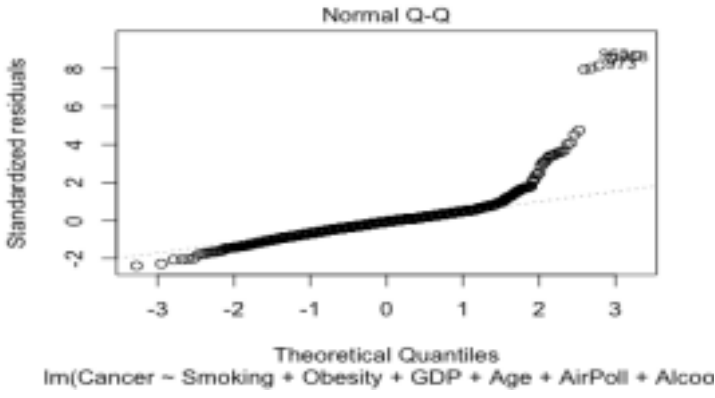
[BACK TO TEXT]

```
##.                 Normality tests
## ------------------------------------------------
##    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
## ------------------------------------------------
```









Missing code output (Model selection)

[BACK TO TEXT]

```
##                            Elimination Summary
## ----------------------------------------------------------------------
##        Variable                  Adj.
## Step   Removed      R-Square     R-Square   C(p)        AIC         RMSE
## ----------------------------------------------------------------------
##  1     Obesity      0.8237       0.8227     8.5024    451.5307    0.3057
##  2     Smoking      0.8231       0.8224     9.5664    452.6010    0.3061
## ----------------------------------------------------------------------
##
```

```
##                            Stepwise Selection Summary
## ------------------------------------------------------------------------------------------
##                      Added/                    Adj.
## Step   Variable     Removed      R-Square     R-Square     C(p)         AIC          RMSE
## ------------------------------------------------------------------------------------------
##  1     Age          addition     0.716        0.715       579.6200      896.8510     0.3875
##  2     GDP          addition     0.811        0.810        71.6310      512.4933     0.3162
##  3     AirPoll      addition     0.821        0.820        19.0720      462.0425     0.3078
##  4     Alcool       addition     0.823        0.822         9.5660      452.6010     0.3061
## ------------------------------------------------------------------------------------------
```

```
##             Best Subsets Regression
## ----------------------------------------------------
## Model Index   Predictors
## ----------------------------------------------------
##  1      Age
##  2      GDP Age
##  3      GDP Age AirPoll
##  4      GDP Age AirPoll Alcool
##  5      Smoking GDP Age AirPoll Alcool
##  6      Smoking Obesity GDP Age AirPoll Alcool
## ----------------------------------------------------
##
```

Subsets Regression Summary

Model	R-Square	Adj. R-Square	Pred R-Square	C(p)	AIC	SBIC	SBC	MSEP	FPE	HSP	APC
1	0.7156	0.7153	0.7142	579.6198	896.8510	-1795.2531	911.4140	142.3524	0.1505	2e-04	0.2856
2	0.8108	0.8104	0.8086	71.6307	512.4933	-2178.2111	531.9107	94.8042	0.1003	1e-04	0.1904
3	0.8210	0.8204	0.8188	19.0720	462.0425	-2228.3576	486.3143	89.7965	0.0951	1e-04	0.1806
4	0.8231	0.8224	0.8207	9.5664	452.6010	-2237.7019	481.7271	88.8134	0.0942	1e-04	0.1788
5	0.8237	0.8227	0.8209	8.5024	451.5307	-2238.7324	485.5112	88.6203	0.0941	1e-04	0.1786
6	0.8243	0.8232	0.8212	7.0000	450.0087	-2240.1947	488.8436	88.3856	0.0939	1e-04	0.1783

AIC: Akaike Information Criteria
SBIC: Sawa's Bayesian Information Criteria
SBC: Schwarz Bayesian Criteria
MSEP: Estimated error of prediction, assuming multivariate normality
FPE: Final Prediction Error
HSP: Hocking's Sp
APC: Amemiya Prediction Criteria