Introduction to Data Science – Project
New York University

I Introduction

In this project, we use the World Health Organization (WHO) dataset. The data is available here. It contains a set of social, economic, health, and political indicators for multiple countries. There are a lot of missing values, especially for the smaller countries. A quick fun visualisation to justify this :

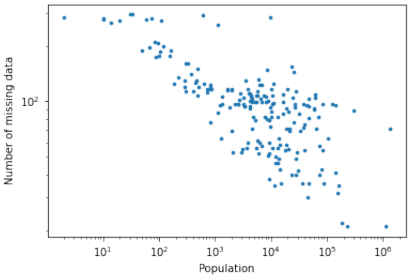
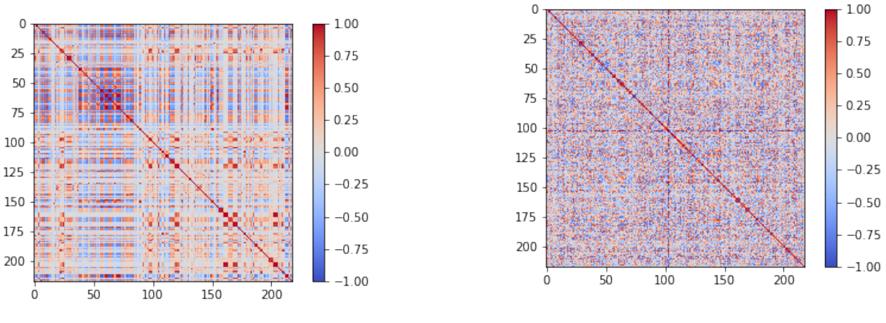


Figure 1: Missing data and population size (in thousands)

Note that the relationship is not linear (it's a log-log plot).

More specifically, the data consists of 202 rows and 357 columns. Each rows summarizes the data for a given country. We will only give the description for our processed data, as many columns will be dropped. The following method is used to process the data : countries with few data in total (< 200) (the smaller countries) are dropped. This leaves 168 entries. Then columns having less than 130 values are discarded. Finally, we are left with 168 rows and 221 columns.

Columns 1-3 summarises the country location (name, identification and continent respectively).
Columns 4-15 relates to demographic quantities.
Columns 16-87 describe health data.
Columns 88-209 are a mixture of purchasing power, expenditures and death causes.
Columns 210-221 are miscellaneous entries.
To impute for the missing values, the mean is used when necessary.



(a) Correlation matrix

(b) Partial correlation matrix (against all the other variables)

Figure 2: Correlations

II Question 1

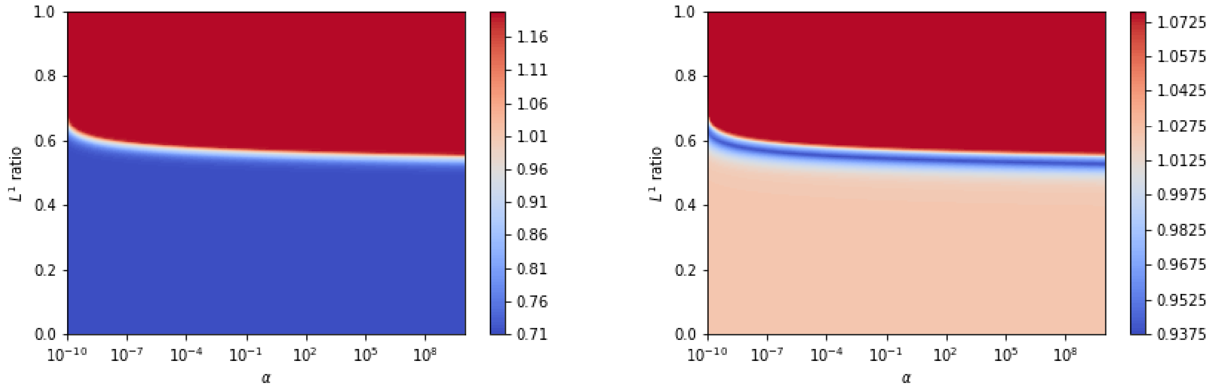
Situation. Assume that you are in the middle of a pandemic. There are signs of a food shortage and the government wants to know how to ration food. And for this objective, you are tasked to predict the population's annual growth rate.

Method. In this hypothetical situation, the intepretability of our model is very important: it is not acceptable to have model to go terribly wrong without us knowing it. Regression is a very safe and conservative method for this task.
To keep the model clean, we have to choose efficiently a handful of predictors. Good thing we computed the partial correlations in the introduction ! We just need to choose the most expressive components. We decide to keep those having a partial correlation (where we control for all other variables) greater than 0.8. This gives us 4 predictors in total but one was discarded due to lack of data. While the first 3 predictors makes logical sense, the 4th one seems quite surprising at first glance. At least, we can say that the first two predictors characterize countries in development, while the last two are symptomatic of developed countries, which also explains the sign of the correlation.

Component	Partial correlation	In reduced dataset
Population living below the poverty line	0.911	False
Registration coverage of births (%)	0.847	True
External resources for health as percentage of total expenditure on health	-0.820	True
Age-standardized mortality rate for cardiovascular diseases (per 100 000 population)	-0.802	True

Table 1: Components for regression

Results. For a simple linear regression, we obtain reasonable values: a train error of 0.712 and test error of 1.023. Had we used all the predictors, we would have obtained $9.057e-5$ for training but 300.914 for testing. For model selection, we do a simple grid search with an elastic net. It is very amusing that the optimal hyperparameters sit on a continuous curve. Interestingly, there is a single small region that concentrates the changes in performances, and this region's width seems constant with respect to the L^1 ratio. Finally, it is possible to obtain quite better performances than linear regression by tuning the hyperparameters. The best test results are given by : $\alpha = 7.75e7$, L^1 ratio $= 0.528$ for a train loss of 0.773 and a test loss of 0.939.



(a) Training loss

(b) Testing loss

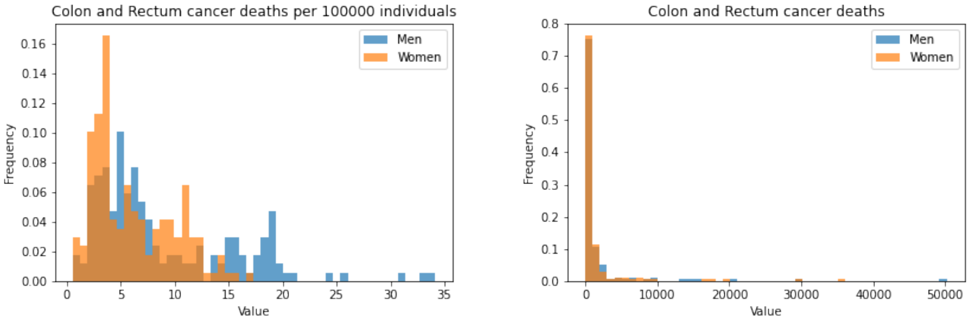
Figure 3: Grid searching for hyperparameters

III Question 2

Situation. Imagine you work for a multinational life insurance company. Per company policy, your prices are equivalent in all places the company is located. You want to adjust the pricing and policies according to gender (sadly for the consumer). For this reason, you task is to find out whether the distributions of death per causes differ for male and female demographics.

Method. For this problem, we need to do hypothesis testing. In our data, there are 4 relevant categories: colon and rectum cancer, liver cancer, lung cancer and stomach cancer. These categories are split between men / women and total number of deaths / deaths per 100000 men / women. The data is not gaussian at all, which means that we should not use a t-test. As the distributions looks roughly similar in shape, two tailed Mann-Whitney U tests will be used. The significance level is set at 0.05. Based on the preprocessing from the introduction (without filling missing values), only few countries have missing components in the categories of interest. Those countries are dropped from our dataset (row-wise).

Results. For illustrative purposes, we will plot the distributions for colorectal cancer deaths. The other graphs look similar.



(a) Deaths per 100000 individual

(b) Total number of deaths

Figure 4: Colon and rectum cancer deaths

	Colon and rectum	Liver	Lung	Stomach
Per 100000 individuals	9.27e-5	2.92e-13	4.39e-22	3.28e-11
Total deaths	0.359	3.58e-2	7.23e-6	3.33e-2

Table 2: p-values for across gender comparison

In the table 2, the p-values for causes of deaths per 100000 individuals are all under the significance level by a large margin so the the null hypothesis is rejected in those cases i.e. the distributions are different for these variables. However it is interesting to note that for the total deaths, the null is not rejected for colon and rectum cancer and are only marginally rejected in the liver and stomach cancer. To have more confidence, we could use some bootstrapping methods, but we did not do it here.

To push the analysis a bit further, let's see if the distributions across disease comes from an unique distribution. For this, we will do Kruskal-Wallis tests as non parametric equivalent of the ANOVA.

	Men	Women
Per 100000 individuals	3.93e-2	8.04e-2
Total deaths	1.48e-44	5.62e-48

Table 3: p-values for across disease comparison

These results show that the total deaths across disease are not likely to come from the same distribution. This means that it may be possible to segment (M/F) the market for liver, lung and stomach cancer (Fig. 2) if the company is pricing on the basis of total deaths. However, if the pricing needs to take into account the proportion of individuals compared to total population dying of a disease, segmenting by disease may be only relevant for men; at least with these results.

IV Question 3

Situation. Say you are a government official in charge of the hospitals and you are concerned with cancers in your continent. You want to know how a change of policy would affect those variables and need a strong model capable of being very accurate. As the situation is very dire, the interpretability although welcome, is secondary.

Methodology. For this problem, neural networks are designated : ”black box” models with high performances. The first step is to z-score the data, as we do not want exploding activations and especially make all components contribute regardless to their absolute value. The whole dataset is used for prediction since the accuracy is what matters. After separating the data into 80-20 for training and testing, it is possible to directly fit a neural network (Fig 5). We chose a ReLU network with a single 256 nodes hidden layer, Adam optimizer and a learning rate of 10^{-4} and as loss, the mean square error.

Results.

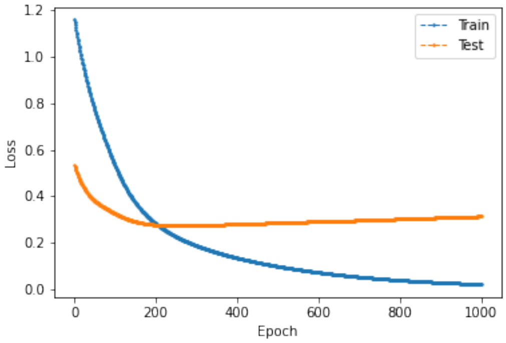
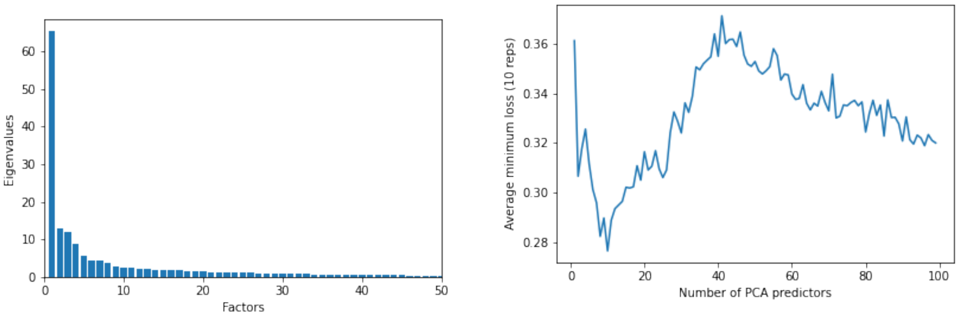


Figure 5: Losses for cancer mortality rate regression

The results are satisfactory; the best testing loss obtained was 0.279, which is much better than naively using the mean (luckily !) which would have on average a loss of 1.125. However, as interpretability would still be nice bonus, we would like to find a way to reduce the number of parameters. A PCA might be a good start, since we can pick the directions that capture the most variance, with those components still staying interpretable. The scree plot is very ”peaked” (Fig. 6a). This suggests that a model with only a few parameters should be able to obtain roughly the same accuracy, while having better interpretability. And indeed, this intuition is correct. Fig. 6b, was obtained by taking the average over 10 repetitions of the minimum testing loss for a set number of PCA components since gradient descent depends on the optimisation. This figure is very interesting, as the testing loss decreases, increases and then decreases again as the model size gets bigger. This phenomenon is called double descent. Note that the minimum loss is obtained for 10 PCA component, for a value of 0.277 which is even less than before !



(a) Scree plot

(b) Average minimum testing loss for different number of PCA components

Figure 6: Double descent

V Conclusions

In this project, we used the WHO dataset to show some demographic properties.
In the first part, we investigated the growth rate of the population. In order to regress this variable, we chose a few components by using partial correlations : registration coverage of births, external resources for health and age-standardized mortality rate for cardiovascular diseases. While we could just dismiss it as simple characteristics of in development / developed countries, this result is already interesting per se. High number cardiovascular diseases is representative of a lifestyle rather than specific countries. With a quick Google search, we find that there may be evidence of link between cardiovascular diseases and infertility[1]. But it is very surprising that it would account for this much in predicting the population growth.

In the second section, we studied the distributions of the deaths for different cancer diseases. We showed that at a significance level of 0.05, most death toll distributions for men and women were different, excepted for the colon and rectum one. Incidentally, this one is also more prevalent in the developed countries[2]. Does this mean we suffered from a selection bias by discarding the countries with fewer entries ? Perhaps. Or it might signify that the data is more accurate for this one. The fact that total death distribution is likely to come from the same distribution but the ratio is not, is surprising and is yet to be fully explained. Is it due to the fact that taking the proportion is heavily biased and has too many confounds (see spurious correlations of ratios) and thus the Mann Whitney U test is not relevant ?
Also, some p-values were marginal of rejection / acceptance. Bootstrapping may be of use to obtain confidence intervals. Those are all directions that could be explored for future work.

Finally, in the third section, we used a neural network to predict different cancer related variables. Surprisingly one of the difficulty faced, was the fact that the data was wider than the number of samples. In general this should not be an issue, but for PCA, it is a problem as the maximum number of principal components is limited by the number of countries, and indicators decomposed into countries is difficult to interpret (if even interpretable at all). This is why we had to discard a few more columns than initially planned by setting the missing data threshold higher.
Thanks to this, we observed – inadvertently – a very interesting, rather well-known phenomenon that is double descent.

As already pointed out, our modelling assumptions can be discussed. First discarding the data necessarily creates a bias and in our case we discarded the values for small and / or poor countries. Furthermore, the missing data was imputed by the mean, which is debatable as the distributions are not gaussian. Perhaps the median might have been more suitable. But then it will raise the question of whether z-scoring (when relevant) before or after imputing for those values. This would needs more thought.

All in all, we only barely scratched the surface of what this data can tell. Although it is a shame that so much data was missing, this data is already very rich. It would have been interesting to build more complex models, by clustering variable and countries. Many columns were not really used: democracy score, sugar per person, tax revenue... All of which could be an interesting topic of discussion.

[1] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5395549/
[2] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6791134/