SurvMeth 727 Term Paper

Chuzhu Zhong and Weining Xu

Contents

Introduction

In this project, we web scraped Wikipedia to capture the data for notable deaths occurred from 1997 to 2020. We also accessed the World Bank API built in R package `wbstats` to download the population, GDP, and unemployment rate data from 1997 to 2020. We then used ggplot2 package to plot the relationship between the number of death per million people in population/number of cases in each country and the year when he/she died; and the chance of dying from getting cancer per million people in population in each year. For further analysis, we conducted the linear regression analysis and compared the significance level for each variable in different countries. GitHub Repository: https://github.com/zczmz/727-Final-Project.git

Motivation and Research Questions

Cancer is a type of disease involving abnormal cell growth with the potential to invade or spread to other parts of the body. These contrast with benign tumors, which do not spread. Therefore, cancer is a chronic disease. It is customary to think that the stronger the country's economic power, the better the medical conditions the country obtains, leading to the smaller the proportion of people who die due to cancer. So we mainly want to explore the relationship between cancer-caused deaths and the economic power of the countries.

There will be two research questions:

1. How does the possibility of getting cancer for each year changes by countries? Do they have same patterns or not?
2. Which variable(s) have the largest relationship with getting cancer or not for different countries?

Setup

In the project, we use `rvest`, `dplyr`, `tidyr`, `stringr`, `useful`, `ggplot2` and other packages to webscrape, clean, preprocess, and implement the data into linear regressions.

Web Scraping from Wikipedia

We scraped the data from Wikipedia, using for loop to reach the recorded death website by looping year from 1997 to 2020, and month from January to December.

```
data = vector("list",12)
mlist = c("January","February","March","April","May","June","July","August",
```

```
      "September","October","November","December")

for (y in 1997:2020){
  for (m in 1:12){
    site = read_html(paste0("https://en.wikipedia.org/wiki/Deaths_in_",mlist[m],"_",y))
    fnames = html_nodes(site,"#mw-content-text h3+ ul li")
    text = html_text(fnames)

    temp<-data.frame(text, stringsAsFactors = FALSE)
    temp$month <- m
    temp$year <- y
    data = rbind(data,temp)
    }
 }
```

Data cleaning

Since the extracted data is not in the version of sentences separated by commas, we first separated the string into three parts (the name of the decease age, and his/her nationality and other texts).

```
ndata<-data%>%
  separate(text,c("name","age","nationality"),sep=",")
```

For further extractions of country and death cause, we then extracted the first two words from the column named nationality for country, and created a new column by extracting everything after the last comma for cause. The rationale of extracting everything after the last comma is we noticed that death causes are typically written at the end of each string.

```
ndata$country <- word(ndata$nationality, 1, 3)
ndata$a <- sub('.*\\,', '', data$text)
```

Because countries for some observations are missing and we noticed that countries are written as the first character being capitalized, we dropped all observations that do not have the first character as capitalized.

```
ndata<- ndata[!upper.case(ndata$country),]
```

We then dropped NAs for the column named country as well.

```
ndata<- ndata[!is.na(ndata$country),]
```

We continued cleaning the column named country by separating it into three columns by space because some countries have two words separated by space (ie. Sri Lankan) while some have only one word (ie. German).

```
ndata<-ndata%>%
  separate(country,c("blank", "c","dummy"),sep=" ")
```

We then mutated a new column containing the information on whether the third separated word has the first character being capitalized.

```
ndata$ifupper<-str_detect(ndata$dummy, "[:upper:].*")
```

Then, we combined the previously separated country into one column named country if the first character is being capitalized. Otherwise, we assigned the second separated word into the new column country.

```
ndata$country_dummy<-paste(ndata$c, ndata$dummy, sep = " ")
ndata$country<-if_else(ndata$ifupper==FALSE,ndata$c,ndata$country_dummy)
```

Now the country is extracted and ready for analysis. However, we still need to extract cause from string. We first cleaned the data by deleting brackets at the end of each string in column named a and mutated the

cleaned data into a new column named cause.

```
ndata$causes <- sub("\\..*", "", ndata$a)
ndata$cause<-sub("\\[.*", "", ndata$causes)
```

Since we observed that all death causes have the first character as not being capitalized, we dropped all observations that start from a capitalized character.

```
ndata_1<- ndata[!str_detect(ndata$cause, "^*[:upper:].*$"), ]
```

However, there are special cases that the death cause is written in abbreviation formats (ie. AIDS). For the completeness of our dataset, we extracted observations with all capitalized characters into a new dataset, named tempdata. Deleting NAs, we recombined the tempdata with the original dataset, named finaldata.

```
tempdata<-ndata[upper.case(ndata$cause),]
tempdata = tempdata[(which(nchar(tempdata$cause) != 2)),]
tempdata <- tempdata[is.na(as.numeric(as.character(tempdata$cause))),]
tempdata<-tempdata[-2,]

finaldata<-rbind(ndata_1,tempdata)
```

Another special case is COVID-19, which is neither starts with a not-capitalized character, nor having all characters being capitalized. Hence, again, we extracted a new dataset for COVID-19 and combined this with finaldata.

```
coviddata<-ndata[ndata$cause==" COVID-19",]
finaldata<-rbind(ndata_1,coviddata)
```

However, looking at the dataset, we found that in many cases, the column named cause contains other information such as works, which often contains "and". Thus, we dropped all observations with "and", dropped NAs, and finally deleted all useless columns.

At the end, the dataset we extracted from wikipedia using web scrapping has age, country, month, year, and cause.

```
finaldata<-finaldata[!grepl("and", finaldata$cause),]
finaldata <- finaldata[!is.na(as.numeric(as.character(finaldata$age))),]

finaldata<-finaldata[,c("age","country","month","year","cause")]
```

Grouping

We regularized death causes into fewer categories and dropped NAs.

```
head(finaldata)
```

	age	country	month	year	cause	group_cause
3	51	Italian	1	1997	colon cancer	cancer
4	59	Canadian	1	1997	lymphoma	cancer
6	25	English	1	1997	traffic accident	crash
10	52	American	1	1997	cardiac arrythmia	heart disease
12	45	American	1	1997	drowned	drowned
13	82	American	1	1997	heart attack	heart disease

Data from the World BankAPI `wbstats`

We are also curious about the relationship in demographics of each country versus proportions of each death cause. To obtain data on population, unemployment rate, and GDP on the year of death, we used the

wbstats package. The wbstats data can be extracted by the same range year as the finaldata has, which is from 1997 to 2020.

```
library(wbstats)
my_indicators <- c("pop"="SP.POP.TOTL","gdp"="NY.GDP.MKTP.CD","unemp"="SL.UEM.TOTL.ZS")
wbdata<-wb_data(my_indicators, start_date = 1997, end_date = 2020)
wbdata$year<-wbdata$date
wbdata<-wbdata[,-c(1,2,4)]
head(wbdata)
```

country	gdp	unemp	pop	year
Aruba	1531944134	NA	85450	1997
Aruba	1665100559	NA	87280	1998
Aruba	1722798883	NA	89009	1999
Aruba	1873452514	NA	90866	2000
Aruba	1920111732	NA	92892	2001
Aruba	1941340782	NA	94992	2002

Before combining two datasets, we first needed to make adjustments for country by regularizing nationality to country names.

Combining data for analysis

For the purpose of analysis, we dropped observations with less than 50 observations for each country and less than 50 observations for each cause.

At the end, the final version of the cleaned data is named ffdata, which is now ready for analysis.

Data Visualization and Analysis

The following plot shows the relationship between the number of death per million people in population/number of cases in each country and the year when he/she died.

The barplot is the exact number of notable people in the dataset, and the red line shows the changes in the proportion of the the number of deaths regardless of causes for each year by countries. We noticed that patterns for each country is different, where some countries have large changes over time, while some remains similar.

However, the proportion is not on the same scale for each country, so to verify whether proportions change significantly for different countries, we plotted the line plot and put proportions on the same scale.

Now we can see a clearer pattern that for countries such as China, India, and Nepal, the line is nearly flat and horizontal. However, for other countries such as United Kingdom, United States, and New Zealand, the line is less flat with many changes. Thus, it confirms that our findings on the unequal proportions across different countries.

To look at what the relationships look like for each countries and the get an answer for research question 1, we first regrouped the group_cause into either “cancer” or “other”. We then added another line indicating the proportion of having cancer per million people for different years by country to the line plot.

```
#the chance of getting cancer and gdp
pdata1<-inner_join(wdata,ffdata)
```

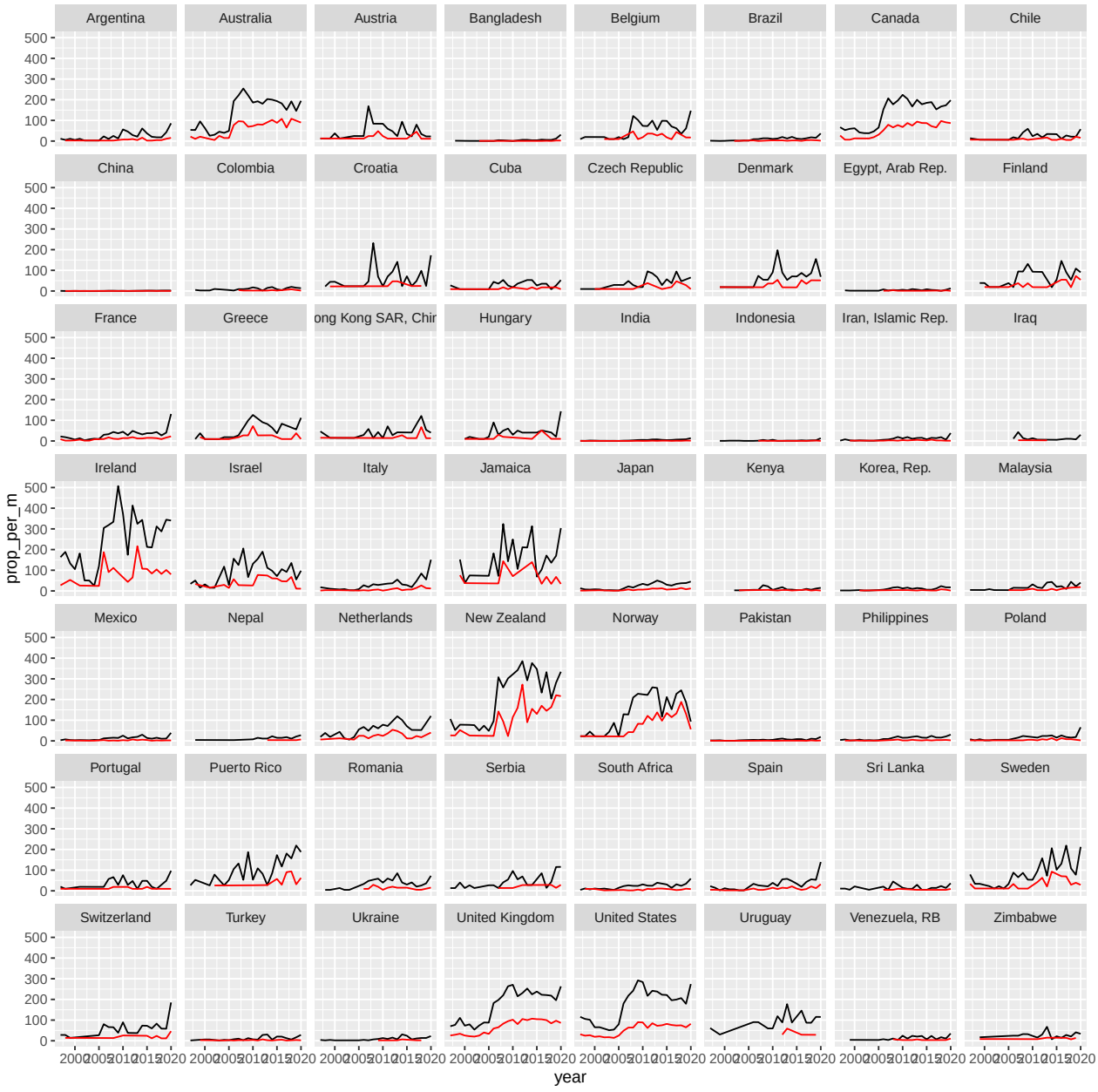
```
## Joining, by = c("country", "year", "pop")
```

```
pdata<-pdata1
pdata$group_cause<- as.factor(ifelse(pdata$group_cause == "cancer", "cancer", "other"))
```

```
cdata<-pdata%>%
  drop_na(pop)%>%
  filter(group_cause=="cancer")%>%
  group_by(country,year,pop, gdp)%>%
  summarise(total=n())%>%
  mutate(cancer_prop_per_m=total/pop*1000000*100)
```

```
## `summarise()` has grouped output by 'country', 'year', 'pop'. You can override using the `.groups` ar
```

```
tcplot+geom_line(data=cdata,aes(x=year,y=cancer_prop_per_m),color="red")+facet_wrap(.~country)
```



The red line is the proportion of people dying due to cancer and the black line is the proportion of people dying due to all reasons.

Compared to the previous line plot and the pattern in this plot itself, we found some different patterns to

answer the research question 1.

There are some countries having the same pattern for the proportion of dying due to total causes (ie. both cancer and other) and the proportion of dying due to cancer.

Among those countries with different patterns, the findings for each country are listed as follow:

For Argentina, since 2016, the red line gets flatter and increases less compared to the black line. It indicates that cancer is getting a less severe cause for deaths when being compared to other causes.

Looking at the red line itself, the cancer neither became more severe not less severe due to its lack of fluctuation.

For Australia, compared to the black line, the more space between two lines since 2007 indicates the less comparative severity than other causes of death.

Looking at the red line itself, it skyrocketed in 2007 and stayed approximately same until 2020, indicating that cancer is gradually becoming a larger cause of leading people's death since 2007 in Australia.

For Austria, compared to the black line, there are less space between two lines, indicating that cancer is getting more severe comparatively when comparing to other causes.

Looking at the red line itself, there are two points that the red line increases and drops afterwards, around 2008 and around 2017. It indicates that in these two years, cancer was a more severe cause comparing to the cancer stands in other years.

For Belgium, the red line is more flat than the black line, meaning that the cancer is gradually a less severe reason of causing death compared to other causes of deaths since 2015.

Looking at the red line itself, the flat pattern itself also indicates that cancer is not getting more severe.

For Canada, the gradually more space between two lines since 2005 indicates that the cancer is increasingly leading to less deaths, compared to other causes.

Looking at the red line itself, cancer itself gets more severe because of the great increase in 2005, while after 2005, the severity of cancer itself does not change much due to the flat pattern.

For Croatia, compared to the black line, the red line does not change much when the black line fluctuates greatly. In around 2008, the difference between two lines is huge, and then intersect with each other immediately. This pattern happens multiple times in 2013 and 2015. It indicates that cancer, when being compared to other causes of death, is fluctuating without a clear pattern of becoming more or less severe comparatively.

Looking at the red line itself, it does not increase a lot and nearly flat, meaning that cancer neither gets more severe nor less severe.

For Denmark, the red line does not change much when the black line increases greatly in 2006-2007 when H1N1 dominated. Obviously, Denmark are mainly suffered from flu, but not really dominated by cancer.

Looking at the red line itself, it increases as the pattern of the all causes, which means that cancer neither gets more severe nor less severe, but the proportion of getting a cancer is increasing as time pass by.

For Finland, in around 2005, cancer is not a severe cause of deaths compared to other causes because there is no difference between two lines. However, in around 2007, the difference between two lines suddenly increases and then decreases to zero in around 2015, indicating that cancer became a severe cause from 2007 to 2015. After 2015, the difference between two lines increases and then decreases again. It indicates that cancer was a severe cause of death compared to other causes from 2007 to 2015. After this comparative severity drops in 2015, cancer became more severe again after 2016.

Looking at the red line itself, the line remains its flat pattern until 2017. After 2017, the red line drops first and then increased, indicating that cancer is becoming more severe since 2018.

For France, the red line increased around 2018 as the black line skyrocketed at the same year, leading to a big difference between lines. However, the slope for the red line is clearly smaller than that for the black line, indicating that the comparative severity of cancer has been decreased since 2018.

Looking at the red line itself, the slight increase in around 2018 indicates that cancer is getting more severe than previous years.

For Greece, the difference between two lines is larger since 2006, even though the difference decreases again in 2016, it increases right after 2016, which indicates that cancer is getting less severe compared to other causes gradually since 2006.

Looking at the red line itself, the line was basically flat, with a increase in 2010 and decrease in 2018. Thus, we conclude that cancer neither becomes more severe not less severe overall.

For Hungary, the difference between two lines increased in around 2006 and continues this pattern until 2016. Afterwards, the difference was nearly zero until the great increase in difference in 2019. The pattern indicates that cancer is getting more severe compared to other causes since 2006 and less severe compared to other causes since 2016.

Looking at the red line itself, it increases in around 2015, even though drops later, we still believe that cancer is getting more severe than before.

For Ireland, in around 2004, the difference between two lines decreases to zero, while increases right after. The difference between two lines is even larger since 2017, compared to the difference in previous years. It indicates that cancer is getting less severe, compared to other causes.

Looking at the red line itself, the red line first increased in around 2004, and then drops in around 2012. This pattern repeated once more in 2013, while the line does not drop to the point as it does in 2012. Thus, we conclude that cancer itself is getting more severe.

For Israel, comparing the red line to the black line, the black line increases and drops more. In around 2002, cancer was not a very serious concern for causing death, while when it comes to 2015, cancer nearly became the only reason for causing death because two lines intersect with each other. Since 2018, the difference between the black line and red line increases, indicating that cancer is getting less severe compared to other causes.

Looking at the red line itself, it first increases gradually and drops in 2018, indicating that cancer is also getting less severe, compared to previous years.

For Italy, cancer was a severe cause of deaths until 2017, for the difference between two lines is small. After 2017, the black line increases a lot while the red line slightly drops. leading to the increase in difference, which indicates that cancer is gradually getting less severe.

Looking at the red line itself, the flat pattern indicates that cancer is neither getting more severe nor less severe.

For Jamaica, before around 2015, the red line has the similar pattern as the black line, which leads to nearly no changes in difference between two lines. However, after 2015, the red line drops when the black line increases, indicating that cancer is getting less severe than other causes.

Looking at the red line itself, the line increased greatly in 2007 and drops greatly in 2015. Hence, cancer became more severe from 2007 to 2015, while it gets less severe after 2015.

For Netherlands, the difference between red line and the black line increases in around 2004. After 2004, the difference does not change much. Thus, cancer became less severe compared to other causes since 2004, and this comparative severity does not change much.

Looking at the red line itself, the line is basically flat without many patterns, indicating the severity of cancer itself remains.

For New Zealand, before around 2006, the comparative severity of cancer did not changed much, because the space between two lines are about the same from 1997 to 2006. From 2006 to 2018, the comparative

severity of cancer became less severe compared to that before 2006, because the space between two lines are large. After 2016, the red line increases faster than the black line, which helps explaining the increase in severity as well. Since 2019, the red line drops again when the black line increases, indicating that cancer is getting less severe comparing to other causes.

Looking at the red line itself, the line continues to increase, indicating that cancer is more and more severe.

For Norway, the difference between two lines increased greatly in around 2004, indicating that cancer became more severe, compared to other causes. The large difference maintains until 2014, where the two lines intersect with each other. After 2014, the difference increases, but not as large as it does in 2004, indicating that the comparative severity of cancer increases after 2014.

Looking at the red line itself, the line increases until the first large decrease in 2018, indicating that cancer was a more severe cause of death until 2018, while became less severe after that.

For Portugal, the difference between two lines increases and then decreases several times, in 2008, 2010, 2012, and 2014. In around 2017, the difference increases greatly compared to previous increases, indicating that the comparative severity of cancer decreases since 2017.

Looking at the red line itself, the flat pattern indicates that cancer is neither getting more severe nor less severe.

For Puerto Rico, the difference between two lines is getting larger since the intersection of two lines in 2014, meaning that the cancer has become a less severe cause of death since 2014, compared to other causes.

Looking at the red line itself, the increase also starts from 2014, which indicates that cancer is gradually becoming more severe.

For Serbia, the red line stays flat while the black line fluctuates a lot in 15 years, meaning that cancer is neither getting more severe nor less severe.

Looking at the red line itself, the flat pattern indicates that cancer is neither getting more severe nor less severe.

For Spain, the difference between two lines greatly increases in 2016, while the difference does not change much before 2016. It indicates that cancer is less severe than other causes of death.

Looking at the red line itself, the flat pattern indicates that cancer is neither getting more severe nor less severe.

For Sweden, the difference between two lines is getting larger since 2005, indicating the less comparative severity for cancer.

Looking at the red line itself, the line keeps increasing until the drop in around 2018. It indicates that cancer has gradually been a severe cause of death until the decrease in 2018.

For Switzerland, the difference between two lines increases in 2005, then decreases to zero in 2010, with a increase right after 2010. Since 2013, the difference increases again and continues getting larger. It indicates that cancer is becoming a less severe cause of death than other causes since 2013.

Looking at the red line itself, the line is basically flat, except a slight increase in 2019. It indicates that cancer started to become more severe since 2019.

For United Kingdom, the difference between two lines gets larger since 2005, and remains its difference until 2019. It indicates that the comparative severity of cancer decreased in 2005, and the severity maintained until 2019 when this comparative severity decreased again.

Looking at the red line, the line gradually increases until 2010 and maintains its flat pattern until 2020. It indicates that cancer had become more severe until 2010, and did not became more severe or less severe after 2010.

For United States, the difference between two lines gets larger in 2005, and remains its difference until 2020. It indicates that the comparative severity of cancer decreased in 2005, and this severity maintained until 2020.

Looking at the red line itself, the line gradually increases until 2009 and maintains its flat pattern until 2020. It indicates that cancer had become more severe until 2009, and did not became more severe or less severe after 2010.

In summary, we analyzed the comparative severity of cancer by comparing the red line (the percentage of per million population dying from cancer) to the black line (the percentage of per million population dying from all causes we coded), and the severity of cancer itself by looking at the red line (the percentage of per million population dying from cancer).

As a result, we found that the comparative severity of cancer has no relationship with the severity of cancer itself.

Based on our visual analysis, some countries have decreasing comparative severity of cancer when the severity of cancer itself remains, such as Argentina, Belgium, Greece, Italy, Netherlands, Portugal, Spain, United Kingdom, and United States.

There are also some countries with the comparative severity decreasing, while the severity of cancer itself is increasing. These countries include: Australia, Canada, France, Hungary, Ireland, New Zealand, Puerto Rico, Sweden, and Switzerland.

For some countries, both the comparative severity and the severity of cancer itself increase, including Austria, Finland, and Norway.

For some countries, both the comparative severity and the severity of cancer itself decrease, including Israel and Jamaica.

It is hard to determine whether Croatia has more comparative severity of cancer, because the black line increases and decreases a lot, which creates a pattern of fluctuation.

Linear regression analysis

To find answers for the second research question, we conducted the linear regression including age, year, month, population, GDP, unemployment rate, and the proportion of people dying from cancer per million people as predictors.

First, we set age as numeric, dropped all NAs, and assigned “cancer” into 1 and “other” into 0.

```
ccdata<-inner_join(cdata,ffdata)
```

```
## Joining, by = c("country", "year", "pop", "gdp")
```

```
ccdata$age<-as.numeric(ccdata$age)
ccdata<-drop_na(ccdata)
ccdata$group_cause_n<-ifelse(ccdata$group_cause == "cancer",1, 0)
```

After readjusting the dataset, we created a for loop for finding predictors that are 95% significant for interpreting whether a person dies from cancer or not.

For Austria, Bangladesh, Canada, Colombia, Croatia, Cuba, Czech Republic, Denmark, Egypt, Finland, Hong Kong SAR, Iraq, Jamaica, Kenya, South Korea, Malaysia, Nepal, Netherlands, Norway, Pakistan, Portugal, Serbia, Sri Lanka, Switzerland, Turkey, Uruguay, and Venezuela, none of these predictors are significant for determining whether a person dies from cancer or not.

For Argentina, only month has a significant relationship with either a person dies from cancer or not.

For Australia, only age has a significant relationship with either a person dies from cancer or not.

For Belgium, only the proportion of people dying from cancer per million people has a significant relationship with either a person dies from cancer or not.

For Brazil, month and the proportion of people dying from cancer per million people have significant relationship with either a person dies from cancer or not.

For Chile, month, unemployment rate, and the proportion of people dying from cancer per million people have significant relationship with either a person dies from cancer or not.

For China, only the proportion of people dying from cancer per million people has significant relationship with either a person dies from cancer or not.

For France, age, year, and population have significant relationship with either a person dies from cancer or not.

For Greece, only the proportion of people dying from cancer per million people has significant relationship with either a person dies from cancer or not.

For Hungary, only the proportion of people dying from cancer per million people has significant relationship with either a person dies from cancer or not.

For India, unemployment rate and the proportion of people dying from cancer per million people have significant relationship with either a person dies from cancer or not.

For Indonesia, only the proportion of people dying from cancer per million people has significant relationship with either a person dies from cancer or not.

For Iran, only the proportion of people dying from cancer per million people has significant relationship with either a person dies from cancer or not.

For Ireland, age and the proportion of people dying from cancer per million people have significant relationship with either a person dies from cancer or not.

For Israel, only the proportion of people dying from cancer per million people has significant relationship with either a person dies from cancer or not.

For Italy, age, year, and the proportion of people dying from cancer per million people have significant relationship with either a person dies from cancer or not.

For Japan, age and the proportion of people dying from cancer per million people have significant relationship with either a person dies from cancer or not.

For Mexico, age, unemployment rate, and the proportion of people dying from cancer per million people have significant relationship with either a person dies from cancer or not.

For New Zealand, age and the proportion of people dying from cancer per million people have significant relationship with either a person dies from cancer or not.

For Philippines, only the proportion of people dying from cancer per million people has significant relationship with either a person dies from cancer or not.

For Poland, only the proportion of people dying from cancer per million people has significant relationship with either a person dies from cancer or not.

For Puerto Rico, only age has significant relationship with either a person dies from cancer or not.

For Romania, only the proportion of people dying from cancer per million people has significant relationship with either a person dies from cancer or not.

For South Africa, age, year, population, and the proportion of people dying from cancer per million people have significant relationship with either a person dies from cancer or not.

For Spain, only month has significant relationship with either a person dies from cancer or not.

For Sweden, only the proportion of people dying from cancer per million people has significant relationship with either a person dies from cancer or not.

For Ukraine, only age has significant relationship with either a person dies from cancer or not.

For United Kingdom, only the proportion of people dying from cancer per million people has significant relationship with either a person dies from cancer or not.

For United States, age, year, GDP and unemployment rate have significant relationship with either a person dies from cancer or not.

For Zimbabwe, only age has significant relationship with either a person dies from cancer or not.

Among all 56 models, age is significant for 12 models, year is significant for 4 models, month is significant for 4 models, population is significant for 2 models, GDP is significant for 1 model, unemployment rate is significant for 4 models, and the proportion of people dying from cancer per million people is significant for 21 models.

Conclusion

In this study, we first using scrapping to extract data on notable deaths from 1997 to 2020 from Wikipedia, followed by data cleaning. We then imported another dataset using westats package that directly connect to the World Bank API for getting data on population, GDP, and unemployment rate for countries from 1997 to 2020. After regularizing and grouping levels, we combined the two dataset into one for further analysis on visualization and linear regression.

From the ggplot, we first explored how data looks by plotting the relationship between the proportion of deaths by all causes per million people and year by countries. We then focused on cancer, to resolve the first research question, and visualized another plot for the relationship between the proportion of deaths due to cancer per million people and year by countries. For this plot, we first analyzed the comparison between the red line (the percentage of per million population dying from cancer) and the black line (the percentage of per million population dying from all causes we coded) to understand the comparative severity of cancer. We also looked at the trend of the red line itself to understand cancer's severity by itself. As a result, we found that the comparative severity of cancer has no relationship with the severity of cancer itself. There countries that the has increasing severity of cancer itself while the comparative severity of cancer is decreasing. There are also countries with decreasing comparative severity when the severity of cancer itself is neither increasing nor decreasing.

We then conducted the linear regression analysis for each country using age, year, month, population, GDP, unemployment rate, and the proportion of people dying from cancer per million people as predictors. From this result, we found that the proportion of people dying from cancer per million people is the most effective predictor for interpreting whether a person dies from cancer or not, for it is significant for 21 out of 56 models with all coefficients being positive. It indicates that as more people in a country dies from cancer, a person who resides in this country will also experience higher possibility of dying due to cancer. The second effective predictor is age, which is significant for 12 models with all coefficients being positive. It means that as people get older or die when they are older, the possibility of them dying from cancer gets higher as well. There are also other less effective predictors, including year, month, and unemployment rate.

Based on our findings, we found that year is effective for not extremely poor countries, such as United States, South Africa, Italy, and France. Unemployment rate is effective for countries with relatively large population and less resources, such as Mexico, India, and Chile. Further researches can be done on this topic of finding whether there are relationships between dying from cancer and socioeconomic index with more extensive data.

Appendix on ML model training

We are curious about whether we can predict one's death cause based on country demographics, and his/her age. The attributes we are using are: age, pop, unemployment rate, GDP, caner_prop_per_m. The outcome

variable is group_cause_n: 0 - not cancer, 1 - cancer

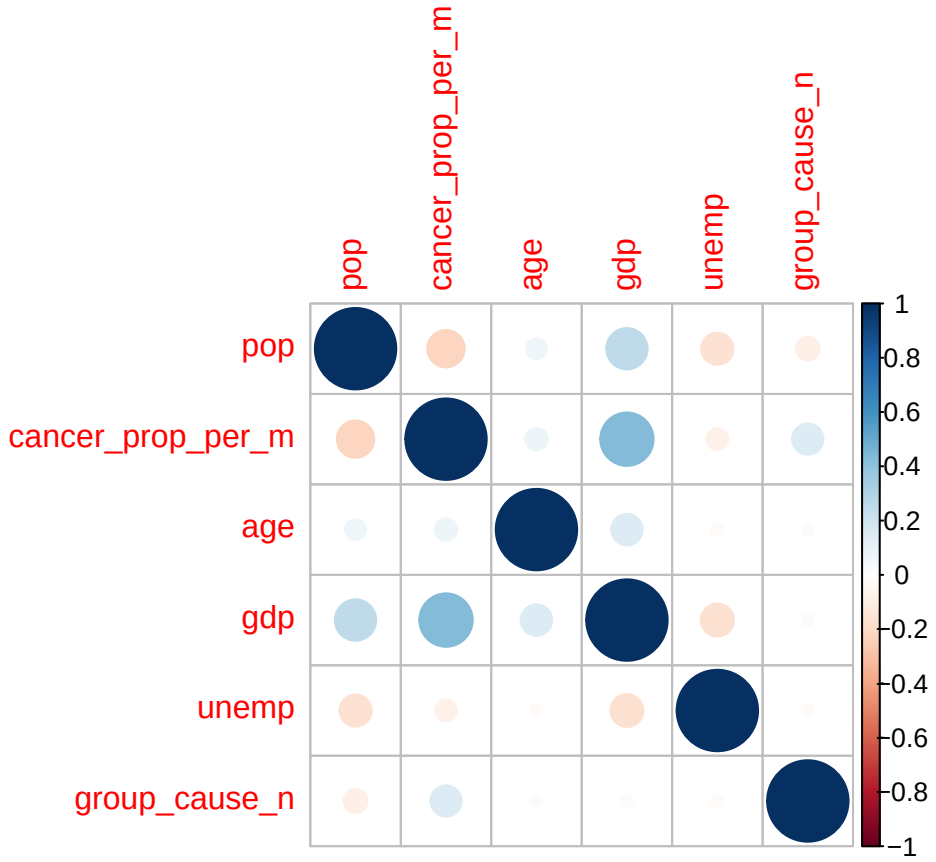
```
attribute <- c("pop", "cancer_prop_per_m", "age", "gdp", "unemp", "group_cause_n")
mldata <- ccdata[,attribute]
head(mldata)
```

pop	cancer_prop_per_m	age	gdp	unemp	group_cause_n
36063451	2.772890	77	298948250000	12.65	1
36063451	2.772890	74	298948250000	12.65	0
36870796	2.712174	77	284203750000	15.00	0
36870796	2.712174	54	284203750000	15.00	1
37275644	2.682717	52	268696750000	17.32	0
37275644	2.682717	23	268696750000	17.32	0

```
table(mldata$group_cause_n)
```

```
##
##     0     1
## 17804  8087
```

```
corrplot::corrplot(cor(mldata[, 1:ncol(mldata)]))
```




```
mldata$group_cause_n <- as.factor(mldata$group_cause_n)
```

From the correlation plot, we can see how each attributes are related among others. GDp and cancer_prop_per_m are highly correlated. Group_cause_n is positively correlated with cancer_prop_per_m, and group_cause_n is negatively correlated with populations of countries.

Train and test split

We first splitted the data into train and test set, using seed=123. Factorized and set levels for "group_cause_n" attributes to not cancer and cancer.

```
set.seed(123)

mldata$group_cause_n <- as.factor(mldata$group_cause_n)
train <- sample(1:nrow(mldata), 0.8*nrow(mldata))
c_train <- mldata[train,]
c_test <- mldata[-train,]
levels(c_train$group_cause_n) <-c("not cancer", "cancer")
```

Decision Tree

First, we applyed Decision Trees to train the model.

Bagging via caret

Then we implemented Bagging. The `train()` function of the `caret` package can be used to call a variety of supervised learning methods and also offers a number of evaluation approaches. For this, we first specify our evaluation method.

```
ctrl  <- trainControl(method = "cv",
                      number = 5)
```

Now we could call `train()`, along with the specification of the model and the evaluation method. Return the cross-validation results.

```
cbag <- train(make.names(group_cause_n) ~ .,
              data = c_train,
              method = "treebag",
              trControl = ctrl)
```

Random Forests

In order to also use random forests for our prediction task, we first specified a set of try-out values for model tuning. For random forest, we primarily had to care about `mtry`, i.e. the number of features to sample at each split point.

```
ncols <- ncol(c_train)
mtrys <- expand.grid(mtry = c(sqrt(ncols)-1,sqrt(ncols),sqrt(ncols)+1))
```

This object could be passed on to `train()`, along with the specification of the model, and the tuning and prediction method. Calling the random forest object lists the results of the tuning process.

```
rf <- train(make.names(group_cause_n) ~ .,
            data = c_train,
            method = "rf",
            trControl = ctrl,
            tuneGrid = mtrys)
```

AdaBoost

In order to build a set of prediction models it is helpful to follow the `caret` workflow and first decide how to conduct model tuning. Here we use 5-Fold Cross-Validation, mainly to keep computation time to a minimum. `caret` offers many performance metrics, however, they are stored in different functions that need to be combined first.

Now we specified the `trainControl` object.

```
evalStats <- function(...) c(twoClassSummary(...),
                             defaultSummary(...),
                             mnLogLoss(...))
```

```
ctrl  <- trainControl(method = "cv",
                      number = 5,
                      summaryFunction = evalStats,
                      #verboseIter = TRUE,
                      classProbs = TRUE)
```

As a first method we tried out AdaBoost as implemented in the `fastAdaboost` package. Specifically, Adaboost.M1 would be used with three try-out values for the number of iterations.

```
grid <- expand.grid(nIter = c(50, 100, 150),
                    method = "Adaboost.M1")
```

Now we passed these two objects on to `train`, along with the specification of the model and the method, i.e. `adaboost`. List the results of the tuning process.

```
#set.seed(744)
#levels(c_train$is_safe) <-c("not safe", "safe")
ada <- train(make.names(group_cause_n) ~.,
             data = c_train,
             method = "adaboost",
             trControl = ctrl,
             tuneGrid = grid,
             metric = "ROC")
```

GBM

For Gradient Boosting as implemented by the `gbm` package, we had to take care of a number of tuning parameters. Now the `expand.grid` is helpful as it creates an object with all possible combinations of our try-out values.

```
grid <- expand.grid(interaction.depth = 1:3,
                    n.trees = c(500, 750, 1000),
                    shrinkage = c(0.05, 0.01),
                    n.minobsinnode = 10)
```

List the tuning grid…

```
grid
```

interaction.depth	n.trees	shrinkage	n.minobsinnode
1	500	0.05	10
2	500	0.05	10
3	500	0.05	10
1	750	0.05	10
2	750	0.05	10
3	750	0.05	10
1	1000	0.05	10
2	1000	0.05	10
3	1000	0.05	10
1	500	0.01	10
2	500	0.01	10

interaction.depth	n.trees	shrinkage	n.minobsinnode
3	500	0.01	10
1	750	0.01	10
2	750	0.01	10
3	750	0.01	10
1	1000	0.01	10
2	1000	0.01	10
3	1000	0.01	10

...and begin the tuning process.

```
gbm <- train(make.names(group_cause_n) ~.,
             data = c_train,
             method = "gbm",
             trControl = ctrl,
             tuneGrid = grid,
             metric = "ROC",
             distribution = "bernoulli",
             verbose = FALSE)
```

Logistic regression

Finally we also added a logistic regression model. Obviously we had no tuning parameter here.

```
set.seed(744)
logit <- train(make.names(group_cause_n) ~.,
             data = c_train,
             method = "glm",
             trControl = ctrl)
```

We might want to take a glimpse at the regression results.

```
summary(logit)
```

```
##
## Call:
## NULL
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0013  -1.3837   0.7724   0.8980   1.7330
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         1.105e+00  7.399e-02  14.935  < 2e-16 ***
## pop                 4.730e-10  7.193e-11   6.575 4.86e-11 ***
## cancer_prop_per_m -8.600e-03  4.894e-04 -17.573  < 2e-16 ***
## age               -2.009e-03  8.641e-04  -2.325 0.020061 *
## gdp                 6.511e-15  2.442e-15   2.666 0.007677 **
## unemp               1.847e-02  4.823e-03   3.829 0.000129 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 25684  on 20711  degrees of freedom
## Residual deviance: 25133  on 20706  degrees of freedom
## AIC: 25145
##
## Number of Fisher Scoring iterations: 4
```

Prediction and Performance

Finally, we predicted the outcome in the test set.

Given predicted class membership, we could use the function `postResample` in order to get a short summary of each models' performance in the test set.

```
## [1] "Accuracy predicted by Decision Trees: 0.682950376520564"
```

```
## [1] "Accuracy predicted by Bagging: 0.640857308360687"
```

```
## [1] "Accuracy predicted by Random Forest: 0.685267426144043"
```

```
## [1] "Accuracy predicted by Adaboosting: 0.645298320139023"
```

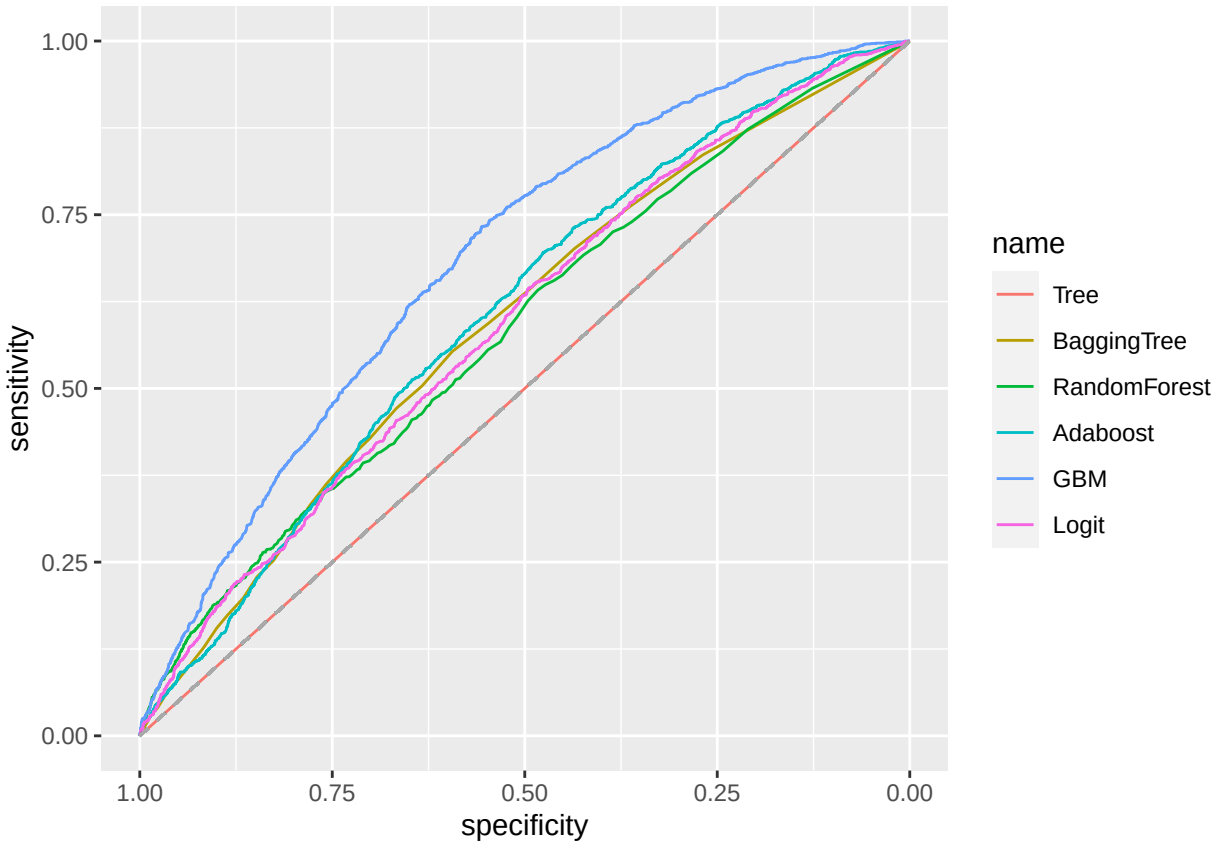
```
## [1] "Accuracy predicted by XGboosting: 0.689708437922379"
```

```
## [1] "Accuracy predicted by Logistic Regreesion: 0.68449507626955"
```

ROC Curve

Creating `ROC` objects based on predicted probabilities using `pROC` package and plotting the ROC curves.

```
ggroc(list(Tree = tree_roc,
           BaggingTree = cbag_roc,
           RandomForest = rf_roc,
           Adaboost = ada_roc,
           GBM = gbm_roc,
           #CART = cart_roc,
           Logit = logit_roc)) +
  geom_segment(aes(x = 1, xend = 0, y = 0, yend = 1),
               color="darkgrey", linetype="dashed")
```



Based on the accuracy score, we could see that all model performed evenly. We could also observe the same result from ROC curve that XGboosing is slightly better than other models.