Suicides in Italy - A comparative analysis among macro regions

Pietro Romozzi and Federico Colombo-Ercole

11 novembre 2014

Research questions

The paper will focus on the relation between suicidal behaviours (an extreme consequence of mental disorders) and various factors concerning climate. In particular, we want to test if the following hypothesis hold true:

- Suicide rate in a certain Italian region increases when the average temperature is low.
- Suicide rate increases when precipitations happen to be frequent.
- Suicide rate increases when GDP per capita is high.
- Suicide rate increases when economic inequality represented by the gini index.

Methodology

Sources and data gathering

For the scope of this work, regional data were necessary; in particular we looked for data about weather (i.e. temperature and precipitations), GDP per capita and inequalities (i.e. gini index). Fortunately, most of the data required were founded in the Italian National Institute of Statistics (ISTAT) database, or in other databases of ISTAT-related agencies. Accessing to the ISTAT database was fairly easy especially because data were clear and well structured. Nevertheless importing data in R, where we intended to carry out our analysis, was more problematic. One necessary step was to export relevant data as .cvs files from the ISTAT website, and then proceed to import them in R using the *read.csv* command. Data from the ISTAT database were not tidy though, as in tidy data:

1. Each variable forms a column.
2. Each observation forms a row.
3. Each type of observational unit forms a table.

Therefore, we have to put our data trough a tidying process, that is structuring our datasets to facilitate analysis (Wickham 2014).

To do so we partly used excel's functions, since the data got from ISTAT were not so many and fairly manageable. Then we concluded our tidying process in R, using more complex functions which Excel does not allow to use easily. What follows is the final part of the tidying process we did on R. As the reader could see, at first we cleaned data for suicides, then we merged them with other data we have cleaned and merged together before, in R as well.

```
# Loading and tidying data for suicides.

suicides_rough <- read.csv("RoughData_suicides.csv", header = T, sep =";", dec = ",", fill = T)
```

```r
suicides_rough$macro_area <- as.character(suicides_rough$macro_area)

# Replacing wrong-named observation with new ones, analogous to observations in other data used.

suicides_rough$macro_area[suicides_rough$macro_area == "northeast"] <- "north"
suicides_rough$macro_area[suicides_rough$macro_area == "northwest"] <- "north"
suicides_rough$macro_area[suicides_rough$macro_area == "islands"] <- "south"

suicides_rough$suicides <- as.numeric(suicides_rough$suicides)

# Then we aggregated data by macro areas.
# For example we sum northeast and northwest to obtain total suicides for north.

suicides_clean <- dplyr::group_by(suicides_rough, macro_area, year)
suicides_clean <- dplyr::summarise(suicides_clean, std_suicides_rate = mean(suicides))

# Merging all data in the final dataset.
# MergedData3 contains NA, MergedData4 does not.

MergedData3 <- merge(x = MergedData2, y = suicides_clean, union("macro_area", "year"), all = T)

MergedData4 <- merge(x = MergedData2, y = suicides_clean, union("macro_area", "year"), all = F)
```

In the end, after an accurate process of data cleaning, we came up with the following dataframe:

macro_area	year	avg_precipitations	avg_temperature	gdp_pc	gini_index	std_suicides_rate
centre	2006	659	13.63	28100	0.2800	6.30
centre	2007	588	13.83	28900	0.2670	6.30
centre	2008	931	13.69	28700	0.2730	6.50
centre	2009	876	13.51	28000	0.2670	6.60
north	2006	640	10.16	30550	0.2655	7.30
north	2007	644	10.59	31600	0.2600	7.30
north	2008	975	10.30	31800	0.2625	7.50
north	2009	880	10.28	30150	0.2655	7.65
south	2006	658	16.01	17050	0.3010	5.55
south	2007	637	16.19	17600	0.2910	5.70
south	2008	642	16.27	17700	0.2960	5.90
south	2009	827	16.18	17200	0.2960	5.80

Data presented above regard:

- Data about climate obtained from CRA-CMA (Department for meteorology applied to agriculture, ISTAT) regarding annual average temperatures, measured in Celsius degrees, and average annual precipitations, measured in millimiters for all years between 2005 and 2009.
- Data about GDP measured in Euros per capita, found in the ISTAT database as well.

- Data about the Gini Index downloaded from the Istat website. Again, data for the three macroregions we are interested in were available. As expected, Gini's index is slightly higher in Southern Regions compared to the Centre and the North.
- Data about suicide rate, measured as number of suicides on 100000 inhabitants.

Unfortunately, as will be explained in the next chapters, after we have carried out our analysis, we found a result problematic to interpret. Conviced that a possible explanation, might reside in too aggregated data, we opted for running our linear model on regional data instead of grouped-by-macro-area ones. It required the construction of another dataframe, whose data were found in the ISTAT database too, and that we processed directly in Excel since we decided to gather data just for 2009. Further developments of our study will try to include a bigger number of observation concerning italian regions or even a panel data analysis.

We believed that a linear model was an effective way to analyze our dataframe because it could give a clear idea of the impact of the selected indipendent variables on suicides which we chose as to be the dependent variable. We also opted for a semplifications of the model, obtained by removing particularly unsignificant variables. Further information about our analysis and about the interpretation of our model, will be discussed in the following sections. In the meantime, we provide the code we used in R to create our complete and simplified model.

```
# Creating linear models to analyze our dataframe.

M1 <- lm(std_suicides_rate ~ gdp_pc + avg_temperature + avg_precipitations + gini_index , data = MergedD

M2 <- lm(std_suicides_rate ~ gdp_pc + avg_temperature + avg_precipitations , data = MergedData4)
```

New model estimation

As we said, we applied this model also to a second dataframe to investigate whether possible to obtain different results using data for regions insted of macro areas. Therefore, what follow are the codes we used to build the complete and the simplified linear model for the second dataframe we used.

```
# Loading a new dataframe trying to use the same model in less aggregated data.

Dataset2009 <- read.csv("TidyDataset2009.csv", header = TRUE, sep = ";", dec = ",", fill = TRUE)
```

We decided to take into account all of the Italian Regions for a total of twenty case studies. Data were obtained from the Istat database. The model remained the same, as we want to test if variations in annual tempreratures, annual precipitations, gini coefficient and levels of GDP per capita can explain variations in suicide rates across Italian Regions. We opted for a multiple linear regression model because of time constraints and thus we selected the most recent gathered data referring to year 2009.

The estimated model looks like this:

$$SuicideRate = \alpha + \beta_1 Temperature + \beta_2 Precipitations + \beta_3 Ginicoefficient + \beta_4 GDPperCap + e$$

Firstly, we upoaded a new CSV file, containing data for all Regions.

```
Dataset2009 <- read.csv("TidyDataset2009.csv", header = TRUE, sep = ";", dec = ",", fill = TRUE)
```

We called LM1 our first model, and by using the lm command, we includeded our dependent and independent variables.

```
LM1 <- lm(std_suicides_rate ~ gdp_pc + avg_precipitations + avg_temperature + gini_index , data = Datas
```

Here there results of the estimation:

% latex table generated in R 3.1.2 by xtable 1.7-4 package % Wed Nov 12 20:49:46 2014

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	11.93	9.76	1.22	0.24
gdp_pc	-0.00	0.00	-0.10	0.92
avg_precipitations	0.00	0.00	0.14	0.89
avg_temperature	-0.27	0.19	-1.41	0.18
gini_index	-8.75	22.71	-0.39	0.71

By looking at the summary table, it is clear that none of the variables included in this model is statistically significant, which means that the model will have to be changed profoundly. Anyway, variables Gdp per capita and gini coefficient seem to be unexpectedly less significant compared to variable temperature: the result of the t test for temperature is much better compared to the results of GDP per capita and Gini coefficient. Since the goal of this short research is exactly that of finding out whether a relationship between climate and suicidial behavour exist, it is interesting to notice that a variable for climate is more significant compared to other socio-econmic indicators, at least in Italy. On the other hand, GDP per capita and Gini coefficient were meant to be our control variables, but they did not succeeded in improving our estimations, given the data we gathered.

Consequently, we run the following commands, which will show simple descriptive statistics operations, in order to interpret our model results:

- summary(LM1)
- coefficients(LM1)
- plot(LM1)
- confint(LM1)
- fitted(LM1)
- residuals(LM1)
- anova(LM1)
- vcov(LM1)
- influence(LM1)

After that, the coefficients command was lunched in order to see if the independent variables have a positive or a negative effect on suicide rate.

```
coefficients(LM1)
```

```
##        (Intercept)             gdp_pc avg_precipitations
##       1.192838e+01      -1.261299e-05       4.619775e-04
##    avg_temperature         gini_index
##      -2.736562e-01      -8.745631e+00
```

A table with the coefficient of all the explanatory variables is obtained,from which it is easy to understand, by simply looking at the coefficient sign, if a specific explanatory variable has a positive or a negative impact on suicide rate. As expected, annual average temperature and Gdp per capita have a negative impact on suicide rate and also annual precipitations seem to have a positive influence on suicides. On the contrart, variable gini has a negative coefficient, which is not what could be expected by theory. In fact, it would be easier to think that unjust societies were wealth is distributed unequally are those experiencing higher suicide

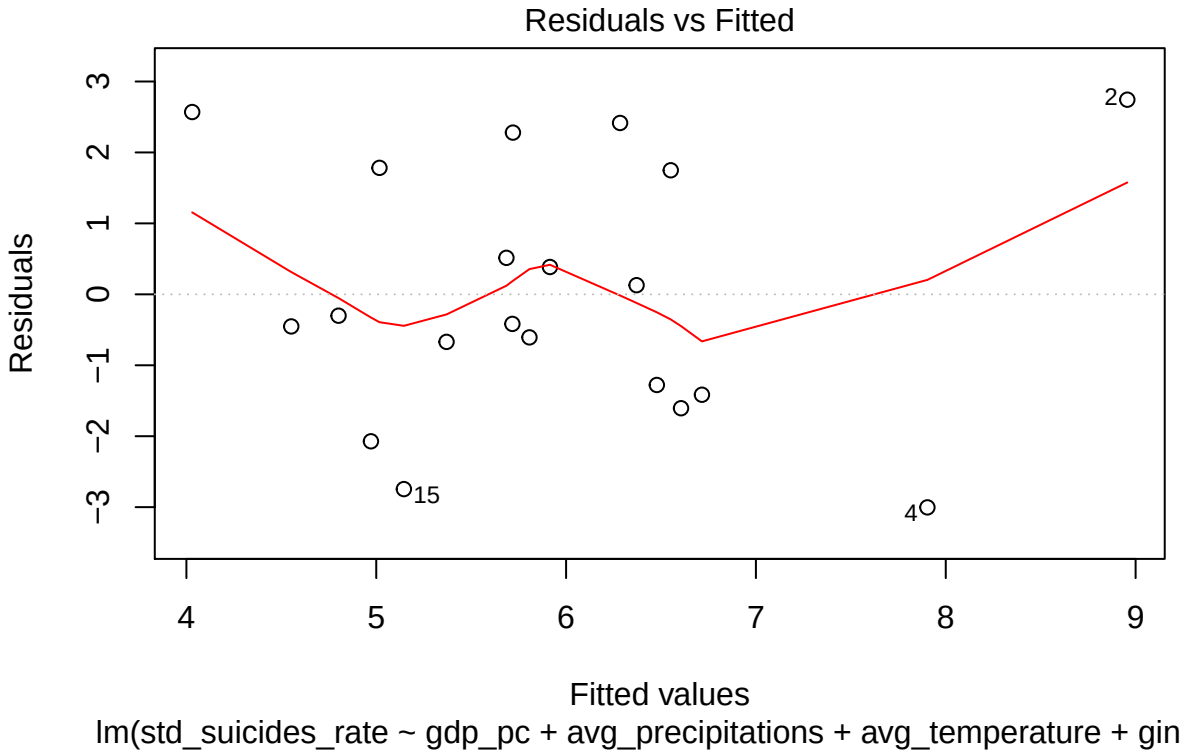
rates. According to the coefficients table, it is not like this in the Italian case. Also by simply looking at the statistics for standardized suicide rate every 100.000 inhabitants, it becomes clear that suicide rates are lower in Southern Italy, where gini coefficient figures are higher, compared to the North, where income is more equally distributed. However, all coefficents are extremely small and since all variables of our model are insignificant, this is meant to be a simple excercise in preparation for the final paper.

Through the anova table variance of our explanatory variables is shown. In combination with command plot(), it is a way to diagnose heteroscedasticity. Errors seem to be fairly normally distributed, but model also suffers from heteroscedasticity, at it can be seen in the Anova table and the residuals/Fitted scatter plot. In fact, variance is not equally distributed throughout the values, but this may be because of the not very large number of observations, so that few outliers manage to modify the results.

```
anova(LM1)
```

```
## Analysis of Variance Table
##
## Response: std_suicides_rate
##                     Df Sum Sq Mean Sq F value  Pr(>F)
## gdp_pc               1 14.785 14.7852  3.6850 0.07413 .
## avg_precipitations   1  0.416  0.4160  0.1037 0.75189
## avg_temperature      1  8.962  8.9623  2.2337 0.15577
## gini_index           1  0.595  0.5949  0.1483 0.70560
## Residuals           15 60.184  4.0122
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
plot(LM1, which=1)
```



Command confint(LM1) is used to calculate the confidence intervals for the coefficients and provide a measure of precision for linear regression coefficient estimates.

```
confint(LM1)
```

```
##                           2.5 %       97.5 %
## (Intercept)         -8.878534178 3.273530e+01
## gdp_pc              -0.000287531 2.623051e-04
## avg_precipitations  -0.006734317 7.658272e-03
## avg_temperature     -0.686107532 1.387952e-01
## gini_index         -57.156140446 3.966488e+01
```

Finally, we also computed the fitted values for our model, the residuals and the variance-covariance table of the main parameters of the fitted model.

```
fitted(LM1)
```

```
##        1        2        3        4        5        6        7        8
## 6.715697 8.956425 6.605571 7.903323 6.371264 6.551352 5.720030 5.915044
##        9       10       11       12       13       14       15       16
## 5.717036 6.284486 5.807077 5.370282 6.478045 5.685625 5.145600 4.972061
##       17       18       19       20
## 5.017007 4.801409 4.030383 4.552283
```

```
residuals(LM1)
```

```
##          1          2          3          4          5          6
## -1.4156974  2.7435745 -1.6055709 -3.0033230  0.1287361  1.7486484
##          7          8          9         10         11         12
##  2.2799697  0.3849562 -0.4170363  2.4155137 -0.6070774 -0.6702822
##         13         14         15         16         17         18
## -1.2780448  0.5143753 -2.7455996 -2.0720611  1.7829933 -0.3014089
##         19         20
##  2.5696170 -0.4522826
```

```
vcov(LM1)
```

```
##                      (Intercept)        gdp_pc avg_precipitations
## (Intercept)         9.529396e+01 -9.910104e-04      -1.159545e-02
## gdp_pc             -9.910104e-04  1.663629e-08       5.003870e-08
## avg_precipitations -1.159545e-02  5.003870e-08       1.139903e-05
## avg_temperature    -7.532384e-01  1.634446e-05       8.921024e-05
## gini_index         -1.694606e+02  1.088957e-03      -3.014456e-03
##                    avg_temperature    gini_index
## (Intercept)          -7.532384e-01 -1.694606e+02
## gdp_pc                1.634446e-05  1.088957e-03
## avg_precipitations    8.921024e-05 -3.014456e-03
## avg_temperature       3.744513e-02 -6.998060e-01
## gini_index           -6.998060e-01  5.158568e+02
```

Limitations

As it has been shown, the explanatory variables of our model are statistically insignificant. For the final paper, it is recommandable to substitute variable for gini index, which has been proved to be particularly

neglegible for our estimation, with other variables that could better explain variation of suicide rates across Italy. In particular, self-rated health, a subjective indicator that assesses health status, unemployment rate and urbanisation rate are factors we had already started considering when setting up the model, but which were not included so far, as we expected variables Gdp per capita and Gini coefficient to be better control variables. At this point, it becomes crucial to take them into consideration once again. Nevertheless, we do not want to drop variables annual average temperature and annual precipitations even if both have resulted to be insignificant after having conducted some tests on our model. The reasons for this are two: firstly, variable average temperature is the most significant variable of our model and maybe it really has an impact on suicidial behaviours in Italy. Apart from that, we originally decided to estimate this model, as we wanted to find out more regarding the relationship between wheater conditions and suicides, even though we knew we could have encountered problems by estimating such a model.

In conclusion, the model will be adjusted thanks to new control variables, without giving up the original goal of our research, namely test the relationship between suicide rates and climate factors.

###Summary

References:

Wickham, Hadley. 2014. "Tidy Data." *Journal of Statistical Software* 59 (2014).