Analysis on the abortion ratio in Europe and the socio-economic variables affecting it

Annamaria Culpo

annamaria.culpo@studenti.unitn.it

Github repository: https://github.com/AnnaCulpo/Abortion_in_Europe

I. Introduction

In June 2022 the US Supreme Court overturned the Roe v Wade ruling, allowing to every state to ban or severely restrict the possibility for pregnant women to get an abortion before 12 weeks[1]. This event has generated several protests among women and feminists and it has reignited the debate in the public opinion. In order to better understand the Italian and European situation on this field, I have decided to undertake at first an exploratory analysis about some of the main variables which could influence a woman's decision to become pregnant or not. Then I have applied linear and non-linear regression methods to evaluate the importance of the variables selected in the prediction of the abortion ratio.

II. Literature review and research question

Nowadays almost all the European countries consider abortion a legal practice if undertaken before the 12th week of pregnancy. The only exceptions are Malta (which bans abortion in all cases), Vatican City, Liechtenstein, Andorra and Poland, where just some particular cases linked to the early age of the mother, her health situation or raping are allowed to access to the abortion practice[2].

The European Journal of Public Health has published research about the induced abortion in Denmark[3]. The authors focused their attention on the socio-economic situation and on the country of birth of women who decided to have an induced abortion. The results highlighted the strong association between their decision and some variables like the marital status, the age of the mother, the number of children and the economic balance. I have taken inspiration from this study and I have analysed the variable importance of some socio-economic factors on the abortion ratio across European countries. The objective of my research is not simply to analyse the abortion situation in Europe. Rather, I would like to suggest that understanding the reasons behind a woman's choice to get an abortion could help governments to implement concrete measures to increase the fertility rate; and the results would probably be more effective than the introduction of anachronistic limitations of the abortion right.

III. Methodology

A. Data and Cleaning

Data has been collected mainly from Eurostat website, from the section "Population and social conditions"[4]. As for the abortion ratio in 2018, other two websites have been useful: "Abort Report.eu"[5], hold by Exelgyn, a global pharmaceutical company specialising in women's healthcare, and "Johnston's Archive"[6]. The variables, contained in different datasets, have been aggregated in a unique, comprehensive one. In addition to the name of the European countries and their alpha-2 codes, the variables present in the final dataset are:

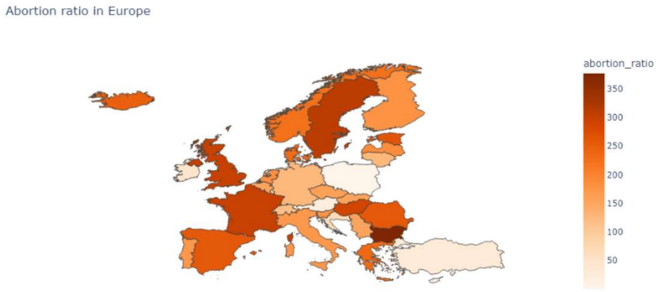
- Population of each country
- Age of the mother at her firstborn
- Proportion of the employed women with respect of the female population of each country
- Social protection benefits: "all interventions from public or private bodies intended to relieve households and individuals of the burden of a defined set of risks or needs" (Eurostat)
- Mean income per inhabitant
- Fertility rate, "the mean number of children that would be born alive to a woman during her lifetime if she were to survive and pass through her childbearing years" (Eurostat)
- Number of live births from:
    - women aged 18
    - women aged 30
    - married women
    - unmarried women
    - women that have a citizenship different from the one of the state where they have given birth
- Abortion ratio, the number of abortions per 1000 live births in a given year

The countries with too many NaN values have been removed. Instead, states with just some missing fields have been adjusted using the mean of the previous years or simple mathematical differences between known values. The year that has been taken as a reference is 2018 for data availability reasons. Finally, some

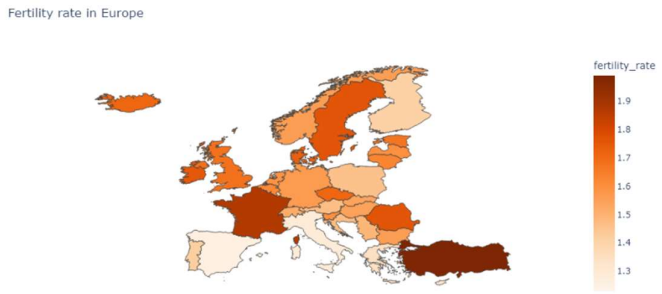
variables have been divided by the population of each country in order to be able to confront them.

B. *Exploratory analysis*

To have a first visual impression of the core phenomenon considered in this analysis, we can take advantage of two choropleth maps: one representing the abortion ratio and the other the fertility rate in Europe.



We can immediately see that the European country with the highest abortion ratio is Bulgaria, followed by Sweden and France. On the contrary, the country with the lowest abortion ratio is Poland, but also Austria, Turkey and Ireland have a limited value of this statistic.

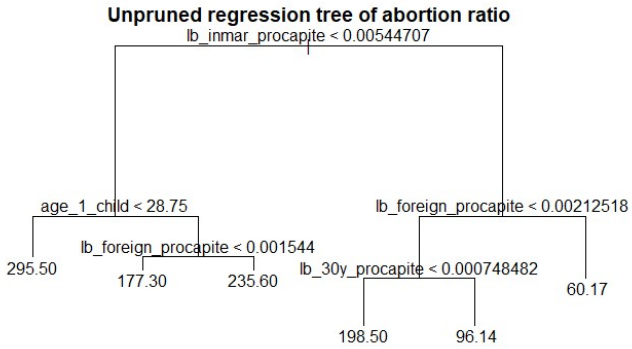


The countries with the highest fertility rate are Turkey and France, while the states where the number of children for every woman is very low are Spain and Italy. A thing worth mentioning is the values related to France: despite this country has a high abortion rate, it has also a considerable fertility rate.

C. *Analytical strategy and techniques*

In order to understand which role every variable of the dataset plays in the prediction of the abortion ratio, two different groups of methods have been used: linear and non-linear methods.

The first group that will be exposed is formed by the tree-based methods for regression. The regression tree is build using a top-down approach, called recursive binary splitting. At the top of the tree all the observations belong to the same region; then at each step, all the predictors and, for every predictor, all the possible cutpoints are considered. The splitting that will be used for the construction of the tree is the one with a combination of the variable X_j and the cutpoint s that minimizes the Residual Sum of Squares (RSS). This process is repeated, splitting each region obtained from the previous steps.

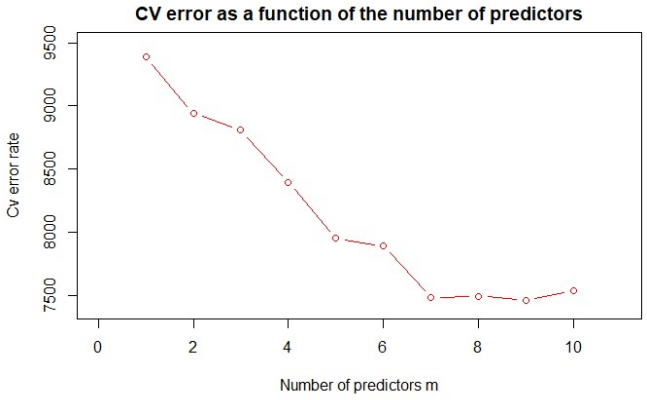


The plot of the model highlights that the predictors that have been used to build the tree are just 4 out of 10 and they are:

- Number of live births from married women per inhabitant. This is also the stump of the tree, so it means that is the most important variable from the ones selected in determining the abortion ratio. If this value is lower than 0.00545, the second variable to consider is the age of the mother when she gave birth to her first child. Instead, if the live births per capita inside marriage is greater than 0.00545, the second predictor to take into account is the live births per capita from foreign women.
- Mean women's age when they had their first child. It is located in the left side of the tree and its cutpoint is nearly 29 years old. If the mean age of the women is less than 29 years old when they have their firstborn, then the predicted abortion ratio of the country is 295.5 (about 295 abortions per 1000 live births), which is pretty high. Instead, if the mean age is lower than this cutpoint, also the variable regarding the live births from foreign women has to be considered.
- Number of live births from foreign women per inhabitant. This predictor is present both on the left and on the right side of the tree with two different cutpoints.
- Number of live births per inhabitant from women aged 30.

In summary, the unpruned regression tree assigns two of the highest values of the abortion ratio to the countries where the number of live births from married women is lower than 0.00545 per inhabitant.

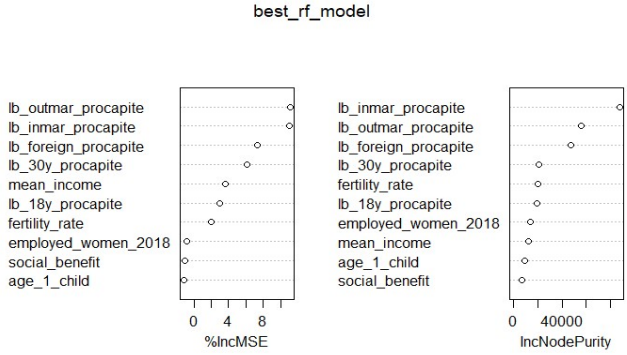
The following methodology applied is random forest, that takes repeated samples from the dataset, fits a regression tree for each sample and averages all the predictions. To avoid the presence of high correlated trees, at every split of each bootstrapped tree, random forest samples from the set of the predictors just a random subset of them ad selects the best one. At each split, a new sample of m predictors is taken. As a result, the correlation between the tree is reduced, as well as the variance of the model. A k-fold cross-validation algorithm could be used to select the best number m of predictors to pick at each split of a tree. The best m is the one associated to the lowest cross-validation error. The number of folds has been set as k=10.



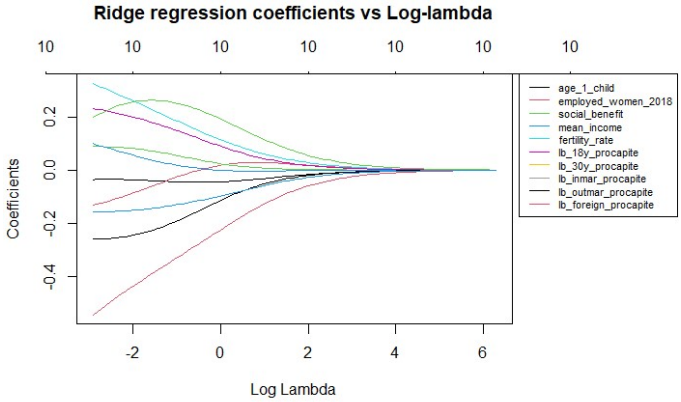
As we can see from the plot above, the best number of predictors to randomly pick at each split is 9 and we can use it to fit a random forest model to the dataset. The division of the dataset between training and test set has not been performed for tree and random forest because the main objective of their usage is to understand the variable importance of the predictors with respect to the abortion ratio. More emphasis on the accuracy of the model has been destined to the next ridge and lasso regressions, presented in the last part of the analysis.

	%IncMSE	IncNodePurity
age_1_child	-1.1407669	9456.823
employed_women_2018	-0.8273014	13598.531
social_benefit	-1.0875688	7168.569
mean_income	3.5774617	12127.467
fertility_rate	1.9551835	19955.762
lb_18y_procapite	3.0199188	19485.210
lb_30y_procapite	6.1229826	20620.559
lb_inmar_procapite	11.0485759	87458.091
lb_outmar_procapite	11.1074594	55691.015
lb_foreign_procapite	7.3373655	47277.974

The first column represents the mean increase of the Mean Squared Error (in percentage) if the variable is removed from the model. The lowest the MSE is, the highest the accuracy of the model becomes. The first three variables listed (age of the mother at her firstborn, number of employed women and social benefits) have negative values. This means that if we remove them, the MSE decreases and the accuracy of the random forest model improves. On the contrary, all the other predictors present positive values in the first column of the table, so their removal from the model would augment the prediction error. In particular, the two most important variables are the number of live births per capita from married and non-married women. The second column of the table measures the increment of the node purity that results from splits over that variable. As in the previous case, the greatest increase of the purity is associated to the live births per capita inside and outside marriage, while the lowest value is referred to the social benefits per inhabitant provided by the government. A visual representation of what has been explained above could be rendered by the following plot.



In order to understand not only the importance of the predictors with respect to the abortion ratio, but also the sign (positive or negative) of their association, we can perform a lasso and a ridge regression. They are called shrinkage methods because instead of performing a traditional linear regression, they constrain the coefficient estimates according to their relative importance with the response variable.



100 different values of lambda have been applied to the 10 predictors (plus the intercept), generating a corresponding number of ridge regression models. It can be seen that as long as the value of lambda increases, the

coefficients are shrunk towards zero, even if they are never completely zeroed. When lambda is equal to zero, on the very left side of the plot, the coefficients are equal to the ones of a linear regression. In order to find the optimal value of lambda we can divide the dataset into train and test sets and perform a 10-fold-cross-validation on the training set. Then we use the optimal lambda to fit a ridge regression model on the training set.

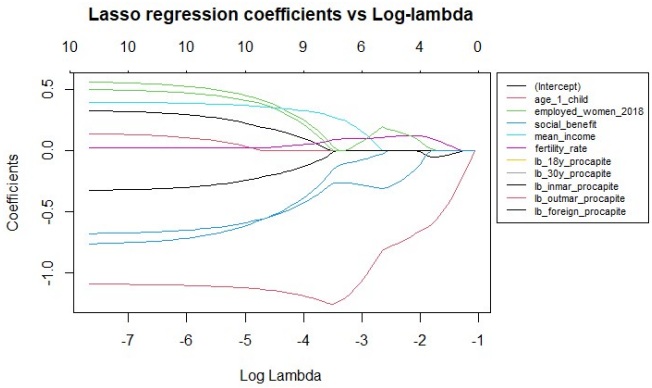
```
(Intercept)           0.000000000000000006543711
age_1_child          -0.045053952145868915446 50
employed_women_2018   0.015091525391011073944 92
social_benefit        0.027746670750649685827 09
mean_income           0.000996255338872606335 12
fertility_rate        0.122713655465564258628 67
lb_18y_procapite      0.097298201936891370866 79
lb_30y_procapite     -0.123761953025941812134 29
lb_inmar_procapite   -0.237840915158901961001 80
lb_outmar_procapite   0.201769034998209750852 99
lb_foreign_procapite -0.101838654674482706474 77
```

The coefficients with negative values imply that an increase of their referring variable leads to a decrease of the abortion ratio, keeping all the other elements fixed. This happens for:

- The age of the mother at her firstborn. If in a country women tend to become mothers when they are not too young, the number of abortions is lower
- The number of live births per inhabitant from mothers aged 30. This coefficient is one of the greatest, considering the absolute value. A woman in her 30s has probably a job or at least an economic and sentimental condition which is rather stable. Hence, she could be more stimulated to have a child. The opposite situation could regard a woman aged 18, who has just finished her course of study and she may not have enough resources to grow a child. In fact, the coefficient of this variable is positive, meaning that an increase of the number of live births from 18-years-old women in a country leads to a higher abortion ratio.
- The number of live births per inhabitant from married women. This is also the coefficient with the highest absolute value, so its variable has the major effect in predicting the abortion ratio of a country. Marriage is a social institution which guarantees several rights to their members and to the offspring; rights that are often not available for unmarried partners or single women. This could be one of the reasons that justify the negative coefficient of the in-marriage live births and the positive one of the out-marriage live births.
- The number of live births from foreign women has a negative value, so the increase of the variable results in a decrease of the abortion ratio.

Note that no coefficient is exactly equal to zero, because ridge regression shrinks all the coefficients without completely eliminate them.

We repeat the procedure adopted for the ridge regression also for the lasso. The main difference is the possibility for the lasso to perform variable selection.



At a first glance, the plot of the lasso coefficients could seem similar to the one of the ridge. However, in this case some coefficients are set to zero in correspondence of increasing lambda values. It is visible also on the horizontal axis on the top of the graph, which shows the number of coefficients that are greater than zero. While in the ridge plot this axis has always values equal to ten (total number of the predictors), in the lasso graph the values decrease along with the rise of lambda. We use the best lambda (found through cross-validation) to fit a lasso regression on the full dataset and we analyse the coefficients.

```
(Intercept)           0.000000000000000160858 5
(Intercept)           .
age_1_child           .
employed_women_2018   .
social_benefit        .
mean_income           .
fertility_rate        0.370156013288145402384 5
lb_18y_procapite      0.224165935605208505032 6
lb_30y_procapite     -0.162658117498205995188 3
lb_inmar_procapite   -0.640423989241336588662 5
lb_outmar_procapite   .
lb_foreign_procapite -0.020181013905620465154 2
```

Not all the coefficients are shown because some of them are shrunk to zero. The variables that are present in the model, which can also be considered as the most relevant in predicting the abortion ratio of a country, are:

- The fertility rate. Its coefficient is positive, meaning that when the fertility rate of a country grows, also the abortion ratio tends to increase
- The number of live births per inhabitant by women aged 18. Also in this case the variable is positively associated with the abortion ratio.
- The number of live births per inhabitant from married women. As already suggested in the ridge regression analysis, the negative value of this coefficient could be due to the rights and stability that a marriage provides. Hence, a country with a high number of births from married couples seems to have a lower abortion rate. The opposite situation is faced by countries with significant number of live births from unmarried women. In fact, the coefficient of this predictor is positive.

- The number of live births per inhabitant from foreign women, which has a negative coefficient, similarly to ridge regression.

IV. Results

The accuracy of the ridge and the lasso have been evaluated fitting the models on a training set and calculating the Mean Squared Error on a test set (validation set approach). As a result, the MSE of the ridge regression is lower than the one of the lasso, meaning that ridge performs better in terms of predictive accuracy. Generally speaking, the lasso is indicated in contexts where a small number of predictors has substantial coefficients and the remaining ones have very small coefficients. Instead, the ridge performs better when the response variable is associated to many variables and the coefficients have more or less all the same magnitude. In this case, if we exclude the intercept and the mean income, all the coefficients of the ridge regression are of the order of 10^{-2} or 10^{-3}. Hence, this could be one of the reasons why ridge seems to overperform lasso.

As regards the variable importance of the predictors, both linear (ridge and lasso) and non-linear methods (decision tree and random forest) are in general consistent. The main advantage of a decision tree is its intuitively description of the predictors and their significance with respect to the abortion ratio. Unfortunately, trees tend to overfit data. Therefore, an improvement that exploits and better the regression tree methodology is random forest.

A drawback of non-linear methods is the lack of the coefficient estimates, that does not allow to understand the sign and the intensity of the association between the predictors and the abortion ratio. This could be overcome with lasso and ridge regressions, able to reduce the variance when multiple predictors are present, at the cost of slightly increasing the bias.

V. Conclusions

The variable that all the methods highlight as the most important to determine the abortion ratio in Europe is the number of live births from married women. Also the live births from unmarried women is present in the majority of the models. This testifies the importance of the situation of the household in which the woman lives. As mentioned before, the rights granted to a married couple offer a stability that encourages women to become pregnant. Other important factors to take into consideration are the age of the mother and their citizenship: the number of live births from women aged 30 and from citizens of the country considered are associated to a lower abortion rate; the reason could probably be ascribed to the socio-economic stability reached at the age of 30 and to the easier access to public services and social rights destined to local women.

Of course, there are a lot of other variables that could be considered and could lead to a deeper comprehension of the phenomenon. But just with this simple analysis it is evident that, talking about the social and not the ethical sphere, the efforts of the states could be channelled into the extension of the women's rights rather than their reduction.

References

[1] R. Levinson-King, C. Kim, P. Sargeant, 2022, 'Abortion: What does overturn of Roe v Wade mean?', *The BBC website*, 29 June, Available: https://www.bbc.com/news/world-us-canada-61804777

[2] 'Abortion in Europe', Wikipedia, available: https://en.wikipedia.org/wiki/Abortion_in_Europe

[3] V. Rasch, T. Gammeltoft, L.B. Knudsen, C. Tobiassen, A. Ginzel, L. Kempf - European journal of Public Health, available: https://academic.oup.com/eurpub/article/18/2/144/455490

[4] Eurostat Data browser, available: https://ec.europa.eu/eurostat/databrowser/explore/all/popul?lang=en&display=list&sort=category

[5] Abort Report.eu website, available: https://abort-report.eu/europe/

[6] Johnston's Archive website, Wm. R. Johnston, available: http://www.johnstonsarchive.net/index.html

[7] G. James, D. Witten, T. Hastie, R. Tibshirani, 2021, *An Introduction to Statistical Learning*, Springer, New York