“C” FUNCTION – DESCRIPTION

Remarks on data and basic structure

The “C” function has the aim to provide a computational way to model the probability that agent *i* will commit a crime at time *t*: the way in which this function works is presented in this document. Its structure revolves around two types of data source: the first source are official statistics gathered from both the Italian National Institute of Statistics (ISTAT) and the Palermo registry office. The second source are several systematic reviews from which effect sizes (in different forms, e.g. odds ratios) of relevant factors that can explain the risk of committing an offense/getting involved in delinquency are retrieved.

The first source type (official data) provides information to:

1. empirically distribute gender and age classes across the whole simulated population, along with education and socio-economic status data;

2. Estimate the probabilities of committing a crime in each year for all individuals based on their gender and age class (Table 1).

Table 1. Gender and age class probabilities of committing a crime in a given year (source: authors' elaboration on ISTAT data)

(Gender , Age Class)[1]	**Probability**	**Odds Ratio[2]**	**Coefficient[3]**
(Female , <13)	0.001	0.00	-3.28
(Female , 14-17)	0.026	0.03	-1.58
(Female , 18-24)	0.056	0.06	-1.22
(Female , 25-34)	0.067	0.07	-1.14
(Female , 35-44)	0.066	0.07	-1.15
(Female , 45-54)	0.050	0.05	-1.28
(Female , 55-64)	0.031	0.03	-1.49
(Female , 65+)	0.011	0.01	-1.94
(Male , <13)	0.002	0.00	-2.60
(Male , 14-17)	0.168	0.20	-0.69
(Male , 18-24)	0.327	0.48	-0.31
(Male , 25-34)	0.320	0.47	-0.33
(Male , 35-44)	0.285	0.40	-0.40
(Male , 45-54)	0.204	0.26	-0.59
(Male , 55-64)	0.027	0.03	-1.56
(Male , 65+)	0.054	0.06	-1.24

[1] These figures will be updated in the following weeks since there was a little bug in the probability calculation sheets.
[2] Calculated in the standard way as OR/(1+OR)
[3] The coefficient is simply calculated through the log of the Odds Ratio.

These data are fundamental since they allow to estimate the average probability for each subclass of the population of committing a crime. These figures have been calculated using two different datasets within the ISTAT repository: both are related to the Sicilian region (in absence of a much specific geographic detail, e.g. Palermo province) and take into account the gender and age class of all known authors of crimes in the years 2012-2016 and the gender and age distribution of the overall Sicilian population in the same period. Probabilities are then calculated via the ratio of the two, and the provided figures are the average of these ratios (probability that a man/woman in a given age class is a known author of a crime) across the considered time-span.

The additional factors retrieved from the systematic reviews in accordance with the theoretical structure of C will allow to tune these values, increasing or decreasing the additive probability based on the presence or non-presence of a given characteristic. To maintain a compact and non-overwhelmingly expensive structure, we have selected few risk factors to test the way in which the function works and the emergent structure that it creates within the model. The factors are presented below:

Table 2. Risk factors for committing a crime

Risk Factor	**Odds Ratio**	**Coefficient**	**Probability**	**Official data to be matched**	**Operazionalization**
Unemployment	1.30	0.11	0.57	Yes	Having/not having a job
Education	0.94	-0.03	0.48	Yes	Having/not having an high school diploma
Natural propensity	1.97	0.29	0.66	No	Having a propensity higher than a certain value *x*
Criminal history	1.62	0.21	0.62	Emergent from the model	Having/not having committed a crime in the past
Criminal family	1.45	0.16	0.59	Emergent from the model	Having a share of criminal family ties which is higher or equal to 0.5. A criminal family tie is a direct link with a family member which has committed at least one crime in the last 2 years.
Criminal friends & co-workers	1.81	0.26	0.64	Emergent from the model	Having a share of criminal friends ties which is higher or equal to 0.5. A criminal friendship/professional tie is a direct link with a family member which has committed at least one crime in the last 2 years.
OC membership	*Assessed otherwise*			Emergent from the model	*To be calculated in the next weeks*

These are the risk factors that will be in dichotomous form, given that the odds ratio gathered from the systematic reviews mapped the risk of committing a crime being in a category (e.g. unemployed) vs not being in that same category. Therefore, we cannot make - at this stage - any further assumption regarding the way in which this odds changes when a more hierarchical structure is imposed (e.g.: three, or four classes instead of two).

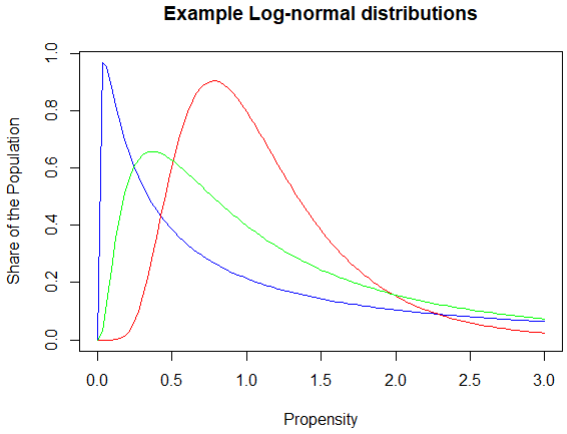
Furthermore, the third column in the table provides additional information on the origin and nature of the data at our disposal. Unemployment and education distribution within the population are available (from official statistics, as already mentioned), natural propensity cannot be data-driven modelled, while criminal history, deviant family ties, peer ties and OC membership are emergent from the model itself and will be originated by both C and R (the relational dimension that will capture the embeddedness in - among the others - OC communities).

Specifically, the number of deviant (criminal) family ties is dependent upon the C function of parents/relatives, and the same applies to (criminal) peer ties. At this stage, these latter two are again modelled as binary, but we will think of a way to control for the intensity/frequency/absolute number of criminal family and peer ties, in order to enrich the model and respect the hypothesis for which the higher the number of criminal ties (regardless of familiar or friendship nature), the higher the probability of getting involved into criminal activities.

Criminal propensity, which has the highest weight in terms of odds ratio and probability cannot be retrieved from real data. For this reason, we can include criminal propensity assuming that it behaves as a lognormal distribution. A positive random variable X is log-normally distributed when its logarithm is normally distributed:

$$\ln(X) \sim \mathcal{N}(\mu, \sigma^2)$$

This type of distribution is well-known and used in different scientific areas, including economics. Indeed, there is evidence that the income distribution usually follows the properties of a lognormal density function.[4]



In our case, allows to distribute the criminal propensity in accordance with the assumption that most of the population will not commit a crime and has not the “intrinsic” characteristics to offend: the magnitude of this propensity will depend upon the parameters that will be tested (i.e.: sample mean and standard deviation).

The Individual C Function and the Population Average Constraint

To derive the probability of committing a crime, we calculate the following function for each individual in the form:

$$P(\overline{C})_{[0,1]_i} = \left[(C|Gender, Age\ Class) \left(\sum_{j=1}^{m} risk_j \right) + 1 \right] + \varepsilon$$

Where the outcome variable is indeed C, the probability of committing a crime for individual *i* at time *t* which is calculated summing multiplying the average baseline probability for the individual given their age and gender with the summation of the risks ($risk_j$)[5] derived from the Odds Ratios included

[4] Fabio Clementi & Mauro Gallegati, 2005. "Pareto's Law of Income Distribution: Evidence for Germany, the United Kingdom, and the United States," Microeconomics 0505006, University Library of Munich, Germany.
[5] In order not to impose a fixed determinism into the model, we can decide that each risk derived from the odds ratios can float in a 95% CI range, so that we include heterogeneity within subpopulations with same characteristics.

in Table 2. Specifically, given the odds ratio of a risk factor, we increase/decrease the baseline risk by the percentage provided by the OR itself (e.g.: if the OR is equal to 1.41, and an individual has it among their characteristics, and their baseline is 0.15, it means that the final value has to be the product between the baseline and 0.41, namely the increase of the risk in percentage given that risk factor). An error term is also included and related to the need for bounding the individual probabilities of committing a crime to the population average. Indeed, at each time of reference (to be decided: a year? Every month?), the following equation shall hold:

$$C_{gender,age} \cong \frac{\sum_{i=1}^{n} P(\overline{C})}{n_{gender|age}}$$

The equation means that at each time of reference, the average probability of committing a crime for all individuals belonging to the same (gender,age) class shall be approximately similar to the fixed average values presented in Table 1, where approximately means that we can allow the model to float in a 95% confidence interval in order not to set overly deterministic mechanics to the model.

This is the type of function that has to be fitted for the individuals belonging to most (gender,age) classes. Indeed, there are four exceptions, specifically (Female,<13), (Female,>65), (Male,<13), (Male,>65). In these four cases the simple probability based on gender and age class is sufficient to model the risk of committing a crime, adding an error term stochastically distributed to make the low probabilities to float in order to prevent strict determinism. The decision is based on the assumption that all the risk factors that have been retrieved from literature do not play a role in the crime commission process when individuals are either too young or too old.