ANDRU Kilian
LACOUR Xavier

MASSIN Keryann
VEILLON Juliette

Applied Statistics Project - On the predictability of life course

Summary note

1 Introduction

The work presented in this document has been realized in the framework of the *Applied Statistics Project* by the joint participation of Keryann Massin, Kilian Andru, Xavier Lacour and Juliette Veillon. It is dedicated to the question of the predictability of the life course, a sociological question approached from a data science perspective. This executive summary will present our methodology and results without going into the technical details.

In the era of overly present information, new technical tools reveal links and hidden relationships between variables that used to be hard to tackle. Data science techniques spread among experts in all fields, and social sciences see major breakthroughs thanks to artificial intelligence. The question of the predictability of life course comes in this context as a challenging one: there is an obvious relationship between age, body-fat ratio and health for instance, but how could one exploit these links and other available information for prediction purposes? For instance, is it possible to forecast one's overall health for the upcoming year based on the socioeconomic and genetic data at our disposal? That is the problem we wish to address with this project.

The main reference of our project is the *Fragile Families Challenge*. The database comprises 4,242 families, who were interrogated in six waves to collect a total of 12,942 variables over 15 years. The objective of this challenge was to predict six life outcomes from the last wave, using the information contained in the previous waves. Several groups of researchers participated in the task and reported both their methods and results, so we were able to use their methods as an inspiration to find our own.

2 Presentation of the database

Our project used three databases, which all have the same source: the *Health and Retirement Study* (HRS). The HRS is sponsored by the National Institute on Aging (grant number NIA U01AG009740) and is conducted by the University of Michigan and focuses on Americans over 50. More than 20,000 individuals are included in the survey. The original goals of the survey were to help facilitate research and to guide policymakers in their decisions.

The first dataset is about participants' lives data (their marital status, age, health issues...). It contains fourteen waves of survey at the moment, which means that the survey was conducted at fourteen different periods in time. At each new wave, people from previous waves must respond again, and new people representative of American society are also added. The data here revolves around 10,000 variables.

The two other datasets correspond to the genetic ones. They are similar and just differ as one contains people with European ancestry while the other people with African ancestry.

3 Cleaning of the database

The database gathers 42,233 individuals and 15,104 variables and contains a lot of missing values. Those missing values can be replaced by a likely value if the individual responded to the given wave but did not answer a question, but we cannot replace them if the individual did not respond to the wave. Due to the large amount of data and the missing values, we need to find methods to reduce the dimension and impute (replace) missing values.

First, we drop the variables related to Spouses because most of them were also interrogated. This means that some information appears twice in the dataset. Then we drop the variables with too many missing values (more than 65%). Finally, since the point of this project is to predict health, we do not want to predict it using itself: we drop health-related variables. We now have 4,147 variables, which is more manageable.

4 Definition of the objective

As explained in the introduction, we want to see whether it is possible to forecast someone's health with the socioeconomic and genetic data we have at our disposal. We formally need to summarize the health-related information contained in the data into one number, since it is not possible to predict different variables at the same time. We call this number the *Global Health Index*.

To create this index, we selected 27 health-related variables. Among them are age, body mass index, the number of overnight stays at hospitals or nursing homes since the last survey... These are variables represented by numbers; we also used categorical variables, which are valued in a finite set of modalities. For instance, the variables which indicated whether the respondent had hypertension, cancer, diabetes, heart disease, etc. take "yes" or "no" as modalities.

The variables were selected based on their ability to differentiate individuals and the existence of relationships between them. A given variable can be useful to differentiate individuals if the values are sufficiently spread across individuals. Regarding the relationships, we want to know if an increase of a given variable is associated with an increase or a decrease of another one, or if one value of a variable is more likely to appear when another variable has a given value. For example, we noted that there seems to be more individuals with heart issues among those with hypertension than among those without.

We used these variables to compute an index for each wave using a statistical method called the *t-distributed Stochastic Neighbor Embedding*. We checked the performance of this algorithm by testing how it changed when we made changes in the data, like only using part of it or deleting values at random. This showed that the index is robust. In addition, higher values of the index refer to fragile health, while lower ones refer the healthier individuals. Indeed, an older age will generally correspond to more fragile health, and so will a high number of overnight hospital stays, for instance.

5 First procedure to format the database

The prediction methods we want to use do not work well when the number of variables approaches the number of individuals, which is the case here. Hence, we need to reduce the number of variables, using another method than deletion to lose as little information as possible in the process. There are many methods to achieve this, but we decided to use a Lasso regression. It has the advantage

of being quick to apply. Applying a Lasso regression allows us to detect variables that are not very useful to predict one's health and drop them. It is possible to use different parameters to keep the wanted number of variables. However, the function that is already implemented does not work with so many missing values; we need to find a way to address this issue. We tried three different methods for imputing missing values, but the first two failed. We finally used a High Missing rate Lasso (HMLasso), which is apt to efficiently deal with high missing rates.

6 Second procedures: prediction

6.1 Predictions using different lassos

First, we wanted to explore the prediction power of the usual lasso and the HMLasso presented above. For this purpose, we computed a regression for each wave. For each regression, we used the variable from the current wave and the previous predicted indexes. With this process, we wanted to account for the temporal correlation between the indexes. We only kept the 15,487 people who responded to the last wave, because it is the one we are interested in. To separate the effect of socioeconomic and genetic data, we applied this algorithm with and without genetic data.

To measure the quality of the prediction, we use the R^2. It is a value comprised between 0 and 1. The closer it is to 1, the better the prediction. However, in social sciences, a R^2 of 0.2 is considered satisfying. It is important to keep in mind that when we add explanatory variables, the R^2 necessarily increases. We look at how much it increases.

With this algorithm, we obtained a R^2 of 0.284 with genetic data, and 0.277 without. Incorporating genetic data leads to a marginal increase in the R^2. Therefore, we can say that this method is not efficient to analyze the role genes play in someone's health over time.

Then, we tried to use fixed effects models for the prediction, which we applied to the same data as previously. We used two slightly different algorithms, with the first, we proceed to select variables by cumulative Lasso regressions, we apply the *Within* transformation to those variables and finally proceed to a pooled OLS regression. The second algorithm is quite similar, we just apply the *Within* transformation before the cumulative selection. With the *Within* model, we can proceed to consistent estimations under the assumption of strict exogeneity which we tested and validated in our case.

With the first fixed effects method, we obtain a R^2 of 0.053 with genetic data and 0.049 without. With the second, we get a R^2 of 0.051 with genetic data and 0.051 without. In both cases, the genetic variables only have a slight impact on the quality of predictions which was already poor.

6.2 XGBoost

The Gradient Boosting paradigm is one of the methods which obtained the best results in the Fragile Families Challenge.

First, we applied the HMLasso making sure that we did not drop a variable in a wave while keeping it in another one. We kept a total of 1,500 variables for 42,233 individuals. However, a major issue of our dataset is the fact that some individuals have not responded to every wave. As explained before, when an individual has not been interrogated, it is not possible to impute the missing values. For this algorithm, we decided to keep only individuals present in each wave (3,396 individuals).

To separate the effect of socioeconomic data, genetic data, and previous predicted indexes, we performed the algorithm on different subsets of the data: socioeconomic data, socioeconomic and genetic data, socioeconomic data and previous indexes, only previous indexes and on all data.

Without using health data, we obtained a R^2 of 0.21, which is close to the best results of the Fragile Families Challenge. However, if we include the indexes from waves 1 to 13 to predict the one from wave 14, we obtain a better result with a R^2 of 0.37.

6.3 Random Forest

Random Forest is an alias path of Gradient Boosting, that we decided to explore considering the encouraging results.

We used the same reduced dataset as before. Furthermore, this algorithm does not support missing values, so we decided to impute them with the mean value of each variable. Using all data (socioeconomic, genetic, and previous indexes), we obtained slightly better predictions, with a R^2 of 0.397. However, the predictions using only socioeconomic data and only socioeconomic and genetic data are worse than with XGBoost (respectively a R^2 of 0.161 and 0.166).

7 Conclusion

In conclusion, our project's goal was to predict one's health by using socioeconomic and genetic data, while verifying whether the latter was useful. We used different machine learning methods, from lasso regressions to discreet models such as XGBoost and random forests. We overall obtained some significant results in explaining one's health but did not manage to show the usefulness of the genetic data to our purpose.

Testing more deeply some of our models or using deep learning may be solutions to get better results considering both issues.