Identification of causal factors leading people to choose high protein food
[Poster]

Irne Demongeot

UMR MIA-Paris, AgroParisTech, INRA, Universit Paris-Saclay, 75005, France

Abstract. Dietary guidelines are poorly followed in France. This is especially true for animal products, which are our primary source of proteins. A better understanding of what leads people to eat high-protein food would help nutrition experts to formulate better food recommendations. The aim of this project was therefore to identify causal factors of our food choices. We compared in this project several methods of causal inference on our data.

Keywords: causal inference, eating behavior, matching, propensity score

1 Motivation/Introduction

Nowadays in France, most of people follow badly the food based dietary guidelines. One hypothesis is that food recommendations are too far away from people's eating habits. It is especially true for animal products. For example, the results of the French survey INCA3 show that less than half of the people are following the recommendations for meat and seafood. Nutritionists believe that people tend to choose more animal food because these products are perceived as better to meet the demand for proteins of our organisms. But we don't quite understand the mechanisms leading people to prefer high protein food. In order to formulate better nutritional recommendations, we have to understand what are the causes leading people to prefer high protein food.

2 Methods

In nutrition and eating behavior, we are often interested in questions which are not associational but causal. One way to prove causal relationships would be to perform an experiment. For example if we want to prove the effectiveness of a drug, we could implement a controlled trial in which patient would be given the drug or not independently of any factors. Such controlled are not always possible, for practicable and ethical reasons. The main problem is then to estimate causal effects from observational data.

This problem has been well described by the Neyman-Rubin model or potential outcome framework [6]. We can formalize it as follow. The hypothesis of the model are : ignorability that is all the confounding factors are measured and that our data are randomized. Let an individual be represented by its features x and a treatment $t \in \{0, 1\}$. We want to measure the effect of the treatment for an individual. For each individual, there are two potentiel outcome : Y_0 if $t = 0$ and Y_1 if $t = 1$. The individual

effect of t for an individual i is then $\tau_i = Y_1 - Y_0$. The main problem is that for an individual we only measure one of the two potential outcomes. Under the assumptions of ignorability and randomization, we can estimate the average treatment effect with x being a vector of covariate :

$$ATE = \mathbb{E}(Y_1 - Y_0) = \mathbb{E}(\mathbb{E}(Y_1|x, t = 1) - \mathbb{E}(Y_0|x, t = 0$$

When we have observational data that are not randomized several algorithms exist such as matching or propensity score. The purpose of matching is to find for each treated example, the nearest example in the untreated group and to consider it as his counterfactual [7, 1]. The propensity score is $p(T = 1|x)$ [5, 2]. When $(Y(0), Y(1)) \perp X|T)$ then $(Y(0), Y(1)) \perp X|P(X = 1|U)$. We can then estimate a conditional causal effect $\mathbb{E}(Y(1) - Y(0)|P(X = 1|U))$.

3 Experiments and results

We organized an online survey to collect data about people's food choices. The survey was composed of three parts a socio-demographic questionnaire, an auto-evaluation by the person of his hunger and thirst and a series of 20 binary choices between two food items chosen among 211 food items selected to be representative of the food consumptions. From the choices, we inferred a score representing the person's willingness to eat high protein food. These data aren't randomized. Our goal was then to measure the effect of six variables : hunger, thirst, gender, weight control (whether the person is controlling it's weight or not), the level of education and the next meal anticipated (a snack or a complete meal)

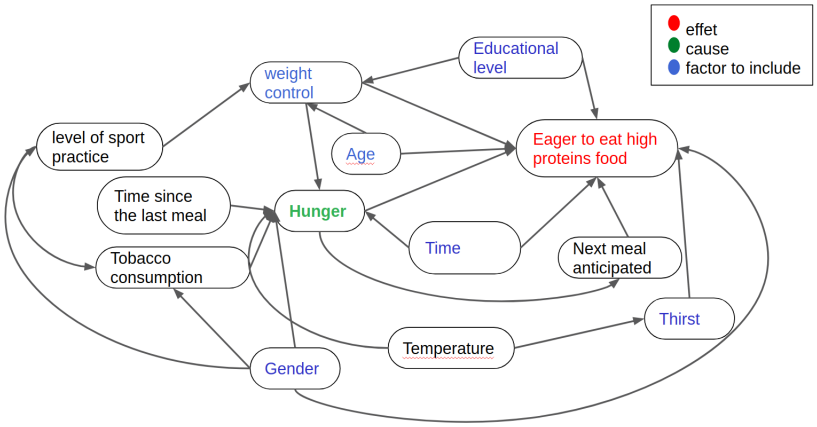

Fig. 1.

The first problem is to determine which confounding factors we have to include in order to have an identifiable effect. With the help of experts in nutrition and food behavior, we drew a causal graph representing the relationships between the different variables which could have an impact on the willingness to eat high protein food. The graph presenting the relationship between the variables implied when we want to test the effect of hunger is represented in 1. We use the back-door criterion to determine a sufficient set of variables from the graph [3, 4]. The back-door criterion states that to estimate the effect of X on Y, a set S of variables is sufficient for adjustment if no

element of S is a descendant of X and the element of S d-separate all "back-door" paths from X to Y (ending with an arrow pointing to X). According to this criteria, for the variable "hunger", a sufficient set is :

$$\{Age, Level\ of\ education, gender, weight\ control, time, thirst\}$$

We estimate the effect of the different variables using two different methods : first we estimate the effect using a matching method with a Mahalanobis distance. We use a maximum of 3 matches when we estimate the effect of hunger because there was about three times more hungry people than satiated people. Using this kind of matching introduces a bias in the analysis. We correct it by assuming a linear relationship between Y_0 and X For all the other variables we use a maximum number of match of 1. The second method we use is the propensity score method. We estimate it with a logistic regression on the covariate. We group then the data by propensity score. The number of group is calculated with the algorithm defined by Imbens and Rubin [1]. In each group we estimate the causal effect with a linear regression and the global effect in the population as the mean effect on every group.

The results show that the causal variable for the willingness to eat high protein food are hunger, thirst and gender. The principal cause, with the highest causal effect is hunger. The difference between the two methods is that with the propensity score method, the 95% confidence interval we calculate is shorter. As hunger was measured as a continuous score, we measured the effect of hunger by setting a limit. We tested different limit and show that the more the people are hungry the more they want to eat high protein food.

4 Discussion

Our results confirm what we see in the literature in eating behavior. The experiments were done on the basis of the graph given by the experts. It gives to experts directions to set a randomized experiment. As the volume of data was rather small (851 questionnaires were collected), the confidence interval stay quite large. It would be interesting to repeat this experiment on a larger population. On a larger dataset it would be also interesting to look for heterogeneous effect because in eating behavior the variables can have different effect depending on which part of the population we work on.

References

1. Guido W Imbens. Matching methods in practice: Three examples. *Journal of Human Resources*, 50(2):373–419, 2015.
2. Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.
3. Judea Pearl. Causal inference in statistics: An overview. *Statist. Surv.*, 3:96–146, 2009.
4. Judea Pearl. *Causality*. Cambridge university press, 2009.
5. Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.
6. Donald B Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.
7. Elizabeth A. Stuart. Matching methods for causal inference: A review and a look forward. *Statist. Sci.*, 25(1):1–21, 02 2010.