Estimation of the Causal Effect of *Income* on *Vaccination*

Neel Rajendra Barge
Copenhagen Business School
neba22ac@student.cbs.dk

1 Introduction

The qualitative knowledge of causal relationships in the domain is represented by a causal model shown in Fig. 1. The treatment variable is $Income$ and the outcome variable is $Vaccination$. The causal effect $do(Income = income)$ on $Vaccination$, written as $P_{Income}(Vaccination)$, is identifiable from the distribution over the observed variables $P(Afford, Age, Education, EssentialWorker, FamilyWealth, Fear, GovtMando$ The identifiability of $P_{Income}(Vaccination)$ allows us to estimate the quantity of $P_{Income}(Vaccination)$ from the given observational dataset *UploadDataSet.xlsx*.

The details of the dataset, input parameters, and the estimation result are presented in the following three sections. Section 2 provides basic information on the dataset, such as statistics of the variables. Section 3 lists a set of input parameters necessary for estimating $P_{Income}(Vaccination)$. Section 4 demonstrates the result of the estimation from both "experimental" and observational distribution over the variables.

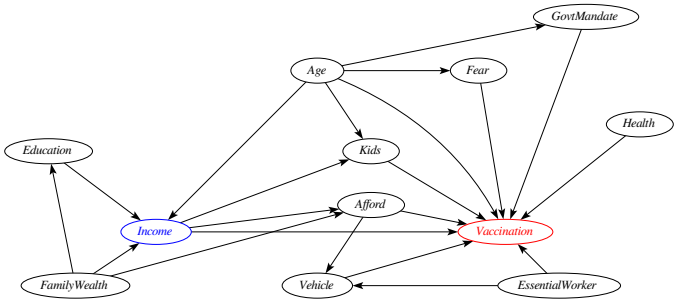

Figure 1: The causal model. $Income$ is the treatment variable, and $Vaccination$ is the outcome variable.

2 Dataset

The dataset *UploadDataSet.xlsx* has twelve variables $\{Afford, Age, Education, EssentialWorker, FamilyWealth, Fe$ The statistics of the variables are shown in Table 1. Note that $Income$ is made a continuous variable (i.e., smoothed out) since $Income$ can take many different values.

3 Parameters

Three parameters are necessary for estimating $P_{Income}(Vaccination)$.

1. Prediction algorithm: Generalized Additive Models.

   The following parameter specifies a list of machine learning algorithms and/or regression methods used for the estimation. We used the algorithms implemented in SuperLearner [Polley 2010] package written in R.

2. P-value: 0.05.

	$Afford$	Age	$Education$	$EssentialWorker$	$FamilyWealth$	$Fear$	$GovtMandate$	He
Type	Discrete	Continuous	Discrete	Binary	Continuous	Binary	Binary	Cont
μ	1.263	44.205	1.001	0.512	19153.803	0.824	0.47	5.
σ	0.808	14.846	0.705	0.5	4916.997	0.381	0.499	2.
Min	0	18	0	0	4234	0	0	
25%	1	32	1	0	15764.5	1	0	
Median	1	44	1	1	19220	1	0	
75%	2	57	1	1	22517.5	1	1	
Max	2	70	2	1	37098	1	1	

Table 1: Statistics of variables of *UploadDataSet.xlsx*.

3. Number of folds: 3.

   The number of folds refers to the number of groups that the dataset will be split into. Splitting up the dataset is one of the procedures of k-fold cross-validation, a resampling technique used for evaluating the performance of a trained machine learning model.

4 Estimation

This section presents the result of the estimation from both "experimental" and observational distribution over the variables. The "experimental" distribution over the variables is generated from the observational distribution using reweighting methods, such as inverse probability weighting. The result is shown in Fig. 2.

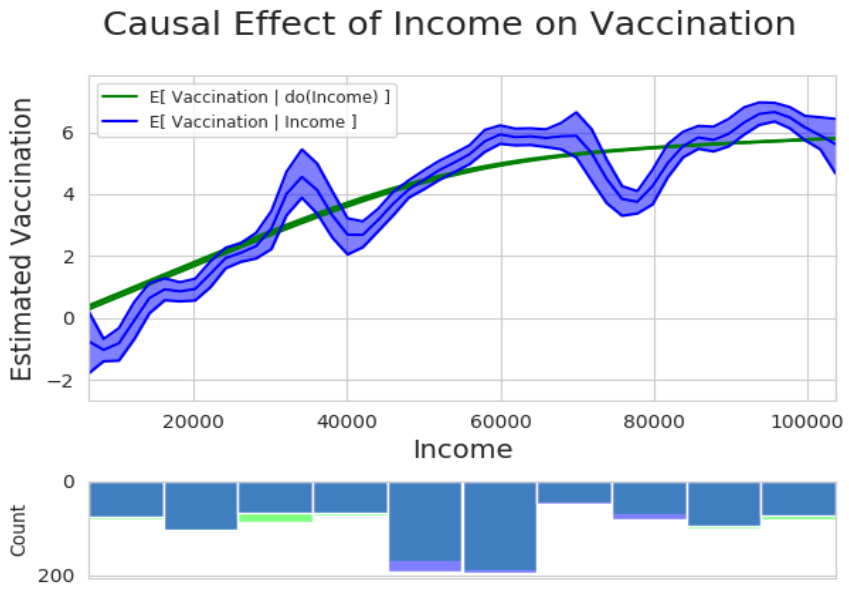

Figure 2: The estimation of the causal effect of *Income* on *Vaccination*. The experimental distribution over the variables is generated by reweighting the observational distribution.

The statistics associated with the result are shown in Table 2.

Experimental Distribution

	$Income$								
	12323.5	22901.5	35949	46848.7	51905.2	57462	62445.2	72633.7	85908.6
$E(Vaccination\|do(Income))$	0.7394	1.7434	2.7299	3.6379	4.3877	4.9289	5.2735	5.4877	5.6354
$\sigma(Vaccination\|do(Income))$	0.3179	0.3167	0.3045	0.2664	0.2056	0.1375	0.084	0.0545	0.0402
Min	0.3373	1.3425	2.3419	3.2947	4.1191	4.7469	5.1621	5.4163	5.5833
25%	0.5383	1.5433	2.5386	3.4724	4.262	4.8463	5.2233	5.4547	5.6107
Median	0.7394	1.7438	2.7329	3.6443	4.3963	4.9371	5.2787	5.4902	5.6366
75%	0.9405	1.944	2.9243	3.8097	4.522	5.0196	5.3289	5.5233	5.6613
Max	1.1415	2.1435	3.1118	3.9682	4.6389	5.0944	5.3746	5.5543	5.685

Observational Distribution

	$Income$									
	12323.5	22901.5	35949	46848.7	51905.2	57462	62445.2	72633.7	85908.6	97
$E(Vaccination\|Income)$	-0.3406	1.1523	3.2912	3.7261	4.6112	5.7206	5.284	4.4393	6.0748	6
$\sigma(Vaccination\|Income)$	0.9425	0.4087	1.2633	0.5728	0.4872	0.3285	0.6175	0.8052	0.4774	0
Min	-1.4862	0.7646	1.9495	3.2489	3.8209	5.1794	4.2902	3.715	5.5846	6
25%	-1.0197	0.8307	2.2741	3.3653	4.4596	5.6725	5.068	3.7284	5.7118	6
Median	-0.3532	1.0294	3.067	3.4281	4.8615	5.801	5.6493	4.2411	5.9374	6
75%	0.3642	1.4056	4.2977	3.9616	4.9526	5.9586	5.6869	4.9475	6.4474	6
Max	0.7921	1.7311	4.8679	4.6264	4.9613	5.9914	5.7258	5.5643	6.6929	6

Table 2: Statistics of variables in the experimental and observational distribution over the variables of *UploadDataSet.xlsx*. $Income$ is divided into ten equidistant intervals.

References

[Polley 2010] E. C. Polley and M. J. van der Laan. "Super Learner in Prediction." 2010. "Super Learner in Prediction." https://cran.r-project.org/web/packages/SuperLearner/vignettes/Guide-to-SuperLearner.html