CentraleSupélec

Case Study : Causal Inference

Toward Ethical Algorithms ?

Mohammed FELLAJI, Ahmed BEN AISSA

Supervisor : Frédéric PENNERATH

September, 2020

Contents

1 Introduction

One of the most challenging questions in every problem is the one related to understanding the reason(s) why an action happened and whether or not we can explain it with the information we have at our disposal. Another interesting question one might ask is what would be the outcome if the conditions of the experiment were different ? What would happen if we have more/less information ? What will we get if we change completely the set of inputs ?

When thinking about these questions, having a time machine seems as the perfect solution : we can then repeat the same experiment with different initial conditions and record the outcome in every scenario. A more realistic and possible solution would be to use Causal inference, which aims to estimate the likelihood of an event under static conditions and also under dynamic changing conditions.

While searching about causal inference, one will definitely come across some of the work of Judea Pearl who is credited for developing a theory of counterfactuals and causal inference based on structured models.[1]

In this document, we will try to study the different literatures about causal inference and collect the results in a simple and yet detailed way. Many papers and book are mentioned in the references section, some were used in writing these document, some are not. Those interested more in the subject may take a look at them.

[1] After watching dozens of Judea Pearl lectures and reading many of his papers, we can only recommend doing the same. One of his main ideas is that even if machine learning is shaping millions of industries around the globe, this is done without attention to fundamental theoretical impediments. This might lead machine learning algorithms, very soon, to reach the barriers of impossibility. According to Judea Pearl, the goal is to use the knowledge acquired from causal inference and combine it with the success of machine learning in order to achieve more general models.

2 Introductory example : Simpson's Paradox

Simpson's paradox, also called Yule-Simpson effect, in statistics, is an effect that occurs when the marginal association between two categorical variables is qualitatively different from the partial association between the same two variables after controlling for one or more other variables.

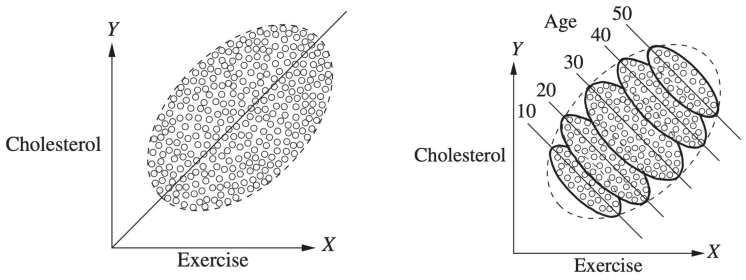


Figure 1: Illustration of the Simpson's Paradox[1]

Here is a more detailed illustration of the paradox :

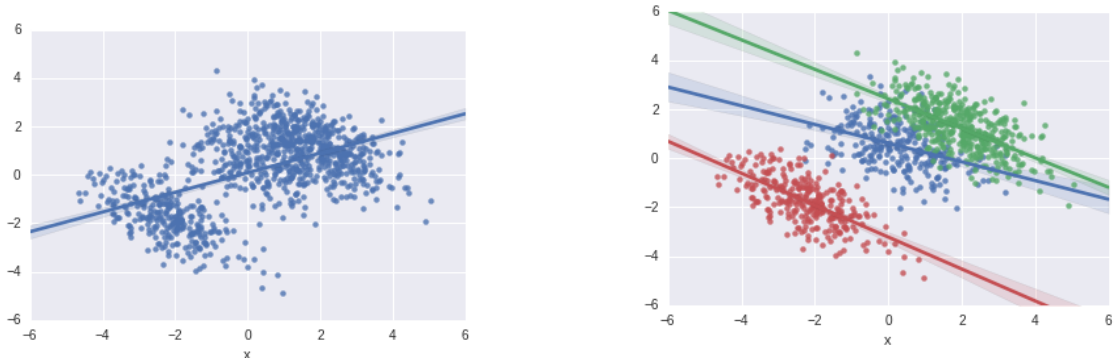
Imagine we are trying to work out whether a certain drug is an effective treatment for a disease. To decide whether it is effective, we compare people who took the drug (call this x, if someone took the drug, x=1, if not x=0), by examining how many of each recovered from the disease (call this y: y=1 means they got better, y=0 means they did not).

z	X=0		X=1	
0	1/22	0.05	**18/109**	**0.17**
1	1/5	0.20	**13/36**	**0.36**
2	2/4	0.50	**3/5**	**0.60**
3	61/91	0.67	**51/72**	**0.71**
4	112/128	0.88	**25/28**	**0.89**
total:	**177/250**	**0.71**	110/250	0.44

Figure 2: Results of a study on the effectiveness of a new drug, layered by age group

When we just look at the last row of Figure 2, we find that of 250 people who took the drug, 110 recovered (44%), whereas out of the 250 people who did not take the drug 177 recovered (71%). From these results it looks like there is a clear advantage to not taking the drug. Unfortunately, this drug was not administered as part of a random controlled trial. This means that the decision of whether or not to take the drug may have been confounded.

Now let's look at the recovery rate per age group : In every row representing each age group, the



(a) Positive Correlation

(b) Negative correlation

Figure 3: Reversal of correlation when a confounding variable, here age, is taken into account

patients who were administered the drug got noticeably better. Thus our conclusion about the drug effectiveness is reversed.

This can be illustrated by Figure 3 above.

It is clear now, more often than not, without proper ground knowledge (here it is medical expertise), we can ignore confounders such as age, income, .. etc. And without proper randomized experimentation, which is very costly, our understanding of causation can be be contrary to reality.

3 Fundamental notions

3.1 Causation vs Association

3.1.1 Example

One of the most common phrases in statistics is "correlation does not imply causation". To understand it well, let's take a look a the following example :

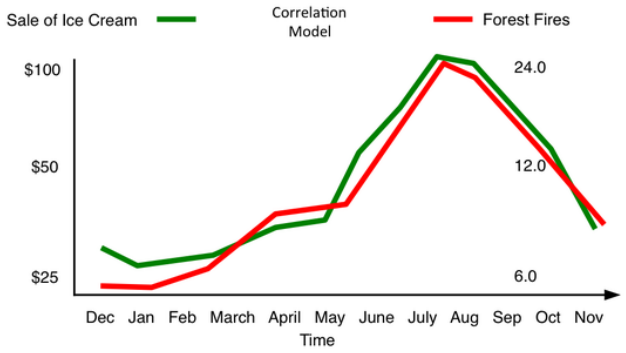


Figure 4: an example of correlation (source)

Without having any information about what we are trying to model, we might conclude that there is a cause-effect relationship between these two measures. The graph also shows a correlation close to 1 for the two curves. When we start to analyse the data in details, the first thing that comes to our mind is that there is no logical relationship (and thus no cause-effect relationship) between the sale of ice cream and the forest fires. However, we can see clearly in the graph that the values are higher between May and November (with a peak in July) compared to the rest of the year. In fact, the heat is the reason behind forest fires and the sale of ice cream. We can then conclude the following :

- a causation between the heat and the sale of ice cream;
- a causation between the heat and the forest fire;
- a correlation between the sale of ice cream and the sale of ice cream.

This simple example shows us that, in general and without having a good knowledge about the problem, we tend to assume simple correlations between 2 variables when in fact there is a third variable that causes both of them. This is a core idea in causal inference : unlike for correlation, we can not rely only on the distribution of the data, even at the population level, but we also should rely on causal assumption that is always not testable in observational studies.[2]

3.1.2 Definitions

The previous example shows a clear difference between causation and correlation. More generally, correlation is a special example of an association concept. The definition of these two concepts is given by Judea Pearl[2, 3] as follows :

– An **associational concept** is any relationship that can be defined in terms of a joint distribution of observed variables. For example: regression, correlation.

– A **causal concept** is any relationship that cannot be defined from the distribution alone. One should also rely on causal assumption that explains the different relationships. For example: randomisation, confounding (it will be detailed later).

3.1.3 The difference between the 2 notions

To have a better understanding of how causation and association differ, the following example is considered. The goal of a study is to test the effectiveness of a vaccine on a sick population. Let's denote by A the action of injecting the vaccine and by Y the effect of the vaccine on the population. In an association approach, the population will be divided into 2 groups : only one group will be injected with the vaccine. The effectiveness of the treatment is thus measured on the treated population ($E[Y|A = 1]$) and it is also possible to see if the untreated population will be treated without the vaccine ($E[Y|A = 0]$). However, in a causation approach, we suppose that is possible to inject the whole population and study the effect of the vaccine ($E[Y^{a=1}]$) and at the same time, have the same population **not** injected ($E[Y^{a=0}]$).

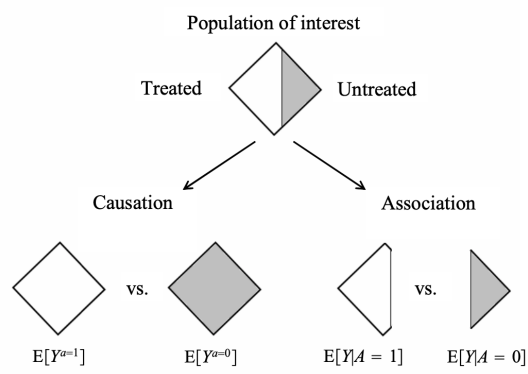


Figure 5: a visualisation of the difference between causation and association

One can see clearly the impossibility of applying in the causation approach : it is impossible to go back in time and inject an individual who was not injected; we can only observe one of the different potential outcomes. Meanwhile, there is some technics so that such an experiment can be possible: randomised experimentation is one example that will be detailed later.

3.2 Counterfactuals

This is an important concept for causation. Counterfactuals could be defined as the answer to the question "what if ?" or simply as the unobserved outcome. If we consider the previous example in the case of an association, the unobserved outcomes will be the observations of "what if we have injected the proportion of the population that was not injected ?" and of "what if we have not injected the proportion of the population that was injected ?".

3.3 Confounders

A confounding variable is an unmeasured variable that influences both the supposed cause and the supposed effect. Ignoring the confounder may lead to conclude an association between 2 variables when actually it is not the case. Confounders can also increase the variance or introduce bias.

Different technics exist to reduce the effect of the confounders (random samples, control variable ...)[2]. However, knowing in advance all the the confounders in the data/model is still very important.

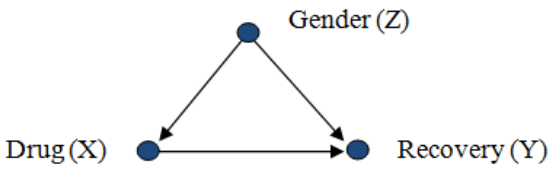


Figure 6: An example of a confounding variable : in this example, the Gender is not observed and it is affecting the variables Drug and Recovery

4 A mathematical model of causal inference

In the this section, we suppose the following :

- Y_i : random variable of the outcome for the "unit" i (a unit is a physical object, for example, a person, at a particular point in time [4]),
- N : the number of units,
- Y : (Y_1, Y_2, ..., Y_n),
- A : treatment (a treatment is an action that can be applied or withheld from the unit [4]),
- Y_i^a : controlled treatment

For simplicity, we assume that the treatment (A) can take only 2 values : {0,1}.

4.1 Individual level Causal Effect : ICE

The ICE is defined as the difference between an individual's two potential outcomes :

$$ICE = \delta = Y_i^{a=1} - Y_i^{a=0}$$

4.2 Average Causal Effect : ACE / Average Treatment Effect : ATE

The ACE or ATE is the population average of the individual level causal effects :

$$ACE = E[\delta] = E[Y^{a=1}] - E[Y^{a=0}]$$

[2] An interesting article about confounders could be found at www.statisticshowto.com

ACE is computed over the entire population. If the treatment has a causal effect on the outcome, the value of the ACE will be different than 0 and it will be equal to zero in the other case. This is a mesure of the causation.

4.3 Standard estimator : S*

$$S^* = E[Y^{a=1}|A = 1] - E[Y^{a=0}|A = 0]$$

S* is computed from the treatment and control groups, and not from the entire population as for the ACE. This is a mesure of the association.

Important special case **: Randomized experimentation**

A randomised experiment is when the investigator carried out the action of interest and it wasrandomised because the decision to act on any study subject was made by a random device.Randomisation results in convincing causal inferences, the downsides being mostly finances and logistics.

In the case of randomised experiment, we have the following :

- $E[Y^{a=1}|A = 1] = E[Y^{a=1}|A = 0] = E[Y^{a=1}]$
- $E[Y^{a=0}|A = 1] = E[Y^{a=0}|A = 0] = E[Y^{a=0}]$

Based on the assertions above, it is easy to prove that a randomised expriment leads to :

$$ACE = S^*$$

which means that it is possible to determine causation from the existing data. This also suppose comparability conditions so that the individuals in both groups are identical.

5 A working example : Evaluating the Econometric Evaluations of Training Programs

All results are obtained through the python notebook added to the git repository of the project [5]. It is inspired by the models in [6] , [7], and [8].

5.1 Dataset

The dataset used to exhibit a model of causal inference is the famous Lalonde datatset, based on the paper : "Evaluating the Econometric Evaluations of Training Programs" by Robert Lalonde[9].

It studies the effect of education/job training (the "treatment") on the revenue of students after four years in 1978 ("the outcome").

The goal is to estimate the causal effect of the education on revenue, and how it compares to other confounding variables, here it is likely age, education, race and marital status.

	age	educ	black	hisp	married	nodegr	re74	re75	re78	u74	u75	treat
0	37	11	1	0	1	1	0.0	0.0	9930.05	1	1	1
1	22	9	0	1	0	1	0.0	0.0	3595.89	1	1	1
2	30	12	1	0	0	0	0.0	0.0	24909.50	1	1	1
3	27	11	1	0	0	1	0.0	0.0	7506.15	1	1	1
4	33	8	1	0	0	1	0.0	0.0	289.79	1	1	1

Figure 7: Dataset overview

5.2 Problem

- Y represents the response, here is is 1978 earnings ("re78")
- D represents the treatment: the job training program ("treat")
- X represents the confounding variables, here it likely is age, education, race and marital status.

Problem statement : How can we disentangle the pure effect of the job training (Treatment D) from that of age, education, race, and marital status (Confounders X), all have a likely influence on earnings Y ?

What we want to know here is the Average Treatment Effect (ATE):

$$\Delta = ACE = E[Y^{D=1}] - E[Y^{D=0}]$$

However, since we do not have a randomized setting, if we try to estimate this quantity from the row observational distribution, we get:

$$\Delta^{'} = E[Y|D=1] - E[Y|D=0] \neq \Delta$$

This difference is due to the fact that we do not know, outside a randomized setting, that the treatment D itself is independent from the covariate X. However for the rest of this case study, and to be able to use the packages DoWhy and CausalIinference on Python, we make the assumption that D is independent from X.

5.2.1 Data analysis, Causal discovery

At first, let look at the statistical description of our database :

	age	educ	black	hisp	married	nodegr	re74	re75	re78	u74	u75
count	445.000000	445.000000	445.000000	445.00000	445.000000	445.000000	445.000000	445.000000	445.000000	445.000000	445.000000
mean	25.370787	10.195506	0.833708	0.08764	0.168539	0.782022	2102.265533	1377.138638	5300.765138	0.732584	0.649438
std	7.100282	1.792119	0.372762	0.28309	0.374766	0.413337	5363.583863	3150.961433	6631.493362	0.443109	0.477683
min	17.000000	3.000000	0.000000	0.00000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	20.000000	9.000000	1.000000	0.00000	0.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000
50%	24.000000	10.000000	1.000000	0.00000	0.000000	1.000000	0.000000	0.000000	3701.810000	1.000000	1.000000
75%	28.000000	11.000000	1.000000	0.00000	0.000000	1.000000	824.389000	1220.840000	8124.720000	1.000000	1.000000
max	55.000000	16.000000	1.000000	1.00000	1.000000	1.000000	39570.700000	25142.200000	60307.900000	1.000000	1.000000

Figure 8: Statistical description of the dataset

Let's have a deeper look at how revenue is distributed between the group who got the job training and the one who did not :

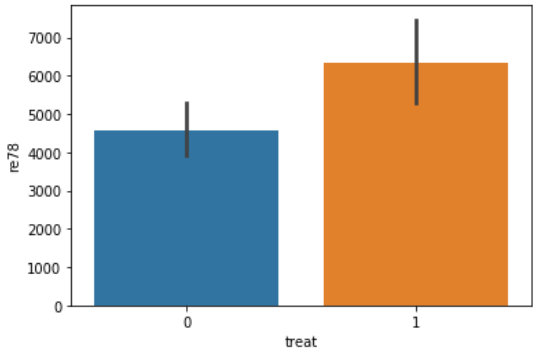
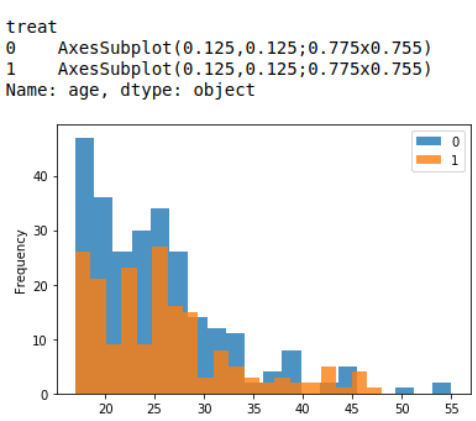


Figure 9: Revenue average for those who followed the job training program versus those who didn't

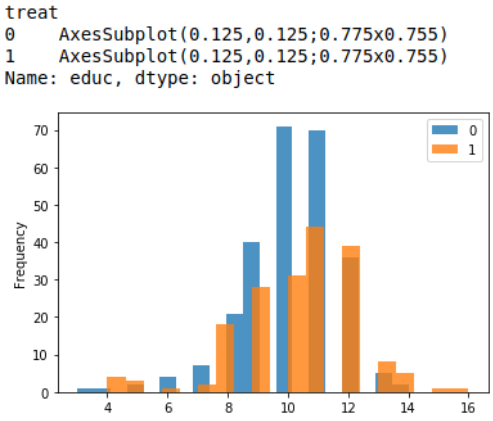
As we could have expected, the group who benefited from the job training program have a higher average revenue after 4 years. However, in order to quantify the net effect of the job training program, we have to check the balance of our data set, and cross-examine the influence of each and every confounder of X.

We start with age (left figure) and education (right figure) represented in Figure 10 :

At first look, it is clear that our dataset is imbalanced. This imbalance in age distribution, as well as other confounders, makes drawing any conclusion erroneous.



(a) Age imbalance



(b) education imbalance

Figure 10: Dataset imbalances

5.2.2 Naive estimate

At first, let's have a naive estimate of the impact of the job training program on revenue. We ignore any potential confounder of X. We obtain through the mean formula :

$$\Delta^{'} = ACE_0 = E[Y1] - E[Y0] = \$1794.34$$

5.2.3 Linear Model

An intuitive approach to the estimating the impact of the confounders X would be the linear model, assuming the law of the revenue Y follows :

$$Y = \alpha + \beta D + \gamma X$$

A simple multi-variable linear regression allows us to obtain specific estimate for each and every coefficient. Using the CausalModel library on python, we compute the mean using the obtained coefficients, and we obtain a rough estimate of the Average Causal Effect :

```
Treatment Effect Estimates: OLS

                  Est.       S.e.          z      P>|z|      [95% Conf. int.]
-------------------------------------------------------------------------------
           ATE   1676.902    642.870      2.608      0.009    416.877   2936.927
```

Figure 11: Average job training effect according to the matching model

Thus, according to the linear model, the net increase in revenue due to the job training is $ 1676.90, a significant decrease from the $1794.34 in the naive model. This decrease is due to X, the confounder variable, being taken into account.

However, the linear model has it limits. First of all, it is heavily influenced with data imbalances. Moreover, the OLS estimator is a simplistic point of view and may not capture all the intricacies of the interaction of the outcome Y with the confounders.

In the next section, we offer a more sophisticated approach, that improves on the linear model.

5.2.4 A more sophisticated approach : propensity score, stratification and matching

The **Propensity Score** is the probability of receiving the treatment, conditional on the covariates.

$$p(X) = P(D = 1|X)$$

It has two main uses :

- Instead of using X as the covariate, $p(X)$ summarises the information contained in X that is relevant to causal inference.
- It allows us to drop samples with extreme propensity scores, and to focus on matching on similar samples

Both the CausalInference and DoWhy allows for a quick estimation of p(X) using statistical likelihood tests, and for trimming the dataset using this estimation.

Stratification is clustering the samples into layers of similar propensity scores, which allow for more consistent estimations.

Matching estimation exploits the specific distribution within the stratified layers, using nearest-neighbour methods, to improve on the quality of estimates and detect the slightest of effect confounding variables have.

Applying these three methods to our dataset, we obtain the following output to our algorithm:

```
Treatment Effect Estimates: Matching

                    Est.       S.e.          z      P>|z|      [95% Conf. int.]
-------------------------------------------------------------------------------
           ATE   1604.359   1071.583      1.497      0.134   -495.944   3704.661
           ATC   1383.124   1232.951      1.122      0.262  -1033.460   3799.709
           ATT   1916.973   1145.752      1.673      0.094   -328.702   4162.647
```

Figure 12: Average job training effect according to the matching model

The average treatment effect of the job training program is an $1604 increase in revenue over the following four years.

5.2.5 Conclusion

While our model is a very simplified approach to the causal inference problem, it offers a solid base to build on, using more sophisticated model. Deep learning models in particular, as cited in [10], do not assume linearity, which is far more reflective of reality, and allow for better estimation of the Average Causal Effect of a perceived cause on a specific outcome

References

[1] J. Pearl, M. Glymour, and N. P. Jewell, *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.

[2] J. Pearl, "The mathematics of causal relations," *Causality and Psychopathology: Finding the Determinants of Disorders and their Cures (P. Shrout, K. Keyes and K. Ornstein, eds.). Oxford University Press, Corvallis, OR*, pp. 47–65, 2010.

[3] J. Pearl *et al.*, "Causal inference in statistics: An overview," *Statistics surveys*, vol. 3, pp. 96–146, 2009.

[4] D. B. Rubin, "Causal inference using potential outcomes: Design, modeling, decisions," *Journal of the American Statistical Association*, vol. 100, no. 469, pp. 322–331, 2005.

[5] "Github repository."

[6] "Lalonde pandas api example."

[7] "Causal inference 1 potential outcomes."

[8] "Causal inference in python."

[9] "R: Lalonde dataset."

[10] L. Yao, Z. Chu, S. Li, Y. Li, J. Gao, and A. Zhang, "A survey on causal inference," *arXiv preprint arXiv:2002.02770*, 2020.

[11] A. Marx and J. Vreeken, "Testing conditional independence on discrete data using stochastic complexity," *arXiv preprint arXiv:1903.04829*, 2019.

[12] B. Schölkopf, "Causality for machine learning," *arXiv preprint arXiv:1911.10500*, 2019.

[13] M. A. Hernán and J. M. Robins, "Causal inference: what if," *Boca Raton: Chapman & Hill/CRC*, vol. 2020, 2020.

[14] P. Judea, "Causality: models, reasoning, and inference," *Cambridge University Press. ISBN 0*, vol. 521, no. 77362, p. 8, 2000.

[15] J. Pearl, "The science and ethics of causal modeling," *Handbook of ethics in quantitative methodology*, pp. 383–416, 2011.

[16] H. Matute, F. Blanco, I. Yarritu, M. Díaz-Lago, M. A. Vadillo, and I. Barberia, "Illusions of causality: how they bias our everyday thinking and how they could be reduced," *Frontiers in Psychology*, vol. 6, p. 888, 2015.

[17] C. Glymour, K. Zhang, and P. Spirtes, "Review of causal discovery methods based on graphical models," *Frontiers in Genetics*, vol. 10, p. 524, 2019.

[18] "DoWhy: A Python package for causal inference." https://github.com/microsoft/dowhy, 2019.

[19] J. Pearl, "The seven tools of causal inference, with reflections on machine learning," *Communications of the ACM*, vol. 62, no. 3, pp. 54–60, 2019.

[20] P. K. Lam, *Estimating individual causal effects*. PhD thesis, 2013.

[21] P. W. Holland, "Statistics and causal inference," *Journal of the American statistical Association*, vol. 81, no. 396, pp. 945–960, 1986.