Bayesian Networks & Causal Inference — Assignment 2

Daan Brugmans (s1080742)
Radboud University
Nijmegen, Netherlands
daan.brugmans@ru.nl

Maarten Berenschot (s1017215)
Radboud University
Nijmegen, Netherlands
maarten.beerenschot@ru.nl

ABSTRACT

This paper presents the execution of two causal inference analyses on the causal relationship between the success of a bank's past campaigns that advertised opening long-term deposits to their clients and the likelihood of a client from that bank opening a new long-term deposit. The causal relationship was analyzed and quantified using Covariate Adjustment analysis and Instrumental Variable analysis within the context of a Directed Acyclic Graph (DAG). The result of the Covariate Adjustment analysis proved fruitful for quantifying the causal relationship, while the results of the Instrumental Variable analysis proved to be potentially insufficient, unless further data preprocessing is applied to the data analyzed. These findings have given the authors greater insight into the mechanisms of the causal inference techniques used and the data analyzed.

1 INTRODUCTION

This paper contains a realization of the Assignment 2 project for the course "Bayesian Networks and Causal Inference" [4] offered by the Radboud University. The Assignment 2 project focuses on applying causal inference techniques on a dataset within the context of a Directed Acyclic Graph (DAG). In the Assignment 1 project, students of the course constructed a DAG and calculated the strength of causal relationships between all variables of the DAG. For the Assignment 2 project, a single causal relationship in the DAG is further explored: one variable is deemed the Exposure, another variable the Outcome, and the causal relationship between the Exposure and Outcome are studied using multiple causal inference techniques.

We, the authors of this paper, have already constructed a DAG for a dataset as per the Assignment 1 project. A more thorough description of our DAG, dataset, and findings may be found in the Assignment 1 paper [1], but are shortly given here. Our fulfillment of Assignment 1 and Assignment 2 use a preprocessed dataset taken from the works of Moro et al. [2] and is on the topic of bank marketing. Specifically, the dataset we use originates from research performed by Moro et al. on the topic of using bank marketing data to construct a prediction algorithm that could predict whether a customer is expected to subscribe to a long-term deposit at the bank or not. The results of this paper were aggregated in the dataset used for the projects Assignment 1 and Assignment 2. A DAG was constructed for the Assignment 1 project that modelled the causal relationships between the variables of the dataset by Moro et al.. Figure 1 shows this DAG.

For our realization of the Assignment 2 project, we analyze a specific causal relationship within the DAG of figure 1. We wish to further analyze the relationship between the likelihood that a banking customer will subscribe to a long-term deposit and the success of previous campaigns performed by the bank that promoted subscribing to long-term deposits. In the DAG of figure 1, these variables are respectively named "HasSubscribedToDeposit" and "PreviousCampaignOutcome". In alternative terms, we wish to analyze the influence of the success of previous campaigns on the success of the latest campaign. We take the "PreviousCampaignOutcome" variable to be the Exposure of the causal relationship and the "HasSubscribedToDeposit" variable to be the Outcome. We ask ourselves:

How can the causal relationship between the success of past banking campaigns promoting long-term deposits and the likelihood that a banking customer subscribes to a long-term deposit be quantified using causal inference techniques?

2 METHODOLOGY

In order to analyze the causal relationship between the "HasSubscribedToDeposit" and "PreviousCampaignOutcome" variables, we apply two causal inference techniques, namely **Covariate Adjustment** analysis and **Instrumental Variable** analysis. Both of these techniques aim to quantify a causal relationship between two variables. We performed these analyses using the R programming language and the dagitty [5] program. We provide the reader an elaboration on how we used these techniques in the rest of this chapter.

For our analysis, we used the fully preprocessed dataset and the final tested DAG from Brugmans and Berenschot [1]. We did not apply any further preprocessing to the dataset for our IV analysis, nor did we alter the structure of the DAG. However, we did remove any existing edge coefficients of the DAG in Assignment 1 and defined the Exposure and Outcome variables for the causal relationship we wish to analyze. Furthermore, we performed IV analysis on the dataset multiple times with different sets of outliers removed to measure the effect of the outliers on the analysis result. This will be discussed later. The setup for our analyses is reflected in the DAG in figure 1.

2.1 Covariate Adjustment

In order to perform a Covariate Adjustment analysis on the causal relationship between our Exposure and Outcome, we started by finding an appropriate adjustment set. An adjustment set may be found by applying the Back-Door Criterion [3]: given a DAG and its dataset, the adjustment set for the causal relationship between the Exposure and Outcome is the set of variables in the DAG that d-separate all paths from the Exposure to the Outcome, starting with the Exposure's incoming paths, excluding variables that are descendants of the Exposure. Within the context of our DAG and for our choice of Exposure and Outcome, such an adjustment set exists. It comprises of only one variable, "JobCategory".

BNCI'23-'24, Radboud University, Nijmegen, Netherlands
2023.

We then constructed two Generalized Linear Models (GLMs) with binomial distributions, one for an unadjusted Outcome and one for an adjusted Outcome. We do this in order to model the Outcome in the cases where we do and do not adjust for covariates. Our reasoning towards the choice of a GLM is that the logistic link function is frequently used to model the relationship between predictors and the log-odds of success. We extracted the coefficient of the GLMs and took it to be the coefficient of the causal relationship from the Exposure to the Outcome. Furthermore, we calculated the 95%-confidence interval and z-values for these coefficients. The results of our analysis may be found in figure 2.

2.2 Instrumental Variable

In addition to a Covariate Adjustment analysis, we perform an Instrumental Variable (IV) analysis. This type of causal inference analysis requires that the data analyzed contains a variable that can be taken as an Instrumental Variable for a given causal relationship. Such a variable must adhere to the assumptions of exogeneity and exclusion restriction. This means that, in the DAG, the candidate IV cannot have any incoming paths from other variables, only outgoing, and its outgoing paths to the Outcome must all traverse through the Exposure. Fortunately, as figure 1 shows, our DAG contains an IV candidate variable for the causal relationship we wish to analyze, that variable being "PreviousCampaignsCalls".

Since the "PreviousCampaignsCalls" variable is a non-conditional IV, we did not condition on any variables with paths towards the IV for our debiasing procedure. In order to come to an unbiased IV coefficient, we first regressed the Exposure onto the IV, resulting in a linear regression model. The predictions made on this regression were used to construct the "Adjusted Exposure", the Exposure adjusted for the IV. The Outcome was then regressed onto the Adjusted Exposure. The coefficient of the Adjusted Exposure of this regression was taken to be the coefficient of the causal relationship between the Exposure and Outcome. This coefficient, alongside its 95%-confidence interval, is given in figure 3.

3 RESULTS

Figure 1 shows the DAG that our causal inference analyses were based upon.

Figure 2 shows the results of our Covariate Adjustment analysis. When adjusted, we found that the path between the Exposure and Outcome attains a coefficient of 2.70 with a 95%-confidence interval of (2.59, 2.82). When unadjusted, we found that the coefficient estimated to 2.81 (2.70, 2.92). The estimated coefficients were found to be statistically significant with $Pr(>|z|) < 2e-16$.

Figure 3 shows the results of our Instrumental Variable analysis. When no outliers are excluded and the full dataset is taken into account, we came to a coefficient of -0.061, or more precisely, -0.06094688. The 95%-confidence interval for this coefficient was $(-0.067, -0.055)$, more precisely $(-0.06718429, -0.05470946)$. We may put these coefficients into the context of the regression by analysing the Residuals against the Fitted Values plot of figure 3. This plot shows a single major outlier for the data. When this data point is removed from the dataset, the IV coefficient changes to -0.062. If all outliers are removed, then no IV coefficient can be estimated.

4 DISCUSSION

The results of our Covariate Adjustment analysis meet our expectations, as we expected the estimated coefficients to be positive. The results show that the unadjusted model overestimates the coefficients by a factor of four percent. We find this difference to be minor and expect that the covariate adjustment does not noticeably influence the causal relationship between the Exposure and Outcome.

The results of the Instrumental Variable analysis ought to be discussed. The estimated coefficient is near-zero, which goes against our initial expectations, which was that the coefficient would show a noticeable correlation between the Exposure and Outcome. We attempt to explain this gap between expectation and result by analyzing the Residuals against the Fitted Values plot of the IV regression, which shows some interesting properties:

- The data points follow a linear relationship where the residuals decrease as the fitted values increase;
- Although all data points adhere to a certain linear relationship, there are clearly two trend lines that may be drawn in the graph;
- These is a single major outlier that extends the range of the data points notably.

We suspect that many of these properties can be explained by elaborating on the distribution of the IV's data.

The IV described the amount of times that a banking customer was called by the bank as part of one of the bank's past long-term deposit campaigns. The value distribution of the IV consists of near exclusively 0. In other terms, almost all banking customers had not been contacted in a prior campaign. Because almost all data points for the IV are 0, any value that is not 0 is considered a statistical outlier. This means that all customers who had been called in a previous campaign are, statistically, outliers. The major outlier of the plot in figure 3 is a person who was called 275 times, meaning that they are also a major outlier in the original data. If this single outlier is removed, indicated by "Single Major Outlier Removed" in figure 3, the IV coefficient slightly changes. As a consequence of the fact that all people who were called at least once by the bank are considered statistical outliers, should we remove the outliers from the data, then the only data points that remain are customers who had never been called: the IV without outliers only has data with the same value (0). Because the distribution of the IV without outliers is 100% 0, the IV cannot be meaningfully used for IV analysis. This fact is reflected in figure 3, where no coefficients can be provided when all outliers are removed.

We did not expect to run into this problem when selecting the IV. It may explain the reason why the estimated IV coefficient is so close to zero: almost all values of the IV are zero themselves, and only outliers give meaningful information. This information imbalance is carried over into the estimated coefficient. To mitigate this, we suggest a rebalancing of the data on the IV: a set of people who had never been called for a campaign in the past should be excluded from the dataset, such that the rebalanced dataset both balances the IV distribution better and maintains the representativeness of the unbalanced dataset. Data may also be imputed using the distribution of the IV or the "CurrentCampaignCalls" variable, which, within the context of the dataset, is the same as the IV, but for the latest

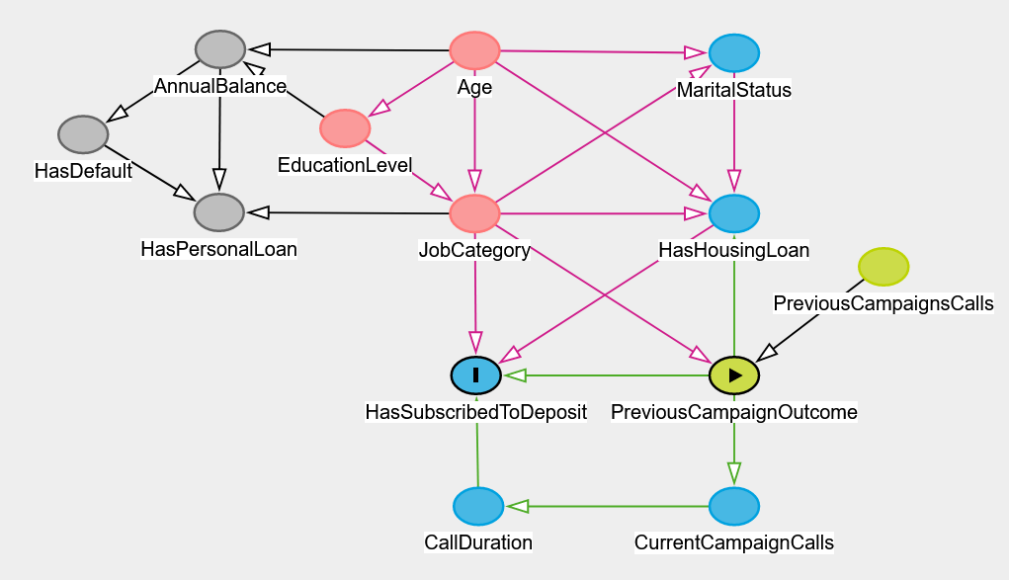
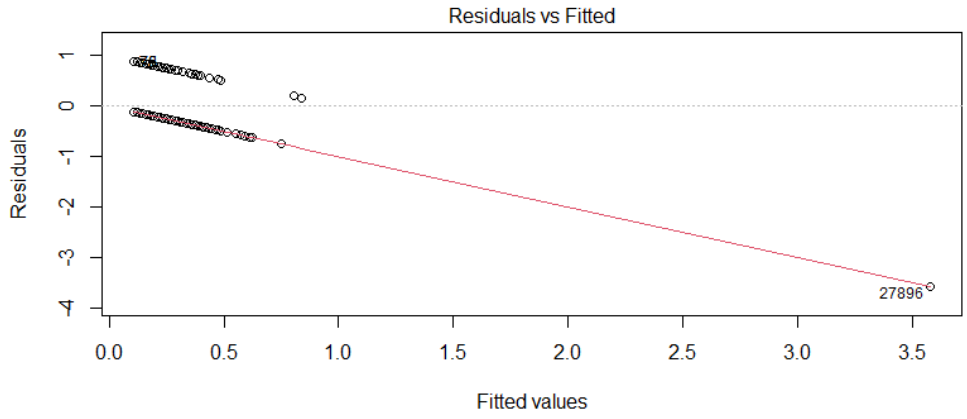


Figure 1: DAG of Brugmans and Berenschot [1] modelled in the dagitty [5] program. Exposure and Outcome are marked with symbols, with the Exposure in yellow and the Outcome in blue.

Estimated Covariate Adjustment Coefficients

Adjustment	**2.5%**	**median**	**97.5%**	$Pr(>\|z\|)$
Adjusted	2.59	2.70	2.82	$< 2e-16$
Unadjusted	2.70	2.81	2.92	$< 2e-16$

Figure 2: Results of our Covariate Adjustment analysis. The 95%-confidence interval and z-values are given for the coefficients.



Estimated Instrumental Variable Coefficients

Outliers	$\mu - 2\sigma$ **(2.5%)**	μ **(median)**	$\mu + 2\sigma$ **(97.5%)**
All	−0.067	−0.061	−0.055
Single Major Outlier Excluded	−0.067	−0.062	−0.057
None	—	—	—

Figure 3: The estimated IV coefficients between the Exposure and Outcome. The 95%-confidence interval is included for all coefficients. A plot of the IV regression of the Outcome onto the Adjusted Exposure is also included for the case where all outliers are included, showing the residuals of the regression line onto the fitted values.

campaign. This imputation should maintain the representativeness of the dataset, however.

As an answer to our research question, we found that both Covariate Adjustment analysis and Instrumental Variable analysis are valid causal inference techniques that can be applied to our Exposure and Outcome. However, we learned that although our choice of IV may be valid, it may not be well-suited as an IV, due to the imbalance of the IV's data. We have not encountered such an issue during our Covariate Adjustment analysis. It is for this reason that we think that, between the two causal inference analyses, the results for the Covariate Adjustment are more plausible and reflect the dataset more accurately. This conclusion is supported by our presumption that the relationship between our Exposure and Outcome should be positive: we expected the success of previous campaigns to positively influence the success of the current campaign.

REFERENCES

[1] Daan Brugmans and Maarten Berenschot. 2023. Bayesian Networks Assignment 1. (nov 2023).

[2] Sérgio Moro, Paulo Cortez, and Paulo Rita. 2014. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems* 62 (2014), 22–31. https://doi.org/10.1016/j.dss.2014.03.001

[3] Judea Pearl. 2009. *Causality.* Cambridge University Press, Cambridge, UK. https://doi.org/10.1017/CBO9780511803161

[4] Johannes Textor, Ankur Ankan, Ioan Gabriel Bucur, Wieske de Swart, and Marco Loog. 2023. NWI-IMC012 Bayesian Networks and Causal Inference. https://www.ru.nl/courseguides/science/vm/osirislinks/imc/nwi-imc012/

[5] Johannes Textor, Benito van der Zander, Mark S Gilthorpe, Maciej Liśkiewicz, and George TH Ellison. 2017. Robust causal inference using directed acyclic graphs: the R package 'dagitty'. *International Journal of Epidemiology* 45, 6 (01 2017), 1887–1894. https://doi.org/10.1093/ije/dyw341 arXiv:https://academic.oup.com/ije/article-pdf/45/6/1887/11120744/dyw341.pdf