AMERICAN CAUSAL INFERENCE CONFERENCE DATA CHALLENGE 2022

Alexandre Monti

Supervised by Amit Sawant and Prof. Mats Stensrud

Contents

1 Introduction to the challenge **2**

1.1 How the datasets were generated . . . 2
1.2 How the datasets are structured . . . 3
1.3 Objective of this challenge . . . 3
1.4 Assumptions for the data generating process (DGP) to make the estimands identifiable . . . 3

2 Exploratory Data Analysis **5**

2.1 First oberservations . . . 5
2.2 How the covariates are correlated . . . 5
2.3 Link between X1, X2, X3, X4, X5, X6, X7, Z and year. . . . 7
2.4 Link between X1, X2, X3, X4, X5, X7, Z, year and V1_avg, V3_avg, V4_avg, Y . . . 10
2.5 Y against Z for each year . . . 12
2.6 Link between categorical variables . . . 12
2.7 Data transformation . . . 13
2.8 Monthly average medical expenditure (Y) against each numeric variables colored by year. . . . 14
2.9 Monthly average medical expenditure (Y) against each numeric variables colored by Z . . . 14
2.10 What we know for now . . . 15
2.11 Conditional independence between categorical variables . . . 15
2.12 Conditional independence between numerical variables . . . 17
2.13 Relation between V1_avg, V3_avg, V4_avg and Y . . . 19
2.14 Conditional independence between categorical variables and numerical variables . . . 21

3 Directed Acylcic Graphs **22**

3.1 general DAG . . . 22

4 Monthly average medical expenditures without the effect of time **23**

4.1 Model for monthly average medical expenditures . . . 23
4.2 Link between Y_mult and numeric variables . . . 26
4.3 Link between Y_mult and categorical variables . . . 26
4.4 Time-invariance of the relations between variables . . . 27
4.4.1 Categorical variables . . . 27
4.4.2 numeric variables . . . 28

5 Verifications **30**

5.1 Dependence and independence . . . 30
5.2 Conditional dependence and independence . . . 31

6 Conclusion **32**

7 References **33**

1 Introduction to the challenge

The ACIC data challenge is a challenge in which we want to estimate causal impacts in simulated datasets based on real-world data from fields such as health care or education. This year, it focuses on the effect of a treatment on monthly average medical expenditures of patients.

Here are some interesting comments on the datasets. These comments comes from the website of the challenge.

Explanations for the given datasets :

- Participation in the intervention is not randomly assigned, making it hard to tell whether the patients or practices that were motivated to join the intervention were already on a favorable trajectory or truly benefited from the program.
- Impacts can be highly heterogeneous, with the intervention potentially lowering expenditures for some patients while increasing expenditures for others.
- The data have a hierarchical structure, with repeated observations of patients over time, and with patients clustered in primary care practices.
- The outcome of interest, Medicare spending, is highly variable and skewed. The average patient incurs around 1,000 in costs per month, but some patients incur tens of thousands of dollars in costs.

Track 1 is a patient-level track. It is very big. Track 2 is a practice-level track which took the patient level track 1 and averaged it up to the practice-year level. It is smaller to simplify the calculations. "Each dataset in Track 2 was calculated by taking means of patient covariates and outcomes from the corresponding dataset in Track 1".

Each track consists of 200 realizations each of 17 data generating processes (DGPs). Note that each realization of the DGP creates a new sample of practices, meaning that practice 1 in the first realization is not the same practice as practice 1 in the second realization. The former could be a small, rural, high-spending practice, whereas the latter might be a large, urban, low-spending practice.

1.1 How the datasets were generated

" A dataset from an evaluation of a real intervention informs the DGP. We have kept the intervention generic for the purposes of this competition to avoid inadvertently disclosing any sensitive information. Although the DGPs use summaries of this real dataset, covariates, outcomes, and treatment assignment are all simulated: the observations in the simulated datasets do not correspond directly to observations in the real data at either the patient or practice level.

The marginal and joint distributions of many of the covariates are constructed to mirror those in the real Medicare dataset and are based on summaries of the real dataset such as empirical CDFs, relationships between covariates, and intra-class correlations. The datasets also contain constructed covariates that do not have a counterpart in the real data, to give us fine-grained control over the degree of confounding. All covariates are held fixed at their baseline values; importantly, this implies that the intervention does not affect any covariates. The marginal distribution of the outcome is calibrated to mirror that of the real dataset. Both

the treatment and response surfaces are constructed and do not necessarily correspond to features of the real data (beyond the marginal distribution of the outcome)."

1.2 How the datasets are structured

"The data have a longitudinal structure, with repeated annual observations of patients over time. Although patients enter and leave the sample throughout the four-year observation period, all primary care practices are observed for all four years. The first two years are a baseline period during which no one receives the intervention. Then, at the beginning of Year 3, treated practices begin receiving the intervention. They continue receiving the intervention through the end of Year 4."

1.3 Objective of this challenge

From the thousands files that are given in the track 2 folder, we will only focus on one dataset to really learn and understand how to make data analysis. We thus choose the first dataset of the track 2 folder. To have all the information of the dataset, we merge the first file of the "practice" folder with the first file of the "practice-year" folder as it is mentioned in the instructions of the challenge.

Our aim is to estimate the average effect of the intervention on monthly spending per patient (ATT). The formula for ATT is given by

$$ATT = E(Y^{z=1} - Y^{z=0}) = E(Y^{z=1}) - E(Y^{z=0})$$

For that, we will first make a DAG representing the experiment and the link between the covariates. We will then begin with a part of exploratory data analysis to find as many as possible relations of dependence and independence between the covariates.

1.4 Assumptions for the data generating process (DGP) to make the estimands identifiable

conditional exchangeability :

$$Y_{t_1} - Y_{t_0} \perp\!\!\!\perp Z|X_i \quad \forall t_1 \text{ such that } p_t = 1, t_0 \text{ such that } p_t = 0$$

Where X_i are the five categorical and binary variables $X1, \ldots, X5$.

"This means that there are no unobserved covariates that relate to both treatment assignment and to the change in the untreated potential outcome $Y^{Z=0}$ from the period before the intervention (Years 1 and 2) to the intervention period (Years 3 and 4)".

Positivity :

$$P(z = 1|X_i) > 0 \quad \forall i \mid z = 1$$

SUTVA : stable unit treatment value assumption, i.e. there is only one form of treatment for all the treated individuals and that the outcome does not depend on the treatment assigned to other individuals.

We have now all the information to begin to work. We took the decision to only work on a subset of the datasets as I don't have any background knowledge on causal inference. Here is a small reminder of what each column corresponds to :

- id.practice
- X1, categorical variable with two levels (0 and 1)
- X2, categorical variable with three levels (A, B and C)
- X3, categorical variable with two levels (0 and 1)
- X4, categorical variable with three levels (A, B and C)
- X5, categorical variable with two levels (0 and 1)
- X6, X7, X8, X9 numerical variables
- year, categorical variable with four levels (1, 2, 3 and 4)
- Y, the monthly average medical expenditures, numerical variable
- Z, indicator for treatment status, categorical variable with two levels (0 and 1)
- post, the post period, categorical variable with two levels (0 and 1)
- V1_avg, V2_avg, V3_avg, V4_avg, V5_A_avg, V5_B_avg, V5_C_avg numerical variables
- n.patients, the number of patients, numerical variable

V1_avg,...,V5_C_avg are averages of patient-level covariates (V1,...,V5) for all patients belonging to the practice for the year.

Let's begin the exploratory data analysis !

2 Exploratory Data Analysis

Exploratory data analysis is needed to understand how the covariates are linked. We will begin with simple data analysis to find the dependence/independence between the covariates and then we will focus on conditional dependence. This will help us to draw a nice directed acyclic graph (DAG) for the causal inference part.

2.1 First oberservations

Apparently, there is 816 treated rows and 1184 untreated rows, which correspond to 204 treated practices and 296 untreated practices as we have one row by year for each practice. We can also see that there is exactly the same number (1000) of $post = 0$ and $post = 1$. This means that we misinterpreted the data at the beginning as we considered that only the treated practices ($z = 1$) could have $post = 1$ which would have meant that this practice began the treatment in year 3-4. With the same reasoning, we thought that if the practice was untreated ($z = 0$), *post* would be always zero because if we are untreated, we cannot begin any treatment.

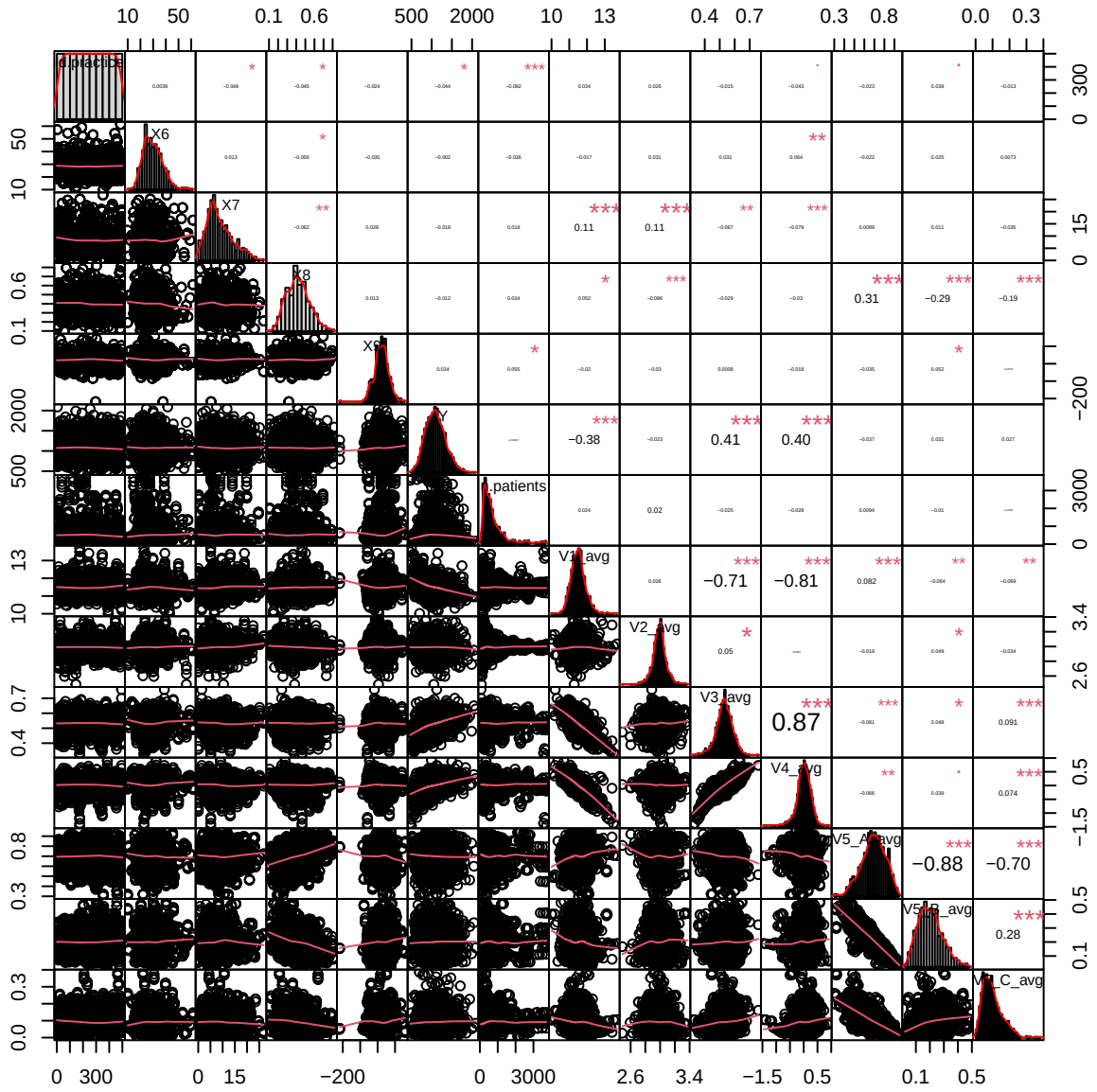
But this interpretation is obviously wrong as we have 184 untreated rows with $post = 1$ (46 practices) which means that this is not just some measurement bias. In fact, the *post* variable seems to only indicate if we are in year 1-2 or 3-4 no matter if we are a treated or an untreated practice. This explain why there is exactly 1000 rows with $post = 0$ and $post = 1$ as we have 500 practices in our datasets but there is always 4 data for each practice, one data per year. This interpretation explains why we can have untreated practices with $post = 1$ because it just means that we are in year 3 or 4. Thus, we can interpret the variable *post* as an indicator for whether we are in year 1-2 or in year 3-4.

Moreover, we learn from the website FAQ that only practices with $z = 1$ and $post = 1$ are actually treated which seems pretty obvious as the the intervention can only begin in year 3-4.

2.2 How the covariates are correlated

Thanks to a nice code found on the web[1], we can plot all the scatterplots with a numeric covariate on one axis and another on the other axis. This code gives us also the correlation between the numeric covariates as well as the histogram for each covariate with its kernel density.

[1] https://www.r-bloggers.com/2012/08/more-on-exploring-correlations-in-r/



This scatterplot matrix gives us a lot of information. We can see that

- V1_avg is correlated with V3_avg ($r = -0.71$)
- V1_avg is correlated with V4_avg ($r = -0.81$)
- V3_avg is correlated with V4_avg ($r = 0.87$)
- V5_A_avg is correlated with V5_B_avg($r = -0.88$)
- V5_A_avg is correlated with V5_C_avg ($r = -0.70$)
- X8 is correlated with V5_A_avg, V5_B_avg and V5_C_avg ($r = 0.31$, $r = -0.29$, $r = -0.19$)
- Y is correlated with V1_avg, V3_avg and V4_avg ($r = -0.38$, $r = 0.41$, $r = 0.46$)

From this, we can infer that V1_avg, V3_avg, V4_avg and Y are all correlated.

We can also say that V5_A_avg, V5_B_avg and V5_C_avg are all correlated but the correlation between V5_B_avg and V5_C_avg is not as strong as the others correlations.

We will come back later on these relations.

2.3 Link between X1, X2, X3, X4, X5, X6, X7, Z and year.

After making all the data analysis, we realized that some covariates were not useful at all for the aim of making a DAG. We will then concentrate our attention to the subset of covariates which have a link with the monthly average medical expenditures for convenience and brievety. This subset is composed of X1 to X7, V1_avg, V3_avg, V4_avg, Z and year.

We focus here on discover the relations between the categorical variables X1, X2, X3, X4, X5, Z and the numeric variables X6, X7 and Y.

By plotting the density curve of a numeric variable w.r.t. levels of a categorical variable, we can determine if these two variables are dependent or independent. Indeed, if the density curve is the same for each level of the categorical variable, it means that the latter has no effect on the numeric variable, i.e. they are independent.

We keep only plots of X6 for brievety.

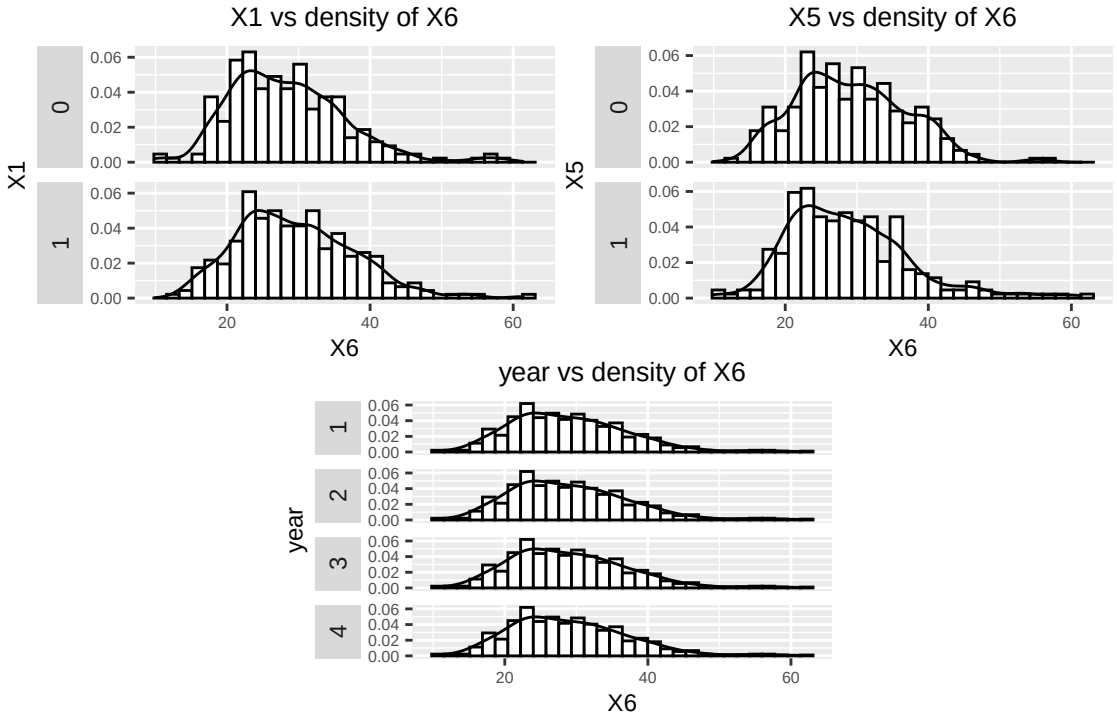


FIGURE 1: The plot of X6 in function of levels of X1 seems to show an independence relation between them. Indeed, $P(X6 \mid X1 = 1) = P(X6 \mid X1 = 0) = P(X6)$. By the same reasoning, we can conclude that X6 is independent of X5 and year.

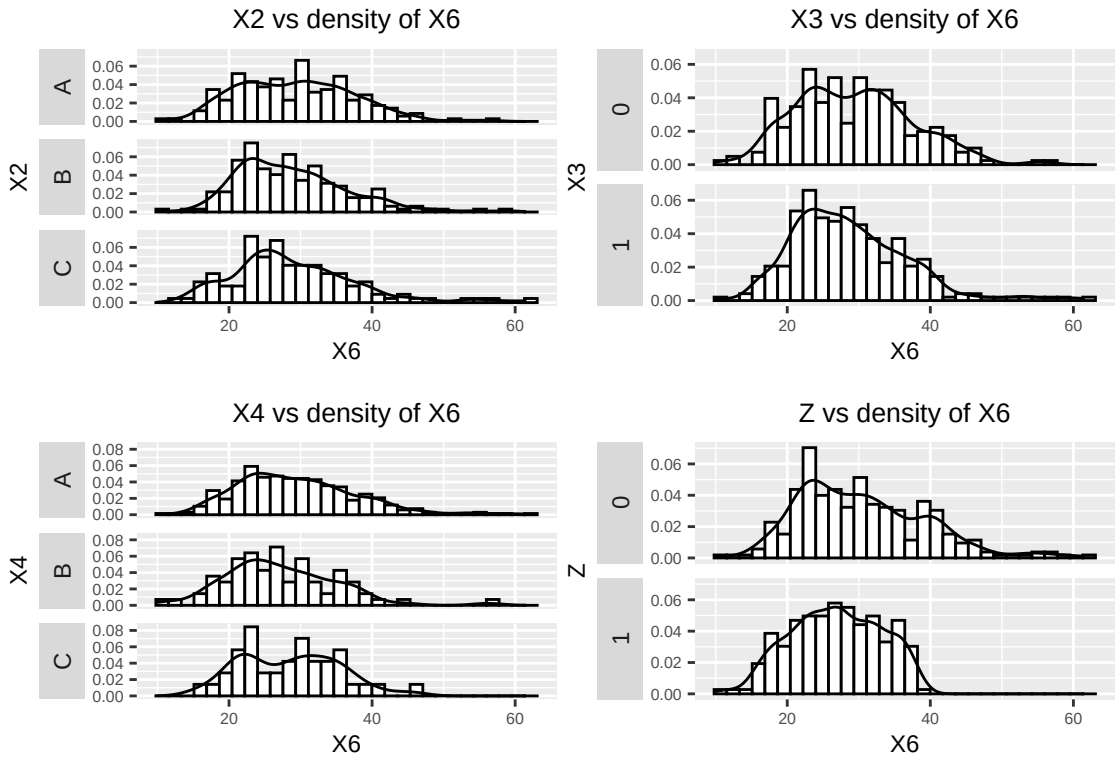


FIGURE 2: Here, the densities of X6 change a lot w.r.t. levels of the categorical variables, i.e. $P(X6 \mid X) \neq P(X6)$ for $X = X2, X3, X4, Z$. This means that X6 is dependent with X2, X3, X4 and Z.

If there is no comment between two variables, it means that we found no independence. Thus, they are dependent.

The dependence between X3 and X6 seems pretty weak. To verify if there is dependence or not, we compute percentage bar chart between the categorical variable and a discretized version of X6. Let's see if this is really a dependence relation :

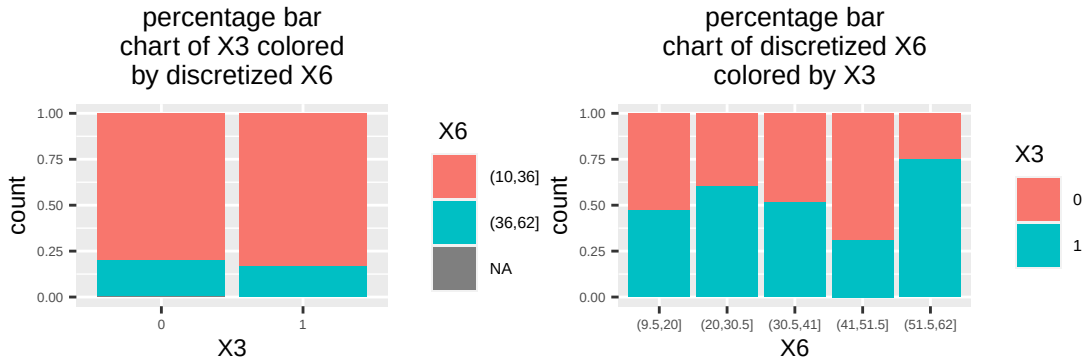


FIGURE 3: As we can see on the first plot, each bar is really similar to the others, i.e. $P(X6 \mid X3 = 0) = P(X6 \mid X3 = 1)$. We could then conclude that X6 is independent of X3, but when we take a look at the second plot, we can see that when we switch the x-axis and y-axis, the independence relation is not so obvious. Thus, what can we conclude from these plots ? We can see on the graph that the first three levels of discretized X6 are pretty similar. We only have a problem with the last two levels of X6. But these

two last levels of X6 contain only 148 observations out of the 2000 that we have. Thus, we can consider these 148 observations as outliers and conclude that X3 and X6 are in fact independent.

We focus now on the relations for X7. By making the same analysis as for X6, we can conclude that X7 is dependent of X1 and X5 and independent of X3 and year. The dependence between X7 and X2,X4 and Z seems pretty weak. To verify if there is dependence or not, we compute percentage bar chart between each categorical variable and a discretized version of X7 as before :

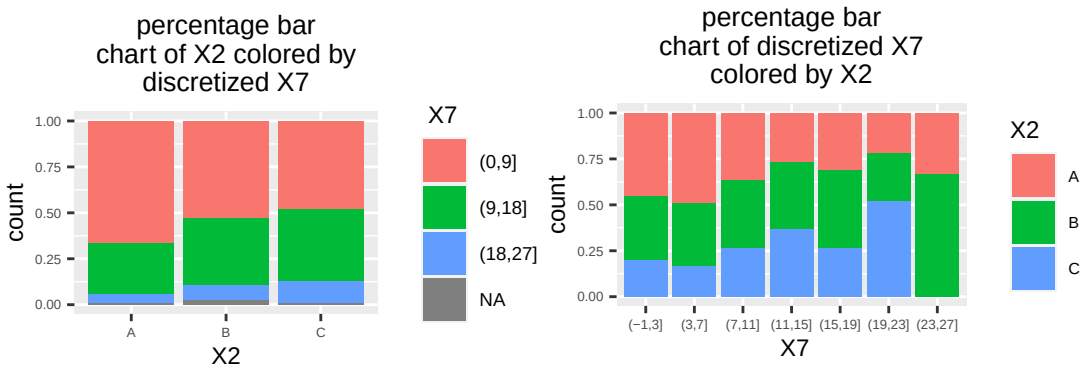


FIGURE 4: Each bar chart is distributed with the same trend except the bars for X7 in (19,23] and (23,27]. But there are only 116 observations of X7 in these interval, thus we conclude that X7 and X2 are weakly dependent.

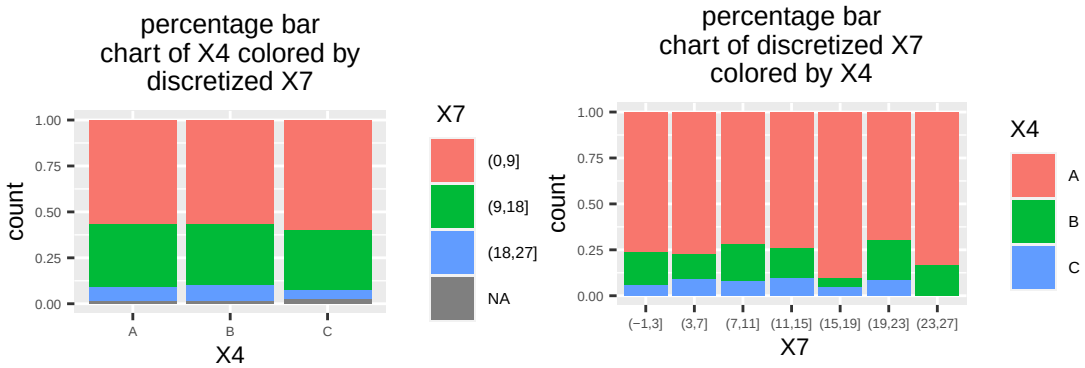


FIGURE 5: Each bar chart is distributed with the same trend. We conclude that X7 $\perp\!\!\!\perp$ X4.

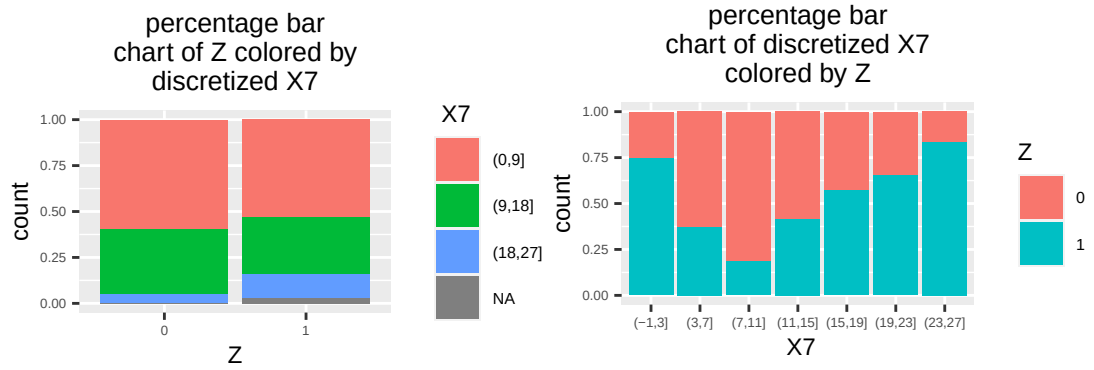


FIGURE 6: The first plot indicates an independence relation while the second plot indicates a dependence relation. Here, every bar is different so we conclude that X7 is dependent with Z.

2.4 Link between X1, X2, X3, X4, X5, X7, Z, year and V1_avg, V3_avg, V4_avg, Y

We have found the first relations between categorical variables and X6, X7. We continue our analysis with the same method to discover relations between V1_avg, V3_avg, V4_avg, Y and the categorical variables. We only keep plots of X1 for brievety.

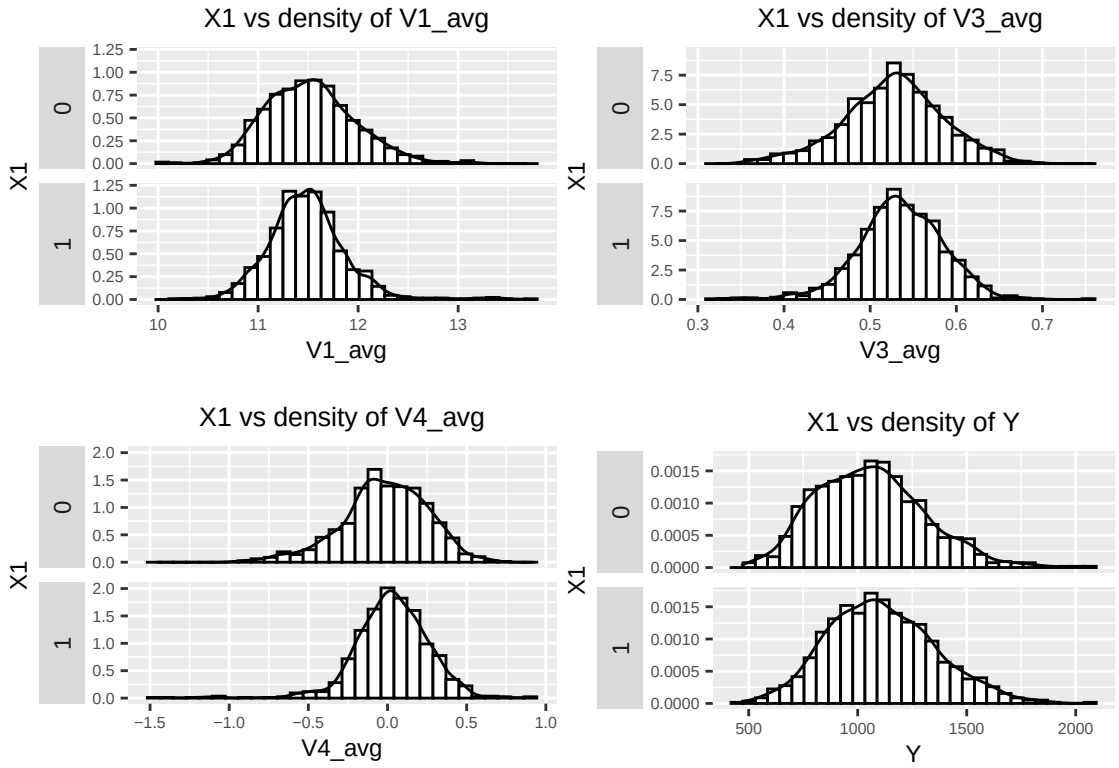


FIGURE 7: X1 seems weakly dependent of V1_avg and V4_avg. On the other hand, it is cleary independent of V3_avg and Y.

We resume here the relations that we found by making the same analysis :

- for X2 : weakly dependent of V3_avg and V4_avg, independent of V1_avg and Y
- for X3 : weakly dependent of V3_avg and V4_avg, independent of V1_avg and Y
- for X4 : weakly dependent of V1_avg, V3_avg and V4_avg, independent of Y
- for X5 : weakly dependent of V3_avg and V4_avg, independent of V1_avg and Y
- for Z : dependent of Y, independent of V1_avg, V3_avg and V4_avg
- for year : dependent of Y, weakly dependent of V1_avg, independent of V3_avg and V4_avg

Now we want to know if the weak dependencies are really dependence relations or not. To verify this, we use the same method as before by discretizing the numeric variable and computing a percentage bar chart colored by the categorical variable. If the percentage bar charts are colored the same, it means that our weak dependence is in fact an independence relation.

Again, we only keep plots of X1 for brievety.

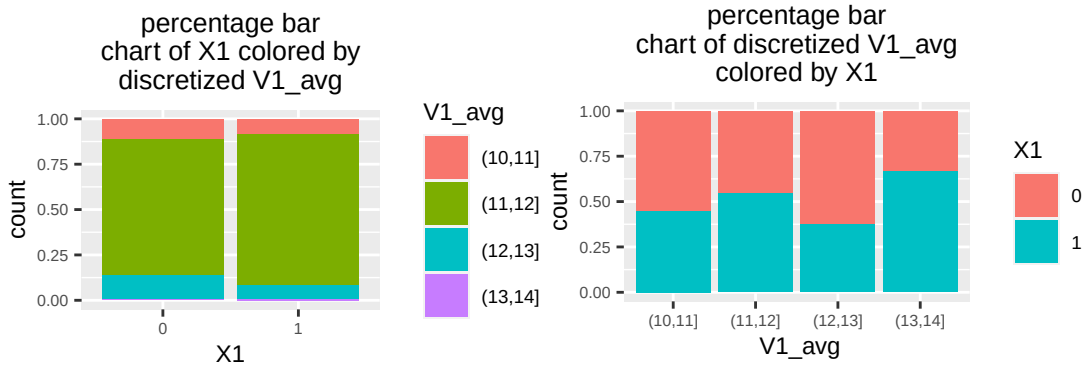


FIGURE 8: The first plot shows an independence relation. The second plot is more ambiguous but there is only 12 observations of V1_avg greater than 13 and the first three bar show the same trend. Thus, X1 $\perp\!\!\!\perp$ V1_avg.

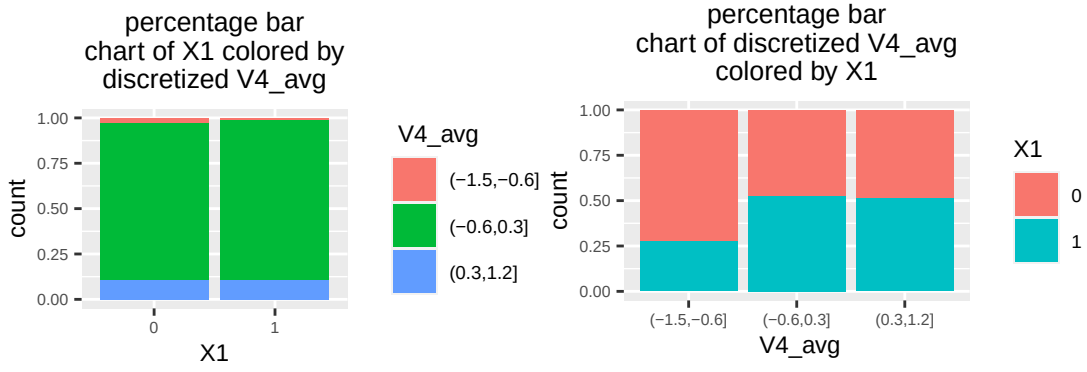


FIGURE 9: The first plot shows independence. The second plot is also more ambiguous but there is only 40 observations of V4_avg smaller than -0.6. Thus, X1 is weakly dependent with V4_avg.

By the same reasoning, we find out that :

- X1 is independent of V1_avg, V3_avg, V4_avg and Y
- X2 is independent of V1_avg, V3_avg, V4_avg and Y
- X3 is dependent of V3_avg, independent of V1_avg, V4_avg and Y
- X4 is dependent of V1_avg and V3_avg, independent of V4_avg and Y
- X5 is independent of V1_avg, V3_avg, V4_avg and Y
- Z is dependent of Y, independent of V1_avg, V3_avg and V4_avg
- year is dependent of Y, independent of V1_avg, V3_avg and V4_avg

2.5 Y against Z for each year

Here, we just verify that treatment has an effect on Y.

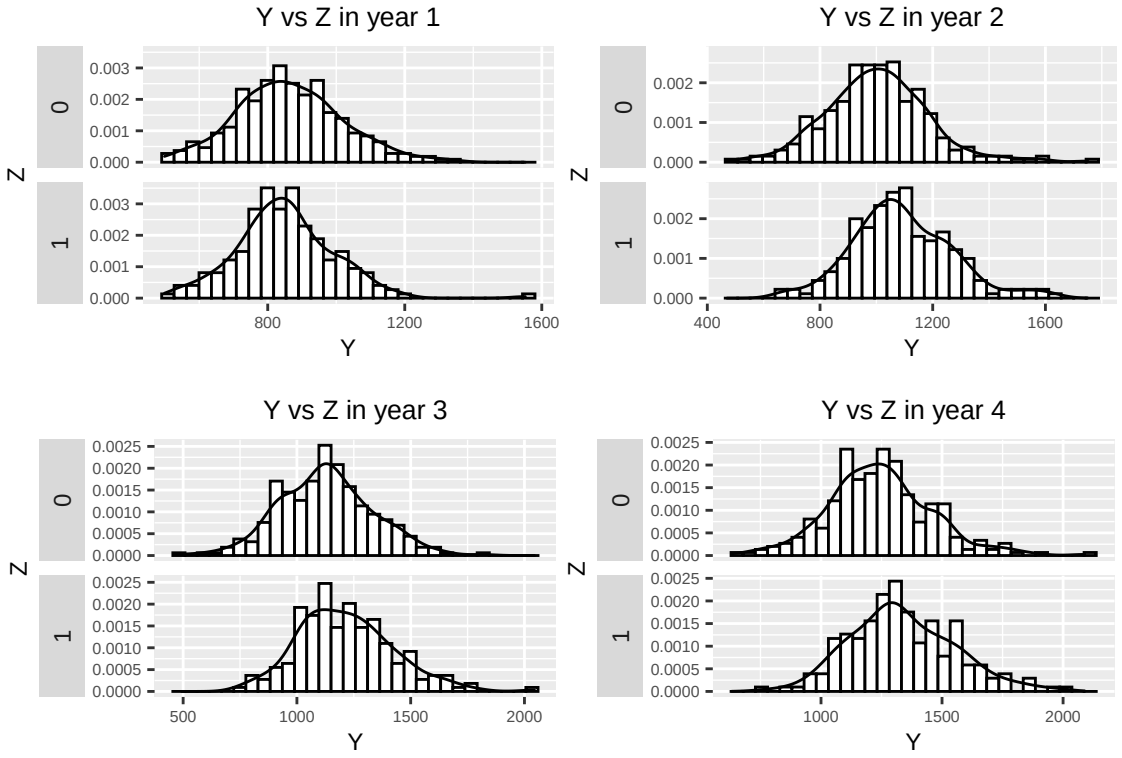


FIGURE 10: As $P(Y \mid Z = 0) \neq P(Y \mid Z = 1)$ for each year, we can conclude that Y depends on Z.

2.6 Link between categorical variables

We would like to discover the relations between categorical variables. For this aim, we plot the bar chart of a categorical variable colored w.r.t. levels of another. If each bar has same proportions of each level of the other categorical variable, we can assume that they are independent.

We only keep plots of X1 for brievety.

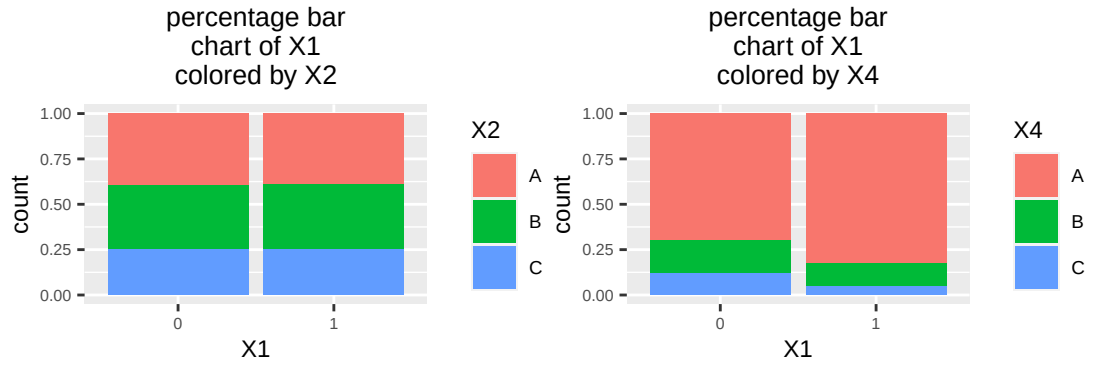


FIGURE 11: We see on the first plot that $P(X2 \mid X1 = 0) = P(X2 \mid X1 = 1)$. We conclude that X1 $\perp\!\!\!\perp$ X2. This conclusion holds for X3, X5, year and Z. On the other hand, $P(X4 \mid X1 = 0) \neq P(X4 \mid X1 = 1)$. Thus, X1 and X4 are dependent.

By the same reasoning, we find out that

- X2 is dependent with X3, X4, X5 and independent of Z and year
- X3 is dependent with X2, X5 and independent of X1, X4, Z and year
- X4 is dependent with X1, X2 and independent of X3, X5, Z and year
- X5 is dependent with X1, X2, X3 and independent of X4, Z and year
- Z is independent of all categorical variables
- year is independent of all categorical variables

2.7 Data transformation

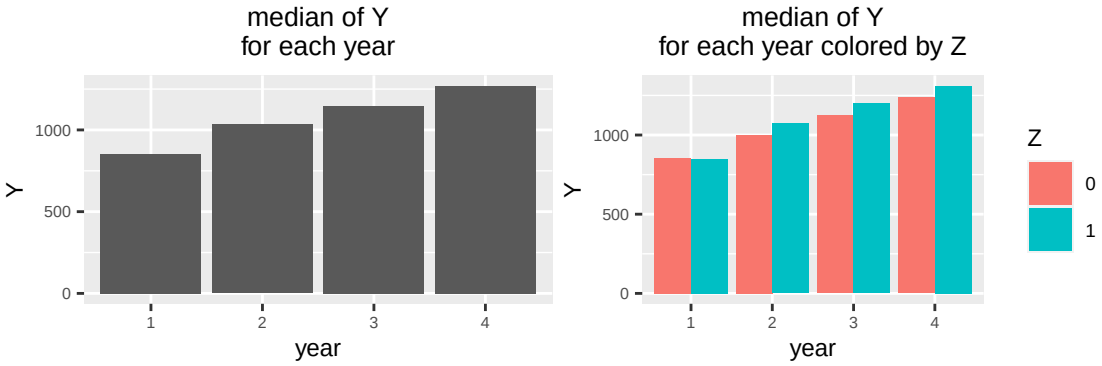


FIGURE 12: There are some things to say about the grouped bar charts. First of all, we can see that the monthly average medical expenditures increase over year. Furthermore, when we color by Z, we can see that $Z = 1$ is more expensive for the last three year, which is interesting as people were treated to decrease their medical expenditures. But we cannot conclude now that the treatment has no effect. Indeed, the treatment could have an effect such that the median of Y in year 2,3 and 4 is greater than non-treated people, but this median would be much greater without treatment. We need to investigate this hypothesis.

We do not plot Y against year colored by other categorical variable because we will see later that there is in fact no dependence between Y and these latter.

2.8 Monthly average medical expenditure (Y) against each numeric variables colored by year.

We would like to understand better the relation between Y and the numeric variables through year. We already plotted each numeric variable against Y in the plots/correlations matrix, but these plots were "time-invariant". We thus plot Y against numeric variables colored by year to have a better idea of this time-varying relation.

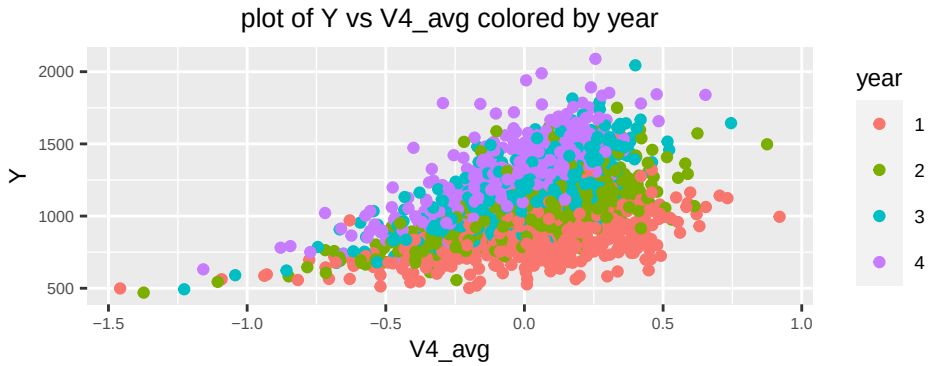


FIGURE 13: No matter which covariate we take for the x-axis, the monthly average medical expenditures becomes greater every year. Here we plotted Y against V4_avg to give an example of this relation.

2.9 Monthly average medical expenditure (Y) against each numeric variables colored by Z

Now that we understand better the link between numeric variables and Y through year, we would like to see if treatment has a significant effect on monthly average medical expenditure. For this aim, we plot Y against each numeric variables colored by Z.

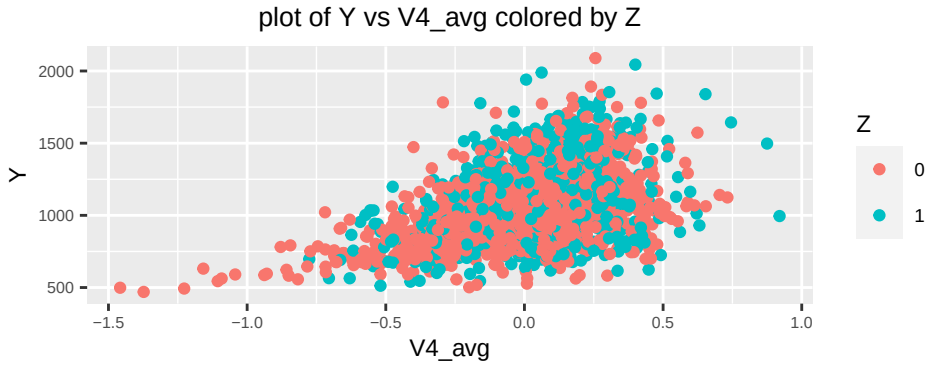


FIGURE 14: There is no trend when we color plots by Z. This seems to indicate that there is no obvious relation between "being treated" and "having low medical expenditures". We keep only the plot of Y against V4_avg as there is no other trend when we replace V4_avg with another numeric variable.

2.10 What we know for now

We represent by a table all the relations between the covariates that we found until now. The table is obviously symmetric. I is for independence and D for dependence.

	X1	X2	X3	X4	X5	X6	X7	Z	year	V1_avg	V3_avg	V4_avg	Y
X1	x	I	I	D	I	I	D	I	I	I	I	I	I
X2	I	x	D	D	D	D	D	I	I	I	I	I	I
X3	I	D	x	I	D	I	I	I	I	I	D	I	I
X4	D	D	I	x	I	D	I	I	I	D	D	I	I
X5	I	D	D	I	x	I	D	I	I	I	I	I	I
X6	I	D	I	D	I	x	I	D	I	I	I	I	I
X7	D	D	I	I	D	I	x	I	I	I	I	I	I
Z	I	I	I	I	I	D	I	x	I	I	I	I	D
year	I	I	I	I	I	I	I	I	x	I	I	I	D
V1_avg	I	I	I	D	I	I	I	I	I	x	D	D	D
V3_avg	I	I	D	D	I	I	I	I	I	D	x	D	D
V4_avg	I	I	I	I	I	I	I	I	I	D	D	x	D
Y	I	I	I	I	I	I	I	D	D	D	D	D	x

Our aim now is to determine some conditional dependence/independence. We recall that if $X \perp\!\!\!\perp Y, Z$ then $X \perp\!\!\!\perp Y \mid Z$ and $X \perp\!\!\!\perp Z \mid Y$. Furthermore, if $X \perp\!\!\!\perp Y$ and $X \perp\!\!\!\perp Y \mid Z$ then $Y \perp\!\!\!\perp X \mid Z$.

Thus we can already say that

- $year \perp\!\!\!\perp X_i \mid X_j$ for all $i, j \in \{1, .., 7\}$
- $X_i \perp\!\!\!\perp year \mid X_j$ for all $i, j \in \{1, .., 7\}$
- $Z \perp\!\!\!\perp year \mid X_i$ for all $i \in \{1, .., 5\}$
- $Z \perp\!\!\!\perp X_i \mid year$ for all $i \in \{2, .., 5\}$
- $year \perp\!\!\!\perp Z \mid X_i$ for all $i \in \{1, .., 5\}$
- $year \perp\!\!\!\perp X_i \mid Z$ for all $i \in \{1, .., 5\}$

2.11 Conditional independence between categorical variables

To find if there is conditional dependence between some variables, we need to make plots in function of levels of a variable. This is simple for categorical variables but for continuous variables, we need to discretize them in order to have different levels for this variable. As two independent variables can be dependent conditioning on a third one, we have to to watch all the possible plots to not miss a dependence relation. We begin here by plotting two categorical variables for different levels of a third categorical variable.

We keep only two plots, one showing conditional independence, another showing conditional dependence.

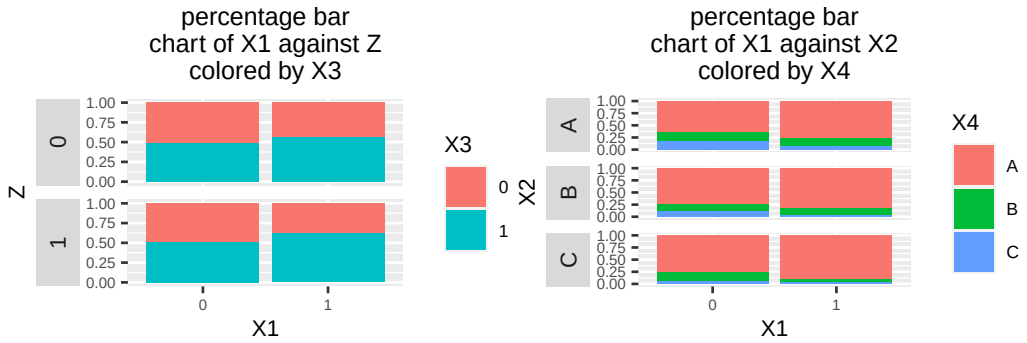


FIGURE 15: We can see on the first plot that $P(X3 \mid X1, Z = 0) = P(X3 \mid X1, Z = 1)$. This means that $X3 \perp\!\!\!\perp Z \mid X1$. On the other hand, we can see on the second plot that levels of X4 are not the same for each level of X2 even if we fix X1. Thus, X4 is dependent on X2 knowing X1.

We summarize here the new conditional independence that we found :

- X3 $\perp\!\!\!\perp$ Z | X1
- Z $\perp\!\!\!\perp$ X3 | X1
- X1 $\perp\!\!\!\perp$ X5 | X2
- X5 $\perp\!\!\!\perp$ Z | X2
- X1 $\perp\!\!\!\perp$ X2 | X3
- X1 $\perp\!\!\!\perp$ X5 | X3
- X2 $\perp\!\!\!\perp$ X1 | X3
- X5 $\perp\!\!\!\perp$ X1 | X3
- X1 $\perp\!\!\!\perp$ X3 | X5
- X2 depends on X4 | X1
- X4 depends on X2 | X1
- X1 depends on X4 | X2

We can see on the following pltos that X1 $\perp\!\!\!\perp$ X2 but X1 $\not\perp\!\!\!\perp$ X2 | X4 and X1 $\not\perp\!\!\!\perp$ X4, X2 $\not\perp\!\!\!\perp$ X4. This means that X4 could be a collider.

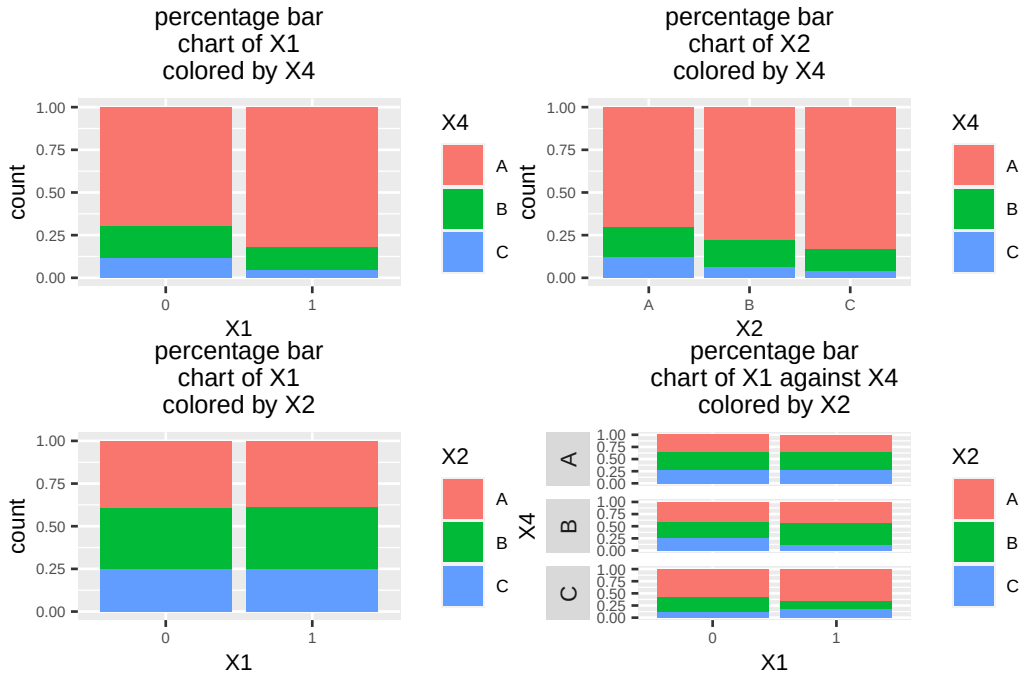


FIGURE 16: The first two plots show the dependence between X1, X4 and X2, X4 respectively. The third plot shows the independence between X1 and X2 and the forth plot reveals that the independence between X1 and X2 disappears when we condition by X4.

We make the same analysis for X2, X3 and X5 to try to find some independence relations.

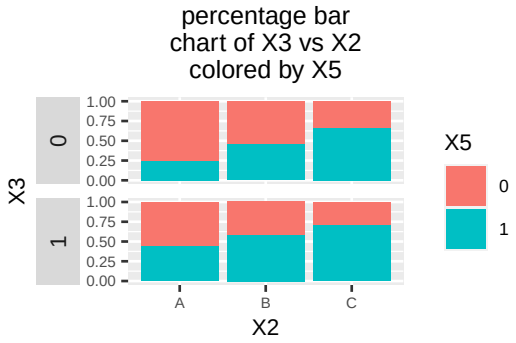


FIGURE 17: We notice that we have a new independence relation. Indeed $P(X5 \mid X2, X3) = P(X5 \mid X2)$.

2.12 Conditional independence between numerical variables

We would like to understand first the different conditional relations between V1_avg, V3_avg and V4_avg. We will then focus our attention on the conditional relations between V1_avg, V3_avg, V4_avg and Y.

We begin first by plotting V3_avg against V1_avg colored by a discretized version of V4_avg. This will help us to discover a conditional independence if there exists one.

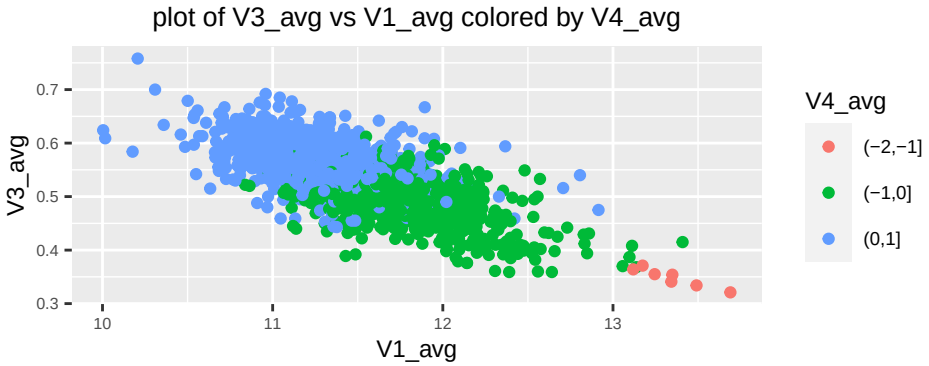


FIGURE 18: We can see on the plot that V3_avg seems dependent of V1_avg | V4_avg as most of the spread of V3_avg is explained by the change in V4_avg. When we fix V4_avg, change in V1_avg does not show change in V3_avg. Thus V3 $\perp\!\!\!\perp$ V4_avg | V1_avg.

To verify what we are claiming, we plot V3_avg against V1_avg for particular values of V4_avg. If V3_avg is really dependent of V1_avg knowing V4_avg, we will have some trend on these plots. If there is no trend, then we will conclude that V3_avg is in fact independent of V1_avg | V4_avg.

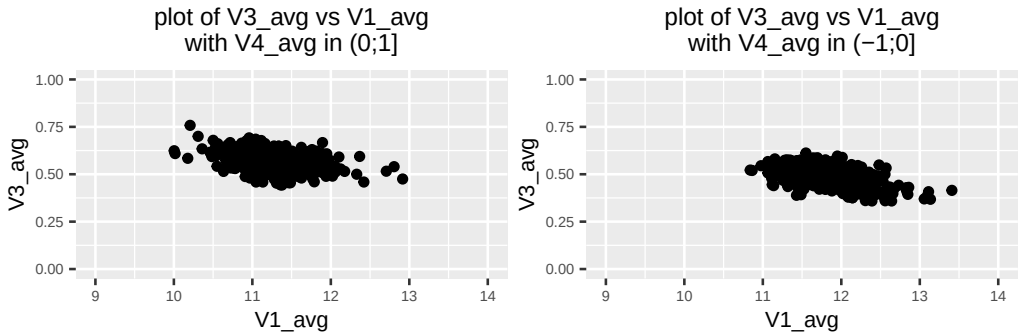


FIGURE 19: These two plots show nothing else than two clouds of points. This means that V3_avg is in fact independent of V1_avg | V4_avg.

We can make the same analysis for the other conditional dependence between these three variables.

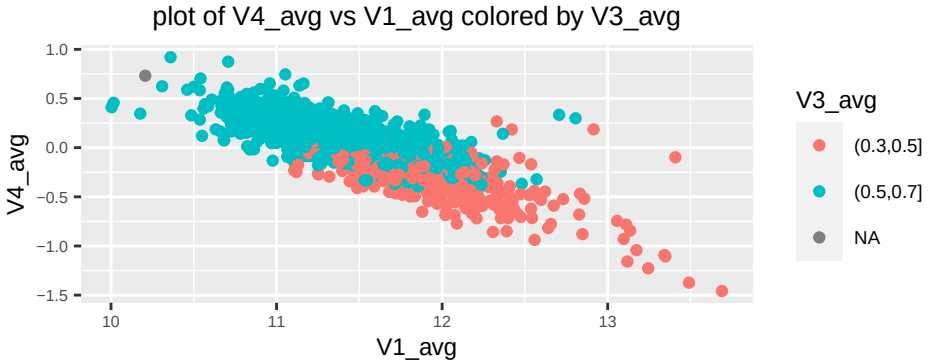


FIGURE 20: There is clearly a negative slope trend.

Let's verify if this trend is still there when we pick only particular values of V3_avg.

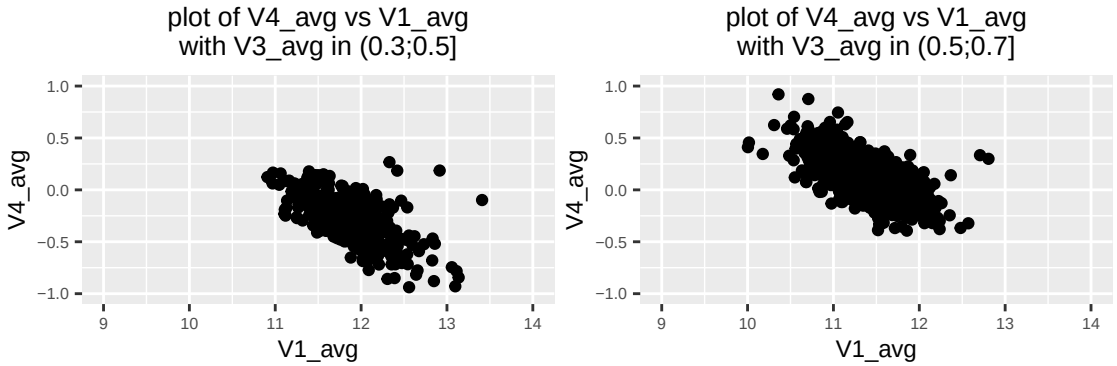


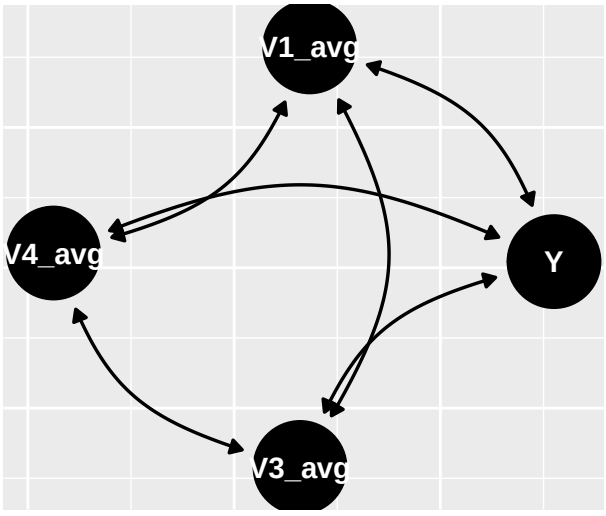
FIGURE 21: The negative trend is still here and the slope seems to be the same (which confirms us that there is no interaction effect of V1_avg/V3_avg on V4_avg) even when we take V3_avg in a particular interval.

This allows us to conclude that V4_avg is dependent of V1_avg | V3_avg.

By making once again the same analysis, we conclude that V4_avg depends on V3_avg | V1_avg.

2.13 Relation between V1_avg, V3_avg, V4_avg and Y

We already saw that these four variables were correlated with different correlation coefficients. We want to know now if we can identify some independence between these variables. Indeed, for now, we cannot have a better "DAG" than the following for these four variables :



We aim to have a a real DAG, i.e. without cycle or double arrows.

It seems reasonable to assume that Y is dependent of V4_avg because they have the biggest correlation coefficient compared to Y/V1_avg and Y/V3_avg. We also know that V1_avg, V3_avg and V4_avg are all dependent. Thus, by plotting Y against V1_avg/V3_avg colored by V4_avg, we will be able to determine if V1_avg and V3_avg really influence Y or if it is only the effect of V4_avg.

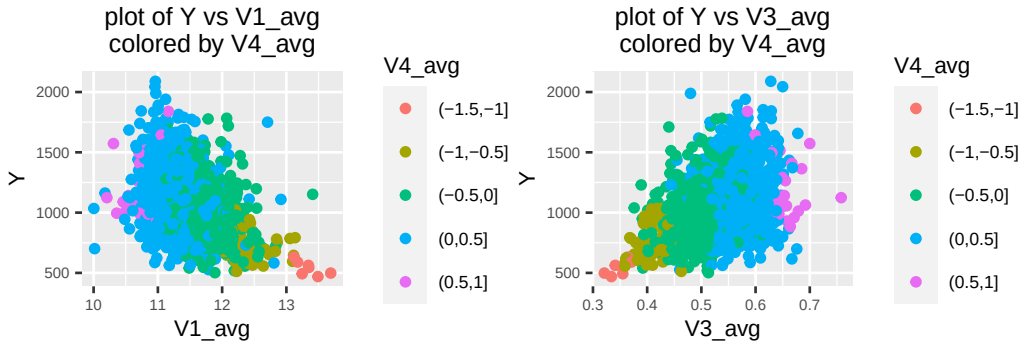
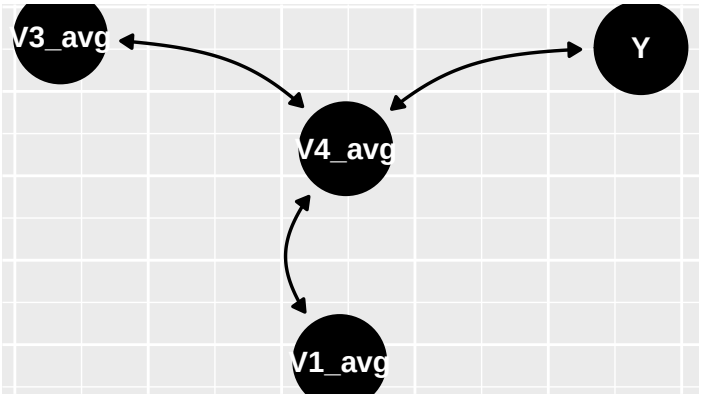
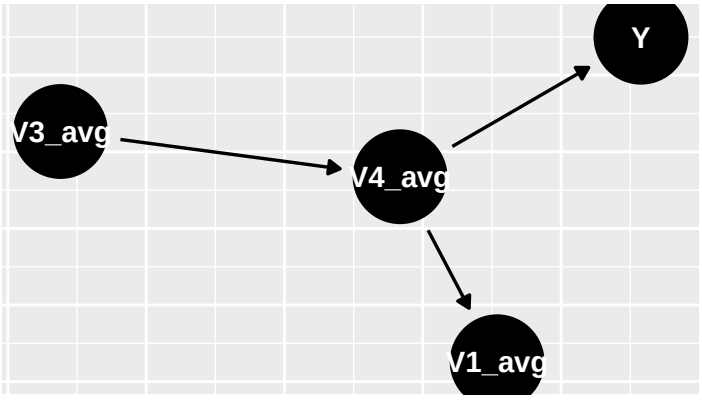


FIGURE 22: We can see on the first plot that the trend in the graph of Y against V1_avg disappears when we consider only patients with V4_avg in a certain interval. Same conclusion for Y against V3_avg. This allows us to say that Y is independent of V1_avg given V4_avg and Y is independent of V3_avg given V4_avg.

With these independence relations that we found, recalling that independence is equivalent to d-separation in a DAG, we are forced to remove the arrows between V3_avg/V1_avg, V1_avg/Y and V3_avg/Y. Indeed, we cannot have conditional independence between two variables if there is a path between them. We then find the following DAG :



Now, we have to determine the direction of the arrows. First, we know that V4_avg cannot be a collider because it would break all the conditional independence above. Thus, one possible DAG would be the following:



This DAG verifies all the relations between V1_avg, V3_avg, V4_avg and Y. Thus, we can keep this DAG in mind as it will be useful when we want to make the complete DAG.

2.14 Conditional independence between categorical variables and numerical variables

We know that there is association between X7 and X5. We want to know now if there is conditional dependence/independence between X7, X5 and another variable. More generally, we want to discover any conditional dependence/independence between a numeric variable and categorical variables.

We only keep one interesting plot, the other plots not showing anything.

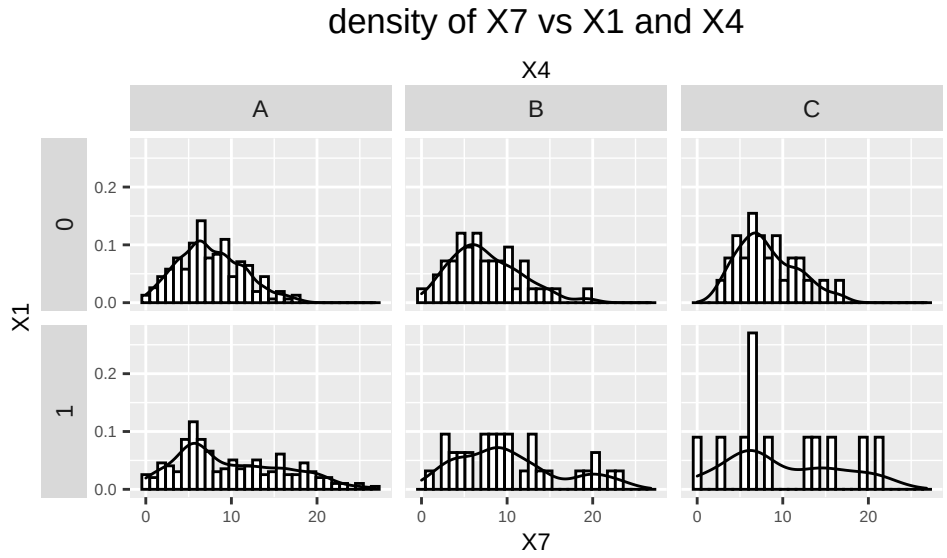


FIGURE 23: One would be tempted to say that $X4 \perp\!\!\!\perp X7 \mid X1$ as the density of X7 | X1 is the same for each level of X4 but as we can see on the graph at the bottom right, there is too few observations with $X1 = 1$ and $X4 = C$ which gives this strange plot without trend. Thus we cannot conclude that X7 | X1 is independent of X4.

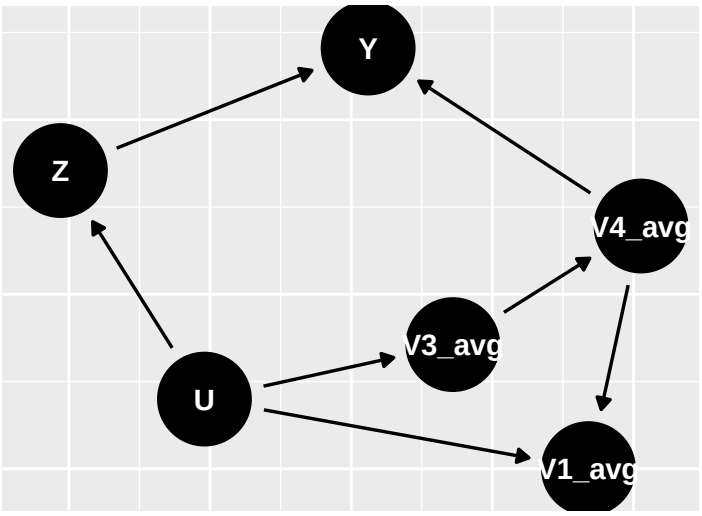
As we did not find any other association between a categorical variable and a numeric variable, we can now conclude our chapter on exploratory data analysis and begin the causal inference part of this project !

3 Directed Acylcic Graphs

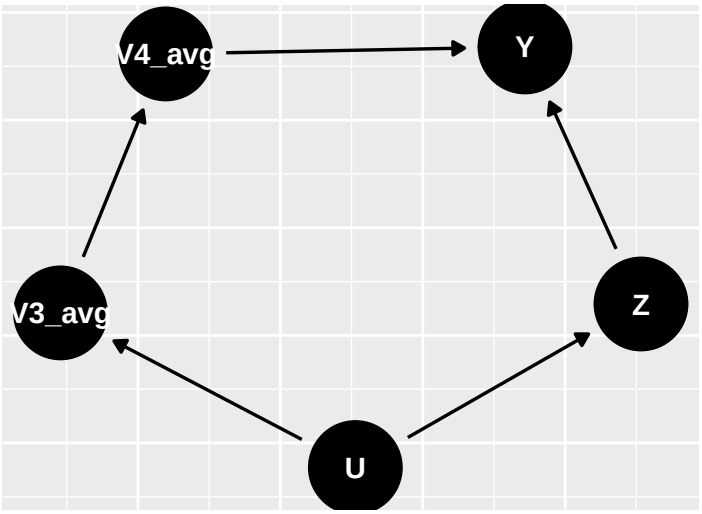
3.1 general DAG

We would like to draw a DAG containing all variables to see how all these work together. But we found out that Y depends only on Z, V1_avg, V3_avg and V4_avg. This means that all the other variables have an effect on Y only because they affect one of these four covariates. We can thus consider the other variables as unmeasured variables (U) to simplify our DAG.

We can then draw a DAG by taking the DAG we made for V1_avg, V3_avg, V4_avg and Y and adding Z and U to it :



We can remove V1_avg from the DAG as it has no direct or indirect effect on Y :



Here it is ! We find the DAG that describes our dataset. But this DAG describes the situation without taking into account time which is a very important variable for the experiment. Our aim now is to see if this DAG changes when we consider that variables vary over time.

4 Monthly average medical expenditures without the effect of time

4.1 Model for monthly average medical expenditures

We have seen several times now that monthly average medical expenditures depend heavily on time.

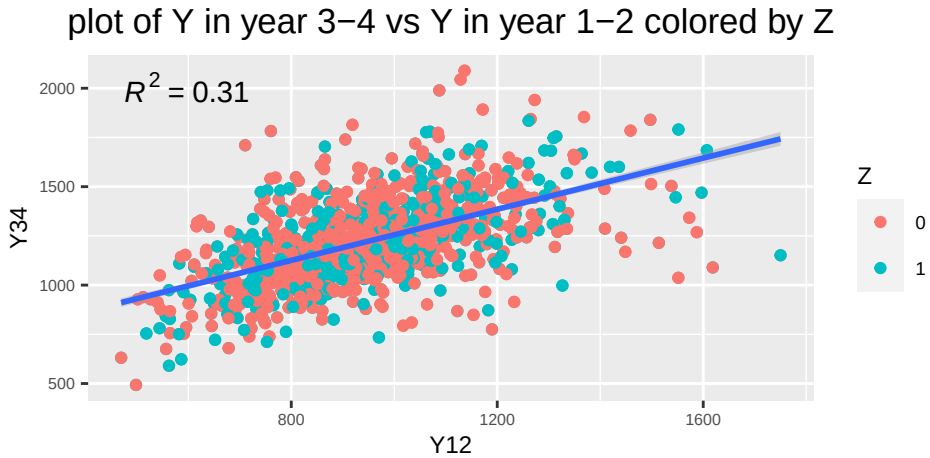


FIGURE 24: This plot shows the strong dependence of Y on time. It seems that the relation is linear as the slope is constant which makes us think that the relation between Y and time is a linear model of the form $Y34 = Y12 + \phi \cdot t$

To understand better this relation, we want to write a model that describes Y without the effect of time. This will help us to understand the link between Y and the other covariates as if the experiment was time-invariant because until now, the relations of dependence and independence between the variables are maybe only due to time. Thus, when we will find a model describing Y without the effect of time, we will be able to say without a doubt that the relations between the covariates and Y are true relations, not depending of time. The only assumption that we have to make to justify the use of a model is that we consider the effect of covariates on Y time-invariant which does not seem to be a big assumption as we can verify it (as we will do later).

We begin with a simple model : a new variable $Y' = Y - \phi \cdot t$. Our goal is then to estimate this ϕ as well as possible. It would be a nonsense to just find a general ϕ. Instead, we will fit a linear model for Y depending on year with the additional condition that $Z = 0$ to not take into account the effect of treatment on Y. This will give us our Y' and we will use it to estimate the real effect of different covariates on Y without the effect of time. We do not need to include interaction terms or to consider that ϕ depends on other variables as we saw that all variables were time-invariant. This means that the effect of time on Y is separated form other effects on Y.

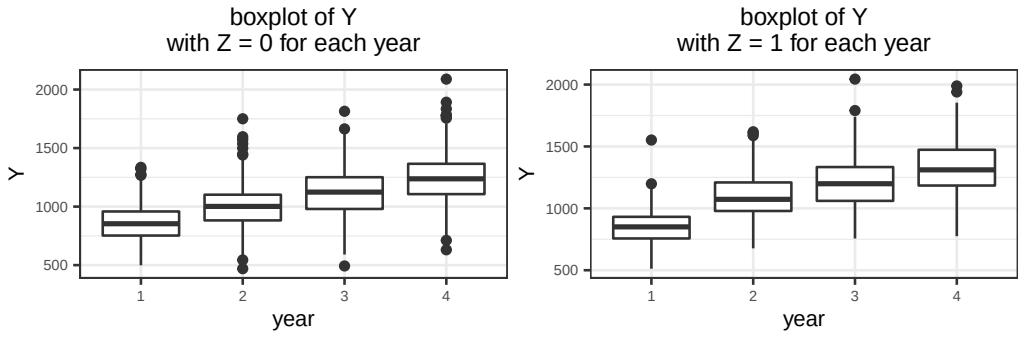


FIGURE 25: Boxplot of Y for each year. The first plot represents only non-treated people, the second only treated people.

We find that the model gives us :

$$\hat{Y} = 864 + 139.43 \cdot I(year = 2) + 263.75 \cdot I(year = 3) + 379.98 \cdot I(year = 4)$$

We create then a new column in our dataset named Y_prime, which gives us average monthly medical expenditures without the effect of time, i.e. we subtract the expected value of Y at a specific year to all values of Y w.r.t. the year of the observation. We can now have a better overview of the change in average monthly medical expenditures without the effect of time.

We plot Y' in year 1-2 against Y' in year 3-4 colored by Z to see if people with high medical expenditures in year 2 were the patients who were treated and thus, the treatment reduced their medical expenditures in year 3-4.

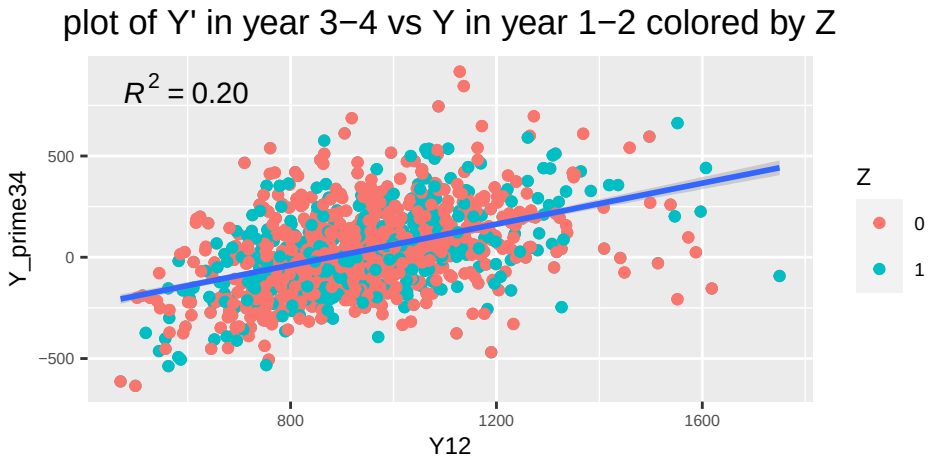


FIGURE 26: The increasing trend present in the first plot of the chapter is also present in this plot, which is surprising as this plot is the medical expenditures in year 1-2 against medical expenditures in year 3-4 without the effect of time. We should have then a flat curve, which is not the case.

This makes us think that the relation between monthly average medical expenditures and time is in fact not linear. Let's see if we can find a model that describes the relationship between Y and time better.

We saw that an increase in Y12 resulted in an increase in Y34. After a linear adjustment, this trend is still

there which makes us think that the increase is equal on percentage terms, i.e. $Y = Y1 \cdot g(t)$. To have a linear model from this expression, we divide by Y1 on the two sides and we compute a new column Y/Y1 from which we fit a linear model for this new variable (say Y_mult).

This gives us the following model :

$$\widehat{\frac{Y}{Y1}} = 1 + 0.179 \cdot I(year = 2) + 0.326 \cdot I(year = 3) + 0.462 \cdot I(year = 4)$$

We plot now Y in year 1-2 against Y_mult in year 3-4 colored by Z as before :

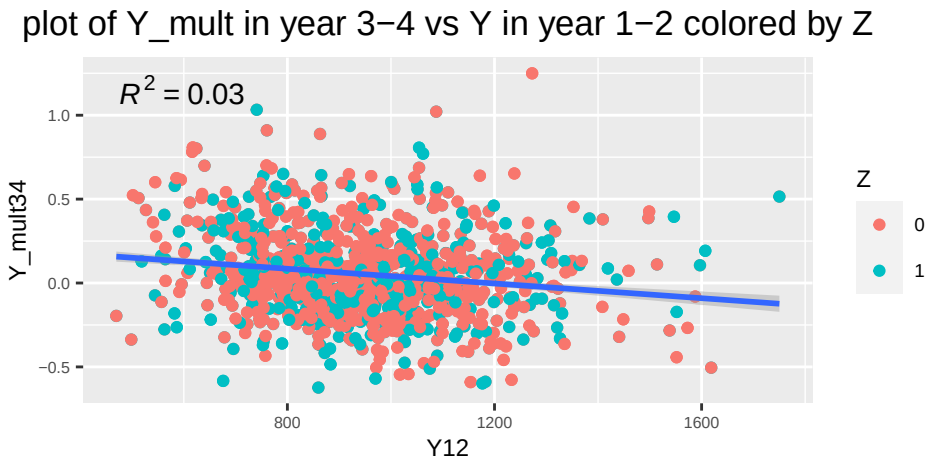


FIGURE 27: The multiplicative model seems convincing. Indeed, there is no increasing trend nor pattern, which indicates that this model could be a plausible one. Furthermore, we can see that the R^2 has a really low value, which indicates that we really removed the effect of time on Y. Indeed, the R^2 of the preceding model was of 0.20 which indicated a trend in the model.

It seems that we find a good approximation of the true model representing monthly average medical expenditures without the effect of time.

We plot now Y_mult against time colored by Z as we did before for Y to see what has changed.

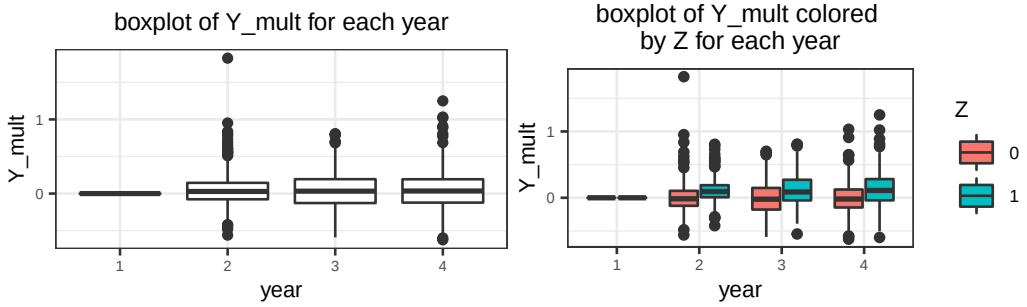


FIGURE 28: We can see with the first plot that the effect of time is really removed as monthly average medical expenditures are all centered around 0. The second plot could make us think that being treated just increases our medical expenditures, but we cannot affirm that as people who were treated could have had even higher expenditures if they were not treated.

4.2 Link between Y_mult and numeric variables

Now that we found the mean average medical expenditures without the effect of time, we can ask ourselves if this variable has other dependence/independence relations with the other variables. We compute again the plots of Y_mult against each numeric variable to see if there is some changes. We keep only the plots with R^2 bigger than 0.01 for convenience.

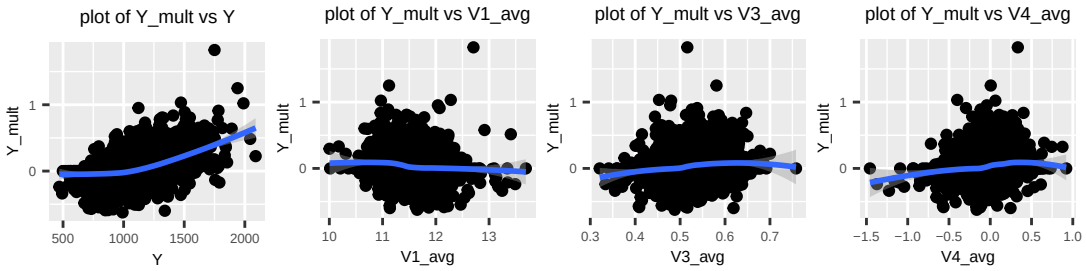


FIGURE 29: We can see that there is no new relations with the numeric variables. There is just small relations between V1_avg, V3_avg, V4_avg and Y_mult which is what we had already for Y.

We can now move on to categorical variables and see if the relations have changed.

4.3 Link between Y_mult and categorical variables

We did not found a lot of relations for Y and categorical variables. In fact, Y depends only on Z and V4_avg. We want to know now if Y_mult has other relations with categorical variables. We filter out year = 1, because it does not give us any information and it could hide some relations.

We only keep two plots for brievety.

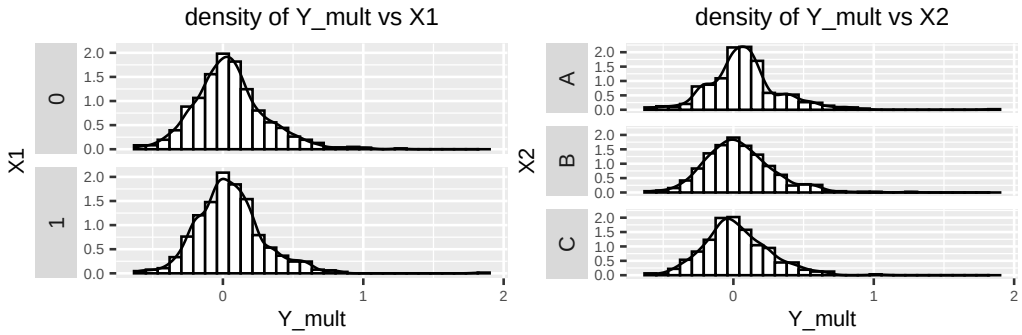


FIGURE 30: The first plot shows that the independence between Y_mult and X1 is still valid. On the other hand, the second plot seems to show a dependence relation between Y_mult and X2 which is surprising as X2 was independent of Y.

With the same reasoning, we can conclude that Y_mult is in fact independent of all categorical variables except X2 and Z. Dependence between Z and Y_mult is a good thing as we already had this relation. Let's verify if the relation between X2 and Y_mult is really a dependence relation.

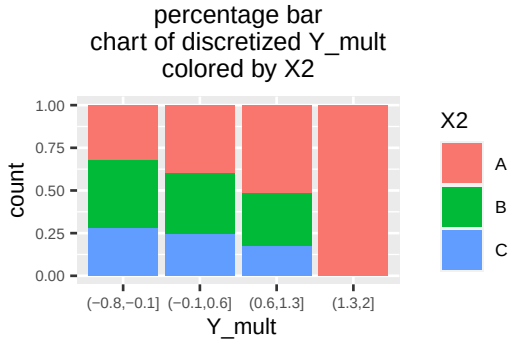


FIGURE 31: The three first bars seem to be distributed the same. The last bar is not really important as there is only one observation of Y_mult between 1.3 and 2. Thus, we can conclude that Y_mult is in fact independent of X2.

We can conclude that there is no new relation between Y_mult and categorical variables.

4.4 Time-invariance of the relations between variables

Until now, we focused on the dependence/independence relations between variables in general. With that, we achieved to make a DAG that represents our experiment. But our experiment depends highly on time. Our aim now is to see if our DAG is still valid if we take into account the fact that our experiment depends on time. We need then to verify that all the relations between variables are valid for each year. We begin with the categorical variables.

4.4.1 Categorical variables

We keep here only one dependence and one independence plot for convenience.

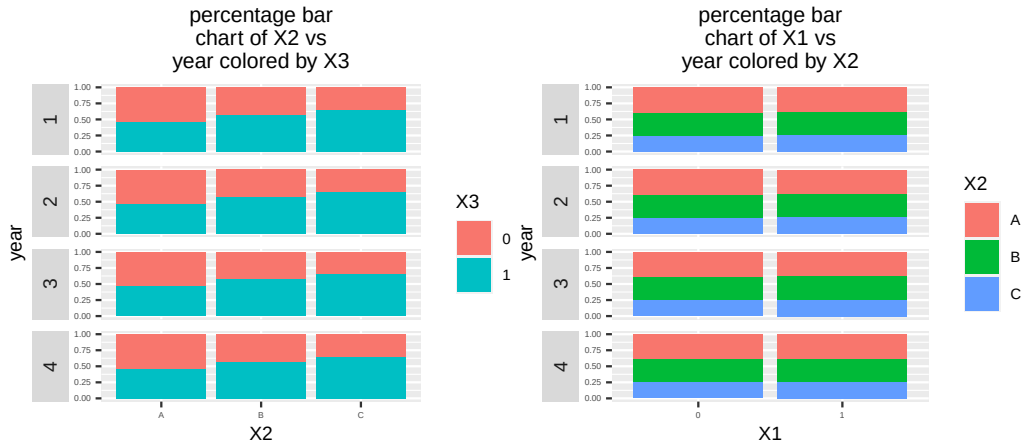


FIGURE 32: The first plot shows us that X2 and X3 are dependent, no matter the year. The second plot shows on the contrary that X1 and X2 are independent, no matter the year.

Following the same reasoning for all categorical variables, we can conclude that the relations between categorical variables are valid for each year. Let's see if this conclusion holds for numeric variables too.

4.4.2 numeric variables

By making a plots/correlations matrix for each year, we can affirm that the relations between the numeric variables that we have from the beginning remain the same for each year. Thus, we focus on finding new relations between Y_mult and these latter.

We filter out year = 1, because it does not give us any information and it could hide some relations as before. We just keep the covariates for which the correlation coefficient is significant compared to Y_mult, i.e. V1_avg, V3_avg and V4_avg. We show plots of year 2 and year 4.

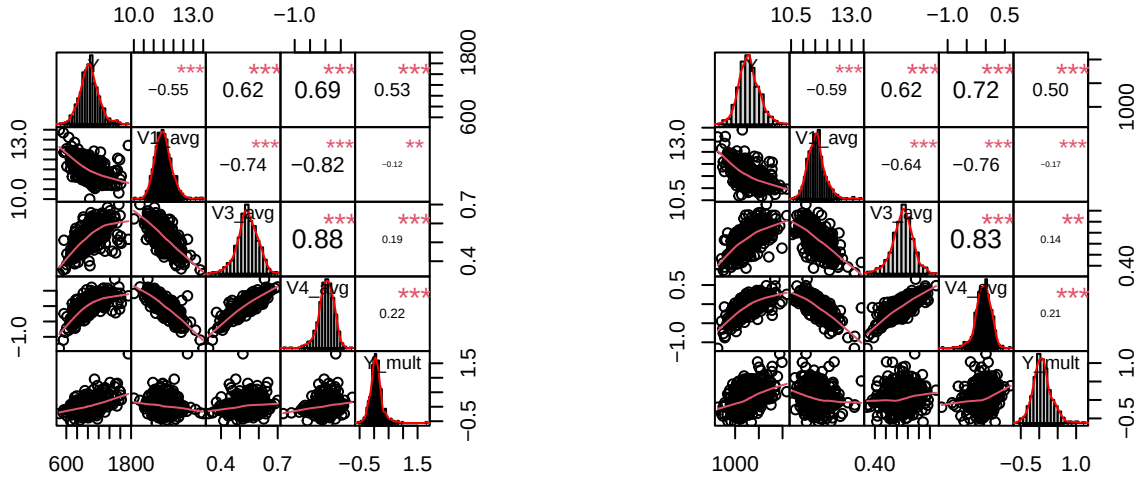
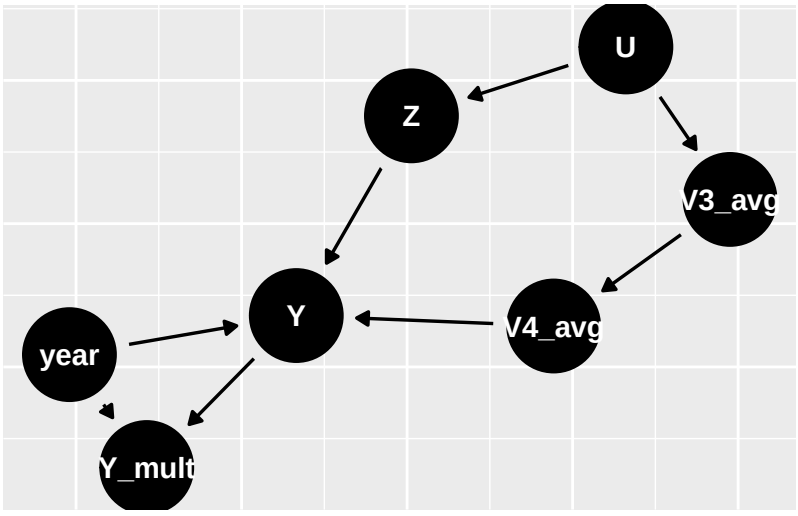
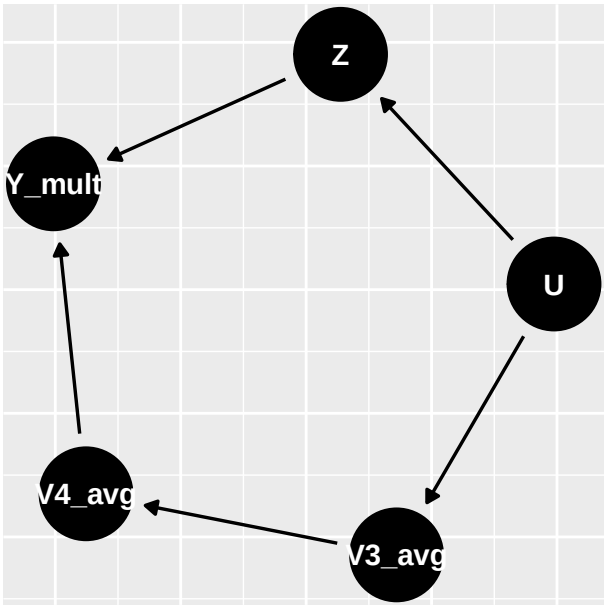


FIGURE 33: We can see some trends between V1_avg, V3_avg, V4_avg and Y_mult but these latter are clearly less obvious than before. This shows us that time was a major factor of the dependence between these three numeric variables and Y. The correlation coefficients being still significant, we can consider that Y_mult depends on V1_avg, V3_avg and V4_avg. As the latters are the same for each year, we can confirm that the relations are time-invariant.

With all these analysis, we can see that the relations of dependence and independence between the variables remain the same each year ! This is very good news because it means that our DAG is still valid. Indeed, we can draw the final DAG of the experiment :



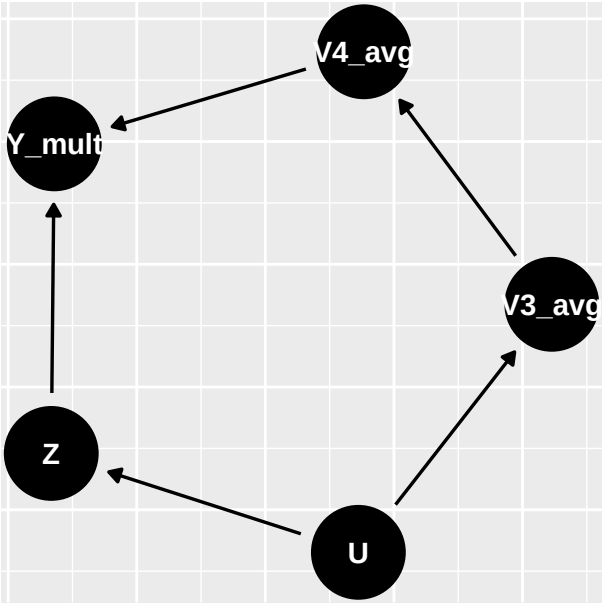
But as Y_mult is just Y adjusted for time, we can just shrink the (Y, Y_mult, time) subgraph to Y_mult. Indeed, no variable is impacted if we make this choice and the previous analysis showed that time impacted nothing but Y. Thus, we can simplify our final DAG to this one :



To conclude, the model describing Y without the effect of time is very useful to justify that our DAG is not just a time-invariant DAG, but really the DAG describing our experiment. The disadvantage of the model is that the relations between the covariates and Y seem weaker than before which could lead to over-interpretation of the different plots.

5 Verifications

After all this work, we are saying that the following DAG is the DAG of the experiment. We want now to verify that our claim is true by verifying the dependence and independence that the DAG implies. For example, The DAG tells us that Y_mult depends directly on V4_avg and Z but also on X6 and V3_avg as there is no collider between them. We also need to have V4_avg independent of Z knowing X6, etc...
To verify these relations, we just do the reversed analysis that we did before.



5.1 Dependence and independence

We have to verify the following relations :

- Y_mult dependent of V4_avg, V3_avg and Z
- V4_avg dependent of V3_avg and Z
- V3_avg dependent of Z

The relations of dependence between V3_avg, V4_avg and Y_mult are already verified as we have used them to construct our DAG. It remains to show the dependence between Z and the numeric variables.

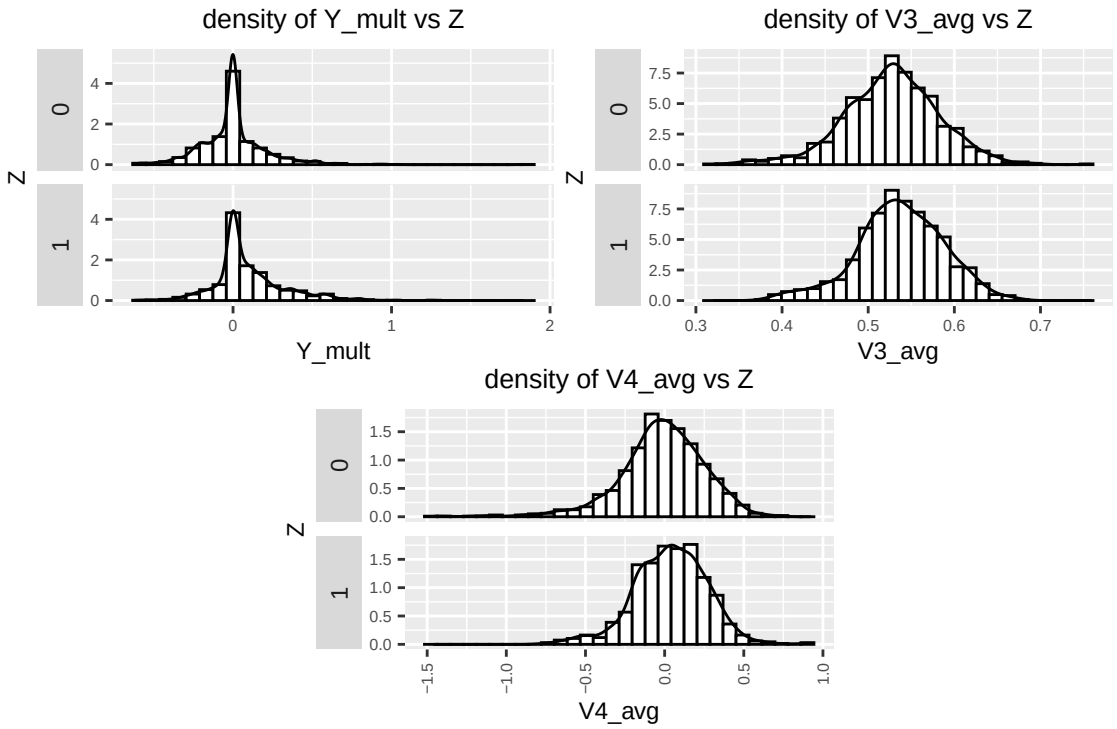


FIGURE 34: Even if they are not obvious, we can see small differences between the density plots for each level of Z which indicates that Y_mult, V3_avg and V4_avg are dependent of Z. We can consider these small perturbations as significant because our DAG should be a "perfect" representation of the experiment and thus, every perturbation has its importance.

The dependence relations seem to be verified by our dataset. Let's continue with the conditional relations.

5.2 Conditional dependence and independence

We need to verify the only one relation : Z $\perp\!\!\!\perp$ of V4_avg | V3_avg.

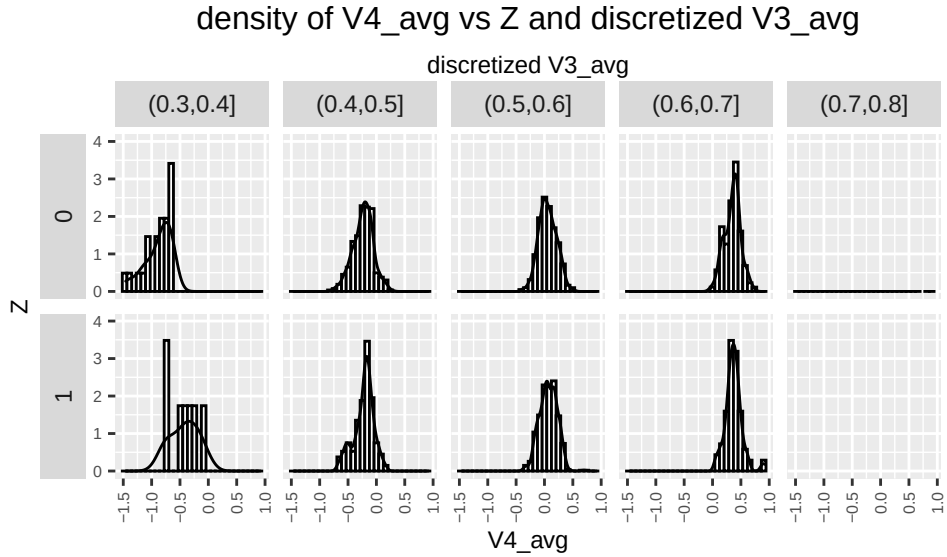


FIGURE 35: This graph shows us that Z is in fact independent of V4_avg | V3_avg as $P(V4_{avg} \mid V3_{avg}, Z = 0) = P(V4_{avg} \mid V3_{avg}, Z = 1)$. This trend is not obvious for V3_avg between 0.3 and 0.4 but there are only 32 observations with V3_avg in this interval (same comment for V3_avg between 0.7 and 0.8). Thus, the independence relation is still valid.

The relation of conditional independence is verified. This means that our DAG is a good approximation of the real situation in which the experiment took place !

6 Conclusion

After making data analysis on our dataset, we found out all the relations of dependence and independence between the covariates which allows us to create a DAG representing the experiment. As we did not take into account the fact that the experiment was time-varying, we verified that the different relations were always true over time and we conclude that the first DAG was in fact the DAG of the time-varying experiment.

To summarize, we achieved to draw a directed acyclic graph from a dataset without knowing the story behind the experiment from which the data come from. Even if the first objective of the challenge was to estimate the effect of treatment, we decided to focus on the analysis of relations between the variables to find a DAG as close as possible to reality, which we managed to make.

7 References

Hernàn MA, Robins JM (2020). Causal Inference: What If. Boca Raton: Chapman & Hall/CRC.