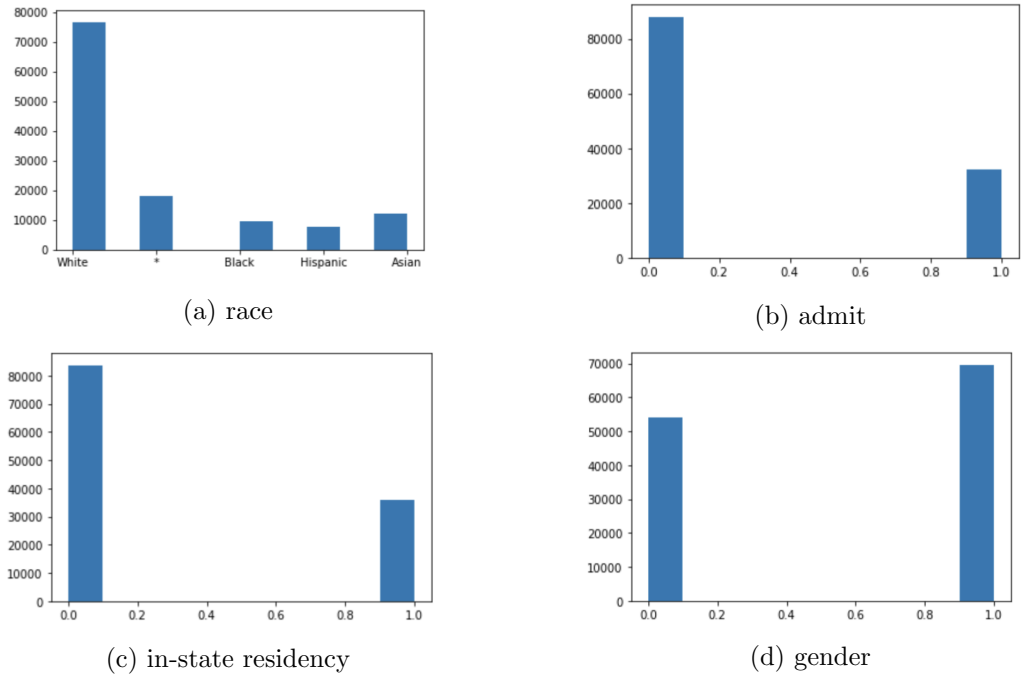
Causal Interpretations between Race and College Admission

Yuxuan(Jonathan) Chen YCHEN349@JHU.EDU
Computer Science
ychen349

1. Introduction

Racial issues have been a main focus recently, and our focus will be on its impact on college admission. We used the FOIA law school admission dataset to study this relationship from Sander et al. (2009). The dataset contains variables we are interested in, race and admission decision, and other variables including the students' grade point average(GPA), LSAT score, gender and in-state residency. It also contains college and time which could also affect admission.
The hypothesis here is that race does not affect admission decision directly. It may have an indirect impact on admission via LSAT score or GPA, or other unmeasured confounder variables like social-economic status.



(a) race (b) admit

(c) in-state residency (d) gender

Figure 1: Distributions of discrete variables in FOIA Dataset

2. Data Overview

FOIA dataset contains relatively fewer variables after cleaning. The only column we removed is *enroll*, which has majority of the data missing and no logical causal effect on admission decision. Different races are represented in an one-hot encoding fashion, so we concatenated all races to one single column named *race*. This column contains five different types of values: Black, Hispanic, Asian, White, and Missing(represented by *). Other variables include LAST score, GPA, in-state residency, college name, year, and gender.
We want to check the distributions of those variables and check if they show significant inbalance. We mainly plot the discrete variables that we are interested in. (See Figure 1)
From the histograms, we see that the *race* variables is not balanced since it contains much more White values than the others. This could potentially be a problem when we are drawing conclusions. However, since we have a total of 124557 rows of data, this problem is mitigated as we should have sufficient data even for those values with fewer data. The other three binary variables does not contain significant inbalance although the numbers of data for 0 and 1 differ by at most 60%.

3. Preliminaries

3.1 Causal Model of a DAG

Causal models of a DAG $\mathcal{G}$ defined over a set of variables V may be interpreted as a tuple consisting of the DAG itself, and a system of non-parametric structural equations with independent errors equipped with the do-operator. Each variable is determined as a function of its parents and an independent error term. This induces a distribution $p(V)$ that factorizes according to the DAG $\mathcal{G}$ as follows,

$$p(V) = \prod_{V_i \in V} p(V_i \mid \mathrm{pa}_{\mathcal{G}}(V_i)),$$

where $\mathrm{pa}_{\mathcal{G}}(V_i)$ denotes the parents of V_i in $\mathcal{G}$. Under this interpretation, a directed edge $V_i \to V_j$ may be interepreted as saying that V_i is potentially a direct cause of V_j. Conditional independences in $p(V)$ can be read off from the DAG via d-separation, i.e., $(X \perp\!\!\!\perp Y \mid Z)_{\text{d-sep}} \implies (X \perp\!\!\!\perp Y \mid Z)_{\text{in } p(V)}$. To facilitate structure learning, we will restrict our analysis to the set of *faithful* distributions where $(X \perp\!\!\!\perp Y \mid Z)_{\text{d-sep}} \iff (X \perp\!\!\!\perp Y \mid Z)_{\text{in } p(V)}$.

3.2 Causal Model of an ADMG

Causal models of an ADMG $\mathcal{G}$ contains all elements in a DAG. Additionally, since ADMG contains bidirected edges that represent the presence of an unmeasured variable in the model. For example, a bidirected edge between A and B looks like A $\leftarrow\!\rightarrow$ B. This represents the model A $\leftarrow$ U $\rightarrow$ B where U is an unmeasured confounder variable that does not exist in the dataset or not defined. The distribution $p(V)$ that factorizes according to the ADMG

$\mathcal{G}$ looks like,

$$p(V) = \prod_{D \in D(\mathcal{G})} q_D(D \mid \text{pa}_{\mathcal{G}}(D)),$$

where D(G) denotes the set of all districts in G(V) and $q_D(D \mid \text{pa}_{\mathcal{G}}(D))$ is a kernel function. Districts are defined as a partition of all vertices into bidirected components. Kernel is a similar representation to ordinary conditional densities that provides a mapping of values of variables to the right of the conditioning bar to normalized densities over variables to the left of the conditioning bar. Conditional independences in $p(V)$ can be read off from the DAG via m-separation, i.e., $(X \perp\!\!\!\perp Y \mid Z)_{\text{m-sep}} \implies (X \perp\!\!\!\perp Y \mid Z)_{\text{in } p(V)}$. To facilitate structure learning, we will restrict our analysis to the set of *faithful* distributions where $(X \perp\!\!\!\perp Y \mid Z)_{\text{m-sep}} \iff (X \perp\!\!\!\perp Y \mid Z)_{\text{in } p(V)}$.

3.3 Tetrad

Tetrad (Scheines et al. (2002)) is a causal discovery software developed by Carnegie Mellon University and University of Pittsburgh to find Markov equivalent DAGs or ADMGs. It supports multiple causal discovery algorithms, including both constraint-based and score-based ones. Tetrad also supports bootstrapping that measures the uncertainty in the process.

3.3.1 FCI algorithm

The FCI algorithm (Spirtes et al. (1993)) is a constraint-based algorithm that takes as input sample data and optional background knowledge and in the large sample limit outputs an equivalence class of CBNs that (including those with hidden confounders) that entail the set of conditional independence relations judged to hold in the population.

3.3.2 FGES algorithm

FGES is an optimized and parallelized version of an algorithm developed by Meek (Meek and Chickering (2002)) called the Greedy Equivalence Search (GES). The algorithm was further developed and studied by Chickering (Chickering (2002)). GES is a Bayesian algorithm that heuristically searches the space of CBNs and returns the model with highest Bayesian score it finds. In particular, GES starts its search with the empty graph. It then performs a forward stepping search in which edges are added between nodes in order to increase the Bayesian score. This process continues until no single edge addition increases the score. Finally, it performs a backward stepping search that removes edges until no single edge removal can increase the score.

3.3.3 GFCI algorithm

GFCI is a combination of the FGES algorithm and the FCI algorithm that improves upon the accuracy and efficiency of FCI. The FGES algorithm is used to improve the accuracy of both the adjacency phase and the orientation phase of FCI by providing a more accurate initial graph that contains a subset of both the non-adjacencies and orientations of the final output of FCI. The initial set of nonadjacencies given by FGES is augmented by

FCI performing a set of conditional independence tests that lead to the removal of some further adjacencies whenever a conditioning set is found that makes two adjacent variables independent. After the adjacency phase of FCI, some of the orientations of FGES are then used to provide an initial orientation of the undirected graph that is then augmented by the orientation phase of FCI to provide additional orientations.

3.4 Odds Ratio

Given any pair of reference values x_0 and y_0, the conditional odds ratio of X and Y given Z is defined as:

$$OR(X = x, Y = y \mid Z) = \frac{p(X = x \mid Y = y, Z) \cdot p(X = x_0 \mid Y = y_0, Z)}{p(X = x_0 \mid Y = y, Z) \cdot p(X = x \mid Y = y_0, Z)}$$

We make use of odds ratio test based on the property that, $X \perp\!\!\!\perp Y \mid Z$ if and only if $\mathrm{OR}(X = x, Y = y \mid Z) = 1$ for all x,y,z. This provides the theoretical basis to derive conditional independences from odds ratio tests.

4. Methods

4.1 Tetrad

According to our hypothesis, the graphical model could be entirely based on the variables we have in the dataset, or could contain unmeasured variables that are common causes of measured ones. Therefore, we approach the problem separately when we are choosing algorithms. To take unmeasured variables into consideration, we choose algorithms that allow latent common causes. To only consider existing variables in the dataset, we chose algorithms that forbid latent common causes.

We first load the dataset into Tetrad. The dataset has been properly pre-processed to only include variables that we are going to use in the model. We replaced the missing values by an asterisk and that can be treated easily by Tetrad. In the FOIA dataset, we set the number to distinguish between continuous and discrete variables at 25 (distinct values in *colleges*).

After loading the dataset, in order to forbid arrows that are not logically possible, we add some prior knowledge to the model. The standard is that all attributes that is based on a person's origin is placed in the first tier. This includes race, gender and residency. We forbid within tiers among these since we consider them as independent of each other.

4.1.1 ADMG using GFCI

Since the dataset is relatively big, GFCI is the more efficient way to derive an ADMG for the dataset. We select Conditional Gaussian Likelihood Ratio Test (CG-LRT) as the test, and Conditional Gaussian BIC Score (CG-BIC) as the score. Most default parameters work well on the dataset. We assume faithfulness, and the default penalty discount of 1 produces a graph that is not too dense or too sparse. To stabilize the result, we consider using bootstrapping for 10 iterations. We choose preserved as the ensemble method as this keeps more edges for us.

4.1.2 DAG USING FGES

To derive a DAG, we use FGES algorithm directly. The settings are mostly the same for the test and score. Since we are working on the same dataset, we don't change the parameters to make sure the cutoff and penalty discount are the same. We again apply bootstrapping for 10 iterations with preserved as the ensemble method. PC algorithm does not work here since the dataset contains both discrete and continuous variables, which contradicts with the assumption PC makes that all variables are either discrete or continuous.

4.2 Odds Ratio Test

To further test the existence of several specific edges, we conducted odds ratio tests on the dataset. In order to do so, we assigned different values of *race* variable between 0 and 1 to distinguish between different *race* types. The four odds ratios we tested are OR(*race*, *admit*), OR(*race*, *admit* | *gpa*), OR(*race*, *admit* | *lsat*), OR(*race*, *admit* | *gpa*, *lsat*).

Since we have a dataset with a big number of rows, we again used bootstrapping to construct a confidence interval for the odds ratios. We are able to obtain pretty narrow confidence intervals that could tell us if one conditional independence we are interested exists.

I also tried the fast conditional independence test (Chalupka et al. (2018)) for the same purpose as odds ratio test. Unfortunately, the size of the dataset (124557 rows) is too big to conduct FCIT as Python crashes when it is running. A more efficient implementation that supports a higher dimension is needed to perform FCIT on this dataset.

5. Results

5.1 ADMG derived by GFCI

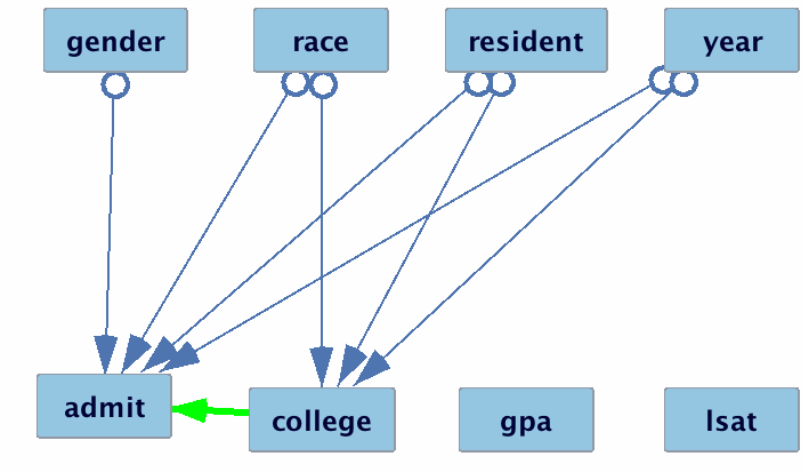


Figure 2: ADMG generated from the two datasets by GFCI

The ADMG we derived from the FOIA dataset is shown in Figure 2. The one-sided circle-arrow from A to B indicates that A could be a cause of B, or there could exist an unmeasured confounder between A and B. Most connected variables are connected via the circle arrow. This includes the connection between all four first-tier variables and *admit*, three first-tier variables and *college*. The arrow from *college* to *admit* is green, indicating that there is no latent confounders between these two variables. No arrows are connected to *lsat* or *gpa*.

5.2 DAG derived by FGES

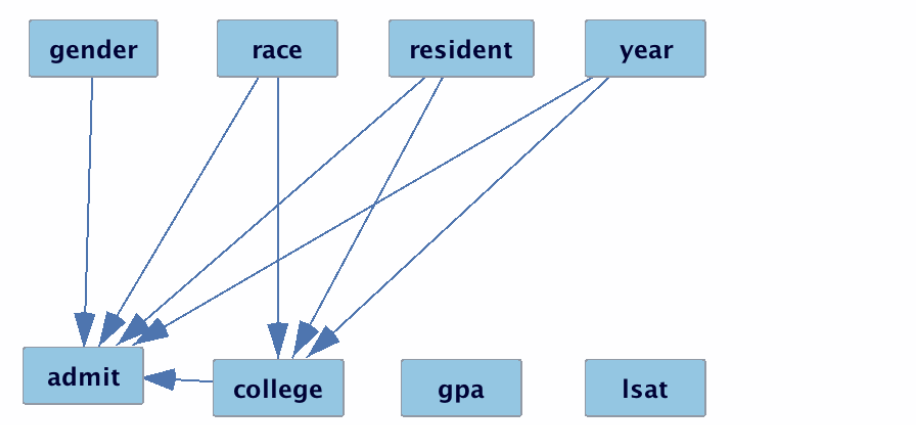


Figure 3: DAG generated from foia datasets by FGES

The DAG we derived from the FOIA dataset is shown in Figure 3. The DAG using DGES has exactly the same structure as the ADMG in Figure 2. This implies that the one-sided circle-arrows in the ADMG are more likely to be causation relationship and no confounders exist.

5.3 Odds Ratio Tests

```
OR(race, admit) 0.9059681854213831 (0.8930397941261121, 0.9166569347790962)
OR(race, admit | gpa) 1.0151255557984287 (1.0019035147089632, 1.0287907321584313)
OR(race, admit | lsat) 1.1871617843528481 (1.1690079026808833, 1.2051835872100738)
OR(race, admit | gpa,lsat) 1.2294002430115951 (1.209370420517068, 1.24554553624103)
```

Figure 4: Odds ratio test and corresponding confidence intervals

From the four odds ratios and the confidence intervals(See Figure 4), we could safely assert that the only possible conditional independence here is that $race \perp\!\!\!\perp admit \mid gpa$. The triplet between *race-gpa-admit* could only be a chain or a fork. Since *gpa* cannot cause *race* according to our assumption, the triplet could only be a chain. Thus, it's possible that *race* has an indirect effect on *admit* via *gpa* according to the odds ratio we obtained.

Based on the fact that *race* $\not\perp\!\!\!\perp$ *admit* marginally, we are not certain if there exists a direct edge from *race* to *admit*. If the prior triplet indeed has a chain structure, the marginal independence does not hold no matter the direct edge exists or not.

6. Discussion and Conclusion

Combining the results from the causal discovery and the conditional independence testing, we could arrive at some consistent conclusions and some inconsistent ones that require further study.

Based on the initial hypothesis that race does not affect admission decision directly, it is reasonable to overturn this statement. From the ADMG and DAG, we observed the edge between *race* and *admit*. Although the edge in the ADMG could either be a directed edge or a bidirected one, we assume that race is an internal characteristic of an individual that cannot be affected or caused by any variables (measured and unmeasured) included in this study. Therefore, the only possibility is that *race* has a direct edge to *admit*, which means that race is a direct cause of admission decision.

The next question is, does *race* affect *admit* via a third measured variable? In our hypothesis, we proposed that this third variable can logically be *lsat* or *gpa*. From the graphs, these two variables are not connected to any other variable. This indicates that *lsat* and *gpa* may be independent of all other variables. Since the criterion we used when creating the graphs is based on conditional likelihood ratio, the models where *lsat* and *gpa* have no edges have a higher likelihood.

Nevertheless, when we approach the problem through odds ratio, we found that $OR(race, admit \mid gpa)$ is very close to 1. According to the property of odds ratio, this implies that $race \perp\!\!\!\perp admit \mid gpa$. The most common interpretation to this is that the triplet of *race-gpa-admit* is a chain. But according to the ADMG and DAG we reached, this is not very likely. We need to consider if other possibilities here could explain the seemingly contradicting results.

The most likely explanation to this, is that *race* is not completely independent of *admit* given *gpa*, although the odds ratio is very close to 1. We see that $OR(race, admit)$ is around 0.9, which is not very far from 1. Conditional on *gpa* makes it much closer to 1, but in the causal discovery task this may still not be good enough to deduct that a chain exists on this path. Even if we assume that the conditional independence actually holds here, there are many different models in which $race \perp\!\!\!\perp admit \mid gpa$ but $race \not\perp\!\!\!\perp admit$. Without more information, it's hard to make an assertion of what the exact model looks like. At this stage, we could only stop at all Markov equivalent models that satisfy these requirements.

Apart from the problem we are interested in, in the ADMG and the DAG, there exists a directed edge from *college* to *admit*. This tells us that certain law schools have a higher admission rate than others, which is commonsense in school application. Additionally, *race*, *resident*, and *year* are possible to have some relationship with *college*. This provides the basis for phenomena like certain race of students tend to prefer certain schools or in some years certain colleges receive a higher number of applications than usual.

Furthermore, we hope to estimate the causal effect of *race* have on *admit*. Traditional Average Causal Effect(ACE) approach does not work well here because the *race* variable has 5 different values. One workaround is to calculate the ACE among all pairs of values in *race*. Another way is to use a contingency table to record the data and perform tests using corresponding test statistics.

While the dataset has a decent amount of data, the number of useful variables is relatively small. It may be sufficient for our study, but a more comprehensive study would be helpful if we hope to find another source to research on. I also hope to find a dataset which contains the social-economic status data like family income or parents' education level. However, this kind of information is less available online since it touches some privacy of students that could not be legally disclosed.

References

Krzysztof Chalupka, Pietro Perona, and Frederick Eberhardt. Fast conditional independence test for vector variables with large sample sizes. *arXiv preprint arXiv:1804.02747*, 2018.

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554, 2002.

Christopher Meek and David Maxwell Chickering. Finding optimal bayesian networks. 2002.

Richard Sander, Roger Bolus, Riley DaCosta, Wayne Grove, Joe Hicks, Andrew Hussey, E. Douglass Williams, and Robert Zelnick. The scale and effects of admissions preferences in higher education. 2009.

Richard Scheines, Peter Spirtes, Clark Glymour, Christopher Meek, and Thomas Richardson. The tetrad project: Constraint based aids to causal model specification. *Multivariate Behavioral Research*, 2002.

Peter Spirtes, Clark Glymour, and Richard Scheines. Causation, prediction, and search. 1993.