Factors that might cause labor condition application (LCA) to be approved

Viveksinh Solanki
Department of Computer Science
Stevens Institute Of Technology
Hoboken, NJ 07030
vsolank3@stevens.edu

Vardaan Kishore Kumar
Department of Computer Science
Stevens Institute Of Technology
Hoboken, NJ 07030
vkishore@stevens.edu

Abstract

The labor condition application (LCA) is a form that requires U.S. Department of Labor (DOL) approval in order for an employer to file an H1B petition for a temporary professional worker. Without approved LCA, you cannot file an H1B petition. In this paper, we try to find the factors that might cause the LCA to be approved. Provided positive results of our experiment, it would help in understanding criteria for LCA approval. Results from two causal inference methods: 1) Rubin Causal Model and 2)Bayesian Networks are negative suggesting not much effect of variables on LCA approval. There are couple of variables being weak causes, but overall there are no effects.

1 Introduction

H-1B is a visa in the United States which allows U.S. employers to temporarily employ foreign workers in various occupations. Number of H-1B visas, that are issued each year, are limited. Hence, it is extremely important for an employer to know the factors which might affect H-1b approval. But, to file petition for H-1B in first place, labor condition application (LCA) should be approved/certified. Hence, we have decided to find the factors which might cause LCA to be approved. LCA is a 6 pages long form. We chose two methods from Causal inference literature to understand the factors : 1) Rubin Causal Mode(RCM) 2) Bayesian Networks.

Why did we choose Rubin Causal Mode(RCM)?[13] We are trying to find causality on observational study, so we don't have control over randomized experiments. For observational study, we cannot assign treatments to research subjects at random. Hence, we decided to use RCM to estimate probabilistic causal effect (average treatment effect) by randomly creating treatment and control groups.

Why did we choose bayesian network[9] :

- They shows relations clearly and can be verified intuitively.
- They represent cause-effect relationships.
- They are probabilistic in nature

32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

Table 1: Discretized variables for RCM tests

Variable Name	Description
case_status	'CERTIFIED' = 1, 'DENIED' = 0
employer_name	employer_name = 1, if employer is in forbes2015 top 100 companies employer_name = 0, Otherwise
prevailing_Wage_encoded	prevailing_Wage_encoded = 1, if prevailing_Wage_encoded >= 90000 prevailing_Wage_encoded = 0, Otherwise
encoded_wage_rate_From	encoded_wage_rate_From = 1, if encoded_wage_rate_From >= 90000 encoded_wage_rate_From = 0, Otherwise
h1b_dependent	'Y' = 1, 'N' = 0

1.1 Dataset

Table 1 shows all the variables which, we believe, might affect outcome variable case_status. We combined data from 4 different H-1B datasets: H-1b2014, H-1b2015, H-1B2016, H-1b2017 into one dataset. This combined dataset has about 1 million records. We want to find the variables which might cause value of our outcome variable "case_status" to be "CERTIFIED", which implies approval of LCA. The dataset was obatined from the United States of Labor Website[5]. The website has data from 2008 but we restrict our range to the above as it has some extra variables such as Wilful Violator and H1-B Dependent.

2 Rubin Causal Model

We are using the framework of Rubin's potential outcome model [13] to find causal effect of different factors on LCA approval. We have used causal inference package[4] for estimating average treatment effect(ATE) except for counterfactual estimator.

For simplicity, we have described all the methods using single treatment variable 'employer_name'. we have mentioned the results for other variables in the discussion section of this part.

2.1 Setting and Notation

Let $Y(0)$ denote the potential outcome of case_status(either 'CERTIFIED' or 'DENIED') in the absence of treatment, and let $Y(1)$ denote the unit's potential outcome when it is treated. Let D denote treatment status, with $D = 1$ indicating treatment(employer_name = 1) and $D = 0$ indicating control(employer_name = 0)l, and let X be a K-column vector of covariates(prevailing_wage, wage_rate_of_pay_from, h1b_dependent). We have taken sample size of 10000 from which 9120 records are in control group and 876 records are in treatment group. As we can see there is huge imbalance between both groups, which is tackled in our testing.

For unit i, $i = 1, 2, \ldots, N$, the observed outcome can be written as

$$Y_i = (1 - D_i)Y_i(0) + D_iY_i(1).$$

2.2 Assumptions

We are assuming that treatment is *strongly ignorable*, as defined in Rosenbaum and Rubin[12]. That is, for all x in the support of X, we have taken from

(i) Unconfoundedness: D is independent of $\big(Y(0), Y(1)\big)$ conditional on $X = x$; (There are no latent variables)

(ii) Overlap: $c < \mathrm{P}(D = 1|X = x) < 1 - c$, for some $c > 0$.(Treatment and control groups are matching well)

2.3 Propensity Score Estimation

Propensity score is the probability of getting treatment conditional on the covariates, $p(X_i) = \mathrm{P}(D_i = 1|X_i)$. Here, we have calculated propensity scores by running a logistic regression

of the treatment indicator D(employer_name) on functions of the covariates(prevailing_wage, wage_rate_of_pay_from, h1b_dependent). This scores will be used in upcoming methods to estimate the average treatment effect.

Results 2: Estimated Parameters of Propensity Score

	Coef.	S.e.	z	P>|z|	[95% Conf.	int.]
Intercept	-2.440	0.115	-21.180	0.000	-2.666	-2.214
X0	0.000	0.000	1.271	0.204	-0.000	0.000
X1	-0.198	0.076	-2.615	0.009	-0.347	-0.050
X2	0.000	0.000	0.105	0.917	-0.000	0.000

2.4 Stratifying the sample

To strengthen the credibility of treatment effect estimates, the treatment and control groups within each propensity bin should be more comparable. To achieve this we have stratified the sample into blocks with units being more similar in terms of their covariates. To stratify, we have used a divide-and-conquer algorithm that recursively divides the sample into two until there is no significant advantage of doing so.(cost = $O(N \log N)$)

Results 3: Stratification Summary

	Propensity Score		Sample Size		Ave. Propensity		Outcome
Stratum	Min.	Max.	Controls	Treated	Controls	Treated	Raw-diff
1	0.071	0.077	2316	186	0.075	0.075	0.003
2	0.077	0.089	2286	212	0.083	0.083	-0.001
3	0.089	0.320	4518	478	0.096	0.097	0.000

2.5 OLS

We start with one of the simplest treatment effect estimator which is Ordinary Least Squares[3] estimator. We are using the following regression:

$$Y_i = \alpha + \beta D_i + \gamma'(X_i - \bar{X}) + \delta' D_i(X_i - \bar{X}) + \varepsilon_i.$$

Results 1: Treatment Effect Estimates: OLS

	Est.	S.e.	z	P>|z|	[95\% Conf.	int.]
ATE	0.002	0.003	0.480	0.631	-0.005	0.008
ATC	0.002	0.003	0.502	0.615	-0.005	0.008
ATT	0.001	0.004	0.257	0.797	-0.006	0.008

Here ATE, ATC, and ATT stand for, respectively, average treatment effect, average treatment effect for the controls, and average treatment effect for the treated.

2.6 Matching

Next, we used the nearest neighborhood matching estimator of Abadie and Imbensr[6] to estimate average treatment effect(ATE). We paired treatment(employer_name=1) and control(employer_name=0)l units by matching directly on the covariate vectors themselves. More specifically, each unit i in the sample is matched with a unit $m(i)$ in the opposite group, where

$$m(i) = \underset{j:D_j \neq D_i}{\operatorname{argmin}} \|X_j - X_i\|,$$

and $\|X_j - X_i\|$ is some measure of distance between the covariate vectors X_j and X_i.

Results 4: Treatment Effect Estimates: Matching

	Est.	S.e.	z	P>|z|	[95% Conf.	int.]
ATE	0.002	0.006	0.350	0.727	-0.010	0.014
ATC	0.002	0.006	0.385	0.701	-0.010	0.014
ATT	-0.001	0.007	-0.115	0.909	-0.014	0.012

2.7 Weighting

We used the Horvitz-Thompson weighting estimator[7](modified for adjusting covariates) to strengthen our results, by running following weight least squares regression:

$$Y_i = \alpha + \beta D_i + \gamma' X_i + \varepsilon_i,$$

where the weight for unit i is $1/\hat{p}(X)$ if i is in the treatment group, and $1/\big(1-\hat{p}(X)\big)$ if i is in the control group.

Results 5: Treatment Effect Estimates: Weighting

	Est.	S.e.	z	P>|z|	[95% Conf.	int.]
ATE	0.001	0.003	0.427	0.669	-0.005	0.008

2.8 Counterfactual Estimator

We used same propensity scores as displayed in Results 2 for matching dataset. To find the optimal matching of treatment and control groups, we used linear_sum_assignment[1] from scipy package, which is based on the Hungarian algorithm[2]. We correctly found 876 records which matched on both groups.

To estimate average treatment effect using counterfactual, first we build following estimators :

$$\hat{Y}_1(1)\& = \alpha_1 + \beta_1 X_i(1) + \epsilon_i, \quad \hat{Y}_0(0)\& = \alpha_0 + \beta_0 X_i(0) + \epsilon_i$$

where the coefficients of $\hat{Y}\_1$ are learned exclusively on samples of X(prevailing_wage + wage_rate_of_pay_from + h1b_dependent) where we've observed employer_name=1, and those of $\hat{Y}\_2$ are likewise are learned from samples where employer_name=0. Next, we just swapped data for both estimators to convert them into counterfactual estimators:

$$\hat{Y}_1(0) = \alpha_1 + \beta_1 X_i(0) + \epsilon_i \,, \quad \hat{Y}_0(1) = \alpha_0 + \beta_0 X_i(1) + \epsilon_i$$

Hence, $\hat{\tau}_i = Y_i - \hat{Y}_1(0)$ for control group and $\hat{\tau}_i = \hat{Y}_0(1) - Y_i$ for treatment group

From this, average treatment would be :

$$\hat{\tau} = \frac{1}{N} \sum_{i=1}^{N} \hat{\tau}_i$$

By using counterfactual estimator, ATE = -0.005

2.9 Discussion

ATE using OLS estimator = 0.002
ATE using the nearest neighborhood matching estimator = 0.002
ATE using Horvitz-Thompson weighting estimator = 0.001
ATE using counterfactual estimator = -0.00536

As we can see from all four average treatment effect estimators, ATE for every model is near to 0. Hence, there is no strong causal effect of treatment variable employer_name on outcome variable case_status provided prevailing_wage, wage_rate_of_pay_from, h1b_dependent as covariates.

Also, we have displayed results for prevailing_wage/wage_rate_of_pay_from/ h1b_dependent as treatment variable and their causal effect on outcome variable case_status in following results section. Overall, there is no causal effect of tested variables on outcome variable.

```
\centering
Results 6: treatment variable - prevailing_Wage_encoded
Summary Statistics
```

Variable	Controls (N_c=3702) Mean	S.d.	Treated (N_t=1298) Mean	S.d.	Raw-diff
Y	0.988	0.107	0.985	0.123	-0.004

Variable	Controls (N_c=3702) Mean	S.d.	Treated (N_t=1298) Mean	S.d.	Nor-diff
X0	0.465	0.499	0.214	0.410	-0.548
X1	0.085	0.279	0.109	0.311	0.080
X2	74413.743	18995.036	129012.452	43852.379	1.616

Treatment Effect Estimates: OLS

	Est.	S.e.	z	P>|z|	[95% Conf.	int.]
ATE	-0.003	0.005	-0.533	0.594	-0.013	0.007
ATC	0.001	0.006	0.140	0.889	-0.011	0.012
ATT	-0.013	0.007	-1.779	0.075	-0.027	0.001

Treatment Effect Estimates: Matching

	Est.	S.e.	z	P>|z|	[95% Conf.	int.]
ATE	0.009	0.058	0.148	0.882	-0.104	0.121
ATC	0.016	0.078	0.203	0.839	-0.136	0.168
ATT	-0.012	0.009	-1.302	0.193	-0.030	0.006

Estimated Parameters of Propensity Score

	Coef.	S.e.	z	P>|z|	[95% Conf.	int.]
Intercept	-10.341	0.313	-32.992	0.000	-10.956	-9.727
X0	0.118	0.108	1.087	0.277	-0.095	0.330
X1	-0.140	0.169	-0.829	0.407	-0.471	0.191
X2	0.000	0.000	32.113	0.000	0.000	0.000

Stratification Summary

	Propensity Score		Sample Size		Ave. Propensity		Outcome
Stratum	Min.	Max.	Controls	Treated	Controls	Treated	Raw-diff
1	0.000	1.000	3702	1298	0.105	0.700	-0.004

Counterfactual estimator
ATE = 0.006

\centering
Results 7: treatment variable - encoded_wage_rate_of_pay_from
Summary Statistics

Variable	Controls (N_c=218) Mean	S.d.	Treated (N_t=151) Mean	S.d.	Raw-diff
Y	0.986	0.117	0.993	0.081	0.007

Variable	Controls (N_c=218) Mean	S.d.	Treated (N_t=151) Mean	S.d.	Nor-diff
X0	0.408	0.493	0.351	0.479	-0.118
X1	0.142	0.350	0.119	0.325	-0.068
X2	77442.110	6009.161	86854.669	8227.241	1.307

Treatment Effect Estimates: OLS

	Est.	S.e.	z	P>|z|	[95% Conf.	int.]
ATE	0.010	0.008	1.350	0.177	-0.005	0.026
ATC	0.017	0.009	2.000	0.046	0.000	0.034
ATT	0.001	0.013	0.040	0.968	-0.025	0.026

Treatment Effect Estimates: Matching

	Est.	S.e.	z	P>|z|	[95% Conf.	int.]
ATE	0.005	0.022	0.235	0.814	-0.038	0.048
ATC	0.014	0.021	0.669	0.504	-0.027	0.054
ATT	-0.007	0.039	-0.188	0.851	-0.084	0.069

Estimated Parameters of Propensity Score

	Coef.	S.e.	z	P>|z|	[95% Conf.	int.]
Intercept	-13.165	0.929	-14.164	0.000	-14.986	-11.343
X0	-1.344	0.242	-5.559	0.000	-1.818	-0.870
X1	0.460	0.352	1.309	0.190	-0.229	1.149
X2	0.000	0.000	14.033	0.000	0.000	0.000

Stratification Summary

Stratum	Propensity Score Min.	Max.	Sample Size Controls	Treated	Ave. Propensity Controls	Treated	Outcome Raw-diff
1	0.083	0.371	153	32	0.195	0.213	0.013
2	0.372	0.485	35	11	0.431	0.417	0.029
3	0.489	0.576	10	13	0.522	0.559	0.000
4	0.587	0.668	4	19	0.627	0.618	0.000
5	0.671	0.917	16	76	0.728	0.826	-0.013

Counterfactual estimator
ATE = 0.017

Results 7: treatment variable - h1b_dependent

Summary Statistics

Variable	Controls (N_c=573) Mean	S.d.	Treated (N_t=374) Mean	S.d.	Raw-diff
Y	0.988	0.110	0.997	0.052	0.010

Variable	Controls (N_c=573) Mean	S.d.	Treated (N_t=374) Mean	S.d.	Nor-diff
X0	88045.826	27730.242	76957.053	18309.776	-0.472
X1	0.086	0.280	0.072	0.259	-0.049
X2	79026.522	25886.228	71559.575	17800.724	-0.336

Treatment Effect Estimates: OLS

	Est.	S.e.	z	P>|z|	[95% Conf.	int.]
ATE	0.011	0.006	2.045	0.041	0.000	0.022
ATC	0.010	0.005	2.072	0.038	0.001	0.020
ATT	0.013	0.007	1.988	0.047	0.000	0.026

Treatment Effect Estimates: Matching

	Est.	S.e.	z	P>|z|	[95% Conf.	int.]
ATE	0.007	0.009	0.794	0.427	-0.011	0.026
ATC	0.009	0.011	0.835	0.404	-0.012	0.030
ATT	0.005	0.012	0.463	0.643	-0.018	0.029

Treatment Effect Estimates: Blocking

	Est.	S.e.	z	P>|z|	[95% Conf.	int.]
ATE	0.011	0.006	1.897	0.058	-0.000	0.022
ATC	0.010	0.005	2.019	0.043	0.000	0.020
ATT	0.011	0.007	1.670	0.095	-0.002	0.025

Estimated Parameters of Propensity Score

	Coef.	S.e.	z	P>|z|	[95% Conf.	int.]
Intercept	0.984	0.239	4.124	0.000	0.516	1.452
X0	-0.000	0.000	-5.086	0.000	-0.000	-0.000
X1	-0.140	0.250	-0.560	0.576	-0.630	0.350
X2	0.000	0.000	2.172	0.030	0.000	0.000

Stratification Summary

	Propensity Score		Sample Size		Ave. Propensity		Outcome
Stratum	Min.	Max.	Controls	Treated	Controls	Treated	Raw-diff

```
-------------------------------------------------------------------------------
         1     0.128     0.327       199        39     0.248     0.251     0.010
         2     0.328     0.385        80        38     0.359     0.361     0.000
         3     0.385     0.420        62        56     0.403     0.405     0.000
         4     0.421     0.611       232       241     0.489     0.471     0.017

Counterfactual estimator
ATE = 0.013
```

3 Bayesian Networks:

3.1 Introduction:

A bayesian belief network is a class of graphical models that represent probabilistic structure of data. It is a combination of:
1. A set of random variables X which we are interested in finding the relationship about.
2. A Directed Acyclic Graph(DAG) which encodes the causal relationship between each node present in the graph.

We are interested in finding a causal bayesian network and for this the DAG must satisfy the following conditions:

- Causal Markov Condition (CMC): This entails that each r.v is conditionally independent of its non-effects given the direct causes.
- Faithfulness: This means that the structure represented by the data matches the true DAG or in other words, the dependencies are true to the data.
- Causal Sufficiency : This condition specifies that we have observed all the variables and that there are no latent confounding variables present that might influence the data.

All the above conditions are not always satisfied in a real world setting, we try and satisfy all the above conditions by using multiple methods to find the best one that can represent the data.

The one condition that is satisfied outright is the causal sufficiency one as all this is a computerized process and the variables given are the variables included in the selection process.
There are different methods to do Bayesian Learning:

3.1.1 Structure Based learning:

The most common type of structure based algorithm is the search and score method:
In this method we assign a score to each BN in the search space, one that helps describe the data (d) in a good way.Given a structure X, the:

$$Score(X, d) = Pr(X|d) \tag{1}$$

We try to maximize the numerator. Here we use the BIC criterion to measure how well we have done, the BIC score [14]:

$$BICscore(X, d) = \log Pr(X|\hat{p}, X) - \frac{d}{2} \log N \tag{2}$$

Here the term $\hat{p}$ is the set of maximum-likelihood estimates of the parametersp of the BN, while d is the number of free parameters of the multivariate Gaussian.

The basic goal is to minimize the BIC score.

For the search and score we use two methods:

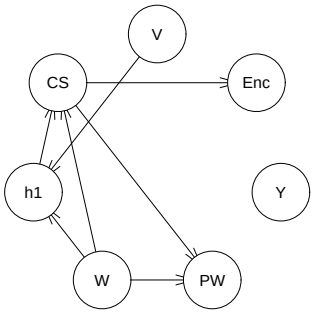

Figure 1: PC algorithm

3.1.2 Hill climbing:

In the Hill climbing method[15]we start with an unconnected graph.The search is started from either an empty, full, or possibly random network, although if there exists background knowledge it can be used to seed the initial candidate network. The algorithm's main loop consists of attempting every possible single-edge addition, removal, or reversal, making the network that increases the score the most the current candidate, and iterating. The process stops when there is no single-edge change that increases the score[9]. This method has it's issues because it can get stuck at a local optimum, this can be solved by performing random restart or by bootstrapping.

3.2 Constraint Based Algorithms:

These are based on the Inductive Causation(IC) algorithm [11] . This type of method learns the DAG using conditional independence tests. The IC algorithm cannot be applied directly as such because of the number of conditional independence relationships it has to test. Thus for our experiments we use the modified version of the above algorithms:

- PC : This is a practical algorithm developed at Carnegie melon university by Sprites. The algorithm assumes faith fullness that there is a directed acyclic graph, G,such that the independence relationships among the variables in X are exactly those represented by G by means of the d-separation criterion [10]. PC algorithm is based on the existence of a procedure which is able of saying when I(A, B|C) is verified in graph G. It first tries to find the skeleton (underlying undirectedgraph) and on a posterior step makes the orientation of the edges.[9]

- GS : the grow shrink algorithm is a forward selection algorithm based on Markov blanket detection, it is considered to be better than PC at run time[9].

In our methodology, we select the discretized parameters as shown in table 2. The networks obtained from the above methods are Shown in the Figures 1,2,3:

Figure 1 is the implementation of the PC algorithm, it gives some edges which make sense intuitively and the BIC = -23159.96.

Figure 2 is the Grow Shrink Algorithm which some nodes similar to the PC algorithm with a BIC= -23677.96.

We implement the Hill Climbing algorithm with bootstrapping 100 times and present the averaged network with BIC = -23530.48.

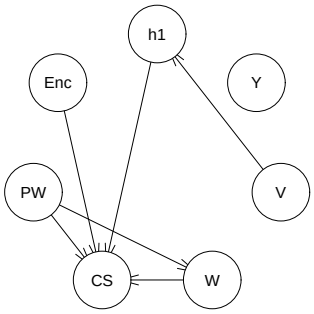

Figure 2: GS algorithm

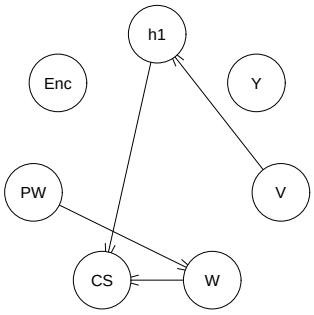

Figure 3: HC algorithm with bootstrapping

As we can see the Grow Shrink(GS) performs the best among the three based on the lower BIC score, thus we select the network produced by it to measure the causal strength.

The causal strength of each relation to ensure the graph represents the causes correctly this is done using the ida function present in the R package pcalg [8]:

The ida() algorithm estimates the causal effect of one variable on another, it is assumed that the distribution is true and under this assumption we estimate the the total causal effect of x on y, this causal effect is measured using pearl's do calculus as follows E(Y|do(X=z+1)) - E(Y|do(X=z)), this is performed using a simple linear regression model.

The threshold is generally set at 0, if the values are below zero then it is usually considered there is no causal relationship. From table 3 we can see that some causal relationships which are strong and some are close to zero which suggest little to no causal effect.

Table 2: Discretized variables for BN

Variable Name	Description
case_status	'CERTIFIED' , 'DENIED'
employer_name	employer_name = A-D,for 25 increments in Forbes 100 ranking, 'E' for everything else
prevailing_Wage_	'high' if >= 90,000 l'low' if <60,000 l'aboveaverage' otherwise.
wage_rate_From	'high' if >= 90,000 l'low' if <60,000 l'aboveaverage' otherwise.
Year submitted	After 2015=A , Before 2015= B.
Dependent	'Yes', 'No'
Wilful Violator	'Yes','No'

Relation	Causal Strength
Violator->Dependent	0.000363632
Dependent->Case Status	0.02272216
Employer->Case Status	0.0118146
Prevailing Wage->Wage	0.5667624
Wage->Case Status	~0
Prevailing Wage->Case Status	~0

Table 3: Causal Values based on IDA

3.3 Discussion

From the bayesian network models we can find that there are weak causes in terms of causal effect but no concrete cause for the case status, but we do find one causal relationship i.e between Prevailing wage and Wage Filed.

4 General Discussion

Hence, we conclude our results as non causal based on results from rubin causal model and bayesian network models.

Our results couldn't find any strong cause for LCA application to be approved. Though, we found prevailing_wage and wage_rate_of_pay to be causally related. But, this relationship doesn't provide any effect of these two variables on our outcome variable case_status.

We cannot ignore the possibility of latent variable(s), which might be unobserved in dataset. There can be many latent variables such as Name, Age, Race, Ethnicity, Gender of the people are not disclosed, these might affect the model as both the models we tested are based on the assumption that we have full knowledge of the variables and there are no confounding factors.

4.1 Future Work

Future work would be to see if it would be possible to get the Complete data set from the Department of labor which include information like age, race , ethnicity.

We could use Structural Equation modelling to test on the data and cross verify the results.

For Bayesian networks we could try tweaking the present models more and test on a more complete dataset.

5 Responsibility Division

Viveksinh Solanki worked on the Rubin Causal Model and interpreted it's results.
Vardaan Kishore Kumar worked on modelling and understanding the Bayesian Directed Acyclic Graphs.
Each have written our own parts in the paper and compared our results.

References

[1] https://docs.scipy.org/doc/scipy-0.18.1/reference/generated/scipy.optimize.linear_sum_assignment.html.

[2] https://en.wikipedia.org/wiki/hungarian_algorithm.

[3] https://en.wikipedia.org/wiki/ordinary_least_squares.

[4] https://github.com/laurencium/causalinference.

[5] https://www.foreignlaborcert.doleta.gov/performancedata.cfm.

[6] Alberto Abadie and Guido Imbens. Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74:235–267, 2006.

[7] D. J. Horvitz, D. G.; Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47:663–685, 1952.

[8] Markus Kalisch, Martin Mächler, Diego Colombo, Marloes H. Maathuis, and Peter Bühlmann. Causal inference using graphical models with the R package pcalg. *Journal of Statistical Software*, 47(11):1–26, 2012.

[9] Dimitris Margaritis. Learning bayesian network model structure from data. Technical report, Carnegie-Mellon Univ Pittsburgh Pa School of Computer Science, 2003.

[10] Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Elsevier, 2014.

[11] Judea Pearl and Thomas S Verma. A theory of inferred causation. In *Studies in Logic and the Foundations of Mathematics*, volume 134, pages 789–811. Elsevier, 1995.

[12] PAUL R. ROSENBAUM and DONALD B. RUBIN. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70:41–55, 1983.

[13] Donald B. Rubin. Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology*, 66:688–701, 1974.

[14] Gideon Schwarz et al. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.

[15] Marco Scutari and Jean-Baptiste Denis. *Bayesian networks: with examples in R*. Chapman and Hall/CRC, 2014.