ATE computations from Baysian Networks in RCTs

This notebook aims to study the capabilities of Bayesian Networks for computing Average Treatment Effects (ATE) in Randomized Control Trials (RCT) under the Neyman-Rubin potential outcome framework.

Consider a set of n independent and identically distributed subjects. an observation on the i-th subject is given by the tuple (T_i, X_i, Y_i) where:

- T_i taking values in $\{0, 1\}$ is a binary random variable representing the treatment.
- X_i is the covariate vector.
- $Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0)$ is the outcome of the treatment on the i-th subject, with $Y_i(1)$ and $Y_i(0)$ representing the treated and untreated outcomes, respectively.

We are interested in quantifying the effect of a given treatment on the population, namely the quantity $\Delta_i = Y_i(1) - Y_i(0)$. Althought this number cannot be directed calculated due to the presence of counterfactuals, there exists methods for approximating its expected value, the Avereage Treatment Effect:

$$\tau = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\Delta_i\right] = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$$

To achive this, we suppose the Stable-Unit-Treatment-Value Assumption (SUTVA) is verified and further assume ignorability between the observations:

- $Y_i = Y_i(T_i)$ (SUTVA)
- $T_i \perp\!\!\!\perp \{Y_i(0), Y_i(1)\}$ (Ignorability)

We will proceed to present estimators of τ using Baysian Networks on generated and real data through three different methods:

- "Exact" Computation
- Parameter Learning
- Structure Learning

1 - Generated Data

We will first consider two generative models in this notebook:

- A linear generative model described by the equation:

$$Y = 3X_1 + 2X_2 - 2X_3 - 0.8X_4 + T(2X_1 + 5X_3 + 3X_4)$$

- And a non-linear generative model described by the equation:

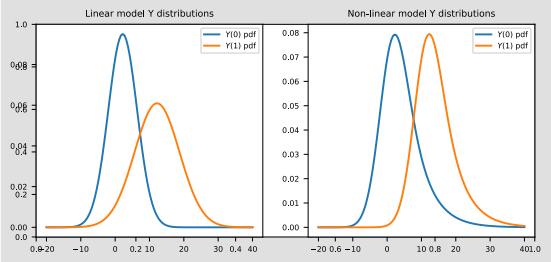
$$Y = 3X_1 + 2X_2^2 - 2X_3 - 0.8X_4 + 10T$$

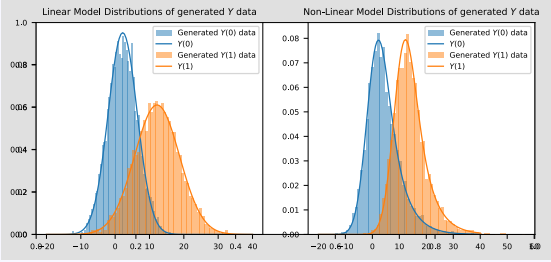
Where $(X_1, X_2, X_3, X_4) \sim \mathcal{N}_4((1, 1, 1, 1), I_4)$, $T \sim \mathcal{B}er(1/2)$ and (X_1, X_2, X_3, X_4, T) are jointly independent in both of the models.

Data from the models can be generated by the functions given below.

Furthermore, the expected values of $Y(0)$ and $Y(1)$ can be explicitly calculated, providing us the theoretical ATE which enables performance evaluations of the estimators.

Both models have an ATE of $\tau = 10$



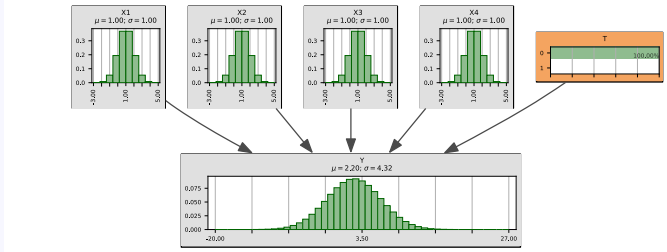


1.1 - "Exact" Computation

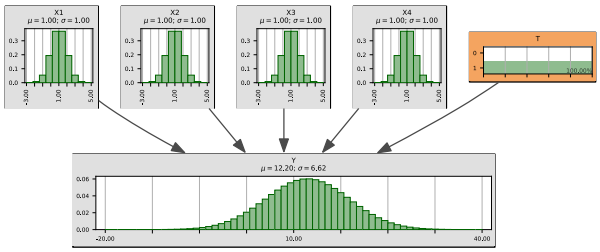
Exact theoretical expected values can be calculated using Bayesian Networks by inputting the data-generating distribution directly into the network. However, since pyAgrum does not support continuous variables as of July 2024, a discretization of continuous distributions is necessary. Consequently, the calculated value will not be exact in a strict sense, but with a sufficient number of discrete states, a close approximation can be achieved.

By employing a 10-bin discretization for the covariates and a 60-bin discretization for the outcome, the Bayesian Network estimator accurately approximates the true distribution for both the treated and untreated outcome.

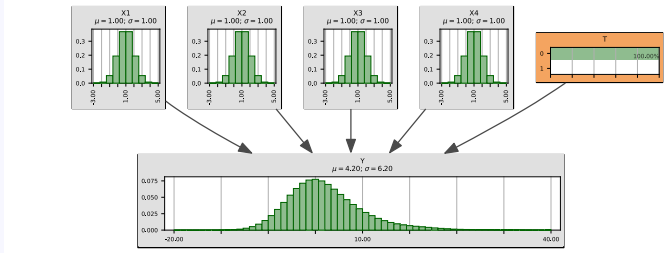
```
BN{nodes: 6, arcs: 5, domainSize: 10^6.07918, dim: 1180037, mem: 9Mo 159Ko 336o}
```



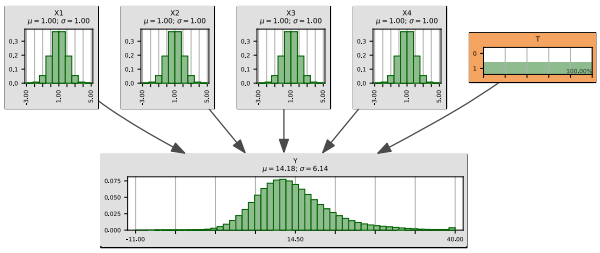
Linear Y(0)



Linear Y(1)



Non-Linear Y(0)



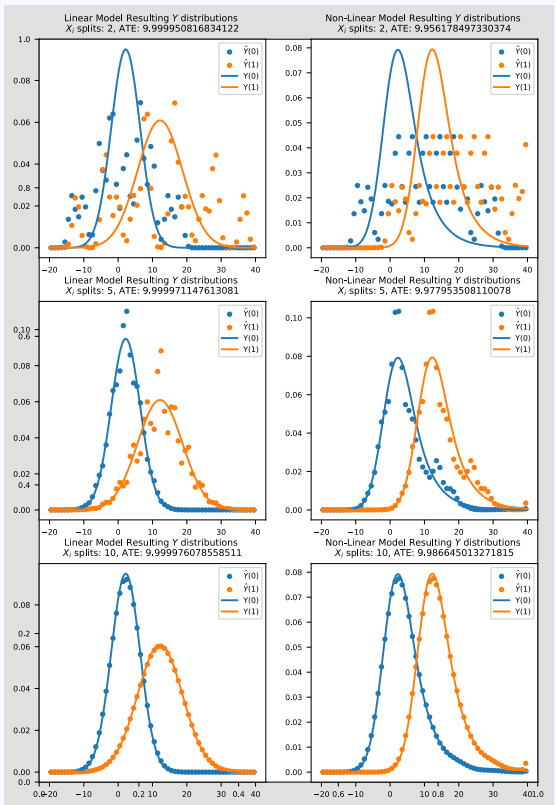
Non-Linear Y(1)

From now on, obtaining the average treatment effect is straightforward using pyAgrum's LazyPropagation exact inference. This involves setting the evidence of the treatment T to 0 and then to 1 to compute the respective posterior CPTs. The ATE is subsequently obtained by performing manipulations on the difference between the CPTs.

```
BN{nodes: 6, arcs: 5, domainSize: 10^6.07918, dim: 1180037, mem: 9Mo 159Ko 336o}
ATE(lin_exbn) = 9.999976078558515

BN{nodes: 6, arcs: 5, domainSize: 10^6.07918, dim: 1180037, mem: 9Mo 159Ko 336o}
ATE(nl_exbn) = 9.986645013271817
```

Let's examine how the fineness of covariate discretization impacts the outcome distribution.

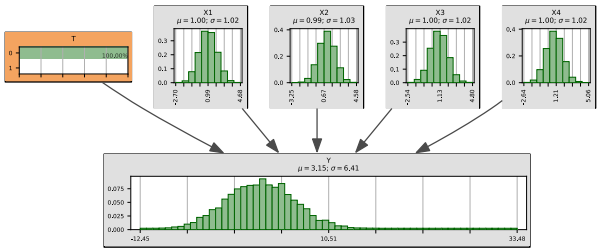


finer discretization with a greater number of bins results in improved approximations of the probability density functions. However, despite the use of rough discretization, the estimations of the ATE remain remarkably accurate.

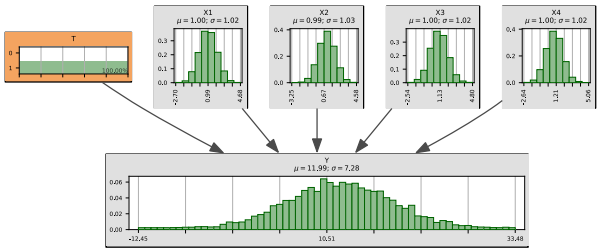
1.2 - Parameter Learning

Given the data generating function defined above, parameter learning methods can be employed to infer the underlying distribution based on the given structure of the Bayesian network. The default Mutual Information based Inference of Causal Networks (MIIC) algorithm utilized in the `BNLearner` class effectively performs this task.

```
Filename       : /var/folders/r1/pj4vdx_n4_d_xpsb04kzf97r0000gp/T/tmpefis6l6e.csv
Size           : (10000,6)
Variables      : T[2], X1[10], X2[10], X3[10], X4[10], Y[60]
Induced types  : False
Missing values : False
Algorithm      : MIIC
Score          : BDeu  (Not used for constraint-based algorithms)
Correction     : NML  (Not used for score-based algorithms)
Prior          : Smoothing
Prior weight   : 0.000001
```

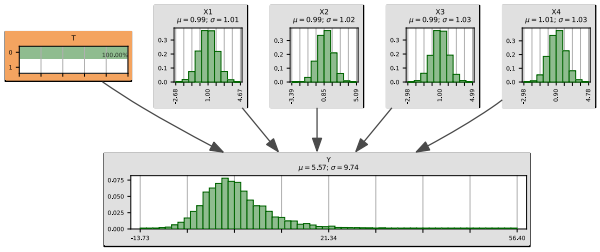


Linear Y(0)

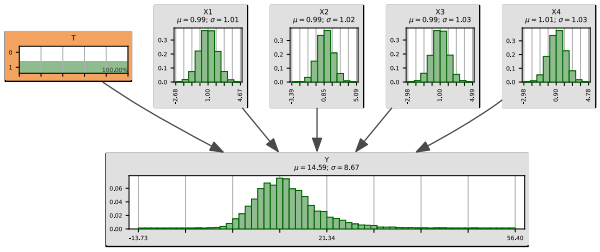


Linear Y(1)

```
Filename       : /var/folders/r1/pj4vdx_n4_d_xpsb04kzf97r0000gp/T/tmp8bxhctbx.csv
Size           : (10000,6)
Variables      : T[2], X1[10], X2[10], X3[10], X4[10], Y[60]
Induced types  : False
Missing values : False
Algorithm      : MIIC
Score          : BDeu  (Not used for constraint-based algorithms)
Correction     : NML  (Not used for score-based algorithms)
Prior          : Smoothing
Prior weight   : 0.000001
```

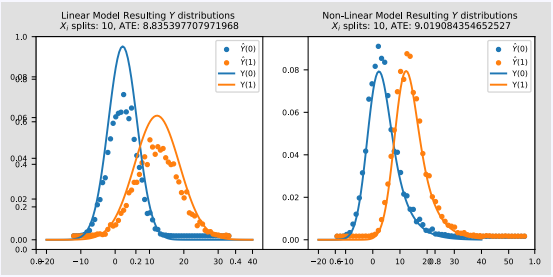


Non-Linear Y(0)

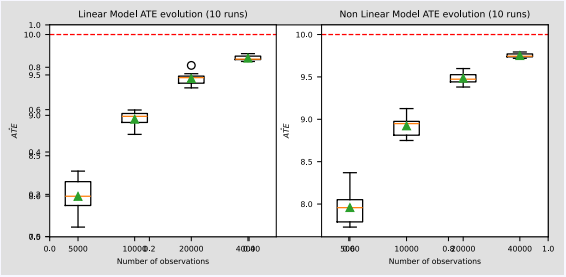


Non-Linear Y(1)

We observe that the inferred outcome distribution generally matches the exact distribution. However, the ATE seems to be biased, as it is consistently smaller.



This underestimation can be further observed with varying numbers of observations in both models.

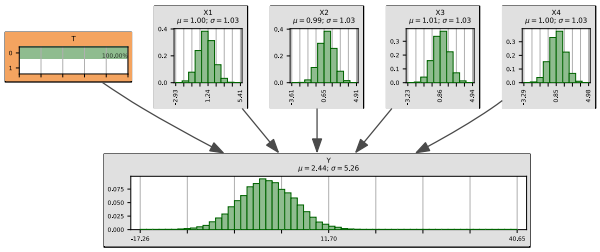


With an increasing number of observations, we observe a convergence of the estimation towards the true value of the ATE and a corresponding reduction in variance.

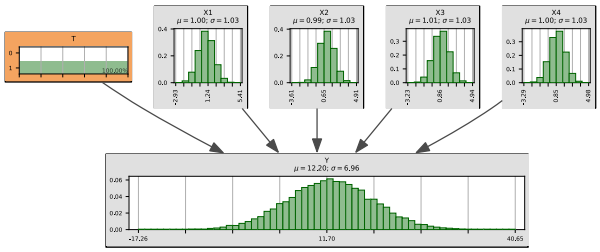
1.3 - Structure Learning

The network's structure and the distributions of the variables can be derived from a sufficiently large dataset through non-parametric learning methods. However, to ensure the integrity of the process, we will impose a slice order on the learner. This ensures that no node is an ancestor of the treatment variable, and no node is a descendant of the outcome variable.

```
Filename               : /var/folders/r1/pj4vdx_n4_d_xpsb04kzf97r0000gp/T/tmpuw7ro_nf.csv
Size                   : (40000,6)
Variables              : T[2], X1[10], X2[10], X3[10], X4[10], Y[60]
Induced types          : False
Missing values         : False
Algorithm              : MIIC
Score                  : BDeu  (Not used for constraint-based algorithms)
Correction             : NML  (Not used for score-based algorithms)
Prior                  : Smoothing
Prior weight           : 0.000001
Constraint Slice Order : {X3:1, X1:1, X4:1, T:0, X2:1, Y:2}
```

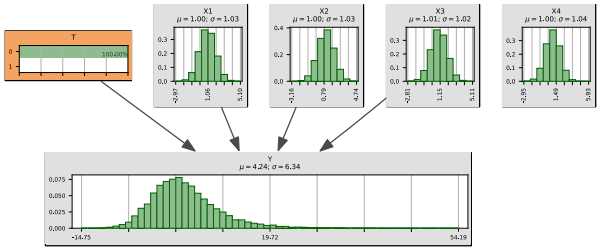


Linear Y(0)

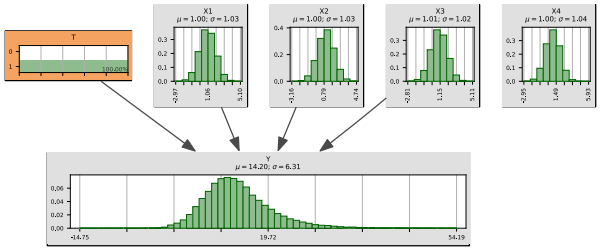


Linear Y(1)

```
Filename              : /var/folders/r1/pj4vdx_n4_d_xpsb04kzf97r0000gp/T/tmp2djddqwq.csv
Size                  : (40000,6)
Variables             : T[2], X1[10], X2[10], X3[10], X4[10], Y[60]
Induced types         : False
Missing values        : False
Algorithm             : MIIC
Score                 : BDeu  (Not used for constraint-based algorithms)
Correction            : NML  (Not used for score-based algorithms)
Prior                 : Smoothing
Prior weight          : 0.000001
Constraint Slice Order : {X3:1, X1:1, X4:1, T:0, X2:1, Y:2}
```

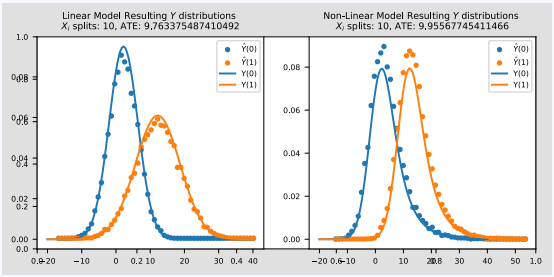


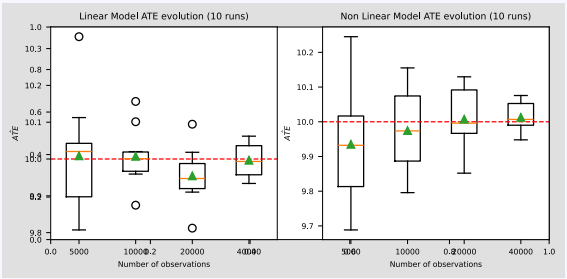
Non-Linear Y(0)



Non-Linear Y(1)

With 10,000 observations, structure learning yields a more accurate estimation of the Average Treatment Effect (ATE) compared to parameter learning. This improvement can be attributed to the use of a less complex structure, as opposed to the structure used previously, which features a higher in-degree on the outcome node. The increased number of parameters to be estimated in the outcome model due to this higher in-degree is suboptimal for smaller datasets.





2 - Real Data

After evaluating various estimation methods using generated data, we will now direct our attention to real data from the Tennessee Student/Teacher Achievement Ratio (STAR) trial. This randomized controlled trial, initiated in 1985, is a pioneering study in the field of education, designed to assess the effects of smaller class sizes in primary schools (T) on students' academic performance (Y).

The covariates in this study include:

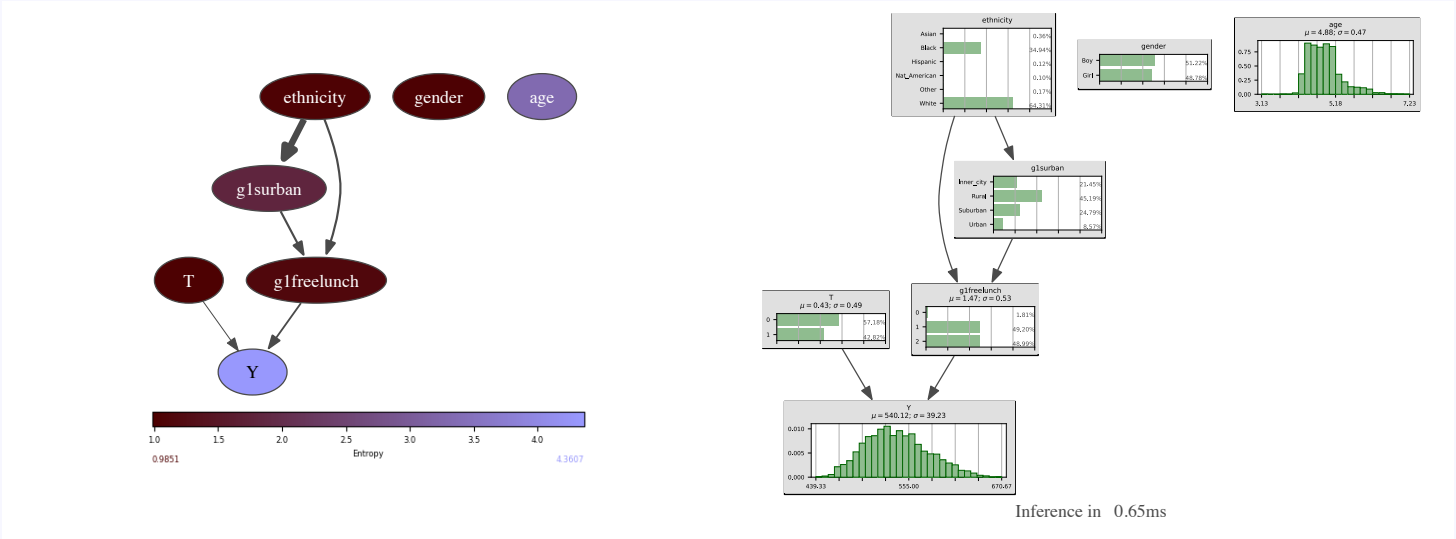
- `gender`
- `age`
- `g1freelunch` being the number of lunchs provided to the child per day
- `g1surban` the localisation of the school (inner city or rural)
- `ethnicity`

	Y	T	gender	ethnicity	age	g1freelunch	g1surban
0	514.000000	0	Boy	White	4.596851	2	Rural
1	512.666667	0	Girl	Black	5.694730	1	Inner_city
2	470.333333	1	Girl	Black	4.180698	1	Suburban
3	500.666667	1	Girl	White	5.963039	2	Urban
4	516.333333	0	Boy	Black	5.867214	1	Inner_city

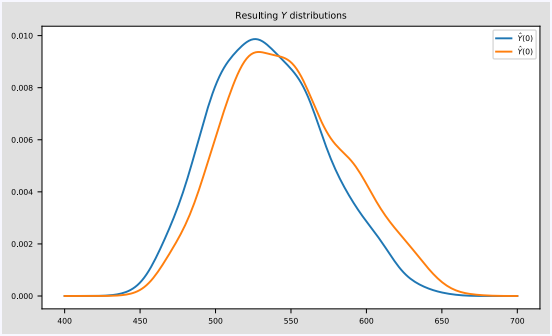
2.1 - Structure Learning

n the absence of prior knowledge regarding the underlying distributions of the variables and their relationships, causal inference can be challenging. Consequently, we will first utilize structure learning to automatically identify the network's underlying structure. To assist the learning process, we will impose a slice order on the variables once again.

```
Filename               : /var/folders/r1/pj4vdx_n4_d_xpsb04kzf97r0000gp/T/tmpm23i3zqq.csv
Size                   : (4215,7)
Variables              : Y[30], T[2], gender[2], ethnicity[6], age[24], g1freelunch[3], g1surban[4]
Induced types          : False
Missing values         : False
Algorithm              : MIIC
Score                  : BDeu  (Not used for constraint-based algorithms)
Correction             : NML  (Not used for score-based algorithms)
Prior                  : Smoothing
Prior weight           : 0.000001
Constraint Slice Order : {ethnicity:0, T:0, g1surban:1, age:0, gender:0, g1freelunch:1, Y:2}
```



This initial approach appears promising, as the inferred causal relationships are somewhat consistent with what might be expected from an non-expert perspective.



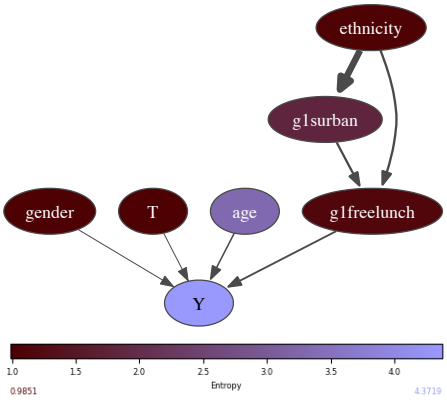

```
Estimated ATE : 11.51375626395145
```

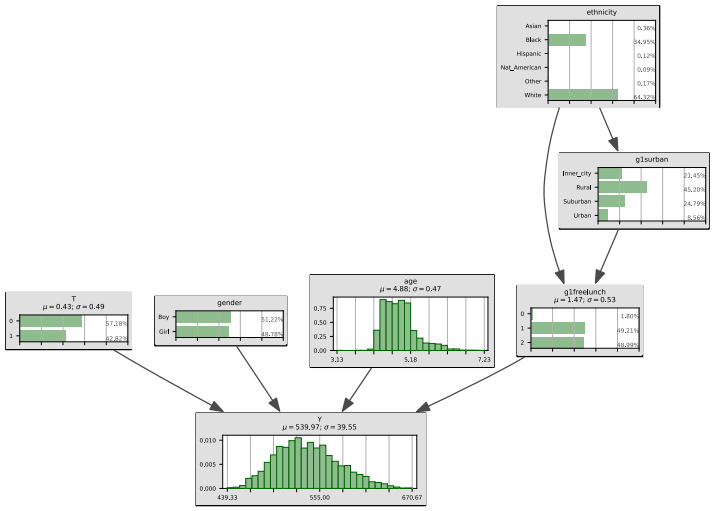
We observe a slight change in the outcome distribution. However, since the outcome takes values in the hundreds, this results in a non-negligeable impact on the treatment effect, given that the outcome is defined as the average of the students' three grades.

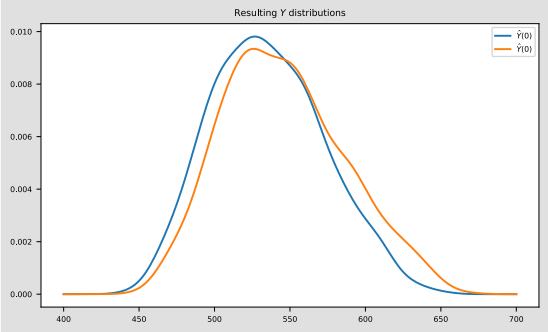
2.2 - Parameter Learning

Using different structures when conducting parameter learning can yield varying results. For the sake of illustration, we will examine how the estimation performs when arcs from the `age` and `gender` covariates are added to the outcome.

```
Filename       : /var/folders/r1/pj4vdx_n4_d_xpsb04kzf97r0000gp/T/tmp3a_80_au.csv
Size           : (4215,7)
Variables      : Y[30], T[2], gender[2], ethnicity[6], age[24], g1freelunch[3], g1surban[4]
Induced types  : False
Missing values : False
Algorithm      : MIIC
Score          : BDeu  (Not used for constraint-based algorithms)
Correction     : NML  (Not used for score-based algorithms)
Prior          : Smoothing
Prior weight   : 0.000001
```








```
Estimated ATE : 10.344241933416356
```

As anticipated, there are observable differences between the parameter learning method and the structure learning method. When compared to direct estimation methods, such as the Difference in Means (DM) estimator and the Ordinary Least Squares (OLS) estimator, which yield average treatment effects of 12.81 and 10.77, respectively, our findings remain largely consistent. These results suggest that incorporating age and gender variables into the outcome model may deteriorate the final estimation accuracy.