Analyzing the Effect of Salary on Employee Attrition

University of California, Berkeley, PH252D

Calvin Chi, Daniel Lee, Dario Cantore, Josiah Davis

May 8, 2017

1 Specify the causal question

The causal question we want to answer is the effect salary has on the probability of leaving a company. Specifically, we are interested in the difference in probability of leaving if employees have a high salary versus if employees have a low salary. It is not clear how high and low salary is defined. For example, it may be high and low relative to the department of the employee, high and low relative to the company, or high and low compared to the market value for a particular type of role.

Beyond the causal question, we are also interested in ranking the relative importance of an employee's salary, employee's score on the most recent evaluation, employee's satisfaction level, and employee's average monthly hours in affecting the employee's probability of leaving a company. To accomplish this, we will estimate a variable importance measure (VIM) for each of the mentioned variables with the G-computation formula, controlling for all the other variables. The VIMs will only have a statistical interpretation for comparing the importance between variables.

Our population is a simulated dataset of 14,999 observations (employees) from Kaggle. We hope to be able to draw inference on the entire population of employees. It is not clear from the documentation whether this population is from a single company or from multiple companies.

2 Specify the causal model

Let us begin by describing our thought processes in constructing our causal directed acyclic graph (DAG). We started out with a complete graph with edges between all nodes, where each node is a separate variable. The variables in our dataset were:

- whether the employee was promoted in the last 5 years (*promotion*)
- the number of projects an employee has completed while at the company (*#projects*)

- whether the employee had a work accident or not (*work_accident* or *accident*)
- the time an employee has spent at the company (*time_spend_company*)
- the average monthly hours (*avg_monthly_hours* or *hours*)
- the employee's department (*department*)
- the employee satisfaction level (*satisfaction*)
- the score on the employee's last (most recent) evaluation (*last_evaluation* or *evaluation*)
- the employee's salary (*salary*)
- whether the employee left (*left*)

In the original graph we could identify 4 exclusion restrictions: between *department* and *promotion*, between *work_accident* and *promotion*, between *salary* and *work_accident* and between *department* and *last_evaluation*. For example, we thought that the one between *department* and *last_evaluation* warranted an exclusion restriction because the department an employee is in should not affect the last evaluation an employee has received (e.g. if all departments are assumed to give out the same "grade scale", imagine their evaluations to be "curved" like in college), and our experience told us this is true in practice. Conversely, the last evaluation received should not affect the department one works in. As another example, we justified that there could be an exclusion restriction between *department* and *promotion* because, generally, it seemed reasonable to assume that employees receive promotions independently of which department they are in. We also assumed that, based on our understanding of the workplace settings, *work_accident* and *promotion* do not affect each other, nor do *salary* and *work_accident*. Except for these four edges, our initial graph would be fully connected.

However, we soon stumbled on other difficulties. For example, we found reason for an edge between *avg_monthly_hours* and *work_accident* (working more hours could potentially increase the risk of a work accident) but also the other way around (after an accident people might work less). We therefore decided that we do not want to assume directionality between these two nodes. To resolve this issue, we decided to combine these two variables.

It is important to note that we did not combine them because we knew that the interaction is bidirectional; we did so because we could not exclude the possibility that it is. We stumbled upon the same problem when looking at the nodes *#projects* and *average_monthly_hours*, namely, that both could potentially influence each other. Because we already combined *work_accident* and *avg_monthly_hours*, we now had to combine all three of those variables. Because of the way *#projects* is defined (number of projects completed while in company) it might be influenced by *time_spend_in_company* (someone who has been at the company for a longer time might have been able to complete more projects). However, the time spent in the company is probably influenced by whether someone had a work accident. Because *work_accident* and *#projects* have been combined into a node, there would now be an arrow

between *time_spend_in_company* and the combined node, and between the combined node and *time_spend_in_company*. The resulting graph would therefore not be acyclic anymore, and we had to combine the already combined node and *time_spend_in_company*. The same logic applied for the *promotion* variable.

Additionally, we felt that we should also combine the variables *last_evaluation* and $satisfaction$, again because we could not say for sure that the edge connecting the two would be one directional. So, our "true" causal graph consisted of the following five nodes:

1. *department* as W_1
2. *promotion*, *#projects*, *work_accident*, *time_spend_company* and *avg_monthly_hours* as W_2
3. *salary* as A
4. *last_evaluation* and $satisfaction$ as Z
5. $left$ as Y

Furthermore, we decided that *department* might influence *avg_monthly_hours* but not the other way around, and that *department* might influence $satisfaction$, *salary*, and $left$. The variable $left$ is in turn also affected by *salary*, *department* and the combined node W_2. We also decided that at least in the initial, "true" causal graph, no independence assumptions were warranted.

We were aware of the fact that drawing a simple "W-A-Y" DAG, which includes only three nodes, might have saved us a lot of work while at the same time contain our graph as a special case. However, we decided that breaking up the nodes into W_1, W_2, and Z in our model would prove its worth later in the roadmap.

Our endogenous variables are $X = \{W_1, W_2, A, Z, Y\}$, our exogenous errors are the following:

$$U = (U_{W_1}, U_{W_2}, U_A, U_Z, U_Y)$$

The structural equations that follow from this are:

$$\begin{aligned} f_{W_1} &= f(U_{W_1}) \\ f_{W_2} &= f(W_1, U_{W_2}) \\ f_A &= f(W_1, W_2, U_A) \\ f_Z &= f(W_1, W_2, A, U_Z) \\ f_Y &= f(W_1, W_2, A, Z, U_Y) \end{aligned}$$

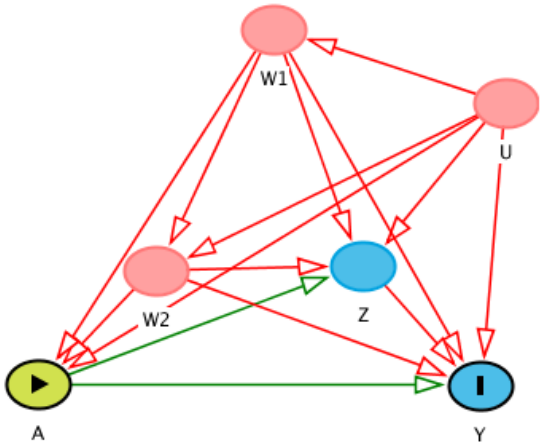


Figure 1: This is our causal directed acyclic graph (DAG).

3 Specify the causal parameter of interest

We denote our target causal parameter $\Psi^F(P_{U,X})$ and define it by the Average Treatment Effect (ATE):

$$\Psi^F(P_{U,X}) = E_{U,X}(Y_{high} - Y_{low}),$$

where Y_{high} and Y_{low} are counterfactual outcomes for *salary*.

4 Specify your observed data and its link to the causal model

The observed data O consists of the following:

- Baseline covariates W_1 and W_2
- Exposure A
- Mediator Z
- Outcome Y

where W_1, W_2, A, Z, and Y are defined above. The random variable O has distribution P_0:

$$O = (W_1, W_2, A, Z, Y) \sim P_0$$

This gives us $n = 14,999$ i.i.d copies $O_1, O_2, ..., O_n$ drawn from probability distribution P_0, which is the underlying probability distribution that are contained in the set of possible distributions implied by the structural causal model. In other words, we assume the observed data were generated by sampling n times from a data generating system contained in the

structural causal model $\mathcal{M}^F$. This provides a link between the causal model $\mathcal{M}^F$ and the statistical model $\mathcal{M}$, which is the set of possible observed data distributions. Another way of thinking about this is that the causal structural model accurately captures or describes the process that gave rise to the observed data. We have not placed any restrictions on the statistical model, which is thereby non-parametric.

5 Identify

5.1 Assessment of the Back-Door Criterion

Our target causal parameter $\Psi^F(P_{U,X}) = E_{U,X}(Y_{high} - Y_{low})$ is not identified under the initial causal model because there are several unblocked back-door paths from outcome Y to exposure A. See Figure 1.

To satisfy the back-door criteria, we would need to place independence assumptions on the distribution of unmeasured factors P_U. Specifically, there are four ways that the causal parameter can be identified as some parameter of our observed data distribution. Then, when we condition on W_1 and W_2, we would meet the back-door criteria. See Table 1 and Figure 2.

Table 1: Four possible sets of independence assumptions to satisfy back-door criteria

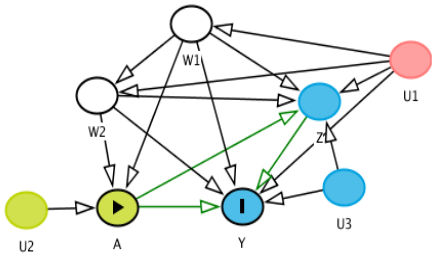
Set 1	Set 2	Set 3	Set 4
$U_{W_2} \perp U_Z$	$U_{W_2} \perp U_Z$	$U_{W_1} \perp U_{W_2}$	$U_A \perp U_{W_2}$
$U_{W_2} \perp U_Y$	$U_{W_2} \perp U_Y$	$U_{W_1} \perp U_Z$	$U_A \perp U_Z$
$U_{W_1} \perp U_{W_2}$	$U_{W_1} \perp U_Z$	$U_{W_1} \perp U_Y$	$U_A \perp U_Y$
$U_A \perp U_Z$	$U_{W_1} \perp U_Y$	$U_A \perp U_{W_2}$	$U_A \perp U_{W_1}$
$U_A \perp U_{W_1}$	$U_A \perp U_Z$	$U_A \perp U_Z$	
$U_A \perp U_Y$	$U_A \perp U_Y$	$U_A \perp U_Y$	

It is hard to say whether one set of independence assumptions is more plausible than the other. W_1, W_2, A, Z, and Y all have unmeasured common causes. For example, ethnicity might affect an employee's *last_evaluation*, *salary*, and $left$. Other examples are age, sex, personality, physical appearance, and intellectual capability. However, for the purposes of trying to attain a causal effect between A and Y, we will have to make these independence assumptions. Set 4 states that the unmeasured variables affecting A are independent of all other unmeasured variables. This means that *salary* is randomly assigned to the employees. This is unrealistic as *salary* is not assigned randomly in practice. So we eliminate set 4.

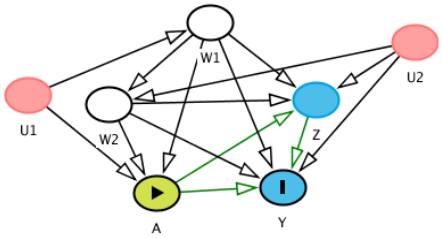
The remaining three sets of independence assumptions do not seem realistic as well. However, if we have to choose one of the remaining three sets of independence assumptions to satisfy the back-door criterion, we would choose set 3. The reason for this is because set 3 is making the most amount of independence assumptions between the unmeasured variables

affecting *department* and the other unmeasured variables. Here, we are stating that the unmeasured variables affecting *department* is independent of unmeasured variables affecting other variables. This is unrealistic, as we know that some unmeasured variables that affect W_1 may also affect W_2, Z, and Y. For example, intellectual capability may affect one's choice of *department*, *avg_monthly_hours*, *satisfaction*, and $left$. However, it is more reasonable than stating that unmeasured variables that affect W_2 is independent of the unmeasured variables that affect Z. For example, whether an individual's work ethic may directly affect *promotion* (W_2). Likewise, it would be difficult to assume that work ethic is independent of *last_evaluation* (Z). However, it may be more reasonable to assume that work ethic is independent of *department* (W_1). That is, we can assume that every department will have individuals with varying levels of work ethic. Again, this may not necessarily be true in real life, but when compared to the other options, this would be the most reasonable assumption. So, we will choose set 3 over set 1 or set 2 as the independence assumptions we make to satisfy the back-door criteria.

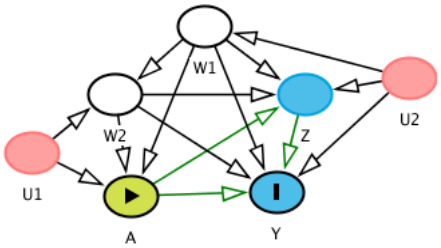
This is one of the limitations we have given our data. In order to satisfy the back-door criterion, we would have to measure more variables as well as change the design of the study. In the context of *salary* and$left$, additional measurements such as sex, age, ethnicity, socioeconomic background, and education level may help in achieving identifiability. Further, what we can consider is doing a longitudinal study in which we follow employees over a period of time, taking measures of variables at different time periods. This may make it more plausible to satisfy the back-door criterion.



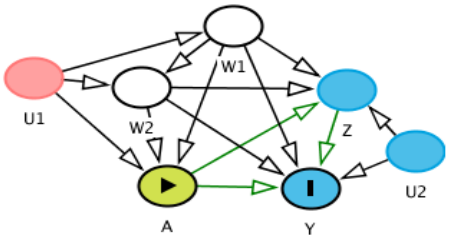
(a) Set 1



(b) Set 2



(c) Set 3



(d) Set 4

Figure 2: Possible sets of assumptions that meet the back-door criteria after conditioning on W_1 and W_2

5.2 Practical Positivity Assumption Evaluation

The positivity assumption states that for every combination of covariates in the observed data distribution, all levels of treatment must be included, at least once.

$$\min_{a \in A} \quad P_0(A = a|W = w) > 0$$

This is important for evaluation of the Average Treatment Effect using the G-computation formula. If the positivity assumption is not met, the G-computation formula will not be well-defined.

One of the interesting aspects of our project is that it has high dimensionality. This makes satisfying the positivity assumption from a practical perspective more difficult. Our approach for handling positivity assumption violations was to remove those strata where positivity assumptions were violated [1]. Dropping strata effectively shrinks our target population for the study.

We have a mixture of both qualitative (e.g., *department*) and quantitative variables (e.g., *avg_monthly_hours*). In order to evaluate the positivity assumption from a practical perspective, we split up the quantitative variables into quantile buckets. We also used these quantile buckets for the estimation of the average conditional probability and treatment mechanism.

We selected to split the *avg_monthly_hours* into 5 quantile buckets and *#projects* and *time_spend_company* into 2 quantile buckets.

How did we choose this? Our goal was to maximize the amount of information contained in the approximated variables while also removing a minimal amount of observations due to strata that violated the practical positivity assumption.

Out of 8 total covariates, we had five that were quantitative.

There are two variables that were not approximated because they are mediator variables in our causal graph. This is because we do not need to satisfy the positivity assumption for variables that we are not conditioning on for the G-computation formula.

- $satisfaction$
- $last\_evaluation$

There is one continuous quantitative variable that is not a mediator variable. Quantiles were used to approximate this variable using buckets of size $k = 2, 3, ..., 10$:

- $avg\_monthly\_hours$

There are two discrete quantitative variables. Quantiles are used to approximate the variables using buckets of size $k = 2, 3$:

- *#projects*
- *time_spend_company*

The results from our evaluation are presented in Table 2. We decided to split the *avg_monthly_hours* into 5 quantile buckets and *#projects* and *time_spend_company* into 2 quantile buckets. We remove the strata (approximately 8.1 percent of the sample) that did not satisfy the practical positivity assumption.

Table 2: Information Loss due to Practical Positivity Assumption Violation

Number of Quantile Buckets			
#projects	*time_spend_company*	*avg_monthly_hours*	% Dropped
2	2	2	3.2
2	2	3	4.9
2	2	4	7
2	**2**	**5**	**8.1**
2	2	6	11.6
2	2	7	12.5
2	2	8	13.9
2	2	9	15.5
2	2	10	16.5
3	3	2	8.9
3	3	3	11.4
3	3	4	14.9
3	3	5	18.8
3	3	6	21.4
3	3	7	24.9
3	3	8	26.5
3	3	9	28.4
3	3	10	30.5

5.3 $\mathcal{M}^{\mathcal{F}*}$

With the back-door criterion satisfied and the positivity assumption met, we can achieve identifiability. We use $\mathcal{M}^{\mathcal{F}*}$ to denote the original SCM, augmented with the independence assumptions and positivity assumptions needed for identifiability. This notation is used to make it explicit that we are making assumptions to be able to answer our causal question. That is, we need the additional assumptions to our original causal structural model in order for our target causal parameter, average treatment effect, to be turned into a parameter of the observed data distribution. However, these assumptions do not reflect our real knowledge of the complex relationship between *salary* and $left$.

Under the working SCM $\mathcal{M}^{\mathcal{F}*}$, the average treatment effect $\Psi^F(P_{U,X})$ is identified using the G-Computation formula:

$$\begin{aligned}\Psi^F(P_{U,X}) &= \Psi(P_0) \\ &= E_w[E_0(Y|A=1,W_1,W_2) - E_0(Y|A=0,W_1,W_2)] \\ &= \sum_{w1,w2}[E_0(Y|A=1,w_1,w_2) - E_0(Y|A=0,w_1,w_2)]P_0(w_1,w_2)\end{aligned}$$

The statistical estimand $\Psi(P_0)$ is the difference in the strata-specific conditional probability of an employee leaving a company when receiving high salary and under low salary, averaged with respect to the distribution of the baseline covariates, which are *department*, *#projects*, *work_accident*, *time_spend_company*, and *avg_monthly_hours*.

6 Estimate

6.1 Super Learning

Super Learning [2] was used to estimate both the conditional probability for leaving the company: $\bar{Q}_0(Y|A,W)$ and the treatment mechanism of having a high salary: $g(A_i|W_i) = P(A_i|W_i)$. The Super Learning Approach incorporated techniques from a variety of different types of data-adaptive algorithms [3]:

- Tree-based/Additive methods
- Linear methods
- Prototype methods
- Neural Networks

Where it was thought that a data-adaptive algorithm was sensitive to certain hyper-parameters, training was conducted on a range of values for this hyper-parameter. Some of the algorithms already contained a procedure internally for selecting the best hyper-parameters. The results from running the Super Learner are presented in Table 3.

Super Learner and Discrete Super Learner Risk are obtained from cross-validation of the Super Learner. The other risk values are obtained from cross-validation of the algorithm. The coefficient values are associated with the relative weight each algorithm is given in making the final prediction.

- **Tree-based methods**: The data-adaptive algorithms with the lowest cross-validated risk was the Gradient Boosted Tree algorithm. The other tree-based methods had low cross-validated risk as well. When the results were cross-validated in order to evaluate the super learner, the Gradient Boosted Tree algorithm was chosen as the Discrete Super Learner and performed as well on average as the Super Learner (although the maximum risk was lower for the Super Learner than for the discrete Super Learner).

[1]Super Learner and Discrete Super Learner do not have coefficient values by definition.

Table 3: Super Learning Results for estimator of $\bar{Q}_0(Y|A,W)$

Algorithm	Risk	Coefficient[1]
Super Learner	0.1088	—
Discrete Super Learner	0.1088	—
GLM	0.1698	0
Ridge	0.1697	0
LASSO	0.1696	0
GAM	0.1539	0
Gradient Boosting	0.1092	0.4521
Random Forest	0.1236	0
CART	0.1145	0.0296
MARS	0.1097	0.3513
Neural Net (nodes = 2)	0.2445	0.0377
Neural Net (nodes = 3)	0.1585	0.0105
Neural Net (nodes = 4)	0.1483	0
Neural Net (nodes = 5)	0.2081	0
Neural Net (nodes = 6)	0.1463	0.0112
KNN (k = 10)	0.1123	0.0897
KNN (k = 15)	0.1143	0.0177
KNN (k = 20)	0.1166	0
KNN (k = 25)	0.1181	0
Mean	0.1977	0

- **Linear methods**: Many of the linear methods had high cross-validated risk values. None of the linear methods were given any weight in making the final prediction.
- **Prototype methods**: K-nearest neighbors had a lower cross-validated risk than all of the Linear and Neural Network-based methods and was generally competitive with the Tree-based/additive methods. This is somewhat surprising given the simplicity of the K-nearest neighbors model.
- **Neural Networks**: There was a sizable difference in the cross-validated risk and coefficient values for the Neural Network algorithms, depending on how many hidden nodes were used in the training process. Some Neural Networks did worse than the mean, suggesting that the Neural Networks may be over-fitting the data.

6.2 Estimators of the ATE

We applied the simple substitution estimator, inverse probability of treatment weighted (IPTW) estimator, the modified Horvitz-Thompson estimator and targeted maximum likelihood estimator (TMLE) to estimate the average treatment effect (ATE) as specified by the G-computation formula.

Each estimator will require different factors of the observed data distribution $P_0(O)$. The simple substitution estimator only requires calculation of $\bar{Q}_n(A, W)$, which is an estimator of $E_0(Y|A, W)$.

$$\hat{\Psi}_{SSE}(P_n) = \frac{1}{n}\sum_{i=1}^{n}\left(\bar{Q}_n(1, W_i) - \bar{Q}_n(0, W_i))\right)$$

The standard IPTW estimator (also called the Horvitz-Thompson estimator) will only require an estimate of $g(A_i|W_i)$, the treatment mechanism.

$$\hat{\Psi}_{IPTW}(P_n) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{I(A_i = 1)}{g(A_i|W_i)}Y_i - \frac{I(A_i = 0)}{g(A_i|W_i)}Y_i\right)$$

Similar to the standard IPTW estimator, stabilized IPTW estimator (also called the modified Horvitz-Thompson estimator) only requires an estimate of $g(A_i|W_i)$.

$$\hat{\Psi}_{ST.IPTW}(P_n) = \left(\frac{\sum_{i=1}^{n}\frac{I(A_i=1)}{g(A_i|W_i)}Y_i}{\sum_{i=1}^{n}\frac{I(A_i=1)}{g(A_i|W_i)}} - \frac{\sum_{i=1}^{n}\frac{I(A_i=0)}{g(A_i|W_i)}Y_i}{\sum_{i=1}^{n}\frac{I(A_i=0)}{g(A_i|W_i)}}\right)$$

Lastly, like the simple substitution estimator, TMLE estimates $E_0(Y|A, W)$ with $\bar{Q}^*_n(A, W)$. However, in order to estimate $\bar{Q}^*_n(A, W)$, TMLE requires an initial estimate $\bar{Q}^0_n(A, W) = E_n(Y|A, W)$ as well as $g_n(A_i|W_i)$. The formula for the TMLE estimator is:

$$\hat{\Psi}_{TMLE}(P_n) = \frac{1}{n}\sum_{i=1}^{n}\left(\bar{Q}^*_n(1, W_i) - \bar{Q}^*_n(0, W_i))\right)$$

where

$$\bar{Q}^*_n(1, W) = expit(logit(\bar{Q}^0_n(1, W_i)) + \epsilon_n H_n(1, W_i))$$
$$\bar{Q}^*_n(0, W) = expit(logit(\bar{Q}^0_n(0, W_i)) + \epsilon_n H_n(0, W_i))$$

and

$$H_n(A_i, W_i) \equiv \left(\frac{I(A_i = 1)}{g(A_i|W_i)} - \frac{I(A_i = 0)}{g(A_i|W_i)}\right)$$

All estimates in this project are performed using Superlearner as a data-adaptive estimation technique according to the previous description.

With the estimates $\bar{Q}^0_n(A, W)$ and $g_n(A_i|W_i)$ produced by Superlearner, the simple substitution estimator, IPTW and stabilized IPTW are implemented as described by their respective formulas. The TMLE package is used to find the updated estimate $\bar{Q}^1_n(A, W)$ using both $\bar{Q}^0_n(A, W)$ and $g_n(A_i|W_i)$ as input. The estimates $\bar{Q}^1_n(1, W)$ and $\bar{Q}^1_n(0, W)$ are then plugged into the G-computation formula to obtain the estimate for our estimand, the ATE. TMLE offers several advantages over the other two estimators, namely:

1. Serves as a plug-in estimator and provides numerical stability not provided by IPTW.

2. Consistent estimator of ATE if either the outcome or treatment mechanism is consistently estimated.

3. Provides the most efficient estimator of ATE if both the outcome and treatment mechanisms are estimated consistently.

Inference can be performed with all 4 estimators. Variance estimates for the simple substitution estimator, IPTW, Stabilized IPTW and TMLE are implemented using the non-parametric bootstrap. The pseudocode for the non-parametric bootstrap can be described as:

procedure NONPARAMETRIC-BOOTSTRAP(X = {W, A, Y}, B)
  for i in 1 to B **do**
    Draw n samples from X with replacement to form bootstrap sample X_i^*
    Estimate Ψ_i^* from X_i^*
  Return $\Psi_1^*, ..., \Psi_n^*$

The variance of the estimator can then be approximated with

$$\sigma^2_{\hat{\Psi}(P_n)} = \frac{1}{B}\sum_{i=1}^{B}(\Psi_i^* - \bar{\Psi}_i^*)^2$$

The variance of the TMLE estimator is also estimated from the variance of the estimating influence curve (IC)

$$\sigma^2_{\hat{\Psi}(P_n)} = \frac{\text{var}(IC_n)}{n}$$

where the IC is a function that maps every single observation to a value. A 95% confidence interval can be constructed as

$$\hat{\Psi}(P_n) \pm 1.96\, \sigma_{\hat{\Psi}(P_n)}$$

where ± 1.96 comes from the z-score covering the middle 95% area of the standard normal curve.

The results from implementing the four estimators can be found in Table 4:

Table 4: Estimator Results

Estimator	$\hat{\Psi}(P_n)$	**p-value**	**Influence Curve 95% CI**	**Bootstrap 95% CI**
Simple Substitution	-0.1599	0.0000	—	[-0.1633, -0.1563]
IPTW	-0.1927	0.0000	—	[-0.2174, -0.1663]
IPTW Stabilized	-0.2100	0.0000	—	[-0.2313, -0.1870]
TMLE	-0.1650	8.48e-6	[-0.2376, -0.0924]	[-0.1753, -0.1560]

For the non-parametric bootstrap, confidence intervals were created using the normality assumption as well as the quantile method. These were found to be nearly identical. The

results for the bootstrapped confidence interval reported in the table are from the normality assumption. Ultimately, all estimators yield a statistically significant average treatment effect. However, there are differences in the confidence intervals. The confidence intervals for the the IPTW estimator (even after stabilization) is larger than the confidence interval for the Simple Substitution estimator and for the TMLE estimator. This suggests that the using the IPTW is not an efficient estimator of the average treatment effect. The confidence interval using the Influence Curve is much more conservative than the confidence interval using the bootstrapping procedure. This may be due to either of $\bar{Q}_n^0(A, W)$ or $g_n(A_i|W_i)$ not being estimated consistently by the Super Learner, however, we do not know this for sure.

6.3 Sensitivity Analysis

After the estimation step, we were interested in the effect of the number of covariates conditioned on the target estimand $\hat{\Psi}(P_n)$. To answer this question, we performed a short sensitivity analysis by computing $\hat{\Psi}(P_n)$ with TMLE with an increasing number of covariates. The order in which we increased the number of covariates conditioned on is

- $W_1 = \{avg\_monthly\_hours, \#projects, work\_accident, time\_spend\_company\}$
- $W_2 = W_1 + \{promotion\}$
- $W_3 = W_2 + \{department\}$
- $W_4 = W_3 + \{satisfaction\}$
- $W_5 = W_4 + \{last\_evaluation\}$

We started with multiple variables in the first round because some statistical learning algorithms in Superlearner run into problems when the number of features is too low. The results are summarized in the graph below, where 95% confidence intervals are calculated as described above.

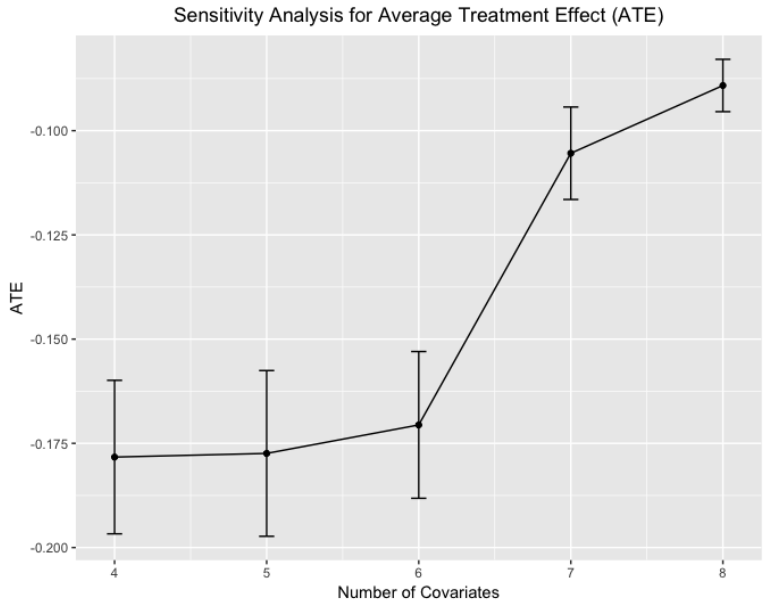


Figure 3: Sensitivity Analysis

In terms of the absolute value of estimated ATE, we observe the most dramatic reduction in ATE for salary when the last 2 covariates, $last\_evaluation$ and $satisfaction$, are conditioned on. In our causal DAG, these last 2 covariates are considered mediator variables of $salary$, and our sensitivity analysis provides further supports for this. Prior to conditioning on the last 2 covariates, the estimated ATE does not change appreciably with an increasing number of covariates conditioned on.

6.4 Variable Importance Measure

Given all the available variables in our dataset, a natural question to ask is what is the relative importance of some of the variables with respect to affecting the probability of an employee leaving a company. In this part of the project, we decided to use the coefficient β_1 in the marginal structural model (MSM)

$$\ln\left(\frac{P_{U,X}(Y_a)}{1-P_{U,X}(Y_a)}\right) = \beta_0 + \beta_1 a$$

as the measure of variable importance, because some variables are multi-level. From the use of logistic regression, the coefficient β_1 can be interpreted as the odds ratio for variable A. We will estimate Ψ for each variable by adjusting for all variables in W without considering the back-door criteria, thus the VIM only takes on a statistical interpretation. For this project we estimate the VIM for variables $salary$, $last\_evaluation$, $satisfaction$, and $avg\_monthly\_hours$. We will iterate through each of the 4 variables in turn as A, with the rest of the variables except for the outcome treated as W, in order to estimate each variable's VIM.

We implemented the MSM with stabilized weights specified as

$$st.wt = \frac{g_n^*(A)}{g_n(A|W)}, \text{ where } g_n^*(A) = \frac{1}{n}\sum_{i=1}^{n} I(A_i = a)$$

where $g_n(A|W)$ is estimated using multinomial regression. Logistic regression is performed with the `glm` package with weights specified by *st.wt*. The 95% confidence intervals are derived from variance estimates for β_1 in logistic regression. Results for estimating the VIMs are summarized in Table 5

Table 5: VIM Results

Variable	**VIM**	**p-value**	**95% CI**
Salary	0.1717	2.2381e-63	[0.1390, 0.2097]
Last evaluation	1.5149	3.9643e-103	[1.4590, 1.5734]
Satisfaction level	0.5264	1.4508e-230	[0.5063, 0.5471]
Average monthly hours	1.4956	7.6736e-74	[1.4324, 1.5623]

7 Interpret Results

The causal question we are interested in is the difference in the counter-factual probability of leaving a company with high salary versus probability of leaving with low salary. Under $\mathcal{M}^{F*}$ for identifiability, the covariates $W = (W_1, W_2)$ satisfy the back-door criteria. These covariates correspond to *#projects*, *avg_monthly_hours*, *time_spend_company*, *work_accident*, *promotion*, and *department*. Looking at the TMLE estimated ATE, a change in *salary* from low to high corresponds to a 0.1650 reduction in probability of leaving, with a 95% CI of [-0.2376, -0.0924], indicating significance of effect. If the assumptions of the study are considered plausible - and we have reason to believe that they are not - then our recommendation for this company is to pay their employees high salaries to reduce employee attrition.

The estimated ATE for *salary* is similar for the simple substitution estimator (-0.1599) but noticeably different for the IPTW estimator (-0.1927). However, all 3 estimators agree on the direction of the effect that *salary* has on the probability of leaving. A potential reason for the IPTW-estimated ATE to be different could be due to IPTW being a high variance estimator from near positivity violations.

In addition to answering the causal question of the effect of *salary* on probability of leaving, we also estimated the VIM for *salary*, *last_evaluation*, *satisfaction*, and *avg_monthly_hours*. This measure only has a statistical interpretation as we conditioned on all covariates without considering the back-door criteria. From the results, both an increase in *salary* and *satisfaction* is associated with a decrease in probability of leaving. On the other hand, an increase in *last_evaluation* and *avg_monthly_hours* is associated with increased probability of leaving. Because we did not analyze these factors from a causal framework, we do not recommend policy decisions based on these results.

A major limitation of our project is the absence of information on time frame of exposure for many variables. For example, neither the length of *salary* nor the length of company stay are provided. Under these limitations, we assume that *salary* is constant from when an employee enters a company to when he/she leaves. The time of stay in a company is treated as an unmeasured confounder. Another major limitation of our project is the necessary independence assumptions we need to make to satisfy the back-door criterion. As mentioned previously, the required independence assumptions do not reflect our knowledge of the complex relationship between salary and whether an employee stays or leaves the job. However, we still choose to make these convenience assumptions in order to attempt to answer our causal question.

A potential future direction for this project includes applying TMLE with continuous treatment A rather than binary treatment. There were originally 3 levels of *salary* (low, medium, high), but we restricted our samples to those with low and high salary levels as a limitation. Additionally, it is more realistic to consider *salary* as a continuous variable. Another potential future direction is to repeat the causal road map for other variables for which we estimated their VIMs. Furthermore, we could look at the causal effect within strata of the company, for example, we could consider the causal impact of *salary* within each individual *department* of the organization. Lastly, we can consider changing the design of the study such that it becomes a longitudinal study where we follow employees and measure variables at different time points to measure the total effect that *salary* has on whether the employee stays or leaves a company.

References

[1] M. L. Petersen, K. E. Porter, S. Gruber, Y. Wang, and M. J. V. D. Laan, "Diagnosing and responding to violations in the positivity assumption," *Statistical Methods in Medical Research*, vol. 21, no. 1, p. 3154, 2012.

[2] A. E. H. Mark J. van der Laan, Eric C. Polley, "Super learner," *Berkeley Division of Biostatistics Working Paper Series*, 2007.

[3] T. Hastie, R. Tibshirani, and J. H. Friedman, *The elements of statistical learning data mining, inference, and prediction.* Springer, 2016.

Individual Contributions

1. **Specify the causal question**: Dario Cantore wrote this section and Josiah Davis reviewed it.

2. **Specify the causal model**: Dario Cantore wrote this section and Josiah Davis reviewed it.

3. **Specify the causal parameter of interest**: Dario Cantore wrote this section and Josiah Davis reviewed it.

4. **Specify the observed data and its link to the causal model**: Daniel Lee wrote this section and Calvin Chi reviewed it.

5. **Identify**: Daniel Lee wrote Section *5.1 Assessment of the back-door criteria* and *5.3* $\mathcal{M}^{\mathcal{F}*}$ and Calvin Chi reviewed it. Josiah Davis wrote *5.2 Practical Positivity Assumption Evaluation* and Dario Cantore reviewed it.

6. **Estimate**: Josiah Davis wrote Section *6.1 Super Learner* and Dario Cantore reviewed it. Calvin Chi and Josiah Davis Section *6.2 Estimators of the ATE* and Dario Cantore reviewed it. Calvin wrote Section *6.3 Sensitivity Analysis* and Section *6.4 Variable Importance Measure* and Daniel Lee reviewed it.

7. **Interpret**: Calvin Chi wrote the initial draft for this section and all team members contributed to it.