Heterogeneous Treatment Effects

Prof. Mike Aguilar

https:/www.linkedin.com/in/mike-aguilar-econ

Outline

1. Motivation - Case Study
2. CATE
3. Meta Learners
   - S-Learner
   - T-Learner
   - X-Learner
4. Causal Random Forests

Heterogeneous Treatment Effects (HTE) - Summary

- Individual Treatment Effects (ITE) are useful, but we can't estimate, so...

Heterogeneous Treatment Effects (HTE) - Summary

- Individual Treatment Effects (ITE) are useful, but we can't estimate, so...
- ... we resort to averages such as the ATE, but....

Heterogeneous Treatment Effects (HTE) - Summary

- Individual Treatment Effects (ITE) are useful, but we can't estimate, so...
- ... we resort to averages such as the ATE, but....
- ... the ATE imposes constant marginal effects on entire domain

Heterogeneous Treatment Effects (HTE) - Summary

- Individual Treatment Effects (ITE) are useful, but we can't estimate, so...
- ... we resort to averages such as the ATE, but....
- ... the ATE imposes constant marginal effects on entire domain
- Our solution: Use CATE

Heterogeneous Treatment Effects (HTE) - Summary

- Individual Treatment Effects (ITE) are useful, but we can't estimate, so...
- ... we resort to averages such as the ATE, but....
- ... the ATE imposes constant marginal effects on entire domain
- Our solution: Use CATE (conditional average treatment effect) as a proxy of ITE based upon observable characteristics

Heterogeneous Treatment Effects (HTE) - Summary

- Individual Treatment Effects (ITE) are useful, but we can't estimate, so...
- ... we resort to averages such as the ATE, but....
- ... the ATE imposes constant marginal effects on entire domain
- Our solution: Use CATE (conditional average treatment effect) as a proxy of ITE based upon observable characteristics
- Non linear models (e.g. Causal Random Forests and Meta Learners) are options to estimate CATEs

Housekeeping

- Throughout the slide deck you will see "Q", which indicates a question to you, the reader
- You will also see "A", which indicates the associated answer
- It is generally most efficient to learn this material through active participation. Whenever you encounter a "Q", be sure to try and develop your answer before turning the page to the provided "A" answer
- I recommend reading through these slides before engaging in associated coding exercises
- This slide deck is part of a broader curriculum in causal estimation and inference
- Suggested pre-requisite: random forests, potential outcomes, CATE

Outline

1. Motivation - Case Study
2. CATE
3. Meta Learners
   - S-Learner
   - T-Learner
   - X-Learner
4. Causal Random Forests

Case Study - Retail Discounting



Case Study - Background

- A retail grocery chain is attempting to increase sales by optimizing its pricing strategy

Case Study - Background

- A retail grocery chain is attempting to increase sales by optimizing its pricing strategy
- The retailer hopes that the price discounts given to their "loyalty club" members will drive sales volume sufficiently to offset the decline in unit prices

Case Study - Background

- A retail grocery chain is attempting to increase sales by optimizing its pricing strategy
- The retailer hopes that the price discounts given to their "loyalty club" members will drive sales volume sufficiently to offset the decline in unit prices
- Stakeholders tell you that customer income drives sales as well as loyalty club membership

Case Study - Background

- A retail grocery chain is attempting to increase sales by optimizing its pricing strategy
- The retailer hopes that the price discounts given to their "loyalty club" members will drive sales volume sufficiently to offset the decline in unit prices
- Stakeholders tell you that customer income drives sales as well as loyalty club membership
- The Chief Strategy Officer (CSO) wants you to determine the impact of the loyalty club membership

Data

- The unit of observation (i) is at the customer-level

Data

- The unit of observation (i) is at the customer-level
- Data on over 50,000 customers is available

Data

- The unit of observation (i) is at the customer-level
- Data on over 50,000 customers is available
- Data is cross sectional such that sales during 2022 are aggregated for each (i) (e.g. Customer #201 purchased $\$3,000$ in groceries from this chain in 2022)

Data

- The unit of observation (i) is at the customer-level
- Data on over 50,000 customers is available
- Data is cross sectional such that sales during 2022 are aggregated for each (i) (e.g. Customer #201 purchased $\$3,000$ in groceries from this chain in 2022)
- Sales, measured in dollars of purchases by customer (i)

Data

- The unit of observation (i) is at the customer-level
- Data on over 50,000 customers is available
- Data is cross sectional such that sales during 2022 are aggregated for each (i) (e.g. Customer #201 purchased $\$3,000$ in groceries from this chain in 2022)
- Sales, measured in dollars of purchases by customer (i)
- Binary variable indicating if member of loyalty program. For simplicity we can assume that loyalty program membership started at the beginning of 2022 for all customers

Data

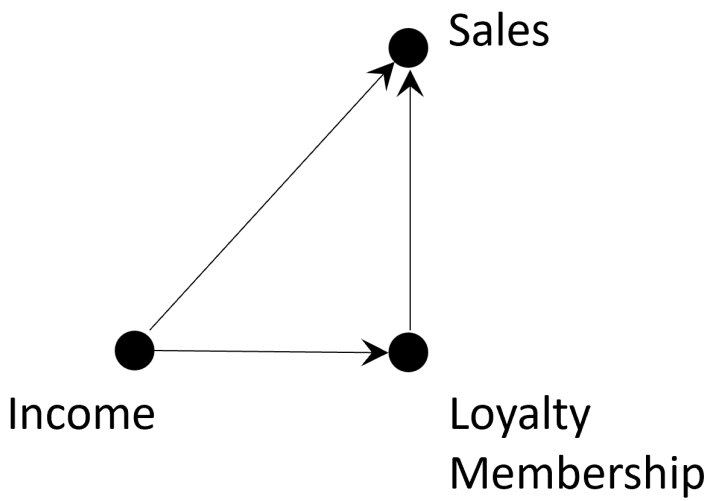
- The unit of observation (i) is at the customer-level
- Data on over 50,000 customers is available
- Data is cross sectional such that sales during 2022 are aggregated for each (i) (e.g. Customer #201 purchased $\$3,000$ in groceries from this chain in 2022)
- Sales, measured in dollars of purchases by customer (i)
- Binary variable indicating if member of loyalty program. For simplicity we can assume that loyalty program membership started at the beginning of 2022 for all customers
- A large set of customer specific features is available, including customer income

Mapping the Causal Path

Q: Can you draw a DAG and define the treatment effect?

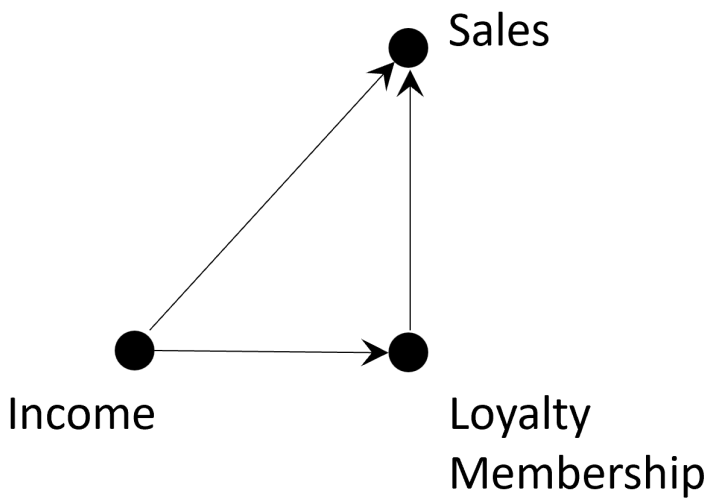
Mapping the Causal Path

Q: Can you draw a DAG and define the treatment effect?



Mapping the Causal Path

Q: Can you draw a DAG and define the treatment effect?

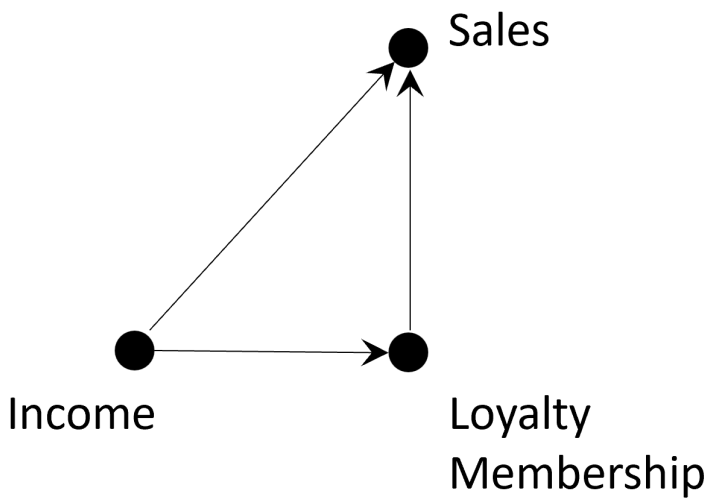


A:

- Outcome:

Mapping the Causal Path

Q: Can you draw a DAG and define the treatment effect?

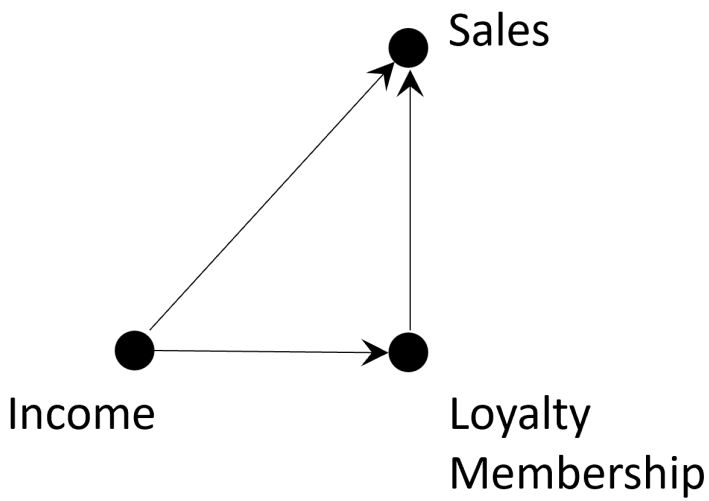


A:

- Outcome: Sales

Mapping the Causal Path

Q: Can you draw a DAG and define the treatment effect?

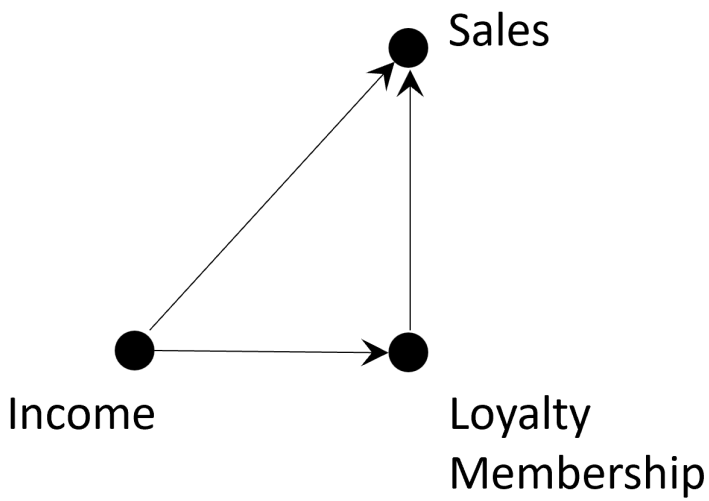


A:

- Outcome: Sales
- Treatment:

Mapping the Causal Path

Q: Can you draw a DAG and define the treatment effect?

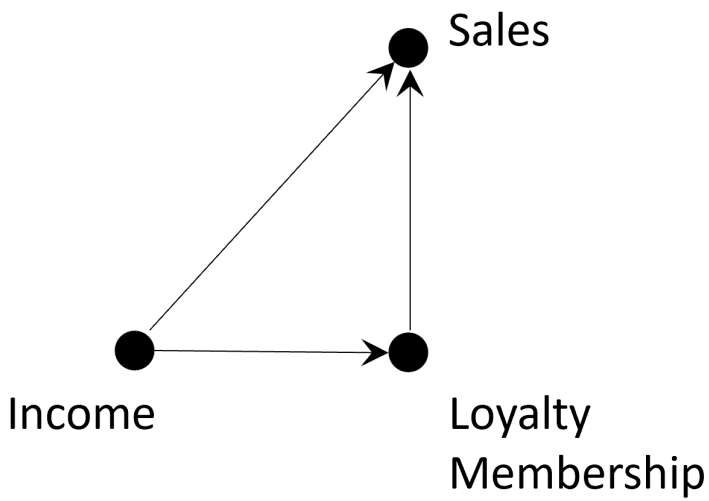


A:

- Outcome: Sales
- Treatment: Loyalty club membership

Mapping the Causal Path

Q: Can you draw a DAG and define the treatment effect?

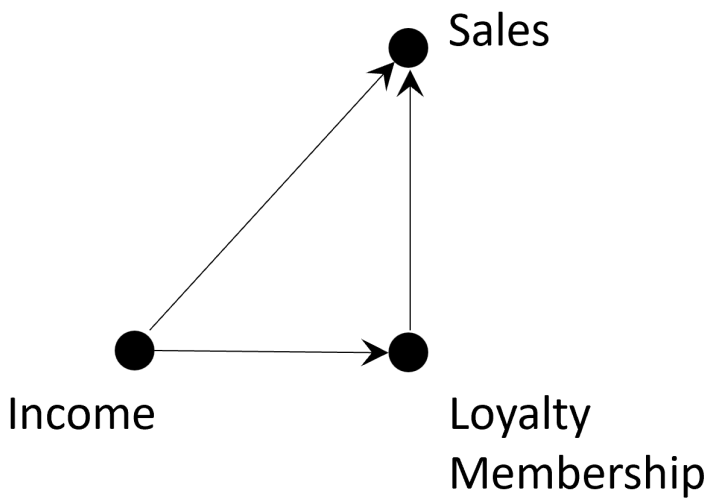


A:

- Outcome: Sales
- Treatment: Loyalty club membership
- Confounder:

Mapping the Causal Path

Q: Can you draw a DAG and define the treatment effect?

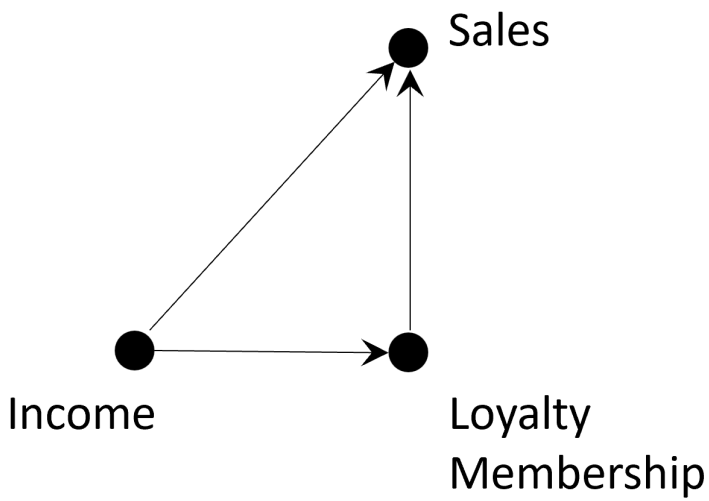


A:

- Outcome: Sales
- Treatment: Loyalty club membership
- Confounder: Income

Mapping the Causal Path

Q: Can you draw a DAG and define the treatment effect?

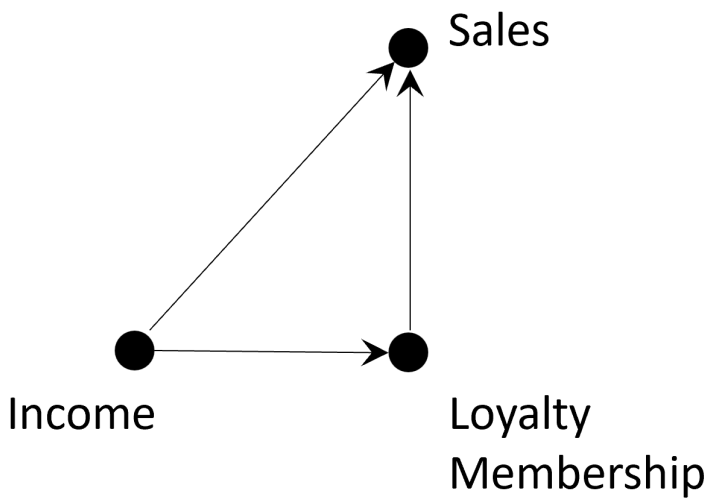


A:

- Outcome: Sales
- Treatment: Loyalty club membership
- Confounder: Income
- Treatment Effect: impact of loyalty club upon sales

Mapping the Causal Path

Q: Can you draw a DAG and define the treatment effect?



A:

- Outcome: Sales
- Treatment: Loyalty club membership
- Confounder: Income
- Treatment Effect: impact of loyalty club upon sales

As usual, there certainly other elements to a complete DAG, but we focus on this simplified structure in order to highlight the important elements of our new tool

The Impossible Ask

- Suppose you estimated the Treatment Effect via typical methods (RA, Matching, etc..).

The Impossible Ask

- Suppose you estimated the Treatment Effect via typical methods (RA, Matching, etc..).
- You report the ATT, ATC, ATE to the CSO.

The Impossible Ask

- Suppose you estimated the Treatment Effect via typical methods (RA, Matching, etc..).
- You report the ATT, ATC, ATE to the CSO.
- The CSO is not satisfied. He/she doesn't want an "average". He/she wants a customized impact measure of the loyalty membership for each and every customer.

The Impossible Ask

- Suppose you estimated the Treatment Effect via typical methods (RA, Matching, etc..).
- You report the ATT, ATC, ATE to the CSO.
- The CSO is not satisfied. He/she doesn't want an "average". He/she wants a customized impact measure of the loyalty membership for each and every customer.
- Q: Why is this an impossible request?

The Impossible Ask

- Suppose you estimated the Treatment Effect via typical methods (RA, Matching, etc..).
- You report the ATT, ATC, ATE to the CSO.
- The CSO is not satisfied. He/she doesn't want an "average". He/she wants a customized impact measure of the loyalty membership for each and every customer.
- Q: Why is this an impossible request?
- A: We can't estimate the impact for a single customer.

The Impossible Ask

- Suppose you estimated the Treatment Effect via typical methods (RA, Matching, etc..).
- You report the ATT, ATC, ATE to the CSO.
- The CSO is not satisfied. He/she doesn't want an "average". He/she wants a customized impact measure of the loyalty membership for each and every customer.
- Q: Why is this an impossible request?
- A: We can't estimate the impact for a single customer. We report the impact averaged across many individuals.

Interviewing the CSO

Asking smart questions

Interviewing the CSO

Asking smart questions

- Analyst: "I can't give you the impact of the loyalty membership for each customer. But I can tell you the impact for groups of customers. Which types of customers do you suspect are impacted the most by the loyalty club?"

Interviewing the CSO

Asking smart questions

- Analyst: "I can't give you the impact of the loyalty membership for each customer. But I can tell you the impact for groups of customers. Which types of customers do you suspect are impacted the most by the loyalty club?"
- Q: What measure of causality are we trying to elicit from these questions?

Interviewing the CSO

Asking smart questions

- Analyst: "I can't give you the impact of the loyalty membership for each customer. But I can tell you the impact for groups of customers. Which types of customers do you suspect are impacted the most by the loyalty club?"
- Q: What measure of causality are we trying to elicit from these questions?
- A: Conditional Average Treatment Effect (CATE)

Interviewing the CSO

Asking smart questions

- Analyst: "I can't give you the impact of the loyalty membership for each customer. But I can tell you the impact for groups of customers. Which types of customers do you suspect are impacted the most by the loyalty club?"
- Q: What measure of causality are we trying to elicit from these questions?
- A: Conditional Average Treatment Effect (CATE)
- CSO: "My suspicion is that customer income is again the key driver. Income drives sales. Income drives loyalty membership. Income drives the impact of the loyalty membership on sales."

Interviewing the CSO

Asking smart questions

- Analyst: "I can't give you the impact of the loyalty membership for each customer. But I can tell you the impact for groups of customers. Which types of customers do you suspect are impacted the most by the loyalty club?"
- Q: What measure of causality are we trying to elicit from these questions?
- A: Conditional Average Treatment Effect (CATE)
- CSO: "My suspicion is that customer income is again the key driver. Income drives sales. Income drives loyalty membership. Income drives the impact of the loyalty membership on sales."
- Note that the CSO confirmed the believe that income is a confounder (i.e.

Interviewing the CSO

Asking smart questions

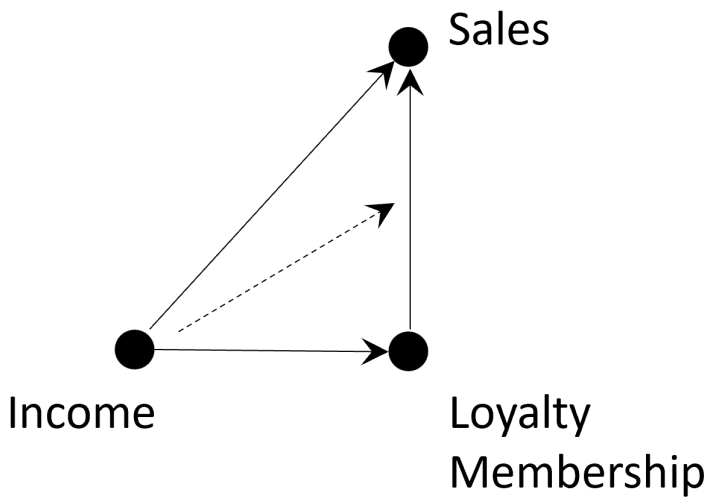
- Analyst: "I can't give you the impact of the loyalty membership for each customer. But I can tell you the impact for groups of customers. Which types of customers do you suspect are impacted the most by the loyalty club?"
- Q: What measure of causality are we trying to elicit from these questions?
- A: Conditional Average Treatment Effect (CATE)
- CSO: "My suspicion is that customer income is again the key driver. Income drives sales. Income drives loyalty membership. Income drives the impact of the loyalty membership on sales."
- Note that the CSO confirmed the believe that income is a confounder (i.e. income drives loyalty club membership (treatment) and sales (outcome).

Interviewing the CSO

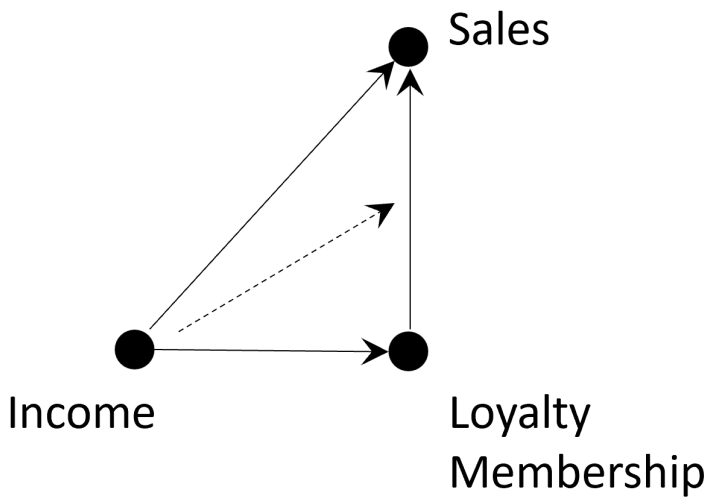
Asking smart questions

- Analyst: "I can't give you the impact of the loyalty membership for each customer. But I can tell you the impact for groups of customers. Which types of customers do you suspect are impacted the most by the loyalty club?"
- Q: What measure of causality are we trying to elicit from these questions?
- A: Conditional Average Treatment Effect (CATE)
- CSO: "My suspicion is that customer income is again the key driver. Income drives sales. Income drives loyalty membership. Income drives the impact of the loyalty membership on sales."
- Note that the CSO confirmed the believe that income is a confounder (i.e. income drives loyalty club membership (treatment) and sales (outcome).
- But the CSO hinted at a dynamic between income the treatment effect itself.

Moderating the Treatment Effect

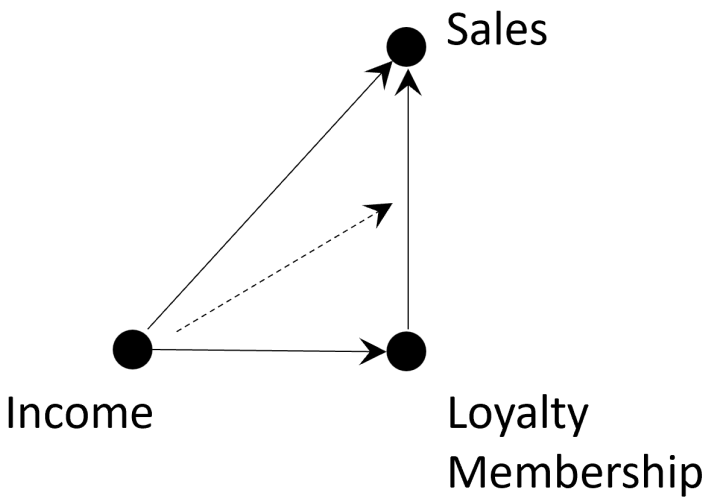


Moderating the Treatment Effect



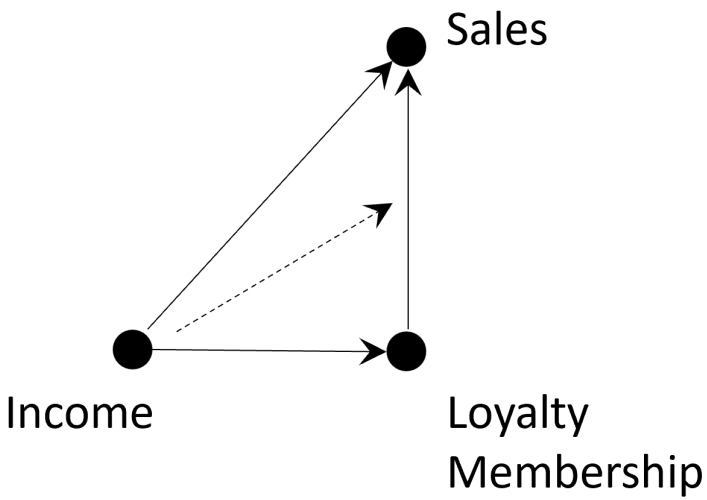
- The CSO is telling us that income is a Confounder &

Moderating the Treatment Effect



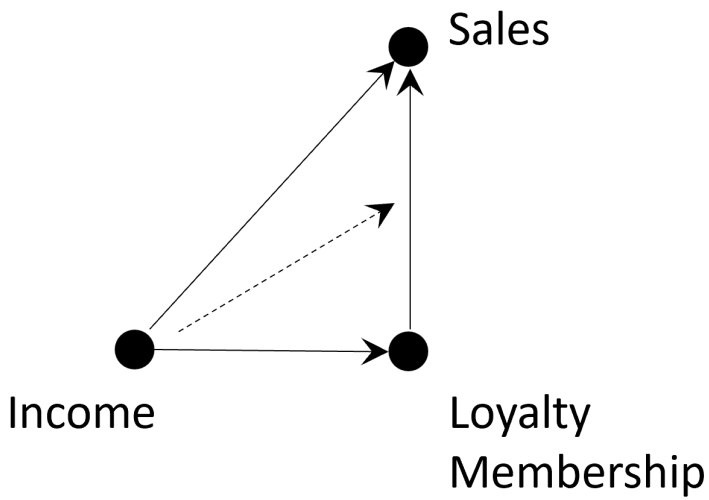
- The CSO is telling us that income is a Confounder & Moderator

Moderating the Treatment Effect



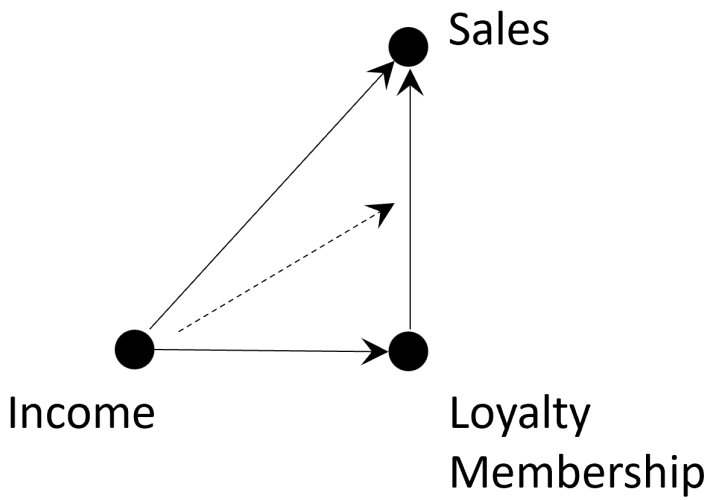
- The CSO is telling us that income is a Confounder & Moderator
- The Dashed Line is used to represent a moderating pathway

Moderating the Treatment Effect



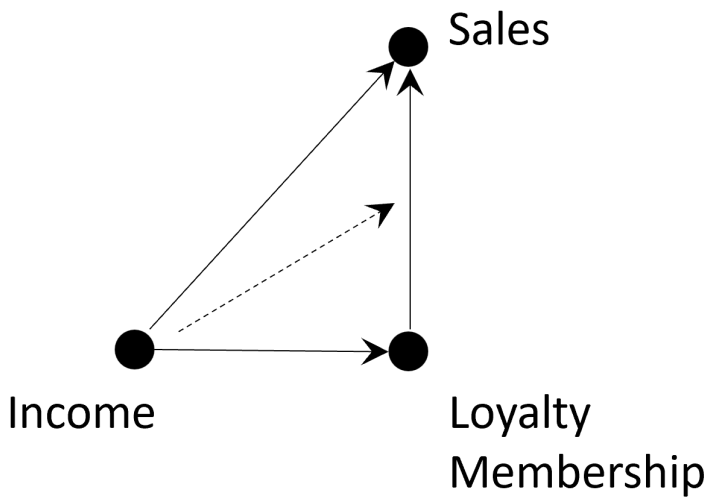
- The CSO is telling us that income is a Confounder & Moderator
- The Dashed Line is used to represent a moderating pathway
- Moderators influence the size/sign of the treatment effect

Moderating the Treatment Effect



- The CSO is telling us that income is a Confounder & Moderator
- The Dashed Line is used to represent a moderating pathway
- Moderators influence the size/sign of the treatment effect
- Moderators need not always be confounders

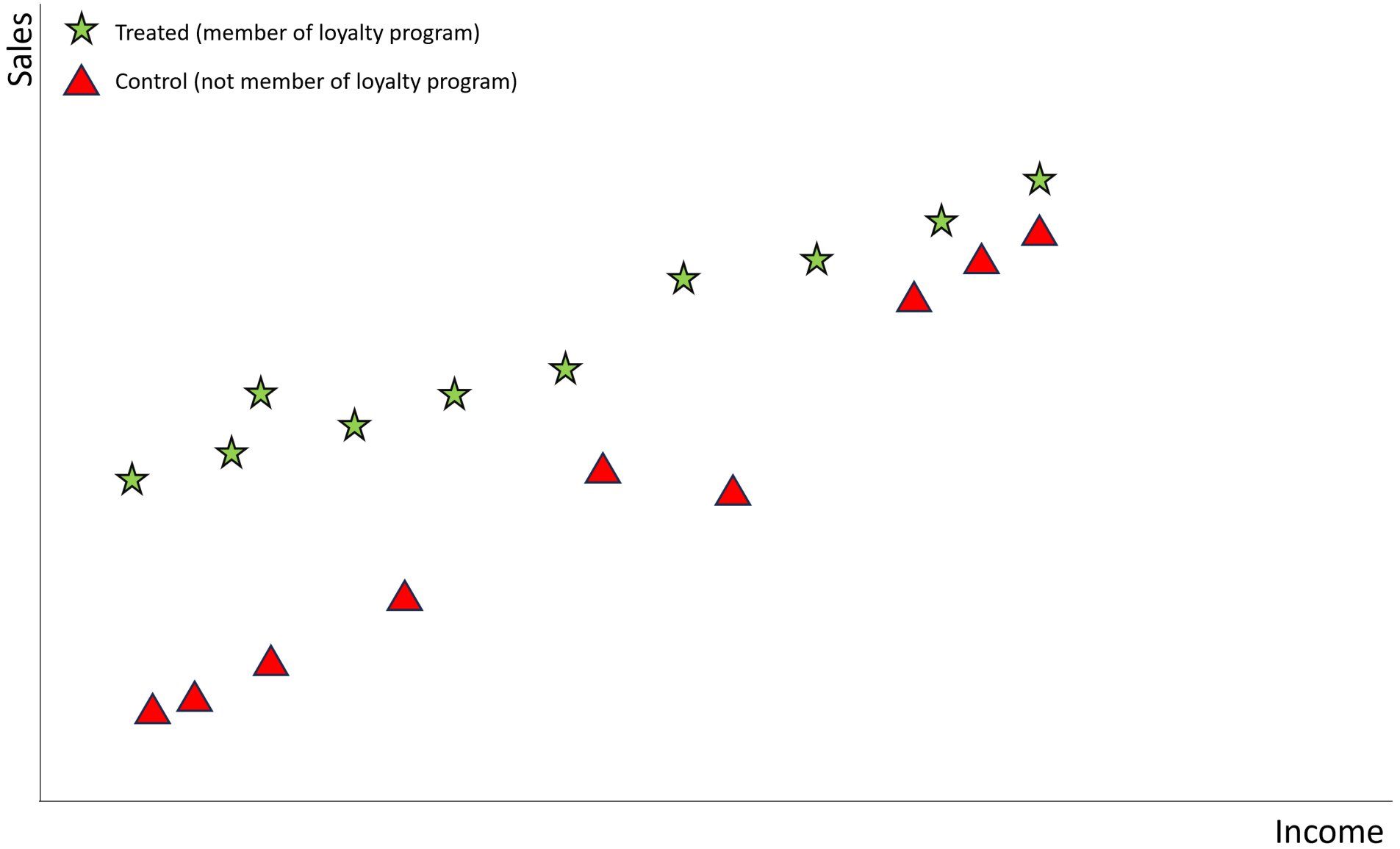
Moderating the Treatment Effect



- The CSO is telling us that income is a Confounder & Moderator
- The Dashed Line is used to represent a moderating pathway
- Moderators influence the size/sign of the treatment effect
- Moderators need not always be confounders

Note: There is controversy over if/how to include moderators in DAGs. See Weinberg 2007

Peaking at one dimension of our data



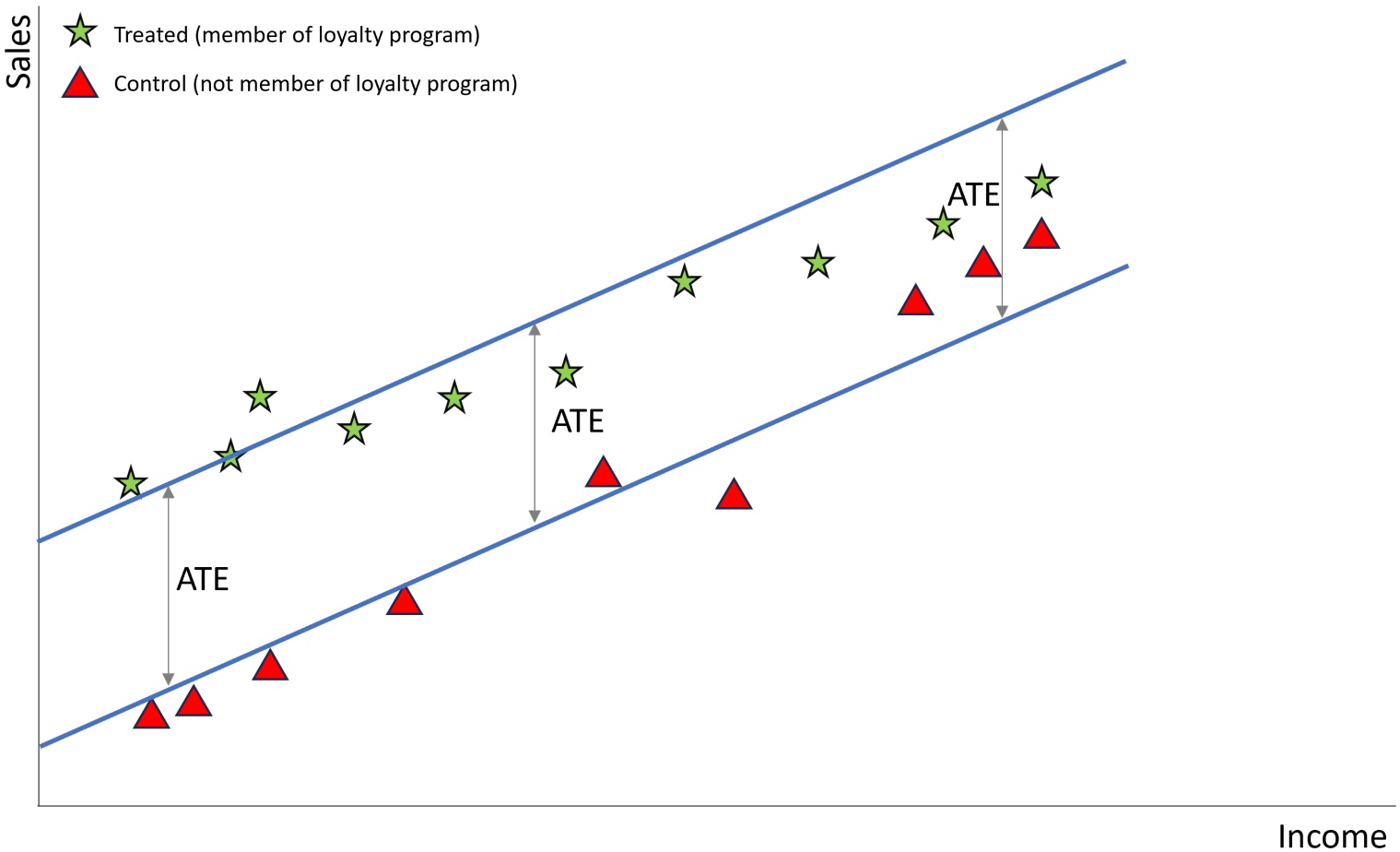
Simple RA model: $Sales = \alpha + \beta Loyalty + \gamma Income + u$

Q: How do we visualize the estimated ATE?

Simple RA model: $Sales = \alpha + \beta Loyalty + \gamma Income + u$

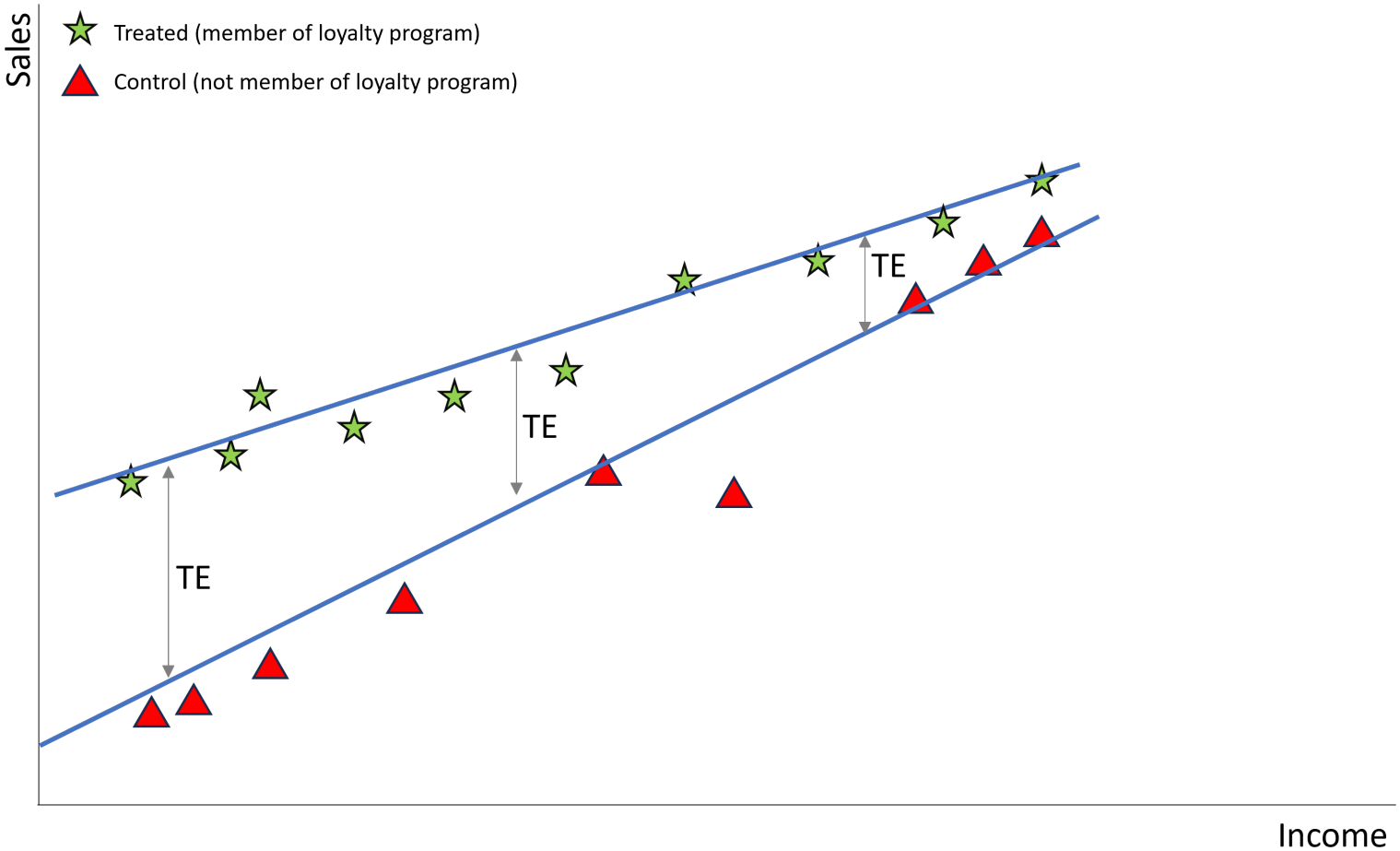
Q: How do we visualize the estimated ATE?

A:



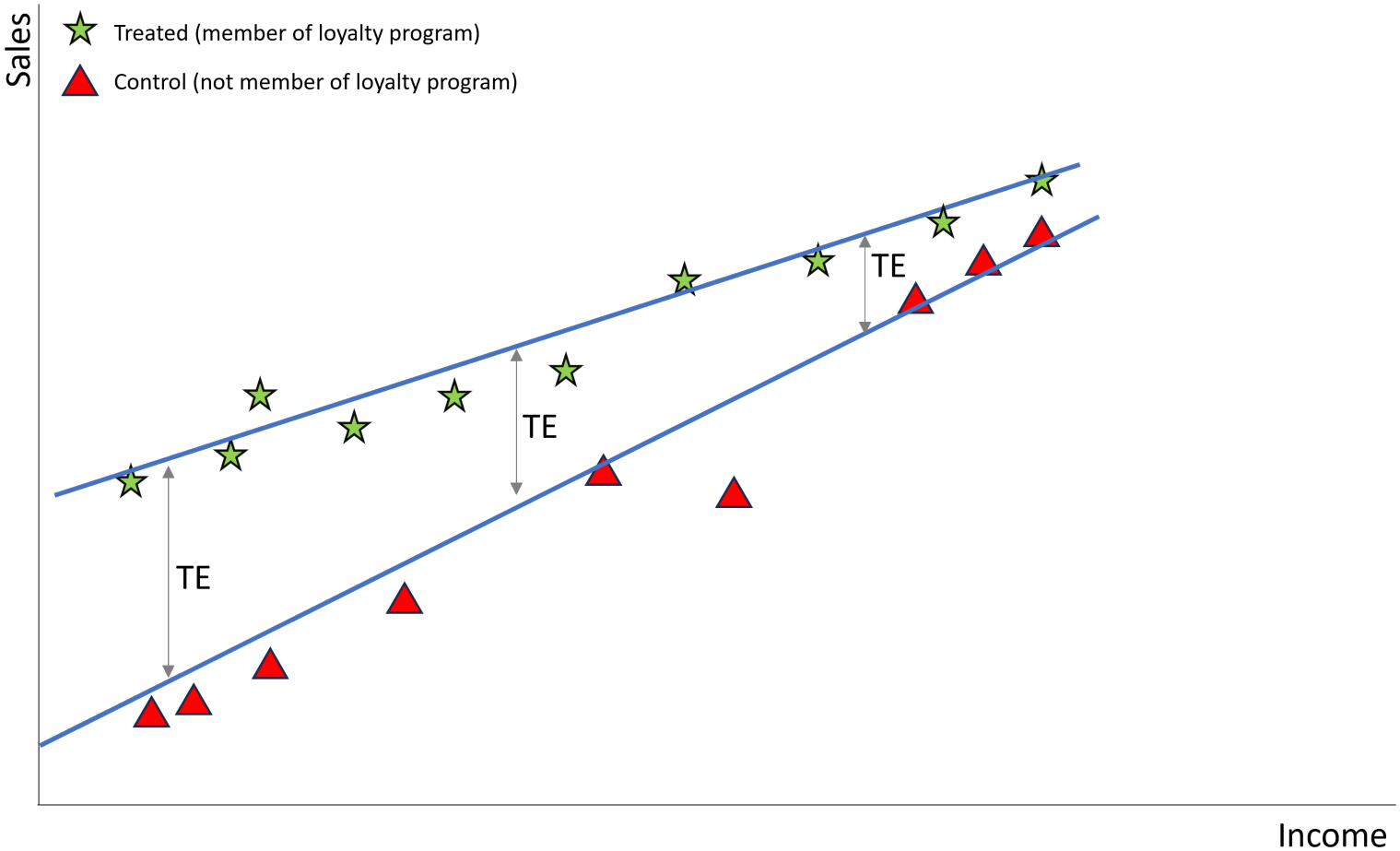
Suppose DGP is not $Sales = \alpha + \beta Loyalty + \gamma Income + u$

Q: Looking at the data, why doesn't ATE seem sufficient?



Suppose DGP is not $Sales = \alpha + \beta Loyalty + \gamma Income + u$

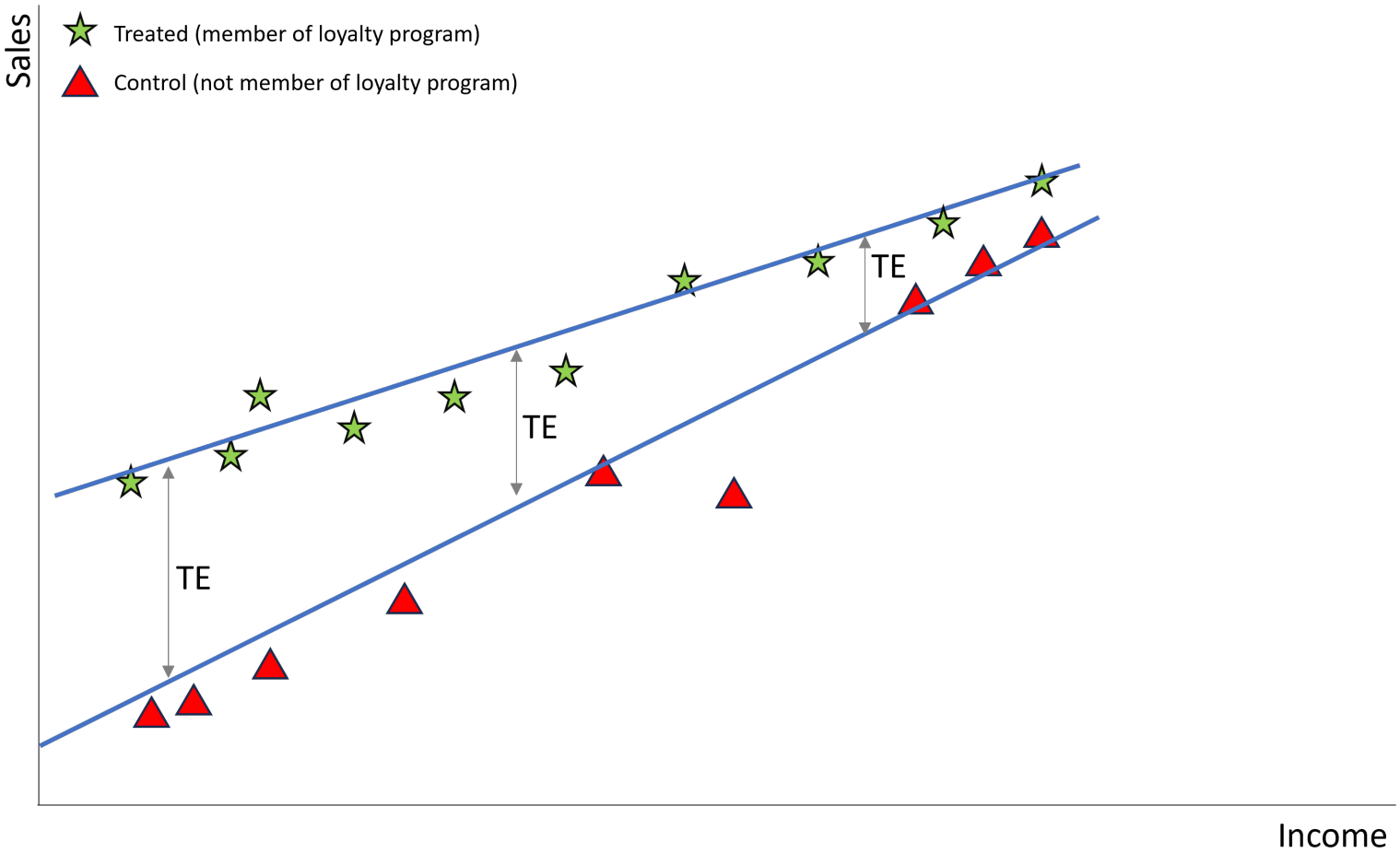
Q: Looking at the data, why doesn't ATE seem sufficient?



A: The TE seems to vary with income (i.e. its heterogeneous). But our simple linear model didn't permit TE to vary.

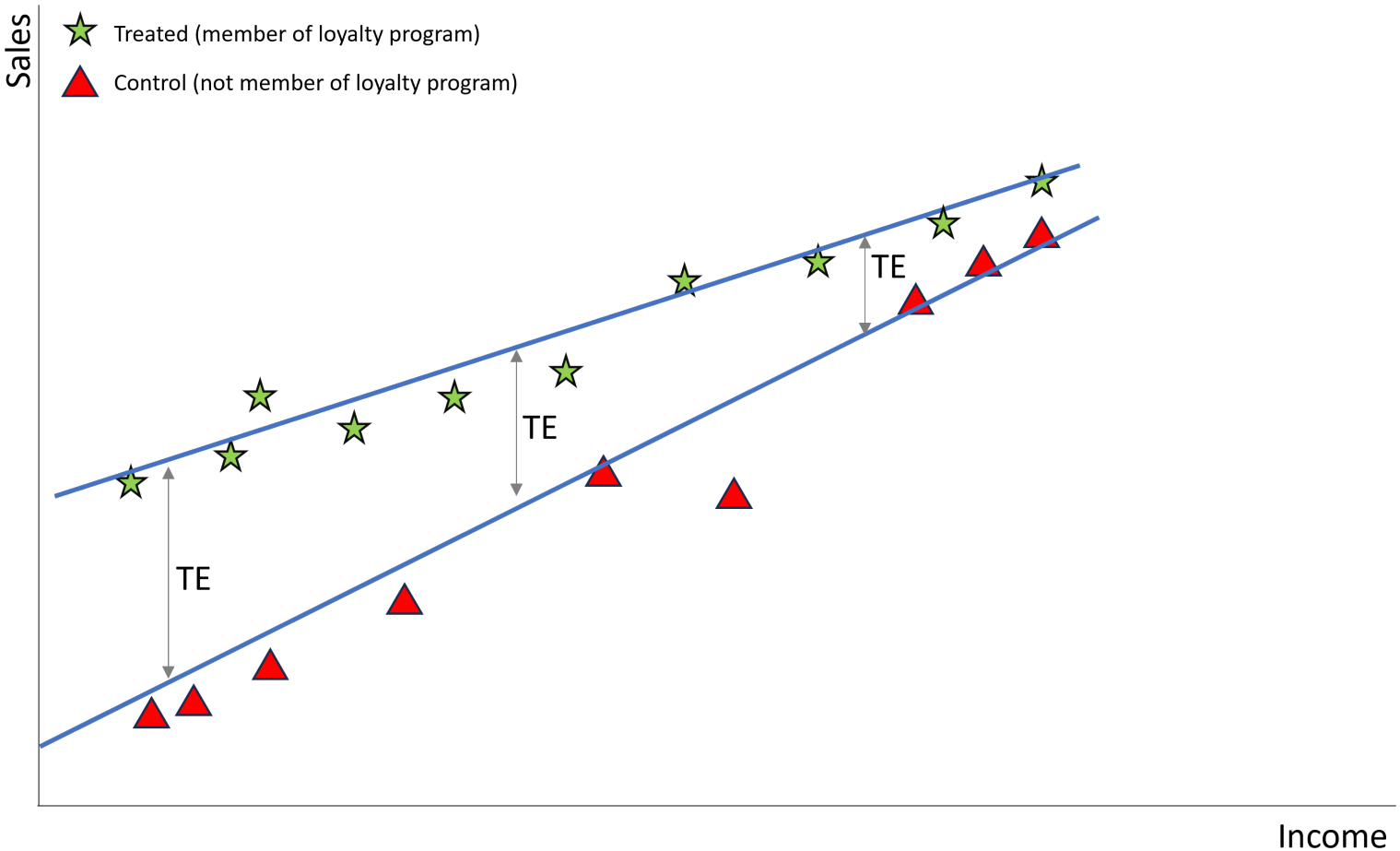
Heterogeneous Treatment Effects (HTE)

Q: How augment RA to accommodate HTE, instead of ATE?



Heterogeneous Treatment Effects (HTE)

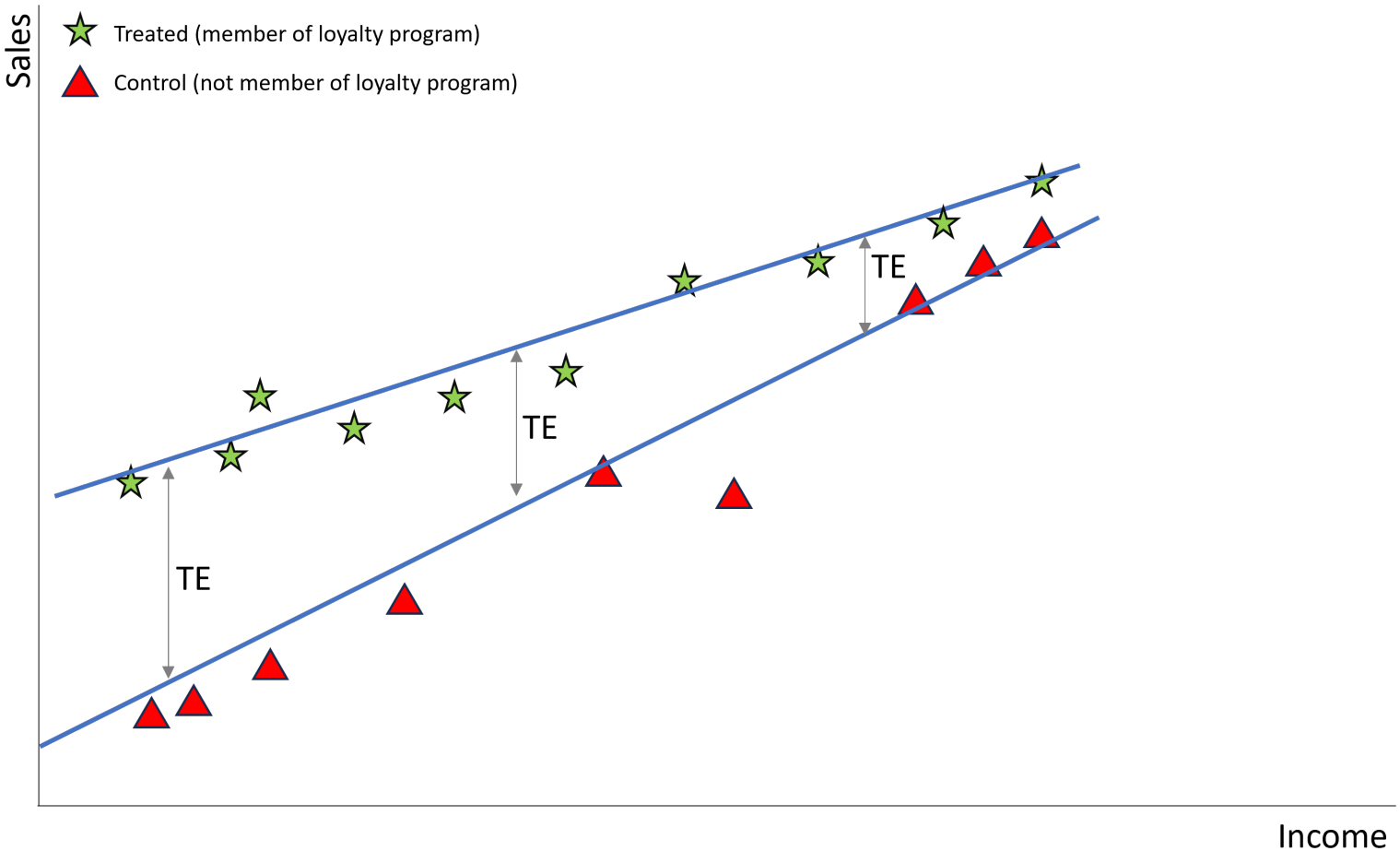
Q: How augment RA to accommodate HTE, instead of ATE?



A: $Sales = \alpha + \beta Loyalty + \gamma Income + \delta Loyalty \times Income + u.$

Heterogeneous Treatment Effects (HTE)

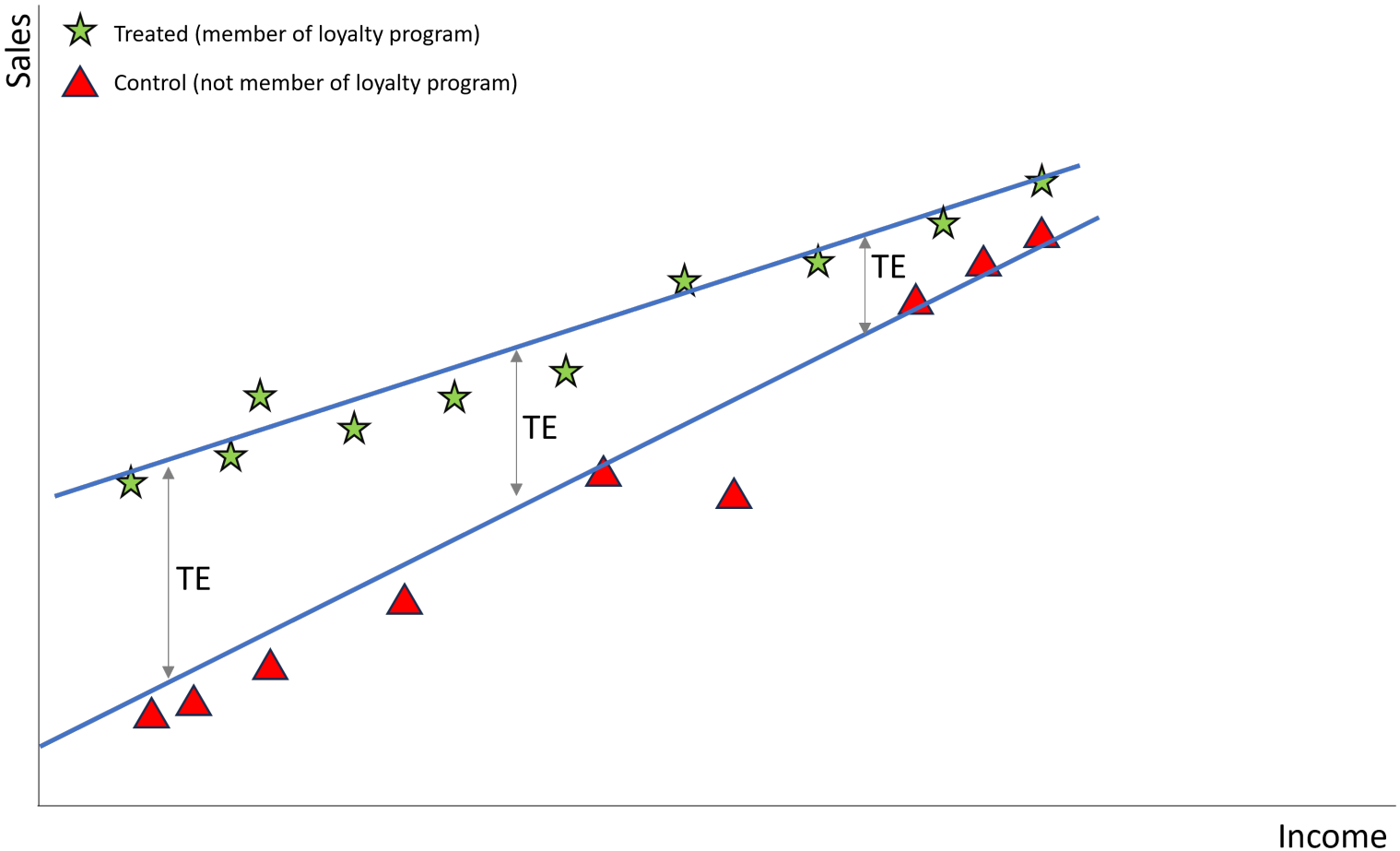
Q: How augment RA to accommodate HTE, instead of ATE?



A: $Sales = \alpha + \beta Loyalty + \gamma Income + \delta Loyalty \times Income + u$. Notice that the estimated treatment effect is a function of income

Heterogeneous Treatment Effects (HTE)

Q: How augment RA to accommodate HTE, instead of ATE?



A: $Sales = \alpha + \beta Loyalty + \gamma Income + \delta Loyalty \times Income + u$. Notice that the estimated treatment effect is a function of income

$E[Sales|Loyalty = 1] - E[Sales|Loyalty = 0] = \beta + \delta \times Income$

Business Context

Q: Why would the chief strategy officer (CSO) care about HTE?

Business Context

Q: Why would the chief strategy officer (CSO) care about HTE?

A:

- Under a constant marginal effects model, we are telling the CSO that we expect all customers to react to the loyalty program in the same way.

Business Context

Q: Why would the chief strategy officer (CSO) care about HTE?

A:

- Under a constant marginal effects model, we are telling the CSO that we expect all customers to react to the loyalty program in the same way.
- The implication is that we could/should target all customers equally. Targeting customers requires money and effort. Moreover, offering the loyalty discount "costs" us the decline in unit price.

Business Context

Q: Why would the chief strategy officer (CSO) care about HTE?

A:

- Under a constant marginal effects model, we are telling the CSO that we expect all customers to react to the loyalty program in the same way.
- The implication is that we could/should target all customers equally. Targeting customers requires money and effort. Moreover, offering the loyalty discount "costs" us the decline in unit price.
- In our case, it appears that the treatment effect is heterogeneous. Specifically, higher income customers aren't impacted by the loyalty program as much as lower income customers.

Business Context

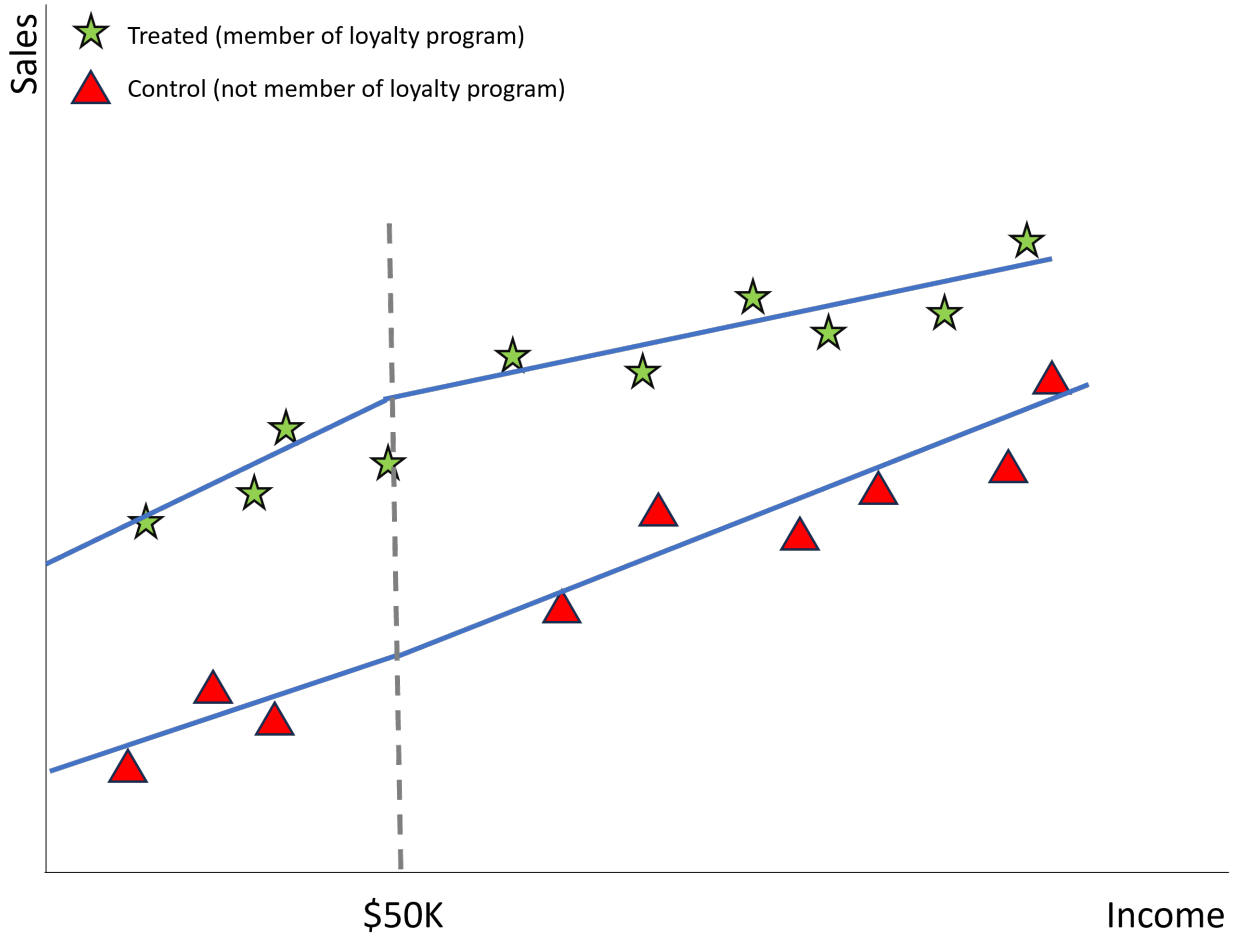
Q: Why would the chief strategy officer (CSO) care about HTE?

A:

- Under a constant marginal effects model, we are telling the CSO that we expect all customers to react to the loyalty program in the same way.
- The implication is that we could/should target all customers equally. Targeting customers requires money and effort. Moreover, offering the loyalty discount "costs" us the decline in unit price.
- In our case, it appears that the treatment effect is heterogeneous. Specifically, higher income customers aren't impacted by the loyalty program as much as lower income customers.
- We can save money by targeting our loyalty program to those customers with the highest expected impact (lower income).

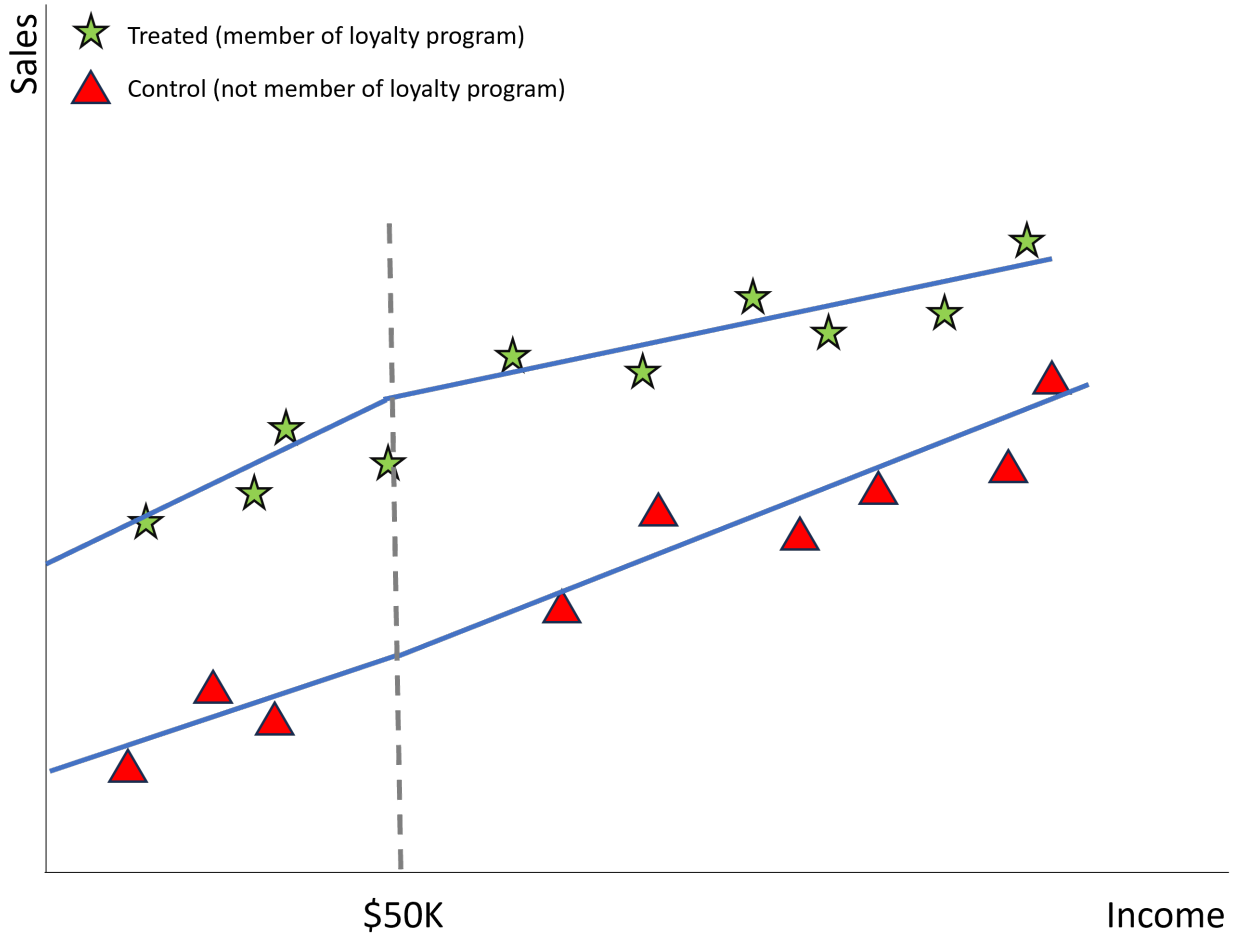
Threshold Effects

Suppose the CSO believes that customers with income greater than $\$50K$ respond differently. Q: Why not run RA for those $>$ $\$50K$ and again for those $<$ $\$50K$?



Threshold Effects

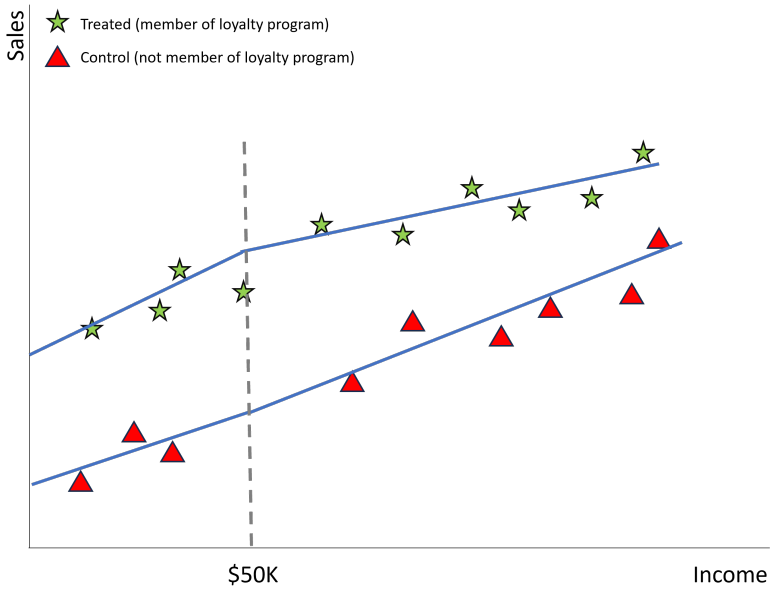
Suppose the CSO believes that customers with income greater than $\$50K$ respond differently. Q: Why not run RA for those $>$ $\$50K$ and again for those $<$ $\$50K$?



A: Can't compare them statistically.

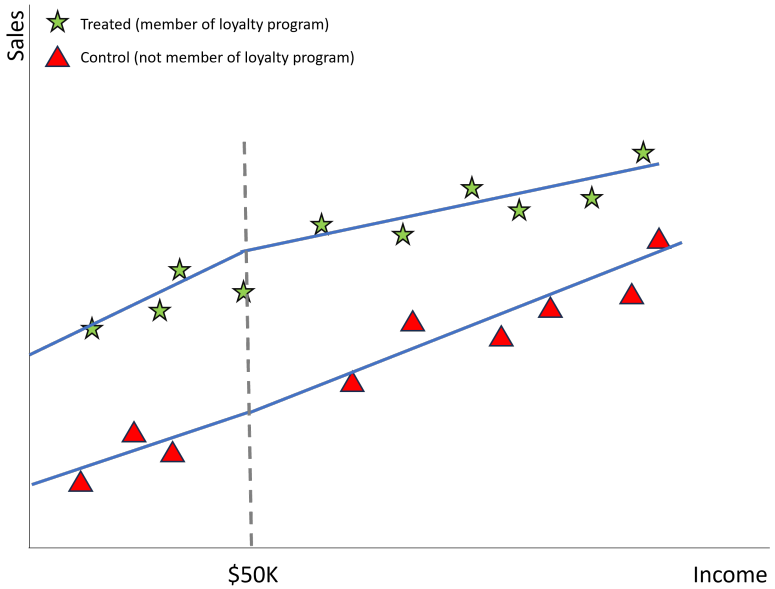
Income Thresholds

Q: How augment RA to accommodate an income threshold (e.g. $\$50K$ in HTE)?



Income Thresholds

Q: How augment RA to accommodate an income threshold (e.g. $\$50K$ in HTE)?

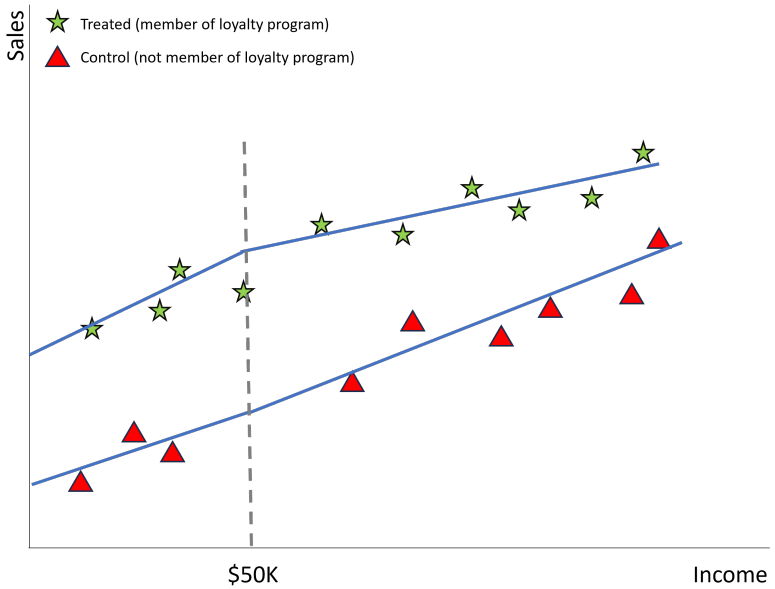


A:

$$Sales = \alpha + \beta Loyalty + \gamma Income + \delta Loyalty \times Income +$$
$$\pi_0 Threshold + \pi_1 Threshold \times Loyalty + \pi_2 Threshold \times Income$$
$$+ \pi_3 Threshold \times Loyalty \times Income + u$$

Income Thresholds

Q: How augment RA to accommodate an income threshold (e.g. $\$50K$ in HTE)?



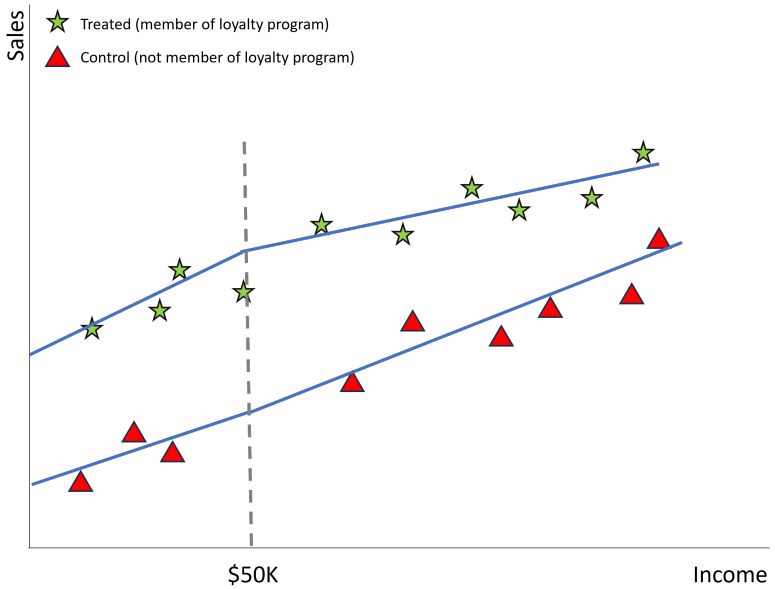
A:

$$Sales = \alpha + \beta Loyalty + \gamma Income + \delta Loyalty \times Income +$$
$$\pi_0 Threshold + \pi_1 Threshold \times Loyalty + \pi_2 Threshold \times Income$$
$$+ \pi_3 Threshold \times Loyalty \times Income + u$$

Then test

Income Thresholds

Q: How augment RA to accommodate an income threshold (e.g. $\$50K$ in HTE)?



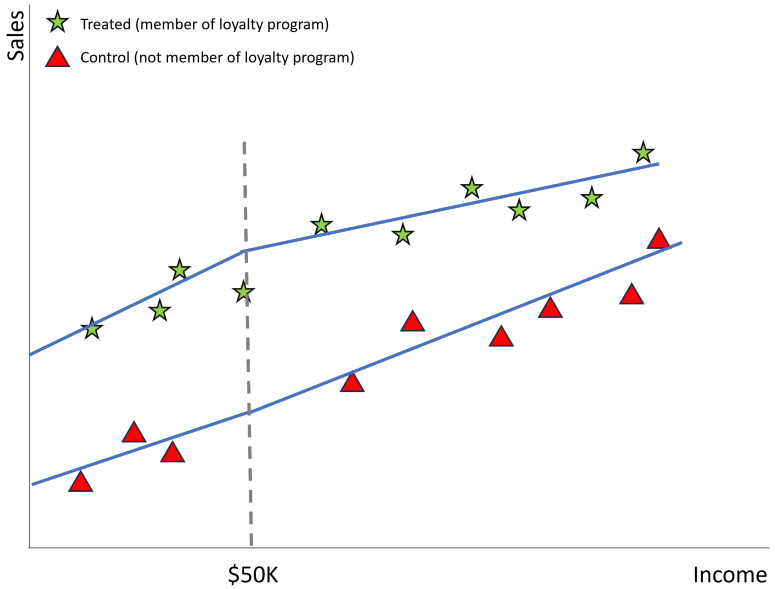
A:

$$Sales = \alpha + \beta Loyalty + \gamma Income + \delta Loyalty \times Income +$$
$$\pi_0 Threshold + \pi_1 Threshold \times Loyalty + \pi_2 Threshold \times Income$$
$$+ \pi_3 Threshold \times Loyalty \times Income + u$$

Then test $\pi_3 = 0$. If reject, then

Income Thresholds

Q: How augment RA to accommodate an income threshold (e.g. $\$50K$ in HTE)?



A:

$$Sales = \alpha + \beta Loyalty + \gamma Income + \delta Loyalty \times Income +$$
$$\pi_0 Threshold + \pi_1 Threshold \times Loyalty + \pi_2 Threshold \times Income$$
$$+ \pi_3 Threshold \times Loyalty \times Income + u$$

Then test $\pi_3 = 0$. If reject, then HTE.

Risks of increased complexity

Q: What are some weaknesses of these RA approaches?

Risks of increased complexity

Q: What are some weaknesses of these RA approaches?

A:

- Linear regression is a global model, hence the ATE is constant over the entire domain of income

Risks of increased complexity

Q: What are some weaknesses of these RA approaches?

A:

- Linear regression is a global model, hence the ATE is constant over the entire domain of income
- We can attempt to accommodate treatment effect heterogeneity via interactions, but...

Risks of increased complexity

Q: What are some weaknesses of these RA approaches?

A:

- Linear regression is a global model, hence the ATE is constant over the entire domain of income
- We can attempt to accommodate treatment effect heterogeneity via interactions, but...
  - since the model is parametric we must specify the "thresholds" in advance; i.e. the model doesn't "discover" the thresholds

Risks of increased complexity

Q: What are some weaknesses of these RA approaches?

A:

- Linear regression is a global model, hence the ATE is constant over the entire domain of income
- We can attempt to accommodate treatment effect heterogeneity via interactions, but...
  - since the model is parametric we must specify the "thresholds" in advance; i.e. the model doesn't "discover" the thresholds
  - linear regression is sensitive to the reduced degrees of freedom that accompanies many interactions

Risks of increased complexity

Q: What are some weaknesses of these RA approaches?

A:

- Linear regression is a global model, hence the ATE is constant over the entire domain of income
- We can attempt to accommodate treatment effect heterogeneity via interactions, but...
  - since the model is parametric we must specify the "thresholds" in advance; i.e. the model doesn't "discover" the thresholds
  - linear regression is sensitive to the reduced degrees of freedom that accompanies many interactions
  - multiple interactions are difficult to interpret

Risks of increased complexity

Q: What are some weaknesses of these RA approaches?

A:

- Linear regression is a global model, hence the ATE is constant over the entire domain of income
- We can attempt to accommodate treatment effect heterogeneity via interactions, but...
  - since the model is parametric we must specify the "thresholds" in advance; i.e. the model doesn't "discover" the thresholds
  - linear regression is sensitive to the reduced degrees of freedom that accompanies many interactions
  - multiple interactions are difficult to interpret
- Note: The same basic critiques hold for DiM, Matching, etc..

Outline

1. Motivation - Case Study
2. CATE
3. Meta Learners
   - S-Learner
   - T-Learner
   - X-Learner
4. Causal Random Forests

Reminder on ATE

Our typical object of interest is

$$ATE = E[Y_{1i} - Y_{0i}]$$

which we may estimate via a sample counterpart

$$\widehat{ATE} = (\frac{1}{N_1}\sum_{i=1}^{N_1} Y_i|D_i = 1) - (\frac{1}{N_0}\sum_{i=1}^{N_0} Y_i|D_i = 0)$$

Individual Treatment Effects

Q: What's the difference between ITE and ATE?

Individual Treatment Effects

Q: What's the difference between ITE and ATE?

A: The Individual Treatment Effect (ITE) is

$$ITE_i \equiv Y_{1i} - Y_{0i}$$

Individual Treatment Effects

Q: What's the difference between ITE and ATE?

A: The Individual Treatment Effect (ITE) is

$$ITE_i \equiv Y_{1i} - Y_{0i}$$

Notice that ITE is indexed by i, emphasizing its variation by individual.

Individual Treatment Effects

Q: What's the difference between ITE and ATE?

A: The Individual Treatment Effect (ITE) is

$$ITE_i \equiv Y_{1i} - Y_{0i}$$

Notice that ITE is indexed by i, emphasizing its variation by individual.

Q: Why have we resorted to averages instead of individualized treatment effects?

Individual Treatment Effects

Q: What's the difference between ITE and ATE?

A: The Individual Treatment Effect (ITE) is

$$ITE_i \equiv Y_{1i} - Y_{0i}$$

Notice that ITE is indexed by i, emphasizing its variation by individual.

Q: Why have we resorted to averages instead of individualized treatment effects?

A:

- ITE_i is a difficult (impossible?) object to estimate since we can't observe both potential outcomes simultaneously. Instead, we resorted to sub populations of treated and control groups.

Individual Treatment Effects

Q: What's the difference between ITE and ATE?

A: The Individual Treatment Effect (ITE) is

$$ITE_i \equiv Y_{1i} - Y_{0i}$$

Notice that ITE is indexed by i, emphasizing its variation by individual.

Q: Why have we resorted to averages instead of individualized treatment effects?

A:

- ITE_i is a difficult (impossible?) object to estimate since we can't observe both potential outcomes simultaneously. Instead, we resorted to sub populations of treated and control groups.
- The CATE is a way to proxy the ITE.

CATE

Defining the Conditional Average Treatment Effect (CATE)

$$CATE(x) \equiv E[Y_{1i}|X_i = x] - E[Y_{0i}|X_i = x]$$
$$CATE(x) \equiv E[Y_{1i} - Y_{0i}|X_i = x]$$

CATE and ITE

Q: Why use CATE?

CATE and ITE

Q: Why use CATE?

A: The best estimator (in terms of MSE) for CATE is also the best estimator for ITE

CATE and ITE

Q: Why use CATE?

A: The best estimator (in terms of MSE) for CATE is also the best estimator for ITE (see Kunzel, Sekhon, Pickel, Yu (2019))

$$E[(ITE_i - \widehat{ITE}_i)^2 | X_i = x]$$
$$= E[(ITE_i - CATE(x))^2 | X_i = x] + E[(CATE(x) - \widehat{ITE}_i)^2]$$

CATE and ITE

Q: Why use CATE?

A: The best estimator (in terms of MSE) for CATE is also the best estimator for ITE (see Kunzel, Sekhon, Pickel, Yu (2019))

$$
\begin{aligned}
& E[(ITE_i - \widehat{ITE}_i)^2 | X_i = x] \\
&= E[(ITE_i - CATE(x))^2 | X_i = x] + E[(CATE(x) - \widehat{ITE}_i)^2]
\end{aligned}
$$

Notice that the first term on the right hand side is comprised of population concepts, not estimators. So, all we can minimize is the second term on the right hand side.

CATE and ITE

Q: Why use CATE?

A: The best estimator (in terms of MSE) for CATE is also the best estimator for ITE (see Kunzel, Sekhon, Pickel, Yu (2019))

$$
\begin{aligned}
&E[(ITE_i - \widehat{ITE}_i)^2 | X_i = x] \\
&= E[(ITE_i - CATE(x))^2 | X_i = x] + E[(CATE(x) - \widehat{ITE}_i)^2]
\end{aligned}
$$

Notice that the first term on the right hand side is comprised of population concepts, not estimators. So, all we can minimize is the second term on the right hand side.

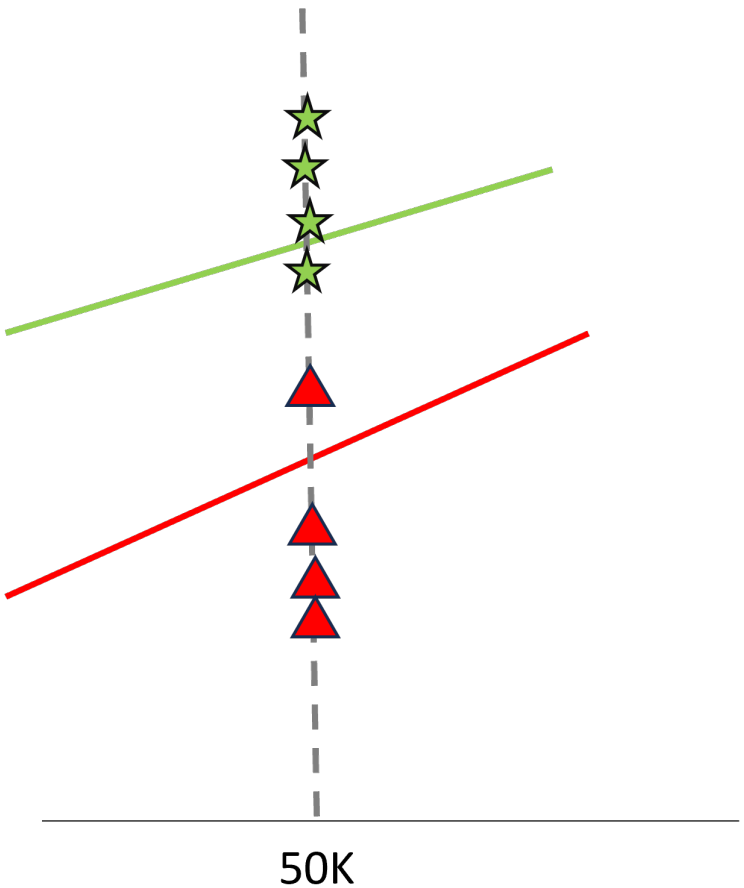
This implies that minimizing MSE for CATE at $X_i = x$ also minimizes MSE for ITE.

CATE vs ITE

Q: We are identifying individual i via X_i. So why isn't $CATE_i(x) = ITE_i$?

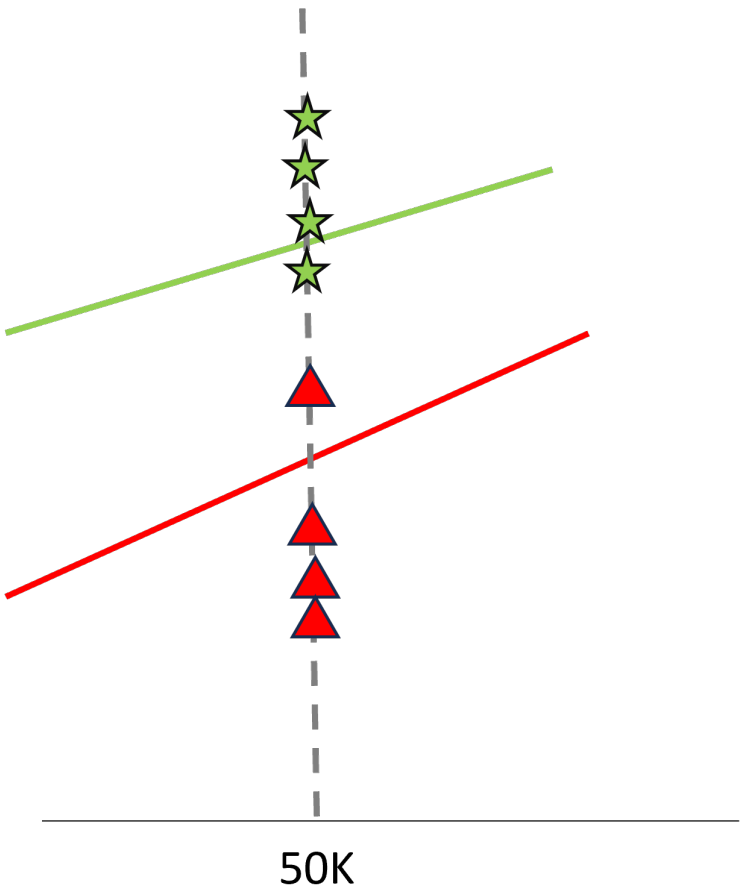
CATE vs ITE

Q: We are identifying individual i via X_i. So why isn't $CATE_i(x) = ITE_i$?



CATE vs ITE

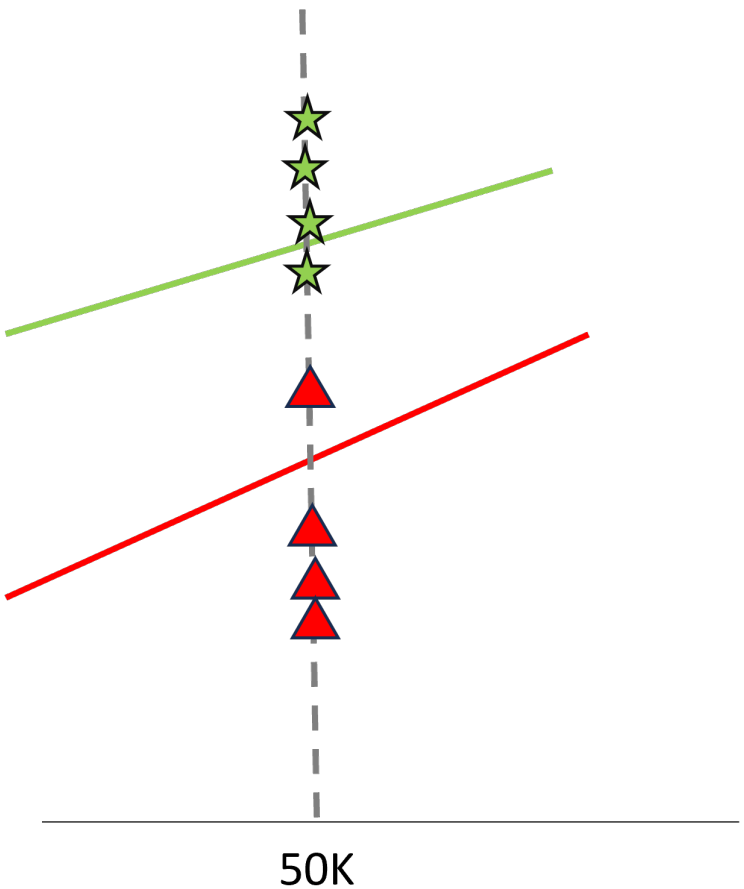
Q: We are identifying individual i via X_i. So why isn't $CATE_i(x) = ITE_i$?



A: If we zoom in on $X = \$50K$ we can see that CATE is not looking at a single person.

CATE vs ITE

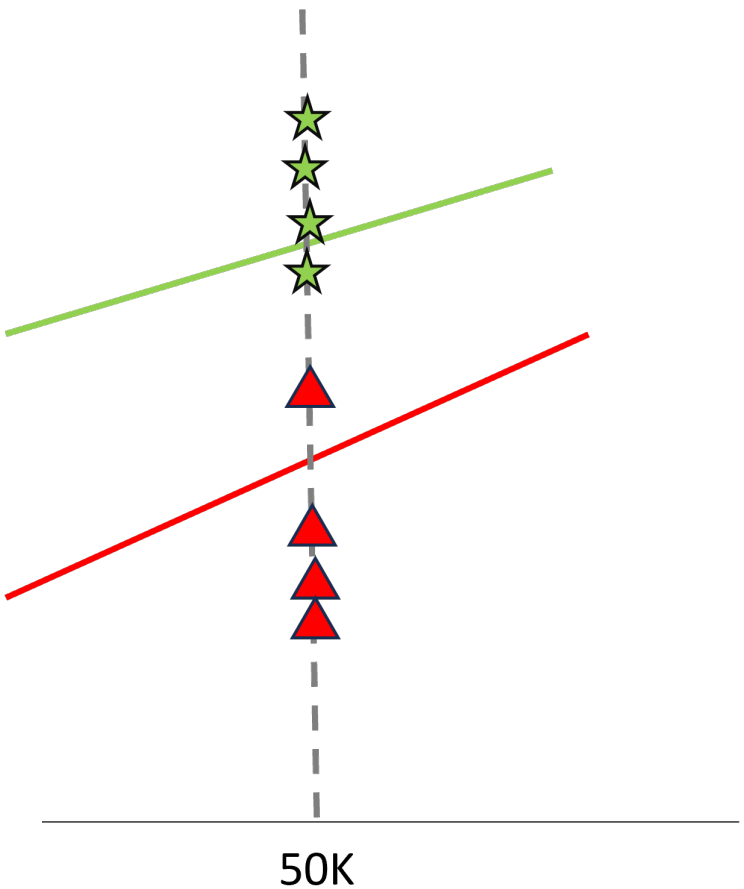
Q: We are identifying individual i via X_i. So why isn't $CATE_i(x) = ITE_i$?



A: If we zoom in on $X = \$50K$ we can see that CATE is not looking at a single person.
CATE is an average across several individuals, and therefore not an ITE_i.

CATE vs ITE

Q: We are identifying individual i via X_i. So why isn't $CATE_i(x) = ITE_i$?



A: If we zoom in on $X = \$50K$ we can see that CATE is not looking at a single person.
CATE is an average across several individuals, and therefore not an ITE_i.
We are still taking an average across individuals (like ATE), but we are clustering them by their characteristics X.

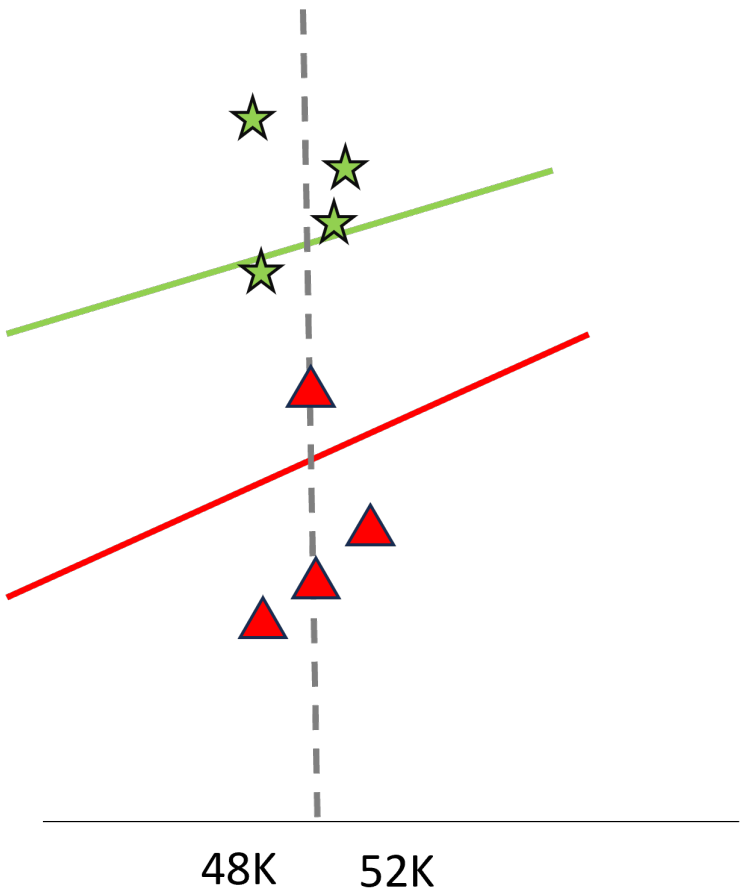
CATE's neighbors

Q: What if there aren't (m)any observations with exactly $X = \$50,000$?

CATE's neighbors

Q: What if there aren't (m)any observations with exactly $X = \$50,000$?

A: We can permit some jitter by considering a caliper around x.



Wide Range of Uses for CATE

Q: How use CATE in an RCT?

Wide Range of Uses for CATE

Q: How use CATE in an RCT?

A:

- We can include CATE within RCTs. Just because our treated and control groups have similar average characteristics (e.g. Income), doesn't mean that the treatment effect is necessarily constant as that characteristic varies.

Wide Range of Uses for CATE

Q: How use CATE in an RCT?

A:

- We can include CATE within RCTs. Just because our treated and control groups have similar average characteristics (e.g. Income), doesn't mean that the treatment effect is necessarily constant as that characteristic varies.
- e.g. the average age of our treated and control patients could be the same, but the "medicine" could be less effective as the patients age

Wide Range of Uses for CATE

Q: How use CATE in an RCT?

A:

- We can include CATE within RCTs. Just because our treated and control groups have similar average characteristics (e.g. Income), doesn't mean that the treatment effect is necessarily constant as that characteristic varies.
- e.g. the average age of our treated and control patients could be the same, but the "medicine" could be less effective as the patients age
- In fact, just like ATE, ATT, etc.., CATE can be applied to (m)any of our typical causal models.

Outline

1. Motivation - Case Study
2. CATE
3. Meta Learners
   - S-Learner
   - T-Learner
   - X-Learner
4. Causal Random Forests

Outline

1. Motivation - Case Study
2. CATE
3. Meta Learners
   - S-Learner
   - T-Learner
   - X-Learner
4. Causal Random Forests

S-Learner

The algorithm

1. Apply any base learner (e.g. regression, ML, etc..) to the entire dataset. Estimate $\mu = E[Y|X = x, D = d]$.

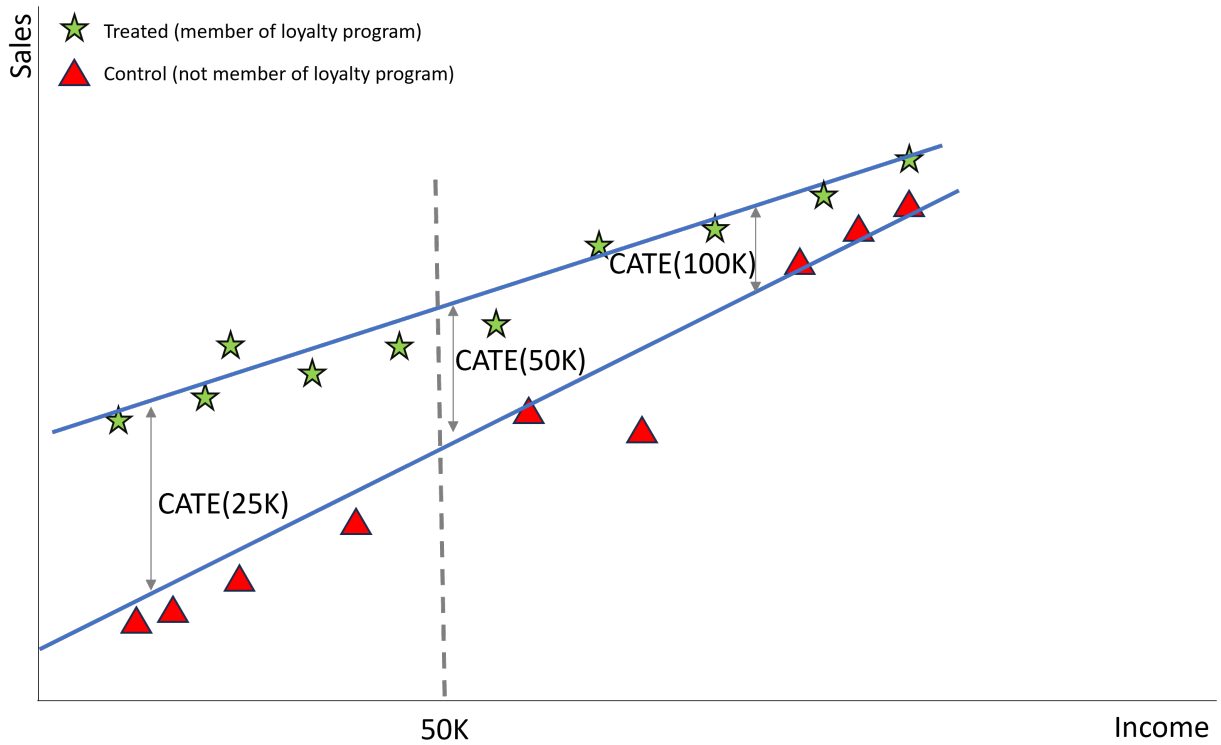
S-Learner

The algorithm

1. Apply any base learner (e.g. regression, ML, etc..) to the entire dataset. Estimate $\mu = E[Y|X = x, D = d]$.
2. Compute $\widehat{CATE} = \hat{\mu}(X = x, D = 1) - \hat{\mu}(X = x, D = 0)$

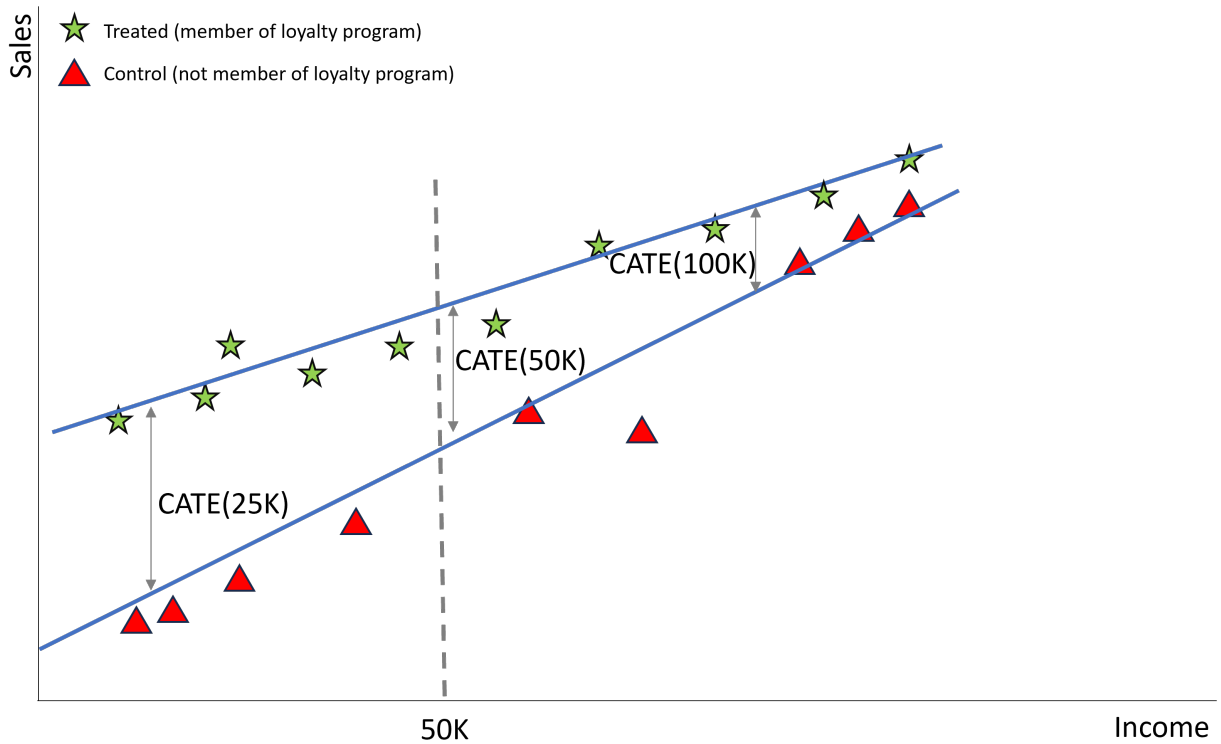
S-Learner - Intuition

Using linear regression with 1 interaction for illustration only



S-Learner - Intuition

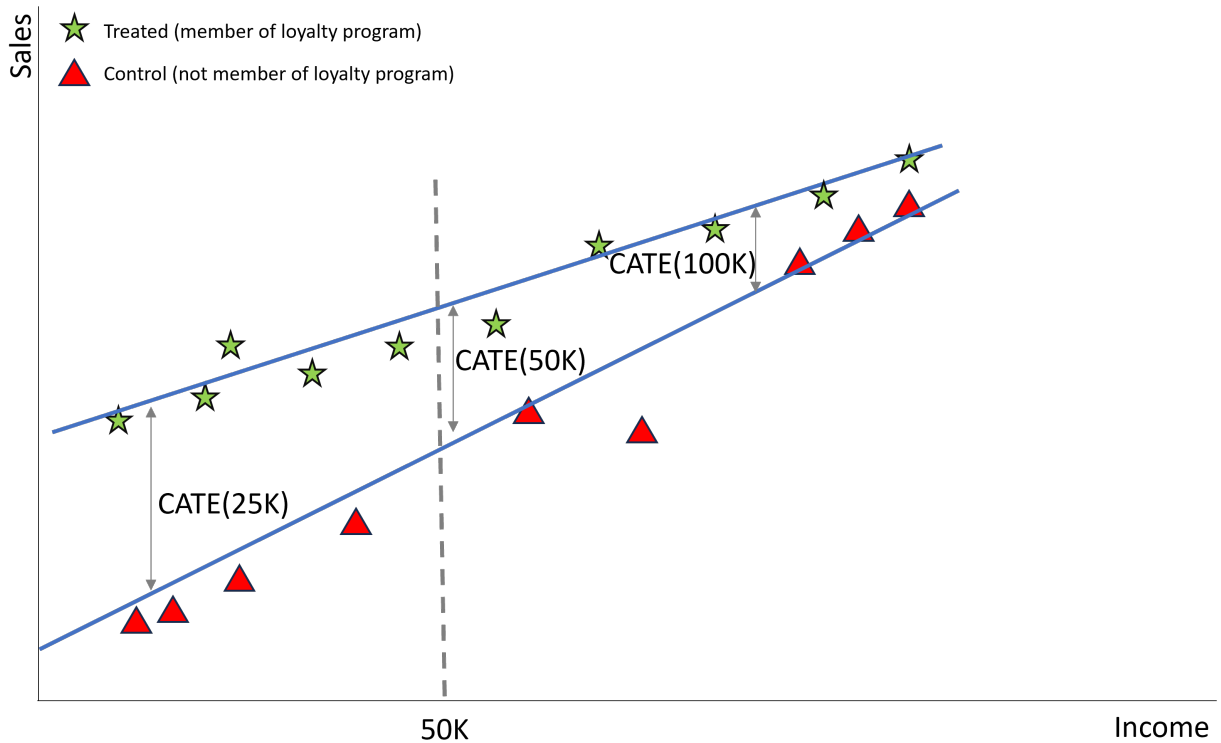
Using linear regression with 1 interaction for illustration only



$$Y = \alpha + \beta Loyalty + \gamma Income + \delta Loyalty \times Income + u$$

S-Learner - Intuition

Using linear regression with 1 interaction for illustration only

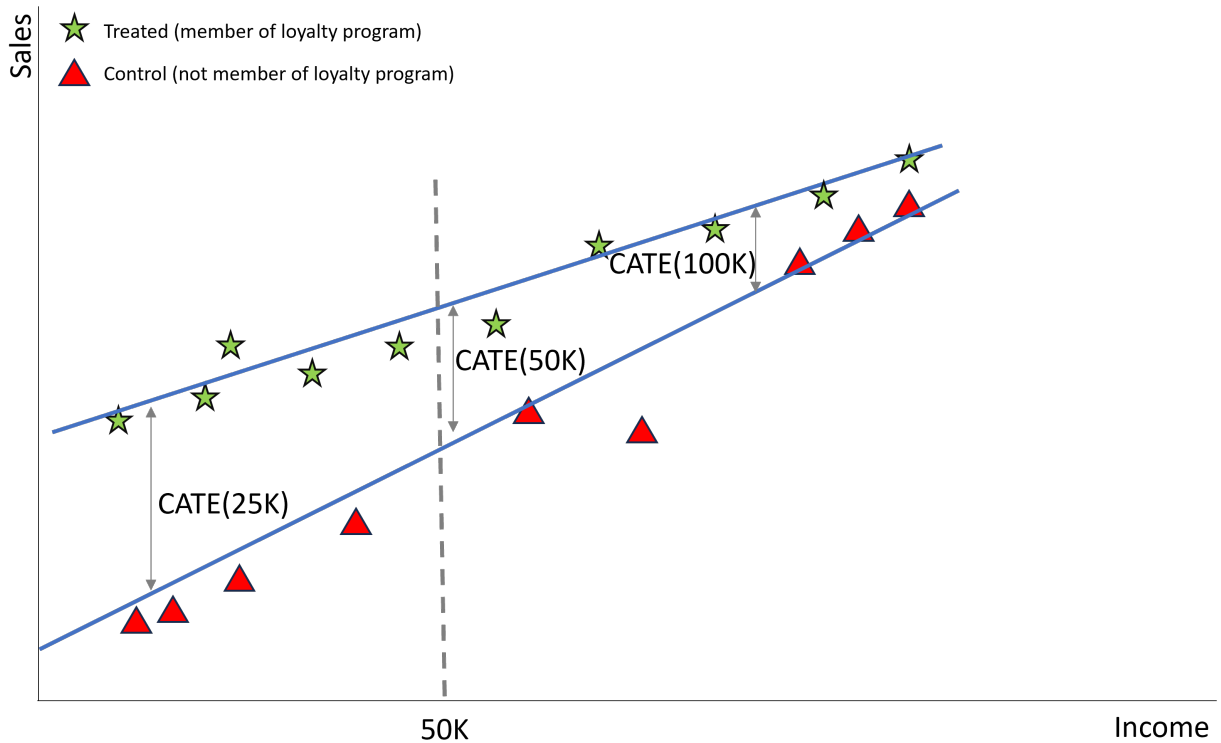


$$Y = \alpha + \beta Loyalty + \gamma Income + \delta Loyalty \times Income + u$$

$$CATE(Income = 50K) = (\hat{\alpha} + \hat{\beta} + \hat{\gamma} \times 50K + \hat{\delta} \times 50K) - (\hat{\alpha} + \hat{\gamma} \times 50K)$$

S-Learner - Intuition

Using linear regression with 1 interaction for illustration only



$$Y = \alpha + \beta Loyalty + \gamma Income + \delta Loyalty \times Income + u$$

$$\begin{aligned} CATE(Income = 50K) &= (\hat{\alpha} + \hat{\beta} + \hat{\gamma} \times 50K + \hat{\delta} \times 50K) - (\hat{\alpha} + \hat{\gamma} \times 50K) \\ &= \hat{\beta} + \hat{\delta} \times 50K \end{aligned}$$

Outline

1. Motivation - Case Study
2. CATE
3. Meta Learners
   - S-Learner
   - T-Learner
   - X-Learner
4. Causal Random Forests

T-Learner

The algorithm

1. Apply any base learner (e.g. regression, ML, etc..) to the control population only. Estimate $\mu_0 = E[Y_0|X = x]$. Note: the treatment variable is NOT included in the model.

T-Learner

The algorithm

1. Apply any base learner (e.g. regression, ML, etc..) to the control population only. Estimate $\mu_0 = E[Y_0|X = x]$. Note: the treatment variable is NOT included in the model.
2. Apply any, possibly different, base learner (e.g. regression, ML, etc..) to the treated population only. Estimate $\mu_1 = E[Y_1|X = x]$. Note: the treatment variable is NOT included in the model.

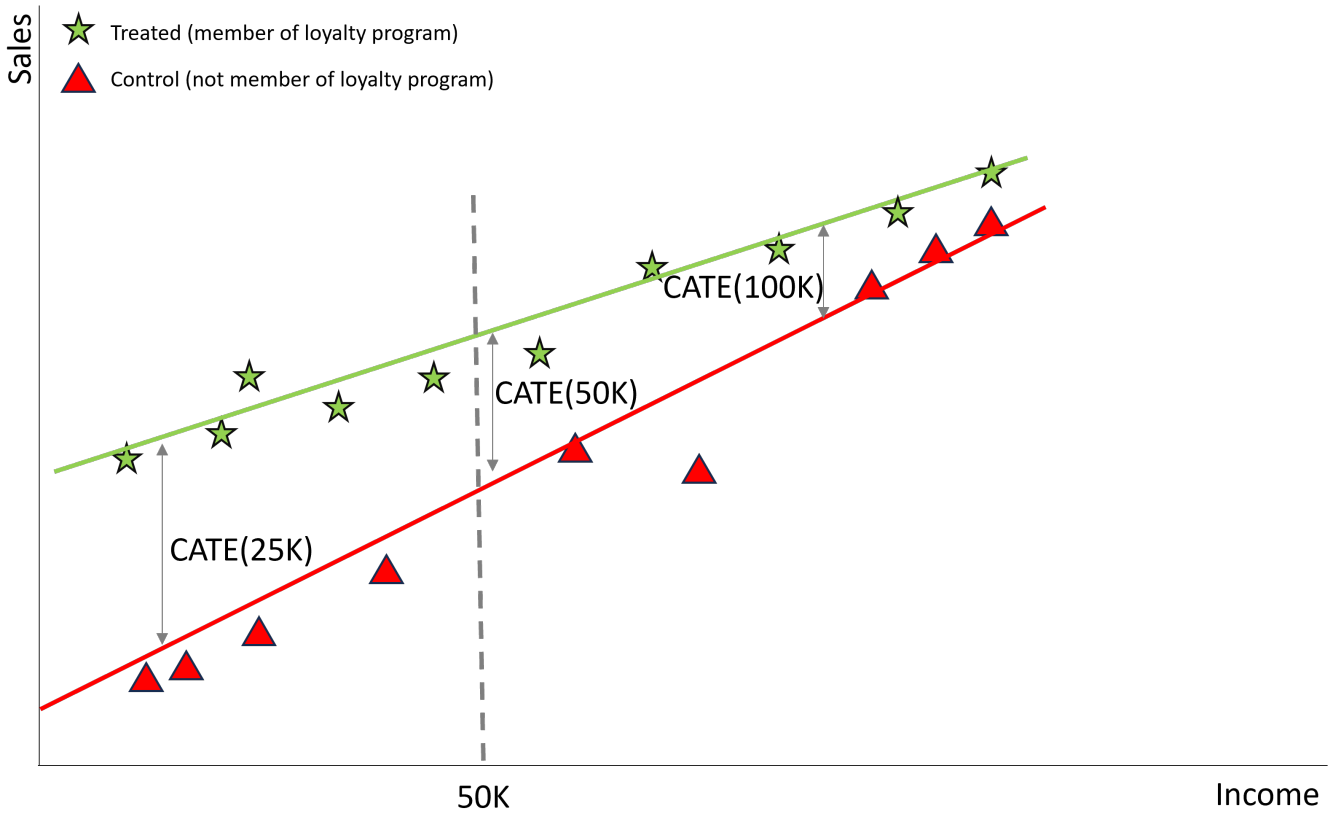
T-Learner

The algorithm

1. Apply any base learner (e.g. regression, ML, etc..) to the control population only. Estimate $\mu_0 = E[Y_0|X = x]$. Note: the treatment variable is NOT included in the model.
2. Apply any, possibly different, base learner (e.g. regression, ML, etc..) to the treated population only. Estimate $\mu_1 = E[Y_1|X = x]$. Note: the treatment variable is NOT included in the model.
3. Compute $\widehat{CATE} = \hat{\mu}_1 - \hat{\mu}_0$

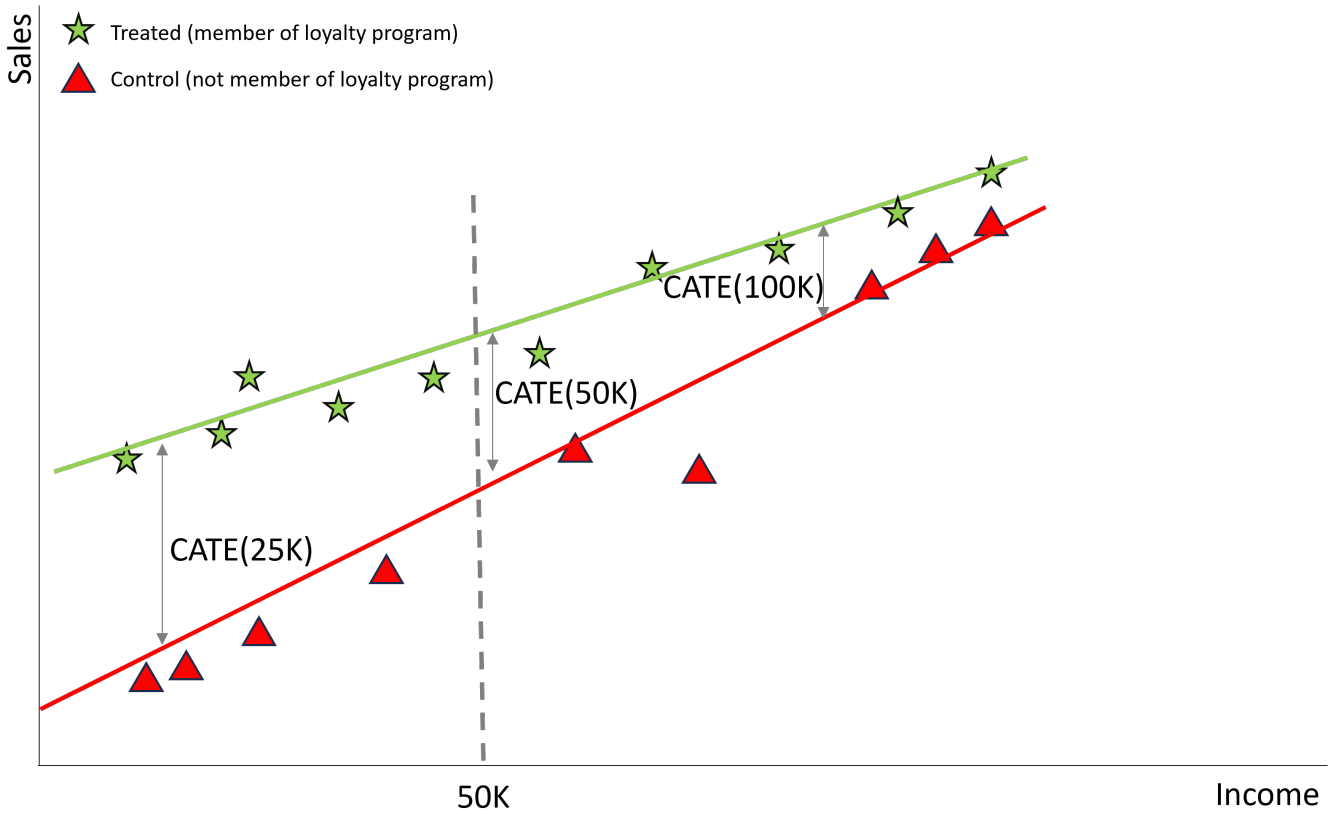
T-Learner - Intuition

Using linear regression with 1 interaction for illustration only



T-Learner - Intuition

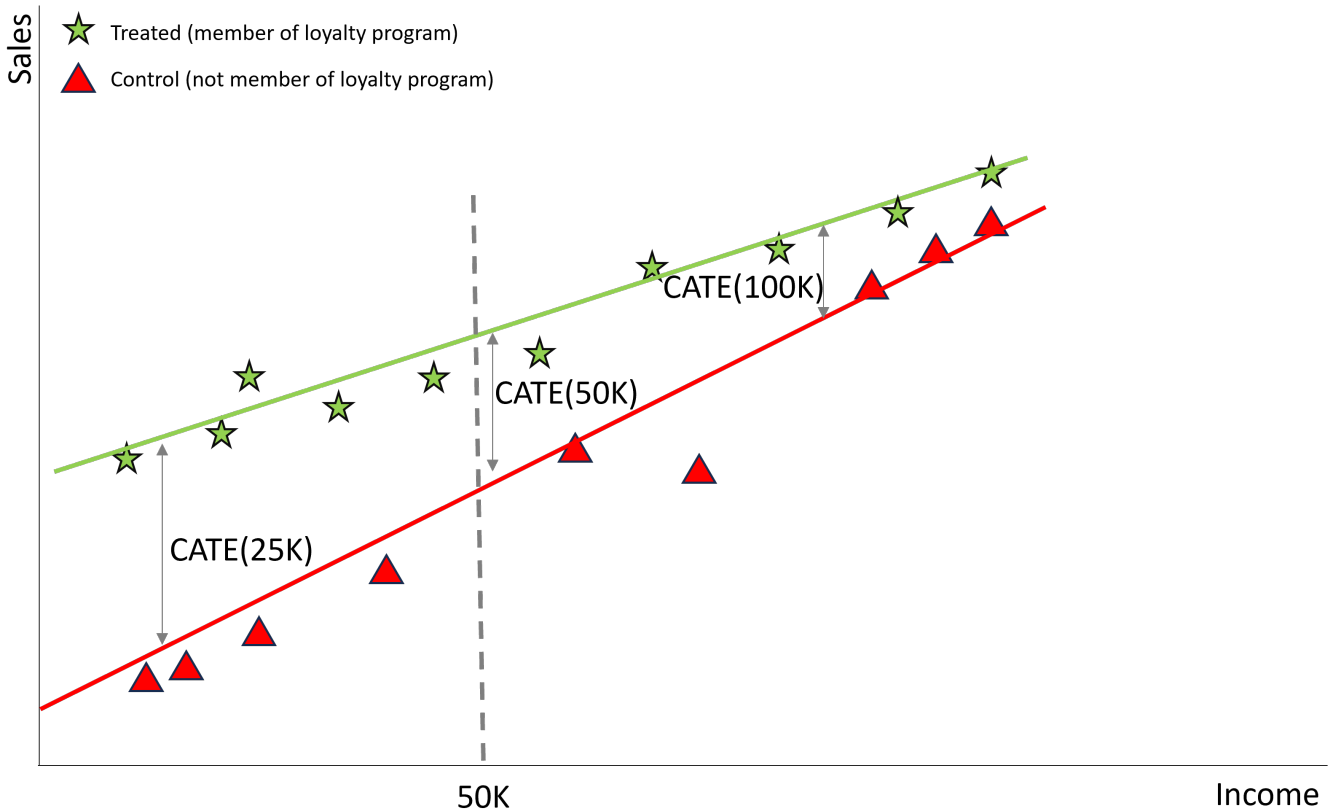
Using linear regression with 1 interaction for illustration only



$$Y = \alpha + \gamma Income + u$$

T-Learner - Intuition

Using linear regression with 1 interaction for illustration only

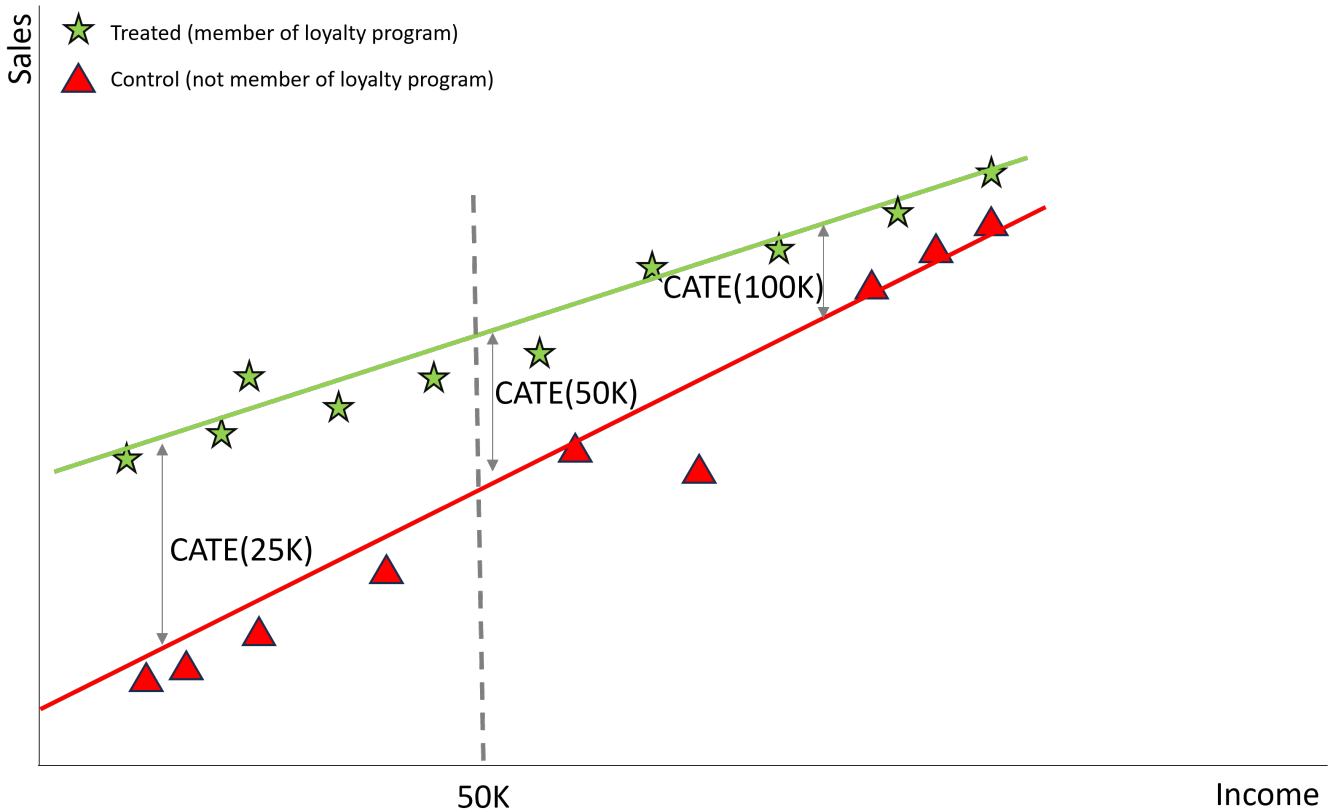


$$Y = \alpha + \gamma Income + u$$

$$CATE(Income = 50K) = (\hat{\alpha}_{Treat} + \hat{\gamma}_{Treat} \times 50K)$$

T-Learner - Intuition

Using linear regression with 1 interaction for illustration only



$$Y = \alpha + \gamma Income + u$$

$$CATE(Income = 50K) = (\hat{\alpha}_{Treat} + \hat{\gamma}_{Treat} \times 50K) - (\hat{\alpha}_{Control} + \hat{\gamma}_{Control} \times 50K)$$

Outline

1. Motivation - Case Study
2. CATE
3. Meta Learners
   - S-Learner
   - T-Learner
   - X-Learner
4. Causal Random Forests

X-Learner

The Algorithm

1. First Stage Learners

X-Learner

The Algorithm

1. First Stage Learners
   - Apply any base learner (e.g. regression, ML, etc..) to the control population only. Estimate $\mu_0 = E[Y_0|X = x]$. Note: the treatment variable is NOT included in the model.

X-Learner

The Algorithm

1. First Stage Learners
   - Apply any base learner (e.g. regression, ML, etc..) to the control population only. Estimate $\mu_0 = E[Y_0|X = x]$. Note: the treatment variable is NOT included in the model.

X-Learner

The Algorithm

1. First Stage Learners
   - Apply any base learner (e.g. regression, ML, etc..) to the control population only. Estimate $\mu_0 = E[Y_0|X = x]$. Note: the treatment variable is NOT included in the model.
   - Apply any, possibly different, base learner (e.g. regression, ML, etc..) to the treated population only. Estimate $\mu_1 = E[Y_1|X = x]$. Note: the treatment variable is NOT included in the model.

2. Second Stage Learners

2. Second Stage Learners
   - For each individual in treated group, impute the counterfactual untreated outcome using the control group's first stage learner and the treated individual's X. $\tilde{ITE}_i \equiv Y_{1i} - \hat{\mu}_0(X_{1i})$

2. Second Stage Learners
   - For each individual in treated group, impute the counterfactual untreated outcome using the control group's first stage learner and the treated individual's X. $\tilde{ITE}_i \equiv Y_{1i} - \hat{\mu}_0(X_{1i})$

2. Second Stage Learners
   - For each individual in treated group, impute the counterfactual untreated outcome using the control group's first stage learner and the treated individual's X. $\tilde{ITE}_i \equiv Y_{1i} - \hat{\mu}_0(X_{1i})$
   - For each individual in control group, impute the counterfactual treated outcome using the treated group's first stage learner and the control individual's X. $\tilde{ITE}_i \equiv \hat{\mu}_1(X_{0i}) - Y_{0i}$

2. Second Stage Learners
   - For each individual in treated group, impute the counterfactual untreated outcome using the control group's first stage learner and the treated individual's X. $\tilde{ITE}_i \equiv Y_{1i} - \hat{\mu}_0(X_{1i})$
   - For each individual in control group, impute the counterfactual treated outcome using the treated group's first stage learner and the control individual's X. $\tilde{ITE}_i \equiv \hat{\mu}_1(X_{0i}) - Y_{0i}$

2. Second Stage Learners
   - For each individual in treated group, impute the counterfactual untreated outcome using the control group's first stage learner and the treated individual's X. $\tilde{ITE}_i \equiv Y_{1i} - \hat{\mu}_0(X_{1i})$
   - For each individual in control group, impute the counterfactual treated outcome using the treated group's first stage learner and the control individual's X. $\tilde{ITE}_i \equiv \hat{\mu}_1(X_{0i}) - Y_{0i}$
   - Apply any base learner (e.g. regression, ML, etc..) to the imputed treatment effects $\tilde{ITE}_i$ for the treated population only. Estimate $CATE_1(x)$. Note: the treatment variable is not included in the model.

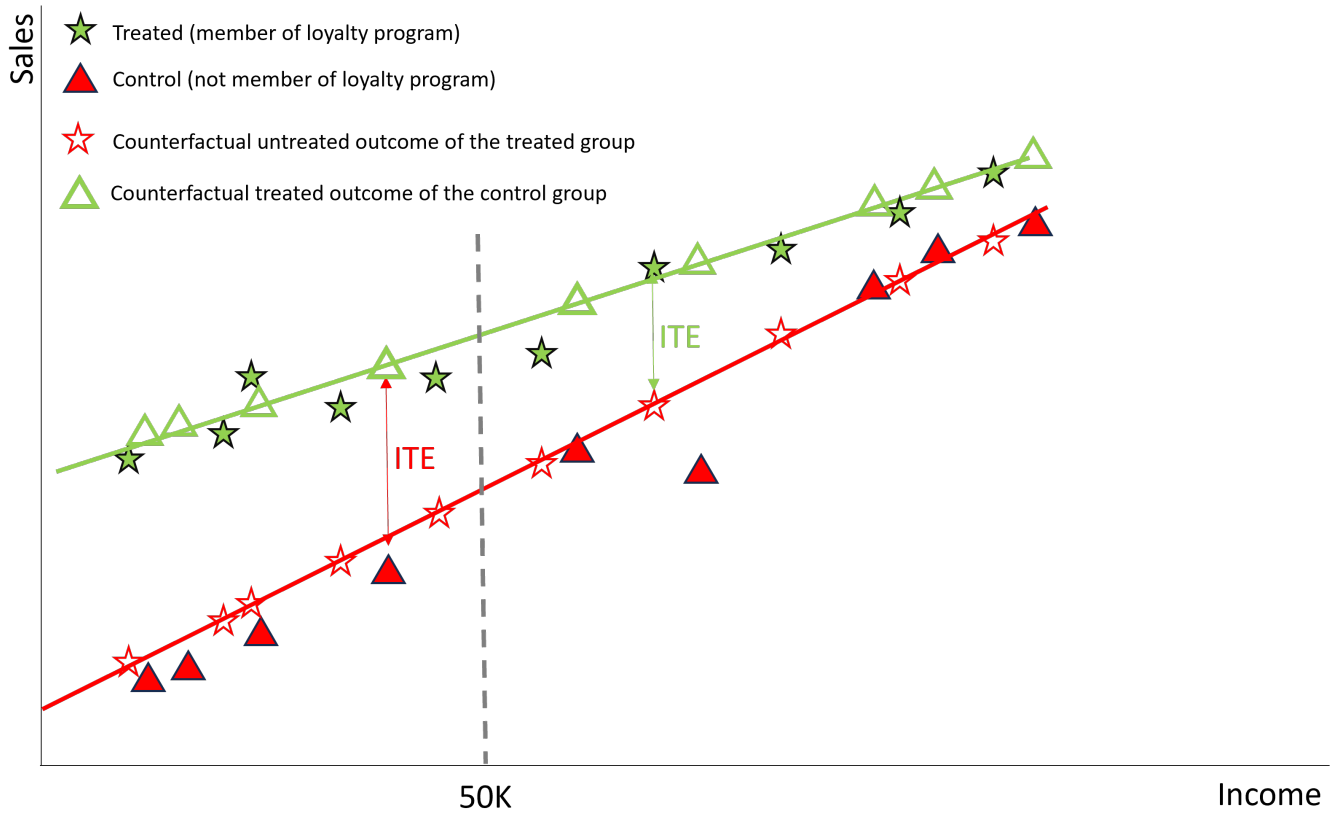
2. Second Stage Learners
   - For each individual in treated group, impute the counterfactual untreated outcome using the control group's first stage learner and the treated individual's X. $\tilde{ITE}_i \equiv Y_{1i} - \hat{\mu}_0(X_{1i})$
   - For each individual in control group, impute the counterfactual treated outcome using the treated group's first stage learner and the control individual's X. $\tilde{ITE}_i \equiv \hat{\mu}_1(X_{0i}) - Y_{0i}$
   - Apply any base learner (e.g. regression, ML, etc..) to the imputed treatment effects $\tilde{ITE}_i$ for the treated population only. Estimate $CATE_1(x)$. Note: the treatment variable is not included in the model.

2. Second Stage Learners
   - For each individual in treated group, impute the counterfactual untreated outcome using the control group's first stage learner and the treated individual's X. $\tilde{ITE}_i \equiv Y_{1i} - \hat{\mu}_0(X_{1i})$
   - For each individual in control group, impute the counterfactual treated outcome using the treated group's first stage learner and the control individual's X. $\tilde{ITE}_i \equiv \hat{\mu}_1(X_{0i}) - Y_{0i}$
   - Apply any base learner (e.g. regression, ML, etc..) to the imputed treatment effects $\tilde{ITE}_i$ for the treated population only. Estimate $CATE_1(x)$. Note: the treatment variable is not included in the model.
   - Apply any base learner (e.g. regression, ML, etc..) to the imputed treatment effects $\tilde{ITE}_i$ for the control population only. Estimate $CATE_0(x)$. Note: the treatment variable is not included in the model.

3. Compute $\widehat{CATE}(x) = g(x)CATE_1(x) + (1 - g(x))CATE_0(x)$ where $g \in [0, 1]$ is a weight function, usually set to the propensity score.

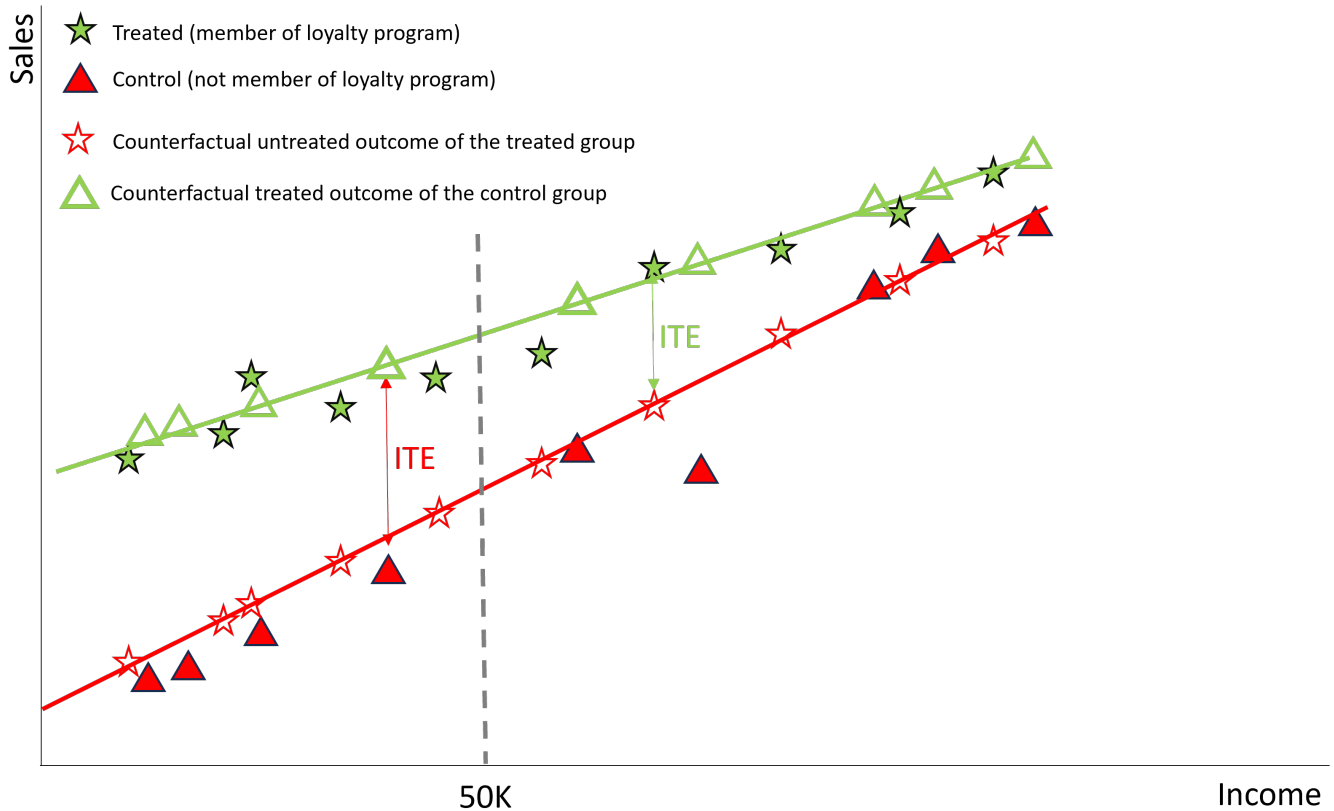
X-Learner Intuition

Using linear regressions with 1 interaction for illustration only



X-Learner Intuition

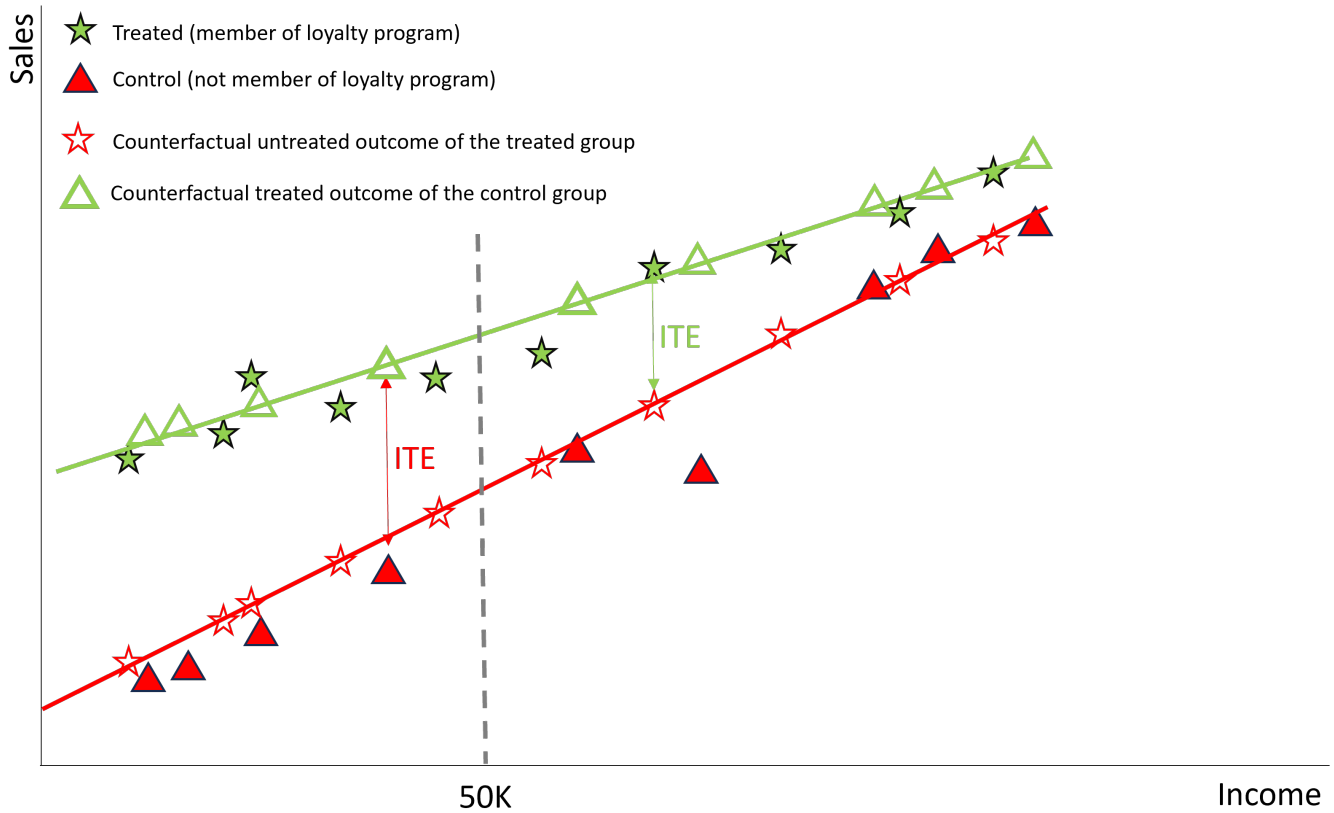
Using linear regressions with 1 interaction for illustration only



$$\tilde{ITE}_i^{Treated} = \alpha + \beta X_i + u_i \Rightarrow$$

X-Learner Intuition

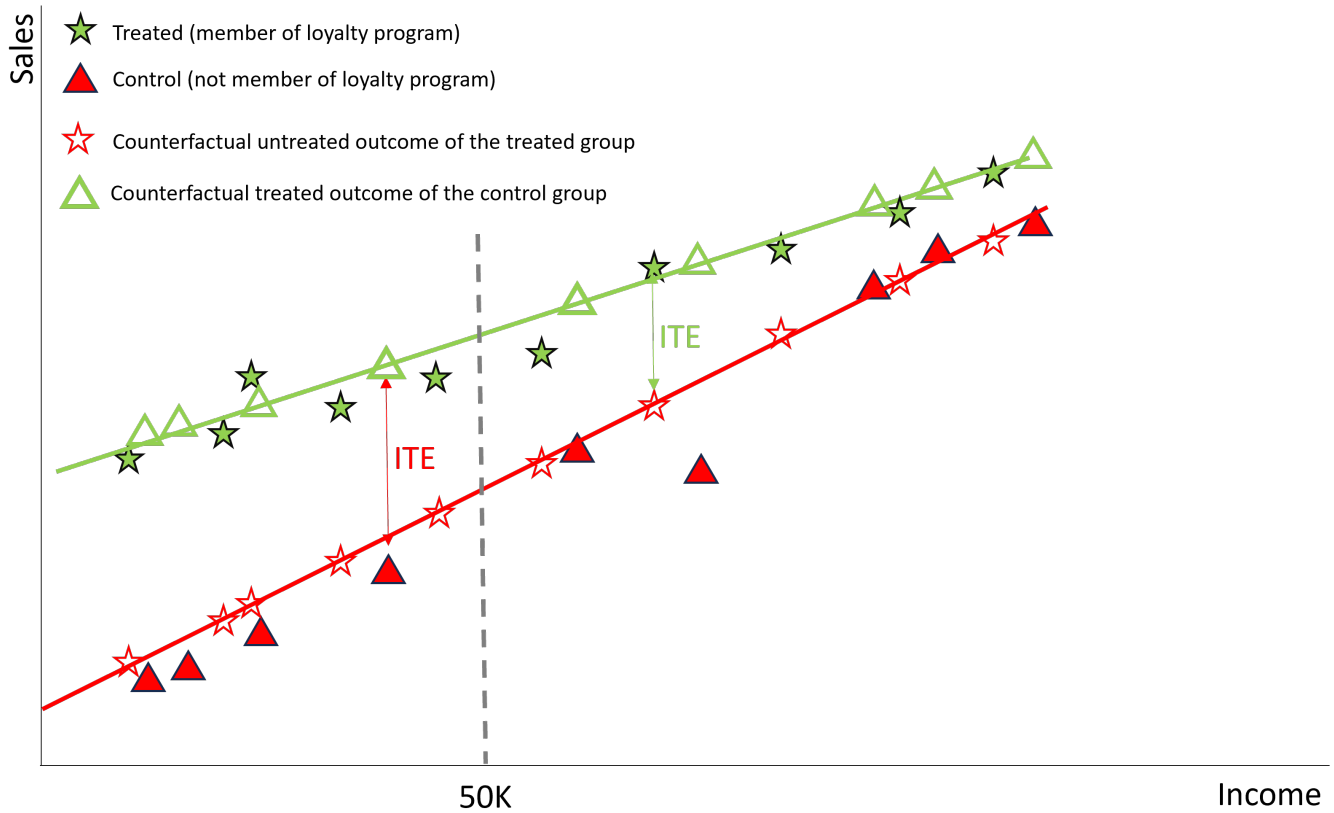
Using linear regressions with 1 interaction for illustration only



$$\tilde{ITE}_i^{Treated} = \alpha + \beta X_i + u_i \Rightarrow \hat{CATE}^{Treated}(x) = \hat{\alpha} + \hat{\beta} x$$

X-Learner Intuition

Using linear regressions with 1 interaction for illustration only

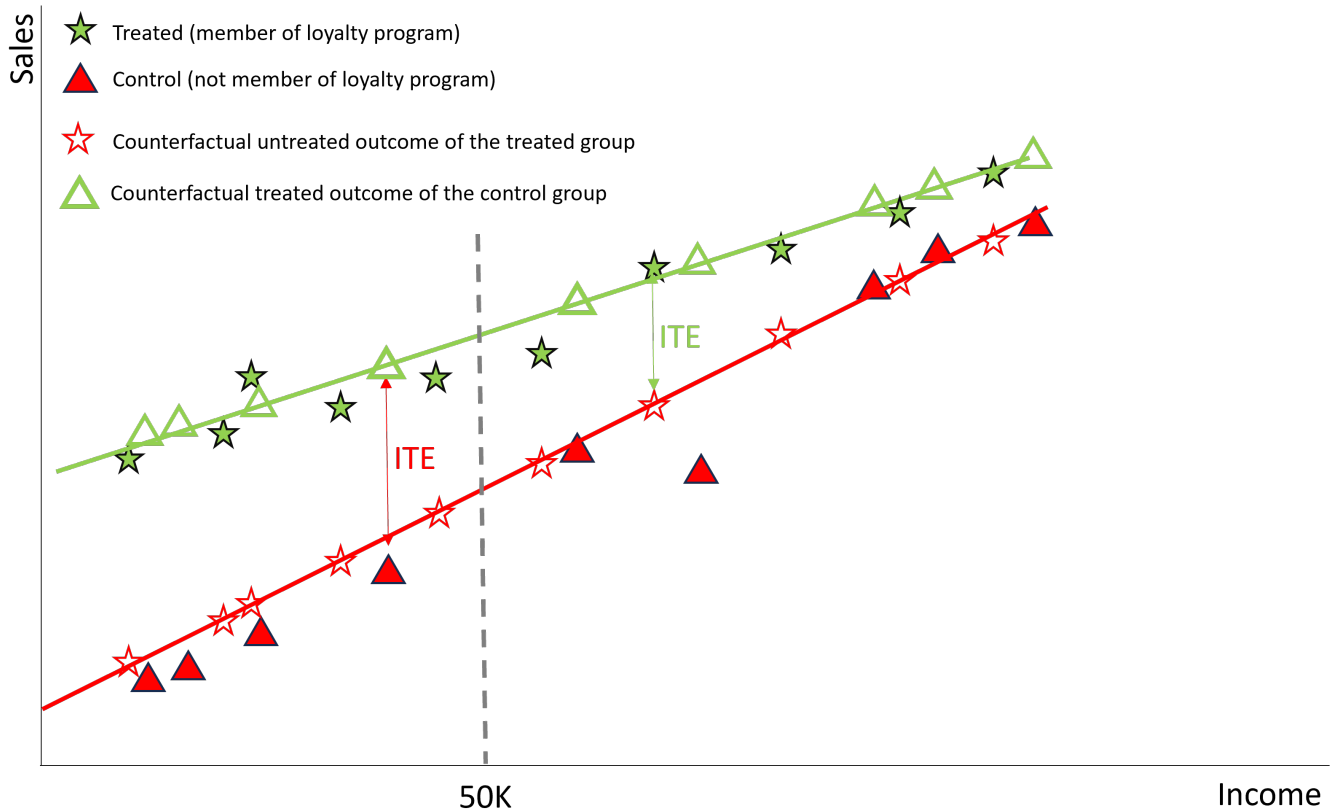


$$\tilde{ITE}_i^{Treated} = \alpha + \beta X_i + u_i \Rightarrow \quad \hat{CATE}^{Treated}(x) = \hat{\alpha} + \hat{\beta} x$$

$$\tilde{ITE}_i^{Control} = \alpha + \beta X_i + u_i \Rightarrow$$

X-Learner Intuition

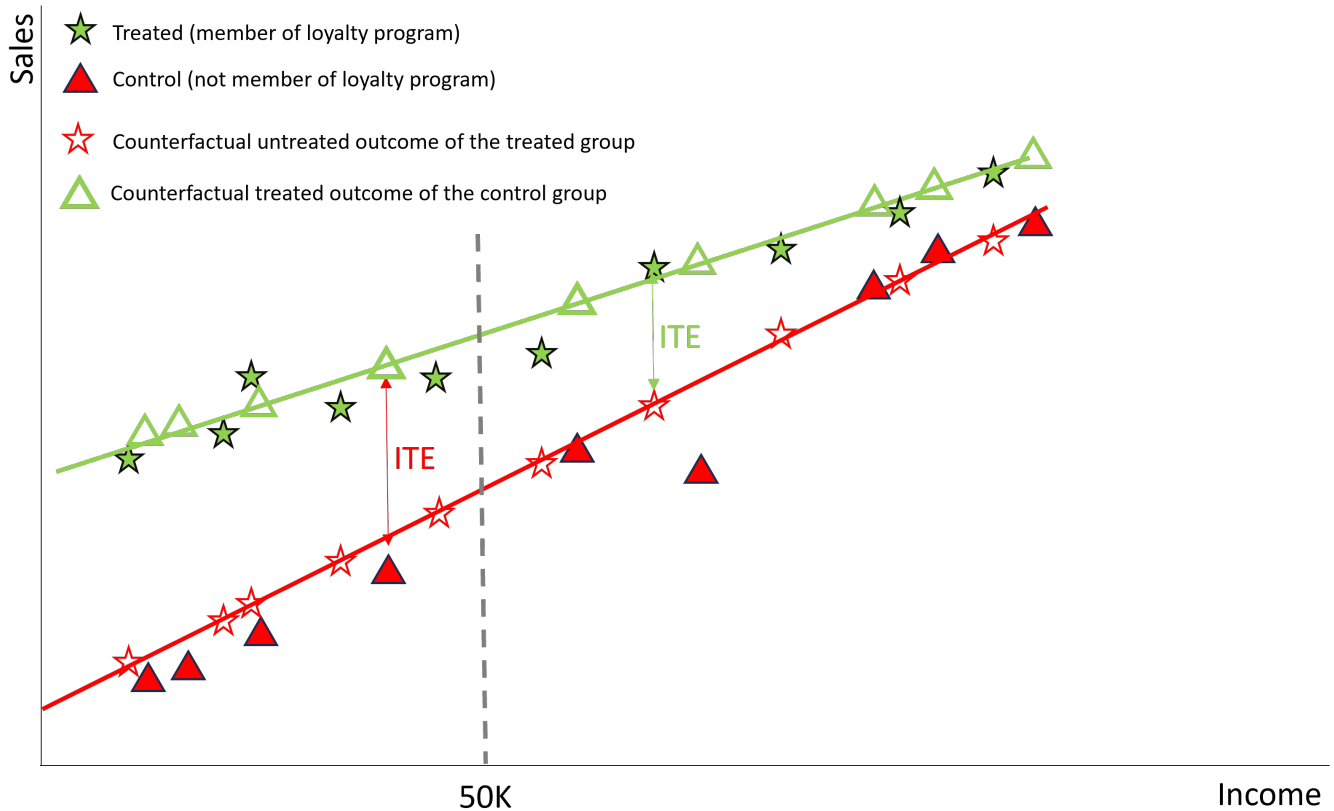
Using linear regressions with 1 interaction for illustration only



$$\tilde{ITE}_i^{Treated} = \alpha + \beta X_i + u_i \Rightarrow \hat{CATE}^{Treated}(x) = \hat{\alpha} + \hat{\beta}x$$
$$\tilde{ITE}_i^{Control} = \alpha + \beta X_i + u_i \Rightarrow \hat{CATE}^{Control}(x) = \hat{\alpha} + \hat{\beta}x$$

X-Learner Intuition

Using linear regressions with 1 interaction for illustration only



$$\tilde{ITE}_i^{Treated} = \alpha + \beta X_i + u_i \Rightarrow \hat{CATE}^{Treated}(x) = \hat{\alpha} + \hat{\beta} x$$

$$\tilde{ITE}_i^{Control} = \alpha + \beta X_i + u_i \Rightarrow \hat{CATE}^{Control}(x) = \hat{\alpha} + \hat{\beta} x$$

$$\hat{CATE}(x) = \text{weighted avg of } \hat{CATE}^{Treated}(x) \text{ and } \hat{CATE}_i^{Control}(x)$$

Outline

1. Motivation - Case Study
2. CATE
3. Meta Learners
   - S-Learner
   - T-Learner
   - X-Learner
4. Causal Random Forests

RA to RF

Recall our attempt to accommodate nonlinearities by augmenting the RA model with numerous interaction terms.

$$\begin{aligned} \mathit{Sales} = \alpha + \beta\,\mathit{Loyalty} + \gamma\,\mathit{Income} + \delta\,\mathit{Loyalty} \times \mathit{Income} + \\ \pi_0\,\mathit{Threshold} + \pi_1\,\mathit{Threshold} \times \mathit{Loyalty} + \pi_2\,\mathit{Threshold} \times \mathit{Income} \\ + \pi_3\,\mathit{Threshold} \times \mathit{Loyalty} \times \mathit{Income} + u \end{aligned}$$

RA to RF

Recall our attempt to accommodate nonlinearities by augmenting the RA model with numerous interaction terms.

$$\begin{aligned} Sales = \alpha + \beta\, Loyalty + \gamma\, Income + \delta\, Loyalty \times Income + \\ \pi_0\, Threshold + \pi_1\, Threshold \times Loyalty + \pi_2\, Threshold \times Income \\ + \pi_3\, Threshold \times Loyalty \times Income + u \end{aligned}$$

The econometric "roots" of Regression Trees can be thought of as an RA model with many interactions.

Causal Random Forest Intuition

We know

- Trees capture non linearities well

Causal Random Forest Intuition

We know

- Trees capture non linearities well
- Trees typically form leaves/branches by minimizing the MSE of predicted vs realized outcomes

Causal Random Forest Intuition

We know

- Trees capture non linearities well
- Trees typically form leaves/branches by minimizing the MSE of predicted vs realized outcomes
- Whereas linear models require us to set thresholds and interactions in advance

Causal Random Forest Intuition

We know

- Trees capture non linearities well
- Trees typically form leaves/branches by minimizing the MSE of predicted vs realized outcomes
- Whereas linear models require us to set thresholds and interactions in advance
- Forests tend to reduce overfitting relative to individual trees

Causal Random Forest Intuition

We know

- Trees capture non linearities well
- Trees typically form leaves/branches by minimizing the MSE of predicted vs realized outcomes
- Whereas linear models require us to set thresholds and interactions in advance
- Forests tend to reduce overfitting relative to individual trees

What's different about Causal Random Forests

Causal Random Forest Intuition

We know

- Trees capture non linearities well
- Trees typically form leaves/branches by minimizing the MSE of predicted vs realized outcomes
- Whereas linear models require us to set thresholds and interactions in advance
- Forests tend to reduce overfitting relative to individual trees

What's different about Causal Random Forests

- For Heterogenous Treatment Effects (HTE) causal trees form leaves/branches by

Causal Random Forest Intuition

We know

- Trees capture non linearities well
- Trees typically form leaves/branches by minimizing the MSE of predicted vs realized outcomes
- Whereas linear models require us to set thresholds and interactions in advance
- Forests tend to reduce overfitting relative to individual trees

What's different about Causal Random Forests

- For Heterogenous Treatment Effects (HTE) causal trees form leaves/branches by maximizing treatment effect heterogeneity, not prediction accuracy

Tree Splitting

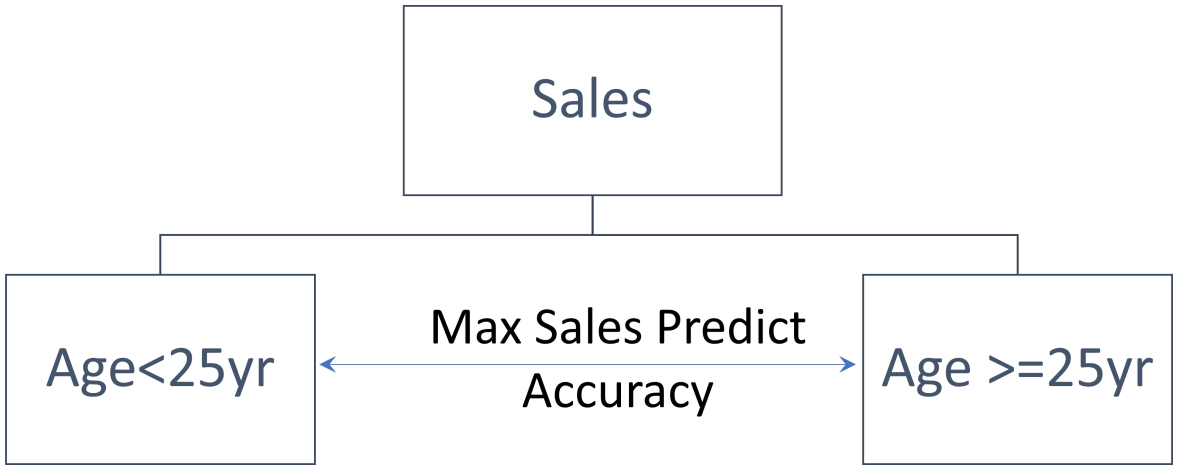


Figure: Regression Tree

Tree Splitting

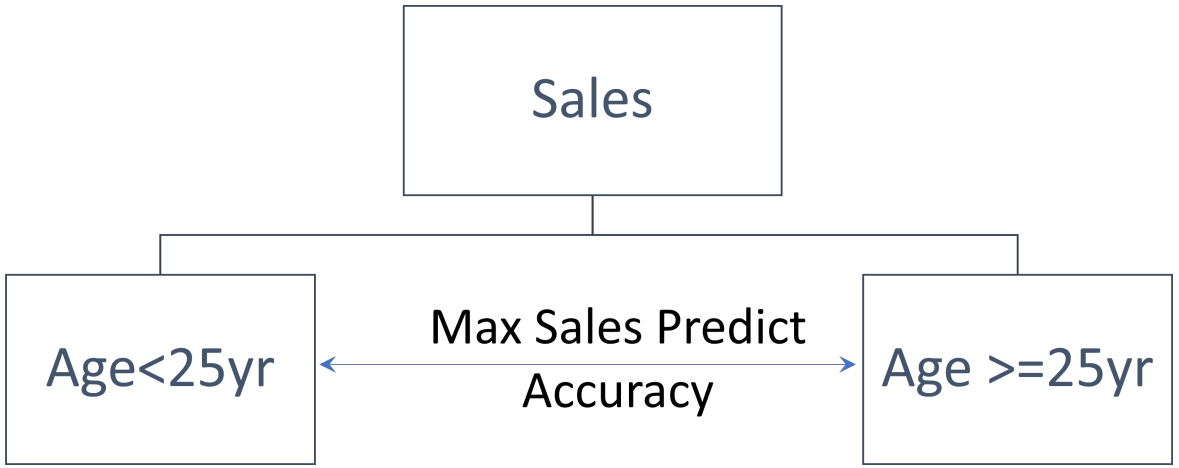


Figure: Regression Tree

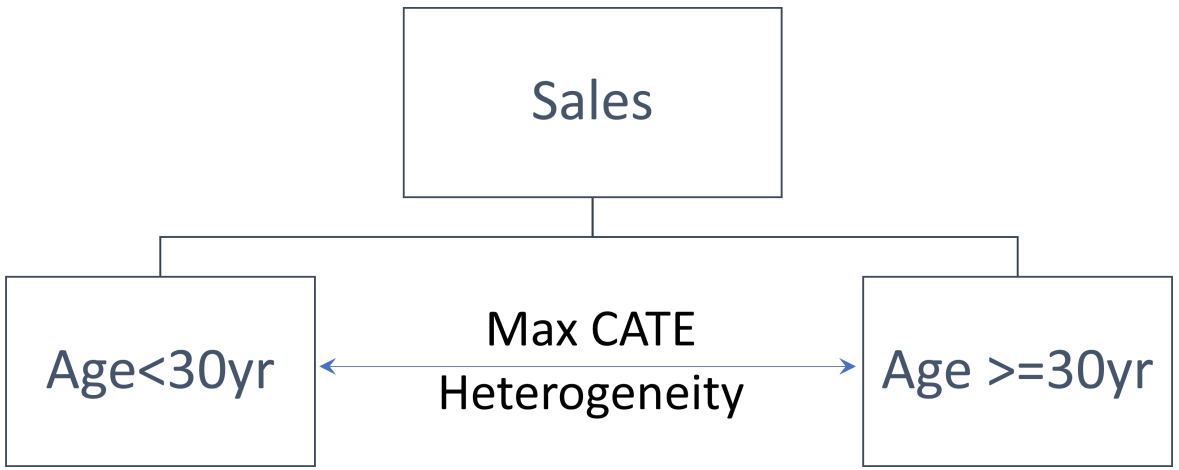


Figure: Causal Tree might split differently because objective is different

Causal Random Forest Assumptions

- We assume treatment effect homogeneity within each terminal leaf such that $CATE = \tau \ \forall x \in X$ some neighborhood of X values

Causal Random Forest Assumptions

- We assume treatment effect homogeneity within each terminal leaf such that $CATE = \tau \; \forall x \in X$ some neighborhood of X values
- Potential outcomes are not observable, so typical optimization criteria not available. Instead, causal forests are "honest" (ala Athey and Ibens (2016)) such that the population used to construct the tree partition is distinct from the population used to estimate the treatment effect.

Causal Random Forest: Meta-Algorithm

- For each tree in the forest:
  - Draw a random subsample of the data.

Causal Random Forest: Meta-Algorithm

- For each tree in the forest:
  - Draw a random subsample of the data.
  - Partition this subsample into two parts:

Causal Random Forest: Meta-Algorithm

- For each tree in the forest:
  - Draw a random subsample of the data.
  - Partition this subsample into two parts:
    - **Structure sample** (determine tree splits)

Causal Random Forest: Meta-Algorithm

- For each tree in the forest:
  - Draw a random subsample of the data.
  - Partition this subsample into two parts:
    - **Structure sample** (determine tree splits)
    - **Estimation sample** (estimate treatment effects)

Causal Random Forest: Meta-Algorithm

- For each tree in the forest:
  - Draw a random subsample of the data.
  - Partition this subsample into two parts:
    - **Structure sample** (determine tree splits)
    - **Estimation sample** (estimate treatment effects)
- Using the structure sample, choose **tree splits** that maximize treatment effect heterogeneity (not outcome prediction).

Causal Random Forest: Meta-Algorithm

- For each tree in the forest:
  - Draw a random subsample of the data.
  - Partition this subsample into two parts:
    - **Structure sample** (determine tree splits)
    - **Estimation sample** (estimate treatment effects)
- Using the structure sample, choose **tree splits** that maximize treatment effect heterogeneity (not outcome prediction).
- Within each leaf, compute the **average treatment effect** as the difference in mean outcomes between treated and control units.

Causal Random Forest: Meta-Algorithm

- For each tree in the forest:
  - Draw a random subsample of the data.
  - Partition this subsample into two parts:
    - **Structure sample** (determine tree splits)
    - **Estimation sample** (estimate treatment effects)
- Using the structure sample, choose **tree splits** that maximize treatment effect heterogeneity (not outcome prediction).
- Within each leaf, compute the **average treatment effect** as the difference in mean outcomes between treated and control units.
- Assign that leaf's treatment effect estimate to all units in the leaf.

Causal Random Forest: Meta-Algorithm

- For each tree in the forest:
  - Draw a random subsample of the data.
  - Partition this subsample into two parts:
    - **Structure sample** (determine tree splits)
    - **Estimation sample** (estimate treatment effects)
- Using the structure sample, choose **tree splits** that maximize treatment effect heterogeneity (not outcome prediction).
- Within each leaf, compute the **average treatment effect** as the difference in mean outcomes between treated and control units.
- Assign that leaf's treatment effect estimate to all units in the leaf.
- Repeat for many trees.

Causal Random Forest: Meta-Algorithm

- For each tree in the forest:
  - Draw a random subsample of the data.
  - Partition this subsample into two parts:
    - **Structure sample** (determine tree splits)
    - **Estimation sample** (estimate treatment effects)
- Using the structure sample, choose **tree splits** that maximize treatment effect heterogeneity (not outcome prediction).
- Within each leaf, compute the **average treatment effect** as the difference in mean outcomes between treated and control units.
- Assign that leaf's treatment effect estimate to all units in the leaf.
- Repeat for many trees.
- For each covariate value x (equivalently, for each unit i with covariates X_i), average its tree-level estimates to obtain the estimated Conditional Average Treatment Effect evaluated at X_i.

Causal Forests

Algorithm as per Athey and Wagner (2020)

1. First Stage
   - Fit a regression forest to $m(x) = E[Y_i|X_i = x]$

Causal Forests

Algorithm as per Athey and Wagner (2020)

1. First Stage
   - Fit a regression forest to $m(x) = E[Y_i|X_i = x]$
   - Fit a regression forest to $e(x) = Pr[D_i|X_i = x]$ for the propensity score

Causal Forests

Algorithm as per Athey and Wagner (2020)

1. First Stage
   - Fit a regression forest to $m(x) = E[Y_i|X_i = x]$
   - Fit a regression forest to $e(x) = Pr[D_i|X_i = x]$ for the propensity score
2. Second Stage
   - Use the First Stage forests to make out-of-the-bag (OOB) forecasts $\hat{m}^{(-i)}$ and $\hat{e}^{(-i)}$ (OOB implies that

Causal Forests

Algorithm as per Athey and Wagner (2020)

1. First Stage
   - Fit a regression forest to $m(x) = E[Y_i|X_i = x]$
   - Fit a regression forest to $e(x) = Pr[D_i|X_i = x]$ for the propensity score
2. Second Stage
   - Use the First Stage forests to make out-of-the-bag (OOB) forecasts $\hat{m}^{(-i)}$ and $\hat{e}^{(-i)}$ (OOB implies that i wasn't used to to construct predictors with $(-i)$ subscript)

Causal Forests

Algorithm as per Athey and Wagner (2020)

1. First Stage
   - Fit a regression forest to $m(x) = E[Y_i|X_i = x]$
   - Fit a regression forest to $e(x) = Pr[D_i|X_i = x]$ for the propensity score
2. Second Stage
   - Use the First Stage forests to make out-of-the-bag (OOB) forecasts $\hat{m}^{(-i)}$ and $\hat{e}^{(-i)}$ (OOB implies that i wasn't used to to construct predictors with $(-i)$ subscript)
   - Use these forecasts to grow a causal forest, where $\alpha_i(x)$ that capture how often i falls in the same leaf as x

Causal Forests

Algorithm as per Athey and Wagner (2020)

1. First Stage
   - Fit a regression forest to $m(x) = E[Y_i|X_i = x]$
   - Fit a regression forest to $e(x) = Pr[D_i|X_i = x]$ for the propensity score
2. Second Stage
   - Use the First Stage forests to make out-of-the-bag (OOB) forecasts $\hat{m}^{(-i)}$ and $\hat{e}^{(-i)}$ (OOB implies that i wasn't used to to construct predictors with $(-i)$ subscript)
   - Use these forecasts to grow a causal forest, where $\alpha_i(x)$ that capture how often i falls in the same leaf as x
   - Estimate CATE as

$$\hat{\tau}(x) = \frac{\sum_{i=1}^{N} \hat{\alpha}_i(x)(Y_i - \hat{m}^{(-i)}(X_i))(D_i - \hat{e}^{(-i)}(X_i))}{\sum_{i=1}^{N} \hat{\alpha}_i(x)(D_i - \hat{e}^{(-i)}(X_i))^2}$$

Common Support

Q: This algo doesn't explicitly compare Treat/Control, so we don't need common support. True or False?

Common Support

Q: This algo doesn't explicitly compare Treat/Control, so we don't need common support. True or False?

- A: False

Common Support

Q: This algo doesn't explicitly compare Treat/Control, so we don't need common support. True or False?

- A: False
- Common Support remains a requirement

Common Support

Q: This algo doesn't explicitly compare Treat/Control, so we don't need common support. True or False?

- A: False
- Common Support remains a requirement
- In a crude interpretation, we need treated and control observations within each terminal node of the tree

Common Support

Q: This algo doesn't explicitly compare Treat/Control, so we don't need common support. True or False?

- A: False
- Common Support remains a requirement
- In a crude interpretation, we need treated and control observations within each terminal node of the tree
- This overlap requirement is imposed via propensity score $e(x)$ being uniformly bounded away from 0 and 1

Unconfoundedness

Q: This algo doesn't impute potential outcomes, and so doesn't require unconfoundedness. True/False?

Unconfoundedness

Q: This algo doesn't impute potential outcomes, and so doesn't require unconfoundedness. True/False?

- A: False
- Unconfoundedness (aka Conditional Independence Assumption, CIA) says that treatment assignment is essentially random once we condition on covariates

Unconfoundedness

Q: This algo doesn't impute potential outcomes, and so doesn't require unconfoundedness. True/False?

- A: False
- Unconfoundedness (aka Conditional Independence Assumption, CIA) says that treatment assignment is essentially random once we condition on covariates

$$\{Y_{i1}, Y_{i0}\} \perp D_i | X_i$$

Heterogeneous Treatment Effects (HTE) - Summary

- Individual Treatment Effects (ITE) are useful, but we can't estimate, so...

Heterogeneous Treatment Effects (HTE) - Summary

- Individual Treatment Effects (ITE) are useful, but we can't estimate, so...
- ... we resort to averages such as the ATE, but....

Heterogeneous Treatment Effects (HTE) - Summary

- Individual Treatment Effects (ITE) are useful, but we can't estimate, so...
- ... we resort to averages such as the ATE, but....
- ... the ATE imposes constant marginal effects on entire domain

Heterogeneous Treatment Effects (HTE) - Summary

- Individual Treatment Effects (ITE) are useful, but we can't estimate, so...
- ... we resort to averages such as the ATE, but....
- ... the ATE imposes constant marginal effects on entire domain
- Our solution: Use CATE

Heterogeneous Treatment Effects (HTE) - Summary

- Individual Treatment Effects (ITE) are useful, but we can't estimate, so...
- ... we resort to averages such as the ATE, but....
- ... the ATE imposes constant marginal effects on entire domain
- Our solution: Use CATE (conditional average treatment effect) as a proxy of ITE based upon observable characteristics

Heterogeneous Treatment Effects (HTE) - Summary

- Individual Treatment Effects (ITE) are useful, but we can't estimate, so...
- ... we resort to averages such as the ATE, but....
- ... the ATE imposes constant marginal effects on entire domain
- Our solution: Use CATE (conditional average treatment effect) as a proxy of ITE based upon observable characteristics
- Non linear models (e.g. Causal Random Forests and Meta Learners) are options to estimate CATEs