# 5: DAGs and Potential Outcomes

## Do()ing Observational Causal Inference

### Structural models

* A DAG (directed acyclic graph) can be expressed as a series of nested models/relationships.
  + These don’t have to exist in real-life as it’s a conceptual model of the relationship between the treatment and outcome.
  + Nested because one function will often be an input to another.
* dagify() in the **ggdag** package forces the user to think of the relationship in terms of these nested functions.

### Causal identification

* These relationships are important because they explain the data-generating process and hence the relationship between treatment and outcome.
* Goal is to identify the relationships abstracting from all the other influences/confounders.
* A causal effect is identified if the association between treatment and outcome is properly stripped and isolated.
* Associations between variables can be redirected/controlled for by “adjusting”/”conditioning”.

### Three types of associations

* Confounding (common cause): some variable affects both treatment and outcome. Can apply some fancy statistics to remove this path and isolate the relationship of interest.
* Causation (mediation): treatment affects outcome indirectly through this variable. Don’t need to necessarily control/adjust for this.
* Collision (selection/endogeneity): treatment and outcome both affect this variable. Definitely don’t ever want to control for this variable as it’ll open up a back-door/indirect path and the relationship between treatment and outcome will no longer be isolated i.e. the causal effect will be distorted.
* Ideally, want to be able to cut/remove some of the relationships. But the only way to do this is to have complete control over the program/intervention i.e. a RCT.

### Interventions

* If you can do() a node, you have control over the node and hence can force subjects to ‘do’ it or not ‘do’ it.
* or means the probability/expected value/causal effect of the outcome given a node (treatment) is set to a specific value.
* Any causal intervention can be expressed using this syntax.
* When you do() X, you delete all arrows entering that node.
  + Relationship between X and Y is fully identified.
  + The confounder is still affecting the outcome, but has no impact on the treatment decision.
  + No more back-doors. (Why RCTs are so powerful).
* Can calculate the causal effect of treatment on the outcome (assuming the data exists) using a simple regression model.
  + Don’t need to control for anything.
* Simplifies DAG’s significantly and lets you isolate specific arrows straight away.

### Undoing things

* What happens if we only have observational data?
* A plain regression analysis won’t reveal the causal affect between treatment and outcome as controlling treatment is not the same as treatment just existing.
  + Correlation does not equal causation.
* Need to get rid of the do() operator so produces a valid causal effect.
  + Achieved via fancy logic / algebra i.e. special rules.
* Official term is do-calculus. 3 rules that are proven to be true.

### Special cases of do-calculus

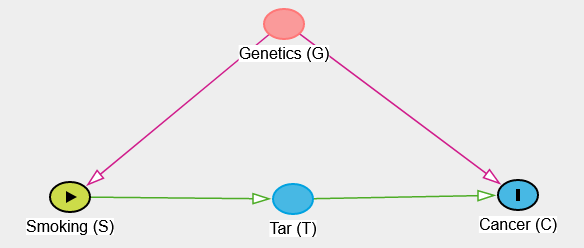
* Backdoor and front-door adjustment.
* Much more intuitive than the formal rules.

### Backdoor adjustment

* – the probability of the outcome given treatment can be expressed as the effect of the treatment on the outcome after adjusting for “Z”.
* Doing something statistically to close the backdoor between treatment and outcome, which is the same thing as ‘do()’ X (treatment).

### Front-door adjustment

* Used to prove the link between smoking and cancer.



* Tobacco industries argued that you couldn’t prove that there was a direct relationship between smoking and cancer.
* Genetics meant that some people are pre-disposed to smoke and some are pre-disposed to cancer, so genetics caused cancer not smoking.
* Once you controlled for genetics, there was no relationship between smoking and cancer.
* Epidemiologists used front-door adjustment instead.
  + Knew that the chemicals in tar caused cancer, genetics caused cancer, and smoking caused tar build-up.
  + Can estimate the causal effect of smoking on tar build-up as there’s no confounder i.e genetics isn’t a confounder of any sort.
  + Can also do the same between tar build-up and cancer as there’s also no confounders (ignoring smoking altogether).
  + According to the rules of do()-calculus, you can add those two effects together to get the causal effect of smoking on cancer.
* Can’t randomise/control for genetics.

### Moral of the story

* **If you can transform do() expressions to do-free versions, you can validly make causal inferences from observational data!!!**
  + By closing back or front-doors for example.
* Dagitty has backdoor adjustment built-in.
* Algorithms in R’s *causaleffect* package can also perform the official do-calculus for you.
  + Really helpful for a complicated DAG.
  + For example, even if it doesn’t look immediately obvious.
* Not all DAGs can be isolated. This means that you’re then stuck and can’t estimate the causal effect between treatment and outcome given only observational data.

## Potential Outcomes

### Program effect

* Many different ways of expressing the difference what would have happened under the treatment and the absence of the treatment:
* Do() operator language
  + – expected value of do()ing X vs not do()ing X
* Potential outcomes language
  + – Do() built in
  + – shorthand way of expressing the above
* The tricky thing is that it’s often difficult or impossible to measure the effect of treatment absence for an individual subject.

### Fundamental problem of causal inference

* in real-life is
* We have no clue of what the outcome would be if a subject hadn’t undergone treatment (or had undergone treatment in the case of a control).
* Impossible to observe individual-level effects simultaneously i.e. impossible to know the counterfactual for any particular subject.

### Average treatment effects

* Solution: use averages instead
* Looking at group-level effects instead of individual-level effects, which gets around the fundamental problem of causal effect.
* In a do() framework, ATE is the difference between the average/expected value when a treatment is ‘on’ vs the expected value when the treatment is ‘off’.
* BUT, this doesn’t exactly work all the time. You can’t just look at the difference in the averages of subjects undergoing treatment vs not undergoing treatment due to built-in selection bias.
* Consider a fake dataset:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Subject** | **Age** | **Treated** | **Outcome with program** | **Outcome without program** | **Effect** |
| 1 | Old | 1 | 80 | 60 | 20 |
| 2 | Old | 1 | 75 | 70 | 5 |
| 3 | Old | 1 | 85 | 80 | 5 |
| 4 | Old | 0 | 70 | 60 | 10 |
| 5 | Young | 1 | 75 | 70 | 5 |
| 6 | Young | 0 | 80 | 80 | 0 |
| 7 | Young | 0 | 90 | 100 | -10 |
| 8 | Young | 0 | 85 | 80 | 5 |

* + We have a time machine, so we can calculate the individual causal effects.

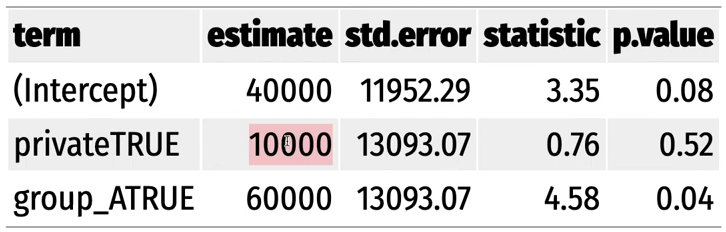
### Conditional Average Treatment Effect (CATE)

* ATE for sub-groups
* Means we can answer the question about whether treatment is more effective for specific covariates/characteristics.
* (and similarly for young subjects)
* Treatment helps old subjects by 10 units and does nothing for young subjects (0 units).
* Average treatment on the treated (ATT) effect – effect of the program on those that were treated.
  + extra years
* Averaged treatment on the untreated (ATU) effect – effect of the program on those that were untreated.
  + extra years
* Can determine the ATE by just averaging the ATT and ATU.
* In real life, we don’t see the counterfactual for each observation:

|  |  |  |  |
| --- | --- | --- | --- |
| **Subject** | **Age** | **Treated** | **Actual outcome** |
| 1 | Old | 1 | 80 |
| 2 | Old | 1 | 75 |
| 3 | Old | 1 | 85 |
| 4 | Old | 0 | 60 |
| 5 | Young | 1 | 75 |
| 6 | Young | 0 | 80 |
| 7 | Young | 0 | 100 |
| 8 | Young | 0 | 81 |

* + Can calculate the ATE using the weighted-averaged of confounder-based CATEs (as long as we can reasonably assume that treatment was randomly assigned within each level of the confounder).
  + No way of mathematically figuring whether this random assignment within confounder levels assumption is reasonable. Need to use judgement and subject matter expertise.
* Very tempting to do something like the following, but it ISN’T VALID unless treatment has been randomly assigned.
  + Not accounting for confounding/selection bias.
* Age was used in this example as it was correlated with and confounded the outcome according to the DAG.
* For the estimation to be valid, need to assume treatment is randomly assigned within the confounder sub-groups.

### Regression example

* Does attending private university cause an increase in income?
* Table shows the students grouped together based on application-acceptance similarity. We know their application history, acceptance results, what college they were admitted to (highlighted in grey) and their 1996 income.
* The tempting thing would be to subtract the average income of public students from the average income of private students. This is wrong as explained previously.
  + (Unless it was a true RCT in which case this calculation is actually correct).
* Need to draw the DAG or match them on something similar (student characteristics or Applicant Group).
* Gets us closer to the actual population-level ATE.
* Can also match with regression instead of CATE formulae:
  + model\_earnings <- lm(earnings ~ private + group\_A, data = schools\_small)
* 
  + The coefficient on private is the estimated impact of going to a private school on income.
  + The coefficient on group\_A is the estimated impact of being in Group A on income i.e. being is Group A increases average income by $60,000 compared to being in any of the other groups.
  + The intercept is the baseline income predicted by the model i.e. going to a public school and not being in Group A.
* The big advantage of a regression model are the additions of standard errors and p-values instead of just a point estimate.