Ethics of Data Science – Part II

Understanding bias

What is bias?

- The word “bias” carries many meanings in this course.
- It is often taken to mean “systematic error”, where “error” is a discrepancy between an estimator (which is a function of the data) and the quantity it is trying to estimate

$$E[\hat{\beta}] \neq \beta$$

- “Systematic” means that the error does not cancel out on average, or even over time.

- Sometimes, we will refer to process that render our estimators biased, as “biases”.
- In this lecture we will discuss one particular source of bias and how to mitigate it.

Treatment selection bias

- Recall our example of a Randomised Clinical Trial (RCT). It showed that younger patients benefit more from that treatment. One way to summarise that information is via the sample Average Treatment Effect (ATE):

$$\hat{\mathrm{ATE}} = \frac{1}{\#\{i : t_i = 1\}} \sum_{i \in \{i : t_i = 1\}}^{n} y_i - \frac{1}{\#\{i : t_i = 0\}} \sum_{i \in \{i : t_i = 0\}}^{n} y_i$$

```
RCT statistics:
             Younger    Older  Average
Placebo    1.4942583 1.447421 1.470840
Treatment 0.4713745 1.549656 1.010515
Sample ATE for RCT:  -0.46
```

Treatment selection bias

```
RCT statistics:
             Younger    Older  Average
Placebo    1.4942583 1.447421 1.470840
Treatment 0.4713745 1.549656 1.010515
Sample ATE for RCT:  -0.46
```

The sample ATE estimates ATE, which cannot be directly observed as it relies on counterfactuals:

$$\text{ATE} = \frac{1}{n} \sum_i \left(y_i^1 - y_i^0 \right)$$

Here y_i^1 is the outcome for patient i assuming they were treated and y_i^0 their outcome if they were not. In any given dataset we would only ever observe one of these quantities, never both simultaneously.

We often refer to these as *potential outcomes.*

Treatment selection bias

- The drug is approved and has some mild side effects. Clinicians hence prescribe it mostly to younger patients who are more likely to benefit.
- Real-World Data (RWD) are collected from the electronic health record of 100 adults. The sample ATE in that case is twice that of the RCT.

```
RCT statistics:
              Younger    Older  Average
Placebo    1.4942583 1.447421 1.470840
Treatment 0.4713745 1.549656 1.010515
Sample ATE for RCT:  -0.46

RWD statistics:
               Younger    Older   Average
Placebo    1.19698075 1.535119 1.4582692
Treatment -0.08857169 1.291250 0.1413986
Sample ATE for RWD:  -1.317
```

Treatment selection bias

- The drug is approved and has some mild side effects. Clinicians hence prescribe it mostly to younger patients who are more likely to benefit.
- Real-World Data (RWD) are collected from the electronic health record of 100 adults. The sample ATE in that case is twice that of the RCT.
- In reality, what is happening is that younger patients are over-represented among the treated. Since they do better than older patients, the treated group appears to benefit even more.

```
RCT statistics:
              Younger    Older  Average
Placebo    1.4942583 1.447421 1.470840
Treatment 0.4713745 1.549656 1.010515
Sample ATE for RCT:  -0.46

RWD statistics:
               Younger    Older   Average
Placebo    1.19698075 1.535119 1.4582692
Treatment -0.08857169 1.291250 0.1413986
Sample ATE for RWD:  -1.317

RCT treatment counts:
          Younger Older Total
Placebo        10    10    20
Treatment      10    10    20

RWD treatment counts:
          Younger Older Total
Placebo         5    17    22
Treatment      15     3    18
```

Treatment selection bias

- The fact that age is interacting with the treatment is not the source of the problem here.
- Assume instead that the drug did not have an interaction with age, but that younger patients just did better overall, both on and off treatment (as is often the case), by way of an additive effect.
- The RWD sample ATE is still biased upwards.

```
RCT statistics:
             Younger     Older   Average
Placebo   1.5852680 1.4949430 1.5401055
Treatment 0.9050052 0.9873792 0.9461922
Sample ATE for RCT:  -0.594

RWD statistics:
             Younger    Older Average
Placebo   1.1166439 1.265091 1.23282
Treatment 0.2879676 1.025789 0.37477
Sample ATE for RWD:  -0.858
```

Propensity to treat

- It is easy to reason about this problem as a standard *non-representative sample* problem: the observed treated population is not representative of the theoretical treated population (which would have included all patients, young and old).
- An easy “trick” to correct for non-representative samples in the sample survey literature is to weight each datapoint by the inverse of the probability of its treatment indicator
- In this case we don’t know that probability, but we can estimate it via another model.

$$\hat{p}_i = \hat{P}(T = t_i \mid X = x_i)$$

$$\hat{\mathrm{ATE}} = \left(\frac{1}{\#\{i : t_i = 1\}} \sum_{i \in \{i : t_i = 1\}}^{n} \frac{y_i}{\hat{p}_i} \right) - \left(\frac{1}{\#\{i : t_i = 0\}} \sum_{i \in \{i : t_i = 0\}}^{n} \frac{y_i}{1 - \hat{p}_i} \right)$$

Inverse propensity to treat weighting - IPTW

```
## Create propensity model
propensity_model = glm(treatment ~ age, data=Drwd, family='binomial')
```

Inverse propensity to treat weighting - IPTW

```
## Create propensity model
propensity_model = glm(treatment ~ age, data=Drwd, family='binomial')

## Predict probabilities of getting treated for each patient
probas = predict(propensity_model,newdata = Drwd, type='response')
```

Inverse propensity to treat weighting - IPTW

```
## Create propensity model
propensity_model = glm(treatment ~ age, data=Drwd, family='binomial')

## Predict probabilities of getting treated for each patient
probas = predict(propensity_model,newdata = Drwd, type='response')

## For convenience create boolean index of treated
ind_treated = Drwd$treatment==1
```

Inverse propensity to treat weighting - IPTW

```
## Create propensity model
propensity_model = glm(treatment ~ age, data=Drwd, family='binomial')

## Predict probabilities of getting treated for each patient
probas = predict(propensity_model,newdata = Drwd, type='response')

## For convenience create boolean index of treated
ind_treated = Drwd$treatment==1

## Compute IPTW ATE (difference of weighted means per group)
iptw_ate = weighted.mean(Drwd$y[ind_treated],1/probas[ind_treated])-
  weighted.mean(Drwd$y[!ind_treated],1/(1-probas[!ind_treated]))
cat('IPTW-adjusted sample ATE: ', round(iptw_ate,2), '\n')
```

Inverse propensity to treat weighting - IPTW

```r
## Create propensity model
propensity_model = glm(treatment ~ age, data=Drwd, family='binomial')

## Predict probabilities of getting treated for each patient
probas = predict(propensity_model,newdata = Drwd, type='response')

## For convenience create boolean index of treated
ind_treated = Drwd$treatment==1

## Compute IPTW ATE (difference of weighted means per group)
iptw_ate = weighted.mean(Drwd$y[ind_treated],1/probas[ind_treated])-
  weighted.mean(Drwd$y[!ind_treated],1/(1-probas[!ind_treated]))
cat('IPTW-adjusted sample ATE: ', round(iptw_ate,2), '\n')
```

```
RCT statistics:
            Younger     Older   Average
Placebo   1.5852680 1.4949430 1.5401055
Treatment 0.9050052 0.9873792 0.9461922
Sample ATE for RCT:  -0.594

RWD statistics:
            Younger    Older Average
Placebo   1.1166439 1.265091 1.23282
Treatment 0.2879676 1.025789 0.37477
Sample ATE for RWD:  -0.858
```

```
IPTW-adjusted sample ATE:  -0.6
```

Knowing when IPTW is enough is very hard

- In the presence of an interaction effect (our original example), the IPTW adjustment does not seem to suffice to completely correct for treatment bias, although it does help a little bit.
- Understanding under which conditions IPTW might be sufficient is itself difficult. We will study these conditions in the next lecture.

```
RCT statistics:
             Younger    Older  Average
Placebo   1.4942583 1.447421 1.470840
Treatment 0.4713745 1.549656 1.010515
Sample ATE for RCT:  -0.46

RWD statistics:
              Younger    Older   Average
Placebo    1.19698075 1.535119 1.4582692
Treatment -0.08857169 1.291250 0.1413986
Sample ATE for RWD:  -1.317
```

```
IPTW-adjusted sample ATE:  -0.57
```

Conclusions

- There are many sources of bias that can introduce systematic error into our estimators.
- No amount of data can reduce the size of a systematic error.
- One common source of error when assessing the effect of any intervention (e.g., a medical treatment, a change of policy, and advertising campaign) is *selection bias*, whereby the intervention may not be applied at random, but rather to a population selected on the basis of criteria that might themselves correlate with the outcome.
- In those cases, the average effect will be biased.
- One possible mitigation is to apply weights to each example according to the probability of having been treated by the intervention, as estimated by a separate model.
- Under certain assumptions this can sufficiently mitigate selection bias, which would allow us to make causal conclusions from non-experimental, observational data.