# Inverse Probability of Treatment Weighting (IPTW)

## Intuition

### Example Part 1

* Assume there is a binary confounder X.
* Suppose that
  + Among subjects with , only 10% will receive the treatment i.e. the value of the propensity score for subjects with is 0.1.
* Suppose that
  + Among subjects with , 80% will receive the treatment i.e. the value of the propensity score for subjects with is 0.8.
* If we focus on the sub-population, we note that only 1 out of every 10 subjects on average would be in the treatment group, compared to 50% in a RCT.
  + If we were going to perform propensity score matching, since there is only a single covariate, every subject would have the same propensity score.
  + The treated subject represents one subject as there was only one subject in the treatment group to start with, but the matched control represents all nine original subjects in the control group.
* Rather than using matched pairs to create balance, we could use all the data in both groups, but change the weights of individual subjects (i.e. down-weight some and up-weight others).
  + The treated subject should be weighted 9x more than any single control subject.

### Weighting

* Accomplished by weighting by the inverse of the probability of treatment received.
  + Treatment group: weight by the inverse of
  + Control group: weight by the inverse of i.e. not getting the treatment.
* i.e. weighting by the inverse of the treatment actually received

### Example Part 2

* weight for the treated subject when :
* weight for each control subject when :
* Count the collection of treated subjects the same as the collection of control subjects, which means that among the subjects with the same value of the propensity score, the treated and control groups end up collectively contributing equally to the analysis even though the raw sample sizes are uneven.
* Now if we look at the other sub-population of subjects with , where the chance of getting treated was 0.8.
  + 1 subject in the control group counts the same as 4 people from the treatment group with propensity score matching.
  + weight for the treated subject when :
  + weight for each control subject when :
* Weighting essentially accomplishes the same thing as matching as one subject in the control group counts the same as four subjects in the treatment group.

## More Intuition

### Motivation

* In surveys, common to oversample some groups relative to the population.
* To estimate a population mean, need to weight the data to account for the oversampling – Horvitz-Thompson estimation.
* Oversampling is fine if you know the true proportions of each group and can use sampling weights to return to the original population proportions.

### Observational Studies

* Certain groups are typically ‘oversampled’ relative to the hypothetical sample from a randomised trial.
  + At a particular covariate value, you may have more subjects in the treated group or control group.
  + Hence, could use ‘weighting’ to get back to the original population.
  + IPTW creates a pseudo-population where treatment assignment no longer depends on X, and hence removes the confounding.

### Pseudo-Population

* Suppose there are 9 treated subjects and 1 control subject.
* Apply weighting to balance the population: vs
* End up with 10 subjects in both the treated and control groups:
  + In the treated group, each subject counts as of a subject so .
  + In the control group, each subject counts as 10 subjects so
* In the original population, some subjects were more likely to get treated than others based on their covariates.
* In the pseudo-population, everyone is equally likely to be treated, irrespective of their covariates.
  + Closer to the case with the RCT.

### Estimator

* Under the assumptions of exchangeability and positivity:

Where is the propensity score and is the expected value of treatment i.e. the average value of Y if all subjects had been treated.

* would be estimated in a similar manner.
* is the sample size.
* is a dummy indicator that takes a value of 1 if the subject was treated and 0 otherwise. Fancy way to say you’re only interested in the treated subjects.
* If there wasn’t any confounding, the estimate of Y amongst the treated subjects would just be the sample mean.
* But since there’s confounding, we need to calculate the sample mean of the pseudo-population and not the original population:
  + Numerator is the sum of the Y’s in the treated pseudo-population.
  + Denominator is the number of subjects in the treated pseudo-population.
* End up with a valid estimate of the mean of the potential outcome, assuming that the typical causal assumptions are met:
  + Exchangability/ignorability – the covariates fully capture the confounding.
    - Treatment assignment is random given the covariates.
  + Positivity – the propensity score is going to be strictly between 0 and 1 exclusive of those upper and lower bounds.
  + From a technical point of view – the positivity assumption is important to avoid a divide-by-zero error in the estimation of the potential outcome.

## Marginal Structural Models

* Thus far, IPTW has been discussed in the context of estimating simple causal effects, such as an average causal effect. But IPTW estimation can be used more generally to estimate causal effect parameters from models.
* Model for the mean of the potential outcomes.
* Marginal: model not conditional on the confounders (population average) not something conditional on the covariates, given some sub-population etc.
  + Averaging over some population.
* Structural: model for potential, not observed, outcomes.

### Linear MSM

* Model is for a potential outcome and some a, and it’s linear in itself.
* So, is the average causal effect i.e. the difference in potential outcomes.
* Typically use this kind of model with a continuous outcome and looks like a linear regression except that we’re modelling potential outcomes instead of observed outcomes.

### Logistic MSM

* Binary outcome
* If you take the mean of a binary outcome, that’s just a probability and in fact, it’s
* is the causal odds ratio: , assuming everyone had been given the treatment/control for the numerator/denominator respectively.

### MSM with Effect Modification

* MSMs can also include effect modifiers.
* Also known as heterogeneity of treatment effects i.e. considering the effect of treatment effects varying across sub-populations of interest.
* Suppose V is a variable modifying the effect of A e.g. a comorbidity or a demographic characteristic.
* Linear MSM with effect modification:
  + Conditioning on V, and not all the confounders.
  + Most confounders you just want to control for/marginalise. There may be a small subset that you’re interested in observing how the treatment effect varies as they vary.
  + All the other terms cancel out in the differencing of the means.
* If the MSM was correct, then you’d know the estimated causal/treatment effect given a value for V.

### General MSM

* is a link function.
* is a function specifying a parametric form of and (typically linear).
* The key is that we’re modelling potential outcomes, so if we’re able to estimate the s, we have estimates of causal effects.
* But since potential outcomes aren’t the same as observed outcomes, it’s not quite as simple as just estimating a linear regression model i.e. the LHS is a potential outcome and the RHS is observed data.

## IPTW Estimation

### Estimation in Regression Models

* Consider the estimation of parameters from a linear regression model:
* Involves solving for
* Value of that minimises the sum of squared deviations (least squares estimator).

### Estimation in Generalised Linear Models

* Consider the estimation of parameters from a generalised linear model:
  + is the inverse of the link function.
  + For example, in a logistic regression model, is the inverse of the logit function.
* Involves solving for .

### Estimation in MSMs

* MSMs look a lot like GLMs except that we’re modelling the mean potential outcome instead of the mean of observed data e.g.
* But it’s NOT equivalent to the regression model due to confounding.
  + In a regression mode, we’re conditioning on the observed random variable treatment, A i.e. we’re restricting to the sub-population of subjects that received treatment.
* In an MSM, we’re setting A i.e. we can attach whatever value we want/make treatment whatever we want.
* In general, setting and conditioning on aren’t the same due to confounding.
* In an RCT, fitting the regression model is fine because there isn’t any confounding and hence the parameters represent a causal effect, but otherwise it’s not fine.
* But recall that the pseudo-population created with IPTW is free from confounding (assuming ignorability and positivity), which means we can estimate MSM parameters by solving estimating equations for the observed data of the pseudo-population:
  + Apply the weights to the observed population to remove the confounding in the pseudo-population, which means that you can then use the standard estimating equations as per a GLM.
  + where
    - Key addition here is the , which are the weights or 1 over the probability of the observed treatment.
    - Rather than analyse the original outcome data, we’re analysing the pseudo-population outcome data, which means we can use standard regression to estimate causal effects.
    - Note that A is binary i.e. takes a value of 0 or 1.
* Steps
  + Estimate propensity score – the probability of treatment given the confounders.
    - Calculate the predicted probabilities:
    - Treated subjects: probability of receiving treatment.
    - Control subjects: probability of not receiving treatment.
  + Create weights:
    - 1 divided by the propensity score for treated subjects.
    - 1 divided by 1 minus the propensity score (1 divided by the probability of not receiving treatment) for control subjects.
  + Specify the MSM of interest e.g. a standard MSM with/with no effect modification, a continuous/binary outcome variable.
  + Fit a weighted GLM.
  + Use asymptotic (sandwich) variance estimator (or bootstrapping).
    - Weighting can artificially inflate the pseudo-population sample size, so we need to account for that.
    - i.e. need to account for weighting when estimating the variances.

## Assessing Balance

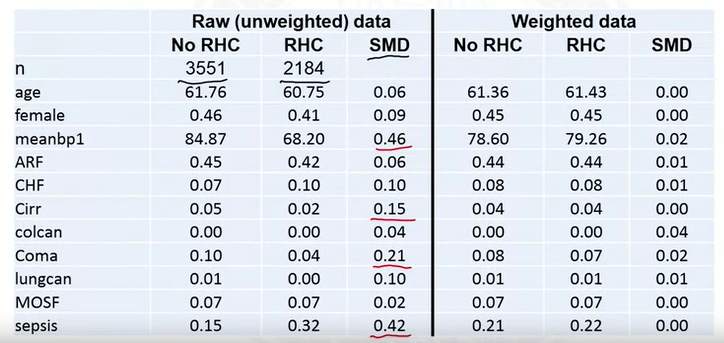
### Balance After Weighting

* Need to check if there is balance in the covariate distribution in the same way you’d expect to have balance in a RCT.
* Covariate balance can be checked on the weighted sample using standardised differences using things like plots or a table 1.

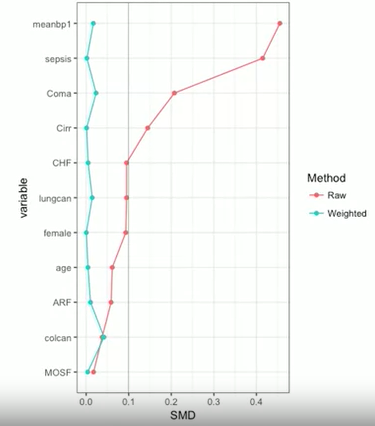
### Standardised Differences After Weighting

* Same idea as previously, except using weighted means and variances.
  + Stratify on treatment group (for each covariate)
    - Calculate weighted means and variances for each group.
  + Take difference in weighted means and divide by an estimate of the pooled (weighted) standard deviations.
* The goal is to have low standardised differences following the weighting.

### Table 1

* 
* Weighted data refers to the pseudo-population.
* Notice how all the SMDs in the weighted data are less than 0.1, indicating good balance between the treatment and control groups.

### SMD Plot

* 
* Common output for either matching or IPTW.
* In this example, sorted by the degree of imbalance in the raw data.

### If Imbalance After Weighting

* Refine propensity score model e.g. interactions, non-linearity.
  + Haven’t considered the outcome at all, so wouldn’t be cheating to go back and modify the model.
  + Need to ensure balance is as good as it can be in the covariates.
  + Iterate back and forth until this is achieved e.g. model and/or variable selection.

## Distribution of Weights

### Why Do Weights Matter?

* Large weights are problematic in that they can lead to large standard errors.
* Larger weights lead to noisier estimates of causal effects.
  + Suppose 1 subject has a weight of 10,000 i.e. it essentially represents 10,000 other subjects.
  + If the outcome is binary, whether the subject has the event or not (yes/no) could have a large impact on the parameter estimate. Hence, a large standard error.
    - Highly variable estimate = large standard error.
* Prefer a single subject not to have a large weight relative to all the other subjects.

### Bootstrapping

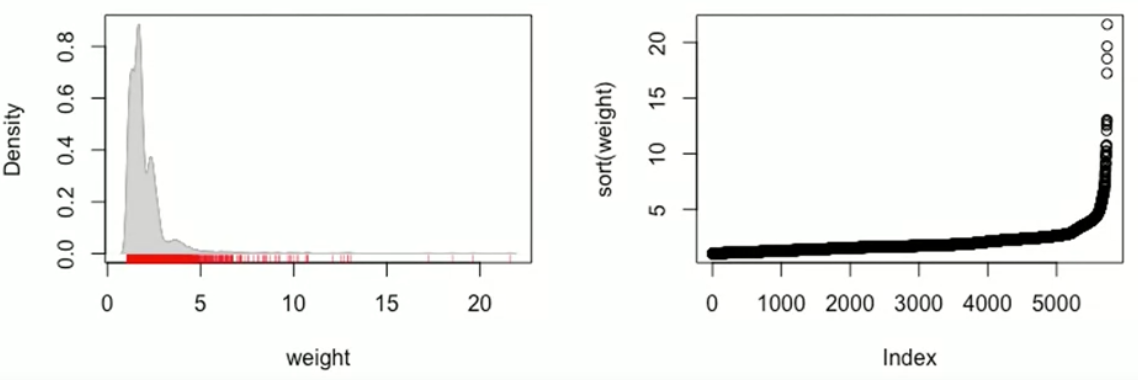
* One way to estimate standard errors is bootstrapping:
  + Randomly sample, with replacement, from the original sample.
  + Estimate parameters.
  + Repeat the first two steps many times.
  + Standard deviation of the bootstraps estimates is an estimate of the standard error.
* One subject with a very large weight included in some samples but not others.
* Whether or not they’re included will have a significant impact on the parameter estimates.
  + If they’re in, will pull the parameter estimate towards them. If they’re in, this won’t occur.
* Hence, much of the variability in the estimator will be due to this subject.

### Relationship with Positivity Assumption

* Very large weights imply the probability of treatment was very small, which indicate near violations of the positivity assumption. i.e subjects with particular covariate values are very unlikely to get one of the treatments.

### Checking Weights

* One of the simplest methods is to just plot them.

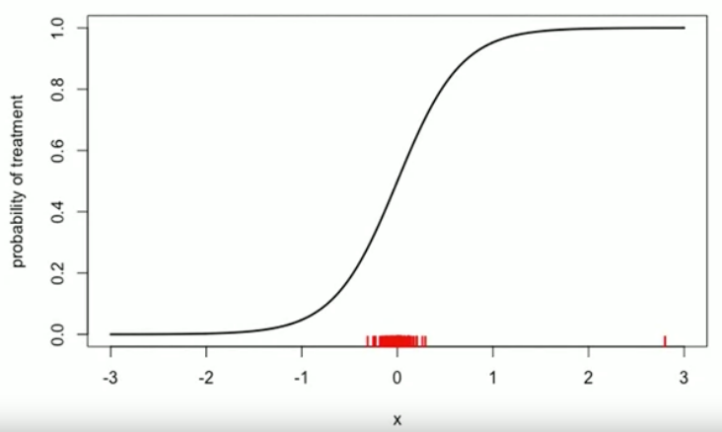


* LHS is a density plot where the rug marks indicate the actual weights.
  + Most weights between 0 and 10, with some larger ones.
  + An upper limit of about 25, which isn’t so large to be a reason for alarm in this case.
* RHS is a sorted rank plot.
  + Ranks on the x-axis and weight values on the y-axis.
  + Can easily see when the values get large and how many there are.
  + Only 4 subjects above a value of 15.
* Could also calculate summary statistics, plus head and tail of distribution.

## Remedies for Large Weights

### Investigation

* A good first step is to understand why the weights are large.
* Identify the subjects with large weights.
  + What’s unusual about those subjects? e.g. is there a variable with an extreme value, or a collection of variables with ‘outlying’ values.
  + Is there an issue with the data? i.e is there a data error?
  + Is there a problem with the propensity score model?
* Could actually be nothing wrong with any of the above and the subject just happens to have a large weight.
* Suppose there is a single confounder and a logistic propensity score model is fitted:



* + Red tick marks correspond to the actual observations.
  + X-axis is the actual covariate.
  + Almost all the data lies in a narrow range about 0, but there’s a single very large outlier.
  + Model indicates the outlier should be near 100% certain to be treated, and if they weren’t, they would have a very large weight.
  + But also don’t really know what the true propensity score curve should look like outside the narrow range of most of the data given the lack of information about the relationship between the probability of treatment and the covariate i.e. lots of extrapolation going on.
    - Curve defined by the assumption that the relationship is linear on the logit scale, but don’t know if it’s actually true or not.
    - Alternative curves where the tails only go to 0.2 and 0.8 respectively (for example) are just as plausible.
  + High weight could be due to a bad model assumption, incorrect data entry etc.
    - The latter should hopefully have been taken care of through data cleaning etc.

### Trimming the Tails

* May want to trim the tails based on the propensity score distribution prior to any IPTW weighting.
  + Helps make the positivity assumption more plausible because it’s removing the extreme values (of the propensity score) and leaving just the subjects who had a reasonable change of getting treatment.
  + In other words, trimming the tails means removing data points.
* Common trimming strategy:
  + Remove treated subjects (+ control if necessary) whose propensity scores are above the 98th percentile of the control distribution.
  + Remove control subjects (+ treated if necessary) whose propensity scores are below the 2nd percentile of the treated distribution.
  + Recall that treated subjects are likely to have higher propensity scores than control subjects on average, and vice versa for control subjects vs treated subjects.
  + The bounds are somewhat arbitrary and can be adjusted.
* Most of the most extreme weights should be eliminated because they occur in the tails.
* Remember that this will change the population of interest i.e. the population we are making inferences about i.e. the sub-population that had a reasonable chance of receiving either treatment where ‘reasonable’ is defined by the trimming strategy.

### Weight Truncation

* Alternative/addition to tail trimming.
* Steps
  + Determine a maximum allowable weight based on a specific value/percentile etc.
  + If the weight is greater than the maximum allowable, set it to the maximum.
* Involves a classic bias-variance trade-off:
  + Truncation: biased (not properly weighting), but smaller variance.
    - Hopefully introduces only a small amount of bias but greatly reduces the variance.
  + No truncation: unbiased but larger variance i.e. noisy estimators.
* Most of the time, whether the trade-off was worth is based on MSE (bias squared + variance).
  + Has been proven in some simulation studies that truncation results in estimators with lower MSE.
  + Too much truncation can result in extreme bias.
* Intention with this technique is just to remove the most extreme values.

## Doubly Robust Estimators (Augmented IPTW)

### Background – IPTW Estimation

* The equation for (the expected value of the potential outcome among control subject) looks very similar except the weights are different.
* If the propensity score model is correctly specified, this estimator is unbiased. i.e. if the true probability of treatment is roughly equivalent to the estimated probability of treatment.
* Recall is the propensity score for subject and is a binary variable indicating if subject had actually received treatment or not.

### Background – Regression-based Estimation

* Could also estimate using an outcome regression model before averaging over the distribution of :
  + is a conditional mean, so you need to average over the distribution of the confounders to get the mean of the potential outcomes.
  + For treated subjects with , use the observed (by the consistency assumption if , )
  + For controls, essentially predict what their outcome would have been had they been treated i.e. use a predicted .
* If the outcome model is correctly specified then this estimator is unbiased.

### Doubly Robust Estimators

* Essentially attempt to use both approaches.
* Unbiased estimator if either the propensity score model or the outcome regression model are correctly specified.
* For example:
  + The first part looks like the standard IPTW estimator.
  + The second part is the ‘augmentation’ and involves a regression model.
  + Together they produce an estimator with this double-robustness property.
* Consider the situation in which only the propensity score model is correctly specified:
  + is ‘wrong’ i.e. where is a model such that (the expectation) doesn’t line up with .
  + or .
  + Outside the curly brackets is a sum divided by , which is a sample average that approaches an expectation as the sample size increases. And we are interested in whether the inside expectation is equal to .
    - The expectation of is equal to the propensity score, which means the entire second half of the equation has expectation 0 i.e. approach 0.
    - This leaves the first part, which is just the IPTW estimator and we’ve already stated that this is a valid estimator of given the propensity score is correctly specified.
* Consider the situation in which only the outcome model is correctly specified:
  + Note that these are models, so we don’t know for sure if either one is correct, but it would be nice if we could have robustness and allow for one to be wrong.
  + but
    - Just rearranged terms via algebraic manipulation.
  + goes to 0 because and the other terms are essentially constants that won’t force the entire fraction not to approach 0.
  + The part that is left is just the average of the regression model over all subjects, which again is just the .
* Can use semi-parametric theory to identify the best estimators.
* In general, AIPTW estimators should be more efficient than regular IPTW estimators i.e they have a smaller associated variance and hence tend to perform better.
* However, they are more complicated to implement in practice.