Potential Outcomes & Treatment Effect

An Observational Study is an empirical investigation in which "the objective is to elucidate cause & effect relationships" in which it is not feasible to use controlled experimentation, e.g. assigning people to smoking.

Consider a treatment Z, e.g. $Z \in K = \{T, C\}$
T is treatment (e.g. smoking) C is control (e.g. non smoking)

For unit i, the Potential response Y_{Ti} is it's outcome if he received T, e.g. $Y_{T1} = 1$ if I would get lung cancer and he/she smoked. Likewise, Y_{Ci} is it's outcome if he received C.

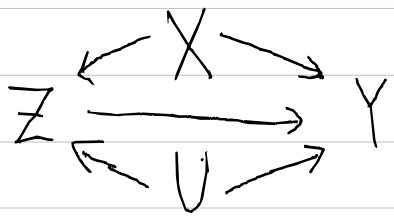
We are interested in $Y_{Ti} - Y_{Ci}$, the causal effect of T on the outcome. However, this is impossible, as only one can be observed. The unobserved outcome is called Counterfactual. So instead, we imagine selecting a patient at random, and taking expected value, to get the Average treatment effect $ATE = E[Y_T - Y_C] = E[Y_T] - E[Y_C]$.

Causal Assumptions

Independence: The treatment actually assigned must be independent of the person's potential outcomes under treatment: $E[Y_k] = E[Y_k | Z = \ell] \quad \forall k, \ell \in K = \{T, C\}$

SUTVA: Stable Unit Treatment Value Assumption, this means Y_{iz_i} varies with Z_i but not with any Z_j, "no interference between units"

No Unmeasured Confounders: To explain this assumption, we invoke the DAG (Directed Acyclic Graph), used to show causal relationships, where an arrow indicates causality and no arrow indicates independence. No unmeasured confounders says that if X are measured and might have an effect on Z and the observed outcome $Y = Y_z$, then no unmeasured variable U exists such that:



e.g. no gene exists that makes you more likely to smoke and more likely to get lung cancer

Common ATE Tests

We wish to test the null hypothesis $H_0: Y_{Ti} = Y_{Ci}, \forall i$
Some common tests against this in different settings:

1. Fisher's Exact, if outcome is binary
2. Mantel-Haenszel, Outcome is binary but two or more Strata. If the Strata are matched pairs, it's called McNemar's
3. t-test, Paired or Unpaired, assume normal data
4. Wilcoxon's Rank Sum, Rank observed responses Small to large, Under No effect the Sum of ranks of treated has Known distribution, if $Y_T > Y_C$, Sum Should be large

5. Wilcoxon's Signed Rank, Used for Paired data, rank absolute differences, Sum ranks with $Y_{T_{i_1}} > Y_{C_{i_2}}$
6. Stratified Rank Sum, Wilcoxon for many Strata
7. Aligned Rank, Superior to 6 for large # of Strata
8. Median Test, Count # of treated to exceed within Stratum Median

Often, the exact p-values are difficult to calculate in large samples, so approximations (eg. based on the Normal using CLT) are often used

Treatment Effect Models

Additive treatment effect: The treatment raises or lowers the response by a fixed value ζ for all i, Obviously we cannot check it directly, but we can find the set of ζ's such that some test of the null hypothesis $H_0: Y_{Ti} = Y_{Ci} + \zeta$ is not rejected. This is the confidence set for ζ under this model.

Dilated Treatment effects:

Multiplicative treatment effect: The treatment scales up or down the response by δ, the confidence set is generated from

$H_0: Y_{Ti} = \delta Y_{Ci}$, this is additive on log-scale:

$H_0: \log(Y_{Ti}) = \log(Y_{Ci}) + \log(\delta)$

Linear treatment effect: treat linear model of control:

$H_0: Y_{Ti} = \beta + \delta Y_{Ci}$

Some-Susceptible effect: Treat effect is 0 if $Y_{Ci} < \tilde{Y}$, but exists if $Y_{Ci} > \tilde{Y}$

Matching

Often, in observational data, there exist many measured confounders, that is $Z \xrightarrow{X} Y$, to determine the ATE using previously mentioned methods, we need to address these confounders, because Z is not assigned randomly, and we need to construct a "random" Z, Z that is independent of Y except through the causal relationship $Z \to Y$, by addressing these confounders.

One strategy is matching. The idea is that, assuming no unmeasured confounders, if treated person i's covariates X_i are the same as control person j's covariates X_j, then it is effectively "random" that person j was treated and person i was control instead or vice-versa.

A good matching balances the covariates between the matched treated group and matched control's. Often, this is assisted by a "propensity-score model".

A good matching is also often measured by standardized differences between matched T and C, in covariate X:

$$d = \frac{\bar{X}_T - \bar{X}_C}{\left(\frac{s^2_{X,T} + s^2_{X,C}}{2}\right)}$$

s diffs of >0.2 desired

Propensity Score & IPW

The <u>Propensity Score</u> $\pi(x_i) = P(Z_i = T \mid X = x_i)$ is the probability that i is treated given his covariates X_i. Often matching procedures are based on or involve a propensity score model of some kind, usually a logistic model: $\text{logit}(\pi(x_i)) = X\beta$, easy to fit

The IPW or <u>inverse probability weighting</u> estimator of $E[Y|Z]$ addresses X through $\pi(x)$, the idea is to down weight over represented (high probability of being observed) observations of Y and upweight underrepresented (low probability of being observed) observations of Y so that the resultant estimates of $E[Y_{Ti}]$, $E[Y_{Ci}]$ are unbiased. The resultant estimator of ATE is

$$\hat{\gamma}_{IPW} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Z_i Y_i}{\hat{\pi}(X_i)} - \frac{(1-Z_i)Y_i}{1-\hat{\pi}(X_i)}\right)$$ since

$$E[\hat{\gamma}_{IPW}] = \frac{1}{n}\sum_{i=1}^{n}\frac{\pi(x_i)E[Y_i|Z_i=1,x_i]}{\pi(x_i)} - \frac{(1-\pi(x_i))E[Y_i|Z_i=0,x_i]}{(1-\pi(x_i))} = E[Y|Z=1] - E[Y|Z=0]$$

assuming $E[\hat{\pi}(x_i)] = \pi(x_i)$, the propensity score model is good

Since X_i is given, $E[Z_i|X_i] = \pi(x_i)$

AIPW

The IPW can have high variance due to extreme weights, and requires $\hat{\pi}(x_i)$ to be correctly specified. The AIPW involves an IPW and a second model, and AIPW is consistent if either is correctly specified and is more robust than IPW. Let $\hat{\mu}_1(x_i)$ be a model for Y_i based only on x_i, and $\hat{\mu}_0(x_i)$ be likewise, $\hat{\mu}_1$ for treated, $\hat{\mu}_0$ for control.

$$\hat{\tau}_{AIPW} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{z_i(y_i-\hat{\mu}_1(x_i))}{\hat{\pi}(x_i)} + \hat{\mu}_1(x_i) - \frac{(1-z_i)(y_i-\hat{\mu}_0(x_i))}{\hat{\pi}(x_i)} - \hat{\mu}_0(x_i)\right]$$

Thus, $\hat{\tau}_{AIPW}$ is $\hat{\mu}_1(x_i) - \hat{\mu}_0(x_i)$ except that the residuals $(y_i - \hat{\mu}_1(x_i))$ and $(y_i - \hat{\mu}_0(x_i))$ inverse weighted by the probability of being treated are added to the model. If both models are accurate, residuals cancel and we get $E[\hat{\mu}_1(x_i) - \hat{\mu}_0(x_i)] = ATE$, if only μ is accurate the residuals will spread "evenly" up and down, since $E[y_i - \hat{\mu}(x_i)]$ terms $= 0$ and result is consistent. Likewise, if only $\hat{\pi}$ is correct, the residual weights "fix" the bias in μ. e.g., if $\hat{\mu}_1(x_i) = c_1$, $\hat{\mu}_0 = c_0$, AIPW can be shown to be just IPW.

Sensitivity Analysis

Often, we are concerned that perhaps there exists a hidden bias, an unmeasured confounder U, that affects both the treatment and the outcome, violating our assumptions. This assumption is untestable, but there are many ways to address the concern. One way is Sensitivity Analysis, which tells us how strong a hidden bias would have to be in order to change our conclusion.

For example: Concerned about gene X that both causes people to smoke and get lung cancer. Sensitivity analysis says that for our test to lose statistical significance the gene would have to be 4 times more present in smokers than nonsmokers

Let i and k be units with same covariates X, one T one C, where $\pi_i \neq \pi_k$ due to a hidden bias U. Imagine we knew the odds ratio for units with matching X is atmost Γ When the test being used is a marked sum, it can be shown that

Sensitivity Analysis

$$\frac{1}{\Gamma} \le \frac{\pi_j(1-\pi_k)}{\pi_k(1-\pi_j)} \le \Gamma \quad \forall j,k \text{ s.t. } x_j = x_k \quad \Rightarrow \frac{1}{1+\Gamma} \le P(Z_i = 1 \mid x_i \text{ aka } x_j) \le \frac{\Gamma}{1+\Gamma}$$

The change in Γ leads to a change in the distribution of the rank-sum null distribution. Suppose we are working with paired data, again T is a ranked sum. Let s be a pair, C_{s1} and C_{s2} be the treatment status of units in the pair, $d_s = Y_{sT}^T - Y_{sC}^C$, define

$$P_s^+ = \begin{cases} 0 & C_{s1} = C_{s2} = 0 \\ 1 & C_{s1} = C_{s2} = 1 \\ \frac{\Gamma}{1+\Gamma} & C_{s1} \neq C_{s2} \end{cases} \qquad P_s^- = \begin{cases} 0 & C_{s1} = C_{s2} = 0 \\ 1 & C_{s1} = C_{s2} = 1 \\ \frac{1}{1+\Gamma} & C_{s1} \neq C_{s2} \end{cases}$$

min/max prob that the pair contributes to the ranked sum. Define T^+ as sum of S random variables where j^{th} r.v. takes on value d_s with probability P_s^+, and T^- as same except j takes on d_s with probability P_s^-. Mean and variance are easily calculated, and can often be normal approximated (R_i's ranks of d_i's):

$$E[T^+] = \sum_{s=1}^{\hat{n}} P_s^+ R_s \xrightarrow{\text{if no } s \text{ concordant}} \sum_{s=1}^{\hat{n}} P_s^+ \frac{\Gamma}{\Gamma+1}$$

$$Var(T^+) = \sum_{s=1}^{\hat{n}} P_s^+(1-P_s^+) R_s^2$$

replace P_s^+ w/ P_s^- to get result for T^-

Sensitivity Interval

A <u>sensitivity interval</u> of Γ for a parameter β is the set of β's for which the null hypotheses for which (say $H_0: \tau_{T_i} = \tau_{C_i} + \beta$ in additive model) cannot be rejected even with sensitivity parameter Γ (a hidden bias could be Γ times more present or $\frac{1}{\Gamma}$ times as present in treatment compared to control) working against the null hypothesis

Analogue of confidence interval when Γ known to have some maximum value. For example:

$$SI_\beta: \left\{ \sup_\beta \left(\frac{T_{\beta,obs} - E[T_\beta^-]}{\sqrt{Var(T_{\beta,\Gamma}^-)}} \right) \leq 1.96 \right\}, \inf_\beta \left(\frac{T_{\beta,obs} - E[T_{\beta,\Gamma}^+]}{\sqrt{Var(T_{\beta,\Gamma}^+)}} \geq -1.96 \right)$$

Using normal approx. for Wilcoxon Rs Sensitivity Analysis.

Nested Hypotheses

Another method for testing while accounting for a specific hidden bias is to use Nested Hypotheses. First, test the treatment group against one control group that you can assume is not subject to one kind of hidden bias. Then, only if the test is significant, test against a different control group that has a different potential hidden bias. If this is also significant, draw the conclusion.

Example: Miners of lead/zinc in a town might be subject to genetic damage from radon exposure. Use two control groups: 1) Non-miners who live near the mine 2) matched/related field workers distant from the mine. The hidden bias controlled by 1) is effect of lead pollution on local environment. The hidden bias controlled by 2) is that of difference in occupation / class that could contribute to increased genetic damage

Difference in Differences

Used to estimate the effect of an intervention by comparing the differences between the treated group and control group before and after the intervention. This is often used to evaluate a sudden policy change.

Compute the treatment effect as
$\delta = (Y_{T,post} - Y_{T,post}) - (Y_{C,post} - Y_{C,pre})$

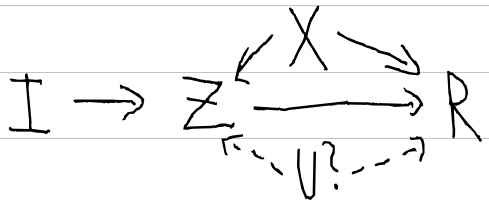
Example: Philadelphia babies, between 1997 and 2007 many hospitals closed obstetric units (12 of 19) with no similar reduction across U.S., did this lead to more newborn deaths? Matched Philadelphia is constructed for Y_C using similar neighborhoods from across U.S.

Again, Y_{it}'s should be matched to Y_{ic}'s so that $\hat{\delta} = (\bar{Y}_{T,post} - \bar{Y}_{T,pre}) - (\bar{Y}_{C,post} - \bar{Y}_{C,pre})$ is not affected by measured confounders

Instrumental Variables

Often, we cannot control the treatment of interest and suspect that there is unmeasured confounding of treatment, but there is another variable I associated with the treatment we know to be independent of Z. If I can be randomized, we call this a <u>randomized encouragement design</u>.

Ex: Randomly encourage some mothers not to smoke, cannot assign to smoking or non-smoke but can assign encouragement against smoking, which will affect treatment and enc. against smoking can be assigned randomly. Let R be baby weight (outcome), $Z = 0, 1$, non-smoke or smoke, X = covariates (e.g. mom age), $I = 1, 0$, encourage or don't



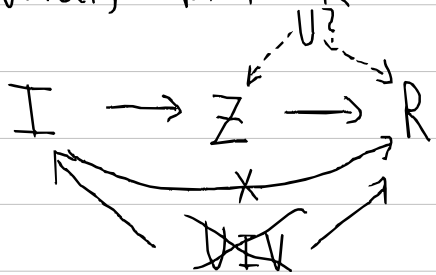
We call I an <u>instrumental variable</u> (IV) and can use IV to eliminate the effect of U

Instrumental Variables

Assumptions needed for a valid IV:

1. Exclusion Restriction: I affects R only through Z

2. Unmeasured Confounding of IV: I has no unmeasured confounders with R



Note in 3. that $Z_i(1)$ is i's potential treatment if encouraged, $Z_i(0)$ is is potential treatment if not encouraged

Sometimes the I.V. is not from an encouragement design, instead some other naturally constructed "random encouragement" can be used as an I.V., eg. living near 2-year or 4-year college

3. NO "defiers" whose potential outcomes are flipped, i.e. encouragement causes them to not be in treated group when they otherwise would be, $Z(1)=0$, $Z(0)=1$.

Other terminology: compliers have $Z(1)=1$, $Z(0)=0$, always takers have $Z(1)=1$, $Z(0)=1$, never takers have $Z(1)=0$, $Z(0)=0$

2SLS

From our I.V. analysis, we wish to estimate $\beta = E[R_i | Z_i = 1, Z_i(1) = 1, Z_i(0) = 0] - E[R_i | Z_i = 0, Z_i(1) = 1, Z_i(0) = 0]$

The effect of treatment on compliers. Since we do not have counterfactuals for always/never takers, no conclusions can be drawn about treatment effect for them.

Two-Stage Least Squares is an unbiased estimate of β, (if component models are correct) and first estimates $E[Z_i | X_i, I_i] = \alpha$ and then uses $\hat{\alpha}$ as a predictor of R_i, and creates a notion model with X_i, $\hat{\alpha}$,

Stage 1: $Z_i = \pi_0 + \pi_1 I_i + \pi_2 X_i + \gamma_i$ for example
$\hat{\alpha}_i = \hat{E}[Z_i | I_i, X_i]$ from this model

Stage 2: $R_i = \beta_0 + \beta_1 \hat{\alpha}_i + \beta_2 X_i + \varepsilon_i$ where our found $\hat{\beta}_1$ is the estimator of interest

$$\hat{\beta}_{TSLS} = \frac{\hat{E}[R_i | I_i = 1] - \hat{E}[R_i | I_i = 0]}{\hat{E}[Z_i | I_i = 1] - \hat{E}[Z_i | I_i = 0]} \xrightarrow{p} \frac{P(Z_i(0) = 0, Z_i(1) = 1)\,\beta}{P(Z_i(0) = 0, Z_i(1) = 1)}$$

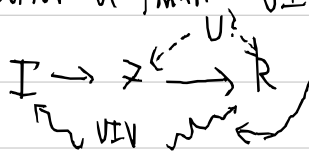
$= \beta$, as before tests exist to form CI's for β

Instrumental Variables

The <u>Prevalence ratio</u> of a binary characteristic X_i among the compliers compared to the full population is $P(X_i=1 \mid Z_i(1)=1, Z_i(0)=0) / P(X_i=1)$, ratio of the prevalence of the trait in the "complier" group compared to the whole population. This helps tell us for what types of units are the IV results legit.

Assuming monotonicity, the prev ratio is $\frac{E[Z_i \mid I_i=1, X=1] - E[Z_i \mid I_i=0, X_i=1]}{E[Z_i \mid I_i=1] - E[Z_i \mid I_i=0]}$

We can "guess" whether the IV is independent of unmeasured confounders by checking if it has a strong relationship with measured confounders. Hopefully, the IV is independent of everything. We also want to check the "strength" of the IV, this is done through Proportion of compliers. Ideally, we would like $P(\text{complier}) \approx 0.4$, a $P(\text{complier}) < 0.2$ is very weak. Weak IVs allow a small UIV confounder to have a massive effect:



very weak connections cause problems if the IV is "weak"

Slightly biased strong IV generally better than a weak IV