CI : RANDOMIZED CONTROL TRIALS

In this chapter, we discuss the estimation of Average Treatment Effects (ATE) in Randomized Control Trials (RCT). RCTs are commonly referred to as the "gold standards" for the estimation of causal effects and represent the foundation of modern statistical causal inference.

• The Potential Outcome (P.O.) framework

For a set of units/individuals $i = 1, ..., n$, we consider a binary treatment assignment $W_i \in \{0,1\}$ and we denote $Y_i(1)$ and $Y_i(0)$ the P.O. that the i-th unit would experience if they had received the treatment $(=1)$ or not $(=0)$.

The CAUSAL EFFECT of the i-th unit is

$$\Delta_i = Y_i(1) - Y_i(0)$$

Throughout this chapter, we assume CONSISTENCY i.e if $W_i = w$, we observe $Y_i = Y_i(w)$ for unit i

$$\Leftrightarrow Y_i = W_i Y_i(1) + (1-W_i) Y_i(0)$$

$$\Leftrightarrow Y_i = Y_i(W_i)$$

x Remark = the notation $Y_i(0), Y_i(1)$ implies that the causal effect Δ_i does not depend on the treatment value of unit $j \neq i$ i.e. there are no interference between units (the Stable Unit Treatment Value

Assumption (SUTVA) (Rubin, 1978). More generally, the observed outcome Y_i is a function of the whole assignment vector $W = (W_1, ..., W_n)$ through the P.O. $Y_i = Y_i(W)$: the outcome of unit i depends on the treatment assignment of all other units in the sample. While the SUTVA assumption is usually satisfied in medical applications, it is often violated in other fields.

x Ex: Suppose that Uber wants to test a new pricing algorithm so that customers in the treatment group are more likely to opt-in for a ride. As there will be fewer drivers on the road, the price for the control group will go up leading to fewer drivers. The difference between the treatment & control is likely to be overestimated ∎

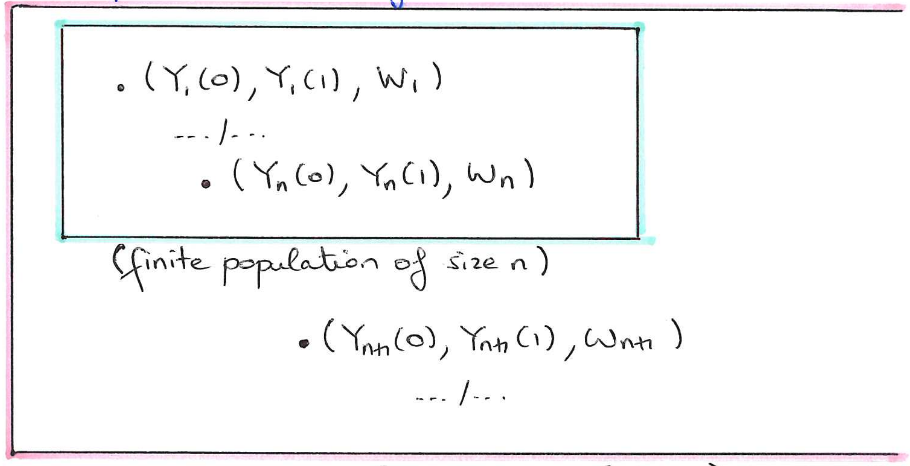
We assume that SUTVA holds throughout this chapter. The general case is discussed in CI : TREATMENT EFFECTS UNDER INTERFERENCE.

Since either one of the P.O. $Y_i(1)$ or $Y_i(0)$ is observed for each unit i, the causal effect Δ_i is not directly observable (the fundamental problem in causal inference) ⇒ we generally estimate aggregate effects, i.e. the average causal effect in a population of individuals.

I_ THE DIFFERENCE ESTIMATOR

I.1. Definition of the ATE estimand

The estimand definition depends on the assumptions of the population model. Either the population is finite (and consists in N individuals, from which $n \leq N$ are sampled from at random – we assume here for simplicity that $n = N$), or the set of n individuals are assumed sampled from an infinite super-population (there are theoretical benefits for assuming so, leading to simpler statistical arguments).


• $(Y_1(0), Y_1(1), W_1)$
…/…
• $(Y_n(0), Y_n(1), W_n)$
(finite population of size n)
• $(Y_{n+1}(0), Y_{n+1}(1), W_{n+1})$
…/…

(∞ - super population)

In the finite population model, the P.O. are generally assumed fixed (as opposed to Random Variables (RVs)). The ATE is defined as

$$\Delta^n := \frac{1}{n} \sum_{i=1}^{n} (Y_i(1) - Y_i(0))$$

no need to generalize the ATE to other individuals

On the other hand, the infinite population model induces a distribution over the P.O. $(Y_i(0), Y_i(1)) \sim \mathbb{P}$.

assumed iid

The ATE is defined as

$$\Delta^\infty := \mathbb{E}\{Y_i(1) - Y_i(0)\}$$

where $\mathbb{E}(.)$ denotes the expectation under $\mathbb{P}$

For convenience, we write $\begin{cases} \mu_j := \mathbb{E}\, Y_i(j) \\ \sigma_j^2 := \operatorname{Var} Y_i(j) \end{cases}$, $j = 0, 1$

This brings us to naturally define the difference estimator as an estimator of the ATE (either Δ^n or Δ^∞ depending on the population model).

$$\hat{\Delta} := \underbrace{\frac{1}{n_1} \sum_{i=1}^{n} W_i \overset{Y_i(1)}{\overset{\shortparallel}{Y_i}}}_{=: \overline{Y}^{(1)}} - \underbrace{\frac{1}{n_0} \sum_{i=1}^{n} (1-W_i) \overset{Y_i(0)}{\overset{\shortparallel}{Y_i}}}_{=: \overline{Y}^{(0)}}$$

where n_j = number of units in the treatment ($j=1$) and control groups ($j=0$).

× Remark : In the finite population model, the only source of randomness is the treatment assignment W_i since $Y_i := W_i \underbrace{Y_i(1) + (1-W_i) Y_i(0)}_{\text{fixed}}$

⇒ we talk about "design-based" uncertainty. In the infinite population model, the P.O. are RVs and contribute to the variance of $\hat{\Delta}$ ⇒ we talk about "population-based" uncertainty. We thus expect that the variance of $\hat{\Delta}$ is larger when assuming an infinite population model.

I.2. The Random Assignment

We consider a completely randomized experiment, where a fixed number of n_1 units are drawn at random from the set of n units and allocated to a treatment group, while the remaining $n_0 := n - n_1$ are allocated to the control group:

$$\boxed{\begin{array}{c}\mathbb{P}(W = w \mid Y(0), Y(1)) = 1 / \binom{n}{n_1} \\ \forall w \in \mathbb{R}^n \text{ s.t. } \sum_{i=1}^{n} w_i = n_1\end{array}}$$

"units are allocated to trt & ctl completely at random".

↳ Other assignment mechanisms (such as a stratified or paired random assignment) are discussed in Athey & Imbens (2016).

I.3. Moments of $\hat{\Delta}$

In this section, we derive the first and second moments of $\hat{\Delta}$ under (a) SUTVA and (b) a treatment assignment completly at random. We consider both the finite population model [$\mathbb{E}(.)$ under W, conditionally on the P.O.] and the infinite population model [$\mathbb{E}(.)$ under the P.O. distribution, conditionally or not on W].

Theorem 1: First moment of $\hat{\Delta}$

- [finite pop] $\quad \mathbb{E}(\hat{\Delta} \mid Y(0), Y(1)) = \Delta^n$
- [∞ pop] $\quad \mathbb{E}(\hat{\Delta} \mid W \qquad) = \Delta^\infty$

proof: The case [∞ pop] is trivial, and we focus on the [finite pop] case.

Note that $\bar{Y}^{(1)} = \frac{1}{n_1} \sum W_i Y_i = \frac{1}{n_1} \sum w_i Y_i(1)$

$$\Rightarrow \mathbb{E}\bar{Y}^{(1)} = \sum_{j=1}^{\binom{n}{n_1}} \bar{Y}_j(1)\, \mathbb{P}(W = w_j)$$

where → $w_j \in \mathbb{R}^n$ contains exactly n_1 "1" values

→ $\mathbb{P}(W = w_j) = 1/\binom{n}{n_1}$ "random assignt"

→ $\bar{Y}_j(1) := \frac{1}{n_1} \sum_{i=1}^{n} w_{ji} Y_i(1)$

Thus

$$\mathbb{E}\bar{Y}^{(1)} = \frac{1}{n_1} \frac{1}{\binom{n}{n_1}} \sum_{j=1}^{\binom{n}{n_1}} \sum_{i=1}^{n} w_{ji} Y_i(1)$$

$$= \frac{1}{n_1} \frac{1}{\binom{n}{n_1}} \sum_{i=1}^{n} Y_i(1) \underbrace{\sum_{j=1}^{\binom{n}{n_1}} w_{ji}}$$

we need to count how many times $w_{ji} = 1$ for each i. Whenever $w_{ji} = 1$, there remains $(n-1)$ entries available for the rest of the sample, and $(n_1 - 1)$ trt units to fill in → the numbe we are looking at is $\binom{n-1}{n_1-1}$

After simplification, we get $\mathbb{E}\bar{Y}^{(1)} = \frac{1}{n} \sum_{i=1}^{n} Y_i(1)$, and similarly for $\mathbb{E}\bar{Y}^{(0)}$. ■

Theorem 2 : Second moment of $\hat{\Delta}$

- [finite pop] $$\mathrm{Var}(\hat{\Delta} \mid Y(0), Y(1)) = \frac{S_0^2}{n_0} + \frac{S_1^2}{n_1} - \frac{S_{01}^2}{n_0+n_1}$$

where $$S_j^2 := \frac{1}{n-1} \sum_{i=1}^{n} (Y_i(j) - \bar{Y}(j))^2$$

$$S_{01}^2 := \frac{1}{n-1} \sum_{i=1}^{n} \left\{ Y_i(1) - Y_i(0) - (\bar{Y}(1) - \bar{Y}(0)) \right\}^2$$

- [∞ pop] $$\mathrm{Var}(\hat{\Delta} \mid \omega) = \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}$$

proof: The [∞ pop] case is trivial. The [finite pop] case is more tedious to show, but we can make use of Theorems 2.2 and 2.3, Section 2.5 in Cochran (1977) to derive the result. See also Neyman (1923, 1990) who originally derived the variance of the difference estimator in the finite population model. ∎

× Remarks (i) Alternative Expression for S_{01}^2.

We can easily show that

$S_{01}^2 = S_0^2 + S_1^2 - 2\rho_{01} S_0 S_1$, where

$$\rho_{01} = \left(\frac{1}{n-1} \sum_{i=1}^{n} (Y_i(0) - \bar{Y}(0))(Y_i(1) - \bar{Y}(1)) \right) / S_0 S_1$$

$= \text{correlation coefficient} \in [-1, 1]$

Therefore

$$\mathrm{Var}(\hat{\Delta} \mid Y(0), Y(1)) = \frac{n_1}{n n_0} S_0^2 + \frac{n_0}{n n_1} S_1^2 + \frac{2}{n} \rho_{01} S_0 S_1$$

$n = n_0 + n_1$

We readily see that this expression is maximized when $\rho_{01} = +1$, in which case

$$\mathrm{Var}(\hat{\Delta} \mid Y(0), Y(1) ; \rho_{01} = +1) = \frac{S_0^2}{n_0} + \frac{S_1^2}{n_1} - \frac{(S_0 - S_1)^2}{n}$$

This situation occurs when the P.O. are perfectly positively correlated. Likewise, the variance is minimum when the P.O. are perfectly negatively correlated.

A special case occurs when the treatment effect is constant across units and additive: $Y_i(0) - Y_i(1) = \Delta \quad \forall i$, in which case

$$\mathrm{Var}(\hat{\Delta} \mid Y(0) - Y(1) = \text{constant}, Y(0), Y(1)) = \frac{S_0^2}{n_0} + \frac{S_1^2}{n_1}$$

since $S_0^2 = S_1^2$ in this case and the third term $(S_0 - S_1)^2$ vanishes.

In particular, this expression holds trues when there is exactly no effect across units $Y_i(0) = Y_i(1)$ (the SHARP NULL of Fisher)

For a complete discussion on the behaviour of $\mathrm{Var}(\hat{\Delta} \mid Y(0), Y(1))$ under various scenario, see Reichardt & Gollob (1999).

(ii) Estimating the variance in the finite pop model

It can be shown that

$$\hat{s}_j^2 := \frac{1}{n_j - 1} \sum_{i|W_i=j} \left(Y_i - \overline{Y}^{(j)} \right)^2 \qquad j=0,1$$

is an unbiased estimator for s_j^2. On the other hand, the third term s_{01}^2 is generally impossible to estimate empirically since we never observe both $Y_i(0)$ and $Y_i(1)$ at the same time. However, under the assumption that the treatment effect $Y_i(1) - Y_i(0)$ is contant across units, an unbiased estimator of the variance is

$$\hat{v}_{Neyman} := \frac{\hat{s}_0^2}{n_0} + \frac{\hat{s}_1^2}{n_1}$$

(see remark (i) above)

Using $\hat{v}_{Neyman}$ leads to confidence intervals with coverage at least as large as the nominal coverage. Alternatively, we can make use of the worst case scenario ($\rho_{01}=1$) described in remark (i) and consider

$$\hat{v}_{\rho=1} := \frac{\hat{s}_0^2}{n_0} + \frac{\hat{s}_1^2}{n_1} - \frac{(s_0 - s_1)^2}{n}$$

without risks of underestimating the variance.

(iii) Estimating the variance in the ∞ pop model

It is straightforward to see that $\hat{s}_j^2$ (page 9) is unbiased for σ_j^2. Therefore, $\hat{v}_{Neyman}$ is always an unbiased estimator of the variance in the ∞-population model.

A property that is not shared with $\hat{v}_{\rho=1}$. For this reason, $\hat{v}_{Neyman}$ remains the most popular estimator of the variance in both the finite & infinite population models.

(iv)

The first and second moments of $\hat{\Delta}$ in the ∞-population model remain unchanged when removing the conditioning on W. In addition, we may use bootstrap resamples to get an approximation of the $\hat{\Delta}$ distribution, as an alternative to the mean and variance expressions derived above. These will yield similar results however, as a CLT holds for $\hat{\Delta}$:

$$\sqrt{n}\left(\hat{\Delta} - \Delta^{\infty}\right) \xrightarrow{d} \mathcal{N}(0, V)$$

where $V = \dfrac{\sigma_0^2}{P(W=0)} + \dfrac{\sigma_1^2}{P(W=1)}$,

$$\frac{n_j}{n} \xrightarrow{P} \mathbb{P}(W=j).$$

As an alternative to the bootstrap, we may consider permutation distributions to characterize the variability in the difference estimator. We discuss this next.

I.4. Randomization Tests.

Randomization tests arise naturally in the finite population model, where the only source of randomness in

$$\hat\Delta = \frac{1}{n_1}\sum_{i=1}^{n} W_i \underbrace{\boxed{Y_i(1)}}_{\text{fixed}} - \frac{1}{n_0}\sum_{i=1}^{n}(1-W_i)\underbrace{\boxed{Y_i(0)}}_{\text{fixed}} \text{ is } W \sim \mathbb{P}(W=w) = 1/\binom{n}{n_1}$$

If both $Y_i(0)$ and $Y_i(1)$ were observed $\forall i$, we could consider all $b = 1, \dots, \binom{n}{n_1}$ possible treatment allocations $w^{(b)} = (w_1^{(b)}, \dots, w_n^{(b)})$; $\sum_{i=1}^{n} w_i^{(b)} = n_1$, and compute for each b the mean difference

$$\hat\Delta^{(b)} = \frac{1}{n_1}\sum_{i=1}^{n} w_i^{(b)} Y_i(1) - \frac{1}{n_0}\sum_{i=1}^{n}(1-w_i^{(b)}) Y_i(0)$$

$\hat\Delta^{(b)}$ may not be computed in general, but there is one notable exception where we can infer all missing P.O.: when the treatment has exactly no effect on each of the units; i.e. when $H_0: Y_i(0) = Y_i(1)$ holds true. Under Ho, we can compute the "randomization distribution" $\mathbb{P}_w(\hat\Delta \le x) = \frac{1}{\binom{n}{n_1}} \sum_{b=1}^{\binom{n}{n_1}} \mathbb{1}(\hat\Delta^{(b)} \le x)$ of $\hat\Delta$ induced by w, and compute exact p-values

$$p = \mathbb{P}_w\left(|\hat\Delta(w)| \geq |\hat\Delta(w_{obs})| \right),$$

where w_{obs} denotes the actual original treatment assignment vector. $H_0: Y_i(0) = Y_i(1)$ is commonly referred to as a SHARP NULL, and credit is

usually given to Fisher (1935). However, some authors have noted that Pitman (1937) and Welch (1937) played a more decisive role in the development of a theory of randomization tests; see e.g. Onghena (2018) who provides a detailed historical perspective on the topic.

Rather than referring to "Fisher's Randomization tests", better to refer to "Fisher-Pitman test" or "Pitman-Welch test".

Neyman disagreed with the sharp null approach, which he considered to be only of a theoretical interest. Instead, Neyman considered the average null

$$H_0: \frac{1}{n}\sum_{i=1}^{n}(Y_i(1) - Y_i(0)) = 0,$$

with the implicit alternative that $H_1: \Delta^n \neq 0$. Neyman used a normal approximation for the distribution of $\hat\Delta$ under Ho, and computed the t-statistic

$$t = \frac{\bar Y^{(1)} - \bar Y^{(0)} - \overbrace{\Delta^n}^{=0 \text{ under } H_0}}{\sqrt{\hat V_{\text{Neyman}}}}$$

- Summary

[Fisher/Pitman/Welch]	[Neyman]
Finite pop. model	Finite pop. model
Sharp null $Y_i(0) = Y_i(1)$	Avg null $\Delta^n = 0$
Exact p-value	Approximate based on the $\mathcal{N}$ distribution

↘ Randomization Tests or Permutation Tests ?

These two terms are sometimes used interchangeably. However, there are conceptual reasons for making a distinction between them :

- [Randomization Tests] = based on a random assignment in the finite population model. Under the sharp null of no treatment effect, the random assignment procedure produces just a random shuffle of the response

  – Shuffling is conditionally on the P.O. $Y_i(0)$ and $Y_i(1)$.

- [Permutation tests] = based on a random sampling in the infinite population model. Under the null hypothesis of equal distribution, all permutations of the response are equally likely

  – Permutation is conditionally on Y_i –

  $H_0 : F_0 = F_1$

  where F_j = distribution of the P.O. $Y(j)$.

↑

The numerical procedure is exactly the same, but the interpretation of the results differ.

In particular, permutation distributions donot provide satisfactorily inferences about population means : a result due to Romano (1989) show that when testing for $H_0 : \mu_0 = \mu_1$, (as opposed to $H_0 : F_0 = F_1$ earlier stated)

the permutation distribution of $\sqrt{n_1}\,\hat{\Delta}$ is asymptotically normal $\mathcal{N}\left(0, \sigma_0^2 + \frac{1-\lambda}{\lambda}\sigma_1^2\right)$, where $\frac{n_0}{n} \to \lambda$. On the other hand, the sampling distribution of $\sqrt{n_1}\,\hat{\Delta}$ is asymptotically normal $\mathcal{N}\left(0, \underset{\uparrow}{\sigma_1^2} + \underset{\uparrow}{\frac{1-\lambda}{\lambda}}\, \underset{\uparrow}{\sigma_0^2}\right)$

Unless $\lambda = 1/2$ or $\sigma_0^2 = \sigma_1^2$, these two asymptotic normal distributions differ.

⇒ the permutation distribution does not control the Type I error in general, and does not yield satisfactory confidence intervals for $\mu_1 - \mu_0$. Romano (1989) show that permutation tests are not satisfactory as well when comparing medians. A general theory for robust permutation tests constructed on studentized versions of the test statistic is given in Chung & Romano (2013).

x Remark = Permutation tests & the Bootstrap.

While permutation tests sample the data without replacemt, the bootstrap is another numerical procedure that samples the data with replacement. However, bootstrap significance tests & permutation tests have different theoretical foundations & have different statistical properties (e.g. the bootstrap yields valid conf. intervals for $\mu_1 - \mu_0$ in large sample situations).

II - COVARIATE ADJUSTMENTS.

In this section, assume that we collect for each unit i a vector of covariates $X_i \in \mathbb{R}^d$ (unaffected by the treatment).

↘ we observe $(X_i, Y_i, W_i) \quad \forall i$

$$Y_i = W_i Y_i(1) + (1-W_i) Y_i(0).$$

II.1. The ∞-population model.

× Goal: Estimation of $\Delta^\infty = \mathbb{E}\{Y_i(1) - Y_i(0)\}$ based on $\mathcal{L}_n = \{(X_1, Y_1, W_1), \dots, (X_n, Y_n, W_n)\}$, where each (X_i, Y_i, W_i) is iid; making use of X_i to reduce the variance of the difference estimator.

× Assumption: $X \perp W$ by randomization

× Notation: ↘ $n_1 = \sum_{i=1}^n W_i = $ # units in the treatment group

↘ $n_0 = n - n_1$

↘ $\overline{Y}^{(1)} = \frac{1}{n_1} \sum_{i=1}^n W_i Y_i \qquad \overline{Y}^{(0)} = \frac{1}{n_0} \sum_{i=1}^n (1-W_i) Y_i$

↘ $\overline{W} = \frac{1}{n} \sum_{i=1}^n W_i = \frac{n_1}{n}$

Leon et al (2003) and Davidian et al (2005) derive the class of all consistent estimators of Δ^∞, and show that they can be written exactly or are asymptotically equivalent to an expression of the form

$$(\Delta) \quad \boxed{\underbrace{\overline{Y}^{(1)} - \overline{Y}^{(0)}}_{\hat{\Delta}} - \underbrace{\sum_{i=1}^n (W_i - \overline{W}) \left\{ \frac{h^{(0)}(X_i)}{n_0} + \frac{h^{(1)}(X_i)}{n_1} \right\}}_{\text{correction}}}$$

where $h^{(j)} : \mathbb{R}^d \to \mathbb{R}$ are arbitrary functions.

[Note that the correction term $\xrightarrow{P} 0$, since $X \perp W$]

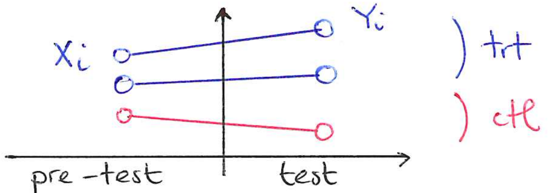
↘ Different estimators of Δ^∞ have different expression for the functions $h^{(j)}$ & different statistical properties. We discuss some examples below, all taken from Tsiatis et al (2008).

× Example 1: the difference estimator.

Trivially obtained for $h^{(j)} \equiv 0 \quad \forall j$. ∎

× Example 2: the diff-in-diff estimator

Take $X_i \in \mathbb{R}$ = pre-treatment outcome for unit i.



The difference-in-difference estimator is

$$(\overline{Y}^{(1)} - \overline{X}^{(1)}) - (\overline{Y}^{(0)} - \overline{X}^{(0)})$$

$$= (\overline{Y}^{(1)} - \overline{Y}^{(0)}) - \underbrace{(\overline{X}^{(1)} - \overline{X}^{(0)})}_{\parallel}$$

$$\frac{1}{n_1} \sum W_i X_i - \frac{1}{n_0} \sum (1-W_i) X_i$$

$$= \frac{1}{n_1} \sum W_i X_i - \frac{1}{n_0} \sum X_i + \frac{1}{n_0} \sum W_i X_i$$

$$= \left(\frac{1}{n_1} + \frac{1}{n_0}\right) \sum W_i X_i - \frac{1}{n_0} \frac{n}{n_1} \sum \overline{W} X_i$$

$$= \left(\frac{1}{n_0} + \frac{1}{n_1}\right) \sum (W_i - \overline{W}) X_i$$

Obtained for $h^{(j)}(x) = x$, $j=0,1$ ∎

x Example 3 = ANCOVA

A standard approach to estimate Δ^∞ is to postulate the linear regression model

$$\boxed{Y = \beta_0 + \beta^t X + \Delta^\infty W + \varepsilon} \quad (*)$$

$\in \mathbb{R}^d$ residual error

We show in Appendix 1 that the least squares estimator of Δ^∞ is asymptotically equivalent to an expression of the form (Δ) page 15, with

$$h^{(j)}(X_i) = \Sigma_{XY}^t \, \Sigma_{XX}^{-1} X_i \qquad j=0,1 ,$$

where

$$\Sigma_{XY} = \mathbb{E}\{(X-\mathbb{E}X)(Y-\mathbb{E}Y)\}$$

$$\Sigma_{XX} = \mathbb{E}\{(X-\mathbb{E}X)(X-\mathbb{E}X)^t\}$$

Let $\hat\Delta_{ANCOVA}$ denote the LS estimator of Δ^∞ in the model $(*)$.

Note that denoting $\Sigma_{XY}^{(j)} := \mathbb{E}\{(X-\mathbb{E}X)(Y-\mathbb{E}Y) \mid W=j\}$,

and $p := \mathbb{P}(W=1)$,

we can rewrite the $h^{(j)}(.)$ as:

$$h^{(j)}(X_i) = \left\{ p\,\Sigma_{XY}^{(1)} + (1-p)\,\Sigma_{XY}^{(0)} \right\}^t \Sigma_{XX}^{-1} X_i$$

This will be useful when comparing the expression of the $h^{(j)}$ in the next example.

x Example 4: ANCOVA with interactions.

Another popular linear model includes interaction terms between X and W:

$$(**) \quad \boxed{Y - \bar{Y} = \beta^t (X - \bar{X}) + \bar{\gamma}(X-\bar{X})(W-\bar{W}) + \Delta^\infty (W - \bar{W}) + \varepsilon}$$

variables should be centered here

Let $\hat\Delta_{ANCOVA\,2}$ denote the least squares estimator of Δ^∞ in the linear model $(**)$. We show in Appendix 2 that $\hat\Delta_{ANCOVA2}$ is asymptotically equivalent to an expression of the form (Δ) with

$$(\Diamond) \quad h^{(j)}(X_i) = \left\{ p\,\Sigma_{XY}^{(0)} + (1-p)\,\Sigma_{XY}^{(1)} \right\}^t \Sigma_{XX}^{-1} X_i$$

Compare with the expression obtained page 17 for $\hat\Delta_{ANCOVA}$: expressions are identical for $p = 1/2$, and differ otherwise $\Rightarrow$ For a 50/50 split treatment/control, expect similar asymptotic precision for the model with and without interaction.

Tsiatis et al (2008) show that the $h^{(j)}(.)$ in $(\Diamond)$ yield estimators of the form (Δ) with smallest possible asymptotic variance, among all estimators for which $h^{(j)}(X_i)$ is linear in X_i.

$\Rightarrow$ ANCOVA with interaction terms is (in general) asymptotically more precise than the linear model without interaction, the diff in diff & the difference estimators.

× Remark / Example 5: Koch et al (1998) propose an estimator of the form

$$\boxed{\hat{\Delta}_{KOCH} := \underbrace{\bar{Y}^{(1)} - \bar{Y}^{(0)}}_{\hat{\Delta}} - \underbrace{V_{XY}^{t} V_{XX}^{-1} \left(\bar{X}^{(1)} - \bar{X}^{(0)} \right)}_{\text{correction}}}$$

where

- $\bar{X}^{(1)} := \frac{1}{n_1} \sum_{i=1}^{n} W_i X_i$
- $\bar{X}^{(0)} := \frac{1}{n_0} \sum_{i=1}^{n} (1 - W_i) X_i$
- $V_{XY} := \frac{1}{n_0} \hat{\Sigma}_{XY}^{(0)} + \frac{1}{n_1} \hat{\Sigma}_{XY}^{(1)}$
- $V_{XX} := \frac{1}{n_0} \hat{\Sigma}_{XX}^{(0)} + \frac{1}{n_1} \hat{\Sigma}_{XX}^{(1)}$
- $\hat{\Sigma}_{XY}^{(j)} := \frac{1}{n_j - 1} \sum_{i=1}^{n} \mathbb{1}(W_i = j)(X_i - \bar{X}^{(j)})(Y_i - \bar{Y}^{(j)})$
- $\hat{\Sigma}_{XX}^{(j)} := \frac{1}{n_j - 1} \sum_{i=1}^{n} \mathbb{1}(W_i = j)(X_i - \bar{X}^{(j)})(X_i - \bar{X}^{(j)})^t$

All the above quantities are computable in practice

We show in Appendix 3 that $\hat{\Delta}_{KOCH}$ is asymptotically equivalent to an expression of the form (Δ) with $h^{(j)}(X_i)$ as in ($\Diamond$) page 18.

⇒ $\hat{\Delta}_{KOCH}$ and $\hat{\Delta}_{ANCOVA\ 2}$ are asymptotically equivalent.

⇒ Unless $p = 1/2$, $\hat{\Delta}_{KOCH}$ is asymptotically more precise than $\hat{\Delta}_{ANCOVA}$, the difference and the diff-in-diff estimators.

Remarks: (i) Tsiatis et al (2008) show that estimators of the form (Δ) with smallest asymptotic variance have $h^{(j)}(X_i) = \mathbb{E}\{ Y_i \mid X_i, W_i = j \}$, $j = 0, 1$.

the true conditional expectation (may or may not be linear in X)

⇒ This result suggests considering separate (non-parametric?) models for each j (eg. Random Forests). The authors argue that failing to specify these models correctly (like in the linear case) will not affect consistency = the resulting estimators will have larger variance than the optimal ones, but remain consistent and asymptotically normal.

↘ In particular, restricting to the class of linear models like ANCOVA will improve on the unadjusted difference estimator $\hat{\Delta}$, even when the true conditional expectation of Y given (X, W) is non-linear.

(ii) Deng et al (2013) consider adjustments of the form $\tilde{Y} = Y - \theta X$; where

- $\theta \in \mathbb{R}$
- X = pre-test variable.

They observe that the θ minimizing the variance of $\tilde{Y}$ is $\operatorname{cov}(X, Y) / \operatorname{var} X$, equal to the LS estimator of the coefficient when regressing $Y - \bar{Y}$ on $X - \bar{X}$. Their estimator of Δ^{∞} is then $\tilde{Y}^{(1)} - \tilde{Y}^{(0)}$, and is widely known as the CUPED estimator. Note however that the CUPED estimator is nothing else but $\hat{\Delta}_{KOCH}$, which was introduced 15 years earlier.

II. 2. The finite population model.

Lin (2013) analyses the behaviour of the LS regression-adjusted estimates when the random assignment is the only source of randomness. Similarly to the ∞-population model, Lin shows that "adjustments cannot hurt asymptotic precision when a full set of treatment x covariate interactions is included", even when the linear model is misspecified.

↘ Lin's contribution offers an alternative perspective to Freedman (2008)'s influencial paper, who argued that adjustments can worsen asymptotic precision. However, Freedman did not consider interaction terms in his paper.

- Model : $$Y_i = \underset{\text{rdm}}{W_i}\, \underset{\text{fixed}}{Y_i(1)} + \underset{\text{rdm}}{(1-W_i)}\, \underset{\text{fixed}}{Y_i(0)}$$

- Let ↘ $\hat{\Delta}_{ADJ}$ = LS estimate of the coefficient of W_i in the linear model $Y_i = \beta_0 + \beta_1 X_i + \delta W_i$

  ↘ $\hat{\Delta}_{INTERACT}$ = LS estimate of the coeff of W_i in $Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_i (X_i - \bar{X}) + \delta W_i$

It follows from Freedman (2008) and Lin (2013) that both $\hat{\Delta}_{ADJ}$ and $\hat{\Delta}_{INTERACT}$ are consistent and asymptotically normal. In addition, it follows from Corollary 1.1 that $\hat{\Delta}_{INTERACT}$ is at least as efficient as $\hat{\Delta}$ (the difference estimator), and more efficient unless the covariates are uncorrelated with the weighted average $\frac{n_0}{n} Y_i(1) + \frac{n_1}{n} Y_i(0)$. On the other hand, Corollary 1.2 ensures that $\hat{\Delta}_{INTERACT}$ is at least as efficient as $\hat{\Delta}_{ADJ}$, and more efficient unless (i) the two groups have equal size or (ii) the covariates are uncorrelated with $Y_i(1) - Y_i(0)$.

↖ Results are proved under general moment conditions.

↖ Similar with the ∞-population model: the $h^{(j)}(X_i)$ pages 17 and 18 are identical if either (i) $p = 1/2$ or (ii) $\Sigma_{XY} = 0$

III. AN ALTERNATIVE (BUT EQUIVALENT) VIEW

x Observations : $(\underset{\in \mathbb{R}^d}{X_i}, \underset{\in \mathbb{R}}{Y_i}, \underset{\in \{0,1\}}{W_i})$ $\forall i$ (same as Section II)

The difference-in-means estimator can be re-expressed as the solution to an optimization problem

$$\hat{\gamma}_0, \hat{\gamma}_1 = \underset{\gamma_0, \gamma_1}{\text{argmin}} \sum_{i=1}^{n} (Y_i - \gamma_0 - \gamma_1 W_i)^2$$

$$\hookrightarrow \hat{\gamma}_1 = \bar{Y}^{(1)} - \bar{Y}^{(0)}.$$

Alternatively, we may consider two separate models for units with $W_i = 1$ and units with $W_i = 0$:

$$(\hat{c}_{(0)}, \hat{\beta}_{(0)}) = \underset{c_{(0)}, \beta_{(0)}}{\text{argmin}} \sum_{i|W_i=0} \left(Y_i - c_{(0)} - \beta_{(0)}^t (X_i - \bar{X}^{(0)})\right)^2$$

$$(\hat{c}_{(1)}, \hat{\beta}_{(1)}) = \underset{c_{(1)}, \beta_{(1)}}{\text{argmin}} \sum_{i|W_i=1} \left(Y_i - c_{(1)} - \beta_{(1)}^t (X_i - \bar{X}^{(1)})\right)^2$$

Since the X_i are centered, we immediatly get that $\hat{c}_{(0)} = \bar{Y}^{(0)}$ and $\hat{c}_{(1)} = \bar{Y}^{(1)}$

⇒ We recover the difference-in-means estimator $\hat{c}_{(1)} - \hat{c}_{(0)} = \bar{Y}^{(1)} - \bar{Y}^{(0)}$.

Alternatively, we can make use of the fitted models in the two subpopulations and consider the following difference estimator (∗)

$$\boxed{\left\{\bar{Y}^{(1)} + \hat{\beta}_{(1)}^t (\bar{X} - \bar{X}^{(1)})\right\} - \left\{\bar{Y}^{(0)} + \hat{\beta}_{(0)}^t (\bar{X} - \bar{X}^{(0)})\right\}}$$

Model fitted on the trt group applied to all individuals ...

... and similarly for the model fitted on the ctl group

$$= (\bar{Y}^{(1)} - \bar{Y}^{(0)}) + [\text{covariate adjustment}].$$

In fact, we can show that this difference estimator corresponds exactly to the LS estimate of the coefficient of W_i when regressing Y_i on $1, W_i, X_i, W_i(X_i - \bar{X})$

Model with interaction term

If instead we consider two intercepts and a single slope,

$$(\hat{c}_{(0)}, \hat{c}'_{(1)}, \hat{\beta}) = \underset{(c_{(0)}, c'_{(1)}, \beta)}{\text{argmin}} \sum_{i=1}^{n} \left(Y_i - c_{(0)} - \underset{\uparrow \, c_{(1)} - c_{(0)}}{c'_{(1)}} W_i - \beta^t X_i\right)^2$$

then

$$\boxed{\left\{\bar{Y}^{(1)} - (\bar{X} - \bar{X}^{(1)})^t \hat{\beta}\right\} - \left\{\bar{Y}^{(0)} - (\bar{X} - \bar{X}^{(0)})^t \hat{\beta}\right\}}$$

is exactly the LS estimate of the coefficient of W_i (∗∗) when regressing Y_i on $1, W_i, X_i$.

No interaction model.

- Consequence: Efficiency results discussed in Section II pages 20 and 22 directly apply to estimators (∗) and (∗∗) due to the equivalence

Two separate models with common slope	⇔	No interaction model
Two separate models with two different slopes	⇔	Model with interaction

In this section, we derive the asymptotic variance of the estimator (∗) under correct specification and misspecification of the linear model; and discuss an asymptotically efficient procedure for estimating the ATE using cross-fitting. The rest of this section is based on the lecture notes of Wager (2020).

Assume that $W_i \sim \mathrm{Bern}(\pi) \in \{0,1\}$

$W_i \perp \{Y_i(0), Y_i(1)\}$ (RCT)

$Y_i = Y_i(W_i)$ (SUTVA)

In addition, we consider a population model for the Potential Outcomes $\{Y_i(0), Y_i(1)\} \overset{d}{=} \mathbb{P}$ (iid)

Then $n^{1/2}(\hat{\Delta} - \Delta^{\infty}) \xrightarrow{d} \mathcal{N}(0, V_\Delta)$

$\mathbb{E}(Y_i(1) - Y_i(0))$

difference-in-means

$\hat{\Delta} = \frac{1}{n_1} \sum W_i Y_i - \frac{1}{n_0} \sum (1-W_i) Y_i$ ← introduced p.4

where $V_\Delta = \frac{\mathrm{var}(Y_i(0))}{1-\pi} + \frac{\mathrm{var}(Y_i(1))}{\pi}$

Can we do better?

"Hypothesis for intuition": suppose that

$$Y(w) = \underbrace{c_{(w)} + X^t \beta_{(w)}}_{\text{linear in } X} + \underbrace{\varepsilon(w)}_{\text{error term}} \qquad w = 0,1$$

with $\mathbb{E}(\varepsilon(w) \mid X) = 0 \qquad \mathbb{E} X = 0$

$\mathrm{var}(\varepsilon(w) \mid X) = \sigma^2 \qquad \mathrm{var}\, X = A$

and $\pi = 1/2$ (for convenience; w.l.o.g.)

In this case, $\Delta^{\infty} = c_{(1)} - c_{(0)}$

Under our linear model, we get

$$V_\Delta = 2\,\mathrm{var}(Y_i(0)) + 2\,\mathrm{var}(Y_i(1))$$

$$= 2\,\mathrm{var}(X^t \beta_{(0)}) + 2\,\mathrm{var}(X^t \beta_{(1)}) + 4\sigma^2$$

$$V_\Delta = 2\beta_{(0)}^t A \beta_{(0)} + 2\beta_{(1)}^t A \beta_{(1)} + 4\sigma^2$$

↓ defining $\|u\|_A^2 = u^t A u$

$$V_\Delta = 2\|\beta_{(0)}\|_A^2 + 2\|\beta_{(1)}\|_A^2 + 4\sigma^2$$

Asymptotic variance of the difference-in-means estimator under a linear model for the P.O.

Q: What's the best we could hope for?

Knowing that the true generating process is linear, we can certainly make use of a linear regression approach and improve on the difference estimator:

LS estimator $\hat{c}_{(w)}, \hat{\beta}_{(w)} \leftarrow LS(Y_i \sim X_i \mid W_i = w)$

we consider here two separate linear models, one for each group.

- $\hat{\Delta}_{LS} = \frac{1}{n} \sum_{i=1}^{n} (\hat{c}_{(1)} + X_i^t \hat{\beta}_{(1)}) - \frac{1}{n} \sum_{i=1}^{n} (\hat{c}_{(0)} + X_i^t \hat{\beta}_{(0)})$
  $= (\hat{c}_{(1)} - \hat{c}_{(0)}) + \overline{X}^t (\hat{\beta}_{(1)} - \hat{\beta}_{(0)})$
- Toolbox: $n_w^{1/2} \begin{pmatrix} \hat{c}_{(w)} - c_{(w)} \\ \hat{\beta}_{(w)} - \beta_{(w)} \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(0, \sigma^2 \begin{pmatrix} 1 & 0 \\ 0 & A^{-1} \end{pmatrix}\right)$

$$\begin{aligned} \Rightarrow \hat{\Delta}_{LS} - \Delta^{\infty} &= \hat{c}_{(1)} - c_{(1)} && \approx \mathcal{N}(0, \sigma^2/n_1) \\ &\quad - (\hat{c}_{(0)} - c_{(0)}) && \approx \mathcal{N}(0, \sigma^2/n_0) \\ &\quad + \overline{X}^t (\beta_{(1)} - \beta_{(0)}) && \approx \mathcal{N}\left(0, \tfrac{1}{n} \mathrm{var}\, X_i^t (\beta_{(1)} - \beta_{(0)})\right) \\ &\quad + \overline{X}^t (\hat{\beta}_{(1)} - \beta_{(1)} - (\hat{\beta}_{(0)} - \beta_{(0)})) && \approx O_p(1/n) \end{aligned}$$

Thus, $n^{1/2}(\hat{\Delta}_{LS} - \Delta^{\infty}) \xrightarrow{d} \mathcal{N}(0, V_{LS})$

where $V_{LS} = 4\sigma^2 + \| \beta_{(1)} - \beta_{(0)} \|_A^2$
$= V_{\Delta} - \| \beta_{(1)} + \beta_{(0)} \|_A^2$

↖ Linear regression helps here (as expected since the linear model is correctly specified)

* Summary so far :

	general setting	linear model holds
$\hat{\Delta}$	OK	OK but ...
$\hat{\Delta}_{LS}$	?	good !

Consider now the general case where

$Y_i(w) = \mu_{(w)}(X_i) + \varepsilon_i(w)$, $\mathbb{E}(\varepsilon_i(w)) = 0$
$\text{var}(\varepsilon_i(w)) = \sigma^2(X_i)$
$\pi = 1/2$

Potentially non-linear in X_i

More generally, we could allow the variance to depend on w

The asymptotic variance of the difference in means is $V_{\Delta} = 2\,\text{var}(\mu_{(0)}(X_i)) + 2\,\text{var}(\mu_{(1)}(X_i)) + 4\,\mathbb{E}\,\sigma^2(X_i)$

Now, run LS in the misspecified setting (i.e. considering LS as an algo, not a ML estimator)

$\hat{c}_{(w)}, \hat{\beta}_{(w)} \leftarrow LS(Y_i \sim X_i \mid W_i = w)$
$= \text{argmin} \left(\frac{1}{n_w} \sum_{W_i = w} (Y_i - c_{(w)} - X_i^t \beta_{(w)})^2 \right)$

Denote

$c^*_{(w)}, \beta^*_{(w)} = \text{argmin}\ \mathbb{E}\left([Y_i - c_{(w)} - X_i^t \beta_{(w)}]^2 \mid W_i = w \right)$

↑ ↑ pseudo-true parameters

- Toolbox = we still get a CLT :

$$n_w^{1/2} \begin{pmatrix} \hat{c}_{(w)} - c^*_{(w)} \\ \hat{\beta}_{(w)} - \beta^*_{(w)} \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(0, \begin{pmatrix} MSE^*_{(w)} & 0 \\ 0 & \boxed{\cdots} \end{pmatrix}\right)$$

this term does not matter for us

where $MSE^*_{(w)} = \mathbb{E}\left\{ (Y_i - c^*_{(w)} - X_i^t \beta^*_{(w)})^2 \mid W_i = w \right\}$

$MSE^*_{(w)}$ would be equal to σ^2 if the linear model was correctly specified.

Then $\tau = c^*_{(1)} - c^*_{(0)}$ (equal to the mean difference of the $\mu_{(w)}$'s)
$c^*_{(w)} = \mathbb{E}(\mu_{(w)}(X_i))$.
(the intercept in the linear model with $\mathbb{E} X_i = 0$ captures the average response variable)

As before,

$$\begin{aligned} \hat{\Delta}_{LS} - \Delta^{\infty} = {} & \hat{c}_{(1)} - c^*_{(1)} && \approx \mathcal{N}(0, MSE^*_{(1)}/n_1) \\ & - (\hat{c}_{(0)} - c^*_{(0)}) && \approx \mathcal{N}(0, MSE^*_{(0)}/n_0) \\ & + \bar{X}^t(\beta^*_{(1)} - \beta^*_{(0)}) && \approx \mathcal{N}(0, \tfrac{1}{n} \| \beta^*_{(1)} - \beta^*_{(0)} \|_A^2) \\ & + \bar{X}^t\left(\hat{\beta}_{(1)} - \beta^*_{(1)} - (\hat{\beta}_{(0)} - \beta^*_{(0)})\right) && \approx O_p(1/n) \end{aligned}$$

$$\Rightarrow \boxed{\begin{array}{l} n^{1/2}(\hat{\Delta}_{LS} - \Delta^{\infty}) \xrightarrow{d} \mathcal{N}(0, V_{LS}) \\ \text{where } V_{LS} = 2MSE^*_{(1)} + 2MSE^*_{(0)} + \|\beta^*_{(1)} - \beta^*_{(0)}\|^2_A \end{array}}$$

To be compared with

$$V_\Delta = 2\,var(\mu_0(X_i)) + 2\,var(\mu_{(1)}(X_i)) + 4\,\mathbb{E}(\sigma^2(X_i))$$

Note that

$$MSE^*_{(\omega)} = \mathbb{E}\Big[\big(\underset{\mu_{(\omega)}(X_i)+\varepsilon_i(\omega)}{\underset{\|}{Y_i}} - c^*_{(\omega)} - X_i^t\beta^*_{(\omega)}\big)^2\Big]$$

$$= \mathbb{E}(\varepsilon_i^2(\omega)) + \mathbb{E}\big(\mu_{(\omega)}(X_i) - c^*_{(\omega)} - X_i^t\beta^*_{(\omega)}\big)^2 + 2\,\mathbb{E}\big\{\varepsilon_i(\omega)\big(\mu_{(\omega)}(X_i) - c^*_{(\omega)} - X_i^t\beta^*_{(\omega)}\big)\big\}$$

(first, condition on X_i)

$$= \mathbb{E}(\sigma^2(X_i)) + var\big(\mu_{(\omega)}(X_i) - c^*_{(\omega)} - X_i^t\beta^*_{(\omega)}\big)$$

(since ↑ is zero mean due to $c^*_{(\omega)} = \mathbb{E}(\mu_{(\omega)}(X_i))$ and $\mathbb{E}X_i = 0$)

$$= var\,\mu_{(\omega)}(X_i) + var(X_i^t\beta^*_{(\omega)}) - 2\,\underbrace{cov(\mu_{(\omega)}(X_i), X_i^t\beta^*_{(\omega)})}_{\overset{\|}{var(X_i^t\beta^*_{(\omega)})}}$$

since $\beta^*_{(\omega)}$ is the projection of $\mu_{(\omega)}(X_i)$ onto X_i

$$= \mathbb{E}(\sigma^2(X_i)) + var(\mu_{(\omega)}(X_i)) - \|\beta^*_{(\omega)}\|^2_A$$

Thus

$$V_{LS} = 4\,\mathbb{E}(\sigma^2(X_i)) + 2\,var(\mu_{(0)}(X_i)) + 2\,var(\mu_{(1)}(X_i)) + \|\beta^*_{(1)} - \beta^*_{(0)}\|^2_A - 2\|\beta^*_{(0)}\|^2_A - 2\|\beta^*_{(1)}\|^2_A.$$

$$\boxed{V_{LS} = V_\Delta - \|\beta^*_{(1)} + \beta^*_{(0)}\|^2_A}$$

The LS estimator beats the simple difference estimator under misspecification. By how much depends on the amount of linearity in the data, captured by the term $\|\beta^*_{(1)} + \beta^*_{(0)}\|^2_A$.

× Remark: Non-parametric version with cross-fitting:

(i) Split the data in A/B samples of sizes $n/2$.
On sample A, estimate $\hat{\mu}^A_{(\omega)}(X)$ via any method
On sample B, —— $\hat{\mu}^B_{(\omega)}(X)$ ——

(ii) $$\hat{\Delta}_A := \frac{1}{n/2}\sum_{i\in A}\Big\{\big(\hat{\mu}^B_{(1)}(X_i) - \hat{\mu}^B_{(0)}(X_i)\big) + \frac{W_i}{\pi}\big(Y_i - \hat{\mu}^B_{(1)}(X_i)\big) + \frac{1-W_i}{1-\pi}\big(Y_i - \hat{\mu}^B_{(0)}(X_i)\big)\Big\}$$

(iii) $$\hat{\Delta} := \frac{1}{2}\hat{\Delta}_A + \frac{1}{2}\hat{\Delta}_B$$

If $\hat{\mu}^{A/B}_{(0/1)}(\cdot)$ is asymptotically consistent in L_2, then a CLT holds $n^{1/2}(\hat{\Delta} - \Delta^{\infty}) \xrightarrow{d} \mathcal{N}(0, V^*)$ with $V^* = Var(\Delta(X_i)) + \frac{1}{\pi(1-\pi)}\mathbb{E}(\sigma^2(X_i))$ where

$\Delta(X_i) = \mathbb{E}(Y_i(1) - Y_i(0) \mid X_i)$,
and this is the best you can do.

- proof for the expression V^*

- Step I = Consider the oracle estimators $\hat{\Delta}^*_{A/B}$:

$$\hat{\Delta}^*_A = \frac{1}{n/2} \sum_{i \in A} \Big\{ (\mu_{(1)}(X_i) - \mu_{(0)}(X_i)) + \frac{W_i}{\pi}(Y_i - \mu_{(1)}(X_i)) + \frac{1-W_i}{1-\pi}(Y_i - \mu_{(0)}(X_i)) \Big\}$$

Three uncorrelated terms

$$\operatorname{var} \hat{\Delta}^*_A = \frac{4}{n^2} \times \frac{n}{2} \times \Big\{ \operatorname{Var}\Big[\overbrace{(\mu_{(1)}(X_i) - \mu_{(0)}(X_i))}^{= \Delta(X_i)} + \frac{W_i}{\pi}\overbrace{(Y_i - \mu_{(1)}(X_i))}^{= \varepsilon_i(1)} + \frac{1-W_i}{1-\pi}\underbrace{(Y_i - \mu_{(0)}(X_i))}_{= \varepsilon_i(0)} \Big]$$

$$= \frac{2}{n} \Big\{ \operatorname{var}(\Delta(X_i)) + \frac{1}{\pi^2} \underbrace{\operatorname{var}(W_i \, \varepsilon_i(1))}_{= \pi \mathbb{E} \sigma^2(X_i)} + \frac{1}{(1-\pi)^2} \underbrace{\operatorname{var}((1-W_i)\,\varepsilon_i(0))}_{\text{\& similarly}} \Big\}$$

$$= \frac{2}{n} \Big\{ \operatorname{var}(\Delta(X_i)) + \frac{1}{\pi(1-\pi)} \mathbb{E}\, \sigma^2(X_i) \Big\}$$

$\Rightarrow \operatorname{var}\left(\frac{1}{2}\hat{\Delta}^*_A + \frac{1}{2}\hat{\Delta}^*_B\right) = n^{-1} V^*$

and $n^{1/2}(\hat{\Delta}^* - \Delta^\infty) \xrightarrow{d} \mathcal{N}(0, V^*)$.

- Step II = We proved in step I an CLT for $\hat{\Delta}^*$.
To show that $n^{1/2}(\hat{\Delta} - \Delta^\infty) \xrightarrow{d} \mathcal{N}(0, V^*)$, it remains to show that $n^{1/2}(\hat{\Delta} - \hat{\Delta}^*) \xrightarrow{P} 0$ $(*)$
We show that $(*)$ holds provided $\hat{\mu}^{A/B}_{(w)}(\cdot)$ is asymp. consistent in L_2.

$$\hat{\Delta}_A - \hat{\Delta}^*_A = \frac{1}{n/2} \sum_{i \in A} \Big\{ \hat{\mu}^B_{(1)}(X_i) + \cancel{\frac{W_i}{\pi}}(Y_i - \hat{\mu}^B_{(1)}(X_i)) - \mu_{(1)}(X_i) - \cancel{\frac{W_i}{\pi}}(Y_i - \mu_{(1)}(X_i)) + \text{terms involving } \hat{\mu}^B_{(0)} \text{ \& } \mu_{(0)} \Big\}$$

$$= \frac{1}{n/2} \sum_{i \in A} \Big\{ (\hat{\mu}^B_{(1)}(X_i) - \mu_{(1)}(X_i))\left(1 - \frac{W_i}{\pi}\right) + \cdots \Big\}$$

$$\mathbb{E}\{(\ldots)^2\} = \mathbb{E}\,\underbrace{\mathbb{E}\Big(\frac{1}{n/2}\sum_{i \in A} \text{—''—}\Big)^2 \mid I_B}_{\text{zero mean conditionally on } I_B \text{ since } \mathbb{E}W_i = \pi}$$

$$= \mathbb{E}\operatorname{var}\Big(\frac{1}{n/2}\sum_{i \in A} \text{—''—} \mid I_B\Big)$$

omitting terms involving $\hat{\mu}^B_{(0)}$ and $\mu_{(0)}$

$$= \frac{1}{n/2} \underset{\text{var}}{\mathbb{E}}\Big\{ (\hat{\mu}^B_{(1)}(X_i) - \mu_{(1)}(X_i))\left(1 - \frac{W_i}{\pi}\right) \mid I_B \Big\}$$

$$= \frac{1}{n/2}\left(\frac{1}{\pi} - 1\right) \mathbb{E}\left(\hat{\mu}^B_{(1)}(X_i) - \mu_{(1)}(X_i)\right)^2$$

$$= \frac{o(1)}{n}$$

$\hat{\mu}^B_{(w)}$ is L_2- consistent

so that $n^{1/2}(\hat{\Delta}_A - \hat{\Delta}^*_A) \xrightarrow{P} 0$, and similarly for $\hat{\Delta}_B$.

- Appendix 1 : ANCOVA with no interactions

Model : $Y_i = \beta_0 + \beta^t X_i + \Delta^\infty W_i + \varepsilon_i$

Instead of fitting the ANCOVA model with intercept directly, center the variables $X_i - \overline{X}$, $W_i - \overline{W}$ so that the least squares estimate of β_0 is $\hat{\beta}_0 = \overline{Y}$

$\Rightarrow$ Consider the linear model without intercept on all centered variables :

$$Y_i - \overline{Y} = \beta^t (X_i - \overline{X}) + \Delta^\infty (W_i - \overline{W}) + \varepsilon_i .$$

$$\Longleftrightarrow \underset{(n\times 1)}{\tilde{Y}} = \underset{n\times(d+1)}{\tilde{X}} \underset{(d+1)\times 1}{\gamma} + \underset{(n\times 1)}{\varepsilon}$$

with • $\tilde{Y} = Y - \overline{Y} \in \mathbb{R}^n$ $\quad Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$

• $\tilde{X} = \Big(\underbrace{X - \overline{X}}_{d} \quad \underbrace{W - \overline{W}}_{1} \Big) \updownarrow n$ $\quad X = \begin{pmatrix} X_1^t \\ \vdots \\ X_n^t \end{pmatrix}$

The least squares estimator of γ is $\hat{\gamma} = (\tilde{X}^t \tilde{X})^{-1} \tilde{X} \tilde{Y}$

(d+1)x1 vector ; we are interested in the last entry, corresponding to the LS estimate of Δ^∞.

Note that

$$\tilde{X}^t \tilde{X} = \left(\begin{array}{c|c} n\hat{\Sigma}_{XX} & d_1 \\ \hline d_1^t & \frac{n_0 n_1}{n} \end{array} \right) \begin{array}{l} \updownarrow d \\ \updownarrow 1 \end{array}$$

$\overleftrightarrow{d}$ $\quad \overleftrightarrow{1}$

$\frac{n_0 n_1}{n} = \sum (W_i - \overline{W})^2$

$d_1 = \sum (W_i - \overline{W}) X_i$

$n \hat{\Sigma}_{XX} = \sum_{i=1}^{n} (X_i - \overline{X})(X_i - \overline{X})^t$

and $\tilde{X}^t \tilde{Y} = \begin{pmatrix} n \hat{\Sigma}_{XY} \\ \boxed{\sum (W_i - \overline{W}) Y_i} \end{pmatrix}$

$\hookrightarrow = \frac{n_0 n_1}{n} \left(\overline{Y}^{(1)} - \overline{Y}^{(0)} \right)$

Toolbox : Block matrix inversion

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} (A - BD^{-1}C)^{-1} & 0 \\ 0 & (D - CA^{-1}B)^{-1} \end{pmatrix} \begin{pmatrix} I & -BD^{-1} \\ -CA^{-1} & I \end{pmatrix}$$

$\hookrightarrow$ apply for $\begin{cases} A = n \hat{\Sigma}_{XX} \\ B = d_1 \\ C = d_1^t \\ D = n_0 n_1 / n \end{cases}$

$\hookrightarrow (A - BD^{-1}C) = n \hat{\Sigma}_{XX} - \frac{n}{n_0 n_1} d_1 d_1^t$

$\hookrightarrow (D - CA^{-1}B) = \frac{n_0 n_1}{n} - n^{-1} d_1^t \hat{\Sigma}_{XX}^{-1} d_1$

$$\Rightarrow \begin{pmatrix} (A - BD^{-1}C)^{-1} & 0 \\ 0 & \boxed{(D - CA^{-1}B)^{-1}} \end{pmatrix} \underbrace{\begin{pmatrix} I & -BD^{-1} \\ -CA^{-1} & I \end{pmatrix} \begin{pmatrix} n \hat{\Sigma}_{XY} \\ \frac{n_0 n_1}{n} (\overline{Y}^{(1)} - \overline{Y}^{(0)} \end{pmatrix}}$$

$$\begin{pmatrix} n \hat{\Sigma}_{XY} - \frac{\cancel{n}}{n_0 n_1} d_1 \frac{\cancel{n_0 n_1}}{\cancel{n}} (\overline{Y}^{(1)} - \overline{Y}^{(0)} \\ \boxed{-n^{-1} d_1^t \hat{\Sigma}_{XX}^{-1} \hat{\Sigma}_{XY} + \frac{n_0 n_1}{n} (\overline{Y}^{(1)} - \overline{Y}} \end{pmatrix}$$

The product of these two terms is the quantity of interest :

$$\frac{n}{n_0 n_1} \Big(\underbrace{1 - \frac{n^2}{n_0 n_1} (n^{-1} d_1)^t \hat{\Sigma}_{XX}^{-1} (n^{-1} d_1)}_{P \downarrow 1} \Big)^{-1} \left(\frac{n_0 n_1}{n} \left[\overline{Y}^{(1)} - \overline{Y}^{(0)} \right] - \hat{\Sigma}_{XY}^t \hat{\Sigma}_{XX} d \right.$$

$$\rightarrow \overline{Y}^{(1)} - \overline{Y}^{(0)} - \frac{n}{n_0 n_1} \sum (W_i - \overline{W}) \hat{\Sigma}_{XY}^t \hat{\Sigma}_{XX} X_i$$

∎

- Appendix 2: ANCOVA with interaction terms.

We proceed as in Apendix 1, with $\tilde Y = \tilde X \gamma + \varepsilon$,

$$\tilde X = \begin{pmatrix} | & | & | \\ X-\bar X & (X-\bar X)(w-\bar w) & w-\bar w \\ | & | & | \end{pmatrix}$$

- $$\tilde X^t \tilde X = \left(\begin{array}{c|c} \overbrace{\begin{matrix} n\hat\Sigma^{(0)}_{XX} & n\hat\Sigma^{(1)}_{XX} \\ n\hat\Sigma^{(1)}_{XX} & n\hat\Sigma^{(2)}_{XX} \end{matrix}}^{/\!/\; n\hat D} & \overbrace{\begin{matrix} d_1 \\ \Sigma (W_i-\bar W)^2 (X_i-\bar X) \end{matrix}}^{/\!/\; d_2} \\ \hline \underbrace{\begin{matrix} d_1^t & \Sigma (W_i-\bar W)^2 \\ & \times (X_i-\bar X)^t \end{matrix}}_{d_2^t \;/\!/} & n_0 n_1/n \end{array}\right)$$

with $\hat\Sigma^{(\ell)}_{XX} := \frac{1}{n}\sum_{i=1}^n (W_i-\bar W)^\ell (X_i-\bar X)(X_i-\bar X)^t$

- $$\tilde X^t \tilde Y = \begin{pmatrix} n\,\hat\Sigma_{XY} \\ n\,\hat\Sigma_{XYW} \\ \Sigma (W_i-\bar W) Y_i \end{pmatrix}$$

with $\hat\Sigma_{XYW} = \frac{1}{n}\sum_{i=1}^n (X_i-\bar X)(Y_i-\bar Y)(W_i-\bar W)$

We proceed as before, except that d_1 is replaced with d_2, $\hat\Sigma_{XX}$ with $\hat D$, and $\hat\Sigma_{XY}$ with $\begin{pmatrix}\hat\Sigma_{XY} \\ \hat\Sigma_{XYW}\end{pmatrix}$

The least squares estimate for Δ^∞ is

$$\underbrace{\left(1-\frac{n^2}{n_0 n_1}(n^{-1}d_2)^t \hat D^{-1}(n^{-1}d_2)\right)^{-1}}_{\substack{\downarrow p \\ 1}}\left(\bar Y^{(1)}-\bar Y^{(0)}-\frac{n}{n_0 n_1} d_2^t \hat D^{-1}\begin{pmatrix}\hat\Sigma_{XY} \\ \hat\Sigma_{XYW}\end{pmatrix}\right)$$

Note that $\hat\Sigma_{XYW} \sim p(1-p)\left\{\Sigma^{(1)}_{XY} - \Sigma^{(0)}_{XY}\right\}$

$$\Rightarrow \hat D^{-1}\begin{pmatrix}\hat\Sigma_{XY} \\ \hat\Sigma_{XYW}\end{pmatrix} \sim \begin{pmatrix} \Sigma_{XX}^{-1} & 0 \\ 0 & \frac{1}{p(1-p)}\Sigma_{XX}^{-1}\end{pmatrix}\begin{pmatrix}\Sigma_{XY} \\ p(1-p)\left[\Sigma^{(1)}_{XY}-\Sigma^{(0)}_{XY}\right]\end{pmatrix}$$

$$= \begin{pmatrix} \Sigma_{XX}^{-1}\Sigma_{XY} \\ \Sigma_{XX}^{-1}\left(\Sigma^{(1)}_{XY}-\Sigma^{(0)}_{XY}\right)\end{pmatrix}$$

$$\Rightarrow d_2^t \hat D^{-1}\begin{pmatrix}\hat\Sigma_{XY} \\ \hat\Sigma_{XYW}\end{pmatrix} \sim d_1^t \Sigma_{XX}^{-1}\Sigma_{XY} + \sum_{i=1}^n (W_i-\bar W)^2 \underbrace{(X_i-\bar X)^t \Sigma_{XX}^{-1}\left(\Sigma^{(1)}_{XY} - \Sigma^{(0)}_{XY}\right)}_{\substack{\in \mathbb{R} \\ \Rightarrow \text{ equal to its transpose}}}.$$

and

$$= \sum_{i=1}^n (W_i-\bar W)\, \Sigma_{XY}^t \Sigma_{XX}^{-1} X_i + \sum_{i=1}^n (W_i-\bar W)^2 \left(\Sigma^{(1)}_{XY}-\Sigma^{(0)}_{XY}\right)^t \Sigma_{XX}^{-1}(X_i-\bar X)$$

- Toolbox $= \sum_{i=1}^n (W_i-\bar W)^2 (X_i-\bar X) = \sum_{i=1}^n (W_i-\bar W) X_i (1-2\bar W)$

proof $=$

$$\overset{\parallel}{\sum_{i=1}^n} (W_i-\bar W)(W_i X_i - \bar X W_i - \bar W X_i + \cancel{\bar W \bar X})$$

$$\downarrow$$

$$= \Sigma (W_i-\bar W) W_i \bar X = \underbrace{\bar X \,\Sigma W_i}_{\substack{\parallel \\ \bar W \,\Sigma X_i}} - \underbrace{\bar X \bar W\, \Sigma W_i}_{\substack{\parallel \\ (\bar W)^2\, \Sigma X_i}}$$

$$= \Sigma X_i \bar W (1-\bar W)$$

$$\ldots = \sum_{i=1}^n X_i \left\{ -\bar W(1-\bar W) + (W_i-\bar W)(W_i-\bar W) \right\}$$

$$= \sum_{i=1}^n X_i (W_i-\bar W)(1-2\bar W)$$

☒

We get that

$$d_2^t \hat{D}^{-1} \begin{pmatrix} \hat{\Sigma}_{XY} \\ \hat{\Sigma}_{XYW} \end{pmatrix} \sim \sum_{i=1}^{n} (W_i - \bar{W}) \Sigma_{XY}^t \Sigma_{XX}^{-1} X_i$$

$$+ \sum_{i=1}^{n} (W_i - \bar{W}) \left(\Sigma_{XY}^{(1)} - \Sigma_{XY}^{(0)} \right)^t \Sigma_{XX}^{-1} X_i \underbrace{(1-2\bar{Z})}_{\sim 1-2p}$$

$$\sim \sum_{i=1}^{n} (W_i - \bar{W}) \left\{ \Sigma_{XY} + (1-2p)\left(\Sigma_{XY}^{(1)} - \Sigma_{XY}^{(0)}\right) \right\} \Sigma_{XX}^{-1} X_i$$

$$= \sum_{i=1}^{n} (W_i - \bar{W}) \left(p\, \Sigma_{XY}^{(0)} + (1-p)\, \Sigma_{XY}^{(1)} \right) \Sigma_{XX}^{-1} X_i$$

∎

Appendix 3 = Asymptotic equivalence between $\hat{\Delta}_{KOCH}$ and $\hat{\Delta}_{ANCOVA\ 2}$.

We derive asymptotic equivalents for $n V_{XX}$ and $n V_{XY}$:

↘ $$n V_{XX} = \frac{n}{n_0} \hat{\Sigma}_{XX}^{(0)} + \frac{n}{n_1} \hat{\Sigma}_{XX}^{(1)}$$

$$\sim \frac{n}{n_0} \frac{1}{n_0} \sum (1-W_i)(X_i - \bar{X}^{(0)})(X_i - \bar{X}^{(0)})^t$$

$$+ \frac{n}{n_1} \frac{1}{n_1} \sum W_i (X_i - \bar{X}^{(1)})(X_i - \bar{X}^{(1)})$$

$$\sim (1-p)^{-1} \mathbb{E}\{(X - \mathbb{E}X)(X - \mathbb{E}X)^t \mid W = 0\}$$

$X \perp W$ ↓

$$+ p^{-1} \mathbb{E}\{(X - \mathbb{E}X)(X - \mathbb{E}X)^t \mid W = 1\}$$

$$= \frac{1}{p(1-p)} \Sigma_{XX}$$

↘ $$n V_{XY} = \frac{n}{n_0} \hat{\Sigma}_{XY}^{(0)} + \frac{n}{n_1} \hat{\Sigma}_{XY}^{(1)}$$

$$\sim (1-p)^{-1} \mathbb{E}\{(X - \mathbb{E}X)(Y - \mathbb{E}Y) \mid W = 0\}$$

$$+ p^{-1} \mathbb{E}\{(X - \mathbb{E}X)(Y - \mathbb{E}Y) \mid W = 1\}$$

$$= \frac{1}{p(1-p)} \left\{ (1-p) \Sigma_{XY}^{(1)} + p\, \Sigma_{XY}^{(0)} \right\}$$

The result follows since $\bar{X}^{(1)} - \bar{X}^{(0)} = \frac{n}{n_0 n_1} \sum X_i (W_i - \bar{W})$.

∎

References.

- Susan Athey & Guido Imbens (2016). The Economics of Randomized Experiments (ArXiv)
- Chung & Romano (2013). Exact and Asymptotic Robust Permutation Tests. Ann. Stats 41(2): 484 - 507
- Cochran, W. G. (1977). Sampling Techniques. 3rd Edition. John Wiley & Sons, New York.
- Davidian M, Tsiatis AA, Leon S. Semiparametric Estimation of Treatment Effect in a Pretest - Posttest Study With Missing Data (with Discussion). Stat. Science (2005). 20: 261: 301
- Deng A, Xu Y, Kohavi R, Walker T (2013). Improving the Sensitivity of Online Controlled Experiments by Utilizing Pre- Experiment Data. WSDM (proceedings).
- Fisher R. (1935). The Design of Experiments.
- Freedman D.A. (2007). On Regression Adjustments to Experimental Data. Adv. in Applied Maths 40: 180-193.
- Koch & al (1998). Issues for Covariance Analysis of Dichotomous ... Stats in Medicine 17: 1863: 1892.
- Leon S, Tsiatis AA, Davidian M. Semiparametric Estimation of Treatment Effect in a Pretest - Posttest Study. Biometrics 59: 1048-1057.
- Lin W (2013). Agnostic Notes On Regression Adjustments To Experimental Data... Ann of Applied Stats 7(1): 295: 318.
- Neyman J (1923). Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. Roczniki Nauk Rolniczych 10: 1-51. & Statistical Science (1990) 5(4). 465: 480
- Onghena P. (2018). Randomization tests or permutation tests? A historical and terminological clarification. In V. Berger (Ed.), Randomization, masking, and allocation concealment (209: 227). Boca Raton/FL: Chapman & Hall/CRC Press.
- Pitman, E.J.G. (1937). Significance tests which may be applied to samples from any populations. JRSS B (4). 119 - 130.
- Reichardt CS and Gollob HF (1999). Justifying the use and increasing the power of a t test for a randomized experiment with a convenience sample. Psychological Methods. 4(1) 117-128.
- Romano J.P. (1989). On the behaviour of Randomization Tests without a group invariance assumption. Technical Report.
- Rubin DB (1978). Bayesian inference for causal effects: The Role of Randomization. Annals of Statistics 6: 34-58.
- Tsiatis et al (2008). Covariate adjustments... Stat Med.
- Welch BL (1937). On the z-test in randomized... Biometrika.
- Wager S (2020). Causal Inference (Lecture Notes). Standford.