Randomization “should” balance all baseline covariates (observed and unobserved) on average...

But covariates may be imbalanced by random chance, and sometimes better balance is desired

Why is covariate balance important? Because better balance

- Provides more meaningful estimates of the causal effect – internal validity
- Increases precision (reduces variance) of estimator, if covariates correlated with outcome (outcome less variable for similar values of covariates)
Covariate Balance

- Option 1: force better balance on important covariates \textit{by design}
  - stratified randomized experiments
  - paired randomized experiments
  - rerandomization

- Option 2: correct imbalance in covariates by \textit{analysis} – covariate adjustment
  - outcome: gain scores
  - separate analysis within subgroups
  - regression (ANCOVA)
  - propensity score weighting
  - model-based imputation
Stratified Experiments

▶ Units are stratified (grouped, blocked) according to covariate(s)

▶ Subdivide sample into $J$ homogeneous strata (blocks)

▶ Randomize units to treatment groups within each strata

▶ Often used with important categorical covariates (or discretized quantitative)

▶ (similar to stratified sampling)
Stratified Experiments: Bee Stings Example


If you are stung by a bee, does that make you more likely to get stung again? (Might bees leave behind a chemical message that tells other bees to attack you?)

Scientists dangled 16 muslin-wrapped cotton balls over a beehive, where half of the balls had been previously stung and the other half were fresh.

Outcome: total number of new stingers

Repeated for a total of nine trials.
Stratified Experiments: Bee Stings Example

- Scientists expect the number of stings to vary by trial (different number of bees in the hive, different times of day, different weather, etc.)

- Each trial is a different strata \((J = 9)\)
Stratified Experiments: Fisherian inference

▶ In Fisher’s mode (randomization) of inference, what to use for a test statistic?

▶ A common choice: weighted average of the strata-specific differences

\[
T = \sum_{j=1}^{J} \frac{N_j}{N} \left( \bar{Y}_{1,j}^{obs} - \bar{Y}_{0,j}^{obs} \right)
\]

where \( \bar{Y}_{z,j}^{obs} \) is the average outcome of group \( z \) in \( J \)th strata, \( N_j \) is the size of \( J \)th strata

▶ Note: another common choice of weights by the variance of each strata – minimize the total variance of \( T \)
Stratified Experiments: Fisherian inference

- Fisher randomization test
- Easier to simulate randomizations, rather than enumerate all possible allocations
- For each simulated randomization, just randomize within strata
- Always for randomization test, just simulate randomization schemes compatible with the randomization design – think about what is the randomization space $S$ in stratified experiments
Bee Stings Example: Fisher’s test

- All trials are equal size ($N_j$ constant), $T^{obs} = 12$
- p-value = 0.04
Stratified Experiments: Neymanian inference

- In Neyman’s mode (repeated sampling) of inference, what to use for an estimator?

- In general, \( \hat{\tau} = \sum_j \lambda_j \hat{\tau}_j \), where \( \hat{\tau}_j \) and \( \lambda_j \) is the estimator and weight of strata \( j \), respectively

- Variance of the general estimator \( \hat{\tau} \) is \( \var(V(\hat{\tau}) = \sum_j \lambda_j^2 \var(V(\hat{\tau}_j)) \)

- One common choice is \( \hat{\tau}_{ave} = \sum_{j=1}^{J} \frac{N_j}{N} \left( \bar{Y}_{1,j}^{obs} - \bar{Y}_{0,j}^{obs} \right) \)

- For the specific estimator \( \hat{\tau}_{ave} \):

\[
\var(\hat{\tau}_{ave}) = \sum_j \frac{N_j^2}{N^2} \left( \frac{s_{1,j}^2}{N_{1,j}} + \frac{s_{0,j}^2}{N_{0,j}} \right)
\]
Stratified Experiments

- “Block what you can; randomize what you cannot” - George Box
- No harm – can only help
- Can stratify on more than one covariate
- Strata can be any size - more balanced design (i.e. similar strata size and similar sizes of treatment and control per strata) has more power
- Difficult to extend to many covariates or continuous covariates
Paired Experiments

- Units are matched into pairs (based on covariate(s))

- Special case of stratified randomized experiment with $N_j = 2$ for each strata

- Useful when expect difference in potential outcomes within a pair to be much smaller than differences across pairs

- Test statistic /estimator: average of differences across all pairs (note: this is the same as the difference of the averages within treatment groups)
Paired Experiments

- Randomization test: randomize sign of each difference
- Neyman inference: analyze differences as a single variable
- When variability within pairs is much smaller than variability across pairs, can get huge gains in precision
- Translate to higher power for tests (lower p-values) and narrower confidence intervals
Chance Imbalance in Randomized Experiments

- Randomization balances all covariates only in expectation, not necessarily in each single trial, particularly when the sample size is small.

- Chance imbalance in baseline covariates commonly exists.

- Some baseline covariates can be predictive of the outcome. For example, in randomized clinical trials, such a covariate is the baseline outcome.

- Chance imbalance in covariates that are predictive of the outcome can
  - reduce the power of the unadjusted estimator $\hat{\tau}_{\text{unadj}}$ detecting true difference if there is
  - threat the face validity of an experiment
Balance by Analysis: Covariate Adjustment

- Solution to chance imbalance: covariate adjustment
  - Regression adjustment
  - Propensity score weighting (later)
  - Model-based imputation/Bayesian (later)
- Goal: improve the precision (power) in estimating the causal effects
Covariate Adjustment: ANCOVA

- Regression adjustment: fit a regression model of the observed $Y$ on the treatment indicator $Z$, covariates $X$, and possibly interactions $Z \times X$; use the ordinary least square (OLS) estimator of the coefficient of $Z$ as the adjusted estimator of ATE $\tau$
- Also known as ANCOVA estimator
- Two different ANCOVA models:
  - ANCOVA 1 (without interactions): $Y_i \sim \beta_0 + \beta_z Z_i + \beta_x X_i$
  - ANCOVA 2 (with interactions):
    - $Y_i \sim \beta_0 + \beta_z Z_i + \beta_x X_i + \beta_{zx} Z_i \times X_i$
- In both models, the OLS estimator of $\beta_z$ is taken as the estimator of $\tau$, often denoted as the adjusted estimator $\hat{\tau}^{adj}$
- Question: Any difference between the two models? Which one is better?
To obtain a ANCOVA estimator that is more efficient than the unadjusted estimator of the ATE $\tau$ in randomized experiments, one must

1. include a full set of treatment-covariates interaction, that is, use ANCOVA2
2. center $X$ (so that you can directly use the OLS estimate of coefficient of $Z$ in the ANCOVA as the adjusted estimator for $\tau$)
Why do interactions matter?

- For simplicity, consider a single continuous covariate $X$
- Fit a regression model of the potential outcomes in treatment and control group separately

\[
Y_i(1) = \alpha_1 + \beta_1 X_i + \epsilon_{i1}
\]
\[
Y_i(0) = \alpha_0 + \beta_0 X_i + \epsilon_{i0}
\]

- Use the equation $Y = Y_i(1)Z_i + Y_i(0)(1 - Z_i)$ to connect the models on the potential outcomes and the observed outcome

\[
Y = (\alpha_1 + \beta_1 X_i)Z_i + (\alpha_0 + \beta_0 X_i)(1 - Z_i) + \epsilon_i
\]
\[
= \alpha_0 + (\alpha_1 - \alpha_0)Z_i + \beta_0 X_i + (\beta_1 - \beta_0)X_i \ast Z_i + \epsilon_i
\]

(1)
OLS Covariate-Adjusted Estimator

- In practice, we can first fit a linear model, $\text{lm}(Y \sim \tilde{X})$, separately in each group, and let $\hat{\beta}_1$ and $\hat{\beta}_0$ be the OLS estimate of the coefficient of $X$ in these two regressions.

- The covariate-adjusted estimator is:

  $$\hat{\tau}_{\text{adj}} = \left\{ \bar{Y}_1 - \hat{\beta}_1' (\bar{X}_1 - \bar{X}) \right\} - \left\{ \bar{Y}_0 - \hat{\beta}_0' (\bar{X}_0 - \bar{X}) \right\}$$

- Numerically, $\hat{\tau}_{\text{adj}}$ is equivalent to the OLS estimate of the coefficient of $Z$ in the following model with interactions with centered $\tilde{X}$

  $$\text{lm}(Y \sim Z + \tilde{X} + Z \ast \tilde{X})$$

- Asymptotically, $\hat{\tau}_{\text{adj}}$ (i.e. ANCOVA2 estimator) is more efficient than $\hat{\tau}_{\text{unadj}}$; in fact it is the most efficient estimator among the semiparametric family of covariate adjustment estimators (Tsiatis et al, 2008).
Model-free validity for model-based statistics

- $\hat{\tau}_{\text{adj}}$ is biased for $\tau$ in finite samples (the bias is usually small and negligible), but is unbiased asymptotically (consistent)

- **Model-free validity for model-based statistics**: Consistency of $\hat{\tau}_{\text{adj}}$ holds even if the linear regression assumption is wrong

- **Intuition**
  - By randomization, the super-population correlation between the treatment indicator and covariates is zero
  - The correlation vanishes in large sample, but may not in finite sample

- Only holds in randomization - not true in observational studies!
To estimate the s.e. of $\hat{\tau}_{\text{adj}}$ under randomization, use the Huber–White sandwich s.e. (also known as the heteroscedasticity-consistent estimator)

$$\hat{\nabla}_{\text{HW}} \left( \hat{\tau}_{\text{adj}} \right) = (X'X)^{-1}(X'diag(\hat{u}_1^2, \cdots, \hat{u}_n^2)X)(X'X)^{-1},$$

where $\hat{u}_i$ are the regression residuals.

Huber–White s.e. is justified by randomization, but the traditional OLS s.e. is not.

When the randomization probability is 1/2, ANCOVA also returns consistent variance (besides point) estimate even if the outcome model is misspecified.
OLS Adjustment: History

- Classical idea dating back to Laplace (cf. Cochran 1977) - Laplace was, among many other things, the first (Bayesian) statistician
- Fisher (1935) proposed ANCOVA estimator in randomized experiments, argued that $\hat{\tau}_{\text{adj}}$ improves the precision (variance) over the unadjusted estimator $\hat{\tau}_{\text{unadj}}$
- Freedman (2008a,b) provided an antagonist view: gave examples where $\hat{\tau}_{\text{adj}}$ decrease precision (unbalanced experiments with heterogeneity, non-continuous outcome)
- Lin (2013) provided an agnostic view: re-examined Freedman’s critique and provided simple fixes – case closed
- Tsiatis and co-authors made similar findings to Lin earlier (Yang and Tsiatis, 2001; Tsiatis et al., 2008)
Regression Adjustment: Limitations

- Limitations of regression adjustment in randomized experiments
  - Invitation for a *fishing expedition*: search for model specifications that give ‘significant’ result
  - Linear regression has nice properties justified by randomization, but other models (e.g., Logit or Probit) do not (Freedman 2008b)
    - Non-continuous (binary or count) outcomes: cannot directly use the coefficient of treatment in a linear model as the causal estimate
    - Rare non-continuous outcomes: a GLM may fail to converge
  - Some purists just disdain the invasion of a model(!) in the sacred temple of randomization

- Propensity score weighting methods can bypass these issues (later)

- All belong to the same semiparametric class of estimators (Tsiatis et al. 2008)
Words of Wisdom on OLS Adjustment

One does not need to believe in the classical linear model to tolerate or even advocate OLS adjustment, just as one does not need to believe in the Four Noble Truths of Buddhism to entertain the hypothesis that mindfulness meditation has causal effects on mental health.

— Winston Lin, Agnostic notes on regression adjustments to experimental data: Reexamining Freedman’s critique (AOAS 2013, p 314)
Model-Based Imputation

- Impute missing potential outcomes
- Use information from control units to impute $Y(0)$ for treated units
- Use information from treated units to impute $Y(1)$ for control units
- Ideally, this should reflect the uncertainty in the imputation
- Similar to multiple imputation (Rubin, 1987)
## Model-Based Imputation

<table>
<thead>
<tr>
<th>Y(1)</th>
<th>Y(0)</th>
<th>X</th>
</tr>
</thead>
<tbody>
<tr>
<td>observed</td>
<td>?</td>
<td>observed</td>
</tr>
<tr>
<td>observed</td>
<td>?</td>
<td>observed</td>
</tr>
<tr>
<td>observed</td>
<td>?</td>
<td>observed</td>
</tr>
<tr>
<td>observed</td>
<td>?</td>
<td>observed</td>
</tr>
<tr>
<td>observed</td>
<td>?</td>
<td>observed</td>
</tr>
<tr>
<td>?</td>
<td>observed</td>
<td>observed</td>
</tr>
<tr>
<td>?</td>
<td>observed</td>
<td>observed</td>
</tr>
<tr>
<td>?</td>
<td>observed</td>
<td>observed</td>
</tr>
<tr>
<td>?</td>
<td>observed</td>
<td>observed</td>
</tr>
<tr>
<td>?</td>
<td>observed</td>
<td>observed</td>
</tr>
</tbody>
</table>

Use treated units to model \( Y(1) \mid X \)

Use control units to model \( Y(0) \mid X \)
Model-Based Imputation

- Options without covariates:
  - Impute observed mean (doesn’t reflect uncertainty)
  - Sample with replacement (bootstrap)

- Options with covariates:
  - Sample with replacement from “donor pools” with similar covariate values
  - Build models to predict potential outcomes based on covariates, i.e. outcome regression
    - MLE can work, but doesn’t fully reflect uncertainty
    - BEST: be fully Bayesian and draw from posterior distribution for $Y(1)$ or $Y(0)$ given the covariates
    - With imputations, one can estimate individual treatment effects (ITE)
  - More on outcome regression in next chapter
In randomized experiments, covariates may be included in the analysis for the following reasons:

- The assignment mechanism depends on covariates (in a known way): e.g., block randomized experiments, cluster randomized experiments.
- Covariates are useful in model-based inference for improving the prediction of the missing outcomes.

In a model-based approach the presence of covariates is usually tackled by specifying a regression model.

Inference can then proceed following either a frequentist or a Bayesian approach.

Randomization ensures *model-free validity for model-based statistics*. 

Cluster Randomized Experiments

- treatment is assigned at cluster level but outcomes are measured for individuals within clusters
- individuals in the same cluster receive the same treatment
- also known as cluster randomized trials (CRT) or group randomized trials (GRT)

Justified when

- treatment is naturally delivered at the cluster level, e.g. villages, classrooms, clinics
- prevent treatment contamination within clusters
- ethical reasons
Estimand: Average Treatment Effect (CRT)

- Individual $j$ ($j = 1, \ldots, m_i$) in cluster $i$ ($i = 1, \ldots, n$) has two P.O’s, $Y_{ij}(1)$ and $Y_{ij}(0)$
- The ATE estimand is characterized as an average over all clusters

\[ \tau = \mathbb{E}[Y_{ij}(1) - Y_{ij}(0)] \]

- Under cluster randomization, can show

\[ \tau = \mathbb{E}[Y_{ij}(1)|Z_i = 1] - \mathbb{E}[Y_{ij}(0)|Z_i = 0] = \mathbb{E}[Y_{ij}|Z_i = 1] - \mathbb{E}[Y_{ij}|Z_i = 0] \]

- The simple difference-in-means estimator is still consistent under cluster randomization

\[ \hat{\tau}^{\text{unadj}} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m_i} Z_i Y_{ij}}{\sum_{i=1}^{n} Z_i m_i} - \frac{\sum_{i=1}^{n} \sum_{j=1}^{m_i} (1 - Z_i) Y_{ij}}{\sum_{i=1}^{n} (1 - Z_i) m_i} \]
Randomization-based Inference
(Gail et al. 1996)

- Randomization test can be carried out in the usual way to test the sharp null $H_0: Y_{ij}(1) = Y_{ij}(0) \forall i, j$

- The unit of randomization is the cluster, and hence the randomization distribution should be created by permuting clusters (instead of individuals)

- Attractive because CRTs often recruit a few clusters ($\leq 24$), guarantee to control of test size at $\alpha$ level

- The weak null $H_0^*: \mathbb{E}[Y_{ij}(1) - Y_{ij}(0)] = 0$ is often more relevant

- Gail et al. (1996) showed that randomization test may have empirical test size exceeding the nominal level when $H_0^*$ holds but $H_0$ violated
Randomization-based Inference

- A pair-matched design often substantially improve the power of CRT (Imai et al. 2009)
- When using pair-matched designs, the randomization test may have near nominal level even under the weak null
- Again the randomization test should respect the randomization design
  - in a pair-matched design, the possible number of randomization schemes is $2^{n/2}$
  - in a completely randomized design, the possible number of randomization schemes is $\binom{n}{n/2}$
  - only the former can be used to create the randomization distribution, otherwise the test may be invalid
Covariate Adjustment in CRT

- Both cluster-level and individual-level covariates can be collected in CRT
- CRT often has small to moderate number of clusters, covariate imbalance is common
- How to use regression to adjust covariate imbalance to gain efficiency? Regression at the cluster level or individual level?
- Su and Ding (2020): main conclusions
  - In large sample (in terms of number of clusters), regression of the scaled cluster total outcome on scaled cluster total covariate and cluster size is more efficient than the regression on individual data
  - Use the Huber-White (heteroskedasticity-robust) estimator of standard errors
  - Consistency and asymptotic normality holds even if the regression is misspecified
Important topics not covered

- Rerandomization
- Cluster randomization (later)
- Multi-arm (more than 2) experiments (later)
- Generalizability of random experiments: internal versus external validity (later)
- Noncompliance (later)
- interference
Covariate Adjustment: Additional Notes on Theory

- Two different ANCOVA models for covariate adjustment in RCT:
  - **ANCOVA1** (without interactions): \( Y_i \sim \beta_0 + \beta_z Z_i + \beta_x^T X_i \)
  - **ANCOVA2** (with interactions):
    \[
    Y_i \sim \beta_0 + \beta_z Z_i + \beta_x^T (X_i - \bar{X}) + \beta_{xz}^T (X_i - \bar{X}) Z_i
    \]

- In both models, the OLS estimator of \( \beta_z \) is taken as the estimator of \( \tau \), often denoted as the adjusted estimator \( \hat{\tau}_{adj} \)

- Question: Any difference between the two OLS models? Which one is better?
Recall ATE $\tau = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] = \mathbb{E}[Y|Z = 1] - \mathbb{E}[Y|Z = 0]$

Sample size: $N_1 = \sum_{i=1}^{N} Z_i$, $N_0 = \sum_{i=1}^{N} (1 - Z_i)$, $N = N_1 + N_0$

Randomization probability: $\tilde{Z} = N^{-1} \sum_{i=1}^{N} Z_i = N_1/N$

$\bar{Y}^{(1)} = N_1^{-1} \sum_{i=1}^{N} Z_i Y_i$, $\bar{Y}^{(0)} = N_0^{-1} \sum_{i=1}^{N} (1 - Z_i)Y_i$

Can show all “reasonable” consistent and asymptotically normal estimators for either can be expressed exactly as or are asymptotically

$$\hat{\tau}_h = \bar{Y}^{(1)} - \bar{Y}^{(0)} - \sum_{i=1}^{N} (Z_i - \tilde{Z}) \left\{ \frac{h^{(0)}(X_i)}{N_0} + \frac{h^{(1)}(X_i)}{N_1} \right\}$$
Covariate Adjustment: General Theory

(Tsiatis et al. 2008)

- General form of the ATE estimator

\[ \hat{\tau}_h = \bar{Y}^{(1)} - \bar{Y}^{(0)} - \sum_{i=1}^{N} (Z_i - \bar{Z}) \left\{ \frac{h^{(0)}(X_i)}{N_0} + \frac{h^{(1)}(X_i)}{N_1} \right\} \]

- What happens if let \( h^{(0)}(X_i) = h^{(0)}(X_i) \equiv 0 \)? (Unadjusted!)

- Consistency for arbitrary choice of \( h \):
  - \( \bar{Y}^{(1)} - \bar{Y}^{(0)} \) is obviously unbiased for \( \tau \)
  - Invoke the Slutsky’s theorem to show that the second piece \( \xrightarrow{p} 0 \);
    intuitively this is because \( \mathbb{E}(Z_i - \bar{Z}) = 0 \)

- Augmenting unadjusted estimator by the second term, which incorporates \( X \) and thereby implements the adjustment
Model structure: $\mathbb{E}(Y_i|Z_i, X_i) = \beta_0 + \beta_z Z_i + \beta_x^T X_i$

Can show that the least squares estimator $\hat{\beta}_z$ is

$$\left\{1 - \frac{1}{N_0N_1} d_1^T \hat{\Sigma}_{XX}^{-1} d_1\right\}^{-1} \left\{\bar{Y}^{(1)} - \bar{Y}^{(0)} - \frac{N}{N_0N_1} \sum_{i=1}^{N} (Z_i - \bar{Z}) \hat{\Sigma}_{XY} \hat{\Sigma}_{XX}^{-1} X_i\right\}$$

$$d_1 = \sum_{i=1}^{N} (Z_i - \bar{Z}) X_i$$

$\hat{\Sigma}_{XX} = N^{-1} \sum_{i=1}^{N} (X_i - \bar{X})(X_i - \bar{X})^T$ the sample covariance matrix of $X$

$\hat{\Sigma}_{XY} = N^{-1} \sum_{i=1}^{N} (X_i - \bar{X})(Y_i - \bar{Y})$ sample covariance between $X$ and $Y$
Covariate Adjustment: ANCOVA1

- OLS estimator $\hat{\beta}_z$ in ANCOVA1 is

$$
\left\{1 - \frac{1}{N_0N_1} d_1^T \hat{\Sigma}_{XX}^{-1} d_1\right\}^{-1} \left\{\bar{Y}^{(1)} - \bar{Y}^{(0)} - \frac{N}{N_0N_1} \sum_{i=1}^N (Z_i - \bar{Z}) \hat{\Sigma}_{XY} \hat{\Sigma}_{XX}^{-1} X_i\right\}
$$

- Weak Law of Large Numbers, $N^{-1} d_1 = N^{-1} \sum_{i=1}^N (Z_i - \bar{Z}) X_i \xrightarrow{p} 0$

- $N^2/(N_0N_1) \rightarrow \{\pi(1 - \pi)\}^{-1}$, with $N_1/N \rightarrow \pi = P(Z = 1)$

$$
\frac{1}{N_0N_1} d_1^T \hat{\Sigma}_{XX}^{-1} d_1 = \frac{N^2}{N_0N_1} (N^{-1} d_1)^T \hat{\Sigma}_{XX}^{-1} (N^{-1} d_1) \xrightarrow{p} 0
$$

- The first term simply converges to 1
Covariate Adjustment: ANCOVA1

Augmentation term

\[
\frac{N}{N_0N_1} \sum_{i=1}^{N} (Z_i - \bar{Z}) \hat{\Sigma}_{XY} \hat{\Sigma}_{XX}^{-1} X_i \approx \sum_{i=1}^{N} (Z_i - \bar{Z}) \left( \frac{1}{N_0} + \frac{1}{N_1} \right) \Sigma_{XY} \Sigma_{XX}^{-1} X_i
\]

- \( \Sigma_{XY} = \mathbb{E}[(X - \bar{X})(Y - \bar{Y})^T] \), \( \Sigma_{XX} = \mathbb{E}[(X - \bar{X})(X - \bar{X})^T] \) true covariance matrices

- \( \hat{\beta}_z \) corresponds to \( \hat{\tau}_h \) with \( h^{(0)}(X_i) = h^{(1)}(X_i) = \Sigma_{XY} \Sigma_{XX}^{-1} X_i \)

- Regression coefficient estimator from ANCOVA1 is consistent for ATE and asymptotically normal under entirely unrestricted conditions

- This is because the OLS estimator does not assume normality, constant variance or even the correctness of the model structure!!
Covariate Adjustment: ANCOVA2

- Model structure:
  \[
  E(Y_i|Z_i, X_i) = \beta_0 + \beta_z Z_i + \beta_x^T (X_i - \bar{X}) + \beta_{xz}^T (X_i - \bar{X}) Z_i 
  \]

- Run OLS regression on centered covariates

- ANCOVA2 may seem inappropriate for estimating ATE \( \tau \), as the interaction implies that the conditional effect depends on the \( X \) and hence cannot equal \( \tau \)

- But the OLS estimator \( \hat{\beta}_z \) is again a special case of \( \hat{\tau}_h \) and consistent to ATE under entirely unrestrictive conditions
Covariate Adjustment: ANCOVA2

- OLS estimator of $\beta_z$ in ANCOVA2 is

$$
\left\{1 - \frac{1}{N_0N_1} d_2^T D^{-1} d_2\right\}^{-1} \left\{\bar{Y}^{(1)} - \bar{Y}^{(0)} - \frac{N}{N_0N_1} d_2^T D^{-1} \begin{pmatrix} \hat{\Sigma}_{XY} \\ \hat{\Sigma}_{XYZ} \end{pmatrix} \right\}
$$

- $d_2 = \{d_1^T, \sum_{i=1}^N (Z_i - \bar{Z})^2 (X_i - \bar{X})^T\}^T$

- $\hat{\Sigma}_{XYZ} = N^{-1} \sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y})(Z_i - \bar{Z})$

- $D = \begin{pmatrix} \hat{\Sigma}_{XX}^{(0)} & \hat{\Sigma}_{XX}^{(1)} \\ \hat{\Sigma}_{XX}^{(1)} & \hat{\Sigma}_{XX}^{(2)} \end{pmatrix}$ with

$$
\hat{\Sigma}_{XX}^{(l)} = N^{-1} \sum_{i=1}^N (Z_i - \bar{Z})^l (X_i - \bar{X})(X_i - \bar{X})^T
$$

- As $N^{-1} d_2 \xrightarrow{p} 0$, the first term (inside the inverse) converges to 1
Covariate Adjustment: ANCOVA2

Can show that augmentation term

\[
\frac{N}{N_0 N_1} d^T D^{-1} \left( \begin{array}{c} \hat{\Sigma}_{XY} \\ \hat{\Sigma}_{XYZ} \end{array} \right)
\approx \sum_{i=1}^{N} (Z_i - \bar{Z}) \left( \frac{1}{N_0} + \frac{1}{N_1} \right) \left\{ \pi \Sigma_{XY}^{(0)} + (1 - \pi) \Sigma_{XY}^{(1)} \right\}^T \Sigma_{XX}^{-1} X_i
\]

- \( \Sigma_{XY}^{(l)} = \mathbb{E}[(X - \bar{X})(Y - \bar{Y})^T | Z = l] \), for \( l = 0, 1 \) is the conditional covariance

- This clearly shows that \( \hat{\beta}_z \) corresponds to \( \hat{\tau}_h \) with
  \( h^{(0)}(X_i) = h^{(1)}(X_i) = \left\{ \pi \Sigma_{XY}^{(0)} + (1 - \pi) \Sigma_{XY}^{(1)} \right\}^T \Sigma_{XX}^{-1} X_i \)

- ANCOVA2 is consistent for \( \tau \) and asymptotically even if the model structure is an incorrect representation of \( \mathbb{E}(Y|Z, X) \)
Comparing ANCOVA1 and ANCOVA2

- Regression coefficient both consistent to ATE

- Can write the augmentation term of ANCOVA1,
  \[ h^{(0)}(X_i) = h^{(1)}(X_i) = \Sigma_{XY}^T \Sigma_{XX}^{-1} X_i = \{(1 - \pi)\Sigma_{XY}^{(0)} + \pi \Sigma_{XY}^{(1)}\}^T \Sigma_{XX}^{-1} X_i \]

- Compare with the augmentation term in ANCOVA2,
  \[ h^{(0)}(X_i) = h^{(1)}(X_i) = \{\pi \Sigma_{XY}^{(0)} + (1 - \pi)\Sigma_{XY}^{(1)}\}^T \Sigma_{XX}^{-1} X_i \]

- ANCOVA1 and ANCOVA2 are asymptotically equivalent equal randomization \( \pi = 1/2 \)!
Comparing ANCOVA1 and ANCOVA2

- For general cases, ANCOVA2 has smaller asymptotic variance than ANCOVA1.
- By appealing to the semiparametric efficiency theory, ANCOVA2 estimator $\hat{\beta}_z$ has the smallest asymptotic variance among the class of estimators $\hat{\tau}_h$ for which $h^{(l)}(X_i)$ are linear functions of $X_i$.
- More interestingly, ANCOVA2 is asymptotically equivalent to propensity score weighting estimators (later).
- Cautionary note is that all of these interesting results are established under linear models.


