Missing data in RCTs

Rasmus Waagepetersen

October 21, 2025

Outline:

1. MCAR, MAR and MNAR
2. Likelihood estimation under MAR
3. Case study: TRACK test results
4. Missing data due to drop out and MMRM

Never, never forget.... CLT

If X_1, X_2,... are *iid* with mean μ and variance σ^2 then

$$\lim_{n\to\infty} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu) \to N(0, \sigma^2)$$

where the convergence is in distribution.

This implies for large n,

$$\frac{1}{n} \sum_{i=1}^{n} X_i \approx N(\mu, \sigma^2/n)$$

Conditions can be relaxed, e.g. weakly dependent instead of independent, different variances and means,...

Applications of CLT

- asymptotic normality of empirical variance estimate (using influence functions) (influence function $\phi(X_i) = (X_i - \mu)^2 - \sigma^2$).
- asymptotic normality of treatment effect in RCT (see tutorial).

  Given *iid* observations $(Y_i, X_i) \in \mathbb{R} \times \mathbb{R}^p$, influence function for regression parameter β is $\phi((Y_i, X_i)) = (\mathbb{E}X_1^T X_1)^{-1}(Y_i - X_i\beta)$.

Donald Rubin's framework for missing data

Consider an $n \times 1$ data vector Z. This could in general consist of both outcome variables and explanatory variables. Further let R be a binary vector with $R_i = 1[Z_i \text{ not missing}]$. That is, we observe those Z_i for which $R_i = 1$.

We use the short hand notation $Z_r = (Z_i)_{i:r_i=1}$ and $\neg r = (1 - r_i)_{i=1}^n$. Then Z_r and $Z_{\neg r}$ represent the observed and unobserved values.

One option is to model joint distribution of (Z, R) and base inference on likelihood $p(z, r)$. Then, we can hope that usual good properties of MLE are valid.

However, in practice we have often only specified a model $p(z)$ for Z and would rather not have to model $p(r|z)$

Can we avoid modeling missingness mechanism?

Rubin's classification of missingness

Consider $P(R = r|Z = z)$. If

- $P(R = r|Z = z) = P(R = r)$ (R independent of Z) then data are said to be missing completely at random (MCAR).
- $P(R = r|Z = z) = P(R = r|Z_r = z_r)$ then data are missing at random (MAR).
- Otherwise data are missing not at random (MNAR).

(of course MCAR implies MAR)

Examples

Suppose we flip a coin to decide whether Z_i is observed or not. Then we have MCAR with $P(R = r) = 0.5^n$. The observed data represents a representative sample of the total data set.

Suppose we for an indidual have $Z = (Y, X_1, X_2)$ for an outcome variable Y and covariates X_1 and X_2 where X_2 is sometimes missing whereas Y and X_1 are always observed. Thus $R_1 = R_2 = 1$ always. If $P(R_3 = 0|Y = y, X_1 = x_1, X_2 = x_2) = P(R_3 = 0|Y = y, X_1 = x_1)$ then we have MAR (note this is equivalent to that R_3 is conditionally independent of X_2 given (Y, X_1)).

Suppose in the previous example that $P(R_3 = 1|Y = y, X_1 = x_1, X_2 = x_2) = \frac{\exp(x_2)}{1+\exp(x_2)}$. Then data is MNAR (the probability that an observation is missing depends on the observation itself). This obviously results in a biased sample since small values of X_2 are less likely to be observed.

MAR condition

On previous slide we considered example where MAR follows from conditional independence. In general MAR *is not conditional independence*.

Note that in $P(R = r|Z_r = z_r)$, conditioning on right hand side of $|$ actually depends on outcome r on left hand side, which is a bit weird. We are allowed to make the assumption that $P(R = r|Z_r = z_r)$ only depends on r and z_r but it is in general not conditional independence.

MAR is a minimal condition that allows us to base likelihood inference on the observed data z_r - next slide.

Maximum likelihood estimation under MAR

Consider the joint density of the observed data (r, z_r):

$$f(r, z_r) = \int f(z, r) \mathrm{d}z_{\neg r} = \int P(R = r|Z = z) f(z) \mathrm{d}z_{\neg r}$$

(assuming here for convenience that z is a continuous random vector with density $f(z)$. We here use convenient but a bit sloppy generic notation for densities)

Under MAR we have

$$f(r, z_r) = P(R = r|Z_r = z_r) \int f(z) \mathrm{d}z_{\neg r} = P(R = r|Z_r = z_r) f(z_r)$$

Suppose the data generating mechanism depends on a parameter $\theta = (\psi, \xi)$ that lives in the product space $\Theta = \Psi \times \Xi$ (separability) such that
$P(R = r|Z_r = z_r; \theta) = P(R = r|Z_r = z_r; \psi)$ and $f(z; \theta) = f(z; \xi)$.
Then the missing data mechanism is *ignorable* for inferring the parameter ξ governing the distribution of the data vector Z.

That is we can ignore the factor $P(R = r|Z_r = z_r; \psi)$ and just use the marginal density (*"observed data"* likelihood)

$$f(z_r; \xi)$$

for inference regarding ξ. Usual nice properties for MLE hold for resulting estimate

$$\hat{\xi} = \underset{\xi}{\text{argmax}}\, f(z_r; \xi)$$

However, the integration needed to obtain $f(z_r; \xi)$ may not be trivial (EM-algorithm, numerical integration, Monte Carlo,...)

Verifiability of MAR?

Unfortunately, MAR can only be an assumption. To verify it empirically we would need to compare $P(R = r|Z = z)$ and $P(R = r|Z_r)$ which is not possible since $Z_{\neg r}$ is not observed.

We can also in general not disprove MAR empirically for the same reason.

For a specific data set we may argue that MAR holds/does not hold based on subject matter knowledge or by imposing further assumptions that allow us to study MAR based on the data.

We can try to disprove more strict assumption MCAR - e.g. if one covariate is always observed we can compare its distribution among subjects with missing values and subjects without missing values.

Back to RCT

Consider again the RCT where we have two observations for each individual (Y_b, Y_e). Without any subject matter knowledge regarding the reasons for possible missingnes of Y_b or Y_e we can not tell whether MAR is valid or not.

If MAR is valid and our linear mixed model is valid too, then we can simply use the marginal likelihood of the observed outcomes which is available in closed form. Same holds in case of baseline adjusting if only endpoint measurements are missing. Otherwise one migh try to impute possible missing baseline values.

MAR and linear mixed model

Consider a baseline and endpoint variable (Y_b, Y_e) and assume the linear mixed model

$$Y_b = \mu + U + \epsilon_b \quad Y_e = \mu + \gamma + U + \epsilon_e$$

where U, ϵ_b and ϵ_e are independent zero-mean normal variables.

Let $R = (R_b, R_e)$ indicate whether Y_b and Y_e are observed or not.

It is now reasonable to assume that possible dependence between R and $Y = (Y_b, Y_e)$ arise through dependence on U. E.g. if Y represents math test results and U represents the math ability of a student then it might be the case that lower performing students have a higher probability of missing a test.

We assume that R conditionally independent of (Y_b, Y_e) given U.

We don't have MAR since:

$$P(R = (1,0)|Y_b, Y_e) = \mathbb{E}_{U|Y_b,Y_e} P(R = (1,0)|U)$$

which in general differs from

$$P(R = (1,0)|Y_b) = \mathbb{E}_{U|Y_b} P(R = (1,0)|U)$$

We could also consider the triple (Y_b, Y_e, U) and $R = (R_b, R_e, R_U)$ in which case $R_u = 0$ always. Then

$$\begin{aligned} P(R = (1,0,0)|Y_b, Y_e, U) &= P(R_b = 1, R_e = 0|Y_b, Y_e, U) \\ &= P(R_b = 1, R_e = 0|U) \neq P(R_b = 1, R_e = 0|Y_b) \end{aligned}$$

so same conclusion

Case study: missing TRACK test results

For each test type (fraction, arithmetic, problem) $\sim$25% missing tests.

Concern: is missingness of test result correlated with student performance ?

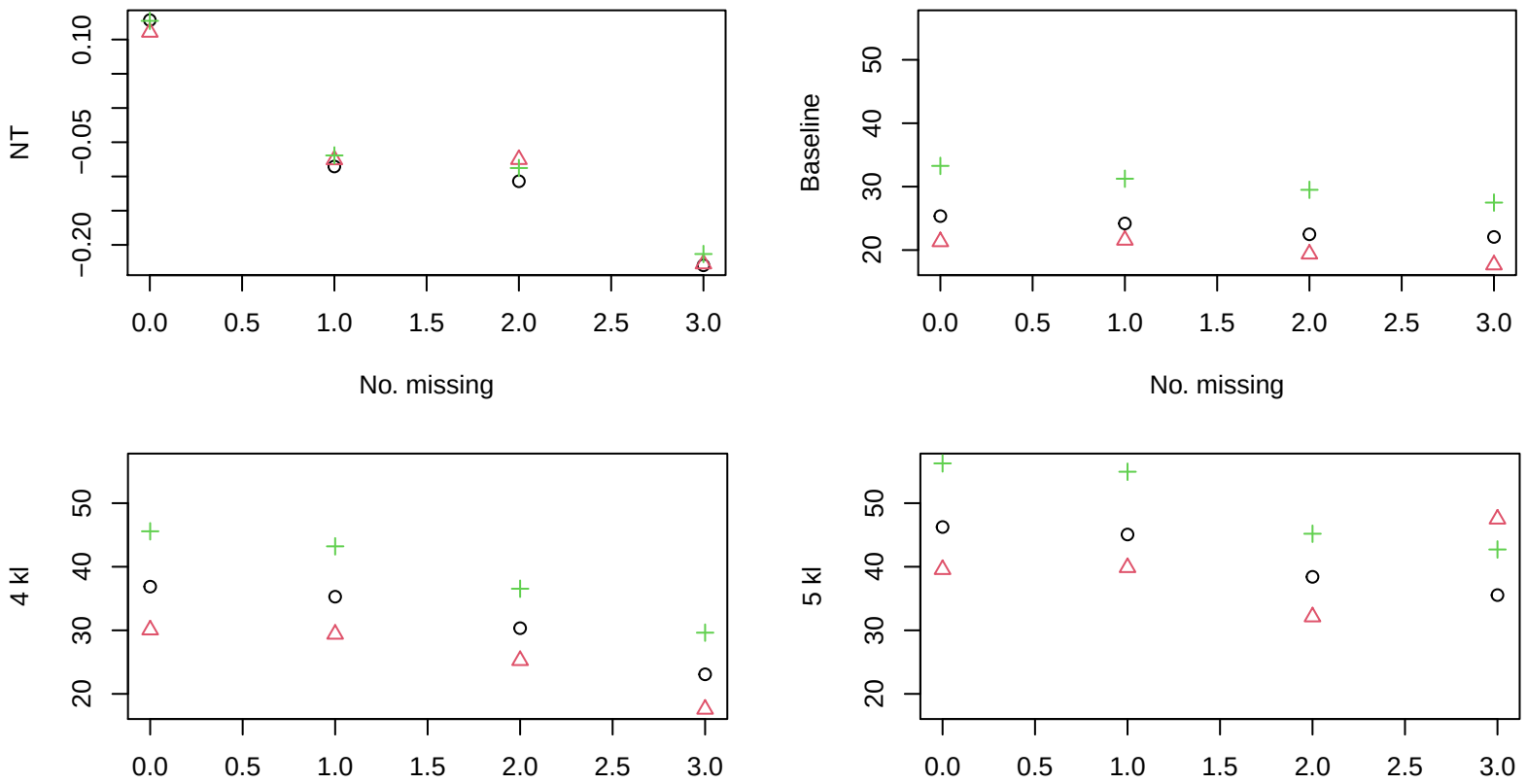
16 possible missing test patterns for a student when considering baseline, 4th, 5th for one type of test and including also national test:

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
NT									M	M	M	M	M	M	M	M
Base					M	M	M	M					M	M	M	M
4th			M	M			M	M			M	M			M	M
5th		M		M		M		M		M		M		M		M

No. missing:	0	1	2	3	4
Frequency:	50%	26 %	11 %	10 %	3 %

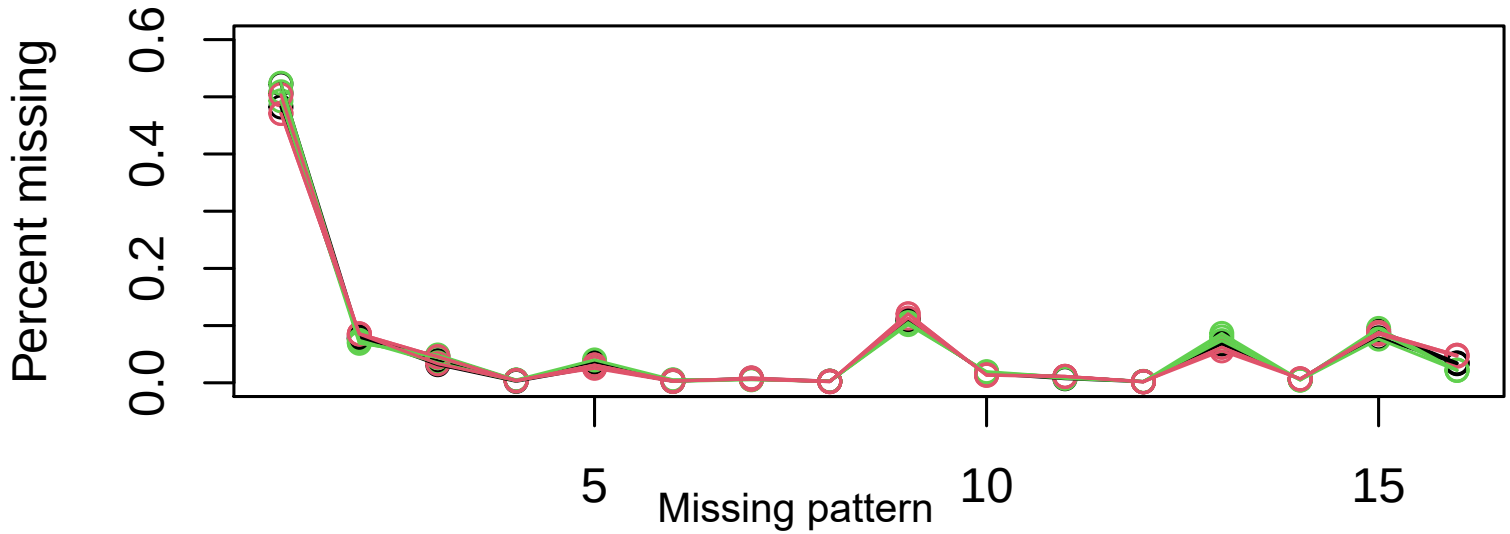
Missing data versus test score

For each number 0,1,2,3 of missing tests for a student, mean scores for tests not missing - each test time (national, base, 4th, 5th) and type (fraction,arithmetic, problem):



Frequencies of 16 missing test patterns - intervention and control

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
NT									M	M	M	M	M	M	M	M
Base					M	M	M	M					M	M	M	M
4th			M	M			M	M			M	M			M	M
5th		M		M		M		M		M		M		M		M



(Black: all data, red: intervention, green: control)

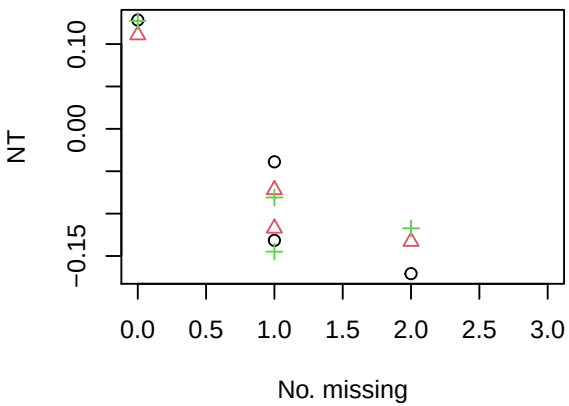
Similar missing frequencies for intervention and control

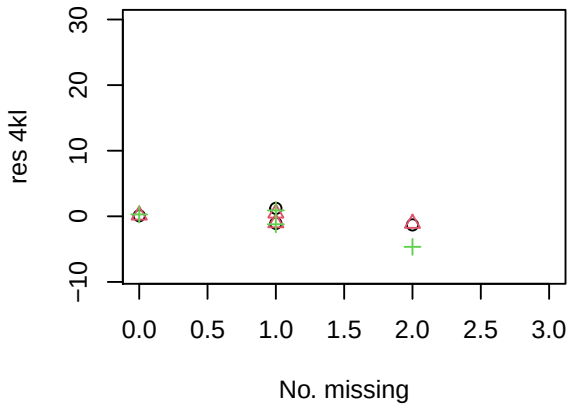
Helpful if same missingness mechanism in intervention and control (exercise)

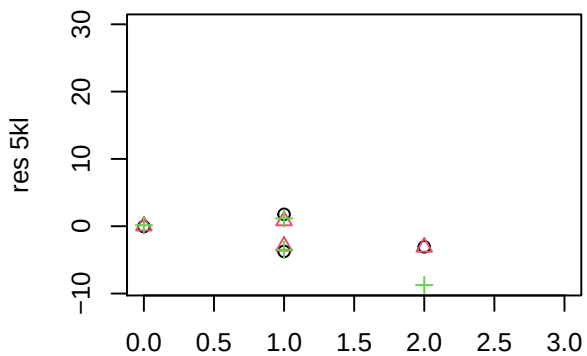
Missing data and residuals from baseline-adjusted analyses

Mixed models or baseline adjusted analyses: estimate treatment effect from residuals after adjusting for baseline - remove student heterogeneity correlated with missingness.

Mean scores for national test and 4th+5th grade residuals:







Differencing

Differencing $\Delta = Y_e - Y_b$ can be viewed as baseline adjusting with regression coefficient 1.

Under the previously mentioned mixed model, Δ is free of student ability U. Hence if dependence between missingness R and Y is due to common dependence on U we may have MCAR after differencing.

Similar effect may be in place for baseline adjustment and mixed model analysis.

Missingness due to dropout and MMRM

Missing data due to dropout happens for longitudinal data (repeated measurements on a subject) when a subject drops out of a study and all measurements after drop out are missing.

This can be due to inter current events (ICE). E.g. a patient's condition deteriorates and rescue medication has to be used.

E.g. if we consider a sequence $Z = (Z_1, \ldots, Z_n)$ of outcome variables recorded at times $t_1 < t_2 < \cdots < t_n$ then if dropout time is $D = d$ we have $R_1 = R_2 = \cdots = \mathbb{R}_{d-1} = 1$ and $R_d = R_{d+1} = \ldots, R_n = 0$.

We have MAR if

$$P(D = d|Z) = P(D = d|Z_1, \ldots, Z_{d-1})$$

Simple example

$Z = (Z_1, Z_2)$. C is intercurrent event. If $C = 1$ then Z_2 is not observed ($R_2 = 0$) and $R_1 = 1$ always. Thus $R = (1, 1 - C)$ and $D = 3 - C$.

Write

$$p(z_1, z_2, c) = p(z_2|z_1, c)p(c|z_1)p(z_1)$$

with no conditional/unconditional independencies.

Then easy to check that not MAR (exercise).

If conditional independence of C and Z_2 given Z_1,

$$p(z_1, z_2, c) = p(z_2|z_1)p(c|z_1)p(z_1)$$

then MAR. But conditional independence may not be valid.

Mixed model for repeated measurements (MMRM)

Let $Z_1 \sim N(\beta_1, \sigma^2)$ and

$$Z_2|Z_1 = z_1 \sim N(\beta_2 + \rho(z_1 - \beta_1), \sigma^2(1 - \rho^2))$$

Then product $p(z_2|z_1)p(z_1)$ equivalent to likelihood for bivariate normal

$$N\left((\beta_1, \beta_2)^\top, \sigma^2 \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right)$$

- simple example of MMRM.

We want to estimate endpoint mean β_2 (see more details on next slide).

We will investigate simple approach and observed data likelihood approach.

Structural causal model

In practice we observe C that could be 0 or 1 but we want to estimate mean of Z_2 in counterfactual 'world' where $C = 0$ always.

Full SCM:

$$\begin{aligned} Z_1 &\sim N(\beta_1, \sigma^2) \\ C|Z_1 = z_1 &\sim \text{logistic regression (for example)} \\ Z_2|Z_1 = z_1, C = c &\sim N(\beta_2 + \Delta c + \rho(z_1 - \beta_1), \sigma^2(1 - \rho^2)) \end{aligned}$$

Then previous MMRM is precisely the distribution of (Z_1, Z_2) if we fix $C = 0$ ($\text{do}(C = 0)$) in above SCM model.

By backdoor formula

$$\mathbb{E}_{\text{do}(C=1)}[Z_2] = \mathbb{E}_{Z_1}\mathbb{E}[Z_2|C = 0, Z_1] = \mathbb{E}_{Z_1}[\beta_2 + \rho(Z_1 - \beta_1)] = \beta_2$$

so β_2 is indeed our estimand of interest !

Naive approach is just take average of endpoint variables Z_{2i} for which $C_i = 0$, $\hat{\beta}_2 = \sum_{i=1}^{n} Z_{2i}(1 - C_i) / \sum_{i=1}^{n}(1 - C_i)$.

Apply law of large numbers:

$$\frac{1}{n}\sum_{i=1}^{n}(1 - C_i) \to P(C_1 = 0) := p_0$$

Further,

$$\begin{aligned}&\mathbb{E}[Z_{2i}(1 - C_i)] = \mathbb{E}[(1 - C_i)\mathbb{E}[Z_{2i}|C_i, Z_{1i}]] \\ =&\mathbb{E}[(1 - C_i)(\beta_2 + \Delta C_i + \rho(Z_{1i} - \beta_1)] = \beta_2 p_0 + \rho\mathbb{E}[(1 - C_i)(Z_{1i} - \beta_1)]\end{aligned}$$

The latter expectation is only zero if $\rho = 0$ or C_i and Z_{1i} uncorrelated.

Likelihood-based approach

Observed data likelihood is

$$\prod_{c_i=1} p(z_{i1}) \prod_{c_i=0} p(z_{2i}|z_{1i})p(z_{1i})$$

Let's assume β_2 is only unknown parameter. Then log likelihood is equivalent to

$$\sum_{c_i=0} \log p(z_{2i}|z_{1i}) \equiv \sum_{i=1}^{n}(1-c_i)(z_{2i}-\beta_2-\rho(z_{1i}-\beta_1))^2$$

Differentiating wrt to β_2, setting equal to zero and solving for β_2 we obtain

$$\tilde{\beta}_2 = [\sum_{i=1}^{n}(1-c_i)]^{-1} \sum_{i=1}^{n}(1-c_i)(z_{2i}-\rho(z_{1i}-\beta_1))$$

Proceeding as on a previous slide we obtain

$$\mathbb{E}(1 - C_i)(Z_{2i} - \rho(Z_{1i} - \beta_1) = \mathbb{E}[1 - C_i]\beta_2 = p_0\beta_2$$

Thus estimate $\tilde{\beta}_2$ is consistent for β_2!

Note: we are again effectively using baseline adjustment - i.e. using observed residuals

$$(1 - c_i)(z_{2i} - \Delta c_i - \rho(z_{1i} - \beta_1)) = (1 - c_i)(z_{2i} - \rho(z_{1i} - \beta_1))$$

Exercises

1. Consider the case

$$P(R = (0,0)|Z_1 = z_1, Z_2 = z_2) = c$$
$$P(R = (1,0)|Z_1 = z_1, Z_2 = z_2) = a(z_1)$$
$$P(R = (0,1)|Z_1 = z_1, Z_2 = z_2) = b(z_2)$$
$$p(R = (1,1)|Z_1 = z_1, Z_2 = z_2) = 1 - c - a(z_1) - b(z_2)$$

   1.1 Is this MAR?
   1.2 Does there exist any conditional independence relation for this model ?

   Hint: or three random variables X, Y, Z with joint density $p(x, y, z)$, X and Y are conditionally independent given Z if and only if there exists a factorization $p(x, y, z) = f(x, z)g(y, z)$ for some functions f and g and all (x, y, z).

2. Assume endpoint observations in a randomized trial are modelled as

$$Y_i = \mu + \psi 1[A_i = 1] + \epsilon_i$$

and R_i is the indicator of whether Y_i is observed or not. We assume that the (Y_i, A_i, R_i) are *iid*.

Can we use the simple averages of non-missing observations in each group to consistently estimate ψ ? Which conditions would be sufficient for this?

Hint: the average of treated non-missing observations can be written as
$\mu + \psi + \sum_{i=1}^{n} 1[A_i = 1]\epsilon_i R_i / (\sum_{i=1}^{n} 1[A_i = 1]R_i)$. Then use law of large numbers.

Note: missingness happens after randomization so we can not in general assume R_i independent of A_i.

3. Would it be reasonable to assume missingness independent of treatment allocation in TRACK experiment?
4. Check that we indeed do not have MAR on slide 20