Introduction to Missing Data Part 1

Jarvis T. Chen (jarvis@hsph.harvard.edu)

10 April 2023

Department of Social and Behavioral Sciences
Harvard T. H. Chan School of Public Health

Developing data intuition

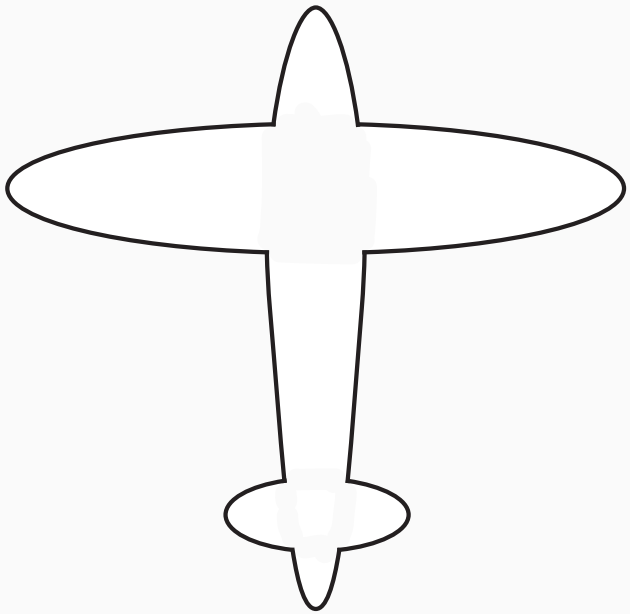
- studying human beings is complicated
    - rarely are we in a position to do tightly controlled experiments
    - especially in observational studies, we have to collect and analyze data on multiple variables to understand a phenomenon well
- we need to develop good intuition about biases and threats to the validity of our inferences

Selection and selection bias

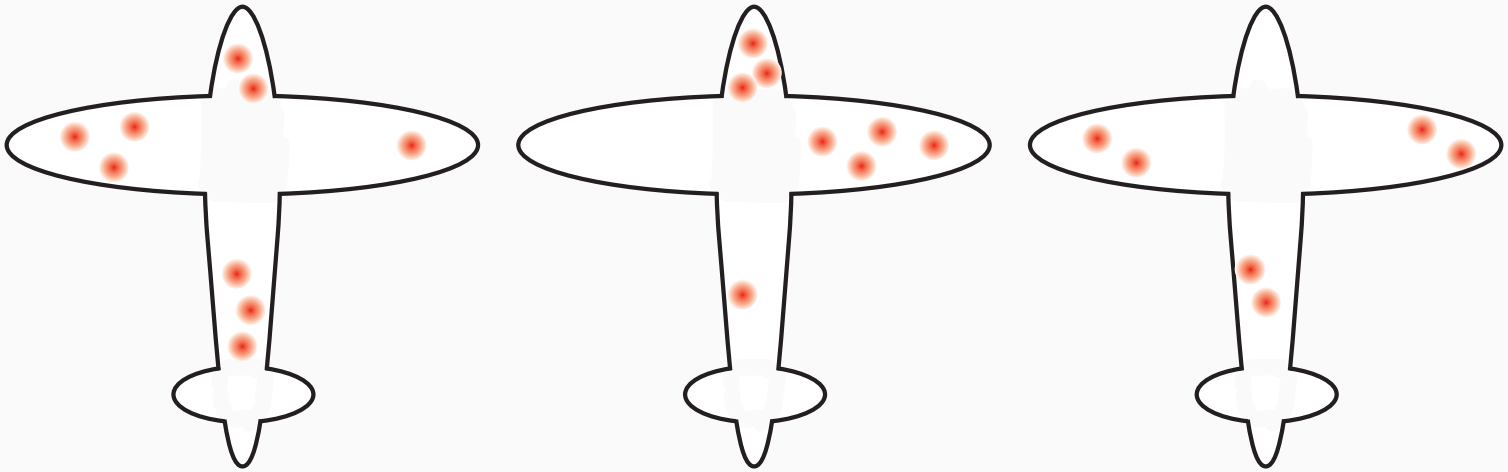
Selection can occur when the study sample we have observed is not representative of the target population from which it was drawn. This *can* lead to bias in estimates of causal effects.

- it's useful to distinguish **selection** from **selection bias**
  - **selection** refers to the non-representativeness of the observed cases available for analysis
  - as we discussed in PHS2000A, **bias** needs to be considered in relation to whether a particular estimator provides a consistent estimate of an estimand of interest
- Note that there are differences in how economists and epidemiologists use the term **"selection"**
  - "selection into treatment"
  - "selection into the study population"

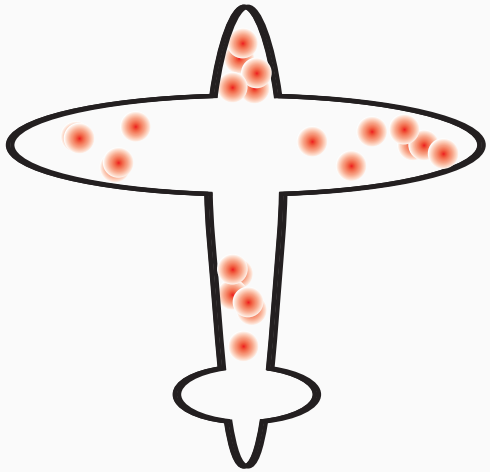
An intuitive example of selection bias



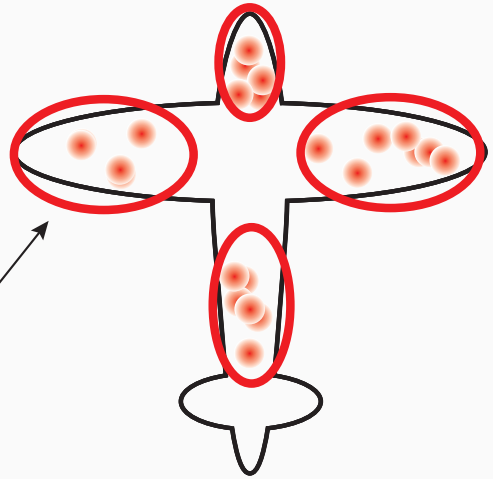
An intuitive example of selection bias



An intuitive example of selection bias

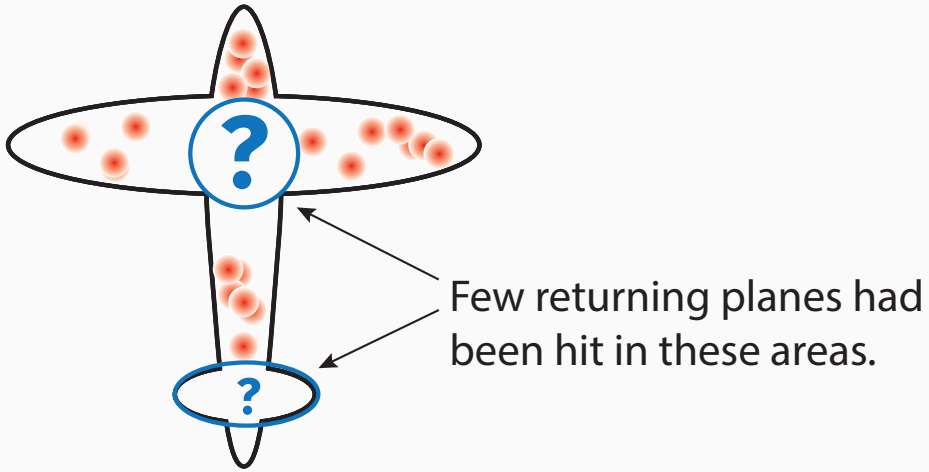


An intuitive example of selection bias

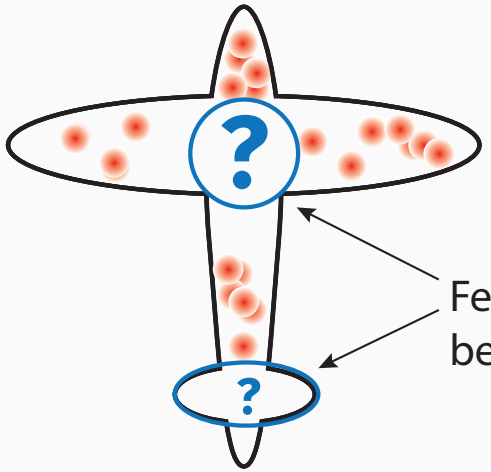


Initial conclusion:
reinforce the armor in these most frequently hit areas

An intuitive example of selection bias



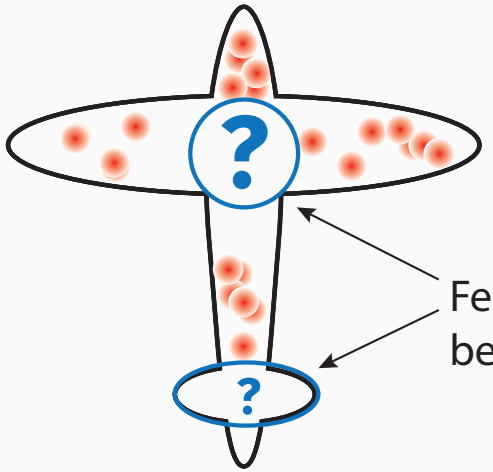
An intuitive example of selection bias



Few returning planes had been hit in these areas.

...but what about planes that did not return?

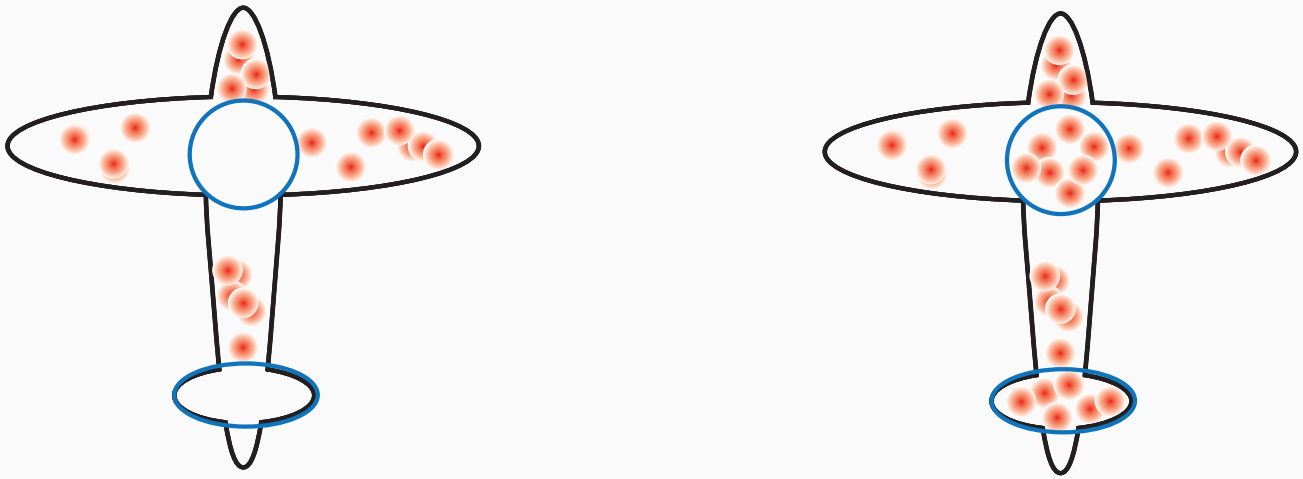
An intuitive example of selection bias



Few returning planes had been hit in these areas.

Abraham Wald recommended reinforcing these areas.
The British RAF followed suit and markedly fewer casualties were observed after the changes were made.

An intuitive example of selection bias



Observed

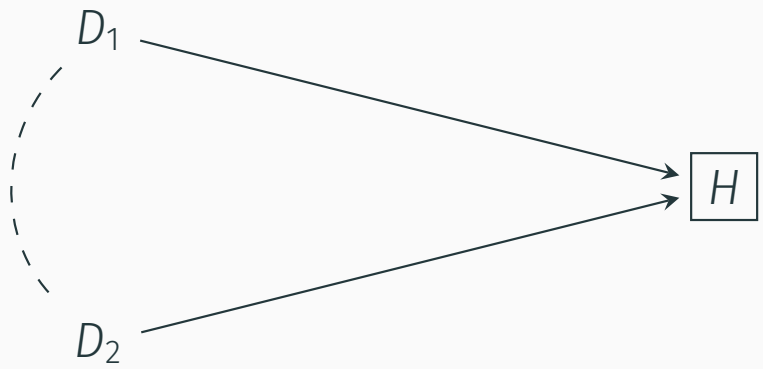
Unobserved

Key intuition: the wrong conclusion was made because observation was restricted only to planes that made it back.

The distribution of bullet holes on plane bodies is uniform, but because hits to the tail or the center part of the plane are catastrophic, planes that are observed tend not to have bullet holes in these places.

Berkson's fallacy

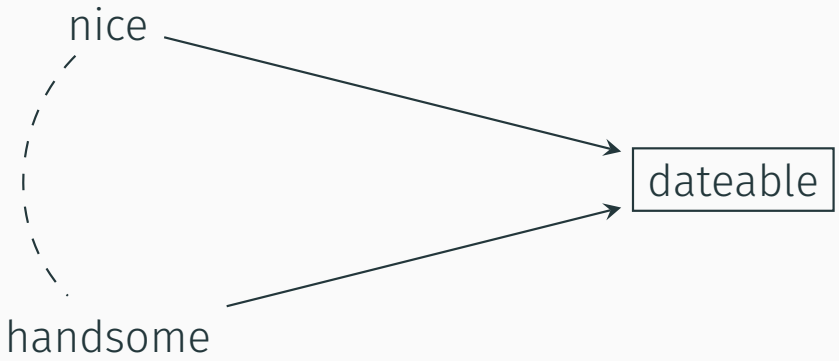
Berkson's fallacy: when two diseases, D_1 and D_2, that lead to hospitalization (H) are observed *among hospital patients*, they appear correlated even though they are independent.



Berkson's original example: is diabetes (D_1) a risk factor for cholecystitis (D_2) in a sample from a hospital in-patient population? (What direction do you expect the association to go?)

Are handsome men more likely to be rude?

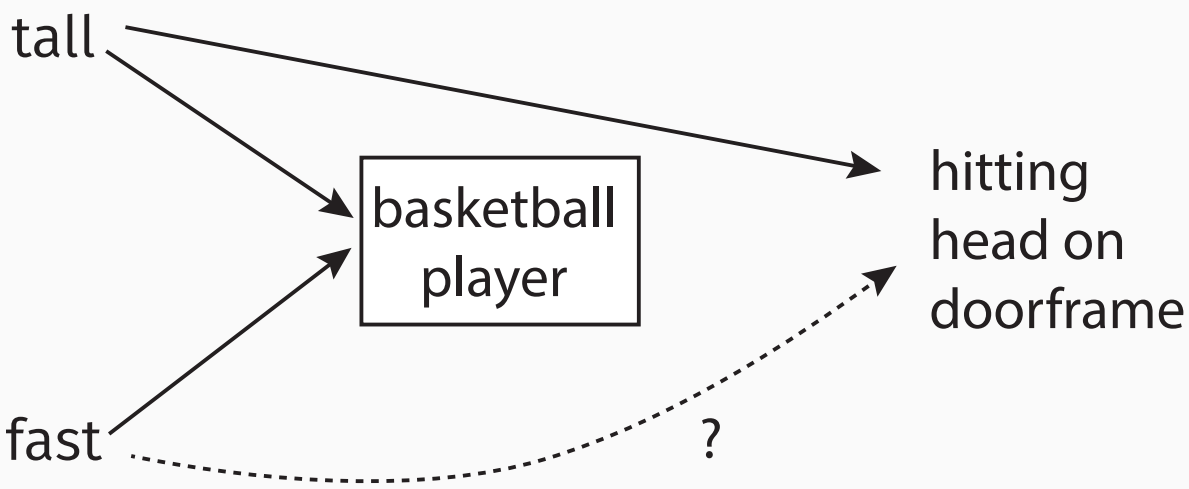
Suppose that Jarvis will only date a man if his niceness plus his handsomeness exceeds some threshold. That is, niceness and handsomeness are both predictors of being in Jarvis' dating pool.



Among the men that Jarvis dates, Jarvis may observe that the nicer ones are less handsome on average and vice versa, even if these traits are uncorrelated in the general population.

Example attributed to the mathematician Jordan Ellenberg

Another example

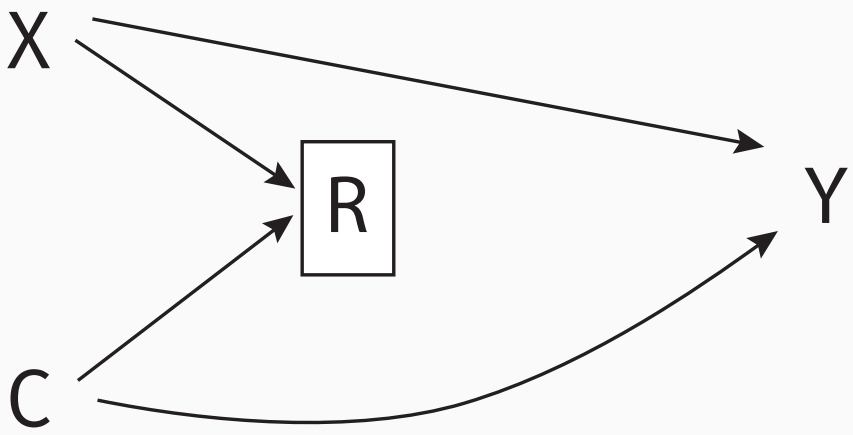


Being tall and being fast are associated with being a basketball player.

Given that you are *not* tall, if you are a basketball player, you must be fast.

Among basketball players, being fast appears to be protective against hitting your head on doorframes.

Collider stratification



Selection shares this general structure of *conditioning* on a common effect (here, restricting observation to the stratum where R=1).

Structurally, this is known as “collider-stratification bias.”

Note that this is not a situation where we have “inappropriately” conditioned on a collider in the analysis! Rather, restriction to $R = 1$ in the **design** of the study has induced collider stratification bias (the data came to us this way).

Overview

- Introduction to Missing Data Theory
    - classification of missing data (MCAR, MAR, MNAR)
    - ignorability/non-ignorability
    - other considerations
- Ad Hoc Methods
    - Complete case analysis
    - Available case analysis
    - Missing indicator
    - Dropping covariates
    - Single imputation
- Statistically principled methods
    - Weighting Methods
    - Multiple Imputation

Missing Data

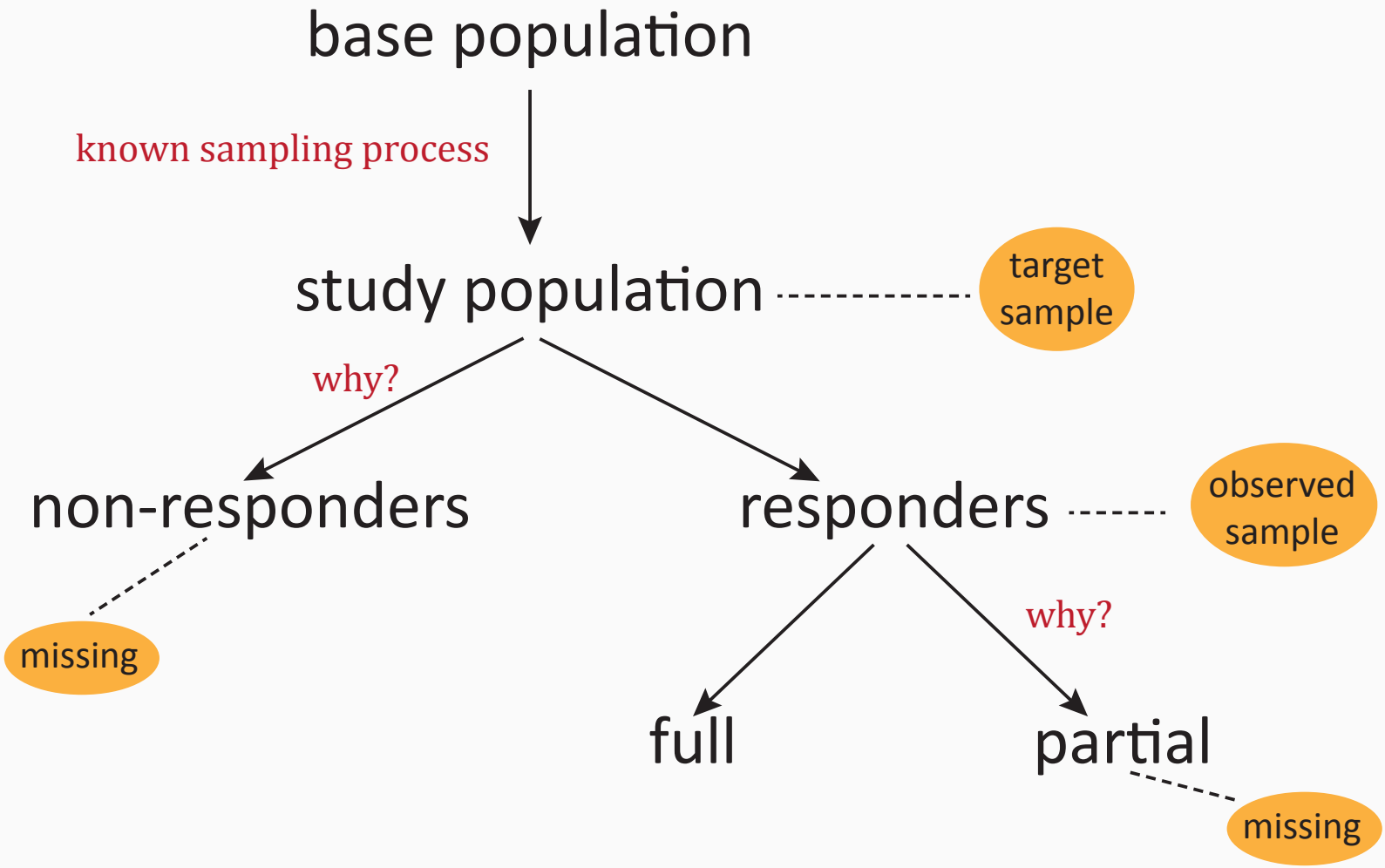
In many studies, measurements of every variable are not available for every subject. Data may be missing by chance or by design.

- unavailability of covariate measurements
- survey non-response
- some study subjects fail
  to report to a clinic for monthly evaluations
- respondents refuse
  to answer certain items on a questionnaire
- loss of data
- some information may be purposely
  suppressed, e.g. to protect confidentiality

Missing data are common

- Often inadequately handled in both observational and experimental research.
- For example, Wood et al. (2004) reviewed 71 recently published BMJ, JAMA, Lancet, and NEJM papers.
    - 89% had partly missing outcome data.
    - In 37 trials with repeated outcome measures, 46% performed complete case analysis.
    - Only 21% reported sensitivity analysis
- Sterne et al. (2009) reviewed articles using Multiple Imputation in BMJ, JAMA, Lancet, and NEJM from 2002 to 2007.
    - 59 articles found, with use doubling over 6 year period
    - However, the reporting was almost always inadequate.

How do missing data arise?



Missing data in longitudinal settings

Time	t_1	t_2	t_3	t_4	t_5	t_6	
Subject 1	y_1	y_2	y_3	y_4	y_5	y_6	*completely observed*
Subject 2	y_1	y_2	y_3	*	*	*	*lost to follow up at t_4*
Subject 3	y_1	*	y_3	y_4	*	y_6	*intermittent missingness*
Subject 4	y_1	*	*	y_4	*	*	*measurements at t_1 and t_4 by design*
Subject 5	*	*	*	y_4	y_5	y_6	*outcome not defined until t_4*

* *denotes missing value*

Consequences of missing data

- Intuitively, when the subjects with missing covariate values differ systematically from those with complete data with respect to the outcome of interest, results from a traditional data analysis omitting the missing cases may no longer be valid.
- If subjects with any missing covariates are excluded from the analysis, analyses can be **biased** and **inefficient**.

Classification of Missing Data

Following Little and Rubin (1987), missing data are generally classified into 3 types.

- Missing Completely At Random (MCAR)
- Missing At Random (MAR)
- Missing Not At Random (MNAR or NMAR)

Classification of Missing Data

- Let's assume that data are collected on a sample of n subjects and that primary interest relates to the parameters governing the conditional distribution, $f(Y_i|\mathbf{X}_i, \boldsymbol{\beta})$. To simply exposition, we suppress the subject indicator i.
- For full generality, we'll use $\mathbf{Z} = (Y, \mathbf{X})$ to denote the set of all variables of interest, including outcomes (Y) and predictors ($\mathbf{X}$).
- In the literature on missing data, papers sometimes focus on completely observed Y and partially missing $\mathbf{X}$, while other papers focus on partially observed Y and fully observed $\mathbf{X}$. The details can sometimes vary, but the substantive concepts are the same.

Classification of Missing Data

- For a given subject, we can partition $\mathbf{Z}$ into components denoting observed variables ($\mathbf{Z}^{obs}$) and those that are missing for that subject ($\mathbf{Z}^{mis}$).
- We denote by $\mathbf{R}$ a set of response indicators where $R_j = 1$ if the jth element of $\mathbf{Z}$ is observed and equals 0 otherwise.
- Little and Rubin proposed a classification for missingness in terms of probability models for $\mathbf{R}$.

Missing Completely At Random

- Data are said to be **MCAR** if the failure to observe a value **does not depend on any data, either observed or missing**.
- The missing values of **Z** are MCAR if the probability of observing **Z** is independent of the values of **Z** that are observed or would have been observed.
- Under MCAR, **the observed data are just a random sample of all the data.**

Missing Completely At Random

- A trivial example would be, suppose that in a survey, each survey respondent decides whether to answer a question on earnings by rolling a die and refusing to answer if a “6” shows up.
- Formally,

$$P(\mathsf{R}|\mathsf{Z}) = P(\mathsf{R}|\mathsf{Z}^{obs}, \mathsf{Z}^{mis}) = P(\mathsf{R}|\boldsymbol{\theta})$$

  where $\boldsymbol{\theta}$ are a set of parameters governing the probability distribution of the response indicators.
- In this trivial example, every respondent has a 1 in 6 chance of having missing data on earnings (regardless of any of their other data values or the missing earnings variable itself).

Missing At Random

- The data are said to be MAR if, conditional on the observed data, the failure to observe a value **does not depend on the data that are unobserved**. However, the conditional probability of missingness may depend on any observed data.
- At first glance, the term ”Missing At Random” may seem confusing, since missingness can actually be predicted (but is random after controlling for missingness due to observed quantities).

Missing At Random

- The missing values of $\mathbf{Z}$ are MAR if, conditional on the observed data, the probability of observing $\mathbf{Z}$ is independent of the values of $\mathbf{Z}$ that would have been observed.
- Note that this probability is not necessarily independent of the observed values of $\mathbf{Z}$.
    - Recall that $\mathbf{Z}$ includes Y and $\mathbf{X}$
    - So we could have, for example, that higher levels of $\mathbf{X}$ predict higher levels of Y and also increase the probability of not being observed. Then the values of Y that are observed are **not** a random sample of all values of Y in the population.

Missing At Random

- E.g. In a survey on earnings, gender, race, education, and age are recorded for all people in the survey. We find that the proportion of subjects missing the earnings variable varies by gender, race, education, and age, but conditional on those variables, we believe that missingness on the earnings variable is random.
- MAR is a more realistic assumption than MCAR, but usually adjustments must be made in the analysis because the subjects with fully observed data are no longer a random sample of all subjects.

Missing At Random

- Note that if the data are MCAR, then they are by definition MAR (i.e. MCAR is a special case of MAR).
- Note also that we can test whether missingness follows MCAR or MAR in a given dataset by examining the data.
- Formally,

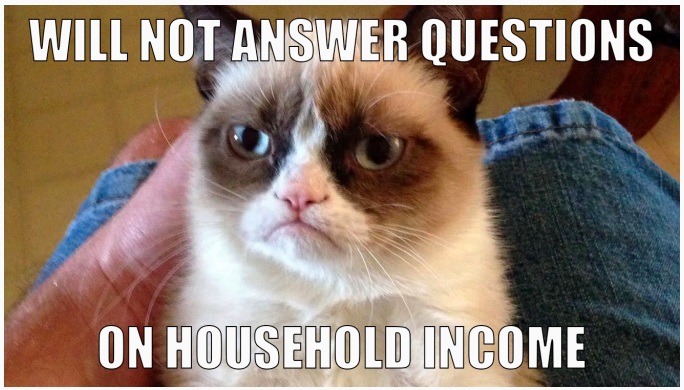
$$P(\mathbf{R}|\mathbf{Z}) = P(\mathbf{R}|\mathbf{Z}^{obs}, \mathbf{Z}^{mis}) = P(\mathbf{R}|\mathbf{Z}^{obs}, \boldsymbol{\theta})$$

Missing Not At Random

- The missing data mechanism is said to be Missing Not At Random (MNAR) if the failure to observe a value **depends on the value that would have been observed**.
- The missing values of Z are MNAR if, conditional on the observed data, the probability that Z is missing depends on the missing values of Z.
- This can occur because missingness depends on the missing value itself or because missingness depends on unobserved predictors (which in turn affect the missing value).

Missing Not At Random

- For example, suppose in a survey on earnings that people with higher earnings are less likely to reveal them. Here, the probability of missingness on earnings varies with the value of the missing variable itself.
- Alternatively, suppose that ”surly” people are less likely to respond to the earnings question, that ”surliness” is predictive of earnings, and ”surliness” is unobserved.



- Or suppose that people with college degrees are less likely to reveal their earnings, having a college degree is predictive of earnings, and there is also some non-response to the education question. Then the earnings variable is Not Missing At Random.

Missing Not At Random

- Here, the expression for the probability of missingness cannot be further simplified.

$$P(\mathbf{R}|\mathbf{Z}) = P(\mathbf{R}|\mathbf{Z}^{obs}, \mathbf{Z}^{mis})$$

  Valid inferences generally require specifying the correct model for the missing-data mechanism, distributional assumptions for the missing $\mathbf{Z}$, or both. The resulting estimators and tests are typically sensitive to these assumptions.

- **Note that we cannot generally distinguish between MAR and MNAR based on the data at hand.**

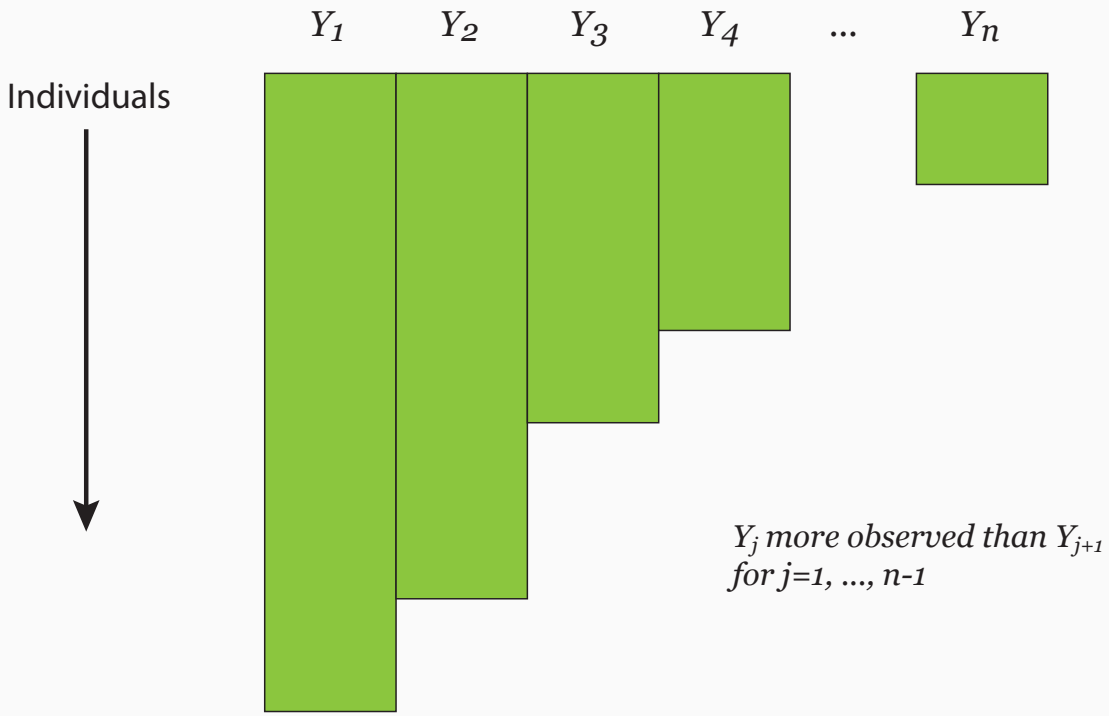
Patterns of missing data

Another important concept regarding missing data, particularly where there are multiple variables with missing values, relates to the **pattern** of missing data. If the data matrix can be rearranged in such a way that there is a hierarchy of missingness, so that observing a particular variable X_b for a subject implies that X_a is observed, for $a < b$, then the missingness is said to be **monotone**.

This has implications for some of the procedures for missing data imputation.

Dropout

Monotonic patterns of missingness are particularly relevant for longitudinal data since this is the pattern of missingness we expect to see with dropout or attrition.



(Recall that we talked about weighting methods for dealing with attrition in the context of marginal structural models)

Unit vs. Item Non-Response

In longitudinal settings, it is also useful to distinguish between **unit** non-response vs. **item** non-response.

- **unit non-response:** If subject i is not measured at time t, all of subject i's responses are missing at time t.
- **item non-response:** If subject i is measured at time t, but fails to answer one of the questions, then only that item is missing at time t.

In general, we expect to see both kinds of missingness in longitudinal datasets.

Ignorable/Nonignorable missingness

The missing data mechanism is said to be **ignorable** if

1. the missing data are MCAR or MAR and
2. the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ are **distinct.**[1]

In many situations, this is intuitively reasonable, as knowing $\boldsymbol{\beta}$ will provide little information about $\boldsymbol{\theta}$ and vice-versa.

- Ignorable means we can ignore the missingness but does not necessarily mean we can ignore the missing data! For inferences to be valid, we need to condition on those factors that influence the missingness of the data.
- If the missing data mechanism is non-ignorable, then we cannot ignore the model of missingness.

Distinct in the sense that the joint parameter space of $(\boldsymbol{\beta}, \boldsymbol{\theta})$ is the product of the parameter space of $\boldsymbol{\beta}$ and the parameter space of $\boldsymbol{\theta}$. The MAR condition is generally considered to be the more important condition here.

Ignorable/Nonignorable missingness

- Note that the distinction between ignorable and non-ignorable missingness is not just based on the data. Rather, the terms apply jointly to the data and to the analysis that is being done.
- For example, suppose one develops a smoking prevention intervention and has a treatment group and a control group.
    - Suppose that one measures smoking status at time 2, one year after implementation of the prevention intervention.
    - Suppose that some people have missing data for smoking at time 2, and that missingness on smoking at time 2 depends on smoking status measured at time 1, just before the program implementation.

Ignorabile/Non-ignorable missingness

- If one includes smoking at time 1 as a predictor of missing data in one of the acceptable missing data procedures (e.g. multiple imputation or weighting or likelihood-based methods), then the missingness on smoking at time 2 is conditioned on smoking at time 1 and is thus MAR. However, if the researcher tested a model in which treatment alone predicted smoking at time 2, then the missingness would be MNAR because the research failed to condition on smoking at time 1, the cause of missingness.

MCAR/MAR/MNAR: Additional comments

- It is tempting to think that the missingness affecting a variable always fits neatly into one of the three categories.
- From a practical viewpoint, the reality is more complex.
    - MNAR can take different forms and depart from MAR by different amounts.
    - It depends on what else has been observed!
- If enough additional information is collected, MNAR missingness can be converted (close) to MAR
    - Often used as justification for methods which assume MAR.
- However, the MAR or MNAR classification will **still be an assumption.**
    - Remember that it is not possible to distinguish between MAR and MNAR from the observed data.
    - Sensitivity analyses may be required.

Methods for partially missing data

Little and Rubin (1987) describe a broad taxonomy of methods for analyses with partially missing data (not mutually exclusive):

1. **Procedures based on completely recorded units** (complete case analysis). This is generally easy to carry out and may be satisfactory with small amounts of missing data. However, it can lead to serious biases and is not usually very efficient.
2. **Imputation-based procedures**. The missing values are "filled in" and the resulting completed data are analyzed by standard methods. These include *hot deck* imputation, where recorded units in the sample are substituted; *mean* imputation, where means from sets of recorded values are substituted; and *regression* imputation, where the missing variables for a unit are estimated by predicted values from the regression of known variables for that unit.

Methods for partially missing data

3. **Weighting procedures**. Complete cases are analyzed using weighted regressions where the weights vary inversely with the estimated probability of being fully observed. The approach is an extension of survey methodology without non-response where design weights are inversely proportional to the probability of selection. Here, the weights are modified in an attempt to adjust for non-response.
4. **Model-based procedures**. A broad class of procedures is generated by defining a model for the partially missing data and basing inferences on the likelihood under that model, with parameters estimated by procedures such as maximum likelihood. Advantages of this approach are flexibility; and the avoidance of ad hoc methods, in that model assumptions underlying the resulting methods can be displayed and evaluated.

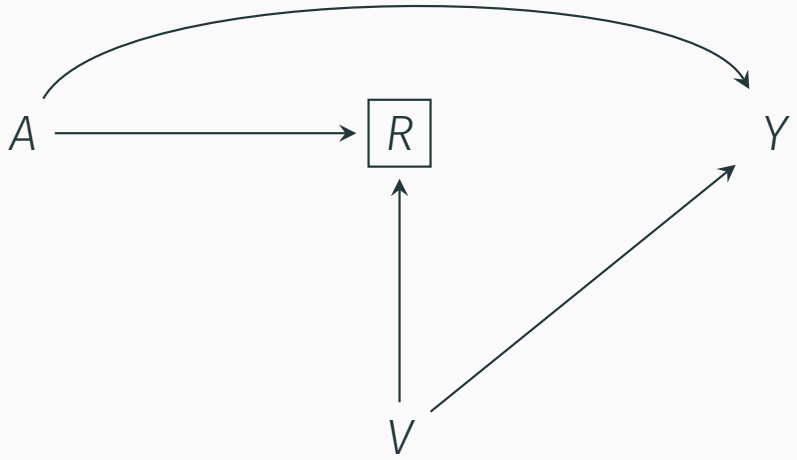
Complete Case Analysis

- missing data are ignored and only complete cases analyzed
- many computer packages do this by default
- **Advantages**: simple
- **Disadvantages**
  - often introduces bias
  - inefficient as data from incomplete cases discarded
  - the amount of discarded data can be potentially large if there are several variables that exhibit missingness.

Complete Case Analysis

- A complete case analysis may lose efficiency, but no bias is introduced when the data are MCAR.
- **In many if not most MAR scenarios, a complete case analysis will be both inefficient and biased.**
    - In data that are MAR, if missingness depends only on $\mathbf{X}$ and not on Y, then a complete case analysis will lead to unbiased estimates.
    - However, if the missingness depends on Y (and not necessarily on $\mathbf{X}$), then a complete case analysis will result in biased estimates.

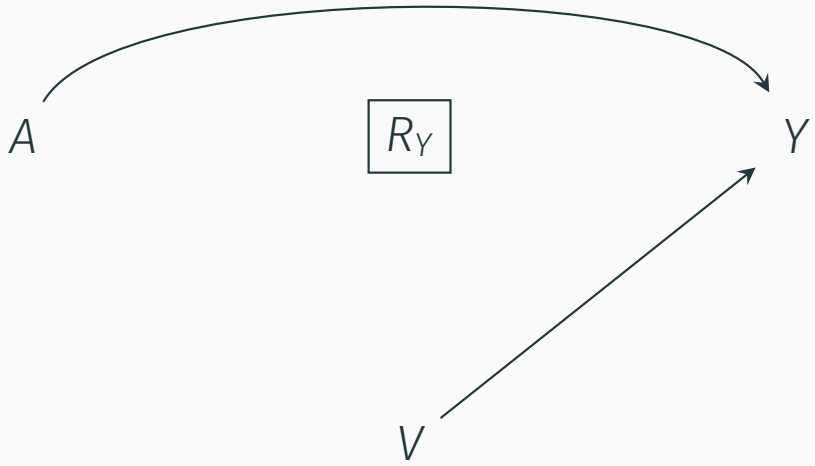
Complete Case Analysis As Selection



A **complete case analysis** is restricted to the stratum where $R = 1$ (observed cases).

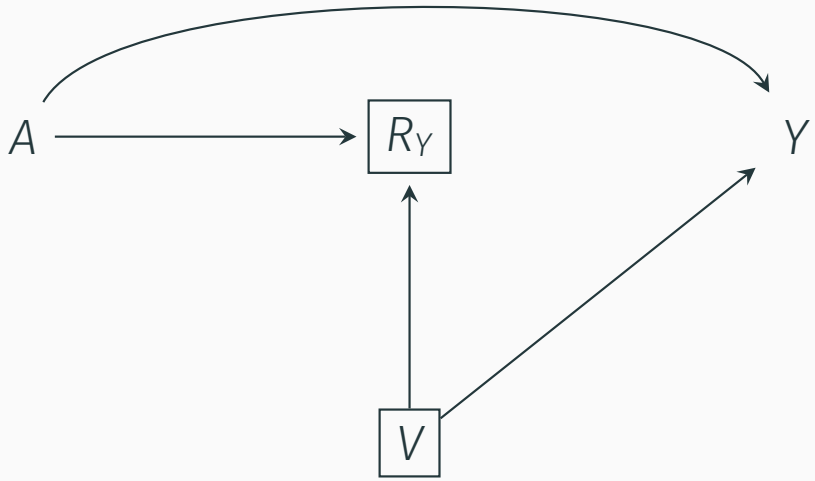
Under what conditions can we recover an unbiased estimate of the effect of A on Y?

Example: Complete Case Analysis Under MCAR



Under MCAR, observed cases are a random sample from the population, i.e. the probability of being observed is unrelated to any of the variables. Thus, an analysis restricted to complete cases can be unbiased, although depending on the amount of missing data, it may be inefficient.

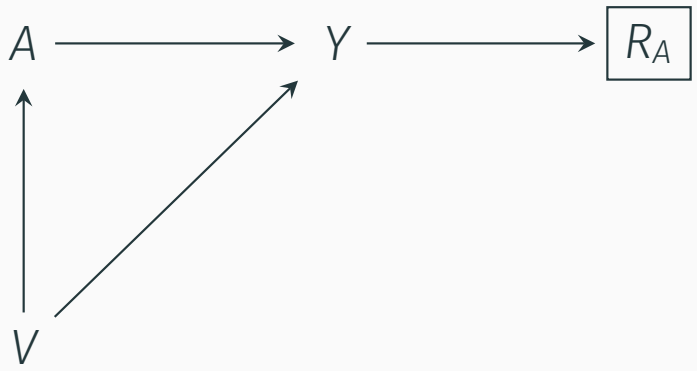
Example: Complete Case Analysis Under MAR



Under MAR, the probability of being observed can depend on observed variables (here it depends on A and V). In this simple situation, unbiased estimate of the effect of A on Y can be obtained in a complete case analysis by conditioning on V (e.g. via regression adjustment).

Note that conditioning in a complete case analysis will not solve all MAR situations; it depends on what variables are unobserved!

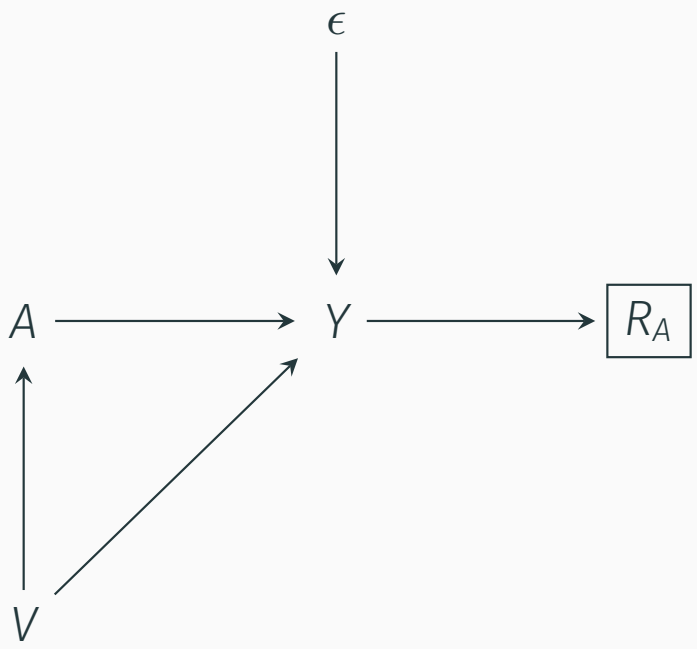
Example: Complete Case Analysis Under MAR



Consider the situation where the exposure A is partially observed and the probability of being observed depends on the outcome Y. We are interested in estimating the effect of A on Y.

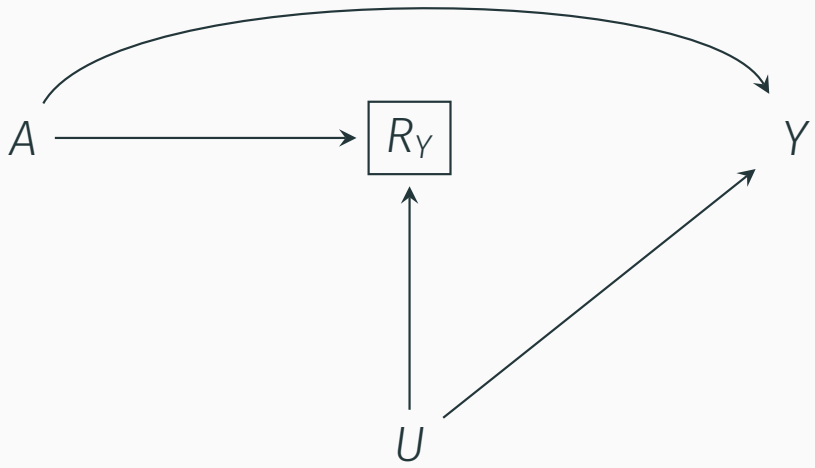
The data are MAR, but it is not immediately obvious from this DAG why regression-based estimates of the effect of A on Y will be biased.

Example: Complete Case Analysis Under MAR



Here, we explicitly depict the random error term ϵ as a "cause" of Y. How does conditioning on R_A affect the correlation between A and ϵ?

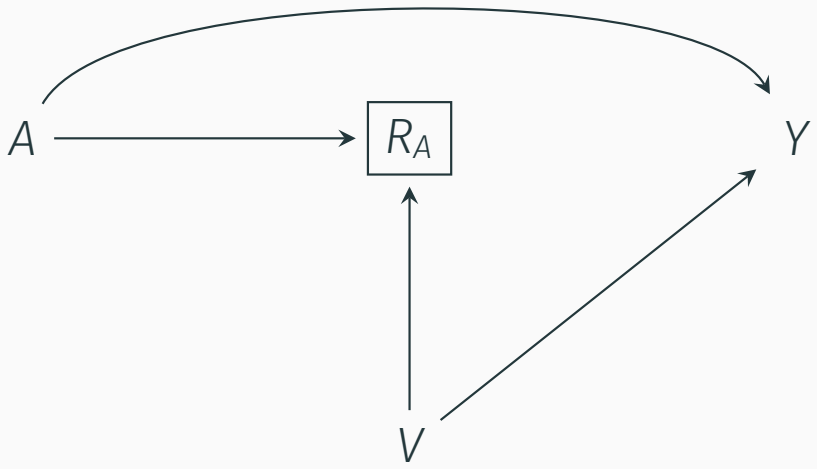
Example: Complete Case Analysis Under MNAR



Here, missingness in the outcome Y depends on U, which is unobserved.

This is an example of Missing Not At Random.

Example: Complete Case Analysis Under MNAR



Here, missingness in the exposure A depends on the values of A itself.

This is another example of Missing Not At Random.

Available case analysis (e.g. pairwise deletion)

- Available case methods include all cases where the variable(s) of interest is (are) present. This is often done in exploratory analyses of datasets where simple univariate summaries are generated based on the set of non-missing values for each variable. Similarly, correlation matrices may be computed based on fully observed data for pairs of variables.
- A **disadvantage** of this is that the sample base changes from variable to variable. Consequently, the resulting correlation matrices can be misleading. (Pairwise deletion is not recommended).

Dropping covariates

Another ad hoc approach involves dropping variables from the analysis that have a large proportion of missing values.

This is not attractive since it may lead to the exclusion of important confounders from the regression model, with consequent bias for estimating the effects of interest.

Single imputation

- Aim to ”fill in” missing values to create a single ”completed” (rectangular) dataset with imputed values, which is then analyzed using standard methods.
- Types of single imputation include
    - mean imputation
    - last observation carried forward (LOCF) (in longitudinal settings)
- Computationally simple but makes strong assumptions about missingness mechanism (usually MCAR in the case of mean imputation)
- ignores uncertainty about imputed missing values

Last observation carried forward (LOCF)

- widely used in clinical trial settings
- Assumption: all unseen measurements = last seen measurement
- is not even valid under the strong assumption of MCAR!

Missing indicator method

- For each variable with missing values, one creates a missing-value indicator to accompany the variable in all analyses. The missing value indicator takes the value 1 wherever the original variable is missing, 0 otherwise.
- For example, if X_1 is age at menarche and some subjects have X_1 missing, one creates a missing age-at-menarche indicator M_1 such that $M_1 = 1$ among subjects with X_1 missing, $M_1 = 0$ among the rest.
- One then replaces age at menarche in the regression with the missingness indicator, M_1, and the product of age at menarche and one minus the indicator, $X_1(1 - M_1)$ (note that this is zero if X_1 is missing).
- For categorical variables, this involves creating an additional category for missing values.

Missing indicator method

- Unfortunately, this method is biased under most conditions, **including under MCAR** (whereas the complete case analysis is unbiased under MCAR).
- Though the missing indicator has been discredited in the literature, it is still frequently used in practice. **We don't recommend that you use it!**
- See the simulation at the end of today's presentation to see why!

Statistically Principled Methods

The ad hoc methods mentioned above can be contrasted with **statistically principled** methods.

- In contrast to ad hoc methods, principled methods are
  - based on a well-defined statistical model for the complete data and explicit assumptions about the missing value mechanism
  - the subsequent analysis, inferences, and conclusions are valid under these assumptions
  - does not mean the assumptions are necessarily true but it does allow the dependence of the conclusions on these assumptions to be investigated.

Statistically Principled Methods

- Incorporate statistical (stochastic) information about the missing values and/or the missingness mechanism, e.g.
  - **Multiple imputation (MI)** – generate $K > 1$ imputed values for missing observations from appropriate probability distribution.
  - **Fully model based (e.g. Bayesian)** – write down statistical model for full data (including missingness mechanism) and base analysis on this model.
- **Likelihood-based methods**: based on factorization of the likelihood into the analysis model and the missing data model.

Weighting

- Another statistically principled technique involves weighted regression using complete case data where the weights are the inverse of the probability of being fully observed. This requires the assumption of MAR missingness.
- Intuitively, weighting allows us to recover the population that we would have observed if missingness had not occurred.
- We discussed this approach in the lecture on marginal structural models where we talked about inverse probability of censoring weights.

Weighting

1. Fit a model for the probability of being observed, e.g. a logistic regression model

$$\text{logit}(P(R = 1)) = \theta_0 + \theta_1 X_1 + \ldots \theta_k X_k$$

2. Predict the probability of being fully observed for every subject in the dataset, $\hat{\pi}_i$.
3. Assign weights $r_i/\hat{\pi}_i$ to each observation. (Note that if $r_i = 0$ then the weight is zero, otherwise it is the inverse probability weight).
4. Fit the weighted regression.
5. Standard errors should be based on a robust variance estimator.

Weighting

- This works well for simple patterns of missingness (e.g. one covariate with missing values). It becomes considerably more complicated for multiple missing variables, particularly when the pattern is non-monotone.
- Note that by construction, this inverse probability of weighting scheme only uses the data from the complete cases and disregards data from individuals for whom the variables are missing. Intuitively, this is likely to result in inefficiency.

Simulation

To explore some of the implications of missing data in a simple, cross-sectional setting, I've taken an existing dataset and simulated missingness under MCAR and MAR so that we can compare the performance of

- complete case analysis
- missing indicator
- weighting

in estimating an exposure-outcome relationship.

In this dataset, we have measurements of $Y = log.FEV1$, $X_1 = height$, and $X_2 = age$ among 300 subjects at the Topeka site of the Six Cities Study (a study of the effects of air pollution on human health).

Complete case analysis under MCAR and MAR

Let's begin by comparing MCAR (where the probability of missing height does not depend on other variables) to MAR (where the probability of missing height depends on age and log.FEV1). Under both scenarios, I simulated 30% missingness.

- For MCAR, $P(R = 0) = 0.3$.
- For MAR,

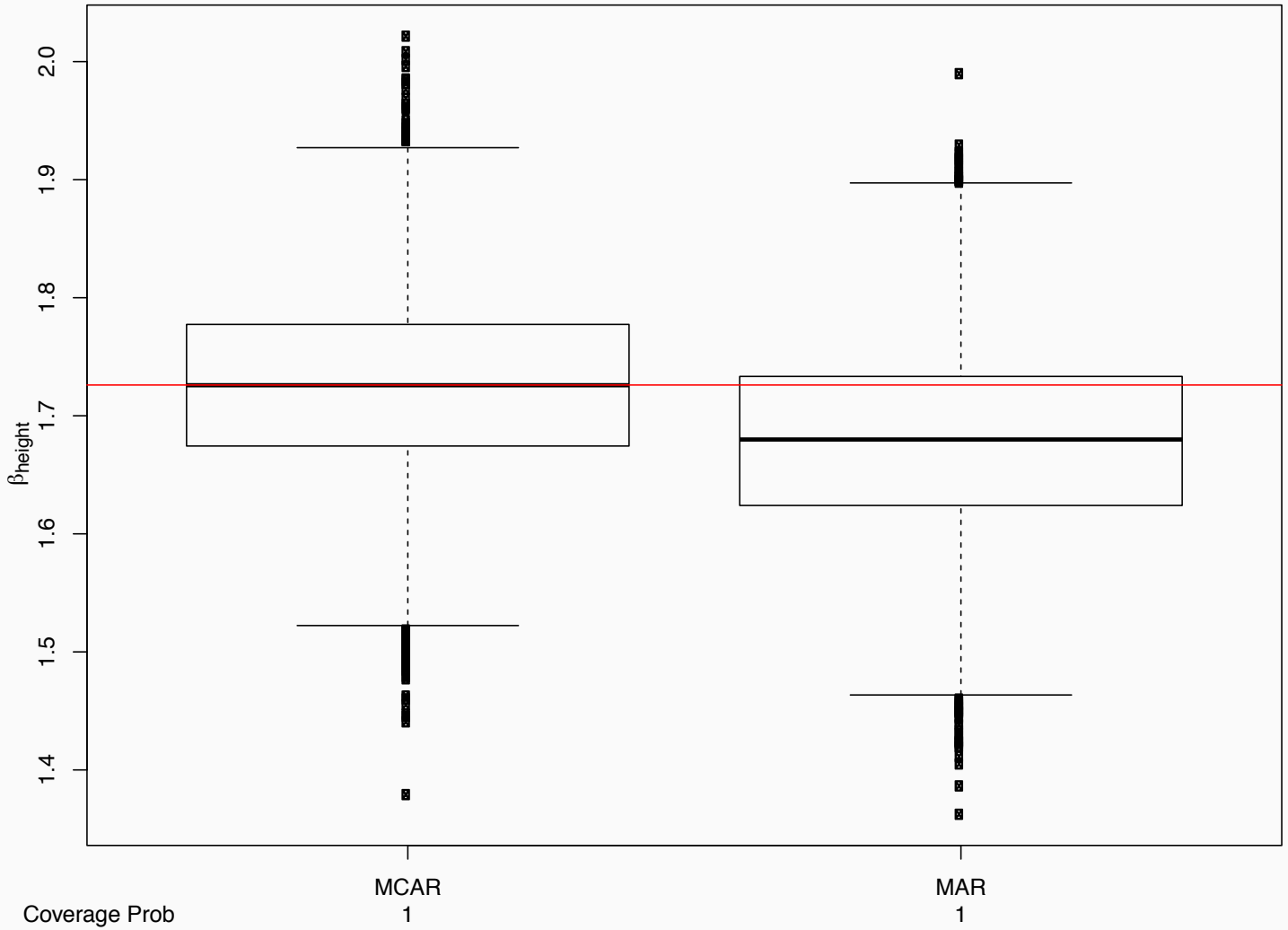
$$logit(P(R = 0|X_2, Y)) = \log(0.3/0.7) + 0.4 * (X_2 - \bar{X}_2) + 0.1 * (Y - \bar{Y})$$

I ran 5000 iterations of each simulation, estimating this linear regression model in each iteration:

$$Y_i = \hat{\beta}_0 + \hat{\beta}_{height} X_1 + \hat{\beta}_{age} X_2 + \epsilon_i$$

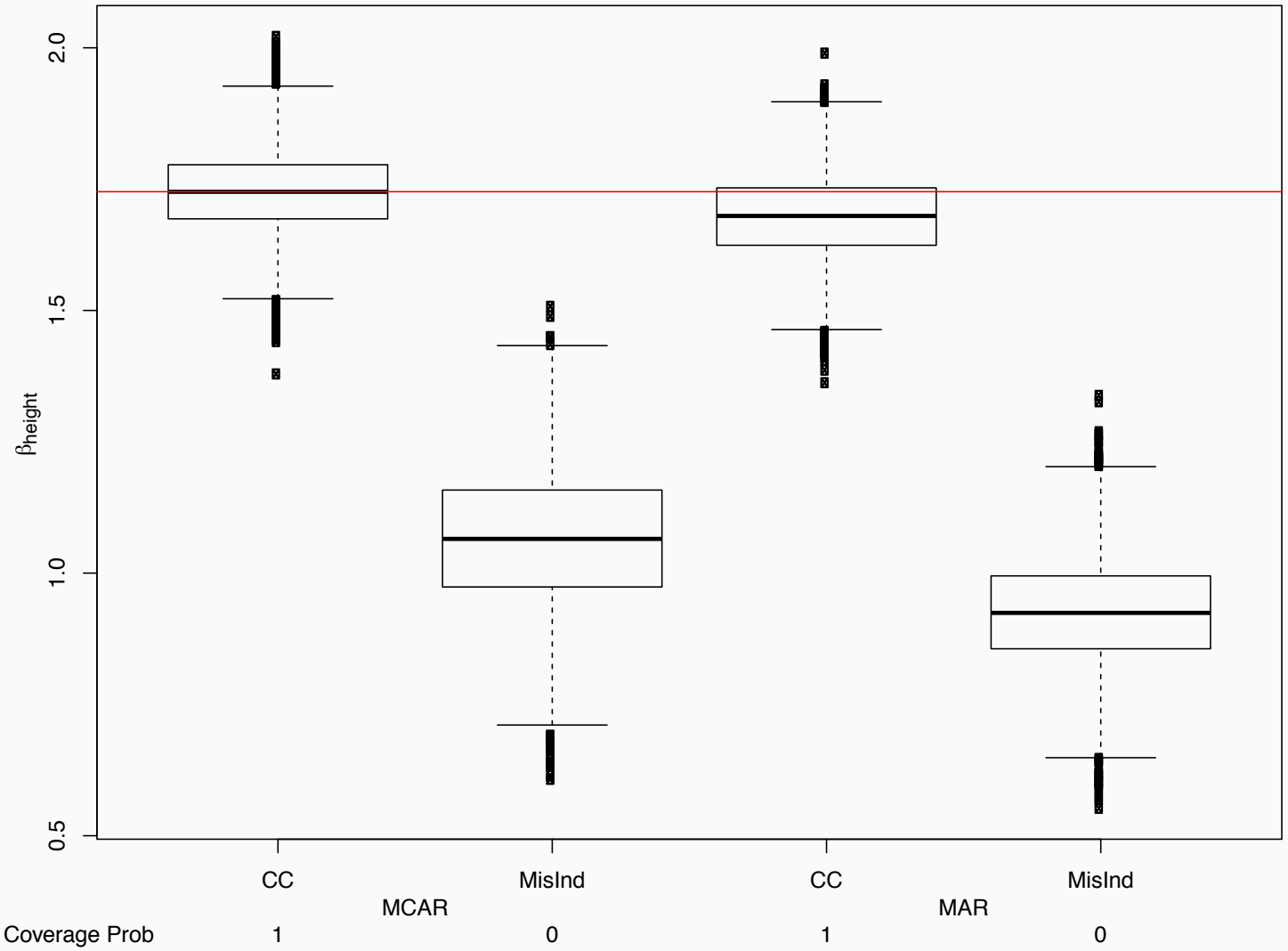
In each plot, I also summarize the coverage probability (proportion of times the true β_{height} was included in the confidence interval).

Complete case analysis under MCAR and MAR



Missing indicator analysis under MCAR and MAR

Now consider an analysis using the missing indicator method.

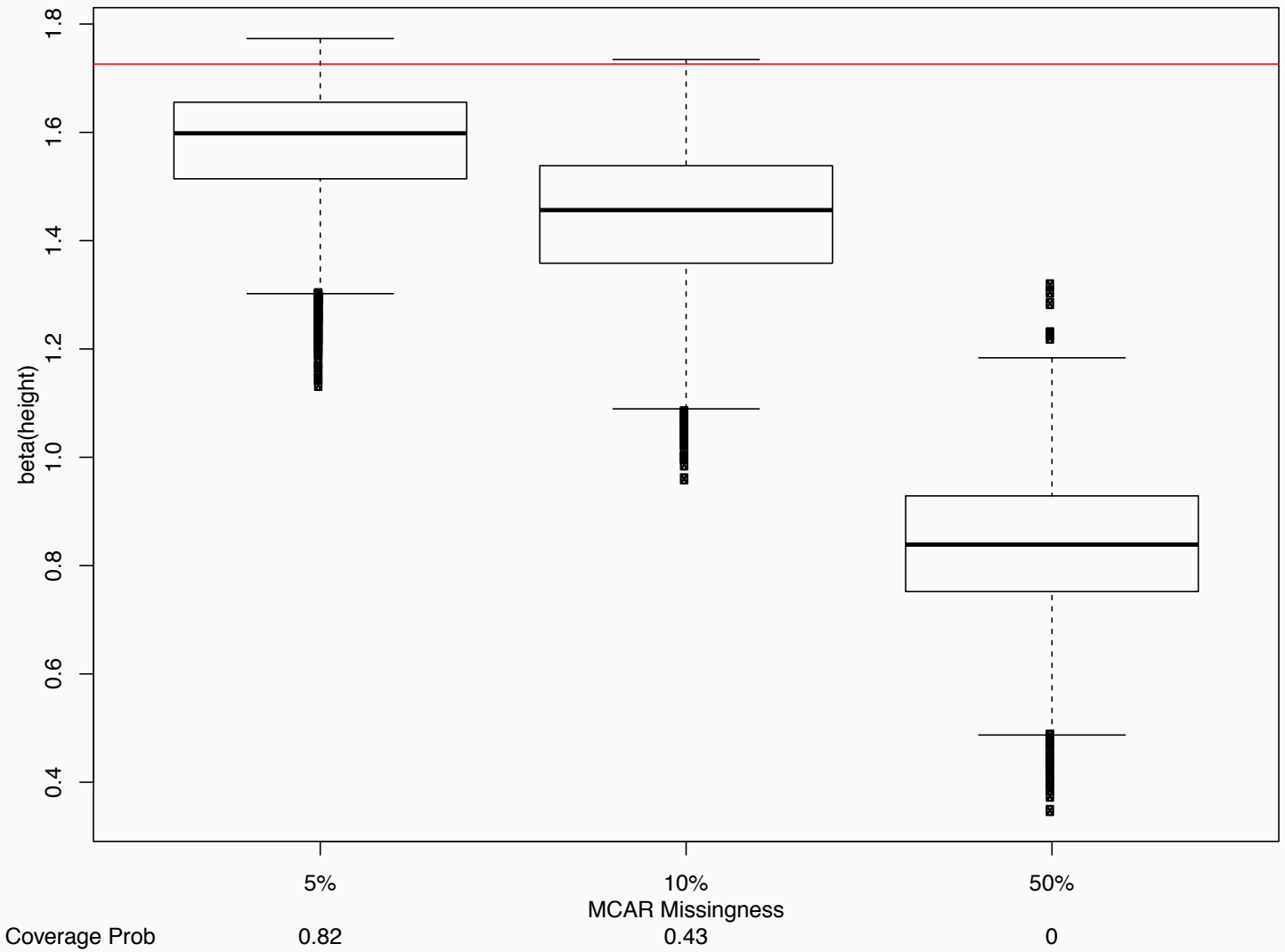


Missing indicator: MCAR scenarios

Now let's consider the missing indicator under the following MCAR missingness scenarios:

1. MCAR with 5% missing
2. MCAR with 10% missing
3. MCAR with 50% missing

Missing indicator: MCAR scenarios



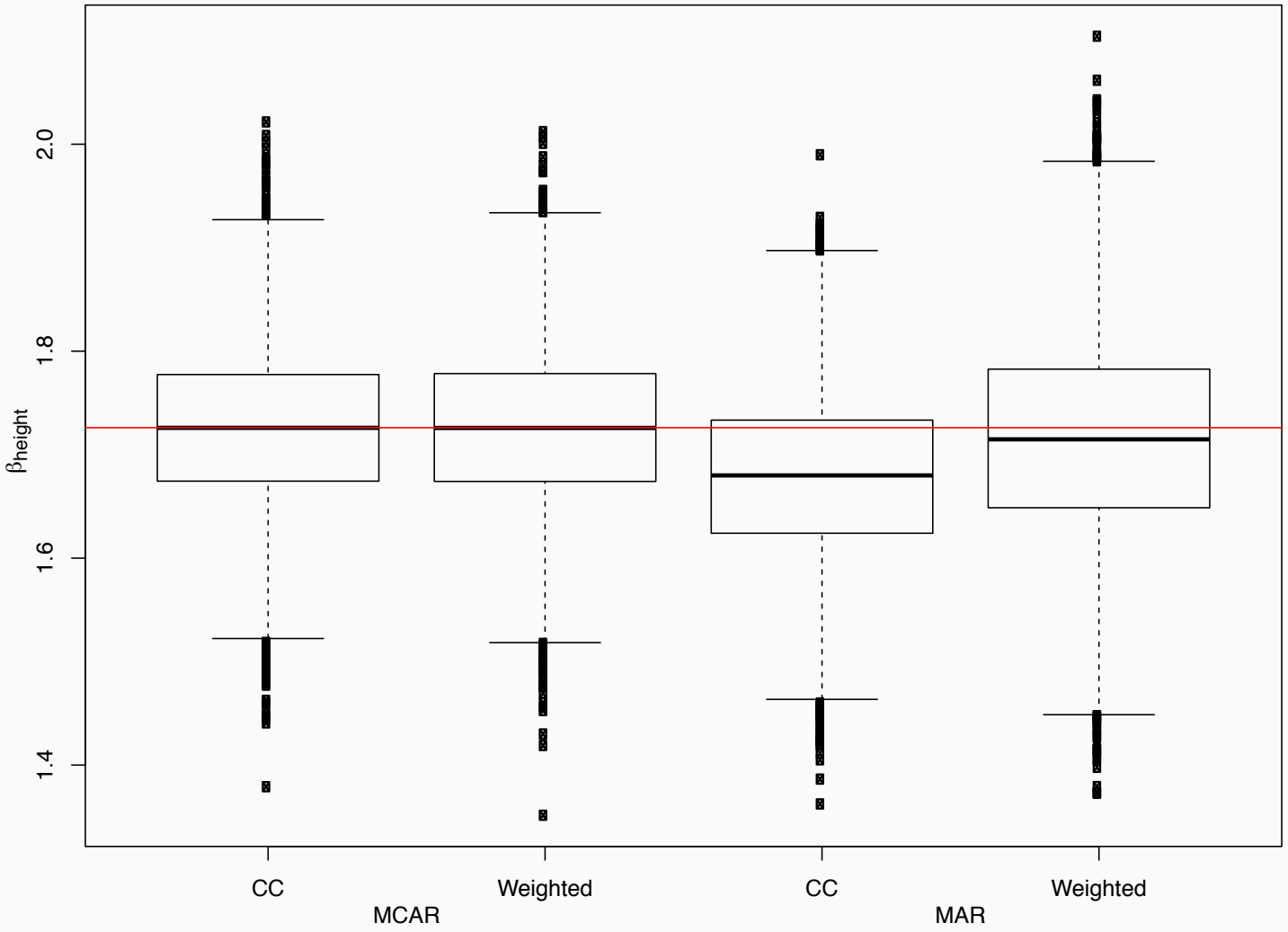
Weighted analysis: complete case vs. weighted under MAR

To compute the inverse probability weights, I fit a logistic regression model for

$$logit(P(R = 1)) = \theta_0 + \theta_1 age + \theta_2 log.FEV1$$

Note that this model included the fully observed Y variable as a predictor of missingness on height.

Weighted analysis: complete case vs. weighted under MAR



Simulation Summary

- Complete case analysis under MCAR gives an unbiased estimate of $\hat{\beta}_{height}$.
- Complete case analysis under MAR gives a biased estimate of $\hat{\beta}_{height}$.
- Weighted regression can recover an unbiased estimate of $\hat{\beta}_{height}$ under MAR.
- The missing indicator method yields biased estimates of $\hat{\beta}_{height}$ **even under MCAR!**

Summary

- We developed an intuition for missing data as a form of collider stratification bias, like selection bias.
- We talked about classifying missingness patterns according to
  - Missing Completely At Random (MCAR)
  - Missing At Random (MAR)
  - Missing Not At Random (MNAR or NMAR)
- We discussed several popular ad hoc methods of dealing with missing data (that all have serious drawbacks).
- We introduced the notion of **statistically principled** methods for handling missing data.

References

Two important textbooks:

1. Little RJA, Rubin DB. (1987) *Statistical Analysis with Missing Data*. New York: John Wiley & Sons.
2. Schafer JL. (1997) *Analysis of Incomplete Multivariate Data*. Boca Raton: Chapman & Hall/CRC.

References

Some helpful papers:

- Graham, JW. (2009) Missing Data: Making It Work in the Real World. *Annu Rev Psychol* 60:549-576.
- Greenland, S., and Finkle, W. D. (1995), A Critical Look at Methods for Handling Missing Covariates in Epidemiologic Regression Analyses, *American Journal of Epidemiology*, 142, 1255-1264.
- **Hernan MA, Hernandez-Diaz S, Robins JM. A structural approach to selection bias. *Epidemiology* 2004;15:615-25.**
- Horton NJ, Kleinman KP. (2007) Much Ado About Nothing: A Comparison of Missing Data Methods and Software to Fit Incomplete Data Regression Models. *The American Statistician* 61:79-90.
- Ibrahim, J. G., Chen, M.-H., Lipsitz, S. R., and Herring, A. H. (2005), Missing-Data Methods for Generalized Linear Models: A Comparative Review, *Journal of the American Statistical Association*, 100, 332-346.
- **Smith LH. (2020) Selection mechanisms and their consequences: understanding and addressing selection bias. *Current Epidemiology Reports*. doi: 10.1007/s40471-020-00241-6**