Analysis with missing data

Missing data can lead to problems

"The occurrence of missing data complicates statistical analysis, because we cannot perform the analysis originally intended for complete data without having to handle the missing values first. A direct consequence of this is that inappropriate handling of missing values can lead to bias and incorrect conclusions."

The best solution

The best solution to missing data is to not have any.

Missing observations occur because

- Participants drop out of studies
- Participants do not answer specific questions
- Data is lost or corrupted
- Measurements fail
- Data is not recorded
- Data is not available (e.g., expensive measurements)
- Data is censored (e.g., detection limits)

ADDICTION

SSA SOCIETY FOR THE STUDY OF ADDICTION

Research Report | Full Access

Evaluation of adding the community reinforcement approach to motivational enhancement therapy for adults aged 60 years and older with DSM-5 alcohol use disorder: a randomized controlled trial

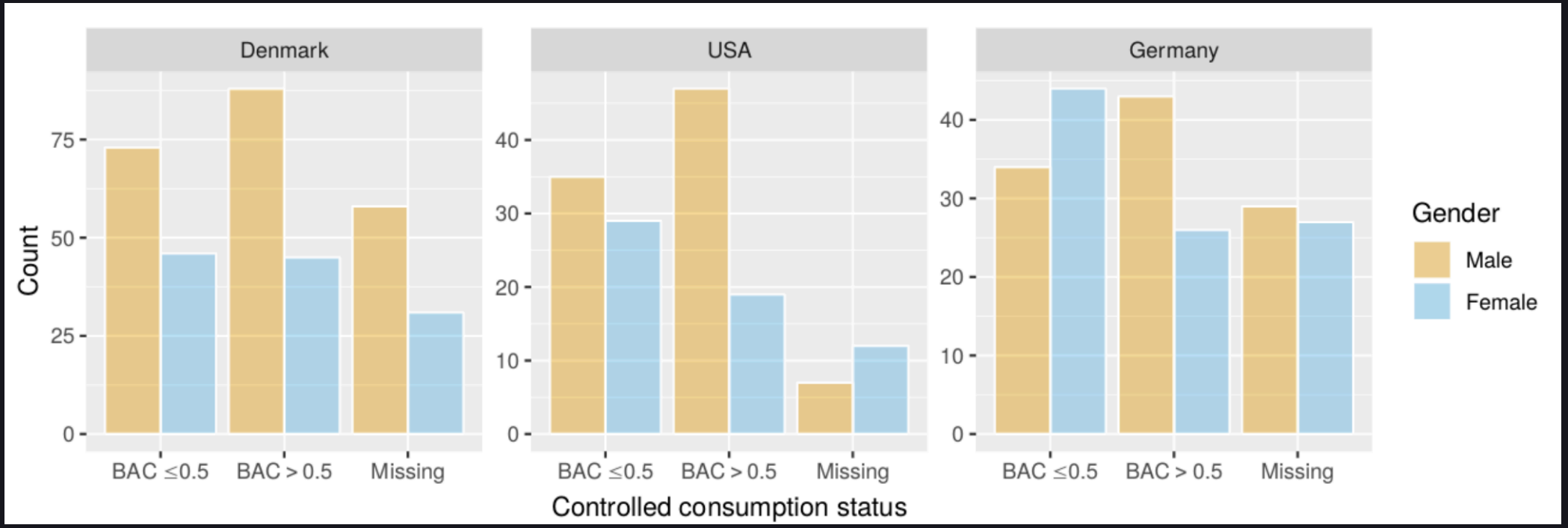
Kjeld Andersen ✉, Silke Behrendt, Randi Bilberg, Michael P. Bogenschutz, Barbara Braun, Gerhard Buehringer, Claus Thorn Ekstrøm, Anna Mejldal, Anne Helby Petersen, Anette Søgaard Nielsen

First published: 27 August 2019 | https://doi.org/10.1111/add.14795 | Citations: 41

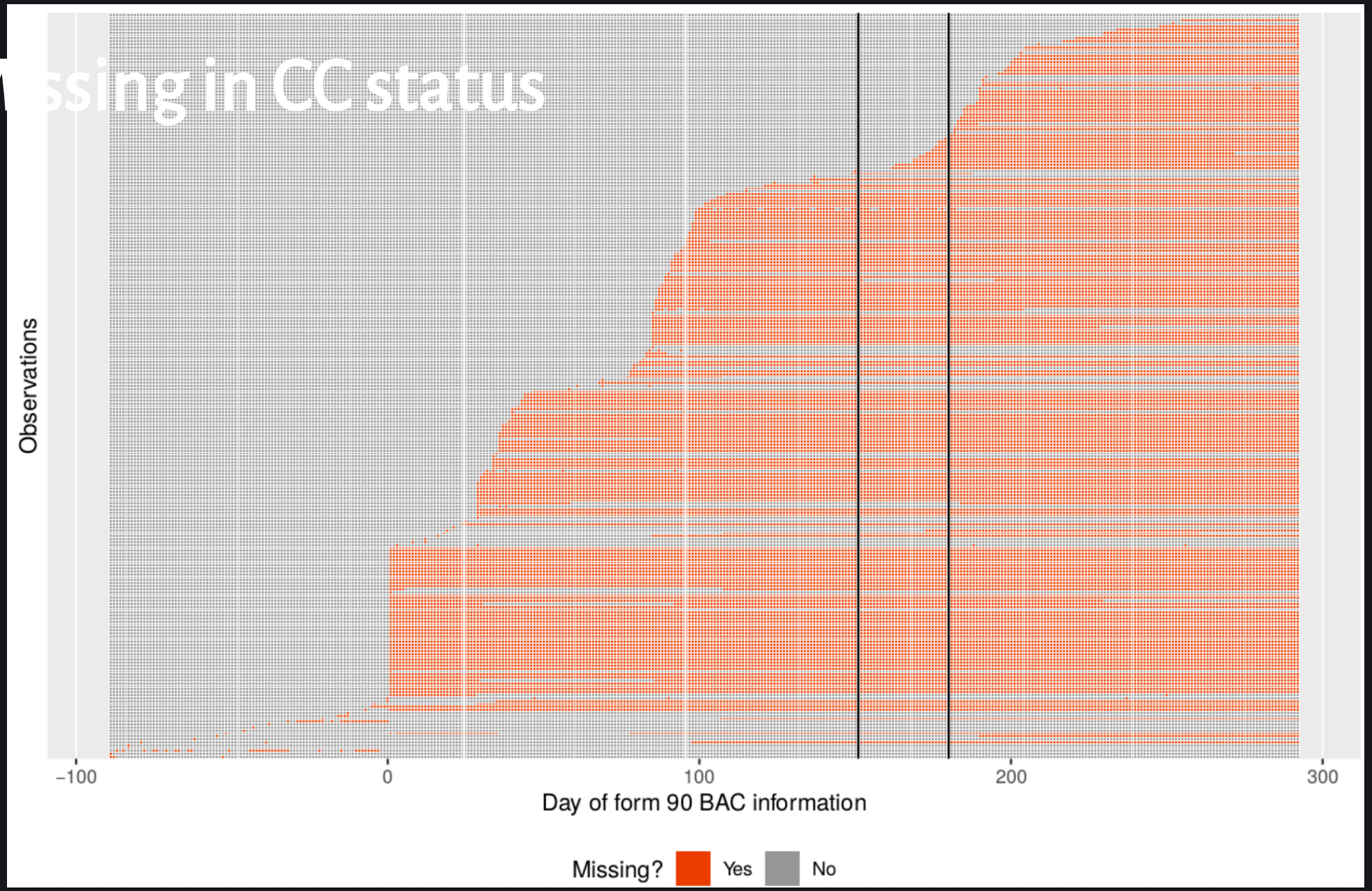
- RCT on 693 patients from USA, DK and DE, suffering from alcohol dependency, all 60+ years old.
- Purpose: Compare usual treatment (MET) with new treatment that contains an additional element (MET+CRA).
- Outcome: Controlled consumption (CC) status after approx. 6 months of treatment (blood alcohol level ≤ 0.05% at all times during 30 days).
- Seven baseline covariates: country (DK, DE, USA), sex (male/female), age (years older than 60), education (no degree/at most undergrad./grad. or post grad.), cohabiting with partner (yes/no), alcohol dependence severity (low/intermediate/severe), number of previous treatments (0/1-2/3+)

Missing information in two variables

- Dependency severity: 3 patients (0.43%) (excluded)
- Controlled consumption status: 164 patients (23.77%)



Missing in CC status



Non-response analysis

We fitted a logistic regression model with:

- Outcome: Indicator of whether the patient has missing CC status
- Predictors: All the remaining variables

	Change in df	AIC	LRT	p-value
Full model		740.8		
Treatment	1	739.2	0.3952	0.5296
Gender	1	738.8	0.0032	0.9551
Country	2	748.9	12.0718	0.0024
Age	1	746.6	7.8397	0.0051
Education	2	741.8	4.9461	0.0843
Partner status	1	744.5	5.7452	0.0165
Alcohol dependence severity	2	739.1	2.3205	0.3134
Previous treatment history	2	744.7	7.8811	0.0194

In practice

We estimated the difference between the two treatments using logistic regression.

- The primary model was fitted on complete cases only.
- In sensitivity analyses, we compared this model to
  - A model using multiple imputation
  - Several best case/worst case scenario models

Table 3.1: Estimated log odds ratios from the model of controlled consumption status using all full covariate adjustment. The reported estimates are on log odds ratio scale and they are computed relative to the following reference category: Treament MET; Gender male; Country Denmark; Age 60; Education none; No partner; Low ADS; Previous treatments 0. The mean log odds of having a controlled alcohol consumption in this reference group is represented by the intercept estimate. The reported p-values correspond to two-sided z-tests of the null-hypothesis of a zero parameter value.

	Estimate	Std. error	z statistic	p-value
Intercept	-0.3507	0.3050	-1.1499	0.2502
Treatment: MET+CRA	0.2028	0.1801	1.1260	0.2602
Country: USA	0.0736	0.2327	0.3164	0.7517
Country: Germany	-0.0351	0.2522	-0.1392	0.8893
Gender: Female	-0.5543	0.1906	-2.9085	0.0036
Age	0.0677	0.0211	3.2038	0.0014
Married or cohabiting: Yes	0.2270	0.1877	1.2094	0.2265
Severity: Intermediate	-0.0777	0.2307	-0.3367	0.7363
Severity: Substantial or severe	-0.2767	0.4096	-0.6755	0.4994
Education: At most undergraduate degree	0.0518	0.2286	0.2268	0.8206
Education: Graduate or post-graduate	-0.4463	0.2872	-1.5537	0.1202
Previous treatments: 1-2	0.2655	0.2187	1.2140	0.2247
Previous treatments: 3+	0.2938	0.3087	0.9517	0.3413

Hypothetical follow-up study

- Outcome: Diagnosis of liver disease during 10 years of follow-up (no censoring, no death).
- Key variable of interest: CC status after 6 months.
- Other variables: Country, gender, age, education, partner status, alcohol dependency severity, previous treatment history.
- You observe 24% missing in CC status.

24% have missing CC information ...

1. ... due to a fire in the storing facility.
2. ... because those patients were embarassed to tell the treatment facility that they had started drinking again.
3. ... and they are the 24% of the patients with the most severe alcohol dependencies, and they dropped out of the study
4. ... and they are the 24% females, and they dropped out of the study.
5. those patients all had last names starting with "A". Their records were lost because someone dropped a cup of coffee on that folder.
6. ... because those patients dropped out of the study since they were not drinking and felt safe that they wouldn't start again.
7. ... and they are the 24% that have red hair. They are missing in CC because a data manager accidentially deleted their information - and the variable containing hair color.
8. and they all had undiagnosed pre-stages to liver disease during the study and dropped out due to illness.

Discuss the missing information scenarios

Assume that we carry out a statistical analysis (e.g. logistic reg.) using only patients with no missing information (complete case analysis).

For each of the eight scenarios, discuss with your neighbors:

- Will this affect the estimate of the effect of CC status on liver disease risk?
- Will this affect the precision (e.g., width of confidence intervals) of our effect estimate?
- Can this problem be solved using statistical methods? And do you have any suggestions for how?

Missing information in R

First rule of missing information handling in R:

Always represent missing values by `NA`.

Second rule of missing information handling in R:

Always represent missing values by `NA`.

A list of bad ideas

Do *not* represent missing information with

- A dot (`.`)
- An empty string (`""`)
- A string with a space (`" "`)
- A string with a dash (`"-"`)
- A special numeric value such as `Inf`
- A specific number (e.g., `-9`, `99`, `0`)
- Anything else that is not `NA`

A quick check

You can use the `dataReporter` package in `R` to look for problems:

```
library("dataReporter")
testData$miscodedMissingVar
```

```
 [1] "."    ""     "nan"  "NaN"  "NAN"  "na"   "NA"
 [8] "Na"   "Inf"  "inf"  "-Inf" "-inf" "-"    "9"
[15] "9"
```

```
identifyMissing(testData$miscodedMissingVar)
```

The following suspected missing value codes enter as regular values: , -, -inf, -Inf, ., …, Na, NA, nan, NaN, NAN (4 values omitted).

Or get a full report

```
library("dataReporter")
makeDataReport(testData)
```

Evil schemes for missingness

Imagine we had a fully observed dataset, but wish to induce missing information in one variable. How can we make data go missing?

	country	gender	prevTreat	CC
1	Denmark	Male	0	Yes
2	Denmark	Male	0	No
3	Denmark	Male	1-2	No
4	Denmark	Male	1-2	Yes
5	Denmark	Female	0	No
7	Denmark	Female	1-2	Yes
8	Denmark	Male	1-2	Yes
9	Denmark	Male	3+	Yes
10	Denmark	Female	0	No
11	Denmark	Male	3+	No
12	Denmark	Male	3+	Yes
13	Denmark	Female	0	No
14	Denmark	Male	0	No

Missing completely at random (MCAR)

Choose who is missing by random dice roll.

	country	gender	prevTreat	CC
1	Denmark	Male	0	Yes
2	Denmark	Male	0	No
3	Denmark	Male	1-2	No
4	Denmark	Male	1-2	Yes
5	Denmark	Female	0	No
7	Denmark	Female	1-2	Yes
8	Denmark	Male	1-2	Yes
9	Denmark	Male	3+	Yes
10	Denmark	Female	0	No
11	Denmark	Male	3+	No
12	Denmark	Male	3+	Yes
13	Denmark	Female	0	No
14	Denmark	Male	0	No

	country	gender	prevTreat	CC
1	Denmark	Male	0	Yes
2	Denmark	Male	0	
3	Denmark	Male	1-2	No
4	Denmark	Male	1-2	Yes
5	Denmark	Female	0	No
7	Denmark	Female	1-2	
8	Denmark	Male	1-2	Yes
9	Denmark	Male	3+	
10	Denmark	Female	0	No
11	Denmark	Male	3+	No
12	Denmark	Male	3+	Yes
13	Denmark	Female	0	No
14	Denmark	Male	0	No

We lose information and hence precision (wider confidence intervals).

Missing at random (MAR)

Choose who is missing by separate random draws for males and females. Females have missing probability 75%, males have probability 5%.

	country	gender	prevTreat	CC
1	Denmark	Male	0	Yes
2	Denmark	Male	0	No
3	Denmark	Male	1-2	No
4	Denmark	Male	1-2	Yes
5	Denmark	Female	0	No
7	Denmark	Female	1-2	Yes
8	Denmark	Male	1-2	Yes
9	Denmark	Male	3+	Yes
10	Denmark	Female	0	No
11	Denmark	Male	3+	No
12	Denmark	Male	3+	Yes
13	Denmark	Female	0	No
14	Denmark	Male	0	No

	country	gender	prevTreat	CC
1	Denmark	Male	0	Yes
2	Denmark	Male	0	No
3	Denmark	Male	1-2	No
4	Denmark	Male	1-2	Yes
5	Denmark	Female	0	
7	Denmark	Female	1-2	
8	Denmark	Male	1-2	Yes
9	Denmark	Male	3+	Yes
10	Denmark	Female	0	No
11	Denmark	Male	3+	No
12	Denmark	Male	3+	Yes
13	Denmark	Female	0	
14	Denmark	Male	0	No

Underrepresentation of females may lead to biased estimates.

Missing not at random (MNAR)

Choose who is missing by looking at CC status itself. Relapsers are more likely to have missing information.

	country	gender	prevTreat	CC
1	Denmark	Male	0	Yes
2	Denmark	Male	0	No
3	Denmark	Male	1-2	No
4	Denmark	Male	1-2	Yes
5	Denmark	Female	0	No
7	Denmark	Female	1-2	Yes
8	Denmark	Male	1-2	Yes
9	Denmark	Male	3+	Yes
10	Denmark	Female	0	No
11	Denmark	Male	3+	No
12	Denmark	Male	3+	Yes
13	Denmark	Female	0	No
14	Denmark	Male	0	No

	country	gender	prevTreat	CC
1	Denmark	Male	0	Yes
2	Denmark	Male	0	No
3	Denmark	Male	1-2	No
4	Denmark	Male	1-2	Yes
5	Denmark	Female	0	No
7	Denmark	Female	1-2	Yes
8	Denmark	Male	1-2	Yes
9	Denmark	Male	3+	Yes
10	Denmark	Female	0	No
11	Denmark	Male	3+	No
12	Denmark	Male	3+	Yes
13	Denmark	Female	0	No
14	Denmark	Male	0	No

Underrepresentation of relapsers may lead to biased estimates. And the worst part: *Whether an observation is missing depends on the very information*

Three categories

- **MCAR**: The missingness is independent of all other observed or unobserved variables. (Dice roll)
- **MAR**: The missingness depends on other observed variables, but not on the missing value itself.
- **MNAR**: The missingness depends on the value that is missing. Or rather: not one of the other two conditions.

Better names?

- Unrelated missingness.
- Data-dependent missingness
- Unobserved data missingness

Exercise

We will now:

- Classify each of the 8 scenarios as MCAR/MAR/MNAR
- Discuss: Would it had been possible to detect this by looking at the data alone (i.e. not knowing why the data went missing)?

24% have missing CC information ...

1. ... due to a fire in the storing facility.
2. ... because those patients were embarassed to tell the treatment facility that they had started drinking again.
3. ... and they are the 24% of the patients with the most severe alcohol dependencies, and they dropped out of the study
4. ... and they are the 24% females, and they dropped out of the study.
5. those patients all had last names starting with "A". Their records were lost because someone dropped a cup of coffee on that folder.
6. ... because those patients dropped out of the study since they were not drinking and felt safe that they wouldn't start again.
7. ... and they are the 24% that have red hair. They are missing in CC because a data manager accidentially deleted their information - and the variable containing hair color.
8. and they all had undiagnosed pre-stages to liver disease during the study and dropped out due to illness.

Distinguishing between MCAR/MAR/MNAR

Strategy 1: Try to rule out MCAR

If you can find a variable - or combination of variables - that gives you information about whether CC is more or less likely to be missing, *the mechanism is not MCAR*.

- Note: Statistical testing doesn't always produce the correct answer - sometimes, we find false positives.

Strategy 2: NA

- Nothing more can be done using data and statistics alone.

We will now start looking at data with missing information.

- We have a dataset consisting of the baseline covariates from the Elderly study and an additional variable, drinks, with the mean number of drinks consumed per day in the month before the study started.
- We wish to model how drinks depends on the other baseline covariates.
- However, an evil person made some of the data go missing.
- Today's goal is to find out what happened to the data and try to obtain a correct analysis despite the evil scheme.

Exercise

Now wouldn't life be good if ...

... there were no missing data?

Perhaps use imputation to fill in the blank/missing slots in the data with plausible values.

Simple missing information setup

Imagine that we wish to estimate the effect of X on Y, controlling for Z.

- X suffers from missing information (MCAR). Assume that we order the observations such that $X_1, \ldots, X_d$ have missing information, while $X_{d+1}, \ldots, X_n$ are fully observed.
- Assume that Y and Z are fully observed.

Note: Complete case analysis would produce an unbiased, but inefficient estimate.

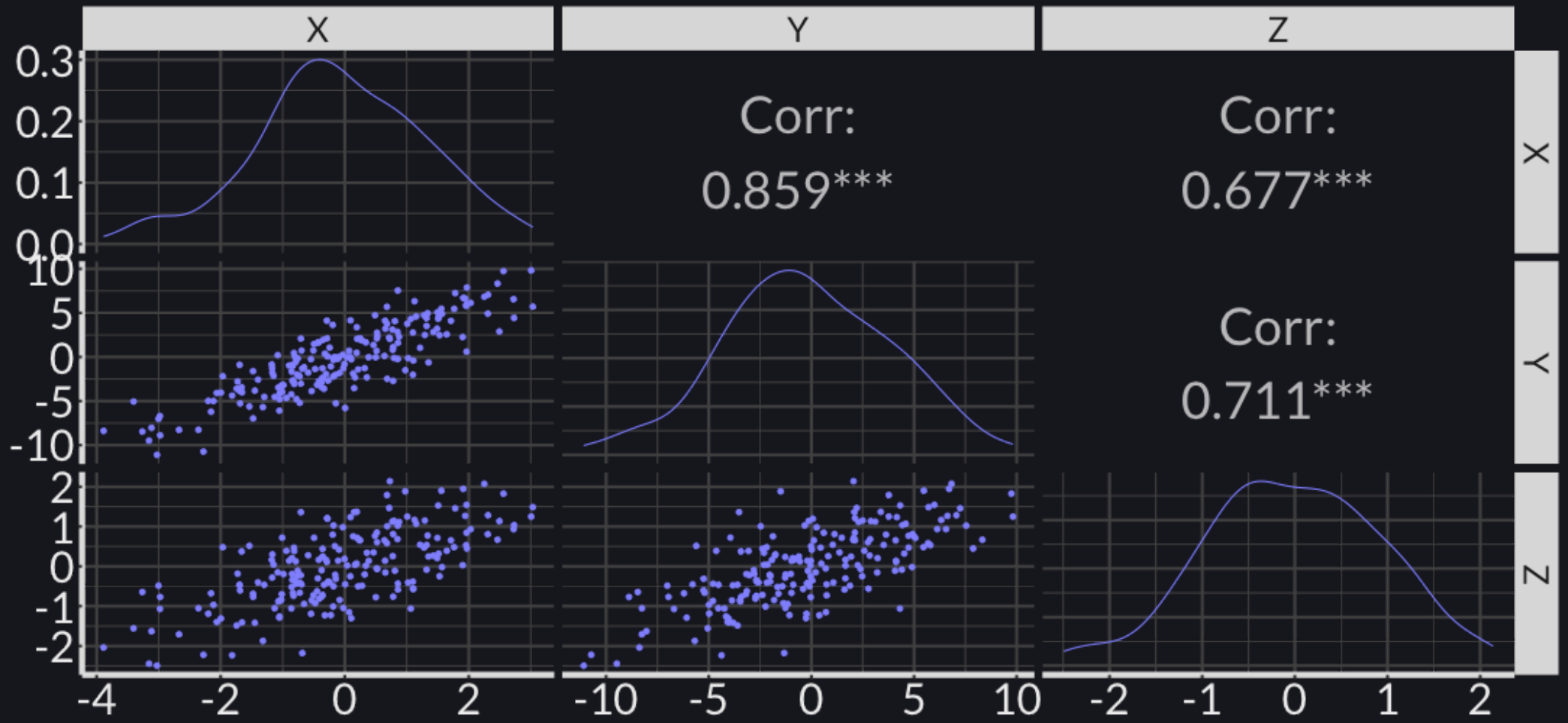
```r
n <- 200
set.seed(1331)
Z <- rnorm(n, sd = 1)
X <- Z + rnorm(n, sd = 1)
Y <- 2*X + Z + rnorm(n, sd = 2)


true_X <- X
true_xmean <- mean(X)
true_xsd <- sd(X)
true_model <- lm(Y ~ X + Z)
```

```r
d <- 40
X[1:d] <- NA

X[38:43]
```

```
[1]         NA         NA         NA -0.9404489
[5]  0.7807026  1.9016603
```


X
Y
Z
Corr:
0.859***
Corr:
0.677***
Corr:
0.711***
X
Y
Z
0.3
0.2
0.1
0.0
10
5
0
-5
-10
2
1
0
-1
-2
-4
-2
0
2
-10
-5
0
5
10
-2
-1
0
1
2

Mean imputation: Insert the mean (or mode) of $X_{d+1}, \ldots, X_n$ into all $X_1, \ldots, X_d$.

```
X_meanimp <- X
xobs_mean <- mean(X[(d+1):n])
X_meanimp[1:d] <- xobs_mean
#Compare mean for full X with mean of mean imputed X
c(true_xmean, mean(X_meanimp))
```

```
[1] -0.07813252 -0.15518740
```

```
#Compare sd for full X with sd of mean imputed X
c(true_xsd, sd(X_meanimp))
```

```
[1] 1.368721 1.240263
```

Compare model coefficients

```
round(summary(true_model)$coefficients,4)
```

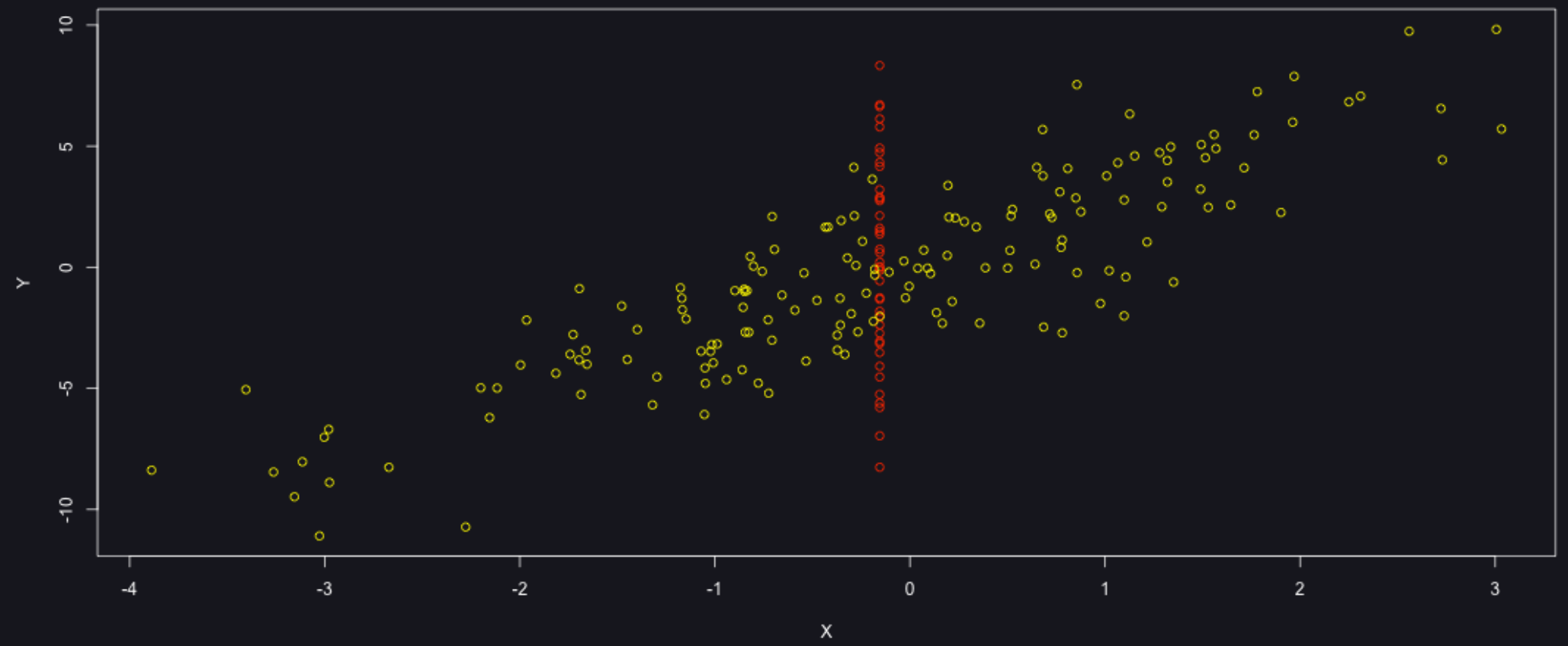
```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.0532     0.1392 -0.3822   0.7028
X             2.0756     0.1382 15.0158   0.0000
Z             1.0260     0.2000  5.1297   0.0000
```

```
round(summary(lm(Y ~ X_meanimp + Z))$coefficients,4)
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.0709     0.1623  0.4371   0.6626
X_meanimp     1.7887     0.1639 10.9102   0.0000
Z             1.6266     0.2150  7.5675   0.0000
```

Conclusion: Do not mean impute!

```
plot(Y ~ X_meanimp, xlab = "X",
col = c(rep("red", 40), rep("yellow", 160)), col.lab = "white", col.ax
```



Hot deck imputation (simplest version): For each missing value, pick and insert a random value among the observed values $X_{d+1}, \ldots, X_n$.

```
X_hdimp <- X
set.seed(13)
X_hdimp[1:d] <- sample(X[(d+1):n], size = d, replace = TRUE)

#Compare mean for full X with mean of mean imputed X
c(true_xmean, mean(X_hdimp))
```

```
[1] -0.07813252 -0.20307655
```

```
#Compare sd for full X with sd of mean imputed X
c(true_xsd, sd(X_hdimp))
```

```
[1] 1.368721 1.387267
```

Comparing model coefficients:

```
round(summary(true_model)$coefficients,4)
```

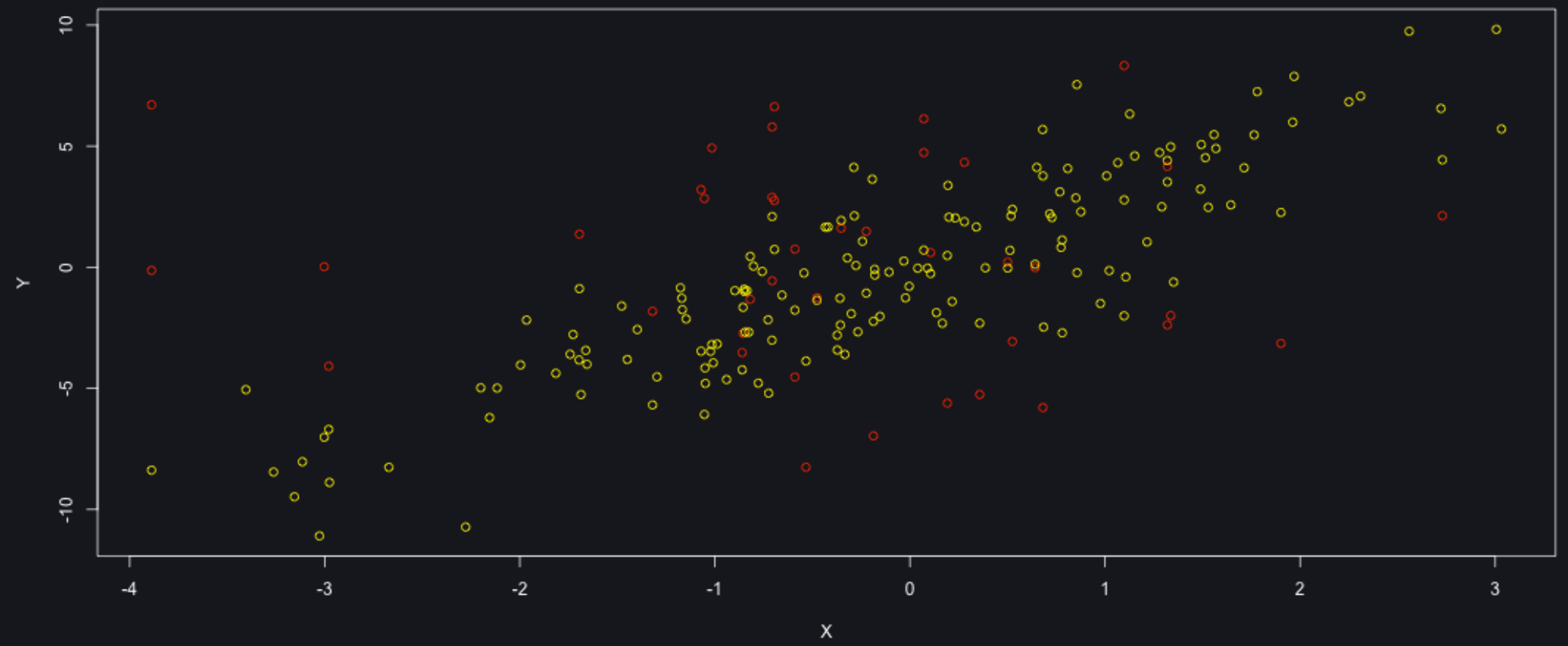
```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.0532     0.1392 -0.3822   0.7028
X             2.0756     0.1382 15.0158   0.0000
Z             1.0260     0.2000  5.1297   0.0000
```

```
round(summary(lm(Y ~ X_hdimp + Z))$coefficients,4)
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.0484     0.1781  0.2720   0.7859
X_hdimp       1.2207     0.1492  8.1828   0.0000
Z             2.1195     0.2188  9.6879   0.0000
```

Conclusion: Don't do hot deck imputation!

```
plot(Y ~ X_hdimp, xlab = "X",
col = c(rep("red", 40), rep("yellow", 160)), col.lab = "white", col.ax
```



Regression imputation: Fit a regression model for all the observations, e.g., $X_i = \alpha + \beta_1 Y_i + \beta_2 Z_i + \epsilon_i$ for $i = d+1, \ldots, n$ and use this model to predict values for the remaining $X_1, \ldots, X_d$.

```
obsdata <- data.frame(X = X[(d+1):n],
                      Y = Y[(d+1):n],
                      Z = Z[(d+1):n])
m_regimp <- lm(X ~ Y + Z, obsdata)
X_regimp <- X
X_regimp[1:d] <- predict(m_regimp,
                         newdata = data.frame(Y = Y[1:d],
                         Z = Z[1:d]))
```

```
#Compare mean for full X with mean of reg. imputed X
true_xmean; mean(X_regimp)
```

```
[1] -0.07813252
```

```
[1] -0.1177343
```

```
#Compare sd for full X with mean of reg. imputed X
true_xsd; sd(X_regimp)
```

```
[1] 1.368721
```

```
[1] 1.352262
```

```
round(summary(true_model)$coefficients,4)
```

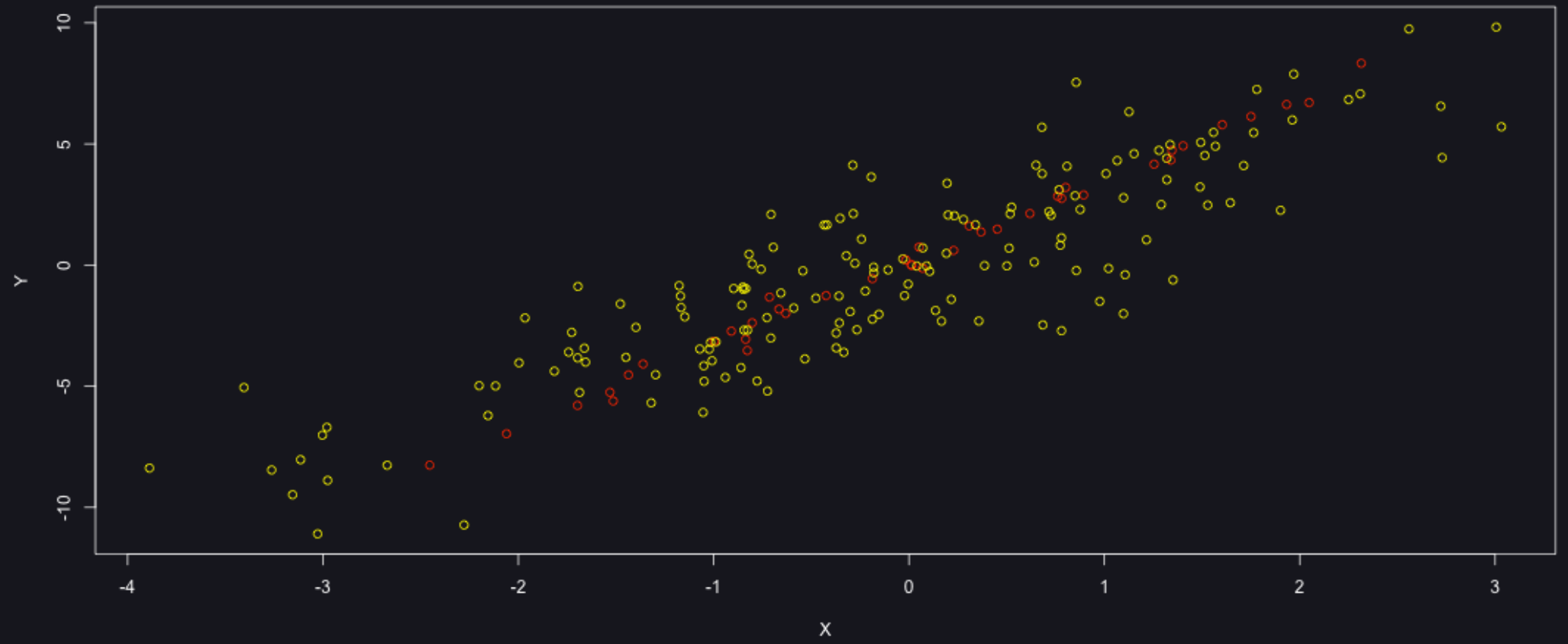
```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.0532     0.1392 -0.3822   0.7028
X             2.0756     0.1382 15.0158   0.0000
Z             1.0260     0.2000  5.1297   0.0000
```

```
round(summary(lm(Y ~ X_regimp + Z))$coefficients,4)
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.0505     0.1256  0.4024   0.6878
X_regimp      2.2815     0.1264 18.0525   0.0000
Z             0.8383     0.1807  4.6401   0.0000
```

Conclusion: Don't do regression imputation!

```
plot(Y ~ X_regimp, xlab = "X",
col = c(rep("red", 40), rep("yellow", 160)), col.lab = "white", col.ax
```



Stochastic regression imputation: Perform regression imputation, but add noise to the predictions by sampling from the residuals from the fitted model.

```
X_stocregimp <- X; set.seed(2)
X_stocregimp[1:d] <- X_regimp[1:d] +
sample(residuals(m_regimp), size = d,
replace = TRUE)
```

```
#Estimate from model with full X
round(summary(true_model)$coefficients,4)[2,]
```

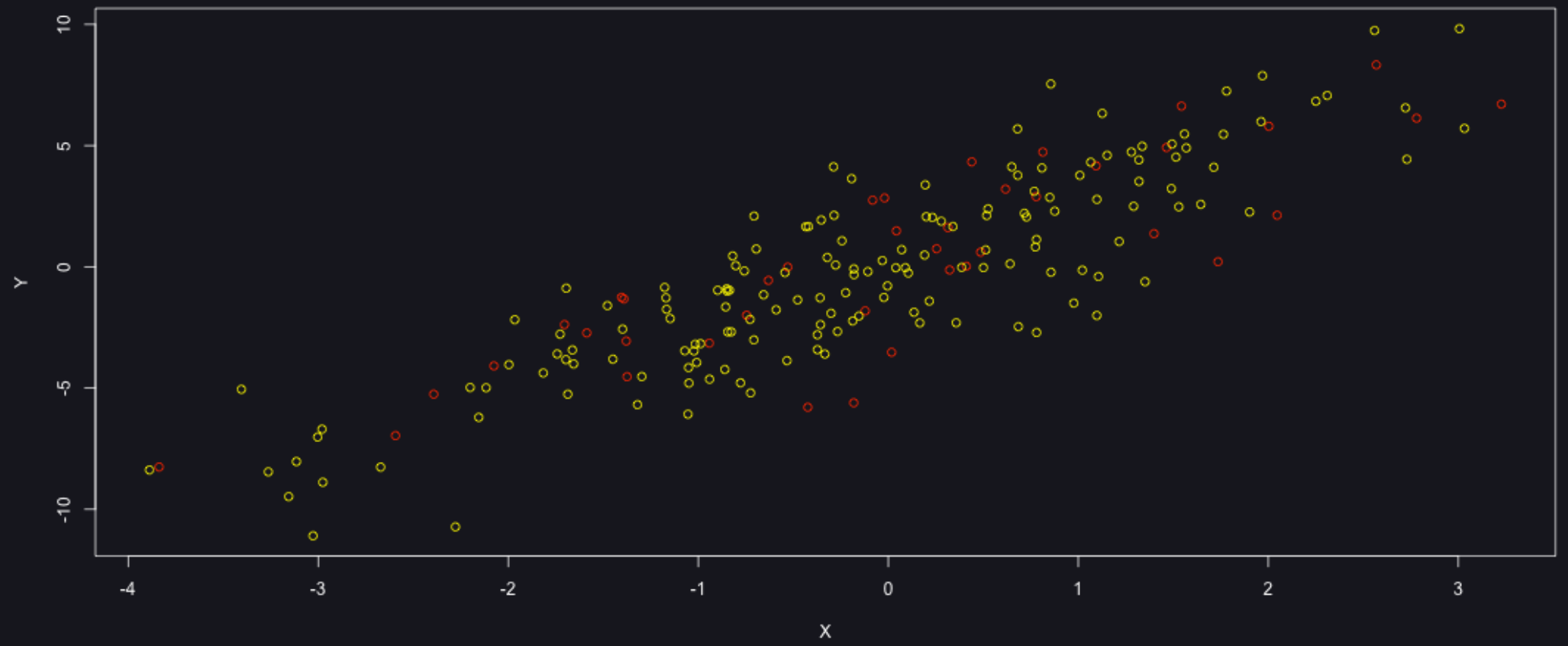
```
 Estimate Std. Error    t value   Pr(>|t|)
   2.0756     0.1382    15.0158     0.0000
```

```
#Estimate from model with X imputed by stochastic regression
round(summary(lm(Y ~ X_stocregimp + Z))$coefficients,4)[2,]
```

```
 Estimate Std. Error    t value   Pr(>|t|)
   2.0430     0.1309    15.6060     0.0000
```

Problem: The variance is still underestimated.

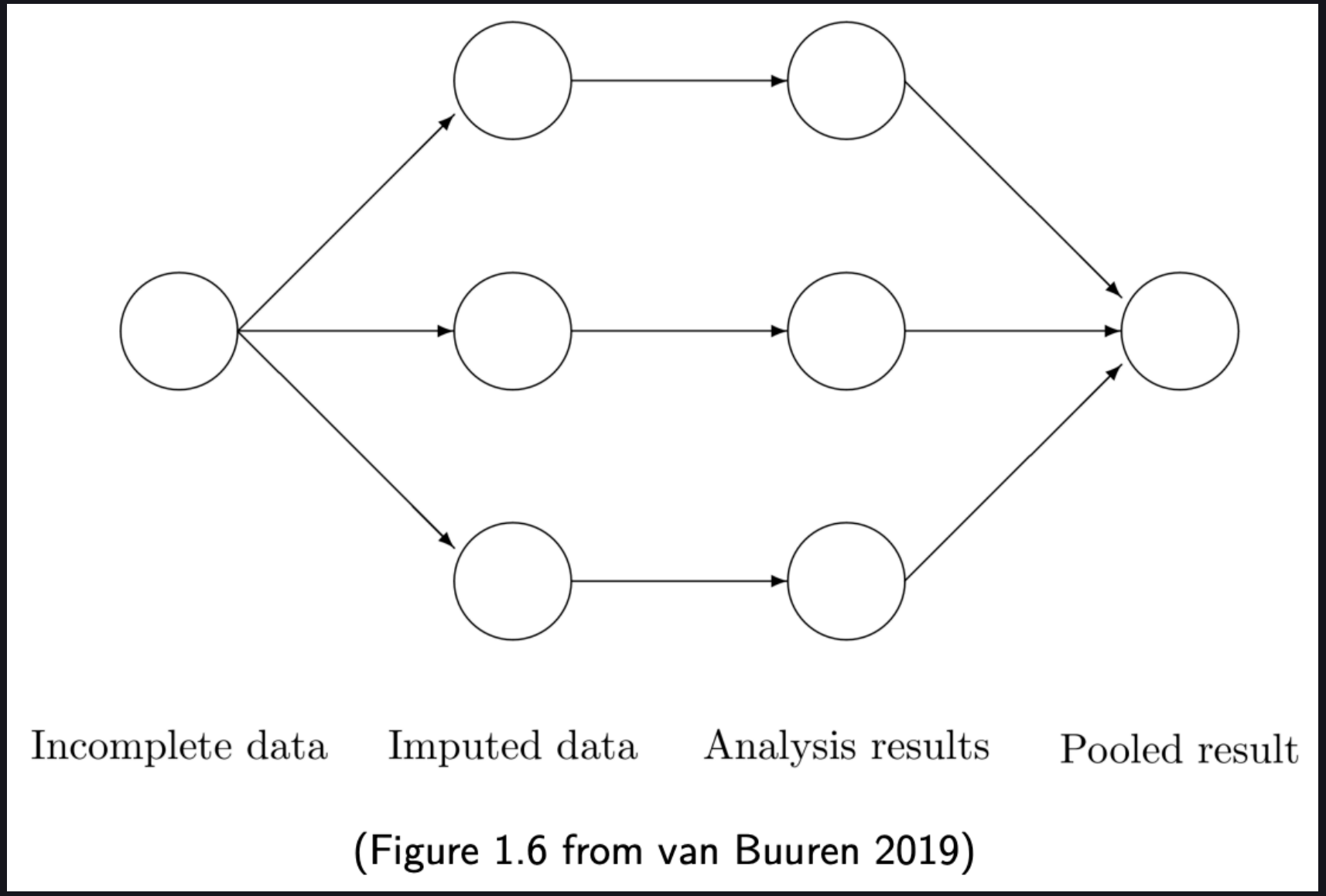
```
plot(Y ~ X_stocregimp, xlab = "X",
col = c(rep("red", 40), rep("yellow", 160)), col.lab = "white", col.ax
```



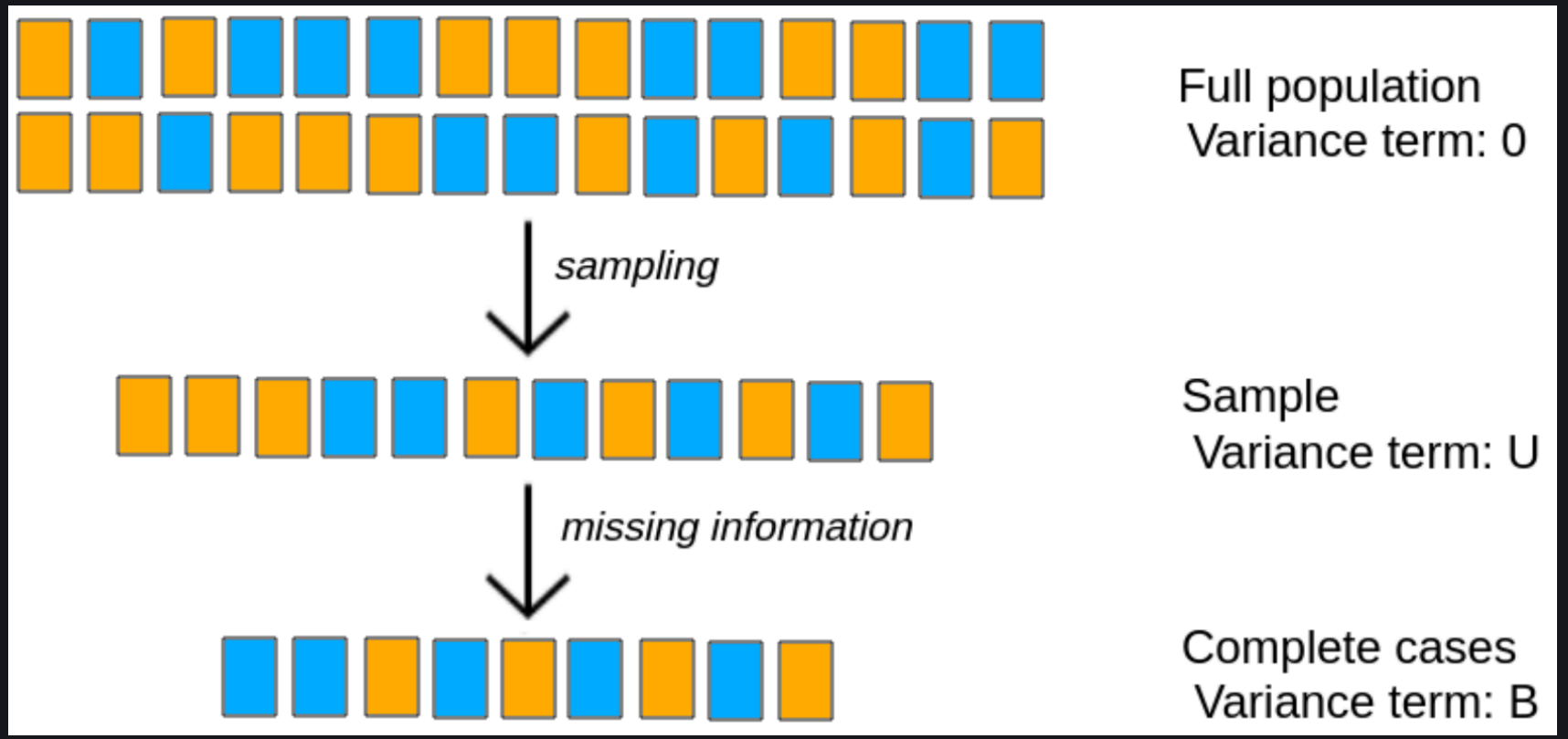
The problem with single imputation strategies

Imputing one value for a missing datum cannot be correct in general, because we don't know what value to impute with certainty (if we did, it wouldn't be missing).

— Donald B. Rubin



(Figure 1.6 from van Buuren 2019)


Full population
Variance term: 0
sampling
Sample
Variance term: U
missing information
Complete cases
Variance term: B

Total variance (Rubin's rule)

It can be shown mathematically that

$$\text{Total variance} = U + B + B\frac{1}{m}$$

- m is the number of imputed datasets
- U is the variance due to sampling
- B is the variance due to missing values
- $B\frac{1}{m}$ is the extra variance due to using a finite number of imputations and the need to estimate the missing model.

MICE

Multiple imputation by chained equations: A specific algorithm (method) for performing data analysis with missing information.

Also known as imputation with fully conditional specification (FCS).

Specifies imputation models variable-by-variable for each variable with missing information.

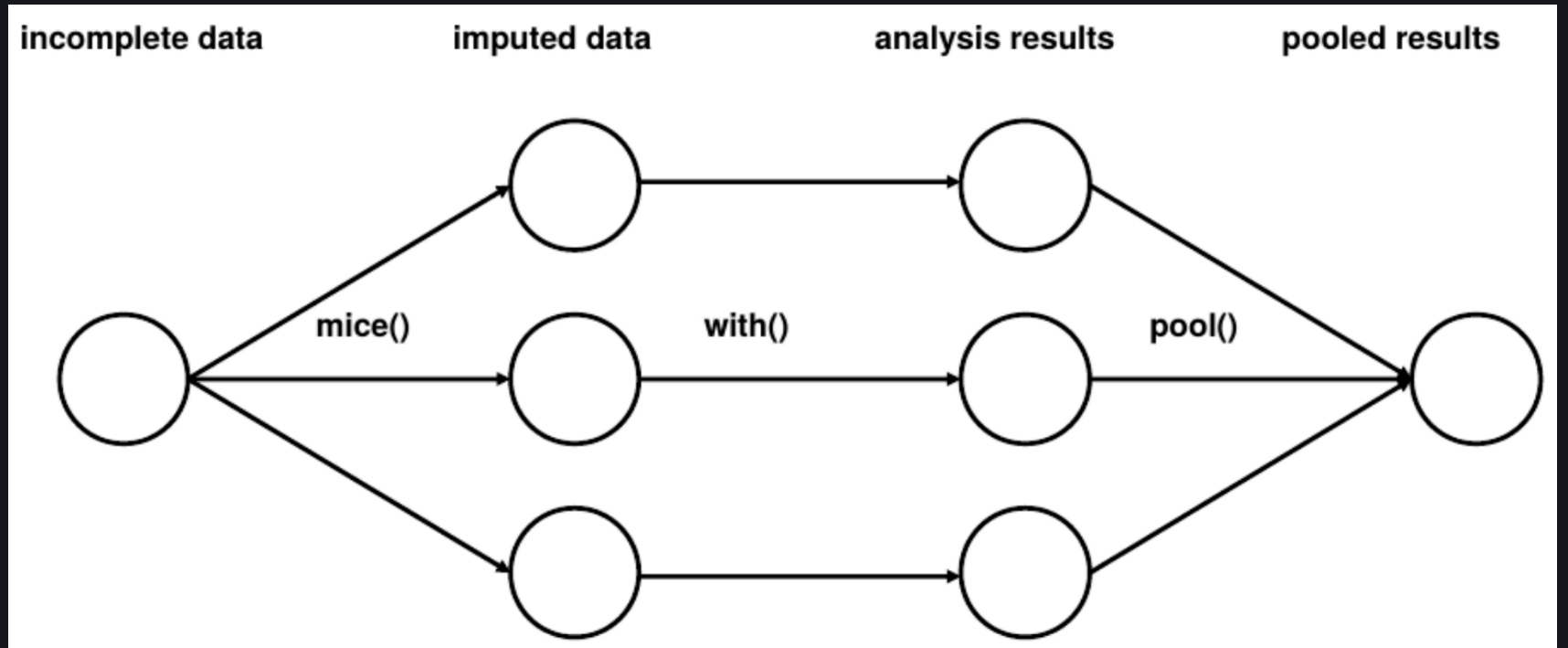
Iteratively updates best guesses to allow all variables (even those with missing information) to inform the imputation of the others.

MICE in R

MICE is implemented in the `mice` package in R:

```
library(mice)
data <- data.frame(X = X, Y = Y, Z = Z)
set.seed(22)
imps <- mice(data, print = FALSE, m = 10)
fits <- with(imps, lm(Y ~ X + Z))
res <- pool(fits)
summary(res)[, c(1:3,6)]
```

```
         term     estimate std.error      p.value
1 (Intercept) -0.007041764 0.1409787 9.602336e-01
2           X  2.072830228 0.1365646 2.759248e-30
3           Z  1.030463434 0.1992245 7.903696e-07
```



(Figure 1 from van Buuren & Groothuis-Oudshoorn 2011)

```r
#Estimate from complete case analysis
round(summary(lm(Y ~ X + Z, data))$coefficients,4)[2,]
```

```
 Estimate Std. Error    t value   Pr(>|t|)
   2.0358     0.1452    14.0252     0.0000
```

```r
#Estimate from model with X imputed by stochastic regression
round(summary(lm(Y ~ X_stocregimp + Z))$coefficients,4)[2,]
```

```
 Estimate Std. Error    t value   Pr(>|t|)
   2.0430     0.1309    15.6060     0.0000
```

```r
#Estimate from mice model (default settings)
round(summary(res)[2, c(2,3,4,6)],4)
```

```
  estimate std.error statistic p.value
2   2.0728    0.1366   15.1784       0
```

Output from MICE

`mice` gives estimates of B (b), U (ubar), T (t = std.error ^{2}), as well as $\lambda = \frac{B(1+1/m)}{T}$ (proportion of variance due to missing), and more

```
summary(res, type = "all")
```

```
         term  m     estimate std.error   statistic
1 (Intercept) 10 -0.007041764 0.1409787 -0.04994913
2           X 10  2.072830228 0.1365646 15.17839086
3           Z 10  1.030463434 0.1992245  5.17237362
        df       p.value       riv    lambda       fmi
1 141.1110 9.602336e-01 0.1228197 0.1093851 0.1217452
2 126.6437 2.759248e-30 0.1540145 0.1334598 0.1468278
3 138.9950 7.903696e-07 0.1271987 0.1128450 0.1253405
        ubar           b          t dfcom
1 0.01770097 0.001976389 0.01987499   197
```

Variable level imputation models

- Numerical variables : Predictive mean matching (PMM). A fusion between regression imputation and hot deck imputation: Use regression to find a selection of plausible “donor values”, choose one at random among them.
- Categorical variables : Logistic regression (binary) or multinomial logistic regression (polyreg).
- Other types of variables : See `?mice::mice` for details and options.

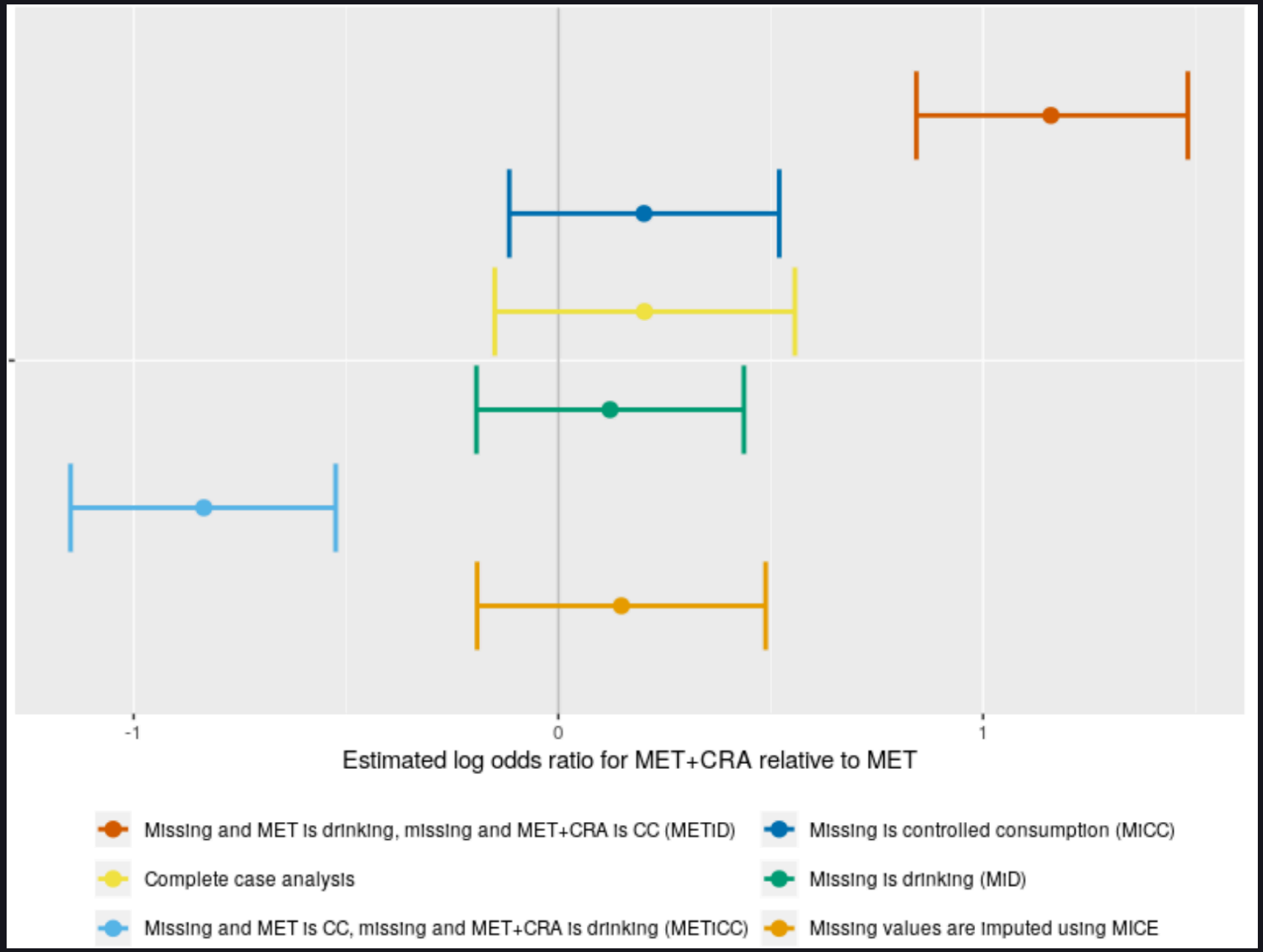
Exercise

Table 3.1: Estimated log odds ratios from the model of controlled consumption status using all full covariate adjustment. The reported estimates are on log odds ratio scale and they are computed relative to the following reference category: Treament MET; Gender male; Country Denmark; Age 60; Education none; No partner; Low ADS; Previous treatments 0. The mean log odds of having a controlled alcohol consumption in this reference group is represented by the intercept estimate. The reported p-values correspond to two-sided z-tests of the null-hypothesis of a zero parameter value.

	Estimate	Std. error	z statistic	p-value
Intercept	-0.3507	0.3050	-1.1499	0.2502
Treatment: MET+CRA	0.2028	0.1801	1.1260	0.2602
Country: USA	0.0736	0.2327	0.3164	0.7517
Country: Germany	-0.0351	0.2522	-0.1392	0.8893
Gender: Female	-0.5543	0.1906	-2.9085	0.0036
Age	0.0677	0.0211	3.2038	0.0014
Married or cohabiting: Yes	0.2270	0.1877	1.2094	0.2265
Severity: Intermediate	-0.0777	0.2307	-0.3367	0.7363
Severity: Substantial or severe	-0.2767	0.4096	-0.6755	0.4994
Education: At most undergraduate degree	0.0518	0.2286	0.2268	0.8206
Education: Graduate or post-graduate	-0.4463	0.2872	-1.5537	0.1202
Previous treatments: 1-2	0.2655	0.2187	1.2140	0.2247
Previous treatments: 3+	0.2938	0.3087	0.9517	0.3413

We fitted five additional models:

- MiD Missing is drinking approach: Treating all missing observations as relapsers (non-controlled consumption).
- MiCC Missing is CC approach: Treating all missing as CC.
- METiD MET is drinking approach: Treating missing observations for patients treated with MET as drinking, while missing obsevrations from MET+CRA-patients are treated as controlled consumption.
- METiCC MET is CC: Treating missing for MET+CRA patients as drinking, while missing from MET-patients are treated as CC.
- MICE Multiple imputation of missing observation using all variables from the primary model and controlled consumption information from previous time points.


-1
0
1
Estimated log odds ratio for MET+CRA relative to MET
Missing and MET is drinking, missing and MET+CRA is CC (METID)
Missing is controlled consumption (MICC)
Complete case analysis
Missing is drinking (MID)
Missing and MET is CC, missing and MET+CRA is drinking (METICC)
Missing values are imputed using MICE

Tutorial in Biostatistics

Statistics in Medicine

Received 3 September 2009, Accepted 14 July 2010 Published online 30 November 2010 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.4067

Multiple imputation using chained equations: Issues and guidance for practice

Ian R. White,[a*†] Patrick Royston[b] and Angela M. Wood[c]

Multiple imputation by chained equations is a flexible and practical approach to handling missing data. We describe the principles of the method and show how to impute categorical and quantitative variables, including skewed variables. We give guidance on how to specify the imputation model and how many imputations are needed. We describe the practical analysis of multiply imputed data, including model building and model checking. We stress the limitations of the method and discuss the possible pitfalls. We illustrate the ideas using a data set in mental health, giving Stata code fragments. Copyright © 2010 John Wiley & Sons, Ltd.

Keywords: missing data; multiple imputation; fully conditional specification

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Article

Statistical Methods in Medical Research
2015, Vol. 24(4) 462–487

DOI: 10.1177/0962280214521348
smm.sagepub.com
SAGE

Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model

Jonathan W Bartlett,[1] Shaun R Seaman,[2] Ian R White[2] and James R Carpenter[1,3] for the Alzheimer's Disease Neuroimaging Initiative*

Abstract

Missing covariate data commonly occur in epidemiological and clinical research, and are often dealt with using multiple imputation. Imputation of partially observed covariates is complicated if the substantive model is non-linear (e.g. Cox proportional hazards model), or contains non-linear (e.g. squared) or interaction terms, and standard software implementations of multiple imputation may impute covariates from models that are incompatible with such substantive models. We show how imputation by fully conditional specification, a popular approach for performing multiple imputation, can be modified

https://doi.org/10.1177/0962280214521348