# 432 Class 12 Slides

github.com/THOMASELOVE/2020-432

2020-02-25

# Setup

```
library(magrittr); library(janitor); library(here)
library(knitr)
library(naniar)
library(broom)

library(mice)
  # mice = multiple imputation through chained equations

library(tidyverse)

theme_set(theme_bw())
```

# Today's Goals

Use multiple imputation to deal with missing data in fitting:

- linear regression with `lm`
- logistic regression with `glm`

using the `mice` package. (MICE = Multiple Imputation through Chained Equations)

## Useful (if somewhat dated) Sources

- https://thomasleeper.com/Rcourse/Tutorials/mi.html.
- https://stats.idre.ucla.edu/r/faq/how-do-i-perform-multiple-imputation-using-predictive-mean-matching-in-r/

# Multiple Imputation: Potential and Pitfalls

# Sterne et al. 2009 *BMJ*

Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls

> *In this article, we review the reasons why missing data may lead to bias and loss of information in epidemiological and clinical research. We discuss the circumstances in which multiple imputation may help by reducing bias or increasing precision, as well as describing potential pitfalls in its application. Finally, we describe the recent use and reporting of analyses using multiple imputation in general medical journals, and suggest guidelines for the conduct and reporting of such analyses.*

- https://www.bmj.com/content/338/bmj.b2393
- https://doi.org/10.1136/bmj.b2393

# Types of Missing Data (from Sterne et al.)

- **Missing completely at random** There are no systematic differences between the missing values and the observed values. For example, blood pressure measurements may be missing because of breakdown of an automatic sphygmomanometer.
- **Missing at random** Any systematic difference between the missing values and the observed values can be explained by differences in observed data. For example, missing blood pressure measurements may be lower than measured blood pressures but only because younger people may be more likely to have missing blood pressure measurements.
- **Missing not at random** Even after the observed data are taken into account, systematic differences remain between the missing values and the observed values. For example, people with high blood pressure may be more likely to miss clinic appointments because they have headaches

"Missing at random" is an assumption that justifies the analysis, not a property of the data.

# Pitfalls When Using Multiple Imputation (Sterne et al.)

**Data that are missing not at random**

- Some data are inherently missing not at random because it is not possible to account for systematic differences between the missing values and the observed values using the observed data.
- In such cases multiple imputation may give misleading results. Those results can be either more or less misleading than a complete case analysis.
- For example, consider a study investigating predictors of depression. If individuals are more likely to miss appointments because they are depressed on the day of the appointment, then it may be impossible to make the missing at random assumption plausible, even if a large number of variables is included in the imputation model.
- Where complete cases and multiple imputation analyses give different results, the analyst should attempt to understand why, and this should be reported in publications.

# Ways to Deal with Missing Data (from Sterne et al.)

- There are circumstances in which analyses of **complete cases** will not lead to bias. Missing data in predictor variables do not cause bias in analyses of complete cases if the reasons for the missing data are unrelated to the outcome. Specialist methods to address missing data may lessen the loss of precision and power resulting from exclusion of individuals with incomplete predictor variables but are not required in order to avoid bias.
- **Single Imputation** of missing values usually causes standard errors to be too small, since it fails to account for the fact that we are uncertain about the missing values.
- If we assume data are missing at random, then unbiased and statistically more powerful analyses (compared with analyses based on complete cases) can generally be done by including individuals with incomplete data.

# Multiple Imputation (from Sterne et al.)

- Multiple imputation . . . aims to allow for the uncertainty about the missing data by creating several different plausible imputed data sets and appropriately combining results obtained from each of them.
- The first stage is to create multiple copies of the dataset, with the missing values replaced by imputed values. The imputation procedure must fully account for all uncertainty in predicting the missing values by injecting appropriate variability into the multiple imputed values; we can never know the true values of the missing data.

# Multiple Imputation (from Sterne et al.)

- The second stage is to use standard statistical methods to fit the model of interest to each of the imputed datasets. Estimated associations in each of the imputed datasets will differ because of the variation introduced in the imputation of the missing values, and they are only useful when averaged together to give overall estimated associations. Standard errors are calculated using Rubin's rules, which take account of the variability in results between the imputed datasets, reflecting the uncertainty associated with the missing values.
  - Valid inferences are obtained because we are averaging over the distribution of the missing data given the observed data.

# A Small Example (Sterne et al.)

Consider, for example, a study investigating the association of systolic blood pressure with the risk of subsequent coronary heart disease, in which data on systolic blood pressure are missing for some people.

The probability that systolic blood pressure is missing is likely to:

- decrease with age (doctors are more likely to measure it in older people),
- decrease with increasing body mass index, and
- decrease with history of smoking (doctors are more likely to measure it in people with heart disease risk factors or comorbidities).

If we assume that data are missing at random and that we have systolic blood pressure data on a representative sample of individuals within strata of age, smoking, body mass index, and coronary heart disease, then we can use multiple imputation to estimate the overall association between systolic blood pressure and coronary heart disease.

# Today's Data

```
fram_raw <- read_csv(here("data/framingham.csv")) %>%
    clean_names()
```

See https://www.framinghamheartstudy.org/ for more details.

This particular data set has been used by lots of people, in many different settings, and variations of it are all over the internet. I don't know who the originators were.

# Data Cleanup

```
fram_10 <- fram_raw %>%
    mutate(educ = fct_recode(factor(education),
                             "Some HS" = "1",
                             "HS grad" = "2",
                             "Some Coll" = "3",
                             "Coll grad" = "4")) %>%
    mutate(obese = as.numeric(bmi >= 30)) %>%
    rename(smoker = "current_smoker",
           cigs = "cigs_per_day",
           stroke = "prevalent_stroke",
           highbp = "prevalent_hyp",
           chol = "tot_chol",
           sbp = "sys_bp", dbp = "dia_bp",
           hrate = "heart_rate",
           chd10 = "ten_year_chd") %>%
    select(sbp, chd10, educ, smoker, cigs, bp_meds,
           chol, bmi, obese, glucose)
```

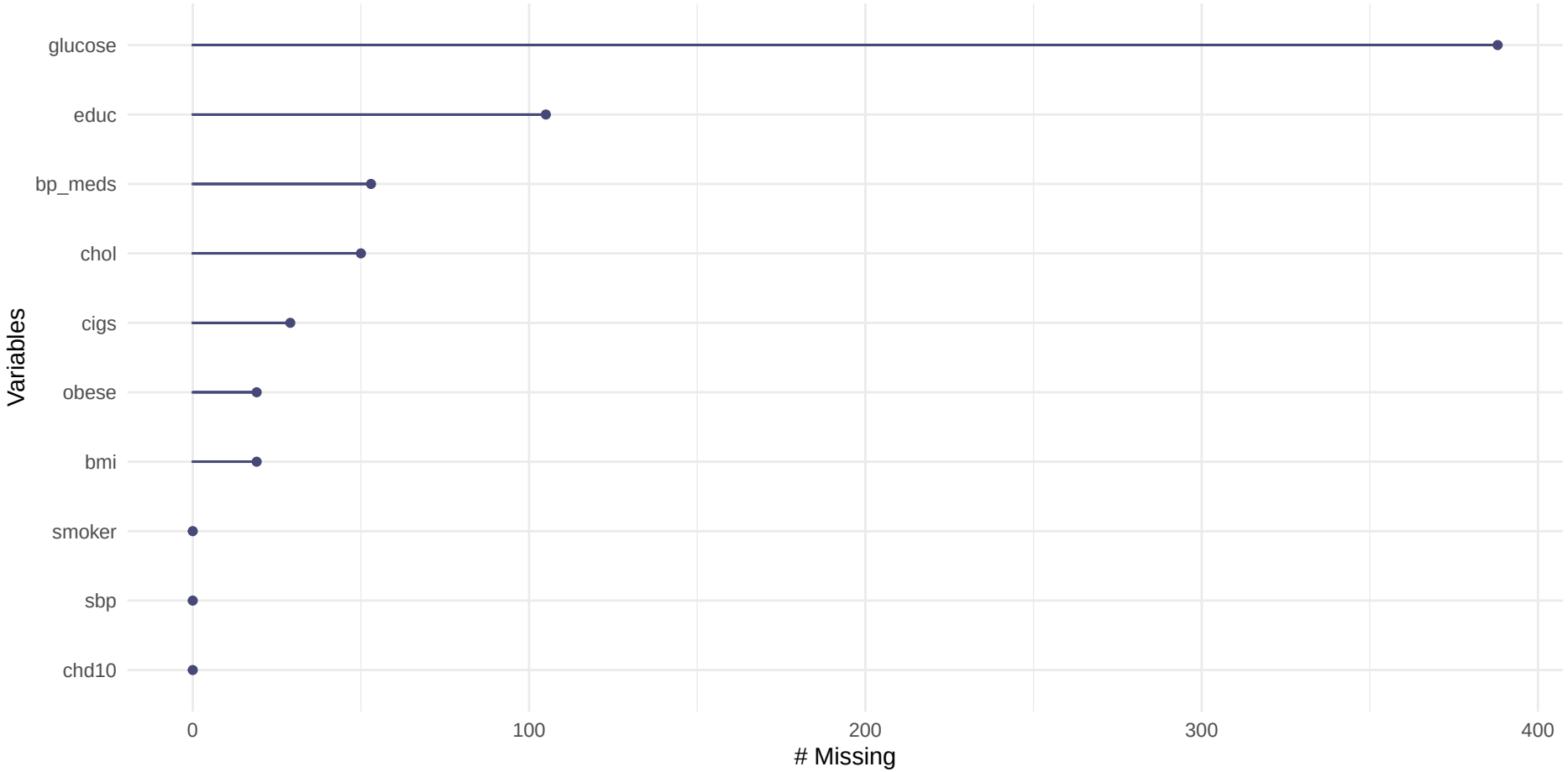
# Data Descriptions (variables we'll use today)

The variables describe n = 4238 adult subjects who were examined at baseline and then followed for ten years to see if they developed incident coronary heart disease during that time.

| Variable | Description |
|---|---|
| `educ` | four-level factor: educational attainment |
| `smoker` | 1 = current smoker at time of examination, else 0 |
| `cigs` | number of cigarettes smoked per day |
| `bp_meds` | 1 = using anti-hypertensive medication at time of exam |
| `chol` | total cholesterol (mg/dl) |
| `sbp` | systolic blood pressure (mm Hg) |
| `bmi` | body mass index in $kg/m^2$ |
| `obese` | 1 if subject's `bmi` is 30 or higher, else 0 |
| `glucose` | blood glucose level in mg/dl |
| `chd10` | 1 = coronary heart disease in next 10 years |

# Which variables are missing data?

```
gg_miss_var(fram_10)
```



# Counts of Missing Data, by Variable

```
miss_var_summary(fram_10) %>%
    filter(n_miss > 0)
```

```
# A tibble: 7 x 3
  variable n_miss pct_miss
  <chr>     <int>    <dbl>
1 glucose     388    9.16
2 educ        105    2.48
3 bp_meds      53    1.25
4 chol         50    1.18
5 cigs         29    0.684
6 bmi          19    0.448
7 obese        19    0.448
```

## Track missingness with shadow

```
fram_10_sh <- bind_shadow(fram_10)
```

# Two Key Settings for Multiple Imputation

- Use linear regression to predict `sbp` accounting for missingness via multiple imputation
  - Predictors include `glucose`, `obese`, `educ`, and `smoker`.
- Use logistic regression to predict `chd10` accounting for missingness via multiple imputation
  - Predictors include `glucose`, `bp_meds`, `chol`, `bmi`, `cigs` and `educ`

# Setting 1: Linear Model for `sbp`

# Model 2 (CC): Two-predictor model for sbp

Suppose we ignore the missingness and just run the model on the data with complete information on sbp, glucose and obese.

```
m2_cc <- fram_10_sh %$% lm(sbp ~ glucose + obese)

tidy(m2_cc, conf.int = TRUE) %>% select(-statistic) %>%
    kable(digits = 3)
```

| term | estimate | std.error | p.value | conf.low | conf.high |
|---|---|---|---|---|---|
| (Intercept) | 121.671 | 1.244 | 0 | 119.232 | 124.110 |
| glucose | 0.111 | 0.015 | 0 | 0.082 | 0.139 |
| obese | 13.532 | 1.045 | 0 | 11.484 | 15.580 |

# Edited Summary of Model 2 (CC)

Suppose we ignore the missingness and just run the model.

```
Residual standard error: 21.42 on 3833 degrees of freedom
  (402 observations deleted due to missingness)
Multiple R-squared:  0.05857,   Adjusted R-squared:  0.05808
F-statistic: 119.2 on 2 and 3833 DF,  p-value: < 2.2e-16
```

```
glance(m2_cc) %>%
    select(r.squared, adj.r.squared, AIC, BIC) %>%
    kable(digits = c(4, 4, 0, 0))
```

| r.squared | adj.r.squared | AIC | BIC |
|---|---|---|---|
| 0.0586 | 0.0581 | 34401 | 34426 |

# Model 4 (CC): Four-predictor model for sbp

```
m4_cc <- fram_10_sh %$%
    lm(sbp ~ glucose + obese + smoker + educ)

tidy(m4_cc, conf.int = TRUE) %>% select(-statistic) %>%
    kable(digits = 3)
```

| term | estimate | std.error | p.value | conf.low | conf.high |
|---|---|---|---|---|---|
| (Intercept) | 127.107 | 1.388 | 0 | 124.385 | 129.829 |
| glucose | 0.106 | 0.015 | 0 | 0.078 | 0.135 |
| obese | 12.304 | 1.066 | 0 | 10.213 | 14.395 |
| smoker | -4.704 | 0.699 | 0 | -6.075 | -3.332 |
| educHS grad | -3.698 | 0.833 | 0 | -5.332 | -2.065 |
| educSome Coll | -4.724 | 1.010 | 0 | -6.704 | -2.744 |
| educColl grad | -5.954 | 1.158 | 0 | -8.225 | -3.683 |

# Edited Summary of Model 4 (CC)

Suppose we ignore the missingness and just run the model.

```
Residual standard error: 21.2 on 3733 degrees of freedom
  (498 observations deleted due to missingness)
Multiple R-squared:  0.08257,   Adjusted R-squared:  0.0811
F-statistic:    56 on 6 and 3733 DF,  p-value: < 2.2e-16
```

```
glance(m4_cc) %>%
    select(r.squared, adj.r.squared, AIC, BIC) %>%
    kable(digits = c(4, 4, 0, 0))
```

| r.squared | adj.r.squared | AIC | BIC |
|---|---|---|---|
| 0.0826 | 0.0811 | 33466 | 33516 |

# Subset of Variables to be used in our models 2 and 4

```
fram_sub <- fram_10 %>%
    select(sbp, glucose, obese, educ, smoker)

miss_var_summary(fram_sub)
```

```
# A tibble: 5 x 3
  variable n_miss pct_miss
  <chr>     <int>    <dbl>
1 glucose     388    9.16
2 educ        105    2.48
3 obese        19    0.448
4 sbp           0    0
5 smoker        0    0
```

# Create multiple imputations for this subset

```
set.seed(4322020)
fram_mice24 <- mice(fram_sub, m = 20)
```

```
 iter imp variable
  1   1  glucose  obese  educ
  1   2  glucose  obese  educ
  1   3  glucose  obese  educ
  1   4  glucose  obese  educ
  1   5  glucose  obese  educ
  1   6  glucose  obese  educ
  1   7  glucose  obese  educ
  1   8  glucose  obese  educ
  1   9  glucose  obese  educ
  1   10  glucose  obese  educ
  1   11  glucose  obese  educ
  1   12  glucose  obese  educ
  1   13  glucose  obese  educ
```

# Summary Information about Imputation Process

```
summary(fram_mice24)
```

```
Class: mids
Number of multiple imputations:  20
Imputation methods:
      sbp   glucose     obese      educ    smoker
       ""     "pmm"     "pmm" "polyreg"        ""
PredictorMatrix:
        sbp glucose obese educ smoker
sbp       0       1     1    1      1
glucose   1       0     1    1      1
obese     1       1     0    1      1
educ      1       1     1    0      1
smoker    1       1     1    1      0
```

# Run Model 2 on each imputed data frame

```
m2_mods <- with(fram_mice24, lm(sbp ~ glucose + obese))
summary(m2_mods)
```

```
# A tibble: 60 x 5
   term        estimate std.error statistic  p.value
   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
 1 (Intercept)  122.        1.17      104.   0.
 2 glucose        0.111     0.0137      8.10 7.07e-16
 3 obese         13.4       0.986      13.6  4.54e-41
 4 (Intercept)  121.        1.16      104.   0.
 5 glucose        0.120     0.0136      8.78 2.32e-18
 6 obese         13.2       0.984      13.4  2.13e-40
 7 (Intercept)  122.        1.17      104.   0.
 8 glucose        0.108     0.0136      7.89 3.90e-15
 9 obese         13.5       0.984      13.7  6.50e-42
10 (Intercept)  121.        1.18      103.   0.
# ... with 50 more rows
```

# Pool Results across the 20 imputations

```
m2_pool <- pool(m2_mods)
summary(m2_pool, conf.int = TRUE, conf.level = 0.95)
```

```
                estimate  std.error statistic       df
(Intercept) 121.1956852 1.24145482  97.62392 1251.744
glucose       0.1154547 0.01466823   7.87107 1050.433
obese        13.2926318 0.99045467  13.42074 4115.498
                 p.value       2.5 %     97.5 %
(Intercept) 0.00000e+00 118.7601235 123.631247
glucose     8.65974e-15   0.0866723   0.144237
obese       0.00000e+00  11.3508052  15.234458
```

# Model 2 (Complete Cases vs. Multiple Imputation)

```
tidy(m2_cc, conf.int = TRUE) %>% kable(digits = 3)
```

| term | estimate | std.error | statistic | p.value | conf.low | conf.high |
|---|---|---|---|---|---|---|
| (Intercept) | 121.671 | 1.244 | 97.792 | 0 | 119.232 | 124.110 |
| glucose | 0.111 | 0.015 | 7.577 | 0 | 0.082 | 0.139 |
| obese | 13.532 | 1.045 | 12.954 | 0 | 11.484 | 15.580 |

```
summary(m2_pool, conf.int = TRUE, conf.level = 0.95) %>%
    select(-df) %>% kable(digits = 3)
```

| | estimate | std.error | statistic | p.value | 2.5 % | 97.5 % |
|---|---|---|---|---|---|---|
| (Intercept) | 121.196 | 1.241 | 97.624 | 0 | 118.760 | 123.631 |
| glucose | 0.115 | 0.015 | 7.871 | 0 | 0.087 | 0.144 |
| obese | 13.293 | 0.990 | 13.421 | 0 | 11.351 | 15.234 |

# More Details on Multiple Imputation Modeling

```
m2_pool
```

```
Class: mipo    m = 20
                 estimate          ubar             b
(Intercept) 121.1956852 1.3856592382 1.481436e-01
glucose       0.1154547 0.0001906054 2.338243e-05
obese        13.2926318 0.9718971045 8.669863e-03
                       t dfcom       df          riv
(Intercept) 1.5412100706  4235 1251.744 0.112257637
glucose     0.0002151569  4235 1050.433 0.128808274
obese       0.9810004604  4235 4115.498 0.009366584
                  lambda          fmi
(Intercept) 0.100927729 0.102360806
glucose     0.114109966 0.115791878
obese       0.009279665 0.009760773
```

- `fmi` = fraction of missing information due to nonresponse

# Model 4 run on each imputed data frame

```
m4_mods <- with(fram_mice24, lm(sbp ~ glucose +
                                  obese + smoker + educ))
summary(m4_mods)
```

```
# A tibble: 140 x 5
   term           estimate std.error statistic  p.value
   <chr>             <dbl>     <dbl>     <dbl>    <dbl>
 1 (Intercept)     127.        1.29       98.4  0.
 2 glucose           0.102     0.0135      7.54 5.62e-14
 3 obese            11.9       0.985      12.0  6.82e-33
 4 smoker           -4.42      0.656      -6.74 1.80e-11
 5 educHS grad      -3.70      0.780      -4.75 2.10e- 6
 6 educSome Coll    -5.34      0.945      -5.65 1.68e- 8
 7 educColl grad    -6.21      1.09       -5.70 1.26e- 8
 8 (Intercept)     127.        1.29       98.5  0.
 9 glucose           0.112     0.0135      8.31 1.28e-16
10 obese            11.7       0.984      11.9  3.71e-32
# ... with 130 more rows
```

# Pool Results across the five imputations

```
m4_pool <- pool(m4_mods)
summary(m4_pool, conf.int = TRUE, conf.level = 0.95)
```

```
                  estimate std.error statistic       df
(Intercept)   126.9270253 1.3588718 93.406183 1571.582
glucose         0.1075469 0.0145208  7.406405 1064.584
obese          11.7897138 0.9896541 11.912964 4146.672
smoker         -4.4254492 0.6561702 -6.744362 4213.845
educHS grad    -3.6929816 0.7939553 -4.651372 3139.413
educSome Coll  -5.2738134 0.9571700 -5.509798 3704.931
educColl grad  -6.1994411 1.0983288 -5.644431 3830.221
                    p.value        2.5 %      97.5 %
(Intercept)   0.000000e+00 124.26163280 129.5924178
glucose       2.633449e-13   0.07905428   0.1360396
obese         0.000000e+00   9.84946106  13.7299665
smoker        1.745493e-11  -5.71188874  -3.1390097
educHS grad   3.433553e-06  -5.24970562  -2.1362576
educSome Coll 3.837089e-08  -7.15044529  -3.3971816
```

# Complete Cases Result (Model 4)

```
tidy(m4_cc, conf.int = TRUE) %>% select(-statistic) %>%
    kable(digits = 3)
```

| term | estimate | std.error | p.value | conf.low | conf.high |
|---|---|---|---|---|---|
| (Intercept) | 127.107 | 1.388 | 0 | 124.385 | 129.829 |
| glucose | 0.106 | 0.015 | 0 | 0.078 | 0.135 |
| obese | 12.304 | 1.066 | 0 | 10.213 | 14.395 |
| smoker | -4.704 | 0.699 | 0 | -6.075 | -3.332 |
| educHS grad | -3.698 | 0.833 | 0 | -5.332 | -2.065 |
| educSome Coll | -4.724 | 1.010 | 0 | -6.704 | -2.744 |
| educColl grad | -5.954 | 1.158 | 0 | -8.225 | -3.683 |

# Multiple Imputation Result (Model 4)

```
summary(m4_pool, conf.int = TRUE) %>%
  select(-statistic, -df) %>% kable(digits = 3)
```

| | estimate | std.error | p.value | 2.5 % | 97.5 % |
|---|---|---|---|---|---|
| (Intercept) | 126.927 | 1.359 | 0 | 124.262 | 129.592 |
| glucose | 0.108 | 0.015 | 0 | 0.079 | 0.136 |
| obese | 11.790 | 0.990 | 0 | 9.849 | 13.730 |
| smoker | -4.425 | 0.656 | 0 | -5.712 | -3.139 |
| educHS grad | -3.693 | 0.794 | 0 | -5.250 | -2.136 |
| educSome Coll | -5.274 | 0.957 | 0 | -7.150 | -3.397 |
| educColl grad | -6.199 | 1.098 | 0 | -8.353 | -4.046 |

# More Details on Multiple Imputation Modeling

```
m4_pool
```

```
Class: mipo    m = 20
                   estimate          ubar             b
(Intercept)   126.9270253 1.6900562422 0.1490250146
glucose         0.1075469 0.0001870152 0.0000227033
obese          11.7897138 0.9721197883 0.0069480605
smoker         -4.4254492 0.4295449894 0.0009660582
educHS grad    -3.6929816 0.6069005757 0.0223471066
educSome Coll  -5.2738134 0.8950889474 0.0200814547
educColl grad  -6.1994411 1.1828536383 0.0223547953
                        t dfcom        df         riv
(Intercept)   1.8465325075  4231 1571.582 0.092586425
glucose       0.0002108536  4231 1064.584 0.127468096
obese         0.9794152518  4231 4146.672 0.007504696
smoker        0.4305593505  4231 4213.845 0.002361478
educHS grad   0.6303650376  4231 3139.413 0.038662778
educSome Coll 0.9161744748  4231 3704.931 0.023556907
```

# Estimate $R^2$ or adjusted $R^2$?

```
pool.r.squared(m2_mods)
```

```
           est      lo 95      hi 95 fmi
R^2 0.05923026 0.04601681 0.07387826 NaN
```

```
pool.r.squared(m2_mods, adjusted = TRUE)
```

```
               est      lo 95      hi 95 fmi
adj R^2 0.05878594 0.04561876 0.07339091 NaN
```

```
pool.r.squared(m4_mods)
```

```
           est     lo 95      hi 95 fmi
R^2 0.08144661 0.0661401 0.09804029 NaN
```

```
pool.r.squared(m4_mods, adjusted = TRUE)
```

```
               est      lo 95      hi 95 fmi
adj R^2 0.08014394 0.06494473 0.09663925 NaN
```

# Compare Model 4 to Model 2 after imputation

The models must be nested for this to be appropriate. We'll use the Wald test after a linear regression fit.

```
fit4 <- with(fram_mice24,
             expr = lm(sbp ~ glucose + obese + smoker + educ))
fit2 <- with(fram_mice24,
             expr = lm(sbp ~ glucose + obese))

pool.compare(fit4, fit2, method = "wald")$pvalue
```

```
     [,1]
[1,]    0
```

# Setting 2: Logistic Model for `chd10`

# Model 3 (CC): Three-predictor model for chd10

Suppose we ignore the missingness and just run the model on the data with complete information on glucose, bp_meds and cigs.

```
m3_cc <- fram_10_sh %$% glm(chd10 ~ glucose + bp_meds + cigs,
                            family = binomial)

tidy(m3_cc, exponentiate = TRUE, conf.int = TRUE) %>%
    select(-statistic) %>% kable(digits = 3)
```

| term | estimate | std.error | p.value | conf.low | conf.high |
|---|---|---|---|---|---|
| (Intercept) | 0.063 | 0.151 | 0 | 0.046 | 0.084 |
| glucose | 1.011 | 0.002 | 0 | 1.007 | 1.014 |
| bp_meds | 2.736 | 0.211 | 0 | 1.791 | 4.105 |
| cigs | 1.015 | 0.004 | 0 | 1.007 | 1.022 |

# Model 6 (CC): Six-predictor model for chd10

```
m6_cc <- fram_10_sh %$% glm(chd10 ~ glucose + bp_meds + cigs +
                              educ + chol + bmi,
                            family = binomial)

tidy(m6_cc, exponentiate = TRUE, conf.int = TRUE) %>%
    select(-statistic) %>% kable(digits = 3)
```

| term | estimate | std.error | p.value | conf.low | conf.high |
|---|---|---|---|---|---|
| (Intercept) | 0.012 | 0.410 | 0.000 | 0.005 | 0.026 |
| glucose | 1.009 | 0.002 | 0.000 | 1.006 | 1.013 |
| bp_meds | 2.418 | 0.217 | 0.000 | 1.564 | 3.676 |
| cigs | 1.016 | 0.004 | 0.000 | 1.009 | 1.023 |
| educHS grad | 0.595 | 0.117 | 0.000 | 0.472 | 0.747 |
| educSome Coll | 0.634 | 0.144 | 0.001 | 0.475 | 0.835 |
| educColl grad | 0.753 | 0.157 | 0.070 | 0.550 | 1.017 |
| chol | 1.005 | 0.001 | 0.000 | 1.003 | 1.007 |
| bmi | 1.034 | 0.011 | 0.003 | 1.012 | 1.057 |

# Subset of Variables to be used in our models 3 and 6

```
fram_sub36 <- fram_10_sh %>%
    select(chd10, glucose, bp_meds, cigs, educ, chol, bmi)

miss_var_summary(fram_sub36)
```

```
# A tibble: 7 x 3
  variable n_miss pct_miss
  <chr>     <int>    <dbl>
1 glucose     388    9.16
2 educ        105    2.48
3 bp_meds      53    1.25
4 chol         50    1.18
5 cigs         29    0.684
6 bmi          19    0.448
7 chd10         0    0
```

# Create multiple imputations for this subset

```
set.seed(432202036)
fram_mice36 <- mice(fram_sub36, m = 10)
```

```
 iter imp variable
  1   1  glucose  bp_meds  cigs  educ  chol  bmi
  1   2  glucose  bp_meds  cigs  educ  chol  bmi
  1   3  glucose  bp_meds  cigs  educ  chol  bmi
  1   4  glucose  bp_meds  cigs  educ  chol  bmi
  1   5  glucose  bp_meds  cigs  educ  chol  bmi
  1   6  glucose  bp_meds  cigs  educ  chol  bmi
  1   7  glucose  bp_meds  cigs  educ  chol  bmi
  1   8  glucose  bp_meds  cigs  educ  chol  bmi
  1   9  glucose  bp_meds  cigs  educ  chol  bmi
  1   10  glucose  bp_meds  cigs  educ  chol  bmi
  2   1  glucose  bp_meds  cigs  educ  chol  bmi
  2   2  glucose  bp_meds  cigs  educ  chol  bmi
  2   3  glucose  bp_meds  cigs  educ  chol  bmi
```

# Summary information about Imputation Process

```
summary(fram_mice36)
```

```
Class: mids
Number of multiple imputations:  10
Imputation methods:
    chd10   glucose   bp_meds      cigs      educ      chol
       ""     "pmm"     "pmm"     "pmm" "polyreg"     "pmm"
      bmi
    "pmm"
PredictorMatrix:
        chd10 glucose bp_meds cigs educ chol bmi
chd10       0       1       1    1    1    1   1
glucose     1       0       1    1    1    1   1
bp_meds     1       1       0    1    1    1   1
cigs        1       1       1    0    1    1   1
educ        1       1       1    1    0    1   1
chol        1       1       1    1    1    0   1
```

# Run Model 3 on each imputed data frame

```
m3_mods <- with(fram_mice36,
                glm(chd10 ~ glucose + bp_meds + cigs,
                    family = binomial))
summary(m3_mods)
```

```
# A tibble: 40 x 5
   term        estimate std.error statistic  p.value
   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
 1 (Intercept) -2.89       0.147     -19.7  1.87e-86
 2 glucose      0.0117     0.00156     7.52 5.68e-14
 3 bp_meds      1.05       0.201       5.23 1.71e- 7
 4 cigs         0.0151     0.00343     4.38 1.16e- 5
 5 (Intercept) -2.72       0.138     -19.7  3.03e-86
 6 glucose      0.00974    0.00146     6.68 2.39e-11
 7 bp_meds      1.06       0.198       5.34 9.35e- 8
 8 cigs         0.0148     0.00344     4.30 1.73e- 5
 9 (Intercept) -2.85       0.148     -19.3  5.08e-83
10 glucose      0.0112     0.00157     7.11 1.15e-12
```

# Pool Results across the 10 imputations

```
m3_pool <- pool(m3_mods)
summary(m3_pool, exponentiate = TRUE,
        conf.int = TRUE, conf.level = 0.95) %>%
    select(-df) %>% kable(digits = c(3, 3, 2, 2, 3, 3))
```

| | estimate | std.error | statistic | p.value | 2.5 % | 97.5 % |
|---|---|---|---|---|---|---|
| (Intercept) | 0.061 | 0.156 | -17.99 | 0 | 0.045 | 0.082 |
| glucose | 1.011 | 0.002 | 6.35 | 0 | 1.007 | 1.014 |
| bp_meds | 2.865 | 0.202 | 5.21 | 0 | 1.927 | 4.259 |
| cigs | 1.015 | 0.003 | 4.41 | 0 | 1.008 | 1.022 |

# Comparing Model 3 Results

## Complete Cases

| term | estimate | std.error | p.value | conf.low | conf.high |
|---|---|---|---|---|---|
| (Intercept) | 0.063 | 0.151 | 0 | 0.046 | 0.084 |
| glucose | 1.011 | 0.002 | 0 | 1.007 | 1.014 |
| bp_meds | 2.736 | 0.211 | 0 | 1.791 | 4.105 |
| cigs | 1.015 | 0.004 | 0 | 1.007 | 1.022 |

## After Multiple Imputation

| | estimate | std.error | p.value | 2.5 % | 97.5 % |
|---|---|---|---|---|---|
| (Intercept) | 0.061 | 0.156 | 0 | 0.045 | 0.082 |
| glucose | 1.011 | 0.002 | 0 | 1.007 | 1.014 |
| bp_meds | 2.865 | 0.202 | 0 | 1.927 | 4.259 |
| cigs | 1.015 | 0.003 | 0 | 1.008 | 1.022 |

# Run Model 6 on each imputed data frame

```
m6_mods <- with(fram_mice36, glm(chd10 ~ glucose +
                bp_meds + cigs + educ + chol + bmi,
                     family = binomial))
summary(m6_mods)
```

```
# A tibble: 90 x 5
   term           estimate std.error statistic  p.value
   <chr>             <dbl>     <dbl>     <dbl>    <dbl>
 1 (Intercept)    -4.29      0.379       -11.3  1.23e-29
 2 glucose         0.0109    0.00154       7.04 1.92e-12
 3 bp_meds         0.917     0.205         4.47 7.91e- 6
 4 cigs            0.0166    0.00344       4.82 1.47e- 6
 5 educHS grad    -0.547     0.109        -5.02 5.18e- 7
 6 educSome Coll  -0.443     0.133        -3.33 8.61e- 4
 7 educColl grad  -0.245     0.146        -1.68 9.29e- 2
 8 chol            0.00417   0.000952      4.38 1.16e- 5
 9 bmi             0.0270    0.0106        2.56 1.05e- 2
10 (Intercept)    -4.19      0.376       -11.1  7.78e-29
```

# Pool Results across the 10 imputations

```
m6_pool <- pool(m6_mods)
summary(m6_pool, exponentiate = TRUE,
        conf.int = TRUE, conf.level = 0.95) %>%
    select(-df) %>% kable(digits = c(3, 3, 2, 2, 3, 3))
```

| | estimate | std.error | statistic | p.value | 2.5 % | 97.5 % |
|---|---|---|---|---|---|---|
| (Intercept) | 0.013 | 0.384 | -11.23 | 0.00 | 0.006 | 0.028 |
| glucose | 1.010 | 0.002 | 6.03 | 0.00 | 1.007 | 1.013 |
| bp_meds | 2.510 | 0.207 | 4.45 | 0.00 | 1.674 | 3.764 |
| cigs | 1.017 | 0.003 | 4.87 | 0.00 | 1.010 | 1.024 |
| educHS grad | 0.591 | 0.110 | -4.76 | 0.00 | 0.476 | 0.734 |
| educSome Coll | 0.675 | 0.134 | -2.93 | 0.00 | 0.519 | 0.879 |
| educColl grad | 0.811 | 0.147 | -1.42 | 0.15 | 0.608 | 1.082 |
| chol | 1.004 | 0.001 | 4.40 | 0.00 | 1.002 | 1.006 |
| bmi | 1.031 | 0.011 | 2.84 | 0.00 | 1.009 | 1.052 |

# Comparing Model 6 Results

## Complete Cases

| term | estimate | std.error | p.value | conf.low | conf.high |
|---|---|---|---|---|---|
| (Intercept) | 0.012 | 0.410 | 0.000 | 0.005 | 0.026 |
| glucose | 1.009 | 0.002 | 0.000 | 1.006 | 1.013 |
| bp_meds | 2.418 | 0.217 | 0.000 | 1.564 | 3.676 |
| cigs | 1.016 | 0.004 | 0.000 | 1.009 | 1.023 |
| educHS grad | 0.595 | 0.117 | 0.000 | 0.472 | 0.747 |
| educSome Coll | 0.634 | 0.144 | 0.001 | 0.475 | 0.835 |
| educColl grad | 0.753 | 0.157 | 0.070 | 0.550 | 1.017 |
| chol | 1.005 | 0.001 | 0.000 | 1.003 | 1.007 |
| bmi | 1.034 | 0.011 | 0.003 | 1.012 | 1.057 |

# Comparing Model 6 Results

## After Multiple Imputation

| | estimate | std.error | p.value | 2.5 % | 97.5 % |
|---|---|---|---|---|---|
| (Intercept) | 0.013 | 0.384 | 0.000 | 0.006 | 0.028 |
| glucose | 1.010 | 0.002 | 0.000 | 1.007 | 1.013 |
| bp_meds | 2.510 | 0.207 | 0.000 | 1.674 | 3.764 |
| cigs | 1.017 | 0.003 | 0.000 | 1.010 | 1.024 |
| educHS grad | 0.591 | 0.110 | 0.000 | 0.476 | 0.734 |
| educSome Coll | 0.675 | 0.134 | 0.003 | 0.519 | 0.879 |
| educColl grad | 0.811 | 0.147 | 0.155 | 0.608 | 1.082 |
| chol | 1.004 | 0.001 | 0.000 | 1.002 | 1.006 |
| bmi | 1.031 | 0.011 | 0.005 | 1.009 | 1.052 |

# Compare Model 6 to Model 3 after imputation

Again, these models need to be nested. We'll use the likelihood ratio test after a logistic regression fit.

```
fit6 <- with(fram_mice36,
          expr = glm(chd10 ~ glucose + bp_meds + cigs +
                       educ + chol + bmi, family = binomial))
fit3 <- with(fram_mice36,
          expr = glm(chd10 ~ glucose + bp_meds + cigs,
                       family = binomial))

pool.compare(fit6, fit3, method = "likelihood")$pvalue
```

```
[1] 8.610224e-12
```

# Pitfalls When Using Multiple Imputation (Sterne et al.)

**Omitting the outcome variable from the imputation procedure**

Often an analysis explores the association between one or more predictors and an outcome but some predictors have missing values.

- Here, the outcome carries information about the missing values of the predictors and this information must be used.
- Consider a model relating systolic blood pressure to time to coronary heart disease, fitted to data that have some missing values of systolic blood pressure.
  - When missing systolic blood pressure values are imputed, individuals who develop coronary heart disease should have larger values, on average, than those who remain disease free.
  - Failure to include the coronary heart disease outcome and time to this outcome when imputing the missing systolic blood pressure values would falsely weaken the association between systolic blood pressure and coronary heart disease.

# Pitfalls When Using Multiple Imputation (Sterne et al.)

**Dealing with non-normally distributed variables**

Many multiple imputation procedures assume that data are normally distributed, so including non-normally distributed variables may introduce bias.

- A pragmatic approach here is to transform such variables to approximate normality before imputation and then transform the imputed values back to the original scale.
- Different problems arise when data are missing in binary or categorical variables. Some procedures handle these types of missing data better than others.

# Options within `mice` for imputation approaches

Default methods include:

- `pmm` predictive mean matching (default choice for quantitative variables)
- `logreg` logistic regression (default for binary categorical variables)
- `polyreg` polytomous logistic regression (for nominal multi-categorical variables)
- `polr` proportional odds logistic regression (for ordinal categories)

but there are `cart` methods and many others available, too.

# Pitfalls When Using Multiple Imputation (Sterne et al.)

**Plausibility of missing at random assumption**

- For example, the missing at random assumption may be reasonable if a variable that is predictive of missing data in a covariate of interest is included in the imputation model, but not if the variable is omitted from the model.
- Multiple imputation analyses will avoid bias only if enough variables predictive of missing values are included in the imputation model.
- It is sensible to include a wide range of variables in imputation models, including all variables in the substantive analysis, plus, as far as computationally feasible, all variables predictive of the missing values themselves and all variables influencing the process causing the missing data.

# Guidelines for reporting, I (Sterne et al.)

How should we report on analyses potentially affected by missing data?

- Report the number of missing values for each variable of interest, or the number of cases with complete data for each important component of the analysis. Give reasons for missing values if possible, and indicate how many individuals were excluded because of missing data when reporting the flow of participants through the study. If possible, describe reasons for missing data in terms of other variables (rather than just reporting a universal reason such as treatment failure.)
- Clarify whether there are important differences between individuals with complete and incomplete data—for example, by providing a table comparing the distributions of key exposure and outcome variables in these different groups
- Describe the type of analysis used to account for missing data (eg, multiple imputation), and the assumptions that were made (eg, missing at random)

# Guidelines for reporting, II (Sterne et al.)

How should we report on analyses that involve multiple imputation?

- Provide details of the imputation modeling (software used, key settings, number of imputed datasets, variables included in imputation procedure, etc.)
- If a large fraction of the data is imputed, compare observed and imputed values.
- Where possible, provide results from analyses restricted to complete cases, for comparison with results based on multiple imputation. If there are important differences between the results, suggest explanations.
- It is also desirable to investigate the robustness of key inferences to possible departures from the missing at random assumption, by assuming a range of missing not at random mechanisms in sensitivity analyses.

# Next Up

- Minute Paper after Class 12 due Wednesday at 2 PM
- For those of you who still need to do work on your proposal, the next revision deadline is 9 AM Wednesday
- You'll have access to Quiz 1 at 5 PM Wednesday
- No class Thursday. Next Tuesday's class will be about ridge regression and the lasso.