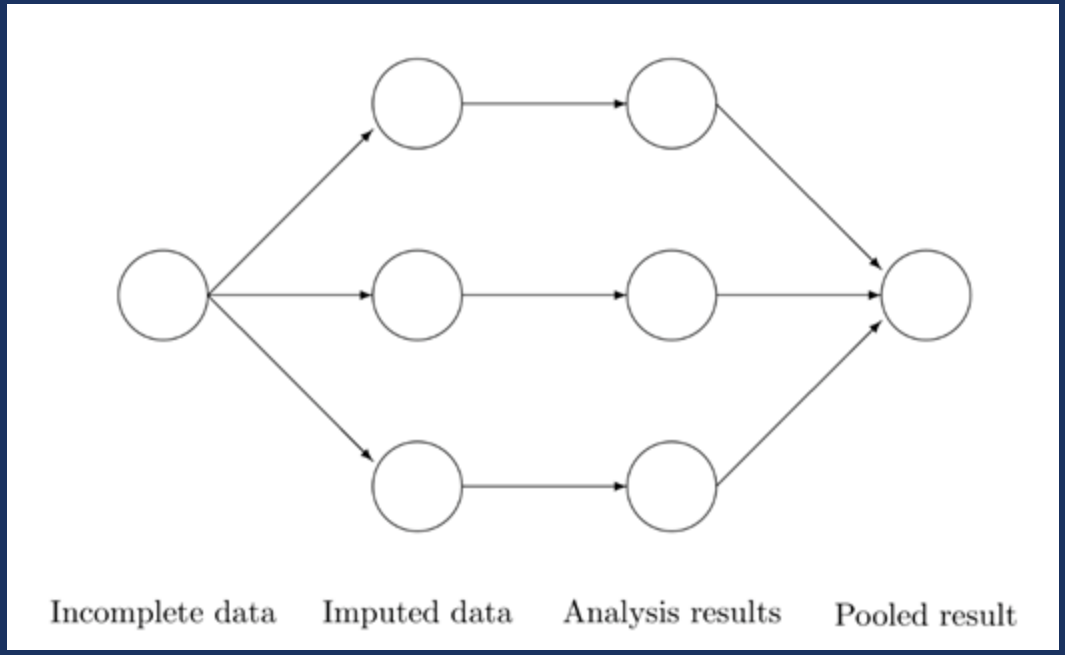
DEALING WITH MISSING DATA

MULTIPLE IMPUTATION

ESTELLE HIGGINS, SUMMER 2022

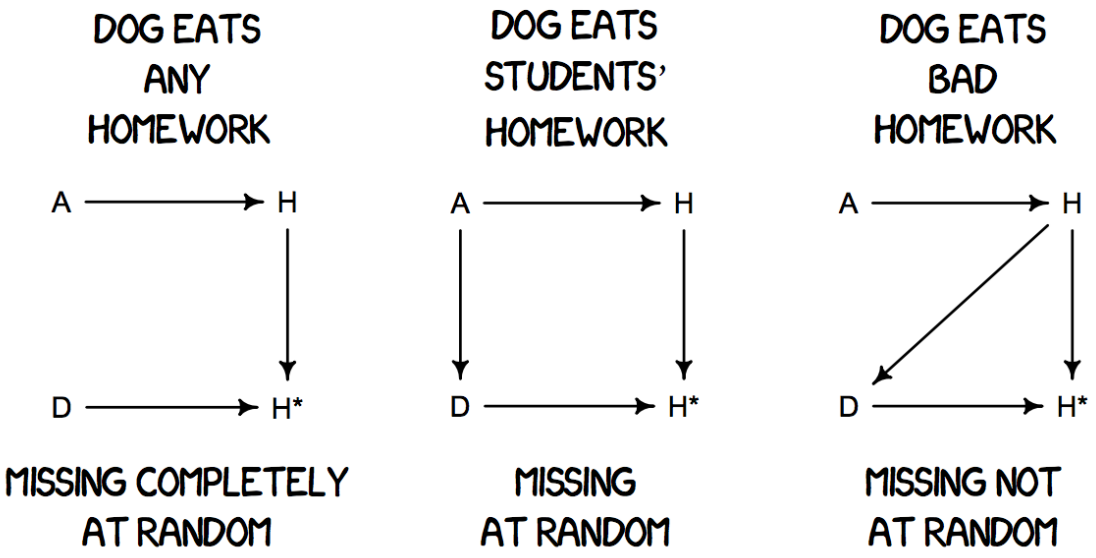


MISSING DATA

ASSUMPTIONS:

- Missing Completely at Random (MCAR)
  - *Cause of missingness is unrelated to data*
- Missing at Random (MAR)
  - *Missingness predicted from other information about subject*
- Missing Not at Random (MNAR)
  - *Missingness is related to what is missing*

H: Homework
H*: Homework with missing values
A: Attribute of student
D: Dog (missingness mechanism)



INTENTION-TO-TREAT ANALYSIS

Complete-Case/Per-Protocol Analysis:

- Include only those who complete study
  - "listwise deletion"
  - May bias intervention effect estimates

Pairwise Deletion / Available-Case Analysis

- Use all observed data

Intention-to-Treat Analysis:

- Include all randomized subjects (including missing)
  - "once randomized, always analyzed"
  - Unbiased – but best if no missing data

R PACKAGES

- mice
  - broom.mixed *(needed for viewing pooled results)*
- lattice
- VIM *(for visuals)*
- Amelia *(I didn't use this)*

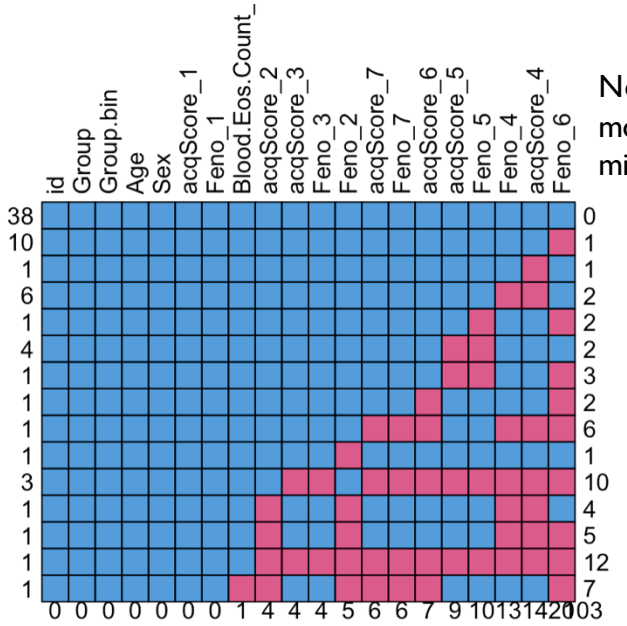
MISSING DATA

Visualizing Missing Data:

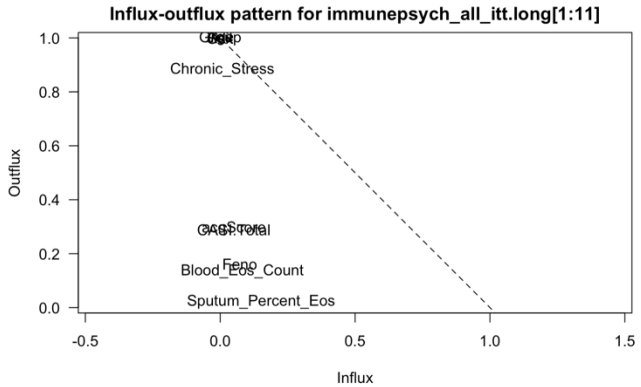
- md.pattern()
  - md.pairs() $rr, $rm, $mr, $mm
- fluxplot()
  - influx & outflux
- aggr()

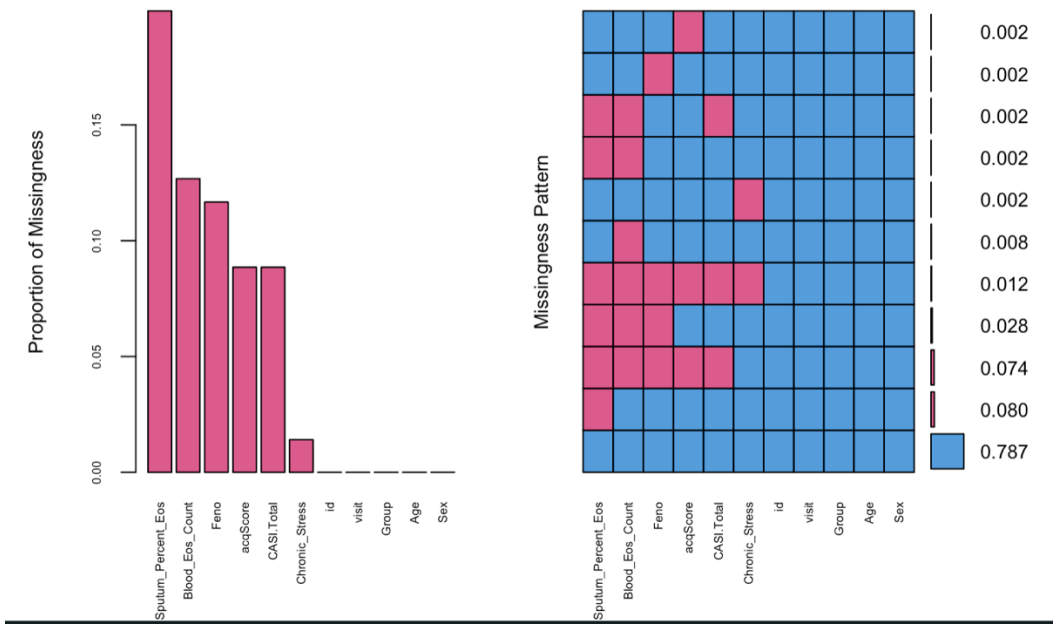
- ≤ 50% missing *(some say 20%)*

→ More info



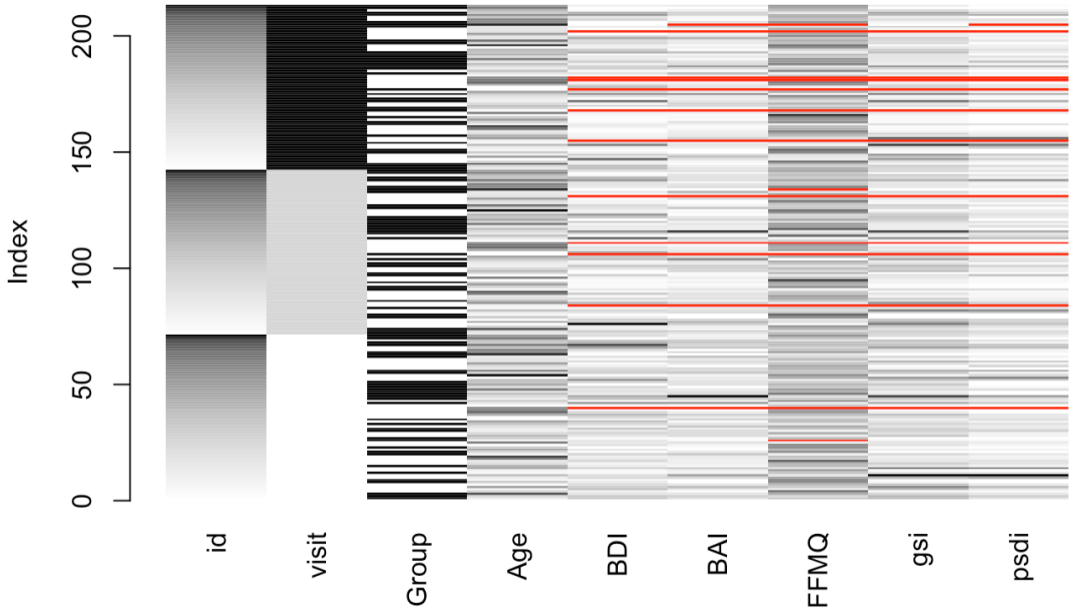
Non-monotone missingness





MISSING DATA

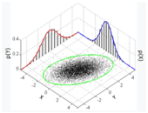
- matrixplot()
  - sort by different variables to look for patterns in missing data



SOME OPTIONS*

- **Single Imputation:**
  - Last or Baseline Observation Carried Forward
  - Mean Imputation

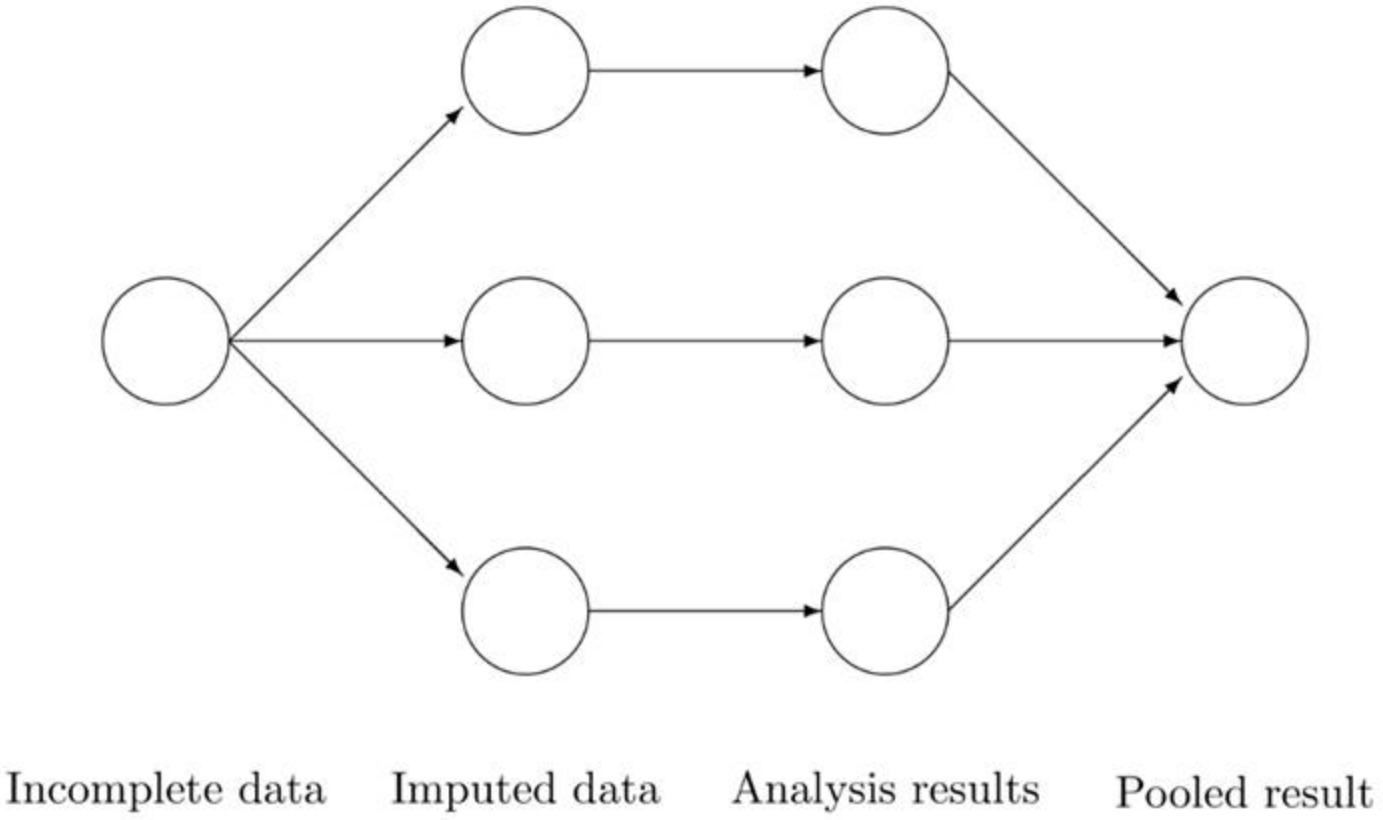
- **Multiple Imputation** – *account for within- & between-dataset variability; uncertainty in imputations*
  - Joint Modeling *(assumes joint multivariate normality)*
    - Multivariate Normal Imputation



  - Fully Conditional Specification / Two-Fold Fully Conditional Specification / Multiple Imputation by Chained Equations
    - Specifies model variable-by-variable, using a distribution conditional on all other variables

**not a comprehensive list*

MULTIPLE IMPUTATION



MULTIPLE IMPUTATION BY CHAINED EQUATIONS (MICE)

- Iterative predictive models, using observed data (all variables)
- Series of regressions: each variable with missing data is modeled conditional on other variables
  - i.e. DV[variable with missings] ~ IV[all other variables] ➔ predictions from regression model replace missings
  - imputed values from variable 1 are used as "predictors" for other variables' missing values *(L → R)*
  - Draws from posterior predictive distribution of missing data, given observed data and imputation model parameters

- Markov Chain Monte Carlo (MCMC) method
  - Gibbs sampler (Bayesian): if conditionals are compatible, sample from conditional distributions to obtain samples from joint distribution

MICE

Algorithm 4.3 (MICE algorithm for imputation of multivariate missing data.♠)

1. Specify an imputation model $P(Y_j^{\text{mis}}|Y_j^{\text{obs}}, Y_{-j}, R)$ for variable Y_j with $j = 1, \ldots, p$.
2. For each j, fill in starting imputations $\dot{Y}_j^0$ by random draws from Y_j^{obs}.
3. Repeat for $t = 1, \ldots, M$.
4. Repeat for $j = 1, \ldots, p$.
5. Define $\dot{Y}_{-j}^t = (\dot{Y}_1^t, \ldots, \dot{Y}_{j-1}^t, \dot{Y}_{j+1}^{t-1}, \ldots, \dot{Y}_p^{t-1})$ as the currently complete data except Y_j.
6. Draw $\dot{\phi}_j^t \sim P(\phi_j^t|Y_j^{\text{obs}}, \dot{Y}_{-j}^t, R)$.
7. Draw imputations $\dot{Y}_j^t \sim P(Y_j^{\text{mis}}|Y_j^{\text{obs}}, \dot{Y}_{-j}^t, R, \dot{\phi}_j^t)$.
8. End repeat j.
9. End repeat t.

https://stefvanbuuren.name/fimd/sec-FCS.html#sec:MICE

MICE

Variables to Include:

- ALL variables and interactions of interest (in any final model) should be included – *3way, 2way, etc.*
  - i.e., imputation model must be at least as general as (or more general than) analysis model
- Any variables that have predictive utility *(especially if complete)*
  - "auxiliary variables" – *increase in explained variance negligible after the best ~15 variables are included* (Buuren & Goothuis-Oudshoorn, 2021)
  - See how well variables are correlated using cor(data, use="pair")
  - quickpred()
    - Minpuc,mincor

MICE

- Longitudinal Data: how can we preserve hierarchical / longitudinal structure?
  - Wide format
- Multilevel Models: *it's complicated…*
  - *2l.lmer?*

MICE

- Multilevel/Longitudinal data & models (see Grund et al., 2018; Nevalainen et al., 2009; Zaninotto & Sacker, 2017)
  - Huque et al. (2018) suggest modeling longitudinal structure is only necessary sometimes, e.g. irregularly spaced data
- Potential options (may not be feasible with small clusters)
  - *for models with only random intercepts, dummy-code the clustering variable and include it as a predictor in imputation models*
  - *for models with clustered data AND random slopes, impute missing data within each cluster*

MICE

What about interaction terms (or squares, or other transformed variables) that have missing values?

MICE

What about interaction terms (or squares, or other transformed variables) that have missing values?

- Transform → Impute (*aka just another variable, J.A.V.)*
  - Calculate with incomplete data, then impute transformations
  - Bartlett et al., 2015; Von Hippel, 2009; Seaman et al., 2012
- Impute → Transform *(bias?)*
- Passive Imputation
  - Transformation is done within imputation algorithm
  - Substantive Model Compatible Fully Conditional Specification (smcfs) (see van Buuren, 2018)
- Standardized vs Raw

MICE

	id	Group	Age	Sex	BDI_1	BDI_3	BDI_7	BAI_1	BAI_3	BAI_7	FFMQ_1	FFMQ_3	FFMQ_7	gsi_1
1	3000	WL	27.0	1	14	9	15	8	6	6	102	106	111	0
2	3002	WL	32.0	2	13	14	16	10	11	4	88	84	94	0
3	3003	WL	61.3	1	7	5	4	6	3	0	113	108	110	0
4	3004	MBSR	33.4	1	4	5	4	1	4	2	133	122	134	0

CSxvisit_2	CSxvisit_3	CSxvisit_4	CSxvisit_5	CSxvisit_6	CSxvisit_7	Groupxvisit_1	Groupxvisit_2	Groupxvisit_3	Grou
1.3192436	1.9788654	2.6384872	3.2981090	3.9577308	4.6173526	0	0	0	
1.3261500	1.9892250	2.6523000	3.3153750	3.9784499	4.6415249	0	0	0	
3.2048625	4.8072937	6.4097249	8.0121561	9.6145874	11.2170186	0	0	0	
-3.4962098	-5.2443146	-6.9924195	-8.7405244	-10.4886293	-12.2367342	1	2	3	
-2.8058213	-4.2087319	-5.6116426	-7.0145532	-8.4174639	-9.8203745	1	2	3	

- nccam3 data: 87 predictors in iterative regressions

MICE

Problem: collinearity & overspecification

- No widely-accepted effective solution
  - Divide data into time blocks; impute independently? *(bias)*
  - Two-fold FCS? *(debated) – e.g. Welch et al., 2014; Zaninotto & Sacker, 2017*
- One option: use slopes
  - outcome ~ visit for each subject

MICE

- Predictor Matrix: id, group, age, sex, slopes, interactions
  - Default: any missing value is imputed
  - *User-specified: matrix of logicals indicating where imputations are needed*
    - *nccam3: identify subjects with ≤2 observations for any given variable*

```
where.na <- is.na(Immune_Psych_itt.wide.slope)
# replace any rows (subj) -other than the ones who
## dropped or only have 2- with NA to FALSE
## row names = 10,13,26,35,39,40,57,60,63,67

where.na[1:9,] <- FALSE #3010
where.na[11:12,] <- FALSE #3013
where.na[14:25,] <- FALSE #3026
where.na[27:34,] <- FALSE #3035
where.na[36:38,] <- FALSE #3039, #3041
where.na[41:56,] <- FALSE
where.na[58:59,] <- FALSE #3058
where.na[61:62,] <- FALSE #3060
where.na[64:66,] <- FALSE #3063
where.na[68:71,] <- FALSE #3067
```

MICE

- Impute using mice()
- !! SET SEED FOR REPRODUCIBILITY !!

```
# Dry run with no iterations in order to change prediction matrix
slope.mice0 <- mice(Immune_Psych_itt.wide.slope,maxit=0,
                    print=F,
                    seed=14243)
slope.mice0$loggedEvents

# Define predictor matrix
slope.pred <- slope.mice0$predictorMatrix

# Set all columns after slopes and interactions to 0,
## meaning that they (i.e., acq_1, acq_2, acq_3, etc.)
##will NOT be used as predictors for missing variables
slope.pred[,62:111] <- 0
```

MICE

Impute using mice()

- m = number of imputations: *depends – usually >5 (I used 10 per Dan Bolt)*
  - m = 3-10 (Rubin, 1987)? approx % cases missing? 100 * fmi (White et al., 2010)? other (von Hippel, 2018)?

$$M = 1 + \frac{1}{2}\left(\frac{FMI}{CV(SE)}\right)^2$$

- maxit = number of iterations – *increasing can help with convergence (I used 10)*
- where = missingness indicator matrix
  - Default is where = is.na(data) – any NA will be imputed
- pred = predictor matrix
  - Matrix of 0/1 of predictors to be used for each target column
    - *(see here for discussion of 2/-2/etc. for multilevel data, using 2l.lmer)*

```
slope.mice <- mice(Immune_Psych_itt.wide.slope,
                   m=10,maxit=10,
                   seed=14243,
                   where=where.na,
                   pred=slope.pred,
                   print=F)
```

```
# Complete: add imputed data to original dataframe in long (wide) format
slope.mice.complete <- complete(slope.mice,"long",inc=T)
```

MICE

Impute using mice()

- method = string (e.g. "PMM"; "" not to impute) to apply to all, or matrix of imputation method for each column
  - Predictive Mean Matching (PMM): *imputed values constrained to set of observed values*

MICE

- mids.object $
  - formulas = formulas used to impute variables
  - nmis = number of missing observations per variable
  - visitSequence = order in which columns are visited
  - method = imputation method for each block/vector
  - chainMean = means of imputations (no observed data included) per variable and iteration
  - chainVar= variances of imputed values per variable and iteration
  - seed = seed value of solution
  - loggedEvents = matrix of automatic removals
    - *Variables with missing values that are not imputed but used as predictors, constant, and collinear automatically removed*
    - remove.collinear = FALSE to override

```
acqScore_1 ~ id + acq.slope.est + spu.slope.est + feno.slope.est +
    casi.slope.est + blood.slope.est + bdi.slope.est + bai.slope.est +
    ffmq.slope.est + gsi.slope.est + psdi.slope.est + Group.bin +
    Age + Sex + Chronic.Stress_1 + CSxgroup_1 + BDIxgroupxvisit_1 +
    BDIxvisit_1 + BAIxgroupxvisit_1 + BAIxvisit_1
```

	1	2	3	4	5	6
id	NaN	NaN	NaN	NaN	NaN	NaN
acq.slope.est	-0.030000000	-5.357143e-02	0.00000000	-0.04750000	0.077500000	-0.12871429
spu.slope.est	-0.383342318	-5.862534e-02	0.08035714	0.36607143	-0.119642857	-0.49409938
feno.slope.est	-1.535714286	3.142857e-01	0.64285714	3.14285714	-0.391304348	-2.28571429
casi.slope.est	-0.257142857	2.857143e-01	0.07142857	-0.21428571	0.035714286	-0.17857143
blood.slope.est	5.997934596	4.116429e+01	-44.12571429	4.69412399	21.057142857	43.05714286
bdi.slope.est	1.830357143	1.214286e+00	-1.15178571	1.15178571	2.035714286	0.70535714
bai.slope.est	-0.535714286	-5.357143e-01	-0.57142857	-0.53571429	-1.017857143	-0.41964286
ffmq.slope.est	1.107142857	8.392857e-01	1.63690476	0.66071429	-0.744047619	3.08333333
gsi.slope.est	-0.007321429	2.678571e-04	0.01616071	-0.03205357	-0.004375000	0.01750000
psdi.slope.est	0.009196429	4.937500e-02	-0.04767857	0.02276786	-0.001785714	0.02125000
Group	NaN	NaN	NaN	NaN	NaN	NaN
Group.bin	NaN	NaN	NaN	NaN	NaN	NaN
Age	NaN	NaN	NaN	NaN	NaN	NaN
Sex	NaN	NaN	NaN	NaN	NaN	NaN
Chronic.Stress_1	-0.227669606	-6.292452e-01	0.65789521	-2.46279917	-1.137635784	0.08458584
CSxgroup_1	0.000000000	0.000000e+00	0.00000000	-1.40291064	0.000000000	0.00000000

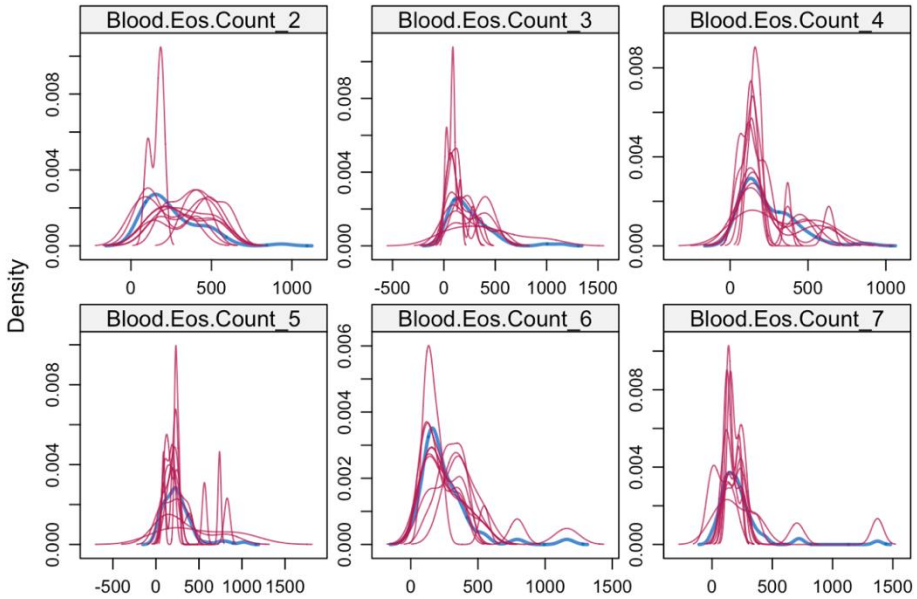
MICE

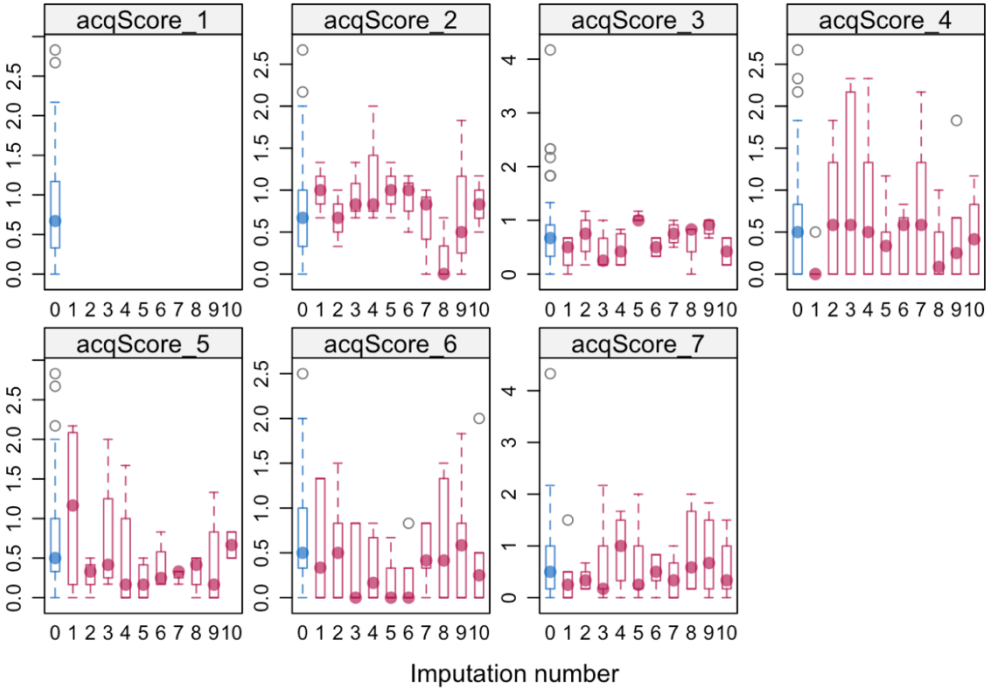
- loggedEvents
  - it = iteration number
  - im = imputation number
  - co = column number in data
  - dep = name of variable being imputed
  - meth = imputation method used
  - out = names of altered/removed preedictors

MICE

Diagnostics:

- Often useful to focus on distributional discrepancy (difference between observed and imputed)
  - Examine imputations using **densityplot(), bwplot()**, stripplot(), xyplot()

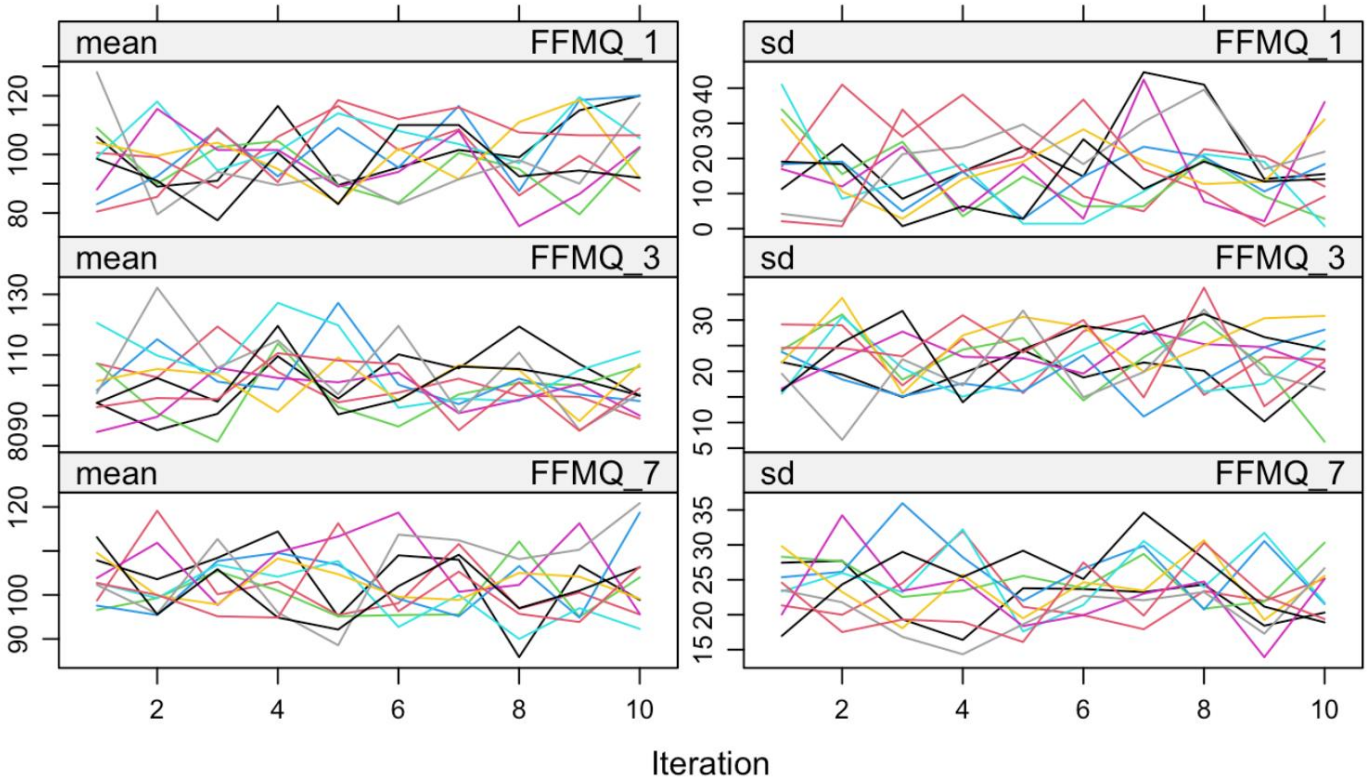




MICE

Diagnostics:

- Check convergence with plot()



MICE

Diagnostics to consider:

- *ANOVA: outcome = variable being imputed; factors = response stratum, indicator for observed/imputed status, and their interaction*
  - *Reject an imputation model is ANOVA is rejected in 2 of 5 imputed datasets (alpha 0.05) (Bondarenko & Raghunathan, ; Nguyan et al., 2017)*
- *Perform regression diagnostics (e.g. residuals vs fitted) for proposed regression imputation model (Marchenko & Eddings, 2011)*
- *Plot residuals (difference between observed OR imputed value and prediction from analysis model) against fitted values for each dataset (Nguyan et al., 2017)*
- *Posterior Predictive Checking (see Ngyuan et al., 2017 for example)*
- ***Something to consider:*** *how important is it that the imputations are realistic? How much weight are you putting on imputed estimates?*

MICE

- Analysis: use fully-specified final models, for each imputed dataset
  - with()
    - Output is a set of estimates/statistics for each imputed dataset (n=10)
    - Look at each model individually to assess convergence/singularity errors

```
[[6]]
Linear mixed model fit by REML ['lmerModLmerTest']
Formula: Feno ~ BAI + Group * visit + Age + Sex + (1 + visit + BAI ||      id)
REML criterion at convergence: 1856.918
Random effects:
 Groups   Name        Std.Dev.
 id       (Intercept) 2.303e+01
 id.1     visit       9.415e-01
 id.2     BAI         1.761e-04
 Residual             1.332e+01
Number of obs: 213, groups:  id, 71
Fixed Effects:
  (Intercept)           BAI        GroupWL          visit            Age            Sex
      68.5247        -0.4338        -8.6011        -1.8279        -0.2383        -8.7758
GroupWL:visit
       1.7643
optimizer (nloptwrap) convergence code: 0 (OK) ; 0 optimizer warnings; 1 lme4 warnings
```

MICE

- Pool estimates and statistics
  - pool()
  - combines sample (within-imputation) variance with variance caused by missing data (between-imputation variance)
    - Robin's rules (Robin, 1987)
  - pool() does not provide random effects variance estimates
    - only pool analyses without singularity/convergence errors

```
example.imputed.mods <- with(mice.MID.data,
                             lmer(dv~iv + (1 | id)))

example.imputed.mods

pool(example.imputed.mods$analyses[c(1,3,4,5,7,8,9)])
```

MICE

- Pool()
  - $pooled
    - ubar = mean of variances
    - t = total variance-covariance matrix
    - b = within-imputation variance
    - dfcom = df in complete-data analysis
    - df = residual df for hypothesis testing
    - riv = relative increase in variance due to nonresponse
    - lambda = proportion of total variance due to missingness
    - fmi = fraction of missing information
- summary(pooled.object)

```
Class: mipo    m = 10
                   term  m     estimate         ubar            b            t dfcom       df
1           (Intercept) 10  0.865697376 7.171868e-02 7.845479e-04 7.258168e-02   472 461.0606
2         Chronic.Stress 10 -0.026168464 2.360767e-03 2.133110e-05 2.384231e-03   472 463.0679
3                GroupWL 10 -0.109323004 1.779578e-02 4.132540e-04 1.825036e-02   472 444.2698
4                  visit 10 -0.045037338 1.911512e-04 1.197612e-05 2.043250e-04   472 365.4816
5                    Age 10 -0.007169221 1.686991e-05 5.024731e-07 1.742263e-05   472 433.0624
6                    Sex 10  0.152652804 1.231791e-02 7.415931e-05 1.239948e-02   472 465.8743
7 Chronic.Stress:GroupWL 10  0.211398491 4.800187e-03 2.194041e-05 4.824321e-03   472 467.0539
8          GroupWL:visit 10  0.047677595 3.985968e-04 3.033848e-05 4.319691e-04   472 336.8245
           riv      lambda         fmi
1 0.012033165 0.011890090 0.016148628
2 0.009939231 0.009841415 0.014090404
3 0.025544229 0.024907974 0.029268170
4 0.068917872 0.064474431 0.069552164
5 0.032763679 0.031724275 0.036165271
6 0.006622492 0.006578923 0.010816395
7 0.005027815 0.005002663 0.009236208
8 0.083724530 0.077256284 0.082686991
```

```
                    term     estimate   std.error  statistic       df     p.value
1            (Intercept)  0.865697376 0.269409872  3.2133098 461.0606 0.001404107
2          Chronic.Stress -0.026168464 0.048828588 -0.5359251 463.0679 0.592267667
3                 GroupWL -0.109323004 0.135093883 -0.8092373 444.2698 0.418811859
4                   visit -0.045037338 0.014294228 -3.1507359 365.4816 0.001762909
5                     Age -0.007169221 0.004174043 -1.7175726 433.0624 0.086589489
6                     Sex  0.152652804 0.111352963  1.3708913 465.8743 0.171068912
7 Chronic.Stress:GroupWL  0.211398491 0.069457332  3.0435734 467.0539 0.002469742
8           GroupWL:visit  0.047677595 0.020783866  2.2939715 336.8245 0.022407203
```

- pool()
  - $glanced
    - nobs, sigma, logLik, AIC, BIC, REMLcrit, df.rersidual

TRY IT!

- https://missingdata.shinyapps.io/mi2variables/
- /study/nccam3_rosenkranz/analyses/Estelle/ITT/ITT_MI_example.Rmd
- https://rmisstastic.netlify.app/tutorials/erler_course_multipleimputation_2018/erler_practical_mice_2018#getting_to_know_the_data

OTHER CONSIDERATIONS

- Conditional Imputation: *restrict imputations (within min/max and/or conditional on other data)*
  - See here https://stefvanbuuren.name/fimd/sec-knowledge.html#conditional-imputation
- Sensitivity analysis
  - Best / Worst case scenario
  - Particularly when data are MNAR

REFERENCES

- Buuren, S. van, & Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, *45*(3). https://doi.org/10.18637/jss.v045.i03
- Grund, S., Lüdtke, O., & Robitzsch, A. (2018). Multiple Imputation of Missing Data for Multilevel Models: Simulations and Recommendations. *Organizational Research Methods*, *21*(1), 111–149. https://doi.org/10.1177/1094428117703686
- Nevalainen, J., Kenward, M. G., & Virtanen, S. M. (2009). Missing values in longitudinal dietary data: A multiple imputation approach based on a fully conditional specification. *Statistics in Medicine*, *28*(29), 3657–3669. https://doi.org/10.1002/sim.3731
- Nguyen, C. D., Carlin, J. B., & Lee, K. J. (2017). Model checking in multiple imputation: An overview and case study. *Emerging Themes in Epidemiology*, *14*(1), 8. https://doi.org/10.1186/s12982-017-0062-6
- Rubin, D. B. (1976). Inference and Missing Data. *Biometrika*, *63*(3), 581–592. https://doi.org/10.2307/2335739
- Seaman, S. R., Bartlett, J. W., & White, I. R. (2012). Multiple imputation of missing covariates with non-linear effects and interactions: An evaluation of statistical methods. *BMC Medical Research Methodology*, *12*(1), 46. https://doi.org/10.1186/1471-2288-12-46
- van Buuren, S. (2018). *Flexible Imputation of Missing Data* (2nd ed.). Chapman & Hall/CRC. https://stefvanbuuren.name/fimd/
- von Hippel, P. T. (2009). 8. How to Impute Interactions, Squares, and other Transformed Variables. *Sociological Methodology*, *39*(1), 265–291. https://doi.org/10.1111/j.1467-9531.2009.01215.x
- Welch, C. A., Petersen, I., Bartlett, J. W., White, I. R., Marston, L., Morris, R. W., Nazareth, I., Walters, K., & Carpenter, J. (2014). Evaluation of two-fold fully conditional specification multiple imputation for longitudinal electronic health record data. *Statistics in Medicine*, *33*(21), 3725–3737. https://doi.org/10.1002/sim.6184
- White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*, *30*(4), 377–399. https://doi.org/10.1002/sim.4067
- Zaninotto, P., & Sacker, A. (2017). Missing Data in Longitudinal Surveys: A Comparison of Performance of Modern Techniques. *Journal of Modern Applied Statistical Methods*, *16*(2), 378–402. https://doi.org/10.22237/jmasm/1509495600