Handling Missing Data Under Difficult and Ordinary Circumstances

JEFF GILL
Distinguished Professor
Department of Government, Department of Mathematics & Statistics
Member, Center for Neuroscience and Behavior
Founding Director, Center for Data Science
American University

Recent Discussion

- In Jan 2012, the top UK medical journal, the *British Medical Journal* disclosed that "missing data is a serious problem in clinical research given that it distorts the scientific record and prevents clinical decisions from being based on the best evidence available.
- As part of an in-depth BMJ review on the subject, experts on `bmj.com` warn that patients can be harmed through missing clinical trial data, leading to unnecessary costs to health systems.
- U.S. Food and Drug Administration requested that the *National Research Council* convene an expert panel in 2008. The current report focuses primarily on phase 3 clinical trials assessing the safety and efficacy of drugs, biologic products and some medical devices.
- The panel found that a major cause of missing data is participants who stop taking their assigned treatment because it's not working, the side effects are troubling or the drug regimen is too inconvenient.
- But the panel suggested that researchers should continue to gather follow-up information on them anyway.

What Is the Problem?

- You download some data of interest:

```
library(foreign)
s.failure <- read.dta("Article.Lasso/Data.State.Failures/sort11v3.dta")
```

- It's big:

```
dim(s.failure)
[1] 8580 1231
```

- It has lots of missing data:

```
sum(is.na(s.failure))
[1] 7019387
sum(is.na(s.failure))/prod(dim(s.failure))
[1] 0.66459
```

- So how do you handle it? Not casewise deletion in this example:

```
sum(apply(apply(s.failure,1,is.na),2,sum) == 0)
[1] 0
```

- Many people deal with this problem incorrectly.

Big Data Relevance

- The problem is that as p gets larger the more rows that listwise deletion removes until there are none.
- Missing data is a phenomenon in all realistic social science datasets, as collected.
- Quote from Wang, J. Sophia, and Peter M. Aronow. "Listwise Deletion in High Dimensions" (*Political Analysis* 31.1, 2023, 149-155):

  > We show that when (i) all data have some idiosyncratic missingness and (ii) the number of variables grows superlogarithmically in n, then, for large n, listwise deletion will drop all rows wit

Background

- Goal: $L(\boldsymbol{\beta}|\mathbf{X}, \mathbf{Y})$ unbiased, efficient, and with the correct standard error.
- Define:

$$Z_{mis} = (X_{mis}, Y_{mis})$$

$$Z_{obs} = (X_{obs}, Y_{obs})$$

- We stipulate a $n \times k$ matrix, $\mathtt{R}$, corresponding to $\mathbf{X}$ that contains 0 when the $\mathbf{X}$ matrix data value is *not* missing, and 1 when it is missing.
- Stipulate a probability model for $\mathtt{R}$ where ϕ is a parameter in this distribution of $\mathtt{R}$.
- Standard Terms from Rubin (1979) for the missing data:

MCAR	$p(\mathtt{R}\|\mathbf{Z}_{obs}, \mathbf{Z}_{mis}) = p(\mathtt{R}\|\phi)$ missingness not related to observed or unobserved
MAR	$p(\mathtt{R}\|\mathbf{Z}_{obs}, \mathbf{Z}_{mis}) = p(\mathtt{R}\|\mathbf{Z}_{obs}, \phi)$ missingness depends only on observed data
Non-Ignorable (NMAR)	$p(\mathtt{R}\|\mathbf{Z}_{obs}, \mathbf{Z}_{mis}) = p(\mathtt{R}\|\mathbf{Z}_{obs}, \mathbf{Z}_{mis}, \phi)$ missingness depends on unobserved data

More On the Types of Missing Data

- If the joint parameter space of $\boldsymbol{\beta}$ and ϕ is orthogonal (*distinctiveness*) and MAR holds, then the missing data mechanism is called Ignorable.
- MCAR can be thought of as a special case of MAR where the observed data do not have to be conditioned on.
- There exist weak statistical procedures for determining MCAR versus MAR (c.f. Little 1988), but not against NMAR.
- Therefore you need to make a defensible assumption about which type you have before proceeding.
- There is no robust statistical procedure for handling non-ignorable missing data: you must add assumptions, restrict the data, or change the research question.
- Most of the tools here deal with MAR data.

Why Case-wise Deletion is Evil, Simple Statistic Version

- Consider the estimation of a true mean μ from a sample,$\mathbf{y}$, where some data are not missing completely at random.
- When μ_O is the mean of respondents and μ_M is the mean of missing data, we write the overall mean as:

$$\mu = \pi_R \mu_O + (1 - \pi_R)\mu_M.$$

  where π_R is the *proportion* of observed responses.
- The bias produced by casewise deletion is the fraction of missing data times the difference in means for observed and missing data (Little and Rubin, 2002, p.43):

$$\mu_O - \mu = (1 - \pi_R)(\mu_O - \mu_M).$$

- In the special case of MCAR missingness, $\mu_O = \mu_M$ and the statistic is unbiased:

$$\mu_O - \mu = (1 - \pi_R)(0) \quad \longrightarrow \quad \mu_O = \mu,$$

  but this is commonly violated in the social sciences.

Why Case-wise Deletion is Evil, Model Version

- Preliminaries: we are interested in obtaining the posterior mode of an unknown k-dimensional $\boldsymbol{\theta}$ coefficient vector, given an outcome variable vector $\mathbf{y}$, and an observed $\mathbf{X}$ matrix of explanatory data values assumed to be distributed iid according to $f(\mathbf{X}|\boldsymbol{\theta})$.
- Normally we would then:
  - as a Likelihoodist: find $\boldsymbol{\theta}$ that maximizes $\ell(\boldsymbol{\theta}|\mathbf{X})$.
  - as a Bayesian: produce a posterior distribution from $\pi(\boldsymbol{\theta}|\mathbf{X}) \propto p(\boldsymbol{\theta})p(\mathbf{X}|\boldsymbol{\theta})$.

  Neither of these approaches are directly possible if there happen to be missing data in $\mathbf{X}$ unless the data are considered "missing completely at random,".

Why Case-wise Deletion is Evil, Model Version

- First, segment the $\mathbf{X}$-matrix into two constituent parts: $\mathbf{X} = [\mathbf{X}_{obs}, \mathbf{X}_{mis}]$, and restate the distribution function:

$$f(\mathbf{X}|\boldsymbol{\theta}) = f(\mathbf{X}_{obs}, \mathbf{X}_{mis}|\boldsymbol{\theta}) = f(\mathbf{X}_{obs}|\boldsymbol{\theta}) f(\mathbf{X}_{mis}|\mathbf{X}_{obs}, \boldsymbol{\theta}).$$

- Now segment the log likelihood function into two distinct components:

$$\begin{aligned} \ell(\boldsymbol{\theta}|\mathbf{X}) &= \ell(\boldsymbol{\theta}|\mathbf{X}_{obs}, \mathbf{X}_{mis}) \\ &= \ell(\boldsymbol{\theta}|\mathbf{X}_{obs}) + \log f(\mathbf{X}_{mis}|\mathbf{X}_{obs}, \boldsymbol{\theta}) \end{aligned}$$

- Rearrange this form to create a statement with both unknowns collected on the right-hand side:

$$\ell(\boldsymbol{\theta}|\mathbf{X}_{obs}) = \ell(\boldsymbol{\theta}|\mathbf{X}_{obs}, \mathbf{X}_{mis}) - \log f(\mathbf{X}_{mis}|\mathbf{X}_{obs}, \boldsymbol{\theta}).$$

Why Case-wise Deletion is Evil, Model Version

- We can average over this uncertainty by taking expectations with respect to $\mathbf{X}_{mis}$ on both sides:

$$\begin{aligned}\int \ell(\boldsymbol{\theta}|\mathbf{X}_{obs}) f(\mathbf{X}_{mis}|\mathbf{X}_{obs},\boldsymbol{\theta}) d\mathbf{X}_{mis} \\ = \int \ell(\boldsymbol{\theta}|\mathbf{X}_{obs},\mathbf{X}_{mis}) f(\mathbf{X}_{mis}|\mathbf{X}_{obs},\boldsymbol{\theta}) d\mathbf{X}_{mis} \\ - \int \log f(\mathbf{X}_{mis}|\mathbf{X}_{obs},\boldsymbol{\theta}) f(\mathbf{X}_{mis}|\mathbf{X}_{obs},\boldsymbol{\theta}) d\mathbf{X}_{mis}.\end{aligned}$$

- LHS simplifies back to $\ell(\boldsymbol{\theta}|\mathbf{X}_{obs})$ because the integral ends up operating over just the isolated complete PDF for $\mathbf{X}_{mis}$:

$$\begin{aligned}\int \ell(\boldsymbol{\theta}|\mathbf{X}_{obs}) f(\mathbf{X}_{mis}|\mathbf{X}_{obs},\boldsymbol{\theta}) d\mathbf{X}_{mis} &= \ell(\boldsymbol{\theta}|\mathbf{X}_{obs}) \int f(\mathbf{X}_{mis}|\mathbf{X}_{obs},\boldsymbol{\theta}) d\mathbf{X}_{mis} \\ &= \ell(\boldsymbol{\theta}|\mathbf{X}_{obs}).\end{aligned}$$

Why Case-wise Deletion is Evil, Model Version

- Now we have an expression based only the observed data that relates the obtainable likelihood to two quantities that can be manipulated.

$$\ell(\boldsymbol{\theta}|\mathbf{X}_{obs}) = \int \ell(\boldsymbol{\theta}|\mathbf{X}_{obs},\mathbf{X}_{mis}) f(\mathbf{X}_{mis}|\mathbf{X}_{obs},\boldsymbol{\theta}) d\mathbf{X}_{mis} - \int \log f(\mathbf{X}_{mis}|\mathbf{X}_{obs},\boldsymbol{\theta}) f(\mathbf{X}_{mis}|\mathbf{X}_{obs},\boldsymbol{\theta}) d\mathbf{X}_{mis}$$

$$= \text{piece that cannot integrate out missingness} - \text{integrated effect of missingness}$$

- MCAR assumes that second term is zero and that $\mathbf{X}_{obs}$ and $\mathbf{X}_{mis}$ come from the same distribution and are therefore iid.
- MAR assumes that $\mathbf{X}_{mis}|\mathbf{X}_{obs}$ contains useful information.
- So the treatment of $\ell(\boldsymbol{\theta}|\mathbf{X}_{obs})$ is different, and casewise deletion under MAR is biased.
- Unfortunately case-wise deletion (also called "complete-case analysis" and "list-wise deletion") is quite common in many literatures.

Common Ways to Deal with Missingness

- Mean Imputation: fill in missing values with column means in the data matrix.
- Last Value Carried Forward: in research designs that apply some intervention/treatment, if the treatment outcome value is missing then substitute the pretreatment observed value.
- Using Information From Related Observations: old-style hot-decking (more on this later).
- Indicator Variables for Missingness of Categorical Predictors: if the missing data is strictly nominal, then define missingness as a new category.
- Imputation Based On Logical Rules: sometimes there is a deterministic relationship with observed data (eg. refuse income but report having no job).
- Inverse Probability Weighting: case-wise delete then use weights to rebalance the smaller dataset such that it is representative of the whole sample.
- Within Bayesian Stochastic Simulation: `JAGS` and `BUGS` treat missing data as unknown parameters to be estimated based on a specified prior.

Common Ways to Deal with Missingness: Random Imputation

- A *very* simple but somewhat limited approach is to impute missing values from observed chosen uniformly randomly with replacement.
- Assumes that $p(\mathtt{R}|\mathbf{Z}_{obs}, \mathbf{Z}_{mis}) = p(\mathtt{R}|\phi)$ (MCAR).
- Consider the function:

```
random.imp.vec <- function(V)  {
        gone <- is.na(V)
        there <- V[!gone]
        V[gone] <- sample(x=there,size=sum(gone),replace=TRUE)
        return(V)
}
X <- c(1,2,NA,4,5,NA)
random.imp.vec(X)
[1] 1 2 5 4 5 1
```

  which could be used on a matrix with the `apply` function.

- Sampling with replacement is important since it continues to favor values with higher incidence (preserving the MCAR empirical distribution).

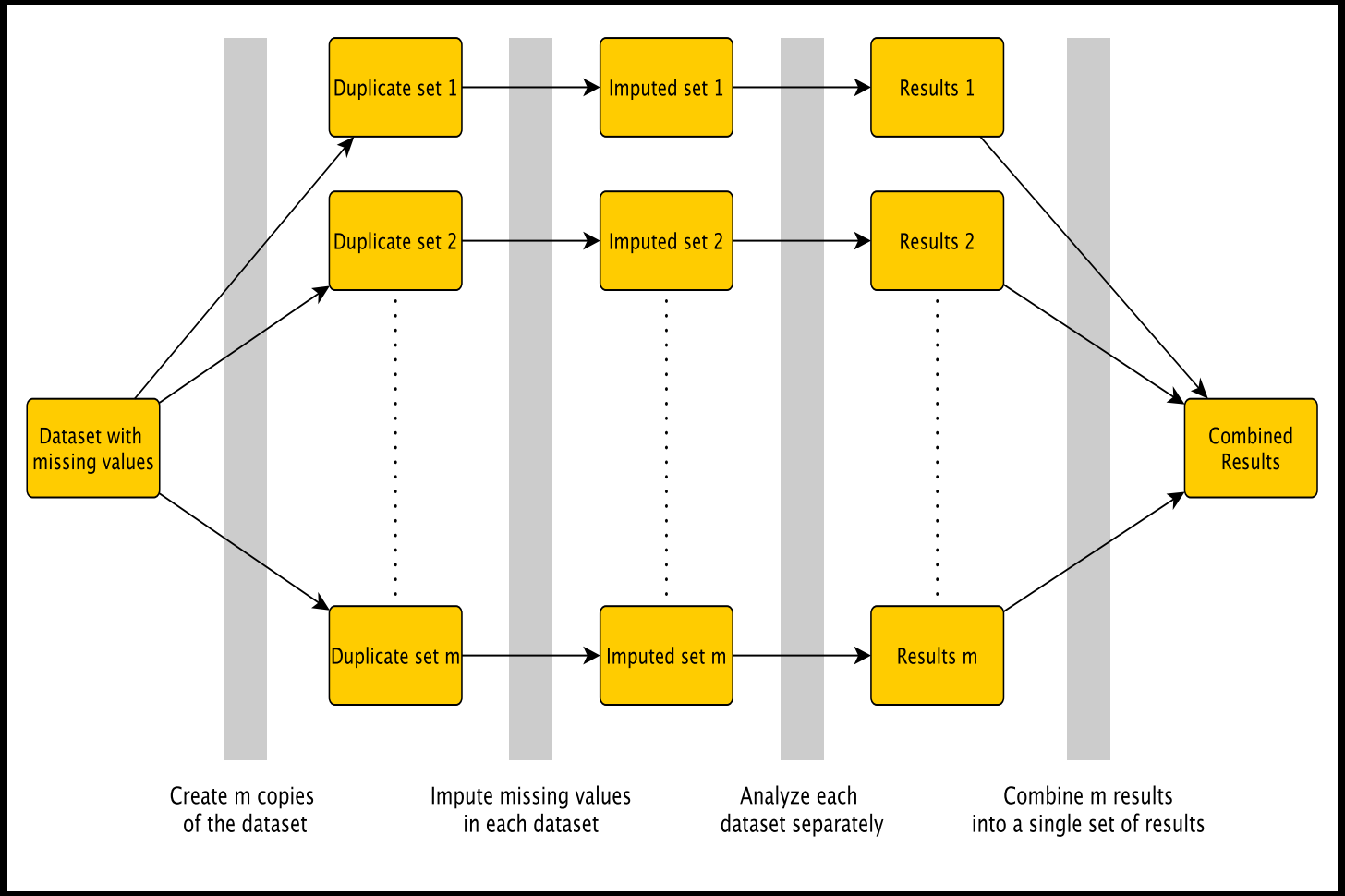
Common Ways to Deal with Missingness: Maximum Likelihood

- Similar to the Bayesian simulation approach, missing values are mathemetically removed from the likelihood function and estimated along with the coefficients.
- This makes estimation more complex, and most implementations use the EM Algorithm.
- Standard errors can be computed with the delta method, bootstrapping, or jackknifing.
- These are more common in structural equation model implementations using FIML (Mplus, Amos, LISREL, EQS, etc.).

Multiple Imputation (Rubin 1979)

- 3 Steps:
  - impute values for the missing data (Z_{mis}) M conditional on the observed values times to get M complete replicate datasets,
  - analyze/regress each dataset separately,
  - combine results with summary process.
- Imputation step assumes a conditional *posterior* distribution for the missing data conditioning on observed values.
- The more information in other variables the smaller the variance of this distribution.
- Oddly enough $M = 5, \ldots, 10$ is often sufficient.
- Combining process uses means for coefficients and an intuitive ANOVA approach for standard errors.

Multiple Imputation (Rubin 1979)



Multiple Imputation, Combining Results

- A single estimate of θ,

$$\bar{\theta}_M = \frac{1}{M}\sum_{m=1}^{M}\hat{\theta}_m,$$

  is created from the mean of $\hat{\theta}_m$ over the m values.

- The total variance is composed of the variation of the coefficient estimates *within* each imputed dataset and the variation of the coefficient estimates *between* the imputed datasets.

- The within imputation variance is the mean of individual coefficient variances across models:

$$W_M = \frac{1}{M}\sum_{m=1}^{M}\Sigma_m.$$

- The between imputation variance is the variance of the M coefficient estimates:

$$B_M = \frac{1}{M-1}\sum_{m=1}^{M}(\hat{\theta}_m - \bar{\theta}_M)^2.$$

Multiple Imputation, Combining Results

- The total variance for $\bar{\theta}_M$ is an ANOVA-style weighted sum:

$$T_M = W_M + \left(1 + \frac{1}{M}\right) B_M,$$

where $[1 + 1/M]$ is the adjustment for finite M, and the new degrees of freedom are:

- The degrees of freedom are calculated:

$$df_{MI} = (M-1)\left[1 + \frac{1}{M+1}\frac{W_M}{B_M}\right]^2.$$

- "Asymptotically" this become $M - 1$.

A Function To Put the Results Together

```
mice.output <- function(mean.mat,se.mat,N,var.names=NULL)  {
    m <- ncol(mean.mat)
    mean.vec <- apply(mean.mat,1,mean)
    between.var <- apply(mean.mat,1,var)
    within.var <- apply(se.mat^2,1,mean)
    se.vec <- sqrt(within.var + ((m+1)/m)*between.var)
    impute.df <- (m-1)*(1 + (1/(m+1)) * within.var/between.var)^2
    if (max(impute.df) > N) {   print("##### adjusting df")
        gamma <- (m/(m+1)) * between.var/(within.var + ((m+1)/m)*between.var)
        full.df <- N - 1
        adj.full.df <- ((full.df+1)/(full.df+3)) * full.df*(1-gamma)
        impute.df <- 1/(1/impute.df + 1/adj.full.df)
    }
    out.table <- round( cbind( mean.vec, se.vec, mean.vec/se.vec,
        1-pt(abs(mean.vec/se.vec),impute.df) ),5 )
    dimnames(out.table) <- list(var.names,
        c("Estimate","Std. Error","t value","Pr(>|t|)"))
    return(out.table)
}
```

Some Useful Quantities

- The *relative increase in variance due to nonresponse* (Rubin 1987) is:

$$r = \frac{m}{(m+1)}\frac{B}{W},$$

  where: B is the coefficient between variance and W is the average within imputation variance.

- The *fraction of information missing for the coefficient estimates* is

$$\lambda = \frac{r + \frac{2}{df+3}}{r+1}.$$

- *relative efficiency* of using m imputations rather an infinite number is:

$$re = \left(1 + \frac{\lambda}{m}\right)^{-1}.$$

Chronic Bronchitis and Dust Concentration Study

- The file contains data from a study of the Deutsche Forschungsgemeinschaft. The data were recorded during the years 1960 and 1977 in a Munich plant (1246 workers).
- Objective: dose response model for `cbr` with covariates `dust`, `expo` and `smoking`, and assessment of threshold limiting value under which dust has no influence on `cbr`.
- Description of the variables:

`cbr`	Chronic Bronchitis Reaction
	`1`: Yes
	`0`: No
`dust`	dust concentration at working place (in mg/m)
`smoking`	does worker smoke?
	`1`: Yes
	`0`: No
`expo`	duration of exposure in work-years

Chronic Bronchitis and Dust Concentration Study, Read Data

```
dust2.df <- read.table("https://pages.wustl.edu/files/pages/imce/jgill/dust2.txt",
    header=TRUE)
summary(dust2.df)
      cbr              dust            smoking           expo
 Min.   :0.0000   Min.   :  0.900   Min.   :0.0000   Min.   : 3.00
 1st Qu.:0.0000   1st Qu.:  1.945   1st Qu.:0.0000   1st Qu.:16.00
 Median :0.0000   Median :  5.065   Median :1.0000   Median :25.00
 Mean   :0.2343   Mean   :  4.822   Mean   :0.7392   Mean   :25.06
 3rd Qu.:0.0000   3rd Qu.:  6.260   3rd Qu.:1.0000   3rd Qu.:33.00
 Max.   :1.0000   Max.   : 24.000   Max.   :1.0000   Max.   :66.00
                  NA's   :578.000

sum(is.na(dust2.df))/prod(dim(dust2.df))
0.1159711
```

Chronic Bronchitis and Dust Concentration Study, Casewise Deletion

```
dust2.glm <- glm(cbr ~ dust+smoking+expo, family = binomial(link = logit),
        data=dust2.df,na.action=na.omit)
summary(dust2.glm)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.442991   0.382028  -9.012  < 2e-16
dust         0.143878   0.037892   3.797 0.000146
smoking      0.835298   0.235169   3.552 0.000382
expo         0.037629   0.008409   4.475 7.64e-06

    Null deviance: 762.07  on 667  degrees of freedom
Residual deviance: 709.61  on 664  degrees of freedom
  (578 observations deleted due to missingness)
AIC: 717.6
```

Chronic Bronchitis and Dust Concentration Study, Random Imputation

```
dust2.df$dust <- random.imp.vec(dust2.df$dust)
dust2.glm <- glm(cbr ~ dust+smoking+expo, family = binomial(link = logit),
        data=dust2.df,na.action=na.fail)
summary(dust2.glm)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.154958   0.275542 -11.450  < 2e-16
dust         0.073567   0.027229   2.702 0.006896
smoking      0.674027   0.173820   3.878 0.000105
expo         0.040802   0.006169   6.614 3.74e-11

    Null deviance: 1356.8  on 1245  degrees of freedom
Residual deviance: 1286.4  on 1242  degrees of freedom
AIC: 1294.4
```

Chronic Bronchitis and Dust Concentration Study, Full Data

```
dust.df <- read.table( "https://pages.wustl.edu/files/pages/imce/jgill/
    dust.asc__1.txt",header=TRUE)
dust.glm <- glm(cbr ~ dust+smoking+expo, family = binomial(link = logit),
        data=dust.df);
summary.glm(dust.glm)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.04785    0.24813 -12.283  < 2e-16
dust         0.09189    0.02323   3.956 7.63e-05
smoking      0.67683    0.17407   3.888 0.000101
expo         0.04016    0.00620   6.476 9.40e-11

    Null deviance: 1356.8  on 1245  degrees of freedom
Residual deviance: 1278.3  on 1242  degrees of freedom
AIC: 1286.3
```

Chronic Bronchitis and Dust Concentration Study, Full Data

```
# CASE-WISE DELETION
(Intercept) -3.442991   0.382028  -9.012  < 2e-16
dust         0.143878   0.037892   3.797 0.000146
smoking      0.835298   0.235169   3.552 0.000382
expo         0.037629   0.008409   4.475 7.64e-06

# RANDOM IMPUTATION
(Intercept) -3.154958   0.275542 -11.450  < 2e-16
dust         0.073567   0.027229   2.702 0.006896
smoking      0.674027   0.173820   3.878 0.000105
expo         0.040802   0.006169   6.614 3.74e-11

# FULL DATA
(Intercept) -3.04785    0.24813 -12.283  < 2e-16
dust         0.09189    0.02323   3.956 7.63e-05
smoking      0.67683    0.17407   3.888 0.000101
expo         0.04016    0.00620   6.476 9.40e-11
```

Chronic Bronchitis and Dust Concentration Study, Multiple Imputation

```
dust2.df <- read.table("https://pages.wustl.edu/files/pages/imce/jgill/dust2.txt",
    header=TRUE)
library(mice);  m <- 5;  imp.dust2 <- mice(dust2.df,m)
dust3.glm <- glm.mids(cbr ~ dust+smoking+expo, family = binomial(link = logit),
                      data=imp.dust2)
summary(pool(dust3.glm))
                   est          se         t          df     Pr(>|t|)       lo 95
(Intercept) -3.37686445 0.349777612 -9.654318  168.55060 0.000000e+00 -4.06737387
dust         0.12613589 0.051039923  2.471318   51.20283 1.682791e-02  0.02367882
smoking      0.64561295 0.180901959  3.568855 1014.44146 3.753238e-04  0.29062809
expo         0.04004765 0.006517554  6.144583  931.29865 1.186852e-09  0.02725686
                  hi 95 missing        fmi
(Intercept) -2.68635503      NA 0.36815695
dust         0.22859295     578 0.70018227
smoking      1.00059781       0 0.06132988
expo         0.05283845       0 0.07686594
```

Chronic Bronchitis and Dust Concentration Study, `dust` coefficient

- complete data: 0.09189
- case-wise deletion: 0.14387
- random imputation: 0.07356
- multiple imputation: 0.12613
- This means that there was more information within the dust column to help impute than there was across the rows using the other variables.

Reflections on the Multiple Imputation Process

- Try to have the largest data matrix possible, even including variables that will not go into the model.
- Although `mice` cannot handle very highly correlated variables as it does an $(\mathbf{X}'\mathbf{X})^{-1}$ operation
- If the imputation uses more "correct" information than the actual model, then it has "supereffi-ciency" (Rubin, 1996), called "uncongeniality" (Meng, 2001; Robins & Wang, 2000) and leads to conservative inferences.
- For incredibly large datasets, multiple imputation may need to proceed in separate groups of variables or cases.
- Remember that there is no statistical solution for nonignorable missing data, you have to make assumptions or impose restrictions that are defensible so that your data can be treated as MAR.

Example: Terrorism in Israel

- Provider: *The International Policy Institute for Counter-Terrorism*, Herzlia, Israel.
- Posted in an online database with details of attacks in Israel since September, 2000.
- Subsetted by Mark Harrison to give 103 suicide attacks over a three-year period from November 6, 2000 to November 3, 2003 when there was a steep drop.
- Information provided: date and place of the attack, attack type, the type of target and device employed, organizational affiliation of the attacker, and the number of casualties, along with a written description of the attack.
- Casualties are given personal attributes such as name, age, sex, nationality, and religion.

Terrorism Data

```
harr <- read.table("ARTICLES/Article.Dirichlet/Data.Israel/harrison4.txt",header=TRUE)
apply(harr[,-1],2,table)

$NumberKilled
 0  1  2  3  5  6  7  8  9 11 15 17 19 21 23 24 30
44 13  9  8  3  2  3  2  2  3  4  3  1  3  1  1  1

$NumberInjured
  0   1   2   3   4   5   6   8   9  11  13  14  16  17  20  21  22  26  27  30
 28   1   5   4   4   3   1   2   2   2   1   1   1   1   3   1   1   1   1   5
 40  42  47  50  52  57  58  59  60  65  69  86  90 100 102 120 130 150 188
  3   1   1   7   1   1   1   2   5   1   1   1   1   3   1   1   1   2   1

$TotalCasualties
  0   1   2   3   4   5   6   8   9  10  12  13  15  17  20  21  26  27  29  30
 22   5   6   4   3   3   2   2   1   1   2   2   1   1   2   1   1   2   1   1
 31  32  35  38  45  49  50  51  52  53  57  58  59  61  62  63  65  67  71  75
  1   1   1   1   1   2   1   1   1   1   2   1   1   2   1   1   2   2   3   2
 81  91  93 105 106 123 126 141 145 151 180 199
  1   1   1   1   1   1   1   1   1   1   1   1
```

Terrorism Data

```
$ResponsibleHamas
 0  1
59 44

$ResponsibleisMartyrs
 0  1
78 25

$ResponsibleisPIJ
 0  1
79 24

$ResponsibleisOther
 0  1
99  4

$TargetisMilitary
 0  1
76 10

$TargetisCivilian
 0  1
10 76

$TargetisBus
 0  1
89 14

$TargetisCafe
 0  1
89 14

$TargetisCheckpoint
 0  1
87 16

$TargetisResidence
  0   1
102   1
```

Terrorism Data

```
$TargetisOffshore
  0   1
101   2

$TargetisStore
  0  1
 96  7

$TargetisStreet
 0  1
71 32

$TargetisTravelstop
  0  1
 88 15

$DeviceisCar
 0  1
89 14

$DeviceisBoat
  0   1
101   2

$AttackisPrevented
  0   1
101   2

$AttackerisChallenged
 0  1
63 40

$FirstAttackerisMale
 0  1
 7 92

$FirstAttackerisFemale
 0  1
92  7
```

Terrorism Data

```
$AgeofFirstAttacker
16 17 18 19 20 21 22 23 24 25 26 27 29 31 43 45 48
 1  8  7 10 15 11 10 12  2  3  2  1  3  1  1  1  1
```

- Data Notes:
  - measurement is very nongranular,
  - some dichotomous variables very lopsided,
  - filtered through a government reporting source,
  - and the real data generating process is never observed: motivations, planning, and training.
- We will worry about the missing data here.

Terrorism Data Analysis

Attacker is Challenged

Device is Car

Target is Military

Hamas Responsible

Terrorism Data GLM With Casewise Deletion

- Read the data:

```
harr <- read.table("./Article.Dirichlet/Data.Israel/harrison3.txt",header=TRUE)
```

- Look at pattern of missingnes:

```
apply(apply(harr,2,is.na),2,sum)
                    0                     0                     0
      TotalCasualties      ResponsibleHamas  ResponsibleisMartyrs
                    0                     0                     0
     ResponsibleisPIJ    ResponsibleisOther      TargetisMilitary
                    0                     0                    17
     TargetisCivilian           TargetisBus          TargetisCafe
                   17                     0                     0
   TargetisCheckpoint     TargetisResidence      TargetisOffshore
                    0                     0                     0
       TargetisStore        TargetisStreet    TargetisTravelstop
                    0                     0                     0
          DeviceisCar          DeviceisBoat     AttackisPrevented
                    0                     0                     0
 AttackerisChallenged    AgeofFirstAttacker   FirstAttackerisMale
                    0                    14                     4
FirstAttackerisFemale
                    4
```

Run Two Models

- Run a standard LM with casewise deletion, note `na.action=na.omit` below:

```
harr.lm <- lm(NumberKilled ~ log(AgeofFirstAttacker) + log(as.numeric(Date)) +
                AttackerisChallenged + FirstAttackerisFemale + DeviceisCar +
                TargetisCafe + TargetisMilitary + ResponsibleHamas, data=harr,
                na.action=na.omit)
```

- Impute and run 10 models:

```
library(mice)
attach(harr)
harr2 <- cbind(NumberKilled, NumberInjured, AgeofFirstAttacker, Date,
              ResponsibleisMartyrs, AttackerisChallenged, FirstAttackerisFemale,
              ResponsibleisPIJ, TargetisBus, TargetisCheckpoint, DeviceisCar,
              TargetisCafe, TargetisMilitary, ResponsibleHamas)
detach(harr)
imp.harr <- mice(harr2,m=10)
harr.mids <- lm.mids(NumberKilled ~ log(AgeofFirstAttacker) +
                     log(as.numeric(Date)) + AttackerisChallenged +
                     FirstAttackerisFemale + DeviceisCar + TargetisCafe +
                     TargetisMilitary + ResponsibleHamas, data=imp.harr)
```

Terrorism Data GLM, Comparing Casewise Deletion (left) With Multiple Imputation (right)

```
cbind(summary(harr.glm)$coef[,1:2], summary(pool(harr.mids))[,1:2])
```

	Estimate	Std. Error	Estimate	Std. Error
(Intercept)	12.19545	12.79263	9.873602	9.49898
log(AgeofFirstAttacker)	-3.36921	4.16985	-2.061098	3.09964
log(as.numeric(Date))	0.58241	0.86409	0.067666	0.63617
AttackerisChallenged	-3.69990	1.60370	-3.131864	1.23538
FirstAttackerisFemale	3.20166	3.04025	2.815036	2.37085
DeviceisCar	-2.01778	2.45341	-0.177107	1.75987
TargetisCafe	3.96730	2.01286	4.893596	1.72607
TargetisMilitary	-5.44396	2.49457	-3.888428	1.48603
ResponsibleHamas	5.36309	1.64870	3.933507	1.26970

Fit a Generalized Additive model (tensor product smooth)

- There is no `mids` process for the GAM, so run `m=10` separate models according to:

```
harr.gam1 <- gam(NumberKilled ~ te(log(AgeofFirstAttacker),log(Date),k=3) +
                 AttackerisChallenged + FirstAttackerisFemale +
                 DeviceisCar + TargetisCafe + TargetisMilitary +
                 ResponsibleHamas, data=complete(imp.harr,1))
```

- These results can be stored separately or in a common list.
- The results are combined with the `mice.output` function plus recording of the smoothing output from each model.

Fit a Generalized Additive model (tensor product smooth)

- Combined results for the parameteric coefficients:

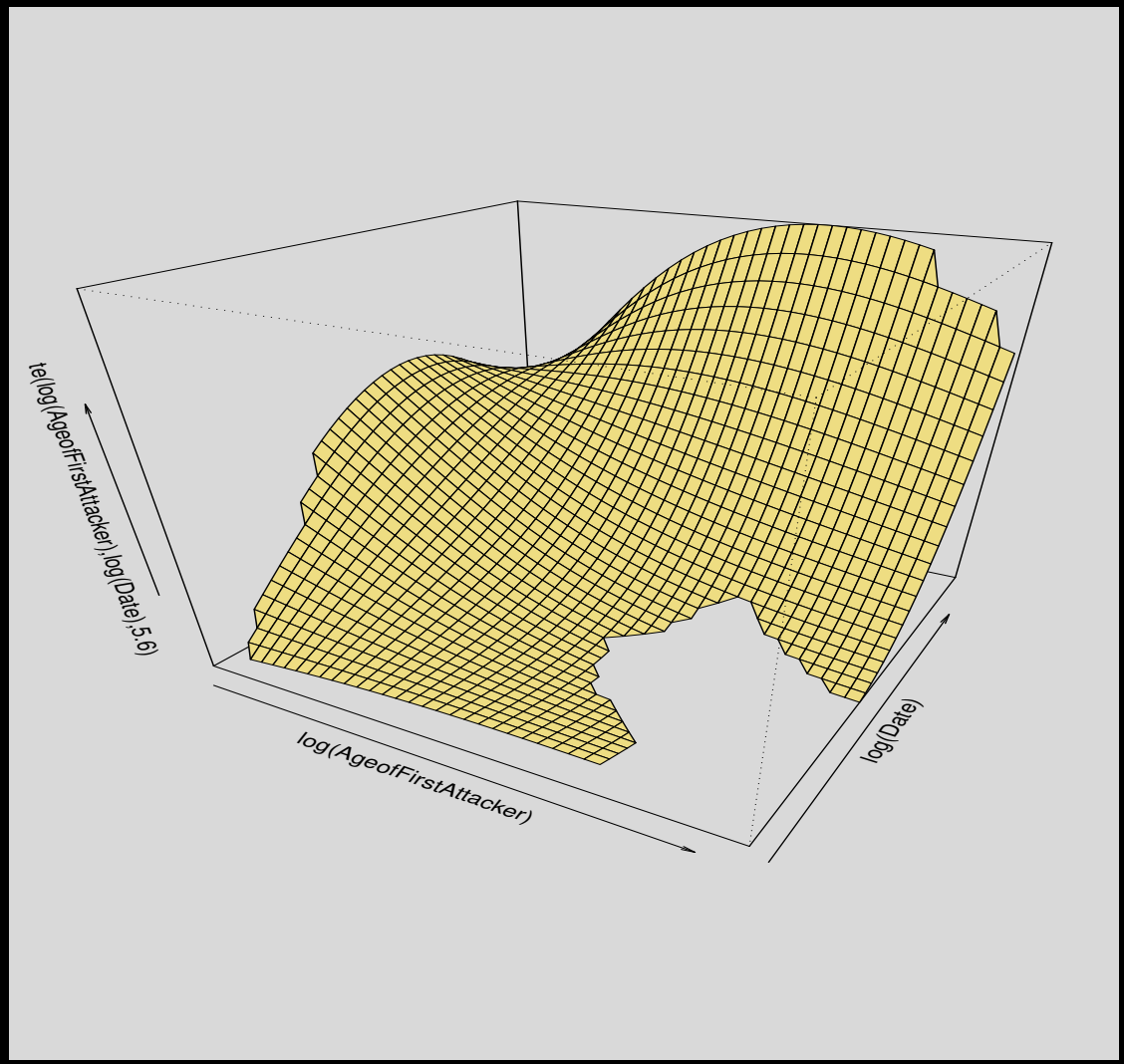
```
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)              4.377      1.070   4.091  9.3e-05
AttackerisChallenged    -4.059      1.144  -3.549 0.000616
FirstAttackerisFemale    1.286      2.255   0.571 0.569737
DeviceisCar              1.204      1.677   0.718 0.474763
TargetisCafe             3.824      1.638   2.335 0.021752
TargetisMilitary        -4.772      1.384  -3.448 0.000860
ResponsibleHamas         4.027      1.184   3.400 0.001003
```

- Now look a the summary for the smooth terms:

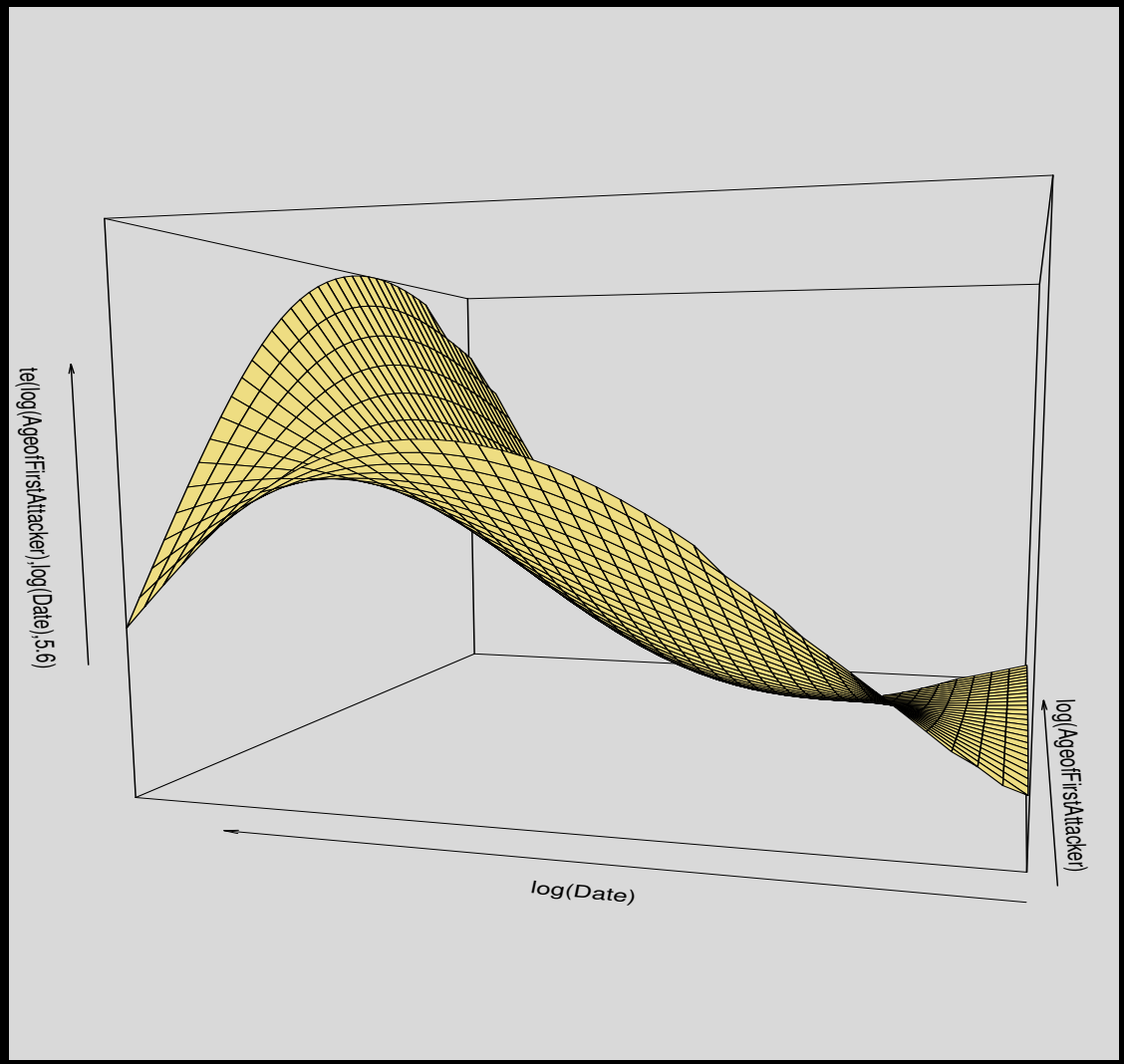
```
Approximate significance of smooth terms:
                                       edf Ref.df     F p-value
te(log(AgeofFirstAttacker),log(Date)) 5.613  5.613 3.794 0.00255
```

- We can also average the smoothed effects graphically...

Viewing the Nonparametric Component



Viewing the Nonparametric Component

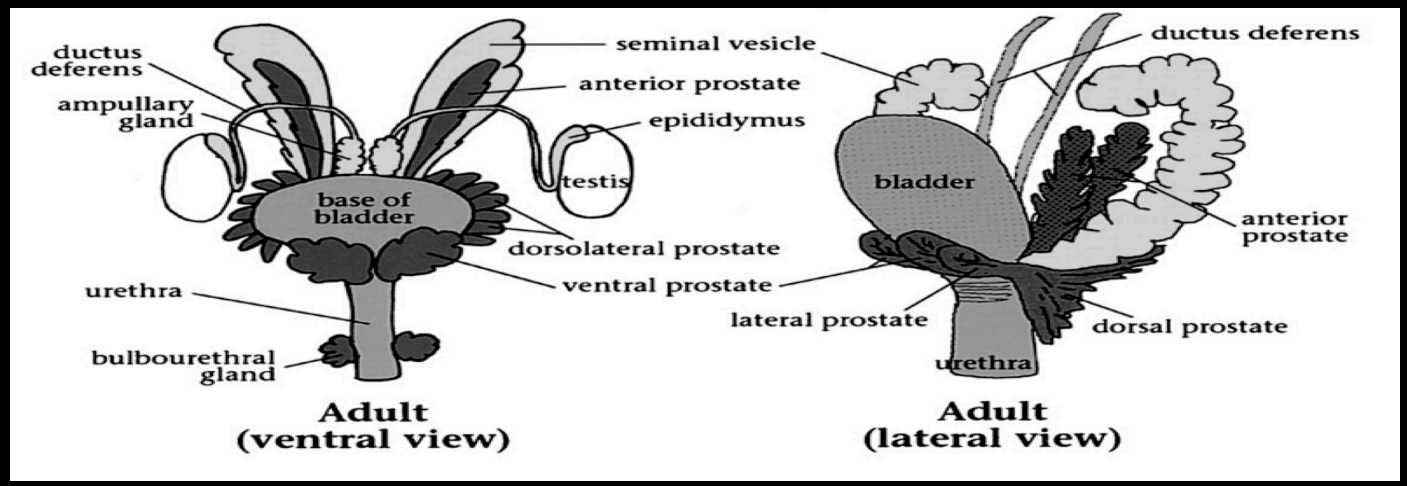


Applying `mice` To Actual Mice

- This research project looks at the impact of high-fat diet of the mother on prostate cancer outcomes in male pups.
- Treatment: the "dam" is fed a high fat diet ranging between 54 and 209 weeks.
- Control: regular mouse chow.
- The pups are sacrificed between 116 and 446 weeks after birth (includes 15 weeks of weaning).
- Emily C. Benesh, Jeff Gill, Laura E. Lamb, and Kelle H. Moley. "Maternal Obesity, Cage Density, and Age Contribute to Prostate Hyperplasia in Mice." *Reproductive Sciences*, Volume 23, Issue 4, 176-185, (February) 2016.

Applying `mice` To Actual Mice

- Since prostate cancer takes a long time to develop, the outcome is instead a count of hyperproliferative cells is obtained at pathology.
- Other variables: age when sacrificed, body weight, urogenital sinus weight, prostate weight, number of male pups in cage, cage number, diet of parents (regular vs. high-fat), age of parents at birth, time parents on diet before birth.
- Challenges: low-n, few covariates to choose from, high missingness, by conventional regression modeling standards.



Applying `mice` To Actual Mice

- Data and libraries:

```
lapply(c("lme4","mice"),library, character.only=TRUE)
mouse <- read.table("/Users/jgill/Grant.TREC/CompiledMouseData.csv",header=TRUE)
```

- Run multiple imputation:

```
m <- 10
mouse.imp <- mice(mouse,m)
mouse.array <- array(NA,c(dim(mouse),m))
for (i in 1:m) mouse.array[,,i] <- as.matrix(complete(mouse.imp,i))
```

Applying `mice` To Actual Mice, Multilevel Model

```
lmer.out.mean <- lmer.out.se <- NULL
for (i in 1:m)  {
    current.mouse.dat <- data.frame(mouse.array[,,i])
    names(current.mouse.dat) <- names(mouse)
    M1 <- lmer (Number_Hyperproliferative ~
        Age_when_used
        + Body_weight
        + UGS_weight
        + Prostate_Weight
        + Time_Parents_on_diet_before_birth
        + (1 | Diet_Treatment), family="poisson", data=current.mouse.dat)
    lmer.out.mean <- cbind(lmer.out.mean, summary(M1)$coef[,1])
    lmer.out.se   <- cbind(lmer.out.se,   summary(M1)$coef[,2])
}
```

Applying `mice` To Actual Mice, Multilevel Model

```
mice.output(lmer.out.mean,lmer.out.se,var.names=names(fixef(M1)))
```

```
                                Estimate Std. Error  t value Pr(>|t|)
(Intercept)                      6.24279    5.93493  1.05187  0.15513
Age_when_used                   -0.10440    0.04853 -2.15109  0.02534
Body_weight                      0.28757    0.11858  2.42509  0.01077
UGS_weight                      -2.61970    1.97876 -1.32391  0.09605
Prostate_Weight                 48.52982   48.98844  0.99064  0.17132
Time_Parents_on_diet_before_birth -0.00283    0.01081 -0.26198  0.39709

  AIC  BIC logLik deviance
 27.5 33.7  -6.73     13.5

Random effects:
Groups         Name        Variance   Std.Dev.
Diet_Treatment (Intercept) 1.7293e-18 1.315e-09
Number of obs: 18, groups: Diet_Treatment, 2
```

Old-Style Hot-Deck Imputation

- An early remedy developed by market researchers and census takers was called "hot decking" in literal reference to taking draws from a deck of computer punch-cards.
- If a respondent has a missing value, a "similar" respondent is found in the sample and that respondent's value for the variable with missingness is used for the respondent with a missing value.
- It is not always clear what is meant by "similar."
- Also the method does not reflect uncertainty in the imputed values: because only one value is imputed for each incidence of missing data, all imputations are treated as factual responses rather than best probabilistic imputations.
- Hot deck imputation methods range from simple random sampling to complicated algorithms in attempts to find respondents similar to those with missing responses.

Multiple Hot-Deck Imputation (Cranmer & Gill 2012)

- Focuses strictly on missing discrete values, keeping measurement discrete: for example a $\{-2, -1, 0, 1, 2\}$ variable will not get imputations like 1.2834.
- First create several replicate copies of the dataset, and then perform the steps:

1. Search down columns of the data sequentially looking for missing observations in one replicate.
   (a) When a missing value is found, compute a vector of affinity scores relative to all other rows (cases) for that case with the missing value.
   (b) Create an empirical distribution of potential donors using affinity scores and draw randomly from it to produce a vector of imputations.
   (c) Impute one of these values into the appropriate cell of each duplicate dataset.
2. Repeat Step 2 until no missing observations remain.
3. Estimate the statistic of interest for each dataset.
4. Combine the estimates of the statistic into a single estimate as in multiple imputation.

Multiple Hot-Deck Imputation, Affinity Score Definition

- For each respondent y_i indicates the outcome variable and $\mathbf{x}_i$ is a k-length vector of only discrete explanatory variables.
- If the ith case under consideration has q_i missing values in $\mathbf{x}_i$, then a potential donor vector, $\mathbf{x}_j, j \neq i$, will have between 0 and $k - q_i$ exact matches with i.
- Define $z_{i,j}$ as the number of variables for which the potential donor j and the recipient i have different values.
- Thus $k - q_i - z_{i,j}$ is the number of variables on which j and i are perfectly matched.
- This value, scaled by the highest number of possible matches $(k - q_i)$ is then the affinity score:

$$\alpha_{i,j} = \frac{k - q_i - z_{i,j}}{k - q_i}.$$

- The affinity score has the desirable properties that $\alpha_{i,j} = 1$ for $i \in \mathbf{D}_R$ (data with responses) and $\alpha_{i,j} = 0$ for $i \in \mathbf{D}_{NR}$ (data missing responses).
- Cases where the recipient and the donor are *both* missing values in the same covariate are deducted from k and q_i prior to the calculation of $\alpha_{i,j}$.

What About Missingness On the Outcome Variable?

- Controversy arises from performing the latter procedure since the goal of the regression model is to explain variance in **y** that is attributable to levels of **X** variables.
- This is an unsupported concern:
  - Different Model: the model that produces the conditional posterior for imputation draws is different than the model that will be specified for research purposes.
  - Different Data: the set of explanatory variables available for the imputation process on **y** is almost always different than the set of explanatory variables used in the final model specification (the ANES 2012 Direct Democracy Study has 1037 variables).
  - Helper Variables: when there is missingness in both **y** and **X**, the non-missing **y** values can contribute to the prediction of missing **X** values, as well as the reverse.
  - Bias: Graham (2009) and Raghunathian (2016) both note that leaving **y** completely out of the imputation procedure imposes a zero correlation between missing **y** and all of the other variables which biases coefficients in the resulting model (except under MCAR).

What About Missingness On the Outcome Variable?

- The easy (and unrealistic) case: missingness only in $\mathbf{y}$ and none in $\mathbf{X}$:
  - Under MAR $f(y|\mathbf{X}, R = 0)$ is the same as $f(y|\mathbf{X}, R = 1)$(Raghunathian 2016).
  - There is no regression information in the rows with missing y.
  - So imputing in this case only adds noise to the data used in the estimated regression model.
  - Casewise deletion with a MAR assumption is appropriate here.

Multiple Imputation Then Deletion (MID)

- Steps (von Hipple 2007):
  - run multiple imputation adding $\mathbf{y}$ on as an additional column of $\mathbf{X}$
  - listwise delete the rows that have imputed values for y
  - run M models
  - combine results as usual.
- The rows with missing y still help in the imputation process for $\mathbf{X}$, and MID *can* be more efficient than MI for small M and a large amount of missing data.
- Two relatively strict assumptions required for this approach:
  - missing $\mathbf{y}$ values are ignorable: unobserved $\mathbf{y}$ values are similar to observed $\mathbf{y}$ values from cases with similar values for $\mathbf{X}$
  - missing $\mathbf{X}$ values are ignorable in cases with missing $\mathbf{y}$.

Multiple Imputation Then Deletion (MID)

- The first assumption states that missing outcome variable values are not related (correlated) to each other and that there is an fixed relationship between the $\mathbf{X}$ matrix rows and the corresponding $\mathbf{y}$ values.
- If that sounds like an assumption of a standard regression model, it is because that *is* an assumption of a standard regression model: $y_i = g^{-1}(\mathbf{X}_i\boldsymbol{\beta}) + e_i$ with iid data.
- So one criticism of the von Hippel procedure is that it conflates the imputation procedure with the modeling procedure in a way that standard multiple imputation does not.
- The second assumption is less distasteful: all of the missing $\mathbf{X}_i$ values for a case with missing $\mathbf{y}_i$ must be ignorable, so a missing y_i imposes restrictions only on the rest of that case.

Expectation-Maximization

- Little and Rubin (2002) give a solution for linear models using the EM algorithm (*Expectation-Maximization*, Dempster, Laird and Rubin 1977) *when all of the missingness is confined to the outcome variable.*

- Using starting values, $\boldsymbol{\beta}_0$ and σ_0^2, for $j = 1$ until convergence do the following iterations:

  - **E-Step**: $E[\mathbf{y}_i|\mathbf{X}, \mathbf{y}_{\text{obs}}, \boldsymbol{\beta}_j, \sigma_j^2] = \begin{cases} y_i & \text{if } y_i \text{ observed} \\ \mathbf{X}\boldsymbol{\beta}^{(j)} & \text{if } y_i \text{ not observed} \end{cases}$

    Produces a complete outcome vector: $\mathbf{y}^{(j)}$.

  - **M-Step**: $\boldsymbol{\beta}^{(j+1)} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}\mathbf{y}^{(j)}$.
    Produces a new coefficient estimate: $\boldsymbol{\beta}^{(j+1)}$.

- When there is also missingness in $\mathbf{X}$ they make a multivariate normal assumption for $[\mathbf{X}, \mathbf{y}]$ and use a slightly more involved version of the EM algorithm with a maximum likelihood M-Step.

Application to a Political Science Problem

- We evaluate the performance of the CC, MI, MID, and MIP approaches via a replication of Hetherington and Husser (2012).
- They develop a theory concerning the changing nature of political trust during the post-911 era, and its effects on U.S. citizens' foreign policy preferences.
- Theory: political trust affects individuals' foreign policy preferences.
- The authors use data from the 2004 National Election Study (NES) cross-section to empirically evaluate the effects of contemporaneous political trust on a variety of outcomes related to U.S. foreign policy support, including individuals' (i) support for the Iraq War, (ii) support for the war in Afghanistan, and (iii) U.S. defense spending preferences.
- Hetherington and Husser specify a linear model with OLS and listwise delete in an effort to assess the effects of political trust on individual preferences towards U.S. defense spending while controlling for individuals' demographic, psychological, and political characteristics.

Replication of Hetherington & Husser, Estimates and 95% CI

Variable	CC	MI	MID	MIP
Political Trust	0.338	0.314	0.322	0.310
	[0.034↔0.641]	[-0.083↔0.712]	[-0.051↔0.695]	[-0.120↔0.741]
Party Id	0.854	0.681	0.683	0.733
	[0.551↔1.158]	[0.351↔1.010]	[0.400↔0.967]	[0.352↔1.115]
Conservatism	0.053	0.267	0.313	0.141
	[-0.426↔0.533]	[-0.288↔0.822]	[-0.163↔0.789]	[-0.581↔0.863]
Authoritarianism	0.029	0.015	-0.009	-0.013
	[-0.235↔0.293]	[-0.247↔0.278]	[-1.869↔1.850]	[-1.999↔1.973]
Education	-0.920	-0.879	-0.915	-0.971
	[-1.239↔-0.601]	[-1.168↔-0.590]	[-4.385↔2.555]	[-4.458↔2.516]
South	0.122	0.118	0.135	0.142
	[-0.028↔0.273]	[-0.041↔0.277]	[-1.321↔1.591]	[-1.419↔1.703]
Age	0.270	0.132	0.105	0.163
	[-0.105↔0.645]	[-0.203↔0.468]	[-0.507↔0.716]	[-0.396↔0.722]
Female	-0.286	-0.272	0.105	0.163
	[-0.422↔-0.149]	[-0.397↔-0.146]	[-3.018↔2.512]	[-3.040↔2.528]
Black	-0.056	-0.021	-0.033	-0.095
	[-0.277↔0.165]	[-0.216↔0.174]	[-0.968↔0.902]	[-1.168↔0.978]
Patriotism	2.085	2.279	2.306	2.312
	[1.629↔2.54]	[1.802↔2.756]	[-2.228↔6.840]	[-1.948↔6.571]
Moral Tradit.	1.109	1.141	1.147	1.168
	[0.629↔1.589]	[0.575↔1.706]	[-1.047↔3.341]	[-1.100↔3.437]
Import. Def. Spend.	1.494	1.265	1.244	1.346
	[1.113↔1.875]	[0.918↔1.611]	[-0.234↔2.723]	[-0.269↔2.961]
Isolationism	-0.259	-0.254	-0.294	-0.313
	[-0.445↔-0.073]	[-0.442↔-0.065]	[-3.107↔2.519]	[-3.213↔2.588]
Constant	0.890	0.879	0.884	0.877
	[0.369↔1.412]	[0.344↔1.415]	[0.374↔1.394]	[0.214↔1.54]
N	872	1212	1061	1212

Note: All estimates from OLS, and 95% CIs are one-tailed (in brackets).