Missing Data

Joseph Kush

11/6/2022

Missing data

We know that missing data has the potential to be problematic when estimating models, depending on the missing data mechanism. The default for most statistical software is to use listwise deletion when handling missing data. Alternatives such as multiple imputation may provide better, more robust estimates in the presence of missing data. How do our estimates differ across the different techniques?

High school and beyond

Work through an applied example using empirical data from the High School and Beyond (HSB) study.

Let's read in and view our data.

```
d <- read.csv('https://stats.idre.ucla.edu/wp-content/uploads/2016/02/hsbdemo.dat', header = F)
```

Let's also load any packages we need.

```
library("mice")
library("naniar")
```

Now, let's view our data

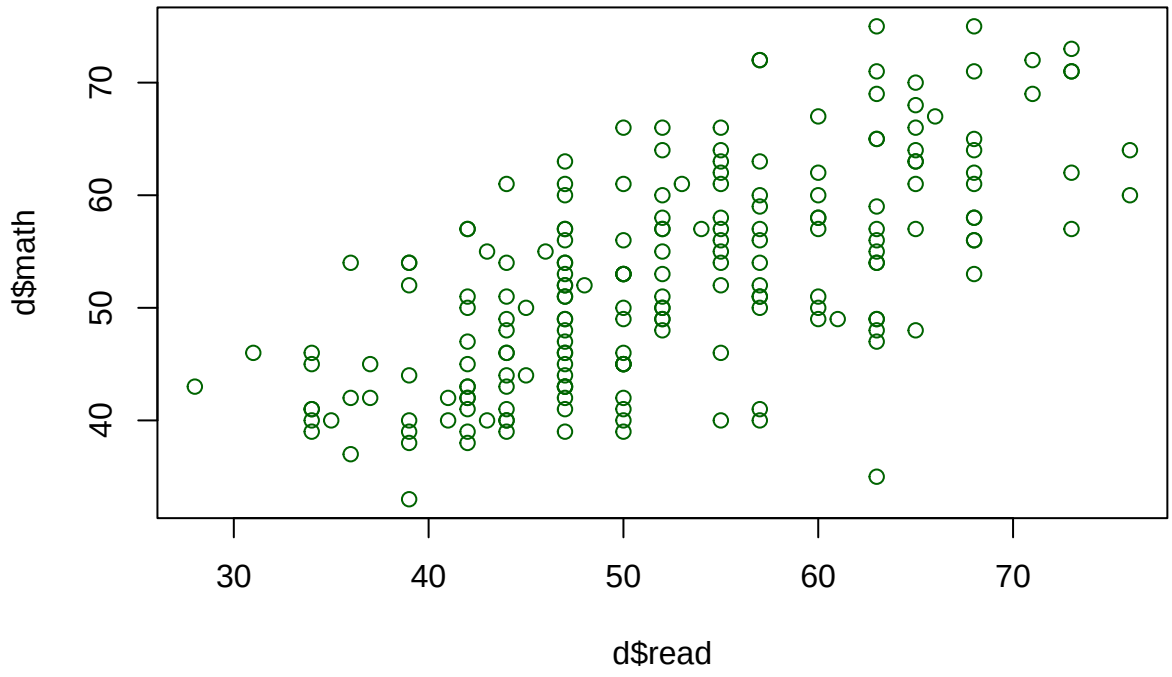
```
head(d)
```

```
##   read write math
## 1   34    35   41
## 2   34    33   41
## 3   39    39   44
## 4   37    37   42
## 5   39    31   40
## 6   42    36   42
```

```
summary(d)
```

```
##       read           write           math
##  Min.   :28.00   Min.   :31.00   Min.   :33.00
##  1st Qu.:44.00   1st Qu.:45.75   1st Qu.:45.00
##  Median :50.00   Median :54.00   Median :52.00
##  Mean   :52.23   Mean   :52.77   Mean   :52.65
##  3rd Qu.:60.00   3rd Qu.:60.00   3rd Qu.:59.00
##  Max.   :76.00   Max.   :67.00   Max.   :75.00
```

```r
plot(d$read, d$math, col = "darkgreen")
```



Looks like there is a positive, linear relationship between reading scores and math scores.

Using all available observed data (n = 200), let's estimate a linear regression model, in which our dependent/outcome variable 'math' is modeled as a function of two independent/predictor variables, 'read' and 'write'.

```r
full_model <- lm(math ~ 0 + read + write, data = d)
summary(full_model)
```

```
##
## Call:
## lm(formula = math ~ 0 + read + write, data = d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -20.963  -4.239   1.057   5.134  20.500
##
## Coefficients:
##       Estimate Std. Error t value Pr(>|t|)
## read   0.51058    0.05519   9.252  < 2e-16 ***
## write  0.48565    0.05478   8.865 4.45e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 6.875 on 198 degrees of freedom
## Multiple R-squared:  0.9836, Adjusted R-squared:  0.9835
## F-statistic:  5949 on 2 and 198 DF,  p-value: < 2.2e-16
```

Our estimates show that reading and writing scores are both statistically significant predictors of math scores. For a 1-unit increase in reading scores, math scores increase by 0.51 points, on average. For a 1-unit increase in writing scores, math scores increase by 0.49 points, on average. Unfortunately, we do not know the true linear relationship between reading scores and math scores for all individuals in the population; however, based on our sample of n = 200, this is our best guess.

Missing Completely at Random (MCAR)

What happens if we purposefully delete values for some observations? Remember that for missing data to be MCAR, the reason for the missing value is not dependent on any other variable.

Let's first make a binary indicator for whether or not an individual's math score should be missing, with a probability set to a specific level (and dependent on nothing else).

```
set.seed(681)
d$na_math <- rbinom(n = nrow(d), size = 1, prob = 0.7)
head(d)
```

```
##   read write math na_math
## 1   34    35   41       1
## 2   34    33   41       1
## 3   39    39   44       1
## 4   37    37   42       1
## 5   39    31   40       1
## 6   42    36   42       0
```

Before we delete values for some observations, let's conduct a two-sample t-test, comparing the means across our two groups.

```
# two-sample t-test ?
t.test(math ~ na_math, data = d)
```

```
##
##  Welch Two Sample t-test
##
## data:  math by na_math
## t = -0.012128, df = 126.61, p-value = 0.9903
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -2.736063  2.702730
## sample estimates:
## mean in group 0 mean in group 1
##        52.63333        52.65000
```

Here, we see our two groups are virtually indistinguishable from one another (m1 = 52.63, m2 = 52.65).

We see that some observations have a missing indicator equal to one, while others do not. Based on the missing value indicator, we can overwrite the math score as missing.

```r
d$math <- ifelse(d$na_math == 1, NA, d$math)
head(d)
```

```
##   read write math na_math
## 1   34    35   NA       1
## 2   34    33   NA       1
## 3   39    39   NA       1
## 4   37    37   NA       1
## 5   39    31   NA       1
## 6   42    36   42       0
```

What is the proportion of missingness?

```r
prop.table(table(d$na_math))["1"]
```

```
##   1
## 0.7
```

Is our data MCAR? Let's use Little's (1988) MCAR test to see if our data meets the assumptions of being MCAR.

```r
# Little's (1988) MCAR test
naniar::mcar_test(d[,c("math", "read", "write")])
```

```
## # A tibble: 1 x 4
##   statistic    df p.value missing.patterns
##       <dbl> <dbl>   <dbl>            <int>
## 1     0.629     2   0.730                2
```

We could interpret this as follows: "Patterns of missingness were further probed. Little's (1988) multivariate test of Missing Completely at Random (MCAR) indicated the data did meet the assumptions of the MCAR missing data mechanism ($\chi^2_{(2)} = 0.629$, $p = .730$). Thus, listwise deletion was used to drop observations with missing math scores when estimating linear regression models, as listwise deletion has been shown to produce unbiased estimates when missingness is MCAR".

Now let's estimate a linear regression using listwise deletion when 70% of our data is MCAR.

```r
missing_model <- lm(math ~ 0 + read + write, data = d)
summary(missing_model) # observations deleted due to missingness
```

```
##
## Call:
## lm(formula = math ~ 0 + read + write, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.9180  -3.9252   0.5042   4.6714  15.3231
##
## Coefficients:
##       Estimate Std. Error t value Pr(>|t|)
## read   0.55191    0.09430   5.853 2.38e-07 ***
```

```
## write  0.45114    0.09505   4.747 1.40e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.381 on 58 degrees of freedom
##   (140 observations deleted due to missingness)
## Multiple R-squared:  0.9862, Adjusted R-squared:  0.9857
## F-statistic:  2065 on 2 and 58 DF,  p-value: < 2.2e-16
```

We see our new estimates indicate that reading and writing scores are both significant predictors. How do our estimates from the missing data model using listwise deletion compare to those from the full model?

```
coef(full_model)
```

```
##      read     write
## 0.5105818 0.4856488
```

```
coef(missing_model)
```

```
##      read     write
## 0.5519102 0.4511375
```

Not off by too much, even with 70% missing. What if 20% were missing? Let's re-read in our data (so that everything is observed).

```
d <- read.csv('https://stats.idre.ucla.edu/wp-content/uploads/2016/02/hsbdemo.dat', header = F)
```

Let's again make a binary indicator for whether or not an individual's math score should be missing, where the probability is not dependent on anything).

```
set.seed(681)
d$na_math <- rbinom(n = nrow(d), size = 1, prob = 0.2)
head(d)
```

```
##   read write math na_math
## 1   34    35   41       0
## 2   34    33   41       0
## 3   39    39   44       0
## 4   37    37   42       0
## 5   39    31   40       0
## 6   42    36   42       1
```

```
d$math <- ifelse(d$na_math == 1, NA, d$math)
head(d)
```

```
##   read write math na_math
## 1   34    35   41       0
## 2   34    33   41       0
## 3   39    39   44       0
## 4   37    37   42       0
## 5   39    31   40       0
## 6   42    36   NA       1
```

What is the proportion of missingness?

```
prop.table(table(d$na_math))["1"]
```

```
##    1
## 0.21
```

Let's estimate a linear regression using listwise deletion when 20% of our data is MCAR.

```
missing_model <- lm(math ~ 0 + read + write, data = d)
```

How do our 20% missing estimates compare to the full model?

```
coef(full_model)
```

```
##      read     write
## 0.5105818 0.4856488
```

```
coef(missing_model)
```

```
##      read     write
## 0.5256626 0.4697000
```

Not off by much at all! This is great news, indicating that if data is truly MCAR, listwise deletion doesn't harm parameter estimates too much.

Missing at Random (MAR)

Remember that for missing data to be MAR, the reason the value is missing depends on other variables that are observed.

To begin, let's start with our full data.

```
d <- read.csv('https://stats.idre.ucla.edu/wp-content/uploads/2016/02/hsbdemo.dat', header = F)
```

We will again make a binary indicator for whether or not an individual's math score should be missing. However, this time, the probability is going to depend on that individual's reading score.

```
set.seed(681)

d$na_math <- rbinom(n = nrow(d), size = 1,
                    prob = ifelse(d$read > quantile(d$read, probs = .7),
                                  1, 0))
```

Before we delete values for some observations, let's again conduct a two-sample t-test.

```
# two-sample t-test ?
t.test(math ~ na_math, data = d)
```

```
##
##  Welch Two Sample t-test
##
## data:  math by na_math
## t = -8.1167, df = 96.183, p-value = 1.591e-12
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -13.145554  -7.979446
## sample estimates:
## mean in group 0 mean in group 1
##         49.6875         60.2500
```

Now we see our two groups are very different from one another (m1 = 49.7, m2 = 60.3).

```
# overwrite the income value as missing if the missing indicator == 1
d$math <- ifelse(d$na_math == 1, NA, d$math)

# view our data
d[100:110,]
```

```
##     read write math na_math
## 100   57    41   57       0
## 101   44    52   51       0
## 102   57    52   40       0
## 103   52    55   50       0
## 104   42    41   57       0
## 105   63    49   NA       1
## 106   61    59   NA       1
## 107   60    54   NA       1
## 108   52    54   55       0
## 109   52    59   48       0
## 110   57    55   52       0
```

```
# proportion of missing?
prop.table(table(d$na_math))["1"]
```

```
##    1
## 0.28
```

Is our data MCAR?

```
# Little's (1988) MCAR test
naniar::mcar_test(d[,c("math", "read", "write")])
```

```
## # A tibble: 1 x 4
##   statistic    df p.value missing.patterns
##       <dbl> <dbl>   <dbl>            <int>
## 1      131.     2       0                2
```

Unfortunately, no ($p < 0.001$).

Let's estimate a linear regression using listwise deletion when 30% of our data is MAR.

```
missing_model <- lm(math ~ 0 + read + write, data = d)
summary(missing_model) # observations deleted due to missingness
```

```
##
## Call:
## lm(formula = math ~ 0 + read + write, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.5701  -4.7156   0.8655   4.6873  18.0961
##
## Coefficients:
##       Estimate Std. Error t value Pr(>|t|)
## read    0.6459     0.0677   9.541  < 2e-16 ***
## write   0.3799     0.0631   6.021 1.41e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.498 on 142 degrees of freedom
##   (56 observations deleted due to missingness)
## Multiple R-squared:  0.9836, Adjusted R-squared:  0.9833
## F-statistic:  4247 on 2 and 142 DF,  p-value: < 2.2e-16
```

We see our new estimates indicate that reading and writing scores are both significant predictors. How do our estimates from the missing data model using listwise deletion compare to those from the full model?

```
coef(full_model)
```

```
##      read     write
## 0.5105818 0.4856488
```

```
coef(missing_model)
```

```
##      read     write
## 0.6458689 0.3799148
```

When our data is MAR, even with only 30%, our estimates using listwise deletion are now very different from those we got with the full sample.

What if we try to use multiple imputation to help us estimate our model? We will use the "mice" package in R, with $m = 50$ imputed datasets estimated using the predictive mean matching.

```
imps <- mice(d, m = 50, method = "pmm")
```

```
## Warning: Number of logged events: 250
```

```
res <- with(imps, lm(math ~ 0 + read + write))
```

Before we look at our multiple imputation estimates, let's remember what our estimates were for the full model and the missing model using listwise deletion.

```
coef(full_model)
```

```
##      read     write
## 0.5105818 0.4856488
```

```
coef(missing_model)
```

```
##      read     write
## 0.6458689 0.3799148
```

And now, our multiple imputation estimates

```
summary(pool(res))[,2]
```

```
## [1] 0.5192563 0.4762216
```

Wow, our multiple imputation estimates were nearly identical to those obtained from the full model. Thus, when missing data are MAR, listwise deletion really warps our estimates, while multiple imputation seems to do a great job.

Missing Not at Random (MNAR)

When data is MNAR, that is extremely problematic. MNAR data indicates that the reason variable x is missing is because of someone's value on variable x directly. When MCAR, the reason is completely random (i.e., not due to x, nor any other variables). When MAR, the reason x is missing is due to other variables in the dataset (e.g., a, b, etc.) When MNAR, the only thing we can use to help us predict the missing x value is the missing x value. Thus, this is extremely problematic.

Let's again start with our full data.

```
d <- read.csv('https://stats.idre.ucla.edu/wp-content/uploads/2016/02/hsbdemo.dat', header = F)
```

We will again make a binary indicator for whether or not an individual's math score should be missing. However, this time, the probability is going to depend on that individual's math score itself.

```
set.seed(681)

d$na_math <- rbinom(n = nrow(d), size = 1,
                    prob = ifelse(d$math > quantile(d$math, probs = 0.75) &
                                  d$math < quantile(d$math, probs = 0.95) |
                                  d$math < quantile(d$math, probs = 0.1),
                                  1, 0))
```

Before we delete values for some observations, let's again conduct a two-sample t-test.

```
# two-sample t-test ?
t.test(math ~ na_math, data = d)
```

```
##
##  Welch Two Sample t-test
##
## data:  math by na_math
## t = -4.0965, df = 66.776, p-value = 0.0001157
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -10.381189  -3.578811
## sample estimates:
## mean in group 0 mean in group 1
##           50.90           57.88
```

Now we see our two groups are very different from one another (m1 = 50.9, m2 = 57.9).

```
# overwrite the income value as missing if the missing indicator == 1
d$math <- ifelse(d$na_math == 1, NA, d$math)

# view our data
d[100:110,]
```

```
##     read write math na_math
## 100   57    41   57       0
## 101   44    52   51       0
## 102   57    52   40       0
## 103   52    55   50       0
## 104   42    41   57       0
## 105   63    49   NA       1
## 106   61    59   49       0
## 107   60    54   50       0
## 108   52    54   55       0
## 109   52    59   48       0
## 110   57    55   52       0
```

```
# proportion of missing?
prop.table(table(d$na_math))["1"]
```

```
##    1
## 0.25
```

Is our data MCAR?

```
# Little's (1988) MCAR test
naniar::mcar_test(d[,c("math", "read", "write")])
```

```
## # A tibble: 1 x 4
##   statistic    df  p.value missing.patterns
##       <dbl> <dbl>    <dbl>            <int>
## 1      15.4     2 0.000447                2
```

Unfortunately, no ($p = 0.0004$).

Let's estimate a linear regression using listwise deletion when 25% of our data is MNAR.

```
missing_model <- lm(math ~ 0 + read + write, data = d)
summary(missing_model) # observations deleted due to missingness
```

```
##
## Call:
## lm(formula = math ~ 0 + read + write, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.1326  -4.0538   0.6926   4.8849  19.9739
##
## Coefficients:
##       Estimate Std. Error t value Pr(>|t|)
## read   0.58543    0.06338   9.237 2.52e-16 ***
## write  0.40122    0.06290   6.379 2.15e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.894 on 148 degrees of freedom
##   (50 observations deleted due to missingness)
## Multiple R-squared:  0.9823, Adjusted R-squared:  0.9821
## F-statistic:  4115 on 2 and 148 DF,  p-value: < 2.2e-16
```

We see our new estimates indicate that reading and writing scores are both significant predictors. How do our estimates from the missing data model using listwise deletion compare to those from the full model?

```
coef(full_model)
```

```
##      read     write
## 0.5105818 0.4856488
```

```
coef(missing_model)
```

```
##      read     write
## 0.5854334 0.4012227
```

When our data is MNAR, even with only 25%, our estimates using listwise deletion are now very different from those we got with the full sample.

What if we try to use multiple imputation to help us estimate our model? We will use the "mice" package in R, with $m = 50$ imputed datasets estimated using the predictive mean matching.

```
library("mice")

imps <- mice(d, m = 50, method = "pmm")
```

```
## Warning: Number of logged events: 250
```

```
res <- with(imps, lm(math ~ 0 + read + write))
```

Before we look at our multiple imputation estimates, let's remember what our estimates were for the full model and the missing model using listwise deletion.

```
coef(full_model)
```

```
##      read     write
## 0.5105818 0.4856488
```

```
coef(missing_model)
```

```
##      read     write
## 0.5854334 0.4012227
```

And now, our multiple imputation estimates.

```
summary(pool(res))[,2]
```

```
## [1] 0.5796665 0.3982474
```

Our multiple imputation estimates are still extremely off from those obtained from the full model. Thus, when missing data are MNAR, listwise deletion really warps our estimates, but multiple imputation is not able to help us.