Lab #2 Discussion

Biostatistics 210

0. BACKGROUND

This lab is meant to give you further understanding of the missing at random assumption and how it can be handled by multiple imputation.

2. EXPLORING MISSINGNESS

What fraction of spider values are missing?

1. What fraction of non-missing data has spiders present: how does that compare with the actual spiders data

```
. tab spiders miss_mar, col

+-------------------+
| Key               |
|-------------------|
|     frequency     |
| column percentage |
+-------------------+

           |     Missing Data:
  Presence |      Scenario #3
of Spiders | Not Missi    Missing |     Total
-----------+----------------------+----------
    Absent |     1,281      1,290 |     2,571
           |     51.45      51.39 |     51.42
-----------+----------------------+----------
   Present |     1,209      1,220 |     2,429
           |     48.55      48.61 |     48.58
-----------+----------------------+----------
     Total |     2,490      2,510 |     5,000
           |    100.00     100.00 |    100.00
```

The finding of spiders is 1209/2490 (48.6%) in the non-missing data and 2429/5000 (48.6%) in the overall data -- so we don't see that the finding of spiders is any more or less frequent in the available data.

2. Can you see any association between other data points and whether the spiders is missing?

When you do this look at hepatomegaly, death and their interaction

If we look at the association of missingness with hepatomegaly and death without the interaction, we get

```
logistic miss_mar dead hepatom

Logistic regression                                     Number of obs   =       5000
                                                        LR chi2(2)      =       0.58
                                                        Prob > chi2     =     0.7499
Log likelihood = -3465.4081                             Pseudo R2       =     0.0001

------------------------------------------------------------------------------
    miss_mar | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        dead |   .9590319   .0547155    -0.73   0.463     .8575701    1.072498
     hepatom |   1.016541   .0579901     0.29   0.774     .9090059    1.136797
------------------------------------------------------------------------------
```

If we include the interaction, we get

```
. logistic miss_mar dead hepatom prod

Logistic regression                                     Number of obs   =       5000
                                                        LR chi2(3)      =    1344.61
                                                        Prob > chi2     =     0.0000
Log likelihood = -2793.3916                             Pseudo R2       =     0.1940

------------------------------------------------------------------------------
    miss_mar | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        dead |   9.545004   .9156519    23.52   0.000     7.908978    11.51945
     hepatom |   9.322838   .8731806    23.84   0.000     7.759331    11.20139
        prod |   .0107197   .0014421   -33.72   0.000     .0082352    .0139538
------------------------------------------------------------------------------
```

Hence, the missingness depends on the outcome and on hepatomegaly combined. This is not evident unless we include the interaction in the logistic model. Just to elaborate on what's going on, check out the missingness scenario.

Below, I tabulate the association between death and hepatomegaly for the available data and for the data overall.

```
. tab hepat dead if miss_mar==0, col

  Presence |
        of |
Hepatomega |         Died
        ly |         0          1 |     Total
-----------+----------------------+----------
    Absent |     1,016        220 |     1,236
           |     80.96      17.81 |     49.64
-----------+----------------------+----------
   Present |       239      1,015 |     1,254
           |     19.04      82.19 |     50.36
-----------+----------------------+----------
     Total |     1,255      1,235 |     2,490
           |    100.00     100.00 |    100.00


 tab hepat dead, col


  Presence |
        of |
Hepatomega |         Died
        ly |         0          1 |     Total
-----------+----------------------+----------
    Absent |     1,420      1,055 |     2,475
           |     55.80      42.97 |     49.50
-----------+----------------------+----------
   Present |     1,125      1,400 |     2,525
           |     44.20      57.03 |     50.50
-----------+----------------------+----------
     Total |     2,545      2,455 |     5,000
           |    100.00     100.00 |    100.00
```

We can see that hepatomegaly is strongly associated with death in the available data but quite a bit less strongly in the full data. That's because the missingness in spiders disproportionately is among those who had hepatomegaly and lived. This is going to distort the assoication between death and hepatomegaly in the available data.

3. COMPLETE CASE ANALYSIS

The underlying data has been generated to fit a logistic regression model with OR: 1.5 for copper_50, 1.5 for bilir_2, and 2.0 for spiders and hepatomegaly. Verify this on the full data

The analysis using the complete data gives

```
logistic dead spiders hepat bilir copper
```

```
logistic dead spiders hepat bilir copper

Logistic regression                                     Number of obs   =       5000
                                                        LR chi2(4)      =     996.98
                                                        Prob > chi2     =     0.0000
Log likelihood = -2966.4366                             Pseudo R2       =     0.1439

------------------------------------------------------------------------------
        dead | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     spiders |   1.979995   .1251107    10.81   0.000     1.749359    2.241038
     hepatom |   1.837629   .1157109     9.66   0.000     1.624276    2.079006
     bilir_2 |   1.523843   .0340713    18.84   0.000     1.458506    1.592106
   copper_50 |   1.502033   .0306591    19.93   0.000     1.443128    1.563341
------------------------------------------------------------------------------
```

The data were simulated so that the OR for spiders and hepatom is 2.0 and for bilir_2 and copper_50 it is 1.50. We see results which are very close to that in the complete data. In the complete cases analysis, we see

```
. logistic dead spider_mar hepat copper_50 bilir_2

Logistic regression                                     Number of obs   =       2490
                                                        LR chi2(4)      =    1382.06
                                                        Prob > chi2     =     0.0000
Log likelihood = -1034.8263                             Pseudo R2       =     0.4004

------------------------------------------------------------------------------
        dead | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  spider_mar |   2.093397   .2352538     6.57   0.000     1.679555    2.609208
     hepatom |   21.87057   2.545566    26.51   0.000     17.40952    27.47473
   copper_50 |   1.503422   .0543631    11.28   0.000      1.40056    1.613838
     bilir_2 |   1.561737   .0612294    11.37   0.000     1.446224    1.686475
------------------------------------------------------------------------------
```

not only do we lose half of the observations but the OR for hepatom is badly biased. That is because the spiders is available on people with hepatom who died and without hepatom who lived -- thus this distorts the relationship between hepatom and death in the available data. This illustrates that a complete cases analysis may give a biased estimator of an effect for any of the variables in the regression. We tried two different scenarios for multiple imputation.

First, we tried imputation using the real model for missingness

```
 mi estimate, or: logistic dead hepat spider_mar copper_50 bilir_2

Multiple-imputation estimates                     Imputations     =         50
Logistic regression                               Number of obs   =       5000
                                                  Average RVI     =     0.3362
DF adjustment:   Large sample                     DF:     min     =     131.24
                                                          avg     =  105142.91
                                                          max     =  288745.60
Model F test:       Equal FMI                     F(   4, 2236.6) =     133.03
Within VCE type:          OIM                     Prob > F        =     0.0000

------------------------------------------------------------------------------
        dead | Odds Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     hepatom |   1.820717   .1177552     9.27   0.000     1.603935    2.066798
  spider_mar |   1.911794   .1928231     6.43   0.000     1.565994    2.333953
   copper_50 |   1.492272   .0305179    19.57   0.000      1.43364    1.553301
     bilir_2 |   1.516759   .0340898    18.53   0.000     1.451394    1.585068
------------------------------------------------------------------------------
```

Does imputation appear to give results which match the complete case analysis and the true odds ratios (hepat & spider = 2, copper_50 & bilir_2 = 1.5)

Note, everyone's imputation results are likely to be different since there is an element of randomness in the imputations. We see that the bias in hepat is gone also we see that we recover substantial efficiency for copper and bilirubin relative to the complete case analysis.

However, in reality we will not know with certainty the model for missingness and we created imputations by simply putting in all the outcome and other predictors into the imputation model (but forgetting the interaction)

```
.  mi estimate, or: logistic dead spider_mar hepat copper_50 bilir_2

Multiple-imputation estimates                     Imputations     =         50
Logistic regression                               Number of obs   =       5000
                                                  Average RVI     =     0.5912
DF adjustment:   Large sample                     DF:     min     =      90.08
                                                          avg     =   11157.92
                                                          max     =   28966.05
Model F test:       Equal FMI                     F(   4, 1066.9) =     111.08
Within VCE type:          OIM                     Prob > F        =     0.0000

------------------------------------------------------------------------------
        dead | Odds Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  spider_mar |    2.11085   .2617353     6.03   0.000     1.649967    2.700469
     hepatom |   1.831701   .1210256     9.16   0.000     1.609173    2.085001
   copper_50 |   1.499572   .0312858    19.42   0.000     1.439488    1.562165
     bilir_2 |   1.534544   .0353762    18.58   0.000     1.466747    1.605474
------------------------------------------------------------------------------
```

Does imputation appear to give results which match the complete case analysis and the true odds ratios (hepat & spider = 2, copper_50 & bilir_2 = 1.5)

It is striking that there the two give fairly similar results even thought the model for imputations was not very similar. It is worthwhile to compare the relative efficiencies.