

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

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

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 ly	Died		
	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 ly	Died		
	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 association 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 only the predictors in the model but excluding the outcome variable “dead”. When we

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

Multiple-imputation estimates	Imputations	=	50
Logistic regression	Number of obs	=	5,000
	Average RVI	=	0.1534
	Largest FMI	=	0.4350
DF adjustment: Large sample	DF: min	=	264.12
	avg	=	1726028.45
	max	=	4372985.66
Model F test: Equal FMI	F(4, 7335.3)	=	145.96
Within VCE type: OIM	Prob > F	=	0.0000

dead	Odds Ratio	Std. Err.	t	P> t	[95% Conf. Interval]	
hepatom	1.777246	.1111801	9.19	0.000	1.572166	2.009078
spider_mar	1.262077	.1043488	2.82	0.005	1.072468	1.485208
copper_50	1.487807	.0299445	19.74	0.000	1.430259	1.54767
bilir_2	1.510691	.0335118	18.60	0.000	1.446416	1.577822
_cons	.1970397	.015597	-20.52	0.000	.1687134	.2301218

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 but that we now have bias in spider s which should have an OR of 2.0 but which is instead 1.26 (and the value 2 is not contained in the confidence interval).

Now, lets look at a the estimated based on imputations which include the outcome

Multiple-imputation estimates	Imputations	=	50
Logistic regression	Number of obs	=	5,000
	Average RVI	=	0.4921
	Largest FMI	=	0.7063
DF adjustment: Large sample	DF: min	=	99.86
	avg	=	28,634.73
	max	=	112,174.52
Model F test: Equal FMI	F(4, 1324.2)	=	118.47
Within VCE type: OIM	Prob > F	=	0.0000

dead	Odds Ratio	Std. Err.	t	P> t	[95% Conf. Interval]	
hepatom	1.82876	.1194442	9.24	0.000	1.608995	2.078541
spider_mar	2.04652	.2370656	6.18	0.000	1.626315	2.575297
copper_50	1.497341	.0308596	19.59	0.000	1.438062	1.559063
bilir_2	1.532889	.0352926	18.55	0.000	1.46525	1.60365
_cons	.1497758	.0148213	-19.19	0.000	.1232838	.1819606

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)

This model has unbiased estimates of all the coefficient. Inclusion of the outcome as a variable in the imputations. It has fixed the bias in spiders compared to imputations without the outcome and the bias in hepatomegaly. I also want to note that the standard errors for copper and bilirubin and lower after imputation than in the complete case analysis.