Assignment 4

Joan Shim and Beimnet Taye

2024-04-17

```
##                         chain:1 chain:2 chain:3
## cvd                       1.199   1.199   1.199
## sex                       1.427   1.427   1.427
## educ                      1.982   1.982   1.982
## age                      49.518  49.518  49.518
## cursmoke.miss             1.503   1.497   1.500
## missing_cursmoke.miss     0.228   0.228   0.228

## mean_cursmoke.miss   sd_cursmoke.miss
##           1.019747           1.000237
```

Q1

Table 1: Descriptive statistics (means/proportions) for each fully observed covariate by indicator of whether smoking is observed or missing.

Covariate	Smoking Missing	Smoking Observed
N	931	3152
CVD	197 (21.2)	617 (19.6)
Age	50.55 (8.65)	49.21 (8.51)
Male	275 (29.5)	1469 (46.6)
< HS grad	376 (40.4)	1319 (41.8)
HS grad	291 (31.3)	946 (30.0)
Some college	171 (18.4)	510 (16.2)
College grad	93 (10.0)	377 (12.0)

Q2

Table 2: Parameter values and standard errors (in parentheses) from logistic regression model of CVD on smoking, age, sex, and education applying several missing data methods.

Model Coefficient	Full Analysis	Complete Case Analysis	IP Weighting	Multiple Imputation	Fully Bayesian
Intercept	-3.7616 (0.28234)	-3.9589 (0.32587)	-3.9478 (0.32285)	-3.8074 (0.28291)	-3.8197 (0.28878)
Smoking	0.2546 (0.08442)	0.2897 (0.09683)	0.2807 (0.09662)	0.2933 (0.08409)	0.2944 (0.09666)
Age	0.0458 (0.00492)	0.0479 (0.00568)	0.0483 (0.00563)	0.0463 (0.00492)	0.0465 (0.00494)
Male	0.3042 (0.08289)	0.3534 (0.09461)	0.341 (0.09354)	0.3001 (0.08252)	0.3004 (0.08257)
HS grad	-0.2813 (0.09745)	-0.2411 (0.1122)	-0.2056 (0.1133)	-0.2774 (0.09729)	-0.2801 (0.09778)
Some college	-0.4917 (0.12398)	-0.4802 (0.14485)	-0.492 (0.14574)	-0.4859 (0.12366)	-0.4933 (0.12326)
College grad	-0.5185 (0.14267)	-0.4705 (0.15986)	-0.4941 (0.15932)	-0.5115 (0.14218)	-0.5225 (0.14109)

Q3

A complete case analysis is valid when either the missing variable is missing completely at random or missing at random and is not dependent on the outcome and when a few observations are missing data (<10%). This is not the case here since according to table 1 22.8% are missing smoking status.

Q4

All of these methods require the assumption that the data is **not** not missing at random. This means that the probability of having missing smoking data is not dependent on being a smoker. I don't think that this is the case here since the percentage missing seems similar to the percentage observed across the other covariates and outcome. So I think all three methods are valid. That being said, while I don't think this is the case here, in general I can see how patients who do smoke might be less likely to report smoking status due to some social desirability bias. This would lead to the missingness of smoking data being NMAR and thus we would not be able to use IP weighting, imputation, or Bayesian methods.

Q5

IPW relies on modeling the outcome with weights that are the inverse of the probability of being observed. These probabilities are first derived by modeling them as a function of the outcome and covariates.

Imputation/Bayesian methods on the other hand rely on modeling the missing covariate/covariates themselves.

I would use IPW when the sample size is large and I'm not too sure about the underlying distribution of my missing covariate. I would use imputation/Bayesian methods when my sample size is small or I have multiple missing covariates.

Q6

i.

In terms of the coefficients, all had higher values/point estimates that are farther from the null than the full data model. When comparing the precision all, except imputation, had large standard errors than the full model. Multiple imputation actually had a very slighter smaller standard error than the full model.

ii.

I would use the multiple imputation model since it seems to be more precise than the other methods.