Imputation of Missing Values in Serum Cholesterol

Rokhsona Pervin

Introduction:

Cholesterol is very important in our body. It is a waxy material that obtains from two sources: our body and food. Our body, and especially our liver, produces all the cholesterol we need and spreads it through the blood. However cholesterol is also found in foods from animal sources, such as meat, poultry and full-fat dairy products. Our liver prepares more cholesterol when we eat a diet high in saturated and *trans* fats.

Too much cholesterol can produce plaque between layers of artery walls, making it harder for our heart to circulate blood. Plaque can break open and cause blood clots. If a clot blocks an artery that feeds the brain, it causes a stroke. If it blocks an artery that feeds the heart, it causes a heart attack. So it is very important to know about which other factors that have significant effects on cholesterol. That's why we are interested in this topic.

Data Collection:

The data and variables information are obtained from

https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/processed.va.data

This site includes the raw data and column information needed to read the data into SAS.

This database contains 76 attributes, but all published experiments refer to using a subset of 14 of them. Since our focus is to impute missing values and based on this imputation we fits a model that is appropriate for our data, so we limit our study three variables which has some missing values.

Objectives:

The goals of our project are

- to estimate the effect of serum cholesterol on age and gender, and
- to estimate missing values using MI technique.

Multiple Imputation:

Multiple imputation provides a convenient approach for dealing with data sets with missing values. Instead of filling in a single value for each missing value, Rubin's (1987) multiple imputation procedure replaces each missing value with a set of plausible values that represent the uncertainty about the right value to impute. These multiple imputed data sets are then analyzed by using standard procedures for complete data and combining the results from these analyses. No matter which complete-data analysis is used, the process of combining results from different imputed data sets is essentially the same. This results in valid statistical inferences that properly reflect the uncertainty due to missing values.

A multiple imputation analysis consists of three distinct steps:

1) **Imputation phase:** The imputation phase creates multiple copies of the data set (e.g., $m = 5$), each of which contains different estimates of the missing values. In fact, this step is an iterative version of stochastic regression imputation, although its mathematical underpinnings rely heavily on Bayesian estimation principles.

2) A**nalysis phase:** The purpose of the analysis phase is to analyze the filled-in data sets. This step uses the similar statistical procedures that you would have used had the data been complete. Procedurally, the only difference is that you perform each analysis m times, once for each imputed data set. The analysis phase yields m sets of parameter estimates and standard errors, so the purpose of the

3) **pooling phase :** The goal of the pooling phase is to combine everything into a single set of results. Rubin (1987) outlined relatively straightforward formulas for pooling parameter estimates and standard errors. For example, the pooled parameter estimate is simply the arithmetic average of the *m* estimates from the analysis phase. Combining the standard errors is kind of more difficult but follows the similar philosophy.

MCMC method:

In the general case where the missing data pattern is arbitrary, one may incorporate different kind of variables (nominal, ordinal, continuous), it is very difficult to estimate the joint posterior distribution. In that situations, one iterative simulation procedure which is designed by statisticians, allow us to generate complex joint posterior distribution. The PROC MI MCMC method is similar kind of algorithm that generates the joint posterior distribution when the data follows arbitrary pattern and the underlying data follows multivariate normal distribution.

In MCMC,a Markov chain has been designed for the distribution of elements to converge to a common distribution. By repeatedly simulating steps of the chain, it generates draws from the distribution of interest. In Bayesian inference, information about the unknown parameters is expressed in the form of posterior distribution. MCMC is applied to the Bayesian inference as a method for analyzing posterior distribution. In other words, using MCMC, the joint posterior distribution of unknown parameters is generated to get simulation based estimates of posterior parameter which are of interest.

The MCMC approach consists of the following two steps:

I-Step (imputation step): Using the estimated mean vector and covariance matrix, the I-step generates the missing values for each observation independently. In other words, if you denote the variables with missing values for observation i by $Y_{i(mis)}$ and the variables with observed values by $Y_{i(obs)}$, then the I-step draws values for $Y_{i(mis)}$ from a conditional distribution $Y_{i(mis)}$ given $Y_{i(obs)}$.

P-step (posterior step): The P-step simulates the posterior population mean vector and covariance matrix from the complete sample estimate. These new estimates are then used in the I-step. Without prior information about the parameters, a noninformative prior is used. You can also use other informative priors. For example, a prior information about the covariance matrix may be helpful to stabilize the inference about the mean vector for a near singular covariance matrix.

The two steps are iterated long enough for the results to be reliable for a multiply imputed data set (Schafer 1997, p.72). The goal is to converge to their stationary distribution and then to generates an approximately independent draw of the missing values.

That is, with a current parameter estimate $\theta^{(t)}$ at t^{th} iteration, the I-step draws $Y^{(t+1)}_{mis}$ from $p(Y_{mis}|Y_{obs,}\theta^{(t)})$ and the P-step draws $\theta^{(t+1)}$ from $p(\theta|Y_{obs,}\theta^{(t)}, Y^{(t+1)}_{mis})$.

This creates a markov chain ($Y^{(1)}_{mis}, \theta^{(1)}$), ($Y^{(2)}_{mis}, \theta^{(2)}$),…,

Which converges in distribution to $p(Y_{mis}, , \theta\ |Y_{obs,})$.

Data Analysis:

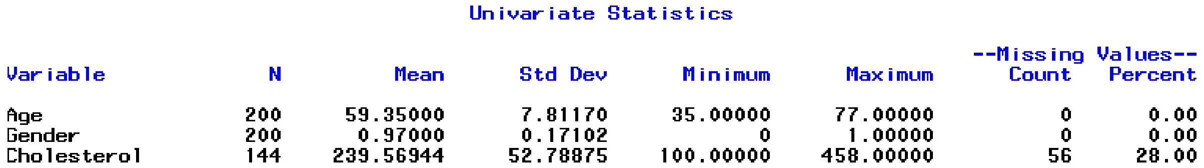
Variables used in this project are,

AGE: Age in years; continuous and fully observed

GENDER: Gender (1=male; 0=female); binary and fully observed

CHOLESTEROL: Serum cholesterol in mg/dl; continuous with some missing data

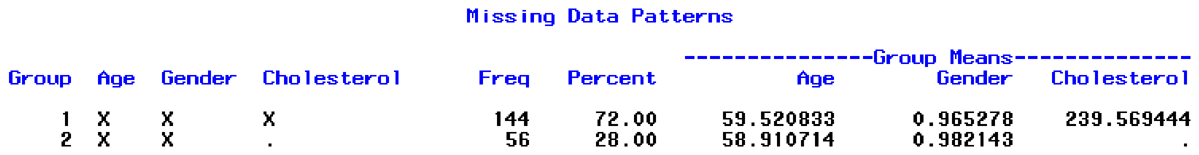
Summary statistics:



```
                                        Univariate Statistics

                                                                                      --Missing Values--
Variable              N          Mean       Std Dev       Minimum       Maximum        Count     Percent

Age                 200      59.35000       7.81170      35.00000      77.00000            0        0.00
Gender              200       0.97000       0.17102             0       1.00000            0        0.00
Cholesterol         144     239.56944      52.78875     100.00000     458.00000           56       28.00
```

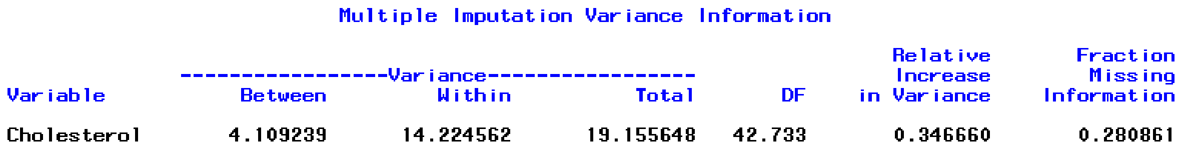
The results show that about 28% data are missing on Cholesterol. We will impute our missing data using multiple imputation techniques. This project includes a number of techniques, including use of multiple imputation to impute missing data for the serum cholesterol variable with the default MCMC method. The data set generated by PROC MI are performed by a linear regression analysis using PROC GLM. This analysis estimates the linear regression of a person's serum cholesterol on age and gender. The ODS OUTPUT statement creates a parameter estimates data set named glmest. This data set contains the regression parameter estimates and standard errors that are the necessary inputs for use by PROC MIANALYZE in the MI "combining step". The data set is next sorted by _IMPUTATION_ and finally as input to PROC MIANALYZE.

Missing Data Patterns

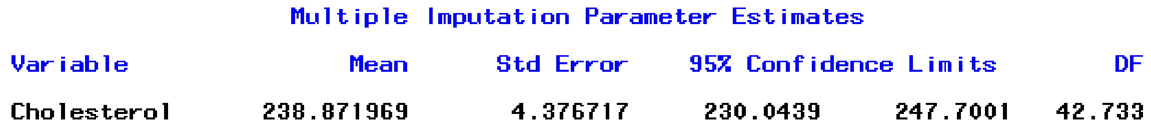


```
                                      Missing Data Patterns

                                                         ---------------Group Means--------------
Group   Age   Gender   Cholesterol        Freq    Percent             Age         Gender    Cholesterol

    1   X     X        X                   144      72.00       59.520833       0.965278     239.569444
    2   X     X        .                    56      28.00       58.910714       0.982143              .
```

There are two missing data groups: Group 1 with with 144 records (72.0%) fully observed, group 2 with 56 records (28.0%) with missing data on just Cholesterol. The missing data pattern is arbitrary.

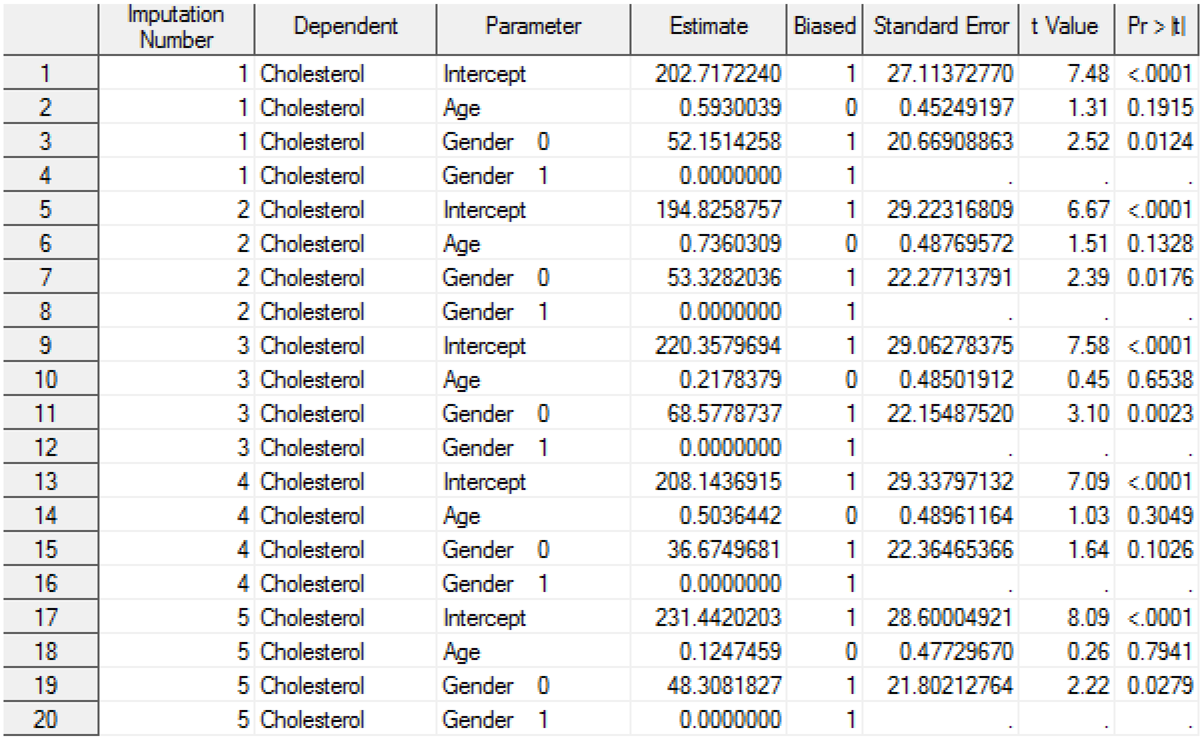


```
                        Multiple Imputation Variance Information

                                                                    Relative        Fraction
                  -----------------Variance-----------------        Increase         Missing
Variable              Between          Within           Total        DF    in Variance     Information

Cholesterol          4.109239       14.224562       19.155648    42.733       0.346660        0.280861
```



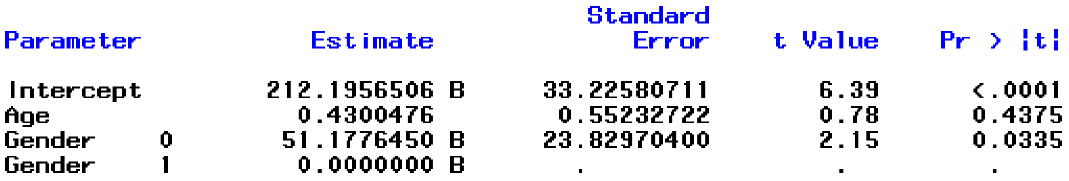
```
                    Multiple Imputation Parameter Estimates

Variable                 Mean       Std Error     95% Confidence Limits         DF

Cholesterol        238.871969        4.376717     230.0439       247.7001    42.733
```

Listing of Processed glmest Data Set



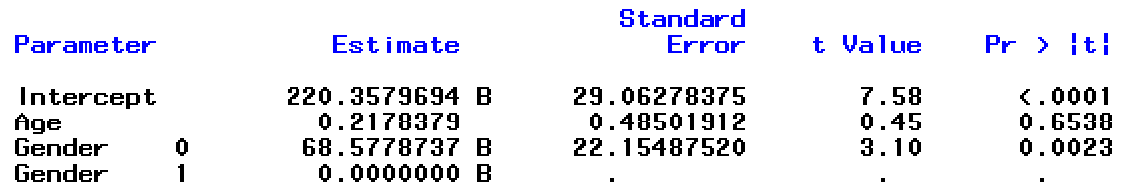
	Imputation Number	Dependent	Parameter	Estimate	Biased	Standard Error	t Value	Pr > |t|
1	1	Cholesterol	Intercept	202.7172240	1	27.11372770	7.48	<.0001
2	1	Cholesterol	Age	0.5930039	0	0.45249197	1.31	0.1915
3	1	Cholesterol	Gender 0	52.1514258	1	20.66908863	2.52	0.0124
4	1	Cholesterol	Gender 1	0.0000000	1	.	.	.
5	2	Cholesterol	Intercept	194.8258757	1	29.22316809	6.67	<.0001
6	2	Cholesterol	Age	0.7360309	0	0.48769572	1.51	0.1328
7	2	Cholesterol	Gender 0	53.3282036	1	22.27713791	2.39	0.0176
8	2	Cholesterol	Gender 1	0.0000000	1	.	.	.
9	3	Cholesterol	Intercept	220.3579694	1	29.06278375	7.58	<.0001
10	3	Cholesterol	Age	0.2178379	0	0.48501912	0.45	0.6538
11	3	Cholesterol	Gender 0	68.5778737	1	22.15487520	3.10	0.0023
12	3	Cholesterol	Gender 1	0.0000000	1	.	.	.
13	4	Cholesterol	Intercept	208.1436915	1	29.33797132	7.09	<.0001
14	4	Cholesterol	Age	0.5036442	0	0.48961164	1.03	0.3049
15	4	Cholesterol	Gender 0	36.6749681	1	22.36465366	1.64	0.1026
16	4	Cholesterol	Gender 1	0.0000000	1	.	.	.
17	5	Cholesterol	Intercept	231.4420203	1	28.60004921	8.09	<.0001
18	5	Cholesterol	Age	0.1247459	0	0.47729670	0.26	0.7941
19	5	Cholesterol	Gender 0	48.3081827	1	21.80212764	2.22	0.0279
20	5	Cholesterol	Gender 1	0.0000000	1	.	.	.

Estimated parameters without imputation:



```
                                           Standard
Parameter               Estimate              Error    t Value    Pr > |t|

Intercept            212.1956506 B      33.22580711       6.39      <.0001
Age                    0.4300476         0.55232722       0.78      0.4375
Gender      0         51.1776450 B      23.82970400       2.15      0.0335
Gender      1          0.0000000 B        .                .         .
```

Estimated parameters using MI technique:



```
                                          Standard
Parameter              Estimate              Error    t Value    Pr > |t|

Intercept            220.3579694 B     29.06278375       7.58      <.0001
Age                    0.2178379        0.48501912       0.45      0.6538
Gender      0         68.5778737 B     22.15487520       3.10      0.0023
Gender      1          0.0000000 B       .                .          .
```

Interpretations: The parameter estimates results suggest that, all other variables held constant, women (GENDER=0) have significantly greater odds than men of cholesterol. In other words women have on an average 36 unit more effect on cholesterol than male. Moreover, if look at the results of standard error, we see that result without imputation of missing values have more standard error in all variables than imputed values. And the result is more significant in Multiple Imputation case. Overall we can say that Multiple Imputation works better than complete case analysis.

SAS Codes

```
libname xx 'D:\Math 6820 Missing data analysis\Project\project data';

data xx.New_Va_data_set;
keep Age Gender  Cholesterol;
set  FINAL ;
run;



proc glm data=xx.New_Va_data_set;
class Gender;
model Cholesterol = Age Gender / solution;
run;


proc mi nimpute=0 data=New_Va_data_set simple;
var Age Gender Cholesterol;
run;

proc mi data=New_Va_data_set out=Najib3 seed=2012 nimpute=5;
var  Age Gender Cholesterol;
run;
proc print data=Najib3;
```

```
run;


proc glm data=Najib3;
class Gender;
model Cholesterol = Age Gender / solution;
ods output parameterestimates=glmest;
by _imputation_;
run;
proc print data=glmest;
run;

proc sort data=glmest; by _imputation_; run;
proc print data=glmest; run;

proc mianalyze parms=glmest;
class gender;
 modeleffects Intercept Age gender;
run;
```