**Multiple Imputation and Missing Data**

M. Murrietta

1/21/2018

**MY NOTES (ERASE BEFORE TURNING IN):**

Make sure we specify that class variables as appropriate; if there is missing data in a class variable it my be imputed incorrectly if we don’t specify that it is a class (or discrete) variable. (DONE)

Now need to go back and fix any terminology that may be incorrect (example: we no longer have used the EM algorithm. Also should check if FCS still incorporates MCMC methods or if it is something else completely, and, if it is, we need specify why it is suitable for our case.)

**Abstract**

We were given a subset of the **carmpg** data set that was missing data. The task was to figure out the pattern of missingness and then apply the appropriate imputation method to come up with a better model than would be produced using the incomplete data set.

**1. Introduction**

Multiple imputation, as opposed to single imputation, provides better filling of missing data points because it accounts for the variability inherent in the data set that is provided. The pattern by which data is missing affects how the missing value are filled: in a monotone missing pattern a parametric regression method that assumes multivariate normality or some nonparametric method utilizing propensity scores is used. The cases where the missing data has an arbitrary patter use the Markov Chain Monte Carlo (MCMC) method. Whatever method is chosen, some imputation iterations are performed, the estimates are combined, and finally the combined estimates are used in a final model that utilizes all observations of the data set.

**2. Methods**

The data was imported into SAS and an initial model was produced revealing that 20 of the 38 observations were missing and therefore omitted from the analysis via list-wise deletion. The specific model used was:

Where MPG, CYL, SIZE, HP, WGT, ACC, and ENG are the MPG, CYLINDER, SIZE, HP, WEIGHT, ACCEL, and ENG\_TYPE variables, respectively. This initial model had an R-square of 0.9243 and an adjusted R-square of 0.8830.

Analysis of the missing patterns was performed, and the missing data were found to have an arbitrary missing pattern so MCMC was used with the initial estimate being derived using the expectation maximization (EM) algorithm with 100 iterations of bootstrap resampling in effort to provide different starting values. There is some discussion on the relation between number of imputations, proportion of missing data, and relative efficiency [1, 2], but thankfully Horton and Lipsitz [3] specify that trying different numbers of imputations and checking for stability between sets is practically effective; 5 and 10 imputations were tried and 5 was found to be suitable by comparing the variance information for each (shown in Table 1 and Table 2). The CYLINDERS and ENG\_TYPE variables were specified to be class variables since they hold discrete values in this dataset. The 5 imputed data sets were then analyzed using a general regression procedure and the model specified in equation 1.

| **Variance Information (5 Imputations)** | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **Variance** | | | **DF** | **Relative Increase in Variance** | **Fraction Missing Information** | **Relative Efficiency** |
| **Between** | **Within** | **Total** |
| **SIZE** | 0.579116 | 207.684273 | 208.379212 | 35.029 | 0.003346 | 0.003341 | 0.999332 |
| **HP** | 0.333381 | 18.379661 | 18.779718 | 34.267 | 0.021766 | 0.021525 | 0.995714 |
| **WEIGHT** | 0.000066624 | 0.013158 | 0.013238 | 34.927 | 0.006076 | 0.006058 | 0.998790 |
| **ACCEL** | 0.003130 | 0.063185 | 0.066942 | 32.333 | 0.059453 | 0.057599 | 0.988611 |

| **Table 2: Variance Information (10 Imputations)** | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **Variance** | | | **DF** | **Relative Increase in Variance** | **Fraction Missing Information** | **Relative Efficiency** |
| **Between** | **Within** | **Total** |
| **SIZE** | 1.319392 | 208.764398 | 210.215729 | 34.901 | 0.006952 | 0.006915 | 0.999309 |
| **HP** | 0.602177 | 18.639405 | 19.301800 | 33.794 | 0.035537 | 0.034570 | 0.996555 |
| **WEIGHT** | 0.000038497 | 0.013171 | 0.013213 | 35.036 | 0.003215 | 0.003207 | 0.999679 |
| **ACCEL** | 0.009537 | 0.067363 | 0.077854 | 28.655 | 0.155734 | 0.138219 | 0.986367 |



Figure 1 Residual Diagnostic plots from Imputations 1 and 2

**3. Results**

The models for each imputation were examined and were mostly found to be as good or better than the initial model that was made using the missing data based on the r-square and adjusted r-squared values (lowest adjusted r-square was 0.9003, lowest r-square was 0.9165 ). The qq-plot of the residuals for each model showed a good fit to the normal distribution but this was expected since the initial model did so as well.

There was one observation (24) that was consistently flagged as an outlier for each imputation as visible in the Cook’s D chart; some of the residual diagnostics for the first and second imputation are shown in Fig. 1. Table 3 shows the estimates from the initial model and the combined model side-by-side which shows an appreciable improvement in the standard error for each of the estimates.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Table 3: Side-by-Side Estimate Comparison** | | | | |
| **Parameter** | **Parameter Estimates (Initial)** | | **Parameter Estimates (5 Imputations)** | |
| **Estimate** | **Standard Error** | **Estimate** | **Standard Error** |
| **Intercept** | 70.14772 | 8.03838 | 68.388562 | 4.933833 |
| **CYLINDERS** | -3.33403 | 1.56072 | -2.384973 | 0.857294 |
| **SIZE** | 0.02280 | 0.03207 | 0.038598 | 0.021495 |
| **HP** | -0.19546 | 0.08065 | -0.144995 | 0.049796 |
| **WEIGHT** | -0.30623 | 5.13263 | -5.090662 | 3.173186 |
| **ACCEL** | -0.78199 | 0.58264 | -0.631053 | 0.307801 |
| **ENG\_TYPE** | 6.59880 | 3.59008 | 4.578992 | 1.792868 |

The change in relative efficiency shown in Table 4 is somewhat concerning as relative efficiencies >0.99 were observed when comparing the 5 and 10 imputation models however these values still seem to be what is expected given the fraction of missing information for each parameter.

| **Table 4: Variance Information (5 Imputations)** | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Parameter** | **Variance** | | | **DF** | **Relative Increase in Variance** | **Fraction Missing Information** | **Relative Efficiency** |
| **Between** | **Within** | **Total** |
| **Intercept** | 4.361865 | 19.108476 | 24.342713 | 86.515 | 0.273922 | 0.232561 | 0.955555 |
| **CYLINDERS** | 0.145978 | 0.559778 | 0.734953 | 70.411 | 0.312935 | 0.259098 | 0.950733 |
| **SIZE** | 0.0000789 | 0.000367 | 0.000462 | 95.316 | 0.257633 | 0.221031 | 0.957665 |
| **HP** | 0.000654 | 0.001695 | 0.002480 | 39.954 | 0.462864 | 0.348238 | 0.934887 |
| **WEIGHT** | 1.638278 | 8.103178 | 10.069111 | 104.93 | 0.242613 | 0.210156 | 0.959664 |
| **ACCEL** | 0.005758 | 0.087832 | 0.094741 | 752.03 | 0.078669 | 0.075387 | 0.985147 |
| **ENG\_TYPE** | 0.708318 | 2.364396 | 3.214377 | 57.205 | 0.359492 | 0.288867 | 0.945382 |

The final model found that all parameter estimates including the intercept were significantly different than 0 except for the SIZE and WEIGHT parameters (see Table 5).

| **Table 5: Parameter Estimates (5 Imputations)** | | | |
| --- | --- | --- | --- |
| **Parameter** | **Theta0** | **t for H0: Parameter=Theta0** | **Pr > |t|** |
| **CYLINDERS** | 0 | -3.28 | 0.0028 |
| **SIZE** | 0 | 0.68 | 0.4992 |
| **HP** | 0 | -3.96 | 0.0004 |
| **WEIGHT** | 0 | 0.07 | 0.9431 |
| **ACCEL** | 0 | -2.71 | 0.0125 |
| **ENG\_TYPE** | 0 | 4.07 | <.0001 |
| **Intercept** | 0 | 13.83 | <.0001 |

**4. Conclusion**

The data set provided had less than 50% complete observations and it was concerning whether any model based on such incomplete data could accurately model the variability of the data set. Multiple imputation was used to fill in the missing information while accounting for variability inherent in the other variables of the data set, this was done 5 times and utilizing a 100-iteration bootstrap method to help provide sufficiently different initial values for the expectation maximization algorithm used with the MCMC method. The regression models based on those 5 imputation iterations were at least as good as the initial model by way of R-square and adjusted R-square values. Finally, the combined model was shown to have parameter estimates with smaller standard error than the initial, non-imputed data model.

**A. Code**

/\* reading in the data \*/

**data** carmpg;

infile 'C:\Users\m\Documents\SMU\qtw\hw01\carmpgdata\_2.txt' dlm='09'x dsd truncover firstobs=**2**;

input Auto $ MPG CYLINDERS SIZE HP WEIGHT ACCEL ENG\_TYPE;

**proc** **print** data=carmpg; **run**;

/\* The professor explained a more simple basic model but since we were given more variables we will use what was given to us \*/

TITLE 'Predicting MPG';

ods rtf file="C:\Users\m\Documents\SMU\qtw\hw01\initial\_model.rtf";

**PROC** **REG** DATA = carmpg;

MODEL MPG = CYLINDERS SIZE HP WEIGHT ACCEL ENG\_TYPE;

**RUN**;

**QUIT**;

ods rtf close;

/\* look at the missing data patterns to figure out which MI method to use, we use NIMPUTE=0 to not perform any imputation but just provide the missing data patterns\*/

ods rtf file="C:\Users\m\Documents\SMU\qtw\hw01\missing\_pattern.rtf";

**PROC** **MI** DATA=carmpg NIMPUTE=**0** simple seed = **35399**;

em itprint outem=outem;

VAR MPG CYLINDERS SIZE HP WEIGHT ACCEL ENG\_TYPE;

**RUN**;

ods rtf close;

ods rtf file="C:\Users\m\Documents\SMU\qtw\hw01\ods\_multiple\_imputation\_10.rtf";

**PROC** **MI** DATA=carmpg NIMPUTE=**10** out = miout seed = **35399**;

VAR MPG CYLINDERS SIZE HP WEIGHT ACCEL ENG\_TYPE;

MCMC initial=em (bootstrap=**100**) displayinit;

**RUN**;

ods rtf close;

ods rtf file="C:\Users\m\Documents\SMU\qtw\hw01\ods\_multiple\_imputation\_5.rtf";

**PROC** **MI** DATA=carmpg NIMPUTE=**5** out = miout seed = **35399**;

VAR MPG CYLINDERS SIZE HP WEIGHT ACCEL ENG\_TYPE;

MCMC initial=em (bootstrap=**100**) displayinit;

**RUN**;

ods rtf close;

/\* Try proc reg with the imputed data, create output dataset to use in MIANALYZE \*/

ods graphics on;

ods rtf file="C:\Users\m\Documents\SMU\qtw\hw01\ods\_imputed\_models.rtf";

**PROC** **REG** DATA = miout outest=outreg covout plots(label)=(CooksD RStudentByLeverage DFFITS DFBETAS);

MODEL MPG = CYLINDERS SIZE HP WEIGHT ACCEL ENG\_TYPE;

by \_Imputation\_;

**RUN**;

**quit**;

ods rtf close;

**proc** **print** data=outreg; **run**;

/\* this will output the combined estimate from all imputation runs, we should

get the original estimates from the first (non-imputed) proc reg and compare side-by-side like in the video\*/

ods rtf file="C:\Users\m\Documents\SMU\qtw\hw01\ods\_mi\_analyze\_10.rtf";

**PROC** **MIANALYZE** data = outreg;

MODELEFFECTS CYLINDERS SIZE HP WEIGHT ACCEL ENG\_TYPE Intercept;

**RUN**;

ods rtf close;

**References**

1. Allison, Paul. Why You Probably Need More Imputation Than You Think. <https://statisticalhorizons.com/more-imputations>, retrieved 1/21/2018 at 15:15 PST.
2. Pan, Qiyuan. Wei, Rong. Fraction of Missing Information (γ) at Different Missing Data Fractions in the 2012 NAMCS Physician Workflow Mail Survey\*. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4934387/>, retrieved 1/21/2018 at 15:20 PST.
3. SAS 9.2 User’s Guide, 2nd Edition. Multiple Imputation Efficiency. <https://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_mianalyze_sect013.htm>, retrieved 1/21/2018 at 15:30 PST.