Unit 2: Case Study on Multiple Imputation

Damon Resnick

January 25, 2018

Abstract

The missing values in the data set *carmpg* were imputed using multiple imputation. The missing data appears to exhibit a missing at random pattern. Using a simple linear regression model, the imputed data shows a reduced error for the parameter estimates compared to modeling the data with only the rows with no missing data. This also helps to trim the regression model and reduce overfitting.

1 Introduction

This case study focuses on the use of multiple imputation of the given *carmpg* data set using a SAS function such as *PROC MI*.

The *carmpg* data set consists of 8 different fields:

Auto – Car Make: Categorical descriptive variable (Buick Estate Wagon, Ford Country Sq. Wagon)
MPG – Miles per Gallon: Continuous variable (15.5 – 37.3)
CYLINDERS – Number of cylinders of the car's engine: Classification/Cont. variable (4, 5, 6, 8)
SIZE – Volume of the engine: Continuous variable (85 - 360)
HP – Horse Power: Continuous variable (65-155)
WIEGHT – Weight of the car: Continuous variable (1.915 – 4.36)
ACCEL – Acceleration of the car: Continuous variable (11.3-19.2)
ENG_TYPE – Type of engine: Classification 1 or 0; presumably automatic transmission or manual

Every field except for Auto and MPG have missing values. The missing values appear to be mainly missing at random. The only pattern or correlation found was that all missing SIZE values are of the 0 ENG_TYPE.

The data set has 38 total rows with 20 rows with missing data in at least one field.

Table 1 shows the missing data patterns using SAS PROC MI:

```
ODS SELECT MISSPATTERN;
PROC MI DATA = cars NIMPUTE = 0;
      VAR MPG CYLINDERS SIZE HP WEIGHT ACCEL ENG_TYPE;
RUN; quit;
```

Missing Data Patterns																
										Group Means						
Group	MPG	CYLINDERS	SIZE	HP	WEIGHT	ACCEL	ENG_TYPE	Freq	Percent	MPG	CYLINDERS	SIZE	HP	WEIGHT	ACCEL	ENG_TYPE
1	X	X	X	X	X	X	X	18	47.37	26.605556	5.333333	177.055556	101.888889	2.795333	14.355556	0.333333
2	X	X	X	X	X	X	.	2	5.26	31.350000	4.000000	95.000000	70.000000	2.125000	16.850000	.
3	X	X	X	X	X	.	X	1	2.63	18.200000	8.000000	318.000000	135.000000	3.830000	.	1.000000
4	X	X	X	X	X	.	.	1	2.63	17.600000	8.000000	302.000000	129.000000	3.725000	.	.
5	X	X	X	X	.	X	X	3	7.89	28.133333	4.666667	128.000000	72.666667	.	16.166667	0
6	X	X	X	X	.	.	X	1	2.63	21.500000	4.000000	121.000000	110.000000	.	.	0
7	X	X	X	.	X	X	X	5	13.16	22.320000	5.400000	182.800000	.	3.009800	15.240000	0.400000
8	X	X	.	X	X	X	X	2	5.26	19.100000	6.000000	.	115.000000	3.112500	15.150000	0
9	X	X	.	X	.	X	X	1	2.63	30.500000	4.000000	.	78.000000	.	14.100000	0
10	X	.	X	X	X	X	X	2	5.26	21.100000	.	176.000000	110.000000	3.087500	15.750000	0
11	X	.	X	X	X	.	X	1	2.63	18.100000	.	258.000000	120.000000	3.410000	.	0
12	X	.	X	X	.	X	X	1	2.63	17.000000	.	305.000000	130.000000	.	15.400000	1.000000

Table 1: You can see that there are no obvious patterns to the missing data.

2 Methods

A basic multiple imputation method is used with SAS to replace the missing data. Five different imputation data sets are created with PROC MI using seed 35399. The CYLINDERS and ENG-TYPE are treated as classification variables in order to make sure no unphysical values are chosen.

```
PROC MI DATA = cars NIMPUTE = 5
      OUT = MIOUT seed = 35399;
      VAR MPG CYLINDERS SIZE HP WEIGHT ACCEL ENG_TYPE;
      class CYLINDERS ENG_TYPE;
      FCS;
RUN; quit;
```

This FCS statement applies a regression method to the imputation for the continuous variables and a discriminant function method for the classification variables, filling in the missing observations with a predicted value. Slightly different estimated parameters are used for the five different imputed sets. FCS was used to include CYLINDERS and ENG_TYPE as classification variables. Since these two variables are not continuous variables, especially ENG_TYPE, then using FCS made more sense than using MCMC.

3 Results

The results of the imputation are judged by the effect of the imputation on the parameters of a simple linear regression model of the data.

Before any imputation the data was modeled using a simple linear regression technique using the model:

$$MPG = \beta_0 + \beta_1 \cdot CYLINDERS + \beta_2 \cdot SIZE\ HP + \beta_3 \cdot WEIGHT + \beta_4 \cdot ACCEL + \beta_5 \cdot ENG\_TYPE \quad (1)$$

```
PROC REG DATA = cars;
   MODEL MPG = CYLINDERS SIZE HP WEIGHT ACCEL ENG_TYPE;
RUN; quit;
```

After imputation the five imputed data sets were also subjected to the same model and then the parameter estimates were combined. The pre and post imputed parameter estimates were then compared. The figures below show some of the output from the linear regression.

Figure 1 shows the some of the fits diagnostics for the model on the pre-imputed data. As you can see the residuals follow a fairly normal distribution, histogram and qq-plot show this, and there are no abnormally high Cook's D values, although observation 10 stands out a bit. The model provides a very good fit to the data with an adjusted r^2 value of 0.883. However, there are a few small anomalies and an imputation of the data could smooth some of these out and provide more confidence in the parameter estimates.

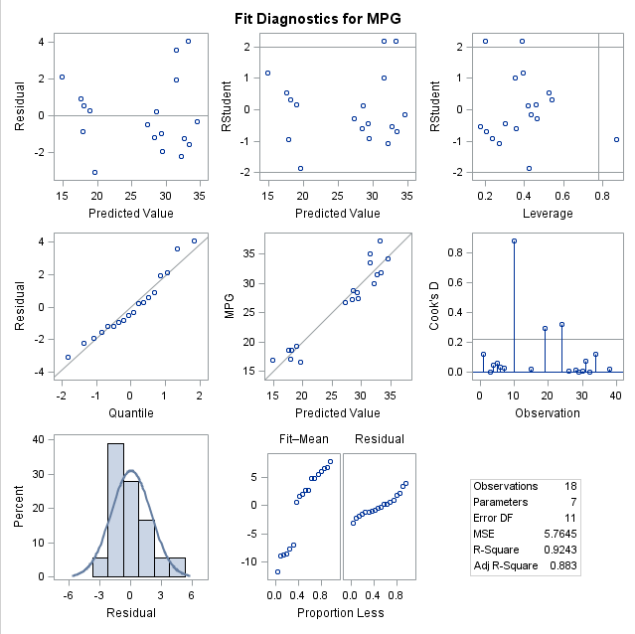

Figure 1: Fit Diagnostics for the Model above.

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	70.14772	8.03838	8.73	<.0001
CYLINDERS	1	-3.33403	1.56072	-2.14	0.0560
SIZE	1	0.02280	0.03207	0.71	0.4918
HP	1	-0.19546	0.08065	-2.42	0.0338
WEIGHT	1	-0.30623	5.13263	-0.06	0.9535
ACCEL	1	-0.78199	0.58264	-1.34	0.2066
ENG_TYPE	1	6.59880	3.59008	1.84	0.0932

Table 2: Parameter Estimate Table (pre-impute).

This first regression model of the pre-imputed data uses only 18 of the 38 rows because 20 of those rows have missing observations. Since it had such few data points for the full regression the stand errors are fairly high and it is hard to determine with the p values whether most of the variables play a significant role in the model. Basically, it is hard to know if this is a good model and if the model overfits the data.

After the data was imputed the five imputed sets were fit to the same model:

```
PROC REG data = miout outest = outreg covout;
     MODEL MPG = CYLINDERS SIZE HP WEIGHT ACCEL ENG_TYPE;
     by _Imputation_;
RUN; quit;
```

Now all 38 rows are used to fit the model to the data. Figure 2 and Table 3 show the fit diagnostics and parameter estimates for the first imputation set. The other imputed sets show similar but not identical results. You can clearly see there are more than twice as many modeled observations. In For most parameter estimates the standard error has been reduced by more than half their pre-imputed values, but the size and sign of the parameters estimates are roughly the same except for the WEIGHT variable which seems to have the least significance, largest p value, for the model.

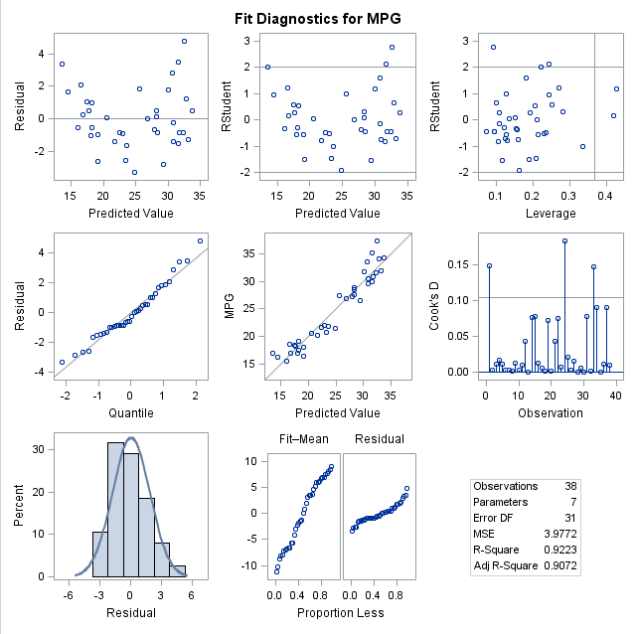

Figure 2: Fit Diagnostics for the Model above (post-impute).

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	70.89044	4.41548	16.05	<.0001
CYLINDERS	1	-2.94141	0.83384	-3.53	0.0013
SIZE	1	0.03638	0.01953	1.86	0.0720
HP	1	-0.18281	0.04168	-4.39	0.0001
WEIGHT	1	-3.03047	3.05690	-0.99	0.3292
ACCEL	1	-0.73881	0.29692	-2.49	0.0184
ENG_TYPE	1	5.98166	1.65253	3.62	0.0010

Table 3: Parameter Estimate Table (post-impute).

The parameter estimates from the five different imputed sets were combined:

```
PROC MIANALYZE data = outreg;
    MODELEFFECTS CYLINDERS SIZE HP WEIGHT ACCEL ENG_TYPE Intercept;
RUN; quit;
```

Variance Information (5 Imputations)							
	Variance						
Parameter	Between	Within	Total	DF	Relative Increase in Variance	Fraction Missing Information	Relative Efficiency
CYLINDERS	0.145978	0.559778	0.734953	70.411	0.312935	0.259098	0.950733
SIZE	0.000078874	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
Intercept	4.361865	19.108476	24.342713	86.515	0.273922	0.232561	0.955555

Table 4: Variance Information for the parameter estimates of all five imputation sets.

It is interesting to note that the between variance is substantially smaller than the within variance for the parameter estimates. This is an indication that the 5 imputed data sets are providing parameters estimates that are consistent with each other.

Parameter Estimates (5 Imputations)										
Parameter	Estimate	Std Error	95% Confidence Limits		DF	Minimum	Maximum	Theta0	t for H0: Parameter=Theta0	Pr > |t|
CYLINDERS	-2.384973	0.857294	-4.0946	-0.67533	70.411	-2.941410	-2.022784	0	-2.78	0.0069
SIZE	0.038598	0.021495	-0.0041	0.08127	95.316	0.030150	0.049214	0	1.80	0.0757
HP	-0.144995	0.049796	-0.2456	-0.04435	39.954	-0.182810	-0.117053	0	-2.91	0.0059
WEIGHT	-5.090662	3.173186	-11.3826	1.20123	104.93	-6.541305	-3.030466	0	-1.60	0.1117
ACCEL	-0.631053	0.307801	-1.2353	-0.02680	752.03	-0.738814	-0.558836	0	-2.05	0.0407
ENG_TYPE	4.578992	1.792868	0.9891	8.16887	57.205	3.849270	5.981657	0	2.55	0.0133
Intercept	68.388562	4.933833	58.5813	78.19587	86.515	66.131582	70.890441	0	13.86	<.0001

Table 5: Combined parameter estimates information.

The addition of more values in the data set allows the confidence in the parameter estimates to be higher. This is not surprising since the imputation method used was a regression technique. It makes sense that the confidence in the significance of the parameters would be better. This allows a sharper look at the model and suggests that a simpler model would be more appropriate. It was much more difficult to tell which parameter estimates were significant with more than half the data not being used in the regression model.

Comparing the pre and post imputed parameter estimates the change in the error is striking:

Parameter	**Pre-Imp**	Error	**Post-Imp**	Error
Intercept	70.14772	8.0384	68.38856	4.9338
CYLINDERS	-3.33403	1.5607	-2.38497	0.8573
SIZE	0.0228	0.0321	0.038598	0.0215
HP	-0.19546	0.0807	-0.145	0.0498
WEIGHT	-0.30623	5.1326	-5.09066	3.1732
ACCEL	-0.78199	0.5826	-0.63105	0.3078
ENG_TYPE	6.5988	3.5901	4.578992	1.7929

Table 6: Comparison of the parameter estimates for the pre and post imputed data. The post imputed values are a combination of the parameter estimates of all five different imputed data sets.

As stated above, imputation allows for a clearer look at the significance of the parameter estimates. In this case a model without SIZE and WEIGHT would seem to make more sense. These two variables have the largest p values. Applying this new model to the imputed data we see in Table 7 that all the parameter estimates are now statistically significant. This increases the confidence in the value of the parameter estimates and further reduces their error.

Parameter Estimates (5 Imputations)										
Parameter	Estimate	Std Error	95% Confidence Limits		DF	Minimum	Maximum	Theta0	t for H0: Parameter=Theta0	Pr > |t|
CYLINDERS	-2.172904	0.550870	-3.25385	-1.09196	1042.3	-2.358410	-2.005701	0	-3.94	<.0001
HP	-0.187218	0.028817	-0.24395	-0.13049	273.2	-0.198885	-0.177009	0	-6.50	<.0001
ACCEL	-0.940926	0.249268	-1.42953	-0.45233	13580	-0.970328	-0.907218	0	-3.77	0.0002
ENG_TYPE	5.881197	1.441269	3.04955	8.71284	503.4	5.467242	6.349363	0	4.08	<.0001
Intercept	68.091340	4.111516	60.03265	76.15003	36465	67.530927	68.609424	0	16.56	<.0001

Table 7: Combined parameter estimates for a model without SIZE and WEIGHT.

4 Conclusion

The missing values in the data set *carmpg* were imputed using multiple imputation with the FCS statement which uses a regression method to impute the continuous variables and a discriminant function method to impute the classification variables. The missing data appears to exhibit a missing at random pattern. Using a simple linear regression model, the imputed data shows a reduced error for the parameter estimates compared to modeling the data with only the rows with no missing data.

Multiple-imputation also made it easier to note the more significant variables for the linear regression model. This made trimming the model to a simpler version helped to reduce overfitting and reduce the error in the parameter estimates even more.

A Code

```
/* Initial analysis of non-imputed data, only 18 of 38 rows were used */
PROC REG DATA = cars;
     MODEL MPG = CYLINDERS SIZE HP WEIGHT ACCEL ENG_TYPE;
RUN; quit;

/* Use multi-imputation to create 5 data sets with missing values imputed and
output to miout*/
PROC MI DATA = cars NIMPUTE = 5
     OUT = MIOUT seed = 35399;
     VAR MPG CYLINDERS SIZE HP WEIGHT ACCEL ENG_TYPE;
       class CYLINDERS ENG_TYPE;
       FCS;
RUN; quit;


/* Analyze all 5 imputed data sets and output to outreg */
PROC REG data = miout outest = outreg covout;
     MODEL MPG = CYLINDERS SIZE HP WEIGHT ACCEL ENG_TYPE;
     by _Imputation_;
RUN; quit;
```

```
/* Combine parameter estimates from all 5 sets from proc reg using output
outreg */
PROC MIANALYZE data = outreg;
     MODELEFFECTS CYLINDERS SIZE HP WEIGHT ACCEL ENG_TYPE Intercept;
RUN; quit;


/* Analyze all 5 imputed data sets and output to outreg2 */
PROC REG data = miout outest = outreg2 covout;
     MODEL MPG = CYLINDERS HP ACCEL ENG_TYPE;
   by _Imputation_;
RUN; quit;

proc print data = outreg2; run;


/* Combine parameter estimates from all 5 sets from proc reg using output
outreg2 */
PROC MIANALYZE data = outreg2;
    MODELEFFECTS CYLINDERS HP ACCEL ENG_TYPE Intercept;
RUN; quit;
```