Case Study: Miles per Gallon Estimate

Introduction

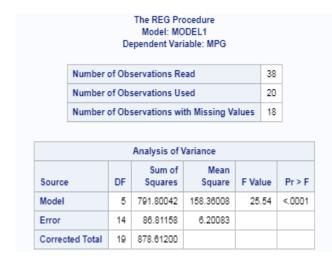
For this case study, we will be looking at the mileage data for 38 cars that were measured in 2005 to estimate miles per gallon. The variables for the dataset include Cylinders (number of cylinders), Size (engine displacement), HP (horsepower), and weight of the car (car weight). However, this dataset is missing values on numerous variables of the 38 records that it contains. In order to acquire a more complete dataset and therefore a more powerful analysis in theory, we will be using multiple imputations to explore this study.

Literature review

From the initial information received on this dataset and the videos on 2ds, we know that this dataset is incomplete and will need figure out the best method to analyze the data with this in mind. This is a scenario that will likely come up on numerous occasions throughout the career of a Data Scientist and one must now how to combat this issue. Multiple imputations will likely be the technique used to properly analyze this information but we will need a comparison to single imputation to verify which method has more power.

Method

First we will run a linear regression of the data in its current state using PROC REG in our SAS code; by default this uses list-wise deletion.



| Parameter Estimates | | | | | | | | | | | |
|---------------------|----|-----------------------|-------------------|---------|---------|--|--|--|--|--|--|
| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > t | | | | | | |
| Intercept | 1 | 67.61816 | 7.12819 | 9.49 | <.0001 | | | | | | |
| CYLINDERS | 1 | -1.19508 | 1.13851 | -1.05 | 0.3116 | | | | | | |
| SIZE | 1 | 0.05221 | 0.02938 | 1.78 | 0.0973 | | | | | | |
| HP | 1 | -0.15009 | 0.07848 | -1.91 | 0.0765 | | | | | | |
| WEIGHT | 1 | -6.71776 | 3.98252 | -1.69 | 0.1138 | | | | | | |
| ACCEL | 1 | -0.68451 | 0.44024 | -1.55 | 0.1423 | | | | | | |

From the image above, we can see that 38 observations were read but only 20 were use due to list-wise deletion and there only 19 degrees of freedom meaning that our analysis has lower power than expected. Next we will attempt imputation in order provide a more complete analysis of the dataset for cars.

The first step in looking at the data would be discover any missing value patterns using the SAS command PROC MI and then deciding which MI option to use.

| Group | MPG | CYLINDERS | SIZE | HP | WEIGHT | ACCEL | Freq | Percent |
|-------|-----|-----------|------|----|--------|-------|------|---------|
| 1 | Х | X | Х | Х | Х | Х | 20 | 52.63 |
| 2 | Х | Х | Х | Х | Х | | 2 | 5.26 |
| 3 | Х | Х | Х | Х | | Х | 3 | 7.89 |
| 4 | Х | Х | Х | Х | - | - | 1 | 2.63 |
| 5 | Х | Х | Х | - | Х | Х | 5 | 13.16 |
| 6 | Х | Х | | Х | Х | Х | 2 | 5.26 |
| 7 | Х | Х | | Х | | Х | 1 | 2.63 |
| 8 | Х | | Х | Х | Х | Х | 2 | 5.26 |
| 9 | Х | | Х | Х | Х | | 1 | 2.63 |
| 10 | Х | | Х | Х | | Х | 1 | 2.63 |

From the image above, we can determine that pattern look likes it is non- monotone being that the values seem to missing randomly within the dataset provided. Now with this information we can use MCMC on the data to proceed with using multiple imputations due to its arbitrary nature.

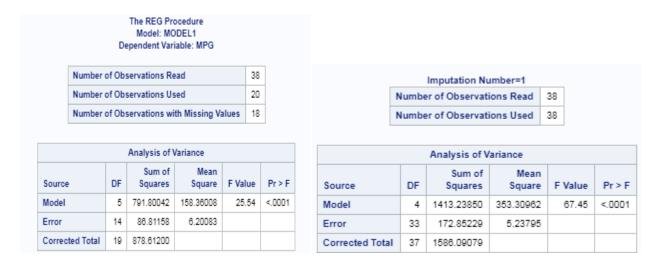
| Model Information | | | | | | | | |
|----------------------------------|-------------------|--|--|--|--|--|--|--|
| Data Set | WORK.CARMPG | | | | | | | |
| Method | MCMC | | | | | | | |
| Multiple Imputation Chain | Single Chain | | | | | | | |
| Initial Estimates for MCMC | EM Posterior Mode | | | | | | | |
| Start | Starting Value | | | | | | | |
| Prior | Jeffreys | | | | | | | |
| Number of Imputations | 25 | | | | | | | |
| Number of Burn-in Iterations | 200 | | | | | | | |
| Number of Iterations | 100 | | | | | | | |
| Seed for random number generator | 3599 | | | | | | | |

After running SAS code to create the imputation data, we can see that the MCMC method was used with a single imputation chain. Also, that the number of imputations is 25 meaning that there were 25 different datasets created from this imputation.

Now, that we have our imputed data we can run a regression analysis on each of the 25 full created datasets in order to estimate miles per gallon.

Results

From imputation #1 below, we can see that all 38 cars are now included within the regression and 37 degrees of freedom are being used meaning that this observation has more power than our initial evaluation which only included only 19 degrees of freedom



Next we will combine the results of all 25 imputations using PROC MIANALYZE for a single analysis

| | | | | | The MIA | NALYZE | Procedu | re | | | | | | |
|---------------------------------------|-----------|-------------------|----------|----------------|--------------|------------|-----------|-------------------------|----------------|-----------|-------|-------------------------------|-------|---------|
| | | Model Information | | | | | | | | | | | | |
| | | | Data Set | | | | WORK. | OUTR | EG | | | | | |
| | | | | Number | of Imp | utations | 25 | | | | | | | |
| | | | | | | | | | | | | | | |
| Variance Information (25 Imputations) | | | | | | | | | | | | | | |
| | | | Variance | | | | | Relative | | | | | | |
| | Paramete | er I | Between | een Within | | Total | DF | Increase in Variance | | | | Relative Efficiency | | |
| | CYLINDE | RS (| .061197 | 0.5668 | 72 0. | 2 0.630517 | | 0.1 | 0.112273 | | 01703 | 0.995948 | | |
| | SIZE | 0.00 | 0085537 | 0.0004 | 18 0. | 000507 | 779.12 | 0.2 | 0.212871 | | 77619 | 0.992945 | | |
| | HP | | 0.000379 | | 12 0. | 002408 | 895.13 | 0.195804 | | 0.165605 | | 0.993419 | | |
| | WEIGHT | 2 | .189402 | 8.2028 | 32 10. | 479840 | 508.4 | 0.277583 | | 0.220333 | | 0.991264 | | |
| | ACCEL | 0 | .026085 | 0.1087 | 0.108795 0.1 | | 602.51 | 0.249349 | | 0.202227 | | 0.991976 | | |
| | Intercept | rcept 3. | | 24.816680 28.3 | | 352982 | 1542.8 | 0.142497 | | 0.125857 | | 0.994991 | | |
| | | | | | | | | | | | | | | |
| | | | | Parar | neter Es | timates (| 25 Imput | ations |) | | | | | |
| Parameter | Estimate | Std Error | 95% C | 95% Confidence | | DF | Minir | Minimum Maxim | | um Theta0 | | t for H0: Parameter=Theta0 | | Pr > t |
| CYLINDERS | -1.533464 | 0.794051 | -3.0 | 906 | 0.02365 | 2355.5 | -1.90 | 4471 | -0.671 | 335 | 0 | | -1.93 | 0.0536 |
| SIZE | 0.055369 | 0.022514 | 0.0 | 112 | 0.09956 | 779.12 | 0.03 | 0.033092 0.06 | | 195 | 0 | | 2.46 | 0.0141 |
| HP | -0.108087 | 0.049050 | -0.2 | 044 - | 0.01182 | 895.13 | -0.145143 | | 3 -0.074101 | | 0 | -2.20 | | 0.0278 |
| WEIGHT | -8.246574 | 3.237258 | -14.6 | 066 - | 1.88652 | 508.4 | -10.19 | 0276 -4.864 | | 179 0 | | -2.55 | | 0.0111 |
| ACCEL | -0.684092 | 0.368678 | -1.4 | D81 (| 0.03996 | 602.51 | -1.063360 | | 3360 -0.355830 | | 0 | | -1.86 | 0.0640 |
| Intercept | 68.052123 | 5.324752 | 57.6 | 076 7 | 3.49864 | 1542.8 | 64.05 | 9281 71.562043 | | 043 | 0 | | 12.78 | <.0001 |

While we do not expect the combined estimates to be similar to the original estimate, we do have confidence they are a better representation of the estimates for our parameters due reduced p values.

Conclusion

In conclusion, using multiple imputations allowed us to provide analysis that represents the uncertainty of the missing value within the cars data. In theory, using this method as opposed to single imputation yields valid stats based inferences that reflect uncertainty of absent data.