Car MPG Case Study

**Amit Patel, Sanjay Jagarlamudi, Murali Parthasarathy**

*MSDS 7333 Quantifying the World, May-20-2017*

**ABSTRACT**

Environment and consumer awareness are driving car OEMs to shift focus on increasing the miles per gallon offered by the vehicle. There are several factors, which affect the MPG of a vehicle. The data collected is from the various sub systems of the car and have a lot of missing values. The missing values are caused by the devices, which transfer the data from the car into a human consumable format. The Paper studies the various options available for data imputation to increase the effectiveness of the MPG data analysis.

**Introduction**

The vehicle mpg is highly dependent on the built of the vehicle and the driving style. Vehicles mpg is drastically reduced in city drives where the driver has to slower speed limits and multiple signal lights. The driver has to accelerate to gain momentum after every slow down or signal light. The MPGs in city drives are affected by driving styles and vehicle attributes. The data set provided has a lot of missing values, which affect the overall statistical capability of the data set. The objective of the study is to find affective ways of data imputation.

**Data Analysis**

|  |  |  |
| --- | --- | --- |
| **Attribute Name** | **Attribute Type** | **Characteristics** |
| Mpg | Continuous | Miles Per gallon. This is the reading from the car during the drive |
| Cylinders | Categorical - Ordinal | The number of cylinders in the car determines the mpg |
| Hp | Continuous | Horse Power. The horse power generated by the engine |
| Weight | Continuous | Weight of the car. A heavier car will need more displacement to gain momentum |
| Acceleration | Continuous | Acceleration is a driving style. It depends on the pressure exerted by the driver on the accelerator gear |
| Eng\_type | Categorical - Ordinal | The Engine type, the type of engine could be a V4 or V6 |

Table 1: Data attributes

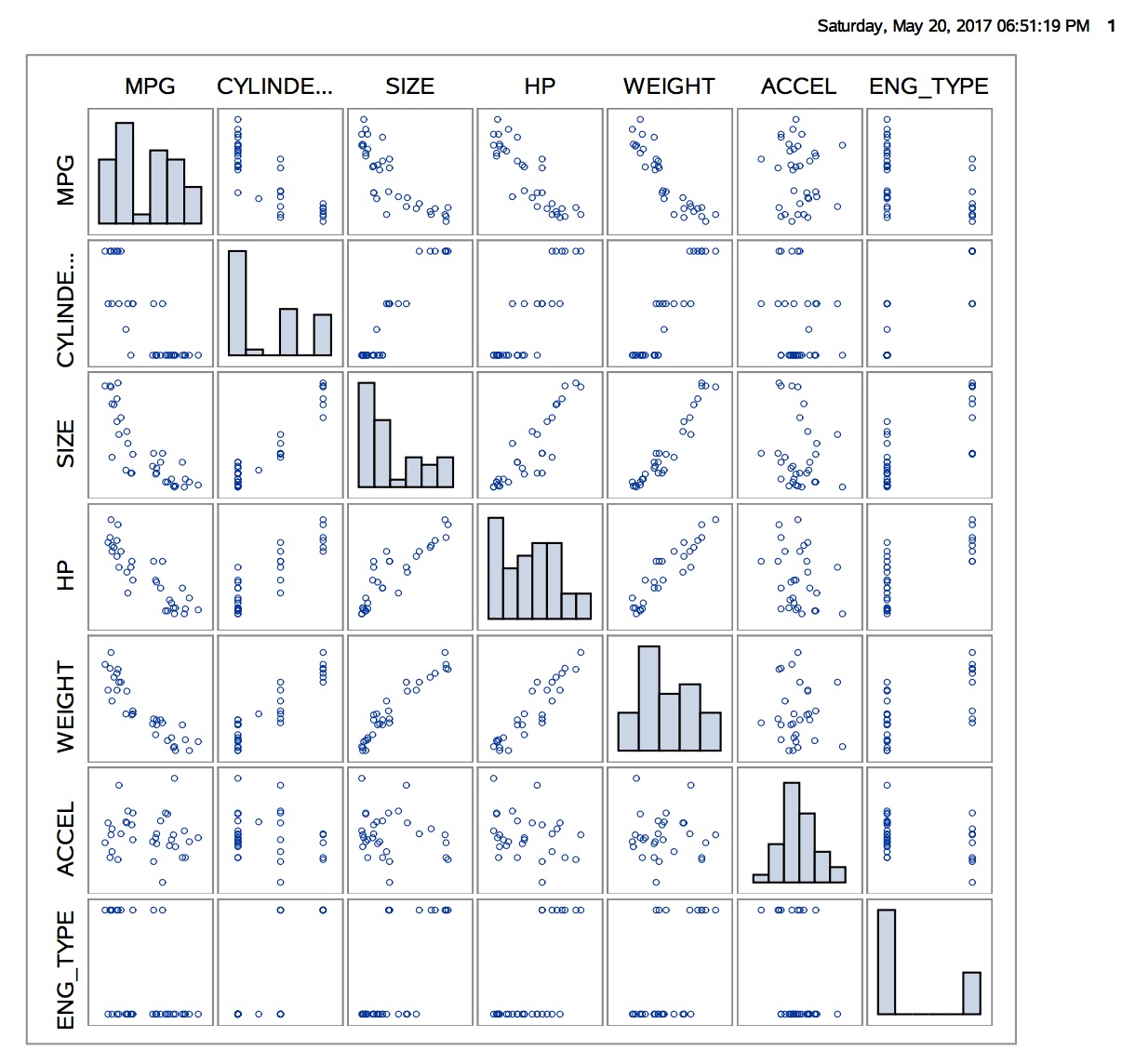


Figure : Scatter Plot

The section below comments a few relationships

|  |  |
| --- | --- |
| **Relationship** | **Comments** |
| Mpg-weight | The mpg decreases as the weight of the vehicle increases. |
| Hp-mpg | The mpg decreases as the vehicle offers higher Horsepower |
| Weight-hp | The horse power increases as the weight of the vehicle increases |

Table 2: Data correlation

**Missing Values**

The data set has a lot of missing values. A cursory inspection suggests that weight, HP and engine type are missing on some data records. The missing values display a monotonous pattern; the figure below provides insight to the missing value monotonous pattern.

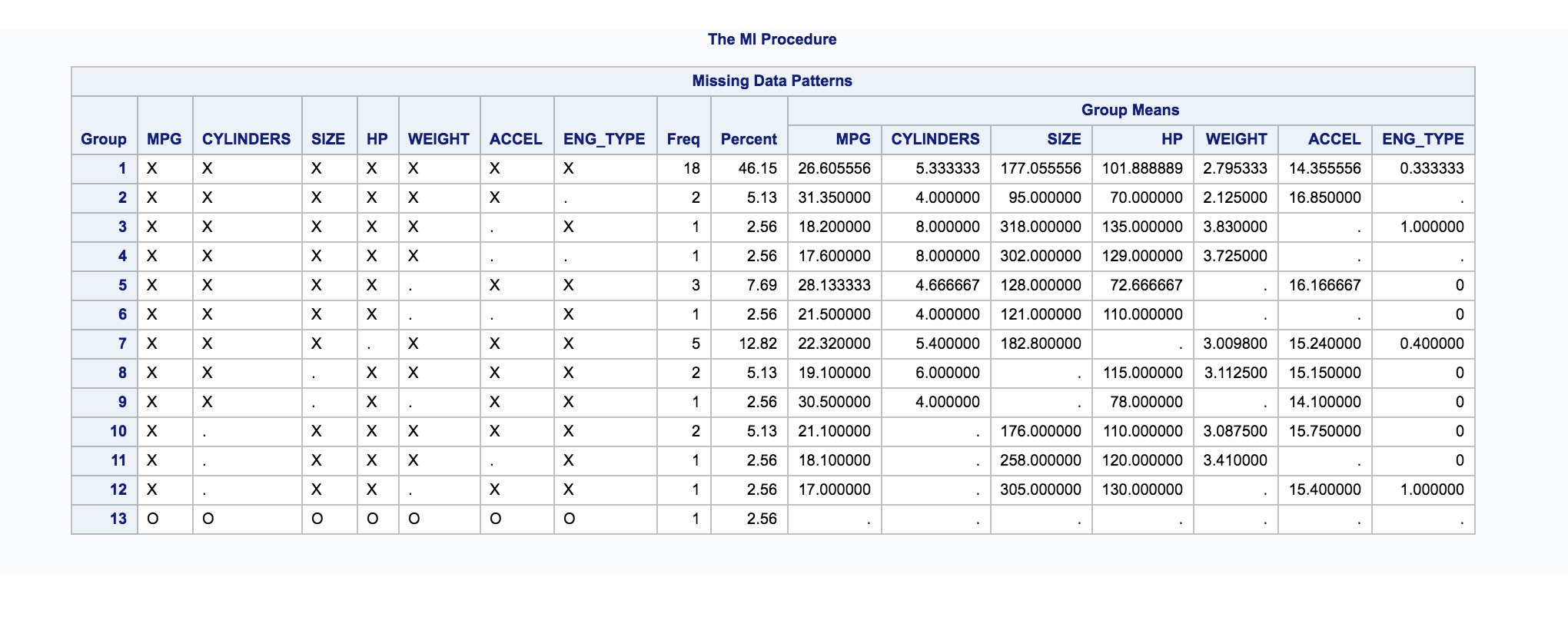


Figure : Missing attribute Pattern

**Missing data pattern**

There are four data missing mechanisms

Missingness completely at random - The probability of missingness is the same for all variables

Missingness at random - The probability of missingness is not the same for all variables

Missingness depends on observed predictors –The missingness is not at random and depends on the observed values

The data set shows a patterns of missing at random, the probability of missingness is not the same for all the variables.

|  |  |  |
| --- | --- | --- |
| Pattern of Missingness | Type of Imputed Value | Recommended Algorithm |
| Missingness at random (MAR) | Continuous | MCMC |
|  | Categorical | MCMC |

Also through observation we are assuming the data is multivariate normal distribution.

**Imputation Methods**

The imputation methods can be categorized as

* Simple imputation
* Multiple imputations

The methods can be further categorized as approaches, which keep all the data, and approaches will remove records with missing data.

**Simple imputation**

Complete case Analysis

In the method the records with missing data is completely ignored. This method will have bias as certain good responses with missing values will be ignored

The other options available are

Last Value Carry Forward – Use the last observed value. This approach will not be suitable for the car data set as the attributes are highly driven by the driving style. This method will introduce a lot of bias to the data set

Hot Deck imputation – This method uses the nearest neighbor concept where the closest value is substituted. Again this approach may not work as the missing attributes are dependent on the driving style.

Single Imputation Methods

The dataset is highly co related but there are certain attributes, which will qualify for simple imputation methods and will not affect the overall predictability of the model.

Single imputation approaches can be done for engine type and weight. The most likely value can be determined from the data available. The approach is not comprehensive and cannot be employed to all the missing attributes.

**Multiple Imputation Approach**

The multiple imputation technique determines the missing values by taking into account the available data sets have to be employed. The multiple imputation technique takes into account the uncertainty in the data set due to the missing values. The multiple imputation technique is easy to employ.

The only disadvantage of the imputation model is the team now has to consider both the analysis model and imputation model.

**Multiple data sets approach**

There are two approaches they are

* Maximum likelihood estimation
* Multiple imputations

**Maximum likelihood estimation**

The maximum likelihood estimation does not create multiple data sets like the multiple imputation method. The method estimates the missing value based on the available attributes. This approach fits linear models very well. The approach estimates the values based on a range of values.

**Regression Analysis (List wise/Complete)**

The regression analysis on the data set shows only 17 out of the 39 data records are used. **The p-values of the coefficients are greater than 0.05, which reduces the power of our model to estimate Mpg.** To increase the power of the model, we will add observations with missing values to see if it improves our estimates.

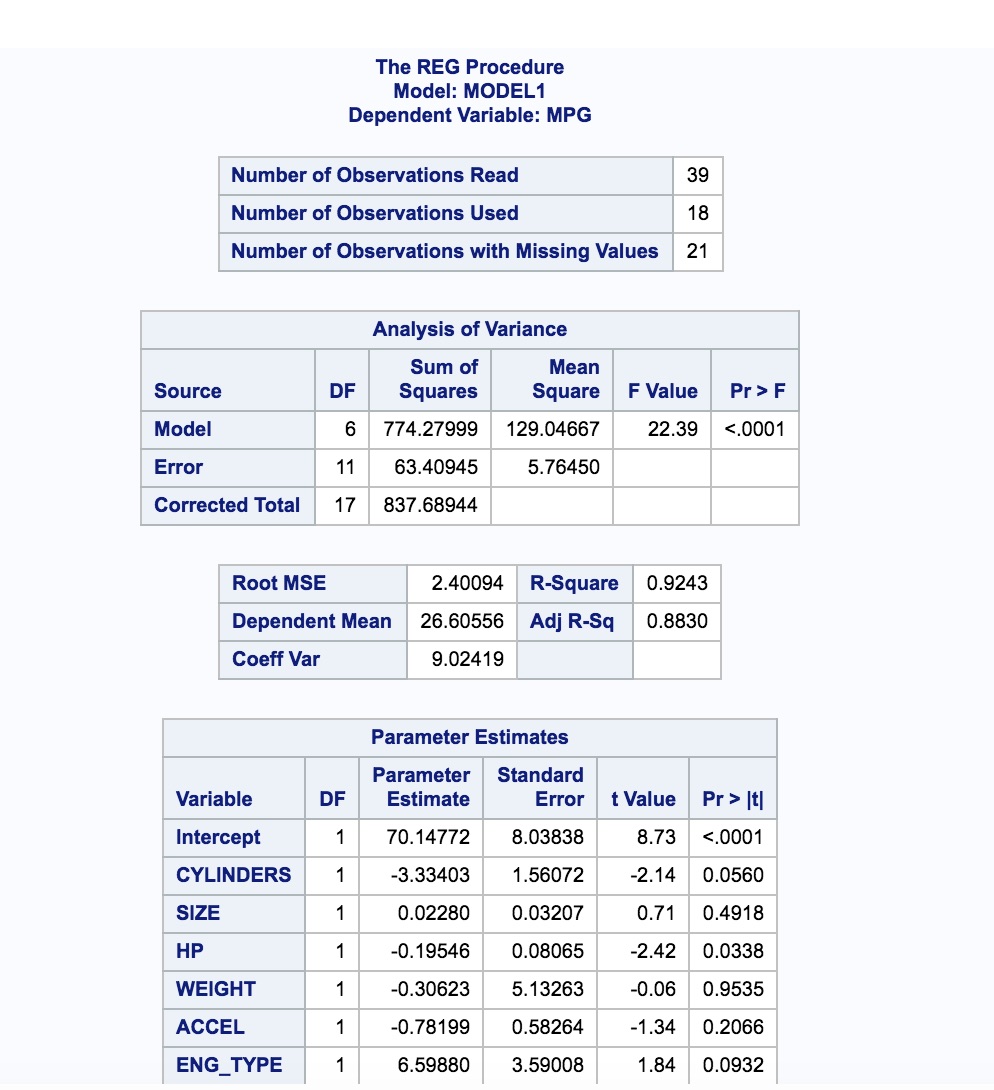


Figure Linear Regression

**Multiple Imputation Details**

## Step 1 – Create the Data Sets

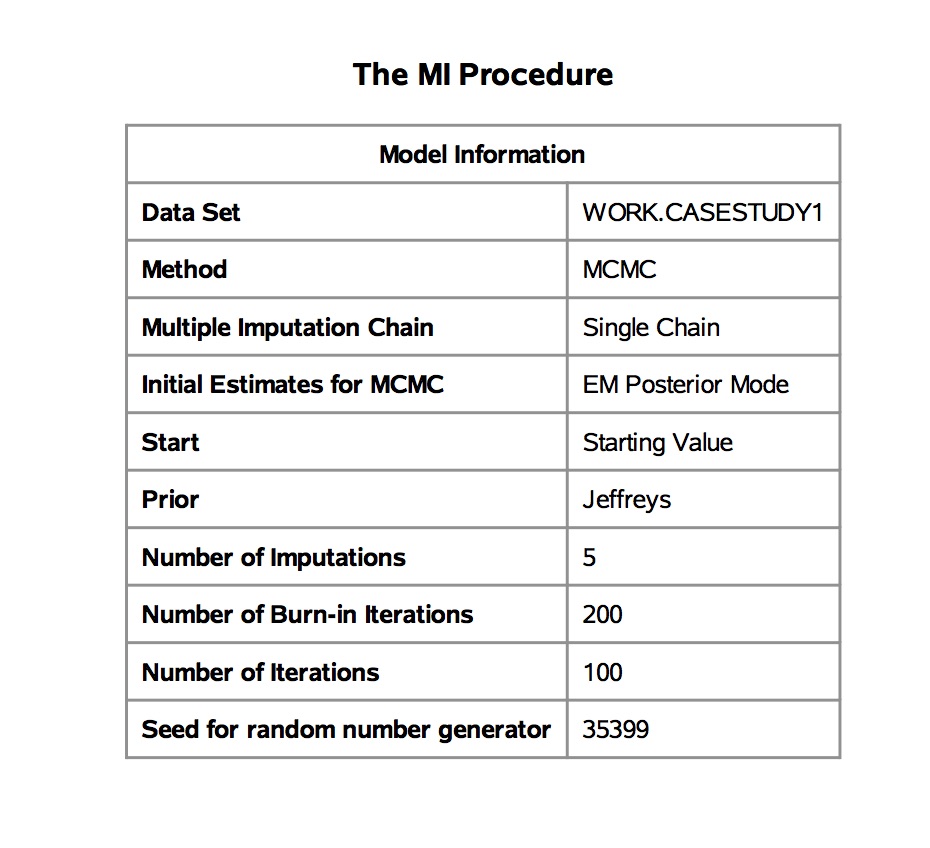


Figure Multiple Imputation Procedures

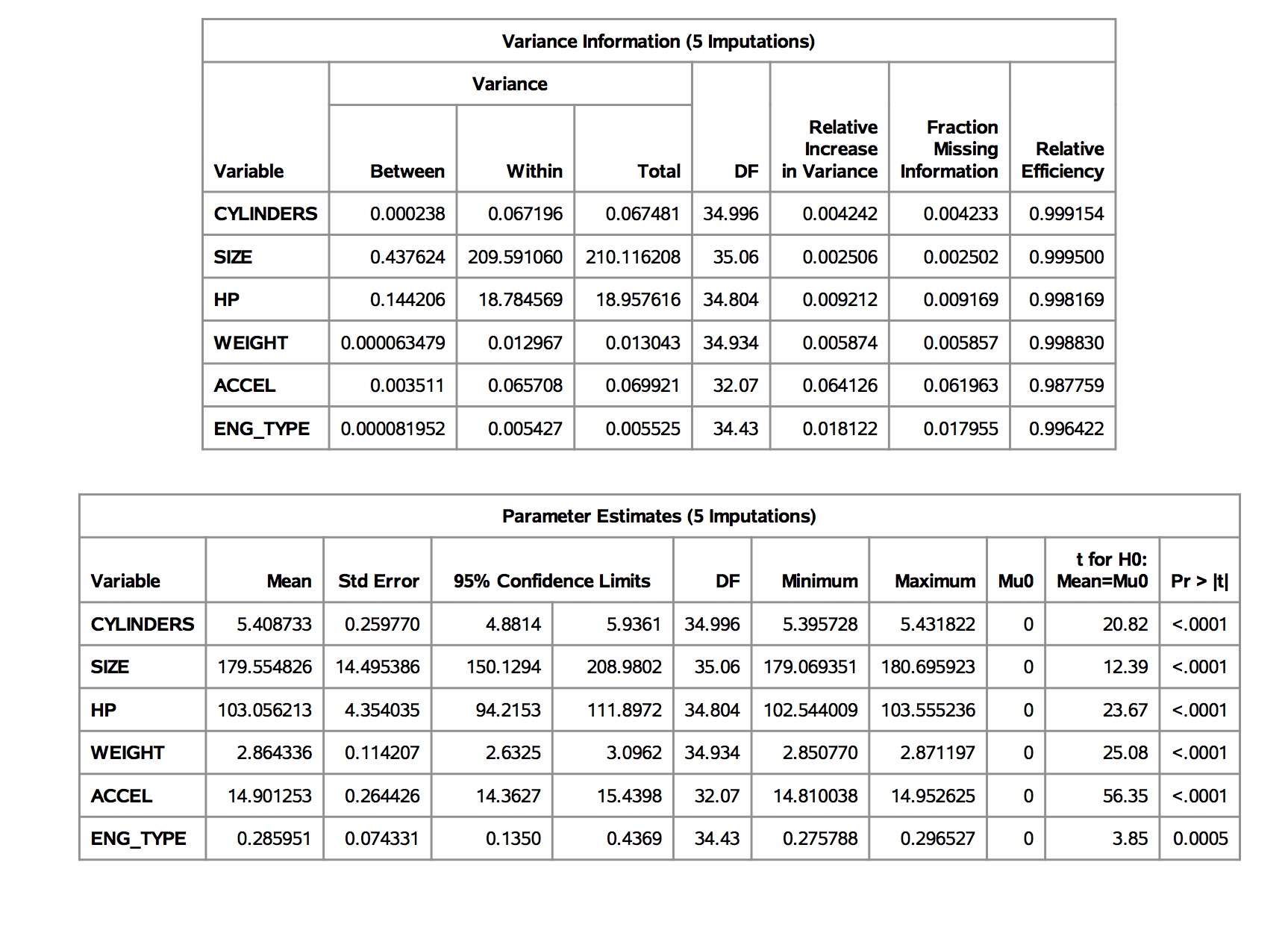


Figure Parameter and Variance Estimation on Imputed value

## Step 2 - Analyze the imputed data set

In this step a regression analysis is done on the imputed data sets.

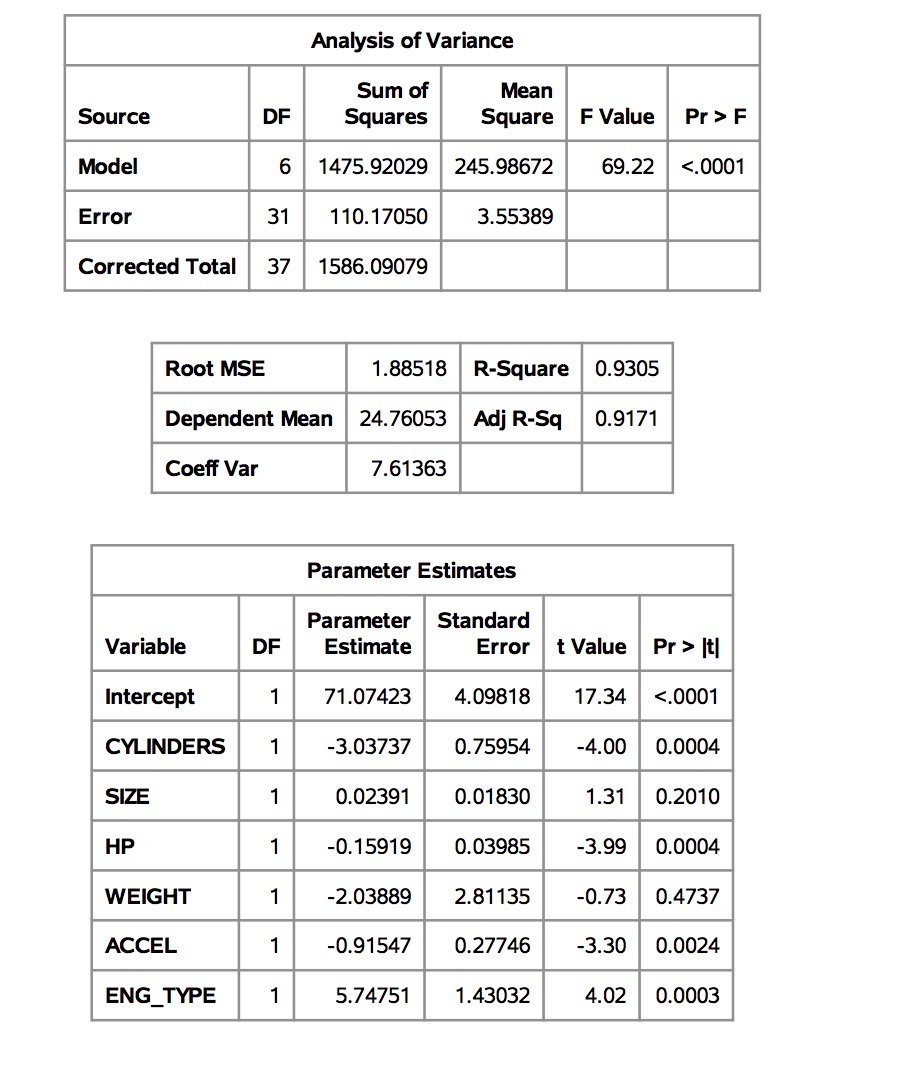


Figure Imputation 1

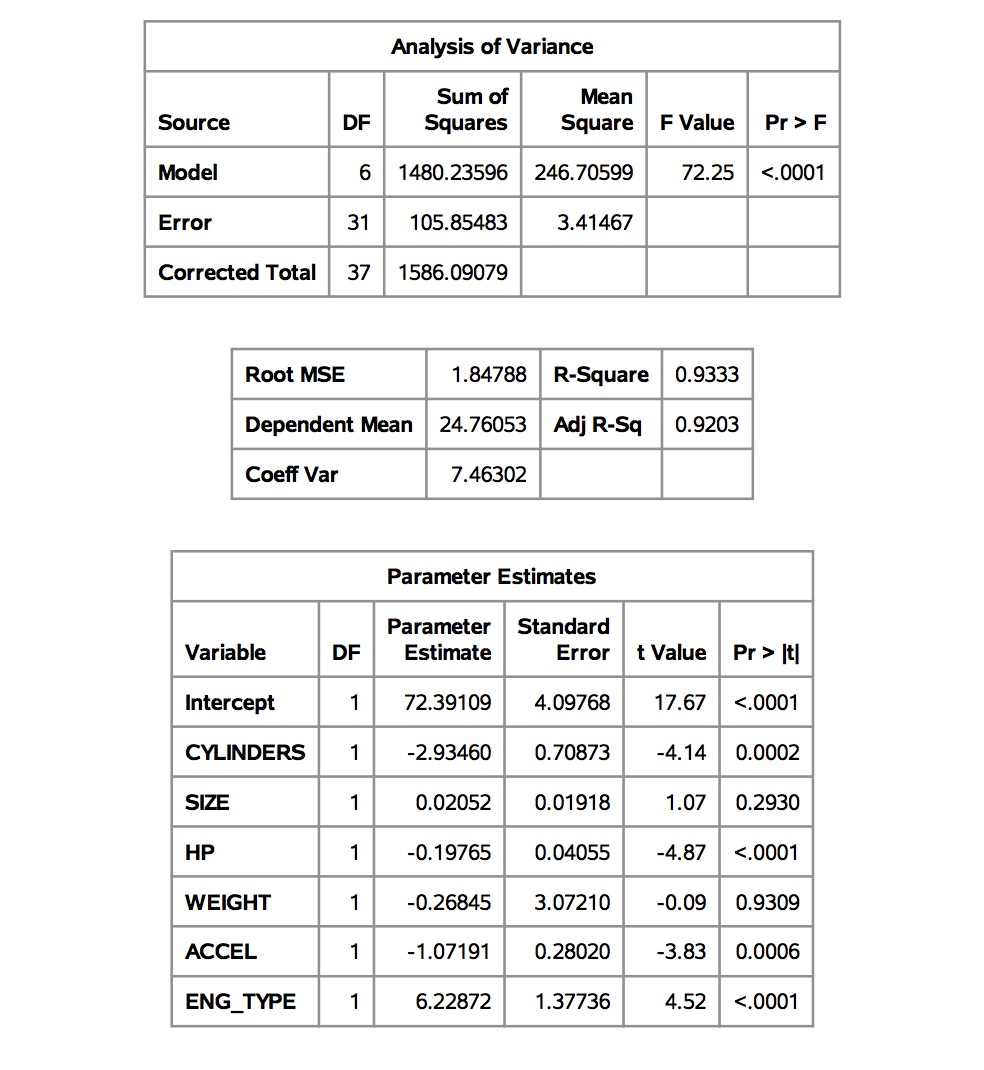


Figure Imputation 2

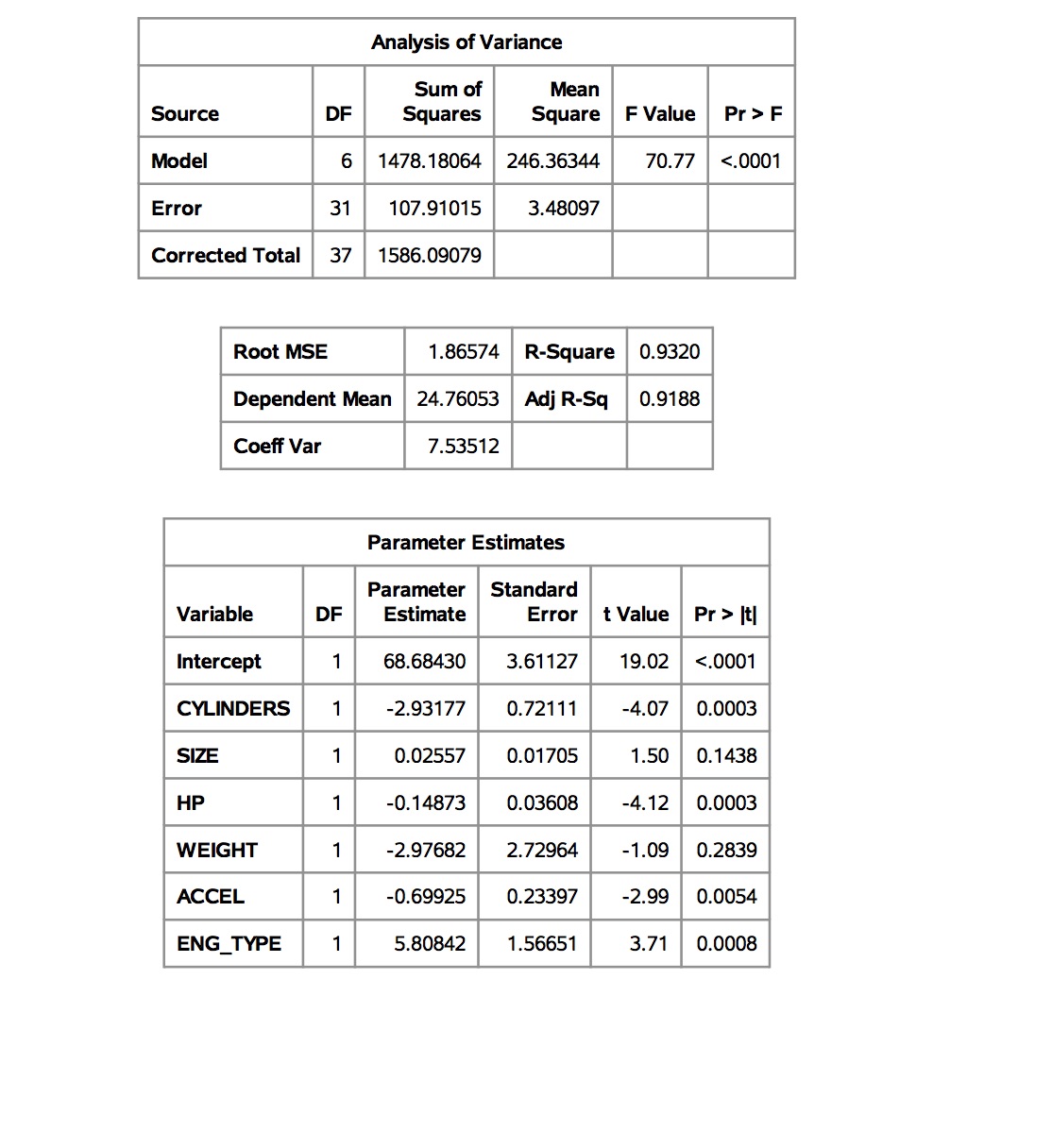


Figure Imputation 3

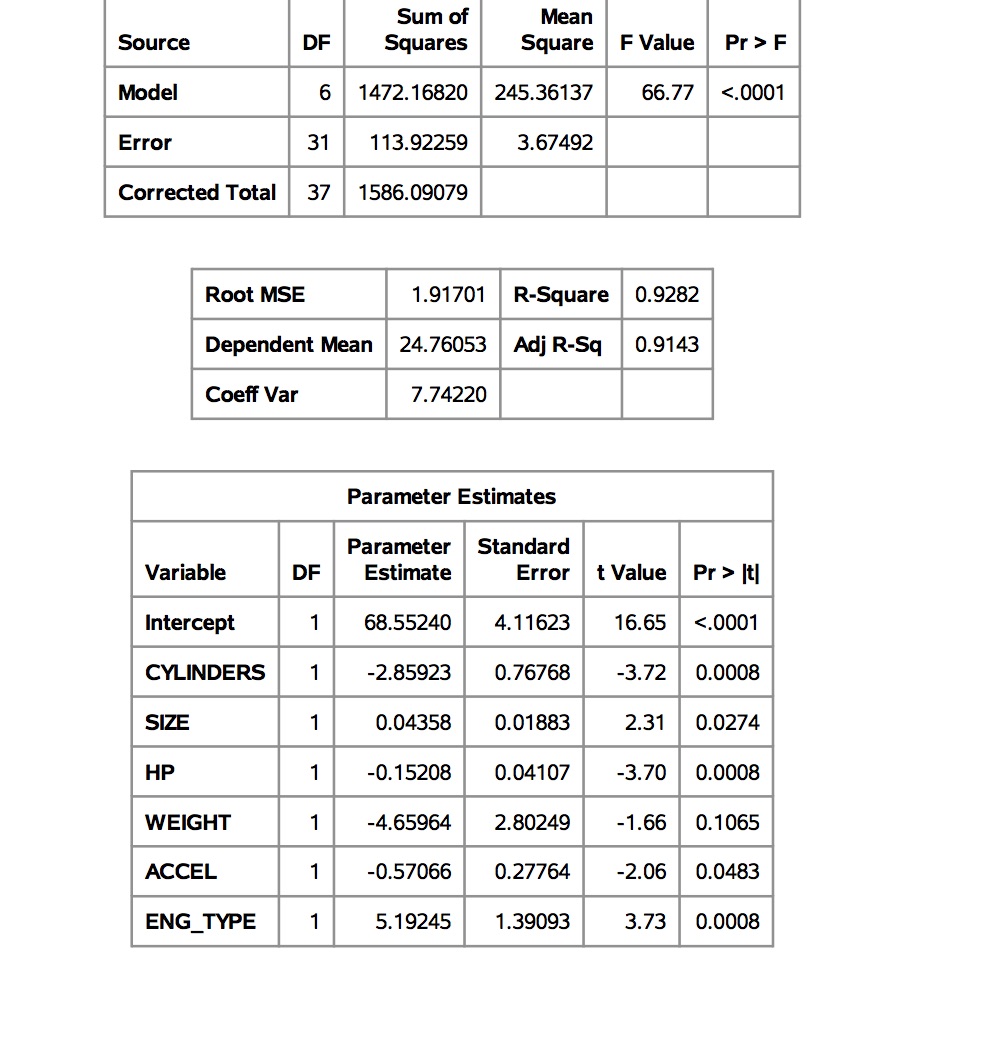


Figure Imputation 4

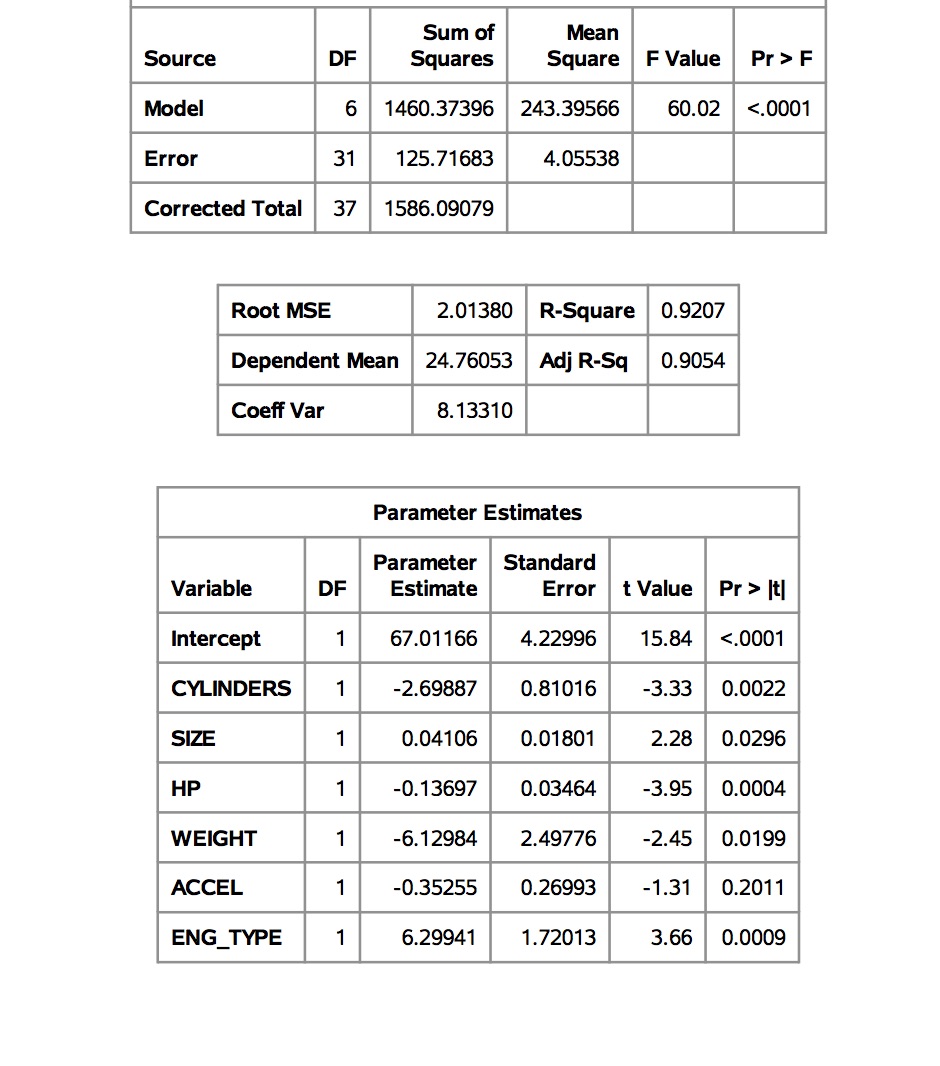


Figure Imputation 5

## Step 3 – Combine Analysis Results

This step averages the results from the multiple analyses.

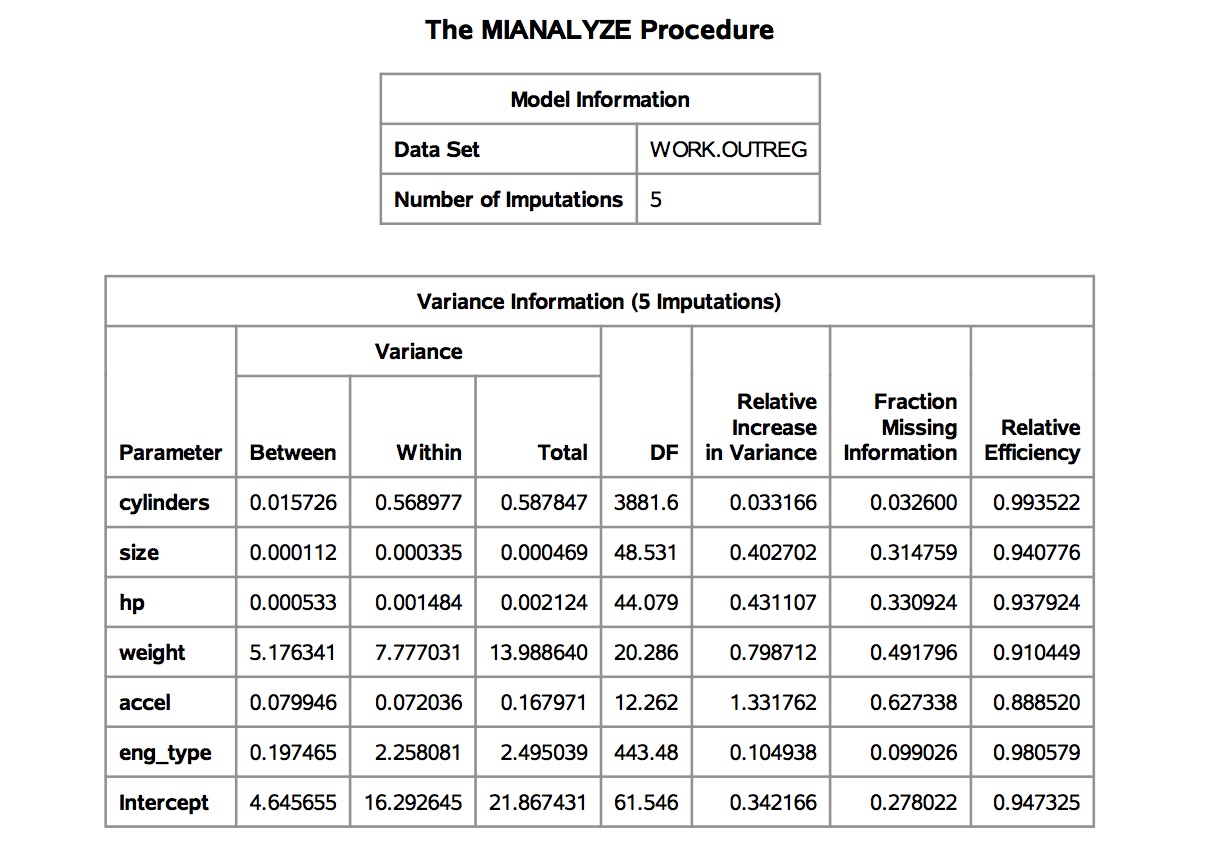


Figure Combined Analysis

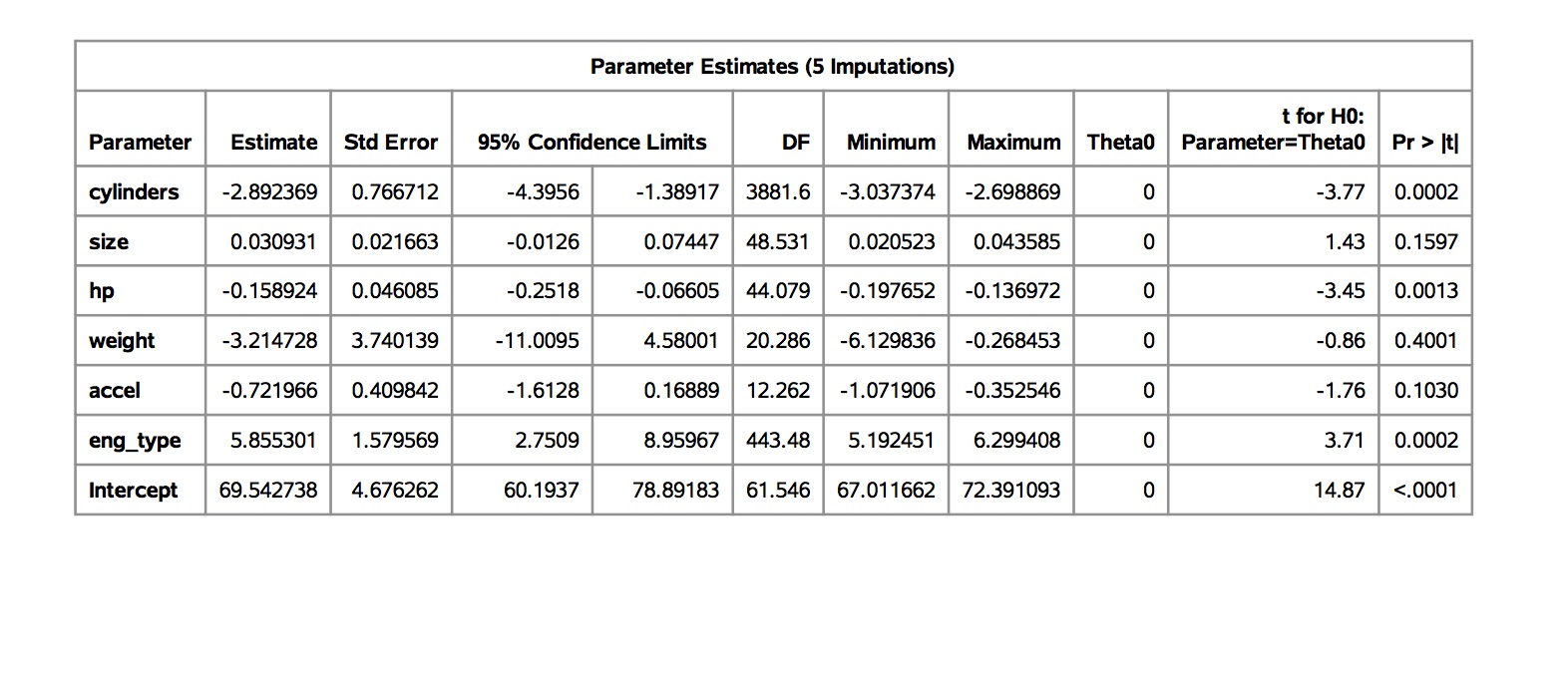


Figure Combined Analysis

**CONCLUSION**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Original** | **Original** | **Paramter Estimates using 5 datasets** |  |
|  | **Estimate** | **Std Error** | **Combined Estimate** | **Combined Std Error** |
|  |  |  |  |  |
| Intercept | 70.14772 | 8.03838 | 69.542738 | 4.676262 |
| cylinders | -3.33403 | 1.56072 | -2.892369 | 0.766712 |
| size | 0.0228 | 0.03207 | 0.030931 | 0.021663 |
| hp | -0.19546 | 0.08065 | -0.158924 | 0.046085 |
| weight | -0.30623 | 5.13263 | -3.214728 | 3.740139 |
| accel | -0.78199 | 0.58264 | -0.721966 | 0.409842 |
| eng\_type | 6.5988 | 3.59008 | 5.855301 | 1.579569 |

Figure Combined Results

In all cases, the coefficients of the combined analysis estimates gave us more precise (with less standard error) and with better p-values (< .05). The multiple imputation method MCMC greatly improved the estimates and also combining the normally distributed imputations gave us better results.

**REFERENCES**

Allison P. (2012). Why You Probably Need More Imputations Than You Think. <http://www.statisticalhorizons.com/more-imputations>. Retrieved 11/10/2013.

<http://home.cerge-ei.cz/mittag/papers/Imputations.pdf>

<https://support.sas.com/resources/papers/proceedings11/351-2011.pdf>

<http://www.stat.columbia.edu/~gelman/arm/missing.pdf>

<https://communities.sas.com/t5/SAS-Statistical-Procedures/Maximum-Likelihood-Estimation-or-OLS-PROC-REG-vs-PROC-MIXED/td-p/197429>

<http://www.biostat.umn.edu/~john-c/5421/notes.019>

<https://statisticalhorizons.com/ml-better-than-mi>

<http://www.theanalysisfactor.com/missing-data-two-recommended-solutions/>

<http://www2.sas.com/proceedings/sugi30/111-30.pdf>

http://home.cerge-ei.cz/mittag/papers/Imputations.pdf