APM 630 Regression Analysis
HW3 – Multiple Linear Regression
Name: Ruta Basijokaite

Project Report

1. Introduction

A study was conducted to investigate the relationship between car gasoline mileage (MPG) and the features of the car. The variables in the study included:

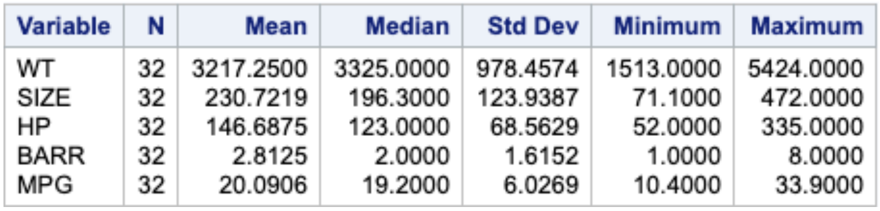
Y = car gasoline mileage (MPG)
X_1 = car weight in pounds (WT)
X_2 = car engine power rating in cubic inches (SIZE)
X_3 = car engine horse power (HP)
X_4 = the number of barrels in carburetor (BARR)

The purpose of the study is to (1) analyze descriptive statistics of all variables and compute their correlations, (2) fit a multiple linear regression model with all independent variables involved (full model), (3) find the "best" model (reduced model) using stepwise model selection method, and (4) compare the characteristics of the "best" model against the full model.

2. Methods

A random sample of cars (n=32) was selected. Data obtained from each car included one dependent variable (Y) and four independent variables ($X_1 – X_4$) as defined above. The descriptive statistics of all variables are listed in Table 1.

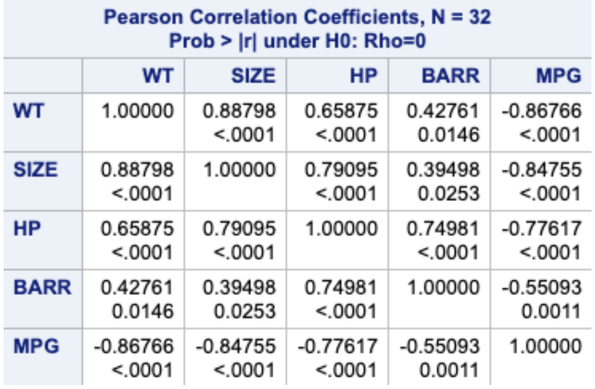
Table 1. Descriptive statistics of all variables. Answers contain four significant digits.



Variable	N	Mean	Median	Std Dev	Minimum	Maximum
WT	32	3217.2500	3325.0000	978.4574	1513.0000	5424.0000
SIZE	32	230.7219	196.3000	123.9387	71.1000	472.0000
HP	32	146.6875	123.0000	68.5629	52.0000	335.0000
BARR	32	2.8125	2.0000	1.6152	1.0000	8.0000
MPG	32	20.0906	19.2000	6.0269	10.4000	33.9000

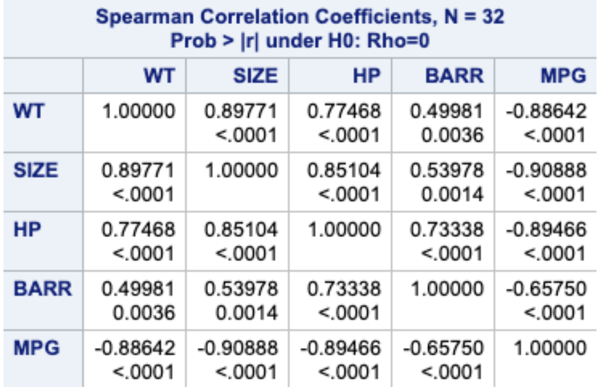
Table 2 represents Pearson and Spearman correlation coefficients between Y and each X, as well as among the four X variables. Pearson correlation coefficients indicate that Y (MPG) has significant (p-value < 0.0001) negative correlations with three predictive variables X_1 (WT), X_2 (SIZE) and X_3 (HP). Although fourth variable variable (BARR) has higher p-value, it is still significant according to α=0.05 and is negatively correlated with dependent variable. Meanwhile, Spearman correlation coefficients indicate that all predictive variables have significant (p-value < 0.0001) negative correlations with Y.

Table 2. Pearson and Spearman correlation coefficients among all variables in the study.



Pearson Correlation Coefficients, N = 32
Prob > |r| under H0: Rho=0

	WT	SIZE	HP	BARR	MPG
WT	1.00000	0.88798 <.0001	0.65875 <.0001	0.42761 0.0146	-0.86766 <.0001
SIZE	0.88798 <.0001	1.00000	0.79095 <.0001	0.39498 0.0253	-0.84755 <.0001
HP	0.65875 <.0001	0.79095 <.0001	1.00000	0.74981 <.0001	-0.77617 <.0001
BARR	0.42761 0.0146	0.39498 0.0253	0.74981 <.0001	1.00000	-0.55093 0.0011
MPG	-0.86766 <.0001	-0.84755 <.0001	-0.77617 <.0001	-0.55093 0.0011	1.00000

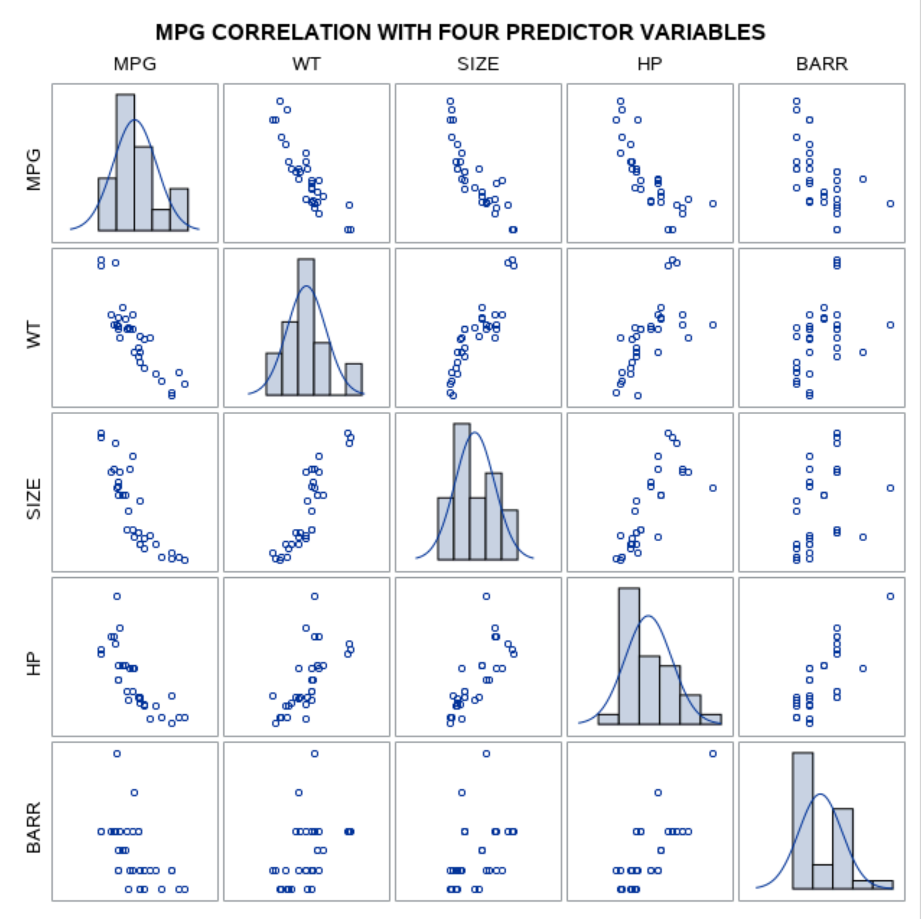


Spearman Correlation Coefficients, N = 32
Prob > |r| under H0: Rho=0

	WT	SIZE	HP	BARR	MPG
WT	1.00000	0.89771 <.0001	0.77468 <.0001	0.49981 0.0036	-0.88642 <.0001
SIZE	0.89771 <.0001	1.00000	0.85104 <.0001	0.53978 0.0014	-0.90888 <.0001
HP	0.77468 <.0001	0.85104 <.0001	1.00000	0.73338 <.0001	-0.89466 <.0001
BARR	0.49981 0.0036	0.53978 0.0014	0.73338 <.0001	1.00000	-0.65750 <.0001
MPG	-0.88642 <.0001	-0.90888 <.0001	-0.89466 <.0001	-0.65750 <.0001	1.00000

Figure 1 illustrates the relationship between Y and each X.

Figure 1. Matrix scatterplot of MPG and four predictive variables.



Least-squares method was applied to fit the following full multiple linear regression model using the dataset described in section 1 using Statistical Analysis System (SAS):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon \quad [1]$$

where Y and X_1 – X_4 are as described in section 1, β_0 - β_4 are regression coefficients to be estimated, and ε is the model error. Stepwise selection method was used to find the "best" model using SLE = 0.15 (SLE = significance level to enter) and SLS=0.05 (SLS = significance level to stay). The characteristics of the "best" model was compared with the full model by comparing significance of the coefficients (α=0.05) and model fitting statistics like STB (produces the standardized regression coefficients), R^2, adjusted R^2, RMSE, predicted sum of squares (PRESS), AIC, and BIC.

3. Results and Discussion

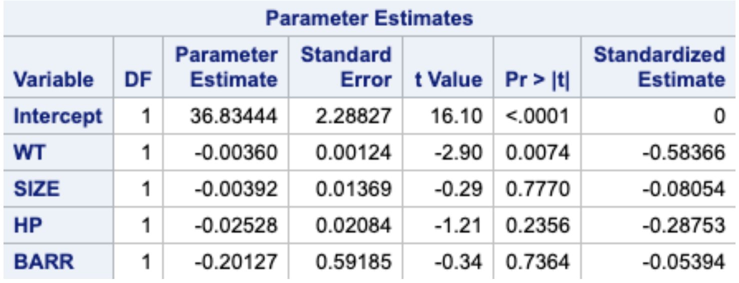
3.1 Full Model

Equation 1 was edited to fit the data using least-square method. Full model can be represented using the following equation:

$$\hat{Y} = 36.83444 – 0.0036\, X_1 – 0.00392\, X_2 – 0.02528\, X_3 – 0.20127\, X_4 \quad [2]$$

Full model R^2 value was equal to 0.82757 indicating that 82.757% of the total variation can be explained by this model. However, p-value analysis revealed that only one slope coefficient (β_1) was statistically significant at α=0.05, while p-values for the other coefficients were higher indicating that they are not statistically significant (Table 3).

Table 3. Estimated regression coefficients for the full model



Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate
Intercept	1	36.83444	2.28827	16.10	<.0001	0
WT	1	-0.00360	0.00124	-2.90	0.0074	-0.58366
SIZE	1	-0.00392	0.01369	-0.29	0.7770	-0.08054
HP	1	-0.02528	0.02084	-1.21	0.2356	-0.28753
BARR	1	-0.20127	0.59185	-0.34	0.7364	-0.05394

Standardized coefficient values of the full model indicate that weight of the car has the biggest negative impact of car gas mileage. Another variable that has a strong negative effect on model output is horse power of the car. However, weight has two times stronger effect on the gas mileage compared to horse power of the car.

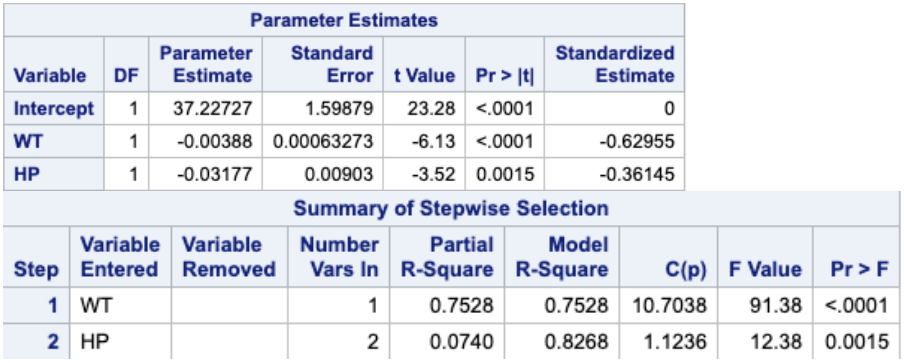
3.2 Reduced Model

After stepwise selection method was implemented, reduced model was selected to represent data analyzed. The "best" model that was produced by stepwise selection method had only two predictive variables (WT and HP) and can be represented using the following equation:

$\hat{Y}$ = 37.22727 – 0.00388 X_1 – 0.03177 X_3 [3]

“Best” model R^2 value was equal to 0.82679. Both slope coefficients were statistically significant at α=0.05 (Table 4).

Table 4. Estimated regression coefficients for the “best” model. Below is a summary of stepwise selection



Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate
Intercept	1	37.22727	1.59879	23.28	<.0001	0
WT	1	-0.00388	0.00063273	-6.13	<.0001	-0.62955
HP	1	-0.03177	0.00903	-3.52	0.0015	-0.36145

Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	WT		1	0.7528	0.7528	10.7038	91.38	<.0001
2	HP		2	0.0740	0.8268	1.1236	12.38	0.0015

The “best” model shows that car gasoline mileage is negatively related to car weight (in pounds) and car engine horse power. However, car weight has 1.7 times stronger impact on the mileage compared to horse power of the car according to the standardized coefficients of the two variables in Table 4.

3.3 Model Comparison

Model fitting statistics and characteristics of the full and “best” model are listed in Table 5. The results indicate that the “best” model represented in equation 3 differs from full model represented in equation 2. R^2 values of both models vary only by less than 1%. On the other hand, adjusted R^2 value produced by “best” model is higher by 1% as compared to full model with four predictive variables as opposed to two predictive variables in “best” model.

Model fitting criteria AIC and BIC were both lower in reduced model. Since AIC values produced by both models differed by more than 2, model differences are distinguishable. Since “best” model produced lower AIC value by more than 2, it can be concluded that “best” model has a better model fit as compared to full model.

Although RMSE values were very close for both models, PRESS value produced by “best” model was lower than the one produced by full model, which indicates that “best” model will perform better at predicting model output Y that full model.

Table 5. Comparison between the full and reduced (“best”) models

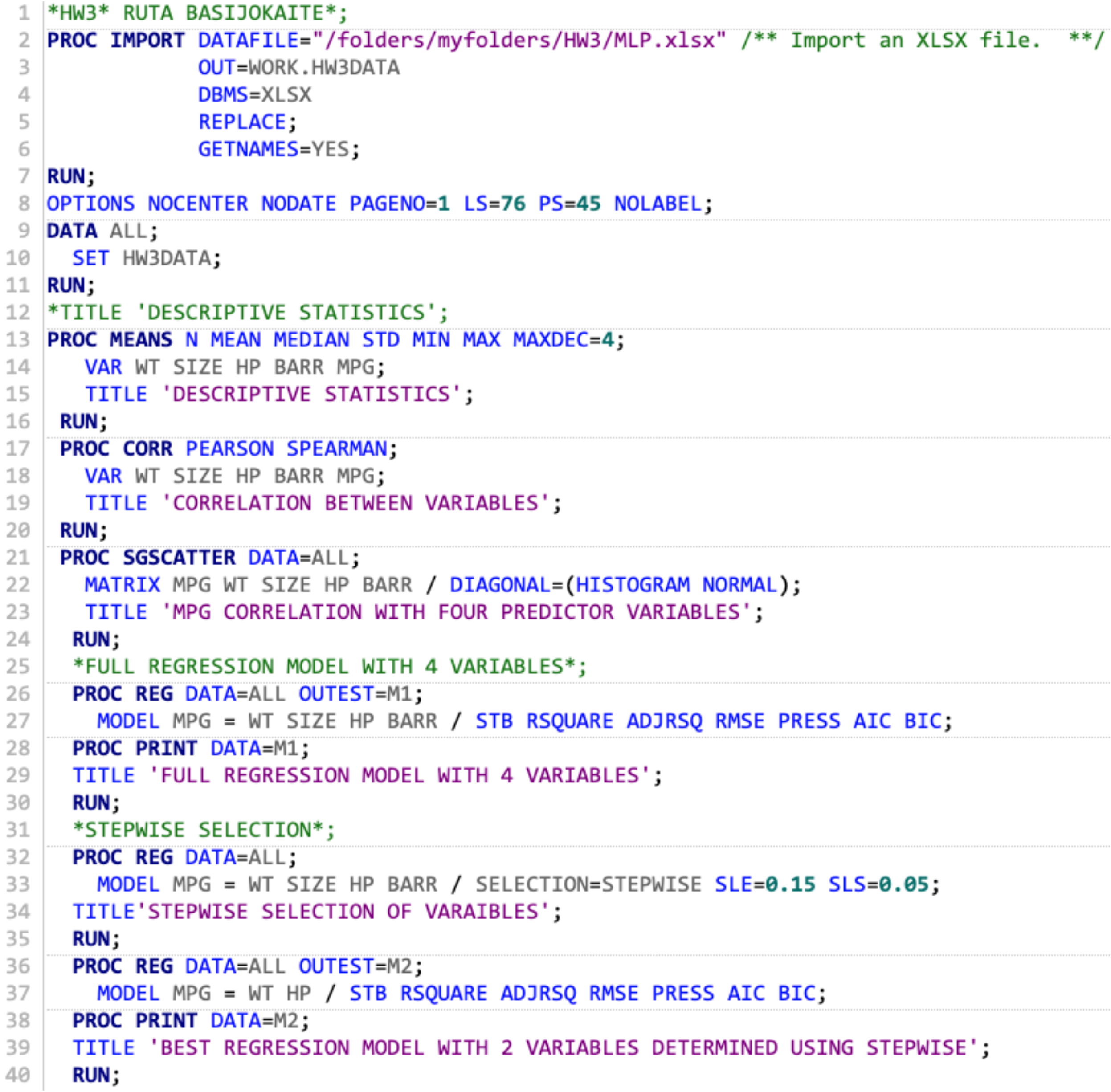
Model	R^2	Adj. R^2	RMSE	PRESS	AIC	BIC
Full	0.82757	0.80203	2.68162	274.246	67.6941	71.4774
“Best”	0.82679	0.81484	2.59341	246.506	63.8403	66.4396

4. Summary

Two multiple linear regression models were developed for the relationship between car gas mileage and four features of the car. Stepwise model selection method resulted in the “best” model with two car feature variables, which can explain about 82.679% of the total variation in the gas mileage data. The results indicated that the “best” model would have better model fit as well as prediction capacity than the full model. The “best” model suggests that both horse power of the car as well as car weight have significant negative effect on gas mileage. In other words, the higher car weigh and higher horse power would reduce car’s fuel efficiency (although car weight affects gas mileage 1.7 more than horse power).

The non-significance of other two car features may be due to correlation among some X variables in the data or the relationship between Y and X might not be linear. Therefore, robust regression techniques should be explored in further studies to account for any outliers or multicollinearity in the data as well as test if non-linear regression model would fit the data better.

5. SAS programs



```
*HW3* RUTA BASIJOKAITE*;
PROC IMPORT DATAFILE="/folders/myfolders/HW3/MLP.xlsx" /** Import an XLSX file.  **/
            OUT=WORK.HW3DATA
            DBMS=XLSX
            REPLACE;
            GETNAMES=YES;
RUN;
OPTIONS NOCENTER NODATE PAGENO=1 LS=76 PS=45 NOLABEL;
DATA ALL;
  SET HW3DATA;
RUN;
*TITLE 'DESCRIPTIVE STATISTICS';
PROC MEANS N MEAN MEDIAN STD MIN MAX MAXDEC=4;
   VAR WT SIZE HP BARR MPG;
   TITLE 'DESCRIPTIVE STATISTICS';
 RUN;
 PROC CORR PEARSON SPEARMAN;
   VAR WT SIZE HP BARR MPG;
   TITLE 'CORRELATION BETWEEN VARIABLES';
 RUN;
 PROC SGSCATTER DATA=ALL;
   MATRIX MPG WT SIZE HP BARR / DIAGONAL=(HISTOGRAM NORMAL);
   TITLE 'MPG CORRELATION WITH FOUR PREDICTOR VARIABLES';
  RUN;
  *FULL REGRESSION MODEL WITH 4 VARIABLES*;
  PROC REG DATA=ALL OUTEST=M1;
    MODEL MPG = WT SIZE HP BARR / STB RSQUARE ADJRSQ RMSE PRESS AIC BIC;
  PROC PRINT DATA=M1;
  TITLE 'FULL REGRESSION MODEL WITH 4 VARIABLES';
  RUN;
  *STEPWISE SELECTION*;
  PROC REG DATA=ALL;
    MODEL MPG = WT SIZE HP BARR / SELECTION=STEPWISE SLE=0.15 SLS=0.05;
  TITLE'STEPWISE SELECTION OF VARAIBLES';
  RUN;
  PROC REG DATA=ALL OUTEST=M2;
    MODEL MPG = WT HP / STB RSQUARE ADJRSQ RMSE PRESS AIC BIC;
  PROC PRINT DATA=M2;
  TITLE 'BEST REGRESSION MODEL WITH 2 VARIABLES DETERMINED USING STEPWISE';
  RUN;
```