Examining Factors that Contribute to a Vehicle's MSRP Value

By Peter Chapman

STAT611: Regression Analysis

Professor Patrick DeFeo

May 2nd, 2025

Introduction

How do manufacturers come up with a price for vehicles on the open market? The manufacturer's suggested retail price (MSRP) of a vehicle is the sale price that manufacturers recommend to dealers when they go to sell the vehicle to consumers. A number of factors contribute to the MSRP of a vehicle, but which ones are most prominent? Are there certain variables that influence this price more heavily than others?

To answer these questions, data was collected from a sample of 428 automotive vehicles from the same production year. Figure 1 below shows the first ten entries of this data set. There are two response variables; manufacturer's suggested retail price (MSRP) is the response of interest, and dealer cost is an alternative response variable that is not of interest in this study. Both of these response variables are measured in the United States dollar (USD). The predictor or independent variables of interest are engine size, number of cylinders in the engine, horsepower of the engine, miles per gallon for city driving, miles per gallon for highway driving, vehicle weight, wheelbase of the vehicle, vehicle length, vehicle width, drive wheels of each vehicle, and vehicle type. Appendix A shows a full data dictionary for this data set, which includes the units and further information about each variable.

Obs	Vehicle	driveWheels	vehicleType	MSRP	dealerCost	engineSize	numCylinders	Horsepower	cityMPG	hwyMPG	Weight	Wheelbase	Length	Width
1	Chevrolet Aveo 4dr	FWD	Sedan	11690	10965	1.6	4	103	28	34	2370	98	167	66
2	Chevrolet Aveo LS 4dr hatch	FWD	Sedan	12585	11802	1.6	4	103	28	34	2348	98	153	66
3	Chevrolet Cavalier 2dr	FWD	Sedan	14610	13697	2.2	4	140	26	37	2617	104	183	69
4	Chevrolet Cavalier 4dr	FWD	Sedan	14810	13884	2.2	4	140	26	37	2676	104	183	68
5	Chevrolet Cavalier LS 2dr	FWD	Sedan	16385	15357	2.2	4	140	26	37	2617	104	183	69
6	Dodge Neon SE 4dr	FWD	Sedan	13670	12849	2.0	4	132	29	36	2581	105	174	67
7	Dodge Neon SXT 4dr	FWD	Sedan	15040	14086	2.0	4	132	29	36	2626	105	174	67
8	Ford Focus ZX3 2dr hatch	FWD	Sedan	13270	12482	2.0	4	130	26	33	2612	103	168	67
9	Ford Focus LX 4dr	FWD	Sedan	13730	12906	2.0	4	110	27	36	2606	103	168	67
10	Ford Focus SE 4dr	FWD	Sedan	15460	14496	2.0	4	130	26	33	2606	103	168	67

Figure 1: The first ten entries of the data from the sample of 428 automotive vehicles are shown. This visual was created using SAS On Demand for Academics.

Methodology

SAS and Initial Hypothesis

To determine the factors that contribute to a vehicle's MSRP, SAS OnDemand for Academics was used to find the best predictive model for price as a function of the vehicle descriptive variables. SAS OnDemand for Academics is an online statistical analysis software platform that allows for complex statistical analysis of large data sets like the sample of automotive vehicles.

The initial hypothesis is that the predictive model from the sample data will include horsepower of the engine, number of cylinders in the engine, drive wheels of each vehicle, and both miles per gallon for city driving and highway driving. In the context of SAS, the expectation is that these vehicle descriptive variables will have statistically significant relationships with MSRP, even in the presence of other variables. Despite this initial expectation, several models will be analyzed and considered.

Indicator Variables

There are three different variables in the data set that are associated with qualitative data instead of quantitative data. This means that instead of numbers, the values associated with these variables are words. In order for SAS to consider these variables in a predictive model, they must be converted into quantitative variables.

The first variable to consider is the Vehicle variable, which is the name of the vehicle being analyzed for the study. There are two parts to the name: the manufacturer and the model. Because there are so many different types of vehicles included in the data set, this variable will not be considered as a quantitative variable in SAS.

The second qualitative variable, called driveWheels in SAS, is the drive wheels of each vehicle. This variable may sound confusing to those not very familiar with cars; it is essentially the number or position of wheels that exert force on the road for the vehicle to move. There are three different values found in the data set for this variable: front-wheel drive (FWD), rear-wheel drive (RWD), and all-wheel drive (AWD). From prior knowledge, it is generally understood that all-wheel drive is superior for a vehicle on the road because of its ability to handle tough weather conditions like snow and ice. Therefore, a new variable called updatedDriveWheels has been created. If the vehicle is equipped with the preferred all-wheel drive (AWD), then the variable takes on a value of 1. If the vehicle has either front-wheel drive or rear-wheel drive, then the variable takes on a value of 0.

The third qualitative variable is the vehicleType variable, which has six different values in the sample data set. These types are 'Sedan', 'Sports Car', 'SUV', 'Wagon', 'Minivan', and a 'Pickup'. Because there are six different categories of vehicleType, five more indicator variables will be created for the data set. Only five will be created in order to avoid perfect multicollinearity, also known as the dummy variable trap. The five indicator variables created in SAS will be called isSedan, isSportsCar, isSUV, isWagon, and isMinivan. The final category, the 'Pickup' type, will be the reference group and will not get its own indicator variable. Each of these new indicator variables will take on a value of 1 if the vehicles have the matching type in the data set. Otherwise, they will take on a value of 0 because they do not match. Appendix B shows the SAS code that relates to the creation of all six indicator variables: updatedDriveWheels, isSedan, isSportsCar, isSUV, isWagon, and isMinivan.

Multiple Linear Regression

Since the indicator variables have now been created, a core model can now be tested using SAS OnDemand for Academics. The first model that will be considered, the base specification model, is a multiple linear regression model with all of the relevant vehicle descriptive variables, including the six indicator variables created above. Multiple linear regression, also called multivariate regression, is when several predictor variables are related to the response variable. Figure 2 shows the SAS code for the first multiple linear regression model considered.

```
36 proc reg data=codedVehicles plots=diagnostics;
37 model MSRP = engineSize numCylinders Horsepower
38 cityMPG hwyMPG Weight Wheelbase Length
39 Width updatedDriveWheels isSedan isSportsCar
40 isSUV isWagon isMinivan;
41 run;
```

Figure 2: The SAS code for the base specification model tested first. The response variable of interest, MSRP, is listed before the "=" statement. The predictor variables, including the six new indicator variables, are located after.

Because fitting a model is a hypothesis test, the null hypothesis for this step is that the overall regression model is not statistically significant. The alternative hypothesis is that the base specification model is statistically significant in explaining the response variable of interest. A confidence level of 10% will be used for this F test. It is anticipated that this null hypothesis will be rejected, meaning the base specification model will be relevant in explaining the dependent variable. In other words, it is expected

that SAS will assert that a model with all relevant vehicle descriptive variables has a statistically significant relationship with vehicle MSRP.

Despite the fact that this initial model as a whole will most likely be considered statistically significant by SAS, it most likely is not the best model to explain vehicle MSRP. While the base specification model as a whole might be relevant, there will most likely be individual predictor variables in this model that are not significant in the presence of the other predictors. Fortunately, from the code above, SAS will automatically test each of the predictor variables individually for its significance in the model while being in the presence of the other predictors. The null hypothesis for each individual predictor is that it does not have a statistically significant relationship with the dependent variable in the presence of other regressors. The alternative hypothesis is that the predictor is relevant with other predictors also in the model. A confidence level of 10% will be used for these individual predictor variable tests, which will be t tests. While variables will not be removed until a future step, this will be a good spot to identify predictors that may not be statistically significant in the presence of other regressors.

Stepwise Selection

Now that the base model is considered, it is time to find the best multiple linear regression model in order to fit MSRP as a function of the vehicle descriptive variables. In order to do this, stepwise selection will be utilized in SAS OnDemand for Academics. Stepwise selection is a regression algorithm that starts with only the intercept included in the model. Predictor variables will be added to the model if they are considered statistically significant, and predictors previously added to the model will be reassessed to see if they are still statistically significant in the new model. In terms of criteria, predictor variables will be added and kept in the model based on a confidence level of 10% for the partial F tests. A p-value will be calculated for each predictor, which represents the probability that the predictor is not statistically significant in the model given the sample data. If this probability is smaller than 0.10, then the predictor variable is added or kept in the model. All of the vehicle descriptive variables and indicator variables will be considered. Figure 3 below shows the SAS code used for this stepwise selection step.

```
45 proc reg data=codedVehicles;
46 model MSRP = engineSize numCylinders Horsepower cityMPG
47 hwyMPG Weight Wheelbase Length
48 Width updatedDriveWheels isSedan isSportsCar
49 isSUV isWagon isMinivan / selection=stepwise slentry=0.1 slstay=0.1;
50 run;
```

Figure 3: The SAS code for stepwise selection algorithm. A confidence level of 10% is used for both the potential entry and removal of predictor variables in this multiple linear regression model.

Multicollinearity Checks

Once the stepwise selection algorithm finalizes a subset of predictor variables for the model, the remaining predictor variables must be checked for multicollinearity. This is defined as a condition where the predictor variables are correlated with each other. In other words, multicollinearity between predictors prevents the ability to estimate the change in MSRP with respect to one predictor, given that the other predictors are held fixed. Applying this to the model, there must be a check to see if the remaining vehicle descriptive variables in the model have near linear dependence among them.

SAS code will be utilized for this check for multicollinearity. Figure 4 below shows the SAS code that will be used for the multicollinearity check. During this step, SAS will output the condition number (referred to as condition index by SAS) as a general indicator of the impact of multicollinearity on the regression coefficients. If this value is above 1000, then multicollinearity is present in the regression

model. SAS will also output Variance Inflation Factors (VIFs) for each predictor variable, which is the amount of increase in the variance of a coefficient above the ideal orthogonal case. As a general rule of thumb, a VIF over 10 indicates multicollinearity. Generally, the condition number indicates multicollinearity in the model, and the VIF indicates the severity of the effect and which coefficients are affected from the impact. If there is a sense of multicollinearity, the regression model can be estimated but it is ill-advised. Therefore, if multicollinearity is evident during this step, predictors will be removed from the model or an alternative method for fitting the regression model will be utilized. If there is no multicollinearity, then the multiple linear regression model will be moved to the next step.

```
53 proc reg data=codedVehicles;
54 model MSRP = Horsepower Wheelbase hwyMPG Weight
55 Width isSUV numCylinders engineSize isSportsCar
56 / vif collinoint;
57 run;
```

Figure 4: SAS code showing multicollinearity checks relating to the remaining predictor variables in the multiple linear regression model. The phrase 'vif' refers to the Variance Inflation Factor of each predictor.

Interaction Checks

After the checks for collinearity, it is important to check for interaction between the remaining variables to consider the idea that the effect of a predictor variable on the response variable is influenced by one of the other predictors. To do this, the remaining predictor variables will be used to create interaction terms that need to be checked for significance. In SAS, a new multiple linear regression procedure will be created, which will include each of the relevant predictor variables that remain. Additionally, new interaction terms will be created from the remaining variables. Appendix C shows this code in SAS. There won't be two-way interaction terms for all two-way combinations of remaining predictors, as there are too many to put in one model. This would lead to overfitting in the model, as there would be several interaction terms that need to be created and analyzed. Rather than including all of them, two-way interaction terms will be selectively created in a theoretical sense based on the individual predictor variables remaining. To avoid overfitting, there will most likely only be a few of these interaction terms. The null hypothesis is that there is no interaction between two given predictors, and the alternative hypothesis is that the two predictors do have a statistically significant interaction relating to the effect on the response variable. A confidence level of 10% will be used for these tests. If an interaction term is significant, it will then be kept in the model to retain its significant effects.

Final Model and Residual Analysis

Once the final model is selected and finalized, a series of tests will be conducted to determine the appropriateness of the multiple linear regression model. The first test is the same F test that was used for the base specification model used at the very beginning. The null hypothesis for this step is that the overall regression model is not statistically significant. The alternative hypothesis is that this final MLR model is statistically significant in explaining the response variable of interest, MSRP. A confidence level of 10% will be used for this F test. An additional statistic that will be examined at this part is R-squared, which is the variability in the response variable MSRP that is explained by the multiple linear regression model.

After the final model is analyzed for its strength and appropriateness, a residual analysis will be performed on the model. There are four main assumptions made in regression that will have to be tested. The first assumption is that the model is correctly specified. To test this assumption, the residuals of the model will be plotted against the predicted values. If there is random fluctuation around 0, then it implies

a correctly specified model. A curve pattern implies that a higher order model may be needed. The second assumption is that there is an independence of errors across the observations. To test this, the residuals will be plotted in the time order that they appear. If there is random variation around 0, then it is safe to say that the error terms are independent from each other. If there is a particular pattern, then there may be a violation of the independence of errors assumption. The third assumption is the constant variance of errors assumption. To test this, the scatter plot of residuals versus predicted values will be looked at again. If there is random fluctuation around 0, then the errors are homoscedastic. If there isn't obvious fluctuation, then the errors may be heteroscedastic. The fourth and final assumption is the normality assumption of errors. To test this, a normal probability plot of the residuals will be constructed. If a lot of points are away from the reference line, then the residuals may not be normally distributed. Appendix D shows the SAS code used for these residual analysis tests.

Results and Discussion

Core Model Analysis

Figure 5 below shows the ANOVA table from SAS for the multiple linear regression of the base specification model, which includes all of the vehicle descriptive variables, including the six indicator variables. As expected, the complete regression model here was considered relevant by SAS. The F test statistic is very large, at a value of 78.14. In addition, SAS calculates the p-value to be an extremely small number less than 0.0001. This p-value is the probability that the multiple linear regression model does not have a statistically significant relationship with MSRP from the sample data. Because this probability is so low, the null hypothesis is rejected at the 10% confidence level. This base model, with all of the vehicle descriptive variables, has a statistically significant relationship with MSRP, the response variable.

Number of Observations Read	428
Number of Observations Used	405
Number of Observations with Missing Values	23

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	15	1.150655E11	7671035775	78.14	<.0001
Error	389	38187692707	98168876		
Corrected Total	404	1.532532E11			

Root MSE	9908.02078	R-Square	0.7508
Dependent Mean	32858	Adj R-Sq	0.7412
Coeff Var	30.15425		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	45462	18730	2.43	0.0157
engineSize	1	-3410.92250	1499.54149	-2.27	0.0235
numCylinders	1	2598.88420	885.62058	2.93	0.0035
Horsepower	1	225.05022	14.74785	15.26	<.0001
cityMPG	1	65.34459	345.10141	0.19	0.8499
hwyMPG	1	592.76815	367.17824	1.61	0.1073
Weight	1	11.51234	2.23491	5.15	<.0001
Wheelbase	1	-620.68930	171.41018	-3.62	0.0003
Length	1	86.06277	89.35228	0.96	0.3361
Width	1	-1048.86246	315.78412	-3.32	0.0010
updatedDriveWheels	1	-155.97680	1510.47237	-0.10	0.9178
isSedan	1	2232.87826	3173.20622	0.70	0.4821
isSportsCar	1	7700.82647	4206.38732	1.83	0.0679
isSUV	1	-1015.93834	3204.47744	-0.32	0.7514
isWagon	1	1759.18046	3540.15692	0.50	0.6195
isMinivan	1	4901.76059	3615.00655	1.36	0.1759

Figure 5: Above are the ANOVA and Parameter Estimates tables for the multiple linear regression model that includes all vehicle descriptive variables, including the six indicator variables. The overall model results are shown on the left, and the tests of each predictor variable are on the right.

Despite the overall model's significance, the table on the right in Figure 5 shows that not all of the predictor variables had a statistically significant relationship with the response variable in the presence of the other predictors. The p-values on the right represent the probability that each individual predictor variable is not statistically significant in the presence of other predictors using the sample data. Because a confidence level of 10% was used for these tests, the null hypothesis is not rejected if the probability for a given predictor variable is higher than 0.10. In other words, any predictor variable that has a probability of greater than 0.10 is not considered statistically significant in this model. The vehicle descriptive variables that are not statistically significant during this step are miles per gallon for city driving, miles per gallon for highway driving, vehicle length, the drive wheels of each vehicle, and whether or not the vehicle is a sedan, SUV, wagon, or minivan.

Model from Stepwise Selection

All variables left in the model are significant at the 0.1000 level.

No other variable met the 0.1000 significance level for entry into the model.

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	Horsepower		1	0.6873	0.6873	87.2172	885.63	<.0001
2	Wheelbase		2	0.0296	0.7169	42.9986	42.04	<.0001
3	hwyMPG		3	0.0038	0.7207	39.0920	5.43	0.0203
4	Weight		4	0.0103	0.7310	24.9850	15.34	0.0001
5	Width		5	0.0050	0.7359	19.2249	7.51	0.0064
6	isSUV		6	0.0046	0.7405	14.0360	7.06	0.0082
7	numCylinders		7	0.0026	0.7431	12.0359	3.96	0.0473
8	engineSize		8	0.0029	0.7460	9.5533	4.48	0.0350
9	isSportsCar		9	0.0032	0.7492	6.5059	5.09	0.0246

Figure 6: Above is the summary table for the stepwise selection analysis performed by SAS. The nine variables are the statistically significant predictors in the presence of other predictors at a confidence level of 10%.

Figure 6 above shows the multiple linear regression model calculated by SAS from the stepwise selection step with the entry and removal confidence level of 10%. As seen above, the relevant vehicle descriptive variables are horsepower of the engine, wheelbase of the vehicle, miles per gallon for highway driving, vehicle weight, vehicle width, number of cylinders in the engine, engine size, whether or not the vehicle is an SUV, and whether or not the vehicle is a sports car. These predictor variables all have a statistically significant relationship with MSRP, even in the presence of other predictors in the model. This subset of vehicle descriptive variables becomes the multiple linear regression model for future steps.

No Severe Multicollinearity

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	40600	16951	2.40	0.0171	0
Horsepower	1	223.92266	14.00379	15.99	<.0001	3.98007
Wheelbase	1	-546.59194	136.03262	-4.02	<.0001	4.55443
hwyMPG	1	665.91451	157.18602	4.24	<.0001	3.33103
Weight	1	11.40611	2.04171	5.59	<.0001	9.24662
Width	1	-841.64193	275.86739	-3.05	0.0024	3.89412
isSUV	1	-3773.36485	1887.85778	-2.00	0.0463	1.84628
numCylinders	1	2671.11194	842.85292	3.17	0.0016	6.61614
engineSize	1	-3423.87966	1442.39908	-2.37	0.0181	9.29074
isSportsCar	1	4981.57181	2207.49599	2.26	0.0246	2.00329

Number	Eigenvalue	Condition Index
1	5.27368	1.00000
2	1.49304	1.87941
3	0.98196	2.31744
4	0.44297	3.45039
5	0.31240	4.10869
6	0.18644	5.31846
7	0.15153	5.89943
8	0.08491	7.88094
9	0.07307	8.49569

Figure 7: Above are the summary tables for the multicollinearity analysis performed by SAS. The Variance Inflation Factors (VIFs) are on the left and the condition indices are on the right.

Figure 7 above shows the results from the multicollinearity test performed by SAS OnDemand for Academics. The image on the right shows the condition indices on the regression coefficients. Notably, all of these condition indices are well below 1000, which shows that there isn't much of a problem with multicollinearity for this regression model. Additionally, the image on the left shows the Variance Inflation Factors (VIFs) on the associated regression coefficients. All of these VIFs are below 10, which shows that there is not a problem with severe multicollinearity for this multiple linear regression model. To conclude, the multicollinearity analysis in SAS proved that linear dependence was not a significant problem among the remaining vehicle descriptive variables in the model. Therefore, no predictors were removed here and an alternative method for fitting the regression model was not needed.

Some Significant Interaction

As mentioned above, none of the predictors from the stepwise selection step were removed during the check for multicollinearity. At this point, two-way interaction terms were created based on the remaining vehicle descriptive variables in the model. These terms come from background knowledge about vehicles, which is limited. The first interaction term was created between horsepower and whether or not the vehicle is a sports car. Generally, sports cars are equipped with higher horsepower, and this interaction could lead to an increase in MSRP. The second and final two-way interaction term created was between miles per gallon for highway driving and vehicle weight. Lighter vehicles typically have better miles per gallon than heavier vehicles, and this interaction could affect MSRP of a vehicle. Appendix C shows the creation of these interaction terms in SAS OnDemand for Academics.

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	45699.64803	18298.91846	2.50	0.0129
Horsepower	181.84381	15.20135	11.96	<.0001
Wheelbase	-442.37737	132.38236	-3.34	0.0009
hwyMPG	414.64127	294.50139	1.41	0.1599
Weight	10.84378	2.69432	4.02	<.0001
Width	-928.97081	269.48019	-3.45	0.0006
isSUV	-3451.00070	1886.16091	-1.83	0.0681
numCylinders	3226.87985	820.80150	3.93	<.0001
engineSize	-3923.03640	1426.91427	-2.75	0.0062
isSportsCar	-22904.40362	5117.33631	-4.48	<.0001
hwyMPG*Weight	0.06030	0.10520	0.57	0.5669
Horsepowe*isSportsCa	113.55523	18.89544	6.01	<.0001

Figure 8: Above is the parameter estimates table for the test for significant interaction between a few of the vehicle descriptive variables.

Figure 8 above shows the results of the SAS test for significant interaction for the two-way interaction terms included in the model. The p-values on the right represent the probability that the two-way interaction terms are not statistically significant from the sample data. A confidence level of 10% was used for testing these interaction effects. The null hypothesis is not rejected for the interaction term between miles per gallon for highway driving and weight of the vehicle. The probability of 0.5669 suggests that there is not enough evidence to conclude that there is a statistically significant interaction between the effects of weight and miles per gallon for highway driving on MSRP. On the other hand, the null hypothesis is rejected for the interaction term between horsepower and whether or not the vehicle is a sports car. The probability is extremely small, so there is enough evidence to conclude that there is a statistically significant interaction between the effects of horsepower and being a sports car on MSRP. The interaction term between weight and miles per gallon for highway driving is removed from the model, as it is not statistically significant.

Final Model

The GLM Procedure

Dependent Variable: MSRP

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	118040162622	11804016262	132.08	<.0001
Error	394	35213066706	89373265.751		
Corrected Total	404	153253229328			

R-Square	Coeff Var	Root MSE	MSRP Mean
0.770230	28.77169	9453.743	32857.80

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	40888.74122	16246.36530	2.52	0.0122
Horsepower	183.00487	15.05292	12.16	<.0001
Wheelbase	-435.25870	131.68607	-3.31	0.0010
hwyMPG	559.27753	151.69065	3.69	0.0003
Weight	11.90326	1.95853	6.08	<.0001
Width	-900.30188	264.57153	-3.40	0.0007
isSUV	-3753.39530	1809.32391	-2.07	0.0387
numCylinders	3283.21923	814.19958	4.03	<.0001
engineSize	-4111.95807	1387.13764	-2.96	0.0032
isSportsCar	-22955.49588	5112.19788	-4.49	<.0001
Horsepowe*isSportsCa	113.30075	18.87411	6.00	<.0001

Figure 9: Above are the ANOVA and Parameter Estimates tables for the finalized model. The overall model results are shown on the left, and the tests of each predictor variable are on the right.

Figure 9 above shows the ANOVA table from SAS for the multiple linear regression of the finalized model, which was considered relevant by SAS. The F test statistic is very large, at a value of 132.08. In addition, SAS calculates the p-value to be an extremely small number less than 0.0001. This p-value is the probability that the final model does not have a statistically significant relationship with MSRP from the sample data. Because this probability is so low, the null hypothesis is rejected at the 10% confidence level. This final model, with a few of the vehicle descriptive variables and an interaction term, has a statistically significant relationship with MSRP, the response variable. The R-squared for this model is 0.770230, which means that just over 77% of the variability in MSRP is explained by this model. This may not be very high, but it is reasonable considering the multifactorial nature of MSRP.

The p-values on the right represent the probability that each individual predictor variable is not statistically significant in the presence of other predictors. Because a confidence level of 10% was used for these tests, the null hypothesis is rejected for all of the predictor variables in the model. The vehicle descriptive variables that are statistically significant in explaining MSRP are horsepower of the engine, wheelbase of the vehicle, miles per gallon for highway driving, vehicle width, vehicle weight, the number of cylinders in the engine, engine size, whether or not the vehicle is a sports car, whether or not the vehicle is an SUV, and the interaction between being a sports car and horsepower of the engine.

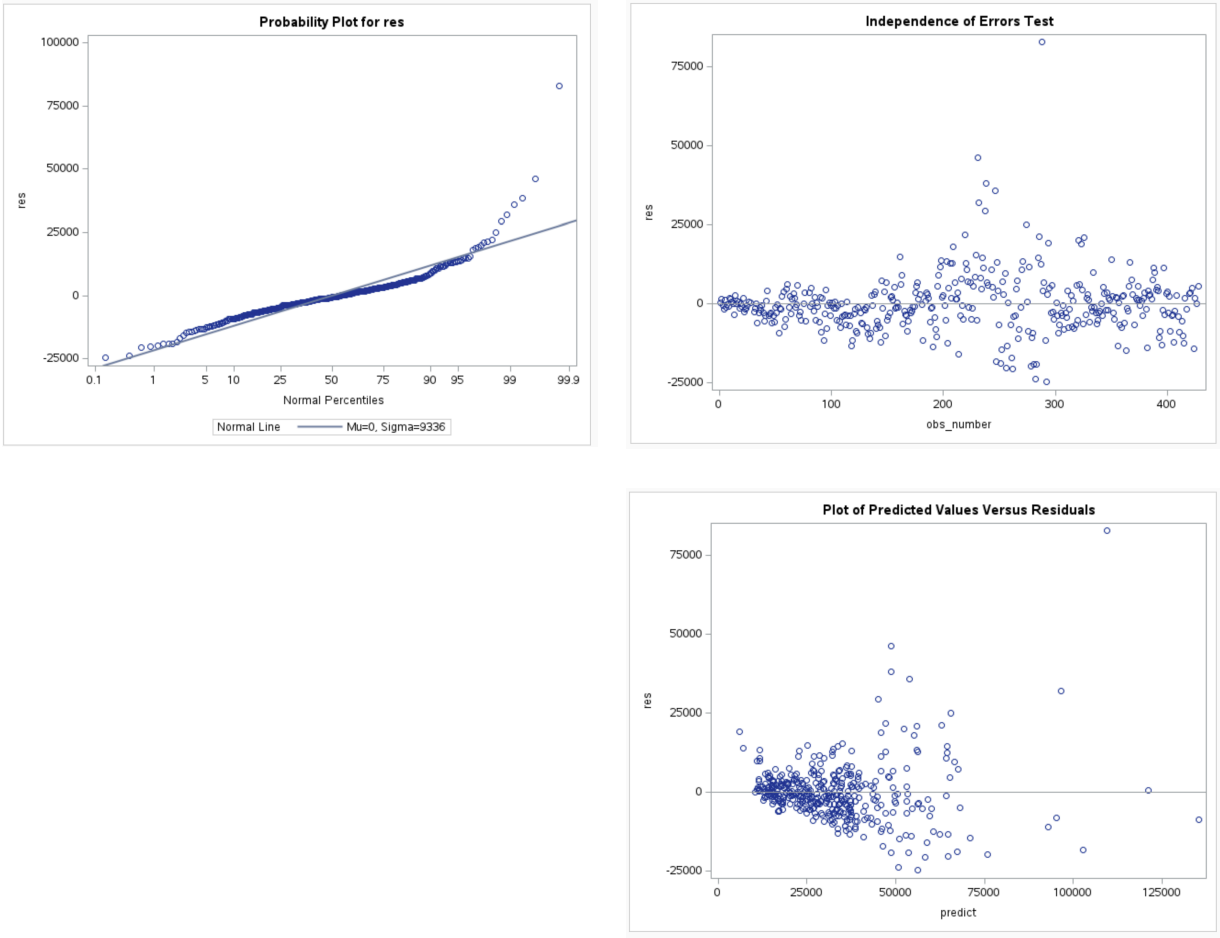

Figure 10: Above are the results from the residual analysis performed in SAS. The graph in the top left is the normal probability plot of residuals. The graph in the top right plots the residuals versus the time order that they appear in. Finally, the graph in the bottom right corner shows residuals plotted versus predicted values.

Figure 10 above shows the results from the residual analysis performed in SAS. To start, the normal probability plot in the top left corner shows the residuals plotted along the reference line. For the most part, the residuals are along the reference line. There are a few data points away from the reference line on the right tail, which could suggest that the errors are not normally distributed. However, because there are over 400 data points, it makes sense that there are a few data points away from the reference line. Therefore, the normality of errors assumption is reasonable to make; there are just a few unusual data points.

The graph in the top right is the plot of the residuals versus the time order that they come in. One could argue that there is a slight triangular pattern in the center, which would suggest a violation of the independence of errors assumption and a violation of the constant variance of errors assumption. This suggests that a generalized least squares method or a time series approach is needed. However, this pattern is not extremely strong, as there is still a decent amount of fluctuation of data points around 0. Therefore, it is reasonable to conclude that there is an independence of errors.

Finally, the graph in the bottom right corner is the graph of the predicted values versus the residuals. There isn't too much randomness around 0, which may suggest that an incorrectly specified model was used. However, there isn't an obvious curve pattern or any other pattern, so it is reasonable to say that the correct model specification was used. However, the lack of randomness around 0 does suggest that there may not be a constant variance of errors (homoscedasticity). Additionally, there is one data point that is significantly away from the rest, which is more evidence of a possible unusual data point among the 400+ observations.

Conclusion

The finalized model for MSRP as a function of vehicle descriptive variables includes the following predictors: horsepower of the engine, wheelbase of the vehicle, miles per gallon for highway driving, vehicle width, vehicle weight, the number of cylinders in the engine, engine size, whether or not the vehicle is a sports car, and whether or not the vehicle is an SUV. Additionally, there is significant interaction between horsepower of the engine and whether or not the vehicle is a sports car when it comes to the effects on MSRP, the response variable of interest. These predictor variables, including the interaction term, are the best model for MSRP as a function of vehicle descriptive variables.

From a multicollinearity test, there was no severe linear dependence among predictor variables in the model. Notably, this test was done without the interaction term in the model. For the residual analysis, the results were a little bit inconclusive. The normality of errors assumption was deemed reasonable, but there were a few data points that seemed unusual. There was a slight problem with the independence of errors assumption; there was decent randomness to the errors but a slight pattern suggested that a generalized least squares method or a time series approach could be needed. For the model specification assumption, it was concluded that the model used was the correct specification. Finally, the test for homoscedasticity proved that there wasn't a perfect constant variance of errors and that a unique data point exists.

Appendices

Appendix A: Data Dictionary for Data Set

Variable	Description and Units
Vehicle	Name of the vehicle, including manufacturer and model
driveWheels	Drive wheels of each vehicle; front wheel drive (FWD), right wheel drive (RWD), or all wheel drive (AWD)
MSRP	Manufacturer's suggested retail price, measured in the United States dollar (USD, $)
dealerCost	Dealer's price for the vehicle, measured in the United States dollar (USD, $)
engineSize	Size of the engine, measured in liters
numCylinders	Number of cylinders in the engine
Horsepower	Horsepower of the engine
cityMPG	Miles per gallon for city driving
hwyMPG	Miles per gallon for highway driving
Weight	Weight of the vehicle in pounds
Wheelbase	Wheelbase (difference between the centers of the front and rear wheel) of the vehicle, measured in inches
Length	Length of the vehicle in inches
Width	Width of the vehicle in inches

Table A.1: The data dictionary for the sample data of 428 cars. The variable names match the analysis conducted in SAS.

Appendix B: SAS Code for Creation of Indicator Variables

```
/* transform the driveWheels variable to be a numeric, indicator variable */
/* transform the vehicleType variable to be numeric indicator variables */
data codedVehicles;
    set vehicles;
    if upcase(driveWheels) = 'AWD' then updatedDriveWheels = 1;
    else updatedDriveWheels = 0;
    if upcase(vehicleType) = "SEDAN" then isSedan = 1;
    else isSedan = 0;
    if upcase(vehicleType) = "SPORTS CAR" then isSportsCar = 1;
    else isSportsCar = 0;
    if upcase(vehicleType) = "SUV" then isSUV = 1;
    else isSUV = 0;
    if upcase(vehicleType) = "WAGON" then isWagon = 1;
    else isWagon = 0;
    if upcase(vehicleType) = "MINIVAN" then isMinivan = 1;
    else isMinivan = 0;
run;
```

Figure B.1: The code above shows the creation of the six indicator variables considered for a multiple linear regression model for the MSRP of vehicles.

Appendix C: SAS Code for Interaction Checks

```
60 proc glm data=codedVehicles;
61 model MSRP = Horsepower Wheelbase hwyMPG Weight
62 Width isSUV numCylinders engineSize isSportsCar
63 hwyMPG*Weight isSportsCar*Horsepower;
64 run;
```

Figure C.1: The code above shows the creation and analysis of the two-way interaction terms considered for a multiple linear regression model for the MSRP of vehicles.

Appendix D: SAS Code for Residual Analysis

```
/* normality assumption check; normal probability plot of residuals */
proc univariate data=residuals;
    var res;
    probplot res / normal (mu=est sigma=est);
    title "Normal Probability Plot of Residuals";
run;

/* scatter plot of predicted versus residuals for constant variance test */
/* also the model specification test */
proc sgplot data=residuals;
    scatter x=predict y=res;
    refline 0 / axis=y;
    title "Plot of Predicted Values Versus Residuals";
run;

data residuals_time;
   set residuals;
   obs_number + 1;  /* creates a time-like variable starting at 1 */
run;

/* scatter plot of residuals versus the time that they come in */
/* independence of errors test */
proc sgplot data=residuals_time;
    scatter x=obs_number y=res;
    refline 0 / axis=y;
    title "Independence of Errors Test";
run;
```

Figure D.1: The SAS code above shows the residual analysis tests for the four assumptions made about linear regression.