Deliverable 3

Team Members: Henry Bayly|Jacqueline Lin|Jiasong Huang|Maysen Pagan|Peiqi Lu|Phil Ledoit

Group Number 1

Lab session: C2

Introduction

The dataset that we are examining comprises of data of 4,843 actual BMW cars that were sold via B2B in 2018. The purpose of analyzing the dataset is to identify the variables that might have influenced the sale price of a car. The possible affecting factors includes model key, mileage, engine power, age and also 8 criteria based on the equipment of car which have been labeled feature_1 to feature_8 in the dataset.

In our first lab we built a simple linear model that predicted the price of each BMW from the mileage. Simple linear regression revealed a statistically significant linear relationship between price and mileage. However, regression diagnostics revealed that our model might not be good. Therefore, in this report we addressed some of those issues. Specifically we looked into removing outliers and unrealistic points from the dataset. Moreover, we used an AIC based model selection criteria to perform multiple linear regression.

All source code and figures can be found in the appendix.

Data Cleaning

In this section we discuss relevant ways in which we altered the dataset. When we performed simple linear regression using mileage as a predictor we noticed a few unrealistic points. We defined an 'unrealistic' mileage to be one that either had negative miles or more than 500,000. We removed 2 cars from our data that met this criteria. Therefore, we were left with $N = $4,841 cars. Furthermore, when analyzing the Model Keys we noticed that there was a single car with model type 'ActiveHybrid 5'. Because there was only one car of this type, we removed it from the data set. Thus, our final data comprised of $N = $4,840 cars.

Variable Inclusion Justifications

Before discussing our model selection techniques, we will discuss rationale and hypotheses for inclusion of each variable in the full model.

The first predictor we used was Mileage. Typically, cars with lower mileage tend to be more valuable since they have undergone less wear and tear, potentially making them more dependable and long-lasting, albeit more costly. Conversely, cars with higher mileage usually command a lower price due to their greater likelihood of requiring frequent repairs and maintenance, leading to increased ownership expenses, hence cheaper than those with lower mileage. Thus, we hypothesize that both higher mileage result in the lower sale price of a car.

The next predictor we used was the Age of the car. We defined the age of the car as the difference in time between when the car was sold and when it was registered. The age of a used car also has a similar effect on car value as mileage. However, an older car with a low mileage might place the car value in a better position than a newer car with a high mileage. Overall, we hypothesize that older cars will be worth less than newer cars.

Furthermore, different fuel types and paint colors have an impact on the sale price of a car. Fuel price is a crucial consideration for consumers when purchasing a vehicle since they are subject to fluctuation, making it challenging to estimate monthly fuel expenses. The average price of petrol per gallon has risen from $3.14 to $3.49 since the beginning of 2022, while diesel has experienced a 38% increase from $3.61 to $5.31 during the same period. The cars in our data set had 4 different fuel types: hybrid, electric, diesel, and petrol. To represent this in our model, we categorized cars as using new energy (hybrid or electric) versus not (diesel or petrol). We hypothesize that cars using new energy will cost more than those that don't use new energy due to the fact that upkeep on the new energy vehicles costs less. As for paint color, we categorized the paint color data into two categories: common colors and less common colors. Common colors, including white, black, gray, and silver, account for 77% of new car colors and typically experience a 15% depreciation rate per year over the first three years of ownership. However, cars with less common paint colors tend to depreciate less, indicating that cars with the common colors listed above are likely to have lower sale prices than those with less common colors.

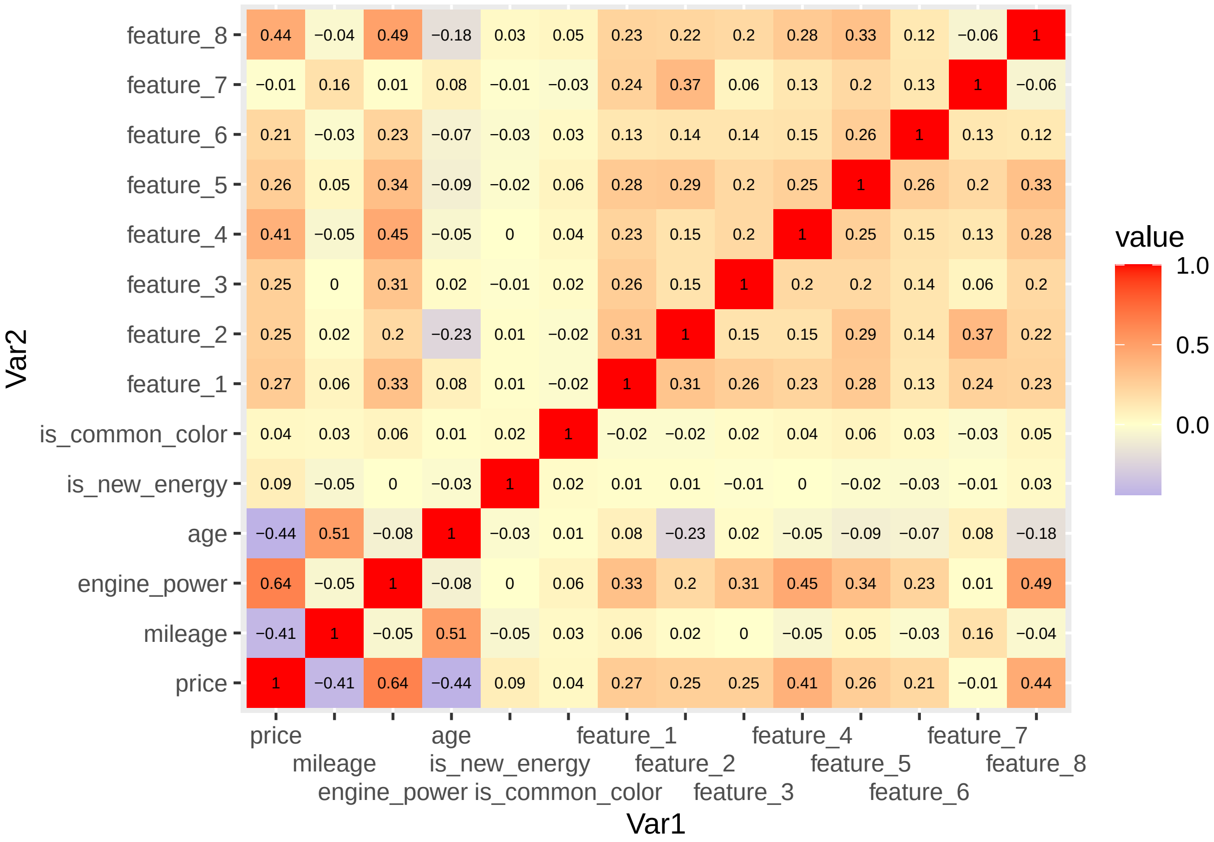
The data contained information on whether or not each car contained certain features. There were 8 features in total. No other information on the actual description of the features was provided. As a result, they were included in the full model. However, as described below, our model selection process informed us on the most relevant features for our model. Therefore, we had no initial rationale or hypotheses about their inclusion in the model but were able to tell through model selection and our model results which features were significant linear predictors of price.

The data also contained data on the model key as well as the type of car. In our research we found that as the 'number' of the model key increased, so did the size of the car. Moreover, the letters in the model key loosely translated to the fuel type of the car. Because we already included a variable for the type of energy that the car used, we created a new variable to capture the size and type of car. This variable was the number of seats that each car had. We represented this as a categorical variable that had three levels: 2 seats, 5 seats, or 7 seats. We are unsure whether or not this variable will have an effect on the price of the car. This is because a 2 seat car is likely to be a sports car and maybe more expensive. However, a 7 seat car is bigger and therefore might cost more to produce resulting a higher price. We will explore this relationship in the next section using boxplots.

Last but not least, it is worth noting that higher engine power translates to better acceleration and overall performance of a car. Based on this, we speculate that cars with better performance will command a higher sale price.

Data Exploration

```
##              Var1  Var2 value
## 1           price price  1.00
## 2         mileage price -0.41
## 3    engine_power price  0.64
## 4             age price -0.44
## 5   is_new_energy price  0.09
## 6 is_common_color price  0.04
```



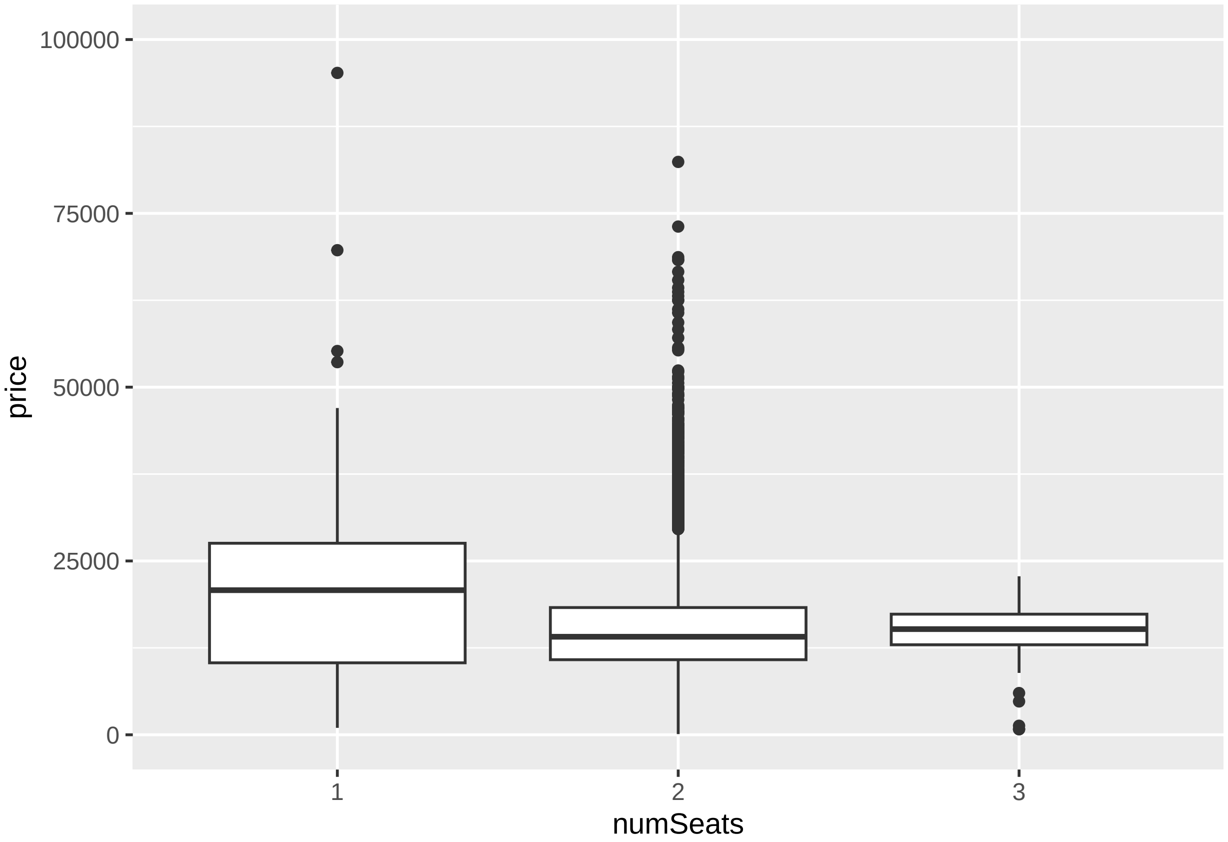
Var2	price	mileage	engine_power	age	is_new_energy	is_common_color	feature_1	feature_2	feature_3	feature_4	feature_5	feature_6	feature_7	feature_8
feature_8	0.44	−0.04	0.49	−0.18	0.03	0.05	0.23	0.22	0.2	0.28	0.33	0.12	−0.06	1
feature_7	−0.01	0.16	0.01	0.08	−0.01	−0.03	0.24	0.37	0.06	0.13	0.2	0.13	1	−0.06
feature_6	0.21	−0.03	0.23	−0.07	−0.03	0.03	0.13	0.14	0.14	0.15	0.26	1	0.13	0.12
feature_5	0.26	0.05	0.34	−0.09	−0.02	0.06	0.28	0.29	0.2	0.25	1	0.26	0.2	0.33
feature_4	0.41	−0.05	0.45	−0.05	0	0.04	0.23	0.15	0.2	1	0.25	0.15	0.13	0.28
feature_3	0.25	0	0.31	0.02	−0.01	0.02	0.26	0.15	1	0.2	0.2	0.14	0.06	0.2
feature_2	0.25	0.02	0.2	−0.23	0.01	−0.02	0.31	1	0.15	0.15	0.29	0.14	0.37	0.22
feature_1	0.27	0.06	0.33	0.08	0.01	−0.02	1	0.31	0.26	0.23	0.28	0.13	0.24	0.23
is_common_color	0.04	0.03	0.06	0.01	0.02	1	−0.02	−0.02	0.02	0.04	0.06	0.03	−0.03	0.05
is_new_energy	0.09	−0.05	0	−0.03	1	0.02	0.01	0.01	−0.01	0	−0.02	−0.03	−0.01	0.03
age	−0.44	0.51	−0.08	1	−0.03	0.01	0.08	−0.23	0.02	−0.05	−0.09	−0.07	0.08	−0.18
engine_power	0.64	−0.05	1	−0.08	0	0.06	0.33	0.2	0.31	0.45	0.34	0.23	0.01	0.49
mileage	−0.41	1	−0.05	0.51	−0.05	0.03	0.06	0.02	0	−0.05	0.05	−0.03	0.16	−0.04
price	1	−0.41	0.64	−0.44	0.09	0.04	0.27	0.25	0.25	0.41	0.26	0.21	−0.01	0.44

In order to understand the relationships between our covariates with each other and the outcome variable we created a correlation matrix. Due to the large amount of covariates, instead of creating scatterplots to visualize the correlations, we elected to use a heat map. This is a simple way to assist in the interpretation and differentiation of correlation coefficients. The aim of this step is to understand whether or not our covariates have any linear association with the outcome. Moreover, we want to check our model assumption to see confirm the covariates are independent of each other.

We will first look at the correlation between our covariates and the outcome. We elected to use a correlation cutoff of $|r| > 0.25$ as being high enough for us to include our model, where r is the correlation coefficient. Thus, if the correlation between a covariate and the outcome is not higher than 0.25 we conclude that it is not associated enough with the outcome to include. Three covariates met this criteria: whether or not the car used new energy ($r = 0.09$), whether or not the car had a common color ($r = 0.04$), and whether or not the car had feature 7 ($r = -0.01$). These variables were removed from the model selection process. We note here that engine power is strongly correlated with our outcome of price ($r = 0.64$). All other covariates can be categorized as having a moderate linear association with our outcome.

Section on covariate to covariate correlation

Section on what this means.



Here we visually compare whether or not price differs by the number of seats present in the car. We can see that all 3 categories (2 seats, 5 seats, and 7 seats) overlap in their interquartile range. As a result of this, we conclude that the price of the car does not have a significant relationship with the number of seats in the car. Therefore, we will not include this in the model selection process.

Model Selection

In order to identify the best model for our data we used an AIC (Akaike Information Criterion) based model selection approach. Recall that AIC is a measure that helps to balance the goodness of fit and complexity of the model. We wanted to capture as much variation in the price of the cars as possible, but wanted to avoid overfitting our model.

After performing our AIC based selection (without the dummy variables for Model Series), we conclude that 9 covariates should be used to predict the price of BMWs in our dataset. Specifically we found that Price should be predicted using: mileage, engine power, age of the car, and whether or not they had features 1,2,3,4,6, and 8. This combination of predictors from our full model resulted in the lowest AIC score, which is ideal.

To cross validate this result we also performed model selection using Mallow's Cp as an indicator of good performance. Recall that Mallow's Cp is similar to AIC but different. The idea of Mallow's Cp is to compare the expected mean square error of multiple candidate models to a 'perfect' model. To utilize this statistic we used forward stepwise selection.

After performing forward stepwise model selection, the use of Mallow's Cp statistic generated a different model. This model had 7 covariates: mileage, engine power, age of the car, and whether or not they had features 1,3,4,6, and 8. Recall that for a good model Mallow's Cp statistic should be close to the number of predictors in the model. Figure 1 displays a graph of the Mallow Cp score for multiple candidate models. We can see here that the model with 8 predictors has the lowest Cp score. However, the Cp score for the model with 8 predictors is 15.83. While this is the lowest of the candidate models, it is still rather high – indicating potential bias in the model.

```
## # A tibble: 8 x 2
##        V1 predictors
##     <dbl>      <dbl>
## 1 3247.           1
## 2 1142.           2
## 3  521.           3
## 4  295.           4
## 5  154.           5
## 6   53.3          6
## 7   26.9          7
## 8   15.8          8
```

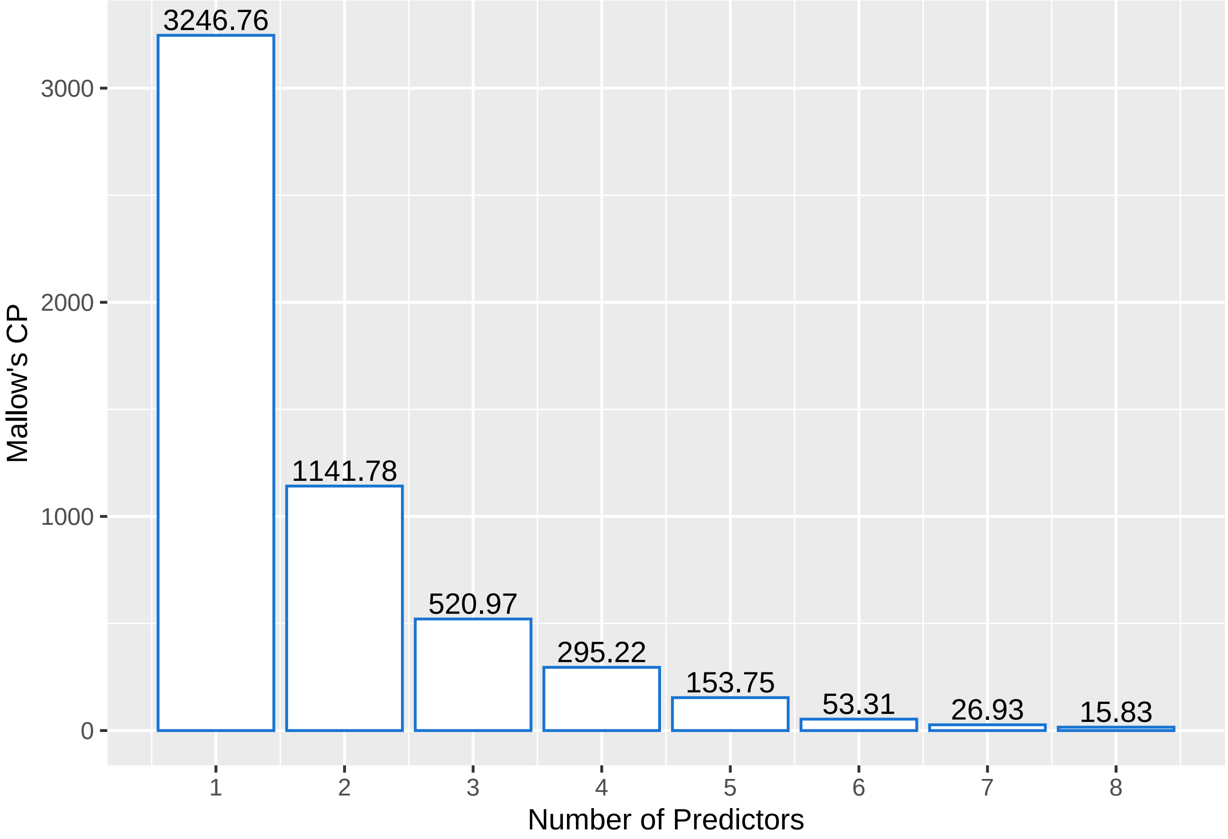


Figure 1 - Mallow's Cp Statistics For Candidate Models

Comparing The Two Previous Models

Table 1 displays the outputs from our two models selected by an AIC and Mallow's Cp based criteria. The purpose of this section is to choose a better candidate model (regression diagnostics and parameter interpretations will be performed on the final model later). The AIC based model has an adjusted R^2 of 0.646 compared to an adjusted R^2 of 0.645 for the Mallow's Cp based model. The AIC based model had an F Statistic of 980.123 (df=9 and 4830,p<0.01). The Mallow's based model model had an F Statistic of 1,100.391(df=8 and 4831,p<0.01).

Overall both models seem relatively comparable. Even though the adjusted R^2 is 0.001 higher in the model that the AIC based selection criteria picked compared to the forward selection model, we elect to continue with the model picked using Mallow's Cp statistic. This is because an increase in 0.001 is not significant. Therefore, to make our model less complicated, we will drop Feature 2 from the covariates.

Table 1: AIC Selected Model Versus Forward Selection Model

	Overall Price	
	AIC	Forward Selection
	(1)	(2)
Mileage (SE)	-0.040^{***}	-0.040^{***}
	(0.002)	(0.002)
Engine Power (SE)	106.064^{***}	106.126^{***}
	(2.614)	(2.616)
Age (SE)	-904.638^{***}	-933.898^{***}
	(38.526)	(36.932)
Feature 1 (SE)	$1{,}568.382^{***}$	$1{,}694.457^{***}$
	(178.589)	(172.239)
Feature 2 (SE)	574.246^{***}	
	(216.762)	
Feature 3 (SE)	$1{,}034.047^{***}$	$1{,}065.750^{***}$
	(210.687)	(210.478)
Feature 4 (SE)	$2{,}788.027^{***}$	$2{,}807.179^{***}$
	(222.716)	(222.738)
Feature 6 (SE)	657.731^{***}	688.364^{***}
	(190.584)	(190.352)
Feature 8 (SE)	$1{,}794.148^{***}$	$1{,}831.096^{***}$
	(185.703)	(185.293)
Constant	$9{,}549.985^{***}$	$9{,}956.149^{***}$
	(396.309)	(365.679)
Observations	4,840	4,840
R^2	0.646	0.646
Adjusted R^2	0.646	0.645
Residual Std. Error	5,466.282 (df = 4830)	5,469.686 (df = 4831)
F Statistic	980.123*** (df = 9; 4830)	1,100.391*** (df = 8; 4831)
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 2: Coefficient-Level Estimates For Full Model

Predictor	Beta	SE	t	p
Mileage	9956.15	365.679	27.23	0.00000000
Engine Power	-0.04	0.002	-25.58	0.00000000
Age (SE)	106.13	2.616	40.57	0.00000000
Feature 1	-933.90	36.932	-25.29	0.00000000
Feature 2	1694.46	172.239	9.84	0.00000000
Feature 3	1065.75	210.478	5.06	0.00000043
Feature 4	2807.18	222.738	12.60	0.00000000
Feature 6	688.36	190.352	3.62	0.00030192
Feature 8	1831.10	185.293	9.88	0.00000000

Full Model (including Model Series)

Multiple linear regression was performed using the model parameters in Table 2 to predict the price of each car. Our model revealed a statistically significant linear association between the predictors in Table 2 and price (F = 1,100, df = 8 and 4831, p < $2e^{-16}$). Our model had an adjusted R^2 of 0.6451. This means that, when adjusting for the number of predictors, our model captures about 64.51% of the variation in price.

Insert interpretations for beta estimates – also note that these are only valid if our model assumptions are met, these will be examined in the next section

Model Performance

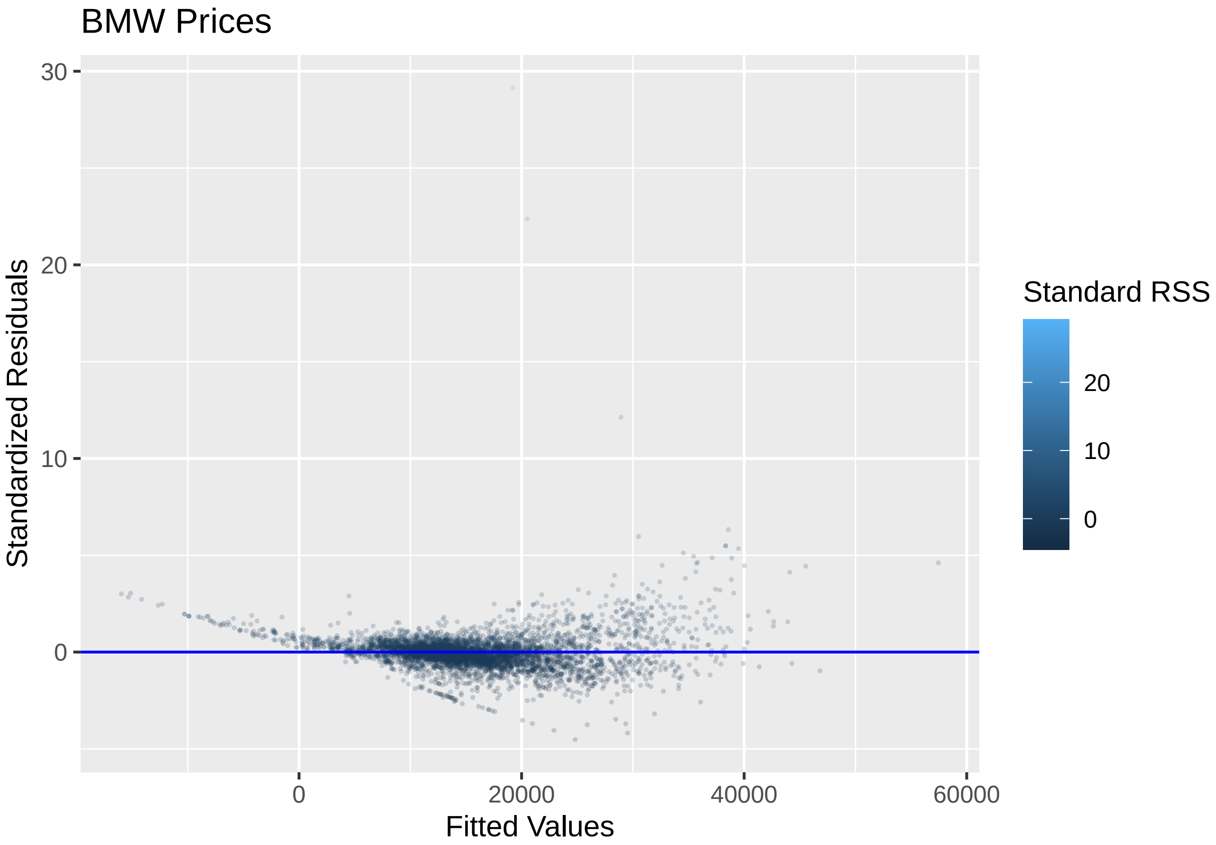


Figure 2 - Standardized Residuals vs Fitted Values

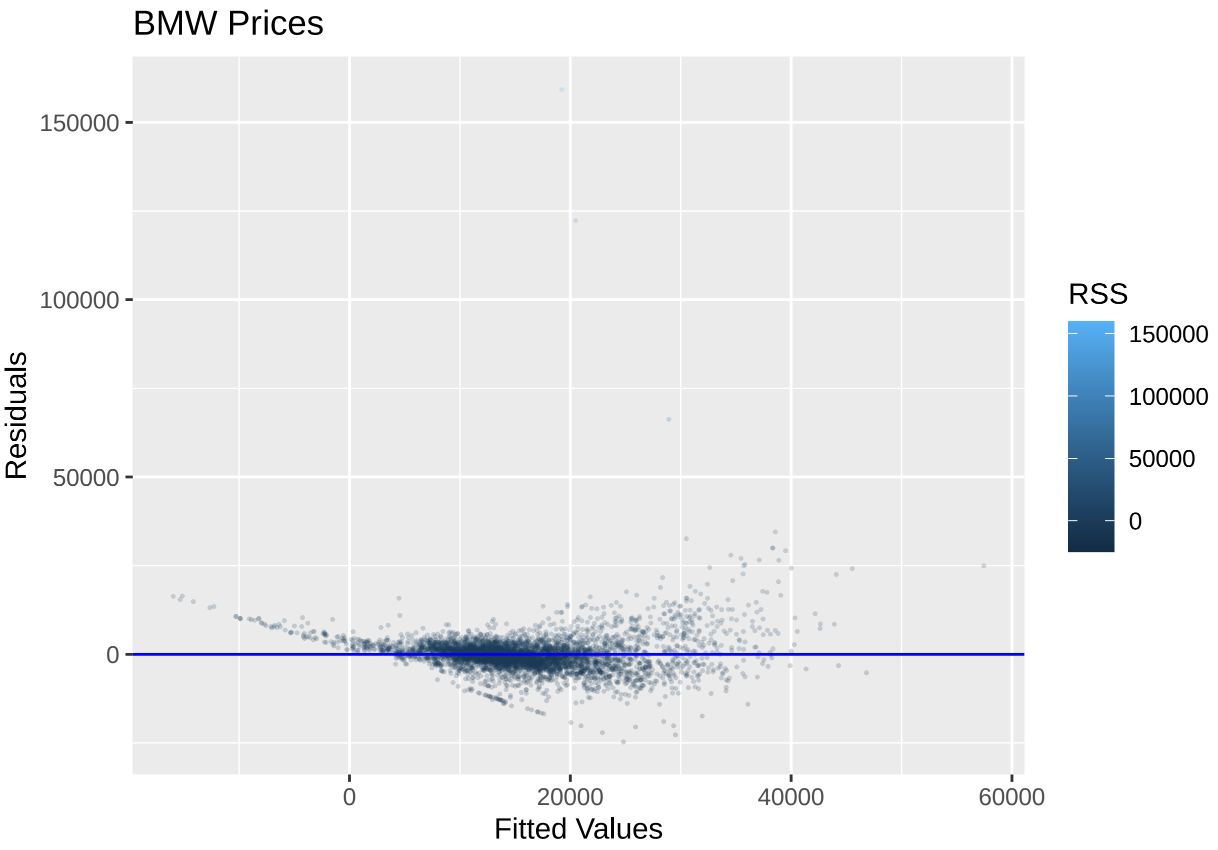


Figure 3 - Residuals vs Fitted Values

Add in interpretation here

	VIF
mileage	1.360321
engine_power	1.678732
age	1.421077
feature_1	1.188139
feature_3	1.155562
feature_4	1.277203
feature_6	1.073222
feature_8	1.379316

Table 3 - Variance Inflation Factors for Full Model

In table 3 we examine the variance inflation factors (VIF) for the full model. Recall that the VIF for an individual predictor variable in a multiple regression analysis measures the inflation in the variance in the estimates of the slopes that is due to collinearities in the set of predictor variables. Values over 10 should be considered indicators of collinearity. We can see that none of our predictors have VIF greater than 10 and thus we can safely conclude that our model satisfied the assumption that our predictors are independent one another.

Conclusions

Next Steps

Appendix