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.

Data Exploration

Need a simpler way to represent this – it's too big right now and doesn't render well. One idea is to stratify this command into smaller combinations

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 information about the 'series' of each car. Based on BMW's criteria for their model series, we categorized the model key data into 11 categories, including 1-7 series, M series, X series, Z series, and i series. Within the 1 to 7 series, we observed that cars with higher model numbers tend to command higher prices. The M series, which represents the highest-performance edition of each series, is the most expensive model. The X series represents a line of SUV vehicles and crossovers(BMW also refers to them as Sports Activity Vehicles). The Z series, which stands for "future" in German, contains models such as roadster, coupé, sports car, and concept variants. The i series models are new energy vehicles(electric vehicles and hybrid vehicles). We hypothesize that M series car will result in the highest sale price since M series models are the highest-performing vehicles, while higher model numbers in each letter series will result in higher sale price of a car. We represented this as a dummy variable in our data. Model series 3 was used as an arbitrary reference category. Because the model series variable had 11 levels, we decided to exclude it from our model selection procedure. The reason for this was because it was likely that certain levels would be excluded while others would not. This creates challenges from an interpretation point of view. We believe that the series of the car is an integral part in determining the value of the car. Therefore, we elected to force its inclusion in the final model.

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.

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 11 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, whether or not they used new energy (as defined before), whether or not they were a common color (as defined before), 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 8 covariates: mileage, engine power, age of the car, whether or not they used new energy (as defined before), and whether or not they had features 1,3,4, 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 30.56. 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 3354.           1
## 2 1221.           2
## 3  592.           3
## 4  363.           4
## 5  220.           5
## 6  118.           6
## 7   58.6          7
## 8   30.6          8
```

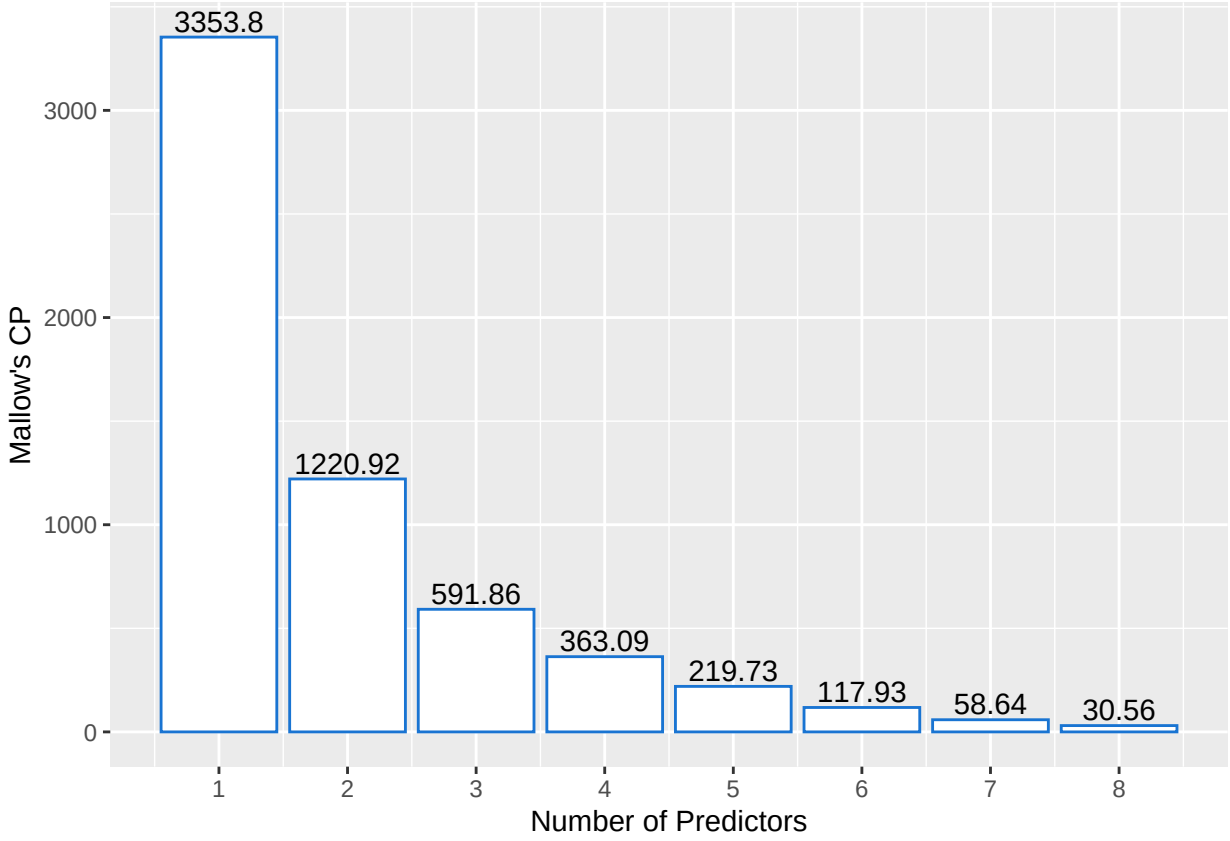

Figure 1 - Mallow's Cp Statistics For Candidate Models

Comparing The Two Previous Models

Table 1 displays the outputs from our two models selected by an AIC and Mallow's Cp based criteria. Note this does not yet include the dummy variable for Model Series. The purpose of this section is to choose a better candidate model (regression diagnostics and parameter interpretations will be performed on the final model including Model Series later). The AIC based model has an adjusted R^2 of 0.651 compared

Table 1: AIC Selected Model Versus Forward Selection Model

	Overall Price	
	AIC	Forward Selection
	(1)	(2)
Mileage (SE)	−0.040***	−0.039***
	(0.002)	(0.002)
Engine Power (SE)	106.197***	107.789***
	(2.599)	(2.578)
Age (SE)	−903.904***	−941.768***
	(38.271)	(36.673)
New Energy? (SE)	13,734.440***	13,631.530***
	(1,723.639)	(1,726.486)
Common Color? (SE)	297.482	
	(183.350)	
Feature 1 (SE)	1,556.070***	1,702.362***
	(177.622)	(171.151)
Feature 2 (SE)	571.131***	
	(215.396)	
Feature 3 (SE)	1,056.347***	1,143.613***
	(209.314)	(208.950)
Feature 4 (SE)	2,786.683***	2,853.193***
	(221.290)	(221.367)
Feature 6 (SE)	694.311***	
	(189.427)	
Feature 8 (SE)	1,734.181***	1,768.723***
	(184.660)	(184.456)
Constant	9,242.692***	9,844.363***
	(412.413)	(364.128)
Observations	4,840	4,840
R^2	0.651	0.649
Adjusted R^2	0.650	0.649
Residual Std. Error	5,430.087 (df = 4828)	5,442.084 (df = 4831)
F Statistic	818.700*** (df = 11; 4828)	1,117.722*** (df = 8; 4831)
Note:		*p<0.1; **p<0.05; ***p<0.01

to an R^2 of 0.650 for the Mallow's Cp based model. The AIC based model had an F Statistic of 818.700 (df=11,p<$2.2e^{-16}$). The Mallow's based model model had an F Statistic of 1,117.722(df=8,p<$2.2e^{-16}$).

Overall both models seem relatively comparable. The AIC model has a slightly larger adjusted R^2. Also, looking at the variables excluded in the Mallow's based model, we can see, with the exception of common color, all predictors were found to be significant in the AIC based model. Because of this, we elected to continue forward with the AIC based model.

Table 2: Coefficient-Level Estimates For Full Model

Predictor	Beta	SE	t	p
Intercept	9821.40	459.904	21.36	0.00000
Mileage	-0.04	0.002	-23.26	0.00000
Engine Power	93.59	2.924	32.01	0.00000
Age	-934.57	37.402	-24.99	0.00000
New Energy?	7343.70	3053.516	2.40	0.01621
Common Color?	161.70	177.972	0.91	0.36363
Feature 1	1233.99	177.439	6.95	0.00000
Feature 2	739.23	211.901	3.49	0.00049
Feature 3	898.57	204.028	4.40	0.00001
Feature 4	1490.68	241.083	6.18	0.00000
Feature 6	999.74	187.676	5.33	0.00000
Feature 8	1485.56	183.797	8.08	0.00000
Model Series 1	-363.12	268.590	-1.35	0.17645
Model Series M	3997.29	1132.823	3.53	0.00042
Model Series 4	4431.11	548.334	8.08	0.00000
Model Series Z	234.24	2173.956	0.11	0.91420
Model Series 2	-203.61	756.441	-0.27	0.78781
Model Series 6	5985.44	1017.583	5.88	0.00000
Model Series i	9812.23	3656.505	2.68	0.00731
Model Series 5	1048.45	223.624	4.69	0.00000
Model Series X	3675.59	243.017	15.12	0.00000
Model Series 7	6111.45	768.518	7.95	0.00000

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 = 473.1, df = 21 and 4,818, p $< 2e^{-16}$). Our model had an adjusted R^2 of 0.672. This means that, when adjusting for the number of predictors, our model captures about 67.2% 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

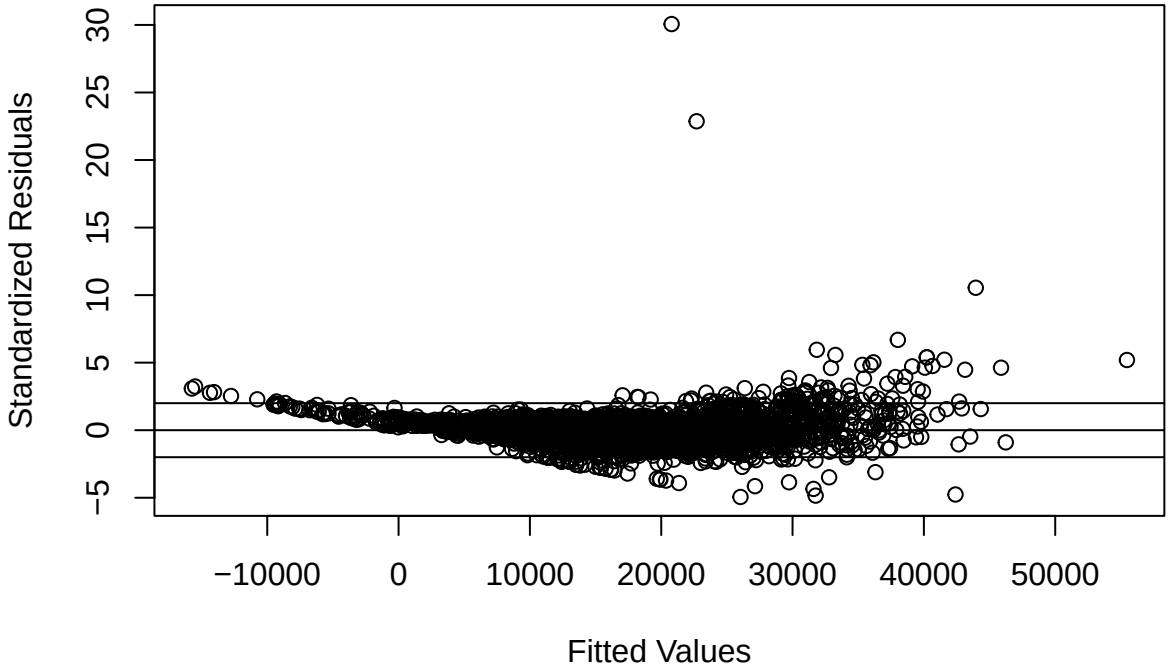

Figure 2 - Standardized Residuals vs Fitted Values

Add in interpretation here

Conclusions

Next Steps

Appendix