Inference and Modeling of Car Price in Ukraine

Team member: Steven Xu, Xiaolin Cao, You Sen Wang, Jiarui Yang
Team leader: Steven Xu

Introduction:

Car price has always been an interesting topic in the area of data analysis. Most people, without much hesistation, would claim that a car's price is in direct proportion to its recency and luxury. With that being true, however, there are certainly other factors that determine the value of a car. Moreover, we want to know how exactly, in a quantitative sense, each factor affects the price.

The dataset we are using comes from the well-known crowdsourcing analytics platform Kaggle and is collected from private car sale advertisements in Ukraine. It contains information of 9576 samples and 10 variables. Of the 10 variables 7 are categorical, which include the manufacture brand (*car* e.g. Ford, BMW, etc), the car body type (*body* e.g. crossover, sedan, etc), the type of fuel (*engType* e.g. Gas, Diesel, etc), registration in Ukraine (*registration* e.g. yes or no), year of production (*year* e.g. 2013, 2009, etc), the specific model name (*model* e.g. X5, Passat B5, etc) and drive type (*drive* e.g. front, rear, full, etc). The 3 continuous variables are the mileage mentioned in advertisement (*mileage*, in $10^3 km$), sellers's price in advertisement (*price*, in USD) and the rounded enginve volume (*engV*, in $10^3 cm^3$). Our goal is to use various resampling and regression method to analysize the data and make meaningful inference on the relation of the variables. In specific, we want to regress car price on the other variables to generalize a model that can be used for prediction task.

After looking at the data from Ukrainian car sale advertisement, we decided to focus on four questions regarding this dataset. First of all, we want to know what the best model to predict the price of car using other variables is. In order to solve this question we will have to start by examining whether each variable has any influence on the car price, which can be found out by setting multiple hypothesis tests. After eliminating the relatively meaningless varaibles we want to find out the marginal importance of each variable on the car price, after all it is very likely that a couple of the variables will have major impacts on the response with others having minor impacts. Assuming linear regression is sufficient, this will be done by estimating the regression coefficients and comparing the values. Thirdly, we want to study the decreasing of car price over time. When thinking about decreasing over time, it is intuitive to consider the exponential distribution for its monotonously decreasing characteristic. We will pick and focus on a fixed model of car to see whether the exponential distribution appropriately depicts this relation or not. Lastly, we want to use regression bootstrap and cross-validation to test the feasibility of our model and increase its accuracy.

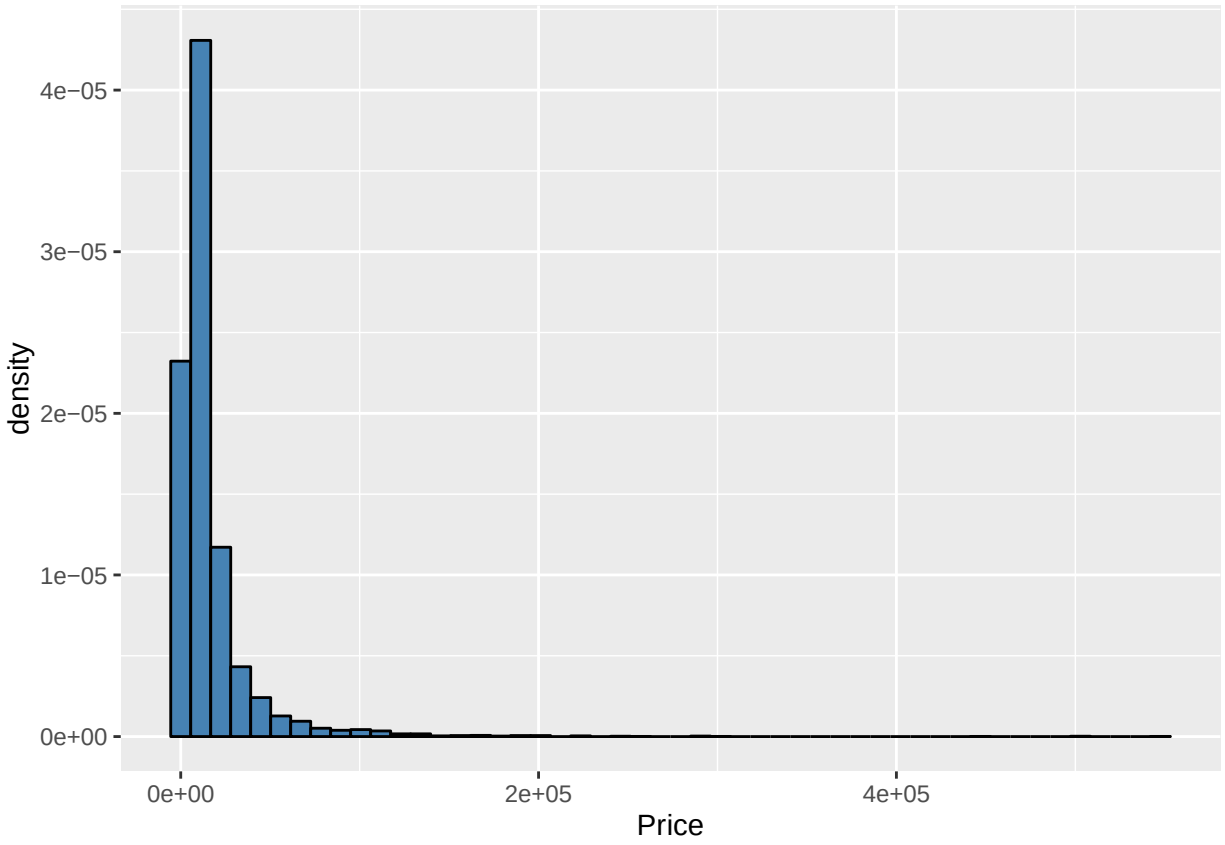
As a side notice, since our data set is a real raw data set, entries with missing values exsist. For example, price could be recorded as zero and rounded engine volume (*engV*) as NA. We will omit the rows of data that contain missing values and decide whether we should replace prices that are zero with appropriate values or omit them. Also, we changed the text of car models that contain Ukranian since R will throw warnings when reading them.

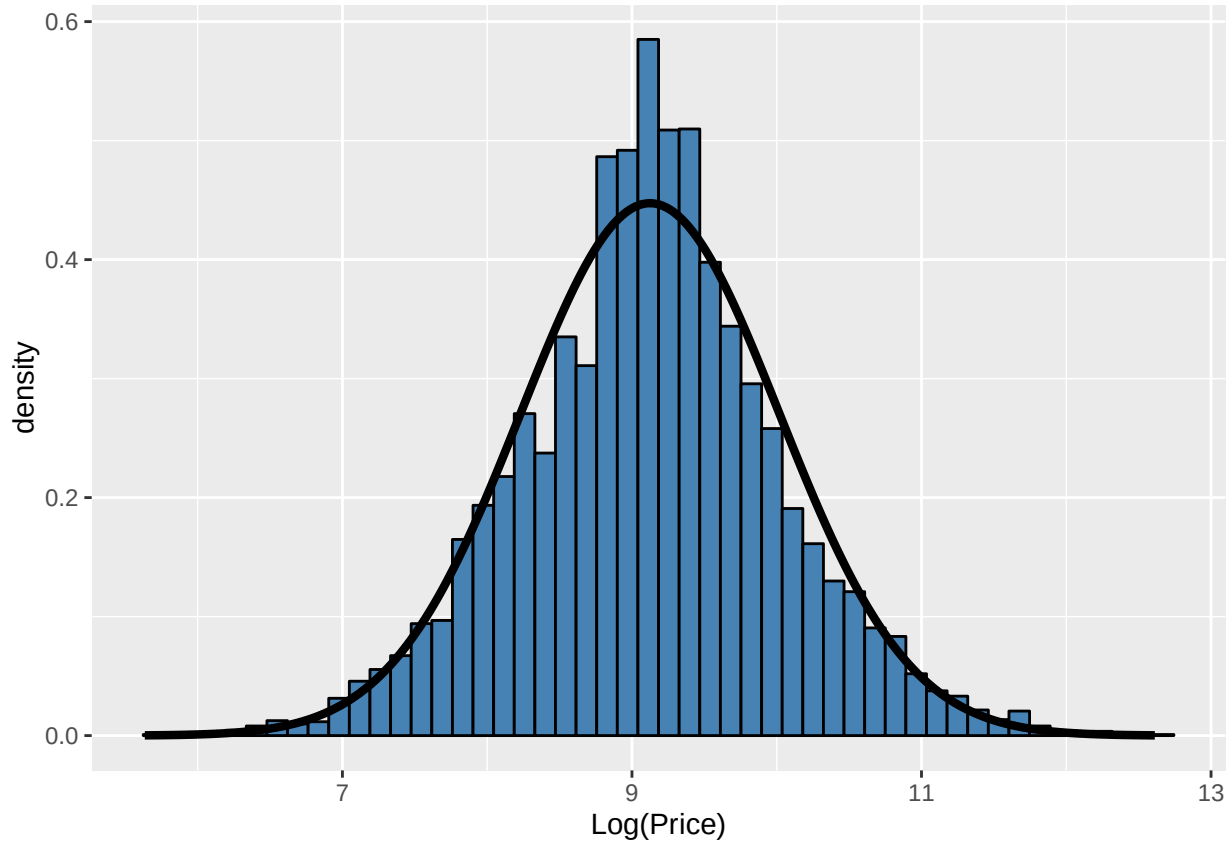
Below we showed a small portion of our dataset to give an idea to readers who are interested.

car	price	body	mileage	engV	engType	registration	year	model	drive
Ford	15500	crossover	68	2.5	Gas	yes	2010	Kuga	full
Mercedes-Benz	20500	sedan	173	1.8	Gas	yes	2011	E-Class	rear
Mercedes-Benz	35000	other	135	5.5	Petrol	yes	2008	CL 550	rear
Mercedes-Benz	17800	van	162	1.8	Diesel	yes	2012	B 180	front
Mercedes-Benz	33000	vagon	91	NA	Other	yes	2013	E-Class	NA
Nissan	16600	crossover	83	2.0	Petrol	yes	2013	X-Trail	full

Data preprocessing:

As we mentioned above our dataset is a real-world raw dataset, therefore it might be incomplete and contains a lot of errors. By browsing the dataset, we found out that lots of entries are left blank or marked as NAs. Since R has a built-in function that omits NAs, we assigned NAs to all blank entries in order to get rid of them. The dataset also contain samples that have unreasonable values at several variables. For example, an engine volume of 99 L is definitely unreasonable for ordinary cars. Samples with unreasonable value of variables are considered as outliers and are excluded from modeling. The distribution of response variable is also an important thing to look at since a highly skewed distribution will make the mean a bad measure of central tendency, thus making ordinary least squares regression rather inappropriate. In our case, we applied a log-transformation on the response variable – *price* to reduce the skewness. The plots below show the distribution of *price* before and after log-transformation. A normal curve based on the mean and variance of price is also plotted to compare with the histogram. We can see that $Log(price)$ has an approximately normal distribution.





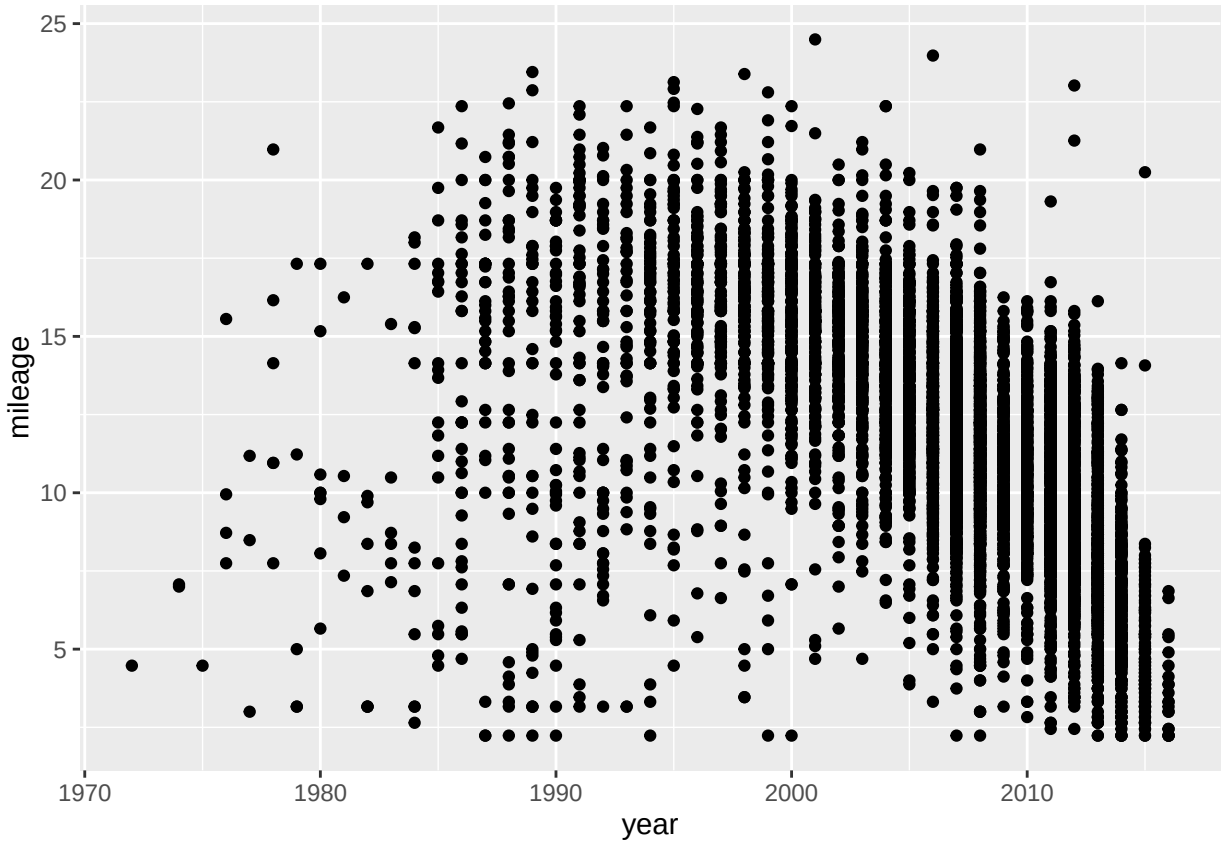
Feature Importance:

Assuming a generalized linear regression would be appropriate, we startd by checking whether each feature by itself has strong impact on the response. We use training MSE and R^2 as the evaluation criterion and we wish to find out the most influential variable. The training MSE gives information about how well the trained model fits the training data, i.e. the bias of our model. Lower MSE indicates lower bias. R^2 tells us the percentage of variation of response variable explained by the variation of the fitted regressors. Their formulas are given below. An important notice is that in calculating the feature importance we exclude the variable *model*, because as we know the model of a car almost exclusively determines its price. Also as a factor with hundreds of levels, *model* is highly dependent to car brand, and different car brands rarely share the same model. Therefore we think that the variable *model* should not be used in modeling.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$$

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

We also want to check if there exists high collinearity between any regressors. If strong collinearity exists, then it would create redundancy in the model since they are just the same thing presented in different ways, and we would need to decide whether we should eliminate any regressors. A simple way to detect collinearity is to look at the scatterplots. Using this method we suspect that the variable *year* and *mileage* might have strong collinearty. This is intuitive since older cars tend to be used longer thus having higher mileage.



To make a more accurate decision, we decide to use Variance Inflation Factor(VIF) as a criterion for high collinearity. VIF measures how much the variance of the estimated regression coefficient is inflated by collinearity. Generally, a VIF greater than 2 indicates significant collinearity.

	GVIF	Df	GVIF^(1/(2*Df))
car	13.711538	82	1.016093
body	6.825088	5	1.211744
mileage	1.937179	1	1.391826
engV	2.342930	1	1.530663
engType	2.017810	3	1.124122
registration	1.184438	1	1.088319
year	1.900186	1	1.378472
drive	8.320449	2	1.698387

From the last column we can see that none of the variables have strong collinearity, therefore we can now compare the individual impact that each variable has on the response.

	car	body	mileage	engV	engType	registration	year	drive
Training MSE	0.49	0.63	0.68	0.57	0.78	0.73	0.43	0.59
R squared	0.39	0.21	0.15	0.29	0.02	0.09	0.46	0.26

We can see from the table that the variable *year* has the lowest training MSE and highest R squared, making it the most impactful variable when doing univariate linear regression.

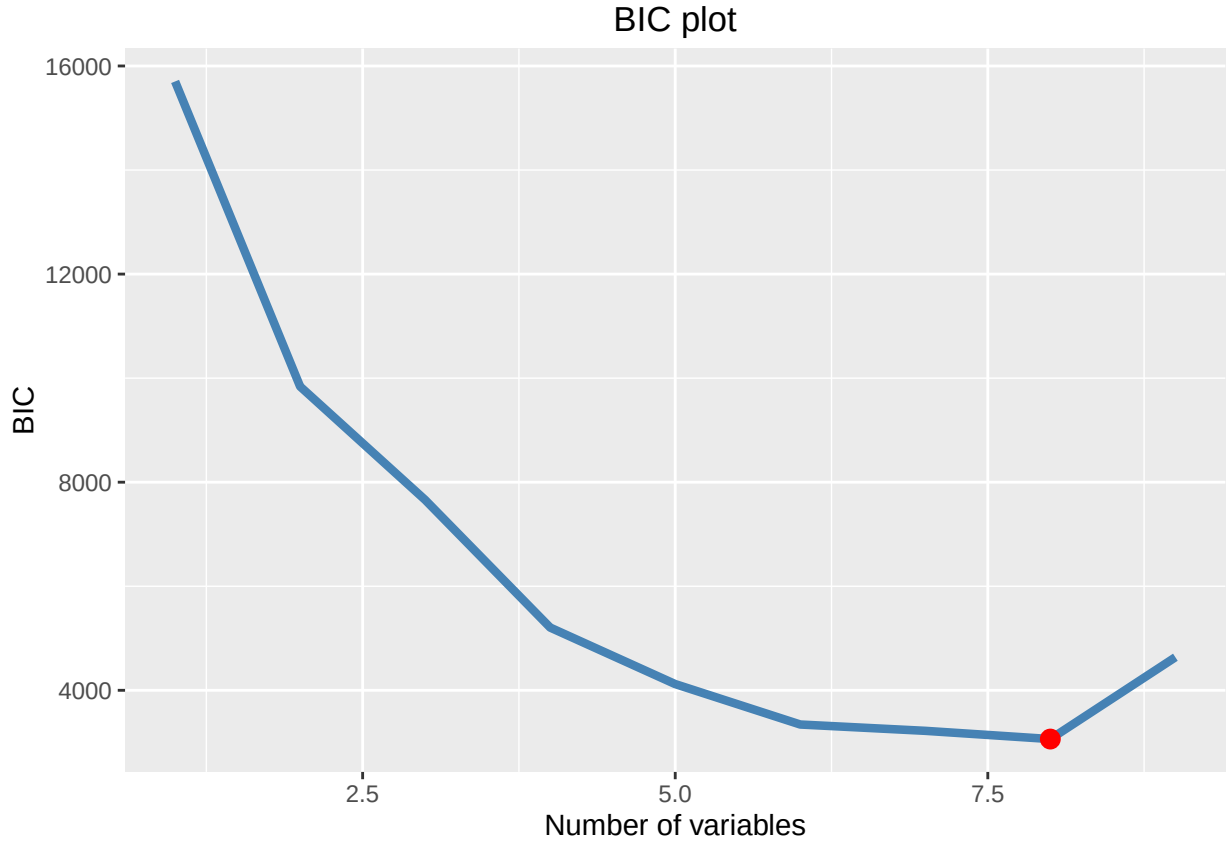
Model Selection:

Our goal is to create a multivariate regression model that best characterizes the relationship between price and other variables, therefore we are more interested in the combined rather than individual effect of the regressors. Thus we need a model selection method to help us determine the best combination of the variables . There are many algorithms used to compare the performance of different combinations of variables such as calculating the AIC, BIC or Cross-Validation. For our first model we will use BIC to determine the best model.

Bayesian information criterion(BIC) is a criterion for model selection that introduces a penalty term to prevent overfitting when fitting a dataset. As we know although as the model gets more complex the training error will decrease continuously, the testing performance will go down at some point. This problem is referred to as the Bias-Variance Tradeoff. Since we are more interested in prediction and interpretation we prefer a model with both low bias and variance. In this case the model with the smallest BIC is favored. We built a nested for-loop that goes through every combination and calculate the BIC. The lowest BIC with its corresponding variable combination is recorded.

$$BIC = \frac{1}{n}(RSS + log(n)d\hat{\sigma}^2)$$

In the formula above, n denotes number of samples, d denotes number of regressors and $\hat{\sigma}^2$ denotes the estimated variance of residual error.



In calculating the BIC we included the variable *model* to further demonstrate how it affects the performance of the model in a negative way. From the above plot we can see that the best model contains 8 variables, which are all variables except *model*. Thus the model we are using is:

$$log(\hat{price}) = \hat{\beta}_0 + \hat{\beta}_1 car + \hat{\beta}_2 body + \hat{\beta}_3 mileage + \hat{\beta}_4 engV + \hat{\beta}_5 engType + \hat{\beta}_6 registration + \hat{\beta}_7 year + \hat{\beta}_8 drive$$

Note that the formula of this model is not that rigorous. Because many of the variables are factors and factors are transformed into numbers of dummy variables each assosciating a coefficient. This means that a factor with 10 levels will have 10 coefficients. Therefore the above formula is only a simplified representation of the true model.

Reduced Model

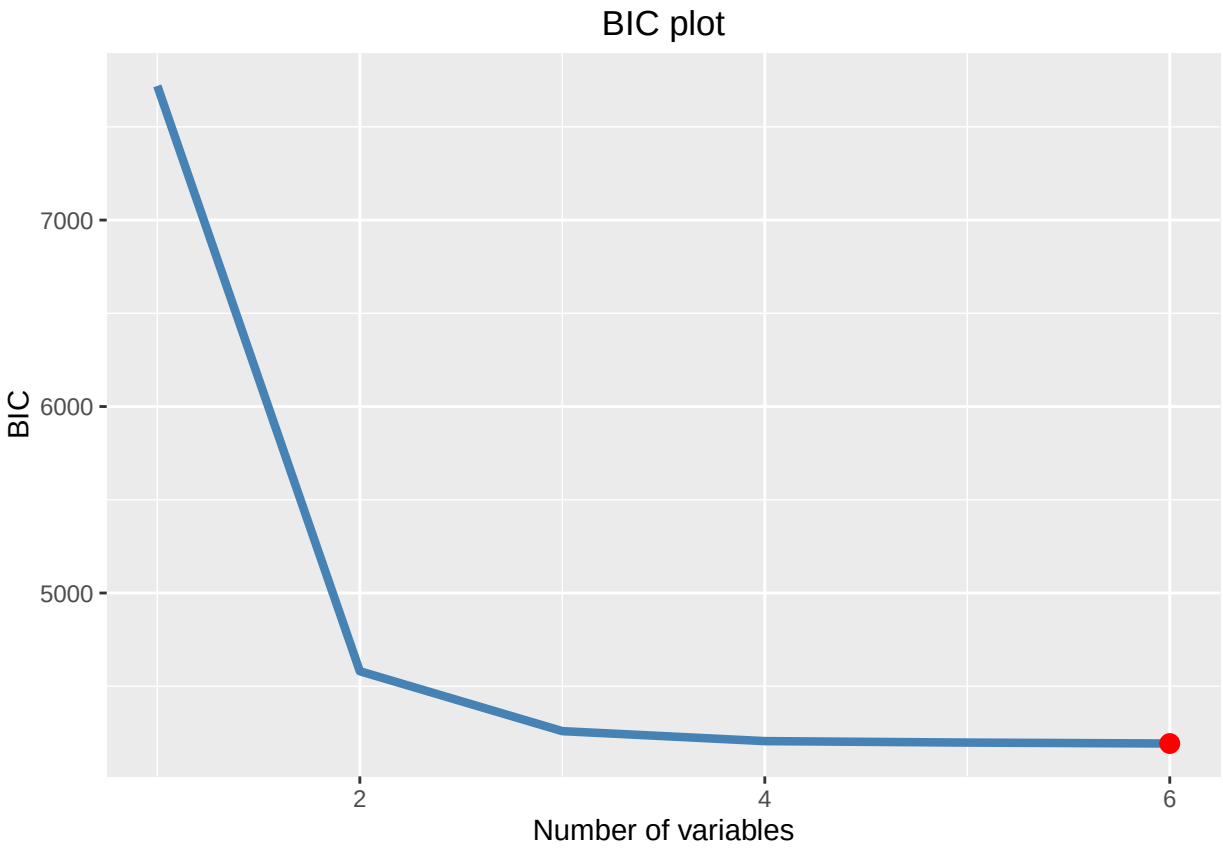
The first model was fitted into the whole dataset with little constraint. For our second model we wish to focus on a more specific subset of the original dataset and build a more generalized and reduced model. The constraints are set as below.

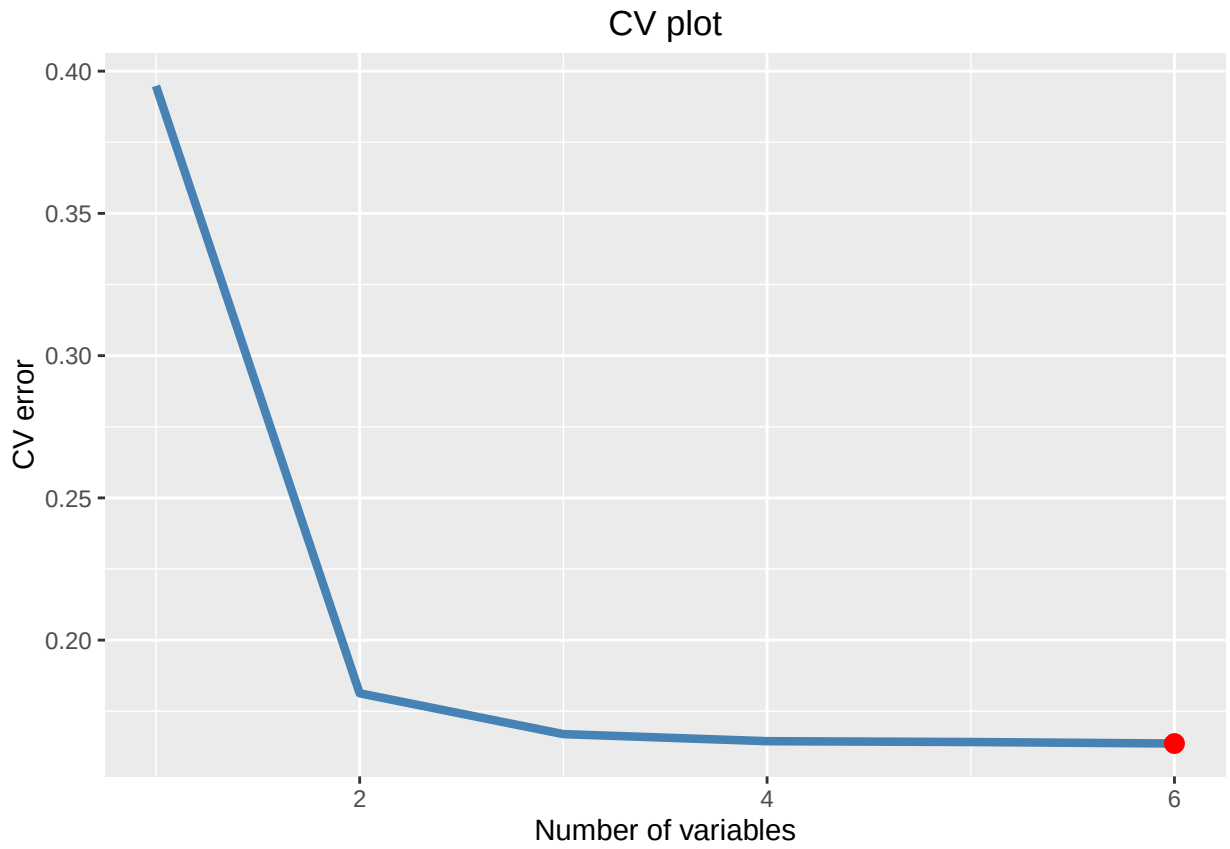
- Body: not van or vagon
- Registration: yes
- Engine Type: Gas or petrol
- Drive: Front or full

We also exclude the variable *car* because we want to create a model for price prediction without knowing the exact brand of the car. This reduced dataset is called Family Car because we think that it presents the most common family cars.

Again we want to determince the best model that characterizes the relationship between price and the other variables. However, this time we will use both BIC and Cross-Validation as the model selection approach since we are curious about whether they yield the same result.

Cross-validation is another model selection method that partitions the dataset into training and testing datasets. The model is fitted onto the training set and MSE is calculated by comparing the predicted values and the testing set. There are many kinds of Cross-validation and we chose 10-fold cross-validation, where the dataset is partitioned into 10 folds. Each iteration one fold is randomly selected as the testing set and the MSE is calculated. The average MSE is taken into comparison.


BIC plot
BIC
7000
6000
5000
2
4
6
Number of variables



From the above two plots we can see that Cross-Validation and BIC yield the same result. The result indicates that when all the variables are included, the model performs best.

$$\hat{log(price)} = \hat{\beta}_0 + \hat{\beta}_1 body + \hat{\beta}_2 mileage + \hat{\beta}_3 engV + \hat{\beta}_4 engType + \hat{\beta}_5 year + \hat{\beta}_6 drive$$

Again this is only a simplified representation.

Bootstrapping

Now we have our model and we want to test its validity. A simple method to test the validicty of the model is Bootstrapping, which simulates new samples by sampling with replacement from the original sample. If this model is indeed valid, then the estimated coefficients should fall in the the bootstrapped confidence interval.

	5%	95%
Intercept	-167.0391997	-134.5327721
bodycrossover	-0.0208969	0.2240048
bodyhatch	-0.1749025	-0.0546812
bodyother	-0.0479700	0.1069615
mileage	-0.0360812	-0.0111263
engV	0.1861749	0.2847651
year	0.0716533	0.0877492
drivefront	-0.2290835	0.0138296
engTypePetrol	-0.0192846	0.0636910

```
##   (Intercept) bodycrossover     bodyhatch     bodyother       mileage
## -151.96811050    0.11185898   -0.11317095    0.03178960   -0.02296722
##          engV          year    drivefront engTypePetrol
##    0.22758651    0.08028883   -0.10233603    0.02302731
```

We can see that the estimated coefficients are indeed contained in the bootstrapped confidence intervals, which proves the model's validity.

Another question we are interested in is whether we can use the data of a representative brand to sufficiently charaterize the pattern in Family Car. We chose Honda as our representative brand. Since Honda has no more than 200 samples, it is extremely small compared to the Family Car. Therefore we will use bootstrapping to create simulated samples and use them to build confidence intervals of the coefficients. We will then compare the estimated coefficients when fitting Family Car to the confidence intervals. If the confidence intervals do contain the estimated coefficients then we will have evidence to say that Honda sufficiently charaterize the pattern in Family Car.

```
##   (Intercept) bodycrossover     bodyhatch     bodyother       mileage
## -1.421454e+02  9.746834e-02 -3.061617e-02  1.077734e-01 -9.471276e-03
##          engV          year    drivefront engTypePetrol
##  3.790918e-01  7.505803e-02 -2.693172e-01  1.084696e-01
```

We can see that some of the estimated coefficients do not fall in the bootstrapped confidence intervals. This proves that the Honda data itself can not sufficiently characterize the pattern of Family Car.

Conclusion

In this project we successfully built a model that quantitatively depicts the relationship between price and other factors of car of data collected from Ukrainian market. Using model selection approaches such as calculating BIC and Cross-validation we are able to find the optimal combination of variables that creates the model. Using resampling method like bootstrapping we are able to validate our chosen model and show whether a specific subset can sufficiently characterize the pattern of the whole dataset. This is not a perfect data analysis project as there are many improvements we could make. In the future we could use more rigorous tests such as Sequential F test to test whether adding or removing a variable varies the performance of the model. We also didn't take in account of interaction effects which migh possibly exsist in our dataset. Moreover, our project focuses largely on the inference part and lacks prediction tasks. In the future we should come up with questions that allow us to show more straightforwardly the prediction power of our model.