Department of Statistical Sciences

Methods of Applied Statistics I

Final Project

Nigel Petersen

Fall 2022

1 Non-Technical

1.1 The Problem of Interest

Over the last several decades, there have been many significant advancements in our society, and as a result, the cost of living has drastically increased. The increased cost of living, together with a lack-luster increase in wages, has forced a larger proportion of people into a tougher financial situation. With vehicles and insurance among the top expenses in the modern day, many have looked beyond the traditional notion of purchasing a new or pre-owned vehicle from a dealership, and have taken a more digital approach. Buying vehicles online has become increasingly popular over the last decade, and this rise in popularity has introduced the need to accurately estimate the price of a used vehicle listed for sale online. As with most online purchasing of used objects, there is always an underlying risk of over-paying, and the ability to estimate the value of a particular used vehicle, given several of its features, is one possible way to avoid paying more than necessary. Conversely, when listing an item for sale online, it is often quite useful to have a sense of its value in the current market. With cars in particular, where there are potentially many factors that play a role in its value, the ability to accurately estimate the price of the vehicle, knowing information about its features, is very beneficial.

1.2 How and Why the Data was Collected

The data set was sourced from Kaggle [Lepchenkov, 2019], a popular data set website. The data itself was sourced from online listings for used cars that appeared across several of the most popular listing sites in Belarus in late 2019. The primary motivation for the collection of the data was to address the problem of interest introduced in the previous section, namely the problem of buying and or selling a used vehicle online for a reasonable price.

1.3 Preliminary Description of the Data

For terminology sake, refer to the data as obtained directly from Kaggle [Lepchenkov, 2019] as the "raw data". The raw data consists of 37679 observations, on 29 variables. The response of interest is `price-usd`, which represents the price, in US dollars, of a vehicle listed for sale online. The raw data consists of a mix of both continuous and categorical variables, with the latter more commonly occurring, and among the categorical variables, there is a mix of binary and multi-level factors. Upon obtaining summary statistics for the response, the minimum value listed was \$1, which seemed rather strange, so all (11) observations with a listed value less than \$100 were removed when constructing the working data set. The entirety of the

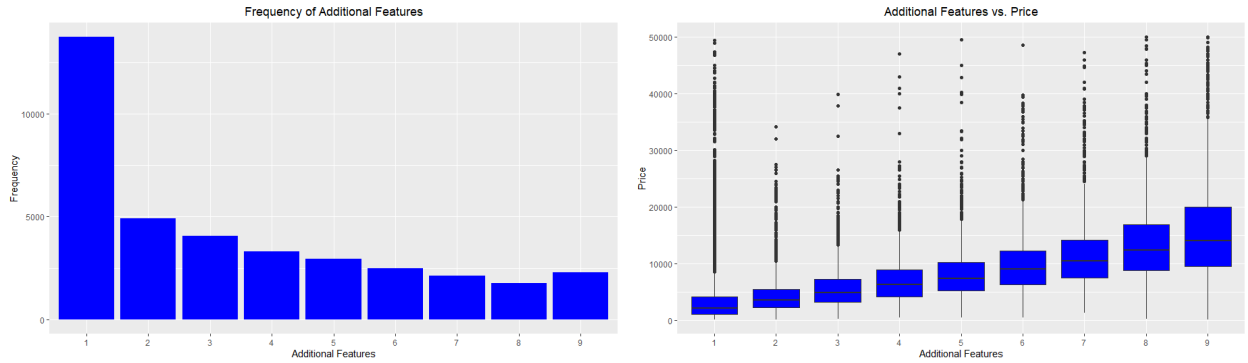
data-cleaning process is outlined in the Appendix. The cleaned data, which will be referred to as simply the data, consists of 37660 observations on 17 variables. There are 6 continuous variables and 11 categorical variables, which include commonly considered characteristics like `odometer-value` (in kilometers), `year-produced`, `manufacturer-name`, `engine-capacity` (in litres), `body-type`, `color` and `engine-type`, among others. As the data was sourced from European listing websites, the features of the vehicles listed can potentially be unique to the European market. Most notably, there is a larger proportion of European and Asian manufacturers present compared to those seen in listings in North America, and the proportion of manual transmission vehicles is significantly higher than that of North America. Regardless of location, there are common variables that seem immediately influential on the price of a vehicle. One would expect that features like `year-produced` and `odometer-value` are among the most influential, and Figure 1 shows these two variables plotted against the response.

Figure 1: `odometer-value` and `year-produced` variables plotted against `price-usd`

For the `odometer` vs. `price-usd` plot, the range for the `odometer-value` variable has been restricted to a maximum value of 800000 (as opposed to the original maximum of 1000000) for better visibility, and as the proportion of observations with `odometer-value` exceeding the plotted maximum is approximately 0.006. Similarly, `year-produced` is plotted beyond 1980, with the proportion of observations produced prior to 1980 approximately 0.001. The shapes of the above plots against the response, particularly for `year-produced`, convey an exponential relationship, suggesting `log(price-usd)` be considered for the response when modelling.

Also among influential variables is `additional-features`, the number of additional features present in a listed vehicle. This variable was coded as a factor with 9 levels, representing the

number of additional features present.



(a) Distribution of `additional-features` (b) `additional-features` against `price-usd`

Figure 2: Analysis of `additional-features`

In Figure 2a, with each increase in additional features, a smaller proportion of observations contain that many additional features. Furthermore, as one would expect, an increase in additional features present would lead to an increase in the price of the vehicle, and this behaviour is present in the data as well, as seen in Figure 2b. Additionally, often times when purchasing items from online listings, the length an item has been up for sale can say a lot about its value compared to its listed price. A higher duration of time listed with no purchase can indicate that consumers are not willing to pay what the lister thinks the item is worth. With cars in particular, the listing duration can vary more significantly than other commonly sold items online. Figure 3 displays the relationship between `duration-listed` and `price-usd`.

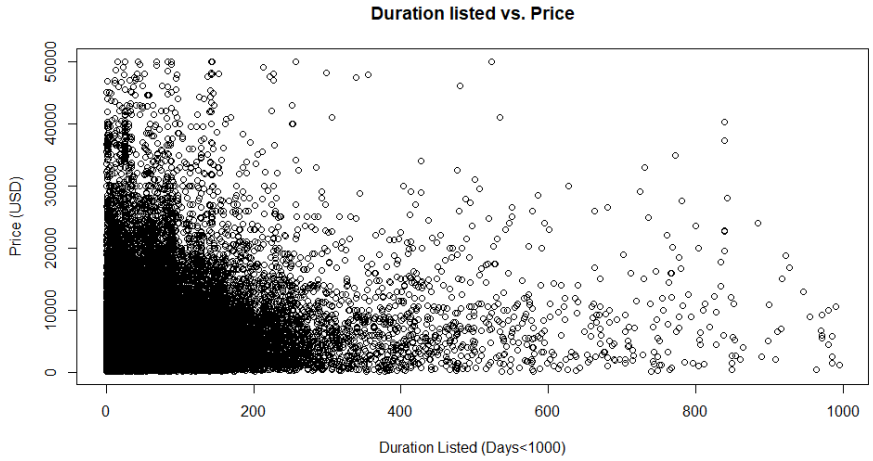


Figure 3: `duration-listed` plotted against `price-usd`

There is a general decrease in `price` with an increase in `duration-listed`, which looks quite vaguely exponential, perhaps further suggesting the use of `log(price-usd)` as the response

when modelling. As with the other plots of continuous variables, the range for `duration-listed` has been reduced for visibility, and as the proportion of observations listed for more than 1000 days is approximately 0.002.

1.4 Summary for the Non-Statistician

One possible approach to gain insight towards the problem of interest stated above is to fit a model on the collected data, and use it to learn information and make predictions. The ability to make predictions, and inference in general, depends directly on the quality of the data and model used. Addressing each of these needs individually, the raw data sourced from Kaggle [Lepchenkov, 2019] was cleaned and reformatted to be easier to use when fitting a model and generally gainining insight. After the data was taken care of, the model selection procedure began with assuming that all of the possible variables included in the data set were important for predicting the value of the variable of interest, `price-usd`. After fitting a basic initial model, a new model was proposed that used all of the variables in the data set, as well as new variables that measure how carefully selected pairs of the original variables influence `price-usd` together. In fitting the new model, statistical test were conducted to determine the optimal model, and together with other measures of performance, the second model introduced, with the additional variables that account for the influence of pairs of variables, was deemed the best fit.

Once the optimal model was fit, it was used to make inference on the relationship between the selected variables and the response variable, `price-usd`. During the initial model selection, it was found that some of the most influential variables in predicting values of the response were `odometer-value`, `year-produced`, `duration-listed` and `additional-features`. Using the selected model, it was found that there was an expected increase of $1543 in average listing price for an increase of a single additional feature. Additionally, for vehicles with an odometer value of at most 400000 kilometres, which account for approximately of 91% of observations in the testing set, an increase of 50000 kilometres on the odometer resulted in an expected $1503 decrease in average listing price. Lastly, for vehicles produced after 1980, which account for approximately 99% of observations in the test set, a decrease of 5 years in the age of a vehicle, namely a 5 year increase in the year produced, resulted in an expected $6029 increase in average listing price.

2 Technical

2.1 Models and Analysis

As prediction performance is of interest here, a 70-30 train-test split was performed on the data prior to any modelling. The model selection process begins by first fitting a main-effects model that contains all 17 explanatory variables, and has `log(price-usd)` as the response, as per the findings in the preliminary analysis of the data. To obtain a reduced model, the backwards selection approach is used. The backwards selection procedure is an iterative, criterion based model selection algorithm that depends on the Akaike Information Criterion (AIC), defined by

$$\text{AIC} = -2\ell\left(\hat{\beta}_{\text{mle}}\right) + 2p$$

where $\hat{\beta}_{\text{mle}}$ is the maximum likelihood estimator of β, the vector of regression coefficients in the linear model $y = X\beta + \varepsilon$, ℓ is the log-likelihood function of the sample, defined by the natural log of the joint density of the sample, and p is the number of predictor variables used in the model [Faraway, 2005]. In the case of the candidate linear model, $p = 18$, and as the model is linear, the expression for the AIC can be computed in terms of the Residual Sum of Squares (RSS) as

$$\text{AIC} = n\log\left(\frac{\text{RSS}}{n}\right) + 2p \qquad \text{RSS} = \hat{\varepsilon}^T\varepsilon = \sum_{i=1}^{n}(y_i - x_i^T\hat{\beta})^2$$

where n is the number of observations, and x_i is the i^{th} observation in the sample [Reid, 2022 (b)]. The goal of the model selection algorithm is to decrease the AIC by iteratively removing insignificant explanatory variables from the model, until further removing variables no longer decreases the AIC [Faraway, 2005]. Using the `step` function in `R`, a reduced model was obtained using the backwards selection procedure illustrated above. Observing residual and quantile-quantile (qq) plots for the full main effects model in Figure 4, there is a rather weak trend in the plot of the fitted values against the residuals, suggesting that the observations are approximately independent. However, there is a significant right-skew, and moderate left-skew in the qq-plot, suggesting that the distribution of the residuals is likely non-normal.

To compare the fit of the full main effects model to the reduced model, an Analysis of Variance (ANOVA) test was used. The ANOVA test for comparing the fit of a nested model tests the hypothesis that all of the regression coefficients unique to the larger model are 0, against the alternative that at least one of them is non-zero. The test statistic used is

$$F^* = \frac{(\text{RSS}_{\text{reduced}} - \text{RSS}_{\text{full}})/(p-q)}{\text{RSS}_{\text{full}}/(n-p)}$$

where p and q are the number of predictor variables in the full and reduced models, respectively, n is the number of observations, and F^* has an F-distribution with $p-q$ and $n-p$ degrees of

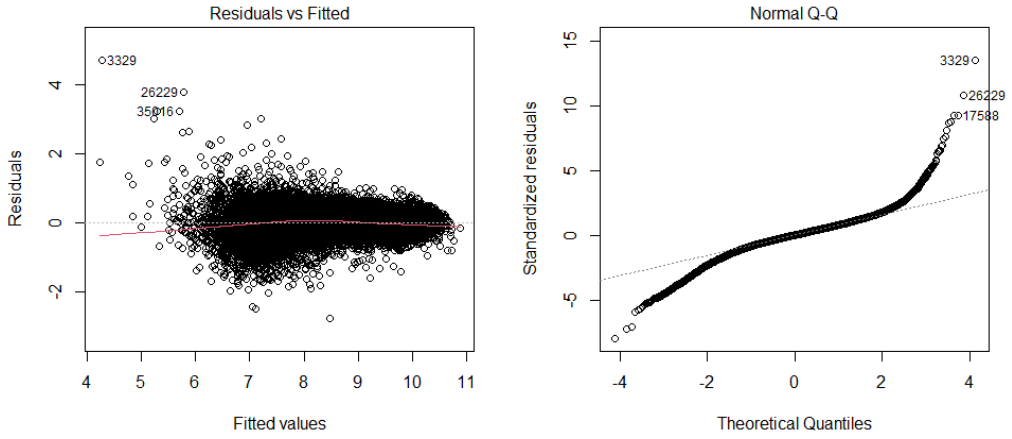


Figure 4: Residual and qq-plots for the full main effects model

freedom, under the null hypothesis [Reid, 2002 (a)]. Applying the ANOVA test to the full and reduced main effects model, a p-value of 0.8636 was obtained, which is quite insignificant, suggesting that there is not evidence that the reduced model has a better fit. To improve the model performance, interaction effects were introduced in a new model, in addition to the main effects. More specifically, a model with all main effects and all pairwise interaction effects between the variables `odometer-value`, `year-produced`, `duration-listed` and `additional-features` was fit with `log(price-usd)` as the response. Analogously to the main effects model, backwards stepwise selection using the AIC was run to produce a reduced model. Using an ANOVA test on the full and reduced interaction effects models, a p-value of 0.5256 was obtained, which is also insignificant, suggesting that there is not evidence that the reduced interactions model has a better fit. However, ANOVA tests comparing each of the interaction models to each of the main effects models produced p-values that suggest that both interaction effects models are a better fit than either of the main effects models at the $\alpha = 0.01$ level. The Root Mean-Squared Error (RMSE) metric was used to determine a comparison of the performance of each of the four models. The RMSE is defined by

$$\text{RMSE} = \sqrt{\text{MSE}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - x_i^T\hat{\beta}\right)^2}$$

and is a metric used to evaluate model performance, namely the ability of a model to predict the response [Faraway, 2005]. Similarly, a weaker, though moderately useful, performance metric is the Coefficient of Determination, or multiple R^2-squared, defined by

$$R^2 = \sum_{i=1}^{n}\left(\overline{y} - x_i^T\hat{\beta}\right)^2 \Bigg/ \sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2 \qquad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

which represents the percentage of explained variation [Faraway, 2005]. Table 1 shows the

values of both metrics across all four models

Model	R-Squared	Root Mean Squared Error
Full	0.8784931	2295.1320
Reduced	0.8784917	2296.3104
Interaction	0.8794653	147.8330
Reduced Interaction	0.8794235	148.2189

Table 1: Performance metrics across all models

As mentioned prior, ANOVA tests displayed statistically significant evidence that the overall fit of both interaction effects models are better than both main effects models, which both performance metrics suggest in Table 1. As the R-squared values for all four models are equal up to 2 significant digits, RMSE will determine the optimal model between the two interaction effects models. As the RMSE is lower for the full interaction effects model, it was selected for further testing and inference.

2.2 Summary for the Statistician

With the problem of interest as stated in the introduction, one possible approach is through the use of linear models. In this particular case, after cleaning and preprocessing the data, an initial main effects model was fit using all of the explanatory variables and the log-transformed response, `log(price-usd)`. Upon using backwards selection with the AIC, a reduced model was obtained, but showed no statistically significant evidence of a better fit. An interaction effects model was fit, including all main effects present in the initial model, as well as all pairwise interaction effects for the variables `odometer-value`, `year-produced`, `duration-listed` and `additional-features`, which were among the most individually significant in the initial model. Backwards selection on the interaction effects model did not produce a new model with statistically significantly better fit, however, ANOAVA tests showed that each of the interaction effects models had statistically significantly better fit at the $\alpha = 0.01$ level. Ultimately, as the R-squared values for the full and reduced interaction effects models were identical up to 4 significant digits, the RMSE performance metric was used to determine the use of the full interaction effects model as the final model.

Under the full interaction effects model, inference was made on `log(price-usd)` in relation to each of the significant variables `additional-features`, `odometer-value`, `year-produced` and `duration-listed`. With the exception of `duration-listed`, linear changes in each variable correspond to sign indefinite changes in the response. With `duration-listed`, an increase of

50 days listed lead to a decrease in price for maximum of listing duration of 150 days, and the opposite was true for listing durations between 150 and 350 days. For each of the remaining significant variables, the expected change in price induced by changes in the variables are shown in Table 2

Variable	Change	Expected change in `price-usd`
`additional-features`	+1 feature	+$1543
`odometer-value`	+50000 km	−$1503
`year-produced`	+5 years	+$6029

Table 2: Performance metrics across all models

Where an increase of 5 years in `year-produced` means the car was produced 5 years earlier, namely the car is newer, which leads to an expected increase of $6029 in `price-usd`. It is worth noting that the model used to perform inference may not be the best possible model for the data, perhaps more advanced techniques in supervised learning, possibly beyond the scope of linear models, may be superior in predicting, and making inference on the response. In any case, the proposed model provides useful insight towards addressing the problem of interest.