 doi:10.20944/preprints201911.0119.v1

Article

Statistical Analysis and Forecasting of Price and Mileage Correlation for Second-hand Cars in Australia

Chinedu G. Agokei [1,†] and Bomonyo J. Afa [2,†*]

[1] Stanbic IBTC Bank Plc, Ilupeju, Lagos state, Nigeria; cgoke11@gmail.com
[2] University of Orleans, Château de la Source, 45100 Orléans, France
* Correspondence: bomonyo.jessica.afa@gmail.com; Tel.: +234-802-613-3920 (B.A.)
† These authors contributed equally to this work.

Abstract: In developed countries, especially the big-sized ones like Australia and the USA, a car is almost an inevitable necessity to carry out daily activities. Due to this, used cars have become a great alternative to brand new cars because of their cost effectiveness. In this work, estimation of prices of used cars based on numerous factors is studied statistically. Data is based on prices of used cars sold across Australia. Statistical methods like correlation and permutation tests using linear regression model, exact tests and non-parametric bootstrapping is implemented to study the relationship of price with mileage and year of manufacture of the car using p-values and null hypothesis. Predictions are also made on the price by calculating a 95% confidence interval (CI) of median prices in small portions of the dataset. The study presents potential ideas for understanding correlation between variables and parameters in business studies.

Keywords: second-hand cars; p-values; confidence interval; non-parametric bootstrapping; correlation

1. Introduction

The cost of buying a new car in industrialized countries like Australia can be quite exorbitant in comparison to present day low growth in wages. As a result, used cars purchases, even though still risky, have started to gain popularity in these areas. However, most buyers and car sellers are sceptical of the conditions to buy a used car or put it up for sale. The first factor to attract a buyer to a car is of course the aesthetics of the car, however, for a car expert or someone with a minimal knowledge of how car operates, there are other factors to take into account, for instance, the brand, model, the year of manufacture and the mileage [1, 2].

Mileage refers to the total distance that this car has covered over the period of its life cycle up until it was put up for sale, so for instance, the number of kilometres driven as seen on its odometer. A general rule to cars is that their performance and integrity is inversely dependent on its mileage. This means the more the car has been used the less trusted it will be. The other factor is age or in other words, year of manufacture. Factors like outdated parts, rusting of parts over time and cost of maintenance could be higher for older cars. The question is would a buyer offer a car for sale based on its mileage or manufacture year. How does this factor of mileage correlate with the selling price and can statistics foster calculation of expected price of a vehicle based on these variables like manufacturing year and mileage [3, 4].

Kuiper [5] introduced multiple regression as a way of estimating the worth of a car by developing a multivariate regression model to predict the retail price of General Motors used cars in 2005. The goal of this study was to describe the relationship between variables like mileage, engine size, number of doors and to predict the contribution of each variable. He used the T-statistic for the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 November 2019 doi:10.20944/preprints201911.0119.v1

slope coefficient to answer predictor questions based on the correlated variables and p-values. These p-values are needed to weigh the strength of evidence about the population based on given data. Small p-values that are greater or equal to 0.05 entail rejecting the null hypothesis, which means accepting the alternative hypothesis and for larger p-values greater than 0.05, the null hypothesis is not rejected. The latter case means that there is no relationship between the measured phenomena or variables [6 - 8].

Bootstrapping is a popular approach to statistical inference in performing permutation, randomization and cross-validation tests based on resampling methods [9]. There are numerous methods based on parametric and non-parametric bootstrapping. Of interest to this work is non-parametric bootstrapping which permits estimation of sampling distribution of a statistic empirically avoiding any prior deduction of population and derivation of this distribution. The work is divided into section 2, which is the description of the data set. The next section is a presentation of the results and statistical analysis and finally the conclusion.

2. Dataset

The data set is obtained from Kaggle (its original source was www.carsale.com.au) [10]. It is about second-hand car prices in Australia. The dataset has 55.9k rows and 13 columns. Hence, this could be considered to be quite a large dataset. The columns cover the registration Id, the brand of the car (Honda, Mazda, Nissan and Toyota), the model of the car, the description of the car type, the price of the car, the discount issued, the mileage on the odometer, the body and nature of the car, transmission (manual or automatic), engine, state, seller and year. It is even seen that the minimum year recorded was in 1968. Due to the size of this data set, a filter of the information is done and focus is made on particular brand, models and year. A summary of the original data set is shown in figure 1.

```
> View(car)
> summary(car)
       id            brand                    model
 Min.   :    1   honda : 3859   hilux           : 4523
 1st Qu.:13968   mazda :15226   3               : 4381
 Median :27936   nissan:11750   corolla         : 3328
 Mean   :27936   toyota:25035   navara          : 2977
 3rd Qu.:41903                  landcruiser     : 2964
 Max.   :55870                  landcruiserprado: 2750
                                (Other)         :34947
                                                   title           price
 2018 Mazda 3 Maxx Sport BN Series Auto               :  324   Min.   :   250
 2018 Mazda 3 SP25 BN Series Auto                     :  323   1st Qu.: 12500
 2016 Toyota Corolla Ascent Auto                      :  289   Median : 20990
 2015 Toyota Landcruiser Prado GXL Auto 4x4           :  231   Mean   : 24610
 2018 Mazda CX-5 Maxx Sport KF Series Auto i-ACTIV AWD:  224   3rd Qu.: 32000
 2015 Toyota Camry Altise Auto                        :  220   Max.   :399000
 (Other)                                              :54259   NA's   :544
                discount       odometres                 body
 Drive Away         :19505   Min.   :     0   SUV           :19624
 Excl. Govt. Charges:15801   1st Qu.: 25609   Sedan         :10140
 Get Price*         :  544   Median : 81000   Hatch         : 8929
 None               :20020   Mean   : 96940   Ute           : 6885
                             3rd Qu.:145917   Cab Chassis:  3258
                             Max.   :999999   Automatic     : 2187
                                              (Other)       : 4847
              transmission                      engine           state
 Automatic             :39164   4cyl 2.5L Petrol      : 6989   NSW    :15397
 Manual                :13900   4cyl 2.0L Petrol      : 6795   VIC    :14281
 4cyl 2.0L Petrol      :  628   4cyl 3.0L Turbo Diesel: 5518   QLD    :11366
 4cyl 2.5L Petrol      :  416   4cyl 1.8L Petrol      : 5130   WA     : 5426
 4cyl 2.5L Turbo Petrol:  301   6cyl 3.5L Petrol      : 4244   SA     : 4351
 4cyl 1.5L Petrol      :  246   4cyl 2.4L Petrol      : 2734   None   : 2800
 (Other)               : 1215   (Other)               :24460   (Other): 2249
               seller           year
 Demo              : 2941   Min.   :1968
 In Stock          : 3244   1st Qu.:2009
 New Car           : 2800   Median :2013
 Private Seller Car:20020   Mean   :2012
 Used Car          :26865   3rd Qu.:2016
                            Max.   :2018
```

Figure 1. Prices of second-hand cars sold in different states in Australia, *data source in* [10]

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 November 2019 doi:10.20944/preprints201911.0119.v1

The goal of the study is to perform statistics in understanding the relationship between price and influential variables by carrying out Pearson and Spearman correlation coefficient to perform a classical T-test and permutation test based on the T-statistic. The next step is to use a multiple regression model on the original data set and then narrowed down to linear regression model to check the correlation of price with mileage. An exact permutation test would be performed to verify the importance of this variable on the price. The idea of employing these numerous methods is that it enables us to determine the correctness of these statistical approaches based on the data provided and making comparison using the p-values calculated [6, 7]. The final step is to calculate the 95% confidence interval for the mean and standard deviation of price for the sampled data. The percentile and simple methods are employed to make comparison and finally perform CI on price prediction for a given mileage.

3. Results and Analysis

3.1. Data Visualization

The data set is visualized below using dot plots and histogram to check the correlation between price and these numerous variables. Also to estimate the range of prices of used cars between these years till date. Figure 2 shows the car prices as a function of variables like the year the cars were manufactured in comparison to the distance covered by these cars over the years. A seller or buyer would want to understand what it entails to sell or purchase a car based on their desired asking price or budget.

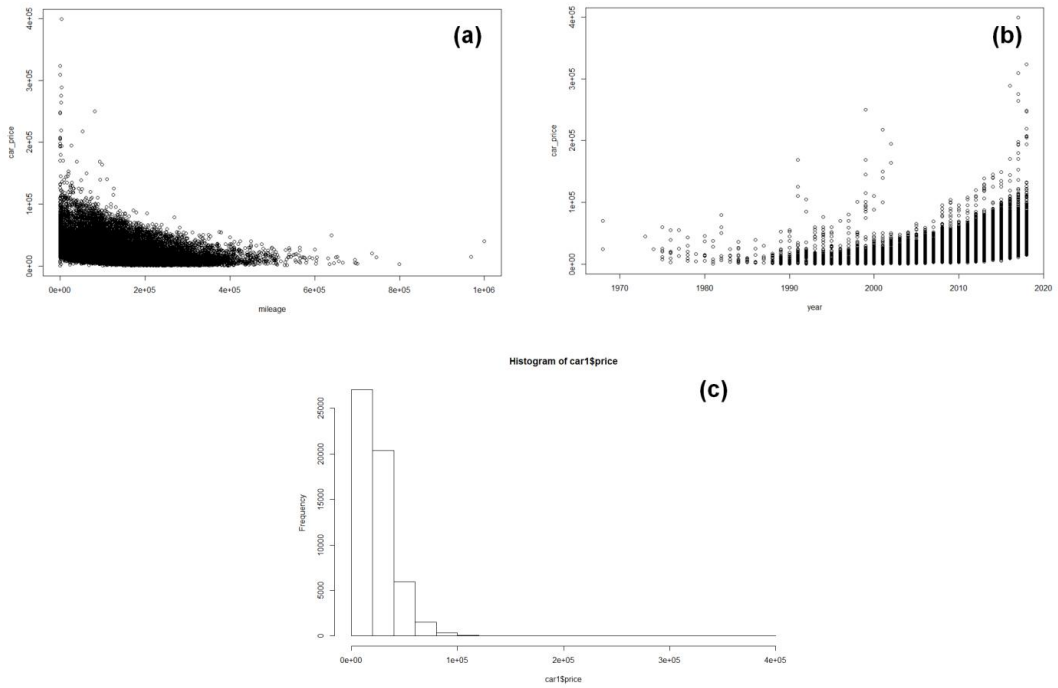

Figure 2. Visualization of original data set of used car price estimation (a) price as a function of mileage (b) price as a function of year (c) histogram of price

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 November 2019 doi:10.20944/preprints201911.0119.v1


```
Call:
lm(formula = car_price ~ mileage + year, data = car1)

Residuals:
   Min     1Q Median     3Q    Max
-24128  -9404  -4019   5186 365710

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.989e+06  3.794e+04  -78.77   <2e-16 ***
mileage     -1.188e-02  1.167e-03  -10.18   <2e-16 ***
year         1.498e+03  1.881e+01   79.63   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14900 on 55323 degrees of freedom
  (544 observations deleted due to missingness)
Multiple R-squared:  0.2637,    Adjusted R-squared:  0.2636
F-statistic:  9905 on 2 and 55323 DF,  p-value: < 2.2e-16
```

A linear regression model is done to estimate the p-value and determine the correlation between price and mileage and year of manufacture for the original data set, we can see that the p-value is far less than 0.05. We therefore reject the null hypothesis in this case and accept the alternative hypothesis which we believe that there is a correlation between the price of the cars and the mileage and year of manufacture. We can also see from figure 2 that prices are more dependent on the mileage. The year correspond to the mileage also in the sense that cars in 2018 are less used in comparison to cars in 1999, so it would be difficult to base our estimation only on the year of manufacture but rather on mileage.

3.2. Data Sampling

3.2.1. Sample data

The sample is done randomly but particularly filtered for the year 2010, Honda accord as the brand and model of the car, most of the cars listed in table 1 correspond to the type V6 luxury Auto and VTi Auto. Table 1 shows the price of the cars listed in the sampled data with the mileage in km. Sample size is set to 6 and the study is carried out based on the reduced data.

Table 1. Price of used cars in Australian Dollars and Mileage in km as displayed in the odometer.

Price (in AUD)	Odometer (km)
8999	210125
12990	151895
12100	95000
11555	121529
8250	142651
9500	146000

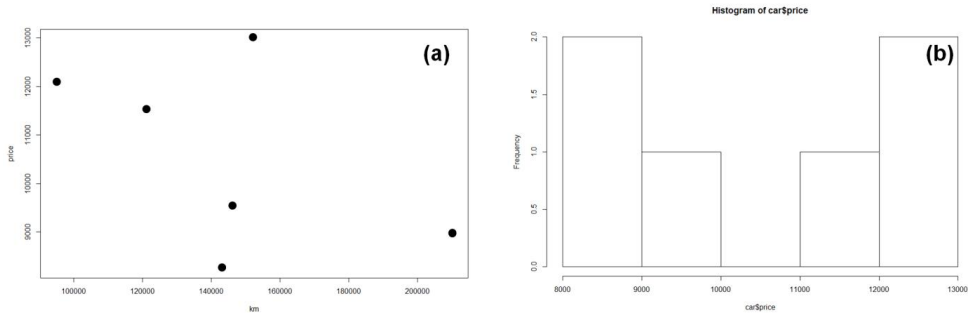

Figure 3. Using the sample data (a) Plot of Price as a function of Mileage in km and (b) histogram of price of data seen in table 1

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 November 2019 doi:10.20944/preprints201911.0119.v1

3.2.1. Correlation and Permutation tests

Four methods were used to calculate the p-value and find a correlation between variables, price and mileage. These methods were (a) Pearson correlation test (b) Spearman correlation test (c) Linear Regression model (d) Exact permutation test.

I. Pearson method

We see that in the case of Pearson method, the p-value is 0.8305556, hence, this means that the p-value is far greater than 0.05 indicating weak evidence to reject the null hypothesis, so in this case we fail to reject the H_0.

II. Spearman method

In this case of Spearman, just like in using the Pearson method, the p-value is 0.6430556. Even though smaller than the previous method, this p-value is also considerably larger than 0.05 and therefore the null hypothesis is also not rejected. A comparison between the Pearson (blue) and Spearman (red) method is shown in figure 4.

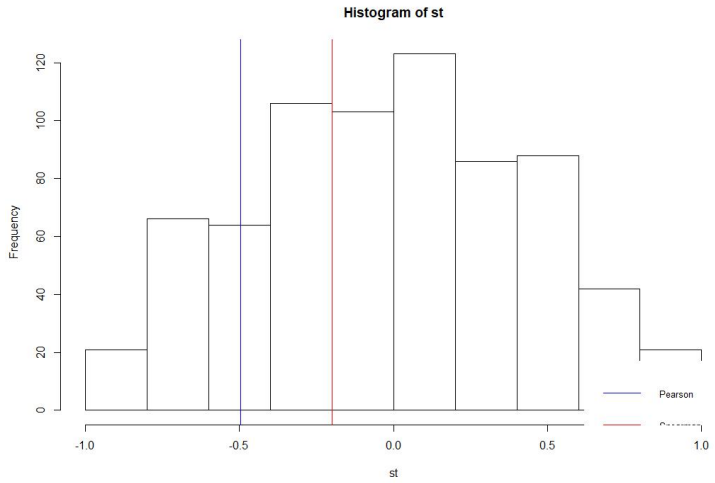

Figure 4. Histogram comparing statistics done using Pearson and Spearman methods

III. Linear Regression model

Linear regression model was used to verify the correlation between price and mileage for the sampled set of data. An estimate was performed and the summary was obtained with a p-value still greater than 0.05

```
Call:
lm(formula = price ~ km, data = car)

Residuals:
     17       16       44        3       37        6
   48.6  2605.6   314.5   422.8 -2362.0 -1029.5

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.412e+04  3.212e+03   4.397   0.0117 *
km          -2.463e-02  2.160e-02  -1.140   0.3179
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1851 on 4 degrees of freedom
Multiple R-squared:  0.2452,    Adjusted R-squared:  0.05655
F-statistic:   1.3 on 1 and 4 DF,  p-value: 0.3179
```

 doi:10.20944/preprints201911.0119.v1

As seen above, the p-value is 0.3179 which is also far greater than 0.05, however, this is much less than the case of Pearson and Spearman methods for verifying the p-value. We also reject the null hypothesis in this case.

IV. Exact Permutation test

An exact permutation test was performed for mileage (in odometer) and the p-value was calculated, just as in the case of Pearson, it is equal to 0.8305556. This is also a significantly high p-value. Hence, the null hypothesis is also not rejected.

3.3. Price Prediction and Confidence Interval Calculation for mean and standard deviation

The estimation and confidence interval for the mean and standard deviation of price is done using the Percentile and simple methods. Using the Percentile or quantile method, the 95% confidence interval of median of price is between 9134.00 for the 2.5% and 11937.33 for the 97.5%, this means the 95% CI is given as 10535.665±1401.665 for the sampled data prices. In the case of the simple method, 95% CI is estimated between 9194 for the 2.5% and 11997.33 for the 97.5%, which is expressed as 10595.665±1401.665 for the same data. The 95% CI for standard deviation of price using percentile method is 1443.965±831.592 (that is, 612.3726, 2275.5571). Using the simple method, the 95% CI is estimated as 2367.8855±831.5925 (1536.293, 3199.478).

Based on the closeness in results, the simple method was used to make prediction on price (95% CI) based on the mileage was covered for 20000 km on the odometer, the result obtained was 3811.85. However, there is a real doubt in the estimation as we can see from previous results that the p-value shows poor correlation between these two variables due to data partitioning.

4. Conclusions

A statistical study is carried out on data set related to prices of second hand cars sold in Australia. Numerous correlation and permutation test methods were implemented to estimate the p-values to verify the correlation between the price and mileage and other variables like year of manufacture. It is seen that from a limited data, there is no correlation between the variables as this is a usual problem with small data sets, however, for the case of the large data set using the linear regression model, we can see that there is better correlation as the p-values are far less than 0.05 and the relationship of these variables with price becomes relevant. The present study shows implementation of useful methods in calculating 95% confidence interval of statistical data and making prediction using these numerous statistical methods. This opens doors in setting criteria for rejection of the null hypothesis which is an important concept to consider in most fields of science studying correlation and dependence of variables ad phenomena on each other.

Supplementary Materials: The R script is available online.

Author Contributions: C.G.A and B.J.A analysed the data, wrote the codes and carried out simulations for visualization and price prediction based on the provided data.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

 doi:10.20944/preprints201911.0119.v1

References

1. RAC - Buying a used car - the ultimate checklist. Available online (accessed on 15 March 2019): https://www.rac.co.uk/drive/advice/buying-and-selling-guides/buying-a-used-car/
2. Used Car Buying Tips. Available online (accessed on 23 February 2019): https://www.immihelp.com/newcomer/used-car-buying-tips.html
3. Autotrader - Mileage vs Age. Available online (accessed on 15 March 2019): https://www.autotrader.com/car-shopping/whats-more-important-when-buying-car-miles-or-age-240611
4. Minitab Blog - can regression and statistics help you find a great deal on a used car. Available online: http://blog.minitab.com/blog/understanding-statistics/can-regression-and-statistical-software-help-you-find-a-great-deal-on-a-used-car (accessed on 15 March 2019)
5. Kuiper, S. (2008) "Introduction to Multiple Regression: How much is your car worth?", Journal of Statistics Education, Vol. 16 Iss. 3. DOI: 10.1080/10691898.2008.11889579
6. James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *An introduction to statistical learning; with Applications in R*. NY, USA: Springer, ISBN 978-1-4614-7137-0. 1st ed. 2013.
7. Altman, D.G.; Machin, D.; Bryant, T.N.; Gardner, M.J. *Statistics with Confidence; Confidence Intervals and Statistical Guidelines*. NJ, USA: Wiley, ISBN 978-1-7279-1375-3. 2nd ed. 2000.
8. Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, et al. (2016) "Statistical tests, *P* values, confidence intervals, and power: a guide to misinterpretations". European Journal of Epidemiology. 31(4):337–350. DOI: 10.1007/s10654-016-0149-3
9. Fox J (2002). "Bootstrapping Regression Models." In "An R and S-PLUS Companion to Applied Regression: A Web Appendix to the Book.", Sage, Thousand Oaks, CA. Available online (accessed on 15 March 2019): http://CRAN.R-project.org/doc/contrib/Fox-Companion/appendix-bootstrapping.pdf.
10. Kaggle datasets - Second-hand price estimation. Available online (accessed on 15 March 2019): https://www.kaggle.com/bahamutedean/secondhand-car-price-estimation