**GEELY AUTO CAR PRICE PREDICTION AND ESTIMATION WITH MULTIPLE LINEAR REGRESSION**

**Abstract**

GEELY auto is one of the new automobiles in china, this organization is interested in joining the American market, but to create a standard plan of boosting their product to be more effective than American and European counterparts. We analysed factors that might influence car price prediction, the results show that the factors significantly predict car price, the forecasted price is higher than the actual price thus this reveals that those factors could trigger the price of cars exponentially. We recommend the GEELY auto to considered this factor and strategize for a better future.

**Keyword: Multiple regression, Descriptive, car price, and Forecasting**.

**Introduction**

Car price prediction is an interesting factor and one of the major popular problems of Civilization, as per historical record gotten from the agency for statistics of BiH over 921,456 vehicles were registered in the year 2014 and of which 84% of it are personal usage cars, Bih (2018) the record show that the number increased by 2.7% since 2013 and it is likely that the trend will continue and the number of the cars will increase in the future. This adds additional significance to the problem of the car price prediction. Car price prediction accuracy involves expert knowledge because mostly price varies and usually depends on many important factors and features or characteristics. Typically, mostly the significant factors are brand and model, age, mileage, horsepower, and Consumption of fuel varies among car per mile highly affect the price of a car due to frequent changes in the price of fuel. Various features like door number, the color of the exterior, transmission type, safety e.g airbag, air condition, heating, dimensions, interior, and whether it has navigation, map or not will as well influence the car price. In this research, we applied multiple regression to estimate the factor which Influences the GEELY auto car price prediction.

**Study Background**

Geely Auto Group is a leading automobile manufacturer based in Hangzhou, China, and was founded in 1997 as a subsidiary of Geely holding group. Geely auto group sells vehicles under the Geely auto brand and holds a 50% stake in the LYNK & Co brand. The group employs more than 50,000 people, operates 12 plants, five global R&D centers in Hangzhou, Ningbo, Gothenburg, Coventry, and Frankfurt. The Group also boasts five global design studios in Shanghai, Gothenburg, Barcelona, California, and Coventry respectively with over 900 members of staff in total. The Geely Auto brand has been listed on the Hong Kong stock exchange since 2005. In 2019, the brands under Geely Auto Group management sold over 1.46 million units, with Geely Auto retaining its position as the best-selling Chinese brand for three consecutive years, Lynk & Co setting a new annual sales record, and a revitalized PROTON returning to second place in its home market of Malaysia. The group is committed to world-class research and development in the auto industry. It has established the Geely Automotive Research Institute in Hangzhou, the renowned China Euro Vehicle Technology (CEVT) in Gothenburg, and the world-leading Geely Research Institute in Hangzhou Bay, Ningbo to focus on the development of vehicles, engines, transmission, and vehicle bound electronics. Other facilities include Design Centres in Barcelona, Gothenburg, Coventry California, and Shanghai.

**Problem description**

GEELY Auto aspires to enter the US market by setting up their manufacturing unit by producing cars locally to the Chinese to arise for competition to US and European counterparts. To evaluate the pricing of the car in the American market, they contracted an automobile consulting company to understand various factors that influence car price and what depends on the pricing of cars in the American market and European markets. To build a standard market that will influence other counterparts. GEELY Auto studies the various variables which can significantly predict car price, in what way those variables can describe the price of a car, and finally, different cars in the American market are considered.

**Research Questions**

The study aims to evaluate the factors influencing the price of car prediction and the various question arose such as

* Which variables are significant in predicting the price of a car?
* How well those variables describe the price of a car?
* What variation does the factor explain in predicting the car price?
* What is the forecasted estimated for the new set of car prices?

**Literature review**

Listiani (2009) study research on predicting the price of a used car in an extended way with a regression model which she built using a support vector machine in predicting the prediction of the car that has been leased and used with better precision, the Support vector machine is known as the best method is dealing with datasets with more dimensions and it is said to be less prone to overfitting or Underfitting. The research weakness is that the method used was not shown in basic statistical indicators such as mean, variance, or standard deviation. Richardson (2009) approach to the prediction of car price was based on that car producers produce more durable cars and this might as well influence the price, he applied multiple regression model estimation and he concluded that traditional cars have roots in environmental concerns about the climate and gives higher fuel efficiency. Wu et al (2009) as well conducted a car price prediction study using the neuro-fuzzy knowledge-based system, they analyzed it by considering attributes such as brand, year of production, and type of engine. Their result for the prediction model produced similar results as the simple regression model. Moreover, they applied an expert system named ODAV (optimal distribution of Auction Vehicles) this system gives a very clear insight of the best price for vehicle prediction as well there is high demand for selling the cars at the end of the leasing year by car dealers. The system produces a Regression model based on the k-nearest neighbor of the machine learning algorithm which was used to predict the price of a car. This system has a very tendency to be exceptionally and successful since there are more than two million vehicles that were exchanged through it Du et al (2009). Gonggi (2011) also proposed a model which is built using ANN (Artificial Neural Networks) for car prediction. The study considered several attributes: estimated car life, miles move, and brand. This proposed model was built so that it could deal with any non-linear relations in data and this approach was not used in the simple linear regression techniques proposed and utilize in his previous model for car prediction.

**Data collection and Source of data**.

The data is extracted from GEELYAuto car prediction data from Kaggle for estimating the factor influencing car price prediction, factors such as wheelbase, car length, car width, car height, curb weight, engine size, bore ratio, stroke, compression ratio, horsepower, peak rpm, city mpg, and highway mpg. The source of the data is from <https://www.kaggle.com/hellbuoy/car-price-prediction>.

**Data Analysis**

Table 1: Summary statistics of the various factor influencing the car price prediction

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Variables | N | Minimum | Maximum | Sum | Mean | Std. Deviation |
| Wheel base | 205 | 86,6 | 120,9 | 20245,1 | 98,757 | 6,0218 |
| car length | 205 | 141,1 | 208,1 | 35680,1 | 174,049 | 12,3373 |
| car width | 205 | 60,3 | 72,3 | 13511,1 | 65,908 | 2,1452 |
| car height | 205 | 47,8 | 59,8 | 11013,6 | 53,725 | 2,4435 |
| engine size | 205 | 61 | 326 | 26016 | 126,91 | 41,643 |
| curb weight | 205 | 1488 | 4066 | 523891 | 2555,57 | 520,680 |
| stroke | 205 | 2,070 | 4,170 | 667,360 | 3,25541 | ,313597 |
| bore ratio | 205 | 2,54 | 3,94 | 682,60 | 3,3298 | ,27084 |
| compression ratio | 205 | 7,00 | 23,00 | 2079,22 | 10,1425 | 3,97204 |
| horsepower | 205 | 48 | 288 | 21344 | 104,12 | 39,544 |
| City mpg | 205 | 13 | 49 | 5170 | 25,22 | 6,542 |
| Peak rpm | 205 | 4150 | 6600 | 1050650 | 5125,12 | 476,986 |
| Highway mpg | 205 | 16 | 54 | 6304 | 30,75 | 6,886 |
| price | 205 | 5118,000 | 45400,000 | 2721725,667 | 13276,71057 | 7988,852332 |

The table above shows the average, the minimum, maximum, the standard deviation, and a sum of the various car features, the average wheelbase is 98.757 with a maximum of 120.9 and 86.6, while the average car length and width is 174.049 and 65.908. the highest and the lowest car length and width are 72.3, 59.8 and 60.3, and 47.8. The maximum engine size is 326 and minimum is 61, also the maximum compression ratio is 23 and lowest 7, the features shows that there is 49 and 54 maximum mpg for city and high way and 13 and 16 for the lowest mpg of the car. The average car price worth $13276.71 with a maximum price of $454000 and a minimum price of $5118.

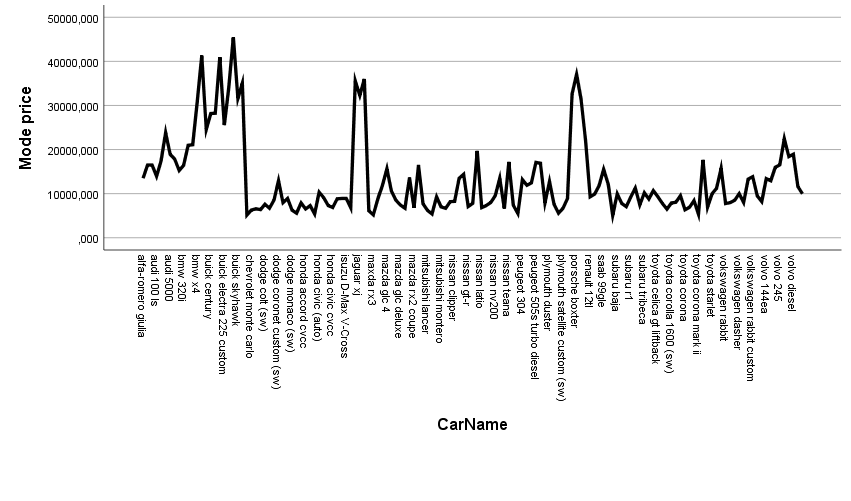


Figure 1: Show various the trend of various car and their price.

The result of the plot shows that the Buick regal sport coupe (turbo) has the highest car price followed by the Buick century special, these two cars cost $45000 and $40000 and the car with the lowest price is Subaru Baja.

**Multiple regression summary estimates**

**Car price prediction**

Table 2: showing the multiple linear estimations of the car price prediction

Dependent: Car Price.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variables | Beta | Std. Error | T stat | P value |
| (Constant) | -47495,741 | 15269,286 | -3,111 | ,002 |
| Wheel base | 122,617 | 100,465 | 1,220 | ,224 |
| Car length | -94,675 | 55,557 | -1,704 | ,090 |
| Car width | 505,572 | 246,013 | 2,055 | ,041 |
| Car height | 163,180 | 135,721 | 1,202 | ,231 |
| Curb weight | 1,885 | 1,737 | 1,085 | ,279 |
| Engine size | 117,346 | 13,837 | 8,481 | ,000 |
| Bore ratio | -1002,565 | 1195,798 | -,838 | ,403 |
| stroke | -3034,606 | 778,604 | -3,897 | ,000 |
| Compression ratio | 298,137 | 82,914 | 3,596 | ,000 |
| Peak rpm | 2,375 | ,671 | 3,540 | ,001 |
| Horse power | 30,809 | 16,216 | 1,900 | ,059 |
| City mpg | -320,355 | 177,769 | -1,802 | ,073 |
| Highway mpg | 202,822 | 159,760 | 1,270 | ,206 |
| R Square | 0.851 | | | |
| R | 0.922 | | | |

**Interpretation**

**Hypothesis testing**

**Rejection Rule**

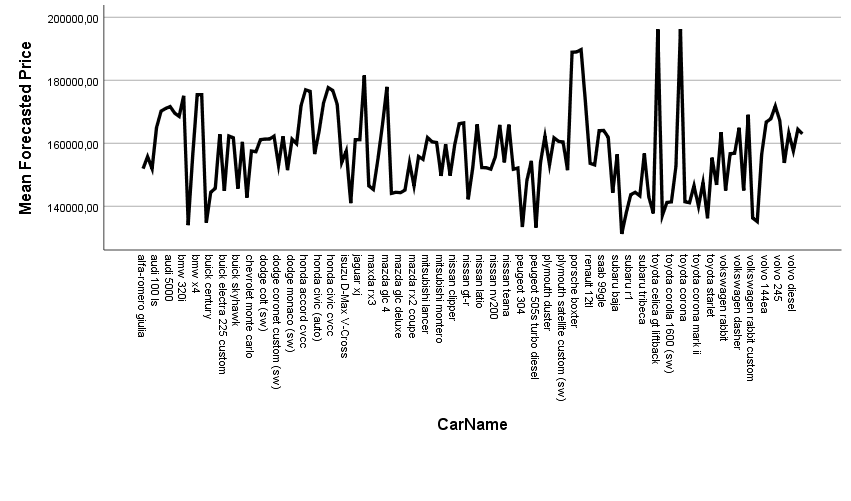
Reject if p -value , alpha =0.05.

**Conclusion**

The result of Multiple regression shows that the car width, engine size, stroke, compression ratio, and peak rpm is statistically significant to influence the price of a car since the P value <0.05, we reject the null hypothesis and conclude that the variables are significant, and the R square shows that there is 85.1% of the total variation of the Car price which is explained by the factors and not by the error. Thus, there is 92.2% of how close or relationship between the car price and the variables in predicting the car price. Car length, bore ratio, stroke, and city mpg contribute a negative effect on the car price, for every increase in the car length, bore ratio, stroke, and city mpg there is a decrease in the price of the car, while other variables contribute a positive effect on the car price and thus it led to increase in car price.

**The car price prediction model is given as follows;**

Car price = -47495,741 + 122,617 Wheel base – 94.675 Car length, 505.572 Car width + 163.180 Car height +1.885 Curb Height + 117.346 Engine size – 1002.565 bore ratio -3034.606 stroke + 298.137 compression ratio + 2.375 Peak rpm +30.809 Horse power – 320.55 City mpg +202.822 Highway mpg.



The forecasted car prediction shows that Toyota corolla and Toyota Celica has the highest car price among the selected car from the Geely Auto and Volvo with the lowest car price. So comparing to the actual price from the American market the expensive car was Buick regal sport while the forecasted price after comparing all the factors the price of each car increases and Toyota corolla and Celica is the expensive car.

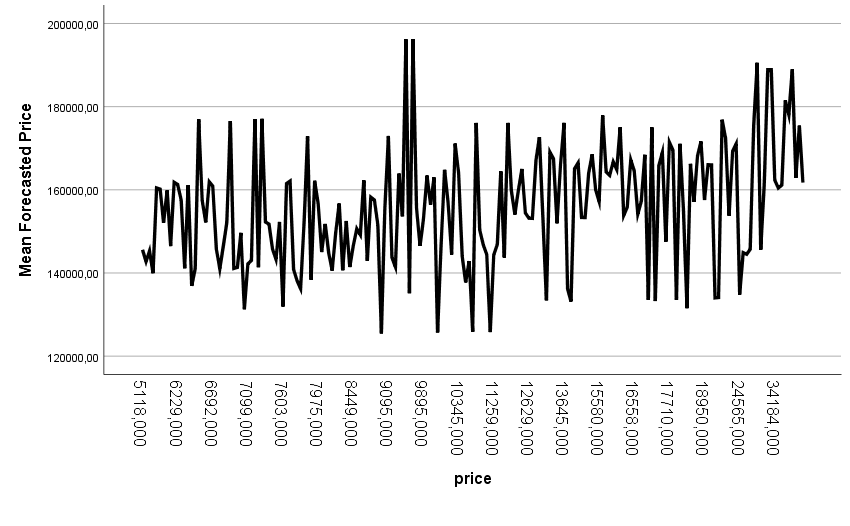


Figure 3: showing the forecasted price for each actual cost of the price after considering all factors.

The results show that when the actual price was $9895 the forecasted price of the car after the factors are considered rises to $190k all of the prices go up.

**Error report**

The data is partition into 60% training data and 40% validation data.

Table3: showing the Training data scoring- summary report

|  |  |  |
| --- | --- | --- |
| Total sum of squared errors | RMS error | Average Error |
| 1385407222.631 | 12260240.908 | 228.08169 |

Table 4: Showing the Validation data-summary report.

|  |  |  |
| --- | --- | --- |
| Total sum of squared errors | RMS error | Average Error |
| 973751304,239 | 13910732,918 | 461.7600 |

**Discussion and Conclusion**

Based on the results of the analysis, the average wheelbase is 98.757 with a maximum of 120.9 and 86.6, while the average car length and width are 174.049 and 65.908. the highest and the lowest car length and width are 72.3, 59.8 and 60.3, and 47.8. The maximum engine size is 326 and minimum is 61, also the maximum compression ratio is 23 and lowest 7, the features shows that there is 49 and 54 maximum mpg for city and high way and 13 and 16 for the lowest mpg of the car. The average car price worth $13276.71 with a maximum price of $454000 and a minimum price of $5118. Furthermore, that the car width, engine size, stroke, compression ratio, and peak rpm is statistically significant to influence the price of a car and the R square shows that there is 85.1% of the total variation of the Car price which is explained by the factors and not by the error. Thus, there is 92.2% of how close or relationship between the car price and the variables in predicting the car price. Car length, bore ratio, stroke, and city mpg contribute a negative effect on the car price, for every increase in the car length, bore ratio, stroke, and city mpg there is a decrease in the price of the car, while other variables contribute a positive effect on the car price and after the forecasting, the price of the car goes up. We can conclude that the variables can predict the increase in the car price.

**Recommendation**

We recommend GEELY auto should be careful with their decision, if they want a standard plan and price to hold a competition against the American and Europeans counterparts, they should consider those factors influencing car price and as well the forecasted price, those factors triggers the car price upwardly, this result will help them to plan how to create their portfolio and prepare them for the future.

References

Agencija za statistiku BiH., retrieved from: http://www.bhas.ba. [accessed July 18, 2018.]

Du, J., Xie, L., & Schroeder, S. (2009). Practice Prize Paper—PIN Optimal Distribution of Auction Vehicles System: Applying Price Forecasting, Elasticity Estimation, and Genetic Algorithms to Used-Vehicle Distribution. Marketing Science, 28(4), 637-644.

Gongqi, S., Yansong, W., & Qiang, Z. (2011, January). New Model for Residual Value Prediction of the Used Car Based on BP Neural Network and Nonlinear Curve Fit. In Measuring Technology and Mechatronics Automation (ICMTMA), 2011 Third International Conference on (Vol. 2, pp. 682-685). IEEE.

Listiani, M. (2009). Support vector regression analysis for price prediction in a car leasing application (Doctoral dissertation, Master thesis, TU Hamburg-Harburg).

Richardson, M. S. (2009). Determinants of used car resale value. Retrieved from: https://digitalcc.coloradocollege.edu/islandora/object /coccc%3A1346

Wu, J. D., Hsu, C. C., & Chen, H. C. (2009). An expert system of price forecasting for used cars using adaptive neuro-fuzzy inference. Expert Systems with Applications, 36(4), 7809-7817.

**Appendix**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model Summary** | | | | |
| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
| 1 | ,922a | ,851 | ,841 | 3189,170849 |
| a. Predictors: (Constant), highway mpg, stroke, peak rpm, car height, compression ratio, bore ratio, engine size, car width, wheelbase, horsepower, car length, curb weight, city mpg | | | | |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **ANOVA** | | | | | | |
| Model | | Sum of Squares | df | Mean Square | F | Sig. |
| 1 | Regression | 11077014517,081 | 13 | 852078039,775 | 83,777 | ,000b |
| Residual | 1942624844,912 | 191 | 10170810,706 |  |  |
| Total | 13019639361,993 | 204 |  |  |  |
| a. Dependent Variable: price | | | | | | |
| b. Predictors: (Constant), highway mpg, stroke, peak rpm, car height, compression ratio, bore ratio, engine size, car width, wheelbase, horsepower, car length, curb weight, city mpg | | | | | | |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Coefficients** | | | | | | |
| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. |
| B | Std. Error | Beta |
| 1 | (Constant) | -47495,741 | 15269,286 |  | -3,111 | ,002 |
| wheelbase | 122,617 | 100,465 | ,092 | 1,220 | ,224 |
| carlength | -94,675 | 55,557 | -,146 | -1,704 | ,090 |
| carwidth | 505,572 | 246,013 | ,136 | 2,055 | ,041 |
| carheight | 163,180 | 135,721 | ,050 | 1,202 | ,231 |
| curbweight | 1,885 | 1,737 | ,123 | 1,085 | ,279 |
| enginesize | 117,346 | 13,837 | ,612 | 8,481 | ,000 |
| boreratio | -1002,565 | 1195,798 | -,034 | -,838 | ,403 |
| stroke | -3034,606 | 778,604 | -,119 | -3,897 | ,000 |
| compressionratio | 298,137 | 82,914 | ,148 | 3,596 | ,000 |
| peakrpm | 2,375 | ,671 | ,142 | 3,540 | ,001 |
| horsepower | 30,809 | 16,216 | ,153 | 1,900 | ,059 |
| citympg | -320,355 | 177,769 | -,262 | -1,802 | ,073 |
| highwaympg | 202,822 | 159,760 | ,175 | 1,270 | ,206 |
| a. Dependent Variable: price | | | | | | |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model Summary** | | | | |
| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
| 1 | ,907a | ,823 | ,804 | 3501,462681 |
| a. Predictors: (Constant), highwaympg, stroke, peakrpm, carheight, compressionratio, boreratio, carwidth, horsepower, wheelbase, carlength, curbweight, citympg | | | | |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **ANOVAa** | | | | | | |
| Model | | Sum of Squares | df | Mean Square | F | Sig. |
| 1 | Regression | 6435575065,418 | 12 | 536297922,118 | 43,743 | ,000b |
| Residual | 1385407222,631 | 113 | 12260240,908 |  |  |
| Total | 7820982288,049 | 125 |  |  |  |
| a. Dependent Variable: price | | | | | | |
| b. Predictors: (Constant), highwaympg, stroke, peakrpm, carheight, compressionratio, boreratio, carwidth, horsepower, wheelbase, carlength, curbweight, citympg | | | | | | |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Coefficientsa** | | | | | | |
| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. |
| B | Std. Error | Beta |
| 1 | (Constant) | -72395,776 | 20415,307 |  | -3,546 | ,001 |
| wheelbase | 146,687 | 131,531 | ,112 | 1,115 | ,267 |
| carlength | -99,357 | 79,480 | -,153 | -1,250 | ,214 |
| carwidth | 934,402 | 319,505 | ,259 | 2,925 | ,004 |
| carheight | -89,261 | 191,949 | -,028 | -,465 | ,643 |
| curbweight | 7,375 | 2,073 | ,486 | 3,558 | ,001 |
| boreratio | 416,657 | 1664,040 | ,014 | ,250 | ,803 |
| stroke | -2454,143 | 1067,960 | -,100 | -2,298 | ,023 |
| compressionratio | 153,442 | 114,984 | ,070 | 1,334 | ,185 |
| horsepower | 85,682 | 20,183 | ,405 | 4,245 | ,000 |
| peakrpm | ,787 | ,986 | ,044 | ,798 | ,426 |
| citympg | -129,136 | 252,191 | -,102 | -,512 | ,610 |
| highwaympg | 266,968 | 229,905 | ,226 | 1,161 | ,248 |
| a. Dependent Variable: price | | | | | | |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model Summary** | | | | |
| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
| 1 | ,913a | ,834 | ,806 | 3729,709495 |
| a. Predictors: (Constant), highwaympg, stroke, carheight, peakrpm, compressionratio, boreratio, carwidth, horsepower, wheelbase, carlength, curbweight, citympg | | | | |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **ANOVAa** | | | | | | |
| Model | | Sum of Squares | df | Mean Square | F | Sig. |
| 1 | Regression | 4909112481,936 | 12 | 409092706,828 | 29,408 | ,000b |
| Residual | 973751304,239 | 70 | 13910732,918 |  |  |
| Total | 5882863786,175 | 82 |  |  |  |
| a. Dependent Variable: price | | | | | | |
| b. Predictors: (Constant), highwaympg, stroke, carheight, peakrpm, compressionratio, boreratio, carwidth, horsepower, wheelbase, carlength, curbweight, citympg | | | | | | |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Coefficientsa** | | | | | | |
| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. |
| B | Std. Error | Beta |
| 1 | (Constant) | -3667,791 | 28578,789 |  | -,128 | ,898 |
| wheelbase | 452,090 | 198,991 | ,352 | 2,272 | ,026 |
| carlength | 21,282 | 124,436 | ,029 | ,171 | ,865 |
| carwidth | -577,267 | 487,219 | -,145 | -1,185 | ,240 |
| carheight | -78,471 | 278,503 | -,022 | -,282 | ,779 |
| curbweight | 1,957 | 3,661 | ,125 | ,535 | ,595 |
| boreratio | -951,161 | 2734,953 | -,030 | -,348 | ,729 |
| stroke | -1072,728 | 1362,271 | -,045 | -,787 | ,434 |
| compressionratio | 239,532 | 155,852 | ,117 | 1,537 | ,129 |
| horsepower | 140,229 | 25,818 | ,726 | 5,432 | ,000 |
| peakrpm | -,620 | 1,214 | -,035 | -,511 | ,611 |
| citympg | -285,654 | 348,252 | -,213 | -,820 | ,415 |
| highwaympg | 198,945 | 281,721 | ,159 | ,706 | ,482 |
| a. Dependent Variable: price | | | | | | |