# US Auto Production and Price Prediction in the Context of Multiple Regression Analysis

Qilin Li*

Department of Economics, University of California, Irvine, the U.S.

*Corresponding author: qilinl2@uci.edu

**Abstract.** After the eruption of COVID-19, the worldwide supply chain had been seriously affected and further impaired. Logistics and semiconductors production are not going smoothly during this period and the supply and demand supply balance had been weakened by this pandemic. There is a noticeable trend that the US auto production rate is declining year after year, whereas the car purchasing price as well as the gasoline price is rising each year. This paper investigates US auto prices based on multiple factors regression model. According to the analysis, there is a direct correlation between semiconductor production and the auto car production rate. In addition, US car production rate follows along with CPI (Consumer Price Index). By comparing this paper's results with the movements of CPI and unemployment rate, the study will derive an approximate prediction upon the future auto car prices in the US. This methodology for car price prediction is not very precise, yet it can still reflect some general opinions or human being's faith in the future car industry. The discussed model is the one to capture this attitude and translate it into figures and graphs which can give a direct assumption regarding to the people's attitude. Otherwise, it can simply reflect possible future investing move one rational investor may take. These results shed light on guiding further exploration of stock price variation through multiple factors varies from raw material to people's faith in the stock markets.

**Keywords:** Regression; FRED; Semiconductor; Price; Car.

## 1. Introduction

The price of US car is a very crucial indicator to most people upon determining whether to purchase a major asset or not. Contemporarily, the COVID-19 pandemic has strongly negative affected a large number of fields. Thereinto, automobile industry is one of the industries that has been strongly influenced. After the major pandemic of COVID-19, the auto market had an overall tendency of rising tendency along with many visible car related products or services, e.g., heightened oil prices, or Uber fees [1-4].

This paper is aimed at investigating how current supply of raw materials of auto and market faith in auto would lead to different prices or variation in auto prices. In addition, this paper is intended to use test results to predict future auto price tendency based on the constructed factors. There are many factors required consideration before actually buying a new auto. For instance, car fuel, the capacity of car, brands, and age of car, or even mileage are all very important and crucial determinants of future car price.

In this paper, the variables and determinants used are not from actual prices of car components and car mileage or anything directly related to the production of the auto itself. This study focuses on other external and broader Macroscopic factors, e.g., the CPI index, Oil prices, Semiconductor prices in general to make an overall estimation of future car price fluctuation and current faith on US auto markets.

One precise prediction on US auto prices asks a series of differentiated features and components [5], from computing chips made from semiconductor to COVID-19 that causes the shortage in many cars' raw materials as well as potential customers. Another critical distinguisher in auto prices is the car brands and whether it's used or not. Luxury and limited car types normally would gain a far higher predicted prices than actual cars. In consideration of this, a bigger data quantity is required [6].

This paper had the following assumption. The US auto car production rate is negatively affected by the supply amount of semiconductor and positively related to Production Price Index. Because semiconductor is vital for many car chips and a higher consumer surplus would cause more production. The Gasoline price and the Consumer Producer Surplus would have an inversely plotted graph with US auto production [7, 8].

This paper assumes a higher complimentary product price would lower the urges of car demands and in the end car production. A rising consumer price index would also indicate a decline in demands and in the end producer's willingness to produce more cars. The US unemployment rate is not positively correlated with US auto car production for there won't be as many consumers in the car purchasing markets when there is a heightened rate of job loss.

To be specific, this research will use multiple regression accompanied with some machine learning algorithms like Sklearn to better market price prediction [9]. Since it's hard to precisely predict the actual tendency of price variation simply using the multiple regression, this paper added a little machine learning algorisms to better facilitate and accelerate the production of reliable test results and related graphs and formulae [10, 11].

In the field of supervised machine learning, the regression-based method has been proven rather useful and solid in predicting continuous variables and numerous datasets [12, 13]. The data used will be from FRED and mainly this paper is using R-squared to see changing tendency. Variables and formulas will be discussed in later sections with the visualization of test results and prediction tendencies.

Based on utilizing US auto prices as the independent variable, this paper can run regression and see changing trajectory of this factor in response of other influential factors. Limitations and graphs will be at the later sections of this paper and datasets along with results will also be presented. The rest part of the paper is organized as follows. The Sec. 2 will descirbe the basic information of the data and introduce the setup in US auto production multiple regression estimation. Subsequently, the procedures for carrying out the whole research and the coding will be clarified. Afterwards, the empirical analysis will be presented with corresponding discussions of limitations and prospects. Eventually, a brief summary of the whole research as well as guideline for further study will be given in Sec. 5.

## 2. Setup in US auto production multiple regression estimation

In Auto production estimation, this study used datasets from FRED (Federal Reserve Economic Data) to support our estimation and prediction. To be specific, this research has chosen and set following independent variables:

- PCU33443344 which is Producer Price Index by Industry: Semiconductor and Other Electronic Component Manufacturing;
- CPALTT01USM657N which is Consumer Price Index: Total All Items for the United States;
- UNRATE which is Unemployment Rate;
- PCU2111112111111 which is Producer Price Index by Industry: Crude Petroleum and Natural Gas Extraction: Crude Petroleum;
- PNGASEUUSDM which is Global price of Natural gas, EU.

In addition, the dependent variable is DAUPSA which stands for Domestic Auto Production. In order to split the train and test set, this paper set DAUPSA as Y and all other independent variables as X, and used a test size of 30% and random state of 0 in our splitting of dataset based on the python scikit-learn API.

Based on above operations, the whole original dataset is divided into four subsets named x_train, x_test, y_train, y_test. For better visualization, the evolutions of the above variables are given in Fig. 1. Apparently, it can be noticed that some of the variables are correlated with each other.

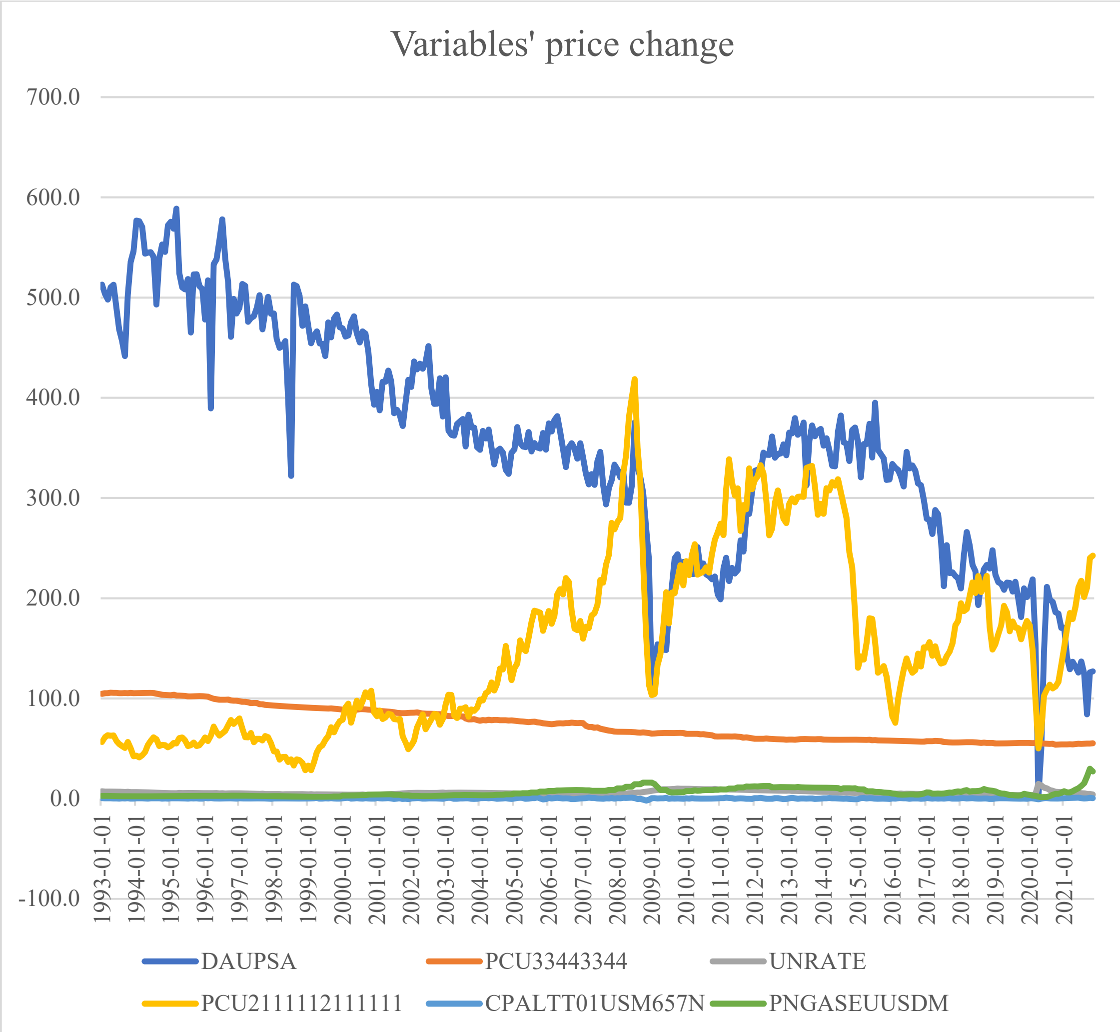


**Fig. 1** Independent Variables evolution for the dataset.

## 3. The Procedures in US auto production multiple regression estimation

To give a better routine of the whole study, the procedures will be given as follows. The Step one of this estimation is gathering data from the FRED database. This paper compiled six different indexes from the same periods from 1993-01-01 to 2021-11-01 which totals up to 347 rows of data in an excel sheet.

Then, to analyze the data, the research imported the excel sheet from our file section and imported several Python build-in libraries like Pandas and Sklearn for further usage in our coding process. This paper used pandas to receive and generate our subject data frame and then imported statsmodels.formula.api to be able to further run our regression.

This paper also imported NumPy in the next line for further calculation. The DAUPSA (Domestic Auto Production) is isolated and become Y in our test and all other independent variables are X in our test. The test used a test size of 0.3 and random state of 0 in our splitting of dataset. The original dataset is divided into four subsets named x_train, x_test, y_train, y_test. Subsequently, the research imported rain-test split as well as LinearRegression from Sklearn model selection. Furthermore, this test used matplotlib to draw the scatterplot of reality and prediction to show our accuracy of prediction. In the end, this study can show the differences in test values and predict values.

## 4. Results & Discussion

The results were able to generate a simple yet comprehensive model summary table which provides this study with the R squared, F-statistics as well as adjusted R squared. According to the results, R-squared is considered one of the measures upon the overall quality of dependent variable prediction. Here, the paper has an R squared greater than 0.76 which then can be considered a good level of prediction. The R squared is called the coefficient of determination and this research have a value of 0.803 which indicates our independent variables can suggest 80.3% of our dependent variable. An adjusted R squared of 0.80 would strengthen this argument here. The details for the regression results are summarized in Table. 1. According to the results, the OLS Regression equation is then:

$$DAUPSA = -110.8466 + PCU33443344 + CPALTT01USM657N + UNRATE + PCU2111112111111 + PNGASEUUSDM \quad (1)$$

**Table 1**. Regression results.

| Dep. Variable: | DAUPSA |
|---|---|
| Model: | OLS |
| Method: | Least Squares |
| Date: | Thu, 04 Aug 2022 |
| Time: | 00:58:55 |
| No. Observations: | 347 |
| Df Residuals: | 341 |
| Df Model: | 5 |
| R-squared: | 0.803 |
| Adj. R-squared: | 0.800 |
| F-statistic: | 277.9 |
| Prob (F-statistic): | 6.50e-118 |
| Log-Likelihood: | -1857.7 |
| AIC: | 3727 |
| BIC: | 3750 |

As already mentioned in above statements, the specific data and test results are presented in Table. 1. Seen from the results, taking UNRATE (unemployment rate) as specific example, the coefficient of -13.1114. This value actually indicates that there will be 13.11% decrease in our dependent variable for 1% increase in unemployment rate. The p values of all independent variables chosen here are all 0 in this case which means the paper have statistically significant coefficients.

This paper did a model prediction after train-test split of the dataset, and then took five values of independent variables from the first row to see our effect of prediction. Actual values of independent variables: [104.6, 7.3, 56.6, 0.493305, 2.70], the final predicted value is then 498.71 which is close to the actual value of 512.9. The reality versus prediction as shown in Fig. 2 denoting that theprediction is roughly in line with reality data.

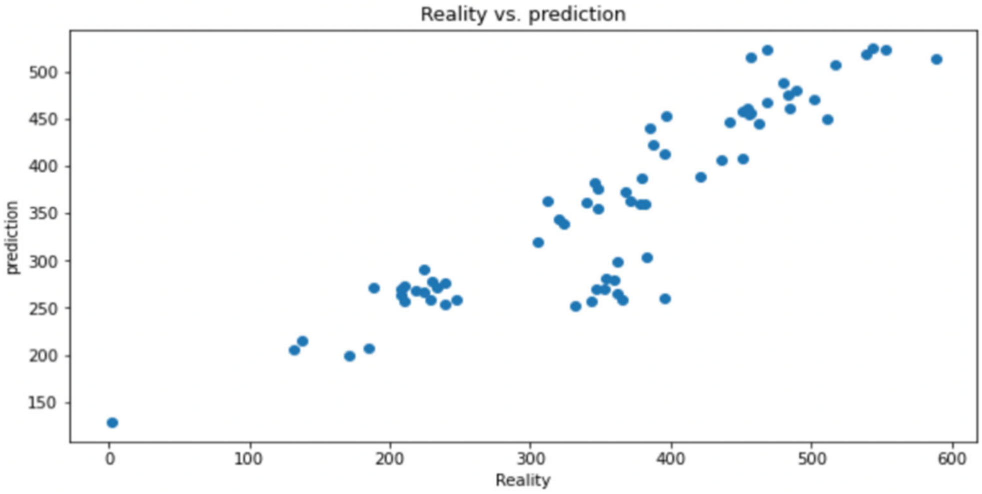


**Fig. 2** Prediction of Auto prices

## 5. Conclusion

In conclusion, this paper investigates the U.S. automobile industry based on multiple factor analysis and regression analysis. To be specific, the multiple regression is utilized for US auto production and results that are very statistically significant as well as a high R squared value are obtained. These predicted values are roughly in line with actual values. The US auto car production rate is negatively affected by the supply amount of semiconductor and positively related to Production Price Index. The Gasoline price and the Consumer Producer Surplus would have an inversely plotted graph with US auto production. The US unemployment rate is not positively correlated with US auto car production.

Nevertheless, the multiple regression used on the paper is still limited and can be improved. For example, there is a relatively small sample size in regard to a huge auto market. There are also not many dependent variables being considered like gold prices or car brands status in people's mind. In the future, there can be more usage of deep learning scenarios and the state-of-art machine learning approaches in predicting the general auto prices using more substantial and up to date data, including Lightgbm, Xgboost, Neural network based-model (CNN, RNN). Overall, these results offer a guideline for how one might be able to use multiple regression to roughly make a prediction on how the future car prices might change in relation to the many supplies of raw materials or other Macro components' index fluctuation.

## References

[1] Wen W, Yang S, Zhou P, et al. Impacts of COVID-19 on the electric vehicle industry: Evidence from China. Renewable and Sustainable Energy Reviews, 2021, 144: 111024.

[2] Belhadi A, Kamble S, Jabbour C J C, et al. Manufacturing and service supply chain resilience to the COVID-19 outbreak: Lessons learned from the automobile and airline industries. Technological Forecasting and Social Change, 2021, 163: 120447.

[3] Ramani V, Ghosh D, Sodhi M M S. Understanding systemic disruption from the Covid-19-induced semiconductor shortage for the auto industry. Omega, 2022, 113: 102720.

[4] Chandrasekar R, Chandrasekhar S, Sundari K K, et al. Development and validation of a formula for objective assessment of cervical vertebral bone age. Progress in orthodontics, 2020, 21(1): 1-8.

[5] DATAMONITOR, Used Cars in Europe, Industry Profile, Reference Code: 0201-0750, December 2007.

[6] GELMAN A., Hill, J. Data Analysis Using Regression and Multilevel Hierarchical Models. Cambridge University Press, New York, USA, 2006.

[7] Gegic E, Isakovic B, Keco D, et al. Car price prediction using machine learning techniques. TEM Journal, 2019, 8(1): 113.

[8] Kiran S. Prediction of resale value of the car using linear regression algorithm[J]. Int. J. Innov. Sci. Res. Technol, 2020, 6(7): 382-386.

[9] Noor K., Jan S. Vehicle Price Prediction System using Machine Learning Techniques. International Journal of Computer Applications, 2017, 167(9), 27-31.

[10] Richardson M S. Determinants of used car resale value. Colorado College., 2009. Retrieved from: https://digitalcc.coloradocollege.edu/islandora/object /coccc%3A1346

[11] Pudaruth S. Predicting the price of used cars using machine learning techniques. Int. J. Inf. Comput. Technol, 2014, 4(7): 753-764.

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

[13] Yan X, Su X. Linear regression analysis: theory and computing. world scientific, 2009.