USED CAR PRICE ESTIMATION: MOVING FROM LINEAR REGRESSION TOWARDS A NEW S-CURVE MODEL

Fadzilah Salim*
Faculty of Electrical and Electronics Engineering Technology, Universiti Teknikal Malaysia Melaka

Nur Azman Abu
Faculty of Information and Communications Technology, Universiti Teknikal Malaysia Melaka

ABSTRACT

A simple linear regression is commonly used as a practical predictive model on a used car price. It is a useful model which carry smaller prediction errors around its central mean. Practically, real data will hardly produce a linear relationship. A non-linear model has been observed to better forecast any price appreciation and manage prediction errors in real-life phenomena. In this paper, an S-curve model shall be proposed as an alternative non-linear model in estimating the price of used cars. A dynamic S-shaped Membership Function (SMF) is used as a basis to build an S-curve pricing model in this research study. Real used car price data has been collected from a popular website. Comparisons against linear regression and cubic regression are made. An S-curve model has produced smaller error than linear regression while its residual is closer to a cubic regression. Overall, an S-curve model is anticipated to provide a better and more practical estimate on used car prices in Malaysia.

Keywords: S-shaped curve, S-curve model, used car prices, price modelling.

1. INTRODUCTION

A simple linear regression is the oldest type of regression technique popularly used specifically in prediction, forecasting, or estimation. A least square method for fitting an equation from linear parameters has been developed by C.F. Gauss in the 1800s. It describes a relationship between an independent predictor ($X$) and a dependent predicted variable ($Y$). Today, fitting a linear model is still one of most frequent activities by many researchers, be it in statistical practices or other research areas.

A linear regression has also been used to build a pricing model. A pricing model, or financial modelling, has been developed centuries ago to drive greater profits and revenues. One approach to fix a correct price of an item is by an estimation. Many researchers have used a linear regression

* Corresponding Author: Universiti Teknikal Malaysia Melaka, Faculty of Electrical and Electronics Engineering Technology, 76100 Durian Tunggal, Melaka, Malaysia; H/phone: 0196691967; E-mail: fadzilah@utem.edu.my
to explain pricing model and found a positive linear relationship between a target variable and its predictor (Balcilar et al., 2018; Pamane & Vikpossi, 2014).

However, due to the nature of a linear regression, its accuracy in describing the linear relationship in real scenarios can be imprecise. Unfortunately, real data will just barely produce a linear relationship. A valuable item does not appreciate or depreciate in a consistently linear fashion with respect to time. Prediction errors grow larger as it moves further from a central mean in a linear regression model. Mamipour and Jezeie (2015) in their study have found that a non-linear relationship exists between independent and dependent variables in economics and finance. Ferrari, Marchese and Tei (Ferrari et al., 2018) has evaluated various effects of business elements on shipbuilding activities in relation to changing economic-cycle phases and found that using a non-linear econometric approach can easily outperform a linear approach. Seifbarghy et al. (2013) have come up with two-estimated cost functions and assessed their accuracy through a simulation. They discovered that a non-linear regression has performed better than linear regression techniques in estimating the total cost function of a two-echelon inventory system with one central warehouse and several non-identical retailers. Thus, a linear regression may not appear to be the best technique to approach price prediction, forecasting, or estimation.

An immediate objective of this research study is to construct a new S-curve model as a practical tool to estimate on a price of a used car. A preliminary research prior to this research study has been carried out to establish a new S-curve model for better and practical estimation on a used car price (Salim & Abu, 2018). Hence, this study is an extension to Salim and Abu (2018) by using a different data set to validate the model. In addition, this paper will highlight an incremental development on an S-curve model based on the S-shaped Membership Function.

There have been many research studies on price prediction previously. However, in the context of Malaysia, many researchers who have conducted serious studies with regard to automotive market have focused more on consumers’ behaviour (Teoh & Mohd Noor, 2015; Islam et al., 2016; Jayaraman et al., 2018; Mashahadi & Mohayidin, 2015; Abdul Wahab et al., 2017). Therefore, developing an S-curve model which offers a more accurate and practical estimate on item pricing at a particular make year has yet to be tapped into. Thus, this S-curve model will provide a significant contribution to an auto market in general.

In the next following sections, some background studies on a linear regression on used car prices in Section 2, and non-linear models and an S-shaped curve in Section 3 will be investigated. Section 4 deals with some experimental details that had been carried out previously. In Section 5 a comparison between linear regression model, cubic regression model, and S-curve models using real data will be highlighted, as well as the proposed algorithm on a new S-curve construction will be introduced. Finally, a brief conclusion is made in Section 6.

2. LINEAR REGRESSION ON USED CAR PRICES

A pricing model in microeconomics is a model on how prices are set within a market for a given product. According to Dolgui and Proth, pricing models are tools that help to better understand an ongoing dynamic pricing is, rather than something used to solve real-life problems. They have
introduced a dynamic pricing model called *the selling curve* which denotes the relationship between price and volume sold at a given period of time (Dolgui & Proth, 2010).

Nowadays, many families own more than one car per household due to current lifestyles. Since the cost of a new car becomes more expensive each year, customers are determined to purchase a good used car. It was reported that 55% of Romanians would buy used cars upon deciding upon buying cars (Oprea, 2010). It was also reported by Wall Street Journal that there was a huge price difference between new and used cars in 2018 (Roberts, 2018). Therefore, used car markets will grow exponentially in order to gain benefit from this situation.

Many research studies on used car price prediction have been carried out previously. Pudaruth (2014) used a machine learning technique to predict the price of used cars. In his research, he made a comparison of the results using four different techniques, that is, multiple linear regression, k-nearest neighbours, Naive Bayes, and decision tree. Ozgur et al. (2016) tried to explore and establish linear regression in predicting the price of used cars. Chen et al. (2017) performed a comparative study of different models to identify the optimal used car price prediction model. They used normalized mean squared error (NMSE) to determine the predictive effect of the models and found that random forest outperformed linear regression. Pal et al. (2018) also used random forest to predict used car prices in his research study.

Oprea (2010) tried to establish a linear regression algorithm for the prediction task, and found that class, year of manufacture, and kilometres (mileage) were the most important attributes that affected the price of used cars. Meanwhile, Puteri and Safitri (2020) analysed the characteristics of a used car, such as age of a car, its mileage, its colour, its type and transmission, in predicting the best selling price of a used car. The result showed that the accuracy of the prediction of used car price would depend on more variables used in the study. In the meantime, Costa et al. (2020) showed good result in terms of predicting true value of used cars using multiple linear regression. At the same time, Rane et al. (2021) developed an efficient and effective used car price predictive model by comparing three regression models namely linear, Lasso, and Ridge regressions.

3. **NON-LINEAR MODEL AND AN S-SHAPED CURVE**

Studies on non-linear models have been carried out for many years. In fact, fitting non-linear models would be the best solution when a linear model is not sufficient. Real data will never produce a linear relationship. A non-linear econometric model has been observed to outperform a linear predictive model. A non-linear model is therefore expected to solve the prediction problem of item appreciation, as well as manage the prediction with better accuracy or better estimation for model pricing in describing real-life phenomena.

A study on a non-linear S-shaped curve has been identified throughout modern computational science. An S-shaped curve is a well-established function, used in many areas, such as neural networks, fuzzy logic, artificial intelligence and so on. One of the S-shaped curve functions, known as *sigmoid function*, is widely used in artificial neural networks to introduce nonlinearity models (Tsai et al., 2015) and back-propagation learning. A non-linear S-shaped curve has also been used to predict how much money could be saved from an amount available in a current account balance.
An S-shaped curve often refers to the special case of a continuous logistic function. A standard sigmoid logistic function has the following mathematical equation:

$$\varphi(t) = \frac{L}{1 + e^{-k(t-t_0)}}$$  \hspace{1cm} (1)

This function was popularized by a mid-19th-century scientist by the name of Pierre François Verhulst, who used it to model population growth in one of his studies. In statistics, this function is quite well-known to describe a cumulative distribution function (cdf) from 0 to 1. It is a continuous and increasing function for all values of $x$. Its slope is always non-negative for all real values of $x$. Its second derivative changes from upward concavity to downward concavity. A typical graph of a sigmoid function from Equation (1) is shown in Figure 1.

![Figure 1: The sigmoid logistic “S”-curve function where $L=1$, $k=1$, and $t_0=0$](image)

An S-shaped curve can be divided into three stages; the graph starts slowly in the first stage, and then accelerates in the second stage, before slowing down again to reach its saturation level in the final stage. Such curves can be used to capture a diffusion process (Meade & Islam, 2003). The logistic S-curve of natural growth is a basic model of the Volterra-Lotka equations, which are reliable for describing and forecasting different forms of competition and technology substitution (Kucharavy & De Guio, 2011).

Many research studies which use S-shaped curves in forecasting and prediction analysis have been carried out. The S-shaped curve has many applications with regard to prediction and forecasting (Kucharavy & De Guio, 2011). A leading forecasting journal, the *International Journal of Technological Forecasting and Social Change*, has published many articles mentioning an S-shaped curve. Further reading on the types of forecasting methods that are related to an S-shaped curve can be referred to Armstrong in his principles of forecasting handbook, which covers all types of forecasting methods from various experts (Armstrong, 2001).
A spline S-shaped Membership Function (SMF) is a mathematical function that has an “S” shaped curve, presented in the following equation:

\[
f(x) = \begin{cases} 
0, & x \leq a \\
2 \left( \frac{x-a}{b-a} \right)^2, & a < x \leq \frac{a+b}{2} \\
1 - 2 \left( \frac{x-b}{b-a} \right)^2, & \frac{a+b}{2} < x < b \\
1, & b \leq x 
\end{cases}
\]  

This function goes from 0 to 1 along its domain from left to right. This SMF has been used in an image enhancement through a fuzzy logic (Patel et al., 2015). An example of the SMF graph with parameters \( a = 0 \) and \( b = 9 \) is shown in Figure 2.

**Figure 2:** A graph of an S-shaped Membership Function with parameters \( a = 0 \) and \( b = 9 \)

4. EXPERIMENTAL DETAILS

Data on used car prices from Mudah.my website have been collected. A car model Perodua Alza 1.5 with automatic transmission has been chosen in this research study. A total of 42 prices was collected, based only on the price of the used cars and their year of manufacture, from 2009 until 2018. Other features, such as mileage, colour, and sale location, which hardly affect the price of the used cars were not considered. The data have been analysed using three types of models, namely simple linear regression, cubic regression, and S-curve model.
Figure 3: (a) A linear regression on used Perodua Alza; (b) A cubic regression on used Perodua Alza; (c) An S-shaped curve on used Perodua Alza using direct application of an S-shaped Membership Function.
5. RESULTS AND DISCUSSION

Three graphs were produced based on the data collected, as shown in Figure 3. Figure 3 (a) and (b) show the graphs of a linear regression and a cubic regression from used Perodua Alza, respectively. Figure 3 (c) shows the graph of an S-shaped Membership Function (SMF) on used Perodua Alza. The equation will be taken as a piecewise function, symmetric at the centre point, as stated in Equation (2). The minimum and maximum values of the data need to be estimated and adjusted to produce an accurate result. The SMF model on Perodua Alza, as shown in Figure 3 (c), will then be established in Equation (3).

\[
f(x) = \begin{cases} 
0, & x \leq 2008 \\
2 \left(\frac{x-2008}{11}\right)^2, & 2008 < x \leq 2013.5 \\
1 - 2 \left(\frac{x-2019}{11}\right)^2, & 2013.5 < x < 2019 \\
1, & x \geq 2019 
\end{cases}
\]

The minimum price for the model was estimated at RM28,600, while the maximum price was estimated at RM50,000 on both end by their local means. Thus, \(\hat{y} = 28.6 + 21.4f(x)\).

In developing a new S-curve model, a piecewise function provides an avenue to provide flexible concavity to the second derivative in the best-fitting option. The S-curve dynamically moves towards the lowest possible residual or mean squared error (MSE) within the degree range of [2, 3]. The first equation on the left of the centre point will carry a lower degree than the second equation to the right of the centre point. This S-curve model construction algorithm is proposed as follows:

- Let \(x_0, x_1, \ldots, x_{n-1}\) be independent random variables.
- i) Compute the central mean point, \((\bar{x}_m, \bar{y}_m)\).
- ii) Compute the average/mean value within the neighbourhood values near or closer to \(x_0\) and \(x_{n-1}\), in other words, compute for \((x_0, \bar{y}_0)\) and \((x_{n-1}, \bar{y}_{n-1})\).
- iii) The curve to the left of the central mean point will be regressed by
  \(\hat{y} = 2 \left(\frac{x-a}{b-a}\right)^\alpha, \quad a < x \leq \bar{x}_m, \quad 0.1 < \alpha < 4\)
- iv) The curve to the right of the central mean point will be regressed by
  \(\hat{y} = 1 - 2 \left(\frac{x-b}{b-a}\right)^\beta, \quad \bar{x}_m < x < b = x_{n-1}, \quad 0.1 < \beta < 4\)

Therefore, the proposed new S-curve model will be:

\[
\hat{y} = f(x) = \begin{cases} 
2 \left(\frac{x-a}{b-a}\right)^\alpha, & a < x \leq \bar{x}_m, \quad 0.1 < \alpha < 4 \\
1 - 2 \left(\frac{x-b}{b-a}\right)^\beta, & \bar{x}_m < x < b, \quad 0.1 < \beta < 4 
\end{cases}
\]

where, \((\bar{x}_m, \bar{y}_m)\) is the central mean point, and \(\bar{x}_m = \frac{a+b}{2}\).
To achieve a better performance using Equation (4), exponents $\alpha$ and $\beta$ have been estimated at 2 and 3, respectively. Minimum and maximum prices remained the same levels at RM28,600 and RM50,000 respectively. Thus, a new S-curve model has the following equation:

$$f(x) = \begin{cases} 
0, & x \leq 2008 \\
2 \left(\frac{x-2008}{11}\right)^2, & 2008 < x \leq 2013.5 \\
1 - 2 \left(\frac{x-2019}{11}\right)^3, & 2013.5 < x < 2019 \\
1, & x \geq 2019
\end{cases} \tag{5}$$

Hence, $\hat{y} = 28.6 + 21.4f(x)$. Figure 4 depicts the resulting graph from an Equation (5).

**Figure 4:** A dynamic S-curve model on used Perodua Alza from Equation (5)

To evaluate the performance of those models, mean squared errors on a simple linear regression, a cubic regression, an S-shaped Membership Function (SMF) and a new S-curve model using Equation (5) have been computed. Table 1 compares all residuals on all the four models.

<table>
<thead>
<tr>
<th>Type of regression/model</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear</td>
<td>17.08</td>
</tr>
<tr>
<td>Cubic</td>
<td>12.87</td>
</tr>
<tr>
<td>SMF</td>
<td>19.93</td>
</tr>
<tr>
<td>New S-curve</td>
<td>14.61</td>
</tr>
</tbody>
</table>
On one hand, Table 1 shows that the mean squared error from a cubic regression is the smallest among all models. In other words, for this data set, a non-linear cubic regression performed the best fit among all models. However, looking back on both far ends of the cubic regression, the behaviour of the curve has gone into negative derivative territories, as shown in Figure 3 (b). The price prediction at both ends of non-linear cubic regression infringe on maintaining a positive derivative throughout a car make year and thus makes it impractical. In the meantime, the new S-curve model was found to be quantitatively closer to a cubic regression. The smaller the mean squared error, the closer it is in finding the line of best fit. Depending on the data, it may be inconceivable to achieve an extremely small value for the mean squared error due to the data’s practical nature. From this practical perspective, the performance of a new S-curve model is practically better and more efficient than a non-linear cubic regression in this automotive market.

Table 2. Target price of used Perodua Alza for make year 2010, 2015 and 2020

<table>
<thead>
<tr>
<th>Type of regression model</th>
<th>Make Year</th>
<th>Make Year</th>
<th>Make Year</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2010</td>
<td>2015</td>
<td>2020</td>
</tr>
<tr>
<td>Linear</td>
<td>RM 28,381</td>
<td>RM 41,878</td>
<td>RM 55,375</td>
</tr>
<tr>
<td>SMF</td>
<td>RM 30,014</td>
<td>RM 44,339</td>
<td>RM 49,646</td>
</tr>
<tr>
<td>New S-curve</td>
<td>RM 30,014</td>
<td>RM 52,058</td>
<td>RM 49,967</td>
</tr>
</tbody>
</table>

Table 2, on the other hand, shows the target price for a salesman to sell on make year 2010, 2015 and 2020 respectively, based on Figure 3 (a, c) and Figure 4. For make year 2010 and 2015, both SMF and a new S-curve models offer much higher price estimation on a used Perodua Alza. The curve on the right of the central mean point has been pushed up further so that a salesman is expected to be able to push the price up in order to earn more profit since the car is still relatively not that old.

On the contrary, for the make year 2020, a simple linear regression will be practical to project a price of an incoming new model. However, a newly purchased car will drop in value significantly upon sale. At the same time, a car which is 20 years of age practically will have a minimum junkyard market value even though the car is not yet considered a classic or vintage. However, with linear regression, the price will be going further down below the x-axis which estimated at a negative value. In Malaysia, a spare part still carries a significant market value even though the car is not running well anymore. Therefore, the S-curve model provides a more accurate estimation of the selling price of used car at any particular make year.

In this research study, a real-life scenario needs to be imitated in a regression model. It should be noted a basic prediction model has always been a simple linear regression. However, a valuable item does not depreciate or appreciate linearly over time in a real scenario. Thus, linear regression may not be the best practical tool for prediction analysis on a real phenomenon. Furthermore, a linear model assumes that an occurrence of price variation (standard deviation) progressively increases as a make year (variable x) move away from a central mean. A simple linear regression will provide a predictable price near the central mean with a lower standard error. However, its standard error grows and a price estimate becomes more unpredictable on both further ends. The variation of prices on both ends gets larger as (the standard deviation) of the normal distributions on the predicted prices get wider, the make year moves further away from the center as shown in Figure 5.
On the contrary, an S-curve model controls the expansion of the standard deviation and pays more attention to one side of the distribution at each end. In other words, as a car gets older, the distribution is skewed to the right, and as a car is newer, the distribution is skewed to the left, as shown in Figures 6, 7, and 8. This S-curve model will also help in projecting the price of an upcoming new car, instead of using a linear price projection from the linear regression model.
The price of used cars is expected to increase based on how newer the car is. Even though a cubic regression has produced the smallest MSE in this research study, the behaviour of the curve at both far ends has already gone into the negative derivative. These phenomena contradict the principle of price increase over make year of a car. Even though the new S-curve model shows a slightly higher MSE, it gives a closer average price estimate in a real-world situation. Thus, this research aims at developing an S-curve model for a better pricing. Based on the outcome and some findings from this research, the authors have further developed the S-curve model by concentrating only at the maximum price estimate on an auto pricing scheme along its domain make year (Salim & Abu, 2020).

6. CONCLUSION

Buying a good used car is becoming a trend in most countries across the globe due to an increase in prices of new cars every year. A used car market plays a more crucial and significant role in
daily life activities. However, there are hardly any specific pricing model which can be used by a car dealer in estimating the price on used cars.

A simple linear regression over a make year, although is useful and powerful, it is not practical or dynamical to portray a real auto market phenomenon. An S-shaped curve model does give some significant proximity to a real-world situation. An S-shaped curve model has the potential to represent a better model in the world of prediction or forecasting in real-life scenarios. This research study is intended to achieve realistic pricing on a used car. It is reckoned that this S-curve model would give a more accurate price estimation on used cars in Malaysia.

ACKNOWLEDGMENT

The authors would like to thank Universiti Teknikal Malaysia Melaka (UTeM) for the encouragement and support to carry out this research study.

REFERENCES


