A study of automobile market structure prediction based on multivariate logit model with embedded Newton's method

Jiao Yufan*, Zhang Bingyu, Liang Panchun

Automotive Data of China Co., Ltd., Dongli District, Tianjin, China

Abstract: In order to deeply understand the factors that consumers focus on when choosing models with different technological routes, and to predict the structure of the future automobile market accordingly, a multivariate logit model incorporating Newton's method of optimization is used, which is able to accurately measure the weights of the various factors that influence the process of choosing models by consumers, in order to reflect the degree of heterogeneity of consumers' preferences for these factors, and to predict the future market share of the different technological routes on the basis of this model. Based on this, it predicts the share of different technological routes in the future market. The results of this study show that when choosing a vehicle of different technological routes, convenience of replenishment is the most important factor, with a preference weight of 0.39. In contrast, the imitation coefficient has the lowest preference weight of 0.11, which suggests that consumers are more likely to take into account their personal experience rather than the choices made by others in their vehicle purchasing decisions. In addition, in the prediction of future technological routes in the automobile market, it is also found that by 2025, consumers' tendency to choose new energy vehicles will be more than 50%, and by 2030, the proportion will be more than 70%. Meanwhile, the market share of sales of traditional fuel vehicles is expected to fall to 17% by 2035.

1. Introduction

In 2009, the incubation of new energy vehicles started from “10 cities and 1,000 vehicles” to the current “1,000 counties and 10,000 towns” demonstration activities, the rapid expansion of new energy vehicles has brought more choices for consumers to buy vehicles. Although new energy vehicles are still unable to reach the level of traditional fuel vehicles in terms of technological maturity[1-2], supporting infrastructure construction, residual value of end-of-life recycling, and mileage, they have demonstrated greater advantages in terms of cost of use, subsidies for vehicle purchases, and intelligent assisted driving[3]. In addition to the fierce competition between battery electric Vehicles (BEVs) and plug-in hybrid electric vehicles (PHEVs) and internal combustion engine fuel vehicles (ICEs) and the gradual emergence of consumer preference for new energy vehicles[4], the competitive relationship between BEVs and PHEVs has been a hot topic of industry concern[5]. Currently, research on the new energy vehicle market focuses on the sales forecast of the overall new energy market from a time-series perspective[6-8], and little research has been done on the structural changes of different technological routes within the automobile market. As “economists”, consumers will make decisions based on the principle of utility maximization when they are faced with the differences of different fuel types in the process of purchasing a vehicle[9].

Logit model, as a more classical model in consumer choice theory, has unique advantages for solving preference and probabilistic choice problems. In 2009, Ma Jun et al. used a combination of AHP and logit regression model to predict the market of new energy vehicles, but the AHP model unavoidably interferes with subjective factors when determining the weights of influencing factors[10]. In 2020, Liu Bin et al. used the Nested multivariate logit model to predict the market share of each technological routes of new energy vehicles, but when judging the influencing factors, it still did not use an effective assessment method[11]. Based on the above analysis, this study embeds Newton's method on the basis of logit model to improve it, and maximizes the avoidance of human factor interference in determining the weights of influencing factors.

2. Model Introduction

2.1. Multivariate Logit Model

In statistics and machine learning, Logistic Regression is a widely used statistical method for classification problems. However, when classification problems involve multiple categories, the standard binary logit model is no longer applicable. For this purpose, the multivariate logit model, which is a generalized linear model, is introduced to deal with the case of discrete

*Corresponding author: jiaoyufan@catarc.ac.cn

random variables with multiple classes of dependent variables. Firstly, the binary logit model is introduced.

In empirical research, it is common to encounter situations where the explanatory variables are categorical variables. In some contexts, the explanatory variables are either/or binary choice variables, i.e., the familiar 0-1 variables, in which case they should be estimated using a binary logit model, and in order to introduce the binary logit model, the 0-1 distribution is introduced first.

For a given sample, y_i can be regarded as a realized value of the random variable Y since the explanatory variable Y takes the value of 0 or 1:The probability of Y taking 1 is π_j and the probability of Y taking 0 is $1-\pi_j$. In this case, the random variable Y is obeying a 0-1 distribution with parameter π_j.

The distribution law of Y can be calculated as: $P=(Y=y_j)=\pi_j^{y_j}(1-\pi_j)^{1-y_j}, y_j=0,1$. It is easy to show that both the expectation and variance of Y are determined by π_j. Any factor that affects the probability affects not only the mean but also the variance of the observations. This shows that the linear regression model cannot be used to analyse dichotomous variables because the linear regression model assumes that the variance is fixed.

In order to make the above model more resilient, it is assumed that the probability π_j is affected by a set of variables, set to X_j^T .A very intuitive approach is to set up the relationship between the two as a linear function: $\pi_j=X_j^T\beta$, where β is a vector of coefficients, which is often referred to as a linear probability model. The main drawback is that since π_j on the left side of the equation represents the probability, it must lie between 0 and 1, and the linear combination term on the right side may take on any value, it is difficult to ensure that the model's predicted values lie within a reasonable range without imposing strict constraints on the model. Therefore, the probabilities must be transformed to remove the constraints on the range of their values, and then the transformed values are set as a linear function of the explanatory variables. The process consists of the following two steps:

(1) Define the odds ratio according to the probability π_j: $\Omega_j=\frac{\pi_j}{1-\pi_j}$, and it is clear that the odds ratio can take any non-negative value, eliminating the upper bound constraint.

(2) Taking the logarithm of the odds ratio yields: $\ln(\Omega_j)=\ln(\frac{\pi_j}{1-\pi_j})$, eliminating the lower bound constraint.

At this point it is assumed that the logit transform of the probability π_j obeys a linear model, that is, $\ln(\frac{\pi_j}{1-\pi_j})=X_j^T\beta$. Since the logit transform is consistent with one-to-one correspondence, taking the inverse function of the logit function yields:

$$P(Y=y_j)=\pi_j(X_j)=\frac{\exp(X_j^T\beta)}{1+\exp(X_j^T\beta)} \quad (1)$$

However, in many cases the explanatory variables involve more than three categorical variables. In these contexts, a multivariate logit model is needed for estimation, similarly it can be obtained if Y has multiple values, defined: $P(Y=y_j)=\pi_j, j=1,2,...,J$.

Assuming that the probability π_j is affected by a set of variables set to X_j^T , the multivariate logit model can be viewed as implementing a joint estimation of multiple binary logit models constituted by pairing the various types of choice behaviors in the explanatory variables two by two, yielding the predicted probability of each choice as: $\pi_j=P(Y=y_j\mid X)=\frac{\exp(X_j^T\beta_j)}{\sum_{j=1}^{J}\exp(X_j^T\beta_j)}$. At this point, Y is said to be a multivariate logit model.

2.2. Newton method

Most equations do not have a root form, so finding exact roots is very difficult or even unsolvable, making it especially important to find approximate roots of equations. Newton's method is a classical iterative algorithm for solving numerical optimization problems and roots of nonlinear equations, which searches for the optimal solution of a function by continuously approximating its roots or minima. Its greatest advantage is that it has square convergence in the neighborhood of a single root of an equation, and the method can also be used to find heavy and complex roots of an equation.

Let r be a root of the function $f(x)=0$. x_0 is chosen as the initial approximation of r . After the Taylor expansion of $f(x)$ in a neighborhood of x_0 , the first two terms of the linear expansion are taken and made 0, that is: $f(x_0)+f'(x_0)(x-x_0)=0$. Taking this as an approximate equation for the equation $f(x)=0$, if $f'(x_0)\neq 0$, then its solution is : $x_1=x_0-\frac{f(x_0)}{f'(x_0)}$. This yields an iterative relation for Newton's method: $x_{n+1}=x_n-\frac{f(x_n)}{f'(x_n)}$. In the case of multivariate functions, it is sufficient to replace the first-order derivatives with Jacobi matrices and the second-order derivatives with Hessian matrices.

3. Empirical analysis

3.1. Selection of indicators

Before the multivariate logit model based on Newton's method predicts the share of each technological routes in the automobile market, the influencing factors need to be screened and quantified. Taking into account the current considerations of consumers when choosing different fuel types of automobiles and the research results of current scholars[10], the influencing factors in this study are finally determined as the total cost, the convenience of replenishment, the mileage and the imitation coefficient of each technological route.

(1) Total cost

The total cost includes the whole process from the acquisition to the use to the end-of-life recycling of this technological route, specifically involving the acquisition cost, use cost and depreciation cost. Among them, the acquisition cost takes into account the impact of changes in the prices of major parts and components of the model, the decrease in R&D cost brought about by the increase in scale effect, and the increase in cost brought about by the withdrawal of the purchase tax on the acquisition cost. The utilization cost mainly refers to the fuel cost paid by different fuel types in the use of vehicles, and takes into account the decline in fuel consumption rate brought about by technological upgrading. Depreciation cost mainly refers to the cost lost when a vehicle is traded as a used vehicle, using (1 - salvage rate) as the equivalent.

(2) Convenience of replenishment

At present, the service capacity of new energy vehicles replenishment infrastructure is relatively poor compared with that of traditional fuel vehicles, considering that the difference in the convenience of replenishment between new energy vehicles and traditional fuel vehicles in short-distance travel scenarios such as daily commuting and short-distance trips is relatively small. The current pain point of new energy vehicles is that it is difficult to satisfy the demand for public charging piles during long-distance trips, so this study uses the inverse of the service radius of public charging piles (stations) and the service radius of gasoline stations as the indicator of the convenience of replenishment. In this case, assuming that the service radius of gas stations is maintained and the public charging pile (station) infrastructure is accelerated [11], the difference in the convenience of replenishment between traditional fuel vehicles and new energy vehicles is predicted as shown in Table 1:

Table 1. Trends in the convenience of replenishment

Year	Traditional Fuel Vehicles	New Energy Vehicle
2025	1	0.75
2030	1	0.86
2035	1	0.97

(3) Mileage

When consumers choose models with different technological routes, mileage is also an important consideration. Due to the limited capacity of power batteries, low-temperature degradation, and low energy density of new energy vehicles, their mileage is relatively low compared with that of traditional fuel vehicles, and the issue of “mileage anxiety” has always been a major concern for consumers.

(4) Imitation coefficient

The imitation coefficient refers to the fact that the rapid development of the new energy vehicle market has brought about a demonstration and guidance effect among consumers, which will lead to the existence and development of the imitation behavior of “buying new energy vehicles as a trend”. That is, the number of consumers accepting new energy vehicles in the next year is positively related to that of the previous year. This article uses the method of reference [12] to predict the imitation coefficient.

Table 2. Trends in imitation coefficients

Year	ICE	HEV	PHEV	BEV
2025	1	0.31	0.22	0.18
2030	0.69	0.68	0.66	0.53
2035	0	1	1	1

3.2. Market forecasts by technological routes

Assuming the annual sales volume of the passenger vehicle market as a whole consumer, each vehicle consumer has a kind of decision-making scheme according to the fuel type of the vehicle, the correspondence between the serial number of the decision-making scheme and the fuel type is shown in Table 3 (due to the relatively small proportion of fuel cell vehicles in the passenger vehicle technological routes and the unknown policy incentives, it is not taken into account in this study for the time being):

Table 3. Correspondence table between decision-making options and technological routes

Decision-making program	technological route
1	BEV
2	PHEV
3	ICE
4	HEV

Denote the K cost of the model of the j th fuel type as X_j ,and define the fixed utility of the consumer's choice decision option as:

$$V_{nj} = \beta_j^T \cdot X_j = (\beta_{j1}, \beta_{j2}, \beta_{j3}, \beta_{j4}) \cdot (X_{j1}, X_{j2}, X_{j3}, X_{j4})^T \quad (2)$$

Define the probability that consumer n chooses each of the four options:

$$P_{n1} = \frac{e^{V_{n1}}}{\sum_{i=1}^{4} e^{V_{ni}}} \quad (3)$$

$$P_{n2} = \frac{e^{V_{n2}}}{\sum_{i=1}^{4} e^{V_{ni}}} \quad (4)$$

$$P_{n3} = \frac{e^{V_{n3}}}{\sum_{i=1}^{4} e^{V_{ni}}} \quad (5)$$

$$P_{n4}=\frac{e^{V_{n4}}}{\sum_{i=1}^{4}e^{V_{ni}}} \tag{6}$$

In the above equation, $\overrightarrow{\beta_1},\overrightarrow{\beta_2},\overrightarrow{\beta_3},\overrightarrow{\beta_4}$ are vectors of parameters to be estimated, which correspond to the utility weights of the four costs for each of the four fuel type models, and the difference between them corresponds to the consumer bias between different fuel type models. Noting that the values taken in the above equation P_{nj} are independent of n, it is abbreviated to P_{nj} as P_j. The parameters in the model are estimated using the great likelihood estimation method based on data from previous years.

Considering each year's data including all costs, the sum of passenger vehicle sales for that year, N and the sales of the four fuel types for that year are n_1,n_2,n_3,n_4 respectively. Suppose $n-th$ consumer buys the model Z_n, then according to the modeling assumptions, Z_n obeys a discrete distribution with values in the range $\{1,2,3,4\}$ and the probability that j is taken is $P(Z_n=j)=P_j$.

According to $n_j=\sum_{n=1}^{N}I(Z_n=j)$, where is $I(\bullet)$ schematic function, each consumer choice is independent of each other, it is obtained that (n_1,n_2,n_3,n_4) obeys a multinomial distribution with a density function of:

$$f(n_1,n_2,n_3,n_4)=\frac{N!}{\prod_{j=1}^{4}n_j!}\prod_{j=1}^{4}P_j^{n_j} \tag{7}$$

Its log-likelihood function is:

$$l(\overrightarrow{\beta_1},\overrightarrow{\beta_2},\overrightarrow{\beta_3},\overrightarrow{\beta_4})=C+\sum_{j=1}^{4}n_j\times\ln(P_j) \tag{8}$$

C is a constant in the above equation and does not affect the estimation of the parameters, which are omitted later. Here the parameters are included in all P_j. Based on all costs and sales in each year, the log-likelihood function corresponding to the current year's data can be obtained. The log-likelihood function corresponding to all previous years' data is a summation of the likelihood function of each previous year, which includes all previous years' information, and therefore the estimates obtained are also the most explanatory. The final log-likelihood function obtained is:

$$L(\overrightarrow{\beta_1},\overrightarrow{\beta_2},\overrightarrow{\beta_3},\overrightarrow{\beta_4})=\sum_{Year=2016}^{2023}\sum_{j=1}^{4}n_j\times\ln(P_j) \tag{9}$$

The maximum value of the above log-likelihood function, i.e., the consumer's utility weight for each cost, is obtained according to Newton's method. The calculation results are shown in Table 4. The results show that consumers pay more attention to the convenience of replenishment when making model purchases, and the preference of this factor reaches 0.39, followed by the total cost and mileage of comparable degree of preference, respectively, 0.26 and 0.24, and finally, consumers have the lowest degree of preference for the imitation coefficient, which is only 0.11. This result reflects that when choosing different types of vehicles, consumers pay more attention to their own use experience rather than others' choice.

Table 4. Results of weighting of influencing factors

	$\overrightarrow{\beta_1}$	$\overrightarrow{\beta_2}$	$\overrightarrow{\beta_3}$	$\overrightarrow{\beta_4}$
Calculation Result	0.26	0.39	0.24	0.11

After deriving the utility weights of consumers for each cost, the probability of consumers choosing each fuel type can be obtained by bringing it into the probability formula, and this is used as an estimate of the market share for each energy type of model to fit the results of the combined assessment model. In predicting future years, since the prediction has been made for each influencing factor, it is not necessary to repeat the calculation of preferences, and after substituting the predicted values of the influencing factors for any year, the market share of the segments of different fuel types can be predicted for that year.

3.3. Analysis of projected results

Based on the above analysis, an improved multivariate logit model is used to predict the consumer's choice of different technological routes in the automobile market. Based on the calculation results, the accuracy of the model is examined by using the real data from 2017-2023 to compare with the predicted data (as shown in Figure 1). The results show that the prediction results of the multivariate logit model embedded with Newton's method are controlled to be within 0.05 error from the real value, and the box plot of the error distribution is shown in Figure 2.

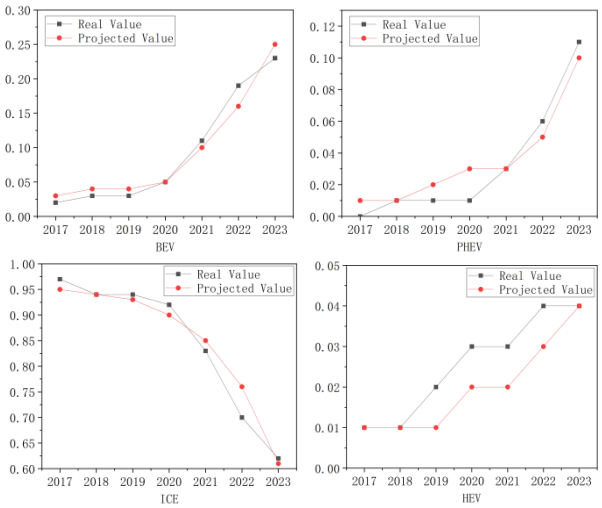

Fig. 1. Probability of consumer choice for different technological routes, 2017-2023

The key time node prediction is shown in Table 5, it is expected that in 2025, the consumer preference for BEV will reach 33%, PHEV will reach 21%, the preference for traditional fuel vehicles will drop to 42%, and the consumer preference for new energy vehicles will exceed that of traditional fuel vehicles. 2030, the probability of consumers choosing new energy vehicles will be close to

70%, with a preference for BEV of 40%, a preference for PHEV of 29%, and a preference for traditional fuel vehicles of 27%. preference 29%, choose fuel vehicles preference 27%. in 2035, consumers' choice tendency for traditional fuel vehicles drops to below 20%, the preference for new energy vehicles rises to 80%, and the market penetration rate of new energy vehicles rises rapidly. consumer preference for HEVs stays at around 4%, with a slight decreasing trend (Figure 3).

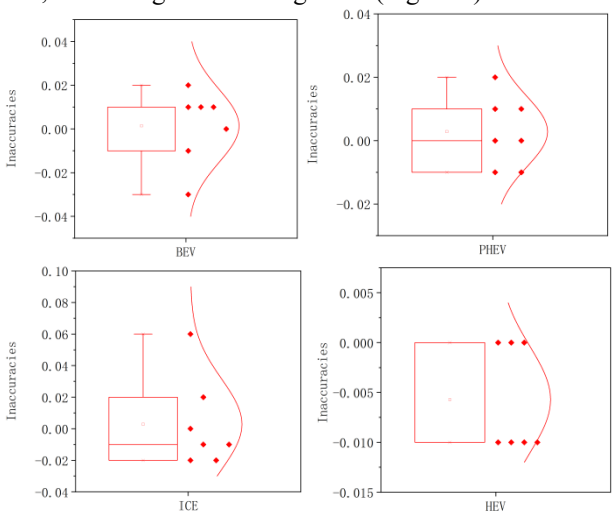

Fig. 2. Box plots of prediction error distribution for different technological routes

Table 5. Consumer preferences for different technological routes, 2025-2035

Consumer Preference	BEV	PHEV	ICE	HEV
2025	33%	21%	42%	4%
2030	40%	29%	27%	4%
2035	55%	25%	17%	3%

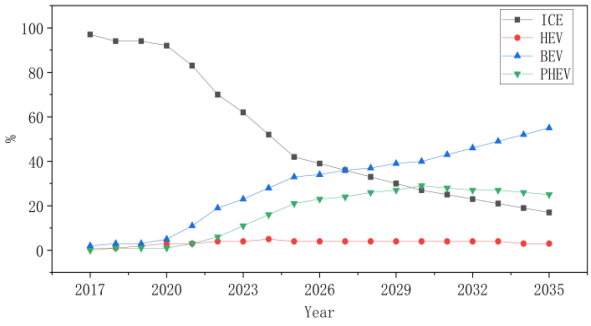

Fig. 3. Consumer preferences for different technological routes, 2017-2035

4. Conclusion

This study quantitatively measures the heterogeneity of consumers' preferences when choosing different technological routes based on the multivariate logit model improved by Newton's method, and the results show that consumers' preferences for the total cost and mileage of different types of vehicles are comparable with about 0.25, and their preferences for the convenience of replenishment are the most important, with a weight of 0.39. The preference for the imitation coefficient is the lowest at 0.11. This result reflects that consumers, as "economic beings", place more importance on their personal experience when making decisions about different vehicle types, and others' decisions have less influence on the judgment of outcomes. This result reflects that consumers, as "economic man", pay more attention to their personal experience when facing the decision of different models, and the decision of others has less influence on the judgment of the result. In addition, the heterogeneity of consumers' choices of different models based on different preferences of influencing factors is also very obvious. The probability of consumers choosing new energy vehicles exceeds 50% in 2025, approaches 70% in 2030, and reaches 80% in 2035, in which the probability of being chosen for PHEVs shows a trend of increasing and then decreasing, while the propensity of choosing traditional fuel vehicles decreases year by year, and is only 17% by 2035, with the probability of being chosen for HEVs remaining at around 4%. The results of this study quantitatively assess the heterogeneous preferences of consumers in choosing models of different technological routes, and provide a theoretical basis for the study of the share of the automobile market.

References

1. Wu JM, Liu C, Ma CQ. Effectiveness of Policy Mix of Battery Electric Vehicle to Sales. Soft Science, 35(03): 129-135 (2021).
2. Yuan X, Liu X, Zuo J. The development of new energy vehicles for a sustainable future: A review. Renewable & Sustainable Energy Reviews, 42: 298-305 (2015).
3. Dudziak A, Drodziel P, Stoma M, et al. Market Electrification for BEV and PHEV in Relation to the Level of Vehicle Autonomy. Energies, 15 (2022).
4. Tal G , Nicholas M A. Studying the PEV market in california: Comparing the PEV, PHEV and hybrid markets.In:Electric Vehicle Symposium & Exhibition.Barcelona. pp. 1-10 (2014).
5. Bjornsson, Lars-Henrik, Karlsson. Electrification of the two-car household: PHEV or BEV. Transportation research, Part C. Emerging technologies,85C(Dec.):363-376 (2017).
6. Liu YQ, Wang M, Wang JY. China New Energy Vehicle Market Forecast Study. Research on Economics and Management, 37(04): 86-91 (2016).
7. Yang XT,Zhang Y. Research on New Energy Vehicle Market Forecasting Based on Bass Model--Taking Shanghai as an Example. Management and Administration, 10: 133-136 (2015).
8. Pei L L, Li Q. Forecasting Quarterly Sales Volume of the New Energy Vehicles Industry in China Using a Data Grouping Approach-Based Nonlinear Grey Bernoulli Model. Sustainability, 11(5) (2019).
9. Liu B, Liu KX, Shi H. China New Energy Passenger Vehicle Market Forecast 2021-2035 - Analysis Based on Discrete Choice Modeling. China Economic & Trade Herald, 05: 44-49 (2020).

10. Ma J, Wang N, Kong DX. Market Forecasting Modeling Study for New Energy Vehicle Based on AHP and Logit Regression. Journal of Tongji University(Natural Science), 37(08): 1079-1084 (2009).
11. Liu L, Liu FH, Gong T. Discussion on the Problem of Charging Station(pile) Site Selection and Constant CapacityOptimization Strategy Based on Charging Demand. Auto Time, 14: 116-118 (2022).
12. Liu YQ, Wang M, Wang JY. The Predictive Research on China's New Energy Vehicles Market. Research on Economics and Management, 37(04):86-91 (2016)