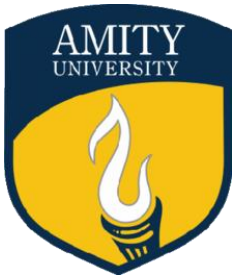
NTCC REPORT



STATISTICAL ANALYSIS OF CAR TYPE PREFERENCE OF EARNERS IN KOLKATA

Name: SHUBHAM KUMAR

B.Sc – Statistics (Sem 5)

Batch: 2017-2020

Enrollment Number: A90555917012

CONTENTS:

ACKNOWLEDGEMENT:

I have got the opportunity to work on discussed topic. So, I acknowledge the valuable contribution of my internal guide **Soumya Banerjee Sir,** external guide **Kuntal Sir,** HOD sir and other respected faculties without whose guidance, support and co-operation my project would not have been possible.

This report, based on two months study, is a part of my BSc. Statistics (H) program, which helped me to gather practical field knowledge.

My sincere gratitude towards Dr. Susanto Tewari, Head of Department, Statistics Dept.(AIASK), who gave me the opportunity to do NTCC project. This has been a great learning experience for me, which helped me enhance my skills, strength and confidence for future opportunities.

ABSTRACT:

Unlike a traditional society in the past, Indian life nowadays has no longer a certain form. Increasing level of complexity in people's activity and diversity of a unit of household composition clearly contribute much to the shift of people's existing traditional travel patterns and vehicle selection behaviors. While vehicle type choice in households has so far been studied extensively, relatively few studies have been devoted to the effect of diverse age and household earnings on vehicle type choice. As vehicle ownership by vehicle type is of much interest to transportation planners and policy makers, the aim of this study is not only to understand the principle of vehicle type choice analysis by following existing methods such as multinomial logit model but also to provide further detailed analysis of vehicle type choice, especially about the influence of various age groups and household earning groups, using the data from a self conducted survey and some secondary sources. This study attempts to identify the feature of four vehicle types; hatchbacks, sedans, SUVs, and luxury cars.

This project has been a great experience for me. There were various information's which I have gathered and has given me a broader picture to this field. This experience and exposure has helped my personal development. This experience has shown me a glimpse of how life is and will be in days to come.

INTRODUCTION:

Origin of the Report:

This report, based on two months study, is a part of my BSc. Statistics (H) program, which helped me to gather practical information, which is necessary for my future life. I would like to express my deep respect to my institute guides Mr. Soumya Banerjee and Dr. Susanta Tewari, Amity Institute of Applied Science Kolkata for giving me their valuable time and all the necessary guidance, which helped me to prepare this report.

Aim:

Prediction of Car Type Preference of income earners in accordance to their Yearly Income and Age.

Organization:

The remaining part of the report has been organized as follows. Section 1 starts with a literature reviews summarizing and comparing the data sets, model structures, and explanatory variables which are been used. Section 2 introduces the description of the data that have been used and furnishes details of the methodology of the modeling framework used in this study. Section 3 gives an overview of the results of vehicle type choice analysis using a multinomial logit model. Finally, conclusion and the scope for further research are summarized in Section 4.

Literature Overview:

Review:

Within the society where automobile plays a pivotal role in daily life, the analysis and modeling of vehicle type choice in households has been much of interests to social scientists and transportation planners.

Data set:

I have a dataset consisting of 155 data points which were obtained by combining primary surveys as well as secondary data collection. The secondary 115 data points were collected from www.cars.com and One Automobiles, Kolkata. And a primary survey was conducted upon 40 respondents through online Google docs forms.

Car Type (Y)	Income per Annum (Rs/lakhs) (x1)	Age (x2)
s	15	41
h	12	28
h	7.2	45
l	50	60
s	18	25
h	6.5	71
suv	20	49
s	3.6	26
h	25	36
h	5	65
h	10	52
suv	12	33
h	6	75
h	12.5	59
l	42	63
h	60	38
suv	55	71
l	47	40
h	3	28
h	12	49
s	15	33
suv	55	71
s	50	54
h	6	46
h	7.5	48
suv	36	51
s	40	40
h	10	39
s	25	66
suv	35	44
s	20	53
suv	40	29
s	20	37

Structure of Model:

For the given data, I had fitted a multinomial logistic regression for predicting the car type preferences.

Let us start with what a multinomial logistic regression is. **Multinomial logistic regression** is a simple extension of binary **logistic regression** that allows for more than two categories of the dependent or outcome variable. Like binary **logistic regression**, **multinomial logistic regression** uses maximum likelihood estimation to evaluate the probability of categorical membership.

In case of 2 classes 1 vs 0 (suppose), we use to develop one logistic model :- $\ln\{p/(1-p)\}=b0+b1X1+b2X2+\ldots\ldots$

Decision rule: If $p\geq 0.5$, then class 1, else otherwise.

Similarly we can extend this for multiclass: For k classes we can develop k-1 models.

Proof: Suppose we have 3 classes A, B & C

Let us take C as the reference class.

Now, let us develop 2 models:-

$\ln\{P(A)/P(C)\}=b0+b1X1+b2X2+\ldots..=Y1$(say)

Again, $\ln\{P(B)/P(C)\}=co+c1X1+c2X2+\ldots..=Y2$(say)

Hence, $P(A)/P(C)=\exp(Y1)$ ----------equation(i)

& $P(B)/P(C)=\exp(Y2)$ ----------equation(ii)

& we know, $P(A)+P(B)+P(C)=1$----equation(iii)

So, by solving equations (i), (ii), (iii) we can get P(A), P(B) & P(C). Therefore, it is sufficient to develop only 2 models for 3 classes.

Similarly, for k classes we can develop k-1 models. (proved)

Therefore in general, for a k class scenario,

$$P(R)=\exp(Yr)/[1+\exp(Y1)+\ldots..+\exp(Yk-1)]$$

The advantage of using this model is, k-1 model equations simultaneously results in smaller standard errors for the parameter estimates than fitting them separately.

Explanatory Variables:

In this analysis, I am taking **age** and **income (per annum)** as the explanatory variables.

Research Methodology:

Research Design:

The purpose of this paper is to examine the relationship between household income, age and type of newly purchased vehicles by analyzing the multinomial logistic model with regard to the recently purchased vehicles as a dependent variable. Because the scope for this study is a national level, the data for this analysis are geographically restricted to national sample only.

Modeling Methodology:

This section describes the fundamental mathematical framework, methodology of estimation and application of the model structures of multinomial logit modeling frameworks that are used for the analysis.

The dependent variable, a vehicle type recently purchased, consists of 4 mutually exclusive categories, so a multinomial logit model which is most widely used for estimating discrete choice models in this field is developed for vehicle type choice.

Random utility maximization:

Discrete choice models are based on random utility maximization theory. The random utility theory which assumes that the decision-maker's preference for an alternative is captured by a utility, an indicator of value to an individual. The utility maximization rule is mentioned that an individual will select the alternative which maximizes his or her utility out of the available alternative set. Probability of choice 'i' is equal to the probability that the utility of alternative 'i' is greater than or equal to the utilities of all other alternatives in the choice set.

(or) $P(i|Cn) = Pr [U_{in} \geq U_{jn}$, all $j \in Cn]$

where, Cn is the set of alternatives available for the nth choice maker (choice set).

The utility maximization rule also implies there is no uncertainty in the individual's decision process, that is, the decision maker is certain to choose the highest ranked alternative under the observed condition. However, the analyst is unlikely to know specific circumstances of the individual's decision or the individual may have incomplete or incorrect information about the attributes of alternatives. To take account of the lack of information on the part of analyst, random utility models are applied by introducing an error term in the utility of each alternative. Thus, the utility of an alternative 'U_i' is split into a deterministic term 'V_i' and a random error term 'ε_i'. Then,

$P(i|Cn) = Pr [V_{in}+\varepsilon_{in} \geq V_{jn}+\varepsilon_{jn}$, all $j \in Cn]$

The deterministic utility V_{in} expressed as linear function of explanatory variables is given by

$V_{t,i} = V(S_t)+V(X_i)+V(S_t, X_i)$

where $V_{t,i}$ is the deterministic utility of alternative *i* for individual *t*,

V(S*t*) is the utility associate with characteristics of individual *t*,
V(X*i*) is the utility associated with attributes of alternative *i,*
V(S*t*, X*i*) is the utility which results from interaction between attributes
of alternative *i* and characteristics of individual *t.*

Assumption that the disturbances are 'Gumbel' distributed leads to the multinomial logit model with the 'Independence of Irrelevant Alternatives' (IIA) property. Multinomial logit model gives the choice probabilities of each alternative as a function of the systematic portion of the utility of all the alternatives. The general expression for the probability of choosing an alternative 'i' (i=1,2,…,J) from a set of J alternatives is: $P(i)$= exp(Vi)/ Σexp(Vj)
Where P(i) is the probability of the decision maker choosing alternative *i* and Vi is the systematic component of the utility of alternative *j.*
After computing the P(i) s, we use the method of maximum likelihood to estimate the parameters of the model.

RESULTS:

Here, there are 4 classes of the dependent variable.
Let us fix the factor (class) hatchback as the reference level.
Then, by running multinomial regression in R we can get 4-1=3 logit models to solve for the probabilities.

Coefficients output (in R):

```
Coefficients:
    (Intercept) Income.per.Annum..Rs.lakhs.         Age
l    -4.7209032                  0.12158415 -0.01738050
s     0.4648983                  0.04300962 -0.02785325
suv  -1.4223254                  0.08140251 -0.02247062
```

Using the output coefficients we can write,
ln[P(l)/P(h)]=-4.721+0.121(Income)-0.017(Age)=Y1(say)
ln[P(s)/P(h)]=0.465+0.043(Income)-0.022(Age)=Y2(say)
ln[P(suv)/P(h)]=-1.422+0.081(Income)-0.022(Age)=Y3(say)

& P(h)+P(l)+P(s)+P(suv)=1

Hence, solving the above simultaneous equations we get,
P(h)=1/[expY1+expY2+expY3+1]
P(l)=Y1/[expY1+expY2+expY3+1]
P(s)=Y2/[expY1+expY2+expY3+1]
P(suv)=Y3/[expY1+expY2+expY3+1]

And, these probabilities can be easily calculated in any software.

Probability table (in R):

	h	l	s	suv
1	0.430837479	0.011657960	0.4172936	0.14021091
2	0.386438157	0.009101364	0.4725268	0.13193366
3	0.559618310	0.005471949	0.3466883	0.08822142
4	0.116153232	0.159235495	0.2985968	0.42601444
5	0.296966331	0.015282641	0.5109945	0.17675650
6	0.724573913	0.004141101	0.2111278	0.06015718
7	0.414460420	0.017923496	0.3983212	0.16929486
8	0.478507448	0.004202075	0.4310483	0.08624218
9	0.275255247	0.027404333	0.4711236	0.22621682
10	0.708095022	0.003742909	0.2286205	0.05954160
11	0.569359813	0.006928647	0.3273854	0.09632618
12	0.418317516	0.009032152	0.4450102	0.12764016
13	0.750191306	0.003763661	0.1913856	0.05465942
14	0.581853329	0.008496610	0.3065543	0.10309572
15	0.201779672	0.099269224	0.3382292	0.36072193
16	0.034912403	0.236637320	0.2546510	0.47379931
17	0.380245938	0.010840590	0.4661001	0.14281339
18	0.549382394	0.005920625	0.3522614	0.09243554
19	0.472003206	0.017456224	0.3530423	0.15749830
20	0.378443946	0.015946690	0.4381293	0.16748005
21	0.302984871	0.041228813	0.4004734	0.25531293
22	0.523395585	0.008557425	0.3565740	0.11147295
23	0.380091650	0.011819075	0.4600327	0.14805654
24	0.102781720	0.213752035	0.2411584	0.44230788
25	0.102490750	0.155949074	0.3113993	0.43016086
26	0.581298254	0.004827663	0.3326099	0.08126417
27	0.575412201	0.005538924	0.3321548	0.08689412
28	0.220382660	0.064398207	0.3986541	0.31656507
29	0.142021441	0.081713610	0.4145240	0.36174093
30	0.483621162	0.007377246	0.3994224	0.10957919
31	0.450431501	0.026623471	0.3342971	0.18864794
32	0.202113425	0.059064426	0.4256089	0.31321325
33	0.644895852	0.005354988	0.2722903	0.07745882
34	0.750191306	0.003763661	0.1913856	0.05465942

So for any individual, suppose for the 6th individual we can see that the probability of buying a hatchback is nearly 73%, so we will predict that 6th person will buy a hatchback. And similarly all the predictions will be done accordingly using the method of maximum likelihood.

Error percentage in prediction:
For the purpose of calculating the error in the prediction we can compute the confusion matrix.

Confusion matrix (in R):

	h	l	s	suv
h	43	0	24	9
l	0	0	0	1
s	13	0	23	16
suv	4	8	8	6

In the matrix, the columns denote the observed values and the rows denote the predicted values.

So the error in prediction can be calculated by,
= [[1-{Trace(Confusion matrix)/Sum(Confusion matrix)}]*100]%
= 53.54%

Predicted & Original percentages:
And, Predicted percentage of people having hatchback cars=49.03% (approximately)
Predicted percentage of people having luxury cars=0.64% (approximately)
Predicted percentage of people having sedans=33.55% (approx)
Predicted percentage of people having SUVs=16.78% (approx)

Whereas originally,
Percentage of people having hatchbacks=38.71%
Percentage of people having luxury cars=5.16%
Percentage of people having sedans=35.48%
Percentage of people having SUVS=20.65%

Hence, we can say that our model is a quite well fit model.

Significance test:
We, want to test the significance of the coefficients of the model.
So, H0:Bi=0 vs H1:Bi ≠0 ; for all i=0(1)2

As, the sample size is large so the appropriate tests here is two-tailed Z test, as for large sample size, t distribution converges to normal (asymptotic distribution)

P-Value Outputs (in R):

	(Intercept)	Income.per.Annum..Rs.lakhs.	Age
l	0.006514257	4.361049e-06	0.52214506
s	0.499829127	5.367459e-03	0.03859688
suv	0.098874905	3.117024e-06	0.16973098

So, At 5% level of significance,

For luxury cars and SUVs, coefficient of age variable is not significant.
But as the variable is significant for atleast one class of dependent variable, we will consider it.

Therefore, we got a quite well fitted model for our purpose which can be used extensively in future.

CONCLUSION:

A disaggregate discrete model (specifically, a multinomial logit model) for the type of recently acquired vehicle to estimate the effect of diverse explanatory variables on the probability of choosing each vehicle type provided insights into the recent trends in the vehicle ownership patterns. The final model is created with hatchback cars as base case.

To summarise the results of the estimation, hatchback cars are most frequently selected for the median group of age and median yearly income group. The next dominant car type with high chance of being selected after hatchbacks are sedans and then SUVs,and with a minimal chance for sports which is predominant intersection of lower age groups and higher income groups.

This model is not only statistically significant but also consistent with expectation and the results of the previous leading studies, however, more developments are still necessary.

Goals Achieved:

- From the fitted multinomial logistic regression model; the probability of a person for buying a particular type of car can be predicted and hence these type of models can be extensively used by automobile companies and others for further profit analysis.
- If the values independent variables are known for an individual, the most suitable car type for him can be predicted, which can be used by car consultancy companies.

Bibliography:

- www.cars.com
- www.un.statistics.in
- www.realstatistics.com
- www.wikipedia.org
- www.youtube.in
- www.mathstackexchange.com

References:

- Logistic Regression – Scott Menard
- The Art Of R Programming – Norman Matloff
- Mohammadian, A., & Miller, E. J. (2003). empirical investigation of household vehicle type choice decisions. Traveler Behavior and Values 2003(1854), 99-106.
- Lave, C. A., & Train, K. (1979). A Disaggregate Model of Auto-Type Choice. Transportation Research Part a-Policy and Practice, 13(1), 1-9.