R Programming Language Written for Estimation of the Integrated Multinomial Logit-Linear Regression Model Based on a Copula Approach: A Technical Article

Monorom RITH [a,b*], Fengqi LIU [c], Pai-Hsien HUNG [d], Kento YOH [e],
Alexis M. FILLONE [f], Jose Bienvenido M. BIONA [g]

[a] Graduate School, *Mechanical Engineering Department, De La Salle University, 1004, Metro Manila, Philippines; Email: rith_monorom@dlsu.edu.ph*
[b] *Research and Innovation Center, Institute of Technology of Cambodia, Russian Conf, Norodom Blvd, Phnom Penh, Cambodia*
[c] *CTI Engineering Co., Ltd., Osaka Prefecture 565-0871, Japan;*
Email: liu-fengqi@ctie.co.jp
[d] Graduate School, *Civil Engineering Department, Osaka University, Suita, Osaka Prefecture 565-0871, Japan; Email: pshung@gmail.com*
[e] Graduate School, *Civil Engineering Department, Osaka University, Suita, Osaka Prefecture 565-0871, Japan; Email: yoh.kento@civil.eng.osaka-u.ac.jp*
[f] *Civil Engineering Department, De La Salle University, Taft Ave, 1004, Metro Manila, Philippines; Email: alexis.fillone@dlsu.edu.ph*
[g] *Center for Engineering and Sustainable Development, De La Salle University, 1004, Metro Manila, Philippines; Email: jose.bienvenido.biona@dlsu.edu.ph*

Abstract: The integrated multinomial logit-linear regression framework based on a copula function can be applied to estimate a joint discrete-continuous choice model. Most of the previous studies have used the GAUSS program that requires users to pay for the license. An alternative tool is the R program, which is open-source. This technical paper attempts to provide some tips to write R programming language to estimate a discrete-continuous choice model. The sample data of household vehicle ownership and energy consumption gathered in 2017 in Metro Manila were employed. The output of this technical paper is expected to contribute to a reduction in research and computation cost. It is very informative for some young researchers and postgraduate students having not enough research budget to support their research works, specifically those students originated in developing countries.

Keywords: R Program, Copula, Multinomial logistic regression, Log-linear regression, Discrete-continuous choice model, Metro Manila

1. INTRODUCTION

The Multinomial Logit (MNL) model proposed by McFadden (1974) is widely applied to develop discrete choice models in various disciplines as the formula for the choice probability takes a closed form and is readily interpretable. However, the MNL model is constrained to a case of a single discrete choice, or a choice maker can choose only one discrete choice from a finite choice set. In the real world, a discrete choice and a continuous choice may be made simultaneously. Some instances include travel mode choice and commuting trip timing (Habib et al., 2009), residential neighborhood choice and household

* Corresponding author

vehicle usage (Bhat and Eluru, 2009), individual vehicle type choice and usage (Nguyen et al., 2017).

Bhat and Eluru (2009) proposed the "Copula" function to couple the binary logistic regression and the log-linear regression by allowing non-linear and asymmetric dependency. After that, Spissu et al. (2009) modified this concept to integrate the MNL and the log-linear regression for vehicle type ownership and usage. Later on, the copula-based discrete-continuous choice model was applied by Nguyen et al. (2017). Some copula types have been applied including Frank, Gaussian, FGM, Clayton, Gumbel, and Joe. Most of the previous studies have used GAUSS program to write the programming language for estimation of the joint discrete-continuous choice model (Bhat and Eluru, 2009; Spissu et al., 2009; Habib et al., 2009; Nguyen et al., 2017). However, that software requires users to pay for the license. Some young research or postgraduate students have a limited research budget and cannot afford the license, especially those students residing in developing countries. An alternative open-source tool is the R program, and this program has been applied for the MNL-linear regression model by Rith et al., (2018a, 2018b). To the best of our knowledge, there is no technical or research article presenting how to write the R programming language for estimation of the MNL-linear regression model.

Consequently, this study attempts to provide some technical tips to write the R programming language to estimate the joint discrete-continuous choice model based on the copula function. The Gaussian copula was applied as an example, and the sample data of household vehicle ownership and energy consumption gathered through various areas in Metro Manila in April through May 2017 were utilized. The output of this technical paper is expected to facilitate the research fields related to discrete-continuous choice model for research students having not enough budget for research work. Furthermore, it is consistent with a reduction in research and computation cost.

The remainder of the paper is structured as the following. Section 2 provides a brief description of the data source and the mathematical framework. Section 3 is related to some technical tips for writing the R programming language. The penultimate section illustrates the estimation results. The last section provides concluding thoughts and the direction for further research.

2. DATA SOURCE AND MATHEMATICAL FRAMEWORK

2.1 Data Source

The data sample of household travel survey was gathered through various areas in Metro Manila in April through May 2017, using a simple random sampling technique because the distribution of household vehicle ownership was known. Some questionnaire forms with incomplete and inappropriate responses were deliberately removed to avoid data inconsistency. After cleaning the data, there were 1,795 households for model development. We classified a number of vehicles owned by a household into three main categories: no vehicle, one vehicle, and two vehicles. The model formulation is illustrated in Figure 1. The zero-vehicle choice has no household energy consumption requirement, and the discrete choice, therefore, has no continuous choice. This is the peculiarity of our study, which is different from the previous literature because all the discrete choices have the corresponding continuous choices (see Bhat and Eluru, 2009; Spissu et al., 2009; Habib et al., 2009; Nguyen et al., 2017; Rith et al., 2018a; Rith et al., 2018b). The explanatory variables are listed in

Table 1. Those factors have remained unexplored mainly in Metro Manila, but we hypothesized that they affect the dependent variables on the matter at hand.

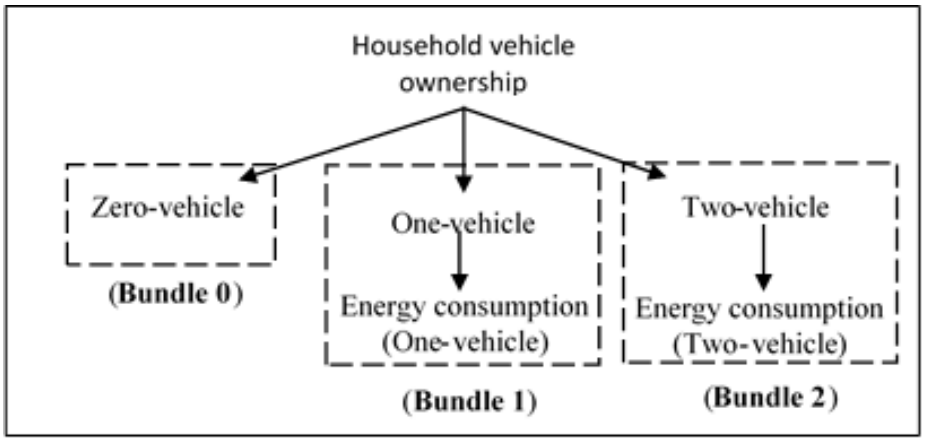

Figure 1. Model formulation

Table 1. Description of explanatory variables

Variable	Description
Socioeconomic characteristics	
Age of household head	1 = aged 40 years and above; 0 = otherwise
Educational level of household head	1 = bachelor and higher; 0 = otherwise
No. of working adults (persons)	Continuous variable
Monthly household income (10^4 Php)	Continuous variable
Built environment attributes	
Population density [a] (10^3 persons/km^2)	Continuous variable
Road density [a] (km/km^2)	Continuous variable
Road public transport line density [a] (km/km^2)	Continuous variable
Distance from home to the shortest railway station (km)	Continuous variable
Mixed facility index [b]	Continuous variable

[a] MUCEP (2015)
[b] see discussion below

Primary schools, secondary schools, colleges, hospitals, markets, and recreation centers (shopping malls) located in the vicinity of residential areas were hypothesized to influence household vehicle ownership and energy consumption decision. The mixed facility index was calculated using Equation 1:

$$Mixed\ facility\ index = \frac{1}{F}\sum_{f=1}^{F} D_f \quad (1)$$

Let f (f = 1, 2…F) be the index representing integral facility types, and D_f [D_f = 1] is the dummy variable of the presence of at least one facility type f located less than 1 km in a walk distance from the residential area. The mixed facility index ranges from zero (least preferably) to one (most preferably).

2.2 Mathematical Framework

For the discrete choice, the implied cumulative distribution of the random error term ε_{nt} of a chosen alternative $F(\varepsilon_{nt})$ can be shown as Equation 2 (Train, 2003):

$$F(\varepsilon_{nt}) = Pr(t) = \frac{exp({\beta_t}' x_{nt})}{\sum_{T=0}^{2} exp({\beta_T}' x_{nT})} \quad (2)$$

where *Pr* stands for the probability. Let *n* (*n* = 1, 2 … *N*) and *T* (*T* = 0, 1, 2) be the indices representing households and vehicle ownership levels, respectively, and $t \in T$. x_{nt} is a column vector of the explanatory variables including a constant, and β_t is a column vector of the corresponding coefficients.

For the continuous choice, the probability density function $f(\eta_{nt})$ and the cumulative distribution function $F(\eta_{nt})$ can be expressed as Equations 3 and 4, respectively (Johnson et al., 1994). The respective ϕ and Φ are the probability density function and the cumulative distribution function of the standard normal distribution.

$$f(\eta_{nt}) = Pr\big(ln(L_{nt}) = ln(l_{nt})\big) = \frac{1}{\sigma_{nt}} \phi\left(\frac{ln(l_{nt}) - \alpha_t' y_{nt}}{\sigma_{nt}}\right) \quad (3)$$

$$F(\eta_{nt}) = Pr\big(ln(L_{nt}) \leq ln(l_{nt})\big) = \Phi\left(\frac{ln(l_{nt}) - \alpha_t' y_{nt}}{\sigma_{nt}}\right) \quad (4)$$

where y_{nt} is a column vector of the explanatory variables including a constant, and α_t is a column vector of the corresponding coefficients.

The Gaussian copula was applied to couple the discrete and continuous choices as a single bundle. Let "0, 1 and 2" be the indices representing the choices of bundle 0, bundle 1, and bundle 2, respectively. Bundle zero has no continuous choice (see Equation 5).

$$Pr_0 = Pr(T = 0) = F(\varepsilon_{n0}) \quad (5)$$

$$Pr_1 = Pr\big(T = 1, ln(L_{n1}) = ln(l_{n1})\big) = \left(\frac{\partial C_\theta\big(F(\varepsilon_{n1}), F(\eta_{n1})\big)}{\partial F(\eta_{n1})}\right)\big(f(\eta_{n1})\big) \quad (6)$$

$$Pr_2 = Pr\big(T = 2, ln(L_{n2}) = ln(l_{n2})\big) = \left(\frac{\partial C_\theta\big(F(\varepsilon_{n2}), F(\eta_{n2})\big)}{\partial F(\eta_{n2})}\right)\big(f(\eta_{n2})\big) \quad (7)$$

The partial derivative of the Gaussian copula function is expressed as Equation 8 (Lee, 1983; Bhat and Eluru, 2009).

$$\frac{\partial C_\theta\big(F(\varepsilon_{nt}), F(\eta_{nt})\big)}{\partial F(\eta_{nt})} = \Phi\left[\frac{\Phi^{-1}\big(F(\varepsilon_{nt})\big) - \theta\Phi^{-1}\big(F(\eta_{nt})\big)}{\sqrt{1-\theta^2}}\right] \quad (8)$$

where θ is the dependency parameter representing the linkage between the two univariate distributions, and the dependency parameter of the Gaussian function ranges from -1 to +1. The joint model was estimated using Equation 9:

$$LL = \sum_{n=1}^{N} \sum_{T=0}^{2} R_{nt}[Pr_T] \quad (9)$$

where *LL* is the log-likelihood function.

3. R PROGRAMMING LANGUAGE

3.1 Package Requirement

R program is an open-source software widely applied for data manipulation, calculation, and graphical display (Venable and Smith, 2018). After installing the R program, users should install RStudio because RStudio can allow users to simultaneously view graphs, tables, R code, and other outputs.

Before writing the code language, the users had better install some packages. Beforehand, the package ***devtools*** is installed, and then we have to call the package using the command ***library(devetools)***. After that, we must install the package ***spcopula*** by writing as ***install_github("BenGraeler/spcopula")***. Finally, we install the package ***maxLik***, developed by Henningsen and Toomet (2011). This package is employed to estimate the model parameters using the maximum likelihood estimation approach. Out of five optimization algorithms, we could select one, and those algorithms are Newton-Raphson, Berndt-Hall-Hall-Hausman, Broyden-Fletcher-Goldfarb-Shanno, Nelder-Mead, and Simulated-Annealing. The ***spcopula*** package was developed by Graeler (2014). This package is used to compute the derivative of copula function, and various copula types are available, e.g., Frank, Gaussian, t, Clayton, Gumble, Joe.

3.2 Written Code Language

After installing the recommended packages, the R programming language was written as the following. Figure 2 illustrates some commands for calling the data sample and the packages. The commands can be explained as follows:

- Line 1: To remove all the objects from the workspace to avoid duplicated objects from the previous works;
- Lines 2-3: To call the data from the folder;
- Line 5: As we mentioned earlier, one of our discrete choices has no continuous choice. It means that natural logarithm of zero does not exist. Therefore, we add 0.01 for households having no vehicle. "carenergy[,4]" refers to the column of energy consumption (see Figure 3). "1000" is used to convert GJ to MJ;
- Line7: To call the packages ***maxLik*** and ***spcopula***.

```
1 rm(list=ls())
2 carenergy<-read.csv("E:\\Journal and Conference\\Energy and Society in Transition\\
3                      Vehicle number and energy.csv", header=TRUE); head(carenergy)
4
5 carenergy[,4]<- 1000*(carenergy[,4]+carenergy[,1]*0.01)
6
7 library(maxLik); library(spcopula)
```

Figure 2. Commands for calling the data sample and packages

The data frame of the data sample is illustrated in Figure 3. There are many columns, and each column stands for each variable. Columns 1 (Veh0) through 4 (HhenegyGJ) are the output variables, column 5 (intercept) is used for the constant coefficient, columns 6 through 29 are the explanatory variables, and the rest is for the additional information. However, there are only nine explanatory variables listed in Table 1, and those variables were used for model estimation.

	Veh0	Veh1	Veh2	HhenergyGJ	Intercept	Age_40	Edu_level	Hhsize	Working_adult	OFW	Preschooler	Child_PS	Child_SS	Ustu	Senior	Homeown	Hh_inc	Pop_den
1	1	0	0	0.000	1	1	0	3	2	0	0	0	0	0	0	0	1.25	7.079
2	1	0	0	0.000	1	1	0	1	1	0	0	0	0	0	0	0	0.25	7.079
3	1	0	0	0.000	1	1	0	4	4	0	0	0	0	0	2	0	5.00	7.079
4	0	1	0	3.119	1	0	1	2	2	0	0	0	0	0	0	0	5.00	7.079
5	0	1	0	2.495	1	1	1	3	1	0	0	0	0	1	0	1	3.50	22.539
6	1	0	0	0.000	1	1	0	2	1	0	0	0	0	0	0	0	0.75	7.079

	Road_den	Line_den	Train_km	Hospital_den	Elem_den	Secon_den	College_den	Market_den	Recreat_den	Mixed_facility	CBD_dis	Hhinc	Family_ID	Barangay
1	12.951	69.27361	0.461	0.318	1.273	3.501	2.865	1.592	0.637	1	0.361	3	1	659
2	12.951	69.27361	0.461	0.318	1.273	3.501	2.865	1.592	0.637	1	0.361	1	2	659
3	12.951	69.27361	0.461	0.318	1.273	3.501	2.865	1.592	0.637	1	0.361	7	3	659
4	12.951	69.27361	0.461	0.318	1.273	3.501	2.865	1.592	0.637	1	0.361	7	5	659
5	19.419	78.81537	0.162	0.955	1.273	0.637	2.228	2.228	0.318	1	2.205	6	6	693
6	12.951	69.27361	0.461	0.318	1.273	3.501	2.865	1.592	0.637	1	0.361	2	7	659

	City	City_barangay	MUCEP_ID	ID
1	Manila	Manila_659	37	1
2	Manila	Manila_659	37	2
3	Manila	Manila_659	37	3
4	Manila	Manila_659	37	4
5	Manila	Manila_693	43	5
6	Manila	Manila_659	37	6

Figure 3. Data frame of the data sample

As stated earlier, there are three discrete choices in the choice set. We used zero-vehicle as the reference category, and all of its parameters estimates, therefore, become zero. Figure 4 demonstrates the R programming language used to estimate the joint MNL-linear regression model based on the Gaussian copula function. The description is made as follows:

- Line 10: "carenergyfunc" is the name of the function created by the authors. "function" is the command of the "maxLik" package. "estpar" is the name of the parameter set created by the authors. The parameter set consists of 104 parameters to be estimated. However, some of them will be removed, and more detail will be shown later on;
- Line 12: The dependency parameter of Bundle 1;
- Lines 13-14: To limit the dependency parameter from -1 to +1.
- Line 19: The standard deviation of energy consumption for Bundle 1;
- Lines 20-21: To limit the standard deviation of energy consumption for Bundle 1;
- Line 26: The parameters of the discrete choice for Bundle 1 to be estimated;
- Line 29: The parameters of the continuous choice for Bundle 1 to be estimated;
- Line 32: The utility function of the discrete choice for bundle 1;
- Line 35: The probability function of the discrete choice for Bundle 0;
- Line 36: The probability function of the discrete choice for Bundle 1;
- Line 39: The function used to calculate the probability density function and the cumulative distribution function for the continuous choice of Bundle 1;
- Line 42: The partial derivative of the copula function with respect to $F(\eta_{nt})$ (see Bhat and Eluru, 2009).
- Line 45: The probability density function of ε_{n1};
- Lines 48-50: The log-likelihood function of the joint model;
- Lines 56-60: The starting values of parameters to be estimated;
- Lines 61-73: The command "activePar" was used to fix or not fix the parameters to be estimated. "FALSE" and "TRUE" mean fix and not fix, respectively. This concept is applied to remove some explanatory variables or insignificant variables.
- Line 75: The command used to show the model estimation results.

```r
carenergyfunc<- function(estpar){

  dependency1<- estpar[1]
  if(dependency1 < -1 | dependency1 > 1)
    return(NA)
  dependency2<- estpar[2]
  if(dependency2 < -1 | dependency2 > 1)
    return(NA)

  sigmaenergy1<- estpar[3]
  if (sigmaenergy1 <= 0)
    return(NA)
  sigmaenergy2<- estpar[4]
  if (sigmaenergy2 <= 0)
    return(NA)

  veh1<-estpar[5:29]
  veh2<- estpar[30:54]

  energy1<- estpar[55:79]
  energy2<- estpar[80:104]

  Uveh1<- (rowSums(t(veh1*t(carenergy[,c(5:29)]))))
  Uveh2<- (rowSums(t(veh2*t(carenergy[,c(5:29)]))))

  Pveh0<- 1/(exp(Uveh2)+exp(Uveh1)+1)
  Pveh1<- exp(Uveh1)/(exp(Uveh2)+exp(Uveh1)+1)
  Pveh2<- exp(Uveh2)/(exp(Uveh2)+exp(Uveh1)+1)

  veh1energy<- ((log(carenergy[,4])-rowSums(t(energy1*t(carenergy[,c(5:29)]))))/sigmaenergy1)
  veh2energy<- ((log(carenergy[,4])-rowSums(t(energy2*t(carenergy[,c(5:29)]))))/sigmaenergy2)

  PDCveh1energy<- ddvCopula(cbind(Pveh1, pnorm(veh1energy)), normalCopula(dependency1))
  PDCveh2energy<- ddvCopula(cbind(Pveh2, pnorm(veh2energy)), normalCopula(dependency2))

  Pveh1energy <- dnorm(veh1energy)
  Pveh2energy <- dnorm(veh2energy)

  sum(log(Pveh0)*carenergy[,1]+
        (log(PDCveh1energy)+log(Pveh1energy)-log(sigmaenergy1))*carenergy[,2]+
        (log(PDCveh2energy)+log(Pveh2energy)-log(sigmaenergy2))*carenergy[,3])

}


mlecarenergy<- maxLik(logLik=carenergyfunc, start=c(dependency1=-0.5, dependency2=-0.5, sigmaenergy1=2, sigmaenergy2=2,
                                                veh1=rep(0,25),
                                                veh2=rep(0,25),
                                                energy1=rep(0,25),
                                                energy2=rep(0,25)),
                    activePar=c(rep(TRUE,4),
                                c(TRUE, TRUE, TRUE, FALSE, TRUE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, TRUE),
                                c(TRUE, TRUE, TRUE, TRUE,      FALSE, FALSE, FALSE, FALSE, FALSE, FALSE,      TRUE, FALSE),

                                c(TRUE, TRUE, TRUE, FALSE, TRUE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, TRUE),
                                c(TRUE, TRUE, TRUE, TRUE,      FALSE, FALSE, FALSE, FALSE, FALSE, FALSE,      TRUE, FALSE),

                                c(TRUE, TRUE, TRUE, FALSE, TRUE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, TRUE),
                                c(TRUE, TRUE, TRUE, TRUE,      FALSE, FALSE, FALSE, FALSE, FALSE, FALSE,      TRUE, FALSE),

                                c(TRUE, TRUE, TRUE, FALSE, TRUE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, TRUE),
                                c(TRUE, TRUE, TRUE, TRUE,      FALSE, FALSE, FALSE, FALSE, FALSE, FALSE,      TRUE, FALSE)
                    ))

summary(mlecarenergy)
```

Figure 4. The written R programming language

4. MODEL ESTIMATION RESULTS AND MODEL CALIBRATION

The model estimation results present in Table 2. The zero-vehicle alternative was used as the reference category. The last row of the table presents the dependency parameters between a number of vehicles owned and energy consumption. Spissu et al. (2009) theoretically explained that a negative dependency parameter implies a positive correlation, while a positive sign means a negative correlation. To the best of our knowledge, the negative and positive dependency parameters are associated with under-estimation and over-estimation of the total continuous output variables, respectively (see Table 3). The estimated coefficients of the discrete and continuous choice components are interpreted below.

Rows 3 through 12 of Table 2 present the estimation results of the discrete choice component. The intercept coefficients are included to capture the average unobserved effect. The results suggest that older households (household heads aged 40 years and above) are more likely to own more vehicles, and a similar finding is found for households with the presence of well-educated household head. Presence of more working adults in family encourages the acquisition of two vehicles, and households with higher income prefer to

acquire more vehicles. The coefficients of population density and road public transport line density imply that households residing in higher population density areas and residential areas with higher public transport line density are less likely to own more vehicles. These findings suggest that urban densification and improvement of road public transport line can discourage household vehicle acquisition. However, an increase in road density is associated with the higher motivation of vehicle ownership among households since on-road parking in residential areas is rampant in Metro Manila. However, improvement of mixed land use has a negative impact on private vehicle ownership.

Table 2. Model estimation results – coefficient (standard error)

Parameters	**One vehicle**	**Two vehicles**
Discrete choice		
Intercept	-3.573 (0.539)***	-10.33 (1.205)***
Age of hosuehold head	0.4768 (0.165)**	1.612 (0.444)***
Education level of household head	1.939 (0.175)***	2.67 (0.529)***
No. of working adults	-0.0116 (0.094)	0.3227 (0.158)*
Monthly household income (10^4 Php)	0.1673 (0.016)***	0.2972 (0.022)***
Population density (10^3persons/km^3)	-0.0262 (0.004)***	-0.0415 (0.008)***
Road density (km/km^3)	0.2949 (0.025)***	0.3364 (0.038)***
Line density (km/km^3)	-0.0126 (0.003)***	-0.0098 (0.005)*
Railway station (km)	0.5702 (0.064)***	0.6159 (0.093)***
Mixed facility index	-1.513 (0.465)**	-0.7086 (0.821)
Continuous choice		
Intercept	7.836 (0.156)***	9.942 (0.529)***
Age of hosuehold head	0.0277 (0.04)	-0.1751 (0.133)
Education level of household head	0.0017 (0.069)	-0.0768 (0.152)
No. of working adults	-0.0009 (0.023)	-0.0554 (0.048)
Monthly household income (10^4 Php)	0.0238 (0.003)***	-0.0047 (0.007)
Population density (10^3persons/km^2)	-0.0005 (0.001)	-0.0013 (0.002)
Road density (km/km^2)	0.0074 (0.005)	-0.0137 (0.009)
Line density (km/km^2)	-0.002 (0)**	-0.0004 (0.001)
Railway station (km)	0.0232 (0.013)	0.0098 (0.019)
Mixed facility index	-0.1314 (0.1)	0.0398 (0.232)
Standard deviation	0.4622 (0.013)***	0.3971 (0.075)***
Dependency parameter	-0.1756 (0.143)	0.8185 (0.125)***

Log-likelihood value at convergence: -1276.87
Zero-vehicle was used as the reference category for the discrete choice model
* significance at 5% level, ** significance at 1% level, *** significance at 0.1% level

Rows 14 through 23 of Table 2 show the estimated coefficients of the energy consumption. The intercept coefficient of one-vehicle alternative was found higher than that of two-vehicle alternative, which implies that households holding more vehicles are likely to consume more energy. All the factors were found to have a statistically insignificant relationship with energy consumption, other than the household income and the road public transport line density. One-vehicle households having higher income are more likely to

consume energy, but one-vehicle households living in higher road public transport line density areas are less likely to consume energy.

The developed model was then used to estimate the output variables. The estimated percentage shares of the discrete choice component and the estimated total energy consumption of the continuous choice component are shown in Table 3. The estimate percentage shares exactly match the actual percentage shares. The total estimated energy consumptions of Bundle 1 and Bundle 2 were underestimated and overestimated, respectively. On a separate note, the estimated and actual total energy consumptions for all the surveyed households were highly comparable (see the second last column). The root mean square error (RMSE) was 1,621.

Table 3. The actual and estimated output variables

	Percentage share (%)			**Total energy consumption (MJ/month)**			**RMSE**
	Zero vehicles	One vehicle	Two vehicles	One vehicle	Two vehicles	All households	
Actual	56.38	38.66	4.96	2,465,150	679,436	3,144,586	–
Estimated	56.38	38.66	4.96	2,110,949	1,029,483	3,140,432	1,621

5. CONCLUSIONS AND RECOMMENDATIONS

The thrust of this technical paper intends to provide some technical tips to write the R programming language for estimation of the MNL-linear regression model. The data sample of household vehicle ownership and energy consumption in Metro Manila in 2017 were employed. Evident from the model estimation results and model calibration, the written R programming language can be applied to estimate the discrete-continuous choice model. This technical work contributes to a reduction in computation and research cost, which is essential for young research students having not enough research budget to support their research work related to discrete-continuous choice modeling.

Future work should focus on how to develop a package as the final product for the discrete-continuous choice model. Therefore, it is more convenient for practitioners to apply it.

ACKNOWLEDGMENTS

The outcomes of this research paper is funded by (1) Japan International Cooperation Agency (JICA) under AUN/SEED-Net project for a Ph.D. Sandwich program at De La Salle University, Philippines and Osaka University, Japan; (2) Newton Project of the University of Oxford, London; and (3) Big Data Analytics and Applications (MTNN:BDAA) project funded by the Department of Science and Technology (DOST), Philippines.

REFERENCES

Bhat, C. R., & Eluru, N. (2009). A copula-based approach to accommodate residential self-selection effect in travel behaviour modeling. *Transportation Research Part B, 43*, 749–765.

Graeler, B. (2014). Modeling skewed spatial random fields through the spatial Vine Copula. *Spatial Statistics*, *10*, 87–102.

Habib, K. N., Day, N., & Miller, E. J. (2009). An investigation of commuting trip timing and mode choice in the Greater Toronto Area: Application of a joint discrete-continuous model. *Transportation Research Part A*, *43*(2009), 639–653.

Henningsen, A., & Toomet, O. (2011). A package for maximum likelihood estimation in R. *Computational Statistics*, *26*(3), 443–458.

Johnson, N. L., Kotz, S., & Balakrisnan, N. (1994). *Continuous univariate distributions.* New York: John Wiley & Sons Inc.

McFadden, D. (1974). *Conditional logit analysis of qualitative choice behaviour*. (P. Zarembka, Ed.) New York: Academic Press.

Nguyen, N. T., Miwa, T., & Morikawa, T. (2017). Vehicle type choice, usage, and CO2 emissions in Ho Chi Minh city: Analysis and simulation using a discrete-continuous model. *Asian Transport Studies*, *4*, 499–517.

Rith, M., Biona, J. B., Fillone, A., Doi, K., & Inoi, H. (2018a). Joint model of private passenger vehicle type ownership and fuel consumption in Metro Manila: Analysis and application of discrete-continuous model. *Philippine Transportation Journal*, *1*(2), 32–47.

Rith, M., Biona, J. B., & Fillone, A. (2018b). Impact of dependency parameters of each discrete-continuous choice on model estimation results using frank copula-based discrete-continuous model. *Proceeding of 11th ATRANS Annual Conference: Young Researcher's Forum 2018 "Transportation for a better life: Lessons learned from global experience to local best practice"*, (pp. 7–16). Bangkok, Thailand, August 24.

Spissu, E., Pinjari, A. R., Pendyala, R. M., & Bhat, C. R. (2009). A copula-based joint multinomial discrete–continuous model of vehicle type choice and miles of travel. *Transportation*, *36*(2009), 403–422.

Train, K. (2003). *Discrete choice methods with simulation.* New York, NY: Cambridge University Press.

Venables, W. N., & Smith, D. M. (2018). *An introduction to R.* Retrieved from CRAN-R Project: https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf