Choosing among modes of transport: a case study

Luiz R. Nakamura [1], **Thiago G. Ramires** [2], **Ana J. Righetto** [3], **Elisa Henning** [4], **Andréa C. Konrath** [1]

1 Introduction

Urban mobility is an increasingly problem in small and bigger centres, especially with the growing number of fleet vehicles. Authorities are working hard to solve this huge public problem by attaining the integration of different modes of transport. One of the possible discussed solutions is to reduce the number of vehicles by encouraging mobility through public and/or human-powered modes of transport.

Hence, the main aim of this survey is to find out which are the factors that lead a user to choose a certain mode of transport, especifically with people going to the university. In order to do this task, a multinomial logistic regression model will be considered in this study.

2 Material and methods

2.1 Data set

The response variable regards which mode of transport people use to go to the university, especifically two located in Joinville, Santa Catarina State, Brazil, namely Universidade do Estado de Santa Catarina and Universidade da Região de Joinville. This variable has three different levels: i) human-powered (on foot or by bike); ii) small-vehicle (by car as the driver or as a passenger; or by motorcycle); and iii) large-vehicle (by bus or van). This data was introduced by Shubert et al. (2018), however they categorised the response variable into only two levels: by bike or not by bike.

The explanatory variables (and their respectively levels) available in the dataset that may be considered to model the response are listed in Table 1.

Table 1: Explanatory variables and their respective levels

Variable	Levels
Position/Status	P_1: undegraduate student; P_2: postgraduate student; P_3: alumni; P_4 : lecturer/professor
Sex	male; female
Age	A_1: up to 17 years old; A_2 : from 17 up to 25 years old; A_3 : from 25 up to 35 years old; A_4 : 35 years old or older
Year	up to 2000; 2010; 2011; 2012; 2013; 2014; 2015; 2016; 2017
Average journey time	T_1 : up to 15 minutes; T_2 : from 15 minutes up to 30 minutes T_3 : from 30 minutes up to 45 minutes; T_4: 45 minutes or more

[1]Universidade Federal de Santa Catarina. e-mail: {*luiz.nakamura, andrea.ck*}*@ufsc.br*
[2]Universidade Tecnológica Federal do Paraná. e-mail: *thiagogentil@gmail.com*
[3]Instituto Agronômico do Paraná. e-mail: *ajrighetto@gmail.com*
[4]Universidade do Estado de Santa Catarina. e-mail: *elisa.henning@udesc.br*

2.2 Multinomial logistic regression model

The multinomial logistic regression model can be used as a classification tool which models the odds of the response (nominal outcomes) as a function of a set of explanatory variables, generalising the logistic regression (Hosmer and Lemeshow, 2000).

Mathematically, let us consider K possible outcome levels, in which one level is chosen as the referent or baseline level and the other $K-1$ outcomes are separately regressed against the referent outcome. As in Righetto et al. (2019), if $K=3$, i.e., the response presents three different levels as stated in Section 2.1, and letting the last outcome (i.e. large-vehicle level) to be the referent outcome, we shall write

$$\frac{P(Y=1)}{P(Y=3)}=\mu \quad \text{and} \quad \frac{P(Y=2)}{P(Y=3)}=\sigma,$$

where $P(Y=m)$, $m=1,2,3$, represents the probability of $Y=m$ and μ and σ are unknown parameters. Note that μ and σ represent the odds ratio between the levels $Y=1$ versus $Y=3$ and $Y=2$ versus $Y=3$, respectively.

If we solve for the probabilities and use the fact that all K probabilities must sum to one, we have

$$\begin{aligned}
P(Y=1)=P(Y=\text{Human-powered})&=\frac{\mu}{1+\mu+\sigma}=\frac{\exp(\boldsymbol{X}_1\boldsymbol{\beta}_1)}{1+\exp(\boldsymbol{X}_1\boldsymbol{\beta}_1)+\exp(\boldsymbol{X}_2\boldsymbol{\beta}_2)}\\
P(Y=2)=P(Y=\text{Small-vehicle})&=\frac{\sigma}{1+\mu+\sigma}=\frac{\exp(\boldsymbol{X}_2\boldsymbol{\beta}_2)}{1+\exp(\boldsymbol{X}_1\boldsymbol{\beta}_1)+\exp(\boldsymbol{X}_2\boldsymbol{\beta}_2)}\\
P(Y=3)=P(Y=\text{Large-vehicle})&=\frac{1}{1+\mu+\sigma}=\frac{1}{1+\exp(\boldsymbol{X}_1\boldsymbol{\beta}_1)+\exp(\boldsymbol{X}_2\boldsymbol{\beta}_2)}. \quad (1)
\end{aligned}$$

In order to select different regression structures for the above mentioned parameters we used a stepwise-based selection method described in Nakamura et al. (2017). Regarding the estimation and diagnostic processes of the model, we adopted here the `gamlss` package (Stasinopoulos and Rigby, 2007) available in `R` software (R Core Team, 2018). For further details, please check Stasinopoulos et al. (2017).

3 Results and discussion

The regression structures obtained by the stepwise-based selection method for both parameters μ and σ only considered the covariates sex and time. We shall highlight here that the interactions between variables were considered in this process, but were not selected in the final model. The maximum likelihood estimates (MLEs) and their corresponding standard errors (SEs) of the final fitted model are given in Table 2.

As expected, the closer a person lives to the university, the higher will be its probability of choosing a human-powered or a small-vehicle instead of a large-vehicle mode of transport. Furthermore, males have a higher probability of choosing both human-powered or small-vehicle instead of a large-vehicle to go the university when compared to females. These statements are supported by the negative coefficients in the average journey time levels and the positive coefficient for male presented in Table 2.

We can easily obtain odds ratio through the fitted model, e.g, $\exp(-2.21)=0.11$, i.e. if a female takes from 15 up to 30 minutes to arrive to her destination, the chance she goes to the university on foot or by bike is 89% smaller than going by a large-vehicle. Moreover,

Table 2: The MLEs and corresponding (SEs) of the estimates from the fitted model

$\log(\mu)$			$\log(\sigma)$		
Variable	Level	Estimate	Variable	Level	Estimate
Intercept		2.12 (0.40)	Intercept		1.80 (0.40)
Sex	Female	0.00 (–)	Sex	Female	0.00 (–)
	Male	1.27 (0.38)		Male	0.69 (0.36)
Time	T_1	0.00 (–)	Time	T_1	0.00 (–)
	T_2	-2.21 (0.47)		T_2	-1.21 (0.46)
	T_3	-4.75 (0.74)		T_3	-3.02 (0.57)
	T_4	-6.41 (0.82)		T_4	-4.57 (0.57)

the chance she goes to the university by a small-vehicle (by car or by motorcycle) is 70% smaller than by a large-vehicle ($\exp(-1.21) = 0.30$).

In order to calculate the associated probabilities of a given person to choose between any of the three modes of transport we can use Equation (1). For instance, let us consider a male that lives from 15 up to 30 minutes far away from the university. Hence,

$$\begin{aligned} P(Y=1) &= P(Y = \text{Human-powered}) = 0.197 \\ P(Y=2) &= P(Y = \text{Small-vehicle}) = 0.299 \\ P(Y=3) &= P(Y = \text{Large-vehicle}) = 0.504. \end{aligned}$$

It is noteworthy that any set of characteristics can be easily used to calculate these probabilities. Apart from the baseline level case (females that take up to 15 minutes to the university) where these probabilities are equal, i.e. $P(Y=1) = P(Y=2) = P(Y=3)$, all other levels present a higher probability to large-vehicle mode of tranport.

Figure 1 displays the worm plot (van Buuren and Fredriks, 2001) obtained from the final fitted multinomial logistic regression model. Worm plots are detrended normal Q-Q residual plots based on the normalised randomised quantile residuals (Dunn and Smyth, 1996). If the model for the response variable is correct, then these residuals must follow a standard normal distribution, i.e., within the worm plot if the model is correct, then 95% of the (black) points must lie between the elliptical 95% pointwise interval band curves and no linear, quadratic or cubic shape may be observed. As we can see, no particular trend or shape can be seen in Figure 1, hence based on the residual diagnostics we can say that the multinomial logistic regression provides a reasonable fit to the data set in the study.

References

HOSMER, D. W.; LEMESHOW, S. *Applied Logistic Regression.* 2ed. New York: Wiley-Interscience Publication. 2000. 375 p.

NAKAMURA, L. R.; RIGBY, R. A.; STASINOPOULOS, D. M.; LEANDRO, R. A.; VILLEGAS, C.; PESCIM, R. R. Modelling location, scale and shape parameters of the Birnbaum-Saunders generalized t-distribution. *Journal of Data Science*, v. 15, p.221–238, 2017.

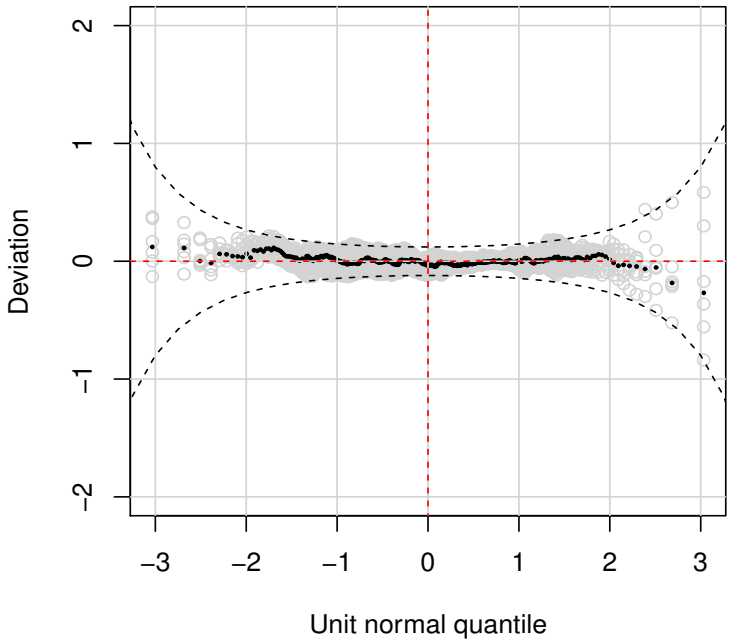

Figure 1: *Worm plot for the fitted model*

R CORE TEAM. *R*: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2018. URL `http://www.R-project.org/`.

RIGHETTO, A. J.; RAMIRES, T. G.; NAKAMURA, L. R.; CASTANHO, P. L. D. B.; FAES, C.; SAVIAN, T. V. Predicting weed invasion in a sugarcane cultivar using multispectral image. *Journal of Applied Statistics*, v. 46, p.1–12, 2019.

SHUBERT, T. F.; MACIEL, A. C.; HENNING, E. Elaboration of a model for representing the transport mode by university students. In: 1st Latin American SDEWES Conference on Sustainable Development of Energy, Water and Environment Systems, 2018, Rio de Janeiro. *Proceedings...*, Rio de Janeiro, RJ, 2018. p. 1–6.

STASINOPOULOS, D. M.; RIGBY, R. A. Generalized additive models for location, scale and shape (GAMLSS) ir R. *Journal of Statistical Software*, v. 23, p. 1–10, 2007.

STASINOPOULOS, D. M.; RIGBY, R. A.; HELLER, G. Z.; VOUDOURIS, V.; DE BASTIANI, F. *Flexible Regression and Smoothing: Using GAMLSS in R*. London: Chapman and Hall/CRC. 2017. 549 p.