Robustness of Student link function in multinomial choice models

Jean Peyhardi

Version of Record: https://www.sciencedirect.com/science/article/pii/S1755534520300270
Manuscript_feb27b6d5041d65397307645e97ad158

Robustness of Student link function in multinomial choice models

Dr Jean Peyhardi[a,*,1] (Assistant Professor)

[a]*Institut Montpellierain Alexander Grothendieck, 34000 Montpellier*

ARTICLE INFO

Keywords:
Discrete choice model
Generalized linear model
Link function
Influence function

ABSTRACT

The Student distribution has already been used to obtain robust maximum likelihood estimator (MLE) in the framework of binary choice models. But, until recently, only the logit and probit binary models were extended to the case of multinomial choices, resulting in the multinomial logit (MNL) and the multinomial probit (MNP). The recently introduced family of reference models, well defines a multivariate extension of any binary choice model, i.e. for any link function. In particular, this is the first extension of the binary robit to the case of multinomial choices. These models define the choice probability for category j relative to an (interchangeable) reference category. This paper highlights the robustness of reference models with Student link function, by showing that the influence function is bounded. Inference of the MLE is detailed through the Fisher's scoring algorithm, which is appropriated since reference models belong to the family of generalized linear models (GLMs). These models are compared to the MNL on the benchmark dataset of travel mode choice between Sydney and Melbourne. The results obtained on this dataset with reference models are completely different compared with those usually obtained with MNL, nested logit (NL) or MNP that failed to select relevant attributes. It will be shown that the travel mode choice is totally deterministic according to the terminal waiting time. In fact, the use of Student link function allow us to detect the total artificial aspect of this famous dataset.

1. Introduction

The use of Student distribution to obtain robust estimations has been introduced by Lange, Little, and Taylor [1989] in the framework of linear regression. For a binary response variable, the use of Student distribution was suggested by Albert and Chib [1993] as an alternative to logit or probit regression. Liu [2004] called this model the robit regression model and demonstrated the robustness of the MLE. Koenker and Yoon [2009] studied the importance of the link function in binary choice models with a focus on Student link function. But the variety of link functions defined in the literature, is considerably decreasing when the number of alternative choices J is more than two. Only the logit and probit binary models were extended to the case of multinomial choices, resulting in the MNL and the MNP. Contrarily to the MNL, the probabilities of alternative choices have no analytic forms with the MNP. Their computation has to be made through approximations and this complexity often leads the practitioner to neglect the MNP when $J > 3$. Despite these difficulties, the extension of the MNP with multivariate Student error distribution in the context of random utility maximisation (RUM), has recently been introduced by Dubey, Bansal, Daziano, and Guerra [2020]. But this multinomial choice model ($J \geq 2$) is not the natural extension of the binary robit model since the difference of two Student independent errors does not follow a Student distribution. This is the case only with MNL and MNP since the difference of two independent Gumbel errors follows a logistic distribution and the difference of two Gaussian errors follows a Gaussian distribution.

An alternative way to extend link functions when $J > 2$, has been recently introduced by Peyhardi, Trottier, and Guédon [2015] in the context of GLMs. The family of reference models was thereby introduced, for which all alternatives are compared to a reference alternative. Since each of these $J-1$ comparisons is binary, the link function through the linear predictor can be made with a cumulative distribution function (cdf). All usual econometric outputs (willingness-to-pay, elasticities, ...) of reference models have been determined by Bouscasse, Joly, and Peyhardi [2019]. The goal of the present paper is to demonstrate the robustness of reference models defined with Student link function. A usual way to study the robustness is to study the influence function [Hampel, 1974]. This approach is well established in the GLM framework for the class of M estimators [Künsch, Stefanski, and Carroll, 1989]. We propose to study the influence function for a reference model based on the score $\psi = \partial l / \partial \beta$. We extend the results on robustness

thumbnails/cas-email.jpeg
thumbnails/cas-url.jpeg
jean.peyhardi@umontpellier.fr (J. Peyhardi)
https://www.researchgate.net/profile/Jean_Peyhardi (J. Peyhardi)
ORCID(s): 0000-0001-7511-2910 (J. Peyhardi)

of Student link function, shown by Liu [2004], to the case of $J > 2$ alternatives. More precisely the influence function for a reference model is shown to be unbounded for the logistic and normal link functions and bounded for the Student link function.

Student and logistic link function are then compared on the well known dataset of travel mode choice between Sydney and Melbourne. This is certainly the most used benchmark dataset to compare different families of discrete choice models [Greene, 2003, Hensher and Greene, 2002]. The MNL, the NL and the MNP models have been extensively studied among this dataset, principally to highlight the limitation of the independence of irrelevant alternatives (IIA) property. This paper presents the IIA property for reference models and relates it to invariance property under permutations of alternatives. It will be shown that, even if the NL or the MNP model do not share this property, they fail, as the MNL, to select relevant attributes. The model selection is clearly in favour of the Student link function against the MNL, NL and MNP. It reveals that the travel mode choice is driven only by the terminal time. Plotting the observed choices according to the terminal time, revealed that all individuals made the same choice for each given terminal time value. Moreover, using a three dimensional plot, it is noticed that the design experiment takes a geometric form of a cross. In fact, the use of Student link function allow us to highlight the total artificial aspect of this famous dataset and thus reconsider classical results obtained in the usual literature [Greene, 2003, Hensher and Greene, 2002].

The present paper is organized as follows. The second section presents the family of reference models as an extension of the MNL. The third section describes the invariance property of such models under permutations of the alternatives and relates it to the IIA property. It is highlighted that a reference model is depending on the reference alternative. Otherwise, the reference model is also depending on the degree of freedom parameter of the Student distribution. In Section 4, the Fisher's scoring algorithm is therefore firstly detailed for a given reference alternative and a given degree of freedom. Then the inference procedure is described when these two parameters are unknown. The influence function of a reference model is described according to the chosen link function, i.e. the chosen cdf. It is thus easily seen that the influence function is bounded with Student cdf and unbounded with logistic or normal cdf. Section 5 briefly presents the dataset. The model selection is then detailed, leading to a simple model with only the terminal time as attribute. This section then backs to the dataset and highlights its artificial aspect, using a well chosen plot of the dataset. It makes this dataset the perfect candidate to study the sensitivity of a model to different kind of noise. Indeed the general cost, for instance, can be considered as a noise in attribute (i.e., the corresponding parameter is null) and some observations can be considered as outliers (i.e., prediction and observation are the opposite). Section 6.1 presents a simulation that firstly studies the inference accuracy of the degree of freedom estimator and then the sensitivity of the inference to the noise in attributes.

2. Reference models

We first recall the notations used all along the paper. The individual subscript will be omitted for convenience, without loss of generality. Let Y denote the response variable corresponding to the choice with J the number of alternatives. Let $\boldsymbol{x} \in \mathbb{R}^p$ denote the vector of p individual attributes and $\boldsymbol{\omega} = \{\boldsymbol{\omega}_j\}_{j=1,\dots,J} \in \mathbb{R}^{qJ}$ the vectors of q alternative specific attributes. Let $\pi_j = P(Y = j)$ denote the probability of choosing the alternative j given the attributes $\boldsymbol{x}$ and $\boldsymbol{\omega}$.

2.1. Multinomial logit models

Reference models can be presented as an extension of the MNL. The MNL is classically presented by the J equations

$$P(Y = j) = \frac{\exp(\eta_j^*)}{\sum_{k=1}^{J} \exp(\eta_k^*)}, \quad j = 1, \dots, J, \tag{1}$$

where η_j^* is the predictor associated with alternative j generally assumed to be linear in attributes. Different logit models are obtained according to the form of the linear predictors. They can be defined using individual attributes $\boldsymbol{x}$ and/or alternative specific attributes $\boldsymbol{\omega}$. In the following we will use the more general parametrization (containing both individual and alternative specific attributes), i.e.,

$$\eta_j^* = \alpha_j^* + \boldsymbol{\omega}_j^t \boldsymbol{\gamma}^* + \boldsymbol{x}^t \boldsymbol{\delta}_j^*$$

for $j = 1, \dots, J$. But these predictors η_j^* are not identifiable and thus the intercept and slope parameters no more. By convention, the last alternative J is considered as the reference alternative. The numerator and denominator of the

fraction in equations (1) are divided by $\exp(\eta_J^*)$. Thanks to the exponentiation identity $\exp(a+b) = \exp(a)\exp(b)$, the J equations (1) are equivalent to the $J-1$ equations

$$P(Y=j) = \frac{\exp(\eta_j)}{1+\sum_{k=1}^{J-1}\exp(\eta_k)}, \quad j = 1, \dots, J-1, \tag{2}$$

where $\eta_j = \eta_j^* - \eta_J^*$ for all $j = 1, \dots, J-1$, and consequently the probability of the reference alternative is given by

$$P(Y=J) = \frac{1}{1+\sum_{k=1}^{J-1}\exp(\eta_k)}.$$

This entails a translation of with respect to the reference alternative:

- alternative constants: $\alpha_j = \alpha_j^* - \alpha_J^*$,
- individual attributes parameter: $\boldsymbol{\delta}_j = \boldsymbol{\delta}_j^* - \boldsymbol{\delta}_J^*$,
- alternative specific attributes: $\boldsymbol{\omega}_j - \boldsymbol{\omega}_J$,

Remark that the alternative specific parameter $\boldsymbol{\gamma} = \boldsymbol{\gamma}^*$ stays unchanged. The translated predictors are given by

$$\eta_j = \alpha_j + (\boldsymbol{\omega}_j - \boldsymbol{\omega}_J)^t\boldsymbol{\gamma} + \boldsymbol{x}^t\boldsymbol{\delta}_j, \quad j = 1, \dots, J-1,$$

where $\alpha_1, \dots, \alpha_{J-1}$, $\boldsymbol{\delta}_1, \dots, \boldsymbol{\delta}_{J-1}$ and $\boldsymbol{\gamma}$ are identifiable parameters.

2.2. Extension of logit models

The equations (2) can be rewritten as

$$\frac{\pi_j}{\pi_j + \pi_J} = \frac{\exp(\eta_j)}{1+\exp(\eta_j)}, \quad j = 1, \dots, J-1. \tag{3}$$

The right part of these equations corresponds to the logistic cdf $F(x) = e^x/(1+e^x)$. It is proposed to use another cdf. In this paper, focus is made on the use of Student cdf denoted by F_ν, where ν is the degree of freedom. The model is described by the $J-1$ equations

$$\frac{\pi_j}{\pi_j + \pi_J} = F_\nu(\eta_j), \quad j = 1, \dots, J-1, \tag{4}$$

with

$$F_\nu(x) = \frac{1}{2} + \frac{x\Gamma\left(\frac{\nu+1}{2}\right)\,{}_2F_1\left(\frac{1}{2}, \frac{\nu+1}{2}; \frac{3}{2}; -\frac{x^2}{\nu}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)},$$

where ${}_2F_1$ denotes the hypergeometric function.

Otherwise, the vector of linear predictors $\boldsymbol{\eta}^t = (\eta_1, \dots, \eta_{J-1})$ can be written as the product of the design matrix Z and the vector of parameters $\boldsymbol{\beta}^t = (\alpha_1, \dots, \alpha_{J-1}, \boldsymbol{\gamma}^t, \boldsymbol{\delta}_1^t, \dots, \boldsymbol{\delta}_{J-1}^t)$ where

$$Z = \begin{pmatrix} 1 & & & (\boldsymbol{\omega}_1 - \boldsymbol{\omega}_J)^t & \boldsymbol{x}^t & & \\ & \ddots & & \vdots & & \ddots & \\ & & 1 & (\boldsymbol{\omega}_{J-1} - \boldsymbol{\omega}_J)^t & & & \boldsymbol{x}^t \end{pmatrix}.$$

The parametrization of the linear predictors $\eta_1, \dots, \eta_{J-1}$ is characterized by the design matrix Z. A reference model is therefore specified by the choice of the cdf F and the design matrix Z. More generally, Peyhardi et al. [2015] introduced an unifying specification of GLMs for categorical responses, including reference models as a special case. This specification is based on three components (r, F, Z) that characterize the equations:

$$r_j(\boldsymbol{\pi}) = F_\nu(\eta_j), \quad j = 1, \dots, J-1,$$

where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_{J-1})$. Three other ratios (adjacent, cumulative and sequential) can also be used but only if a total ordering assumption is made among the J alternatives. In the following we will denote a model by its corresponding (r, F, Z) triplet. For instance the MNL is fully characterized by the triplet (reference, logistic, Z) and can thus be viewed as a special case of reference models. It is considered as the canonical reference model since it corresponds to the canonical link function in the framework of GLMs. All along this paper we will only use the reference ratio $r_j(\boldsymbol{\pi}) = \pi_j/(\pi_j + \pi_J)$ since it is appropriate for non-ordered alternatives. It can be ensure by study the invariance properties under permutations of the alternatives.

Otherwise, an interesting property of the family of Student distributions is that Cauchy, logistic, and normal distributions can be viewed as special cases. First, the Cauchy distribution is exactly the Student distribution with $\nu = 1$. Albert and Chib [1993] note that "logistic quantiles are approximately a linear function of $\mathcal{T}(8)$ quantiles." In application we will see that the (reference, F_8, Z) model and the MNL give similar results. Finally, it is know that F_ν converges towards the normal cdf when ν tends to infinity. In application we will see that the (reference, F_{20}, Z) model and the (reference, normal, Z) model give similar results. It should be remarked, to avoid confusion, that the (reference, normal, Z) model and the MNP are different models. They are equivalent only in the binary case $J = 2$.

3. Invariance property under permutations of alternatives

3.1. Transposition of the reference alternative

It can be shown that the MNL, or equivalently the (reference, logistic, Z) model is invariant under all permutations of alternatives. It means that a permutation of alternatives only implies a (linear) transformation of the parameter vector $\boldsymbol{\beta}$ such that the fitted probabilities stay unchanged. In particular, the MNL is invariant under any transposition of the reference alternative. It means that the choice of the reference alternative has no impact on the fitted probabilities. This is quite different for (reference, F_ν, Z) models. They are invariant under all permutation except under transposition of the reference alternative [Peyhardi et al., 2015]. We will therefore denote by (reference, F_ν, $Z)_{j_0}$ the model defined with j_0 as reference alternative, i.e., such that

$$\frac{\pi_j}{\pi_j + \pi_{j_0}} = F_\nu(\eta_j), \quad \forall j \neq j_0.$$

Otherwise, the $J-1$ non-reference alternatives $j \neq j_0$ can be permuted without modifying the model. We will see in the application that this invariance property of reference models will be crucial. The reference alternative has thus to be selected as a part of the link function.

3.2. Independence of irrelevant alternatives

Let us remark that the total invariance property of the MNL is related to the IIA property. For non-canonical reference models the IIA property partially holds, i.e., we have

$$\frac{\pi_j}{\pi_{j_0}} = \frac{F_\nu}{1 - F_\nu}\left\{\alpha_j + (\boldsymbol{\omega}_j - \boldsymbol{\omega}_{j_0})^t\boldsymbol{\gamma} + \boldsymbol{x}^t\boldsymbol{\delta}_j\right\}, \quad \forall j \neq j_0. \tag{5}$$

These $J-1$ ratio of probabilities are independent of other alternatives. On the contrary for two non-reference alternatives $j \neq j_0$ and $k \neq j_0$ we have

$$\frac{\pi_j}{\pi_k} = \frac{\pi_j/\pi_{j_0}}{\pi_k/\pi_{j_0}},$$
$$\frac{\pi_j}{\pi_k} = \frac{\frac{F_\nu}{1-F_\nu}\left\{\alpha_j + (\boldsymbol{\omega}_j - \boldsymbol{\omega}_{j_0})^t\boldsymbol{\gamma} + \boldsymbol{x}^t\boldsymbol{\delta}_j\right\}}{\frac{F_\nu}{1-F_\nu}\left\{\alpha_k + (\boldsymbol{\omega}_k - \boldsymbol{\omega}_{j_0})^t\boldsymbol{\gamma} + \boldsymbol{x}^t\boldsymbol{\delta}_k\right\}}.$$

Therefore these $\binom{J-1}{2}$ ratios of probabilities are depending on the reference alternative specific attributes $\boldsymbol{\omega}_{j_0}$. In the case of canonical link (i.e., logistic cdf), the function $F/(1-F)$ turns out to be the exponential function and thus the dependence disappears:

$$\frac{\pi_j}{\pi_k} = \frac{\exp\left\{\alpha_j + (\boldsymbol{\omega}_j - \boldsymbol{\omega}_{j_0})^t\boldsymbol{\gamma} + \boldsymbol{x}^t\boldsymbol{\delta}_j\right\}}{\exp\left\{\alpha_k + (\boldsymbol{\omega}_k - \boldsymbol{\omega}_{j_0})^t\boldsymbol{\gamma} + \boldsymbol{x}^t\boldsymbol{\delta}_k\right\}},$$

$$\frac{\pi_j}{\pi_k} = \exp\left\{\alpha_j + (\boldsymbol{\omega}_j - \boldsymbol{\omega}_k)^t\boldsymbol{\gamma} + \boldsymbol{x}^t(\boldsymbol{\delta}_j - \boldsymbol{\delta}_k)\right\}.$$

4. Inference procedure

The inference procedure is first described for a (reference, F_ν, $Z)_{j_0}$ model with a fixed reference alternative j_0 and a fixed degree of freedom ν.

4.1. Fixed alternative reference j_0 and degree of freedom ν

4.1.1. Fisher's scoring algorithm

Reference models belong to the set of GLMs and thus the Fisher's scoring algorithm is easily computed. For maximum likelihood estimation, the $t+1^{\text{th}}$ iteration of Fisher's scoring algorithm is given by

$$\boldsymbol{\beta}^{[t+1]} = \boldsymbol{\beta}^{[t]} - \left\{ \mathrm{E}\left(\frac{\partial^2 l}{\partial\boldsymbol{\beta}^t\partial\boldsymbol{\beta}}\right)_{\boldsymbol{\beta}=\boldsymbol{\beta}^{[t]}} \right\}^{-1} \left(\frac{\partial l}{\partial\boldsymbol{\beta}}\right)_{\boldsymbol{\beta}=\boldsymbol{\beta}^{[t]}},$$

where l denotes the log-likelihood among the dataset $(\boldsymbol{y}_i, \boldsymbol{\omega}_i, \boldsymbol{x}_i)_{i=1,\dots,n}$. Using the chain rule and properties of the exponential family of distributions, the score becomes the following product of matrices

$$\begin{aligned}
\frac{\partial l}{\partial\boldsymbol{\beta}} &= \sum_{i=1}^{n} \frac{\partial l_i}{\partial\boldsymbol{\beta}} \\
&= \sum_{i=1}^{n} \frac{\partial\boldsymbol{\eta}_i}{\partial\boldsymbol{\beta}} \ \frac{\partial\boldsymbol{\theta}_i}{\partial\boldsymbol{\eta}_i} \ \frac{\partial l_i}{\partial\boldsymbol{\theta}_i} \\
\frac{\partial l}{\partial\boldsymbol{\beta}} &= \sum_{i=1}^{n} Z_i^t \ \frac{\partial\boldsymbol{\theta}_i}{\partial\boldsymbol{\eta}_i} \ (\boldsymbol{y}_i - \boldsymbol{\pi}_i)
\end{aligned}$$

where $\boldsymbol{\theta}_i := (\theta_{i,j})_{j\neq j_0}$, $\boldsymbol{\eta}_i := (\eta_{i,j})_{j\neq j_0}$, $\boldsymbol{\pi}_i := (\pi_{i,j})_{j\neq j_0}$ and $\boldsymbol{y}_i := (y_{i,j})_{j\neq j_0}$ such that $y_{i,j} = 1$ if the alternative j is chosen by the individual i and $y_{i,j} = 0$ otherwise. The parameter $\boldsymbol{\theta}_i$ is the natural parameter of the multinomial distribution, viewed as member of the exponential family. In this framework , it can be shown that $\theta_{i,j} = \ln(\pi_{i,j}/\pi_{i,j_0})$. Now, using the equation (5) we obtain $\theta_{i,j} = \ln F_\nu(\eta_{i,j}) - \ln\{1 - F\nu(\eta_{i,j})\}$ and thus

$$\frac{\partial\boldsymbol{\theta}_i}{\partial\boldsymbol{\eta}_i} = D_i = \mathrm{diag}_{j\neq j_0}\left[\frac{f_\nu(\eta_{i,j})}{F_\nu(\eta_{i,j})\{1 - F_\nu(\eta_{i,j})\}}\right],$$

where f_ν denotes the probability density function (pdf) of the Student distribution with ν degree of freedom, i.e.,

$$f_\nu(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)\left(1 + \frac{x^2}{\nu}\right)^{(\nu+1)/2}}.$$

Following the same way, the Fisher's information matrix is given by

$$\mathrm{E}\left(\frac{\partial^2 l}{\partial\boldsymbol{\beta}^t\partial\boldsymbol{\beta}}\right) = -\sum_{i=1}^{n} Z_i^t \ D_i \ \mathrm{Cov}(Y_i) \ D_i \ Z_i,$$

with $\mathrm{Cov}(Y_i) = \mathrm{diag}(\boldsymbol{\pi}_i) - \boldsymbol{\pi}_i\boldsymbol{\pi}_i^t$. It could be remarked that this algorithm is simplified if the logistic cdf F is used (instead of the Student cdf F_ν) since $f = F(1-F)$ and therefore D_i becomes the identity matrix of size $J-1$.

4.1.2. Robustness of estimator

An usual way to study the robustness of the estimator is to study the influence function [Hampel, 1974]. It measures the influence of a new observation on a M estimator. According to Künsch et al. [1989], the influence function of a

new observation $(\boldsymbol{y}^*, \boldsymbol{\omega}^*, \boldsymbol{x}^*)$ on the MLE $\boldsymbol{\beta}$ of a GLM, is given by

$$\mathrm{IF}\left\{(\boldsymbol{y}^*, \boldsymbol{x}^*), \hat{\boldsymbol{\beta}}\right\} = \left\{ \mathrm{E}\left(\frac{\partial^2 l}{\partial \boldsymbol{\beta}^t \partial \boldsymbol{\beta}} \right)_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} \right\}^{-1} \left(\frac{\partial l^*}{\partial \boldsymbol{\beta}} \right)_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}},$$

where l (respectively l^*) denotes the log-likelihood computed on the full dataset $\left\{(\boldsymbol{y}_i, \boldsymbol{\omega}_i, \boldsymbol{x}_i)\right\}_{i=1,\dots,n}$ (respectively on the new observation $(\boldsymbol{y}^*, \boldsymbol{\omega}^*, \boldsymbol{x}^*)$). Since the left factor on the right hand side is not depending on the new observation, it is sufficient to show that the score vector $\partial l / \partial \boldsymbol{\beta}$ is bounded according to $(\boldsymbol{y}, \boldsymbol{\omega}, \boldsymbol{x}) \in \{0,1\}^{J-1} \times \mathbb{R}^{qJ+p}$. Without loss of generality, focus is made on one coordinate of this vector, i.e., for $k \in \{1, \dots, p\}$ focus is made on quantity

$$\sup_{y_j \in \{0,1\}, \boldsymbol{\omega} \in \mathbb{R}^{qJ}, \boldsymbol{x} \in \mathbb{R}^p} x_k \frac{f_\nu(\eta_j)}{F_\nu(\eta_j)\{1 - F_\nu(\eta_j)\}} (y_j - \pi_j).$$

The right part $(y_j - \pi_j)$ lies in $(-1, 1)$. The parameter $\boldsymbol{\beta}$ being constant, we can consider that x_k and η_j are equivalent. Focus is thus made on the superior bound on $\mathbb{R}$ of the function

$$\frac{\eta f_\nu(\eta)}{F_\nu(\eta)\{1 - F_\nu(\eta)\}}. \tag{6}$$

This function being continuous on $\mathbb{R}$, we only study limit when $\eta \to -\infty$ or $\eta \to +\infty$.

$$\lim_{\eta \to -\infty} \frac{\eta f_\nu(\eta)}{F_\nu(\eta)\{1 - F_\nu(\eta)\}} = \lim_{\eta \to -\infty} \frac{\eta f_\nu(\eta)}{F_\nu(\eta)}.$$

As noted by Liu [2004], we have

$$\lim_{\eta \to -\infty} \frac{\eta f_\nu(\eta)}{F_\nu(\eta)} = 1 - \lim_{\eta \to -\infty} \frac{(1+\nu)\eta^2}{\nu + \eta^2} = -\nu.$$

The other limit is given by

$$\lim_{\eta \to +\infty} \frac{\eta f_\nu(\eta)}{F_\nu(\eta)\{1 - F_\nu(\eta)\}} = \lim_{\eta \to +\infty} \frac{\eta f_\nu(\eta)}{1 - F_\nu(\eta)}.$$

By symmetry of the Student distribution we have

$$\lim_{\eta \to +\infty} \frac{\eta f_\nu(\eta)}{1 - F_\nu(\eta)} = \lim_{\eta \to +\infty} \frac{\eta f_\nu(-\eta)}{F_\nu(-\eta)} = \nu.$$

The influence function of a (reference, F, Z) model is therefore bounded if F is the Student cdf. On the contrary it is unbounded when F is the logistic cdf since

$$\frac{\eta f(\eta)}{F(\eta)\{1 - F(\eta)\}} = \eta,$$

or if F is the normal cdf since

$$\lim_{\eta \to -\infty} \frac{\eta f(\eta)}{F(\eta)\{1 - F(\eta)\}} = \lim_{\eta \to -\infty} \eta^2.$$

Let us remark that the influence function for Student link function is bounded since the degree of freedom ν is fixed. If ν tends to infinity then the influence function is not bounded anymore. In fact, it turns out to be the case of the normal link function. The three different behaviours of function 6 according to the three cdfs (Student, logistic and normal) are well represented in Figure 1.

4.2. Unknown alternative reference j_0 and degree of freedom ν

The Fisher's scoring algorithm for a (reference, F_ν, $Z)_{j_0}$ model has been detailed for a given link function, i.e., for a given reference alternative j_0 and degree of freedom ν. The link function is here selected, by simply comparing the log-likelihood of models for all values of j_0 and ν. First assume that the reference alternative is fixed. The aim is to select ν among $(0, +\infty)$. Let us note that Student distributions with different degrees of freedom F_{ν_1} and F_{ν_2} are not connected by a linear transformation and thus lead to different models (reference, F_{ν_1}, Z) and (reference, F_{ν_2}, Z), i.e, different likelihood maxima. That is why we have to estimate ν contrarily to the location or scale parameter of the cdf. In practice a finite discretization of ν values is proposed and all corresponding models are compared using the log-likelihood since the same parametrization is used (i.e., the same design matrix Z).

5. Application

We use the benchmark dataset of travel modes between Sydney and Melbourn used by Louviere, Hensher, and Swait [2000], Greene [2003]. The dataset contains 210 observations of choice among $J = 4$ travel modes: *air*, *bus*, *car* and *train*. The two considered alternative specific attributes were the general cost (GC) denoted by $c = (c_{\text{air}}, c_{\text{bus}}, c_{\text{car}}, c_{\text{train}})$ and the terminal waiting time (TTime) denoted by $t = (t_{\text{air}}, t_{\text{bus}}, t_{\text{car}}, t_{\text{train}})$. The two considered individual attributes were the household income (Hinc) denoted by h and the number of people travelling (PSize) denoted by p.

5.1. Model selection

Let us first remark that the design matrix Z is depending on the reference alternative $Z = Z_{j_0}$. The linear predictors are given by

$$\eta_j = \alpha_j + (t_j - t_{j_0})\gamma^1 + (c_j - c_{j_0})\gamma^2 + (h\delta^1_{\text{air}} + p\delta^2_{\text{air}})\mathbf{1}_{(j=\text{air})} \tag{7}$$

for all $j \neq j_0$. This is the design proposed by Louviere et al. [2000] page 157. Using the bus as the reference alternative, for instance, the design matrix is given by

$$Z_{\text{bus}} = \begin{pmatrix} 1 & 0 & 0 & t_{\text{air}} - t_{\text{bus}} & c_{\text{air}} - c_{\text{bus}} & h & p \\ 0 & 1 & 0 & t_{\text{car}} - t_{\text{bus}} & c_{\text{car}} - c_{\text{bus}} & 0 & 0 \\ 0 & 0 & 1 & t_{\text{train}} - t_{\text{bus}} & c_{\text{train}} - c_{\text{bus}} & 0 & 0 \end{pmatrix}.$$

Note that, using the car as reference alternative, gives the design matrix

$$Z_{\text{car}} = \begin{pmatrix} 1 & 0 & 0 & t_{\text{air}} & c_{\text{air}} - c_{\text{car}} & h & p \\ 0 & 1 & 0 & t_{\text{bus}} & c_{\text{bus}} - c_{\text{car}} & 0 & 0 \\ 0 & 0 & 1 & t_{\text{train}} & c_{\text{train}} - c_{\text{car}} & 0 & 0 \end{pmatrix},$$

since the terminal time of the car is null $t_{\text{car}} = 0$. In the following we will denote (reference, F, $Z)_{j_0}$ instead of (reference, F, $Z_{j_0})_{j_0}$ to simplify the notations.

Selection of the link function For each reference alternative $j_0 \in \{\text{air, bus, car, train}\}$, the cdf F was selected among the family of Student cdfs $(F_\nu)_{\nu>0}$. In practice, the values of ν was discretized into a fine grid between 0 and 2 (scale of 0.05) and coarse grid between 2 and 20 (unit scale). The result was compared with the (reference, logistic, $Z)_{j_0}$ model, i.e., the MNL, and the (reference, normal, $Z)_{j_0}$ model. The results are represented in Figure 2, corresponding to the four reference alternatives. As expected we can check that:

- the log-likelihood of the (reference, F_ν, $Z)_{j_0}$ model converges towards those of the (reference, normal, $Z)_{j_0}$ model when $\nu \to +\infty$, in the four situations $j_0 \in \{\text{air, bus, car, train}\}$,
- the log-likelihood of the (reference, F_ν, $Z)_{j_0}$ model equals those of the MNL for values of ν around 8, in the four situations $j_0 \in \{\text{air, bus, car, train}\}$,
- the MNL gives the same result, in the four situations $j_0 \in \{\text{air, bus, car, train}\}$ (invariance under permutation),
- the best (reference, F_{ν^*}, $Z)_{j_0}$ models outperforms the MNL in the four situations $j_0 \in \{\text{air, bus, car, train}\}$.

The AIC obtained with the MNL is 385.83 as obtained by Louviere et al. [2000] page 157 (the log-likelihood is equal to −185.91). The AICs obtained with the (reference, F_{ν^*}, $Z)_{j_0}$ models were 387.3, 383.58, 300, 382.98 respectively with the four reference alternatives j_0 =air, j_0 =bus, j_0 =car, j_0 =train and corresponding degree of freedom $\nu^* = 3$, $\nu^* = 20$, $\nu^* = 0.2$, $\nu^* = 1.35$. The results are clearly in favour of the reference alternative j_0 =car since the gain in AIC is 85.83 compared to MNL results, i.e., 22% of the AIC. It is a very huge difference compared to the results given in the literature [Louviere et al., 2000, Greene, 2003], obtained with MNL and also with NL. Note that AIC is used here instead of log-likelihood because the additional parameter ν has to be take into account in the comparison between Student and logistic link functions.

Selection of relevant attributes After the selection of the link function (reference alternative and degree of freedom), focus is made on parameter estimations and variables selection. Before to compare parameters of different models, a first step is to choose a normalized space $\mathfrak{F}$ of cdfs, i.e., with fixed location and scale values. The normalized space $\mathfrak{F}_{q_{0.95}}$ is proposed as the standard case; see [Bouscasse et al., 2019]. The estimates obtained for the (reference, normal, $Z)_{\text{car}}$ and (reference, logistic, $Z)_{\text{car}}$ models have comparable scales compared to results obtained with the normalized space $\mathfrak{F}_1$; see the Supplementary Material. This step is useful to obtain comparable scale values but not necessary since ratios of parameters stay constant for any normalized space. For instance, the willingness to pay for the terminal time, given by $\text{WTP} = -\gamma^1/\gamma^2$ (see [Bouscasse et al., 2019] for demonstration), stay constant when we change the normalized space (e.g., $F \in \mathfrak{F}_{q_{0.95}}$ or $F \in \mathfrak{F}_1$).

Let us now focus on parameter estimations obtained with (reference, F_ν, $Z)_{\text{car}}$ models. The behaviour of the log-likelihood suggests that the MLE for ν lies in $(0, 1)$; see Figure 2.(C). Estimates of ν are more chaotic near to 0 because of computation instabilities; see Figure 3. The log-likelihood is maximal for $\nu = 0.2$ but parameter estimates are very high, especially for alternative constants; see Table 1. We thus chose the smallest $\nu > 0.2$ such that at least the alternative constant estimates were significant (p-value less than 0.01). Results obtained for the (reference, $F_{0.45}$, $Z)_{\text{car}}$ model are summarized in Table 2. Even if the parameter estimates obtained with $\nu = 0.2$ have a different scale compared to those obtained with $\nu = 0.45$ or the MNL, the ratio of parameters are comparable. For instance, the ratios of alternative constant are nearly identical between the two models: $\alpha_{\text{air}}/\alpha_{\text{bus}}$ and $\alpha_{\text{bus}}/\alpha_{\text{train}}$ are equal to 1.9357 and 0.91843 with $\nu = 0.2$ compared to 2.042 and 0.822 with the MNL; see Table 3 for the MNL results. The two models specially differ by their slope parameters. The terminal time effect is favoured by the Student distribution. The willingness to pay for the terminal time is −146.142 for the (reference, $F_{0.45}$, $Z)_{\text{car}}$ versus −4.263 for the MNL. Moreover, regarding the p-values and their associated test, the three other variables have no significant effect (p-values more than 0.1).

We thus propose to use a more simple design with only the terminal time variable, i.e., we have $\eta_j = \alpha_j + t_j\gamma$ for $j \in \{\text{air, bus, train}\}$. Equivalently the design matrix is given by

$$Z' = \begin{pmatrix} 1 & 0 & 0 & t_{\text{air}} \\ 0 & 1 & 0 & t_{\text{bus}} \\ 0 & 0 & 1 & t_{\text{train}} \end{pmatrix}.$$

Estimates of the (reference, logistic, $Z')_{\text{car}}$ and (reference, $F_{0.45}$, $Z')_{\text{car}}$ models are summarized in Tables 4 and 5. The (reference, logistic, $Z')_{\text{car}}$ model poorly fits the data, obtaining a log-likelihood of −206.82 using only the TTime versus −185.91 using all the variables. The (reference, $F_{0.45}$, $Z')_{\text{car}}$ model obtains log-likelihood of −146.68 versus −145.89 using all variables, confirming the essential role of the terminal time in travel mode choices for this dataset. To confirm these result, several (reference, F_ν, $Z')_{\text{car}}$ models were estimated for $\nu \in (0, 20]$. The corresponding log-likelihoods are represented in Figure 4.(A). Looking at the result obtained near to 0 in Figure 4.(B), we see that more the degree of freedom is small better is the fit. But some problems occur in the standard error estimation; see the estimates of the (reference, $F_{0.05}$, $Z')_{\text{car}}$ model in Table 6. As noticed by Koenker and Yoon [2009], it may be due to poor evaluations of the pdf f_ν and the cdf F_ν when the degree of freedom ν is small, impacting the Fisher's scoring algorithm computation. Nevertheless, the ratios of parameters stay consistent even if estimates are very small; see for instance $\alpha_{\text{air}}/\alpha_{\text{bus}} = 1.943$ and $\alpha_{\text{bus}}/\alpha_{\text{train}} = 0.979$ whereas estimates are less than 10^{-4}.

Finally, using the Student link function, it is concluded that only the terminal time is discriminant in travel mode choice. This conclusion is totally different of those obtained with the MNL, which states that the general cost, the household income and the number of people travelling are discriminant attributes. It should be noted that the NL (see [Hensher and Greene, 2002] pages 169-170 or [Greene, 2003] pages 732-733) and the MNP (see [Greene, 2003] page

735) also selects these attributes as relevant. To go a step further, we will see that these attributes can be considered as noise regarding the travel mode choice.

5.2. Back on the dataset

The variables selection is totally different using the Student distribution instead of the logistic distribution. Using only the terminal time attribute, the model fitting is represented in Figure 5. Indeed, the observed and predicted probability $P(Y = j|Y \in \{j, j_0\})$ are represented according to the terminal time of alternative j, where j_0 is the reference alternative car and j is the air alternative in Figure 5.(A), the bus alternative in Figure 5.(B) and the train alternative in Figure 5.(C). Looking at these graphs, the dataset is very surprising since all observed proportions are either 0 or 1. It means that all individuals, sharing the same terminal time value, take the same decision. In other words, knowing the terminal time value, the travel mode choice is totally deterministic.

To check this surprising fact, the dataset has been reduced by conserving only the columns corresponding to the terminal time of air, bus and train. Then, all individuals being in the same situation, i.e., with the same terminal time values (for air, bus and train), have been aggregated and the weight (number of observations) has been added as a new column. Looking at the three dimensional representation of terminal time values $(t^i_{\text{air}}, t^i_{\text{bus}}, t^i_{\text{train}})_{i=1,\ldots,n=210}$, in Figure 6 (left), we clearly see two groups with the same geometric form of a cross. For a better view, these two groups have been represented separately. Let us focus on the first cross (similar comments can be made on the second cross). This cross has a middle point with coordinates $t_0 = (69, 35, 34)$. All other points are obtained by fixing two coordinates and letting free the last one. It corresponds to a theoretical design that aims to study the effect of terminal time modification for a given alternative when the two other are fixed. Now let see in Figure 6 (right) the travel mode chosen by individuals in each terminal time situation of the first cross. Recall that there is potentially several observations for the same situation (the size of points are represented proportionally to the number of individuals). For each terminal time situation, remark that all individual made the same travel mode choice. For instance, for the middle point $t_0 = (69, 35, 34)$, all the 22 individuals have chosen the car alternative.

Now imagine that this point corresponds to the starting point of stated preference study. Since all the 22 individual chose the car alternative in this situation, the terminal time values $t_0 = (69, 35, 34)$ can be considered as a threshold value. Now let increase the terminal time value only for air alternative. The individual behaviour should take the car again. But the data show that all individuals in situation $t = (t_{\text{air}}, 35, 34)$ with $t_{\text{air}} > 69$ have chosen the air alternative. More generally we observe that all individuals in situation $t = (t_{\text{air}}, 35, 34)$ have chosen the air alternative for any value of $t_{\text{air}} \neq 69$. The same observation is made when modifying the terminal time of bus or train and fixing the two others. This cross design is totally deterministic, each of the three axes corresponding to exactly one travel mode choice (air, bus or train) except the intersection that corresponds to the car choice. The same observation is made for the other cross; see Figure 6 (left) that represents all the dataset. We can conclude that neither the design experiment, or the travel mode choice corresponds to real observations in this dataset. In other words, this is totally improbable to observe these data.

5.3. Sensitivity of the link function to the noise

One can distinguish two kinds of noise: a noise concerning the attributes and another concerning the choice of alternative. The first kind corresponds to the addition of columns (variables) in the dataset and the second to the addition of rows (individuals). The influence function measure the sensitivity of the model to the addition of an outlier, i.e., an observation $(\boldsymbol{y}, \boldsymbol{w}, \boldsymbol{x})$ for which the distance between $\boldsymbol{y}$ and the prediction $\boldsymbol{\pi}$ is high. Such an observation is also called bad leverage point or contamination. It can be considered as a noise in the choice of alternative (response variable in regression). But it is also interesting to study the impact on the model, of a noise in attributes (explanatory variables in regression). It corresponds to the addition of attributes that have no influence on alternative choice, i.e., whose the true parameter is null. One talk about over-fitting when the model is to sensitive to this kind of noise. As previously seen, the benchmark dataset of travel mode choice between Sydney and Melbourne is totally artificial. It is perfect to highlight the sensitivity of the model to these two kind of noise. We propose to compare the sensitivity of the logistic and Student link functions.

5.3.1. Noise in attributes

As previously seen, the alternative choice is totally deterministic knowing the terminal time. It means that the general cost, the household income and the number of people traveling can be considered as noise in attributes. Figure 7 represents the probability of choosing alternative j versus j_0 before and after the addition of theses three non-

informative attributes. The results are here presented for alternative j = air and j_0 = car, but similar results are holding with alternatives bus and train. The first model is estimated, using only the terminal time; see Tables 4 and 5. It corresponds to the equations

$$\frac{\pi_j}{\pi_j + \pi_{j_0}} = F(\hat{\alpha}_j + \hat{\gamma} t_j). \tag{8}$$

This probability is represented for all values of terminal time t between 0 and 100. The second model is estimated, using the terminal time, the general cost, the household income and the number of people traveling; see Tables 2 and 3. It corresponds to the equations

$$\frac{\pi_j}{\pi_j + \pi_{j_0}} = F\left\{\hat{\alpha}_j + \hat{\gamma}^1 t_j + \hat{\gamma}^2 (c_j - c_{j_0}) + \hat{\delta}^1 h + \hat{\delta}^2 p\right\}.$$

This probability is represented for all values of terminal time t between 0 and 100 and for median (and also 25% and 75% percentiles) values of c, h and p. It can be seen that amplitude in prediction error is small for Student link function and high for logistic link function. The Student link function is clearly less sensitive to noise in attributes.

5.3.2. Noise in alternative choice

Now assume that only the terminal time is used as an attribute, i.e., the model is described by equation (8). Here we focus on the conditional choice between bus and car alternatives since it is more representative, but similar results are holding for the two other conditional probabilities. As previously seen, based on the Figure 5 (B), it can be seen that the individual behaviour is to choose the bus alternative while $t < 35$ and the car alternative when $t \geq 35$. Note that the car alternative is chosen for two terminal time values: $t = 35$ and $t = 53$.

The perfect situation is therefore to observe only bus travel choice when $t < 35$ and only car travel choice when $t \geq 35$. These observation are represented by crosses in Figure 5 and are corresponding to a complete separation. According to Albert and Anderson [1984], for logistic link function, there is not a finite MLE but the maximum of likelihood is attained at infinity on the boundary of the parameter space. Therefore, if we stop the Fisher's scoring algorithm when the likelihood is closed to 1, we obtain a perfect classification of these completely separated data. The model estimated on this dataset (complete separation) is represented with a dotted line in Figure 5. The second situation corresponds to the addition of bus observations (resp. air and train) for terminal time values $35 \leq t \leq 53$ (resp. for $64 \leq t \leq 69$ and $34 \leq t \leq 44$). These additional observations are represented by stars in Figure 5 and are corresponding to an overlap situation. According to Albert and Anderson [1984], for logistic link function, the MLE is finite and unique. The model estimated on this completed dataset (overlap) is represented with a dashed line in Figure 5. The third situation corresponds to the addition of bus observations (resp. air and train) for terminal time higher than $t = 53$ (resp. 69 and 44). These additional observations are represented by points in Figure 5 and are corresponding to outliers. This is an overlap situation with outliers. The model estimated on the full dataset (overlap with outliers) is represented with a line in Figure 5. It can be seen that amplitude in prediction error is high for logistic link function (Figure 5 left) and small for Student link function (Figure 5 right). The Student link function is clearly less sensitive to outliers (noise in alternative choice).

6. Simulation study

To test the inference procedure of the degree of freedom, a simulation study is here proposed. The same attributes values than in the benchmark dataset were used. Only the travel mode choice has been simulated according to the (reference, Student$_{\nu^*}$, Z')$_{\text{car}}$ model with a fixed value for ν^*. Only the terminal time attribute was used to simulate the travel mode. Then the three other attributes (CG, Hinc and Psize) were added as noise and taken into account in the estimation procedure. Note that, in order to choose coherent value for $\boldsymbol{\beta}^t = (\alpha_{\text{air}}, \alpha_{\text{bus}}, \alpha_{\text{train}}, \gamma)$ with respect to "real" observations, the model was first estimated using the true degree of freedom ν leading to the MLE $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$. For each fixed value of the true parameter ν^*, ten datasets were simulated. For each simulation, the (reference, Student, Z_{car}) and (reference, logistic, Z_{car}) models were estimated. It means that the slopes of the four attributes (TTime, GC, Hinc and Psize) were estimated. This procedure of estimation/simulation/estimation was repeated with different values for $\nu^* \in \{0.1; 0.5; 0.8; 1; 2; 8\}$. Let see in Figure 8 a plot of two different datasets simulated with the same value $\nu^* = 0.5$. This figure also represents the estimated models (Student with a blue curve and logistic with a dark curve). This plot

only focus on alternatives air and car. The points represents the observed proportions of air alternatives among air and car (the point size is proportional to the number of individual in this situation). Note that these simulated proportions lie in interval [0, 1] whereas the real proportions are either 0 or 1 (see Figure 5(A) of the paper). It means that simulated datasets are more realistic than the true dataset.

6.1. Inference accuracy of $\hat{\nu}$

For each fixed value of the true parameter ν^*, ten datasets have been simulated and so ten value of $\hat{\nu}$ have been estimated. The corresponding box-plots are represented in Figure 9.a) for the six values of $\nu^* \in \{0.1; 0.5; 0.8; 1; 2; 8\}$. It can be seen that, contrarily to what can be expected, more the degree of freedom is small, more the inference accuracy is good. A zoom of boxplots is proposed in Figure 9.b). When the true parameter ν^* is too high (e.g., $\nu^* = 8$) then the MLE $\hat{\nu}$ tends to infinity. In application, the procedure gives the maximum value of the grid, i.e., $\hat{\nu} = 20$; see Figure 9.a). Figure 10 (resp. 11, 12 and 13) represents the log-likelihood obtained with the (reference, F_ν, Z_{car}) model for a grid of ν values between 0 and 20, when $\nu^* = 0.1$ (resp. $\nu^* = 0.5$, $\nu^* = 2$ and $\nu^* = 8$) was used for the simulation. The blue line represents the log-likelihood obtained with (reference, normal=F_∞, Z_{car}) model and the dark line those obtained with the (reference, logistic, Z_{car}) model. These two log-likelihood are very similar whatever the ν^* value used for the simulation. The true value ν^* used for the simulation is represented by a green vertical line. The MLE $\hat{\nu}$ is represented by a red vertical line. When $\nu^* = 8$, regarding the shape of the log-likelihood of Student models, it seems that the MLE $\hat{\nu}$ tends to infinity. The inference accuracy is therefore very bad in this case. But the difference of fitting between the (reference, F_{ν^*}, Z_{car}) and (reference, F_∞, Z_{car}) models in this case is so small that use one model rather than the other does not really matter. Even if $\hat{\nu}$ is very distant from ν^*, the interpretation of attributes effects will be very similar and the prediction of alternative choices to. On the contrary, when ν^* is near to zero, the inference accuracy is really better and the gain in fitting is much higher. It is important since interpretation and prediction will be very different between those obtained by the (reference, $F_{\hat{\nu}}$, Z_{car}) model and the (reference, F_∞, Z_{car}) model in this case.

Finally, remark that one can distinguish two kind of log-likelihood profile according to the order between the blue line and the dark line (normal and logistic). When the blue line is above the dark line, the log-likelihood of Student models is strictly increasing according to the ν value, the MLE $\hat{\nu}$ is very distant from the true value ν^* and there is no gain to use Student instead of normal or logistic distribution. On the contrary, when the blue line is below the dark line, the log-likelihood of Student models seems locally concave around the maximum, the MLE $\hat{\nu}$ is closed to the true value ν^* and more the ν^* value is close to zero more the fit of the Student model is better compared to the fit of normal or logistic models. We can see these two profiles in the "real" application : the first profile in Figure 2.(B) and second profile in Figure 2.(A), (C) and (D).

6.2. Sensitivity to noise

It should be first noted that an additional adjustment of the $\hat{\nu}$ value is necessary when zero is too close because of numerical accuracy problems in standard error estimations. It implies that all p-values are too high and thus cannot be interpreted. This adjustment consists in increasing the ν value until the p-value of intercepts be at least less than 0.01. The true value ν^* is therefore over estimated.

Recall that the true models used for simulations, only contains the TTime as relevant attribute ($\delta > 0$). The three other attributes (GC, Hinc and Psize) are not used for the simulation but added as noise variables for the estimation. In order to compare the sensitivity to this noise of the Student and logistic models, one will count the number of estimated model that only select the TTime as relevant attribute (i.e., the p-value of TTime is less than 0.05 and the p-values of GC, Hinc and Psize are higher than 0.05) among the ten simulations. Results are summarized in Table 7. Student is less sensitive than logistic, especially when the true value ν^* is close to zero.

6.3. Conclusions of the simulation study

Firstly, after a comparison between plots of simulated and real observed proportions of alternatives (Figure 8 and Figure 5), it is clear that, for a given transfer time, the travel mode choice is randomly distributed in simulated dataset whereas it is deterministic in real dataset. There is no longuer any doubt about the total artificial aspect of the benchmark dataset of travel mode choice between Sydney and Melbourne.

Secondly, the inference accuracy of the MLE $\hat{\nu}$ is good for small values of the true parameter (i.e., $\nu^* \in (0, 1]$) and is degrading when the true parameter is higher than 1. Two profiles of log-likelihood have been identified for Student models. In one profile, the MLE $\hat{\nu}$ is reached at infinity and results obtained with Student link function are really close

to those obtained with normal or logistic link functions. The use of Student link instead of logistic has a little interest in this case. On the contrary, in the second profile, the MLE $\hat{\nu}$ is reached close to zero and results obtained with Student link function are really different of those obtained with normal or logistic link functions. The fit is much better and the model is less sensitive to noise in attributes.

7. Discussion

After the introduction of reference models by Peyhardi et al. [2015] and the description of their economic outputs by Bouscasse et al. [2019], this paper studies the robustness of reference models defined with Student link function. More precisely, their influence function is shown to be bounded contrarily to the MNL, and they are thus less sensitive to outliers. It is also empirically shown that Student link functions are less sensitive to over-fitting since non-informative attributes are not selected. The empirical comparison with classical models has been made on a benchmark dataset. It is thus easy to compare our results with several other models, well studied in the literature. To our knowledge, the reference model with Student link function, is the first model that detects the total artificial aspect of this dataset, simply because it is less sensitive to outliers and non-relevant attributes. A limitation of the present paper is the inference procedure that separates the estimation of the link function (alternative reference and degree of freedom) and the selection of relevant attributes through a test. It leads to problem with non-significant p-values for small values of the degree of freedom. It could be improved by using penalized criteria, such as the LASSO for instance.

Furthermore, reference models take several advantages of the GLM framework, such as the simple probability formulation, simple Fisher's scoring algorithm. The GLM framework also eases the addition of random effects in reference models. Finally, extension of reference models to the case of nested alternative has already been introduced by Peyhardi, Trottier, and Guédon [2016]. All these properties make the reference model with Student link function, a good candidate among the huge family of discrete choice models.

References

Adelin Albert and John A Anderson. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 71(1):1–10, 1984.

James H Albert and Siddhartha Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American statistical Association*, 88(422):669–679, 1993.

Hélène Bouscasse, Iragaël Joly, and Jean Peyhardi. A new family of qualitative choice models: An application of reference models to travel mode choice. *Transportation Research Part B: Methodological*, 121:74–91, 2019.

Subodh Dubey, Prateek Bansal, Ricardo A Daziano, and Erick Guerra. A generalized continuous-multinomial response model with a t-distributed error kernel. *Transportation Research Part B: Methodological*, 133:114–141, 2020.

William H. Greene. *Econometric Analysis*. Pearson Education, New York, fifth edition, 2003.

Frank R Hampel. The influence curve and its role in robust estimation. *Journal of the american statistical association*, 69(346):383–393, 1974.

David A Hensher and William H Greene. Specification and estimation of the nested logit model: alternative normalisations. *Transportation Research Part B: Methodological*, 36(1):1–17, 2002.

Roger Koenker and Jungmo Yoon. Parametric links for binary choice models: A fisherian–bayesian colloquy. *Journal of Econometrics*, 152(2): 120–130, 2009.

Hans R Künsch, Leonard A Stefanski, and Raymond J Carroll. Conditionally unbiased bounded-influence estimation in general regression models, with applications to generalized linear models. *Journal of the American Statistical Association*, 84(406):460–466, 1989.

Kenneth L Lange, Roderick JA Little, and Jeremy MG Taylor. Robust statistical modeling using the t distribution. *Journal of the American Statistical Association*, 84(408):881–896, 1989.

Chuanhai Liu. Robit regression: a simple robust alternative to logistic and probit regression. *Applied Bayesian Modeling and Casual Inference from Incomplete-Data Perspectives*, pages 227–238, 2004.

Jordan J Louviere, David A Hensher, and Joffre D Swait. *Stated choice methods: analysis and applications*. Cambridge University Press, 2000.

Jean Peyhardi, Catherine Trottier, and Yann Guédon. A new specification of generalized linear models for categorical responses. *Biometrika*, 102 (4):889–906, 2015.

Jean Peyhardi, Catherine Trottier, and Yann Guédon. Partitioned conditional generalized linear models for categorical responses. *Statistical Modelling*, 2016.

Table 1
Summary results for the (reference, $F_{0.2}$, $Z)_{\text{car}}$ model using the normalized space $\mathfrak{F}_{q_{0.95}}$.

Variables	Estimate	Std. Error	p-value	Test
Air	338.29	233.69	0.015492	*
Bus	174.78	124.53	0.025326	*
Train	190.29	129.36	0.021186	*
GC	−0.41235	0.29453	0.031887	*
TTime	−5.524	3.7516	0.020684	*
Hinc(air)	1.3504	1.2499	0.26482	
PSize(air)	−38.208	27.798	0.055836	·
Loglikelihood	−141.998			
AIC	323.43			
Pseudo R^2	0.49958			

Table 2
Summary results for the (reference, $F_{0.45}$, $Z)_{\text{car}}$ model using the normalized space $\mathfrak{F}_{q_{0.95}}$.

Variables	Estimate	Std. Error	p-value	Test
Air	6.27604	2.21017	0.00384	**
Bus	3.00083	1.00849	0.00134	**
Train	3.04039	1.03155	0.00139	**
GC	−0.00066	0.00246	0.682	
TTime	−0.09698	0.03289	0.00251	**
Hinc(air)	0.00166	0.0081	0.689	
PSize(air)	−0.21431	0.26504	0.433	
Loglikelihood	−145.89			
AIC	307.79			
Pseudo R^2	0.48585			

Table 3
Summary results for the (reference, logistic, $Z)_{\text{car}}$ model.

Variables	Estimate	Std. Error	p-value	Test
Air	7.3347943	0.945712	8.8810^{-15}	***
Bus	3.5916978	0.4754201	4.19610^{-14}	***
Train	4.3719054	0.477643	$< 10^{-16}$	***
GC	−0.0235074	0.005081	3.71810^{-6}	***
TTime	−0.1002126	0.010531	$< 10^{-16}$	***
Hinc(air)	0.0238154	0.011184	0.03322	*
PSize(air)	−1.1738153	0.2580201	5.38210^{-6}	***
Loglikelihood	−185.915			
AIC	385.83			
Pseudo R^2	0.34481			

Table 4
Summary results for the (reference, logistic, $Z')_{\text{car}}$ model.

Variables	Estimate	Std. Error	p-value	Test
Air	5.9486392	0.6667107	$< 10^{-16}$	***
Bus	3.1236823	0.4499032	3.84 $10^{-}12$	***
Train	3.5354086	0.4170212	$< 10^{-16}$	***
TTime	-0.101083	0.0104822	$< 10^{-16}$	***
Log-likelihood	-206.817			
AIC	421.634			
Pseudo R^2	0.27115			

Table 5
Summary results for the (reference, $F_{0.45}$, $Z')_{\text{car}}$ model using the normalized space $\mathfrak{F}_{q_{0.95}}$.

Variables	Estimate	Std. Error	p-value	Test
Air	6.59149	2.29819	0.00299	**
Bus	3.4303	1.23521	0.00425	**
Train	3.43999	1.23142	0.00393	**
TTime	-0.107856	0.037306	0.0026	**
Log-likelihood	-146.68			
AIC	303.37			
Pseudo R^2	0.48307			

Table 6
Summary results for the (reference, $F_{0.05}$, Z') model using the normalized space $\mathfrak{F}_{q_{0.95}}$.

Variables	Estimate	Std. Error	p-value	Test
Air	$2.74\ 10^{-5}$	$8.25\ 10^{-5}$	0.59	
Bus	$1.41\ 10^{-5}$	$4.32\ 10^{-5}$	0.564	
Train	$1.44\ 10^{-5}$	$4.49\ 10^{-5}$	0.561	
TTime	$-5\ 10^{-7}$	$1.4\ 10^{-6}$	0.607	
Log-likelihood	-129.76			
AIC	269.52			
Pseudo R^2	0.54271			

ν^*	0.1	0.5	0.8	1	2	8
Student	6	10	8	9	9	8
logistic	0	4	8	8	9	8

Table 7
Number of estimated model (with Student or logistic link function) that only select the TTime as relevant attribute among the ten simulated dataset and this according to each value of ν^*.

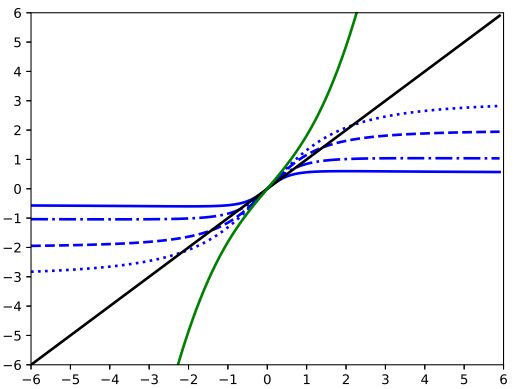

Figure 1: Function (6) with $\nu = 0.5$ (blue line), $\nu = 1$ (blue dashed-dotted line), $\nu = 2$ (blue dashed line), $\nu = 3$ (blue dotted line) and also with the logistic cdf (dark line) and the normal cdf (green line).

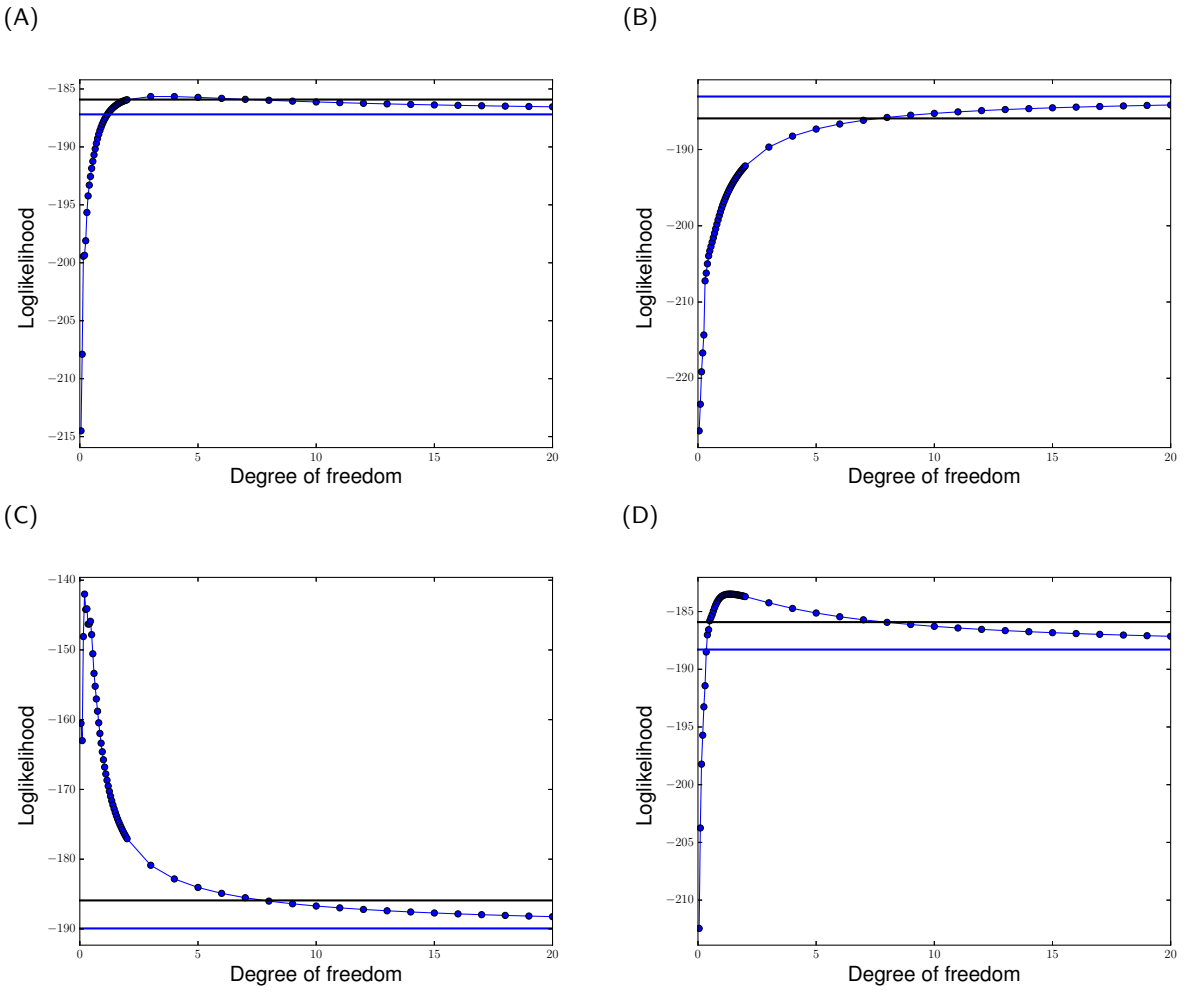

Figure 2: Log-likelihood of the (reference, normal, $Z)_{j_0}$ model (blue line), the (reference, logistic, $Z)_{j_0}$ model (dark line) and the (reference, F_ν, $Z)_{j_0}$ models with $\nu \in (0, 20]$ (blue points) in the four cases: (A) j_0 = air, (B) j_0 = bus, (C) j_0 = car and (D) j_0 = train.

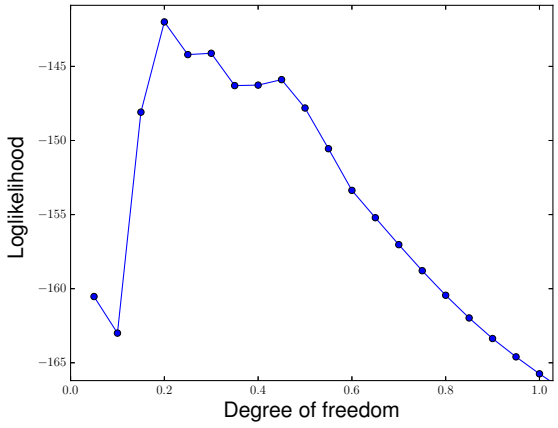

Figure 3: Log-likelihood of the (reference, $\mathcal{T}(\nu)$, $Z_0)_{\text{car}}$ models with $\nu \in \{0.05, 0.1, \ldots, 1\}$.

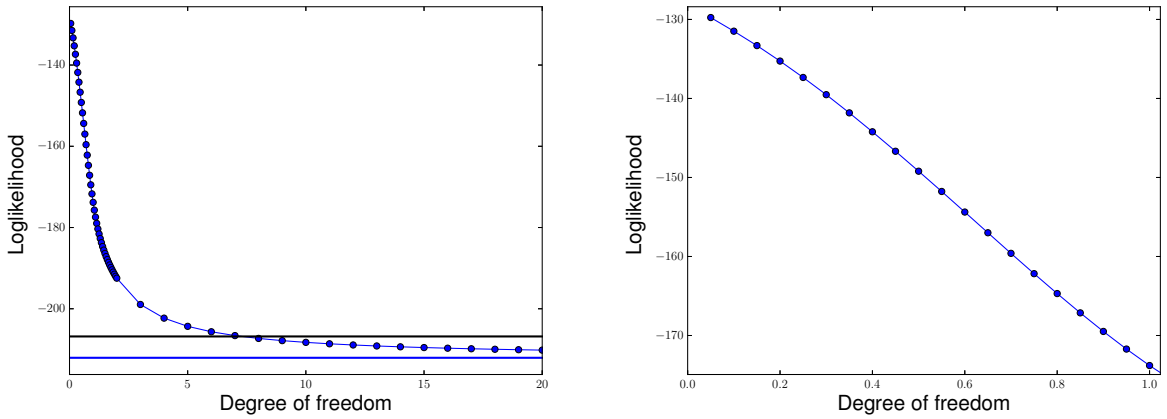

Figure 4: Log-likelihood of (reference, F_ν, $Z')_{\text{car}}$ models for (A) $\nu \in (0, 20]$ (blue points), (reference, normal, $Z')_{\text{car}}$ model (blue line) and (reference, logistic, $Z')_{\text{car}}$ model (dark line) on the real dataset; (B) zooms on curve for $\nu \in (0, 1)$.

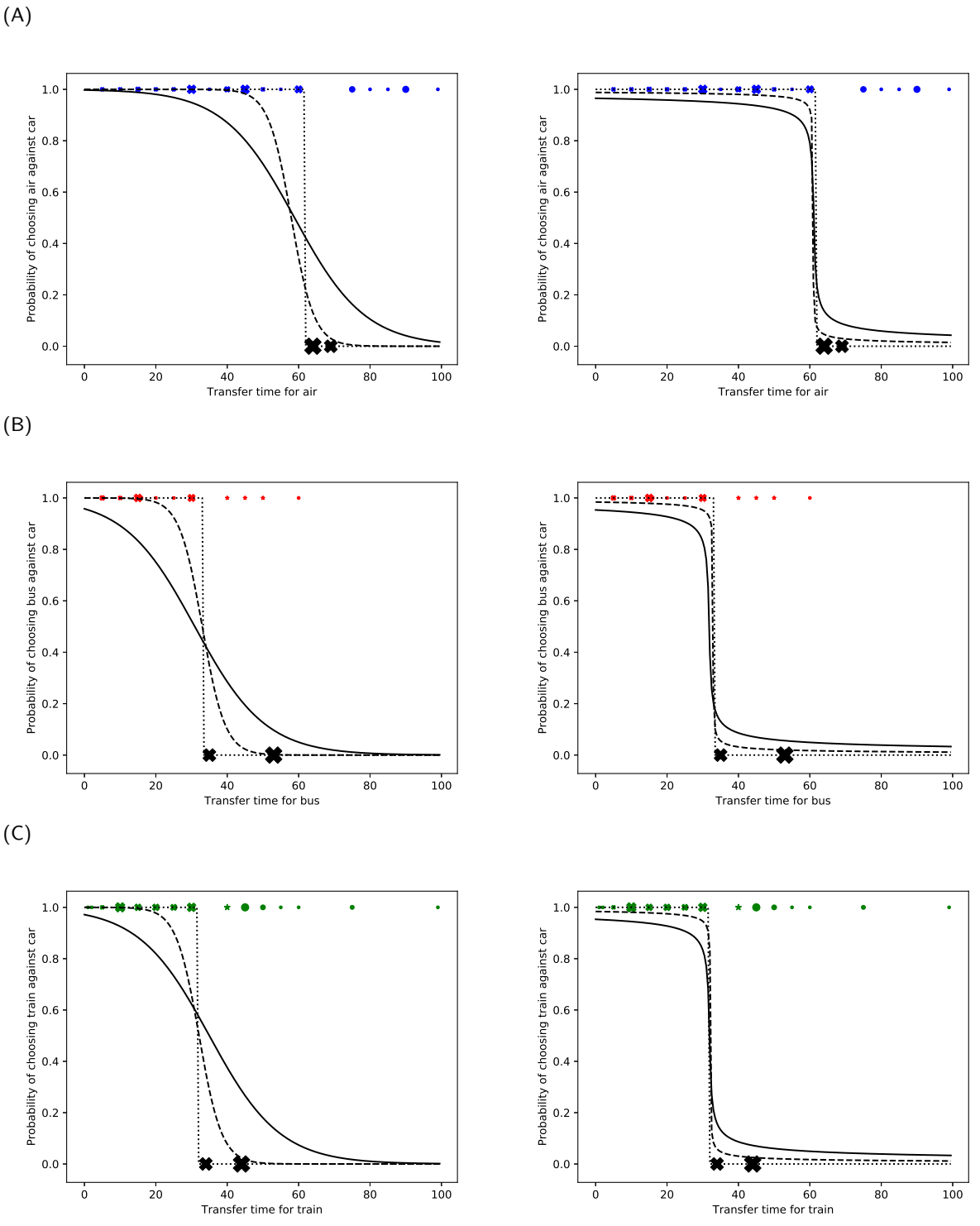

Figure 5: Conditional probabilities $P(Y = j|Y \in \{j, j_0\}; t_j)$ with $j_0 = car$ and (A) j = air (B) j = bus and (C) j = train, according to (reference, F, Z') model where F is the logistic cdf (left) or Student $F_{0.45}$ cdf (right). The corresponding observed proportions are represented by crosses (stars and points), whose the size depends on the number of observations. The color indicates the chosen alternative (blue for air, red for bus, green for train and dark for car).

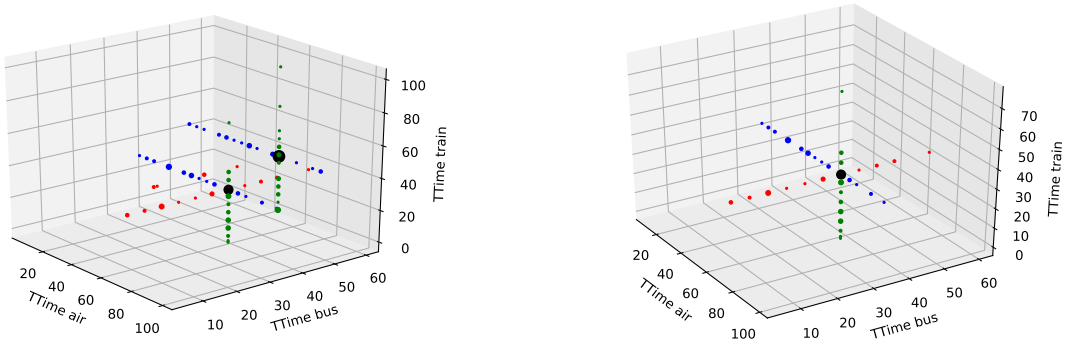

Figure 6: Three dimensional representation of observed terminal time values for A) the whole dataset B) the first group, and their associated travel mode choice (blue for air, red for bus, green for train and black for car). The size of each point is proportional to the number of individuals.

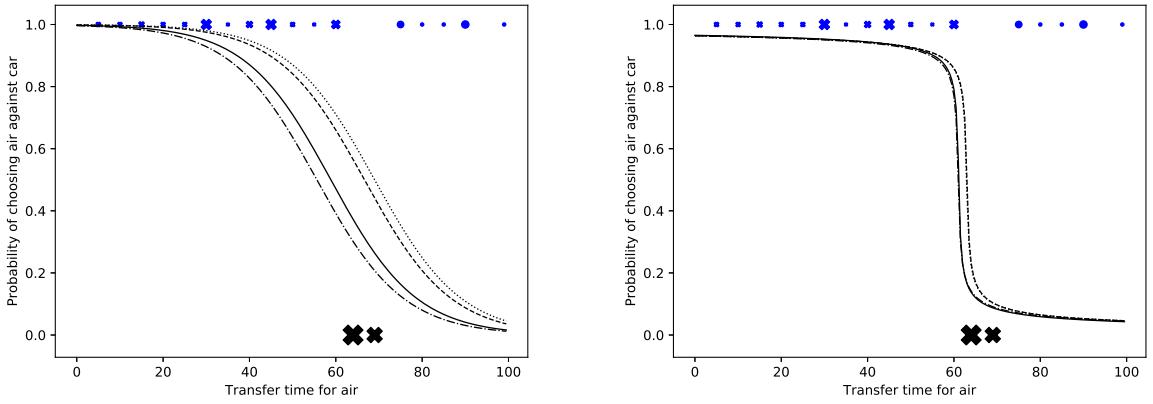

Figure 7: Conditional probabilities $P(Y = j|Y \in \{j, j_0\}; t_j)$ with $j_0 = car$ and $j =$ air, according to (reference, F, Z') model where F is the logistic cdf (left) or Student $F_{0.45}$ cdf (right) and after addition of non informative attributes, with median value (dashed line), 25% and 75% percentiles (resp. dotted and dashed-dotted line). The corresponding observed proportions are represented by crosses (stars and points), whose the size depends on the number of observations. The color indicates the chosen alternative (blue for air, red for bus, green for train and dark for car).

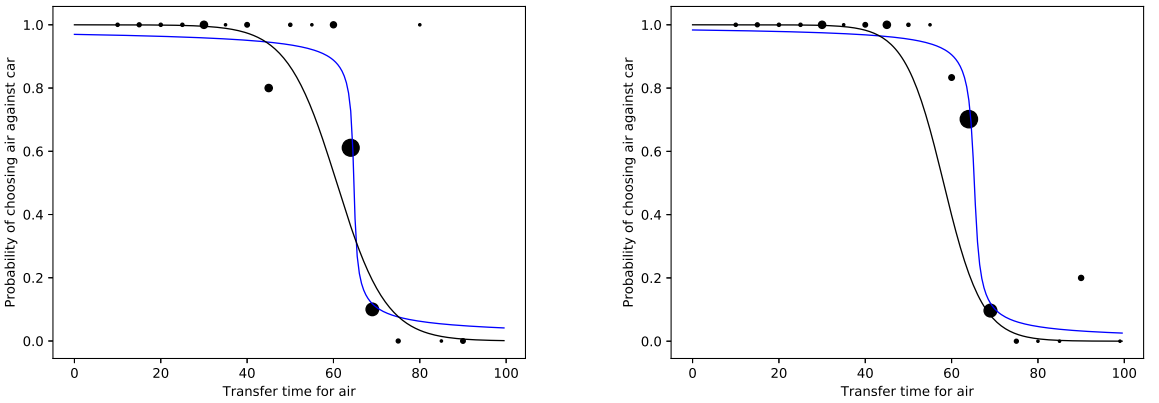

Figure 8: Plot of air proportion (represented by points), among air and car alternatives, according to the transfer time. The blue (resp. dark) curve represents the proportion estimated with the Student (resp. logistic) link function. The two datasets (left and right) was simulated with $\nu^* = 0.5$.

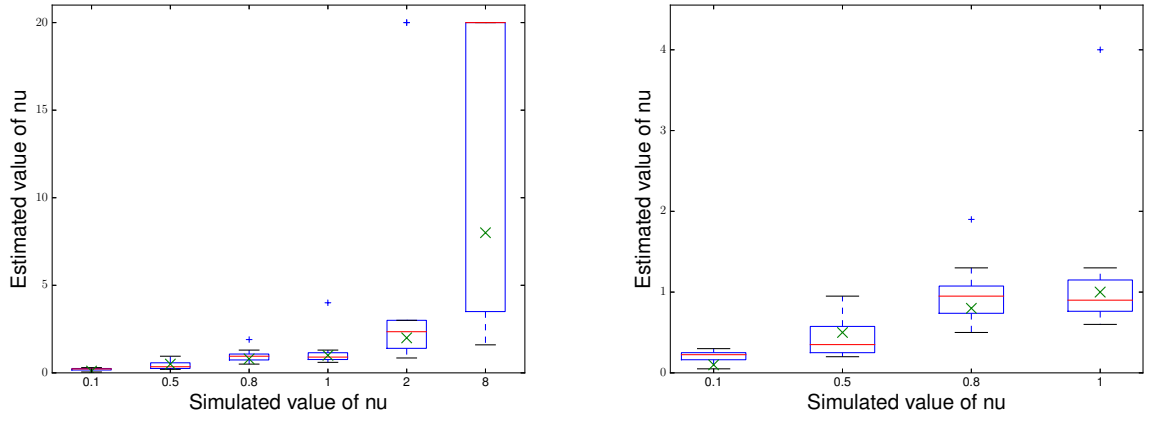

Figure 9: Boxplots of estimated values $\hat{\nu}$ according to a) $\nu^* \in \{0.1, 0.5, 0.8, 1, 2, 8\}$ and b) $\nu^* \in \{0.1, 0.5, 0.8, 1\}$ (zoom). The true value ν^* (used for simulations) is also represented by a green cross. The median of each boxplot is represented by a red line.

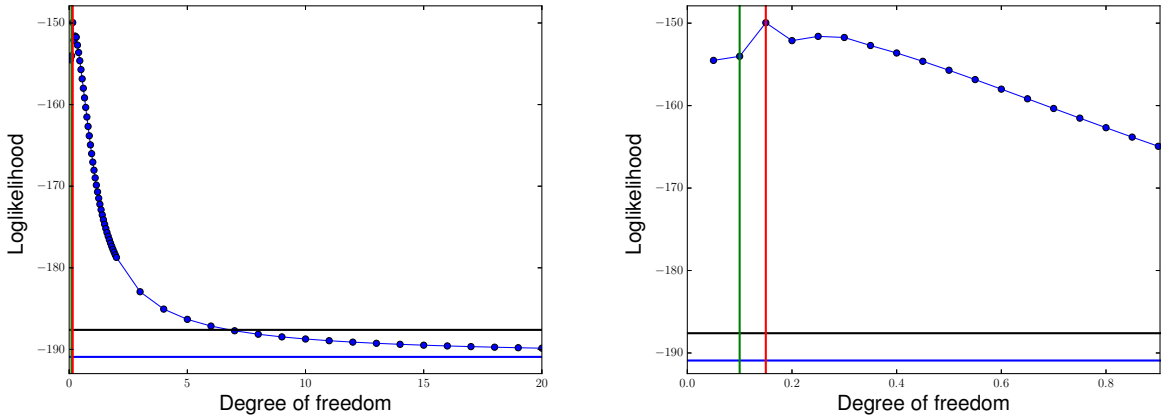

Figure 10: Log-likelihood of the (reference, normal, $Z)_{\text{car}}$ model (blue line), the (reference, logistic, $Z)_{\text{car}}$ model (dark line) and the (reference, F_ν, $Z)_{\text{car}}$ models (blue points) for A) $\nu \in (0, 20]$ and B) $\nu \in \{0.05, 0.1, \dots, 0.8\}$ (zoom) estimated on a simulated dataset. The simulation was made with the (reference, F_{ν^*}, $Z)_{\text{car}}$ model with $\nu^* = 0.1$ (represented by a green vertical line). The MLE $\hat{\nu}$ is represented by a red vertical line.

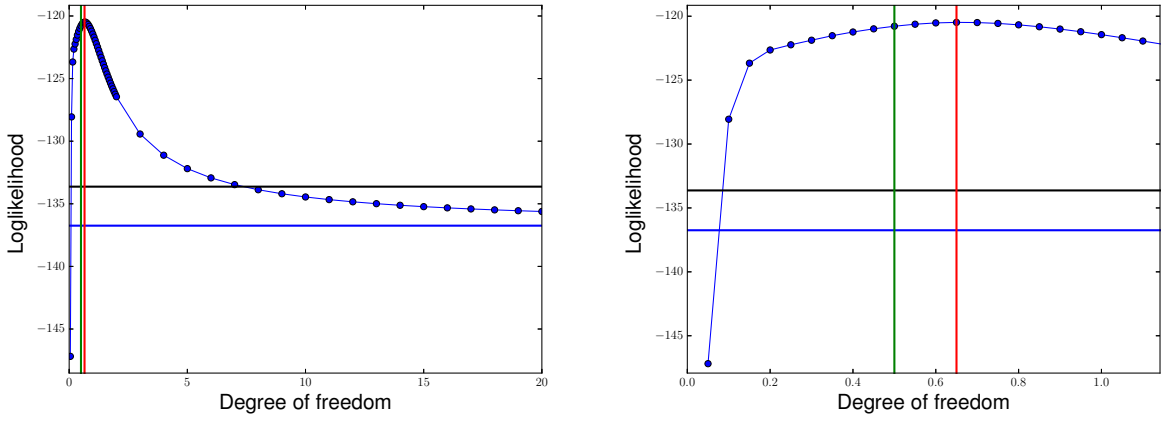

Figure 11: Log-likelihood of the (reference, normal, $Z)_{\text{car}}$ model (blue line), the (reference, logistic, $Z)_{\text{car}}$ model (dark line) and the (reference, F_ν, $Z)_{\text{car}}$ models (blue points) for a) $\nu \in (0, 20]$ and b) $\nu \in \{0.05, 0.1, \dots, 1\}$ (zoom) estimated on a simulated dataset. The simulation was made with the (reference, F_{ν^*}, $Z)_{\text{car}}$ model with $\nu^* = 0.5$ (represented by a green vertical line). The MLE $\hat{\nu}$ is represented by a red vertical line.

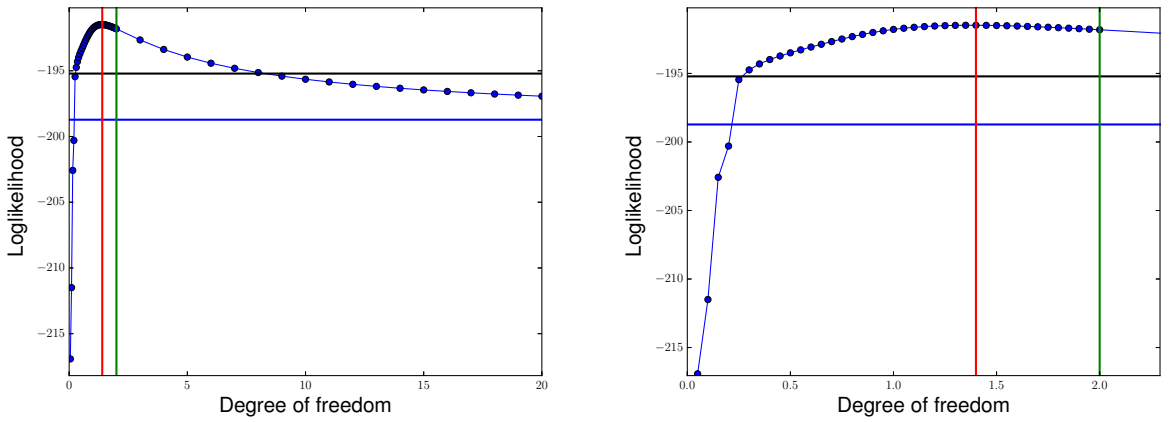

Figure 12: Log-likelihood of the (reference, normal, $Z)_{\text{car}}$ model (blue line), the (reference, logistic, $Z)_{\text{car}}$ model (dark line) and the (reference, F_ν, $Z)_{\text{car}}$ models (blue points) for a) $\nu \in (0, 20]$ and b) $\nu \in \{0.05, 0.1, \ldots, 2\}$ (zoom) estimated on a simulated dataset. The simulation was made with the (reference, F_{ν^*}, $Z)_{\text{car}}$ model with $\nu^* = 2$ (represented by a green vertical line). The MLE $\hat{\nu}$ is represented by a red vertical line.

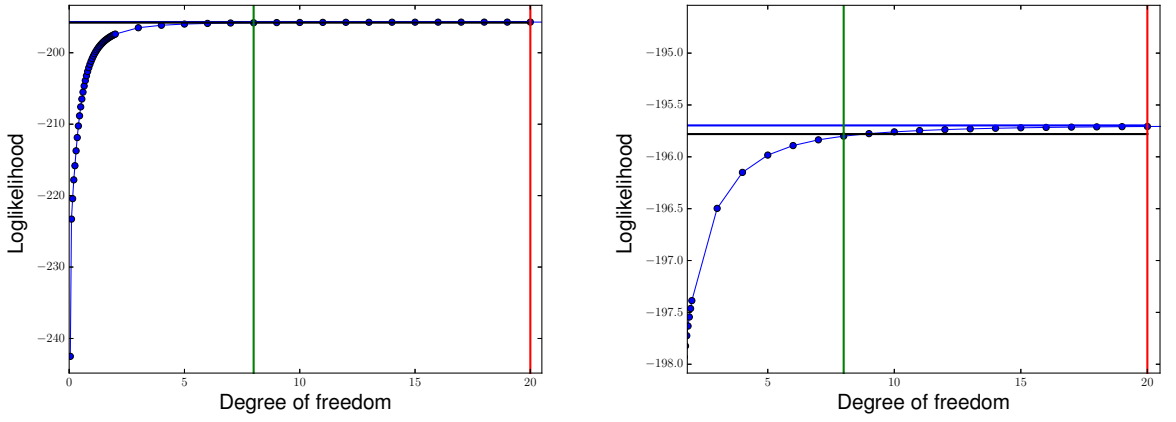

Figure 13: Log-likelihood of the (reference, normal, $Z)_{\text{car}}$ model (blue line), the (reference, logistic, $Z)_{\text{car}}$ model (dark line) and the (reference, F_ν, $Z)_{\text{car}}$ models (blue points) for a) $\nu \in (0, 20]$ and b) $\nu \in \{2, 3 \ldots, 20\}$ (zoom) estimated on a simulated dataset. The simulation was made with the (reference, F_{ν^*}, $Z)_{\text{car}}$ model with $\nu^* = 8$ (represented by a green vertical line). The MLE $\hat{\nu}$ is represented by a red vertical line.