An example of discrete choice modeling in \texttt{GLMcat}

Lorena León$^*$  Jean Peyhardi$^†$  Catherine Trottier$^‡$

In econometrics there is a wide variety of binary models with different functions, but none of them can be extended to the multivariate case because the differences of such random variables are not usually known. The very well known Random Utility models (RUM) don’t have closed-form solutions, thus, for more than four categories ($J > 4$), these models are difficult to estimate. On the other side, Generalized Linear Models (GLM) have an analytic solution, which make it easier the estimation process, and there is a great flexibility at using different cdf functions for the link function.

Following the approach of J. Peyhardi, Trottier, and Guédon (2015) for writing GLMs, qualitative choice models can be written as \textit{(Reference, Logistic, Z)} models. The function \texttt{discrete_cm} in the \texttt{GLMcat} package is available to implement these models.

Dataset

The choice of travel mode of $n = 210$ passengers in Australia was investigated by Louviere et al. (2000), Greene (2003) and Tutz (2011). The alternatives of travel mode are: air, train, bus, and car. As \textit{category – specific} variables they considered the travel time in vehicle (\textit{inv}t) and the general cost (\textit{gc}), and, as the \textit{global – variables} they considered the household income (\textit{hinc}), and the number of people traveling (\textit{psize}).

As example data, \texttt{GLMcat} includes the database: \texttt{TravelChoice}; which we can load as follows:

\begin{verbatim}
# devtools::load_all()
library(GLMcat)
data("TravelChoice")
head(TravelChoice)

## indv mode choice ttme invc invt gc hinc psize
## 1 1 air FALSE 69 59 100 70 35 1
## 2 1 train FALSE 34 31 372 71 35 1
## 3 1 bus FALSE 35 25 417 70 35 1
## 4 1 car TRUE 0 10 180 30 35 1
## 5 2 air FALSE 64 58 68 68 30 2
## 6 2 train FALSE 44 31 354 84 30 2

str(TravelChoice)

## 'data.frame': 840 obs. of 9 variables:
## $ indv : Factor w/ 210 levels "1","2","3","4",...: 1 1 1 1 2 2 2 2 3 3 ...
## $ mode : Factor w/ 4 levels "air","bus","car",...: 1 4 2 3 1 4 2 3 1 4 ...
## $ choice: logi FALSE FALSE FALSE TRUE FALSE FALSE ...
\end{verbatim}

$^*$Université de Montpellier, ylorenaleonv@gmail.com
$^†$Université de Montpellier, jean.peyhardi@umontpellier.fr
$^‡$Université de Montpellier, catherine.trottier@umontpellier.fr
To execute the model proposed by Tutz (2011) (Example 8.4), we execute the `discrete_cm` function with the specific parameters as follows:

```r
exp_8.4 <- discrete_cm(
  formula = choice ~ hinc + gc + invt,
  case_id = "indv",
  alternatives = "mode",
  reference = "air",
  data = TravelChoice,
  alternative_specific = c("gc", "invt"),
  cdf = "logistic")
summary(exp_8.4)
```

According to Tutz (2011), the income seems to be influential for the preference of train and bus over airplane. And, time in vehicle seems to have an impact for the choice of travel mode. Also, cost turns out to be non-influential if income is in the predictor.

To replicate the results of Louviere et al. (2000) we can use the following lines of code. Note that for the variables `hinc` and `psize` the effect is specified only for category `air`.

```r
(constant_model <- discrete_cm(
  formula = choice ~ 1,
  case_id = "indv",
  alternatives = "mode",
  reference = c("air", "train", "bus", "car"),
  data = TravelChoice,
  cdf = "logistic")
))
```

According to Tutz (2011), the income seems to be influential for the preference of train and bus over airplane. And, time in vehicle seems to have an impact for the choice of travel mode. Also, cost turns out to be non-influential if income is in the predictor.
D. J. Peyhardi (2020) demonstrated that the use of the reference category and the choice of the cumulative cdf function, highly affects the fit of the model. Through an experiment in which they used different reference categories as well as different cumulative cdf functions (including Student varying the degrees of freedom) they found that for this case, car as the reference category, and Student(0.2) will result in the best fit.
The results are clearly in favour of the reference alternative $j_0 = \text{car}$ together with $\text{Student}(0.2)$ since the gain in LogLikelihood is 43.92 compared to the multinomial logit model (MNL) results, i.e., 24% of the LogLikelihood. It is a considerable difference compared to the results given in the literature [(Louviere et al. 2000); (Greene 2003)], obtained with MNL and with the nested model.
Conclusion

Until recently, only the logit and probit binary models were extended to the case of multinomial choices, resulting in the multinomial logit and the multinomial probit. The recently introduced family of reference models, defines a multivariate extension of any binary choice model, i.e. for any link function. The GLMcat library through the `discrete_cm` function offers this whole range of models, as demonstrated in the example above.

References


