An example of discrete choice modeling in `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 distribution functions for the link function.

Following the approach of Peyhardi, Trottier, and Guédon (2015) for writing GLMs, qualitative choice models can be written as *(Reference, Logistic, Z)* models. The function `Discrete_CM` in the `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 $category-specific$ variables they considered the travel time in vehicle *(invt)* and the general cost *(gc)*, and, as the $global-variables$ they considered the household income *(hinc)*, and the number of people traveling *(psize)*. As example data, `GLMcat` includes the database: `TravelChoice`; which we can load as follows:

```
# 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 ...
```

[*]Université de Montpellier, ylorenaleonv@gmail.com
[†]Université de Montpellier, jean.peyhardi@umontpellier.fr
[‡]Université de Montpellier, catherine.trottier@umontpellier.fr

```
##  $ ttme  : int  69 34 35 0 64 44 53 0 69 34 ...
##  $ invc  : int  59 31 25 10 58 31 25 11 115 98 ...
##  $ invt  : int  100 372 417 180 68 354 399 255 125 892 ...
##  $ gc    : int  70 71 70 30 68 84 85 50 129 195 ...
##  $ hinc  : int  35 35 35 35 30 30 30 30 40 40 ...
##  $ psize : num  1 1 1 1 2 2 2 2 1 1 ...
```

To execute the model proposed by Tutz (2011) (Example 8.4), we execute the `Discrete_CM` function with the specific parameters as follows:

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

```
##                        Estimate Std. Error z value Pr(>|z|)
## X.Intercept. bus      2.2782935  0.7176686  3.1746 0.001501 **
## X.Intercept. car      1.5334984  0.7065856  2.1703 0.029985 *
## X.Intercept. train    3.5250538  0.6549825  5.3819 7.37e-08 ***
## hinc bus             -0.0355781  0.0131492 -2.7057 0.006816 **
## hinc car             -0.0023652  0.0104475 -0.2264 0.820898
## hinc train           -0.0569415  0.0124103 -4.5882 4.47e-06 ***
## gc                   -0.0016225  0.0055279 -0.2935 0.769128
## invt                 -0.0031266  0.0009548 -3.2746 0.001058 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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*.

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

```
## $`Nb. iterations`
## [1] 4
##
## $coefficients
##                           [,1]
```

```
## (Intercept) air   -0.01709443
## (Intercept) train  0.06559728
## (Intercept) bus   -0.67634006
##
## $LogLikelihood
## [1] -283.7588
##
## $LogLikIter
## [1]    0.0000 -291.1218 -283.9549 -283.7591 -283.7588 -283.7588
##
## $X_M_i
##      [,1] [,2] [,3]
## [1,]    1    0    0
## [2,]    0    1    0
## [3,]    0    0    1
##
## $stderr
## [1] 0.1849068 0.1811689 0.2242376
##
## attr(,"class")
## [1] "glmcat"
```

```
(car_0 <- Discrete_CM(
  formula = choice ~ hinc[air] + psize[air] + gc + ttme,
  case_id = "indv",
  alternatives = "mode",
  reference = c("air", "train", "bus", "car"),
  alternative_specific = c("gc", "ttme"),
  data = TravelChoice,
  distribution = "logistic"
))
```

```
## $`Nb. iterations`
## [1] 5
##
## $coefficients
##                             [,1]
## X.Intercept. air    7.33480684
## X.Intercept. train  4.37191345
## X.Intercept. bus    3.59170206
## hinc air            0.02381549
## psize air          -1.17381658
## gc                 -0.02350742
## ttme               -0.10021278
##
## $LogLikelihood
## [1] -185.9149
##
## $LogLikIter
## [1]    0.0000 -291.1218 -189.4114 -186.0075 -185.9150 -185.9149 -185.9149
##
## $X_M_i
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7]
## [1,]    1    0    0   50    3  -27   64
```

```
## [2,]    0    1    0    0    0   67   44
## [3,]    0    0    1    0    0   45   40
##
## $stderr
## [1] 0.946436165 0.478124444 0.475770576 0.011189102 0.258133127 0.005083637
## [7] 0.010542861
##
## attr(,"class")
## [1] "glmcat"
```

Peyhardi (2020) demonstrated that the use of the reference category and the choice of the cumulative distribution function, highly affects the fit of the model. Through an experiment in which they used different reference categories as well as different cumulative distribution 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.

```r
mod_1 <- Discrete_CM(
  formula = choice ~ hinc[air] + psize[air] + gc + ttme,
  case_id = "indv",
  alternatives = "mode",
  reference = "air",
  alternative_specific = c("gc", "ttme"),
  data = TravelChoice,
  distribution = "logistic"
)
logLik(mod_1)
```

```
## 'log Lik.' -185.9149 (df=NULL)
```

```r
mod_2 <- Discrete_CM(
  formula = choice ~ hinc[air] + psize[air] + gc + ttme,
  case_id = "indv",
  alternatives = "mode",
  reference = "bus",
  alternative_specific = c("gc", "ttme"),
  data = TravelChoice,
  distribution = "student",
  freedom_degrees = 30
)
logLik(mod_2)
```

```
## 'log Lik.' -183.7852 (df=NULL)
```

```r
mod_3 <- Discrete_CM(
  formula = choice ~ hinc[air] + psize[air] + gc + ttme,
  case_id = "indv",
  alternatives = "mode",
  reference = "car",
  alternative_specific = c("gc", "ttme"),
  data = TravelChoice,
  distribution = "student",
  freedom_degrees = 0.2
```

```r
)
logLik(mod_3)
```

```
## 'log Lik.' -141.9978 (df=NULL)
```

```r
mod_4 <- Discrete_CM(
  formula = choice ~ hinc[air] + psize[air] + gc + ttme,
  case_id = "indv",
  alternatives = "mode",
  reference = "train",
  alternative_specific = c("gc", "ttme"),
  data = TravelChoice,
  distribution = "student",
  freedom_degrees = 1.35
)
logLik(mod_4)
```

```
## 'log Lik.' -183.4886 (df=NULL)
```

The results are clearly in favour of the reference alternative $j_0 = car$ together with *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

Greene, W. H. 2003. *Econometric Analysis*. Pearson Education.

Louviere, Jordan J., David A. Hensher, Joffre D. Swait, and Wiktor Adamowicz. 2000. *Stated Choice Methods: Analysis and Applications*. Cambridge University Press. https://doi.org/10.1017/CBO9780511753831.

Peyhardi, Dr Jean. 2020. "Robustness of Student Link Function in Multinomial Choice Models." *Journal of Choice Modelling* 36: 100228. https://doi.org/https://doi.org/10.1016/j.jocm.2020.100228.

Peyhardi, J., C. Trottier, and Y. Guédon. 2015. "A new specification of generalized linear models for categorical responses." *Biometrika* 102 (4): 889–906. https://doi.org/10.1093/biomet/asv042.

Tutz, Gerhard. 2011. *Regression for Categorical Data*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press. https://doi.org/10.1017/CBO9780511842061.