Postestimation functions for R package mixl

Author(s):
Schmid, Basil ⓘD

mixl: An open-source R package for estimating complex choice models on large datasets

Joseph Molloy [a,*], Felix Becker [a], Basil Schmid [a], Kay W. Axhausen [a]

[a] IVT, ETH Zurich, Switzerland

ARTICLE INFO

ABSTRACT

This paper introduces mixl, a new R package for the estimation of advanced choice models. The estimation of such models typically relies on simulation methods with a large number of random draws to obtain stable results. mixl uses inherent properties of the log-likelihood problem structure to greatly reduce both the memory usage and runtime of the estimation procedure for specific types of mixed multinomial logit models. Functions for prediction and posterior analysis are included. Parallel computing is also supported, with near linear speedups observed on up to 24 cores. mixl is directly accessible from R, available on CRAN. We show that mixl is fast, easy to use, and scales to very large datasets. This paper presents the architecture and performance of the package, details its use, and presents some results using real world data and models.

Postestimation functions for *R* package *mixl*

Basil Schmid

Working Paper 1775   **October 2022**

Institut für Verkehrsplanung und Transportsysteme
Institute for Transport Planning and Systems

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Postestimation functions for *R* package *mixl*

Basil Schmid
IVT
ETH Zürich
`basil.schmid@ivt.baug.ethz.ch`

October 2022

Abstract

This paper describes the (post)estimation framework and folder structure to ensure an efficient workflow when using the *R* package *mixl*. By running the postestimation script, several useful choice model indicators, such as elasticities, prediction accuracy, marginal effects and confusion matrices, are automatically calculated and saved as data sheets within the proposed folder structure.

Keywords

Suggested Citation

Contents

1   Introduction

This working paper introduces the estimation and postestimation framework specific to the *R* package *mixl* (Molloy *et al.*, 2021), a specialized software tool for estimating flexible choice models (Multinomial Logit, Mixed Logit and Hybrid Choice models) on large datasets. Several authors have recently used *mixl* to conduct their discrete choice analyses, including Marra *et al.* (2022), Krauss *et al.* (2022), Schmid *et al.* (2022b), Schmid *et al.* (2022a), Schmid *et al.* (2019), Schmid *et al.* (2021), Dias *et al.* (2022), Gschwendtner and Krauss (2022), Lopez-Carmona and Garcia (2022), Puppateravanit *et al.* (2022) and more. This (post)estimation framework has been developed to ensure an efficient workflow and to automatically provides the most relevant and widely used indicators based on an estimated model, such as elasticities, marginal probability effects, prediction accuracy and predicted market shares (e.g. Train, 2009; Schmid *et al.*, 2022a).

All information on how to install and run *mixl* can be found in Molloy *et al.* (2021) or on **github.com/joemolloy/fast-mixed-mnl/blob/master/vignettes/user-guide.md**. It is highly recommended that you first install the most recent version of *R* and *Rstudio.*

The paper is structured as follows: Section 2 explains the architecture and folder structure of the estimation framework. Section 3 shows how to use the framework to obtain the choice model indicators and describes the relevant postestimation functions. Section 4 summarizes and discusses scope for further developments.

2   (Post)estimation framework

The (post)estimation framework is publicly available and can be downloaded from polybox: **polybox.ethz.ch/index.php/s/ryEVgOi05rkdLw1**. The framework is at the same time highly flexible (it can be easily extended or modified), but also very organized (predefined folder structure with all the output files that include the predefined model name). The *R* code is commented whenever appropriate and guides through the scripts.

2.1   Store your dataset

Store your dataset in the **0_data** folder. One row has to correspond to one choice situation. In cases where the availabilities of alternatives vary, all the attribute values

of the non-available alternatives should be set to a numeric value, –99, –97 or 9999, to ensure that the postestimation functions work properly.

Before you continue, open **Zz_source_code.R** and choose your path settings. You can add paths for multiple machines easily, so *R* automatically recognizes where you are working from. Define the number of cores on these machines (note that *mixl* has linear speedups with increasing number of cores; however, parallel computing only works on UNIX/LINUX operating systems). There are different subfolder options that can be added to the main (parent) path (see also Listing 2).

2.2   Estimation, utility scripts and output files

Create your utility script and place it in the **1_estimation** folder. Note: For the sake of good organization, you have to add **_utilities** at the end of the *R* filename for automatic file detection. Use the guidelines provided in Molloy *et al.* (2021) to create your utility script, or just use one of the examples[1] provided in the **1_estimation** folder to get familiar with the notation.

Open the model script **0_RUN_MODEL.R** and choose your settings at the top (see Listing 1), such as the model name (important for saving the output files), the number of draws (in case of Mixed Logit models), the starting values and more. Note: You can either use zero starting values (**Zz_start_values_zero.R**), manually defined starting values (**Zz_start_values_manual.R**), or estimates from a previous model. Also define the columns containing the availabilities of each alternative (they have to be specified as dummy variables; variables should be named *av_*1 to *av_J*, where *J* is the total number of alternatives, to facilitate the postestimation routines with test/validation datasets; see also Listing 3), and do some final data processing if needed. More information is provided in the *R* script.

---

[1]The examples are modified versions of the models presented in Schmid (2019), Chapter 4.

Listing 1: Example of global model settings in **0_RUN_MODEL.R**.

```
1  source("../Zz_source_code.R") # loads the source code with basic functions and
       settings
2
3  Ndraws = 0 # 0 draws for a MNL, > 0 for a MIXL model
4
5  startingkit = 1 # if 0: Zero starting values; if 1: Starting values defined below
6  manual = 1 # if 1: Define starting values manually in Zz_start_values_manual.R
7
8  modelname <- "1_mnl_pooled" # Defines your model name
9
10 modelfile = paste0(modelname,"_utilities.R")
11 source(modelfile) # loads utility script (1_mnl_pooled_utilities.R)
12
13 # defines the starting values you defined above
14
15 if(startingkit==0){
16     file_startingvalues <- "../Zz_start_values_zero.R"
17 }else{
18   if(manual==0){
19     file_startingvalues <- "1_mnl_pooled__est.Rdata"
20     }else{
21     file_startingvalues <- "../Zz_start_values_manual.R"
22   }
23 }
```

When executing **0_RUN_MODEL.R**, the output files are stored in the corresponding folders (see Listing 2) as defined in **Zz_source_code.R** and include the model name as defined in **0_RUN_MODEL.R**. This includes the following:

⋄ **___model.txt** output table with the estimation results

⋄ **___est.Rdata** file with the parameter estimates (can be subsequently used as starting values)

⋄ **___texreg.Rdata** file for later use with LATEX.
In the folder **1_estimation/Xx_multitable/**, **multi_tex.R** also allows you to combine different models into one consolidated LATEX table

⋄ **___mod.Rdata** file which includes all model information (and the training dataset) required for postestimation

⋄ **___posteriors.csv** file with the posteriors includes the posterior means of all parameters that were specified as ...**__RND** in the model file/utility script (for more

information, see also Molloy *et al.*, 2021)

Listing 2: Predefined folder/path structure in **Zz__source__code.R**.

```
1  datafolder <- "0_data/"
2  subfolder <- "1_estimation/"
3  texfolder <- "1_estimation/Xx_multitable/"
4  outputfolder <- "2_postestimation/"
5
6  if(grepl("eu",Sys.info()["nodename"]) & Sys.info()["user"] =="maxmuster"){
7    parentpath <- paste0("/cluster/home/maxmuster/r-input/005_estimation/")
8    cores <- 24
9  }else if(Sys.info()["nodename"]=="IVT-THKPD-30"){
10    parentpath <- "I:/Lehre/MSc Messung und Modellierung/2022/005_estimation/"
11    cores <- 1
12 }
13
14 inputpath <- paste0(parentpath,subfolder)
15 datapath <- paste0(parentpath,datafolder)
16 outputpath <- paste0(parentpath,outputfolder)
17 latexpath <- paste0(parentpath,texfolder)
```

3   Postestimation functions

Open **0__POSTESTIMATION__ANALYSIS.R** in the **2__postestimation** folder.
Load the model file **____mod.Rdata**, define the names each choice alternative (no special
characters allowed, ONLY letters and numbers), define the availabilities of each alternative,
if you are working with a test dataset, etc. (see also Listing 3). Run the **run__analysis(...)**
function, which automatically generates the most relevant figures and numbers. The code
is defined in **Zz__postestimation.R** and can easily be modified by the user if necessary.
All the output files are stored in the **2__postestimation** folder, again including the model
name as part of the file name (except **000__alt__choice__frequencies.png**, a plot of the
choice frequencies for each alternative, and **000__choices__per__ID.png**, a plot of the
number of choices for each decision maker, which are model-independent).

Listing 3: Example of global settings in **1_POSTESTIMATION_ANALYSIS.R**.

```
1  source("../Zz_source_code.R") # loads the source code
2  source("../Zz_postestimation.R") # loads the postestimation code
3
4  modelname <- "1_mnl_pooled"
5  load(paste0(outputpath,modelname,"__mod.Rdata")) # loads the model data file
     (1_mnl_pooled__mod.Rdata)
6
7  # number of repetitions for calculation of prediction accuracy and 95% CI
8  bootn <- 50
9
10 # define if working with test (=0)/training (=1) dataset
11 training <- 1
12
13 if(training==1){
14   dat <- model$data
15   availabilities <- model$availabilities # choose appropriate columns for
       availabilities in training dataset
16   head(availabilities) # check
17 } else {
18   dat <- fread(paste0(datafolder,"testdata.csv"),sep = ";") # define the name of your
       test dataset which stored in the datafolder
19   colnames(model$data)[!colnames(model$data)%in%colnames(dat)] # Check, if the same
       columns are in the training- and testdata
20   availabilities <- as.matrix(dat[,c(grep(pattern = "av_",colnames(dat)))]) # choose
       appropriate columns for availabilities in the test dataset
21   head(availabilities) # check
22 }
23
24 # Proper labeling of alternatives (pooled RP/SP mode and route choice alternatives)
25 alternatives <- c("WRP","BRP","CRP","PTRP", # RP mode choice
26                 "WSP","BSP","CPSP","CSSP","PTSP", # SP mode choice
27                 "RCC1","RCC2","RCC3", # SP route choice carsharing
28                 "RCPT1","RCPT2","RCPT3") # SP route choice public transport
29 alternatives_names <- alternatives
```

⋄ **_confusion_econ.csv** sheet with the %-shares of observed and predicted choices (sampling according to the alternative-specific choice probabilities as described in Train (2009) and applied e.g. in Schmid *et al.* (2022a)). As discussed in Train (2009), this measure is more appropriate than the percent of correctly predicted choices according to the highest probabilities (first preference recovery; e.g., Ortúzar and Willumsen, 2011), since it better reproduces the market shares and reflects the

probabilistic nature (uncertainty) of the Logit model (see also Palma *et al.*, 2016). If the choices were to be repeated many times, or observed by many individuals with the same attributes, each alternative would be chosen by a certain fraction. The structure of the output table is: Observed choices (from left to right); predicted choices (from top to bottom)

◇ **__confusion__opt.csv** sheet with the %-shares of observed and predicted choices (first preference recovery as described in Ortúzar and Willumsen (2011) and applied e.g. in Schmid *et al.* (2022a)). The structure of the output table is: Observed choices (from left to right); predicted choices (from top to bottom)

◇ **__hitrates.csv** sheet with the prediction accuracy (in %) using both methods (i.e. sum of diagonal elements in the confusion matrices, where the predicted equals the observed choices)

◇ **__mean__vars.csv** sheet with the mean of each variable in the dataset (can be used e.g. for a partworth analysis, as discussed and applied e.g. in Schmid *et al.* (2022a))

◇ **__change__bin__MPE.csv** sheet with the marginal probability effects (in %-points), only reported for discrete variables (see e.g. Schmid *et al.*, 2022a)[2]. The marginal probability effects show the average %-point change in the choice probabilities if attribute x_k changes (discrete = jump from 0 to 1) to x_{k^*} while keeping all other attributes $x_l \neq x_k$ fixed:

$$MPE_{i,k} = \%\text{-point change in } P_i = \overline{P}_{i,k^*} - \overline{P}_{i,k} \text{ with } \sum_i MPE_{i,k} = 0 \ \forall \ x_k \tag{1}$$

where $\overline{P}_{i,k}$ is the average (simulated in case of a Mixed Logit type model) alternative-specific predicted probability before the change and $\overline{P}_{i,k^*}$ after the change, conditional on the vector of estimated parameters $\widehat{\mathbf{\Omega}}$.

◇ **__change__perc__ELAST.csv** sheet with the arc-elasticities (in %), only reported for continuous variables (see e.g. Schmid *et al.*, 2022a). The elasticities show the relative % change in the choice probabilities if attribute x_k changes by 1% to x_{k^*} while keeping all other attributes $x_l \neq x_k$ fixed:

$$E_{i,k} = \frac{\%\text{-change in } P_i}{\%\text{-change in } x_k} = \frac{\dfrac{\overline{P}_{i,k^*} - \overline{P}_{i,k}}{(\overline{P}_{i,k} + \overline{P}_{i,k^*})/2}}{\dfrac{\overline{x}_{k^*} - \overline{x}_k}{(\overline{x}_k + \overline{x}_{k^*})/2}} \tag{2}$$

where $\overline{P}_{i,k}$ is the average (simulated in case of a Mixed Logit type model) alternative-specific predicted probability before the change and $\overline{P}_{i,k^*}$ after the change, conditional on the vector of estimated parameters $\widehat{\mathbf{\Omega}}$.

---

[2]Note that the code automatically detects binary variables and variables with more than two categories (which are processed as continuous).

◇ **__insig__pars.csv** sheet with a list of all insignificant parameters (p > 0.05)

◇ **__95ci.csv** sheet with the prediction accuracy (PA; in %) and the 95% confidence bounds using **bootn** draws (number can be adjusted in **0__POSTESTIMA-TION__ANALYSIS.R**; see also Listing 3) from the multivariate normal distribution of the estimated parameters and robust covariance matrix as described in Bierlaire (2017). The PA is obtained by simulating how many choices are, on average, predicted correctly. Specifically, we draw **bootn** times from $\mathcal{N}(\widehat{\mathbf{\Omega}}, \widehat{\mathbf{\Sigma}})$, where $\widehat{\mathbf{\Omega}}$ is the vector of estimated parameters and $\widehat{\mathbf{\Sigma}}$ is the robust variance-covariance matrix of a model, to predict the alternative-specific probabilities. In each repetition, we use a probabilistic calculation of the PA by sampling the predicted choices according to the probabilities of each alternative. Finally, the 95% confidence interval is approximated by calculating the 2.5% and 97.5% quantiles of the resulting distribution as a lower and upper bound, respectively (Bierlaire, 2017).

◇ **__loglike__by__ID.csv** sheet with the (panel) log-likelihood of each decision maker

◇ **__worst__probs__total.csv** sheet with the 10 (if required, number can be adjusted in **Zz__postestimation.R**) decision makers (including their IDs) who exhibit the worst average choice probabilities (outlier detection)

Non-generic (i.e. application-dependent) postestimation examples are additionally included in **0__POSTESTIMATION__ANALYSIS.R** after the **run__analysis(...)** function.

◇ calculation of standard errors for a function of estimated coefficients using the delta method (e.g. Daly *et al.*, 2012). A wide range of the most common functions (e.g. sum, subtraction, multiplication, division of two coefficients, etc.) are already included in **Zz__source__code.R** and can be easily extended.

◇ posterior distribution and willingness-to-pay (WTP) analysis to create a consolidated list of means, medians, standard deviations and/or interquartile ranges for different models

◇ partworth analysis to assess the relative importance of each variable in the choice model (see e.g. Schmid *et al.*, 2022a):

$$VI_k = \sum_i \widehat{|\widehat{\beta}|}_{k,i} \cdot \overline{|x|}_{k,i} \tag{3}$$

where the absolute value of the estimated parameter $\widehat{|\beta|}_{k,i}$ is multiplied with the sample mean of the absolute values of the corresponding variable, $\overline{|x|}_{k,i}$, and summed up over all alternatives.

4 Summary

The introduced estimation and postestimation framework ensures a structured and efficient workflow when using the *R* package *mixl*. It allows a user to automatically obtain relevant indicators based on an estimated model without additional coding efforts. While a lot of tedious work is automatically done when using the postestimation functions, the user still has to do manual adjustments of basic inputs, such as e.g. the name of the alternatives or the definition of availabilities. Also, the partworth analysis (variable importance) is not automatically conducted, since it is highly model-specific. Therefore, the framework also consists of several examples that provide a jump-start.

5   References

Bierlaire, M. (2017) Calculating indicators with PythonBiogeme, *Working Paper*, **170517**, Transport and Mobility Laboratory, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne.

Daly, A., S. Hess and G. de Jong (2012) Calculating errors for measures derived from choice modelling estimates, *Transportation Research Part B: Methodological*, **46** (2) 333–341.

Dias, C., M. Abdullah, R. Lovreglio, S. Sachchithanantham, M. Rekatheeban and I. Sathyaprasad (2022) Exploring home-to-school trip mode choices in Kandy, Sri Lanka, *Journal of Transport Geography*, **99**, 103279.

Gschwendtner, C. and K. Krauss (2022) Coupling transport and electricity: How can vehicle-to-grid boost the attractiveness of carsharing?, *Transportation Research Part D: Transport and Environment*, **106**, 103261.

Krauss, K., M. Krail and K. W. Axhausen (2022) What drives the utility of shared transport services for urban travellers? A stated preference survey in German cities, *Travel Behaviour and Society*, **26**, 206–220.

Lopez-Carmona, M. A. and A. P. Garcia (2022) Adaptive cell-based evacuation systems for leader-follower crowd evacuation, *Transportation Research Part C: Emerging Technologies*, **140**, 103699.

Marra, A. D., L. Sun and F. Corman (2022) The impact of COVID-19 pandemic on public transport usage and route choice: Evidences from a long-term tracking study in urban area, *Transport Policy*, **116**, 258–268.

Molloy, J., F. Becker, B. Schmid and K. W. Axhausen (2021) mixl: An open-source R package for estimating complex choice models on large datasets, *Journal of Choice Modelling*, **39**, 100284.

Ortúzar, J. d. D. and L. G. Willumsen (2011) *Modelling Transport*, John Wiley and Sons, West Sussex.

Palma, D., J. de Dios Ortúzar, L. I. Rizzi, C. A. Guevara, G. Casaubon and H. Ma (2016)

Modelling choice when price is a cue for quality: a case study with Chinese consumers, *Journal of Choice Modelling*, **19**, 24–39.

Puppateravanit, C., K. Sano and K. Hatoyama (2022) Attitude-based segmentation of residential self-selection and travel behavior changes affected by COVID-19, *Future Transportation*, **2** (2) 541–566.

Schmid, B. (2019) Connecting time-use, travel and shopping behavior: Results of a multi-stage household survey, Ph.D. Thesis, IVT, ETH Zurich, Zurich.

Schmid, B., F. Aschauer, S. Jokubauskaite, S. Peer, R. Hössinger, R. Gerike, S. R. Jara-Diaz and K. W. Axhausen (2019) A pooled RP/SP mode, route and destination choice model to investigate mode and user-type effects in the value of travel time savings, *Transportation Research Part A: Policy and Practice*, **124**, 262–294.

Schmid, B., F. Becker, J. Molloy, K. W. Axhausen, J. Lüdering, J. Hagen and A. Blome (2022a) Modeling train route decisions during track works, *Journal of Rail Transport Planning & Management*, **22**, 100320.

Schmid, B., J. Molloy, S. Peer, S. Jokubauskaite, F. Aschauer, R. Hössinger, R. Gerike, S. R. Jara-Diaz and K. W. Axhausen (2021) The value of travel time savings and the value of leisure in Zurich: Estimation, decomposition and policy implications, *Transportation Research Part A: Policy and Practice*, **150**, 186–215.

Schmid, B., T. Schatzmann, C. Winkler and K. W. Axhausen (2022b) A two-stage RP/SP survey to estimate the value of travel time in Switzerland: Short-versus long-term choice behavior, *Arbeitsberichte Verkehrs-und Raumplanung*, **1724**.

Train, K. E. (2009) *Discrete Choice Methods with Simulation*, Cambridge University Press, New York.