Chapter 9
Software

Abstract This chapter describes and compares suitable software for the analysis of basic and advanced discrete choice models. Software packages are classified into proprietary and non-proprietary, according to the operating system required and modelling capabilities. Abilities of both selected commercial (Stata, SAS and Latent Gold, e.g.) and open-source packages (Biogeme and R-libraries) are considered. Finally, some user-written estimation packages for Gauss, Matlab, R and Stata are presented.

There are many software packages for statistical computing and data analysis but not so many for the analysis of basic and advanced discrete choice models. The general statistical software packages can be classified into proprietary and non-proprietary (open-source, public-domain, freeware), by the operating system support (Windows, Mac OS, Linux, BSD, Unix, Cloud) or whether they are menu driven or non-menu driven.

The computing capabilities of new technologies and the dramatic increase of users and disciplines in which discrete choice has been used over the last two decades have positively influenced the number of software packages available today. Writing own codes of complex discrete choice models could and still can only be done by experienced users. Given that open-source and freeware concepts are relatively new, historically the commercial packages were very successful in spite of their limitations regarding the possibility for customisation or delays in the incorporation of the latest methodological approaches. The pioneers worth mentioning in this regard are Limdep-Nlogit (2016) and Alogit (2016). Other commercial packages that include more advanced discrete choice models are Stata (2019), SAS (2020) or Latent Gold (2020). All these commercial packages differ with respect to pricing, estimation speed, possibilities of various model options (constraints, covariates), flexibility of data structures (varying number of choice tasks or alternatives per individual), modelling in preference or WTP-space or the availability of other models.

Probably the most prominent examples for open-source packages are Biogeme (Bierlaire 2020) and several libraries in R (R Core Team 2020). Biogeme is an open-source Python package designed for the maximum likelihood estimation of parametric models in general, with a special emphasis on discrete choice models

P. Mariel et al., *Environmental Valuation with Discrete Choice Experiments*,
SpringerBriefs in Economics, https://doi.org/10.1007/978-3-030-62669-3_9

(Bierlaire 2020). There are several versions of Biogeme that have been developed over the years (Gnu, Bison, Python, Pandas). The latest version called Pandas Biogeme is not a standalone executable, but a Python package. The package is written in Python, with the exception of the core calculations of the models written in C++ for the sake of efficiency. The management of the data relies on the Python data analysis library Pandas, which has become the workhorse of data scientists in recent years.

There are several R packages (libraries) available for the estimation for the discrete choice models. The mlogit (Croissant 2013) package belongs to the oldest and it includes only some extensions of MNL such as nested logit or heteroskedastic logit. The mnlogit (Hasan et al. 2016) package provides significant speed improvements over mlogit with very fast computations of the Hessian of the log-likelihood function. Therefore, it is preferable for the estimation of large-scale multiclass classification problems. Another, more flexible package for large-scale models is the mixl (Molloy 2020) package. It reduces markedly both the memory usage and runtime of the estimation allowing for estimation of more complex models such as MXL and HCM. The gmnl (Sarrias and Daziano 2017) package is one of the most complete packages offering estimation of a wide scale of models including MNL, MXL, G-MXL, and the mixed-mixed multinomial logit. It also offers many different post-estimation analysis procedures. Apollo (Hess and Palma 2019) is currently one of the most flexible packages as it allows estimation of a wide range of models and is fully customisable to support many more. Finally, the RSGHB (Dumont et al. 2019) package allows for estimation of MNL, RP- MXL, EC-MXL, LCM and Nested Logit by the use of the Hierarchical Bayesian framework.

In addition, there are many user-written estimation packages for Matlab, R, Gauss, Ox, C and others. These are typically available from researchers' websites or public repositories. MATLAB codes for estimation of a wide variety of discrete choice models can be found at Czajkowski (2020) or Train (2020). Similarily Gauss codes are at Train (2020b). Some STATA codes for numerous choice models and different postestimation analysis are at Hole (2020). Codes for RRM estimation can be found at very comprehensive website created by van Cranenburgh (2020). It includes codes for Pandas Biogeme, Apollo R, Python Biogeme, Bison Biogeme, MATLAB and Latent Gold.

The advantages of user-written estimation packages include the possibility of studying and modifying the code (e.g. to come up with a new specification). Some of them are also much faster and more precise than commercial packages. Finally, even though the simulated maximum likelihood is the preferred estimator of most researchers dealing with discrete choice models, some of the user-written packages offer the possibility of using other estimation frameworks, such as Bayesian framework (Train and Sonnier 2005), Expectation–Maximisation (EM) algorithm (Train 2007), Laplace approximation (Harding and Hausman 2007) or Maximum Approximate Composite Marginal Likelihood (Bhat and Sidharthan 2011). However, the EM algorithm is also used, in combination with Newton–Raphson, in Latent Gold, for example.

Many commercial packages are menu-driven and that make it easy for the user to input data, estimate a model and get some results. They are therefore usually a first step for beginners and will often be sufficient for many practical purposes. One drawback of this is that they are typically less concerned with the quality of the estimation. This could be problematic, particularly with more complex models (e.g. HCM, models in WTP-space, RP-MXL with correlated parameters). Experienced users tend to switch to user-written packages in R or MATLAB. In addition to greater estimation speed (partially through parallel computing) and higher flexibility of the model specification, they offer the possibility to choose from a wide selection of optimisers, investigate convergence criteria, use different strategies for starting values, etc. to make sure that the results are robust. The applied optimiser is an important issue when estimating nonlinear models, determining speed, robustness and precision.

When focusing on just speed and precision, Czajkowski et al. (2018) compare some of the available estimation packages. They found that in their specific setting their MATLAB implementation outperforms other packages, with R being approximately 5–10 times slower, Python Biogeme—approximately 20 times slower, NLOGIT—60 times slower and Stata—over 100 times slower.

Beginners should start with software that offers a user-friendly interface and gain some experience in the estimation of discrete choice models before moving on to more advance settings. The aforementioned packages offer a wide scale of models and practitioners who stick to standard models can pick the most convenient one. Advanced practitioners looking for the newest methodological approaches will probably code their own estimation procedures. For both groups, it is advisable to not only rely on one package but to estimate models in two environments and compare results. This is, of course, less important when MNL models are estimated but becomes more important when more complex models are estimated and the optimisers and starting values used in the estimation process become more influential.

Researchers should always bear in mind that the code for estimating models is a key part of his or her research. McCullough and Vinod (2003, p. 888) state that "Replication is the cornerstone of science. Research that cannot be replicated is not science, and cannot be trusted either as part of the profession's accumulated body of knowledge or as a basis for policy". Apart from replicability, that is repeating an entire study, independently of the original investigator without the use of original data, the reproducibility should be always guaranteed. A reproducibility seems to be an easy requirement to fulfil because it requires that we can take the original data and the computer code and reproduce all of the numerical outcomes from the study. Nevertheless, it is not, because the researchers are not always careful when organising and documenting their research.

So far not many journals publishing papers concerned with environmental valuation require that the data and estimation code are made available to readers. In other sciences, replicability is regarded as a fundamental principle for research and it should also be a top priority for the environmental valuation research agenda. Even if the journals in which we publish do not require the publication of code and data we should use other methods to make them public. This should be done in spite of the difficulties such as the time it takes to make the research reproducible, knowing that code and data are not universally recognised as research products or that there is not a well-established etiquette for working with code written by other researchers. Finally, the markdown concept of using dynamic analysis documents brings together modelling, documentation and publishing which helps to improve the replicability of research findings. Markdown software is available for R (rmarkdown) or Stata (markstat), for example.

References

Alogit (2016) ALOGIT 4.3. ALOGIT Software & Analysis Ltd. www.alogit.com

Bhat CR, Sidharthan R (2011) A simulation evaluation of the maximum approximate composite marginal likelihood (MACML) estimator for mixed multinomial probit models. Transp Res B Methodol 45:940–953. https://doi.org/10.1016/j.trb.2011.04.006

Bierlaire M (2020) Biogeme. https://biogeme.epfl.ch/. Accessed 21 May 2020

Croissant Y (2013) mlogit: Multinomial Logit models. Version 1.0-3.1URL https://CRAN.R-project.org/package=mlogit

Czajkowski M (2020) Models for discrete choice experiments. https://github.com/czaj/dce. Accessed: 21 May 2020

Czajkowski M, Buczyński M, Budziński W (2018) Replicability, simulation error and robustness to non-parametric treatment of preference heterogeneity in discrete choice models. In: The 25'th Ulvön Conference on Environmental Economics. 20.06.2018. Ulvön

Dumont J, Keller J, Carpenter C (2019) RSGHB: functions for hierarchical bayesian estimation: a flexible approach. Version 1.2.2URL https://CRAN.R-project.org/package=RSGHB

Harding MC, Hausman J (2007) Using a Laplace approximation to estimate the random coefficients logit model by nonlinear least squares. Int Econ Rev 48:1311–1328. https://doi.org/10.1111/j.1468-2354.2007.00463.x

Hasan A, Wang Z, Mahani AS (2016) Fast estimation of multinomial logit models: R package mnlogit. J Stat Softw 75:1–24. https://doi.org/10.18637/jss.v075.i03

Hess S, Palma D (2019) Apollo: a flexible, powerful and customisable freeware package for choice model estimation and application—ScienceDirect. J Choice Model 32:100170. https://doi.org/10.1016/j.jocm.2019.100170

Hole AR (2020) Stata modules. https://www.sheffield.ac.uk/economics/people/hole/stata/software.html. Accessed: 21 May 2020

Latent Gold (2020) Statistical Innovations, Arlington, USA. https://www.statisticalinnovations.com/. Accessed 12 June 2020

LIMDEP (2016) LIMDEP, Econometric Software, Inc. https://www.limdep.com/. Accessed 12 June 2020

McCullough BD, Vinod HD (2003) Verifying the solution from a nonlinear solver: a case study. Am Econ Rev 93:873–892. https://doi.org/10.1257/000282803322157133

Molloy J (2020) mixl: simulated maximum likelihood estimation of mixed logit models for large datasets. Version 1.1.2URL https://CRAN.R-project.org/package=mixl

R Core Team (2020) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Sarrias M, Daziano R (2017) Multinomial logit models with continuous and discrete individual heterogeneity in R: the gmnl package. J Stat Softw 79:1–46. https://doi.org/10.18637/jss.v079.i02

SAS (2020) SAS Institute Inc., Cary, NC, USA. https://www.sas.com/. Accessed 12 June 2020

StataCorp (2019) Stata statistical software: Release 16. StataCorp LLC, College Station, TX

Train K (2020a) MATLAB codes. https://eml.berkeley.edu/~train/software.html. Accessed: 21 May 2020

Train K (2020b) GAUSS codes. https://eml.berkeley.edu/~train/software.html. Accessed: 21 May 2020

Train K (2007) A recursive estimator for random coefficient models

Train K, Sonnier G (2005) Mixed logit with bounded distributions of correlated partworths. In: Scarpa R, Alberini A (eds) Springer. The Netherlands, Dordrecht, pp 1–16

van Cranenburgh S (2020) Advanced random regret minimization models. https://www.advancedrrmmodels.com. Accessed: 21 May 2020