Symbolic Regression for Choice Models: An Interpretable Data-Driven Machine Learning for Discovering Utility Specifications

Georges Sfeir*[1], Stephane Hess[2], and Francisco Camara Pereira[3]

[1]Research Fellow, Choice Modelling Centre, Institute for Transport Studies, University of Leeds, United Kingdom
[2]Professor, Choice Modelling Centre, Institute for Transport Studies, University of Leeds, United Kingdom
[3]Professor, Intelligent Transportation Systems, DTU Management, Technical University of Denmark

Short summary

This paper proposes symbolic regression (SR) as an interpretable, data-driven approach to utility specification in discrete choice models. The aim is to address the limitations of traditional manual specification search and the lack of interpretability in common machine learning methods. The methodology formulates the systematic utility difference as an unknown function learned via SR, optimised using a logit negative log-likelihood, ensuring consistency with random utility theory. SR generates a Pareto frontier of candidate specifications, balancing goodness-of-fit and complexity. Using synthetic binary choice data, we evaluate the ability of SR to recover known functional forms under varying search spaces, including missing transformations, irrelevant features, and noise variables. The results show that SR consistently recovers the true specification, reconstructs nonlinear effects endogenously, and remains robust to irrelevant inputs. These findings demonstrate the potential of SR to support systematic, interpretable specification discovery in choice modelling.
Keywords: Choice Modelling; Machine Learning; Model Specification; Symbolic Regression.

1 Introduction

Discrete choice models (DCMs) are widely used across numerous fields for understanding individual choice behaviour (Hess & Daly, 2024; Mariel et al., 2021; Liebe et al., 2023; Buckell et al., 2022). Modellers use these structures to mathematically represent individuals' choices and forecast demand in future scenarios. They often propose a candidate specification, estimate its parameters, examine both goodness-of-fit statistics and theoretical plausibility, and then refine the specification, balancing parsimony, model fit, and behavioural assumptions (cf. Nova et al., 2024; Van Cranenburgh et al., 2022). Even with few variables, the number of possible specifications grows combinatorially (Rodrigues et al., 2024). In practice, modellers explore this vast model space through incremental trial-and-error, guided by intuition, convention, and tractability constraints. As a result, specification search remains a largely manual, iterative, labour-intensive, subjective, and inevitably incomplete process. This introduces substantial misspecification risk, which could lead to biased parameter estimates, lower predictive power, false conclusions, and incorrect forecasts (Bentz & Merunka, 2000; Ortelli et al., 2021). Machine learning (ML) methods have been proposed as a solution to these specification challenges as they can flexibly learn nonlinearities and interactions. However, widely used approaches (e.g. neural networks, tree ensembles) are often criticised as "black boxes" whose outputs are hard to validate behaviourally and economically, as they do not yield explicit utility representations (Sfeir, 2021). Without interpretable and economically consistent expressions, it becomes difficult to derive welfare measures, interpret marginal effects, compute willingness-to-pay, or ensure behavioural plausibility.

This paper proposes symbolic regression (SR) (Cranmer, 2023) as an interpretable machine learning approach to utility specification. SR searches over mathematical expressions to discover functional forms that best describe the data. Unlike neural networks or ensemble methods, it returns explicit algebraic equations. When applied to utility specification, symbolic regression can automatically identify nonlinear transformations, interaction structures, and non-additive components while preserving a transparent and closed-form representation of utility.

2 Methodology

Symbolic regression is a growing class of ML for inferring a Pareto front of interpretable mathematical equations directly from data, historically associated with genetic programming (Makke & Chawla, 2024). More recently, it has gained traction in deep-learning (cf. Udrescu & Tegmark, 2020; Mundhenk et al., 2021; Kamienny et al., 2022) as a data-driven model-discovery approach for deriving tractable equations and, therefore, interpretable ML models. SR has also been applied with promising results across fundamental and applied sciences, including medicine, chemistry, and astrophysics, to name a few (Makke & Chawla, 2024). This paper introduces SR to choice modelling, specifically to the specification process, where it produces a Pareto frontier of candidate equations/utility specifications.

Problem formulation

In this paper, we consider the case of a binary logit model, where an individual $n = 1, ..., N$ chooses between two travel alternatives $i \in \{1, 2\}$. Under the random utility maximisation framework, the utility of alternative i for individual n is:

$$U_{ni} = V_{ni} + \varepsilon_{ni}, \tag{1}$$

where V_{ni} is the systematic (observed) component of the utility and ε_{ni} is an unobserved error term assumed to be independently and identically distributed (i.i.d.) extreme-value. With two alternatives, the specification can be expressed in terms of systematic utility difference:

$$\Delta V_n = V_{n1} - V_{n2}. \tag{2}$$

The probability that individual n chooses alternative 1 is then given by the standard binary logit form:

$$P_{n1} = \frac{1}{1 + \exp(-\Delta V_n)}. \tag{3}$$

In conventional discrete choice modelling, the systematic utility difference ΔV_n is specified a priori as a function of observed attributes of the alternatives and characteristics of individual n (x_n). In this work, we instead treat it as an unknown function to be discovered from the data:

$$\Delta V_n = f(x_n), \tag{4}$$

where $f(x_n)$ belongs to a large space of candidate mathematical expressions. Conventional specification search restricts f to be mainly linear-in-parameters with manually selected transformations and interactions. Symbolic regression instead treats the discovery of f as an optimisation problem:

$$\min_{f \in \mathcal{F}} \mathcal{L}(f), \tag{5}$$

where $\mathcal{F}$ is a space of algebraic expressions built from variables and operators, and $\mathcal{L}$ is a loss function measuring fit to observed choices.
For logit models, the natural objective is the negative log-likelihood. For observations n with choice $y_n \in \{0, 1\}$, the element-wise loss is:

$$\ell\big(f(x_n), y_n\big) = \log\big(1 + \exp\big(f(x_n)\big)\big) - y_n\, f(x_n), \tag{6}$$

which corresponds to the binary logit log-likelihood contribution when $f(x_n) = \Delta V_n$. Loss minimisation therefore yields a function consistent with random utility maximisation.

Pareto-optimal specification search

SR does not return a single equation but a Pareto of expressions trading off goodness-of-fit and complexity. Complexity is typically measured as the number of nodes or operations in the expression/equation. Formally, SR approximates the Pareto set:

$$\mathcal{P} = \left\{ f : \big(\mathcal{L}(f), C(f)\big) \text{ is Pareto-optimal} \right\}, \tag{7}$$

where $C(f)$ denotes complexity.

In the context of utility specification, this corresponds closely to the modeller's traditional trade-off between fit and parsimony. The Pareto front can therefore be interpreted as an automatically generated specification frontier, from which the modeller may select behaviourally meaningful, data-driven candidate utility functions.

3 Results and discussion

Synthetic data

To illustrate the proposed SR approach in a controlled setting where the data-generation process is known, we consider a synthetic binary choice experiment consistent with the formulation introduced in Section 2. Individuals $n = 1, ..., N$ are assumed to face a choice between two alternatives, car ($i = 1$) and public transport ($i = 2$). The synthetic dataset contains 2000 individuals, each with two observations. The true systematic utility difference is specified as follows:

$$\begin{aligned}\Delta V_n = \Delta ASC + \beta_{\Delta time}\,\Delta\text{time}_n + \beta_{\Delta time^2}\,(\Delta\text{time}_n)^2 + \beta_{\Delta cost}\,\Delta\text{cost}_n \\ + \beta_{income}\,\text{income}_n + \beta_{age}\,\text{age}_n.\end{aligned} \tag{8}$$

Choices are simulated by drawing i.i.d. Gumbel errors and applying the logit model described above. The modeller is assumed to observe the four explanatory variables $x_n = (\Delta\text{time}_n, \Delta\text{cost}_n, \text{income}_n, \text{age}_n)$ but not the true functional form of ΔV_n. The task is therefore to recover a behaviourally plausible specification for ΔV_n from the data using SR, without prior knowledge of the true underlying specification.

Symbolic regression setup

We apply symbolic regression using PySR (Cranmer, 2023), which performs a multi-population evolutionary search over mathematical expressions. The search explores combinations of input features and mathematical operators to identify candidate utility specifications. To ensure compatibility with random utility theory, we use the binary logit negative log-likelihood as the SR objective function:

$$\mathcal{L}(f) = \sum_{n=1}^{N} \left[\log\big(1 + \exp(f(x_n))\big) - y_n f(x_n)\right]. \tag{9}$$

This loss directly interprets the symbolic expression as the utility difference ΔV_n. Consequently, any discovered function yields a valid logit model.

To systematically evaluate the ability of symbolic regression to recover and discover utility specifications, we design five experiments with different configurations of the search space (Table 1). These experiments aim to assess (i) recovery of a known specification when the relevant variables are available, (ii) the ability to construct missing functional forms, and (iii) robustness to irrelevant operators, transformations, noise variables, and/or engineered features. In the first experiment, we include the true nonlinear transformation of travel time (Δtime^2) as well as the four explanatory variables x_n in the feature/variable space (i.e. set of featues/variables for SR to consider), and a restricted operator set $\{+, -, \times\}$. This serves as a benchmark setting, where the correct specification is directly accessible to the search, as the search space contains only relevant features and simple algebraic operators. In the second experiment, we remove the squared travel-time term from the feature space while retaining the same restricted operator set. This setting evaluates whether SR can reconstruct the missing nonlinear component indirectly, despite the absence of both explicit feature engineering and nonlinear operators. The third experiment expands the operator set to include nonlinear transformations $\{\log(x), \exp(x), \sqrt{x}\}$, while keeping the original feature space without the squared travel-time term. Although these transformations are not part of the true data-generating process, this setting evaluates whether SR can still recover the correct specification when the search space includes potentially irrelevant operators. In the fourth experiment, we further expand the feature space by including additional squared transformations and interaction terms that are not part of the true specification. This setting tests the robustness of SR to irrelevant or misleading feature engineering. Finally, in the fifth experiment, we add three randomly generated variables $\{r_1, r_2, r_3\}$ to the feature space. This setting evaluates the ability of SR to ignore irrelevant input and to identify behaviourally meaningful structures in the presence of both noise variables and irrelevant operators/transformations. Across all experiments, the SR search is conducted for 100 iterations with a population size of 30 candidate expressions, ensuring a consistent basis for comparing the resulting Pareto frontiers and the recovered utility specifications.

Table 1: Experiments.

Experiment	Feature Space	Binary Operators	Unary Operators
1	Δtime, Δcost, income, age, Δtime^2	$+, -, \times$	
2	Δtime, Δcost, income, age	$+, -, \times$	
3	Δtime, Δcost, income, age	$+, -, \times$	$\log(x), \exp(x), \sqrt{x}$
4	Δtime, Δcost, income, age, Δtime^2, Δcost^2, Δincome^2, Δage^2, $\Delta\text{time}\times\text{income}$, $\Delta\text{time}\times\text{age}$	$+, -, \times$	$\log(x), \exp(x), \sqrt{x}$
5	Δtime, Δcost, income, age, Δtime^2, Δcost^2, Δincome^2, Δage^2, $\Delta\text{time}\times\text{income}$, $\Delta\text{time}\times\text{age}$, r_1, r_2, r_3	$+, -, \times$	$\log(x), \exp(x), \sqrt{x}$

Results

In each experiment, the SR search produced between 10 and 20 sequence of candidate utility specifications of increasing complexity along the Pareto frontier. Across all experimental settings, the set of Pareto-optimal expressions consistently included the true underlying specification. Moreover, the evolution of the Pareto frontier follows a similar pattern across experiments, suggesting a stable search process despite differences in the search space. At low levels of complexity (approximately 1 to 7), the recovered expressions are simple, typically involving a constant term or single explanatory variables, and gradually extending to basic bilinear combinations. As complexity increases (around 7–11), the expressions begin to stabilise into a more structured and interpretable form, with the utility difference approximated as an additive function of key variables such as travel time, travel cost, income, and age. At moderate levels of complexity (around 9-14), symbolic regression introduces a quadratic transformation of travel time, thereby recovering the true underlying specification. Beyond this point, further increases in complexity lead to more elaborate expressions, incorporating additional transformations and interaction terms. While these higher complexity specifications may offer marginal improvements in fit, they do so at the cost of reduced parsimony and interpretability.

Experiments 2 and 3 further demonstrate that SR is able to recover the squared travel time term even when it is not provided explicitly in the feature space. In these settings, the nonlinear effect emerges endogenously from the search process, indicating that SR can reconstruct missing functional forms rather than relying on predefined transformations. Experiments 4 and 5 highlight the robustness of the approach to an expanded and potentially misleading search space. Despite the presence of additional irrelevant transformations, interaction terms, and randomly generated noise variables, SR consistently recovers the true underlying specification. This suggests that the method is able to effectively navigate complex search spaces and suggest behaviourally meaningful structures. For brevity, we only report the equation corresponding to the true data-generating specification recovered by SR in experiment 3:

$$0.230 - 2.26 \times \Delta time - 3.73 \times \Delta time^2 - 0.180 \times \Delta cost + 0.220 \times income - 0.009 \times age. \quad (10)$$

To further assess the consistency of the recovered specification, we estimate a binary logit model in Apollo (Hess & Palma, 2019) using the true specification and compare the resulting parameter estimates with those implied by the SR-derived expression. Table 2 reports the true parameter values used in the data-generating process, the estimates obtained with Apollo, and the coefficients recovered by symbolic regression. The results show a close correspondence between the SR-derived coefficients and those obtained from maximum likelihood estimation in Apollo, with only minor deviations across parameters. This provides additional evidence that symbolic regression not only recovers the correct functional form, but also yields quantitatively consistent parameter estimates. The small differences can be attributed to the fact that SR optimises a joint search over functional form and parameters, rather than performing conditional maximum likelihood estimation for a fixed specification.

Table 2: Results.

Parameter	True Estimate	True Estimate from Apollo	Estimate from SR	% Difference Apollo - SR
ΔASC	0.250	0.244	0.230	5.74%
$\beta_{\Delta time}$	-2.50	-2.27	-2.26	0.50%
$\beta_{\Delta time^2}$	-4.00	-3.83	-3.73	2.51%
$\beta_{\Delta cost}$	-0.200	-0.178	-0.180	-1.14%
β_{income}	0.250	0.221	0.220	0.28%
β_{age}	-0.010	-0.009	-0.009	0.02%

4 Conclusions

This paper introduced symbolic regression as an interpretable, data-driven approach to discovering utility specifications within the logit framework. Using synthetic binary choice data, the results show that SR is able to successfully recover the true underlying specification across a range of experimental settings. In particular, SR was able to reconstruct nonlinear effects endogenously, even when the relevant transformations were not provided a priori, and to recover the correct specification in the presence of irrelevant variables, engineered features, and additional operators. These findings illustrate the potential of symbolic regression to support the specification search process by automatically generating behaviourally interpretable candidate utility functions. Rather than replacing the modeller, the approach is best understood as a complementary tool that can assist in exploring the specification space more systematically, while preserving the transparency and theoretical consistency required in discrete choice modelling. Future work will extend the approach to multinomial logit models, incorporate behavioural constraints such as monotonicity with respect to cost and travel time, and evaluate performance on real datasets.

Acknowledgements

The authors acknowledge the financial support by: 1) the European Research Council through the advanced grant 101020940-SYNERGY; and 2) the Horizon Europe Framework Programme under the Marie Skłodowska-Curie Postdoctoral Fellowship MSCA2021-PF-01 project No 101063801.

References

Bentz, Y., & Merunka, D. (2000). Neural networks and the multinomial logit for brand choice modelling: a hybrid approach. *Journal of Forecasting*, *19*, 177-200. doi: https://doi.org/10.1002/(SICI)1099-131X(200004)19:3<177::AID-FOR738>3.0.CO;2-6

Buckell, J., Vasavada, V., Wordsworth, S., Regier, D. A., & Quaife, M. (2022). Utility maximization versus regret minimization in health choice behavior: Evidence from four datasets. *Health Economics*, *31*(2), 363–381.

Cranmer, M. (2023). *Interpretable machine learning for science with pysr and symbolicregression.jl.* Retrieved from `https://arxiv.org/abs/2305.01582`

Hess, S., & Daly, A. (2024). *Handbook of choice modelling.* Edward Elgar Publishing.

Hess, S., & Palma, D. (2019). Apollo: A flexible, powerful and customisable freeware package for choice model estimation and application. *Journal of choice modelling*, *32*, 100170.

Kamienny, P.-A., d'Ascoli, S., Lample, G., & Charton, F. (2022). End-to-end symbolic regression with transformers. In *Proceedings of the 36th international conference on neural information processing systems.* Red Hook, NY, USA: Curran Associates Inc.

Liebe, U., van Cranenburgh, S., & Chorus, C. (2023). Maximizing utility or avoiding losses? uncovering decision rule-heterogeneity in sociological research with an application to neighbourhood choice. *Sociological Methods & Research.*

Makke, N., & Chawla, S. (2024). Interpretable scientific discovery with symbolic regression: a review. *Artificial Intelligence Review*, *57*, 177-200. doi: https://doi.org/10.1007/s10462-023-10622-0

Mariel, P., Hoyos, D., Meyerhoff, J., Czajkowski, M., Dekker, T., Glenk, K., ... others (2021). *Environmental valuation with discrete choice experiments: Guidance on design, implementation and data analysis*. Springer Nature.

Mundhenk, T. N., Landajuela, M., Glatt, R., Santiago, C. P., Faissol, D. M., & Petersen, B. K. (2021). Symbolic regression via neural-guided genetic programming population seeding. In *Proceedings of the 35th international conference on neural information processing systems*. Red Hook, NY, USA: Curran Associates Inc.

Nova, G., van Cranenburgh, S., & Hess, S. (2024). Understanding the decision-making process of choice modellers. *arXiv preprint arXiv:2411.01704*.

Ortelli, N., Hillel, T., Pereira, F. C., de Lapparent, M., & Bierlaire, M. (2021). Assisted specification of discrete choice models. *Journal of choice modelling*, *39*, 100285.

Rodrigues, F., Krueger, R., & Pereira, F. C. (2024). Model building, inference and interpretation: developing discrete choice models in the age of machine learning. In *Handbook of choice modelling* (pp. 74–115). Edward Elgar Publishing.

Sfeir, G. (2021). *Integration of machine learning and discrete choice models to better predict and describe decision makers' choices with applications to travel decisions* (PhD thesis, American University of Beirut, Beirut, Lebanon). Retrieved from `http://hdl.handle.net/10938/22801`

Udrescu, S.-M., & Tegmark, M. (2020). Ai feynman: A physics-inspired method for symbolic regression. *Artificial Intelligence Review*, *6(16)*. doi: https://doi.org/10.1126/sciadv.aay2631

Van Cranenburgh, S., Wang, S., Vij, A., Pereira, F., & Walker, J. (2022). Choice modelling in the age of machine learning-discussion paper. *Journal of choice modelling*, *42*, 100340.